IEDA 2540 Statistics for Engineers
Point Estimation

Wei YOU

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Spring, 2025

Introduction: Distributions for Modeling Common Random Experiments

In the last topic, we saw distributions for modeling common random experiments. These distributions serve as a reasonable simplification/model for the experiments.

- **Example:** Bernoulli distribution to characterize experiments with binary counts.
- **Example:** Poisson distribution to characterize integer counts.
- **Example:** Normal distribution to characterize continuous data arised from taking average or sum (CLT).

These distributions are usually specified by one or more parameters.

- **Example:** Bernoulli distribution is specified by the success probability p.
- **Example:** Poisson distribution is specified by the rate λ.
- **Example:** Normal distribution is specified by the mean μ and variance σ^2.

Introduction: Statistical Inference

- Assumes that the population follows a certain family of distributions (e.g. Bernoulli with some unknown p).
- Focuses on drawing conclusions about the unknown parameters of the distribution.
- An important part of this process is obtaining **estimates** (a reasonable value) of the parameters.

Random Sample

Throughout this course, we shall assume that the data we collected forms a random sample.

Random sample

If $X_1, X_2, \ldots, X_n$ are independent random variables having a common distribution F, then we say that they constitute a random sample from the distribution F. We also say that $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.).

Why random sample? The i.i.d. property allows us to write the joint PMF/PDF as

$$f(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

We will see that such a product form simplifies many calculations.

Statistic

Statistic

A **statistic** is a random variable whose value is determined by the sample. In other words, it maps the collection of observations to some real number.

Statistics are constructed from

- anything computable from the data; and
- anything assumed to be known.

You **CANNOT** construct statistics using

- unknown parameters.

Statistics vs. Parameters

You need to be able to tell the exact value of a statistics, once the data is given.

Example: The sample mean $\bar{X}$ is a statistic. Once the data is given, you can calculate the value of $\bar{X}$. But $(X_1 + \mu)/2$ is not a statistic, because you don't know the value of μ.

	Value known?	Value random?	Example
Parameter	unknow	deterministic	Population mean μ, variance σ^2
Statistic	known	random	Sample mean $\bar{X}$, sample variance S^2

Point Estimator

Point Estimator

A statistic $\hat{\theta}$ intended to estimate an unknown quantity of interest (a parameter of the population) θ is called a **point estimator** of θ.

The hat notation

The point estimator used to estimate a parameter θ is usually denoted as $\hat{\theta}$.

Point Estimate

After the sample has been selected, $\hat{\theta}$ takes on a particular numerical value; this value is called the **point estimate** of θ.

In this topic, we shall see techniques to find reasonable point estimators of the population parameters.

Example: Common Point Estimators

Parameter	Point estimator
$\mu = \mathbb{E}[X]$	Sample mean $\bar{X}$
	Any observation X_i
	Use 0 as estimator (Is this reasonable?)
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$
$\sigma^2 = \mathrm{Var}(X)$	Sample variance S^2
	$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$
$\theta = \mathrm{Cov}(X, Y)$	$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
correlation coefficient ρ	$\frac{1}{n-1}\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{S_X S_Y}$

Sample Mean: Estimating the Mean

Suppose distribution F has mean μ. This is the most common parameter one wishes to estimate.

A natural estimator: the sample mean of a sample $X_1, X_2, \ldots, X_n$

$$\bar{X} = \frac{1}{n}\big(X_1 + X_2 + \cdots + X_n\big).$$

The sample mean is a statistic, hence it is a random variable!

Is this a reasonable estimator? More importantly, what makes a good estimator?

Visualizing the Distribution of the Sample Mean

We have the intuition that the sample mean should be close to the true population mean, especially when the sample size is large.

How to visualize?

- We want to understand *the distribution of the sample mean*.
- From descriptive statistics, we can plot the *histogram of the sample mean*.
- This requires us to take many i.i.d. observations of the sample mean.
    1. For a certain sample size n, we draw n i.i.d. samples from the original distribution F and calculate one observation of the sample mean.
    2. Repeat the first step for m times, then we obtain m i.i.d. samples of the sample mean. The total number of samples from the original distribution F is $n \times m$.

Table: $m = 100$ observations of sample mean $\bar{X}_4$ for sample size $n = 4$

Sample	x_1	x_2	x_3	x_4	Sample Mean $\bar{x}_4$
1	0.2146	0.6409	0.5702	1.2420	0.6669
2	-1.1453	-0.7257	0.5799	0.8147	-0.1191
3	2.5040	-0.6341	-0.7738	-1.5476	-0.1128
4	-0.2335	-2.0745	-0.2115	-1.3960	-0.9789
5	0.4958	0.6800	-0.0545	0.1364	0.3144
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1000	-0.1441	1.1047	0.9348	-0.1594	0.4340

- There are $m = 1000$ observations of the sample mean, i.e., the values in the last column of each table.
- We use the $m = 1000$ observations for the sample mean to plot the histogram.

Table: $m = 100$ observations of sample mean $\bar{X}_{50}$ for sample size $n = 50$

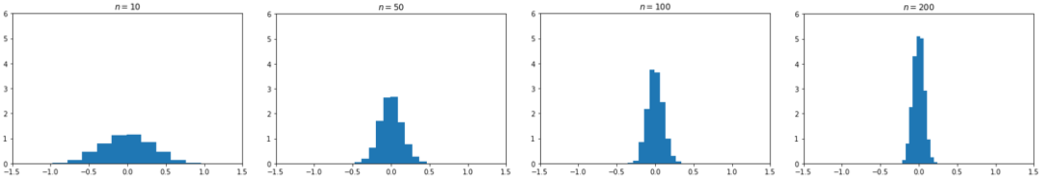
Group	x_1	x_2	$\cdots$	x_{50}	Sample Mean $\bar{x}_{50}$
1	-0.6197	0.3715	$\cdots$	-0.0018	-0.0352
2	-0.0436	-1.1562	$\cdots$	-0.3941	0.1770
3	-0.9820	-0.2195	$\cdots$	1.2833	-0.0760
4	1.6812	-1.1482	$\cdots$	0.0877	-0.1705
5	1.0325	0.2447	$\cdots$	0.8609	-0.0913
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
1000	0.3022	-1.2556	$\cdots$	-0.0183	0.0984

- There are $m = 1000$ observations of the sample mean, i.e., the values in the last column of each table.
- We use the $m = 1000$ observations for the sample mean to plot the histogram.

Histogram Plots for Different Sample Sizes

Consider $n = 10$, $n = 50$, $n = 100$, and $n = 200$.

