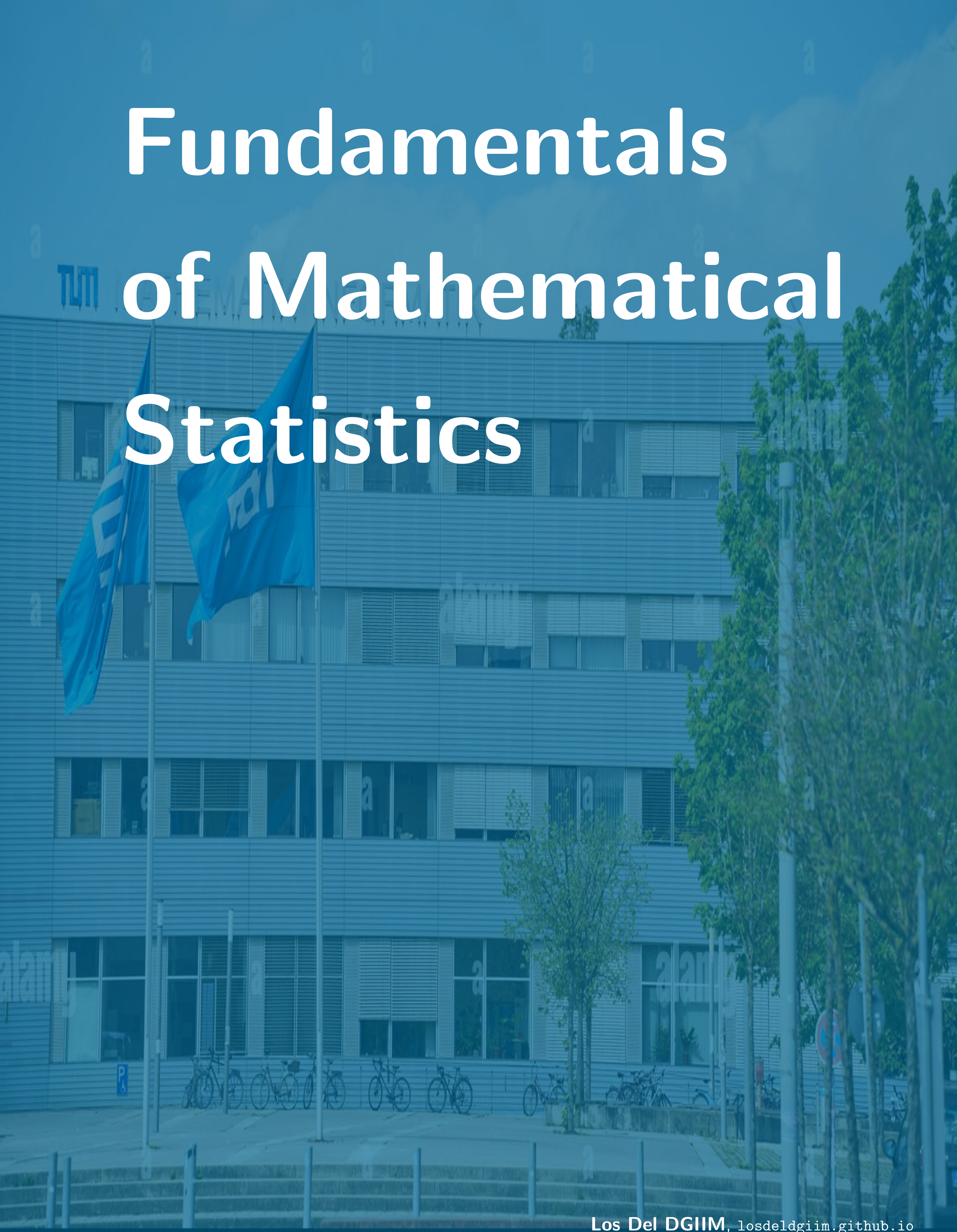
Fundamentals of Mathematical Statistics



Los Del DGIIM, `losdeldgiim.github.io`

Doble Grado en Ingeniería Informática y Matemáticas
Universidad de Granada

Fundamentals of Mathematical Statistics

Los Del DGIIM, losdeldgiim.github.io

Laura Mandow Fuentes

Granada, 2025/26

Contents

Preface

These notes are a personal LaTeX adaptation of the notes and lectures given by Prof. Mathias Drton for the course "*Fundamentals of Mathematical Statistics*". The original material is licensed under the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 license. Any errors or omissions in this version are entirely my own.

1. Construction of Estimators

The goal of estimation is to approximate an unknown characteristic $\gamma(\mathsf{P})$ (some statistical parameter) of a data-generating distribution P using observed data. An *estimator* is a function that maps a sample $(x_1, \dots, x_n)$ to a numerical value intended to approximate $\gamma(\mathsf{P})$.

1.1 Plug-in Estimator

A fundamental idea is the *plug-in principle*: if $\gamma(\mathsf{P})$ is a functional of the distribution P, then one can estimate P first and evaluate γ at this estimate.

1.1.1 Empirical Distribution and Distribution Function

If I want to estimate the probability of some set A under my data generating distribution, I can just look at the n data points that have been collected as draws from the underlying data generating distribution. Then I can say, what's the probability of an event A?

It's simply. I'm going to count up how often do I see a data point in A, and then I put that into perspective. So it's the empirical proportion of how many data points fall in the set A. You want to estimate the probability, you just look in your data set how often did that event that you're interested in happen out of all.

The **empirical distribution** of $x_1, \dots, x_n \in \mathbb{R}$ is the probability distribution $\hat{\mathsf{P}}_n$ given by

$$\hat{\mathsf{P}}_n(A) := \frac{1}{n} \#\{i : x_i \in A\} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \in A\}}, \quad A \subseteq \mathbb{R}.$$

where $\mathbf{1}_{\{x_i \in A\}}$ is an indicator function that equals 1 if $x_i \in A$ and 0 otherwise. Thus, probabilities are estimated by empirical frequencies. The empirical distribution assigns mass $1/n$ to each observation and makes minimal assumptions about the underlying model.

The **empirical distribution function (ecdf)** of $x_1, \dots, x_n$ is the distribution function of $\hat{\mathsf{P}}_n$, which is

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq t\}}, \quad t \in \mathbb{R}$$

Remark.

- The empirical distribution could be defined for domains other than $\mathbb{R}$.
- We may write $\hat{\mathsf{P}}_n(A; x_1, \ldots, x_n)$ or $\hat{F}_n(t; x_1, \ldots, x_n)$ to highlight the dependence on the given data.

Example of an Empirical CDF

Let us start with normal sample of size $n = 101$.

```
set.seed(2020)
ggplot(data.fram(x = rnorm(10)), aes(x)) +
 stat_ecdf(geom = "step") +
 geom_funtion(data = data.frame(x = c(-3.5, 3.5)),
     fun = pnorm, colour = "red")
```

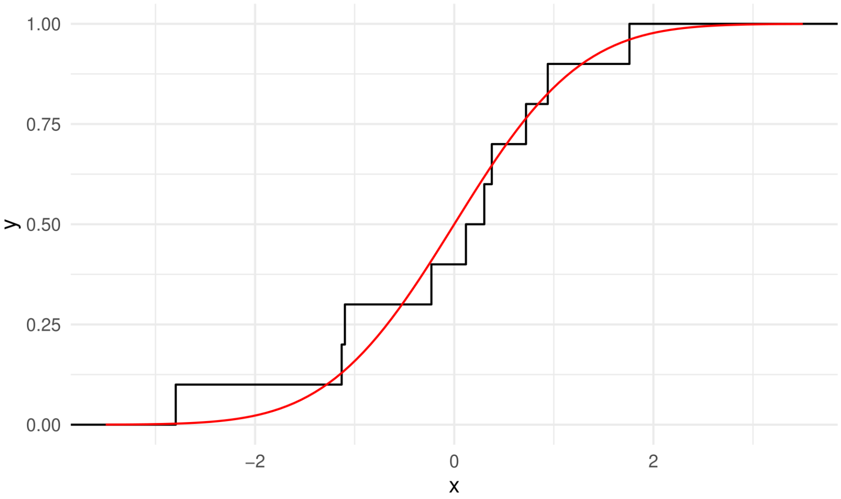


Figure 1.1

In Figure 1.1, the black curve represents the empirical cumulative distribution function (ECDF) based on 10 observations drawn from a standard normal distribution. The ECDF is a step function that increases by increments of $1/10$ at each observed data point, taking the value 0 for sufficiently small arguments and 1 for sufficiently large ones. The locations of the jumps correspond to the unordered sample values, while their heights reflect the sample size. Since the data are drawn from a continuous distribution, ties occur with probability zero. The empirical distribution (black) roughly follows the true standard normal cumulative distribution function (red), illustrating how empirical proportions approximate theoretical probabilities in accordance with the law of large numbers.

Note. From the ecdf we can recover the data points $x_1, \ldots, x_n$ up to their order, i.e., we can recover the order statistics.

Here is the plot for a normal sample of size $n = 500$.

```
set.seed(2020)
ggplot(data.fram(x = rnorm(500)), aes(x)) +
 stat_ecdf(geom = "step") +
 geom_funtion(data = data.frame(x = c(-3.5, 3.5)),
     fun = pnorm, colour = "red")
```

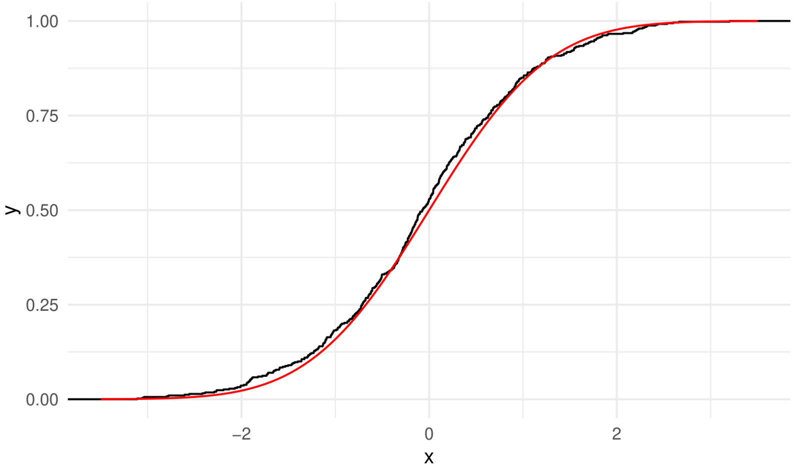


Figure 1.2

For a larger sample size, the empirical cumulative distribution function (ECDF) becomes smoother, since the jump size decreases as the number of observations increases. Although the ECDF may locally deviate from the true distribution function, it follows it closely overall. As the number of i.i.d. observations grows, the ECDF converges to the underlying distribution function. This shows that the ECDF is a consistent estimator of the true distribution function and highlights its connection to fundamental results such as the law of large numbers and the central limit theorem.

1.1.2 Glivenko-Cantelli Theorem

Let $X_1, X_2, \ldots$ be real valued r.v. (random variables) that are i.i.d. with cdf F.

For any fixed $t \in \mathbb{R}$, the probability $F(t)$ can be interpreted as the success probability of the indicator r.v. $\mathbf{1}_{\{X_i \le t\}}$, which is then clearly a Bernoulli r.v. with parameter $F(t)$, so we have

$$\mathsf{P}(X_i \le t) = F(t) \quad \text{and} \quad \mathbf{1}_{\{X_i \le t\}} \sim \text{Bernoulli}(F(t)),$$

so

$$\mathsf{E}[\mathbf{1}_{\{X_i \le t\}}] = F(t), \quad \mathsf{Var}[\mathbf{1}_{\{X_i \le t\}}] = F(t)(1 - F(t)).$$

The empirical distribution function evaluated at t can therefore be interpreted as the average of these indicator variables, linking it directly to fundamental results such as the law of large numbers.

Strong law of large numbers:

Since these indicators have finite expectation, the Strong Law of Large Numbers applies.

$$\begin{aligned}&\hat{F}_n(t) \equiv \hat{F}_n(t; X_1, \ldots, X_n) \xrightarrow{a.s.} F(t),\\ &\text{i.e.,}\quad \mathsf{P}\left(\lim_{n\to\infty} \hat{F}_n(t) = F(t)\right) = 1.\end{aligned}$$

Thus, for each fixed t, the value of the empirical distribution function converges almost surely to the corresponding value of the true distribution function. In particular, as the sample size increases, the empirical proportion of observations less than or equal to t converges with probability one to $F(t)$.

Central limit theorem:

For a fixed $t \in \mathbb{R}$, the Central Limit Theorem describes the fluctuation (the **error**) of the empirical distribution function around the true distribution function. After scaling by $\sqrt{n}$, the estimation error $\hat{F}_n(t) - F(t)$ converges in distribution to a centered normal distribution with variance $F(t)(1-F(t)$. This variance is inherited from the underlying Bernoulli indicator variables and quantifies the typical size of the estimation error.

$$\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow{d} \mathcal{N}\left(0, F(t)(1 - F(t))\right).$$

Thus, for large n, the estimation error at a fixed point t behaves approximately like a centered normal random variable with variance $F(t)(1 - F(t))$. This variance is maximized when $F(t) = 1/2$ and decreases as $F(t)$ approaches 0 or 1, indicating smaller fluctuations in the tails of the distribution, and bigger on the middle of the distribution.

Note. Convergence in distribution: $Y_n \xrightarrow{d} Y$ if and only if $F_{Y_n}(y) \underset{n\to\infty}{\longrightarrow} F_Y(y)$ at all points $y \in \mathbb{R}$ at which F_Y is continuous.

The Strong Law of Large Numbers (SLLN) ensures that, for each fixed $t \in \mathbb{R}$,

$$\hat{F}_n(t) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le t\}} \quad \xrightarrow{\text{a.s.}} \quad F(t).$$

In other words, the empirical CDF (black curve) converges almost surely to the true CDF (red curve) at any individual point.

However, this is *pointwise convergence*: it applies separately for each t, and does not automatically guarantee that the empirical CDF approximates the true CDF uniformly over all points. In particular, some regions of the distribution—often the tails—may require much larger sample sizes for the approximation to be accurate.

This motivates *plug-in estimation.* If we want to estimate a functional of the underlying distribution (e.g., the median), we can compute it using the empirical CDF:

$$\text{Estimate of } \theta(F) \quad = \quad \theta(\hat{F}_n),$$

where θ is any functional (mean, median, quantile, etc.). The idea is that the empirical CDF provides a good pointwise approximation to the true distribution, which can then be used to estimate various characteristics of the underlying data-generating process.

While the SLLN and CLT provide guarantees at individual points, they do not automatically control the empirical CDF as a whole. Since there are uncountably many points $t \in \mathbb{R}$, the number of samples required for a good approximation may vary across the domain. For points in the tails ($F(t)$ very small or very large), much larger sample sizes are needed for the CLT to provide an accurate normal approximation.

This motivates the question: can classical limit theorems (SLLN) be strengthened to show that the empirical CDF estimates the entire true distribution function simultaneously? Yes, and this is the content of the *Glivenko–Cantelli theorem.* It provides a uniform convergence guarantee, ensuring that the black ECDF approaches the red true CDF across the whole domain.

Theorem 1.1 (Glivenko-Cantelli)**.** *If $X_1, X_2, \dots$ are i.i.d. r.v. with cdf F, then the empirical CDF*

$$\hat{F}_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq t\}}$$

satisfies

$$\|\hat{F}_n - F\|_\infty := \sup_{t \in \mathbb{R}} \mid \hat{F}_n(t; X_1, \dots, X_n) - F(t) \mid \xrightarrow{a.s.} 0.$$

The Glivenko-Cantelli theorem strengthens the Strong Law of Large Numbers to uniform convergence over all $t \in \mathbb{R}$. It also addresses the global accuracy of the empirical CDF. Instead of evaluating convergence at a single point t, it considers the maximal deviation across all $t \in \mathbb{R}$.

By taking the supremo of the norm, we measure the worst-case difference between $\hat{F}_n(t)$ and $F(t)$. The theorem guarantees that this maximal deviation converges to zero almost surely. Intuitively, the empirical CDF (black curve) approximates the true CDF (red curve) uniformly across the entire real line as n grows.

Glivenko-Cantelli can be seen as a uniform Law of Large Numbers for the indicator functions $\mathbf{1}_{\{X_i \leq t\}}$, and Donsker's theorem is the corresponding functional analog of the Central Limit Theorem.

Remark.

- Glivenko-Cantelli is a LLN for $\hat{F}_n$ as element of a function space. The "functional generalization" of the CLT statement is known as Donsker's theorem.

- Empirical process theory generalizes these results to more general settings. For example, Glivenko-Cantelli results of the form

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathsf{E}[f(X_i)]\right| \xrightarrow{a.s.} 0.$$

Proof (Theorem 1.1). We first prove the theorem in the special case where $X_1, X_2, \ldots$ are i.i.d Unif(0,1), so $F(t) = t$ for $t \in [0,1]$.

Idea: Partition domain [0,1] *finite* number of blocks and use monotonicity of cdf to control behaviour in each block.

The main difficulty in the theorem is that $\hat{F}_n(t)$ defines a random variable for each $t \in \mathbb{R}$, so we are dealing with an uncountable collection of random variables. However, the key property we exploit is **monotonicity**:

- The true CDF F is strictly increasing on $[0,1]$.
- The empirical CDF $\hat{F}_n$ is a non-decreasing step function.

This monotonicity allows us to control the sup-norm deviation of $\hat{F}_n$ from F over an entire interval by examining finitely many points.

For any $M \in \mathbb{N}$, we partition $[0,1]$ into M subintervals of equal length:

$$I_j := \left[\frac{j-1}{M}, \frac{j}{M}\right], \quad j = 1, \ldots, M.$$

Within each interval, monotonicity implies that the maximum deviation of $\hat{F}_n(t)$ from $F(t)$ occurs at the endpoints. Then, by the monotonicity of $\hat{F}_n$ and F,

$$\begin{aligned}
\|\hat{F}_n - F\|_\infty &= \max_{1\le j\le M} \sup_{t\in\left[\frac{j-1}{M},\frac{j}{M}\right]} \mid \hat{F}_n - F(t) \mid \\
&= \max_{1\le j\le M} \sup_{t\in\left[\frac{j-1}{M},\frac{j}{M}\right]} \max\{\hat{F}_n - F(t), F(t) - \hat{F}_n(t)\} \\
&\overset{\hat{F}_n\uparrow, F\uparrow}{\le} \max_{1\le j\le M} \max\left\{\hat{F}_n\left(\frac{j}{M}\right) - F\left(\frac{j-1}{M}\right), F\left(\frac{j}{M}\right) - \hat{F}_n\left(\frac{j-1}{M}\right)\right\} \\
&= \max_{1\le j\le M} \max\left\{\hat{F}_n\left(\frac{j}{M}\right) - \frac{j-1}{M}, \frac{j}{M} - \hat{F}_n\left(\frac{j-1}{M}\right)\right\} \\
&= \max_{1\le j\le M} \max\left\{\hat{F}_n\left(\frac{j}{M}\right) - \frac{j}{M}, \frac{j-1}{M} - \hat{F}_n\left(\frac{j-1}{M}\right)\right\} + \frac{1}{M}.
\end{aligned}$$

Since there are only finitely many endpoints, we can apply the Strong Law of Large Numbers at each endpoint. For all $j = 1, \ldots, M$,

$$\hat{F}_n\left(\frac{j}{M}\right) \xrightarrow{a.s.} \frac{j}{M}.$$

Hence, with probability one,

$$\lim_{n\to\infty} \max_{1\le j\le M} \max\left\{\hat{F}_n\left(\frac{j}{M}\right) - \frac{j}{M}, \frac{j-1}{M} - \hat{F}_n\left(\frac{j-1}{M}\right)\right\} = 0.$$

On the same event of probability one,

$$\lim_{n\to\infty} \sup \|\hat{F}_n - F\|_\infty \leq \frac{1}{M}.$$

Since M was arbitrary, we conclude that

$$\|\hat{F}_n - F\|_\infty \xrightarrow{a.s.} 0.$$

We will finish the proof for general r.v. at the end of the next section ... □

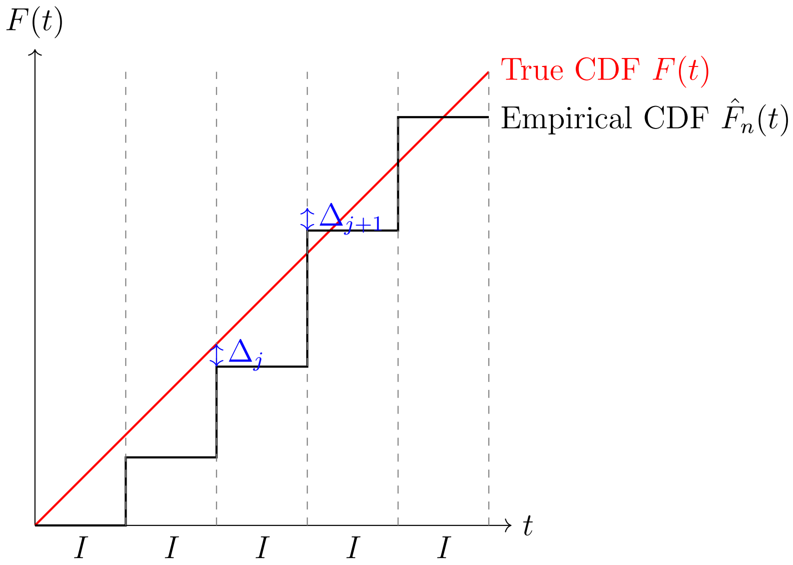


Figure 1.3: Illustration of the partition argument in the proof of the Glivenko-Cantelli theorem. The black curve is the empirical CDF, the red curve is the true CDF, and dashed gray lines indicate the partition intervals. The blue arrows show maximal deviation in each interval.

Remark.

- This argument reduces the problem from an uncountable collection of random variables to finitely many points using the monotonicity of the CDF and empirical CDF.
- The extension to a general distribution F is straightforward via the probability integral transform: if $X_i \sim F$, then $F(X_i) \sim \mathrm{Unif}(0, 1)$.

1.1.3 Quantile Function and Uniform Representation

Having established the Glivenko-Cantelli theorem for $\mathrm{Unif}(0, 1)$ random variables, we can now discuss how to extend this result to general distributions and motivate the *plug-in principle* for estimation.

For uniform random variables, the proof is convenient because the distribution function is explicit: $F(t) = t$ on $[0, 1]$. This allows us to evaluate the empirical CDF $\hat{F}_n$ at specific grid points and directly control its deviation from F.

To generalize, consider a random variable X with cumulative distribution function F. Recall that F is non-decreasing:

$$F(t) = \Pr[X \leq t], \quad t \in \mathbb{R}.$$

If F is strictly increasing, it admits a well-defined inverse. Even if F is flat over some interval $[a, b]$, meaning $F(a) = F(b)$, this simply indicates that X does not take values in $[a, b]$ with positive probability.

This motivates the definition of the *generalized inverse* (or *quantile function*).

The **quantile function** of a cdf F is the left-continuous function

$$F^{-1}(q) = \inf\{x \in \mathbb{R} : F(x) \geq q\}, \quad q \in (0, 1).$$

This definition assigns to each $q \in (0, 1)$ the smallest value x such that the distribution function reaches or exceeds q.

The quantile function allows us to map uniform random variables to a general distribution and thereby extend results proven for the uniform case to arbitrary distributions. This forms the basis of the plug-in principle: given an estimator $\hat{F}_n$ of F, we can estimate functionals of F by $\phi(\hat{F}_n)$.

Recall that $F(t)$ is right-continuous.

Remark.

- If F is strictly increasing, the quantile function coincides with the usual inverse.
- In general, F^{-1} is a well-defined function for any distribution function F.
- By the right-continuity of F (a property of all distribution functions), the quantile function F^{-1} is *left-continuous.*

Question: What is the quantile function of this cdf?

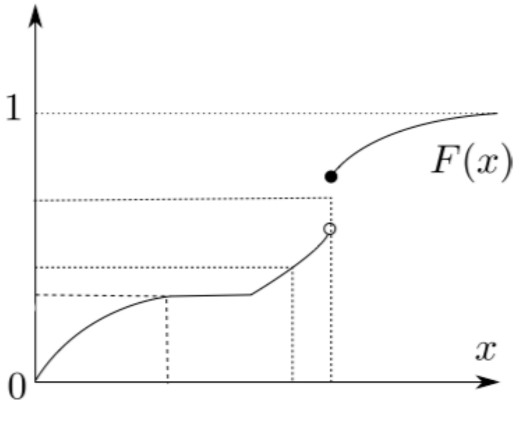


Figure 1.4

Suppose we are given a cdf F as a graph. The graph may have the following features:

- an initial flat segment at 0,
- a strictly increasing portion,
- a flat interval,
- jumps (discontinuities) due to the right-continuity of F,
- and it eventually converges to 1.

The quantile function F^{-1} can be interpreted as a "generalized inverse" of this graph. Graphically:

- A strictly increasing segment of F becomes an increasing segment of F^{-1}.
- A flat segment of F (where F does not increase) becomes a *jump* in F^{-1} — we take the lower endpoint of the interval.
- A jump in F (where F jumps upward due to right-continuity) becomes a flat segment in F^{-1}.

In other words, to obtain the graph of the quantile function, one can "reflect" F across the diagonal $y = x$, and then interpret flat segments and jumps according to the rules above. This construction ensures that F^{-1} is left-continuous and correctly represents the distribution of quantiles.

Exercise: Draw F with an increasing segment, a flat interval, and a jump, and then sketch the corresponding F^{-1} to see these transformations in action.

Uniform Representation

Lemma 1.1. *If $U \sim Unif(0,1)$, then $X := F^{-1}(U)$ has cdf F.*

Quantile functions provide a powerful tool for generating random variables with a desired distribution. Specifically, if we want a random variable X with cdf F, we can construct it by transforming a uniform random variable $U \sim \text{Unif}(0,1)$ through the quantile function:

$$X = F^{-1}(U).$$

This works because the quantile function maps the uniform probability space $[0,1]$ into the distribution defined by F. For example, if F is the cdf of a standard normal distribution, then $F^{-1}(U)$ produces a standard normal random variable.

This construction is the basis of one of the simplest and most fundamental methods for simulating random variables on a computer: start with a uniform random number generator, then transform the output via the quantile function of the desired distribution.

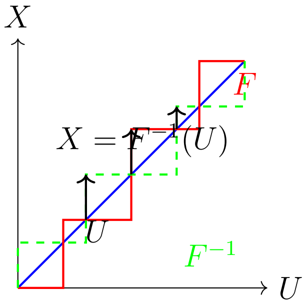


Figure 1.5: Graphical interpretation of the uniform representation: mapping a uniform random variable U to a random variable X with CDF F using the quantile function F^{-1}.

While this is conceptually straightforward, in practice there are often more efficient simulation methods implemented in statistical software packages such as R. Nevertheless, the uniform representation is extremely useful for both theoretical and practical purposes, as it allows us to write any random variable of interest explicitly as a function of a uniform random variable. Such representations are widely used in probability theory, simulation, and Monte Carlo methods.

Proof (Lemma 1.1). For any $u \in (0,1)$ and $x \in \mathbb{R}$, the quantile function F^{-1} satisfies

$$F^{-1}(u) \leq x \iff u \leq F(x).$$

($\Longleftarrow$) : Follows from the definition of $F^{-1}(u)$ as an infimum.

($\Longrightarrow$) : Follows by observing that

$$F^{-1}(u) \leq x \implies F(x+\epsilon) \geq u \quad \forall \epsilon > 0$$

taking $\epsilon \to 0$ and using the right-continuity of F gives $F(x) \geq u$.

Since $U \sim \text{Unif}(0,1)$, using Theorem 1.1, we obtain that for all $x \in \mathbb{R}$,

$$\mathsf{P}(X \leq x) = \mathsf{P}(F^{-1}(U) \leq x) = \mathsf{P}(U \leq F(x)) = F(x).$$

□

Proof (Theorem 1.1 Glivenko-Cantelli for general r.v.) Let $U_1, U_2, \ldots$ be i.i.d Unif(0,1) with cdf $G(t) = t, t \in [0,1]$.

By the uniform (quantile) representation, the sequence

$$X_i := F^{-1}(U_i), \qquad i = 1, 2, \ldots,$$

is i.i.d. with cdf F. In particular, the joint distribution of $(X_1, X_2, \ldots)$ coincides with that of $(F^{-1}(U_1), F^{-1}(U_2), \ldots)$.

Let $\hat{G}_n$ be the ecdf of $U_1, \dots, U_n$. Then for every n,

$$\begin{aligned}\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \le t\}} &\underset{d}{=} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{F^{-1}(U_i) \le t\}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{U_i \le F(t)\}} = \hat{G}_n(F(t)),\end{aligned}$$

where Lemma 1.1 was used in the second last equality.

Noting that $G(F(t)) = F(t)$, we conclude that

$$\|\hat{F}_n - F\|_\infty \overset{d}{=} \|\hat{G}_n \circ F - G \circ F\|_\infty = \sup_{t\in\mathbb{R}} \big|\hat{G}_n(F(t)) - G(F(t))\big|.$$

Since $F(t) \in [0,1]$ for all t, we obtain the bound

$$\|\hat{F}_n - F\|_\infty = \sup_{t\in\mathbb{R}} \big|\hat{G}_n(F(t)) - G(F(t))\big| \le \sup_{t\in[0,1]} \big|\hat{G}_n(t) - G(t)\big| = \|\hat{G}_n - G\|_\infty.$$

By the Glivenko-Cantelli Theorem 1.1 for the uniform distribution,

$$\|\hat{G}_n - G\|_\infty \xrightarrow{\text{a.s.}} 0.$$

Therefore,

$$\|\hat{F}_n - F\|_\infty \xrightarrow{\text{a.s.}} 0,$$

which proves the theorem for a general distribution function F. □

1.1.4 Plug-in Estimation

The Glivenko–Cantelli theorem suggests replacing the unknown cdf F with the ecdf $\hat{F}_n$ when estimating a statistical parameter of the form $\gamma(F)$, where γ is a functional defined on distributions and $X_1, \dots, X_n$ are i.i.d. r.v. with distribution F.

Definition 1.1. The **plug-in estimator** of the parameter $\gamma(F)$ is defined as

$$\hat{\gamma} := \gamma(\hat{F}_n).$$

Note. The functional γ must be well defined for discrete distributions, since the empirical distribution function $\hat{F}_n$ is discrete.

The principle of plug-in estimation is conceptually straightforward. Rather than attempting to estimate $\gamma(F)$ directly, we first estimate the underlying distribution F using the empirical distribution $\hat{F}_n$. We then compute the parameter of interest by applying the same functional γ to $\hat{F}_n$.

This approach provides a general and intuitive framework for constructing estimators. Many parameters of interest, such as moments or quantiles, can be expressed as functionals of the distribution. Plug-in estimation therefore offers a natural way to estimate such quantities by substituting the unknown distribution with its empirical counterpart.

Example 1.1. Consider the mean (expectation) functional

$$\gamma(F) = \int x\, dF(x).$$

The corresponding plug-in estimator is obtained by replacing F with the empirical distribution function $\hat{F}_n$:

$$\gamma(\hat{F}_n) = \int x\, d\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} X_i =: \bar{X}_n,$$

which is the sample mean.

This example illustrates the plug-in principle in a particularly simple setting. The population mean is a functional that integrates the identity function with respect to the distribution F. To estimate it, we apply the same functional to the empirical distribution $\hat{F}_n$.

Since $\hat{F}_n$ is a discrete distribution that assigns probability $1/n$ to each observation $X_1, \ldots, X_n$, integration with respect to $\hat{F}_n$ reduces to taking the average of the observed data points. Thus, the mean of the empirical distribution coincides with the usual sample mean.

Many other estimators can be interpreted in the same way. Sample moments, variances, and correlations arise by applying the corresponding functionals to the empirical distribution.

The main limitation of this approach is that the functional must be well defined for discrete distributions. In particular, functionals involving densities cannot be directly estimated via plug-in methods, since the empirical distribution does not admit a density with respect to Lebesgue measure.

1.2 M-Estimators

An estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ that maximizes a criterion function of the form

$$\theta \mapsto \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i),$$

where m_θ is a known real-valued function, is called an **M-estimator** (maximum-likelihood type).

M-estimation is a general principle for constructing estimators based on optimization. The parameter estimate $\hat{\theta}$ is obtained by maximizing (or minimizing) an empirical objective function, often interpreted as an average (empirical) loss or criterion evaluated at the observed data.

This framework is broad and includes many classical estimators. For instance, in parametric models such as the normal distribution with unknown mean and variance, parameter values can be chosen by optimizing an appropriate criterion function. In practice, such criteria are typically expressed as averages of the form $m_\theta(X_i)$, with one contribution from each data point.

Example 1.2.

- For $\theta \in \mathbb{R}$, let
$$m_\theta(x) = -(x-\theta)^2.$$
The resulting M-estimator $\hat{\theta}$ is the sample mean $\bar{X}_n$. Indeed, for any random variable X with distribution function F and finite variance, the expectation satisfies
$$\mathsf{E}[X] = \arg\min_{\theta\in\mathbb{R}} \mathsf{E}_{X\sim F}\big[(X-\theta)^2\big].$$
Replacing F by the empirical distribution $\hat{F}_n$ yields
$$\begin{aligned}\bar{X}_n = \gamma(\hat{F}_n) &= \arg\min_{\theta\in\mathbb{R}} \mathsf{E}_{X\sim\hat{F}_n}\left[(X-\theta)^2\right] \\ &= \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n}(X_i-\theta)^2.\end{aligned}$$

- Replacing the squared loss by the absolute loss, $m_\theta(x) = -|x-\theta|$, yields the sample median as the corresponding M-estimator.

- The **Huber estimator** is the M-estimator defined by
$$m_\theta(x) = -\rho_c(x-\theta),$$
where
$$\rho_c(x) = \begin{cases} x^2, & \text{if } |x| \le c, \\ c(2|x|-c), & \text{if } |x| > c. \end{cases}$$

Connection to plug-in estimation

M-estimation can be viewed as an application of the plug-in principle. A parameter is defined as the optimizer of a *population* objective function, typically an expected loss under the data-generating distribution. The corresponding estimator is obtained by replacing this expectation with an empirical average computed from the sample.

For instance, the population mean can be characterized as the minimizer of the expected squared deviation. Replacing the expectation by an average over the data leads to the empirical squared loss, whose minimizer is the sample mean. Thus, the sample mean can be interpreted both as a plug-in estimator and as an M-estimator.

More generally, whenever a parameter can be expressed as the minimizer (or maximizer) of an expected loss function, replacing the expectation by an empirical mean yields an estimator of M-estimation type. Absolute loss leads to the sample median, while other loss functions give rise to robust estimators such as the Huber estimator.

Throughout, it is important to distinguish between *population* quantities—defined in terms of the true data-generating distribution—and *sample* quantities, which are computed from the observed data.

1.3 Method of Moments (MOM)

Suppose we observe real-valued data from a parametric model

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathsf{P}_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}^k.$$

For $j = 1, \dots, k$, define the j-th moment of the distribution as

$$\mu_j(\theta) = \mathsf{E}_\theta[X_i^j].$$

Whenever it exists, the j-th moment can be estimated empirically by

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j.$$

Definition 1.2. The **method of moments (MOM) estimator** $\hat{\theta}$ is defined as a solution to the system of equations

$$\mu_j(\theta) = \hat{\mu}_j, \quad j = 1, \dots, k.$$

Subtleties: A solution to this system need not exist, and if it exists, it may not be unique.

The intuition behind the method of moments is simple. For each parameter value θ, the model implies specific values for the moments $\mu_1(\theta), \dots, \mu_k(\theta)$. On the other hand, these moments can be estimated directly from the data by taking empirical averages. The method of moments chooses the parameter value $\hat{\theta}$ for which the theoretical moments of the model match the empirical moments computed from the sample.

In practice, this leads to a system of k equations in k unknowns. Whether this system admits a solution, and whether that solution is unique, depends on the specific model under consideration.

Example 1.3 (Negative binomial model)**.** Let $\mathrm{NegBin}(k, \theta)$ denote the distribution of the number of failures observed until the k-th success in independent Bernoulli trials with success probability θ.

Suppose

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathsf{P}_\theta = \mathrm{NegBin}(k, \theta),$$

where $k \in \mathbb{N}$ is known and $\theta \in (0, 1)$ is unknown. The probability mass function is given by

$$\mathsf{P}_\theta(X_i = x) = \binom{k + x - 1}{x} \theta^k (1 - \theta)^x, \quad x = 0, 1, 2, \dots$$

Since there is a single unknown parameter, we use the first moment. The expectation of X_i is

$$\mu_1(\theta) = \mathsf{E}_\theta[X_i] = \frac{k(1 - \theta)}{\theta}.$$

The corresponding method of moments equation is

$$\frac{k(1-\theta)}{\theta} \overset{!}{=} \hat{\mu}_1 = \bar{X}_n.$$

Solving for θ yields the MOM estimator

$$\hat{\theta} = \frac{k}{\bar{X}_n + k} = \frac{nk}{\sum_{i=1}^n X_i + nk}.$$

This estimator admits the interpretation

$$\hat{\theta} = \frac{\text{total number of successes}}{\text{total number of trials}},$$

since across the n experiments exactly nk successes are observed and $\sum_{i=1}^n X_i$ failures occur in total.

Example 1.4 (Gaussian model)**.** Suppose

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

where both the mean $\mu \in \mathbb{R}$ and the variance $\sigma^2 \in (0, \infty)$ are unknown. The parameter is

$$\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty).$$

The density of the normal distribution is

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad x \in \mathbb{R}.$$

Since there are two unknown parameters, we use the first two moments. The population moments are

$$\begin{aligned}
\mu_1(\theta) &= \mathsf{E}_\theta[X_1] = \mu, \\
\mu_2(\theta) &= \mathsf{E}_\theta[X_1^2] = \mathsf{Var}_\theta(X_1) + \big(\mathsf{E}_\theta[X_1]\big)^2 = \sigma^2 + \mu^2.
\end{aligned}$$

The corresponding method of moments equations are

$$\begin{aligned}
\mu &= \bar{X}_n, \\
\sigma^2 + \mu^2 &= \hat{\mu}_2,
\end{aligned}$$

where

$$\hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n X_i^2.$$

Solving this system yields the MOM estimators

$$\begin{aligned}
\hat{\mu} &= \bar{X}_n, \\
\hat{\sigma}^2 &= \hat{\mu}_2 - (\bar{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2.
\end{aligned}$$

which are actually the sample mean and the empirical variance.

Example 1.5 (Pareto model)**.** Suppose P_θ has the density

$$p_\theta(x) = \theta \cdot (1+x)^{-(1+\theta)}, \quad x > 0,$$

with unknown parameter $\theta \subseteq \Theta = (0, \infty)$.

Density for $\theta \in \{0.5, 1, 2, 3\}$:

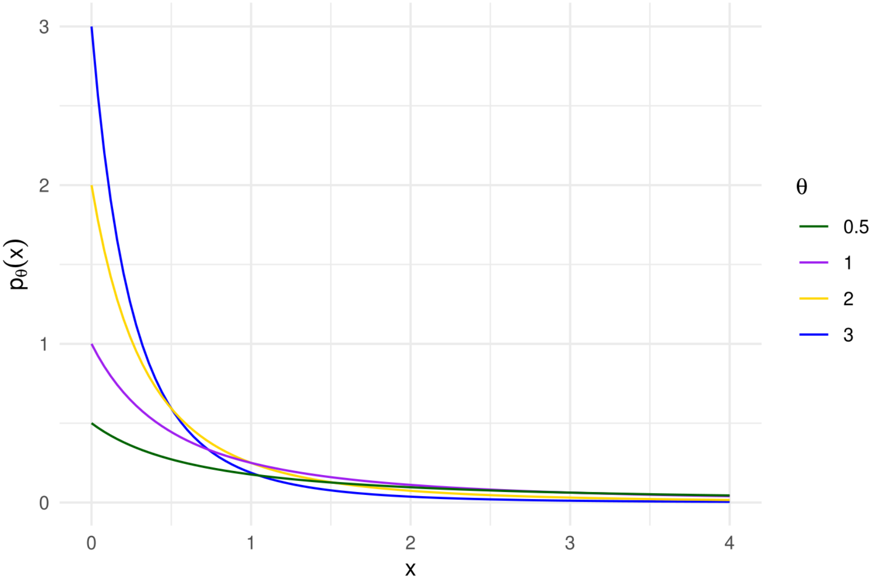


Figure 1.6

First moment $\mu_1(\theta) = \mathsf{E}_\theta[X_1]$ exists for which θ?

We have

$$\begin{aligned}\mathsf{E}_\theta[X_1] &= \int_0^\infty x\theta \cdot (1+x)^{-(1+\theta)} dx \\ &= x \cdot (1+x)^{-\theta} \mid_0^\infty + \int_0^\infty (1+x)^{-\theta} dx \\ &= \frac{1}{1-\theta}(1+x)^{-\theta+1} \mid_0^\infty \\ &= \lim_{a\to\infty} \frac{1}{1-\theta}(1+a)^{-\theta+1} + \frac{1}{\theta-1},\end{aligned}$$

which is finite for $\theta > 1$.

If $\mu_1(\theta)$ exists, then the MOM is obtained by equating the sample mean $\bar{X}_n$ with $\mu_1(\theta)$

$$\mu_1(\theta) = \bar{X}_n = \frac{1}{\theta - 1}.$$

MOM estimator is

$$\hat{\theta} = \frac{1 + \bar{X}_n}{\bar{X}_n}.$$

How do you think $\hat{\theta}$ behaves if $\mu_1(\theta)$ doesn't exists and we get a large sample (i.e., if n is large)?

As sample size n increases, the MOM estimator is heavily influenced by extreme values and does not converge to the true θ. Instead, it tends to values around 1 or slightly above.

Remark.

- The estimator $\hat{\theta}$ is always well-defined for any sample because $\bar{X}_n > 0$, even if the true mean does not exist ($\theta \leq 1$).
- If $\theta > 1$, $\hat{\theta}$ converges to the true parameter as $n \to \infty$.
- If $\theta \leq 1$, the population mean does not exist. In this case, $\hat{\theta}$ will still produce a value greater than 1, typically overestimating θ, because extreme values dominate the sample mean.
- This illustrates a general limitation of the method of moments: it assumes that the relevant population moments exist. Heavy-tailed distributions with non-existent moments can produce misleading estimates.

Comments on the Method of Moments (MOM)

- By definition, the method of moments uses the first k population moments to estimate a k–dimensional parameter. Generally, higher-order moments (μ_j with large j) are estimated less accurately by their sample counterparts $\hat{\mu}_j$.
- For multivariate observations $X_i = (X_{i1}, \ldots, X_{id})$, one can consider expectations of each coordinate, their powers, and cross-products, e.g.,

$$\mathsf{E}[X_{i1}],\ \mathsf{E}[X_{i2}],\ \mathsf{E}[X_{i1}^2],\ \mathsf{E}[X_{i1}X_{i2}],\ \mathsf{E}[X_{i2}^2], \ldots$$

- More generally, one can use arbitrary functions $h_1, \ldots, h_k$ of the data to construct the equations

$$\mathsf{E}_\theta[h_j(X_i)] = \frac{1}{n}\sum_{i=1}^{n} h_j(X_i), \quad j = 1, \ldots, k,$$

  potentially leading to improved estimators. The choice of the functions h_j is, however, not always obvious.
- In practice, powers of the observations are typically used because they often allow for closed-form calculations of expectations under the assumed model. This simplifies solving the moment equations.
- MOM can be extended to multivariate data by including powers and cross-products of the coordinates, forming monomials that define the estimating equations.

- While maximum likelihood estimation often produces more efficient estimators and is preferred in many classical examples, the method of moments remains valuable. It can provide computationally simple and tractable estimators, especially in modern applications where closed-form or robust solutions are desired.

Note. Although sometimes considered a "classical" or "old-fashioned" method, MOM continues to be useful in practical and theoretical settings, albeit with some limitations in accuracy and uniqueness of solutions.

1.4 Maximum Likelihood Estimation (MLE)

Consider a parametric statistical model for an observed random element

$$X \sim \mathsf{P}_\theta, \quad \theta \in \Theta \subseteq \mathbb{R}^k.$$

The collection

$$\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$$

is referred to as a *parametric model* with parameter space Θ.

We assume that the model is *dominated*, meaning that there exists a σ-finite measure ν such that

$$\mathsf{P}_\theta \ll \nu \quad \text{for all } \theta \in \Theta.$$

This assumption implies that every distribution in the model admits a density with respect to ν, given by

$$p_\theta(x) = \frac{d\mathsf{P}_\theta}{d\nu}(x), \quad x \in \mathcal{X}.$$

This framework encompasses both continuous and discrete models. For instance, when ν is Lebesgue measure, p_θ is an ordinary probability density function, while if ν is counting measure, p_θ is a probability mass function. Throughout, we use P_θ to denote the distribution of the data and p_θ to denote its corresponding density.

The maximum likelihood principle is one of the most widely used methods for inference in parametric models. For simplicity of exposition, we present the method for a single observation X; the same framework applies when X represents a sample, a vector, or a more general data object.

Definition 1.3 (Likelihood Function)**.** Let $x \in \mathcal{X}$ denote an observed data point. The function

$$L_x(\theta) = p_\theta(x), \qquad \theta \in \Theta,$$

is called the **likelihood function** of the model $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ for the data x.

Recall that, for a fixed parameter value θ, the density $p_\theta(x)$ is a function of the data x and fully characterizes the distribution P_θ. Probabilities of events are obtained by integrating this density with respect to the underlying measure.

In likelihood-based inference, this perspective is reversed: the observed data x is treated as fixed, and the density is viewed as a function of the parameter θ. The resulting function $L_x(\theta)$ evaluates how compatible different parameter values are with the observed data.

In discrete models, where p_θ is a probability mass function, $L_x(\theta)$ coincides with the probability of observing the data x under the distribution P_θ. More generally, the likelihood function provides a basis for comparing the relative plausibility of different parameter values given the data.

Definition 1.4 (Maximum Likelihood Estimation (MLE))**.** The **maximum likelihood estimate (MLE)** of θ based on observed data x is defined as

$$\hat{\theta}(x) = \arg\max_{\theta \in \Theta} L_x(\theta),$$

where $L_x(\theta) = p_\theta(x)$ is the likelihood function.

If $\hat{\theta}(x)$ can be chosen as a measurable function of the observation X, then $\hat{\theta}(X)$ is called the *maximum likelihood estimator (MLE)* of θ.

The principle underlying maximum likelihood estimation is to select the parameter value under which the observed data are most compatible with the model. In discrete models, this corresponds to choosing the parameter that maximizes the probability of observing the data. In continuous models, probabilities of exact observations are zero, and the likelihood is interpreted in terms of density values rather than probabilities. In both cases, the guiding principle is to choose the parameter value that maximizes the likelihood of the observed data.

Subtleties: The definition raises several important questions, including whether a maximizer exists, whether it is unique, and whether the likelihood function is bounded. These issues depend on the specific model under consideration.

Log–Likelihood Function

It is often convenient to work with the **log–likelihood function**, defined by

$$\ell_x(\theta) = \log L_x(\theta),$$

where $L_x(\theta)$ denotes the likelihood function.

The use of the log–likelihood is motivated by both computational and theoretical considerations:

- **Numerical stability.** Likelihood functions often involve products of probabilities or density values, which can be extremely small. Taking logarithms avoids numerical underflow and improves numerical stability in optimization algorithms.

- **Analytical convenience.** Logarithms transform products into sums, which simplifies differentiation and optimization. This additive structure is essential for analytical tractability and for studying the asymptotic behaviour of maximum likelihood estimators using tools such as the law of large numbers and the central limit theorem.

In particular, suppose the observation consists of an i.i.d. sample $\mathbf{X} = (X_1, \dots, X_n)$, where each $X_i \sim \mathsf{P}_\theta$ has density p_θ. Then the likelihood function is given by

$$L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} p_\theta(X_i),$$

and the corresponding log–likelihood function is

$$\ell_{\mathbf{X}}(\theta) = \sum_{i=1}^{n} \log p_\theta(X_i).$$

Example 1.6 (Negative Binomial model)**.** Check that $\hat{\theta}_{\mathrm{ML}} = \hat{\theta}_{MOM}$, where $\hat{\theta}_{MOM}$ is the MOM estimator from Example 1.3.

Example 1.7 (Gaussian model; MLE = MOM)**.** Let $X_1, \dots, X_n$ be i.i.d. random variables with distribution $\mathcal{N}(\mu, \sigma^2)$, where $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. Assume $n \geq 2$ and that

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 > 0 \quad \text{a.s.}$$

The log–likelihood function is given by

$$\begin{aligned}
\ell_{\mathbf{X}}(\mu, \sigma^2) &= \sum_{i=1}^{n} \log\biggl(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\biggl\{ -\frac{(X_i - \mu)^2}{2\sigma^2} \biggr\} \biggr) \\
&= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2.
\end{aligned}$$

We first maximize $\ell_{\mathbf{X}}(\mu, \sigma^2)$ with respect to μ. For any fixed $\sigma^2 > 0$, this is equivalent to minimizing $\sum_{i=1}^{n} (X_i - \mu)^2$, which yields

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (X_i - \mu)^2 = \bar{X}_n.$$

Substituting $\hat{\mu} = \bar{X}_n$ into the log–likelihood and maximizing (set derivative to 0) with respect to σ^2 gives

$$\begin{aligned}
\hat{\sigma}^2 &= \arg\max_{\sigma^2 > 0} \left\{ -\log(\sigma^2) - \frac{1}{\sigma^2} \cdot \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right\} \\
&= \arg\min_{\sigma^2 > 0} \left\{ \log(\sigma^2) + \frac{1}{\sigma^2} \cdot \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \right\} \\
&= \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2.
\end{aligned}$$

Hence, in the Gaussian model, the maximum likelihood estimators coincide with the method–of–moments (MOM) estimators: the sample mean and the empirical variance.

If $\sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = 0$, the log–likelihood is unbounded above as $\sigma^2 \downarrow 0$, and the MLE for σ^2 does not exist. This corresponds to the degenerate case in which all observations are identical and the variance is not identifiable.

Example 1.8 (Pareto model; MLE $\neq$ MOM)**.** Let $X_1, \dots, X_n$ be i.i.d. random variables with density

$$p_\theta(x) = \theta(1+x)^{-(1+\theta)}, \qquad x > 0,$$

where the parameter space is $\Theta = (0, \infty)$.

The log–likelihood function based on the sample $\mathbf{X} = (X_1, \dots, X_n)$ is

$$\begin{aligned} \ell_{\mathbf{X}}(\theta) &= \sum_{i=1}^{n} \log p_\theta(X_i) \\ &= \sum_{i=1}^{n} [\log\theta - (1+\theta)\log(1+X_i)] \\ &= n\log\theta - (1+\theta)\sum_{i=1}^{n}\log(1+X_i). \end{aligned}$$

Differentiating with respect to θ gives

$$\frac{d}{d\theta}\ell_{\mathbf{X}}(\theta) = \frac{n}{\theta} - \sum_{i=1}^{n}\log(1+X_i).$$

Setting this derivative equal to zero yields the unique critical point

$$\hat{\theta}_{\text{ML}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\log(1+X_i)}.$$

Since the second derivative of $\ell_{\mathbf{X}}$ is negative, this critical point corresponds to a maximum, and $\hat{\theta}_{\text{ML}}$ is the maximum likelihood estimator.

In contrast, the method–of–moments estimator is obtained by matching the sample mean to the theoretical mean and is given by

$$\hat{\theta}_{\text{MOM}} = \frac{1+\bar{X}_n}{\bar{X}_n},$$

provided the mean exists. Hence, for this Pareto model, the MLE and MOM estimators do not coincide.

The MLE remains well defined for all $\theta > 0$ and is asymptotically more efficient than the method–of–moments estimator. In this sense, maximum likelihood estimation can be interpreted as automatically identifying an appropriate transformation of the data—here, $\log(1+X)$—that leads to an effective estimating equation.

1.5 Bayes Estimators

Bayesian estimation incorporates prior knowledge into statistical inference. This approach is particularly useful in situations where only limited data are available.

For instance, when estimating the incidence of a rare disease based on a single observed individual, the resulting estimate becomes extreme: either the entire population is inferred to have the disease if the individual is affected, or no one is inferred to have it if the individual is unaffected.

The Bayesian framework provides a principled way to combine prior knowledge with observed data, leading to more stable and interpretable estimators when data alone are insufficient.

Consider an observation (dataset) modeled as $X \sim P_\theta, \theta \in \Theta \subseteq \mathbb{R}^k$.

Idea of Bayesian Inference

- Treat θ as a random variable and choose a **prior distribution** for θ. The prior distribution summarizes all information available before observing the data.
- Regard P_θ as the conditional distribution of the data, X, given the parameter, θ.
$$X \mid \theta \sim \mathsf{P}_\theta$$
- The prior distribution of θ and the conditional distribution of $X \mid \theta$ together define a joint distribution for (X, θ).
- After observing data $X = x$, inference is based on the **posterior distribution**
$$\theta \mid X = x,$$
which is the conditional distribution of θ given the observed data.

  The posterior distribution combines prior beliefs with information from the data.
- In Bayesian inference, all estimation, uncertainty quantification, and decision-making about θ are derived from the posterior distribution.

Theorem 1.2 (Bayes Theorem)**.** *Suppose the prior distribution of θ has density π w.r.t to a measure ν, and that $P_\theta \ll \nu \quad \forall \theta \in \Theta$, with densities $p_\theta(x) = p(x|\theta)$. Interpreting $p(x \mid \theta)$ as the conditional density of X given θ, the posterior distribution of θ given $X = x$ has density*

$$p(\theta \mid x) = \frac{p(x \mid \theta)\,\pi(\theta)}{p(x)},$$

where

$$p(x) = \int_\Theta p(x \mid \theta)\,\pi(\theta)\,d\nu(\theta)$$

is the marginal (prior predictive) density of X.

Note. The posterior density satisfies

$$p(\theta \mid x) \propto L_x(\theta)\,\pi(\theta) = \text{Likelihood} \cdot \text{prior}.$$

Since the normalization constant $p(x)$ does not depend on θ, it can usually be ignored when identifying the form of the posterior distribution. This is analogous to working with the log-likelihood and dropping terms that do not depend on θ.

Bayes Estimators of θ are obtained as characteristics of the posterior distribution. The most commonly used estimator is the **posterior mean**, defined by

$$\hat{\theta} = \mathsf{E}[\theta \mid X = x] = \int_\Theta \theta\, p(\theta \mid x)\, d\nu(\theta).$$

Example 1.9 (Gaussian model, variance known)**.** Assume we observe $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known and μ is unknown. For mathematical convenience, we take a Gaussian prior on μ:

$$\mu \sim \mathcal{N}(m, \tau^2), \quad \pi(\mu) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(- \frac{(\mu - m)^2}{2\tau^2} \right),$$

reflecting prior belief that μ is near m with uncertainty τ.

The likelihood function is

$$L_X(\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(X_i - \mu)^2}{2\sigma^2} \right).$$

Posterior distribution. Using Bayes' theorem (up to proportionality):

$$p(\mu \mid X) \propto L_X(\mu)\pi(\mu) \propto \exp\left(- \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 - \frac{1}{2\tau^2}(\mu - m)^2 \right),$$

which is quadratic in μ. Since only Gaussian densities are exponentials of quadratics, the posterior is a normal distribution:

$$\mu \mid X \sim \mathcal{N}\Big(\mathsf{E}[\mu \mid X], \mathsf{Var}[\mu \mid X]\Big).$$

Completing the square. Rewriting:

$$\begin{aligned} p(\mu \mid X) &\propto \exp\left\{ - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i^2 - 2X_i\mu + \mu^2) - \frac{1}{2\tau^2}(\mu^2 - 2m\mu + m^2) \right\} \\ &\propto \exp\left\{ - \frac{n}{2\sigma^2}\mu^2 + \frac{n\bar{X}_n}{\sigma^2}\mu - \frac{1}{2\tau^2}\mu^2 + \frac{m}{\tau^2}\mu \right\} \\ &= \exp\Big(a\mu^2 - 2b\mu \Big), \quad a = -\frac{n}{2\sigma^2} - \frac{1}{2\tau^2}, \quad b = -\frac{n\bar{X}_n}{2\sigma^2} - \frac{m}{2\tau^2}. \end{aligned}$$

Completing the square gives

$$a\mu^2 - 2b\mu = a\Big(\mu - \frac{b}{a}\Big)^2 - \frac{b^2}{a},$$

so $p(\mu \mid X) \propto \exp\left\{ a\Big(\mu - \frac{b}{a}\Big)^2 - \frac{b^2}{a} \right\}$. Therefore the posterior mean and variance are

$$\mathsf{E}[\mu \mid X] = \frac{b}{a}, \quad \mathsf{Var}[\mu \mid X] = -\frac{1}{2a}.$$

Posterior moments. After simplification:

$$\begin{aligned}\mathsf{E}[\mu \mid X] &= \frac{b}{a} = \frac{-\frac{n\bar{X}_n}{2\sigma^2} - \frac{m}{2\tau^2}}{-\frac{n}{2\sigma^2} - \frac{1}{2\tau^2}} = \frac{\frac{n}{\sigma^2}\bar{X}_n + \frac{1}{\tau^2}m}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} = \bar{X}_n \cdot \frac{{}^{n}\!/\!{}_{\sigma^2}}{{}^{n}\!/\!{}_{\sigma^2} + {}^{1}\!/\!{}_{\tau^2}} + m \cdot \frac{{}^{1}\!/\!{}_{\tau^2}}{{}^{n}\!/\!{}_{\sigma^2} + {}^{1}\!/\!{}_{\tau^2}}\\ &= \bar{X}_n \cdot \frac{\tau^2}{\tau^2 + {}^{\sigma^2}\!/\!{}_{n}} + m \cdot \frac{{}^{\sigma^2}\!/\!{}_{n}}{\tau^2 + {}^{\sigma^2}\!/\!{}_{n}}\\ &= w\bar{X}_n + (1-w)m, \quad w = \frac{\tau^2}{\tau^2 + \sigma^2/n},\\ \mathsf{Var}[\mu \mid X] &= \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}} = \underbrace{\frac{\sigma^2}{n}}_{\text{Var of } \bar{X}_n} \cdot \frac{\tau^2}{\tau^2 + \sigma^2/n} < \frac{\sigma^2}{n}.\end{aligned}$$

Posterior precision. Defining precision as the inverse of variance, $\frac{1}{\text{variance}}$:

$$\frac{1}{\mathsf{Var}[\mu \mid X]} = \frac{n}{\sigma^2} + \frac{1}{\tau^2} = \text{precision of } \bar{X}_n + \text{prior precision}.$$

Interpretation.

- The posterior mean is a *convex combination* of the sample mean $\bar{X}_n$ and prior mean m, weighted by their respective precisions.
- The posterior variance is always smaller than the variance of $\bar{X}_n$, reflecting increased certainty after observing data.
- As $n \to \infty$, the posterior mean converges to $\bar{X}_n$ and the influence of the prior diminishes. Meaning the more data we have, the more certain we are about our estimate of μ.
- As $\tau^2 \to 0$, the prior dominates; as $\tau^2 \to \infty$, the prior becomes uninformative.

This example illustrates the convenience of conjugate priors: a Gaussian prior combined with Gaussian likelihood yields a Gaussian posterior, allowing closed-form Bayesian updating and an intuitive interpolation between prior beliefs and observed data.

Example 1.10 (Newcomb's speed of light data: Bayesian updating)**.** Newcomb collected several datasets measuring deviations (in millionths of a second) from the value 24.8 for the speed of light. The second dataset consists of 20 observations with sample mean 28.55 and sample standard deviation approximately 5.12 (see Section 1.3). Let's look at the third dataset, which consists of 26 observations, using the estimates from the second dataset to define a prior distribution for the unknown mean deviation μ.

Specifically, we take a normal prior

$$\mu \sim \mathcal{N}(\mu_0, \tau^2),$$

with prior mean $\mu_0 = 28.55$ and prior standard deviation $\tau = 2$. The value $\tau = 2$ is chosen as a conservative approximation to the standard error of the sample mean from the second dataset,

$$\frac{5.12}{\sqrt{20}} \approx 1.14,$$

rounded up to reflect additional uncertainty.

```r
# Prior parameters
mu0 <- 28.55
tau <- 2  # standard error of old sample mean (rounded up)
```

We now analyze the third dataset, consisting of $n = 26$ observations:

```r
data3 <- c(36,27,26,28,29,23,31,32,24,27,
           27,27,32,25,28,27,26,24,32,29,
           28,33,39,25,16,23)
n <- length(data3)   # n = 26
sigma <- 5           # assumed SD of measurement error
```

Assuming a normal observation model with known measurement error standard deviation $\sigma = 5$, conjugacy implies that the posterior distribution for μ is again normal. The posterior mean is a convex combination of the sample mean $\bar{x}_3$ and the prior mean μ_0, with weight

$$w = \frac{\tau^2}{\tau^2 + \sigma^2/n}.$$

The posterior standard deviation is given by

$$\sqrt{w}\,\frac{\sigma}{\sqrt{n}}.$$

```r
w <- tau^2 / (tau^2 + sigma^2 / n)
post_mean <- w * mean(data3) + (1 - w) * mu0
post_sd <- sqrt(w) * (sigma / sqrt(n))
c(post_mean, post_sd)
# [1] 27.98077 0.8804509
```

The posterior mean, 27.98, is slightly smaller than the prior mean, reflecting the lower average observed in the third dataset. At the same time, the posterior distribution is narrower than the prior, since the data provide additional information and reduce uncertainty.

Figure Figure 1.7 illustrates this Bayesian updating process: the prior distribution (blue) summarizes beliefs informed by earlier experiments, while the posterior distribution (red) combines these beliefs with the evidence provided by the new data.

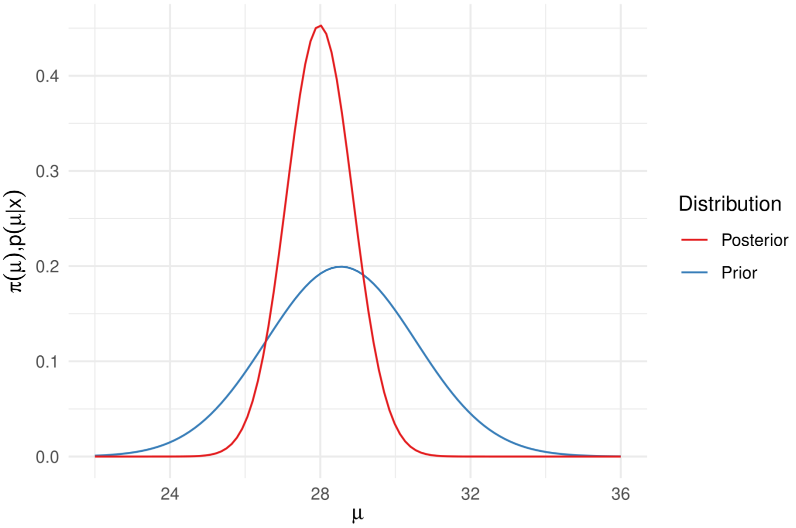


Figure 1.7: Prior and posterior distributions for the mean deviation μ based on Newcomb's third dataset.

```r
ggplot(data.frame(x = c(22, 36)), aes(x)) +
  geom_function(fun = dnorm,
                args = list(mean = mu0, sd = tau),
                aes(colour = "Prior")) +
  geom_function(fun = dnorm,
                args = list(mean = post_mean, sd = post_sd),
                aes(colour = "Posterior")) +
  scale_color_brewer(palette = "Set1") +
  labs(colour = "Distribution",
       x = expression(mu),
       y = expression(paste(pi(mu), ", ", p(mu * "|" * x))))
```

This example illustrates a key feature of Bayesian inference: information can be incorporated sequentially. Updating the prior using one dataset and then treating the resulting posterior as the prior for subsequent data yields the same result as combining all datasets at once, provided the observations are conditionally independent. The posterior distribution thus represents a complete summary of current knowledge, and any further inference—such as point estimates, credible intervals, or decision-making under loss—can be carried out directly from it.

Credible Intervals

> *Given the data and the prior assumptions, which values of μ are most plausible?*

In the Bayesian framework, interval estimates are obtained by identifying regions of the parameter space that carry a prescribed amount of posterior probability. Such intervals are called **credible intervals**.

For the posterior distribution of the mean deviation μ obtained in the Newcomb example, which is normal with mean `post_mean` and standard deviation `post_sd`, a 95% credible interval can be computed directly from the posterior quantiles:

```r
qnorm(c(0.025, 0.975), mean = post_mean, sd = post_sd)
# [1] 26.25691 29.70821
```

Thus, the interval (26.26, 29.71) contains μ with posterior probability 95%.

In contrast to confidence intervals, credible intervals do not rely on long-run frequency guarantees over repeated samples. Instead, they summarize uncertainty about μ directly through the posterior distribution. For this reason, credible intervals are specific to Bayesian inference and should not be confused with confidence intervals, even though in some cases their numerical values may be similar.

In the Gaussian case considered here, the posterior density is symmetric and unimodal. A natural choice of credible interval is therefore obtained by trimming 2.5% of the posterior probability mass from each tail, yielding the central 95% interval. Equivalently, this corresponds to selecting the 2.5th and 97.5th percentiles of the posterior distribution. For more general or skewed posterior distributions, one may instead consider highest posterior density (HPD) regions, which give the shortest interval containing a specified posterior probability.

In simple models such as the Gaussian–Gaussian setting, the posterior distribution can be derived in closed form, and posterior summaries such as means, variances, or credible intervals are easy to compute analytically. However, for more complex likelihood–prior combinations, the posterior density may not have a tractable normalization constant or an explicit formula.

In such cases, direct analytical calculation is replaced by simulation. If we are able to generate samples from the posterior distribution, then posterior expectations, quantiles, and credible intervals can be approximated numerically using Monte Carlo averages. This observation motivates the use of *Markov Chain Monte Carlo (MCMC)* methods, which allow us to sample from the posterior distribution whenever we can evaluate a function proportional to it. But *remember*, Bayesian computation relies on the identity

$$\text{posterior} \;\propto\; \text{likelihood} \times \text{prior}.$$

Once the posterior distribution is available—either analytically or via simulation—it represents a complete summary of current knowledge about the parameter. Point estimates (such as posterior means or medians), interval estimates (credible intervals), and decision-theoretic procedures can all be derived from it in a unified and coherent way.

1.6 Mean Square Error, Bias and Variance

Define

$$\mathrm{MSE}_\theta[\hat\theta] := \mathsf{E}_\theta[(\hat\theta - \theta)^2] = \int_{\mathcal{X}} (\theta(x) - \theta)^2 d\mathsf{P}_\theta(x).$$

Theorem 1.3. *The* ***mean square error*** *decomposes as*

$$MSE_\theta[\hat\theta] = (Bias_\theta[\hat\theta])^2 + \mathsf{Var}_\theta[\hat\theta],$$

where $Bias_\theta[\hat\theta] := \mathsf{E}_\theta[\hat\theta] - \theta$ is the ***bias*** *of $\hat\theta$.*

Proof. Write $\mathrm{MSE}_\theta[\hat\theta] = \mathsf{E}_\theta\left[(\hat\theta - \mathsf{E}_\theta[\hat\theta] + \mathsf{E}_\theta[\hat\theta] - \theta)^2\right]$ and expand

$$\begin{aligned}\mathrm{MSE}_\theta[\hat\theta] &= \mathsf{E}_\theta\left[(\hat\theta - \mathsf{E}_\theta[\hat\theta])^2\right] + 2\mathsf{E}_\theta\left[(\hat\theta - \mathsf{E}_\theta[\hat\theta])(\mathsf{E}_\theta[\hat\theta] - \theta)\right] + \mathsf{E}_\theta\left[(\mathsf{E}_\theta[\hat\theta] - \theta)^2\right] \\ &= \mathsf{Var}_\theta[\hat\theta] + \mathsf{E}_\theta\left[(\mathrm{Bias}_\theta[\hat\theta])^2\right] + 2\cdot \mathrm{Bias}_\theta[\hat\theta]\cdot \underbrace{\mathsf{E}_\theta\left[\hat\theta - \mathsf{E}_\theta[\hat\theta]\right]}_{=0} \\ &= \mathsf{Var}_\theta[\hat\theta] + (\mathrm{Bias}_\theta[\hat\theta])^2.\end{aligned}$$

□

Interpretation.
The bias–variance decomposition shows that the mean square error arises from two distinct sources:

- *Bias*, which captures systematic deviation of the estimator from the true parameter, and
- *Variance*, which quantifies the variability of the estimator across repeated samples.

An estimator with small MSE must therefore balance these two components: low bias alone is insufficient if variance is large, and conversely, low variance does not compensate for substantial systematic bias.

This trade-off is fundamental and holds universally, independent of the complexity of the estimator or the model under consideration. In many settings, especially in high-dimensional problems or with limited data, it can be advantageous to accept a small amount of bias in order to achieve a substantial reduction in variance.

Optimality with respect to mean square error

Question. Does there exist an estimator that is optimal in the sense of minimizing the mean square error?

More precisely, for a fixed $\theta \in \Theta$, which estimator minimizes

$$\mathrm{MSE}_\theta[\hat\theta]?$$

While for a fixed value of θ such an estimator can always be constructed, there exists *no* estimator that minimizes $\mathrm{MSE}_\theta[\hat\theta]$ uniformly over all $\theta \in \Theta$, provided the statistical model contains at least two distinct distributions.

Explanation. Fix $\theta \in \Theta$. The constant estimator $\hat\theta(x) \equiv \theta$ has mean square error

$$\mathrm{MSE}_\theta[\hat\theta] = 0,$$

since it has zero bias and zero variance under P_θ. Thus, for each individual value of θ, there exists an estimator that is optimal at that specific point.

Suppose, however, that there were an estimator $\hat{\theta}^\star$ that minimized $\mathrm{MSE}_\theta[\hat{\theta}]$ simultaneously for all $\theta \in \Theta$. Then $\hat{\theta}^\star$ would have to achieve zero mean square error for every θ, since it must outperform all constant estimators. In particular,

$$\mathrm{MSE}_\theta[\hat{\theta}^\star] = 0 \quad \text{for all } \theta.$$

By the bias–variance decomposition, this implies both

$$\mathsf{Var}_\theta[\hat{\theta}^\star] = 0 \quad \text{and} \quad \mathrm{Bias}_\theta[\hat{\theta}^\star] = 0 \quad \text{for all } \theta.$$

Zero variance implies that $\hat{\theta}^\star$ is almost surely constant under every P_θ. However, a constant estimator cannot be unbiased for more than one value of θ. This contradiction shows that a uniformly MSE-optimal estimator cannot exist.

Consequences. Since uniform optimality with respect to mean square error is impossible, optimality theory proceeds by modifying the criterion. Common approaches include:

- Restricting attention to a class of estimators with additional properties, such as unbiasedness or equivariance, and seeking optimality within that class.
- Summarizing the function $\theta \mapsto \mathrm{MSE}_\theta[\hat{\theta}]$ by a single number, for example by averaging with respect to a probability measure on Θ (leading to Bayes risk), or by considering the worst-case value (minimax risk).

Gaussian mean: sample mean versus Bayes estimator

Consider the Gaussian location model $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known. We compare two estimators of the unknown mean μ and study their mean square error (MSE) as a function of μ.

1. *Sample mean:* $\hat{\mu}_1 = \overline{X}_n$. Since $\mathsf{E}_\mu[\overline{X}_n] = \mu$, the sample mean is unbiased:

$$\mathrm{Bias}_\mu[\hat{\mu}_1] = 0.$$

Its variance is $\mathsf{Var}_\mu[\overline{X}_n] = \sigma^2/n$, and therefore

$$\mathrm{MSE}_\mu[\hat{\mu}_1] = \mathsf{Var}_\mu[\hat{\mu}_1] = \frac{\sigma^2}{n}.$$

The MSE of the sample mean is constant in μ: it depends only on the noise level σ^2 and the sample size n.

2. *Bayes estimator:* now consider a Bayesian estimator obtained as the posterior mean under a Gaussian prior $\mu \sim \mathcal{N}(m, \tau^2)$ with $m = 0$ and $\tau^2 = 1$.

The resulting Bayes estimator $\hat{\mu}_2$ is a shrinkage estimator that pulls the sample mean toward the prior mean. A direct calculation yields

$$\mathrm{Bias}_\mu[\hat{\mu}_2] = -\frac{\mu\,\sigma^2}{n + \sigma^2}, \qquad \mathsf{Var}_\mu[\hat{\mu}_2] = \frac{n\sigma^2}{(n + \sigma^2)^2}.$$

Unlike the sample mean, the Bayes estimator is biased, but its variance is smaller. Combining bias and variance gives the mean square error

$$\mathrm{MSE}_\mu[\hat{\mu}_2] = \frac{\sigma^2(\mu^2\sigma^2 + n)}{(n + \sigma^2)^2},$$

which is a quadratic function of μ.

Comparison of MSE curves. To illustrate the bias–variance trade-off, consider $n = 10$ and $\sigma^2 = 1$. Figure 1.8 shows the MSE of both estimators as a function of μ.

```r
ggplot(data.frame(x = c(-3, 3)), aes(x)) +
  geom_function(fun = function(mu) 1 / 10,
                aes(colour = "Sample mean")) +
  geom_function(fun = function(mu) (mu^2 + 10) / (10 + 1)^2,
                aes(colour = "Bayes")) +
  scale_color_brewer(palette = "Set1") +
  labs(colour = "Estimator",
       x = expression(mu),
       y = "MSE")
```

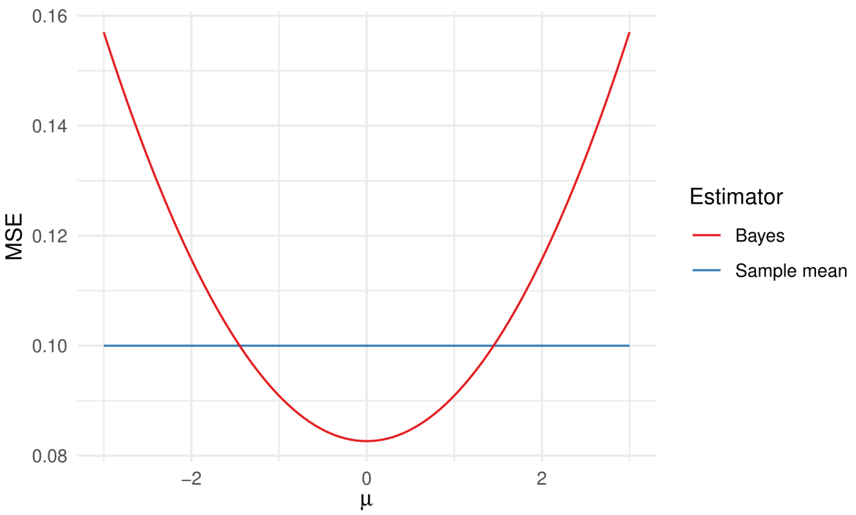


Figure 1.8: Mean square error as a function of μ for the sample mean and a Bayes estimator with prior $\mathcal{N}(0, 1)$.

The sample mean has constant MSE and is equally accurate for all values of μ. In contrast, the Bayes estimator performs better when μ is close to the prior mean, where the reduction in variance outweighs the introduced bias. However, when μ is far from the prior mean, the bias dominates and the MSE increases substantially. This example illustrates how incorporating prior information can reduce MSE locally, at the price of worse performance when the prior is badly misspecified.

1.7 Discussion

- **Plug–in principle.**
  - A general and flexible approach for estimating complex functionals by replacing unknown distributions with their empirical counterparts.
  - Applicable beyond simple examples, e.g. for measuring dependence via Hoeffding's statistic

    $$D = \int \big(F(x,y) - F(x)F(y)\big)^2 \, \mathrm{d}F(x,y),$$

    which vanishes under independence and is positive for nonlinear dependence.
- **Maximum likelihood estimation (MLE).**
  - MLEs remain useful even when closed-form solutions are unavailable and numerical optimization is required.
  - In regular models, asymptotic theory ensures optimality of the MLE for large sample sizes.
  - Asymptotic normality allows uncertainty quantification (confidence intervals, p-values) using Hessian information at the optimum.
  - This methodology underlies standard output in statistical software (e.g. regression models).
- **Method of moments (MoM).**
  - Some models admit simple, closed-form MoM estimators when MLEs are difficult or computationally expensive.
  - Particularly useful for theoretical analysis, such as deriving lower bounds on estimation accuracy.
  - While sometimes suboptimal, MoM estimators can provide insight and practical solutions.
- **Bayesian estimation.**
  - Especially valuable when data are scarce and prior information can meaningfully improve inference.
  - Suitable priors can encode low-dimensional structure in high-dimensional problems.
  - The Bayesian framework naturally provides uncertainty quantification through the posterior distribution.

2. Sufficiency

2.1 Sufficient Statistics

Setting.

- Sample space: $\mathcal{X} \subset \mathbb{R}^n$, equipped with its Borel σ–algebra.
- Observation: a random element $\boldsymbol{X}$ taking values in $\mathcal{X}$ with distribution $\boldsymbol{X} \sim \mathsf{P}_\theta$.
- Statistic: a measurable mapping $T : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}^k$ is equipped with its Borel σ–algebra.

