Describing Data
With
Numerical Measures

The world is becoming more and more quantitative. Many professions depend on numerical measurements to make decisions in the face of uncertainty. Statisticians use quantitative abilities, statistical knowledge, and communication skills to work on many challenging problems.

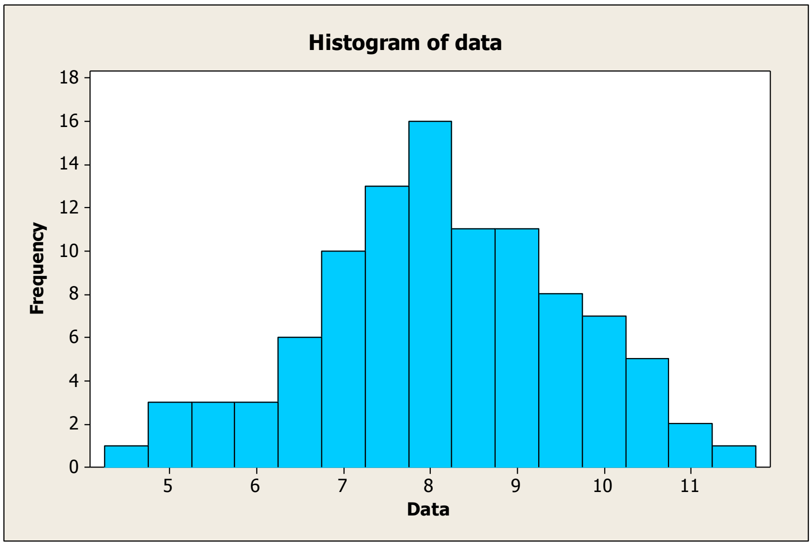
The use of Statistical methods in manufacturing development of computer software and other areas involves the gathering of information or scientific data.

- **Measures of Central Tendency**

- **Measures of Variability**

Measures of Center

In Chapter 2, we introduced histograms, stem and leaf plots, and Dotplots to describe the distribution of a set of measurements on a quantitative variable x.



Data

9.8527	10.8311	9.5529	5.3030	9.8878	8.3768	7.1962
7.9904	8.9761	9.6169	9.6660	7.6093	9.2290	8.2131
8.6503	7.5884	8.1406	6.6059	8.6010	8.9540	10.6517
6.9135	8.7902	7.9193	9.6792	10.7252	6.1061	6.6777
7.6287	8.1044	8.6553	6.1636	8.7772	7.1000	8.3479
8.3065	6.0013	8.0544	4.8606	5.3941	6.8944	9.9000
7.4928	7.8192	6.5690	7.5477	9.4773	6.6852	7.5299
10.3566	8.2186	7.2995	7.3761	8.0283	8.7425	6.7988
7.7562	9.1686	10.2419	5.0469	10.1405	7.1094	10.9362
8.9549	9.3190	8.6137	9.8774	6.8356	6.9017	8.9835
8.3461	10.4107	8.7552	10.3868	4.3714	9.4246	11.2942
9.0678	6.4766	7.3116	8.0751	7.2378	8.1734	7.4499
5.0099	7.6257	7.9819	8.7123	7.2403	8.7219	7.6983
9.8232	6.5852	7.2597	7.9998	5.5064	8.2329	8.7539
7.9505	9.4665					

The given data ranged from a low 4.3714 to a high of 11.2942 with the center of the histogram located in the vicinity of 8.

Let's consider some rules for locating the center of a distribution of measurements.

Definition The arithmetic mean or average of a set of n measurement is equal to the sum the measurements divided by n.

