

Mathematical Statistics

Course Information

- ▶ **Lecturer:** Oana Lang, e-mail: oana.lang@ubbcluj.ro
- ▶ **Lectures:** Wednesday between 8:00 - 10:00 Amf. Goangă (Level 1, Kogălniceanu).
- ▶ **Examination:**
 - ▶ 70% written examination in January/February
 - ▶ 30% Lab & seminar project - Deadline: 25th of January 2026
 - ▶ extra 10% possible based on seminar activity - minimal requirement: 3 'stars' per term where 1 ★ is given for a demonstration at the board/ 1 assignment
- ▶ **Pre-requisites:** a 1st course in probability theory

Further literature

1. M. Baron, *Probability and Statistics for Computer Scientists* (2019).
2. G. Casella, R. L. Berger, *Statistical inference*, 2nd edition (2002).
3. H. Lisei, *Mathematical Statistics - Lecture Notes* (2023-2024).
4. S. Micula, *Probability and Statistics for Computational Sciences* (2009).
5. S. Chongchitnan, *Exploring University Mathematics with Python*.
Python codes: [here](#).

Labs

- ▶ Lab 1: [this link - click!](#)
- ▶ Lab 2: [this link - click!](#)
- ▶ Lab 3: [this link - click!](#)
- ▶ Lab 4: [this link - click!](#)
- ▶ Lab 5: [this link - click!](#)
- ▶ Lab 6: [this link - click!](#)
- ▶ Lab 7: [this link - click!](#)

LECTURE 1

What's Stats All About?

- ▶ **Statistics** is the science of collecting, analyzing, and interpreting data to uncover patterns and inform decisions.
 - ▶ Bridging Data and Models
 - ▶ **Models** alone are abstract and need real data to validate or refine them.
 - ▶ **Data** alone lacks context or structure, which models provide.
- ⇒ by combining both **data and models** we can make better informed decision. **Statistics** enables powerful insights through **data analysis**.

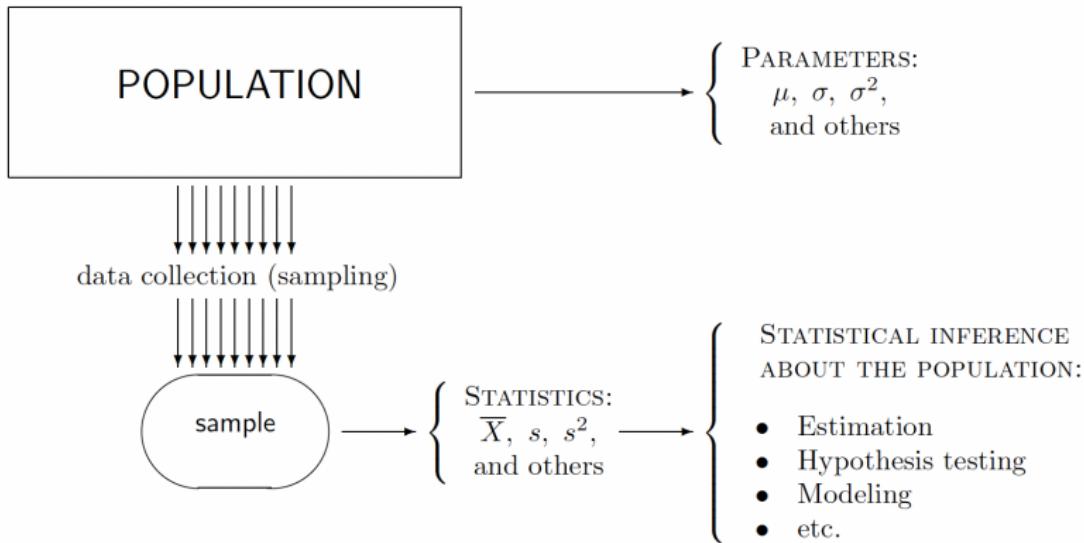
What's Stats All About?

Descriptive Statistics:

- ▶ Provides tools for summarizing and understanding data.
- ▶ Focuses on central tendencies and variability.

Inferential Statistics:

- ▶ Allows for making predictions and inferences about populations based on samples.
- ▶ Includes hypothesis testing and confidence intervals.



Population, samples, parameters, inference ([Baron])

Definitions

- ▶ A **population** consists of all units of interest.
- ▶ Any numerical characteristic of a population is a **parameter**.
- ▶ A **sample** consists of observed units collected from the population.
- ▶ Any function of a sample is called **statistic**.

- ▶ Data collection is a crucially important step in Statistics. We use the collected and observed sample to make statements about a much larger set - the population.

Population and Sampling

- ▶ In practical situations, we aim to draw conclusions about the population. To calculate probabilities, expectations, and make informed decisions in the face of uncertainty, it is essential to understand the population parameters. The only way to ascertain these parameters accurately is through measuring the entire population, which requires conducting a census.
- ▶ Instead of resorting to a census, we can gather data through a random sample from the population. This sample serves as our data. We can analyze this data, perform calculations, and estimate the unknown population parameters with a certain degree of measurable accuracy.

Notation

θ = population parameter

$\hat{\theta}$ = its estimator, obtained from a sample

Example: Employee Engagement

Consider a company that claims 75% of its employees are engaged in their work. However, this does not guarantee that exactly 75 out of every 100 employees in your selected sample are engaged.

For instance, there can be a 0.025 probability that only 60 out of a sample of 100 employees are engaged. This implies a 2.5% chance that a random sample could indicate that, contrary to the claim, only 60% (or less) of employees are actually engaged.

⇒ a sample may sometimes provide misleading insights about the population, even if such occurrences are statistically rare → sampling errors cannot be ignored.

Sampling and Non-Sampling Errors

Sampling and non-sampling errors refer to any discrepancies between a collected sample and the overall population.

Sampling Errors:

- ▶ Arise from the fact that we observe only a subset of the population (one selected sample).
- ▶ For most reasonable statistical methods, sampling errors decrease (and converge to zero) as the sample size increases.

Non-Sampling Errors:

- ▶ Occur due to inappropriate sampling methods or flawed statistical techniques.
- ▶ No sound statistical methods can rectify a poorly collected data sample.

Example 1: Inadequate Population Representation

Surveying the Wrong Group: To assess the customer service of a local cafe, a survey is conducted among attendees of a yoga class nearby. This sample is unlikely to represent the entire population of cafe customers. For instance, regular patrons and occasional visitors may have very different experiences and opinions about the service.

Example 2: Dependent Observations

Dependent Feedback: A manager asks a couple of close team members which project management tool they prefer and generalizes these responses to conclude which tool is better. Since these employees frequently collaborate, their preferences may be influenced by discussions with one another.

While dependent observations do not always result in non-sampling errors, independence cannot be assumed in this context.

Example 3: Unequal Sampling Probability

Inequitable Sampling Method: A study on commuter preferences for public transport is conducted by randomly selecting bus routes and then surveying passengers on those buses. If one bus route has significantly more frequent riders compared to another, then passengers on the bus with higher ridership are over-represented in the sample. This unequal probability must be accounted for to avoid non-sampling errors.

Example 4: Historical Bias in Polling

Polling Error in 1948: In the 1948 U.S. presidential election, a major newspaper published a prediction favoring Thomas Dewey over Harry Truman based on a poll. The sampling methodology was flawed because it relied heavily on responses from individuals with telephones, a demographic skewed towards wealthier voters. This resulted in a significant bias, as many supporters of Truman were not included in the sample, leading to an inaccurate prediction.

Simple Random Sampling

We are going to work mainly with simple random sampling, to avoid non-sampling errors.

Definition 1

Simple random sampling is a sampling technique in which units are selected from the entire population independently, with each unit having an equal chance of being included in the sample.

Observations obtained through simple random sampling are considered iid (independent, identically distributed) random variables.

Descriptive Statistics

Suppose we have collected the following random sample

$$\mathcal{S} = (X_1, X_2, \dots, X_n)$$

For instance, to assess the effectiveness of a marketing campaign, we recorded the number of new customers gained in $n = 30$ randomly selected weeks:

20	34	45	60	25	38	40	28	32	10
55	22	47	50	33	15	26	30	42	18
35	39	54	29	46	27	12	24	31	14.

What insights can we derive from this data collection? We recognize that X , the number of new customers in a given week, is a random variable, and its value may not be one of the thirty recorded observations. We will utilize the collected data to characterize the distribution of X .

Characteristics of Descriptive Statistics

Simple descriptive statistics that measure location, spread, variability, and other features can be calculated from this data. In this section, we will discuss the following statistics:

- ▶ **Mean:** Measures the average value of the sample.
- ▶ **Median:** Indicates the central value of the data set.
- ▶ **Quantiles and Quartiles:** Show where certain portions of the data/sample are located.
- ▶ **Variance, Standard Deviation, and Interquartile Range:**
Measure variability and the spread of the data.

Each statistic is a random variable, as it is computed from random data. Each one estimates a corresponding population parameter and provides insights into the distribution of X , the variable of interest.

Sample Mean

The sample mean \bar{X} serves as an estimator for the population mean $\mu = \mathbb{E}[X]$.

Definition 2

The sample mean \bar{X} is defined as the arithmetic average:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n}.$$

As the average of the sampled observations, \bar{X} estimates the overall average value of the distribution of X . Since \bar{X} is calculated from random data, it does not always equal μ . However, we expect it to approach μ with larger sample sizes.

Remark: This holds true provided that the population has finite mean and variance.

Unbiasedness

Definition 3

An estimator $\hat{\theta}$ is considered **unbiased** for a parameter θ if its expectation equals that parameter:

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

for all possible values of θ . The bias of $\hat{\theta}$ is defined as:

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

Unbiasedness implies that if we were to collect a large number of samples and compute $\hat{\theta}$ from each, the average of these estimators would converge exactly to the unknown parameter θ . That is, in the long run, unbiased estimators neither overestimate nor underestimate the parameter.

For instance, **the sample mean is an unbiased estimator for the population mean**:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{X_1 + \dots + X_n}{n}\right] = \frac{\mathbb{E}[X_1] + \dots + \mathbb{E}[X_n]}{n} = \frac{n\mu}{n} = \mu.$$

Consistency

Definition 4

An estimator $\hat{\theta}$ is **consistent** for a parameter θ if the probability of its sampling error being any magnitude, converges to 0 as the sample size approaches infinity"

$$\mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) \rightarrow 0 \text{ as } n \rightarrow \infty$$

for any $\varepsilon > 0$.

This means that as we estimate θ using larger samples, the estimation error $|\hat{\theta} - \theta|$ is increasingly unlikely to exceed ε , and this probability decreases as the sample size grows. The consistency of the sample mean

\bar{X} can be derived using Chebyshev's inequality. To apply this inequality, we first compute the variance of \bar{X} :

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{\text{Var}(X_1) + \dots + \text{Var}(X_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Using Chebyshev's Inequality

Applying Chebyshev's inequality to the random variable \bar{X} , we have:

$$\mathbb{P}(|\bar{X} - \mu| > \varepsilon) \leq \frac{\text{Var}(\bar{X})}{\varepsilon^2} = \frac{\sigma^2/n}{\varepsilon^2} \rightarrow 0$$

as $n \rightarrow \infty$. Therefore, the sample mean is consistent: as the sample size increases, the likelihood of a large sampling error decreases.

Asymptotic Normality

According to the Central Limit Theorem, the sum of observations, and hence the sample mean, approaches a normal distribution if calculated from a sufficiently large sample. Specifically, the distribution of

$$Z := \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

converges to the standard normal distribution as n increases. This property is sometimes referred to as **Asymptotic Normality**.

NOTATION:

μ = population mean

\bar{X} = sample mean, estimator of μ

σ = population standard deviation

s = sample standard deviation, estimator of σ

σ^2 = population variance

s^2 = sample variance, estimator of σ^2

Median

- ▶ One limitation of the sample mean is its susceptibility to outliers. For instance, if the first job in our dataset is exceptionally demanding, taking 30 minutes to process instead of the usual 70 seconds, this extreme value can inflate the sample mean from 48.2333 seconds to 105.9 seconds.
- ▶ This raises the question: can we consider such an estimator as "reliable"?
- ▶ An alternative measure of central tendency is the [sample median](#), which serves to estimate the population median.
- ▶ Unlike the sample mean, the median is considerably less affected by extreme observations.

Definition of Median

Definition 5

The median represents a "central" value.

The sample median \widehat{M} is the value such that at most half of the observations are greater than it and at most half are less than it. The population median M is defined as the value that has a probability of being exceeded no greater than 0.5 and a probability of being preceded no greater than 0.5. That is, M satisfies:

$$\begin{cases} \mathbb{P}(X > M) \leq 0.5 \\ \mathbb{P}(X \leq M) \leq 0.5 \end{cases}$$

Mean vs Median

Mean: The sum of all values divided by the number of values.

$$\text{Mean} = \frac{\sum_{i=1}^n x_i}{n}$$

- ▶ Sensitive to outliers.

Median: The middle value when data is ordered.

- ▶ Not affected by outliers.
- ▶ For an odd number of values: middle value.
For an even number: average of the two middle values.

Example

Given the data set:

1, 2, 3, 4, 100

- ▶ **Mean:**

$$\frac{1 + 2 + 3 + 4 + 100}{5} = 22$$

- ▶ **Median:** Since there are 5 values, the middle one is 3.

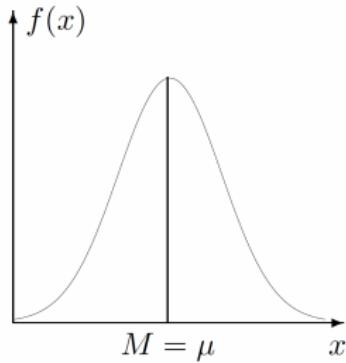
Conclusion: The mean is distorted by the outlier 100, while the median provides a better central value.

Understanding the Shape of a Distribution

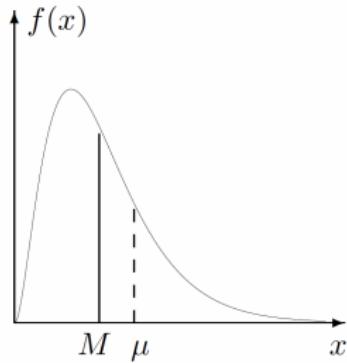
By comparing the mean μ and the median M , we can determine the skewness of the distribution of X :

- Symmetric distribution $\Rightarrow M = \mu$
- Right-skewed distribution $\Rightarrow M < \mu$
- Left-skewed distribution $\Rightarrow M > \mu$

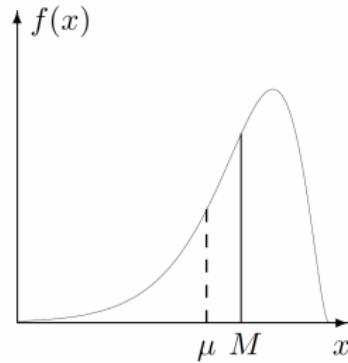
(a) symmetric



(b) right-skewed



(c) left-skewed



Mean and median [Baron]

Computation of the Population Median

For continuous distributions, calculating the population median involves solving the following:

$$\begin{cases} \mathbb{P}(X > M) = 1 - F(M) \leq 0.5 \\ \mathbb{P}(X < M) = F(M) \leq 0.5 \end{cases} \Rightarrow F(M) = 0.5$$

where F is the cumulative distribution function.

See Seminar 1 for examples.

Quantiles, Percentiles, and Quartiles

To generalize the concept of a median, we replace 0.5 in the previous definition with a value $0 < p < 1$.

Definition 6

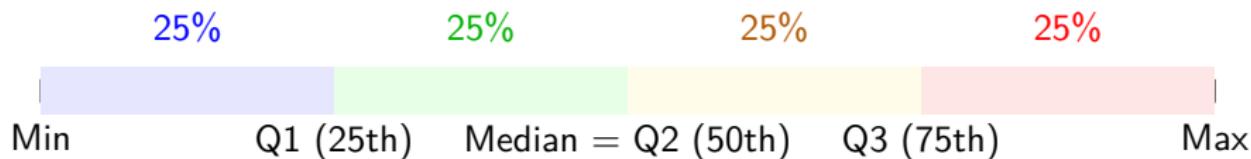
A p -quantile of a population is a number x that satisfies the following conditions:

$$\begin{cases} \mathbb{P}(X \leq x) \leq p \\ \mathbb{P}(X > x) \leq 1 - p. \end{cases}$$

Sample p-Quantile and Percentiles

- ▶ A sample p -quantile is any number that exceeds at most $100p\%$ of the sample and is exceeded by at most $100(1 - p)\%$.
- ▶ A γ -percentile is defined as the (0.01γ) -quantile.
- ▶ The first, second, and third quartiles are the 25th, 50th, and 75th percentiles, respectively, which divide a population or sample into four equal parts.
- ▶ A median serves as a 0.5-quantile, the 50th percentile, and the second quartile.

Sample p-Quantile and Percentiles



- ▶ A *p*-quantile is a cutoff point on this line. A cutoff point is the value that cuts the dataset so that a fraction p (or $100(1 - p)\%$) of the data is below it.
- ▶ Percentiles are just quantiles scaled to 100.
- ▶ Quartiles split the data into 4 equal parts.
- ▶ The median = 0.5-quantile = 50th percentile = Q2.

Example: p-Quantiles and Percentiles

Consider a sorted sample of queueing times (in minutes):

$$\mathcal{S} = \{5, 8, 12, 15, 18, 22, 25, 30, 35, 40\}$$

- ▶ 25th Percentile (First Quartile or 0.25-Quantile): The value that exceeds at most 25% of the sample is the 3rd value. Thus, the 25th percentile is **12**.
- ▶ 50th Percentile (Median or 0.5-Quantile): The value that exceeds at most 50% of the sample (the median) is the average of the 5th and 6th values:

$$\frac{18 + 22}{2} = \mathbf{20}.$$

- ▶ 75th Percentile (Third Quartile or 0.75-Quantile): The value that exceeds at most 75% of the sample is the 8th value, so the 75th percentile is **30**.

Notation for Quantiles and Quartiles

Notation:

q_p = population p -quantile

\hat{q}_p = sample p -quantile, estimator of q_p

π_γ = population γ -percentile

$\hat{\pi}_\gamma$ = sample γ -percentile, estimator of π_γ

Q_1, Q_2, Q_3 = population quartiles

$\hat{Q}_1, \hat{Q}_2, \hat{Q}_3$ = sample quartiles, estimators of Q_1, Q_2 , and Q_3

M = population median

\hat{M} = sample median, estimator of M

Example: p-Quantiles and Percentiles

See percentile calculation methods in Seminar 1.

- ▶ **Given:** A dataset of 10 ordered values:

12, 15, 18, 20, 22, 25, 28, 30, 35, 40

- ▶ **25th Percentile (Q1):**

- ▶ The 25th percentile corresponds to $p = 0.25$.
- ▶ Rank: $P_{25} = 0.25 \times (10 + 1) = 2.75$.
- ▶ Interpolating between the 2nd and 3rd values:

$$Q_1 = 15 + 0.75 \times (18 - 15) = 15 + 2.25 = 17.25$$

- ▶ Thus, the 25th percentile (Q1) is 17.25.
- ▶ **50th Percentile (Median, Q2):**

- ▶ The median corresponds to $p = 0.50$.
- ▶ Rank: $P_{50} = 0.50 \times (10 + 1) = 5.5$.
- ▶ Interpolating between the 5th and 6th values:

$$Q_2 = 22 + 0.5 \times (25 - 22) = 22 + 1.5 = 23.5$$

- ▶ Thus, the 50th percentile (Q2) is 23.5.

Example: p-Quantiles and Percentiles

► 75th Percentile (Q3):

- The 75th percentile corresponds to $p = 0.75$.
- Rank: $P_{75} = 0.75 \times (10 + 1) = 8.25$.
- Interpolating between the 8th and 9th values:

$$Q_3 = 30 + 0.25 \times (35 - 30) = 30 + 1.25 = 31.25$$

- Thus, the 75th percentile (Q3) is 31.25.

► Summary:

- $Q_1 = 17.25$
- $Q_2 = 23.5$ (Median)
- $Q_3 = 31.25$

- Quartiles divide the dataset into four equal parts.

Relationship Between Quantiles, Quartiles, and Percentiles

The relationships among quantiles, quartiles, and percentiles can be expressed as follows:

$$q_p = \pi_{100p}$$

$$Q_1 = \pi_{25} = q_{1/4} \quad Q_3 = \pi_{75} = q_{3/4}$$

$$M = Q_2 = \pi_{50} = q_{1/2}$$

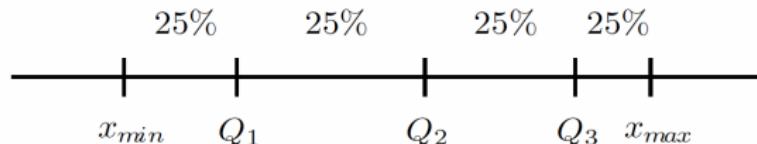
Let X be a random variable.

- ▶ Percentiles $\pi_1, \pi_2, \dots, \pi_{99}$ divide the data into 100 equal parts.
- ▶ For $k = 1, \dots, 99$, π_k satisfies:

$$\mathbb{P}(X < \pi_k) \approx \frac{k}{100} \approx \mathbb{P}(X \leq \pi_k).$$

- ▶ Quartiles Q_1, Q_2, Q_3 divide the data into 4 equal parts:

$$Q_1 = \pi_{25}, \quad Q_2 = \pi_{50}, \quad Q_3 = \pi_{75}.$$



Example: Quantiles, Quartiles, and Percentiles

- ▶ Consider a dataset of 12 ordered values:

5, 8, 10, 12, 15, 18, 20, 22, 25, 28, 30, 32

- ▶ **First Quartile Q_1 (25th Percentile, $q_{1/4}$):**

$$P_{25} = 0.25 \times (12 + 1) = 3.25$$

Interpolating between the 3rd and 4th values:

$$Q_1 = 10 + 0.25 \times (12 - 10) = 10.5$$

- ▶ **Second Quartile Q_2 (50th Percentile, Median, $q_{1/2}$):**

$$P_{50} = 0.50 \times (12 + 1) = 6.5$$

Interpolating between the 6th and 7th values:

$$Q_2 = 18 + 0.5 \times (20 - 18) = 19$$

Example: Quantiles, Quartiles, and Percentiles

- ▶ **Third Quartile Q_3 (75th Percentile, $q_{3/4}$):**

$$P_{75} = 0.75 \times (12 + 1) = 9.75$$

Interpolating between the 9th and 10th values:

$$Q_3 = 25 + 0.75 \times (28 - 25) = 27.25$$

- ▶ Summary of Results:

- ▶ $Q_1 = 10.5$
- ▶ $Q_2 = 19$
- ▶ $Q_3 = 27.25$

Interquartile Range and Related Measures

Let X be the studied variable with quartiles Q_1 , Q_2 , and Q_3 .

- ▶ The **Interquartile Range (IQR)** is the difference between the third and the first quartile:

$$\text{IQR} = Q_3 - Q_1$$

- ▶ The **Interquartile Deviation (IQD)** or the **Semi-Interquartile Range** is the value

$$\text{IQD} = \frac{\text{IQR}}{2} = \frac{Q_3 - Q_1}{2}$$

- ▶ The **Interquartile Deviation Coefficient (IQDC)** or the **Relative Interquartile Deviation** is the value

$$\text{IQDC} = \frac{\text{IQD}}{Q_2} = \frac{Q_3 - Q_1}{2Q_2}.$$

Properties of the Interquartile Deviation and Coefficient

► **Interquartile Deviation (IQD):**

- The IQD is an absolute measure of variation.
- Important property: the range $Q_2 \pm \text{IQD}$ contains approximately 50% of the data.

► **Interquartile Deviation Coefficient (IQDC):**

- The IQDC ranges from -1 to 1 .
- Values close to 0 indicate a symmetrical distribution with little variation.
- Values close to ± 1 indicate skewed data with large variation.

Outliers

An **outlier** is an *atypical* value, a value which is “far away” from the rest of the data, that does not seem to belong to the distribution of the rest of the values in the data set. For example, in a set of data where all values but one are between 0 and 1, a value of 100 is an outlier.

- ▶ A **mild outlier** is any data point outside the range:

$$[Q_1 - 1.5 \times (Q_3 - Q_1), \quad Q_3 + 1.5 \times (Q_3 - Q_1)]$$

- ▶ An **extreme outlier** is any data point outside the range:

$$[Q_1 - 3 \times (Q_3 - Q_1), \quad Q_3 + 3 \times (Q_3 - Q_1)]$$

Outliers may occur for two main reasons: they can be legitimate observations with values that are simply unusually high or low compared to the rest of the data, or they may result from errors in measurement, poor experimental methods, or mistakes in recording or entering data. Regardless of the cause, outliers can distort certain measures of central tendency and variability, potentially leading to incorrect conclusions in data analysis.

Variance and Standard Deviation

- ▶ The statistics discussed so far provide insights into the average value and specific percentiles within a population.
- ▶ We will also focus on quantifying the variability of our variable—specifically, how much it can fluctuate and how significantly the actual values can deviate from their expected values.
- ▶ This will allow us to evaluate the reliability of our estimates and the precision of our predictions.

Definition of Sample Variance

Definition 7

For a sample (X_1, X_2, \dots, X_n) , the **sample variance** is defined as follows:

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

This formula quantifies the variability among the observations and serves as an estimator for the population variance $\sigma^2 = \text{Var}(X)$.

Sample Standard Deviation

The sample standard deviation is simply the square root of the sample variance:

$$s = \sqrt{s^2}$$

This measure of variability is expressed in the same units as the variable X and estimates the population standard deviation $\sigma = \text{Std}(X)$.

Why Divide by $n - 1$?

Variance of a Population vs. Sample Variance

- ▶ The variance of a population σ^2 is given by:

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

- ▶ In practice, we often deal with samples from a population. The sample variance s^2 is:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- ▶ Dividing by $n - 1$ corrects the bias caused by using the sample mean \bar{x} instead of the population mean μ . This is called **Bessel's correction**.

Degrees of Freedom:

- ▶ The degrees of freedom refer to the number of independent values in a dataset.
- ▶ Since the sample mean \bar{x} is calculated from the data, only $n - 1$ values are free to vary. Hence, we divide by $n - 1$ to account for this loss in freedom.

Why Divide by $n - 1$?

Example:

- ▶ Consider the dataset: $\{2, 4, 6\}$.
- ▶ The sample mean \bar{x} is:

$$\bar{x} = \frac{2 + 4 + 6}{3} = 4$$

- ▶ The squared deviations from the sample mean are:

$$(2 - 4)^2 = 4, \quad (4 - 4)^2 = 0, \quad (6 - 4)^2 = 4$$

- ▶ The sample variance s^2 is:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{2} \cdot (4 + 0 + 4) = 4$$

- ▶ If we divided by n instead of $n - 1$, the variance would be $s^2 = 2.66$, underestimating the true variability in the population.

Why Divide by $n - 1$?

Why $n - 1$ Avoids Underestimation:

- ▶ Dividing by n leads to a biased estimate of the population variance, as it tends to underestimate the spread.
- ▶ This bias occurs because the sample mean \bar{x} is already fitted to the data, so the deviations $(x_i - \bar{x})$ tend to be smaller than deviations from the true population mean μ .
- ▶ The correction by $n - 1$ adjusts for this and provides an **unbiased** estimator for the population variance.

LECTURES 2 and 3

Parameter Estimation

Introduction to Parameter Estimation

Parameter estimation is a fundamental aspect of statistics, where we aim to **infer the values of population parameters based on sample data**. The key concepts include:

- ▶ **Population parameter:** A characteristic or measure of a whole population (e.g., mean μ , variance σ^2).
- ▶ **Sample statistic:** A characteristic or measure **computed** from a sample (e.g., sample mean \bar{x} , sample variance s^2).

When we sample from a population described by a pdf $f(x, \theta)$, the parameter θ yields knowledge of the entire population, so a good estimator for θ improves our knowledge about the whole population.

Types of Estimators

Estimators can be categorized as:

- ▶ **Point estimators:** A single value estimate of a population parameter.
 - ▶ Example: Sample mean as an estimate of the population mean:

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- ▶ **Interval estimators:** A range of values within which the parameter is expected to lie.
 - ▶ Example: Confidence interval for mean:

$$\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Properties of Estimators

A good estimator should possess the following properties:

- ▶ **Unbiasedness:**

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

where $\hat{\theta}$ is the estimator and θ is the true parameter.

- ▶ **Consistency:**

$$\lim_{n \rightarrow \infty} \hat{\theta} = \theta \text{ in probability}$$

i.e. the estimator converges to the true parameter as the sample size increases.

- ▶ **Efficiency:**

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta}) \quad \forall \tilde{\theta}$$

where $\tilde{\theta}$ is any other unbiased estimator, i.e. the estimator has the smallest variance among all unbiased estimators.

Method of Moments

Moments

Definition The k -th population moment is defined as:

$$\mu_k = \mathbb{E}[X^k].$$

The k -th sample moment is defined as:

$$m_k = \frac{1}{n} \sum_{i=1}^n X_i^k.$$

Remark: Note that m_k estimates μ_k from a sample (X_1, \dots, X_n) , and the first sample moment is the sample mean \bar{X} .

Central Moments

Central moments are computed by *centralizing the data* that is, by subtracting the mean.

Definition For $k \geq 2$, the k -th population central moment is defined as:

$$\mu'_k = \mathbb{E}[(X - \mu)^k]$$

The k -th sample central moment:

$$m'_k = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k$$

estimates μ'_k from a sample (X_1, \dots, X_n) .

Second Central Moment: Variance

Remark: The second population central moment is the variance $\text{Var}(X)$. The second sample central moment is the sample variance, although the denominator $(n - 1)$ is now replaced by n .

For an unbiased estimation of $\sigma^2 = \text{Var}(X)$, we use:

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

Estimation via Method of Moments

The method of moments is based on the fact that, since our sample comes from a family of distributions $(F(\theta))_\theta$, we choose such a member of this family whose properties are close to those of our data. That is, we *match the moments*.

To estimate k parameters, we form the following system, by equating the first k population and sample moments:

$$\left\{ \begin{array}{rcl} \mu_1 & = & m_1 \\ \mu_2 & = & m_2 \\ \dots & \dots & \dots \\ \mu_k & = & m_k \end{array} \right.$$

The left-hand sides depend on the distribution parameters (θ) , while the right-hand sides can be *computed from the data*. The "method of moments estimator" is the solution of this system.

Example: Poisson Distribution

To estimate the parameter λ of a $\text{Poisson}(\lambda)$ distribution, recall that:

$$\mu_1 = \mathbb{E}[X] = \lambda.$$

There is only one unknown parameter, so we write the equation:

$$\mu_1 = m_1 \Leftrightarrow \lambda = \bar{X}.$$

Solving it for λ we obtain:

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

→ the *method of moments estimator* for λ is \bar{X} in this case.

Example: Gamma Distribution of CPU Times

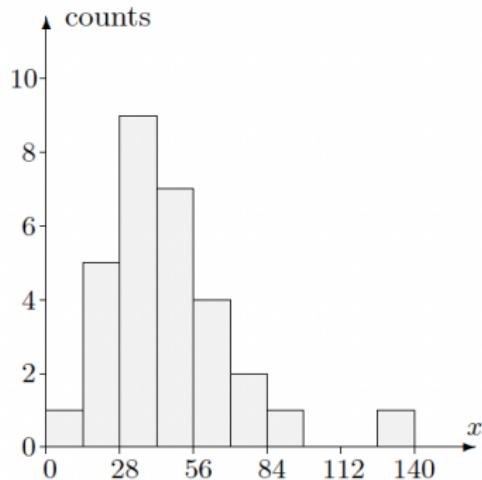
Suppose we collect a random sample to evaluate effectiveness of a processor for a certain type of tasks

$$\mathcal{S} = (X_1, X_2, \dots, X_n).$$

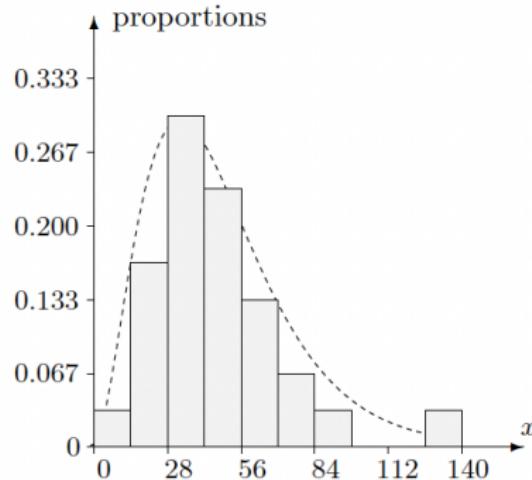
We record the CPU time in seconds for $n = 30$ randomly chosen jobs:

70	36	43	69	82	48	34	62	35	15
59	139	46	37	42	30	55	56	36	82
38	89	54	25	35	24	22	9	56	19.

The frequency and relative frequency histograms are below. The frequency histogram contains one column for each bin, with the height determined by the number of observations in the bin, while the relative frequency histogram contains columns with heights given by the proportion of all data that appears in each bin (i.e. on the right each count is divided by the sample size $n = 30$).



(a) Frequency histogram



(b) Relative frequency histogram

Histograms for CPU data [Baron]

Example: Gamma Distribution of CPU Times

The histogram for this suggests that CPU times have a Gamma distribution with parameters α and λ .

