Statistical Modelling and Inferencing Complete Study Material

BITS Pilani Digital

Table of Contents

Continuous Random Variables

- Can take any value within a continuous range
- Examples: Height, weight, temperature
- Infinite possibilities in any interval<

Question: What is the probability of rain given there are dark clouds in the sky?

- A = rain occurs
- B = dark clouds present
- P(A|B) = how the probability of rain changes with evidence of dark clouds

Continuous Random Variables

- Can take any value within a continuous range
- Examples: Height, weight, temperature
- Infinite possibilities in any interval<

  Question: What is the probability of rain given there are dark clouds in the sky?

  - A = rain occurs
  - B = dark clouds present
  - P(A|B) = how the probability of rain changes with evidence of dark clouds

Random Variables and Distributions

Definition of Random Variables

A **random variable** is a variable that can take on any value randomly, with unknown outcomes in advance, but with known probabilities for each possible outcome.

Types of Random Variables

Discrete Random Variables

- Take on countable values
- Examples: Number of heads in coin flips, number of defects
- Can be listed: {0, 1, 2, 3, ...}

Continuous Random Variables

- Can take any value within a continuous range
- Examples: Height, weight, temperature
- Infinite possibilities in any interval

Probability Functions

Probability Mass Function (PMF)

Used for **discrete** random variables. Assigns a specific probability to each distinct outcome.
Example: Coin toss has P(Head) = 0.5, P(Tail) = 0.5

Probability Density Function (PDF)

Used for **continuous** random variables. The probability is for an interval or range of values, not a specific point.
Example: Probability of height being exactly 5.5 feet is zero, but we can find probability of height being between 5.4 and 5.6 feet.

Discrete Probability Distributions

Binomial Distribution

Models the number of successes in a fixed number of independent trials, where each trial has the same probability of success.

Key Characteristics

Fixed number of trials (n)

- Each trial has only two outcomes: success or failure
- Probability of success (p) remains constant
- All trials are independent

Parameters

- **n:** number of trials
- **p:** probability of success on each trial

Binomial Formula:

```
P(X = k) = C(n,k) × p^k × (1-p)^(n-k)where C(n,k) is the number of combinations
```

Properties

- **Mean:** $\mu = np$
- **Variance:** $\sigma^2 = np(1-p)$
- **Standard deviation:** $\sigma = \sqrt{[np(1-p)]}$

Example: Getting 70 Heads in 100 Coin Flips

Given parameters:

- n = 100 (trials)
- p = 0.5 (fair coin)

- k = 70 (desired heads)

Applying the Binomial Formula:

```
P(X = 70) = C(100,70) × (0.5)^70 × (1-0.5)^(100-70)
P(X = 70) = C(100,70) × (0.5)^70 × (0.5)^30P(X = 70) = C(100,70) × (0.5)^100Where C(100,70) = 100!/(70! × 30!)
```

Distribution properties:

- Expected heads: μ = 100 × 0.5 = 50
- Standard deviation: σ = √[100 × 0.5 × 0.5] = 5
- Note: 70 heads is 4 standard deviations above the mean, making it extremely unlikely

Poisson Distribution

Models the number of events occurring within a fixed interval of time or space, when you know the average rate of occurrence.

Key Requirements

- ## Two pieces of information: average (λ) and interval
- Events occur independently
- Average rate is constant

Poisson Formula:

```
P(X = k) = (e^(-λ) × λ^k) / k!where λ is the average number of events per interval
```

Properties

Unique Property: For Poisson distribution, Mean = Variance = λ

Example 1: Phone Calls

Average of 15 calls per day. What's the probability of getting exactly 20 calls?

- λ = 15 (average)

- k = 20 (desired number)
- P(X = 20) = (e^(-15) × 15^20) / 20!

Example 2: Bakery Planning

A bakery serves an average of 100 customers per day. During a festival, what's the probability that 200 customers will visit?

- λ = 100 (average customers per day)
- This is useful for planning resources and inventory

Example 3: Call Center

Call center employees receive an average of 80 calls per day. Estimate calls in next 2-3 hours using Poisson distribution for better planning and staffing.

Continuous Probability Distributions

Normal Distribution

The most important continuous distribution, characterized by its bell-shaped curve. Many natural phenomena follow this distribution.

Example: Heights of People in India

If we plot heights of the population, we get a bell-shaped curve where:

- Very few people are extremely short or tall
- Most people cluster around the average height
- We can find probability of height being between any two values

Parameters

- **Mean (μ):** Center of the distribution
- **Standard Deviation (σ):** Spread of the distribution

Normal PDF:

```
f(x) = (1 / (σ√(2π))) × e^(-1/2 × ((x-μ)/σ)²)
```

The Empirical Rule (68-95-99.7 Rule)

- **68%** of values lie within 1 standard deviation (μ ± σ)
- **95%** of values lie within 2 standard deviations (μ ± 2σ)
- **99.7%** of values lie within 3 standard deviations (μ ± 3σ)

Parametric vs Non-Parametric Methods

Parametric Methods

Statistical methods that make assumptions about the population distribution and focus on estimating specific parameters.

Advantages

- Higher statistical power
- More likely to detect true effects
- More precise estimates

Disadvantages

- Very sensitive to violations
- Like a tailored suit - fits specific situations only
- May give misleading results if assumptions are wrong

Non-Parametric Methods

Distribution-free methods that make no assumptions about the population distribution.

Advantages

- Very robust and flexible
- Work with any data distribution
- Perfect for ordinal/ranked data

Disadvantages

- Lower statistical power
- Less reliable for detecting effects
- Less precise estimates

When to use which method:

Decision Framework

1. **Ordinal/Ranked data:** Use Non-Parametric
2. **Normally distributed data:** Use Parametric
3. **Non-normally distributed data:** Use Non-Parametric

Example: Competition Judging

When judges rank contestants as 1st, 2nd, 3rd place (ordinal data), non-parametric tests work perfectly because they handle ranked data effectively.

Summary

Key Takeaways:

- **Random Variables:** Discrete (countable) vs Continuous (infinite range)
- **Probability Functions:** PMF for discrete, PDF for continuous
- **Distributions:**
  - Binomial: Fixed trials, binary outcomes
  - Poisson: Events in fixed intervals, known average
  - Normal: Bell curve, defined by mean and standard deviation
- **Methods:** Parametric (assumes distribution) vs Non-parametric (distribution-free)

This foundation prepares us for statistical inference, hypothesis testing, and advanced modeling techniques.

"> learning-objectives

Lesson: Foundations of Statistical Inference

Foundations of Statistical Inference:

A Detailed Guide to Sampling, Estimation, and the Central Limit Theorem

Contents

1 The Why and How of Sampling

Statistical inference is the art and science of making conclusions about a whole population based on a small piece of it. This process begins with the fundamental act of sampling.

1.1 Populations, Samples, Parameters, and Statistics

- **Population:** The entire group of individuals or objects we wish to study. For example, all students at BITS Pilani, all smartphones produced by a factory, or all voters in India.
- **Parameter:** A numerical value describing a characteristic of the population. Parameters are typically unknown constants that we want to estimate. We denote them with Greek letters (e.g., μ for mean, σ for standard deviation, p for proportion).
- **Sample:** A subset of the population that we actually collect data from.
- **Statistic:** A numerical value describing a characteristic of the sample. We calculate statistics from our data. We denote them with Roman letters (e.g., $\bar{x}$ for sample mean, s for sample standard deviation, $\hat{p}$ for sample proportion).

The core goal is to use a known statistic to infer the value of an unknown parameter. For this leap of faith to be valid, the sampling method is critical.

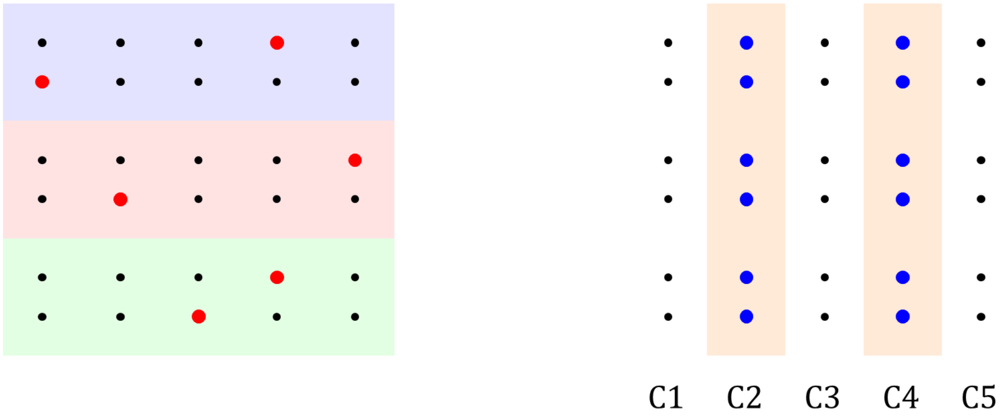
1.2 Types of Sampling Methods

A good sample should be representative of the population. The best way to achieve this is through probability sampling, where every member of the population has a known chance of being selected.

- **Simple Random Sampling (SRS):** Every individual and every possible sample of size n has an equal chance of being selected. This is the ideal gold standard.
- **Stratified Sampling:** The population is first divided into non-overlapping subgroups, or strata, based on a shared characteristic (e.g., year of study, campus location). A simple random sample is then taken from each stratum. This ensures representation from all key subgroups.
- **Cluster Sampling:** The population is divided into subgroups, or clusters, often based on geography. We then randomly select entire clusters and sample every individual within the chosen clusters. This is often more practical and cost-effective than SRS.
- **Systematic Sampling:** A starting point is chosen randomly, and then every k-th individual is selected from a list.
- **Convenience Sampling:** Selecting individuals who are easiest to reach. While simple, it is highly prone to bias and is not a valid method for statistical inference.

2 Point Estimation: The Best Single Guess

Stratified Sampling Cluster Sampling



Sample from every stratum Take all from selected clusters

Figure 1: A clean comparison of Stratified and Cluster Sampling.

A point estimate is our best single-value guess for an unknown population parameter, based on our sample data. For example, x̄ is the point estimate for μ.

2.1 Properties of Good Estimators

- **Unbiasedness:** An estimator is unbiased if the mean of its sampling distribution is equal to the true value of the parameter being estimated. The sample mean x̄ is an unbiased estimator of μ.
- **Efficiency:** An efficient estimator is one that has a small variance in its sampling distribution. A more efficient estimator is more likely to produce an estimate close to the true parameter value.

2.2 The Inherent Limitation: Sampling Error

The great weakness of a point estimate is that it provides no information about its precision. It is a single number that is almost guaranteed to be at least slightly wrong. To overcome this, we must understand how this estimate behaves across many samples.

3 The Central Limit Theorem (CLT)

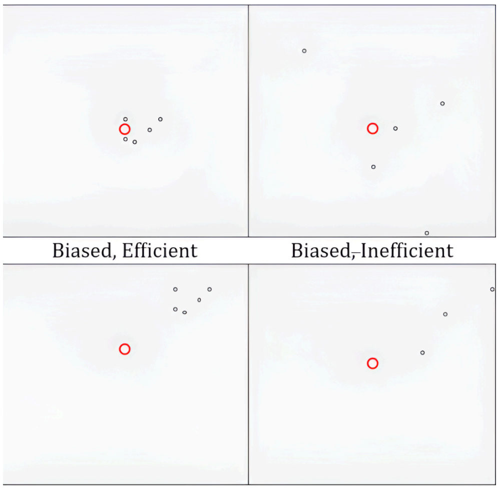


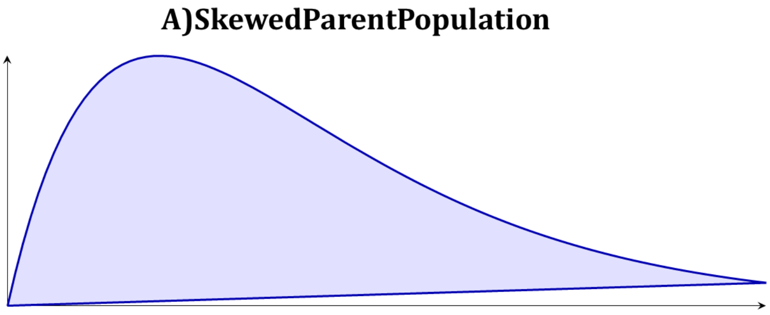
Figure 2: The target shooting analogy for estimators. The red bullseye is the true parameter. Each black dot is a point estimate from a different sample.

The Central Limit Theorem is arguably the most important result in all of statistics. It describes the shape, center, and spread of the sampling distribution of the sample mean, and it works even when the original population is not normally distributed.

The Central Limit Theorem States:

- **Center:** Its mean will be equal to the population mean ($\mu\bar{x} = \mu$).
- **Spread:** Its standard deviation (the standard error) will be $\sigma/\sqrt{n}$.
- **Shape:** It will be approximately normally distributed.

A common rule of thumb considers $n \geq 30$ to be "sufficiently large." This allows us to use the mathematics of the normal distribution to make inferences about the population mean with large samples.



