

# Hypothesis

Unit IV

# T-test

- A t-test is a statistical test that compares the means of two samples.
- It is used in hypothesis testing, with a null hypothesis that the difference in group means is zero and an alternate hypothesis that the difference in group means is different from zero.

# T-test

- How to perform T-tests in R
- In the T-test, for specifying equal variances and a pooled variance estimate, we set `var.equal=True`. We can also use `alternative="less"` or `alternative="greater"` for specifying one-tailed test.

## Different Types

- one-sample,
- paired sample, and
- independent samples T-test

# T-test

- Take a sample from both sets and establish the problem assuming a null hypothesis that the two means are the same.

## Classification of T-tests

- One Sample T-test
- Two sample T-test
- Paired sample T-test

# One-Sample T-test

- **One-Sample T-test** is a T-test which compares the mean of a vector against a theoretical mean. There is a following formula which is used to compute the T-test :

$$t = \frac{m - \mu}{\frac{s}{\sqrt{n}}}$$

# One-Sample T-test

- Here,
- $M$  is the mean.
- $\mu$  is the theoretical mean.
- $s$  is the standard deviation.
- $n$  is the number of observations.
- For evaluating the statistical significance of the t-test, we need to compute the p-value. The p-value range starts from 0 to 1, and is interpreted as follow

# One-Sample T-test

- If the p-value is lower than 0.05, it means we are strongly confident to reject the null hypothesis.
- If the p-value is higher than 0.05, then it indicates that we don't have enough evidence to reject the null hypothesis.
- We construct the pvalue by looking at the corresponding absolute value of the t-test

# One Sample T – Test Approach

- The One-Sample T-Test is used to test the statistical difference between a sample mean and a known or assumed/hypothesized value of the mean in the population.
- So, for performing a one-sample t-test in R, we would use the syntax `t.test(y, mu = 0)`
- where `x` is the name of the variable of interest and
- `mu` is set equal to the mean specified by the null hypothesis.



# One Sample T – Test Approach

One Sample t-test

data: sweetSold

t = -15.249, df = 49, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 150

95 percent confidence interval:

138.8176 141.4217

sample estimates:

mean of x

140.1197

- $t = -15.249$ ,  $df = 49$ , and a  $2.2e-16$  p-value: provides the p-value, degrees of freedom (df), and test statistic (t). The computed t-value in this instance is -15.249, there are 49 degrees of freedom, and the p-value is very small ( $2.2e-16$ ), indicating strong evidence that the null hypothesis is false.
- The true mean is not equal to 150, as an alternative explains the alternative theory, which contends that the population's actual mean is not 150.
- The confidence interval, which ranges from 138.8176 to 141.4217, shows that there is a 95% chance that the genuine population mean is located between those two numbers.
- provides the sample estimate, in this example the sample mean ( $\bar{x}$ ) of 140.1197, or "sample estimates: mean of  $\bar{x}$  140.1197."

# Two sample T-Test Approach

- It is used to help us to understand whether the difference between the two means is real or simply by chance.
- The general form of the test is `t.test(y1, y2, paired=FALSE)`. By default, R assumes that the variances of y1 and y2 are unequal, thus defaulting to Welch's test. To toggle this, we use the flag `var.equal=TRUE`.

# Two sample T-Test Approach

- `> shopOne <- rnorm(50, mean = 140, sd = 4.5)`
- `> shopTwo <- rnorm(50, mean = 150, sd = 4)`
- `> t.test(shopOne, shopTwo, var.equal = TRUE)`

## Two Sample t-test

data: shopOne and shopTwo

t = -13.158, df = 98, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-11.482807 -8.473061

sample estimates:

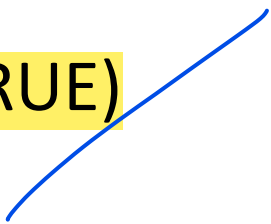
mean of x mean of y

140.1077 150.0856

# Two sample T-Test Approach

- Sample estimates: 140.1077 for the mean of x and 150.0856 for the mean of y the sample means (x and y), which are the sample estimates. In this instance, shopOne's mean is 140.1077, whereas shopTwo's mean is 150.0856