Followings are the histogram plots of 1000 observations of the sample means (for a normal population) for each of the four choices of n.



Properties of Sample Mean

Expectation (center)

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n}\Big(\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]\Big) = \mu.$$

The sample mean is close to the mean of F.

Variance (dispersion)

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)\right] = \frac{1}{n^2}\Big(\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)\Big) = \frac{1}{n}\sigma^2.$$

This variance depends on sample size n. It decreases to 0 as n increases.

Key observations: The sample mean is centered around the true mean μ and becomes less and less random as the sample size increases.

Bias

Bias

Suppose $\hat{\theta}$ is used to estimate θ, then is the **bias** of $\hat{\theta}$ is defined as

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

- Bias measures the distance between the "center" of the (random) estimator to the true parameter. Intuitively, a small bias implies that if you estimate for multiple times, on average you will be closed to the true parameter.

Unbiased Estimator

$\hat{\theta}$ is **unbiased** if the expected value of the estimator is equal to the parameter it intends to estimate, i.e., if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$.

Is Unbiased Estimators Always Good?

Bias tells us how “on average” how close the estimator is to the true parameter.

Example: If we use a single observation X_1 as an estimator of the population mean.

- Is it unbiased?
- What is the variance of this estimator?
- How does the variance compare to that of the sample mean?

Bias says nothing about the “spread” of the estimator.

Example: Paris Olympic Shooting 10m Air Rifle Men

`http://dingyue.ws.126.net/2024/0904/9a08b039g00sjaklm0354d000qo00f0m.gif`

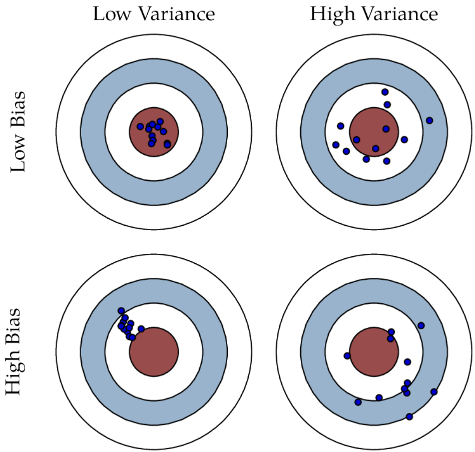
Even though achieving a 10.9 is technically a perfect shot, the player isn't entirely satisfied because this isolated result doesn't guarantee **consistent performance**.

- The inherent instability and variability in shooting means that while one shot might hit the perfect ring, the athlete **cannot confidently predict** similar outcomes in future rounds.
- Consistency is crucial in Olympic air rifle shooting, so a sporadic peak performance can actually undermine overall confidence and hinder the development of a reliable competitive edge.

Bias Variance Tradeoff

- Bias measures the distance between the center of the estimator and the true parameter.
- Variance measures the spread of the estimator.
- A good estimator should have both small bias and small variance.

How do we balance these two goals?



Mean Square Error

Mean Square Error (MSE)

Suppose $\hat{\theta}$ is used to estimate θ. Then

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

is the mean squared error of $\hat{\theta}$.

- Mean Squared Error estimates the average squared distance from the estimator to the parameter.
- MSE is NOT the same as $\mathrm{Var}(\hat{\theta}) = \mathbb{E}_{\theta}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$.
  - If $\hat{\theta}$ is unbiased, then $\mathrm{Var}(\hat{\theta}) = \mathrm{MSE}_{\theta}(\hat{\theta})$.
- MSE is NOT the same as $\mathrm{Var}(X) = \mathbb{E}_{\theta}[(X - \theta)^2]$.

Bias Variance Decomposition

Bias Variance Decomposition

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

- Mean Squared Error *balances between the two goals*: bias and variance.
- In comparing two estimators, we *prefer the one with lower MSE*, i.e., overall smaller bias and smaller variance.

Example: MSE of sample mean.

- $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
- $\mathrm{Bias}(\bar{X}) = 0$.
- $\mathrm{MSE}(\bar{X}) = \frac{\sigma^2}{n}$.

Bias Variance Decomposition

Proof:

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

Relative Efficiency

The mean squared error is used to compare the efficiency of estimators. To achieve the same "accuracy," a more efficient estimator needs less amount of data.

Relative Efficiency

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two estimators of θ. Then the relative efficiency of $\hat{\theta}_2$ to $\hat{\theta}_1$ is

$$\frac{\mathrm{MSE}_\theta(\hat{\theta}_1)}{\mathrm{MSE}_\theta(\hat{\theta}_2)}.$$

- If the relative efficiency is less than 1, then we say that $\hat{\theta}_1$ is a more efficient estimation of θ than $\hat{\theta}_2$.
- The MSE, and hence the relative efficiency, may depend on the sample size n. Then we can look at the **asymptotic relative efficiency** as $n \to \infty$.

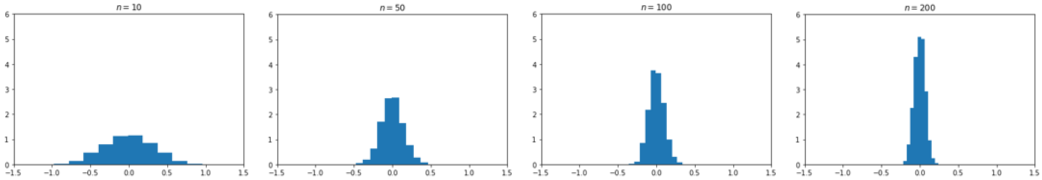
Asymptotic Properties

Consistency

An estimator $\hat{\theta}$ is a *consistent* estimator for θ if $\hat{\theta}$ *converges to* θ *in probability*, i.e.,

$$\mathbb{P}\big(|\hat{\theta} - \theta| > \epsilon\big) \to 0 \quad \text{as} \quad n \to \infty.$$

Example: Sample mean is consistent for the population mean.



This can be understood by the CLT: $\bar{X} \approx \mu + \sqrt{\dfrac{\mathrm{Var}(X_1)}{n}}\, Z.$

Law of Large Numbers (LLN)

Theorem (Weak Law of Large Numbers)

If $X_1, X_2, \ldots, X_n$ are independent random variables having a common distribution F with mean μ, then

$$\mathbb{P}\left(\left|\frac{1}{n}(X_1 + X_2 + \cdots + X_n) - \mu\right| > \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Example: By the LLN, the sample mean is consistent for the population mean μ:

$$\mathbb{P}\left(|\bar{X} - \mu| > \epsilon\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Asymptotic Properties

Consistency implies that having larger sample size helps.