Definition 2.1 (Sufficient statistic)**.** A statistic T is called **sufficient** for the model $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ if there exists a version of the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X}) = t$ that does *not* depend on the parameter θ.

Equivalently, once the value $T(\boldsymbol{X})$ is known, the remaining randomness in $\boldsymbol{X}$ has a distribution that is the same for all $\theta \in \Theta$.

Interpretation and motivation. Sufficiency is a property of a statistic relative to a given statistical model. Intuitively, a statistic is sufficient if it retains all information in the data that is relevant for inference about the parameter θ. In other words, no additional knowledge of the full data set beyond $T(\boldsymbol{X})$ can improve estimation, testing, or the construction of confidence intervals for θ.

A useful way to think about this is through a thought experiment. Suppose the original data set $\boldsymbol{X}$ is lost, but the value of the statistic $T(\boldsymbol{X})$ is retained. If T is sufficient, then one can generate a new random data set $\widetilde{\boldsymbol{X}}$ by sampling from the conditional distribution of $\boldsymbol{X}$ given $T(\boldsymbol{X})$. Because this conditional distribution does not depend on θ, the reconstructed data $\widetilde{\boldsymbol{X}}$ has the same distribution as the original data $\boldsymbol{X}$ under *every* $\mathsf{P}_\theta \in \mathcal{P}$. Thus, although the original data cannot be recovered exactly, it can be reconstructed in a probabilistic sense without loss of information about the parameter.

From this perspective, sufficiency can be viewed as a form of information reduction: the statistic $T(\boldsymbol{X})$ may be lower dimensional than the full data, yet it contains all information relevant for inference within the specified model. This notion is related to, but distinct from, ideas such as data compression or storage efficiency.

Further reading. See Casella and Berger (2002, Section 6.2) for additional discussion and examples.

Example 2.1 (Bernoulli distribution)**.** Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(θ) r.v. with $\theta \in (0,1)$. We interpret $X_i = 1$ as a success (e.g. heads) and $X_i = 0$ as a failure (tails). The sample space is therefore $\mathcal{X} = \{0,1\}^n$.

A natural statistic in this setting is the total number of successes,

$$T(X_1, \ldots, X_n) \;=\; \sum_{i=1}^{n} X_i,$$

which counts how often the outcome 1 is observed. Intuitively, if the goal is to learn about θ, the probability of success in a single trial, then only the total number of successes should matter, not the specific order in which they occur. We now verify that this intuition is correct.

Claim. The statistic $T(X_1, \ldots, X_n) = \sum\limits_{i=1}^{n} X_i$ is sufficient for the Bernoulli(θ) model.

Proof. Consider the conditional distribution of the full data $(X_1, \ldots, X_n)$ given that $T(X_1, \ldots, X_n) = t$. This conditional probability is nonzero only for sequences $(x_1, \ldots, x_n) \in \{0,1\}^n$ satisfying $\sum\limits_{i=1}^{n} x_i = t$. For such a sequence, we have

$$\mathsf{P}_\theta\left(X_1 = x_1, \ldots, X_n = x_n \middle| \sum_{i=1}^{n} X_i = t\right) = \frac{\mathsf{P}_\theta(X_1 = x_1, \ldots, X_n = x_n)}{\mathsf{P}_\theta(\sum_{i=1}^{n} X_i = t)}.$$

Since the X_i are i.i.d. Bernoulli(θ),

$$\mathsf{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^t (1-\theta)^{n-t}.$$

Moreover, $\sum\limits_{i=1}^{n} X_i$ has a Binomial(n, θ) distribution, so

$$\mathsf{P}_\theta\left(\sum_{i=1}^{n} X_i = t\right) \;=\; \binom{n}{t} \theta^t (1-\theta)^{n-t}.$$

Dividing yields

$$\mathsf{P}_\theta\left(X_1 = x_1, \ldots, X_n = x_n \mid \sum_{i=1}^{n} X_i = t\right) = \frac{1}{\binom{n}{t}},$$

which does not depend on θ.

Thus, conditional on $T(\boldsymbol{X}) = t$, all sequences in $\{0,1\}^n$ containing exactly t ones are equally likely (uniform distribution). The conditional distribution of $(X_1, \ldots, X_n)$ given $T(\boldsymbol{X})$ is therefore independent of θ, so T is sufficient. □

Interpretation. Knowing only the total number of successes t, one can probabilistically reconstruct a data set by placing t ones uniformly at random among the n positions. This reconstructed data has the same distribution as the original sample under any P_θ. Hence, for inference about θ in the i.i.d. Bernoulli model, the statistic $T = \sum\limits_{i=1}^{n} X_i$ retains all relevant information in the data.

Example 2.2 (Poisson model and sufficiency)**.** Let $X_1, \dots, X_n$ be i.i.d. Poisson(θ) r.v. with $\theta > 0$. The sample space is $\mathcal{X} = \mathbb{N}_0^n$. A typical interpretation is that X_i counts the number of events (e.g. earthquakes, arrivals, failures) observed in the ith time period or region.

A natural statistic in this model is the total count

$$T(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i,$$

which aggregates all observed events across the n independent observations. Intuitively, if the goal is to learn the underlying rate θ, only the total number of events should matter, not how they are distributed across the individual observations.

Claim. The statistic $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$ is sufficient for the Poisson(θ) model.

Proof. Since the sum of independent Poisson r.v. is again Poisson,

$$T(X_1, \dots, X_n) \sim \text{Poisson}(n\theta).$$

Consider the conditional distribution of $(X_1, \dots, X_n)$ given $T(X_1, \dots, X_n) = t$. This conditional probability is nonzero only for vectors $(x_1, \dots, x_n) \in \mathbb{N}_0^n$ satisfying $\sum\limits_{i=1}^{n} x_i = t$. For such a vector,

$$\begin{aligned}\mathsf{P}_\theta\left(X_1 = x_1, \dots, X_n = x_n \,\middle|\, \sum_{i=1}^{n} X_i = t \right) &= \frac{\prod\limits_{i=1}^{n} e^{-\theta} \cdot \theta^{x_i}/x_i!}{e^{-n\theta} \cdot (n\theta)^t/t!} \\ &= \frac{t!}{x_1! \cdots x_n!} \left(\frac{1}{n}\right)^t = \binom{t}{x_1, \dots, x_n} \left(\frac{1}{n}\right)^t.\end{aligned}$$

This is the probability mass function of a multinomial distribution with parameters t and cell probabilities $(1/n, \dots, 1/n)$, and it does not depend on θ. Hence, the conditional distribution of $(X_1, \dots, X_n)$ given T is independent of θ, so T is sufficient. □

Interpretation. Given the total count t, the individual counts $(X_1, \dots, X_n)$ can be viewed as obtained by randomly allocating t events among n categories, each with equal probability. Thus, once T is known, the remaining randomness in the data is purely combinatorial and unrelated to the unknown rate θ.

Example 2.3 (Continuous i.i.d. model and order statistics)**.** Let $X_1, \dots, X_n$ be i.i.d. real-valued r.v. with a continuous cdf F on $\mathbb{R}$. Here, the parameter of interest is the distribution function itself, $\theta \equiv F$.

Define the *order statistics*

$$T(X_1, \dots, X_n) = (X_{(1)}, \dots, X_{(n)}),$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ is the sorted version of the data. This statistic discards the order in which the observations were collected, retaining only their relative magnitudes.

Claim. The vector of order statistics $(X_{(1)}, \dots, X_{(n)})$ is sufficient for the i.i.d. model with continuous cdf F.

Proof. Fix distinct real numbers $x_1, \dots, x_n$ with order statistics $(x_{(1)}, \dots, x_{(n)})$. Conditioning on the event that the sorted values equal $(x_{(1)}, \dots, x_{(n)})$, each of the $n!$ possible permutations of these values is equally likely. Hence,

$$\mathsf{P}_F\big(X_1 = x_1, \dots, X_n = x_n \,\big|\, X_{(1)} = x_{(1)}, \dots, X_{(n)} = x_{(n)}\big) = \frac{1}{n!},$$

which does not depend on F. Therefore, the conditional distribution of $(X_1, \dots, X_n)$ given the order statistics is independent of the underlying distribution function, and the order statistics are sufficient. □

Interpretation. In an i.i.d. continuous model, the labeling or order of the observations carries no information about the distribution F. If the original ordering is lost but the sorted data are retained, one can reconstruct a probabilistically equivalent data set by randomly permuting the order statistics. Thus, all information relevant for inference about F is contained in the order statistics.

Further reading. A formal measure-theoretic treatment of conditioning on order statistics can be found in Lehmann and Romano (2005, Example 2.4.1).

2.2 Neyman–Fisher Factorization Criterion

In many statistical models, all distributions admit a density (or probability mass function) with respect to a common dominating measure. In this setting, sufficiency can be characterized directly from the form of the likelihood, without explicit computation of conditional distributions.

Theorem 2.1 (Neyman–Fisher factorization criterion)**.** *Let $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ be a statistical model dominated by a σ–finite measure ν. Denote the corresponding densities by*

$$p_\theta(x) = \frac{d\mathsf{P}_\theta}{d\nu}(x), \qquad \theta \in \Theta.$$

A statistic $T : \mathcal{X} \to \mathcal{Y}$ is sufficient for $\mathcal{P}$ if and only if there exist measurable non–negative functions $h : \mathcal{X} \to \mathbb{R}$ and $g_\theta : \mathcal{Y} \to \mathbb{R}$ such that, for all $\theta \in \Theta$,

$$p_\theta(x) = g_\theta(T(x))\, h(x), \qquad x \in \mathcal{X}. \tag{2.1}$$

Interpretation. The factorization (2.1) separates the dependence on the parameter θ from the full data x. All information about θ contained in the data appears through the statistic $T(x)$, while the remaining factor $h(x)$ depends only on the observed sample. Consequently, once $T(x)$ is known, no further information about θ can be extracted from the data.

Remark. If (2.1) holds, any likelihood–based procedure depends on the data only through $T(x)$. In particular, the maximum likelihood estimator satisfies

$$\hat{\theta}(x) = \arg\max_{\theta} p_\theta(x) = \arg\max_{\theta} g_\theta(T(x)),$$

and is therefore a function of the sufficient statistic T alone.

Proof (Theorem 2.1, discrete case). We prove the result when ν is the counting measure. Let

$$\mathsf{Q}_\theta(t) := \mathsf{P}_\theta(T = t) = \sum_{x\in\mathcal{X}\,:\,T(x)=t} p_\theta(x), \qquad t \in \mathcal{Y},$$

denote the probability mass function of T. For $x \in \mathcal{X}$ with $T(x) = t$, the conditional distribution of X given $T = t$ is

$$p_\theta(x \mid T = t) = \frac{p_\theta(x)}{\mathsf{Q}_\theta(t)}.$$

$\Longrightarrow$) Suppose T is sufficient. Then the conditional distribution $p_\theta(x \mid T = t)$ does not depend on θ. Let's denote this common conditional distribution by $p_*(x \mid T = t)$. Using the identity

$$p_\theta(x) = \mathsf{P}_\theta(T = t)\, p_*(x \mid T = t), \qquad t = T(x),$$

we get

$$\begin{aligned} p_\theta(x) &= \mathsf{P}_\theta(X = x) \\ &= \mathsf{P}_\theta(X = x, T = T(x)) \\ &= Q_\theta(T(x)) \cdot p_*(x \mid T = T(x)) \\ &=: g_\theta(T(x)) \cdot h(x). \end{aligned}$$

and so we obtain the factorization

$$p_\theta(x) = g_\theta(T(x))\, h(x), \qquad g_\theta(t) := \mathsf{Q}_\theta(t), \quad h(x) := p_*(x \mid T = T(x)).$$

$\Longleftarrow$) Conversely, suppose (2.1) holds. Then

$$Q_\theta(t) = \sum_{x\in\mathcal{X}\,:\,T(x)=t} p_\theta(x) = \sum_{x\in\mathcal{X}\,:\,T(x)=t} g_\theta(T(x))h(x) = g_\theta(t) \sum_{x\in\mathcal{X}\,:\,T(x)=t} h(x)$$

Therefore, for x and t with $T(x) = t$, it holds that

$$\begin{aligned} p_\theta(x \mid T(x) = t) &= \frac{g_\theta(t)h(x)}{g_\theta(t) \sum\limits_{x\in\mathcal{X}\,:\,T(x)=t} h(x)} \\ &= \frac{h(x)}{\sum\limits_{x\in\mathcal{X}\,:\,T(x)=t} h(x)} \end{aligned}$$

does not depend on θ. We conclude that T is sufficient.

□

Further reading. The theorem holds for general dominated models. A measure–theoretic treatment of conditioning and sufficiency can be found in Lehmann and Romano (2005, Sections 1.9 and 2.6).

Example 2.4 (Uniform model and the German tank problem)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables with distribution Uniform$(0, \theta)$, where $\theta > 0$ is unknown. This model is a continuous analogue of the German tank problem: observing n labels drawn uniformly from $[0, \theta]$, the goal is to infer the unknown upper endpoint θ. (How many tanks do the Germans have if they label them in order, and I have observed n different labels so far?)

Claim. The statistic

$$T(X_1, \ldots, X_n) = \max_{1 \le i \le n} X_i$$

is sufficient for θ.

Proof. The joint density of $X_1, \ldots, X_n$ is

$$\begin{aligned} p_\theta(x_1, \ldots, x_n) &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}_{[0,\theta]}(x_i) \\ &= \underbrace{\frac{1}{\theta^n} \mathbf{1}_{[0,\theta]}(\max_i x_i)}_{g_\theta(T(x))} \underbrace{\mathbf{1}_{[0,\infty)}(\min_i x_i)}_{h(x)} . \end{aligned}$$

By the Neyman–Fisher factorization theorem, T is sufficient. □

Remark.

- Intuitively, in the German tank problem only the largest observed label is informative about the unknown maximum; smaller observations carry no additional information about θ.
- The order statistics are also sufficient in this model. More generally, if a statistic is sufficient for a model $\mathcal{P}$, then it is sufficient for any submodel $\mathcal{P}' \subset \mathcal{P}$.
- Hence, since order statistics are sufficient for the model of i.i.d. observations from an arbitrary continuous distribution, they remain sufficient for the uniform submodel.

2.3 Many Sufficient Statistics

Sufficiency is generally not unique: a given statistical model typically admits many different sufficient statistics. This follows from the fact that sufficiency is preserved under suitable transformations.

Suppose T and T' are two statistics such that

$$T(X) = S\big(T'(X)\big)$$

for some measurable mapping S.

- When Neyman's factorization criterion applies, if T is sufficient for the model, then T' is also sufficient. Indeed, if the likelihood depends on the data only through $T(X)$, and $T(X)$ itself can be recovered from $T'(X)$, then all information about the parameter contained in T is also contained in T'. This agrees with the interpretation of sufficiency given in section 2.1.
- If S is bijective, then T is sufficient if and only if T' is sufficient. In this case, conditioning on T or on T' provides exactly the same information, since each statistic can be reconstructed from the other.

These principles are illustrated by several earlier examples. In Example 2.1 and Example 2.2, the sum $\sum_{i=1}^{n} X_i$ was identified as a sufficient statistic. Equivalently, one could use the sample mean

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

since it is a one–to–one transformation of the sum. More generally, any statistic from which the sum can be recovered is sufficient; for instance, the pair

$$\big(X_1 + X_2,\ \sum_{i=3}^{n} X_i\big)$$

is also sufficient.

In Example 2.3, the vector of order statistics was shown to be sufficient. An equivalent sufficient statistic is the empirical distribution function $\widehat{F}_n$, since the order statistics determine the jump locations of $\widehat{F}_n$, and conversely the empirical distribution function uniquely determines the ordered sample values.

Finally, in Example 2.4, note that the sample maximum

$$\max_i X_i = X_{(n)}$$

is a function of the order statistics, and hence inherits sufficiency whenever the full vector of order statistics is sufficient.

Question: Which sufficient statistics provide greatest data reduction?

2.3.1 Minimal Sufficiency

Sufficient statistics summarize the data without loss of information about the unknown parameter θ. Since sufficiency alone is not unique (for example, the identity map $T(X) = X$ is always sufficient), it is natural to ask whether one can reduce the data *as much as possible* while still retaining all information about θ. This idea leads to the notion of *minimal sufficiency.*

Definition 2.2. A sufficient statistic T is called **minimal sufficient** if T is a function of every other sufficient statistic T'. More precisely, T is minimal sufficient if

$$\exists \text{ a measurable map } S \text{ such that } T(x) = S\big(T'(x)\big) \quad \text{for almost every } x \in \mathcal{X}.$$

Here, "almost every" means that there exists a set $\mathcal{X}^* \subseteq \mathcal{X}$ such that $\mathsf{P}_\theta(\mathcal{X}^*) = 1$ for all $\theta \in \Theta$, and the above equality holds for all $x \in \mathcal{X}^*$.

The existence of such a map S is equivalent to the condition

$$T'(x) = T'(\tilde{x}) \implies T(x) = T(\tilde{x}) \quad \forall x, \tilde{x} \in \mathcal{X}^*. \tag{2.2}$$

Thus, a minimal sufficient statistic cannot be refined further without losing information about θ.

Condition (2.2) shows that minimal sufficiency depends only on how statistics identify data sets: whenever two data sets are indistinguishable under a sufficient statistic T', they must also be indistinguishable under T. In this sense, T induces a coarser classification of the data than any other sufficient statistic.

Partition viewpoint. Any statistic T induces a partition of the sample space,

$$\mathcal{X} = \bigcup_t \{x \in \mathcal{X} : T(x) = t\},$$

where two data sets are grouped together if they yield the same value of T. From this perspective:

- Minimal sufficient statistics correspond to the *coarsest possible* partitions among all sufficient statistics (up to null sets).
- Any two minimal sufficient statistics T and T' induce the same partition of $\mathcal{X}$ (up to null sets), and hence satisfy (2.2) with $\iff$.
- One-to-one transformations of a statistic do not affect sufficiency or minimal sufficiency, since they leave the induced partition unchanged.

2.3.2 Criterion for Minimal Sufficiency

The definition of minimal sufficiency is abstract, as it quantifies over *all* sufficient statistics. In dominated models, however, there is a powerful and practical characterization that allows us both to *verify* minimal sufficiency and to *construct* minimal sufficient statistics directly from the likelihood.

Theorem 2.2 (Criterion for minimal sufficiency)**.** *Let $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ be dominated by a σ–finite measure ν, with densities*

$$p_\theta(x) = \frac{d\mathsf{P}_\theta}{d\nu}(x).$$

Define the support of the model by

$$\operatorname{supp}(\mathcal{P}) = \{x \in \mathcal{X} : \exists\theta \in \Theta \text{ such that } p_\theta(x) > 0\}.$$

Suppose that a statistic T satisfies the following property: for all $x, \tilde{x} \in \operatorname{supp}(\mathcal{P})$,

$$T(x) = T(\tilde{x}) \iff \exists\, c(x,\tilde{x}) > 0 \text{ such that } p_\theta(x) = p_\theta(\tilde{x}) \cdot c(x,\tilde{x}) \quad \forall\theta \in \Theta. \tag{2.3}$$

Then T is minimal sufficient for $\mathcal{P}$.

Interpretation. The right–hand side of (2.3) states that the likelihood functions corresponding to x and $\tilde{x}$ are proportional as functions of θ. Thus, two data sets receive the same value of T if and only if they induce the same likelihood up to a multiplicative constant. This criterion therefore identifies minimal sufficient statistics as those that classify data sets exactly according to likelihood equivalence.

Comments

- Condition (2.3) provides a constructive way to identify minimal sufficient statistics by analyzing when likelihood functions are proportional.
- The criterion depends only on likelihoods and does not require computing conditional distributions.
- The original result is due to Lehmann and Scheffé (1950, Theorem 6.3).

Proof (Discrete case). We show that T is sufficient and minimal sufficient.

Step 1: T is sufficient. For each value t of T, choose a representative element

$$x_t \in T^{-1}(\{t\}) = \{x : T(x) = t\},$$

with $x_t \in \operatorname{supp}(\mathcal{P})$ whenever $T^{-1}(\{t\}) \cap \operatorname{supp}(\mathcal{P}) \neq \emptyset$.

For any $x \in \mathcal{X}$, we have $T(x) = T(x_{T(x)})$, so by assumption (2.3),

$$p_\theta(x) = p_\theta(x_{T(x)})\, c(x, x_{T(x)}), \quad \theta \in \Theta,$$

where we define $c(x, x_{T(x)}) = 0$ if $x \notin \operatorname{supp}(\mathcal{P})$. This representation expresses $p_\theta(x)$ as a product of a term depending on θ only through $T(x)$ and a term independent of θ. Hence, by Neyman's factorization criterion, T is sufficient.

Step 2: T is minimal sufficient. Let T' be any other sufficient statistic and let $x, \tilde{x} \in \operatorname{supp}(\mathcal{P})$ satisfy $T'(x) = T'(\tilde{x})$. We must show that $T(x) = T(\tilde{x})$.

Since T' is sufficient, the factorization criterion yields

$$p_\theta(x) = g'_\theta(T'(x))h'(x),$$
$$p_\theta(\tilde{x}) = g'_\theta(T'(\tilde{x}))h'(\tilde{x}).$$

Because $T'(x) = T'(\tilde{x})$, the g'_θ terms coincide, and thus

$$p_\theta(x) = p_\theta(\tilde{x}) \cdot \frac{h'(x)}{h'(\tilde{x})} \quad \forall \theta \in \Theta. \tag{2.4}$$

Since $\tilde{x} \in \operatorname{supp}(\mathcal{P})$, we have $h'(\tilde{x}) > 0$, so the ratio is well defined. By condition (2.3), this proportionality implies $T(x) = T(\tilde{x})$.

Therefore, T is a function of every sufficient statistic T', and hence T is minimal sufficient. □

Example 2.5 (Gaussian model)**.** Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with parameter $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$. Recall that the maximum likelihood estimators are

$$\hat{\mu} = \overline{X}_n, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Claim. The statistic $(\hat{\mu}, \hat{\sigma}^2)$ is minimal sufficient.

Proof. The joint density of $\boldsymbol{X} = (X_1, \dots, X_n)$ is

$$p_\theta(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right\}.$$

Using the identity

$$\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x}_n + \bar{x}_n - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \mu)^2,$$

we obtain

$$p_\theta(x) \propto \exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{x}_n - \mu)^2 + \hat{\sigma}_x^2\right]\right\}.$$

Let $x, y \in \mathbb{R}^n$ be two data sets. The ratio of likelihoods is

$$\begin{aligned}\frac{p_\theta(x)}{p_\theta(y)} &= \exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{x}_n - \mu)^2 - (\bar{y}_n - \mu)^2 + \hat{\sigma}_x^2 - \hat{\sigma}_y^2\right]\right\} \\ &= \exp\left\{-\frac{n}{2\sigma^2}\left[(\bar{x}_n^2 - \bar{y}_n^2) - 2\mu(\bar{x}_n - \bar{y}_n) + \hat{\sigma}_x^2 - \hat{\sigma}_y^2\right]\right\}.\end{aligned}$$

This ratio is independent of $\theta = (\mu, \sigma^2)$ if and only if

$$\bar{x}_n = \bar{y}_n \quad \text{and} \quad \hat{\sigma}_x^2 = \hat{\sigma}_y^2.$$

By the minimal sufficiency criterion of subsection 2.3.2, $(\hat{\mu}, \hat{\sigma}^2)$ is minimal sufficient. □

Example 2.6 (Uniform location model)**.** Let $X_1, \dots, X_n$ be i.i.d. Uniform$(\theta, \theta + 1)$ for $\theta \in \mathbb{R}$. Applying the minimal sufficiency criterion yields

$$T(x) = \Big(\min\{x_1, \dots, x_n\}, \max\{x_1, \dots, x_n\}\Big)$$

as a minimal sufficient statistic; see Casella and Berger (2002, Example 6.2.15) for details.

This example is noteworthy because the model is 1–dimensional, yet the minimal sufficient statistic is 2–dimensional. In fact, there are many low–dimensional models for which the order statistics are minimal sufficient; see Lehmann and Casella (1998, Example 1.6.15).

3. Exponential Families

Exponential families form a broad and important class of statistical models with many desirable theoretical and computational properties. They provide a unifying framework for a wide range of commonly used distributions and play a central role in statistical inference.

A key reason for the popularity of exponential family models is their favorable optimization structure. In many cases, parameter estimation reduces to maximizing a concave log-likelihood (or equivalently, minimizing a convex objective function), which leads to stable and efficient numerical algorithms.

Several widely used models arise naturally as members of exponential families. For example, logistic regression for binary data is based on a specific parametric form that links predictors to event probabilities through log-odds. This formulation is not only interpretable but also ensures that estimation involves a convex optimization problem. Similarly, the Poisson distribution is a standard choice for modeling count data, partly due to its mathematical convenience and tractable likelihood.

More generally, exponential families provide a common structure underlying these models. Many results that are traditionally derived on a case-by-case basis can be obtained more systematically by working within this general framework. Throughout this course, exponential families will serve as an important source of examples and will be revisited in various contexts.

We now introduce the formal definition of exponential families and discuss their main properties.

3.1 Definition

Definition 3.1. Let $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ be a statistical model on $(\mathcal{X}, \mathcal{A})$ that is dominated by a σ–finite measure ν. The model $\mathcal{P}$ is called an **exponential family** if there exist an integer $k \in \mathbb{N}$ and functions

$$\begin{aligned} B &: \Theta \to \mathbb{R}, & \eta &: \Theta \to \mathbb{R}^k, \\ T &: \mathcal{X} \to \mathbb{R}^k, & h &: \mathcal{X} \to [0, \infty), \end{aligned}$$

such that each distribution P_θ admits a density $p_\theta(x) = \frac{d\mathsf{P}_\theta}{d\nu}(x)$ of the form

$$p_\theta(x) = \exp\{\langle \eta(\theta), T(x) \rangle - B(\theta)\}\, h(x), \quad x \in \mathcal{X} \tag{3.1}$$

where

$$\langle \eta(\theta), T(x) \rangle = \sum_{i=1}^{k} \eta_i(\theta) T_i(x).$$

This representation expresses the log–density as a linear function of the statistic $T(x)$, with coefficients $\eta(\theta)$ depending on the parameter. In particular,

$$\log p_\theta(x) = \langle \eta(\theta), T(x) \rangle - B(\theta) + \log h(x),$$

which simplifies likelihood-based inference by reducing it to linear and convex-analytic operations.

The map $\eta(\theta)$ is called the **natural parameter**, and T is the **sufficient statistic** of the family (by Neyman's factorization criterion). The function $B(\theta)$ acts as a normalizing term ensuring that p_θ integrates to one, while $h(x)$ modifies the dominating measure ν.

The integer k is referred to as the **dimension** of the representation and is, in general, not unique (because the dimension of sufficient statistics of a family aren't). A statistical model is a k–dimensional exponential family if all its densities admit a representation of the form (3.1) with the same statistic T.

Example 3.1 (Binomial model)**.** Let $X \sim \text{Binomial}(n, \theta)$ with $\theta \in (0, 1)$. The distribution of X admits a density with respect to the counting measure on $\mathcal{X} = \{0, 1, \ldots, n\}$ given by

$$p_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Writing the parameter–dependent part on the log scale yields

$$p_\theta(x) = \binom{n}{x} \exp\{x \log\theta + (n-x)\log(1-\theta)\},$$

which is of exponential family form. This representation corresponds to a 2–dimensional exponential family with sufficient statistic $T(x) = (x, n-x)$ and natural parameters $\eta(\theta) = (\log\theta, \log(1-\theta))$.

The representation is, however, redundant since the components of $T(x)$ satisfy the linear relation $x + (n-x) = n$. Using this relation, the density can be rewritten as

$$p_\theta(x) = \binom{n}{x} \exp\left\{x \log\frac{\theta}{1-\theta} + n\log(1-\theta)\right\}.$$

This yields a 1–dimensional representation of the Binomial model, with sufficient statistic $T(x) = x$ and natural parameter

$$\eta(\theta) = \log\frac{\theta}{1-\theta}.$$

The quantity $\theta/(1-\theta)$ is called the *odds* associated with the success probability θ. More generally, for an event A with probability $P(A) = \theta$, the odds are defined as the ratio

$$\frac{P(A)}{P(A^c)} = \frac{\theta}{1-\theta}.$$

While probabilities take values in $(0,1)$, odds range over $(0,\infty)$. Taking logarithms yields the *log–odds*

$$\log\frac{\theta}{1-\theta},$$

which form the natural parameter $\eta(\theta)$ of the Binomial exponential family in its one–dimensional representation. This representation highlights the connection between the Binomial model and generalized linear models, in particular logistic regression.

Example 3.2 (Multinomial model as an exponential family)**.** Consider n independent repetitions of a random experiment with s possible outcomes. Let $X_i \geq 0$ denote the number of times outcome i is observed, and define

$$\boldsymbol{X} = (X_1,\dots,X_s), \qquad \sum_{i=1}^{s} X_i = n.$$

Assume that the probability of outcome i in a single trial is $\theta_i > 0$, with $\sum\limits_{i=1}^{s}\theta_i = 1$. Then

$$\boldsymbol{X} \sim \text{Multinomial}(n,\theta_1,\dots,\theta_s),$$

with probability mass function

$$p_{\boldsymbol{\theta}}(x_1,\dots,x_s) = \frac{n!}{\prod_{i=1}^{s} x_i!}\prod_{i=1}^{s}\theta_i^{x_i}.$$

Writing the density in exponential form yields

$$p_{\boldsymbol{\theta}}(x_1,\dots,x_s) = \exp\left\{\sum_{i=1}^{s} x_i\log(\theta_i)\right\}\cdot\frac{n!}{\prod_{i=1}^{s} x_i!},$$

which shows that the multinomial model is an exponential family with sufficient statistic $T(x) = (x_1,\dots,x_s)$. This representation, however, is redundant, since the counts satisfy the linear constraint $\sum_{i=1}^{s} X_i = n$.

To obtain a minimal representation, write

$$X_s = n - \sum_{i=1}^{s-1} X_i$$

and substitute into the exponent to obtain

$$p_{\boldsymbol{\theta}}(x_1,\dots,x_s) = \exp\left\{\sum_{i=1}^{s-1} x_i\log\left(\frac{\theta_i}{\theta_s}\right) + n\log(\theta_s)\right\}\cdot\frac{n!}{\prod_{i=1}^{s} x_i!}.$$

This shows that the multinomial model forms an $(s-1)$-dimensional exponential family with sufficient statistic

$$T(x) = (x_1,\dots,x_{s-1})$$

and natural parameter vector

$$\eta(\boldsymbol{\theta}) = \left(\log\left(\frac{\theta_1}{\theta_s}\right), \dots, \log\left(\frac{\theta_{s-1}}{\theta_s}\right)\right).$$

The term $\frac{n!}{\prod_{i=1}^{s} x_i!}$ plays the role of the base measure $h(x)$ and does not depend on the parameter $\boldsymbol{\theta}$.

Example 3.3 (Gaussian model)**.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with unknown parameter $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. The density on $\mathcal{X} = \mathbb{R}$ is

$$p_{\boldsymbol{\theta}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}.$$

Expanding the quadratic term yields

$$p_{\boldsymbol{\theta}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right\}.$$

The terms depending on the data x appear linearly as x and x^2, which allows the density to be written in exponential family form. Indeed,

$$p_{\boldsymbol{\theta}}(x) = \exp\left\{\left\langle \begin{pmatrix} x \\ -\frac{x^2}{2} \end{pmatrix}, \begin{pmatrix} \frac{\mu}{\sigma^2} \\ \frac{1}{\sigma^2} \end{pmatrix} \right\rangle - B(\boldsymbol{\theta})\right\},$$

where $B(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)$ is the log-normalizing constant.
Hence, the Gaussian model is a two-dimensional exponential family with

- sufficient statistic $T(x) = \big(x,\, -\frac{x^2}{2}\big)$,
- natural parameter $\eta(\boldsymbol{\theta}) = \big(\frac{\mu}{\sigma^2},\, \frac{1}{\sigma^2}\big)$.

In this representation the base measure $h(x)$ is constant, corresponding to Lebesgue measure.

Why is not the Gaussian model one-dimensional?
Although a single observation X uniquely determines the statistic $T(X) = \big(X, -\frac{X^2}{2}\big)$, the Gaussian model cannot be represented as a one-dimensional exponential family. The reason is geometric: the map

$$x \longmapsto \left(x, -\frac{x^2}{2}\right)$$

traces out a nonlinear curve in $\mathbb{R}^2$. Consequently, there is no nontrivial linear functional that recovers both components of $T(X)$ from a single scalar statistic while preserving the required inner-product structure of an exponential family.

This contrasts with the binomial case, where the corresponding embedding $x \mapsto (x, n-x)$ lies on an affine one-dimensional subspace, allowing a reduction to a one-dimensional representation. For the Gaussian model, the quadratic dependence on the data is essential and cannot be absorbed into a single sufficient statistic without leaving the exponential family framework.

Example 3.4 (Multivariate Gaussian model)**.** Let $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$, where the parameter

$$\theta = (\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^k \times PD(k),$$

and $PD(k)$ denotes the cone of symmetric positive definite $k \times k$ matrices. The density of $\boldsymbol{X}$ with respect to Lebesgue measure on $\mathcal{X} = \mathbb{R}^k$ is

$$p_\theta(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k \det(\Sigma)}} \exp\left\{ -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right\}.$$

Expanding the quadratic form and discarding terms that do not depend on $\boldsymbol{x}$ yields

$$p_\theta(\boldsymbol{x}) \propto \exp\left\{ -\frac{1}{2}\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{x} \right\}.$$

The linear term $\boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{x}$ is an inner product in $\mathbb{R}^k$ between $\boldsymbol{x}$ and the vector $\Sigma^{-1}\boldsymbol{\mu}$. To express the quadratic term in inner-product form, write

$$\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} = \sum_{i=1}^k \sum_{j=1}^k x_i x_j (\Sigma^{-1})_{ij}.$$

Introducing the outer product $\boldsymbol{x}\boldsymbol{x}^T$, whose (i, j)-entry equals $x_i x_j$, this can be rewritten as

$$\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} = \sum_{i,j} (\Sigma^{-1})_{ij} (\boldsymbol{x}\boldsymbol{x}^T)_{ij} = \operatorname{tr}\left(\Sigma^{-1} \boldsymbol{x}\boldsymbol{x}^T\right).$$

The trace induces an inner product on the vector space of symmetric $k \times k$ matrices, defined by

$$\langle A, B \rangle := \operatorname{tr}(AB), \qquad A, B \in \operatorname{Sym}(k).$$

Since both Σ^{-1} and $\boldsymbol{x}\boldsymbol{x}^T$ are symmetric, the quadratic term is an inner product in this space.

Combining both parts, the density can be written in exponential-family form as

$$p_\theta(\boldsymbol{x}) \propto \exp\left\{ \left\langle \Sigma^{-1}\boldsymbol{\mu},\, \boldsymbol{x} \right\rangle + \left\langle \Sigma^{-1},\, -\tfrac{1}{2}\boldsymbol{x}\boldsymbol{x}^T \right\rangle \right\}.$$

Thus, the multivariate Gaussian model is an exponential family with

- sufficient statistic $T(\boldsymbol{x}) = \left(\boldsymbol{x}, -\tfrac{1}{2}\boldsymbol{x}\boldsymbol{x}^T\right)$,
- natural parameter $\eta(\theta) = (\Sigma^{-1}\boldsymbol{\mu},\, \Sigma^{-1})$.

The dimension of this exponential family equals the dimension of the vector space in which $T(\boldsymbol{x})$ takes values. The first component $\boldsymbol{x}$ lies in $\mathbb{R}^k$, contributing k dimensions. The second component lies in the space of symmetric $k \times k$ matrices, whose dimension equals the number of free entries: k diagonal elements and $\frac{k(k-1)}{2}$ off-diagonal elements. Hence,

$$\dim \operatorname{Sym}(k) = k + \frac{k(k-1)}{2} = \frac{k(k+1)}{2}.$$

Therefore, the multivariate Gaussian family forms an exponential family of dimension

$$k + \frac{k(k+1)}{2}.$$

3.2 Nonuniqueness of Sufficient Statistics and Natural Parameters

Exponential-family representations of a statistical model are not unique: one can transform the sufficient statistic and the natural parameter without changing the model.

Let $T(x)$ be a sufficient statistic and $\eta(\theta)$ the corresponding natural parameter. For any invertible matrix $A \in \mathbb{R}^{k\times k}$ and vectors $\boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}^k$, define the affine transformations

$$\tilde{T}(x) = AT(x) + \boldsymbol{b}, \qquad \tilde{\eta}(\theta) = A^{-T}\eta(\theta) + \boldsymbol{c}.$$

Then the inner product between the transformed parameter and statistic expands as

$$\begin{aligned}\langle\tilde{\eta}(\theta), \tilde{T}(x)\rangle &= \langle A^{-T}\eta(\theta) + \boldsymbol{c}, AT(x) + \boldsymbol{b}\rangle \\ &\overset{(*)}{=} \langle\eta(\theta), T(x)\rangle + \langle A^{-T}\eta(\theta), \boldsymbol{b}\rangle + \langle\boldsymbol{c}, AT(x)\rangle + \langle\boldsymbol{c}, \boldsymbol{b}\rangle\end{aligned}$$

where in $(*)$ we have use that for any two vectors $\boldsymbol{u}, \boldsymbol{v}$ their inner product is define as

$$\langle\boldsymbol{u}, \boldsymbol{v}\rangle = \boldsymbol{u}^T\boldsymbol{v}$$

so

$$\begin{aligned}\left\langle A^{-T}\eta(\theta), AT(x)\right\rangle &= (A^{-T}\eta(\theta))^T AT(x) = \eta(\theta)^T(A^{-T})^T AT(x) = \eta(\theta)^T T(x) \\ &= \langle\eta(\theta), T(x)\rangle\,.\end{aligned}$$

This implies that the density can be rewritten as

$$p_\theta(x) = \exp\Big\{\langle\tilde{\eta}(\theta), \tilde{T}(x)\rangle - \tilde{B}(\theta)\Big\} \cdot h(x)\exp\big\{-\langle\boldsymbol{c}, AT(x)\rangle\big\},$$

with the new log-partition function

$$\tilde{B}(\theta) = B(\theta) + \langle A^{-T}\eta(\theta), \boldsymbol{b}\rangle + \langle\boldsymbol{c}, \boldsymbol{b}\rangle.$$

In other words, an affine transformation of the sufficient statistic $T(x)$ can be compensated by a corresponding transformation of the natural parameter $\eta(\theta)$, possibly adjusting the normalizing constant. Linear relations among the coordinates of $T(x)$ can thus be eliminated by choosing a suitable A and $\boldsymbol{b}$, leading to a *minimal representation* of the exponential family.

Definition 3.2 (Order of an Exponential Family)**.** The **order of the exponential family** is the dimension of the minimal sufficient statistic after all linear dependencies among its coordinates have been removed. This dimension is unique for a given family.

Remark (Invertibility and singular cases).

- In univariate or multivariate Gaussian models, assuming Σ positive definite (invertible) ensures the density exists in $\mathbb{R}^k$. If a variance is zero, the distribution degenerates to a point mass; if a covariance matrix is singular, the random vector takes values in a strict subspace of $\mathbb{R}^k$. Such singular cases are not part of the exponential-family representation as discussed here.
- Affine transformations provide a systematic way to expose linear dependencies among coordinates of $T(x)$. For example, if some linear combination of coordinates is constant across all observations, one can redefine T via a suitable matrix A to eliminate the redundant coordinate. This reduction ensures that only the essential dimensions remain, giving the family its unique order.

3.3 Random Sampling

Let $\mathsf{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ be a k–dimensional exponential family with densities

$$p_\theta(x) = \frac{d\mathsf{P}_\theta}{d\nu}(x) = \exp\{\langle \eta(\theta), T(x)\rangle - B(\theta)\}\, h(x), \quad x \in \mathcal{X}.$$

Let $X_1, \dots, X_n$ be i.i.d. random variables with distribution P_θ.

The joint distribution of the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ then has density

$$p_\theta(\boldsymbol{x}) = \prod_{i=1}^n p_\theta(x_i) = \exp\left\{\left\langle \eta(\theta), \sum_{i=1}^n T(x_i)\right\rangle - nB(\theta)\right\} \cdot \prod_{i=1}^n h(x_i), \quad \boldsymbol{x} \in \mathcal{X}^n.$$

Thus, the family of joint distributions of $\boldsymbol{X}$ is again a k–dimensional exponential family. The natural parameter $\eta(\theta)$ remains unchanged, while the sufficient statistic becomes

$$\sum_{i=1}^n T(X_i).$$

In particular, independent sampling preserves the exponential family structure.

Note. The dimension of the sufficient statistic does not depend on the sample size n.[1]

This property has important consequences. Even for very large samples, all relevant statistical information is summarized by a single vector in $\mathbb{R}^k$. As a result, statistical analysis, geometric intuition, and asymptotic arguments for exponential families can be carried out in a fixed finite-dimensional space, regardless of the number of observations.

By contrast, for general statistical models, sufficient statistics may grow in dimension with the sample size (for example, order statistics). Under mild regularity conditions, the invariance of the dimension of sufficient statistics under i.i.d. sampling essentially characterizes exponential families.

[1] This behavior is highly atypical outside the class of exponential families; see Hipp (1974). Recall, however, the Uniform$(0, \theta)$ model as a counterexample without regularity.

3.4 Minimal Sufficiency

For exponential families, minimal sufficiency can be characterized in a particularly transparent way. The key reason is that likelihood ratios in exponential families reduce to differences of linear forms in the sufficient statistic. As a result, the general criterion for minimal sufficiency based on proportional likelihoods admits a simple geometric interpretation in terms of the natural parameter space; see Corollary 6.16 in Lehmann and Casella (1998).

Corollary 3.1. *Let $\mathsf{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ be a k–dimensional exponential family with natural parameter mapping $\eta : \Theta \to \mathbb{R}^k$. If the set*

$$\eta(\Theta) = \{\eta(\theta) : \theta \in \Theta\} \subseteq \mathbb{R}^k$$

contains $k+1$ points $\eta^{(0)}, \dots, \eta^{(k)}$ such that the vectors

$$\eta^{(j)} - \eta^{(0)}, \quad j = 1, \dots, k,$$

are linearly independent, then the sufficient statistic T is minimal sufficient.

This result follows directly from the characterization of minimal sufficiency in terms of proportional likelihoods. In an exponential family, the ratio of two likelihoods corresponding to parameter values θ and θ' can be written as

$$\exp\Big\{\big\langle \eta(\theta) - \eta(\theta'), T(x)\big\rangle - \big(B(\theta) - B(\theta')\big)\Big\},$$

so equality of likelihood ratios across all parameter values is determined entirely by linear constraints on $T(x)$. If the set of possible difference vectors $\eta(\theta) - \eta(\theta')$ spans $\mathbb{R}^k$, then no nontrivial reduction of T can preserve these ratios, and T is therefore minimal sufficient.

An equivalent geometric formulation of the condition in the corollary is the following. The sufficient statistic T is minimal sufficient if and only if the affine hull of $\eta(\Theta)$ has dimension k. The affine hull of a set $A \subseteq \mathbb{R}^k$ is defined as

$$\left\{ \sum_{j=0}^{k} \alpha_j \eta^{(j)} : \eta^{(j)} \in A,\ \alpha_j \in \mathbb{R},\ \sum_{j=0}^{k} \alpha_j = 1 \right\}.$$

In many standard examples, this condition is easily verified.

Univariate Gaussian family. Consider the normal distribution with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. The parameter space $\Theta = \mathbb{R} \times (0, \infty)$ is the open upper half–plane. The corresponding natural parameter space $\eta(\Theta)$ is also an open subset of $\mathbb{R}^2$, hence its affine hull has dimension 2. Therefore, the canonical sufficient statistic is minimal sufficient.

Multivariate Gaussian family. For the multivariate normal distribution with mean vector $\mu \in \mathbb{R}^k$ and covariance matrix Σ positive definite, the natural parameter space is $\mathbb{R}^k \times \mathcal{S}^k_{++}$, where $\mathcal{S}^k_{++}$ denotes the cone of positive definite matrices. This set is open and full–dimensional in the corresponding Euclidean space, implying minimal sufficiency of the usual sufficient statistic.

Binomial family and logistic regression. For the binomial model, the original parameter θ lies in the open unit interval $(0,1)$. The natural parameter is given by the log–odds transformation

$$\eta(\theta) = \log \frac{\theta}{1-\theta},$$

which maps $(0,1)$ onto all of $\mathbb{R}$. Consequently, $\eta(\Theta) = \mathbb{R}$, whose affine hull has dimension 1. This explains both the minimal sufficiency of the binomial sufficient statistic and the central role of the log–odds parameterization in logistic regression, where linear predictors naturally take values in $\mathbb{R}$.

In summary, for exponential families, minimal sufficiency is governed entirely by the geometry of the natural parameter space. If $\eta(\Theta)$ is full–dimensional (or equivalently, has an affine hull of dimension k), then the canonical sufficient statistic is automatically minimal sufficient.

3.5 Canonical Form

By re–parametrizing an exponential family in terms of its natural parameters, the model can be written in **canonical form**. Specifically, a k–dimensional exponential family admits a representation

$$p_\eta(x) = \exp\{\langle \eta, T(x)\rangle - A(\eta)\}\, h(x), \quad x \in \mathcal{X},$$

where $\eta \in \mathbb{R}^k$ is the natural parameter and $T(x) \in \mathbb{R}^k$ is the sufficient statistic. We write P_η for the distribution with density p_η.

Passing to the canonical form amounts to abandoning the original parameterization (e.g. means, variances, probabilities, rates) in favor of the natural parameters. For instance, instead of indexing a Gaussian distribution by its mean and variance, or a binomial distribution by a success probability, we index these distributions by inverse variances, mean–variance ratios, or log–odds. This is purely a re–parametrization: the underlying distributions do not change, only the coordinates used to describe them.

3.5.1 Natural Parameter Space

The density $p_\eta(x)$ is well defined only for those values of η for which the normalizing constant is finite. This leads to the definition of the **natural parameter space**

$$\mathcal{H} = \left\{ \eta \in \mathbb{R}^k : \int_{\mathcal{X}} \exp\{\langle \eta, T(x)\rangle\} h(x)\, d\nu(x) < \infty \right\}.$$

For every $\eta \in \mathcal{H}$, a valid probability density is obtained by setting

$$A(\eta) = \log \int_{\mathcal{X}} \exp\{\langle \eta, T(x)\rangle\} h(x)\, d\nu(x).$$

The function $A(\eta)$ is called the **cumulant generating function** or **log–Laplace transform** of the family.

The role of $\mathcal{H}$ is fundamental. Any nonnegative function with finite integral can be normalized to obtain a probability density, but if the integral diverges, no normalization is possible. Thus, $\mathcal{H}$ is the largest set of natural parameters for which the exponential representation yields a well-defined statistical model.

Examples

Gaussian family. For the univariate Gaussian distribution, a natural sufficient statistic is $T(x) = (x, x^2)$. Only certain choices of η lead to integrable densities: the coefficient multiplying x^2 must be negative. This restriction corresponds exactly to the requirement that the variance be positive. If the quadratic term had the wrong sign, the exponential would not be integrable over $\mathbb{R}$, and no probability density could be defined.

Binomial family and logistic regression. For the binomial model with success probability $\theta \in (0, 1)$, the natural parameter is the log–odds

$$\eta = \log \frac{\theta}{1 - \theta},$$

which ranges over all of $\mathbb{R}$. Thus, the natural parameter space is unconstrained, even though the original parameter θ lies in a bounded interval. This re–parametrization explains why logistic regression works with linear predictors: the natural parameter lives on the entire real line, making linear combinations natural and unconstrained.

Multinomial and symmetry considerations. In multinomial models, it is sometimes convenient to retain a redundant representation of the sufficient statistic in order to preserve symmetry. While redundant coordinates may later be removed to obtain a minimal representation, keeping them can simplify calculations and notation without affecting the underlying statistical structure.

Exponential random graph models. The exponential family framework can also be used constructively to define models. For example, when modeling random graphs, the observation X is a graph, and one may choose a sufficient statistic $T(X)$ that counts features such as the number of edges and the number of triangles. Given parameters η_1, η_2, the probability of observing a particular graph is proportional to

$$\exp\{\eta_1 T_1(X) + \eta_2 T_2(X)\}.$$

Provided the graph space is finite, the normalizing constant $A(\eta)$ always exists. Varying η yields a flexible family of distributions that can encode complex network structure beyond independent edge formation.

Convexity and Likelihood Optimization

A key advantage of the canonical form emerges when studying likelihood functions. In general statistical models, likelihoods may be highly non–convex and multimodal. In contrast, for exponential families in canonical form, the log–likelihood is a concave function of η over the natural parameter space $\mathcal{H}$.

This explains a common empirical observation: for many standard models (binomial, multinomial, Poisson, Gaussian, gamma), maximum likelihood estimators are unique and easy to compute. Although this fact may appear model–specific when working in the original parameters, it has a unified explanation in the canonical parameterization.

Moment and Cumulant Generating Functions

Finally, note that without the logarithm, the integral defining $A(\eta)$ corresponds to a moment generating function. For a random variable X, the moment generating function is $\mathsf{E}[e^{tX}]$; for a random vector, it becomes $\mathsf{E}[\exp(\langle t, X\rangle)]$. Taking the logarithm yields the cumulant generating function, whose derivatives generate moments and cumulants of the sufficient statistic.

Thus, the canonical form not only simplifies likelihood theory and optimization, but also provides a natural bridge between exponential families, convex analysis, and moment theory.

3.5.2 Convexity Properties

One of the most powerful structural features of exponential families is their built–in convexity. This convexity governs both the geometry of the parameter space and the shape of likelihood functions, and it explains why maximum likelihood estimation is particularly well behaved in these models.

Recall that in canonical form an exponential family has densities

$$p_\eta(x) = \exp\{\langle \eta, T(x)\rangle - A(\eta)\}h(x), \quad \eta \in \mathcal{H},$$

where the natural parameter space $\mathcal{H}$ consists precisely of those η for which $A(\eta) < \infty$.

Theorem 3.1 (Convexity Properties)**.**

(i) The natural parameter space $\mathcal{H}$ is a convex subset of $\mathbb{R}^k$.

(ii) The cumulant generating function $A(\eta)$ is convex on $\mathcal{H}$. It is strictly convex if $\mathsf{P}_\eta \neq \mathsf{P}_{\eta'}$ for all $\eta, \eta' \in \mathcal{H}$ with $\eta \neq \eta'$.

(iii) For fixed data x, the log–likelihood function

$$\ell_x(\eta) = \log p_\eta(x)$$

is concave on $\mathcal{H}$. It is strictly concave if $\mathsf{P}_\eta \neq \mathsf{P}_{\eta'}$ for all $\eta, \eta' \in \mathcal{H}$ with $\eta \neq \eta'$.

Comments:
Convexity is the mathematical reason why maximum likelihood estimation in exponential families rarely exhibits pathologies such as multiple local optima. In contrast to general statistical models, likelihood optimization here is a convex optimization problem.

Recall the basic facts:

- Every local maximum of a concave function is a global maximum.
- A strictly concave function has at most one maximizer.

Consequently, in the *full* exponential family indexed by all $\eta \in \mathcal{H}$, the maximum likelihood estimator (if it exists) is unique.

Existence, however, is not guaranteed: a strictly concave function may fail to attain its supremum (for example, the logarithm on $(0, \infty)$). Nevertheless, uniqueness follows automatically once existence is established.

If one restricts η to an arbitrary subset of $\mathcal{H}$, convexity and concavity may be destroyed. An important exception occurs when the restriction is a linear subspace of $\mathcal{H}$: in that case, the model remains an exponential family of lower dimension, and the theorem continues to apply.

Proof. We begin with convexity of the natural parameter space and of the function A. Let $\eta, \eta' \in \mathcal{H}$ and $\alpha \in [0, 1]$. By definition,

$$\begin{aligned}\exp\{A(\alpha\eta + (1-\alpha)\eta')\} &= \int \exp\{\langle \alpha\eta + (1-\alpha)\eta', T(x)\rangle\} h(x)\, d\nu(x) \\ &= \int \left(\exp\{\langle \eta, T(x)\rangle\}\right)^{\alpha} \left(\exp\{\langle \eta', T(x)\rangle\}\right)^{1-\alpha} h(x)\, d\nu(x).\end{aligned}$$

This expression is exactly of the form to which Hölder's inequality applies. Using Hölder's inequality yields

$$\begin{aligned}\exp\{A(\alpha\eta + (1-\alpha)\eta')\} &\leq \left(\int \exp\{\langle \eta, T(x)\rangle\} h(x)\, d\nu(x)\right)^{\alpha} \left(\int \exp\{\langle \eta', T(x)\rangle\} h(x)\, d\nu(x)\right)^{1-\alpha} \\ &= \exp\{\alpha A(\eta) + (1-\alpha) A(\eta')\}.\end{aligned}$$

Taking logarithms gives

$$A(\alpha\eta + (1-\alpha)\eta') \leq \alpha A(\eta) + (1-\alpha) A(\eta'). \tag{3.2}$$

This inequality has two immediate consequences. First, if $A(\eta)$ and $A(\eta')$ are finite, then $A(\alpha\eta + (1-\alpha)\eta')$ is finite as well. Hence, $\mathcal{H}$ is convex. Second, inequality (3.2) is precisely the definition of convexity of the function A.

Equality in Hölder's inequality occurs if and only if the two functions inside the integral are proportional. Here, this would require the existence of a constant $c > 0$ such that

$$\exp\{\langle \eta, T(x)\rangle\} = c \cdot \exp\{\langle \eta', T(x)\rangle\} \quad \text{for all } x.$$

After normalization, this implies $\mathsf{P}_\eta = \mathsf{P}_{\eta'}$. Therefore, if distinct natural parameters always correspond to distinct distributions, the inequality in (3.2) is strict for $\alpha \in (0,1)$, and A is strictly convex, proving (ii).

Finally, for fixed data x,

$$\ell_x(\eta) = \log p_\eta(x) = \langle \eta, T(x) \rangle - A(\eta) + \log h(x).$$

The term $\langle \eta, T(x) \rangle$ is linear in η, while $-A(\eta)$ is concave because A is convex. Adding a linear function does not affect concavity. Hence, $\ell_x(\eta)$ is concave on $\mathcal{H}$, and strictly concave whenever A is strictly convex, proving (iii). □

The convexity of the natural parameter space and the cumulant generating function provides a unified explanation for the uniqueness of maximum likelihood estimators in exponential families such as the Gaussian, Poisson, binomial, and multinomial models. What appears as a collection of model–specific calculations is, in fact, a single geometric phenomenon.

Example 3.5 (Gaussian model and convexity of the cumulant function)**.** Let $X_1, \ldots, X_n$ be i.i.d. random variables with

$$X_i \sim \mathcal{N}(\mu, \sigma^2),$$

where both the mean μ and the variance σ^2 are unknown. We represent this model as a full exponential family by introducing the natural parameters

$$\eta_1 = \frac{\mu}{\sigma^2}, \qquad \eta_2 = \frac{1}{\sigma^2}.$$

With this parametrization, the log-likelihood function can be written in the canonical exponential-family form

$$\ell_x(\eta) = \eta_1 \sum_{i=1}^n x_i + \eta_2 \left(-\frac{1}{2} \sum_{i=1}^n x_i^2 \right) - A(\eta),$$

where the cumulant generating function (log-partition function) is given by

$$A(\eta) = \frac{n}{2}\frac{\mu^2}{\sigma^2} + \frac{n}{2}\log(2\pi\sigma^2) = \frac{n}{2}\left(\frac{\eta_1^2}{\eta_2} - \log \eta_2 + \log(2\pi) \right).$$

The gradient of A coincides with the vector of expectations of the sufficient statistics. A direct calculation yields

$$\nabla A(\eta) = \begin{pmatrix} n\,\dfrac{\eta_1}{\eta_2} \\ -\dfrac{n}{2}\left(\dfrac{\eta_1^2}{\eta_2^2} + \dfrac{1}{\eta_2} \right) \end{pmatrix} = \begin{pmatrix} n\mu \\ -\frac{n}{2}\left(\mu^2 + \sigma^2\right) \end{pmatrix} = \begin{pmatrix} \mathsf{E}_\eta\left[\sum\limits_{i=1}^n X_i \right] \\ \mathsf{E}_\eta\left[-\frac{1}{2} \sum\limits_{i=1}^n X_i^2 \right] \end{pmatrix}.$$

Thus, as predicted by the general theory of exponential families, the gradient of the cumulant function returns the expectations of the sufficient statistics.

To verify the convexity of A directly, we compute its Hessian matrix:

$$\nabla^2 A(\eta) = \begin{pmatrix} \dfrac{n}{\eta_2} & -\dfrac{n\eta_1}{\eta_2^2} \\ -\dfrac{n\eta_1}{\eta_2^2} & n\left(\dfrac{\eta_1^2}{\eta_2^3} + \dfrac{1}{2\eta_2^2}\right) \end{pmatrix}.$$

Since $\eta_2 > 0$ (it is the inverse variance), the upper-left entry is positive. Moreover, the determinant is

$$\det\bigl(\nabla^2 A(\eta)\bigr) = \frac{n^2}{2\eta_2^3} > 0.$$

By Sylvester's criterion, the Hessian is positive definite for all admissible η, and hence A is strictly convex.

This explicit calculation illustrates the general convexity result for cumulant functions in exponential families, which in theory follows from Hölder's inequality but in this Gaussian case can be verified directly by elementary calculus.

3.5.3 Analytic Properties

Theorem 3.2 (Analyticity of exponential-family integrals)**.** *Let η be a point in the interior of the natural parameter space $\mathcal{H}$. Let $f : \mathcal{X} \to \mathbb{R}$ be a measurable function that is integrable with respect to P_η.*

Then the mapping

$$\eta \longmapsto \int f(x)\, \exp\bigl\{\langle \eta, T(x)\rangle\bigr\}\, h(x)\, d\nu(x)$$

admits continuous partial derivatives of all orders. Moreover, all derivatives can be computed by differentiating under the integral sign. In particular, the function is analytic on the interior of $\mathcal{H}$.

Remark. The expression above represents (up to normalization) the expectation of the statistic $f(X)$ when X is drawn from an exponential family distribution with natural parameter η. The theorem therefore shows that expectations with respect to exponential family models depend smoothly on the natural parameters, as long as η lies in the interior of the parameter space.

Remark (Idea of the proof). The result follows from an application of Lebesgue's dominated convergence theorem to difference quotients arising in the definition of partial derivatives. Since derivatives are limits of difference quotients, the key step is to justify the interchange of differentiation and integration. The exponential structure of the density allows the difference quotients to be dominated by an integrable function, ensuring the validity of this interchange. A detailed proof can be found in Lehmann and Romano (2005, Theorem 2.7.1).

3.5.4 Cumulant Generating Transform

A central object in the theory of exponential families is the cumulant generating function $A(\eta)$. Beyond its role as a normalizing constant, A encodes the first- and second-order moments of the sufficient statistic through differentiation.

Corollary 3.2 (Derivatives of the cumulant function)**.** *Let $\{\mathsf{P}_\eta : \eta \in \mathcal{H}\}$ be an exponential family in canonical form with sufficient statistic $T = (T_1, \dots, T_d)$ and cumulant generating function A. For all η in the interior of $\mathcal{H}$,*

$$\frac{\partial}{\partial \eta_j} A(\eta) = \mathsf{E}_\eta[T_j(X)] \,, \qquad \frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta) = \mathsf{Cov}_\eta[T_j(X), T_k(X)] \,.$$

Equivalently,

$$\nabla A(\eta) = \mathsf{E}_\eta[T(X)], \qquad \nabla^2 A(\eta) = \mathsf{Cov}_\eta[T(X)].$$

Proof. By definition of the exponential family in canonical form, the density integrates to one:

$$\int_{\mathcal{X}} \exp\{\langle \eta, T(x) \rangle - A(\eta)\} \, h(x) \, d\nu(x) = 1, \qquad \forall \, \eta \in \mathcal{H}.$$

Although the right-hand side is constant, the left-hand side is a function of η. Since η lies in the interior of $\mathcal{H}$, we may differentiate under the integral sign.

Taking the partial derivative with respect to η_j yields

$$\int_{\mathcal{X}} \exp\{\langle \eta, T(x) \rangle - A(\eta)\} \left(T_j(x) - \frac{\partial}{\partial \eta_j} A(\eta) \right) h(x) \, d\nu(x) = 0.$$

Recognizing the integrand as an expectation under P_η proves the first identity.

Differentiating once more with respect to η_k and differentiating under the integral sign, yields

$$\begin{aligned} 0 = &\int_{\mathcal{X}} \exp\{\langle \eta, T(x) \rangle - A(\eta)\} \Big(T_k(x) - \frac{\partial}{\partial \eta_k} A(\eta) \Big) \Big(T_j(x) - \frac{\partial}{\partial \eta_j} A(\eta) \Big) h(x) \, d\nu(x) \\ &- \int_{\mathcal{X}} \exp\{\langle \eta, T(x) \rangle - A(\eta)\} \, \frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta) \, h(x) \, d\nu(x). \end{aligned}$$

Since the density integrates to one, the second integral equals $\frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta)$, and we obtain

$$\int_{\mathcal{X}} \exp\{\langle \eta, T(x) \rangle - A(\eta)\} \Big(T_j(x) - \mathsf{E}_\eta[T_j(X)] \Big) \Big(T_k(x) - \mathsf{E}_\eta[T_k(X)] \Big) h(x) \, d\nu(x) = \frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta).$$

Recognizing the left-hand side as a covariance under P_η, we conclude that

$$\mathsf{Cov}_\eta[T_j(X), T_k(X)] = \frac{\partial^2}{\partial \eta_j \partial \eta_k} A(\eta).$$

□

Remark (Interpretation)*.* The function A is called the cumulant generating function because its derivatives generate cumulants of the sufficient statistic. In particular, the gradient of A yields the mean vector, while the Hessian yields the covariance matrix. This provides a direct and efficient way to compute moments in exponential families by differentiation rather than integration.

3.5.5 Random Vectors

Let

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_d \end{pmatrix}$$

be a random vector, viewed equivalently as a vector of d real-valued random variables or as a random element of $\mathbb{R}^d$ equipped with its Borel σ-algebra.

Definition 3.3 (Expectation of a random vector)**.** The expectation (mean vector) of X is defined componentwise by

$$\mathsf{E}[X] = \begin{pmatrix} \mathsf{E}[X_1] \\ \vdots \\ \mathsf{E}[X_d] \end{pmatrix},$$

whenever all component expectations exist.

Definition 3.4 (Variance matrix)**.** The variance matrix of X is the $d \times d$ matrix

$$\mathsf{Var}[X] = \big(\mathsf{Cov}(X_i, X_j)\big)_{1 \le i,j \le d},$$

whose diagonal entries are the variances of the coordinates of X and whose off-diagonal entries are the pairwise variances.

Remark. Expectations of vector- or matrix-valued random quantities are always understood entrywise. In particular, whenever $\mathsf{E}[X]$ or $\mathsf{Var}[X]$ is written, all relevant moments are assumed to exist.

Remark (Positive semidefiniteness)*.* The covariance matrix $\mathsf{Var}[X]$ is always positive semidefinite.

In the context of exponential families, the sufficient statistic $T(X)$ is a random vector. Applying the results of Section 3.5.4, the gradient and Hessian of the cumulant generating function A admit a direct probabilistic interpretation:

$$\nabla A(\eta) = \mathsf{E}_\eta[T(X)], \qquad D_2 A(\eta) = \mathsf{Var}_\eta[T(X)].$$

Thus, the first derivative of A yields the mean vector of the sufficient statistic, while the second derivative yields its variance matrix.

3.6 Mean Parametrization

Let $\{\mathsf{P}_\eta : \eta \in \mathcal{H}'\}$ be an exponential family in canonical form, where the natural parameter space $\mathcal{H}' \subset \mathbb{R}^d$ is open and convex. The associated cumulant generating function $A : \mathcal{H}' \to \mathbb{R}$ is strictly convex on $\mathcal{H}'$.

As a consequence, the gradient mapping

$$\eta \longmapsto \nabla A(\eta) = \mathsf{E}_\eta[T(X)]$$

is injective. Hence, each value of the natural parameter η corresponds to a unique value of the expectation of the sufficient statistic. This establishes a one–to–one correspondence between the natural parameter η and the *mean parameter*

$$\theta \;:=\; \mathsf{E}_\eta[T(X)].$$

This alternative parametrization of the exponential family is called the *mean parametrization.* While the natural parametrization is particularly well-suited for studying convexity and optimization properties, the mean parametrization provides a direct probabilistic interpretation in terms of expected sufficient statistics.

Example 3.6 (Gaussian model)**.** Consider the Gaussian exponential family with unknown mean μ and variance σ^2, and natural parameters

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{pmatrix}.$$

The corresponding mean parameters are given by

$$\nabla A(\eta) = \mathsf{E}_\eta[T(X)] = \begin{pmatrix} \mu \\ -\frac{1}{2}(\mu^2+\sigma^2) \end{pmatrix} = \begin{pmatrix} \dfrac{\eta_1}{\eta_2} \\ -\dfrac{1}{2}\left(\dfrac{\eta_1^2}{\eta_2^2} + \dfrac{1}{\eta_2}\right) \end{pmatrix}.$$

This mapping is one–to–one, illustrating explicitly the equivalence between the natural and mean parametrizations in this model.

Remark. The injectivity of the gradient map follows from strict convexity of A: the gradient of a strictly convex function is injective on any convex domain. Consequently, distributions in an exponential family may be uniquely identified either by their natural parameters or by the mean values of their sufficient statistics.

3.7 A Terminological Remark on Exponential Families

Before turning to estimation, we make a brief remark on terminology. An *exponential family* is not a single probability distribution but a *statistical model*, that is, a collection of probability distributions indexed by a parameter. Typical examples include the Gaussian distributions with varying mean and variance, the Poisson distributions with varying rate parameter, or the binomial distributions with varying success probability.

In the literature—particularly in machine learning—it is common to encounter phrases such as "the Poisson distribution is in the exponential family." Such formulations should be interpreted with care. More precisely, one should say that a *family* of Poisson distributions, indexed by their rate parameter, *forms* or *constitutes* an exponential family when written in canonical form.

Throughout these notes, we will therefore avoid referring to individual distributions as belonging to “the” exponential family. Instead, we emphasize that exponential families are collections of distributions that admit a common representation in exponential-family form.

We now proceed to the problem of statistical estimation for exponential family models.

4. Unbiased Estimation

This chapter is devoted to estimators that are *unbiased*, that is, estimators whose expectation equals the target parameter.We will develop a classical optimality theory, including criteria for comparing and improving unbiased estimators.

The material follows the classical theory of point estimation, as presented for example in Lehmann and Casella (1998, Chapter 2).

4.1 Unbiased Estimators

Setting We consider the following framework:

- Observation: X, taking values in a sample space $\mathcal{X}$.
- Statistical model: $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$.
- Parameter of interest: $\gamma(\theta) \in \mathbb{R}$, a real-valued function of θ.

Definition 4.1. An estimator $T = T(X)$ is called **unbiased** for $\gamma(\theta)$ if

$$\mathsf{E}_\theta[T] = \gamma(\theta) \quad \text{for all } \theta \in \Theta,$$

where the expectation is taken with respect to $X \sim \mathsf{P}_\theta$.

If an unbiased estimator exists for $\gamma(\theta)$, the parameter $\gamma(\theta)$ is called **U–estimable**.

Motivation and Intuition
Unbiased estimators are attractive because, on average, they return the correct value of the target parameter. That is, repeated applications of the estimator will not systematically overestimate or underestimate the quantity of interest.

In terms of mean squared error (MSE),

$$\text{MSE}(T) = \big(\text{Bias}(T)\big)^2 + \mathsf{Var}_\theta(T),$$

unbiased estimators have zero bias, so minimizing MSE reduces to minimizing variance. Therefore, when restricting attention to unbiased estimators, the natural question is:

Does there exist an unbiased estimator with minimal variance?

Discussion

The notion of unbiasedness depends on the underlying model $\mathcal{P}$ and the parameter of interest $\gamma(\theta)$. For example:

- X may be a single observation, a vector of independent draws, or even a matrix of observations.
- $\gamma(\theta)$ could be a component of a multivariate parameter, such as the mean of a Gaussian distribution.

Unbiased estimators are often preferred when constructing confidence intervals, because the usual ± margin-of-error construction is centered around the estimator. A biased estimator would systematically shift this interval, potentially distorting coverage probabilities.

Finally, note that not every parameter admits an unbiased estimator. Restricting attention to unbiased estimators removes pathological cases (such as constant estimators that trivially minimize MSE but are useless for estimating a varying parameter), allowing the meaningful study of *optimal unbiased estimators*, i.e., those with minimal variance among all unbiased estimators.

4.2 UMVUE

Definition 4.2. Let T be an unbiased estimator of $\gamma(\theta)$. If for any other unbiased estimator T' of $\gamma(\theta)$, it holds that

$$\mathsf{Var}_\theta[T] \leq \mathsf{Var}_\theta[T'] \quad \forall \theta \in \Theta,$$

then T is called **UMVUE** (*Uniform Minimum Variance Unbiased Estimator*) of $\gamma(\theta)$.

That is, among all unbiased estimators of $\gamma(\theta)$, the estimator T achieves the smallest variance uniformly over the entire parameter space.

Proposition 4.1 (Uniqueness of the UMVUE)**.** *Let T be a UMVUE of $\gamma(\theta)$ such that*

$$\mathsf{Var}_\theta(T) < \infty, \quad \forall \theta \in \Theta.$$

Then T is unique in the following sense: if T' is another UMVUE of $\gamma(\theta)$, then

$$\mathsf{P}_\theta(T = T') = 1, \quad \forall \theta \in \Theta.$$

Proof. Assume that T' is another UMVUE of $\gamma(\theta)$. Since both T and T' are unbiased, their average

$$T^* := \frac{1}{2}(T + T')$$

is also an unbiased estimator of $\gamma(\theta)$.

Because T is a UMVUE, its variance cannot exceed that of any other unbiased estimator. In particular,

$$\mathsf{Var}_\theta(T) \le \mathsf{Var}_\theta(T^*), \quad \forall\theta \in \Theta.$$

A direct computation yields

$$\mathsf{Var}_\theta(T^*) = \frac{1}{4}\mathsf{Var}_\theta(T) + \frac{1}{4}\mathsf{Var}_\theta(T') + \frac{1}{2}\mathsf{Cov}_\theta(T, T').$$

Since both T and T' are UMVUEs, they must have the same variance for every θ. Applying the Cauchy–Schwarz inequality,

$$\mathsf{Cov}_\theta(T, T') \le \sqrt{\mathsf{Var}_\theta(T)\mathsf{Var}_\theta(T')},$$

we obtain

$$\mathsf{Var}_\theta(T^*) \le \mathsf{Var}_\theta(T).$$

Combining this with the earlier inequality shows that equality must hold throughout. In particular, equality in the Cauchy–Schwarz inequality implies

$$\mathrm{Corr}_\theta(T, T') = 1,$$

and therefore there exist constants $a, b \in \mathbb{R}$ such that

$$\mathsf{P}_\theta(T' = aT + b) = 1, \quad \forall\theta \in \Theta.$$

Because T and T' are unbiased for the same target, $\mathsf{E}_\theta[T] = \mathsf{E}_\theta[T']$ implies $b = 0$. Moreover, equality of variances yields

$$\mathsf{Var}_\theta(T') = a^2\mathsf{Var}_\theta(T) = \mathsf{Var}_\theta(T),$$

so $a = 1$. Hence,

$$\mathsf{P}_\theta(T = T') = 1, \quad \forall\theta \in \Theta,$$

which proves uniqueness. □

Example 4.1 (Binomial model)**.** Let $X \sim \mathrm{Binomial}(n, \theta)$, where n is known and $\theta \in (0, 1)$.

For any estimator $T = T(X)$, its expectation is given by

$$\mathsf{E}_\theta[T] = \sum_{x=0}^{n} T(x)\binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Since the binomial probability mass function is a polynomial in θ of degree at most n, the expectation $\mathsf{E}_\theta[T]$ is itself a polynomial in θ of degree at most n. Moreover, by choosing the values $T(0), \dots, T(n)$ appropriately, any polynomial of degree at most n can be represented in this way.

Consequently, the parameters $\gamma(\theta)$ that are unbiasedly estimable in the binomial model are precisely the polynomials in θ of degree at most n. Thus, non-polynomial functions such as $\gamma(\theta) = \sqrt{\theta}$ are not unbiasedly estimable.

As a simple example, the parameter $\gamma(\theta) = \theta$ is unbiasedly estimable. The estimator

$$\widehat{\theta} = \frac{X}{n}$$

is unbiased for θ and, in fact, is the UMVUE of θ in the binomial model; see Lehmann and Casella (1998, Example 2.3.1).

Example 4.2 (Gaussian model)**.** Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$.

In this model, several UMVUEs can be identified explicitly. A later theorem shows that the

- Sample mean
$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$
is the UMVUE of μ;
- Sample variance
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$
is the UMVUE of σ^2.

If the target parameter is changed from σ^2 to σ, unbiasedness is no longer preserved by taking square roots: $s = \sqrt{s^2}$ is biased for σ. Nevertheless, there exists an unbiased estimator of σ, given by

$$2^{1/2}\frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \sqrt{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2},$$

which is the UMVUE of σ.

Note. In this Gaussian example, the maximum likelihood estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

has smaller mean square error than the UMVUE s^2, despite being biased. More generally, among estimators of the form

$$\frac{1}{a}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2,$$

the choice $a = n+1$ minimizes the mean square error. This highlights that UMVUE optimality depends on the unbiasedness criterion and does not necessarily align with MSE optimality.

4.2.1 Behrens–Fisher Problem

Proposition 4.2 (Behrens–Fisher Problem)**.** *Consider two independent samples $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$, where*

$$X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad Y_j \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \tau^2),$$

with unknown parameter $\theta = (\mu, \sigma^2, \tau^2)$. Although unbiased estimators of the common mean μ exist, there is no UMVUE of μ.

Proof. Suppose, for contradiction, that there exists a UMVUE $\hat{\mu}^*$ for μ in the full model.

To analyze optimality, consider a submodel in which the variance ratio

$$\frac{\tau^2}{\sigma^2} = a$$

is fixed and known, with $a > 0$. In this submodel, the optimal way to combine information from the two samples is determined by the relative variances. It can be shown that the UMVUE of μ in this submodel is

$$\hat{\mu}_a = \frac{\sum_{i=1}^{n} X_i + \frac{1}{a}\sum_{j=1}^{m} Y_j}{n + \frac{1}{a}m}.$$

This estimator assigns more weight to the sample with smaller variance and is unbiased for μ.

Importantly, $\hat{\mu}_a$ remains an unbiased estimator of μ even in the full model with arbitrary (μ, σ^2, τ^2). Since $\hat{\mu}^*$ is assumed to be a UMVUE in the full model, it must also be a UMVUE within every submodel. By the uniqueness of UMVUEs, this implies

$$\hat{\mu}^* = \hat{\mu}_a \quad \text{Lebesgue–almost everywhere.}^1$$

However, the estimator $\hat{\mu}_a$ depends explicitly on the value of a. Since $a > 0$ was arbitrary, the above almost sure equality would have to hold simultaneously for all $a > 0$, which is impossible. This contradiction shows that no UMVUE for μ exists in the Behrens–Fisher problem.

The underlying intuition is that an optimal combination of the sample means must weight them according to their relative variances. If the variance ratio were known, the optimal weighting would be clear. When the ratio is unknown, no single unbiased estimator can uniformly dominate all variance–specific combinations, preventing the existence of a UMVUE. □

4.2.2 L_2–Orthogonality

A useful way to understand optimal unbiased estimation is through the geometry of L_2 spaces. Unbiased estimators of a fixed target $\gamma(\theta)$ form an affine space, and differences of such estimators lie in a linear space of estimators with mean zero. Since variance is simply the squared L_2-norm, the problem of finding a UMVUE becomes a problem of orthogonal projection. The following theorem formalizes this idea.

Theorem 4.1 (Characterization via L_2-orthogonality). *Let T be an unbiased estimator of $\gamma(\theta)$ with $\mathsf{Var}_\theta[T] < \infty$ for all $\theta \in \Theta$. For each $\theta \in \Theta$, define*

$$\mathcal{U}(\theta) = \left\{ U \;:\; \mathsf{E}_{\theta'}[U] = 0 \;\forall \theta' \in \Theta,\; \mathsf{Var}_\theta[U] < \infty \right\}.$$

Then T is a UMVUE of $\gamma(\theta)$ if and only if

$$\mathsf{Cov}_\theta[T, U] = \mathsf{E}_\theta[T \cdot U] = 0 \quad \forall \theta \in \Theta,\; \forall U \in \mathcal{U}(\theta).$$

[1] Null sets of any multivariate normal distribution coincide with sets of Lebesgue measure zero.

