Statistical Descriptors

Statistics

- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data

- Two ways to present statistics

  - Presenting Data in Tables and Charts

    Bar Chart, Pie Chart, Histogram, Time Series, Scatter Plot, Box- and-whisker Plot …

  - Numerical measures

    Sample Mean, Sample Median, Sample Mode, Sample Variance, Sample Standard Deviation, Sample Percentiles, Sample Correlation Coefficient

Usage of Capital and Small Letters

- Large capital letter X denote the random variable
- Small letter x denote the sampled value of the random variable. It is called the statistic

Example

$\bar{X}$ sample mean as a random variable

$\bar{x}$ the value of the sample mean of the current sample

S^2 sample variance as a random variable

s^2 the value of the sample variance of the current sample

Population and Sample

- Population: total collection of elements. Often too large to examine each of its members
- Sample: we try to learn about the population by choosing and then examining a subset of its element. This subset is called a sample
- Statistic: information derived from the sample
- Random sampling: Choose the subset at random. It is used to ensure fairness, so that we can get unbiased samples. Otherwise we get biased samples.

  Sampling Method: Two types
  - Random sampling without replacement
  - Random sampling with replacement
  - which method to use depends on the problem

Example

44 students out of 52 hand in Assignment 1. This gives a sample of 44 out of a population of 52

Do you think the sample is random?   Is it a sample with replacement or a sample without replacement?

The sample is probably not random (why?) What bias do you think statistics based on the sample may have?

It is a sample without replacement

Data Sources

- Published Sources: Data available in print or electronic form, including those found in the web. Primary data source are those published by those collecting the data. Secondary data sources refer to those compiled from primary source

- Experiments: A study that examines the effect on a variable of varying the value(s) of another variable or variables, while keeping all other things equal. A typical experiment contains both a treatment group and a control group. The control group does not receive the treatment

- Surveys: Use questionnaires to gather values for the responses from a set of participants

Statistical Uses and Misuses

- Individualistic Fallacy

  When working with data representing individuals, it is important not to fall victim to the individualistic fallacy. This is a mistake that occurs when researchers infer characteristics of members of a group based on information they obtained from one person in that group. It can be thought of as over-generalizing

- Ecological Fallacy

  It is a type of error in which characteristic of an area or a region are believed to be the characteristics of the people in that region. It is a mistake that is made when we impose characteristics of a group upon individuals in that group

Example (Individualistic Fallacy)

A secondary school student comments that she spends a lot of time doing art projects, then you conclude that the secondary school emphasizes art more than science

Example (Ecological Fallacy)

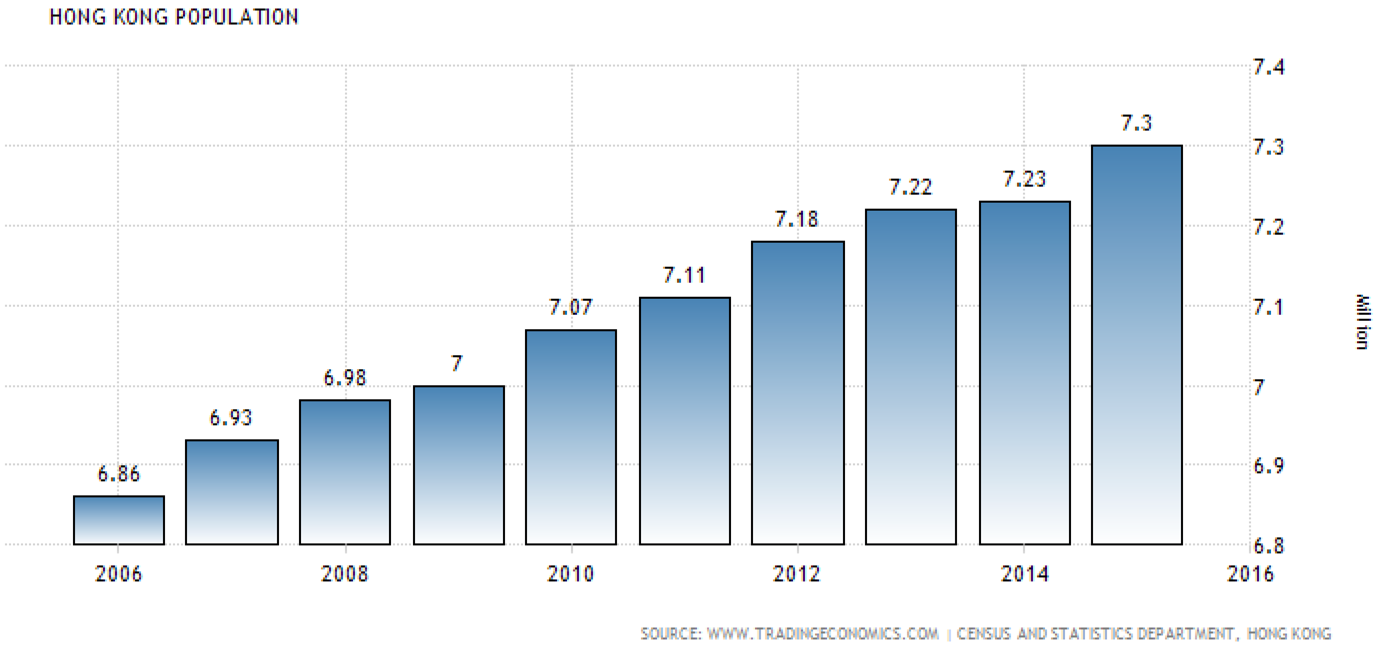
A secondary school obtains nearly the best statistics in entering the universities in Hong Kong. Then any one student in that school will be doing well in the HKDSE

Presenting Data in Tables and Charts

- Categorical variables first sort values according to the categories of the variable. Bar Chart and Pie Chart can be used to display them

- Numerical variables first establishes groups that represent separate ranges of values and then place each value into the proper group. Histogram, Time Series Plot, Scatter Plot and Box-and-Whisker Plot can be used to display them

