# Intro to Business Statistics

## Archived study guides

These archived materials are no longer updated. This archive was exported from a saved copy of these guides.

Export date: 2026-09-05T17:13:38.281Z

## Contents

# Unit 1 – Sampling and Data

## 1.1 Definitions of Statistics, Probability, and Key Terms

Statistics and probability form the foundation of data analysis and decision-making. These tools help us make sense of complex information, from summarizing data to drawing conclusions about populations. By understanding these concepts, we can interpret real-world phenomena and make informed predictions.

Descriptive statistics summarize data, while inferential statistics draw conclusions about populations. Probability quantifies uncertainty, enabling predictions and risk assessment. Population parameters are estimated using sample statistics, bridging the gap between what we can observe and broader trends.

## Introduction to Statistics and Probability

## Descriptive vs inferential statistics

- Descriptive statistics summarize and describe basic features of sample data without drawing conclusions beyond the data
- Involve measures of central tendency (mean, median, mode) and variability (range, variance, standard deviation)
- Example: calculating the average test score of a class
- Inferential statistics use sample data to make inferences, estimates, or predictions about a larger population
- Involve hypothesis testing and confidence intervals to generalize findings from a sample to a population
- Rely on probability theory to determine the likelihood of observations
- Example: using a survey of 1,000 voters to estimate the percentage of all voters who support a candidate
- Statistical inference is the process of drawing conclusions about populations based on sample data

## Probability for predictions

- Probability quantifies the likelihood of random events by assigning a numerical value between 0 (impossible event) and 1 (certain event)
- Example: the probability of rolling a 6 on a fair die is $\frac{1}{6}$
- Probability models and analyzes situations involving uncertainty or randomness
- Helps make predictions about future events based on past observations
- Allows for the quantification of risk in decision-making processes (insurance premiums, investment strategies)
- Probability distributions describe the likelihood of different outcomes in a given situation
- Enable the calculation of probabilities for specific events
- Common distributions include:

• Binomial: discrete probability distribution of the number of successes in a fixed number of independent trials (coin flips)

• Poisson: discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space (number of customers arriving at a store per hour)

• Normal: continuous probability distribution that is symmetric about the mean, with data near the mean more frequent (heights, IQ scores)

• Random variables are variables whose possible values are outcomes of a random phenomenon

## Population parameters vs sample statistics

• Population: the entire group of individuals, objects, or events of interest in a study

• Often large and sometimes impossible to observe entirely (all registered voters in a country)

• Sample: a subset of the population selected for observation and analysis to make inferences about the population

• Example: a random sample of 500 registered voters

• Various sampling methods are used to select representative samples from populations

• Parameters: numerical measures that describe characteristics of a population, usually unknown and estimated using sample statistics

• Denoted by Greek letters ($\mu$ for population mean, $\sigma$ for population standard deviation)

• Statistics: numerical measures computed from sample data, used to estimate population parameters

• Denoted by Latin letters ($\bar{x}$ for sample mean, $s$ for sample standard deviation)

• Example: the sample mean height of 100 students is a statistic used to estimate the population mean height of all students in a school

## Data Analysis Techniques

• Data analysis involves examining, cleaning, transforming, and modeling data to extract useful information

• Correlation measures the strength and direction of the linear relationship between two variables

• Regression analysis is used to model the relationship between a dependent variable and one or more independent variables

# 1.2 Data, Sampling, and Variation in Data and Sampling

Data types and sampling methods shape how business statistics are collected and analyzed. Qualitative data represents non-numerical attributes, while quantitative data consists of measurable values. These distinctions help determine which analysis techniques fit the data.

Random sampling ensures unbiased representation of a population. Simple random, stratified, cluster, and systematic sampling are common methods. Each has advantages for different research scenarios, helping businesses make informed decisions based on reliable data.

# Types of Data and Sampling Methods

## Qualitative vs quantitative data

• Qualitative data represents non-numerical attributes, characteristics, or categories (colors, marital status, product reviews)

• Quantitative data consists of numerical values that can be measured or counted

• Discrete quantitative data involves countable values, often integers (number of employees, defective products)

• Continuous quantitative data includes measurable values that can take on any value within a range (height of students, time to complete a task)

## Types of random sampling

• Simple random sampling ensures each element in the population has an equal chance of being selected, resulting in an unbiased and representative sample

• Can be conducted with or without replacement

• Stratified sampling divides the population into homogeneous subgroups (strata) based on a specific characteristic, then applies simple random sampling within each stratum to ensure representation of all subgroups

• Cluster sampling involves dividing the population into naturally occurring groups (clusters), randomly selecting a sample of clusters, and including all elements within the selected clusters in the sample

• Systematic sampling selects elements from the population at a fixed interval, with the starting point chosen randomly and the interval determined by dividing population size by desired sample size

# Sources of Variation in Data and Sampling

## Sources of data variation

• Sampling errors arise from differences between sample statistics and population parameters due to chance

• Can be reduced by increasing sample size

• Types include selection bias (non-representative sample) and random sampling error (inherent variability in sampling)

• Nonsampling errors occur during data collection, processing, or analysis and are not related to the sampling process

• Measurement errors result from inaccuracies in data collection instruments or methods (poorly worded survey questions, faulty measuring devices)

• Non-response errors occur when some elements in the sample do not respond, potentially leading to biased results if non-respondents differ from respondents

• Coverage errors happen when the sampling frame does not accurately represent the population (outdated telephone directories, incomplete email lists)

• Processing errors are mistakes made during data entry, coding, or analysis (typos, incorrect data entry, programming errors)

## Population, Sample, and Statistical Concepts

• Population refers to the entire group of individuals or objects about which information is desired

• Sample is a subset of the population selected for study

• A parameter is a numerical characteristic of the population, while a statistic is a numerical characteristic of the sample

• Variability refers to the extent to which data points differ from each other

• A sampling frame is the list of all elements in the population from which the sample is drawn

• A representative sample accurately reflects the characteristics of the population it was drawn from

## 1.3 Levels of Measurement

Measurement theory explains how variables are classified before data analysis begins. It organizes information into nominal, ordinal, interval, and ratio scales, each with different levels of precision and interpretation.

Frequency distributions organize data into meaningful groups, allowing you to calculate relative and cumulative frequencies. This approach reveals patterns, trends, and the overall shape of a data distribution.

## Levels of Measurement and Frequency Distributions

## Measurement Theory and Data Classification

• Measurement theory provides a framework for understanding how variables are measured and classified

• Data can be broadly categorized into two types:

• Categorical data: Qualitative information that can be sorted into categories (e.g., nominal and ordinal scales)

• Quantitative data: Numerical information that can be measured (e.g., interval and ratio scales)

• Variables can be further classified as:

• Discrete variables: Can only take on specific, countable values

• Continuous variables: Can take on any value within a given range

## Levels of measurement scales

• Nominal scale categorizes data without any inherent order or numerical meaning (gender, race, religion, zip code)

• Ordinal scale categorizes data with a meaningful order but inconsistent scale, differences between values are not interpretable (letter grades, customer satisfaction ratings, socioeconomic status)

• Interval scale measures numerical data with consistent intervals between values but no true zero point, ratios are not meaningful (temperature in ℃ or ℉, dates, IQ scores)

• Ratio scale measures numerical data with consistent intervals and a true zero point, ratios are meaningful and interpretable (height, weight, age, income, distance)

• Ratio scales offer the highest level of measurement precision

## Relative and cumulative frequency calculations

• Frequency represents the number of observations in each category or class

• Relative frequency represents the proportion or percentage of observations in each category or class

• Calculated by dividing the frequency of each class by the total number of observations

• Formula: $\text{Relative Frequency} = \frac{\text{Class Frequency}}{\text{Total Observations}}$

• Cumulative frequency represents the running total of frequencies up to a given class

• Calculated by adding the frequencies of all classes up to and including the current class

• Cumulative relative frequency represents the running total of relative frequencies up to a given class

• Calculated by adding the relative frequencies of all classes up to and including the current class

• Formula: $\text{Cumulative Relative Frequency} = \frac{\text{Cumulative Frequency}}{\text{Total Observations}}$

## Interpretation of grouped frequency distributions

• Grouped frequency distribution organizes data into classes or intervals, useful for large datasets or continuous variables

• Class intervals define the range of values for each class

• Lower class limit represents the smallest value in a class interval

• Upper class limit represents the largest value in a class interval

• Class midpoint represents the average of the lower and upper class limits

• Formula: $\text{Class Midpoint} = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$

• Class width represents the difference between the lower limits of two consecutive classes, should be consistent throughout the distribution

• Interpreting interval data involves: 1. Identifying the modal class with the highest frequency 2. Determining the shape of the distribution (symmetric, skewed left, or skewed right) 3. Analyzing the spread of the data using range, interquartile range, or standard deviation 4. Comparing relative frequencies or cumulative relative frequencies between classes

# 1.4 Experimental Design and Ethics

Experimental design helps researchers test cause-and-effect relationships. It involves manipulating independent variables, measuring dependent variables, and using random assignment to create comparable groups. These elements help isolate the effects of treatments and minimize confounding factors.

Ethical considerations shape every experimental design. Blinding techniques, such as single-blind and double-blind studies, reduce bias and improve result reliability. Informed consent, confidentiality, and minimizing harm protect participants and maintain research integrity.

# Experimental Design

## Components of randomized experiments

• Randomized experiments establish cause-and-effect relationships between variables by manipulating the independent variable (explanatory variable) and measuring changes in the dependent variable (response variable)

• Independent variable is controlled by the researcher hypothesized to cause changes in the dependent variable (drug dosage)

• Dependent variable changes in response to the independent variable and is measured to determine the effect (blood pressure)

• Treatments are different levels or conditions of the independent variable administered to experimental units (high dose, low dose, placebo)

• Experimental units are subjects or objects to which treatments are applied can be individuals, groups, or objects (patients, petri dishes)

• Random assignment of experimental units to treatment groups helps ensure similarity in all aspects except for the treatment received (coin flip, random number generator)

• Replication of the experiment with multiple trials or groups increases the reliability of results

## Random assignment and control groups

• Random assignment distributes potential confounding variables evenly across treatment groups reducing the likelihood that differences in the dependent variable are due to factors other than the independent variable (age, gender)

• Control groups do not receive the treatment or receive a standard treatment serving as a baseline for comparison with treatment groups (sugar pill)

• Control groups help isolate the effect of the independent variable by controlling for other factors (natural recovery, regression to the mean)

• Random assignment and control groups minimize the impact of confounding variables, increase internal validity, and allow for stronger causal inferences about the relationship between variables (smoking causes lung cancer)

## Statistical Considerations in Experimental Design

• Sample size affects the precision and generalizability of results, with larger samples typically providing more reliable estimates

• Statistical power is the ability of a study to detect a true effect, influenced by sample size, effect size, and significance level

• Validity refers to the extent to which a study measures what it intends to measure and produces accurate, generalizable results

• Research protocols outline the specific procedures, methods, and analyses to be used in an experiment, ensuring consistency and reproducibility

# Ethics in Experimental Design

## Blinding in experimental design

• Blinding conceals information about treatment assignment from participants, researchers, or both to minimize bias and the influence of expectations on results

• Single-blind: Participants are unaware of their treatment group (patient does not know if they received the drug or placebo)

• Double-blind: Both participants and researchers directly involved in the experiment are unaware of treatment assignments (neither patient nor doctor knows who received the drug or placebo)

• Blinding helps minimize the placebo effect where participants experience a perceived improvement due to their belief in the treatment, even if the treatment is inactive (sugar pill reduces pain)

• Blinding prevents participants' expectations from influencing their responses or behavior ensuring that observed effects are due to the treatment itself rather than psychological factors (Hawthorne effect)

• Blinding enhances the reliability and validity of experimental results by reducing bias (confirmation bias, observer bias)

• Other ethical considerations in experimental design include: 1. Informed consent: Participants should be fully informed about the experiment and voluntarily agree to participate (risks, benefits, purpose) 2. Confidentiality: Participants' personal information and data should be kept confidential and secure (anonymized data, encrypted files) 3. Minimizing harm: Experiments should be designed to minimize potential risks or harm to participants (safety monitoring, emergency protocols)

# Unit 2 – Descriptive Statistics

## 2.1 Display Data

Stemplots, histograms, and time series graphs are powerful tools for visualizing data. These methods help us understand patterns, trends, and distributions in datasets of various sizes. Each technique serves a unique purpose, from analyzing small datasets to tracking changes over time.

