R- Stastistics

R - Statistics

R Statistics concerns data; their collection, analysis, and interpretation.

It has the following two types:
Descriptive statistics: It is about providing a description of the data. It deals with the quantitative description of data through numerical representations or graphs.

Example: Normal distribution, Central Tendency, Kurtosis, etc. are some of the statistical techniques in Descriptive Statistics.

Inferential statistics: In inferential statistics, we draw conclusions or 'inferences' from our dataset. Also, a conclusion is drawn about the larger population from a data of a much smaller sample.

Example: Central Limit Theorem, Hypothesis Testing, ANOVA are some of the inferential statistics techniques.

Types of Data in Statistics

Different types of data in Statistics:

- Numerical (discrete and continuous)
- Categorical
- Ordinal

Data is nothing but information that is gathered as a result of a survey.

Data can either be numerical or categorical in nature.

Numerical Data is again of two types –
Discrete
Continuous.

a. **Discrete data –** It represents items that can be counted.

b. Continuous data – It represents measurements. Also, their possible values cannot be counted. Although, it can only be described using intervals on the real number line.

2. Categorical Data

Categorical Data is used to represent characteristics that are present in the data such as a person's gender, marital status, hometown.

3. Ordinal data

In this form of data, the variables have an ordered category which is natural and the distance between these variables is not known. Ordinal Data is similar to categorical data with the only difference that the data is ordered.

For example, Rating a restaurant on a scale of 0 to 4 gives us ordinal data.

ANOVA

ANOVA is used for testing the significance of the differences among more than two sample means.

Assumptions

- Each sample is randomly drawn from normal population
- Each of these population have same variance

Analysis of variance is based on comparison of two different estimates of the variance σ^2 ,of overall population.

Hypothesis:

- H_0 : All means are equal
- H1 : At least two means are not equal.

>aov(formula = Petal.Length ~ Species, data = iris)

```
Call:
   aov(formula = Petal.Length ~ Species, data = iris)

Terms:
                 Species Residuals
Sum of Squares  437.1028   27.2226
Deg. of Freedom        2       147

Residual standard error: 0.4303345
Estimated effects may be unbalanced
```