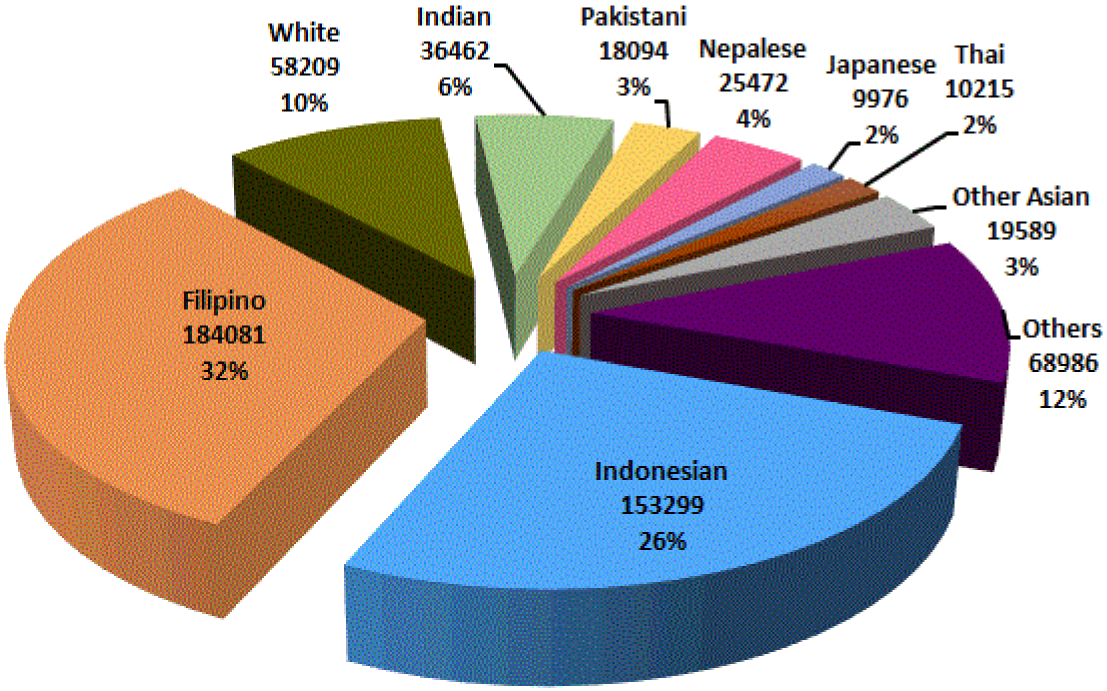
Bar Chart



SOURCE: WWW.TRADINGECONOMICS.COM | CENSUS AND STATISTICS DEPARTMENT, HONG KONG

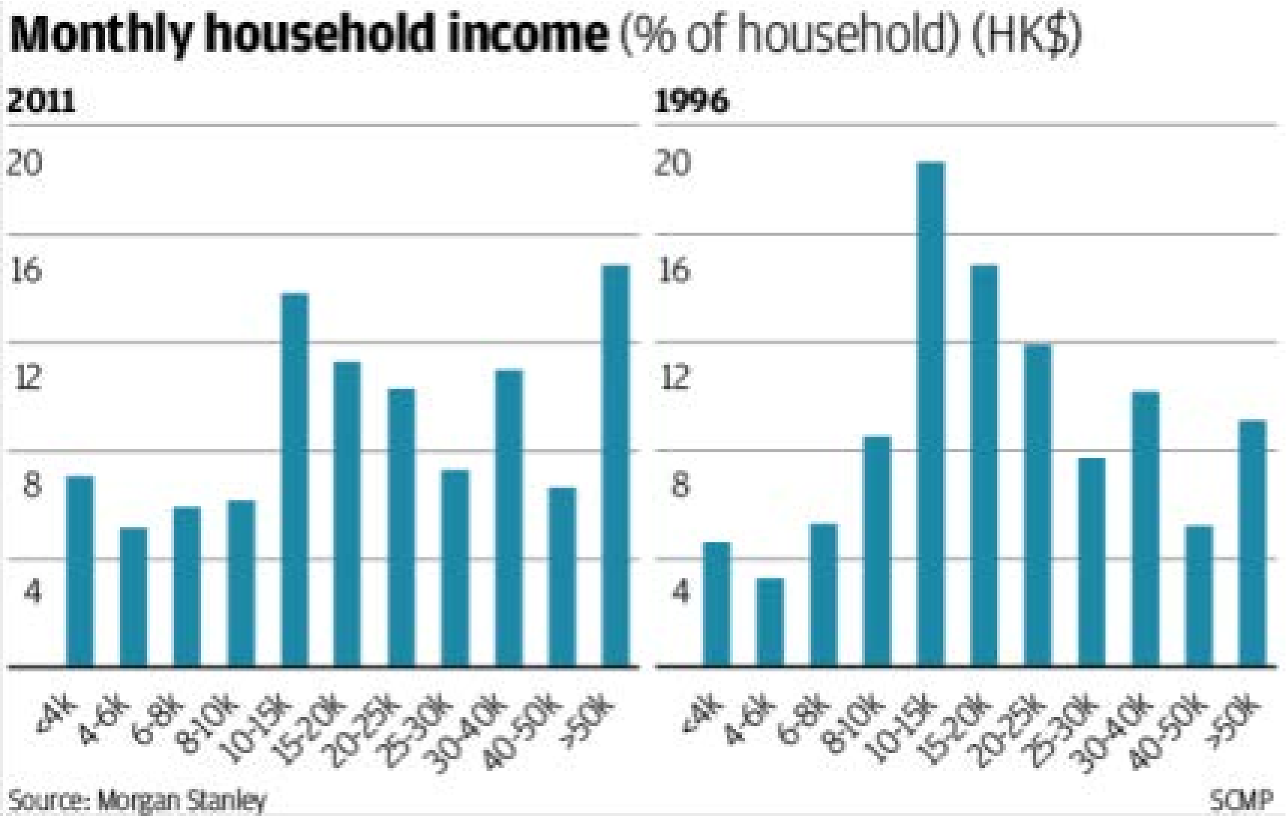
http://www.tradingeconomics.com/hong-kong/population

Pie Chart



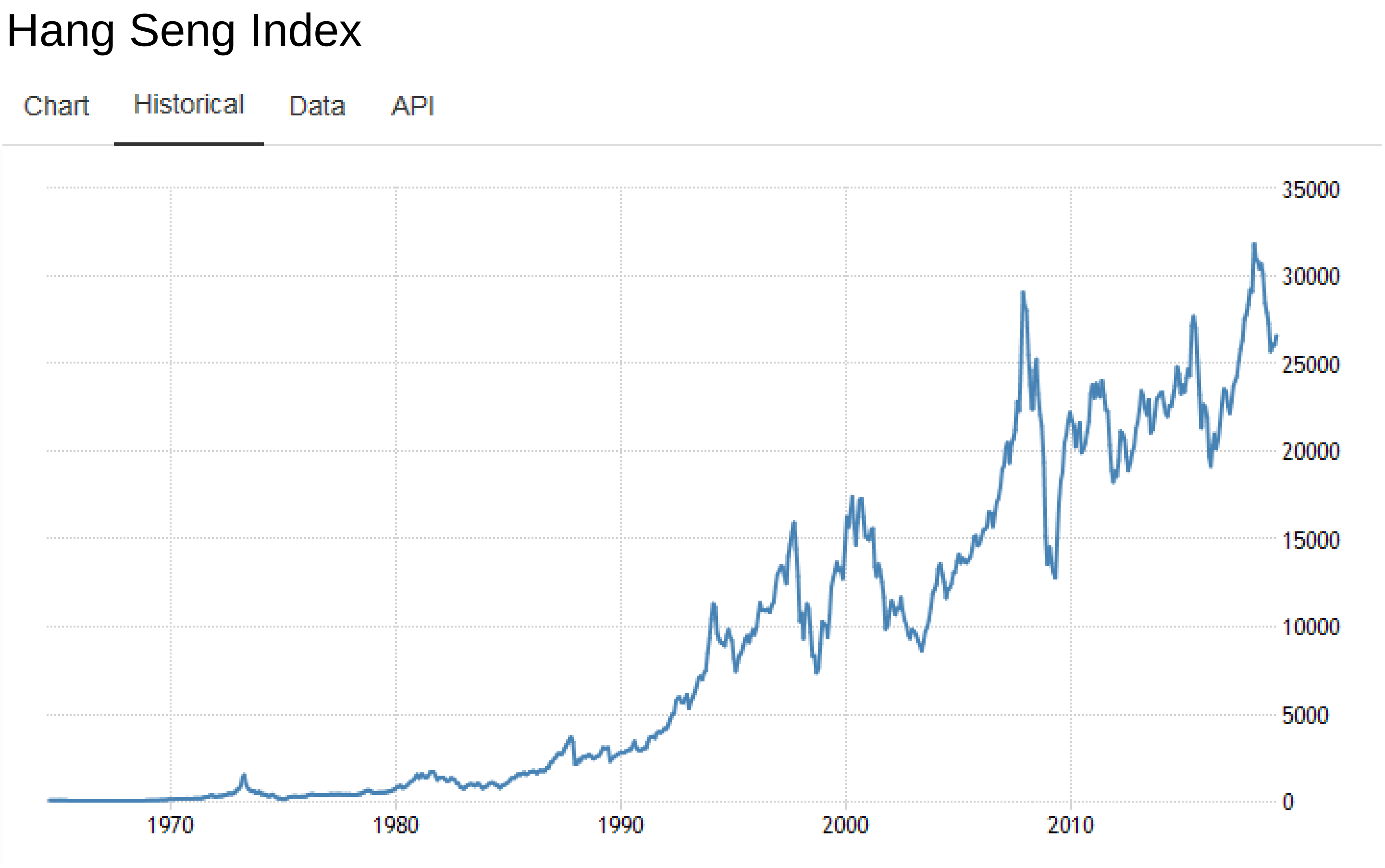
The Demographics : Ethnic Groups
http://www.had.gov.hk/rru/english/info/info_dem.html

