**Programming with R**

**Module 4: Statistics using R**

**Introduction**

Before we attempt to describe data, we need to make sure the data is in the right format. This means

* Making sure all the data is contained in a data frame (or in a vector if it’s a single variable)
* Ensuring that all the variables are of the correct type
* Checking that the values are all processed correctly

library(dplyr)

library(ggplot2)

path <- "D:/CMS/R\_2020/Placement\_Data\_Full\_Class.csv" # copy applicable path

placementds <- read.csv(path)

placementds$sl\_no <- NULL # remove sl.no column

colnames(placementds)

# split the dataset

placementnum <- select(placementds, ends\_with("\_p"), salary)

placementcat <- select(placementds, -(ends\_with("\_p")), -salary)

placedset <- filter(placementds, status == "Placed")

placedset <- na.omit(placementds) # alternative

notplacedset <- subset(placementds, status == "Not Placed")

# basic exploration of the dataset

class(placementds)

anyNA(placementds) # check for NAs

colSums(is.na(placementds))

lapply(placementds, anyNA)

str(placementds)

unique(placementds$degree\_t)

**Descriptive Analytics**

Descriptive analytics is about finding “what has happened” by summarizing the data using innovative methods and analysing the past data using simple queries. Analysing past data can provide insights that can assist organisations to take appropriate decisions. Primary objective of descriptive analytics is simple comprehension of data using data summarisation, basic statistical measures and visualisation.

Statistics is basically the study of what causes variability in the data. Descriptive statistics such as measures of central tendency, measures of variation and measures of shape provide useful insights.

**Measures of central tendency**

Measures of central tendency are the measures that are used for describing the data using a single value. Mean, Median and Mode are the three commonly used measures to compare different data sets. Percentile, Decile and Quantile are frequently used to identify the position of the observation in the data set.

**Measures of variation**

While the measures of central tendency yield information about the center or middle part of a data, measures of variability describe the spread or dispersion of a set of data. Using the measures of variation in conjunction with measures of central tendency makes possible a more complete numerical description of data. Range, Inter Quartile Range, Mean Absolute Deviation, Variance and Standard Deviation are common measures of variability.

**Measures of shape – Skewness and Kurtosis**

Skewness is a measure of symmetry or lack of symmetry. A data set is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. A value of zero for the measure indicates the data is symmetrical.

Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis value of 3 indicates standard normal distribution. The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is obtained by subtracting the value 3 from the kurtosis value.

Note: You can refer to any standard book on Statistics for a detailed explanation to all the above measures

**#** summary statistics

# measures of central tendency

mean(placementnum$degree\_p) # Describing center of the data

placementds %>% group\_by(degree\_t) %>% summarise(mean(degree\_p))

group\_by(placementds, degree\_t, status) %>% summarise(mean(degree\_p))

aggregate(degree\_p ~ status + degree\_t, placementds, FUN = mean) # alternative

sapply(placementnum, mean) # add na.rm if necessary

apply(placementnum, 2, mean) # other apply family fucntion

lapply(placementnum, mean)

tapply(placementds$degree\_p,placementds$gender, mean)

median(placementnum$salary, na.rm = T) # Dealing with missing values

quantile(placementnum$salary, na.rm = T)

quantile(placementnum$salary, 0.25, na.rm = T) # Customized quantiles

fivenum(placementnum$salary, na.rm = T) # same as quantile for odd no. of observations

# The core packages in R don't have a function for calculating the mode

lsr::modeOf(placementds$salary)

# other means

colMeans(placementnum)

rowMeans(placementnum[ , -6])

# skewness and kurtosis

psych::skew(placementnum) # value close to 0 indicates data is symmetrical

ggplot(placementnum, aes(salary)) + geom\_histogram(binwidth = 25000, na.rm = T)

ggplot(placementnum, aes(mba\_p)) + geom\_histogram(binwidth = 5)

psych::kurtosi(placementnum) # negative-flat, positive-pointy, zero-just enough pointy

ggplot(placementnum, aes(ssc\_p)) + geom\_histogram(bins = 10)

ggplot(placementnum, aes(hsc\_p)) + geom\_histogram(bins = 10)

