Robert Davidson

ST1112: Statistics

70% Exam
30% Continuous Assessment (3 parts)

Contents

1 Inferential Statistics - Interval Estimation

The ultimate goal in statistical inference is to estimate population parameters (like the mean μ) based on sample statistics (like the sample mean $\bar{X}$).

1.1 Probability vs Statistics

- **Probability** deals with known underlying processes: one starts with a model (like proportion of red vs. green jelly beans in a jar) and computes probability of specific outcomes
- **Statistics** works in reverse: one observes outcomes (sample data) and attempts to infer the underlying process or population parameters (e.g. proportion of red jellybeans)

1.2 Definitions and Concepts

Definition 1.1: Population

A **population** is the complete set of items (or individuals) of interest.

Definition 1.2: Sample

A **sample** is a subset of that population, intended to represent the population

For example the sample mean $\bar{X}$ is an estimate of the population mean μ.

Definition 1.3: Population Mean (μ)

μ represents the central tendency of a population distribution.

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

where N is the population size and x_i are the individual values in the population.

μ is sometimes called the expected value or average.

Definition 1.4: Population standard deviation (σ)

σ measures the dispersion or spread of values around the mean in a population.

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

where N is the population size and x_i are the individual values in the population.

Concept 1.1: Sampling Variation

When we take multiple samples from the same population, each sample's mean $\bar{X}$ will be different. This is variability is called **sampling variation**.

Larger sample sizes tend to reduce this variation, that is as n gros, the sample mean $\bar{X}$ becomes a better estimate of the population mean μ.

Concept 1.2: Sampling Distributions

The sample mean itself is a **random variable** because different samples yield different mean values.

The distribution of all possible sample means (of a given sample size n) is called the **sampling distribution** of the sample mean ($\bar{X}$).

Definition 1.5: Expected Value of the Sample Mean

$$E(\bar{X}) = \mu$$

This means if you averaged all possible sample means, you would get the population mean μ.

Definition 1.6: Standard Error of the Mean

$$SE = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size.

This value is called the **standard error** of the mean and measures how much the sample mean $\bar{X}$ fluctuates around the population mean μ.

Definition 1.7: Central Limit Theorem

$$\bar{X} \sim N\left(\mu, \frac{\sigma^m}{n}\right)$$

where $\bar{X}$ is the sample mean, μ is the population mean, and σ is the population standard deviation.

The **Central Limit Theorem** states that the sampling distribution of the sample mean $\bar{X}$ (the distribution of all sample means) approaches a normal distribution as the sample size n increases, **regardless of the shape of the population distribution.**

This means that for large enough sample sizes, we can use the normal distribution to make inferences about the population mean μ.

Practically, many apply the rule of thumb $n \geq 30$ to treat $\bar{X}$ as normally distributed.

Definition 1.8: Unbiased Estimators

We say a statistic T is an **unbiased estimator** of a population parameter θ, if $E(T) = \theta$.

For example, the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ because $E(\bar{X}) = \mu$.

The sample standard deviation s (using Bessel's correction, dividing by multiplying by $\frac{1}{n-1}$ rather than $\frac{1}{N}$) is an unbiased estimator of the population standard deviation σ.

1.3 Example

Example 1.1: Weekly rent

If a population mean rent is $\mu = 225$, with $\sigma = 25$ for a population sample size $n = 30$, the sample distribution of the sample mean is approximately:

$$\bar{X} \sim N\left(225, \frac{25^2}{30}\right)$$

This lets us compute probabilities for specific sample mean ranges using the normal distribution (e.g. $P(\bar{X} < 220)$).

1.4 Recap

A **sample statistic** (e.g. the sample mean $\bar{X}$) varies from one sample to another. Understanding this variation (and quantifying it via the standard error) is crucial for knowing how precise (or imprecise) an estimate really is.

If we have a large sample size n from a population with mean μ and standard deviation σ, then our sample distribution of the sample mean $\bar{X}$ is approximately normal:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

In practice, for $n \geq 30$, $\bar{X}$ can be treated as normally distributed even if the original population is not strictly normal.

2 Confidence Intervals

2.1 Confidence Intervals

Concept 2.1: Why confidence intervals?

Why do we need confidence intervals, instead of a single point estimate, like the sample mean $\bar{X}$?

A confidence interval provides a range of plausible values for the population parameter (e.g. μ) based on the sample data.

Analogy: Using a single point estimate is like trying to catch a fish wih a spear; your aim may not be perfect. Using a confidence interval is like using a net; we have a better chance of "catching" (capturing) the true population parameter.

Definition 2.1: Confidence Interval

$$\bar{X} \pm (\text{critical value}) \times SE(\bar{X})$$

where $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ is the standard error of the sample mean, $\pm$ is the margin of error.
The general formula for a desired confidence level 100(1- α)% is:

$$\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

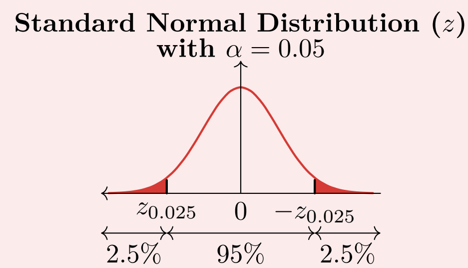
where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

Interpretation:
If we repeat the sampling process many times and construct confidence intervals from each sample, then approximately $100 \times (1-\alpha)\%$ of those intervals will contain the true population parameter μ.
In other words, you do not say *"there is a 95% chance that μ lies in my interval"*. Rather we say, **"on repeated sampling 95% of such intervals will contain the true population mean μ."**

Definition 2.2: Critical Values

The **critical value** is a z-score that corresponds to the desired confidence level.

For example, for a 95% confidence level, the critical value is $Z_{\alpha/2} = 1.96$ (where $\alpha = 0.05$). This means that 95% of the area under the normal curve lies within 1.96 standard deviations of the mean.



Example 2.1: Find crtitical value for the 95% CI

For a confidence interval of 95%, we want to find the z-score that leaves 2.5% in each tail of the normal distribution.
We want to find the z-value where the cumulative area)from the left up to that $z - score$) is $1 - 0.025 = 0.975$.
We look in the z-tables for the value closest to 0.975 and read the row and column headers to find the z-value.
The z-value is 1.96.

Example 2.2: Find the 95% confidence interval for the population mean μ given

A dataset of 103 students, of whom 71 pay rent, was used to estimate the average weekly rent μ.

- **Point estimate**: the sample mean $\bar{X} \approx 546.239$.
- **Sample standard deviation**: $s \approx 187.862$.
- **Sample size**: $n = 71$.

Confidence Interval is given by:

$$\bar{X} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}} \Rightarrow 546.239 \pm 1.96 \times \frac{187.862}{\sqrt{71}}$$

where $z_{\alpha/2} = 1.96$ for a 95% confidence level.The resulting confidence interval is:

$$(502.541, 589.938)$$

Interpretation: We are 95% confident that the true mean weekly rent for all NUI Galway students (population) is roughly 503 to 590 euros.