Histogram



https://bujnoch.me/2012/03/13/salaries-in-hong-kong-the-gap-is-widening/

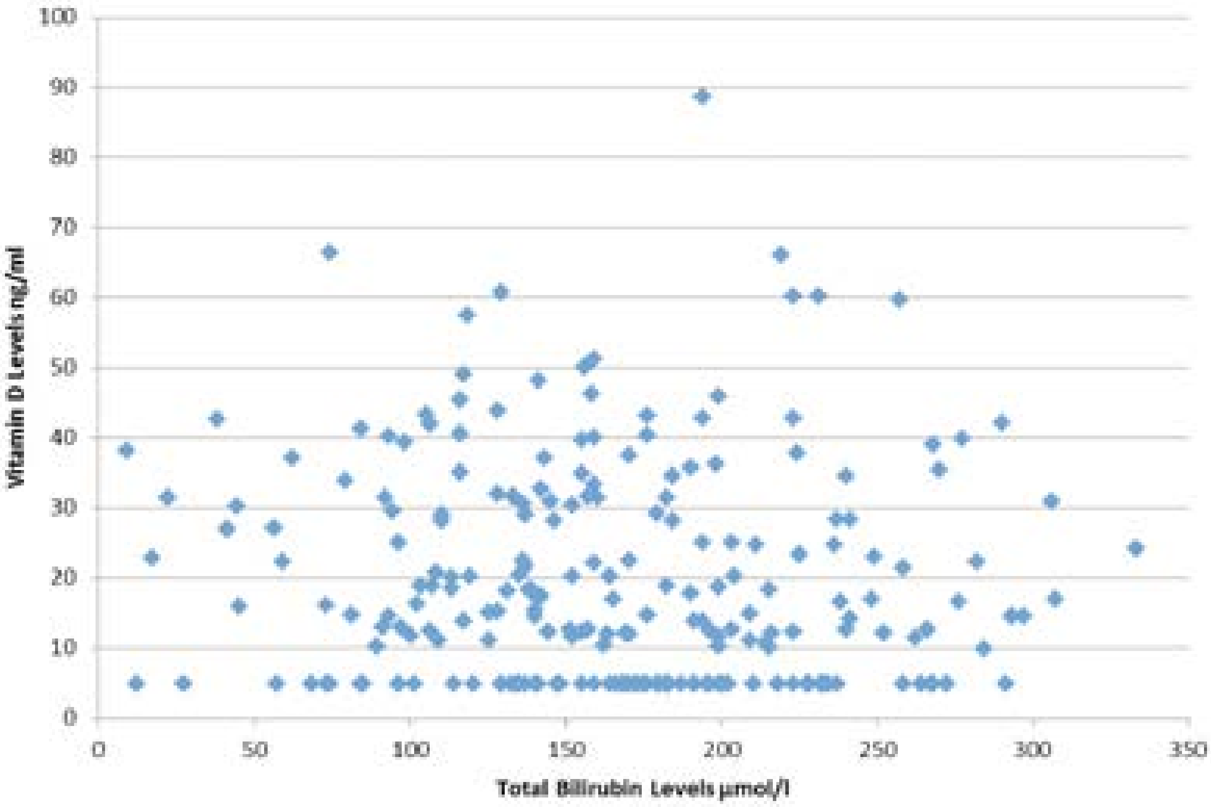
Time Series Plot



Hang Seng Index

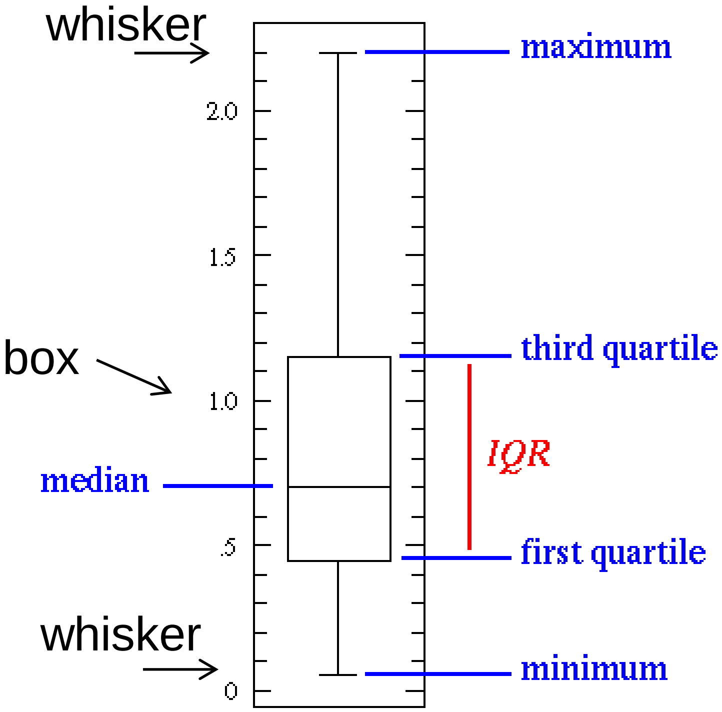
http://www.tradingeconomics.com/hong-kong/stock-market

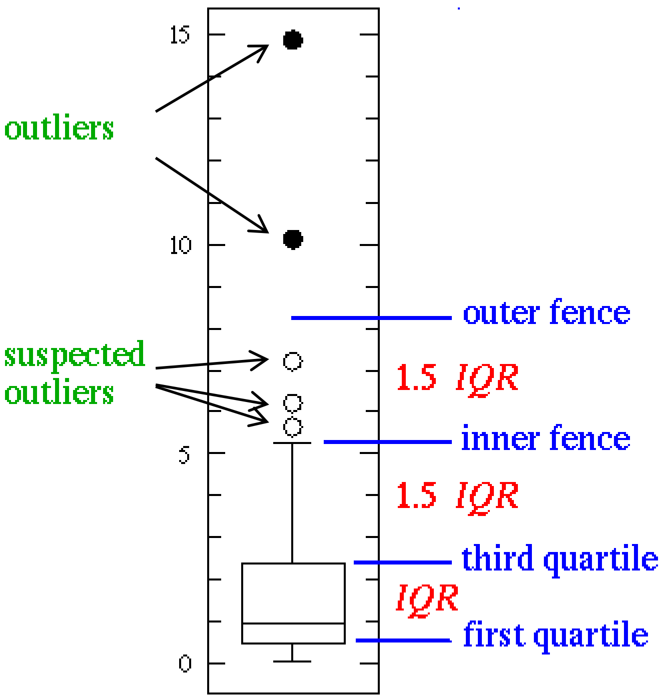
Scatter Plot



http://www.hkjpaed.org/details.asp?id=1032&show=1234

Box-and-whisker plot





IQR stands for interquartile range
First quartile = first 25% starting from minimum
Third quartile = first 75% starting from minimum

Source: http://www.physics.csbsju.edu/stats/box2.html

Numerical Descriptors

- There are various ways to summarize a data set by numbers. This includes
  - Sample Mean, Sample Median, Sample Mode (measures of "centers")
  - Sample Variance and Sample standard deviation (measures of "deviation" from "centers")
  - Sample Percentiles (measure of "positions")

Sample Mean

- Suppose we have a data set consisting of n numerical values $x_1, x_2 \dots, x_n$. The sample mean $\bar{x}$ is the arithmetic average of these values

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

- The computation of the sample mean can often be simplified by noting that if for constants a and b

$$y_i = ax_i + b \quad i = 1, \dots, n$$

then the sample mean of the data set $y_1, \dots, y_n$ is

$$\bar{y} = \frac{\sum_{i=1}^{n}(ax_i + b)}{n} = \frac{\sum_{i=1}^{n} ax_i}{n} + \frac{\sum_{i=1}^{n} b}{n} = a\bar{x} + b$$

Sample Median

- Loosely speaking, it is the middle value when the data set is arranged in increasing order
- Definition: Order the value of a data set of size n from smallest to largest. If n is odd, the sample median is the value in position $\frac{n+1}{2}$; if n is even, it is the average of the values in positions $\frac{n}{2}$ and $\frac{n}{2}+1$
- Using the median has the advantage of avoiding distortion due to outliers. For example, median values are given for Hong Kong Household Income by the Census and Statistics Department

Sample Mode

- Definition: value that occurs with the greatest frequency. If no single value occurs most frequently, then all the values that occur at the highest frequency are called modal values.

