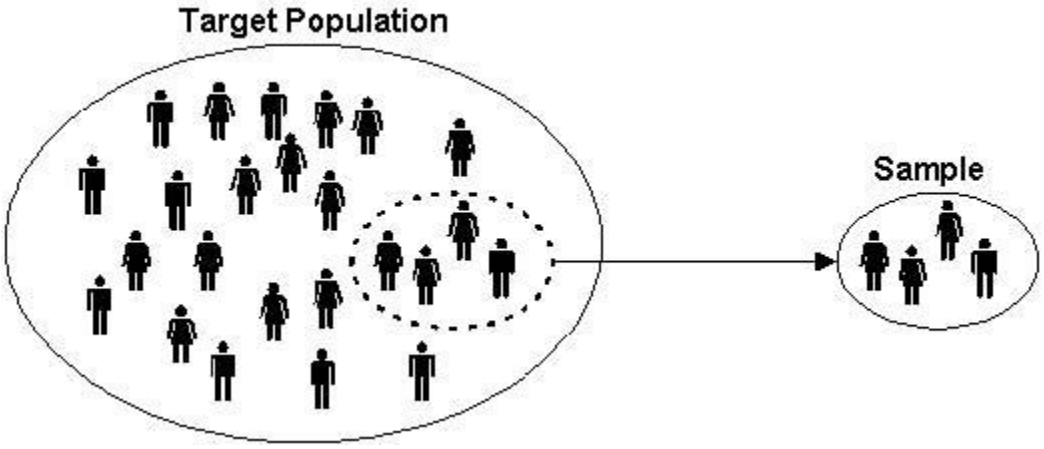
Statistics is the study of how to collect, organize, analyze, and interpret numerical information from data.

population the total set of individuals that we are interested about and a sample a subset of the individuals selected in a prescribed manner of study.



- **Observation**: Result from one trial of an experiment.
- **Sample**: Group of results gathered from separate independent trials.
- **Population**: Space of all possible observations that could be seen from a trial.

We can classify data into two types:

1. Numerical or Quantitative data is data where the observations are numbers. For example, age, height, on a scale from one to ten..., distance, number of,...

2. Categorical or Qualitative data is data where the observations are non-numerical. For example, favorite color, choice of politician, ...

Data is called univariate if it represents one attribute and bivariate if it contains two attributes.

Bivariate data is often used to compare.

Numerical data is called discrete if the number of possible values within every bounded range is finite. Examples include rolling dice, number of times that..., ...

Otherwise, numerical data is called continuous. For example, height, weight, temperature, distance, ...

Data Analysis

With data analysis, two main statistical methods- *Descriptive* and *Inferential*.

- **Descriptive statistics** uses tools like mean and standard deviation on a sample to summarize data.
- **Inferential statistics**, on the other hand, looks at data that can randomly vary, and then draw conclusions from it.
- **Predictive statistics**
- **Prescriptive statistics**

Measure of central tendency is a value that represents a typical, or central, entry of a data set.

The sample mean is the average and is computed as the sum of all the observed outcomes from the sample divided by the total number of events.

We use x as the symbol for the sample mean. In math terms,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x$$

where n is the sample size and the x correspond to the observed valued.

34, 43, 81, 106, 106 and 115

We compute the sample mean by adding and dividing by the number of samples, 6.

$$\frac{34 + 43 + 81 + 106 + 106 + 115}{6} = 80.83$$

The mode of a set of data is the number with the highest frequency.

In the above example 106 is the mode, since it occurs twice, and the rest of the outcomes occur only once.

The mean will be strongly affected by this outcome. Such an outcome is called and outlier.

An alternative measure is the median. The median is the middle score.

The value that lies in the **middle** of the data when the data set is **ordered.**

If the data set has an odd number of entries, then the median is the middle data entry.

If the data has an even number of entries, then the median is obtained by adding the two numbers in the middle and dividing result by two.

Outliers are not just greatest and least values, but values that are very different from the pattern established by the rest of the data. Outliers affect the mean.

