Business Club

Basic Statistics

Analytics Team

October, 2017

INDEX

CLEANING THE DATA :

1. Variable Identification
2. Univariate Analysis
3. Multivariate Analysis
4. Missing values treatment
5. Outlier treatment
6. Variable transformation *
7. Variable creation *

Variable Identification :

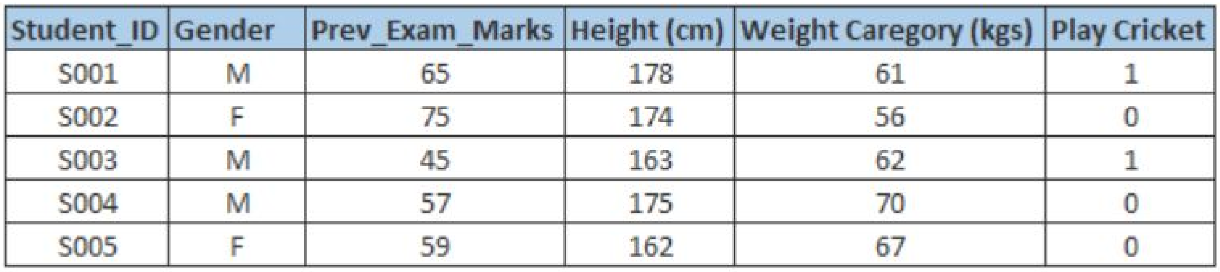
Continuous variable : These variables can take number of numeric values.
Eg :temperature

Categorical variable : These variables contain a finite number of categories or groups .
Eg : Gender

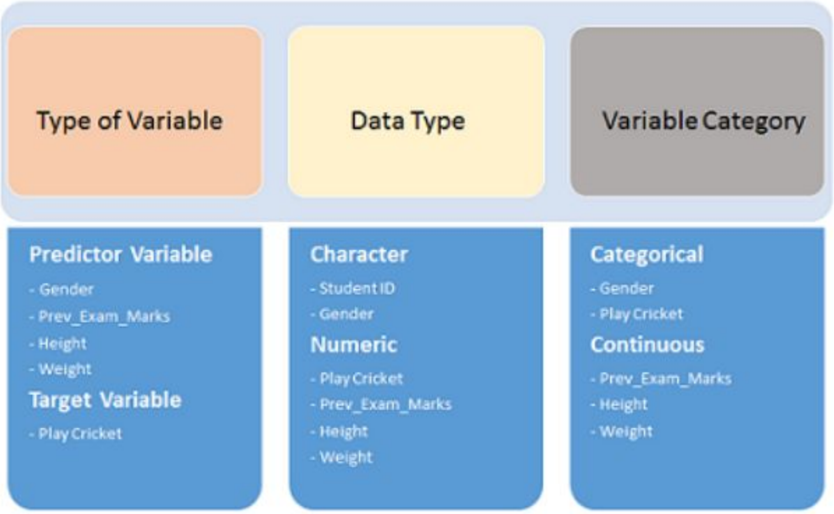
For any dataset perhaps the most important part is identifying Predictor (Input) and Target (output) variables. Next is identifying the data type and category of the variables.

Let’s understand this step more clearly by taking an example.

Example:- Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables.



Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0



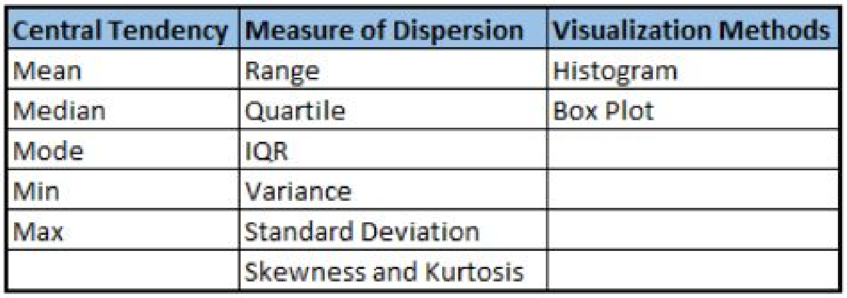
Type of Variable	Data Type	Variable Category
Predictor Variable - Gender - Prev_Exam_Marks - Height - Weight **Target Variable** - Play Cricket	**Character** - Student ID - Gender **Numeric** - Play Cricket - Prev_Exam_Marks - Height - Weight	**Categorical** - Gender - Play Cricket **Continuous** - Prev_Exam_Marks - Height - Weight

Univariate Analysis :

Univariate analysis is the simplest form of analyzing data. This deals with only variable . It is mostly used to find patterns in the data.

Here also, the approaches for analysing the categorical and continuous variables are separate:

Continuous Variables:
Here, we need to understand the central tendency of the data and spread of the variable. These are measured using various statistical metrics visualization methods as shown below:



Central Tendency	Measure of Dispersion	Visualization Methods
Mean	Range	Histogram
Median	Quartile	Box Plot
Mode	IQR	
Min	Variance	
Max	Standard Deviation	
	Skewness and Kurtosis	