Example

From a population of 40 BEngCE graduates in a particular year, a random sample of 8 is taken. Their graduating GPA, in ascending order, are:

2.4   2.8   2.8   3.1   3.2   3.4   3.5   3.8

Find the a) sample mean   b) sample median   c) sample mode

$$a)\ \ \bar{x} = \frac{2.4 + 2.8 + 2.8 + 3.1 + 3.2 + 3.4 + 3.5 + 3.8}{8} = 3.125$$

$$b)\ \ sample\ median = \frac{3.1 + 3.2}{2} = 3.15$$

$$c)\ \ sample\ mode = 2.8$$

Sample Variance

- The previous three measures describes the “centers” (central tendency) of the data set. The following two measures describes its spread or variability
- Sample Variance measures the average value of the squares of the distances between the data values and the sample mean.
- The sample variance s^2 of the data set $x_1, \ldots, x_n$ is

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

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

- The unit of s^2 is the square of the unit of the data values

- To make S^2 an unbiased estimator of the population variance, it divides by $(n-1)$ rather than n. [More will be said about this later]

- The population variance divides by n

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

This should be used when the population consists of the n samples

Sample Standard Deviation

- The positive square root of the sample variance is called the sample standard deviation *s*

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

- The unit of *s* is the same as the unit of the data values

Example

In the GPA example above, find the variance and the standard deviation

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

$$(7)s^2 = (2.4)^2 + (2.8)^2 + \cdots + (3.8)^2 - (8)(3.125^2)$$

$$s^2 = 0.202142857$$
$$s \approx 0.45$$

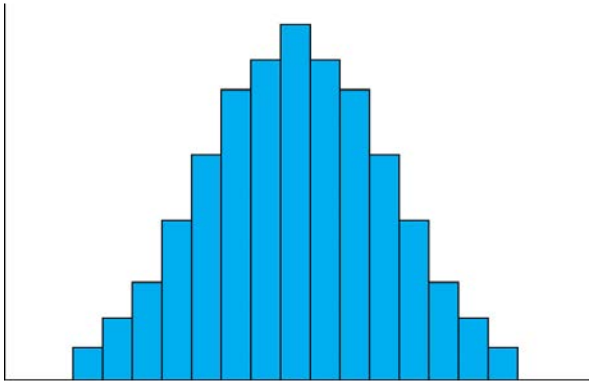
Sample Percentiles

- This measures the position of a piece of data in the data set
- The sample 100p percentile is that data value such that at least $100p$ percent of the data are less than or equal to it and at least $100(1-p)$ percent are greater than or equal to it. If two data values satisfy this condition, then the sample $100p$ percentile is the arithmetic average of these two values
- Sample 25 percentile is called the first quartile
- Sample 50 percentile is called the median
- Sample 75 percentile is called the third quartile

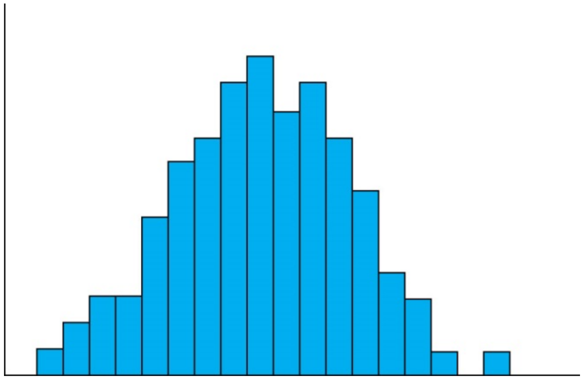
Shape of the Data Set

- Many of the large data sets observed in practice have histograms that are similar in shape. These histograms often reach their peaks at the sample median and then decrease on both sides of this point in a bell-shaped symmetric fashion. Such data sets are said to be "normal" or "approximately normal"
- It follows from the symmetry that a data set that is approximately normal will have its sample mean and sample median approximately equal
- Any data set that is not approximately symmetric about its sample median is said to be skewed.
- It is skewed to the right (left) if it has a long tail to the right (left)

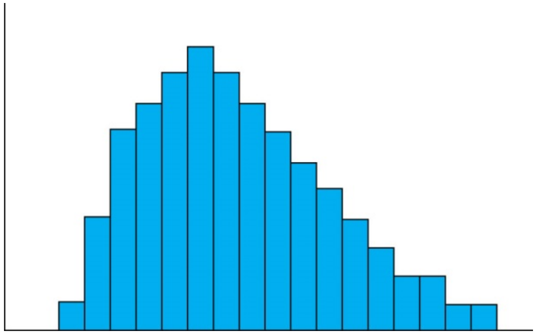
Different Shapes and Their Qualitative Descriptions



