Chapter 8: Estimation and Model Fitting (Rice 8.1–8.5)

Donghyun Ko

May 27, 2026

This chapter formalizes *parametric statistical modeling*: assume data are generated i.i.d. from a distribution $f(x;\theta)$ with unknown parameter(s) θ. We study how to (i) propose a probability model, (ii) estimate θ (method of moments, maximum likelihood), and (iii) evaluate estimator quality via bias/variance/MSE and standard errors, including large-sample approximations used for inference.

Contents

1 Probability models and data

In survey sampling, randomness arises from the *sampling design* applied to a finite population. Here, randomness comes from an assumed *stochastic data-generating mechanism*:

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x; \theta),$$

where θ is a fixed but unknown parameter (or vector).

A statistical model is not the truth itself, but a **deliberate and controlled approximation** of reality. Once a distribution $f(x; \theta)$ is specified, **all inference about unobserved or future data is carried out through the unknown parameter** θ. From this perspective, problems of statistical estimation naturally decompose into two components: i) assessing whether the chosen probabilistic model provides an adequate description of the data (model choice and model adequacy), and ii) estimating the parameter θ within that model using the observed data.

1.1 Parameters, statistics, and estimators

In parametric statistical modeling, it is essential to distinguish clearly between *parameters*, *statistics*, and *estimators*. These concepts play different roles in inference and uncertainty quantification.

A **parameter** θ is a fixed but unknown constant that characterizes the probability distribution governing the data. It is a property of the underlying population or data-generating mechanism, not of the observed sample. For example, in a Poisson model, the rate λ determines the entire distribution of counts.

A **statistic** is any function of the observed data only. Formally, if $X_1, \ldots, X_n$ denote the random variables representing the data, then a statistic has the form

$$T = T(X_1, \ldots, X_n),$$

and does *not* depend on unknown parameters. Before data are observed, a statistic is a random variable; after observation, it takes a numerical value.

An **estimator** $\hat{\theta}$ is a statistic that is specifically chosen to approximate an unknown parameter θ. Thus, every estimator is a statistic, but not every statistic is used as an estimator.

Example

Poisson counting model (radioactive emissions): We are interested in modeling the number of α-particle emissions from a radioactive source over a fixed time interval. Suppose emissions are counted over many disjoint 10-second intervals.

Model assumption. Let X denote the number of emissions observed in a single 10-second interval. A common and scientifically justified model is

$$X \sim \text{Poisson}(\lambda),$$

with probability mass function

$$\mathbb{P}(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

Here, $\lambda > 0$ is the *rate parameter*, representing the average number of emissions per 10 seconds.

Interpretation of the parameter. The parameter λ is a *population parameter*:

$$\mathbb{E}[X] = \lambda, \qquad \mathrm{Var}(X) = \lambda.$$

If λ were known, the distribution of emissions would be completely determined.

Observed data. In the experiment, emissions were counted in $n = 1207$ independent 10-second intervals. Rather than listing all individual counts, the data are summarized as frequencies: for each integer k, the number of intervals in which k emissions occurred was recorded (e.g., how many intervals had 7 emissions, 8 emissions, etc.).

Statistics from the data. Let $X_1, \ldots, X_{1207}$ denote the emission counts from the 1207 intervals. Examples of statistics include:

- the sample mean
$$\bar{X} = \frac{1}{1207} \sum_{i=1}^{1207} X_i,$$
- the sample variance
$$\frac{1}{1206} \sum_{i=1}^{1207} (X_i - \bar{X})^2,$$
- the maximum observed count,
- the proportion of intervals with zero emissions.

Each of these depends only on the observed data.

Estimator of λ. Because $\mathbb{E}[X] = \lambda$ for a Poisson distribution, a natural estimator of λ is the sample mean:

$$\hat{\lambda} = \bar{X}.$$

Here:

- λ is the unknown *parameter*,
- $\bar{X}$ is a *statistic*,
- using $\hat{\lambda} = \bar{X}$ makes $\bar{X}$ an *estimator* of λ.

Goal of inference. Using the observed counts, we aim to:

- estimate the emission rate λ,
- assess the variability (standard error) of $\hat{\lambda}$,
- draw probabilistic conclusions about λ (e.g., confidence intervals or hypothesis tests).

Parameters describe the underlying probabilistic model, statistics summarize observed data, and estimators link the two by providing data-driven approximations to unknown parameters. Once a model such as Poisson(λ) is assumed, all inference about future or unobserved behavior of the system is determined by the parameter λ.

1.2 Sampling distribution and standard error

Before the data are observed, an estimator $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ is a *random variable*, since it is a function of random sample observations. The probability distribution of $\hat{\theta}$ induced by repeated sampling under the same data-generating mechanism is called the **sampling distribution** of $\hat{\theta}$. The sampling distribution describes how the estimator would vary from sample to sample if the data collection process were repeated many times. Its center reflects systematic behavior (bias), while its spread reflects random fluctuation due to sampling variability. A fundamental numerical summary of this spread is the **standard error (SE)**, defined as

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}.$$

The standard error is the standard deviation of the sampling distribution of $\hat{\theta}$. The standard error quantifies the *typical magnitude of random estimation error* caused solely by sampling variability. A smaller standard error corresponds to a tighter sampling distribution and more reliable estimation.

Interpretation. If the sampling distribution of $\hat{\theta}$ is concentrated tightly around the true parameter value θ, then $\hat{\theta}$ will typically be close to θ in repeated samples. Conversely, a wide sampling distribution indicates substantial uncertainty due to random sampling.

Practical implications.

- The standard error sets the natural scale for judging estimation accuracy: differences on the order of one SE are common, while differences of several SEs are relatively unlikely.
- Standard errors decrease as the sample size increases, typically at rate $1/\sqrt{n}$, reflecting diminishing returns from additional data.
- Confidence intervals and hypothesis tests are constructed by combining the estimator with its standard error, making SE a central quantity for inference.

In practice, $\mathrm{Var}(\hat{\theta})$ is rarely known exactly and must be estimated from the data. The resulting *estimated standard error* plays the same conceptual role, serving as a data-based measure of uncertainty in the estimator.

2 Evaluating estimators: bias, variance, MSE, and consistency

In parametric inference, we typically tighten two loops at the same time: (i) we *construct* an estimator $\hat{\theta}$ from the data, and (ii) we *evaluate* how well it behaves under repeated sampling from the assumed model. This section collects the basic criteria used throughout Ch 8: bias, variance, mean squared error, standard error, and consistency. These quantities describe the sampling distribution of $\hat{\theta}$ and clarify why two estimators for the same parameter can behave very differently.

2.1 Bias, variance, and mean squared error (MSE)

Theorem

Let $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator of a fixed parameter θ. All expectations and variances below are taken with respect to the sampling distribution of $\hat{\theta}$ under the assumed model.

(1) Bias.

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

Derivation. Define the estimation error $e = \hat{\theta} - \theta$. Then,

$$\mathbb{E}[e] = \mathbb{E}[\hat{\theta} - \theta] = \mathbb{E}[\hat{\theta}] - \theta,$$

so bias is simply the *mean error* under repeated sampling:

$$\mathrm{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta} - \theta].$$

(2) Variance.

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\Big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\Big].$$

Derivation. By definition, the variance of any random variable Y is

$$\mathrm{Var}(Y) = \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big].$$

Applying this with $Y = \hat{\theta}$ gives the displayed formula. Expand the square:

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\Big[\hat{\theta}^2 - 2\hat{\theta}\,\mathbb{E}[\hat{\theta}] + (\mathbb{E}[\hat{\theta}])^2\Big] = \mathbb{E}[\hat{\theta}^2] - 2(\mathbb{E}[\hat{\theta}])^2 + (\mathbb{E}[\hat{\theta}])^2 = \mathbb{E}[\hat{\theta}^2] - (\mathbb{E}[\hat{\theta}])^2.$$

Hence,

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}[\hat{\theta}^2] - \big(\mathbb{E}[\hat{\theta}]\big)^2.$$

(3) Mean squared error (MSE).

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\Big[(\hat{\theta} - \theta)^2\Big].$$

MSE is the expected squared error under repeated sampling: if $e = \hat{\theta} - \theta$, then $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}[e^2]$.
MSE decomposition (bias–variance identity) is:

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2.$$

Derivation. Add and subtract $\mathbb{E}[\hat{\theta}]$ inside the error:

$$\hat{\theta} - \theta = \big(\hat{\theta} - \mathbb{E}[\hat{\theta}]\big) + \big(\mathbb{E}[\hat{\theta}] - \theta\big).$$

Square both sides:

$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta) + (\mathbb{E}[\hat{\theta}] - \theta)^2.$$

Take expectations term-by-term:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + 2(\mathbb{E}[\hat{\theta}] - \theta)\,\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]] + (\mathbb{E}[\hat{\theta}] - \theta)^2.$$

The middle expectation is zero because $\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]] = 0$. Therefore,

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}_{\mathrm{Var}(\hat{\theta})} + \underbrace{(\mathbb{E}[\hat{\theta}] - \theta)^2}_{\mathrm{Bias}(\hat{\theta})^2},$$

which yields

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2.$$

Implication / Interpretation

Implications. The MSE splits error into two components:

- **Variance** captures random fluctuation from sample to sample. Two estimators can have the same bias but very different variances.
- **Bias** captures systematic deviation from θ under repeated sampling. An unbiased estimator can still be inaccurate if its variance is large.
- **Bias–variance tradeoff:** a slightly biased estimator can have smaller MSE if it reduces variance enough. In practice we often prefer the estimator with smaller MSE, not necessarily the unbiased one.

2.2 Sampling distribution, standard error, and plug-in estimation

Bias and variance are properties of the *sampling distribution* of $\hat{\theta}$. A convenient summary of its spread is the **standard error (SE)**:

$$\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}.$$

When $\mathrm{Var}(\hat{\theta})$ depends on unknown parameters, we replace them with estimates; this produces a **plug-in** (estimated) SE:

$$\widehat{\mathrm{SE}}(\hat{\theta}) = \sqrt{\widehat{\mathrm{Var}}(\hat{\theta})}.$$

Implication / Interpretation

Implications.

- The SE sets the *natural scale* for typical estimation noise: errors of about one SE are common under repeated sampling.
- Plug-in SEs are ubiquitous in practice because exact variances are rarely known. The quality of inference can depend strongly on whether the plug-in approximation is accurate.
- SE is the building block for large-sample approximations, CI, and test statistics.

Example

Poisson mean (SE of $\hat{\lambda} = \bar{X}$). If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\lambda)$ and $\hat{\lambda} = \bar{X}$, then

$$\mathbb{E}[\bar{X}] = \lambda, \qquad \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X_1)}{n} = \frac{\lambda}{n}.$$

Hence

$$\mathrm{SE}(\hat{\lambda}) = \sqrt{\frac{\lambda}{n}} \approx \sqrt{\frac{\hat{\lambda}}{n}} = \sqrt{\frac{\bar{X}}{n}}.$$

Interpretation: the typical deviation of $\bar{X}$ from λ is of order $\sqrt{\lambda/n}$; doubling the sample size reduces SE by a factor of $1/\sqrt{2}$ (diminishing returns).

2.3 Consistency (large-sample behavior)

Theorem

Consistency. An estimator sequence $\hat{\theta}_n$ is **consistent** for θ if $\hat{\theta}_n \xrightarrow{p} \theta$ as $n \to \infty$.

Implication / Interpretation

Implications.

- Consistency is a **large-sample guarantee**: with more data, $\hat{\theta}_n$ concentrates near θ.
- Consistency does *not* imply unbiasedness for finite n, and it does *not* guarantee good performance when n is small.
- Many biased estimators are consistent (their bias vanishes as $n \to \infty$).

2.4 A key derivation: unbiased sample variance

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Define the two common "sample variance" versions:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Theorem

Unbiasedness of S^2 and bias of s_n^2.

$$\mathbb{E}[S^2] = \sigma^2, \qquad \mathbb{E}[s_n^2] = \frac{n-1}{n}\sigma^2.$$

Proof (detailed)

Derivation (finite-sample, no asymptotics). Start from the identity

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\big[(X_i - \mu) - (\bar{X} - \mu)\big]^2.$$

Expand, sum, and use $\sum_{i=1}^{n}(X_i - \mu) = n(\bar{X} - \mu)$:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) + n(\bar{X} - \mu)^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2.$$

Take expectations:

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \sum_{i=1}^{n}\mathbb{E}[(X_i - \mu)^2] - n\,\mathbb{E}[(\bar{X} - \mu)^2].$$

Now, $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$ and $\mathbb{E}[(\bar{X} - \mu)^2] = \mathrm{Var}(\bar{X}) = \sigma^2/n$, so

$$\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2.$$

Divide by $n-1$ to obtain $\mathbb{E}[S^2] = \sigma^2$, and divide by n to obtain $\mathbb{E}[s_n^2] = \frac{n-1}{n}\sigma^2$.

Implication / Interpretation

Implications.

- Using denominator n underestimates σ^2 by the factor $(n-1)/n$. This bias is most visible when n is small and becomes negligible as n grows.
- The $n-1$ correction yields an **exact** unbiasedness result for all $n \geq 2$.
- Despite being biased, s_n^2 is still **consistent** because $(n-1)/n \to 1$.

2.5 Parametric models, goals, and notation

We now place these evaluation criteria into the broader parametric-model setting. Let $X_1, \ldots, X_n$ be i.i.d. from a population distribution

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x; \theta),$$

where $f(x;\theta)$ is a pmf or pdf and θ is an unknown model parameter (scalar or vector). **Knowledge of θ determines everything about the model:** once θ is fixed, the entire probability law for X—and hence the sampling behavior of statistics and estimators—is determined. A family specified up to an unknown parameter is called a **parametric model**.

