Data Science in Aerospace

Bachelor's Degree in Aeronautical and Space Sciences
Bachelor's Degree in Aeronautical Management
Short Cycle in Aircraft Repair and Maintenance

Emanuel A. R. Camacho

emanuel.camacho@iseclisboa.pt

Instituto Superior de Educação e Ciências (ISEC Lisboa)

Bibliography

- Pestana D. D., Velosa S F., Introdução à Probabilidade e à Estatística, vol. 1, Fundação Calouste Gulbenkian, Lisboa, 2010.
- Murteira Bento, Antunes Marilia, Probabilidades e Estatística, vol. 1, Escolar Editora, Lisboa, 2012.
- Andy Field, Jeremy Miles, Zoe Field, Discovering Statistics Using R - SAGE Publications Ltd, 1st edition (April 5, 2012).
- Hadley Wickham, Garret Grolemund, R for Data Science: Import, Tidy, Transform, Visualiza, and Model Data - O'Reilly Media 1st edition (January 17, 2017).
- Murray R. Spiegel, Pedro Consentino, Carlos de Lucena, Estatística, McGraw-Hill, São Paulo, 1976.
- Sheldon M. Ross, Introduction to Probability and Statistics For Engineers and Scientists - Fifth Edition, Elsevier Inc, 2014.

Evaluation

Data Science in Aerospace (100% [20/20])

Frequencies (80% [16/20]) + Project (20% [4/20])

- Frequency 1 (40% [8/20]) - 04/05/2026
- Frequency 2 (40% [8/20]) - 09/06/2026
- Project (20% [4/20]) - Send reports until the end of the semester

or

Exam (100% [20/20])

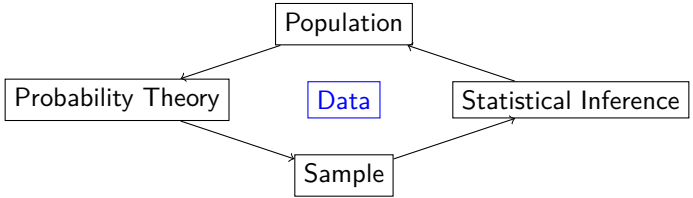
- Exam (100% [20/20])

There is no minimum score for any component of the evaluation. [10/20] is required to pass.

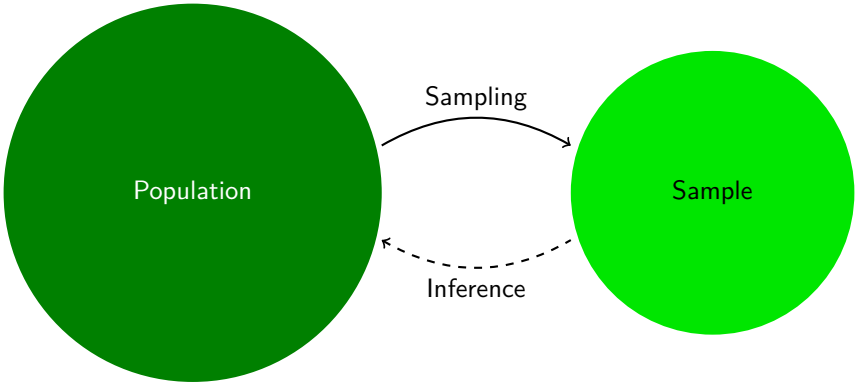
Table of Contents

Outline

Table of Contents



- Data: The information collected which serves as the foundation for analysis.
- Population: Represents the entire group or dataset that is the subject of study.
- Sample: A subset of the population, selected to make inferences about the population.
- Probability Theory: The mathematical framework used to model randomness and uncertainty in data.
- Statistical Inference: The process of drawing conclusions about the population based on sample data.


Population
Sampling
Sample
Inference

Population	**Sample**
A population is the entire collection of subjects affected by your research question.	A sample is a subset of the population you study.
Measurements taken from a whole population are called parameters.	Measurements taken from a sample are called statistics.
Data for an entire population is often very difficult or impossible to collect.	When population data is unavailable, we use sample data to make inferences about the population.
If you do have data for a whole population, your parameters will be "true" measures of some population characteristic.	Sample data yield statistics, which can be used to estimate population parameters.

Probability Sampling

- Simple Random Sampling: Every individual in the population has an equal chance of being selected.
- Stratified Sampling: The population is divided into subgroups, and samples are taken from each subgroup.
- Systematic Sampling: Individuals are randomly selected from a list or sequence at regular intervals.
- Cluster Sampling: The population is divided into clusters, and entire clusters are randomly selected for the sample.
- Multistage Sampling: Combines multiple sampling methods, often used for large populations

Non-probability Sampling

- Convenience Sampling: Participants are chosen based on ease of access.
- Voluntary Response Sampling: Participants self-select to be part of the study.
- Purposive (Judgmental) Sampling: Researchers select participants based on specific criteria or purpose.
- Snowball Sampling: Existing participants recruit others, often used for hard-to-reach populations.
- Quota Sampling: A sample is created to represent certain characteristics or proportions in the population.

Measurement scales, also known as levels of measurement, describe how variables are classified and the type of information they represent. There are four main types of measurement scales, each with increasing levels of precision and mathematical applicability:

- Nominal Scale
- Ordinal Scale
- Interval Scale
- Ratio Scale

Nominal Scale (Categorical, No Order)

Definition: Data is grouped into categories without any inherent ranking.

- Types of aircraft: Boeing, Airbus, Cessna.
- Airport codes: JFK, LAX, ORD.
- Flight status: On-time, delayed, canceled.
- Airline names: Delta, American Airlines, Lufthansa.

Ordinal Scale (Categorical, Ordered)

Definition: Data is grouped into categories with a meaningful order, but the intervals between ranks are not uniform.

- Passenger seating classes: Economy, Business, First Class.
- Pilot experience levels: Trainee, Junior, Senior.
- Turbulence severity: Light, Moderate, Severe.
- Customer satisfaction ratings for a flight: Poor, Fair, Good, Excellent.

Interval Scale (Ordered with Equal Intervals, No True Zero)

Definition: Data is ordered with equal intervals between values but lacks a true zero point.

- Temperature inside the cabin (measured in Celsius or Fahrenheit).
- Time of day for departure or arrival (e.g., 2 PM vs. 3 PM).
- Altitude above sea level in standard atmospheric pressure levels (e.g., flight levels in hundreds of feet).

Ratio Scale (Ordered with Equal Intervals and a True Zero)

Definition: Data is ordered with equal intervals and a true zero point; ratios are meaningful.

- Flight duration (e.g., a flight lasting 0 hours means no flight occurred).
- Distance traveled by an aircraft (e.g., miles or kilometers).
- Fuel consumption during a flight (e.g., gallons or liters).
- Aircraft weight (e.g., kilograms or pounds).

1. **Collection**: Gathering raw data from various sources, ensuring it is relevant and accurate.
2. **Preparation**: Cleaning and organizing the data by removing errors, duplicates, or irrelevant information (also known as data cleansing).
3. **Input**: Entering the prepared data into a processing system, either manually or through automated methods.
4. **Processing**: Transforming the data using techniques such as sorting, filtering, aggregating, and statistical calculations to extract meaningful insights.
5. **Output and Interpretation**: Presenting the processed data in readable formats like graphs, tables, or charts for interpretation and decision-making.
6. **Storage**: Storing the processed data for future use or further analysis.

- Handling missing data
  - Remove incomplete records (listwise deletion).
  - Impute missing values using methods such as: Mean, median, or mode substitution. Regression or machine learning models. Multiple imputation techniques.
  - Use statistical techniques that can handle missing data directly

- Dealing with outliers
  - Identify outliers using statistical methods (e.g., Z-scores, IQR method).
  - Remove the outliers (if they are errors or irrelevant to the analysis).
  - Transform the data (e.g., log transformation) to reduce their impact.
  - Retain them if they represent meaningful phenomena.

- Data normalization and scaling
  - Adjusting data values to ensure consistency across variables, especially when they have different units or scales.
  - Min-Max Scaling: Rescale data to a fixed range (e.g., 0 to 1).
  - Standardization: Transform data to have a mean of 0 and a standard deviation of 1.

- Encoding categorical variables
  - Definition: Converting non-numerical (categorical) data into numerical formats for analysis.
  - One-Hot Encoding: Create binary columns for each category.
  - Label Encoding: Assign numerical labels to categories.

- Removing duplicates
  - Identifying and removing duplicate records that can skew results.
  - Use software tools or algorithms to detect and eliminate duplicate rows or entries.

- Addressing data inconsistencies
  - Resolving discrepancies in data caused by errors in recording or formatting.
  - Standardizing date formats (e.g., MM/DD/YYYY vs. DD/MM/YYYY).
  - Correcting misspelled entries in categorical variables.

- Transforming data
  - Applying mathematical transformations to make the data suitable for statistical methods.
  - Logarithmic transformation for skewed data.
  - Square root transformation for reducing variance.
  - Binning continuous variables into categories.

- Dealing with Multicollinearity
  - Multicollinearity occurs when independent variables are highly correlated.
  - Remove one of the correlated variables.
  - Combine correlated variables using several techniques.

Table of Contents

"Python is a programming language that lets you work quickly and integrate systems more effectively"

Download at: *https://www.python.org*

Library/Module	**Feature**
SciPy Stats	Provides both discrete (e.g., `stats.binom`, `stats.poisson`, `stats.geom`) and continuous distributions (e.g., `stats.norm`, `stats.expon`, `stats.beta`), along with methods like `.pmf()`, `.pdf()`, `.cdf()`, and `.rvs()`.
Python `statistics` Module	Offers summary statistical functions such as `mean`, `variance`, `stdev`, and `quantiles`.

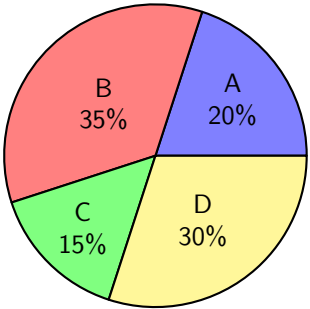
"Python is a programming language that lets you work quickly and integrate systems more effectively"

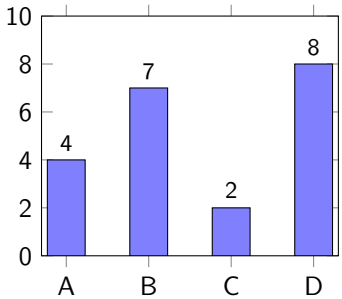
Download at: *https://www.python.org*

NumPy	Facilitates random variable generation through functions like `np.random.normal`, `np.random.poisson`, and `np.random.choice`.
Pandas	Provides descriptive statistics using methods such as `DataFrame.describe()`, `DataFrame.mean()`, and `DataFrame.std()`.
Statsmodels	Includes tools for statistical modeling with functions for regression, hypothesis testing, and time-series analysis.

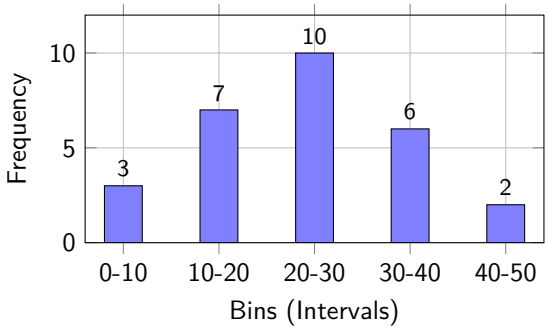
Table of Contents

- Use bar charts or pie charts for categorical comparisons.

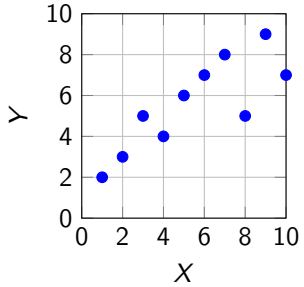


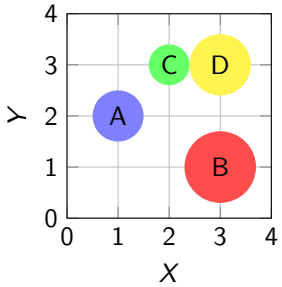


- Use histograms, box plots, or density plots for distributions.



- Use scatter plots or bubble charts for relationships between variables.