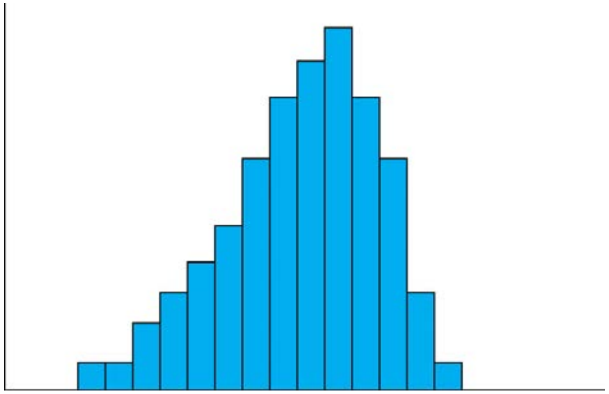
normal



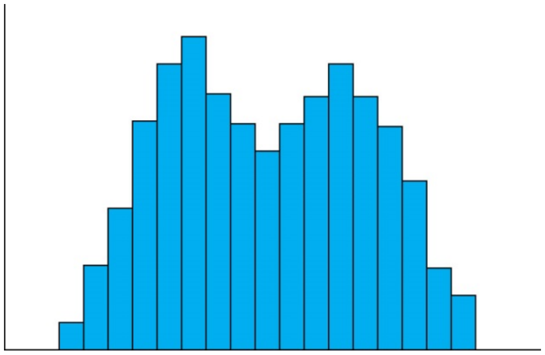
approximately normal



Skewed to right



Skewed to left

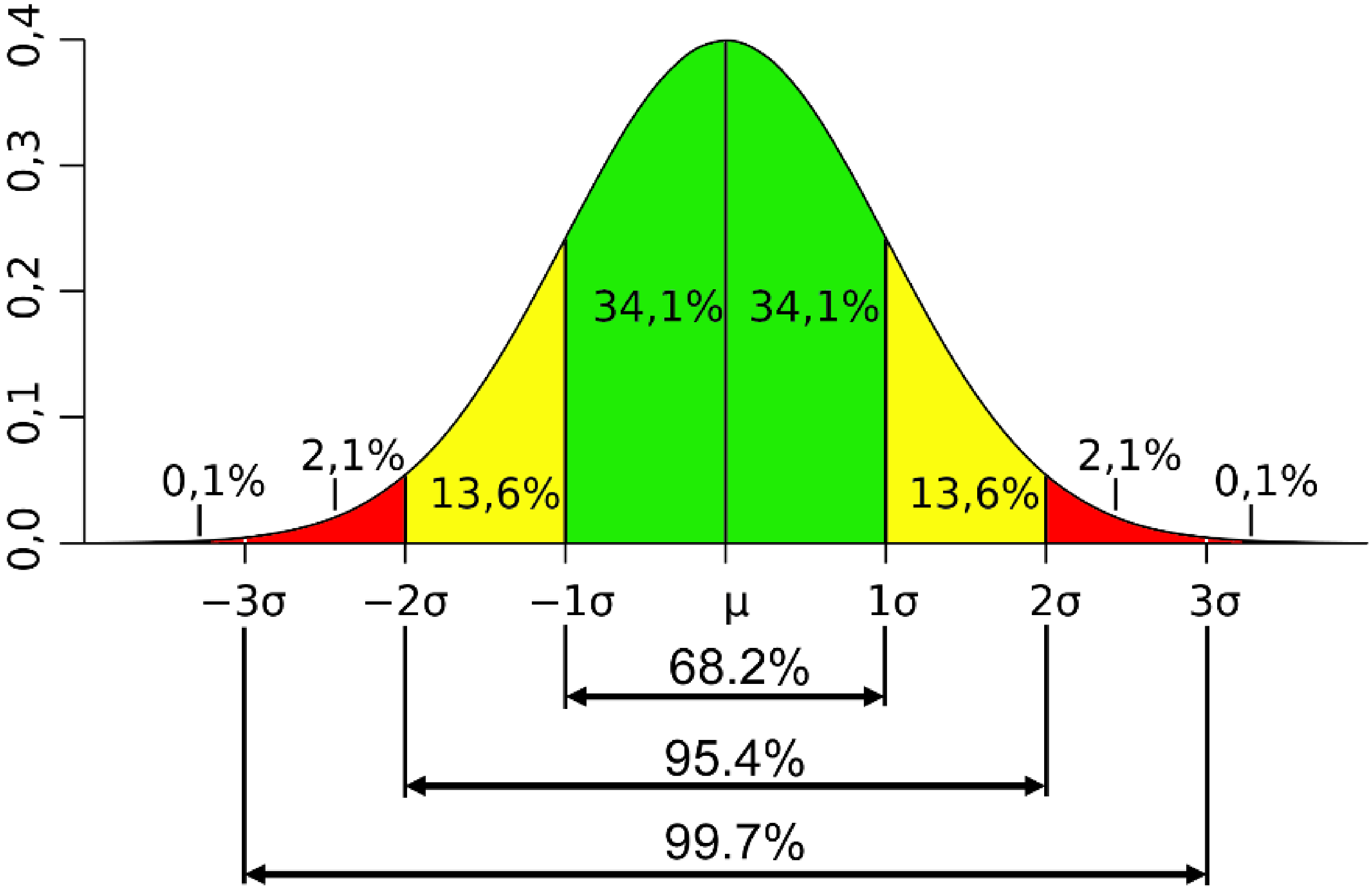


Bimodal

The empirical rule of approximately Normal distribution

- If a data set is approximately normal with sample mean $\bar{x}$ and sample standard deviation s, then

1. Approximately 68.2% of the observations lie within
$$\bar{x} \pm s$$
2. Approximately 95.4% of the observations lie within
$$\bar{x} \pm 2s$$
3. Approximately 99.7% of the observations lie within
$$\bar{x} \pm 3s$$

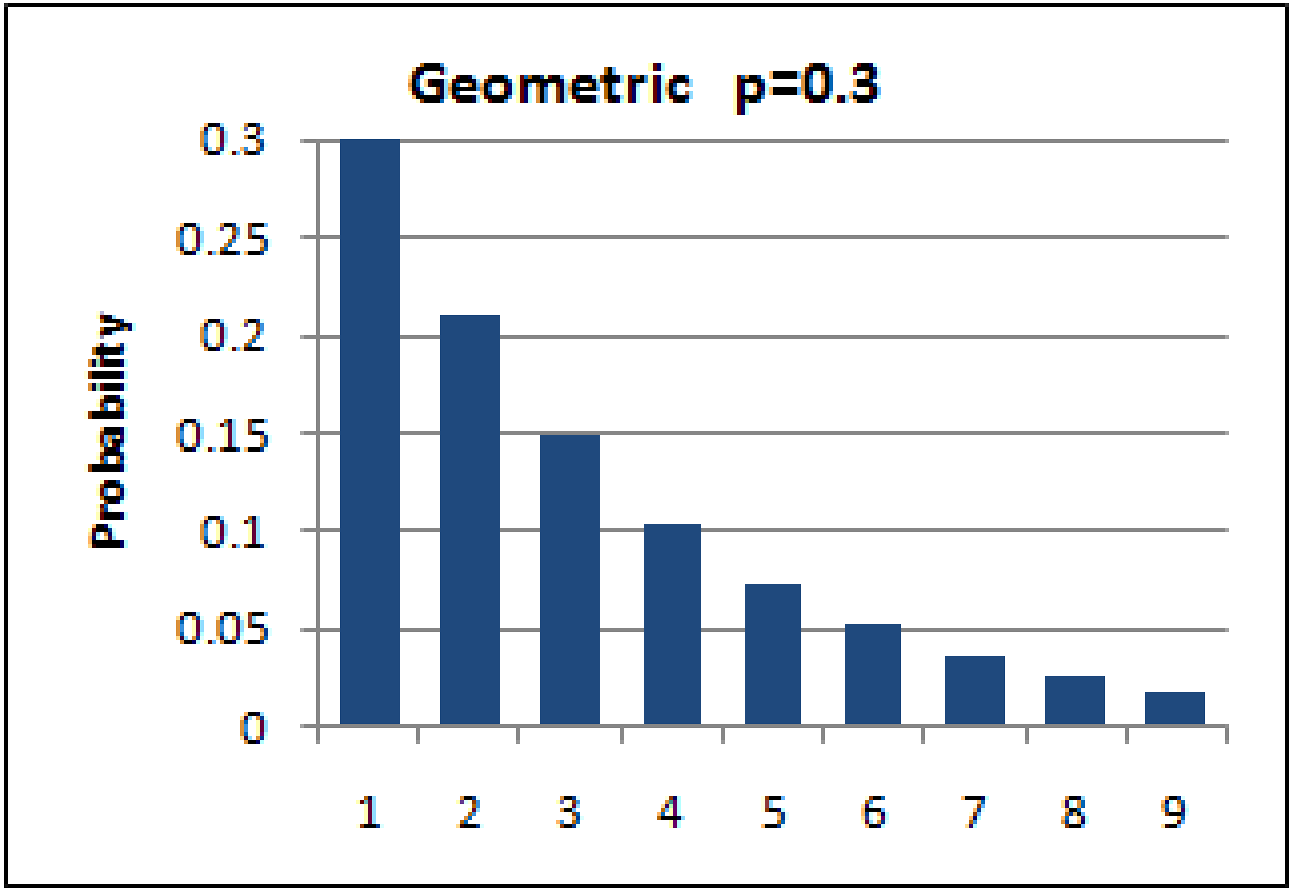

0,0
0,1
0,2
0,3
0,4
0,1%
2,1%
13,6%
34,1%
34,1%
13,6%
2,1%
0,1%
−3σ
−2σ
−1σ
μ
1σ
2σ
3σ
68.2%
95.4%
99.7%

Examples of non-normal distributions

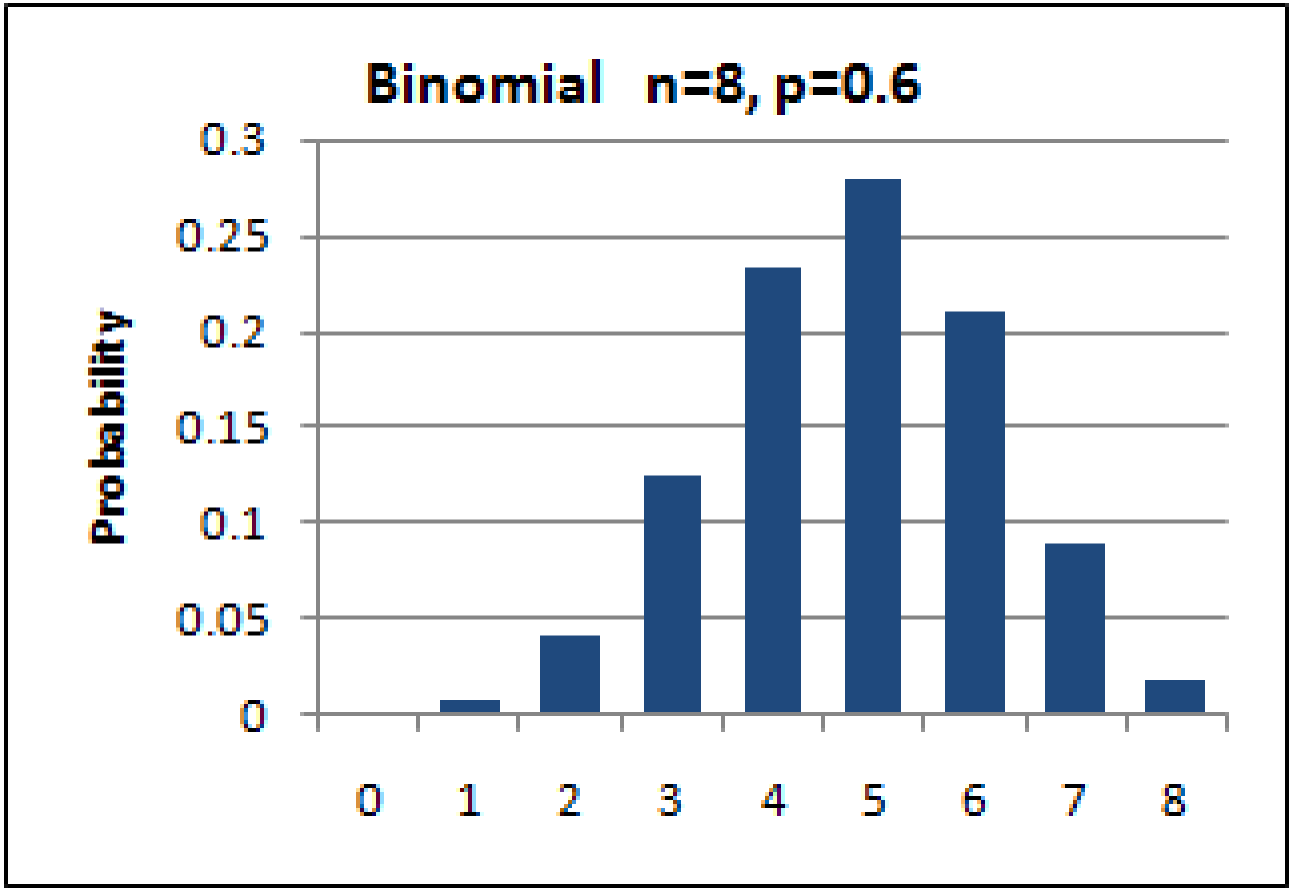
- There are many distributions which by nature is non-normal

