Introductory Statistics

Descriptive statistics is often referred to as "exploratory data analysis". Steps involved in the analysis are:
(i) constructing a **table** from the raw data,
(ii) constructing **graphs** using the information from the table,
(iii) computing **numerical summaries** for the raw data.

Categorical data

Graphs

Pie Chart is a circle divided into pieces ("slice of the pie") according to the number of categories. Each "slice" size is proportional to corresponding category relative frequency.

Bar graph displays a vertical bar for each category and the height of the bar corresponds to category relative frequency (or **percentage**).

Numerical summary for categorical variables is **proportion** (or **percentage**) of specific category.

Numerical (Quantitative) variables

Graphs

Dot plot shows a dot for each observation. Dots are placed just above the value on the number line for that observation.

Stem-and-leaf plot displays each observation in two parts: stem (all the digits except the last one) and leaf (the last digit).

Histogram displays a vertical bar for each class or value of the variable. The height of the bar corresponds to class (or value) relative frequency or **percentage**.

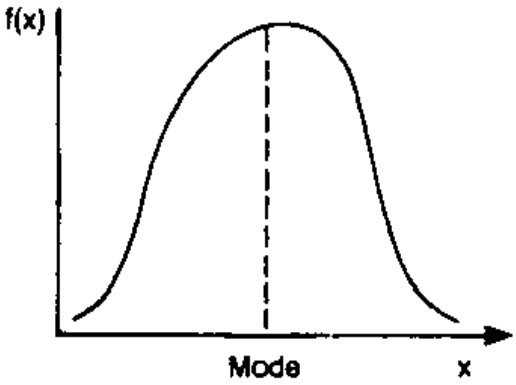
Box plot displays 5 summery numbers of an ordered (lowest to highest) data set on a vertical or horizontal axis. These 5 numbers are: minimum value, 25^{th} percentile point, median, 75^{th} percentile point, and the maximum value. The middle 50% of the data is represented by a box bounded by 25^{th} and 75^{th} percentile points on the sides.

Distribution is the collection of data in the form of a frequency table or graphical form.

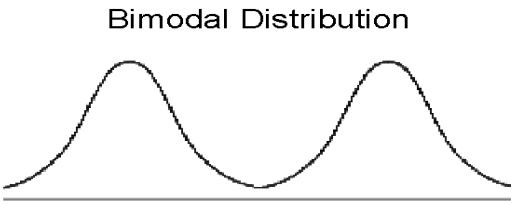
Distribution shapes:

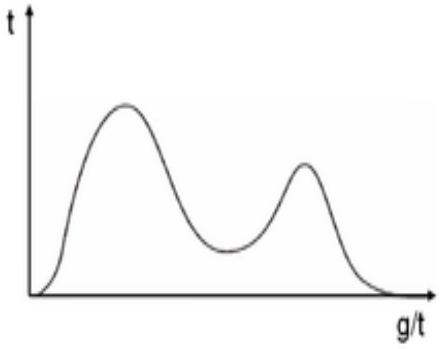
Distribution shape refers the shape of a smooth curve super imposed on top of a histogram.

Unimodal distribution has only one mode.

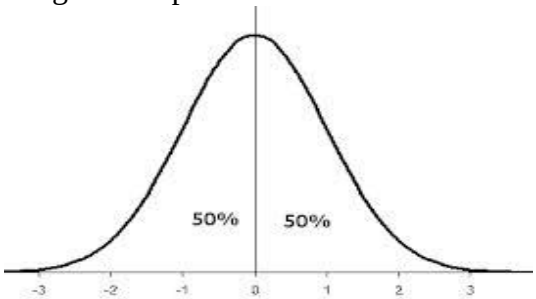


Bimodal distribution has two modes. Multimodal distribution has more than 2 modes.





