TRANSFORMATION OF RANDOM VARIABLES

- If X is an rv with cdf $F(x)$, then $Y=g(X)$ is also an rv.

- If we write $y=g(x)$, the function $g(x)$ defines a mapping from the original sample space of X, Ξ, to a new sample space, Ψ, the sample space of the rv Y.

$$g(x)\colon \Xi \rightarrow \Psi$$

TRANSFORMATION OF RANDOM VARIABLES

- We associate with g an inverse mapping, denoted by g^{-1}, which is a mapping from subsets of Ψ to subsets of Ξ, and is defined by

$$\boldsymbol{g}^{-1}(\boldsymbol{A}) - \{\boldsymbol{x} : \boldsymbol{x} \in \chi : \boldsymbol{g}(\boldsymbol{x}) \in \boldsymbol{A}\}.$$

If X is a discrete rv then Ξ is countable. The sample space for Y=g(X) is Ψ={y:y=g(x),x∈Ξ}, also countable. The pmf for Y is

$$\boldsymbol{f}_Y(\boldsymbol{y}) = \boldsymbol{P}(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{\boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y})} \boldsymbol{P}(\boldsymbol{X} = \boldsymbol{x}) = \sum_{\boldsymbol{x} \in \boldsymbol{g}^{-1}(\boldsymbol{y})} \boldsymbol{f}(\boldsymbol{x})$$

FUNCTIONS OF DISCRETE RANDOM VARIABLE

- If X be a discrete rv and g be a Borel-measurable function on $\Re$. Then, $g(X)$ is also a discrete rv.

Example: Let X be an rv with pmf

$$p(x)=\begin{cases} 1/5, & x=-2 \\ 1/6, & x=-1 \\ 1/5, & x=0 \\ 1/15, & x=1 \\ 11/30, & x=2 \end{cases}$$

Let $Y=X^2$. $\longrightarrow$ $A=\{-2,-1,0,1,2\}$ $\longrightarrow$ $B=\{0,1,4\}$

$$p(y)=\begin{cases} 1/5, & y=0 \\ 7/30, & y=1 \\ 17/30, & y=4 \end{cases}$$

FUNCTIONS OF CONTINUOUS RANDOM VARIABLE

- Let X be an rv of the continuous type with pdf f. Let $y=g(x)$ be differentiable for all x and either $g'(x)>0$ for all x. Then, $Y=g(X)$ is also an rv of the continuous type with pdf given by

$$h(y)=\begin{cases} f\left[g^{-1}(y)\left|\frac{d}{dy}g^{-1}(y)\right|, \ \alpha<y<\beta\right] \\ 0, \qquad\qquad\qquad\qquad\qquad \text{otherwise}\end{cases}$$

where $\alpha=min\{g(-\infty),g(+\infty)\}$ and $\beta=max\{g(-\infty), g(+\infty)\}$.

FUNCTIONS OF CONTINUOUS RANDOM VARIABLE

- **Example:** Let X have the density

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0, & \text{otherwise} \end{cases}$$

Let $Y=e^X$.

$$X=g^{-1}(y)=log\ Y \rightarrow dx=(1/y)dy.$$

$$h(y) = 1.\left|\frac{1}{y}\right|, 0 < \log y < 1$$

$$h(y) = \begin{cases} \frac{1}{y}, & 1 < y < e \\ 0, & \text{otherwise} \end{cases}$$

FUNCTIONS OF CONTINUOUS RANDOM VARIABLE

- **Example:** Let X have the density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, -\infty < x < \infty.$$

Let $Y=X^2$. Find the pdf of Y.

Transformation of Random Variables using the cdf Method

- Let X have cdf $F_X(x)$, let $Y=g(X)$, and let

$$\Xi=\{x: f_X(x)>0\}$$ and

$$\Psi=\{y: y=g(x) \text{ for some } x\in\Xi\}.$$

a) If g is an increasing function on Ξ,

$$F_Y(y) = F_X(g^{-1}(y)) \text{ for } y \in \Psi.$$

b) If g is a decreasing function on Ξ and X is a continuous r.v.,

$$F_Y(y) = 1-F_X(g^{-1}(y)) \text{ for } y \in \Psi.$$

EXAMPLE

Uniform(0,1) distribution

- Suppose $X \sim f_X(x) = 1$ if $0<x<1$ and 0 otherwise. Find the distribution function of $Y=g(X)=-\log X$.

THE PROBABILITY INTEGRAL TRANSFORMATION

- Let X have continuous cdf $F_X(x)$ and define the rv Y as $Y=F_X(x)$. Then, Y is uniformly distributed on (0,1), that is,

$$P(Y \leq y) = y, \ 0<y<1.$$

Describing the Population or The Probability Distribution

- The probability distribution represents a population
- We're interested in describing the population by computing various parameters.
- Specifically, we calculate the population mean and population variance.

EXPECTED VALUES

Let X be a rv with pdf $f_X(x)$ and $g(X)$ be a function of X. Then, the expected value (or the mean or the mathematical expectation) of $g(X)$

$$E[g(X)] = \begin{cases} \sum_x g(x) f_X(x), \; if\ X \;\; is\ discrete \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx, \; if\ X \;\; is\ continuous \end{cases}$$

providing the sum or the integral exists, i.e., $-\infty < E[g(X)] < \infty$.

EXPECTED VALUES

- $\mathrm{E}[g(X)]$ is finite if $\mathrm{E}[|\,g(X)\,|]$ is finite.

$$E\left[\left|g(X)\right|\right]=\begin{cases}\sum_{x}\left|g(x)\right|f_X(x)<\infty, \text{ if } X \text{ is discrete}\\ \int_{-\infty}^{\infty}\left|g(x)\right| f_X(x)dx<\infty, \text{ if } X \text{ is continuous}\end{cases}$$

Population Mean (Expected Value)

- Given a discrete random variable X with values x_i, that occur with probabilities $p(x_i)$, the population mean of X is.

$$E(X) = \mu = \sum_{\text{all } x_i} x_i \cdot p(x_i)$$

Population Variance

– Let X be a discrete random variable with possible values x_i that occur with probabilities $p(x_i)$, and let $E(x_i) = \mu$. The variance of X is defined by

$$V(X) = \sigma^2 = E\left[(X-\mu)^2\right] = \sum_{\text{all } x_i} (x_i - \mu)^2 p(x_i)$$

The standard deviation is

$$\sigma = \sqrt{\sigma^2}$$

EXPECTED VALUE

- The expected value or mean value of a continuous random variable X with pdf $f(x)$ is

$$\mu = E(X) = \int_{all\ x} xf(x)dx$$

- The expected value or mean value of a continuous random variable X with pdf $f(x)$ is

$$\sigma^2 = Var(X) = E(X-\mu)^2 = \int_{\text{all x}} (x-\mu)^2 f(x)dx$$

$$= E(X^2) - \mu^2 = \int_{\text{all x}} (x)^2 f(x)dx - \mu^2$$

EXAMPLE

- The pmf for the number of defective items in a lot is as follows

$$p(x) = \begin{cases} 0.35, x = 0 \\ 0.39, x = 1 \\ 0.19, x = 2 \\ 0.06, x = 3 \\ 0.01, x = 4 \end{cases}$$

Find the expected number and the variance of defective items.

EXAMPLE

- What is the mathematical expectation if we win \$10 when a die comes up 1 or 6 and lose \$5 when it comes up 2, 3, 4 and 5?

X = amount of profit

EXAMPLE

- A grab-bay contains 6 packages worth $2 each, 11 packages worth $3, and 8 packages worth $4 each. Is it reasonable to pay $3.5 for the option of selecting one of these packages at random?

X = worth of packages

EXAMPLE

- Let X be a random variable and it is a life length of light bulb. Its pdf is

$$f(x)=2(1-x),\ 0< x < 1$$

Find $E(X)$ and $Var(X)$.

Laws of Expected Value

- Let X be a rv and a, b, and c be constants. Then, for any two functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

$$a)E\left[ag_1(X)+bg_2(X)+c\right]=aE\left[g_1(X)\right]+bE\left[g_2(X)\right]+c$$

$$b)\ If\ g_1(x)\geq 0\ for\ all\ x,\ then\ E\left[g_1(X)\right]\geq 0.$$

$$c)\ If\ g_1(x)\leq g_2(x)\ for\ all\ x,\ then\ E\left[g_1(x)\right]\leq E\left[g_2(x)\right].$$

$$d)\ If\ a\leq g_1(x)\leq b\ for\ all\ x,\ then\ a\leq E\left[g_1(X)\right]\leq b$$

Laws of Expected Value and Variance

Let X be a rv and c be a constant.

