**Contents**

1. **Descriptive stats: Mean, median, SD, Correlation, covariance, box plots, histograms**
2. **Probability - Simple**
3. **Probability - conditional and joint. Application focused**
4. **Binomial and Normal distribution**
5. **CLT, Construction of confidence intervals, setting up hypothesis**
6. **1 sample testing, 2 sample testing for mean not proportion**
7. **Chi square analysis and ANOVA**

**Suggested Book:** *Business Statistics A First Course by Dr. P. K. Viswanathan and others.*

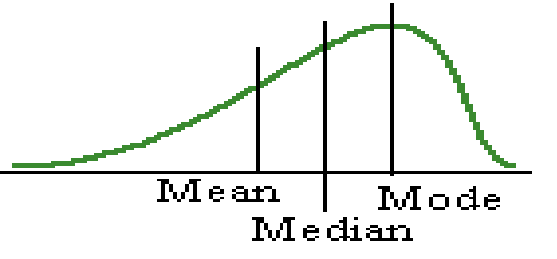
**Please down load all the data sets used in this book by referring to page 529.**

1. **Descriptive stats:** 
   1. **Central Tendency** is the extent to which the values of a numerical variable group around a typical, or central value.

* The arithmetic **mean** or commonly known as mean is the most common measure of central tendency. The mean is the only common measure is which all the values play an equal role.
* The **median** is the middle value in an ordered array of data that has been ranked from smallest to highest. Half the values are larger than or equal to the median and half the values are larger than or equal to the median.
* The **Mode** is the value that appears most frequently.
  1. **Variation** measures the spread or dispersion of the values of a numerical variable.
* **Range** is the difference between the largest and smallest value and is the simplest descriptive measure of variation for a numerical variable.
* **Standard Deviation (SD) or variance** is another measure of variation most commonly used in Statistics. Variance is the average of the squares of difference between the data points and the mean. Standard Deviation is the square root of the variance.
* **Coefficient of variation** is equal to the standard deviation divided by the mean, multiplied by 100%. This measures the scatter in the data relative to the mean. CV is a relative measure of variation that is always expressed as a percentage rather than in terms of units of the particular data.
* **Z score** of a value is the difference between a value and its mean, divided by the standard deviation. A Z score of 0 indicates that the value is the same as mean. If a Z score is positive or negative number, it indicates whether the value is above or below the mean and by how many standard deviation. Z scores help identify outliers (extreme values). As a general rule, a Z score that is less than -3 or greater than 3, indicates an outlier value.
  1. **Shape** of a variable represents a pattern of all the values from the lowest to the highest values.
* **Skewness** measures the extent to which the data values are not symmetrical around the mean:

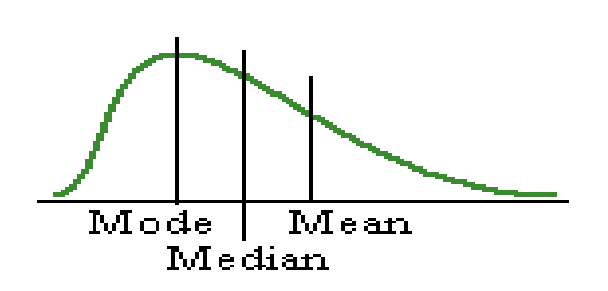
1. **Left-skewed distribution**

**Mean < Median; negative skewness**

****

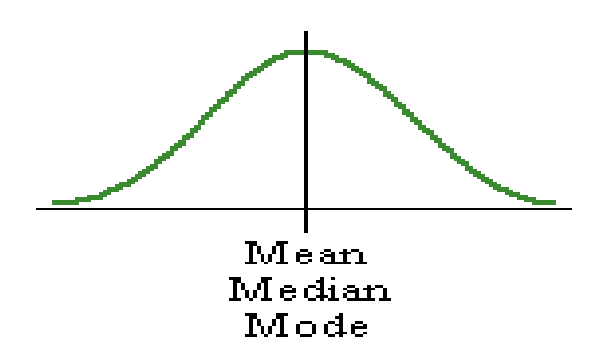
1. **Right skewed distribution**

**Mean > Median; negative skewness**

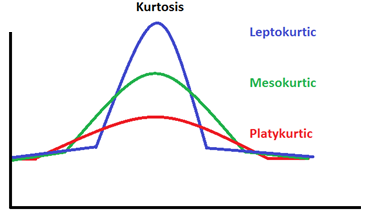
****

1. **Symmetrical distribution**

**Mean = Median; zero skewness**

****

* **Kurtosis** measures the peakedness of the curve of the distribution.