When outliers are present it is best to use the median as the measure of central tendency.

If we have an even number of events, we take the average of the two middles.

2.7, 2.9, 3.1, 3.4, 3.7, 4.1, 4.3, 4.7, 4.7, 40.8

Since there is an even number of outcomes, we take the average of the middle two

$$\frac{3.7 + 4.1}{2} = 3.9$$

Example: -

44, 50, 38, 96, 42, 47, 40, 39, 46, 50.

To find the sample mean, add them and divide by 10:

$$\frac{44 + 50 + 38 + 96 + 42 + 47 + 40 + 39 + 46 + 50}{10} = 49.2$$

To find the median, first sort the data:

38, 39, 40, 42, 44, 46, 47, 50, 50, 96

Notice that there are two middle numbers 44 and 46.

To find the median we take the average of the two.

$$\text{Median} = \frac{44 + 46}{2} = 45$$

Notice also that the mean is larger than all but three of the data points.

The mean is influenced by outliers while the median is robust.

The mean, mode, median, and trimmed mean do a nice job in telling where the center of the data set is,

Measures of Variation:

• **Range:** The difference between the maximum and minimum data entries in the set.

Range = (Max. data entry) – (Min. data entry)

The **standard deviation** measure **variability** and **consistency** of the sample or population.
In most real-world applications, consistency is a great advantage.
In statistical data analysis, <u>less variation</u> is often better.

We define the variance to be

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2$$

and the standard deviation to be

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2}$$

Variance and Standard Deviation: Step by Step

1. Calculate the mean, x.
2. Write a table that subtracts the mean from each observed value.
3. Square each of the differences.
4. Add this column.
5. Divide by n -1 where n is the number of items in the sample This is the *variance*.
6. To get the *standard deviation* we take the square root of the variance.

Example

following data.

44, 50, 38, 96, 42, 47, 40, 39, 46, 50

He calculated the mean by adding and dividing by 10 to get

$$\bar{x} = 49.2$$

x	x - 49.2	$(x - 49.2)^2$
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24
42	-7.2	51.84
47	-2.2	4.84
40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64
Total		**2600.4**

Now

$$\frac{2600.4}{10 - 1} = 288.7$$

Hence the variance is 289 and the standard deviation is the square root of 289 = 17.

Since the standard deviation can be thought of measuring how far the data values lie from the mean,

we take the mean and move one standard deviation in either direction.

The mean for this example was about 49.2 and the standard deviation was 17. We have:

$$49.2 - 17 = 32.2$$

and

$$49.2 + 17 = 66.2$$

- **Percentiles**: Let p be any integer between 0 and 100. The pth percentile of data set is the data value at which p percent of the value in the data set are less than or equal to this value.
- **How to calculate percentiles**: Use the following steps for calculating percentiles for small data sets. For large data sets, we use computers to find percentiles.
  - Step 1: Sort the data in ascending order (from smallest to largest)
  - Step 2: Calculate $i^{th} = \left(\frac{p}{100}\right) n$, where p is the particular percentile you wish to calculate and n is the sample size.
  - Step 3: If i is an integer, the pth percentile is the mean of the data values in positions i and $i + 1$. If i is not an integer, then round up to the next integer and use the value in this position.

Example: Use the following set of stock prices (in dollars): 10, 7, 20, 12, 5, 15, 9, 18, 4, 12, 8, 14 Find the 10th percentile and the 50th percentile

Solutions:

- First sort the data in ascending order: **4, 5, 7, 8, 9, 10, 12, 12, 14, 15, 18, 20**
- There are 12 scores so, $n = 12$.
- To find the 10th percentile, we use the formula

$$i^{th} = \left(\frac{p}{100}\right)n = \left(\frac{10}{100}\right)12 = 1.2 \approx \text{Round Up}(1.2) = 2$$

- The 10th percentile is the number in the 2nd position.