Given observed data $x_1, \ldots, x_n$, our goal is **estimation and inference** about θ: we construct an estimator $\hat{\theta}$ and then study (or approximate) its mean, variance, MSE, and its sampling distribution, especially as n increases.

In the next sections, we develop three major approaches to estimate θ:

- Method of moments (MOM),
- Maximum likelihood estimation (MLE),
- Bayesian analysis (studied in Part II).

We will evaluate the resulting estimators using the criteria introduced above and examine how their distributions behave in large samples.

Convention (notation).

- Use Greek letters for model parameters (e.g., $\omega, \rho, \vartheta, \varpi$).
- Place hats to denote estimators or realized estimates (e.g., $\hat{\omega}, \hat{\rho}, \hat{\vartheta}, \hat{\varpi}$).
- If helpful, show sample-size dependence explicitly via subscripts (e.g., $\hat{\varpi}_n$).
- Context determines whether $\hat{\varpi}$ refers to the random estimator (before observing data) or the numerical estimate (after observing $x_1, \ldots, x_n$).

3 Method of Moments (MOM)

3.1 General construction of the MoM

The method of moments, introduced by Chebyshev (1887), is based on a simple and intuitive idea: *if a probability model is correct, then its theoretical (population) moments should agree with the corresponding moments computed from the data.* Thus, MOM estimates the unknown parameter by forcing agreement between population moments (which depend on θ) and sample moments (which depend on the observed data).

Theorem

Method of Moments (MOM). Let $X_1, \dots, X_n$ be i.i.d. from a distribution $f(x;\theta)$, where $\theta \in \mathbb{R}^d$ is an unknown parameter. Choose d population moments (or functions) $g_1, \dots, g_d$ and define the population moment equations

$$\mathbb{E}_\theta[g_j(X)] = m_j(\theta), \qquad j = 1, \dots, d.$$

The method of moments estimator $\hat{\theta}_{\text{MOM}}$ is obtained by solving the system

$$m_j(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_j(X_i), \qquad j = 1, \dots, d,$$

for θ.

Population moments. Assume X has pdf or pmf $f(x;\theta)$. The k-th *population moment* is defined as

$$\mu_k(\theta) = \mathbb{E}_\theta[X^k] = \begin{cases} \displaystyle\int x^k f(x;\theta)\,dx, & \text{(continuous case)}, \\ \displaystyle\sum_x x^k f(x;\theta), & \text{(discrete case)}. \end{cases}$$

Crucially, $\mu_k(\theta)$ depends *only* on the parameter θ, not on the observed sample.

Sample moments. Given a random sample $X_1, \dots, X_n$, the k-th *sample moment* is

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Because it is a function of the random sample, $\hat{\mu}_k$ is itself a random variable. After observing data $x_1, \dots, x_n$, the realized value

$$\hat{\mu}_k(x_1, \dots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

is a numerical *estimate* of the population moment.

Defining equations of MOM. The method of moments equates population and sample moments:

$$\mu_1(\theta) = \hat{\mu}_1, \qquad \mu_2(\theta) = \hat{\mu}_2, \qquad \dots$$

More generally, for a d-dimensional parameter θ, we solve the system

$$\mu_j(\theta) = \hat{\mu}_j, \qquad j = 1, \dots, d,$$

where $\mu_j(\theta)$ emphasizes dependence on the parameter and $\hat{\mu}_j = \hat{\mu}_j(X_1, \ldots, X_n)$ emphasizes dependence on the data.

Estimator vs. estimate. Solving the moment equations yields

$$\hat{\theta}_{\text{MOM}} = \hat{\theta}_{\text{MOM}}(X_1, \ldots, X_n),$$

which is a *random variable* and hence an estimator. When the observed data $x_1, \ldots, x_n$ are plugged in, the resulting value $\hat{\theta}_{\text{MOM}}(x_1, \ldots, x_n)$ is a numerical *estimate* of θ.

Procedure summary.

- Specify a parametric model $f(x; \theta)$.
- Write down the first d population moments $\mu_1(\theta), \ldots, \mu_d(\theta)$.
- Compute the corresponding sample moments $\hat{\mu}_1, \ldots, \hat{\mu}_d$.
- Solve $\mu_j(\theta) = \hat{\mu}_j$ for θ.

Implication / Interpretation

The method of moments is often algebraically simple and frequently yields closed-form estimators. However, it is not automatically optimal: different choices of moments can lead to different estimators, and MOM estimators are often less statistically efficient (larger variance) than maximum likelihood estimators. Nevertheless, MOM provides a natural starting point and useful benchmarks for more sophisticated methods such as MLE.

3.2 Examples of the MoM

We now illustrate the method of moments through several canonical examples. Each example follows the same structure: (i) identify population moments, (ii) compute sample moments, (iii) equate them to obtain the MOM estimator, and (iv) interpret the result.

3.2.1 Example 1: Bernoulli distribution

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a Bernoulli(p) distribution, with parameter $p \in (0, 1)$. Recall that

$$\mathbb{E}[X] = p, \qquad \text{Var}(X) = p(1 - p).$$

Population and sample moments. The first population moment is $\mu_1(p) = p$. The corresponding sample moment is

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

MOM estimator. Equating population and sample moments,

$$p = \bar{X},$$

so the MOM estimator is

$$\hat{p}_{\text{MOM}} = \bar{X}.$$

Numerical illustration. If $n = 100$ and $\sum_{i=1}^{100} x_i = 55$, then

$$\hat{p}_{\text{MOM}} = \frac{55}{100} = 0.55.$$

The MOM estimator equals the sample proportion of successes, which is intuitive and coincides with the MLE in this case.

3.2.2 Example 2: Normal distribution

Let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, \sigma^2)$. The first two population moments are

$$\mathbb{E}[X] = \mu, \qquad \text{Var}(X) = \sigma^2.$$

Population and sample moments. For $X \sim N(\mu, \sigma^2)$, the first two population moments are

$$\mu_1 = \mathbb{E}[X] = \mu, \qquad \mu_2 = \mathbb{E}[X^2] = \mu^2 + \sigma^2.$$

The second-moment identity follows from $\text{Var}(X) = \mathbb{E}[X^2] - \{\mathbb{E}[X]\}^2$:

$$\mathbb{E}[X^2] = \text{Var}(X) + \{\mathbb{E}[X]\}^2 = \sigma^2 + \mu^2.$$

Given an i.i.d. sample $X_1, \ldots, X_n$, the corresponding sample moments are

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \qquad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

(As functions of the random sample, $\hat{\mu}_1, \hat{\mu}_2$ are random variables; after observing data, they become numerical values.)

MOM estimator. The method of moments sets population moments equal to sample moments and solves for (μ, σ^2):

$$\mathbb{E}[X] = \hat{\mu}_1 \quad \Longrightarrow \quad \mu = \bar{X},$$

and

$$\mathbb{E}[X^2] = \hat{\mu}_2 \quad \Longrightarrow \quad \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Substitute the first equation ($\mu = \bar{X}$) into the second:

$$\bar{X}^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \quad \Longrightarrow \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Thus the MOM estimators are

$$\hat{\mu}_{\text{MOM}} = \bar{X}, \qquad \hat{\sigma}^2_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Finally, note the algebraic identity

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}\big(X_i^2 - 2\bar{X}X_i + \bar{X}^2\big) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - 2\bar{X}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) + \bar{X}^2.$$

Since $\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$, the middle term becomes $2\bar{X} \cdot \bar{X} = 2\bar{X}^2$, so

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Therefore we can equivalently write

$$\hat{\sigma}^2_{\text{MOM}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,$$

which is the usual "denominator n" sample variance (biased but consistent). The MOM estimator for σ^2 uses denominator n and is biased; this contrasts with the unbiased estimator using $n-1$. MOM prioritizes moment matching, not unbiasedness.

3.2.3 Example 3: A nonstandard distribution

Suppose $X_1, \ldots, X_n$ are i.i.d. with density

$$f(x;\omega) = \frac{1+\omega x}{2}, \qquad -1 \le x \le 1, \quad \omega \in [-1,1].$$

Validity of the model. Before applying MOM, we verify that $f(x;\omega)$ is a legitimate density. For $\omega \in [-1,1]$ and $x \in [-1,1]$, we have $1+\omega x \ge 0$, so $f(x;\omega) \ge 0$. Moreover,

$$\int_{-1}^{1} f(x;\omega)\,dx = \frac{1}{2}\int_{-1}^{1}(1+\omega x)\,dx = \frac{1}{2}\left[2+0\right] = 1,$$

so f integrates to one. Hence the model is well-defined for $\omega \in [-1,1]$.

Population moment. To apply the method of moments, we compute the first population moment. By definition,

$$\mathbb{E}[X] = \int_{-1}^{1} x f(x;\omega)\,dx = \frac{1}{2}\int_{-1}^{1} x(1+\omega x)\,dx.$$

Split the integral:

$$\mathbb{E}[X] = \frac{1}{2}\left(\int_{-1}^{1} x\,dx + \omega\int_{-1}^{1} x^2\,dx\right).$$

The first integral vanishes by symmetry:

$$\int_{-1}^{1} x\,dx = 0,$$

while

$$\int_{-1}^{1} x^2\,dx = \left[\frac{x^3}{3}\right]_{-1}^{1} = \frac{2}{3}.$$

Therefore,

$$\mathbb{E}[X] = \frac{1}{2}\left(0 + \omega\cdot\frac{2}{3}\right) = \frac{\omega}{3}.$$

Sample moment. The corresponding first sample moment is the sample mean

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

As a statistic, $\bar{X}$ depends only on the observed data and is a random variable before the data are observed.

MOM estimator. The method of moments equates the population moment to the sample moment:

$$\bar{X} = \frac{\omega}{3}.$$

Solving for ω yields the MOM estimator

$$\hat{\omega}_{\text{MOM}} = 3\bar{X}.$$

Thus, the estimator is a simple linear function of the sample mean. Because each $X_i \in [-1, 1]$, the sample mean must satisfy $\bar{X} \in [-1, 1]$. Consequently,

$$\hat{\omega}_{\text{MOM}} = 3\bar{X} \in [-3, 3].$$

However, the parameter space of the model is $\omega \in [-1, 1]$. Hence, with positive probability, $\hat{\omega}_{\text{MOM}}$ falls outside the admissible parameter range.

Interpretation and implication. This example highlights an important limitation of the method of moments: *MOM does not automatically respect parameter constraints.* The estimator is obtained purely by algebraic moment matching, without enforcing the structural restrictions of the model. In practice, one may truncate the estimator to $[-1, 1]$ or prefer alternative methods (such as MLE), which typically enforce parameter constraints by construction.

3.2.4 Example 4: Poisson distribution

Let $X_1, \ldots, X_n$ be i.i.d. from Poisson(λ). Recall that a Poisson random variable has the special property that its mean and variance are equal:

$$\mathbb{E}[X] = \lambda, \qquad \text{Var}(X) = \lambda.$$

This single–parameter structure makes the Poisson model particularly convenient for illustrating estimation principles.

MOM estimator. The method of moments (MOM) estimator is obtained by equating a population moment to its sample counterpart. Here, the first population moment is $\mathbb{E}[X] = \lambda$, and the corresponding sample moment is the sample mean $\bar{X}$. Setting these equal gives

$$\mathbb{E}[X] = \bar{X} \quad \Rightarrow \quad \hat{\lambda}_{\text{MOM}} = \bar{X}.$$

Thus, the MOM estimator of λ is simply the sample average. Intuitively, this makes sense as λ represents the average rate of occurrence, the empirical average of the observed counts is a natural estimate.

Standard error. Because $X_1, \ldots, X_n$ are independent,

$$\mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n} = \frac{\lambda}{n}.$$

The standard error (SE) measures the typical sampling fluctuation of the estimator around the true parameter value. Therefore,

$$\mathrm{SE}(\hat{\lambda}_{\mathrm{MOM}}) = \sqrt{\mathrm{Var}(\bar{X})} = \sqrt{\frac{\lambda}{n}}.$$

Since the true value of λ is unknown in practice, we replace it with the MOM estimate $\hat{\lambda}_{\mathrm{MOM}} = \bar{X}$, yielding the plug-in approximation

$$\mathrm{SE}(\hat{\lambda}_{\mathrm{MOM}}) \approx \sqrt{\frac{\hat{\lambda}_{\mathrm{MOM}}}{n}}.$$

Large-sample distribution. By the Central Limit Theorem (CLT), the sample mean of i.i.d. observations with finite variance is approximately normally distributed for large n. In particular,

$$\sqrt{n}(\bar{X} - \lambda) \xrightarrow{d} N(0, \lambda).$$

Equivalently, for large sample sizes,

$$\bar{X} \approx N\left(\lambda, \frac{\lambda}{n}\right).$$

Since $\hat{\lambda}_{\mathrm{MOM}} = \bar{X}$, this result provides an approximate sampling distribution for the estimator, which forms the basis for large-sample confidence intervals and hypothesis tests for λ.

