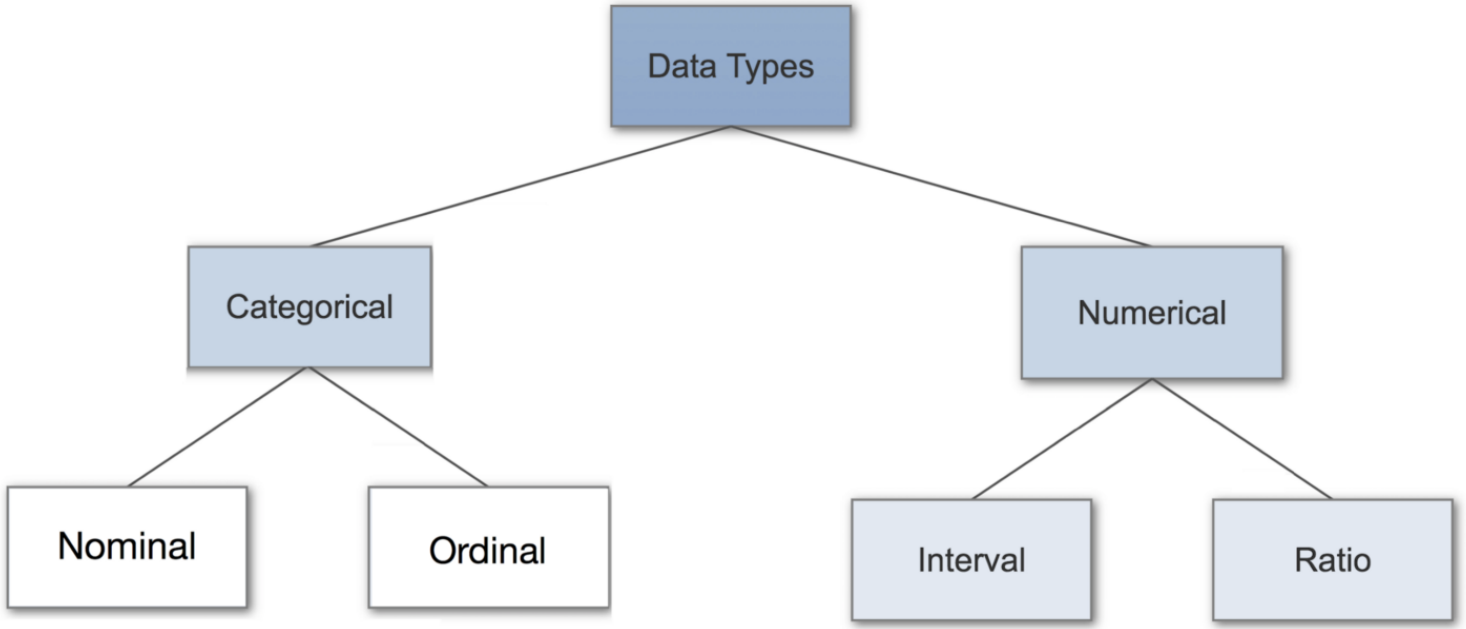
Fundamental to Machine Learning

Basics of Statistics

Definition: Science of collection, presentation, analysis, and reasonable interpretation of data.

Statistics presents a rigorous scientific method for gaining insight into data. For example, suppose we measure the weight of 100 patients in a study. With so many measurements, simply looking at the data fails to provide an informative account. However statistics can give an instant overall picture of data based on graphical presentation or numerical summarization irrespective to the number of data points. Besides data summarization, another important task of statistics is to make inference and predict relations of variables.

Types of data



Some Definitions

Variable - any characteristic of an individual or entity. A variable can take different values for different individuals. Variables can be *categorical* or *quantitative*. Per S. S. Stevens…

- **Nominal** - Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as „labels". Note that nominal data that has no order (e.g., gender, colors). Value may be a numerical, but without numerical value (e.g., I, II, III). The only operation that can be applied to Nominal variables is enumeration.
- **Ordinal** - Variables with an inherent rank or order, e.g. mild, moderate, severe. Can be compared for equality, or greater or less, but not *how much* greater or less. (e.g. rating of a material or application)
- **Interval** - Values of the variable are ordered as in Ordinal, and additionally, differences between values are meaningful, however, the scale is not absolutely anchored. Calendar dates and temperatures on the Fahrenheit scale are examples. Addition and subtraction, but not multiplication and division are meaningful operations.
- **Ratio** - Variables with all properties of Interval plus an absolute, non-arbitrary zero point, e.g. age, weight, temperature (Kelvin). Addition, subtraction, multiplication, and division are all meaningful operations.