# mean v/s median

mean(placementnum$salary, na.rm = T)

median(placementnum$salary, na.rm = T)

mean(placementnum$salary, trim = 0.10, na.rm = T) #Trimmed mean

# measures of variation

min(placementnum$mba\_p)

max(placementnum$mba\_p)

range(placementnum$mba\_p)

IQR(placementnum$salary, na.rm = T)

sd(placementnum$degree\_p)

var(placementnum$degree\_p)

mad(placementnum$degree\_p)

# Summarizing a variable

summary(placementnum$mba\_p)

# Summarizing a complete dataset

summary(placementnum)

summary(placementds)

psych::describe(placementnum)

# Describing Categories

table(placementcat$status)

sapply(placementcat, table)

table(placementcat$specialisation, placementcat$status) # Creating a two-way table

mytab <- with(placementcat, table(specialisation, status)) # alternative

mytab

addmargins(mytab)

prop.table(mytab) # proportion based on total number

prop.table(mytab, margin = 1) # proportions over rows and columns

# visualisation

df\_mytab <- as.data.frame(mytab)

df\_mytab

ggplot(df\_mytab, aes(x=specialisation, y = Freq)) +

geom\_bar(aes(fill = status), stat = "identity")

# Note - when heights of the bars have to represent values (freq) in the data,

# use stat="identity" and map a value to the y aesthetic

tab3 <- xtabs(~gender+specialisation+status, placementds) # 3-way crosstabs

tab3

ftable(tab3)

**Measure of association**

Correlation is a measure of the strength and direction of relationship that exists between two random variables and is measured using correlation coefficient. Correlation is only an association relationship and not a causal relationship.

# Tracking correlations

cor(placementnum$degree\_p, placementnum$mba\_p)

ggplot(placementnum, aes(degree\_p, mba\_p)) + geom\_point()

corcomplete <- cor(placementnum) # correlations for multiple variables

corcomplete

corcomplete["ssc\_p", "mba\_p"]

placementnum %>% cor() # alternative

# Dealing with missing values

cor(placementnum$mba\_p, placementnum$salary, use="complete.obs")

cor(placementnum, use = "pairwise.complete.obs")

**Hypothesis testing**

Hypothesis is a claim made by a person/organisation. Such claims are usually about population parameters like mean and we seek evidence from a sample for the support of the claim.

Hypothesis testing is one of the most important concepts in analytics. Hypothesis testing consists of two complementary statements called null and alternative hypothesis, and only one of them is true. Hypothesis testing is a process used for either rejecting or retaining a null hypothesis.

**Null and alternative hypothesis:** A null hypothesis usually denoted as H0 refers to the statement that there is no relationship or no difference between different groups with respect to the value of a population parameter. At the beginning we assume that the null hypothesis is true and try to retain it unless there is strong evidence against it. Alternative hypothesis usually denoted as H1 is what the researcher believes to be true and would like to reject the null hypothesis.

**One-tailed and two-tailed test:** Statistical hypotheses are written such that they will produce either a one-tailed (left/right tailed) or a two-tailed test. In a one-tailed test the rejection region will be only on one side (left/right) of the distribution. One-tailed tests are directional. For a two-tailed test, the rejection region is on either side of the distribution. Two tailed tests always use the equality and the non-equality symbols in the hypotheses and are directionless.

**Decision Criteria:** After we have stated the null and alternative hypothesis, the next step is to decide what criterion to use for deciding whether to accept or reject the null hypothesis. Significance value, usually denoted by alpha (α), is the criteria used for taking this decision. The chosen value of alpha may depend on the context of the problem. The usual values of alpha are 0.01 (one percent), 0.05 and 0.1.