By mastering these visualization methods, we can effectively communicate complex information and make data-driven decisions. Understanding how to construct and interpret these graphs is crucial for anyone working with data, whether in business, science, or everyday life.

## Displaying Data

## Stemplots for small datasets

- Visualize small datasets (typically less than 50 data points) and identify patterns or outliers
- Consist of a stem (first digit(s) of the data value) and leaves (last digit of the data value)
- Split each data value into a stem and a leaf, then plot accordingly
- Constructing a stemplot involves: 1. Determine the range of the data values 2. Choose an appropriate stem unit based on the range 3. List the stems vertically in ascending order 4. Place each data value's leaf next to the corresponding stem (if a data value has more than one digit in the leaf, place it next to the stem multiple times)
- Interpret a stemplot by identifying the minimum and maximum values, looking for clusters, gaps, or outliers, and assessing the overall shape of the distribution (symmetric, skewed, or bimodal)

## Histograms for large datasets

- Display the distribution of a large dataset using bars representing the frequency or relative frequency of data values within specific ranges (bins)
- Constructing a histogram involves: 1. Determine the range of the data values 2. Choose an appropriate bin width based on the range and desired number of bins 3. Define the bin intervals 4. Count the number of data values that fall within each bin 5. Draw the histogram with bin intervals on the x-axis and frequency or relative frequency on the y-axis
- Analyze a histogram by examining its shape (symmetric, skewed left or right, bimodal, or uniform), locating the approximate mean or median, observing the range and variability of the data, and identifying any outliers or unusual features
- Similar to a bar chart, but used for continuous data rather than categorical data

## Time series graphs for trends

- Display data collected over a specific time period to show trends, patterns, and changes in the data over time (also known as a line graph)
- Components include time on the x-axis, variable of interest on the y-axis, and data points connected by lines to show the progression of the variable over time

• Constructing a time series graph involves: 1. Determine the time period and variable to be analyzed 2. Collect data for the variable at regular intervals over the specified time period 3. Plot the data points on the graph, with time on the x-axis and the variable on the y-axis 4. Connect the data points with lines to show the trend over time

• Interpret a time series graph by identifying the overall trend (increasing, decreasing, or stable), looking for seasonal patterns or cyclical behavior, observing sudden changes or irregularities, and comparing the graph with other relevant variables or events to identify potential relationships or causes of changes

## Additional Data Visualization Methods

• Scatterplot: Used to display the relationship between two continuous variables, with each point representing a pair of values

• Pie chart: Displays the proportion of different categories in a dataset as slices of a circular "pie"

• Box plot: Summarizes the distribution of a dataset using quartiles, showing the median, spread, and potential outliers

• Pareto chart: Combines a bar chart and a line graph to show both individual values and cumulative totals, often used in quality control

• Dot plot: Represents individual data points as dots along a number line, useful for small datasets and comparing distributions

# 2.2 Measures of the Location of the Data

Quartiles and percentiles show where values fall within a dataset. They divide data into ordered positions, making it easier to compare individual values, describe spread, and summarize typical ranges.

Median and interquartile range are robust measures for skewed data. They're less affected by outliers, giving a clearer picture of typical values. Percentile rankings let us compare individual values to the whole dataset, useful in education, healthcare, and business.

## Measures of the Location of the Data

## Quartiles and percentiles for distribution

• Quartiles divide a dataset into four equal parts

• First quartile (Q1) represents the 25th percentile, meaning 25% of the data falls below this value

• Second quartile (Q2) represents the 50th percentile or median, where half of the data is below and half is above this value

• Third quartile (Q3) represents the 75th percentile, indicating 75% of the data is below this value

• Percentiles represent the percentage of data below a specific value

• Calculated by sorting the data in ascending order and using the formula:
$\text{Percentile rank} = \frac{i}{n} \times 100$

• $i$ represents the rank of the value in the sorted dataset (e.g., 1st, 2nd, 3rd)

• $n$ represents the total number of values in the dataset

• Percentiles provide a more detailed understanding of data distribution compared to quartiles

• Interquartile range (IQR) is the difference between Q3 and Q1: $IQR = Q3 - Q1$

• Measures the spread of the middle 50% of the data, excluding the top and bottom 25%

• Useful for identifying the range of typical values and detecting potential outliers

• The five-number summary provides a concise description of the data's distribution, including the minimum, Q1, median, Q3, and maximum values

## Median and interquartile range interpretation

• Median is the middle value in a sorted dataset, separating the higher half from the lower half

• Resistant to outliers, making it a robust measure of central tendency (average) for skewed distributions

• Preferred over the mean when the data is skewed (asymmetrical) or contains extreme values (outliers)

• Example: In a dataset of house prices, the median is less affected by a few extremely expensive houses compared to the mean

• Interquartile range (IQR) measures the spread (dispersion) of the middle 50% of the data

• Resistant to outliers, providing a robust measure of dispersion that is not influenced by extreme values

• A larger IQR indicates greater variability (spread) in the middle 50% of the data

• Example: Comparing the IQR of test scores between two classes can reveal which class has a wider range of performance in the middle 50% of students

• Box plots (also known as box-and-whisker plots) visually represent the five-number summary and IQR, making it easy to identify the data's distribution and potential outliers

## Percentile rankings for data comparison

• Percentile rankings allow for comparisons between individual values and the entire dataset

• A value at the 80th percentile is greater than 80% of the values in the dataset, while a value at the 20th percentile is greater than only 20% of the values

• Percentile rankings provide context for understanding the relative position of a value within a dataset

• Percentile rankings are useful for: 1. Identifying the relative position (rank) of a value within a dataset, regardless of the dataset's size or scale 2. Comparing values from different datasets with different scales or units (e.g., comparing test scores from different exams)

• Percentile rankings can be used in various fields, such as:

• Education: comparing student performance on standardized tests (SAT, ACT) to determine their relative standing among test-takers

• Healthcare: assessing growth and development in children by comparing their measurements (height, weight) to age-specific percentiles

• Business: benchmarking company performance (revenue, market share) against industry peers to identify competitive position

## Additional measures of central tendency and variability

- Mean: The arithmetic average of all values in a dataset, calculated by summing all values and dividing by the number of values
- Mode: The most frequently occurring value in a dataset, useful for identifying the most common category or value
- Standard deviation: A measure of variability that quantifies the average distance between each data point and the mean, providing insight into the spread of the data around its central tendency

# 2.3 Measures of the Center of the Data

Central tendency measures describe typical values in a dataset. The mean, median, and mode each summarize data differently: the mean is sensitive to outliers, while the median is more resistant to extreme values.

Understanding the differences between population and sample means is key for statistical inference. Grouped frequency tables simplify complex datasets, allowing for easier calculation of central tendency measures. These concepts form the foundation for analyzing data distributions and variability.

## Measures of Central Tendency

## Mean and median calculations

- Mean calculates the average value by summing all values and dividing by the total number of values
- Sensitive to extreme values or outliers (income data with a few billionaires)
- Best used when data is symmetrically distributed and without outliers (heights of students in a class)
- Formula: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ where $\bar{x}$ is the mean, $\sum_{i=1}^{n} x_i$ is the sum of all values, and $n$ is the number of values
- Median represents the middle value when data is ordered from smallest to largest
- Robust to extreme values or outliers (house prices in a neighborhood with a few mansions)
- Best used when data is skewed or contains outliers (salaries in a company with a highly paid CEO)
- For an odd number of values, the median is the middle value (median of 1, 2, 3, 4, 5 is 3)
- For an even number of values, the median is the average of the two middle values (median of 1, 2, 3, 4 is 2.5)
- Mode is the value that appears most frequently in a data set
- Useful for categorical data or discrete numerical data

## Sample vs population means

- Population mean ($\mu$) represents the average of all values in a population
- Usually unknown and must be estimated from a sample (average income of all U.S. households)
- Denoted by the Greek letter $\mu$
- Sample mean ($\bar{x}$) represents the average of all values in a sample

• Used to estimate the population mean (average income of 1,000 randomly selected U.S. households)

• Denoted by $\bar{x}$

• Estimating population mean from a sample: 1. The sample mean is an unbiased estimator of the population mean 2. As sample size increases, the sample mean approaches the population mean (law of large numbers) 3. The larger the sample size, the more accurate the estimate of the population mean (smaller margin of error)

## Mean from grouped frequency tables

• Grouped frequency table organizes data into intervals or classes with corresponding frequencies

• Intervals are typically of equal width (age groups: 0-9, 10-19, 20-29, etc.)

• Interval midpoint represents the average of the lower and upper bounds of an interval

• Represents all values within the interval (midpoint of 10-19 is 14.5)

• Formula: $\text{Midpoint} = \frac{\text{Lower bound} + \text{Upper bound}}{2}$

• Calculating the mean from a grouped frequency table: 1. Multiply each interval midpoint by its corresponding frequency 2. Sum the products of midpoints and frequencies 3. Divide the sum by the total frequency

• Formula: $\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$ where $\bar{x}$ is the mean, $m_i$ is the midpoint of the $i$-th interval, $f_i$ is the frequency of the $i$-th interval, and $k$ is the number of intervals

## Data Characteristics and Distribution

• Central tendency measures (mean, median, mode) describe the typical or central value in a distribution

• Dispersion measures the spread or variability of data around the central tendency

• Distribution refers to the pattern of data values and their frequencies

• Symmetry in a distribution occurs when data is evenly distributed around the center

• Variability indicates how much individual data points differ from each other in a data set

# 2.4 Sigma Notation and Calculating the Arithmetic Mean

Sigma notation gives a compact way to write sums in statistics. It is especially useful for calculating the arithmetic mean because the formula for an average requires adding every value in a dataset and dividing by the number of values.

The arithmetic mean, calculated using sigma notation, helps summarize data by finding the typical value in a distribution. Whether working with population or sample data, this method allows for efficient computation of averages across various fields, from finance to scientific research.

# Sigma Notation and Arithmetic Mean

## Arithmetic mean calculation using sigma notation

- Population mean
- Denoted by the Greek letter $\mu$ (mu) represents the average value of all items in a population
- Formula: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ calculates the population mean by summing all values and dividing by the total number of items
- $\sum_{i=1}^{N} x_i$ represents the sum of all values in the population (heights of all students in a school)
- $N$ is the total number of values in the population (total number of students in the school)
- Sample mean
- Denoted by $\bar{x}$ (x-bar) represents the average value of a subset of items selected from a population
- Formula: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ calculates the sample mean by summing the values in the sample and dividing by the sample size
- $\sum_{i=1}^{n} x_i$ represents the sum of all values in the sample (heights of students in a randomly selected classroom)
- $n$ is the total number of values in the sample (number of students in the selected classroom)

## Components of sigma notation

- Sigma ($\sum$) symbol indicates the summation or addition of a series of terms (adding all the values in a dataset)
- Subscript ($i$) represents the index variable or current position in the sequence (first value, second value, etc.)
- $i$ is commonly used, but other letters like $j$ or $k$ can also be used depending on the context
- Lower limit is the starting value of the index variable, typically written below the sigma (start summing from the first value)
- Example: $\sum_{i=1}$ means the summation starts at $i = 1$ (begin adding from the first term)
- Upper limit is the ending value of the index variable, typically written above the sigma (stop summing at the last value)
- Example: $\sum_{i=1}^{n}$ means the summation ends at $i = n$ (add up to and including the nth term)
- Expression is the term or formula being summed, written to the right of the sigma (the values being added together)

• Example: $\sum_{i=1}^{n} x_i$ means summing all $x_i$ values from $i = 1$ to $i = n$ (add up all the data points from the first to the last)

## Sigma notation in real-world averages

1. Identify the dataset and determine if it represents a population or sample (all employees in a company vs. a subset of employees)

2. Write out the appropriate formula for the arithmetic mean using sigma notation - Use $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$ for population data (calculating average salary for all employees) - Use $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ for sample data (calculating average salary for a department)

3. Substitute the given values into the formula - Replace $x_i$ with the actual values from the dataset (individual salaries) - Replace $N$ or $n$ with the total number of values in the dataset (total number of employees or sample size)

4. Perform the summation in the numerator (add up all the salaries)

5. Divide the sum by the total number of values ($N$ or $n$) (divide total salary by number of employees)

6. The result is the arithmetic mean for the given dataset (average salary for the company or department)

## Measures of Central Tendency

• The arithmetic mean is a measure of central tendency, which describes the typical or central value in a distribution

• Other measures of central tendency include:

• Median: the middle value in a sorted dataset

• Mode: the most frequently occurring value in a dataset

• These measures help summarize the central location of a data distribution

## 2.5 Geometric Mean

The geometric mean is a powerful tool for calculating central tendencies, especially useful for data with varying ranges. It's calculated by finding the nth root of the product of all values, making it ideal for stock prices and growth rates.

