Day 5-6 – Sampling Distribution:

The sampling distribution of a statistic (such as a sample mean or sample proportion) is the distribution of all values of the statistic when all possible samples of the same size n are taken from the same population. (The sampling distribution of a statistic is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Sampling Distribution of Mean: -

The sampling distribution of the mean is the distribution of sample means, with all samples having the same sample size n taken from the same population. (The sampling distribution of the mean is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Sampling Distribution of Variance: -

The sampling distribution of the variance is the distribution of sample variances, with all samples having the same sample size n taken from the same population. (The sampling distribution of the variance is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Sampling Distribution of Proportion: -

The sampling distribution of the proportion is the distribution of sample proportions, with all samples having the same sample size n taken from the same population.

If you organise all of those statistics of each different sample into a table, this is a sampling distribution.

❖ Central Limit Theorem: -

The central limit theorem states that if you have a population with mean **μ** and standard deviation **σ** and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

- The mean of the sampling distribution is the mean of the population, μ.
- Standard deviation [standard error] of $\frac{\sigma}{\sqrt{n}}$
- For a population with any distribution, if **n > 30**, then the sample means have a distribution that can be approximated by a normal distribution with mean and standard deviation
- If and the original population has a normal distribution, then the sample means $n \leq 30$ have a normal distribution with mean and standard deviation.
- If and the original population does not have a normal distribution, then the central limit theorem do not apply.

The Central Limit Theorem and the Sampling Distribution of $\bar{x}$

Given

1. The random variable x has a distribution (which may or may not be normal) with mean μ and standard deviation σ.
2. Simple random samples all of the same size n are selected from the population. (The samples are selected so that all possible samples of size n have the same chance of being selected.)

Conclusions

1. The distribution of sample means $\bar{x}$ will, as the sample size increases, approach a *normal* distribution.
2. The mean of all sample means is the population mean μ.
3. The standard deviation of all sample means is $\sigma/\sqrt{n}$.

Practical Rules Commonly Used

1. If the original population is *not normally distributed,* here is a common guideline: For $n > 30$, the distribution of the sample means can be approximated reasonably well by a normal distribution. (There are exceptions, such as populations with very nonnormal distributions requiring sample sizes larger than 30, but such exceptions are relatively rare.) The distribution of sample means gets closer to a normal distribution as the sample size n becomes larger.
2. If the original population is *normally distributed,* then for *any* sample size n, the sample means will be normally distributed.

Estimation: - Use information from the sample to estimate (or predict) the parameter of interest.

Two common estimation methods are point and interval estimates.

- **Point Estimates: -** An estimate for a parameter that is one numerical value. An example of a point estimate is the sample mean or the sample proportion.

p = Population proportion of Success

$\hat{p}$ = Sample proportion of Success

$\hat{p} = \frac{x}{n}$ **(x = Number of Success, n = Number of Trials)**

$\hat{q}$ = Sample proportion of Failure

$\hat{q} = 1 - \hat{p}$

IMP: $\hat{p}$ is a point estimate for P

- **Interval Estimates: -** An **interval estimate** gives you a range of values where the parameter is expected to lie. A **confidence interval** is the most common type of interval estimate.

Statistical Tests: - Use information from the sample to determine whether a certain statement about the parameter of interest is true. Statistical tests are also referred to as ***hypothesis tests***

Statistics And Probability for Data Science And Analyst

Confidence Intervals: - A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter. A confidence interval is sometimes abbreviated as CI.

Note!

We should stop here and explain why we use the estimated standard error and not the standard error itself when constructing a confidence interval. The answer is because, typically, the population values are not known. Take, for example, the standard error of the sample proportion. It is...

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

If the goal is to estimate p and p is unknown, we would also then have to estimate the standard error. In this case the estimated standard error is...