Sample Mean (Arithmetic Average)

The *sample mean*, designated by $\bar{x}$, is defined by

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{1}$$

The computation of the sample mean can often be simplified by noting that if for constants a and b

$$y_i = ax_i + b, \quad i = 1, \ldots, n \tag{2}$$

Sample Mean (Arithmetic Average)

Then the sample mean of the data set $y_1, \ldots, y_n$ is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = \frac{1}{n}\sum_{i=1}^{n} ax_i + \frac{1}{n}\sum_{i=1}^{n} b = a\bar{x} + b \tag{3}$$

- Uses all data points
- Excellent measurement when data is symmetrically distributed
- Highly affected by outliers or skewed data

Sample Mean (Arithmetic Average)

When using a frequency table listing, we have k distinct values $v_1, \ldots, v_k$ having corresponding frequencies $f_1, \ldots, f_k$. Since such a data set consists of

$$n = \sum_{i=1}^{k} f_i$$

observations, with the value v_i appearing f_i times, for each $i = 1, \ldots, k$, it follows that the sample mean of these n data values is

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} v_i f_i \tag{4}$$

Median (Middle Value)

Another statistic used to indicate the center of a data set is the *sample median*. In simple terms, it is the middle value when the data set is arranged in increasing order.

Order the values of a data set of size n from smallest to largest:

- If n is odd, the *sample median* is the value in position $(n+1)/2$
- If n is even, it is the average of the values in positions $n/2$ and $n/2+1$.

- Resistant to outliers
- Useful for skewed distributions or ordinal data
- Does not consider directly all data points

Mode (Most Frequent Value)

Another statistic that has been used to indicate the central tendency of a data set is the *sample mode*, defined to be the value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called *modal values*.

- Useful for categorical data
- May not exist or may have multiple modes in some datasets

Sample Variance

The *sample variance*, call it s^2, of the data set $x_1, \cdots, x_n$ is defined by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{5}$$

- Provides a precise numerical measure of variability
- Units are squared, making it less intuitive than standard deviation
- Sensitive to outliers due to squaring deviations

Sample Standard Deviation

The positive square root of the sample variance is called the *sample standard deviation*. The quantity s, is mathematically defined as

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{6}$$

- Expressed in original units for easy interpretation.

Coefficient of Variation (CV)

The *coefficient of variation* (CV) is a statistical measure that represents the ratio of the standard deviation (s) to the mean ($\bar{x}$) of a data set. It is often expressed as a percentage and is used to assess the relative variability or dispersion of data compared to its mean. The formula is:

$$CV = \frac{s}{\bar{x}} \times 100 \tag{7}$$

- The CV is unitless, allowing for comparisons between data sets with different units or scales.

Pearson's Second Coefficient of Skewness

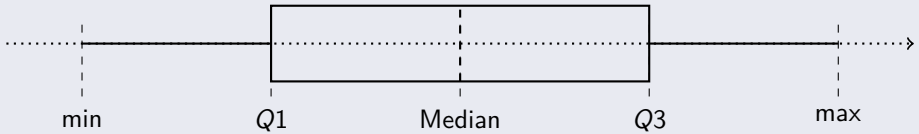
The Pearson's second coefficient of skewness is a measure used to quantify the asymmetry of a probability distribution around its mean. It is given by

$$\frac{3(\text{mean} - \text{median})}{\text{standard deviation}} = \frac{3(\bar{x} - \text{Med})}{s} \tag{8}$$

- Symmetric Distribution: If the data is perfectly symmetric, the mean and median are equal.
- Positive Skew (Right-Skewed): If the mean is greater than the median, the skewness will be positive.
- Negative Skew (Left-Skewed): If the mean is less than the median, the skewness will be negative.

Five-Number Summary

- Describe the center and spread of data
- Identify potential outliers using the interquartile range (IQR $= Q3 - Q1$).



$$\text{range} = \max x - \min x \tag{9}$$

$$\text{IQR} = Q3 - Q1 \tag{10}$$

z-Score

- A z-score is a statistical tool that indicates how many standard deviations a particular data point is from the mean of a dataset.

$$z = \frac{x - \bar{x}}{s} \tag{11}$$

- **Assumes normal distribution for meaningful interpretation.**
- Provides a standardized way to compare values across datasets.
- Helps identify unusual or extreme values in data. Typically, values with a z-score beyond ± 2 or ± 3 indicate outliers.
- Useful statistical tool for hypothesis testing.

Sample Correlation Coefficient

Consider the data pairs $(x_i, y_i), i = 1, \ldots, n$. The *sample correlation coefficient*, call it r, of the data pairs $(x_i, y_i), i = 1, \ldots, n$, is defined by

$$-1 \leq r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \leq 1 \tag{12}$$

- When $r > 0$, we say that the sample data pairs are *positively correlated*
- When $r < 0$, we say that the sample data pairs are *negatively correlated*

Descriptive Statistics

End of Section

Descriptive Statistics Exercises

Table of Contents

In simple terms, probability is a quantification of likelihood of an event occurring, expressed as a number between 0 and 1. The probability of an event A occurring is calculated using

$$P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of outcomes}}. \tag{13}$$

"two outcomes = equal probabilities" only applies when the outcomes are equally likely

$$P(A) = \frac{\text{Likelihood of favorable outcome}}{\text{Sum of likelihoods of all possible outcomes}} \tag{14}$$

The probability of the outcome will then be observable as being the proportion of the experiments that result in the outcome.

Frequently, to compute probabilities it is often necessary to be able to effectively count the number of different ways that a given event can occur. To do this, we will make use of combinatorics.

Generalized Basic Principle of Counting

If r experiments that are to be performed are such that the first one may result in any of n_1 possible outcomes, and if for each of these n_1 possible outcomes there are n_2 possible outcomes of the second experiment, and if for each of the possible outcomes of the first two experiments there are n_3 possible outcomes of the third experiment, and if, ... , then there are a total of $n_1 \cdot n_2 \cdots n_r$ possible outcomes of the r experiments.

$$n_1 \times n_2 \times \cdots \times n_r \tag{15}$$

Permutations

The number of *permutations* of n objects taken r at a time is

$$P(n, r) = \frac{n!}{(n-r)!} \tag{16}$$

where $n!$ is the n *factorial*.

- Boarding Sequences
- Runway Departure Orders
- Flight Schedules
- Cargo Loading

Combinations

The number of *combinations* of n objects taken r at a time is

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{17}$$

where $n!$ is the n *factorial*.

- Air Cargo Palletization
- Crew Pairing and Scheduling
- Aircraft Maintenance Planning
- Airport Slot Allocation

- Basic Principle of Counting

$$n_1 \times n_2 \times \cdots \times n_k \tag{18}$$

- Factorials

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 1 \tag{19}$$

- Permutations

$$P(n,r) = \frac{n!}{(n-r)!} \tag{20}$$

- Combinations

$$C(n,r) = \binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{21}$$

Probability

Pause

Combinatorial Analysis Exercises

Consider an experiment whose outcome is not predictable with certainty in advance. Although the outcome of the experiment will not be known in advance, let us suppose that the set of all possible outcomes is known. This set of all possible outcomes of an experiment is known as the *sample space* of the experiment and is denoted by $\mathcal{S}$.

Example: Takeoff Outcome

$$\mathcal{S} = \{S, A, E, R\} \tag{22}$$

- Successful Takeoff (S)
- Aborted Takeoff (A)
- Engine Failure during Takeoff (E)
- Runway Excursion (R)

Any subset E of the sample space, $\mathcal{S}$, is known as an *event*.

For any two events E and F

- The new event $E \cup F$ is called the *union* of the events E and F
- The new event $E \cap F$ is called the *intersection* of E and F
- When $E \cap F = \emptyset$ (cannot both occur), E and F are said to be *mutually exclusive*
- The new event $\overline{E}$ (or E^c) is referred to as the *complement* of E.
- If all of the outcomes in E are also in F, then $E \subset F$, meaning that E is *contained* in F.
- If $E \subset F$ and $F \subset E$, then the two events are *equal*, $E = F$.

Commutative law

- $E \cup F = F \cup E$
- $E \cap F = F \cap E$

Associative law

- $(E \cup F) \cup G = E \cup (F \cup G)$
- $(E \cap F) \cap G = E \cap (F \cap G)$

Distributive law

- $(E \cup F) \cap G = (E \cap G) \cup (F \cap G)$
- $(E \cap F) \cup G = (E \cup G) \cap (F \cup G)$

Axiom 1

The probability of any event is always some number between 0 and 1

$$0 \leq P(E) \leq 1 \tag{23}$$

This means that probabilities cannot be negative, reflecting the intuitive idea that an event cannot occur less than "never". Similarly, it cannot occur more than always.

Axiom 2

The probability of the entire sample space $\mathcal{S}$, which represents all possible outcomes of an experiment, is equal to 1:

$$P(\mathcal{S}) = 1 \tag{24}$$

This axiom ensures that something in the sample space will occur with certainty

Axiom 3

For any countable sequence of mutually exclusive events $E_1, E_2, E_3, \ldots$, the probability of their union is equal to the sum of their individual probabilities:

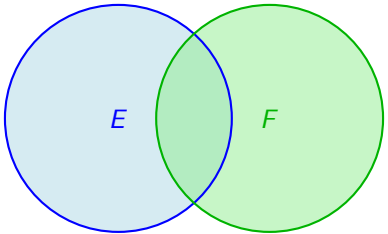
$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) \qquad n = 1, 2, \ldots, \infty \tag{25}$$

This axiom governs how probabilities combine for events that cannot occur simultaneously

Prepositions

$$P(\bar{E}) = 1 - P(E) \tag{26}$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{27}$$



When some partial information concerning the result of the experiment is available, or there is the need to recalculate a probability in light of additional information, we are interested in *conditional probability*.

Conditional probability is the probability of an event occurring given that another event has already occurred. The conditional probability of event E given event F (assuming $P(F) > 0$) is defined as:

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{28}$$

This formula automatically adjusts the sample space to consider only the outcomes where F has occurred.

Bayes' Formula

Bayes's theorem indicates that for events E and F, where $P(F) > 0$,

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)} \tag{29}$$

- $P(E|F)$ is the probability of observing event E given that F occured
- $P(F|E)$ is the probability of observing event F given that E occured
- $P(E)$ and $P(F)$ are the probabilities of observing events E and F, respectively, without any given conditions.

Law of Total Probability

Suppose that $F_1, F_2, \ldots, F_n$ are mutually exclusive events such that

$$\bigcup_{i=1}^{n} F_i = \mathcal{S} \tag{30}$$

This means that exactly one of the events $F_1, F_2, \ldots, F_n$ must occur. Considering now the fact that events $E \cap F_i$ are mutually exclusive

$$E = \bigcup_{i=1}^{n} EF_i \tag{31}$$

Law of Total Probability

This means that exactly one of the events $F_1, F_2, \ldots, F_n$ must occur. Considering now the fact that events $E \cap F_i$ are mutually exclusive

$$E = \bigcup_{i=1}^{n} EF_i \tag{32}$$

$$P(E) = \sum_{i=1}^{n} P(E \cap F_i) \tag{33}$$

$$P(E) = \sum_{i=1}^{n} P(E|F_i)P(F_i) \tag{34}$$

We have seen before that $P(E|F)$, the conditional probability of E given F, may not be equal to $P(E)$, the *unconditional probability* of E. What happens when these are the same?

$$P(E|F) = P(E) \tag{35}$$

$$\frac{P(E \cap F)}{P(F)} = P(E) \tag{36}$$

$$P(E \cap F) = P(E)P(F) \tag{37}$$

In this conditions, the two events E and F are said to be *independent*. These conditions mean that the probability of both events happening together is simply the product of their individual probabilities.

Independent Events

$$P(E \cap F) = P(E)P(F) \tag{38}$$

Important Notes

- Independence is different from mutual exclusivity. Two mutually exclusive events cannot happen at the same time ($P(E \cap F) = 0$), whereas independent events can occur simultaneously.
- For more than two events ($E, F, G, \ldots$), pairwise independence (each pair is independent) does not necessarily imply mutual independence (all combinations of events are independent).

Probability

$$P(E) \in [0,1] \tag{39}$$

$$P(\overline{E}) = 1 - P(E) \tag{40}$$

$$P(E \cup F) = P(E) + P(F) \text{ (when mutually exclusive)} \tag{41}$$

$$P(E \cup F) = P(E) + P(F) - P(E \cap F) \tag{42}$$

$$P(E \cap F) = P(E)P(F) \text{ (when independent)} \tag{43}$$

$$P(E \cap F) = P(E|F)P(F) = P(F|E)P(E) \tag{44}$$

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(F|E)P(E)}{P(F)} \tag{45}$$

Probability

Pause

Elementary Probability Theory Exercises

A *random variable* is a mathematical concept used in probability and statistics to represent numerical outcomes of random phenomena or experiments. It is a function that assigns a numerical value to each possible outcome in the sample space of a random process.