3.2.5 Example 5: Gamma distribution

Let $X_1, \ldots, X_n$ be i.i.d. from Gamma(α, β) with shape $\alpha > 0$ and rate $\beta > 0$. The Gamma distribution is flexible and commonly used to model nonnegative, right-skewed data such as waiting times and lifetimes. Its first and second population moments are

$$\mathbb{E}[X] = \frac{\alpha}{\beta}, \qquad \mathbb{E}[X^2] = \frac{\alpha(\alpha+1)}{\beta^2}.$$

Moment equations. The method of moments (MOM) equates population moments with their sample counterparts. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

denote the first and second sample moments, respectively. Equating moments gives the system

$$\bar{X} = \frac{\alpha}{\beta}, \qquad \overline{X^2} = \frac{\alpha(\alpha+1)}{\beta^2}.$$

Solving. From the first equation,

$$\alpha = \beta \bar{X}.$$

Substituting into the second moment equation yields

$$\overline{X^2} = \frac{(\beta\bar{X})(\beta\bar{X}+1)}{\beta^2} = \bar{X}^2 + \frac{\bar{X}}{\beta}.$$

Rearranging,

$$\overline{X^2} - \bar{X}^2 = \frac{\bar{X}}{\beta}.$$

Noting that $\overline{X^2} - \bar{X}^2$ is the sample analogue of the variance, we solve for

$$\hat{\beta}_{\text{MOM}} = \frac{\bar{X}}{\overline{X^2} - \bar{X}^2}, \qquad \hat{\alpha}_{\text{MOM}} = \beta\bar{X} = \frac{\bar{X}^2}{\overline{X^2} - \bar{X}^2}.$$

The quantity $\overline{X^2} - \bar{X}^2$ measures dispersion relative to the mean. When the second moment is close to $\bar{X}^2$, variability is small, leading to a large estimate of the shape parameter α and a more concentrated Gamma distribution. Larger gaps between $\overline{X^2}$ and $\bar{X}^2$ indicate greater dispersion and result in smaller $\hat{\alpha}$. The rate parameter β governs the overall scale of the distribution: larger values correspond to faster decay of the density.

Implication / Interpretation

- MOM can be formulated using raw moments rather than central moments.
- Using $\mathbb{E}[X^2]$ avoids introducing variance explicitly.
- The resulting estimators coincide algebraically with variance-based MOM.
- Exact sampling distributions are typically unavailable.
- Bootstrap methods provide a practical tool for inference.

3.2.6 Properties of the MOM estimator

Let $\hat{\theta}$ denote a method of moments (MOM) estimator of a population parameter θ. A natural question is: *How reliable is $\hat{\theta}$ as an estimator of θ?* To address this, we study the *sampling distribution* of $\hat{\theta}$, or an approximation to it.

Sampling distribution. The sampling distribution of an estimator is the probability distribution of $\hat{\theta}$ induced by repeated sampling from the underlying population. It describes how $\hat{\theta}$ varies from sample to sample and forms the basis for assessing estimator accuracy.

Bias. The mean of the sampling distribution is $\mathbb{E}[\hat{\theta}]$. If

$$\mathbb{E}[\hat{\theta}] = \theta,$$

then $\hat{\theta}$ is called an *unbiased* estimator of θ. The bias of $\hat{\theta}$ is defined as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

An estimator with small or zero bias is centered, on average, at the true parameter value.

Variance and standard deviation. The variability of the estimator is measured by its variance,

$$\mathrm{Var}(\hat{\theta}),$$

and the standard deviation of the estimator (often called the *standard error*) is

$$\mathrm{SD}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}.$$

Smaller variance indicates that $\hat{\theta}$ is more concentrated around its mean and therefore more stable across samples.

Accuracy and reliability. Together, bias and variance determine the accuracy of an estimator. Even an unbiased estimator may be unreliable if its variance is large, while a slightly biased estimator with small variance may perform well in practice.

Large-sample behavior. For many MOM estimators, exact sampling distributions are difficult or impossible to obtain in finite samples. However, under mild regularity conditions, MOM estimators are often *consistent* and *asymptotically normal*, with their large-sample behavior derived using the Central Limit Theorem and the delta method. This provides approximate standard errors and confidence intervals.

Bootstrap as an alternative. When the sampling distribution of a MOM estimator is not readily accessible, a popular and practical alternative is the *bootstrap*. Bootstrap methods approximate the sampling distribution of $\hat{\theta}$ by repeatedly resampling (with replacement) from the observed data and recomputing the estimator. This approach is especially useful for complex estimators or small-sample settings where analytic approximations may be inaccurate.

3.3 Bootstrap for MOM Estimators

Suppose $X_1, \ldots, X_n$ are i.i.d. from a distribution $F(\cdot;\theta)$, where the family F is known but the parameter θ is unknown. Let $\hat{\theta} = T(X)$ be an estimator computed from the observed sample $X = (X_1, \ldots, X_n)$. To interpret $\hat{\theta}$, we need its *sampling distribution* which is the distribution of $\hat{\theta}$ across repeated samples from $F(\cdot;\theta)$, because it determines variability/accuracy (e.g., standard error) and supports confidence intervals. In practice, the exact sampling distribution is rarely available in closed form. Bootstrap approximates it computationally by mimicking repeated sampling, replacing the unknown θ (or unknown F) with an estimate.

Example. Let X be precipitation amount for a randomly selected storm, and we observe $n = 227$ storms. A Gamma model is assumed:

$$X \sim \mathrm{Gamma}(\alpha, \lambda),$$

where $\alpha > 0$ is the *shape* and $\lambda > 0$ is the *rate*. Using the MOM approach, we equate population moments to sample moments. Define the first two *sample moments*

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Then, the sample variance can be written as

$$\widehat{\mathrm{Var}}(X) = \hat{\mu}_2 - \hat{\mu}_1^2.$$

For Gamma(α, λ),

$$\mathbb{E}[X] = \frac{\alpha}{\lambda}, \qquad \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}.$$

By equating $\mathbb{E}[X] = \hat{\mu}_1$ and $\mathrm{Var}(X) = \hat{\mu}_2 - \hat{\mu}_1^2$ gives

$$\frac{\alpha}{\lambda} = \bar{X}, \qquad \frac{\alpha}{\lambda^2} = \hat{\mu}_2 - \bar{X}^2.$$

From the first equation, $\alpha = \lambda \bar{X}$. Substitute into the second:

$$\frac{\lambda \bar{X}}{\lambda^2} = \frac{\bar{X}}{\lambda} = \hat{\mu}_2 - \bar{X}^2 \quad \Rightarrow \quad \hat{\lambda}_{\mathrm{MOM}} = \frac{\bar{X}}{\hat{\mu}_2 - \bar{X}^2}.$$

Finally,

$$\hat{\alpha}_{\mathrm{MOM}} = \hat{\lambda}_{\mathrm{MOM}} \bar{X} = \frac{\bar{X}^2}{\hat{\mu}_2 - \bar{X}^2}.$$

In the notes, $\bar{x} = 0.224$ and $\hat{\sigma}^2 = 0.1338$ produce $\hat{\alpha} = 0.375$ and $\hat{\lambda} = 1.674$.

Key question. What is the *variability (accuracy)* of $\hat{\alpha}_{\mathrm{MOM}}$ and $\hat{\lambda}_{\mathrm{MOM}}$?
If we knew the true parameters (α_0, λ_0), we could simulate many samples from Gamma(α_0, λ_0) and empirically approximate the sampling distribution of the estimators. But, (α_0, λ_0) are unknown; bootstrap replaces them with estimates (parametric bootstrap) or replaces F by the empirical distribution (nonparametric bootstrap). More precisely, the variability of an estimator is described by its sampling distribution. From this distribution, we can compute

$$\mathrm{Var}(\hat{\alpha}_{\mathrm{MOM}}), \quad \mathrm{Var}(\hat{\lambda}_{\mathrm{MOM}}),$$

their standard errors, and construct confidence intervals. Since closed-form expressions are difficult to obtain here, we approximate the sampling distribution computationally.

Parametric bootstrap:

1. Use the fitted model Gamma$(\hat{\alpha}, \hat{\lambda})$ as an estimate of the true distribution.
2. Generate B bootstrap samples of size $n = 227$ from this fitted Gamma model.
3. For each bootstrap sample, compute $\hat{\alpha}^{*(b)}$ and $\hat{\lambda}^{*(b)}$.
4. Use the empirical distribution of $\{\hat{\alpha}^{*(b)}\}_{b=1}^{B}$ and $\{\hat{\lambda}^{*(b)}\}_{b=1}^{B}$ to estimate standard errors and construct confidence intervals.

Nonparametric bootstrap:

1. Draw B samples of size $n = 227$ by sampling "with replacement" from the observed data.
2. Compute $\hat{\alpha}^{*(b)}$ and $\hat{\lambda}^{*(b)}$ for each resampled dataset.
3. Approximate the sampling distribution using the empirical distribution of these bootstrapped replicates.

In both cases, the bootstrap distribution provides an estimate of the estimator's variability, allowing us to quantify uncertainty and interpret the MOM estimates more cautiously.

A. Parametric bootstrap (Model-based resampling)

Idea. Assume that the parametric model, $F(\cdot;\theta)$, given to us is correct. But we don't know the values of parameters. Now we are going to approximate $F(\cdot;\theta)$ by $F(\cdot;\hat{\theta})$, then repeatedly sample from $F(\cdot;\hat{\theta})$ to emulate repeated sampling from the true model.

Algorithm. Fix a bootstrap replication count B.

1. Compute $\hat{\theta} = T(X)$ from the observed data.
2. For $b = 1, \ldots, B$:
   (a) Draw 'n' bootstrap samples $X^{*b} = (X_1^{*b}, \ldots, X_n^{*b})$ i.i.d. from $F(\cdot;\hat{\theta})$.
   (b) Compute the bootstrap estimate $\hat{\theta}_b^* = T(X^{*b})$.
3. Use the empirical distribution of $\{\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_B^*\}$ as an approximation to the sampling distribution of $\hat{\theta}$.

Gamma MoM specialization. Here $\theta = (\alpha, \lambda)$ and $F(\cdot;\hat{\theta}) = \text{Gamma}(\hat{\alpha}, \hat{\lambda})$. Each bootstrap replication produces

$$(\hat{\alpha}_b^*, \hat{\lambda}_b^*) = \left(\frac{\bar{X}_{*b}^2}{\hat{\mu}_{2,*b} - \bar{X}_{*b}^2}, \ \frac{\bar{X}_{*b}}{\hat{\mu}_{2,*b} - \bar{X}_{*b}^2} \right),$$

where $\bar{X}_{*b}$ and $\hat{\mu}_{2,*b}$ are computed from the bootstrap sample X^{*b}.

Example. Suppose that $X_1, \ldots, X_n$ are precipitation amounts from a storm sample with the size of $n = 227$ and we model their distribution as

$$X \sim \text{Gamma}(\alpha, \lambda),$$

where $\alpha > 0$ is the shape and $\lambda > 0$ is the rate. We estimate (α, λ) by MOM and then use a *parametric bootstrap* to approximate the sampling distribution of these estimators.

Step 0: Define the statistic.
Let $\theta = (\alpha, \lambda)$ and let $T(\cdot)$ be the MOM estimator computed from a dataset $X = (X_1, \ldots, X_n)$. Using sample moments

$$\hat{\mu}_1 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

the MOM estimators are

$$\hat{\lambda}_{\text{MOM}} = \frac{\hat{\mu}_1}{\hat{\mu}_2 - \hat{\mu}_1^2}, \qquad \hat{\alpha}_{\text{MOM}} = \hat{\lambda}_{\text{MOM}}\hat{\mu}_1 = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2}.$$

Thus, $T(X) = (\hat{\alpha}_{\text{MOM}}, \hat{\lambda}_{\text{MOM}})$.

Step 1: Compute the estimate from data.
From the observed sample $x_1, \ldots, x_n$, compute

$$\hat{\theta} = (\hat{\alpha}, \hat{\lambda}) = T(x_1, \ldots, x_n).$$

(For example, in the notes $\bar{x} = 0.224$ and $\hat{\sigma}^2 = 0.1338$ yield $\hat{\alpha} = 0.375$ and $\hat{\lambda} = 1.674$.)

Step 2: Generate bootstrap samples from the fitted model.
Fix a bootstrap replication count as B (e.g., $B = 1000$ or 5000). For $b = 1, \ldots, B$:

1. Generate a bootstrap dataset

$$X_1^{*(b)}, \ldots, X_n^{*(b)} \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\hat{\alpha}, \hat{\lambda}),$$

   i.e., simulate $n = 227$ precipitation amounts from the *fitted* Gamma model.

2. Compute the MOM estimators on the bootstrap dataset:

$$\hat{\theta}_b^* = (\hat{\alpha}_b^*, \hat{\lambda}_b^*) = T\Big(X_1^{*(b)}, \ldots, X_n^{*(b)}\Big).$$

Step 3: Approximate the sampling distribution and quantify variability.
The empirical distribution of the bootstrap replicates

$$\{(\hat{\alpha}_1^*, \hat{\lambda}_1^*), \ldots, (\hat{\alpha}_B^*, \hat{\lambda}_B^*)\}$$

approximates the joint sampling distribution of $(\hat{\alpha}, \hat{\lambda})$. From these replicates, we can estimate standard errors:

$$\widehat{\text{SE}}(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\alpha}_b^* - \bar{\hat{\alpha}}^*)^2}, \qquad \widehat{\text{SE}}(\hat{\lambda}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\lambda}_b^* - \bar{\hat{\lambda}}^*)^2},$$

where

$$\bar{\hat{\alpha}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\alpha}_b^*, \qquad \bar{\hat{\lambda}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\lambda}_b^*.$$

Step 4: Bootstrap confidence intervals (percentile method). A simple $100(1-\gamma)\%$ bootstrap CI for α is given by the empirical quantiles

$$\Big[q_{\gamma/2}\big(\hat{\alpha}^*\big),\ q_{1-\gamma/2}\big(\hat{\alpha}^*\big)\Big],$$

where $q_p(\hat{\alpha}^*)$ denotes the pth sample quantile of $\{\hat{\alpha}_1^*, \ldots, \hat{\alpha}_B^*\}$. Similarly for λ:

$$\Big[q_{\gamma/2}\big(\hat{\lambda}^*\big),\ q_{1-\gamma/2}\big(\hat{\lambda}^*\big)\Big].$$

If the bootstrap distribution is narrow, the MOM estimates are relatively precise; if it is wide or highly skewed, the estimates have substantial variability. The bootstrap, hence, provides a practical, data-driven way to assess estimator accuracy when analytic sampling distributions are unavailable.