Suppose there are n measurements on the variable x

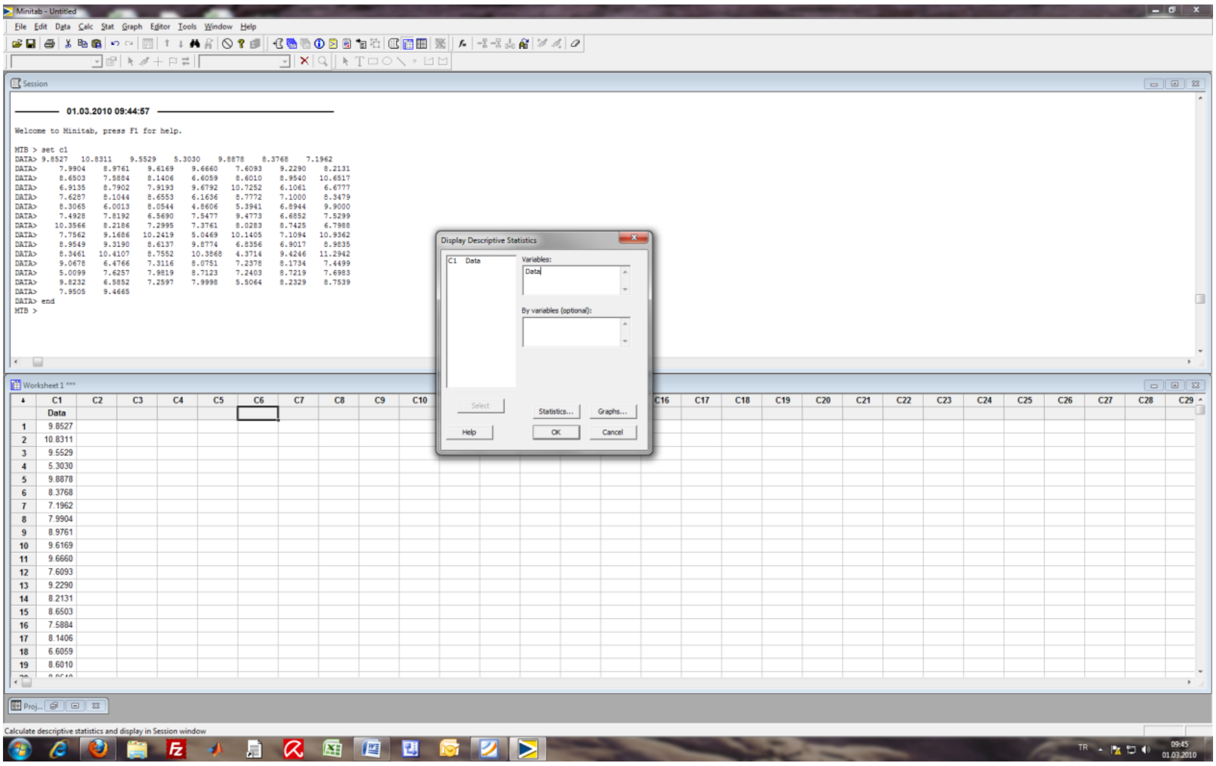
$$x_1, x_2, \ldots, x_n$$

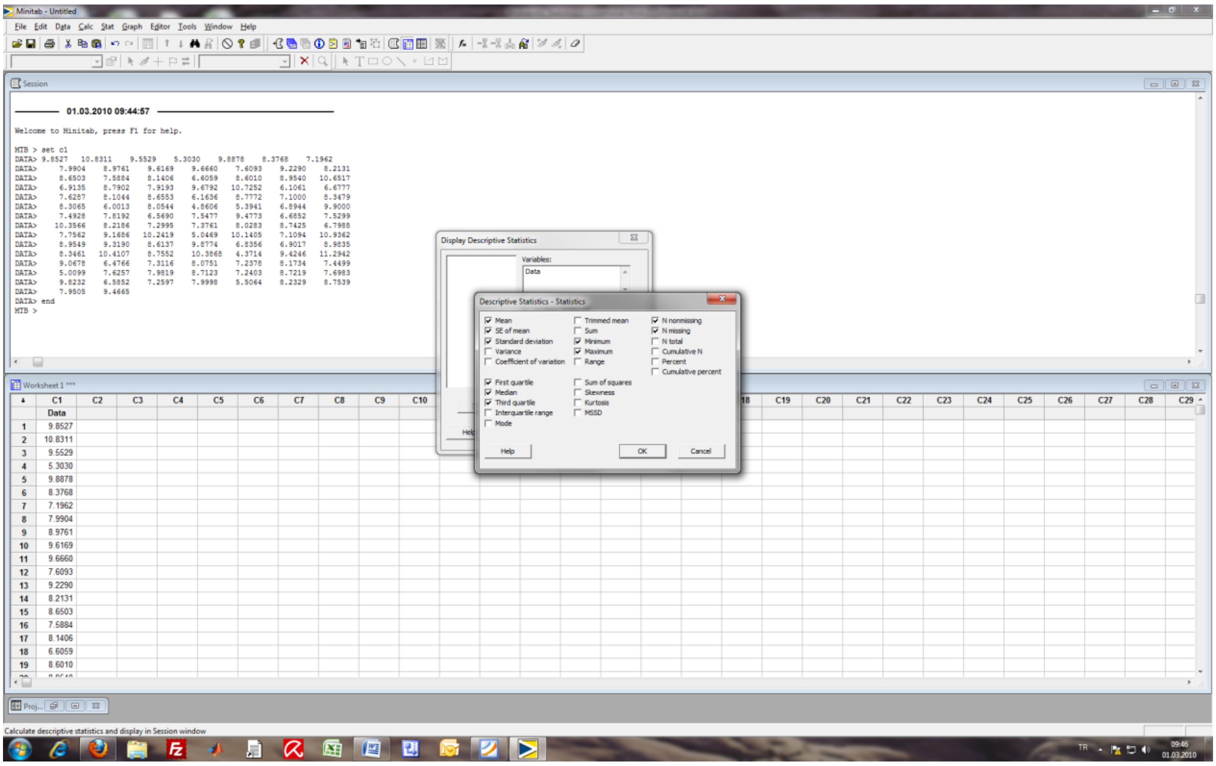
$\sum_{i=1}^{n} x_i$ the sum of all the x measurements