- *Discrete Random Variables*: These take on a countable set of values, such as integers. For example, the number of landings per day is discrete.
- *Continuous Random Variables*: These can take any value within a continuous range, such as measurements of altitude or temperature.

For a discrete random variable X, we define the *probability mass function* $p(a)$ of X by

$$p(a) = P\{X = a\} \tag{46}$$

The probability mass function $p(a)$ is positive for at most a countable number of values of a. That is, if X must assume one of the values $x_1, x_2, \ldots$, then

$$p(x_i) > 0, \quad i = 1, 2, \ldots \tag{47}$$

$$p(x) = 0, \quad \text{all other values of } x \tag{48}$$

Since X must take on one of the values x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1 \tag{49}$$

The *cumulative distribution function* F can be expressed in terms of $p(x)$ by

$$F(a) = \sum_{\text{all } x \leq a} p(x) \tag{50}$$

For a continuous random variable X, there is a nonnegative function $f(x)$, defined for all real $x \in]-\infty, \infty[$, having the property that for any set B of real numbers

$$P\{X \in B\} = \int_B f(x)\, dx \tag{51}$$

The function $f(x)$ is called the *probability density function* of the random variable X. Since X must assume some value, $f(x)$ must satisfy

$$P\{X \in]-\infty, \infty[\} = \int_{-\infty}^{\infty} f(x)\, dx = 1 \tag{52}$$

All probability statements about X can be answered using $f(x)$. For instance, when $B = [a, b]$, we obtain that

$$P\{X \in B\} = \int_B f(x)\, dx \tag{53}$$

$$P\{a \leq X \leq b\} = \int_a^b f(x)\, dx \tag{54}$$

When considering that $a = b$, then

$$P\{X = a\} = \int_a^a f(x)\, dx = 0 \tag{55}$$

The relationship between the cumulative distribution F and the probability density function f is expressed by

$$F(a) = P\{X \in]-\infty, a]\} = \int_{-\infty}^{a} f(x)\, dx \tag{56}$$

Differentiating both sides yields

$$\frac{d}{da}F(a) = f(a) \tag{57}$$

Expectation

One of the most important concepts in probability theory is that of the expectation of a random variable.

If X is a discrete random variable taking on the possible values $x_1, x_2, \ldots$, then the *expectation* or *expected value* of X, denoted by $E[X]$, is defined by

$$E[X] = \sum_i x_i P\{X = x_i\} \tag{58}$$

Hence, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it.

Expectation

We can also define the expectation of a continuous random variable. Suppose that X is a continuous random variable with probability density function f. Since, for dx small,

$$f(x)\, dx \approx P\{x < X < x + dx\} \tag{59}$$

it follows that a weighted average of all possible values of X, with the weight given to x equal to the probability that X is near x, is just the integral over all x of $xf(x)\, dx$. Hence, it is natural to define the expected value of X by

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx \tag{60}$$

Expectation of a Function of a Random Variable

- If X is a discrete random variable with probability mass function $p(x)$, then for any real-valued function g,

$$E[g(X)] = \sum_x g(x)p(x) \tag{61}$$

- If X is a continuous random variable with probability density function $f(x)$, then for any real-valued function g,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx \tag{62}$$

Property

Let X be a discrete (or continuous) random variable with probability mass function (or probability density function), $f(x)$, and let $a \neq 0$ and b be real constants. The expected value of $aX + b$ is:

$$E[aX + B] = aE[X] + b \tag{63}$$

In the discrete case,

$$E[aX + b] = \sum_x (ax + b)p(x) = a\sum_x xp(x) + b\sum_x p(x) = aE[X] + b. \tag{64}$$

Property

Let X be a discrete (or continuous) random variable with probability mass function (or probability density function), $f(x)$, and let $a \neq 0$ and b be real constants. The expected value of $aX + b$ is:

$$E[aX + B] = aE[X] + b \tag{65}$$

In the continuous case,

$$E[aX + b] = \int_{-\infty}^{\infty} (ax + b) f(x) dx = a \int_{-\infty}^{\infty} x f(x) dx + b \int_{-\infty}^{\infty} f(x) dx = aE[X] + b. \tag{66}$$

Simple Moments

Simple moments provide information about the distribution's location and scale.

The expected value of a random variable X, $E[X]$, is also referred to as the *mean* or the *first moment* of X. The quantity $E[X^n]$, $n \geq 1$, is called the nth moment of X.

$$\textbf{Discrete case: } E[X^n] = \sum_x x^n p(x) \tag{67}$$

$$\textbf{Continuous case: } E[X^n] = \int_{-\infty}^{\infty} x^n f(x)\, dx \tag{68}$$

Central Moments

Central moments are calculated about the mean of the distribution, focusing on deviations from this central point, describing variability and shape.

Let X be a random variable and let n be a positive integer. The expected value of $(X - E(X))^n$, known as the *central moment* of order n of X, if it exists, is given by

$$\textbf{Discrete case: } E[(X - E(X))^n] = \sum_x (x - E(X))^n p(x) \tag{69}$$

$$\textbf{Continuous case: } E[(X - E(X))^n] = \int_{-\infty}^{\infty} (x - E(X))^n f(x)\, dx \tag{70}$$

Variance as the Second Central Moment

If X is a random variable with mean μ, then the *variance* of X, denoted by $\text{Var}(X)$, is defined by

$$\text{Var}(X) = E[(X - \mu)^2] \tag{71}$$

$$\textbf{Discrete case: } \text{Var}(X) = E[(X - E[X])^2] = \sum_x (x - E[X])^2 p(x) \tag{72}$$

$$\textbf{Continuous case: } \text{Var}(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)\, dx \tag{73}$$

Alternative Formula

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= E[X^2 - 2\mu X + \mu^2] \\ &= E[X^2] - E[2\mu X] + E[\mu^2] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - \mu^2 \end{aligned} \tag{74}$$

$$\text{Var}(X) = E[X^2] - (E[X])^2 \tag{75}$$

Property

$$\begin{aligned}\text{Var}(aX+b) &= E[(aX+b-E[aX+b])^2] \\ &= E[(aX+b-a\mu-b)^2] \\ &= E[(aX-a\mu)^2] \\ &= E[a^2(X-\mu)^2] \\ &= a^2E[(X-\mu)^2] \\ &= a^2\text{Var}(X)\end{aligned} \tag{76}$$

Property

Specifying particular values for a and b in $\text{Var}(aX + b) = a^2\text{Var}(X)$ leads to some interesting corollaries. For instance, by setting $a = 0$ we obtain that

$$\text{Var}(b) = 0 \tag{77}$$

meaning that the variance of a constant is 0.

Similarly, by setting $a = 1$ we obtain

$$\text{Var}(X + b) = \text{Var}(X) \tag{78}$$

The quantity $\sigma = \sqrt{\text{Var}(X)}$ is called the *standard deviation* of X. The standard deviation has the same units as does the mean.

Probability

Pause

Random Variables Exercises

Jointly Distributed Random Variables

For a given experiment, we are often interested not only in probability distribution functions of individual random variables but also in the relationships between two or more random variables

To specify the relationship between two random variables, we define the *joint cumulative probability distribution function* of X and Y by

$$F(x, y) = P\{X \leq x, Y \leq y\} \tag{79}$$

One can also define the distribution functions of X and Y, respectively, as

$$F_X(x) = P\{X \leq x, Y < \infty\} \text{ and } F_Y(x) = P\{X < \infty, Y \leq y\} \tag{80}$$

Jointly Distributed Random Variables

When X and Y are both discrete random variables whose possible values are, respectively, $x_1, x_2, \ldots$, and $y_1, y_2, \ldots$, we define the *joint probability mass function* of X and Y, $p(x_i, y_j)$, by

$$p(x_i, y_j) = P\{X = x_i, Y = y_j\} \tag{81}$$

Since Y must take on some value y_j, it follows that the event $\{X = x_i\}$ can be written as the union, for all j, of the mutually exclusive events $\{X = x_i, Y = y_j\}$.

$$\{X = x_i\} = \bigcup_j \{X = x_i, Y = y_j\} \tag{82}$$

Using Axiom 3 of the probability function, we obtain

$$P\{X = x_i\} = P\left(\bigcup_j \{X = x_i, Y = y_j\}\right) \tag{83}$$

which results in

$$P\{X = x_i\} = \sum_j p(x_i, y_j) \tag{84}$$

Similarly, $P\{Y = y_j\}$ can be obtained by

$$P\{Y = y_j\} = \sum_i p(x_i, y_j) \tag{85}$$

When X and Y are jointly continuous random variables, there is a function $f(x, y)$ defined for all real x and y having the property that for every set C of pairs of real numbers

$$P\{(X, Y) \in C\} = \iint_{(x,y)\in C} f(x, y)\, dx\, dy \tag{86}$$

The function $f(x, y)$ is called the *jointly probability density function* of X and Y. If A and B are any sets of real numbers, then by defining $C = \{(x, y) : x \in A, y \in B\}$, we see that

$$P\{X \in A, Y \in B\} = \int_A \int_B f(x, y)\, dx\, dy \tag{87}$$

Jointly Distributed Random Variables

Based on $f(x, y)$, the cumulative distribution F is defined as

$$F(a, b) = P\{X \leq a, Y \leq b\} \tag{88}$$
$$= P\{X \in]-\infty, a], Y \in]-\infty, b]\} \tag{89}$$
$$= \int_{-\infty}^{b} \int_{-\infty}^{a} f(x, y)\, dx\, dy \tag{90}$$

Upon differentiation

$$f(a, b) = \frac{\partial^2}{\partial a \partial b} F(a, b) \tag{91}$$

Jointly Distributed Random Variables

Concerning the individual probability density functions of X, f_X, and Y, f_Y,

$$P\{X \in A\} = P\{X \in A, Y \in]-\infty, \infty[\} \tag{92}$$

$$= \int_A \int_{-\infty}^{\infty} f(x,y)\, dy\, dx \tag{93}$$

$$= \int_A f_X(x)\, dx \tag{94}$$

Hence, the probability density functions of X and Y are given, respectively, by

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \text{ and } f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx \tag{95}$$

Independent Random Variables

The random variables X and Y are said to be independent if for any two sets of real numbers A and B,

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\} \tag{96}$$

In other words, X and Y are independent if, for all A and B, the events $E_A = \{X \in A\}$ and $F_B = \{Y \in B\}$ are independent. Using the three axioms of probability, it can be shown that

$$P\{X \leq A, Y \leq B\} = P\{X \leq A\}P\{Y \leq B\} \tag{97}$$

$$F(a, b) = F_X(a)F_Y(b) \tag{98}$$

Independent Random Variables

When X and Y are discrete random variables, the condition of independence is equivalent to

$$p(x, y) = p_X(x)p_Y(y) \tag{99}$$

where p_X and p_Y are the probability mass functions of X and Y.

In the jointly continuous case, the condition of independence is equivalent to

$$f(x, y) = f_X(x)f_Y(y) \tag{100}$$

where f_X and f_Y are the probability density functions of X and Y.

We can also define joint probability distributions for n random variables

$$F(a_1, a_2, \ldots, a_n) = P\{X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n\} \tag{101}$$

If these random variables are discrete, we define their joint probability mass function $p(x_1, x_2, \ldots, x_n)$ by

$$p(x_1, x_2, \ldots, x_n) = P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \tag{102}$$

Further, the n random variables are said to be jointly continuous if there exists a function $f(x_1, x_2, \ldots, x_n)$

$$P\{(X_1, \ldots, X_n) \in C\} = \int \int_{(x_1, \ldots, x_n) \in C} \ldots \int f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n \tag{103}$$

The concept of independence may, of course, also be defined for more than two random variables. In general, the n random variables $X_1, X_2, \ldots, X_n$ are said to be independent if, for all sets of real numbers $A_1, A_2, \ldots, A_n$,

$$P\{X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\} = \prod_{i=1}^{n} P\{X_i \in A_i\} \tag{104}$$

As before, it can be shown that this condition is equivalent to

$$P\{X_1 \leq a_1, X_2 \leq a_2, \ldots, X_n \leq a_n\} = \prod_{i=1}^{n} P\{X_i \leq a_i\} \tag{105}$$

for all $a_1, a_2, \ldots, a_n$.