B. Nonparametric bootstrap (Empirical-distribution resampling)

Idea. The parametric form of the distribution of our interest $F(\cdot;\theta)$ is not assumed. We only have a dataset. We are going to approximate this unknown CDF $F(\cdot;\theta)$ by the empirical distribution F_n, which assigns probability $1/n$ to each observed x_i

$$F_n(x) = \frac{\#\text{ of } \{x_i \le x,\ i = 1, \ldots, n\}}{n}, \qquad x \in \mathbb{R}.$$

Sampling from F_n is equivalent to sampling *"with replacement"* from $\{x_1, \ldots, x_n\}$.

Algorithm. Fix a bootstrap replication count B.

1. For $b = 1, \ldots, B$:
   (a) Draw X^{*b} by sampling n observations with replacement from $\{x_1, \ldots, x_n\}$.
   (b) Compute $\hat{\theta}_b^* = T(X^{*b})$.
2. Use $\{\hat{\theta}_b^*\}_{b=1}^B$ to approximate the sampling distribution of $\hat{\theta}$.

This is more generally applicable because $T(X)$ need not even be a parametric estimator. The bootstrap is typically called nonparametric because, in the resampling step, it does not make any assumption about the underlying distribution. The nonparametric bootstrap does not require that the underlying distn is parametric, nor that the statistic T(X) is an estimator for the model parameter, and can be applied more generally.

Example. Suppose $x_1, \ldots, x_{227}$ are the observed precipitations. We do not assume any parametric model for the resampling step. Instead, we approximate the unknown population distribution F by the empirical distribution $\hat{F}_n$, which places probability $1/n$ on each observed value.

Step 1: original estimate. From the observed data compute the MOM estimators

$$\hat{\theta} = (\hat{\alpha}, \hat{\lambda}) = T(x_1, \ldots, x_n).$$

Step 2: resample with replacement. Fix a bootstrap replication count B (e.g. $B = 1000$). For $b = 1, \ldots, B$:

1. Draw a bootstrap sample
   $$X^{*(b)} = (X_1^{*(b)}, \ldots, X_n^{*(b)})$$
   by sampling *with replacement* from $\{x_1, \ldots, x_n\}$. Each observation has probability $1/n$ of being selected at each draw.
2. Compute the bootstrap estimator
   $$\hat{\theta}_b^* = (\hat{\alpha}_b^*, \hat{\lambda}_b^*) = T\left(X^{*(b)}\right).$$

Step 3: approximate the sampling distribution. The empirical distribution of

$$\{(\hat{\alpha}_1^*, \hat{\lambda}_1^*), \ldots, (\hat{\alpha}_B^*, \hat{\lambda}_B^*)\}$$

approximates the sampling distribution of $(\hat{\alpha}, \hat{\lambda})$. From these replicates, we can compute:

- Estimated standard errors:
  $$\widehat{\mathrm{SE}}(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\alpha}_b^* - \bar{\hat{\alpha}}^*)^2},$$
  and similarly for $\hat{\lambda}$.

- Bootstrap percentile confidence intervals:

$$\Big[q_{\gamma/2}(\hat{\alpha}^*),\ q_{1-\gamma/2}(\hat{\alpha}^*)\Big],$$

  and similarly for λ.

Unlike the parametric bootstrap, we do not assume a Gamma distribution (or any other parametric form) in the resampling step. The method only assumes the data are i.i.d. Therefore, the nonparametric bootstrap can be applied to:

- nonparametric statistics (e.g., median, quantiles),
- complicated estimators without closed-form sampling distributions,
- situations where the underlying population model is unknown.

Thus, the nonparametric bootstrap provides a broadly applicable, model-free way to assess estimator variability.

C. Evaluate the accuracy of the estimator: Bootstrap standard error

The *standard error* of $\hat{\theta}$ is $\mathrm{SE}(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$ under repeated sampling from the true distribution. Bootstrap estimates it by the sample standard deviation of the bootstrap replicates:

$$\widehat{\mathrm{SE}}_B(\hat{\theta}) = \left(\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2\right)^{1/2}, \quad \text{where} \quad \bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_b^*.$$

The larger the variance the smaller the accuracy of the estimator. Likewise, an estimator with a small variance will have high accuracy.

D. Bootstrap confidence intervals: three common constructions

Let $\alpha \in (0,1)$ be the error level; we want a $(1-\alpha)100\%$ CI for θ.

1) Bootstrap Percentile CI. Compute the empirical $\alpha/2$ and $(1-\alpha/2)$ quantiles of the sampling distribution of $\{\hat{\theta}_b^*\}_{b=1}^B$, denoted $\hat{\theta}^*_{(\alpha/2)}$ and $\hat{\theta}^*_{(1-\alpha/2)}$. Then,

$$\mathrm{CI}_{\mathrm{perc}} = \Big(\hat{\theta}^*_{(\alpha/2)},\ \hat{\theta}^*_{(1-\alpha/2)}\Big).$$

Bootstrap Percentile CI Example

Using the parametric bootstrap (Gamma model assumed), we generated B bootstrap replicates of the MOM estimators $\hat{\alpha}^*$ and $\hat{\lambda}^*$.

95% percentile CI for α:

$$\big[q_{0.025}(\hat{\alpha}^*),\ q_{0.975}(\hat{\alpha}^*)\big] = (0.312,\ 0.585).$$

We are 95% confident that the true shape parameter α lies between 0.312 and 0.585.

95% percentile CI for λ:

$$\left[q_{0.025}(\hat{\lambda}^*),\ q_{0.975}(\hat{\lambda}^*)\right] = (1.447,\ 3.065).$$

We are 95% confident that the true rate parameter λ lies between 1.447 and 3.065.

Limitation. These intervals rely on the approximation that the sampling distribution of $\hat{\theta}$ under the true model $F(\cdot;\theta_0)$ is well approximated by the bootstrap distribution of $\hat{\theta}^*$ generated from $F(\cdot;\hat{\theta})$. If this approximation is poor (e.g., small sample size or highly skewed sampling distribution), percentile intervals may be inaccurate. More refined bootstrap intervals (e.g., bias-corrected or BCa intervals) can provide improved coverage accuracy.

2) Revised Bootstrap Percentile CI. The revised percentile interval is motivated by comparing the true estimation error

$$\hat{\theta} - \theta$$

(with randomness coming from $F(\cdot;\theta)$) to the bootstrap error

$$\hat{\theta}^* - \hat{\theta}$$

(with randomness coming from $F(\cdot;\hat{\theta})$ where $\hat{\theta}$ is fixed).

The bootstrap assumes that the distribution of the true error $\hat{\theta} - \theta$ is approximately the same as the distribution of $\hat{\theta}^* - \hat{\theta}$. Symbolically,

$$\hat{\theta} - \theta \ \approx\ \hat{\theta}^* - \hat{\theta}.$$

If this approximation is valid, then

$$P\Big(\hat{\theta}^*_{(\alpha/2)} \le \hat{\theta}^* \le \hat{\theta}^*_{(1-\alpha/2)}\Big) \approx 1-\alpha,$$

where $\hat{\theta}^*_{(p)}$ denotes the pth quantile of the bootstrap distribution. Subtracting $\hat{\theta}$ inside the probability:

$$P\Big(\hat{\theta}^*_{(\alpha/2)} - \hat{\theta} \le \hat{\theta}^* - \hat{\theta} \le \hat{\theta}^*_{(1-\alpha/2)} - \hat{\theta}\Big) \approx 1-\alpha.$$

Now we can replace $\hat{\theta}^* - \hat{\theta}$ by $\hat{\theta} - \theta$ (bootstrap approximation principle):

$$P\Big(\hat{\theta}^*_{(\alpha/2)} - \hat{\theta} \le \hat{\theta}^* - \hat{\theta} \le \hat{\theta}^*_{(1-\alpha/2)} - \hat{\theta}\Big) \approx P\Big(\hat{\theta}^*_{(\alpha/2)} - \hat{\theta} \le \hat{\theta} - \theta \le \hat{\theta}^*_{(1-\alpha/2)} - \hat{\theta}\Big) \approx 1-\alpha.$$

Rearranging to isolate θ gives:

$$P\Big(2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)} \le \theta \le 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\Big) \approx 1-\alpha.$$

Thus, the *reflected percentile interval* is

$$\mathrm{CI}_{\mathrm{refl}} = \Big(2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)},\ 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}\Big).$$

Unlike the ordinary percentile interval $(\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)})$, which directly uses bootstrap quantiles, the reflected interval recenters the bootstrap distribution around $\hat{\theta}$. In effect, it adjusts for potential bias in the bootstrap distribution. Geometrically, if the bootstrap distribution is slightly shifted

relative to $\hat{\theta}$, the reflection corrects this shift. If the bootstrap distribution of $\hat{\theta}^*$ is biased (i.e., its mean differs from $\hat{\theta}$), the ordinary percentile interval may inherit that bias. The reflected interval partially corrects for this bias by centering around $2\hat{\theta} - \hat{\theta}^*$. However, for heavily skewed sampling distributions, even better methods (e.g., BCa intervals) may provide improved coverage accuracy.

Example. In the notes, the 95% reflected-percentile CI for α is approximately

$$(0.265,\ 0.537).$$

We are approximately 95% confident that the true shape parameter α lies between 0.265 and 0.537.

3) Bootstrap t interval. This interval is inspired by the classical pivot

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx t_{n-1} \quad \text{(moderate } n\text{)},$$

and more generally considers the *studentized* statistic

$$T = \frac{\hat{\theta} - \theta}{\widehat{\mathrm{SE}}(\hat{\theta})}.$$

Because T is typically not exactly t-distributed, bootstrap estimates the distribution of T and uses its quantiles rather than t-quantiles. This requires *two-level* bootstrapping:

Bootstrap-t Algorithm. Fix Level-1 replication size B and Level-2 replication size B_2.

1. **Original estimate.** Compute $\hat{\theta} = T(X)$ from the observed data.
2. **Level 1 bootstrap (outer loop).** For $b = 1, \ldots, B$:
   (a) Draw a bootstrap sample $X^{*(b)} = (X_1^{*(b)}, \ldots, X_n^{*(b)})$.
   (b) Compute $\hat{\theta}_b^* = T(X^{*(b)})$.
   (c) **Level 2 bootstrap (inner loop).** From $X^{*(b)}$, draw B_2 resamples $X^{*(b,1)}, \ldots, X^{*(b,B_2)}$ and compute
   $$\hat{\theta}_{b,j}^{**} = T(X^{*(b,j)}), \quad j = 1, \ldots, B_2.$$
   (d) Estimate the conditional bootstrap standard error:
   $$\widehat{\mathrm{SE}}_b^* = \mathrm{SD}\big(\hat{\theta}_{b,1}^{**}, \ldots, \hat{\theta}_{b,B_2}^{**}\big).$$
   (e) Form the studentized statistic
   $$T_b^* = \frac{\hat{\theta}_b^* - \hat{\theta}}{\widehat{\mathrm{SE}}_b^*}.$$
3. **Quantiles of the studentized statistics.** Let $t_{\alpha/2}^*$ and $t_{1-\alpha/2}^*$ denote the empirical quantiles of $\{T_b^*\}_{b=1}^B$ such that
   $$P(T^* > t_{\alpha/2}^*) = \alpha/2.$$
4. **Final confidence interval.**
   Let $\widehat{\mathrm{SE}}(\hat{\theta}) = \mathrm{SD}\big(\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\big)$, then, the standard deviation of the Level-1 bootstrap estimates. Then the $(1-\alpha)100\%$ bootstrap-t CI is
   $$\mathrm{CI}_{t\text{-boot}} = \Big(\hat{\theta} - t_{1-\alpha/2}^*\,\widehat{\mathrm{SE}}(\hat{\theta}),\ \hat{\theta} - t_{\alpha/2}^*\,\widehat{\mathrm{SE}}(\hat{\theta})\Big).$$

4 Maximum Likelihood Estimation (MLE)

Let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution with density/pmf $f(x;\theta)$, and suppose we observe data $x_1, \ldots, x_n$. The goal is to estimate the (unknown) parameter value θ that generated the observed sample.

Likelihood function. The *likelihood function* treats the data as fixed and views $L(\theta)$ as a function of the parameter:

$$L(\theta) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\theta).$$

For discrete models, $f(x;\theta) = P_\theta(X = x)$, so $L(\theta; x)$ is the joint probability of the observed sample. Likelihood does not require identical distributions. If X_i are independent with $X_i \sim f_i(x;\theta)$, then

$$L(\theta) = \prod_{i=1}^{n} f_i(x_i;\theta).$$

Likelihood methods can also be defined for dependent data by using the appropriate joint density/pmf.

Example 1: Binomial(n, p)

Let $X \sim \text{Binomial}(n, p)$ and we observe a single value x (number of successes). The pmf is

$$P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n.$$

Step 1 (construct likelihood). Treat x as fixed and view the pmf as a function of p:

$$L(p) = L(p; x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad 0 < p < 1.$$

Step 2 (optional simplification). Since $\binom{n}{x}$ does not depend on p, maximizing $L(p)$ is equivalent to maximizing

$$\tilde{L}(p) = p^x (1-p)^{n-x}.$$

Example 2: Poisson(λ)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ and observe $x_1, \ldots, x_n$. The pmf is

$$P_\lambda(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

Step 1 (construct likelihood).

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}, \qquad \lambda > 0.$$

Step 2 (log-likelihood).

$$\ell(\lambda) = \log L(\lambda) = -n\lambda + \Big(\sum_{i=1}^{n} x_i\Big) \log \lambda - \sum_{i=1}^{n} \log(x_i!).$$

Example 3: Exponential(λ)

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exponential}(\lambda)$ with pdf

$$f(x;\lambda) = \lambda e^{-\lambda x}\mathbf{1}\{x > 0\}, \qquad \lambda > 0.$$

Given observed data $x_1, \dots, x_n$ (assume all $x_i > 0$):
Step 1 (construct likelihood).