- This is a desired property we want for most estimators.
- In general, the condition $\mathbb{P}\big(|\hat{\theta} - \theta| > \epsilon\big) \to 0$ is not straightforward to check.

MSE allow us to prove consistency without the need of calculating $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon)$.

Theorem (Sufficient condition for consistency)

If the MSE of $\hat{\theta}_n$ converges to 0 as $n \to \infty$, then $\hat{\theta}_n$ is consistent.

Example: Sample mean is consistent for the population mean.

$$\mathrm{MSE}(\bar{X}) = \frac{\sigma^2}{n} \to 0, \quad \text{as} \quad n \to \infty.$$

A Quick Summary: Evaluating Estimators

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

The Dispersion: $\mathrm{Var}(\hat{\theta})$.

An Overall Measure: $\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big] = \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta})^2$.

- When multiple estimators are available, we prefer the one with smaller MSE.

The Convergence: Consistency. MSE converging to 0 implies consistency.

Example: The sample mean is unbiased and consistent.

Example: Consider a Poisson population, we know that $\mathbb{E}[X] = \lambda$. So sample mean can be used to estimate λ. It is unbiased and consistent.

Estimating the variance σ^2 when μ is known

If the mean μ is known, then we can use

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

- It is a valid statistic only when μ is known.
- It is the sample mean of $(X_i - \mu)^2$.

Bias

$$\mathbb{E}[\hat{\sigma}^2] = \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2\Big] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i - \mu)^2] = \frac{1}{n}\sum_{i=1}^{n}\sigma^2 = \sigma^2.$$

Hence, $\hat{\sigma}^2$ is unbiased.

Variance of $\hat{\sigma}^2$ and Its MSE

$$\mathrm{Var}(\hat{\sigma}^2) = \mathrm{Var}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\Big) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\Big((X_i-\mu)^2\Big) = \frac{1}{n}\mathrm{Var}\Big((X_1-\mu)^2\Big).$$

How do we calculate $\mathrm{Var}\Big((X_1-\mu)^2\Big)$?

$$\mathrm{Var}\Big((X_1-\mu)^2\Big) = \mathbb{E}\Big[(X_1-\mu)^4\Big] - \Big(\mathbb{E}\Big[(X_1-\mu)^2\Big]\Big)^2 = \mu_4 - \sigma^4,$$

where $\mu_4 = \mathbb{E}\Big[(X_1-\mu)^4\Big]$ is the 4th central moment (kurtosis).

Hence, $\mathrm{Var}(\hat{\sigma}^2) = \frac{\mu_4}{n} - \frac{\sigma^4}{n}$. Together with the unbiasedness of $\hat{\sigma}^2$, we have

$$\mathrm{MSE}(\hat{\sigma}^2) = \frac{\mu_4}{n} - \frac{\sigma^4}{n} + 0^2 \to 0, \quad \text{which implies that } \hat{\sigma}^2 \text{ is consistent.}$$

Variance of $\hat{\sigma}^2$ and Its MSE

Example: For a normal population $\mathcal{N}(\mu, \sigma^2)$, the 4th central moment is

$$\mu_4 = 3\sigma^4.$$

Thus,

$$\mathrm{MSE}(\hat{\sigma}^2) = \mathrm{Var}(\hat{\sigma}^2) = \frac{2\sigma^4}{n}.$$

Reference: the 4th Central Moment for a Normal Distribution

Consider a random variable $X \sim N(\mu, \sigma^2)$. Define $Z = \frac{X-\mu}{\sigma}$ so that $Z \sim N(0,1)$. Then, the 4th central moment of X is

$$\mu_4 = \mathbb{E}\left[(X-\mu)^4\right] = \mathbb{E}\Big[(\sigma Z)^4\Big] = \sigma^4 \mathbb{E}\left[Z^4\right].$$

Derivation for $\mathbb{E}[Z^4]$: For a standard normal variable, the even moments are given by the formula

$$\mathbb{E}[Z^{2k}] = \frac{(2k)!}{2^k k!}.$$

For $k = 2$, we have $\mathbb{E}[Z^4] = \frac{4!}{2^2 \cdot 2!} = \frac{24}{4 \cdot 2} = \frac{24}{8} = 3$. Hence, the 4th central moment of X is

$$\mu_4 = \sigma^4 \cdot 3 = 3\sigma^4.$$

Estimating the Variance when μ is Unknown

If the mean μ is **unknown**

- the previous estimator is no longer valid because it contains the unknown quantity μ. An estimator must be a statistic.
- we would not be able to determine its realized value even with all observations in hand.

Alternatively, we can use the sample variance defined by

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

Question: But why use $\frac{1}{n-1}$ instead of $\frac{1}{n}$?

Sample variance is unbiased

The use of $\frac{1}{n-1}$ is to make S^2 an **unbiased estimator** of σ^2.

Sample Variance is Unbiased

Proof:

Recall from **Descriptive Statistics (Slide 47)** that $S^2 = \frac{1}{n-1}\sum_{i=1}^n X_i^2 - \frac{n}{n-1}\bar{X}^2$.

Then

$$\begin{aligned}(n-1)\mathbb{E}[S^2] &= \mathbb{E}\left[\sum_{i=1}^n X_i^2 - n\bar{X}^2\right] = n\,\mathbb{E}[X_1^2] - n\,\mathbb{E}[\bar{X}^2] \\ &= n\Big(\mathrm{Var}(X_1) + \mathbb{E}[X_1]^2\Big) - n\Big(\mathrm{Var}(\bar{X}) + \mathbb{E}[\bar{X}]^2\Big) \\ &= n\Big(\sigma^2 + \mu^2\Big) - n\Big(\sigma^2/n + \mu^2\Big) = (n-1)\sigma^2.\end{aligned}$$

Divide both sides by $(n-1)$, we see that S^2 is unbiased.

Variance of the Sample Variance S^2

It turns out that the variance of the sample variance can be quite complicated because the summands are mutually dependent through $\bar{X}$: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Fortunately, an explicit expression exists:

$$\mathrm{Var}(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)},$$

where $\mu_4 = \mathbb{E}[(X_1 - \mu)^4]$ is the 4th central moment (kurtosis). Since S^2 is unbiased, the MSE of S^2 is given by

MSE of the sample variance

$$\mathrm{MSE}(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}.$$

For a detailed calculation, see Section 2.5 in `https://hal.science/hal-02012458`.

Consistency of the Sample Variance

Theorem

Assume that the 4th central moment is finite. Since

$$\mathrm{MSE}(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)} \to 0, \quad \text{as} \quad n \to \infty.$$

The sample variance S^2 is a consistent estimator of the population variance σ^2.