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

For the case for estimating the population mean, the population standard deviation, σ, may also be unknown. When it is unknown, we can estimate it with the sample standard deviation, s. Then the estimated standard error of the sample mean is...

$$\frac{s}{\sqrt{n}}$$

They Have:

1} Confidence Level: - This tells you how confident you are that the actual population parameter will be in the range(interval) you are giving me.

1 - α, α denotes the complement of a confidence level.

Most Common are,

Confidence Level	α	Critical Value, $z_{\alpha/2}$
90%	0.10	1.645
95%	0.05	1.96
99%	0.01	2.575

2} Critical Value: - A **Z score** that separates the 'Likely' region from the 'Unlikely' region**.**

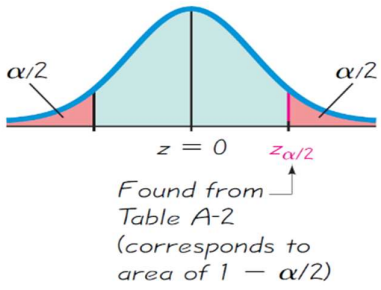


Figure 7-2 Critical Value $z_{\alpha/2}$ in the Standard Normal Distribution

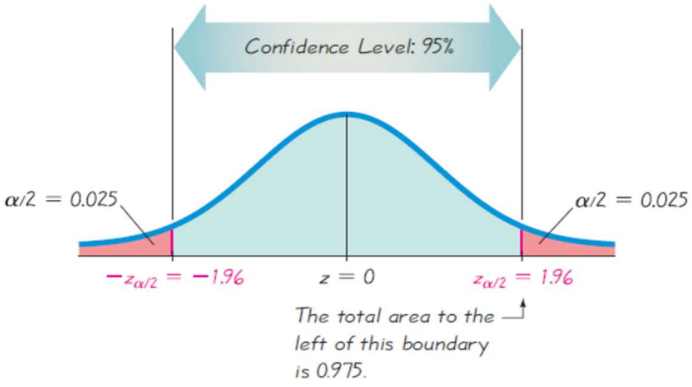


Figure 7-3 Finding $z_{\alpha/2}$ for a 95% Confidence Level

Point Estimate for the Population Proportion: -

Point Estimate of the Population Proportion

$\hat{p}$ = # of successes in the sample of size n

Confidence Interval for the Population Proportion: -

If np and $n(1-p)$ are greater than five, then $\hat{p}$ is approximately normal with mean, p , standard error $\sqrt{\frac{p(1-p)}{n}}$

Construct Confidence Intervals: -

To construct a confidence interval, we're going to use the following 3 steps:

- ***Check Conditions: -*** Check all conditions before using the sampling distribution of the sample proportion.
  We previously used np and $n(1-p)$. But is not known. Therefore, for the confidence interval, We will use

$$n\,p > 5 \; And$$
$$n(1-\hat{p}) > 5$$

  Condition Should Be satisfied: -
  If the sample comes from a Normal distribution, then the sample mean will also be normal. In this case $\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$, will follow a t-distribution with $n-1$ degrees of freedom.

If the sample does not come from a normal distribution but the sample size is large ($\boldsymbol{n \geq 30}$), we can apply the Central Limit Theorem and state that $\overline{X}$ is approximately normal. Therefore, $\frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$ will follow a t-distribution with $\boldsymbol{n-1}$ degrees of freedom.

- ***General Form: -*** The general form of the confidence interval is **'point estimate** $\pm$ $\boldsymbol{M \times \hat{SE}}$**(estimate)'**. The point estimate is the sample proportion, $\hat{\boldsymbol{p}}$**,** and the estimated standard error is $\hat{\boldsymbol{SE}}(\hat{\boldsymbol{p}}) = \sqrt{\frac{\hat{\boldsymbol{p}}(\boldsymbol{1}-\hat{\boldsymbol{p}})}{\boldsymbol{n}}}$. If the conditions are satisfied, then the sampling distribution is approximately normal. Therefore, the multiplier comes from the normal distribution. This interval is also known as the one-sample *z*-interval for $\boldsymbol{p}$, or the Normal Approximation confidence interval for $\boldsymbol{p}$.