2.2 Higher Confidence Levels means Wider Intervals

- To achieve a **higher confidence level**, we need to increase the critical value $z_{\alpha/2}$, which in turn increases the margin of error.
- This results in a wider confidence interval, which means we are more certain that the true population parameter lies within that interval.
- Conversely a **lower confidence level** results in a smaller critical value, leading to a narrower confidence interval.

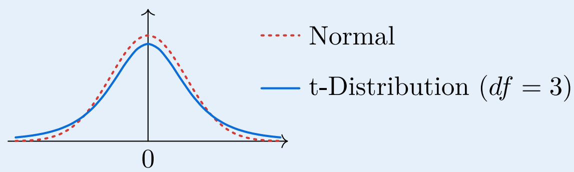
2.3 t-Distribution

Concept 2.2: Why the $t-$distribution

When the sample size is small ($n < 30$) and the population standard deviation σ is unknown, simply substituting the sample standard deviation s no longer suffices because the standard error is itself estimated with more uncertainty.

The **t-distributution has thicker** tails than the normal distributution. This extra "fatness" in the tails accounts for the additional uncertainty in using s instead of σ.

Normal vs. t-Distribution



Concept 2.3: Degrees of Freedom (df)

A t-distribution is characterized by its degrees of freedom, where

$$df = n - 1 \quad \text{for a sample mean}$$

As the sample size n increases, the t-distribution approaches the standard normal distribution. For example, for $n = 30$, $df = 29$ and the t-distribution is very close to the normal distribution.

Definition 2.3: Confidece Intervals (t-based)

$$\bar{X} \pm t_{\alpha/2,df} \times \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2,df}$ is the critical value from the t-distribution with $df = n - 1$ degrees of freedom - or from a function like `qt()` in R.

Assumption: The population itself should be approximately normally distributed when using t-based methods for small sample sizes.

Example 2.3: Finding t-critical values

Find the critical value for a 95% confidence interval with $n = 12$ (so $df = 11$).
We look for the row associated with $df = 11$ and the column associated with $\alpha/2 = 0.025$.
The critical value is:

$$t_{0.025,11} \approx 2.201$$

2.4 CI with large n, and σ unknown

The z-based critical interval is given as:

$$\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the critical value from the standard normal distribution. However, if the population standard deviation σ is unknown, we can use the sample standard deviation s as an estimate.
This gives us the following confidence interval:

$$\bar{X} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

Interpretation: around 95% of all possible 95% confidence intervals will contain the true population mean μ. We can visualize that if we drew many repeated samples, sample means will form an overlapping μ and a small fraction will not.

2.5 CI with small n, and σ unknown

If the sample size is small ($n < 30$) and the population standard deviation σ is unknown, we use the t-distribution to construct the confidence interval. This gives us the following confidence interval:

$$\bar{X} \pm t_{\alpha/2,df} \times \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2,df}$ is the critical value from the t-distribution with $df = n - 1$ degrees of freedom.
Interpretation: around 95% of all possible 95% confidence intervals will contain the true population mean μ. We can visualize that if we drew many repeated samples, sample means will form an overlapping μ and a small fraction will not.

Example 2.4: Turin Shroud

A historical cloth's age was tested by carbon dating on 12 pieces ($n = 12$). The sample mean was $x \approx 1261\ AD$ and the sample standard deviation was $s \approx 61.2\ AD$. Find the 95% confidence interval for the population mean age of the cloth.

The standard error is given by:

$$SE = \frac{s}{\sqrt{n}} = \frac{61.2}{\sqrt{12}} \approx 17.67$$

For a 95% confidence interval with $n - 1 = 11$ degrees of freedom, the critical value is $t_{0.025,11} \approx 2.201$. The confidence interval is given by:

$$\bar{X} \pm t_{\alpha/2,df} \times SE = 1261 \pm 2.201 \times 17.67$$

The resulting confidence interval is:

$$(1222, 1300)$$

Interpretation: The cloth's true average carbon-dated age is plausibly within about 1222–1300 AD. This range casts doubt on claims that the cloth dates from centuries earlier.

Example 2.5: Unathorized Computer Acess

Find 95% CI given:

- **Data**: 18 times between keystrokes
- **Sample mean**: $\bar{X} = 0.29$ seconds
- Sample standard deviation: $s = 0.0074$ seconds

$$n = 18 \Rightarrow df = 17$$

For a 95% confidence interval with $n - 1 = 18$ degrees of freedom, the critical value is $t_{0.025,17} \approx 1.740$. The resulting confidence interval is:

$$(0.2532, 0.3268)$$

Interpretation: We are 95% confident that the true mean time between keystrokes is between 0.2532 and 0.3268 seconds.

3 Transformations and the Bootstrap.

3.1 When normality is questionable

Recall that for small n, the t-distribution-based confidence interval requires data to be approximately normally distributed in the population. But many real datasets violate this assumption. - e.g. skewed data, heavily tailed data etc.

Two broad remedies exist:

- **Data transformation**: Apply a mathematical transformation to make the data more symmetric or bell shape (e.g.log-transformation). Then use t-based or z-based methods on the transformed scale.
- **Non-parametric methods**: Rely less on strict distributional assumptions. The bootstrap is a common and versatile non-parametric method approach to estimating confidence intervals and sampling variability.

3.2 Data Transformations

Purpose:

- If the data has a strongly skewed or otherwise non-normal distribution, applying a suitable transformation (e.g. $\log(x)$, $\sqrt{x}$) can help to make the data more symmetric and bell-shaped.
- After the transformation, we can apply t-based or z-based methods can be applied more safely.

Cautions:

- Finding the write transformation can be tricky; sometimes no simple transformation works well.
- Interpretation of results becomes more complex; if you compute a CI for the transformed mean, you must convert (e.g. exponentiate) the results back to the original scale.
- Despite these challenges, transformation often prove very useful in practice.

3.3 The Bootstrap

Motivation:

- Bootstrap methods do not require normality assumptions or a large n. They rely on the principle that the observed sample can server a reasonable proxy for the populations shape.
- By resampling with replacement from the original sample (many times), one creates a "bootstrap distribution" that mimics the statistic (e.g. mean, median) of interest.
- This bootstrap distribution is then used to estimate hpw the statistic varies, allowing for confidence interval construction and hypothesis testing without explicit formulas.

Basic Steps (Bootstrap Scheme)

1. **Resample with replacement**: Take a bootstrap sample of the same size n as the original dataset, but drawn from the dataset with replacement.
2. **Calculate Bootstrap statistic**: Compute the same summary measure of interest (e.g. mean, median) on the bootstrap sample.
3. **Repeat**: Repeat steps (1) and (2) many times (e.g. 1000 times) to create a distribution of the bootstrap statistic.
4. **Construct CI**: The bootstrap distribution of the resampled statistics can be used to determine the middle 95% (or chosen confidence level) as the CI bounds.

Advantages:

- Works for all kinds of statistics (mean, median, proportion, regression coefficients, etc.) even when no closed-form CI exists.
- Far fewer assumptions about the underlying population distribution.

