DS JULY 2022 Batch
Module 10 – Statistics Advanced

Topics

- Population and Sample
- Sampling Distribution and Central Limit Theorem
- Standard Error
- Confidence Interval
- Hypothesis testing: One tail, Two tail and p-value
- Z-test, t-test

Population and Sample

Population Vs Sample

Population

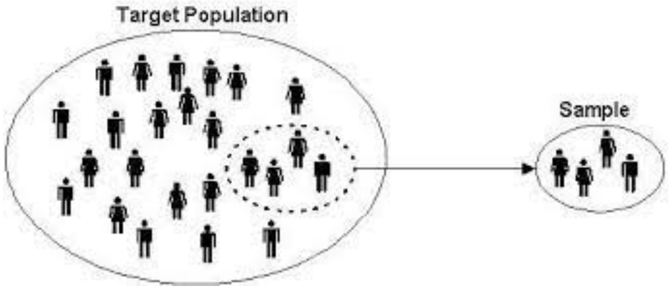
Generally, population refers to the people who live in a particular area at a specific time.

But in statistics, population refers to data on your study of interest. It can be a group of individuals, objects, events, organizations, etc.

Sample

A sample is defined as a smaller and more manageable representation of a larger group.

A subset of a larger population that contains characteristics of that population. A sample is used in statistical testing when the population size is too large for all members or observations to be included in the test.



Parameter vs Statistic

Parameter

Parameters are numbers that describe the properties of entire populations.

Statistic

Statistics are numbers that describe the properties of samples.

Example :

The average income for the India is a population parameter.

The average income for a sample drawn from the India is a sample statistic.

Inferential Statistics

- Descriptive statistics describes data (for example, a chart or graph) and inferential statistics allows you to make predictions ("inferences") from that data.
- With inferential statistics, you take data from samples and make generalizations about a population.
- For example, you might stand in a mall and ask a sample of 100 people if they like shopping on weekends. You could make a bar chart of yes or no answers (that would be descriptive statistics) or you could use your research (and inferential statistics) to reason that around 75-80% of the population (all shoppers in all malls) like shopping on weekends.

There are two main areas of inferential statistics:

- **Estimating parameters :** This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).
- **Hypothesis tests :** This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Sampling Distribution

Sampling distribution

- A sampling distribution is a probability distribution of a statistic that is obtained through repeated sampling of a specific population.
- The sampling distribution of a given population is the distribution of frequencies of a range of different outcomes that could possibly occur for a statistic of a population.
- Each sample has its own sample mean, and the distribution of the sample means is known as the sample distribution.

Sample Mean as a Random Variable

- Whenever we want to do some statistical investigation we take different random samples.
- Because each sample is different, each sample will have a different mean and standard deviation. Therefore, sample statistics are random variables that can be described with distributions.

Central Limit Theorem

Central limit theorem

Assume that $\boldsymbol{x_1}, \boldsymbol{x_2}$,..., $\boldsymbol{x_n}$ are n observations from a random sample.

1) If $\boldsymbol{x_1}, \boldsymbol{x_2}$,..., $\boldsymbol{x_n}$ are from a normal distribution with mean μ and standard deviation **σ** then sample mean will follow normal distribution with mean = μ and standard deviation = $\frac{\sigma}{\sqrt{n}}$
2) If $\boldsymbol{x_1}, \boldsymbol{x_2}$,..., $\boldsymbol{x_n}$ are not from a normal distribution but sample size is large then also central limit theorem holds.

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

Point Estimate

Point Estimate

Point Estimate

A point estimate is a single statistic which is computed from sample data and is used to estimate a population parameter.

Example

Suppose Best Buy Inc. wants to estimate the mean age of buyer who purchase LCD HDTV televisions.

They select a random sample of 75 recent purchases, determine the age of each buyer, and compute the mean age of the buyers in the sample.

The **mean of this sample is a point estimate** of the population mean.

Here

Sample size : 75

Population : All the buyer who purchase LCD HDTV televisions

Sample : All the buyer of 75 recent purchases

Parameter : mean age of all buyer who purchase LCD HDTV televisions

Statistic : mean age of all the buyer of 75 recent purchases

Standard Error

Standard Error

Standard Error

- The standard error (SE) of a statistic is the approximate standard deviation of a statistical sample population.
- The standard error is a statistical term that measures the accuracy with which a sample distribution represents a population by using standard deviation.
- In statistics, a sample mean deviates from the actual mean of a population; this deviation is the standard error of the mean.

The smaller the standard error, the more representative the sample will be of the overall population.

The formula for standard deviation is given by

$$SE = \frac{\sigma \;\longleftarrow \text{Standard deviation}}{\sqrt{n} \;\longleftarrow \text{Number of samples}}$$

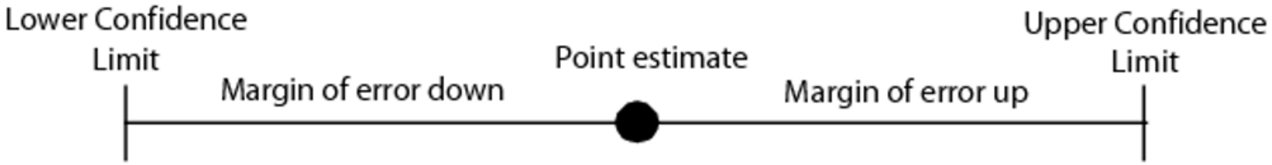
Confidence Interval

Confidence Interval

A point estimate is a single value estimate of a parameter. For instance, a sample mean is a point estimate of a population mean. 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.

- A confidence interval displays the probability that a parameter will fall between a pair of values around the mean.
- Confidence intervals measure the degree of uncertainty or certainty in a sampling method.
- They are also used in hypothesis testing and regression analysis.
- Statisticians often use p-values in conjunction with confidence intervals to gauge statistical significance.
- They are most often constructed using confidence levels of 95% or 99%.

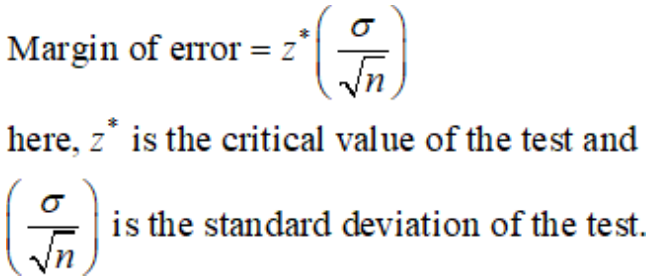


Margin of Error

The margin of error is defined as the range of values below and above the sample statistic in a confidence interval.

A margin of error tells you how many percentage points your results will differ from the real population value.