Sample mean $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Population mean μ

$$\bar{x} = \frac{813.669}{100} = 8.137$$





Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
C1	100	0	8.137	0.147	1.471	4.371	7.238	8.122	9.143

Variable	Maximum
C1	11.294

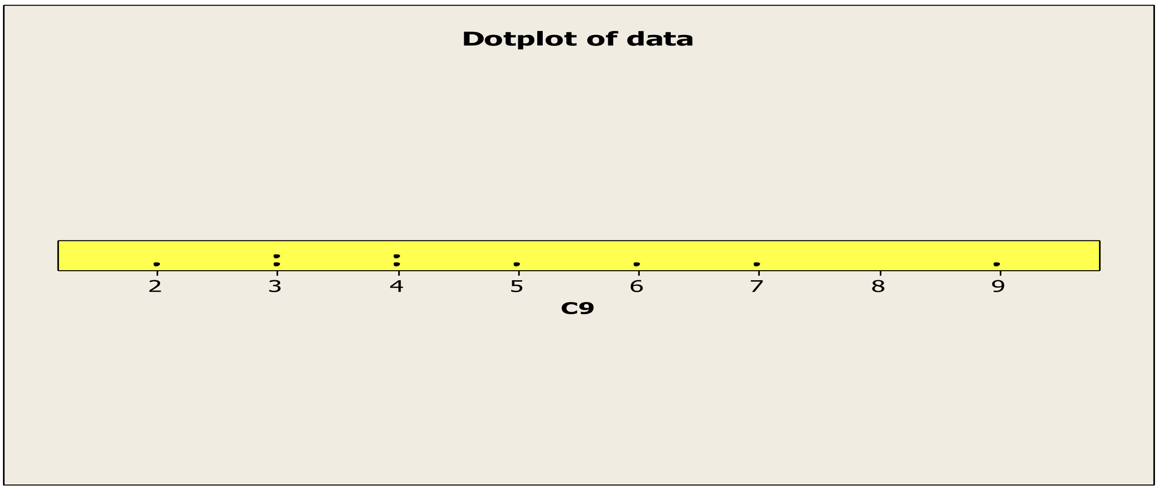
MTB > aver data

Mean of Data

Mean of Data = 8.13669

Data Display

4 2 5 3 6 4 7 3 9



$$\bar{x} = \frac{4+2+5+3+6+4+7+3+9}{9} = 4.7778$$

Mean Data

Mean of Data = 4.77778

Some Properties of arithmetic mean

1. **Uniqueness** For a given set of data there is one and only one arithmetic mean.
2. **Simplicity** The arithmetic mean is easily understood any to compute.
3. Since each and every value in a set of data enters into the computation of the mean, it is **affected by each value. Extreme values** therefore have an influence on the mean and, in some cases can so distort it that it becomes undesirable as a measure of central tendency.