Comparing the Cases with and without Knowledge of μ

With knowledge of μ

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2, \quad \mathrm{MSE}(\hat{\sigma}^2) = \frac{\mu_4}{n} - \frac{\sigma^4}{n}.$$

Without knowledge of μ

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad \mathrm{MSE}(S^2) = \frac{\mu_4}{n} - \frac{n-3}{n-1}\frac{\sigma^4}{n}.$$

Notice that $\mathrm{MSE}(S^2) > \mathrm{MSE}(\hat{\sigma}^2)$. **Having more information (the knowledge of μ) helps reduce the variability of our estimator.**

- The ratio between the two MSE converges to 1 as $n \to \infty$, hence as the sample size increases, they are asymptotically indistinguishable.

Example: Assume in addition that $X_i \sim N(0,1)$ follows a standard normal distribution. Then the 4th central moment is $\mu_4 = 3\sigma^4$. (Slide 27)

Hence,

MSE of the sample variance under normal population

$$\text{MSE}(S^2) = \frac{3\sigma^4}{n} - \frac{n-3}{n-1}\frac{\sigma^4}{n} = \frac{2\sigma^4}{n-1}.$$

We have

$$\text{MSE}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n} = \text{MSE}(\hat{\sigma}^2).$$

However,

$$\frac{\text{MSE}(S^2)}{\text{MSE}(\hat{\sigma}^2)} \to 1 \quad \text{as } n \to \infty.$$

Comparing the Estimators for Variance

We have just compared the two estimators for the variance σ^2:

- $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ when μ is known.
- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ when μ is unknown.
- We see that knowing more information is beneficial.
- However, this comparison is **unfair** as they use different information.

Suppose we compare the following two estimators, both assuming no knowledge of μ:

- $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- $S^2_{\text{alt}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
- Which one is better?

Alternative Estimator when μ is Unknown

Now, let's investigate an alternative estimator:

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

- S^2_{alt} is **biased**:

$$\mathrm{Bias}(S^2_{\text{alt}}) = \mathbb{E}[S^2_{\text{alt}}] - \sigma^2 = \frac{n-1}{n}\mathbb{E}[S^2] - \sigma^2 = -\frac{\sigma^2}{n}.$$

- But it is **asymptotically unbiased**:

$$\mathrm{Bias}(S^2_{\text{alt}}) \to 0 \quad \text{as } n \to \infty.$$

Asymptotically Unbiased Estimator

An estimator is **asymptotically unbiased** if its bias converges to 0 as the sample size increases.

Variance and MSE of the Alternative Estimator S^2_{alt}

Variance of S^2_{alt}:

$$\operatorname{Var}(S^2_{\text{alt}}) = \operatorname{Var}\left(\frac{n-1}{n}S^2\right) = \left(\frac{n-1}{n}\right)^2 \operatorname{Var}(S^2) = \left(\frac{n-1}{n}\right)^2 \left(\frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}\right).$$

MSE of S^2_{alt}:

$$\operatorname{MSE}(S^2_{\text{alt}}) = \operatorname{Var}(S^2_{\text{alt}}) + \operatorname{Bias}^2(S^2_{\text{alt}}) = \left(\frac{n-1}{n}\right)^2 \left(\frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}\right) + \frac{\sigma^4}{n^2}.$$

Example: For a standard normal population, $\mu_4 = 3\sigma^4$. Then,

$$\operatorname{Var}(S^2_{\text{alt}}) = \left(\frac{n-1}{n}\right)^2 \left(\frac{2\sigma^4}{n-1}\right),$$

$$\operatorname{MSE}(S^2_{\text{alt}}) = \left(\frac{n-1}{n}\right)^2 \left(\frac{2\sigma^4}{n-1}\right) + \frac{\sigma^4}{n^2} = \frac{\sigma^4(2n-1)}{n^2} < \frac{2\sigma^4}{n-1} = \operatorname{MSE}(S^2).$$

Comparing S^2_{alt} and S^2 for Normal Population

- $S^2_{\text{alt}} = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is biased. $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ is unbiased.
- In the Gaussian (normal) case,

$$\mathrm{MSE}(S^2_{\text{alt}}) < \mathrm{MSE}(S^2),$$

hence S^2_{alt} is considered a better estimator.

Bias-Variance Tradeoff:

Usually, we can considerably decrease the variance of an estimator by introducing a little bit of bias. Overall, we may reduce the MSE, hence obtaining a better estimator.

- One may check that

$$\frac{\mathrm{MSE}(S^2)}{\mathrm{MSE}(S^2_{\text{alt}})} \to 1 \quad \text{as } n \to \infty.$$

So as the sample size grows, the efficiency is *almost the same*.

Example: Estimation for Poisson

Recall the Poisson(λ) distribution with PMF

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

We have

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

We can use either the sample mean or the sample variance to estimate λ.

Both are unbiased and consistent. Which one is better?

- The MSE of $\bar{X}$: $\mathrm{Var}(\bar{X}) = \frac{\lambda}{n}$.
- The MSE of S_n^2: $\mathrm{Var}(S_n^2) = \frac{1}{n}\left(\mathbb{E}[X^4] - \mathbb{E}[X^2]^2\frac{n-3}{n-1}\right) = \frac{\lambda}{n}(1 + 2\lambda\frac{n}{n-1})$. This one needs tedious computation so the steps are omitted.
- The asymptotic relative efficiency of $\bar{X}$ to S_n^2 is $1 + 2\lambda$. So $\bar{X}$ is better.

Two Systematic Ways to Construct Estimators

We have now studied in detail the properties of the sample mean and sample variance, as estimators of the population mean and variance, respectively.

Question: How do we construct point estimators for other parameters of interest?

Next, we look at two more ways to construct point estimators

- Method of moments (MoM)
- Maximum likelihood estimators (MLE)

Sample Moments

Recall that the jth population moment defined as

$$\mu_j = \mathbb{E}[X^j].$$

Sample moments

Let $X_1, \ldots, X_n$ be a random sample, the jth sample moment is defined as

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

- Sample moments have the desired properties of unbiasedness and consistency (under very mild technical conditions).
- **Why?**

Example: Estimating μ and σ^2 using Sample Moments

Sample moments can be used to estimate parameters other than the population moments. To demonstrate the main idea, consider the following example.

0 **Example:** Find estimators of μ and σ^2 for a general population using sample moments.