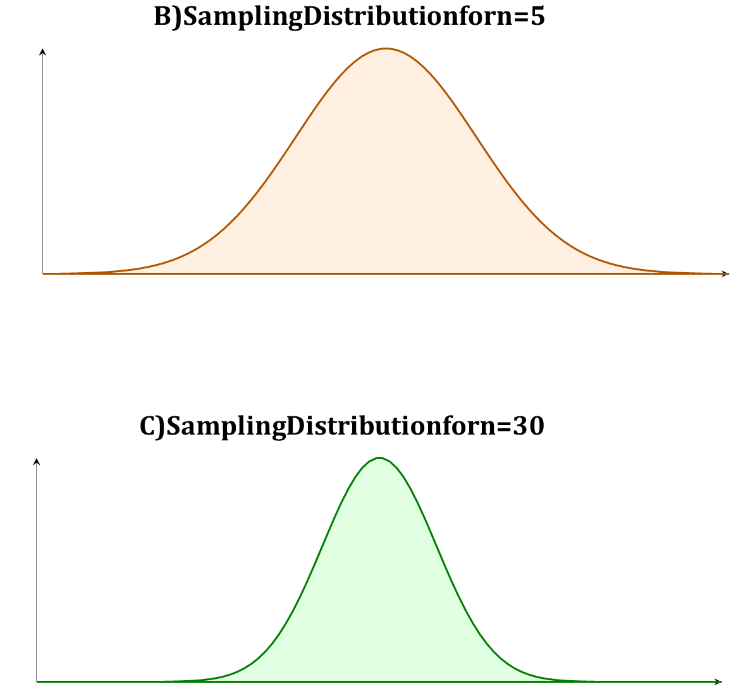


Figure 3: Visualizing the CLT. As sample size n increases, the distribution of sample means becomes narrower and more normal, regardless of the parent population’s shape.

4 Interval Estimation: A Range of Confidence

Now that we know the sampling distribution of ¯x is approximately normal (thanks to the CLT), we can now build a confidence interval around it.

4.1 Structure of a Confidence Interval

$CI = \bar{x} \pm \text{Margin of Error} = \bar{x} \pm (\text{Critical Value} \times \text{Standard Error})$

4.2 Case 1: When σ is Known (The Z-Interval)

Standard error = $\sigma/\sqrt{n}$, critical value comes from the standard normal (Z) distribution. $CI = \bar{x} \pm Z\alpha/2 \times (\sigma/\sqrt{n})$

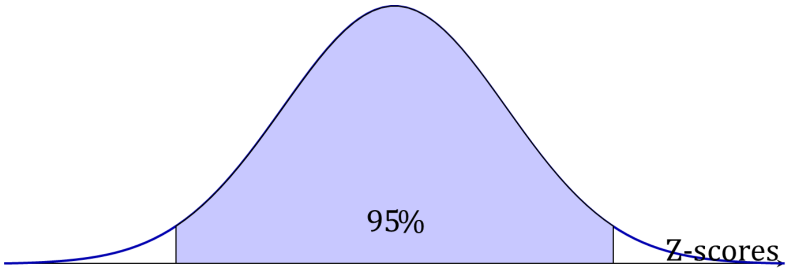


Figure 4: A 95% Confidence Interval on the Z-distribution.

4.3 Case 2: When σ is Unknown (The T-Interval)

Standard error = $s/\sqrt{n}$. Critical values come from the t-distribution with $df = n-1$. The t-distribution has fatter tails, creating wider intervals, especially for small n.

4.4 The Correct Interpretation: Net Fishing

Imagine the true population mean μ is a stationary fish in a pond. A confidence interval is like a net you throw into the pond. A 95% confidence level means that in the long run, 95% of the nets will capture the fish. It does not mean there is a 95% chance that your specific net has caught the fish. Confidence is in the process, not in the single outcome.

"> lesson-foundations-of-statistical-inference

Inferential Statistics

Week 2 Reading Material

Core Concept: Inferential statistics is the process of learning about a population through sampling. Rather than examining every individual in a population (which is often impractical), we take representative samples and use them to make inferences about the entire population.

Key Analogy: Like tasting a spoonful of soup to judge the entire pot, we use samples to understand populations.

1. Understanding Population vs. Sample

Population Parameter vs. Sample Estimate

- **Population Parameter:** The true value we want to know (e.g., average height of all Indians)
- **Sample Estimate:** Our approximation based on a subset (e.g., average height of 500 randomly selected Indians)

> **Example:** To find the average height of 1.4 billion Indians, instead of measuring everyone (time-consuming, expensive, often infeasible), we measure a sample of 1,000 people and use that to estimate the population average. >2. The Art of Sampling
>
> ### Why Sample?

Three main reasons to use sampling instead of complete enumeration:

- Time efficiency: Examining entire populations takes too long
- Resource conservation: Full population studies are expensive
- Feasibility: Sometimes impossible to access entire population

Types of Sampling

A. Probability Sampling

Every member of the population has a known, non-zero chance of being selected.

Simple Random Sampling

Every individual has an equal chance of selection, like drawing names from a hat.

Stratified Sampling

Divide population into groups (strata) first, then sample from each group. Example: Dividing India by states, then sampling from each state.

Systematic Sampling

Select every nth individual (e.g., every 5th person in a line).

⚠ **Important Caveat:** Avoid systematic sampling when dealing with periodic patterns. For instance, if you always sample on the 7th day of each week (always a Sunday), you might miss weekly variations in behavior, leading to biased results. The sampling interval should not align with natural cycles in your data.

Cluster Sampling

Two-step randomness: (1) Randomly select clusters/groups, (2) Sample within selected clusters. Example: Randomly select 5 states, then randomly sample within those states.

B. Non-Probability Sampling

Convenience-driven but prone to bias.

Convenience Sampling

Select whoever is easily accessible (e.g., surveying only your neighbors).

Quota Sampling

Set quotas for different groups, then conveniently fill them (e.g., first 50 males and first 50 females you encounter).

Snowball Sampling

Participants recruit other participants, creating a chain effect. Useful for hard-to-reach populations.

3. Sample Size Determination: The Goldilocks Problem

Not too big, not too small! Sample size depends on three factors:

- **Confidence Level:** How sure do you want to be? Higher confidence = larger sample needed
- **Margin of Error:** How much inaccuracy can you tolerate? Less error tolerance = larger sample needed

- **Population Variability:** How diverse is your population?
  - → More variation = larger sample needed
  - → Less variation (homogeneous/similar population) = smaller sample needed

Historical Example of Sampling Bias: 1936 U.S. Presidential Election

A survey of 2.4 million people predicted the wrong winner because samples were drawn only from car owners and telephone users, who represented a wealthy subset of the population. This demonstrates how **biased sampling leads to biased results**, regardless of sample size.

4. Point Estimation

Point estimation provides a single "best guess" value for a population parameter. While convenient, remember: the best guess may sometimes be wrong!

Common Point Estimates

- **Population Mean:** Average value (e.g., average age, average height)
- **Population Proportion:** Percentage or ratio (e.g., % voting for a party)
- **Population Variance:** Measure of spread in the data

Properties of Good Estimators

1. Unbiased

On average, the estimator should hit the true value. Individual estimates may vary, but they should center around the truth.

2. Efficient (Consistent)

Less variability in estimates; they cluster tightly around the true value rather than being scattered.

Think of it like archery:

- **Unbiased:** Your arrows center around the bullseye (even if individual shots vary)
- **Efficient:** Your arrows are tightly grouped (not scattered all over)

5. Understanding Degrees of Freedom

Why n-1 in Variance Formula?

Variance Formula: Variance = $\Sigma(x - \text{mean})^2 / (n - 1)$

Football Analogy: Imagine filling 10 positions on a football field with a fixed formation (4 forwards, 3 mids, 3 defenders):

- First 9 players can freely choose from available positions
- The 10th player has no choice; only one position remains

Similarly, when calculating variance using the sample mean, once n-1 deviations are determined, the last one is fixed (since all deviations must sum to zero). This "loss of freedom" is why we divide by n-1 instead of n, ensuring an unbiased estimate of population variance.

Key Takeaways

- Good sampling is crucial for valid inferences about populations
- Probability sampling reduces bias but requires more effort
- Sample size depends on desired confidence, acceptable error, and population variability
- Point estimates give single values but come with uncertainty
- Good estimators should be both unbiased and efficient

"> week-2-reading-material

The Hypothesis Testing Framework

1 The Core Idea: Making Decisions with Data
Hypothesis testing is a formal procedure for using sample data to evaluate a claim about a population. It provides a structured framework for making decisions in the face of uncertainty. The core logic of hypothesis testing is analogous to a criminal trial.

1.1 The Courtroom Analogy

- **The Presumption of Innocence:** In a trial, the defendant is presumed innocent until proven guilty. This is the default position, the status quo.
- **The Burden of Proof:** The prosecution must present compelling evidence to convince the jury to overturn this presumption of innocence.
- **The Verdict:** The jury does not prove the defendant is innocent; they either find enough evidence to convict (”guilty”) or not enough evidence (”not guilty”).

In statistics, the roles are played by:

- **The Null Hypothesis (H0):** This is the ”presumption of innocence.” It is a statement of no effect, no difference, or the status quo. We assume H0 is true unless the evidence suggests otherwise. Example: A new drug has no effect on recovery time. (μnew = μold)
- **The Alternative Hypothesis (HA or H1):** This is the claim the researcher is trying to find evidence for. It is the ”guilty” verdict. Example: The new drug reduces recovery time. (μnew < μold)

- **The Sample Data:** This is the ”evidence” we collect.
- **The Conclusion:** We don’t ”prove” the null hypothesis. We either find enough evidence to reject the null hypothesis in favor of the alternative, or we fail to reject the null hypothesis due to insufficient evidence.

2 The 4-Step Hypothesis Testing Framework

Every formal hypothesis test follows a consistent, four-step process.

2.1 Step 1: State the Hypotheses

First, clearly state the null (H0) and alternative (HA) hypotheses in terms of the popula tion parameter(s) of interest (e.g., μ,p,σ2). The alternative hypothesis can be one-tailed (<,>) or two-tailed (≠).

- H0 : μ = 100
- HA : μ≠ 100 (Two-tailed)

2.2 Step 2: Set the Decision Criteria

Before collecting data, we must define what constitutes ”strong evidence.” This is done by setting a significance level, α.

- α is the probability of making a Type I error (rejecting a true null hypothesis).
- It is our threshold for ”reasonable doubt.”
- Common values for α are 0.05 (5%), 0.01 (1%), and 0.10 (10%).

2.3 Step 3: Collect Data and Calculate the Test Statistic

After setting the criteria, we collect our random sample and summarize the evidence into a single number called a test statistic. The general form of a test statistic is:

Test Statistic = (Sample Statistic − Null Hypothesis Value) / Standard Error

This value measures how many standard errors our sample result is from the value claimed in the null hypothesis. Common test statistics are the Z-statistic, t-statistic, χ2-statistic, and F-statistic.

2.4 Step 4: Make a Decision

Finally, we use the test statistic to determine how likely our observed sample result is, assuming the null hypothesis is true. This likelihood is quantified by the p-value.

- **P-value:** The probability of observing a test statistic as extreme or more extreme than the one calculated, given that the null hypothesis (H0) is true.

We then compare our p-value to our pre-determined significance level, α.

The Decision Rule:

- If p-value ≤ α, the result is statistically significant. The evidence is strong enough to reject the null hypothesis.
- If p-value > α, the result is not statistically significant. The evidence is not strong enough; we fail to reject the null hypothesis.

We then state our conclusion in the context of the original research question.

reading-material-the-hypothesis-testing-framework

Errors, P-values, and Significance

1 The Nature of Statistical Decisions

Hypothesis testing uses sample data to

- **Random Experiment:** An experiment with random outcomes (e.g., coin flip)
- **Sample Space:** All possible outcomes of a random experiment
- **Event:** A subset of the sample space

Example: Rolling a Six-sided Die

- **Sample Space:** {1, 2, 3, 4, 5, 6}
- **Event (odd numbers):** {1, 3, 5}
- **Probability of odd number:** 3/6 = 0.5
- **Probability of even number:** 3/6 = 0.5

Law of Large Numbers

As you repeat an experiment multiple times, the observed probability approaches the theoretical probability. For example, flipping a fair coin many times will result in approximately 50% heads and 50% tails.

Conditional Probability

Conditional Probability is the probability of occurrence of an event given another event has already occurred.

$$P(A|B) = P(A \cap B) / P(B)$$

Example: Weather Prediction

Question: What is the probability of rain given there are dark clouds in the sky?

- A = rain occurs
- B = dark clouds present
- P(A|B) = how the probability of rain changes with evidence of dark clouds

Random Variables and Distributions

Definition of Random Variables

A **random variable** is a variable that can take on any value randomly, with unknown outcomes in advance, but with known probabilities for each possible outcome.

Types of Random Variables

Discrete Random Variables

- Take on countable values
- Examples: Number of heads in coin flips, number of defects
- Can be listed: {0, 1, 2, 3, ...}

Continuous Random Variables

- Can take any value within a continuous range
- Examples: Height, weight, temperature
- Infinite possibilities in any interval

Probability Functions

Probability Mass Function (PMF)

Used for **discrete** random variables. Assigns a specific probability to each distinct outcome.
Example: Coin toss has P(Head) = 0.5, P(Tail) = 0.5

Probability Density Function (PDF)

Used for **continuous** random variables. The probability is for an interval or range of values, not a specific point.
Example: Probability of height being exactly 5.5 feet is zero, but we can find probability of height being between 5.4 and 5.6 feet.

Discrete Probability Distributions

Binomial Distribution

Models the number of successes in a fixed number of independent trials, where each trial has the same probability of success.

Key Characteristics

- ## Fixed number of trials (n)

- Each trial has only two outcomes: success or failure
- Probability of success (p) remains constant
- All trials are independent

Parameters

- **n:** number of trials
- **p:** probability of success on each trial

Binomial Formula:

```
P(X = k) = C(n,k) × p^k × (1-p)^(n-k)where C(n,k) is the number of
combinations
```

Properties

- **Mean:** $\mu = np$
- **Variance:** $\sigma^2 = np(1-p)$

- **Standard deviation:** σ = √[np(1-p)]

Example: Getting 70 Heads in 100 Coin Flips

Given parameters:

- n = 100 (trials)
- p = 0.5 (fair coin)
- k = 70 (desired heads)

Applying the Binomial Formula:

```
P(X = 70) = C(100,70) × (0.5)^70 × (1-0.5)^(100-70)
P(X = 70) = C(100,70) × (0.5)^70 × (0.5)^30P(X = 70) = C(100,70) ×
(0.5)^100Where C(100,70) = 100!/(70! × 30!)
```

Distribution properties:

- Expected heads: μ = 100 × 0.5 = 50
- Standard deviation: σ = √[100 × 0.5 × 0.5] = 5
- Note: 70 heads is 4 standard deviations above the mean, making it extremely unlikely

Poisson Distribution

Models the number of events occurring within a fixed interval of time or space, when you know the average rate of occurrence.

