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.

1.5 Confidence Intervals

Concept 1.3: 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 1.9: 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:

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

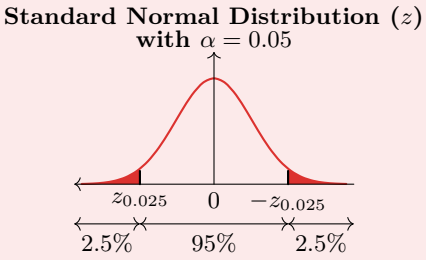
where $z_{\alpha/2}$ is the critical value from the standard normal distribution.

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 1.10: 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 1.2: 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 1.3: 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.

1.6 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.

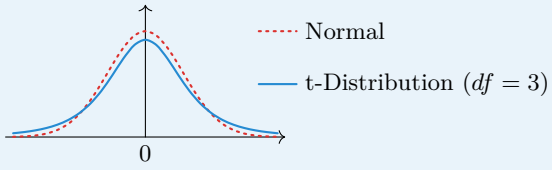
1.7 t-Distribution

Concept 1.4: 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 1.5: 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 1.11: 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 1.4: 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:

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

1.8 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.

1.9 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 1.5: 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 1.6: Unathorized Computer Acess

Find 95% CI given:

- **Data**: 18 times between keystokes
- **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.

1.10 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.

1.11 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.

1.12 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.

2 Inferential Statistics - Hypothesis Testing

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.

2.1 Proportions

Definition 2.1: 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 2.1: 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.

2.1.1 Binomial Distribution

Concept 2.2: Bernoulli Trials

When we repeat an experiment or observation, each trial is assumed to be independant 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.

2.1.2 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.

2.1.3 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 2.1: 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 2.3: Proportion CI Test IN R

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

2.1.4 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.

2.2 Confidence Intervals for Counts

2.2.1 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 $Var(X) = \lambda$.

2.2.2 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 2.2: 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.