For example, a 95% confidence interval with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time.



$$\text{Margin of error} = z^{*}\left(\frac{\sigma}{\sqrt{n}}\right)$$

here, z^{*} is the critical value of the test and

$\left(\frac{\sigma}{\sqrt{n}}\right)$ is the standard deviation of the test.

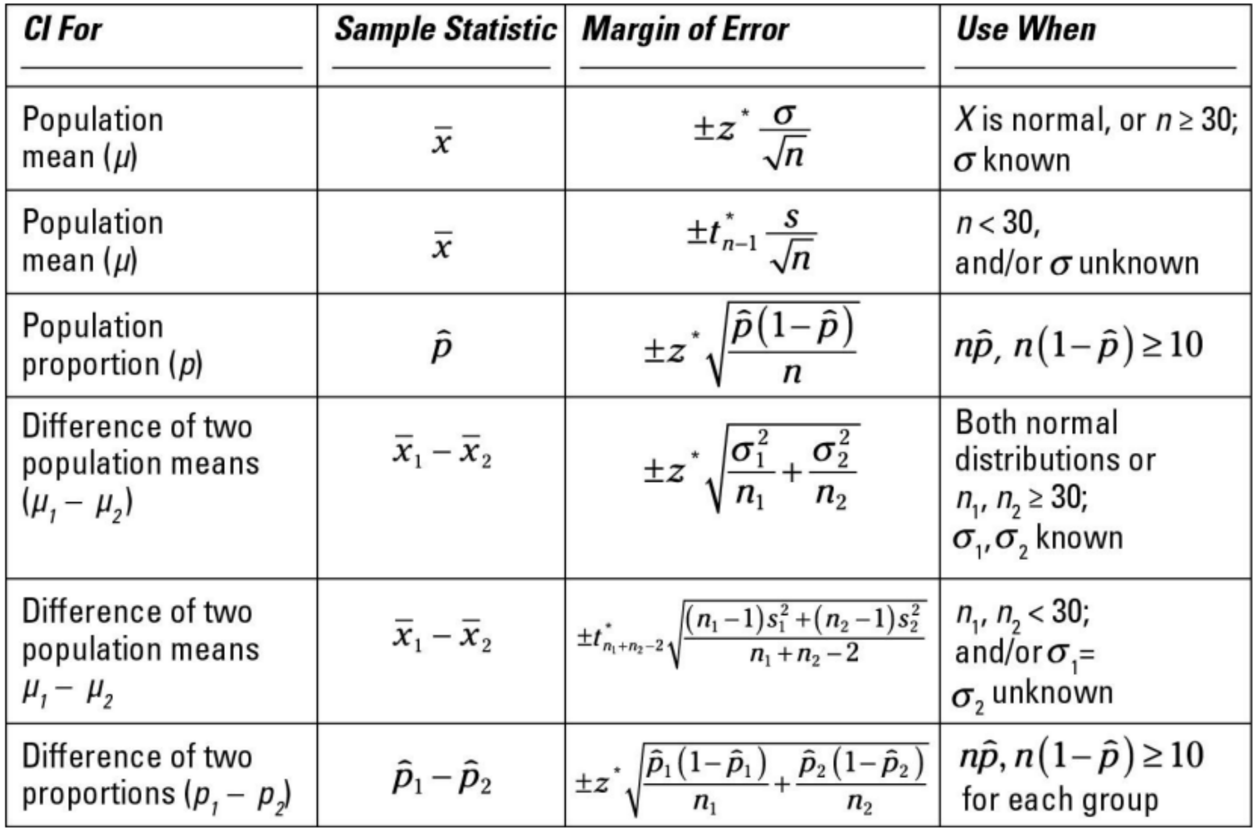
Example

A poll might report that a certain candidate is going to win an election with 51 percent of the vote. Plus, the confidence level is 95 percent and the error is 4 percent. If we assume that the poll was repeated using the same techniques, then the pollsters would expect the results to be within 4 percent of the stated result (51 percent) 95 percent of the time. In other words, 95 percent of the time they would expect the results to be between:

•51 – 4 = 47 percent and

•51 + 4 = 55 percent

Confidence Intervals



CI For	*Sample Statistic*	*Margin of Error*	*Use When*
Population mean (μ)	$\bar{x}$	$\pm z^{*}\frac{\sigma}{\sqrt{n}}$	X is normal, or $n \geq 30$; σ known
Population mean (μ)	$\bar{x}$	$\pm t^{*}_{n-1}\frac{s}{\sqrt{n}}$	$n < 30$, and/or σ unknown
Population proportion (p)	$\hat{p}$	$\pm z^{*}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$n\hat{p},\ n(1-\hat{p}) \geq 10$
Difference of two population means ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$\pm z^{*}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	Both normal distributions or $n_1, n_2 \geq 30$; σ_1, σ_2 known
Difference of two population means $\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\pm t^{*}_{n_1+n_2-2}\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$	$n_1, n_2 < 30$; and/or $\sigma_1 =$ σ_2 unknown
Difference of two proportions ($p_1 - p_2$)	$\hat{p}_1 - \hat{p}_2$	$\pm z^{*}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$	$n\hat{p}, n(1-\hat{p}) \geq 10$ for each group

Hypothesis Testing

Hypothesis Testing

Hypothesis Testing

Statistical Hypothesis: It is a claim about the value of a parameter or population characteristic

- Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically validating an assumption that we make about the population parameter.
- Hypothesis-testing procedures rely on using the information in a random sample from the population of interest.

Example:

- H: μ = 75 cents, where μ is the true population average of daily per-student candy expenses in US high schools

we assume that we need some statistical way to prove those. we need some mathematical conclusion what ever we are assuming is true.

Need of Hypothesis Testing

Need of Hypothesis Testing

- Hypothesis testing is an essential procedure in statistics.
- It is done to confirm our observation about the population using sample data, within the desired error level.
- A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.
- Through hypothesis testing, we can determine whether we have enough statistical evidence to conclude if the hypothesis about the population is true or not.

Components of a hypothesis test

Components of a hypothesis test

1.Formulate the hypothesis to be tested.

2. Determine the appropriate test statistic and calculate it using the sample data.

3. Comparison of test statistic to critical region to draw initial conclusions.

4. Calculation of p-value.

5. Conclusion, written in terms of the original problem.

Null vs Alternative Hypothesis

Null vs Alternative Hypotheses

- In any hypothesis-testing problem, there are always two competing hypotheses under consideration:

  1.The status quo (null) hypothesis

  2. The research (alternative) hypothesis
- The objective of hypothesis testing is to decide, based on sample information, if the alternative hypotheses is actually supported by the data.
- We usually do new research to challenge the existing (accepted) beliefs

Is there strong evidence for the alternative?