Key Requirements

- ## Two pieces of information: average (λ) and interval

- Events occur independently
- Average rate is constant

Poisson Formula:

```
P(X = k) = (e^(-λ) × λ^k) / k!where λ is the average number of events
per interval
```

Properties

Unique Property: For Poisson distribution, Mean = Variance = λ

Example 1: Phone Calls

Average of 15 calls per day. What's the probability of getting exactly 20 calls?

- λ = 15 (average)
- k = 20 (desired number)
- P(X = 20) = (e^(-15) × 15^20) / 20!

Example 2: Bakery Planning

A bakery serves an average of 100 customers per day. During a festival, what's the probability that 200 customers will visit?

- λ = 100 (average customers per day)
- This is useful for planning resources and inventory

Example 3: Call Center

Call center employees receive an average of 80 calls per day. Estimate calls in next 2-3 hours using Poisson distribution for better planning and staffing.

Continuous Probability Distributions

Normal Distribution

The most important continuous distribution, characterized by its bell-shaped curve. Many natural phenomena follow this distribution.

Example: Heights of People in India

If we plot heights of the population, we get a bell-shaped curve where:

- Very few people are extremely short or tall
- Most people cluster around the average height
- We can find probability of height being between any two values

Parameters

- **Mean (μ):** Center of the distribution
- **Standard Deviation (σ):** Spread of the distribution

Normal PDF:

$$f(x) = (1 / (\sigma\sqrt{(2\pi)})) \times e\hat{}(-1/2 \times ((x-\mu)/\sigma)^2)$$

The Empirical Rule (68-95-99.7 Rule)

- **68%** of values lie within 1 standard deviation (μ ± σ)
- **95%** of values lie within 2 standard deviations (μ ± 2σ)
- **99.7%** of values lie within 3 standard deviations (μ ± 3σ)

Parametric vs Non-Parametric Methods

Parametric Methods

Statistical methods that make assumptions about the population distribution and focus on estimating specific parameters.

Advantages

- **Higher statistical power**
- **More likely to detect true effects**
- **More precise estimates**

Disadvantages

- **Very sensitive to violations**
- **Like a tailored suit - fits specific situations only**
- **May give misleading results if assumptions are wrong**

Non-Parametric Methods

Distribution-free methods that make no assumptions about the population distribution.

Advantages

- **Very robust and flexible**
- **Work with any data distribution**
- **Perfect for ordinal/ranked data**

Disadvantages

- **Lower statistical power**
- **Less reliable for detecting effects**
- **Less precise estimates**

When to use which method:

Decision Framework

1. **Ordinal/Ranked data: Use Non-Parametric**
2. **Normally distributed data: Use Parametric**
3. **Non-normally distributed data: Use Non-Parametric**

Example: Competition Judging

When judges rank contestants as 1st, 2nd, 3rd place (ordinal data), non-parametric tests work perfectly because they handle ranked data effectively.

Summary

Key Takeaways:

- **Random Variables:** Discrete (countable) vs Continuous (infinite range)
- **Probability Functions:** PMF for discrete, PDF for continuous
- **Distributions:**
  - Binomial: Fixed trials, binary outcomes
  - Poisson: Events in fixed intervals, known average
  - Normal: Bell curve, defined by mean and standard deviation
- **Methods:** Parametric (assumes distribution) vs Non-parametric (distribution-free)

This foundation prepares us for statistical inference, hypothesis testing, and advanced modeling techniques.

"> learning-objectives

Lesson: Foundations of Statistical Inference

Foundations of Statistical Inference:

A Detailed Guide to Sampling, Estimation, and the Central Limit Theorem

Contents

1 The Why and How of Sampling

Statistical inference is the art and science of making conclusions about a whole population based on a small piece of it. This process begins with the fundamental act of sampling.

1.1 Populations, Samples, Parameters, and Statistics

- **Population:** The entire group of individuals or objects we wish to study. For example, all students at BITS Pilani, all smartphones produced by a factory, or all voters in India.
- **Parameter:** A numerical value describing a characteristic of the population. Parameters are typically unknown constants that we want to estimate. We denote them with Greek letters (e.g., μ for mean, σ for standard deviation, p for proportion).
- **Sample:** A subset of the population that we actually collect data from.
- **Statistic:** A numerical value describing a characteristic of the sample. We calculate statistics from our data. We denote them with Roman letters (e.g., $\bar{x}$ for sample mean, s for sample standard deviation, $\hat{p}$ for sample proportion).

The core goal is to use a known statistic to infer the value of an unknown parameter. For this leap of faith to be valid, the sampling method is critical.

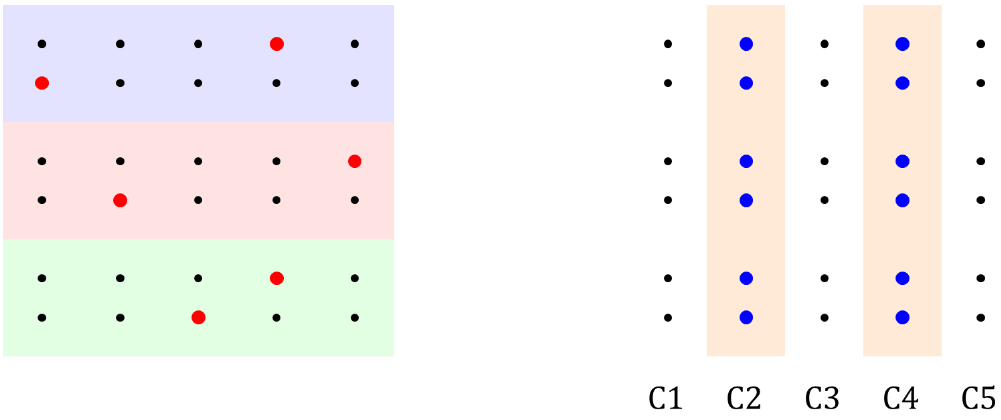
1.2 Types of Sampling Methods

A good sample should be representative of the population. The best way to achieve this is through probability sampling, where every member of the population has a known chance of being selected.

- **Simple Random Sampling (SRS):** Every individual and every possible sample of size n has an equal chance of being selected. This is the ideal gold standard.
- **Stratified Sampling:** The population is first divided into non-overlapping subgroups, or strata, based on a shared characteristic (e.g., year of study, campus location). A simple random sample is then taken from each stratum. This ensures representation from all key subgroups.
- **Cluster Sampling:** The population is divided into subgroups, or clusters, often based on geography. We then randomly select entire clusters and sample every individual within the chosen clusters. This is often more practical and cost-effective than SRS.
- **Systematic Sampling:** A starting point is chosen randomly, and then every k-th individual is selected from a list.
- **Convenience Sampling:** Selecting individuals who are easiest to reach. While simple, it is highly prone to bias and is not a valid method for statistical inference.

2 Point Estimation: The Best Single Guess

Stratified Sampling Cluster Sampling



Sample from every stratum Take all from selected clusters

Figure 1: A clean comparison of Stratified and Cluster Sampling.

A point estimate is our best single-value guess for an unknown population parameter, based on our sample data. For example, $\bar{x}$ is the point estimate for μ.

2.1 Properties of Good Estimators

- **Unbiasedness:** An estimator is unbiased if the mean of its sampling distribution is equal to the true value of the parameter being estimated. The sample mean $\bar{x}$ is an unbiased estimator of μ.

- **Efficiency:** An efficient estimator is one that has a small variance in its sampling distribution. A more efficient estimator is more likely to produce an estimate close to the true parameter value.

2.2 The Inherent Limitation: Sampling Error

The great weakness of a point estimate is that it provides no information about its precision. It is a single number that is almost guaranteed to be at least slightly wrong. To overcome this, we must understand how this estimate behaves across many samples.

3 The Central Limit Theorem (CLT)

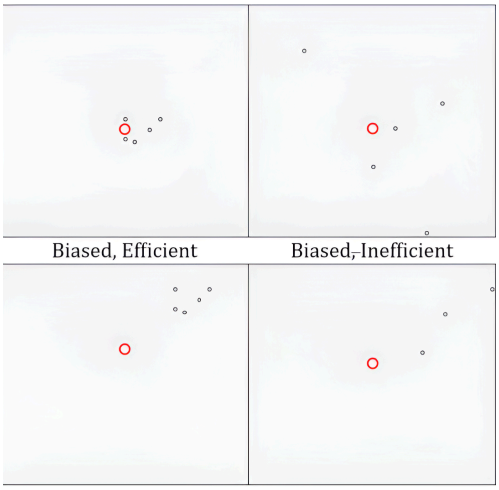


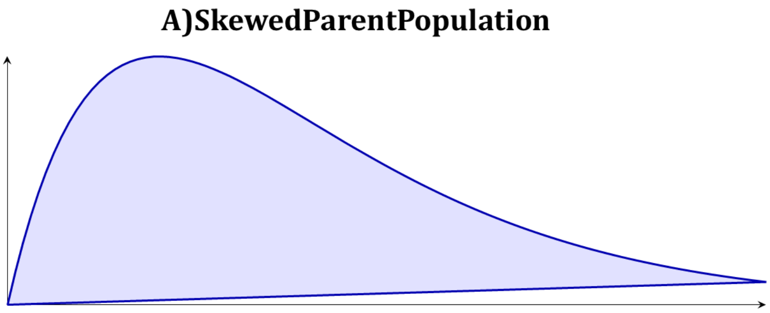
Figure 2: The target shooting analogy for estimators. The red bullseye is the true parameter. Each black dot is a point estimate from a different sample.

The Central Limit Theorem is arguably the most important result in all of statistics. It describes the shape, center, and spread of the sampling distribution of the sample mean, and it works even when the original population is not normally distributed.

The Central Limit Theorem States:

- **Center:** Its mean will be equal to the population mean $(\mu\bar{x} = \mu)$.
- **Spread:** Its standard deviation (the standard error) will be $\sigma/\sqrt{n}$.
- **Shape:** It will be approximately normally distributed.

A common rule of thumb considers $n \geq 30$ to be "sufficiently large." This allows us to use the mathematics of the normal distribution to make inferences about the population mean with large samples.



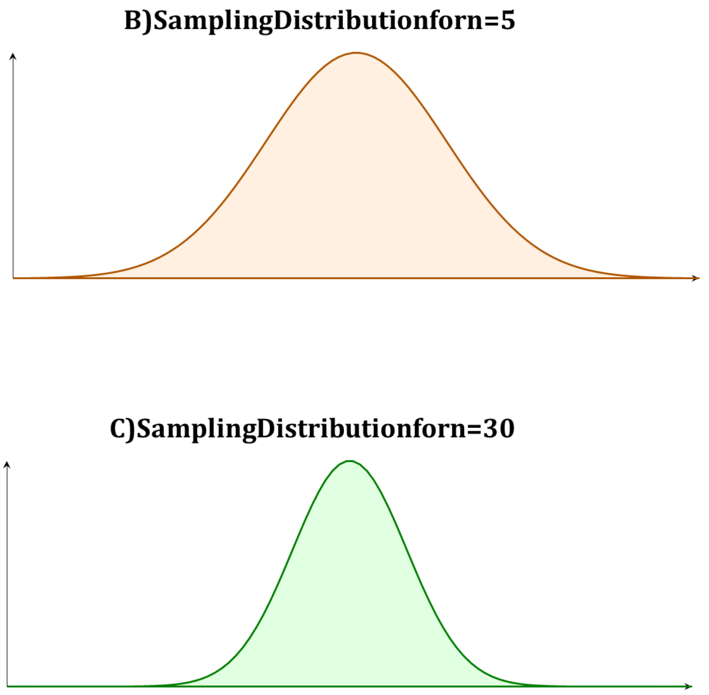


Figure 3: Visualizing the CLT. As sample size n increases, the distribution of sample means becomes narrower and more normal, regardless of the parent population's shape.

4 Interval Estimation: A Range of Confidence

Now that we know the sampling distribution of $\bar{x}$ is approximately normal (thanks to the CLT), we can now build a confidence interval around it.

4.1 Structure of a Confidence Interval

$CI = \bar{x} \pm \text{Margin of Error} = \bar{x} \pm (\text{Critical Value} \times \text{Standard Error})$

4.2 Case 1: When σ is Known (The Z-Interval)

Standard error = $\sigma/\sqrt{n}$, critical value comes from the standard normal (Z) distribution. $CI = \bar{x} \pm Z\alpha/2 \times (\sigma/\sqrt{n})$

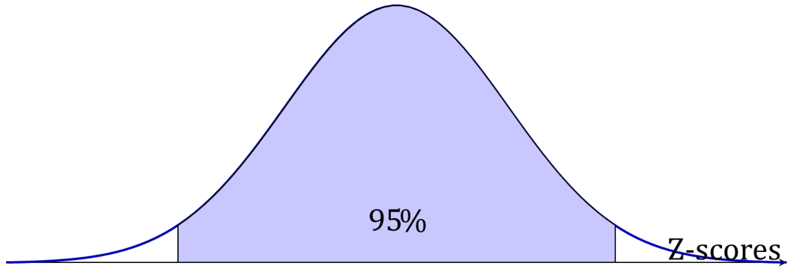


Figure 4: A 95% Confidence Interval on the Z-distribution.

4.3 Case 2: When σ is Unknown (The T-Interval)

Standard error = $s/\sqrt{n}$. Critical values come from the t-distribution with $df = n-1$. The t-distribution has fatter tails, creating wider intervals, especially for small n.

4.4 The Correct Interpretation: Net Fishing

Imagine the true population mean μ is a stationary fish in a pond. A confidence interval is like a net you throw into the pond. A 95% confidence level means that in the long run, 95% of the nets will capture the fish. It does not mean there is a 95% chance that your specific net has caught the fish. Confidence is in the process, not in the single outcome.