**Test statistic:** Test statistic is the standardised difference between the estimated value of the parameter being tested calculated from the sample and the hypothesis value in order to establish the evidence in support of the null hypothesis. The test statistic that we will mostly use in this module will be **Z** and **t**

The cases in which the test statistic is Z are

* Population standard deviation is known, and the population is normal
* Population standard deviation is known, and the sample size is at least 30. Note that the population need not be normal

The formula for calculating Z is –

is the sample mean

μ is the hypothesised value of population mean

σ is the population standard deviation and

n is the sample size

We use the t-statistic when the population is normal, and population standard deviation is unknown but the sample standard deviation ‘S’ is known. t-test is a robust test for violation of normality of the data as long as the data is close to symmetry and there are no outliers.

The formula for calculating t statistic is –

**Concluding Step – Decision of Hypothesis test:** In hypothesis test we end up with one of the below decision –

* Reject the Null Hypothesis
* Fail to Reject (Retain) the Null Hypothesis

We can reach a statistical conclusion using either p-value or critical value method. We will prefer the p-value approach.

p-value is the evidence in support of null hypothesis. The decision to reject or retain the null hypothesis will depend on whether the calculated p-value crosses the threshold value alpha or not. The decision criteria is as shown below –

p-value < α Reject the null hypothesis

p-value ≥ α Retain the null hypothesis

**Statistical Tests and Assumptions**

Many of statistical tests including correlation, regression, t-test, and analysis of variance (ANOVA) assume some certain characteristics about the data.

Generally, they assume that:

* the data are normally distributed
* and the variances of the groups to be compared are homogeneous (equal)

These tests are called parametric tests, because their validity depends on the distribution of the data. Before using a parametric test, we should perform some preliminary tests to make sure that the test assumptions are met. In the situations where the assumptions are violated, non-parametric tests are recommended.

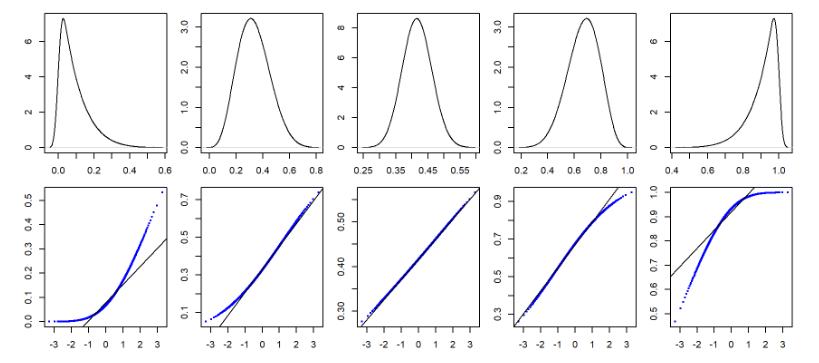
**How to assess the normality?**

If the sample size is large enough (n > 30), we can ignore the distribution of the data and use parametric tests. The central limit theorem tells us that no matter what distribution things have, the sampling distribution tends to be normal if the sample is large enough (n > 30).

Normality can be checked by visual inspection using normal plots (histogram), Q-Q plot (quantile-quantile plot) or by significance tests.

geom\_qq and stat\_qq produce quantile-quantile plots. geom\_qq\_line and stat\_qq\_line compute the slope and intercept of the line connecting the points at specified quartiles of the theoretical and sample distributions.

In the following figure, the top row shows the different density distribution plots; the bottom row shows the normal q-q plots for each distribution.



**#** Assess the normality - Visual methods

ggplot(placementds) + geom\_density(aes(degree\_p)) # density plot

ggplot(placementds) + geom\_histogram(aes(degree\_p), bins = 10) # histogram

qqbase <- ggplot(placementds, aes(sample=degree\_p))

qqbase + geom\_qq()

qqbase + geom\_qq() + geom\_qq\_line() # qq plot

qqbase + stat\_qq() + stat\_qq\_line() # alternative

qqbase1 <- ggplot(placementds, aes(sample=degree\_p, color = status))

qqbase1 + geom\_qq() + geom\_qq\_line()

Visual inspection is usually unreliable. It is possible to use a significance test comparing the sample distribution to a normal one in order to ascertain whether data show or not a serious deviation from normality.

There are several methods for normality test such as Kolmogorov-Smirnov (K-S) normality test and Shapiro-Wilk’s test.