Statistical Description of Data

- Statistics describes a numeric set of data by its
  - Center
  - Variability
  - Shape
- Statistics describes a categorical set of data by
  - Frequency, percentage or proportion of each category

Some Definitions

Distribution - (of a variable) tells us what values the variable takes and how often it takes these values.

- Unimodal - having a single peak
- Bimodal - having two distinct peaks
- Symmetric - left and right half are mirror images.

Frequency Distribution

Consider a data set of 26 children of ages 1-6 years. Then the frequency distribution of variable 'age' can be tabulated as follows:

Frequency Distribution of Age

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2

Grouped Frequency Distribution of Age:

Age Group	1-2	3-4	5-6
Frequency	8	12	6

Cumulative Frequency

Cumulative frequency of data in previous page

Age	1	2	3	4	5	6
Frequency	5	3	7	5	4	2
Cumulative Frequency	5	8	15	20	24	26

Age Group	1-2	3-4	5-6
Frequency	8	12	6
Cumulative Frequency	8	20	26

Data Presentation

Two types of statistical presentation of data - graphical and numerical.

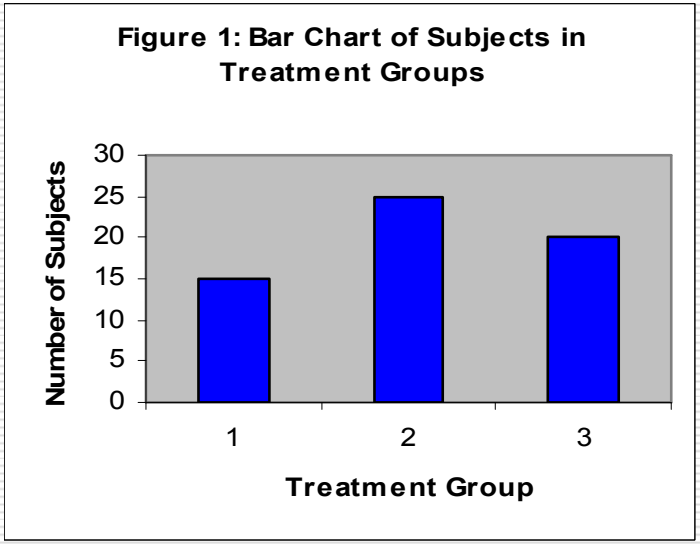
Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Over all pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an ***outlier***.

Bar diagram and Pie charts are used for categorical variables.

Histogram, stem and leaf and Box-plot , scatter plot are used for numerical variable.

Data Presentation –Categorical Variable

Bar Diagram: Lists the categories and presents the percent or count of individuals who fall in each category.

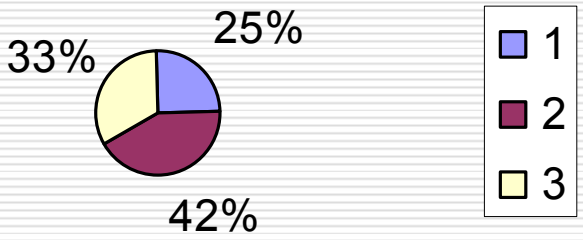


Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)=0.333	41.7
3	20	(20/60)=0.417	33.3
Total	60	1.00	100

Data Presentation –Categorical Variable

Pie Chart: Lists the categories and presents the percent or count of individuals who fall in each category.

Figure 2: Pie Chart of Subjects in Treatment Groups

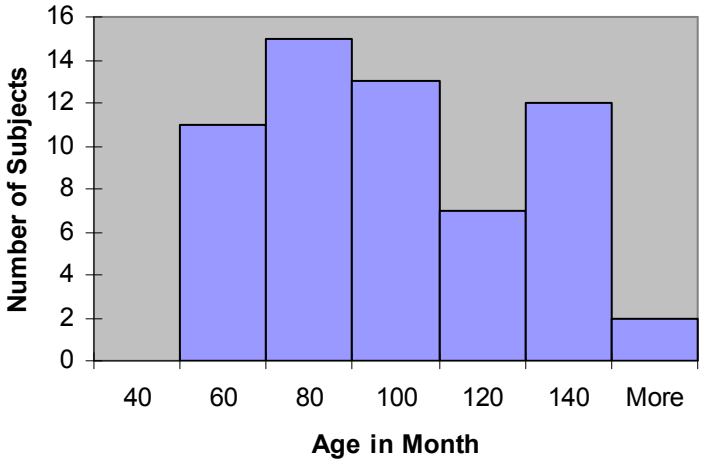


Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)=0.333	41.7
3	20	(20/60)=0.417	33.3
Total	60	1.00	100