$(1-\alpha)$100% Confidence Interval for the Population Mean, μ

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where the *t*-distribution has $df = n-1$. This interval is also known as the one-sample t-interval for the population mean.

$(1-\alpha)100\%$ confidence interval for the population proportion, p

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

where $z_{\alpha/2}$ represents a z-value with $\alpha/2$ area to the right of it.

General notes about the confidence interval...

- **The $\pm$ in the formula above means "plus or minus". It is a shorthand way of writing $(\hat{\boldsymbol{p}} - \boldsymbol{z_{\alpha/2}}\sqrt{\frac{\hat{\boldsymbol{p}}(\boldsymbol{1}-\hat{\boldsymbol{p}})}{\boldsymbol{n}}}, \hat{\boldsymbol{p}} + \boldsymbol{z_{\alpha/2}}\sqrt{\frac{\hat{\boldsymbol{p}}(\boldsymbol{1}-\hat{\boldsymbol{p}})}{\boldsymbol{n}}})$.**
- **It is centered at the point estimate, $\hat{\boldsymbol{p}}$.**
- **The width of the interval is determined by the margin of error.**
- **You must determine the multiplier.**

- **Interpret the confidence interval: -** Applying the template from earlier in the lesson we can say we are $(1-\alpha)100\%$ confident that the population proportion is between $\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ and $\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. The examples will go into more detail regarding the interpretation of the confidence interval.

We are $\boldsymbol{(1-\alpha)100\%}$ confident that the population mean, μ , is between $\boldsymbol{\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}}}$ **and** $\boldsymbol{\overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}}$**.**

Estimating a Population Mean σ Is Known: -

The confidence interval for $\boldsymbol{p}$, the $(\mathbf{1}-\boldsymbol{\alpha})100\%$ confidence interval for the population mean $\boldsymbol{\mu}$ is...

$$\boldsymbol{P}(|\frac{\overline{\boldsymbol{x}}-\boldsymbol{\mu}}{\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}}| \leq \boldsymbol{z_{\alpha/2}}) = \mathbf{1}-\boldsymbol{\alpha}$$

A little bit of algebra will lead you to...

$$\boldsymbol{P}(\overline{\boldsymbol{x}}-\boldsymbol{z_{\alpha/2}}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} \leq \boldsymbol{\mu} \leq \overline{\boldsymbol{x}}+\boldsymbol{z_{\alpha/2}}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}) = \mathbf{1}-\boldsymbol{\alpha}$$

The $(1-\alpha)$100% confidence interval for isμ:

$$\overline{\boldsymbol{x}} \pm \boldsymbol{z_{\alpha/2}}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

Estimating a Population Mean σ is Unknown: - If you don't know σ , you can't use a Z-score. Instead, we use T Score.

T-Score: we need to have random sample size.

$n \geq 30$ or sample is from a normally distributed population.

$$\boldsymbol{t} = \frac{\overline{\boldsymbol{X}}-\boldsymbol{\mu}}{\frac{\boldsymbol{s}}{\sqrt{\boldsymbol{n}}}}$$

Degree of Freedom: - The number of independent pieces of information used to calculate a statistic.

df = n – 1

df = sample size – 1

Note! When the sample size is larger than 30, the t-values are not that different from the z-values. Thus, a crude estimate for $t_{0.05}$ with 34 degrees of freedom is $z_{0.05} = 1.645$. Although it is a crude estimate, when software is available, it is best to find the t values rather than use the z.

T-distribution: -

Properties of the t-distribution

1. t is symmetric about 0
2. t-distribution is more variable than the Standard Normal distribution
3. t-distributions are different for different degrees of freedom (d.f.).
4. The larger n gets (or as n goes to infinity), the closer the t-distribution is to the z.
5. The meaning of t_α is the t-value having the area "α" to the right of it.