****

1. **Leptokurtic:** In this case, the distribution has a sharper-rising centre peak than the peak of the normal distribution and has a positive kurtosis.
2. **Platykurtic:** In this case, the distribution has a flatter centre peak than the peak of the normal distribution and has a negative kurtosis.
3. **Mesokurtic:** In this case, the distribution has kurtosis which is similar to the kurtosis of the normally distributed data set**.**
   1. **Quartiles**

Quartiles split the values into four equal parts:

* The **first quartile (Q1)** divides the smallest 25% of the values from the other 75% that are larger.
* The **second quartile (Q2)** is the median; 50% of the values are smaller or equal to the median.
* The **third quartile (Q3)** divides the smallest 75% of the values from the largest 25%.
* **Interquartile Range (IQR)** also called the mid-spread, measures the difference in the centre of the distribution between the third and first quartile.
* **A five number summary** for a variable consists of theMinimum or smallest value, the first quartile, median, third quartile and the largest value.
  1. **Histogram**: Histogram accurately represents the distribution of numerical data.

It was first introduced by Karl Pearson.

[](http://en.wikipedia.org/wiki/Karl_Pearson)

* 1. **Boxplot**: Boxplot (aka box and whisker plot) is a graph that depicts groups of numerical data through their quartiles. It is useful to compare distributions across groups.

John Tukey introduced this plot in 1970.****

**Please read the chapter 2, 3.1, 3.2 and 3.3**

**Example 1:**

A bank branch located in a commercial district of a city has the business objective of developing an improved process for serving customers during the noon-to-1 P.M. lunch hour. The waiting time, in minute, collected from a sample of 15 customers during this hour are stored in Bank2 data-set.

9.66, 5.90, 8.02, 5.79, 8.73, 3.82, 8.01, 8.35, 10.49, 6.68, 5.64, 4.08, 6.17, 9.91, 5.47

a. Compute the mean and median.

b. Compute the variance, standard deviation, range, coefficient of variation, Z scores. Draw a box-plot. Are there any outliers? If so, how?

c. Draw a histogram plot. Are the data skewed? If so, how?

d. As a customer walks into the branch office during the lunch hour, he asks the branch manager how long he can expect to wait. The branch manager replies, “Almost certainly less than 5 minutes”. Based on the results from a through c, evaluate the accuracy of the statement.

**Answer:**

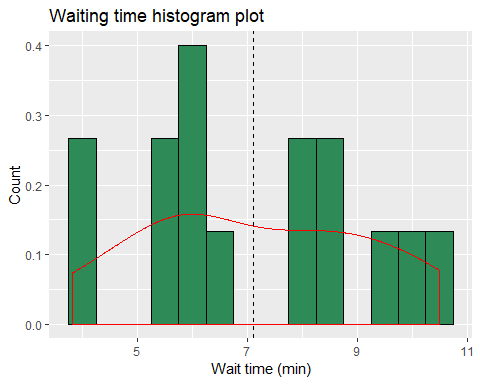
1. **Mean = 7.114667, Median = 6.68**
2. **Variance = 4.336, SD = 2.082189, Range = 6.67, CV = 29.27%**

|  |  |
| --- | --- |
| ## Wait Z ## 1 9.66 1.2224 ## 2 5.90 -0.5834 ## 3 8.02 0.4348 ## 4 5.79 -0.6362 ## 5 8.73 0.7758 ## 6 3.82 -1.5823 ## 7 8.01 0.4300 ## 8 8.35 0.5933 ## 9 10.49 1.6211 ## 10 6.68 -0.2088 ## 11 5.64 -0.7082 ## 12 4.08 -1.4574 ## 13 6.17 -0.4537 ## 14 9.91 1.3425 ## 15 5.47 -0.7899 |  |

**There are no outliers since for any of the Wait values, Z score is neither less than -3 or greater than 3. From the Box plot also, we don’t find any outlier.**

1. **We observe that Mean > Median indicating that there are some usually high values and the distribution is right skewed. This can be seen in the below histogram:**