The condition above has a natural geometric interpretation. For $U \in \mathcal{U}(\theta)$, we have $\mathsf{E}_\theta[U] = 0$, and therefore

$$\mathsf{Cov}_\theta[T, U] = \mathsf{E}_\theta[(T - \mathsf{E}_\theta[T])U] = \mathsf{E}_\theta[T \cdot U].$$

Thus the covariance plays the role of an inner product in $L_2(\mathsf{P}_\theta)$. The theorem states that a UMVUE is precisely an unbiased estimator that is orthogonal, at every θ, to the subspace of unbiased estimators of zero. As in Euclidean geometry, orthogonality characterizes the point of minimal distance, here corresponding to minimal variance.

Remark (Uniqueness of the UMVUE). The orthogonality characterization immediately implies uniqueness. If both T and T' are UMVUEs of $\gamma(\theta)$, then $T - T' \in \mathcal{U}(\theta)$, and hence

$$\begin{aligned}\mathsf{Var}_\theta[T - T'] &= \mathsf{Cov}_\theta[T - T', T - T'] \\ &= \mathsf{Cov}_\theta[T, T - T'] - \mathsf{Cov}_\theta[T', T - T'] = 0.\end{aligned}$$

Consequently, $\mathsf{P}_\theta(T = T') = 1$ for all $\theta \in \Theta$.

Proof.

$\Longrightarrow$) Assume that T satisfies the orthogonality condition. Let T' be any other unbiased estimator of $\gamma(\theta)$, so that $\mathsf{E}_\theta[T - T'] = 0$ for all $\theta \in \Theta$.

Fix $\theta \in \Theta$. If $\mathsf{Var}_\theta[T'] = \infty$, then trivially $\mathsf{Var}_\theta[T] \leq \mathsf{Var}_\theta[T']$. Otherwise, suppose $\mathsf{Var}_\theta[T'] < \infty$. Since both T and T' have finite variance under P_θ, their difference satisfies

$$\mathsf{Var}_\theta[T - T'] = \mathsf{Var}_\theta[T] + \mathsf{Var}_\theta[T'] - 2\mathsf{Cov}_\theta[T, T'] < \infty,$$

and hence $T - T' \in \mathcal{U}(\theta)$.

By orthogonality,

$$0 = \mathsf{Cov}_\theta[T, T - T'] = \mathsf{Var}_\theta[T] - \mathsf{Cov}_\theta[T, T'].$$

Therefore,

$$\mathsf{Var}_\theta[T] = \mathsf{Cov}_\theta[T, T'] \leq \sqrt{\mathsf{Var}_\theta[T]\,\mathsf{Var}_\theta[T']}$$

by the Cauchy–Schwarz inequality. Dividing by $\sqrt{\mathsf{Var}_\theta[T]}$ yields $\mathsf{Var}_\theta[T] \leq \mathsf{Var}_\theta[T']$. Since T' was arbitrary, T is a UMVUE.

$\Longleftarrow$) Suppose T is a UMVUE and let $U \in \mathcal{U}(\theta)$. For any $a \in \mathbb{R}$, the estimator $T + aU$ is unbiased for $\gamma(\theta)$. Optimality of T implies that the function $a \mapsto \mathsf{Var}_\theta[T + aU]$ is minimized at $a = 0$, which forces $\mathsf{Cov}_\theta[T, U] = 0$.

□

This characterization is often useful in practice. In some models, such as the binomial family, one can directly verify the orthogonality condition to identify a UMVUE. Conversely, the failure of such an orthogonality relation may be used to show that no UMVUE exists. In this sense, the theorem provides both a constructive and a diagnostic tool in the theory of unbiased estimation.

Some Preparation on Conditioning

Conditioning provides a powerful and often simpler alternative to direct L_2-orthogonality arguments when constructing good estimators. Beyond its interpretation as an averaging or integration operation, conditional expectation can be understood as the optimal prediction of one random variable based on the information contained in another. This viewpoint naturally leads to variance decompositions and explains why conditioning typically reduces variability.

Throughout this section, let Y and Z be random variables such that $\mathsf{Var}[Y] < \infty$.

Lemma 4.1 (Basic identities for conditioning)**.** *The following properties hold:*

i) Law of total expectation (tower rule):

$$\mathsf{E}[Y] = \mathsf{E}[\mathsf{E}[Y \mid Z]].$$

ii) Law of total variance:

$$\mathsf{Var}[Y] = \mathsf{Var}[\mathsf{E}[Y \mid Z]] + \mathsf{E}[\mathsf{Var}[Y \mid Z]].$$

iii) Degeneracy conditions:

$$\mathsf{Var}[\mathsf{E}[Y \mid Z]] = 0 \iff \mathsf{E}[Y \mid Z] \textit{ is a.s. constant},$$

and

$$\mathsf{E}[\mathsf{Var}[Y \mid Z]] = 0 \iff Y = \mathsf{E}[Y \mid Z] \quad \textit{a.s.}$$

These identities reveal how the variability of Y decomposes into two distinct components. The term $\mathsf{Var}[\mathsf{E}[Y \mid Z]]$ measures the variability explained by the information in Z, while $\mathsf{E}[\mathsf{Var}[Y \mid Z]]$ captures the remaining randomness after conditioning. In particular, conditioning can never increase variance.

Proof.

i) The first identity is the law of total expectation.

ii) Observe that

$$\begin{aligned}\mathsf{Var}[\mathsf{E}[Y \mid Z]] &= \mathsf{E}\big[(\mathsf{E}[Y \mid Z])^2\big] - (\mathsf{E}[\mathsf{E}[Y \mid Z]])^2 \\ &= \mathsf{E}\big[(\mathsf{E}[Y \mid Z])^2\big] - (\mathsf{E}[Y])^2,\end{aligned}$$

while

$$\begin{aligned}\mathsf{E}[\mathsf{Var}[Y \mid Z]] &= \mathsf{E}\big[\mathsf{E}[Y^2 \mid Z] - (\mathsf{E}[Y \mid Z])^2\big] \\ &= \mathsf{E}[Y^2] - \mathsf{E}\big[(\mathsf{E}[Y \mid Z])^2\big].\end{aligned}$$

Adding these two expressions yields

$$\mathsf{Var}[\mathsf{E}[Y \mid Z]] + \mathsf{E}[\mathsf{Var}[Y \mid Z]] = \mathsf{E}[Y^2] - (\mathsf{E}[Y])^2 = \mathsf{Var}[Y].$$

iii) Note that

$$\begin{aligned}\mathsf{E}\big[(Y-\mathsf{E}[Y\mid Z])^2\big] &= \mathsf{E}[Y^2]-2\mathsf{E}[Y\,\mathsf{E}[Y\mid Z]]+\mathsf{E}\big[(\mathsf{E}[Y\mid Z])^2\big]\\ &= \mathsf{E}[Y^2]-2\mathsf{E}[\mathsf{E}[Y\,\mathsf{E}[Y\mid Z]\mid Z]]+\mathsf{E}\big[(\mathsf{E}[Y\mid Z])^2\big]\\ &= \mathsf{E}[Y^2]-\mathsf{E}\big[(\mathsf{E}[Y\mid Z])^2\big]\\ &= \mathsf{E}[\mathsf{Var}[Y\mid Z]].\end{aligned}$$

Thus $\mathsf{E}[\mathsf{Var}[Y\mid Z]]=0$ if and only if $Y=\mathsf{E}[Y\mid Z]$ almost surely. Similarly, $\mathsf{Var}[\mathsf{E}[Y\mid Z]]=0$ holds if and only if $\mathsf{E}[Y\mid Z]$ is almost surely constant.

□

From a geometric perspective, $\mathsf{E}[Y\mid Z]$ is the best approximation to Y among all functions of Z in the L_2-sense. That is, it minimizes

$$\mathsf{E}\big[(Y-g(Z))^2\big]$$

over all measurable functions g. This variance-minimization property explains why conditioning is such an effective tool in estimation theory: replacing a statistic by its conditional expectation given additional information can only reduce variance.

4.2.3 Rao–Blackwell Theorem

A central principle in estimation theory is that *conditioning reduces variance.* The Rao–Blackwell theorem formalizes this idea and shows how an arbitrary unbiased estimator can be systematically improved by conditioning on a sufficient statistic. This result provides one of the main practical tools for constructing optimal unbiased estimators.

Theorem 4.2 (Rao–Blackwell)**.** *Let T be an unbiased estimator of $\gamma(\theta)$ with $\mathsf{Var}_\theta[T]<\infty$ for all $\theta\in\Theta$. Suppose that S is a sufficient statistic for θ. Define*

$$T^* := \mathsf{E}_*[T\mid S]\equiv \mathsf{E}_\theta[T\mid S].$$

Then T^* is an unbiased estimator of $\gamma(\theta)$ and satisfies

$$\mathsf{Var}_\theta[T^*]\le \mathsf{Var}_\theta[T]\quad \forall\theta\in\Theta.$$

Moreover,

$$\mathsf{Var}_\theta[T^*]<\mathsf{Var}_\theta[T]\quad \textit{unless } T=T^* \textit{ a.s. under } \mathsf{P}_\theta.$$

Remark. The logic of the theorem has two complementary components. First, conditioning on any statistic can only decrease variance, since

$$\mathsf{Var}_\theta[T]=\mathsf{Var}_\theta[\mathsf{E}_\theta[T\mid S]]+\mathsf{E}_\theta[\mathsf{Var}_\theta[T\mid S]],$$

and the second term is nonnegative. Second, sufficiency of S ensures that the conditional expectation $\mathsf{E}_*\,[T|S]\equiv\mathsf{E}_\theta[T\mid S]$ does not depend on the unknown parameter θ, so that T^* is a **well-defined statistic**, i.e. a function of the data alone.

Proof. Unbiasedness follows from the law of total expectation:

$$\mathsf{E}_\theta[T^*] = \mathsf{E}_\theta[\mathsf{E}_*[T \mid S]] = \mathsf{E}_\theta[\mathsf{E}_\theta[T \mid S]] = \mathsf{E}_\theta[T] = \gamma(\theta), \quad \forall\theta \in \Theta.$$

To compare variances, apply the law of total variance by part (ii) of Lemma 4.1:

$$\mathsf{Var}_\theta[T] = \mathsf{Var}_\theta[T^*] + \mathsf{E}_\theta\underbrace{[\mathsf{Var}_*[T \mid S]]}_{\geq 0} \geq \mathsf{Var}_\theta[T^*].$$

Equality holds if and only if $\mathsf{E}_\theta[\mathsf{Var}_*[T \mid S]] = 0$, which by (iii) of Lemma 4.1 occurs precisely when $T = \mathsf{E}_\theta[T \mid S]$ almost surely. □

The Rao–Blackwell theorem thus provides a universal variance-improvement procedure: starting from any unbiased estimator, conditioning on a sufficient statistic yields a new estimator that is at least as good, and typically strictly better. This result will serve as the foundation for constructing uniformly minimum variance unbiased estimators in subsequent sections.

4.2.4 Complete Statistics

Completeness is an additional structural property of statistics that plays a central role in identifying estimators that cannot be further improved by conditioning. As with sufficiency, completeness is always defined relative to a fixed statistical model.

Notation. *For a model $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ and random variables X and Y, we write*

$$X = Y \ [\mathcal{P}\text{–}a.e.]$$

if $\mathsf{P}_\theta(X = Y) = 1$ for all $\theta \in \Theta$. That is, the equality holds almost surely under every distribution in the model.

Definition 4.3 (Complete statistic)**.** A statistic S is said to be **complete** for the model $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ if for every measurable function h,

$$\mathsf{E}_\theta[h(S)] = 0 \quad \forall\theta \in \Theta \implies h(S) = 0 \ [\mathcal{P}\text{–a.e.}].$$

Thus, completeness requires that the only function of S whose expectation vanishes under all distributions in the model is the trivial one. In particular, while a random variable may have expectation zero under many different distributions, completeness rules out such nontrivial cancellations when the random variable is a function of a complete statistic.

The importance of completeness becomes apparent in conjunction with the Rao–Blackwell theorem. Conditioning an unbiased estimator on a sufficient statistic can only reduce variance; however, unless the sufficient statistic is complete, further variance reduction may still be possible. Complete sufficient statistics are precisely those for which this process cannot be continued, leading to estimators that are optimal among all unbiased estimators.

As a structural result, known as **Bahadur's theorem**,

$$\text{complete and sufficient} \implies \text{minimal sufficient.}$$

The converse does not hold in general: a minimal sufficient statistic need not be complete.

Several examples of complete statistics will be discussed in the following sections.

4.2.5 Lehmann–Scheffé Theorem

The Lehmann–Scheffé theorem formalizes the idea that conditioning on a statistic that is both sufficient and complete does not merely improve an unbiased estimator, but actually produces the *unique optimal* one.

Theorem 4.3 (Lehmann–Scheffé). *In the model $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$, let T be an unbiased estimator of $\gamma(\theta)$ with $\mathsf{Var}_\theta(T) < \infty$ for all $\theta \in \Theta$. If S is a sufficient and complete statistic, then*

$$T^* := \mathsf{E}_*[T \mid S]$$

is the UMVUE of $\gamma(\theta)$.

Motivation and intuition.
Conditioning on a sufficient statistic S (via Rao–Blackwell) always improves variance among unbiased estimators. However, conditioning alone does not guarantee optimality: if we condition on a merely sufficient (or even minimal sufficient) statistic, different unbiased estimators may still lead to different conditional expectations.

Completeness is the additional assumption that rules this out. It says that any function of S whose expectation is zero for all θ must be the zero function. This forces all unbiased estimators that are functions of S to coincide almost surely, leaving only a single candidate. That candidate must therefore be the uniformly minimum variance unbiased estimator.

Proof. As in the proof of the Rao–Blackwell Theorem 4.2, the estimator T^* is unbiased because

$$\mathsf{E}_\theta[T^*] = \mathsf{E}_\theta[\mathsf{E}_*[T \mid S]] = \mathsf{E}_\theta[\mathsf{E}_\theta[T \mid S]] = \mathsf{E}_\theta[T] = \gamma(\theta).$$

Now let T' be any other unbiased estimator of $\gamma(\theta)$. By sufficiency of S, using Rao–Blackwell yields an improved unbiased estimator

$$T'' := \mathsf{E}_*[T' \mid S],$$

which is a function of S.

Both T^* and T'' are functions of S and satisfy

$$\mathsf{E}_\theta[T^* - T''] = \gamma(\theta) - \gamma(\theta) = 0 \quad \text{for all } \theta \in \Theta.$$

Hence, $T^* - T''$ is a function of S with expectation zero for all θ. By completeness of S, this implies

$$T^* = T'' \qquad [\mathcal{P}\text{-a.e.}].$$

Therefore, every unbiased estimator, after conditioning on S, collapses to the same estimator T^*. This estimator is thus the $\mathcal{P}$-almost surely unique UMVUE of $\gamma(\theta)$. □

Note. If an unbiased estimator T is already a function of the complete and sufficient statistic S, then it is automatically the UMVUE, since

$$T = \mathsf{E}_*[T \mid S].$$

Example 4.3 (Binomial model)**.** Let $X \sim \text{Binomial}(n, \theta)$ with $\theta \in (0, 1)$. Since the entire sample reduces to the total number of successes, X is (trivially) a sufficient statistic.

To verify completeness, let h be any function such that

$$\mathsf{E}_\theta[h(X)] = 0 \quad \text{for all } \theta \in (0, 1).$$

Then

$$\mathsf{E}_\theta[h(X)] = \sum_{x=0}^{n} h(x) \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

which is a polynomial in θ of degree at most n. If this polynomial vanishes for all $\theta \in (0, 1)$, then it must be the zero polynomial, and hence all of its coefficients are zero.

In particular, the constant term satisfies

$$h(0)\binom{n}{0} = h(0) = 0.$$

Proceeding inductively, one concludes that $h(x) = 0$ for all $x = 0, \ldots, n$. Therefore, X is a complete statistic.

We conclude that X is both sufficient and complete.

Now consider estimation. Since $\mathsf{E}_\theta[X] = n\theta$, the estimator

$$\hat{\theta} = \frac{1}{n} X$$

is unbiased for θ. As a function of the complete and sufficient statistic X, it follows from the Lehmann–Scheffé theorem that $\hat{\theta}$ is the UMVUE of θ.

Similarly, to estimate θ^2, note that

$$\mathsf{E}_\theta\left[\frac{X(X-1)}{n(n-1)}\right] = \theta^2.$$

Thus,

$$\widehat{\theta^2} = \frac{X(X-1)}{n(n-1)}$$

is an unbiased estimator and, being a function of the complete and sufficient statistic X, is the UMVUE of θ^2.

Example 4.4 (Uniform distributions)**.** Let $X_1, \dots, X_n$ be i.i.d. Uniform$(0, \theta)$ with $\theta > 0$. The parameter θ is the unknown upper endpoint of the support.

As seen earlier, the statistic

$$T(X) = \max\{X_1, \dots, X_n\}$$

is sufficient. Intuitively, all information about the endpoint θ is contained in the largest observation.

To determine the distribution of T, we compute its distribution function. For $0 \leq t \leq \theta$,

$$\mathsf{P}_\theta(T \leq t) = \mathsf{P}_\theta(X_i \leq t \ \forall i) = \left(\frac{t}{\theta}\right)^n,$$

since the X_i are i.i.d. Uniform$(0, \theta)$. Differentiating yields the density [2]

$$p_\theta^T(t) = \frac{nt^{n-1}}{\theta^n} \mathbf{1}_{[0,\theta]}(t).$$

Claim. The estimator $\frac{n+1}{n}T$ is the UMVUE of θ.

Proof. We first show unbiasedness. Using the density of T,

$$\mathsf{E}_\theta[T] = \int_0^\theta t \cdot \frac{nt^{n-1}}{\theta^n}\, dt = \frac{n}{n+1}\theta,$$

and hence $\frac{n+1}{n}T$ is unbiased for θ.

To apply the Lehmann–Scheffé theorem, it remains to verify that T is complete. Let $h : \mathbb{R} \to \mathbb{R}$ satisfy

$$\mathsf{E}_\theta[h(T)] = 0 \quad \text{for all } \theta > 0.$$

Then

$$\begin{aligned} \mathsf{E}_\theta[h(T)] = \int_0^\theta h(t)\, \frac{nt^{n-1}}{\theta^n}\, dt = 0 \quad \forall\, \theta > 0 \\ \iff \int_0^\theta h(t) t^{n-1}\, dt = 0 \quad \forall\, \theta > 0. \end{aligned}$$

Since the integral of $h(t)t^{n-1}$ over every interval $[0, \theta]$ vanishes, it follows that

$$h(t)t^{n-1}\mathbf{1}_{[0,\infty)}(t) = 0 \quad \text{Lebesgue-a.e.}^3$$

Consequently, $h(T) = 0$ P_θ-a.s. for all $\theta > 0$, and T is complete.

Since T is both sufficient and complete, the Lehmann–Scheffé theorem implies that $\frac{n+1}{n}T$ is the UMVUE of θ. □

[2] Equivalently, $\frac{1}{\theta}T \sim \text{Beta}(n, 1)$.

[3] If a locally integrable function integrates to zero over all intervals $[0, \theta]$, then it must vanish almost everywhere; see, e.g., Shorack (2017), p. 44.

4.3 Completeness in Exponential Families

A large and important class of models in which completeness arises almost automatically is given by exponential families. In this setting, completeness is closely related to uniqueness properties of Laplace and characteristic transforms; see, for instance, Lehmann and Romano (2005, Theorem 4.3.1).[4]

Theorem 4.4. *Let $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$ be an exponential family with natural parameter $\eta(\theta)$ and sufficient statistic $T(X)$. That is, each P_θ admits a density (with respect to a σ–finite measure ν) of the form*

$$p_\theta(x) = \exp\{\langle \eta(\theta), T(x)\rangle - B(\theta)\}\, h(x).$$

If the set $\{\eta(\theta) : \theta \in \Theta\}$ has non–empty interior, then T is complete.

Interpretation. The theorem states that, for a *full* exponential family, sufficiency already implies completeness. The condition that the natural parameter space has non–empty interior means that the model is not artificially restricted to a lower–dimensional subset of the natural parameter space.

Together with the Lehmann–Scheffé theorem, this result immediately yields UMVUEs in many classical models, including the binomial model (Example 4.3). In contrast, the uniform model (Example 4.4) is not an exponential family, but still admits a complete sufficient statistic via a separate argument.

Why varying support rules out exponential families. In an exponential family, the density is strictly positive wherever $h(x) > 0$, and the function h does not depend on θ. Consequently, all distributions in the family share the same support. Models in which the support depends on θ, such as the Uniform$(0, \theta)$ family, cannot be exponential families.

Role of the interior condition. If the natural parameter space $\{\eta(\theta)\}$ contains a full–dimensional open set, then completeness follows. For example, in the normal family with unknown mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$, the natural parameters

$$\left(\frac{\mu}{\sigma^2}, \frac{1}{\sigma^2}\right)$$

range over an open subset of $\mathbb{R}^2$, and the corresponding sufficient statistics are complete.

However, if the model is restricted, completeness may fail. For instance, if one considers only normal distributions satisfying a constraint that links mean and variance (e.g. $\mu = \sigma^2$), then the natural parameters lie on a lower–dimensional subset of $\mathbb{R}^2$. In this case, the parameter space has empty interior, and the theorem no longer applies.

[4] The proof reduces the problem to uniqueness of characteristic functions.

Remark. Completeness is a property of all measurable functions of the statistic. Even if a statistic takes values only in $(0, \infty)$, this alone does not imply completeness, since arbitrary functions of the statistic may change signs. Completeness requires that *every* function with zero expectation for all parameter values must vanish almost surely.

Although the result may appear abstract, it is ultimately a consequence of familiar uniqueness principles for Laplace and characteristic transforms.

Example 4.5 (Estimating a Gaussian probability)**.** Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, 1)$ with unknown mean μ and known variance, and assume $n \geq 2$.

Parameter of interest. For a fixed $x \in \mathbb{R}$, we aim to estimate

$$\gamma(\mu) := \mathsf{P}_\mu(X_1 \leq x) = \Phi(x - \mu),$$

where Φ denotes the distribution function of $\mathcal{N}(0, 1)$.

Unbiased estimation. Since probabilities are expectations of indicator functions, an unbiased (but inefficient) estimator is immediately available:

$$T(X_1, \ldots, X_n) = \mathbf{1}_{(-\infty, x]}(X_1), \qquad \mathsf{E}_\mu[T] = \gamma(\mu).$$

Towards a UMVUE. The normal location model with known variance is a one–parameter exponential family. Its sufficient statistic $\overline{X}_n$ is also complete (cf. section 4.3). Hence, by the Lehmann–Scheffé theorem, the UMVUE of $\gamma(\mu)$ is obtained by conditioning T on $\overline{X}_n$:

$$\begin{aligned} T^*(\bar{x}_n) &= \mathsf{E}_*\big[T(\boldsymbol{X}) \mid \overline{X}_n = \bar{x}_n\big] \\ &= \mathsf{P}_*(X_1 \leq x \mid \overline{X}_n = \bar{x}_n) \\ &= \mathsf{P}_*(X_1 - \overline{X}_n \leq x - \bar{x}_n \mid \overline{X}_n = \bar{x}_n). \end{aligned}$$

Evaluation. The pair $(X_1 - \overline{X}_n, \overline{X}_n)$ is jointly Gaussian, being a linear transformation of $(X_1, \ldots, X_n)$. A direct computation shows that

$$\mathsf{Cov}(X_1 - \overline{X}_n, \overline{X}_n) = 0,$$

and hence $X_1 - \overline{X}_n$ and $\overline{X}_n$ are independent. Therefore,

$$T^*(\bar{x}_n) = \mathsf{P}(X_1 - \overline{X}_n \leq x - \bar{x}_n).$$

Since $X_1 - \overline{X}_n \sim \mathcal{N}\big(0, 1 - \frac{1}{n}\big)$, we obtain

$$\mathsf{P}\left(\frac{X_1 - \overline{X}_n}{\sqrt{1 - \frac{1}{n}}} \leq \frac{x - \bar{x}_n}{\sqrt{1 - \frac{1}{n}}}\right) = \Phi\left((x - \bar{x}_n)\frac{1}{\sqrt{1 - \frac{1}{n}}}\right).$$

Conclusion. The UMVUE of $\gamma(\mu) = \mathsf{P}_\mu(X_1 \le x)$ is

$$T^*(\overline{X}_n) = \Phi\left((x - \overline{X}_n)\frac{1}{\sqrt{1-\frac{1}{n}}}\right).$$

4.4 Nonparametric Unbiased Estimation

Let $X_1, \dots, X_n$ be i.i.d. random variables taking values in $\mathbb{R}$ with distribution function (d.f.) F.

Motivation and Model
So far we have mostly discussed parametric models, where we assume a specific distributional shape but unknown finite-dimensional parameters. For example:

- $X_i \sim \mathsf{Normal}(\mu, \sigma^2)$ with unknown μ and σ^2,
- $X_i \sim \mathsf{Poisson}(\lambda)$ with unknown rate λ.

In contrast, *nonparametric* models are infinite-dimensional: we assume less about the shape of the distribution. For instance, we might only assume that F has a density, is smooth, or log-concave, without specifying a particular parametric family.

Here we consider a general nonparametric model. Let F belong to a class $\mathcal{F}$ of distribution functions with the following properties:

- $\mathcal{F}$ is *convex*: any convex combination of distributions in $\mathcal{F}$ also belongs to $\mathcal{F}$,
- All $F \in \mathcal{F}$ are *absolutely continuous* (i.e., they have a density f with respect to Lebesgue measure),
- $\mathcal{F}$ contains all *uniform distributions.*

This is a very large class of distributions: all have densities, but otherwise the assumptions are minimal.

Order Statistics

We know from general theory that for i.i.d. samples, the order of the data points contains no additional information. Therefore, the *order statistics*

$$T(X_1, \dots, X_n) = (X_{(1)}, \dots, X_{(n)})$$

are *sufficient* statistics. Here, $X_{(1)}$ is the smallest observation, $X_{(2)}$ the second smallest, and so on up to $X_{(n)}$.

Interestingly, in this nonparametric model, the order statistics are not just sufficient—they are also *complete.*

Theorem 4.5. *The order statistics*

$$T(X_1, \dots, X_n) = (X_{(1)}, \dots, X_{(n)})$$

are complete for the class $\mathcal{F}$.

Proof. To show completeness, we must prove that if $h(T)$ satisfies

$$\mathsf{E}_F[h(T)] = 0 \quad \text{for all } F \in \mathcal{F},$$

then $h(T) = 0$ almost surely for all $F \in \mathcal{F}$.

Step 1: Consider mixture distributions. Let $F_1, \dots, F_n \in \mathcal{F}$ with densities $f_1, \dots, f_n$ and weights $\alpha_1, \dots, \alpha_n > 0$. Form the mixture

$$F = \frac{1}{\sum_{i=1}^n \alpha_i} \sum_{i=1}^n \alpha_i F_i,$$

which belongs to $\mathcal{F}$ by convexity. Its density is

$$f = \frac{1}{\sum_{i=1}^n \alpha_i} \sum_{i=1}^n \alpha_i f_i.$$

Step 2: Express expectation as a polynomial. By assumption, $\mathsf{E}_F[h(T)] = 0$, i.e.,

$$\int h(T(x_1, \dots, x_n)) \prod_{j=1}^n f(x_j)\, dx = 0.$$

Substituting the mixture density gives

$$\int h(T(x_1, \dots, x_n)) \prod_{j=1}^n \Big(\sum_{i=1}^n \alpha_i f_i(x_j) \Big) dx = 0.$$

This is a polynomial in $\alpha_1, \dots, \alpha_n$. Since it vanishes for all $\alpha_i > 0$, all coefficients must be zero.

Step 3: Focus on the coefficient of $\prod_i \alpha_i$. This coefficient is

$$\sum_{\pi \in S_n} \int h(T(x_1, \dots, x_n)) \prod_{i=1}^n f_i(x_{\pi(i)})\, dx.$$

Reindexing the integration variables using the permutation π gives

$$n! \int h(T(x_1, \dots, x_n)) \prod_{i=1}^n f_i(x_i)\, dx = 0.$$

Step 4: Reduce to uniform distributions. Choose $f_i(x) = \frac{1}{b_i - a_i}\mathbf{1}_{[a_i,b_i]}(x)$, i.e., $F_i \sim \text{Uniform}(a_i, b_i)$. Then

$$0 = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} h(T(x_1, \ldots, x_n))\, dx_1 \ldots dx_n.$$

Step 5: Conclude by rectangles. Since this holds for all rectangles $[a_1, b_1] \times \cdots \times [a_n, b_n]$, a standard result from integration theory implies

$$h(T) \equiv 0 \quad \text{almost everywhere.}$$

Because all $F \in \mathcal{F}$ are absolutely continuous, we conclude

$$\mathsf{P}_F(h(T) = 0) = 1 \quad \text{for all } F \in \mathcal{F}.$$

□

Discussion. This shows that in this very general nonparametric model, the order statistics are both sufficient and complete. By the Lehmann–Scheffé theorem, any unbiased estimator of a parameter can be improved (or obtained) by conditioning on the order statistics. The completeness property here relies on the richness of $\mathcal{F}$ and the convexity/mixture argument.

4.4.1 U–Statistics

In statistics, we often want to estimate parameters that are *functions of several observations at a time*, not just of single data points. This leads naturally to the notion of **U–statistics**, which generalize many familiar estimators such as the sample mean or variance.

A U–statistic is built from a *kernel function* h that acts on m data points at a time:

$$h \,:\, \mathcal{X}^m \to \mathbb{R}, \quad \text{for } \mathcal{X} \subseteq \mathbb{R}.$$

We usually assume that h is *symmetric*, meaning that permuting its arguments does not change its value:

$$h(x_{\pi(1)}, \ldots, x_{\pi(m)}) = h(x_1, \ldots, x_m) \quad \forall \pi \in S_m,\ \forall x \in \mathcal{X}.$$

If h is not symmetric, we can symmetrize it by averaging over all permutations:

$$\tilde{h}(x_1, \ldots, x_m) = \frac{1}{m!} \sum_{\pi \in S_m} h(x_{\pi(1)}, \ldots, x_{\pi(m)}).$$

Definition 4.4. For $n \geq m$, the **U–statistic** of order m with kernel h is

$$U_n(x_1, \ldots, x_n) = \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} h(x_{i_1}, \ldots, x_{i_m}).$$

Intuitively, U_n computes h on *every subset of size* m of the data and averages the results.

Intuition Think of your dataset as a vector in $\mathbb{R}^n$. A U–statistic of order m examines all possible subsets of size m, applies a function h to each, and averages the outputs. This approach captures interactions among m points and leads to unbiased estimators for parameters that are expectations of such functions.

4.4.2 U–Statistics and UMVUE

We now connect U–statistics to *optimal unbiased estimation.* Suppose we want to estimate a parameter that is the expectation of a kernel function h applied to m i.i.d. observations.

Theorem 4.6. *Let $\mathcal{F}$ be a class of distributions satisfying:*

- *$\mathcal{F}$ is convex,*
- *$\mathcal{F}$ contains only absolutely continuous distributions,*
- *$\mathcal{F}$ contains all uniform distributions.*

If $X_1, \ldots, X_n$ are i.i.d. with d.f. $F \in \mathcal{F}$ and $\mathsf{E}_F[h(X_1, \ldots, X_m)^2] < \infty$ for all $F \in \mathcal{F}$, then the U–statistic U_n is the ***UMVUE*** *of*

$$\gamma(F) = \mathsf{E}_F[h(X_1, \ldots, X_m)].$$

Proof.

i) **Unbiasedness:** By linearity of expectation,

$$\mathsf{E}_F[U_n(X_1, \ldots, X_n)] = \frac{1}{\binom{n}{m}} \sum_{1 \le i_1 < \cdots < i_m \le n} \mathsf{E}_F[h(X_{i_1}, \ldots, X_{i_m})] = \gamma(F),$$

because each term in the sum has expectation $\gamma(F)$ and there are exactly $\binom{n}{m}$ terms.

ii) **Optimality via Lehmann–Scheffé:** U_n is symmetric in its arguments, so it is a function of the *order statistics* $X_{(1)}, \ldots, X_{(n)}$. Since order statistics are complete and sufficient for $\mathcal{F}$, the Lehmann–Scheffé theorem guarantees that U_n is UMVUE.

□

Examples

- **Sample mean:** If $\gamma(F) = \mathsf{E}_F[X_1]$, take $m = 1$ and $h(x_1) = x_1$. Then

$$U_n(\boldsymbol{X}) = \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- **Squared mean:** For $\gamma(F) = (\mathsf{E}_F[X_1])^2$, use $m = 2$ and $h(x_1, x_2) = x_1 x_2$. Then
$$U_n(\boldsymbol{X}) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} X_i X_j,$$
which is an unbiased estimator of $(\mathsf{E}[X_1])^2$ because $\mathsf{E}[X_i X_j] = (\mathsf{E}[X_1])^2$ for independent X_i, X_j.

- **Variance:** For $\gamma(F) = \mathsf{Var}_F[X_1]$, noting that $\mathsf{Var}_F[X_1] = \mathsf{E}_F[X_1(X_1 - X_2)]$ and symmetrizing $x_1(x_1 - x_2)$ yields this kernel
$$h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2, \quad m = 2.$$
The corresponding U–statistic
$$U_n(\boldsymbol{X}) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \frac{(X_i - X_j)^2}{2}$$
is the unbiased sample variance.

Takeaways U–statistics provide a systematic way to construct unbiased estimators of parameters that are expectations of functions of multiple observations. The kernel h encodes which combinations of observations are relevant, and the averaging over subsets ensures unbiasedness. In large nonparametric models, U–statistics often produce the *best possible unbiased estimators*, capturing quantities like the mean, variance, or higher-order moments in a unified framework.

5. Cramér-Rao Lower Bound

This chapter presents a lower bound on the variance of an estimator, which offers a different perspective on optimality in the context of unbiased estimation. While previous chapters focused on characterizing optimal estimators—such as those found through complete statistics or conditioning—this “lower bound” approach provides a complementary perspective by identifying fundamental barriers to estimation accuracy.

The core idea is to establish a threshold that any estimator (under specific assumptions like unbiasedness) must respect; if an estimator’s variance matches this lower bound, it is proven to be optimal as no better performance is possible. This bound also illustrates how estimation accuracy is impacted by the presence of nuisance parameters (i.e., parameters that are unknown but not of direct interest). For further reading, consider, e.g., Wellner (2018, chap. 3).

Remark. The study of fundamental barriers is common in statistical literature. For instance, in minimax estimation, one seeks to prove lower bounds on the minimax risk. Proving a lower bound and subsequently exhibiting an estimator that achieves it (a matching upper bound) is a standard method to establish optimality. In this course, we focus on the simplest instance: the Cramér-Rao lower bound for unbiased estimation.

Our starting point is the fundamental question:

How well can one estimate a parameter?

Setup

- $X \sim P_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^k$. We assume Θ is an **open subset** to allow for differentiation with respect to the parameters.
- Densities: $p_{\boldsymbol{\theta}}(x) = \frac{dP_{\boldsymbol{\theta}}}{d\nu}(x), x \in \mathcal{X}$. We assume the model is smoothly parametrized so that these densities are differentiable.
- Target Parameter: $\gamma(\boldsymbol{\theta}) \in \mathbb{R}^r$ for a differentiable map $\gamma : \Theta \to \mathbb{R}^r$.
- The Jacobian of γ is the $r \times k$ matrix:

$$\dot{\gamma}(\boldsymbol{\theta}) = \left(\frac{\partial \gamma_i(\boldsymbol{\theta})}{\partial \theta_j} \right)_{ij}.$$

Definition 5.1 (Bias)**.** Suppose we have an estimator $T : \mathcal{X} \to \mathbb{R}^r$ such that $\mathsf{E}_{\boldsymbol{\theta}}[|T|] < \infty$ for all $\boldsymbol{\theta} \in \Theta$. The **bias** of the estimator is defined as

$$b(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}[T] - \gamma(\boldsymbol{\theta}).$$

Question: How small can $\mathsf{Var}_{\boldsymbol{\theta}}[T]$ be in relationship to its bias?

Specifically, for unbiased estimators, this question asks for the minimum achievable Mean Square Error (MSE).

5.1 Score Function and Fisher Information

Definition 6.1. Assuming existence of the derivatives and expectations, we define

i. **Score function**

$$\dot{\ell}_{\boldsymbol{\theta}}(x) = \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(x) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\theta}_1} \log p_{\boldsymbol{\theta}}(x) \\ \vdots \\ \frac{\partial}{\partial \boldsymbol{\theta}_k} \log p_{\boldsymbol{\theta}}(x) \end{pmatrix}.$$

ii. **Fisher Information**

$$I(\theta) = \mathsf{Var}_{\theta}[\dot{\ell}_{\theta}(X)].$$

In the sequel we derive a lower bound on $\mathsf{Var}_{\theta}[T]$, the Cramér–Rao bound or also information inequality, in which the Fisher information will play a key role.

5.2 Cramér–Rao Bound

If $r = k = 1$, the Cramér–Rao (CR) bound will take the form

$$\mathsf{Var}_{\boldsymbol{\theta}}[T] \geq \frac{\left(\dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta})\right)^2}{\mathrm{I}(\boldsymbol{\theta})}, \quad \boldsymbol{\theta} \in \Theta.$$

If T is **unbiased**, then

$$\mathsf{Var}_{\boldsymbol{\theta}}[T] \geq \frac{(\dot{\gamma}(\boldsymbol{\theta}))^2}{\mathrm{I}(\boldsymbol{\theta})}, \quad \boldsymbol{\theta} \in \Theta. \tag{5.1}$$

If $r = 1$ and $k > 1$, the bound in (5.1) generalizes to

$$\mathsf{Var}_{\boldsymbol{\theta}}[T] \geq \dot{\gamma}(\boldsymbol{\theta}) \mathrm{I}^{-1}(\boldsymbol{\theta}) \dot{\gamma}(\boldsymbol{\theta})^{\top}.$$

If $r \geq 2$, a lower bound on the $r \times r$ covariance matrix $\mathsf{Var}_{\theta}[T]$ may be given in terms of the Löwner (or positive semidefinite) ordering:

$$A \succeq B \iff A - B \succeq 0 \iff y^{\top}(A - B)y \geqslant 0 \quad \forall y \in \mathbb{R}^r.$$

Assumptions:

(M1) Θ is an open subset of $\mathbb{R}^k$;

(M2) $\exists B \subset \mathcal{X}$ with $\nu(B) = 0$ s.t. $\forall x \notin B : \theta \mapsto p_\theta(x)$ is differentiable;

(M3) $A := \{x : p_\theta(x) = 0\}$ does not depend on θ;[1]

(M4) $I(\theta)$ is finite and positive definite for all $\theta \in \Theta$;

(M5) The functions $\boldsymbol{\theta} \mapsto \int p_\theta(x) dv(x)$ and $\boldsymbol{\theta} \mapsto \int T_i(x) p_\theta(x) dv(x)$, $1 \leq i \leq r$, can be differentiated (wrt. $\boldsymbol{\theta}$) under the integral sign. (Of course, $\int p_\theta(x) dv(x) \equiv 1$.)

Theorem 5.1 (Cramér–Rao bound / Information inequality)**.** *If (M1)-(M5) holds, then*

$$\mathsf{Var}_{\boldsymbol{\theta}}[T(X)] \succeq \left(\dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta})\right) \mathbf{I}^{-1}(\boldsymbol{\theta}) \left(\dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta})\right)^\top, \quad \boldsymbol{\theta} \in \Theta.$$

Proof.

$$\mathsf{Var}_{\boldsymbol{\theta}}\left[\begin{pmatrix} T(X) \\ \dot{\ell}_{\boldsymbol{\theta}}(X) \end{pmatrix}\right] = \begin{pmatrix} \mathsf{Var}_{\boldsymbol{\theta}}[T(X)] & \mathsf{Cov}_{\boldsymbol{\theta}}[T(X), \dot{\ell}_{\boldsymbol{\theta}}(X)] \\ \mathsf{Cov}_{\boldsymbol{\theta}}[T(X), \dot{\ell}_{\boldsymbol{\theta}}(X)] & \mathsf{Var}_{\boldsymbol{\theta}}[\dot{\ell}_{\boldsymbol{\theta}}(X)] \end{pmatrix}$$

By definition, $\mathsf{Var}_\theta[\dot{\ell}_\theta(X)] = I(\theta)$. Under our assumptions:

$$\boxed{\text{Fact: } \mathsf{E}_{\boldsymbol{\theta}}[\dot{\ell}_{\boldsymbol{\theta}}(X)] = 0} \tag{5.2}$$

Indeed, for $j = 1, \ldots, k$:

$$\begin{aligned} \mathsf{E}_{\boldsymbol{\theta}}[\dot{\ell}_{\boldsymbol{\theta},j}(X)] &= \int \left(\frac{\partial}{\partial \boldsymbol{\theta}_j} \log p_{\boldsymbol{\theta}}(x)\right) p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\nu(x) \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}_j} p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\nu(x) \overset{(M5)}{=} \frac{\partial}{\partial \boldsymbol{\theta}_j} \int p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\nu(x) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}_j} 1 = 0. \end{aligned}$$

It also holds that $\mathsf{Cov}_{\boldsymbol{\theta}}[T(X), \dot{\ell}_{\boldsymbol{\theta}}(X)] = \dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta})$ as seen from

$$\begin{aligned} \mathsf{Cov}_{\boldsymbol{\theta}}\big[T_i(X), \dot{\ell}_{\boldsymbol{\theta},j}(X)\big] &\overset{(5.2)}{=} \mathsf{E}_{\boldsymbol{\theta}}\big[T_i(X)\dot{\ell}_{\boldsymbol{\theta},j}(X)\big] \\ &= \int T_i(x)\Big(\frac{\partial}{\partial \boldsymbol{\theta}_j} \log p_{\boldsymbol{\theta}}(x)\Big) p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\nu(x) \\ &= \int \frac{\partial}{\partial \boldsymbol{\theta}_j} T_i(x) p_{\boldsymbol{\theta}}(x)\, \mathrm{d}\nu(x) \\ &\overset{(M5)}{=} \frac{\partial}{\partial \boldsymbol{\theta}_j} \mathsf{E}_{\boldsymbol{\theta}}\big[T_i(X)\big] = \big(\dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta})\big)_{ij}. \end{aligned}$$

[1] This is essentially the complement of the support of P_θ , where the support is the smallest closed set C with $\mathsf{P}_\theta(C) = 1$.

Therefore,

$$\mathsf{Var}_{\boldsymbol{\theta}}\left[\begin{pmatrix} T(X) \\ \dot{\ell}_{\boldsymbol{\theta}}(X) \end{pmatrix}\right] = \begin{pmatrix} \mathsf{Var}_{\boldsymbol{\theta}}[T(X)] & \dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta}) \\ \left(\dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta})\right)^\top & \mathrm{I}(\boldsymbol{\theta}) \end{pmatrix}.$$

Since this covariance matrix is positive semidefinite and $I(\theta)$ is invertible, it follows that the Schur complement (see Definition 5.2) of the Fisher information is equal to

$$\mathsf{Var}_{\boldsymbol{\theta}}[T(X)] - (\dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta}))I^{-1}(\boldsymbol{\theta})(\dot{\gamma}(\boldsymbol{\theta}) + \dot{b}(\boldsymbol{\theta}))^\top \succeq 0,$$

and positive semidefinite, which completes the proof. □

Schur Complement

Consider a partitioned matrix

$$A = \begin{array}{cc} & \begin{array}{cc} k & m \end{array} \\ \begin{array}{c} k \\ m \end{array} & \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \end{array}$$

in which the $m \times m$ block A_{22} is invertible.

Definition 5.2 (Schur complement)**.** The Schur complement of A_{22} in A is

$$A_{11.2} := A_{11} - A_{12}A_{22}^{-1}A_{21}.$$

Lemma 5.1. *If $A \succeq 0$ and in particular A symmetric, then $A_{11.2} \succeq 0$. And if $A \succ 0$, then $A_{11.2} \succ 0$.*[2]

Proof. By symmetry of A,

$$\begin{pmatrix} \mathbf{I} & -A_{12}A_{22}^{-1} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}^\mathsf{T} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -A_{22}^{-1}A_{21} & \mathbf{I} \end{pmatrix},$$

and, thus,

$$\begin{pmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & -A_{12}A_{22}^{-1} \\ 0 & I \end{pmatrix}^\top = \begin{pmatrix} A_{11.2} & 0 \\ 0 & A_{22} \end{pmatrix}. \tag{5.3}$$

For any $\boldsymbol{y} \in \mathbb{R}^k$, let $\boldsymbol{z} = \begin{pmatrix} \boldsymbol{I} & -A_{12}A_{22}^{-1} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix}^\top \begin{pmatrix} \boldsymbol{y} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^{k+m}$. Then

[2] Notation: $\succeq 0$ means positive semidefinite, and $\succ 0$ means positive definite.

$$\begin{aligned}\mathbf{0} &\leqslant \boldsymbol{z}^\top A\boldsymbol{z}\\ &= \begin{pmatrix}\boldsymbol{y}^\top & \mathbf{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & -A_{12}A_{22}^{-1}\\ \mathbf{0} & \boldsymbol{I}\end{pmatrix}\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & -A_{12}A_{22}^{-1}\\ \mathbf{0} & \boldsymbol{I}\end{pmatrix}^\top\begin{pmatrix}\boldsymbol{y}\\ \mathbf{0}\end{pmatrix}\\ &= \begin{pmatrix}\boldsymbol{y}^\top & \mathbf{0}\end{pmatrix}\begin{pmatrix}A_{11.2} & \mathbf{0}\\ \mathbf{0} & A_{22}\end{pmatrix}\begin{pmatrix}\boldsymbol{y}\\ \mathbf{0}\end{pmatrix}\\ &= \boldsymbol{y}^\top A_{11.2}\boldsymbol{y},\end{aligned}$$

for all $y \in \mathbb{R}^k$, and thus $A_{11.2} \succeq 0$.

If $A \succ 0$, then $\boldsymbol{z}^\top A\boldsymbol{z} > 0$ for all $\boldsymbol{z} \neq 0$. Since $\boldsymbol{z} \neq 0$ implies $\boldsymbol{y} \neq 0$ and hence $\boldsymbol{y}^\top A_{11.2}\boldsymbol{y} > 0$ for all $\boldsymbol{y} \neq 0$, which shows $A_{11.2} \succ 0$. □

Inverting a Partitioned Matrix

Lemma 5.2. *Assuming all inverses exists:*

$$A^{-1} = \begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix}^{-1} = \begin{pmatrix}(A_{11.2})^{-1} & -A_{11.2}^{-1}A_{12}A_{22}^{-1}\\ -A_{22}^{-1}A_{21}A_{11.2}^{-1} & (A_{22.1})^{-1}\end{pmatrix},$$

where $A_{11,2}^{-1}A_{12}A_{22}^{-1} = A_{11}^{-1}A_{12}A_{22,1}^{-1}$.

Proof. Check $A \cdot A^{-1} = I$ or invert (5.3) and rearrange the pieces. □

5.3 More on Fisher Information

5.3.1 Fisher Information and Curvature of the Log–likelihood Function

Under (M1)–(M5),

$$\mathrm{I}(\boldsymbol{\theta}) := \mathsf{Var}_{\boldsymbol{\theta}}\big[\dot{\ell}_{\boldsymbol{\theta}}(X)\big] = \mathsf{E}_{\boldsymbol{\theta}}\big[\dot{\ell}_{\boldsymbol{\theta}}(X)\dot{\ell}_{\boldsymbol{\theta}}(X)^\top\big]$$

because $\mathsf{E}_{\boldsymbol{\theta}}[\dot{\ell}_{\boldsymbol{\theta}}(X)] = 0$.

If in addition it holds that **(M6)**

$$\int p_\theta(x)d\nu(x) \qquad \text{can be differentiated twice under the integral sign,}$$

then,

$$\mathrm{I}(\boldsymbol{\theta}) = -\mathsf{E}_{\boldsymbol{\theta}}\big[\ddot{\boldsymbol{\ell}}_{\boldsymbol{\theta}}(\boldsymbol{X})\big]$$

where $\ddot{\ell}_{\boldsymbol{\theta}}(x) = \left(\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\log p_{\boldsymbol{\theta}}(x)\right)_{ij}$ is the Hessian of the log–likelihood function.

NB: The Fisher information captures expected curvature of the log–likelihood function!

Proof. We have that

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\log p_{\boldsymbol{\theta}}(x) &= \frac{1}{p_{\boldsymbol{\theta}}(x)}\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}p_{\boldsymbol{\theta}}(x) - \frac{1}{p_{\boldsymbol{\theta}}(x)^2}\left(\frac{\partial}{\partial\boldsymbol{\theta}_i}p_{\boldsymbol{\theta}}(x)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}_j}p_{\boldsymbol{\theta}}(x)\right)\\ &= \frac{1}{p_{\boldsymbol{\theta}}(x)}\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}p_{\boldsymbol{\theta}}(x) - \left(\frac{\partial}{\partial\boldsymbol{\theta}_i}\log p_{\boldsymbol{\theta}}(x)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}_j}\log p_{\boldsymbol{\theta}}(x)\right).\end{aligned}$$

Take expectations,

$$\begin{aligned}&\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}\log p_{\boldsymbol{\theta}}(X)\right]\\ &= \int \frac{\partial^2}{\partial\boldsymbol{\theta}_i\partial\boldsymbol{\theta}_j}p_{\boldsymbol{\theta}}(x)\mathrm{d}\nu(x) - \mathsf{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}_i}\log p_{\boldsymbol{\theta}}(X)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}_j}\log p_{\boldsymbol{\theta}}(X)\right)\right]\\ &= 0 - \left(\mathsf{E}_{\boldsymbol{\theta}}\left[\dot{\ell}_{\boldsymbol{\theta}}(X)\dot{\ell}_{\boldsymbol{\theta}}(X)^\top\right]\right)_{ij}.\end{aligned}$$

□

5.3.2 Fisher Information and Random Samples

Proposition 5.1. *Suppose $\boldsymbol{X} = (X_1, ..., X_n)$ with $X_1, ..., X_n \overset{i.i.d.}{\sim} P_\theta, \theta \in \Theta$. Let $I_n(\theta) = \mathsf{Var}_\theta[\dot{\ell}_\theta(X)]$ be the Fisher information. Then under (M1)–(M5):*

$$I_n(\boldsymbol{\theta}) = nI_1(\boldsymbol{\theta}).$$

Proof. By the assumed independence, it holds that

$$\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{X}) = \nabla_{\boldsymbol{\theta}}\sum_{i=1}^n \log p_{\boldsymbol{\theta}}(X_i) = \sum_{i=1}^n \dot{\ell}_{\boldsymbol{\theta}}(X_i).$$

Taking variances gives

$$\mathrm{I}_n(\boldsymbol{\theta}) = \mathsf{Var}_{\boldsymbol{\theta}}\left[\dot{\ell}_{\boldsymbol{\theta}}(\boldsymbol{X})\right] = \sum_{i=1}^n \mathsf{Var}_{\boldsymbol{\theta}}\left[\dot{\ell}_{\boldsymbol{\theta}}(X_i)\right] = n\mathrm{I}_1(\boldsymbol{\theta}).$$

□

5.4 Examples

Example 5.1 (Gaussian model). Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2), \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$. Some calculus gives

$$I_1(\mu, \sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \text{ and } I_1^{-1}(\mu, \sigma^2) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

CR bound for unbiased estimation of μ :

$$\mathsf{Var}_\theta[T] \geq \frac{\sigma^2}{n},$$

achieved by $T(\boldsymbol{X}) = \overline{X}_n$.

CR bound for unbiased estimation of σ^2:

$$\mathsf{Var}_{\boldsymbol{\theta}}[T] \geq \frac{2\sigma^4}{n},$$

not achieved because the UMVUE is the sample variance s^2 with

$$\mathsf{Var}_{\boldsymbol{\theta}}[s^2] = \frac{2\sigma^4}{n-1}.$$

But note that bound is achieved asymptotically for $n \to \infty$.

Example 5.2 (Reparametrization of Gaussian model)**.** In Example 5.1, $I_1(\mu, \sigma^2)$ is diagonal, which need not stay true if we reparametrize.

Let

$$\nu = \mathsf{E}_{\mu,\sigma^2}[X_1^2].$$

Then

$$\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mu \\ \nu - \mu^2 \end{pmatrix}.$$

and [check!]

$$\mathrm{I}_1(\mu,\nu) = \begin{pmatrix} \frac{\mu^2+\nu}{(\mu^2-\nu)^2} & \frac{-\mu}{(\mu^2-\nu)^2} \\ \frac{-\mu}{(\mu^2-\nu)^2} & \frac{1}{2(\mu^2-\nu)^2} \end{pmatrix},$$

$$\mathrm{I}_1^{-1}(\mu,\nu) = \begin{pmatrix} \nu-\mu^2 & 2\mu(\nu-\mu^2) \\ 2\mu(\nu-\mu^2) & 2(\nu^2-\mu^4) \end{pmatrix}.$$

Since $\nu - \mu^2 = \sigma^2$, the sample mean $\overline{X}_n$ still achieves the CR bound obtained from this parametrization (as it should...).

Moreover, $\hat{\nu} = \frac{1}{n}\sum_{i=1}^n X_i^2$, with $\mathsf{Var}_{(\mu,\nu)}[\hat{\nu}] = \frac{1}{n}\mathsf{Var}_{(\mu,\nu)}[X_1^2]$, can be shown to achieve the CR bound for unbiased estimation of ν .

Example 5.3 (Exponential family in natural parametrization)**.** Consider an observation X that follows a distribution P_η from an exponential family in canonical form with density

$$p_\eta(x) = \exp\{\langle \eta, T(x)\rangle - A(\eta)\}h(x), \quad x \in \mathcal{X}.$$

Then $\dot{\ell}_\eta(x) = T(x) - \nabla_\eta A(\eta)$ and

$$\begin{aligned} \mathrm{I}(\eta) &= \mathsf{Var}_\eta[\dot{\ell}_\eta(X)] = \mathsf{Var}_\eta[T(X)] \\ &= D_2 A(\eta) = -\mathsf{E}_\eta[\ddot{\ell}_\eta(X)]. \end{aligned}$$

Consider estimating the vector of mean parameters $\gamma(\eta) = \mathsf{E}_\eta[T(X)]$. Then $\dot{\gamma}(\eta) = D_2 A(\eta)$ and the CR bound takes the form:

$$\begin{aligned} \dot{\gamma}(\eta)\boldsymbol{I}^{-1}(\eta)\dot{\gamma}(\eta)^\top &= D_2A(\eta)\left(D_2A(\eta)\right)^{-1} D_2A(\eta) \\ &= D_2A(\eta) \\ &= \mathsf{Var}_\eta[T(X)]. \end{aligned}$$

We observe that T achieves CR bound for unbiased estimation of $\gamma(\eta) = \mathsf{E}_\eta[T(X)]$. This provides a second proof that T is UMVUE (besides using the theorem of Lehmann-Scheffé).

Example 5.4 (Gamma model)**.** Let $X_1, \ldots, X_n$ be i.i.d. Gamma (α, β), $\boldsymbol{\theta} = (\alpha, \beta) \in (0, \infty)^2$. The Lebesgue density is

$$p_{\boldsymbol{\theta}}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \mathbf{1}_{(0,\infty)}(x)$$

$$= \exp\left\{\left\langle \begin{pmatrix} \alpha - 1 \\ -\beta \end{pmatrix}, \begin{pmatrix} \log x \\ x \end{pmatrix} \right\rangle\right.$$

$$\left. - \left(\log\Gamma(\alpha) - \alpha\log\beta\right)\right\} \mathbf{1}_{(0,\infty)}(x).$$

The Fisher information can be shown to be

$$I_1(\alpha, \beta) = \begin{pmatrix} \psi^{(1)}(\alpha) & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix},$$

with inverse

$$I_1^{-1}(\alpha, \beta) = \frac{1}{\alpha\psi^{(1)}(\alpha) - 1} \begin{pmatrix} \alpha & \beta \\ \beta & \beta^2\psi^{(1)}(\alpha) \end{pmatrix},$$

where $\psi^{(1)}(\alpha) = \frac{d^2}{d\alpha^2}\log\Gamma(\alpha)$.

Then

$$I_n(\alpha, \beta) = nI_1(\alpha, \beta) \qquad \text{and} \qquad I_n(\alpha, \beta)^{-1} = \frac{1}{n} I_1(\alpha, \beta)^{-1}$$

.

Remark.

- Cramér–Rao bound need not be achieved by UMVUE.
- Bound is attained for mean parameters of exponential families.
- Under a differentiability assumption, achievement of Cramér–Rao bound requires an exponential family (and consideration of their mean parameters).
- We will later see that MLE achieves Cramér-Rao bound asymptotically.
- Differentiability requirements can be weakened using the concept of differentiability in quadratic mean; see, e.g., van der Vaart (1998, sec. 7.2) for further reading if you are interested.

5.5 Nuisance Parameters

In many statistical problems, the parameter vector $\boldsymbol{\theta}$ contains components of primary interest, alongside other components that are unknown but of secondary interest. These secondary parameters are often required to fully specify the model but are not the direct target of the estimation. We refer to these as **nuisance parameters**.

Example 5.5 (Speed of Light)**.** Consider measuring the speed of light (a physical constant) using a device with unknown accuracy. We might model the measurements using a normal distribution $N(\mu, \sigma^2)$.

- The mean μ represents the speed of light. This is the **parameter of interest**.
- The variance σ^2 represents the measurement error spread. This is unknown, since we did not collect data specifically to learn about the device's accuracy. Thus, σ^2 is a **nuisance parameter**.

A fundamental question arises:

What is the "price" we pay in terms of estimation accuracy for not knowing the nuisance parameter?

We can answer this using the Cramér-Rao bound.

5.5.1 Partitioned Information

Suppose the model is given by

$$\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \Theta_1 \times \Theta_2\},$$

where $\boldsymbol{\theta}_1$ is the parameter of interest and $\boldsymbol{\theta}_2$ is the nuisance parameter. The Fisher information matrix can be naturally partitioned into four blocks:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{I}_{11}(\boldsymbol{\theta}) & \boldsymbol{I}_{12}(\boldsymbol{\theta}) \\ \boldsymbol{I}_{21}(\boldsymbol{\theta}) & \boldsymbol{I}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

Here, $\boldsymbol{I}_{11}$ corresponds to partial derivatives with respect to $\boldsymbol{\theta}_1$, $\boldsymbol{I}_{22}$ to $\boldsymbol{\theta}_2$, and the off-diagonal blocks represent cross-covariances.

Consider the estimation of the target parameter

$$\gamma(\boldsymbol{\theta}) = \boldsymbol{\theta}_1 = \begin{pmatrix} I & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix}.$$

The Jacobian of this mapping is the projection matrix:

$$\dot{\gamma}(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \end{pmatrix},$$

where $\boldsymbol{I}$ is the identity matrix corresponding to the dimension of $\boldsymbol{\theta}_1$.

The Cramér-Rao (CR) bound at $\boldsymbol{\theta}_0$ for an unbiased estimator of $\boldsymbol{\theta}_1$ is given by the top-left block of the inverse information matrix:

$$\begin{aligned} \text{CR Bound} &= \dot{\gamma}(\boldsymbol{\theta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\dot{\gamma}(\boldsymbol{\theta}_0)^\top \\ &= \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \end{pmatrix} \boldsymbol{I}^{-1}(\boldsymbol{\theta}_0) \begin{pmatrix} \boldsymbol{I} \\ \mathbf{0} \end{pmatrix} \\ &= \left(\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\right)_{11}. \end{aligned}$$

Using the formula for the inverse of a block matrix, this quantity is the inverse of the **Schur complement**:

$$\left(\boldsymbol{I}^{-1}(\boldsymbol{\theta}_0)\right)_{11} = \left(\boldsymbol{I}_{11}(\boldsymbol{\theta}_0) - \boldsymbol{I}_{12}(\boldsymbol{\theta}_0)\boldsymbol{I}_{22}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}_{21}(\boldsymbol{\theta}_0)\right)^{-1}.$$

We define the **effective information** for $\boldsymbol{\theta}_1$ in the presence of unknown $\boldsymbol{\theta}_2$ as:

$$\boldsymbol{I}_{11.2}(\boldsymbol{\theta}_0) := \boldsymbol{I}_{11}(\boldsymbol{\theta}_0) - \boldsymbol{I}_{12}(\boldsymbol{\theta}_0)\boldsymbol{I}_{22}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{I}_{21}(\boldsymbol{\theta}_0).$$

Thus, the CR bound is $\boldsymbol{I}_{11.2}(\boldsymbol{\theta}_0)^{-1}$.

5.5.2 Comparison: Known vs. Unknown Nuisance Parameters

To understand the cost of ignorance, consider a submodel where the nuisance parameter is known to be $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$:

$$\mathcal{P}(\boldsymbol{\theta}_{20}) = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}), \boldsymbol{\theta}_1 \in \Theta_1\}.$$

In this scenario, estimating $\boldsymbol{\theta}_1$ relies only on the partial derivatives with respect to $\boldsymbol{\theta}_1$. The Fisher information for this submodel is simply the top-left block $\boldsymbol{I}_{11}(\boldsymbol{\theta}_0)$. Consequently, the CR bound becomes $\boldsymbol{I}_{11}(\boldsymbol{\theta}_0)^{-1}$.

We can now compare the information available in both scenarios:

- **With $\boldsymbol{\theta}_2$ known:** Information is $\boldsymbol{I}_{11}(\boldsymbol{\theta}_0)$.

- **With $\boldsymbol{\theta}_2$ unknown:** Information is $\boldsymbol{I}_{11.2}(\boldsymbol{\theta}_0) = \boldsymbol{I}_{11}(\boldsymbol{\theta}_0) - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21}$.

  Since $\boldsymbol{I}(\boldsymbol{\theta})$ is a covariance matrix, it is symmetric and positive semi–definite. The term being subtracted, $\boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21}$, is of the form $ABA^\top$ with B positive definite, meaning the subtraction term is positive semi–definite. Therefore:

  $$\boldsymbol{I}_{11}(\boldsymbol{\theta}_0) \succeq \boldsymbol{I}_{11.2}(\boldsymbol{\theta}_0).$$

  Remark (The Information Inequality). In general, not knowing the nuisance parameter $\boldsymbol{\theta}_2$ strictly reduces the information available for estimating $\boldsymbol{\theta}_1$ (and thus increases the variance bound), unless:

  $$\boldsymbol{I}_{12}(\boldsymbol{\theta}_0) = \mathbf{0}.$$

  If the Fisher information matrix is block diagonal, the parameters are orthogonal. in this "lucky" case, there is no price to pay for not knowing the nuisance parameter. For example, when estimating the mean μ of a normal distribution, the sample mean $\overline{X}$ is the optimal estimator regardless of whether the variance σ^2 is known or unknown.

5.6 Reparametrization

Consider a statistical model $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\theta}\}$ with Fisher information $I(\boldsymbol{\theta}_0)$ at $\boldsymbol{\theta}_0 \in \boldsymbol{\theta}$.

In practice, different software or researchers may parametrize the same distribution differently (e.g., using variance versus precision, or rate versus scale). Suppose we reparametrize the model using a **diffeomorphism**[3] that maps each new parameter $\boldsymbol{\lambda} \in \Lambda$ to a point $\boldsymbol{\theta}(\boldsymbol{\lambda}) \in \boldsymbol{\theta}$.

Proposition 5.2. *The Fisher information of the reparametrized model $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}(\boldsymbol{\lambda})} : \boldsymbol{\lambda} \in \boldsymbol{\lambda}\}$ at the point $\boldsymbol{\lambda}_0$ with $\boldsymbol{\theta}(\boldsymbol{\lambda}_0) = \boldsymbol{\theta}_0$ is equal to*

$$\boldsymbol{I}(\boldsymbol{\lambda}_0) = D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^\top\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} \cdot \boldsymbol{I}(\boldsymbol{\theta}_0) \cdot D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0},$$

where $D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})$ denotes the Jacobian matrix of the map $\boldsymbol{\lambda} \mapsto \boldsymbol{\theta}(\boldsymbol{\lambda})$.

Proof. By the chain rule, the reparametrized model has the score function:

$$\dot{\ell}_{\boldsymbol{\lambda}}(X) \equiv \nabla_{\boldsymbol{\lambda}} \log p_{\boldsymbol{\theta}(\boldsymbol{\lambda})}(X) = \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(X)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\boldsymbol{\lambda})} \cdot D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda}) = \dot{\ell}_{\boldsymbol{\theta}(\boldsymbol{\lambda})}(X) \cdot D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda}).$$

Hence, the Fisher information transforms as a quadratic form:

$$\boldsymbol{I}(\boldsymbol{\lambda}) \equiv \mathsf{Var}_{\boldsymbol{\lambda}}[\dot{\ell}_{\boldsymbol{\lambda}}(X)] = D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^\top \boldsymbol{I}(\boldsymbol{\theta}(\boldsymbol{\lambda})) D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda}).$$

□

[3] A differentiable map with a differentiable inverse.

Remark. The transformation of the Fisher information follows the logic of the chain rule. If one researcher parametrizes in terms of $\boldsymbol{\theta}$ and another in terms of $\boldsymbol{\lambda}$, the information matrices are related simply by picking up the Jacobian of the change of parametrization. This changes the numerical values of the matrix entries (e.g., information about the standard deviation looks different than information about the variance), but it respects the geometry of the problem.

Example 5.6 (Gamma Distribution Ambiguity)**.** Why does this matter? Consider the Gamma distribution. Some software packages parametrize it using a *rate* parameter β (density $\propto e^{-\beta x}$), while others use a *scale* parameter $1/\beta$ (density $\propto e^{-x/\beta}$). This is a diffeomorphic change of variables. While the definitions differ, the chain rule ensures we can always convert the Fisher information from one coordinate system to the other.

Despite the information matrix changing forms, the fundamental limit on estimation accuracy remains constant.

Proposition 5.3. *The Cramér–Rao bound is invariant under reparametrization.*

Proof. In the original model $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\theta}\}$, the CR bound for estimating $\boldsymbol{\theta}$ at the distribution given by $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is $\boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}$.

In the reparametrized model $\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}(\boldsymbol{\lambda})} : \boldsymbol{\lambda} \in \boldsymbol{\lambda}\}$, the considered distribution is $\mathsf{P}_{\boldsymbol{\theta}_0} = \mathsf{P}_{\boldsymbol{\theta}(\boldsymbol{\lambda}_0)}$ for $\boldsymbol{\lambda}_0 := \boldsymbol{\theta}^{-1}(\boldsymbol{\theta}_0)$. In the new parametrization, we estimate the function $\gamma(\boldsymbol{\lambda}) = \boldsymbol{\theta}(\boldsymbol{\lambda})$.

By Proposition 5.2, the CR bound for estimating this function is:

$$\begin{aligned}
&D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda}) \cdot \boldsymbol{I}(\boldsymbol{\lambda})^{-1} \cdot D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^\top \Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} \\
&= D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} \cdot \left(D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^\top\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} \cdot \boldsymbol{I}(\boldsymbol{\theta}_0) \cdot D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} \right)^{-1} D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^\top\Big|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_0} \\
&= D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda}) \cdot \left(D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^{-1} \cdot \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1} \cdot (D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^\top)^{-1} \right) \cdot D_{\boldsymbol{\lambda}}\boldsymbol{\theta}(\boldsymbol{\lambda})^\top \\
&= \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1}.
\end{aligned}$$

□

Note. Intuitively, this invariance *must* hold. We are estimating the same characteristic of the same data-generating distribution $\mathsf{P}_{\boldsymbol{\theta}_0}$. Whether we label the distribution using $\boldsymbol{\theta}$ or $\boldsymbol{\lambda}$ is merely a naming convention.

For example, if the true data follows a Gaussian distribution with mean 5 and variance 7, the minimum variance for estimating the mean is a fixed number. Renaming the parameters does not change the difficulty of the statistical problem. The chain rule terms in the derivative of the target parameter $\gamma(\boldsymbol{\lambda})$ exactly cancel out the chain rule terms from the Fisher information transformation.

6. Equivariant Estimation

Sometimes, there is no optimal unbiased estimator. In such cases, instead of insisting on unbiasedness and seeking the estimator with the minimum Mean Squared Error (MSE), we look for alternative criteria. Specifically, we look for **invariance properties**.

In this chapter, we treat optimal estimation when restricting ourselves to **equivariant estimators**. In brief, an equivariant estimator has the (desirable) property that an estimate computed from transformed data coincides with an appropriate transformation of the estimate computed from the original data.

Note. Unbiasedness is not always a "super fruitful" recipe. There are settings that are more naturally dealt with by other considerations. Here, natural invariances (equivariance) replace unbiasedness as the primary guide.

We will explore this concept in the context of a very specific invariant property: **Location Models**. In this setting, the natural invariances we respect are **shifts** in location.

For reading on location models, see Lehmann and Casella (1998, sec. 3.1). For additional reading on more general settings (beyond simple shifts), consider Lehmann and Casella (1998, secs. 3.2–3.4), and Lehmann and Romano (2005, sec. 6.2).

6.1 Location Models

In this section, we focus on estimating the location (the center) of a distribution while remaining invariant with respect to shifts.

Consider an observation vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ taking values in $\mathbb{R}^n$.

Definition 6.1 (Location Model)**.** The location model is defined by the structural equation:

$$X_i = \theta + \epsilon_i, \quad i = 1, \ldots, n, \tag{6.1}$$

where:

- $\theta \in \Theta = \mathbb{R}$ is the unknown parameter (the location).
- $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. error terms with a **known** distribution P_0.

We write P_θ for the distribution of X_i (and $\mathcal{P}_\theta$ for the joint distribution of $\boldsymbol{X}$).

Note (Intuition: Signal plus Noise)**.** This setup is analogous to the "Speed of Light" example.

- θ represents the constant physical quantity we wish to locate (the signal).
- ϵ_i represents the chance measurement error (the noise) that typically fluctuates around zero.

Crucially, in a location model, the *only* unknown is the location θ. The shape of the distribution (the density of the errors) is known. Mentally, you can visualize a fixed density curve sliding along the real line; we know exactly what the curve looks like, we just do not know where its peak (or center) is positioned.

Example 6.1. We list three standard examples of location models. Note that in all cases, the shape of the distribution is fixed, and only the center shifts.

- **Normal:** Taking $\mathsf{P}_0 = \mathcal{N}(0,1)$, the location model is $\{\mathcal{N}(\theta,1) : \theta \in \mathbb{R}\}$.
- **Uniform:** Fix an interval length $a > 0$ and take $\mathsf{P}_0 = \text{Uniform}(-a/2, a/2)$. Then the location model is
$$\{\text{Uniform}(\theta - a/2, \theta + a/2) : \theta \in \mathbb{R}\}.$$
- **Cauchy:** Taking $\mathsf{P}_0 = \text{Cauchy}$ gives the model in which P_θ has density
$$p_\theta(x) = p_0(x-\theta) = \frac{1}{\pi\left[1 + (x-\theta)^2\right]}, \qquad x \in \mathbb{R}.$$

Note (Symmetry and Positive Data)**.** You may notice that the examples above (Normal, Uniform, Cauchy) are all symmetric around θ.

One might ask: *What if the quantity we are measuring must be positive?* For example, if we are measuring the weight of an animal, an additive error model like $X = \theta + \epsilon$ is usually inappropriate. An error of 1 kg means something very different when weighing an elephant versus weighing a mouse. In such cases, errors tend to be multiplicative rather than additive.

However, location models remain relevant because we can simply **take the logarithm** of the data. If the physics suggests a multiplicative structure, working on the log–scale converts it to an additive structure:

$$\log(\text{Weight}) = \log(\text{True Value}) + \log(\text{Error}).$$

This returns us to the location model framework, where the Central Limit Theorem often justifies assuming normality on the log–scale (Log–Normal models).

6.1.1 Changing units

Additive change of units
Suppose we change units for our observations in an additive way (e.g., Celsius to Kelvin), giving transformed observations:

$$X_i' = X_i + c \quad \text{for } c \in \mathbb{R}.$$

Example 6.2 (Temperature: Kelvin vs. Celsius)**.** Consider analyzing temperature data. Suppose a physicist friend sends you a data–set where the units are in Kelvin. You compute an estimate of the temperature (e.g., the sample mean) based on these Kelvin measurements.

Now, imagine a second analyst receives the same data but prefers the Celsius (centigrade) scale.

1. They first transform the raw data by subtracting the constant 273.15 (an additive shift).
2. Then, they perform the statistical estimation (e.g., taking the average) on the transformed data.

Intuitively, the results should be consistent. The estimate derived by the second analyst should simply be the first analyst's estimate minus 273.15.

Note (Commutativity)**.** This concept is the essence of **equivariance**. We require the following diagram to commute:

- **Path A:** Estimate on raw data → Transform the result.
- **Path B:** Transform the raw data → Estimate on transformed data.

If these two paths yield the same result, the estimator respects the natural invariances of the problem. While the sample mean satisfies this property for additive shifts, not all estimators do. In this chapter, we restrict our search to estimators that satisfy this logical consistency.

Invariance of the Location Model
The location model is invariant under additive transformations. This means that if we shift the data, we remain within the same family of distributions; we simply move to a different parameter value.

Mathematically, for any shift $c \in \mathbb{R}$, the transformed observations $X_1', \dots, X_n'$ are still i.i.d. with marginal distribution $\mathsf{P}_{\theta'}$ for some $\theta' \in \mathbb{R}$. Specifically:

$$X_i' = X_i + c = (\theta + c) + \epsilon_i \sim \mathsf{P}_{\theta'},$$

where the new parameter is $\theta' = \theta + c$.

Note. There is an induced action on the parameter space. By transforming the data ($X \to X + c$), we induce a corresponding transformation on the parameter ($\theta \to \theta + c$). The set of all distributions $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \mathbb{R}\}$ remains unchanged as a whole; the distributions are just permuted.

6.2 Location Equivariance

We now formalize the requirement that our estimator should respect the symmetries of the location model.

Notation. *For a vector $\boldsymbol{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ and a scalar $c \in \mathbb{R}$, we write $\boldsymbol{x} + c$ to denote the vector where c is added to every component:*

$$\boldsymbol{x} + c := (x_1 + c, \dots, x_n + c).$$

Definition 6.2 (Location Equivariance)**.** An estimator $T : \mathbb{R}^n \to \mathbb{R}$ is called **location equivariant** if for all $\boldsymbol{x} \in \mathbb{R}^n$ and all $c \in \mathbb{R}$:

$$T(\boldsymbol{x} + c) = T(\boldsymbol{x}) + c.$$

Note (The Commutative Property)**.** Intuitively, this definition implies that the diagram commutes. It should not matter whether you:

1. Transform the data first ($\boldsymbol{x} + c$) and then estimate; or
2. Estimate first ($T(\boldsymbol{x})$) and then transform the result ($+c$).

If this holds, the estimator "respects" the shift.

Example 6.3. Many common statistics satisfy this property.

- **Sample Mean:** $T(\boldsymbol{X}) = \overline{X}_n$.
- **Sample Median:** $T(\boldsymbol{X}) = \text{med}(\boldsymbol{X})$.
- **The Maximum:** $T(\boldsymbol{X}) = X_{(n)} = \max(X_1, \dots, X_n)$.
- **Single Observation:** $T(\boldsymbol{X}) = X_n$ (simply taking the last observation).

Note (Equivariance vs. Invariance)**.** It is crucial to distinguish between *equivariance* and *invariance*.

- **Equivariance:** The output transforms suitably with the input (e.g., input $+c \implies$ output $+c$).
- **Invariance:** The output does *not* change when the input transforms (e.g., input $+c \implies$ output unchanged).

Example 6.4 (Estimating Variance)**.** Consider the sample variance S^2. If we shift the data by c, the spread of the data does not change. Therefore:

$$S^2(\boldsymbol{x} + c) = S^2(\boldsymbol{x}).$$

The sample variance is **location invariant**, not equivariant. This makes sense: we do not want our estimate of the spread to change just because we switched from Kelvin to Celsius (an additive shift). However, for estimating the *location* θ, we specifically require equivariance.

Note (Scale Transformations)**.** One could also consider scaling transformations (e.g., Fahrenheit to Celsius involves both a shift and a scale factor). While we could define *scale equivariance* (where $T(a\boldsymbol{x}) = aT(\boldsymbol{x})$), in this chapter we restrict our attention strictly to location shifts.

6.3 Judging Estimators by Their Risk

Consider a model $\{P_\theta : \theta \in \Theta\}$ and a parameter $\gamma : \Theta \to \Gamma$.

Going beyond MSE, we may evaluate estimators of $\gamma(\theta)$ using the notions of loss and risk.

Motivation: Why move beyond MSE?
In previous chapters, we focused on unbiased estimators and sought the one with the minimum variance (MSE). Now, we are asking a similar question for **equivariant estimators**: is there a "best" one within this class?

To answer this, we must generalize how we measure quality. MSE is just one specific way to penalize errors (squaring them). However, depending on the scientific or business context, the "cost" of an error might be different. For example, is an error of $+10$ equally as bad as -10? Is a small error negligible, or does it still incur a cost?

Definition 6.3 (Risk Function)**.** The **risk** of an estimator T of the parameter $\gamma(\theta)$ is the expected loss:

$$R(\theta, T) = \mathsf{E}_\theta\big[L(\theta, T(X))\big],$$

where $L : \Theta \times \Gamma \to [0, \infty)$ is a **loss function**.

The loss function $L(\theta, a)$ quantifies the "price" paid when the true state of nature is θ and the estimate is a. It must satisfy:

$$L(\theta, \gamma(\theta)) = 0 \quad \text{for all } \theta.$$

(i.e., if you estimate the target perfectly, you pay nothing).