Method	Conditions
Use normal (z) distribution.	σ known and normally distributed population *or* σ known and $n > 30$
Use t distribution.	σ not known and normally distributed population *or* σ not known and $n > 30$
Use a nonparametric method or bootstrapping.	Population is not normally distributed and $n \leq 30$.

❖ **Hypothesis Testing: -** Hypothesis testing, sometimes called significance testing, is an act in statistics whereby an analyst tests an assumption regarding a population parameter.
Z = ($\bar{x}$ – μ0) / (σ /√n).

Types of Hypothesis testing: -

1}. Null Hypothesis Testing: - The null hypothesis is the claim that there's no effect in the population. The null hypothesis is typically denoted as $\boldsymbol{H_0}$.

Eg: - H_0: μ = 5, H_0: p = 0.5 We test the null hypothesis directly in the sense that we assume (or pretend) it is true and reach a conclusion to either reject it or fail to reject it.

Note : How to test a hypothesis

1.Begin by assuming the assume H_0 is the true statement.

2. Then use evidence to reach a conclusion

a. Reject H_0 - I have enough evidence to prove H_0 is wrong. (Note: Cannot accept H_0- We may reject the hypothesis but it doesn't necessarily mean we accept the hypothesis).

b. Fail to reject H_0 I don't have enough evidence to prove H_0 is wrong.

2}. **Alternate Hypothesis Testing:** - Alternate Hypothesis claims that there's an effect in the population. The alternative hypothesis is typically denoted as $\boldsymbol{H_a}$ **or** $\boldsymbol{H_1}$.

The symbolic form of the alternative hypothesis must use one of these symbols: < , >, ≠

Errors in Hypothesis Testing:

When testing a null hypothesis, we arrive at a conclusion of rejecting it or failing to reject it. Such conclusions are sometimes correct and sometimes wrong (even if we do everything correctly). The table summarises the two different types of errors that can be made, along with the two different types of correct decisions. We distinguish between the two types of errors by calling them type I and type II errors.

- **Type I error -** A Type I error means rejecting the null hypothesis when it's actually true. The symbol **α (Alpha)** is used to represent the probability of a type I error. Is also Known as significance level.
- **Type II error -** When we fail to reject the null hypothesis when the null hypothesis is false. The symbol **β** (beta) is used to represent the probability of a type II error.

(Routine For Fun), we can easily remember that a type I error is RTN: Reject True Null (hypothesis), whereas a type II error is FRFN: Fail to Reject a False Null (hypothesis).

α (alpha) probability of a type I error (the probability of rejecting the null hypothesis when it is true).
β (beta) probability of a type II error (the probability of failing to reject a null hypothesis when it is false).

		True State of Nature	
		The null hypothesis is true	The null hypothesis is false
Decision	We decide to reject the null hypothesis	**Type I error** (rejecting a true null hypothesis) $P(\text{type I error}) = \alpha$	Correct decision
	We fail to reject the null hypothesis	Correct decision	**Type II error** (failing to reject a false null hypothesis) $P(\text{type II error}) = \beta$

 # How do we decide whether to reject the null hypothesis?

- If the sample data are consistent with the null hypothesis, then we do not reject it.
- If the sample data are inconsistent with the null hypothesis, but consistent with the alternative, then we reject the null hypothesis and conclude that the alternative hypothesis is true.

Six Steps for Hypothesis Test: -

1. **Set up the hypotheses and check conditions**:
2. **Decide on the significance level, α:**
3. **Calculate the test statistic:**
4. **Calculate probability value (p-value), or find the rejection region:**
5. **Make a decision about the null hypothesis:**
6. **State an overall conclusion**:

One-sample Hypothesis: -

One Sample Proportion

Research Question	Is the population proportion different from p_0?	Is the population proportion greater than p_0?	Is the population proportion less than p_0?
Null Hypothesis, H_0	$p = p_0$	$p = p_0$	$p = p_0$
Alternative Hypothesis, H_a	$p \neq p_0$	$p > p_0$	$p < p_0$
Type of Hypothesis Test	Two-tailed, non-directional	Right-tailed, directional	Left-tailed, directional

