**Course 1 - Intro to Data Science**

**Becoming a Data Scientist?**

• Data Explosion

• Why Data Science?

• What is Data Science?

• Type of Analytics

• Data Science Portfolio

• Data Science Process

• Career in Data Science

**What is Data Science, Deep Learning, Machine Learning?**

• Data Science

• Machine Learning

• Deep Learning overview

**Introduction to Data Science Process - CRISP-DM**

• Solutions Methodologies (Macro vs. Micro)

* Scientific Research Method
* Operations Research Method | Water Fall Method | CRISP-DM

• CRISP-DM Process

* Business Understanding
* Data Understanding
* Data Preparation
* Modeling
* Evaluation
* Deployment

**This course is a combination of lecture and heavy hands-on exercises. The goal of the Statistics in R course is not to teach Statistics but to show the participant the R programming code to implement such statistics.**

**Statistics** – This process enables to confirm the assumptions. Hence, it uses the assumption to test using a statistical model.

Summary Commands with Single Value Results in R

### R Summary Commands Producing Multiple Results

### R Cumulative Statistics

### Simple Cumulative Commands in R

### Descriptive Statistics in R for Data Frames

### Distributions –

### Normal – 4 inbuilt functions and parameters

x <- seq(-20, 20, by = .1)

> y <- dnorm(x, mean = 5.0, sd = 1.0)

> plot(x,y, main = "Normal Distribution", col = "blue")

### Binomial - single variable discrete data where results are the no. of “successful outcomes”

### binomial-distribution-in-R-formula

### where, x = No. of success. (n-x) = No. of failures. (1-p) = Probability of failure. p^x = Probability of success.

### Poisson - distribution of rare events.

|  |  |
| --- | --- |
| Binomial | Poisson |
| Fixed no. of Trials (n) [10 pie throws], although, only two possible outcomes are possible.A probability of success is constant(p).Each trial is independent.Also, it predicts no.s of successes within a set no. of trials.We use it to test for independence. | * Infinite no. of trials. * Also, it has unlimited no. of outcomes possible. * The mean of the distribution is the same for all intervals. * No. of occurrence in any given interval independent of others. * Also, it predicts no. of occurrences per unit, time, space. * We use it to test for independence. |

### T-tests - We use t.test() which provides a variety of T-tests:

### One sample

### Paired sample

### Independent samples

### ANOVA – Analysis of Variance - measure the statistical difference between the means.

There are two ways of implementing ANOVA in R:

* One-way ANOVA
* Two-way ANOVA
* > attach(InsectSprays)
* > data(InsectSprays)
* > str(InsectSprays)
* > oneway.test(count~spray)
* Default is equal variances not assumed that is Welch’s correction applied and this explains why the denom df (which is k\*{n-1}) is not a whole number in the output O.
* In order to alter this, we set the “var.equal =” option to TRUE.
* Oneway.test( ) corrects the non-homogeneity but doesn’t give much information.
* > AOV\_Output <- aov(count ~ spray, data=InsectSprays)
* > summary(AOV\_Output)

#### Descriptive Statistics

1. With the help of descriptive statistics, we calculate the mean, variance and number of elements in each cell.
2. Visualise the data – boxplot; look at the distribution for outliers.

### Two-way ANOVA in R

Two-way Analysis of Variance

### Hypothesis Testing

### > qf(.97, 5, 23)

### Chi-Square

**Syntax of a chi-square test:**

chisq.test(data)

* Case Study – Effectiveness of a drug treatment
* Purpose and math of Chi-Square statistic
* Chi-Square Test
* R Code
* To test the effectiveness of a drug for a certain medical condition, we will consider a hypothetical case.
* Suppose we have 105 patients under study and 50 of them were treated with the drug. Moreover, the remaining 55 patients were kept under control samples. Thus, the health condition of all patients was checked after a week.
* With the following table, we can assess if their condition has improved or not. By observing this table, one can you tell if the drug had a positive effect on the patient?
* Here in this example, we can see that 35 out of the 50 patients showed improvement. Suppose if the drug had no effect, the 50 will split the same proportion of the patients who were not given the treatment. Here, in this case, improvement of the control case is high as about 70% of patients showed improvement, since both categorical variables which we have already defined must have only 2 levels. Also, it was sort of perceptive today that the drug treatment and health condition are dependent.

Particularly in this test, we have to check the p-values. Moreover, like all statistical tests, we assume this test as a null hypothesis and an alternate hypothesis.

The main thing is, we reject the null hypothesis if the p-value that comes out in the result is less than a predetermined significance level, which is 0.05 usually, then we reject the null hypothesis.

H0: The two variables are independent.  
H1: The two variables relate to each other.

In the case of a null hypothesis, a chi-square test is to test the two variables that are independent.

### 4. R Code

We will work on R by doing a chi-squared test on the treatment (X) and improvement (Y) columns in treatment.csv

First, read in the treatment.csv data.

> #Author DataFlair

> data\_frame <- read.csv("https://goo.gl/j6lRXD") #Reading CSV

> table(data\_frame$treatment, data\_frame$improvement)

Let’s do the chi-squared test using the chisq.test() function. It takes the two **vectors** as the input. We also set `correct=FALSE` to turn off Yates’ continuity correction.

### > chisq.test(data\_frame$treatment, data\_frame$improvement, correct=FALSE)

### We have a chi-squared value of 5.5569. Since we get a p-Value less than the significance level of 0.05, we reject the null hypothesis and conclude that the two variables are in fact dependent.

### Insert mini-challenge here

### Particularly for this challenge, first, find out if the ‘cyl’ and ‘carb’ variables are in ‘mtcars’ dataset and whether it is dependent or not.

> data("mtcars")

> table(mtcars$carb, mtcars$cyl)

Since there are more levels, therefore, it’s too hard to figure out if they relate to each other. Let’s use the chi-squared test instead.

**# Chi-sq test**

> chisq.test(mtcars$carb, mtcars$cyl)

### # independent 2-group T-test

### t.test(y~x) # where y is numeric and x is a binary factor

### # independent 2-group T-test

### t.test(y1,y2) # where y1 and y2 are numeric

### # paired T-test

### t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric

### # one sample T-test

### t.test(y,mu=3) # Ho: mu=3

We can use the var.equal = TRUE option to specify equal variances and a pooled variance estimate.

You can use them:

alternative=”less” or  
alternative=”greater”, option to specify one-tailed test.

# Contingency Tables in R (using a vector, table(), sort()

### Creating R Contingency Tables from Data table(), addmargins(), as.data.frame.matrix() and prop.table()

table0 <- table(toy\_data$c1, toy\_data$c2)

print.table(table0)

table1 <- as.data.frame.matrix(table0) # convert it to dataframe

print.data.frame(table1)

### Di

**Course 2: Basics of R Programming**

* How to use RStudio, a free and open-source development environment for R,
* Learn the fundamentals of R syntax,
* How to assign and manipulate variables,
* Learn the data types,
* Learn the data structures; vectors, matrices, lists, arrays, and data frames, factors
* Basic functions: str, summary, head, tail, View
* Packages - Import csv data into RStudio
* Explain the significance of Exploratory Data Analysis in Data Science,
* Demonstrate basic functions to create plots, graphs
* RMarkdown