Example 6.5.

- **Squared Error Loss:** This leads to the Mean Squared Error (MSE) risk.

  $$L(\theta, a) = (a - \gamma(\theta))^2 \implies R(\theta, T) = \mathsf{E}_\theta[(T(X) - \gamma(\theta))^2].$$

  Comment: We often use this because it is mathematically convenient (differentiable, easy to work with), not necessarily because it perfectly reflects the real–world cost of errors.

- **Absolute Error Loss:** This leads to the Mean Absolute Error risk.

  $$L(\theta, a) = |a - \gamma(\theta)| \implies R(\theta, T) = \mathsf{E}_\theta\big[|T(X) - \gamma(\theta)|\big].$$

  Comment: This is less sensitive to outliers than the squared error but harder to differentiate.

Who chooses the Loss Function?
Ideally, the loss function should come from the domain expert (the physicist, the business analyst), not the statistician. They know the actual cost of making a mistake. For example, in a Kaggle competition, you must optimize for the specific scoring metric provided (the loss function).

However, in practice, clients often do not know their specific loss function, so statisticians frequently default to MSE because "who doesn't love a square?" It is convenient and generally reasonable. But be aware: the choice of loss function dictates which estimator is optimal.

6.3.1 Location Invariant Loss Functions

Since we are enforcing equivariance on our estimators (shifting the input shifts the output), we should also require our loss function to respect this symmetry.

Motivation: Consider the temperature example again. Suppose your estimate a is in Kelvin, and the true value θ is also in Kelvin. You incur some loss $L(\theta, a)$.

If your colleague converts everything to Centigrade (shifting both the estimate and the true value by c), the "price" of the error should not change. Being off by 1 degree costs the same whether that degree is Kelvin or Centigrade. Therefore, we require:

$$L(\theta + c, a + c) = L(\theta, a).$$

Definition 6.4 (Location Invariant Loss)**.** A loss function $L : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ is **location invariant** if

$$L(\theta + c, a + c) = L(\theta, a)$$

for all $\theta, a, c \in \mathbb{R}$.

This condition holds precisely when the loss depends only on the *difference* between the estimate and the parameter:

$$L(\theta, a) = \rho(\theta - a),$$

where ρ is a function such that $\rho(0) = 0$.

Example 6.6.

- Squared Error: $L(\theta, a) = (\theta - a)^2$. Here $\rho(u) = u^2$.
- Absolute Error: $L(\theta, a) = |\theta - a|$. Here $\rho(u) = |u|$.

Risk of Equivariant Estimators A remarkable property emerges when we combine an equivariant estimator with an invariant loss function: the risk becomes constant. It no longer depends on the true parameter θ.

Lemma 6.1 (Constant Risk)**.** *Let T be an equivariant estimator of θ in the location model. If the loss L is location invariant, then the risk $R(\theta, T)$ does not depend on θ. That is,*

$$R(\theta, T) = R(0, T) \quad \forall \theta \in \Theta = \mathbb{R}.$$

Proof. The proof follows by utilizing the invariance of the loss and the equivariance of the estimator to shift the problem to $\theta = 0$.

By definition, the risk is the expected loss under P_θ:

$$R(\theta, T) = \mathsf{E}_\theta\big[L(\theta, T(\boldsymbol{X}))\big].$$

Since the loss L is location invariant, we can shift both arguments by $-\theta$ (effectively setting the first argument to 0):

$$L(\theta, T(\boldsymbol{X})) = L(\theta - \theta, T(\boldsymbol{X}) - \theta) = L(0, T(\boldsymbol{X}) - \theta).$$

Since the estimator T is location equivariant, shifting the output by $-\theta$ is equivalent to shifting the input by $-\theta$:

$$T(\boldsymbol{X}) - \theta = T(\boldsymbol{X} - \theta).$$

Substituting this back into the expectation:

$$R(\theta, T) = \mathsf{E}_\theta\big[L\big(0, T(\boldsymbol{X} - \theta)\big)\big].$$

Finally, consider the distribution of the random variable $\boldsymbol{Y} = \boldsymbol{X} - \theta$. If $\boldsymbol{X} \sim \mathsf{P}_\theta$ (i.e., $X_i = \theta + \epsilon_i$), then $\boldsymbol{X} - \theta$ is simply the error term ϵ, which follows the distribution P_0. Therefore, taking the expectation of $f(\boldsymbol{X} - \theta)$ under P_θ is the same as taking the expectation of $f(\boldsymbol{X})$ under P_0:

$$\mathsf{E}_\theta\big[L\big(0, T(\boldsymbol{X} - \theta)\big)\big] = \mathsf{E}_0\big[L(0, T(\boldsymbol{X}))\big] = R(0, T).$$

□

6.3.2 Minimum Risk Equivariant (MRE) Estimators

In general statistical problems, comparing two estimators is difficult. Often, one estimator is better for certain parameter values (e.g., low temperatures) while another is better for others (e.g., high temperatures). Unless we specify a prior (Bayesian approach), we cannot say which is strictly "better."

However, for **equivariant estimators** with **invariant loss**, the risk is constant across all θ. This reduces the comparison to a single number: the risk at $\theta = 0$. Since real numbers are ordered, we can now meaningfully search for a global "best" estimator in this class.

Definition 6.5 (Minimum Risk Equivariant (MRE))**.** An equivariant estimator T is **Minimum Risk Equivariant (MRE)** for a location invariant loss function L if, for any other equivariant estimator T', it holds that:

$$R(\theta, T) \leq R(\theta, T') \quad \text{for all } \theta \in \Theta = \mathbb{R}.$$

Since the risk of an equivariant estimator is constant according to Lemma 6.1, the condition above is equivalent to checking the risk at a single point (conventionally $\theta = 0$):

$$R(0, T) \leq R(0, T').$$

Example 6.7 (Squared Error Loss)**.** For the squared error loss $L(\theta, a) = (\theta - a)^2$, minimizing the risk reduces to minimizing the second moment under P_0:

$$R(0, T) = \mathsf{E}_0\big[(T(X) - 0)^2\big] = \mathsf{E}_0\big[T(X)^2\big].$$

Optimization Strategy Finding the MRE estimator is a constrained optimization problem: minimize $R(0, T)$ subject to the constraint that T is equivariant.

To solve this, we will convert the constrained problem into an unconstrained one. The strategy is to find a way to **represent** or **parametrize** the set of *all* equivariant estimators. Once we have a general form, we can optimize over the parameters of that form.

6.3.3 Location Invariance and Maximal Invariants

Motivation: Group Actions and Equivalence To systematically represent equivariant estimators, we first look at invariants. Our data sets live in $\mathbb{R}^n$. The group action encoding our transformations is the addition of a real number c to the vector.

The action of $c \in \mathbb{R}$ on vectors $\boldsymbol{x} \in \mathbb{R}^n$ yields an equivalence relation:

$$\boldsymbol{x} \sim \boldsymbol{x}' \iff \exists c \in \mathbb{R} : \boldsymbol{x} = \boldsymbol{x}' + c.$$

Intuitively, two data sets are equivalent if one is simply a shifted version of the other (like two columns in a spreadsheet differing by a constant).

Definition 6.6 (Location Invariant Statistic)**.** A statistic $u : \mathbb{R}^n \to \mathcal{U}$ is (location) invariant if

$$\boldsymbol{x} \sim \boldsymbol{x}' \implies u(\boldsymbol{x}) = u(\boldsymbol{x}').$$

If the reverse implication also holds, i.e.,

$$u(\boldsymbol{x}) = u(\boldsymbol{x}') \iff \boldsymbol{x} \sim \boldsymbol{x}',$$

then u is **maximal (location) invariant**.

Intuition: Partitions An invariant statistic takes the same value for all equivalent inputs. A *maximal* invariant distinguishes between every distinct equivalence class. It provides a "fine partition" of the sample space: its values tell you exactly which orbit (equivalence class) the data belongs to.

Proposition 6.1. *If u is maximal invariant, then any other statistic u' will be invariant if and only if it is a function of u, i.e., $u'(\boldsymbol{x}) = v(u(\boldsymbol{x}))$ for some map v and all $\boldsymbol{x}$.*

Proof. First, if $u'(\boldsymbol{x}) = v(u(\boldsymbol{x}))$ for all $\boldsymbol{x} \in \mathbb{R}^n$, then for any equivalent $\boldsymbol{x}' \sim \boldsymbol{x}$:

$$u'(\boldsymbol{x}') = v(u(\boldsymbol{x}')) = v(u(\boldsymbol{x})) = u'(\boldsymbol{x}),$$

which shows that u' is invariant.

Conversely, assume that u' is invariant. If $u(\boldsymbol{x}) = u(\boldsymbol{x}')$, then by the definition of maximal invariance, $\boldsymbol{x} \sim \boldsymbol{x}'$ (i.e., $\boldsymbol{x} = \boldsymbol{x}' + c$). Since u' is invariant, this implies $u'(\boldsymbol{x}) = u'(\boldsymbol{x}')$. Thus, u' is constant on the sets where u is constant, so we can define v such that $u'(\boldsymbol{x}) = v(u(\boldsymbol{x}))$. □

Differences give a maximal invariant

Motivation: Constructing the Invariant
How do we find a maximal invariant in our location model? We need a statistic that eliminates the shift c but preserves all relative information. Differences between coordinates are the natural candidate.

Example 6.8. The statistic $Y : \mathbb{R}^n \to \mathbb{R}^n$ given by the differences with respect to the last coordinate:

$$Y(\boldsymbol{x}) = \boldsymbol{x} - x_n = (x_1 - x_n, \dots, x_{n-1} - x_n, 0)$$

is clearly invariant. If we shift $\boldsymbol{x}$ by c, both x_i and x_n increase by c, so their difference remains constant.

In fact, Y is maximal invariant because

$$Y(\boldsymbol{x}) = Y(\boldsymbol{x}') \implies \boldsymbol{x} - x_n = \boldsymbol{x}' - x'_n \implies \boldsymbol{x} = \boldsymbol{x}' + (x_n - x'_n).$$

Since $x_n - x'_n$ is just a scalar, $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ only by a constant shift.

We can generalize this result using any equivariant estimator.

Lemma 6.2. *For any equivariant estimator T, the difference $\boldsymbol{x} - T(\boldsymbol{x})$ is maximal invariant.*

Proof. First, we check invariance. Let $c \in \mathbb{R}$.

$$(\boldsymbol{x} + c) - T(\boldsymbol{x} + c) = (\boldsymbol{x} + c) - [T(\boldsymbol{x}) + c] = \boldsymbol{x} - T(\boldsymbol{x}).$$

Thus, the statistic is invariant.

To show it is maximal invariant, assume $\boldsymbol{x} - T(\boldsymbol{x}) = \boldsymbol{x}' - T(\boldsymbol{x}')$. Rearranging terms gives:

$$\boldsymbol{x} = \boldsymbol{x}' + \underbrace{T(\boldsymbol{x}) - T(\boldsymbol{x}')}_{c}.$$

Since $T(\boldsymbol{x})$ and $T(\boldsymbol{x}')$ are scalars, their difference c is a scalar. Thus, $\boldsymbol{x} = \boldsymbol{x}' + c$, which means $\boldsymbol{x} \sim \boldsymbol{x}'$. □

6.3.4 Representation of Location Equivariant Estimators

Motivation: The Affine Structure
We now have a powerful way to generate *all* possible equivariant estimators. The structure is analogous to solving a non–homogeneous linear equation (or a differential equation): the general solution is the sum of a **particular solution** (a base equivariant estimator T_0) and the general solution to the **homogeneous equation** (an invariant term).

Since the difference between any two equivariant estimators is invariant, we can represent any T by starting with a fixed T_0 and adding an arbitrary invariant function.

Lemma 6.3 (Representation of equivariant estimators)**.** *Let T_0 be a fixed equivariant estimator, and let u be a maximal invariant statistic. For every other equivariant estimator T, there exists a function v such that*

$$T(\boldsymbol{x}) = T_0(\boldsymbol{x}) + v(u(\boldsymbol{x})), \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{6.2}$$

Proof. The proof relies on the property that differences of equivariant estimators are invariant.

1. **The difference is invariant:** Let $D(\boldsymbol{x}) = T(\boldsymbol{x}) - T_0(\boldsymbol{x})$. We check if D is invariant under the shift $\boldsymbol{x} \to \boldsymbol{x} + c$:

   $$D(\boldsymbol{x} + c) = T(\boldsymbol{x} + c) - T_0(\boldsymbol{x} + c).$$

   Since both T and T_0 are equivariant (i.e., $T(\boldsymbol{x} + c) = T(\boldsymbol{x}) + c$), the shifts cancel out:

   $$D(\boldsymbol{x} + c) = (T(\boldsymbol{x}) + c) - (T_0(\boldsymbol{x}) + c) = T(\boldsymbol{x}) - T_0(\boldsymbol{x}) = D(\boldsymbol{x}).$$

   Thus, the difference $T - T_0$ is an invariant statistic.

2. **Representation via Maximal Invariant:** Since u is a **maximal invariant**, by Proposition 6.1, any invariant statistic can be written as a function of u. Therefore, there exists a function v such that:

   $$D(\boldsymbol{x}) = v(u(\boldsymbol{x})).$$

   Rearranging terms yields (6.2).

□

Example 6.9 (Using a specific base estimator)**.** Let's apply this to a concrete case.

1. **Base Estimator:** Choose the simple projection to the last coordinate, $T_0(\boldsymbol{x}) = x_n$.

2. **Maximal Invariant:** Use the difference vector $Y(\boldsymbol{x}) = \boldsymbol{x} - x_n$ (from Example 6.8).

Then, according to Lemma 6.3, *any* equivariant estimator T can be written as:

$$T(\boldsymbol{x}) = T_0(\boldsymbol{x}) + v(Y(\boldsymbol{x})) = x_n + v(\boldsymbol{x} - x_n).$$

Alternatively, we can derive this directly using equivariance:

$$T(\boldsymbol{x}) = T(\boldsymbol{x} - x_n + x_n) = T(Y(\boldsymbol{x}) + x_n).$$

By equivariance, $T(Y(\boldsymbol{x}) + x_n) = T(Y(\boldsymbol{x})) + x_n$. In this form, the function v is simply the estimator T itself applied to the invariant differences.

6.4 Construction of MRE Estimators

Motivation: Decomposing the Risk
We have established that any equivariant estimator can be written as $T(\boldsymbol{X}) = T_0(\boldsymbol{X}) + v(\boldsymbol{Y})$, where $\boldsymbol{Y}$ is a maximal invariant. Finding the best estimator is therefore equivalent to finding the best function v.

How do we find this optimal v?

We use the law of iterated expectations (the "Tower Rule"). The total risk is the expectation of the conditional risk given $\boldsymbol{Y}$:

$$R(0,T) = \mathsf{E}_0\big[L(0,T(\boldsymbol{X}))\big] = \mathsf{E}_0\Big[\mathsf{E}_0\big[L(0,T_0(\boldsymbol{X})+v(\boldsymbol{Y})) \mid \boldsymbol{Y}\big]\Big].$$

Since the outer expectation sums over all possible values of $\boldsymbol{Y}$, we can minimize the total risk by minimizing the inner term *point-wise* for every specific value $\boldsymbol{y}$. Effectively, for each observed difference vector $\boldsymbol{y}$, we find the constant c that minimizes the loss, and set $v(\boldsymbol{y}) = c$.

Formal Construction For $\boldsymbol{X} = (X_1, ..., X_n)$ from the location model, we define the specific maximal invariant (as discussed in Example 6.9):[1]

$$\boldsymbol{Y} = Y(\boldsymbol{X}) = \boldsymbol{X} - X_n.$$

Theorem 6.1. *Suppose T_0 is an equivariant estimator with finite risk $R(0,T_0) < \infty$ under a location invariant loss L. For any $\boldsymbol{y} \in \mathbb{R}^n$ (where implicitly $y_n = 0$), define the function v^* by:*

$$v^*(\boldsymbol{y}) = \arg\min_{c\in\mathbb{R}} \mathsf{E}_0\big[L\big(0,T_0(\boldsymbol{X})+c\big) \mid \boldsymbol{Y}=\boldsymbol{y}\big]. \tag{6.3}$$

If v^* is well-defined, then the estimator

$$T^*(\boldsymbol{X}) := T_0(\boldsymbol{X}) + v^*(\boldsymbol{Y})$$

is a Minimum Risk Equivariant (MRE) estimator of θ.

Proof. The proof follows directly from the properties of conditional expectation and the definition of the minimum.

Let T be any other equivariant estimator. By Lemma 6.3, we can write $T(\boldsymbol{X}) = T_0(\boldsymbol{X}) + v(\boldsymbol{Y})$ for some function v. The risk at $\theta = 0$ is:

$$\begin{aligned} R(0,T) &= \mathsf{E}_0\big[L\big(0,T_0(\boldsymbol{X})+v(\boldsymbol{Y})\big)\big] \\ &= \mathsf{E}_0\Big[\mathsf{E}_0\big[L\big(0,T_0(\boldsymbol{X})+v(\boldsymbol{Y})\big) \mid \boldsymbol{Y}\big]\Big] \quad \text{(Tower Rule).} \end{aligned}$$

[1] Instead of Y we could consider any other maximal invariant.

Inside the inner expectation, we condition on $\boldsymbol{Y}$, so $v(\boldsymbol{Y})$ acts as a constant c. By definition, $v^*(\boldsymbol{Y})$ chooses the constant that minimizes this specific inner expectation. Therefore, for every $\boldsymbol{y}$:

$$\mathsf{E}_0\big[L\big(0, T_0(\boldsymbol{X}) + v(\boldsymbol{y})\big) \mid \boldsymbol{Y} = \boldsymbol{y}\big] \geqslant \mathsf{E}_0\big[L\big(0, T_0(\boldsymbol{X}) + v^*(\boldsymbol{y})\big) \mid \boldsymbol{Y} = \boldsymbol{y}\big].$$

Taking the expectation over $\boldsymbol{Y}$ preserves this inequality:

$$\mathsf{E}_0\Big[\mathsf{E}_0\big[L\big(0, T_0(\boldsymbol{X}) + v(\boldsymbol{Y})\big) \mid \boldsymbol{Y}\big]\Big] \geqslant \mathsf{E}_0\Big[\mathsf{E}_0\big[L\big(0, T_0(\boldsymbol{X}) + v^*(\boldsymbol{Y})\big) \mid \boldsymbol{Y}\big]\Big].$$

This simplifies to:

$$R(0, T) \geqslant R(0, T^*).$$

Thus, T^* minimizes the risk among all equivariant estimators. □

6.4.1 Existence and Uniqueness

Motivation: Minimizing the Risk We have reduced the problem of finding the MRE estimator to solving an optimization problem for each value $\boldsymbol{y}$ of the maximal invariant:

$$\min_{c \in \mathbb{R}} \mathsf{E}_0\big[L(0, T_0(\boldsymbol{X}) + c) \mid \boldsymbol{Y} = \boldsymbol{y}\big].$$

Does a solution always exist? Is it unique?

Since the loss function is location invariant, it depends only on the difference $\theta - a$. We can write $L(\theta, a) = \rho(\theta - a)$ with $\rho(0) = 0$. The properties of the function ρ (convexity, monotonicity) will determine whether the optimization problem is well-behaved.

Proposition 6.2. *Suppose there exists an equivariant estimator T_0 with finite risk $R(0, T_0)$.*

1. *If ρ is convex and not monotone, then an MRE estimator exists.*
2. *If ρ is strictly convex and not monotone, then the MRE estimator is a.e.-unique.*[2]

Proof. The conditional risk function we need to minimize is:

$$f(c) = \mathsf{E}_0\big[L\big(0, T_0(\boldsymbol{X}) + c\big) \mid \boldsymbol{Y} = \boldsymbol{y}\big] = \mathsf{E}_0\big[\rho\big(-T_0(\boldsymbol{X}) - c\big) \mid \boldsymbol{Y} = \boldsymbol{y}\big].$$

By the linearity and monotonicity of expectation:

- If ρ is convex, then $f(c)$ is convex (since a linear combination of convex functions is convex).
- If ρ is not monotone, then $f(c)$ is not monotone.

[2] Here, “a.e.” (almost everywhere) refers to all distributions in the model, similar to the definition for UMVUEs. See also Lehmann and Casella (1998, Theorem 1.7.15).

A convex, non-monotone function on $\mathbb{R}$ must have a global minimum (think of a parabola or a "U" shape). Thus, a minimizer $v^*(\boldsymbol{y})$ exists for every $\boldsymbol{y}$, making v^* well–defined. By Theorem 6.1, an MRE estimator exists.

If ρ is *strictly* convex, then $f(c)$ is strictly convex, which implies the minimizer is unique. Thus, $v^*(\boldsymbol{y})$ is uniquely determined for almost all $\boldsymbol{y}$. □

Example 6.10 (Gaussian Location Model)**.** Let $X_1, ..., X_n$ be i.i.d. $\mathcal{N}(\theta, \sigma_0^2)$ with σ_0^2 known. We consider the squared error loss $L(\theta, a) = (\theta - a)^2$, so $\rho(u) = u^2$.

Candidate Estimator The sample mean $\overline{X}_n$ is a natural candidate. We know:

- It is the MLE and UMVUE.
- It is equivariant (shifting data shifts the mean).
- It has constant risk (variance):

$$R(0, \overline{X}_n) = \mathsf{E}_0\left[\left(\overline{X}_n\right)^2\right] = \mathsf{Var}_0[\overline{X}_n] = \frac{\sigma_0^2}{n},$$

  where the equality with the variance holds because $\mathsf{E}_0\left[\overline{X}_n\right] = 0$.

Is it the MRE estimator?

Claim. $\overline{X}_n$ is the unique MRE estimator (unique up to Lebesgue null sets).

Proof. We apply Theorem 6.1 using $T_0(\boldsymbol{X}) = \overline{X}_n$ as our base estimator. We need to find:

$$v^*(\boldsymbol{y}) = \arg\min_{c\in\mathbb{R}} \mathsf{E}_0\left[(\overline{X}_n + c)^2 \mid \boldsymbol{Y} = \boldsymbol{y}\right].$$

Step 1: Independence. The maximal invariant is $\boldsymbol{Y} = (X_1 - X_n, \ldots, X_{n-1} - X_n, 0)$. The joint vector $(\overline{X}_n, \boldsymbol{Y})$ is multivariate normal because it consists of linear combinations of independent normal variables X_i.

For joint normals, independence is equivalent to zero covariance. Let's check:

$$\mathsf{Cov}_0[\overline{X}_n, X_i - X_n] = \mathsf{Cov}_0[\overline{X}_n, X_i] - \mathsf{Cov}_0[\overline{X}_n, X_n].$$

We know that $\mathsf{Cov}_0[\overline{X}_n, X_i] = \frac{\sigma_0^2}{n}$ for any i (by symmetry). Thus, the difference is zero.

Since $\overline{X}_n$ is uncorrelated with every component of $\boldsymbol{Y}$, it is independent of $\boldsymbol{Y}$.

Step 2: Optimization. Because of independence, the conditional expectation equals the unconditional expectation:

$$v^*(\boldsymbol{y}) = \arg\min_{c\in\mathbb{R}} \mathsf{E}_0\left[(\overline{X}_n + c)^2\right].$$

This is the standard problem of minimizing the second moment about a point c. The minimum of $\mathsf{E}[(Z + c)^2]$ occurs at $c = -\mathsf{E}[Z]$.

Here, $\mathsf{E}_0[\overline{X}_n] = 0$ (since the data is centered at 0). Therefore, the optimal shift is:

$$c^* = -\mathsf{E}_0[\overline{X}_n] = 0.$$

Conclusion: The optimal adjustment function is $v^*(\boldsymbol{y}) = 0$. The MRE estimator is:

$$T^*(\boldsymbol{X}) = T_0(\boldsymbol{X}) + 0 = \overline{X}_n.$$

□

6.4.2 Squared and absolute error loss

Theorem 6.1 is, in a way, the "end of the story" for the general theory – unless we bring in concrete loss functions. The optimization problem defined by $v^*(y)$ changes depending on the choice of L.

We now apply the general result to two specific, standard loss functions. The problem reduces to finding the constant c that is "closest" to the random variable $T_0(\boldsymbol{X})$ (conditional on $\boldsymbol{Y}$) under different metrics.

Squared Error Loss For squared error loss $L(\theta, a) = (\theta - a)^2$, we are looking for the constant that minimizes the expected squared deviation. Theorem 6.1 considers:

$$v^*(\boldsymbol{y}) = \arg\min_c \mathsf{E}_0\big[L(0, T_0(\boldsymbol{X}) + c) \mid \boldsymbol{Y} = \boldsymbol{y}\big]$$

Substituting the specific loss function:

$$v^*(\boldsymbol{y}) = \arg\min_c \mathsf{E}_0\big[(T_0(\boldsymbol{X}) + c)^2 \mid \boldsymbol{Y} = \boldsymbol{y}\big].$$

From basic probability theory (or our earlier discussions), we know that the quantity $\mathsf{E}[(Z-k)^2]$ is minimized when $k = \mathsf{E}[Z]$. Here, we are effectively minimizing expected square distance with respect to $-c$. Thus, the optimal c is the negative of the expectation:

$$v^*(\boldsymbol{y}) = -\mathsf{E}_0[T_0(\boldsymbol{X}) \mid \boldsymbol{Y} = \boldsymbol{y}].$$

The MRE estimator is, thus, the base estimator adjusted by its conditional expectation:

$$T^*(\boldsymbol{X}) = T_0(\boldsymbol{X}) - \mathsf{E}_0[T_0(\boldsymbol{X}) \mid \boldsymbol{Y}]. \tag{6.4}$$

Absolute Error Loss If we change to **absolute error loss** $L(\theta, a) = |\theta - a|$, the optimization problem becomes minimizing the mean absolute deviation:

$$\min_c \mathsf{E}_0\big[|T_0(\boldsymbol{X}) + c| \mid \boldsymbol{Y} = \boldsymbol{y}\big].$$

Recall from earlier lectures that the constant minimizing the mean absolute error $\mathsf{E}[|Z - k|]$ is the **median** of the distribution of Z.

Applying this to our conditional setup, we obtain the MRE estimator by subtracting the conditional median:

$$T^*(\boldsymbol{X}) = T_0(\boldsymbol{X}) - \text{median}_0[T_0(\boldsymbol{X}) \mid \boldsymbol{Y}]. \tag{6.5}$$

This explicitly shows how changing the loss function changes the optimal estimator – different loss functions emphasize different types of errors, and the MRE estimator adapts accordingly.

6.5 Pitman Estimator

We refer back to the model specification in Definition 6.1, with $X_i = \theta + \epsilon_i$ for additive error ϵ_i. When we specifically adopt **squared error loss**, the MRE estimator has a specific form and name: the **Pitman estimator**.

Theorem 6.2 (Pitman Estimator)**.** *Consider squared error loss $L(\theta, a) = (\theta - a)^2$, and suppose there exists an equivariant estimator with finite risk. Let $\epsilon_1, \ldots, \epsilon_n$ have joint density p_0 with respect to Lebesgue measure on $\mathbb{R}^n$. Then the MRE estimator is given by*

$$T^*(\boldsymbol{X}) = \frac{\int_{\mathbb{R}} z\, p_0(X_1 - z, \ldots, X_n - z)\, \mathrm{d}z}{\int_{\mathbb{R}} p_0(X_1 - z, \ldots, X_n - z)\, \mathrm{d}z}.$$

*In this form, the MRE estimator is known as the **Pitman estimator**.*

Proof. We use the construction from section 6.4 (specifically Theorem 6.1) applied to squared error loss. As derived in (6.4), using the base estimator $T_0(\boldsymbol{X}) = X_n$, the MRE estimator is:[3]

$$T^*(\boldsymbol{X}) = X_n - \mathsf{E}_0[X_n \mid \mathbf{Y}],$$

where $\boldsymbol{Y} = (X_1 - X_n, \ldots, X_{n-1} - X_n)$ is the maximal invariant. Note that implicitly $Y_n = X_n - X_n = 0$, but the randomness is driven by $\epsilon_1, \ldots, \epsilon_n$.

Step 1: Density Transformation. Consider the transformation from errors to the invariant/base statistics:

$$g : (\epsilon_1, \ldots, \epsilon_n) \mapsto (Y_1, \ldots, Y_{n-1}, \epsilon_n).$$

The Jacobian matrix of this transformation is upper triangular with 1s on the diagonal (since $Y_i = X_i - X_n = \epsilon_i - \epsilon_n$), so the Jacobian determinant is 1. The joint density of $(Y_1, ..., Y_{n-1}, \epsilon_n)$ is simply the original density p_0 evaluated at the pre–images:

$$\begin{aligned} f(y_1, \ldots, y_{n-1}, \epsilon_n) &= p_0(g^{-1}(y_1, \ldots, y_{n-1}, \epsilon_n)) \cdot \frac{1}{|\det(\dot{g}(g^{-1}(y_1, \ldots, y_{n-1}, \epsilon_n)))|} \\ &= p_0(y_1 + \epsilon_n, \ldots, y_{n-1} + \epsilon_n, \epsilon_n). \end{aligned}$$

Step 2: Conditional Expectation. The conditional density of X_n (which equals ϵ_n under $\theta = 0$) given $\boldsymbol{Y}$ is the joint density divided by the marginal density of $\mathbf{Y}$:

$$f(\epsilon_n \mid \boldsymbol{Y}) = \frac{p_0(Y_1 + \epsilon_n, \ldots, Y_{n-1} + \epsilon_n, \epsilon_n)}{\int_{\mathbb{R}} p_0(Y_1 + u, \ldots, Y_{n-1} + u, u)\, \mathrm{d}u}.$$

Now we compute the expectation $\mathsf{E}_0[X_n \mid \boldsymbol{Y}]$:

$$\mathsf{E}_0[X_n \mid \boldsymbol{Y}] = \mathsf{E}_0\left[\epsilon_n | \boldsymbol{Y}\right] = \int_{\mathbb{R}} \epsilon_n\, f(\epsilon_n \mid \mathbf{Y})\, \mathrm{d}\epsilon_n = \frac{\int u\, p_0(Y_1 + u, \ldots, Y_{n-1} + u, u)\, \mathrm{d}u}{\int p_0(Y_1 + u, \ldots, Y_{n-1} + u, u)\, \mathrm{d}u}.$$

[3] Here, $\mathsf{E}_0\left[|X_n| \mid \boldsymbol{Y}\right] < \infty$ by the assumed existence of an equivariant estimator with finite risk; compare Lehmann and Casella (1998, Problem 3.1.21).

Step 3: Substitution. We start with the expression for the conditional expectation derived above. We first substitute the definition of the maximal invariant components, $Y_i = X_i - X_n$, back into the density arguments. Note that the last argument u can be written as $X_n - X_n + u$:

$$p_0(X_1 - X_n + u, \dots, X_{n-1} - X_n + u, X_n - X_n + u).$$

Next, we perform a change of variables to simplify the arguments. Let $z = X_n - u$. This implies:

$$u = X_n - z \quad \text{and} \quad \mathrm{d}u = -\mathrm{d}z.$$

The limits of integration $(-\infty, \infty)$ for u correspond to $(\infty, -\infty)$ for z. The negative sign from the differential $\mathrm{d}u$ flips the limits back to $(-\infty, \infty)$.

Now, we transform the arguments inside the density function using this substitution:

$$X_i - X_n + u = X_i - X_n + (X_n - z) = X_i - z.$$

This holds for all coordinates $i = 1, \dots, n$ (including the last one).

We also substitute $u = X_n - z$ into the linear term in the numerator. The expression becomes:

$$\mathsf{E}_0[X_n \mid \boldsymbol{Y}] = \frac{\int (X_n - z)\, p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z}{\int p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z}.$$

Using linearity, the numerator splits: $X_n \cdot (\text{denominator}) - \int zp_0(\dots)\mathrm{d}z$. Thus:

$$\begin{aligned}\mathsf{E}_0[X_n \mid \boldsymbol{Y}] &= \frac{X_n \int p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z - \int z\, p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z}{\int p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z} \\ &= X_n - \frac{\int z\, p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z}{\int p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z}.\end{aligned}$$

Finally, substituting this back into $T^*(\boldsymbol{X})$

$$T^*(\boldsymbol{X}) = X_n - \mathsf{E}_0[X_n \mid \boldsymbol{Y}] = \frac{\int z\, p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z}{\int p_0(X_1 - z, \dots, X_n - z)\,\mathrm{d}z}.$$

yields the Pitman estimator. □

Example 6.11 (Uniform Model)**.** Let $X_1, \dots, X_n$ be i.i.d. Uniform $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathbb{R}$. It is known that no UMVUE exists for this model (Lehmann and Casella 1998, Ex. 2.1.9). However, we can find the optimal equivariant estimator.

The joint density under $\theta = 0$ is:

$$p_0(x_1, \dots, x_n) = 1_{\{|x_1| < \frac{1}{2}, \dots, |x_n| < \frac{1}{2}\}}.$$

To apply the Pitman formula, we evaluate this density at shifted points $X_i - z$. The condition for the density to be non-zero is:

$$|X_i - z| < \frac{1}{2} \quad \forall i.$$

Rearranging this inequality for z:

$$z - \frac{1}{2} < X_i < z + \frac{1}{2} \iff X_i - \frac{1}{2} < z < X_i + \frac{1}{2}.$$

For this to hold for *all* i, z must be greater than the largest lower bound and smaller than the smallest upper bound. Let $X_{(1)} = \min(X_i)$ and $X_{(n)} = \max(X_i)$. The range of integration is:

$$X_{(n)} - \frac{1}{2} < z < X_{(1)} + \frac{1}{2}.$$

Computing the Estimator: The Pitman estimator is the ratio of two integrals over this interval $[a, b]$, where $a = X_{(n)} - 1/2$ and $b = X_{(1)} + 1/2$.

- **Denominator:** $\int_a^b 1 \, \mathrm{d}z = b - a$.

- **Numerator:** $\int_a^b z \, \mathrm{d}z = \left[\frac{z^2}{2}\right]_a^b = \frac{1}{2}(b^2 - a^2) = \frac{1}{2}(b-a)(b+a)$.

The ratio is:

$$T^*(\boldsymbol{X}) = \frac{\frac{1}{2}(b-a)(b+a)}{b-a} = \frac{a+b}{2}.$$

Substituting a and b back:

$$\frac{(X_{(n)} - 1/2) + (X_{(1)} + 1/2)}{2} = \frac{X_{(1)} + X_{(n)}}{2}.$$

Thus, the Pitman estimator is the **mid-range** of the data.

6.6 MRE and Unbiasedness

In the previous section, we saw that for location models like the Uniform distribution, a UMVUE may not exist, yet the Pitman estimator provides a unique optimal solution. This raises a natural question: what is the relationship between this MRE estimator and unbiasedness?

Proposition 6.3. *Fix squared error loss $L(\theta, a) = (\theta - a)^2$ and suppose there exists an equivariant estimator of finite risk. Then:*

(i) The bias of any equivariant estimator is constant.

(ii) The MRE estimator is unbiased.

(iii) If an equivariant estimator T is unbiased and $T(\boldsymbol{X})$ is independent of $\boldsymbol{X} - T(\boldsymbol{X})$, then T is the MRE estimator.

Proof.

(i) The bias of an estimator T is $b(\theta) = \mathsf{E}_\theta[T(\boldsymbol{X})] - \theta$. Using the equivariance property $T(\boldsymbol{x}+c) = T(\boldsymbol{x}) + c$, we can write:

$$\mathsf{E}_\theta[T(\boldsymbol{X})] - \theta = \mathsf{E}_\theta[T(\boldsymbol{X}) - \theta] = \mathsf{E}_\theta[T(\boldsymbol{X} - \theta)].$$

When $\boldsymbol{X} \sim \mathsf{P}_\theta$, the shifted variable $\boldsymbol{X} - \theta$ is distributed according to P_0 (the error distribution). Thus:

$$\mathsf{E}_\theta[T(\boldsymbol{X} - \theta)] = \mathsf{E}_0[T(\boldsymbol{X})].$$

This expectation depends only on the fixed distribution P_0 and is therefore constant for all $\theta \in \mathbb{R}$.

(ii) Suppose the MRE estimator T^* has a non–zero bias $b = \mathsf{E}_0[T^*] \neq 0$. We can construct a new estimator $T'(\boldsymbol{X}) = T^*(\boldsymbol{X}) - b$. This estimator is still equivariant (shifting by a constant does not break equivariance).

Consider the risk at $\theta = 0$ (which is the constant risk for equivariant estimators). The risk of T' is:

$$R(0, T') = \mathsf{E}_0[(T^* - b)^2] = \mathsf{Var}_0[T^*].$$

The risk of the original estimator T^* is its second moment:

$$R(0, T^*) = \mathsf{E}_0[(T^*)^2] = \mathsf{Var}_0[T^*] + (\mathsf{E}_0[T^*])^2 = \mathsf{Var}_0[T^*] + b^2.$$

Since $b \neq 0$, we have $b^2 > 0$, which implies $R(0, T') < R(0, T^*)$. This contradicts the assumption that T^* minimizes the risk. Therefore, the bias b must be zero.

(iii) From the construction in Theorem 6.1 (and the squared error derivation), we know that the MRE estimator is unique (a.e.) and is given by adjusting any equivariant estimator T by its conditional expectation given a maximal invariant.

The vector of residuals $\boldsymbol{U} = \boldsymbol{X} - T(\boldsymbol{X})$ is a maximal invariant. Thus, the MRE estimator T^* is:

$$T^*(\boldsymbol{X}) = T(\boldsymbol{X}) - \mathsf{E}_0[T(\boldsymbol{X}) \mid \boldsymbol{X} - T(\boldsymbol{X})].$$

Therefore, T is the MRE estimator if and only if the adjustment term is zero:

$$\mathsf{E}_0\big[T(\boldsymbol{X}) \mid \boldsymbol{X} - T(\boldsymbol{X})\big] = 0. \tag{6.6}$$

If $T(\boldsymbol{X})$ is independent of $\boldsymbol{X} - T(\boldsymbol{X})$, the conditional expectation becomes the unconditional expectation:

$$\mathsf{E}_0[T(\boldsymbol{X}) \mid \boldsymbol{X} - T(\boldsymbol{X})] = \mathsf{E}_0[T(\boldsymbol{X})].$$

Since T is unbiased, $\mathsf{E}_0[T(\boldsymbol{X})] = 0$. Thus, condition (6.6) is satisfied, and T is the MRE estimator.

□

6.7 General In–/Equivariance

Setting:
We generalize the location model to general group actions. The setting consists of:

- X: observation (data).
- $\mathcal{X}$: sample space.
- $\mathcal{P} = \{\mathsf{P}_\theta \mid \theta \in \Theta\}$: model with $\mathsf{P}_\theta \neq \mathsf{P}_{\theta'}$ if $\theta \neq \theta'$ (identifiability).
- $\mathcal{G}$: a group acting on $\mathcal{X}$.

Definition 6.7 (Invariant Model)**.** The model $\mathcal{P}$ is **invariant** under $\mathcal{G}$ if for all $g \in \mathcal{G}$:

1. If $X \sim \mathsf{P}_\theta$ for $\theta \in \Theta$, then the transformed data $gX \sim \mathsf{P}_{\theta'}$ for some $\theta' \in \Theta$. We denote this induced parameter as $\theta' := \bar{g}\theta$.
2. The mapping on the parameter space is surjective: $\{\bar{g}\theta \mid \theta \in \Theta\} = \Theta$.

Fact It follows that:

- The induced map $\bar{g} : \Theta \to \Theta$ is bijective.
- The collection of induced maps $\bar{\mathcal{G}} = \{\bar{g} \mid g \in \mathcal{G}\}$ forms a group acting on the parameter space.

Definition 6.8. An estimator $T : \mathcal{X} \to \Theta$ is **equivariant** if it preserves the group action structure:

$$T(gx) = \bar{g}T(x) \quad \forall g \in \mathcal{G},\ \forall x \in \mathcal{X}.$$

A loss function $L(\cdot, \cdot)$ is **invariant** if the loss remains unchanged when both the truth and the estimate are transformed:

$$L(\bar{g}\theta, \bar{g}a) = L(\theta, a) \quad \forall a, \theta \in \Theta,\ \bar{g} \in \bar{\mathcal{G}}.$$

Examples:

- **Location–scale model:** The group action on $\boldsymbol{x} \in \mathbb{R}^n$ is given by affine transformations:

  $$\boldsymbol{x} \mapsto a\boldsymbol{x} + c \quad \text{for } c \in \mathbb{R},\ a > 0.$$

  Here, the group acts by both shifting (location) and scaling (variance). See Lehmann and Casella (1998, Sec. 3.3).

- **Linear regression:** See Lehmann and Casella (1998, Sec. 3.4).

  Note. Linear regression assumes the expectation vector lies in a linear subspace. This subspace is invariant under various coordinate transformations. The standard **F–test**, which is used globally for hypothesis testing in regression (e.g., testing if coefficients are zero or if two groups share the same slope), can be derived and justified as the optimal invariant procedure under the relevant group of transformations.

Outlook: Decision Theory
This concludes our discussion on uniformly optimal procedures (like UMVUE and MRE). In many general settings, no single estimator is uniformly best for all θ. The next chapters will focus on **Decision Theory**, where we summarize performance via:

- **Average performance:** Bayes estimators (averaging over a prior).
- **Worst–case performance:** Minimax estimators.

7. Decision Theory

Decision theory provides a common framework to discuss different statistical tasks (estimation, testing, confidence sets, etc.). We will adopt this framework to introduce the concepts of Bayes and Minimax optimality.

Previously, we restricted ourselves to specific families of estimators (e.g., unbiased or equivariant estimators) and sought **uniform optimality**. Uniform optimality requires that an estimator T^* dominates all others for every possible parameter value:

$$R(\theta, T^*) \leqslant R(\theta, T) \quad \forall \theta \in \Theta.$$

As we have discussed, finding such an estimator without strong restrictions is often impossible.

Now, we shift our perspective. Instead of restricting the class of estimators, we change how we measure "goodness." We will evaluate procedures based on an **aggregated risk**:

- **Average Risk (Bayes):** Weighting the risk across different θ values (using a prior).
- **Worst–case Risk (Minimax):** Looking at the maximum risk over all θ.

7.1 Decision Rules

Setup We begin with the standard statistical setup:

- X: The observation (data), which could be a vector, matrix, image, etc.
- $\mathcal{X}$: The sample space (the set of all possible outcomes).
- $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$: The statistical model, where θ is the unknown parameter.

Actions In decision theory, we formally define what we "do" with the data as an **action**. Our goal is to unify different statistical tasks (estimation, testing, etc.) under this common terminology.

Let $\mathcal{A}$ denote the **action space**, the set of all possible actions available to the statistician.

Example 7.1. This language allows us to unify various statistical tasks:

i. **Point Estimation:** The action is to return a value (an estimate) for the parameter.

$$\mathcal{A} = \Theta \quad (\text{or } \mathcal{A} \subseteq \mathbb{R}^r).$$

For example, if we are estimating the speed of light, the action is simply providing a number.

ii. **Hypothesis Testing:** The goal is to decide between two competing hypotheses which partition the parameter space:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1,$$

where $\Theta_0, \Theta_1 \subseteq \Theta$ and $\Theta_0 \cap \Theta_1 = \emptyset$.

Possible action spaces include:

- **Deterministic Tests:** $\mathcal{A} = \{0, 1\}$, where usually 0 denotes "Fail to reject H_0" (Accept) and 1 denotes "Reject H_0".
- **Randomized Tests:** $\mathcal{A} = [0, 1]$. Here, the action $a \in \mathcal{A}$ represents the *probability* to reject H_0. If the procedure returns 0.7, you flip a biased coin that lands heads with probability 0.7 to make the final decision.

iii. **Confidence Sets:** The action is to select a subset of the parameter space.

$$\mathcal{A} = \{C \subseteq \Theta : C \text{ is a valid set}\}.$$

For interval estimation of a parameter $\gamma(\theta) \in \mathbb{R}$, the action space is the set of all intervals (ℓ, u) with $-\infty \leq \ell \leq u \leq \infty$.

Once we have defined the available actions, we need a rule for choosing one based on the observation.

Definition 7.1. A **decision rule** (or statistical procedure) is a measurable mapping from the sample space to the action space:

$$d : \mathcal{X} \to \mathcal{A}.$$

In the context of estimation, $d(X)$ is what we previously called an estimator $T(X)$. In testing, $d(X)$ is the test function.

7.1.1 Risk of a Decision Rule

To evaluate decision rules, we must quantify the cost of making a specific decision when the state of nature is θ.

Loss Function A loss function $L : \Theta \times \mathcal{A} \to [0, \infty)$ quantifies the "loss" or "price" of taking an action $a \in \mathcal{A}$ when nature is in state $\theta \in \Theta$.

Definition 7.2. The **risk** of a decision rule d is the expected loss:

$$R(\theta, d) = \mathsf{E}_\theta[L(\theta, d(X))], \quad \theta \in \Theta.$$

As in earlier chapters, E_θ indicates that the expectation is taken with respect to the random observation $X \sim \mathsf{P}_\theta$.

Optimality?

How do we choose the best rule? There are two main approaches:

Approach I: Uniform Optimality Find a rule d^* that minimizes the risk for *every* parameter value simultaneously:

$$R(\theta, d^*) \leq R(\theta, d) \quad \forall d, \forall \theta \in \Theta.$$

- This corresponds to finding uniformly optimal rules within a restricted class.
- Examples: UMVUE (Unbiasedness restriction), MRE (Equivariance restriction).

Approach II: One-number Summary of Risk Since uniform optimality is often impossible without restrictions, we can instead summarize the risk function $R(\cdot, d)$ into a single scalar value and minimize that.

- **Average-case optimality (Bayes):** We minimize the weighted average of the risk (weighted by a prior).
- **Worst-case optimality (Minimax):** We minimize the maximum possible risk.

Example 7.2 (continued)**.**

i. **Estimation of a parameter** $\gamma(\theta) \in \mathbb{R}^r$: Consider the squared error loss:

$$L(\theta, a) = \|\gamma(\theta) - a\|^2.$$

The risk is the Mean Squared Error (MSE):

$$R(\theta, d) = \mathsf{E}_\theta[\|\gamma(\theta) - d(X)\|^2].$$

ii. **Hypothesis Testing (Non-randomized)**: Let $\mathcal{A} = \{0, 1\}$. We test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. We can define the "Neyman-Pearson" loss function as:

$$L(\theta, a) = \begin{cases} 1, & \text{if } \theta \in \Theta_0 \text{ and } a = 1 \quad (\text{"Type I error"}), \\ c, & \text{if } \theta \in \Theta_1 \text{ and } a = 0 \quad (\text{"Type II error"}), \\ 0, & \text{otherwise.} \end{cases}$$

Here, $c > 0$ is a constant that weights the relative severity of the errors.

Note. The constant c allows us to model asymmetric costs. For example, in a clinical trial, approving a harmful drug (Type I error) might be considered more "grave" than failing to approve a beneficial one (Type II error), or vice versa depending on the context.

The risk function becomes the probability of making an error, weighted by the cost:

$$R(\theta, d) = \begin{cases} \mathsf{P}_\theta\big(d(X) = 1\big), & \text{if } \theta \in \Theta_0, \\ c\, \mathsf{P}_\theta\big(d(X) = 0\big), & \text{if } \theta \in \Theta_1, \\ 0, & \text{if } \theta \notin \Theta \backslash (\Theta_0 \cup \Theta_1). \end{cases}$$

(The last case is typically not of interest as the hypotheses usually partition the parameter space).

7.1.2 Admissibility

We now introduce a property that essentially formalizes the idea of "don't be stupid." If there exists a decision rule that performs as well as yours in every scenario and strictly better in at least one, there is no justification for keeping your current rule.

Definition 7.3. A decision rule d' is said to be **strictly better** (or dominates) a decision rule d if:

1. $R(\theta, d') \leq R(\theta, d)$ for all $\theta \in \Theta$;
2. There exists at least one $\theta \in \Theta$ such that $R(\theta, d') < R(\theta, d)$.

A decision rule d is **admissible** if no other rule d' is strictly better than d.
If there exists a rule d' that is strictly better than d, then d is **inadmissible**.

Admissible rules are those that lie on the "Pareto frontier" of risk. If you plot the risk functions of two admissible estimators, their curves will typically cross; neither is uniformly better than the other. An inadmissible estimator's risk curve lies strictly above another estimator's curve (or touches it but goes higher elsewhere).

Example 7.3 (Gaussian mean)**.** Let $X_1, \ldots, X_n$ be i.i.d. observations from the statistical model $\mathcal{P} = \{\mathcal{N}(\mu, \sigma^2) : \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)\}$. We consider the estimation of $\gamma(\theta) = \mu$ under squared error loss.

a) **Inadmissibility of the "Partial" Mean:** Suppose someone is lazy and uses only the first m observations ($m < n$) to estimate the mean:

$$\overline{X}_m = \frac{1}{m} \sum_{i=1}^{m} X_i.$$

This is clearly inadmissible. Comparing it to the full sample mean $\overline{X}_n$:

$$R(\theta, \overline{X}_m) = \mathsf{Var}_\theta[\overline{X}_m] = \frac{\sigma^2}{m} > \frac{\sigma^2}{n} = R(\theta, \overline{X}_n) \quad \forall \theta.$$

The risk of $\overline{X}_n$ is strictly lower for all θ, so $\overline{X}_m$ is dominated.

b) **Admissibility of the Sample Mean (1D vs 3D):**

- In **1 dimension**, the standard sample mean $\overline{X}_n$ is admissible. (We will prove this later).
- **Stein's Paradox:** Surprisingly, if we estimate a vector of means $\boldsymbol{\mu} \in \mathbb{R}^p$ for $p \geq 3$ (e.g., estimating blood pressure averages for 3 different metrics simultaneously) under the combined squared error loss, the sample mean vector is **inadmissible**. It is dominated by "shrinkage estimators" (like the James–Stein estimator).

c) **Admissibility of Constant Estimators:** Consider a trivial estimator that ignores the data and always predicts a constant $T(X) = \mu_0$ (e.g., "17").

Is this admissible? Yes.

Reasoning: If the true state of nature happens to be $\theta = (\mu_0, \sigma^2)$, the estimator is perfect:

$$R((\mu_0, \sigma^2), T) = \mathsf{E}[(\mu_0 - \mu_0)^2] = 0.$$

For another rule T' to be *strictly better*, it must never have higher risk than T. This means T' must also have risk 0 at $\theta = (\mu_0, \sigma^2)$. Since risk is variance plus bias squared, a risk of 0 implies T' must essentially equal μ_0 with probability 1. Therefore, T' cannot be different from T, so T cannot be dominated.

7.2 Bayes Rules

To move beyond the constraints of uniform optimality, we introduce the concept of "average risk." To compute an average, we must specify how to weight the different parameter values. This weighting is provided by a prior distribution.

Let Π be a probability measure on the parameter space Θ. We refer to Π as the **prior distribution**.

Definition 7.4. The **Bayes risk** of a decision rule d (with respect to the prior distribution Π) is the weighted average of its risk function:

$$r(\Pi, d) = \int_\Theta R(\theta, d) \, \mathrm{d}\Pi(\theta).$$

Definition 7.5. A decision rule d is called a **Bayes rule** (with respect to Π) if it minimizes the Bayes risk:

$$r(\Pi, d) = \inf_{d'} r(\Pi, d').$$

Essentially, a Bayes rule achieves the "smallest average risk."

Densities and Posterior Risk

In most practical applications, we work with models where distributions have densities (e.g., with respect to Lebesgue measure).

Setup Assume the following densities exist:

- **Sampling density:** $p_\theta(x) = \frac{d\mathsf{P}_\theta(x)}{d\nu}$, corresponding to the data–generating distribution P_θ.
- **Prior density:** $\pi(\theta) = \frac{d\Pi(\theta)}{d\mu}$, corresponding to the prior Π.

Posterior Distribution Using Bayes' theorem, the **posterior density** of θ given the observation $X = x$ is:

$$\pi(\theta|x) = \frac{p_\theta(x)\pi(\theta)}{\int_\Theta p_\theta(x)\pi(\theta)\,\mathrm{d}\theta} = \frac{p_\theta(x)\pi(\theta)}{m(x)},$$

where $m(x)$ is the marginal density of X (sometimes called the *prior predictive density*).

Posterior Risk This framework allows us to define a new type of risk. Instead of averaging over the data X (as in the frequentist risk $R(\theta, d)$), we condition on the observed data x and average over the parameter θ.

Definition 7.6. Under the prior Π and having observed $X = x$, the **posterior risk** of taking a specific action $a \in \mathcal{A}$ is the expected loss with respect to the posterior distribution:

$$\begin{aligned}\ell(x, a) &:= \mathsf{E}\big[L(\theta, a) \,|\, X = x\big] \\ &= \int_\Theta L(\theta, a)\pi(\theta|x)\,\mathrm{d}\theta.\end{aligned}$$

Consider the distinction between the two risks:

- **Frequentist Risk** $R(\theta, d)$: Fixes θ (state of nature) and averages over random X.
- **Posterior Risk** $\ell(x, a)$: Fixes x (observed data) and action a, and averages over random θ.

7.2.1 Construction of Bayes Rules

We now turn to the practical construction of Bayes rules. The following theorem provides a constructive method: to minimize the average risk (Bayes risk), one simply needs to minimize the posterior risk for every observed x.

Theorem 7.1. *Let $\ell(x, a)$ be the posterior risk defined in the previous section. Assume there exists a rule d_0 with finite risk. Then, the decision rule d defined as the minimizer of the posterior risk,*

$$d(x) := \arg\min_{a \in \mathcal{A}} \ell(x, a), \quad x \in \mathcal{X},$$

is a Bayes rule.

Going forward, we will make the mild assumption that the above construction is well–defined for any prior distribution (i.e., the posterior risk admits a minimizer).

Remark. If S is a sufficient statistic, then by the Factorization Theorem (Neyman's criterion), we can write $p_\theta(x) = g_\theta(S(x))h(x)$. Substituting this into the posterior risk definition:

$$\begin{aligned}\ell(x,a) &= \int_\Theta L(\theta,a)\pi(\theta|x)\,\mathrm{d}\mu(\theta)\\ &= \int_\Theta L(\theta,a)\frac{p_\theta(x)\pi(\theta)}{p_\pi(x)}\,\mathrm{d}\mu(\theta)\\ &= \frac{h(x)}{p_\pi(x)}\int_\Theta L(\theta,a)g_\theta(S(x))\pi(\theta)\,\mathrm{d}\mu(\theta).\end{aligned}$$

Since the term $\frac{h(x)}{p_\pi(x)}$ does not depend on the action a, minimizing $\ell(x,a)$ is equivalent to minimizing the integral term, which depends on data only through $S(x)$. Thus, the constructed Bayes rule depends only on the sufficient statistic S.

Proof. The proof relies on Fubini's theorem to swap the order of integration. We want to show that our constructed rule d has a Bayes risk $r(\Pi,d)$ that is no larger than that of any other rule d'.

For any decision rule d' with finite risk:

$$\begin{aligned}r(\Pi,d') &= \int_\Theta R(\theta,d')\pi(\theta)\,\mathrm{d}\mu(\theta)\\ &= \int_\Theta\left[\int_\mathcal{X} L(\theta,d'(x))p_\theta(x)\,\mathrm{d}\nu(x)\right]\pi(\theta)\,\mathrm{d}\mu(\theta)\\ &\overset{\text{Fubini}}{=} \int_\mathcal{X}\left[\int_\Theta L(\theta,d'(x))\underbrace{p_\theta(x)\pi(\theta)}_{=\pi(\theta|x)p_\pi(x)}\,\mathrm{d}\mu(\theta)\right]\mathrm{d}\nu(x)\\ &= \int_\mathcal{X}\left[\int_\Theta L(\theta,d'(x))\pi(\theta|x)\,\mathrm{d}\mu(\theta)\right]p_\pi(x)\,\mathrm{d}\nu(x)\\ &= \int_\mathcal{X}\ell(x,d'(x))p_\pi(x)\,\mathrm{d}\nu(x).\end{aligned}$$

Since $d(x)$ is defined as the minimizer of $\ell(x,a)$ for every x, we have $\ell(x,d(x)) \leq \ell(x,d'(x))$ pointwise. Therefore:

$$r(\Pi,d') = \int_\mathcal{X}\ell(x,d'(x))p_\pi(x)\,\mathrm{d}\nu(x) \geqslant \int_\mathcal{X}\ell(x,d(x))p_\pi(x)\,\mathrm{d}\nu(x) = r(\Pi,d).$$

□

Example 7.4. The form of the Bayes rule depends entirely on the chosen loss function.

i. **Squared Error Loss** (Estimation of $\gamma(\theta)\in\mathbb{R}$):

$$L(\theta,a) = (\gamma(\theta)-a)^2$$

The Bayes rule minimizes the posterior expected squared error:

$$d(x) = \arg\min_{a \in \mathcal{A}} \mathsf{E}\big[(\gamma(\theta) - a)^2 \,|\, X = x\big].$$

As shown in introductory statistics, this is minimized by the expectation:

$$d(x) = \mathsf{E}[\gamma(\theta) \,|\, X = x] \quad \longrightarrow \quad \textbf{Posterior Mean}.$$

ii. **Absolute Error Loss** (Estimation of $\gamma(\theta) \in \mathbb{R}$):

$$L(\theta, a) = |\gamma(\theta) - a|$$

The Bayes rule minimizes the posterior expected absolute deviation:

$$d(x) = \arg\min_{a \in \mathcal{A}} \mathsf{E}[|\gamma(\theta) - a| \,|\, X = x] \quad \longrightarrow \quad \textbf{Posterior Median}.$$

iii. **0/1 "Hit or Miss" Loss** (Estimation of $\gamma(\theta) = \theta \in \mathbb{R}$): Consider a loss that penalizes being further than distance c from the truth:

$$L(\theta, a) = \mathbf{1}_{\{|\theta - a| > c\}} \quad \text{for some given } c > 0.$$

The posterior risk is the probability of the error being large:

$$\ell(x, a) = \Pi(|\theta - a| > c \,|\, X = x) = 1 - \Pi(|\theta - a| \leq c \,|\, X = x).$$

Minimizing this risk is equivalent to maximizing the probability of being within the interval $[a - c, a + c]$. If we consider the limit as $c \downarrow 0$:

$$\frac{1 - \ell(x, a)}{2c} = \frac{\Pi(|\theta - a| \leq c \,|\, X = x)}{2c} \approx \pi(a|x).$$

Hence, for small c, the constructed Bayes rule maximizes the posterior density:

$$d_{\text{MAP}}(x) = \arg\max_{\theta} \pi(\theta|x) = \arg\max_{\theta} p_\theta(x)\pi(\theta).$$

This is known as the **Maximum A Posteriori (MAP)** estimator.

Remark. If the prior is uniform, $\pi(\theta) = \text{const}$ (which strictly requires Θ to be bounded to integrate to 1), then maximizing the posterior is equivalent to maximizing the likelihood. In this specific case, MAP = MLE.

7.3 Minimax Decision Rules

Another major perspective in decision theory is focusing on the "worst–case" scenario. While Bayes rules optimize for the average case (given a prior), minimax rules optimize for the worst possible state of nature.

Definition 7.7. A decision rule d is **minimax** if

$$\sup_{\theta \in \Theta} R(\theta, d) = \inf_{d'} \sup_{\theta \in \Theta} R(\theta, d').$$

In other words, the worst–case risk of d is the minimal worst–case risk achievable by any rule.

Remark (Not relevant for this course). In asymptotic statistics (where $n \to \infty$), strict minimaxity is often hard to derive. You may encounter the following relaxations in research literature:

- A sequence of rules d_n is **asymptotically minimax** if
$$\sup_\theta R_n(\theta, d_n) \sim \inf_{d'} \sup_\theta R_n(\theta, d').$$
(Here $a_n \sim b_n$ denotes $a_n/b_n \to 1$).

- A sequence achieves the **minimax rate** if
$$\sup_\theta R_n(\theta, d_n) \asymp \inf_{d'} \sup_\theta R_n(\theta, d').$$
(Here $\asymp$ denotes that the ratio is bounded away from 0 and ∞).

Example 7.5 (Gaussian Mean)**.** If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$, then the sample mean $\overline{X}_n$ **is** the minimax estimator of μ under squared error loss. (We will prove this later).

Example 7.6 (Binomial Proportion: Counter–Intuitive Result)**.** Let $X \sim \text{Bin}(n, p)$. We wish to estimate $p \in (0, 1)$ under squared error loss.

Consider the natural estimator $d(X) = \frac{X}{n}$ (which is the MLE and UMVUE). Its risk is:
$$R(p, d) = \mathsf{E}_p\left[\left(\frac{X}{n} - p\right)^2\right] = \mathsf{Var}_p\left[\frac{X}{n}\right] = \frac{p(1-p)}{n}.$$
The worst-case risk occurs at $p = 1/2$:
$$\sup_{p\in(0,1)} R(p, d) = \frac{1}{n} \cdot \frac{1}{4} = \frac{1}{4n}.$$

Surprisingly, $d(X) = \frac{X}{n}$ is **not** minimax under squared error loss. We can construct a rule with a lower worst-case risk.

Demonstration: A Randomized Estimator of a Binomial probability

To see why X/n is not minimax, we can construct a randomized rule that achieves a smaller maximum risk.

Consider two base estimators:

1. $d(X) = \frac{X}{n}$: Risk is maximal at $p = 1/2$.
2. $d'(X) \equiv \frac{1}{2}$: A constant estimator. Risk is
$$R(p, d') = \mathsf{E}_p\left[\left(\frac{1}{2} - p\right)^2\right] = \left(p - \frac{1}{2}\right)^2.$$
This risk is zero at $p = 1/2$ and maximal at $p \in \{0, 1\}$.

We can create a trade–off by mixing these two. Define a randomized rule d_r that chooses d with probability $1-\epsilon$ and d' with probability ϵ. Formally, let $U \sim \text{Uniform}(0,1)$ independent of X:

$$d_r(X,U) = \frac{X}{n}\mathbf{1}_{(0,1-\epsilon]}(U) + \frac{1}{2}\mathbf{1}_{(1-\epsilon,1)}(U).$$

The risk of this randomized rule is the convex combination of the individual risks:

$$\begin{aligned} R(p,d_r) &= \mathsf{E}_{X,U}\left[\left(d_r(X,U)-p\right)^2\right] \\ &= (1-\epsilon)R(p,d) + \epsilon R(p,d') \\ &= (1-\epsilon)\frac{p(1-p)}{n} + \epsilon\left(p-\frac{1}{2}\right)^2. \end{aligned}$$

By choosing $\epsilon = \frac{1}{n+1}$, the risk becomes constant (flat) across all p:

$$R(p,d_r) = \frac{1}{4(n+1)}.$$

Since $\frac{1}{4(n+1)} < \frac{1}{4n}$, this randomized rule has a strictly lower worst-case risk than X/n. Therefore, X/n cannot be minimax.

Remark. We used a randomized rule here for intuition. Later, we will derive a deterministic minimax estimator for the Binomial model (the Hodges–Lehmann estimator) which relates closely to this logic.

7.4 Decision Rules With Constant Risk

Finding minimax rules directly can be difficult. However, there is a powerful strategy that leverages Bayes rules. If we can find a Bayes rule whose risk function turns out to be constant (flat) across all parameters, we can conclude it is minimax.

Lemma 7.1. *Suppose a decision rule d has constant risk, i.e., $R(\theta,d) \equiv R(d)$ for all $\theta \in \Theta$. Then:*

(i) If d is admissible, then d is minimax.

(ii) If d is Bayes (with respect to some prior Π), then d is minimax.

Proof.

(i) Let d' be any other decision rule. Since d is admissible, d' cannot be strictly better than d. This implies that either d' has higher risk for some θ, or it has equal risk everywhere. In either case:

$$\sup_\theta R(\theta,d') \geq R(d) = \sup_\theta R(\theta,d).$$

Thus, d minimizes the worst–case risk.

(ii) This is the more practical condition. Let d be Bayes for prior Π. For any rule d':

$$\begin{aligned}\sup_\theta R(\theta, d') &\geq \int_\Theta R(\theta, d')\,\mathrm{d}\Pi(\theta) = r(\Pi, d') \quad \text{(Max } \geq \text{ Average)}\\ &\geq r(\Pi, d) \quad \text{(Since } d \text{ is Bayes)}\\ &= \int_\Theta R(d)\,\mathrm{d}\Pi(\theta) = R(d) \quad \text{(Risk is constant)}\\ &= \sup_\theta R(\theta, d).\end{aligned}$$

Therefore, d is minimax.

□

Example 7.7 (Binomial Model Minimax Estimator)**.** Consider $X \sim \mathrm{Bin}(n, \theta)$. We wish to estimate $\theta \in [0, 1]$ under squared error loss $L(\theta, a) = (\theta - a)^2$.

We previously saw that the MLE X/n is not minimax. We will construct a minimax estimator by finding a Bayes rule with constant risk.

The Bayes Rule Family Consider the Conjugate Prior $\Pi = \mathrm{Beta}(\alpha, \beta)$. The Bayes estimator (posterior mean) is:

$$\hat{\theta}_{\alpha,\beta}(X) = \frac{X + \alpha}{n + \alpha + \beta}.$$

The Risk Function Using the bias–variance decomposition:

$$\begin{aligned}R(\theta, \hat{\theta}_{\alpha,\beta}) &= \mathsf{E}_\theta\left[\left(\hat{\theta}_{\alpha,\beta}(X) - \theta\right)^2\right] = \mathsf{Var}_\theta[\hat{\theta}_{\alpha,\beta}(X)] + (\mathsf{Bias}_\theta[\hat{\theta}_{\alpha,\beta}(X)])^2\\ &= \mathsf{Var}_\theta\left[\frac{X}{n + \alpha + \beta}\right] + \left(\mathsf{E}\left[\frac{X + \alpha}{n + \alpha + \beta}\right] - \theta\right)^2\\ &= \frac{n\theta(1 - \theta)}{(n + \alpha + \beta)^2} + \left(\frac{n\theta + \alpha}{n + \alpha + \beta} - \theta\right)^2.\end{aligned}$$

Simplifying the bias term:

$$\frac{n\theta + \alpha - \theta(n + \alpha + \beta)}{n + \alpha + \beta} = \frac{\alpha - \theta(\alpha + \beta)}{n + \alpha + \beta}.$$

Grouping terms by powers of θ, the risk becomes:

$$R(\theta, \hat{\theta}_{\alpha,\beta}) = \frac{[(\alpha + \beta)^2 - n]\theta^2 + [n - 2\alpha(\alpha + \beta)]\theta + \alpha^2}{(n + \alpha + \beta)^2}.$$

Enforcing Constant Risk For the risk to be constant (independent of θ), the coefficients of θ^2 and θ must be zero:

1. Coefficient of θ^2: $(\alpha + \beta)^2 - n = 0 \implies \alpha + \beta = \sqrt{n}$.
2. Coefficient of θ: $n - 2\alpha(\alpha + \beta) = 0 \implies n - 2\alpha\sqrt{n} = 0 \implies \alpha = \sqrt{n}/2$.

Since $\alpha + \beta = \sqrt{n}$, it follows that $\beta = \sqrt{n}/2$.

Conclusion The estimator corresponding to prior parameters $\alpha = \beta = \frac{\sqrt{n}}{2}$ is:

$$\hat{\theta}_{\sqrt{n}/2,\sqrt{n}/2}(X) = \frac{X + \sqrt{n}/2}{n + \sqrt{n}}.$$

This estimator is Bayes and has constant risk. Therefore, by Lemma 7.1, it is minimax.

The constant risk value is:

$$R(\hat{\theta}_{\text{minimax}}) = \frac{\alpha^2}{(n + \alpha + \beta)^2} = \frac{n/4}{(n + \sqrt{n})^2} = \frac{1}{4(\sqrt{n} + 1)^2}.$$

This is strictly smaller than the worst–case risk of the MLE ($1/4n$).

7.5 Least Favorable Distributions

In Example 7.7, we derived the Bayes rule $\hat{\theta}_{\alpha,\beta}(X)$ for the Binomial model. We found that for the specific choice $\alpha = \beta = \frac{\sqrt{n}}{2}$, the risk of the Bayes rule was constant. By Lemma 7.1, this implied that the estimator was minimax.

To apply this strategy generally for determining minimax estimators, we must ask:

For which prior distribution is the Bayes rule likely to be minimax?

A minimax decision rule attempts to minimize the risk in the *worst–case* scenario. Intuitively, we might expect the minimax rule to be the Bayes rule corresponding to the "worst possible" prior distribution—the one that makes the average risk as high as possible. We characterize this distribution as being **least favorable**.

Notation. *We use d_Π to denote a Bayes rule constructed under a given prior distribution Π.*

Definition 7.8. A prior distribution Π is **least favorable** if

$$r(\Pi, d_\Pi) \geq r(\Pi', d_{\Pi'}) \quad \text{for all prior distributions } \Pi'.$$

In other words, the least favorable prior maximizes the minimal achievable Bayes risk.

Example 7.8 (Binomial Model)**.** Let $X \sim \text{Binomial}(n, \theta)$ with $\theta \in [0, 1]$. The prior Beta $\left(\frac{\sqrt{n}}{2}, \frac{\sqrt{n}}{2}\right)$ is least favorable under squared error loss. This prior puts significant mass where estimation is "hardest" in an average sense to maximize the Bayes risk.

The connection between minimaxity and least favorable priors is formalized in the following lemma.

Lemma 7.2. *Let d_Π be a (unique) Bayes rule for prior Π. If*

$$r(\Pi, d_\Pi) = \sup_{\theta \in \Theta} R(\theta, d_\Pi),$$

then:

1. d_Π *is (unique) minimax.*

2. Π *is a least favorable prior.*

Note. The assumption $r(\Pi, d_\Pi) = \sup_\theta R(\theta, d_\Pi)$ is weaker than requiring constant risk. For example, Π could be a discrete distribution that puts all its mass on the specific values of θ where the risk $R(\theta, d_\Pi)$ is maximal.

Proof. **1. Minimaxity of d_Π**
This follows the same argument as Lemma 7.1 (ii). For any competitor rule d':

$$\begin{aligned}\sup_\theta R(\theta, d') &\geqslant \int_\Theta R(\theta, d')\,\mathrm{d}\Pi(\theta) = r(\Pi, d') \\ &\overset{d_\Pi \text{ Bayes}}{\geqslant} r(\Pi, d_\Pi) \\ &= \sup_\theta R(\theta, d_\Pi) \quad \text{(by assumption).}\end{aligned}$$

Thus, d_Π has the smallest worst–case risk. If d_Π is the unique Bayes rule, strict inequality holds for $d' \neq d_\Pi$ (on a set of positive measure), implying uniqueness.

2. Π is Least Favorable
Let $\widetilde{\Pi}$ be any other prior distribution, and let $d_{\widetilde{\Pi}}$ be its corresponding Bayes rule. By definition of the Bayes risk for $\widetilde{\Pi}$:

$$r(\widetilde{\Pi}, d_{\widetilde{\Pi}}) = \inf_{d'} r(\widetilde{\Pi}, d') \leq r(\widetilde{\Pi}, d_\Pi).$$

Now, expand the risk of d_Π under $\widetilde{\Pi}$:

$$r(\widetilde{\Pi}, d_\Pi) = \int_\Theta R(\theta, d_\Pi)\,\mathrm{d}\widetilde{\Pi}(\theta).$$

Since the average of a function cannot exceed its supremum, we have:

$$\int_\Theta R(\theta, d_\Pi)\,\mathrm{d}\widetilde{\Pi}(\theta) \leqslant \sup_\theta R(\theta, d_\Pi).$$

Using the assumption that $\sup_\theta R(\theta, d_\Pi) = r(\Pi, d_\Pi)$, we combine the inequalities:

$$r(\widetilde{\Pi}, d_{\widetilde{\Pi}}) \leq r(\widetilde{\Pi}, d_\Pi) \leq \sup_\theta R(\theta, d_\Pi) = r(\Pi, d_\Pi).$$

Therefore, $r(\Pi, d_\Pi) \geq r(\widetilde{\Pi}, d_{\widetilde{\Pi}})$, proving that Π is least favorable. □

Remark.

- **Non-uniqueness of Priors:** A least favorable distribution is not unique in general. In the Binomial example, the Bayes estimator (posterior mean) depends only on the first $n+1$ moments of θ under Π.

  $$d_\Pi(x) = \mathsf{E}[\theta|x] = \frac{\int \theta^{x+1}(1-\theta)^{n-x}\,\mathrm{d}\Pi(\theta)}{\int \theta^{x}(1-\theta)^{n-x}\,\mathrm{d}\Pi(\theta)}.$$

  Different priors sharing these moments will yield the same estimator and same constant risk, thus qualifying as least favorable.

- **Dependence on Loss Function:** Minimax rules are highly sensitive to the chosen loss function.
  - Under **Squared Error Loss** $L(\theta, a) = (\theta - a)^2$, the standard MLE $\frac{X}{n}$ is **not** minimax.
  - However, consider the weighted loss function:
  $$L(\theta, a) = \frac{(\theta - a)^2}{\theta(1-\theta)}.$$
  This loss penalizes errors near the boundaries (where variance is usually low) more heavily. Under this specific loss, the risk of $\frac{X}{n}$ becomes constant. Since $\frac{X}{n}$ is Bayes for the Uniform prior (Beta(1,1)), it implies that $\frac{X}{n}$ **is** a minimax estimator for this weighted loss.

7.6 Extended Bayes

A least favorable distribution does not always exist.

Consider estimating the mean θ of a normal distribution $\mathcal{N}(\theta, 1)$ where $\theta \in \mathbb{R}$. Intuitively, no value of θ is harder to estimate than another, so a least favorable prior should be uniform over $\mathbb{R}$. However, a uniform distribution on the entire real line is improper (unbounded measure).

To handle such cases, we generalize the concept of a least favorable distribution to a *sequence* of prior distributions.

Definition 7.9. A decision rule d is **extended Bayes** if there exists a sequence of prior distributions $(\Pi_m)_{m=1}^{\infty}$ such that

$$\lim_{m\to\infty} \left(r(\Pi_m, d) - \inf_{d'} r(\Pi_m, d') \right) = 0.$$

In simple terms, d acts like a limit of Bayes rules; the gap between its performance and the optimal Bayes performance vanishes along the sequence.

Remark. Extended Bayes rules are connected to least favorable sequences of prior distributions. Let $(\Pi_m)_{m=1}^{\infty}$ be a sequence of priors with minimal average risks $r_{\Pi_m} = \inf_{d'} r(\Pi_m, d')$. Then, $(\Pi_m)_{m=1}^{\infty}$ is a least favorable sequence of priors if $\lim\limits_{m\to\infty} r_{\Pi_m} =: r < \infty$ and

$$r \geq r(\Pi', d_{\Pi'})$$

for any prior Π'.

Lemma 7.3. *Suppose a decision rule d has constant risk $R(d)$. If d is extended Bayes, then d is minimax.*

Proof. By the definition of extended Bayes, for any $\epsilon > 0$, there exists an m such that:

$$r(\Pi_m, d) \leq \inf_{d'} r(\Pi_m, d') + \epsilon.$$

Let d' be any competitor rule. Since d has constant risk, its average risk under any prior is just that constant value: $r(\Pi_m, d) = R(d)$.

Using the fact that the average risk of d' is bounded by its worst-case risk ($r(\Pi_m, d') \leq \sup_\theta R(\theta, d')$), we have:

$$\begin{aligned} R(d) = \sup_\theta R(\theta, d) = r(\Pi_m, d) &\leq r(\Pi_m, d') + \epsilon \\ &\leq \sup_\theta R(\theta, d') + \epsilon. \end{aligned}$$

Since this holds for any $\epsilon > 0$, letting $\epsilon \to 0$ yields $R(d) = \sup\limits_\theta R(\theta, d) \leq \sup_\theta R(\theta, d')$. Thus, d is minimax. □

Example 7.9 (Multivariate Gaussian Means)**.** Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. random vectors in $\mathbb{R}^p$, distributed as $\mathcal{N}_p(\boldsymbol{\theta}, \sigma^2\mathbf{I})$ with $\sigma^2 > 0$ known.

So here we have multivariate observations $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip})^T$ where all X_{ij} are independent with $\mathsf{E}\left[X_{ij}\right] = \theta_j$ and $\mathsf{Var}\left[X_{ij}\right] = \sigma^2$.

We wish to estimate the vector $\boldsymbol{\theta}$ under the squared Euclidean error loss:

$$L(\boldsymbol{\theta}, \mathbf{a}) = \|\boldsymbol{\theta} - \mathbf{a}\|^2 = \sum_{j=1}^{p} (\theta_j - a_j)^2.$$

Claim: The sample mean vector $\overline{\mathbf{X}}_n$ is minimax.

Step 1: Verify Constant Risk

The risk of $\overline{\mathbf{X}}_n$ is the sum of the variances of its components:

$$R(\boldsymbol{\theta}, \overline{\mathbf{X}}_n) = \mathsf{E}_{X\sim\mathsf{P}_\theta}\left[\sum_{j=1}^{p} (\overline{X}_{j,n} - \theta_j)^2\right] = \sum_{j=1}^{p} \mathsf{Var}(\overline{X}_{j,n}) = \sum_{j=1}^{p} \frac{\sigma^2}{n} = \frac{p\sigma^2}{n}.$$

This risk is independent of $\boldsymbol{\theta}$, so it is constant.