- Some examples are
- geometric, binomial, Poisson, uniform, exponential, and many more …

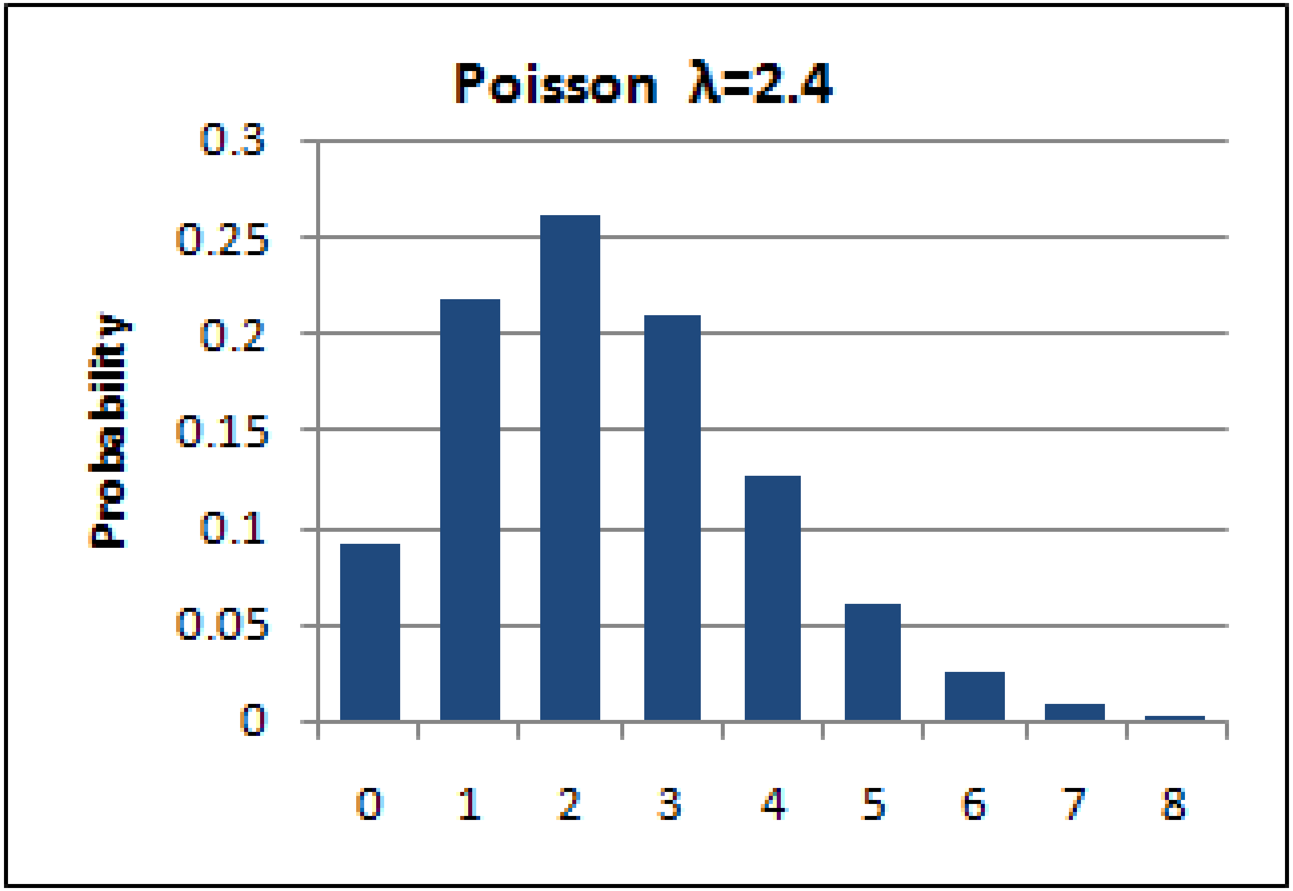
- Some distributions, for example, binomial and Poisson, may sometimes by approximated by a normal distribution



Example: Number of attempts before passing an exam

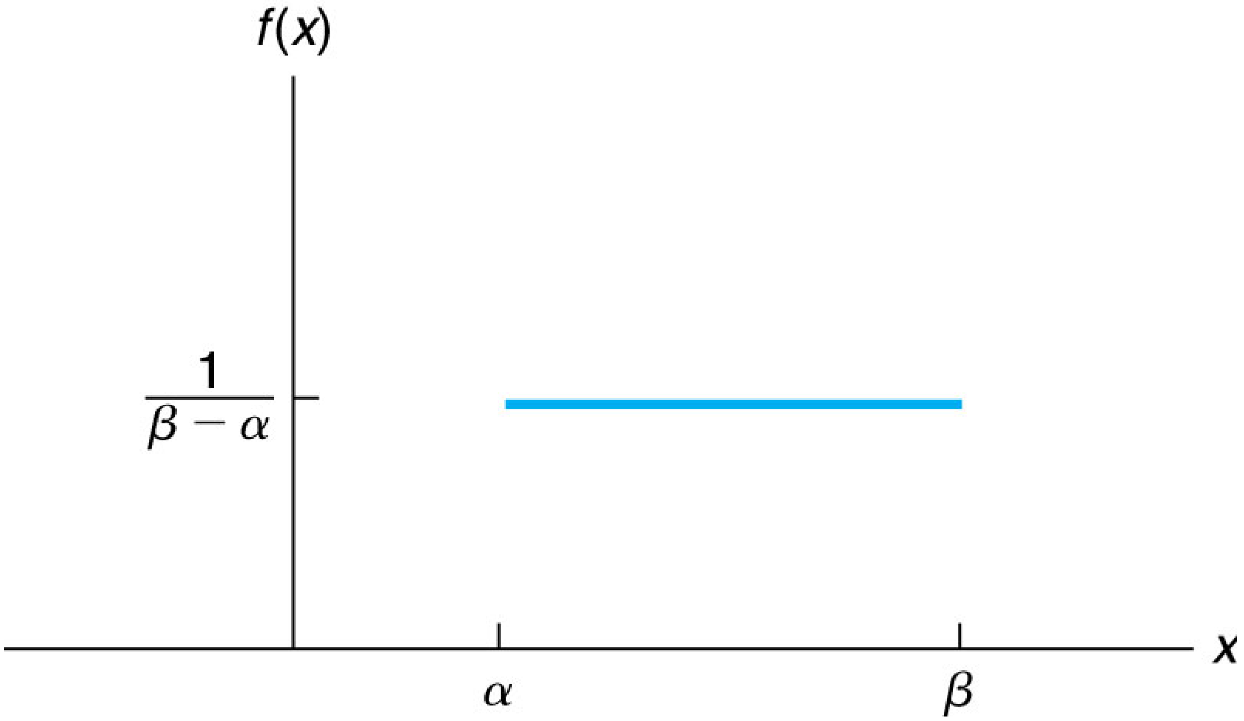


Example: Given a biased coin with probability 0.6, the number of heads in 8 throws