Laws of Expected Value

- $E(c) = c$
- $E(X + c) = E(X) + c$
- $E(cX) = cE(X)$

Laws of Variance

- $V(c) = 0$
- $V(X + c) = V(X)$
- $V(cX) = c^2V(X)$

SOME MATHEMATICAL EXPECTATIONS

- Population Mean: $\mu = E(X)$
- Population Variance:

$$\sigma^2 = Var(X) = E(X-\mu)^2 = E(X)^2 - \mu^2 \geq 0$$

(measure of the deviation from the population mean)

- Population Standard Deviation: $\sigma = \sqrt{\sigma^2} \geq 0$
- Moments:

$$\mu_k^* = E\left[X^k\right] \rightarrow \textit{the k-th moment}$$

$$\mu_k = E\left[X-\mu\right]^k \rightarrow \textit{the k-th central moment}$$

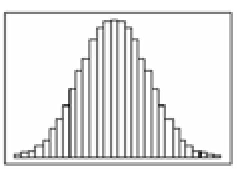
SKEWNESS

- Measure of lack of symmetry in the pdf.

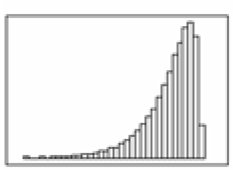
$$Skewness = \frac{E\left(X - \mu\right)^3}{\sigma^3} = \frac{\mu_3}{\mu_2^{3/2}}$$

If the distribution of X is symmetric around its mean μ,

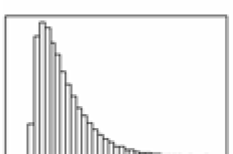
$\mu_3 = 0 \rightarrow$ Skewness=0



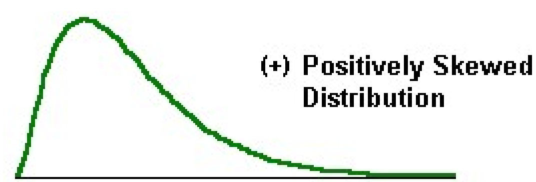
Symmetric
Bell shaped



Skewed to
the Left



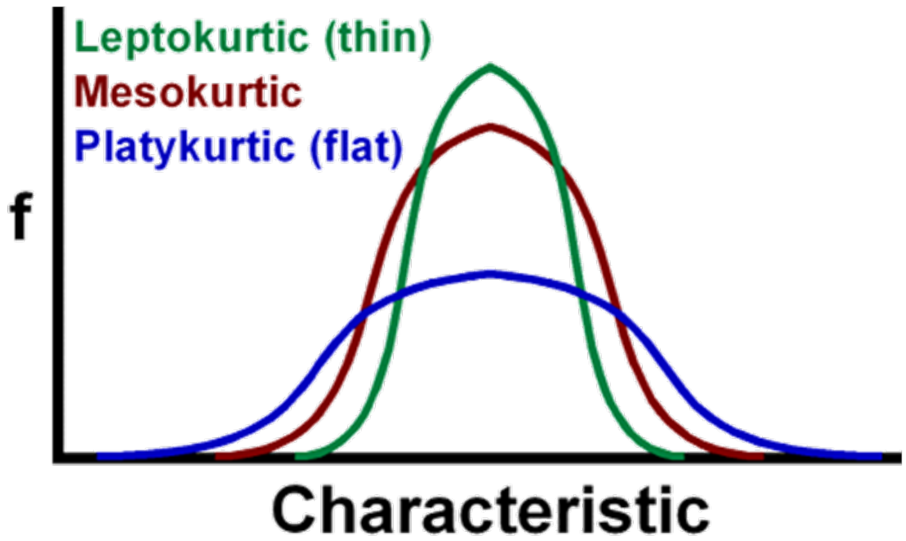
Skewed to
the Right



KURTOSIS

- Measure of the peakedness of the pdf. Describes the shape of the r,v.

$$Kurtosis = \frac{E(X-\mu)^4}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$



Kurtosis=3 $\rightarrow$ Normal

Kurtosis >3 $\rightarrow$ Leptokurtic (peaked and fat tails)

Kurtosis<3 $\rightarrow$ Platykurtic (less peaked and thinner tails)

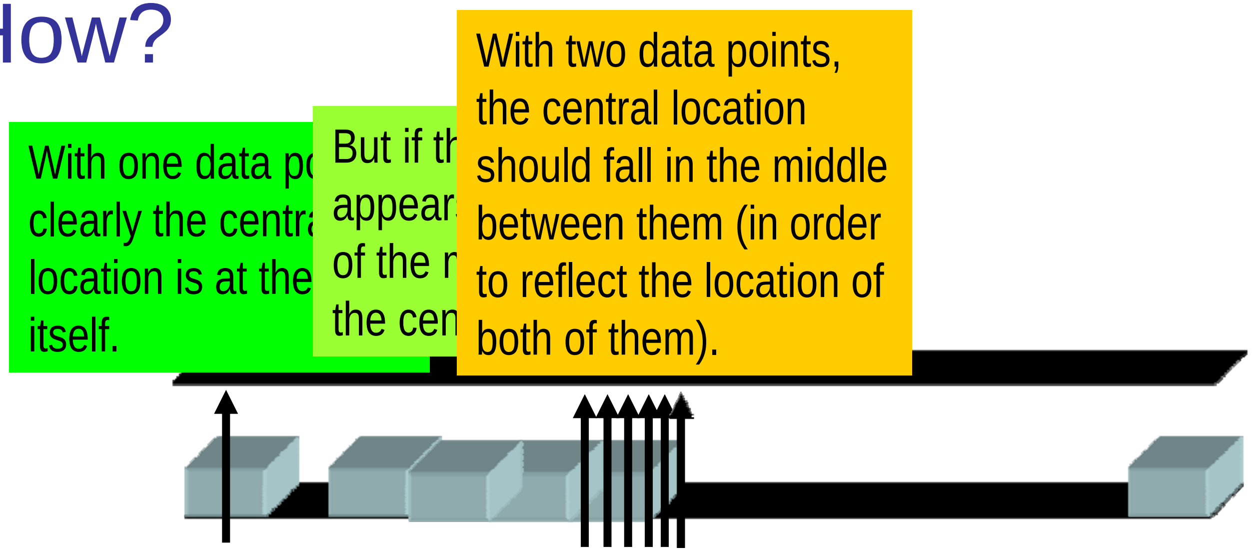
Measures of Central Location

- Usually, we focus our attention on two types of measures when describing population characteristics:
  - Central location
  - Variability or spread

The measure of central location reflects the locations of all the actual data points.

Measures of Central Location

- The measure of central location reflects the locations of all the actual data points.
- How?



The Arithmetic Mean

- This is the most popular and useful measure of central location

$$\text{Mean} = \frac{\text{Sum of the observations}}{\text{Number of observations}}$$

The Arithmetic Mean

Sample mean

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

Sample size

Population mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Population size

The Arithmetic Mean

- Example 1

The reported time on the Internet of 10 adults are 0, 7, 12, 5, 33, 14, 8, 0, 9, 22 hours. Find the mean time on the Internet.

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{0 + 7 + \ldots + 22}{10} = 11.0$$

- Example 2

Suppose the telephone bills represent the *population* of measurements. The population mean is

$$\mu = \frac{\sum_{i=1}^{200} x_i}{200} = \frac{42.19 + 38.45 + \ldots + 45.77}{200} = 43.59$$

The Arithmetic Mean

- Drawback of the mean:

  It can be influenced by unusual observations, because it uses all the information in the data set.

The Median

- The **Median** of a set of observations is the value that falls in the middle when the observations are arranged in order of magnitude. It divides the data in half.

Example 3

Find the median of the time on the internet for the 10 adults of example 1

Even number of observations

0, 0, 0, 5, 7, 8, 9, 12, 14, 22, 33

8.5

Comment

Suppose only 9 adults were sampled (exclude, say, the longest time (33))

Odd number of observations

0, 0, 5, 7, 8 9, 12, 14, 22

The Median

- Median of

8 2 9 11 1 6 3

n = 7 (odd sample size). First order the data.

1 2 3 6 8 9 11

Median

- For odd sample size, median is the $\{(n+1)/2\}^{th}$ ordered observation.

The Median

- The engineering group receives e-mail requests for technical information from sales and services person. The daily numbers for 6 days were