The Mean Computed from Grouped Data

To find the mean we multiply each midpoint by the corresponding frequency, sum these products, and divide by the sum of the frequencies. If the data represent a sample of observations, the computation of the mean may be shown symbolically as

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$$

Where

k= the number of class intervals,

m_i= the midpoint of the i^{th} class interval and,

f_i= the frequency of the i^{th} class interval.

$$n = \sum_{i=1}^{k} f_i$$

Class Interval	Midpoint m_i	Frequency f_i	$m_i f_i$
10-19	14.5	5	72.5
20-29	24.5	19	465.5
30-39	34.5	10	345.0
40-49	44.5	13	578.5
50-59	54.5	4	218.0
60-69	64.5	4	258.0
70-79	74.5	2	149.0
Total		57	2086.5

$$\overline{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i} = \frac{2086.5}{57} = 36.6$$

Geometric mean

The geometric mean is an average that is useful for sets of numbers that are interpreted according to their product and not their sum (as is the case with the arithmetic mean). For example rates of growth.

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Harmonic mean

The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, for example speed (distance per unit of time).

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

Weighted mean

The weighted mean is used, if one wants to combine average values from samples of the same population with different sample sizes:

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

```
macro
#*****************************************************
#   written by: Serdar Korukoglu                          *
#   geometric mean                                        *
#   Purpose: This macro calculates  geometric  means for each *
 #   row in a set of columns.  The answers are stored in sonuc   *
#    column.
#****************************************************
rgmean x.1-x.n sonuc
mcolumn x.1-x.n sonuc
mconstant con1 i j total n
noecho
brief 0
let con1=count(x.1)
do i=1:con1
let total=0.0
do j=1:n
let total=total+loge(x.j(i))
enddo
let total=total/n
let total=exp(total)
let sonuc(i)=total
enddo
endmacro
```

```
macro
#*****************************************************
#    written by: Serdar Korukoglu                       *
#    geometric mean                                     *
#    Purpose: This macro calculates  geometric  means for each  *
 #    row in a set of columns.  The answers are stored in sonuc   *
#     column.
#*****************************************************
rgmean1 x.1-x.n sonuc
mcolumn x.1-x.n y.1-y.n sonuc
mconstant con1 i j total n
let con1=count(x.1)
do i=1:n
let y.i=loge(x.i)
enddo
rmean y.1-y.n sonuc
let sonuc=exp(sonuc)
endmacro
```

```
macro
#*****************************************************
#    written by: Serdar Korukoglu                        *
#    harmonic mean                                       *
#    Purpose: This macro calculates  harmonic  means for each *
 #    row in a set of columns.  The answers are stored in sonuc   *
#     column.
#****************************************************
rhmean x.1-x.n sonuc
mcolumn x.1-x.n y.1-y.n sonuc
mconstant con1 i j total n
let con1=count(x.1)
do i=1:n
let y.i=1/(x.i)
enddo
rsum y.1-y.n sonuc
let sonuc=n/sonuc
endmacro
```

```
macro
rowmeans x.1-x.n son;
type a.
mcolumn x.1-x.n y.1-y.n son
mconstant con1 i j a
default a=1
if a=1
rmean x.1-x.n son
endif
if a=2
%rgmean1 x.1-x.n son
endif
if a=3
%rhmean x.1-x.n son
endif
if a < 1 or a >3
write "Type must be valid"
exit
endif
endmacro
```

Median

***The median* m of a set of n measurements is the value of x that falls in the middle position when the measurements are ordered from smallest to largest.**

In probability theory and statistics, the **median** is a number that separates the higher half of a sample, a population, or a probability distribution from the lower half. It is the middle value in a distribution, above and below which lie an equal number of values. This states that 1/2 of the population will have values less than or equal to the ***median*** and 1/2 of the population will have values equal to or greater than the median.

To find the median of a finite list of numbers, arrange all the observations from lowest value to highest value and pick the middle one. If there are an even number of observations, one often takes the mean of the two middle values

Find the median for the set of measurement

2, 9, 11, 5, 7

Rank the n=5 measurements from smallest to largest:

2 5 7 9 11

Find the median for the set of measurement

2, 9, 11, 5, 7, 32

Rank the n=6 measurements from smallest to largest:

2 5 7 9 11 32

m=(7+9)/2=8

Data Display

Data

2 9 11 5 7 32

Median of data

Median of data = 8

The value 0.5(n+1) indicates the position of the median in the ordered data set. If the position of the median is a number that ends in the value .5, we need to average the two adjacent values.