- This initially favored claim (H0) will not be rejected in favor of the alternative claim (Ha or H1) unless the sample evidence provides significant support for the alternative assertion.
- If the sample does not strongly contradict H0, we will continue to believe in the plausibility of the null hypothesis.
- The two possible conclusions:

1) Reject H0.
2) Fail to reject H0.

Examples

Example of Hypotheses

The alternative to the null hypothesis H0: θ = m will look like one of the following three assertions:

1. Ha: θ ≠ m
2. Ha: θ > m (in which case the null hypothesis is θ ≤ m)
3. Ha: θ < m (in which case the null hypothesis is θ ≥ m)

- The equality sign is always with the null hypothesis.
- The alternate hypothesis is the claim for which we are seeking statistical proof

Test Statistic

Test Statistic

- A test statistic is a rule, based on sample data, for deciding whether to reject H0.
- The test statistic is a function of the sample data that will be used to make a decision about whether the null hypothesis should be rejected or not.

Example: Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards.

H0: p = .10 vs Ha: p < .10
Our data is a random sample of n = 200 boards from company B.

What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

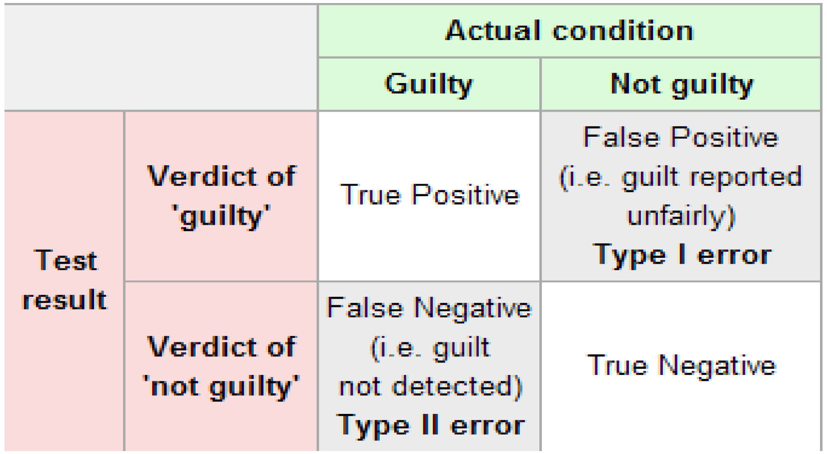
Errors in Hypothesis Testing

Definition

- A type I error is when the null hypothesis is rejected, but it is true.
- A type II error is not rejecting H0 when H0 is false.

This is very similar in spirit to our diagnostic test examples

- False negative test = type I error
- False positive test = type II error



		Actual condition	
		Guilty	**Not guilty**
Test result	**Verdict of 'guilty'**	True Positive	False Positive (i.e. guilt reported unfairly) **Type I error**
	Verdict of 'not guilty'	False Negative (i.e. guilt not detected) **Type II error**	True Negative

Critical Region and Critical Value

Critical Region and Critical Value

Critical Region

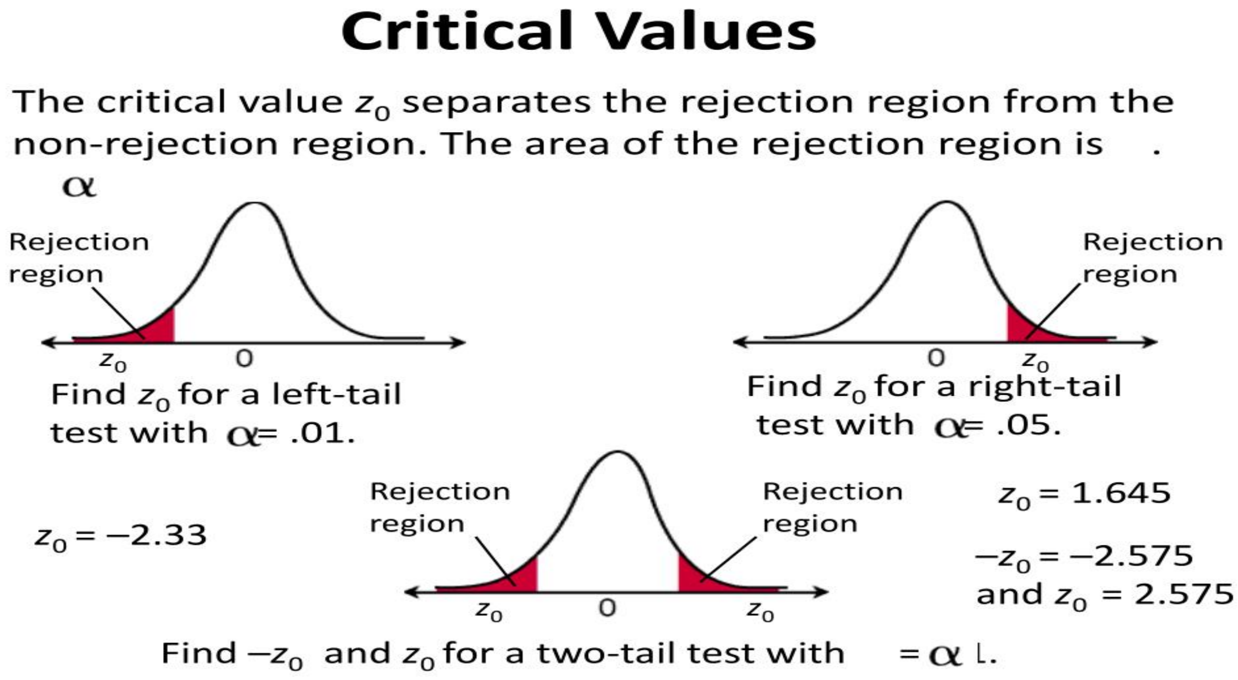
- A critical region, also known as the rejection region, is a set of values for the test statistic for which the null hypothesis is rejected.
- if the observed test statistic is in the critical region, then we reject the null hypothesis and accept the alternative hypothesis.

Critical Values

- The critical value at a certain significance level can be thought of as a cut-off point.
- If a test statistic on one side of the critical value results in accepting the null hypothesis, a test statistic on the other side will result in rejecting the null hypothesis.

Critical Region and Critical Value

Critical values and Region in case of Normal distribution



Critical Values

The critical value z_0 separates the rejection region from the non-rejection region. The area of the rejection region is .

α

Rejection
region
z_0
0
Rejection
region
0
z_0
Rejection
region
Rejection
region
z_0
0
z_0

Find z_0 for a left-tail test with α= .01.

$z_0 = -2.33$

Find z_0 for a right-tail test with α= .05.