Step 2: Show it is Extended Bayes

We construct a sequence of priors $\Pi_m = \mathcal{N}_p(\mathbf{0}, \tau_m^2\mathbf{I})$ where $\tau_m^2 \to \infty$. These priors become "flatter" and approach a uniform distribution over $\mathbb{R}^p$.

The Bayes rule for Π_m is the posterior mean (a shrinkage estimator):

$$d_{\Pi_m}(\mathbf{X}) = \mathsf{E}_{X\sim\mathsf{P}_\theta}\left[\boldsymbol{\theta}|\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\right] = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau_m^2}\overline{\mathbf{X}}_n.$$

The Bayes risk (minimum average risk) for Π_m is the sum of the posterior variances. Since the posterior covariance matrix is diagonal with entries $(n/\sigma^2 + 1/\tau_m^2)^{-1}$, the Bayes risk is, using that $\mathsf{Var}_{\Pi_m}\left[\theta_j|\boldsymbol{X}_1, \dots, \boldsymbol{X}_n\right] = \left(\frac{n}{\sigma^2} + \frac{1}{\tau_m^2}\right)^{-1}$ does not depend on the data

$$\begin{aligned}
r(\Pi_m, d_{\Pi_m}) &= \mathsf{E}_{\theta\sim\Pi_m}\Big[R(\theta, d_{\Pi_m}(\mathbf{X}_1,\ldots,\mathbf{X}_n))\Big] \\
&= \mathsf{E}_{\theta\sim\Pi_m}\Big[\mathsf{E}_{\mathbf{X}\sim\mathsf{P}_\theta}\big[\|d_{\Pi_m}(\mathbf{X}_1,\ldots,\mathbf{X}_n)-\theta\|^2\big]\Big] \\
&\overset{\text{(Fubini)}}{=} \mathsf{E}_{\mathbf{X}\sim\mathsf{P}_\Pi}\Big[\mathsf{E}_{\theta\sim\Pi_m(\cdot|\mathbf{X})}\big[\|d_{\Pi_m}(\mathbf{X}_1,\ldots,\mathbf{X}_n)-\theta\|^2 \,\big|\, \mathbf{X}_1,\ldots,\mathbf{X}_n\big]\Big] \\
&= \mathsf{E}_{\mathbf{X}\sim\mathsf{P}_\Pi}\Big[\mathsf{E}_{\theta\sim\Pi_m(\cdot|\mathbf{X})}\big[\|\theta-\mathsf{E}[\,\theta\,|\,\mathbf{X}_1,\ldots,\mathbf{X}_n]\|^2 \,\big|\, \mathbf{X}_1,\ldots,\mathbf{X}_n\big]\Big] \\
&= \mathsf{E}_{\mathbf{X}\sim\mathsf{P}_\Pi}\left[p\cdot\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_m^2}}\right] = p\cdot\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_m^2}}
\end{aligned}$$

As $m\to\infty$ (so $\tau_m\to\infty$), the term $1/\tau_m^2\to 0$. Thus:

$$\lim_{m\to\infty} r(\Pi_m, d_{\Pi_m}) = p\frac{\sigma^2}{n}.$$

Since the minimum Bayes risk converges to the constant risk of $\overline{\mathbf{X}}_n$, the sample mean is extended Bayes.

$$\begin{aligned}
r(\Pi_m, \overline{X}_n) = \mathsf{E}_{\theta\sim\Pi_m}\big[\,R(\theta,\overline{X}_n)\,\big] &= \mathsf{E}_{\theta\sim\Pi_m}\big[\mathsf{E}_{X\sim\mathsf{P}_\theta}\big[\|\overline{X}_n-\theta\|^2\big]\big] \\
&= \mathsf{E}_{\theta\sim\Pi_m}\left[p\frac{\sigma^2}{n}\right] = p\,\frac{\sigma^2}{n}.
\end{aligned}$$

By Lemma 7.3, $\overline{\mathbf{X}}_n$ is minimax.

Example 7.10 (Gaussian Model with unknown variance)**.** Suppose $X_1,\ldots,X_n\sim\mathcal{N}(\mu,\sigma^2)$ where both $\mu\in\mathbb{R}$ and $\sigma^2\in(0,\infty)$ are unknown. If we estimate μ under standard squared error loss $(\mu-a)^2$:

- The risk of $\overline{X}_n$ is σ^2/n.
- Since σ^2 can be arbitrarily large, $\sup_{\mu,\sigma} R((\mu,\sigma),\overline{X}_n)=\infty$.
- In fact, the minimax risk for this problem is infinite; no estimator has a finite worst-case risk over the unbounded variance space.

Solutions:

(i) **Bounded Variance:** Assume $\sigma^2\le M$. The minimax analysis becomes feasible.

(ii) **Scaled Loss Function:** Is squared error loss reasonable? Change the loss function to penalize errors relative to the noise level:

$$L((\mu,\sigma),a)=\frac{(\mu-a)^2}{\sigma^2}.$$

Under this loss, the risk of $\overline{X}_n$ is $1/n$ (constant). It can be shown that $\overline{X}_n$ is minimax for this specific loss.

7.7 Admissibility

We have previously discussed strict requirements for estimators, such as being Minimax (best worst-case risk) or Bayes (best average risk). Admissibility is a "bare minimum" requirement: an estimator is admissible if no other estimator is uniformly strictly better.

This section connects these concepts, providing tools to prove that an estimator is admissible by showing it is Unique Bayes, Unique Minimax, or Extended Bayes.

7.7.1 Admissibility of Unique Bayes and Minimax Rules

It is often easier to prove that a rule is Unique Bayes or Unique Minimax than to prove admissibility directly. Lemma 7.4 provides the bridge.

Lemma 7.4.

(i) If d is unique minimax, then d is admissible.

(ii) If d is unique Bayes, then d is admissible.

Remark. Lemma 7.2 gives conditions for a rule to be unique minimax. See the exercises for conditions regarding unique Bayes rules.

Proof. The proof relies on contradiction. Suppose d is not admissible. Then there exists a rule d' that is strictly better than d. This means:

$$R(\theta, d') \leq R(\theta, d) \quad \forall \theta \in \Theta,$$

with strict inequality for at least one θ.

(i) **Minimax Case:** If $R(\theta, d') \leq R(\theta, d)$ for all θ, then the worst-case risks satisfy:

$$\sup_\theta R(\theta, d') \leq \sup_\theta R(\theta, d).$$

Since d is unique minimax, no other rule d' can match its worst-case risk unless $d' = d$ almost everywhere. If d' were strictly better, it would either have a lower max risk (contradicting d's optimality) or equal max risk (contradicting d's uniqueness).

(ii) **Bayes Case:** Similarly, integrating the risk inequality with respect to the prior Π:

$$r(\Pi, d') = \int R(\theta, d') \, \mathrm{d}\Pi(\theta) \leq \int R(\theta, d) \, \mathrm{d}\Pi(\theta) = r(\Pi, d).$$

Since d is unique Bayes, it is the *unique* minimizer of the Bayes risk. Therefore, d' cannot satisfy this inequality unless $d' = d$ almost surely. Thus, no strictly better rule d' can exist.

□

7.7.2 Admissibility of Bayes Rules

We can relax the "uniqueness" requirement if we introduce continuity assumptions on the risk function and the prior.

Theorem 7.2. *Consider a model $\{\mathsf{P}_\theta : \theta \in \Theta\}$ with an open parameter space $\Theta \subset \mathbb{R}^k$. Suppose that:*

(i) All decision rules d have a continuous risk function $\theta \mapsto R(\theta, d)$.

(ii) Π is a prior distribution with $\Pi(\mathcal{U}) > 0$ for all open subsets $\mathcal{U} \subseteq \Theta$. (The prior has "full support").

If d_Π is a Bayes rule for prior Π with finite Bayes risk $r(\Pi, d_\Pi) < \infty$, then d_Π is admissible.

Note. The continuity assumption (i) holds for exponential families and squared error loss.

Proof. Assume for the sake of contradiction that d' is strictly better than d_Π. This implies:

1. $R(\theta, d') \leq R(\theta, d_\Pi)$ for all θ.
2. There exists some $\theta_0 \in \Theta$ such that $R(\theta_0, d') < R(\theta_0, d_\Pi)$.

Let $\eta = R(\theta_0, d_\Pi) - R(\theta_0, d') > 0$ be the risk gap at θ_0.

By the continuity of the risk functions, this gap must persist in a neighborhood around θ_0. There exists an $\epsilon > 0$ such that for all θ with $\|\theta - \theta_0\| < \epsilon$:

$$R(\theta, d') \leq R(\theta, d_\Pi) - \frac{\eta}{2}.$$

Now, consider the Bayes risk of d':

$$\begin{aligned} r(\Pi, d') &= \int_\Theta R(\theta, d') \, \mathrm{d}\Pi(\theta) \\ &= \int_{\|\theta-\theta_0\|<\epsilon} R(\theta, d') \, \mathrm{d}\Pi(\theta) + \int_{\|\theta-\theta_0\|\geq\epsilon} R(\theta, d') \, \mathrm{d}\Pi(\theta). \end{aligned}$$

Substituting the bounds:

- Inside the ϵ-ball: $R(\theta, d') \leq R(\theta, d_\Pi) - \eta/2$.
- Outside the ϵ-ball: $R(\theta, d') \leq R(\theta, d_\Pi)$.

$$\begin{aligned} r(\Pi, d') &\leqslant \int_{\|\theta-\theta_0\|<\epsilon} \Big(R(\theta, d_\Pi) - \frac{\eta}{2}\Big) \, \mathrm{d}\Pi(\theta) + \int_{\|\theta-\theta_0\|\geq\epsilon} R(\theta, d_\Pi) \, \mathrm{d}\Pi(\theta) \\ &= \int_\Theta R(\theta, d_\Pi) \, \mathrm{d}\Pi(\theta) - \int_{\|\theta-\theta_0\|<\epsilon} \frac{\eta}{2} \, \mathrm{d}\Pi(\theta) \\ &= r(\Pi, d_\Pi) - \frac{\eta}{2} \cdot \Pi\big(\{\theta : \|\theta - \theta_0\| < \epsilon\}\big). \end{aligned}$$

By assumption (ii), the prior mass of the open ϵ-ball is strictly positive. Therefore:

$$r(\Pi, d') < r(\Pi, d_\Pi).$$

This contradicts the fact that d_Π is Bayes (minimizes the Bayes risk). Thus, d_Π must be admissible. □

Admissibility of Extended Bayes Rules

We can further relax the requirement to extended Bayes rules. However, we need a condition to ensure the prior mass on open sets does not vanish "too fast" relative to the convergence of the risk gap.

Theorem 7.3. *Consider a model $\{\mathsf{P}_\theta : \theta \in \Theta\}$ with open parameter space $\Theta \subset \mathbb{R}^k$. Suppose that:*

(i) All decision rules d' have continuous risk functions $\theta \mapsto R(\theta, d')$.

(ii) d is extended Bayes with respect to a sequence of priors $(\Pi_m)_{m=1}^\infty$ such that for all open subsets $\mathcal{U} \subseteq \Theta$:

$$\lim_{m\to\infty} \frac{r(\Pi_m, d) - \inf_{d'} r(\Pi_m, d')}{\Pi_m(\mathcal{U})} = 0.$$

Then d is admissible.

Proof. Assume there exists a rule $\tilde{d}$ that is strictly better than d. By the same continuity argument as Theorem 7.2, there exists an $\eta > 0$ and an open subset $\mathcal{U} \subseteq \Theta$ such that the risk gap on $\mathcal{U}$ is at least η. This implies:

$$r(\Pi_m, \tilde{d}) \leq r(\Pi_m, d) - \eta \cdot \Pi_m(\mathcal{U}) \quad \forall m \geq 1.$$

Rearranging this inequality:

$$r(\Pi_m, d) - r(\Pi_m, \tilde{d}) \geq \eta \cdot \Pi_m(\mathcal{U}).$$

Dividing by $\Pi_m(\mathcal{U})$ (assuming it is non-zero):

$$\frac{r(\Pi_m, d) - r(\Pi_m, \tilde{d})}{\Pi_m(\mathcal{U})} \geq \eta.$$

Since the optimal Bayes risk $\inf_{d'} r(\Pi_m, d')$ is certainly less than or equal to $r(\Pi_m, \tilde{d})$, we have:

$$\frac{r(\Pi_m, d) - \inf_{d'} r(\Pi_m, d')}{\Pi_m(\mathcal{U})} \geq \frac{r(\Pi_m, d) - r(\Pi_m, \tilde{d})}{\Pi_m(\mathcal{U})} \geq \eta.$$

This contradicts assumption (ii), which states that this ratio must converge to 0. Therefore, d is admissible. □

7.8 Sample and Shrinkage Estimators

Consider a single observation $X \sim \mathcal{N}(\theta, 1)$ with known variance. (Note: You can think of having $X_1, \ldots, X_n$ i.i.d. normal and reducing to the sufficient and complete statistic $X \equiv \overline{X}_n$ by the Rao-Blackwell Theorem).

We wish to estimate the mean $\theta \in \mathbb{R}$ under squared error loss $L(\theta, a) = (\theta - a)^2$.

Proposition 7.1. *The affine estimator $d_{a,b}(X) = aX + b$ with $a, b \in \mathbb{R}$ is admissible if and only if:*

i) $0 \leq a < 1$, *or*

ii) $a = 1$ *and* $b = 0$ *(so* $d_{a,b}(X) = X$*).*

Remark (Shrinkage Estimators). In case i) where $a \in [0, 1)$, the estimator can be rewritten as:

$$d_{a,b}(X) = aX + b = aX + (1-a)\frac{b}{1-a}.$$

This is a convex combination of the data X and a constant $\theta_0 = \frac{b}{1-a}$. This is known as a **shrinkage estimator** because it "shrinks" the observation towards the value θ_0.

Proof. First, we derive the risk function for any affine estimator $d_{a,b}(X)$.

$$\begin{aligned} R(\theta, d_{a,b}) &= \mathsf{Var}_\theta[aX + b] + (\mathsf{bias}_\theta(aX + b))^2 \\ &= a^2 \cdot \mathsf{Var}[X] + (a\theta + b - \theta)^2 \\ &= a^2 + (b - \theta(1-a))^2. \end{aligned}$$

1. Inadmissibility (Necessity)

We show that if the conditions are not met, the estimator is dominated.

- **Case** $a > 1$**:** The risk is $R(\theta, d_{a,b}) \geq a^2 > 1$. Since the standard estimator X (where $a = 1, b = 0$) has constant risk 1, X strictly dominates $d_{a,b}$.
- **Case** $a = 1, b \neq 0$**:** The risk is $R(\theta, d_{a,b}) = 1^2 + b^2 > 1$. Again, X strictly dominates $d_{a,b}$.
- **Case** $a < 0$**:** Here $1 - a > 1$. Compare $d_{a,b}$ to the constant estimator $\tilde{d}(x) \equiv \frac{b}{1-a}$. The risk of $d_{a,b}$ is:

  $$R(\theta, d_{a,b}) > (b - \theta(1-a))^2 = (1-a)^2 \left(\frac{b}{1-a} - \theta\right)^2.$$

  Since $(1-a)^2 > 1$, this is strictly larger than $(\frac{b}{1-a} - \theta)^2$, which is the risk of the constant estimator $\tilde{d}$. Thus, $d_{a,b}$ is inadmissible.

2. Admissibility (Sufficiency)

Case i): $0 \leq a < 1$

- If $a = 0$, then $d_{a,b}(X) = b$. This is a constant estimator, which is admissible (it is the unique Bayes rule for a point mass prior at b, or see Example 8.1).

- If $0 < a < 1$, we show $d_{a,b}$ is a Bayes estimator. Consider a normal prior $\Pi = \mathcal{N}(c, \tau^2)$. The Bayes rule is:

$$d_\Pi(X) = \frac{\tau^2}{\tau^2+1}X + c\frac{1}{\tau^2+1}.$$

We can match coefficients by setting $\frac{\tau^2}{\tau^2+1} = a$ (which implies $\tau^2 = \frac{a}{1-a}$) and choosing c such that the intercept matches b. Since $d_{a,b}$ is a Bayes estimator with finite risk on an open parameter space (with continuous risk), it is admissible by Theorem 7.2 .

Case ii): $a = 1, b = 0$ **(The Sample Mean)**

The estimator $d_{1,0}(X) = X$ is not Bayes for any proper prior (requires infinite variance). We prove admissibility using Theorem 7.3 by showing it is extended Bayes w.r.t. $(\Pi_m)_{m=1}^\infty$ for $\Pi_m = \mathcal{N}(0, \tau_m^2)$ with $\tau_m^2 \to \infty$; as we have shown when proving minimaxity of the Gaussian sample mean.

Specifically, we take $\tau_m^2 = m$.

Consider then the sequence of priors $\Pi_m = \mathcal{N}(0, m)$. The Bayes rule for Π_m is

$$d_{\Pi_m}(X) = \frac{m}{m+1}X.$$

The rule d_{Π_m} has risk

$$R(\theta, d_{\Pi_m}) = \left(\frac{m}{m+1}\right)^2 \underbrace{\mathsf{Var}_\theta\left[X\right]}_{=1} + \underbrace{\left(\frac{m}{m+1}\theta - \theta\right)^2}_{\text{bias}} = \frac{m^2+\theta^2}{(m+1)^2},$$

and, since $\mathsf{E}_{\Pi_m}\left[\theta^2\right] = \mathsf{Var}_{\Pi_m}\left[\theta\right] = m$, the Bayes risk for d_{Π_m} is:

$$r(\Pi_m, d_{\Pi_m}) = \frac{m^2 + \mathsf{E}_{\Pi_m}\left[\theta^2\right]}{(m+1)^2} = \frac{m^2+m}{(m+1)^2} = \frac{m}{m+1}.$$

The risk of our estimator $d_{1,0}(X) = X$ is constant $R(\theta, X) = \mathsf{Var}_\theta\left[X\right] = 1$, so its average risk is also $r(\Pi_m, X) = 1$ for all m.

We now need to show that

$$\frac{r(\Pi_m, d_{1,0}) - r(\Pi_m, d_{\Pi_m})}{\Pi_m(\mathcal{U})} \underset{m\to\infty}{\longrightarrow} 0 \quad \forall \mathcal{U} \subseteq \mathbb{R} \text{ open}.$$

The "risk gap" is:

$$r(\Pi_m, X) - r(\Pi_m, d_{\Pi_m}) = 1 - \frac{m}{m+1} = \frac{1}{m+1}.$$

To apply Theorem 7.3, we must show that for any open set $\mathcal{U} \subseteq \mathbb{R}$:

$$\lim_{m\to\infty} \frac{1/(m+1)}{\Pi_m(\mathcal{U})} = 0.$$

It suffices to consider an open interval $\mathcal{U} = (u, u+h)$ with $h > 0$.

Let $\phi(x)$ and $\Phi(x)$ be the density and d.f. of the standard normal distribution, respectively. Then the probability of this interval under $\Pi_m = \mathcal{N}(0, m)$ is:

$$\Pi_m((u, u+h)) = \Phi\left(\frac{u+h}{\sqrt{m}}\right) - \Phi\left(\frac{u}{\sqrt{m}}\right)$$

Using a Taylor expansion (or the Mean Value Theorem) around 0 as $m \to \infty$:

$$\Pi_m((u, u+h)) \approx \phi(0) \cdot \left(\frac{u+h}{\sqrt{m}} - \frac{u}{\sqrt{m}}\right) = \frac{1}{\sqrt{2\pi}} \frac{h}{\sqrt{m}}.$$

Thus, for large m, $\Pi_m(\mathcal{U})$ behaves like $C/\sqrt{m}$. Substituting this into the limit condition:

$$\frac{r(\Pi_m, X) - r(\Pi_m, d_{\Pi_m})}{\Pi_m(\mathcal{U})} \approx \frac{\frac{1}{m+1}}{\frac{C}{\sqrt{m}}} \approx \frac{\sqrt{m}}{m} = \frac{1}{\sqrt{m}} \xrightarrow{m \to \infty} 0.$$

Therefore, X is admissible. □

7.8.1 Estimation in Higher Dimensions

Suppose we have a multivariate observation $\mathbf{X} = (X_1, \dots, X_p)^\top \sim \mathcal{N}_p(\boldsymbol{\theta}, \mathbf{I})$ with $\boldsymbol{\theta} \in \mathbb{R}^p$. This implies that the components $X_1, \dots, X_p$ are independent with $X_j \sim \mathcal{N}(\theta_j, 1)$.

Consider the estimation of the mean vector $\boldsymbol{\theta}$ under the squared error loss:

$$L(\boldsymbol{\theta}, \mathbf{a}) = \|\boldsymbol{\theta} - \mathbf{a}\|^2 = \sum_{j=1}^{p} (\theta_j - a_j)^2.$$

The natural estimator (which is both the MLE and UMVUE) is simply the observation vector itself, $\mathbf{X}$. Its risk is the sum of the variances of its components:

$$R(\boldsymbol{\theta}, \mathbf{X}) = \mathsf{E}_{\boldsymbol{\theta}}\left[\|\mathbf{X} - \boldsymbol{\theta}\|^2\right] = \sum_{j=1}^{p} \mathsf{E}_{\theta_j}\left[(X_j - \theta_j)^2\right] = \sum_{j=1}^{p} \mathsf{Var}[X_j] = p.$$

We have previously shown that $\mathbf{X}$ is admissible if $p = 1$. It turns out that $\mathbf{X}$ is also admissible if $p = 2$ (see Lehmann and Casella 1998, Exercise 5.4.5).

Stein's Paradox The situation changes drastically in higher dimensions.

If $p \geq 3$, then $\mathbf{X}$ is inadmissable.

This result is known as **Stein's Paradox**. It was considered paradoxical because, for a long time, the Maximum Likelihood Estimator (MLE) was believed to be the optimal estimator in such standard settings. While $\mathbf{X}$ has desirable properties like unbiasedness, for $p \geq 3$ there exist estimators that strictly dominate it (i.e., have lower risk for all $\boldsymbol{\theta}$).

Example 7.11 (Batting averages in baseball)**.** To illustrate a scenario where independent parameters are estimated simultaneously, consider the 1990 batting averages of 18 Major League Baseball players. Let Z_i be the batting average of player i after n_i times at bat. We apply a variance-stabilizing transformation[1] to obtain X_i:

$$X_i = \sqrt{n_i} \arcsin(2Z_i - 1).$$

To a good approximation, we can model $X_i \sim \mathcal{N}(\theta_i, 1)$. We treat the true mean θ_i as the transformed career batting average π_i:

$$\theta_i = \sqrt{n_i} \arcsin(2\pi_i - 1).$$

Player	n_i	Z_i	π_i
Baines	415	0.284	0.289
Barfield	476	0.246	0.256
Bell	583	0.254	0.265
Biggio	555	0.276	0.287
Bonds	519	0.301	0.297
Bonilla	625	0.280	0.279
Brett	544	0.329	0.305
Brooks Jr.	568	0.266	0.269
Browne	513	0.267	0.271

Table 7.1: Batting averages for $p = 9$ baseball players (subset). n_i is times at bat, Z_i is 1990 average, π_i is career average. Taken from Samworth(2012)

In this example, even though the players' performances are independent, Stein's result suggests that estimating their averages jointly (using information from the whole group) can lead to a lower total squared error than estimating each one individually using only their own average X_i.

7.9 Inadmissibility of the MLE for $p \geq 3$

TODO: Improve this section.

7.9.1 Geometric Intuition

In lower dimensions ($p = 1, 2$), the Maximum Likelihood Estimator (MLE) $\boldsymbol{X}$ is admissible. However, in higher dimensions ($p \geq 3$), the squared length of the observation vector $\boldsymbol{X}$ tends to be significantly larger than the squared length of the true parameter $\boldsymbol{\theta}$.

Consider the expectation of the squared norm of $\boldsymbol{X}$:

$$\mathsf{E}_\theta[\|\boldsymbol{X}\|^2] = \mathsf{E}_\theta\left[\sum_{j=1}^{p} X_j^2\right] = \sum_{j=1}^{p} \mathsf{E}_\theta\left[X_j^2\right] = \sum_{j=1}^{p} (\theta_j^2 + 1) = \|\boldsymbol{\theta}\|^2 + p.$$

[1] This transformation improves the accuracy of the normal approximation for Binomial counts.

The term $+p$ represents a systematic bias in the length. As p grows, $\|\boldsymbol{X}\|^2$ becomes an increasingly poor estimate of $\|\boldsymbol{\theta}\|^2$ because the "noise" (variance) accumulates across all dimensions, pushing the vector $\boldsymbol{X}$ further away from the origin than $\boldsymbol{\theta}$ is.

Idea: Since $\boldsymbol{X}$ is "too long" on average, we should construct an estimator that shrinks $\boldsymbol{X}$ towards the origin (or another target). Consider the shrinkage estimator:

$$T^b(\boldsymbol{X}) := \left(1 - \frac{b}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X}, \quad \text{for } b > 0.$$

Here, the shrinkage factor $1-b/\|\boldsymbol{X}\|^2$ is data-dependent: we shrink more when $\|\boldsymbol{X}\|$ is small and less when it is large.

Theorem 7.4 (James and Stein, 1961)**.** *If $p \geq 3$ and $0 < b < 2(p-2)$, then for all $\boldsymbol{\theta} \in \mathbb{R}^p$, the risk under squared error loss is:*

$$R(\boldsymbol{\theta}, T^b) = p - b\Big(2(p-2) - b\Big) \cdot \mathsf{E}_{\boldsymbol{\theta}}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right].$$

Since the term subtracted is strictly positive, $R(\boldsymbol{\theta}, T^b) < p = R(\boldsymbol{\theta}, \boldsymbol{X})$ for all $\boldsymbol{\theta}$. Thus, $\boldsymbol{X}$ is inadmissible.

Proof (of Theorem 7.4). Let $p \geq 3$ and $0 < b < 2(p-2)$. Let $\boldsymbol{\theta} \in \mathbb{R}^p$. Define the helper function $f(\boldsymbol{X}) := \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|^2}$. We can write the estimator as $T^b(\boldsymbol{X}) = \boldsymbol{X} - bf(\boldsymbol{X})$.

The risk is the expected squared Euclidean distance:

$$\begin{aligned} R(\boldsymbol{\theta}, T^b) &= \mathsf{E}_{\boldsymbol{\theta}}\left[\|T^b(\boldsymbol{X}) - \boldsymbol{\theta}\|^2\right] \\ &= \mathsf{E}_{\boldsymbol{\theta}}\left[\|(\boldsymbol{X} - \boldsymbol{\theta}) - bf(\boldsymbol{X})\|^2\right] \\ &= \underbrace{\mathsf{E}_{\boldsymbol{\theta}}\left[\|\boldsymbol{X} - \boldsymbol{\theta}\|^2\right]}_{=p} + b^2\mathsf{E}_{\boldsymbol{\theta}}\left[\|f(\boldsymbol{X})\|^2\right] - 2b\mathsf{E}_{\boldsymbol{\theta}}\left[(\boldsymbol{X} - \boldsymbol{\theta})^\top f(\boldsymbol{X})\right]. \end{aligned}$$

Note that $\|f(\boldsymbol{X})\|^2 = \frac{\|\boldsymbol{X}\|^2}{\|\boldsymbol{X}\|^4} = \frac{1}{\|\boldsymbol{X}\|^2}$. Thus, the second term is $b^2\mathsf{E}_{\boldsymbol{\theta}}[\frac{1}{\|\boldsymbol{X}\|^2}]$.

The proof hinges on evaluating the cross term $\mathsf{E}_{\boldsymbol{\theta}}[(\boldsymbol{X} - \boldsymbol{\theta})^\top f(\boldsymbol{X})]$. We use **Stein's Lemma** (Integration by Parts), which states that for $X \sim \mathcal{N}(\theta, 1)$ and a differentiable function g, $\mathsf{E}[(X-\theta)g(X)] = \mathsf{E}[g'(X)]$.

Consider the j-th component of the dot product:

$$\mathsf{E}_{\boldsymbol{\theta}}\left[(X_j - \theta_j)\frac{X_j}{\|\boldsymbol{X}\|^2}\right].$$

Applying Stein's Lemma with respect to X_j:

$$\mathsf{E}_{\boldsymbol{\theta}}\left[(X_j - \theta_j)\frac{X_j}{\|\boldsymbol{X}\|^2}\right] = \mathsf{E}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial X_j}\left(\frac{X_j}{\|\boldsymbol{X}\|^2}\right)\right].$$

We calculate the partial derivative using the quotient rule (noting $\frac{\partial}{\partial X_j}\|\boldsymbol{X}\|^2 = 2X_j$):

$$\frac{\partial}{\partial X_j}\left(\frac{X_j}{\|\boldsymbol{X}\|^2}\right) = \frac{1 \cdot \|\boldsymbol{X}\|^2 - X_j \cdot 2X_j}{(\|\boldsymbol{X}\|^2)^2} = \frac{\|\boldsymbol{X}\|^2 - 2X_j^2}{\|\boldsymbol{X}\|^4} = \frac{1}{\|\boldsymbol{X}\|^2} - \frac{2X_j^2}{\|\boldsymbol{X}\|^4}.$$

Now, sum over all indices $j = 1, \dots, p$:

$$\begin{aligned}
\mathsf{E}_{\boldsymbol{\theta}}\big[(\boldsymbol{X} - \boldsymbol{\theta})^\top f(\boldsymbol{X})\big] &= \sum_{j=1}^{p} \mathsf{E}_{\boldsymbol{\theta}}\left[\frac{1}{\|\boldsymbol{X}\|^2} - \frac{2X_j^2}{\|\boldsymbol{X}\|^4}\right] \\
&= \mathsf{E}_{\boldsymbol{\theta}}\left[\sum_{j=1}^{p} \frac{1}{\|\boldsymbol{X}\|^2} - \frac{2\sum X_j^2}{\|\boldsymbol{X}\|^4}\right] \\
&= \mathsf{E}_{\boldsymbol{\theta}}\left[\frac{p}{\|\boldsymbol{X}\|^2} - \frac{2\|\boldsymbol{X}\|^2}{\|\boldsymbol{X}\|^4}\right] \\
&= \mathsf{E}_{\boldsymbol{\theta}}\left[\frac{p-2}{\|\boldsymbol{X}\|^2}\right] = (p-2)\mathsf{E}_{\boldsymbol{\theta}}\left[\frac{1}{\|\boldsymbol{X}\|^2}\right].
\end{aligned}$$

□

Comments on the Estimator

- **Optimal Shrinkage:** The quadratic term $b(2(p-2)-b)$ is maximized at $b = p-2$. This yields the standard **James-Stein Estimator**:

  $$d_{JS}(\boldsymbol{X}) = \left(1 - \frac{p-2}{\|\boldsymbol{X}\|^2}\right)\boldsymbol{X}.$$

- **Risk at the Origin:** The improvement in risk is maximal at $\boldsymbol{\theta} = 0$. At this point, $\|\boldsymbol{X}\|^2 \sim \chi_p^2$, and it can be shown that $\mathsf{E}_0[1/\|\boldsymbol{X}\|^2] = 1/(p-2)$. Thus:

  $$R(0, d_{JS}) = p - (p-2)^2 \cdot \frac{1}{p-2} = p - (p-2) = 2.$$

  Comparing this to $R(0, \boldsymbol{X}) = p$, the improvement is substantial (e.g., risk of 2 vs 18 for $p = 18$).

- **Positive-Part Estimator:** The factor $(1 - \frac{p-2}{\|\boldsymbol{X}\|^2})$ can be negative if the observed $\boldsymbol{X}$ is very close to zero. Shrinking "past" zero reverses the sign, which is geometrically nonsensical and increases error. The **Positive-Part James-Stein Estimator** fixes this:

  $$d_{JS}^{+}(\boldsymbol{X}) = \left(1 - \frac{p-2}{\|\boldsymbol{X}\|^2}\right)_+ \boldsymbol{X}, \qquad (x)_+ := \max\{x, 0\}.$$

  This estimator strictly dominates the standard James-Stein estimator, though it is still inadmissible (Shao and Strawderman, 1994), because it is not smooth (non-differentiable at the cut-off).

Choice of Shrinkage Target

Why shrink to zero? The choice of the origin is arbitrary.

1. **Arbitrary Target $\boldsymbol{\theta}_0$:** We can shrink towards any vector $\boldsymbol{\theta}_0$ (e.g., a theoretical prediction):

$$d_{JS}(\boldsymbol{X};\boldsymbol{\theta}_0) = \boldsymbol{\theta}_0 + \left(1 - \frac{p-2}{\|\boldsymbol{X} - \boldsymbol{\theta}_0\|^2}\right)(\boldsymbol{X} - \boldsymbol{\theta}_0).$$

2. **Shrinkage to the Mean (Lindley's Estimator):** In many problems (like the baseball example below), the components θ_i are conceptually similar. It is reasonable to assume they cluster around a common mean. We can shrink towards the *grand mean* $\overline{X} = \frac{1}{p}\sum X_j$:

$$d_j(\boldsymbol{X}) = \overline{X} + \left(1 - \frac{p-3}{\sum_{j=1}^{p}(X_j - \overline{X})^2}\right)(X_j - \overline{X}).$$

This is highly effective if the true parameters θ_i are close to each other.

Example 7.12 (Baseball Batting Averages)**.** We analyze the 1990 batting averages of 18 MLB players ($p = 18$). Let n_i be the number of at-bats and Z_i be the batting average. We apply a variance-stabilizing transformation so that the data is approximately normal with variance 1:

$$X_i = \sqrt{n_i}\arcsin(2Z_i - 1).$$

We compare three estimators:

1. **Naive MLE:** X_i (using only individual data).
2. **JS to Fixed Target:** Shrinking to $\pi_0 = 0.275$ (a reasonable global average).
3. **JS to Mean:** Shrinking to the observed average of the 18 players.

```r
library(dplyr)
library(readr)

# Load and transform data
baseball <- read_csv("data/baseball.txt") %>%
  mutate(
    X = sqrt(n_i) * asin(2 * Z_i - 1),
    theta = sqrt(n_i) * asin(2 * pi_i - 1) # 'Truth' based on career average
  )
p <- nrow(baseball) # p = 18

# 1. JS Estimator shrinking towards fixed target (pi_0 = 0.275)
pi_0 <- 0.275
t0 <- sqrt(mean(baseball$n_i)) * asin(2 * pi_0 - 1)

```

```r
baseball <- baseball %>%
  mutate(
    JS_t0 = t0 + max(1 - (p - 2) / sum((X - t0)^2), 0) * (X - t0)
  )

# 2. JS Estimator shrinking towards the group mean
m <- mean(baseball$X)

baseball <- baseball %>%
  mutate(
    JS_mean = m + max(1 - (p - 3) / sum((X - m)^2), 0) * (X - m)
  )

# Calculate Total Squared Errors (Risk)
risk_results <- baseball %>%
  summarize(
    MSE_MLE = sum((X - theta)^2),
    MSE_JS_t0 = sum((JS_t0 - theta)^2),
    MSE_JS_mean = sum((JS_mean - theta)^2)
  )

print(risk_results)
```

R Code Implementation

Results The James-Stein estimator shrinking towards the group mean performs best. By “borrowing strength” from the other players, we reduce the noise inherent in individual observation.

The reduction from 2.56 (MLE) to 1.17 (JS-Mean) represents a variance reduction of over 50%, confirming the inadmissibility of the standard estimator in practice.

8. Hypothesis Tests

8.1 Testing Problems

Observation: X, values in sample space $\mathcal{X} \subset \mathbb{R}^d$

Model: $\{P_\theta : \theta \in \Theta\}$ with parameter space $\Theta \subset \mathbb{R}^k$.

A **testing problem** amounts to deciding between two competing hypotheses regarding the true parameter θ:

$$H_0 : X \sim P_\theta \text{ with } \theta \in \Theta_0 \quad \text{versus} \quad H_1 : X \sim P_\theta \text{ with } \theta \in \Theta_1,$$

where Θ_0 and Θ_1 form a partition of the parameter space Θ (i.e., they are disjoint).

The Action Space and Asymmetry

The decision problem involves an action space $\mathcal{A} = \{0, 1\}$. While formally symmetric, the Neyman–Pearson theory (which we will cover later) adopts an **asymmetric view**:

- $\boldsymbol{H_0}$ (**Null Hypothesis**): Represents the default scenario, status quo, or "no effect."
- $\boldsymbol{H_1}$ (**Alternative Hypothesis**): Represents a discovery, a new effect, or a deviation from the norm.

Consequently, the actions are interpreted as:

- $a = 0$: **Do not reject** H_0**.**
    - This is *not* a confirmation that H_0 is true.
    - It is interpreted as a "missed opportunity" to find a signal, which is generally considered less severe than a false discovery.
- $a = 1$: **Reject** H_0 (Accept H_1).
    - This is a claim that the default assumption is false.
    - Making this claim erroneously is considered "something bad" (a false positive) and is strictly controlled.

Remark ("p–value crisis" and Reproducibility)**.** In many scientific fields (e.g., medicine, psychology), there is a strong culture of supporting claims with statistical significance (p–values).

- A rejection of H_0 (e.g., "The drug does nothing") is often required to publish results.
- **Effect Size vs. Significance:** A significant result (tiny p–value) only tells us that H_0 is false; it does not tell us if the effect is *meaningful.* A drug might statistically extend life by 1 second (rejecting "no effect"), but this is clinically irrelevant. Modern statistics emphasizes reporting **confidence intervals** and **effect sizes** alongside tests to capture the magnitude of the benefit.
- **Replicability:** If a researcher repeats an experiment many times and only reports the one successful test (hiding the failures), the statistical guarantees of the test are invalidated. This "file drawer problem" contributes to the replication crisis in science.

Example 8.1 (Television ads and NBC Guidelines)**.** The following requirements for television advertisements (based on historical NBC guidelines) illustrate how testing formulations change based on the claim being made.

a) **Superiority Claim ("Product A is better than B")** To air a commercial claiming superiority, NBC requires a study with $n \geq 300$ participants. Let $X \sim \text{Binomial}(n, \theta)$ be the count of participants preferring A. The null hypothesis must represent the opposite of the claim (i.e., A is not better).

$$H_0 : \theta \leq 0.5 \quad \text{vs.} \quad H_1 : \theta > 0.5$$

The claim is allowed only if the test rejects H_0 at a 95% confidence level.

b) **Parity Claim ("Product A is as good as B")** To claim parity, the requirements are stricter ($n \geq 500$). If the observed preference rate $X/n < 0.5$, the claim is allowed only if we **fail to reject** the hypothesis of equality:

$$H_0 : \theta = 0.5 \quad \text{vs.} \quad H_1 : \theta \neq 0.5$$

at a 90% confidence level.

Discussion on "Gaming" the System: In practice, a company might conduct a study for Scenario A and fail to reject H_0. If they simply discard that study and repeat the experiment until they get a "lucky" rejection (without telling the network about the failures), they are exploiting random chance. This highlights the importance of pre-registration in clinical trials (e.g., FDA requirements for vaccines) to prevent such "p–hacking."

8.1.1 Tests and Their Errors

Definition 8.1 (Non-randomized test)**.** A **non-randomized test** is a decision rule $d : \mathcal{X} \to \{0, 1\}$.

- The value $d(x) = 1$ corresponds to rejecting H_0 (accepting H_1).
- The value $d(x) = 0$ corresponds to accepting H_0 (rejecting H_1).

The **rejection region** is the set of observations leading to rejection:

$$S_1 = \{x \in \mathcal{X} : d(x) = 1\}.$$

The **acceptance region** is its complement $\mathcal{X} \setminus S_1$.

Type I and Type II Errors

In this binary decision framework, there are two specific ways to make an incorrect decision. While other fields use descriptive terms like "False Positive" (Sensitivity/Specificity), statistics traditionally uses the labels Type I and Type II.

- **Type I error (False Positive):** Rejecting H_0 when it is actually true ($\theta \in \Theta_0$).
- **Type II error (False Negative):** Accepting H_0 when it is actually false ($\theta \in \Theta_1$).

Randomized Tests

It is often mathematically convenient to allow **randomized decision rules**. In practice, you typically want a clear "Yes/No" answer, but allowing a test to output a *probability* of rejection simplifies the theoretical search for optimal tests (see convexity below).

Definition 8.2 (Randomized Test)**.**

(i) A **randomized decision rule** maps observed values x to probability distributions on the action space $\mathcal{A}$.

(ii) A **randomized test** is a function $d : \mathcal{X} \to [0, 1]$, where $d(x)$ represents the probability of rejecting H_0 given observation x.

In the sequel, the term "test" will generally refer to randomized tests, with non-randomized tests being the special case where $d(x) \in \{0, 1\}$.

Remark (Convexity). The set of randomized tests is **convex**. If d_1 and d_2 are tests (mapping into $[0, 1]$) and $\alpha \in [0, 1]$, then the convex combination

$$d_{\text{new}}(x) = \alpha d_1(x) + (1 - \alpha) d_2(x)$$

is also a valid test, as the output remains in $[0, 1]$. This property is crucial for optimization theory in hypothesis testing.

8.1.2 Neyman–Pearson Loss and Risk of Tests

To apply decision theory to hypothesis testing, we define a loss function that penalizes incorrect decisions. The **Neyman–Pearson (NP) loss** is the natural choice for the binary action space $\mathcal{A} = \{0, 1\}$. It focuses solely on whether the decision matches the truth state of θ.

The Loss Function

Let $c_1 > 0$ be the cost of a Type I error and $c_0 > 0$ be the cost of a Type II error.

Definition 8.3 (Neyman-Pearson Loss)**.** The loss function $L(\theta, a)$ for a deterministic action $a \in \{0, 1\}$ is defined as:

State of Nature $(L(\theta, a))$	**Accept** H_0 $(a = 0)$	**Reject** H_0 $(a = 1)$
H_0 True $(\theta \in \Theta_0)$	0	c_1
H_1 True $(\theta \in \Theta_1)$	c_0	0
Indifferent $(\theta \notin \Theta_0 \cup \Theta_1)$	0	0

Extension to Randomized Tests As discussed previously, a randomized test outputs a probability $p = d(x) \in [0, 1]$ rather than a strict action. To define the loss for a probability p, we consider the **expected loss** over the internal randomization of the test.

Let $A \sim \text{Bernoulli}(p)$ be the random action generated by the test. The extended loss function $L : \Theta \times [0, 1] \to \mathbb{R}$ is:

$$L(\theta, p) = \mathsf{E}_{A\sim\text{Bernoulli}(p)}[L(\theta, A)] = pL(\theta, 1) + (1 - p)L(\theta, 0).$$

Substituting the values from the NP table, this simplifies to a linear function of p:

$$L(\theta, p) = \begin{cases} p \cdot c_1 & \text{if } \theta \in \Theta_0 \\ (1 - p) \cdot c_0 & \text{if } \theta \in \Theta_1 \end{cases}$$

Intuitively, if the Null is true ($\theta \in \Theta_0$), the "mistake" is rejecting it (action 1). You pay the cost c_1 weighted by the probability p that you actually make that mistake.

The Risk Function

The risk $R(\theta, d)$ is the expected loss over the randomness of the data X.

$$R(\theta, d) = \mathsf{E}_\theta\big[L(\theta, d(X))\big].$$

Using the linearity of the expectation, we can express the risk directly in terms of the test's behavior.

For a Non-Randomized Test ($d(X) \in \{0, 1\}$):

$$R(\theta, d) = \begin{cases} c_1 \cdot \mathsf{P}_\theta(d(X) = 1) & \text{if } \theta \in \Theta_0, \\ c_0 \cdot \mathsf{P}_\theta(d(X) = 0) & \text{if } \theta \in \Theta_1. \end{cases}$$

For a Randomized Test ($d(X) \in [0, 1]$): Since $\mathsf{E}_\theta[d(X)]$ represents the expected probability of rejection:

$$R(\theta, d) = \begin{cases} c_1 \cdot \mathsf{E}_\theta[d(X)] & \text{if } \theta \in \Theta_0, \\ c_0 \cdot (1 - \mathsf{E}_\theta[d(X)]) & \text{if } \theta \in \Theta_1. \end{cases}$$

The Power Function

Notice that in both cases above, the risk is entirely determined by a single quantity: the expected value of the test function.

Definition 8.4 (Power Function)**.** The **power function** of a test d, denoted $\beta_d(\theta)$ (or simply $\beta(\theta)$), is the probability (or expected probability) of rejecting the null hypothesis as a function of θ:

$$\beta_d(\theta) := \mathsf{E}_\theta[d(X)].$$

This definition unifies both cases:

- If d is non–randomized, $\mathsf{E}_\theta[d(X)] = 1 \cdot \mathsf{P}(d(X) = 1) + 0 = \mathsf{P}_\theta(\text{Reject } H_0)$.
- If d is randomized, it is the average rejection probability.

Relationship between Risk and Power: We can now rewrite the risk compactly using the power function:

$$R(\theta, d) = \begin{cases} c_1\beta(\theta) & \text{if } \theta \in \Theta_0 \quad \text{(Type I Error Risk)} \\ c_0(1 - \beta(\theta)) & \text{if } \theta \in \Theta_1 \quad \text{(Type II Error Risk)} \end{cases}$$

Thus, characterizing the performance of a test is equivalent to studying its power function. Ideally, we want $\beta(\theta) \approx 0$ for $\theta \in \Theta_0$ and $\beta(\theta) \approx 1$ for $\theta \in \Theta_1$.

8.1.3 Tests in Neyman-Pearson Form

In practice, we rarely construct arbitrary randomized mappings from $\mathcal{X}$ to $[0, 1]$. Instead, nearly all practical tests follow the specific **Neyman-Pearson form** (or "0-1 form"). This structure reduces the dimensionality of the data to a single scalar and applies a threshold.

Definition 8.5 (Neyman-Pearson Form)**.** A test d is in Neyman-Pearson form if it is defined by a **test statistic** $T : \mathcal{X} \to \mathbb{R}$, a **critical value** $k \in \mathbb{R}$, and a randomization constant $\gamma \in [0, 1]$ such that:

$$d(x) = \begin{cases} 1 & \text{if } T(x) > k, \\ \gamma & \text{if } T(x) = k, \\ 0 & \text{if } T(x) < k. \end{cases}$$

Intuition and Components

- **The Test Statistic $T(x)$:** This function summarizes the complex dataset (e.g., a vector of observations) into a single number relevant to the hypothesis.
  - We choose T such that it tends to take **small to moderate values** when H_0 is true.
  - It tends to take **large values** when H_0 is false (i.e., when H_1 is true).
- **The Critical Value k:** This is the threshold of evidence. If the signal $T(x)$ exceeds k, we reject H_0. The "art" of designing a statistical test lies in calibrating this k to achieve a specific risk profile (e.g., ensuring the Type I error probability does not exceed 5%).
- **The Randomization γ:** If the statistic lands exactly on the threshold ($T(x) = k$), we may need "wiggle room" to satisfy strict error constraints. In this boundary case, we flip a coin with probability γ to decide whether to reject. (Note: For continuous data, $P(T(X) = k) = 0$, so γ is often irrelevant).

Example 8.2 (Application to Television Ads)**.** Revisiting Example 8.1 (Television ads), we can define the components as follows:

- **Data:** The raw data might be a spreadsheet of n rows, where each entry is a preference for Product A (1) or Product B (0).
- **Test Statistic:** We choose $T(x) = \sum x_i$ (the total count of participants preferring A).
- **Logic:**
  - If H_0 is true (Product A is not better), we expect $T(x)$ to be around $n/2$ or lower.
  - If H_1 is true (Product A is better), we expect $T(x)$ to be significantly higher.
- **Decision:** We calculate a threshold k based on the Binomial distribution. If the count $T(x) > k$, we reject H_0 and air the ad.

8.1.4 Bayes Tests

In the Bayesian framework, we treat the parameter θ as a random variable. Let Π be a **Prior distribution** on the parameter space, which is partitioned into $\Theta = \Theta_0 \dot{\cup} \Theta_1$. Let $\Pi(\cdot|x)$ denote the posterior distribution given the data $x \in \mathcal{X}$.

Minimizing Posterior Risk

Unlike the frequentist risk (which averages over data for a fixed θ), the Bayesian approach fixes the data x and averages the loss over the uncertainty in θ (the posterior).

Under the Neyman-Pearson (NP) loss structure (defined in the previous section), the **posterior risk** of taking a randomized action $p \in [0,1]$ (where p is the probability of rejecting H_0) is:

$$\begin{aligned}\ell(x,p) &= \int_\Theta L(\theta,p)\, d\Pi(\theta|x) \\ &= \int_{\Theta_0} \underbrace{c_1 p}_{\text{Loss if } H_0 \text{ true}} d\Pi(\theta|x) + \int_{\Theta_1} \underbrace{c_0(1-p)}_{\text{Loss if } H_1 \text{ true}} d\Pi(\theta|x) \\ &= p \cdot c_1 \cdot \Pi(\Theta_0|x) + (1-p) \cdot c_0 \cdot \Pi(\Theta_1|x).\end{aligned}$$

Rearranging terms to group by p:

$$\ell(x,p) = p\left[c_1\Pi(\Theta_0|x) - c_0\Pi(\Theta_1|x)\right] + c_0\Pi(\Theta_1|x).$$

Optimization: A **Bayes test** (or Bayes rule) is the decision rule $d(x)$ that minimizes this posterior risk for every x:

$$d(x) = \arg\min_{p\in[0,1]} \ell(x,p).$$

Note that $\ell(x,p)$ is a **linear function** of p on the interval $[0,1]$.

- If the slope (term in brackets) is negative, the minimum is at $p = 1$.
- If the slope is positive, the minimum is at $p = 0$.
- If the slope is zero, any p works (the risks are balanced).

Theorem 8.1 (Bayes Test Form). *A Bayes test under Neyman-Pearson loss has the form:*

$$d(x) = \begin{cases} 1, & \text{if } c_0\Pi(\Theta_1|x) > c_1\Pi(\Theta_0|x), \\ \gamma(x), & \text{if } c_0\Pi(\Theta_1|x) = c_1\Pi(\Theta_0|x), \\ 0, & \text{if } c_0\Pi(\Theta_1|x) < c_1\Pi(\Theta_0|x), \end{cases}$$

where $\gamma(x) \in [0,1]$ is arbitrary.

Bayes Tests of Simple Hypotheses

Consider the specific case where we compare exactly two distributions (Simple vs. Simple). Let $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$.

- **Model:** P_{θ_0} has density $p_0(x)$ and P_{θ_1} has density $p_1(x)$.

- **Priors:** Let $\pi_0 := \Pi(\Theta_0)$ and $\pi_1 := \Pi(\Theta_1)$ with $\pi_0 + \pi_1 = 1$.

Using Bayes' theorem, the posterior probabilities are:

$$\Pi(\Theta_0|x) = \frac{\pi_0 p_0(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}, \qquad \Pi(\Theta_1|x) = \frac{\pi_1 p_1(x)}{\pi_0 p_0(x) + \pi_1 p_1(x)}.$$

Substituting these into Theorem 8.1, the denominators cancel out. We reject H_0 (action $d(x) = 1$) if:

$$c_0 \pi_1 p_1(x) > c_1 \pi_0 p_0(x).$$

The Likelihood Ratio Test

We can rearrange the inequality above to separate the data from the external factors (costs and priors). The Bayes test rejects H_0 when:

$$\underbrace{\frac{p_1(x)}{p_0(x)}}_{\text{Likelihood Ratio}} > \underbrace{\frac{c_1}{c_0}}_{\text{Cost Ratio}} \cdot \underbrace{\frac{\pi_0}{\pi_1}}_{\text{Prior Odds}}$$

Interpretation:

- **Likelihood Ratio (p_1/p_0):** This quantifies how much more likely the observed data x is under H_1 compared to H_0. It is the sufficient statistic for this problem.

- **Cost Ratio (c_1/c_0):** Adjusts the threshold based on consequences. If Type I errors are very expensive (c_1 is high), the threshold increases, requiring stronger evidence to reject.

- **Prior Odds (π_0/π_1):** Adjusts based on prior belief. If H_0 is extremely likely a priori (e.g., "Trump will not win" in the lecturer's 2016 example), π_0 is high, pushing the threshold up.

Remark. If costs are equal ($c_1 = c_0$) and priors are equal ($\pi_0 = \pi_1$), the test simply compares which density is higher: reject if $p_1(x) > p_0(x)$.

Example 8.3 (Testing Binomial Probability)**.** Consider the test of a simple null hypothesis against a composite alternative:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0,$$

where the data follows a Binomial distribution, $X \sim \text{Binomial}(n, \theta)$, with density:

$$p_\theta(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}, \quad x = 0, 1, \ldots, n.$$

Constructing the Prior To construct a Bayes test, we must assign a prior distribution Π over $\Theta = [0, 1]$.

Note (Prior Selection)**.** A continuous prior (e.g., $\theta \sim \text{Uniform}[0, 1]$) would be inappropriate here because it assigns zero probability to any single point, meaning $\Pi(\{\theta_0\}) = 0$. *If the prior probability of H_0 is zero, the posterior probability will also be zero regardless of the data.*

Therefore, we must assign a **positive probability mass** to θ_0. We define the prior as a mixture:

$$\Pi = \pi_0 \cdot \delta_{\theta_0} + (1 - \pi_0) \cdot \text{Beta}(a, b),$$

where $\pi_0 \in (0, 1)$ is the prior probability that H_0 is true, and δ_{θ_0} is a point mass (Dirac delta) at θ_0.

The Two-Stage Draw We can think of drawing a parameter θ from this prior as a two-step random process:

1. **Hypothesis Selection:** Flip a biased coin that shows Heads with probability π_0.
2. **Parameter Assignment:**
   - If Heads (H_0): Set $\theta = \theta_0$ (the coin is “fair” or fixed).
   - If Tails (H_1): Draw θ randomly from the alternative distribution, $\theta \sim \text{Beta}(a, b)$.

Prior Predictive Probability To compute the posterior, we first need the marginal likelihood (prior predictive probability) of observing the data x. By the Law of Total Probability, we integrate the likelihood against the prior mixture:

$$\begin{aligned}
\mathsf{P}(X = x) &= \int p_\theta(x)\, \mathrm{d}\Pi(\theta) \\
&= \pi_0 \underbrace{\int p_\theta(x)\, \mathrm{d}\delta_{\theta_0}(\theta)}_{\text{Likelihood under } H_0} + (1 - \pi_0) \underbrace{\int p_\theta(x)\, \mathrm{dBeta}(a, b)}_{\text{Likelihood under } H_1}.
\end{aligned}$$

The first integral is simply the evaluation at the point mass. The second integral exploits the conjugacy of the Beta distribution with the Binomial likelihood (“lazy integration”):

$$\begin{aligned}
\mathsf{P}(X = x) &= \pi_0 p_{\theta_0}(x) + (1 - \pi_0) \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \binom{n}{x} \int_0^1 \theta^{x+a-1}(1 - \theta)^{n-x+b-1}\, \mathrm{d}\theta \\
&= \pi_0 \binom{n}{x} \theta_0^x (1 - \theta_0)^{n-x} + (1 - \pi_0) \binom{n}{x} \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(x + a)\Gamma(n - x + b)}{\Gamma(n + a + b)}.
\end{aligned}$$

Posterior Probability and Decision The posterior probability of the null hypothesis $\Theta_0 = \{\theta_0\}$ is given by Bayes' Rule:

$$\Pi(\Theta_0|x) = \frac{\mathsf{P}(X = x|\theta = \theta_0)\,\mathsf{P}(\theta = \theta_0)}{\mathsf{P}(X = x)} = \frac{\pi_0\binom{n}{x}\theta_0^x(1-\theta_0)^{n-x}}{\mathsf{P}(X = x)}.$$

The posterior for the alternative is simply the complement, $\Pi(\Theta_1|x) = 1 - \Pi(\Theta_0|x)$.

A Bayes test under Neyman-Pearson loss compares the ratio of these posterior probabilities (the posterior odds) against the cost ratio:

$$\text{Reject } H_0 \iff \frac{\Pi(\Theta_1|x)}{\Pi(\Theta_0|x)} > \frac{c_1}{c_0}.$$

8.2 Most Powerful Tests

In the previous sections on Decision Theory and Bayes tests, we assumed we could quantify the loss of making errors (via costs c_0, c_1) and potentially assign prior probabilities. In practice, however, these values are often unknown or controversial (e.g., what is the exact cost of a medical side effect versus a cure?).

The Neyman-Pearson Approach

The Neyman-Pearson (NP) framework avoids assigning specific costs by treating the hypotheses asymmetrically.

- **Null Hypothesis (H_0):** Represents the status quo, current scientific standard, or "no effect."
- **Alternative Hypothesis (H_1):** Represents a new discovery, effect, or change.

We consider rejecting H_0 when it is true (Type I Error) to be the "worse" mistake. Therefore, we impose a strict constraint on this error probability.

Definition 8.6 (Significance Level)**.** We require a test $d(X)$ to have a **significance level** $\alpha \in (0, 1)$. This means the probability of rejecting H_0 when it is true is bounded by α:

$$\mathsf{E}_\theta[d(X)] \le \alpha, \quad \forall\theta \in \Theta_0.$$

Commonly, $\alpha = 0.05$ (a convention popularized by Fisher).

Subject to this constraint, we seek to maximize the test's ability to detect the alternative.

Definition 8.7 (Power)**.** The **power** of a test is the probability of correctly rejecting H_0 when H_1 is true:

$$\beta(\theta) = \mathsf{E}_\theta[d(X)] \quad \text{for } \theta \in \Theta_1.$$

Most Powerful (MP) Tests for Simple Hypotheses

Consider the simplest case where we test two simple hypotheses:

$$\Theta_0 = \{\theta_0\} \quad \text{and} \quad \Theta_1 = \{\theta_1\}.$$

Definition 8.8 (Most Powerful Test). A test d is a **Most Powerful (MP) level α test** if:

1. **Constraint:** It satisfies the size constraint:
$$\mathsf{E}_{\theta_0}[d(X)] \leqslant \alpha.$$
2. **Maximization:** For any other test d' that satisfies the size constraint, d has greater or equal power:
$$\mathsf{E}_{\theta_1}[d(X)] \geqslant \mathsf{E}_{\theta_1}[d'(X)].$$

Remark (Existence of Densities). To derive these tests, we typically work with probability densities p_0 and p_1 corresponding to P_{θ_0} and P_{θ_1}.

Assuming the existence of these densities is **not** a restrictive assumption.

Even if P_{θ_0} and P_{θ_1} are singular with respect to Lebesgue measure (or discrete), we can always define a dominating measure $\nu = \mathsf{P}_{\theta_0} + \mathsf{P}_{\theta_1}$.

By the Radon–Nikodym theorem, densities $p_0 = \frac{dP_{\theta_0}}{d\nu}$ and $p_1 = \frac{dP_{\theta_1}}{d\nu}$ always exist.

The Neyman-Pearson Lemma

Theorem 8.2 (Neyman-Pearson Lemma, Lehmann and Romano 2005, Theorem 3.2.1). *Consider testing a simple null hypothesis against a simple alternative:*

$$H_0 : \theta = \theta_0 \quad vs. \quad H_1 : \theta = \theta_1.$$

Let p_0 and p_1 be the densities of P_{θ_0} and P_{θ_1} with respect to a dominating measure ν.We have:

a) ***Existence:*** *For all $\alpha \in (0,1)$, there exists a Most Powerful (MP) level α test of the form:*
$$d(x) = \begin{cases} 1, & \textit{if } p_1(x) > kp_0(x), \\ \gamma, & \textit{if } p_1(x) = kp_0(x), \\ 0, & \textit{if } p_1(x) < kp_0(x), \end{cases}$$
where $k \geq 0$ and $\gamma \in [0,1]$ are constants chosen such that $\mathsf{E}_{\theta_0}[d(X)] = \alpha$.

b) ***Characterization:*** *Let d be any test of level α. Then d is MP level α if and only if there exists $k \geqslant 0$ such that:*
$$d(x) = \begin{cases} 1, & \textit{if } p_1(x) > kp_0(x), \\ 0, & \textit{if } p_1(x) < kp_0(x), \end{cases} \qquad (\mathsf{P}_{\theta_0} + \mathsf{P}_{\theta_1})\text{-}a.e.$$
and if $k > 0$, then $\mathsf{E}_{\theta_0}[d(X)] = \alpha$.

Proof Strategy: Lagrangian Duality The optimization problem is to maximize Power $\mathsf{E}_{\theta_1}[d(X)]$ subject to the constraint $\mathsf{E}_{\theta_0}[d(X)] \leq \alpha$. We approach this using the logic of **Lagrange Multipliers**. We introduce a multiplier k and maximize the quantity:

$$\mathcal{L}(d,k) = \mathsf{E}_{\theta_1}[d(X)] - k(\mathsf{E}_{\theta_0}[d(X)] - \alpha).$$

Proof. **Step 1: Establishing an Upper Bound**

We derive a bound on the power of *any* level α test. Pick any $k \geq 0$. Define the weighted difference of densities:

$$v_k(x) := p_1(x) - kp_0(x).$$

We decompose this into positive and negative parts:

$$v_k^+(x) = \max\{v_k(x), 0\}, \quad v_k^-(x) = \max\{-v_k(x), 0\}.$$

Consider the power of an arbitrary test $d(x)$:

$$\begin{aligned}\mathsf{E}_{\theta_1}[d(X)] &= \int_{\mathcal{X}} d(x)p_1(x)\mathrm{d}\nu(x) \\ &= \int_{\mathcal{X}} d(x)[p_1(x) - kp_0(x) + kp_0(x)]\mathrm{d}\nu(x) \\ &= \int_{\mathcal{X}} d(x)v_k(x)\mathrm{d}\nu(x) \ + \ k\mathsf{E}_{\theta_0}[d(X)].\end{aligned}$$

Using the decomposition $v_k = v_k^+ - v_k^-$:

$$\mathsf{E}_{\theta_1}[d(X)] = \int_{\mathcal{X}} d(x)v_k^+(x)\mathrm{d}\nu(x) \ - \ \int_{\mathcal{X}} d(x)v_k^-(x)\mathrm{d}\nu(x) \ + \ k\mathsf{E}_{\theta_0}[d(X)].$$

We now apply upper bounds to each term:

- Since $d(x) \leq 1$ and $v_k^+ \geq 0$, the first integral is $\leq \int v_k^+$.
- Since $d(x) \geq 0$ and $v_k^- \geq 0$, the second integral is ≥ 0 (so subtracting it makes the total smaller or equal).
- Since d is level α, $\mathsf{E}_{\theta_0}[d(X)] \leq \alpha$.

Combining these, we get an upper bound $g(k)$:

$$\mathsf{E}_{\theta_1}[d(X)] \leqslant \int_{\mathcal{X}} v_k^+(x)\mathrm{d}\nu(x) \ + \ k\alpha =: g(k).$$

Step 2: Conditions for Equality (Sufficiency)

To achieve the maximum power $g(k)$, the inequalities above must become equalities. This happens if and only if:

i) $d(x) = 1$ whenever $v_k^+(x) > 0$ (i.e., $p_1 > kp_0$).

ii) $d(x) = 0$ whenever $v_k^-(x) > 0$ (i.e., $p_1 < kp_0$).

iii) If $k > 0$, we must fully exhaust the Type I error budget: $\mathsf{E}_{\theta_0}[d(X)] = \alpha$.

These conditions describe exactly the Neyman–Pearson test form defined in the theorem. Thus, any test of this form achieves the upper bound $g(k)$ and is Most Powerful.

Step 3: Construction (Existence)

How do we practically find such a test? We need to find k and γ.

Choosing k: We interpret the test condition $p_1(x) > kp_0(x)$ as a rejection region based on the Likelihood Ratio statistic $T(X) = \frac{p_1(X)}{p_0(X)}$. We reject when $T(X)$ is large. To ensure level α, we choose k to be the $(1-\alpha)$**-quantile** of the distribution of the Likelihood Ratio under H_0:

$$k = \inf \left\{ t : \mathsf{P}_{\theta_0} \left(\frac{p_1(X)}{p_0(X)} > t \right) \leq \alpha \right\}.$$

Choosing γ (Randomization): If the distribution of the likelihood ratio is continuous, $\mathsf{P}_{\theta_0}(p_1/p_0 = k) = 0$, so γ does not matter (set to 0). However, if the distribution is **discrete**, the probability mass at k might prevent us from achieving exactly α. We might jump from a rejection probability of, say, 0.04 to 0.06 when we include the point k.

We set γ to "fill the gap" exactly:

$$\gamma = \frac{\alpha - \mathsf{P}_{\theta_0}\left(\frac{p_1(X)}{p_0(X)} > k\right)}{\mathsf{P}_{\theta_0}\left(\frac{p_1(X)}{p_0(X)} = k\right)}.$$

This ensures $\mathsf{E}_{\theta_0}[d(X)] = 1 \cdot \mathsf{P}(LR > k) + \gamma \cdot \mathsf{P}(LR = k) + 0 = \alpha$.

Step 4: Necessity

If d is any MP level α test, it must have power equal to the optimal test constructed above (since both are MP). Therefore, d must achieve the upper bound $g(k)$. As shown in Step 2, achieving this bound implies d must have the Neyman-Pearson form almost everywhere. □

Example 8.4 (Binomial Proportion)**.** Let $\boldsymbol{X} = (X_1, ..., X_n)$ be independent Bernoulli trials where $X_i \sim \text{Bern}(\theta)$. We observe the sequence of successes and failures. Consider the simple hypotheses:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta = \theta_1, \quad \text{with } 0 < \theta_0 < \theta_1 < 1.$$

Deriving the Test

According to the Neyman-Pearson Lemma, the MP test rejects H_0 for large values of the likelihood ratio:

$$\Lambda(\boldsymbol{x}) = \frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} > k.$$

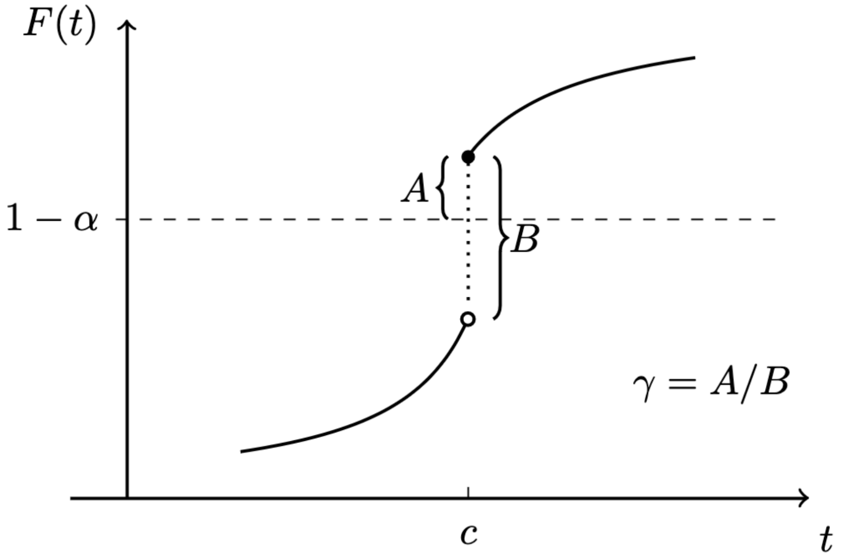


Figure 8.1: Randomization constant

The likelihood for a binary sequence $\boldsymbol{x} \in \{0,1\}^n$ is:

$$p_\theta(\boldsymbol{x}) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.$$

Substituting this into the ratio:

$$\frac{p_{\theta_1}(\boldsymbol{x})}{p_{\theta_0}(\boldsymbol{x})} = \frac{\theta_1^{\sum x_i}(1-\theta_1)^{n-\sum x_i}}{\theta_0^{\sum x_i}(1-\theta_0)^{n-\sum x_i}} = \left(\frac{\theta_1}{\theta_0} \cdot \frac{1-\theta_0}{1-\theta_1}\right)^{\sum\limits_{i=1}^{n} x_i} \left(\frac{1-\theta_1}{1-\theta_0}\right)^n.$$

Simplification via Monotonicity The term $\left(\frac{1-\theta_1}{1-\theta_0}\right)^n$ is a constant (does not depend on data). Because we assumed $\theta_1 > \theta_0$, the base of the exponent term is strictly greater than 1:

$$\frac{\theta_1(1-\theta_0)}{\theta_0(1-\theta_1)} > 1.$$

Therefore, the likelihood ratio is a strictly **increasing function** of the sufficient statistic $T(\boldsymbol{x}) = \sum\limits_{i=1}^{n} X_i$.

Instead of defining a threshold k on the likelihood ratio scale, we can define an equivalent threshold c on the scale of $T(\boldsymbol{x})$. The test becomes:

$$d(\boldsymbol{x}) = \begin{cases} 1, & \text{if } T(\boldsymbol{x}) > c, \\ \gamma, & \text{if } T(\boldsymbol{x}) = c, \\ 0, & \text{if } T(\boldsymbol{x}) < c. \end{cases}$$

Determining Constants c and γ

We need $\mathsf{E}_{\theta_0}[d(\boldsymbol{X})] = \alpha$. Under H_0, the statistic follows a Binomial distribution:

$$T(\boldsymbol{X}) \sim \text{Bin}(n, \theta_0).$$

Since this distribution is discrete, there may not be a integer c such that $\mathsf{P}(T > c) = \alpha$.

We choose c as the $(1-\alpha)$-quantile of $\text{Bin}(n, \theta_0)$. Specifically:

1. Find c such that $\mathsf{P}_{\theta_0}(T > c) \leq \alpha$ but $\mathsf{P}_{\theta_0}(T \geq c) > \alpha$.

2. Set γ to account for the difference:

$$\gamma = \frac{\alpha - \mathsf{P}_{\theta_0}(T > c)}{\mathsf{P}_{\theta_0}(T = c)}.$$

In practice (e.g., in R), c is found via `qbinom(1-alpha, n, theta0)`.

8.2.1 Always Some Power

We can prove that an optimal test is always strictly better than a naive guess, provided the distributions are distinct.

Corollary 8.1 (Lehmann and Romano 2005, Corollary 3.2.1)**.** *Let $\alpha \in (0,1)$, and let d be a most powerful level α test for a simple vs. simple problem where $\mathsf{P}_{\theta_0} \neq \mathsf{P}_{\theta_1}$. Then the power β satisfies:*

$$\beta = \mathsf{E}_{\theta_1}[d(X)] > \alpha \geq \mathsf{E}_{\theta_0}[d(X)].$$

Proof. Consider a trivial competitor test $d^*(x) \equiv \alpha$ (the "Three Monkeys" test: ignore data, reject with probability α regardless of observation).

- The size of d^* is $\mathsf{E}_{\theta_0}[d^*(X)] = \alpha$. Thus, d^* is a valid level α test.
- The power of d^* is $\mathsf{E}_{\theta_1}[d^*(X)] = \alpha$.

Since d is the Most Powerful test, it must be at least as powerful as d^*. Thus $\beta \geq \alpha$.

Proof of Strict Inequality (by contradiction): Assume $\beta = \alpha$. Then d^* is also a Most Powerful test (it achieves the maximum power α). By the uniqueness part of the Neyman–Pearson Lemma (Theorem 8.2), any MP test must satisfy the form:

$$d^*(x) = \begin{cases} 1 & p_1(x) > kp_0(x) \\ 0 & p_1(x) < kp_0(x) \end{cases} \quad \text{a.e.}$$

However, $d^*(x)$ is constantly α (never 0 or 1, assuming $\alpha \in (0,1)$). This implies that we are almost always in the "tie" case where $p_1(x) = kp_0(x)$.

Thus, $p_1(x) = kp_0(x)$ almost everywhere. Integrating both sides:

$$\int p_1(x)d\nu = k\int p_0(x)d\nu \implies 1 = k(1) \implies k = 1.$$

If $k = 1$, then $p_1(x) = p_0(x)$ almost everywhere, which means $P_{\theta_1} = P_{\theta_0}$. This contradicts the assumption that the distributions are distinct. Therefore, $\beta > \alpha$. □

8.3 Uniformly Most Powerful Tests

In the previous section, we found the optimal test for a *simple* alternative (e.g., $\theta = \theta_1$). Now we consider **composite** hypotheses, where the alternative hypothesis contains multiple distributions.

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

where $|\Theta_1| > 1$.

Definition 8.9 (Size and Level)**.** The **size** of a test d is the maximum probability of Type I error over the null hypothesis space:

$$\text{Size}(d) = \sup_{\theta \in \Theta_0} \mathsf{E}_\theta[d(X)].$$

A test d has **level** α if its size is at most α.

Definition 8.10 (Uniformly Most Powerful)**.** Let $D(\alpha)$ be the set of all level α tests. A test $d \in D(\alpha)$ is a **Uniformly Most Powerful (UMP)** level α test if it is more powerful than any other level α test at *every* point in the alternative hypothesis:

$$\mathsf{E}_\theta[d(X)] \geq \mathsf{E}_\theta[d'(X)] \qquad \forall d' \in D(\alpha) \text{ and } \forall \theta \in \Theta_1.$$

Example 8.5 (Binomial Proportion, One–Sided)**.** Recall the Binomial example where $X_i \sim \text{Bin}(1, \theta) = \text{Bern}(\theta)$ and $T(X) = \sum_{i=1}^{n} X_i$. We have shown that the MP level α test of

$$H_0 \ : \ \theta = \theta_0 \text{ vs. } H_1 \ : \ \theta = \theta_1 \quad \text{ for } 0 < \theta_0 < \theta_1 < 1$$

is of the form

$$d_{\theta_0,\theta_1,\alpha}(\boldsymbol{x}) = \mathbf{1}_{(c,\infty)}(T(\boldsymbol{x})) + \gamma \mathbf{1}_{\{c\}}(T(\boldsymbol{x})).$$

Consider the composite alternative:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0.$$

From the Neyman-Pearson Lemma, for any *specific* alternative $\theta_1 > \theta_0$, the Most Powerful (MP) test rejects for large values of $T(X)$:

$$d_{\theta_0,\theta_1}(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) > c \\ \gamma & \text{if } T(\mathbf{x}) = c \\ 0 & \text{if } T(\mathbf{x}) < c \end{cases}$$

where c and γ are determined by the equation $\mathsf{E}_{\theta_0}[d(X)] = \alpha$.

Key Observation: Notice that the constants c and γ depend **only** on the null distribution $\text{Bin}(n, \theta_0)$ and the significance level α. They do **not** depend on the specific value of θ_1, as long as $\theta_1 > \theta_0$ (which ensures the likelihood ratio is increasing in T).

> *Intuition:* If Student A tests $H_0 : \theta = 0.5$ vs $H_1 : \theta = 0.75$, and Student B tests $H_0 : \theta = 0.5$ vs $H_1 : \theta = 0.9$, they will derive the **exact same test** (same threshold c, same γ). Since this single test is optimal for every single $\theta > \theta_0$ individually, it is **Uniformly Most Powerful** for the entire set $\theta > \theta_0$.

Expanding the Null Hypothesis

The test d_{θ_0} derived above is UMP for $H_0 : \theta = \theta_0$ vs $H_1 : \theta > \theta_0$.

Can we expand the null hypothesis to include all values less than θ_0?

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0$$

For the test to be valid for this expanded null, we must ensure its size is still α. That is, the probability of rejection must not exceed α for any $\theta < \theta_0$.

This holds true because the power function $\beta(\theta) = \mathsf{E}_\theta[d(X)]$ is **monotonically increasing** in θ.

$$\beta(\theta) = \sum_{t=c+1}^{n} \binom{n}{t} \theta^t (1-\theta)^{n-t} + \gamma \binom{n}{c} \theta^c (1-\theta)^{n-c}.$$

One can show by differentiation that $\beta'(\theta) > 0$. Therefore:

$$\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha.$$

Conclusion: The test d_{θ_0} is the UMP level α test for the one–sided problem $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$.

8.4 Monotone Likelihood Ratio

The Binomial example generalizes to a broader class of problems. For $\theta_0, \theta_1 \in \Theta$, define the **likelihood ratio**:

$$L_{\theta_1/\theta_0}(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \cdot \mathbf{1}_{(0,\infty)}(p_{\theta_0}(x)) + \infty \cdot \mathbf{1}_{\{0\}}(p_{\theta_0}(x)),$$

where p_{θ_0} and p_{θ_1} are densities of P_{θ_0} and P_{θ_1} with respect to a measure ν.

Definition 8.11 (Monotone Likelihood Ratio)**.** A one-parameter family $\{\mathsf{P}_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ has a **monotone likelihood ratio (MLR)** in a statistic $T : \mathcal{X} \to \mathbb{R}$ if for all $\theta_0 < \theta_1$:

i) $\mathsf{P}_{\theta_0} \neq \mathsf{P}_{\theta_1}$, and

ii) The likelihood ratio $L_{\theta_1/\theta_0}(x)$ is almost everywhere a non-decreasing function of $T(x)$. That is, there exists a non-decreasing function $H_{\theta_1/\theta_0} : \mathbb{R} \to [0, \infty]$ such that:
$$L_{\theta_1/\theta_0}(x) = H_{\theta_1/\theta_0}(T(x))$$
almost surely under both P_{θ_0} and P_{θ_1}.

Example 8.6. Many common statistical families satisfy the MLR property. Note that for i.i.d. observations, the relevant statistic $T(x)$ is often the sufficient statistic we are familiar with.