Graphical Presentation –Numerical Variable

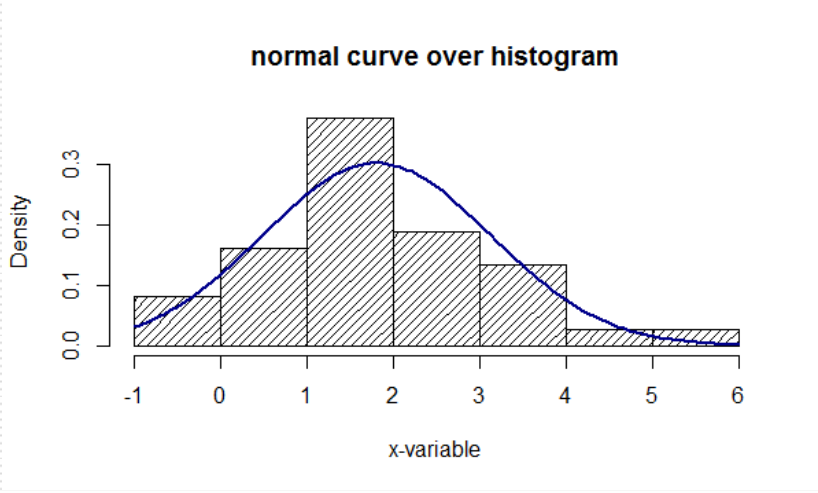
Histogram: Overall pattern can be described by its **shape**, **center**, and **spread**. The following age distribution is **right skewed**. The **center** lies between **80 to 100**. **No outliers**.