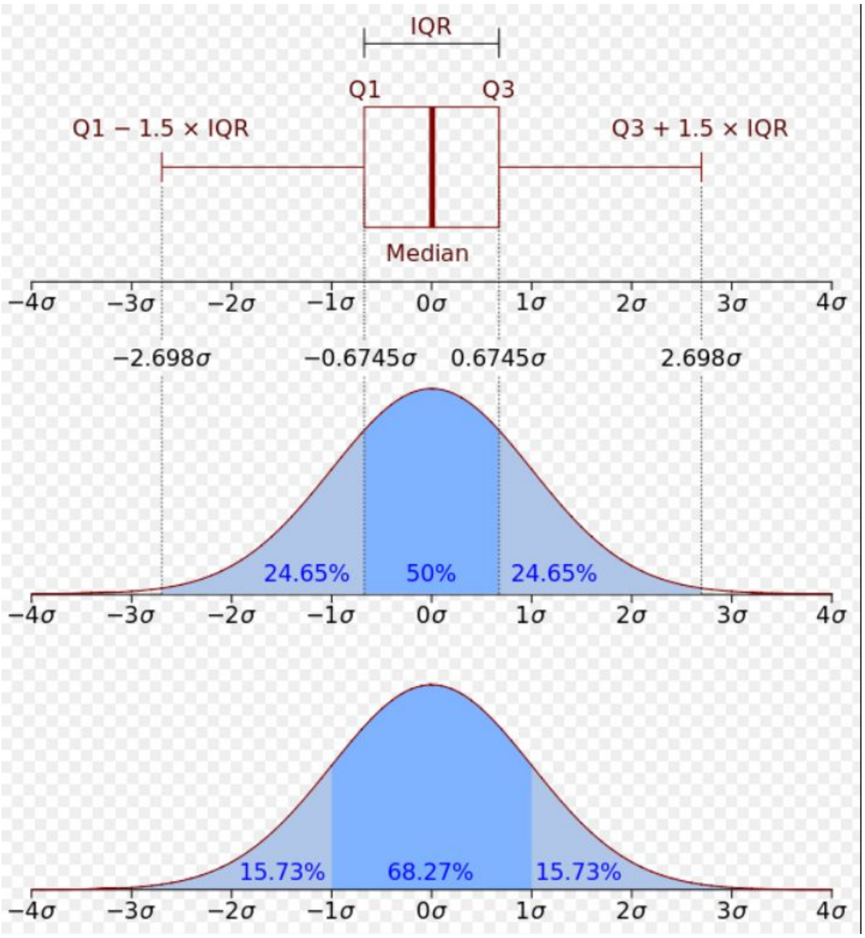
Most terms in the table are self explanatory, except for some which we'll discuss here:

Quartile and InterQuartile Range(IQR):

Quartiles are data values which divide the dataset in four equal parts. The first quartile(Q1) is the middle point of initial value and median whereas Q2 is the median of the data. Q3 is the middle value between median and the highest value.

IQR as the name suggests is the the measure of dispersion between upper and lower quartile or;

IQR=Q3-Q1



Kurtosis: In a similar way to the concept of skewness, *kurtosis* is a descriptor of the shape of a probability distribution and, just as for skewness, there are different ways of quantifying it for a theoretical distribution and corresponding ways of estimating it from a sample from a population.

Mathematical tools regarding quantifying Kurtosis would be discussed later.

Categorical Variables:

For categorical variables, we can use frequency table to understand distribution of each category. We can also read as percentage of values under each category. It can be be measured using two metrics, Count and Count% against each category. Bar chart can be used as visualization.

Multivariate Analysis :

Any method that is used to analyse data with more than one variable is called multivariate analysis.

In multivariate Analysis we'll discuss bi-variate analysis in greater detail as it is the most frequently used:

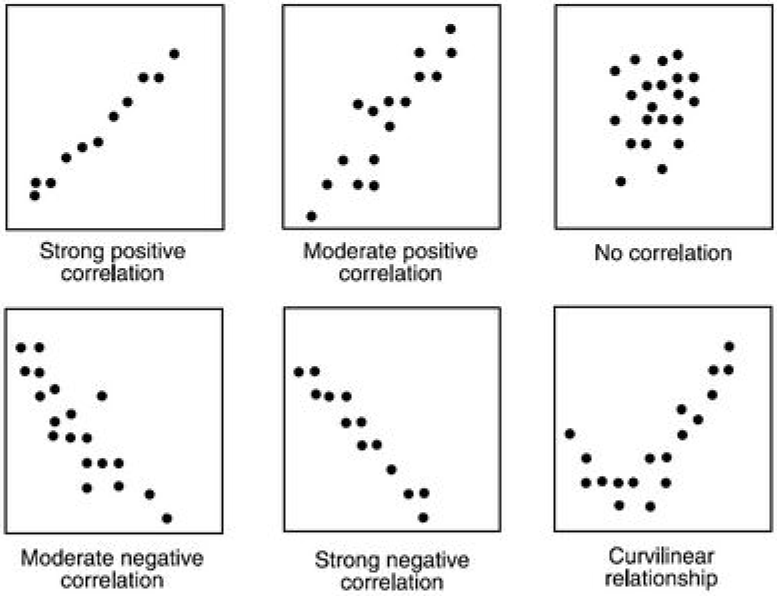
Bi-Variate Analysis:

Bi-variate Analysis finds out the relationship between two variables. Here, we look for association and disassociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables. The combination can be: Categorical & Categorical, Categorical & Continuous and Continuous & Continuous. Different methods are used to tackle these combinations during analysis process.

Correlation

It is a statistical measure that indicates the extent to which two or more variables fluctuate together. A positive correlation indicates the extent to which those variables increase or decrease in parallel; a negative correlation indicates the extent to which one variable increases as the other decreases.

Continuous & Continuous: While doing bi-variate analysis between two continuous variables, we should look at scatter plot. The pattern of scatter plot indicates the relationship between variables. The relationship can be linear or non-linear.



Strong positive correlation

Moderate positive correlation

No correlation

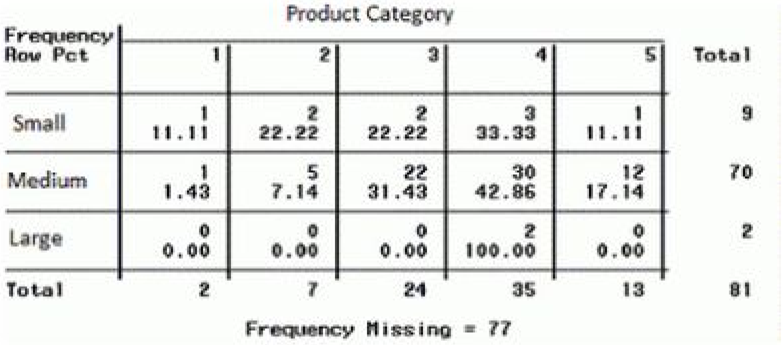
Moderate negative correlation

Strong negative correlation

Curvilinear relationship

Categorical & Categorical: To find the relationship between two categorical variables, we can use following methods:

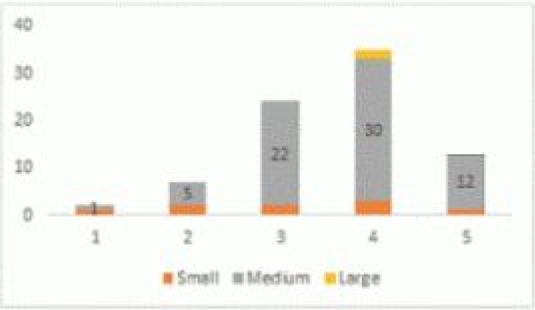
- Two-way table: We can start analyzing the relationship by creating a two-way table of count and count%. The rows represents the category of one variable and the columns represent the categories of the other variable. We show count or count% of observations available in each combination of row and column categories.



Frequency Row Pct	Product Category 1	2	3	4	5	Total
Small	1 11.11	2 22.22	2 22.22	3 33.33	1 11.11	9
Medium	1 1.43	5 7.14	22 31.43	30 42.86	12 17.14	70
Large	0 0.00	0 0.00	0 0.00	2 100.00	0 0.00	2
Total	2	7	24	35	13	81

Frequency Missing = 77

- Stacked Column Chart: This method is more of a visual form of Two-way table.



- **Chi-Square Test:** This test is used to derive the statistical significance of relationship between the variables. Also, it tests whether the evidence in the sample is strong enough to generalize that the relationship for a larger population as well. Chi-square is based on the difference between the expected and observed frequencies in one or more categories in the two-way table. It returns probability for the computed chi-square distribution with the degree of freedom.

  Probability of 0: It indicates that both categorical variable are dependent
  Probability of 1: It shows that both variables are independent.
  Probability less than 0.05: It indicates that the relationship between the variables is significant at 95% confidence. The chi-square test statistic for a test of independence of two categorical variables is found by:

$$X^2 = \sum (O - E)^2 / E$$

  where *O* represents the observed frequency. *E* is the expected frequency under the null hypothesis and computed by:

$$E = \frac{row\ total \times column\ total}{sample\ size}$$

  Different data science language and tools have specific methods to perform chi-square test. In SAS, we can use **Chisq** as an option with **Proc freq** to perform this test.

  Categorical & Continuous: While exploring relation between categorical and continuous variables, we can draw box plots for each level of categorical variables. If levels are small in number, it will not show the statistical significance. To look at the statistical significance we can perform Z-test, T-test or ANOVA. These will be discussed in detail later.

Missing values treatment

Missing values are categorised broadly into four types :

- **Missing completely at random:** This is a case when the probability of missing variable is same for all observations. For example: respondents of data collection process decide that they will declare their earning after tossing a fair coin. If an head occurs, respondent declares his / her earnings & vice versa. Here each observation has equal chance of missing value.
- **Missing at random:** This is a case when variable is missing at random and missing ratio varies for different values / level of other input variables. For example: We are collecting data for age and female has higher missing value compare to male.
- **Missing that depends on unobserved predictors:** This is a case when the missing values are not random and are related to the unobserved input variable. For example: In a medical study, if a particular diagnostic causes discomfort, then there is higher chance of drop-out from the study.
- **Missing that depends on the missing value itself:** This is a case when the probability of missing value is directly correlated with missing value itself. For example: People with higher or lower income are likely to provide non-response to their earning.

1. **Deletion:**

   Here ,we delete observations where any of the variable is missing. Even though this method is easy to handle but the sample size reduces .

   Deletion methods are used when the nature of missing data is "**Missing completely at random**" else non random missing values can bias the model output.

2. Mean/ Mode/ Median Imputation: Imputation is a method to fill in the missing values with estimated ones. In this method the missing values are replaced with mean / mode / median of the known values under that variable.

It can be of two types:-

- ○ **Generalized Imputation:** In this case, we calculate the mean or median for all non missing values of that variable then replace missing value with mean or median,

- **Similar case Imputation:** In this case, we calculate average for gender “**Male”** (29.75) and “**Female**” (25) individually of non missing values then replace the missing value based on gender. For “**Male**“, we will replace missing values of manpower with 29.75 and for “**Female**” with 25.

3.. **Prediction Model**: Prediction model is one of the sophisticated method for handling missing data. Here, we create a predictive model to estimate values that will substitute the missing data.

In this case the data is divided into two sets –one with no missing values for the variable and another one with missing values. The first is treated as a training data set to predict the missing values of the other set .Major drawbacks of this method are :

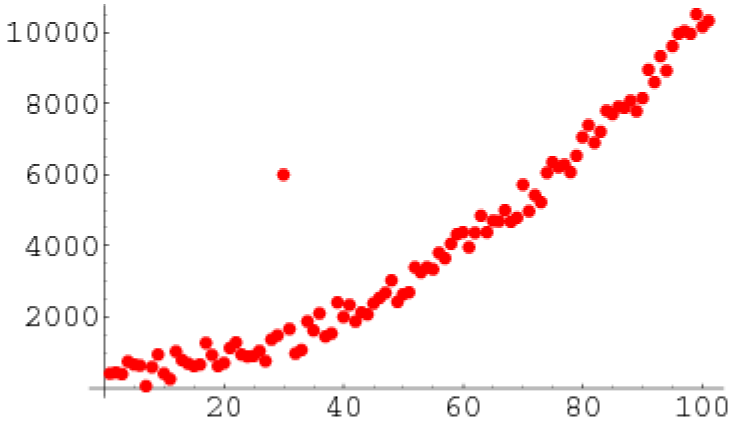
1. The model estimated values are usually more well-behaved than the true values
2. If there are no relationships with attributes in the data set and the attribute with missing values, then the model will not be precise for estimating missing values.