- Recall that $\mu = \mu_1$ and $\sigma^2 = \mu_2 - \mu_1^2$, where μ_1 and μ_2 are the first and second population moments, respectively. Equivalently, $\mu_1 = \mu$ and $\mu_2 = \sigma^2 + \mu^2$.
- We have natural estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ for μ_1 and μ_2, namely,

$$\hat{\mu}_1 \approx \mu_1 = \mu \quad \text{and} \quad \hat{\mu}_2 \approx \mu_2 = \sigma^2 + \mu^2.$$

- These give us a set of two equations in the two unknowns μ and σ^2, which yield the estimators:

$$\mu \approx \hat{\mu}_1 = \bar{X}, \quad \sigma^2 \approx \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Derivation of $\hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$$
\begin{aligned}
\hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 &= \frac{1}{n}\sum_{i=1}^{n}\Big[(X_i - \bar{X}) + \bar{X}\Big]^2 - \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[(X_i - \bar{X})^2 + 2\bar{X}(X_i - \bar{X}) + \bar{X}^2\Big] - \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 + \frac{2\bar{X}}{n}\sum_{i=1}^{n}(X_i - \bar{X}) + \bar{X}^2 - \bar{X}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2,
\end{aligned}
$$

since

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0.$$

The Method of Moments generalizes this idea of constructing estimators based on sample moments.

Method of Moments

Suppose we have p numbers of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Suppose we also have a random sample $X_1, \ldots, X_n$ of size n. To estimate the parameters $\boldsymbol{\theta}$:

1. **Step 1: Calculate the first p population moments as functions of $\boldsymbol{\theta}$.**

$$\mu_1 = \mu_1(\theta_1, \ldots, \theta_p) = \mathbb{E}[X^1], \quad \ldots, \quad \mu_p = \mu_p(\theta_1, \ldots, \theta_p) = \mathbb{E}[X^p].$$

2. **Step 2: Equate the p population moments to the p corresponding sample moments,** we get a set of p equations to solve for p unknown variables:

$$\mu_1(\theta_1, \ldots, \theta_p) = \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \ldots, \quad \mu_p(\theta_1, \ldots, \theta_p) = \frac{1}{n}\sum_{i=1}^{n} X_i^p.$$

3. **Step 3:** The solutions $\hat{\theta}_1, \ldots, \hat{\theta}_p$ are called the **MoM estimators** for $\theta_1, \ldots, \theta_p$.

MoM for Mean and Variance

Example: To estimate the mean μ, the sample mean is the MoM estimator for μ.

Example: To estimate the mean μ and variance σ^2 at the same time, so $\boldsymbol{\theta} = (\mu, \sigma^2)$.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Thus, the MoM estimators are $\hat{\mu} = \bar{X}$ and

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

Method of Moments

Example: (Normal) $\mathcal{N}(\mu, \sigma^2)$. The unknown parameters are $\boldsymbol{\theta} = (\mu, \sigma^2)$.

$$\mathbb{E}[X] = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

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

So the estimators are

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

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

Example: Binomial (k, p). The unknown parameters are $\boldsymbol{\theta} = (k, p)$.

$$\mathbb{E}[X] = kp = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

$$\mathbb{E}[X^2] = kp(1-p) + k^2p^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

So the estimators are

$$\hat{k} = \frac{\bar{X}^2}{\bar{X} - \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \in \mathbb{Z}_+?$$

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

Remarks

Advantages

- Simple to use and easy to construct.
- Usually gives consistent estimators, though often biased.

However,

- It sometimes give estimates outside the domain, like $\hat{k}$ in the last slide.
- It is quite arbitrary depending on which moments to use. By default, we use the first p moments, if p parameters are to be estimated.

Maximum Likelihood Estimator

In estimating parameters of a Normal distribution and general distributions, Gauss and R.A. Fisher (widely regarded as the father of modern statistics) summarized the following magnificent intuition:

> A proper estimator of the true parameter is the one that makes the given observation $x_1, x_2, \ldots, x_n$ most likely to occur.

Figure: Carl Friedrich Gauss (left) and R.A. Fisher (right)

Maximum Likelihood Estimator

Example: 10 coin tosses gives you 9 heads and 1 tail. Can you confidently estimate that the probability of getting a head is $p = 0.5$? No!

Because the probability of getting "9 heads and 1 tail" under $p = 0.5$ is

$$\mathbb{P}_{p=0.5}(\text{9 heads out of 10}) = \binom{10}{1} \times 0.5^9 \times 0.5^1 = \frac{10}{1024} \approx 0.01.$$

How about $p = 0.9$?

The probability of getting "9 heads and 1 tail" under $p = 0.9$ is

$$\mathbb{P}_{p=0.9}(\text{9 heads out of 10}) = \binom{10}{1} \times 0.9^9 \times 0.1^1 \approx 0.39 \gg 0.01.$$

$p = 0.9$ is the scenario where the data "9 heads and 1 tail" has much more chance of being observed.

Intuition for Maximum Likelihood Estimation

Intuition:

- To find the estimator, we search among all possible values of the parameter for the one that makes the data we have the most likely to occur.

Implementation:

- Find a reasonable function to represent the “likelihood” of observing a data set.
- Maximize the “likelihood function.”

The goal of maximum likelihood estimation is to determine the parameters for which the observed data have the highest joint probability.

Likelihood Function

To measure the change of occurance, we introduce the likelihood function.

Let $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ be an *observed* sample from a pdf $f(x|\theta)$ with parameter θ.

Likelihood function

The likelihood function of the sample $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ is the joint probability density function of $\boldsymbol{X}$, evaluated at $\boldsymbol{X} = \boldsymbol{x}$, i.e.,

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

- The likelihood function is a function of θ.
- It measures how likely an experiment produces the *observed* $\boldsymbol{x}$ as a sample, if the parameter takes the value of θ.

Comparing Joint Density with Likelihood

- **Joint Density:** $f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$.
  - This is viewed as a function of $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ with the parameter θ fixed.
  - **Probability Theory Perspective:** How likely is it to observe $\boldsymbol{X} = \boldsymbol{x}$ under a true, fixed (but unknown) parameter θ?
- **Likelihood:** $L(\theta|\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i|\theta)$.
  - Interpreted as a function of θ with the observed data $\boldsymbol{x}$ held fixed.
  - **Statistical Perspective:** Now that we have observed $\boldsymbol{x}$, what is the chance of observing $\boldsymbol{x}$ if the unknown parameter takes on a particular value of θ?

This view provides a tool to assess the plausibility of different guesses for θ, under the observed $\boldsymbol{x}$, leading to the principle of choosing the most plausible value as our estimate.