*Note: Average waiting time is shown as a dashed line in the histogram.*



1. **Mean and Median are greater than 5 minutes. There are usually some high values since mean > median. Out of 15 observations, there are only two observations (3.82 and 4.08) are below 5 and 13 observations are above 5. Hence the customer is likely to experience a waiting time in excess of 5 minutes. So the manager’s statement is not accurate.**

**R code:**

library(ggplot2)

Bank2 <- c(9.66, 5.90, 8.02, 5.79, 8.73, 3.82, 8.01, 8.35, 10.49, 6.68, 5.64, 4.08, 6.17, 9.91, 5.47)

df <- data.frame(Wait = Bank2)

## Compute the mean and median

Bank2\_mean <- mean(Bank2)

Bank2\_median <- median(Bank2)

cat("\nMean: ", mean(Bank2\_mean))

cat("\nMedian: ", median(Bank2\_median))

## Compute the variance, standard deviation, range, coefficient of variation, Z scores.

Bank2\_sd <- sd(Bank2)

Bank2\_var <- Bank2\_sd \* Bank2\_sd

Bank2\_hi <- max(Bank2)

Bank2\_lo <- min(Bank2)

Bank2\_range <- Bank2\_hi - Bank2\_lo

Bank2\_cv <- round((Bank2\_sd / Bank2\_mean)\*100,2)

Bank2\_Z <- round((Bank2 - Bank2\_mean) / Bank2\_sd,4)

cat("\nVariance: ", Bank2\_var)

cat("\nStandard Deviation: ", Bank2\_sd)

cat("\nHighest: ", Bank2\_hi)

cat("\nLowest: ", Bank2\_lo)

cat("\nRange: ", Bank2\_range)

cat("\nCV: ", Bank2\_cv)

df$Z <- Bank2\_Z

cat("\nZ Score: ")

print(df)

## Histogram

ggplot(df, aes(x = Wait)) +

geom\_histogram(aes(y=..density..), binwidth = 0.5, fill = 'seagreen', colour="black") +

geom\_density(color = 'red') +

labs(title="Waiting time histogram plot",x="Wait time (min)", y = "Count") +

geom\_vline(data = df, aes(xintercept = Bank2\_mean),

linetype="dashed")

## Boxplot

ggplot(df, aes(x = "", y = Wait)) +

geom\_boxplot(color = 'red') +

labs(title="Waiting time Box plot",x="Wait time (min)", y = "Count")

**Exercise 1:** Do the exercise 3.60 in page 140 of the prescribed book.

* 1. **Covariance and Correlation**

The covariance measures the strength of the linear relationship between two numerical variables (X and Y).