Some Properties of Median

1. **Uniqueness** As is true with the mean, there is only one median for a given set of data.
2. **Simplicity** The median is easy to calculate.
3. It is **<u>not as drastically affected</u>** by extreme values as in the mean.

Do It Yourself!: How Extreme Values Affect the Mean and Median

http://metalab.uniten.edu.my/~abdrahim/matb344/beaver/dotInfluence.html

Use your mouse to drag the data point marked in green. Watch how the mean and the median change as this point takes on new values.

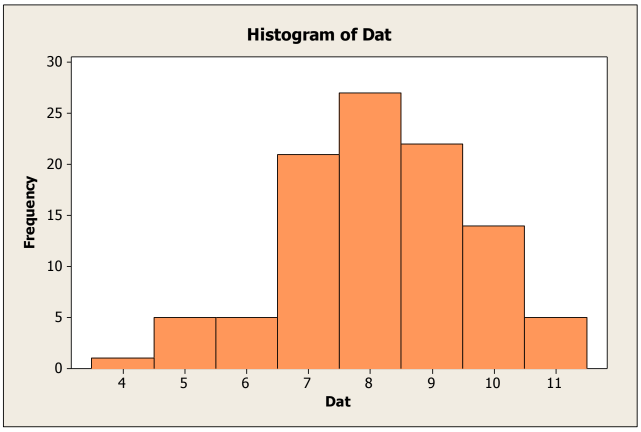
The Mode

The mode of a set of values is that value which occurs most frequently. If all the values are different there is no mode; on the other hand, a set of values may have more than one mode.

The mode is generally used to describe large data sets, whereas the mean and median are used for both large ans small data sets.

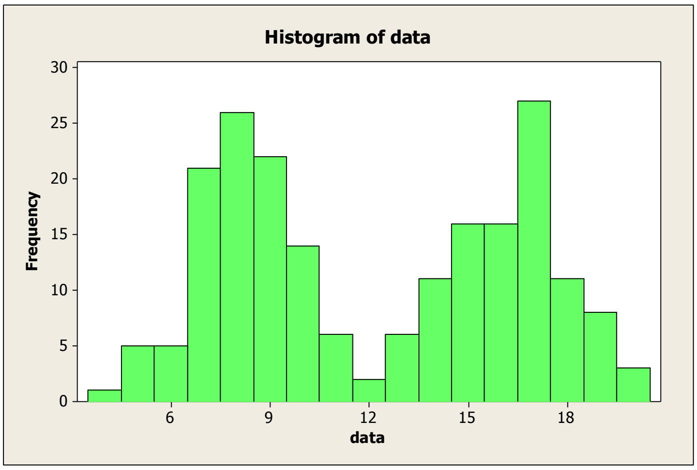
The mode may be used for describing qualitative data.

```
Data Display
Data
   2   9  11   5   7  32
MTB > Describe 'data';
SUBC>   Mode.
 Descriptive Statistics: data
              N for
Variable  Mode   Mode
data         *      0
```