"> lesson-foundations-of-statistical-inference

Inferential Statistics

Week 2 Reading Material

Core Concept: Inferential statistics is the process of learning about a population through sampling. Rather than examining every individual in a population (which is often impractical), we take representative samples and use them to make inferences about the entire population.

Key Analogy: Like tasting a spoonful of soup to judge the entire pot, we use samples to understand populations.

1. Understanding Population vs. Sample

Population Parameter vs. Sample Estimate

- **Population Parameter:** The true value we want to know (e.g., average height of all Indians)
- **Sample Estimate:** Our approximation based on a subset (e.g., average height of 500 randomly selected Indians)

> **Example:** To find the average height of 1.4 billion Indians, instead of measuring everyone (time-consuming, expensive, often infeasible), we measure a sample of 1,000 people and use that to estimate the population average. >2. The Art of Sampling
>
> ### Why Sample?

Three main reasons to use sampling instead of complete enumeration:

- Time efficiency: Examining entire populations takes too long
- Resource conservation: Full population studies are expensive
- Feasibility: Sometimes impossible to access entire population

Types of Sampling

A. Probability Sampling

Every member of the population has a known, non-zero chance of being selected.

Simple Random Sampling

Every individual has an equal chance of selection, like drawing names from a hat.

Stratified Sampling

Divide population into groups (strata) first, then sample from each group. Example: Dividing India by states, then sampling from each state.

Systematic Sampling

Select every nth individual (e.g., every 5th person in a line).

⚠ **Important Caveat:** Avoid systematic sampling when dealing with periodic patterns. For instance, if you always sample on the 7th day of each week (always a Sunday), you might miss weekly variations in behavior, leading to biased results. The sampling interval should not align with natural cycles in your data.

Cluster Sampling

Two-step randomness: (1) Randomly select clusters/groups, (2) Sample within selected clusters. Example: Randomly select 5 states, then randomly sample within those states.

B. Non-Probability Sampling

Convenience-driven but prone to bias.

Convenience Sampling

Select whoever is easily accessible (e.g., surveying only your neighbors).

Quota Sampling

Set quotas for different groups, then conveniently fill them (e.g., first 50 males and first 50 females you encounter).

Snowball Sampling

Participants recruit other participants, creating a chain effect. Useful for hard-to-reach populations.

3. Sample Size Determination: The Goldilocks Problem

Not too big, not too small! Sample size depends on three factors:

- **Confidence Level:** How sure do you want to be? Higher confidence = larger sample needed
- **Margin of Error:** How much inaccuracy can you tolerate? Less error tolerance = larger sample needed

- **Population Variability:** How diverse is your population?
  - → More variation = larger sample needed
  - → Less variation (homogeneous/similar population) = smaller sample needed

Historical Example of Sampling Bias: 1936 U.S. Presidential Election

A survey of 2.4 million people predicted the wrong winner because samples were drawn only from car owners and telephone users, who represented a wealthy subset of the population. This demonstrates how **biased sampling leads to biased results**, regardless of sample size.

4. Point Estimation

Point estimation provides a single "best guess" value for a population parameter. While convenient, remember: the best guess may sometimes be wrong!

Common Point Estimates

- **Population Mean:** Average value (e.g., average age, average height)
- **Population Proportion:** Percentage or ratio (e.g., % voting for a party)
- **Population Variance:** Measure of spread in the data

Properties of Good Estimators

1. Unbiased

On average, the estimator should hit the true value. Individual estimates may vary, but they should center around the truth.

2. Efficient (Consistent)

Less variability in estimates; they cluster tightly around the true value rather than being scattered.

Think of it like archery:

- **Unbiased:** Your arrows center around the bullseye (even if individual shots vary)
- **Efficient:** Your arrows are tightly grouped (not scattered all over)

5. Understanding Degrees of Freedom

Why n-1 in Variance Formula?

Variance = $\Sigma(x - \text{mean})^2 / (n - 1)$

Football Analogy: Imagine filling 10 positions on a football field with a fixed formation (4 forwards, 3 mids, 3 defenders):

- First 9 players can freely choose from available positions
- The 10th player has no choice; only one position remains

Similarly, when calculating variance using the sample mean, once n-1 deviations are determined, the last one is fixed (since all deviations must sum to zero). This "loss of freedom" is why we divide by n-1 instead of n, ensuring an unbiased estimate of population variance.

Key Takeaways

- Good sampling is crucial for valid inferences about populations
- Probability sampling reduces bias but requires more effort
- Sample size depends on desired confidence, acceptable error, and population variability
- Point estimates give single values but come with uncertainty
- Good estimators should be both unbiased and efficient

"> week-2-reading-material

The Hypothesis Testing Framework

1 The Core Idea: Making Decisions with Data
Hypothesis testing is a formal procedure for using sample data to evaluate a claim about a population. It provides a structured framework for making decisions in the face of uncertainty. The core logic of hypothesis testing is analogous to a criminal trial.

1.1 The Courtroom Analogy

- **The Presumption of Innocence:** In a trial, the defendant is presumed innocent until proven guilty. This is the default position, the status quo.
- **The Burden of Proof:** The prosecution must present compelling evidence to convince the jury to overturn this presumption of innocence.
- **The Verdict:** The jury does not prove the defendant is innocent; they either find enough evidence to convict (”guilty”) or not enough evidence (”not guilty”).

In statistics, the roles are played by:

- **The Null Hypothesis (H0):** This is the ”presumption of innocence.” It is a statement of no effect, no difference, or the status quo. We assume H0 is true unless the evidence suggests otherwise. Example: A new drug has no effect on recovery time. (μnew = μold)

- **The Alternative Hypothesis (HA or H1):** This is the claim the researcher is trying to find evidence for. It is the ”guilty” verdict. Example: The new drug reduces recovery time. ($\mu new < \mu old$)
- **The Sample Data:** This is the ”evidence” we collect.
- **The Conclusion:** We don’t ”prove” the null hypothesis. We either find enough evidence to reject the null hypothesis in favor of the alternative, or we fail to reject the null hypothesis due to insufficient evidence.

2 The 4-Step Hypothesis Testing Framework

Every formal hypothesis test follows a consistent, four-step process.

2.1 Step 1: State the Hypotheses

First, clearly state the null (H0) and alternative (HA) hypotheses in terms of the popula tion parameter(s) of interest (e.g., $\mu, p, \sigma 2$). The alternative hypothesis can be one-tailed ($<,>$) or two-tailed ($\neq$).

- $H0 : \mu = 100$
- $HA : \mu \neq 100$ (Two-tailed)

2.2 Step 2: Set the Decision Criteria

Before collecting data, we must define what constitutes ”strong evidence.” This is done by setting a significance level, α.

- α is the probability of making a Type I error (rejecting a true null hypothesis).
- It is our threshold for ”reasonable doubt.”
- Common values for α are 0.05 (5%), 0.01 (1%), and 0.10 (10%).

2.3 Step 3: Collect Data and Calculate the Test Statistic

After setting the criteria, we collect our random sample and summarize the evidence into a single number called a test statistic. The general form of a test statistic is:

Test Statistic = (Sample Statistic − Null Hypothesis Value) / Standard Error

This value measures how many standard errors our sample result is from the value claimed in the null hypothesis. Common test statistics are the Z-statistic, t-statistic, $\chi 2$-statistic, and F-statistic.

2.4 Step 4: Make a Decision

Finally, we use the test statistic to determine how likely our observed sample result is, assuming the null hypothesis is true. This likelihood is quantified by the p-value.

- **P-value:** The probability of observing a test statistic as extreme or more extreme than the one calculated, given that the null hypothesis (H0) is true.

We then compare our p-value to our pre-determined significance level, α.

The Decision Rule:

- If p-value $\leq \alpha$, the result is statistically significant. The evidence is strong enough to reject the null hypothesis.

- If p-value > α, the result is not statistically significant. The evidence is not strong enough; we fail to reject the null hypothesis.

We then state our conclusion in the context of the original research question.

reading-material-the-hypothesis-testing-framework

Errors, P-values, and Significance

1 The Nature of Statistical Decisions

Hypothesis testing uses sample data to make an inference about an entire population. Because a sample is only a small snapshot of the population, we can never be 100% certain that our conclusion is correct. There is always a risk of making an error. A core part of statistics is understanding and managing this risk.

2 The Two Types of Errors

In any hypothesis test, there are four possible outcomes, which can be summarized in a confusion matrix. Two outcomes represent a correct decision, and two represent an error.

Decision

True State of RealityH0 is True H0 is FalseFail to Reject H0 Correct Decision Type II Error (β)Reject H0 Type I Error (α) Correct Decision (Power)*Table 1: Outcomes of a Hypothesis Test*

2.1 Type I Error: The ”False Alarm”

A Type I Error occurs when we reject a true null hypothesis.

- Analogy: Convicting an innocent person; a fire alarm ringing when there is no fire; a medical test diagnosing a healthy person with a disease (a false positive).
- The probability of committing a Type I error is denoted by α, the significance level of the test.
- We, the researchers, have direct control over this error rate. By setting α (e.g., to 0.05), we are explicitly defining our tolerance for a false alarm.

2.2 Type II Error: The ”Missed Detection”

A Type II Error occurs when we fail to reject a false null hypothesis.

- Analogy: Acquitting a guilty person; a fire alarm failing to ring during a fire; a medical test failing to detect a disease that is present (a false negative).
- The probability of committing a Type II error is denoted by β.

- The complement of this, $1 - \beta$, is called the Power of the test—the probability of correctly detecting a real effect.

There is an inverse relationship between α and β. Decreasing the risk of a Type I error (making α smaller) generally increases the risk of a Type II error (making β larger).

3 The P-Value: Measuring the Strength of Evidence

The p-value is the primary tool for evaluating the evidence against the null hypothesis. It is calculated from the sample data.**Definition:** The p-value is the probability of obtaining a test statistic as extreme or more extreme than the one observed, under the assumption that the null hypothesis is true.

A p-value is a measure of surprise.

- A small p-value (e.g., 0.01) means that our observed data is very surprising if the null hypothesis were true. This strong evidence leads us to doubt the null hypothesis.
- A large p-value (e.g., 0.45) means our observed data is not surprising. It is consistent with what we might see due to random chance if the null hypothesis were true. This provides weak evidence against the null.

Crucial Misconception: The p-value is NOT the probability that H0 is true.

4 Significance: Where Evidence Meets the Standard

The final decision in a hypothesis test is a direct comparison between the evidence and our pre-set standard.Evidence → P-Value | Standard → α

The decision rule is simple:

- If P-Value ≤ α, we say the result is statistically significant. We reject the null hypothesis. The evidence has met our standard of proof.
- If P-Value > α, we say the result is not statistically significant. We fail to reject the null hypothesis. The evidence was not strong enough.

4.1 Statistical Significance vs. Practical Importance

It is critical to distinguish between statistical significance and real-world importance.

- **Statistical Significance** simply means the observed effect is unlikely to be due to random chance.
- **Practical Importance** refers to the magnitude of the effect. Is the difference large enough to matter in a real-world context?

With a very large sample size, a tiny and practically meaningless effect (e.g., a new drug lowering blood pressure by 0.01 mmHg) can be statistically significant. Always consider the effect size, not just the p-value.

reading-material-errors-p-values-and-significance

Interval Estimation of the Mean

1 From Point Estimates to Interval Estimates

In statistics, we often want to estimate a population parameter, such as the population mean (μ). A common approach is to take a random sample and calculate the sample mean (¯x). This single value, ¯x, is our point estimate for μ.

However, a point estimate has a significant limitation: it provides no information about its own accuracy. Due to sampling error—the natural, random variation between samples—it is virtually certain that our sample mean ¯x will not be exactly equal to the true population mean μ. If we took a different sample, we would get a different ¯x.

To address this uncertainty, we use an interval estimate, more commonly known as a confidence interval. Instead of a single number, a confidence interval provides a range of plausible values that is likely to contain the true population parameter.

1.1 The Core Idea

An interval estimate is more informative because it reflects the uncertainty inherent in the sampling process. For example, stating that the average student score is between 72 and 78 is more realistic and useful than claiming it is exactly 75.

2 Anatomy of a Confidence Interval

Every two-sided confidence interval has the same fundamental structure:
Confidence Interval = Point Estimate ± Margin of Error

Let’s break down the components:

- Point Estimate: This is our best single guess for the parameter. For a population mean, the point estimate is the sample mean, ¯x. It forms the center of our interval.
- Margin of Error (E): This is the ”radius” of our interval. It quantifies the precision of our estimate at a given level of confidence. A smaller margin of error implies a more precise estimate.

The margin of error is itself a product of two factors:
Margin of Error = (Critical Value) × (Standard Error)

- Critical Value: A multiplier determined by our desired confidence level. It reflects how certain we want to be. Common confidence levels are 90%, 95%, and 99%.

- Standard Error: The standard deviation of the sampling distribution of the point estimate. For the sample mean, it measures the typical amount of error between $\bar{x}$ and μ.

3 The Ideal Case: Population Standard Deviation (σ) is Known

In the rare scenario where we know the true standard deviation of the population, σ, we can use the standard normal (Z) distribution to construct our confidence interval. This is known as a Z-interval.

The standard error of the mean is given by $\sigma \sqrt{n}$. The critical value, $Z\alpha/2$, is found from the Z-distribution, where α is the significance level ($1 - \text{confidence level}$). For a 95% confidence interval, $\alpha = 0.05$, and we look for the Z-score that leaves an area of $\alpha/2 = 0.025$ in each tail. This value is $Z0.025 = 1.96$.

The complete formula for a Z-interval for the mean is:
$CI = \bar{x} \pm Z\alpha/2 \ \sigma \sqrt{n}$

4 The Realistic Case: Population Standard Deviation (σ) is Unknown

In nearly all real-world applications, the population standard deviation σ is unknown. We must therefore estimate it using the sample standard deviation, s.