Disadvantages:

- Computationally intensive; requires many resamples (e.g. 1000) to get a good approximation.
- Requires the sample itself to be a good representation of the population; if the sample is biased, the bootstrap may not work well.

4 Confidence Intervals for Population Proportions and Counts

Recap: Confidence Intervals for a Population Mean

- A **Confidence interval (CI)** provides a range of plausible values for a population parameter
- For a large sample ($n \geq 30$) or a known σ, we often use a z-based interval:

$$\bar{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

  or replacing σ with s if σ is unknown.
- For a small sample ($n < 30$) and unknown σ, we use a t-based interval:

$$\bar{X} \pm t_{\alpha/2,df} \times \frac{s}{\sqrt{n}}$$

  where $df = n - 1$. Provided the population is approximately normal.
- If normality is questionable, we may use transformations or bootstrapping.

4.1 Proportions

Definition 4.1: Proportion

The **proportion** is a way to express the frequency of a specific outcome (labeled as "success") relative to the total number of trials or observations.

$$p = \frac{X}{n}$$

where p is the proportion, X is the number of successes, and n is the total number of trials.

Concept 4.1: Why proportions?

Many outcomes are binary or categorical with two possible outcomes (e.g. success/failure, yes/no). Examples:

- Whether a student has a part-time job
- Whether a business has fallen victim to a scam

In such cases, we often estimate a population proportion π of successes rather than a mean μ.

4.1.1 Binomial Distribution

Concept 4.2: Bernoulli Trials

When we repeat an experiment or observation, each trial is assumed to be independent and has two possible outcomes. If each trial has a probability of π success, these trials are called **Bernoulli trials**.

If we perform n independent Bernoulli trials, the number of successes X follows a **binomial distribution** with n, the number of trials and π,the probability of success on each trial.

$$X \sim B(n, \pi)$$

This tells us how likely we are to observe a certain number of successes in n trials.

Link to Proportion:
The sample proportion p is just the normalized version of X, calculated by $p = \frac{X}{n}$. It provides a direct, interpretable measure of success rate in the sample.

4.1.2 Normal Approximation of the Sample Proportion

When is the normal approximation valid?
The approximation of the distribution of p by a normal distribution is valid when both of the following conditions are met:

$$n\pi \geq 5 \quad \text{and} \quad n(1-\pi) \geq 5$$

These conditions ensure there are enough successes and failures for the approximation to hold.

How does it work? Since X is binomially distributed, its mean is $n\pi$ and its variance is $n\pi(1-\pi)$. When we convert X into the proportion p, the mean and variance transform as follows:

- Mean of p: $E(p) = \frac{E(X)}{n} = \pi$
- Variance of p: $Var(p) = \frac{Var(X)}{n^2} = \frac{\pi(1-\pi)}{n}$

For large n (above conditions), the distribution of p can be approximated by a normal distribution:

$$p \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$$

Interpretation:
This approximation means if we were to make many samples of size n, the distribution of the same proportions would cluster around the true proportion, pi, with variability decreasing as the sample size n increases. This normality is what allows statisticians to construct confidence intervals and perform hypothesis tests on population proportions.

4.1.3 Confidence Intervals for Proportion π

For a large sample size where np and $n(1-p)$ are both greater or equal to 5, a 95% C.I for the population proportion π is given by:

$$p \pm z_{\alpha/2} \times \sqrt{\frac{p(1-p)}{n}}$$

where:

- p is the sample proportion (e.g. $\frac{X}{n}$)
- $z_{\alpha/2}$ is the critical value from the standard normal distribution (e.g. 1.96 for 95% confidence)
- The quantity under the square root is the standard error of the sample proportion.

Example 4.1: Financial Scams

A survey of $n = 80$ small businesses found that $X = 16$ had fallen victim to a financial scam. Find the 95% confidence that all small businesses have fallen victim to this scam.

- Sample proportion: $p = \frac{X}{n} = \frac{16}{80} = 0.20$
- Standard Error $= SE = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.20(1-0.20)}{80}} = \sqrt{\frac{0.20 \times 0.80}{80}} \approx 0.05$
- For a 95% confidence interval $\alpha = 0.05, z_{\alpha/2} = 1.96$.

The 95% confidence interval is given by:

$$p \pm z_{\alpha/2} \times SE = 0.20 \pm 1.96 \times 0.05$$

The resulting confidence interval is:

$$\approx (0.10, 0.30)$$

Interpretation: We are 95% confident that between 10% and 30% of all small businesses have fallen victim to this scam.

Concept 4.3: Proportion CI Test IN R

The function `prop.test(x, n, conf.level, correct=False)` gives a confidence interval for a proportion.

4.1.4 Maximizing the Standard Error

The standard error for a proportion p is given by:

$$SE = \sqrt{\frac{p(1-p)}{n}}$$

This maximizes at $p = 0.5$. Thus the worst-case margin of error for a 95% confidence interval is:

$$\approx \pm 2 \times \sqrt{\frac{0.5 \times 0.5}{n}} = \pm \frac{1}{\sqrt{n}}$$

Rule of thumb: for $n = 1000$, the margin of error is about $1/\sqrt{1000} \approx 0.03$, i.e. 3% error.

4.2 Confidence Intervals for Counts

4.2.1 Possion Setup

A count variable X, over a fixed interval (e.g. "number of emails per day") often follows a **Poisson** distribution. with parameter λ.
Recall $X \sim Poisson(\lambda)$ implies $E(X) = \lambda$ and $Var(X) = \lambda$.

4.2.2 Central Limit Theorem Approximation

For large enough λ, the Central Limit Theorem, implies the sample mean of a Possion variable is approximately normally distributed:

$$X \sim N(\lambda, \frac{\lambda}{n})$$

If we have n observations of some Poisson process, the overall mean $\bar{\lambda}$ is used to estimate the population mean λ.

Criteria: The product $n\lambda$ should be sufficiently large (e.g. ≥ 50) for the approximation to hold well.

Example 4.2: Emails per Day

Given a sample of $n = 64$ students with a mean of $\bar{\lambda} = 53$ emails per day, find the 95% confidence interval for the population mean λ.

Standard Error:

$$SE = \sqrt{\bar{\lambda}} = \sqrt{53} \approx 7.28$$

For a 95% confidence interval $\alpha = 0.05, z_{\alpha/2} = 1.96$. The 95% confidence interval is given by:

$$\bar{\lambda} \pm z_{\alpha/2} \times SE = 53 \pm 1.96 \times 7.28$$

The resulting confidence interval is:

$$(38.7, 67.3)$$

Interpretation: We are 95% confident that the true mean number of emails per day for all students is between 39 and 67.

5 Hypothesis Tests

A **hypothesis test** is a statistical framework used to evaluate claims (hypotheses) about population parameters (e.g. means, proportions).

Example scenario: A claim is made that college students have been in, on average, at least 4 exclusive relationships. Observing a random sample mean of 3.2 (with 95% CI [2.7, 3.7]) suggests that 4 is not within the plausible range, thus casting doubt on the claim.