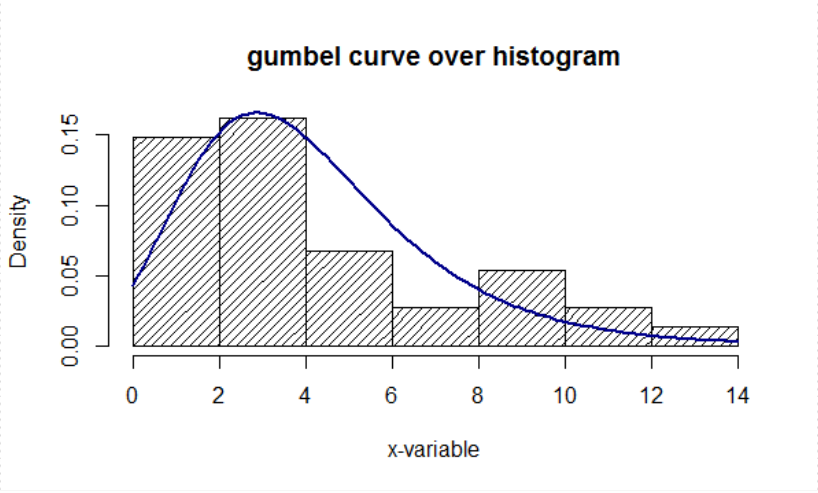
Figure 3: Age Distribution



Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60

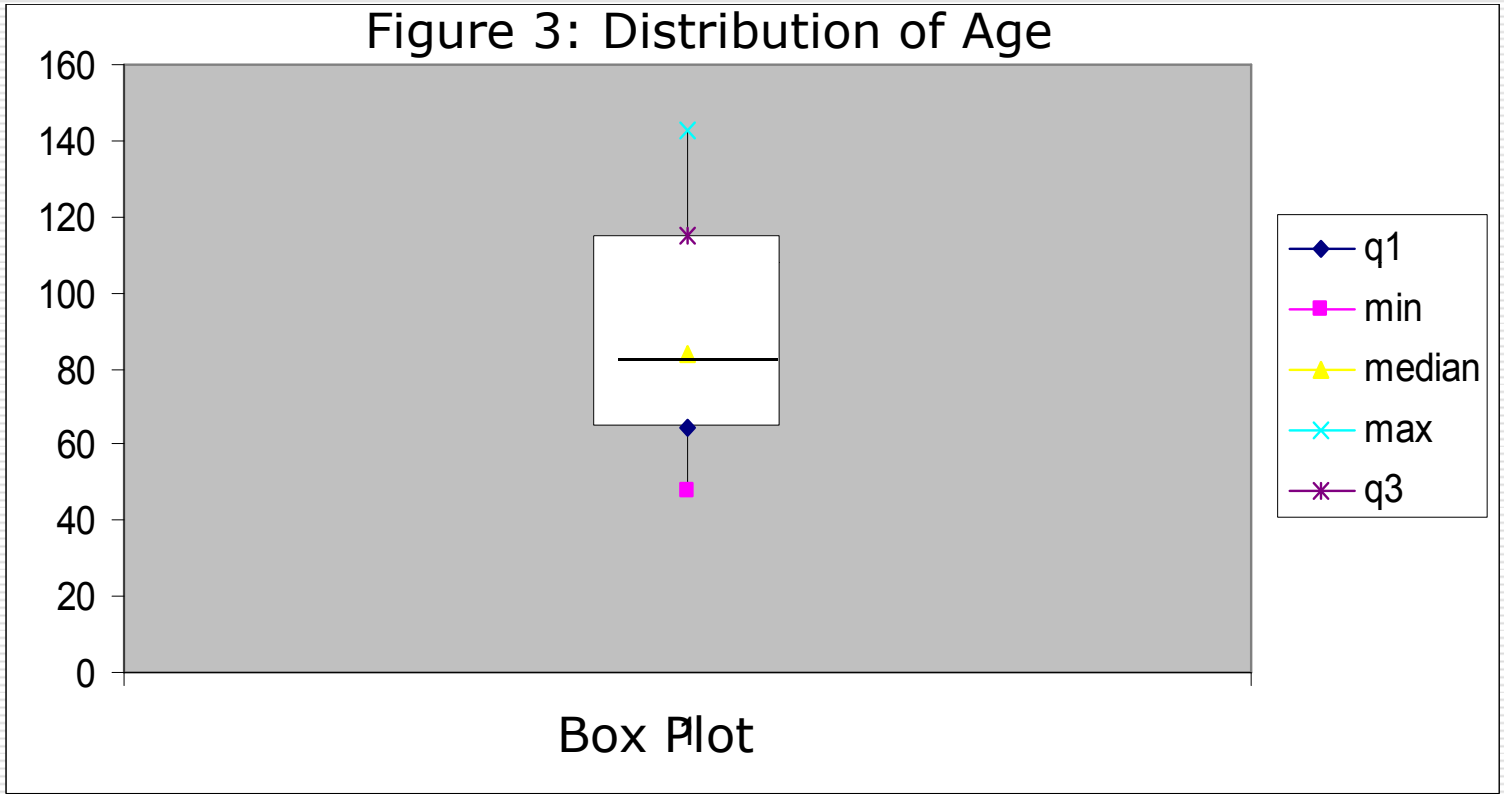
Curves over histogram to know approximate distribution





Graphical Presentation –Numerical Variable

Box-Plot: Describes the five-number summary

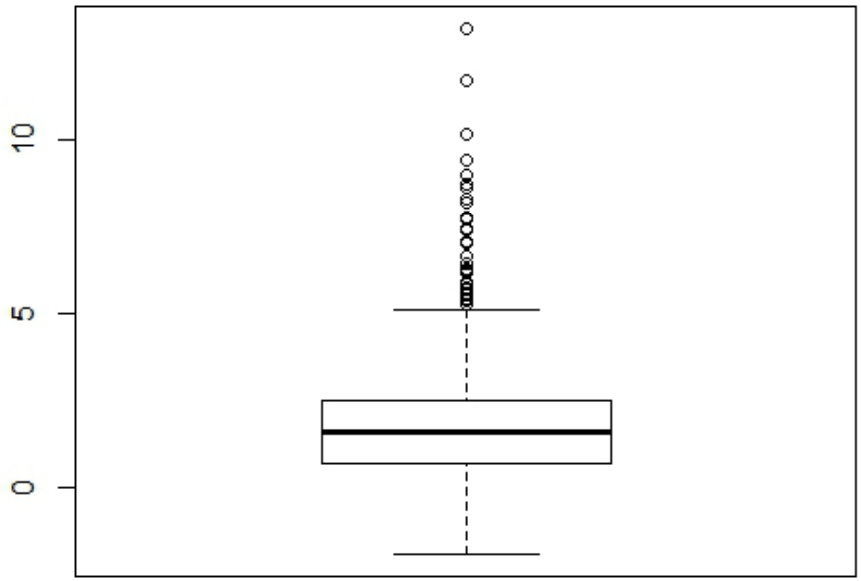


About boxplot

- Boxplot descibes some prominent features of data set:
  - Center
  - Spread
  - The extent and nature of any departure from symmetry
  - Identification of outliers
- Order the n observations from smallest to largest and separate the smallest half from the largest half: the median is included in both halves if n is odd. Then the lower fourth is the median of the smallest half and the upper fourth is the median of the largest half. A measure of spread that is resistant to outliers is the fourth spread f_s, given by