Using an estimate (s) instead of the true value (σ) introduces an extra layer of uncertainty. To account for this, we cannot use the Z-distribution. Instead, we use the Student’s t-distribution. The resulting interval is known as a T-interval.

4.1 The t-Distribution

The t-distribution is a family of curves that are similar to the Z-distribution (bell-shaped, symmetric about zero) but have ”fatter tails.” These fatter tails reflect the increased uncertainty of using s to estimate σ. The precise shape of the t-distribution is determined by its degrees of freedom (df), which for a one-sample mean test is $df = n - 1$. As the sample size n (and thus the df) increases, the t-distribution converges to the Z distribution.

The formula for a T-interval for the mean is:
$CI = \bar{x} \pm t\alpha/2{,}df \ s \sqrt{n}$

Here, $t\alpha/2{,}df$ is the critical value from the t-distribution with $n - 1$ degrees of freedom that leaves an area of $\alpha/2$ in the upper tail.

5 The Correct Interpretation of a Confidence Interval

This is one of the most misunderstood concepts in statistics.

- INCORRECT: ”There is a 95% probability that the true population mean μ is in my calculated interval [72, 78].” This is wrong because the true mean μ is a fixed, unknown constant. It is either in the interval or it is not. The probability is either 1 or 0.

- CORRECT: ”I am 95% confident in the method used to construct this interval. If I were to repeat this sampling process many times, I would expect 95% of the intervals I create to successfully capture the true population mean, μ.”

Confidence is in the long-run success rate of the procedure, not in any single result.

reading-material-interval-estimation-of-the-mean

One-Sample Hypothesis Tests (Z, T)

1 Applying the Hypothesis Testing Framework
The general hypothesis testing framework provides the logic for making statistical de cisions. One-sample tests are the direct application of this framework to a common scenario: testing a claim about a single population mean (μ).The goal is to determine if the mean of a population is statistically different from a specific, hypothesized value ($\mu 0$). For example:

- Is the average weight of a product from our factory equal to the claimed 500 grams? ($H0 : \mu = 500$)
- Is the average response time for our customer service center less than 3 minutes? ($H0 : \mu = 3$, $HA : \mu < 3$)

To conduct the test, we apply the 4-step framework. The primary decision we must make is choosing the correct tool to calculate our test statistic in Step 3. This choice leads to two different tests: the Z-test and the t-test.

2 The Critical Decision: Is σ Known?
The choice between a Z-test and a t-test for a single mean is the simplest in statistics. It depends on one piece of information:Is the population standard deviation, σ, known?

2.1 Scenario 1: The Z-Test (When σ is Known)
We use the one-sample Z-test in the rare but ideal situation where the true population standard deviation, σ, is known. This might occur in industrial processes with extensive historical data.The test statistic is calculated as:

$Z = (¯x - \mu 0) / (\sigma/\sqrt{n})$

Where:

- ¯x is the sample mean.
- $\mu 0$ is the hypothesized population mean from H0.
- σ is the known population standard deviation.
- n is the sample size.

The resulting Z-statistic is then compared to the standard normal distribution to find the p-value.

2.2 Scenario 2: The T-Test (When σ is Unknown)

In virtually all real-world research, σ is unknown. We must estimate it using our sample standard deviation, s. The uncertainty of using an estimate requires us to use the one sample t-test.The test statistic is calculated as:

$t = (\bar{x} - \mu 0) / (s/\sqrt{n})$

The formula's structure is identical to the Z-test, but the substitution of s for σ means the statistic follows a t-distribution, not a Z-distribution.
The p-value is found using the t-distribution with $df = n - 1$ degrees of freedom.

3 Assumptions for One-Sample Tests for a Mean

For the results of these tests to be valid, certain assumptions must be met:

1. **Random Sample:** The data must be collected from a random sample to ensure it is representative of the population.
2. **Normality:** The underlying population should be approximately normally dis tributed. The t-test is fairly "robust" to violations of this assumption, especially as the sample size increases (due to the Central Limit Theorem). However, for very small sample sizes ($n < 15$), the test should be used with caution if the data shows strong skewness or outliers.

reading-material-one-sample-hypothesis-tests-z-t

Determining Sample Size

1 The Importance of Sample Size

Before conducting any research, poll, or experiment, one of the most critical questions to answer is: "How much data do we need?" The choice of sample size (n) is a fundamental part of experimental design and represents a crucial trade-off between precision and resources.

- **Too Small a Sample:** A sample that is too small will have a large margin of error, leading to imprecise estimates. This increases the risk of making a Type II error (failing to detect a real effect), rendering the study's conclusions unreliable.
- **Too Large a Sample:** While a large sample provides more precision, collecting data costs time and money. An unnecessarily large sample is an inefficient use of resources, as the gains in precision become marginal after a certain point.

Our goal is to find the minimum sample size required to achieve a desired level of statistical precision and confidence.

2 The Key Ingredients of Sample Size Calculation

To calculate the necessary sample size, we must first define our objectives by specifying three key inputs:

1. **The Margin of Error (E):** This is the maximum amount of error we are willing to tolerate in our estimate. It defines the desired precision. For a mean, it's the ”plus or minus” value (e.g., ±3 kg). For a proportion, it's the ”plus or minus” percentage (e.g., ±4%). A smaller, more desirable margin of error requires a larger sample size.
2. **The Confidence Level ($1 - \alpha$):** This is the level of certainty we require. It is typically expressed as 90%, 95%, or 99%. A higher confidence level requires a larger sample size. The confidence level determines the critical value (e.g., $Z\alpha/2$) used in the calculation.
3. **The Population Variability (σ or p):** This is an estimate of the standard de viation or proportion in the population we are studying. A more heterogeneous (variable) population requires a larger sample to capture its characteristics accu rately. This is often the most challenging component to estimate.

3 Sample Size for a Population Mean (μ)

The formula for the sample size required to estimate a population mean is derived from the margin of error formula for a Z-interval:

$$E = Z\alpha/2 \ \sigma / \sqrt{n}$$

By algebraically solving for n, we get:

$$n = (Z\alpha/2 \cdot \sigma / E)^2$$

Note: Since the result of this calculation is often not a whole number, we always round up to the next integer to ensure our sample size is sufficient.

3.1 The Challenge: Estimating σ

The formula requires an estimate for the population standard deviation, σ, before we've even collected data. We can obtain this estimate in a few ways:

- **Pilot Study:** Conduct a small preliminary study to calculate a sample standard deviation, s, and use it as an estimate for σ.
- **Previous Research:** Use the standard deviation reported in similar studies or historical data.
- **Range Rule of Thumb:** A rough estimate for sigma is the expected range of the data (Maximum − Minimum) divided by 4.

4 Sample Size for a Population Proportion (p)

Similarly, the formula for the sample size required to estimate a population proportion is derived from its margin of error formula:

$$n = p(1 - p) \ (Z\alpha/2 / E)^2$$

4.1 The Challenge: Estimating p

This formula requires a preliminary estimate for the population proportion, p.

- If we have a reasonable estimate from prior knowledge or a pilot study, we can use that.
- If we have no idea what p might be, we must use the most conservative estimate. The term $p(1 - p)$ is maximized when $p = 0.5$. Using $p = 0.5$ will yield the largest possible sample size for a given confidence level and margin of error, guaranteeing our sample will be large enough.

Therefore, if p is unknown, the safe formula to use is:
$n = 0.25 (Z\alpha/2 / E)^2$

reading-material-determining-sample-size

---

1 Interval Estimation of the Mean

Estimation and Hypothesis Testing

Module 3: Statistical Inference and Modeling

1 Interval Estimation of the Mean

▼

What You Learned

Moving beyond single point estimates to construct **confidence intervals** that provide a range of credible values for the population mean.

> **Key Concepts:**
>
> - **Point Estimate:** Single best guess (e.g., sample mean $\bar{x}$) but never perfectly correct
> - **Sampling Error:** Difference between sample mean and true population mean (μ)
> - **Confidence Interval:** Point Estimate ± Margin of Error
>
> ### Two Scenarios

Case 1: When σ is Known (rare in practice)

$$CI = \bar{x} \pm z(\alpha/2) \times (\sigma/\sqrt{n})$$

Uses standard normal (Z) distribution

Case 2: When σ is Unknown (more practical)

$$CI = \bar{x} \pm t(\alpha/2,\ df) \times (s/\sqrt{n})$$

Uses Student's t-distribution with degrees of freedom (df = n-1)
Wider intervals due to estimation uncertainty

Why This Matters:
Important Insights

- Higher confidence level (95% → 99%) = wider interval
- Larger sample size = narrower interval (more precision)
- As sample size grows, t-distribution approaches normal distribution

2 Determining Sample Size

▼

What You Learned

How to calculate the minimum sample size needed to achieve desired precision without wasting resources.

The Balancing Act:

- **Too small:** Results unreliable, large margin of error
- **Too large:** Wasting time, money, and resources
- **Just right:** Adequate precision with efficient resource use

Three Key Ingredients

1. Margin of Error (E): Your tolerance for imprecision
Smaller margin requires larger sample size

2. Confidence Level: How certain you want to be
90% vs 95% vs 99% confidence (determines Z value)

3. Population Variability (σ): How diverse the population is
More variability requires larger sample

The Formula

$$n = (Z^2\sigma^2) / E^2$$

Critical Insight:
For Proportions

$$n = (Z^2p(1-p)) / E^2$$

Maximized when $p = 0.5$ (most conservative estimate)

Why This Matters:
Practical Solutions When σ is Unknown

- Conduct a pilot study to estimate σ
- Use data from similar previous studies
- Apply rule of thumb: σ ≈ Range/4

3 Hypothesis Testing Framework

▼

What You Learned

A systematic, four-step process to test claims about populations using sample data - like a courtroom trial for statistics.

The Courtroom Analogy:

- **Null Hypothesis (H_0):** Defendant is innocent (status quo)
- **Alternative Hypothesis (H_a):** Defendant is guilty (claim to prove)
- **Evidence:** Sample data we collect
- **Verdict:** Reject H_0 or Fail to Reject H_0

The Four-Step Process

Step 1: State the Hypotheses

H_0: Status quo (contains "=")H_a: What you suspect/want to prove

Step 2: Set Significance Level (α)

Usually 5% (0.05), sometimes 1% or 10%

Step 3: Calculate Test Statistic

Z-test or t-test depending on available information

Step 4: Make Decision

Compare p-value to α

Example Application

> **Claim:** Phone battery lasts 40 hours on average
>
> **H_0:** $\mu = 40$
>
> **H_a:** $\mu < 40$ (suspicion: it's less)
>
> **Sample:** $n=36$, $\bar{x}=38.8$, $\sigma=3$
>
> **Z** = $(38.8-40)/(3/\sqrt{36}) = -2.40$
>
> **p-value** = $0.082 < 0.05$ → Reject H_0

Why This Matters:

Hypothesis testing moves you from "I think this is true" to "The data statistically supports this conclusion." It's used everywhere: quality control in manufacturing, medical trials, A/B testing in tech companies, and academic research.

>4 Errors, P-values & Significance ▼

What You Learned

Understanding the types of mistakes we can make in hypothesis testing and how to interpret statistical evidence.

Two Types of Errors

Type I Error (False Alarm):

Rejecting H_0 when it's actually trueExample: Medical test says you have disease when you don'tProbability = α (significance level)

Type II Error (Miss Detection):

Failing to reject H_0 when it's actually falseExample: Medical test says you're healthy when you have diseaseUnderstanding P-value

P-value = Probability of getting results as extreme (or more) as observed, assuming H_0 is true

Coin Example:

- Assume coin is fair (H_0)
- Flip 10 times, get 6 heads → p-value high (not surprising)
- Flip 10 times, get 10 heads → p-value tiny (very surprising!)
- Conclusion: Coin likely biased

Decision Rule

If p-value ≤ α: Reject H_0 (statistically significant)
If p-value > α: Fail to reject H_0 (not significant)

Why This Matters:
Key Insight

α is your standard (set before testing)
p-value is your evidence (calculated from data)

5 One Sample Tests (Z-test & T-test)

▼

What You Learned

How to choose between Z-test and t-test when testing claims about a population mean using sample data.

The Critical Question

Is the population standard deviation (σ) known?

Decision Tree:

- **If σ is known:** Use Z-test (rare in practice)
- **If σ is unknown:** Use t-test (common in practice)

>Z-test (σ known)

$$Z = (\bar{x} - \mu_0) / (\sigma/\sqrt{n})$$

Uses standard normal distribution
Example: Manufacturing with well-documented historical variability

>T-test (σ unknown)

```
formula
```

$t = (\bar{x} - \mu_0) / (s/\sqrt{n})$

Uses t-distribution with df = n-1
Uses sample standard deviation (s) as estimateWider critical values to account for estimation uncertainty

Practical Example: Pizza Delivery

Scenario: Pizza shop claims delivery ≤ 30 min

Suspicion: Actually takes longer

H_0: $\mu \leq 30$$H_a$: $\mu > 30$Sample: $n=16$, $\bar{x}=33.5$, $s=6$$t = (33.5-30)/(6/\sqrt{16}) = 2.33$p-value = $0.017 < 0.05$ → Reject H_0

Why This Matters:

Key Similarities & Differences

- **Similar:** Both test hypotheses about population mean
- **Similar:** Formula structure is identical
- **Different:** Z uses σ, t uses s
- **Different:** t has fatter tails (accounts for estimation uncertainty)
- **Convergence:** As n increases, t → Z distribution

"> 1-interval-estimation-of-the-mean

Inferences for Two Population Means

1 The Core Question: Comparing Two Groups

While one-sample tests are useful, many research questions involve comparing two differ ent groups. For example, we might want to know if a new drug is more effective than a placebo, or if one teaching method leads to better exam scores than another.When comparing the means of two groups, our first and most critical step is to determine the nature of our samples.Are the samples independent or paired (dependent)?This choice of experimental design dictates the entire statistical analysis.