5.1 The purposes of Hypothesis Testing

Definition 5.1: Null hypotheses (H_0)

A baseline or status-quo assumption about the population parameter (often an equality claim such as $\mu = 4$")

Definition 5.2: Alternative (or Research) Hypothesis (H_1)

A competing claim that contradicts H_0. It can be **one sided** (e.g. $\mu > 4$ or $\mu < 4$) or **two sided** (e.g. $\mu \neq 4$).

The test uses sample data to decided whether the evidence strongly contradict H_0. If so, we reject H_0 in favor of H_1. If not, we do not reject H_0 - but we conclude that the data does not provide enough evidence to reject H_0

5.2 Stages in Hypothesis Testing

Concept 5.1: Stages in Hypothesis Testing

A typical hypothesis test follows these steps:

1. State the hypotheses
   - $H0$ The null hypothesis (e.g., $\mu = \mu_0$).
   - H_1 The alternative hypothesis (e.g., $\mu \neq \mu_0$, $\mu < \mu_0$ or $\mu > \mu_0$).
2. Collect a random sample and compute the test statistic:
   - The test statistic measures how far the observed sample statistic is from the hypothesized parameter, in standardized units.
3. Identify the sampling distribution of the test statistic (usually via the Central Limit Theorem or a t-distribution):
   - For large n, use a z-approximation.
   - For smaller nn (and unknown σ), use a t-distribution with $n - 1$ degrees of freedom, assuming approximate normality of the population.
4. Decide whether the observed test statistic would be rare or common if H_0 were true:
   - p-value approach: Probability of obtaining a result at least as extreme as the actual sample result, given H_0 is true.
   - Rejection region approach: Compare the test statistic to a critical value derived from the chosen distribution and significance level α
5. Make a decision:
   - If p-value $< \alpha$, we reject H_0
   - If p-value $> \alpha$, we do not reject H_0 (We do not conclude H_0 is "proven," just that the sample does not contradict H_0 strongly.)
6. Draw a conclusion:
   - Summarize the practical meaning and state whether there is "sufficient evidence" that H_1 holds

5.3 The Test Statistic for a Mean

Concept 5.2: Formulating the Hypothesis

For example, suppose someone claims $\mu = 6.5$ hours of weekly study time for students. We can test:

- **One-sided**: $H_0 : \mu = 6.5$ **vs** $H_1 : \mu > 6.5$ (or $H_1 : \mu < 6.5$)
- **Two-sided**: $H_0 : \mu = 6.5$ **vs** $H_1 : \mu \neq 6.5$

Definition 5.3: Test Statistic

If the sample mean is $\bar{X}$ (with sample standard deviation s and sample size n), and the hypothesized mean ius μ_0, the test statistic is given by:

$$T_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

- If $n \geq 30$, T_0 is compared to a normal distribution - $N(0, 1)$.
- If $n < 30$, T_0 is compared to a t-distribution with $n - 1$ degrees of freedom, assuming the population is approximately normal.

Definition 5.4: Rejection Region

H_1	Rejection region if $n \geq 30$	Rejection region if $n < 30$
$u < \mu_0$	$T_0 < -Z_\alpha$	$T_0 < -t_{\alpha,df}$
$u > \mu_0$	$T_0 > Z_\alpha$	$T_0 > t_{\alpha,df}$
$u \neq \mu_0$	$\lvert T_0 \rvert > Z_{\alpha/2}$	$\lvert T_0 \rvert > t_{\alpha/2,df}$

5.4 Test using p-values

Instead of a formal "rejection region" many prefer the p-value approach.

Concept 5.3: Steps when testing using p-values

1. **Compute** T_0 from the data
2. **Compute p-value** probability under H_0 of of observing a test statistic as or more extreme than T_0.
3. **Compare p-value to α**:
   - p-value $< \alpha$: reject H_0 in favor of H_1
   - p-value $> \alpha$: fail to reject H_0

5.4.1 Significance Levels and p-values

We say *"the result is statistically significant at the α level"* when $p \leq \alpha$. Common values for α are: 0.05 and 0.01 (5% and 1% significance levels).

A small p-value means: *"Given H_0, it would be unlikely to observe data this extreme."* **It does not mean**: *"There is 5% chance H_0 is true.:* (p-values are not the probability of the null hypothesis itself.) If the p-value is not small, that does not prove H_0 is correct - only that the data fails to provide strong evidence against it.

5.5 Connection to Confidence Intervals

- **CI approach**: If the hypothesized value μ_0 lies outside the $(1 - \alpha)$% confidence interval, the data suggests rejecting H_0.
- If μ_0 ies inside the CI, the data is consistent with H_0.

Hence, hypothesis testing and confidence examples are closely linked. For example, if the 95% CI for a mean $[2.7, 3.7]$ and $\mu_0 =$, we see that 4 is not in the interval $\Rightarrow$ strongly consider rejecting H_0.

Example 5.1: Study time in NUI Galway

Problem:

- **Claim**: The average study time for students is $\mu = 6.5$ hours per week.
- **Sample**: 102 students, with sample mean $\bar{X} = 6.77$

Solution: We're conducting a two-sided test with $H_0 : \mu = 6.5$ and $H_1 : \mu \neq 6.5$.

1. **Hypotheses**

$$H_0 = \mu = 6.5 \quad \text{vs} \quad H_1 : \mu \neq 6.5$$

2. **Compute the test statistic**

$$SE = \frac{s}{\sqrt{n}} \approx 0.65$$

$$T_0 = \frac{\bar{X} - \mu_0}{SE} = \frac{6.77 - 6.5}{0.65} \approx 0.41$$

3. **Identify the sampling distribution**

   Since $n > 30$ we can use a a t-distribution with $df = n - 1 = 101$ degrees of freedom.

4. **Decide whether test statistic is rare or common**

   Rejection Region Approach:
   For a two-tailed test at the 5% significant level, the critical values are approximately ± 1.984. Since:

$$|T_0| \approx 0.41 \not> 1.984$$

   The test statistic is not in the rejection region.

   p-value Approach:
   Since $T_0 \approx 0.41$, we need the probability of obtaining a value as extreme as 0.41 or more, given H_0 is true. That is: $P(T > 0.41)$.

   - Find the one-tailed probability. Look up the value for $Z = 0.41$ This value is approximately $P(z > 0.41) = 1 - 0.6591 \approx 0.3409$.
   - Compute two-tailed p-value: $p = 2 \times P(z > 0.41) \approx 2 \times 0.3409 \approx 0.6818$.

   The p corresponding to $T_0 \approx 0.41$ is approximately 0.6818, which is much larger than the significance level $\alpha = 0.05$.

5. **Make a Decision**

   - If p-value $< \alpha$, reject H_0.
   - Since p-value $\approx (0.68)$ is greater than $\alpha = 0.05$ and T_0 does not lie in the rejection region, we do not reject H_0.

6. **Conclusion**

   - The data does not provide sufficient evidence to reject the claim the the true mean study time is 6.5 hours per week.
   - The sample results are consistent with a true mean study time of 6.5 hours per week. Additionally, the 95% confidence interval is:

$$6.77 \pm 1.98 \times 0.65 \approx (5.5, 8.0)$$

     which includes 6.5, reinforcing our conclusion.