1. **Binomial / Bernoulli:**
   - $\{\bigotimes_{i=1}^{n} \mathrm{Bin}(1, \theta) : \theta \in (0,1)\}$ has MLR in $T(x) = \sum_{i=1}^{n} x_i$.
   - $\{\mathrm{Bin}(n, \theta) : \theta \in (0,1)\}$ has MLR in $T(x) = x$.
2. **Normal (Known Variance):**
$$\{\bigotimes_{i=1}^{n} \mathcal{N}(\theta, \sigma^2) : \theta \in \mathbb{R}\} \quad \text{has MLR in } T(x) = \overline{x}.$$
3. **Hypergeometric:**
$$\{\mathrm{Hypergeometric}(N, \theta, n) : \theta \in \{0, ..., n\}\} \quad \text{has MLR in } T(x) = x.$$
4. **Poisson:**
$$\{\bigotimes_{i=1}^{n} \mathrm{Poi}(\theta) : \theta > 0\} \quad \text{has MLR in } T(x) = \sum_{i=1}^{n} x_i.$$
5. **Uniform (Continuous):**
$$\{\bigotimes_{i=1}^{n} \mathrm{Uniform}(0, \theta) : \theta > 0\} \quad \text{has MLR in } T(x) = \max_{1 \le i \le n} x_i.$$
   Remark. This is not an exponential family, but it still possesses MLR.
6. **Advanced Distributions:** Further examples include noncentral χ^2, noncentral F, or noncentral t-distributions (often parameterized by a non-centrality parameter).

Counter-Example: The Cauchy Distribution

The **Cauchy location family** defined by the density:
$$p_\theta(x) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}, \quad \theta \in \mathbb{R},$$
does **not** have a Monotone Likelihood Ratio in x (or in any other statistic $T(x)$).

Why? Although shifting the center of a Cauchy distribution increases the probability of values near the new center, the heavy tails mean that extremely large values of x do not necessarily provide stronger evidence for the larger θ, causing the likelihood ratio to oscillate or decrease eventually.

8.4.1 Exponential Families and MLR

One-parameter exponential families provide a powerful source of models with the Monotone Likelihood Ratio property.

Proposition 8.1. *If $\{\mathsf{P}_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ forms a one-parameter exponential family with densities:*

$$p_\theta(x) = c(\theta) \cdot \exp\{\eta(\theta)T(x)\} \cdot h(x), \quad x \in \mathcal{X},$$

and the natural parameter $\eta(\theta)$ is a ***strictly increasing*** *function of θ, then the family has MLR in the sufficient statistic $T(x)$.*

Proof. Let $\theta_1 > \theta_0$. The likelihood ratio is:

$$L_{\theta_1/\theta_0}(x) = \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = \frac{c(\theta_1)e^{\eta(\theta_1)T(x)}h(x)}{c(\theta_0)e^{\eta(\theta_0)T(x)}h(x)} = \frac{c(\theta_1)}{c(\theta_0)} \exp\left\{[\eta(\theta_1) - \eta(\theta_0)] \cdot T(x)\right\}.$$

Since η is strictly increasing, $[\eta(\theta_1) - \eta(\theta_0)] > 0$. The function $y \mapsto e^{ay}$ with $a > 0$ is increasing. Therefore, $L_{\theta_1/\theta_0}(x)$ is an increasing function of $T(x)$. □

Example 8.7 (Gaussian Mean)**.** Let $X_1, ..., X_n \overset{\text{iid}}{\sim} \mathcal{N}(\theta, \sigma^2)$ with $\theta \in \mathbb{R}$ unknown and σ^2 known. We wish to test:

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0.$$

Checking MLR

First, we verify the MLR property. For any $\theta_1 > \theta_0$, the likelihood ratio is:

$$\begin{aligned}
L_{\theta_1/\theta_0}(\boldsymbol{x}) &= \frac{\prod_{i=1}^{n} \exp\left\{\frac{-1}{2\sigma^2}(x_i - \theta_1)^2\right\}}{\prod_{i=1}^{n} \exp\left\{\frac{-1}{2\sigma^2}(x_i - \theta_0)^2\right\}} \quad \text{(constants cancel)} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i - \theta_1)^2 - (x_i - \theta_0)^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[x_i^2 - 2x_i\theta_1 + \theta_1^2 - (x_i^2 - 2x_i\theta_0 + \theta_0^2)\right]\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[2(\theta_0 - \theta_1)\sum_{i=1}^{n} x_i + n(\theta_1^2 - \theta_0^2)\right]\right\} \\
&= \exp\left\{\frac{n(\theta_1 - \theta_0)}{\sigma^2}\overline{x}_n\right\} \cdot C(\theta_0, \theta_1).
\end{aligned}$$

Since $\theta_1 > \theta_0$, the coefficient of $\overline{x}_n$ is positive. Thus, the family has MLR in $T(\boldsymbol{x}) = \overline{x}_n$.

The UMP Test

Since the family has MLR in $\overline{x}_n$, the UMP level α test rejects for large values of $\overline{x}_n$:

$$d(\boldsymbol{x}) = \begin{cases} 1, & \text{if } \overline{x}_n \geq c, \\ 0, & \text{if } \overline{x}_n < c. \end{cases}$$

Remark. Since the distribution is continuous, $\mathsf{P}(\overline{x}_n = c) = 0$, so the randomization constant γ is irrelevant and we can simply reject at equality.

Determining c: We require size α at the boundary θ_0:

$$\mathsf{P}_{\theta_0}(\overline{X}_n \geq c) = \alpha.$$

Under θ_0, $\overline{X}_n \sim \mathcal{N}(\theta_0, \sigma^2/n)$. Standardizing:

$$\mathsf{P}\left(\frac{\overline{X}_n - \theta_0}{\sigma/\sqrt{n}} \geq \frac{c - \theta_0}{\sigma/\sqrt{n}}\right) = \alpha \implies \frac{c - \theta_0}{\sigma/\sqrt{n}} = z_{1-\alpha}.$$

Thus, $c = \theta_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$.

Expanding the Null Hypothesis

This test is derived for $H_0 : \theta = \theta_0$. Is it valid for $H_0 : \theta \leq \theta_0$?

Yes, because the power function $\beta(\theta) = \mathsf{P}_\theta(\text{Reject})$ is strictly increasing.

Proof of Monotonicity: Let $\theta < \tilde{\theta}$. We use the fact that if $Z \sim \mathcal{N}(0, \sigma^2/n)$, then $\overline{X}_n \overset{d}{=} Z + \theta$.

$$\begin{aligned} \beta(\theta) &= \mathsf{P}_\theta(\overline{X}_n > c) \\ &= \mathsf{P}(Z + \theta > c) \\ &= \mathsf{P}(Z > c - \theta). \end{aligned}$$

Since $\theta < \tilde{\theta}$, we have $c - \theta > c - \tilde{\theta}$. Therefore, $\mathsf{P}(Z > c - \theta) < \mathsf{P}(Z > c - \tilde{\theta})$, which implies $\beta(\theta) < \beta(\tilde{\theta})$.

Since the power is increasing, $\sup\limits_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0) = \alpha$. The test is valid and UMP for the composite null.

8.4.2 UMP Tests and Monotone Likelihood Ratios

We can now formalize the generalization of the Binomial and Gaussian examples into a powerful theorem.

Theorem 8.3 (Lehmann and Romano 2005, Thm 3.4.1)**.** *Suppose $\{\mathsf{P}_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}$, has **Monotone Likelihood Ratio (MLR)** in $T(x)$. Let $\theta_0 \in \Theta$ and $\alpha \in (0, 1)$. Consider the test $d(x)$ defined by:*

$$d(x) = \begin{cases} 1, & \textit{if } T(x) > c, \\ \gamma, & \textit{if } T(x) = c, \\ 0, & \textit{if } T(x) < c, \end{cases}$$

where $c \in \mathbb{R}$ and $\gamma \in [0,1]$ are chosen such that $\mathsf{E}_{\theta_0}[d(X)] = \alpha$ (i.e., size is exactly α at the boundary).

Then the following properties hold:

i) ***UMP Property:*** *The test d is the Uniformly Most Powerful (UMP) level α test for:*

$$H_0 : \theta \leqslant \theta_0 \quad \textit{vs.} \quad H_1 : \theta > \theta_0.$$

ii) ***Monotonicity:*** *The power function $\beta(\theta) = \mathsf{E}_\theta[d(X)]$ is non-decreasing. It is strictly increasing for all θ where $0 < \beta(\theta) < 1$.*

iii) ***Minimizing Error under Null:*** *For all $\theta \leq \theta_0$, the test minimizes the probability of Type I error among all tests that have size α at the boundary:*

$$\beta(\theta) = \min\{\mathsf{E}_\theta[\tilde{d}] : \tilde{d} \textit{ is a test with } \mathsf{E}_{\theta_0}[\tilde{d}] = \alpha\}.$$

Proof.

i) This follows directly from the Neyman-Pearson Lemma and the definition of MLR.

1. Pick any specific alternative $\theta_1 > \theta_0$.
2. The NP Lemma says the MP test rejects when the likelihood ratio $L_{\theta_1/\theta_0}(x)$ is large.
3. Since the family has MLR, $L_{\theta_1/\theta_0}(x)$ is increasing in $T(x)$. Therefore, "large likelihood ratio" is equivalent to "$T(x) > c$".
4. Since this test form depends only on the direction ($\theta > \theta_0$) and not the specific value θ_1, it is optimal for *all* $\theta > \theta_0$ simultaneously.

ii) The power function must be increasing. If we treat θ' as a null and θ'' as an alternative (where $\theta' < \theta''$), the test d is Most Powerful. By the corollary "MP tests always have power $\geq$ size," we must have $\beta(\theta'') \geq \beta(\theta')$. This ensures that testing $H_0 : \theta \leq \theta_0$ is valid, because the maximum error rate (size) occurs at the boundary θ_0.

iii) This part is proven by a "flipping" argument. Consider the test $1 - d(x)$, which rejects when $T(x)$ is *small*. This corresponds to testing the reverse hypothesis (e.g., $H_0 : \theta = \theta_0$ vs $H_1 : \theta < \theta_0$). Since d maximizes power for $\theta > \theta_0$, the test $1 - d$ maximizes power for $\theta < \theta_0$ (relative to the reverse problem). Maximizing $\mathsf{E}[1 - d]$ is equivalent to minimizing $\mathsf{E}[d]$. Therefore, for $\theta < \theta_0$, our original test d has the minimal possible rejection probability.

Note. This means not only is the test good at rejecting the null when it's false (power), but it is also the best at accepting the null when it is true (minimizing Type I error) compared to any other test that spends the same error budget α at the boundary θ_0.

□

8.5 Discussion of Two-Sided Testing Problems

Consider the standard two-sided testing problem:

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

Non-Existence of UMP Tests

Typically, a Uniformly Most Powerful (UMP) test **does not exist** for this problem.

Why? To be UMP for the two-sided alternative, a test d would need to have a power function $\beta_d(\theta)$ that is at least as high as *any* other level α test for *every* $\theta \neq \theta_0$.

- There exists a UMP test d_1 for $H_1 : \theta > \theta_0$ (high power on the right, low on the left).
- There exists a UMP test d_2 for $H_1 : \theta < \theta_0$ (high power on the left, low on the right).

A hypothetical two-sided UMP test would need to match the power of d_1 on the right and d_2 on the left. This would require the power function to be extremely high on both sides while instantaneously dropping to α at exactly θ_0. Due to the smoothness of power functions (especially in exponential families), such a function usually cannot exist without violating the level constraint.

The Solution: Unbiasedness

Since we cannot satisfy the UMP condition everywhere, we restrict our search to a smaller, more "reasonable" class of tests: **Unbiased Tests**.

Definition 8.12 (Unbiased Test)**.** A test with power function $\beta(\theta)$ is **unbiased** at level α if:

$$\begin{aligned} \beta(\theta) &\leq \alpha \quad \forall \theta \in \Theta_0 \quad \text{(Control Type I Error)} \\ \beta(\theta) &\geq \alpha \quad \forall \theta \in \Theta_1 \quad \text{(Power never worse than random guessing)} \end{aligned}$$

By restricting the competition to unbiased tests, we can often find a **Uniformly Most Powerful Unbiased (UMPU)** test.

- For the Normal mean (variance known), the UMPU test rejects when $|\overline{x} - \theta_0|$ is large.
- This corresponds to symmetric rejection regions (e.g., using $\alpha/2$ on both tails).

8.6 p–Values

In practice, data analysis often requires checking multiple significance levels or reporting the strength of evidence rather than a simple "Yes/No" decision.

Suppose we have a family of non-randomized tests d_α for every level $\alpha \in (0, 1)$ with **nested rejection regions** (i.e., rejecting at a smaller α implies rejecting at a larger α).

Definition 8.13 (p–value)**.** The p**–value** $p(x)$ is the smallest significance level at which the observed data leads to rejection:

$$p(x) = \inf\{\alpha \in (0, 1) : d_\alpha(x) = 1\}.$$

For standard tests that reject for large values of a statistic $T(X)$, the p–value corresponds to the probability of observing a statistic as extreme or *more extreme* than the observed value $T(x_{obs})$ under the null hypothesis:

$$p(x_{obs}) = \sup_{\theta \in \Theta_0} \mathsf{P}_\theta \left(T(X) \geq T(x_{obs})\right).$$

If the null distribution is unique (e.g., $T(X) \sim \mathcal{N}(0, 1)$ under H_0), this simplifies to:

$$p(x_{obs}) = \mathsf{P}_0(T(X) \geq T(x_{obs})).$$

Practical Interpretation

The p–value serves as a continuous measure of evidence against the null hypothesis.

- **Small p–value (e.g., < 0.05):** Strong evidence *against* H_0.
- **Large p–value:** No evidence against H_0.

  Remark. This does **not** prove H_0 is true; it merely indicates the data is consistent with it.

The "Lottery" Analogy: If you observe a very small p–value (e.g., $p = 0.0001$), you are faced with two possibilities:

a) The null hypothesis is true, and you have just witnessed a highly improbable "lottery win" event.

b) The null hypothesis is false.

Scientific testing generally proceeds by betting on the second option.

Decision Rule

If a specific significance level α_{target} is fixed in advance, the decision rule using the p–value is:

$$\text{Decision} = \begin{cases} \text{Reject } H_0 & \text{if } p(x) \leq \alpha_{\text{target}}, \\ \text{Do not reject } H_0 & \text{if } p(x) > \alpha_{\text{target}}. \end{cases}$$

8.6.1 Uniformity of p–Values

A crucial property of p–values is their behavior when the null hypothesis is actually true. Understanding this helps distinguish genuine signals from random noise.

Informally: **Under the null hypothesis, the p–value is uniformly distributed on $(0,1)$.**

This means that if H_0 is true (e.g., "red wine has no effect on health") and you repeat the experiment many times, you are equally likely to get a p–value of 0.01, 0.45, or 0.99.

- **Under H_0:** The histogram of p–values is flat.
- **Under H_1:** The histogram of p–values spikes near 0 (small p–values are more likely).

Lemma 8.1 (Lehmann and Romano 2005, Lemma 3.3.1).). *Suppose $X \sim \mathsf{P}_\theta$ with $\theta \in \Theta_0$ (the null is true).*

(i) If the test size is controlled for all levels, i.e.,

$$\sup_{\theta \in \Theta_0} \mathsf{P}_\theta(d_\alpha(X) = 1) \leqslant \alpha \qquad \forall \alpha \in (0,1),$$

then the p–value is ***stochastically larger*** *than the Uniform(0,1) distribution:*

$$\mathsf{P}_\theta(p(X) \leq u) \leq u \quad \forall u \in [0,1].$$

(ii) If the test size is exactly equal to α for all levels, i.e.,

$$\mathsf{P}_\theta(d_\alpha(X) = 1) = \alpha \qquad \forall \alpha \in (0,1),$$

then:

$$p(X) \sim \textit{Unif}(0,1).$$

Proof.

i) By the definition of the p–value ($p(x) = \inf\{\alpha : d_\alpha(x) = 1\}$) and the nested rejection regions assumption: If $p(x) \leq u$, then for any significance level $\nu > u$, the test must reject ($d_\nu(x) = 1$). Therefore, for any $\nu > u$:

$$\{x : p(x) \leq u\} \subseteq \{x : d_\nu(x) = 1\}.$$

Taking probabilities under $\theta \in \Theta_0$:

$$\mathsf{P}_\theta(p(X) \leq u) \leq \mathsf{P}_\theta(d_\nu(X) = 1) \leq \nu.$$

Since this holds for all $\nu > u$, we take the limit as $\nu \downarrow u$:

$$\mathsf{P}_\theta(p(X) \leq u) \leq u.$$

ii) Conversely, if a test rejects at level u ($d_u(x) = 1$), then by definition the p–value must be less than or equal to u:

$$\{x : d_u(x) = 1\} \subseteq \{x : p(x) \leq u\}.$$

Taking probabilities:

$$\mathsf{P}_\theta(p(X) \leq u) \geq \mathsf{P}_\theta(d_u(X) = 1) = u.$$

Combining the results from (i) and (ii) (since $\leq u$ and $\geq u$ implies $= u$), we get:

$$\mathsf{P}_\theta(p(X) \leq u) = u.$$

This is the CDF of a Uniform(0,1) random variable.

□

Connection to Probability Integral Transform

This result is related to the Probability Integral Transform. For a continuous random variable T with CDF F_T, the random variable $Y = F_T(T)$ is uniformly distributed. Since a one–sided p–value is essentially $1 - F_T(T(X))$ (calculating the tail area), it follows the same uniform distribution logic under the null.

8.7 Confidence Sets

Setup We consider the standard statistical setup:

- Observation X in sample space $\mathcal{X} \subset \mathbb{R}^d$.
- Statistical model $\mathcal{P}$ for the distribution of X.
- Target parameter $\gamma : \mathcal{P} \to \Gamma$ (e.g., the mean).
- Significance level $\alpha \in (0, 1)$, representing our “budget for mistakes.”

Definition 8.14 (Confidence Set)**.** A family of subsets $S(x)$ is a **confidence set of confidence level** $1-\alpha$ if the probability that the random set covers the true parameter is at least $1-\alpha$ for all distributions in the model:

$$\inf_{\mathsf{P}\in\mathcal{P}} \mathsf{P}(\gamma(\mathsf{P}) \in S(X)) \geqslant 1-\alpha.$$

The probability $\mathsf{P}(\gamma(\mathsf{P}) \in S(X))$ is called the **coverage probability**.

Example 8.8 (Normal Mean)**.** Let $\boldsymbol{X} = (X_1, ..., X_n)$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. The target parameter is the mean μ.

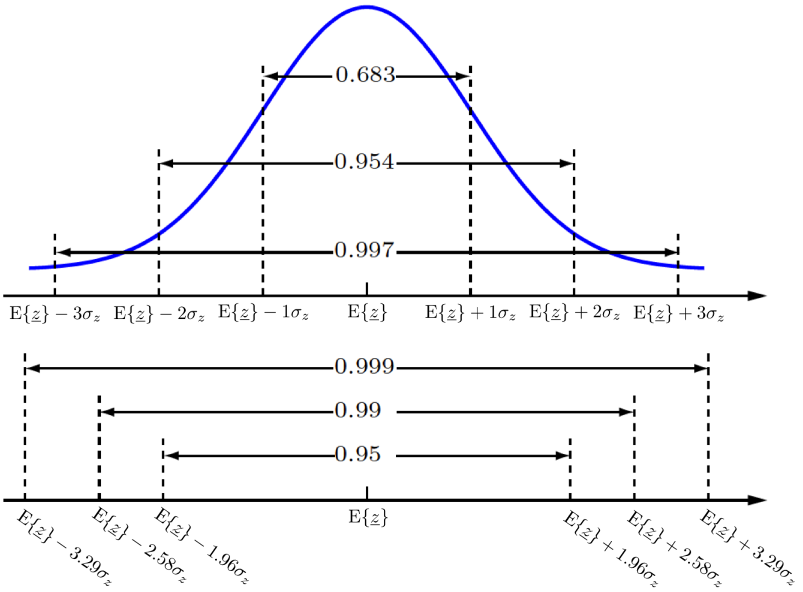

0.683
0.954
0.997
$\mathrm{E}\{\underline{z}\} - 3\sigma_z$
$\mathrm{E}\{\underline{z}\} - 2\sigma_z$
$\mathrm{E}\{\underline{z}\} - 1\sigma_z$
$\mathrm{E}\{\underline{z}\}$
$\mathrm{E}\{\underline{z}\} + 1\sigma_z$
$\mathrm{E}\{\underline{z}\} + 2\sigma_z$
$\mathrm{E}\{\underline{z}\} + 3\sigma_z$
0.999
0.99
0.95
$\mathrm{E}\{\underline{z}\} - 3.29\sigma_z$
$\mathrm{E}\{\underline{z}\} - 2.58\sigma_z$
$\mathrm{E}\{\underline{z}\} - 1.96\sigma_z$
$\mathrm{E}\{\underline{z}\}$
$\mathrm{E}\{\underline{z}\} + 1.96\sigma_z$
$\mathrm{E}\{\underline{z}\} + 2.58\sigma_z$
$\mathrm{E}\{\underline{z}\} + 3.29\sigma_z$

Figure 8.2: Standard normal distribution confidence interval

a) **Variance $\sigma^2 = \sigma_0^2$ is Known:** When the variance is known, we use the quantiles of the standard normal distribution $\mathcal{N}(0,1)$. The $(1-\alpha)$ confidence interval is:

$$S(\mathbf{x}) = \left(\overline{x}_n - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \quad \overline{x}_n + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right)$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$.

Derivation via Pivot: The construction relies on the random variable:

$$Z = \frac{\overline{X}_n - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1)$$

This variable Z is a **pivot**: it depends on the data and the parameter, but its distribution (Standard Normal) is known and fixed. We construct the interval by requiring:

$$\mathsf{P}\left(-z_{\alpha/2} \leq \frac{\overline{X}_n - \mu}{\sigma_0/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

Rearranging the inequalities to isolate μ yields the interval above.

b) **Variance σ^2 is Unknown:** If σ^2 is unknown, we must estimate it using the sample standard deviation $s(\boldsymbol{x})$. We replace the normal quantile $z_{\alpha/2}$ with the quantile from the Student's t-distribution.

$$S(\boldsymbol{x}) = \left(\overline{x}_n - t_{\alpha/2}(n-1)\frac{s(\boldsymbol{x})}{\sqrt{n}}, \quad \overline{x}_n + t_{\alpha/2}(n-1)\frac{s(\boldsymbol{x})}{\sqrt{n}}\right)$$

- $s(\boldsymbol{x}) = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{x}_n)^2}$
- $t_{\alpha/2}(n-1)$ is the $(1-\alpha/2)$-quantile of the t-distribution with $n-1$ degrees of freedom.
- **Pivot:** $\frac{\overline{X}_n - \mu}{s(\boldsymbol{X})/\sqrt{n}} \sim t(n-1)$.

c) **One–Sided Intervals** We can also form upper or lower confidence bounds (e.g., "The mean is at least L").

$$\text{Lower Bound: } \left(-\infty, \quad \overline{x}_n + t_{\alpha}(n-1)\frac{s(\boldsymbol{x})}{\sqrt{n}}\right)$$

$$\text{Upper Bound: } \left(\overline{x}_n - t_{\alpha}(n-1)\frac{s(\boldsymbol{x})}{\sqrt{n}}, \quad \infty\right)$$

Note the use of t_α (all error budget on one side) rather than $t_{\alpha/2}$.

8.7.1 Pivots

A **pivot** is a random variable that depends on the data and the parameter of interest, but whose probability distribution does *not* depend on any unknown parameters.

Definition 8.15 (Pivot)**.** A function $V : \mathcal{P} \times \mathcal{X} \to \mathbb{R}$ is a **pivot** for parameter γ if:

i) $V(\mathsf{P}, X) = U(\gamma(\mathsf{P}), X)$ for some function $U : \Gamma \times \mathcal{X} \to \mathbb{R}$. (It depends on the underlying distribution only through the parameter of interest).

ii) For all $\mathsf{P}_1, \mathsf{P}_2 \in \mathcal{P}$, the distribution of $V(\mathsf{P}_1, X)$ is the same as $V(\mathsf{P}_2, X)$. (The distribution is fixed/known).

Examples:

- **Z-Statistic (Known Variance):**

$$V = \frac{\overline{X}_n - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1).$$

- **T-Statistic (Unknown Variance):**

$$V = \frac{\overline{X}_n - \mu}{s(X)/\sqrt{n}} \sim t(n-1).$$

Constructing Confidence Sets Using Pivots

If V is a pivot, we can construct a confidence set by "inverting" the probability statement.

1. Choose a set A (based on the known distribution of V) such that:
$$\mathsf{P}(V(P, X) \in A) \geq 1 - \alpha.$$
2. Define the confidence set $S(x)$ as the set of all parameter values that keep the pivot inside A:
$$S(x) = \{\gamma(P) : V(P, x) \in A\}.$$

8.7.2 Duality: Confidence Sets and Hypothesis Tests

There is a fundamental equivalence between confidence sets and hypothesis tests. You can convert one into the other.

1. Confidence Sets → Tests

If $S(x)$ is a $(1 - \alpha)$ confidence set, we can define a level α test for $H_0 : \gamma(P) = \gamma_0$ as:
$$d_{\gamma_0}(x) = \begin{cases} 1(\text{Reject}) & \text{if } \gamma_0 \notin S(x), \\ 0(\text{Accept}) & \text{if } \gamma_0 \in S(x). \end{cases}$$

Intuition: If the hypothesized value γ_0 is not inside the "plausible range" (the confidence set), we reject it.

2. Tests → Confidence Sets (Test Inversion)

If we have a family of level α tests d_{γ_0} for every possible parameter value γ_0, we can build a confidence set by collecting all the values we do **not** reject:
$$S(x) = \{\gamma_0 \in \Gamma : d_{\gamma_0}(x) = 0\}.$$

The coverage probability is guaranteed:
$$\mathsf{P}(\gamma(\mathsf{P}) \in S(X)) = \mathsf{P}(d_{\gamma(\mathsf{P})}(X) = 0) \geq 1 - \alpha.$$

9. Preliminaries for Large Sample Theory

We review basic facts about stochastic convergence and limit theorems that we will draw on in our later discussions of large-sample theory for statistical methodology. For additional reading, consider the initial chapters of Ferguson (1996) and van der Vaart (1998).

9.1 Stochastic Convergence

Let X_n be a sequence of k–dimensional random vectors on a probability space $(\Omega, \mathcal{A}, \mathsf{P})$. We distinguish between different "modes" of how this sequence approaches a limit X.

1. Almost Sure Convergence ($\xrightarrow{a.s.}$)

This is the strongest form of pointwise convergence. The sequence converges for "all" outcomes ω, except possibly for a set of probability zero.

$$X_n \xrightarrow{a.s.} X \iff \mathsf{P}\left(\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\right) = 1$$

Characterization via Supremum:

$$X_n \xrightarrow{a.s.} X \iff \forall\epsilon > 0 : \mathsf{P}\left(\sup_{m\geqslant n} \|X_m - X\| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

2. Convergence in r–th Mean ($\xrightarrow{r}$)

Often used with $r = 2$ (Quadratic Mean). It measures convergence using the expected value of the distance raised to the power r.

$$X_n \xrightarrow{r} X \iff \mathsf{E}[\|X_n - X\|^r] \to 0 \quad \text{as} \quad n \to \infty.$$

- Requires $\mathsf{E}[\|X_n\|^r] < \infty$ and $\mathsf{E}[\|X\|^r] < \infty$.
- **Implication:** If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{r'} X$ for any $0 < r' \leq r$.

3. Convergence in Probability ($\xrightarrow{p}$)

This is the standard mode for statistical consistency. It demands that the probability of the estimator X_n deviating from X by any fixed amount ϵ goes to zero.

$$X_n \xrightarrow{p} X \iff \forall\epsilon > 0 : \mathsf{P}(\|X_n - X\| > \epsilon) \to 0 \text{ as } n \to \infty.$$

4. Convergence in Distribution ($\xrightarrow{d}$)

This is the weakest form. It does not require X_n and X to be close to each other value-wise; it only requires their *distribution functions* (CDFs) to become similar.

$$X_n \xrightarrow{d} X \iff F_n(x) \to F(x) \quad \forall x \text{ where } F \text{ is continuous.}$$

Note. This is the mode of convergence used in the Central Limit Theorem.

Relationships and Properties

Implications

- **Top Tier:** $\xrightarrow{a.s.}$ and $\xrightarrow{r}$ are the strongest. They do not imply each other directly (without extra conditions).
- **Middle Tier:** Both $\xrightarrow{a.s.}$ and $\xrightarrow{r}$ imply $\xrightarrow{p}$.
- **Bottom Tier:** $\xrightarrow{p}$ implies $\xrightarrow{d}$.

Useful Tools

1. Subsequence Characterization $X_n \xrightarrow{p} X$ if and only if every subsequence (n_k) contains a further sub-subsequence (n_{k_ℓ}) such that $X_{n_{k_\ell}} \xrightarrow{a.s.} X$.

> *Usage:* If you find proving $\xrightarrow{p}$ directly is hard, but you can prove $\xrightarrow{a.s.}$ for subsequences, you can use this equivalence.

2. Convergence to a Constant If the limit is a constant c (non-random), then convergence in distribution is equivalent to convergence in probability:

$$X_n \xrightarrow{d} c \iff X_n \xrightarrow{p} c$$

3. The Cramér-Wold Device (for $k > 1$) Used to prove convergence in distribution for random **vectors** by reducing the problem to scalar (1D) linear combinations.

$$X_n \xrightarrow{d} X \iff a^\top X_n \xrightarrow{d} a^\top X \quad \forall a \in \mathbb{R}^k$$

> *Why it works:* It relies on Characteristic Functions (Fourier transforms of distributions). If the characteristic functions match for all linear combinations $t(a^\top X)$, the joint distributions must match.

9.2 Limit Theorems

We consider an i.i.d. sequence $(X_i)_{i=1}^{\infty}$ with finite mean $\mu = \mathsf{E}[X_i]$. Let $\overline{X}_n$ be the sample mean.

1. Law of Large Numbers (LLN)

The LLN guarantees that sample averages converge to the population mean.

- **Strong LLN:** $\overline{X}_n \xrightarrow{a.s.} \mu$. (Convergence with probability 1).
- **Weak LLN:** $\overline{X}_n \xrightarrow{p} \mu$. (Convergence in probability).

Remark. Since $\xrightarrow{a.s.}$ implies $\xrightarrow{p}$, the Strong LLN implies the Weak LLN.

2. Central Limit Theorem (CLT)

The CLT describes the distribution of the error $\overline{X}_n - \mu$.
Univariate ($k = 1$): If $\sigma^2 = \mathsf{Var}[X_i] \in (0, \infty)$, then:

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Multivariate ($k > 1$): If the covariance matrix $\Sigma = \mathsf{Var}[X_i]$ exists, then:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}_k(0, \Sigma).$$

3. CLTs for Triangular Arrays

Used when the data generating distribution changes with n. Let $X_{n,1}, \dots, X_{n,n}$ be independent random variables for each row n, with:

$$\mu_{n,i} = \mathsf{E}[X_{n,i}] \quad \text{and} \quad \sigma_{n,i}^2 = \mathsf{Var}[X_{n,i}].$$

Let $\sigma_n^2 = \sum_{i=1}^{n} \sigma_{n,i}^2$ be the total variance of the sum in row n.
Lindeberg-Feller CLT: The standardized sum converges to a standard normal:

$$\frac{1}{\sigma_n} \sum_{i=1}^{n} (X_{n,i} - \mu_{n,i}) \xrightarrow{d} \mathcal{N}(0, 1)$$

*Provided the **Lindeberg Condition** holds:*

$$\forall \epsilon > 0 : \frac{1}{\sigma_n^2} \sum_{i=1}^{n} \mathsf{E}\left[(X_{n,i} - \mu_{n,i})^2 \cdot \mathbf{1}_{\{|X_{n,i} - \mu_{n,i}| > \sigma_n \epsilon\}}\right] \underset{n \to \infty}{\longrightarrow} 0.$$

Intuition: The contribution of the “tails” (extreme values) to the total variance must be negligible.

Lyapunov CLT (Simpler Condition): The Lindeberg condition is satisfied (and thus the CLT holds) if the **Lyapunov Condition** holds:

$$\frac{1}{\sigma_n^3}\sum_{i=1}^{n}\mathsf{E}[|X_{n,i}-\mu_{n,i}|^3] \underset{n\to\infty}{\longrightarrow} 0.$$

Usage: It is often easier to calculate the 3rd moments than to evaluate the indicator function in the Lindeberg condition.

The Skorokhod Representation Theorem

The **Skorokhod Representation Theorem** is a powerful theoretical tool. While convergence in distribution ($X_n \xrightarrow{d} X$) is the "weakest" form of convergence (implying only that CDFs converge), this theorem allows us to "upgrade" this convergence to almost sure convergence ($\xrightarrow{a.s.}$) by constructing new random variables on a different probability space.

Think of this as a "technical device" or a bridge. It allows us to prove properties about weak convergence ($\xrightarrow{d}$) using the stronger machinery of almost sure convergence ($\xrightarrow{a.s.}$). It essentially says: "If the distributions converge, I can simulate these random variables in a way that the values actually converge point-wise."

Theorem 9.1 (Skorokhod Representation)**.** *Suppose $X_n \xrightarrow{d} X$. Then there exist random vectors $\tilde{X}_n$ and $\tilde{X}$ defined on a common probability space $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$ such that:*

1. ***Distributions Match:*** *$\tilde{X}_n \overset{d}{=} X_n$ for all n, and $\tilde{X} \overset{d}{=} X$.*
2. ***Strong Convergence:*** *$\tilde{X}_n \xrightarrow{a.s.} \tilde{X}$ as $n \to \infty$.*

Proof (The Quantile Transform). For the 1–dimensional case ($k = 1$), the proof is constructive and relies on the **Probability Integral Transform**.

1. Take a single random source $U \sim \text{Unif}(0, 1)$.
2. Define the random variables using the quantile functions (generalized inverses) of the CDFs:
$$\tilde{X}_n := F_{X_n}^{-1}(U) \quad \text{and} \quad \tilde{X} := F_X^{-1}(U).$$
3. Since $X_n \xrightarrow{d} X$, we know $F_{X_n}(x) \to F_X(x)$ at continuity points. This implies the quantile functions converge point-wise.
4. Because $\tilde{X}_n$ and $\tilde{X}$ are built from the *exact same* U, they are highly dependent (coupled), which forces them to converge almost surely.

Warning: We lose all information about the joint relationships (e.g., independence) of the original sequence X_n. Skorokhod preserves the *marginal* distributions of X_n and X, but creates a specific dependence structure to force path-wise convergence.

For $k > 1$, the construction is significantly more complex; see Billingsley (1986). □

The Continuous Mapping Theorem (Mann–Wald)

This theorem answers a fundamental question:

> If X_n converges to X, does $g(X_n)$ converge to $g(X)$?

This is the stochastic analog of the definition of continuity. For deterministic numbers, continuity means $x_n \to x \implies g(x_n) \to g(x)$. This theorem states that this preservation of convergence holds for our three stochastic modes.

Theorem 9.2 (Continuous Mapping Theorem / Mann–Wald)**.** *Let $g : \mathbb{R}^k \to \mathbb{R}^m$ be a function that is continuous almost everywhere with respect to the distribution of X (i.e., $\mathsf{P}(X \in D_g) = 0$, where D_g is the set of discontinuity points of g). Then:*

1. $X_n \xrightarrow{a.s.} X \implies g(X_n) \xrightarrow{a.s.} g(X)$
2. $X_n \xrightarrow{p} X \implies g(X_n) \xrightarrow{p} g(X)$
3. $X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$

Proof.

1. **Almost Sure Convergence ($\xrightarrow{a.s.}$):** There is nothing to do.
   - By definition, there exists a set A with $\mathsf{P}(A) = 1$ where $X_n(\omega) \to X(\omega)$ pointwise.
   - By the definition of continuity, for every $\omega \in A$, $g(X_n(\omega)) \to g(X(\omega))$.

   Thus, $g(X_n) \xrightarrow{a.s.} g(X)$.

2. **Convergence in Probability ($\xrightarrow{p}$):** Use the Subsequence Characterization.
   - Instead of struggling with ϵ-δ definitions for probability, recall: $Y_n \xrightarrow{p} Y$ if and only if every subsequence has a further sub-subsequence that converges almost surely.
   - Take any subsequence of X_n. It has a sub-subsequence converging a.s.
   - By result (1), applying g preserves this a.s. convergence.

   Therefore, $g(X_n) \xrightarrow{p} g(X)$.

3. **Convergence in Distribution ($\xrightarrow{d}$):** Use Skorokhod as a bridge.
   - Proving this directly with integrals and CDFs is messy.
   - Instead, apply the **Skorokhod Representation Theorem**.
   - Switch from $X_n \xrightarrow{d} X$ to the coupled version $\tilde{X}_n \xrightarrow{a.s.} \tilde{X}$.
   - Apply result (1) (continuous mapping for a.s.) to get $g(\tilde{X}_n) \xrightarrow{a.s.} g(\tilde{X})$.
   - Since almost sure convergence implies convergence in distribution, we have $g(\tilde{X}_n) \xrightarrow{d} g(\tilde{X})$.
   - Since distributions are preserved ($\tilde{X}_n \overset{d}{=} X_n$), we conclude $g(X_n) \xrightarrow{d} g(X)$.

□

Slutsky's Theorem

Slutsky's theorem is a workhorse in asymptotic statistics. It allows us to perform algebraic manipulations on sequences of random variables (like addition and multiplication) while preserving their convergence properties, *provided* that one of the sequences converges to a constant.

Theorem 9.3 (Slutsky's Theorem)**.** *Let $X_n \xrightarrow{d} X$.*

a) ***Univariate case*** *($k = 1$)**:** If $A_n \xrightarrow{p} a$ and $B_n \xrightarrow{p} b$ for constants $a, b \in \mathbb{R}$, then the affine transformation converges:*

$$A_n X_n + B_n \xrightarrow{d} aX + b.$$

b) ***General case*** *($k \geq 1$)**:** If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ for a constant vector c, then the joint vector converges in distribution:*

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}.$$

Consequently, by the Continuous Mapping Theorem, for any continuous function g:

$$g(X_n, Y_n) \xrightarrow{d} g(X, c).$$

The "Stacking" Argument

Why is the condition $Y_n \xrightarrow{p} c$ (constant) so important?

- **The Problem:** generally, if $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$ (random), we know **nothing** about the convergence of the pair (X_n, Y_n). They could be independent, perfectly correlated, or have a dependency structure that changes with n. Without knowing their joint dependence, we cannot determine the limit of $g(X_n, Y_n)$.
- **The Solution:** If Y_n converges to a *constant* c, the dependence "decouples" in the limit. The "stacking" of a random vector X_n and a vector converging to a constant always converges jointly. This allows us to treat Y_n as if it were the constant c for the purposes of asymptotic algebra.

  Practical use in Statistics: We often derive a limit theorem involving a true parameter (e.g., standardizing by σ). In practice, we must replace σ with an estimator S_n. Slutsky's theorem guarantees that as long as S_n is **consistent** ($S_n \xrightarrow{p} \sigma$), the asymptotic distribution of our test statistic remains unchanged.

Example 9.1 (Asymptotic Normality of the t–Statistic)**.** This is the classic application of Slutsky's theorem. We want to show that replacing the true standard deviation σ with the sample standard deviation S_n in the CLT does not change the limit distribution.

Setup: Let $X_1, X_2, \ldots$ be i.i.d. with $\mathsf{E}[X_i] = \mu$ and $\mathsf{Var}[X_i] = \sigma^2 \in (0, \infty)$.

Consistency of Sample Moments (LLN)

By the Weak Law of Large Numbers (WLLN):

$$\overline{X}_n \xrightarrow{p} \mu \quad \text{and} \quad \overline{X_n^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathsf{E}[X_i^2] = \mu^2 + \sigma^2.$$

Since both converge in probability to constants, their joint vector converges:

$$\left(\overline{X}_n, \overline{X_n^2}\right) \xrightarrow{p} (\mu, \mu^2 + \sigma^2).$$

Consistency of Variance Estimator (Continuous Mapping)

We express the sample variance using the identity:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \overline{X_n^2} - (\overline{X}_n)^2.$$

Define the function $g(u, v) = v - u^2$, which is continuous. Applying the Continuous Mapping Theorem to the result from Step 1:

$$\overline{X_n^2} - (\overline{X}_n)^2 \xrightarrow{p} (\mu^2 + \sigma^2) - (\mu)^2 = \sigma^2.$$

Remark (on Bias Correction). We typically use $S^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$. Since $\frac{n}{n-1} \to 1$, the factor is asymptotically negligible. Thus:

$$S_n^2 \xrightarrow{p} \sigma^2 \implies S_n \xrightarrow{p} \sigma.$$

This formally proves that S_n is a consistent estimator for σ.

The t–Statistic (CLT + Slutsky)

We want the limit of the t-statistic: $T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n}$. We can rewrite this as a product:

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} = \underbrace{\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}}_{\xrightarrow{d}\mathcal{N}(0,1) \text{ by CLT}} \cdot \underbrace{\frac{\sigma}{S_n}}_{\xrightarrow{p}1 \text{ by CMT}}.$$

- The first term converges in distribution to $\mathcal{N}(0, 1)$ by the Central Limit Theorem.
- The second term converges in probability to 1 because $S_n \xrightarrow{p} \sigma$.

Applying Slutsky's Theorem (product rule):

$$T_n \xrightarrow{d} \mathcal{N}(0, 1) \cdot 1 = \mathcal{N}(0, 1).$$

Note. This confirms that for large samples, we can effectively "ignore" the fact that we estimated the variance; the t–statistic behaves exactly like a standard normal variable.

9.3 The Delta Method

While the Continuous Mapping Theorem tells us that applying a continuous function preserves convergence (i.e., $X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X)$), it does not tell us about the *rate* of convergence or the resulting distribution of the fluctuations.

Error Propagation Think of this as the statistical equivalent of "error propagation" in numerical analysis. Suppose X_n is an estimator for a parameter b. As $n \to \infty$, the estimator becomes more accurate, meaning the error $X_n - b \to 0$. However, if we scale this error by a sequence a_n (often $\sqrt{n}$), the scaled error stabilizes to a limit distribution Z (e.g., a standard normal). The Delta Method answers the question:

> If we transform our estimator via a function g, how do the estimation errors of $g(X_n)$ behave relative to the target $g(b)$?

The method relies on a **linearization** (first-order Taylor approximation) of the function g. If g is differentiable, linearizing it locally around the parameter b allows us to transfer the central limit behavior of X_n to $g(X_n)$.

Theorem 9.4 (The Delta Method / Cramér's Theorem)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random vectors in $\mathbb{R}^m$. Let $a_n \in \mathbb{R}$ and $b \in \mathbb{R}^m$ be such that $a_n \to \infty$ and*

$$Z_n := a_n(X_n - b) \xrightarrow{d} Z.$$

If $g : \mathbb{R}^m \to \mathbb{R}^k$ is differentiable at b, with Jacobian matrix $g'(b) \in \mathbb{R}^{k\times m}$, then

$$a_n(g(X_n) - g(b)) \xrightarrow{d} g'(b)Z.$$

Note (on Dimensions)**.** If g maps $\mathbb{R}^m \to \mathbb{R}^k$, the Jacobian $g'(b)$ is a $k \times m$ matrix. The limit is the random vector Z linearly transformed by this matrix.

Note (The Normal Case)**.** The most common application is when the limit Z is multivariate normal. If $Z \sim \mathcal{N}_m(0, \Sigma)$, then by the properties of linear transformations of Gaussian vectors:

$$g'(b)Z \sim \mathcal{N}_k(0, g'(b)\Sigma g'(b)^T).$$

In this specific context, the result is often referred to as *Cramér's Theorem.*

Proof. We consider $m = 1$ and show an explicit difference quotient. The case $m > 1$ is proven analogously; see van der Vaart (1998, Theorem 3.1).

Skorokhod Representation. Working directly with convergence in distribution is abstract because the random variables as functions can behave erratically. To make the analysis "pedestrian" (standard calculus), we use the Skorokhod representation theorem. There exist $\tilde{X}_n, \tilde{Z}_n, \tilde{Z}$ on a common probability space such that:

- $\tilde{Z}_n \overset{d}{=} Z_n$ and $\tilde{Z} \overset{d}{=} Z$.

- $\tilde{Z}_n \xrightarrow{a.s.} \tilde{Z}$.

We define $\tilde{X}_n$ via the relation $\tilde{Z}_n = a_n(\tilde{X}_n - b)$. Since $\tilde{Z}_n$ converges to a finite limit and $a_n \to \infty$, it must be that $\tilde{X}_n \xrightarrow{a.s.} b$.

Linearization. We can now analyze the difference quotient directly using pathwise convergence.

$$a_n\big(g(\tilde{X}_n) - g(b)\big) = \begin{cases} \frac{g(\tilde{X}_n)-g(b)}{\tilde{X}_n-b}(\tilde{X}_n - b)a_n, & \text{if } \tilde{X}_n \neq b \\ 0, & \text{if } \tilde{X}_n = b \end{cases}$$

As $n \to \infty$:

- The term $(\tilde{X}_n - b)a_n = \tilde{Z}_n \xrightarrow{a.s.} \tilde{Z}$.
- Since $\tilde{X}_n \to b$ and g is differentiable, the difference quotient $\frac{g(\tilde{X}_n)-g(b)}{\tilde{X}_n-b} \xrightarrow{a.s.} g'(b)$.

Thus, the product converges almost surely to $g'(b) \cdot \tilde{Z}$. Since distributional convergence is preserved under this representation,

$$a_n(g(X_n) - g(b)) \xrightarrow{d} g'(b)Z.$$

□

Example 9.2 (Squared Sample Mean). Let $X_1, \dots, X_n$ be i.i.d. with $\mathsf{E}[X_i] = \mu$ and $\mathsf{Var}[X_i] = \sigma^2 < \infty$. By the CLT, the estimation error of the mean behaves as:

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

We wish to estimate μ^2 using the estimator $(\overline{X}_n)^2$.

Case 1: $\mu \neq 0$ (Standard Delta Method) Let $g(x) = x^2$. The derivative is $g'(x) = 2x$, so at the point of interest, $g'(\mu) = 2\mu$. Applying the Delta Method:

$$\sqrt{n}\Big((\overline{X}_n)^2 - \mu^2\Big) \xrightarrow{d} \mathcal{N}(0, \underbrace{(2\mu)^2\sigma^2}_{\text{Variance}}).$$

Interpretation: The limiting variance $4\mu^2\sigma^2$ comes from two sources:

1. σ^2: The inherent error in estimating the mean itself.
2. $(2\mu)^2$: The "sensitivity" of the squaring function. Small errors in $\overline{X}_n$ are amplified by the slope 2μ.

Case 2: $\mu = 0$ (Singularity / Second-Order Delta Method) If $\mu = 0$, the derivative $g'(\mu) = 2(0) = 0$. The Delta Method gives:

$$\sqrt{n}((\overline{X}_n)^2 - 0) \xrightarrow{d} 0 \cdot Z = 0.$$

This means the estimator is "super-accurate"—the errors vanish faster than $1/\sqrt{n}$. This happens because the function x^2 is very flat near 0; small deviations in estimation result in quadratically smaller deviations in the output.

To get a non-trivial distribution, we must scale more aggressively. Since $\sqrt{n}$ yields 0, we try scaling by n:

$$n(\overline{X}_n)^2 = \left(\sqrt{n}(\overline{X}_n - 0)\right)^2.$$

From the CLT, we know $\sqrt{n}\overline{X}_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. We can rewrite this as σZ, where $Z \sim \mathcal{N}(0, 1)$. By the Continuous Mapping Theorem:

$$n(\overline{X}_n)^2 \xrightarrow{d} (\sigma Z)^2 = \sigma^2 Z^2 \sim \sigma^2 \chi^2_1.$$

Key Takeaway: At a point where $g'(b) = 0$ (a singularity), the linear approximation fails. We essentially rely on a quadratic approximation, requiring a scaling of a_n^2 (here n) instead of a_n (here $\sqrt{n}$), often resulting in a Chi–square type distribution.

Example 9.3 (Reciprocal of the Sample Mean)**.** Assume $\mu \neq 0$. Consider the estimator $g(\overline{X}_n) = \frac{1}{\overline{X}_n}$ for the quantity $1/\mu$. The derivative is $g'(x) = -1/x^2$. Applying the Delta Method with $g'(\mu) = -1/\mu^2$:

$$\sqrt{n}\left(\frac{1}{\overline{X}_n} - \frac{1}{\mu}\right) \xrightarrow{d} \mathcal{N}\left(0, \left(-\frac{1}{\mu^2}\right)^2 \sigma^2\right) = \mathcal{N}\left(0, \frac{\sigma^2}{\mu^4}\right).$$

Qualitative Check:

- If σ^2 is small, the input is accurate, so the output variance is small.
- If μ is close to 0, $1/x$ is very steep (high derivative). Small errors in $\overline{X}_n$ lead to massive errors in $1/\overline{X}_n$. This matches the $\frac{1}{\mu^4}$ term in the variance, which blows up as $\mu \to 0$.

Note. This result holds purely via asymptotic error propagation. It does not require the expectation $\mathsf{E}[1/\overline{X}_n]$ to exist (indeed, for normal variables, the inverse mean has no first moment). This is an "apples and oranges" comparison: the Delta Method describes the *limiting distribution* of the estimator, not the convergence of its moments.

9.4 Stochastic Landau Symbols

Just as Landau symbols ($o(1)$, $O(1)$) are essential in deterministic analysis for handling error terms and convergence rates, we define their stochastic counterparts to streamline asymptotic arguments in statistics. The subscript p denotes that the property holds "in probability."

Definition 9.1 (Little-o in probability)**.** We write $X_n = o_p(1)$ if the sequence converges to zero in probability:

$$X_n \xrightarrow{p} 0.$$

Intuition: This is the stochastic equivalent of a sequence of numbers converging to zero.

Definition 9.2 (Big-O in probability)**.** We write $X_n = O_p(1)$ if the sequence is **bounded in probability** (also called **uniformly tight**). Formally:

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathsf{P}(\|X_n\| \geq M) \longrightarrow 0$$

Equivalently,

$$\forall \epsilon > 0 \, \exists M_\epsilon > 0 \; : \; \sup_{n\in\mathbb{N}} \mathsf{P}(\|X_n\| > M_\epsilon) < \epsilon.$$

For a single random variable, we know that probability mass is tight—we can always find a bounds $[-M, M]$ that contain 99% of the mass. For a *sequence* X_n, "boundedness in probability" means we can find a **single** M that works for **all** n simultaneously.

- If the sequence "drifts off" to infinity (mass escapes), it is not bounded.
- **Key Result:** If $X_n \xrightarrow{d} X$, then $X_n = O_p(1)$. Just as a convergent sequence of numbers is bounded, a sequence converging in distribution is tight.

General Orders

We extend these definitions to compare a random sequence X_n against a deterministic rate sequence R_n (e.g., $R_n = n^{-1/2}$).

- $X_n = o_p(R_n)$ means $X_n = Y_n R_n$ where $Y_n = o_p(1)$.
- $X_n = O_p(R_n)$ means $X_n = Y_n R_n$ where $Y_n = O_p(1)$.

Informally: $X_n = o_p(R_n)$ means $X_n/R_n \xrightarrow{p} 0$.

Calculus of Stochastic Symbols

These symbols allow us to manipulate "remainders" in proofs algebraically. These equations should be read directionally (from left to right): "A sequence with the property on the left also has the property on the right."

i) $o_p(1) \implies O_p(1)$:

If a sequence goes to zero in probability, it is necessarily bounded in probability. The converse is false.

ii) $o_p(1) + o_p(1) = o_p(1)$:

The sum of two sequences converging to zero also converges to zero.

iii) $o_p(1) \cdot O_p(1) = o_p(1)$

If you multiply something bounded (O_p) by something shrinking to zero (o_p), the result shrinks to zero. (This is Slutsky's theorem in disguise).

iv) $O_p(1) + O_p(1) = O_p(1)$

The sum of two bounded sequences is bounded.

Recommended Literature

For further details on these asymptotic tools:

- **Ferguson (1996):** *A Course in Large Sample Theory.* (The lecturer's "favorite" for this topic).
- **van der Vaart (1998):** *Asymptotic Statistics.* (More general and advanced).
- **Billingsley:** For probability refreshers and convergence results (e.g., Skorokhod).

10. Large Sample Theory for Moment Estimators

This chapter discusses the asymptotic distribution of estimators obtained through the Method of Moments (MOM). From a mathematical perspective, the treatment here is straightforward: we simply combine the standard **Central Limit Theorem (CLT)** and the **Delta Method**.

Motivation Suppose you live in 1870. You are an informed statistician, and you know one thing for sure: **how to estimate expectations**. The Law of Large Numbers (LLN) guarantees that sample averages converge to true expectations. However, not every estimation problem presents itself directly as an expectation. The Method of Moments (introduced in Chapter 2) is essentially a "vehicle" for relating arbitrary estimation tasks *back* to the estimation of expectations. Once we do that, we can leverage the powerful tools of the LLN and CLT.

10.1 Estimating an Expectation

The foundation of the Method of Moments is the estimation of a simple expectation.

Setup: Let $X_1, X_2, \dots$ be i.i.d. observations from the statistical model:

$$\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}.$$

Let h be a function such that $\mathsf{E}_\theta[|h(X)|] < \infty$ for all $\theta \in \Theta$. We consider the estimation of the statistical parameter defined by the expectation of this function:

$$\gamma(\theta) = \mathsf{E}_\theta[h(X)].$$

Consistency (Law of Large Numbers)

We employ the "plug-in" estimator, which replaces the population expectation with the sample average:

$$\hat{\gamma}_n := \frac{1}{n}\sum_{i=1}^{n} h(X_i).$$

By the Weak Law of Large Numbers (LLN), this estimator converges in probability to the true parameter:

$$\hat{\gamma}_n \xrightarrow{p} \gamma(\theta) \quad \forall \theta \in \Theta.$$

Terminology: We say that $\hat{\gamma}_n$ is a **consistent estimator** of $\gamma(\theta)$. As you collect more data ($n \to \infty$), the estimator converges to the target.

Asymptotic Normality (Central Limit Theorem)

To understand the *estimation error*, we look to the Central Limit Theorem. If the variance is finite (i.e., $\mathsf{Var}_\theta[h(X)] < \infty$), the CLT implies that the standardized error follows a standard normal distribution asymptotically:

$$\sqrt{n}(\hat{\gamma}_n - \gamma(\theta)) \xrightarrow{d} \mathcal{N}(0, \mathsf{Var}_\theta[h(X)]).$$

This result serves as the building block for more complex Method of Moments estimators, where we will estimate parameters θ that are functions of these expectations $\gamma(\theta)$.

10.2 Method of Moments (General Formulation)

While the introductory Method of Moments (section 1.3) typically uses simple powers ($T(x) = x, x^2, \dots$), the general theory allows for any set of chosen "helper functions."

To estimate a parameter vector $\boldsymbol{\theta} \in \mathbb{R}^k$, we choose k functions $T_1, \dots, T_k$. We define the vector of these statistics as:

$$T(X) = (T_1(X), \dots, T_k(X))^\top.$$

We then define the **theoretical expectation map** τ, which links the parameters to the moments:

$$\tau(\boldsymbol{\theta}) = \mathsf{E}_{\boldsymbol{\theta}}[T(X)] = \begin{pmatrix} \mathsf{E}_{\boldsymbol{\theta}}[T_1(X)] \\ \vdots \\ \mathsf{E}_{\boldsymbol{\theta}}[T_k(X)] \end{pmatrix}.$$

The Estimator Construction: The Method of Moments (MOM) operates by matching the theoretical expectations to the empirical data averages. Let $\overline{T}_n$ be the vector of sample means:

$$\overline{T}_n = \frac{1}{n} \sum_{i=1}^{n} T(X_i).$$

We obtain the estimator $\hat{\boldsymbol{\theta}}_n$ by solving the system:

$$\overline{T}_n = \tau(\hat{\boldsymbol{\theta}}_n).$$

If the mapping τ is injective (one-to-one), we can explicitly define the estimator using the inverse mapping τ^{-1}:

$$\hat{\boldsymbol{\theta}}_n = \tau^{-1}(\overline{T}_n).$$

Remark. This estimator is well-defined provided the sample average $\overline{T}_n$ falls within the image of τ.

Asymptotic Distribution (Delta Method Applied to MOM)

Since the estimator is defined via a differentiable transformation of sample means ($\hat{\boldsymbol{\theta}}_n = \tau^{-1}(\overline{T}_n)$), its asymptotic behavior follows directly from the Delta Method combined with the Central Limit Theorem.

Theorem 10.1 (Asymptotic Normality of MOM)**.** *Let $X_1, X_2, \ldots$ be i.i.d. observations generated by P_{θ_0}. Assume the following conditions:*

i) ***Finite Variance:*** *The chosen statistics have finite second moments, i.e., $\mathsf{E}_{\theta_0}[\|T(X)\|_2^2] < \infty$.*

ii) ***Invertibility:*** *The mapping τ is injective and continuously differentiable in an open neighborhood of θ_0.*

iii) ***Non-Singularity:*** *The Jacobian matrix $\tau'(\boldsymbol{\theta}) = \left(\frac{\partial \tau_j(\boldsymbol{\theta})}{\partial \theta_l}\right)_{jl} \in \mathbb{R}^{k\times k}$ is invertible at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.*

Then, the Method of Moments estimator $\hat{\boldsymbol{\theta}}_n$ exists with probability tending to 1 as $n \to \infty$, and satisfies:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_k\Big(0, \Sigma_{MOM}\Big),$$

where the asymptotic covariance matrix is:

$$\Sigma_{MOM} = \tau'(\boldsymbol{\theta}_0)^{-1} \operatorname{Var}_{\boldsymbol{\theta}_0}[T(X)]\, \tau'(\boldsymbol{\theta}_0)^{-\top}.$$

Anatomy of the Variance

The asymptotic variance formula can be understood as having two distinct components:

1. **Inherited Variance ($\operatorname{Var}_{\theta_0}[T(X)]$):** This is the "middle" term. It represents the unavoidable error in estimating the expectations of the helper functions T. If $T(X)$ has high variance, your input to the equation system is noisy.
2. **Error Propagation ($\tau'(\theta_0)^{-1}$):** This term (and its transpose) comes from the Delta Method. It represents the sensitivity of the inverse map τ^{-1}. It answers: "How much does the solution $\hat{\theta}$ wiggle if the moments $\overline{T}_n$ wiggle?" If the derivative of τ is small (flat), the inverse derivative is large, amplifying the error.

TODO: improve this proof

Proof (Application of Inverse Function Theorem). By the **Inverse Function Theorem**, if τ is differentiable with an invertible Jacobian at $\boldsymbol{\theta}_0$, then locally around $\tau(\boldsymbol{\theta}_0)$, the inverse function τ^{-1} exists and is differentiable. Crucially, the derivative of the inverse is the inverse of the derivative:

$$(\tau^{-1})'(t)\Big|_{t=\tau(\boldsymbol{\theta}_0)} = [\tau'(\boldsymbol{\theta}_0)]^{-1}.$$

By the **CLT**, the sample moments are asymptotically normal:

$$\sqrt{n}(\overline{T}_n - \tau(\theta_0)) \xrightarrow{d} Z \sim \mathcal{N}_k(0, \mathsf{Var}_{\theta_0}[T(X)]).$$

By the **Delta Method** applied to the function $g = \tau^{-1}$:

$$\sqrt{n}(\tau^{-1}(\overline{T}_n) - \tau^{-1}(\tau(\theta_0))) \xrightarrow{d} [\tau'(\theta_0)]^{-1} Z.$$

The covariance of the linear transformation AZ is $A\,\mathrm{Cov}(Z)A^\top$, yielding the result. □

10.3 Application to Exponential Families

Equivalence of MOM and MLE

We now consider the specific case where the data follows an **exponential family**. This setting is particularly elegant because there is a natural connection between the Method of Moments (MOM) and Maximum Likelihood Estimation (MLE).

Consider an exponential family with densities given by:

$$p_{\boldsymbol{\theta}}(x) = \exp\{\langle \boldsymbol{\theta}, T(x)\rangle - A(\boldsymbol{\theta})\}h(x), \quad \boldsymbol{\theta} \in \boldsymbol{\theta}.$$

Here, $\boldsymbol{\theta}$ represents the canonical parameters. Suppose the parameter space Θ is open.

Injectivity of the Mean Map: Recall from the properties of exponential families that the cumulant generating function $A(\theta)$ is strictly convex. Consequently, its gradient—which defines the **mean parametrization** τ—is an injective (one-to-one) mapping:

$$\tau(\boldsymbol{\theta}) := \mathsf{E}_{\boldsymbol{\theta}}[T(X)] = \nabla A(\boldsymbol{\theta}).$$

Because this mapping is injective, we can uniquely estimate θ using the Method of Moments estimator based on the sufficient statistics $T(X)$:

$$\hat{\boldsymbol{\theta}}_{\text{MOM}} = \tau^{-1}(\overline{T}_n).$$

Claim. In this setting, the Method of Moments estimator is identical to the Maximum Likelihood Estimator. That is,

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}|\boldsymbol{x}).$$

Proof. Since Θ is open, the MLE (if it exists) must satisfy the first-order condition $\dot{\ell}_n(\theta|\boldsymbol{x}) = 0$. The log-likelihood function for n i.i.d. observations is:

$$\begin{aligned}
\ell_n(\boldsymbol{\theta}|\boldsymbol{X}) &= \sum_{i=1}^n \log p_\theta(X_i) \\
&= \sum_{i=1}^n [\langle \boldsymbol{\theta}, T(X_i)\rangle - A(\boldsymbol{\theta}) + \log h(X_i)] \\
&= n\langle \boldsymbol{\theta}, \overline{T}_n\rangle - nA(\boldsymbol{\theta}) + \text{const.}
\end{aligned}$$

Intuition: The logarithm "eats" the exponential, exposing the linear structure of the sufficient statistics and the cumulant function. The score function (gradient) is:

$$\nabla_\theta \ell_n(\boldsymbol{\theta}|\boldsymbol{X}) = n\overline{T}_n - n\nabla A(\boldsymbol{\theta}).$$

Recalling that $\nabla A(\theta) = \tau(\theta)$, the condition for the gradient to be zero is:

$$n\left[\overline{T}_n - \tau(\boldsymbol{\theta})\right] = 0 \iff \overline{T}_n = \tau(\boldsymbol{\theta}).$$

Thus, solving for the MLE is mathematically equivalent to solving the Method of Moments equation. □

Fisher Information Gives Asymptotic Variance

Since we have established that $\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MOM}}$, we can apply the Delta Method results from the previous section to derive the asymptotic distribution of the MLE.

Corollary 10.1 (Asymptotic Normality of MLE)**.** *In the considered exponential family setting, the MLE $\hat{\boldsymbol{\theta}}_n$ exists with probability tending to 1 as $n \to \infty$, and satisfies:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_k(0, I(\boldsymbol{\theta}_0)^{-1}),$$

where $I(\boldsymbol{\theta}_0)$ is the Fisher Information matrix for a single observation.

Derivation: The "Sandwich" Cancellation

The general Method of Moments variance (from Theorem 10.1) takes the form of a "sandwich":

$$\text{Asymp. Var} = \underbrace{\tau'(\theta_0)^{-1}}_{\text{Bread}} \underbrace{\text{Var}_{\theta_0}[T(X)]}_{\text{Meat}} \underbrace{\tau'(\theta_0)^{-\top}}_{\text{Bread}}.$$

In the exponential family context, a miraculous simplification occurs:

1. **The Meat (Variance):** The variance of the sufficient statistic is the Hessian of the cumulant function:

$$\text{Var}_{\boldsymbol{\theta}_0}[T(X)] = \nabla^2 A(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0).$$

2. **The Bread (Jacobian):** The mapping $\tau(\boldsymbol{\theta})$ is the gradient $\nabla A(\boldsymbol{\theta})$. Therefore, its Jacobian $\tau'(\boldsymbol{\theta})$ is the Hessian $\nabla^2 A(\boldsymbol{\theta})$:

$$\tau'(\boldsymbol{\theta}_0) = \nabla^2 A(\boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0).$$

Substituting these into the sandwich formula:

$$I(\boldsymbol{\theta}_0)^{-1} \cdot I(\boldsymbol{\theta}_0) \cdot I(\boldsymbol{\theta}_0)^{-1} = I(\boldsymbol{\theta}_0)^{-1}.$$

This confirms that the MLE achieves the **Cramér–Rao bound** asymptotically. While the estimator may have finite-sample bias, it reaches the optimal variance level as $n \to \infty$.

From Theory to Practice: Software and Approximation

The theoretical result $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathcal{N}(0, I(\boldsymbol{\theta}_0)^{-1})$ motivates the approximation:

$$\hat{\boldsymbol{\theta}} \mathrel{\dot\sim} \mathcal{N}\left(\boldsymbol{\theta}_0, \frac{1}{n} I(\boldsymbol{\theta}_0)^{-1}\right).$$

However, this is not directly usable because the true parameter $\boldsymbol{\theta}_0$ (and thus the true variance) is unknown.

Implementation Strategy (Slutsky's Theorem): In practice (and in statistical software), we replace the unknown information matrix $I(\boldsymbol{\theta}_0)$ with a consistent estimator, typically the observed information at the estimated value, $\hat{I} = I(\hat{\boldsymbol{\theta}}_n)$. By Slutsky's Theorem, the convergence still holds:

$$\sqrt{n}\, \hat{\boldsymbol{I}}(\hat{\boldsymbol{\theta}}_n)^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_k(0, I_k).$$

This makes MLE extremely "software-friendly." An optimizer finds the peak of the log-likelihood ($\hat{\boldsymbol{\theta}}$), and the Hessian at that peak automatically provides the estimated Fisher Information, which yields standard errors and confidence intervals.

Note (on Dependent Data)**.** The standard derivation assumes i.i.d. data, where the total information is $I_n(\boldsymbol{\theta}) = nI_1(\boldsymbol{\theta})$. For dependent data (e.g., Markov Chains), the Fisher Information does not necessarily scale linearly with n. However, generalized theorems exist (see Billingsley, 1961) showing that $\hat{\boldsymbol{\theta}} \mathrel{\dot\sim} \mathcal{N}(\boldsymbol{\theta}_0, I_{\text{total}}^{-1})$, provided one calculates the total information inherent in the dependent sequence correctly.

Invariance: If the model uses a different parametrization $\boldsymbol{\lambda}$ (via a diffeomorphism $\boldsymbol{\theta}(\boldsymbol{\lambda})$), the asymptotic normality is preserved. The asymptotic variance becomes $I(\boldsymbol{\lambda}_0)^{-1}$, which can be calculated using the Delta Method and the appropriate Jacobians.

11. Pearson's Chi-Square

In this chapter, we explore Pearson's Chi-Square statistic/test from an asymptotic point of view and derive the "usual" chi-square approximation that is typically used to set critical values and compute p-values. We also demonstrate how useful insights on the power of the test can be obtained from asymptotics along local alternatives. For further reading, see Ferguson (1996, chaps 9-10).

Motivation: Is the Dice Fair?

Consider the canonical problem: you are at a casino, and someone hands you a die. You want to know, is this die **fair**? That is, does each of the six faces show up with equal probability 1/6?

To test this, you roll the die $n = 100$ times and observe the counts for each face. Suppose you get a sequence of counts like $20, 10, 25, \ldots$. If the die were truly fair, you would expect each face to appear $100/6 \approx 16.7$ times. The observed vector of counts differs from the expected vector of counts. The core statistical question is: *How different is too different?*

We need to calculate the probability of seeing such a deviation (or a more extreme one) under the assumption that the die is fair. This is the rationale behind p-values. If this probability is very low, it casts doubt on the fairness of the die.

Example 11.1 (Random Number Generation)**.** Another practical application arises in computational statistics. Suppose you have an algorithm designed to generate random numbers from a standard normal distribution. Since computer-generated numbers are deterministic (pseudo-random), you might want to verify their statistical properties. One way to check this is to discretize the problem:

1. Bin the real line into intervals.
2. Calculate the theoretical probability of a standard normal variable landing in each interval.
3. Generate n samples using your algorithm and count how many fall into each bin.

This reduces the problem to the same structure as the dice roll—comparing observed counts in bins to expected counts derived from the theoretical probabilities. This is often called a "goodness-of-fit" test.

Note (on Distances)**.** When comparing observed counts to expected counts, we are essentially measuring a distance between two vectors. However, simple Euclidean distance might not be appropriate. Consider the following:

- **Scenario A:** Expected 16, Observed 20.
- **Scenario B:** Expected 0, Observed 4.

In both cases, the absolute difference is 4. However, Scenario B represents a much more drastic deviation (an impossible event becoming possible) than Scenario A. Similarly, a deviation of 4 is large if the expected count is small, but negligible if the expected count is 1000. Pearson's Chi-Square addresses this by looking at distances relative to the expected magnitude (perturbing the coordinate system).

11.1 Background on Normal and Chi–Square Distributions

Before diving into the Chi-Square test itself, let's establish some fundamental facts about normal distributions. Many statistical problems, when the sample size n is pushed to infinity, approximate problems involving normal distributions.

11.1.1 Multivariate Normal Distribution

If $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$ is a multivariate normal vector in $\mathbb{R}^k$, it is fully characterized by its mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. An important property is stability under affine transformations. If A is a matrix and $\boldsymbol{b}$ is a vector of suitable dimensions, then:

$$A\boldsymbol{X} + \boldsymbol{b} \sim \mathcal{N}_d(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^\top).$$

Proof (Sketch). One definition of a multivariate normal distribution is that it is the joint distribution of linear combinations of independent standard normal variables. Since linear combinations of linear combinations remain linear combinations, the affine transformation of a normal vector remains normal. The parameters follow from linearity of expectation and the definition of covariance. □

11.1.2 Chi-Square Distribution

Let $X_1, \ldots, X_k$ be i.i.d. $\mathcal{N}(0, 1)$. The chi-square distribution with k degrees of freedom, denoted χ^2_k, is defined as the distribution of the sum of their squares:

$$\sum_{j=1}^{k} X_j^2 \sim \chi^2_k.$$

In vector notation, if $\boldsymbol{X} = (X_1, \ldots, X_k)^\top \sim \mathcal{N}_k(0, I_k)$ represents a "standard normal point" in $\mathbb{R}^k$, then its squared Euclidean norm follows a chi-square distribution:

$$\|\boldsymbol{X}\|_2^2 \sim \chi_k^2.$$

This distribution is well-studied; we have analytical expressions for its density and efficient numerical methods for computing probabilities.

11.1.3 Mahalanobis Distance

We are interested in measuring the distance between a random vector $\boldsymbol{X}$ and its mean $\boldsymbol{\mu}$. If the coordinates of $\boldsymbol{X}$ are highly correlated, a simple Euclidean distance might be misleading. To account for correlations, we use the covariance matrix to "standardize" the distance.

Lemma 11.1 (Mahalanobis Distance)**.** *If $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, \Sigma)$ and Σ is invertible, then the quadratic form*

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_k^2.$$

Proof. Since Σ is positive definite, we can decompose it (e.g., via Cholesky decomposition) as $\Sigma = LL^\top$, where L is invertible. Then $\Sigma^{-1} = (L^{-1})^\top L^{-1}$. Substituting this into the quadratic form:

$$(\boldsymbol{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = (\boldsymbol{X} - \boldsymbol{\mu})^\top (L^{-1})^\top L^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) = \|L^{-1}(\boldsymbol{X} - \boldsymbol{\mu})\|_2^2.$$

Let $\boldsymbol{Z} = L^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$. Since $\boldsymbol{X}$ is normal, $\boldsymbol{Z}$ is normal.

- **Mean:** $\mathsf{E}[\boldsymbol{Z}] = L^{-1}(\mathsf{E}[\boldsymbol{X}] - \boldsymbol{\mu}) = \mathbf{0}$.
- **Covariance:** $\mathrm{Cov}(\boldsymbol{Z}) = L^{-1}\mathrm{Cov}(\boldsymbol{X})(L^{-1})^\top = L^{-1}\Sigma(L^{-1})^\top = L^{-1}(LL^\top)(L^\top)^{-1} = I$.

Thus, $\boldsymbol{Z} \sim \mathcal{N}_k(\mathbf{0}, I)$, and $\|\boldsymbol{Z}\|_2^2 \sim \chi_k^2$. □

The transformation $L^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ effectively de-correlates the components of $\boldsymbol{X}$ and scales them to unit variance. By measuring distance in this way (the Mahalanobis distance), we obtain a statistic with a fixed, known distribution (χ_k^2) regardless of the specific values of $\boldsymbol{\mu}$ or Σ. This invariance is crucial for constructing test statistics.

Remark. While the Cholesky decomposition gives a lower-triangular L, one could alternatively use the symmetric square root $\Sigma^{1/2}$ derived from the spectral decomposition:

$$\Sigma^{1/2} = \boldsymbol{Q} \begin{pmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_k} \end{pmatrix} \boldsymbol{Q}^\top,$$

where $\lambda_j > 0$ are the eigenvalues of Σ and the columns of $\boldsymbol{Q}$ are the eigenvectors. The result holds for any matrix square root L such that $LL^\top = \Sigma$.

11.1.4 Hotelling's T^2

We now apply the asymptotic theory of normal distributions to a specific hypothesis testing problem: comparing a sample mean vector to a hypothesized population mean.

Setup: Suppose we observe n data points, where each data point is a vector in $\mathbb{R}^k$. For example, imagine a clinical study with n patients, where we measure k characteristics (e.g., iron, protein, vitamin levels) for each patient. We denote these random vectors as $X_1, X_2, \ldots, X_n$. We assume they are drawn i.i.d. from a distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Note that we do not strictly require the underlying data to be normally distributed; normality will emerge asymptotically via averaging (CLT).

Goal: We wish to test whether the population mean $\boldsymbol{\mu}$ is equal to a specific target vector $\boldsymbol{\mu}_0$.

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0.$$

To do this, we compute the sample mean $\overline{X}_n$. We need a way to measure the "distance" between the observed $\overline{X}_n$ and the hypothesized $\boldsymbol{\mu}_0$.

Decorrelation and the Mahalanobis Distance Simply calculating the Euclidean distance $\|\overline{X}_n - \boldsymbol{\mu}_0\|^2$ is often misleading because the coordinates of X might be highly correlated or have vastly different variances (e.g., one variable is order 10^3 while another is 10^{-1}). To gauge if a deviation is statistically significant, we must account for this structure. We essentially need to "decorrelate" the data.

If we knew the true covariance Σ, we would use the Mahalanobis distance based on Σ^{-1}. Since Σ is unknown, we replace it with the **sample covariance matrix** S_n:

$$S_n = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)(X_i - \overline{X}_n)^\top.$$

This leads us to **Hotelling's T^2 statistic**.

Analogy to the Univariate t-statistic: It is helpful to view this as the high-dimensional generalization of the squared Student's t-statistic. In 1D ($k = 1$), the t-statistic is $t = \frac{\overline{X}_n - \mu_0}{s/\sqrt{n}}$. Squaring this gives:

$$t^2 = \frac{n(\overline{X}_n - \mu_0)^2}{s^2} = n(\overline{X}_n - \mu_0)(s^2)^{-1}(\overline{X}_n - \mu_0).$$

Hotelling's T^2 has the exact same form, but with vectors and matrices:

$$T^2 = n(\overline{X}_n - \boldsymbol{\mu}_0)^\top S_n^{-1}(\overline{X}_n - \boldsymbol{\mu}_0).$$

Theorem 11.1 (Asymptotic Distribution of Hotelling's T^2)**.** *Let $X_1, X_2, \dots$ be i.i.d. random vectors in $\mathbb{R}^k$ with mean vector $\boldsymbol{\mu} = \mathsf{E}[X_i]$ and finite invertible covariance matrix $\Sigma = \mathsf{Var}[X_i]$. Let S_n be the sample covariance matrix. Then $S_n \xrightarrow{p} \Sigma$, and*

$$T^2(\boldsymbol{\mu}) := n(\overline{X}_n - \boldsymbol{\mu})^\top S_n^{-1}(\overline{X}_n - \boldsymbol{\mu}) \xrightarrow{d} \chi^2_k.$$

Proof. The proof relies on Slutsky's Theorem and the Continuous Mapping Theorem. First, consider the decomposition of the sample covariance matrix:

$$S_n = \frac{n}{n-1}\left[\underbrace{\frac{1}{n}\sum_{i=1}^{n}(X_i - \boldsymbol{\mu})(X_i - \boldsymbol{\mu})^\top}_{\xrightarrow{p}\Sigma \text{ by LLN}} - \underbrace{(\overline{X}_n - \boldsymbol{\mu})}_{\xrightarrow{p}0}\underbrace{(\overline{X}_n - \boldsymbol{\mu})^\top}_{\xrightarrow{p}0}\right].$$

Thus, $S_n \xrightarrow{p} \Sigma$, which implies $S_n^{-1} \xrightarrow{p} \Sigma^{-1}$ (since matrix inversion is continuous at invertible matrices).