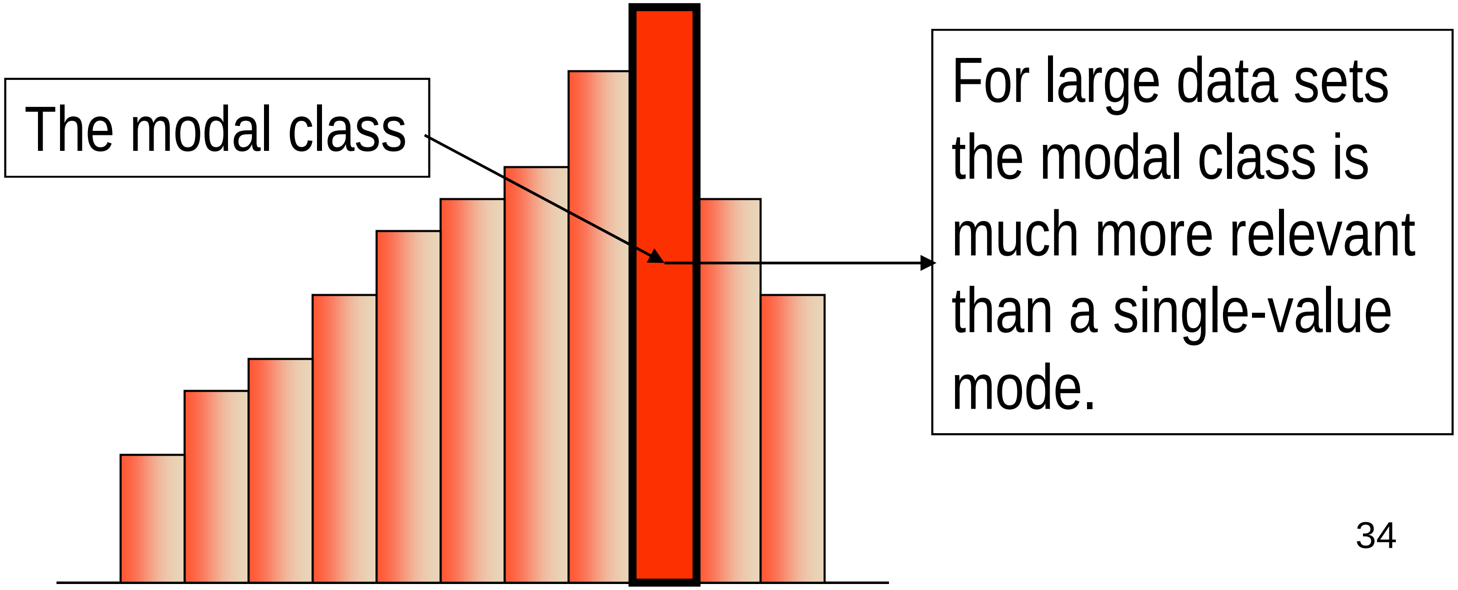
  11, 9, 17, 19, 4, and 15.

  What is the central location of the data?

- For even sample sizes, the median is the average of $\{n/2\}^{th}$ and $\{n/2+1\}^{th}$ ordered observations.

The Mode

- The **Mode** of a set of observations is the value that occurs most frequently.
- Set of data may have one mode (or modal class), or two or more modes.



The Mode

- Find the mode for the data in Example 1. Here are the data again: 0, 7, 12, 5, 33, 14, 8, 0, 9, 22

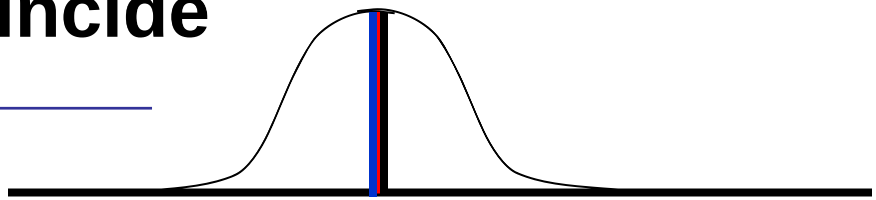
Solution

- All observation except “0” occur once. There are two “0”. Thus, the mode is zero.
- Is this a good measure of central location?
- The value “0” does not reside at the center of this set (compare with the mean = 11.0 and the median = 8.5).

Relationship among Mean, Median, and Mode

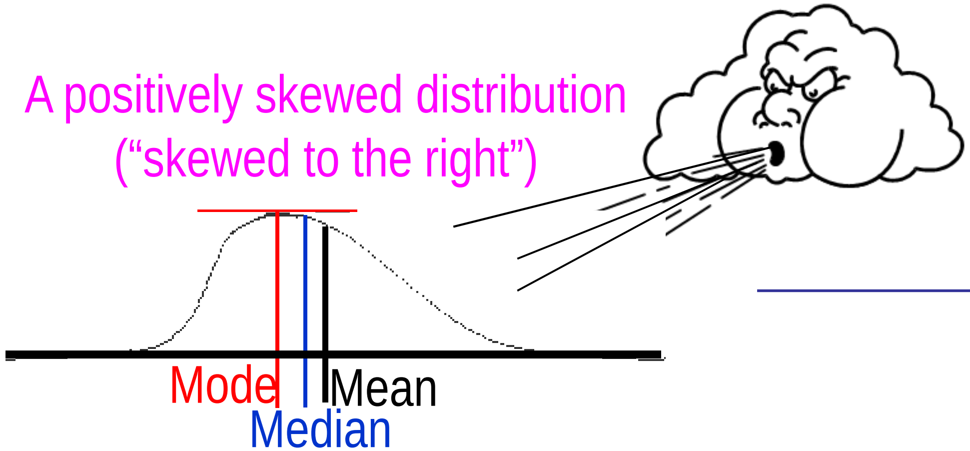
- If a distribution is symmetrical, the mean, median and mode coincide

Mean = Median = Mode



- If a distribution is asymmetrical, and skewed to the left or to the right, the three measures differ.

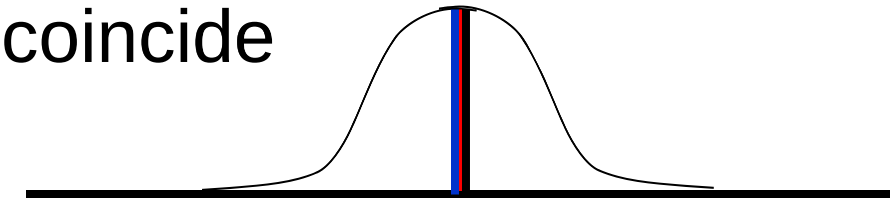
A positively skewed distribution ("skewed to the right")



Mode < Median < Mean

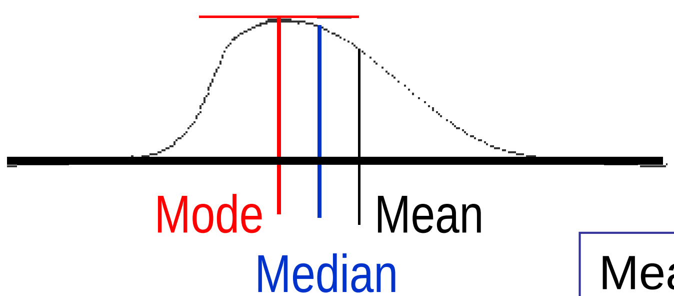
Relationship among Mean, Median, and Mode

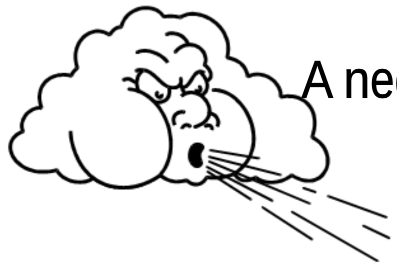
- If a distribution is symmetrical, the mean, median and mode coincide



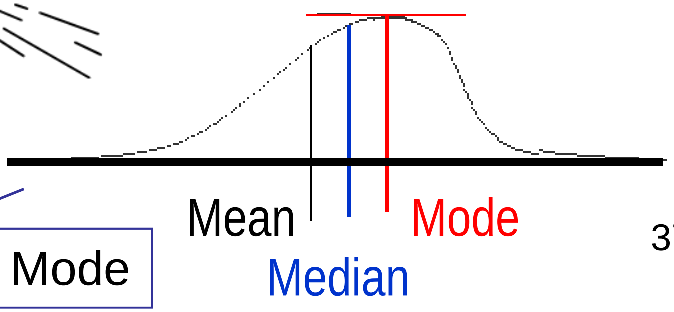
- If a distribution is non symmetrical, and skewed to the left or to the right, the three measures differ.

A positively skewed distribution ("skewed to the right")





A negatively skewed distribution ("skewed to the left")



Mean < Median < Mode

MEAN, MEDIAN AND MODE

- Why are the mean, median, and mode like a valuable piece of real estate?

LOCATION! LOCATION! LOCATION!

- *All you beginning students of statistics just remember that measures of central tendancy are all POINTS on the score scale as oppposed to measures of variability which are all DISTANCES on the score scale. Understand this maxim and you will always know where you are LOCATED!

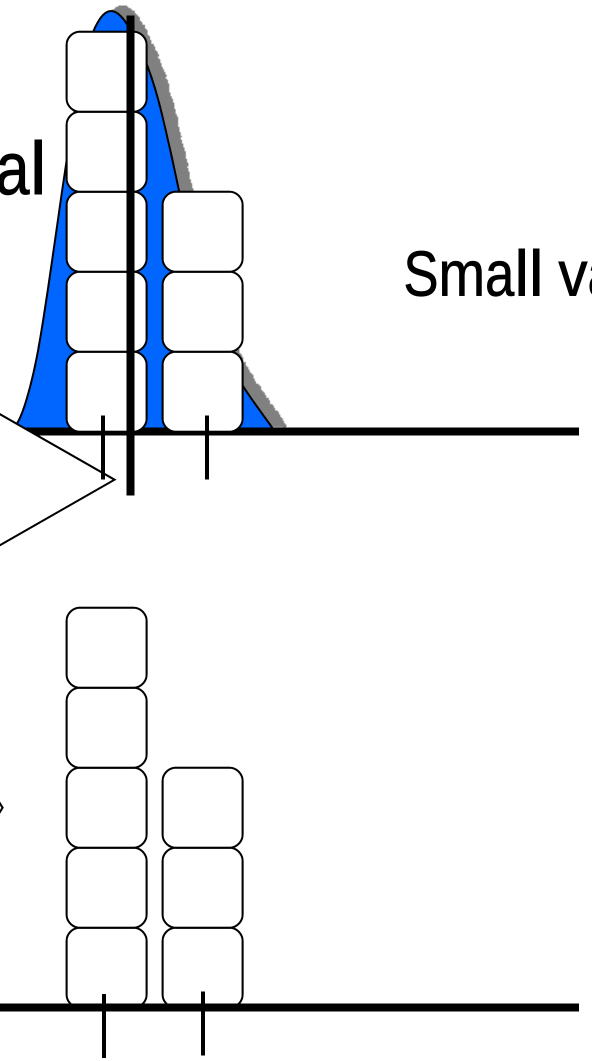
Measures of variability

- Measures of central location fail to tell the whole story about the distribution.
- A question of interest still remains unanswered:

How much are the observations spread out around the mean value?