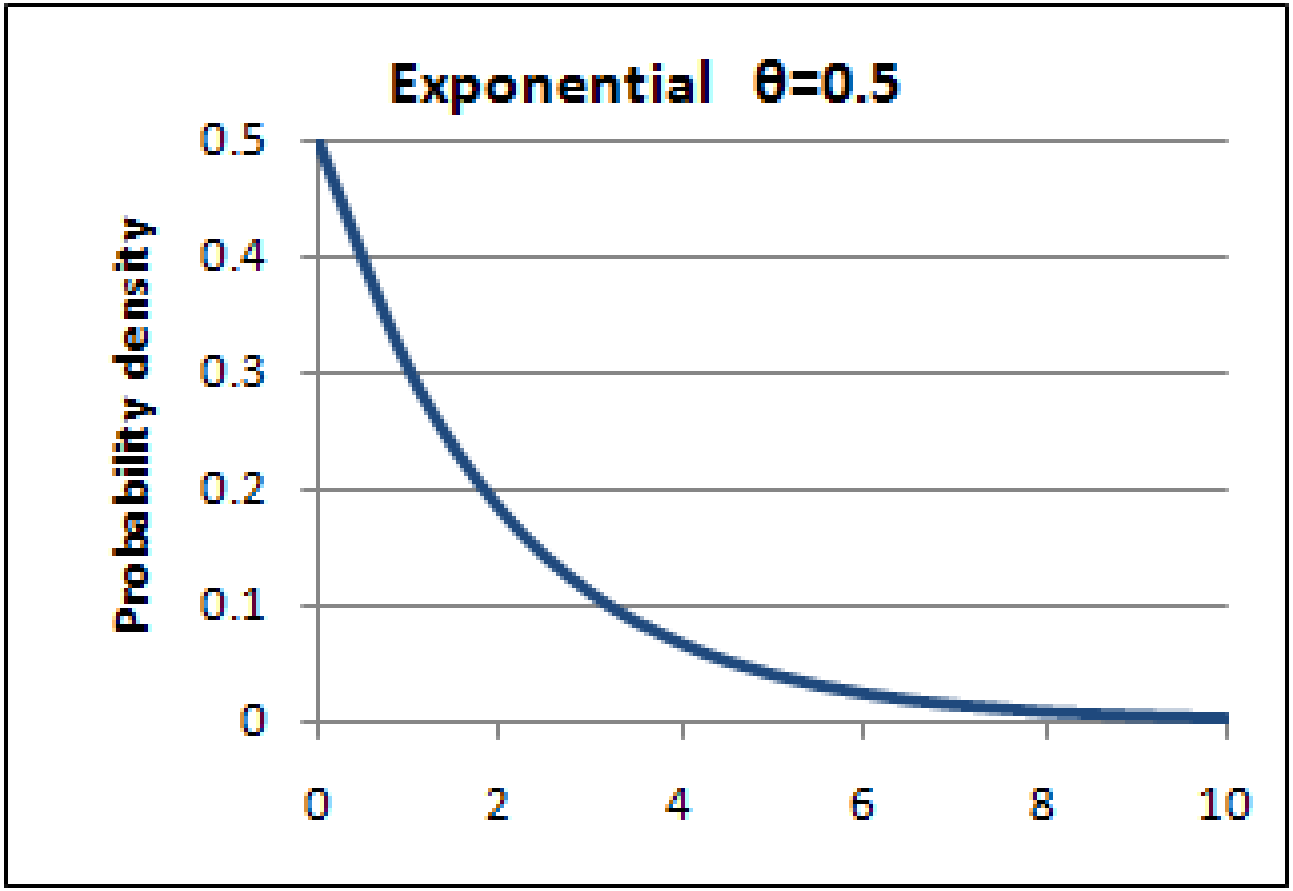


Example: The number of misprints in the 5th edition of a book

Uniform Distribution



Example: Calculator generating a pseudo-random number between $\alpha = 0$ and $\beta = 1$



Example: the time that takes a radioactive particle to decay. The time it takes for the next phone call also follows an approximate exponential distribution

Paired Data Sets

- We are often concerned with data sets that consist of pairs of values that have some relationship to each other.
- We wish to know whether the pairs are correlated to each other.
- This is measured by sample correlation coefficient r
- r is also called Pearson's coefficient

Sample Correlation Coefficient r

- Consider the data pairs (x_i, y_i), $i = 1, \dots, n$,

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

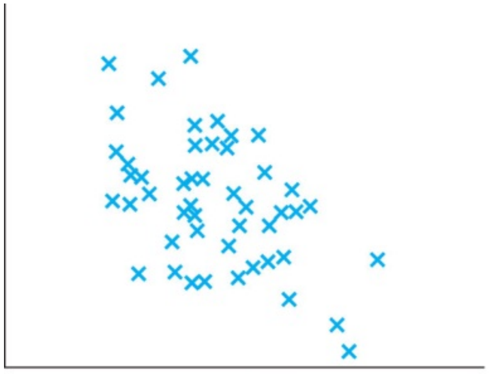
- $\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$ is the sample covariance

- s_x and s_y are the sample standard deviations of the x values and the y values respectively. They are used to normalize the random variables X and Y such that $-1 \le r \le 1$

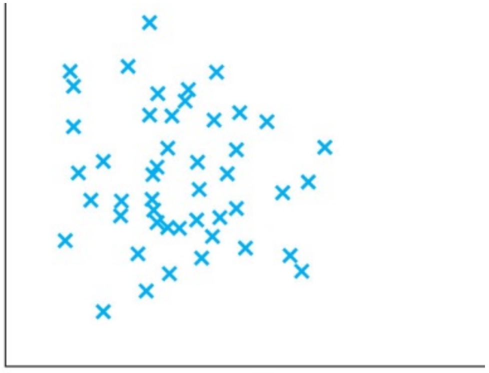
when $r > 0$, the sample data pairs are positively correlated;

when $r < 0$, the sample data pairs are negatively correlated;

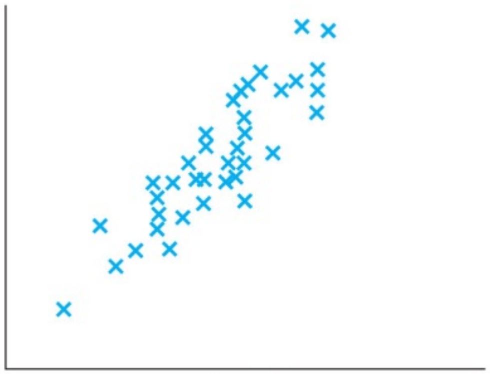
when $r = 0$, the data pairs are not correlated



$r = -.50$



$r = 0$



$r = .90$

Properties of r

1. $$-1 \leq r \leq 1$$

2. If for constants a and b, with $b > 0$,

   $$y_i = a + bx_i \qquad i = 1, \ldots, n$$

   then $r = 1$

3. If for constants a and b, with $b < 0$,

   $$y_i = a + bx_i \qquad i = 1, \ldots, n$$

   then $r = -1$

4. If r is the sample correlation coefficient for the data pairs $x_i, y_i, i = 1, \ldots, n$ then it is also the sample correlation coefficient for the data pairs

   $$a + bx_i \qquad c + dy_i \qquad i = 1, \ldots, n$$

Proof of property 1

$$\sum \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0$$

$$\Longrightarrow \sum \frac{(x_i - \bar{x})^2}{{s_x}^2} + \sum \frac{(y_i - \bar{y})^2}{{s_y}^2} - 2 \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \geq 0$$

$\Longrightarrow (n-1) + (n-1) - 2(n-1)r \geq 0$

$\Longrightarrow r \leq 1$

Similarly, use

$$\sum \left(\frac{x_i - \bar{x}}{s_x} + \frac{y_i - \bar{y}}{s_y} \right)^2 \geq 0$$

to show $r \geq -1$

Proof of property 2

$$r = 1 \Leftrightarrow \sum \left(\frac{x_i - \bar{x}}{s_x} - \frac{y_i - \bar{y}}{s_y} \right)^2 = 0$$

$\Leftrightarrow$ for all i

$$\frac{x_i - \bar{x}}{s_x} = \frac{y_i - \bar{y}}{s_y}$$

or equivalently, for all i

$$y_i = \left(\bar{y} - \frac{s_y}{s_x} \bar{x} \right) + \left(\frac{s_y}{s_x} \right) x_i = a + b x_i$$

Thus $r = 1$ if and only if the data lies on a straight line, i.e., is linearly related

Property 3 can be proved similarly using

$$\sum\left(\frac{x_i-\bar{x}}{s_x}+\frac{y_i-\bar{y}}{s_y}\right)^2 \geq 0$$

Property 2 says that $r = 1$ when the relationship is linear and is positively correlated such that larger x gives larger y

Property 3 says that $r = -1$ when the relationship is linear and is negatively correlated such that larger x gives smaller y

Property 4 can be easily proved by using the definition of r

Property 4 states that the value of r is unchanged when we add a constant to x (or y) or multiply x (or y) by a constant.

