Data Science Notes (Practical Statistics) (RStudio) (Day 1)

Date: July 11th 2019

* Descriptive Statistics
  + Have a well-described data without generalization
  + The Centre of Data
    - Mean
      * Most commonly used method
      * Descriptive Statistics: Mean is commonly used for the centre of data.
      * Statistical Inference: Mean is commonly used for moving average data-time series.
      * Pro and Con:
        + (+)
        + (-) : Not very accurate for making is as an expectation (Not very representative)
      * Function(s) to get it: Use the *‘mean()’* function
    - ‘Trimmed Mean’
      * The ‘more representative version’ of mean.
      * Function(s) to get it: add a 'trim' parameter at the 'mean' function. the number as the parameter input has to be in percentage.
      * NOTE: You can't cut more than 5% of total data.
    - Median
      * Most common alternative method
      * Data division concept:
        + divided by 2: Min |---------------| Median |----------------| Max
        + divided by 4: Min |--| 1st Q |--| Median |--| 3rd Q |--| Max
      * Pros and Cons;
        + (+): Rarely get some outlier (Robust)
        + (-): Too robust
      * Function(s) to get it:
        + *‘median()’*
        + *‘fivenum()’*
        + *‘quantile()’*

NOTE: You can display a specified part of *‘quantile()’* by adding the name as the next parameter

* + - * + *‘summary()’*
    - Mode
      * Unfortunately, there is no official function to get mode.
      * However, there is a custom function to get mode.
  + ‘Common’ Range
    - Interquartile Range (IQR)
      * Difference of 1st quartile and 3rd quartile
      * Function(s) to get it:
        + Way #1: Use ‘IQR()’ function
        + Way #2: Subtract ‘quantile(\*database\*, 0.75)’ with ‘quantile(\*database\*, 0.25)’
      * Upper Whisker / outer Fence: Q3 + k (1.5) \* IQR
      * Lower Whisker / Inner Fence: Q1 - k (1.5) \* IQR
    - Variance and Standard Deviation
      * Variance:
        + Function(s) to get it:

Way #1: use var() function

Way #2: sum((\*database\*-mean(\*database\*))^2)/(length(\*database\*)-1)

* + - * Standard Deviation
        + Basically the square root of Variance
        + Function(s) to get it:

Way #1: use sd() function

Way #2: square root the variance by using the sqrt() function

* + Distribution
    - Covariance and Correlation
      * Covariance
        + Covariance usually used only to compare if both values have the same behaviour
        + Function(s) to get it: use the cov() function
      * Correlation:
        + Correlation is similar to probability. If the value is getting into 1, the behaviour of both values become more equal and so the opposite.
        + Function(s) to get it: use the cor() function
    - Normal Distribution
      * Characteristics:
        + Bell curve
        + Symmetrical
        + Mean = Median = Mode
        + Uni-Mode
* Statistical Inference