Recalling the conditional probability of E given F, provided that $P(F) > 0$, is defined by

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{106}$$

Hence, if X and Y are discrete random variables, it is natural to define the conditional probability mass function of X given that $Y = y$, by

$$p_{X|Y}(x|y) = P\{X = x|Y = y\} \tag{107}$$

$$= \frac{P\{X = x, Y = y\}}{P\{Y = y\}} = \frac{p(x, y)}{p_Y(y)} \tag{108}$$

for all values of y such that $p_Y(y) > 0$.

If X and Y have a joint probability density function $f(x, y)$, then the conditional probability density function of X, given that $Y = y$, is defined for all values of y such that $f_Y(y) > 0$, by

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} \tag{109}$$

The use of conditional densities allows us to define conditional probabilities of events associated with one random variable when we are given the value of a second random variable. That is, if X and Y are jointly continuous, then, for any set A,

$$P\{X \in A|Y = y\} = \int_A f_{X|Y}(x|y)\, dx \tag{110}$$

The two-dimensional version states that if X and Y are random variables and g is a function of two variables, then

$$\textbf{Discrete case: } E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y) \tag{111}$$

$$\textbf{Continuous case: } E[g(X,Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)\,dx\,dy \tag{112}$$

In the continuous case, if $g(X,Y) = X + Y$ then

$$E[X+Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f(x,y)\,dx\,dy \tag{113}$$

If $g(X, Y) = X + Y$ then

$$E[X + Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x + y) f(x, y)\, dx\, dy \tag{114}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y)\, dx\, dy + \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y)\, dx\, dy \tag{115}$$

$$= E[X] + E[Y] \tag{116}$$

In general, for any n,

$$E[X_1 + X_2 + \cdots + X_n] = E[X_1] + E[X_2] + \cdots + E[X_n] \tag{117}$$

Variance of Sums of Random Variables

As we saw before, the expectation of a sum of random variables is equal to the sum of their expectations. The corresponding result for variances is, however, not generally valid. Consider

$$\begin{aligned} \mathrm{Var}(X+X) &= \mathrm{Var}(2X) \\ &= 2^2\,\mathrm{Var}(X) \\ &= 4\,\mathrm{Var}(X) \\ &\neq \mathrm{Var}(X)+\mathrm{Var}(X) \end{aligned} \tag{118}$$

There is, however, an important case in which the variance of a sum of random variables is equal to the sum of the variances occuring when the random variables are independent.

Covariance

Before proving this, however, let us define the concept of the covariance of two random variables.

The *covariance* of two random variables X and Y, written $\text{Cov}(X, Y)$, is defined by

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] \tag{119}$$

where μ_X and μ_Y are the means of X and Y, respectively.

Covariance & Correlation

- $\text{Cov}(X, Y) > 0$ indicates that Y tends to increase as X increases
- $\text{Cov}(X, Y) < 0$ indicates that Y tends to decrease as X increases

The strength of the relationship between X and Y is indicated by the *correlation* between X and Y, a dimensionless quantity obtained by dividing the covariance by the product of the standard deviations of X and Y. That is,

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}. \tag{120}$$

Covariance Properties

$$\begin{aligned}
\text{Cov}(X, Y) &= E[XY - \mu_x Y - \mu_y X + \mu_x \mu_y] && (121)\\
&= E[XY] - \mu_x E[Y] - \mu_y E[X] + \mu_x \mu_y && (122)\\
&= E[XY] - \mu_x \mu_y - \mu_y \mu_x + \mu_x \mu_y && (123)\\
&= E[XY] - E[X]E[Y] && (124)
\end{aligned}$$

$$\text{Cov}(X, X) = \text{Var}(X) \quad (125)$$

$$\text{Cov}(X, Y) = \text{Cov}(Y, X) \quad (126)$$

$$\text{Cov}(aX, Y) = a\text{Cov}(X, Y) \quad (127)$$

Covariance Properties

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n} \text{Cov}(X_i, Y) \tag{128}$$

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j) \tag{129}$$

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \text{Cov}(X_i, X_j) \tag{130}$$

Covariance Properties

If X and Y are independent random variables, then

$$\mathrm{Cov}(X, Y) = 0 \tag{131}$$

and so for independent $X_1, \ldots, X_n$,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i). \tag{132}$$

!!! Zero covariance does not imply independence !!!

Probability

Pause

Jointly Distributed Random Variables Exercises

- **Discrete Distributions**
  - Binomial Distribution
  - Negative Binomial Distribution
  - Geometric Distribution
  - Hypergeometric Distribution
  - Poisson Distribution
- **Continuous Distributions**
  - Uniform Distribution
  - Normal Distribution
  - Exponential Distribution
  - Chi-Square Distribution
  - t-Distribution
  - F-Distribution

Bernoulli Random Variable

Suppose that a trial, or an experiment, whose outcome can be classified as either a "success" or as a "failure" is performed. If we let $X = 1$ when the outcome is a success and $X = 0$ when it is a failure, then the probability mass function of X is given by

$$P\{X = 0\} = 1 - p \tag{133}$$

and

$$P\{X = 1\} = p \tag{134}$$

where $p, 0 \leq p \leq 1$, is the probability that the trial is a "success."

Bernoulli Random Variable

A random variable X is said to be a Bernoulli random variable if its probability mass function is given by previous equations for some $p \in [0, 1]$. Its expected value is

$$E[X] = \sum_i x_i P\{X = x_i\} \tag{135}$$

$$E[X] = 1 \cdot P\{X = 1\} + 0 \cdot P\{X = 0\} = p \tag{136}$$

That is, the expectation of a Bernoulli random variable is the probability that the random variable equals 1.

Binomial Random Variable

Suppose now that n independent trials, each of which results in a "success" with probability p and in a "failure" with probability $1 - p$, are to be performed. If X represents the number of successes that occur in the n trials, then X is said to be a *binomial random variable* with parameters (n, p).

The probability mass function of a binomial random variable with parameters n and p is given by

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n \tag{137}$$

Binomial Distribution Function ($X \sim B(n, p)$)

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n \tag{138}$$

- Expected Value

$$E[X] = np \tag{139}$$

- Variance

$$\mathrm{Var}(X) = npq = np(1-p) \tag{140}$$

Binomial Distribution Function ($X \sim B(n, p)$)

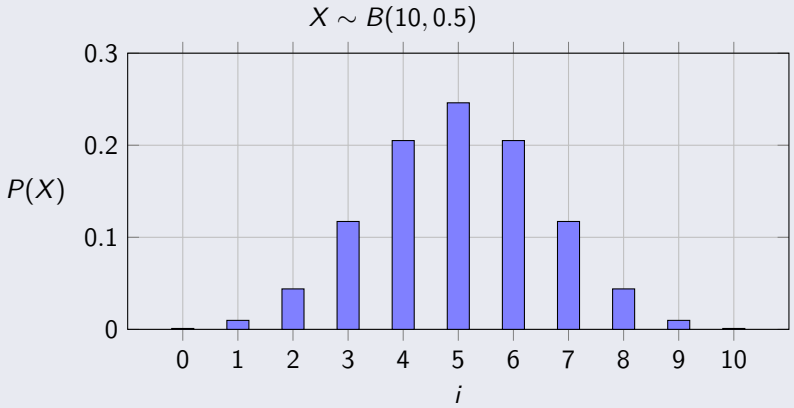
The probability mass function of a binomial random variable with parameters n and p is given by

$$P\{X = i\} = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0, 1, \ldots, n \tag{141}$$

The cumulative probability function for a binomial random variable is given by

$$P\{X \leq i\} = \sum_{k=0}^{i} \binom{n}{k} p^k (1-p)^{n-k}, \quad i = 0, 1, \ldots, n. \tag{142}$$

Binomial Distribution ($X \sim B(n, p)$)



Negative Binomial Distribution ($X \sim NB(r, p)$)

If X represents the number of trials required to achieve r successed in a series of independent Bernoulli trials, each with a constant probability of success, p, then X is said to be a *negative binomial random variable* with parameters (r, p).

$$P\{X = i\} = \binom{i + r - 1}{i} p^r (1 - p)^i \tag{143}$$

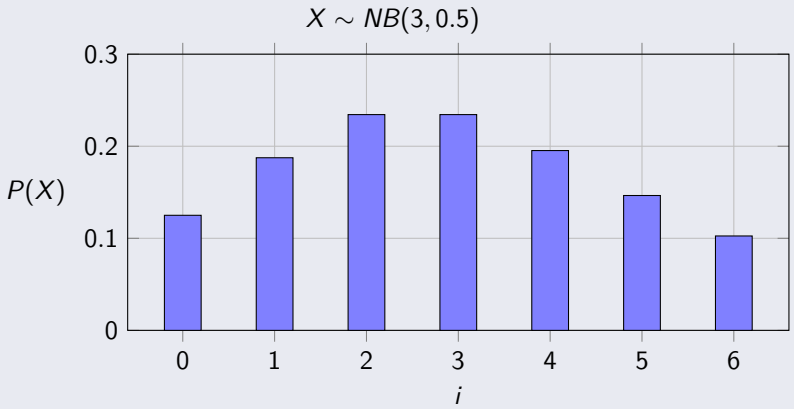
- Expected Value

$$E[X] = \frac{r(1 - p)}{p} \tag{144}$$

- Variance

$$\text{Var}(X) = \frac{r(1 - p)}{p^2} \tag{145}$$

Negative Binomial Distribution ($X \sim NB(r, p)$)



Geometric Distribution ($X \sim G(p)$)

A *geometric random variable* represents the number of trials in a sequence of draws to get one success. It is governed by the geometric distribution, a discrete probability distribution that models the probability that the first occurence of success required i independent trials, each with success probability, p.

$$P\{X = i\} = (1-p)^{i-1}p, \quad i = 0, 1, \ldots \tag{146}$$

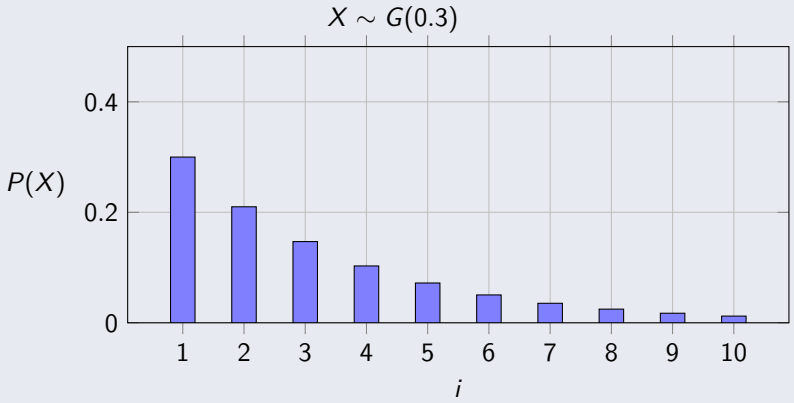
- Expected Value

$$E[X] = \frac{1}{p} \tag{147}$$

- Variance

$$\text{Var}(X) = \frac{1-p}{p^2} \tag{148}$$

Geometric Distribution ($X \sim G(p)$)



Hypergeometric Distribution ($X \sim H(N, M, n)$)

A *hypergeometric random variable* represents the number of successes in a sequence of draws from a finite population without replacement. It is governed by the hypergeometric distribution, a discrete probability distribution that models scenarios where the probability of success changes with each draw due to the lack of replacement.

- Population Size ($N + M$): The total number of items in the population.
- Number of Successes in Population (N): The count of items classified as "successes."
- Sample Size (n): The number of items drawn from the population.
- Number of Observed Successes (i): The count of successes in the sample.