$z_0 = 1.645$

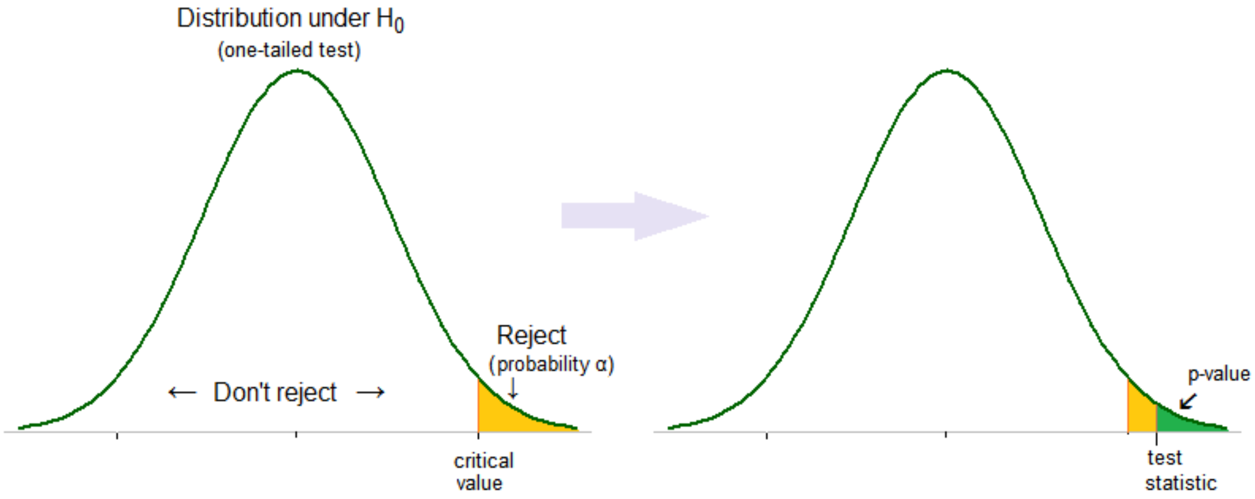
Find $-z_0$ and z_0 for a two-tail test with $= \alpha$ l.

$-z_0 = -2.575$
and $z_0 = 2.575$

p-Value

p-value

- In statistics, the p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.
- The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.
- A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.



Tests For Means (Large samples, Standard Deviation Known)

1.1 One Sample Z-Test

Assumptions:

- Population data is continuous.

- Population follows a standard normal distribution.

- The **standard deviation** of the population is known.

- Samples are independent of each other.

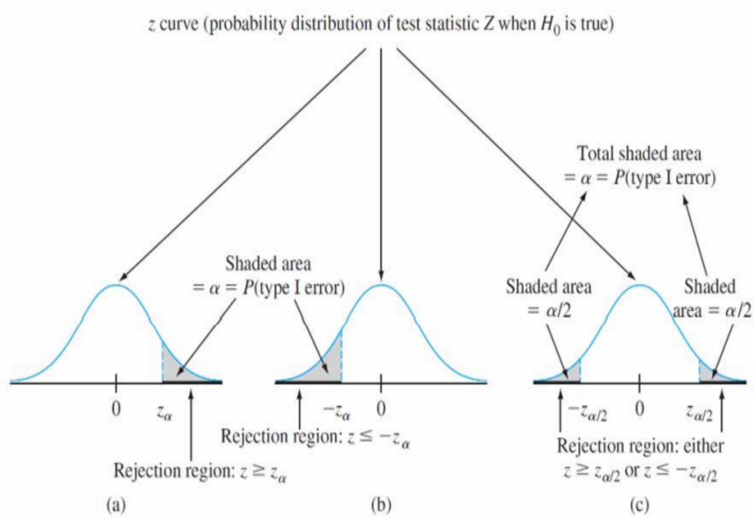
- The sample should be randomly selected from the population.

1.1 One Sample Z-Test

Null hypothesis: $H_0: \mu = \mu_0$

Test statistic value : $z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a: \mu > \mu_0$	$z \geq z_\alpha$ (upper-tailed test)
$H_a: \mu < \mu_0$	$z \leq -z_\alpha$ (lower-tailed test)
$H_a: \mu \neq \mu_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)



z curve (probability distribution of test statistic Z when H_0 is true)
Shaded area $= \alpha = P$(type I error)
Total shaded area $= \alpha = P$(type I error)
Shaded area $= \alpha/2$
Shaded area $= \alpha/2$
0
z_α
$-z_\alpha$
0
$-z_{\alpha/2}$
0
$z_{\alpha/2}$
Rejection region: $z \leq -z_\alpha$
Rejection region: either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$
Rejection region: $z \geq z_\alpha$
(a)
(b)
(c)

Rejection regions for z tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

1.2 Two Sample Z-Test

Assumptions:

- Population data is continuous.

- Each population follow standard normal distribution.

- The **standard deviation** of each of the population is known.

- Samples within both populations are independent of each other.

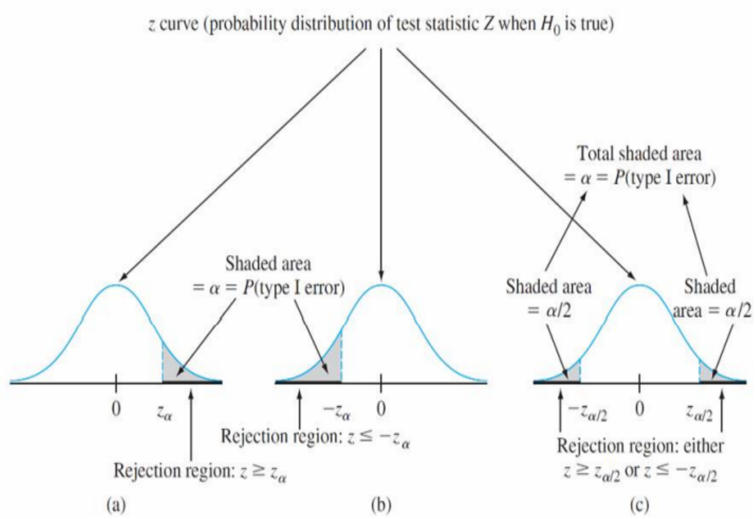
- The samples should be randomly selected from the population.