Shapiro-Wilk’s method is widely recommended for normality test and it provides better power than K-S. It is based on the correlation between the data and the corresponding normal scores.

Note: The null hypothesis of these tests is that “sample distribution is normal”. The p-value ≥ 0.05 (0.05 is our choosen α value) implies that the distribution of the data are not significantly different from normal distribution. In other words, we can assume the normality.

**#** Shapiro-Wilk's test to test normality

shapiro.test(placementds$degree\_p)

# Note: p>= 0.05 which is our choosen alpha, we can assume normality

shapiro.test(placementds$mba\_p)

**How to assess the equality of variances?**

If the samples, being compared, follow normal distribution, then it’s possible to use:

* F-test to compare the variances of two samples
* Bartlett’s Test or Levene’s Test to compare the variances of multiple samples.

**Research Questions**

The most popular research questions include:

1. whether two variables (n = 2) are correlated (i.e., associated)
2. whether multiple variables (n > 2) are correlated
3. whether two groups (n = 2) of samples differ from each other
4. whether multiple groups (n >= 2) of samples differ from each other
5. whether the variability of two samples differ

The above questions can be answered using the following statistical tests:

1. Correlation test between two variables
2. Correlation matrix between multiple variables
3. Comparing the means of two groups
   * Student’s t-test (parametric)
   * Wilcoxon rank test (non-parametric)
4. Comparing the means of more than two groups
   * ANOVA test (analysis of variance, parametric): extension of t-test to compare more than two groups
   * Kruskal-Wallis rank sum test (non-parametric): extension of Wilcoxon rank test to compare more than two groups
5. Comparing the variances
   * Comparing the variances of two groups: F-test (parametric)
   * Comparison of the variances of more than two groups: Bartlett’s test (parametric), Levene’s test (parametric) and Fligner-Killeen test (non-parametric)
6. **Correlation Test Between Two Variables**

Correlation test is used to evaluate the association between two or more variables. For instance, if we are interested to know whether there is a relationship between degree percentage and MBA percentage, a correlation coefficient can be calculated to answer this question.

There are different methods to perform correlation analysis:

* + Pearson correlation (r), which measures a linear dependence between two variables (x and y). It is also known as a parametric correlation test because it depends to the distribution of the data. It can be used only when x and y are from normal distribution
* Kendall tau and Spearman rho, which are rank-based correlation coefficients (non-parametric)

Use the function cor.test(x, y) to analyze the correlation coefficient between two variables and to get significance level of the correlation.

Note: The null hypotheis is that there is no correlation between the two random varaiables.

**#** Correlation test between two variables

ggplot(placementnum, aes(degree\_p, mba\_p)) + geom\_point() + geom\_smooth()

cor(placementnum$degree\_p, placementnum$mba\_p)

# Pearson correlation test

cor.test(placementnum$degree\_p, placementnum$mba\_p)

# Note: p < 0.05, we can conclude the variables are significantly correlated

1. **Correlation Test Between Multiple Variables**

A correlation matrix is used to analyze the correlation between multiple variables at the same time. There are different methods for correlation analysis: Pearson parametric correlation test, Spearman and Kendall rank-based correlation analysis.

The function rcorr() [in Hmisc package] can be used to compute the significance levels for pearson and spearman correlations. It returns both the correlation coefficients and the p-value of the correlation for all possible pairs of columns in the data table.

**#** Correlation test between multiple variables

round(cor(placementnum, use = "pairwise.complete.obs"), 2)

GGally::ggpairs(placementnum)

library(Hmisc)

rcorr(as.matrix(placementnum))

1. **Comparing the means of two groups**
2. **Comparing one-sample mean to a standard known mean**

One-sample t-test (parametric): One-sample t-test is used to compare the mean of one sample (m) to a known standard (or theoretical/hypothetical) mean (μ).

Typical research questions are:

1. whether the mean (m) of the sample *is equal* to the theoretical mean (μ)?

(two-tailed test; H0: m = μ and H1 m ≠ μ)

1. whether the mean (m) of the sample *is less than* the theoretical mean (μ)?