Next, by the Central Limit Theorem (CLT):

$$\sqrt{n}(\overline{X}_n - \boldsymbol{\mu}) \xrightarrow{d} Y \sim \mathcal{N}_k(0, \Sigma).$$

We can now apply Slutsky's Theorem to the product of the vector convergence and the matrix convergence. The statistic converges in distribution to the quadratic form of the limiting normal variable:

$$T^2(\boldsymbol{\mu}) \xrightarrow{d} Y^\top \Sigma^{-1} Y.$$

By Lemma **??**, $Y^\top \Sigma^{-1} Y \sim \chi^2_k$. □

Application: Hypothesis Testing

To perform a test with asymptotic level $\alpha \in (0,1)$, we use the quantile of the chi-square distribution as our critical value.

Decision Rule: Reject $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ if

$$T^2(\boldsymbol{\mu}_0) \geq \chi^2_{k,1-\alpha},$$

where $\chi^2_{k,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2_k distribution. Under H_0, the probability of type I error converges to α as $n \to \infty$.

Example 11.2 (Women's Health and Nutrition)**.** Consider a dataset regarding the nutrient intake of $n = 40$ women. We observe four variables ($k = 4$): Iron, Protein, Vitamin A, and Vitamin C. We want to test if their average intake matches the recommended daily values.

Data Preview:

```
X <- readr::read_csv("data/nutrient.cs")
# # A tibble: 40 x 4
#    Iron Protein `Vit A` `Vit C`
#   <dbl>   <dbl>   <dbl>   <dbl>
# 1  10.2    42.6    349.    54.1
# 2  13.7    59.9    668.   155.
...
n <- nrow(X) # 40
```

Hypothesis: The suggested daily intake vector is $\boldsymbol{\mu}_0 = (15, 75, 800, 90)^\top$.

```
mu0 <- c(15, 75, 800, 90)
```

Observed Means:

```
Xbar <- colMeans(X)
#     Iron  Protein     Vit A     Vit C
# 13.48640 81.52893 769.4960 97.69060
```

Notice that Vitamin A has a much larger scale (variance) than Iron. A simple difference would be dominated by Vitamin A, but Hotelling's T^2 handles this scaling automatically.

Test Statistic Calculation:

```
# T2 = n * (Xbar - mu0)' * S_inv * (Xbar - mu0)
S_inv <- solve(cov(X))
diff <- Xbar - mu0
T2 <- n * t(diff) %*% S_inv %*% diff
# [1] 18.66158
```

Conclusion: Under H_0, $T^2 \approx \chi^2_4$. The expected value of a χ^2_4 variable is 4. The observed value of 18.66 is unusually large (in the far right tail).

```
1 - pchisq(T2, df=4)
# [1] 0.000915846
```

With a p-value of ≈ 0.0009, we reject the null hypothesis. It is highly implausible that the women's average intake conforms to the recommended guidelines.

11.1.5 Non-Central Chi-Square Distribution

We previously established that under the null hypothesis ($H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$), the test statistic follows a central χ^2 distribution. This allows us to control the Type I error rate. However, if we want to analyze the **power** of a test, we must ask:

> *What happens when the null hypothesis is false?*

In this scenario, the data is centered around a true mean $\boldsymbol{\mu}$ that differs from the hypothesized $\boldsymbol{\mu}_0$. Consequently, we are looking at the squared norm of a random vector with a non-zero mean. This leads to the **non-central chi-square distribution**.

Definition 11.1 (Non-Central Chi-Square)**.** Let $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\mu}, I_k)$. That is, $X_1, \ldots, X_k$ are independent with unit variance, but $X_j \sim \mathcal{N}(\mu_j, 1)$. The distribution of the squared norm $\|\boldsymbol{X}\|_2^2 = \sum_{j=1}^k X_j^2$ is called the **non-central chi-square distribution** with k degrees of freedom and non-centrality parameter $\lambda = \|\boldsymbol{\mu}\|_2^2$. We denote this as:

$$\|\boldsymbol{X}\|_2^2 \sim \chi_k^2(\|\boldsymbol{\mu}\|_2^2).$$

Geometric Intuition (Rotational Invariance): The definition above claims that the distribution depends on $\boldsymbol{\mu}$ *only* through its squared length $\|\boldsymbol{\mu}\|^2$. Why is this well-defined? Consider two mean vectors $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ with the same length ($\|\boldsymbol{\mu}\| = \|\boldsymbol{\delta}\|$).

Since they lie on the same hypersphere, there exists an orthogonal matrix $\boldsymbol{Q}$ (a rotation) such that $\boldsymbol{\delta} = \boldsymbol{Q}\boldsymbol{\mu}$. If we define $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{X}$, then:

- Length is preserved: $\|\boldsymbol{Y}\|^2 = \|\boldsymbol{Q}\boldsymbol{X}\|^2 = \|\boldsymbol{X}\|^2$.
- Distribution is preserved: $\boldsymbol{Y}$ is still normal with identity covariance (since $\boldsymbol{Q}I\boldsymbol{Q}^\top = I$), but its mean is rotated to $\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{\delta}$.

Thus, the distribution of the squared norm depends only on the distance of the mean vector from the origin, not its direction.

Note (R Implementation)**.** In `R`, the chi-square family functions (`rchisq`, `dchisq`, `pchisq`, `qchisq`) accept an optional argument `ncp` (non-centrality parameter).

```r
# Generate 100 draws from non-central chi-square (df=5, lambda=2.1)
rchisq(n=100, df=5, ncp=2.1)
```

Projection Matrices and Singular Covariance

In the previous sections, we assumed the covariance matrix Σ was invertible to define the Mahalanobis distance. However, in many applications—specifically the upcoming dice roll problem—variables satisfy linear constraints (e.g., counts summing to N). This forces the random vector to lie in a lower-dimensional subspace, resulting in a **singular** (non-invertible) covariance matrix.

We need to understand when $\|\boldsymbol{X}\|^2$ follows a chi-square distribution even if Σ is singular. The key concept is the **projection matrix**.

Definition 11.2 (Projection Matrix)**.** A symmetric matrix Σ is a projection matrix if $\Sigma^2 = \Sigma$.

- **Eigenvalues:** Since $\lambda^2 = \lambda$, the eigenvalues must be either 0 or 1.

- **Intuition:** If you project an object onto the floor (Σ), it lands on the floor. If you try to project it onto the floor again (Σ^2), it doesn't move—it's already there.

Lemma 11.2. *Let $\boldsymbol{X} \sim \mathcal{N}_k(\boldsymbol{\delta}, \Sigma)$ where Σ is a projection matrix of rank r. If the mean vector lies in the projection space (i.e., $\Sigma\boldsymbol{\delta} = \boldsymbol{\delta}$), then:*

$$\|\boldsymbol{X}\|_2^2 \sim \chi_r^2(\|\boldsymbol{\delta}\|_2^2).$$

TODO: improve proof

Proof (Sketch). Using the spectral decomposition, we can rotate the space such that Σ becomes a diagonal matrix with r ones and $k-r$ zeros. This effectively isolates r active standard normal variables (plus means), while the remaining $k-r$ variables are identically zero. The sum of squares is then simply the sum of r squared normals. □

11.2 Multinomial Experiments

Suppose we conduct n independent replications of an experiment with c possible outcomes (e.g., rolling a die). Let U_i be the random variable representing the categorical outcome of the i-th trial:

$$U_i = j \quad \text{if the } i\text{-th trial results in outcome } j \in \{1, \dots, c\}.$$

The variables $U_1, \dots, U_n$ are i.i.d., and their distribution is completely determined by the probabilities:

$$p_j = \mathsf{P}(U_i = j), \quad j = 1, \dots, c.$$

We assume $p_j > 0$ for all j, and naturally, $\sum_{j=1}^{c} p_j = 1$.

Indicator Variables (One-Hot Encoding) While the raw data U_i are integers, theoretical analysis (specifically the Central Limit Theorem) is much easier if we work with sums of vectors. We introduce binary indicators X_{ij} to represent the data algebraically.

$$X_{ij} = \mathbf{1}_{\{j\}}(U_i) = \begin{cases} 1 & \text{if } U_i = j \\ 0 & \text{otherwise} \end{cases}$$

We can view the outcome of the i-th trial as a vector $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ic})^\top$. This vector is a canonical basis vector $\boldsymbol{e}_j$ (often called a "one-hot" vector in machine learning) with a 1 at position j and 0 elsewhere.

$$\boldsymbol{X}_i = \boldsymbol{e}_j = (0, \dots, 1, \dots, 0)^\top \iff U_i = j.$$

This notation "blows up" a single column of categorical data into a matrix of binary indicators, but it simplifies the math: the column sums of this matrix are exactly the counts we need.

Example 11.3 (Rolling a four-sided die)**.** Consider $n = 8$ rolls of a 4-sided die ($c = 4$).

```r
set.seed(2025)
c <- 4
n <- 8
# Simulate raw outcomes
U <- sample(1:c, n, replace=TRUE)
print(U)
# [1] 1 4 4 2 1 3 2 3

# Convert to Indicator Matrix (One-Hot Encoding)
X <- matrix(0, n, c)
for(i in 1:n){ X[i, U[i]] = 1 }

# Display as Data Frame
data.frame(X)
#   X1 X2 X3 X4
# 1  1  0  0  0  <-- U[1]=1
# 2  0  0  0  1  <-- U[2]=4
# 3  0  0  0  1
# 4  0  1  0  0
# 5  1  0  0  0
# 6  0  0  1  0
# 7  0  1  0  0
# 8  0  0  1  0
```

11.2.1 The Multinomial Distribution

Sufficient Statistics: The Counts From the indicator vectors, we derive the aggregate counts for each category. Let N_j be the number of times outcome j occurred in the n trials. In terms of our indicator matrix, N_j is simply the column sum:

$$N_j = \sum_{i=1}^{n} X_{ij}.$$

The vector of counts $\boldsymbol{N} = (N_1, \ldots, N_c)^\top$ follows a **Multinomial Distribution** with parameters n and probability vector $\boldsymbol{p} = (p_1, \ldots, p_c)^\top$.

- If $c = 2$, this simplifies to the familiar **Binomial distribution**.
- Note that $\sum_{j=1}^{c} N_j = n$. This linear constraint means the random vector $\boldsymbol{N}$ is confined to a subspace of dimension $c-1$. Consequently, the covariance matrix of $\boldsymbol{N}$ is **singular**. This connects back to our previous discussion on projection matrices—we will rely on that theory to analyze test statistics involving $\boldsymbol{N}$.

We formally define the distribution of the count vector $\boldsymbol{N}$.

Definition 11.3 (Multinomial Distribution)**.** The random vector of counts $\boldsymbol{N} = (N_1, \ldots, N_c)^\top$ follows a **Multinomial Distribution** with parameters n (number of trials) and $\boldsymbol{p} = (p_1, \dots, p_c)^\top$ (probabilities). The probability mass function is given by:

$$\mathsf{P}(N_1 = n_1, \dots, N_c = n_c) = \binom{n}{n_1 \dots n_c} p_1^{n_1} \dots p_c^{n_c}$$

where $n_j \geq 0$ and $\sum_{j=1}^{c} n_j = n$. The term $\binom{n}{n_1 \dots n_c} = \frac{n!}{n_1! \dots n_c!}$ is the *multinomial coefficient*, counting the number of ways to partition n items into c labeled groups with sizes $n_1, \dots, n_c$.

Moments and Covariance Structure

Using the indicator variable representation $\boldsymbol{N} = \sum_{i=1}^{n} \boldsymbol{X}_i$, we can easily derive the moments. Recall that $\boldsymbol{X}_i$ is a "one-hot" vector where $X_{ij} \sim \text{Bernoulli}(p_j)$.

Expectation: Since expectation is linear:

$$\mathsf{E}[N_j] = \sum_{i=1}^{n} \mathsf{E}[X_{ij}] = \sum_{i=1}^{n} p_j = np_j.$$

Thus, the expected vector is $\mathsf{E}[\boldsymbol{N}] = n\boldsymbol{p}$.

Variance and Covariance: Since the trials are independent, the variance of the sum is the sum of the variances: $\text{Var}(\boldsymbol{N}) = n\text{Var}(\boldsymbol{X}_i)$. We focus on determining the covariance matrix of a single trial, $\boldsymbol{X}_i$.

- **Variance (Diagonal):** For a single category j, X_{ij} is a Bernoulli variable.

$$\text{Var}(X_{ij}) = p_j(1 - p_j).$$

- **Covariance (Off-diagonal):** For distinct categories $j \neq \ell$, we look at the interaction.

$$\mathsf{Cov}(X_{ij}, X_{i\ell}) = \mathsf{E}[X_{ij}X_{i\ell}] - \mathsf{E}[X_{ij}]\mathsf{E}[X_{i\ell}].$$

  Crucially, a single trial cannot result in both outcome j and outcome ℓ simultaneously (mutually exclusive). Thus, the product $X_{ij}X_{i\ell}$ is always 0.

$$\mathsf{E}[X_{ij}X_{i\ell}] = 0 \implies \mathsf{Cov}(X_{ij}, X_{i\ell}) = 0 - p_j p_\ell = -p_j p_\ell.$$

Note. The covariance is negative. This makes intuitive sense: if we know trial i resulted in category j (so $X_{ij} = 1$), then it *cannot* be category ℓ (so $X_{i\ell}$ must be 0). If one indicator goes up, the others must yield.

Matrix Representation and Singularity

We can express the covariance matrix of a single trial $\boldsymbol{X}_i$, denoted as $\Sigma_{\boldsymbol{X}}$, in a compact matrix form. Let $\mathbf{P} = \operatorname{diag}(p_1, \dots, p_c)$ be a diagonal matrix of probabilities.

$$\operatorname{Var}(\boldsymbol{X}_i) = \begin{pmatrix} p_1(1-p_1) & -p_1p_2 & \dots & -p_1p_c \\ -p_2p_1 & p_2(1-p_2) & & \vdots \\ \vdots & & \ddots & \\ -p_cp_1 & \dots & & p_c(1-p_c) \end{pmatrix} = \mathbf{P} - \boldsymbol{p}\boldsymbol{p}^\top.$$

Consequently, the covariance matrix for the total counts $\boldsymbol{N}$ is:

$$\operatorname{Var}(\boldsymbol{N}) = n(\mathbf{P} - \boldsymbol{p}\boldsymbol{p}^\top).$$

Rank and Kernel: The matrix $\mathbf{P} - \boldsymbol{p}\boldsymbol{p}^\top$ is **singular**. To see this, consider the vector of all ones, $\mathbf{1} = (1, \dots, 1)^\top$.

$$(\mathbf{P} - \boldsymbol{p}\boldsymbol{p}^\top)\mathbf{1} = \mathbf{P}\mathbf{1} - \boldsymbol{p}(\boldsymbol{p}^\top\mathbf{1}).$$

Since $\mathbf{P}\mathbf{1} = \boldsymbol{p}$ (the vector of probabilities) and $\boldsymbol{p}^\top\mathbf{1} = \sum p_j = 1$, we get:

$$\boldsymbol{p} - \boldsymbol{p}(1) = \mathbf{0}.$$

The vector $\mathbf{1}$ is in the kernel (null space) of the covariance matrix. This implies the rank is at most $c-1$ (it is exactly $c-1$ provided $p_j > 0$). This algebraic singularity corresponds exactly to the physical constraint that the counts must sum to n ($\sum N_j = n$). The data does not vary freely in c dimensions; it is confined to a $(c-1)$-dimensional subspace.

11.3 Pearson's Chi-Square Test

We now address the testing problem: given a vector of counts $\boldsymbol{N} \sim \text{Multinomial}(n, \boldsymbol{p})$, is the underlying probability vector $\boldsymbol{p}$ equal to a specific hypothesized vector $\boldsymbol{p}_0$?

The Hypotheses:

$$H_0 : \boldsymbol{p} = \boldsymbol{p}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{p} \neq \boldsymbol{p}_0$$

The Test Statistic: To test this, we compare the **observed counts** (N_j) with the **expected counts** under the null hypothesis. Under H_0, $\mathsf{E}[N_j] = np_{0j}$. Pearson's statistic aggregates the squared deviations, normalized by the expected counts:

$$W_n(\boldsymbol{p}_0) := \sum_{j=1}^{c} \frac{(N_j - np_{0j})^2}{np_{0j}}.$$

Decision Rule: We reject H_0 if the statistic is sufficiently large, specifically if $W_n \geq k_\alpha$, where k_α is the critical value derived from the asymptotic distribution.

11.3.1 Asymptotic Distribution

While the exact distribution of W_n is discrete (and complex to calculate for large n), Pearson's great insight was that it approaches a known limit.

Theorem 11.2. *If $\boldsymbol{N} \sim$ Multinomial$(n, \boldsymbol{p}_0)$ and $n \to \infty$, then:*

$$W_n(\boldsymbol{p}_0) \xrightarrow{d} \chi^2_{c-1}.$$

Consequently, the test that rejects when $W_n \geq \chi^2_{c-1;1-\alpha}$ has asymptotic level α:

$$\mathsf{P}_{\boldsymbol{p}_0}(W_n(\boldsymbol{p}_0) \geq k_\alpha) \underset{n\to\infty}{\longrightarrow} \alpha.$$

Proof.
Reformulation as a Quadratic Form
Let $\boldsymbol{P}_0 = \operatorname{diag}(p_{01}, \ldots, p_{0c})$. We can rewrite the sum notation using vectors. Let $\overline{\boldsymbol{X}}_n = \frac{1}{n}\boldsymbol{N} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i$ be the vector of relative frequencies. Algebraic manipulation shows:

$$W_n(\boldsymbol{p}_0) = n\sum_{j=1}^{c} \frac{(N_j/n - p_{0j})^2}{p_{0j}} = n(\overline{\boldsymbol{X}}_n - \boldsymbol{p}_0)^\top \boldsymbol{P}_0^{-1}(\overline{\boldsymbol{X}}_n - \boldsymbol{p}_0).$$

Multivariate CLT
The vector $\overline{\boldsymbol{X}}_n$ is an average of i.i.d. indicator vectors $\boldsymbol{X}_i$. Under H_0, $\mathsf{E}[\boldsymbol{X}_i] = \boldsymbol{p}_0$ and $\mathsf{Var}[\boldsymbol{X}_i] = \boldsymbol{P}_0 - \boldsymbol{p}_0\boldsymbol{p}_0^\top$ (from Section 11.2.1). Applying the Multivariate Central Limit Theorem:

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{p}_0) \xrightarrow{d} \boldsymbol{Y} \sim \mathcal{N}_c(\boldsymbol{0}, \boldsymbol{\Sigma}), \quad \text{where } \boldsymbol{\Sigma} = \boldsymbol{P}_0 - \boldsymbol{p}_0\boldsymbol{p}_0^\top.$$

Continuous Mapping
By the Continuous Mapping Theorem, the statistic converges to a quadratic form of the Gaussian vector $\boldsymbol{Y}$:

$$W_n(\boldsymbol{p}_0) \xrightarrow{d} \boldsymbol{Y}^\top \boldsymbol{P}_0^{-1}\boldsymbol{Y}.$$

We can factor $\boldsymbol{P}_0^{-1}$ as $\boldsymbol{P}_0^{-1/2}\boldsymbol{P}_0^{-1/2}$. Let $\boldsymbol{Z} = \boldsymbol{P}_0^{-1/2}\boldsymbol{Y}$. Then:

$$W_n(\boldsymbol{p}_0) \xrightarrow{d} \|\boldsymbol{Z}\|_2^2.$$

Analyzing the Covariance of Z
The vector $\boldsymbol{Z}$ is a linear transformation of a Gaussian, so $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{Z}})$. We compute its covariance:

$$\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \mathsf{Var}[\boldsymbol{P}_0^{-1/2}\boldsymbol{Y}] = \boldsymbol{P}_0^{-1/2}\boldsymbol{\Sigma}\boldsymbol{P}_0^{-1/2}.$$

Substituting $\boldsymbol{\Sigma} = \boldsymbol{P}_0 - \boldsymbol{p}_0\boldsymbol{p}_0^\top$:

$$\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{P}_0^{-1/2}(\boldsymbol{P}_0 - \boldsymbol{p}_0\boldsymbol{p}_0^\top)\boldsymbol{P}_0^{-1/2} = \boldsymbol{I}_c - (\boldsymbol{P}_0^{-1/2}\boldsymbol{p}_0)(\boldsymbol{P}_0^{-1/2}\boldsymbol{p}_0)^\top.$$

Let $\boldsymbol{u} = \boldsymbol{P}_0^{-1/2}\boldsymbol{p}_0$. This vector $\boldsymbol{u}$ has entries $\sqrt{p_{0j}}$. Its squared norm is:

$$\|\boldsymbol{u}\|^2 = \boldsymbol{u}^\top\boldsymbol{u} = \sum_{j=1}^{c}(\sqrt{p_{0j}})^2 = \sum_{j=1}^{c} p_{0j} = 1.$$

Thus, $\boldsymbol{\Sigma}_{\boldsymbol{Z}} = \boldsymbol{I}_c - \boldsymbol{u}\boldsymbol{u}^\top$. Since $\|\boldsymbol{u}\| = 1$, this is a **projection matrix** projecting onto the orthogonal complement of $\boldsymbol{u}$. Its rank is:

$$\operatorname{rank}(\boldsymbol{I}_c - \boldsymbol{u}\boldsymbol{u}^\top) = \operatorname{tr}(\boldsymbol{I}_c) - \operatorname{tr}(\boldsymbol{u}\boldsymbol{u}^\top) = c - 1.$$

Conclusion:
By Lemma 11.2 (distribution of squared norm of projected Gaussian),

$$\|\boldsymbol{Z}\|_2^2 \sim \chi^2_{c-1}.$$

□

11.3.2 Asymptotic Power

We now consider the behavior of the test when the null hypothesis is false. A good test should eventually detect any fixed deviation from the null as the sample size increases.

Theorem 11.3 (Consistency of Chi-Square Test)**.** *The Chi-Square test is **consistent**. That is, if the true distribution is $\boldsymbol{N} \sim$ Multinomial$(n, \boldsymbol{p})$ with $\boldsymbol{p} \neq \boldsymbol{p}_0$, then the probability of rejecting H_0 converges to 1:*

$$\mathsf{P}_{\boldsymbol{p}}(W_n(\boldsymbol{p}_0) \geq k_\alpha) \underset{n\to\infty}{\longrightarrow} 1.$$

Proof. We analyze the behavior of the statistic scaled by $1/n$. Recall that $W_n(\boldsymbol{p}_0) = n(\overline{\boldsymbol{X}}_n - \boldsymbol{p}_0)^\top \boldsymbol{P}_0^{-1}(\overline{\boldsymbol{X}}_n - \boldsymbol{p}_0)$.

Convergence of the Scaled Statistic
By the Law of Large Numbers (LLN), the vector of observed relative frequencies converges in probability to the true probability vector:

$$\overline{\boldsymbol{X}}_n \xrightarrow{p} \boldsymbol{p}.$$

By the Continuous Mapping Theorem, the quadratic form converges:

$$\frac{1}{n}W_n(\boldsymbol{p}_0) = (\overline{\boldsymbol{X}}_n - \boldsymbol{p}_0)^\top \boldsymbol{P}_0^{-1}(\overline{\boldsymbol{X}}_n - \boldsymbol{p}_0) \xrightarrow{p} (\boldsymbol{p} - \boldsymbol{p}_0)^\top \boldsymbol{P}_0^{-1}(\boldsymbol{p} - \boldsymbol{p}_0) =: \Delta.$$

Positivity of the Limit
The matrix $\boldsymbol{P}_0 = \operatorname{diag}(\boldsymbol{p}_0)$ is a diagonal matrix with positive entries, so its inverse

$\boldsymbol{P}_0^{-1}$ is positive definite. Since $\boldsymbol{p} \neq \boldsymbol{p}_0$, the vector $(\boldsymbol{p} - \boldsymbol{p}_0)$ is non-zero. Therefore, the quadratic form Δ must be strictly positive:

$$\Delta > 0.$$

Asymptotic Probability
The critical value k_α is a fixed constant (the quantile of a χ^2_{c-1} distribution) and does not grow with n. Thus, $k_\alpha/n \to 0$. We can lower bound the rejection probability for sufficiently large n (such that $k_\alpha/n < \Delta/2$):

$$\mathsf{P}(W_n(\boldsymbol{p}_0) \geq k_\alpha) = \mathsf{P}\left(\frac{1}{n}W_n(\boldsymbol{p}_0) \geq \frac{1}{n}k_\alpha\right).$$

Since $\frac{1}{n}W_n(\boldsymbol{p}_0) \xrightarrow{p} \Delta$ and $\frac{1}{n}k_\alpha \to 0$, for any $\epsilon > 0$, the probability mass concentrates around Δ. Specifically:

$$\mathsf{P}\left(\frac{1}{n}W_n(\boldsymbol{p}_0) \geq \frac{1}{2}\Delta\right) \to 1.$$

Intuitively, under the alternative hypothesis, the statistic W_n grows linearly with n (it is roughly $n\Delta$), while the critical threshold k_α stays constant. Thus, the statistic eventually exceeds the threshold with certainty. □

11.3.3 Local Power Analysis

While consistency (Theorem 11.3) assures us that the test will eventually reject any fixed false null hypothesis as $n \to \infty$, it does not help us approximate the power for a specific finite sample size n. To derive a useful approximation for power and sample size planning, we consider **local alternatives** that approach the null hypothesis at a rate of $1/\sqrt{n}$.

Sequence of Alternatives: Consider a sequence of probability vectors $\boldsymbol{p}_n$ defined by:

$$\boldsymbol{p}_n = \boldsymbol{p}_0 + \frac{1}{\sqrt{n}}\boldsymbol{\delta},$$

where $\boldsymbol{p}_0$ is the null hypothesis vector ($p_{0j} > 0, \sum p_{0j} = 1$) and $\boldsymbol{\delta} \in \mathbb{R}^c$ is a fixed direction vector such that $\sum_{j=1}^c \delta_j = 0$ (ensuring $\sum p_{nj} = 1$).

Theorem 11.4. *Let $\boldsymbol{N}^{(n)} \sim$ Multinomial$(n, \boldsymbol{p}_n)$. Then, as $n \to \infty$:*

$$W_n(\boldsymbol{p}_0) = \sum_{j=1}^c \frac{(N_j^{(n)} - np_{0j})^2}{np_{0j}} \xrightarrow{d} \chi^2_{c-1}(\lambda),$$

where the limit is a ***non-central chi-square distribution*** *with $c-1$ degrees of freedom and non-centrality parameter:*

$$\lambda = \sum_{j=1}^c \frac{\delta_j^2}{p_{0j}}.$$

Application (Power Calculation): This theorem allows us to approximate the power of the test for a specific alternative $\boldsymbol{p}$ and sample size n.

1. Set $\boldsymbol{\delta} = \sqrt{n}(\boldsymbol{p} - \boldsymbol{p}_0)$.
2. Calculate $\lambda = \sum \delta_j^2/p_{0j} = n\sum(p_j - p_{0j})^2/p_{0j}$.
3. The power is approximately $\mathsf{P}(Y \geq k_\alpha)$, where $Y \sim \chi^2_{c-1}(\lambda)$.

Proof. The proof relies on establishing the asymptotic normality of the scaled count vector under the sequence of local alternatives.

Asymptotic Normality of the Counts
We claim that:

$$\sqrt{n}\left(\frac{1}{n}\boldsymbol{N}^{(n)} - \boldsymbol{p}_0\right) \xrightarrow{d} \boldsymbol{Y} \sim \mathcal{N}_c(\boldsymbol{\delta}, \boldsymbol{\Sigma}), \tag{11.1}$$

where $\boldsymbol{\Sigma} = \boldsymbol{P}_0 - \boldsymbol{p}_0\boldsymbol{p}_0^\top$. To see this, we decompose the term:

$$\sqrt{n}\left(\frac{1}{n}\boldsymbol{N}^{(n)} - \boldsymbol{p}_0\right) = \sqrt{n}\left(\frac{1}{n}\boldsymbol{N}^{(n)} - \boldsymbol{p}_n\right) + \sqrt{n}(\boldsymbol{p}_n - \boldsymbol{p}_0).$$

By definition of $\boldsymbol{p}_n$, the second term is exactly $\boldsymbol{\delta}$. It remains to show that the first term converges to $\mathcal{N}_c(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Using the Cramér-Wold device, we check convergence for linear combinations $\boldsymbol{a}^\top\boldsymbol{N}^{(n)}$ for any $\boldsymbol{a} \in \mathbb{R}^c$. Specifically, we restrict attention to $\boldsymbol{a} \perp \boldsymbol{1}$ (since the sum of counts is fixed). Let $\boldsymbol{X}_i^{(n)}$ be the indicator vector for the i-th trial in the n-th experiment.

$$\frac{1}{n}\boldsymbol{N}^{(n)} \overset{d}{=} \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^{(n)}.$$

The variance of the projection is $\sigma_n^2 = n \cdot \boldsymbol{a}^\top\boldsymbol{\Sigma}(\boldsymbol{p}_n)\boldsymbol{a}$. As $n \to \infty$, $\boldsymbol{p}_n \to \boldsymbol{p}_0$, so $\boldsymbol{\Sigma}(\boldsymbol{p}_n) \to \boldsymbol{\Sigma}(\boldsymbol{p}_0)$. To apply the Lyapunov CLT, we bound the third moment. Let $A = \max_j |a_j|$. Then $|\boldsymbol{a}^\top\boldsymbol{X}_i^{(n)}| \leq A$ and $|\boldsymbol{a}^\top\boldsymbol{p}_n| \leq A$.

$$\mathsf{E}[|\boldsymbol{a}^\top\boldsymbol{X}_i^{(n)} - \boldsymbol{a}^\top\boldsymbol{p}_n|^3] \leq 8A^3.$$

The Lyapunov condition requires:

$$\frac{1}{\sigma_n^3}\sum_{i=1}^{n}\mathsf{E}[\dots] \leq \frac{8nA^3}{n^{3/2}(\boldsymbol{a}^\top\boldsymbol{\Sigma}(\boldsymbol{p}_n)\boldsymbol{a})^{3/2}} \xrightarrow{n\to\infty} 0.$$

Thus, the CLT holds, and $\sqrt{n}(\frac{1}{n}\boldsymbol{N}^{(n)} - \boldsymbol{p}_n) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Adding the constant shift $\boldsymbol{\delta}$ yields Equation (11.1).

Transformation to Non-Central Chi-Square
As in Theorem 12.2, the test statistic is a quadratic form:

$$W_n(\boldsymbol{p}_0) \xrightarrow{d} \boldsymbol{Y}^\top\boldsymbol{P}_0^{-1}\boldsymbol{Y} = \|\boldsymbol{Z}\|^2,$$

where $\boldsymbol{Z} = \boldsymbol{P}_0^{-1/2}\boldsymbol{Y}$. Since $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\delta}, \boldsymbol{\Sigma})$, the linear transformation $\boldsymbol{Z}$ is also Normal:

$$\boldsymbol{Z} \sim \mathcal{N}_c(\boldsymbol{P}_0^{-1/2}\boldsymbol{\delta}, \boldsymbol{P}_0^{-1/2}\boldsymbol{\Sigma}\boldsymbol{P}_0^{-1/2}).$$

The covariance matrix simplifies to a projection matrix $\boldsymbol{I}_c - \boldsymbol{u}\boldsymbol{u}^\top$ (where $\boldsymbol{u} = \sqrt{\boldsymbol{p}_0}$), just as in the null case. The mean vector is $\boldsymbol{\mu}_Z = \boldsymbol{P}_0^{-1/2}\boldsymbol{\delta}$. Since $\sum \delta_j = 0$, we have $\boldsymbol{u}^\top \boldsymbol{\mu}_Z = \boldsymbol{p}_0^\top \boldsymbol{P}_0^{-1}\boldsymbol{\delta} = \boldsymbol{1}^\top \boldsymbol{\delta} = 0$. The vector $\boldsymbol{Z}$ lies in the subspace where the covariance acts as the identity. Thus, $\|\boldsymbol{Z}\|^2$ follows a non-central chi-square distribution.

The non-centrality parameter is the squared norm of the mean:

$$\lambda = \|\boldsymbol{\mu}_Z\|^2 = \|\boldsymbol{P}_0^{-1/2}\boldsymbol{\delta}\|^2 = \sum_{j=1}^{c} \frac{\delta_j^2}{p_{0j}}.$$

□

11.4 R Example

In this section, we apply the theory to a practical example: testing the fairness of a die using R. We will proceed in three steps: first by manually calculating the statistic to understand the mechanics, second by using R's built-in testing function, and third by handling small sample sizes via simulation.

11.4.1 Applying the Test Manually

Consider a scenario where we roll a six-sided die $n = 100$ times. We wish to test the null hypothesis that the die is fair:

$$H_0 : \boldsymbol{p} = \boldsymbol{p}_0 = (1/6, \ldots, 1/6)^\top.$$

First, we define the null probabilities and the observed counts $\boldsymbol{N}$ in R:

```r
# Null hypothesis probabilities (fair die)
p0 <- rep(1/6, 6)

# Observed counts for the six sides
N <- c(20, 19, 16, 21, 15, 9)

# Total number of trials
n <- sum(N)
print(n)
# [1] 100
```

To determine if these counts are unusual, we calculate Pearson's chi-square statistic $W_n(\boldsymbol{p}_0)$ manually:

$$W_n = \sum_{j=1}^{c} \frac{(N_j - np_{0j})^2}{np_{0j}}.$$

```
# Calculate the test statistic manually
W <- sum((N - n * p0)^2 / (n * p0))
print(W)
# [1] 5.84
```

The observed value is $W \approx 5.84$. Is this large or small? Under H_0, W follows a χ^2 distribution with $c - 1 = 5$ degrees of freedom. The expectation of a χ^2_5 variable is 5, so 5.84 is very close to the expected value, suggesting no strong evidence against H_0.

We can calculate specific critical values k_α for $\alpha = 0.05$ and $\alpha = 0.1$:

```
# Critical values for alpha = 0.05 and 0.1
qchisq(c(0.95, 0.90), df = 6 - 1)
# [1] 11.07050  9.23636
```

Since $5.84 < 9.23$, we cannot reject H_0 even at the 10% level. The asymptotic p–value is calculated as $\mathsf{P}(\chi^2_5 \geq W)$:

```
# £p£--value calculation
pchisq(W, df = 6 - 1, lower.tail = FALSE)
# [1] 0.3221003
```

Dedicated Function in R

In practice, we use the dedicated `chisq.test` function. It automatically handles the degrees of freedom and p–value calculation.

```
test_result <- chisq.test(N)
print(test_result)
#
#         Chi-squared test for given probabilities
#
# data:  N
# X-squared = 5.84, df = 5, £p£--value = 0.3221
```

The output matches our manual calculation perfectly. The result is stored as an object of class `htest`. We can inspect its structure to extract specific components:

```
attributes(test_result)
# £names
# [1] "statistic" "parameter" "p.value"   "method"    "data.name" "observed"
# [7] "expected"  "residuals" "stdres"
# £class
# [1] "htest"

test_result$p.value
# [1] 0.3221003
```

Non-Uniform Null Hypotheses We can also test against arbitrary probability vectors. For example, if we hypothesize a loaded die with $\boldsymbol{p}_0 = (0.5, 0.1, 0.1, 0.1, 0.1, 0.1)$:

```
# Testing a biased hypothesis
chisq.test(N, p = c(0.5, 0.1, 0.1, 0.1, 0.1, 0.1))
# X-squared = 44.4, df = 5, £p£--value = 1.921e-08
```

In this case, the p–value is extremely small, leading to a strong rejection of the biased hypothesis.

Small Counts and Simulation

Asymptotic approximations rely on the Central Limit Theorem. If the expected counts np_{0j} are small (a common rule of thumb is < 5), the χ^2 approximation may be inaccurate.

Consider a small experiment with a 3-sided die (outcomes 1, 2, 3) rolled only $n = 12$ times:

```
N_small <- c(3, 2, 7)
# Standard asymptotic test
chisq.test(N_small)
# X-squared = 3.5, df = 2, £p£--value = 0.1738
```

The asymptotic p–value is roughly 0.17. However, because the sample size is so small, we should not trust the continuous χ^2 curve to approximate the discrete distribution of the statistic.

Monte Carlo Simulation Since the null hypothesis is simple (all parameters are fully specified), we can approximate the exact distribution of the statistic by simulation. We generate B datasets from the null distribution, calculate the statistic for each, and see how often the simulated statistic exceeds our observed value.

R provides a built-in argument `simulate.p.value` for this:

```
set.seed(123)
chisq.test(N_small, simulate.p.value = TRUE, B = 20000)
#
#          Chi-squared test for given probabilities with simulated £p£--value
#          (based on 20000 replicates)
#
# data:  N_small
# X-squared = 3.5, df = NA, £p£--value = 0.2637
```

Notice the discrepancy: the asymptotic p–value was ≈ 0.17, while the simulated p–value is ≈ 0.26. While the statistical decision (do not reject) might remain the same here, the numerical difference highlights the error in asymptotic approximation for small n.

Under the Hood To understand what `simulate.p.value = TRUE` does, we can replicate the simulation manually:

```
# 1. Define data and parameters
n <- sum(N_small)
p0 <- c(1/3, 1/3, 1/3)
observed_W <- sum((N_small - n * p0)^2 / (n * p0))

# 2. Generate B synthetic datasets under H0
B <- 20000
set.seed(123)
# rmultinom generates B vectors of counts
simulated_counts <- rmultinom(n = B, size = n, prob = p0)

# 3. Compute statistic for every synthetic dataset
sim_W <- apply(simulated_counts, 2, function(x) {
  sum((x - n * p0)^2 / (n * p0))
})

# 4. Compute £p£--value (proportion of simulated W >= observed W)
mean(sim_W >= observed_W)
# [1] 0.26755
```

This manual simulation confirms the result provided by the dedicated function.

11.4.2 Power of Chi-Square Test

Following the theoretical framework of local alternatives, we can perform practical power calculations. This is often required in research proposals to justify that a study has a sufficient chance of detecting an effect if it exists.

Example 11.4 (Ferguson 1996)**.** Suppose we wish to test the fairness of a six-sided die ($\boldsymbol{p}_0 = 1/6 \cdot \mathbf{1}$) with a sample size of $n = 300$ at a significance level $\alpha = 0.05$.

First, we determine the critical value k_α. The test rejects H_0 if the statistic exceeds this value.

```r
# 1. Setup Null Hypothesis and Critical Value
p0 <- rep(1/6, 6)
n <- 300
alpha <- 0.05

# Critical value for chi-square with 6-1 = 5 df
crit <- qchisq(1 - alpha, df = 5)
print(crit)
# [1] 11.0705
```

Now, consider a specific alternative hypothesis where the die is biased:

$$\boldsymbol{p} = (0.13, 0.13, 0.17, 0.17, 0.20, 0.20)^\top$$

To approximate the power, we calculate the non-centrality parameter λ. Recall that $\boldsymbol{p} = \boldsymbol{p}_0 + \boldsymbol{\delta}/\sqrt{n}$, which implies $\boldsymbol{\delta} = \sqrt{n}(\boldsymbol{p} - \boldsymbol{p}_0)$. The non-centrality parameter is $\lambda = \sum \delta_j^2 / p_{0j}$.

```r
# 2. Define Alternative Hypothesis
p_alt <- c(0.13, 0.13, 0.17, 0.17, 0.2, 0.2)

# 3. Calculate Non-centrality Parameter (lambda)
# delta corresponds to the local alternative shift
delta <- sqrt(n) * (p_alt - p0)
lambda <- sum(delta^2 / p0)
print(lambda)
# [1] 8.88
```

The power is the probability that a non-central χ^2 variable exceeds the critical value established under the null.

```r
# 4. Calculate Asymptotic Power
power_approx <- pchisq(crit, df = 5, ncp = lambda, lower.tail = FALSE)
print(power_approx)
# [1] 0.6167596
```

With $n = 300$, we have approximately a 62% chance of detecting this specific deviation from fairness.

11.4.3 Sample Size Calculation

A common question in grant applications is the inverse of the power calculation: *"How large a sample size n is needed to achieve a target power (e.g., 90%)?"*

This is essentially a budget justification question: proving to a funding agency that the requested resources (to enroll n patients or perform n experiments) are necessary and sufficient to obtain a statistically significant finding.

We can solve this by reversing the previous calculation. We need to find the specific λ that yields a power of 0.9, and then solve for n.

Step 1: Find required λ We use a root-finding function to find λ_0 such that $P(\chi^2_5(\lambda_0) > k_\alpha) = 0.9$.

```r
target_power <- 0.9

# Function to minimize: Power(lambda) - Target
f_opt <- function(lambda) {
  pchisq(crit, df = 5, ncp = lambda, lower.tail = FALSE) - target_power
}

# Find root
lambda0 <- uniroot(f_opt, interval = c(0, 50))$root
print(lambda0)
# [1] 16.46946
```

Step 2: Solve for n We know that $\lambda = n \sum_{j=1}^{c} \frac{(p_j - p_{0j})^2}{p_{0j}}$. Let $E = \sum \frac{(p_j - p_{0j})^2}{p_{0j}}$ be the "effect size" (or distance from the null). Then $\lambda = n \cdot E$, so $n = \lambda / E$.

```r
# Calculate effect size component
effect_size <- sum((p_alt - p0)^2 / p0)

# Solve for n
n_required <- lambda0 / effect_size
print(n_required)
# [1] 556.4008
```

To achieve 90% power for this specific alternative, we would need $n \approx 557$ trials.

12. Asymptotic Normality of MLE

This chapter and the two chapters that follow develop asymptotic theory for likelihood-based statistical inference. Specifically, we will focus on maximum likelihood estimation (this chapter and Chapter 15) and likelihood ratio tests (Chapter 14).

For reading, consider Wellner (2018, chap. 4), Ferguson (1996, pt. 4), Lehmann and Casella (1998, chap. 6).

Setup

We consider the following standard statistical setup:

- **Model:** We assume a parametric family of distributions:

$$\mathcal{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}, \quad \Theta \subseteq \mathbb{R}^d.$$

- **Data:** We observe data points $X_1, \dots, X_n$ which are independent and identically distributed (i.i.d.) according to the true distribution $\mathsf{P}_{\boldsymbol{\theta}_0}$, where the true parameter $\boldsymbol{\theta}_0$ lies in the interior of the parameter space Θ.
- **Likelihood:** We define the likelihood function $L(\boldsymbol{\theta})$ and the log-likelihood function $\ell(\boldsymbol{\theta})$ based on the densities $p_{\boldsymbol{\theta}}$:

$$L(\boldsymbol{\theta}) \equiv L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(X_i),$$

$$\ell(\boldsymbol{\theta}) \equiv \ell_n(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(X_i).$$

The Maximum Likelihood Estimator (MLE), denoted as $\hat{\boldsymbol{\theta}}_n$, is the parameter value that maximizes the probability of observing the data we actually saw.

$$\hat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta})$$

In sufficiently smooth models, this estimator solves the **likelihood equations**:

$$\dot{\ell}(\boldsymbol{\theta}) \equiv \nabla_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}) = 0.$$

Intuition: The Link to KL Divergence Why do we maximize the likelihood? We can view the MLE as an estimator that attempts to minimize the Kullback-Leibler (KL) divergence between the empirical distribution of the data and the model distribution.

$$\min_{\theta\in\Theta} KL(\widehat{\mathsf{P}}|\mathsf{P}_\theta) = \text{const.} - \max_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i).$$

In other words, finding the MLE is equivalent to finding the parameter θ that makes the model distribution P_θ "closest" (in the KL sense) to the observed data $\widehat{\mathsf{P}}$.

Note (on Model Misspecification)**.** This perspective is particularly powerful because it explains what the MLE does even if our model is **wrong** (misspecified). If the true distribution P is not in our model family $\mathcal{P}$, the MLE will still converge to the parameter θ^* that minimizes the KL divergence $KL(\mathsf{P}|\mathsf{P}_\theta)$. It finds the "best approximating" distribution within our (flawed) model class.

In the sequel, we consider an estimator $\tilde{\theta}_n$ that is a **solution to the likelihood equations**, i.e., $\dot{\ell}(\tilde{\theta}_n) = 0$, allowing for the case that $\tilde{\theta}_n$ is merely a particular local maximum.

Example 12.1 (Gamma distribution model)**.** Suppose $X_1, ..., X_n$ are insurance claim sizes that we model as i.i.d. with a Gamma(α, β) distribution, where $\alpha > 0$ (shape) and $\beta > 0$ (rate) are unknown. The density is given by:

$$p_{\boldsymbol{\theta}}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \cdot \mathbf{1}_{(0,\infty)}(x),$$

where $\boldsymbol{\theta} = (\alpha, \beta)^\top$ and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$.

The log-likelihood function is:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} (\alpha \log(\beta) - \log\Gamma(\alpha) + (\alpha - 1)\log(X_i) - \beta X_i)$$

$$= n\alpha\log(\beta) - n\log\Gamma(\alpha) + (\alpha - 1)\sum_{i=1}^{n}\log(X_i) - \beta\sum_{i=1}^{n} X_i.$$

The MLE depends on the sufficient statistics $\sum\limits_{i=1}^{n}\log(X_i)$ and $\sum\limits_{i=1}^{n} X_i$. However, taking derivatives and setting them to zero results in a system that cannot be solved "in closed form" for α and β. This is a common situation in practice! Fortunately, it is no problem to compute the MLE via numerical optimization.

Illustration of consistency of MLE:
We can define the negative log-likelihood function in R and minimize it using 'optim'.

```
negloglik <- function(ab, data){
    a <- ab[1]
    b <- ab[2]
    n <- length(data)
    return(
        -(n * a * log(b) - n * log(gamma(a))
        + (a - 1) * sum(log(data))
        - b * sum(data))
    )
}
```

We simulate Gamma data with true parameters $\alpha = 5, \beta = 2$:

```
set.seed(22)
n <- 1000
x <- rgamma(n, shape = 5, rate = 2)
```

We minimize the negative log-likelihood function based on $n \in \{10, 100, 1000\}$ observations, starting with initial values $\alpha = \beta = 1$.

```
opt_10   <- optim(c(1, 1), negloglik, data = x[1:10])
opt_100  <- optim(c(1, 1), negloglik, data = x[1:100])
opt_1000 <- optim(c(1, 1), negloglik, data = x)
```

The estimates for $n = 10, 100$ and 1000 are:

```
rbind(opt_10$par, opt_100$par, opt_1000$par)
#           [,1]     [,2]
# [1,] 9.755201 4.360160  <-- n=10 (Very noisy)
# [2,] 5.145909 1.954187  <-- n=100 (Better)
# [3,] 5.038665 2.031913  <-- n=1000 (Very close to truth 5, 2)
```

Illustration of asymptotic normality of MLE:
To visualize the distribution of the estimator, we compute the MLE many times (10,000 simulations) for a fixed sample size $n = 500$.

```
set.seed(22)
Sims <- 10000
n <- 500
df <- data.frame(a = numeric(Sims), b = numeric(Sims))

for(s in 1:Sims){
    x = rgamma(n, shape = 5, rate = 2)
    opt = optim(c(5,2), negloglik, data = x)
    df$a[s] = opt$par[1]
    df$b[s] = opt$par[2]
}
```

We then plot the distribution of the estimates of α and β. To illustrate **joint** asymptotic normality (not just marginals), we also consider a linear combination, e.g., $2\alpha - \beta$.

```r
p1 <- ggplot(df, aes(x = a)) +
  geom_histogram(aes(y = after_stat(density)), colour = "white") +
  geom_function(fun = dnorm,
          args = list(mean = mean(df$b),
                  sd = sd(df$b)),
          colour = "red", linewidth = 1.5) +
  labs(x = expression(paste("MLE of ", beta)))

p3 <- ggplot(df, aes(x = 2 * a - b)) +
 geom_histogram(aes(y = after_stat(density)), colour = "white") +
 geom_function(fun = dnorm,
          args = list(mean = mean(2*df$a - df$b),
                  sd = sd(2*df$a - df$b)),
          colour = "red", linewidth = 1.5) +
 labs(x = expression(paste("MLE of ", 2*alpha - beta)))

p1 + p2 + p3 + plot_layout(ncol = 2, axis_titles = "collect")
```

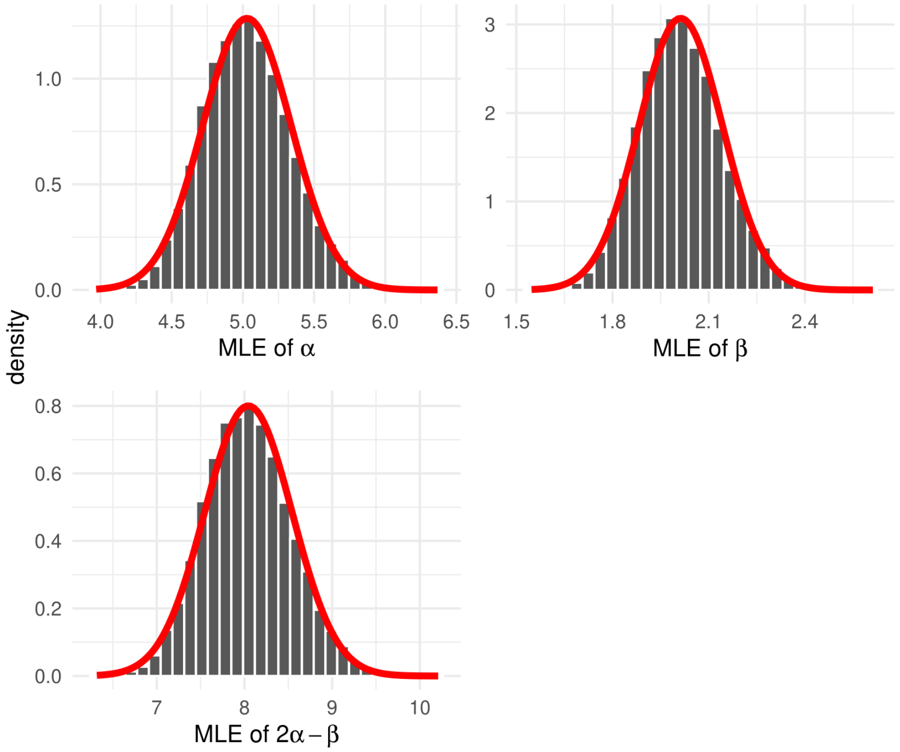


Figure 12.1: Comparison between the empirical distribution (histogram) of the maximum likelihood estimates for α, β and the linear combination $2\alpha - \beta$, and a normal distribution (solid line) with mean and variance being equal to the corresponding empirical moments.

The resulting histograms (Figure 12.1) closely match the normal density curves (solid red lines), providing empirical evidence for the central limit theorem applied to MLEs.

12.1 Consistency and Asymptotic Normality

Assumptions for a general theorem
To prove that the MLE behaves well (converges to the truth and is asymptotically normal), we need to ensure the log-likelihood function is sufficiently "smooth" and well-behaved. We make the following assumptions:

A0 (Identifiability): $\boldsymbol{\theta} \neq \boldsymbol{\theta}^* \implies \mathsf{P}_{\boldsymbol{\theta}} \neq \mathsf{P}_{\boldsymbol{\theta}^*}$. Different parameters must correspond to different distributions; otherwise, we could never distinguish them based on data.

A1 : Densities $p_\theta(x) = \frac{dP_\theta}{d\nu}(x)$ exist w.r.t. a σ-finite measure ν.

A2 (Common Support): $A := \{x : p_{\boldsymbol{\theta}}(x) > 0\}$ does not depend on $\boldsymbol{\theta}$. (This rules out cases like Uniform$(0, \boldsymbol{\theta})$, where the support boundary depends on the parameter, often breaking standard asymptotic theory).

A3 (Smoothness): On an open neighborhood Θ_0 of $\boldsymbol{\theta}_0$, for ν-a.e x:

i) $\ell(\boldsymbol{\theta}|x) \equiv \log p_{\boldsymbol{\theta}}(x)$ is twice continuously differentiable in $\boldsymbol{\theta}$.

ii) $\ell(\boldsymbol{\theta}|x)$ has third-order derivatives

$$\dddot{\ell}_{jkl}(\boldsymbol{\theta}|x) = \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} l(\boldsymbol{\theta}|x)$$

bounded by an integrable function $M(x)$:

$$|\dddot{\ell}_{ikl}(\boldsymbol{\theta}|x)| \leq M_{ikl}(x) \quad \forall \boldsymbol{\theta} \in \boldsymbol{\theta}_0 \quad \forall 1 \leq j,k,l \leq d,$$

where $\mathsf{E}_{\theta_0}[M_{jkl}(X_1)] < \infty \quad \forall j,k,l = 1,\ldots,d.$

Intuition: This control over the third derivative is crucial for the Taylor expansion argument. It ensures that the quadratic approximation of the log-likelihood dominates the higher-order terms (the "wiggles" of the function die out fast enough).

A4 (Fisher Information):

i) The score function has mean zero: $\mathsf{E}_{\boldsymbol{\theta}_0}[\dot{\ell}_j(\boldsymbol{\theta}_0|X_1)] = 0 \quad \forall j = 1, \ldots d.$

ii) The Fisher information $\mathrm{I}(\boldsymbol{\theta}_0) = \mathsf{Var}_{\boldsymbol{\theta}_0}[\dot{\ell}(\boldsymbol{\theta}_0|X_1)]$ is finite.

iii) The Fisher information is positive definite and equals

$$I(\boldsymbol{\theta}_0) = \left(-\mathsf{E}_{\boldsymbol{\theta}_0}\left[\dddot{\ell}_{jk}(\boldsymbol{\theta}_0|X_1)\right]\right)_{jk}$$

Our derivations will be driven by an application of the Central Limit Theorem (CLT) to the score function (gradient of the log-likelihood):

$$\boldsymbol{Z}_n := \frac{1}{\sqrt{n}}\dot{\ell}(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{\ell}(\boldsymbol{\theta}_0|X_i) \xrightarrow{d} \boldsymbol{Z} \sim \mathcal{N}_d\left(\mathbf{0}, \mathsf{Var}_{\boldsymbol{\theta}_0}\left[\dot{l}(\boldsymbol{\theta}_0|X_1)\right]\right) = \boldsymbol{Z} \sim \mathcal{N}_d\big(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\theta}_0)\big).$$

Theorem 12.1 (Consistency and Asymptotic Normality). *If assumptions A0-A4 hold, then:*

i. ***Consistency:*** *There exists a consistent sequence $\tilde{\boldsymbol{\theta}}_n$ of solutions to the likelihood equations. That is,*

$$\tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$$

and $\mathsf{P}_{\boldsymbol{\theta}_0}(\tilde{\boldsymbol{\theta}}_n$ solves likelihood equations$) \longrightarrow 1$ as $n \to \infty$.

ii. ***Asymptotic Normality:*** *If $\tilde{\boldsymbol{\theta}}_n$ satisfies the above, then*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = I(\boldsymbol{\theta}_0)^{-1} Z_n + o_p(1) \xrightarrow{d} \mathcal{N}_d(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1}).$$

Remark (Clarification on Existence). The phrasing "there exists a consistent sequence" can be technically subtle. It does not necessarily mean every root is consistent. However, for a concrete definition of $\tilde{\boldsymbol{\theta}}_n$ used in proofs, one can define:

$$\tilde{\boldsymbol{\theta}}_n = \begin{cases} \text{solution closest to } \boldsymbol{\theta}_0, & \text{if } \exists \text{ solution to likelihood eqn's,} \\ n \cdot \mathbf{1}, & \text{if } \nexists \text{ solution.} \end{cases}$$

In practice, if the likelihood is unimodal (e.g., exponential families), the unique root is the MLE and is consistent. If the likelihood is multimodal, we typically select the root that maximizes the likelihood or is closest to a preliminary consistent estimator.

Definition 12.1 (Asymptotically Linear Estimator). An estimator $\hat{\gamma}_n$ of parameter γ is **asymptotically linear** if there exists a function h (called the **influence function**) such that:

$$\sqrt{n}(\hat{\gamma}_n - \gamma) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} h(X_i) + o_p(1).$$

Remark (on Influence Functions and Asymptotic Linearity). Part ii. proves more than just normality; it proves the estimator is **asymptotically linear**.

Our theorem shows that the consistent root $\tilde{\boldsymbol{\theta}}_n$ is asymptotically linear with influence function:

$$h(X_i) = I(\boldsymbol{\theta}_0)^{-1} \dot{\ell}(\boldsymbol{\theta}_0 | X_i).$$

Why is this useful?

- **Robustness:** The term "influence function" comes from robust statistics. If $h(X_i)$ is bounded, a single outlier cannot destroy the estimator (like the median). If $h(X_i)$ is unbounded (like the mean), outliers can have a large effect.

- **Joint Distributions:** If we have two asymptotically linear estimators, say $\hat{\gamma}_n$ and $\tilde{\boldsymbol{\theta}}_n$, we can easily derive their joint asymptotic distribution. By stacking their influence functions, the vector $(\hat{\gamma}_n, \tilde{\boldsymbol{\theta}}_n)$ behaves like a sum of i.i.d. vectors, allowing us to apply the multivariate CLT to find asymptotic correlations.

TODO: detail this proof more

Proof.

1. **Existence and Consistency.**

   The proof relies on establishing that, with high probability, the log-likelihood function $\ell(\boldsymbol{\theta})$ has a local maximum within an arbitrarily small neighborhood of the true parameter $\boldsymbol{\theta}_0$.

   We define the boundary of a ball of radius a centered at $\boldsymbol{\theta}_0$:

   $$Q_a := \{\boldsymbol{\theta} \in \Theta : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = a\}.$$

   Claim (1)**.** For any sufficiently small $a > 0$,

   $$\mathsf{P}_{\boldsymbol{\theta}_0}(\ell(\boldsymbol{\theta}) < \ell(\boldsymbol{\theta}_0) \quad \forall \boldsymbol{\theta} \in Q_a) \longrightarrow 1 \quad \text{as } n \to \infty.$$

   If this claim holds, then with probability approaching 1, the continuous function $\ell(\boldsymbol{\theta})$ is strictly lower on the boundary Q_a than at the center $\boldsymbol{\theta}_0$. Consequently, $\ell(\boldsymbol{\theta})$ must attain a local maximum at some point $\tilde{\boldsymbol{\theta}}_n$ in the interior of the ball (i.e., $\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| < a$). Since ℓ is differentiable, this local maximum satisfies the likelihood equations $\dot{\ell}(\tilde{\boldsymbol{\theta}}_n) = 0$. Since this holds for any $a > 0$, it implies the existence of a sequence of roots $\tilde{\boldsymbol{\theta}}_n$ such that $\tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$.

   Proof (of Claim 1). We perform a Taylor expansion of the log-likelihood around $\boldsymbol{\theta}_0$. Let $\boldsymbol{\theta} \in Q_a$.

   $$\frac{1}{n}[\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0)] = S_{1n} + S_{2n} + R_n,$$

   where:

   - **Linear term:** $S_{1n} = (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \left(\frac{1}{n}\dot{\ell}(\boldsymbol{\theta}_0)\right)$.
   - **Quadratic term:** $S_{2n} = -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \left[-\frac{1}{n}\ddot{\ell}(\boldsymbol{\theta}_0)\right](\boldsymbol{\theta} - \boldsymbol{\theta}_0)$.
   - **Remainder term:** R_n involves third derivatives evaluated at some intermediate point $\boldsymbol{\theta}^*$ between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$.

   We analyze the asymptotic behavior of each term as $n \to \infty$:

   (a) **Linear Term (S_{1n}):** By assumption A4, $\mathsf{E}[\dot{\ell}(\boldsymbol{\theta}_0)] = 0$. By the Law of Large Numbers (LLN), $\frac{1}{n}\dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{p} 0$. Since $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = a$,

   $$|S_{1n}| \leq a \left\| \frac{1}{n}\dot{\ell}(\boldsymbol{\theta}_0) \right\| \xrightarrow{p} 0.$$

   Thus, for any $\epsilon > 0$ (e.g., a^3), $|S_{1n}| < \frac{1}{2}a^3$ with high probability.

   (b) **Quadratic Term (S_{2n}):** By the LLN, $-\frac{1}{n}\ddot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{p} I(\boldsymbol{\theta}_0)$. Since the Fisher Information matrix $I(\boldsymbol{\theta}_0)$ is positive definite, its smallest eigenvalue $\lambda_d > 0$. For sufficiently large n, the quadratic form behaves like:

   $$S_{2n} \approx -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top I(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Using the eigenvalue bound:

$$(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top I(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) \geq \lambda_d \|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2 = \lambda_d a^2.$$

Therefore, with high probability:

$$S_{2n} < -\frac{\lambda_d}{4}a^2 =: -Ca^2 \quad \text{(where } C > 0\text{)}.$$

(c) **Remainder Term (R_n):** Using the bound on third derivatives $M_{jkl}(x)$ from assumption A3:

$$|R_n| \leq \frac{1}{6}\sum_{j,k,l} |\theta_j - \theta_{0j}||\theta_k - \theta_{0k}||\theta_l - \theta_{0l}| \left(\frac{1}{n}\sum_{i=1}^n M_{jkl}(X_i)\right).$$

Since $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| = a$, the product of differences scales with a^3. By the LLN, $\frac{1}{n}\sum M_{jkl}(X_i) \xrightarrow{p} \mathsf{E}[M_{jkl}(X_1)]$. Thus, with high probability:

$$|R_n| \leq Ba^3 \quad \text{(where } B > 0 \text{ is a constant)}.$$

Combining these terms, on the sphere Q_a:

$$\frac{1}{n}[\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0)] \leq \frac{1}{2}a^3 - Ca^2 + Ba^3 = a^2((B+1/2)a - C).$$

For sufficiently small a (specifically $a < \frac{C}{B+1/2}$), the negative quadratic term $-Ca^2$ dominates the cubic terms. Thus, the expression is strictly negative.

This proves that $\ell(\boldsymbol{\theta}) < \ell(\boldsymbol{\theta}_0)$ for all $\boldsymbol{\theta} \in Q_a$ with probability approaching 1. □

2. **Asymptotic Normality.**

Let $\tilde{\boldsymbol{\theta}}_n$ be the consistent sequence of roots found in part 1. We start by Taylor expanding the score function $\dot{\ell}(\tilde{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}_0$. Since $\tilde{\boldsymbol{\theta}}_n$ is a root, $\dot{\ell}(\tilde{\boldsymbol{\theta}}_n) = 0$.

For the j-th component:

$$0 = \dot{\ell}_j(\tilde{\boldsymbol{\theta}}_n) = \dot{\ell}_j(\boldsymbol{\theta}_0) + \left[\nabla \dot{\ell}_j(\boldsymbol{\theta}_0)\right]^\top (\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \frac{1}{2}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \nabla^2 \dot{\ell}_j(\boldsymbol{\theta}^*_{n,j})(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*_{n,j}$ lies between $\tilde{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. Since $\tilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, we also have $\boldsymbol{\theta}^*_{n,j} \xrightarrow{p} \boldsymbol{\theta}_0$.

Rearranging terms to isolate the difference $(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$:

$$-\frac{1}{\sqrt{n}}\dot{\ell}_j(\boldsymbol{\theta}_0) = \left[\frac{1}{n}\nabla \dot{\ell}_j(\boldsymbol{\theta}_0) + \text{Remainder}_j\right]^\top \sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).$$

We analyze the matrix in the brackets:

- The first term is the Hessian matrix of the log-likelihood (scaled by $1/n$). By the LLN:

$$\frac{1}{n}\nabla \dot{\ell}(\boldsymbol{\theta}_0) = \frac{1}{n}\ddot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{p} -I(\boldsymbol{\theta}_0).$$

- The remainder term involves third derivatives evaluated at $\boldsymbol{\theta}^*_{n,j}$. Using the bound $M_{jkl}(x)$ and the fact that $\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \xrightarrow{p} 0$, this term is $o_p(1)$.

Stacking all components $j = 1, \dots, d$, we get the vector equation:

$$-\frac{1}{\sqrt{n}}\dot{\ell}(\boldsymbol{\theta}_0) = [-I(\boldsymbol{\theta}_0) + o_p(1)]\,\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).$$

Multiplying by -1 and inverting the matrix (which converges to the invertible $I(\boldsymbol{\theta}_0)$):

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = [I(\boldsymbol{\theta}_0) + o_p(1)]^{-1}\,\frac{1}{\sqrt{n}}\dot{\ell}(\boldsymbol{\theta}_0).$$

By the Central Limit Theorem, the normalized score function converges in distribution:

$$\mathbf{Z}_n = \frac{1}{\sqrt{n}}\dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_d(\mathbf{0}, I(\boldsymbol{\theta}_0)).$$

Using Slutsky's Theorem (Theorem 9.3), we conclude:

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} I(\boldsymbol{\theta}_0)^{-1}\mathcal{N}_d(\mathbf{0}, I(\boldsymbol{\theta}_0)) = \mathcal{N}_d(\mathbf{0}, I(\boldsymbol{\theta}_0)^{-1}).$$

□

12.2 One-Step Estimator

In practice, the Maximum Likelihood Estimator (MLE) is often computed via iterative numerical optimization, typically using a Newton-type method. This requires an initialization and multiple iterations until convergence.

Interestingly, asymptotic theory suggests a remarkable shortcut: if we start with a "decent" preliminary estimator (one that is consistent and converges at a reasonable rate), performing just **one single Newton step** is sufficient to produce an estimator that shares the same asymptotic efficiency as the fully converged MLE.

Optimization vs. Statistical Noise

Why stop after one step?

Optimization algorithms are designed to find the peak of the log-likelihood function $\ell_n(\theta)$ with high numerical precision. However, from a statistical perspective, the location of this peak (the MLE $\hat{\theta}_n$) is itself a random variable that fluctuates around the true parameter θ_0 due to sampling noise. There is often little statistical gain in optimizing the likelihood function to a precision far beyond the inherent statistical noise of the problem. A one-step correction often brings the estimator within the range of statistical efficiency.

Let $\bar{\boldsymbol{\theta}}_n$ be a preliminary estimator. The one-step estimator $\check{\boldsymbol{\theta}}_n$ is defined by taking a single Newton step from $\bar{\boldsymbol{\theta}}_n$:

$$\check{\boldsymbol{\theta}}_n := \bar{\boldsymbol{\theta}}_n + \hat{\mathbf{I}}(\bar{\boldsymbol{\theta}}_n)^{-1}\frac{1}{n}\dot{\ell}(\bar{\boldsymbol{\theta}}_n). \tag{12.1}$$

Here, $\frac{1}{n}\dot{\ell}(\bar{\boldsymbol{\theta}}_n)$ is the gradient (score) scaled by sample size, and $\hat{\mathbf{I}}(\bar{\boldsymbol{\theta}}_n)$ is an estimator of the Fisher Information matrix. The Newton step essentially corrects the bias of the preliminary estimator using the curvature of the log-likelihood.

We can choose $\hat{\mathbf{I}}(\boldsymbol{\theta})$ in several ways, all consistent for the true Fisher Information:

$$\hat{\mathbf{I}}(\boldsymbol{\theta}) = \begin{cases} \mathbf{I}(\boldsymbol{\theta}) & \text{(“Expected Fisher Information” – Fisher Scoring),} \\ -\frac{1}{n}\sum_{i=1}^{n}\ddot{\ell}(\boldsymbol{\theta}|X_i) & \text{(“Observed Fisher Information” – Standard Newton),} \\ \frac{1}{n}\sum_{i=1}^{n}\dot{\ell}(\boldsymbol{\theta}|X_i)\dot{\ell}(\boldsymbol{\theta}|X_i)^\top & \text{(Covariance of the Score).} \end{cases} \tag{12.2}$$

Remark. Using the observed Fisher information (the negative Hessian) corresponds to the standard Newton-Raphson method. Using the expected Fisher information is known as the **Fisher Scoring** algorithm. Asymptotically, these methods are equivalent because all three matrices converge to $\mathbf{I}(\theta_0)$.

Theorem 12.2 (Asymptotic Normality of One-Step Estimator)**.** *If assumptions A0-A4 hold and the preliminary estimator satisfies the condition*

$$n^{1/4}(\bar{\theta}_n - \theta_0) = o_p(1),$$

then the one-step estimator satisfies:

$$\sqrt{n}(\check{\theta}_n - \theta_0) = \mathrm{I}(\theta_0)^{-1}\mathbf{Z}_n + o_p(1) \xrightarrow{d} \mathcal{N}_d(\mathbf{0}, \mathrm{I}(\theta_0)^{-1}).$$

This theorem states that $\check{\theta}_n$ is asymptotically efficient (it achieves the Cramér-Rao lower bound), just like the MLE. The condition $n^{1/4}$ is weaker than the standard $\sqrt{n}$-consistency; essentially, the starting guess needs to be “reasonably” close to the truth, but not necessarily optimal.

Proof. See, e.g., van der Vaart (1998, Theorem 5.45). The proof relies on Taylor expanding the score function and utilizing the consistency of the preliminary estimator. □

Example 12.2 (Gamma Distribution)**.** Consider the Gamma distribution model where $X_1, \dots, X_n \sim \text{Gamma}(\alpha, \beta)$ are i.i.d. observations. The parameter vector is $\boldsymbol{\theta} = (\alpha, \beta)^\top \in \mathbb{R}_+^2$.

The Method of Moments (MoM) provides a convenient and consistent preliminary estimator $\overline{\boldsymbol{\theta}}$. By matching the first two theoretical moments ($E[X] = \alpha/\beta$, $Var(X) = \alpha/\beta^2$) with the sample moments, we obtain closed-form estimators:

$$\overline{\boldsymbol{\theta}} = \begin{pmatrix} \bar{\alpha}_n \\ \bar{\beta}_n \end{pmatrix} = \left(\frac{\overline{X}_n^2}{\overline{X^2}_n - \overline{X}_n^2}, \quad \frac{\overline{X}_n}{\overline{X^2}_n - \overline{X}_n^2} \right)^T.$$

These estimators are $\sqrt{n}$-consistent (and thus satisfy the $n^{1/4}$ requirement), but they are not efficient compared to the MLE. We can improve them using the one-step method.

R implementation using ‘OneStep’ package

We simulate a dataset of size $n = 1000$ from a Gamma(5, 2) distribution and compare the MoM estimator with the one-step improved version.

```r
library(OneStep)
n <- 1000
set.seed(1234)

# True parameters
theta_true <- c(5, 2)
# Generate data
x <- rgamma(n, shape = theta_true[1], rate = theta_true[2])

# 1. Compute Method of Moments (Preliminary Estimator)
# Note: mean(x^2) - mean(x)^2 is the biased sample variance
alpha_bar <- mean(x)^2 / (mean(x^2) - mean(x)^2)
beta_bar <- mean(x) / (mean(x^2) - mean(x)^2)

theta_bar <- list(shape = alpha_bar, rate = beta_bar)
print(theta_bar)
# £shape
# [1] 4.925981
# £rate
# [1] 1.950755
```

The MoM estimates ($\approx 4.93, 1.95$) are decent but slightly off from the truth $(5, 2)$. Now we apply the one-step correction:

```r
# 2. Apply One-Step Correction
onestep(x, "gamma", init = theta_bar)
# Parameters:
#         estimate Std. Error
# shape   5.082690         NA
# rate    2.012814         NA
```

The improved estimate for β (≈ 2.01) is closer to the true value of 2. Interestingly, if we run ‘onestep’ without providing an initial value, it often defaults to using the Method of Moments internally:

```r
onestep(x, "gamma")
# Parameters:
#         estimate Std. Error
# shape   5.082690         NA
# rate    2.012814         NA
```

12.3 Asymptotic Confidence Intervals

We have derived the asymptotic normality of the MLE primarily to perform formal uncertainty assessments. Specifically, we want to construct confidence intervals for the true parameter θ_0.

12.3.1 Using Asymptotic Normality as a Pivot

Consider a one-parameter model ($d = 1$), so $\Theta \subseteq \mathbb{R}$. Suppose $\tilde{\theta}_n$ is an estimator (like the MLE) such that for all $\theta_0 \in \Theta$:

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I(\theta_0)^{-1}).$$

By Slutsky's Theorem (Theorem 9.3), we can standardize this convergence. If we multiply by the square root of the Fisher Information $I(\theta_0)^{1/2}$, we obtain a pivotal quantity in the asymptotic sense:

$$\sqrt{n} I(\theta_0)^{1/2}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1). \tag{12.3}$$

Constructing the Interval We treat the left-hand side of (12.3) as a pivot. Let $z_{\alpha/2}$ be the $(1 - \alpha/2)$ quantile of the standard normal distribution (e.g., 1.96 for $\alpha = 0.05$). We define the confidence set S_α as the collection of all parameter values θ for which the pivot falls within the standard normal quantiles:

$$S_\alpha = \left\{ \theta \in \Theta : \left| \sqrt{n} I(\theta)^{1/2}(\tilde{\theta}_n - \theta) \right| \le z_{\alpha/2} \right\}. \tag{12.4}$$

By construction, the coverage probability of this set converges to the target level:

$$\mathsf{P}_\theta(\theta \in S_\alpha) \underset{n \to \infty}{\longrightarrow} \mathsf{P}(|\mathcal{N}(0,1)| \le z_{\alpha/2}) = 1 - \alpha \quad \forall \theta \in \Theta.$$

Thus, S_α is an asymptotic $(1 - \alpha)$ confidence interval.

Example 12.3 (Wilson Interval for Binomial Proportions)**.** If the model is Binomial and θ is the success probability, this approach yields the **Wilson interval**. In this case, the Fisher information $I(\theta)$ depends on θ in a non-linear way ($I(\theta) = \frac{1}{\theta(1-\theta)}$). Solving the inequality in (12.4) requires solving a quadratic equation in θ.

The Wilson interval has interesting properties compared to simpler intervals (like the Wald interval discussed next). For instance, even if you observe extreme data (e.g., 0 successes in n trials), the Wilson interval provides a non-trivial, positive-length interval, reflecting the inherent uncertainty more accurately than degenerate intervals.

Note. While conceptually sound, solving the inequality (12.4) for θ can be analytically difficult because θ appears in both the linear term $(\tilde{\theta}_n - \theta)$ and the information term $I(\theta)^{1/2}$. In general, S_α is not guaranteed to be a single connected interval, though it often is in exponential families like the Binomial.

12.3.2 Wald Intervals

A very tractable construction of confidence intervals, which is the default in most statistical software, results from estimating the Fisher information. By Slutsky's theorem, we can replace the true Fisher information $I(\theta_0)$ with a consistent estimator $\hat{\mathbf{I}}(\tilde{\theta}_n)$ (e.g., the observed Fisher information evaluated at the MLE).

We have:

$$\sqrt{n}\,\hat{\mathbf{I}}(\tilde{\theta}_n)^{1/2}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0,1).$$

Based on this, the **Wald interval** is an asymptotic $(1-\alpha)$ confidence interval defined as:

$$W_\alpha = \left\{\theta : \left|\sqrt{n}\,\hat{\mathbf{I}}(\tilde{\theta}_n)^{1/2}(\tilde{\theta}_n - \theta)\right| \le z_{\alpha/2}\right\} = \left[\tilde{\theta}_n \pm z_{\alpha/2} \cdot \frac{1}{\sqrt{n\hat{\mathbf{I}}(\tilde{\theta}_n)}}\right].$$

This interval is symmetric around the estimator $\tilde{\theta}_n$.

Multivariate Case $(d \ge 2)$ This approach generalizes easily to intervals for individual components of a higher-dimensional parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$. The asymptotic normality result

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_d(0, I(\boldsymbol{\theta}_0)^{-1})$$

implies convergence of the marginal distributions. For the j-th component:

$$\sqrt{n}(\tilde{\theta}_{nj} - \theta_{0j}) \xrightarrow{d} \mathcal{N}\left(0, \left(\mathrm{I}(\boldsymbol{\theta}_0)^{-1}\right)_{jj}\right).$$

Thus, an interval for the j-th component of $\boldsymbol{\theta}$ is formed as:

$$W_{\alpha,j} = \left[\tilde{\theta}_{nj} \pm z_{\alpha/2} \cdot \sqrt{\frac{(\hat{\mathbf{I}}(\tilde{\boldsymbol{\theta}}_n)^{-1})_{jj}}{n}}\right].$$

Here, $(\hat{\mathbf{I}}(\tilde{\boldsymbol{\theta}}_n)^{-1})_{jj}$ is the j-th diagonal element of the *inverse* of the estimated Fisher information matrix.

Example 12.4 (Wald intervals for the Gamma model)**.** We simulate data from a Gamma distribution and compute Wald intervals for its parameters.