# Paired Sample T-test

- This is a statistical procedure that is used to determine whether the mean difference between two sets of observations is zero.
  - In a paired sample t-test, each subject is measured two times, resulting in pairs of observations.
  - The test is run using the syntax `t.test(y1, y2, paired=TRUE)`
- 

# Paired Sample T-test

- `> set.seed(2820)`
- `> sweetOne <- c(rnorm(100, mean = 14, sd = 0.3))`
- `> sweetTwo <- c(rnorm(100, mean = 13, sd = 0.2))`
- `> t.test(sweetOne, sweetTwo, paired = TRUE)`

Paired t-test

data: sweetOne and sweetTwo

t = 29.31, df = 99, p-value < 2.2e-16

alternative hypothesis: true mean difference is not equal to 0

95 percent confidence interval:

0.9892738 1.1329434

sample estimates:

mean difference

1.061109

# Correlation

- Correlation is a statistical measure that indicates how strongly two variables are related.
- It involves the relationship between multiple variables as well.
- For instance, if one is interested to know whether there is a relationship between the heights of fathers and sons, a correlation coefficient can be calculated to answer this question.
- Generally, it lies between -1 and +1.  $\delta$
- It is a scaled version of covariance and provides the direction and strength of a relationship. Correlation coefficient test in R



# Correlation

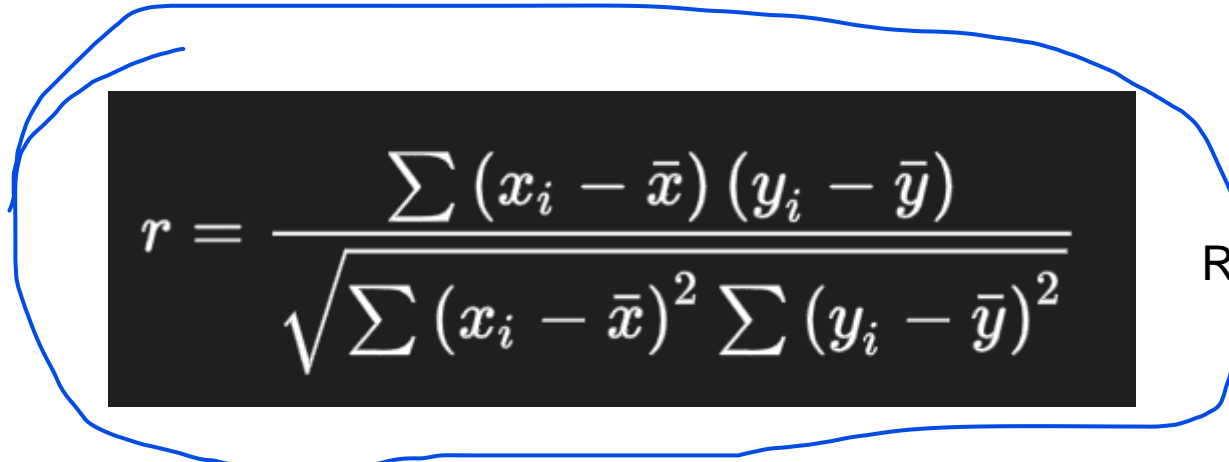
- There are mainly two types of correlation:

**Parametric Correlation** – **Pearson correlation( $r$ )**: It measures a linear dependence between two variables ( $x$  and  $y$ ) is known as a parametric correlation test because it depends on the distribution of the data.

**Non-Parametric Correlation** – **Kendall( $\tau$ )** and **Spearman( $\rho$ )**: They are rank-based correlation coefficients, and are known as non-parametric correlation

# Pearson Rank Correlation Coefficient Formula

- Pearson Rank Correlation is a parametric correlation.
- The Pearson correlation coefficient is probably the most widely used measure for linear relationships between two normal distributed variables and thus often just called “correlation coefficient”.
- The formula for calculating the Pearson Rank Correlation is as follows:


$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Sum  $X_i * Y_i$

Root ( Sum  $X_i$  sq - Sum  $Y_i$  sq )

# Pearson Rank Correlation Coefficient Formula

where,

- $r$ : pearson correlation coefficient
- $x$  and  $y$ : two vectors of length  $n$
- $m_x$  and  $m_y$ : corresponds to the means of  $x$  and  $y$ , respectively

Note:

$r$  takes a value between  $-1$  (negative correlation) and  $1$  (positive correlation).