Likelihood Inferences

- We saw that the likelihood function $L(\theta|\boldsymbol{x})$ is just the probability of obtaining the data $\boldsymbol{x}$ when the true value of the parameter is θ.
- This imposes a belief ordering on possible values of the parameter θ: we believe θ_1 is more plausible than θ_2 if

$$L(\theta_1|\boldsymbol{x}) > L(\theta_2|\boldsymbol{x}).$$

- Maximum likelihood estimation is based on this ordering.
  - It is possible that the value of $L(\theta|\boldsymbol{x})$ is very small for every value of θ.
  - It is not the actual value of the likelihood that tells us how much support to give a particular θ, but rather its value relative to the likelihoods of other possible parameter values.

Maximum Likelihood Estimator

Maximum Likelihood Estimators

The **Maximum Likelihood Estimators** (MLE), denoted as $\hat{\theta}_{\text{MLE}}$, is the value of the parameter θ that maximizes the likelihood function for the given sample $\boldsymbol{X}$, i.e.

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} L(\theta|\boldsymbol{X})$$

- Equivalently, $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \log L(\theta|\boldsymbol{X})$ maximizes the log-likelihood function.
- This is because the log function is strictly monotone.
- We will see why this can help in many of our cases.

Calculating the Maximum Likelihood Estimator

To calculate the MLE, we need to solve the maximization problem.

$$\hat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta} L(\theta|\boldsymbol{X})$$

From calculus, you know that

- It can be solved by
  - If θ is 1-dimensional, set the derivative $\frac{d}{d\theta} L(\theta|\boldsymbol{X}) = 0$ or $\frac{d}{d\theta} \log L(\theta|\boldsymbol{X}) = 0$.
  - If θ is k-dimensional, set the partial derivatives $\frac{\partial}{\partial\theta_i} L(\theta|\boldsymbol{X}) = 0$ or $\frac{\partial}{\partial\theta_i} \log L(\theta|\boldsymbol{X}) = 0$
- Check the second-order derivative to make sure it is not local minimum.

Maximizing $\log L(\theta|\boldsymbol{X})$ is usually much easier. As we see next in the Normal example.

Example: MLE for Bernoulli Population I

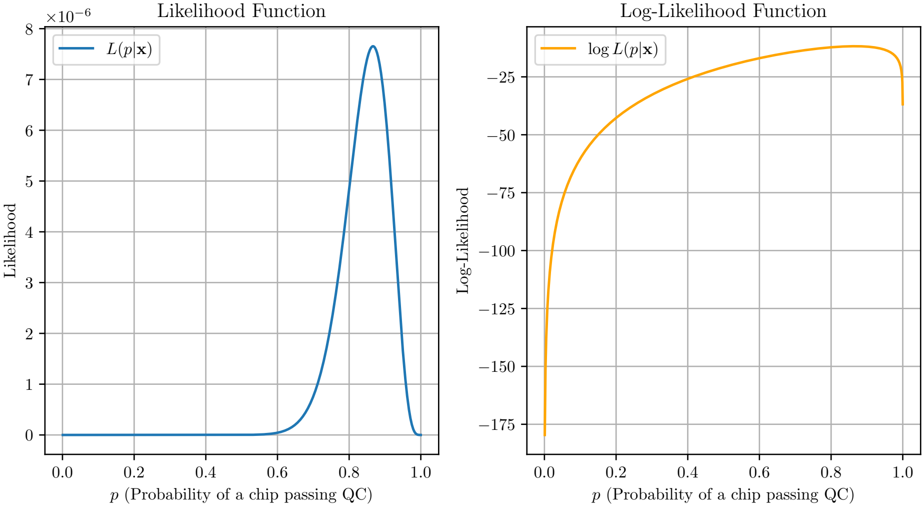
Example: Suppose we randomly choose 30 chips from a production line and find that 26 of them are of acceptable quality. We wish to estimate the probability p that a chip passes quality control (QC).

- **Modeling:** Assume that each chip passes QC with probability p, independent of others.
- **Likelihood Function:**

$$L(p|\boldsymbol{X}) = \prod_{i=1}^{n} f(X_i|p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

- How to maximize $L(p|\boldsymbol{X})$ over $p \in [0,1]$?

Example: MLE for Bernoulli Population II



Example: MLE for Bernoulli Population III

The likelihood function is given by

$$L(p|\boldsymbol{X}) = \prod_{i=1}^{n} f(X_i|p) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} = p^{\sum_{i=1}^{n} X_i}(1-p)^{n-\sum_{i=1}^{n} X_i}.$$

Taking logarithms, we have

$$\begin{aligned} \ell(p|\boldsymbol{X}) &\equiv \log L(p|\boldsymbol{X}) \\ &= \log p^{\sum_{i=1}^{n} X_i} + \log(1-p)^{n-\sum_{i=1}^{n} X_i} \\ &= \left(\sum_{i=1}^{n} X_i\right)\log p + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-p). \end{aligned}$$

Example: MLE for Bernoulli Population IV

To maximize the **log-likelihood**, we differentiate with respect to p:

$$\frac{d}{dp}\ell(p|\boldsymbol{X}) = \frac{1}{p}\Big(\sum_{i=1}^{n} X_i\Big) - \frac{1}{1-p}\Big(n - \sum_{i=1}^{n} X_i\Big).$$

Setting the derivative equal to zero,

$$\frac{1}{p}\Big(\sum_{i=1}^{n} X_i\Big) - \frac{1}{1-p}\Big(n - \sum_{i=1}^{n} X_i\Big) = 0,$$

we solve for p and obtain the MLE:

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

Example: MLE for Bernoulli Population V

To verify that this is a maximum, we check the second derivative:

$$\frac{d^2}{dp^2}\ell(p|\boldsymbol{X}) = -\frac{1}{p^2}\left(\sum_{i=1}^{n} X_i\right) - \frac{1}{(1-p)^2}\left(n - \sum_{i=1}^{n} X_i\right) < 0.$$

Evaluating at $p = \hat{p}_{\text{MLE}}$, we see that the second derivative is negative, confirming a maximum.

In our example, with $n = 30$ chips and $\sum_{i=1}^{30} x_i = 26$ acceptable chips, the MLE is

$$\hat{p}_{\text{MLE}} = \frac{26}{30} \approx 0.8667.$$

Maximum Likelihood Estimator – Bernoulli

$$\frac{d}{dp}\log L(p|\boldsymbol{X}) = \frac{1}{p}\sum_{i=1}^{n} X_i - \frac{1}{1-p}\Big(n - \sum_{i=1}^{n} X_i\Big)$$

MLE for Bernoulli Sample

$$\hat{p}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Example 2 – Bernoulli Population I

Suppose I asked Tony and Paul to proofread “T4.pdf”. The findings are:

- Tony found 10 typos.
- Paul found 12 typos.
- 6 typos were found by both.