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big), \qquad \lambda > 0.$$

Step 2 (log-likelihood).

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Example 4: Beta(α, β) on $(0,1)$

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta)$ with pdf

$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \qquad 0 < x < 1,\ \alpha > 0,\ \beta > 0,$$

where $B(\alpha,\beta)$ is the beta function. Given observed $x_1, \dots, x_n \in (0,1)$:
Step 1 (construct likelihood).

$$L(\alpha,\beta) = \prod_{i=1}^{n} \frac{x_i^{\alpha-1}(1-x_i)^{\beta-1}}{B(\alpha,\beta)} = \frac{\prod_{i=1}^{n} x_i^{\alpha-1}(1-x_i)^{\beta-1}}{\big(B(\alpha,\beta)\big)^n}.$$

Step 2 (log-likelihood).

$$\ell(\alpha,\beta) = (\alpha-1)\sum_{i=1}^{n} \log x_i + (\beta-1)\sum_{i=1}^{n} \log(1-x_i) - n \log B(\alpha,\beta).$$

Step 3 (useful identity). Using $B(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$,

$$\ell(\alpha,\beta) = (\alpha-1)\sum_{i=1}^{n} \log x_i + (\beta-1)\sum_{i=1}^{n} \log(1-x_i) - n\big[\log\Gamma(\alpha) + \log\Gamma(\beta) - \log\Gamma(\alpha+\beta)\big].$$

4.1 Likelihood, log-likelihood, and the MLE

Let $X_1, \dots, X_n$ be an i.i.d. sample from a model with density/pmf $f(x;\theta)$, where $\theta \in \Omega$ and Ω is the **parameter space**. Given observed data $x = (x_1, \dots, x_n)$, the **likelihood function** is defined by

$$L(\theta) = L(\theta; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\theta), \qquad \theta \in \Omega.$$

Equivalently, we often work with the **log-likelihood**

$$\ell(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta),$$

which is typically easier to differentiate and maximize.

Implication / Interpretation

Maximum likelihood estimation (MLE)
Because $\log(\cdot)$ is strictly increasing on $(0, \infty)$, we have for any $\theta_1, \theta_2 \in \Omega$:

$$L(\theta_1) > L(\theta_2) \iff \log L(\theta_1) > \log L(\theta_2).$$

Therefore,

$$\arg\max_{\theta \in \Omega} L(\theta) = \arg\max_{\theta \in \Omega} \ell(\theta),$$

whenever $L(\theta) > 0$ (so $\ell(\theta)$ exists).

Definition of the MLE: Estimate vs. Estimator. The maximum likelihood estimate is the parameter value in Ω that maximizes the likelihood:

$$\hat{\theta}_{\text{MLE}} \in \arg\max_{\theta \in \Omega} L(\theta; x_1, \ldots, x_n) \quad \Big(\text{or equivalently } \arg\max_{\theta \in \Omega} \ell(\theta)\Big).$$

This notation emphasizes that $\hat{\theta}_{\text{MLE}}$ is computed from the observed sample and thus depends on $x_1, \ldots, x_n$. To emphasize randomness, replace the observed values x_i by random variables X_i. Then,

$$\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MLE}}(x_1, \ldots, x_n) \quad \text{but} \quad \hat{\theta}_{\text{MLE}}(X_1, \ldots, X_n) \text{ is a } \textbf{random variable}.$$

We call the latter the **MLE (as an estimator)**; the distinction between *estimate* (a number) and *estimator* (a random variable) should be clear from context.

4.2 Computing the MLE via the log-likelihood

Score equation (first-order condition). In many problems, maximizing $L(\theta)$ directly is inconvenient, so we maximize $\ell(\theta) = \log L(\theta)$ instead. Assume $\ell(\theta)$ is differentiable. Define the score function and differentiate the log-likelihood:

$$U(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) \quad \text{(score)}.$$

$$U(\theta) = \ell'(\theta) = \frac{d}{d\theta}\, \ell(\theta) = \frac{d}{d\theta} \sum_{i=1}^{n} \log f(x_i; \theta).$$

If $\hat{\theta}$ is an interior maximizer (i.e., $\hat{\theta} \in \text{int}(\Omega)$), then a necessary condition is that it is a critical point:

$$U(\hat{\theta}) = \ell'(\hat{\theta}) = 0.$$

(If $\hat{\theta}$ lies on the boundary of Ω, we must also check boundary points separately.)

Curvature test (second-order condition). Assume further that $\ell(\theta)$ is twice continuously differentiable. Let

$$\ell''(\theta) = \frac{d^2}{d\theta^2}\,\ell(\theta).$$

If $\ell'(\hat{\theta}) = 0$ and

$$\ell''(\hat{\theta}) < 0,$$

then $\hat{\theta}$ is a point of **local maximum** of $\ell(\theta)$ (and hence also of $L(\theta)$).

Implication / Interpretation

Heuristic interpretation (sharp peak ⇒ higher precision).
Near a maximizer, larger negative curvature (more negative $\ell''(\hat{\theta})$) means the log-likelihood has a sharper peak, which typically corresponds to smaller variability of the MLE.

Observed information and Fisher information. A common curvature measure is the **observed information**

$$J(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta) = -\ell''(\theta),$$

which is nonnegative near a maximum. The **Fisher information** is typically defined as

$$I(\theta) = \mathbb{E}_\theta[J(\theta)] = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2\right],$$

under regularity conditions that justify exchanging differentiation and expectation. This formalizes the link between curvature and estimator variance in large samples.

Example 1: MLE for Poisson. Let $X_i \sim \text{Poisson}(\lambda)$. Then,

$$L(\lambda) = \prod_{i=1}^{n}\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{\sum x_i}\prod_{i=1}^{n}\frac{1}{x_i!}.$$

Log-likelihood:

$$\ell(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} x_i\right)\log\lambda - \sum_{i=1}^{n}\log(x_i!).$$

Differentiate and set to zero:

$$\frac{d}{d\lambda}\ell(\lambda) = -n + \frac{\sum x_i}{\lambda} = 0 \quad \Rightarrow \quad \hat{\lambda}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}.$$

To check if it is a maximum with the second derivative:

$$\frac{d^2}{d\lambda^2}\ell(\lambda) = -\frac{\sum x_i}{\lambda^2} < 0 \quad (\lambda > 0),$$

so $\ell(\lambda)$ is concave and the critical point is the global maximizer.

Implication / Interpretation

For Poisson, MOM and MLE coincide: both give $\bar{X}$. This is not always true, but it often happens when the mean directly equals a parameter.

Example 2. MLE for Gaussian. Assume $X_i \sim N(\mu, \sigma^2)$ with both μ and σ^2 unknown. The probability density function is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Log-likelihood:

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

MLE for μ (given σ^2). Differentiate w.r.t. μ:

$$\frac{\partial}{\partial\mu}\ell(\mu, \sigma^2) = -\frac{1}{2\sigma^2} \cdot 2\sum_{i=1}^{n}(x_i - \mu)(-1) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu).$$

Set to zero:

$$\sum_{i=1}^{n}(x_i - \mu) = 0 \quad \Rightarrow \quad \hat{\mu} = \bar{X}.$$

MLE for σ^2 (plug $\hat{\mu} = \bar{X}$). Differentiate w.r.t. σ^2:

$$\frac{\partial}{\partial\sigma^2}\ell(\mu, \sigma^2) = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Set to zero and solve:

$$-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum(x_i - \mu)^2 = 0 \quad \Rightarrow \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2.$$

Implication / Interpretation

- The MLE for μ is $\bar{X}$.
- The MLE for σ^2 uses denominator n (not $n-1$), hence is biased downward.
- Many courses prefer S^2 (with $n-1$) for unbiased estimation; MLE prioritizes likelihood optimality.

Example 1: Binomial(n, p)

Let $X \sim \text{Binomial}(n, p)$ and we observe x. Likelihood is

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad 0 < p < 1.$$

The log-likelihood is

$$\ell(p) = \log L(p) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p).$$

MLE (solve score equation). Differentiate:

$$\ell'(p) = \frac{x}{p} - \frac{n-x}{1-p}.$$

Set $\ell'(p) = 0$:

$$\frac{x}{p} = \frac{n-x}{1-p} \quad \Rightarrow \quad x(1-p) = p(n-x) \quad \Rightarrow \quad x = np \quad \Rightarrow \quad \hat{p}_{\text{MLE}} = \frac{x}{n}.$$

MLE as an estimator. Replacing x by the r.v. X gives the *ML estimator*

$$\hat{p}_{\text{MLE}} = \frac{X}{n}.$$

Randomness. Since $X \sim \text{Binomial}(n, p)$,

$$\hat{p}_{\text{MLE}} = \frac{X}{n} \quad \text{takes values} \quad \left\{0, \frac{1}{n}, \ldots, 1\right\},$$

and

$$\mathbb{E}[\hat{p}_{\text{MLE}}] = \frac{\mathbb{E}[X]}{n} = \frac{np}{n} = p, \qquad \text{Var}(\hat{p}_{\text{MLE}}) = \frac{\text{Var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Thus $\hat{p}_{\text{MLE}}$ is unbiased with variance $p(1-p)/n$.

Example 2: Poisson(λ)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Poisson}(\lambda)$ and observe $x_1, \ldots, x_n$.
Likelihood and log-likelihood.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}, \qquad \lambda > 0,$$

$$\ell(\lambda) = \log L(\lambda) = -n\lambda + \Big(\sum_{i=1}^{n} x_i\Big) \log \lambda - \sum_{i=1}^{n} \log(x_i!).$$

MLE (solve score equation). Differentiate:

$$\ell'(\lambda) = -n + \frac{\sum_{i=1}^n x_i}{\lambda}.$$

Set $\ell'(\lambda) = 0$:

$$-n + \frac{\sum_{i=1}^n x_i}{\lambda} = 0 \quad \Rightarrow \quad \hat{\lambda}_{\text{MLE}} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

MLE as an estimator.

$$\hat{\lambda}_{\text{MLE}} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Randomness. Since $\sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$,

$$\hat{\lambda}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^n X_i \quad \text{is a scaled Poisson r.v.}$$

Moreover,

$$\mathbb{E}[\hat{\lambda}_{\text{MLE}}] = \lambda, \qquad \text{Var}(\hat{\lambda}_{\text{MLE}}) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}(n\lambda) = \frac{\lambda}{n}.$$

So $\bar{X}$ is unbiased with variance λ/n.

Example 3: Exponential(λ)

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$ with pdf $f(x;\lambda) = \lambda e^{-\lambda x}\mathbf{1}\{x > 0\}$, and observe $x_1, \ldots, x_n > 0$.
Likelihood and log-likelihood.

$$L(\lambda) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n \exp\Big(-\lambda \sum_{i=1}^n x_i\Big), \qquad \lambda > 0,$$

$$\ell(\lambda) = \log L(\lambda) = n\log\lambda - \lambda\sum_{i=1}^n x_i.$$

MLE (solve score equation). Differentiate:

$$\ell'(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n x_i.$$

Set $\ell'(\lambda) = 0$:

$$\frac{n}{\lambda} = \sum_{i=1}^n x_i \quad \Rightarrow \quad \hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}.$$

MLE as an estimator.

$$\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}}.$$

Randomness. Let $S = \sum_{i=1}^n X_i$. For i.i.d. Exponential(λ),

$$S \sim \text{Gamma}(n, \lambda) \quad \text{(shape } n\text{, rate } \lambda\text{)}.$$

Thus $\hat{\lambda}_{\text{MLE}} = n/S$ is a transformation of a Gamma r.v. In particular, using the moment identity $\mathbb{E}[S^{-1}] = \dfrac{\lambda}{n-1}$ for $n > 1$,

$$\mathbb{E}[\hat{\lambda}_{\text{MLE}}] = n\,\mathbb{E}\left[\frac{1}{S}\right] = n \cdot \frac{\lambda}{n-1} = \lambda\frac{n}{n-1},$$

so $\hat{\lambda}_{\text{MLE}}$ is *biased upward* (for finite n), but the bias vanishes as $n \to \infty$. Also, since $\mathbb{E}[S^{-2}] = \dfrac{\lambda^2}{(n-1)(n-2)}$ for $n > 2$,

$$\text{Var}(\hat{\lambda}_{\text{MLE}}) = n^2\left(\mathbb{E}[S^{-2}] - (\mathbb{E}[S^{-1}])^2\right) = n^2\left(\frac{\lambda^2}{(n-1)(n-2)} - \frac{\lambda^2}{(n-1)^2}\right) = \frac{n^2\lambda^2}{(n-1)^2(n-2)}.$$

(These formulas require $n > 2$ for the variance to be finite.)

4.3 Large-sample approximation: asymptotic normality

Let $X_1, \dots, X_n$ be i.i.d. from a model $f(x;\theta)$ with true parameter θ_0. Define the (log-)likelihood

$$L_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta), \qquad \ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta),$$

and the MLE

$$\hat{\theta} = \arg\max_{\theta} \ell_n(\theta).$$

Define the *score* and *observed information*

$$U_n(\theta) = \ell_n'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i;\theta), \qquad J_n(\theta) = -\ell_n''(\theta).$$

4.3.1 Invariance property of the MLE

Theorem 1 (Invariance of the MLE)**.** *If $\hat{\theta}$ is the MLE of θ and $g(\cdot)$ is continuous, then*

$$g(\hat{\theta}) \text{ is the MLE of } g(\theta).$$

Proof sketch. Let $\eta = g(\theta)$. The likelihood for η is $L_\eta(\eta) = \sup_{\theta: g(\theta)=\eta} L(\theta)$. Since $L(\hat{\theta}) \geq L(\theta)$ for all θ, the maximizer over η is achieved at $\eta = g(\hat{\theta})$. Hence, $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Example

(Geometric odds). Let $X_1, \dots, X_n$ be i.i.d. Geometric(p) on $\{1, 2, \dots\}$ with

$$P_p(X = x) = (1-p)^{x-1}p.$$

Then

$$L(p) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p = p^n (1-p)^{\sum_{i=1}^{n}(x_i - 1)}.$$

Log-likelihood:

$$\ell(p) = n \log p + \Big(\sum_{i=1}^{n} x_i - n \Big) \log(1-p).$$

Differentiate and set to 0:

$$\ell'(p) = \frac{n}{p} - \frac{\sum x_i - n}{1-p} = 0 \implies \hat{p} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{X}}.$$

By invariance, the MLE of the odds $\dfrac{p}{1-p}$ is

$$\widehat{\frac{p}{1-p}} = \frac{\hat{p}}{1-\hat{p}} = \frac{n/\sum x_i}{1 - n/\sum x_i} = \frac{n}{\sum x_i - n} = \frac{1}{\bar{X} - 1}.$$

Example

MLE of $P_\lambda(X=0)$ for Poisson(λ). Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$ with pmf

$$f(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \qquad x = 0, 1, 2, \ldots$$

Step 1: Construct the likelihood.