Example 5.2: Golf Club Design

Problem:

- **Claim**: The true mean coefficient of restitution is $\mu > 0.82$
- **Sample**: $n = 15$ clubs, with sample mean $\bar{X} = 0.83725$ and sample standard deviation $s = 0.02456$.

Solution:
We're conducting a one-sided test with $H_0 : \mu = 0.82$ and $H_1 : \mu > 0.82$.

1. **Hypotheses**

$$H_0 : \mu = 0.82 \quad \text{vs} \quad H_1 : \mu > 0.82$$

2. **Compute the test statistic**

$$SE = \frac{s}{\sqrt{n}} = \frac{0.02456}{\sqrt{15}} \approx 0.00634$$

$$T_0 = \frac{\bar{X} - \mu_0}{SE} = \frac{0.83725 - 0.82}{0.00634} \approx 2.72$$

3. **Identify the sampling distribution**

   Since $n < 30$ and the population standard deviation is unknown, we use a t-distribution with $df = n - 1 = 14$ degrees of freedom.

4. **Decide whether test statistic is rare or common**

   Rejection Region Approach:
   For a one-sides test at the $\alpha = 0.05$ significance level, the critical value is approximately $t_{0.05,14} \approx 1.761$. Since:

$$T_0 \approx 2.72 > 1.761$$

   The test statistic is in the rejection region.

   p-value Approach:
   The p-value associated with $T_0 \approx 2.72 > 0.05$ (from the tables)

5. **Make a Decision**

   - If p-value $< \alpha$, or T_0 exceeds the critical value, reject H_0.

   Since $2.72 > 1.761$ and the p-value is less than $\alpha = 0.05$, we reject H_0.

6. **Conclusion**

   - The evidence from the sample indicates that the true mean coefficient of restitution is greater than 0.82.
   - With a one sides test, the data provides strong evidence to support the claim that the true mean coefficient of restitution is greater than 0.82.
   - The 95% confidence interval is:

$$0.83725 \pm 1.761 \times 0.00634 \approx (0.824, 0.850)$$

   which does not include 0.82, reinforcing our conclusion.

Key Takeaways 5.1

- **Null Hypothesis and Alternative**: Formulate them carefully based on research question/claim
- **Test Statistic**: For means is typically $T_0 = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$.
- **p-value**: Probability (assuming H_0) of observing a result at least as extreme as the actual sample result.
- **Significance level** α: Commonly 0.05, if p-value $< \alpha$, reject H_0.
- **Confidence Intervals Link:** If μ_0 lies outside the CI, that typically corresponds to rejecting H_0
- **Check conditions**: Independence of observations, approximate normality, random sampling, etc.
- **Practical vs Statistical Significance**: Even a small difference can be "statistically significant" with a large enough sample—but might not be practically meaningful.

5.6 Decision Outcomes in Hypothesis Testing

When conducting a hypothesis test, there are two possible decisions:

- **Reject H_0**: conclude evidence contradicts H_0.
- **Fail to reject H_0**: The sample does not provide sufficient evidence to reject H_0.

Because the true status of H_0 (true or false) is unknown in practice, then the table below shows the four outcomes.

Decision	**H_0 true**	**H_0 false**
Reject H_0	**Type I Error** (False Positive)	**Correct Decision**
Fail to reject H_0	**Correct Decision**	**Type II Error** (False Negative)

Definition 5.5: Type I Error (α)

Rejecting H_0 when its actually true

Definition 5.6: Type II Error (β)

Failing to reject H_0 when its actually false

Definition 5.7: Significance Level (α)

The probability of making a Type I error. It is the threshold for rejecting H_0. Commonly set at 0.05 or 0.01.

Definition 5.8: Power of the test ($1 - \beta$)

The probability of correctly rejecting H_0 when it is false.

Example 5.3: Wine Taster (two-sided)

- The population standard deviation of the fill volume is known to be $\sigma = 50ml$
- The sample size is $n = 100$
- Test:
$$H_0 : \mu = 750ml \text{ vs } H_1 : \mu \neq 750ml$$
- Significance level $\alpha = 0.05$
- The test statistic is z-based because n is large and σ is known.
$$Z_0 = \frac{\bar{X} - 750}{50/\sqrt{100}}$$
- The rejection region for a two-sided test at $\alpha = 0.05$ is:
$$|Z_0| > 1.96 \iff \bar{X} < 750 - 196 \times 5 = 740.2 \text{ or } \bar{X} > 750 + 1.96 \times 5 = 759.8$$

Type I Error $\boldsymbol{\alpha}$
By design:
$$P(\text{Type I Error}) = P(\text{reject} H_0 | H_0 \text{true}) = \alpha = 0.05$$

Type II Error $\boldsymbol{\beta}$
The Type II error would be failing to reject H_0, when $\mu \neq 750$. Suppose the true mean is $\mu = 740$. Then under repeated sampling $\bar{X}$ is distributed approximately:
$$\bar{X} \sim N\left(740, \frac{50^2}{100}\right) = N(740, 5^2)$$

The decision rule says "do not reject H_0" if $740.2 < \bar{X} < 759.8$. The probability of making a Type II error is:
$$\beta = P(\text{Type II Error}|\mu = 740) = P(740.2 < \bar{X} < 759.8 | X \sim N(750, 5^2))$$

Converting to standard normal:
$$\beta = P\left(\frac{740.2 - 740}{5} \leq Z \leq \frac{759.8 - 740}{5}\right) = P(0.04 < Z < 3.96) = 0.484$$

Power of the test $\boldsymbol{1 - \beta}$ The power is the probability of correctly rejecting H_0 when in fact $\mu = 740$
So,
$$\text{Power} = 1 - \beta = 1 - 0.484 = 0.516$$

Interpretation
If the true mean is 740 there is 51.6% chance this test (with $\alpha = 0.05$ and $n = 100$) will detect that the process has changed from the target of 750.

Example 5.4: Coffee Machine (One-sided)

- $\sigma = 25$ml and $n = 45$
- Test:

$$H_0 : \mu = 200 \text{ vs } H_1 : mu > 200$$

- Significance level $\alpha = 0.05$
- The test statistic is z-based because n is large and σ is known.

$$Z_0 = \frac{\bar{X} - 200}{25/\sqrt{45}}$$

- The rejection region for a one-sided test at $\alpha = 0.05$ is:

$$Z_0 > Z_0.05 \approx 1.645$$

Type I Error $\boldsymbol{\alpha}$
By definition, for a test with significance level $\alpha = 0.05$:

$$P(\text{Type I error}) = \alpha = 0.05$$

Type II Error $\boldsymbol{\beta}$
Suppose the true mean is 210, then under repeated sampling $\bar{X}$ is distributed approximately:

$$\bar{X} \sim N\left(210, \frac{25^2}{45}\right) = N(210, 3.727)$$

Reject H_0 if $\bar{X} > 206.11$. $(206.11 \approx 200 + 1.64 \times 3.727)$
Therefore, **Type II Error** (β) we do no reject H_0 when $\mu = 210$, i.e $\bar{X} \leq 206.11$:

$$\beta = P(\bar{X} \leq 206.11 | \mu = 210)$$

Converting to z scores:

$$B = P(Z < \frac{206.11 - 210}{3.727}) = P(Z < -1.04) = 0.1492$$

Power of the test $\boldsymbol{1 - \beta}$

$$1 - \beta = 0.8505$$

Interpretation
If $\mu = 210$, the probability of rejecting H_0 (detecting the mean is > 200) is 85.05%.