Measures of variability

Observe two hypothetical data sets:



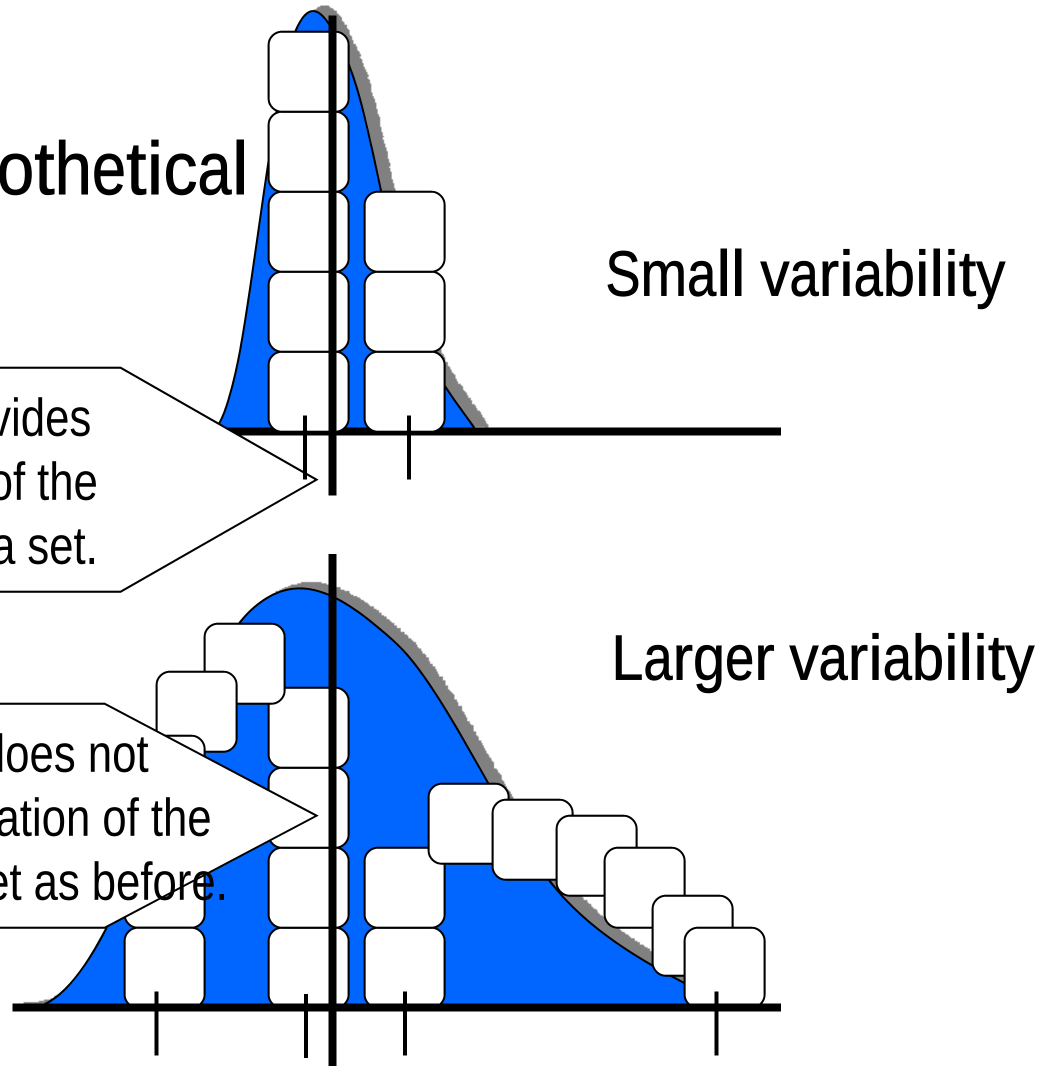
Small variability

The average value provides a good representation of the observations in the data set.

This data set is now changing to...

Measures of Variability

Observe two hypothetical data sets:



The average value provides a good representation of the observations in the data set.

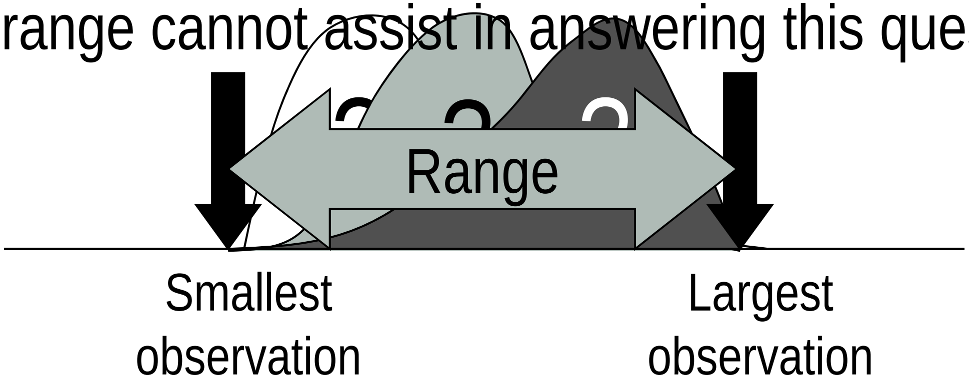
The same average value does not provide as good representation of the observations in the data set as before.

The Range

- The range of a set of observations is the difference between the largest and smallest observations.
- Its major advantage is the ease with which it can be computed.
- Its major shortcoming is its failure to provide information on the dispersion of the observations between the two end points.

But, how do all the observations spread out?

The range cannot assist in answering this question



The Variance

- This measure reflects the dispersion of *all* the observations
- The variance of a population of size N $x_1, x_2, \ldots, x_N$ whose mean is μ is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- The variance of a sample of n observations $x_1, x_2, \ldots, x_n$ whose mean is $\bar{x}$ is defined as

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

Why not use the sum of deviations?

Consider two small populations:

Can the sum of deviations

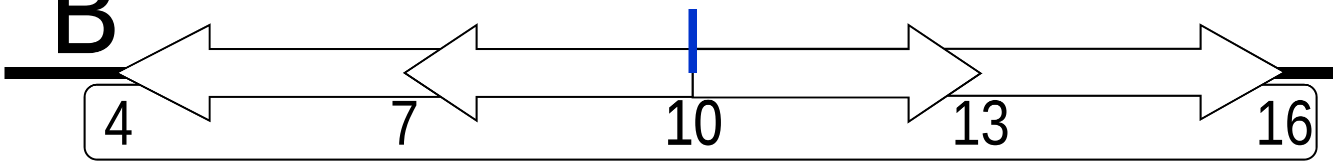
The sum of deviations is zero for both populations, therefore, is not a good measure of dispersion.

A

9-10= -1
11-10= +1
8-10= -2
12-10= +2
Sum = 0

than those in A.

B



4-10 = - 6
16-10 = +6
7-10 = -3
13-10 = +3
Sum = 0

The Variance

Let us calculate the variance of the two populations

$$\sigma_A^2 = \frac{(8-10)^2 + (9-10)^2 + (10-10)^2 + (11-10)^2 + (12-10)^2}{5} = 2$$

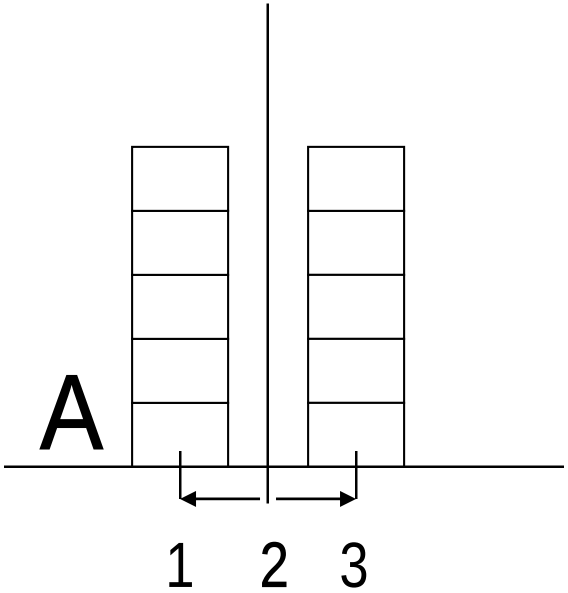
$$\sigma_B^2 = \frac{(4-10)^2 + (7-10)^2 + (10-10)^2 + (13-10)^2 + (16-10)^2}{5} = 18$$

Why is the variance defined as the average squared deviation? Why not use the sum of squared deviations as a measure of variation instead?