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\prod_{i=1}^{n}\frac{1}{x_i!}.$$

Step 2: Log-likelihood.

$$\ell(\lambda) = -n\lambda + \Big(\sum_{i=1}^{n} x_i\Big)\log\lambda - \sum_{i=1}^{n}\log(x_i!).$$

Step 3: Differentiate and solve for MLE.

$$\ell'(\lambda) = -n + \frac{\sum_{i=1}^n x_i}{\lambda}.$$

Setting $\ell'(\lambda) = 0$:

$$-n + \frac{\sum x_i}{\lambda} = 0 \quad \Longrightarrow \quad \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}.$$

Second derivative:

$$\ell''(\lambda) = -\frac{\sum x_i}{\lambda^2} < 0,$$

so this critical point is indeed a maximum.

Step 4: Parameter of interest. We are asked to find the MLE of

$$P_\lambda(X=0).$$

For a Poisson distribution,

$$P_\lambda(X=0) = e^{-\lambda}.$$

Thus the parameter of interest is

$$\theta = g(\lambda) = e^{-\lambda}.$$

Step 5: Apply invariance property of the MLE. Since $\hat{\lambda} = \bar{X}$ is the MLE of λ, and $g(\lambda) = e^{-\lambda}$ is continuous, the invariance property gives:

$$\widehat{P(X=0)} = g(\hat{\lambda}) = e^{-\hat{\lambda}} = e^{-\bar{X}}.$$

$$\boxed{\widehat{P(X=0)} = e^{-\bar{X}}.}$$

4.3.2 Consistency of the MLE

Let $X_1, \ldots, X_n$ be i.i.d. from a distribution with pdf/pmf $f(x;\theta)$. Let θ_0 denote the true (unknown) parameter value. Define the log-likelihood function

$$\ell_n(\theta) = \sum_{i=1}^{n} \log f(X_i;\theta),$$

and the MLE

$$\hat{\theta}_n = \arg\max_{\theta} \ell_n(\theta).$$

The subscript n emphasizes the dependence of the estimator on the sample size.

Theorem. Under appropriate regularity (smoothness and identifiability) conditions on $f(x;\theta)$,

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \qquad \text{as } n \to \infty.$$

Proof. Consider the average log-likelihood:

$$\frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f(X_i;\theta).$$

By the Weak Law of Large Numbers (WLLN), for each fixed θ,

$$\frac{1}{n}\ell_n(\theta) \xrightarrow{p} \mathbb{E}_{\theta_0}[\log f(X_1;\theta)] \equiv Q(\theta), \qquad \text{as } n \to \infty.$$

Thus, the sample average log-likelihood converges in probability to the population quantity

$$Q(\theta) = \mathbb{E}_{\theta_0}[\log f(X;\theta)].$$

Now observe that

$$Q(\theta) - Q(\theta_0) = \mathbb{E}_{\theta_0}\left[\log \frac{f(X;\theta)}{f(X;\theta_0)}\right] = -D_{\mathrm{KL}}\big(f_{\theta_0} \,\|\, f_{\theta}\big) \le 0,$$

with equality if and only if $\theta = \theta_0$ (by identifiability). Therefore, the function $Q(\theta)$ is uniquely maximized at the true parameter value θ_0. Since

$$\frac{1}{n}\ell_n(\theta) \xrightarrow{p} Q(\theta),$$

the maximizer of the left-hand side, $\hat{\theta}_n$, converges in probability to the maximizer of the right-hand side, namely θ_0. Hence,

$$\hat{\theta}_n \xrightarrow{p} \theta_0.$$

4.3.3 Fisher information

For one observation $X \sim f(x;\theta)$, the Fisher information is

$$I_1(\theta) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right).$$

For n i.i.d. observations,

$$I(\theta) = nI_1(\theta).$$

Equivalent form. Under regularity conditions (interchange derivative and integral),

$$I_1(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

Proof. Let $s(X;\theta) = \partial_\theta \log f(X;\theta)$ be the score for one observation. Since $\int f(x;\theta)\,dx = 1$,

$$0 = \frac{\partial}{\partial\theta}\int f(x;\theta)\,dx = \int \partial_\theta f(x;\theta)\,dx = \int f(x;\theta)\,\partial_\theta \log f(x;\theta)\,dx = \mathbb{E}_\theta[s(X;\theta)].$$

Differentiate again:

$$0 = \frac{\partial}{\partial\theta}\mathbb{E}_\theta[s] = \mathbb{E}_\theta[\partial_\theta s] + \mathbb{E}_\theta[s^2] = \mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right] + \mathbb{E}_\theta[s^2].$$

Thus $\mathbb{E}_\theta[s^2] = -\mathbb{E}_\theta[\partial_\theta^2 \log f(X;\theta)]$. Since $\mathbb{E}_\theta[s] = 0$, $\mathrm{Var}_\theta(s) = \mathbb{E}_\theta[s^2]$, proving the equivalence.

4.3.4 Asymptotic normality of the MLE

Theorem 2. *Let $X_1, \ldots, X_n$ be i.i.d. from $f(x;\theta)$ with true parameter θ_0. Assume θ_0 is an interior point of the parameter space and standard regularity conditions hold. Then,*

$$\sqrt{n}\big(\hat\theta_n - \theta_0\big) \xrightarrow{d} N\big(0,\ I_1(\theta_0)^{-1}\big)\,,$$

where $I_1(\theta)$ is the Fisher information for one observation. Equivalently,

$$\hat\theta_n \approx N\left(\theta_0,\ \frac{1}{n}I_1(\theta_0)^{-1}\right) = N\big(\theta_0,\ I(\theta_0)^{-1}\big) \quad \textit{for large } n,$$

since $I(\theta_0) = nI_1(\theta_0)$.

Proof. The MLE satisfies the score equation

$$U_n(\hat\theta_n) = \ell_n'(\hat\theta_n) = 0.$$

Apply a first-order Taylor expansion of $U_n(\hat\theta_n)$ around θ_0:

$$0 = U_n(\theta_0) + (\hat\theta_n - \theta_0)\,\ell_n''(\tilde\theta_n),$$

for some $\tilde\theta_n$ between $\hat\theta_n$ and θ_0. Rearranging,

$$\sqrt{n}(\hat\theta_n - \theta_0) = -\frac{\frac{1}{\sqrt{n}}U_n(\theta_0)}{\frac{1}{n}\ell_n''(\tilde\theta_n)}.$$

Now analyze numerator and denominator separately.

Step 1: Numerator (CLT).

$$\frac{1}{\sqrt{n}}U_n(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i;\theta_0).$$

Since the score has mean zero and variance $I_1(\theta_0)$, the Central Limit Theorem gives

$$\frac{1}{\sqrt{n}}U_n(\theta_0) \xrightarrow{d} N(0,\ I_1(\theta_0))\,.$$

Step 2: Denominator (LLN + consistency). By the Law of Large Numbers,

$$-\frac{1}{n}\ell_n''(\theta_0) \xrightarrow{p} I_1(\theta_0).$$

Since $\hat{\theta}_n \xrightarrow{p} \theta_0$ (consistency), we also have

$$-\frac{1}{n}\ell_n''(\tilde{\theta}_n) \xrightarrow{p} I_1(\theta_0).$$

Step 3: Slutsky's theorem. Combining the CLT result for the numerator and convergence in probability of the denominator,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\big(0,\ I_1(\theta_0)^{-1}\big)\,.$$

Thus, for large n,

$$\hat{\theta}_n \approx N\left(\theta_0,\ \frac{1}{nI_1(\theta_0)}\right).$$

Implications. The result

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\big(0,\ I_1(\theta_0)^{-1}\big)$$

means the following for large n:

- The distribution of $\hat{\theta}_n$ is approximately normal.
- The mean of $\hat{\theta}_n$ is approximately (but not necessarily exactly) equal to θ_0.
- The variance of $\hat{\theta}_n$ is approximately

$$\mathrm{Var}(\hat{\theta}_n) \approx \frac{1}{nI_1(\theta_0)}.$$

- The variance decreases as n increases (precision improves).
- Since θ_0 is unknown, $I_1(\hat{\theta}_n)$ is often a good approximation to $I_1(\theta_0)$.

Why this result is fundamental. This asymptotic normality result is the foundation of

- Wald tests,
- Likelihood ratio tests,
- Confidence intervals,
- Large-sample hypothesis testing,
- Standard error calculations.

Almost all parametric inference in classical statistics is built on this theorem.

Example

Bernoulli(p) — MLE, Fisher information, asymptotic distribution.
Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$ with

$$f(x;p) = p^x(1-p)^{1-x}, \qquad x \in \{0,1\},\ 0 < p < 1.$$

(i) MLE of p. The likelihood is

$$L(p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum X_i}(1-p)^{n-\sum X_i}.$$

Log-likelihood:

$$\ell(p) = \Big(\sum_{i=1}^{n} X_i\Big)\log p + \Big(n - \sum_{i=1}^{n} X_i\Big)\log(1-p).$$

Differentiate and set to zero:

$$\ell'(p) = \frac{\sum X_i}{p} - \frac{n - \sum X_i}{1-p} = 0 \quad \Longrightarrow \quad \hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

Second derivative:

$$\ell''(p) = -\frac{\sum X_i}{p^2} - \frac{n - \sum X_i}{(1-p)^2} < 0,$$

so $\hat{p}$ is a maximizer.

(ii) Fisher information. For one observation $X \sim \text{Bernoulli}(p)$,

$$\log f(X;p) = X\log p + (1-X)\log(1-p).$$

The score for one observation is

$$\ell_1'(p) = \frac{\partial}{\partial p}\log f(X;p) = \frac{X}{p} - \frac{1-X}{1-p}.$$

Method 1: $I_1(p) = \text{Var}\big(\ell_1'(p)\big)$. First note that $\mathbb{E}[\ell_1'(p)] = 0$ (regularity condition), so

$$I_1(p) = \text{Var}(\ell_1'(p)) = \mathbb{E}\big[(\ell_1'(p))^2\big].$$

Evaluate by conditioning on $X \in \{0,1\}$:

$$\ell_1'(p) = \begin{cases} \frac{1}{p}, & X = 1, \\ -\frac{1}{1-p}, & X = 0. \end{cases}$$

Thus

$$I_1(p) = \mathbb{E}\big[(\ell_1'(p))^2\big] = P(X=1)\left(\frac{1}{p}\right)^2 + P(X=0)\left(\frac{1}{1-p}\right)^2 = p\cdot\frac{1}{p^2} + (1-p)\cdot\frac{1}{(1-p)^2}.$$

Hence

$$\boxed{I_1(p) = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.}$$

Method 2: $I_1(p) = -\mathbb{E}\big(\ell_1''(p)\big)$.
Differentiate the score:

$$\ell_1''(p) = \frac{\partial^2}{\partial p^2} \log f(X;p) = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2}.$$

Take expectation:

$$\mathbb{E}[\ell_1''(p)] = -\frac{\mathbb{E}[X]}{p^2} - \frac{\mathbb{E}[1-X]}{(1-p)^2} = -\frac{p}{p^2} - \frac{1-p}{(1-p)^2} = -\left(\frac{1}{p} + \frac{1}{1-p}\right).$$

Therefore,

$$\boxed{I_1(p) = -\mathbb{E}[\ell_1''(p)] = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.}$$

Finally, for n i.i.d. observations,

$$\boxed{I(p) = nI_1(p) = \frac{n}{p(1-p)}.}$$

(iii) Asymptotic distribution of $\hat{p}$. By asymptotic normality of the MLE,

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N\big(0,\ I_1(p)^{-1}\big) = N(0,\ p(1-p))\,,$$

equivalently,

$$\hat{p} \approx N\left(p,\ \frac{p(1-p)}{n}\right).$$

Example

Geometric — $f(x;\theta) = \theta x^{\theta-1}$ on $(0,1)$. Let $X_1, \ldots, X_n \overset{iid}{\sim} f(x;\theta)$ where

$$f(x;\theta) = \theta x^{\theta-1}, \qquad 0 < x < 1,\ \theta > 1.$$

(i) MLE of θ. The likelihood is

$$L(\theta) = \prod_{i=1}^{n} \theta X_i^{\theta-1} = \theta^n \prod_{i=1}^{n} X_i^{\theta-1}.$$

Log-likelihood:

$$\ell(\theta) = n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log X_i.$$

Differentiate and set to zero:

$$\ell'(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log X_i = 0 \quad \Longrightarrow \quad \hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log X_i}.$$

Second derivative:

$$\ell''(\theta) = -\frac{n}{\theta^2} < 0,$$

so $\hat{\theta}$ is a maximizer.