Definition 5.9: The Power Function

The power of a test depends on the actual true value of μ.

- **For one sided-test** $H_0 : \mu = mu_0$ vs $H_1 : \mu > mu_0$ if our rejection region is $\bar{X} > a$ then the power at a given true μ is:

$$\text{Power}(\mu) = P(\text{reject } H_0|\mu) = P(\bar{X} > a|\bar{X} \sim N(\mu, \sigma^2/n)) = 1 - \phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)$$

- **For a two-sided test** $H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$ if our rejection region is $\bar{X} < a$ or $\bar{X} > b$ then the power at a given true μ is:

$$\text{Power}(\mu) = P(\bar{X} < a|\mu) + P(\bar{X} > b|\mu) = \phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right) + 1 - \phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right)$$

By evaluating this function across different values of μ, we get a power curve showing how likely the test is to detect a shift from μ_0 to μ.

Concept 5.4: Balancing α and β

- When designing a test, we typically fix α (Type I error rate) e.g.5%
- This choice influences the probability of Type II error β (and thus the power of the test).

Trade-off: Lowering α typically raises β for a given sample size, because making the threshold for a rejection more stringent also makes it harder to detect real deviations.

Key Takeaways 5.2

- **Type I Error (α):** Rejecting H_0 when H_0 is true; this is set as the significance level.
- **Type II Error (β):** Failing to reject H_0 when it is false.
- **Power $(1-\beta)$:** The probability of rejecting H_0 given that the true parameter is not what H_0 claims. High power (typically > 0.8) is often desired.
- **Relation:** Power $= 1 - \beta$. A large β implies that the test frequently misses a real effect.
- **Implementation:** Once α and the sample size n are chosen, the power depends on the true (unknown) parameter value. The further the true mean is from the hypothesized value, the higher the power.

6 Hypothesis Test for a Proportion

6.1 Possible Forms of the Hypotheses

Definition 6.1: Hypotheses for Proportions

1. **Left-tailed (one-sided)**

$$H_0 = \pi = \pi_0 \quad \text{vs} \quad H_1 : \pi < \pi_0$$

2. **Right-tailed (one-sided)**

$$H_0 = \pi = \pi_0 \quad \text{vs} \quad H_1 : \pi > \pi_0$$

3. **Two-sided**

$$H_0 = \pi = \pi_0 \quad \text{vs} \quad H_1 : \pi \neq \pi_0$$

Where π_0 is the specific hypothesized proportion. The decisions depends on the sample data from n trials and x successes.

6.2 Test Statistic for Proportions

Definition 6.2: Test Statistic for Proportions

Provided $n\hat{\pi} \geq 5$ and $n(1-\hat{\pi}) \geq 5$, we can use a normal approximation to the distribution of the sample proportion. The **z-test statistic** is

$$T_0 = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Where:

- $\hat{p} = \frac{x}{n}$ is the observed sample proportion of successes
- π_0 is the hypothesized population proportion in H_0
- The denominator is the standard error of the sample proportion under H_0.

Under H_0, T_0 approximately follows a standard normal distribution $N(0, 1)$.

6.3 Decision Criteria and p-value

Definition 6.3: Decision Criteria

- **One-sided test**: Depending on H_1, we look for large positive values of T_0 (if $\pi > \pi_0$) or large negative values (if $\pi < \pi_0$).
- **Two-sided test**: $|T_0|$ is compared to the critical value $z_{\alpha/2}$ (often 1.96 for $\alpha = 0.05$).

Definition 6.4: p-value

The **p-value** is the probability (under H_0) of observing a test statistic as extreme or more extreme than the actual T_0.

- For a two-sided test:
$$\text{p-value} = P\big(Z \leq -|T_0|\big) + P\big(Z \geq |T_0|\big)$$
- For a right-tailed test ($\pi > \pi_0$):
$$\text{p-value} = P\big(Z \geq T_0\big)$$
- For a left-tailed test ($\pi < \pi_0$):
$$\text{p-value} = P\big(Z \leq T_0\big)$$

If the p-value $\leq \alpha$, we reject H_0. Otherwise, we fail to reject H_0.

Example 6.1: Online Communication

Setup

- **Claim**: A study suggests $\pi = 0.63$ (63% of college students spend 10+ hours/week communicating online).
- **Sample**: $n = 150$ students, among whom $x = 99$ do so, so $\hat{p} = \frac{99}{150} \approx 0.66$.
- **Hypotheses** (two-sided test):

$$H_0 : \pi = 0.63 \quad \text{vs.} \quad H_1 : \pi \neq 0.63.$$

Solution:
Test Statistic:
$$T_0 = \frac{0.66 - 0.63}{\sqrt{\frac{0.63\,(1-0.63)}{150}}} \approx 0.76.$$

Decision
At $\alpha = 0.05$, a two-sided rejection region is given by $|T_0| > 1.96$. Since $|0.76| < 1.96$, we do not reject H_0.

p-value
$$\text{p-value} = P(Z \geq 0.76) + P(Z \leq -0.76) \approx 0.4466.$$

Since $0.4466 > 0.05$, we again fail to reject H_0.

Conclusion
There is **insufficient evidence** (p-value ≈ 0.45) to conclude that the true proportion differs from 0.63.

Concept 6.1: Using `prop.test()` in R

```
prop.test(x, n, p, alternative = "two.sided", conf.level = 0.95, correct = FALSE)
```

where:

- `x` is the number of successes.
- `n` is the sample size.
- `p` is the hypothesized proportion under H_0.
- `alternative` can be `"less"`, `"greater"`, or `"two.sided"`.
- `correct = FALSE` disables the Yates continuity correction (commonly used for small sample sizes).

The function returns:

- A test statistic (given as `X-squared`, whose square root is the z-value).
- The p-value.
- A confidence interval for the true proportion.

Key Takeaways 6.1

1. **Hypothesis Test Setup**: For proportions, we specify $H_0 : \pi = \pi_0$ and check if the data strongly contradict π_0.
2. **z-Test Statistic**:
$$T_0 = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}.$$
This requires $n\hat{p} \geq 5$ and $n(1 - \hat{p}) \geq 5$ to ensure that the normal approximation is valid.
3. **Decision Rule or p-value**: Compare $|T_0|$ to $z_{\alpha/2}$ for two-sided tests (or use the appropriate one-sided cutoff), or compute the p-value.
4. **Interpretation**: A **small p-value** ($\leq \alpha$) means the data provide strong evidence that π differs from π_0, whereas a **large p-value** indicates insufficient evidence against H_0.
5. **Connection to Confidence Intervals**: If π_0 lies outside the confidence interval for π, this typically corresponds to rejecting H_0. Conversely, if π_0 lies inside, we fail to reject H_0.

