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

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.

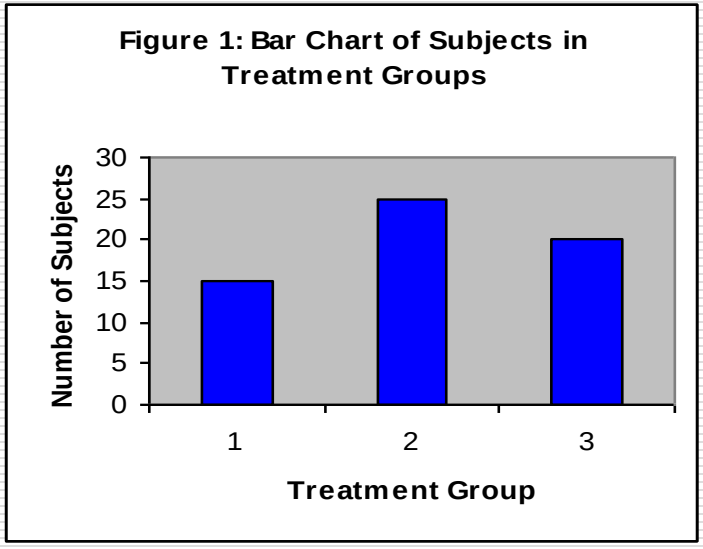
Graphical Presentation: We look for the overall pattern and for striking deviations from that pattern. Over all pattern usually described by shape, center, and spread of the data. An individual value that falls outside the overall pattern is called an ***outlier***.

Bar diagram and Pie charts are used for categorical variables.

Histogram, stem and leaf and Box-plot 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.

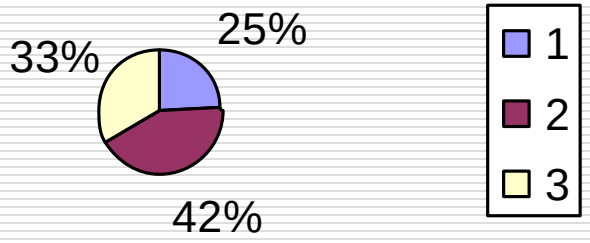


Treatment Group	Frequency	Proportion	Percent (%)
1	15	(15/60)=0.25	25.0
2	25	(25/60)=0.333	41.7
3	20	(20/60)=0.417	33.3
Total	60	1.00	100

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

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.

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.

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

Coefficient of Variation = $s/\bar{x} * 100$

$= 2/5 * 100$

$= 4\%$