Assume they work independently (without collusion).

Goal: Estimate the total number of typos, N.

Modeling: Assume each typo is detected by Tony with probability p_1 and by Paul with probability p_2 (Bernoulli random variables). Suppose there are N typos in total.

Estimation:

Example 2 – Bernoulli Population II

- Tony's detection: Tony finds 10 out of N typos. Thus, by the MLE for Bernoulli,

$$\hat{p}_1 = \frac{10}{N}.$$

- Overlap: Among the 12 typos found by Paul, Tony identified 6. Thus, by the MLE,

$$\hat{p}_1 = \frac{6}{12} = \frac{1}{2}.$$

Equate the two estimates:

$$\frac{10}{N} = \frac{1}{2} \quad \Rightarrow \quad N = 20.$$

Therefore, the estimated total number of typos in "T4.pdf" is 20.

Example: MLE for Normal Population

Example: Normal sample. Likelihood function

$$L(\mu, \sigma | \boldsymbol{X}) = f(\boldsymbol{X} | \mu, \sigma) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[\frac{-\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}\right]$$

Take logarithm (because derivatives of multiplications are usually harder to calculate)

$$\log L(\mu, \sigma | \boldsymbol{X}) = -n \log(\sqrt{2\pi}\sigma) - \frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2}.$$

Take partial derivatives

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma | \boldsymbol{X}) = -\frac{\sum_{i=1}^n (X_i - \mu)}{\sigma^2}, \qquad \frac{\partial}{\partial \sigma} \log L(\mu, \sigma | \boldsymbol{X}) = -\frac{n}{\sigma} + \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma^3}$$

MLE for Normal Sample, $\theta = (\mu, \sigma)$

$$\hat{\mu}_{\text{MLE}} = \frac{\sum_{i=1}^n X_i}{n} \quad \text{and} \quad \hat{\sigma}_{\text{MLE}} = \left[\frac{\sum_{i=1}^n (X_i - \hat{\mu})^2}{n}\right]^{\frac{1}{2}}$$

Normal MLE

For Normal random sample, the MLEs are $\mu_{\mathrm{MLE}} = \bar{X}$ and $\hat{\sigma}^2_{\mathrm{MLE}} = \frac{n-1}{n} S^2$.

We have seen that

- the MLE gives smaller MSE, trading off bias for smaller variance.
- Both S^2 and $\hat{\sigma}^2_{\mathrm{MLE}}$ are asymptotically unbiased and consistent.

MLE for Normal Distribution Parameters (μ, σ^2)

Example: What is the MLE for (μ, σ^2) in a Normal sample?

The log-likelihood function

$$\log L(\mu, \sigma^2 | \boldsymbol{X}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

It has the same form, but understood as a function of σ^2.

Differentiate with respect to σ^2:

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Setting $\frac{\partial \log L}{\partial \sigma^2} = 0$ and substituting $\mu = \bar{X}$ yields:

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

MLE for Normal Distribution Parameters (μ, σ^2)

MLE for normal parameter (μ, σ^2)

The MLE for (μ, σ^2) is

$$\hat{\mu}_{\text{MLE}} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Notice that $\hat{\sigma}^2_{\text{MLE}}$ is exactly the square of the MLE for σ.

Invariance Property of MLEs

For a distribution, we may use different parameters to describe it. Or, we may be interested in a new parameter that is a function of the default parameters.

Example: For Normal distribution, we can may use either standard devation or variance.

Do we need to recalculate the MLE every time we change the parameter? No!

Theorem (Invariance of MLE)

Let $\hat{\theta}_1, \dots, \hat{\theta}_k$ are the MLEs of the parameters $\theta_1, \dots, \theta_k$. Then the MLE of any function $h(\theta_1, \dots, \theta_k)$ is $h(\hat{\theta}_1, \dots, \hat{\theta}_k)$.

Example: If $\bar{X}$ is the MLE for θ, then $\bar{X}^2$ is the MLE for θ^2.

Example: If $\hat{\sigma}$ is the MLE for the standard deviation, then $\hat{\sigma}^2$ is the MLE for the variance, and vice versa.

Example: MLE for Poisson Population

Assume that $X_1, X_2, \dots X_n \sim \text{Poisson}(\lambda)$.

Likelihood

$$L(\lambda|\boldsymbol{X}) = \frac{e^{-\lambda}\lambda^{X_1}}{X_1!}\frac{e^{-\lambda}\lambda^{X_2}}{X_2!}\dots\frac{e^{-\lambda}\lambda^{X_n}}{X_n!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n X_i}}{x_1!\dots X_n!}$$

Take logarithm

$$\log L(\lambda|\boldsymbol{X}) = -n\lambda + \sum_{i=1}^n X_i \log(\lambda) - \log(X_1!\dots X_n!)$$

Differentiate

$$\frac{d}{d\lambda}\log L(\lambda|\boldsymbol{X}) = -n + \frac{1}{\lambda}\sum_{i=1}^n X_i$$

MLE for Poisson Sample

$$\hat{\lambda}_{\text{MLE}} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Example: MLE for Exponential Distribution

Assume that $X_1, X_2, \ldots, X_n \sim$ Exponential(λ). Likelihood

$$L(\lambda|\boldsymbol{X}) = \prod_{i=1}^{n} \lambda e^{-\lambda X_i} = \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} X_i\Big).$$

Take logarithm

$$\log L(\lambda|\boldsymbol{X}) = n\log(\lambda) - \lambda \sum_{i=1}^{n} X_i.$$

Differentiate

$$\frac{d}{d\lambda} \log L(\lambda|\boldsymbol{X}) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i.$$

MLE for Exponential Sample

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

Example: Apply the Invariance Property

Example: Suppose 10 rats are used in a biomedical study where they are injected with cancer cells and given a drug to increase their survival rate. The survival times (months) are 14, 17, 27, 18, 12, 8, 22, 13, 19 and 12. We usually use exponential distribution to model survivals. What is the MLE of the mean survival time?

- For exponential $f(x|\lambda) = \lambda \exp(-\lambda x)$, the MLE is $1/\bar{X}$.
- So the mean survival time is $1/\hat{\lambda} = \bar{X} = 16.2$.

Properties of a MLE

Properties of a MLE

Under very general and not restrictive conditions when the sample size n is large, the MLE $\hat{\theta}_{\mathrm{MLE}}$ enjoys the following properties

- $\hat{\theta}_{\mathrm{MLE}}$ is approximately unbiased estimator for θ.
- The variance of $\hat{\theta}_{\mathrm{MLE}}$ is nearly as small as the variance that could be obtained with any other estimator.
- $\hat{\theta}_{\mathrm{MLE}}$ has an approximate normal distribution.