In economics and finance, the geometric mean shines when analyzing growth rates and investment returns. It accounts for compounding effects, providing a more accurate representation of average growth compared to the arithmetic mean, particularly in exponential growth patterns.

## Geometric Mean

## Calculation of geometric mean

• Calculates central tendency for a set of numbers by finding nth root of product of all values

• Useful when values have different ranges (stock prices, growth rates)

• Formula: $\sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$

• $n$ represents number of values in set

• $x_1, x_2, \ldots, x_n$ represent individual values in set

• Steps to calculate: 1. Multiply all values in set together 2. Take nth root of product, where n is number of values in set

• Examples:

• Geometric mean of 2, 8, and 16 is $\sqrt[3]{2 \times 8 \times 16} = 4$

• Geometric mean of 10, 20, 40, and 80 is $\sqrt[4]{10 \times 20 \times 40 \times 80} = 20$

## Applications in economic growth

• Commonly used to calculate average growth rates over time

• Examples: average annual growth rates for GDP, company revenue, investment returns

• Calculating average growth rate using geometric mean: 1. Convert each period's growth rate to decimal and add 1 (5% becomes 1.05) 2. Multiply these values together 3. Take nth root of product, where n is number of periods 4. Subtract 1 from result and convert back to percentage

• Provides more accurate representation of average growth compared to arithmetic mean when compounding is involved

• Example: if GDP grows by 3% in year 1 and 5% in year 2, geometric mean growth rate is $\sqrt{1.03 \times 1.05} - 1 = 4\%$, while arithmetic mean is $(3\% + 5\%)/2 = 4\%$

• Useful for analyzing exponential growth patterns in economic indicators

## Geometric mean for investment returns

• Calculates average annual return for an investment over multiple periods

• Accounts for compounding effect of returns

• Provides more accurate representation of investment performance compared to arithmetic mean

• Steps to calculate geometric mean rate of return: 1. Convert each annual return to decimal and add 1 (-3% becomes 0.97) 2. Multiply these values together 3. Take nth root of product, where n is number of years 4. Subtract 1 from result and convert back to percentage

• When dealing with negative returns, geometric mean will always be lower than arithmetic mean

• Accounts for compounding effect of losses on overall return

• Example: if an investment has returns of 10%, -5%, and 20% over three years, geometric mean return is $\sqrt[3]{1.10 \times 0.95 \times 1.20} - 1 = 7.8\%$, while arithmetic mean is $(10\% - 5\% + 20\%)/3 = 8.3\%$

## Advanced applications and related concepts

• Logarithms are often used to simplify calculations involving geometric means

• Geometric mean is particularly useful in time series analysis for financial and economic data

• The concept of geometric mean is closely related to compound interest calculations in finance

# 2.6 Skewness and the Mean, Median, and Mode

Data distributions come in different shapes, affecting how we interpret their central tendencies. Symmetrical distributions have equal tails, while skewed ones lean left or right. Understanding these shapes helps us choose the right measure of central tendency.

Mean, median, and mode are key measures of central tendency, each with unique properties. The mean is sensitive to outliers, the median is robust, and the mode shows the most common value. Skewness values help quantify distribution asymmetry, guiding our interpretation of data.

## Measures of Central Tendency and Distribution Shape

## Symmetrical vs skewed distributions

• Symmetrical distributions have histograms that are mirror-symmetric about the center, with the mean, median, and mode being equal or very close to each other (normal distribution, uniform distribution)

• Skewed distributions have asymmetric histograms, with one tail longer than the other and the mean, median, and mode not equal and potentially far apart

• Right-skewed (positively skewed) distributions have a longer tail on the right side of the histogram, with the mean > median > mode (income distribution)

• Left-skewed (negatively skewed) distributions have a longer tail on the left side of the histogram, with the mode > median > mean (exam scores with a difficult test)

## Measures of central tendency

• Mean represents the arithmetic average of all values in a dataset, calculated using the formula $\bar{x} = \frac{\sum_{i=1} x_i}{n}$, and is sensitive to extreme values (outliers)

• In a symmetrical distribution, the mean equals the median and mode

• In a right-skewed distribution, the mean is greater than the median and mode

• In a left-skewed distribution, the mean is less than the median and mode

• Median is the middle value when the dataset is ordered from smallest to largest, calculated as the $\frac{n+1}{2}$th value for odd n or the average of the $\frac{n}{2}$th and $(\frac{n}{2}+1)$th values for even n, and is a robust measure not affected by extreme values

• In a symmetrical distribution, the median equals the mean and mode

• In a skewed distribution, the median lies between the mean and mode

• Mode is the most frequently occurring value in a dataset and can have no mode, one mode (unimodal), or multiple modes (bimodal, multimodal)

• In a symmetrical distribution, the mode equals the mean and median

• In a right-skewed distribution, the mode is less than the median and mean

• In a left-skewed distribution, the mode is greater than the median and mean

## Interpreting skewness values

• Skewness measures the asymmetry of a distribution using Pearson's coefficient of skewness: $SK = \frac{3(\bar{x} - \text{Median})}{s}$, where $\bar{x}$ is the sample mean and s is the sample standard deviation

• Interpretation of skewness values: 1. If SK = 0, the distribution is symmetrical 2. If SK > 0, the distribution is right-skewed (positively skewed), with higher positive values indicating a more severe right skew 3. If SK < 0, the distribution is left-skewed (negatively skewed), with lower negative values indicating a more severe left skew

• Rule of thumb for interpreting skewness values:

• If |SK| < 0.5, the distribution is approximately symmetrical

• If 0.5 ≤ |SK| < 1, the distribution is moderately skewed

• If |SK| ≥ 1, the distribution is highly skewed

## Additional measures and visualizations

• Standard deviation measures the spread of data around the mean, providing insight into the variability of a distribution

• Quartiles divide the dataset into four equal parts, with the second quartile being the median

• Percentiles indicate the relative position of a data point within the distribution

• Box plots visually represent the quartiles, median, and potential outliers of a dataset

• Histograms provide a graphical representation of the frequency distribution of a dataset, helping to visualize its shape and skewness

# 2.7 Measures of the Spread of the Data

Measures of variability help us understand how spread out data is. Standard deviation, the most common measure, tells us how far values typically stray from the average. We use different formulas for samples versus entire populations, and can compare datasets using the coefficient of variation.

Other ways to measure spread include range, interquartile range, and skewness. These tools give us a fuller picture of how data is distributed, which is crucial for making accurate interpretations and predictions in statistical analysis.

## Measures of Variability

## Standard deviation calculation and interpretation

• Quantifies spread of data values around the mean by calculating square root of variance

• Variance averages squared deviations from mean, making all values positive and giving more weight to larger deviations

• Low standard deviation indicates data points clustered closely around mean (test scores)

• High standard deviation indicates data points spread out over wider range (income levels)

• In normal distribution, approximately 68% of values fall within one standard deviation of mean, 95% within two, and 99.7% within three

• Outliers can significantly impact standard deviation (few extremely high or low values)

• Range provides a simple measure of spread by calculating the difference between the highest and lowest values in a dataset

## Sample vs population standard deviation

• Population standard deviation $\sigma$ used when data represents entire population

• Formula: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$

• $x_i$: individual values

• $\mu$: population mean

• $N$: number of values in population (census data)

• Sample standard deviation $s$ used when data represents subset of population

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

• $x_i$: individual values

• $\bar{x}$: sample mean

• $n$: number of values in sample (survey responses)

• Denominator $n - 1$ (degrees of freedom) accounts for using sample mean $\bar{x}$ instead of population mean $\mu$

## Coefficient of variation for comparisons

• Coefficient of variation (CV) standardizes dispersion as ratio of standard deviation to mean, expressed as percentage

• Formula: $CV = \frac{s}{\bar{x}} \times 100\%$ or $CV = \frac{\sigma}{\mu} \times 100\%$

• Compares variability between datasets with different units or means (stock returns vs bond returns)

• Higher CV indicates greater variability relative to mean

• Useful for comparing risk or relative variability across investments, industries, populations (pharmaceutical companies)

• Limitations: sensitive to small changes in mean close to zero, assumes standard deviation is meaningful for data

## Additional Measures of Spread

• Interquartile range (IQR) measures variability by calculating the difference between the third and first quartiles

• Dispersion refers to the overall spread or scatter of data points in a distribution

• Skewness indicates the degree of asymmetry in a distribution, affecting how data is spread around the central tendency

# Unit 3 – Probability Topics

## 3.1 Terminology

Probability is the backbone of statistical analysis, providing a framework to measure uncertainty and make predictions. This section introduces key concepts like experiments, outcomes, and events, laying the groundwork for understanding how probabilities are calculated and interpreted.

Diving deeper, we explore the difference between "and" and "or" events, conditional probability, and independence. These concepts are crucial for solving complex probability problems and form the basis for more advanced statistical techniques used in data analysis and decision-making.

## Probability Terminology and Concepts

## Key probability terminology

- Experiment involves a process or action that generates well-defined outcomes (rolling a die, flipping a coin, drawing a card from a deck)
- Outcome represents a single result of an experiment, also known as a sample point (rolling a 3 on a die, flipping a head on a coin, drawing an ace from a deck)
- Sample space encompasses the set of all possible outcomes of an experiment, denoted by $S$ ( $S = \{1, 2, 3, 4, 5, 6\}$ for rolling a die, $S = \{H, T\}$ for flipping a coin)
- Event consists of a subset of the sample space, a collection of one or more outcomes (rolling an even number on a die, flipping at least one head in two coin flips)
- Random variable is a function that assigns a numerical value to each outcome in the sample space

## Probability calculation for equal outcomes

- Probability of an event measures the likelihood of an event occurring, denoted by $P(A)$ for event $A$
- Formula: $P(A) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$
- Equally likely outcomes occur when each outcome in the sample space has the same probability of occurring (each side of a fair die has a probability of $\frac{1}{6}$)
- Calculating probabilities involves: 1. Identifying the sample space and the total number of possible outcomes 2. Determining the number of favorable outcomes for the event of interest 3. Dividing the number of favorable outcomes by the total number of possible outcomes

## "And" vs "or" events

- "AND" events (Intersection), denoted by $A \cap B$, occur when both event $A$ and event $B$ happen simultaneously (rolling a 3 AND flipping a head)
- Probability of "AND" events: $P(A \cap B) = P(A) \times P(B)$ if $A$ and $B$ are independent
- "OR" events (Union), denoted by $A \cup B$, occur when either event $A$ or event $B$ or both happen (rolling a 3 OR flipping a head)
- Probability of "OR" events: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- Mutually exclusive events cannot occur simultaneously
- If $A$ and $B$ are mutually exclusive, $P(A \cap B) = 0$
- For mutually exclusive events, $P(A \cup B) = P(A) + P(B)$

## Advanced Probability Concepts

- Conditional probability is the probability of an event occurring given that another event has already occurred
- Independent events are events where the occurrence of one does not affect the probability of the other
- Law of large numbers states that as the number of trials increases, the sample mean approaches the expected value
- Expected value is the average outcome of an experiment if it is repeated many times

# 3.2 Independent and Mutually Exclusive Events

Independent and mutually exclusive events describe different relationships between outcomes. Independent events do not affect each other's probabilities, while mutually exclusive events cannot happen at the same time in one trial.

The probability rules change depending on which relationship applies. For independent events, you multiply probabilities; for mutually exclusive events, you add probabilities because there is no overlap.