DEALING WITH MISSING VALUES IN R AND PYTHON :

http://www.statmethods.net/input/missingdata.html
https://www.r-bloggers.com/missing-value-treatment/

OUTLIERS :

Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

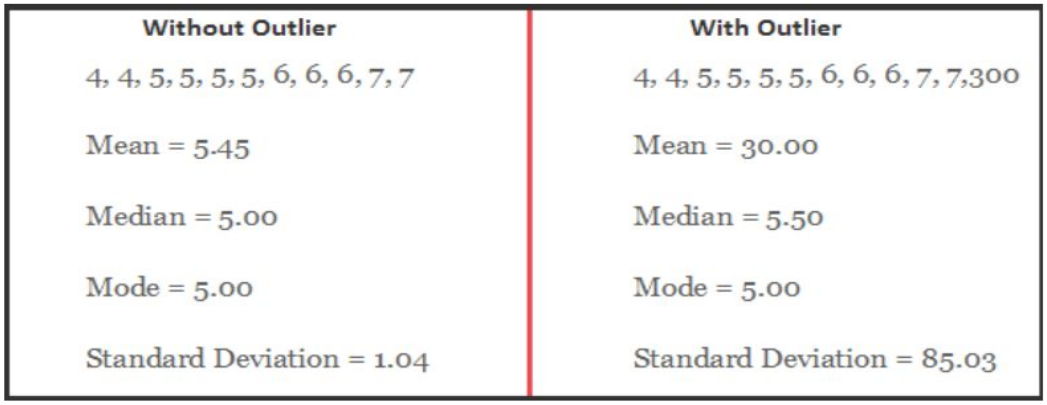


What is the impact of Outliers on a dataset?

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavourable impacts of outliers in the data set:

- It increases the error variance and reduces the power of statistical tests
- They can bias or influence estimates that may be of substantive interest
- They can also impact the basic assumption of Regression,and other statistical model assumptions.

To understand the impact deeply, let’s take an example to check what happens to a data set with and without outliers in the data set.



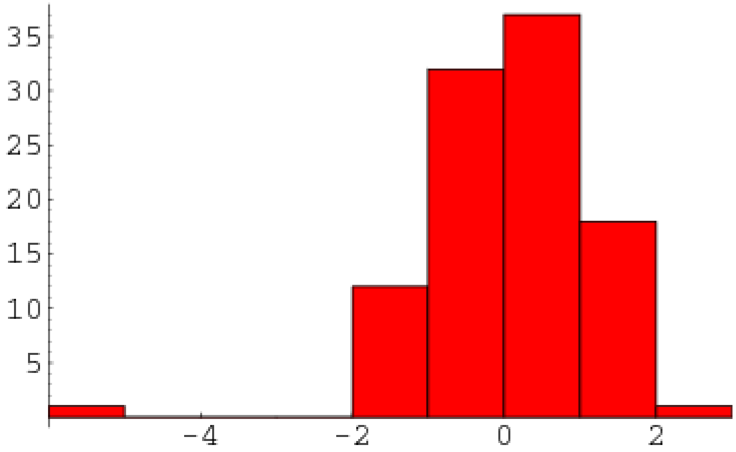
Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

The table probably emphasizes enough how important handling outliers is. Now we should move on to detecting and handling outliers.

HOW TO DETECT OUTLIERS :

Visualization through box plots , histogram :

Outliers are easily detected through visualisations using histogram or other plots .



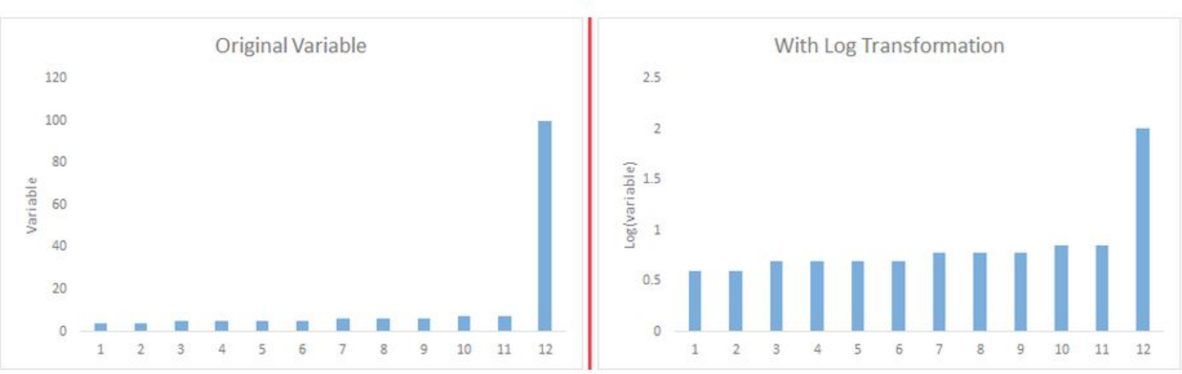
Thumb rules to detect outliers

1. Any value, which is beyond the range of -1.5 x IQR to 1.5 x IQR
2. Any value which out of range of 5th and 95th percentile can be considered as outlier
3. Data points, three or more standard deviation away from mean are considered outlier.

Dealing with outliers

1. **Deletion :** We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.

2. **Transforming and binning values:** Transforming variables can also eliminate outliers. Natural log of a value reduces the variation caused by extreme values. Binning is also a form of variable transformation. Decision Tree algorithm allows to deal with outliers well due to binning of variable. We can also use the process of assigning weights to different observations.



3. **Treating them separately:** If there are significant number of outliers, we should treat them separately in the statistical model. One of the approach is to treat both groups as two different groups and build individual model for both groups and then combine the output

4. **Imputing:** Like imputation of missing values, we can also impute outliers. We can use mean, median, mode imputation methods. Before imputing values, we should analyse if it is natural outlier or artificial. If it is artificial, we can go with imputing values. We can also use statistical

model to predict values of outlier observation and after that we can impute it with predicted values.

***FEATURE ENGINEERING**

Feature engineering is the science of extracting more information from existing data.

2 important steps in feature engineering are :

- Variable transformation.
- Variable / Feature creation.