  f_s = upper fourth – lower fourth

  Any observation farther than $1.5f_s$ from the closest fourth is an outlier. An outlier is extreme if it is more than $3f_s$ from the nearest fourth, and it is mild otherwise.


0
5
10

Numerical Presentation

A fundamental concept in summary statistics is that of a *central value* for a set of observations and the extent to which the central value characterizes the whole set of data. Measures of central value such as the mean or median must be coupled with measures of data dispersion (e.g., average distance from the mean) to indicate how well the central value characterizes the data as a whole.

To understand how well a central value characterizes a set of observations, let us consider the following two sets of data:

A: 30, 50, 70

B: 40, 50, 60

The mean of both two data sets is 50. But, the distance of the observations from the mean in data set A is larger than in the data set B. Thus, the mean of data set B is a better representation of the data set than is the case for set A.

Methods of Center Measurement

Center measurement is a summary measure of the overall level of a dataset

Commonly used methods are **mean**, **median**, **mode**, **geometric mean** etc.

Mean: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is (20+30+40)/3 = 30.

Notation : Let $x_1, x_{2,} \ldots x_n$ are n observations of a variable x. Then the mean of this variable,

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Methods of Center Measurement

Median: The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value. In case of an even number of observations the average of the two middle most values is the median. For example, to find the median of {9, 3, 6, 7, 5}, we first sort the data giving {3, 5, 6, 7, 9}, then choose the middle value 6. If the number of observations is even, e.g., {9, 3, 6, 7, 5, 2}, then the median is the average of the two middle values from the sorted sequence, in this case, (5 + 6) / 2 = 5.5.

Mode: The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated.

Mean or Median

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income. For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270. The median of these four observations is (30+40)/2 =35. Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