1.2 Two Sample Z-Test

Null Hypothesis : $H_0 : \mu_1 = \mu_2$

Test Statistic Value : $Z = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Alternative Hypothesis	**Rejection Region for Level α Test**
$H_a : \mu_1 > \mu_2$	$z \geq z_\alpha$ (upper-tailed test)
$H_a : \mu_1 < \mu_2$	$z \leq -z_\alpha$ (lower-tailed test)
$H_a : \mu_1 \neq \mu_2$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed test)


z curve (probability distribution of test statistic Z when H_0 is true)
Shaded area
$= \alpha = P$(type I error)
Total shaded area
$= \alpha = P$(type I error)
Shaded area
$= \alpha/2$
Shaded
area $= \alpha/2$
0
z_α
Rejection region: $z \geq z_\alpha$
(a)
$-z_\alpha$
0
Rejection region: $z \leq -z_\alpha$
(b)
$-z_{\alpha/2}$
0
$z_{\alpha/2}$
Rejection region: either
$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$
(c)

Rejection regions for z tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

Tests For Means (Small samples, Standard Deviation Unknown)

2.1 One Sample t-test

Assumptions:

- Population data is continuous.
- Population follows a standard normal distribution.
- Samples are independent of each other.
- The sample should be randomly selected from the population.

2.1 One Sample t-test

The One-Sample t Test

Null hypothesis: H_0: $\mu = \mu_0$

Test statistic value: $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

Alternative Hypothesis	**Rejection Region for a Level α Test**
H_a: $\mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
H_a: $\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
H_a: $\mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

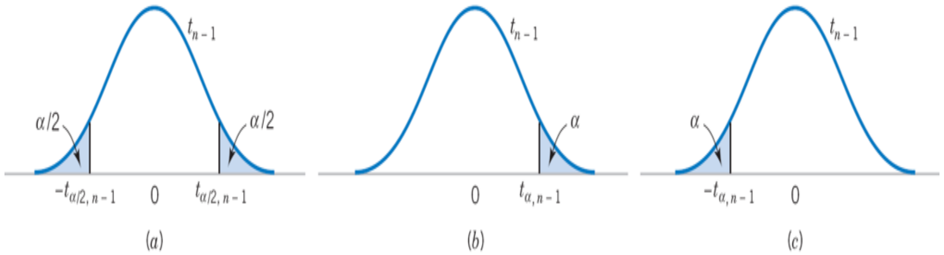


Figure 4-19 The distribution of T_0 when H_0: $\mu = \mu_0$ is true, with critical region for (a) H_1: $\mu \neq \mu_0$, (b) H_1: $\mu > \mu_0$, and (c) H_1: $\mu < \mu_0$.

One Sample Sign Test

What if assumptions are violated?

t-test and Z-test are parametric tests which make certain assumptions about the distribution of the data.
If these assumptions are not satisfied, then we using parametric tests is inappropriate.
In such cases we can use relevant nonparametric test.

One Sample sign test
One sample sign test is used to test hypotheses concerning the median of a continuous distribution.

Requirements:

- A random sample of independent measurements for a population with unknown median.
- The variable of interest is continuous.
- The Variable of interest is measured on at least ordinal scale.
- The observations are independent.

One Sample Sign Test

Null Hypothesis: $H_0 : \theta = \theta_0$

Procedure:

- Find the ‘+’ and ‘–’ signs for the given distribution. Put a ‘+’ sign for value greater than the mean value, a ‘–’ sign for a value smaller than the mean value and a ‘0’ for a value equal to the mean value.
- Let T^+ = Number of positive signs
  T^- = Number of negative signs
  T = min (T^+, T^-)
- Obtain the critical value (k) at appropriate level of significance by using the formula:

$$\sum_{i=0}^{k} \binom{n}{i} (0.5)^{n-i} \leq \alpha$$

Decision Criteria :

1) $H_a : \theta < \theta_0$,Reject H_0 if $T^+ \leq$ k.
2) $H_a : \theta > \theta_0$,Reject H_0 if $T^- \leq$ k.
3) $H_a : \theta \neq \theta_0$,Reject H_0 if T ≤ k.

One Sample Wilcoxon-Signed rank test

One Sample Wilcoxon-Signed rank test
The sign test utilizes only the signs of differences between the observed values and hypothesized median whereas this test uses the magnitude of the differences and hence we first rank the differences in order of absolute size.

Requirements:

- A random sample of independent measurements for a population with unknown median.
- The variable of interest is continuous.
- The Variable of interest is measured on at least interval scale.
- The observations are independent.

One Sample Wilcoxon-Signed rank test

Null Hypothesis : $H_0 : \theta = \theta_0$

Procedure:

- A random sample of size n is taken from the above distribution as X1, X2, …. Xn.
- Then Di = Xi − θ_0 is calculated. Any of the Xi is equal to θ_0 , then delete such observation and reduce n accordingly.
- Rank the differences from smallest to largest without considering their signs. I.e., we rank |Di|. The smallest |Di| will get rank one and so on. If 2 or more |Di| are equal, then average rank is given to them. For e.g., if |Di| in the rank positions 2, 3, and 4 are tied , then we assign the rank as (2+3+4)/ 3 = 3.
- Now signs are given to these ranks as per the actual sign of Di.
- Let T^+ = Sum of the ranks having positive sign.
  T^-= Sum of the ranks having negative sign.
  T = Min (T^+, T^-)

Decision Criteria :

1) $H_a : \theta < \theta_0$,Reject H_0 if $T^+ \leq$ d.

2) $H_a : \theta > \theta_0$,Reject H_0 if $T^- \leq$ d.

3) $H_a : \theta \neq \theta_0$,Reject H_0 if T ≤ d.

Where d can be obtained from table for Wilcoxon Signed Rank Test for one sample

2.2 Independent Samples t-test

Assumptions:

- Population data is continuous.
- Both the Populations follow a standard normal distribution.
- Samples are independent of each other.
- The sample should be randomly selected from the population.

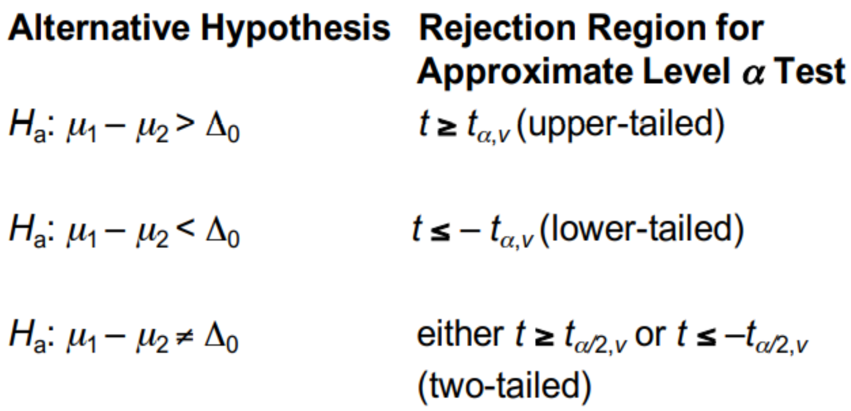
2.2 Independent Samples t-test

2.2 Independent Samples t test

Two-Sample t-test

Null Hypothesis : $H_0 : \mu_1 = \mu_2$

Test Statistic : $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$



Alternative Hypothesis	Rejection Region for Approximate Level α Test
H_a: $\mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,v}$ (upper-tailed)
H_a: $\mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,v}$ (lower-tailed)
H_a: $\mu_1 - \mu_2 \neq \Delta_0$	either $t \geq t_{\alpha/2,v}$ or $t \leq -t_{\alpha/2,v}$ (two-tailed)

Man Whitney U Test

Mann Whitney U Test

When assumptions of independent samples t-test are not satisfied, we use Mann Whitney U test.
This test gives the procedure for testing null hypothesis H0 of equal population location parameters.