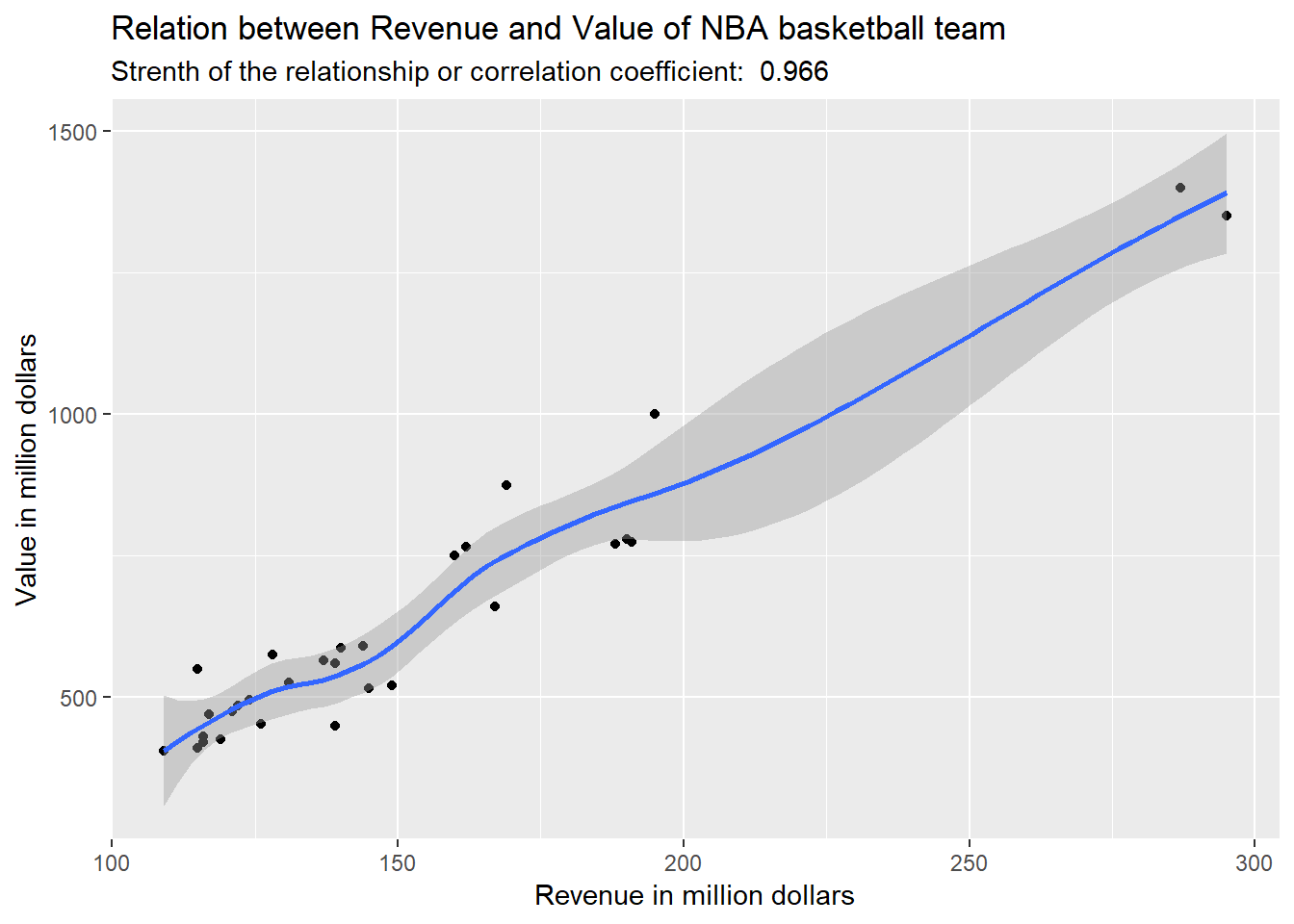
The correlation coefficient measures the relative strength of a linear relationship between two numerical variables. Values of a correlation coefficient ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation).

Scatter plot explores the possible relationship between two numerical variables by plotting the values of one numerical variable on the horizontal axis and the values of a second numerical variable on the vertical axis.

**Please read the chapter 2.5 and 3.5**

**Example 2: Refer to example 2.12 in page number 62 of the book.**

1. *Construct a scatter plot..*
2. *Measure the relative strength of a linear relationship between the revenue and value by determining the sample correlation coefficient.*



*We observe that there is a strong increasing (positive) relationship between revenue and value of a team. When team generates higher revenues, they have higher values. This relationship is highlighted by the addition of a linear prediction line. Strength of the relationship is very good as the correlation coefficient is 0.966.*

R code:

*library(ggplot2)*

*NBA\_Revenue\_value < -data.frame(*

*team\_code = c('ATL','BOS','BKN','CHA','CHI','CLE','DAL','DEN','DET','GSW','HOU','IND','LAC','LAL','MEM',*

*'MIA','MIL','MIN','NOH','NYK','OKC','ORL','PHI','PHX','POR','SAC','SAS','TOR','UTA','WAS'),*

*revenue =*

*c(119, 169, 190, 115, 195, 145, 162, 124, 139, 160, 191, 121, 128, 295, 126, 188, 109, 116, 116, 287, 144, 139, 117, 137, 140, 115, 167, 149, 131, 122),*

*value =*

*c(425, 875, 780, 410, 1000, 515, 765, 495, 450, 750, 775, 475, 575, 1350, 453, 770, 405, 430, 420, 1400, 590, 560, 469, 565, 587, 550, 660, 520, 525, 485))*

*cor\_coeff = round(cor(NBA\_Revenue\_value$value, NBA\_Revenue\_value$revenue),5)*

*qplot(x = revenue,y = value, data = NBA\_Revenue\_value, geom = c("point","smooth")) +*

*labs(*

*x = "Revenue in million dollars",*

*y = "Value in million dollars",*

*title = "Relation between Revenue and Value of NBA basketball team",*

*subtitle = paste("Strength of the relationship or correlation coefficient: ",cor\_coeff)*

*)*

**Exercise 2:** Do the exercise 3.38 in page 135

**Exercise 3:** Do the exercise 3.40 in page 136

**Exercise 4:** Do the exercise 3.41 in page 136

1. **Probability - Simple Probability - conditional and joint. Application focused**

**2.1 A probability** is the numerical value representing the chance, likelihood, or possibility that a particular event will occur, such as the price of a stock increasing, a rainy day, a defective product, etc.