## Independent and Mutually Exclusive Events

## Independence vs mutual exclusivity

- Independent events occur when the outcome of one event does not influence the probability of another event happening (coin flips)
- Probability of event A given event B has occurred is equal to the probability of event A: $P(A|B) = P(A)$
- Probability of event B given event A has occurred is equal to the probability of event B: $P(B|A) = P(B)$
- Mutually exclusive events cannot happen simultaneously in the same trial (rolling even or odd numbers on a die)
- If one mutually exclusive event occurs, the other event(s) cannot occur in that trial
- Probability of the intersection of mutually exclusive events A and B is zero: $P(A \cap B) = 0$
- Venn diagrams can be used to visualize mutually exclusive events as non-overlapping circles

## Probability calculations for event types

- For independent events, calculate the probability by multiplying the individual probabilities of each event
- Probability of the intersection of independent events A and B: $P(A \cap B) = P(A) \times P(B)$
- Probability of getting heads twice when flipping a fair coin: $0.5 \times 0.5 = 0.25$

• For mutually exclusive events, calculate the probability by adding the individual probabilities of each event

• Probability of the union of mutually exclusive events A and B: $P(A \cup B) = P(A) + P(B)$

• Probability of rolling an even or odd number on a fair six-sided die: $0.5 + 0.5 = 1$

## Sampling methods and event dependence

1. Simple random sampling gives each item in the population an equal chance of being selected - Selections are independent of each other (drawing names from a hat with replacement)

2. Systematic sampling selects items at regular intervals from a list - Selections may be dependent on the ordering of the list (choosing every 10th person from an alphabetical list)

3. Stratified sampling divides the population into subgroups (strata) based on a characteristic - Simple random sampling is performed within each stratum (dividing population by age groups and randomly sampling within each group) - Selections within each stratum are independent, but the overall sample may be dependent on the chosen strata

4. Cluster sampling divides the population into naturally occurring groups (clusters) - A random sample of clusters is selected, and all items within the selected clusters are included (randomly selecting city blocks and surveying all households within those blocks) - Selections within clusters are dependent on each other

## Additional Probability Concepts

• Set theory provides a foundation for understanding probability, including operations like union and intersection

• The law of total probability allows for calculating probabilities by considering all possible outcomes

• The complement rule states that the probability of an event not occurring is 1 minus the probability of it occurring

# 3.3 Two Basic Rules of Probability

Probability rules show how events interact and how to calculate their likelihood. These rules help you determine the chances of multiple events occurring together or at least one event happening among several possibilities.

The multiplication rule is used for calculating the probability of combined events, while the addition rule helps determine the likelihood of at least one event occurring. Together, they give you a clear method for solving probability problems in real-world scenarios.

## Probability Rules

## Fundamental Concepts

• Probability axioms form the foundation of probability theory

• Sample space represents all possible outcomes of an experiment

• An event is a subset of the sample space

• Complement of an event A is all outcomes in the sample space not in A

## Multiplication rule for probabilities

• Multiplication rule applies when calculating the probability of two or more events occurring together

• Independent events have no influence on each other's probability

• Probability of both independent events A and B occurring equals the product of their individual probabilities

• Formula: $P(A \text{ and } B) = P(A) \times P(B)$

• Example: Probability of rolling a 6 on a fair die ($\frac{1}{6}$) and flipping a head on a fair coin ($\frac{1}{2}$) equals $\frac{1}{6} \times \frac{1}{2} = \frac{1}{12}$

• Dependent events influence each other's probability

• Probability of both dependent events A and B occurring equals the probability of event A multiplied by the conditional probability of event B given event A has occurred

• Formula: $P(A \text{ and } B) = P(A) \times P(B|A)$

• $P(B|A)$ represents the probability of event B occurring given that event A has already occurred

• Example: Probability of drawing a heart from a standard deck of cards ($\frac{13}{52}$) and then drawing a king from the remaining cards ($\frac{4}{51}$) equals $\frac{13}{52} \times \frac{4}{51} = \frac{1}{51}$

## Addition rule in probability

• Addition rule applies when calculating the probability of at least one of two or more events occurring

• Mutually exclusive events cannot occur simultaneously

• Probability of either mutually exclusive event A or B occurring equals the sum of their individual probabilities

• Formula: $P(A \text{ or } B) = P(A) + P(B)$

• Example: Probability of rolling an even number ($\frac{1}{2}$) or a 3 ($\frac{1}{6}$) on a fair die equals $\frac{1}{2} + \frac{1}{6} = \frac{2}{3}$

• Non-mutually exclusive events can occur simultaneously

• Probability of either non-mutually exclusive event A or B occurring equals the sum of their individual probabilities minus the probability of both events occurring together

• Formula: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

• $P(A \text{ and } B)$ represents the probability of both events A and B occurring, calculated using the multiplication rule

• Example: Probability of drawing a heart ($\frac{13}{52}$) or a face card ($\frac{12}{52}$) from a standard deck equals $\frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{11}{26}$, as there are 3 face cards that are also hearts

## Probability rules for events

• Multiplication rule used when calculating the probability of two or more events occurring together ("and")

• Events can be independent or dependent

• Keywords: "and," "given," "if"

• Addition rule used when calculating the probability of at least one of two or more events occurring ("or")

• Events can be mutually exclusive or non-mutually exclusive

• Keywords: "or," "either," "at least one"

• Examples: 1. Multiplication rule (independent events): Probability of rolling a 4 on a fair die and then flipping a tail on a fair coin

• Conditional probability is used when events are dependent 2. Multiplication rule (dependent events): Probability of drawing a spade from a standard deck and then drawing a queen from the remaining cards 3. Addition rule (mutually exclusive events): Probability of rolling a prime number or a 6 on a fair die 4. Addition rule (non-mutually exclusive events): Probability of drawing a red card or a king from a standard deck

## Visualization Tools

• Venn diagrams help illustrate relationships between events in a sample space

• Probability trees are useful for visualizing sequential events and their probabilities

# 3.4 Contingency Tables and Probability Trees

Contingency tables and probability trees are powerful tools for visualizing and calculating probabilities. These methods help organize data and map out possible outcomes, making it easier to understand complex probability scenarios.

By using these techniques, you can calculate joint, marginal, and conditional probabilities. They also allow you to see relationships between variables and events, helping you make informed decisions based on probability calculations.

## Contingency Tables and Probability Trees

## Probabilities from contingency tables

• Contingency tables organize and display the frequency distribution of two categorical variables

• Rows represent one variable (gender) and columns represent the other (preferred color)

• Each cell shows the frequency or count of observations for a specific combination of the two variables (males who prefer blue)

• Joint probability calculates the probability of two events occurring simultaneously

• Calculate by dividing the frequency in a specific cell by the total number of observations

• Formula: $P(A \cap B) = \frac{frequency(A \cap B)}{total}$

• Example: $P(male \cap blue) = \frac{25}{100} = 0.25$

• Marginal probability calculates the probability of an event occurring regardless of the outcome of the other variable

• Calculate by summing the frequencies in a row or column and dividing by the total number of observations

• Formula for row marginal probability: $P(A) = \frac{frequency(A)}{total}$

• Example: $P(male) = \frac{60}{100} = 0.6$

• Formula for column marginal probability: $P(B) = \frac{frequency(B)}{total}$

• Example: $P(blue) = \frac{40}{100} = 0.4$

• Conditional probability calculates the probability of an event occurring given that another event has already occurred

• Calculate by dividing the joint probability by the marginal probability of the given event

• Formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

• Example: $P(male|blue) = \frac{0.25}{0.4} = 0.625$

• Helps determine if events are independent

## Construction of tree diagrams

• Tree diagrams visualize and represent probability problems graphically

• Each branch represents a possible outcome (heads or tails on a coin flip)

• Probabilities are written along the branches (0.5 for heads, 0.5 for tails)

• Start with an initial node and draw branches for each possible outcome

• Label each branch with the probability of that outcome occurring

• For subsequent stages, draw branches from each previous outcome

• Label these branches with conditional probabilities

• Example: Drawing a second coin flip after the first flip resulted in heads

• The sum of the probabilities of all branches stemming from a single node must equal 1

• Example: $P(heads) + P(tails) = 0.5 + 0.5 = 1$

• Branches represent the sample space of the probability experiment

## Interpretation of tree diagrams

• To find the probability of a specific path, multiply the probabilities along the branches leading to that outcome

• Example: $P(heads \cap heads) = 0.5 \times 0.5 = 0.25$

• To find the overall probability of an outcome, sum the probabilities of all paths leading to that outcome

• Example: $P(1\ heads) = 0.25 + 0.25 = 0.5$

• Conditional probability can be calculated using the tree diagram

• Identify the paths that satisfy the given condition (paths with at least one heads)

• Sum the probabilities of these paths and divide by the probability of the given condition

• Example: $P(at\ least\ 1\ heads|1st\ flip\ heads) = \frac{0.5}{0.5} = 1$

• Bayes' Theorem can be applied using tree diagrams to update the probability of an event based on new information

• Formula: $P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$

• Example: $P(1st\ heads|at\ least\ 1\ heads) = \frac{0.5 \times 0.5}{0.75} = 0.333$

• Tree diagrams can illustrate the law of total probability

## Fundamental Probability Concepts

• Probability axioms form the foundation of probability theory

• Events can be mutually exclusive, meaning they cannot occur simultaneously

• The law of total probability states that the probability of an event is the sum of the probabilities of all possible ways for the event to occur

# 3.5 Venn Diagrams

Venn diagrams are powerful tools for visualizing set relationships and calculating probabilities. They use overlapping circles to show how different sets interact, making it easier to understand concepts like intersections and unions.

Probability rules, like addition and multiplication, help us calculate the likelihood of events occurring. These rules work hand-in-hand with Venn diagrams, allowing us to solve complex probability problems by breaking them down into simpler parts.

## Venn Diagrams and Probability

## Venn diagrams for set relationships

• Venn diagrams visually represent relationships between sets using overlapping circles or ovals

• Each circle or oval represents a set (set A, set B)

• Overlapping regions indicate elements belonging to multiple sets (intersection, $A \cap B$)

• Non-overlapping regions represent elements belonging to only one set (complement, $A'$ or $B'$)

• All sets are contained within a rectangular frame representing the universal set

• Venn diagrams calculate probabilities of events based on set relationships

• $P(A)$ probability of event A occurring (elements in circle A)

• $P(B)$ probability of event B occurring (elements in circle B)

• $P(A \cap B)$ probability of both events A and B occurring (overlapping region)

• $P(A \cup B)$ probability of either event A or B occurring (all elements in both circles)

• Examples of sets in Venn diagrams

• Students enrolled in math and science courses (overlapping region for students in both)

• Customers who purchase product X and product Y (intersection of product purchasers)

## Addition and multiplication rules in probability

- Addition rule calculates probability of either event A or B occurring: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Adds individual probabilities $P(A)$ and $P(B)$
- Subtracts intersection probability $P(A \cap B)$ to avoid double-counting overlapping region
- Example: Probability of selecting a red or blue marble from a bag
- Multiplication rule calculates probability of both events A and B occurring: $P(A \cap B) = P(A) \times P(B|A)$
- Multiplies probability of event A, $P(A)$, by conditional probability of event B given A, $P(B|A)$
- Used when events are dependent (occurrence of A affects probability of B)
- Example: Probability of drawing a king and then a queen from a deck of cards (without replacement)

## Independent vs dependent events

- Independent events: Occurrence of one event does not affect probability of the other
- $P(A|B) = P(A)$ and $P(B|A) = P(B)$ (conditional probabilities equal individual probabilities)
- Multiplication rule simplifies to $P(A \cap B) = P(A) \times P(B)$ for independent events
- Examples: Rolling a die twice (outcome of first roll doesn't affect second), flipping a fair coin
- Dependent events: Occurrence of one event affects probability of the other
- $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$ (conditional probabilities differ from individual probabilities)
- Multiplication rule $P(A \cap B) = P(A) \times P(B|A)$ or $P(B) \times P(A|B)$ for dependent events
- Examples: Drawing cards without replacement, selecting a defective item from a batch
- Conditional probability $P(A|B)$ is probability of event A occurring given that event B has already occurred
- Formula: $P(A|B) = \frac{P(A \cap B)}{P(B)}$, where $P(B) \neq 0$ (probability of B must be non-zero)
- Useful for updating probabilities based on new information or prior events
- Example: Probability of a student passing a test given that they studied for it

## Additional Set Concepts

- Disjoint sets: Sets with no common elements (their intersection is empty)
- Cardinality: The number of elements in a set, often denoted as |A| for set A
- Symmetric difference: Elements in either set, but not in their intersection (A $\triangle$ B = (A ∪ B) - (A ∩ B))

# Unit 4 – Discrete Random Variables

## 4.1 Hypergeometric Distribution

The hypergeometric distribution calculates probabilities for sampling without replacement from finite populations. It's used when the population size and number of successes are known, like drawing cards from a deck or selecting defective items from a batch.