Assumptions:

1. The data consist of random sample of observations X1, X2, … Xn1 from population 1 and another random sample of observations Y1, Y2 … Yn2 from second population.
2. The two samples are independent.
3. The variable under consideration is continuous.
4. The measurement scale is at least ordinal.

Man Whitney U Test

Null Hypothesis :

H_0: The two populations are identical. i.e., F(x) = G(y)

Procedure:

1) Combine both the samples and rank all observations from smallest to largest. If two or more observations are equal, then average rank is given to them.
2) Obtain sum of the ranks of observations from population 1 i.e., of Xi’s

Test Statistic:

$$U = S - \frac{n_1(n_1+1)}{2}$$

where S is the sum of ranks assigned to sample observations from population 1

n_1 is number of observations in sample 1.

Decision Criteria :

$H_a : \theta < \theta_0$, Reject H_0 if U < w_α.
$H_a : \theta > \theta_0$, Reject H_0 if U ≤ $w_{1-\alpha}$ where $w_{1-\alpha} = n_1 n_2 - w_\alpha$.
$H_a : \theta \neq \theta_0$, Reject H_0 if U < $W_{\frac{\alpha}{2}}$ or U > $w_{1-\frac{\alpha}{2}}$.
Where W is the critical value which can be obtained here.

2.3 Paired t-test

The paired t-test is useful for analyzing the same set of items that **were measured under two different conditions**, differences in measurements made on the **same subject before and after a treatment**, or differences between two treatments given to the same subject.

Assumptions:

•Subjects are independent.

•Each of the paired measurements are obtained from the same subject.

•The distribution of differences is normally distributed.

•Differences does not contain any significant outlier observation.

2.3 Paired t-test

2.3 Paired t-test

Null Hypothesis： $H_0 : \mu_d = 0$

Test Statistic： $t = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$

Alternative Hypothesis	Rejection Region for a Level α Test
H_a: $\mu > \mu_0$	$t \geq t_{\alpha,n-1}$ (upper-tailed)
H_a: $\mu < \mu_0$	$t \leq -t_{\alpha,n-1}$ (lower-tailed)
H_a: $\mu \neq \mu_0$	either $t \geq t_{\alpha/2,n-1}$ or $t \leq -t_{\alpha/2,n-1}$ (two-tailed)

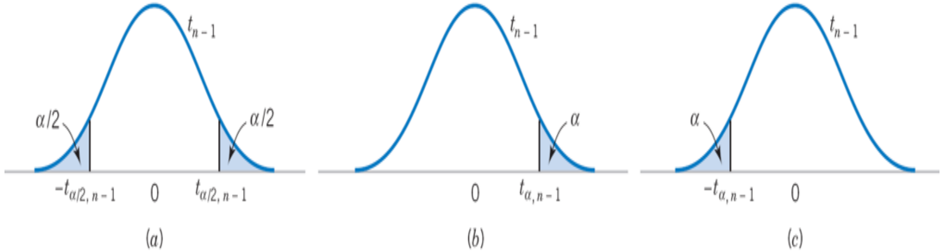


Figure 4-19 The distribution of T_0 when H_0: $\mu = \mu_0$ is true, with critical region for (a) H_1: $\mu \neq \mu_0$, (b) H_1: $\mu > \mu_0$, and (c) H_1: $\mu < \mu_0$.

Tests For Proportions

3.1 One Sample Proportion Test

3.1 One Sample proportion test

Central Limit Theorem

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution.

Assumptions

- Population follows a binomial distribution.
- n*p ≥ 5
- n*(1-p) ≥ 5

Where
n : Sample Size
p : Probability of Success

3.1 One Sample Proportion Test

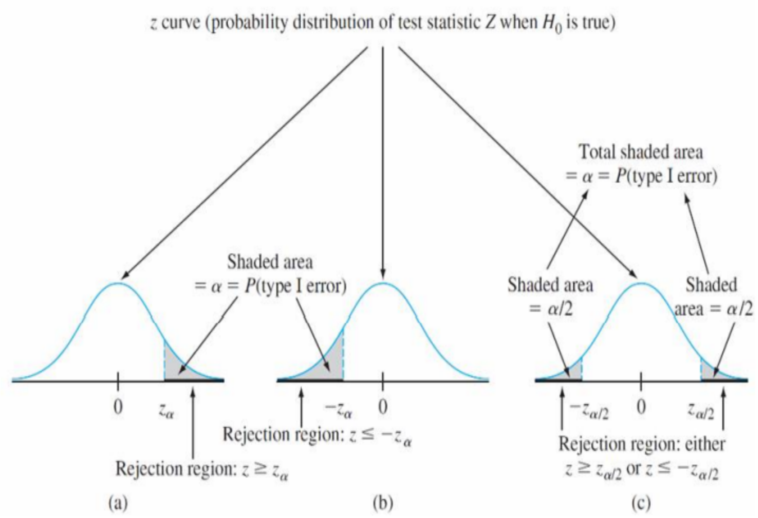
3.1 One Sample proportion test

One Sample Proportion test
Null Hypothesis : $H_0 : p = p_0$

Test Statistic : $Z = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ $\hat{p} = \dfrac{X}{n}$

X is number of events that resulted in success

Alternative Hypothesis	Rejection Region
H_a: $p > p_0$	$z \geq z_\alpha$ (upper-tailed)
H_a: $p < p_0$	$z \leq -z_\alpha$ (lower-tailed)
H_a: $p \neq p_0$	either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$ (two-tailed)



z curve (probability distribution of test statistic Z when H_0 is true)
Total shaded area $= \alpha = P$(type I error)
Shaded area $= \alpha = P$(type I error)
Shaded area $= \alpha/2$
Shaded area $= \alpha/2$
0
z_α
Rejection region: $z \geq z_\alpha$
(a)
$-z_\alpha$
0
Rejection region: $z \leq -z_\alpha$
(b)
$-z_{\alpha/2}$
0
$z_{\alpha/2}$
Rejection region: either
$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$
(c)

Rejection regions for z tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

3.2 Two Sample Proportion Test

3.2 Two Samples proportion test

Assumptions

Both populations follow a binomial distribution.