To estimate these parameters, we need two equations. From the data above, we compute:

$$m_1 = \bar{X} = 48.2333 \quad \text{and} \quad m'_2 = s^2 = 679.7122$$

Using the second central moment, we write two equations, as $\mu = \frac{\alpha}{\lambda}$ and $\text{Var}(X) = \frac{\alpha}{\lambda^2}$ are known (see lecture notes Probability Theory) for a Gamma variable:

$$\begin{cases} \mu_1 &= m_1 \\ \mu_2 &= m'_2 \end{cases}$$

that is

$$\begin{cases} \frac{\alpha}{\lambda} &= \bar{X} \\ \frac{\alpha}{\lambda^2} &= s^2 \end{cases}$$

Gamma Distribution Parameter Estimation

Solving the system of equations in terms of α and λ , we obtain the method of moments estimates:

$$\begin{cases} \hat{\alpha} = 3.4227 \\ \hat{\lambda} = 0.0710 \end{cases}$$

These estimates provide the parameters for the Gamma distribution of CPU times.

Example: Normal Distribution

Suppose we already know the mean μ of a Normal distribution and would like to estimate the variance σ^2 .

Only one parameter σ^2 is unknown, but the first method of moments equation:

$$\mu_1 = m_1$$

does not involve σ^2 . Therefore, it does not produce an estimate for it. We then use the second equation:

$$\mu'_2 = \sigma^2 = m'_2 = s^2$$

which gives us the method of moments estimate: $\hat{\sigma}^2 = s^2$.

Method of moments estimates are typically easy to compute and they serve as a quick tool for parameter estimation. However, these estimators are not always optimal, so we need also other methods.

Binomial Method of Moments

Binomial Model: Let X_1, \dots, X_n be iid binomial with parameters (k, p) :

$$\mathbb{P}(X_i = x \mid k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, \dots, k.$$

- ▶ Here, both k and p are unknown.
- ▶ This approach can estimate crime rates where the reporting rate p and total number of occurrences k are unknown.

Solving for Parameters k and p

Moment Equations:

$$\bar{X} = kp$$

$$\frac{1}{n} \sum X_i^2 = kp(1 - p) + k^2 p^2$$

- ▶ Solving this system yields the **method of moments estimators**:

$$\tilde{k} = \frac{\bar{X}^2}{\bar{X} - (1/n) \sum (X_i - \bar{X})^2}$$

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

Conclusion: These estimators \tilde{k} and \tilde{p} provide estimates for the unknown parameters k and p using the method of moments.

Evaluating the Method of Moments Estimators

- ▶ The method of moments estimators for k and p are not always optimal:
 - ▶ Possible to obtain negative estimates of k and p , which contradicts the positive nature of these parameters.
 - ▶ This arises when the sample mean is smaller than the sample variance, indicating high data variability.
- ▶ Despite limitations, the method of moments offers viable candidate estimators:
 - ▶ It provides a structured approach to point estimation, particularly useful when an intuitive estimator for k is challenging to construct.
- ▶ Practical use of the method of moments includes “moment matching”:
 - ▶ This technique approximates a distribution by aligning sample moments with those of a theoretical distribution.
 - ▶ In practice, moment matching is most effective with distributions that closely resemble each other.

Maximum-Likelihood Estimator: Background

- ▶ For a fixed random sample from a probability distribution, the maximum likelihood method selects the values of the population parameters that make the observed data "more likely" than any other values.
- ▶ Given an observed sample $\mathbf{X} = (X_1, \dots, X_n)$, with the method of Maximum Likelihood Estimation (MLE) we aim to identify parameter values that maximize the likelihood of this sample, that is, to find parameters θ that maximize the probability (or likelihood) of observing \mathbf{X} .
 - ▶ **Discrete case:** Maximize joint probability mass function $\mathbb{P}(X_1, \dots, X_n)$.
 - ▶ **Continuous case:** Maximize joint probability density function $f(X_1, \dots, X_n)$.

Likelihood Function

- ▶ X_1, \dots, X_n are **independent** sample variables for the variable X ;
- ▶ x_1, \dots, x_n are the statistical data (i.e. x_i are realizations of X_i);
- ▶ θ is the unknown parameter to be estimated.

Definition 8

The likelihood function is defined as:

$$L(x_1, \dots, x_n; \theta) = \begin{cases} \mathbb{P}(X = x_1) \cdot \dots \cdot \mathbb{P}(X = x_n), & \text{if } X \text{ is discrete} \\ f(x_1) \cdot \dots \cdot f(x_n), & \text{if } X \text{ is continuous with pdf } f \end{cases}$$

i.e. this tells us how likely it is to observe the data set x_1, x_2, \dots, x_n under specific values of the parameter θ .

- ▶ X is characterized by a pmf or a pdf, depending on θ :

$$f(x; \theta) = \begin{cases} \mathbb{P}(X = x), & \text{if } X \text{ is discrete} \\ f(x), & \text{if } X \text{ is continuous with pdf } f. \end{cases}$$

Likelihood Function Example

The likelihood function can also be rewritten as:

$$L(x_1, \dots, x_n; \theta) = f(x_1; \theta) \cdot \dots \cdot f(x_n; \theta)$$

where $f(x_i; \theta)$ describes the probability of observing the value x_i given the parameter μ .

Example: Consider $X \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, i.e., $X \sim \text{Bernoulli}(p)$, with statistical data 0, 1, 1, 0, 0, 0, 1, 0, 1, 0.

$$\Rightarrow n = 10, \quad x_1 = 0, \quad x_2 = 1, \quad \dots, \quad x_{10} = 0.$$

Since $\mathbb{P}(X = x) = p^x(1 - p)^{1-x}$ for $x \in \{0, 1\}$, the likelihood function becomes:

$$\begin{aligned} L(x_1, \dots, x_{10}; p) &= \mathbb{P}(X = x_1) \cdots \mathbb{P}(X = x_{10}) \\ &= p^{x_1 + \dots + x_{10}} (1 - p)^{10 - (x_1 + \dots + x_{10})}. \end{aligned}$$

Maximum Likelihood Estimator (MLE)

We estimate the parameters by maximizing the likelihood function:

$$L(\mathbf{x}; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

where $f(x_i; \theta)$ is the probability density function (pdf) or probability mass function (pmf), corresponding to independent data points x_1, \dots, x_n .

- The log-likelihood function is given by:

$$\ell(\mathbf{x}; \theta) = \ln L(\mathbf{x}; \theta) = \sum_{i=1}^n \ln f(x_i; \theta)$$

- The MLE is found by solving:

$$\frac{\partial \ell(\mathbf{x}; \theta)}{\partial \theta} = 0$$

and

$$\frac{\partial^2 \ell(\mathbf{x}; \theta)}{\partial \theta^2} < 0.$$

MLE: Single Parameter Case

The parameter θ is a maximum for L if

$$\frac{\partial L(\cdot; \theta)}{\partial \theta} = 0 \text{ and } \frac{\partial^2 L(\cdot; \theta)}{\partial \theta^2} < 0,$$

or equivalently,

$$\frac{\partial \ln L(\cdot; \theta)}{\partial \theta} = 0 \text{ and } \frac{\partial^2 \ln L(\cdot; \theta)}{\partial \theta^2} < 0.$$

In cases where $\frac{\partial L}{\partial \theta} = 0$ (or equivalently $\frac{\partial \ln L}{\partial \theta} = 0$) has no solution, alternative methods may be used.

Log-Likelihood Maximization

- ▶ The second condition ensures that ℓ is a concave function, therefore the stationary point found due to the first condition is indeed a local maximum.
- ▶ Maximizing the likelihood function $L(x; \theta)$ directly can be cumbersome due to multiplicative terms, especially in large samples. The log-likelihood,

$$\ell(x; \theta) = \log L(x; \theta),$$

transforms this product into a sum, simplifying differentiation and ensuring numerical stability.

- ▶ Since the logarithm is a monotonic function, maximizing $\ell(x; \theta)$ is equivalent to maximizing $L(x; \theta)$.

Example: Estimating the Mean of a Normal Distribution

Suppose we have n observations x_1, x_2, \dots, x_n that are independently and identically distributed (i.i.d.) from a normal distribution with mean μ and variance σ^2 . The probability density function (pdf) for each observation x_i is:

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

- ▶ The likelihood function for μ is:

$$L(\mathbf{x}; \mu) = \prod_{i=1}^n f(x_i; \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right).$$

- ▶ Taking the log of the likelihood function to simplify the calculations, we get:

$$\ell(\mathbf{x}; \mu) = \sum_{i=1}^n \ln f(x_i; \mu) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2.$$

- To find the MLE, we take the derivative with respect to μ , set it to zero, and solve for μ :

$$\frac{\partial \ell(\mathbf{x}; \mu)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0.$$

Solving this, we find:

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

- We check the second derivative:

$$\frac{\partial^2 \ell(\mathbf{x}; \mu)}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0.$$

This confirms that $\hat{\mu}$ is indeed a maximum. So the MLE for μ is the sample mean $\hat{\mu}$

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i.$$

MLE: Multiple Parameters Case

In the case of k unknown parameters $(\theta_1, \dots, \theta_k)$, one solves the system

$$\frac{\partial L}{\partial \theta_j} = 0, \quad j = 1, \dots, k,$$

and shows that the matrix

$$\left(\frac{\partial^2 L}{\partial \theta_i \partial \theta_j} \right)_{1 \leq i, j \leq k}$$

is negative definite.¹ Similarly, one can use

$$\frac{\partial \ln L}{\partial \theta_j} = 0, \quad j = 1, \dots, k,$$

and show that the matrix

$$\left(\frac{\partial^2 \ln L}{\partial \theta_i \partial \theta_j} \right)_{1 \leq i, j \leq k}$$

is also negative definite.

¹A matrix M is negative definite if $y^t M y < 0$ for each $y \in \mathbb{R}^k \setminus \{0_k\}$ (equivalent to M having only strictly negative eigenvalues).

Likelihood function of a sample (multi-dim case)

Definition 9

The likelihood function for a sample is:

$$L(X_1, \dots, X_n; \Theta) = \prod_{i=1}^n f(X_i; \Theta),$$

where $\Theta = (\theta_1, \dots, \theta_k)$ is the vector of unknown parameters.

MLE selects values for an estimator $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ that maximizes $L(X_1, \dots, X_n; \Theta)$. To find the maximum likelihood estimates, we solve:

$$\frac{\partial L(X_1, \dots, X_n; \theta_1, \dots, \theta_k)}{\partial \theta_j} = 0, \quad j = 1, \dots, k.$$

Alternatively, maximizing $\ln L$ is computationally simpler:

$$\frac{\partial \ln L(X_1, \dots, X_n; \theta_1, \dots, \theta_k)}{\partial \theta_j} = 0, \quad j = 1, \dots, k.$$

Example - 1-dim case

Consider the pdf

$$f(x; \theta) = \frac{1}{\theta^2} xe^{-\frac{x}{\theta}}, \quad x > 0, \quad \theta > 0,$$

with θ as the unknown parameter.

The likelihood function is given by

$$\begin{aligned} L(X_1, \dots, X_n; \theta) &= \prod_{i=1}^n \left(\frac{1}{\theta^2} X_i e^{-\frac{X_i}{\theta}} \right) \\ &= \left(\prod_{i=1}^n X_i \right) \frac{1}{\theta^{2n}} e^{-\frac{1}{\theta} \sum_{i=1}^n X_i} \\ &= K \frac{1}{\theta^{2n}} e^{-\frac{n\bar{X}}{\theta}}, \end{aligned}$$

where $K = \prod_{i=1}^n X_i$ is a constant with respect to θ .

Finding the MLE

Take the logarithm of the likelihood function for easier computation:

$$\ln L = \ln K - 2n \ln \theta - \frac{n\bar{X}}{\theta}$$
$$\frac{\partial \ln L}{\partial \theta} = -\frac{2n}{\theta} + \frac{n\bar{X}}{\theta^2}.$$

Setting the derivative to zero gives:

$$-\frac{2n}{\theta} + \frac{n\bar{X}}{\theta^2} = 0 \implies \hat{\theta} = \frac{1}{2}\bar{X},$$

which is the same as the method of moments estimator (check!).

Note that $\hat{\theta}$ yields a maximum for $\ln L$ since:

- ▶ $\frac{\partial \ln L(\cdot; \theta)}{\partial \theta} > 0$ for $\theta < \hat{\theta}$ (increasing)
- ▶ $\frac{\partial \ln L(\cdot; \theta)}{\partial \theta} < 0$ for $\theta > \hat{\theta}$ (decreasing)

Check also the 2nd derivative!

Monotonicity (increasing / decreasing). We can write,

$$\frac{\partial \ln L}{\partial \theta} = \frac{n}{\theta^2} (\bar{X} - 2\theta),$$

since $n > 0$ and $\theta^2 > 0$,

$$\begin{cases} \frac{\partial \ln L}{\partial \theta} > 0 & \text{if } \bar{X} - 2\theta > 0 \quad (\theta < \frac{1}{2}\bar{X}), \\ \frac{\partial \ln L}{\partial \theta} < 0 & \text{if } \bar{X} - 2\theta < 0 \quad (\theta > \frac{1}{2}\bar{X}). \end{cases}$$

Thus $\ln L$ is increasing for $\theta < \hat{\theta}$ and decreasing for $\theta > \hat{\theta}$, and therefore we have a (local) maximum at $\hat{\theta}$.

Second derivative test. Differentiate the first derivative:

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{d}{d\theta} \left(-\frac{2n}{\theta} + \frac{n\bar{X}}{\theta^2} \right) = \frac{2n}{\theta^2} - \frac{2n\bar{X}}{\theta^3}.$$

That is:

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{2n}{\theta^3} (\theta - \bar{X}).$$

Evaluate at $\theta = \hat{\theta} = \frac{1}{2}\bar{X}$:

$$\theta - \bar{X} \Big|_{\theta=\hat{\theta}} = \frac{1}{2}\bar{X} - \bar{X} = -\frac{1}{2}\bar{X},$$

so

$$\frac{\partial^2 \ln L}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} = \frac{2n}{\hat{\theta}^3} \left(-\frac{1}{2}\bar{X} \right) = -\frac{n\bar{X}}{\hat{\theta}^3}.$$

Substitute $\hat{\theta} = \frac{1}{2}\bar{X}$ to obtain a simple numerical expression:

$$\frac{\partial^2 \ln L}{\partial \theta^2} \Big|_{\theta=\hat{\theta}} = -\frac{n\bar{X}}{\left(\frac{1}{2}\bar{X}\right)^3} = -\frac{n\bar{X}}{\frac{1}{8}\bar{X}^3} = -\frac{8n}{\bar{X}^2} < 0.$$

\Rightarrow The second derivative at $\hat{\theta}$ is strictly negative (assuming $\bar{X} \neq 0$), so $\hat{\theta} = \frac{1}{2}\bar{X}$ is indeed a (strict) local maximum. Combined with the monotonicity argument, it is the global maximizer on $\theta > 0$.

Example: Poisson Distribution

Poisson pmf:

$$P(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Logarithm of pmf:

$$\ln P(x) = -\lambda + x \ln \lambda - \ln(x!)$$

Log-Likelihood Function:

$$\ln P(\mathbf{X}) = \sum_{i=1}^n (-\lambda + X_i \ln \lambda) + C = -n\lambda + \ln \lambda \sum_{i=1}^n X_i + C$$

where $C = -\sum \ln(x!)$ is constant, independent of λ .

Finding the Critical Point:

$$\frac{\partial}{\partial \lambda} \ln P(\mathbf{X}) = -n + \frac{1}{\lambda} \sum_{i=1}^n X_i = 0$$

Solving gives:

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

Since this is the only critical point (check!) and the likelihood vanishes as $\lambda \rightarrow 0$ or $\lambda \rightarrow \infty$, we conclude $\hat{\lambda}$ is the maximizer, i.e., the MLE for λ .

Remark: For the Poisson distribution, the method of moments and the method of maximum likelihood both yield $\hat{\lambda} = \bar{X}$.

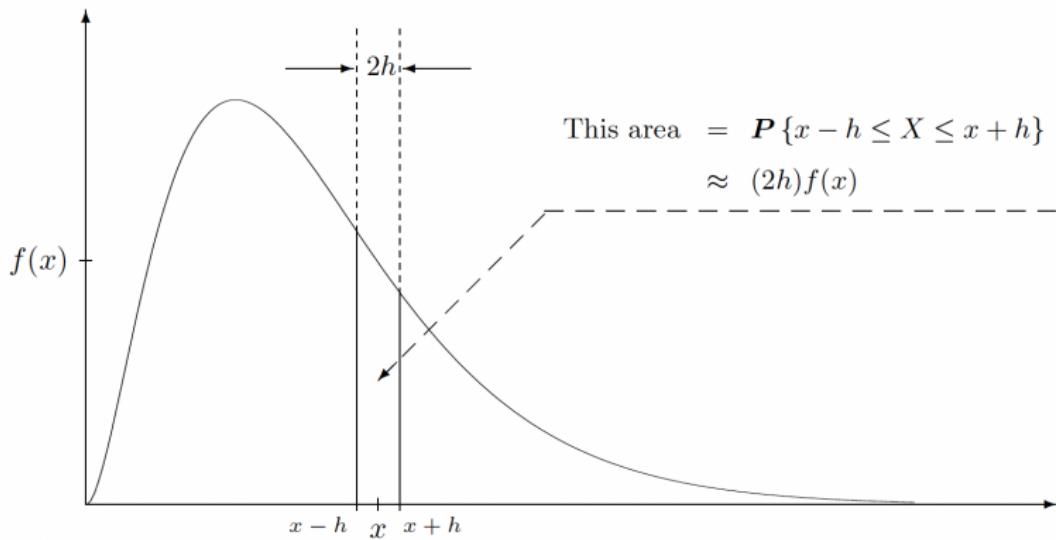
Maximum Likelihood Estimation - Continuous Case

- ▶ In the continuous case, the probability of observing an exact value $X = x$ is zero. Instead, MLE maximizes the probability of observing values close to x .
- ▶ For a small h ,

$$\mathbb{P}(x - h < X < x + h) = \int_{x-h}^{x+h} f(y) dy \approx (2h)f(x),$$

where the probability of observing a value near x is proportional to the density $f(x)$.

- ▶ For a sample $\mathbf{X} = (X_1, \dots, X_n)$, MLE maximizes the joint density $f(X_1, \dots, X_n)$.



Probability of observing almost X=x [Baron]

Example: Exponential Distribution

Exponential Density:

$$f(x) = \lambda e^{-\lambda x}$$

Log-Likelihood of a Sample:

$$\ln f(\mathbf{X}) = \sum_{i=1}^n \ln(\lambda e^{-\lambda X_i}) = \sum_{i=1}^n (\ln \lambda - \lambda X_i) = n \ln \lambda - \lambda \sum_{i=1}^n X_i$$

Finding the MLE for λ :

- ▶ Taking the derivative with respect to λ and setting it to zero:

$$\frac{\partial}{\partial \lambda} \ln f(\mathbf{X}) = \frac{n}{\lambda} - \sum_{i=1}^n X_i = 0$$

Solving, we find:

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

- ▶ Since $\hat{\lambda}$ is the only critical point and $f(\mathbf{X})$ vanishes as $\lambda \rightarrow 0$ or $\lambda \rightarrow \infty$, $\hat{\lambda} = \bar{X}$ is the MLE of λ .

Example: Uniform Distribution

Uniform $(0, b)$ Density:

$$f(x) = \frac{1}{b} \quad \text{for } 0 \leq x \leq b$$

- ▶ To maximize the likelihood of a sample (X_1, \dots, X_n) , we consider the joint density:

$$f(X_1, \dots, X_n) = \left(\frac{1}{b}\right)^n \quad \text{for } 0 \leq X_1, \dots, X_n \leq b$$

- ▶ This density is maximized at the smallest possible b that still satisfies $b \geq X_i$ for all i . Hence, $b \geq \max(X_i)$.
- ▶ If $b < \max(X_i)$, then $f(\mathbf{X}) = 0$, which cannot be a maximum.
- ▶ Therefore, the **maximum likelihood estimator** for b is:

$$\hat{b} = \max(X_i)$$

Note: When estimating multiple parameters, all partial derivatives should equal zero at the critical point. If no critical points exist, the likelihood is maximized at the boundary.

MLEs in the Multivariate Case

Invariance Property of MLEs (Multivariate Case)

If the MLE of

$$\Theta = (\theta_1, \dots, \theta_k)$$

is

$$\hat{\Theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k),$$

and if $\tau(\Theta)$ is any function of the parameters, then the MLE of $\tau(\Theta)$ is

$$\hat{\tau} = \tau(\hat{\Theta}).$$

Example - Normal MLEs, μ and σ^2 Unknown

Let X_1, \dots, X_n be i.i.d. $\mathcal{N}(\theta, \sigma^2)$, with both θ and σ^2 unknown.

The likelihood function is:

$$L(\theta, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right].$$

The log-likelihood is:

$$\ln L(\theta, \sigma^2 | \mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2.$$

Computing the Partial Derivatives

Partial derivative with respect to θ :

$$\frac{\partial \ln L(\theta, \sigma^2 | \mathbf{x})}{\partial \theta} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta).$$

Setting to zero:

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

Partial derivative with respect to σ^2 :

$$\frac{\partial \ln L(\theta, \sigma^2 | \mathbf{x})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \theta)^2.$$

Setting to zero:

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^2 = 0 \quad \Rightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Verifying the maximum via successive maximization

If $\theta \neq \bar{x}$, then

$$\sum_{i=1}^n (x_i - \theta)^2 > \sum_{i=1}^n (x_i - \bar{x})^2.$$

Thus, for any σ^2 ,

$$\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2} \geq \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum (x_i - \theta)^2}.$$

Therefore, verifying the MLEs reduces to a one-dimensional problem:

$$g(\sigma^2) = (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum (x_i - \bar{x})^2\right].$$

Differentiating $g(\sigma^2)$:

$$\frac{\partial \ln g(\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i - \bar{x})^2 = 0.$$

Hence,

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2.$$

MLEs:

$$(\hat{\theta}, \hat{\sigma}^2) = \left(\bar{x}, \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right).$$

Partial Maximization Intuition

For two parameters $\Theta = (\theta, \sigma^2)$, the likelihood

$$L(\theta, \sigma^2 | \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2\right]$$

depends on both θ and σ^2 .

Finding the MLE requires maximizing $L(\theta, \sigma^2)$ *simultaneously* in both parameters. However, this is often simplified by the method of **successive (or partial) maximization**:

1. First, fix σ^2 and maximize the log-likelihood with respect to θ .
2. Substitute this $\hat{\theta}(\sigma^2)$ back into $\ln L(\theta, \sigma^2)$.
3. Then maximize the resulting expression (now only a function of σ^2) with respect to σ^2 .

This technique avoids having to solve the system of equations jointly and reduces a two-dimensional optimization problem to two one-dimensional problems.

Step 1: Maximizing with respect to θ (holding σ^2 fixed)

The log-likelihood is:

$$\ln L(\theta, \sigma^2 | \mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2.$$

Partial derivative with respect to θ :

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \theta).$$

Set to zero (necessary condition for an extremum):

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

For any fixed σ^2 , $\hat{\theta} = \bar{x}$ maximizes $\ln L$ in θ . Hence, we can now treat θ as fixed at \bar{x} and proceed to the next step.

Step 2: Maximizing with respect to σ^2 (profile likelihood)

Substitute $\theta = \bar{x}$ into the log-likelihood:

$$\ln L(\bar{x}, \sigma^2 | \mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2.$$

This function of σ^2 is known as the **profile log-likelihood** for σ^2 .

Derivative:

$$\frac{\partial \ln L(\bar{x}, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Setting to zero gives:

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^2 = 0 \quad \Rightarrow \quad \boxed{\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.}$$

Hence,

$$(\hat{\theta}, \hat{\sigma}^2) = \left(\bar{x}, \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right).$$

Why Successive Maximization Works

If $\theta \neq \bar{x}$, then

$$\sum_{i=1}^n (x_i - \theta)^2 > \sum_{i=1}^n (x_i - \bar{x})^2,$$

so the likelihood decreases when moving away from \bar{x} , *for any fixed σ^2* .

Thus, for any σ^2 ,

$$L(\bar{x}, \sigma^2) \geq L(\theta, \sigma^2),$$

meaning that the global maximization over (θ, σ^2) can be performed by first maximizing over θ , then over σ^2 .

This step-by-step (successive) maximization is justified because the likelihood separates cleanly in θ and σ^2 once the dependence is made explicit through $\sum(x_i - \theta)^2$.

Two-Variable Calculus Verification

To verify a local maximum of a function $H(\theta_1, \theta_2)$ at $(\hat{\theta}_1, \hat{\theta}_2)$, we require:

- a. **First-order conditions:**

$$\frac{\partial H}{\partial \theta_1} = 0, \quad \frac{\partial H}{\partial \theta_2} = 0.$$

- b. **At least one negative second-order partial derivative:**

$$\frac{\partial^2 H}{\partial \theta_j^2} < 0.$$

- c. **Positive Hessian determinant:**

$$\begin{vmatrix} H_{\theta_1 \theta_1} & H_{\theta_1 \theta_2} \\ H_{\theta_2 \theta_1} & H_{\theta_2 \theta_2} \end{vmatrix} = H_{\theta_1 \theta_1} H_{\theta_2 \theta_2} - (H_{\theta_1 \theta_2})^2 > 0.$$

Understanding the Hessian and the Maximum Condition

Intuition: Consider the log-likelihood (or any function) $H(\theta_1, \theta_2)$ as a surface over the two-dimensional parameter space. The point $(\hat{\theta}_1, \hat{\theta}_2)$ corresponds to a stationary point (where all first derivatives are zero). To determine whether it is a *maximum*, we examine the **curvature** of the surface.

The curvature is summarized by the **Hessian matrix**:

$$H = \begin{pmatrix} H_{\theta_1 \theta_1} & H_{\theta_1 \theta_2} \\ H_{\theta_2 \theta_1} & H_{\theta_2 \theta_2} \end{pmatrix} = \begin{pmatrix} \frac{\partial^2 H}{\partial \theta_1^2} & \frac{\partial^2 H}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 H}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 H}{\partial \theta_2^2} \end{pmatrix}.$$

Interpreting the Hessian Determinant

At a stationary point, the sign of the Hessian entries determines the type of extremum.

$$|H| = \begin{vmatrix} H_{\theta_1 \theta_1} & H_{\theta_1 \theta_2} \\ H_{\theta_2 \theta_1} & H_{\theta_2 \theta_2} \end{vmatrix} = H_{\theta_1 \theta_1} H_{\theta_2 \theta_2} - (H_{\theta_1 \theta_2})^2.$$

- ▶ If $|H| > 0$: curvature has the **same sign** in all directions. The surface is either concave (downward) or convex (upward).
- ▶ If $|H| < 0$: curvature has **opposite signs** in orthogonal directions. The surface forms a *saddle point*.

Condition	Interpretation	Stationary point
$H_{\theta_1 \theta_1} < 0, H > 0$	Downward curv. in all directions	Local maximum
$H_{\theta_1 \theta_1} > 0, H > 0$	Upward curv. in all directions	Local minimum
$ H < 0$	Opposite curv. (saddle)	No extremum

Application to the Normal Log-Likelihood

For the normal model:

$$\ln L(\theta, \sigma^2 | \mathbf{x}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \theta)^2.$$

The second derivatives are:

$$\frac{\partial^2 \ln L}{\partial \theta^2} = -\frac{n}{\sigma^2}, \quad \frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \theta)^2,$$

$$\frac{\partial^2 \ln L}{\partial \theta \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \theta).$$

At the MLEs $\hat{\theta} = \bar{x}$, $\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$, we have:

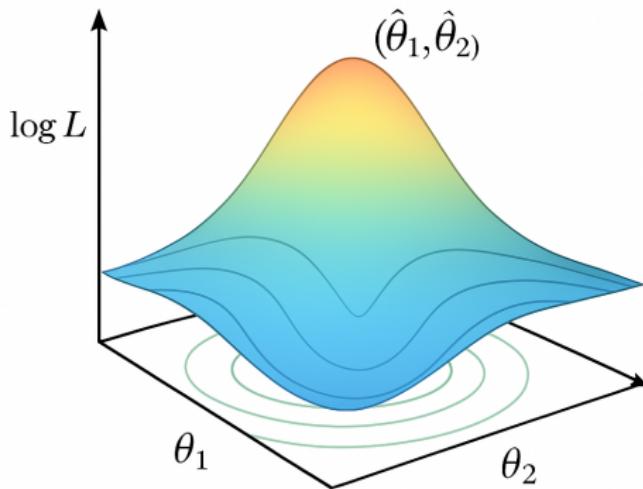
$$H_{\theta\theta} < 0, \quad |H| = \frac{n^2}{2\hat{\sigma}^6} > 0.$$

Hence, the log-likelihood is concave downward in both directions, confirming that $(\hat{\theta}, \hat{\sigma}^2)$ is a **local (and global) maximum**.

Geometric Intuition

At the maximum of the likelihood surface:

- ▶ The slope in both parameter directions is zero ($\frac{\partial \ln L}{\partial \theta_j} = 0$).
- ▶ The surface curves **downward** in every direction ($H_{\theta_j \theta_j} < 0$).
- ▶ The determinant $|H| > 0$ ensures this curvature is consistent in all directions, ruling out saddle points.



Normal Log-Likelihood Second Derivatives

For the normal case:

$$\frac{\partial^2 \ln L(\theta, \sigma^2 | \mathbf{x})}{\partial \theta^2} = -\frac{n}{\sigma^2},$$

$$\frac{\partial^2 \ln L(\theta, \sigma^2 | \mathbf{x})}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (x_i - \theta)^2,$$

$$\frac{\partial^2 \ln L(\theta, \sigma^2 | \mathbf{x})}{\partial \theta \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{i=1}^n (x_i - \theta).$$

At $(\hat{\theta}, \hat{\sigma}^2) = (\bar{x}, \hat{\sigma}^2)$, the cross derivative term vanishes:

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

Evaluating the Hessian Determinant

At $(\theta, \sigma^2) = (\bar{x}, \hat{\sigma}^2)$:

$$\begin{aligned}|H| &= \begin{vmatrix} \frac{-n}{\sigma^2} & -\frac{1}{\sigma^4} \sum (x_i - \theta) \\ -\frac{1}{\sigma^4} \sum (x_i - \theta) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum (x_i - \theta)^2 \end{vmatrix} \\ &= \frac{1}{\sigma^6} \left[-\frac{n^2}{2} + \frac{n}{\sigma^2} \sum (x_i - \theta)^2 - \frac{1}{\sigma^2} \left(\sum (x_i - \theta) \right)^2 \right].\end{aligned}$$

Substituting $\theta = \bar{x}$ and $\sigma^2 = \hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$:

$$|H| = \frac{1}{\hat{\sigma}^6} \left(-\frac{n^2}{2} + n^2 \right) = \frac{n^2}{2\hat{\sigma}^6} > 0.$$

Thus, the calculus conditions for a local maximum are satisfied.

Remarks

- ▶ The solution $(\hat{\theta}, \hat{\sigma}^2) = (\bar{x}, n^{-1} \sum (x_i - \bar{x})^2)$ indeed provides the global maximum of the likelihood function.
- ▶ The second-derivative verification is algebraically heavy even for two parameters; for more parameters, it becomes cumbersome.
- ▶ Practical alternative methods include:
 - ▶ Successive maximization (profile likelihoods),
 - ▶ Numerical optimization methods.
- ▶ Since MLEs are obtained via maximization, they are susceptible to numerical instability: small changes in data x can cause large changes in $\hat{\Theta}$.

Challenges in Finding MLEs

- ▶ Finding the global maximum: Identifying the global maximum can be straightforward but may present challenges even with familiar distributions, due to the fact that solutions for potential MLEs are found by solving:

$$\frac{\partial}{\partial \theta_i} L(\mathbf{x}; \theta) = 0, \quad i = 1, \dots, k$$

and solutions to this equation are only *possible candidates* for MLE: the first derivative equal to zero indicates potential extreme points, but further checks are needed for global maxima: points where the first derivative is zero can be local or global minima, maxima, or inflection points. Our goal is to identify the global maximum.

- ▶ Numerical sensitivity: estimates may vary significantly with small changes in data, raising concerns about the reliability of the MLE.

Comparing Method of Moments and Maximum Likelihood Estimation

We saw that the two methods can lead to the same point estimator, but this is not always the case:

- ▶ **Reliability:** When estimating parameters of a known probability distribution, Maximum Likelihood Estimation (MLE) is generally considered more reliable than the Method of Moments, as it tends to yield estimates closer to the true parameter values.
- ▶ **Computational Complexity:** MLE can involve complex equations that may be challenging to solve without computational tools, whereas the Method of Moments is usually simpler and can often be calculated by hand.
 - ▶ Thus, Method of Moments estimates can serve as useful initial approximations that can later be refined to MLE solutions using iterative methods like Newton-Raphson algorithm (see Lab 3).

- ▶ **Parameter Bounds:** In cases with smaller samples, Method of Moments estimates may sometimes fall outside of plausible parameter values. This issue does not occur with MLE.
- ▶ **Sufficiency:** MLE estimates often incorporate all relevant information in the data (a property known as sufficiency), while Method of Moments estimates may not capture all available information.
⇒ while both methods have their strengths, MLE is often preferred for its consistency and accuracy, especially when computational tools are available.

Bayesian Estimation

In Bayesian estimation, we combine our **prior beliefs** about a parameter (like a guess based on past experience or previous data) with **current data** to improve our estimate of the parameter. This approach contrasts with methods that rely solely on current data, like maximum likelihood estimation.