Example 1

Giving the data y in m or cm will not affect r

Example 2

Giving the data x in pound or kg will not affect r

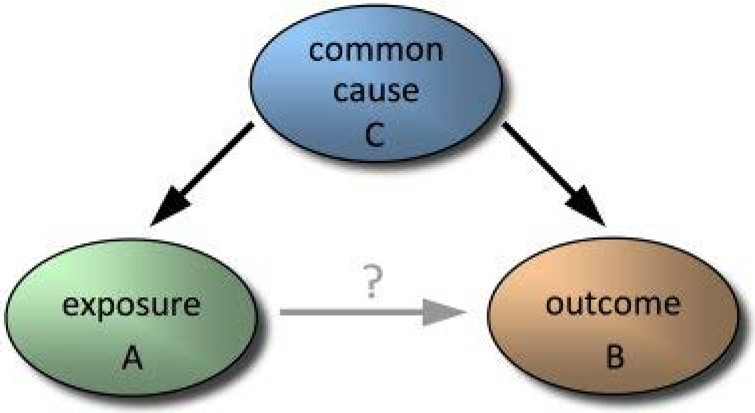
Example 3

Show that giving the data x in Celsius or Fahrenheit will not affect r

$$x({}^{0}C) = \frac{5}{9}[x({}^{o}F) - 32]$$

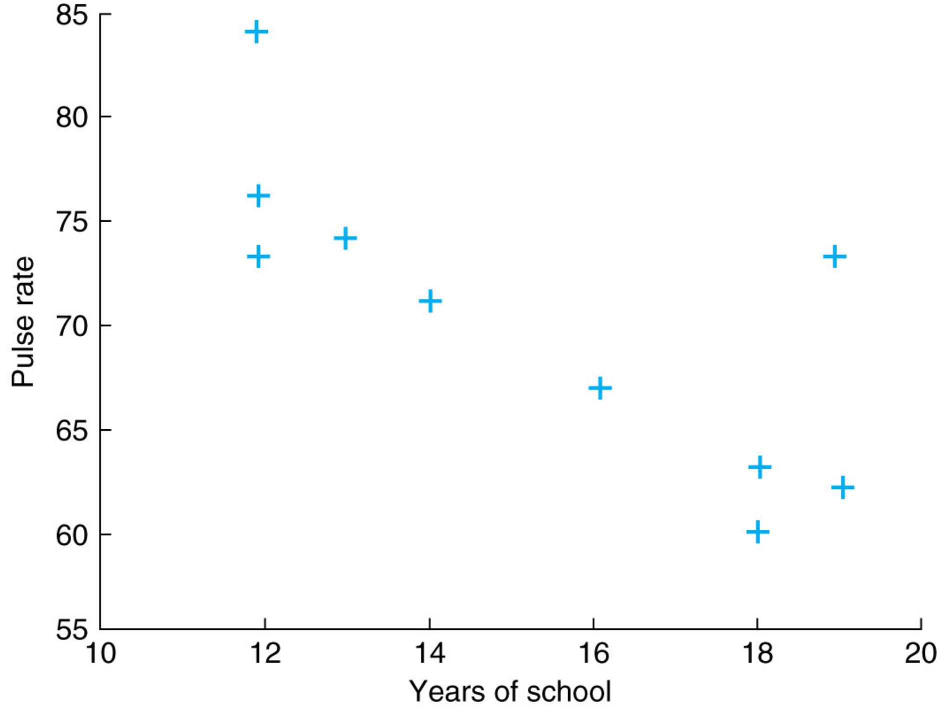
Physical Meaning of Correlation

Correlation does not imply causation



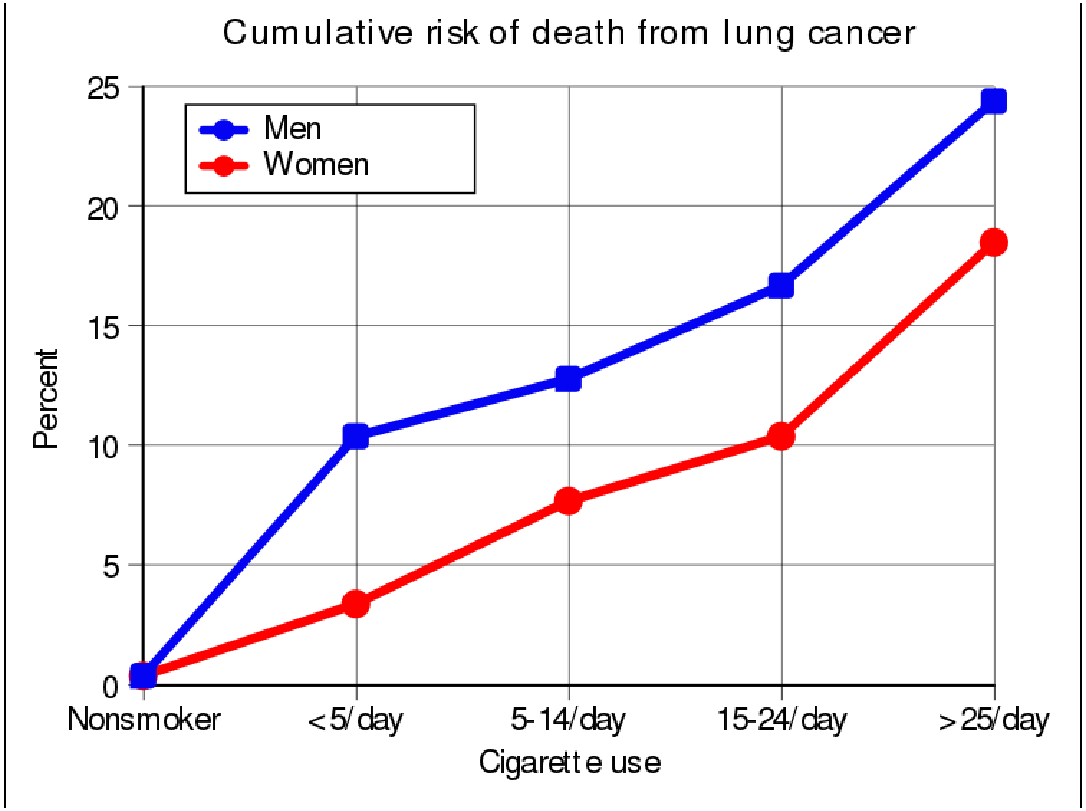
exposure A may be highly correlated with outcome B. However, A does not cause B

Example 1: Scatter plot of the pulse rate and years of school of 10 students show a negative correlation



However, the years of school does not cause the pulse rate to drop

Example 2:



Risk of death from lung cancer is strongly correlated with smoking. In this case, it is generally believed that smoking is a factor that causes lung cancer (https://en.wikipedia.org/wiki/Lung_cancer)

References

- Text book, Ch. 2
- D.M. Levine and D.F, Stephan, Even You can Learn Statistics and Analytics, 3rd Edition, Pearson, 2015.
- E.J. Krieg, Statistics and Data Analysis for Social Science, Pearson, 2012.
- All models are wrong, 7 sources of model risk (accessed 24 Aug. 2016)
- N. Altman and M. Krzywinski, “Points of Significance: Association, correlation and causation,” *Nature Methods*, vol. 12, pp. 899-900, 2015. www.nature.com/nmeth/journal/v12/n10/full/nmeth.3587.html