This distribution differs from others in its handling of independence and sampling. Unlike the binomial distribution, which assumes constant probability and independence, the hypergeometric distribution accounts for changing probabilities as items are removed from the population.

## Hypergeometric Distribution

## Hypergeometric distribution calculations

- Calculates probabilities for scenarios involving sampling without replacement from a finite population
- Hypergeometric distribution formula:
- $P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
- $N$: total population size (number of items in a batch)
- $K$: number of successes in the population (defective items in the batch)
- $n$: sample size (items selected for inspection)
- $k$: number of successes in the sample (defective items found in the sample)
- Binomial coefficient $\binom{n}{k}$ calculated as:
- $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$
- $n!$ represents factorial of $n$, product of all positive integers ≤ $n$ (5! = 5 × 4 × 3 × 2 × 1)
- Steps to calculate probabilities using hypergeometric distribution: 1. Identify values for $N$, $K$, $n$, and $k$ 2. Substitute values into hypergeometric distribution formula 3. Calculate binomial coefficients using formula or calculator 4. Simplify fraction to obtain probability
- The cumulative distribution function can be used to calculate probabilities for ranges of values

## Applications of hypergeometric distribution

- Appropriate when:
- Sampling without replacement from finite population (cards from deck, defective items from batch)
- Population size known (52 cards in standard deck)
- Number of successes in population known (4 aces in deck)
- Sample size fixed (5-card hand)
- Examples of hypergeometric distribution applications:
- Drawing cards from deck without replacing them

• Selecting defective items from batch without replacing them

• Choosing fixed number of individuals with specific characteristic from population (10 left-handed people from group of 50)

• Used to model discrete random variables in scenarios with finite populations

## Hypergeometric vs other distributions

• Hypergeometric distribution differs from other probability distributions in independence and sampling

• Independence:

• Hypergeometric distribution: probability of success changes with each trial due to sampling without replacement

• Other distributions (binomial): probability of success constant for each trial due to sampling with replacement or assumed independence

• Sampling:

• Hypergeometric distribution: sampling without replacement from finite population

• Other distributions (binomial): assume sampling with replacement or infinite population

• Comparison with binomial distribution:

• Binomial distribution assumes constant probability of success for each trial and independence between trials (coin flips)

• Hypergeometric distribution used when population finite and sampling without replacement, leading to dependent trials (cards drawn from deck)

• Hypergeometric distribution incorporates the finite population correction factor, which accounts for the changing probability of success as items are removed from the population

## Additional Concepts

• Variance: Measures the spread of the hypergeometric distribution

• Conditional probability: The hypergeometric distribution can be used to calculate probabilities of events given that certain conditions have been met

# 4.2 Binomial Distribution

The binomial distribution is a powerful tool for analyzing scenarios with fixed trials and two possible outcomes. It helps calculate probabilities for events like coin flips, product defects, or survey responses, using key parameters like number of trials and success probability.

Understanding the binomial distribution's characteristics, mean, and standard deviation is crucial for real-world applications. From quality control to medical research, this concept allows us to make informed decisions based on probability calculations in various fields.

# Binomial Distribution

## Binomial distribution probability calculations

- Binomial distribution discrete probability distribution describes number of successes in fixed number of independent trials, each with same probability of success
- Formula for binomial probability mass function: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
- $n$: total number of trials (coin flips, product tests)
- $k$: number of successes (heads, defect-free products)
- $p$: probability of success on each trial (fairness of coin, manufacturing process reliability)
- Calculate probability of specific number of successes by substituting values for $n$, $k$, and $p$ into formula
- Probability of 3 heads in 5 coin flips with fair coin: $P(X = 3) = \binom{5}{3} (0.5)^3 (1-0.5)^{5-3}$
- Cumulative binomial probability calculated by summing individual probabilities for all values of $k$ up to desired number of successes
- $P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$
- Probability of 2 or fewer defective items in batch of 10 with 10% defect rate: $P(X \leq 2) = \sum_{i=0}^{2} \binom{10}{i} (0.1)^i (1-0.1)^{10-i}$
- Many calculators and software packages have built-in functions for calculating binomial probabilities (BINOM.DIST in Excel)

## Key characteristics of binomial experiments

- Binomial experiment characteristics:
- Fixed number of trials ($n$) (number of coin flips, product tests)
- Each trial independent of others (coin flips don't influence each other, products tested separately)
- Each trial has only two possible outcomes (success or failure) (heads or tails, defective or non-defective)
- Probability of success ($p$) constant for each trial (fairness of coin, manufacturing process consistency)
- Trials exhibit statistical independence
- Real-world applications of binomial distribution include:
- Quality control: Probability of certain number of defective items in batch (electronics manufacturing, food production)
- Marketing: Likelihood of specific number of customers responding to promotional offer (email campaign, product launch)
- Medical research: Probability of certain number of patients experiencing side effects from treatment (drug trials, surgical procedures)

• Finance: Probability of specific number of defaults in loan portfolio (mortgage lending, credit card issuing)

## Mean and standard deviation in binomial distributions

• Mean (expected value) of binomial distribution given by: $\mu = E(X) = np$

• $n$: total number of trials (survey respondents, product tests)

• $p$: probability of success on each trial (response rate, defect rate)

• Expected number of defective items in batch of 100 with 5% defect rate: $\mu = 100 \times 0.05 = 5$

• Standard deviation of binomial distribution given by: $\sigma = \sqrt{np(1-p)}$

• $n$: total number of trials (coin flips, patients in study)

• $p$: probability of success on each trial (heads, treatment effectiveness)

• Standard deviation of number of heads in 50 coin flips with fair coin: $\sigma = \sqrt{50 \times 0.5 \times (1-0.5)} \approx 3.54$

• Mean and standard deviation formulas used to: 1. Describe central tendency and dispersion of distribution 2. Calculate z-scores and percentiles for binomial distribution 3. Approximate binomial distribution with normal distribution when $n$ is large and $p$ not close to 0 or 1 (typically, when $np \geq 10$ and $n(1-p) \geq 10$)

## Additional Concepts in Probability Theory

• Random variable: A variable whose possible values are outcomes of a random phenomenon (e.g., number of successes in a binomial experiment)

• Probability distribution: A description of the probabilities associated with all possible values of a random variable (binomial distribution is an example)

• Mutually exclusive events: Events that cannot occur simultaneously (e.g., success and failure in a single trial of a binomial experiment)

• Combinatorics: The mathematical study of counting, arrangement, and combination of objects, which is crucial in calculating binomial probabilities

• Law of large numbers: As the number of trials in a binomial experiment increases, the observed proportion of successes tends to approach the true probability of success

# 4.3 Geometric Distribution

The geometric distribution models the number of trials until the first success in a series of independent events. It helps analyze scenarios like sales calls or product testing, where you want to know how long it may take to achieve a desired outcome.

Understanding this distribution helps predict the likelihood of success on a specific trial and the average number of attempts needed. The memoryless property is key, meaning past failures don't influence future probabilities, which simplifies calculations in many real-world applications.

# Geometric Distribution

## Geometric distribution probability calculations

- Models the number of trials until the first success in a series of independent Bernoulli trials
- Bernoulli trials have only two possible outcomes (success or failure)
- Probability of success ($p$) remains constant for each trial
- Trials are independent, outcome of one trial does not affect probability of success in subsequent trials (coin flips, rolling dice)
- Probability mass function (PMF) for the geometric distribution: $P(X = x) = (1 - p)^{x-1} \cdot p$
- $X$ represents the number of trials until the first success
- $x$ is the specific number of trials (positive integer)
- $p$ is the probability of success on each individual trial
- Calculate the probability of the first success occurring on a specific trial by substituting the trial number for $x$ and the probability of success for $p$ in the PMF formula
- First success on 3rd trial with $p = 0.3$: $P(X = 3) = (1 - 0.3)^{3-1} \cdot 0.3 = 0.147$
- The failure probability (1 - p) is used in the PMF to account for the unsuccessful trials before the first success

## Geometric distribution parameter interpretation

- Mean (expected value) of a geometric distribution: $E(X) = \frac{1}{p}$
- $E(X)$ represents the average number of trials needed to achieve the first success
- $p$ is the probability of success on each individual trial
- Standard deviation of a geometric distribution: $\sigma = \sqrt{\frac{1-p}{p^2}}$
- $\sigma$ measures the spread or dispersion of the distribution around the mean
- Higher standard deviation indicates greater variability in the number of trials needed to achieve the first success
- Mean and standard deviation provide insights into expected behavior of the geometric distribution in real-world contexts
- Sales representative with 20% chance of making a sale on each call, mean number of calls until first sale: $\frac{1}{0.2} = 5$ calls
- Standard deviation indicates how much the actual number of calls might deviate from this average (variability in sales performance)
- The geometric distribution models the waiting time until the first success occurs in a sequence of trials

## Memoryless property of geometric distributions

- Geometric distribution exhibits the memoryless property, probability of success on the next trial is independent of the number of failures that have occurred before it
- Probability of success remains constant, regardless of the number of previous trials (waiting for a bus, customer arrivals)
- Memoryless property expressed mathematically: $P(X > m + n|X > m) = P(X > n)$
- Probability of waiting more than $m + n$ trials for the first success, given that $m$ failures have already occurred, equals the probability of waiting more than $n$ trials for the first success from the beginning
- Memoryless property simplifies certain probability calculations in the geometric distribution
- Number of previous failures does not need to be considered
- Fair coin tossed repeatedly until first heads appears, probability of getting heads on the next toss is always 0.5, regardless of the number of tails observed before it (memoryless property in action)

## Additional properties of the geometric distribution

- The cumulative distribution function (CDF) of the geometric distribution gives the probability of observing the first success within a certain number of trials
- The geometric distribution assumes independence between trials, meaning the outcome of one trial does not affect the outcomes of subsequent trials

# 4.4 Poisson Distribution

The Poisson distribution models rare events occurring in fixed intervals. It's used to calculate probabilities for things like customer arrivals or product defects. The formula uses the average event rate and desired number of occurrences to determine likelihood.

Poisson experiments have key characteristics like independent events and constant average rates. The distribution can also estimate binomial probabilities when there are many trials with low success rates. This simplifies calculations for large-scale scenarios with rare occurrences.