(ii) Fisher information. For one observation,

$$\log f(X;\theta) = \log\theta + (\theta - 1)\log X.$$

Then

$$\frac{\partial}{\partial\theta}\log f(X;\theta) = \frac{1}{\theta} + \log X, \qquad \frac{\partial^2}{\partial\theta^2}\log f(X;\theta) = -\frac{1}{\theta^2}.$$

Hence using $I_1(\theta) = -\mathbb{E}[\partial_\theta^2 \log f(X;\theta)]$,

$$I_1(\theta) = \frac{1}{\theta^2}, \qquad I(\theta) = nI_1(\theta) = \frac{n}{\theta^2}.$$

(iii) Asymptotic distribution of $\hat{\theta}$. By asymptotic normality of the MLE,

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\big(0,\ I_1(\theta)^{-1}\big) = N\big(0,\ \theta^2\big)\,,$$

equivalently,

$$\hat{\theta} \approx N\left(\theta,\ \frac{\theta^2}{n}\right).$$

Confidence intervals using asymptotic normality of the MLE. From asymptotic normality,

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{d} N\big(0,\ I_1(\theta_0)^{-1}\big)\,.$$

Multiplying both sides by $\sqrt{I_1(\theta_0)}$,

$$\sqrt{nI_1(\theta_0)}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{d} N(0,1).$$

Since $I(\theta_0) = nI_1(\theta_0)$, this can also be written as

$$\sqrt{I(\theta_0)}\big(\hat{\theta}_n - \theta_0\big) \xrightarrow{d} N(0,1).$$

For large n, we therefore approximate

$$\sqrt{nI_1(\theta_0)}(\hat{\theta}_n - \theta_0) \approx N(0,1).$$

Let $z_{1-\alpha/2}$ be the $(1-\alpha/2)$ quantile of $N(0,1)$. Then for large n,

$$\mathbb{P}\Big(-z_{1-\alpha/2} \le \sqrt{nI_1(\theta_0)}(\hat{\theta}_n - \theta_0) \le z_{1-\alpha/2}\Big) \approx 1-\alpha.$$

Dividing through by $\sqrt{nI_1(\theta_0)}$ gives

$$\mathbb{P}\left(\hat{\theta}_n - \frac{z_{1-\alpha/2}}{\sqrt{nI_1(\theta_0)}} \le \theta_0 \le \hat{\theta}_n + \frac{z_{1-\alpha/2}}{\sqrt{nI_1(\theta_0)}}\right) \approx 1-\alpha.$$

Hence, an approximate $(1-\alpha)100\%$ confidence interval for θ_0 is

$$\boxed{\hat{\theta}_n \pm \frac{z_{1-\alpha/2}}{\sqrt{nI_1(\theta_0)}}.}$$

Wald interval. Since θ_0 is unknown, we replace $I_1(\theta_0)$ by $I_1(\hat{\theta}_n)$:

$$\widehat{\mathrm{SE}}(\hat{\theta}_n) = \sqrt{\frac{1}{nI_1(\hat{\theta}_n)}}.$$

This yields the practical Wald confidence interval

$$\boxed{\hat{\theta}_n \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}(\hat{\theta}_n).}$$

Example: Poisson(λ), $n = 23$, $\bar{x} = 24.9$. For Poisson(λ),

$$I_1(\lambda) = \frac{1}{\lambda}.$$

Thus,

$$\mathrm{Var}(\hat{\lambda}) \approx \frac{\lambda}{n}, \qquad \widehat{\mathrm{SE}}(\hat{\lambda}) = \sqrt{\frac{\hat{\lambda}}{n}}.$$

With $\hat{\lambda} = \bar{x} = 24.9$ and $n = 23$,

$$\widehat{\mathrm{SE}}(\hat{\lambda}) = \sqrt{\frac{24.9}{23}} \approx 1.04.$$

For a 95% CI, $z_{0.975} = 1.96$, so

$$\lambda \approx 24.9 \pm 1.96(1.04) = 24.9 \pm 2.04.$$

Therefore the approximate 95% CI is

$$\boxed{(22.86,\ 26.94).}$$

4.3.5 Asymptotic efficiency, Cramér–Rao bound, and comparison of estimators

Cramér–Rao Lower Bound (CRLB). Let $X_1, \ldots, X_n$ be i.i.d. from $f(x;\theta)$ and let T be an unbiased estimator of θ. Under regularity conditions,

$$\mathrm{Var}(T) \geq \frac{1}{nI_1(\theta)},$$

where $I_1(\theta)$ is the Fisher information for one observation.

Proof. Define the score

$$S = \ell_n'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i;\theta).$$

Then,

$$\mathbb{E}[S] = 0, \qquad \mathrm{Var}(S) = nI_1(\theta).$$

One can show that for any unbiased estimator T of θ,

$$\mathrm{Cov}(S, T) = 1.$$

Applying the Cauchy–Schwarz inequality,

$$\mathrm{Cov}(S,T)^2 \le \mathrm{Var}(S)\mathrm{Var}(T),$$

gives

$$1^2 \le nI_1(\theta)\mathrm{Var}(T),$$

which implies the lower bound

$$\boxed{\mathrm{Var}(T) \ge \frac{1}{nI_1(\theta)}.}$$

Efficiency. An unbiased estimator whose variance attains the CR bound is called **efficient**.

- CRLB gives the smallest possible variance among unbiased estimators.
- An estimator achieving the bound is called efficient.
- Efficient estimators do not necessarily exist.
- MLE is said to be asymptotically efficient.
- MLE is not necessarily efficient for a given sample size.

Asymptotic efficiency of the MLE. From asymptotic normality,

$$\mathrm{Var}(\hat{\theta}_n) \approx \frac{1}{nI_1(\theta_0)}.$$

Thus, the MLE achieves the CR lower bound asymptotically:

$$\boxed{\text{The MLE is asymptotically efficient.}}$$

However,

- The MLE is not necessarily unbiased for finite n.
- It may not achieve the CR bound for small samples.
- Normal approximations may be poor in small samples or near parameter boundaries.

Comparing estimators: efficiency. If two estimators T_1 and T_2 are unbiased for θ, the relative efficiency of T_1 to T_2 is

$$eff(T_1,T_2) = \frac{1/\mathrm{Var}(T_2)}{1/\mathrm{Var}(T_1)} = \frac{\mathrm{Precision}(T_1)}{\mathrm{Precision}(T_2)}.$$

- $eff(T_1,T_2) < 1$: T_1 is less precise than T_2.
- $eff(T_1,T_2) = 1$: equal precision.
- $eff(T_1,T_2) > 1$: T_1 is more precise.

UMVUE (Uniform Minimum Variance Unbiased Estimator). The **UMVUE**, also called the *best unbiased estimator*, is an estimator that

- is unbiased for θ,
- has the smallest variance among *all* unbiased estimators of θ.

Thus, if T^* is the UMVUE and T is any other unbiased estimator of θ, then

$$\mathrm{Var}(T) \geq \mathrm{Var}(T^*).$$

Relationship with efficiency and CR bound.

- If an unbiased estimator attains the Cramér–Rao lower bound,

  $$\mathrm{Var}(T) = \frac{1}{nI_1(\theta)},$$

  then it is called **efficient**.
- An efficient estimator (if it exists) is automatically the UMVUE.
- However, a UMVUE does *not* have to be efficient (i.e., it may not attain the CR bound).
- Efficient estimators do not necessarily exist for all models.
- UMVUEs generally exist under standard regularity conditions, especially when a complete sufficient statistic exists.

How to find a UMVUE? Typically, one proceeds as follows:

- Find a sufficient statistic T for θ.
- Show that T is complete.
- Use the Lehmann–Scheffé theorem: any unbiased estimator that is a function of a complete sufficient statistic is the UMVUE.
- Alternatively, if an unbiased estimator attains the CR bound, it is efficient and hence the UMVUE.

Example: Bernoulli(p). Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The estimator

$$\hat{p} = \bar{X}$$

is unbiased and

$$\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

Since $I_1(p) = \frac{1}{p(1-p)}$,

$$\frac{1}{nI_1(p)} = \frac{p(1-p)}{n}.$$

Thus, $\bar{X}$ attains the CR bound and is therefore efficient and UMVUE.

Beyond efficiency: Mean Squared Error (MSE). When comparing estimators that may be biased, variance alone is insufficient. The Mean Squared Error (MSE) of an estimator T of θ is

$$\mathrm{MSE}(T) = \mathbb{E}\big[(T-\theta_0)^2\big] = \mathrm{Var}(T) + \mathrm{Bias}(T)^2.$$

It is possible for a biased estimator to have smaller MSE than an unbiased estimator. Thus, in practice, MSE is often a more meaningful measure of performance than variance alone.

5 Sufficiency and the Factorization Theorem

Let $X_1, \ldots, X_n$ be a random sample from a family of distributions with joint pdf/pmf

$$f(x_1, \ldots, x_n; \theta),$$

where θ (scalar or vector) is an unknown parameter. A central question in inference is: *"Is there a statistic $T = T(X_1, \ldots, X_n)$ that contains all information about θ needed for inference?"* If such a statistic exists, we can base inference on T without losing any information about θ.

5.1 Definition of Sufficiency

Theorem

Definition of a "Sufficient statistic". A statistic $T = T(X_1, \ldots, X_n)$ is **sufficient** for θ if the conditional distribution of the full sample $(X_1, \ldots, X_n)$ given $T = t$ does not depend on θ for any t in the support of T.

If the conditional distribution of the sample given T is free of θ, then once we know T, the remaining variation in the data carries no additional information about θ. Hence, no information about θ is lost by reducing the full data to T.

Note. Even if θ is scalar, a sufficient statistic may be vector-valued (and vice versa).

5.2 Factorization Theorem (Neyman–Fisher)

Theorem

Factorization Theorem.
A statistic $T = T(X_1, \ldots, X_n)$ is sufficient for θ if and only if the joint pdf/pmf can be written as

$$f(x_1, \ldots, x_n; \theta) = g\big(T(x_1, \ldots, x_n); \theta\big)\, h(x_1, \ldots, x_n),$$

where

- g depends on the data only through T and on θ,
- h does not depend on θ.

Proof (detailed)

($\Rightarrow$) Factorization implies sufficiency.
Suppose

$$f(\boldsymbol{x}; \theta) = g(T(\boldsymbol{x}); \theta) h(\boldsymbol{x}).$$

For $T(\boldsymbol{x}) = t$,

$$\mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T = t) = \frac{f(\boldsymbol{x};\theta)}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=t} f(\boldsymbol{y};\theta)}.$$

Substitute the factorization:

$$= \frac{g(t;\theta)h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=t} g(t;\theta)h(\boldsymbol{y})}.$$

Since $g(t;\theta)$ cancels,

$$= \frac{h(\boldsymbol{x})}{\sum_{\boldsymbol{y}:T(\boldsymbol{y})=t} h(\boldsymbol{y})},$$

which does not depend on θ. Thus T is sufficient.

($\Leftarrow$) Sufficiency implies factorization.
If T is sufficient,

$$f(\boldsymbol{x};\theta) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T = T(\boldsymbol{x})) \cdot \mathbb{P}_\theta(T = T(\boldsymbol{x})).$$

Define

$$h(\boldsymbol{x}) = \mathbb{P}_\theta(\boldsymbol{X} = \boldsymbol{x} \mid T = T(\boldsymbol{x})), \quad g(t;\theta) = \mathbb{P}_\theta(T = t).$$

Then

$$f(\boldsymbol{x};\theta) = g(T(\boldsymbol{x});\theta)h(\boldsymbol{x}),$$

which gives the desired factorization.

5.3 Consequence: MLE is a Function of a Sufficient Statistic

Theorem

If T is sufficient for θ, then any MLE $\hat{\theta}$ is a function of T.

Proof (detailed)

If

$$L(\theta;\boldsymbol{x}) = g(T(\boldsymbol{x});\theta)h(\boldsymbol{x}),$$

and h does not depend on θ, then

$$\arg\max_\theta L(\theta;\boldsymbol{x}) = \arg\max_\theta g(T(\boldsymbol{x});\theta).$$

Hence, the maximizer depends only on $T(\boldsymbol{x})$.

Example

Example 1: Bernoulli(θ).

Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(θ):

$$P(X = x; \theta) = \theta^x (1-\theta)^{1-x}, \quad x \in \{0, 1\}.$$

Step 1. Joint pmf.

$$f(\boldsymbol{x}; \theta) = \prod_{i=1}^{n} \theta^{x_i} (1-\theta)^{1-x_i}.$$

$$= \theta^{\sum_{i=1}^n x_i} (1-\theta)^{\sum_{i=1}^n (1-x_i)} = \theta^{\sum x_i} (1-\theta)^{n - \sum x_i}.$$

Step 2. Factorization. All dependence on θ occurs through

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

Thus, we can write

$$f(\boldsymbol{x}; \theta) = \underbrace{\theta^{T(\boldsymbol{x})} (1-\theta)^{n - T(\boldsymbol{x})}}_{g(T(\boldsymbol{x}); \theta)} \cdot \underbrace{1}_{h(\boldsymbol{x})}.$$

Since the joint density factors through T, by the Factorization Theorem,

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

is sufficient for θ.

Step 3. MLE.

Since the likelihood is proportional to

$$\theta^T (1-\theta)^{n-T},$$

maximizing gives

$$\hat{\theta} = \frac{T}{n} = \bar{X},$$

which is a function of the sufficient statistic.