Bayes' Theorem in Parameter Estimation

Bayes' Theorem provides the foundation for Bayesian estimation. It updates our beliefs about a parameter θ after observing data \mathbf{x} . Bayes' theorem states:

$$\mathbb{P}(\theta|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathbf{x})}$$

where each term has a specific meaning in the context of parameter estimation:

- ▶ **Posterior Distribution** $\mathbb{P}(\theta|x)$: This is our updated belief about the parameter θ after taking into account the data x . It represents the probability of θ given the observed data and is the core result of Bayesian estimation.
- ▶ **Likelihood** $\mathbb{P}(x|\theta)$: This is the probability of observing the data x given a specific value of θ . It reflects how well the parameter θ explains the observed data.
- ▶ **Prior Distribution** $\mathbb{P}(\theta)$: This represents our belief about the parameter θ before seeing the current data. It could come from previous studies or expert knowledge.
- ▶ **Marginal Likelihood** $\mathbb{P}(x)$: Also called the "evidence," this term normalizes the posterior to ensure it's a proper probability distribution. It's calculated by integrating the likelihood over all possible values of θ , but in practice, it's often treated as a constant when comparing models or estimating parameters.

Bayesian Estimate

The Bayesian estimate of θ can be derived from the posterior distribution. A common choice is to use the **mean of the posterior distribution**, which represents the expected value of θ given the data:

$$\hat{\theta}_{\text{Bayesian}} = \mathbb{E}[\theta | \mathbf{x}]$$

This Bayesian estimate incorporates both prior information and the observed data, making it a more flexible and informative approach than classical estimates that do not use prior beliefs.

Example

Suppose you want to estimate the average height, θ , of adults in a city, based on a small sample of observed heights. Observed data is stored in x , and the prior beliefs are stored in θ .

- ▶ Define Prior Belief: Based on prior knowledge, you assume that adults in the general population have an average height of about 170 cm, but you recognize there is some variability. This belief can be represented by a Gaussian (Normal) distribution:

$$\theta \sim \mathcal{N}(170, \sigma_{\text{prior}}^2)$$

where:

- ▶ $\mathbb{E}[\theta] = 170$: This is the prior mean height, representing your central belief about θ before observing any data.
- ▶ $\sigma_{\text{prior}} = 10$: This standard deviation allows for variation around the mean of 170 cm, representing the uncertainty. Thus, $\sigma_{\text{prior}}^2 = 100$.

- ▶ Collect Data: To improve the estimate of θ , you gather a small sample of heights from the city population: [169, 167, 169].
 - ▶ Calculate the **sample mean** from this data: $\bar{x} = 168.3$.
 - ▶ Assume $\sigma_{\text{data}}^2 = 4$: This is the measurement variance, or the assumed variance of heights in the sample, representing the precision of your sample data.

Likelihood: You assume that heights follow a normal distribution around the true mean height, θ , with known sample variance, σ_{data}^2 . The likelihood, $\mathbb{P}(\mathbf{x}|\theta)$, represents the probability of observing this sample data, given different possible values of θ .

Posterior Distribution: Using Bayes' theorem, we can update our prior belief with the likelihood, to form a posterior distribution:

$$\mathbb{P}(\theta|\mathbf{x}) = \frac{\mathbb{P}(\mathbf{x}|\theta)\mathbb{P}(\theta)}{\mathbb{P}(\mathbf{x})}.$$

Since both the prior and likelihood are Gaussian, the posterior distribution of θ will also be Gaussian, with posterior mean (why?):

$$\theta_{\text{post}} = \frac{\sigma_{\text{data}}^2 \cdot \mathbb{E}[\theta] + n\sigma_{\text{prior}}^2 \cdot \bar{x}}{n\sigma_{\text{prior}}^2 + \sigma_{\text{data}}^2}$$

where $n = 3$ is the sample size. Plugging in our values:

- ▶ $\mathbb{E}[\theta] = 170$, $\sigma_{\text{prior}}^2 = 100$
- ▶ Sample mean: $\bar{x} = 168.3$, sample variance: $\sigma_{\text{data}}^2 = 4$

The Bayesian estimate of the mean height, θ , is:

$$\hat{\theta}_{\text{Bayesian}} = \frac{4 \cdot 170 + 3 \cdot 100 \cdot 168.3}{3 \cdot 100 + 4} = 168.8$$

The Bayesian estimate for θ , the average height, is 168.8 cm. This value incorporates both our prior belief (170 cm) and the observed data (mean of 168.3 cm), providing a balanced, refined estimate for θ .

Proportionality Symbol (\propto)

In Bayesian analysis, we often write:

$$\mathbb{P}(\theta | \mathbf{x}) \propto \mathbb{P}(\mathbf{x} | \theta) \mathbb{P}(\theta).$$

Interpretation:

- ▶ The symbol \propto means “*proportional to*”.
- ▶ It indicates that the posterior distribution has the same functional shape (in θ) as the product of the likelihood and the prior.
- ▶ The missing factor is the *normalizing constant*:

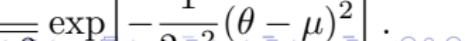
$$\mathbb{P}(\mathbf{x}) = \int \mathbb{P}(\mathbf{x} | \theta) \mathbb{P}(\theta) d\theta,$$

which ensures that $\int \mathbb{P}(\theta | \mathbf{x}) d\theta = 1$.

Practical note:

- ▶ During derivations, we omit this constant because it does not depend on θ .
- ▶ After identifying the posterior’s distributional form (e.g., Gaussian), normalization is automatically restored.

Example:

$$\mathbb{P}(\theta) \propto \exp\left[-\frac{1}{2\sigma^2}(\theta - \mu)^2\right] \Rightarrow \mathbb{P}(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(\theta - \mu)^2\right].$$


115 / 395

Bayesian Updating: Combining Prior and Likelihood

Bayes' Theorem:

$$\mathbb{P}(\theta | \mathbf{x}) = \frac{\mathbb{P}(\mathbf{x} | \theta) \mathbb{P}(\theta)}{\mathbb{P}(\mathbf{x})} \propto \mathbb{P}(\mathbf{x} | \theta) \mathbb{P}(\theta).$$

We assume both the **prior** and the **likelihood** are Gaussian:

$$\theta \sim \mathcal{N}(\mathbb{E}[\theta], \sigma_{\text{prior}}^2), \quad X_i | \theta \sim \mathcal{N}(\theta, \sigma_{\text{data}}^2).$$

Hence:

$$\mathbb{P}(\theta) \propto \exp\left[-\frac{1}{2\sigma_{\text{prior}}^2}(\theta - \mathbb{E}[\theta])^2\right], \quad \mathbb{P}(\mathbf{x} | \theta) \propto \exp\left[-\frac{1}{2\sigma_{\text{data}}^2} \sum_{i=1}^n (x_i - \theta)^2\right].$$

Posterior Distribution: Analytical Derivation

Multiplying the prior and likelihood:

$$\mathbb{P}(\theta | \mathbf{x}) \propto \exp \left[-\frac{1}{2\sigma_{\text{data}}^2} \sum_{i=1}^n (x_i - \theta)^2 - \frac{1}{2\sigma_{\text{prior}}^2} (\theta - \mathbb{E}[\theta])^2 \right].$$

Expanding the terms and collecting those in θ yields (check!):

$$-\frac{1}{2} \left[\left(\frac{n}{\sigma_{\text{data}}^2} + \frac{1}{\sigma_{\text{prior}}^2} \right) \theta^2 - 2 \left(\frac{n\bar{x}}{\sigma_{\text{data}}^2} + \frac{\mathbb{E}[\theta]}{\sigma_{\text{prior}}^2} \right) \theta \right].$$

Completing the square gives a Gaussian posterior (check!):

$$\theta | \mathbf{x} \sim \mathcal{N}(\theta_{\text{post}}, V_{\text{post}}),$$

with

$$V_{\text{post}}^{-1} = \frac{n}{\sigma_{\text{data}}^2} + \frac{1}{\sigma_{\text{prior}}^2}, \quad \theta_{\text{post}} = V_{\text{post}} \left(\frac{n\bar{x}}{\sigma_{\text{data}}^2} + \frac{\mathbb{E}[\theta]}{\sigma_{\text{prior}}^2} \right).$$

Interpretation of the Posterior Mean

Substituting for V_{post} , the posterior mean simplifies to:

$$\theta_{\text{post}} = \frac{\sigma_{\text{data}}^2 \mathbb{E}[\theta] + n \sigma_{\text{prior}}^2 \bar{x}}{n \sigma_{\text{prior}}^2 + \sigma_{\text{data}}^2}.$$

Interpretation:

- ▶ θ_{post} is a *precision-weighted average* of the prior mean and the sample mean.
- ▶ The weights depend on the inverse variances (precisions):

$$\theta_{\text{post}} = \underbrace{\frac{\sigma_{\text{data}}^2}{n \sigma_{\text{prior}}^2 + \sigma_{\text{data}}^2} \mathbb{E}[\theta]}_{\text{Weight on prior}} + \underbrace{\frac{n \sigma_{\text{prior}}^2}{n \sigma_{\text{prior}}^2 + \sigma_{\text{data}}^2} \bar{x}}_{\text{Weight on data}}$$

- ▶ Smaller variance (higher precision) implies greater influence in the posterior mean.
- ▶ The result provides a rational compromise between prior belief and empirical evidence.

LECTURES 4 and 5

Fisher's Information, Cramér-Rao, and Rao-Blackwell Theorems

Definition 10

For a sample of size n , *Fisher's information* relative to the parameter θ , is the quantity

$$I_n(\theta) = \mathbb{E} \left[\left(\frac{\partial \ln L(X_1, \dots, X_n; \theta)}{\partial \theta} \right)^2 \right] \quad (1)$$

if the likelihood function L is differentiable with respect to θ .

- ▶ Fisher's information is a way of measuring the amount of information that a random sample X_1, \dots, X_n carries about an unknown parameter θ , upon which the likelihood function depends. Formally, it is the expected value of the observed information (*score function*).

- ▶ Intuitively, Fisher's information quantifies how much the data tell us about θ . It measures how sensitive the likelihood function is to changes in the parameter θ : if the likelihood changes a lot when θ changes slightly, then the data contains a lot of information about θ .
- ▶ There's another, equivalent way to write Fisher Information — as the negative expected value of the second derivative of the log-likelihood. This form is often easier to compute.
- ▶ The first derivative measures sensitivity (the *score*). The second derivative measures curvature (convex/concave function) — sharper curvature \Rightarrow more information.

Proposition 1

If the range of X does not depend on θ and the likelihood function L is twice differentiable with respect to θ , then

$$I_n(\theta) = -\mathbb{E} \left[\frac{\partial^2 \ln L(X_1, \dots, X_n; \theta)}{\partial \theta^2} \right]$$

Corollary 2

If the range of X does not depend on θ and X_1, X_2, \dots, X_n are i.i.d, then

$$I_n(\theta) = n I_1(\theta)$$

⇒ larger samples carry proportional more information.

Proof. We saw before that

$$\ln L = \sum_{i=1}^n \ln f(X_i; \theta),$$
$$\frac{\partial^2 \ln L}{\partial \theta^2} = \sum_{i=1}^n \frac{\partial^2 \ln f(X_i; \theta)}{\partial \theta^2}.$$

By Proposition 1,

$$I_n(\theta) = - \sum_{i=1}^n \mathbb{E} \left[\frac{\partial^2 \ln f(X_i; \theta)}{\partial \theta^2} \right] = \sum_{i=1}^n I_1(\theta) = nI_1(\theta)$$

since $(X_i)_i$ are iid.

Example 3

$X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 known.

Log-Likelihood

$$\ln L(\mu) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

First-Derivative (Score Function)

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu)$$

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

Second-Derivative Form

$$\frac{\partial^2 \ln L}{\partial \mu^2} = -\frac{n}{\sigma^2} \implies I_n(\mu) = -\mathbb{E}\left[\frac{\partial^2 \ln L}{\partial \mu^2}\right] = \frac{n}{\sigma^2}.$$

- ▶ **Both forms are equivalent**, but the second-derivative form is usually easier because:
 - ▶ It avoids squaring and integrating complicated terms.
 - ▶ Often, the second derivative does not depend on the data, so the expectation is trivial.
 - ▶ It directly measures the *curvature* of the log-likelihood — sharper curvature means more information.
- ▶ Fisher Information reflects how “peaked” the likelihood is around the true parameter value.

Example 4

Let X be a Bernoulli trial. The Fisher information contained in X may be calculated to be

$$\begin{aligned} I_1(\theta) &= -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} \log (\theta^X (1-\theta)^{1-X}) \right] \\ &= -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} (X \log \theta + (1-X) \log(1-\theta)) \right] \\ &= \mathbb{E} \left[\frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2} \right] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)} \implies I_n(\theta) = \frac{n}{\theta(1-\theta)} \end{aligned}$$

- In general we look for unbiased estimators with small variance or at least we hope that the variance becomes smaller as the sample size increases. This is the idea formalised in the following definition.

Definition 11

An estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ is called an **absolutely correct estimator** for θ , if it satisfies the conditions

- (i) $\mathbb{E}[\hat{\theta}] = \theta$,
- (ii) $\lim_{n \rightarrow \infty} V(\hat{\theta}) = 0$.

- ▶ An estimator is absolutely correct if it's unbiased and becomes perfectly accurate as the sample size grows. That is, its variance goes to zero when $n \rightarrow \infty$.

Remark 5

The sample mean \bar{X} is an absolutely correct estimator for the theoretical mean $\mu = \mathbb{E}[X]$. More generally, the sample moment of order k , $\bar{\mu}^k$, is an absolutely correct estimator for the population moment of order k , $\mu^k = \mathbb{E}[X^k]$.

- ▶ A *minimum-variance unbiased estimator* has the lowest variance that an unbiased estimator can possibly have. The next result tells us exactly how low that can be, under certain conditions.
- ▶ Recall: a *characteristic* or *variable* is a certain feature of interest of the elements of a population or a sample, that is analysed statistically. From the probabilistic perspective, a numerical characteristic is a random variable.

- ▶ Among all unbiased estimators, some have smaller variance than others. But there's a theoretical lower limit — the Cramér–Rao bound.
- ▶ No unbiased estimator can have variance smaller than this. If one reaches the bound exactly, we call it *efficient*.

Theorem 6 (Cramér–Rao Inequality)

Let X_1, X_2, \dots, X_n be a random sample with pdf or pmf $f(\cdot; \theta)$, and let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be an estimator of θ , such that

$$\frac{d}{d\theta} \mathbb{E}[\hat{\theta}] = \int \frac{\partial}{\partial \theta} [\hat{\theta}(\mathbf{x}) f(\mathbf{x}; \theta)] d\mathbf{x},$$

and

$$V(\hat{\theta}) < \infty.$$

Then, for every value of θ for which these quantities exist,

$$V(\hat{\theta}) \geq \frac{\left(\frac{d}{d\theta} \mathbb{E}[\hat{\theta}] \right)^2}{\mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta) \right)^2 \right]}.$$

Proof: The proof of this theorem rests on an application of the Cauchy–Schwarz inequality, or, stated statistically, on the fact that for any two random variables X and Y ,

$$[\text{Cov}(X, Y)]^2 \leq V(X)V(Y).$$

Rearranging gives a lower bound on the variance of X ,

$$V(X) \geq \frac{[\text{Cov}(X, Y)]^2}{V(Y)}.$$

The key idea is to choose X to be the estimator $\hat{\theta}(\mathbf{X})$ and Y to be the quantity

$$\frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta),$$

and to apply the inequality above.

Let $\mathbf{X} = (X_1, \dots, X_n)$ denote the random sample, with joint pdf or pmf $f(\mathbf{x}; \theta)$ defined on the sample space $\mathcal{X} \subseteq \mathbb{R}^n$. Differentiating under the integral sign,

$$\frac{d}{d\theta} \mathbb{E}[\hat{\theta}] = \int_{\mathcal{X}} \hat{\theta}(\mathbf{x}) \frac{\partial}{\partial \theta} f(\mathbf{x}; \theta) d\mathbf{x}.$$

Multiplying and dividing by $f(\mathbf{x}; \theta)$ inside the integrand, we obtain

$$\frac{d}{d\theta} \mathbb{E}[\hat{\theta}] = \int_{\mathcal{X}} \hat{\theta}(\mathbf{x}) f(\mathbf{x}; \theta) \frac{\frac{\partial}{\partial \theta} f(\mathbf{x}; \theta)}{f(\mathbf{x}; \theta)} d\mathbf{x} = \mathbb{E}\left[\hat{\theta} \frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta)\right].$$

This expression suggests a covariance between $\hat{\theta}$ and $\frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta)$. To make this explicit, we note that

$$\mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta)\right] = \frac{d}{d\theta} \mathbb{E}[1] = \frac{d}{d\theta} \left(\int_{\mathcal{X}} f(\mathbf{x}; \theta) d\mathbf{x} \right) = 0.$$

Therefore the covariance simplifies to

$$\text{Cov}\left(\hat{\theta}, \frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)\right) = \mathbb{E}\left[\hat{\theta} \frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)\right] = \frac{d}{d\theta} \mathbb{E}[\hat{\theta}].$$

Furthermore, since the expected value of the score function is zero, its variance equals its second moment,

$$V\left(\frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)\right) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)\right)^2\right].$$

Now, applying the Cauchy–Schwarz inequality to $\hat{\theta}$ and the score function, we have

$$\left(\frac{d}{d\theta} \mathbb{E}[\hat{\theta}]\right)^2 = \text{Cov}^2\left(\hat{\theta}, \frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)\right) \leq V(\hat{\theta}) \mathbb{E}\left[\left(\frac{\partial}{\partial\theta} \ln f(\mathbf{X}; \theta)\right)^2\right].$$

Rearranging gives the Cramér–Rao inequality,

$$V(\hat{\theta}) \geq \frac{\left(\frac{d}{d\theta} \mathbb{E}[\hat{\theta}] \right)^2}{\mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta) \right)^2 \right]}.$$

The denominator is the Fisher information,

$$I_n(\theta) := \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta) \right)^2 \right],$$

so that equivalently

$$V(\hat{\theta}) \geq \frac{\left(\frac{d}{d\theta} \mathbb{E}[\hat{\theta}] \right)^2}{I_n(\theta)}.$$

Remark on the unbiased case. If $\hat{\theta}$ is unbiased, $\mathbb{E}[\hat{\theta}] = \theta$, hence $\frac{d}{d\theta}\mathbb{E}[\hat{\theta}] = 1$ and we recover the bound

$$V(\hat{\theta}) \geq \frac{1}{I_n(\theta)}.$$

Equality condition. Equality in the Cauchy–Schwarz step (and thus in the Cramér–Rao bound) holds iff there exist real functions $a(\theta)$ and $b(\theta)$ such that, with probability 1,

$$\hat{\theta} - \mathbb{E}[\hat{\theta}] = a(\theta) U(\mathbf{X}; \theta) + b(\theta),$$

which (after centering) reduces to a linear relation between $\hat{\theta} - \mathbb{E}[\hat{\theta}]$ and the score $U(\mathbf{X}; \theta)$.

Equality condition in Cauchy–Schwarz

Remark: For real random variables A, B with finite variances, equality in the Cauchy–Schwarz inequality

$$[\text{Cov}(A, B)]^2 \leq \text{Var}(A) \text{Var}(B)$$

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

$$A = aB + b \quad \text{almost surely (a.s.)}.$$

That is, A must be an *exact affine function* of B with probability 1.

Interpretation

Equality \iff one variable lies on a straight line (a.s.) in terms of the other.

Applying this to the Cramér–Rao inequality

In some sense, we applied Cauchy–Schwarz with

$$A = \hat{\theta} - \mathbb{E}[\hat{\theta}], \quad B = U(\mathbf{X}; \theta) = \frac{\partial}{\partial \theta} \ln f(\mathbf{X}; \theta)$$

and used that $\mathbb{E}[\dots] = 0$. Equality holds if and only if

$$\hat{\theta} - \mathbb{E}[\hat{\theta}] = a(\theta) U(\mathbf{X}; \theta) + b(\theta) \quad \text{a.s.}$$

The constants $a(\theta)$ and $b(\theta)$ may depend on θ . But since $\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]] = 0$, we must have

$$b(\theta) + a(\theta) \mathbb{E}[U(\mathbf{X}; \theta)] = 0.$$

Because $\mathbb{E}[U(\mathbf{X}; \theta)] = 0$, it follows that $b(\theta) = 0$.

$$\boxed{\hat{\theta} - \mathbb{E}[\hat{\theta}] = a(\theta) U(\mathbf{X}; \theta) \quad \text{a.s.}}$$

Intuition

The score $U(\mathbf{X}; \theta)$ contains all information about how the likelihood changes with θ . If the estimator's deviation from its mean is a linear function of this score, then it fully exploits that information. Such an estimator achieves the Cramér–Rao bound — i.e. it is *efficient*. Asymptotically, the MLE has this property.

Corollary 7 (Cramér-Rao)

Let X be a characteristic with the property that the function $f(\cdot; \theta)$ is differentiable with respect to θ and let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be an absolutely correct estimator for θ . Then

$$V(\hat{\theta}) \geq \frac{1}{I_n(\theta)}.$$

Definition 12

Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be an absolutely correct estimator for θ . The efficiency of $\hat{\theta}$ is the quantity

$$e(\hat{\theta}) = \frac{1}{I_n(\theta)V(\hat{\theta})}$$

The estimator $\hat{\theta}$ is said to be efficient for θ , if $e(\hat{\theta}) = 1$.

Sufficient Statistics and the Rao-Blackwell Theorem

- ▶ Recall that we are estimating a population parameter θ , by an estimator $\hat{\theta}(X_1, \dots, X_n)$, based on a sample and by using the sample variables X_1, X_2, \dots, X_n .
- ▶ Our goal is to find "good" estimators, i.e. unbiased estimators with low variance.
- ▶ We will explore also other means of finding such estimators, by finding "appropriate" statistics (sample functions) to use in our estimation process.

- ▶ So far, we have chosen estimators from a sample on the basis of intuition, like the sample mean \hat{X} for the population mean $\mu = \mathbb{E}(X)$, or the sample variance \hat{s}^2 for the population variance $\sigma^2 = V(X)$.
- ▶ We did not use the actual sample values themselves, but rather, we summarized (reduced) them into a value. Has this process retained all the information about the target parameter θ (e.g., μ or σ^2) or has some important information been lost?
- ▶ We want to find statistics that, in some way, summarize all the information in a sample about a target parameter. These are called **sufficient statistics**.

Definition 13

Let X_1, \dots, X_n denote a random sample of size n from a distribution characterized by the pdf or pmf $f(\cdot; \theta)$ with unknown parameter θ .

Consider x_1, \dots, x_n to be the corresponding statistical data. Let $S = S(X_1, X_2, \dots, X_n)$ be a statistic whose pdf or pmf is $f_S(\cdot; \theta)$.

Then S is a **sufficient statistic** for θ if and only if

$$\frac{f(x_1; \theta) \cdot \dots \cdot f(x_n; \theta)}{f_S(S(x_1, \dots, x_n); \theta)} = h(x_1, \dots, x_n),$$

where $h(x_1, \dots, x_n)$ does not depend on the parameter θ .

- ▶ In fact, the statistic $S = S(X_1, \dots, X_n)$ is called *sufficient* for (the estimation of) θ , if the conditional probability distribution of (X_1, \dots, X_n) , given the value of the statistic $S = s$ does not depend on the parameter θ .
- ▶ Intuitively, this is saying that once the value of the statistic S is known, no other function of X_1, \dots, X_n (so no other statistic) will shed additional light on the possible values of θ .
- ▶ Then the conditional probability distribution of the data does not depend on the unknown parameter except through the sufficient statistic. In that sense, S contains all the information in the sample about θ .

Fisher's Factorization Criterion

An equivalent result for sufficiency is given in the following theorem:

Theorem 8

A statistic $S = S(X_1, \dots, X_n)$ is sufficient for θ , if and only if the likelihood function L can be factored into two nonnegative functions

$$L(x_1, \dots, x_n; \theta) = h(x_1, \dots, x_n) g(s; \theta),$$

such that one factor, h , does not depend on θ and the other factor, g , which does depend on θ , depends on (x_1, \dots, x_n) only through the value $s = S(x_1, \dots, x_n)$.

Example 9

Let us consider X_1, \dots, X_n to be a random sample drawn from a Poisson $\text{Poiss}(\lambda)$ distribution, with $\lambda > 0$ and $s = \sum_{k=1}^n X_k$.

Solution: For each $i = \overline{1, n}$,

$$f(x_i; \lambda) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}$$

Then the likelihood function is given by

$$\begin{aligned} L(x_1, \dots, x_n; \lambda) &= \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\left(\sum_{i=1}^n x_i\right)}}{x_1! \dots x_n!} \cdot e^{-n\lambda} \\ &= \frac{1}{x_1! \dots x_n!} \cdot \lambda^s e^{-n\lambda} = g(x_1, \dots, x_n) h(s; \lambda), \end{aligned}$$

where $s = x_1 + \dots + x_n$, $g(x_1, \dots, x_n) = \frac{1}{x_1! \dots x_n!}$ and $h(s; \lambda) = \lambda^s e^{-n\lambda}$.

Thus, S is a sufficient statistic for the estimation of λ .

Definition 14

Let X and Y be discrete random variables. Define

$$p_{X|Y}(x | y) = \begin{cases} \frac{\mathbb{P}(X=x, Y=y)}{\mathbb{P}(Y=y)} = \mathbb{P}(X = x | Y = y), & \text{if } \mathbb{P}(Y = y) > 0 \\ \mathbb{P}(X = x), & \text{otherwise .} \end{cases}$$

$p_{X|Y}(\cdot | y)$ is called **conditional probability mass function of X given $\{Y = y\}$** (for $y \in \mathbb{R}$). The conditional expectation of the random variable X given $\{Y = y\}$ is given by

$$\mathbb{E}[X | Y = y] = \sum_{x \in X(\Omega)} x p_{X|Y}(x | y),$$

if the series above is absolutely convergent.

Let (X, Y) be a continuous random vector with joint density function $f_{X,Y}$. Then, from Probability Theory it is known that

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) dy \quad \text{for } x \in \mathbb{R}$$

and

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y) dx \quad \text{for } y \in \mathbb{R}$$

are density functions of X , respectively Y .

Definition 15

Consider the function $f_{X|Y}(\cdot | \cdot) : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined for each $x, y \in \mathbb{R}$ by

$$f_{X|Y}(x | y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)}, & \text{if } f_Y(y) > 0 \\ f_X(x), & \text{if } f_Y(y) = 0. \end{cases}$$

$f_{X|Y}(\cdot | y)$ is called the **conditional density function** of the random variable X given $\{Y = y\}$ (for $y \in \mathbb{R}$). The conditional expectation of the random variable X given $\{Y = y\}$ is given by

$$\mathbb{E}[X | Y = y] = \int_{\mathbb{R}} x f_{X|Y}(x | y) dx$$

if the above integral is absolutely convergent. We define the **conditional expectation of X given Y** , denoted by $\mathbb{E}[X | Y]$, as the random variable $H(Y)$, where $H(y) := \mathbb{E}[X | Y = y]$ (for both cases).

Example 10

The joint density function of (X, Y) is $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by

$$f(x, y) = \begin{cases} \frac{1}{2}ye^{-xy}, & 0 < x < \infty, 0 < y < 2 \\ 0, & \text{otherwise} \end{cases}$$

Compute $\mathbb{E}[X \mid Y = 1]$.

Solution: The conditional density function of X , given that $\{Y = 1\}$ is given by

$$f_{X|Y}(x \mid 1) = \frac{f(x, 1)}{f_Y(1)} = \frac{\frac{1}{2}e^{-x}}{\int_0^\infty \frac{1}{2}e^{-x}dx} = e^{-x}, \text{ for } x > 0$$
$$f_{X|Y}(x \mid 1) = 0, \text{ for } x \leq 0$$

Then,

$$\mathbb{E}[X \mid Y = 1] = \int_{\mathbb{R}} xf_{X|Y}(x \mid 1)dx = \int_0^\infty x \cdot e^{-x}dx = 1$$

Theorem 11 (Rao-Blackwell)

Let $\hat{\theta}$ be an unbiased estimator and S be a sufficient statistic for θ . Let $\bar{\theta} = \mathbb{E}[\hat{\theta} | S]$. Then

- a. $\mathbb{E}[\bar{\theta}] = \theta$ (i.e. $\bar{\theta}$ is also an unbiased estimator for θ)
- b. $V(\bar{\theta}) \leq V(\hat{\theta})$.

Proof:

Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be an unbiased estimator of θ , and let $S = S(X_1, \dots, X_n)$ be a sufficient statistic for θ . Define

$$\bar{\theta} := \mathbb{E}[\hat{\theta} | S].$$

(a) **Unbiasedness.** By the properties of the conditional expectation,

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

so $\bar{\theta}$ is unbiased.

(b) Variance reduction. Note that, by the definition of variance,

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}\left[(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2\right].$$

Expanding the square:

$$(\hat{\theta} - \bar{\theta} + \bar{\theta} - \theta)^2 = (\hat{\theta} - \bar{\theta})^2 + (\bar{\theta} - \theta)^2 + 2(\hat{\theta} - \bar{\theta})(\bar{\theta} - \theta).$$

Take expectation. The cross-term vanishes because

$$\begin{aligned}\mathbb{E}[(\hat{\theta} - \bar{\theta})(\bar{\theta} - \theta)] &= \mathbb{E}\left[\mathbb{E}[(\hat{\theta} - \bar{\theta})(\bar{\theta} - \theta) | S]\right] \\ &= (\bar{\theta} - \theta)\mathbb{E}\left[\mathbb{E}[\hat{\theta} - \bar{\theta} | S]\right] = 0,\end{aligned}\tag{2}$$

since $\mathbb{E}[\hat{\theta} - \bar{\theta} | S] = 0$ by definition of conditional expectation.

Thus

$$\text{Var}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \bar{\theta})^2] + \mathbb{E}[(\bar{\theta} - \theta)^2] \geq \mathbb{E}[(\bar{\theta} - \theta)^2] = \text{Var}(\bar{\theta}),$$

showing that $\bar{\theta}$ has variance no larger than $\hat{\theta}$.

Remark Since $\bar{\theta} = \mathbb{E}[\hat{\theta} | S]$, it is a function of S , which in turn is a function of the sample. Moreover, by sufficiency of S , this function does not depend on θ . Therefore $\bar{\theta}$ is a valid estimator.

Combining (a) and (b), $\bar{\theta}$ is an unbiased estimator of θ with variance no larger than that of $\hat{\theta}$.

LECTURE 6

Lehmann-Scheffé Theorem

- ▶ We want to identify estimators that are **unbiased** and have the **minimum variance** among all unbiased estimators.
- ▶ Ingredients we need:
 1. **Sufficient statistic:** captures all the information about the parameter θ contained in the sample X_1, \dots, X_n .
 2. **Complete statistic:** no non-trivial function of the statistic has expectation zero for all θ .
 3. **Rao-Blackwell theorem:** given an unbiased estimator, conditioning on a sufficient statistic improves (or at least does not worsen) its variance.
- ▶ Lehmann-Scheffé insight:

*If S is **sufficient and complete** for θ , then the Rao-Blackwell improvement of any unbiased estimator using S is **the unique minimum-variance unbiased estimator (MVUE)**.*
- ▶ Strategy for finding MVUE:
 1. Identify a sufficient and complete statistic S .
 2. Start with any unbiased estimator $\hat{\theta}$.
 3. Compute $\bar{\theta} = \mathbb{E}[\hat{\theta} | S]$.

Definition 12

The statistic $S = S(X_1, \dots, X_n)$ is said to be *complete* for the family of probability distributions $f(\cdot; \theta), \theta \in A$, if for any measurable function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$

$$\mathbb{E}[\varphi(S)] = 0, \forall \theta \in A \implies \varphi(S) \stackrel{\text{a.s.}}{=} 0, \text{ which means } \mathbb{P}(\varphi(S) = 0) = 1.$$

Example 13

Consider the random sample $X_1, X_2, \dots, X_n \sim N(\theta, 1), \theta \in \mathbb{R}$, (i.i.d.) and the statistic $S(X_1, X_2, \dots, X_n) = X_1 - X_n$.

This is a statistic which is **not** complete for the family of normal distributions $N(\theta, 1), \theta \in \mathbb{R}$, because:

for $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ given by $\varphi(x) = x$, we have $\mathbb{E}[\varphi(S)] = \mathbb{E}[X_1 - X_n] = 0$ for all $\theta \in \mathbb{R}$, but

$$\mathbb{P}(\varphi(S) = 0) = \mathbb{P}(X_1 - X_n = 0) = 0 \neq 1, \text{ since } X_1 - X_n \sim N(0, 2).$$

Example 14

Let us consider the statistic $S = \sum_{i=1}^n X_i$ for the Poiss (λ), $\lambda > 0$ distribution. Is it complete?

Since each sample variable $X_i \sim \text{Poiss}(\lambda)$, the sum S also has a Poisson distribution $\text{Poiss}(n\lambda)$ (result from Probability Theory), with pmf

$$\mathbb{P}(S = s) = \frac{(n\lambda)^s}{s!} e^{-n\lambda}, s \in \{0, 1, \dots\}.$$

We have

$$\mathbb{E}[\varphi(S)] = \sum_{s=0}^{\infty} \varphi(s) \frac{(n\lambda)^s}{s!} e^{-n\lambda} = e^{-n\lambda} \sum_{s=0}^{\infty} \frac{\varphi(s)n^s}{s!} \lambda^s.$$

We check the condition from the previous definition:

$$e^{-n\lambda} \sum_{s=0}^{\infty} \frac{\varphi(s)n^s}{s!} \lambda^s = 0, \forall \lambda > 0 \implies \sum_{s=0}^{\infty} \frac{\varphi(s)n^s}{s!} \lambda^s = 0, \forall \lambda > 0.$$

The only way that this can happen is if all the coefficients in the power series are equal to 0 , i.e.

$$\varphi(s) = 0, \forall s \in \mathbb{N},$$

which means

$$\varphi(S) \stackrel{\text{a.s.}}{=} 0.$$

Thus, S is a complete statistic for the family of Poisson Poiss (λ), $\lambda > 0$, distributions.