Conditional Probability

Conditional probabilities arise naturally in the investigation of experiments where an outcome of a trial may affect the outcomes of the subsequent trials.

We try to calculate the probability of the second event (event B) given that the first event (event A) has already happened. If the probability of the event changes when we take the first event into consideration, we can safely say that the probability of event B is dependent of the occurrence of event A.

Let’s think of cases where this happens:

- Drawing a second ace from a deck given we got the first ace
- Finding the probability of having a disease given you were tested positive
- Finding the probability of liking Harry Potter given we know the person likes fiction

And so on….

Here we can define, 2 events:

- Event A is the probability of the event we’re trying to calculate.
- Event B is the condition that we know or the event that has happened.

We can write the conditional probability as $P\left(\frac{A}{B}\right)$, the probability of the occurrence of event A given that B has already happened.

$$P\left(\frac{A}{B}\right) = \frac{\text{P(A and B)}}{P(B)} = \frac{\text{Probability of the occurence of both A and B}}{\text{Probability of B}}$$

Let’s play a simple game of cards for you to understand this. Suppose you draw two cards from a deck and you win if you get a jack followed by an ace (without replacement). What is the probability of winning, given we know that you got a jack in the first turn?

Let event A be getting a jack in the first turn

Let event B be getting an ace in the second turn.

We need to find $P\left(\frac{B}{A}\right)$

P(A) = 4/52

P(B) = 4/51 {no replacement}

P(A and B) = 4/52*4/51= 0.006

$$P\left(\frac{B}{A}\right) = \frac{\text{P(A and B)}}{P(A)} = \frac{0.006}{0.077} = 0.078$$

Here we are determining the probabilities when we know some conditions instead of calculating random probabilities. Here we knew that he got a jack in the first turn.

Let’s take another example.

Suppose you have a jar containing 6 marbles – 3 black and 3 white. What is the probability of getting a black given the first one was black too.

P (A) = getting a black marble in the first turn

P (B) = getting a black marble in the second turn

P (A) = 3/6

P (B) = 2/5

P (A and B) = ½*2/5 = 1/5

$$P\left(\frac{B}{A}\right) = \frac{P(\text{A and B})}{P(A)} = \frac{0.2}{0.5} = 0.4$$

Bayes Theorem

The Bayes theorem describes the probability of an event based on the prior knowledge of the conditions that might be related to the event. If we know the conditional probability $P\left(\frac{A}{B}\right)$, we can use the bayes rule to find out the reverse probabilities $P\left(\frac{B}{A}\right)$. How can we do that?

$$P\left(\frac{A}{B}\right) = \frac{P\ (A \cap B)}{P\ (B)}$$

$$P\left(\frac{B}{A}\right) = \frac{P\ (A \cap B)}{P\ (A)}$$

$$P\ (A \cap B) = P\left(\frac{A}{B}\right) * P\ (B) = P\left(\frac{B}{A}\right) * P\ (A)$$

$$P\left(\frac{B}{A}\right) = P\left(\frac{A}{B}\right) * \frac{P(B)}{P(A)}$$

The above statement is the general representation of the Bayes rule.

For the previous example – if we now wish to calculate the probability of having a pizza for lunch provided you had a bagel for breakfast would be = 0.7 * 0.5/0.6.

We can generalize the formula further.

If multiple events A_i form an exhaustive set with another event B.

We can write the equation as

$$P\ (A_i/B) = \frac{P(B|Ai)*P(Ai)}{\sum(i=1\ to\ n)\ P(B|Ai)*P(Ai)}$$

Example of Bayes Theorem

Let’s take the example of the breast cancer patients. The patients were tested thrice before the oncologist concluded that they had cancer. The general belief is that 1.48 out of a 1000 people have breast cancer in the US at that particular time when this test was conducted. The patients were tested over multiple tests. Three sets of test were done and the patient was only diagnosed with cancer if she tested positive in all three of them.

Let’s examine the test in detail.

Sensitivity of the test (93%) – true positive Rate

Specificity of the test (99%) – true negative Rate

Let’s first compute the probability of having cancer given that the patient tested positive in the first test.

P (has cancer | first test +)

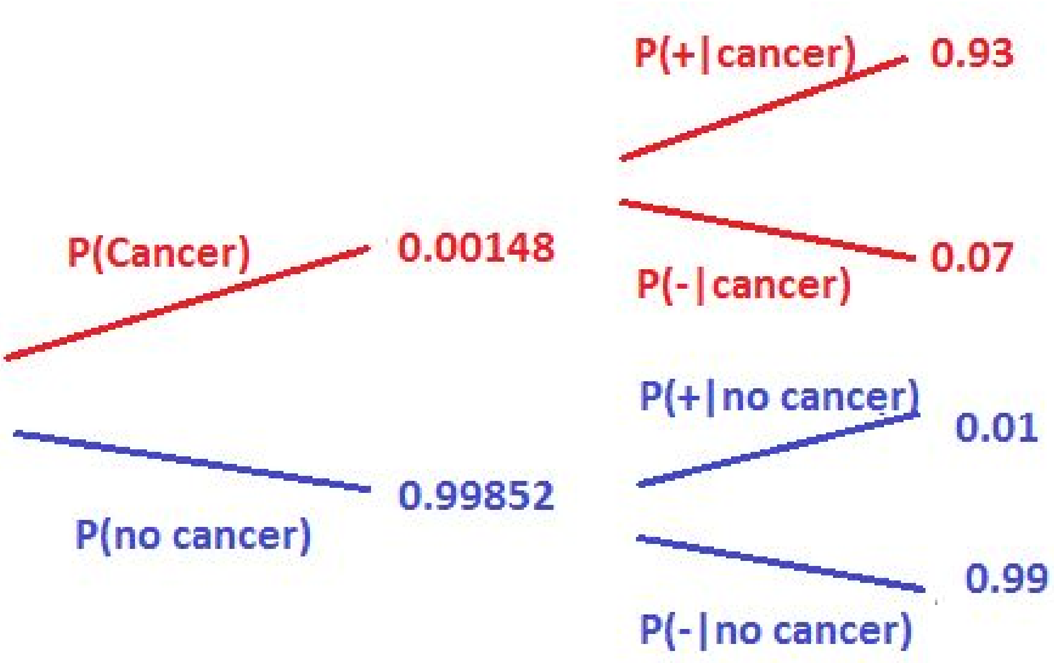
P (cancer) = 0.00148

Sensitivity can be denoted as P (+ | cancer) = 0.93

Specificity can be denoted as P (- | no cancer)

Since we do not have any other information, we believe that the patient is a randomly sampled individual. Hence our prior belief is that there is a 0.148% probability of the patient having cancer.