7 Two Sample Comparisons

Concept 7.1: Recap: One-Sample Inference

- We learned how to make inferences about a single population parameter (mean μ or proportion π) using:
  1. **Confidence Intervals (CIs)** for providing a plausible range of values.
  2. **Hypothesis Tests** for deciding whether the parameter equals a specific value.
- While hypothesis tests yield a yes/no conclusion about a particular value, CIs reveal the magnitude and practical significance of differences.

Concept 7.2: Why Compare Two Samples?

In practice, we often want to compare parameters from **two** different populations or groups. Examples:

- Do female vs. male students differ in average study time?
- Does a new treatment for ankle fractures produce a higher average recovery score than the standard treatment?
- Does one manufacturing process have a higher mean output than another?

In such scenarios, each group represents a distinct population, and we want to compare (for means) the difference $\mu_2 - \mu_1$.

7.1 Comparing Two Independent Population Means

Definition 7.1: The Parameter of Interest

We want to estimate or test hypotheses about:

$$\mu_2 - \mu_1.$$

- A **point estimate** of this difference is $\bar{X}_2 - \bar{X}_1$.
- If $\mu_2 = \mu_1$, then their difference is zero (i.e., no difference in means).

Example 7.1: Ankle Fractures

Context: 60 patients split into two treatment groups (30 each).

- **Treatment A**: Cast immobilization.
- **Treatment B**: Early mobilization.

Outcome: AOFAS scores (a measure of ankle function/pain) at 24 weeks; range 0–100 (higher is better).
Question: Does Treatment B lead to a higher mean AOFAS score than Treatment A? **Exploratory Analysis:**

1. Summary statistics show:
   - Treatment A: $\bar{X} \approx 79.3$, $s \approx 7.0$.
   - Treatment B: $\bar{X} \approx 85.8$, $s \approx 2.8$.
2. Boxplots or violin plots indicate that Treatment B appears to have higher scores on average, with less variation, and the data distribution appears reasonably symmetric.

While these summaries are informative, a **formal two-sample inference** is needed to confirm whether the observed difference is statistically significant in the underlying populations.

7.2 Four Main Cases for Two-Sample Inference on Means

When comparing two means μ_1 versus μ_2, we choose the appropriate approach based on:

- **Sample sizes** (large vs. small).
- **Population variances** (known or unknown).
- **Equality of population variances** (if unknown, are they assumed equal?).

7.2.1 Case 1: Large Samples, Known Variances

If both population variances σ_1^2 and σ_2^2 are known and each sample is large ($n_1, n_2 \geq 30$), then by the Central Limit Theorem:

$$\bar{X}_2 - \bar{X}_1 \sim N\left(\mu_2 - \mu_1,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

$$\begin{aligned}
\textbf{Standard Error (SE) for } \bar{X}_2 - \bar{X}_1 \quad & \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
\textbf{Confidence Interval (CI)} \quad & (\bar{X}_2 - \bar{X}_1) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\
\textbf{Hypothesis Test} \text{ (we commonly test)} \quad & H_0 : \mu_2 - \mu_1 = 0 \quad \text{vs.} \quad H_1 : \mu_2 - \mu_1 \neq 0. \\
\textbf{z-test statistic} \quad & Z_0 = \frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1).
\end{aligned}$$

7.2.2 Case 2: Large Samples, Unknown Variances

If both samples are large but σ_1^2 and σ_2^2 are unknown, we estimate them with s_1^2 and s_2^2. Then:

$$\bar{X}_2 - \bar{X}_1 \approx N\left(\mu_2 - \mu_1,\ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right), \quad \text{with } \textbf{Standard Error} \quad \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

And a **z-interval** or **z-test** is used similarly, substituting s_i^2 for σ_i^2.

7.2.3 Case 3: Small Samples, Unknown Variances, Assumed Equal

When at least one sample is small ($n < 30$) and we assume $\sigma_1^2 = \sigma_2^2 = \sigma^2$:

$$\begin{aligned}
\text{Compute the } \textbf{pooled variance} : \quad & s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \\
\text{The } \textbf{Standard Error} \textit{ is} : \quad & s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\end{aligned}$$

The test statistic follows a **t-distribution** with $n_1 + n_2 - 2$ degrees of freedom:

$$\frac{(\bar{X}_2 - \bar{X}_1) - (\mu_2 - \mu_1)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}.$$

Conditions: Each population is (approximately) normally distributed and the two population variances are equal.

7.2.4 Case 4: Small Samples, Unknown Variances, Not Assumed Equal

If at least one sample is small and we do **not** assume equality of variances:

$$\text{Standard Error} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

with a **t-distribution** whose degrees of freedom are approximated by the **Welch–Satterthwaite** formula:

$$\text{df}^* = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

Conditions: Each population is approximately normal and the variances σ_1^2 and σ_2^2 may differ.

7.3 Back to the Ankle Fracture Example

Example 7.2: Ankle Fractures

- $n_1 = n_2 = 30$ (both large enough), σ_1^2, σ_2^2 unknown $\Rightarrow$ **Case 2** applies.
- Sample means:
$$\bar{X}_A = 79.30, \quad \bar{X}_B = 85.77.$$
- Sample variances:
$$s_A^2 \approx 49.39, \quad s_B^2 \approx 7.84.$$
- **Estimated difference**: $\bar{X}_B - \bar{X}_A = 6.47$.

Confidence Interval for $\mu_B - \mu_A$

$$\text{SE} = \sqrt{\frac{49.39}{30} + \frac{7.84}{30}} \approx 1.38.$$

At 95% confidence ($z_{0.025} = 1.96$):

$$6.47 \pm 1.96 \times 1.38 = (3.77,\ 9.17).$$

Interpretation: We are 95% confident that, on average, Treatment B yields between about 3.8 and 9.2 points higher AOFAS score than Treatment A.

Hypothesis Test Null Hypothesis: $H_0 : \mu_B - \mu_A = 0$. **Alternative Hypothesis**: $H_1 : \mu_B - \mu_A \neq 0$.

$$Z_0 = \frac{6.47 - 0}{1.38} \approx 4.69.$$

Since 4.69 is well above $z_{0.025} = 1.96$, we reject H_0. The **p-value** is effectively zero (less than 10^{-5}), indicating strong evidence that Treatment B's mean AOFAS score is higher than Treatment A's.

Concept 7.3: Using `t.test()` in R:

For two independent samples (Welch's method by default), use:

```
t.test(AOFAS ~ Treatment, data = ankle24.df)
```

The output gives a t-statistic, approximate degrees of freedom, p-value, and a confidence interval. For large samples, the difference between the z-approximation and Welch's t-approximation is negligible.

7.4 Comparing Variances

Sometimes we want to test if $\sigma_1^2 = \sigma_2^2$ for two populations.

- **F-test** (requires approximate normality).
- **Levene's test** (less sensitive to non-normality).