In the next theorem, we will see what happens when we construct the estimator from the Rao-Blackwell Theorem using a statistic that is sufficient and complete. But before that we introduce one more definition, for completeness:

Definition 15

An unbiased estimator $\bar{\theta} = \bar{\theta}(X_1, \dots, X_n)$ for θ is called a **minimum-variance unbiased estimator (MVUE)**, if it has lower variance than any other unbiased estimator for θ , that is

$$V(\bar{\theta}) \leq V(\hat{\theta})$$

for all estimators $\hat{\theta}$ with $\mathbb{E}[\hat{\theta}] = \theta$.

Theorem 16 (Lehmann-Scheffé)

Let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be an unbiased estimator for θ and $S = S(X_1, \dots, X_n)$ be a sufficient and complete statistic for θ . Then the estimator

$$\bar{\theta} = \bar{\theta}(X_1, \dots, X_n) = \mathbb{E}[\hat{\theta} | S]$$

is a MVUE.

Example 17

Consider again X_1, \dots, X_n , a sample drawn from the Poisson distribution $\text{Poisson}(\lambda)$, $\lambda > 0$ and the statistic $S = \sum^n X_k = n\bar{X}$. Find a MVUE for $\theta = e^{-\lambda}$.

To make computations easier, we will find first an unbiased estimator for the parameter $\theta = e^{-\lambda}$. The pmf of each X_i is

$$f(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda} = \frac{1}{x!} \theta \left(\ln \frac{1}{\theta} \right)^x, x = 0, 1, \dots$$

and S has a Poisson pdf with parameter $n\lambda = n \left(\ln \frac{1}{\theta} \right)$. Recall that S is a sufficient and complete statistic for λ (we proved this in examples), hence it is also a sufficient and complete statistic for $\theta = e^{-\lambda}$.

In order to use the theorem, we have to start with an unbiased estimator for θ . For $i = 1, \dots, n$, consider the variables

$$Y_i = \begin{cases} 1, & X_i = 0 \\ 0, & X_i \neq 0 \end{cases}$$

Then

$$\mathbb{P}(Y_i = 1) = \mathbb{P}(X_i = 0) = e^{-\lambda} = \frac{1}{0!}\theta \left(\ln \frac{1}{\theta}\right)^0 = \theta$$

and, obviously $\mathbb{P}(Y_i = 0) = 1 - \theta$. Thus, each $Y_i \sim \text{Bernoulli}(\theta)$ with pmf

$$Y_i \sim \begin{pmatrix} 0 & 1 \\ 1 - \theta & \theta \end{pmatrix}$$

and expectation $E(Y_i) = \theta, i = \overline{1, n}$. Let

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{\text{the number of } X_i \text{'s that are equal to 0}}{n}.$$

Then

$$\mathbb{E}[\hat{\theta}] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i] = \frac{1}{n} \cdot n\theta = \theta,$$

so $\hat{\theta}$ is an unbiased estimator for θ . Hence, by Theorem Lehmann-Scheffé, the estimator $\bar{\theta} = E[\hat{\theta} | S]$ is the MVUE for θ . Let us compute it. By the linearity property of the conditional expectation we have

$$\begin{aligned}\bar{\theta} &= \mathbb{E}[\hat{\theta} | S] = \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n Y_i \middle| S \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[Y_i | S] \\ &= \frac{1}{n} \cdot n \mathbb{E}[Y_1 | S] = \mathbb{E}[Y_1 | S].\end{aligned}\tag{3}$$

Now, to go further with the computation, let us recall a few things about conditional probability:

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \mathbb{P}(B | A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \mathbb{P}(A | B) = \frac{\mathbb{P}(A) \cdot \mathbb{P}(B | A)}{\mathbb{P}(B)}$$

So, we have

$$\begin{aligned}\mathbb{E}[Y_1 \mid S = s] &= \mathbb{P}(Y_1 = 0 \mid S = s) \cdot 0 + \mathbb{P}(Y_1 = 1 \mid S = s) \cdot 1 \\&= \mathbb{P}(X_1 = 0 \mid S = s) \\&= \frac{\mathbb{P}((X_1 = 0) \cap (S = s))}{\mathbb{P}(S = s)} \\&= \frac{\mathbb{P}(X_1 = 0) \cdot \mathbb{P}(S = s \mid X_1 = 0)}{\mathbb{P}(S = s)} \\&= \frac{\mathbb{P}(X_1 = 0) \cdot \mathbb{P}(\sum_{i=1}^n X_i = s \mid X_1 = 0)}{\mathbb{P}(S = s)} \\&= \frac{\mathbb{P}(X_1 = 0) \cdot \mathbb{P}(X_2 + \dots + X_n = s)}{\mathbb{P}(S = s)}\end{aligned}$$

Now, $\mathbb{P}(X_1 = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = e^{-\lambda} = \theta$ and since all X_i 's have $\text{Poiss}(\lambda)$ pmf, the sum $X_2 + \dots + X_n$ will have $\text{Poiss}((n-1)\lambda)$ pmf

$$X_2 + \dots + X_n \sim \left(\frac{\frac{s}{((n-1)\lambda)^s}}{s!} e^{-(n-1)\lambda} \right)_{s=0,1,\dots}.$$

Then,

$$\mathbb{E}[Y_1 \mid S = s] = \frac{\theta \cdot \frac{1}{s!} \left((n-1) \ln \frac{1}{\theta} \right)^s \cdot \theta^{n-1}}{\frac{1}{s!} \left(n \ln \frac{1}{\theta} \right)^s \cdot \theta^n} = \left(\frac{n-1}{n} \right)^s = \left(1 - \frac{1}{n} \right)^s.$$

Thus, the estimator

$$\bar{\theta} = \mathbb{E}[\hat{\theta} \mid S] = \left(1 - \frac{1}{n} \right)^S = \left(1 - \frac{1}{n} \right)^{n\bar{X}}$$

is the MVUE for $\theta = e^{-\lambda}$.

Example 18

Let X_1, \dots, X_n be sample variables for a random sample drawn from a Bernoulli distribution with parameter $p \in (0, 1)$, unknown. Find the MVUE for p .

Recall the $\text{Bern}(p)$ pdf

$$\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

So, for each $i = \overline{1, n}$, the pdmf can be written

$$f(x_i; p) = \mathbb{P}(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}, \quad x_i \in \{0, 1\}$$

Then the likelihood function of the sample is

$$\begin{aligned} L(x_1, \dots, x_n; p) &= \prod_{i=1}^n f(x_i; p) \\ &= p^{x_1 + \dots + x_n} (1-p)^{n - (x_1 + \dots + x_n)} \\ &= p^s (1-p)^{n-s} \end{aligned}$$

where S is the statistic

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

By the previous theorem, with $g(x_1, \dots, x_n) = 1$ and $h(s; p) = p^s(1-p)^{n-s}$, S is sufficient. Now assume that $\mathbb{E}[\varphi(S)] = 0$, for all $p \in (0, 1)$. Recall that since X_1, \dots, X_n are independent and identically distributed with a Bernoulli distribution, S has a Binomial distribution with parameters n and p

$$S \sim \binom{s}{C_n^s p^s (1-p)^{n-s}}_{s=0,n}$$

and $\varphi(S)$ has pmf

$$\varphi(S) \sim \binom{\varphi(s)}{C_n^s p^s (1-p)^{n-s}}_{s=0,n}$$

Then

$$\mathbb{E}[\varphi(S)] = \sum_{s=0}^n \varphi(s) C_n^s p^s (1-p)^{n-s} = (1-p)^n \sum_{s=0}^n \varphi(s) C_n^s \left(\frac{p}{1-p}\right)^s$$

If $\mathbb{E}[\varphi(S)] = 0$, for all $p \in (0, 1)$, then

$$\sum_{s=0}^n \varphi(s) C_n^s \left(\frac{p}{1-p} \right)^s = 0$$

for all $p \in (0, 1)$, which is possible only if $\varphi(s) = 0$, for all $s = \overline{0, n}$, i.e. $\varphi(S) \stackrel{\text{a.s.}}{=} 0$. Thus S is also complete. Now let us consider the estimator $\hat{p} = \bar{X} = \frac{1}{n}S$. Since $S \in \text{Bino}(n, p)$, we know that $\mathbb{E}[S] = np$ and, hence $\mathbb{E}[\hat{p}] = p$, so \hat{p} is an unbiased estimator for p . Then by the previous theorem, the MVUE is given by

$$\bar{p} = \mathbb{E}[\hat{p} | S] = \mathbb{E} \left[\frac{1}{n}S \mid S \right] = \frac{1}{n}\mathbb{E}[S | S] = \frac{1}{n}S = \bar{X}.$$

Thus, the sample mean \bar{X} is the MVUE for the population mean $p = \mathbb{E}[X]$.

LECTURES 7 and 8

Confidence Intervals

Interval Estimation

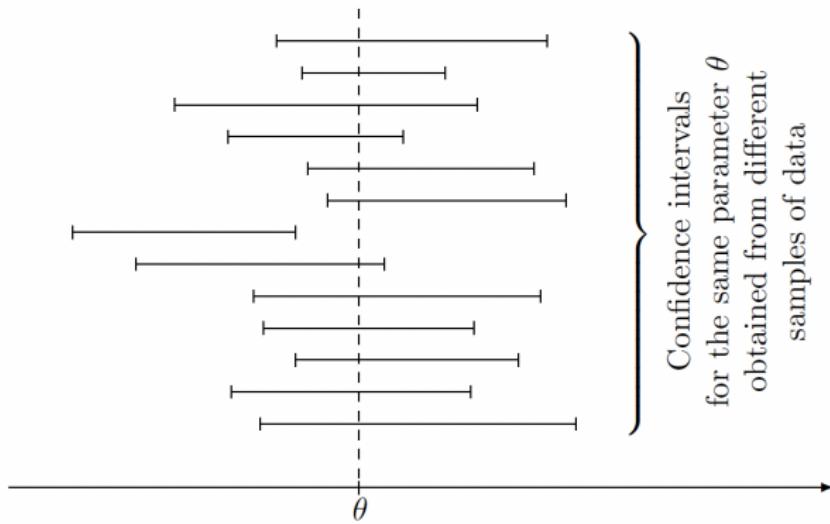
- ▶ Until now, we have considered point estimators, which provide a single value, $\hat{\theta}$, to estimate an unknown parameter θ . However, point estimators do not provide too much information about the accuracy of the estimate, it doesn't have a built-in measure of uncertainty.
- ▶ **Point estimation** answers
What is your best guess for the parameter?
- ▶ **Interval estimation** answers
How accurate/uncertain is this guess?

Interval Estimation

- ▶ An **interval estimator** defines a range of values within which the parameter is likely to lie.
- ▶ More specifically, the sample will produce two functions, $\hat{\theta}_L(X_1, \dots, X_n)$ and $\hat{\theta}_U(X_1, \dots, X_n)$, representing the lower and upper bounds of the confidence interval, such that for a given $\alpha \in (0, 1)$,

$$\mathbb{P}(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha.$$

- ▶ $(\hat{\theta}_L, \hat{\theta}_U)$ is called a **two-sided confidence interval (CI)**, specifically a $100(1 - \alpha)\%$ confidence interval.
- ▶ The values $\hat{\theta}_L$ and $\hat{\theta}_U$ are the **lower and upper confidence limits**.
- ▶ The quantity $1 - \alpha$ is called the **confidence level** or **confidence coefficient**.
- ▶ The value α is called the **significance level**.



Confidence intervals and coverage for parameter θ [Baron]

Coverage Probability and Confidence Intervals

- ▶ The population parameter θ is a fixed value; it is not random. It represents a characteristic of the entire population and remains constant, independent of any random sampling process. In contrast, the confidence interval is derived from random data, meaning that **the interval itself is random**.
- ▶ The *coverage probability* refers to the likelihood that our confidence interval contains the true parameter θ . This is illustrated in the figure above.
- ▶ To better understand this, consider the following process: if we repeatedly collect random samples and construct a confidence interval for each, then, for a confidence level of $1 - \alpha$, we expect approximately $(1 - \alpha)100\%$ of these intervals to contain θ , while $100\alpha\%$ will not.
- ▶ In the previous figure, one interval does not cover θ . This is not due to an error in data collection or interval construction, but rather due to a sampling error.

Additionally, we can construct **one-sided confidence intervals (CIs)**:

- ▶ A $100(1 - \alpha)\%$ lower confidence interval for θ is $(-\infty, \hat{\theta}_U]$, where

$$\mathbb{P}(\theta \leq \hat{\theta}_U) = 1 - \alpha.$$

- ▶ A $100(1 - \alpha)\%$ upper confidence interval for θ is $[\hat{\theta}_L, \infty)$, where

$$\mathbb{P}(\hat{\theta}_L \leq \theta) = 1 - \alpha.$$

Remark:

- ▶ The condition $\mathbb{P}(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$ does not uniquely determine a $100(1 - \alpha)\%$ CI.
- ▶ A smaller value of α results in a narrower confidence interval, which implies a more precise estimate for θ . Increasing the confidence level will make the interval wider, reducing the precision.

To construct a confidence interval for θ , we need a **pivotal quantity**, which is a statistic S that satisfies two conditions:

- ▶ $S = S(X_1, \dots, X_n; \theta)$ is a function of the sample measurements and the unknown parameter θ , which is the only unknown.
- ▶ The distribution of S is known and does not depend on θ .

Quantiles of a Distribution

Definition 19

A quantile of order α for the distribution of the studied characteristic X is the number $q_\alpha \in \mathbb{R}$ for which

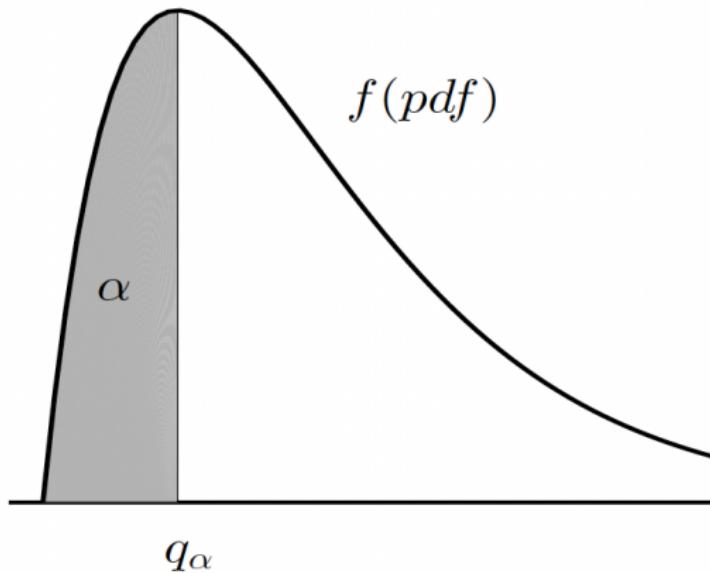
$$\mathbb{P}(X < q_\alpha) \leq \alpha \leq \mathbb{P}(X \leq q_\alpha)$$

- ▶ If X is a continuous random variable: q_α is a quantile of order $\alpha \iff \mathbb{P}(X \leq q_\alpha) = \alpha \iff F_X(q_\alpha) = \alpha$
- ▶ If the cdf F_X admits an inverse function F_X^{-1} , then $q_\alpha = F_X^{-1}(\alpha)$

Example: For $X \sim \begin{pmatrix} 1 & 3 & 5 & 7 \\ 0.2 & 0.35 & 0.35 & 0.1 \end{pmatrix}$, compute $q_{0.5}$, its median.

$$\mathbb{P}(X < 3) = 0.2 \leq 0.5 \leq \mathbb{P}(X \leq 3) = 0.2 + 0.35 = 0.55 \implies q_{0.5} = 3$$

In general



Quantile of order $\alpha \in (0, 1)$.

Quantiles for common distributions

- ▶ Standard Normal distribution $N(0, 1)$: The corresponding cdf is $F_{N(0,1)}(x)$.

Example Quantiles:

- ▶ `norm.ppf(0.01) = -2.3263, norm.ppf(1 - 0.01) = 2.3263` after importing `norm` from `scipy.stats` in Python.

Proposition 20

For the quantiles of the standard normal distribution $N(0, 1)$, we have

$$z_\alpha = -z_{1-\alpha} \quad \text{for each } \alpha \in (0, 1).$$

CI for the mean when the variance is known

Proposition 21

Let X_1, X_2, \dots, X_n be random variables representing a characteristic X with a normal distribution or with a sufficiently large sample size ($n > 30$). If the sample size is large enough, the following holds for the sample mean:

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1),$$

where $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ is the sample mean, $\mu = \mathbb{E}[X]$ is the theoretical mean, and σ is the population standard deviation.

- ▶ Given: $\alpha \in (0, 1)$, σ , and the sample data x_1, \dots, x_n .
- ▶ X_1, \dots, X_n are sample variables for the characteristic X and X follows a normal distribution or $n > 30$.
- ▶ The goal is to construct a confidence interval (CI) for the unknown parameter μ (the theoretical mean), when the variance σ^2 is known.

Constructing the Confidence Interval

Using the previous proposition, we can construct a confidence interval by using the following variable Z (*pivot*):

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

The quantiles of the standard normal distribution $N(0, 1)$ are:

$$z_{1-\frac{\alpha}{2}} = \text{norm. ppf}\left(1 - \frac{\alpha}{2}\right), \quad z_{1-\alpha} = \text{norm. ppf}(1-\alpha), \quad z_\alpha = \text{norm. ppf}(\alpha)$$

after importing norm from scipy.stats in Python. The three quantiles $z_{1-\frac{\alpha}{2}}$, $z_{1-\alpha}$, and z_α are used in different contexts depending on whether you're working with a two-tailed test or a one-tailed test, and whether you're defining a confidence interval or evaluating a hypothesis.

Two-sided Confidence Interval for μ : When the variance σ^2 is known, the $100(1 - \alpha)\%$ confidence interval for μ is:

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \right].$$

Justification for the Confidence Interval

To justify this two-sided confidence interval, we consider the following probability:

$$\mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}\right)$$

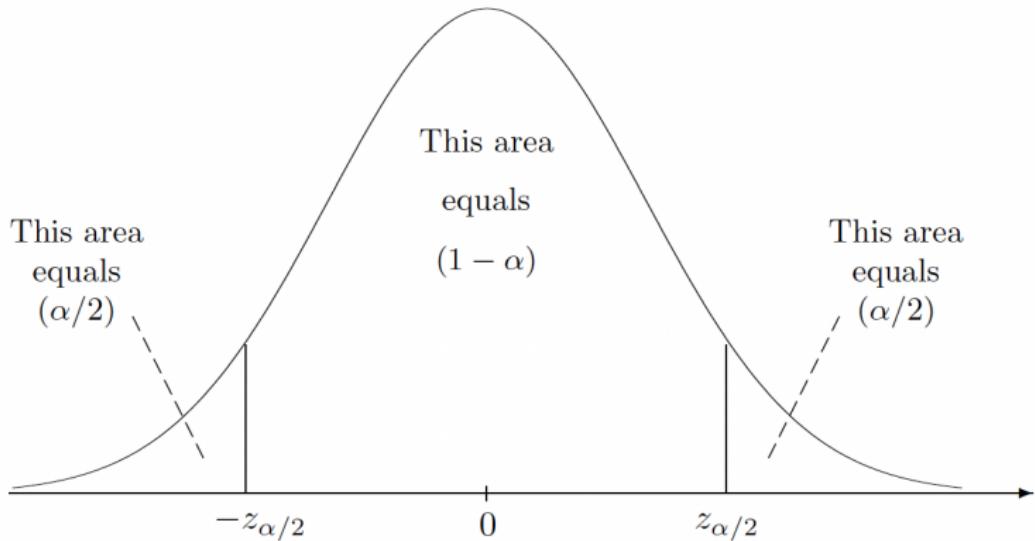
which is equivalent to finding:

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}}\right).$$

This is the probability that the standardized variable lies between $-z_{1-\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$, which is:

$$F_{N(0,1)}\left(z_{1-\frac{\alpha}{2}}\right) - F_{N(0,1)}\left(-z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Thus, the confidence interval has the desired confidence level of $1 - \alpha$.



Standard normal quantiles [Baron]

$$\mathbb{P}(Y \leq z_{\alpha/2}) = \frac{\alpha}{2} + 1 - \alpha = 1 - \frac{\alpha}{2}$$

\Rightarrow this quantile is sometimes denoted by $z_{1-\alpha/2}$ (e.g. on the previous slide).

What is a Z-score?

- ▶ A **z-score** (or **standard score**) is a numerical measurement that describes a data point's position relative to the mean of a dataset, measured in terms of standard deviations.
- ▶ Formula for z-score for a data point X :

$$z = \frac{X - \mu}{\sigma}$$

- ▶ X : the value you are standardizing
- ▶ μ : the mean of the dataset
- ▶ σ : the standard deviation of the dataset

Interpreting Z-scores

- ▶ $z = 0$: The data point is exactly at the mean.
- ▶ $z > 0$: The data point is above the mean.
- ▶ $z < 0$: The data point is below the mean.
- ▶ $z = 1$: The data point is one standard deviation above the mean.
- ▶ $z = -1$: The data point is one standard deviation below the mean.

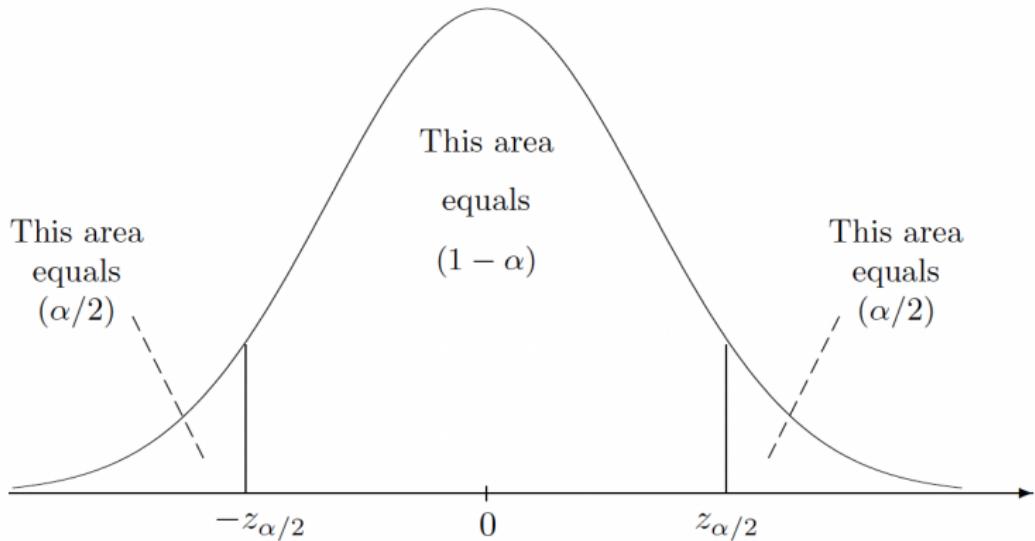
Example: If a student's test score has a z-score of 1.5, it means their score is 1.5 standard deviations above the average score of the group.

Why Z-scores Are Useful

- ▶ **Standardize values across different datasets:** Converts data points into a common scale, allowing comparisons across different distributions.
- ▶ **Determine probabilities in a normal distribution:** z-scores correspond to specific probabilities, helping determine how likely a data point is to occur (via quantiles).

For example:

- ▶ About 68% of values in a normal distribution lie within 1 standard deviation of the mean (z-scores between -1 and 1).
- ▶ About 95% lie within 2 standard deviations (z-scores between -2 and 2).
- ▶ About 99.7% lie within 3 standard deviations (z-scores between -3 and 3).



Standard normal quantiles [Baron]

$$\mathbb{P}(Y \leq z_{\alpha/2}) = \frac{\alpha}{2} + 1 - \alpha = 1 - \frac{\alpha}{2}$$

\Rightarrow this quantile is sometimes denoted by $z_{1-\alpha/2}$ (e.g. on the previous slide).

Understanding the Probability $F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$

- ▶ In a standard normal distribution (mean 0, std. deviation 1), $z_{1-\frac{\alpha}{2}}$ is a z-score corresponding to an upper-tail probability of $\frac{\alpha}{2}$.
- ▶ This means the area under the curve to the right of $z_{1-\frac{\alpha}{2}}$ is $\frac{\alpha}{2}$.
- ▶ The cumulative distribution function $F_{N(0,1)}(z)$ gives the probability that a standard normal variable Z is less than or equal to z :

$$F_{N(0,1)}(z) = \mathbb{P}(Z \leq z).$$

- ▶ Since the total area under the normal curve is 1, if the right-tail area is $\frac{\alpha}{2}$, then the left area is $1 - \frac{\alpha}{2}$:

$$F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}.$$

Remark: If we are interested in finding the **upper tail probability** for a given z-score z , we're asking:

"What is the probability that a random variable Z will take on a value greater than z ?"

i.e. we want to find $\mathbb{P}(Z > z)$.

Example

- ▶ Suppose $\alpha = 0.05$ (for a 95% confidence interval).
 - ▶ Then $\frac{\alpha}{2} = 0.025$ and $1 - \frac{\alpha}{2} = 0.975$.
 - ▶ The z-score for a cumulative probability of 0.975 is approximately 1.96:
- $$F_{N(0,1)}(1.96) = 0.975 = 1 - 0.025$$

- ▶ **General Case:** For any α , $z_{1-\frac{\alpha}{2}}$ is defined so that

$$F_{N(0,1)}(z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$$

- ▶ This property is essential for calculating two-sided confidence intervals.

One-Sided Confidence Intervals

In addition to the two-sided confidence interval, we can also construct one-sided confidence intervals for μ when the variance is known:

- ▶ **Lower confidence interval:** A $100(1 - \alpha)\%$ lower confidence interval for μ is:

$$\left(-\infty, \bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha}\right],$$

such that

$$\mathbb{P}\left(\mu \leq \bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha}\right) = 1 - \alpha.$$

- ▶ **Upper confidence interval:** A $100(1 - \alpha)\%$ upper confidence interval for μ is:

$$\left[\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha}, \infty\right),$$

such that

$$\mathbb{P}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\alpha} \leq \mu\right) = 1 - \alpha.$$

Example: Confidence Interval for Student Test Scores

A teacher has been recording his students' test scores for several years. The scores range from 0 to 100, with a **known** population standard deviation of 10. For a sample of 144 students, the sample mean of the test scores is 68. We are asked to construct a two-sided confidence interval for the population mean $\mu = \mathbb{E}[X]$ of the students' test scores, using a significance level of $\alpha = 0.05$.

Solution: The two-sided confidence interval for the population mean is given by the formula:

$$\left[\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \right]$$

where:

- ▶ $n = 144$: sample size
- ▶ $\sigma = 10$: population standard deviation
- ▶ $\bar{x} = 68$: sample mean
- ▶ $\alpha = 0.05$
- ▶ $z_{1-\frac{\alpha}{2}} = \text{norm. ppf}\left(1 - \frac{0.05}{2}\right) \approx 1.96$

Substituting the values, the confidence interval is:

$$\left[68 - \frac{10}{\sqrt{144}} \cdot 1.96, 68 + \frac{10}{\sqrt{144}} \cdot 1.96 \right]$$

This simplifies to the confidence interval:

$$[66.367, 69.633]$$

The two-sided confidence interval for the theoretical mean test score μ of the students is [66.367, 69.633]. We therefore report the interval [66.367, 69.633] as our estimate for the population mean, with a confidence level of 95% for the method used to construct it. That is, the confidence interval was obtained using a method which, over many repeated samples, yields intervals that contain the true mean in 95% of cases.

- ▶ The length of the interval reflects the precision of the estimate. A smaller interval would indicate a more precise estimate, but this would require a larger sample size or a lower confidence level.

Selecting the sample size for a desired precision

When estimating a population parameter, particularly the mean, we often wish to control the precision of our estimate. Specifically, we might want to know how large a sample is required to ensure that the confidence interval (CI) for the population mean has a length no greater than a specified value Δ .

- ▶ The length of a two-sided confidence interval for the mean, when the variance σ^2 is known, is given by:

$$\frac{2\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}}$$

where n is the sample size and $z_{1-\frac{\alpha}{2}}$ is the critical value from the standard normal distribution corresponding to the confidence level.

The problem we need to solve is: *What sample size n ensures that the length of the confidence interval does not exceed a given margin of error Δ ?*

Solving for the required sample size

To ensure the length of the confidence interval is within the desired precision Δ , we set up the following inequality:

$$\frac{2\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \leq \Delta$$

Rearranging this inequality to solve for n , we get:

$$\left(\frac{2\sigma z_{1-\frac{\alpha}{2}}}{\Delta} \right)^2 \leq n$$

Thus, to guarantee a CI length is no greater than Δ , the required sample size n must satisfy this inequality.

- ▶ The sample size n depends on:
 - ▶ The population standard deviation σ
 - ▶ The desired precision Δ
 - ▶ The confidence level (which determines $z_{1 - \frac{\alpha}{2}}$).
- ▶ As the desired precision Δ decreases, the required sample size n must increase to maintain the same level of confidence.

The formula above a practical way to determine the sample size necessary for estimating the population mean with a specified margin of error.

Example: Confidence Interval and Sample Size Calculation

Consider a sample of measurements: 2.5, 7.4, 8.0, 4.5, 7.4, 9.2, drawn from a normal distribution. We know that the measurement device guarantees a standard deviation $\sigma = 2.2$.

Part (a): Construct a 95% Confidence Interval

- ▶ The sample size is $n = 6$, and the sample mean is $\bar{x} = 6.5$.
- ▶ The desired confidence level is 95%, so $\alpha = 0.05$ and $\alpha/2 = 0.025$.
- ▶ The critical value from the standard normal distribution is $z_{0.975} = 1.96$.

- The formula for the confidence interval is:

$$\left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

- Using the data:

$$\left[6.5 - 1.96 \cdot \frac{2.2}{\sqrt{6}}, 6.5 + 1.96 \cdot \frac{2.2}{\sqrt{6}} \right] \approx [4.74, 8.26]$$

Therefore, the 95% confidence interval for the population mean is [4.74, 8.26].

Determining the Sample Size

Part (b): How many measurements should be taken in order for the length of the 95% confidence interval for the mean to not exceed 1?

- ▶ The length of the confidence interval in part (a) is approximately 3.52, which is relatively large. We want to improve the precision of the estimate.
- ▶ We can use the following formula to determine the required sample size:

$$\frac{2\sigma}{\sqrt{n}} \cdot z_{1-\frac{\alpha}{2}} \leq \Delta$$

where $\Delta = 1$ is the desired margin of error.

- ▶ Substituting the known values $\sigma = 2.2$, $z_{0.975} = 1.96$, and $\Delta = 1$, we solve for n :

$$\left(\frac{2 \cdot 2.2 \cdot 1.96}{1} \right)^2 \leq n.$$

Simplifying:

$$74.37 \leq n$$

Thus, a sample size of at least 75 measurements is required to ensure that the length of the 95% confidence interval does not exceed 1.