Hypergeometric Distribution ($X \sim H(N, M, n)$)

$$P\{X = i\} = \frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}}, \quad i = 0, 1, \ldots, \min(N, n) \tag{149}$$

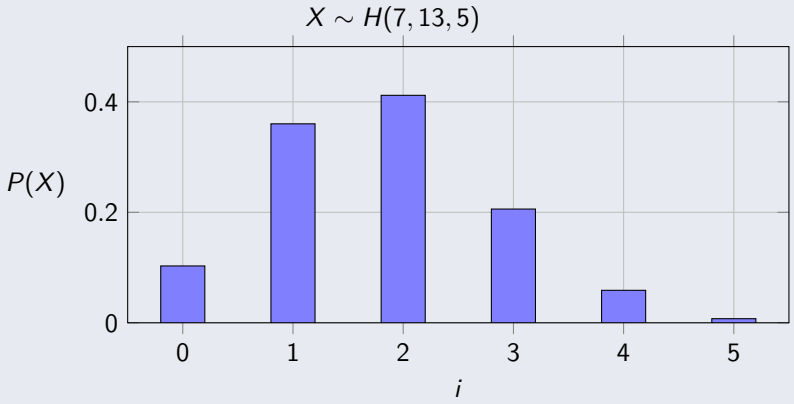
- Expected Value

$$E[X] = \frac{nN}{N+M} = np \tag{150}$$

- Variance

$$\mathrm{Var}(X) = np(1-p)\left[1 - \frac{n-1}{N+M-1}\right] \tag{151}$$

Hypergeometric Distribution ($X \sim H(N, M, n)$)



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

The Poisson distribution is a discrete probability distribution that models the likelihood of a given number of events occurring within a fixed interval of time, space, or other dimensions, provided that these events happen independently and at a constant average rate.

A random variable X, taking on one of the values $0, 1, 2, \ldots$, is said to be a Poisson random variable with parameter $\lambda, \lambda > 0$, if its probability mass function is given by

$$P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, \ldots \tag{152}$$

with $e \approx 2.7183$.

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

$$P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, \ldots \tag{153}$$

$$P\{X \leq i\} = \sum_{k=0}^{i} e^{-\lambda}\frac{\lambda^k}{k!} \tag{154}$$

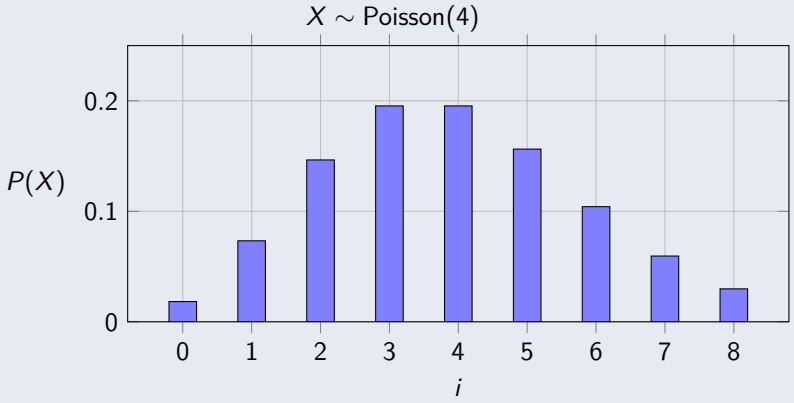
- Expected Value

$$E[X] = \lambda \tag{155}$$

- Variance

$$\text{Var}(X) = E[X] = \lambda \tag{156}$$

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



Probability

Pause

Discrete Distributions Exercises

- **Discrete Distributions**
  - Binomial Distribution
  - Negative Binomial Distribution
  - Geometric Distribution
  - Hypergeometric Distribution
  - Poisson Distribution
- **Continuous Distributions**
  - Uniform Distribution
  - Normal Distribution
  - Exponential Distribution
  - Chi-Square Distribution
  - t-Distribution
  - F-Distribution

Uniform Distribution ($X \sim U(\alpha, \beta)$)

A random variable is said to follow a *uniform distribution* with parameters α and β if its density is

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } x \in [\alpha, \beta] \\ 0 & \text{otherwise} \end{cases} \tag{157}$$

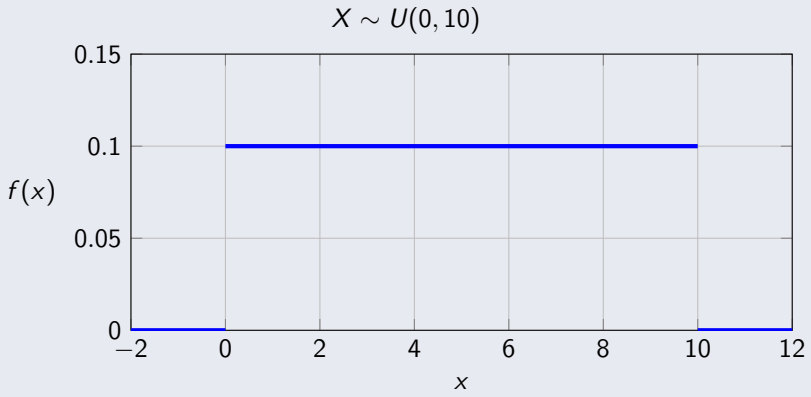
- Expected Value

$$E[X] = \frac{\alpha + \beta}{2} \tag{158}$$

- Variance

$$\text{Var}(X) = \frac{(\beta - \alpha)^2}{12} \tag{159}$$

Uniform Distribution ($X \sim U(\alpha, \beta)$)



Normal Distribution ($X \sim \mathcal{N}(\mu, \sigma^2)$)

A random variable is said to follow a *normal distribution* with parameters μ and σ^2 if its density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{160}$$

- Expected Value

$$E[X] = \mu \tag{161}$$

- Variance

$$\text{Var}(X) = \sigma^2 \tag{162}$$

Standard Normal Distribution

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

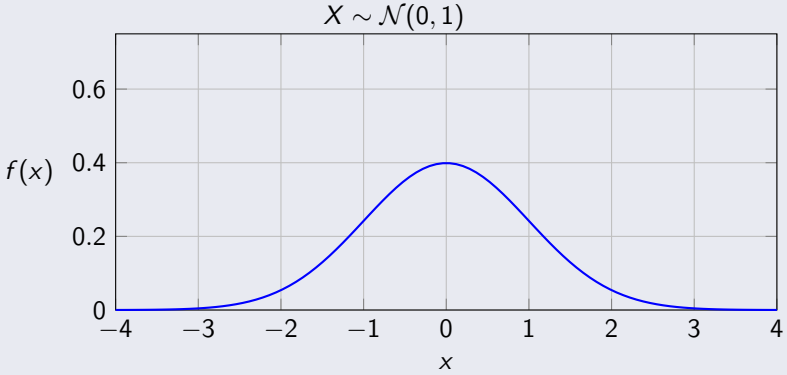
$$Z = \frac{X - \mu}{\sigma} \tag{163}$$

is a normal random variable with mean 0 and variance 1. Such a random variable Z is said to have a standard, or unit, normal distribution.

Its distribution function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \tag{164}$$

Normal Distribution ($X \sim \mathcal{N}(\mu, \sigma^2)$)



Exponential Distribution ($X \sim \text{Exp}(\lambda)$)

A random variable is said to follow an *exponential distribution* with parameter λ if its density is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{165}$$

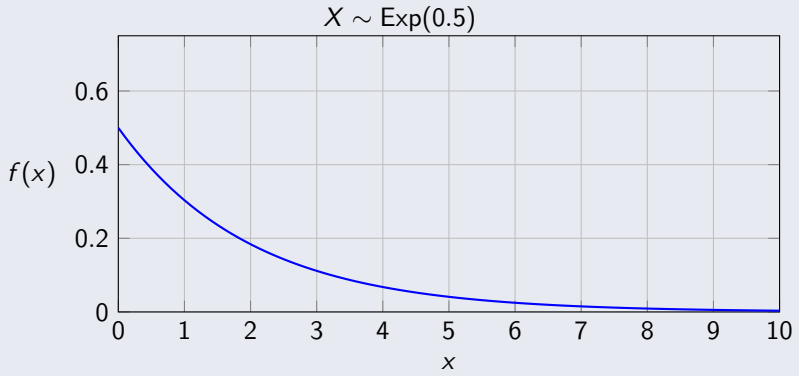
- Expected Value

$$E[X] = \frac{1}{\lambda} \tag{166}$$

- Variance

$$\text{Var}(X) = \frac{1}{\lambda^2} \tag{167}$$

Exponential Distribution ($X \sim \text{Exp}(\lambda)$)



Chi-Square Distribution ($X \sim \chi^2_n$)

If $Z_1, Z_2, \ldots, Z_n$ are independent standard normal random variables, then X, defined by

$$X = Z_1^2 + Z_2^2 + \cdots + Z_n^2 \tag{168}$$

is said to have a *chi-square distribution with n degrees of freedom*.

$$X \sim \chi^2_n \tag{169}$$

Chi-Square Distribution ($X \sim \chi^2_n$)

If X is a chi-square random variable with n degrees of freedom, then for any $\alpha \in [0, 1]$, the quantity $\chi^2_{\alpha,n}$ is defined to be such that

$$P(X \geq \chi^2_{\alpha,n}) = \alpha \tag{170}$$

- Expected Value

$$E[X] = n \tag{171}$$

- Variance

$$\text{Var}(X) = 2n \tag{172}$$

t-Distribution ($T_n \sim t_n$)

If Z and χ^2_n are independent random variables, with Z having a standard normal distribution and χ^2_n having a chi-square distribution with n degrees of freedom, then the random variable T_n defined by

$$T_n = \frac{Z}{\sqrt{\chi^2_n / n}} \tag{173}$$

is said to have a *t-distribution with n degrees of freedom* with

$$\frac{\chi^2_n}{n} = \frac{Z_1^2 + Z_2^2 + \cdots + Z_n^2}{n} \tag{174}$$

t-Distribution ($T_n \sim t_n$)

For α, $0 < \alpha < 1$, let $t_{\alpha,n}$ be such that

$$P(T_n \geq t_{\alpha,n}) = \alpha \tag{175}$$

- Expected Value

$$E[T_n] = 0 \tag{176}$$

- Variance

$$\text{Var}(T_n) = \frac{n}{n-2} \tag{177}$$

F-Distribution ($F_{n,m} \sim F(n, m)$)

If χ^2_n and χ^2_m are independent chi-square random variables with n and m degrees of freedom, respectively, then the random variable $F_{n,m}$ defined by

$$F_{n,m} = \frac{\chi^2_n / n}{\chi^2_m / m} \tag{178}$$

is said to have an *F-distribution with n and m degrees of freedom*.

F-Distribution ($X \sim F_{n,m}$)

For any $\alpha \in [0, 1]$, let $F_{\alpha,n,m}$ be such that

$$P(X \geq F_{\alpha,n,m}) = \alpha \tag{179}$$

- Expected Value

$$E[X] = \frac{m}{m-2} \tag{180}$$

- Variance

$$\text{Var}(X) = \frac{2m(n+m-2)}{n(m-2)^2(m-4)} \tag{181}$$

Probability

Pause

Continuous Distributions Exercises

If $X_1, \cdots, X_n$ are independent random variables having a common distribution F, then we say that they constitute a *sample* (sometimes called a random sample) from the distribution F.

The Sample Mean

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \tag{182}$$

- Expected Value

$$\begin{aligned} E[\overline{X}] &= E\left[\frac{X_1 + \cdots + X_n}{n}\right] \\ &= \frac{1}{n}(E[X_1] + \cdots + E[X_n]) = \mu \end{aligned} \tag{183}$$

The Central Limit Theorem is a fundamental concept in probability and statistics. It states that, under certain conditions, the sampling distribution of the sample mean will approximate a normal distribution (bell-shaped curve) as the sample size becomes sufficiently large, regardless of the original population's distribution.

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables each having mean μ and variance σ^2. Then for n large, the distribution of

$$X_1 + X_2 + \cdots + X_n \tag{184}$$

is approximately normal with mean $n\mu$ and variance $n\sigma^2$.

The Central Limit Theorem

It follows from the central limit theorem that

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \tag{185}$$

is approximately a standard normal random variable. Hence, for n large,

$$P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} < x\right\} \approx P(Z < x), \tag{186}$$

where Z is a standard normal random variable.

Approximate Distribution of the Sample Mean

Let $X_1, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2. The central limit theorem can be used to approximate the distribution of the sample mean

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} = \sum_{i=1}^{n} X_i/n \tag{187}$$

From the central limit theorem, $\overline{X}$ will be approximately normal for a large enough n. Since the sample mean has expected value μ and standard deviation $\sigma/\sqrt{n}$, it then follows that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{188}$$

Is n large enough?

- If the underlying population distribution is normal, then the sample mean $\overline{X}$ will also be normal regardless of the sample size.
- A general rule of thumb is that one can be confident of the normal approximation whenever the sample size n is at least 30.
- That is, practically speaking, no matter how nonnormal the underlying population distribution is, the sample mean of a sample of size at least 30 will be approximately normal.