Example

Example 2: Laplace(ρ). Let $X_1, \ldots, X_n$ be i.i.d. with density

$$f(x; \rho) = \frac{1}{2\rho} \exp\left(-\frac{|x|}{\rho} \right), \qquad \rho > 0.$$

Step 1. Joint density.

$$f(\boldsymbol{x}; \rho) = \prod_{i=1}^{n} \frac{1}{2\rho} \exp\left(-\frac{|x_i|}{\rho} \right) = (2\rho)^{-n} \exp\left(-\frac{1}{\rho} \sum_{i=1}^{n} |x_i| \right).$$

Step 2. Factorization. All ρ-dependence occurs through

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

Thus,

$$f(\boldsymbol{x};\rho) = \underbrace{(2\rho)^{-n}\exp\left(-\frac{T(\boldsymbol{x})}{\rho}\right)}_{g(T(\boldsymbol{x});\rho)} \cdot \underbrace{1}_{h(\boldsymbol{x})}.$$

Hence, by the Factorization Theorem,

$$T = \sum_{i=1}^{n} |X_i|$$

is sufficient for ρ.

Step 3. MLE.

Log-likelihood:

$$\ell(\rho) = -n\log(2\rho) - \frac{T}{\rho}.$$

Differentiate:

$$\ell'(\rho) = -\frac{n}{\rho} + \frac{T}{\rho^2}.$$

Setting $\ell'(\rho) = 0$ gives

$$\hat{\rho} = \frac{T}{n}.$$

Again, the MLE depends only on the sufficient statistic.

Example

Example 3: Normal $N(\mu, \sigma^2)$. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$:

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Step 1. Joint density.

$$f(\boldsymbol{x};\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Step 2. Expand quadratic term.

$$\sum_{i=1}^{n}(x_i-\mu)^2 = \sum x_i^2 - 2\mu\sum x_i + n\mu^2.$$

Step 3. Factorization. All parameter dependence appears through

$$T_1 = \sum_{i=1}^{n} X_i, \qquad T_2 = \sum_{i=1}^{n} X_i^2.$$

Hence, the density can be written as

$$f(\boldsymbol{x};\mu,\sigma^2) = \underbrace{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}[T_2 - 2\mu T_1 + n\mu^2]\right)}_{g(T_1,T_2;\mu,\sigma^2)} \cdot \underbrace{1}_{h(\boldsymbol{x})}.$$

Therefore,

$$(T_1, T_2) = \left(\sum X_i, \sum X_i^2\right)$$

is sufficient for (μ, σ^2).

Step 4. Equivalent sufficient statistics.

Since

$$\bar{X} = \frac{1}{n}\sum X_i$$

and

$$\sum (X_i - \bar{X})^2 = \sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2,$$

we may equivalently use

$$\left(\bar{X}, \sum (X_i - \bar{X})^2\right)$$

as a sufficient statistic for (μ, σ^2).

5.4 Exponential Families and Sufficiency

5.4.1 One-Parameter Exponential Family

Theorem

A model $\{f(x;\theta) : \theta \in \Theta\}$ is in the **one-parameter exponential family** if its pmf/pdf can be written as

$$f(x;\theta) = \exp\Big\{w(\theta)T(x) - b(\theta) + c(x)\Big\},$$

where the support of x does not depend on θ. The idea is a separability. The only term involving *both* data and parameter is $w(\theta)T(x)$. This "separates" data-only and parameter-only parts.

- $T(x)$ and $c(x)$ depend only on the data x.
- $w(\theta)$ and $b(\theta)$ depend only on the parameter θ.
- $b(\theta)$ is the **normalizing function**: the factor $\exp\{-b(\theta)\}$ acts as the normalizing constant that ensures $\int f(x;\theta)\,dx = 1$ (or $\sum_x f(x;\theta) = 1$ in discrete cases).
- If $w(\theta) = \theta$, then θ is called the **natural (canonical) parameter**. In the natural parameterization,

$$f(x;\theta) = \exp\{\theta T(x) - b(\theta) + c(x)\}.$$

Theorem

(**Sufficiency in one-parameter exponential families**) If $X_1, \ldots, X_n$ are i.i.d. from a one-parameter exponential family, then

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

is sufficient for θ.

Proof (detailed)

For i.i.d. sampling,

$$f(\boldsymbol{x};\theta) = \prod_{i=1}^{n} \exp\{w(\theta)T(x_i) - b(\theta) + c(x_i)\}.$$

Combine terms:

$$= \exp\Big\{w(\theta)\sum_{i=1}^{n} T(x_i) - nb(\theta) + \sum_{i=1}^{n} c(x_i)\Big\}.$$

This has the factorized form

$$f(\boldsymbol{x};\theta) = \underbrace{\exp\Big\{w(\theta)\sum T(x_i) - nb(\theta)\Big\}}_{g(\sum T(x_i);\theta)} \cdot \underbrace{\exp\Big\{\sum c(x_i)\Big\}}_{h(\boldsymbol{x})},$$

so by the Factorization Theorem, $\sum_{i=1}^{n} T(X_i)$ is sufficient for θ.

5.4.2 K-Parameter Exponential Family

Theorem

A model is in the **K-parameter exponential family** if its pmf/pdf can be written as

$$f(x;\boldsymbol{\theta}) = \exp\Big\{\sum_{k=1}^{K} w_k(\boldsymbol{\theta})\, T_k(x) - b(\boldsymbol{\theta}) + c(x)\Big\},$$

where the support does not depend on $\boldsymbol{\theta}$.

- The only data–parameter interaction appears through $\sum_{k=1}^{K} w_k(\boldsymbol{\theta})T_k(x)$.
- The vector $(T_1(x), \ldots, T_K(x))$ plays the role of the data-reduction summary.
- If $w_k(\boldsymbol{\theta}) = \theta_k$ for all k, then $\boldsymbol{\theta}$ is a **natural/canonical parameter vector**.

Theorem

(**Sufficiency in K-parameter exponential families**) If $X_1, \ldots, X_n$ are i.i.d. from a K-parameter exponential family, then the vector

$$\Big(\sum_{i=1}^{n} T_1(X_i),\ \sum_{i=1}^{n} T_2(X_i),\ \ldots,\ \sum_{i=1}^{n} T_K(X_i)\Big)$$

is sufficient for $\boldsymbol{\theta}$.

Proof (detailed)

Multiply the i.i.d. densities and collect terms:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{i=1}^{n} \exp\Big\{\sum_{k=1}^{K} w_k(\boldsymbol{\theta})T_k(x_i) - b(\boldsymbol{\theta}) + c(x_i)\Big\}$$

$$= \exp\Big\{\sum_{k=1}^{K} w_k(\boldsymbol{\theta}) \sum_{i=1}^{n} T_k(x_i) - nb(\boldsymbol{\theta}) + \sum_{i=1}^{n} c(x_i)\Big\}.$$

Hence, the dependence on $\boldsymbol{\theta}$ is only through the sums $\sum_i T_k(x_i)$, so the stated vector is sufficient.

5.5 Rao–Blackwell Theorem

The following theorem argues that when a sufficient statistic is available, one should construct an estimator of a parameter, using this sufficient statistic.

Theorem

Let $g(\boldsymbol{X})$ be an estimator of a parameter θ such that $E_\theta[g(\boldsymbol{X})^2] < \infty$ for all θ. Suppose $T(\boldsymbol{X})$ is sufficient for θ, and define the **Rao–Blackwellized estimator**

$$\tilde{g}(\boldsymbol{X}) \;=\; E_\theta[g(\boldsymbol{X}) \mid T(\boldsymbol{X})].$$

Then, for all θ,

$$E_\theta\big(\tilde{g}(\boldsymbol{X}) - \theta\big)^2 \;\le\; E_\theta\big(g(\boldsymbol{X}) - \theta\big)^2,$$

and the inequality is strict unless $\tilde{g}(\boldsymbol{X}) = g(\boldsymbol{X})$ almost surely. Moreover, $\tilde{g}$ is a function of the sufficient statistic T.

Theorem

Let $g(\boldsymbol{X})$ be an estimator of a parameter θ such that $E_\theta[g(\boldsymbol{X})^2] < \infty$ for all θ. Suppose that $T = T(\boldsymbol{X})$ is a sufficient statistic for θ, and define the **Rao–Blackwellized estimator**

$$\tilde{g}(\boldsymbol{X}) = E_\theta[g(\boldsymbol{X}) \mid T(\boldsymbol{X})].$$

Then for all θ,

$$E_\theta\big[(\tilde{g}(\boldsymbol{X}) - \theta)^2\big] \;\le\; E_\theta\big[(g(\boldsymbol{X}) - \theta)^2\big],$$

with strict inequality unless

$$\tilde{g}(\boldsymbol{X}) = g(\boldsymbol{X}) \quad \text{(almost surely)}.$$

Moreover, $\tilde{g}(\boldsymbol{X})$ is a function of the sufficient statistic T.

Equivalent formulation. Let $\hat{\theta} = \hat{\theta}(\boldsymbol{X})$ be an estimator of θ with $E_\theta[\hat{\theta}^2] < \infty$. If $T = T(\boldsymbol{X})$ is sufficient for θ and we define

$$\tilde{\theta} = E_\theta[\hat{\theta} \mid T],$$

then for all θ,

$$E_\theta\big[(\tilde{\theta} - \theta)^2\big] \;\le\; E_\theta\big[(\hat{\theta} - \theta)^2\big],$$

with strict inequality unless

$$\tilde{\theta} = \hat{\theta} \quad \text{(almost surely)}.$$

Thus, the Rao–Blackwellized estimator

$$\tilde{\theta} = E[\hat{\theta} \mid T]$$

is a function of the sufficient statistic T and has MSE no larger than that of $\hat{\theta}$.

What does Rao–Blackwell imply?

- If we already have an estimator $g(\boldsymbol{X})$ that is *not* a function of a sufficient statistic, then conditioning on T produces an **improved estimator** (no worse MSE, usually strictly better).
- The theorem does *not* automatically give the globally best estimator among all possible estimators. It says: **given any estimator**, you can often improve it using sufficiency.
- Therefore, if **MSE is the evaluation criterion**, it is natural to restrict attention to estimators that are functions of sufficient statistics.

Proof (detailed)

(Proof idea.) Use the law of total expectation and total variance (or conditional Jensen):

$$E_\theta[(g-\theta)^2] = E_\theta\big[\,E_\theta[(g-\theta)^2 \mid T]\,\big] \geq E_\theta\big[\,(E_\theta[g \mid T]-\theta)^2\,\big] = E_\theta[(\tilde{g}-\theta)^2].$$

Strictness holds unless g is already a function of T a.s.

Example

Example 1: Poisson(λ). Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$.

Step 1. Identify a sufficient statistic. The joint pmf is

$$f(\boldsymbol{x};\lambda) = \exp\left\{-n\lambda + \left(\sum_{i=1}^n x_i\right)\log\lambda - \sum \log(x_i!)\right\}.$$

By the factorization theorem,

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

is sufficient for λ.

Step 2. Start with any estimator. Take

$$\hat{\lambda}_1 = X_1.$$

This is unbiased since

$$E[X_1] = \lambda.$$

Notice: $\hat{\lambda}_1$ is *not* a function of the sufficient statistic T.

Step 3. Apply Rao–Blackwell. Define the improved estimator

$$\tilde{\lambda} = E[\hat{\lambda}_1 \mid T] = E[X_1 \mid T].$$

Step 4. Compute the conditional expectation. A key fact:

$$X_1 \mid T = t \sim \text{Binomial}\left(t, \frac{1}{n}\right).$$

Why? Because given the total number of events t, each of the t events is equally likely to belong to any of the n observations. Thus,

$$E[X_1 \mid T = t] = \frac{t}{n}.$$

Therefore,

$$\tilde{\lambda} = \frac{T}{n} = \bar{X}.$$

Conclusion. By Rao–Blackwell,

$$\mathrm{MSE}(\bar{X}) \leq \mathrm{MSE}(X_1).$$

So the sample mean is obtained directly by conditioning on the sufficient statistic.

Example

Improving an estimator of $p(\lambda) = P(X = 0) = e^{-\lambda}$.

Step 1. Start with an unbiased estimator. Since

$$P(X_1 = 0) = e^{-\lambda},$$

an unbiased estimator is

$$\hat{p}_1 = \mathbf{1}\{X_1 = 0\}.$$

Again, this depends only on X_1, not on the sufficient statistic T.

Step 2. Rao–Blackwellize. Define

$$\tilde{p} = E[\mathbf{1}\{X_1 = 0\} \mid T].$$

Step 3. Compute conditional probability. Since

$$X_1 \mid T = t \sim \mathrm{Binomial}\left(t, \frac{1}{n}\right),$$

we have

$$P(X_1 = 0 \mid T = t) = \left(1 - \frac{1}{n}\right)^t.$$

Thus,

$$\tilde{p} = \left(1 - \frac{1}{n}\right)^T.$$

Conclusion. The improved estimator is a function of the sufficient statistic T and has smaller (or equal) MSE than $\mathbf{1}\{X_1 = 0\}$.

Example

Example 2: Uniform$[0, \theta]$ Let $X_1, \ldots, X_n \sim \mathrm{Unif}[0, \theta]$.

Step 1. Sufficient statistic.
The joint density factors through

$$T = X_{(n)} = \max X_i.$$

Step 2. Start with an unbiased estimator. Since $E[X_1] = \theta/2$,

$$\hat{\theta}_1 = 2X_1$$

is unbiased. But it is not a function of T.

Step 3. Rao–Blackwellize.

$$\tilde{\theta} = E[2X_1 \mid X_{(n)}].$$

Step 4. Compute conditional expectation. Given $X_{(n)} = t$,
- One observation equals t.
- The remaining $n-1$ are i.i.d. Unif$[0,t]$.
- By symmetry, $P(X_1 = t) = 1/n$.

Thus,

$$E[X_1 \mid X_{(n)} = t] = \frac{1}{n}t + \frac{n-1}{n} \cdot \frac{t}{2} = \frac{n+1}{2n}t.$$

Therefore,

$$\tilde{\theta} = 2E[X_1 \mid X_{(n)}] = \frac{n+1}{n}X_{(n)}.$$

Conclusion.
Rao–Blackwell automatically transforms $2X_1$ into a function of the sufficient statistic $X_{(n)}$ with smaller MSE.