$n_1 * p_1 \geq 5$

$n_1 * (1 - p_1) \geq 5$

$n_2 * p_2 \geq 5$

$n_2 * (1 - p_2) \geq 5$

Where

p_1 is the probability of success in sample 1

p_2 is the probability of success in sample 2

n_1 is size of sample 1

n_2 is size of sample 2

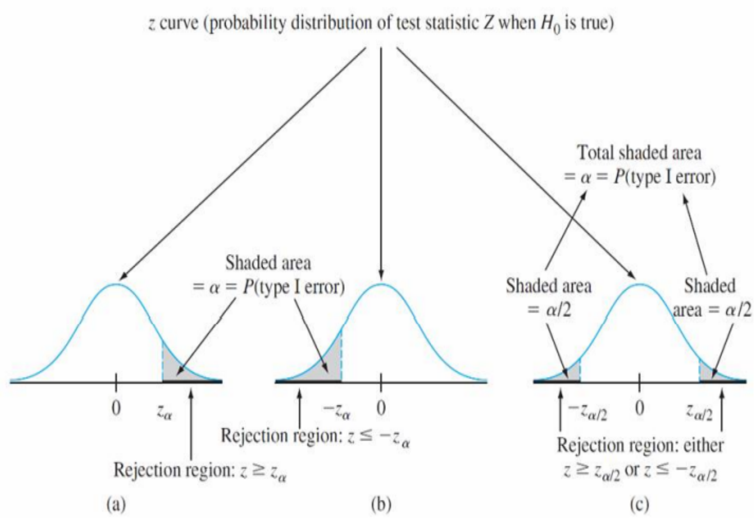
3.2 Two Sample Proportion Test

3.2 Two Samples proportion test

Null Hypothesis : $H_0 : p_1 = p_2$

Test Statistic : $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}}$

Alternative Hypothesis	Rejection Region for Approximate Level α Test
H_a: $p_1 - p_2 > 0$	$z \geq z_a$
H_a: $p_1 - p_2 < 0$	$z \leq -z_a$
H_a: $p_1 - p_2 \neq 0$	either $z \geq z_{a/2}$ or $z \leq -z_{a/2}$


z curve (probability distribution of test statistic Z when H_0 is true)
Shaded area $= \alpha = P$(type I error)
Total shaded area $= \alpha = P$(type I error)
Shaded area $= \alpha/2$
Shaded area $= \alpha/2$
0
z_α
$-z_\alpha$
0
$-z_{\alpha/2}$
0
$z_{\alpha/2}$
Rejection region: $z \geq z_\alpha$
Rejection region: $z \leq -z_\alpha$
Rejection region: either $z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$
(a)
(b)
(c)

Rejection regions for z tests: (a) upper-tailed test; (b) lower-tailed test; (c) two-tailed test

Chi Square Test of Independence

4. Chi Square test of independence

4. Chi Square test of independence

- The Chi-square test of independence checks whether two categorical variables are likely to be related or not.
- We have counts for two categorical or nominal variables.
- We also have an idea that the two variables are not related.
- The test gives us a way to decide if our idea is plausible or not.

Assumptions

•The sampling method is simple random sampling.

•The variables under study are each categorical.

•If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

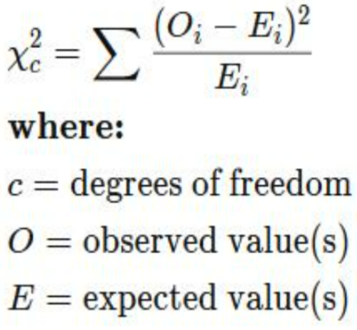
4. Chi Square test of independence

4. Chi Square test of independence

Null Hypothesis : In the population, the two categorical variables are independent.

Alternative hypothesis : In the population, the two categorical variables are dependent.

Test Statistic



$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where:

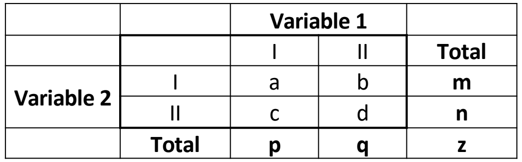
$c =$ degrees of freedom

$O =$ observed value(s)

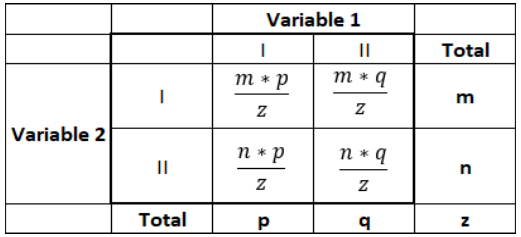
$E =$ expected value(s)

Where
c = (r-1)*(k-1)
r : Number of rows
k : Number of columns

Observed Frequency



		Variable 1		
		I	II	**Total**
Variable 2	I	a	b	**m**
	II	c	d	**n**
	Total	**p**	**q**	**z**



		Variable 1		
		I	II	**Total**
Variable 2	I	$\frac{m*p}{z}$	$\frac{m*q}{z}$	**m**
	II	$\frac{n*p}{z}$	$\frac{n*q}{z}$	**n**
	Total	**p**	**q**	**z**

4. Chi Square test of independence

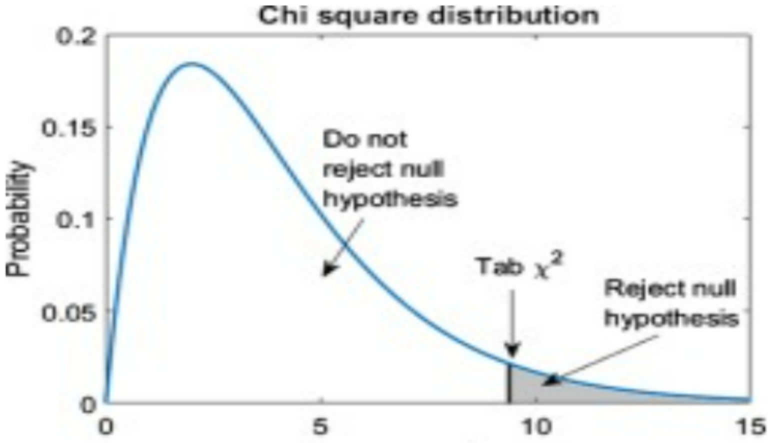
4. Chi Square test of independence

Rejection Criteria

Critical Value : $\chi^2_{c,\alpha}$

Conclusion:

Reject Null Hypothesis if test statistic is greater than the critical value else do not reject.



ANOVA

5. ANOVA

5. Analysis of Variance (ANOVA)

- The analysis of variance or ANOVA is a statistical inference test that lets you compare multiple groups at the same time and conclude if there is difference between means of groups.
- For example, if we wanted to test whether voter age differs based on some categorical variable like race, we have to compare the means of each level or group the variable.