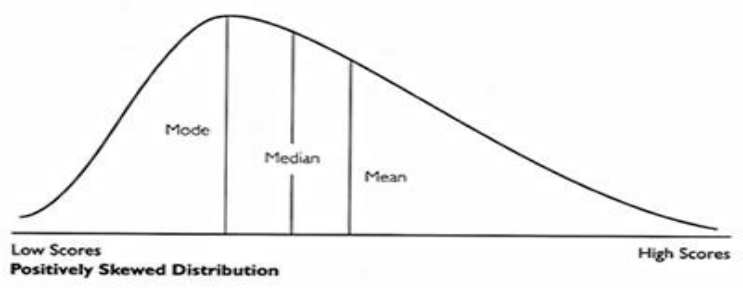
Symmetric (or bell-shaped) distribution has only one mode and can be divided into two equal parts at the center of the distribution. The part below the center value is a mirror image of the part above the center value.



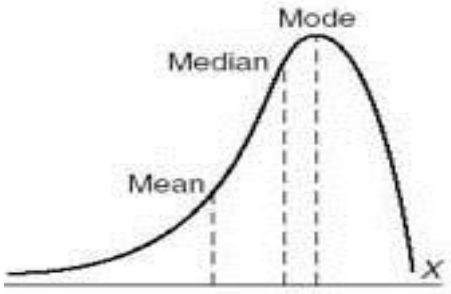
For the distribution above, Mean = Median = Mode = 0.

Asymmetric or skewed distribution is the type of distributions in which one side of the distribution stretches out longer than the other side.

Right skewed distribution Mean $\geq$ Median $\geq$ Mode



Left skewed distribution Mean $\leq$ Median $\leq$ Mode



Measures of center for quantitative data:

Mean is the most common and widely used measure of center for quantitative variables. Mean of a variable is the sum of all observations or values of that variable divided by the number of observations.

Mean = $\bar{y} = \frac{\sum y}{n}$, where *y* represents the variable and $\bar{y}$ (*y* bar) is sample mean of variable *y* and *n* represents the total number of observations and the symbol $\sum$ is for "sum".

Median is also a popular measure of center for quantitative variables. Median is the midpoint of an ordered data set, half of the values are smaller than median and the other half is larger than the median.

How to find median:

1. The observations need to be ordered from lowest to highest.
2. If the number of observations, *n*, is odd then median is the middle observation in the ordered sample.
3. If *n* is even, then median is the average of two middle observations in the ordered sample.

Mode is another measure of center for quantitative variables. Mode is the most frequently occurred value, in other words, mode is the value with the highest frequency.

Outlier: An observation that is unusually small or large compare to the overall bulk of the data.

Comparison between mean and median

Since mean is computed by adding all the values, it is highly influenced by the presence of any outlier in the data. Mean is pulled in the direction of the outlier. On the hand, outlier has no effect on median. It is the midpoint of the ordered data, median depends on the number of observations. That is why median is known as a "**robust**" or resistant measure of the center.
For a perfectly **symmetric** distribution, mean and median are **equal**. For a right skewed distribution mean is usually larger than the median, because extreme high values on the right side makes the mean large compare to median. For a left skewed distribution, mean is usually smaller than the median, because extreme small values on the left side makes the mean small compare to median.

Measures of spread for quantitative data:

Range is the difference between the largest and the smallest observations. Range is easy to compute and also easy to understand. But it uses information from only two (minimum and maximum) values and ignores all other values. Also range is highly affected by outliers in the data set.

Standard deviation is the average deviation of the observations from their mean. The deviation of each observation from the mean is denoted by: $(y - \bar{y})$. Sum of the deviations is equal to zero, $\sum(y - \bar{y}) = 0$. That is why we need to use the squared deviation $(y - \bar{y})^2$ when we are computing standard deviation. The average of all the squared deviations is called **variance**. The square root of the variance is the **standard deviation**.

The computing formula for standard deviation is, $s = \sqrt{\frac{\sum(y - \bar{y})^2}{n-1}}$ = 10.30 (*verify!*)

Exam scores: 62 72 76 85 91 82

Range: Max – Min =

Mean = $\bar{y}$ = 78 (*verify!*)

y	$(y - \bar{y})$	$(y - \bar{y})^2$
60		
72		
76		
85		
91		
82		
$\sum y$	$\sum(y - \bar{y})$	$\sum(y - \bar{y})^2$

Percentile: The ***p***th percentile is a value such that *p* percent of the observations fall below or at that value. For example in a data set, if the value corresponding to ***p*** = 50 is 79, then is the 50th percentile value is 79 and 50% of the data is below or at 79.