Position	1st	**2nd**	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
Data	4	**5**	7	8	9	10	12	12	14	15	18	20

$$10\text{th Percentile} = P_{10} = 5$$

- To find the 50th percentile, we use the formula

$$i^{th} = \left(\frac{p}{100}\right)n = \left(\frac{50}{100}\right)12 = 6$$

- We need to find the 6th and 7th numbers in the sorted data set.
- Since the answer is an integer, we need to find the 6th and 7th number in the data set.

Position	1st	2nd	3rd	4th	5th	**6th**	**7th**	8th	9th	10th	11th	12th
Data	4	5	7	8	9	**10**	**12**	12	14	15	18	20

$$50\text{th Percentile} = P_{50} = \frac{10+12}{2} = 11$$

The frequency of an event is the number of times that the event occurs.

The relative frequency is the proportion of observed responses in the category.

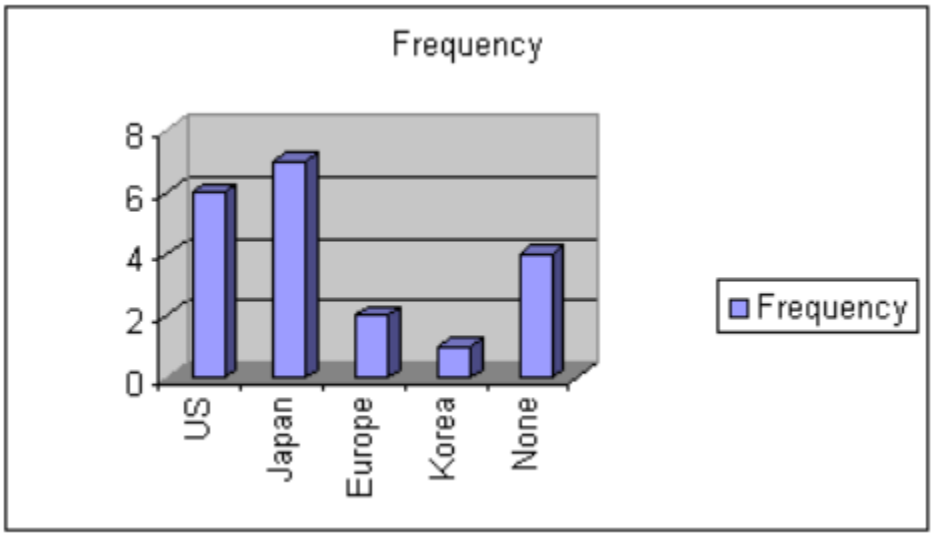
The relative frequency is computed by dividing the frequency by the total number of respondents.

Country	Frequency	Relative Frequency
US	6	0.3
Japan	7	0.35
Europe	2	0.1
Korea	1	0.05
None	4	0.2
Total	20	1

Bar Graph:

This bar graph is called a Pareto chart since the height represents the frequency.

Notice that the widths of the bars are always the same.



Histograms

Histograms are bar graphs whose vertical coordinate is the frequency count and whose horizontal coordinate corresponds to a numerical interval.

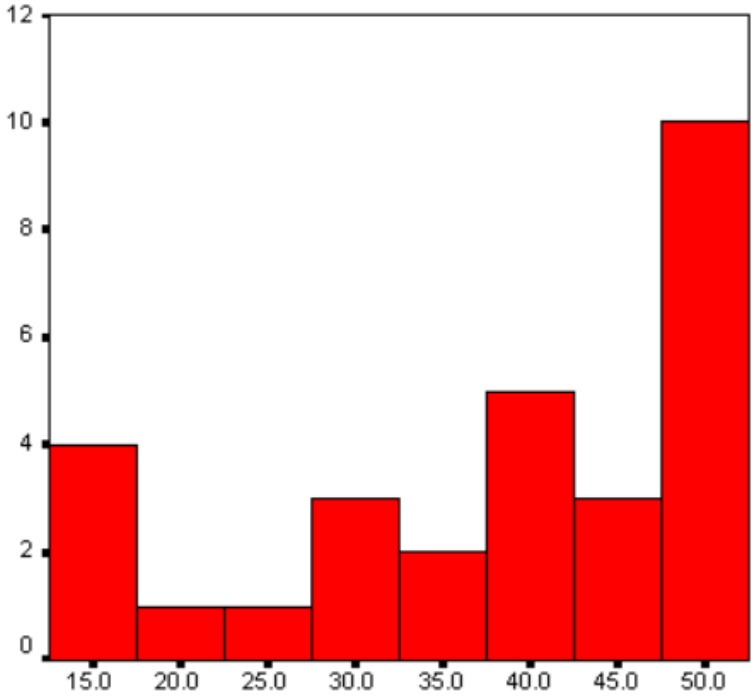
Example:

The depth of clarity of Lake Tahoe was measured at several different places with the results in inches as follows:

15.4, 16.7, 16.9, 17.0, 20.2, 25.3, 28.8, 29.1, 30.4, 34.5, 36.7, 39.1, 39.4, 39.6, 39.8, 40.1, 42.3, 43.5, 45.6, 45.9, 48.3, 48.5, 48.7, 49.0, 49.1, 49.3, 49.5, 50.1, 50.2, 52.3

We use a frequency distribution table with class intervals of length 5.

Class Interval	Frequency	Relative Frequency	Cumulative Relative Frequency
15 -<20	4	0.129	0.129
20 -<25	1	0.032	0.161
25 -< 30	3	0.097	0.258
30 -< 35	2	0.065	0.323
35 -< 40	6	0.194	0.516
40 -< 45	3	0.097	0.613
45 -< 50	9	0.290	0.903
50 -< 55	3	0.097	1.000
Total	31	1.000	



The Shape of a Histogram

A histogram is unimodal if there is one hump, bimodal if there are two humps and multimodal if there are many humps.

A nonsymmetric histogram is called skewed if it is not symmetric.

If the upper tail is longer than the lower tail then it is positively skewed.

If the upper tail is shorter than it is negatively skewed.

There are special percentile that deserve recognition.

1. The second quartile (Q_2) is the median or the 50th percentile
2. The first quartile (Q_1) is the median of the data that falls below the median. This is the 25th percentile
3. The third quartile (Q_3) is the median of the data falling above the median. This is the 75th percentile

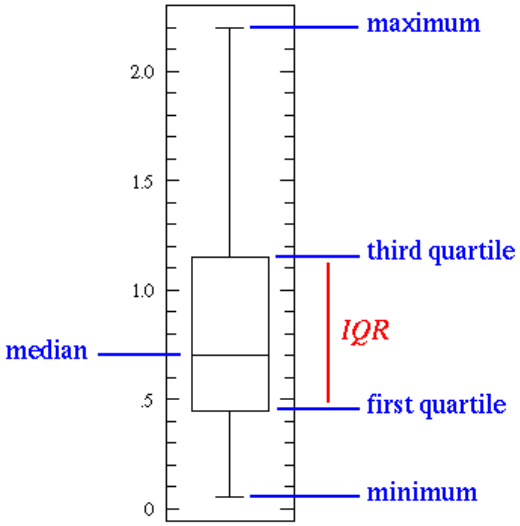
We define the interquartile range as the difference between the first and the third quartile

$$IQR = Q_3 - Q_1$$

Boxplot

Box-and-whisker plot: Requires (**five-number summary**):

- Minimum entry
- First quartile= $Q_1 = P_{25}$
- Median= $Q_2 = P_{50}$
- Third quartile = $Q_3 = P_{25}$
- Maximum entry



Central Limit Theorem (CLT)

CLT is an important finding and pillar in the fields of statistics and probability.

The theorem states that as the size of the sample increases, the distribution of the mean across multiple samples will approximate a Gaussian distribution.

If we calculate the mean of a sample, it will be an estimate of the mean of the population distribution.

If we draw multiple independent samples, and calculate their means, the distribution of those means will form a Gaussian distribution.