2 Scenario 1: Independent Samples

We have independent samples when the individuals in the first group are completely unrelated to the individuals in the second group. The selection of one group has no bearing on the selection of the other.

2.1 Examples

- A sample of male students and a separate sample of female students.
- A control group of patients and a treatment group of different patients.
- A sample of products from Machine A and a separate sample from Machine B.

2.2 Hypotheses and the Two-Sample T-Test

The null hypothesis (H0) is typically that there is no difference between the two popula tion means.$H_0 : \mu_1 = \mu_2$ or $H_0 : \mu_1 - \mu_2 = 0$The alternative hypothesis (HA) can be two-tailed ($\mu_1 /= \mu_2$), left-tailed ($\mu_1 < \mu_2$), or right-tailed ($\mu_1 > \mu_2$).

Since the population standard deviations (σ_1, σ_2) are almost always unknown, we use the two-sample t-test for independent samples. The test statistic is:

$$t = (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0 / \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$$

Where $(\mu_1 - \mu_2)_0$ is the hypothesized difference from H0, which is almost always 0. The denominator is the standard error of the difference between two means. The degrees of freedom for this test are calculated with a complex formula (Welch’s approximation) that is handled by statistical software.

3 Scenario 2: Paired (Dependent) Samples

We have paired samples when each data point in the first sample is naturally and uniquely linked to a data point in the second sample.

3.1 Examples

- **Before-and-After:** The same subjects are measured before and after a treatment (e.g., measuring weight before and after a diet).
- **Matched Pairs:** Subjects are matched based on key characteristics (e.g., twins, or matching subjects by age and gender), and one from each pair is randomly assigned to each group.
- **Paired Conditions:** The same subject is tested under two different conditions (e.g., a person’s reaction time using their left hand vs. their right hand).

3.2 The Paired T-Test: A Clever Simplification

The analysis of paired data is elegant. Instead of comparing two large, variable groups, we simplify the problem:

1. For each pair, we calculate the difference, $d = x_{after} - x_{before}$.

2. This creates a single sample of differences.
3. We then perform a simple one-sample t-test on this sample of differences.

The null hypothesis becomes a test of whether the mean difference in the population is zero.
$H0 : \mu d = 0$

The test statistic is the one-sample t-statistic applied to the differences:
$t = (\bar{d} - 0) / (sd/\sqrt{n})$

Where $\bar{d}$ is the mean of the sample differences, sd is the standard deviation of the sample differences, and n is the number of pairs. The degrees of freedom are $df = n - 1$.

3.3 The Power of Pairing

When a paired design is appropriate, it is generally much more powerful than an indepen dent design. By comparing each subject to themselves or a close match, we control for a vast amount of extraneous, individual-to-individual variability. This ”noise” reduction makes it easier to detect the true ”signal” of the treatment effect.

reading-material-inferences-for-two-population-means

Inferences for Two Population Variances

1 Why Compare Variances?

While tests concerning means are common, there are many situations where the primary interest is in the variability or consistency of a population. The variance, σ^2 (or standard deviation, σ), is the parameter that measures this spread. Comparing the variances of two populations allows us to make judgments about their relative consistency, stability, or risk.

Examples:

- **Quality Control:** A factory manager might want to know if a new machine pro duces parts with the same consistency (variance) as the old machine. Lower variance is often a sign of higher quality.
- **Finance:** An investor might compare two stocks with the same average return. The stock with the lower variance in returns is considered less risky.
- **Education:** An educator might want to see if a new teaching method results in more consistent test scores (lower variance) than the traditional method, even if the average score is the same.

2 The F-Distribution

To compare two variances, we need a new probability distribution. While t-tests use subtraction to compare means, we use a ratio to compare variances. This requires the F-distribution.

Key Properties of the F-Distribution:

- It is a family of distributions, and its specific shape is defined by two sets of degrees of freedom: – df1 = n1 − 1: Degrees of freedom for the numerator.– df2 = n2 − 1: Degrees of freedom for the denominator.
- The F-distribution is non-negative (its values are always ≥ 0).
- It is skewed to the right.
- The F-statistic is the ratio of two independent chi-square variables, each divided by its degrees of freedom. For our purposes, it is the ratio of two sample variances.

3 The F-Test for Equality of Two Variances

The F-test is the formal procedure for testing a claim about the equality of two popu lation variances ($\sigma^2{}_1$ and $\sigma^2{}_2$).

3.1 Hypotheses

The null hypothesis states that the two population variances are equal. The alternative states they are not (for a two-tailed test).H0 : $\sigma^2{}_1 = \sigma^2{}_2$HA : $\sigma^2{}_1 /= \sigma^2{}_2$

3.2 The F-Test Statistic

The F-statistic is the most intuitive test statistic we have encountered. It is simply the ratio of the two sample variances:F = $s1^2$ / $s2^2$

If the null hypothesis is true and the population variances are equal, we would expect the sample variances to be close to each other, making the F-statistic close to 1. A large F-statistic provides evidence against the null hypothesis.

3.3 A Practical Convention

To simplify the use of F-tables and software, we adopt the following convention:Always place the larger sample variance in the numerator.This ensures that our calculated F-statistic is always greater than or equal to 1. For a two-tailed test, this allows us to only find the area in the upper (right) tail and multiply it by two to get the p-value.

3.4 The Critical Assumption: Normality

The F-test for comparing variances has a significant limitation: it is very sensitive to the assumption that both underlying populations are normally distributed.If this assumption is not met, the results of the F-test can be highly inaccurate. The t-tests for means are much more ”robust” to violations of this assumption than the F-test for variances. Therefore, one should always check for the normality of the data before performing this test.

Testing Population Proportions

1 Introduction to Proportions

While many statistical analyses focus on means of continuous data (e.g., height, weight, time), a vast number of questions revolve around categorical data. A population pro portion (p) represents the fraction of a population that possesses a certain characteristic. The data for each individual is binary: ”yes/no,” ”success/failure,” ”for/against.”

Examples:

- The proportion of the electorate that supports a particular policy.
- The defect rate (proportion) of a product from a manufacturing line.
- The percentage of patients who recover after a specific medical treatment.

We use the sample proportion, $\hat{p} = x/n$ (where x is the number of successes and n is the sample size), to make inferences about the true population proportion, p.

2 The One-Sample Z-Test for a Proportion

This test is used to assess a claim about a single population proportion. The hypothesis testing framework is the same, but the parameter and formulas are specific to proportions.

2.1 Hypotheses

The hypotheses are stated in terms of the population proportion p.$H_0 : p = p_0$ vs. $H_A : p /= p_0$ (or $<$ or $>$)Here, p_0 is the specific value claimed in the null hypothesis.

2.2 The Z-Test Statistic

The test statistic measures how many standard errors the sample proportion ($\hat{p}$) is from the hypothesized proportion (p_0).

$Z = (\hat{p} - p_0) / \sqrt{(p_0(1 - p_0)/n)}$

The denominator, $\sqrt{(p_0(1 - p_0)/n)}$, is the standard error of the proportion. Note that it is calculated using the hypothesized value p_0, because we always assume the null hypothesis is true when conducting the test.

2.3 Conditions for Validity

The Z-test for proportions relies on the normal approximation to the binomial distribu tion. This approximation is only valid if two key conditions are met:

1. **Random Sample:** The data must be from a random and representative sample.
2. **Large Counts Condition:** The sample size must be large enough to expect at least 10 successes and 10 failures under the null hypothesis. We check this using the formulas:
   $np_0 \geq 10$ and $n(1 - p_0) \geq 10$

If this condition is not met, the Z-test is not appropriate, and other methods (like an exact binomial test) should be used.

3 The Two-Sample Z-Test for Proportions
This test is used to compare the proportions of two independent populations.

3.1 Hypotheses
The null hypothesis states that the two population proportions are equal.$H0 : p1 = p2$ or $H0 : p1 − p2 = 0$

3.2 The ”Pooled” Proportion and Test Statistic
Because the null hypothesis assumes the two proportions are equal ($p1 = p2 = p$), our best estimate for this common proportion is to combine, or ”pool,” the data from both samples. The pooled sample proportion, ˆppool, is:

ˆppool = (x1 + x2) / (n1 + n2) = Total Number of Successes / Total Combined Sample Size

We use this pooled proportion to calculate the standard error. The test statistic is:

Z = ((ˆp1 − ˆp2) − 0) / √(ˆppool(1 − ˆppool)(1/n1 + 1/n2))

The conditions for this test are similar, requiring large enough counts in both samples.

reading-material-testing-population-proportions

The Chi-Square Test of Independence

1 Testing for Relationships in Categorical Data
The Chi-Square (χ^2) Test of Independence is a non-parametric hypothesis test used to determine if there is a statistically significant association between two categorical variables. In essence, it answers the question: ”Are these two variables related, or are they independent?”

Examples:

- Is there a relationship between a student’s major and their preferred learning style (e.g., visual, auditory, kinesthetic)?
- Is a person’s opinion on a political issue independent of their age group?
- Is the choice of car brand associated with the buyer’s income level?

The data for a chi-square test is typically presented as a contingency table (or two-way table), which displays the frequency distribution of the variables.

2 The Core Logic: Observed vs. Expected Frequencies

The entire test is built on a comparison between the frequencies we actually collected in our sample and the theoretical frequencies we would expect if the variables were perfectly independent.

- **Observed Frequencies (O):** These are the actual counts from our sample data. They represent what really happened.
- **Expected Frequencies (E):** These are the counts we would expect to see in each cell of our table if the null hypothesis of independence were true.

The expected frequency for any given cell is calculated based on the marginal totals of the table:
E = (Row Total × Column Total) / Grand Total

3 The Chi-Square (χ^2) Test

The Chi-Square test follows the standard hypothesis testing framework.

3.1 Hypotheses

- **H0 (Null Hypothesis):** The two categorical variables are independent (there is no association between them).
- **HA (Alternative Hypothesis):** The two categorical variables are dependent (there is an association between them).

3.2 The Chi-Square Test Statistic

The χ^2 test statistic is a single value that summarizes the total discrepancy between the observed and expected frequencies across all cells of the table.

$\chi^2 = \Sigma\ ((O - E)^2 / E\ )$ across all cells

If the observed frequencies are very close to the expected frequencies, the χ^2 value will be small (close to 0). If they are very different, the χ^2 value will be large. This provides a measure of evidence against the null hypothesis of independence.

3.3 The Chi-Square Distribution

The χ^2 statistic follows a Chi-Square distribution, which has the following properties:

- It is a family of distributions, and its specific shape is determined by the degrees of freedom (df).
- For a test of independence, df = (number of rows − 1) × (number of columns − 1).
- It is non-negative and skewed to the right.

We use this distribution to find the p-value associated with our test statistic. The test is always right-tailed because any deviation from the expected counts (positive or negative) becomes positive after squaring, contributing to a larger χ^2 value.

3.4 Conditions for Validity

For the Chi-Square test to be valid, certain conditions should be met:

1. **Counted Data:** The data must be frequencies or counts, not percentages or pro portions.

2. **Random Sample:** The sample must be representative of the population.
3. **Expected Counts Condition:** The sample size must be large enough so that the expected count in every cell is reasonably large. A common rule of thumb is that all expected cell counts should be 5 or greater. If this condition is not met, the test may not be reliable.

reading-material-the-chi-square-test-of-independence

Statistical Modelling and Inferencing

Week 4 Reading Material

1. Inferences for Two Population Means

Key Concept

Testing whether the difference between two groups is statistically significant by comparing their means.

Types of Two-Sample Tests

Independent (Unpaired) Samples

When comparing two different groups or populations:

- Height of people in Lucknow vs. Kashmir
- Online vs. offline book prices
- Engineering vs. Arts students' salaries

Paired (Dependent) Samples

When the same group is measured under different conditions:

- Weight before and after a diet (same people)
- Skin cream on left arm vs. placebo on right arm (same patients)
- Performance of two tires on the same cars

Mathematical Formulas

Independent T-Test:

$$>t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $>\bar{X}_1$, $>\bar{X}_2$ are sample means, and SE is the standard error of the difference

Paired T-Test:

$$>t = \frac{\bar{D} - \mu_D}{s_D / \sqrt{n}}$$

where $>\bar{D}$ = mean of differences
$>\mu_D$ = hypothesized mean difference (usually 0)
$>s_D$ = standard deviation of differences

Decision Rule (Determined During Study Design)

Critical Question: Was the data collection designed such that each observation in group 1 is naturally matched/paired with a specific observation in group 2?

- **YES** ⟶ Use Paired T-Test (same subjects/units measured twice)
- **NO** ⟶ Use Independent T-Test (two different groups)

Note: This decision is made BEFORE data collection based on your experimental design, not after examining the data.

Example Scenarios

- **Independent:** Does a new teaching method lead to higher test scores than the old one? (Different student groups)
- **Paired:** Does a new drug reduce blood pressure? (Same patients measured before and after)

2. Inferences for Two Population Variances

Key Concept

Comparing the variability or consistency between two groups, not their averages.

Why Compare Variances?

- **Manufacturing:** Which machine produces more consistent output?
- **Finance:** Which investment fund has more stable returns (lower risk)?
- **Quality Control:** Is the variability in production acceptable?