Student's distribution $T(n)$

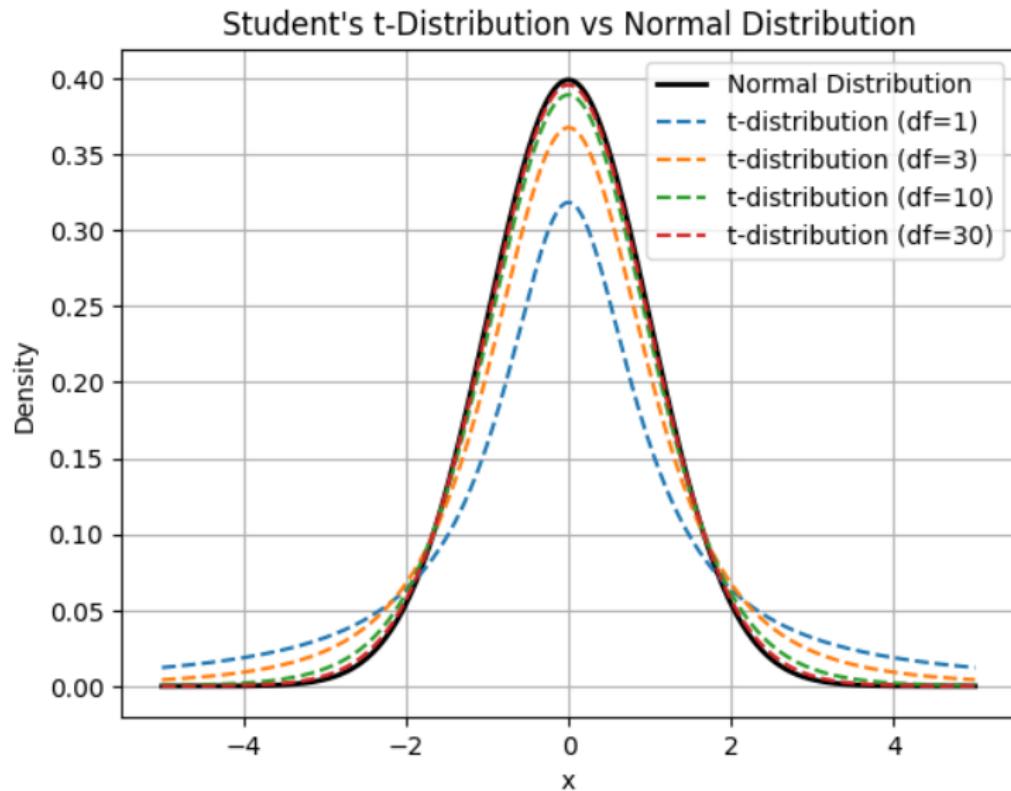
- ▶ Also called the *t-distribution*.
- ▶ It is a continuous probability distribution used mainly to estimate population parameters when the sample size is small and/or the population variance is **unknown**. It is similar to the normal distribution but has heavier tails \Rightarrow is more prone to producing values that fall far from its mean.
- ▶ Symmetric around zero, like the normal distribution; widely used in hypothesis testing (such as t-tests) and for constructing confidence intervals for small sample sizes; helps to account for extra variability in smaller samples

Proposition 22

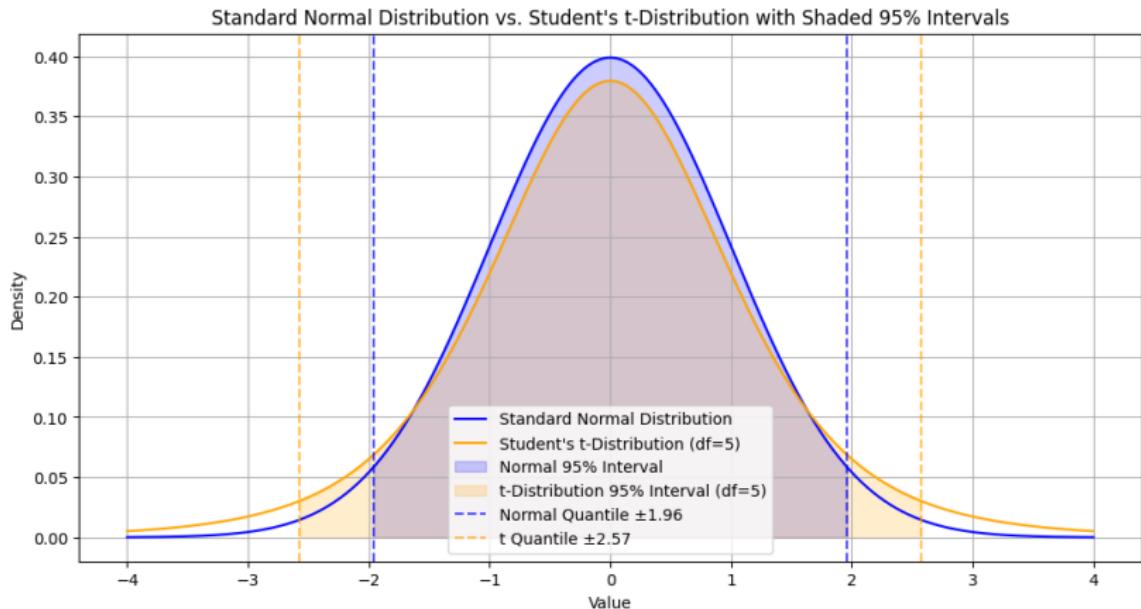
For the quantiles of the Student distribution $T(n)$ it holds that

$$t_\alpha = -t_{1-\alpha} \text{ for each } \alpha \in (0, 1).$$

Student's vs Normal Distribution



Student's vs. Normal - Quantiles



Confidence Interval for the Mean - Unknown Variance

Proposition 23

Let X_1, \dots, X_n be sample variables for a characteristic X . If the sample size is large enough ($n > 30$), then for the sample mean:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim T(n-1),$$

where:

- ▶ $\bar{X} = \frac{1}{n} \sum_{k=1}^n X_k$ is the sample mean.
- ▶ $S = \left(\frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2 \right)^{\frac{1}{2}}$ is the sample standard deviation.

Conditions:

- ▶ Given $\alpha \in (0, 1)$ and statistical data x_1, \dots, x_n .
- ▶ X follows a normal distribution or $n > 30$.
- ▶ Construct a confidence interval (CI) for the unknown mean $\mu = \mathbb{E}[X]$ when variance $V(X)$ is unknown.

Confidence Interval Construction

Using the previous proposition, consider the pivot:

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim T(n-1).$$

Quantiles for the Student's $T(n-1)$ distribution:

$$t_{1-\frac{\alpha}{2}} = \text{t.ppf}\left(1 - \frac{\alpha}{2}, n-1\right), \quad t_{1-\alpha} = \text{t.ppf}(1-\alpha, n-1), \quad t_\alpha = \text{t.ppf}(\alpha, n-1)$$

Two-sided CI for μ (variance unknown):

$$\left[\bar{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{X} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} \right]$$

which implies:

$$\mathbb{P}\left(\bar{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} \leq \mu \leq \bar{X} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha.$$

One-Sided Confidence Intervals for μ (Variance Unknown)

One-sided CI for μ , **when the variance is unknown:**

$$\left(-\infty, \bar{X} - \frac{S}{\sqrt{n}} \cdot t_{\alpha}\right], \quad \left[\bar{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\alpha}, \infty\right),$$

i.e.,

$$\mathbb{P}\left(\mu \leq \bar{X} - \frac{S}{\sqrt{n}} \cdot t_{\alpha}\right) = 1 - \alpha, \quad \mathbb{P}\left(\bar{X} - \frac{S}{\sqrt{n}} \cdot t_{1-\alpha} \leq \mu\right) = 1 - \alpha.$$

Example: Constructing a Two-Sided Confidence Interval

The sample mean of the length of 100 screws is 15.5 cm, and the empirical variance is 0.09 cm^2 . Construct a 99% two-sided confidence interval for the theoretical mean of the length of a screw.

Solution:

$$\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}}, \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} \right]$$

where $\bar{x} = 15.5$, $s = 0.3$ ($s^2 = 0.09$), $\alpha = 0.01$,
 $t_{1-\frac{\alpha}{2}} = \text{scipy.stats.t.ppf}(0.995, 99)$ in Python, and $\sqrt{n} = 10$.

By using the statistical data, the two-sided CI for the theoretical mean of the length of a screw is:

$$[15.421208, 15.578792].$$

Confidence Interval for the Variance

Proposition 24

Let X_1, \dots, X_n be sample variables for a characteristic $X \sim N(\mu, \sigma^2)$. Then for the sample variance we have that

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$$

where:

- ▶ $S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2$ is the sample variance.

Conditions:

- ▶ Given $\alpha \in (0, 1)$ and statistical data x_1, \dots, x_n ,
- ▶ X is normally distributed,
- ▶ We construct a confidence interval (CI) for the unknown variance $\sigma^2 = V(X)$.

$\chi^2(n)$ Distribution

- ▶ The χ^2 (chi-squared) distribution is a continuous probability distribution that arises mainly when dealing with variances, goodness-of-fit tests, and hypothesis testing in general.
- ▶ Used to: construct confidence intervals for population variances, test hypotheses about population variances when samples are drawn from normal distributions, decide how well observed data fits a specific theoretical distribution (e.g., in categorical data analysis), determine whether there is a significant association between two categorical variables.

Definition and Degrees of Freedom

The χ^2 distribution is defined as the sum of the squares of k independent standard normal random variables, where k is the *degrees of freedom* (often written as ν or df).

$$\chi_k^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

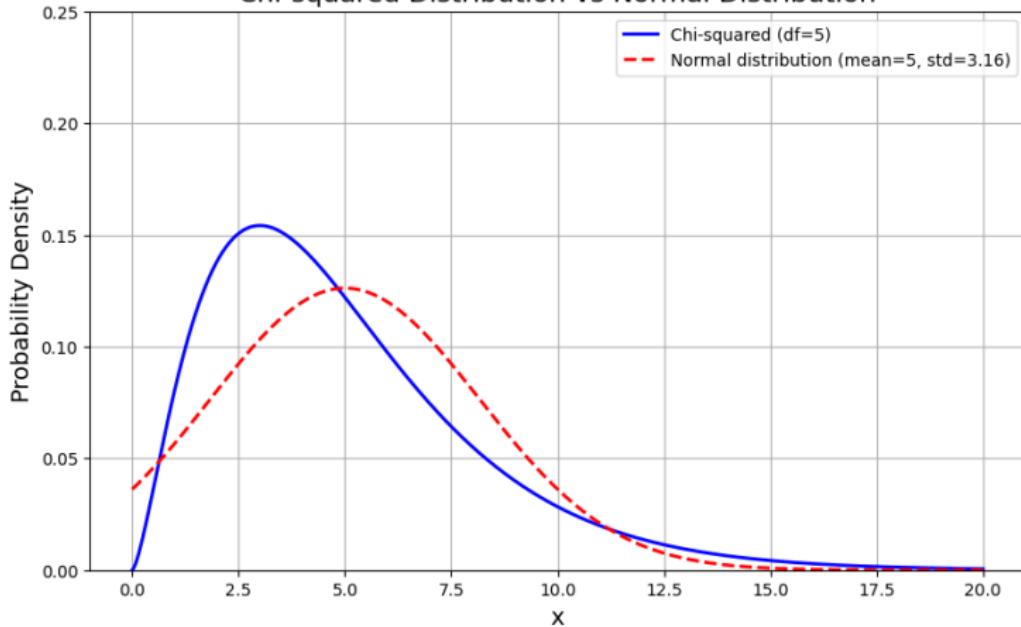
Here, Z_1, Z_2, \dots, Z_k are independent, standard normal random variables, each with mean 0 and variance 1. The degrees of freedom k influence the

shape of the distribution:

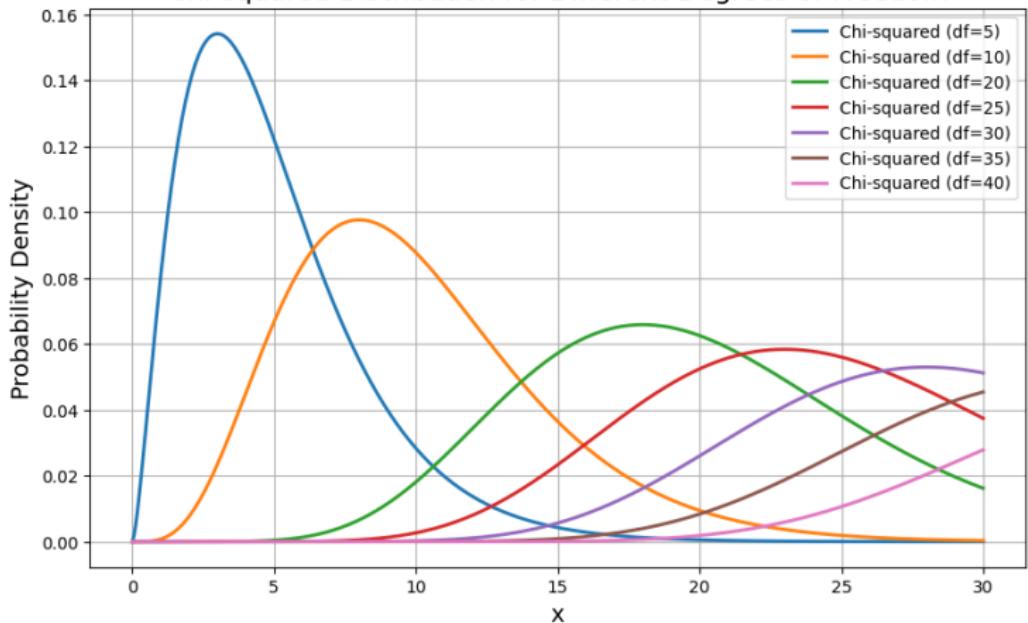
- ▶ Smaller k : The distribution is more skewed.
- ▶ Larger k : The distribution becomes more symmetrical.

The distribution becomes more normal in appearance as k increases.

Chi-squared Distribution vs Normal Distribution



Chi-squared Distribution for Different Degrees of Freedom



Probability Density Function (PDF)

The probability density function of a χ^2 distribution with k degrees of freedom is:

$$f(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)}x^{(k/2)-1}e^{-x/2}$$

where $x \geq 0$, and Γ denotes the gamma function.

- ▶ **Mean:** The mean of a χ^2 distribution with k degrees of freedom is k .
- ▶ **Variance:** The variance is $2k$.
- ▶ **Mode:** For $k > 2$, the mode is $k - 2$; otherwise, the mode is 0.

R: Gamma function Γ

The Gamma function for a real number $x > 0$ is defined as:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

- ▶ Can be seen as a generalisation the concept of a factorial to non-integer values.
- ▶ **Factorial Relation:** For a positive integer n ,

$$\Gamma(n) = (n - 1)!$$

- ▶ **Recursive Property:** The Gamma function satisfies the following recursive relationship:

$$\Gamma(x + 1) = x\Gamma(x).$$

Confidence Interval for the Variance

Proposition 25

Let X_1, \dots, X_n be sample variables for a characteristic $X \sim N(\mu, \sigma^2)$. Then for the sample variance we have that

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$$

where:

- ▶ $S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2$ is the sample variance.

Conditions:

- ▶ Given $\alpha \in (0, 1)$ and statistical data x_1, \dots, x_n ,
- ▶ X is normally distributed,
- ▶ We construct a confidence interval (CI) for the unknown variance $\sigma^2 = V(X)$.

Confidence Interval Construction

Using Proposition 6, we use the pivot:

$$Q = \frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1).$$

Quantiles for the Chi-Squared distribution $\chi^2(n-1)$. To express the quantiles in Python, we use the `scipy.stats.chi2` module:

$$q_{1-\frac{\alpha}{2}} = \text{chi2.ppf}\left(1 - \frac{\alpha}{2}, n-1\right), \quad q_{\frac{\alpha}{2}} = \text{chi2.ppf}\left(\frac{\alpha}{2}, n-1\right),$$

$$q_{1-\alpha} = \text{chi2.ppf}(1 - \alpha, n-1), \quad q_\alpha = \text{chi2.ppf}(\alpha, n-1).$$

Two-sided CI for $\sigma^2 = V(X)$:

$$\left[\frac{n-1}{q_{1-\frac{\alpha}{2}}} \cdot S^2, \frac{n-1}{q_{\frac{\alpha}{2}}} \cdot S^2 \right],$$

i.e.,

$$\mathbb{P}\left(\frac{n-1}{q_{1-\frac{\alpha}{2}}} \cdot S^2 \leq \sigma^2 \leq \frac{n-1}{q_{\frac{\alpha}{2}}} \cdot S^2\right) = 1 - \alpha.$$

One-Sided Confidence Intervals for σ^2

One-sided CI for the theoretical variance $\sigma^2 = V(X)$:

$$\left(0, \frac{n-1}{q_\alpha} \cdot S^2\right], \quad \left[\frac{n-1}{q_{1-\alpha}} \cdot S^2, \infty\right),$$

i.e.,

$$\mathbb{P}\left(\sigma^2 \leq \frac{n-1}{q_\alpha} \cdot S^2\right) = 1 - \alpha, \quad \mathbb{P}\left(\frac{n-1}{q_{1-\alpha}} \cdot S^2 \leq \sigma^2\right) = 1 - \alpha.$$

Example: Confidence Interval for Variance

The sample mean of the length of 100 screws is 15.5 cm, and the sample variance is 0.09 cm². Construct a 99% two-sided confidence interval for the theoretical variance of screw lengths. If the variance is too large (i.e., over 0.099 cm²), the machine producing the screws must be adjusted. The lengths are assumed to follow a normal distribution.

Solution:

$$\left[\frac{n-1}{q_{1-\frac{\alpha}{2}}} \cdot s^2, \frac{n-1}{q_{\frac{\alpha}{2}}} \cdot s^2 \right]$$

where:

- ▶ $\bar{x} = 15.5$,
- ▶ $s^2 = 0.09$,
- ▶ $\alpha = 0.01$,
- ▶ $q_{1-\frac{\alpha}{2}} = \text{scipy.stats.chi2.ppf}(0.995, 99) = 138.99$
- ▶ $q_{\frac{\alpha}{2}} = \text{scipy.stats.chi2.ppf}(0.005, 99) = 66.510$

Calculating the Interval

Using the statistical data, the two-sided CI for the theoretical variance is:

$$[0.064107, 0.133965].$$

Conclusion: This interval contains values above 0.099, so the machine producing the screws must be adjusted.

Confidence Interval for Proportions

We assume a subpopulation A of items that have a certain attribute. By the **population proportion** we mean the probability

$$p = \mathbb{P}\{i \in A\}$$

for a randomly selected item i that has this attribute. A **sample proportion**

$$\hat{p} = \frac{\text{number of sampled items from } A}{n}$$

is used to estimate p .

We can use the indicator variables

$$X_i = \begin{cases} 1 & \text{if } i \in A \\ 0 & \text{if } i \notin A \end{cases}$$

Each X_i has Bernoulli distribution with parameter p . In particular,

$$\mathbb{E}(X_i) = p \quad \text{and} \quad \text{Var}(X_i) = p(1 - p).$$

Then

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

corresponds to a sample mean of X_i .

Confidence Interval for Proportions

Conditions:

- ▶ Given $\alpha \in (0, 1)$ and statistical data x_1, \dots, x_n ; compute $\bar{p} := \bar{x}$
- ▶ Let X_1, \dots, X_n be sample variables for the studied characteristic $X \sim \text{Bernoulli}(p)$
- ▶ We construct a confidence interval for the unknown parameter $p \in (0, 1)$.

Since $X \sim \text{Bernoulli}(p)$, then by a previous proposition (and CLT):

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1), \quad \text{for large enough } n.$$

Quantiles of the Standard Normal Distribution

The quantiles of the standard normal distribution $N(0, 1)$ are:

$$z_{1-\frac{\alpha}{2}} = \text{norminv}\left(1 - \frac{\alpha}{2}\right), \quad z_{1-\alpha} = \text{norminv}(1-\alpha), \quad z_\alpha = \text{norminv}(\alpha).$$

- ▶ Use the `scipy.stats.norm.ppf` to calculate them in Python.

Goal: Construct a two-sided CI for p by finding estimators $\theta_L = \theta_L(X_1, \dots, X_n)$ and $\theta_U = \theta_U(X_1, \dots, X_n)$ such that:

$$\mathbb{P}(\theta_L \leq p \leq \theta_U) = 1 - \alpha.$$

Derivation of the CI for p

Starting from:

$$\mathbb{P}\left(-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha$$

we have:

$$\left| \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| < z_{1-\frac{\alpha}{2}} \Rightarrow \bar{X} - \sqrt{\frac{p(1-p)}{n}} \cdot z_{1-\frac{\alpha}{2}} < p < \bar{X} + \sqrt{\frac{p(1-p)}{n}} \cdot z_{1-\frac{\alpha}{2}}.$$

Since \bar{X} is an unbiased and consistent estimator for p , we replace p by $\bar{p} = \bar{X}$.

Two-sided Confidence Interval for p

The two-sided CI for p is:

$$\left[\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{X} + \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_{1-\frac{\alpha}{2}} \right]$$

i.e.,

$$\mathbb{P}\left(\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_{1-\frac{\alpha}{2}} \leq p \leq \bar{X} + \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_{1-\frac{\alpha}{2}} \right) = 1 - \alpha.$$

One-sided Confidence Interval for p

The one-sided CI for p is:

$$\left(0, \bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_\alpha \right], \quad \left[\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_{1-\alpha}, 1 \right)$$

i.e.

$$\mathbb{P} \left(p \leq \bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_\alpha \right) = 1-\alpha, \quad \mathbb{P} \left(\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \cdot z_{1-\alpha} \leq p \right)$$

Example: Confidence Interval for Population Proportion

A random sample of 2000 residents of a city was surveyed, and 980 people indicated support for a particular candidate in the upcoming election. We wish to construct a two-sided confidence interval at the 95% confidence level for the population proportion p of supporters.

Solution: The two-sided CI for p is:

$$\left[\bar{p} - \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \cdot z_{1-\frac{\alpha}{2}}, \bar{p} + \sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \cdot z_{1-\frac{\alpha}{2}} \right] \cap (0, 1)$$

where:

- ▶ $n = 2000$
- ▶ $\alpha = 0.05$
- ▶ $\bar{p} = \frac{980}{2000} = 0.49$
- ▶ $z_{1-\frac{\alpha}{2}} = \text{scipy.stats.norm.ppf}\left(1 - \frac{0.05}{2}\right) \approx 1.96.$

Using the provided data, the two-sided CI for p is:

$$[0.4678, 0.5122].$$

Conclusion: With 95% confidence, we estimate that the proportion of supporters in the city's population lies between 46.78% and 51.22%.

Remark: In general, $z_\alpha = q_{1-\alpha} = F_{N(0,1)}^{-1}(1 - \alpha)$ is the value of a Standard Normal variable Z that is exceeded with probability α

LECTURES 9 and 10

Hypothesis Testing

Scenario: The Coffee Shop Debate

Imagine you own a coffee shop, and one of your baristas claims that offering free samples of your new latte flavor will increase overall sales. You're skeptical. To test this claim, you decide to run a trial. For one week, you offer free samples of the new latte flavor and record your daily sales. You then compare these sales to the data from the previous week when no samples were offered.

- ▶ How can you determine if the increase in sales (if any) is truly due to the free samples or just random chance?
- ▶ What if some other factor, like better weather or a holiday, caused the change in sales?

This situation introduces the need for **hypothesis testing**:

- ▶ **Null hypothesis (H_0)**: The free samples have no effect on sales.
- ▶ **Alternative hypothesis (H_1, H_A)**: The free samples increase sales.

Hypothesis testing is a structured way to make decisions based on data, rather than assumptions or gut feelings.

► Key Concepts:

- ▶ **Null Hypothesis (H_0)**: A tentative assumption about the population parameter (θ) or distribution of X .
- ▶ **Alternative Hypothesis (H_1)**: Contradicts H_0 , proposing a different assumption.
- ▶ H_0 and H_1 are two mutually exclusive statements. Each test results either in acceptance of H_0 or its rejection in favor of H_1 .

► Procedure:

1. Start with H_0 .
2. Analyze sample data.
3. Decide whether to reject H_0 in favor of H_1 .
4. If H_0 is rejected, conclude H_1 is true.

Example

A radio station *assumes* its average listener age is 30 years.

- ▶ **Null Hypothesis (H_0):** $\mu = 30$.
- ▶ **Alternative Hypothesis (H_1):** $\mu \neq 30$.

Process:

1. Collect a sample of listener ages and calculate the sample mean (\bar{x}).
2. Compare \bar{x} to 30:
 - ▶ If \bar{x} is "close" to 30, H_0 is supported.
 - ▶ If \bar{x} is "not close" to 30, H_0 is rejected, and H_1 is supported.

Applications of Hypothesis Testing

Common Uses:

- ▶ Verify claims such as:
 - ▶ Average connection speed meets provider's promises.
 - ▶ Proportion of defective items is within acceptable limits.
 - ▶ Service times match the stated distribution.

Broader Applications:

- ▶ Proving the efficiency of a new medical treatment.
- ▶ Testing the safety of new vehicles.
- ▶ Determining the innocence of a defendant in legal cases.
- ▶ Verifying authorship of documents.

Key Idea:

- ▶ To reject H_0 , strong evidence in favor of H_1 is required.
- ▶ This evidence must come from data.

Testing a Hypothesis

Key Components:

- ▶ Test Statistic (TS):

- ▶ A sample function used to evaluate the evidence. The choice of the test statistic depends on the data type, the distribution of the population, the sample size, and the specific hypothesis that is being tested; it is a standardized value that helps us compare the observed sample data to what we would expect under the null hypothesis → See also Appendix A.

- ▶ Rejection Region (RR):

- ▶ The set of test statistic values where H_0 is rejected.
- ▶ If the TS falls within RR, reject H_0 .
- ▶ Otherwise, fail to reject H_0 .
- ▶ See also Appendix B.

Types of Hypothesis Tests

Standard Form of Hypotheses:

- ▶ Null Hypothesis: $H_0 : \theta = \theta_0$.
- ▶ Alternative Hypotheses:
 - I. $H_1 : \theta \neq \theta_0$ (**Two-tailed test**).
 - II. $H_1 : \theta > \theta_0$ (**Right-tailed test**).
 - III. $H_1 : \theta < \theta_0$ (**Left-tailed test**).

Special Cases:

- ▶ Consider $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$ for given θ_0 and θ_1 .

Decision Making:

- ▶ If the test statistic lies in the rejection region, reject H_0 .
- ▶ Otherwise, fail to reject H_0 .

Formulating Hypotheses and Designing Tests

- ▶ The first and one of the most crucial steps in a hypothesis testing problem is to define the appropriate null and alternative hypotheses to be evaluated. For instance, to check whether the average broadband internet connection speed is 100 Mbps, we test the following hypothesis:

Two-Sided Test

$$H_0 : \mu = 100$$

$$H_1 : \mu \neq 100$$

One-Sided Test

$$H_0 : \mu = 100$$

$$H_1 : \mu < 100$$

The significance level (or *probability of risk*, α) is the probability of rejecting H_0 when it is true. Typically $\alpha \in (0, 1)$ is given. Designing a hypothesis test means constructing a rejection region $RR \subset \mathbb{R}$ (for each of the cases I, II, III respectively) such that

$$\mathbb{P}(TS \in RR \mid H_0) = \alpha,$$

which is equivalent to

$$\mathbb{P}(TS \notin RR \mid H_0) = 1 - \alpha.$$

Let ts_0 be the value of the test statistic computed by using the statistical data under hypothesis H_0 .

Test Conclusion and Errors

Conclusion of the Test:

- ▶ If $ts_0 \notin RR$: Accept H_0 .
- ▶ If $ts_0 \in RR$: Reject H_0 in favor of H_1 .

Possible Outcomes:

Decision	Actual Situation	
	H_0 true	H_1 true
Reject H_0	Type I error (prob. α)	Correct decision
Accept H_0	Correct decision	Type II error (prob. β)

Key Points:

- ▶ **Type I Error:** Rejecting H_0 when it is true (α).
- ▶ **Type II Error:** Accepting H_0 when H_1 is true (β).

Type I and Type II Errors

Type I Error:

- ▶ Occurs when a true null hypothesis (H_0) is rejected.
- ▶ The probability of a Type I error is:

$$\mathbb{P}(\text{Type I error}) = \mathbb{P}(\text{reject } H_0 \mid H_0) = \mathbb{P}(TS \in RR \mid H_0) = \alpha.$$

where α is the **significance level** or **risk probability**.

Type II Error:

- ▶ Occurs when a false null hypothesis (H_0) is not rejected.
- ▶ The probability of a Type II error is:

$$\mathbb{P}(\text{Type II error}) = \mathbb{P}(\text{accept } H_0 \mid H_1) = \mathbb{P}(TS \notin RR \mid H_1) = \beta.$$

Designing a Hypothesis Test

Remarks:

1. The rejection region (RR) and the test are **not uniquely determined**.
2. Both α and β represent the **risks of making an error**.
 - ▶ Ideally, our goal is to design tests where α and β are both small.
3. Typically, α is **preset**, and an appropriate rejection region is determined based on it.

Z-Test for Theoretical Mean μ

Consider X to be some characteristic (relative to a population). Let \bar{X} be the unbiased estimator for $\theta = \mu = \mathbb{E}[X]$ (where $\mathbb{E}[X]$ is theoretical mean of X), \bar{X} has standard deviation $\frac{\sigma}{\sqrt{n}}$. As for confidence intervals, we start with the case where the test statistic has $N(0, 1)$ distribution. If X has normal distribution or $n > 30$ we consider the test statistic

$$TS := Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

- ▶ For a given significance level $\alpha \in (0, 1)$ and a given value $\theta_0 = \mu_0$, test:

$$H_0 : \mu = \mu_0$$

- ▶ Against one of the alternatives:

- I. $H_1 : \mu \neq \mu_0$
- II. $H_1 : \mu < \mu_0$
- III. $H_1 : \mu > \mu_0$.

Z-Test Procedure

Given Data:

- ▶ Parameters: $\alpha \in (0, 1)$, μ_0 , σ .
- ▶ Sample data: x_1, \dots, x_n , where X_1, \dots, X_n are independent sample variables.
- ▶ Assumption: X is normally distributed or $n > 30$.

Steps:

- ▶ Under H_0 :

$$\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1).$$

- ▶ Compute the test statistic using the data:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

- ▶ Compare z to quantiles of the standard normal distribution $N(0, 1)$.

Rejection Regions for Z-Test

Critical Values:

$$z_{1-\frac{\alpha}{2}}, z_{1-\alpha}, z_\alpha$$

Decision Rules:

	I. $H_1 : \mu \neq \mu_0$	II. $H_1 : \mu < \mu_0$	III. $H_1 : \mu > \mu_0$
Accept H_0 , if	$ z < z_{1-\frac{\alpha}{2}}$	$z > z_\alpha$	$z < z_{1-\alpha}$
Reject H_0 in favor of H_1 , if	$ z \geq z_{1-\frac{\alpha}{2}}$	$z \leq z_\alpha$	$z \geq z_{1-\alpha}$

Connection Between CI and Rejection Regions

Rejection Regions for Hypotheses:

- ▶ I. $H_1 : \mu \neq \mu_0$: Reject H_0 if:

$$z \in RR = \{r \in \mathbb{R} : |r| \geq z_{1-\frac{\alpha}{2}}\}.$$

- ▶ II. $H_1 : \mu < \mu_0$: Reject H_0 if:

$$z \in RR = \{r \in \mathbb{R} : r \leq z_\alpha\}.$$

- ▶ III. $H_1 : \mu > \mu_0$: Reject H_0 if:

$$z \in RR = \{r \in \mathbb{R} : r \geq z_{1-\alpha}\}.$$

Confidence Intervals and Rejection Regions:

- ▶ A $100(1 - \alpha)\%$ CI for μ includes all parameter values μ_0 for which the test statistic does not fall into the rejection region.
- ▶ The CI and the rejection region are **not** complements. The RR pertains to the test statistic, while the CI pertains to the parameter μ .

Example: Testing the Manager's Suspicion

- ▶ A firm expects its employees to average 20 sales per month with $\sigma = 4$.
- ▶ A sales manager fears the average number of monthly sales has dropped below 20 during a recession.
- ▶ A sample of $n = 36$ employees found a mean of $\bar{x} = 19$ sales.
- ▶ Test at $\alpha = 0.05$: Does the data confirm the manager's suspicion?

We have

- ▶ Population: Monthly sales for all employees.
- ▶ Test for the mean μ , using a Z -test ($n > 30$ and σ known).
- ▶ Hypotheses:

$$H_0 : \mu = 20, \quad H_1 : \mu < 20 \quad (\text{left-tailed test}).$$

Type I Error: Conclude $\mu < 20$ when $\mu = 20$ is actually true.

Significance Level: $\alpha = 0.05$.

Test Statistic:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{19 - 20}{\frac{4}{\sqrt{36}}} = -1.5.$$

Rejection Region:

$$RR = (-\infty, z_\alpha] = (-\infty, -1.645].$$

Decision:

- ▶ Observed value of $z = -1.5$.
- ▶ Since $z \notin RR$, we do not reject H_0 .

Conclusion:

- ▶ The evidence from the data is not sufficient to reject H_0 .
- ▶ At the 5% significance level, we accept the null hypothesis.
- ▶ The data does not confirm the manager's suspicion that the average monthly sales are less than 20.

- ▶ While the observed mean is lower, the test does not provide enough evidence to conclude that the true mean is less than 20.

T-Test for Theoretical Mean $\mu = \mathbb{E}[X]$

- ▶ Consider X to be a characteristic of a population.
- ▶ Let X_1, \dots, X_n be the corresponding sample variables.
- ▶ If X follows a normal distribution or $n > 30$, we use the test statistic:

$$T := \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim T(n-1),$$

where \bar{X} is the sample mean and s is the sample standard deviation.

- ▶ For a given level of significance $\alpha \in (0, 1)$ and a given value μ_0 :

$$H_0 : \mu = \mu_0.$$

- ▶ Choose one of the alternative hypotheses:

- I. $H_1 : \mu \neq \mu_0$
- II. $H_1 : \mu < \mu_0$
- III. $H_1 : \mu > \mu_0$

given the significance level α , μ_0 , the hypothesized mean, statistical data x_1, \dots, x_n .

T-Test for Theoretical Mean $\mu = \mathbb{E}[X]$

Under H_0 , the test statistic follows:

$$\frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim T(n-1).$$

Using the sample data x_1, \dots, x_n , compute:

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

Quantiles of the Student's T -distribution:

$$t_{1-\frac{\alpha}{2}}, t_\alpha, t_{1-\alpha}.$$

The T-test is used when the population variance is unknown and the sample size is small or large, as long as X is normally distributed or $n > 30$.

T-Test: Rejection Regions

Decision Rules for Hypothesis Testing:

	I. $H_1 : \mu \neq \mu_0$	II. $H_1 : \mu < \mu_0$	III. $H_1 : \mu > \mu_0$
accept H_0 , if	$ t < t_{1-\frac{\alpha}{2}}$	$t > t_\alpha$	$t < t_{1-\alpha}$
reject H_0 in favour of H_1 , if	$ t \geq t_{1-\frac{\alpha}{2}}$	$t \leq t_\alpha$	$t \geq t_{1-\alpha}$

T-Test: Rejection Regions

Summary of Rejection Regions:

- ▶ I. Reject H_0 in favour of $H_1 : \mu \neq \mu_0$ if:

$$t \in RR = \{r \in \mathbb{R} : |r| \geq t_{1-\frac{\alpha}{2}}\}$$

- ▶ II. Reject H_0 in favour of $H_1 : \mu < \mu_0$ if:

$$t \in RR = \{r \in \mathbb{R} : r \leq t_\alpha\}$$

- ▶ III. Reject H_0 in favour of $H_1 : \mu > \mu_0$ if:

$$t \in RR = \{r \in \mathbb{R} : r \geq t_{1-\alpha}\}$$

Statistical Testing and Confidence Intervals:

- I. $H_1 : \mu \neq \mu_0$:

$$H_0 \text{ is accepted} \Leftrightarrow |t| < t_{1-\frac{\alpha}{2}} \Leftrightarrow \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}} < \mu_0 < \bar{x} + \frac{\sigma}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2}},$$

so μ_0 belongs to the two-sided open confidence interval.