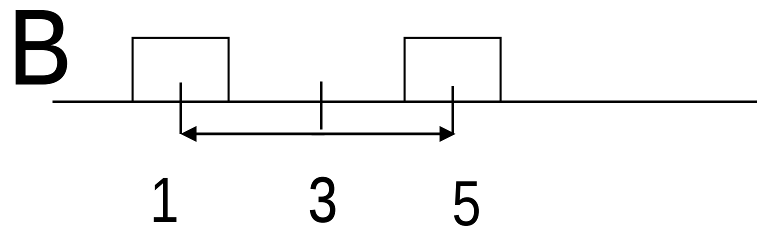
After all, the sum of squared deviations increases in magnitude when the variation of a data set increases!!

The Variance

Let us calculate the sum of squared deviations for both data sets



Data set B
is more dispersed
around the mean

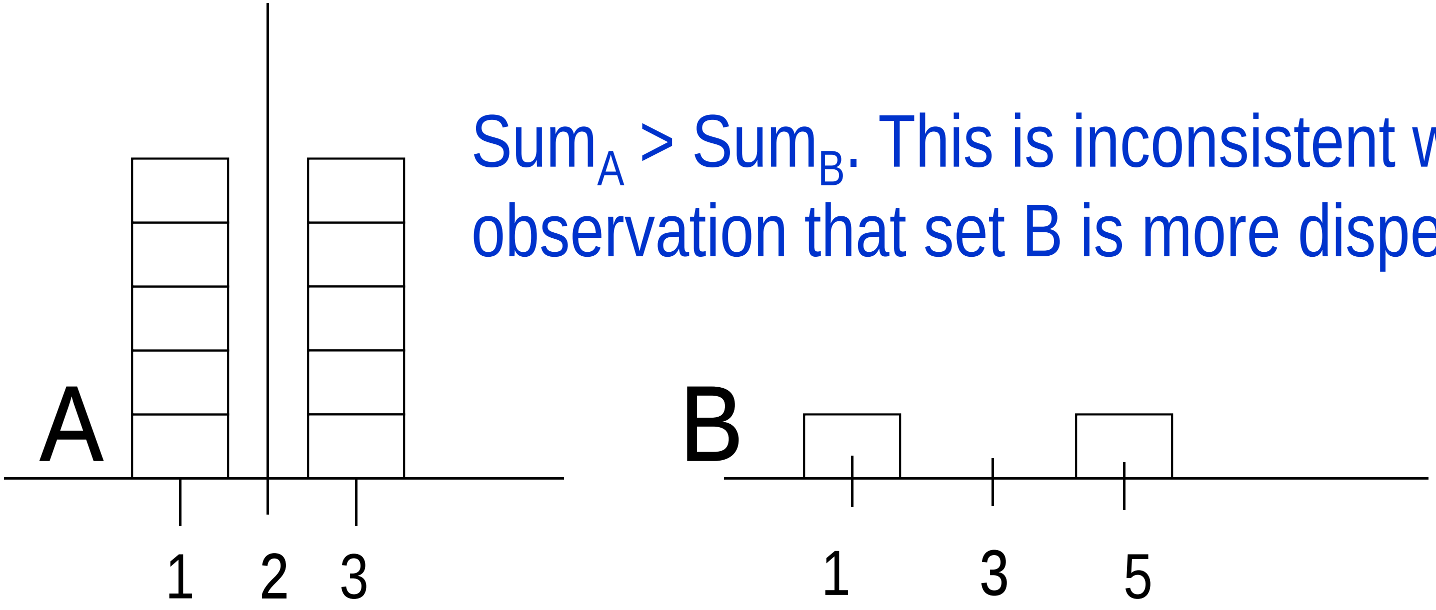


The Variance

$$Sum_A = (1-2)^2 + \ldots + (1-2)^2 + (3-2)^2 + \ldots + (3-2)^2 = 10$$

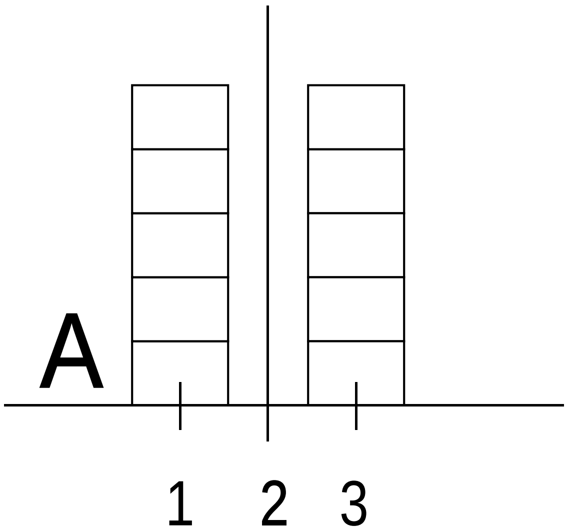
$$Sum_B = (1-3)^2 + (5-3)^2 = 8$$

$Sum_A > Sum_B$. This is inconsistent with the observation that set B is more dispersed.



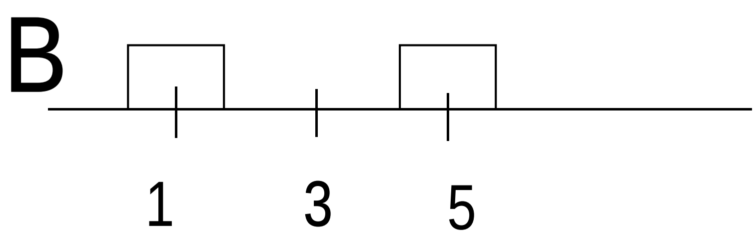
The Variance

However, when calculated on "per observation" basis (variance), the data set dispersions are properly ranked.



$$\sigma_A^2 = Sum_A/N = 10/5 = 2$$

$$\sigma_B^2 = Sum_B/N = 8/2 = 4$$



The Variance

- **Example 4**
  - The following sample consists of the number of jobs six students applied for: **17, 15, 23, 7, 9, 13.** Find its mean and variance
- **Solution**

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{17 + 15 + 23 + 7 + 9 + 13}{6} = \frac{84}{6} = 14 \text{ jobs}$$

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{1}{6-1}\left[(17-14)^2 + (15-14)^2 + \ldots (13-14)^2\right]$$

$$= 33.2 \text{ jobs}^2$$

The Variance – Shortcut method

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] =$$

$$= \frac{1}{6-1}\left[\left(17^2 + 15^2 + \ldots + 13^2\right) - \frac{\left(17 + 15 + \ldots + 13\right)^2}{6}\right] =$$

$$= 33.2 \text{ jobs}^2$$

Standard Deviation

- The standard deviation of a set of observations is the square root of the variance .

$$\text{Sample standard deviation}: s = \sqrt{s^2}$$

$$\text{Population standard deviation}: \sigma = \sqrt{\sigma^2}$$

Standard Deviation

- Example 5
  - To examine the consistency of shots for a new innovative golf club, a golfer was asked to hit 150 shots, 75 with a currently used (7-iron) club, and 75 with the new club.
  - The distances were recorded.
  - Which 7-iron is more consistent?

Standard Deviation

- Example 5 – solution

Excel printout, from the "Descriptive Statistics" sub-menu.

The innovation club is more consistent, and because the means are close, is considered a better club

Current		*Innovation*	
Mean	150.5467	Mean	150.1467
Standard Error	0.668815	Standard Error	0.357011
Median	151	Median	150
Mode	150	Mode	149
Standard Deviation	5.792104	Standard Deviation	3.091808
Sample Variance	33.54847	Sample Variance	9.559279
Kurtosis	0.12674	Kurtosis	-0.88542
Skewness	-0.42989	Skewness	0.177338
Range	28	Range	12
Minimum	134	Minimum	144
Maximum	162	Maximum	156
Sum	11291	Sum	11261
Count	75	Count	75

Interpreting Standard Deviation

- The standard deviation can be used to
  - compare the variability of several distributions
  - make a statement about the general shape of a distribution.
- The empirical rule: If a sample of observations has a mound-shaped distribution, the interval

$(\bar{x} - s, \bar{x} + s)$ contains approximately 68% of the measurements

$(\bar{x} - 2s, \bar{x} + 2s)$ contains approximately 95% of the measurements

$(\bar{x} - 3s, \bar{x} + 3s)$ contains approximately 99.7% of the measurements

Interpreting Standard Deviation

- Example 6
  A statistics practitioner wants to describe the way returns on investment are distributed.
  - The mean return = 10%
  - The standard deviation of the return = 8%
  - The histogram is bell shaped.

Interpreting Standard Deviation

Example 6 – solution

- The empirical rule can be applied (bell shaped histogram)
- Describing the return distribution
  - Approximately 68% of the returns lie between 2% and 18%
    [10 – 1(8), 10 + 1(8)]
  - Approximately 95% of the returns lie between -6% and 26%
    [10 – 2(8), 10 + 2(8)]
  - Approximately 99.7% of the returns lie between -14% and 34% [10 – 3(8), 10 + 3(8)]

The Chebyshev's Theorem

- For any value of k $\geq$ 1, greater than 100(1-1/k^2)% of the data lie within the interval from $\bar{x} - ks$ to $\bar{x} + ks$.