Histogram of Dat
MTB > desc c2;
SUBC> mode.
Descriptive Statistics: Dat

	N for	
Variable	**Mode**	**Mode**
Dat	**8**	**27**



Bimodal Distribution

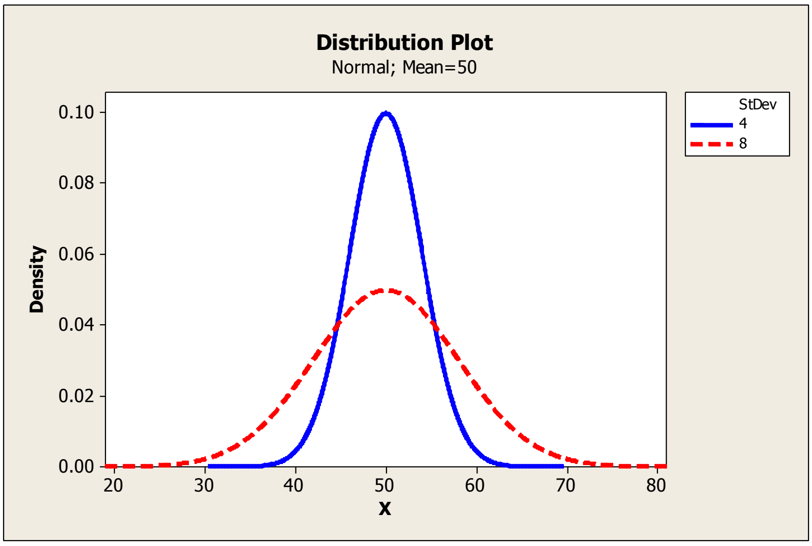
Descriptive Statistics: data

	N for	
Variable	**Mode**	**Mode**
data2	**8; 17**	**27**

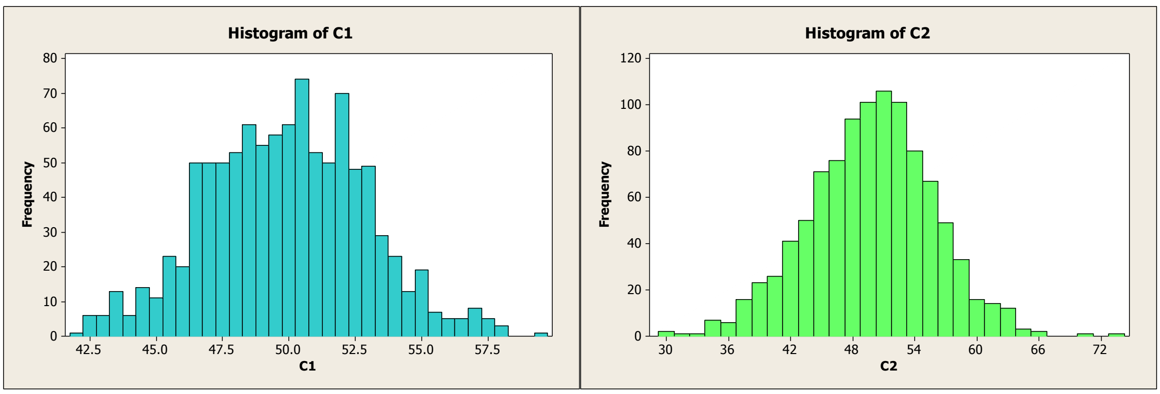
Sometimes bimodal distributions reflect a **mixture of measurements**.

Measures of Variability (Dispersion)

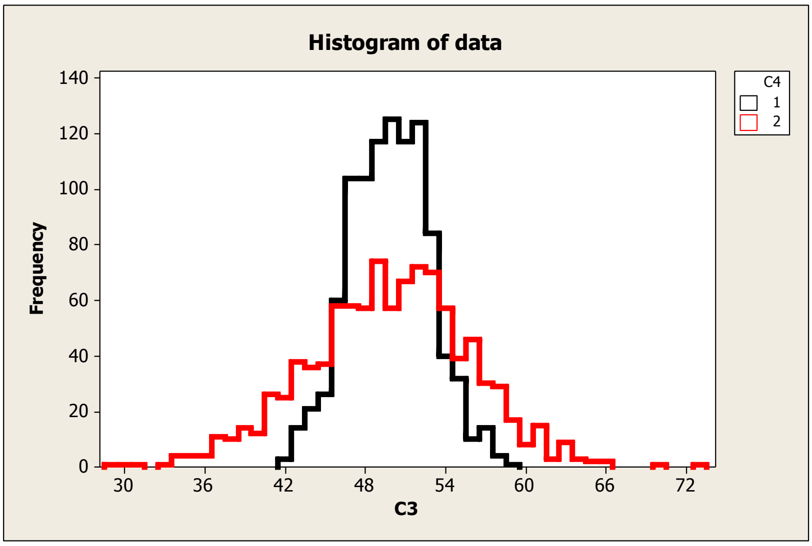
Data sets may have the same center but look different because of the way the numbers spread out from the center.



Distributions with equal means but unequal variability



Same center but different spread



Variability or dispersion is a very important characteristic of data. Variability of a set of observations refers to the variety that they exhibit.