Important percentiles:
p = 25, the 25th percentile value is known as first quartile (Q1).
p = 50, the 50th percentile value is known as second quartile (Q2).

p = 75, the 75th percentile value is known as third quartile (Q3).

Example:
16 68 44 50 23 28 63 41 55 46 53 43 42 44 50 54 46 31 48

Ordered data:
16 23 28 31 41 42 43 44 44 46 46 48 50 50 53 54 55 63 68

Q2 = median = 46
Q1 = 1st quartile = 41
Q3 = 3rd quartile = 53

Inter-quartile range (IQR)

IQR = Q3 – Q1 = 12

Detecting potential outliers:

Q1 – 1.5 (IQR) = 23 [lower boundary for potential outliers]

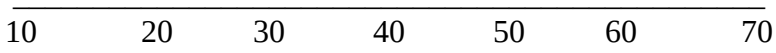
Q3 + 1.5 (IQR) = 71 [upper boundary for potential outliers]

16 is a potential outlier in this data set! This point should be identified by an asterisk on the box plot.

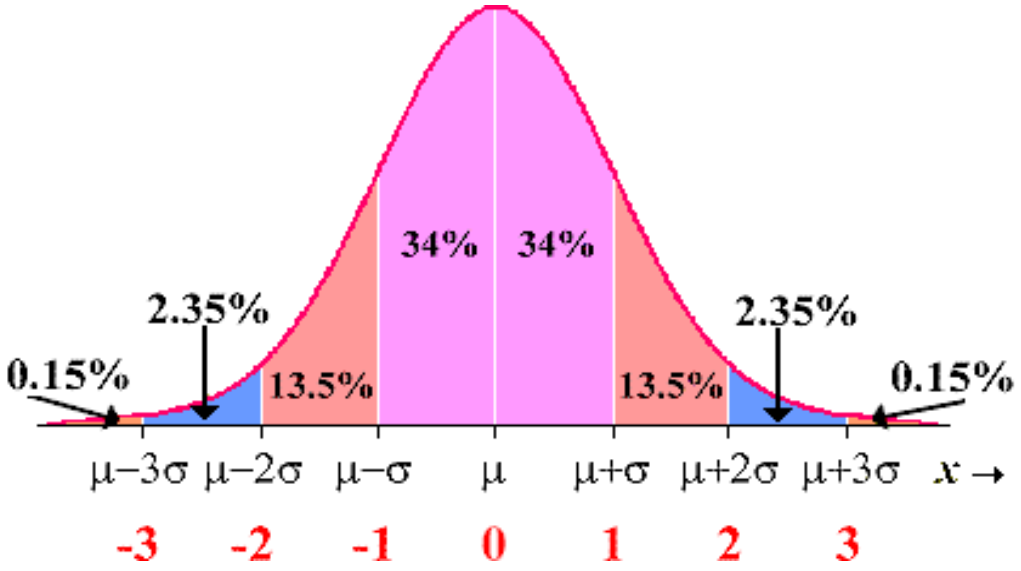
5 number summary plot (Box plot)

Minimum value:
First quartile Q1:
Second quartile Q2:
Third quartile Q3:
Maximum value:
Outliers:

Draw the box-plot using the example data.



Empirical Rule (applicable only for bell-shaped or symmetric data):



Source: oswego.edu

Approximately **68%** of the observations fall within 1 standard deviation of the mean.
Approximately **95%** of the observations fall within 2 standard deviations of the mean.
Almost all **(99.7%)** of the observations fall within 3 standard deviations of the mean.

	Characteristic	Symbol
Population	Variance	σ^2
Population	Standard deviation	σ
Sample	Variance	s^2
Sample	Standard deviation	s