- This theorem is valid for *any* set of measurements (sample, population) of any shape!!

k	Interval	Chebyshev	Empirical Rule
1	$\bar{x} - s, \bar{x} + s$	at least 0% $(1\text{-}1/1^2)$	approximately 68%
2	$\bar{x} - 2s, \bar{x} + 2s$	at least 75% $(1\text{-}1/2^2)$	approximately 95%
3	$\bar{x} - 3s, \bar{x} + 3s$	at least 89% $(1\text{-}1/3^2)$	approximately 99.7%

The Chebyshev's Theorem

- Example 7
  - The annual salaries of the employees of a chain of computer stores produced a positively skewed histogram. The mean and standard deviation are $28,000 and $3,000,respectively. What can you say about the salaries at this chain?

    Solution

    At least 75% of the salaries lie between $22,000 and $34,000

    28000 – 2(3000)   28000 + 2(3000)

    At least 88.9% of the salaries lie between $19,000 and $37,000

    28000 – 3(3000)   28000 + 3(3000)

The Coefficient of Variation

- The coefficient of variation of a set of measurements is the standard deviation divided by the mean value.

$$\text{Sample coefficient of variation: cv} = \frac{s}{\bar{x}}$$

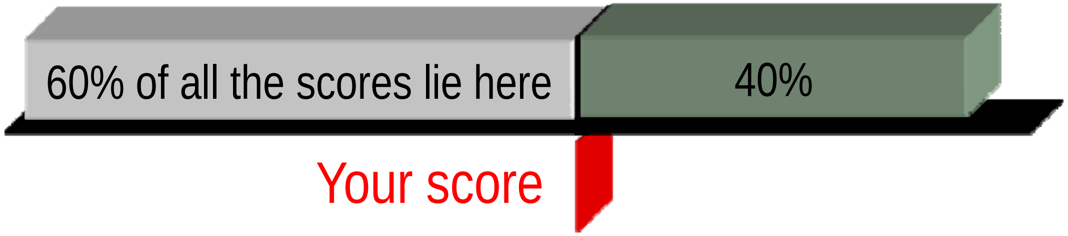
$$\text{Population coefficient of variation: CV} = \frac{\sigma}{\mu}$$

- This coefficient provides a proportionate measure of variation.

A standard deviation of 10 may be perceived large when the mean value is 100, but only moderately large when the mean value is 500

Sample Percentiles and Box Plots

- **Percentile**
  - The *p*th percentile of a set of measurements is the value for which
    - p percent of the observations are less than that value
    - 100(1-p) percent of all the observations are greater than that value.
  - **Example**
    - Suppose your score is the 60% percentile of a SAT test. Then



Sample Percentiles

- To determine the sample 100p percentile of a data set of size n, determine

a) At least np of the values are less than or equal to it.

b) At least n(1-p) of the values are greater than or equal to it.

- Find the 10 percentile of 6 8 3 6 2 8 1
- Order the data: 1 2 3 6 6 8
- Find np and n(1-p): 7(0.10) = 0.70 and 7(1-0.10) = 6.3

A data value such that at least 0.7 of the values are less than or equal to it and at least 6.3 of the values greater than or equal to it. So, the first observation is the 10 percentile.

Quartiles

- Commonly used percentiles
  - First (lower) decile= 10th percentile
  - First (lower) quartile, $\mathbf{Q_1}$ = 25th percentile
  - Second (middle) quartile, $\mathbf{Q_2}$ = 50th percentile
  - Third quartile, $\mathbf{Q_3}$ = 75th percentile
  - Ninth (upper) decile = 90th percentile

Location of Percentiles

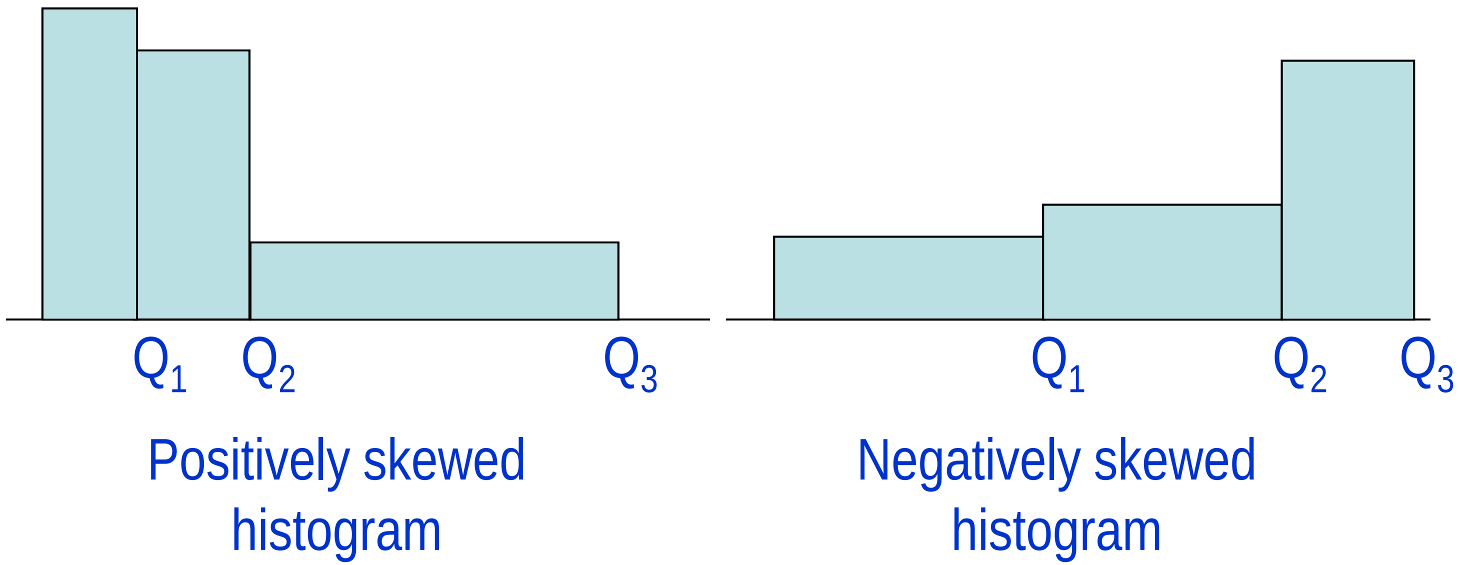
- Find the location of any percentile using the formula

$$L_P = (n+1)\frac{P}{100}$$

where L_P is the location of the P^{th} percentile

Quartiles and Variability

- Quartiles can provide an idea about the shape of a histogram



Interquartile Range

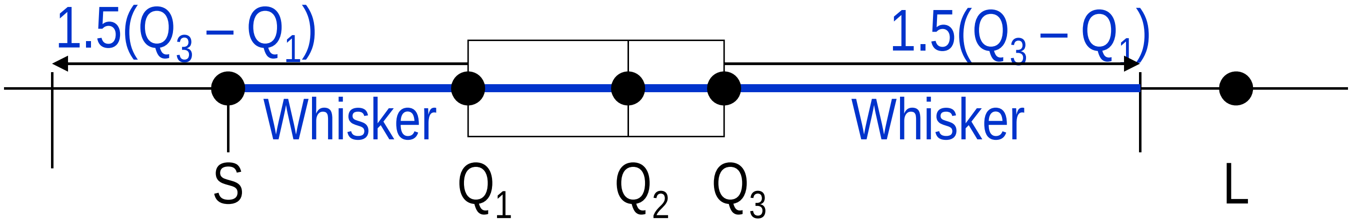
- This is a measure of the spread of the middle 50% of the observations
- Large value indicates a large spread of the observations

$$\text{Interquartile range} = Q_3 - Q_1$$

Box Plot

– This is a pictorial display that provides the main descriptive measures of the data set:

- L - the largest observation
- Q_3 - The upper quartile
- Q_2 - The median
- Q_1 - The lower quartile
- S - The smallest observation



Box Plot

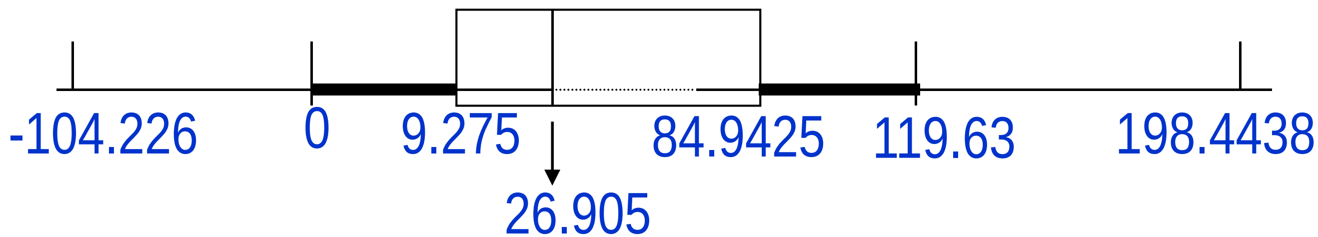
- Example 10

Bills
42.19
38.45
29.23
89.35
118.04
110.46
.