(one-tailed test; H0: m ≤ μ and H1 m > μ)

1. whether the mean (m) of the sample *is greater than* the theoretical mean (μ)?

(one-tailed test; H0: m ≥ μ and H1 m < μ)

To perform one-sample t-test, the function t.test() can be used.

Note: We set the alternative hypotheis as the statement that we would like to prove.

**#** One-sample t-test

# Select a sample of size 100

set.seed(25)

mysample <- sample\_n(placementds, size = 100)

summary(mysample$degree\_p)

# We want to prove that the average degree\_p differs from 60% (two-tailed test)

#H0=60

t.test(mysample$degree\_p, mu = 60, alternative = "two.sided") # default is two sided

# Note: p < 0.05, We can conclude that the mean degree\_p is significantly different from 60%

# We want to prove that the average degree\_p is less than 66% (one-tailed test)

# H0 >=66

t.test(mysample$degree\_p, mu = 66, alternative = "less")

1. **Comparing the means of two independent groups**

The unpaired two-samples t-test is used to compare the mean of two independent groups.

For example, suppose that we have the degree percetange of 100 students: 50 Female (group A) and 50 Male (group B). We want to know if the mean degree percenatge of Female (mA) is significantly different from that of male (mB). In this case, we have two unrelated (i.e., independent or unpaired) groups of samples. Therefore, it is possible to use an independent t-test to evaluate whether the means are different.

Typical research questions are:

1. whether the mean of group A (mA) *is equal* to the mean of group B (mB)?

(two-tailed test; H0: mA = mB and H1 mA ≠ mB)

1. whether the mean of group A (mA) *is less than* the mean of group B (mB)?

(one-tailed test; H0: mA ≤ mB and H1 mA > mB)

1. whether the mean of group A (mA) *is greater than* the mean of group B (mB)?

(one-tailed test; H0: mA ≥ mB and H1 mA < mB)

To perform two-samples t-test comparing the means of two independent samples (x & y), the function t.test() can be used.

Note:  var.equal: a logical variable indicating whether to treat the variances of the two groups as being equal. If TRUE then the pooled variance (classical t-test) is used to estimate the variance otherwise the Welch-t test is used. The Welch t-test is considered as the safer one.

**#** Two-sample t-test

# create two samples based on gender

set.seed(25)

malesample <- placementds %>% filter(gender == "M") %>% sample\_n(size = 50)

femalesample <- placementds %>% filter(gender == "F") %>% sample\_n(size = 50)

# Is there any significant difference between male and female percentage?

t.test(malesample$degree\_p, femalesample$degree\_p, var.equal = T)

# Note: p < 0.05, We can conclude that the mean degree\_p is significantly different between the two groups

# We want to prove that the average percentage of male is less than that of female

#H0: mm ≥ mf

t.test(malesample$degree\_p, femalesample$degree\_p, var.equal = T, alternative = "less")

# Method 2 - If the data are saved in single dataframe

t.test(degree\_p~gender, data = mysample, var.equal = T)

# Welch t-statistic (unequal varaince)

t.test(degree\_p~gender, data = mysample, var.equal = F)

1. **Comparing the means of paired samples**

The paired samples t-test is used to compare the means between two related groups of samples. In this case, we have two values (i.e., pair of values) for the same samples.

In many cases, we would like to analyse whether an intervention (or treatment) such as training programs, marketing promotion, treatment for specific illness etc. may have significantly changed the population parameter values such as a mean before and after the intervention. The objective here is to check whether the difference in the parameter values is statistically significant before and after the intervention or between two different types of interventions. Thus, in a paired t-test, the data related to the parameter is captured twice from the same subject, once before and once after the intervention.

Typical research questions are:

1. whether the mean difference (m) *is equal* to 0?
2. whether the mean difference (m) *is less than* 0?
3. whether the mean difference (m) *is greater than* 0?