- II. $H_1 : \mu < \mu_0$:

$$H_0 \text{ is accepted} \Leftrightarrow t > t_\alpha \Leftrightarrow \mu_0 < \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot t_\alpha,$$

so μ_0 belongs to the one-sided open confidence interval.

- III. $H_1 : \mu > \mu_0$:

$$H_0 \text{ is accepted} \Leftrightarrow t < t_{1-\alpha} \Leftrightarrow \bar{x} - \frac{\sigma}{\sqrt{n}} \cdot t_{1-\alpha} < \mu_0,$$

so μ_0 belongs to the one-sided open confidence interval.

χ^2 -Test for Standard Deviation

Goal: Test for the standard deviation $\text{Std}(X)$ (or variance $V(X)$).

Setup:

- ▶ Let X be some characteristic relative to a population.
- ▶ Let X_1, \dots, X_n be the corresponding sample variables.
- ▶ Assume $X \sim \mathcal{N}(\mu, \sigma^2)$ (normal distribution).

We consider the test statistic

$$TS = Q = \frac{n-1}{\sigma^2} s^2 \sim \chi^2(n-1),$$

where s^2 is the sample variance.

Given Inputs:

- ▶ Significance level $\alpha \in (0, 1)$,
- ▶ Hypothesized value σ_0 ,
- ▶ Observed sample data x_1, \dots, x_n .

χ^2 -Test: Hypotheses

Hypotheses: We test the population standard deviation σ with

$$H_0 : \sigma = \sigma_0$$

against one of the following alternatives:

$$\begin{cases} \text{I. } H_1 : \sigma \neq \sigma_0 & \text{(two-sided test)} \\ \text{II. } H_1 : \sigma < \sigma_0 & \text{(left-tailed test)} \\ \text{III. } H_1 : \sigma > \sigma_0 & \text{(right-tailed test).} \end{cases}$$

How do we do this:

- ▶ Specify α and the hypothesized σ_0 .
- ▶ Compute the test statistic:

$$Q = \frac{n - 1}{\sigma^2} s^2.$$

- ▶ Compare Q to critical values from the χ^2 -distribution with $n - 1$ degrees of freedom.

Test Statistic and Quantiles

We have seen before that

$$\frac{n-1}{\sigma^2} s^2 \sim \chi^2(n-1).$$

Using the statistical data x_1, \dots, x_n , we compute

$$q = \frac{n-1}{\sigma_0^2} \cdot s^2.$$

The quantiles of interest of the $\chi^2(n-1)$ distribution are

$$q_{1-\frac{\alpha}{2}}, q_{\frac{\alpha}{2}}, q_{1-\alpha}, q_\alpha.$$

Acceptance and Rejection Regions

	I. $H_1 : \sigma \neq \sigma_0$	II. $H_1 : \sigma < \sigma_0$	III. $H_1 : \sigma > \sigma_0$
Accept H_0 , if	$q_{\frac{\alpha}{2}} < q < q_{1-\frac{\alpha}{2}}$	$q > q_\alpha$	$q < q_{1-\alpha}$
Reject H_0 in favor of H_1 , if	$q \notin (q_{\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}})$	$q \leq q_\alpha$	$q \geq q_{1-\alpha}$

Remember that the rejection regions for the three alternatives are

1. Rejection Regions for Alternatives:

- ▶ I: $q \in \left\{ r \in \mathbb{R} : r \leq q_{\frac{\alpha}{2}} \text{ or } q_{1-\frac{\alpha}{2}} \leq r \right\}$
- ▶ II: $q \in \{r \in \mathbb{R} : r \leq q_\alpha\}$
- ▶ III: $q \in \{r \in \mathbb{R} : r \geq q_{1-\alpha}\}$

2. Statistical Testing and Confidence Intervals (CI):

- ▶ I: σ_0 belongs to a two-sided open CI.
- ▶ II: σ_0 belongs to a one-sided open CI.
- ▶ III: σ_0 belongs to a one-sided open CI.

3. Variance Test: Similar approach is used for testing $\sigma^2 \neq \sigma_0^2$.

Test for Proportions

Let $X \sim \text{Bernoulli}(p)$ be the studied characteristic, where p is an unknown parameter. Let $p_0 \in (0, 1)$ be given. We consider the hypothesis $H_0 : p = p_0$ with one of the alternatives:

$$\begin{cases} \text{I. } H_1 : p \neq p_0 \\ \text{II. } H_1 : p < p_0 \\ \text{III. } H_1 : p > p_0 \end{cases}$$

- ▶ Given: $\alpha \in (0, 1)$, the statistical data x_1, \dots, x_n .
- ▶ Let X_1, \dots, X_n be sample variables for the studied characteristic $X \sim \text{Bernoulli}(p)$.
- ▶ Since $X \sim \text{Bernoulli}(p)$, we know that:

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1) \quad \text{for large enough } n.$$

- ▶ Using the statistical data x_1, \dots, x_n , we compute $\bar{p} := \bar{x}$ and the test statistic:

$$z = \frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- ▶ Verify if $np_0(1 - p_0) \geq 10$. (See Appendix C.)

Test for Proportions: Hypothesis Testing

	I. $H_1 : p \neq p_0$	II. $H_1 : p < p_0$	III. $H_1 : p > p_0$
accept H_0 , if	$ z < z_{1-\frac{\alpha}{2}}$	$z > z_\alpha$	$z < z_{1-\alpha}$
reject H_0 in favour of H_1 , if	$ z \geq z_{1-\frac{\alpha}{2}}$	$z \leq z_\alpha$	$z \geq z_{1-\alpha}$

Test for Proportions: Statistical Testing and Confidence Interval

Remark:

(1) Statistical Testing and Confidence Interval:

- I. $H_1 : p \neq p_0$:

H_0 is accepted $\Leftrightarrow |z| < z_{1-\frac{\alpha}{2}}$ \Leftrightarrow

$$\bar{p} - \sqrt{\frac{p_0(1-p_0)}{n}} \cdot z_{1-\frac{\alpha}{2}} < p_0 < \bar{p} + \sqrt{\frac{p_0(1-p_0)}{n}} \cdot z_{1-\frac{\alpha}{2}}$$

i.e. p_0 belongs to the two-sided confidence interval (CI).

Test for Proportions: Statistical Testing and CI

- ▶ **II.** $H_1 : p < p_0$:

$$H_0 \text{ is accepted} \Leftrightarrow z > z_\alpha \Leftrightarrow p_0 < \bar{p} - \sqrt{\frac{p_0(1-p_0)}{n}} \cdot z_\alpha$$

i.e. p_0 belongs to the one-sided Confidence Interval (CI).

- ▶ **III.** $H_1 : p > p_0$:

$$H_0 \text{ is accepted} \Leftrightarrow z < z_{1-\alpha} \Leftrightarrow \bar{p} - \sqrt{\frac{p_0(1-p_0)}{n}} \cdot z_{1-\alpha} < p_0$$

i.e. p_0 belongs to the one-sided confidence interval (CI).

Test for Proportions: Example (Part a)

We have a significance level $\alpha = 0.05$, so for the rejection region we need the quantile

$$z_{1-\alpha} = z_{0.95} = 1.6449$$

Thus, the rejection region is $RR = [1.645, \infty)$, and the test statistic is

$$z = \frac{0.12 - 0.1}{\sqrt{\frac{0.1 \cdot 0.9}{200}}} = 0.943.$$

Since $z \notin RR$, we do not reject H_0 at this significance level.

Test for Proportions: Example (Part b)

We are now considering the following two hypotheses:

$$H_0 : p = p_0, \text{ where } p_0 = 0.1,$$

$$H_1 : p = p_1, \text{ where } p_1 = 0.15.$$

From part a), we found the rejection region:

$$z \geq z_{0.95}.$$

This gives us the condition:

$$\frac{\bar{x} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_{0.95}.$$

Rewriting this inequality, we find that H_0 is rejected if:

$$\bar{x} \geq p_0 + z_{0.95} \sqrt{\frac{p_0(1-p_0)}{n}} = 0.1349.$$

So, we reject H_0 if $\bar{x} \geq 0.1349$, and do not reject H_0 if:

$$\bar{x} < 0.1349.$$

Test for Proportions: Example (Part c)

Assuming the true value is $p = p_0$, under hypothesis H_0 , the test statistic is

$$Z_0 := \frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0, 1).$$

We now compute the probability of a Type II error $\beta(p_1)$, which is the probability of not rejecting H_0 when H_1 is true:

$$\begin{aligned}\beta(p_1) &= \mathbb{P}(\text{not reject } H_0 \mid H_1) = \mathbb{P}(\bar{X} < 0.1349 \mid p = p_1) \\ &= \mathbb{P}\left(\frac{\bar{X} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} < \frac{0.1349 - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \mid p = p_1\right) \\ &= \mathbb{P}(Z_1 < -0.598 \mid p = p_1) \quad (\text{since } Z_1 \sim N(0, 1)) \\ &= \mathbb{P}(Z_1 < -0.598) = 0.2749.\end{aligned}$$

This is a very large value! Such a probability of error is not acceptable.

Test for Proportions: Example (Part d, Part 1)

To make the probability of a Type II error acceptable, we need to adjust the sample size n so that $\beta = 0.05$.

The idea is to solve for n by going backwards from the calculations in part b). Specifically, we set a cutoff value for \bar{p} (previously 0.1349) as k , and we find n from the condition:

$$\mathbb{P}(\text{reject } H_0 \mid H_0) = \mathbb{P}(\bar{p} \geq k \mid p = p_0).$$

This becomes

$$\mathbb{P}\left(\frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq \frac{k - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \middle| p = p_0\right).$$

Since $Z_0 \sim N(0, 1)$, we have

$$\mathbb{P}\left(Z_0 \geq z_{1-\alpha} \middle| p = p_0\right) = 1 - F_{N(0,1)}(z_{1-\alpha}) = \alpha = 0.05.$$

Test for Proportions: Example (Part d, Part 2)

We now have

$$\frac{k - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = z_{1-\alpha}.$$

Solving for k , we get

$$k = p_0 + z_{1-\alpha} \sqrt{\frac{p_0(1-p_0)}{n}}.$$

This is the condition that relates k , p_0 , and the sample size n for ensuring that the probability of a Type II error is $\beta = 0.05$.

Type II Error Calculation

On the other hand, as we did in part b), we have

$$\begin{aligned}\mathbb{P}(\text{not reject } H_0 \mid H_1) &= \mathbb{P}(\bar{p} < k \mid p = p_1) \\&= \mathbb{P}\left(\frac{\bar{p} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} < \frac{k - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \middle| p = p_1\right) \\&= \mathbb{P}\left(Z_1 < \frac{k - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \middle| p = p_1\right) \\&= F_{N(0,1)}(z_\beta) = \beta = 0.05 \\&\quad (\text{by the definition of a quantile}).\end{aligned}$$

Solving for Cutoff Value

From the above, we have

$$\frac{k - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} = z_\beta \implies k = p_1 + z_\beta \sqrt{\frac{p_1(1-p_1)}{n}}.$$

Now, from the two equations:

$$p_0 + z_{1-\alpha} \sqrt{\frac{p_0(1-p_0)}{n}} = p_1 + z_\beta \sqrt{\frac{p_1(1-p_1)}{n}}$$
$$z_{0.95} \sqrt{\frac{p_0(1-p_0)}{n}} - z_{0.05} \sqrt{\frac{p_1(1-p_1)}{n}} = p_1 - p_0$$
$$\frac{1.0808}{\sqrt{n}} = 0.05.$$

Final Calculation

Solving for n :

$$\implies n = \left(\frac{1.0809}{0.05} \right)^2 = 467.34.$$

Thus, a sample of size at least 468 items should be selected in order to ensure that $\alpha = \beta = 0.05$.

A Problem in Hypothesis Testing

A potential issue that can arise in hypothesis testing is as follows:

- ▶ We start by setting a significance level α , the probability of committing a type I error, and use this to define the rejection region (RR).
- ▶ When we compute the test statistic, its observed value does not fall within the rejection region, so we fail to reject the null hypothesis H_0 . In other words, we accept H_0 as plausible.

However:

- ▶ When we calculate the probability of obtaining this value of the test statistic under the assumption that H_0 is true, we find this probability to be quite small—comparable to our preset α .
- ▶ This creates a paradox: we accept H_0 , yet under the assumption that it is true, the observed value of the test statistic appears very unlikely.

Example of a Problem in Hypothesis Testing

Consider the following scenario:

- ▶ We test $H_0 : \mu = 10$ vs. $H_1 : \mu \neq 10$ at $\alpha = 0.05$ (two-tailed test).
- ▶ The rejection region (RR) is dictated by $|z| > 1.96$, based on the standard normal distribution, that is

$$RR = \{z \in \mathbb{R} : |z| \geq z_{1-\frac{0.05}{2}} = 1.96\} = (-\infty, -1.96] \cup [1.96, +\infty).$$

Observed Test Statistic: Assume $n = 25$, $\bar{x} = 10.76$, $\sigma = 2$. Then

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = 1.9 \quad (\text{just outside the rejection region}).$$

However:

- ▶ The probability of observing $z = 1.9$ under H_0 is ([exercise!](#)):
$$\mathbb{P}(|Z| > 1.9) = 0.058 \quad (\text{close to } \alpha = 0.05).$$
- ▶ Although $z \notin RR$, the probability is very small, making the observed value surprising under H_0 .

Paradox:

- ▶ We fail to reject H_0 , yet the observed test statistic seems unlikely if H_0 were true.

A Problem in Hypothesis Testing

This raises concerns:

- ▶ Did we set our rejection region appropriately?
- ▶ Were we too quick to “accept” H_0 by dismissing values of the test statistic that, although outside the RR, are still surprising enough to warrant further consideration?

Significance Testing, p -Values

- ▶ We should take a look at how "far-fetched" does the value of the test statistic seem, under the assumption that H_0 is true. If it seems really implausible to occur by chance i.e. if its probability is small, then maybe we should reject the null hypothesis H_0 .
- ▶ To avoid this situation, we perform what is called a **significance test**: for a given random sample X_1, \dots, X_n , we still set up H_0 and H_1 as before and we choose an **appropriate** test statistic.
- ▶ Then, we compute the probability of observing a value at least as extreme (in the sense of the test conducted) of the test statistic TS as the value observed from the sample, ts_0 , under the assumption that H_0 is true.

Significance Testing, p -Values

- ▶ This probability is called the **critical value/descriptive significance level/probability of the test** p -value of the test. If it is small, we reject H_0 , otherwise we do not reject it.
- ▶ The p -value is a numerical value assigned to the test, it depends only on the sample data and its distribution, but not on α .
- ▶ In general, for the three possible alternatives, if $ts_0 = TS(x_1, \dots, x_n; \theta_0)$ is the value of the test statistic $TS(X_1, \dots, X_n; \theta)$ under the assumption that $H_0 : \theta = \theta_0$ is true and F is the cdf of TS , the P -value is computed by:

$H_1 : \theta \neq \theta_0$ (two-tailed test)

$$\begin{aligned} p &= 2 \cdot \min \{ \mathbb{P}(TS \leq ts_0 | H_0), \mathbb{P}(TS \geq ts_0 | H_0) \} \\ &= 2 \cdot \min \{ F(ts_0), 1 - F(ts_0) \} \end{aligned}$$

$H_1 : \theta < \theta_0$ (left-tailed test) $p = \mathbb{P}(TS \leq ts_0 | H_0) = F(ts_0)$

$H_1 : \theta > \theta_0$ (right-tailed test) $p = \mathbb{P}(TS \geq ts_0 | H_0) = 1 - F(ts_0).$

Significance Testing, p -Values

Then the decision will be:

if $p \leq \alpha$, reject H_0

if $p > \alpha$, do not reject H_0 .

- ▶ The p -value of a test is the smallest level at which we could have preset α and still have been able to reject H_0 , or the lowest significance level that forces rejection of H_0 , i.e. the minimum rejection level.

What is a p-value?

A **p-value** (probability value) is a measure used in hypothesis testing to assess the strength of the evidence against the null hypothesis (H_0) based on the sample data. It quantifies the probability of observing a test statistic at least as extreme as the one calculated from your sample, assuming that the null hypothesis is true.

- ▶ **Definition:** The p-value is the probability of obtaining results at least as extreme as the results actually observed, given that the null hypothesis is true.
- ▶ **Interpretation:**
 - ▶ **Small p-value** ($\leq \alpha$): Suggests strong evidence against the null hypothesis. This leads to **rejecting the null hypothesis** (H_0).
 - ▶ **Large p-value** ($> \alpha$): Suggests weak evidence against the null hypothesis. This leads to **failing to reject** the null hypothesis.

Example:

Suppose we are testing the hypothesis that a coin is fair (i.e., the probability of heads is 0.5).

- ▶ **Null hypothesis (H_0):** The coin is fair, $p = 0.5$.
- ▶ **Alternative hypothesis (H_1):** The coin is biased, $p \neq 0.5$.

After flipping the coin 100 times, we observe 60 heads. We calculate a test statistic (like a z-score or t-score), and we obtain a p-value of 0.03.

- ▶ **If the significance level (α) is 0.05:** Since $0.03 < 0.05$, the p-value is less than the threshold. This suggests strong evidence against the null hypothesis, and we would **reject** H_0 , concluding that the coin may not be fair.
- ▶ **If the p-value had been 0.08:** Since $0.08 > 0.05$, we would **fail to reject** the null hypothesis, suggesting that there is not enough evidence to conclude that the coin is biased.

Formula for p-value:

In hypothesis testing, the p-value is calculated based on the chosen test (such as a z-test or t-test) and the observed data. For example, in a **two-tailed z-test**, the p-value is:

$$\text{p-value} = 2 \times \mathbb{P}(Z > |z_{\text{observed}}|)$$

where z_{observed} is the test statistic (z-score) calculated from the sample data, and $\mathbb{P}(Z > |z_{\text{observed}}|)$ is the probability of getting a test statistic at least as extreme as the observed one.

Summary:

- ▶ A **p-value** helps determine whether to reject or fail to reject the null hypothesis.
- ▶ It provides evidence for or against the null hypothesis in the context of your data.
- ▶ The smaller the p-value, the stronger the evidence against H_0 .

Example

Remember a previous example: The number of monthly sales at a firm is known to have a mean of 20 and a standard deviation of 4, and all salary, tax, and bonus figures are based on these values. However, in times of economic recession, a sales manager fears that his employees do not average 20 sales per month, but less, which could seriously hurt the company. For a number of 36 randomly selected salespeople, it was found that in one month they averaged 19 sales. At the 5% significance level, does the data confirm or contradict the manager's suspicion? Let us now perform a significance test.

Solution

We tested a left-tailed alternative for the mean:

$$\begin{aligned} H_0 : \mu &= 20 \\ H_1 : \mu &< 20. \end{aligned}$$

The population standard deviation was given, $\sigma = 4$, and for a sample of size $n = 36$, the sample mean was $\bar{x} = 19$. For the test statistic:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

the observed value was:

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{19 - 20}{\frac{4}{6}} = -1.5$$

Solution

Now, we compute the P -value (case II.):

$$P = \mathbb{P}(Z \leq z) = \mathbb{P}(Z \leq -1.5) = F_{N(0,1)}(-1.5) = \text{normcdf}(-1.5) = 0.0668$$

Since:

$$\alpha = 0.05 < 0.0668 = P,$$

(which is below the minimum rejection level), we do not reject H_0 .
Thus, at the 5% significance level, we conclude that the data contradicts the manager's suspicion.

Two Methods for Hypothesis Testing

Given a hypothesis test:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0 \quad (\text{Two-tailed Test}).$$

At a significance level $\alpha = 0.05$, we can use two equivalent methods:

► **Critical Value Approach:**

- Compute the test statistic (e.g., z).
- Compare it to the critical value $z_{\alpha/2}$ (e.g., ± 1.96).

► **p-Value Approach:**

- Compute the test statistic.
- Find the p-value, which is the probability of observing the test statistic (or more extreme values) under H_0 .
- Compare the p-value to α .

Conclusion: Both methods lead to the same decision about H_0 .

Critical Value vs. p-Value Approach

Critical Value Approach:

- ▶ Define the rejection region
 $|z| > z_{\alpha/2}$.
- ▶ Compute the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

- ▶ Compare $|z|$ to the critical value:

$$|z| > 1.96 \Rightarrow \text{Reject } H_0.$$

p-Value Approach:

- ▶ Compute the test statistic z .
- ▶ Find the p-value:

$$p = 2 \cdot \mathbb{P}(Z > |z|).$$

- ▶ Compare the p-value to α :

$$p < \alpha \Rightarrow \text{Reject } H_0.$$

Equivalence: Both approaches result in the same decision regarding H_0 .

Example: Critical Value and p-Value Equivalence

Problem: Test $H_0 : \mu = 50$ vs. $H_1 : \mu \neq 50$ at $\alpha = 0.05$.

- ▶ Sample data: $n = 36, \bar{x} = 48.7, \sigma = 10$.
- ▶ Compute the test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{48.7 - 50}{10/\sqrt{36}} = -1.9.$$

Critical Value Approach:

$$z_{\alpha/2} = 1.96, \quad |z| = 1.9 < 1.96 \quad \Rightarrow \text{Fail to reject } H_0.$$

p-Value Approach:

$$p = 2 \cdot \mathbb{P}(Z > 1.9) \approx 2 \cdot 0.0287 = 0.0574.$$

Since $p = 0.0574 > 0.05$, we fail to reject H_0 .

Conclusion: Both methods lead to the same decision: Fail to reject H_0 .

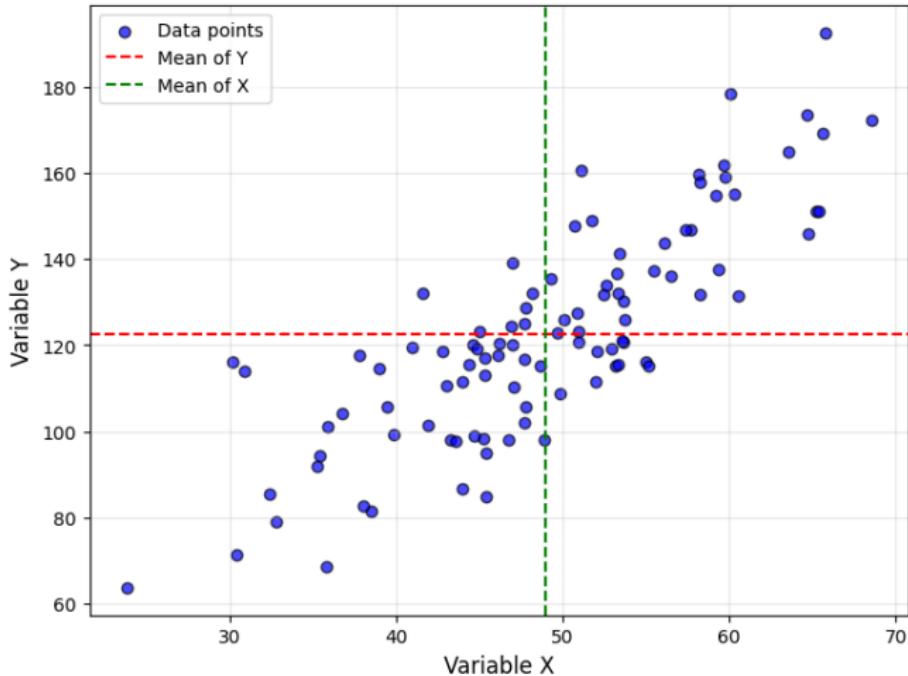
LECTURES 11 and 12

Regression

Correlation and Regression

- ▶ We have focused on various descriptive techniques for describing a single variable. However, a crucial aspect of statistics involves analysing the association between two or more variables, determining whether they are independent, and, if not, *understanding the nature of their dependence*. One of the core concepts in statistical analysis is **correlation**.
- ▶ **Correlation** typically quantifies the relationship between a dependent variable, known as the *response variable*, and one or more independent variables, referred to as *predictor variables* (or simply *predictors*). When two variables are correlated, it implies that knowing the value of one variable can help predict the value of the other.
 - ▶ **Regression** is the statistical method used to establish and quantify this relationship.

Scatter Plot of Two Variables



A scatter plot helps us visualise the relationship between two variables. Each point represents an observation. Here, x is the independent variable, and y is the dependent variable. x is generated from a normal distribution, y is linearly dependent on x with some added random noise for variability. The blue points represent the data, showing how y changes with x : as x increases, y tends to increase.

Correlation, Curves of Regression

We will restrict our discussion to the case of two characteristics, X and Y .

- ▶ We can get a first idea of the relationship between X and Y by plotting them in a **scatterplot or scattergram / scatter diagram** (see the picture above), which is a plot of the points with coordinates:

$$(x_i, y_i)_{i=\overline{1,k}}, x_i \in X, y_i \in Y, i = \overline{1, k}.$$

- ▶ We group the N data into $m \times n$ classes and denote:
 - ▶ class mark: (x_i, y_j)
 - ▶ absolute frequency of the class (i, j) : f_{ij}for $i = \overline{1, m}, j = \overline{1, n}$.

The two-dimensional characteristic (X, Y) is represented in a **correlation table (contingency table)**:

Correlation Table

$X \setminus Y$	y_1	\dots	y_j	\dots	y_n	
x_1	f_{11}	\dots	f_{1j}	\dots	f_{1n}	$f_{1\cdot}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_i	f_{i1}	\dots	f_{ij}	\dots	f_{in}	$f_{i\cdot}$
\vdots	\vdots		\vdots		\vdots	\vdots
x_m	f_{m1}	\dots	f_{mj}	\dots	f_{mn}	$f_{m\cdot}$
	$f_{\cdot 1}$	\dots	$f_{\cdot j}$	\dots	$f_{\cdot n}$	$f_{\cdot \cdot} = N$

Notice that:

$$\sum_{j=1}^n f_{ij} = f_{i\cdot}, \quad \sum_{i=1}^m f_{ij} = f_{\cdot j}, \quad \sum_{i=1}^m f_{i\cdot} = \sum_{j=1}^n f_{\cdot j} = f_{\cdot \cdot} = N.$$

Now we can define numerical characteristics associated with (X, Y) .

Moments of a Two-Dimensional Characteristic

Definition 26

Let (X, Y) be a two-dimensional characteristic/variable whose distribution is given by the previous correlation table and let $k_1, k_2 \in \mathbb{N}$.

(1) The **initial moment of order** (k_1, k_2) is the value

$$\bar{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^n f_{ij} x_i^{k_1} y_j^{k_2}.$$

(2) The **central moment of order** (k_1, k_2) is the value

$$\bar{\mu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^n f_{ij} (x_i - \bar{x})^{k_1} (y_j - \bar{y})^{k_2}$$

where $\bar{x} = \bar{\nu}_{10} = \frac{1}{N} \sum_{i=1}^m f_{i\cdot} x_i$ and $\bar{y} = \bar{\nu}_{01} = \frac{1}{N} \sum_{j=1}^n f_{\cdot j} y_j$ are the means of X and Y , respectively.

Remark: The means and variances of X and Y can be expressed as moments of (X, Y) :

$$\bar{\sigma}_X^2 := \bar{\mu}_{20} = \bar{\nu}_{20} - \bar{x}^2$$

$$\bar{\sigma}_Y^2 := \bar{\mu}_{02} = \bar{\nu}_{02} - \bar{y}^2$$

Covariance and Correlation Coefficient for 2-dim Characteristics

Definition 27

Let (X, Y) be a two-dimensional characteristic whose distribution is given by the previous correlation table.

(1) The covariance of (X, Y) is the value

$$\bar{\mu}_{11} = \frac{1}{N} \sum_{i=1}^m \sum_{j=1}^n f_{ij} (x_i - \bar{x}) (y_j - \bar{y}).$$

(2) The correlation coefficient of (X, Y) is the value

$$\bar{\rho} = \bar{\rho}_{XY} = \frac{\bar{\mu}_{11}}{\bar{\sigma}_X \bar{\sigma}_Y}.$$

Connection to Probability Theory

- ▶ The covariance

$$\text{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))).$$

- ▶ The correlation coefficient

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{V(X)V(Y)}}.$$

- ▶ These notions from Probability Theory are defined similarly for sample data and have the same properties. The covariance gives a rough idea of the relationship between X and Y .
- ▶ If X and Y are independent (so there is no relationship, no correlation between them), then the covariance is 0.
- ▶ An easier computational formula for the covariance is:

$$\bar{\mu}_{11} = \bar{\nu}_{11} - \bar{x} \cdot \bar{y}$$

Interpretation of Correlation Coefficient

- ▶ The correlation coefficient is then:

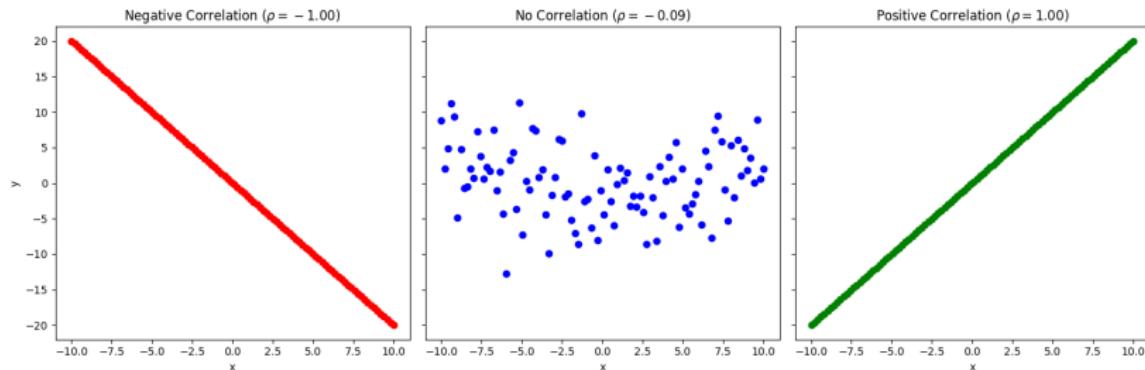
$$\bar{\rho} = \frac{\bar{v}_{11} - \bar{x} \cdot \bar{y}}{\bar{\sigma}_X \bar{\sigma}_Y}$$

- ▶ It satisfies the inequality:

$$-1 \leq \bar{\rho} \leq 1$$

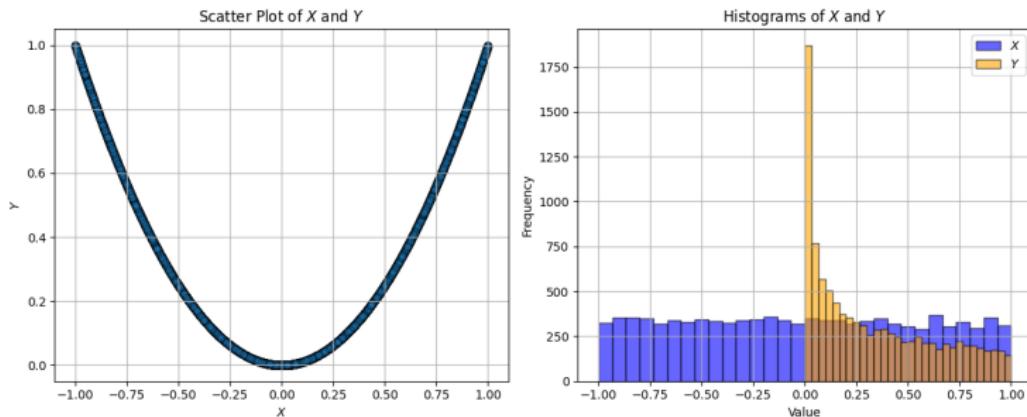
- ▶ By its variation between -1 and 1, the correlation coefficient measures the linear relationship between X and Y .
- ▶ Key Cases:
 - ▶ If $\rho = 1$, there is a perfect positive correlation between X and Y .
 - ▶ If $\rho = -1$, there is a perfect negative correlation between X and Y .
 - ▶ If $\rho = 0$, there is no linear correlation, but X and Y may still be related in a non-linear way.

Interpretation of Correlation Coefficient



Interpretation of Correlation Coefficient

A theoretical example: $X \sim \text{Unif}[-1, 1]$, $Y = X^2$, $\mathbb{E}(X) = 0$, $\mathbb{E}(XY) = \mathbb{E}(X^3) = 0 \Rightarrow \text{cov}(X, Y) = 0 \Rightarrow X$ and Y are uncorrelated, but not independent.



Conditional Mean

- ▶ In our task of finding a relationship between X and Y , we may take the following approach: knowing the value of one of the characteristics/variables, try to find a probable or "expected" value for the other. If the two characteristics are related, there should be a pattern, i.e. the expected value of one characteristic, conditioned by the other, should be a function of the value that the other variable assumes.
- ▶ In other words, we should consider **conditional means**. These are defined similarly to regular means, but taking into account the condition we need to impose.

Conditional Mean

Definition 28

Let (X, Y) be a two-dimensional characteristic whose distribution is given by the previous correlation table.