## Poisson Distribution

## Poisson distribution probability calculations

- Expresses probability of a given number of events occurring within a fixed interval of time or space (minutes, hours, days, area)
- Poisson probability mass function calculates the probability: $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$
- $\lambda$ average number of events per interval (expected value)
- $e$ mathematical constant (2.71828)
- $k$ number of events (non-negative integer)
- $k!$ factorial of $k$
- Calculate probability of a specific number of events occurring within an interval: 1. Identify average number of events per interval ($\lambda$) 2. Determine desired number of events ($k$) 3. Plug values into

Poisson probability mass function

• Examples:

• Probability of 3 customers arriving in a 10-minute interval with an average of 2 customers per 10 minutes

• Probability of 5 defects occurring in a batch of 1000 items with an average defect rate of 0.4%

## Key characteristics of Poisson experiments

• Events occur independently of each other (one event does not affect the probability of another)

• Average rate of occurrence remains constant over the interval (consistent average number of events per unit of time or space)

• Two events cannot occur at exactly the same instant (events are discrete)

• Appropriate when:

• Number of possible occurrences is large (many potential events)

• Probability of an event occurring is small (rare events)

• Events are independent of each other (no influence between events)

• Mean and variance of a Poisson distribution equal $\lambda$ (expected value)

• Examples:

• Number of phone calls received by a call center per hour

• Number of accidents at a busy intersection per day

• Follows the law of rare events, which states that the total number of events will follow a Poisson distribution if there are many opportunities for an event to occur, but the probability of occurrence for each opportunity is small

## Poisson as binomial distribution estimator

• Approximates binomial distribution under certain conditions:

• Number of trials ($n$) is large (many repetitions)

• Probability of success ($p$) is small (rare successes)

• Product of $n$ and $p$ is a constant ($\lambda$) (consistent average number of successes)

• When conditions met, Poisson distribution estimates binomial distribution:

• $\lambda = np$, $n$ number of trials, $p$ probability of success

• Poisson probability mass function approximates binomial probability mass function

• Advantages of Poisson approximation:

• Simplifies calculations for large $n$ and small $p$ (easier computation)

• Requires only one parameter ($\lambda$) instead of two ($n$ and $p$) (less information needed)

• Examples:

• Approximating the probability of 2 defective items in a large batch of 5000 with a 0.1% defect rate

• Estimating the probability of 4 successful sales calls out of 200 with a 1.5% success rate

## Poisson Processes and Related Concepts

• Homogeneous Poisson process: A process where events occur continuously and independently at a constant average rate

• Non-homogeneous Poisson process: A process where the rate of occurrence varies over time or space

• Intensity function: Describes the rate of occurrence in a non-homogeneous Poisson process

• Cumulative distribution function (CDF): Gives the probability that the number of events is less than or equal to a specific value

# Unit 5 – Continuous Random Variables

## 5.1 Properties of Continuous Probability Density Functions

Continuous probability distributions model real-world situations with infinitely many possible outcomes. They use probability density functions (PDFs) to represent the relative likelihood of values, with area under the curve giving probabilities for specific ranges.

Continuous distributions also use cumulative distribution functions (CDFs), which give probabilities for values below a certain point. Together, PDFs and CDFs help analyze complex data and make predictions in fields ranging from finance to physics.

## Continuous Probability Distributions

## Area under curve for probability

- In continuous probability distributions, the probability of a continuous random variable falling within a specific range is represented by the area under the probability density function (PDF) curve within that range
- The total area under the PDF curve always equals 1, representing the total probability of all possible outcomes ($X$ taking any value from $-\infty$ to $\infty$)
- The probability of a random variable taking on a specific value is 0, as there are infinitely many possible values in a continuous distribution (probability of $X = 3.14159...$ is 0)
- To find the probability of a random variable falling within a range $[a, b]$, calculate the definite integral of the PDF $f(x)$ from $a$ to $b$:
- $P(a \leq X \leq b) = \int_a^b f(x)dx$
- Example: For a standard normal distribution, $P(-1 \leq X \leq 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \approx 0.6827$
- The area under the curve can be calculated using integration techniques, such as the fundamental theorem of calculus or numerical methods (trapezoidal rule, Simpson's rule)

## Intervals in cumulative distribution functions

- The cumulative distribution function (CDF) of a continuous random variable $X$, denoted as $F(x)$, gives the probability that $X$ is less than or equal to a specific value $x$
- $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$, where $f(t)$ is the probability density function
- To calculate the probability of a random variable falling within an interval $[a, b]$ using the CDF:
- $P(a \leq X \leq b) = F(b) - F(a)$
- Example: For a standard normal distribution,
$P(-1 \leq X \leq 1) = F(1) - F(-1) \approx 0.8413 - 0.1587 = 0.6826$
- For the probability of a random variable being greater than a value $a$:
- $P(X > a) = 1 - F(a)$

• Example: For an exponential distribution with rate parameter $\lambda = 0.5$, $P(X > 2) = 1 - F(2) = 1 - (1 - e^{-0.5 \cdot 2}) \approx 0.3679$

• CDF properties: 1. Non-decreasing: $F(a) \leq F(b)$ for $a < b$ 2. Limits: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$ 3. Right-continuous: $\lim_{x \to a^+} F(x) = F(a)$ for all $a$ 4. Monotonicity: The CDF is a monotonically increasing function

## Density vs cumulative distribution functions

• Probability Density Function (PDF):

• Denoted as $f(x)$

• Represents the relative likelihood of a random variable taking on a specific value

• The area under the PDF curve between two points represents the probability of the random variable falling within that range

• Properties:

• Non-negative: $f(x) \geq 0$ for all $x$

• Total area under the curve is 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

• Does not directly give probabilities for specific values or ranges (area under a single point is 0)

• The support of the PDF is the set of all possible values the random variable can take

• Cumulative Distribution Function (CDF):

• Denoted as $F(x)$

• Represents the probability that a random variable is less than or equal to a specific value

• Can be obtained by integrating the PDF: $F(x) = \int_{-\infty}^{x} f(t)dt$

• Directly gives probabilities for specific values or ranges ($P(X \leq x) = F(x)$)

• Properties:

1. Non-decreasing: $F(a) \leq F(b)$ for $a < b$

2. Limits: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

3. Right-continuous: $\lim_{x \to a^+} F(x) = F(a)$ for all $a$

• The PDF and CDF are related by differentiation and integration:

• $F(x) = \int_{-\infty}^{x} f(t)dt$

• $f(x) = \frac{d}{dx}F(x)$ (for continuous CDFs)

• Example: For a standard normal distribution, the PDF is $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ and the CDF is $F(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$

## Probability Space and Transformations

• Probability space: A mathematical construct that models a real-world process consisting of events that occur randomly

• It includes the sample space, event space, and probability measure

• Transformation of variables: A method used to derive the probability distribution of a function of a random variable

• This technique is useful when dealing with functions of random variables or when changing coordinate systems

# 5.2 The Uniform Distribution

The uniform distribution assumes equal likelihood for all values within a range. It's used for scenarios like rolling dice or selecting random numbers. The probability density function is constant, and probabilities are calculated using simple formulas based on the range's endpoints.

Understanding the uniform distribution's properties is crucial for various applications. The mean represents the central value, while the standard deviation measures spread. These concepts are fundamental for analyzing and interpreting data in many real-world situations.

## The Uniform Distribution

## Uniform distribution probability calculations

• Continuous probability distribution assumes all values within a given range have an equal likelihood of occurring (rolling a fair die, selecting a random number between 1 and 10)

• Probability density function (PDF) of a uniform distribution is denoted as $f(x) = \frac{1}{b-a}$

• $a$ represents the lower endpoint of the range

• $b$ represents the upper endpoint of the range

• Function is only defined for values between $a$ and $b$ (values outside this range have a probability of 0)

• Calculate the probability of an event occurring within a specific range using the formula $P(c \leq X \leq d) = \frac{d-c}{b-a}$

• $c$ represents the lower bound of the event

• $d$ represents the upper bound of the event

• $a$ and $b$ are the endpoints of the uniform distribution (defining the range of possible values)

• Examples:

• Probability of selecting a number between 3 and 7 from a range of 1 to 10 is $\frac{7-3}{10-1} = \frac{4}{9}$

• Probability of a bus arriving between 5 and 10 minutes, given a uniform distribution between 0 and 15 minutes, is $\frac{10-5}{15-0} = \frac{1}{3}$

## Significance of uniform distribution endpoints

• Endpoints $a$ and $b$ define the range of possible values for the uniform distribution (minimum and maximum values)

• Probability of a value occurring outside the range of $a$ and $b$ is always 0 (impossible events)

• Width of the range $b - a$ directly affects the probability density

• Larger range results in a lower probability density (values are more spread out)

• Smaller range results in a higher probability density (values are more concentrated)

• Examples:

• Uniform distribution between 0 and 1 has a higher probability density than a distribution between 0 and 100

• Waiting time for a bus with endpoints of 5 and 10 minutes has a smaller range and higher probability density than a bus with endpoints of 5 and 30 minutes

## Mean and standard deviation of uniform distributions

• Mean of a uniform distribution is the average of the lower and upper endpoints, calculated using the formula $\mu = \frac{a+b}{2}$

• Represents the central tendency or expected value of the distribution

• Example: Mean of a uniform distribution between 10 and 20 is $\frac{10+20}{2} = 15$

• Standard deviation of a uniform distribution measures the dispersion of values, calculated using the formula $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

• Represents the average distance between each value and the mean

• Example: Standard deviation of a uniform distribution between 0 and 6 is $\sqrt{\frac{(6-0)^2}{12}} = \sqrt{3} \approx 1.73$

• Formulas allow you to determine the central tendency and dispersion of the distribution based on the given endpoints (summarizing key characteristics of the distribution)

## Additional Concepts in Uniform Distributions

• The range (b - a) of a uniform distribution determines the spread of possible values

• Equiprobability is a key characteristic, meaning all outcomes within the range are equally likely

• Cumulative distribution function (CDF) for a uniform distribution is used to calculate probabilities for intervals

• Inverse transform sampling utilizes the uniform distribution to generate random variables from other distributions

• Random number generation often relies on uniform distributions as a basis for creating pseudorandom sequences

# 5.3 The Exponential Distribution

The exponential distribution models the time until an event occurs or the time between events. In business statistics, it helps analyze customer arrivals, machine failures, and product lifetimes by connecting timing with probability.

One of the distribution's key features is its memoryless property, which means the probability of a future event is independent of time already passed. This makes it ideal for modeling systems with constant failure

rates, like electronic components or customer service scenarios.

## Properties and Applications of the Exponential Distribution

## Exponential distribution probability calculations

- Models time until an event occurs or time between events using continuous probability distribution
- Exponential random variable $X$ represents waiting time until event happens
- Probability density function (PDF): $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, where $\lambda > 0$ is rate parameter
- Cumulative distribution function (CDF): $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$
- Calculate probability of event occurring within specific time interval using CDF
- $P(X \leq x) = F(x) = 1 - e^{-\lambda x}$
- If average time between customer arrivals is 10 minutes ($\lambda = \frac{1}{10}$), probability next customer arrives within 5 minutes is $P(X \leq 5) = 1 - e^{-\frac{1}{10} \cdot 5} \approx 0.3935$
- Calculate probability of event occurring after specific time using complement of CDF
- $P(X > x) = 1 - F(x) = e^{-\lambda x}$
- Using same scenario, probability next customer arrives after 15 minutes is $P(X > 15) = e^{-\frac{1}{10} \cdot 15} \approx 0.2231$
- The time between successive events in a Poisson process is called the interarrival time and follows an exponential distribution

## Memoryless property of exponential distribution

- Exponential distribution has memoryless property, meaning probability of future event is independent of time already passed
- Mathematically, $P(X > s + t | X > s) = P(X > t)$ for all $s, t \geq 0$
- Probability of waiting additional time $t$ given you've already waited time $s$ is same as probability of waiting time $t$ from start
- Memoryless property suitable for modeling longevity of devices or systems with constant failure rate
- If light bulb has exponential lifetime distribution with average lifespan of 1000 hours ($\lambda = \frac{1}{1000}$), probability it lasts additional 500 hours given it's been in use for 200 hours is same as probability of new light bulb lasting 500 hours
- $P(X > 700 | X > 200) = P(X > 500) = e^{-\frac{1}{1000} \cdot 500} \approx 0.6065$
- Memoryless property also applies to time between events in Poisson process (customer arrivals, machine failures)

## Exponential vs Poisson distributions

- Exponential and Poisson distributions closely related, both describe occurrence of events over time
- Exponential distribution (continuous):
- Models time until event occurs or time between events

• PDF: $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, where $\lambda > 0$ is rate parameter

• CDF: $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$

• Mean: $E(X) = \frac{1}{\lambda}$

• Variance: $Var(X) = \frac{1}{\lambda^2}$

• Poisson distribution (discrete):

• Models number of events occurring in fixed interval of time or space

• Probability mass function (PMF): $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$, where $\lambda > 0$ is average number of events per interval

• Mean: $E(X) = \lambda$

• Variance: $Var(X) = \lambda$

• Relationship between exponential and Poisson distributions:

• If time between events follows exponential distribution with rate $\lambda$, then number of events in fixed time interval follows Poisson distribution with mean $\lambda t$, where $t$ is length of time interval

• If customer arrivals follow Poisson process with average of 6 customers per hour ($\lambda = 6$), then time between customer arrivals follows exponential distribution with rate $\lambda = 6$ per hour ($\frac{1}{10}$ per minute)

• Applications:

• Exponential distribution: Modeling lifetime of devices, waiting times between events, time until failure in reliability analysis

• Poisson distribution: Modeling number of rare events in fixed interval (defects in product batch, customers arriving at store, accidents at intersection)

## Additional Applications and Related Concepts

• Survival analysis: Uses exponential distribution to model time until an event of interest (e.g., failure, death) occurs

• Hazard function: Represents the instantaneous rate of failure at a given time, which is constant for the exponential distribution

• Exponential decay: Describes processes where a quantity decreases at a rate proportional to its current value, following an exponential distribution pattern

# Unit 6 – The Normal Distribution

## 6.1 The Standard Normal Distribution

The standard normal distribution is a powerful tool for analyzing data. It uses z-scores to measure how far data points are from the mean in terms of standard deviations. This lets us compare values across different datasets easily.