*p_0 is the hypothesized population proportion

One Sample Mean

Research Question	Is the population mean different from μ_0?	Is the population mean greater than μ_0?	Is the population mean less than μ_0?
Null Hypothesis, H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
Alternative Hypothesis, H_a	$\mu \neq \mu_0$	$\mu > \mu_0$	$\mu < \mu_0$
Type of Hypothesis Test	Two-tailed, non-directional	Right-tailed, directional	Left-tailed, directional

*μ_0 is the hypothesized population mean

Test Statistic: -

For Proportion p:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p * q}{n}}}$$

For Mean μ:

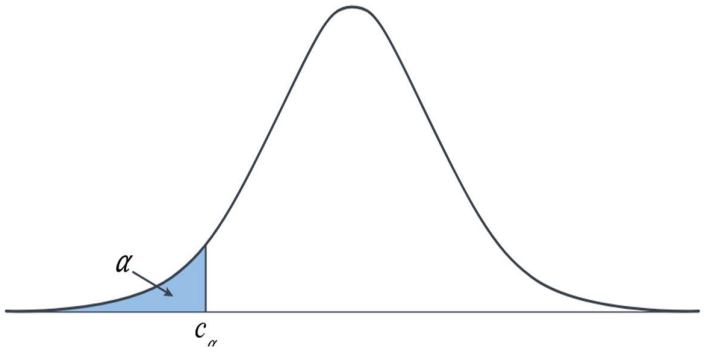
$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad If\ \sigma\ is\ unknown$$

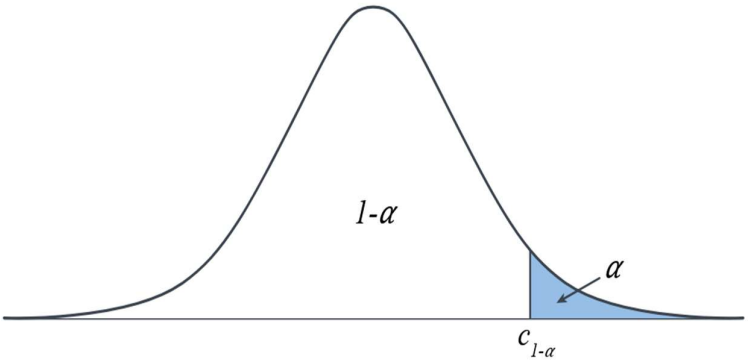
The test statistic is a value used in making a decision about the null hypothesis. It is found by converting the sample statistic. (Such as the sample Proportion, sample mean, sample standard deviation) to a score with the assumption that the null hypothesis is true.

Rejection Region: - If our test statistic falls into this region Reject H_0.

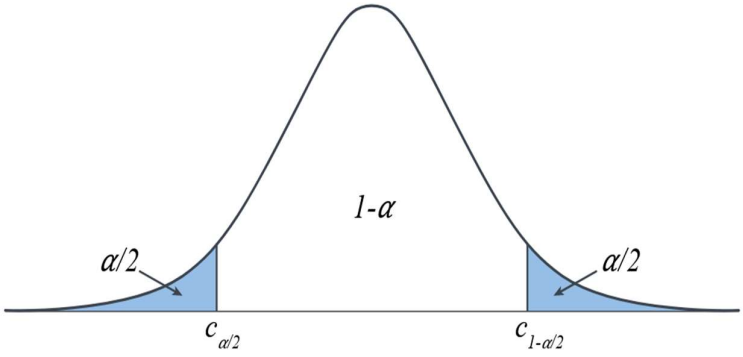
- **Left-Tailed Test: -** Reject H_0 if the test statistics is **less than or equal to** the critical value (c_α).



- **Right-Tailed Test: -** Reject H_0 if the test statistic is **greater than or equal to** the critical value ($c_{1-\alpha}$).