(1) **Conditional Mean of Y given $X = x_i$** is the value

$$\bar{y}_i = \bar{y}(x_i) = \frac{1}{f_i} \sum_{j=1}^n f_{ij} y_j, \quad i = \overline{1, m}$$

(2) **Conditional Mean of X given $Y = y_j$** is the value

$$\bar{x}_j = \bar{x}(y_j) = \frac{1}{f_{\cdot j}} \sum_{i=1}^m f_{ij} x_i, \quad j = \overline{1, n}$$

Conditional Means: Example

Scenario: Suppose X represents the grade level of a student ($x_1 = \text{Grade 1}, x_2 = \text{Grade 2}$) and Y represents the number of books read in a month ($y_1 = 0 - 5 \text{ books}, y_2 = 6 - 10 \text{ books}$). The joint frequency distribution is:

$X \setminus Y$	$y_1 = 0 - 5$	$y_2 = 6 - 10$	Row Total $f_{i\cdot}$
$x_1 = \text{Grade 1}$	20	10	30
$x_2 = \text{Grade 2}$	15	25	40
Column Total $f_{\cdot j}$	35	35	70

(1) Conditional Mean of Y given $X = x_i$:

$$\bar{y}(x_1) = \frac{1}{30} (20 \cdot 2.5 + 10 \cdot 8) = 4.83, \quad \bar{y}(x_2) = \frac{1}{40} (15 \cdot 2.5 + 25 \cdot 8) = 6.75$$

(2) Conditional Mean of X given $Y = y_j$:

$$\bar{x}(y_1) = \frac{1}{35} (20 \cdot 1 + 15 \cdot 2) = 1.43, \quad \bar{x}(y_2) = \frac{1}{35} (10 \cdot 1 + 25 \cdot 2) = 1.71$$

Regression Curves

Definition 29

Let (X, Y) be a two-dimensional characteristic.

(1) **The curve** $y = f(x)$ formed by the points $(x_i, \bar{y}_i), i = \overline{1, m}$ is called *curve of regression of Y on X*.

(2) **The curve** $x = g(y)$ formed by the points $(y_j, \bar{x}_j), j = \overline{1, n}$ is called *curve of regression of X on Y*.

Example: Data and Regression Curves

Data: Suppose we observe the following pairs of (X, Y) values:

$$(1, 2), (2, 4), (3, 6), (4, 8), (5, 10)$$

Regression of Y on X :

$$y = 2x \quad (\text{a perfect linear relationship})$$

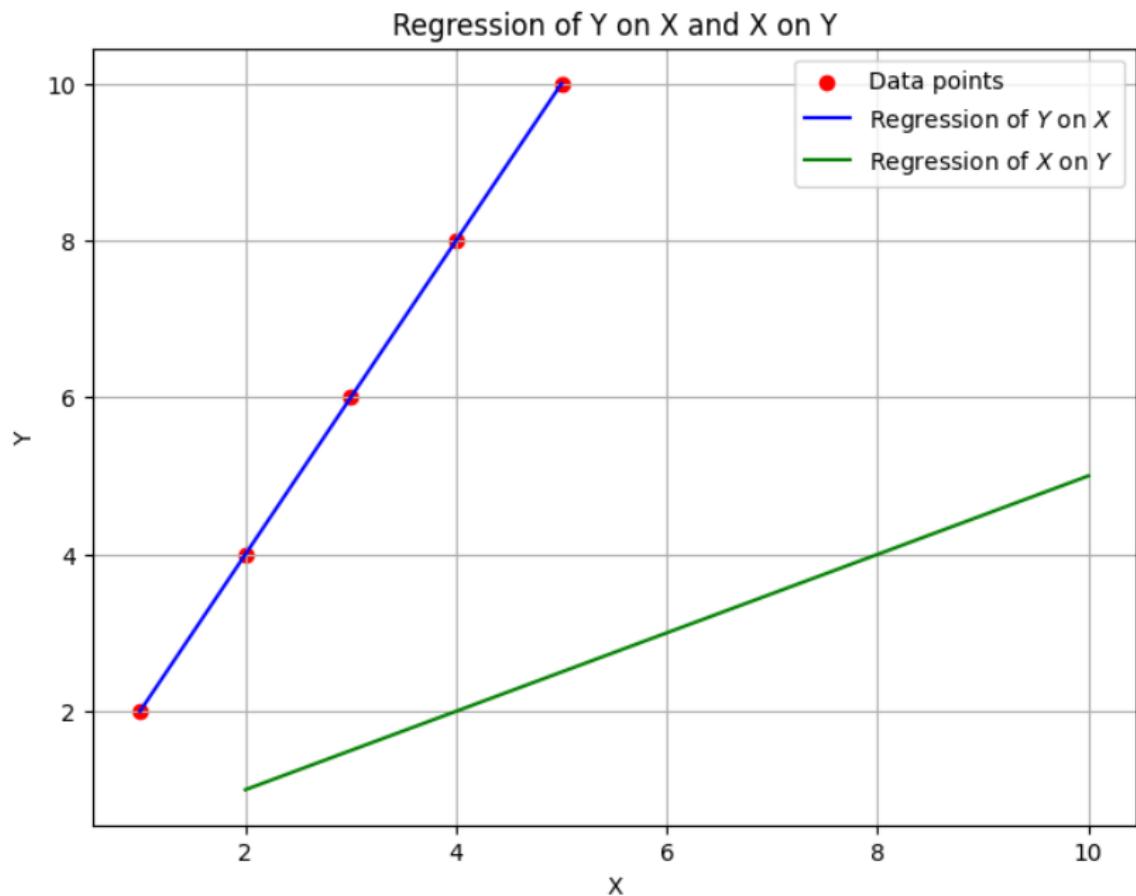
Regression of X on Y :

$$x = \frac{y}{2}$$

We calculate:

- ▶ **Curve of Regression of Y on X :** Formed by the points (x_i, \bar{y}_i) for $i = 1, \dots, m$.
- ▶ **Curve of Regression of X on Y :** Formed by the points (y_j, \bar{x}_j) for $j = 1, \dots, n$.

Example: Data and Regression Curves

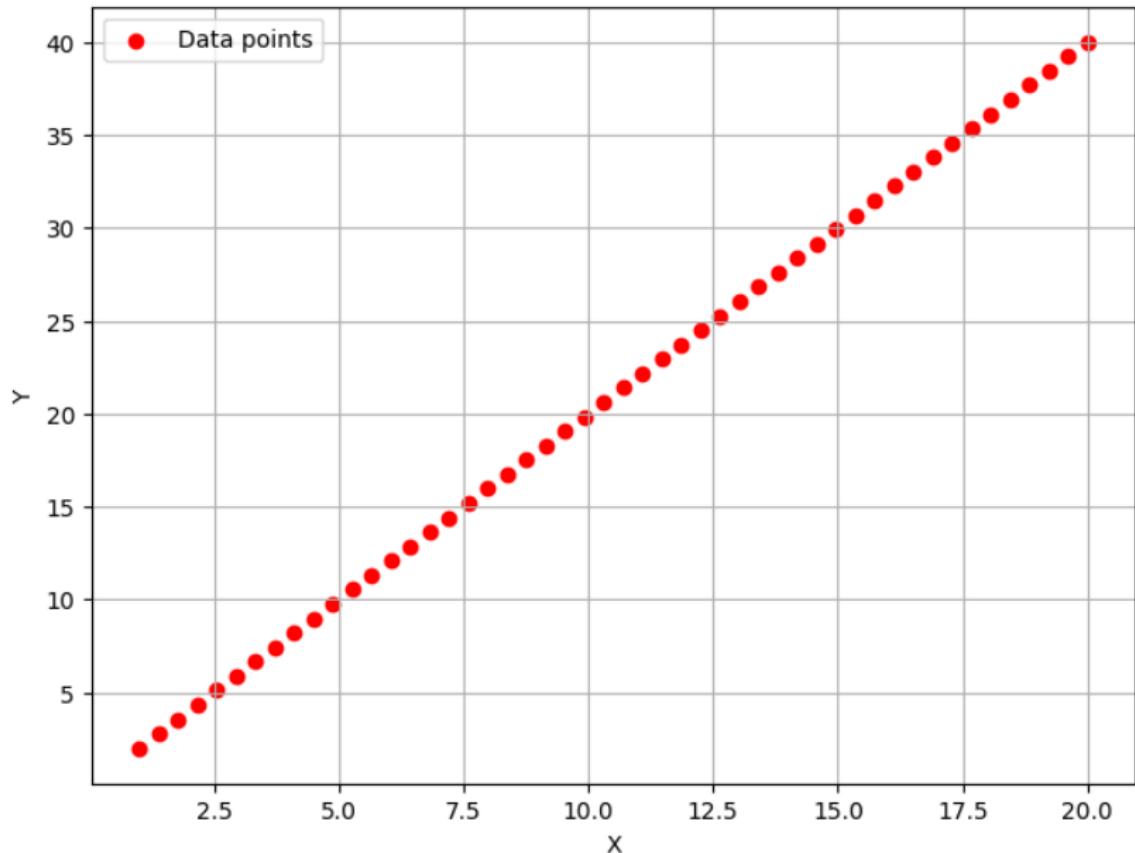


Remark: Regression Curves

The curve of regression of a characteristic Y with respect to another characteristic X represents the mean value of Y , denoted $\bar{y}(x)$, given a specific value $X = x$.

The regression curve is determined to best approximate the scatterplot of (X, Y) .

Scatter Plot of More Data Points (X, Y)



Least Squares Estimation, Linear Regression

One of the most popular ways of finding curves of regression is the **least squares method**.

Assume the curve of regression of Y on X is of the form:

$$y = y(x) = f(x; a_1, \dots, a_s)$$

We determine the unknown parameters a_1, \dots, a_s so that the **sum of squares error (SSE)**, which is the sum of the squares of the differences between the observed responses y_j and their fitted values $y(x_i)$, weighted by their corresponding frequencies, is **minimized**.

The SSE is:

$$SSE = \sum_{i=1}^m \sum_{j=1}^n f_{ij} (y_j - f(x_i; a_1, \dots, a_s))^2.$$

Minimizing SSE and Solving for Parameters

We find the point of minimum $(\bar{a}_1, \dots, \bar{a}_s)$ of *SSE* by solving the system:

$$\frac{\partial SSE}{\partial a_k} = 0, \quad k = \overline{1, s}.$$

This leads to the following equations:

$$-2 \sum_{i=1}^m \sum_{j=1}^n f_{ij} (y_j - f(x_i; a_1, \dots, a_s)) \frac{\partial f(x_i; a_1, \dots, a_s)}{\partial a_k} = 0$$

for every $k = \overline{1, s}$.

- ▶ Regression equation: The equation of the curve of regression of Y on X is then

$$y = f(x; \bar{a}_1, \dots, \bar{a}_s).$$

Linear Regression

Let us consider the case of linear regression and find the equation of the line of regression of Y on X .

We are finding a curve $y = ax + b$, for which the sum of squared errors

$$S(a, b) = \sum_{i=1}^m \sum_{j=1}^n f_{ij} (y_j - ax_i - b)^2$$

is minimized.

Solving the System for Linear Regression

We need to solve the 2×2 system:

$$\begin{cases} \frac{\partial S(a,b)}{\partial a} = 0 \\ \frac{\partial S(a,b)}{\partial b} = 0 \end{cases} \implies \begin{cases} -2 \sum_{i=1}^m \sum_{j=1}^n f_{ij} (y_j - ax_i - b) x_i = 0 \\ -2 \sum_{i=1}^m \sum_{j=1}^n f_{ij} (y_j - ax_i - b) = 0 \end{cases}$$

This becomes:

$$\begin{cases} \left(\sum_{i=1}^m \sum_{j=1}^n f_{ij} x_i^2 \right) a + \left(\sum_{i=1}^m \sum_{j=1}^n f_{ij} x_i \right) b = \sum_{i=1}^m \sum_{j=1}^n f_{ij} x_i y_j \\ \left(\sum_{i=1}^m \sum_{j=1}^n f_{ij} x_i \right) a + \left(\sum_{i=1}^m \sum_{j=1}^n f_{ij} \right) b = \sum_{i=1}^m \sum_{j=1}^n f_{ij} y_j. \end{cases}$$

Solution of System of Equations

After dividing both equations by N , we get the following system:

$$\begin{cases} \bar{\nu}_{20}a + \bar{x}b = \bar{\nu}_{11} \\ \bar{x}a + \bar{\nu}_{00}b = \bar{y} \end{cases}$$

The solution to the system is:

$$\begin{aligned}\bar{a} &= \frac{\bar{\nu}_{11} - \bar{x} \cdot \bar{y}}{\bar{\nu}_{20} - \bar{x}^2} = \frac{\bar{\nu}_{11} - \bar{x} \cdot \bar{y}}{\bar{\sigma}_X \bar{\sigma}_Y} \cdot \frac{\bar{\sigma}_Y}{\bar{\sigma}_X} = \bar{\rho} \frac{\bar{\sigma}_Y}{\bar{\sigma}_X} \\ \bar{b} &= \bar{\nu}_{01} - \bar{\nu}_{10}\bar{a} = \bar{y} - \bar{x} \cdot \bar{a}\end{aligned}$$

Regression Equations

The equation of the line of regression of Y on X is:

$$y - \bar{y} = \bar{\rho} \frac{\bar{\sigma}_Y}{\bar{\sigma}_X} (x - \bar{x}).$$

By analogy, the equation of the line of regression of X on Y is:

$$x - \bar{x} = \bar{\rho} \frac{\bar{\sigma}_X}{\bar{\sigma}_Y} (y - \bar{y}).$$

Remark:

1. The point of intersection of the two lines of regression, (\bar{x}, \bar{y}) , is called the **centroid of the distribution of (X, Y)** .
2. For the angle α between the two lines of regression, we have:

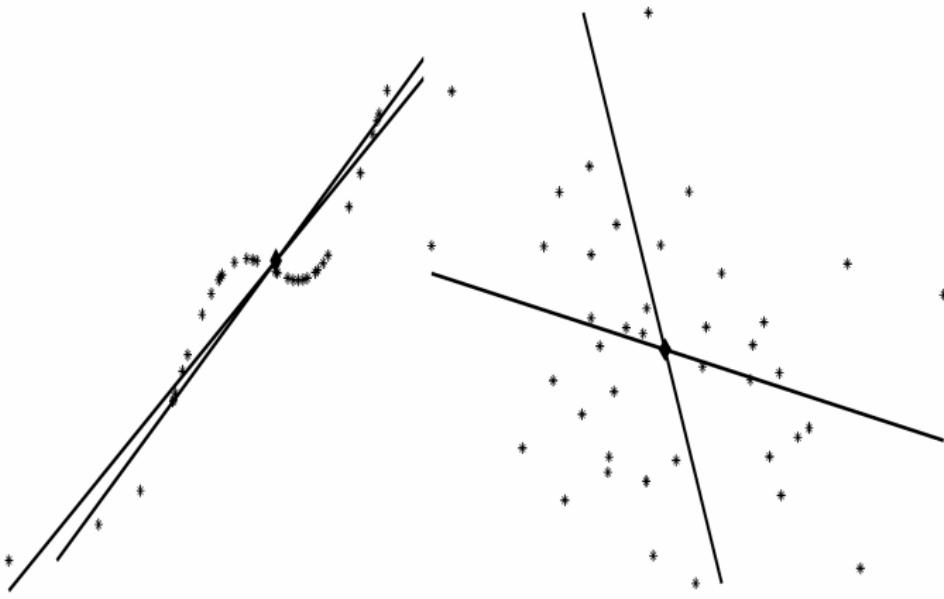
$$\tan \alpha = \frac{1 - \bar{\rho}^2}{\bar{\rho}^2} \cdot \frac{\bar{\sigma}_X \bar{\sigma}_Y}{\bar{\sigma}_X^2 + \bar{\sigma}_Y^2}$$

If $|\bar{\rho}| = 1$, the lines coincide ($\alpha = 0$). If $|\bar{\rho}| = 0$, the lines are perpendicular.

Example

In Figure (a) below, we have $\bar{\rho} = 0.95$, which is positive and very close to 1, suggesting a strong positive linear trend.

- ▶ Most of the points are on or very close to the line of regression of Y on X .
- ▶ The positivity indicates that large values of X are associated with large values of Y .
- ▶ Since the correlation coefficient is so close to 1, the two lines of regression almost coincide.



(a) $\bar{\rho} = 0.95$

(b) $\bar{\rho} = -0.28$

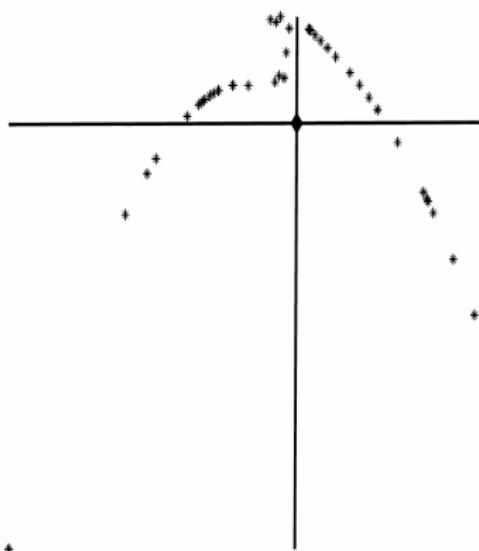
Example

In Figure (b), we have $\bar{\rho} = -0.28$, which is negative and fairly small, close to 0.

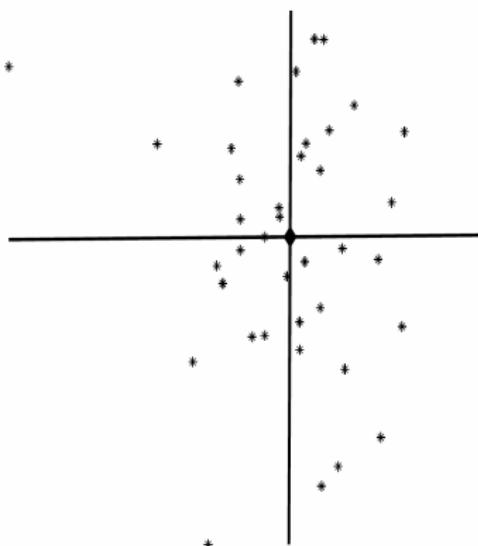
- ▶ If a relationship exists between X and Y , it does not seem to be linear.
- ▶ The points are scattered around the plane with no visible pattern.
- ▶ The two lines of regression are distinct and both have negative slopes, suggesting that large values of X are associated with small values of Y .

Example

- ▶ In Figure (c) below, $\bar{\rho} = 0$, so the two characteristics are uncorrelated and no linear relationship exists between them. However, they are not independent, as they were chosen so that $Y = -X^2 + \sin\left(\frac{1}{X}\right)$.
- ▶ Additionally, the two lines of regression are perpendicular.
- ▶ In Figure (d), $\bar{\rho} = 0$ again, so no linear relationship exists. In fact, the two characteristics are independent, as suggested by their random scatter inside the plane.



(c) $\bar{\rho} = 0$



(d) $\bar{\rho} = 0$

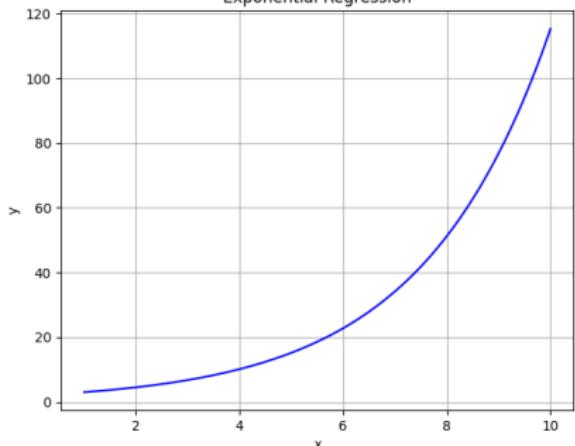
- ▶ Regression tells us the "best guess" of Y for any given X: if we know the value of X, the regression line gives us the predicted value of Y.
- ▶ A regression curve captures the overall pattern: if our data points are scattered, the regression line tries to "best fit" or represent the trend that most of the points follow. This is especially useful when data is noisy or has some variability.
- ▶ It measures the strength and direction of the relationship. The slope of the regression line tells you how strongly Y changes as X changes:
 - ▶ A positive slope means Y increases as X increases.
 - ▶ A negative slope means Y decreases as X increases.
 - ▶ If the line is flat ($\text{slope} = 0$), it indicates no relationship between X and Y.

Different Curves of Regression

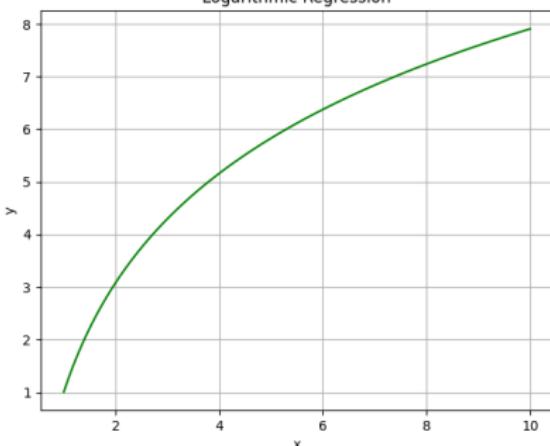
Remark: Other types of curves of regression that are fairly frequently used are:

- ▶ Exponential regression: $y = ab^x$
- ▶ Logarithmic regression: $y = a \log x + b$
- ▶ Logistic regression: $y = \frac{1}{ae^{-x}+b}$
- ▶ Hyperbolic regression: $y = \frac{a}{x} + b$

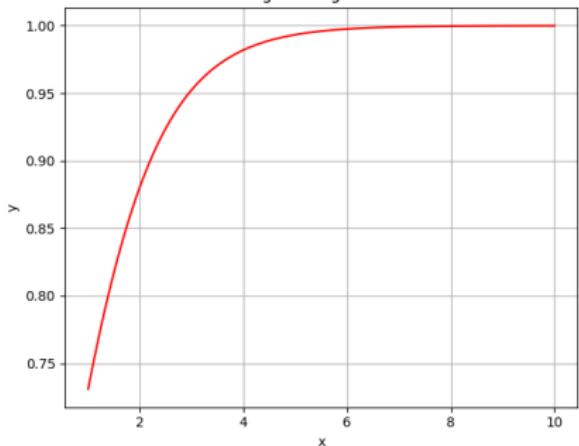
Exponential Regression



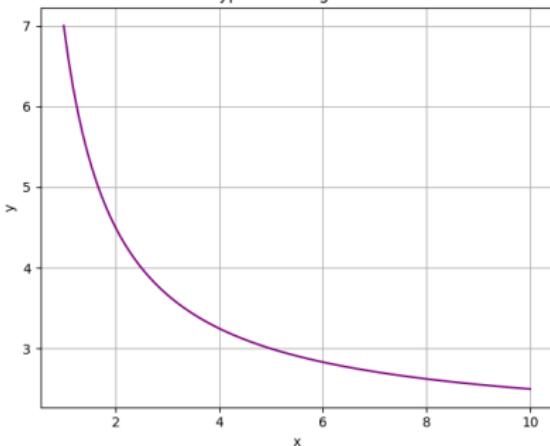
Logarithmic Regression



Logistic Regression



Hyperbolic Regression



Regression, in slightly different words

- ▶ Regression models link a response variable Y to one or more predictors.
- ▶ Using observed predictor values, we can predict Y by calculating its conditional expectation given the predictors.

Definition 30

- ▶ **Response/Dependent Variable (Y)**: *The target variable to be predicted.*
- ▶ **Predictors/Independent Variables ($X^{(1)}, \dots, X^{(k)}$)**: *Variables used to predict the response Y .*

The regression of Y on $X^{(1)}, \dots, X^{(k)}$ is the conditional expectation:

$$G\left(x^{(1)}, \dots, x^{(k)}\right) = \mathbb{E}\left[Y \mid X^{(1)} = x^{(1)}, \dots, X^{(k)} = x^{(k)}\right].$$

This function depends on $x^{(1)}, \dots, x^{(k)}$ and is estimated from data.

Conditional Expectation in Regression

For a random variable Y and predictors $X = (X^{(1)}, \dots, X^{(k)})$, the conditional expectation of Y given $X = x$ is defined as:

$$\mathbb{E}[Y | X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y | x) dy,$$

where:

- ▶ $f_{Y|X}(y | x)$ is the **conditional probability density function (pdf)** of Y given $X = x$.

For discrete variables: If Y is a discrete random variable, the conditional expectation is:

$$\mathbb{E}[Y | X = x] = \sum_y y \mathbb{P}(Y = y | X = x),$$

where $\mathbb{P}(Y = y | X = x)$ is the **conditional probability mass function (pmf)**.

In regression, this expectation is modeled as $G(x) = \mathbb{E}[Y | X = x]$, representing the dependency of Y on predictors.

Conditional pmf and pdf

For discrete random variables Y and $X = (X^{(1)}, \dots, X^{(k)})$, the **conditional probability mass function (pmf)** of Y given $X = x$ is:

$$\mathbb{P}(Y = y \mid X = x) = \frac{\mathbb{P}(Y = y, X = x)}{\mathbb{P}(X = x)},$$

where:

- ▶ $\mathbb{P}(Y = y, X = x)$ is the **joint pmf** of Y and X .
- ▶ $\mathbb{P}(X = x)$ is the **marginal pmf** of X , ensuring that $\mathbb{P}(X = x) > 0$.

For continuous random variables Y and $X = (X^{(1)}, \dots, X^{(k)})$, the **conditional probability density function (pdf)** of Y given $X = x$ is:

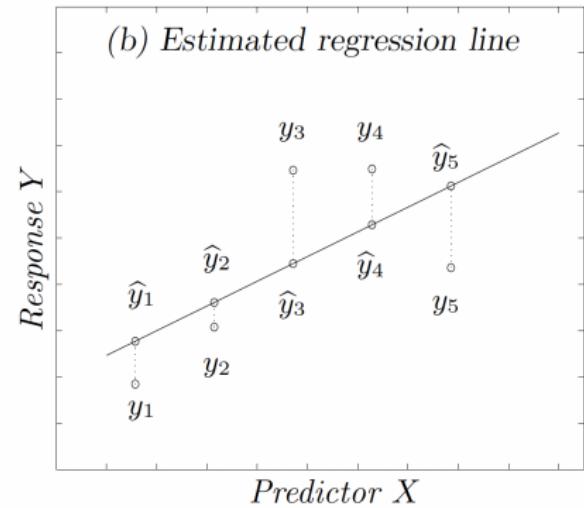
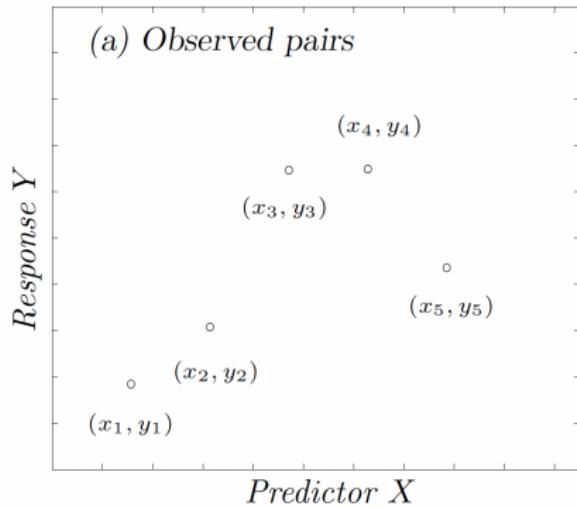
$$f_{Y|X}(y \mid x) = \frac{f_{Y,X}(y, x)}{f_X(x)},$$

where:

- ▶ $f_{Y,X}(y, x)$ is the **joint pdf** of Y and X .
- ▶ $f_X(x)$ is the **marginal pdf** of X , ensuring that $f_X(x) > 0$.

Univariate Regression

- ▶ An important goal in regression is to estimate the regression function G that relates the response variable Y to the predictors $X^{(1)}, \dots, X^{(k)}$.
- ▶ In univariate regression, we observe pairs $(x_1, y_1), \dots, (x_n, y_n)$. To make accurate predictions, we seek the function $\hat{G}(x)$ that best fits the data by minimizing the distance between the observed points y_1, \dots, y_n and the corresponding points on the fitted regression line, $\hat{y}_1 = \hat{G}(x_1), \dots, \hat{y}_n = \hat{G}(x_n)$.
- ▶ The **method of least squares** achieves this by minimizing the sum of squared distances.



Least squares estimation of the regression line [Baron]

Residuals in Regression

Definition 31

The **residuals** $e_i = y_i - \hat{y}_i$ are the differences between observed responses y_i and their fitted values $\hat{y}_i = \hat{G}(x_i)$. The method of least squares finds a regression function $\hat{G}(x)$ that minimizes the sum of squared residuals:

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

The regression function \hat{G} is typically chosen in a suitable form, such as linear, quadratic, logarithmic, etc. The simplest form is linear.

Linear Regression Model

The linear regression model assumes that the conditional expectation of the response variable Y given the predictor X is a linear function of x :

$$G(x) = \mathbb{E}[Y | X = x] = \beta_0 + \beta_1 x$$

Here, β_0 represents the *intercept* and β_1 represents the *slope* of the linear function.

The intercept $\beta_0 = G(0)$ is the value of the regression function when $x = 0$. While it may not always have a physical meaning, in some cases, like e.g. predicting the fuel efficiency of a vehicle, a non-zero intercept would indicate that the vehicle uses a certain amount of fuel even when the speed is zero (e.g., idle fuel consumption).

Slope and Interpretation

- ▶ The slope $\beta_1 = G(x + 1) - G(x)$ represents the predicted change in the response variable when the predictor x changes by 1 unit. This parameter is crucial as it tells us how much the expected value of Y will change in response to a change in X .
- ▶ For example, an increase in the quality of produced computers by Δx will increase customer satisfaction by $\beta_1 \Delta x$. A zero slope indicates no linear relationship between X and Y , meaning that Y is expected to remain constant as X changes.

Regression estimates

By the linear regression method

$$b_0 = \hat{\beta}_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \hat{\beta}_1 = S_{xy}/S_{xx}$$

where

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Sample Covariance and Correlation Coefficient

From the observed data, we estimate the covariance $\text{Cov}(X, Y)$ and the correlation coefficient ρ using the sample covariance:

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

This is an unbiased estimator of the population covariance.
The sample correlation coefficient is given by:

$$r = \frac{s_{xy}}{s_x s_y}$$

where s_x and s_y are the sample standard deviations of X and Y , respectively:

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}, \quad s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n - 1}}.$$

Regression Slope and Correlation Coefficient

- If we compare the sample covariance and correlation coefficient, we see that the estimated regression slope b_1 and the correlation coefficient r are related by the following formula:

$$b_1 = \frac{s_{xy}}{s_{xx}} = \frac{s_{xy}}{s_x^2} = r \left(\frac{s_y}{s_x} \right),$$

where s_{xx} is the sample variance of X .

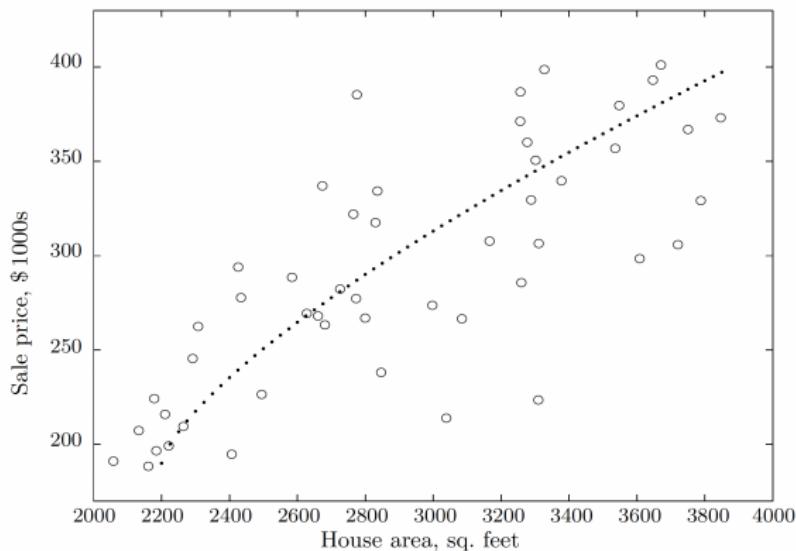
- As in the case of the correlation coefficient, the regression slope is positive for positively correlated X and Y , and negative for negatively correlated variables. However, while r is dimensionless, the slope has units of Y per unit of X . Therefore, the slope's value depends on the units and scale of X and Y and does not indicate the strength of the relationship on its own.

ANOVA and R-square

- ▶ The analysis of variance (ANOVA) explores variation among the observed responses. A part of this variation can be explained by predictors, but the rest of it is attributed to "error."
- ▶ In real-world data analysis, ANOVA helps us determine how much of the variability in data can be explained by our model, allowing us to understand the strength of the relationship between predictors and responses.

Example: Consider the sale prices of 70 houses in a certain county, along with the corresponding house areas (figure on next slide).

Example: House Prices



- ▶ A clear relationship is observed between house area and price, with larger houses generally being more expensive. However, the trend does not appear to be perfectly linear: there is considerable variability around the trend, indicating that area is not the only factor influencing house prices. Houses of the same size may still have different prices.

Analysis of Variance (ANOVA)

- ▶ In the previous figure, there is an obvious variation in house sale prices: the price depends on the house area, larger houses being in general more expensive \Rightarrow some of the price variation can be explained by the variation in house area. However, houses with the same area may still have different prices, and these differences are not explained by the area alone.
- ▶ While a regression model explains a portion of the variation in house prices, the remaining variation (error) can be attributed to factors that are not accounted for, such as location or home condition.

Definition 32

The total variation among observed responses is quantified by the total sum of squares:

$$SS_{TOT} = \sum_{i=1}^n (y_i - \bar{y})^2 = (n - 1)s_y^2$$

This represents the variation of y_i around their sample mean, independent of any regression model.

The total sum of squares represents the overall spread of the observed values, which includes both the explained variation and the unexplained variability.

Partitioning Total Variation in Regression

A portion of the total variation in the data is explained by the predictor X and the regression model that connects the predictor to the response. This portion is measured by the regression sum of squares:

$$SS_{\text{REG}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

This represents the variation explained by the model. It can also be computed as:

$$\begin{aligned} SS_{\text{REG}} &= \sum_{i=1}^n (b_0 + b_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^n (\bar{y} - b_1 \bar{x} + b_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^n b_1^2 (x_i - \bar{x})^2 \\ &= b_1^2 S_{xx} = (n - 1) b_1^2 s_x^2 \end{aligned}$$

The remaining variation is attributed to "error" and is measured by the error sum of squares:

$$SS_{\text{ERR}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$$

This portion of the variation is not explained by the model and is minimized by the least squares method.