<u>We pass *two* arguments to the aov() function:</u>

1. For the formula parameter, we pass Petal.Length ~ Species. **This format is used for describing relationships we are testing.** The format is y ~ x, where the **response variables (e.g. y)** are to the left of the tilde (~) and **the predictor variables (e.g. x)** are to the right of the tilde. In this example, we are asking if petal length is significantly different among the three species.
2. We also need to define from where to find the Petal.Length and Species data, so we pass the variable name of the iris data.frame to the data parameter.

```
>petal.length.aov <- aov(formula = Petal.Length ~ Species, data = iris)
> petal.length.aov
Call:
   aov(formula = Petal.Length ~ Species, data = iris)

Terms:
                 Species Residuals
Sum of Squares  437.1028   27.2226
Deg. of Freedom        2       147

Residual standard error: 0.4303345
Estimated effects may be unbalanced
```

```
> summary(object = petal.length.aov)
            Df    Sum Sq  Mean Sq   F value Pr(>F)
Species      2    437.1    218.55    1180 <2e-16 ***
Residuals  147   27.2      0.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> sepal.width.aov <- aov(formula = Sepal.Width ~ Species, data = iris)
> summary(object = sepal.width.aov)
            Df   Sum Sq   Mean Sq   F value Pr(>F)
Species      2    11.35   5.672      49.16 <2e-16 ***
Residuals  147  16.96   0.115
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The species *do* have significantly different petal lengths ($P < 0.001$)

Simple Linear Regression in R: The **simple linear regression** is used to predict a quantitative outcome y on the basis of one single predictor variable x. The goal is to build a mathematical model (or formula) that defines y as a function of the x variable.

Formula and basics

The mathematical formula of the linear regression can be written as **y = b0 + b1*x + e**, **where**:

b0 and b1 are known as the regression *beta coefficients* or *parameters*:

b0 is the *intercept* of the regression line; that is the predicted value when x = 0.

b1 is the *slope* of the regression line.

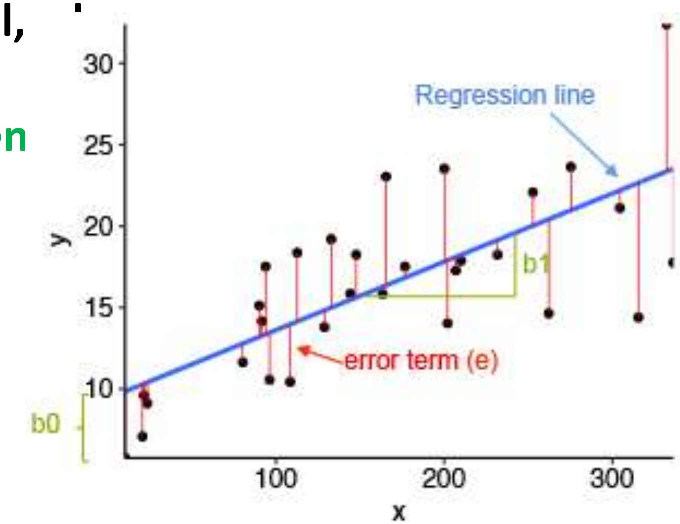
e is the *error term* (also known as the *residual errors*), the part of y that can be explained by the regression model

The figure below illustrates the linear regression model,

the best-fit regression line is in blue

the intercept (b0) and the slope (b1) are shown in green

the error terms (e) are represented by vertical red lines



•From the scatter plot above, it can be seen that not all the data points fall exactly on the fitted regression line. Some of the points are above the blue curve and some are below it; overall, the residual errors (e) have approximately mean zero.

•The sum of the squares of the residual errors are called the **Residual Sum of Squares** or **RSS**.

•The average variation of points around the fitted regression line is called the **Residual Standard Error** (**RSE**). This is one the metrics used to evaluate the overall quality of the fitted regression model. The lower the RSE, the better it is.

Since the mean error term is zero, the outcome variable y can be approximately estimated as follow:

y ~ b0 + b1*x

Once, the beta coefficients are calculated, a t-test is performed to check whether or not these coefficients are significantly different from zero. A non-zero beta coefficients means that there is a significant relationship between the predictors (x) and the outcome variable (y).

```
> linearMod <- lm(dist ~ speed, data=cars)  # build linear regression model on full data
> print(linearMod)

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
(Intercept)       speed
   -17.579       3.932
```

you can notice the ‘Coefficients’ part having two components: *Intercept*: -17.579, *speed*: 3.932 These are also called the beta coefficients. In other words,

$$\boldsymbol{dist = Intercept + (\beta * speed)}$$

=> dist = −17.579 + 3.932∗speed

Linear Regression Diagnostics

```
> summary(linearMod)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
             Estimate   Std. Error  t value  Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324      0.4155    9.464  1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

The summary outputs shows 6 components, including:

Call. Shows the function call used to compute the regression model.

Residuals. Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute value.

Coefficients. Shows the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.

Residual standard error (RSE), **R-squared** (R2) and the **F-statistic** are metrics that are used to check how well the model fits to our data.

Coefficients significance

The coefficients table, in the model statistical summary, shows:

- the estimates of the **beta coefficients**
- the **standard errors** (SE), which defines the accuracy of beta coefficients. For a given beta coefficient, the SE reflects how the coefficient varies under repeated sampling. It can be used to compute the confidence intervals and the t-statistic.
- the **t-statistic** and the associated **p-value**, which defines the statistical significance of the beta coefficients.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-17.5791	6.7584	-2.601	0.0123	*
speed	3.9324	0.4155	9.464	1.49e-12	***