The Sample Variance

Let $X_1, \ldots, X_n$ be a sample from a population having mean μ and variance σ^2. Let $\overline{X}$ be the sample mean, the statistic S^2, defined by

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \tag{189}$$

is called the *sample variance*. $S = \sqrt{S^2}$ is called the *sample standard deviation*.

- Expected value of S^2

$$E[S^2] = \sigma^2 \tag{190}$$

Sample Mean Distribution

Since the sum of independent normal random variables is normally distributed, it follows that X is normal with mean

$$E[\overline{X}] = \mu \tag{191}$$

and variance

$$\text{Var}(\overline{X}) = \sigma^2/n \tag{192}$$

That is, $\overline{X}$, the average of the sample, is normal with a mean equal to the population mean but with a variance reduced by a factor of $1/n$. It follows that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \tag{193}$$

Joint Distribution of $\overline{X}$ and S^2

If $X_1, \ldots, X_n$ is a sample from a normal population having mean μ and variance σ^2, then $\overline{X}$ and S^2 are independent random variables, with

$$\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n) \tag{194}$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{195}$$

If $\overline{X}$ denotes the sample mean and S the sample standard deviation, then

$$\sqrt{n}\frac{\overline{X} - \mu}{S} \sim t_{n-1} \tag{196}$$

Probability

End

Continuous Distributions Exercises

Table of Contents

In statistics, an estimator is a rule or formula used to estimate a population parameter based on sample data. A "good" estimator possesses certain desirable properties that ensure its reliability and accuracy in estimating the true parameter.

The basic properties of estimators are related to notions of accuracy and precision, similar to the characterization of experimental methods for measuring an unknown quantity in terms of the agreement of repeated measurements obtained, where:

- *Accuracy* - agreement of observations with the target value.
- *Precision* - agreement of observations with each other.

Accuracy is associated with systematic errors, for instance, deficiencies in measurement instruments, while precision refers to random errors that are responsible for small, unpredictable variations in the measurements made, whose causes are not fully understood.

Any statistic used to estimate the value of an unknown parameter θ is called an *estimator* of θ. The observed value of the estimator is called the *estimate*. For instance, as we shall see, the usual estimator of the mean of a normal population, based on a sample $X_1, \ldots, X_n$ from that population, is the sample mean $\overline{X}$.

Maximum Likelihood Estimator

The *likelihood function* is defined as:

$$\mathcal{L}(\theta|\mathbf{x}) = f(\mathbf{x}|\theta), \tag{197}$$

where $f(\mathbf{x}|\theta)$ is the probability density (or mass) function of the data $\mathbf{x}$, parameterized by θ.

Maximum Likelihood Estimator

For independent and identically distributed samples, the likelihood function becomes

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta). \tag{198}$$

The maximum likelihood estimate (MLE) is obtained by maximizing the likelihood function

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta|\mathbf{x}), \tag{199}$$

or equivalently, by maximizing the log-likelihood

$$\ell(\theta|\mathbf{x}) = \log \mathcal{L}(\theta|\mathbf{x}), \tag{200}$$

Evaluating an Estimator

Let $\hat{\Theta} = \hat{\Theta}(X_1, \ldots, X_n)$ be an estimator of the parameter θ. One way to determine the "worth" of the estimator $\hat{\Theta}$ is the *Mean Squared Error* (MSE), given by:

$$\text{MSE}(\hat{\Theta}) \equiv E((\hat{\Theta} - \theta)^2) = \underbrace{(E(\hat{\Theta}) - \theta)}_{\text{Bias of } \theta}^2 + \text{Var}(\hat{\Theta}) \tag{201}$$

Let $\hat{\Theta}_1 = \hat{\Theta}_1(X_1, \ldots, X_n)$ and $\hat{\Theta}_2 = \hat{\Theta}_2(X_1, \ldots, X_n)$ be two estimators of the parameter θ. It is said that $\hat{\Theta}_1$ is better than $\hat{\Theta}_2$ when

$$\text{MSE}(\hat{\Theta}_1) < \text{MSE}(\hat{\Theta}_2) \tag{202}$$

Inferential Statistics

Pause

Pontual Estimation Exercises

Suppose that $X_1, \ldots, X_n$ is a sample from a normal population having unknown mean μ and known variance σ^2. It has been shown that $\overline{X}$ is the maximum likelihood estimator for μ. However, we do not expect that the sample mean $\overline{X}$ will exactly equal μ, but rather that it will "be close."

Point Estimates $\rightarrow$ Interval Estimates

In the foregoing, since the point estimator $\overline{X}$ is normal with mean μ and variance σ^2/n, it follows that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n}\frac{(\overline{X} - \mu)}{\sigma} \tag{203}$$

has a standard normal distribution.

Considering

$$\sqrt{n}\frac{(\overline{X}-\mu)}{\sigma} \sim \mathcal{N}(0,1) \tag{204}$$

then,

$$P\left\{-z_{\alpha/2} < \sqrt{n}\frac{(\overline{X}-\mu)}{\sigma} < z_{\alpha/2}\right\} = 1-\alpha \tag{205}$$

$$P\left\{\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha \tag{206}$$

This means that $(1-\alpha)$ percent of the time the value of the sample average $\overline{X}$ will be such that the distance between it and the mean μ will be less than $z_{\alpha/2}\sigma/\sqrt{n}$.

Based on

$$P\left\{\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha, \tag{207}$$

the interval

$$\left[\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{208}$$

is called a $100(1-\alpha)$ *percent confidence interval estimate* of μ where $\overline{x}$ is the observed sample mean.

This is a *two-sided confidence interval* but there can be *one-sided upper* or *lower confidence intervals* for μ.

Suppose now that $X_1, \ldots, X_n$ is a sample from a normal distribution with unknown mean μ and unknown variance σ^2. Since σ is unknown, we can no longer base our interval on the fact that $\sqrt{n}(\overline{X} - \mu)/\sigma$ is a standard normal random variable. However, by using the sample variance, S^2, then it follows that

$$\sqrt{n}\frac{(\overline{X} - \mu)}{S} \tag{209}$$

is a t-random variable with n degrees of freedom. From the symmetry of the t-density function we have that for any $\alpha \in [0, 1/2]$,

$$P\left\{-t_{\alpha/2,n-1} < \sqrt{n}\frac{(\overline{X} - \mu)}{S} < t_{\alpha/2,n-1}\right\} = 1 - \alpha \tag{210}$$

Based on

$$P\left\{\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right\} = 1-\alpha, \tag{211}$$

the interval

$$\left[\overline{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] \tag{212}$$

is called a $100(1-\alpha)$ *percent confidence interval estimate* of μ where $\overline{x}$ is the observed sample mean.

This is a *two-sided confidence interval* but there can be *one-sided upper* or *lower confidence intervals* for μ.

If $X_1, \ldots, X_n$ is a sample from a normal distribution having unknown parameters μ and σ^2, then we can construct a confidence interval for σ^2 by using the fact that

$$(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{213}$$

Hence,

$$P\left\{\chi^2_{1-\alpha/2,n-1} \leq (n-1)\frac{S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha \tag{214}$$

which can be further simplified to

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\} = 1-\alpha \tag{215}$$

Based on

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\} = 1-\alpha \tag{216}$$

the interval

$$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right] \tag{217}$$

is called a $100(1-\alpha)$ *percent confidence interval estimate* of σ^2 where S^2 is the sample variance.

This is a *two-sided confidence interval* but there can be *one-sided upper* or *lower confidence intervals* for σ^2.

When sampling from normal populations

- Confidence interval for μ having unknown mean μ and known variance σ^2

Confidence Interval	Lower Interval	Upper Interval
$\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\left[-\infty, \overline{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}\right]$	$\left[\overline{X} + z_{\alpha} \frac{\sigma}{\sqrt{n}}, \infty\right]$

- Confidence interval for μ having unknown mean μ and unknown variance σ^2

Confidence Interval	Lower Interval	Upper Interval
$\overline{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$	$\left[-\infty, \overline{X} + t_{\alpha,n-1} \frac{S}{\sqrt{n}}\right]$	$\left[\overline{X} - t_{\alpha,n-1} \frac{S}{\sqrt{n}}, \infty\right]$

- Confidence interval for σ^2 having unknown mean μ and known variance σ^2

Confidence Interval	Lower Interval	Upper Interval
$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right]$	$\left[0, \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right]$	$\left[\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, \infty\right]$

Let $X_1, X_2, \ldots, X_n$ be a sample of size n from a normal population having mean μ_1 and variance σ_1^2 and let $Y_1, \ldots, Y_m$ be a sample of size m from a different normal population having mean μ_2 and variance σ_2^2 and suppose that the two samples are independent of each other.

$$\overline{X} \sim \mathcal{N}(\mu_1, \sigma_1^2/n) \tag{218}$$

$$\overline{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2/m) \tag{219}$$

We are interested in estimating $\mu_1 - \mu_2$. Since $\overline{X}$ and $\overline{Y}$ are the maximum likelihood estimators of μ_1 and μ_2, it can be proved that $\overline{X} - \overline{Y}$ is the maximum likelihood estimator of $\mu_1 - \mu_2$.

To obtain a confidence interval estimator, we need the distribution of $\overline{X} - \overline{Y}$.
Considering that

$$\overline{X} \sim \mathcal{N}(\mu_1, \sigma_1^2/n) \tag{220}$$

$$\overline{Y} \sim \mathcal{N}(\mu_2, \sigma_2^2/m) \tag{221}$$

it follows from the fact that the sum of independent normal random variables is also normal, that

$$\overline{X} - \overline{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right) \tag{222}$$

Hence, assuming σ_1^2 and σ_2^2 are known, we have that

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim \mathcal{N}(0,1) \tag{223}$$

Based on this, the confidence interval estimate can be extracted from

$$P\left\{ -z_{\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} < z_{\alpha/2} \right\} = 1 - \alpha \tag{224}$$

When assuming that σ_1^2 and σ_2^2 are unknown but equal, we have that

$$\frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}} \sim t_{n+m-2} \tag{225}$$

Based on this, the confidence interval estimate can be extracted from

$$P\left\{ -t_{\alpha/2,n+m-2} < \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}} < t_{\alpha/2,n+m-2} \right\} = 1 - \alpha \tag{226}$$

$100(1-\alpha)$ Two-Sided Confidence Interval Estimate for $\mu_1 - \mu_2$

- **Assumption:** σ_1, σ_2 known

$$\bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \tag{227}$$

- **Assumption:** σ_1, σ_2 unknown but equal

$$\bar{X} - \bar{Y} \pm t_{\alpha/2,n+m-2}\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}} \tag{228}$$

$100(1-\alpha)$ Lower Confidence Interval Estimate for $\mu_1 - \mu_2$

- **Assumption:** σ_1, σ_2 known

$$\left]-\infty, \bar{X} - \bar{Y} + z_\alpha \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right] \tag{229}$$

- **Assumption:** σ_1, σ_2 unknown but equal

$$\left]-\infty, \bar{X} - \bar{Y} + t_{\alpha, n+m-2} \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right) \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}\right] \tag{230}$$

Upper confidence intervals for $\mu_1 - \mu_2$ are obtained from lower confidence intervals for $\mu_2 - \mu_1$

Inferential Statistics

Pause

Interval Estimation Exercises

Consider a population having distribution F_θ, where θ is unknown, and suppose we want to test a specific hypothesis about θ.We shall denote this hypothesis by H_0 and call it the *null hypothesis*. Suppose now that in order to test a specific null hypothesis H_0, a population sample of size n is to be observed. Based on these n values, we must decide whether or not to accept H_0. A test for H_0 can be specified by defining a region C in n-dimensional space with the requirements

$$\text{accept } H_0 \text{ if } (X_1, X_2, \ldots, X_n) \notin C \tag{231}$$

and

$$\text{reject } H_0 \text{ if } (X_1, X_2, \ldots, X_n) \in C \tag{232}$$

with region C being called the *critical region*.

Tests Concerning the Mean of a Normal Population

Suppose that $X_1, \ldots, X_n$ is a sample of size n from a normal distribution having an unknown mean μ and a known variance σ^2 and suppose we are interested in testing the null hypothesis

$$H_0 : \mu = \mu_0 \tag{233}$$

against the alternative hypothesis

$$H_a : \mu \neq \mu_0 \tag{234}$$

where μ_0 is some specified constant.