This explains why the maximum likelihood estimation technique is widely used.

An Important Remark

In many cases, there will be restrictions on the value a parameter can take, e.g. in a set of possible values Θ.

Example: The rate of the exponential distribution is always larger than zero. So $\Theta = (0, \infty)$ in this case.

Example: The binomial distribution with unknown number of trials k and unknown success probability p. Then $k \in \mathbb{Z}_+$ and $p \in [0, 1]$.

Example: We may only be interested in several potential values of a paremeter. (Perhaps only these values are feasible in your experiments, say due to the restriction of your equiptment.)

An Important Remark

In the case where the parameters are restricted, MLE only maximize over the set of possible values Θ. For this reason, MLE will not produce invalid parameter values, e.g. non-integer k for binomial distribution as in MoM estimator.

MLE is calculated by

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(\theta|\boldsymbol{X}).$$

- In all previous examples, we did not explicitly consider the restriction on the range, e.g. for the variance of Normal. But the solutions turn out to fall in the range. **Example:** Variance of Normal. Recall that

$$\sigma^2_{\mathrm{MLE}} = \frac{\sum_{i=1}^{n}(X_i - \hat{\mu})^2}{n} \geq 0.$$

- Usually, we first maximize regardless of the range, then check if the MLE falls in the range.

MLE for Continuous Uniform Distribution

Assume that $X_1, \ldots, X_n$ are a random sample from Uniform$(0, \theta)$, where $\theta \geq 0$. We want to estimate the parameter θ.

- Suppose you observed a piece of data $x_1 = 1.77$.
  - What can you say about θ?

$$\theta \geq 1.77$$

- Suppose you observed another piece of data $x_2 = 0.46$.
  - What can you say about θ?

$$\theta \geq 1.77$$

- Suppose you observed yet another piece of data $x_3 = 1.82$.
  - What can you say about θ?

$$\theta \geq 1.82$$

Question: Suppose you have $x_1, \ldots, x_n$, can you guess an estimator for θ?

A Different **Example:** MLE for Uniform Population

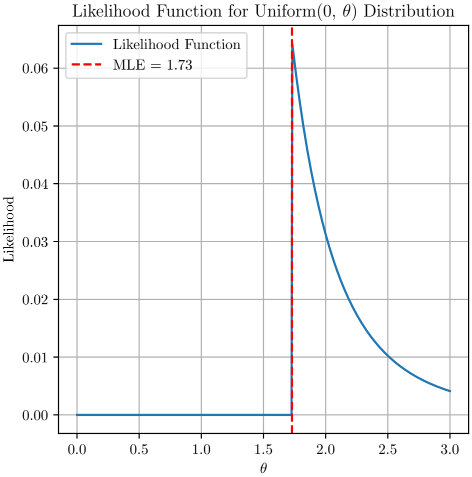
Example: Assume that $X_1, X_2, \ldots, X_n \sim \text{Uniform}([0, \theta])$. Likelihood

$$f(\boldsymbol{X}|\theta) = \frac{1}{\theta} \times \cdots \times \frac{1}{\theta} \mathbb{1}_{0 \le X_1 \le \theta} \cdots \mathbb{1}_{0 \le X_n \le \theta}$$

MLE for Uniform Sample

$$\hat{\theta} = \max(X_1, \ldots, X_n)$$

MoM uses $\hat{\theta} = 2\bar{X}$. The two estimators are very different.



Alternative Estimators for Uniform

Suppose $X_1, \ldots, X_n$ are sampled from Uniform$([0, \theta])$.

- Estimator 1 (MoM)

$$d_1(\boldsymbol{X}) = 2\bar{X}$$

- Estimator 2 (MLE)

$$d_2(\boldsymbol{X}) = \max_i X_i$$

Compare: For the MoM estimator:

$$\begin{aligned}
\text{Bias}_\theta d_1(\boldsymbol{X}) &= 0 \\
\mathrm{MSE}_\theta d_1(\boldsymbol{X}) &= \mathrm{Var}(d_1(\boldsymbol{X}, \theta)) + 0^2 \\
&= \frac{4}{n^2}\left(\mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)\right) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}
\end{aligned}$$

*Alternative Estimators for Uniform

What is the distribution of $d_2(\boldsymbol{X}) = \max_i X_i$?

Note that

$$\mathbb{P}(\max_i X_i < x) = \mathbb{P}(X_1 \leq x, X_2 \leq x, \ldots, X_n \leq x) = \mathbb{P}(X_1 \leq x)\mathbb{P}(X_2 \leq x) \cdots \mathbb{P}(X_n \leq x),$$

where $\mathbb{P}(X_i \leq x) = \frac{x}{\theta} \mathbb{1}_{0 \leq x \leq \theta}$ is the CDF of Uniform$([0, \theta])$.

Taking derivative with respect to x, we have the pdf

$$f_{d_2}(x) = \frac{n}{\theta^n} x^{n-1} \mathbb{1}_{0 \leq x \leq \theta}.$$

*Alternative Estimators for Uniform

So

$$\mathbb{E}[d_2(\boldsymbol{X})] = \frac{n}{n+1}\theta \qquad \mathrm{Var}(d_2(\boldsymbol{X})) = \frac{n\theta^2}{(n+2)(n+1)^2}$$

Hence

$$\begin{aligned}\text{Bias}_\theta d_2(\boldsymbol{X}) &= -\frac{1}{n+1}\theta \\ \mathrm{MSE}_\theta d_2(\boldsymbol{X}) &= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{1}{(n+1)^2}\theta^2 \\ &= \frac{2\theta^2}{(n+1)(n+2)} < \frac{\theta^2}{3n} = \mathrm{MSE}_\theta d_1(\boldsymbol{X})\end{aligned}$$

*Comparison of d_1 and d_2

- d_1 is unbiased; has MSE $\frac{\theta^2}{3n}$; is consistent (WLLN).
- d_2 is negatively biased; has MSE $\frac{2\theta^2}{(n+1)(n+2)}$; is consistent and asymptotically unbiased.
- The asymptotic relative efficiency of d_2 to d_1 is infinite.
- Those MSE decreases on the order of n^{-2} is called super efficient. Recall that the MSE of sample mean (which is a decent estimator) only decreases at n^{-1}.

Extended Readings

- Sections 7.3, 7.4 of Douglas C. Montgomery and George C. Runger, *Applied Statistics and Probability for Engineers, 7th Ed.*