* **Simple probability** refers to the probability of occurrence of a sample event, P(A).
* **Joint probability** refers to the probability of an occurrence involving two or more events.
* **Marginal probability** of an event consists of a set of joint probabilities.
* **Conditional probability** refers to the probability of event A, given the information about the occurrence of another event B.
* Bayes’ theorem is used to revise previously calculated probabilities bases on new information.

**Please read the chapter 4**

**Exercise 5:** Do the exercise 4.12 in page 159

**Exercise 6:** Do the exercise 4.14 in page 159

**Exercise 7:** Do the exercise 4.24 in page 166

**Exercise 8:** Do the exercise 4.26 in page 166

**Exercise 9:** Do the exercise 4.34 in page 171

**Exercise 10:** Do the exercise 4.36 in page 171

**Exercise 11:** Do the exercise 4.48 in page 175

1. **Binomial**

**Please read the chapter 5.1 and 5.2**

**Exercise 12:** Do the exercise 5.14 in page 195

**Exercise 13:** Do the exercise 5.16 in page 195

1. **Normal distribution**

**Please read the chapter 6**

**Exercise 14:** Do the exercise 6.30 in page 228

**Exercise 15:** Do the exercise 6.36 in page 228

**Exercise 16:** Do the exercise Case study in page 226

1. **CLT, Construction of confidence intervals, setting up hypothesis**

**Please read the chapter 7, 8 and 9**

**Example:** Refer the example 7.5 in page 243 and try to do the same..

**Exercise 17:** Do the exercise 7.8 in page 246

**Example:** Refer the example 8.1 and 8.2 in page 263, 264 and try to do the same..

**Exercise 18:** Do the exercise 8.10 in page 265

**Example:** Refer the example 8.3 in page 269 and try to do the same..

**Exercise 19:** Do the exercise 8.22 in page 272

**Exercise 20:** Do the exercise 8.24 in page 272

**Example:** Refer the example 8.4 in page 274

**Exercise 21:** Do the exercise 8.28 in page 275

**Example:** Refer the example 8.4 in page 277 and example 8.6 in page 280.

**Exercise 21:** Do the exercise 8.40 in page 280

**Exercise 22:** Do the exercise 8.48 in page 281

**Example:** Refer the example 9.3 in page 303 and example 9.4 in page 306.

**Exercise 23:** Do the exercise 9.14 in page 308

**Exercise 24:** Do the exercise 9.16 in page 308

**Example:** Refer pages from 309 to 312

**Exercise 25:** Do the exercise 9.24 in page 313

**Exercise 26:** Do the exercise 9.32 in page 314

1. **1 sample testing, 2 sample testing for mean not proportion**

**Example:** Refer the example 9.5 in page 317

**Exercise 27:** Do the exercise 9.50 in page 319

**Example:** Refer the example 9.6 in page 321

**Exercise 28:** Do the exercise 9.56 in page 322

**Example:** Please refer to the Mini Case in page 325

**2-sample testing**

Read chapter 10

**Example:** Refer the example 10.1 in page 339

**Exercise 29:** Do the exercise 10.8 in page 343

**Exercise 30:** Do the exercise 10.12 in page 344

**Example:** Refer the example 10.2 in page 339

**Exercise 31:** Do the exercise 10.24 in page 352

**Example:** Refer the example 10.4 in page 362

**Exercise 32:** Do the exercise 10.50 in page 364

1. **ANOVA**

Read chapter 10.5, 10.6

**Example:** Refer the example 10.5, 10.6 in page 375, 376

**Exercise 33:** Do the exercise 10.58 in page 379

**Exercise 34:** Do the exercise 10.64 in page 381

**Exercise 35:** Do the exercise 10.86 in page 387

1. **Chi square analysis**

Read chapter 11

**Example: Refer the example 11.1 in page 406**

**Exercise 36: Do the exercise 11.10 in page 410**

**Example: Refer the example 11.2 in page 412**

**Exercise 37: Do the exercise 11.18 in page 414**

**Exercise 38: Do the exercise 11.26 in page 420**