Since $\overline{X}$ is a natural point estimator of μ, it seems reasonable to accept H_0 if $\overline{X}$ is not too far from μ_0. That is, the critical region of the test would be of the form

$$C = \{X_1, \ldots, X_n : |\bar{X} - \mu_0| > c\} \tag{235}$$

for some suitably chosen value c. If we desire that the test has significance level α, then c must be such that

$$P_{\mu_0}\{|\bar{X} - \mu_0| > c\} = \alpha \tag{236}$$

where we write P_{μ_0} to mean that the preceding probability is to be computed under the assumption that $\mu = \mu_0$. However, when $\mu = \mu_0$, $\overline{X}$ will be normally distributed with mean μ_0 and variance σ^2/n.

Hypothesis Testing Steps

1. Establish hypotheses
   - Null Hypothesis (H_0): The default assumption, often stating no effect or no difference.
   - Alternative Hypothesis (H_a): The claim being tested, which contradicts the null hypothesis.
2. Choose a significance level
3. Calculate a test statistic
4. Compare the test statistic to a critical value
5. Make a decision
   - Reject H_0 if the evidence supports H_a (*type I error* can occur).
   - Fail to reject H_0 if the evidence is insufficient (*type II error* can occur).

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma^2)$ population with known σ^2

- Test Statistic TS

$$TS = \sqrt{n}\frac{\overline{X} - \mu_0}{\sigma} \tag{237}$$

H_0	H_a	Significance Level α	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	Reject if $\lvert TS \rvert > z_{\alpha/2}$	$2P\{Z \geq \lvert t \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	Reject if $TS > z_\alpha$	$P\{Z \geq t\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	Reject if $TS < -z_\alpha$	$P\{Z \leq t\}$

Z is a standard normal random variable

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma^2)$ population with unknown σ^2

- Test Statistic TS

$$TS = \sqrt{n}\frac{\overline{X} - \mu_0}{S} \tag{238}$$

H_0	H_a	Significance Level α	p-Value if $TS = t$
$\mu = \mu_0$	$\mu \neq \mu_0$	Reject if $\lvert TS \rvert > t_{\alpha/2,n-1}$	$2P\{T_{n-1} \geq \lvert t \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	Reject if $TS > t_{\alpha,n-1}$	$P\{T_{n-1} \geq t\}$
$\mu \geq \mu_0$	$\mu < \mu_0$	Reject if $TS < -t_{\alpha,n-1}$	$P\{T_{n-1} \leq t\}$

T_{n-1} is a t-random variable with $n-1$ degrees of freedom

Tests Concerning the Equality of Means of Two Normal Populations

Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are independent samples from normal populations having unknown means μ_x and μ_y but known variances σ_x^2 and σ_y^2.

Let us consider the problem of testing the hypothesis

$$H_0 : \mu_x = \mu_y \tag{239}$$

against the alternative hypothesis

$$H_a : \mu_x \neq \mu_y \tag{240}$$

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma_1^2)$ population

$Y_1, \ldots, Y_m$ is a sample from a $\mathcal{N}(\mu, \sigma_2^2)$ population

$$H_0 : \mu_x = \mu_y \text{ versus } H_a : \mu_x \neq \mu_y$$

- Assuming known σ_1, σ_2, the test Statistic TS is

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \tag{241}$$

Significance Level α	p-Value if $TS = t$
Reject if $\lvert TS \rvert > z_{\alpha/2}$	$2P\{Z \geq \lvert t \rvert\}$

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma_1^2)$ population

$Y_1, \ldots, Y_m$ is a sample from a $\mathcal{N}(\mu, \sigma_2^2)$ population

$$H_0 : \mu_x = \mu_y \text{ versus } H_a : \mu_x \neq \mu_y$$

- Assuming $\sigma_1 = \sigma_2$, the test Statistic TS is

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}}\sqrt{\frac{1}{n} + \frac{1}{m}}} \tag{242}$$

Significance Level α	p-Value if $TS = t$
Reject if $\lvert TS\rvert > t_{\alpha/2, n+m-2}$	$2P\{T_{n+m-2} \geq \lvert t\rvert\}$

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma_1^2)$ population

$Y_1, \ldots, Y_m$ is a sample from a $\mathcal{N}(\mu, \sigma_2^2)$ population

$$H_0 : \mu_x = \mu_y \text{ versus } H_a : \mu_x \neq \mu_y$$

- Assuming n, m large, the test Statistic TS is

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_1^2}{n}} + \sqrt{\frac{S_2^2}{m}}} \tag{243}$$

Significance Level α	p-Value if $TS = t$
Reject if $\lvert TS \rvert > z_{\alpha/2}$	$2P\{Z \geq \lvert t \rvert\}$

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma_1^2)$ population

$Y_1, \ldots, Y_m$ is a sample from a $\mathcal{N}(\mu, \sigma_2^2)$ population

$$H_0 : \mu_x = \mu_y \text{ versus } H_a : \mu_x \neq \mu_y$$

- When n, m are small, the test Statistic TS is

$$TS = \frac{\overline{X} - \overline{Y}}{\sqrt{S_1^2/n} + \sqrt{S_2^2/m}} \tag{244}$$

Significance Level α	p-Value if $TS = t$
Reject if $\lvert TS\rvert > t_{\alpha/2,df}$	$2P\{T_{df} \geq \lvert t\rvert\}$

$$df = \frac{\left(S_1^2/n + S_2^2/m\right)^2}{\frac{\left(S_1^2/n\right)^2}{n-1} + \frac{\left(S_2^2/m\right)^2}{m-1}}$$

The Paired t-Test

Considering now that data can be described by the n pairs $(X_i, Y_i), i = 1, \ldots, n$ and that $W_i = X_i - Y_i, i = 1, \ldots, n$.

Then one could test the hypothesis of no effect by testing

$$H_0 : \mu_w = 0 \text{ versus } H_a : \mu_w \neq 0$$

where $W_1, \ldots, W_n$ are assumed to be a sample from a normal population having unknown mean μ_w and unknown variance σ_w^2.

$W_1, \ldots, W_n$ is a sample from a $\mathcal{N}(\mu_w, \sigma_w^2)$ population with unknown σ_w^2

$$H_0 : \mu_w = 0 \text{ versus } H_a : \mu_w \neq 0$$

- Test Statistic TS

$$TS = \sqrt{n}\frac{\overline{W}}{S_w} \tag{245}$$

H_0	H_a	Significance Level α	p-Value if $TS = t$
$\mu_w = 0$	$\mu_w \neq 0$	Reject if $\lvert TS \rvert > t_{\alpha/2,n-1}$	$2P\{T_{n-1} \geq \lvert t \rvert\}$

Tests Concerning the Variance of a Normal Population

Let $X_1, \ldots, X_n$ denote a sample from a normal population having unknown mean μ and unknown variance σ^2, and suppose we desire to test the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2$$

versus

$$H_a : \sigma^2 \neq \sigma_0^2$$

for a given σ_0^2.

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu, \sigma^2)$ population with unknown μ and σ^2

$$H_0 : \sigma^2 = \sigma_0^2 \text{ versus } H_a : \sigma^2 \neq \sigma_0^2$$

- Test Statistic TS

$$TS = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \tag{246}$$

Accept H_0 if

$$\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma_0^2} \leq \chi^2_{\alpha/2,n-1}$$

with the test data having a $p-$value of

$$p = 2\min(P\{\chi^2_{n-1} < TS\}, 1 - P\{\chi^2_{n-1} < TS\}) \tag{247}$$

Tests Concerning the Equality of Variances of Two Normal Populations

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ denote independent samples from two normal populations having respective (unknown) parameters μ_x, σ_x^2 and μ_y, σ_y^2 and consider a test of

$$H_0 : \sigma_x^2 = \sigma_y^2$$

versus

$$H_a : \sigma_x^2 \neq \sigma_y^2$$

$X_1, \ldots, X_n$ is a sample from a $\mathcal{N}(\mu_x, \sigma_x^2)$ population with unknown μ_x and σ_x^2

$Y_1, \ldots, Y_m$ is a sample from a $\mathcal{N}(\mu_y, \sigma_y^2)$ population with unknown μ_y and σ_y^2

$$H_0 : \sigma_x^2 = \sigma_y^2 \text{ versus } H_a : \sigma_x^2 \neq \sigma_y^2$$

- Test Statistic TS

$$TS = S_x^2/S_y^2 \sim F_{n-1,m-1} \tag{248}$$

Accept H_0 if

$$F_{1-\alpha/2,n-1,m-1} \leq S_x^2/S_y^2 \leq F_{\alpha/2,n-1,m-1}$$

with the test data having a $p-$value of

$$p = 2\min(P\{F_{n-1,m-1} < TS\}, 1 - P\{F_{n-1,m-1} < TS\}) \tag{249}$$

Suppose that we have n independent samples, each with the probability p of being a "success". Then, the random variable X, representing the number of "successes" in a sample of n items, will follow a binomial distribution with parameters (n, p).

$$H_0 : p = p_0 \text{ versus } H_a : p \neq p_0$$

- Assuming n large, X will follow approximately $\mathcal{N}(np, np(1-p))$
- Test Statistic TS

$$TS = \frac{X - np_0}{\sqrt{np_0(1-p_0)}} = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim\approx \mathcal{N}(0,1) \tag{250}$$

Inferential Statistics

Pause

Hypothesis Testing Exercises

A *nonparametric test* is a hypothesis test where it is not necessary (or not possible) to specify the parametric form of the distribution(s) of the underlying population(s)

χ^2 Test (Chi-Square Goodness of Fit Test)

Suppose that n independent random variables — $Y_1, \ldots, Y_n$, each taking on one of the values $1, 2, \ldots, k$ — are to be observed and we are interested in testing the null hypothesis that $\{p_i, i = 1, \ldots, k\}$ is the probability mass function of the Y_j. That is, if Y represents any of the Y_j, then the null hypothesis is

$$H_0 : P\{Y = i\} = p_i, \quad i = 1, \ldots, k, \tag{251}$$

whereas the alternative hypothesis is

$$H_a : P\{Y = i\} \neq p_i, \quad \text{for some } i = 1, \ldots, k. \tag{252}$$

To test the foregoing hypothesis, let X_i, $i = 1, \ldots, k$, denote the number of the Y_j's that equal i. Then as each Y_j will independently equal i with probability $P\{Y = i\}$, it follows that, under H_0, X_i is binomial with parameters n and p_i.

Hence, when H_0 is true,

$$E[X_i] = np_i \tag{253}$$

and so $(X_i - np_i)^2$ will be an indication as to how likely it appears that p_i indeed equals the probability that $Y = i$.

When this is large, say, in relationship to np_i, then it is an indication that H_0 is not correct.

Indeed such reasoning leads us to consider the following test statistic:

$$T = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i}, \tag{254}$$

and to reject the null hypothesis when T is large. The critical region, at a significance level α, is linked to a critical value c such that

$$P_{H_0}\{T \geq c\} = \alpha \tag{255}$$

That is, we need to determine c so that the probability that the test statistic T is at least as large as c, when H_0 is true, is α. The test is then to reject the hypothesis, at the α level of significance, when $T \geq c$ and to accept when $T < c$.

Assuming that H_0 is true, for a n large, T will have a chi-square distribution with $k-1$ degrees of freedom. Hence, for n large ($n \geq 30 \wedge np_i \geq 5$), c can be taken to equal $\chi^2_{\alpha,k-1}$, and so the approximate α-level test is

$$\text{reject } H_0 \text{ if } T \geq \chi^2_{\alpha,k-1} \text{ or accept } H_0 \text{ otherwise}$$

If the observed value of T is $T = t$, then the preceding test is equivalent to rejecting H_0 if the significance level α is at least as large as the p-value given by

$$p\text{-value} = P_{H_0}\{T \geq t\} \tag{256}$$
$$\approx P(\chi^2_{k-1} \geq t), \tag{257}$$

where χ^2_{k-1} is a chi-square random variable with $k-1$ degrees of freedom.

χ^2 Test with Unspecified Parameters

We can also perform goodness of fit tests of a null hypothesis that does not completely specify the probabilities $\{p_i, i = 1, \ldots, k\}$. The test statistic is now defined as

$$T = \sum_{i=1}^{k} \frac{(X_i - n\hat{p}_i)^2}{n\hat{p}_i}, \tag{258}$$

where $\hat{p}_i$ is now an estimated probability. It can then be proven that when n is large, the test statistic T will have, when H_0 is true, approximately a chi-square distribution with $k - 1 - m$ degrees of freedom. In this context, m is the number of unspecified parameters and that they are to be estimated by the method of maximum likelihood.