F-Test / F-Distribution

F-Statistic:

$$>F = \frac{s_1^2}{s_2^2}$$

where $>s_1^2$ = larger sample variance
$>s_2^2$ = smaller sample variance

Important Properties of F-Distribution

- Always **non-negative** (variance cannot be negative)
- By convention, place the **larger variance in the numerator**
- Characterized by **two degrees of freedom**: df_1 (numerator) and df_2 (denominator)
- If $F \approx 1$ → variances are similar
- If $F >> 1$ → first group has significantly more variability

Four-Step Hypothesis Testing Framework

1. **State Hypotheses:**
   - $>H_0 : \sigma_1^2 = \sigma_2^2$ (variances are equal)
   - $>H_1 : \sigma_1^2 \neq \sigma_2^2$ (variances are different)
2. **Set Significance Level:** $>\alpha$ = 0.05, 0.01, or 0.10

3. **Calculate F-Statistic:** >$F = \frac{s_1^2}{s_2^2}$ with df = >$(n_1 - 1, n_2 - 1)$

4. **Make Decision:** Compare p-value with >α

Example: Ball Bearing Production

Machine A: >$n_1 = 21$, >$s_1^2 = 0.025$

Machine B: >$n_2 = 16$, >$s_2^2 = 0.014$

F-statistic: >$F = \frac{0.025}{0.014} = 1.79$

Degrees of freedom: (20, 15)

Result: p-value = 0.25 > 0.10 → Fail to reject >H_0

Conclusion: No significant difference in variability between machines

3. Testing Population Proportions

Key Concept

Analyzing categorical data such as yes/no, true/false, pass/fail - focusing on **percentages and proportions** rather than averages.

What is a Proportion?

A fraction or percentage of a population that possesses a certain characteristic.

Examples of Proportions

- Percentage of voters supporting a political party
- Proportion of defective products in manufacturing
- Fraction of students passing a certification exam
- Percentage of website visitors who click on an ad

One-Sample Z-Test for Proportion

Z-Statistic:

$$>Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $>\hat{p}$ = sample proportion
$>p_0$ = hypothesized population proportion
$>n$ = sample size

Conditions for Valid Test

1. **Random Sample:** Sample must be randomly selected
2. **Large Sample Size:**
   - $>np_0 \geq 10$ (expected successes)
   - $>n(1 - p_0) \geq 10$ (expected failures)

Two-Sample Z-Test for Proportions

Z-Statistic:

$$>Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $>\hat{p}$ is the pooled proportion: $>\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$

Why Pooled Standard Error?

Under the null hypothesis ($>H_0 : p_1 = p_2$), we assume both populations have the same proportion. Our best estimate is to combine (pool) the data from both samples.

Example Applications

- Do BITS Pilani and BITS Hyderabad have different pass rates?

- Does a new marketing campaign result in a higher click rate?
- Is the proportion of defective products different between two factories?

4. Test of Independence (Chi-Square Test)

Key Concept

Testing whether two categorical variables are related or independent of each other.

Fundamental Question

Does knowing a person's preference for one thing give us information about their preference for another?

Example Questions

- Is there a relationship between campus (Pilani/Hyderabad) and chosen major?
- Is there an association between age group and preferred movie genre?
- Does branch choice depend on home campus?

Chi-Square Test Logic

Compare **observed counts** (what we actually see in the data) with **expected counts** (what we would expect if variables were independent).

Chi-Square Statistic:

$$>\chi^2 = \sum \frac{(O - E)^2}{E}$$

where $>O$ = observed frequency
$>E$ = expected frequency
$>\sum$ = sum over all cells

Interpretation

- **Large** $>\chi^2$ → Big difference between observed and expected → Variables are likely dependent
- **Small** $>\chi^2$ → Small difference → Variables are likely independent

- >$\chi^2 = 0$ → Perfect match → Variables are definitely independent

Four-Step Framework

1. **State Hypotheses:**
    - >H_0 : Variables are independent (no association)
    - >H_1 : Variables are dependent (association exists)
2. **Set Significance Level:** >α (typically 0.05)
3. **Calculate** >χ^2 **Statistic:** Using the formula above
4. **Make Decision:**
    - If p-value < >α → Reject >H_0 (variables are related)
    - If p-value ≥ >α → Fail to reject >H_0 (no evidence of relationship)

Example: Student Branch Preference

Question: Is there an association between campus (Pilani) and branch enjoyment?

Branch	Enjoy	Don't Enjoy
CS	40	40
Mech	60	60

Interpretation: If observed counts differ significantly from expected (equal distribution), >χ^2 will be large, indicating an association.

Why Chi-Square Test is Important

One of the most versatile statistical tools for analyzing relationships between categorical variables. Widely used in:

- Market research
- Medical studies
- Social science research
- Quality control

- Survey analysis

Module Summary

This module extended hypothesis testing from single populations to comparing multiple populations:

- **Two Population Means:** Independent t-tests for different groups, paired t-tests for same group under different conditions
- **Two Population Variances:** F-tests to compare consistency and variability between groups
- **Population Proportions:** Z-tests for categorical data involving percentages and proportions
- **Chi-Square Test:** Testing independence and association between categorical variables

These methods provide practical tools widely applicable across business, research, and real-world decision-making scenarios.

"> 1-inferences-for-two-population-means

Introduction to Experimental Design & ANOVA

Contents

1. The Need for Sound Experimental Design
In statistics, our goal is to draw meaningful conclusions from data. While observational studies allow us to find associations, only a well-designed experiment can establish a true cause-and-effect relationship. The quality of our conclusions is entirely dependent on the quality of our experimental design.

1.1 The Problem of Confounding Variables
Imagine we want to test if a new fertilizer increases plant growth. A naive approach might be to give fertilizer to one group of plants and no fertilizer to another.

Group A
Received FertilizerPlaced in Sunny WindowGroup BNo FertilizerPlaced in Shady Corner

Result: Group A grew taller.
Conclusion: Was it the fertilizer or the sun?

In this scenario, the effect of the fertilizer is mixed up, or confounded, with the effect of sunlight. We cannot separate the two. A proper experimental design aims to eliminate or account for such confounding variables.

2. The Vocabulary of Experiments

To design and discuss experiments, we use a specific set of terms. Let’s continue with our fertilizer example.

- **Experimental Units:** The individuals or objects on which the experiment is performed. – Example: The individual plants.
- **Response Variable:** The outcome of interest that is measured at the end of the experiment. – Example: The final height of each plant (in cm).
- **Factor:** An explanatory variable that is intentionally manipulated by the researcher. – Example: The ”Type of Fertilizer” is our factor.
- **Levels:** The specific values or categories of a factor. – Example: The levels could be ”Fertilizer A,” ”Fertilizer B,” and a ”No Fertilizer” control.

Treatment: The specific experimental condition applied to the units. In a single factor experiment, the treatments are the levels of the factor. – Example: The three treatments are Fertilizer A, Fertilizer B, and No Fertilizer.

3. The Three Pillars of Experimental Design

A scientifically valid experiment is built on three fundamental principles.

3.1 1. Control

Control is the effort to reduce the effects of extraneous variables (lurking variables). The simplest form of control is to use a control group, which receives no treatment or a placebo. This provides a baseline against which the effects of the other treatments can be measured.

3.2 2. Randomization

Randomization is the use of chance to assign experimental units to treatments. This is the most important principle for establishing causation. Random assignment does not eliminate the effects of extraneous variables, but it spreads them evenly among the treatment groups, preventing them from systematically biasing the results.

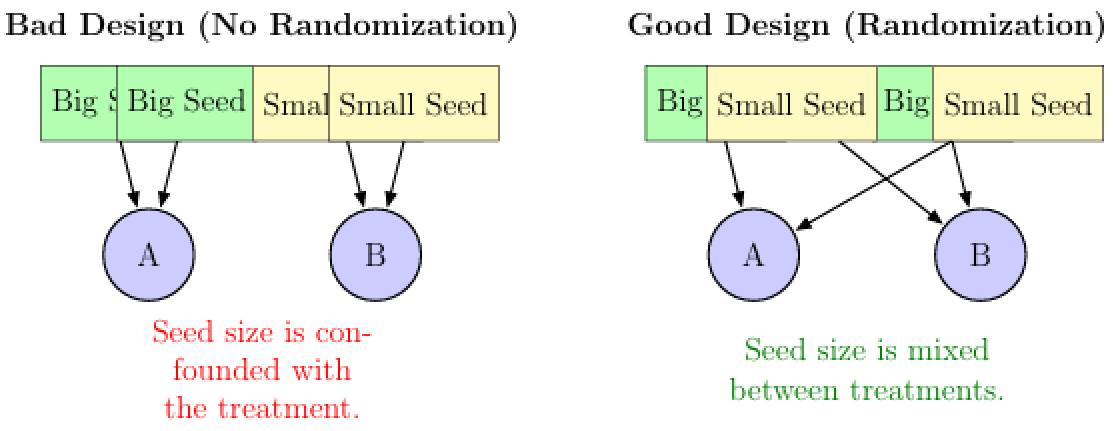


Figure 1: Random assignment prevents systematic bias.

3.3 3. Replication

Replication means applying each treatment to multiple experimental units. Replicating the experiment allows us to see the natural variation within each treatment group. Without replication, we cannot tell if a difference in the response is due to the treatment or just random chance affecting a single unit. The more replication we have, the more precisely we can estimate the true effect of a treatment.

4. From Design to Analysis: ANOVA

After a well-designed experiment is run, we need a statistical tool to analyze the results. When an experiment involves comparing the means of three or more groups (e.g., our three fertilizer treatments), the appropriate tool is Analysis of Variance (ANOVA).

4.1 The Problem with Multiple t-Tests

One might be tempted to simply run a series of two-sample t-tests to compare every pair of groups. This is incorrect. Each test has a chance of a Type I error (a false positive), controlled by α. When we run multiple tests, the overall probability of making at least one false positive (the family-wise error rate) inflates dramatically. ANOVA is a single test that avoids this problem.

4.2 The Core Logic of ANOVA: Signal vs. Noise

ANOVA is a clever test that analyzes variances to make a conclusion about means. It works by partitioning the total variation in the data into two sources and comparing them.

- **Variance BETWEEN Groups (The Signal):** This measures how much the means of our treatment groups differ from one another. If the treatments have a strong effect, the group means will be far apart, and this variance will be large. This is also called the Mean Square for Treatments (MST).
- **Variance WITHIN Groups (The Noise):** This measures the natural, random variation of the data points inside each treatment group. It represents the inherent variability we can’t explain. This is also called the Mean Square for Error (MSE).

The test statistic for ANOVA is the F-statistic, which is an intuitive ratio:
F = Signal / Noise = Variance BETWEEN Groups / Variance WITHIN Groups = MST / MSE

If the signal is much larger than the noise (a large F-statistic), we conclude that the difference between the group means is statistically significant.

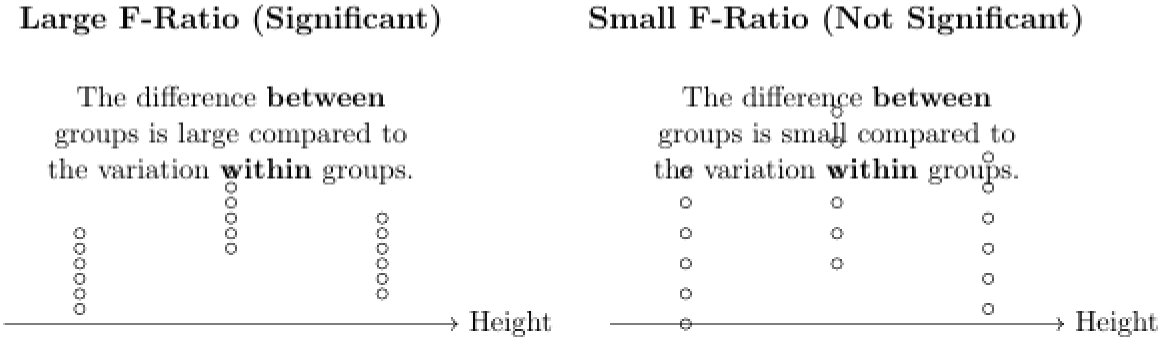

Large F-Ratio (Significant)
The difference **between** groups is large compared to the variation **within** groups.
Height
Small F-Ratio (Not Significant)
The difference **between** groups is small compared to the variation **within** groups.
Height

Figure 2: Visual intuition for the F-Ratio.

4.3 The ANOVA Procedure

1. State Hypotheses:
   • H0 : μ1 = μ2 = ··· = μk (All population means are equal).• HA : At least one population mean is different.
2. Perform the F-Test: The results are summarized in an ANOVA table, which calculates the F-statistic and its corresponding p-value.
3. Conclusion: If the p-value is less than the significance level α, we reject H0 and conclude that there is a significant difference among the group means.
4. Post-Hoc Tests: A significant ANOVA result is a ”green light” to investigate further. To find out which specific groups are different from each other, we must use a post-hoc test (e.g., Tukey's HSD).

"> reading-material-introduction-to-experimental-design-anova

---

One-Way and Two-Way Analysis of Variance (ANOVA)

Contents

1.Introduction to Analysis of Variance (ANOVA)
Analysis of Variance (ANOVA) is a powerful statistical method used to compare the means of two or more groups. It is a cornerstone of experimental data analysis. The primary goal of ANOVA is to determine whether the observed differences between group means are statistically significant or if they could have occurred simply due to random sampling variability.

The core logic of ANOVA involves partitioning the total variability in a dataset into different sources of variation. By comparing the variability between groups to the variability within groups (the “signal” vs. the “noise”), we can make an inference about the population means.

The type of ANOVA used depends on the number of factors (independent categorical variables) in the experimental design.

2.One-Way ANOVA (Single Factor Analysis)
A One-Way ANOVA is used to determine if there are any statistically significant differences between the means

of three or more independent groups. It is called “one way” because the groups are defined by a single categorical factor.

- Factor: One categorical independent variable.
- Response: One continuous dependent variable.

Example Scenario: A researcher wants to compare the effectiveness of three different teaching methods (the factor) on student exam scores (the response).

2.1 Hypotheses for One-Way ANOVA

ANOVA is an “omnibus” test, meaning it tests for an overall difference among the groups.

- Null Hypothesis (H0): There is no difference between the group means. All population means are equal. H0 : μ1 = μ2 = μ3 = ··· = μk
- Alternative Hypothesis (HA): At least one group mean is different from the others. It does not state that all means are different.