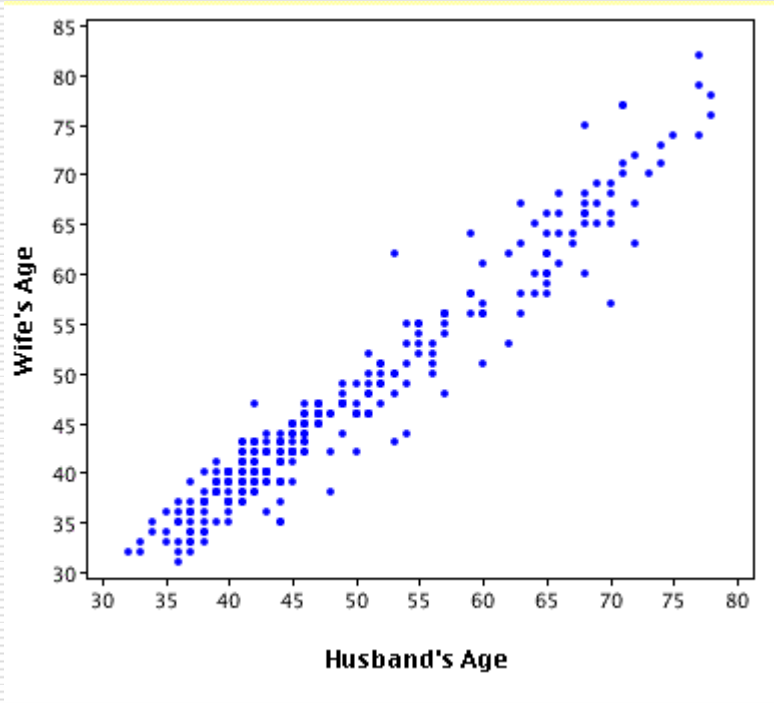
Methods of Variability Measurement

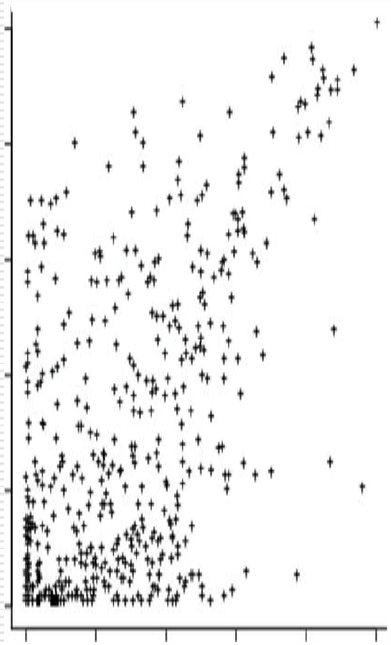
Variability (or dispersion) measures the amount of scatter in a dataset.

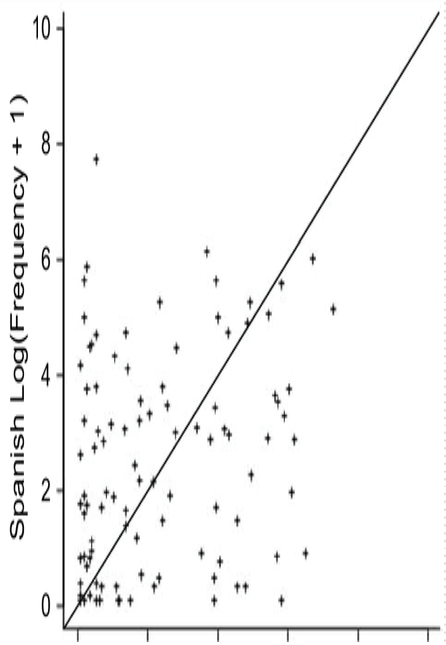
Commonly used methods: *range*, *variance*, *standard deviation*, *interquartile range*, *coefficient of variation etc*.

Range: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is (100-2)=98. It's a crude measure of variability.

Plots showing variability







Methods of Variability Measurement

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations x_1, x_2,…x_n is

$$S^2 = \frac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$$

Variance of 5, 7, 3? Mean is (5+7+3)/3 = 5 and the variance is

$$\frac{(5-5)^2 + (3-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation: Square root of the variance. The standard deviation of the above example is 2.

Methods of Variability Measurement

Quartiles: Data can be divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.

In notations, quartiles of a data is the $((n+1)/4)q^{th}$ observation of the data, where q is the desired quartile and n is the number of observations of data.

The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25^{th} and 50^{th} percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

Methods of Variability Measurement

In the following example Q1= ((15+1)/4)1 =4th observation of the data. The 4th observation is 11. So Q1 is of this data is 11.

An example with 15 numbers

3 6 7 11 13 22 30 40 44 50 52 61 68 80 94

Q1 Q2 Q3

The first quartile is Q1=11. The second quartile is Q2=40 (This is also the Median.) The third quartile is Q3=61.

Inter-quartile Range: Difference between Q3 and Q1. Inter-quartile range of the previous example is 61- 40=21. The middle half of the ordered data lie between 40 and 61.

Deciles and Percentiles

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25^{th} percentile is the Q1, 50^{th} percentile is the Median (Q2) and the 75^{th} percentile of the data is Q3.

In notations, percentiles of a data is the ((n+1)/100)p th observation of the data, where p is the desired percentile and n is the number of observations of data.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

Coefficient of Variation = $\frac{\sigma}{\bar{x}} \times 100$

Five Number Summary

Five Number Summary: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1), The median(Q2), the third quartile, and the largest (Maximum) observation written in order from smallest to largest.

Box Plot: A box plot is a graph of the five number summary. The central box spans the quartiles. A line within the box marks the median. Lines extending above and below the box mark the smallest and the largest observations (i.e., the range). Outlying samples may be additionally plotted outside the range.

Choosing a Summary

The five number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with extreme outliers. The mean and standard deviation are reasonable for symmetric distributions that are free of outliers.

In real life we can't always expect symmetry of the data. It's a common practice to include number of observations (n), mean, median, standard deviation, and range as common for data summarization purpose. We can include other summary statistics like Q1, Q3, Coefficient of variation if it is considered to be important for describing data.