The Empirical Rule is a key feature, showing how data clusters around the mean. It tells us that 68% of data falls within one standard deviation, 95% within two, and 99.7% within three. This helps predict where most data will lie in a normal distribution.

## The Standard Normal Distribution

## Z-score calculation for data deviation

- Calculate z-score using formula $z = \frac{x - \mu}{\sigma}$
- $x$ represents individual data point value
- $\mu$ represents population mean
- $\sigma$ represents population standard deviation
- Z-score indicates number of standard deviations a data point is from the mean
- Positive z-score signifies data point above the mean (right side of distribution)
- Negative z-score signifies data point below the mean (left side of distribution)
- Example calculation: Given $x = 85$, $\mu = 75$, and $\sigma = 5$, $z = \frac{85 - 75}{5} = 2$
- Interpretation: The data point 85 is 2 standard deviations above the mean (75)

## Empirical Rule for normal distributions

- Empirical Rule (68-95-99.7 Rule) applies to normally distributed data
- 68% of data within 1 standard deviation of mean ($\mu \pm 1\sigma$)
- 95% of data within 2 standard deviations of mean ($\mu \pm 2\sigma$)
- 99.7% of data within 3 standard deviations of mean ($\mu \pm 3\sigma$)
- Calculate percentage of data within specific range using z-scores and Empirical Rule
- Example: Percentage of data between 70 and 80 with $\mu = 75$ and $\sigma = 5$
- Calculate z-scores: $z_{70} = \frac{70 - 75}{5} = -1$ and $z_{80} = \frac{80 - 75}{5} = 1$
- Range spans -1 to 1 standard deviations from mean
- Empirical Rule states 68% of data falls within this range (-1σ to 1σ)
- The area under the curve between two z-scores represents the probability of data falling within that range

## Interpretation of positive vs negative z-scores

- Z-scores represent relative position of data point compared to mean
- Positive z-score indicates data point above mean (right side)
- Example: $z = 1.5$ means data point is 1.5 standard deviations above mean
- Negative z-score indicates data point below mean (left side)
- Example: $z = -2$ means data point is 2 standard deviations below mean
- Absolute value of z-score represents distance from mean in standard deviations
- Larger absolute z-scores indicate data points further from mean
- Example: $z = 2.5$ further from mean than $z = 1.2$ (2.5σ vs 1.2σ)
- Z-scores enable comparison of relative positions across different normal distributions
- Example: $z = 1.5$ in distribution A is equivalent to $z = 1.5$ in distribution B

## Additional Concepts in Normal Distribution

- The probability density function describes the shape of the normal distribution curve
- The central limit theorem states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases
- Standard error is the standard deviation of the sampling distribution of a statistic
- A normal probability plot can be used to assess whether a dataset follows a normal distribution

# 6.2 Using the Normal Distribution

The standard normal distribution, with its iconic bell curve shape, is a powerful tool in statistics. It allows us to calculate probabilities for any normally distributed data by converting values to z-scores, which measure how far a data point is from the mean in terms of standard deviations.

Understanding the standard normal distribution helps us interpret data across different scales. By using z-scores and the standard normal table, we can find probabilities for specific values or ranges, making it easier to compare and analyze data from various normal distributions.

## The Standard Normal Distribution

## Probabilities from standard normal distribution

- Normal distribution with mean of 0 and standard deviation of 1 (also known as the bell curve)
- Z-scores measure number of standard deviations from mean
- Positive z-scores are values above mean
- Negative z-scores are values below mean
- Standard normal distribution table provides area (probability) to left of given z-score
- Table organized with z-scores in rows and additional decimal places in columns

• Find probability for specific z-score by locating row with integer and first decimal place, then move to column with second decimal place

• Value at intersection is probability (area) to left of z-score

• For probabilities to right of z-score, subtract table value from 1

• For probabilities between two z-scores, find areas to left of each z-score and subtract smaller area from larger area

## Conversion to z-scores

• Standardizing formula converts values from any normal distribution to z-scores: $z = \frac{x - \mu}{\sigma}$

• $x$ is value being converted

• $\mu$ is mean of original distribution

• $\sigma$ is standard deviation of original distribution

• Convert value to z-score by subtracting mean from value and dividing by standard deviation

• Z-scores allow comparison of values from different normal distributions on common scale

## Interpretation of normal curve area

• Total area under normal curve equals 1, representing 100% probability

• Area between any two points on curve represents probability of value falling within that range

• In standard normal distribution: 1. Approximately 68% of data falls within one standard deviation of mean ($\mu \pm 1\sigma$) 2. Approximately 95% of data falls within two standard deviations of mean ($\mu \pm 2\sigma$) 3. Approximately 99.7% of data falls within three standard deviations of mean ($\mu \pm 3\sigma$)

• Area to left of z-score represents probability of value being less than or equal to that z-score

• Area to right of z-score represents probability of value being greater than that z-score

## Additional Properties of the Normal Distribution

• Symmetry: The normal distribution is perfectly symmetrical about its mean

• Continuous probability distribution: The normal distribution is a continuous function, allowing for infinite possible values

• Central Limit Theorem: As sample size increases, the distribution of sample means approaches a normal distribution

• Percentile: The value below which a given percentage of observations fall in a normal distribution

• Cumulative distribution function: Gives the probability that a random variable is less than or equal to a specific value

## 6.3 Estimating the Binomial with the Normal Distribution

The normal distribution can be used to approximate binomial probabilities when certain conditions are met. This powerful tool simplifies calculations for large sample sizes and moderate probabilities of success, making it invaluable in various fields.

By understanding the requirements and applications of this approximation, you can efficiently estimate probabilities in real-world scenarios. This method bridges the gap between discrete and continuous probability distributions, offering a practical approach to complex problems.

# Estimating the Binomial with the Normal Distribution

## Normal approximation of binomial probabilities

- Binomial probabilities approximated using normal distribution when certain conditions met
- Calculate probabilities using normal approximation: 1. Find mean of binomial distribution $\mu = np$ ($n$ = sample size, $p$ = probability of success) 2. Find standard deviation of binomial distribution $\sigma = \sqrt{np(1-p)}$ 3. Standardize binomial random variable $X$ using Z-score $Z = \frac{X - \mu}{\sigma}$ 4. Use standard normal distribution table or calculator to find probability corresponding to Z-score
- Normal approximation useful for large sample sizes (typically $n \geq 30$) and probabilities not too close to 0 or 1 ($0.1 \leq p \leq 0.9$)
- Approximation accurate when $np \geq 5$ and $n(1-p) \geq 5$ (ensures sufficient expected successes and failures)
- Central Limit Theorem supports the use of normal approximation for large sample sizes

## Applications of normal distribution estimates

- Normal distribution accurately estimates binomial probabilities in various situations
- Suitable when sample size ($n$) large (typically $n \geq 30$) and probability of success ($p$) not too close to 0 or 1 ($0.1 \leq p \leq 0.9$)
- Product of $n$ and $p$ ($np$) and product of $n$ and $(1-p)$ ($n(1-p)$) must be greater than or equal to 5
- Conditions satisfied, shape of binomial distribution approximately normal (bell-shaped curve)
- Applications include quality control (defective products), marketing (customer preferences), and finance (loan defaults)

## Symmetry in binomial distributions

- Symmetry of binomial distribution depends on probability of success ($p$)
- Perfectly symmetric when $p = 0.5$ (equal chance of success and failure)
- Becomes more skewed as $p$ moves away from 0.5 (closer to 0 or 1)
- Normal approximation most accurate when binomial distribution nearly symmetric
- More accurate for $p$ close to 0.5 (balanced outcomes)
- Less accurate for $p$ close to 0 or 1, especially for smaller sample sizes (skewed distribution)
- Highly skewed binomial distributions ($p$ very close to 0 or 1) may require alternative methods (Poisson approximation) for better accuracy

## Properties of the Normal Distribution

- Probability density function describes the shape of the normal distribution curve

- Cumulative distribution function gives the probability of a value falling below a certain point
- Expected value of a normal distribution is equal to its mean
- Variance measures the spread of the distribution around the mean

# Unit 7 – The Central Limit Theorem

## 7.1 The Central Limit Theorem for Sample Means

Sampling distributions and the Central Limit Theorem are key concepts in statistics. They help you understand how sample statistics behave and relate to population parameters. These ideas make it possible to draw reasonable conclusions about populations from sample data.

The Central Limit Theorem is particularly powerful. It states that for large enough samples, the sampling distribution of the mean approaches a normal distribution, regardless of the population's shape. This allows us to use normal distribution properties for many statistical analyses.

## Sampling Distributions and the Central Limit Theorem

## Concept of sampling distribution

• Probability distribution of a sample statistic (sample mean, sample proportion) obtained from repeated sampling of a population

• Describes variability and behavior of the sample statistic across all possible samples of the same size from the population

• Supports statistical inference using sample data to make conclusions about the population

• Understanding properties of the sampling distribution (shape, center, spread) enables making probabilistic statements about the population parameter based on the sample statistic

• Quantifies uncertainty associated with using a sample statistic to estimate a population parameter

• Helps determine likelihood of obtaining a particular sample result if the population parameter takes on a specific value (hypothesis testing, confidence intervals)

• The sampling distribution is closely related to the probability density function of the underlying random variable

## Central Limit Theorem application

• States that under certain conditions, the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the shape of the population distribution

• Conditions for CLT to hold:

• Sample is randomly selected from the population

• Sample size is sufficiently large (typically $n \geq 30$)

• Samples are independent of each other

• When CLT applies, the sampling distribution of the sample mean has the following properties:

• Mean of the sampling distribution equals the population mean: $\mu_{\bar{x}} = \mu$

• Standard deviation of the sampling distribution (standard error) equals the population standard deviation divided by the square root of the sample size: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

• If population is normally distributed, the sampling distribution of the sample mean will also be normally distributed, regardless of the sample size (heights, IQ scores)

• The law of large numbers supports the CLT by stating that as the sample size increases, the sample mean converges to the true population mean

## Probabilities for sample means

• Calculate probabilities for sample means using the standardizing formula (z-score) to convert the sample mean to a standard normal distribution

• Standardizing formula for the sample mean: $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

• $\bar{x}$ is the sample mean

• $\mu_{\bar{x}}$ is the mean of the sampling distribution (equal to the population mean $\mu$)

• $\sigma_{\bar{x}}$ is the standard error of the sampling distribution ($\frac{\sigma}{\sqrt{n}}$)

• Once z-score is calculated, use the standard normal distribution table or calculator to find the probability associated with the z-score

• To find probability of obtaining a sample mean less than or equal to a specific value, calculate z-score for that value and find corresponding area to the left of the z-score in the standard normal distribution (left-tailed test)

• To find probability of obtaining a sample mean greater than or equal to a specific value, calculate z-score for that value and find corresponding area to the right of the z-score in the standard normal distribution (right-tailed test)

## Sampling and Probability Concepts

• Sampling error: The difference between the sample statistic and the true population parameter, which decreases as sample size increases

• Standard deviation: A measure of variability in a dataset that affects the spread of the sampling distribution

• Random variable: A variable whose possible values are determined by chance and follows a probability distribution

• Probability distribution: Describes the likelihood of different outcomes for a random variable, forming the basis for sampling distributions

# 7.2 Using the Central Limit Theorem

The Central Limit Theorem explains why sample means become more predictable as sample size grows. As long as the conditions are met, the distribution of sample means approaches a normal shape even when the original population is not normal.