2.2 Assumptions

For the results of a One-Way ANOVA to be valid, three assumptions must be met:

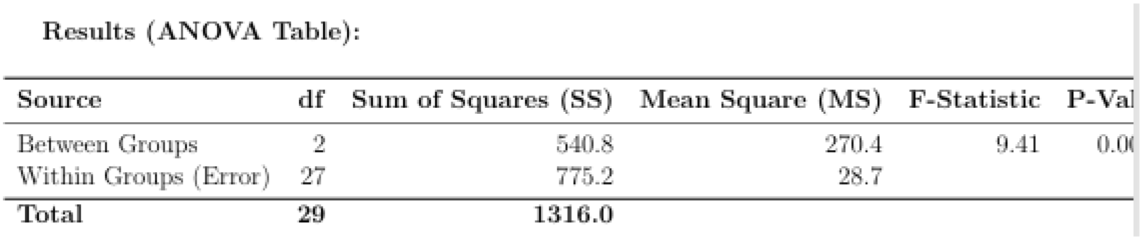
1. Independence: The observations in each group must be independent of one another.
2. Normality: The data within each group should be approximately normally distributed. (ANOVA is robust to moderate violations of this, especially with larger sample sizes).
3. Homogeneity of Variances (Homoscedasticity): The variance of the data should be equal across all groups.

2.3 Example 1: Teaching Methods (Significant Result)

Scenario: A professor wants to know if teaching method affects exam scores. 30 students are randomly assigned to one of three methods: “Lecture,” “Group Project,” and “Interactive.” Their final exam scores are recorded. (α = 0.05)

Hypotheses:

- H0 : μLecture = μGroup = μInteractive
- HA : At least one mean score is different.



Results (ANOVA Table):

Source	df	Sum of Squares (SS)	Mean Square (MS)	F-Statistic	P-Val
Between Groups	2	540.8	270.4	9.41	0.0
Within Groups (Error)	27	775.2	28.7		
Total	**29**	**1316.0**			

Decision & Conclusion: The P-Value (0.0008) is much smaller than our significance level α (0.05). Therefore, we reject the null hypothesis. There is strong statistical evidence to conclude that the choice of teaching method has a significant effect on the average student exam score.

Next Step: Since the result is significant, we would perform a post-hoc test (like Tukey's HSD) to determine which specific pairs of teaching methods are different from each other (e.g., is "Interactive" better than "Lecture"?).

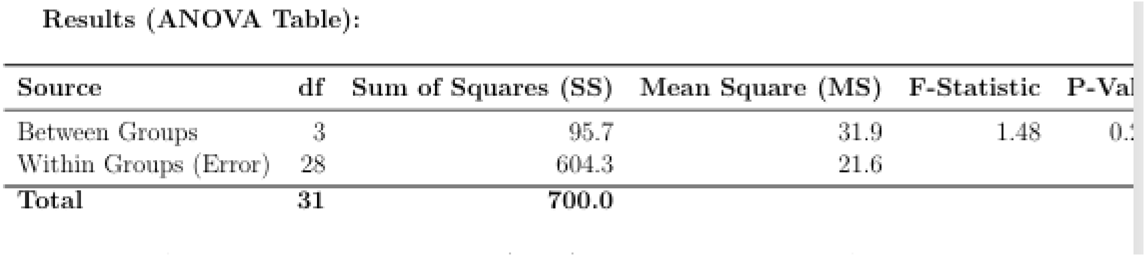
2.4 Example 2: Fertilizer Types (Non-Significant Result)

Scenario: A farmer tests four different fertilizer brands (A, B, C, D) to see if they produce different mean crop yields. Each fertilizer is applied to 8 plots of land. ($\alpha = 0.05$)

Hypotheses:

- $H0 : \mu A = \mu B = \mu C = \mu D$
- HA : At least one mean yield is different.

Results (ANOVA Table):



Results (ANOVA Table):

Source	df	Sum of Squares (SS)	Mean Square (MS)	F-Statistic	P-Val
Between Groups	3	95.7	31.9	1.48	0.
Within Groups (Error)	28	604.3	21.6		
Total	**31**	**700.0**			

Decision & Conclusion: The P-Value (0.241) is larger than our significance level α (0.05). Therefore, we fail to reject the null hypothesis. There is not enough statistical evidence to conclude that there is any difference in mean crop yield among the four fertilizer brands.

3. Two-Way ANOVA (Two Factor Analysis)

A Two-Way ANOVA extends the one-way ANOVA to include two categorical independent variables (factors). This allows us to test not only the effect of each factor individually but also whether the two factors interact with each other.

3.1 Main Effects and Interaction Effects

A two-way ANOVA analyzes three potential effects:

1. Main Effect of Factor A: The effect of the first factor on the response variable, averaged across all levels of the second factor.
2. Main Effect of Factor B: The effect of the second factor on the response variable, averaged across all levels of the first factor.
3. Interaction Effect (A x B): This is the most interesting part. An interaction effect occurs when the effect of one factor depends on the level of the other factor.

Analogy for Interaction: A new drug (Factor 1: Drug vs. Placebo) might significantly reduce blood pressure in adults but have no effect in children (Factor 2: Adult vs. Child). The effect of the drug depends on the age group. This is an interaction.

3.2 Visualizing Interactions

Interaction plots are the best way to visualize and understand interactions.

- Parallel lines suggest NO interaction. The effect of one factor is consistent across the levels of the other.
- Non-parallel or crossing lines suggest an interaction is present. The effect of one factor changes depending on the level of the other.

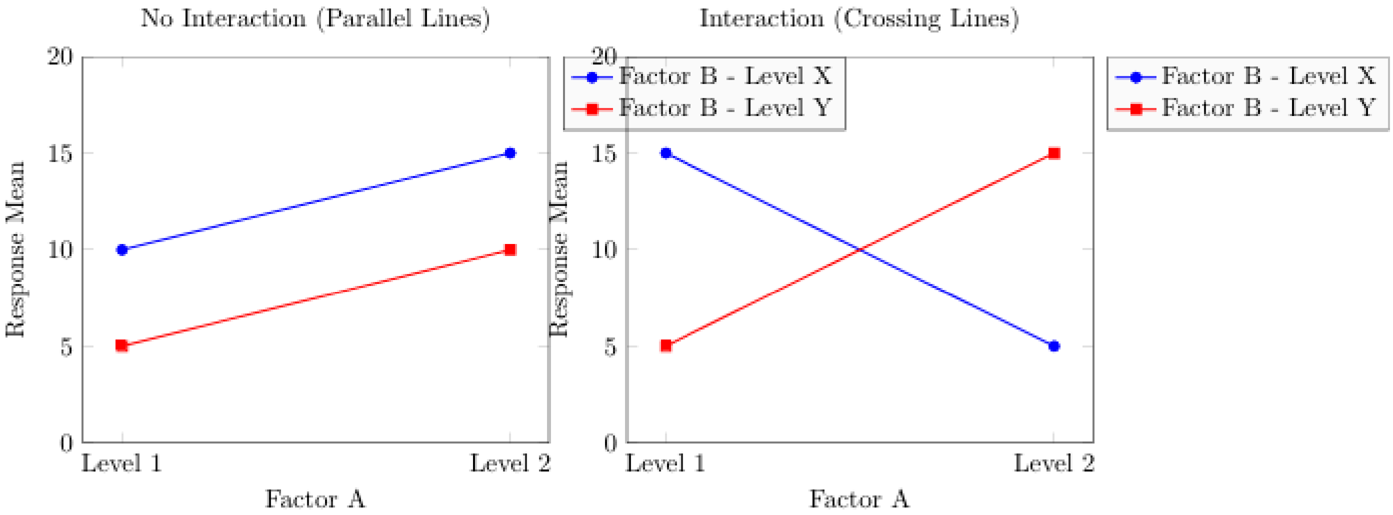


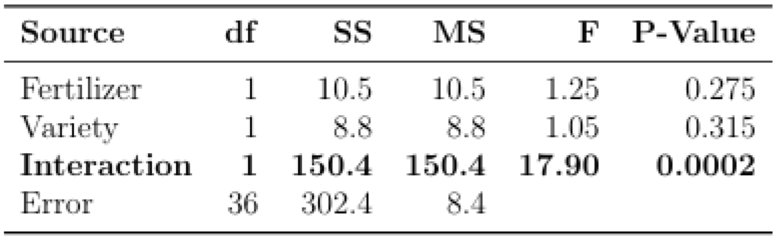
Figure 1: Visualizing Interaction Effects.

3.3 Example 1: Fertilizer & Plant Variety (Interaction)

Scenario: A botanist tests two fertilizers (A, B) on two plant varieties (X, Y) to see their effect on plant height. They test for main effects and an interaction effect. (α = 0.05)

Hypotheses: Three sets are tested: (1) for interaction, (2) for fertilizer main effect, (3) for variety main effect. The interaction is tested first.

Results (ANOVA Table):



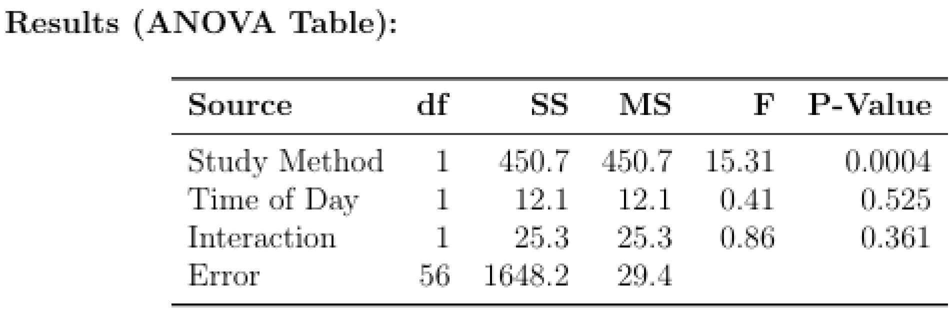
Source	df	SS	MS	F	P-Value
Fertilizer	1	10.5	10.5	1.25	0.275
Variety	1	8.8	8.8	1.05	0.315
Interaction	**1**	**150.4**	**150.4**	**17.90**	**0.0002**
Error	36	302.4	8.4		

Decision & Conclusion: The P-Value for the Interaction (0.0002) is significant. We stop here and interpret the interaction. The main effects are not meaningful in the presence of a significant interaction. The conclusion is:

“The effect of the fertilizer depends on the plant variety. Fertilizer A works best on one variety, while Fertilizer B works best on the other.”

3.4 Example 2: Study Method & Time of Day (No Interaction)

Scenario: A researcher tests two study methods (Visual, Auditory) at two times of day (Morning, Evening) to see the effect on test scores. (α = 0.05)



Results (ANOVA Table):

Source	df	SS	MS	F	P-Value
Study Method	1	450.7	450.7	15.31	0.0004
Time of Day	1	12.1	12.1	0.41	0.525
Interaction	1	25.3	25.3	0.86	0.361
Error	56	1648.2	29.4		

Decision & Conclusion: The P-Value for the Interaction (0.361) is NOT significant. This means the effect of the study method is the same in the morning and evening. We can now interpret the main effects.

- Study Method: The P-Value (0.0004) is significant. There is a difference between the visual and auditory methods.
- Time of Day: The P-Value (0.525) is not significant. There is no evidence of a difference between morning and evening study times.

The final conclusion is: “The Visual study method is significantly better than the Audi tory method, and this is true regardless of whether the studying is done in the morning or evening.”

"> reading-material-one-way-and-two-way-analysis-of-variance-anova

Statistical Modelling and Inferencing

Week 5 Reading Material

1. Introduction to Experimental Design and ANOVA

Module Overview

In previous modules, we learned how to compare one or two populations. Now we extend our analysis to compare **more than two groups** simultaneously using Analysis of Variance (ANOVA).

The "Garbage In, Garbage Out" Principle

Even the best statistical tools cannot fix problems that arise from poorly designed experiments. Good experimental design ensures we start with clean, reliable data.

Why Experimental Design Matters

Bad Example: Testing fertilizer by placing fertilized plants in sunny spots and non-fertilized plants in shady spots.

Problem: This creates a confounded experiment – you cannot distinguish between the effects of sunlight and fertilizer!

Good Design: Isolate effects by ensuring all plants receive the same amount of sunlight, varying only the fertilizer treatment.

Key Terminology

- **Response Variable:** What you are measuring (e.g., plant height, test scores)
- **Experimental Units:** The subjects being studied (e.g., plants, students, cars)
- **Factors:** Explanatory variables whose effects you are testing (e.g., type of fertilizer, teaching method)
- **Levels:** Different values of a factor (e.g., Fertilizer A, Fertilizer B, No Fertilizer)
- **Treatments:** Different levels of the factor being applied

Three Principles of Good Experimental Design

1. Control

Establish a **baseline** for comparison. Include a control group that receives no treatment (e.g., no fertilizer) to measure the effect of the treatment.

2. Randomization

Assign experimental units to groups **by chance** (like flipping a coin). This helps even out individual differences and prevents systematic bias.

3. Replication

Repeat the experiment multiple times. Even with randomization, unexpected factors can affect results. Replication helps ensure findings are reliable and not due to chance.

2. Analysis of Variance (ANOVA) Fundamentals

What is ANOVA?

Analysis of Variance (ANOVA) is a statistical method used to compare the means of **three or more groups** simultaneously.

Why not use multiple t-tests? Running multiple t-tests increases the probability of Type I error (false positives). ANOVA controls for this issue.

The Core Question

Is there a statistically significant difference in the average outcomes across the groups?

Signal vs. Noise: The Foundation of ANOVA

Signal (Between-Group Variation): Differences in means *between* the groups

- Also called **Sum of Squares Treatment (SST)** or **Sum of Squares Between (SSB)**
- Measures how much group means differ from the overall mean

Noise (Within-Group Variation): Natural variability *within* each group

- Also called **Sum of Squares Error (SSE)** or **Sum of Squares Within (SSW)**
- Measures how much individual observations differ from their group mean

F-Statistic (The Heart of ANOVA):

$$>F = \frac{\text{Variance etween roups}}{\text{Variance ithin roups}} = \frac{\text{Signal}}{\text{Noise}}$$