Smallest = 0
Q1 = 9.275
Median = 26.905
Q3 = 84.9425
Largest = 119.63
IQR = 75.6675
Outliers = ()

Left hand boundary = 9.275–1.5(IQR)= -104.226

Right hand boundary=84.9425+ 1.5(IQR)=198.4438



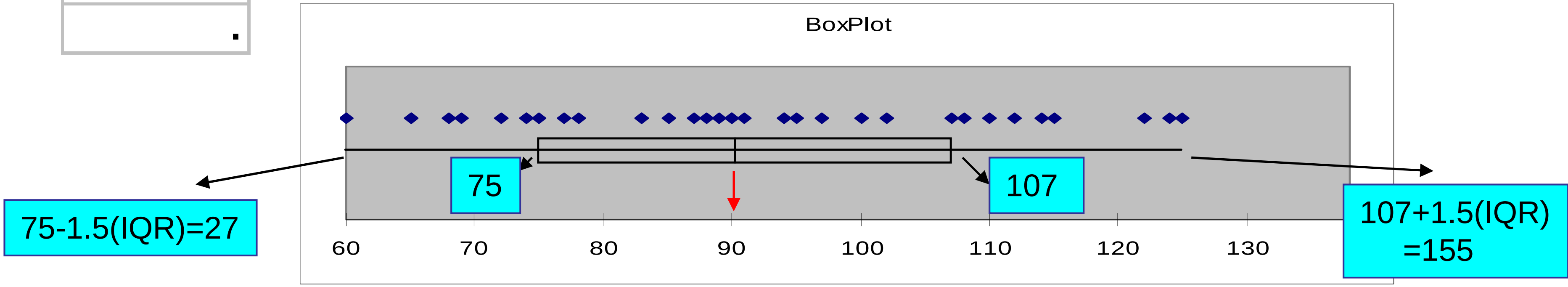
No outliers are found

Box Plot

– The following data give noise levels measured at 36 different times directly outside of Grand Central Station in Manhattan.

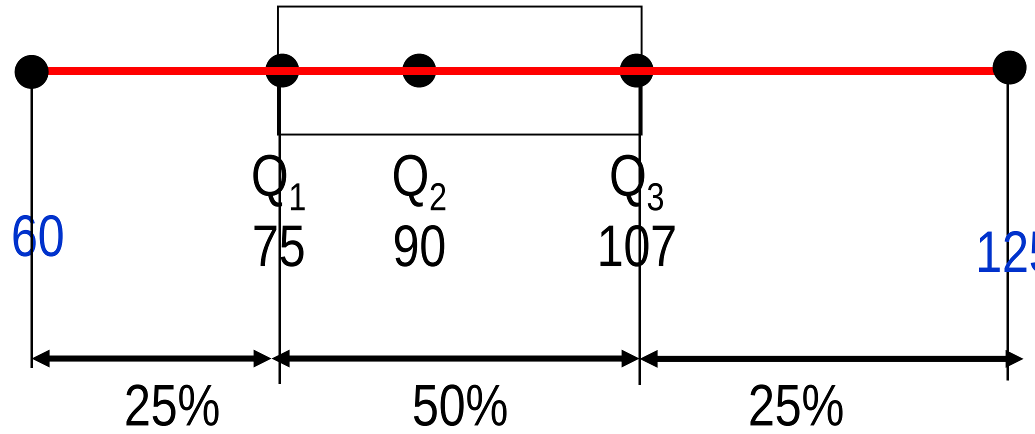
NOISE
82
89
94
110
.
.
.

Smallest = 60		
Q1 = 75		
Median = 90		
Q3 = 107		
Largest = 125		
IQR = 32		
Outliers =		



Box Plot

NOISE - continued

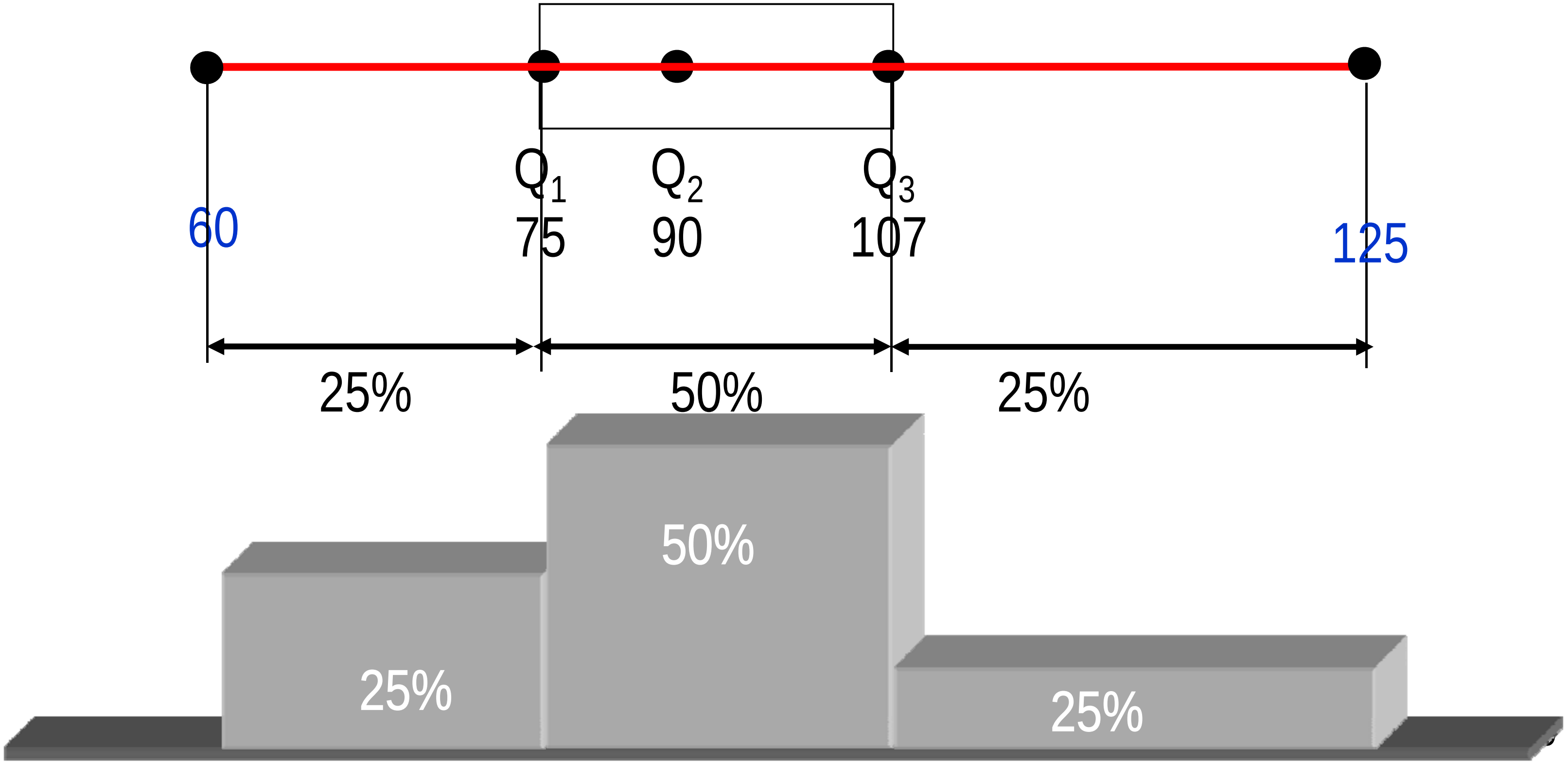


– Interpreting the box plot results
  - The scores range from 60 to 125.
  - About half the scores are smaller than 90, and about half are larger than 90.
  - About half the scores lie between 75 and 107.
  - About a quarter lies below 75 and a quarter above 107.