F-test:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs.} \quad H_1 : \sigma_1^2 \neq \sigma_2^2.$$

Test Statistic:

$$F_0 = \frac{s_1^2}{s_2^2}.$$

Under H_0, $F_0 \sim F_{n_1-1,\, n_2-1}$. If F_0 is too large or too small (i.e., falls outside the critical interval), we reject H_0.

Example 7.3: Paper Mills

- Mill 1: $n_1 = 13$, $\bar{X}_1 = 26.31$, $s_1 = 8.36$.
- Mill 2: $n_2 = 18$, $\bar{X}_2 = 19.89$, $s_2 = 4.85$.

$$F_0 = \frac{8.36^2}{4.85^2} \approx 2.97.$$

If F_0 is outside the interval $(F_{0.025}, F_{0.975})$, we reject H_0. Here, with a p-value ≈ 0.04, we reject the null hypothesis that $\sigma_1^2 = \sigma_2^2$.

Key Takeaways 7.1

- **Two-sample methods** extend the one-sample inference concepts to compare means from two groups.
- Depending on sample size, variance knowledge, and normality assumptions, we choose the appropriate formula (z-approximation or t-approximation).
- **Confidence intervals** remain crucial for assessing the practical importance of observed differences.
- If normality or these assumptions are questionable, consider **transformations** or **bootstrap/non-parametric** approaches.

8 Bootstrap and Permutation Test

8.1 When Normality (or Other Assumptions) Is Questionable

In comparing two populations with standard parametric tests (z-test or t-test), we typically assume:

- The sample sizes are sufficiently large, or
- The underlying data are (approximately) normally distributed.

If these assumptions fail (for instance, if data are skewed, have heavy tails, or sample sizes are small) we can:

1. **Transform the data** to approximate normality (e.g., log-transform), or
2. **Use non-parametric/resampling methods**, such as
   - **Bootstrap** confidence intervals, or
   - **Permutation tests** for hypothesis testing.

These methods rely less on strict distributional assumptions, making them particularly useful in real-world scenarios where normality is questionable.

8.2 Bootstrap CI for the Difference of Two Means

8.2.1 Rationale

The **bootstrap** is a resampling method that constructs an empirical distribution of a statistic (e.g., the mean difference) by sampling *with replacement* from the observed data. For two independent samples, one resamples each sample separately and computes the difference in their means repeatedly, building a "bootstrap distribution" of those differences.

8.3 Steps for Two-Sample Mean Difference

Concept 8.1: Steps for Two-Sample Mean Difference

1. **Separate the data** by group A and group B.
2. **Resample with replacement** from group A's data and from group B's data to create "bootstrap samples" of the same sizes as the originals.
3. **Compute the mean difference** of these two bootstrap samples.
4. **Repeat** the above many times (e.g., 1000 or 10,000 times).
5. **Use percentiles** of the resulting bootstrap distribution (e.g., the 2.5th and 97.5th percentiles for a 95% CI) to form the confidence interval for $\mu_B - \mu_A$.

Example 8.1: Ankle Fractures

Using R's `infer` package, one can execute:

```
specify(AOFAS ~ Treatment) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "diff in means", order = c("B", "A"))
```

- This generates 1000 bootstrap differences in means (Group B minus Group A).
- Then `get_ci(..., level = 0.95)` provides the 95% CI from the empirical distribution of differences.
- A histogram (via `visualize()`) displays the spread of bootstrap differences with shading for the CI bounds.

Concept 8.2: Interpretation of Bootstrap CI

- If the entire bootstrap CI is above zero, it suggests that μ_B is likely larger than μ_A.
- Likewise, if the entire interval is below zero, it suggests that μ_B is likely smaller than μ_A.
- If it straddles zero, there is no strong evidence of a difference.

8.4 Permutation Test for Two-Sample Comparison

8.4.1 Motivation

A **permutation test** is a non-parametric hypothesis test that makes minimal assumptions about the data distribution. It tests whether two samples come from populations with the **same** mean (or more generally, the same distribution).

- **Null hypothesis** (H_0): The two groups are exchangeable—there is no real difference in their underlying population means (or distributions).
- **Alternative hypothesis** (H_1): The groups differ (e.g., $\mu_1 \neq \mu_2$).

Concept 8.3: Steps for Permutation Test

1. **Observed difference**: Compute the actual mean difference $\bar{X}_1 - \bar{X}_2$ in the sample data.
2. **Permute group labels**: Shuffle or reassign the observed data points randomly between the two groups, effectively destroying any "true" difference by mixing the data. Under H_0, any labeling is equally plausible.
3. **Compute new difference**: For each permutation, compute the mean difference in the permuted dataset.
4. **Build the null distribution**: After many permutations, collect the distribution of permuted mean differences; this is the empirical distribution under the null (i.e., assuming no difference).
5. **p-value**: Calculate the fraction of permuted differences that are as or more extreme than the observed difference.

$$\text{p-value} = \frac{\sum(\text{permuted difference} \geq |\text{observed difference}|)}{\text{number of permutations}}.$$

   For a two-sided test, count how many permuted differences are either $\geq +|\text{observed diff}|$ or $\leq -|\text{observed diff}|$.

If the empirical p-value is below the significance threshold (α), we conclude there is evidence of a real difference between groups.

Example 8.2

- Suppose you have 6 observations from group 1 and 6 from group 2. The observed mean difference is 2.5.
- By randomly reassigning the 12 data points to "group 1" vs. "group 2" many times, we generate a **null distribution** of differences typically centered near 0.
- If only about 3% of those permuted differences exceed 2.5 in magnitude, then the p-value is 0.03, providing evidence that the observed 2.5 difference is unlikely to arise under the null hypothesis of no difference.

8.5 Comparison of Bootstrap CI vs. Permutation Test

- **Bootstrap**: Provides a **confidence interval** for the difference of means by resampling each group independently.
- **Permutation**: Provides a **hypothesis test** by shuffling group labels to generate a null distribution for the test statistic.
- Both methods are **non-parametric** (distribution-free) and rely on the sample data to represent the underlying populations well.

9 Key Takeaways

1. **Non-parametric options**: If normality is doubtful or sample sizes are small, bootstrap and permutation methods can be more robust than standard t-tests.
2. **Bootstrap for Confidence Intervals**:
   - Straightforward to implement.
   - Relies on the data themselves to estimate sampling variability.
3. **Permutation Test for Hypothesis Testing**:

- Constructs a null distribution by reassigning the data to groups at random.
- Yields an **empirical p-value** for the difference in means (or other statistics).

4. **Minimal assumptions**: Both methods require fewer assumptions than parametric tests, particularly about the shape of the distribution.
5. **Implementation**:

   - In R, packages like `infer` or custom code loops can handle bootstrapping and permutation procedures.
   - Typically repeated many times (e.g., 1000, 10,000) to obtain stable estimates of intervals or p-values.

In sum, **bootstrap** and **permutation** methods provide powerful alternatives for two-sample comparisons when standard assumptions (like normality or large sample sizes) may not hold. They expand the analyst's toolbox for making statistically sound inferences under less restrictive conditions.