χ^2 Test with Unspecified Parameters

The test is, therefore, to

$$\text{reject } H_0 \text{ if } T \geq \chi^2_{\alpha,k-1-m} \tag{259}$$

$$\text{accept } H_0 \text{ otherwise} \tag{260}$$

An equivalent way of performing the foregoing is to first determine the value of the test statistic T, say $T = t$, and then compute

$$p\text{-value} \approx P\{\chi^2_{k-1-m} \geq t\} \tag{261}$$

The hypothesis would be rejected if $\alpha \geq p$-value.

The Kolmogorov-Smirnov Goodness of Fit Test

The Kolmogorov-Smirnov Goodness of Fit Test provides another way of testing that the Y_j come from the continuous distribution function F that is generally more efficient than discretizing.

After observing $Y_1, \ldots, Y_n$, let F_e be the empirical distribution function defined by

$$F_e(x) = \frac{\#\{i : Y_i \leq x\}}{n} \tag{262}$$

That is, $F_e(x)$ is the proportion of the observed values that are less than or equal to x. Because $F_e(x)$ is a natural estimator of the probability that an observation is less than or equal to x, it follows that, if the null hypothesis that F is the underlying distribution is correct, it should be close to $F(x)$.

The Kolmogorov-Smirnov Goodness of Fit Test

Since this is so for all x, a natural quantity on which to base a test of H_0 is the test quantity

$$D = \max_x |F_e(x) - F(x)| \tag{263}$$

where the maximum is over all values of x from $-\infty$ to $+\infty$. The statistics D is called the *Kolmogorov–Smirnov test statistic* and can be computed using

$$D = \max \left\{ \frac{j}{n} - F(y_{(j)}), F(y_{(j)}) - \frac{j-1}{n}, j = 1, \ldots, n \right\} \tag{264}$$

The Kolmogorov-Smirnov Goodness of Fit Test

Suppose now that the Y_j are observed and their values are such that $D = d$. Since a large value of D would appear to be inconsistent with the null hypothesis that F is the underlying distribution, it follows that the p-value for this data set is given by

$$p\text{-value} = P_F\{D \geq d\} \tag{265}$$

where we have written P_F to make explicit that this probability is to be computed under the assumption that H_0 is correct (and so F is the underlying distribution).

It can be shown that a significance level α test can be obtained by considering the quantity D^* defined by

$$D^* = \left(\sqrt{n} + 0.12 + 0.11/\sqrt{n}\right) D \tag{266}$$

Letting d_α^* be such that

$$P_F\{D^* \geq d_\alpha^*\} = \alpha \tag{267}$$

then the following are accurate approximations for d_α^* for a variety of values:

$$d_{0.10}^* = 1.224, \quad d_{0.05}^* = 1.358, \quad d_{0.025}^* = 1.480, \quad d_{0.01}^* = 1.626$$

The level α test would reject the null hypothesis that F is the distribution if the observed value of D^* is at least as large as d_α^*.

Tests of Independence

Consider a population that can be classified according to two distinct characteristics, which we shall denote as the X-characteristic and the Y-characteristic. We suppose that there are r possible values for the X-characteristic and s for the Y-characteristic, and let

$$P_{ij} = P\{X = i, Y = j\} \tag{268}$$

for $i = 1, \ldots, r$ and $j = 1, \ldots, s$. That is, P_{ij} represents the probability that a randomly chosen member of the population will have X-characteristic i and Y-characteristic j.

Tests of Independence

The different members of the population will be assumed to be independent. Also, let

$$p_i = P\{X = i\} = \sum_{j=1}^{s} P_{ij}, \quad i = 1, \ldots, r \tag{269}$$

and

$$q_j = P\{Y = j\} = \sum_{i=1}^{r} P_{ij}, \quad j = 1, \ldots, s \tag{270}$$

That is, p_i is the probability that an arbitrary member of the population will have X-characteristic i, and q_j is the probability it will have Y-characteristic j.

Tests of Independence

We are interested in testing the hypothesis that a population member's X- and Y-characteristics are independent. That is, we are interested in testing

$$H_0 : P_{ij} = p_i q_j, \quad \text{for all } i = 1, \ldots, r,\ j = 1, \ldots, s \tag{271}$$

against the alternative

$$H_1 : P_{ij} \neq p_i q_j, \quad \text{for some } i, j \quad i = 1, \ldots, r,\ j = 1, \ldots, s. \tag{272}$$

To test this hypothesis, assume that from n members of the population, N_{ij} have simultaneously the X-characteristic i and Y-characteristic j. Since the quantities $p_i, i = 1, \ldots, r$, and $q_j, j = 1, \ldots, s$ are not specified by the null hypothesis, they must first be estimated.

Tests of Independence

The natural estimators of p_i and q_i are, respectively,

$$\hat{p}_i = \frac{N_i}{n} \text{ and } \hat{q}_i = \frac{M_j}{n} \tag{273}$$

where

$$N_i = \sum_{j=1}^{s} N_{ij}, \quad i = 1, \ldots, r \tag{274}$$

and

$$M_j = \sum_{i=1}^{r} N_{ij}, \quad j = 1, \ldots, s \tag{275}$$

Tests of Independence

The test statistic is now defined as

$$T = \sum_{j=1}^{s} \sum_{i=1}^{r} \frac{(N_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j} \tag{276}$$

It can then be proven that when n is large, the test statistic T will have, when H_0 is true, approximately a chi-square distribution with $(r-1)(s-1)$ degrees of freedom. At the approximate significance level α, the test should reject H_0 when

$$T \geq \chi^2_{\alpha,(r-1)(s-1)} \tag{277}$$

$$p\text{-value} \approx P\{\chi^2_{(r-1)(s-1)} \geq T = t\} \tag{278}$$

Inferential Statistics

Pause

Hypothesis Testing Exercises

Table of Contents

Relationship between a set of variables

In many situations, there is a single *response* variable Y, also called the *dependent* variable, which depends on the value of a set of *input*, also called *independent*, variables $x_1, \ldots, x_r$. The simplest type of relationship between the dependent variable Y and the input variables $x_1, \ldots, x_r$ is a linear relationship.

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \tag{279}$$

However, such precision is almost never attainable, meaning that realistically,

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e \tag{280}$$

where e represents a random error.

Relationship between a set of variables

The random error, e, is assumed to be a random variable having mean 0. This implies that

$$E[Y|x] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \tag{281}$$

This equation is called a *linear regression equation* that describes the regression of Y on the set of independent variables $x_1, x_2, \ldots, x_r$. The quantities $\beta_0, \beta_1, \ldots, \beta_r$ are called the *regression coefficients* and must be estimated from a set of data. A *simple linear regression* supposes a linear relationship between the mean response and the value of a single independent variable, expressed as

$$Y = \alpha + \beta x + e \tag{282}$$

Least Squares Estimators of the Regression Parameters

Suppose that the responses Y_i corresponding to the input values $x_i, i = 1, \ldots, n$ are to be observed and used to estimate α and β in a simple linear regression model.

- Determine estimators of α and β
- Considering A an estimator of α and B an estimator of β
- Estimator response to the input variable x_i is $A + Bx_i$
- Since the actual response is Y_i, the squared difference between the actual and estimator reponses is $(Y_i - A - Bx_i)^2$

Least Squares Estimators of the Regression Parameters

With A and B being the estimators of α and β, respectively, then the sum of the squared differences between the estimated responses and the actual response values is given by

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2 \tag{283}$$

The method of least squares chooses as estimators of α and β the values of A and B that minimize SS. To determine these estimators, we differentiate SS first with respect to A and then to B.

Least Squares Estimators of the Regression Parameters

$$\frac{\partial SS}{\partial A} = -2\sum_{i=1}^{n}(Y_i - A - Bx_i) = 0 \tag{284}$$

$$\sum_{i=1}^{n} Y_i = nA + B\sum_{i=1}^{n} x_i \tag{285}$$

$$\left(\sum_{i=1}^{n} Y_i\right) \Big/ n = \left(nA + B\sum_{i=1}^{n} x_i\right) \Big/ n \tag{286}$$

$$\overline{Y} = A + B\overline{x} \Rightarrow A = \overline{Y} - B\overline{x} \tag{287}$$

Least Squares Estimators of the Regression Parameters

$$\frac{\partial SS}{\partial B} = -2\sum_{i=1}^{n} x_i(Y_i - A - Bx_i) = 0 \tag{288}$$

$$\sum_{i=1}^{n} x_i Y_i = A\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 \tag{289}$$

$$\sum_{i=1}^{n} x_i Y_i = (\overline{Y} - B\overline{x})\sum_{i=1}^{n} x_i + B\sum_{i=1}^{n} x_i^2 \tag{290}$$

$$B = \left(\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}\right) \Big/ \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \tag{291}$$

Estimated Regression Line

The least squares estimators of α and β corresponding to the data set $x_i, Y_i, i = 1, \ldots, n$ are, respectively,

$$A = \overline{Y} - B\overline{x} \tag{292}$$

$$B = \left(\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y} \right) \Big/ \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right) \tag{293}$$

The straight line $A + Bx$ is called the *estimated regression line*.

Evaluating the Estimated Regression Line

The value of R^2, called the *coefficient of determination*, is often used as an indicator of how well the regression model fits the data, with a value near 1 indicating a good fit, and one near 0 indicating a poor fit. It is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - A - Bx_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \tag{294}$$

Statistical Inferences about the Regression Parameters

$$Y = \alpha + \beta x + e \qquad e \sim \mathcal{N}(0, \sigma^2) \tag{295}$$

- Statistical Inference about β

$$\sqrt{\frac{(n-2)S_{xx}}{SS_r}}(B - \beta) \sim t_{n-2} \tag{296}$$

- Statistical Inference about α

$$\sqrt{\frac{(n-2)S_{xx}}{\sum_i x_i^2 SS_r}}(A - \alpha) \sim t_{n-2} \tag{297}$$

Statistical Inferences about the Regression Parameters

- Statistical Inference about $\alpha + \beta x_0$

$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}} \sim t_{n-2} \tag{298}$$

- Statistical Inference about $Y(x_0)$

$$\frac{Y(x_0) - (A + Bx_0)}{\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}} \sim t_{n-2} \tag{299}$$

Analysis of Residuals: Assessing the Model

$$Y = \alpha + \beta x + e \qquad e \sim \mathcal{N}(0, \sigma^2) \tag{300}$$

To analyze the residuals, we start by normalizing, or standardizing, the residuals by dividing them by

$$\sqrt{SS_R/(n-2)}, \tag{301}$$

the estimate of the standard deviation of the Y_i.The resulting quantities

$$\frac{Y_i - (A + Bx_i)}{\sqrt{\frac{SS_R}{n-2}}}, \quad i = 1, \ldots, n \tag{302}$$

are called the *standardized residuals*.

Analysis of Residuals: Assessing the Model

When the simple linear regression model is correct, the standardized residuals

$$\frac{Y_i - (A + Bx_i)}{\sqrt{\frac{SS_R}{n-2}}}, \quad i = 1, \ldots, n, \tag{303}$$

are approximately independent standard normal random variables. This implies that they should be randomly distributed about 0 with about 95 percent of their values being between -2 and $+2$. In addition, a plot of the standardized residuals should not indicate any distinct pattern. Indeed, any indication of a distinct pattern should make one suspicious about the validity of the assumed simple linear regression model.

Transforming to Linearity

In many situations, it is clear that the mean response is not a linear function of the input level. In such cases, if the form of the relationship can be determined it is sometimes possible, by a change of variables, to transform it into a linear form.

For instance, let us assume that $W(t)$ is approximately related to t by the functional form

$$W(t) \approx ce^{-dt} \tag{304}$$

Using logarithms (inverses of exponential functions), this equation can be expressed as

$$\log W(t) \approx \log c - dt \tag{305}$$

which can be modeled as a linear regression.

Other Regressions

Polynomial Regression

In situations where the functional relationship between the response Y and the independent variable x cannot be adequately approximated by a linear relationship, it is sometimes possible to obtain a reasonable fit by considering a polynomial relationship.

Multiple Linear Regression

In the majority of applications, the response of an experiment can be predicted more adequately not on the basis of a single independent input variable but on a collection of such variables.

Data Science in Aerospace

Bachelor's Degree in Aeronautical and Space Sciences
Bachelor's Degree in Aeronautical Management
Short Cycle in Aircraft Repair and Maintenance

Emanuel A. R. Camacho

emanuel.camacho@iseclisboa.pt

Instituto Superior de Educação e Ciências (ISEC Lisboa)