Shape of Data

- Shape of data is measured by
  - Skewness
  - Kurtosis

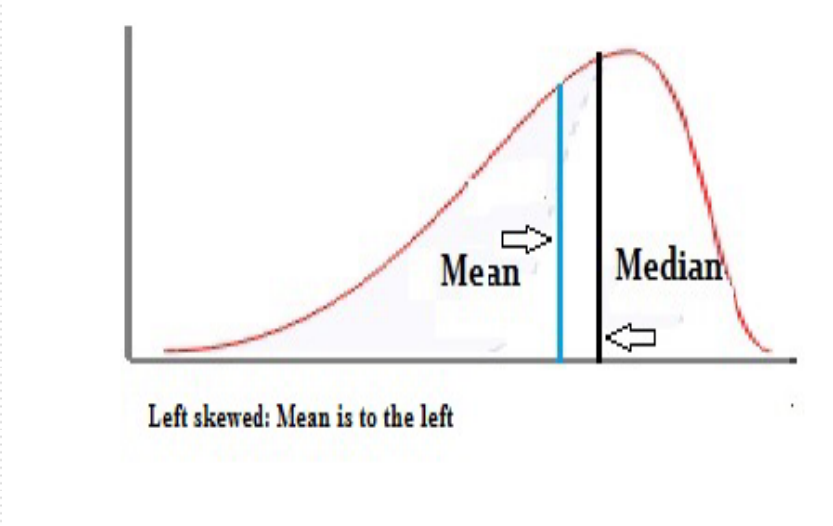
Skewness

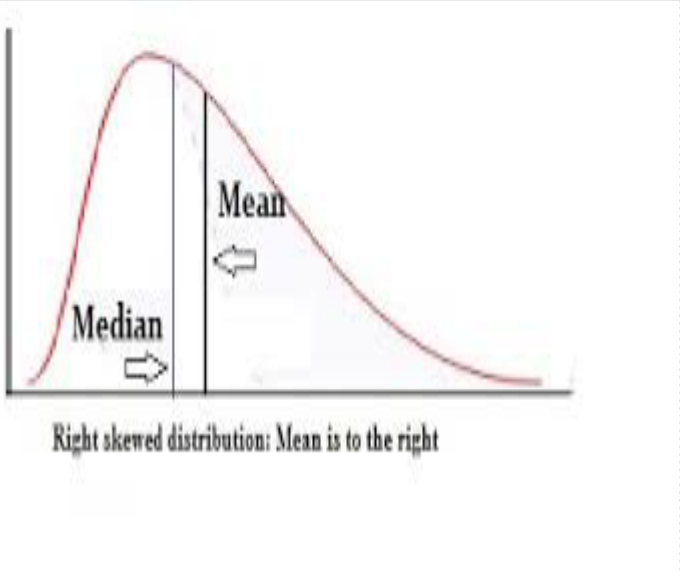
- Measures asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail

Let $x_1, x_2, \ldots x_n$ be n observations. Then,

$$\text{Skewness} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

- Left skewed plot
- Right skewed plot





Kurtosis

- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

Let $x_1, x_2, \ldots x_n$ be n observations. Then,

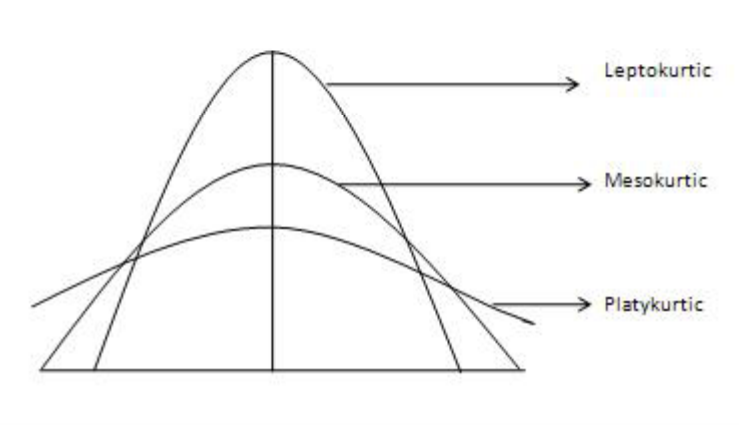
$$\text{Kurtosis} = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