To perform paired samples t-test comparing the means of two paired samples (x & y), the function t.test() can be used as follows-

t.test(x, y, paired = TRUE, alternative = "two.sided")

Example: Check whether the weekly alcohol consumption is more after break up

(Refer data\_paired\_t\_test.csv file)

The null hypothesis here will be that the mean difference is ≤ 0.

**#** Paired sample t-test

pairdata <- read.csv("D:/Material\_CMS/R\_2020/data\_paired\_t\_test.csv")

colnames(pairdata)

t.test(pairdata$After\_Breakup, pairdata$Before\_Breakup, paired = T, alternative = "greater")

# Note: we retain the null and conclude the diff. in alcohol consumption is not greater than 0 before and after breakup

1. **Comparing the means of more than two groups**

ANOVA: In many situations, we may have to conduct a hypothesis test to compare mean values simultaneously for more than two groups (samples). In such cases we use the Analysis of Variance (ANOVA) to understand the differences in population means.

The objective of ANOVA is to check simultaneously whether population mean from more than two populations are different.

**One-way ANOVA**

The one-way analysis of variance (ANOVA), also known as *one-factor ANOVA*, is an extension of independent two-samples t-test for comparing means in a situation where there are more than two categories (levels). In one-way ANOVA, the data is organized into several groups based on one single grouping variable (also called *factor* variable).

In other words, we would like to study the impact of a single treatment (factor) at different levels (thus forming groups) on a continuous response variable (or outcome variable).

Example: The variable degree\_t is the treatment (factor) and CM, ST and Ot are the different levels (3 in this case) which are likely to have varying impact on the mba\_p (outcome/response variable)

ANOVA model is robust and minor violations of the assumptions of normality and equal variance may not result in incorrect decision.

The function aov() can be used. The function summary.aov() is used to summarize the analysis of variance model.

ANOVA test hypotheses:

* Null hypothesis: the means of the different groups are the same
* Alternative hypothesis: At least one sample mean is not equal to the others

**#** one-way ANOVa

levels(placementds$degree\_t)

# compute the summary statistics for understanding

placementds %>% group\_by(degree\_t) %>% summarise(count = n(),

mean(mba\_p), sd(mba\_p))

# visulaise for a better understanding

ggplot(placementds, aes(y=mba\_p)) +geom\_boxplot(aes(fill = degree\_t))

# we want to check whether type of degree had any siginifcant impact on the average MBA %

myaov1 <- aov(mba\_p ~ degree\_t, data = placementds)

summary.aov(myaov1)

# Note: p >= 0.05, We can conclude that the mean MBA % under different degree type are same

# Diagnostic Checking

# a. Homogeneity of variance

plot(myaov1, 1)

bartlett.test(mba\_p ~ degree\_t, data = placementds)

# Null hypothesis is there is homogenity of variances across groups

# b. Check the normality assumption

plot(myaov1, 2)

myaov1residuals <- residuals(myaov1)

shapiro.test(myaov1residuals)

# Note: normality assumption is violated

**Two-way ANOVA**

Two-way ANOVA test is used to evaluate simultaneously the effect of two grouping variables (factors) on a response variable. The number of levels can vary between factors. The level combinations of factors are called cell.

Example: In addition to the type of degree, the specialisation opted by the student in MBA may also play an important role in the MBA %age. Here we would like to understand the impact of both factors (degree\_t and specialisation) simultaneously on MBA %age.

**Nonparametric Statistics**

Parametric statistics are statistical techniques based on assumptions about the population from which the sample data are selected. For example- If a t-statistic is used to conduct a hypothesis test about a population mean, the assumption is that the data being analysed are randomly selected from a normally distributed population. The name parametric statistics refers to the fact that an assumption (here, normally distributed data) is being made about the data used to test or estimate the parameter (in this case, the population mean). In addition, the parametric statistics requires quantitative measurements that yield interval or ratio level data.

For data that do not meet the assumptions about the population, or when the level of data being measured is qualitative, statistical techniques called nonparametric or distribution-free techniques are used. A variety of nonparametric statistics are available for use with nominal or ordinal data. A few important techniques are Mann-Whitney U test, Wilcoxon matched-pairs signed ranks test, Kruskal-Wallis test, Friedman test, chi-square test of goodness-of-fit, and chi-square test of independence.