$r = 0$  means no correlation.

Can not be applied to ordinal variables.

The sample size should be moderate (20-30) for good estimation.

Outliers can lead to misleading values means not robust with outliers

# Pearson Rank Correlation Coefficient Formula

R Programming Language provides two methods to calculate the pearson correlation coefficient.

By using the functions `cor()` or `cor.test()` it can be calculated.

It can be noted that `cor()` computes the correlation coefficient whereas `cor.test()` computes the test for association or correlation between paired samples.

It returns both the correlation coefficient and the significance level(or p-value) of the correlation

# Pearson Rank Correlation Coefficient Formula

**Syntax:** `cor(x, y, method = "pearson")`

`cor.test(x, y, method = "pearson")`

**Parameters:**

- **x, y:** numeric vectors with the same length
- **method:** correlation method

# Pearson Rank Correlation Coefficient Formula

- > `x = c(1, 2, 3, 4, 5, 6, 7)`
- > `y = c(1, 3, 6, 2, 7, 4, 5)`
- > `result = cor(x, y, method = "pearson")`
- > `cat("Pearson correlation coefficient is:", result)`

## Output

- Pearson correlation coefficient is: 0.5357143

# Correlation Coefficient Test In R Using cor.test() method

- > `x = c(1, 2, 3, 4, 5, 6, 7)`
- > `y = c(1, 3, 6, 2, 7, 4, 5)`
- > `result = cor.test(x, y, method = "pearson")`
- > `print(result)`

```
Pearson's product-moment correlation
```

```
data: x and y
```

```
t = 1.4186, df = 5, p-value = 0.2152
```

```
alternative hypothesis: true correlation is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.3643187  0.9183058
```

```
sample estimates:
```

```
cor
```

```
0.5357143
```

# Correlation Coefficient Test In R Using `cor.test()` method

- In the output above:
- `T` is the value of the test statistic ( $T = 1.4186$ )
- `p-value` is the significance level of the test statistic ( $p\text{-value} = 0.2152$ ).
- `alternative hypothesis` is a character string describing the alternative hypothesis (`true` correlation is not equal to 0).
- `sample estimates` is the correlation coefficient. For Pearson correlation coefficient it's named as `cor` (`Cor.coef = 0.5357`)



# Chi Square Test

- A chi-square test is a statistical test used to compare observed results with expected results.
- The purpose of this test is to determine if a difference between observed data
- The chi-square formula is:



- $\chi^2 = \sum (O_i - E_i)^2 / E_i$ ,

- where  $O_i$  = observed value (actual value) and  $E_i$  = expected value.

# Chi Square Test

- The null hypothesis states that there is no relationship between the two variables,
- while the research hypothesis states that there is a relationship between the two variables.
- In a chi-square analysis, the p-value is the probability of obtaining a chi-square as large or larger than that in the current experiment and yet the data will still support the hypothesis.
- It is the probability of deviations from what was expected being due to mere chance.

# Chi Square Test

- The chi-square test of independence evaluates whether there is an association between the categories of the two variables.
- There are basically two types of random variables and they yield two types of data: numerical and categorical.
- In R Programming Language Chi-square statistics is used to investigate whether distributions of categorical variables differ from one another.
- The chi-square test is also useful while comparing the tallies or counts of categorical responses between two(or more) independent groups

# Chi Square Test

- In R Programming Language, the function used for performing a chi-square test is
- `chisq.test()`.