Box Plot

NOISE - continued

The histogram is positively skewed

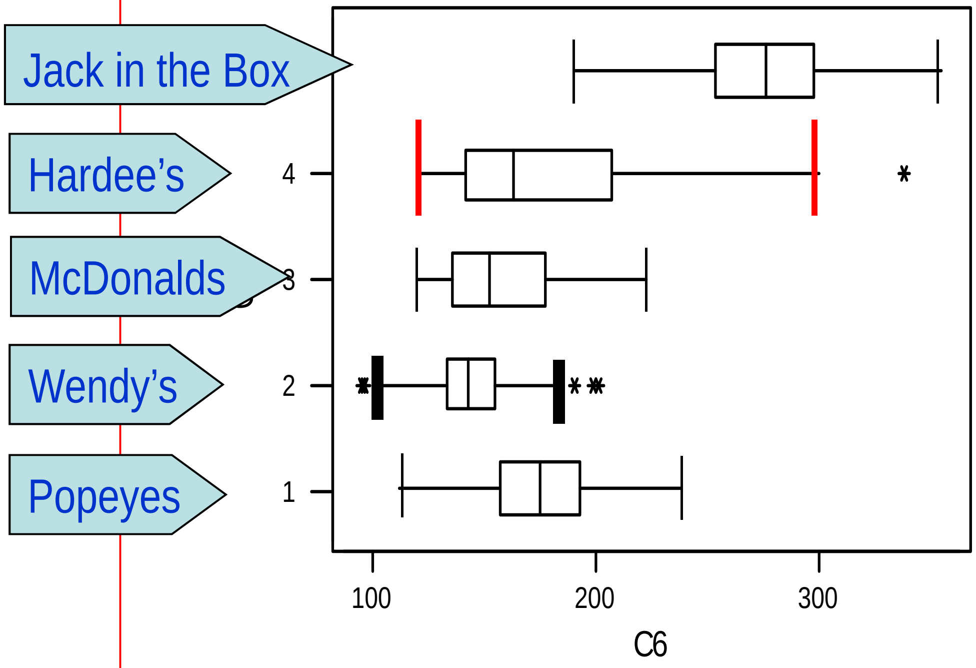


Box Plot

- Example 11
  - A study was organized to compare the quality of service in 5 drive through restaurants.
  - Interpret the results
- Example 11 – solution
  - Minitab box plot (MINITAB 15 demo: http://www.minitab.com/en-US/products/minitab/free-trial.aspx?langType=1033)
  - To download SPSS18, follow the link

  http://bidb.odtu.edu.tr/ccmscontent/articleRead/articleId/articleRead/articleId/390

Box Plot

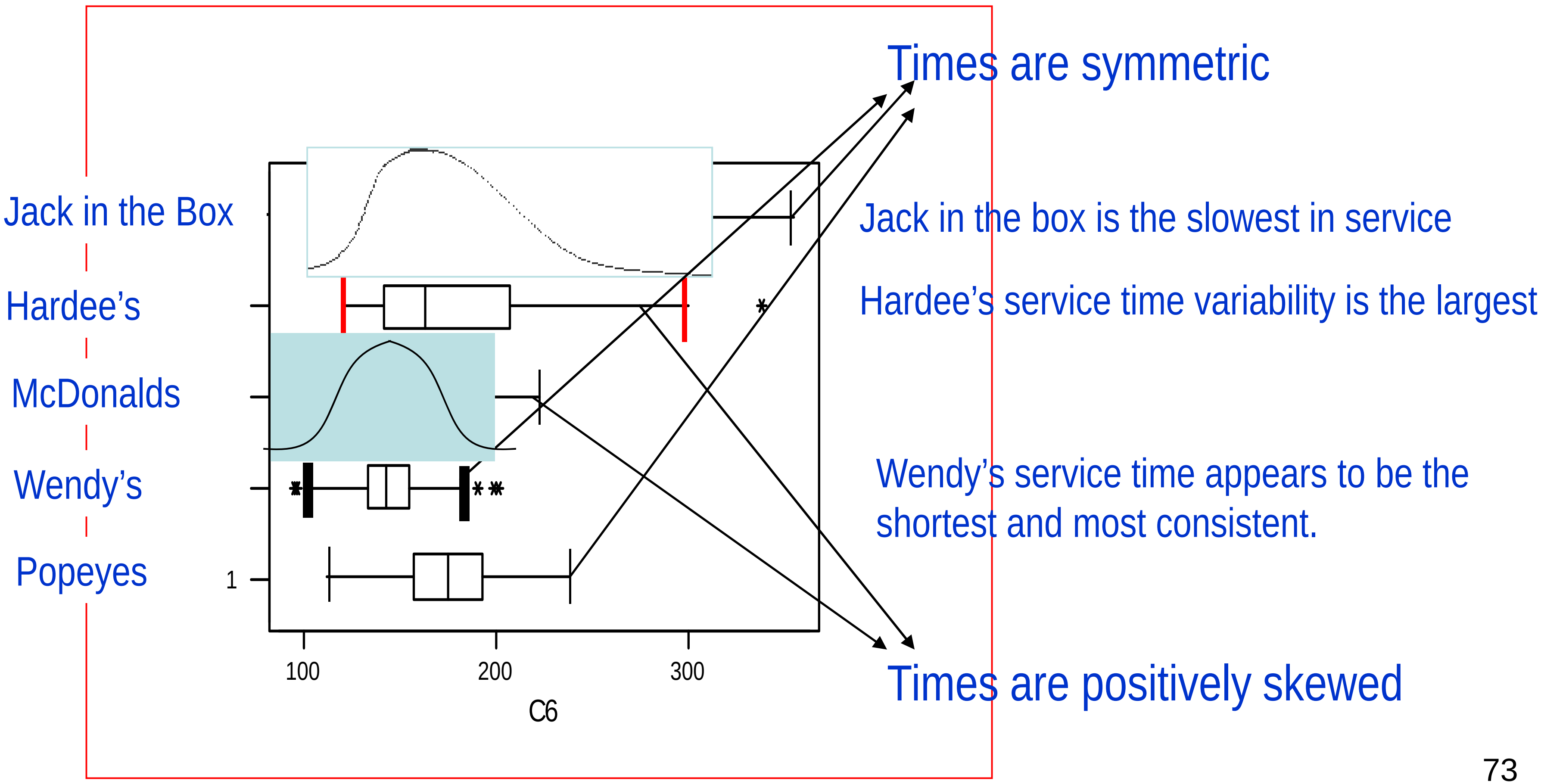


Jack in the box is the slowest in service

Hardee's service time variability is the largest

Wendy's service time appears to be the shortest and most consistent.

Box Plot


Times are symmetric
Jack in the Box
Hardee's
McDonalds
Wendy's
Popeyes
1
100
200
300
C6
Jack in the box is the slowest in service
Hardee's service time variability is the largest
Wendy's service time appears to be the shortest and most consistent.
Times are positively skewed

Paired Data Sets and the Sample Correlation Coefficient

- The covariance and the coefficient of correlation are used to measure the direction and strength of the linear relationship between two variables.
  - *Covariance* - is there any pattern to the way two variables move together?
  - *Coefficient* of c*orrelation* - how strong is the linear relationship between two variables

Covariance

$$\text{Population covariance} = \text{COV}(X, Y) = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$

μ_x (μ_y) is the population mean of the variable X (Y).
N is the population size.

$$\text{Sample covariance} = \text{cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$\bar{x}$ ($\bar{y}$) is the sample mean of the variable X (Y).
n is the sample size.

Covariance

- Compare the following three sets

x_i	y_i	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
2	13	-3	-7	21
6	20	1	0	0
7	27	2	7	14
x=5	y =20			Cov(x,y)=17.5

x_i	y_i	$(x - \bar{x})$	$(y - \bar{y})$	$(x - \bar{x})(y - \bar{y})$
2	27	-3	7	-21
6	20	1	0	0
7	13	2	-7	-14
$\bar{x}$=5	$\bar{y}$ =20			Cov(x,y)=-17.5

x_i	y_i	
2	20	
6	27	Cov(x,y) = -3.5
7	13	
$\bar{x}$=5	$\bar{y}$ =20	

Covariance

- If the two variables move in the same direction, (both increase or both decrease), the covariance is a large positive number.
- If the two variables move in opposite directions, (one increases when the other one decreases), the covariance is a large negative number.
- If the two variables are unrelated, the covariance will be close to zero.

The coefficient of correlation

Population coefficient of correlation

$$\rho = \frac{COV(X,Y)}{\sigma_x \sigma_y}$$

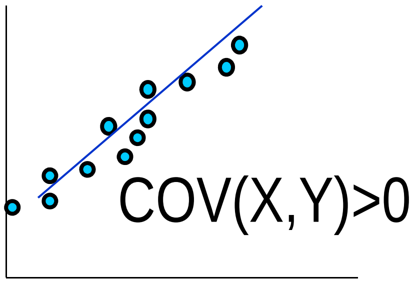
Sample coefficient of correlation

$$r = \frac{cov(X,Y)}{s_x s_y}$$

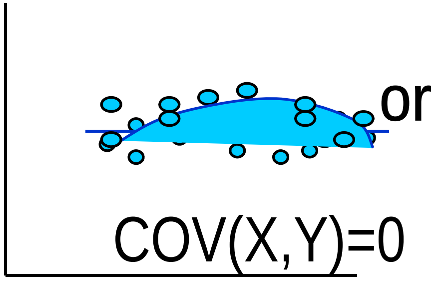
– This coefficient answers the question: How strong is the association between X and Y.

The coefficient of correlation

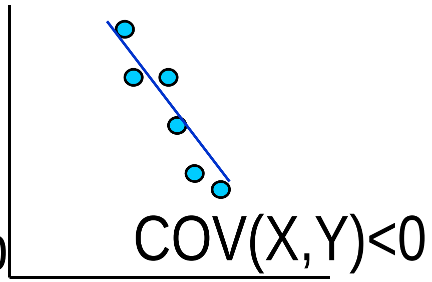
+1 Strong positive linear relationship



ρ or r = 0 No linear relationship



-1 Strong negative linear relationship



The Coefficient of Correlation

- If the two variables are very strongly positively related, the coefficient value is close to +1 (strong positive linear relationship).
- If the two variables are very strongly negatively related, the coefficient value is close to -1 (strong negative linear relationship).
- No straight line relationship is indicated by a coefficient close to zero.