The complement is that there is a 100 – 0.148% chance that the patient does not have CANCER. Similarly we can draw the below tree to denote the probabilities.



Let’s not try to calculate the probability of having cancer given that he tested positive on the first test i.e. P (cancer|+)

$$\text{P (cancer}\,|\,+) = \frac{P(cancer\ and{+})}{P(+)}$$

P (cancer and +) = P (cancer) * P (+) = 0.00148*0.93
P (no cancer and +) = P (no cancer) * P(+) = 0.99852*0.01
To calculate the probability of testing positive, the person can have cancer and test positive or he may not have cancer and still test positive.

$$\text{P (CANCER}\,|\,+) = \frac{P(cancer\ and{+})}{P(cancer\ and{+}) + P(no\ cancer\ and{+})} = 0.12$$

This means that there is a 12% chance that the patient has cancer given he tested positive in the first test. This is known as the **posterior probability.**

TYPES OF PLOTS

Histograms

A histogram is very common plot. It plots the frequencies that data appears within certain ranges.

Dataset used in the examples

1. (w1)

https://docs.google.com/spreadsheets/d/1uYfR4Oi7FIgkjtyVgTI7uPYpDmBhIrtP6O2hGvLFjHM/edit?usp=sharing

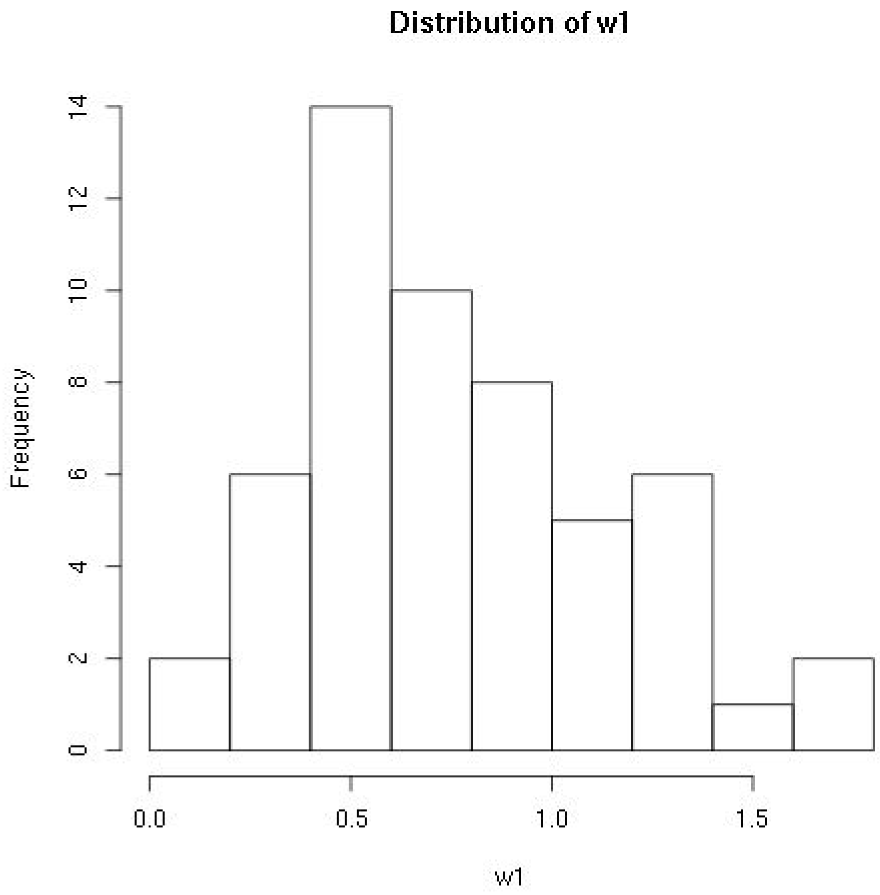
2. Trees

https://docs.google.com/spreadsheets/d/1WhtwoyDd7Xkkeou_NcQkBd_ygQ5ABdjdEgbF_cilKQM/edit?usp=sharing

Implementation in R

To plot a histogram of the data use the “hist” command:

```
> hist(w1$vals)
> hist(w1$vals,main="Distribution of w1",xlab="w1")
```