Why not t-test?

- The t-test works well when dealing with two groups, but sometimes we want to compare more than two groups at the same time.
- We could carry out a separate t-test for each pair of groups, but when you conduct many tests you increase the chances of false positives (Type-1 Error)

Assumptions

1.The responses for each factor level have a **normal population distribution**.
2.These distributions have the **same variance**.
3.The data are **independent**.

5. ANOVA

5. One Way ANOVA

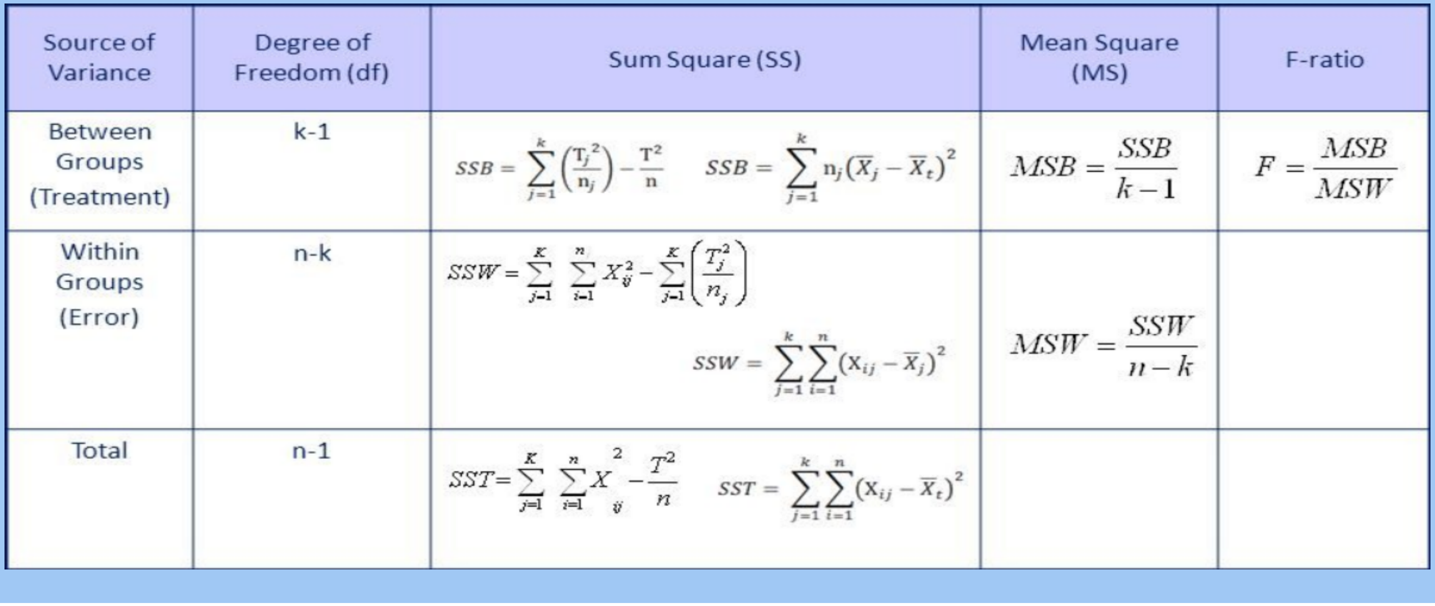
Null Hypothesis : There is no significant difference among the means of k treatments.

$H_0 : \mu_1 = \mu_2 = ... = \mu_k = \mu$

Alternative Hypothesis : At least one of the treatment mean differs significantly from the other treatment means.

$H_a : \mu_i \neq \mu$ For at least one i (i = 1,2,3,,,k)

ANOVA Table



Source of Variance	Degree of Freedom (df)	Sum Square (SS)	Mean Square (MS)	F-ratio
Between Groups (Treatment)	k-1	$SSB = \sum_{j=1}^{k}\left(\frac{T_j^2}{n_j}\right) - \frac{T^2}{n}$ $SSB = \sum_{j=1}^{k} n_j(\overline{X}_j - \overline{X}_t)^2$	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups (Error)	n-k	$SSW = \sum_{j-1}^{K} \sum_{i-1}^{n} X_{ij}^2 - \sum_{j-1}^{K}\left(\frac{T_j^2}{n_j}\right)$ $SSW = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$	$MSW = \frac{SSW}{n-k}$	
Total	n-1	$SST = \sum_{j=1}^{K} \sum_{i=1}^{n} X_{ij}^{2} - \frac{T^2}{n}$ $SST = \sum_{j=1}^{k}\sum_{i=1}^{n}(X_{ij} - \overline{X}_t)^2$		

5. ANOVA

5. One Way ANOVA

Rejection Region:

If Test Statistic F is Greater than Critical Value ($F_{\alpha,k-1,n-k}$) then we reject the Null hypothesis and conclude that at least one of the treatment means is significantly different than others.

Post –hoc test

- ANOVA only helps us to come at the conclusion whether all the treatment group means are equal or not, but it does not give information about which treatment group is statistically different if we reject the null hypothesis.

- To get information about which treatment group mean is statistically different, we can use post-hoc tests like Tukey’s test, Bonferroni Procedure.

5. ANOVA

Kruskal Wallis Test

The most widely used non-parametric technique for testing null hypothesis that several samples have been drawn from the same or identical populations is the Kruskal Wallis (one way ANOVA by ranks) test. When only two samples are considered, this test is equivalent to Mann Whitney test.

This test uses more information than the median test and hence it is more powerful and preferred when the available data are measured on at least ordinal scale.

Requirements:

1. The data consists of K random samples of sizes n1, n2, …. nK for analysis.
2. The variable of interest is continuous.
3. The observations are independent both within and between (or among) the samples.
4. The measurement scale is at least ordinal.

5. ANOVA

Hypothesis

H0: The K population distribution functions are identical. (or All the K populations have same median.)

H1: The K population distribution functions are not identical. (or All the K populations do not have same median.)

Procedure:

1) Let N = n1 + n2 +….+ nK be the total number of observations in the k samples and we rank all the observations from the smallest to the largest. If 2 or more observations are equal, then average rank is given to them.
2) If H0 is true, we expect the distribution of ranks over the groups to be a matter of chance, so that either small or large ranks do not tend to be concentrated in one sample.

Test Statistic : $H = \frac{12}{N\ (N+1)} \sum \frac{R_i^2}{n_i} - 3\ (N+1)$

Decision Criteria :

The test statistic H Follows chi squared distribution with k-1 degrees of freedom (k: number of groups)

Reject H0 if H > $\chi^2_{\alpha,k-1}$ else do not reject H0.

Thank you!