This theorem is closely tied to the Law of Large Numbers, which says sample means get closer to the true population mean as samples get bigger. Together, these ideas support statistical inference, including confidence intervals and hypothesis tests based on sample data.

# Central Limit Theorem

## Central Limit Theorem for inferential statistics

• States sampling distribution of sample means approximates normal distribution as sample size increases, regardless of population distribution shape (uniform, skewed)

• Allows inferential statistics to be applied to non-normal populations (income data, test scores)

• Three main conditions:

• Samples must be independent and randomly selected from population

• Sample size must be sufficiently large (typically n ≥ 30)

• Population must have finite variance

• When conditions met, sampling distribution of sample means approximately normal, centered at population mean ($\mu$), with standard deviation equal to population standard deviation divided by square root of sample size ($\frac{\sigma}{\sqrt{n}}$)

• Enables use of inferential statistics, such as confidence intervals and hypothesis tests, even when population distribution not normal (heights, weights)

• The resulting normal distribution is also known as the standard normal distribution (bell curve)

## Law of Large Numbers and convergence

• States as sample size increases, sample mean converges to population mean

• Larger sample size, closer sample mean to true population mean (polling data, product ratings)

• Fundamental concept in probability theory and statistics, underlies Central Limit Theorem

• As sample size increases, variability of sample means decreases, causing sampling distribution to become more concentrated around population mean

• Can be demonstrated through simulations or by calculating sample means for increasing sample sizes and observing convergence to population mean (coin flips, dice rolls)

• Convergence of sample means to population mean, as stated by Law of Large Numbers, supports accurate inferences about population based on sample data

## Sample size impact on sampling distribution

• Standard deviation of sampling distribution, also known as standard error, directly affected by sample size

• Standard error calculated as population standard deviation divided by square root of sample size ($\frac{\sigma}{\sqrt{n}}$)

• As sample size (n) increases, standard error decreases (survey results, experiment data)

• Larger sample size results in smaller standard error, meaning sampling distribution becomes more concentrated around population mean

• Increased concentration leads to more precise estimates and narrower confidence intervals (political polls, medical studies)

• Conversely, smaller sample size results in larger standard error, causing sampling distribution to be more spread out and less concentrated around population mean

• Leads to less precise estimates and wider confidence intervals (pilot studies, small-scale experiments)

• Relationship between sample size and standard error important when designing studies and determining appropriate sample size to achieve desired level of precision in estimates

## Statistical Inference and Hypothesis Testing

• Z-score: Measures how many standard deviations an observation is from the mean

• Degrees of freedom: Number of independent values that can vary in statistical analysis

• Statistical significance: Likelihood that a relationship between two or more variables is caused by something other than chance

# 7.3 The Central Limit Theorem for Proportions

The Central Limit Theorem for proportions explains why sample proportions become easier to model as sample size grows. When samples are large enough, the distribution of sample proportions becomes approximately normal, no matter what the original population looks like.

This theorem supports estimates about population proportions. It helps you create confidence intervals and run hypothesis tests, which are tools for drawing conclusions from sample data about larger populations.

## The Central Limit Theorem for Proportions

## Central Limit Theorem for proportions

• States sampling distribution of sample proportion $\hat{p}$ is approximately normal when sample size is large enough (typically $n \geq 30$) regardless of population distribution shape

• Requires both $n \cdot p \geq 10$ and $n \cdot (1 - p) \geq 10$, where $n$ is sample size and $p$ is population proportion

• Mean of sampling distribution of $\hat{p}$ equals population proportion $p$

• Standard deviation of sampling distribution of $\hat{p}$ is $\sqrt{\frac{p(1-p)}{n}}$

• Sampling distribution of $\hat{p}$ becomes more normally distributed as sample size increases (Law of Large Numbers)

• Applies to binomial distributions, which model the number of successes in a fixed number of independent trials

## Mean and standard deviation of sampling distributions

• Mean of sampling distribution of $\hat{p}$: $\mu_{\hat{p}} = p$

• Standard deviation of sampling distribution of $\hat{p}$: $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

• Requires population proportion $p$ and sample size $n$

• When population proportion $p$ is unknown, estimate using sample proportion $\hat{p}$

• Estimated standard deviation: $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

• Larger sample sizes result in smaller standard deviations and more precise estimates of population proportion

## Confidence intervals for population proportions

• Formula: $\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

• $\hat{p}$ is sample proportion

• $z_{\alpha/2}$ is critical value from standard normal distribution for desired confidence level ($z_{0.025} = 1.96$ for 95% confidence)

• $n$ is sample size

• Provides range of plausible values for population proportion $p$ based on sample data

• Confidence level (90%, 95%, 99%) is probability that interval contains true population proportion

• Assumptions: 1. Random sample from population 2. Large enough sample size: $n \cdot \hat{p} \geq 10$ and $n \cdot (1-\hat{p}) \geq 10$ 3. Independent observations in sample

• Example: Survey of 500 customers finds 60% satisfaction. 95% confidence interval: $0.60 \pm 1.96 \cdot \sqrt{\frac{0.60(1-0.60)}{500}} = (0.552, 0.648)$

• The margin of error is represented by the term after the ± symbol in the confidence interval formula

## Statistical Inference and Hypothesis Testing

• Statistical inference uses sample data to draw conclusions about population parameters

• Hypothesis testing involves making decisions about population parameters based on sample data

• Both rely on probability theory to quantify uncertainty in estimates and decisions

• The Central Limit Theorem for proportions supports these methods when working with categorical data

# 7.4 Finite Population Correction Factor

When sampling from small populations, the finite population correction factor adjusts standard errors and variances when sample size exceeds 5% of the population. This gives more accurate estimates for means and proportions when sampling without replacement.

This factor accounts for reduced variability in larger samples relative to population size. It is used in scenarios like election polls, quality control, and market research, where sampling without replacement affects independence and estimation precision.

## Finite Population Correction Factor

## Finite population correction factor calculation

• Adjusts standard error of sampling distribution when sample size exceeds 5% of population size

• Formula: $\sqrt{\frac{N-n}{N-1}}$ where $N$ = population size and $n$ = sample size

• Apply to standard error of sampling distribution of means by multiplying $\sigma_{\bar{x}} \cdot \sqrt{\frac{N-n}{N-1}}$ where $\sigma_{\bar{x}}$ = standard error of sampling distribution of means

• Apply to standard error of sampling distribution of proportions by multiplying $\sigma_{\hat{p}} \cdot \sqrt{\frac{N-n}{N-1}}$ where $\sigma_{\hat{p}}$ = standard error of sampling distribution of proportions

• Reduces standard error as it accounts for the reduction in variability when a larger proportion of the population is sampled (election polls, small town surveys)

• Used when sampling without replacement from a finite population

## Adjustment of sampling variances

• Adjusts variance of sampling distribution when sample size exceeds 5% of population size

• Apply to variance of sampling distribution of means by multiplying $\sigma_{\bar{x}}^2 \cdot \frac{N-n}{N-1}$ where $\sigma_{\bar{x}}^2$ = variance of sampling distribution of means

• Apply to variance of sampling distribution of proportions by multiplying $\sigma_{\hat{p}}^2 \cdot \frac{N-n}{N-1}$ where $\sigma_{\hat{p}}^2$ = variance of sampling distribution of proportions

• Reduces variance as it accounts for the reduction in variability when a larger proportion of the population is sampled (quality control testing, market research on niche segments)

## Independence in population sampling

• Assumes each observation is independent of all other observations in the sample

• Violated when sample size exceeds 5% of population size as probability of selecting an individual changes after each selection due to decreasing population size (drawing cards without replacement, selecting marbles from a jar)

• Population proportion near 0 or 1 also affects independence as probability of selecting an individual with the characteristic of interest changes significantly after each selection, especially with larger sample sizes relative to population (rare disease testing, defect rates in manufacturing)

• Lack of independence can lead to biased estimates and incorrect conclusions about the population (overestimating support for an unpopular policy, underestimating failure rates of a product)

## Statistical Inference and Estimation

• Central limit theorem allows for inference about population parameters from sample statistics

• Margin of error quantifies the precision of estimates derived from samples

• Population parameters are estimated using corresponding sample statistics

• Finite population correction factor improves the accuracy of these estimates for small populations

# Unit 8 – Confidence Intervals

## 8.1 A Confidence Interval When the Population Standard Deviation Is Known or Large Sample Size

Confidence intervals help estimate population means using sample data. They work especially well when you know the population's standard deviation or have a large sample size. The Central Limit Theorem allows you to create these intervals, giving you a range of likely values for the true population mean.

Sample size and confidence level impact interval width. Larger samples give more precise estimates, while higher confidence levels widen intervals. Researchers must balance precision and certainty when choosing confidence levels, considering the tradeoffs between narrow and wide intervals.

## Confidence Intervals for Population Mean with Known Standard Deviation or Large Sample Size

## Confidence intervals using Central Limit Theorem

- Central Limit Theorem enables creating confidence intervals for population mean ($\mu$) when population standard deviation ($\sigma$) is known or sample size is large ($n \geq 30$)
- Confidence interval formula: $\bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$
- $\bar{x}$ represents sample mean (point estimate)
- $z_{\alpha/2}$ represents critical value from standard normal distribution based on desired confidence level
- $\sigma$ represents known population standard deviation
- $n$ represents sample size
- Find critical value ($z_{\alpha/2}$) using standard normal distribution table or calculator
- 95% confidence level: $\alpha = 0.05$ and $z_{\alpha/2} = 1.96$
- 99% confidence level: $\alpha = 0.01$ and $z_{\alpha/2} = 2.58$
- Confidence interval provides range of plausible values for population mean based on sample data
- Example: Estimating average height of students in a school with known standard deviation of 5 cm and sample mean of 170 cm ($n = 50$) at 95% confidence level
  - $170 \pm 1.96 \cdot \frac{5}{\sqrt{50}} = (168.6, 171.4)$

## Effects of sample size on intervals

- Sample size ($n$) impacts width of confidence interval
- Increasing sample size decreases width of confidence interval
- Larger sample sizes yield more precise estimates of population mean
- Example: Doubling sample size from 50 to 100 students in height example
  - $170 \pm 1.96 \cdot \frac{5}{\sqrt{100}} = (169.0, 171.0)$

• Interval width reduced from 2.8 cm to 2.0 cm

• Margin of error quantifies range of values added and subtracted from sample mean to create confidence interval

• Margin of error formula: $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

• Increasing sample size or decreasing confidence level reduces margin of error, resulting in narrower interval

## Tradeoffs in confidence level vs width

• Inverse relationship exists between confidence level and interval width

• Increasing confidence level widens interval, while decreasing confidence level narrows interval

• Higher confidence levels (99%) provide more certainty that true population mean falls within interval

• Increased certainty comes at cost of wider, less precise interval

• Lower confidence levels (90%) result in narrower intervals, providing more precise estimate of population mean

• Increased precision comes with lower confidence that true population mean falls within interval

• Researchers must balance desired confidence level with need for precision when selecting confidence level for study

• Example: Comparing 90%, 95%, and 99% confidence intervals for student height example ($n = 50$)

• 90% CI: $170 \pm 1.65 \cdot \frac{5}{\sqrt{50}} = (168.8, 171.2)$

• 95% CI: $170 \pm 1.96 \cdot \frac{5}{\sqrt{50}} = (168.6, 171.4)$

• 99% CI: $170 \pm 2.58 \cdot \frac{5}{\sqrt{50}} = (168.2, 171.8)$

## Statistical Concepts and Confidence Intervals

• Sampling distribution forms the basis for constructing confidence intervals

• Z-score measures the number of standard deviations a data point is from the mean

• Hypothesis testing uses confidence intervals to make inferences about population parameters

• Degrees of freedom affect the shape of the sampling distribution for small sample sizes

# Unit 9 – Hypothesis Testing: Single Sample

# Unit 10 – Two-Sample Hypothesis Testing

# Unit 11 – The Chi–Square Distribution

# Unit 12 – F Distribution and One-Way ANOVA

# Unit 13 – Linear Regression and Correlation