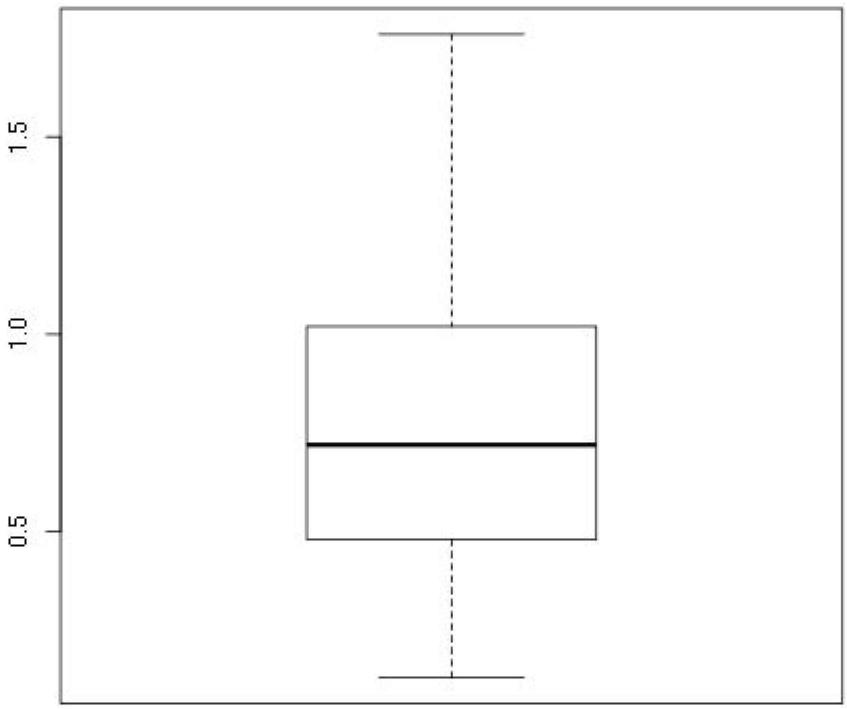
Boxplots

A boxplot provides a graphical view of the median, quartiles, maximum, and minimum of a data set.

Implementation in R

We first use the *w1* data set and look at the boxplot of this data set:

```
> boxplot(w1$vals)
```



Again, this is a very plain graph, and the title and labels can be specified in exactly the same way as in the stripchart and hist commands:

```
> boxplot(w1$vals,
        main='Leaf BioMass in High CO2 Environment',
        ylab='BioMass of Leaves')
```

Note that the default orientation is to plot the boxplot vertically. Because of this we used the ylab option to specify the axis label. There are a large number of options for this command. To see more of the options see the help page:

```
> help(boxplot)
```

Strip Charts

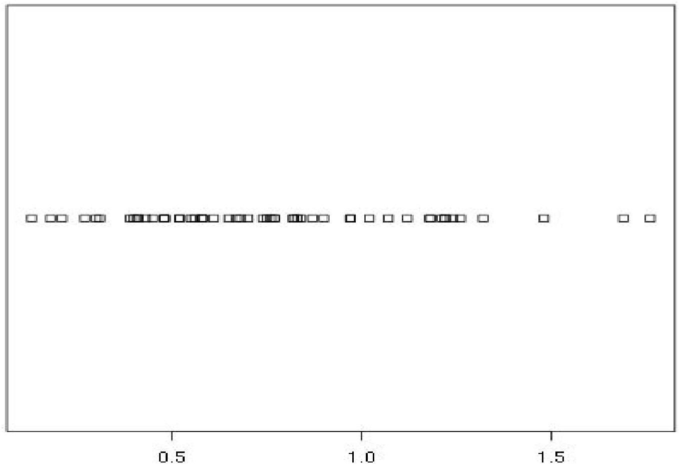
A strip chart is the most basic type of plot available. It plots the data in order along a line with each data point represented as a box

Implementation in R

. Here we provide examples using the *w1* data frame mentioned at the top of this page, and the one column of the data is *w1$vals*.

To create a strip chart of this data use the stripchart command:

```
> help(stripchart)
> stripchart(w1$vals)
```



Scatter Plots

A scatter plot provides a graphical view of the relationship between two sets of numbers. Here we provide examples using the *tree* data frame from the trees91.csv data file which is mentioned at the top of the page. In particular we look at the relationship between the stem biomass (“tree$STBM”) and the leaf biomass (“tree$LFBM”).

The command to plot each pair of points as an x-coordinate and a y-coordinate is “plot:”

```
> plot(tree$STBM,tree$LFBM)
```

you should always annotate your graphs. The title and labels can be specified in exactly the same way as with the other plotting commands:

```
> plot(tree$STBM,tree$LFBM,
      main="Relationship Between Stem and Leaf Biomass",
      xlab="Stem Biomass",
      ylab="Leaf Biomass")
```

Normal QQ Plots

The final type of plot that we look at is the normal quantile plot. This plot is used to determine if your data is close to being normally distributed. You cannot be sure that the data is normally distributed, but you can rule out if it is not normally distributed.

Implementation in R

Here we provide examples using the *w1* data frame mentioned at the top of this page, and the one column of data is *w1$vals*.

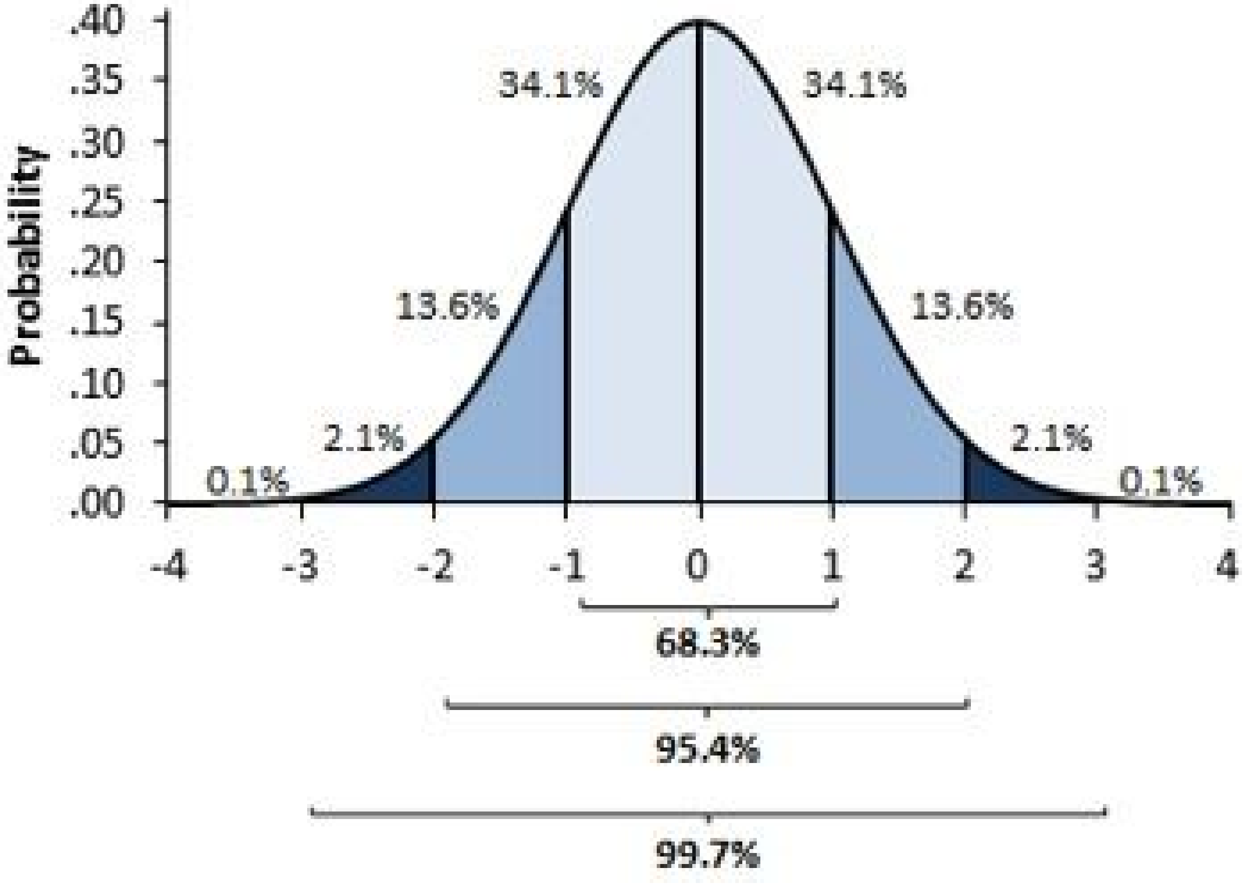
The command to generate a normal quantile plot is qqnorm. You can give it one argument, the univariate data set of interest:

```
> qqnorm(w1$vals)
```