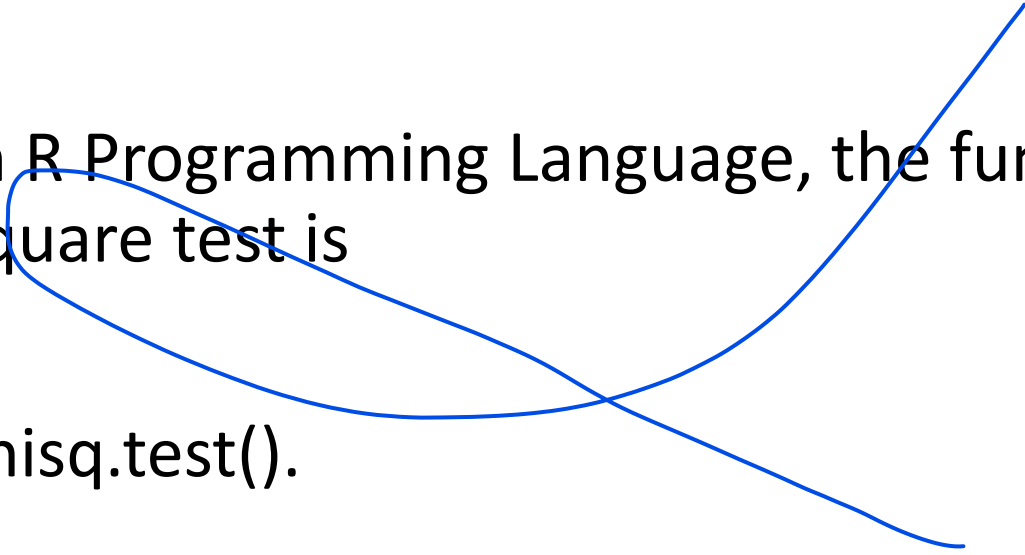
## *Syntax:*

*chisq.test(data)*

## *Parameters:*

*data: data is a table containing count values of the variables in the table.*

# Chi Square Test

- In R Programming Language, the function used for performing a chi-square test is
  - `chisq.test()`.
- 

# Chi Square Test

- We will take the survey data in the MASS library which represents the data from a survey conducted on students.

- library(MASS)
- print(str(survey))

```
'data.frame':    237 obs. of  12 variables:
 $ Sex      : Factor w/ 2 levels "Female","Male": 1 2 2 2 2 1 2 1 2 2 ...
 $ Wr.Hnd   : num  18.5 19.5 18 18.8 20 18 17.7 17 20 18.5 ...
 $ NW.Hnd   : num  18 20.5 13.3 18.9 20 17.7 17.7 17.3 19.5 18.5 ...
 $ W.Hnd    : Factor w/ 2 levels "Left","Right": 2 1 2 2 2 2 2 2 2 2 ...
 $ Fold     : Factor w/ 3 levels "L on R","Neither",...: 3 3 1 3 2 1 1 3 3 3 ...
 $ Pulse    : int  92 104 87 NA 35 64 83 74 72 90 ...
 $ Clap     : Factor w/ 3 levels "Left","Neither",...: 1 1 2 2 3 3 3 3 3 3 ...
 $ Exer     : Factor w/ 3 levels "Freq","None",...: 3 2 2 2 3 3 1 1 3 3 ...
 $ Smoke    : Factor w/ 4 levels "Heavy","Never",...: 2 4 3 2 2 2 2 2 2 2 ...
 $ Height   : num  173 178 NA 160 165 ...
 $ M.I      : Factor w/ 2 levels "Imperial","Metric": 2 1 NA 2 2 1 1 2 2 2 ...
 $ Age      : num  18.2 17.6 16.9 20.3 23.7 ...
NULL
```

# Chi Square Test

- For our model, we will consider the variables “Exer” and “Smoke”.
- The Smoke column records the students smoking habits while the Exer column records their exercise level.
- Our aim is to test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

```
> stu_data = data.frame(survey$Smoke, survey$Exer)  
> stu_data = table(survey$Smoke, survey$Exer)  
> print(stu_data)
```

	Freq	None	Some
Heavy	7	1	3
Never	87	18	84
Occas	12	3	4
Regul	9	1	7

# Chi Square Test

- And finally we apply the `chisq.test()` function to the contingency table `stu_data`.

```
chi_result <- chisq.test(stu_data)  
print(chi_result)
```

- As the p-value 0.4828 is greater than the .05, we conclude that the smoking habit is independent of the exercise level of the student and hence there is a weak or no correlation between the two variables.

Pearson's Chi-squared test

data: stu\_data

X-squared = 5.4885, df = 6, p-value = 0.4828