The total variation is partitioned as follows:

$$SS_{\text{TOT}} = SS_{\text{REG}} + SS_{\text{ERR}}.$$

Definition 33

R-square, or coefficient of determination is the proportion of the total variation explained by the model,

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}}$$

It is always between 0 and 1 , with high values generally suggesting a good fit.

- ▶ A higher R-square indicates that the predictor(s) explain more of the variation in the response, which usually signals a better fit for the model.

Example of ANOVA: Study Hours and Exam Scores

Consider a study on the effect of the number of study hours (X) on exam scores (Y). We collect data from a group of students, recording the number of hours each student studies and their corresponding exam scores.

The total variation in exam scores (Y) can be explained by two components:

- ▶ The variation explained by the relationship between study hours and exam scores: SS_{REG}
- ▶ The unexplained variation due to other factors (e.g., prior knowledge, test anxiety, etc.): SS_{ERR}

We partition the total sum of squares (SS_{TOT}) into these two parts:

$$SS_{\text{TOT}} = \sum_{i=1}^n (y_i - \bar{y})^2$$

This represents the total variation in exam scores, without considering study hours.

The regression sum of squares SS_{REG} represents the portion of this variation explained by the linear model relating study hours to exam scores:

$$SS_{\text{REG}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

This shows how much the exam scores vary based on the number of hours studied.

The error sum of squares SS_{ERR} represents the remaining unexplained variation in exam scores:

$$SS_{\text{ERR}} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

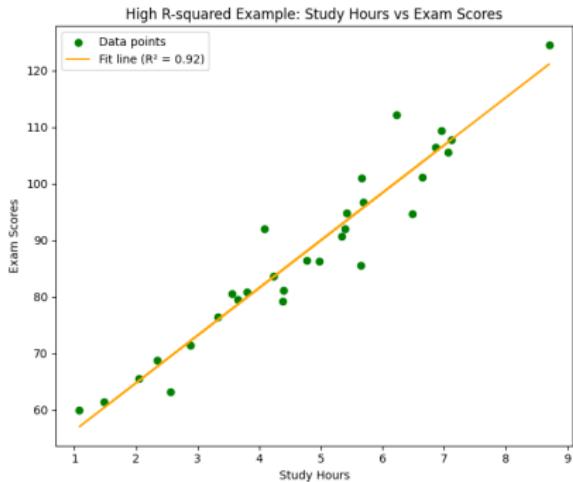
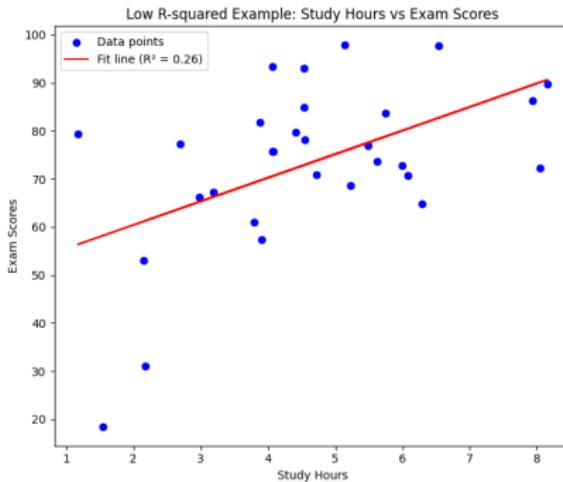
which reflects all the factors not captured by the study hours.

We can then calculate R^2 , the coefficient of determination, which tells us how much of the total variation is explained by the model:

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}}$$

A high R^2 would indicate that study hours are a good predictor of exam scores, meaning the model fits well.

- ▶ We aim to analyze how well the number of hours students study predicts their exam scores. We collect data from 30 students, recording both their study hours and corresponding exam scores. After fitting a regression line to the data, we calculate the R^2 value to assess how well study hours explain the variation in exam scores.
- ▶ In the first model, the R^2 value is 0.26, meaning that only 26% of the variation in exam scores is explained by the number of hours spent studying. The remaining of the variation is attributed to other factors, such as individual aptitude, test anxiety, or the difficulty of the exam.
- ▶ In the second model, after fitting a new regression line to the data, the R^2 value increases to 0.92, indicating that 92% of the variation in exam scores is now explained by study hours. This demonstrates a much stronger relationship between the two variables.



Impact of R-square on Model Fit

Model 1: $R^2 = 0.26$

- ▶ 26% of the variation in exam scores is explained by study hours.
- ▶ Strong scatter around the regression line.

Model 2: $R^2 = 0.92$

- ▶ 92% of the variation in exam scores is explained by study hours.
- ▶ Much better fit, less scatter around the regression line.

A higher R^2 indicates a stronger relationship between the predictor and the response, suggesting a better model fit.

Concluding Remarks on ANOVA

- ▶ Analysis of Variance (ANOVA) is a powerful statistical tool for examining the variation in response variables explained by different factors. It partitions total variation into the portion explained by the model (regression sum of squares) and the portion unexplained (error sum of squares).
- ▶ Key measure: R^2 , or the coefficient of determination, helps quantify how much variation in the response variable is explained by the model. A higher R^2 typically indicates a better fit.
- ▶ ANOVA also reveals residuals (unexplained variation), which suggest areas for potential improvement in the model or indicate the presence of additional factors not considered in the analysis.

Future Directions for Study

- ▶ **Generalized Linear Models (GLMs):** Extend ANOVA to handle non-linear relationships and complex error structures, including binary or count data.
- ▶ **Multivariate Analysis:** MANOVA allows analysis of multiple response variables simultaneously, useful when responses are correlated or when studying multiple outputs.
- ▶ **Mixed-Effects Models:** Incorporate both fixed and random effects (e.g., repeated measures or hierarchical data) for better analysis of complex experimental designs.
- ▶ **Non-Parametric Methods:** When assumptions of ANOVA (e.g., normality) are not met, methods like the Kruskal-Wallis test can be used as alternatives.
- ▶ **Improved Model Diagnostics:** Further focus on residual analysis and model diagnostics can guide model improvements and detect outliers that affect results.
- ▶ **Machine Learning and ANOVA:** Integrate ANOVA with machine learning techniques for feature selection and better model understanding, particularly in complex, high-dimensional datasets.

Recap. and exam discussions

Topics for the exam

- ▶ Population, samples, characteristics, sampling errors, frequencies.
- ▶ Descriptive statistics:
 - ▶ mean, median, mode; shape of a distribution
 - ▶ quantiles, quartiles, and percentiles; IQR & outliers
 - ▶ variance and standard deviation
- ▶ Estimators: unbiasedness, consistency, efficiency; central moments; point estimators vs. interval estimators.
- ▶ Pointwise estimators:
 - ▶ Method of moments
 - ▶ Maximum likelihood estimator
- ▶ Interval estimators: confidence intervals, with given confidence level (*understand how the CI intervals are constructed*).
 - ▶ Confidence interval for the mean: with known and with unknown variance, two-sided and one-sided
 - ▶ Student Distribution
 - ▶ Confidence interval for the variance
 - ▶ χ^2 distribution

- ▶ Foundational topics of statistical inference (with proofs)
 - ▶ Fisher's Information, Fisher's factorization criterion
 - ▶ Cramér-Rao
 - ▶ Rao-Blackwell Theorem
 - ▶ Lehmann - Scheffé Theorem
- ▶ Hypothesis Testing
 - ▶ Null hypothesis, alternative hypothesis
 - ▶ Test statistic, rejection region
 - ▶ Types of error: Type I, Type II
 - ▶ Designing a hypothesis test with a given significance level
(understanding the strategy)
 - ▶ Z-test for theoretical mean
 - ▶ T-test for theoretical mean
 - ▶ p -value; critical value vs p -value approach
 - ▶ χ^2 for standard deviation
 - ▶ F-test for ratio of variances.

- ▶ Regression
 - ▶ Correlation and regression
 - ▶ Curves/lines of regression
 - ▶ Correlation tables, conditional means, correlation coefficient
 - ▶ Least squares estimation, linear regression
 - ▶ Conditional expectation and residuals in regression
 - ▶ ANOVA, coefficient of determination (R^2)

Exam structure for AI

- ▶ 3 parts → 70 marks total; duration 1,5 hrs
 - 1. Theoretical question from descriptive statistics OR a foundational topic of statistical inference (see above) with proof.
 - 2. Problem involving pointwise estimators: method of moments/maximum likelihood estimator.
 - 3. Problem involving confidence intervals and/or hypothesis testing.

Exam structure for MIE

- ▶ 4 parts → 70 marks total; duration 2hrs
 - 1. Theoretical question from descriptive statistics OR a foundational topic of statistical inference (see above) with proof.
 - 2. Problem involving pointwise estimators: method of moments/maximum likelihood estimator.
 - 3. Problem involving confidence intervals and/or hypothesis testing.
 - 4. Problem involving regression.

Appendix A - Choosing the Test Statistic

Choosing the Test Statistic

The choice of the test statistic in hypothesis testing depends on several factors, such as the type of data we have, the nature of our hypothesis, and whether certain parameters (like the population standard deviation) are known. Below is a breakdown of how to choose the test statistic for different scenarios.

1. Data Type

- ▶ **Continuous Data:** For data that can take on any value within a range (e.g., height, weight, temperature), we typically use test statistics like the **z-test** or **t-test**.
- ▶ **Categorical Data:** For data in categories or groups (e.g., yes/no responses, treatment types), we use test statistics like the **chi-squared test**.

2. Parameter Known vs Unknown

- ▶ **Population Standard Deviation Known:** When the population standard deviation (σ) is known, we use a **z-test**. This is common when we are dealing with large samples ($n > 30$) or when we have a good estimate of the population variance.

$$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

- ▶ **Population Standard Deviation Unknown:** When the population standard deviation is not known, we use a **t-test**. This test is used when the sample size is small ($n < 30$) or the population standard deviation is unknown, and we estimate the standard deviation from the sample.

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

3. Sample Size

- ▶ **Large Sample Size ($n > 30$):** When the sample size is large, the **z-test** is often preferred. The Central Limit Theorem ensures that the sampling distribution of the sample mean is approximately normal, even if the underlying population distribution is not.
- ▶ **Small Sample Size ($n < 30$):** For small samples, the **t-test** is generally used, since the t-distribution is more appropriate when the sample size is small and the population standard deviation is unknown. The t-distribution accounts for the increased variability expected in smaller samples.

4. Hypothesis Type

- ▶ **One-Sided (One-Tailed) Test:** If the alternative hypothesis suggests that the parameter is either greater than or less than a certain value (e.g., $H_1 : \mu > \mu_0$ or $H_1 : \mu < \mu_0$), we use a **one-tailed test** with a corresponding test statistic.

Rejection region: $z > z_\alpha$ or $z < -z_\alpha$.

- ▶ **Two-Sided (Two-Tailed) Test:** If the alternative hypothesis suggests that the parameter is not equal to a certain value (e.g., $H_1 : \mu \neq \mu_0$), we use a **two-tailed test** with a corresponding test statistic.

Rejection region: $|z| > z_{\alpha/2}$.

5. Choice of Test Statistic by Specific Scenarios

Test Type	Population Var Known?	Test Statistic
One-Sample Mean	Yes	Z-test
One-Sample Mean Proportion	No	T-test
Two Independent Means	N/A	Z-test for proportions
Two Independent Means	Yes	Z-test
Paired Sample Mean	No	T-test
Categorical Data	N/A	T-test (paired)
		Chi-Square Test

Conclusions for Choosing the Test Statistic

The choice of test statistic depends on:

- ▶ The **type of hypothesis test** (means, proportions, variances).
- ▶ Whether the **population parameters** (e.g., standard deviation) are known.
- ▶ The **sample size**.

Summary Table for Choosing a Test Statistic:

Scenario	Sample Size	Pop. Info Known?	TS
Testing a mean	$n \geq 30$	Known σ	z-test
Testing a mean	$n < 30$	Unknown σ	t-test
Testing proportions	Large	Proportion known	z-test
Comparing two means	Large	Known σ	z-test
Comparing two means	Small	Unknown σ	t-test
Testing a variance	Any	Testing var, itself	χ^2 -test
Comparing variances	Any	Comparing 2 variances	F-test

Summary and Conclusions for Choosing the Test Statistic

The choice of the test statistic depends on the characteristics of the data, such as whether the population standard deviation is known, the sample size, and the type of hypothesis. The test statistic standardizes the data to compare observed outcomes against the null hypothesis, helping us make data-driven decisions. Choosing the correct test statistic is essential for performing valid hypothesis testing and making accurate inferences from the data.

Appendix B - Choosing the Rejection Region

Choosing the Rejection Region

The rejection region is chosen based on:

- ▶ The null hypothesis (H_0) and the alternative hypothesis (H_1).
- ▶ The desired significance level (α).

Purpose: The rejection region is the range of values where H_0 is rejected in favor of H_1 .

Steps to Choose the Rejection Region

1. Define the Hypotheses:

- ▶ H_0 : Null hypothesis (e.g., $\mu = 50$).
- ▶ H_1 : Alternative hypothesis:
 - ▶ Two-tailed: $\mu \neq 50$
 - ▶ Right-tailed: $\mu > 50$
 - ▶ Left-tailed: $\mu < 50$

2. Select the Significance Level (α).

3. Determine the Test Type: z-test, t-test, etc.

4. Find the Critical Value(s) from the appropriate distribution.

5. Decide the Rejection Region based on the alternative hypothesis.

Summary Table of Rejection Regions

Test Type	Alt. Hypothesis	Rejection Region	Critical Value
Two-Tailed Test	$H_1 : \mu \neq \mu_0$	$ z > z_{\alpha/2}$	$\pm z_{\alpha/2}$
Right-Tailed Test	$H_1 : \mu > \mu_0$	$z > z_{\alpha}$	z_{α}
Left-Tailed Test	$H_1 : \mu < \mu_0$	$z < -z_{\alpha}$	$-z_{\alpha}$

Example: Right-Tailed Test

Scenario: Test $H_0 : \mu = 50$ vs $H_1 : \mu > 50$ at $\alpha = 0.05$.

- ▶ Right-tailed test: $H_1 : \mu > 50$.
- ▶ Significance level: $\alpha = 0.05$.
- ▶ Critical value: From the standard normal table, $z_\alpha = 1.645$.
- ▶ Rejection region: $z > 1.645$.

Test Statistic: Compute

$$z = \frac{\bar{z} - 50}{\frac{\sigma}{\sqrt{n}}}.$$

If e.g. $z = 2.1$, it lies in the rejection region.

Conclusion: Reject H_0 .

Visual Representation of Rejection Regions

Two-Tailed Test:

Rejection Region: $|z| > z_{\alpha/2}$

Region: $(-\infty, -z_{\alpha/2}] \cup [z_{\alpha/2}, +\infty)$

Right-Tailed Test:

Rejection Region: $z > z_{\alpha}$

Left-Tailed Test:

Rejection Region: $z < -z_{\alpha}$

Appendix C

Why Choose 10?

The Threshold of 10 for Validity of the Normal Approximation:

- ▶ The condition $np_0(1 - p_0) \geq 10$ is a **rule of thumb** to ensure the normal approximation to the binomial distribution is valid.
- ▶ The central idea is to ensure that both the number of expected successes (np_0) and failures ($n(1 - p_0)$) are sufficiently large to create a symmetric, bell-shaped distribution.
- ▶ **Why 10?**
 - ▶ If $np_0(1 - p_0) \geq 10$, the binomial distribution is close enough to a normal distribution for practical purposes.
 - ▶ This ensures that both the *tail probabilities* and the *central part of the distribution* are well approximated by a normal curve.
 - ▶ For values $np_0(1 - p_0) < 10$, the distribution may become skewed and fail to be well approximated by the normal distribution.
- ▶ **Why not a higher value?**
 - ▶ A threshold of 10 provides a good balance between accuracy and practicality.
 - ▶ Using a higher threshold (e.g., 20) would be overly cautious and unnecessarily strict for most real-world applications.

Appendix D

Comparison of Two Variances: F-Distribution

When comparing two populations, it is often necessary to assess their variances. This analysis is commonly applied to evaluate differences in:

- ▶ Accuracy
- ▶ Stability
- ▶ Uncertainty
- ▶ Risks

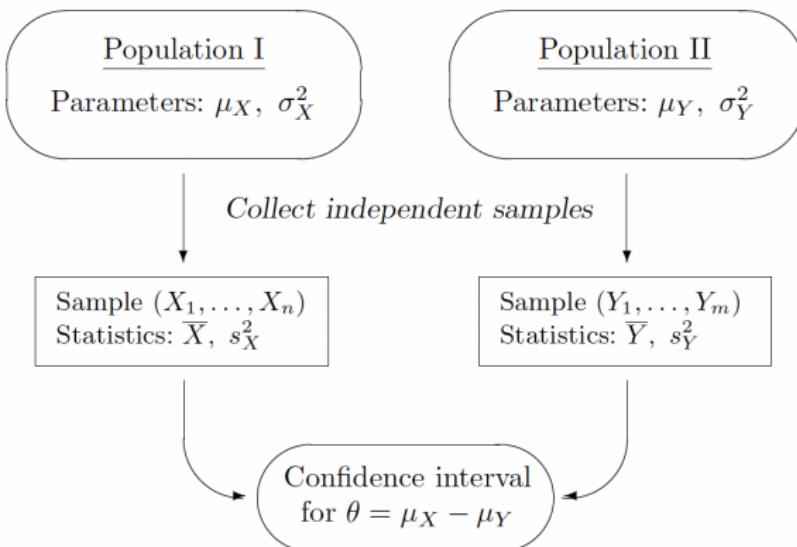
Example 1: A manufacturing process produces items with an average weight of 50 grams. A new production method is proposed to reduce variability in the weight while keeping the mean constant. Reduced variability corresponds to a lower standard deviation. How can we assess, with 90% confidence, the relative improvement in weight consistency achieved by the new method?

Comparison of Two Variances: F-Distribution

Example 2: Two airlines claim to have the same on-time arrival rate, but one of them shows 15% higher variability in arrival times over the past month. For passengers who prioritize reliability, is this sufficient evidence to favor the other airline? (Variability is measured by the standard deviation of arrival delays.)

Comparing Variances or Standard Deviations

To compare the variances or standard deviations of two populations, independent samples $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_m)$ are collected.



Comparing two populations [Baron fig. 9.4]

Unlike population means or proportions, variances are compared through their ratio:

$$\theta = \frac{\sigma_X^2}{\sigma_Y^2}.$$

A natural estimator for this ratio is the ratio of sample variances:

$$\hat{\theta} = \frac{s_X^2}{s_Y^2} = \frac{\sum(X_i - \bar{X})^2/(n-1)}{\sum(Y_i - \bar{Y})^2/(m-1)}.$$

This approach relies on the assumption of independent sampling from each population.

F-Distribution: Historical Background

The distribution of the ratio of sample variances was first derived in 1918 by the renowned English statistician and biologist Sir Ronald Fisher (1890–1962). It was later formalized in 1934 by the American mathematician George Snedecor (1881–1974).

The standardized form of this ratio is known as the Fisher-Snedecor distribution, or simply the ***F-distribution***, with $(n - 1)$ and $(m - 1)$ degrees of freedom.

Distribution of the Ratio of Variances: For independent samples X_1, \dots, X_n from $\mathcal{N}(\mu_X, \sigma_X)$ and Y_1, \dots, Y_m from $\mathcal{N}(\mu_Y, \sigma_Y)$, the standardized ratio:

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{\sum(X_i - \bar{X})^2/\sigma_X^2/(n - 1)}{\sum(Y_i - \bar{Y})^2/\sigma_Y^2/(m - 1)},$$

follows an ***F-distribution*** with $(n - 1)$ and $(m - 1)$ degrees of freedom.

Understanding the F-Distribution

The F -distribution plays a central role in comparing variances between two populations. It is characterized by two key parameters:

- ▶ **Numerator degrees of freedom** (ν_1), which correspond to the sample variance in the numerator of the F -ratio.
- ▶ **Denominator degrees of freedom** (ν_2), which correspond to the sample variance in the denominator.

These degrees of freedom arise directly from the sample sizes of the two groups being compared. For instance, if we are comparing variances from two independent samples, X of size n and Y of size m , the degrees of freedom are $\nu_1 = n - 1$ and $\nu_2 = m - 1$.

Critical Values: The critical values of the F -distribution can be found in statistical tables or calculated in e.g. Python. These values are crucial for two primary tasks:

- ▶ Constructing confidence intervals to estimate the ratio of variances.
- ▶ Performing hypothesis tests to determine whether the variances are significantly different.

The F -distribution thus provides a powerful framework for answering questions about variability and stability between two populations.

Note that in this lecture we consider that the critical value F_α is such that $\alpha = \mathbb{P}(F > F_\alpha)$. The values we consider for critical values should then be consistent with this. This is the case, if we use the statistical tables from [Baron]. However, in Python, 'stats.f.ppf' finds the critical value x such that the probability of observing a value less than or equal to x is equal to a given probability α i.e. $x = F^{-1}(\alpha)$ where F^{-1} is the cdf for the distribution we are working with (F distribution in this case).

Choosing the Ratio and Symmetry in the F-Distribution

Which ratio should we use? When comparing variances, we can compute the F -ratio either as:

$$F = \frac{s_X^2}{s_Y^2}, \quad \text{or} \quad F = \frac{s_Y^2}{s_X^2}.$$

Both approaches are valid, but the choice of ratio determines the specific F -distribution we are working with:

- ▶ If $F = \frac{s_X^2}{s_Y^2}$, the F -statistic follows $F(n - 1, m - 1)$.
- ▶ If $F = \frac{s_Y^2}{s_X^2}$, the F -statistic follows $F(m - 1, n - 1)$.

This difference leads to an important symmetry property of the F -distribution.

Key Insight: If F has an $F(\nu_1, \nu_2)$ distribution, then the reciprocal of F ,

$$\frac{1}{F},$$

follows an $F(\nu_2, \nu_1)$ distribution.

This means that regardless of the ratio chosen, the F -distribution remains a consistent and symmetric tool for comparing variances. However, it is important to clearly state which ratio was used to ensure correct interpretation of the results.

Confidence Interval for the Ratio of Variances

To construct a $(1 - \alpha)100\%$ confidence interval for the ratio of population variances, $\theta = \frac{\sigma_X^2}{\sigma_Y^2}$, we start by using the estimator:

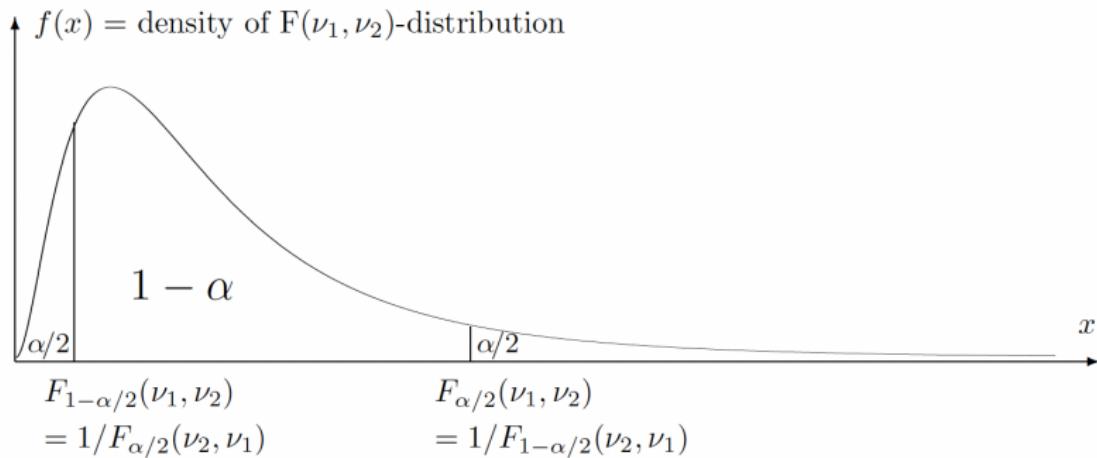
$$\hat{\theta} = \frac{s_X^2}{s_Y^2}.$$

To standardize this estimator, we introduce the F -statistic:

$$F = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{\hat{\theta}}{\theta}.$$

The F -statistic follows an F -distribution with $(n - 1)$ numerator and $(m - 1)$ denominator degrees of freedom.

Critical values for the F-distribution



The critical values $F_{1-\alpha/2}(n - 1, m - 1)$ and $F_{\alpha/2}(n - 1, m - 1)$ come from the F -distribution with degrees of freedom $(n - 1)$ and $(m - 1)$. These values are typically found in statistical tables. Picture from [Baron fig. 9.15].

Then

$$\mathbb{P} \left(F_{1-\alpha/2}(n-1, m-1) \leq \frac{\hat{\theta}}{\theta} \leq F_{\alpha/2}(n-1, m-1) \right) = 1 - \alpha.$$

By solving the inequality for θ , we obtain:

$$\mathbb{P} \left(\frac{\hat{\theta}}{F_{\alpha/2}(n-1, m-1)} \leq \theta \leq \frac{\hat{\theta}}{F_{1-\alpha/2}(n-1, m-1)} \right) = 1 - \alpha.$$

Confidence Interval for the Ratio of Variances

Finally, the $(1 - \alpha)100\%$ confidence interval for $\theta = \frac{\sigma_X^2}{\sigma_Y^2}$ becomes:

$$\left[\frac{\hat{\theta}}{F_{\alpha/2}(n-1, m-1)}, \frac{\hat{\theta}}{F_{1-\alpha/2}(n-1, m-1)} \right] =$$
$$\left[\frac{s_X^2/s_Y^2}{F_{\alpha/2}(n-1, m-1)}, \frac{s_X^2/s_Y^2}{F_{1-\alpha/2}(n-1, m-1)} \right].$$

Critical Values and Confidence Interval

Key Relationship: The critical values of the F -distributions are connected by:

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_\alpha(\nu_2, \nu_1)}.$$

Using this property, we can compute the $(1 - \alpha)100\%$ confidence interval for the ratio of variances $\theta = \frac{\sigma_X^2}{\sigma_Y^2}$ as:

$$\left[\frac{s_X^2}{s_Y^2 F_{\alpha/2}(n-1, m-1)}, \frac{s_X^2 F_{\alpha/2}(m-1, n-1)}{s_Y^2} \right].$$

Example: Efficiency in Machine Processes

Scenario: A factory upgrades its machines to improve the consistency of product quality.

- ▶ Before the upgrade, measurements of production times (in seconds) from 27 random observations yield a standard deviation of 22 seconds.
- ▶ After the upgrade, 16 random measurements show a reduced standard deviation of 14 seconds.

Goal: Construct a 90% confidence interval for the relative change in the standard deviation of production times, assuming normality.

Solution to the Example

Given Data:

- ▶ Before upgrade: $s_Y = 22$, $m = 27$.
- ▶ After upgrade: $s_X = 14$, $n = 16$.
- ▶ Confidence level: 90% $\Rightarrow \alpha = 0.10, \alpha/2 = 0.05$.

Step 1: Critical Values from F -Distribution. Using Python's `scipy.stats.f` library to calculate critical values:

$$F_{0.05}(15, 26) \approx 2.0716, \quad F_{0.05}(26, 15) \approx 2.2722.$$

Alternatively, these values can be approximated from statistical tables.

Step 2: Confidence Interval for the Variance Ratio. Using the formula:

$$\left[\frac{s_X^2}{s_Y^2 \cdot F_{\alpha/2}(n-1, m-1)}, \frac{s_X^2 \cdot F_{\alpha/2}(m-1, n-1)}{s_Y^2} \right],$$

substitute the values:

$$s_X^2 = 14^2 = 196, \quad s_Y^2 = 22^2 = 484.$$

Compute:

$$\left[\frac{196}{484 \cdot 2.0716}, \frac{196 \cdot 2.2722}{484} \right] = [0.20, 0.92].$$

Step 3: Confidence Interval for the Standard Deviation Ratio.

Taking the square root of the variance ratio confidence interval:

$$[\sqrt{0.20}, \sqrt{0.92}] = [0.44, 0.96].$$

Interpretation:

- ▶ We can state with 90% confidence that the new standard deviation is between 44% and 96% of the old standard deviation.
- ▶ The relative reduction in standard deviation is calculated as:

$$1 - \sqrt{\theta}, \quad \text{where } \sqrt{\theta} \in [0.44, 0.96].$$

- ▶ Therefore, the *relative reduction* in standard deviation is between 4% and 56%, corresponding to an improvement in stability of the data transfer.

F-tests for Comparing Two Variances

In this section, we focus on hypothesis testing for the ratio of variances between two populations. We test the null hypothesis

$$H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = \theta_0$$

against either a one-sided or two-sided alternative hypothesis. In many cases, we are only interested in testing if the two variances are equal, in which case we set $\theta_0 = 1$. When comparing variances, we use the F-distribution, which is why this test is called the **F-test**.

F-test Statistic

The test statistic for this hypothesis test is given by:

$$F = \frac{s_X^2}{s_Y^2} / \theta_0$$

Under the null hypothesis H_0 , this statistic simplifies to:

$$F = \frac{s_X^2 / \sigma_X^2}{s_Y^2 / \sigma_Y^2}$$

Where s_X^2 and s_Y^2 are the sample variances, and σ_X^2 and σ_Y^2 are the population variances. As we saw, if \mathbf{X} and \mathbf{Y} are independent samples

drawn from normal distributions, then the test statistic follows an F-distribution with degrees of freedom $(n - 1)$ and $(m - 1)$, where n and m are the sample sizes of \mathbf{X} and \mathbf{Y} , respectively.

Characteristics of the F-test

- ▶ The F-statistic, similar to the χ^2 -statistic, is non-negative. It has a right-skewed, non-symmetric distribution, meaning the values are mostly concentrated on the lower end of the scale, with a long tail to the right.
- ▶ To perform the hypothesis test, we compare the observed value of the F-statistic to the critical values from the F-distribution. For a level α test, the critical value is determined using the degrees of freedom for the numerator and denominator. The process of calculating p-values and making decisions is analogous to χ^2 -tests (see Appendix D).
- ▶ In summary: an F-test allows us to compare variances and assess the significance of their differences.

Example: F-Test Comparing Two Variances

A quality control manager in a factory wants to test if the variability in the production time of two machines, A and B , is the same.

- ▶ Data for machine A : 20 cycles, with sample variance $s_A^2 = 2.5$ (time units squared).
- ▶ Data for machine B : 30 cycles, with sample variance $s_B^2 = 1.8$ (time units squared).

The manager wants to test if the variances are equal at a 5% significance level.

Step 1: Define the Hypotheses

We test the following hypotheses:

$$H_0 : \frac{\sigma_A^2}{\sigma_B^2} = 1 \quad (\text{the variances are equal})$$

$$H_1 : \frac{\sigma_A^2}{\sigma_B^2} \neq 1 \quad (\text{the variances are not equal})$$

This is a two-sided F-test, where H_0 asserts that the ratio of the variances is equal to 1, i.e., $\sigma_A^2 = \sigma_B^2$.

Step 2: Test Statistic

The test statistic for the F-test is given by:

$$F = \frac{s_A^2}{s_B^2}$$

Substitute the values from the data:

$$F = \frac{2.5}{1.8} \approx 1.39$$

This is the observed F-statistic.

Step 3: Degrees of Freedom

The degrees of freedom for both samples are:

- ▶ For machine A : $n_A - 1 = 20 - 1 = 19$
- ▶ For machine B : $n_B - 1 = 30 - 1 = 29$

Thus, the F-statistic follows an F-distribution with 19 and 29 degrees of freedom.

Step 4: Critical Values

At a 5% significance level ($\alpha = 0.05$), the two-tailed critical values are:

$$F_{\alpha/2}(19, 29) = F_{0.025}(19, 29) \approx 0.416$$

$$F_{1-\alpha/2}(19, 29) = F_{0.975}(19, 29) \approx 2.231$$

Step 5: Compare F-Statistic to Critical Values

Now we compare the observed F-statistic $F = 1.39$ with the critical values:

- ▶ If $F < 0.416$ or $F > 2.231$, reject H_0 .
- ▶ If $0.416 \leq F \leq 2.231$, fail to reject H_0 .

Since 1.39 is between 0.416 and 2.231, we fail to reject H_0 .

⇒ There is insufficient evidence at the 5% significance level to conclude that the variances of the production times for machines A and B are different. In other words, we cannot reject the hypothesis that the variances are equal.

Summary of F-tests for the Ratio of Population Variances

Null Hypothesis $H_0 : \frac{\sigma_X^2}{\sigma_Y^2} = \theta_0$		Test Statistic $F_{\text{obs}} = \frac{s_X^2}{s_Y^2} / \theta_0$
Alternative Hypothesis		p-value
$\frac{\sigma_X^2}{\sigma_Y^2} > \theta_0$	$F_{\text{obs}} \geq F_\alpha(n - 1, m - 1)$	$\mathbb{P}(F \geq F_{\text{obs}})$
$\frac{\sigma_X^2}{\sigma_Y^2} < \theta_0$	$F_{\text{obs}} \leq F_\alpha(n - 1, m - 1)$	$\mathbb{P}(F \leq F_{\text{obs}})$
$\frac{\sigma_X^2}{\sigma_Y^2} \neq \theta_0$	$F_{\text{obs}} \geq F_{\alpha/2}(n - 1, m - 1) \text{ or}$ $F_{\text{obs}} < 1/F_{\alpha/2}(m - 1, n - 1)$	$2 \min(\mathbb{P}(F \geq F_{\text{obs}}), \mathbb{P}(F \leq F_{\text{obs}}))$