You can annotate the plot in exactly the same way as all of the other plotting commands given here:

```
> qqnorm(w1$vals,
        main="Normal Q-Q Plot of the Leaf Biomass",
        xlab="Theoretical Quantiles of the Leaf Biomass",
        ylab="Sample Quantiles of the Leaf Biomass")
```

Gaussian Distribution



A general standard distribution of data which is often taken as a reference in many applications.

The normal distribution, also known as the Gaussian or standard normal distribution, is the probability distribution that plots all of its values in a symmetrical fashion, and most of the results are situated around the probability's mean. Values are equally likely to plot either above or below the mean. Grouping takes place at values close to the mean and then tails off symmetrically away from the mean.

Standard Deviation

It depicts how the whole data deviate from the central measures.

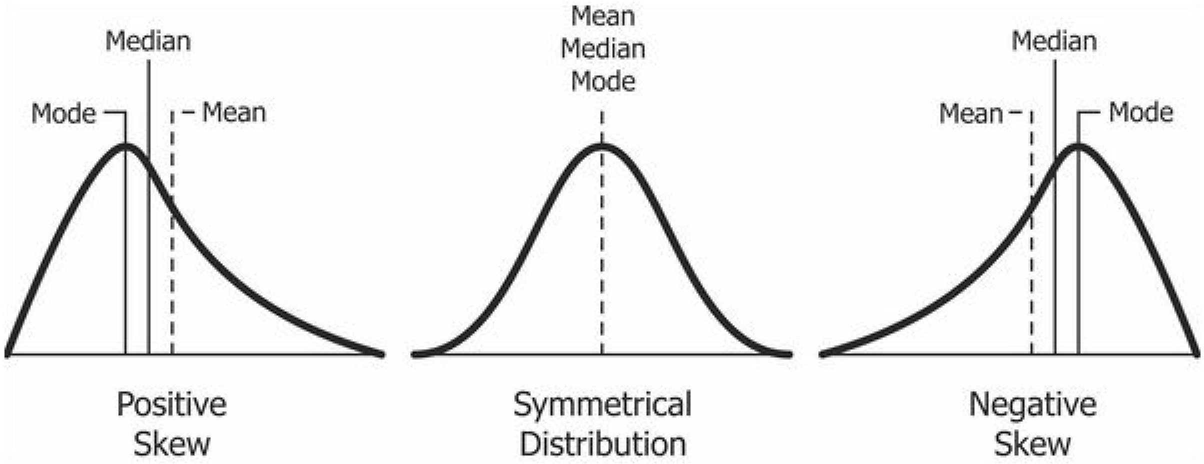
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

PROPERTIES

- ❏ Mean = Median = Mode
- ❏ Symmetry about the center
- ❏ 50% of the data is greater than mean, 50% of the data is less than mean
- ❏ 68% of the data is within 1 standard deviation (Likely)
- ❏ 95% of the data is within 2 standard deviation (Very Likely)
- ❏ 99.7% of the data is within 3 standard deviation (Almost certainly)

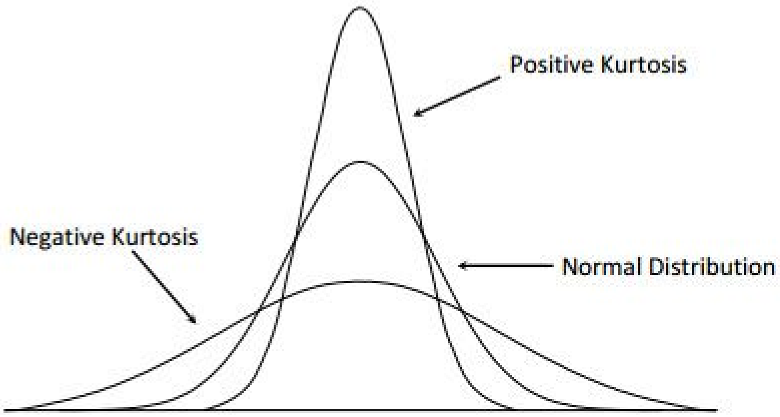
Skewness

Used to depict the asymmetry of the data from the general standard normal distribution.



Kurtosis

Measures how pointed is the peak relative to the normal distribution



SUGGESTED VIDEO
http://www.investopedia.com/terms/n/normaldistribution.asp