1. Computing the Interval: First, we define the negative log-likelihood and simulate 150 observations from a Gamma(5, 2) distribution.

```r
# Define negative log-likelihood (without 1/n scaling)
negloglik <- function(theta, data) {
  alpha <- theta[1]
  beta <- theta[2]
  n <- length(data)
  -sum(dgamma(data, shape = alpha, rate = beta, log = TRUE))
}

set.seed(22)
n <- 150
x <- rgamma(n, shape = 5, rate = 2)
```

We optimize the likelihood and extract the Hessian (observed Fisher information) at the optimum.

```r
# hessian=TRUE requests the matrix of second derivatives
opt <- optim(c(1,1), negloglik, data = x, hessian = TRUE)
hat.theta <- opt$par
```

We calculate the standard errors. Note that ‘`opt$hessian`‘ computes $-\sum \ddot{\ell}$, so we must scale by $1/n$ to get the observed Fisher information $\hat{\mathbf{I}}$, invert it, and then scale by $1/n$ again for the variance of the estimator. Or equivalently, invert the raw Hessian directly. The standard error (SE) is the square root of the diagonal elements of the inverse Hessian.

```r
# Inverse of the observed Fisher information matrix for the full sample
# opt£hessian is roughly n * I(theta)
var_matrix <- solve(opt$hessian)
se <- sqrt(diag(var_matrix))
```

Using the 0.975 quantile of $\mathcal{N}(0, 1)$, we construct the intervals:

```r
z975 <- qnorm(0.975)
cbind(mle = hat.theta,
      lower = hat.theta - z975 * se,
      upper = hat.theta + z975 * se)
#            mle    lower    upper
# [1,] 5.458209 4.258774 6.657645  (alpha)
# [2,] 2.140010 1.647429 2.632591  (beta)
```

2. Checking Coverage Probability: To verify the validity of this interval, we can simulate the process 1000 times and check how often the true parameter $\alpha = 5$ falls within the calculated interval.

```r
wald_ci_alpha_coverage <- function(n = 150, alpha = 5, beta = 2, conf.level = 0.95){
    x <- rgamma(n, shape = alpha, rate = beta)
    opt <- optim(c(1,1), negloglik, data = x, hessian = TRUE)

    hat.theta <- opt$par
    # Standard error for alpha (1st parameter)
    se_alpha <- sqrt(solve(opt$hessian)[1,1])

    quant <- qnorm(1 - (1 - conf.level)/2)

    # Check if true alpha is in interval
    lower <- hat.theta[1] - quant * se_alpha
    upper <- hat.theta[1] + quant * se_alpha

    return(alpha >= lower && alpha <= upper)
}

set.seed(22)
mean(replicate(1000, wald_ci_alpha_coverage()))
# [1] 0.944
```

The estimated coverage is 94.4%, which is very close to the nominal 95% level, confirming the asymptotic theory works well for moderate sample sizes ($n = 150$) in this regular model.

12.3.3 Optimality of the MLE

The MLE achieves the Cramér-Rao bound in an asymptotic sense: Its asymptotic covariance matrix is the inverse Fisher information. An estimator with this property is referred to as being **asymptotically efficient**.

However, making a rigorous statement about the asymptotic optimality of the MLE proved to be surprisingly difficult. While Fisher conjectured this efficiency in the 1920s, formal proofs took decades to materialize. The issue was settled in the 1960s and 70s with the **Hájek-Le Cam convolution theorem**.

Note. The convolution theorem:

- Was proven decades after Fisher's initial conjecture.
- Restricts attention to **regular estimators** to avoid pathological counterexamples (see below).
- For details, see van der Vaart (1998, Chap. 8).

Why all this complexity? The main hurdle was the existence of "superefficient" estimators—estimators that seemingly outperform the MLE (have lower asymptotic variance) at specific points in the parameter space.

Example 12.5 (Hodges' Superefficient Estimator)**.** Suppose $d = 1$. Let $\hat{\theta}_n$ be the standard MLE such that for all $\theta \in \Theta$:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1}).$$

Fix a specific value $\theta_0 \in \Theta$. Define the Hodges estimator $\hat{\theta}_n^H$ as:

$$\hat{\theta}_n^H := \begin{cases} \hat{\theta}_n & \text{if } n^{1/4}|\hat{\theta}_n - \theta_0| > 1, \\ \theta_0 & \text{if } n^{1/4}|\hat{\theta}_n - \theta_0| \leqslant 1. \end{cases}$$

Essentially, this estimator checks if the MLE is close to θ_0. If it is very close (within a $n^{-1/4}$ window), it snaps the estimate exactly to θ_0.

As shown in Ferguson (1996, Example 1 on p. 134), the asymptotic distribution is:

$$\sqrt{n}(\hat{\theta}_n^H - \theta) \xrightarrow{d} \mathcal{N}(0, V(\theta)), \quad \text{where } V(\theta) = \begin{cases} I(\theta)^{-1} & \text{if } \theta \neq \theta_0, \\ 0 & \text{if } \theta = \theta_0. \end{cases}$$

At $\theta = \theta_0$, the asymptotic variance is 0, which is strictly better than the MLE's variance $I(\theta_0)^{-1}$. Everywhere else, it matches the MLE. Thus, $\hat{\theta}_n^H$ appears to be "superefficient."

Why is this considered an artifact? While asymptotically superior at a single point, the Hodges estimator has terrible finite-sample performance near θ_0. The risk (expected squared error) becomes very large in the transition region where the estimator switches between regimes. This superefficiency is an artifact of pointwise asymptotic comparisons that fails to capture the local behavior of the estimator.

Regular Estimators and the Convolution Theorem To exclude pathological cases like the Hodges estimator, the Hájek-Le Cam theorem restricts the class of competitors to **regular estimators**.

Definition 12.2 (Regular Estimator)**.** An estimator T_n is regular if its asymptotic distribution is insensitive to small local perturbations of the parameter. Specifically, for local alternatives of the form $\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \mathbf{t}/\sqrt{n}$, we require that the limiting distribution of $\sqrt{n}(T_n - \boldsymbol{\theta}_n)$ does not depend on $\mathbf{t}$.

Under our standard assumptions, the MLE $\tilde{\boldsymbol{\theta}}_n$ is a regular estimator with asymptotic distribution $\mathcal{N}(0, \mathbf{I}(\boldsymbol{\theta}_0)^{-1})$. The convolution theorem states that among all regular estimators, the MLE has the smallest asymptotic variance (in a matrix sense). Any other regular estimator will have an asymptotic distribution equal to the MLE's distribution *convolved* with independent noise, necessarily increasing the variance.

12.3.4 Kullback-Leibler Divergence

The MLE can also be thought of as a plug-in estimator in the context of finding a distribution in the model that is as close as possible to the true data-generating distribution. This leads us to the concept of Kullback-Leibler (KL) divergence.

Remark. Thinking about minimization of KL divergence offers a perspective that is helpful, in particular, when the model under consideration is **misspecified** (i.e., does not contain the true data-generating distribution). For example, if we model counts as Poisson when they are actually Negative Binomial, what is the MLE estimating?

Definition 12.3 (Kullback-Leibler Divergence)**.** Let P and Q be probability distributions on $(\mathcal{X}, \mathcal{A})$ with densities p and q w.r.t. a σ-finite measure ν. The Kullback-Leibler divergence from P to Q is

$$KL(\mathsf{P}|\mathsf{Q}) = \mathsf{E}_{\mathsf{P}}\left[\log \frac{p(X)}{q(X)}\right].$$

(It does not depend on the choice of ν.)

Note. KL divergence is not a distance/metric (it is not even symmetric, i.e., $KL(\mathsf{P}|\mathsf{Q}) \neq KL(\mathsf{Q}|\mathsf{P})$).

Lemma 12.1. *KL divergence is well-defined (as an integral) and*

$$KL(\mathsf{P}|\mathsf{Q}) \begin{cases} \in [0,\infty] & \textit{always}, \\ = 0 \iff & \mathsf{P} = \mathsf{Q}. \end{cases}$$

TODO: expand proofs with x.

Proof (Well-definedness). Recall that $\log(x) \le x - 1$ for all $x > 0$. The KL divergence is well-defined if the negative part of $\log \frac{p(X)}{q(X)}$ has finite expectation under P.

This holds because:

$$\begin{aligned}\mathsf{E}_\mathsf{P}\Big[\max\Big\{-\log\frac{p(X)}{q(X)},0\Big\}\Big] &= \int_{\{p\leq q\}}\Big(\log\frac{q}{p}\Big)p\,\mathrm{d}\nu\\ &\leqslant \int_{\{p\leq q\}}\Big(\frac{q}{p}-1\Big)p\,\mathrm{d}\nu\\ &\leqslant \int_{\{p\leq q\}}(q-p)\,\mathrm{d}\nu \leqslant \int q\,\mathrm{d}\nu = 1.\end{aligned}$$

□

Proof (Non-negativity).

$$\begin{aligned}\mathsf{E}_\mathsf{P}\Big[\log\frac{p(X)}{q(X)}\Big] &= -\mathsf{E}_\mathsf{P}\Big[\log\frac{q(X)}{p(X)}\Big] \geqslant -\mathsf{E}_\mathsf{P}\Big[\frac{q(X)}{p(X)}-1\Big]\\ &= -\int_{\{p>0\}} q\,\mathrm{d}\nu + 1 \geqslant -1+1 = 0.\end{aligned}$$

□

Note. For $KL(\mathsf{P}|\mathsf{Q}) < \infty$ it is necessary (but not sufficient) that $\mathsf{P} \ll \mathsf{Q}$, i.e., $\mathsf{Q}(A) = 0 \implies \mathsf{P}(A) = 0$ for all $A \in \mathcal{A}$.

Example 12.6.

$$\begin{aligned}p(x) &= \mathbf{1}_{(0,1)}(x),\\ q(x) &\propto e^{-1/x}\mathbf{1}_{(0,1)}(x).\end{aligned}$$

TODO: this proof is way longer.

Proof (Minimizer). Clearly $\mathsf{P} = \mathsf{Q} \implies KL = 0$. For the converse, if $KL(\mathsf{P}|\mathsf{Q}) = 0$, then the non-negativity proof implies $\mathsf{E}_\mathsf{P}[\frac{q(X)}{p(X)} - 1 - \log\frac{q(X)}{p(X)}] = 0$. Since $x - 1 - \log x \geq 0$ with equality if and only if $x = 1$, we must have $q(X)/p(X) = 1$ a.s. $[\mathsf{P}]$, implying $\mathsf{P} = \mathsf{Q}$. □

Back to Estimation

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathsf{P}$ (the true data-generating distribution), and we model this with a family $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}$.

Which P_θ is closest to P in the KL sense?

$$\begin{aligned}\min_{\theta\in\Theta} KL(\mathsf{P}|\mathsf{P}_\theta) &= \min_{\theta\in\Theta}\mathsf{E}_\mathsf{P}\Big[\log\frac{p(X)}{p_\theta(X)}\Big]\\ &= \mathsf{E}_\mathsf{P}[\log p(X)] + \min_{\theta\in\Theta}\{-\mathsf{E}_\mathsf{P}[\log p_\theta(X)]\}\\ &= \text{const.} - \max_{\theta\in\Theta}\mathsf{E}_\mathsf{P}[\log p_\theta(X)].\end{aligned}$$

If we replace the true expectation E_{P} with the empirical average (plug-in the empirical distribution $\widehat{\mathsf{P}}$), we get:

$$\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i).$$

This is exactly the log-likelihood function!

Remark (Interpretation under Misspecification). We can interpret the MLE $\hat{\theta}_{\text{MLE}}$ as the parameter that minimizes the KL divergence between the empirical distribution and the model family.

If the model is **misspecified** (i.e., $\mathsf{P} \notin \mathcal{P}$), the MLE converges to the parameter θ^* that minimizes $KL(\mathsf{P}|\mathsf{P}_\theta)$.

Example 12.7 (Linear Regression)**.** Suppose we fit a linear model $Y = \beta_0 + \beta_1 X + \epsilon$, but the true relationship is non-linear (e.g., quadratic). The MLE for the slope β_1 estimates the slope of the **best linear approximation** to the true curve in the KL sense (which corresponds to the best L^2 approximation for Gaussian errors).

This theory is further developed in the context of **sandwich estimation**, where the asymptotic variance of the MLE takes a different form ("sandwich variance") because the cancellation between the Fisher information and the Hessian of the log-likelihood no longer holds.

13. Likelihood Ratio, Wald, and Score Tests

This chapter introduces a "companion" methodology to maximum likelihood estimation: the theory of likelihood-based testing. Just as MLEs provide a general framework for estimation, likelihood theory provides a unified approach to hypothesis testing. Specifically, we will discuss the **Likelihood Ratio Test**, the **Wald Test**, and **Rao's Score Test** (sometimes called the Lagrange Multiplier Test in economics). These three form the "Trinity of Tests."

13.1 Trinity of Tests for Simple Null Hypotheses

Consider the standard setup where we observe an i.i.d. sample $X_1, \ldots, X_n$ from a distribution in a parametric model $\mathsf{P} = \{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, with $\boldsymbol{\theta} \subseteq \mathbb{R}^d$. Let $\hat{\boldsymbol{\theta}}_n$ be the MLE of $\boldsymbol{\theta}$.

We wish to test a simple null hypothesis:

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

We define three test statistics to address this problem.

a) **Likelihood Ratio Statistic (λ_n)**

This statistic compares the maximum likelihood achievable under the full model to the likelihood under the null hypothesis.

$$\lambda_n \equiv \lambda_n(\hat{\boldsymbol{\theta}}_n) = \frac{L(\hat{\boldsymbol{\theta}}_n)}{L(\boldsymbol{\theta}_0)} = \frac{\sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)}.$$

Intuition: The statistic λ_n measures how much more likely the data are when we allow the parameter to vary freely compared to when it is fixed at θ_0.

- Since the numerator maximizes over all $\boldsymbol{\theta} \in \Theta$ (including $\boldsymbol{\theta}_0$), we always have $\lambda_n \geq 1$.
- If the data are well-explained by $\boldsymbol{\theta}_0$, λ_n will be close to 1.
- If the data are much better explained by some alternative $\theta \neq \theta_0$, λ_n will be large.

Later, we will often work with the log-scale version, $2\log\lambda_n$, for distributional reasons.

b) **Wald Statistic (W_n)**

This statistic directly compares the MLE $\hat{\boldsymbol{\theta}}_n$ to the hypothesized value $\boldsymbol{\theta}_0$. It measures the squared distance between them, weighted by the curvature of the likelihood (information).

$$W_n \equiv W_n(\hat{\boldsymbol{\theta}}_n) = n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).$$

Here, $\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_n)$ is a consistent estimator of the Fisher information (e.g., observed information). This is essentially the squared Mahalanobis distance between the estimate and the hypothesis.

c) **Rao's Score Statistic (R_n)**

This statistic evaluates the slope (gradient) of the log-likelihood function at the hypothesized value $\boldsymbol{\theta}_0$.

$$R_n = \frac{1}{n}\dot{\ell}(\boldsymbol{\theta}_0)^\top \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\dot{\ell}(\boldsymbol{\theta}_0).$$

Intuition: If $\hat{\boldsymbol{\theta}}_n$ is the maximizer, the gradient $\dot{\ell}(\hat{\boldsymbol{\theta}}_n)$ is zero. If $\boldsymbol{\theta}_0$ is close to the maximizer (i.e., H_0 is true), the gradient $\dot{\ell}(\boldsymbol{\theta}_0)$ should be close to zero. If $\boldsymbol{\theta}_0$ is far from the truth, the gradient will be large.

We will study the asymptotic behavior of these statistics as $n \to \infty$ under the regularity assumptions from section 12.1 . Specifically, we will define the statistics using a consistent root of the likelihood equations, $\tilde{\boldsymbol{\theta}}_n$:

$$\widetilde{\lambda}_n := \lambda_n(\widetilde{\boldsymbol{\theta}}_n), \quad \widetilde{W}_n := W_n(\widetilde{\boldsymbol{\theta}}_n).$$

13.1.1 Chi-Square Limits

Under the null hypothesis, all three statistics converge to the same limiting distribution.

Theorem 13.1 (Null distribution). *Suppose H_0 is true, i.e., $X_1, X_2, \dots$ are i.i.d. P_{θ_0}. Suppose further assumptions A0-A4 hold at θ_0, and that the estimator $\tilde{\theta}_n$ satisfies*

$$\tilde{\theta}_n \xrightarrow{p} \theta_0$$

and

$$\mathsf{P}_{\theta_0}(\tilde{\theta}_n \text{ solves the likelihood eq.}) \longrightarrow 1$$

for $n \to \infty$. Then

$$\left.\begin{matrix} 2\log\tilde{\lambda}_n \\ \widetilde{W}_n \\ R_n \end{matrix}\right\} \xrightarrow{d} \mathbf{Z}^\top \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{Z} \sim \chi^2_d,$$

where $\mathbf{Z} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0))$.

Remark.

1. **Asymptotic Equivalence:** The three statistics are asymptotically equivalent. The differences between them converge to zero in probability:
$$2\log\tilde{\lambda}_n - \widetilde{W}_n = o_p(1), \quad 2\log\tilde{\lambda}_n - R_n = o_p(1), \quad R_n - \widetilde{W}_n = o_p(1).$$

2. **Finite Sample Differences:** While asymptotically identical, they can differ significantly for finite n. For example, in high-dimensional settings where n is small relative to d, the likelihood ratio and Wald statistics might be undefined (unbounded likelihood), while Rao's score statistic might still be well-defined.

3. **Computational Differences:**
   - **Likelihood Ratio:** Requires optimizing the likelihood (finding the max).
   - **Wald:** Requires finding the MLE $\hat{\theta}_n$ (optimizing).
   - **Score:** Does **not** require optimization; only requires evaluating the gradient at θ_0. This makes it computationally cheapest.

Proof. Recall that the multivariate CLT gives
$$\mathbf{Z}_n := \frac{1}{\sqrt{n}}\dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{Z} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0)).$$

Rao's score statistic (R_n): By definition, R_n is a quadratic form in the normalized score $\mathbf{Z}_n$:
$$R_n = \mathbf{Z}_n^\top \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{Z}_n.$$

By the Continuous Mapping Theorem and the properties of quadratic forms of normal vectors (Lemma 11.1), this converges directly to a χ^2_d distribution:
$$R_n \xrightarrow{d} \mathbf{Z}^\top \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{Z} \sim \chi^2_d.$$

Wald statistic ($\widetilde{W}_n$): We use the asymptotic linearity of the estimator derived in Chapter 13:
$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = \mathbf{I}(\theta_0)^{-1}\mathbf{Z}_n + o_p(1).$$

Substituting this into the definition of W_n and using the consistency of the information estimator $\hat{\mathbf{I}}(\tilde{\theta}_n) \xrightarrow{p} \mathbf{I}(\theta_0)$:
$$\begin{aligned}\widetilde{W}_n &= \sqrt{n}\big(\widetilde{\theta}_n - \theta_0\big)^\top \hat{\mathbf{I}}(\widetilde{\theta}_n)\sqrt{n}\big(\widetilde{\theta}_n - \theta_0\big) \\ &\approx (\mathbf{I}(\theta_0)^{-1}\mathbf{Z}_n)^\top \mathbf{I}(\theta_0)(\mathbf{I}(\theta_0)^{-1}\mathbf{Z}_n) \\ &= \mathbf{Z}_n^\top \mathbf{I}(\theta_0)^{-1}\mathbf{Z}_n + o_p(1).\end{aligned}$$

This matches the form of the Score statistic, so it also converges to χ^2_d.

Likelihood ratio statistic ($2\log\tilde{\lambda}_n$): Let G_n be the event that $\tilde{\theta}_n$ is a consistent root. On G_n, we perform a second-order Taylor expansion of the log-likelihood $\ell(\theta_0)$ around the maximizer $\widetilde{\theta}_n$:
$$\ell(\boldsymbol{\theta}_0) \approx \ell(\tilde{\boldsymbol{\theta}}_n) + \dot{\ell}(\tilde{\boldsymbol{\theta}}_n)^\top(\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_n)^\top \ddot{\ell}(\boldsymbol{\theta}_n^*)(\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_n).$$

Since $\tilde{\boldsymbol{\theta}}_n$ is a critical point, the linear term vanishes: $\dot{\ell}(\tilde{\boldsymbol{\theta}}_n) = 0$. Rearranging for the log-likelihood ratio:

$$2\log\tilde{\lambda}_n = 2\big[\ell(\tilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\big] \approx -(\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_n)^\top \ddot{\ell}(\boldsymbol{\theta}_n^*)(\boldsymbol{\theta}_0 - \tilde{\boldsymbol{\theta}}_n).$$

By the Law of Large Numbers, $-\frac{1}{n}\ddot{\ell}(\boldsymbol{\theta}_n^*) \xrightarrow{p} \mathbf{I}(\boldsymbol{\theta}_0)$. Thus:

$$2\log\tilde{\lambda}_n \approx \sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{I}(\boldsymbol{\theta}_0)\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).$$

This is exactly the asymptotic form of the Wald statistic, so it also converges to χ^2_d. □

13.1.2 Tests

a) The likelihood ratio (LR) test rejects H_0 if

$$2\log\tilde{\lambda}_n \geq F^{-1}_{\chi^2_d}(1-\alpha). \tag{13.1}$$

b) The Wald test rejects H_0 if

$$\widetilde{W}_n \geq F^{-1}_{\chi^2_d}(1-\alpha). \tag{13.2}$$

c) Rao's score test rejects H_0 if

$$R_n \geq F^{-1}_{\chi^2_d}(1-\alpha). \tag{13.3}$$

Corollary 13.1. *The likelihood ratio test, the Wald test, and Rao's score test have asymptotic size α.*

Proof. All three test statistics are non-negative random variables and converge in distribution, under H_0, to a chi-square distribution. Consequently, the corresponding type I error probabilities converge to α. □

Example 13.1 (Likelihood ratio test for the Gamma model)**.** We compute the p-value for a test of

$$H_0 : \alpha = 4, \quad \beta = 1.5,$$

based on a sample from the Gamma$(5, 2)$ distribution.

Recall the negative log-likelihood for the Gamma model.

```r
negloglik <- function(ab, data){
  a <- ab[1]
  b <- ab[2]
  n <- length(data)
  return(
    -(n * a * log(b)
      + (a - 1) * sum(log(data))
      - b * sum(data)
      - n * lgamma(a))
  )
}
```

We minimize the negative log-likelihood $-\log L(\alpha, \beta)$.

```r
set.seed(22)
n <- 150
x <- rgamma(n, shape = 5, rate = 2)
opt1 <- optim(c(1,1), negloglik, data = x)
```

The likelihood ratio statistic is

$$2\log\tilde{\lambda}_n = 2\big(\log L(\tilde{\alpha}, \tilde{\beta}) - \log L(4, 1.5)\big). \tag{13.4}$$

```r
lambda <- 2 * (-opt1$value - (-negloglik(c(4,1.5), x)))
```

The resulting p-value is

```r
p.value <- 1 - pchisq(lambda, df = 2)
p.value
# [1] 0.01687918
```

Example 13.2 (Null Distribution of the LR Statistic)**.** We simulate the null distribution for a test of

$$H_0 : \alpha = 5, \quad \beta = 2,$$

based on samples from the Gamma$(5, 2)$ distribution.

We first write a function to compute the LR statistic and the corresponding p-value.

```r
lr_test_gamma <- function(alpha = 1, beta = 1, n = 100){
  x <- rgamma(n, shape = alpha, rate = beta)
  opt <- optim(c(1,1), negloglik, data = x)
  stat <- 2 * (-opt$value - (-negloglik(c(alpha,beta), x)))
  pval <- 1 - pchisq(stat, df = 2)
  data.frame(stat = stat, pval = pval)
}
```

We repeat this procedure 10,000 times.

```r
set.seed(22)
df <- 1:10000 |>
  purrr::map_df(\(i) lr_test_gamma(alpha = 5, beta = 2))
```

The empirical distribution function of the LR statistic is compared to its asymptotic χ^2_2 limit, and the histogram of simulated p-values is examined in Figure 13.1.

```r
p1 <- ggplot(df, aes(x = stat)) +
  stat_ecdf(geom = "step") +
  geom_function(
    fun = pchisq,
    args = list(df = 2),
    color = "red",
    linewidth = 1.3,
    linetype = 2
  ) +
  labs(
    title = "Empirical and asymptotic distribution of LR statistic",
    x = expression(2 * log(tilde(lambda)[n])),
    y = "empirical / asymptotic distribution function"
  )
```

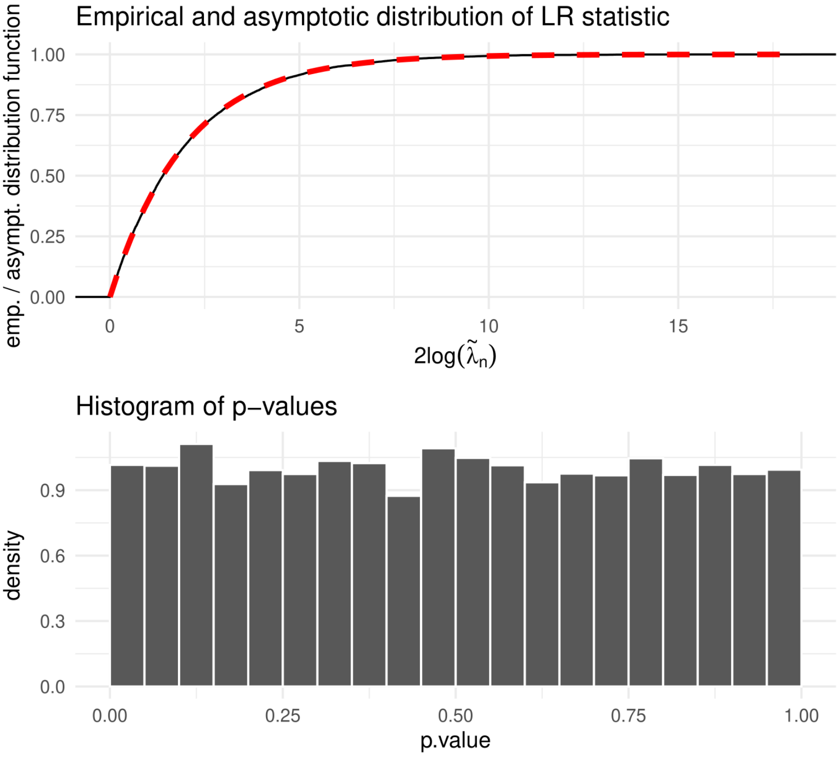


Figure 13.1

Remark. Under H_0, the simulated p-values are approximately uniformly distributed on $[0, 1]$, confirming that the likelihood ratio test is well calibrated even for moderate sample sizes such as $n = 100$.

Interpretation and Use in Regression Models

Note. Likelihood ratio tests are routinely implemented in statistical software. In regression settings, such as logistic regression, they are commonly used to compare nested models by testing whether a subset of regression coefficients is equal to zero.

Example 13.3 (Likelihood ratio tests in logistic regression)**.** Suppose a categorical variable such as age group is encoded using two dummy variables. Testing whether age has an effect corresponds to testing whether both associated regression coefficients are zero. This hypothesis is naturally assessed using a likelihood ratio test with a chi-square reference distribution, where the degrees of freedom equal the number of constrained parameters.

13.1.3 Fixed Alternatives

Theorem 13.2 (Fixed Alternatives)**.** *Suppose H_0 is false and $X_1, X_2, \dots$ are i.i.d. with distribution $\mathsf{P}_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Assume that regularity conditions A0–A4 hold at both $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Then, as $n \to \infty$,*

$$\frac{1}{n} 2 \log \widetilde{\lambda}_n \xrightarrow{p} 2\,\mathsf{KL}(\mathsf{P}_{\boldsymbol{\theta}} \mid \mathsf{P}_{\boldsymbol{\theta}_0}) \equiv 2\,\mathsf{E}_{X \sim \mathsf{P}_{\boldsymbol{\theta}}}\left[\log \frac{p_{\boldsymbol{\theta}}(X)}{p_{\boldsymbol{\theta}_0}(X)}\right] > 0, \tag{13.5}$$

and

$$\frac{1}{n}\widetilde{W}_n \xrightarrow{p} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \mathsf{I}(\boldsymbol{\theta})(\boldsymbol{\theta} - \boldsymbol{\theta}_0) > 0. \tag{13.6}$$

If in addition:

$$A5: \qquad \mathsf{E}_\theta\big[|\dot{\ell}_j(\theta_0 \mid X_1)|\big] < \infty \quad \forall j = 1, \dots, d,$$

then

$$\frac{1}{n} R_n \xrightarrow{p} \mathsf{E}_\theta\big[\dot{\ell}(\theta_0 \mid X_1)\big]^\top \mathsf{I}(\theta_0)^{-1} \mathsf{E}_\theta\big[\dot{\ell}(\theta_0 \mid X_1)\big] > 0. \tag{13.7}$$

Remark. The ordering in the Kullback–Leibler divergence in (13.5) is essential. Since the data are generated from P_θ, expectations are taken with respect to P_θ, and the divergence must be $\mathsf{KL}(\mathsf{P}_\theta \mid \mathsf{P}_{\theta_0})$. If $\theta \neq \theta_0$, this quantity is strictly positive.

Remark. Equations (13.5) and (13.6) show that, under fixed alternatives, the likelihood ratio and Wald statistics grow linearly with n, up to stochastic fluctuations. In contrast, under the null hypothesis, these statistics have non-degenerate stochastic limits.

Proof. **Likelihood ratio statistic.** Write

$$\frac{2}{n} \log \widetilde{\lambda}_n = \frac{2}{n}\big[\ell(\widetilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta})\big] + \frac{2}{n}\big[\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0)\big].$$

Since $2[\ell(\widetilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta})] \xrightarrow{d} \chi^2_d$, we have

$$\frac{2}{n}\big[\ell(\widetilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta})\big] = o_p(1).$$

Moreover,

$$\frac{2}{n}\big[\ell(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta}_0)\big] = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_{\boldsymbol{\theta}}(X_i)}{p_{\boldsymbol{\theta}_0}(X_i)} \xrightarrow{p} 2\,\mathsf{KL}(\mathsf{P}_{\boldsymbol{\theta}} \mid \mathsf{P}_{\boldsymbol{\theta}_0}),$$

by the law of large numbers, proving (13.5).

Wald statistic. By Theorem 9.2,

$$\frac{1}{n}\widetilde{W}_n = (\widetilde{\boldsymbol{\theta}}_n - \theta_0)^\top \widehat{\mathsf{I}}(\widetilde{\boldsymbol{\theta}}_n)(\widetilde{\boldsymbol{\theta}}_n - \theta_0) \xrightarrow{p} (\boldsymbol{\theta} - \theta_0)^\top \mathsf{I}(\boldsymbol{\theta})(\boldsymbol{\theta} - \theta_0),$$

establishing (13.6).

Rao's score statistic. We may write

$$\frac{1}{n}R_n = \left(\frac{1}{n}\dot{\ell}(\boldsymbol{\theta}_0)\right)^\top \mathsf{I}(\boldsymbol{\theta}_0)^{-1}\left(\frac{1}{n}\dot{\ell}(\boldsymbol{\theta}_0)\right).$$

Thus it suffices to show

$$\frac{1}{n}\dot{\ell}(\boldsymbol{\theta}_0) \xrightarrow{p} \mathsf{E}_{\boldsymbol{\theta}}\big[\dot{\ell}(\boldsymbol{\theta}_0 \mid X_1)\big].$$

By assumption A5, the expectation is finite, and the law of large numbers applies. This yields (13.7). □

Note (Interpretation)**.** Under fixed alternatives, all three classical test statistics (likelihood ratio, Wald, and Rao) diverge at rate n. In contrast, under the null hypothesis, they converge in distribution to non-degenerate χ^2 limits.

13.2 Consistency

Corollary 13.2 (Consistency of Likelihood-Based Tests)**.** *Under the assumptions of Theorem 13.2, the likelihood ratio and Wald tests are consistent. That is, if $\boldsymbol{\theta} \neq \theta_0$,*

$$\mathsf{P}_{\boldsymbol{\theta}}(\textit{rejection of } H_0) \longrightarrow 1 \quad \textit{as } n \to \infty.$$

The same conclusion holds for Rao's score test provided $\mathsf{E}_{\boldsymbol{\theta}}[\dot{\ell}(\theta_0 \mid X_1)] \neq 0$, in which case

$$\mathsf{E}_{\boldsymbol{\theta}}\big[\dot{\ell}(\boldsymbol{\theta}_0 \mid X_1)\big]^\top \mathsf{I}(\boldsymbol{\theta}_0)^{-1}\mathsf{E}_{\boldsymbol{\theta}}\big[\dot{\ell}(\boldsymbol{\theta}_0 \mid X_1)\big] > 0.$$

Proof. The argument is analogous to the proof of Theorem 13.2. Under fixed alternatives, the test statistics divided by n converge in probability to strictly positive constants, whereas under the null hypothesis they have non-degenerate stochastic limits. Hence, with a fixed rejection threshold, the probability of rejection converges to one. □

Remark (Why consistency holds). Under H_0, the likelihood ratio, Wald, and Rao statistics fluctuate in a bounded stochastic range (e.g. they converge in distribution to a χ^2 law). If H_0 is false, each statistic behaves like a positive constant times n, up to random noise. Since the rejection threshold is fixed, the statistic eventually exceeds it with probability tending to one.

Example 13.4 (Geometric intuition)**.**

- The likelihood ratio statistic measures separation between $\mathsf{P}_{\boldsymbol{\theta}}$ and $\mathsf{P}_{\boldsymbol{\theta}_0}$ via the Kullback–Leibler divergence.

- The Wald statistic measures the squared Mahalanobis distance $(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^\top\mathsf{I}(\boldsymbol{\theta})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$.

- Both quantities are strictly positive when $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

Local Alternatives

Definition 13.1 (Local Alternatives)**.** Let $t \in \mathbb{R}^d$. A sequence of local alternatives is defined by

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \frac{\boldsymbol{t}}{\sqrt{n}},$$

and the corresponding distributions $\mathsf{P}_{\boldsymbol{\theta}_n}$.

Theorem 13.3 (Asymptotic Distribution of the MLE under Local Alternatives)**.** *Assume that regularity conditions A0–A4 hold at $\boldsymbol{\theta}_0$. Then, under $\mathsf{P}_{\boldsymbol{\theta}_n}$,*

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_d\big(\boldsymbol{t}, \mathsf{I}(\boldsymbol{\theta}_0)^{-1}\big).$$

Theorem 13.4 (Non-Central Limit of Test Statistics)**.** *Under the local alternatives of Definition 13.1, the likelihood ratio, Wald, and Rao statistics converge in distribution to a non-central chi-square law,*

$$\chi^2_d\big(\boldsymbol{t}^\top \mathsf{I}(\theta_0)\boldsymbol{t}\big)\,, \tag{13.8}$$

with non-centrality parameter $\boldsymbol{t}^\top \mathsf{I}(\theta_0)\boldsymbol{t}$.

Remark (Interpretation)*.* Local alternatives make the testing problem harder as n increases: while the sample size grows, the parameter θ_n approaches the null at rate $1/\sqrt{n}$. The non-centrality parameter in (13.8) quantifies how far the alternative is from the null in the metric induced by the Fisher information.

Note (Power and sample size planning)**.** The non-central chi-square limits in Theorem 13.4 provide a basis for power calculations and sample size planning. Given a model (e.g. logistic regression), one can choose $\boldsymbol{t}$ to represent a scientifically meaningful effect size and compute the resulting power.

Remark (Further reading)*.* A systematic treatment of local alternatives relies on contiguity theory, developed by Lucien Le Cam. A key result is Le Cam's third lemma, which yields joint convergence results for likelihood ratios and test statistics. See Wellner (2018, Chapter 4) and van der Vaart (1998, Chapter 16) for details.

13.3 Composite Null Hypotheses

13.3.1 Test Statistics

Often one is interested in testing hypotheses that concern only a subset of the parameters. Let the parameter vector be partitioned as

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} \in \mathbb{R}^m \times \mathbb{R}^{d-m},$$

where $\boldsymbol{\theta}_1$ is the parameter of primary interest and $\boldsymbol{\theta}_2$ is a nuisance parameter. We consider the composite null hypothesis

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10} \quad \text{vs.} \quad H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_{10}, \tag{13.9}$$

where $\boldsymbol{\theta}_{10}$ is a specified value, while $\boldsymbol{\theta}_2$ remains unrestricted.

Remark. The null hypothesis (13.9) is composite because it contains many parameter values: all vectors $\boldsymbol{\theta} = (\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2)$ with arbitrary $\boldsymbol{\theta}_2$. A classical example is testing whether the mean of a Gaussian distribution equals a fixed value when the variance is unknown.

Likelihood Ratio Statistic

Definition 13.2 (Likelihood Ratio Statistic)**.** Let $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\theta}}_{n1}, \hat{\boldsymbol{\theta}}_{n2})$ denote the unrestricted MLE in $\{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, and let $\hat{\boldsymbol{\theta}}^0_n = (\boldsymbol{\theta}_{10}, \hat{\boldsymbol{\theta}}^0_{n2})$ be the MLE under the constraint $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$. The likelihood ratio statistic is defined as

$$2\log\lambda_n, \qquad \lambda_n := \frac{\sup_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta,\, \boldsymbol{\theta}_1=\boldsymbol{\theta}_{10}} L(\boldsymbol{\theta})} = \frac{L(\hat{\boldsymbol{\theta}}_n)}{L(\hat{\boldsymbol{\theta}}^0_n)}.$$

Remark. The numerator optimizes the likelihood over the full parameter space, while the denominator optimizes only over the null model. Hence $\lambda_n \geqslant 1$ and $2\log\lambda_n \geqslant 0$. Large values provide evidence against H_0.

Rao's Score Statistic

Definition 13.3 (Rao's Score Statistic)**.** Let $\hat{\boldsymbol{\theta}}^0_n$ be the MLE under the null hypothesis. Rao's score statistic for the composite null is

$$R_n = \frac{1}{n}\dot{\ell}(\hat{\boldsymbol{\theta}}^0_n)^\top \widehat{\mathsf{I}}(\hat{\boldsymbol{\theta}}^0_n)^{-1}\dot{\ell}(\hat{\boldsymbol{\theta}}^0_n). \tag{13.10}$$

Remark. The statistic (13.10) measures whether the MLE under the null is close to satisfying the first-order optimality conditions of the unrestricted likelihood. Some score components vanish by construction, while others need not; their joint magnitude is assessed through the quadratic form.

Wald Statistic

Definition 13.4 (Partitioned Fisher Information)**.** Partition the Fisher information matrix as

$$\mathsf{I}(\boldsymbol{\theta}) = \begin{pmatrix} \mathsf{I}_{11}(\boldsymbol{\theta}) & \mathsf{I}_{12}(\boldsymbol{\theta}) \\ \mathsf{I}_{21}(\boldsymbol{\theta}) & \mathsf{I}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

The Schur complement of $\mathsf{I}_{22}(\boldsymbol{\theta})$ is

$$\mathsf{I}_{11.2}(\boldsymbol{\theta}) = \mathsf{I}_{11}(\boldsymbol{\theta}) - \mathsf{I}_{12}(\boldsymbol{\theta})\mathsf{I}_{22}(\boldsymbol{\theta})^{-1}\mathsf{I}_{21}(\boldsymbol{\theta}),$$

and satisfies $(\mathsf{I}(\boldsymbol{\theta})^{-1})_{11} = \mathsf{I}_{11.2}(\boldsymbol{\theta})^{-1}$.

Definition 13.5 (Wald Statistic)**.** Let $\hat{\boldsymbol{\theta}}_{n1}$ be the first m components of the unrestricted MLE $\hat{\boldsymbol{\theta}}_n$. The Wald statistic for testing (13.9) is

$$W_n = n(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{10})^\top \widehat{\mathsf{I}}_{11.2}(\hat{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{10}).$$

Remark. The Wald test compares the unrestricted estimate of $\boldsymbol{\theta}_1$ to its hypothesized value $\boldsymbol{\theta}_{10}$, while appropriately accounting for nuisance parameters through the Schur complement $\widehat{\mathsf{I}}_{11.2}(\hat{\boldsymbol{\theta}}_n)$.

Example 13.5 (Gaussian Mean with Unknown Variance)**.** Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. To test $H_0 : \mu = 7$ with unknown σ^2,

- $\boldsymbol{\theta}_1 = \mu$, $\boldsymbol{\theta}_2 = \sigma^2$,
- the likelihood ratio compares the unrestricted MLE $(\hat{\mu}, \hat{\sigma}^2)$ to the constrained MLE $(7, \hat{\sigma}_0^2)$,
- the Wald statistic assesses whether $\hat{\mu}$ is close to 7, accounting for uncertainty in $\hat{\sigma}^2$.

Note (Practical considerations)**.**

- The likelihood ratio test requires two optimizations: one unrestricted and one under the null.
- Rao's test requires only optimization under the null.
- The Wald test uses the unrestricted MLE and an estimate of the partitioned Fisher information.

13.3.2 Chi-Square Limits for Composite Null Hypotheses

As before, assume that the estimator $\tilde{\boldsymbol{\theta}}_n$ is a consistent solution of the likelihood equations for the model $\{\mathsf{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$.

Similarly, assume that $\tilde{\boldsymbol{\theta}}_n^0$ is a consistent solution of the likelihood equations for the null model with $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_{10}$.

We write $\tilde{\lambda}_n$, $\widetilde{W}_n$, and $\widetilde{R}_n$ to indicate the use of $\tilde{\boldsymbol{\theta}}_n$ and $\tilde{\boldsymbol{\theta}}_n^0$ instead of $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n^0$.

Theorem 13.5 (Null distribution / Wilks' Theorem)**.** *Suppose assumptions A0–A4 hold at the true parameter $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{02})^\top \in \mathbb{R}^m \times \mathbb{R}^{d-m}$, and assume that $\boldsymbol{\theta}_{01} = \boldsymbol{\theta}_{10}$, so that H_0 is true. Then*

$$\left.\begin{array}{l} 2\log\tilde{\lambda}_n \\ \widetilde{W}_n \\ \widetilde{R}_n \end{array}\right\} \xrightarrow{d} \chi_m^2.$$

Remark. The degrees of freedom m equal the number of imposed constraints, or equivalently the difference between the dimension of the alternative model (d) and the dimension of the null model ($d - m$).

Proof. **Wald Statistic**

Define

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \end{pmatrix} := \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{Z}, \qquad \boldsymbol{Z} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0)).$$

Then $\boldsymbol{D} \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0)^{-1})$ and

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{10}) \xrightarrow{d} \boldsymbol{D}_1 \sim \mathcal{N}_m(\mathbf{0}, (\mathbf{I}(\boldsymbol{\theta}_0)^{-1})_{11}) = \mathcal{N}_m(\mathbf{0}, \mathbf{I}_{11.2}(\boldsymbol{\theta}_0)^{-1}).$$

Since $\hat{\mathbf{I}}(\tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{I}(\boldsymbol{\theta}_0)$, we also have $\hat{\mathbf{I}}_{11.2}(\tilde{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{I}_{11.2}(\boldsymbol{\theta}_0)$. Therefore,

$$\widetilde{W}_n \xrightarrow{d} \boldsymbol{D}_1^\top \mathbf{I}_{11.2}(\boldsymbol{\theta}_0)\boldsymbol{D}_1 \sim \chi^2_m.$$

Note. The Schur complement $\mathbf{I}_{11.2}$ represents the Fisher information about $\boldsymbol{\theta}_1$ in the presence of the nuisance parameter $\boldsymbol{\theta}_2$. It also corresponds to the inverse asymptotic covariance matrix of $\sqrt{n}(\tilde{\boldsymbol{\theta}}_{n1} - \boldsymbol{\theta}_{10})$.

Rao (Score) Statistic

Partition the score function as

$$\dot{\ell}(\tilde{\boldsymbol{\theta}}^0_n) = \begin{pmatrix} \dot{\ell}_1(\tilde{\boldsymbol{\theta}}^0_n) \\ \dot{\ell}_2(\tilde{\boldsymbol{\theta}}^0_n) \end{pmatrix}.$$

Since $\tilde{\boldsymbol{\theta}}^0_n$ maximizes the likelihood under H_0, we have $\dot{\ell}_2(\tilde{\boldsymbol{\theta}}^0_n) = \mathbf{0}$.
A Taylor expansion around $\boldsymbol{\theta}_0$ yields

$$\frac{1}{\sqrt{n}}\dot{\ell}_1(\tilde{\boldsymbol{\theta}}^0_n) = \frac{1}{\sqrt{n}}\dot{\ell}_1(\boldsymbol{\theta}_0) - \mathbf{I}_{12}(\boldsymbol{\theta}_0)\sqrt{n}(\tilde{\boldsymbol{\theta}}^0_{n2} - \boldsymbol{\theta}_{02}) + o_p(1).$$

Using asymptotic linearity of the MLE under the null model,

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}^0_{n2} - \boldsymbol{\theta}_{02}) = \mathbf{I}_{22}(\boldsymbol{\theta}_0)^{-1}\frac{1}{\sqrt{n}}\dot{\ell}_2(\boldsymbol{\theta}_0) + o_p(1).$$

Hence,

$$\frac{1}{\sqrt{n}}\dot{\ell}_1(\tilde{\boldsymbol{\theta}}^0_n) \xrightarrow{d} \boldsymbol{Z}_1 - \mathbf{I}_{12}(\boldsymbol{\theta}_0)\mathbf{I}_{22}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{Z}_2,$$

where $(\boldsymbol{Z}_1^\top, \boldsymbol{Z}_2^\top)^\top \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta}_0))$.
Now, we have that

$$\boldsymbol{Z}_1 - \mathbf{I}_{12}(\boldsymbol{\theta}_0)\mathbf{I}_{22}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{Z}_2 \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I}_{11.2}(\boldsymbol{\theta}_0)).$$

Therefore,

$$\widetilde{R}_n = \frac{1}{n}\dot{\ell}_1(\tilde{\boldsymbol{\theta}}^0_n)^\top \hat{\mathbf{I}}_{11.2}(\tilde{\boldsymbol{\theta}}^0_n)^{-1}\dot{\ell}_1(\tilde{\boldsymbol{\theta}}^0_n) \xrightarrow{d} \chi^2_m.$$

Remark. The Rao statistic measures whether the MLE under the null is already close to satisfying the full first-order optimality conditions of the unrestricted model.

Likelihood Ratio Statistic

Write

$$2\log\tilde{\lambda}_n = 2\big(\ell(\tilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\big) - 2\big(\ell(\tilde{\boldsymbol{\theta}}_n^0) - \ell(\boldsymbol{\theta}_0)\big).$$

By standard likelihood theory,

$$\begin{aligned} 2\big(\ell(\tilde{\boldsymbol{\theta}}_n) - \ell(\boldsymbol{\theta}_0)\big) &\xrightarrow{d} \boldsymbol{Z}^\top \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{Z}, \\ 2\big(\ell(\tilde{\boldsymbol{\theta}}_n^0) - \ell(\boldsymbol{\theta}_0)\big) &\xrightarrow{d} \boldsymbol{Z}_2^\top \mathbf{I}_{22}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{Z}_2, \end{aligned}$$

jointly. Hence,

$$2\log\tilde{\lambda}_n \xrightarrow{d} \boldsymbol{D}_1^\top \mathbf{I}_{11.2}(\boldsymbol{\theta}_0)\boldsymbol{D}_1 \sim \chi^2_m.$$

□

Remark. All three statistics are asymptotically equivalent. Their differences are of order $o_p(1)$ and vanish asymptotically.

Example 13.6 (Simulation Illustration)**.** For a Gamma$(5,2)$ model with sample size $n = 150$, simulate 10,000 datasets and compute $2\log\tilde{\lambda}_n$ under a composite null hypothesis. The empirical distribution function closely matches the χ^2_m distribution, and the corresponding p-value distribution is approximately uniform, confirming correct calibration.

Example 13.7 (Likelihood ratio test for Gamma model)**.** We compute the p-value for testing

$$H_0 : \alpha = 4$$

based on a sample from the Gamma$(5,2)$ distribution.

Under H_0, we optimize the negative log-likelihood with respect to β while fixing $\alpha = \alpha_0$.

Definition 13.6 (Negative log-likelihood under the null)**.** Let $\alpha_0 > 0$ be fixed. The negative log-likelihood under H_0 is given by

$$\ell_0(\beta; \boldsymbol{x}, \alpha_0) = -\sum_{i=1}^{n} \log f_{\alpha_0,\beta}(x_i),$$

where $f_{\alpha,\beta}$ denotes the Gamma density with shape α and rate β.

```r
negloglik0 = function(b, data, a0){
    a = a0
    n = length(data)
    return(
        -sum(dgamma(data, shape = a, rate = b, log = TRUE))
    )
}
```

Data generation We generate a sample of size $n = 150$ from a Gamma$(5,2)$ distribution.

```
set.seed(22)
n <- 150
x <- rgamma(n, shape = 5, rate = 2)
```

Optimization under null and alternative We optimize the negative log-likelihood under H_0 and under the unrestricted model.

```
opt0 <- optimize(negloglik0, c(0, 100), data = x, a0 = 4)
opt1 <- optim(c(1,1), negloglik, data = x)
```

Definition 13.7 (Likelihood ratio statistic)**.** The likelihood ratio statistic is defined as

$$\Lambda_n = 2\big(\ell(\hat{\alpha}, \hat{\beta}) - \ell(\alpha_0, \hat{\beta}_0)\big),$$

where $(\hat{\alpha}, \hat{\beta})$ denotes the unrestricted MLE and $\hat{\beta}_0$ the MLE under H_0.

```
lambda <- 2*(-opt1$value - (-opt0$objective))
lambda
```

This yields $\Lambda_n \approx 6.99$.

p–value computation

Under regularity conditions, $\Lambda_n \xrightarrow{d} \chi^2_1$ since one restriction is imposed.

```
p.value <- 1 - pchisq(lambda, df = 1)
p.value
```

The resulting p-value is approximately 0.008, indicating strong evidence against H_0.

Remark. Intuitively, the likelihood ratio compares how much more likely the observed data are under the best-fitting Gamma distribution compared to the Gamma distribution with fixed shape $\alpha = 4$. Here, the unrestricted model is about seven times more likely in terms of density.

Example 13.8 (Null distribution of LR statistic and p–value)**.**

Simulation under the null We now investigate the null distribution of the likelihood ratio statistic and the corresponding p-values for testing

$$H_0 : \alpha = 5$$

when the data are generated from a Gamma$(5, 2)$ distribution.

Definition 13.8 (Composite likelihood ratio test function)**.** The following function computes the likelihood ratio statistic and p-value, including optimization under the restricted model.

```r
lr_test_gamma_comp <- function(
    alpha = 1, beta = 1, n = 100, alpha0 = 1){
    x <- rgamma(n = n, shape = alpha, rate = beta)
    opt0 <- optimize(negloglik0, c(0,100),
                     data = x, a0 = alpha0)
    opt1 <- optim(c(1,1), negloglik, data = x)
    stat <- 2*(-opt1$value - (-opt0$objective))
    pval <- 1 - pchisq(stat, df = 1)
    return(data.frame(stat = stat, pval = pval))
}
```

Monte Carlo approximation We repeat the test 10,000 times with sample size $n = 150$.

```r
set.seed(22)

df <- 1:10000 |>
  purrr::map_df(\(i)
    lr_test_gamma_comp(
      alpha = 5, beta = 2, n = 150, alpha0 = 5
    )
  )
```

Remark. For each simulated dataset, we compute $2 \log \Lambda_n$ and the corresponding p-value. This yields an empirical distribution of the test statistic under the null.

Empirical vs asymptotic distribution The empirical distribution function of the simulated statistics is compared to the χ^2_1 distribution function.

```r
p1 <- ggplot(df, aes(x = stat)) +
    stat_ecdf(geom = "step") +
    geom_function(fun = pchisq, args = list(df = 1))
```

Remark. The empirical distribution function is a step function with many small jumps, but it closely approximates the χ^2_1 distribution. This confirms the validity of the asymptotic approximation.

Claim (Calibration of the test)**.** If the likelihood ratio test is well calibrated, the distribution of p-values under H_0 should be uniform on $[0, 1]$.

Remark. In this simulation, the p-value histogram is nearly uniform, indicating excellent calibration for sample size $n = 150$ in the Gamma example.

Note. Throughout, the nuisance parameter β is optimized out under both the null and alternative. The test focuses solely on the parameter of interest α.

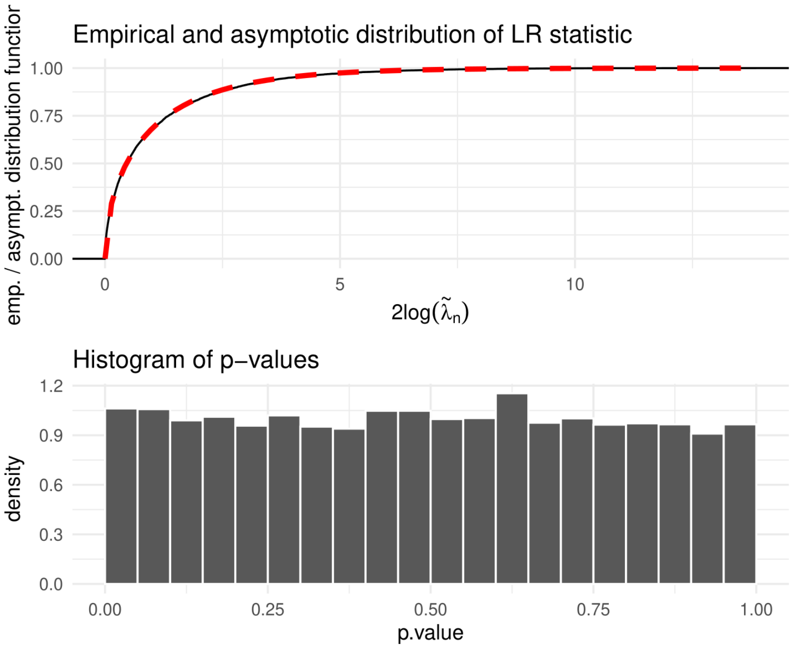

Empirical and asymptotic distribution of LR statistic
emp. / asympt. distribution functior
1.00
0.75
0.50
0.25
0.00
0
5
10
2log($\tilde{\lambda}_n$)
Histogram of p−values
density
1.2
0.9
0.6
0.3
0.0
0.00
0.25
0.50
0.75
1.00
p.value

Figure 13.2

14. Consistency of the MLE – Wald's Theorem

Setup

- **Model:** $\mathcal{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}, \qquad \Theta \subseteq \mathbb{R}^d.$
- **Data:** $X_1, \dots, X_n$ i.i.d. from P_{θ_0}, where $\theta_0 \in \Theta$ denotes the **true parameter** and the sample space is $\mathcal{X}$.
- **Maximum likelihood estimator (MLE):**

$$\hat{\theta}_n = \arg\max_{\theta\in\Theta} \ell_n(\theta) = \arg\max_{\theta\in\Theta} \frac{1}{n}\big[\ell_n(\theta) - \ell_n(\theta_0)\big].$$

Asymptotic log-likelihood and Kullback–Leibler divergence

For every $\theta \in \Theta$,

$$\begin{aligned}\frac{1}{n}\big[\ell_n(\theta) - \ell_n(\theta_0)\big] &= \frac{1}{n}\sum_{i=1}^n \log\frac{p_\theta(X_i)}{p_{\theta_0}(X_i)} \\ &\xrightarrow{\text{a.s.}} \mathsf{E}_{\theta_0}\left[\log\frac{p_\theta(X_1)}{p_{\theta_0}(X_1)}\right] = -\operatorname{KL}(\mathsf{P}_{\theta_0} \mid \mathsf{P}_\theta).\end{aligned}$$

Moreover,

$$\theta_0 = \arg\max_{\theta\in\Theta}\big\{-\operatorname{KL}(\mathsf{P}_{\theta_0} \mid \mathsf{P}_\theta)\big\},$$

and the maximizer is unique under identifiability assumptions (A0).

Question. Under which conditions does $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta_0$ hold?

Why pointwise convergence is not enough

Example 14.1 (Deterministic counterexample)**.** Let $\theta \in [0,1]$ and define

$$f_n(\theta) = \sqrt{2en}\,\theta(1-\theta^2)^n - \Big(\theta - \tfrac{1}{2}\Big)^2.$$

Then

$$f_n(\theta) \longrightarrow f(\theta) = -\Big(\theta - \tfrac{1}{2}\Big)^2 \quad \text{pointwise as } n \to \infty.$$

However,

$$\arg\max_{\theta\in[0,1]} f_n(\theta) \longrightarrow 0 \neq \tfrac{1}{2} = \arg\max_{\theta\in[0,1]} f(\theta).$$

as $n \to \infty$. E.g., note that

$$\lim_{n\to\infty} f_n\left(\frac{1}{\sqrt{2n}}\right) = \frac{3}{4} > 0.$$

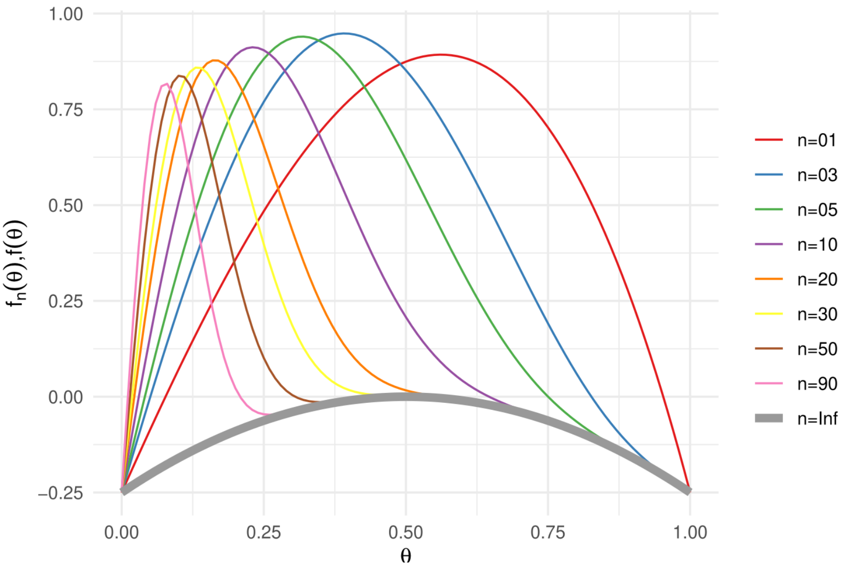


Figure 14.1

Remark. Pointwise convergence of objective functions does *not* guarantee convergence of maximizers. Some form of **uniform control** is required to prevent the maximizer from "escaping".

14.1 Achieving the Maximum

Definition 14.1 (Upper semi-continuity)**.** A function $g : \Theta \to \mathbb{R}$ is called **upper semi-continuous** (u.s.c.) at $\theta_0 \in \Theta$ if

$$\theta_n \to \theta_0 \quad \Longrightarrow \quad \limsup_{n\to\infty} g(\theta_n) \leq g(\theta_0).$$

The function g is called upper semi-continuous if it is u.s.c. at every point of its domain.

Lemma 14.1 (Existence of a maximizer on compact sets)**.** *If $g : \Theta \to \mathbb{R}$ is upper semi-continuous, then g achieves its maximum on any compact subset $\Theta_0 \subseteq \Theta$.*

Proof. Let

$$g^* = \sup_{\theta\in\Theta_0} g(\theta).$$

There exists a sequence $(\theta_n)_{n\geq 1} \subset \Theta_0$ such that $g(\theta_n) \to g^*$. Since Θ_0 is compact, there exists a convergent subsequence $(\theta_{n_k})_{k\geq 1}$ with limit $\theta^* \in \Theta_0$. By upper semicontinuity,

$$g^* = \lim_{k\to\infty} g(\theta_{n_k}) \leq g(\theta^*) \leq \sup_{\theta\in\Theta_0} g(\theta) = g^*.$$

Hence, $g(\theta^*) = g^*$ and the maximum is achieved. □

Note. In the context of maximum likelihood estimation, the function g will typically be the normalized log-likelihood

$$g_n(\theta) = \frac{1}{n}\ell_n(\theta),$$

or a log-likelihood ratio. Compactness of the parameter space Θ and upper semicontinuity of g_n ensure the existence of maximizers for each sample size n.

Remark. Pointwise convergence of $g_n(\theta)$ to a limiting function $g(\theta)$ does *not* in general imply convergence of the maximizers. Additional uniformity conditions are required to prevent the maximizers from "escaping".

Failure of Pointwise Convergence

Example 14.2 (Deterministic counterexample)**.** Consider the sequence of functions $f_n : [0,1] \to \mathbb{R}$ defined by

$$f_n(\theta) = \sqrt{2en}\,\theta(1-\theta^2)^n - \left(\theta - \tfrac{1}{2}\right)^2.$$

Then $f_n(\theta) \to f(\theta) = -(\theta - \frac{1}{2})^2$ pointwise as $n \to \infty$. The limit function has a unique maximizer at $\theta = \frac{1}{2}$. However,

$$\arg\max_{\theta\in[0,1]} f_n(\theta) \longrightarrow 0 \quad \text{as } n \to \infty,$$

and thus the maximizers do not converge to the maximizer of the limiting function.

Remark. This example illustrates the need for uniform control over the log-likelihood surface. In maximum likelihood theory, this role is played by assumptions ensuring that likelihoods cannot take excessively large values away from the true parameter.

14.2 One-Sided Version of a Uniform Law of Large Numbers

We apply the following result to the special case

$$f(x,\theta) := \log p_\theta(x) - \log p_{\theta_0}(x).$$

Theorem 14.1 (One-sided uniform law of large numbers)**.** *Suppose that:*

a) *Θ is compact;*

b) *for all $x \in \mathcal{X}$, the map*

$$\theta \mapsto f(x,\theta) := \log p_\theta(x) - \log p_{\theta_0}(x)$$

is upper semi-continuous;

c) *there exists a measurable function $F : \mathcal{X} \to \mathbb{R}$ such that*

$$f(x,\theta) \le F(x) \quad \forall x \in \mathcal{X},\ \forall \theta \in \Theta, \qquad \mathsf{E}_{\theta_0}[F(X)] < \infty;$$

d) *for all $\theta \in \Theta$ and all sufficiently small $\rho > 0$, the function*

$$\psi(x,\theta,\rho) := \sup_{\|\theta'-\theta\|<\rho} f(x,\theta')$$

is measurable in x.

Then

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n f(X_i,\theta) \overset{a.s.}{\le} \sup_{\theta\in\Theta} \mathsf{E}_{\theta_0}[f(X_1,\theta)].$$

Remark. In maximum likelihood applications, f is the log density ratio comparing p_θ to the true density p_{θ_0}. The theorem ensures that, even after taking a supremum over a continuum of parameter values, the empirical log-likelihood cannot asymptotically exceed the supremum of its population counterpart.

Proof. The argument follows the spirit of the Glivenko–Cantelli theorem:

i) Control expectations of suprema over sufficiently small neighborhoods;

ii) Cover the compact parameter space by finitely many such neighborhoods.

In detail,

i. Fix $\theta \in \Theta$. As $\rho \searrow 0$, upper semi-continuity implies

$$\psi(x,\theta,\rho) \searrow f(x,\theta) \qquad \text{for all } x \in \mathcal{X}.$$

Indeed, the supremum is attained on the closed ball, and the values of the maximizing sequence converge to the value at the center by upper semi-continuity. Consequently,

$$0 \le F(x) - \psi(x,\theta,\rho) \nearrow F(x) - f(x,\theta),$$

and by the monotone convergence theorem,

$$\mathsf{E}_{\theta_0}[F(X_1) - \psi(X_1,\theta,\rho)] \longrightarrow \mathsf{E}_{\theta_0}[F(X_1) - f(X_1,\theta)] \quad \text{as } \rho \to 0.$$

Since $\mathsf{E}_{\theta_0}[F(X_1)] < \infty$, this yields

$$\mathsf{E}_{\theta_0}[\psi(X_1,\theta,\rho)] \longrightarrow \mathsf{E}_{\theta_0}[f(X_1,\theta)] =: g(\theta).$$

Hence, for all $\varepsilon > 0$ and all $\theta \in \Theta$, there exists $\rho_\theta > 0$ such that

$$\mathsf{E}_{\theta_0}[\psi(X_1,\theta,\rho_\theta)] < g(\theta) + \varepsilon.$$

ii. Define the open balls

$$S(\theta, \rho) := \{\theta' \in \Theta : \|\theta' - \theta\| < \rho\}.$$

Then

$$\Theta \subseteq \bigcup_{\theta \in \Theta} S(\theta, \rho_\theta).$$

By compactness, there exist $\theta_1, \ldots, \theta_m \in \Theta$ such that

$$\Theta \subseteq \bigcup_{j=1}^{m} S(\theta_j, \rho_{\theta_j}).$$

Therefore,

$$\begin{aligned} \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) &\leq \sup_{1 \leq j \leq m} \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \theta_j, \rho_{\theta_j}) \\ &\xrightarrow{a.s.} \sup_{1 \leq j \leq m} \mathsf{E}_{\theta_0}[\psi(X_1, \theta_j, \rho_{\theta_j})] \\ &\leq \sup_{1 \leq j \leq m} g(\theta_j) + \varepsilon \leq \sup_{\theta \in \Theta} g(\theta) + \varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, the claim follows.

□

Lemma 14.2. *Under the assumptions of Theorem 14.1, the function*

$$g(\theta) = \mathsf{E}_{\theta_0}[f(X_1, \theta)]$$

is upper semi-continuous on Θ.

Proof. Let $\theta_n \to \theta$. By Fatou–Lebesgue,

$$\begin{aligned} \limsup_{n \to \infty} g(\theta_n) &= \limsup_{n \to \infty} \mathsf{E}_{\theta_0}[f(X_1, \theta_n)] \\ &\leq \mathsf{E}_{\theta_0}\left[\limsup_{n \to \infty} f(X_1, \theta_n)\right] \leq \mathsf{E}_{\theta_0}[f(X_1, \theta)] = g(\theta), \end{aligned}$$

where the last inequality uses upper semi-continuity of $f(x, \cdot)$ for every x. □

Application to Maximum Likelihood

Remark. In the likelihood setting,

$$g(\theta) = \mathsf{E}_{\theta_0}\left[\log \frac{p_\theta(X_1)}{p_{\theta_0}(X_1)}\right] = -\operatorname{KL}(\mathsf{P}_{\theta_0} \,||\, \mathsf{P}_\theta).$$

Thus, upper semi-continuity of the log density ratio implies upper semi-continuity of the negative Kullback–Leibler divergence.

Note. The combination of compactness, upper semi-continuity, and the one-sided uniform law of large numbers forms the technical backbone of Wald's consistency theorem for maximum likelihood estimators.

14.3 Wald's Theorem

We study conditions under which maximum likelihood estimators (MLEs) converge to the true parameter. The key technical ingredient is a uniform (one-sided) law of large numbers that prevents the likelihood from "escaping" to large values away from the truth.

Definition 14.2 (Log density ratio)**.** Let $\{p_\theta : \theta \in \Theta\}$ be a parametric family of densities and let θ_0 denote the true parameter. Define

$$f(x, \theta) := \log p_\theta(x) - \log p_{\theta_0}(x).$$

Definition 14.3 (Identifiability)**.** The model is said to be identifiable if

$$\mathsf{P}_\theta = \mathsf{P}_{\theta_0} \implies \theta = \theta_0.$$

Theorem 14.2 (Wald's Theorem 1949)**.** *Suppose that:*

(a) Θ *is compact;*

(b) $\forall x \in \mathcal{X}$, *the map* $\theta \mapsto p_\theta(x)$ *is upper semi-continuous;*

(c) there exists a measurable function $F : \mathcal{X} \to \mathbb{R}$ *such that*

$$f(x, \theta) \leq F(x) \quad \forall x \in \mathcal{X},\ \forall \theta \in \Theta, \qquad \mathsf{E}_{\theta_0}[F(X_1)] < \infty;$$

(d) for all $\theta \in \Theta$ *and all sufficiently small* $\rho > 0$, *the function*

$$x \mapsto \sup_{\|\theta' - \theta\| < \rho} f(x, \theta')$$

is measurable;

(e) the identifiability condition of Definition 14.3 holds.

Then any sequence of (possibly non-measurable) maximum likelihood estimators $\hat{\theta}_n$ *satisfies*

$$\hat{\theta}_n \xrightarrow{a.s.} \theta_0.$$

Remark. The estimator $\hat{\theta}_n$ need not be measurable. The almost sure convergence statement is interpreted pathwise, that is, as a statement about data sequences outside a null set. Conditions ensuring measurability can be found in Ferguson (1996).

Heuristic discussion

Fix $\theta \in \Theta$. By the law of large numbers,

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \xrightarrow{a.s.} \mathsf{E}_{\theta_0}[f(X_1, \theta)] = -\operatorname{KL}(\mathsf{P}_{\theta_0} \,||\, \mathsf{P}_\theta).$$

The limiting criterion is uniquely maximized at θ_0. However, pointwise convergence is not sufficient to ensure convergence of maximizers: without uniform control, maxima may "run away". Compactness and integrable domination provide the needed uniformity.

Uniform control via small balls

For $\theta \in \Theta$ and $\rho > 0$, define

$$\psi(x, \theta, \rho) := \sup_{\|\theta' - \theta\| < \rho} f(x, \theta').$$

Claim. If $f(x, \cdot)$ is upper semi-continuous, then for every x and θ,

$$\psi(x, \theta, \rho) \searrow f(x, \theta) \quad \text{as } \rho \searrow 0.$$

Proof. The supremum is at least $f(x, \theta)$ since θ belongs to every ball. As $\rho \to 0$, any maximizing sequence must converge to θ, and upper semi-continuity implies that the limit cannot exceed $f(x, \theta)$. □

Claim. For fixed θ,

$$\mathsf{E}_{\theta_0}[\psi(X_1, \theta, \rho)] \longrightarrow \mathsf{E}_{\theta_0}[f(X_1, \theta)] =: g(\theta) \quad \text{as } \rho \to 0.$$

Proof. Since $f(x, \theta) \leq F(x)$ for all θ, we have

$$0 \leq F(x) - \psi(x, \theta, \rho) \nearrow F(x) - f(x, \theta).$$

By the monotone convergence theorem and $\mathsf{E}_{\theta_0}[F(X_1)] < \infty$,

$$\mathsf{E}_{\theta_0}[F(X_1) - \psi(X_1, \theta, \rho)] \to \mathsf{E}_{\theta_0}[F(X_1) - f(X_1, \theta)],$$

which implies the claim. □

Note. For every $\varepsilon > 0$ and $\theta \in \Theta$, there exists $\rho_\theta > 0$ such that

$$\mathsf{E}_{\theta_0}[\psi(X_1, \theta, \rho_\theta)] \leq g(\theta) + \varepsilon.$$

Finite covering argument

By compactness,

$$\Theta \subseteq \bigcup_{j=1}^{m} \{\theta' : \|\theta' - \theta_j\| < \rho_{\theta_j}\}.$$

Hence,

$$\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \leq \sup_{1 \leq j \leq m} \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \theta_j, \rho_{\theta_j}).$$

Each term converges almost surely by the law of large numbers, and the supremum over finitely many indices preserves almost sure convergence.

Proof (of Theorem 14.2). Fix $r > 0$ and define

$$S_r := \{\theta \in \Theta : \|\theta - \theta_0\| \geq r\}.$$

Then S_r is compact.

Claim.

$$\mathsf{P}_{\theta_0}\big(\exists N\ \forall n \geq N:\ \hat{\theta}_n \notin S_r\big) = 1.$$

Proof. Applying the uniform law of large numbers to S_r yields

$$\limsup_{n\to\infty} \sup_{\theta \in S_r} \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \leq \sup_{\theta \in S_r} g(\theta) =: \delta \quad \text{a.s.}$$

By Lemma 14.2, the supremum is attained. Since $\theta_0 \notin S_r$ and $g(\theta) < 0$ for $\theta \neq \theta_0$, we have $\delta < 0$. Hence, eventually

$$\sup_{\theta \in S_r} \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \leq \frac{\delta}{2}.$$

On the other hand,

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i, \hat{\theta}_n) = \sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \geq 0,$$

which implies $\hat{\theta}_n \notin S_r$ eventually. □

Define

$$A_r := \{\exists N\ \forall n \geq N:\ \hat{\theta}_n \notin S_r\}.$$

The events A_r are decreasing as $r \searrow 0$, and $\mathsf{P}_{\theta_0}(A_r) = 1$ for all $r > 0$. Therefore,

$$\mathsf{P}_{\theta_0}\Big(\bigcap_{r>0} A_r\Big) = 1,$$

which is exactly the statement $\hat{\theta}_n \to \theta_0$ almost surely. □

Note. Compactness, upper semi-continuity, and an integrable envelope are indispensable. Without them, global maximizers of the likelihood may fail to be consistent.

14.4 Examples for Wald's Theorem

Example 14.3 (Cauchy location model)**.** Consider the Cauchy location model with density

$$p_\vartheta(x) = \frac{1}{\pi} \frac{1}{1 + (x - \vartheta)^2}, \qquad \vartheta \in \mathbb{R}.$$

Since the parameter space $\mathbb{R}$ is not compact, Wald's theorem cannot be applied directly. We therefore reparametrize by

$$\vartheta = \tan(\theta/2), \qquad \theta \in (-\pi, \pi),$$

which yields the equivalent family

$$p_\theta(x) = \frac{1}{\pi} \frac{1}{1 + \big(x - \tan(\theta/2)\big)^2}, \qquad \theta \in (-\pi, \pi).$$

To compactify the parameter space, define artificial boundary values

$$p_{\pm\pi}(x) := \lim_{\theta\to\pm\pi} p_\theta(x) = \lim_{\vartheta\to\pm\infty} p_\vartheta(x) = 0, \qquad \forall x \in \mathbb{R}.$$

This yields a compact parameter space

$$\Theta = [-\pi, \pi], \qquad \mathsf{P} = \{\mathsf{P}_\theta : \theta \in \Theta\}.$$

Without loss of generality, we may take $\theta_0 = 0$ (otherwise shift the data by $x_i' = x_i - \vartheta_0$).

Verification of Wald's conditions

a) **Compactness.** By construction, $\Theta = [-\pi, \pi]$ is compact.

b) **Identifiability.** The median of P_θ equals $\tan(\theta/2)$, hence

$$\theta \neq \theta_0 \implies \mathsf{P}_\theta \neq \mathsf{P}_{\theta_0}, \qquad \theta_0 \in (-\pi, \pi).$$

c) **Upper semi-continuity.** For every x, the map $\theta \mapsto p_\theta(x)$ is continuous on $(-\pi, \pi)$, and the boundary values were chosen to preserve continuity. Hence $\theta \mapsto p_\theta(x)$ is upper semi-continuous on Θ.

d) **Measurability.** Continuity implies measurability. Indeed, for any $A \subseteq \Theta$ and $t \in \mathbb{R}$,

$$\{x : \sup_{\theta\in A} p_\theta(x) \le t\} = \bigcap_{\theta\in A\cap\mathbb{Q}} \{x : p_\theta(x) \le t\},$$

a countable intersection of measurable sets.

e) **Integrable envelope.** The log-likelihood ratio equals

$$f(x,\theta) = \log p_\theta(x) - \log p_{\theta_0}(x) = \begin{cases} \log\left(\dfrac{1+x^2}{1+\big(x-\tan(\theta/2)\big)^2}\right), & \theta \in (-\pi,\pi), \\ -\infty, & \theta = \pm\pi. \end{cases}$$

Since $1 + (x - \tan(\theta/2))^2 \ge 1$ for all θ, we have

$$f(x,\theta) \le \log(1+x^2) =: F(x), \qquad \forall x, \theta.$$

Moreover,

$$\mathsf{E}_{\theta_0}[F(X_1)] = \int_{\mathbb{R}} \log(1+x^2)\,\frac{1}{\pi}\frac{1}{1+x^2}\,\mathrm{d}x = 2\log 2 < \infty.$$

All assumptions of Wald's theorem are satisfied. Hence the global maximum likelihood estimator in the Cauchy location model is **consistent**, despite the likelihood exhibiting many local optima.

Example 14.4 (Normal distribution with unknown mean and variance)**.** Consider the normal model

$$p_{\boldsymbol{\theta}}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}, \qquad \boldsymbol{\theta} = (\mu, \sigma^2)^\top \in \mathbb{R} \times (0, \infty).$$

For i.i.d. data $X_1, \ldots, X_n \sim \mathsf{P}_{\boldsymbol{\theta}_0}$, the MLE is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2.$$

By the strong law of large numbers,

$$\hat{\boldsymbol{\theta}}_n = (\hat{\mu}, \hat{\sigma}^2)^\top \xrightarrow{a.s.} (\mu_0, \sigma_0^2)^\top.$$

Why Wald's theorem fails

Fix $x \in \mathbb{R}$. Maximizing the likelihood over θ yields

$$\sup_{\mu,\sigma^2} p_\theta(x) = \infty,$$

since setting $\mu = x$ and letting $\sigma^2 \to 0$ makes the density diverge. Consequently, the log-likelihood ratio admits no integrable upper bound, and Wald's theorem does not apply.

Solution via grouping

Group observations into pairs

$$\tilde{X}_1 = (X_1, X_2),\ \tilde{X}_2 = (X_3, X_4),\ \tilde{X}_3 = (X_5, X_6), \ldots$$

and consider the joint density of (X_1, X_2). One can show that

$$\max_\theta p_\theta(x_1, x_2) = p_{(\tilde{\mu}, \tilde{\sigma}^2)}(x_1, x_2) = \frac{2}{e\pi(x_1 - x_2)^2}, \tag{14.1}$$

where

$$\tilde{\mu} = \frac{x_1 + x_2}{2}, \qquad \tilde{\sigma}^2 = \frac{1}{4}(x_1 - x_2)^2.$$

Since $\mathsf{P}_{\theta_0}(X_1 \neq X_2) = 1$, the bound in (14.1) is integrable. After compactifying the parameter space as in Example 14.3, Wald's theorem applies, yielding consistency of the MLE; see van der Vaart (1998, Problem 5.25).

Remark. These examples illustrate that Wald's conditions are sufficient but not necessary. The normal MLE is consistent even when Wald's theorem fails in its basic form, while suitable reformulations may restore applicability.