- **Two-Tailed Test: -** Reject H_0 if the absolute value of the test statistic is **greater than or equal to** the absolute value of the critical value ($c_{\alpha/2}$).



P-Value: - The *p*-value (or probability value) is the probability that the test statistic equals the observed value or a more extreme value under the assumption that the null hypothesis is true. The p-value is a probability statement based on the alternative hypothesis.

- **Left Tailed: -** If H_a is left-tailed, then the *p*-value is the probability the sample data produces a value equal to or less than the observed test statistic.
- **Right Tailed: -** If H_a is right-tailed, then the *p*-value is the probability the sample data produces a value equal to or greater than the observed test statistic.
- **Two-Tailed: -** If H_a is two-tailed, then the *p*-value is **two times** the probability the sample data produces a value equal to or greater than the **absolute value** of the observed test statistic.

Using The significance level: -

- **If our p-value is** less than or equal to α**, then there we** reject the null hypothesis**.**
- **If our p-value is** greater than α **, fail to** reject the null hypothesis.

T-test: - A ***t* test** is a statistical test that is used to compare the means of two groups.

Assumption: -

- **The data are continuous**
- **The distribution is approximately normal.**
- **The sample data have been randomly sampled from a population.**
- **There is homogeneity of variance (i.e., the variability of the data in each group is similar).**

Types of T-test: -

- **One sample: - I**f there is one group being compared against a standard value (e.g., comparing the acidity of a liquid to a neutral pH of 7), perform a **one-sample *t* test**.
- **Two-sample: -** If the groups come from two different populations (e.g., two different species, or people from two separate cities), perform a **two-sample *t* test** (a.k.a. **independent *t* test**).
- **Paired Test: -** If the groups come from a single population (e.g., measuring before and after an experimental treatment), perform a **paired *t* test**.

Z-Test: -

Z-tests can be defined as **statistical hypothesis testing techniques** that are used to quantify the hypothesis testing related to claim made about the **population parameters such as mean and proportion.**

Four Variables are involved in the Z-test.

- An independent variable that is called the “sample” and assumed to be normally distributed;
- **Sample size is greater than 30.**
- A dependent variable that is known as the test statistic (Z) and calculated based on sample data.
- The **standard deviation** and **mean** of the population is **known.**

- Different types of Z-test that can be used for performing hypothesis testing
- A significance level or “alpha” is usually set at 0.05 but can take the values such as 0.01, 0.05, 0.1

Four Types Of Z-test: -

- **One sample z-test: -** We perform the One-Sample z-Test when we want to compare **a sample mean with the population mean.**
- **Two sample z-test: -** We perform a Two Sample z-test when we want to compare **the mean of two samples.**
- **One proportion Z-test: -** there are two groups and compares the value of an observed proportion.
- **Two proportion z-test: -** A two proportion z test is conducted on two proportions to check if they are the same or not.

One-way Anova: - One-Way ANOVA ("analysis of variance") compares the means of two or more independent groups. One-Way ANOVA is a parametric test.

- Dependent variable that is continuous (i.e., interval or ratio level)
- Independent variable that is categorical (i.e., two or more groups)
- These distributions have the **same variance**.

Test Statistic for One-Way ANOVA

For more than two populations, the test statistic, F, is the ratio of between group sample variance and the within-group-sample variance. That is,

$$F = \frac{\text{between group variance}}{\text{within group variance}}$$

Two-way Anova: - A two-way ANOVA is used to estimate how the mean of a quantitative variable changes according to the levels of two categorical variables.

Chi-Square Test: - chi-square test **is a statistical test for categorical data.**

Types of Chi-Square: -

- **Chi-square goodness of fit: -** You can use a **chi-square goodness of fit test** when you have **one** categorical variable. It allows you to test whether the frequency distribution of the categorical variable is significantly different from your expectations.
- **Chi-square test of independence: - chi-square test of independence** when you have **two** categorical variables. It allows you to test whether the two variables are related to each other.