- Types of kurtosis

i. Leptokurtic

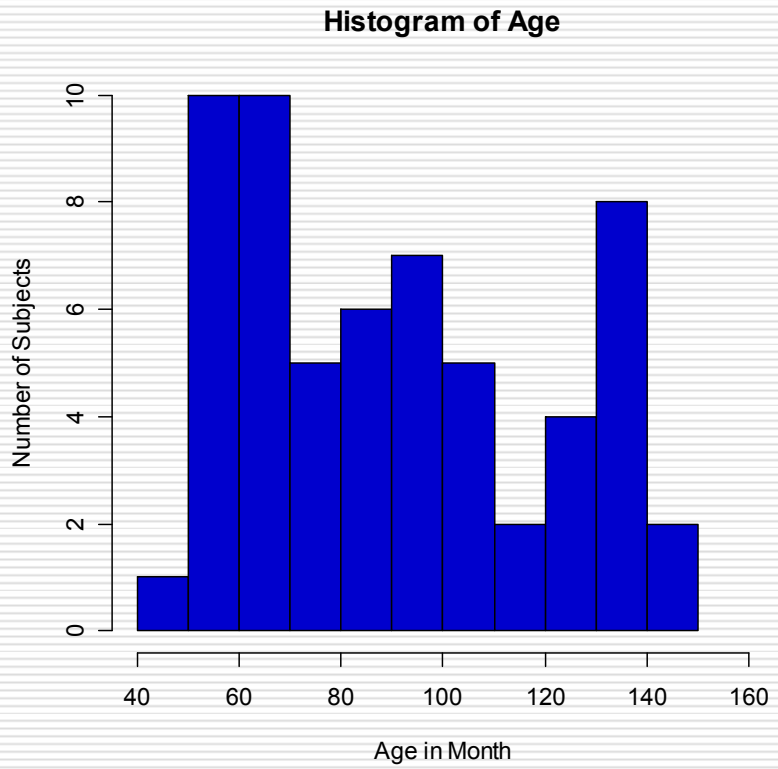
ii. Mesokurtic

iii. Platykurtic

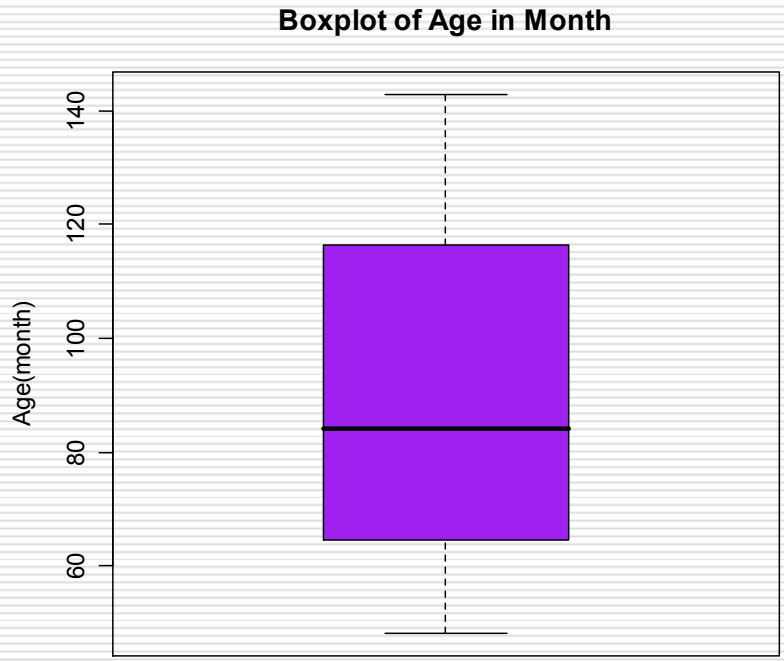


Summary of the Variable 'Age' in the given data set

Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60



Summary of the Variable ‘Age’ in the given data set



Class Summary (First Part)

So far we have learned-

Statistics and data presentation/data summarization

Graphical Presentation: Bar Chart, Pie Chart, Histogram, and Box Plot

Numerical Presentation: Measuring Central value of data (mean, median, mode etc.), measuring dispersion (standard deviation, variance, co-efficient of variation, range, inter-quartile range etc), quartiles, percentiles, and five number summary

Any questions ?

Brief concept of Statistical Softwares

There are many softwares to perform statistical analysis and visualization of data. Some of them are SAS (System for Statistical Analysis), S-plus, R, Matlab, Minitab, BMDP, Stata, SPSS, StatXact, Statistica, LISREL, JMP, GLIM, HIL, MS Excel etc. We will discuss MS Excel and SPSS in brief.

Some useful websites for more information of statistical softwares-

http://www.galaxy.gmu.edu/papers/astr1.html

http://ourworld.compuserve.com/homepages/Rainer_Wuerlaender/statsoft.htm#archiv

http://www.R-project.org