**Chi-square goodness-of-fit test**

The chi-square goodness-of-fit test compares the expected, or theoretical, frequencies of categories from a population distribution to the observed, or actual, frequencies from a distribution to determine whether there is a difference between what was expected and what was observed.

For example- The Director of a B-school theorizes that 60% of the students joining the MBA program come from Commerce & Management degree, 35% from Science and Technology background, and 5% with other degrees. To validate or reject this expected distribution, an actual sample of students is gathered randomly, and the observed results are compared to the expected results with chi-square goodness-of-fit test.

The null and alternative hypotheses in chi-square goodness-of-fit test are

Null: There is no statistically significant difference between the observed frequencies and the expected frequencies (The observed distribution is the same as the expected distribution)

Alternative: There is a statistically significant difference between the observed frequencies and the expected frequencies (The observed distribution is not the same as the expected distribution)

The R function chisq.test() is used as below-

chisq.test(x, p)

where, x: a numeric vector showing the observed frequencies as absolute values (i.e. counts)

p: a vector of probabilities of the same length of x showing the expected frequencies as probabilities or proportions

Note: The chi-square test should be used only when all calculated expected values are greater than 5

**#** chi-square goodness of fit test

deg\_count <- table(placementcat$degree\_t)

deg\_count

chi\_degree <- chisq.test(deg\_count, p = c(0.6, 0.05, 0.35))

chi\_degree

# Note: p >= 0.05, We retain the null

chi\_degree$expected #(should be greater than 5)

**Chi-square test of independence**

Chi-square test of independence is a hypothesis test that can be used to analyse the frequencies of two variables with multiple categories to determine whether the two variables are independent. In a sense, the Chi-square test of independence is a test whether the variables are related.

For example- The Director of a B-school is interested in checking whether or not placement depends on the type of degree or on choice of specialisation

The null and alternative hypotheses in Chi-square test of independence are

Null: The two variables are independent (not related)

Alternative: The variables are dependent (related)

**#** Chi-square test of independence

table(placementcat$degree\_t, placementcat$status)

chi\_degpla <- chisq.test(placementcat$degree\_t, placementcat$status)

chi\_degpla

# Note: p >= 0.05, We retain the null

chi\_degpla$expected # reason for the approximation may be incorrect warning

# different variable

chisq.test(placementcat$specialisation, placementcat$status)

**One-sample Wilcoxon signed rank test**

The one-sample Wilcoxon signed rank test is a non-parametric alternative to one-sample t-test when the data cannot be assumed to be normally distributed. It is used to determine whether the median of the sample is equal to a known standard value (i.e. theoretical value).

**#** One-sample Wilcoxon signed rank test

# first, check for normality

# Shapiro-Wilk's test to test normality

shapiro.test(placementds$salary)

# Note: p<0.05, reject null and hence we cannot assume normality

# We want to prove that the median salary differs from Rs.250000 (two-tailed test)

#H0=250000

wilcox.test(placementds$salary, mu = 250000, alternative = "two.sided")

# Note: p < 0.05, We can conclude that the median salary is significantly different from 250000

# We want to prove that the median salary is less than Rs.275000 (one-tailed test)

# H0 >=275000

wilcox.test(placementds$salary, mu = 275000, alternative = "less")

**Mann Whitney U Test (Wilcoxon Rank Sum Test)**

The Mann Whitney U test, sometimes called the Mann Whitney Wilcoxon Test or the Wilcoxon Rank Sum Test is a non-parametric alternative to the unpaired two-samples t-test, which can be used to compare two independent groups of samples. It is used when data are not normally distributed.

Note: Specify Paired = T so as to get alternative to paired t-test

**#** Mann Whitney U Test

# explore the median

group\_by(placementds, gender) %>% summarise(median(salary, na.rm = T))

# Is there any significant difference between male and female salary?

wilcox.test(salary~gender, data = placementds)

# Note: p < 0.05, We can conclude that the median salary is significantly different between the genders