If all the values are the same, there is no variability or dispersion.

- **The amount of variability may be small, when the values, though different, are close together.**
- **If he values are widely scattered, the dispersion is greater.**

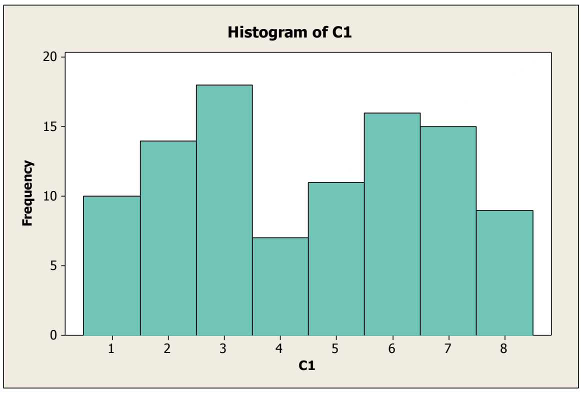
Measure of variability can help us create a mental picture of spread of the data.

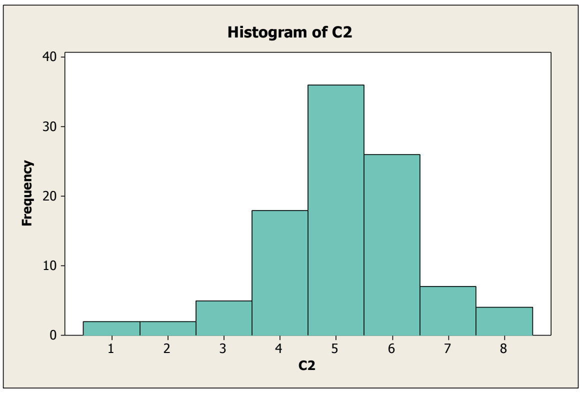
Range

Definition The range R, ***of set of n measurement is defined as the difference between the largest and the smallest measurements.***

In descriptive statistics, the **range** is the length of the smallest interval which contains all the data. It is calculated by subtracting the smallest observations from the greatest and provides an indication of statistical dispersion.

It is measured in the same units as the data. Since it only depends on two of the observations, ***it is a poor and unrobust measure of dispersion except when the sample size is large.***





Distributions with equal range but unequal variability

MTB > range c1
Range of C1
Range of C1 = 7
MTB > range c2
Range of C2
Range of C2 = 7

Is there a measure of variability that is more sensitive than the range?

When the values of a set of observations lie close to their mean, the dispersion is less than when they are scattered over a wide range. Since this is true, it would be intuitively appealing if we could measure dispersion relative to the scatter of the values about their mean. Such a measure is realized in what is known as the variance.

Variance

Definition: ***The Variance of a population of N measurements is the average of the squares of the deviations of the measurements about their mean μ. The population variance is denoted by σ^2 and is given by the formula***

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Definition: ***The Variance of a sample of n measurements is the sum of the squared deviations of the measurements about their mean $\bar{x}$ divided by (n-1). The sample variance is denoted by s^2 and is given by the formula***

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

The variance represents squared units and, therefore, is not an appropriate measure of dispersion when we wish to express this concept in terms of original units.

To obtain a measure of dispersion in original units, we merely take the square root of the variance.

Definition: ***The standard deviation of a set of measurements is equal to the positive square root of the variance.***

Notation

n :number of measurements in the sample

N: Number of measurements in the population

s^2 :**Sample variance** σ^2 :**population variance**

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

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

The Computing Formula for Calculating Variance

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

$$s^2 = \frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}$$

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

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

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

Example:

x	x^2
7	49
4	16
6	36
4	16
2	4
1	1
5	25

```
let c2=c1^2
MTB > sum c1
 Sum of x
Sum of x = 29
MTB > sum c2
Sum of x^2
Sum of x^2 = 147
MTB > name k1 "stan_dev"
MTB > let k1=sqrt(((count(c1)*sum(c2)-sum(c1)^2))/(count(c1)*(count(c1)-1)))
MTB > print k1
Data Display
stan_dev    2.11570
MTB > stan c1
 Standard Deviation of x
Standard deviation of x = 2.11570
```

Or directly

```
MTB > stan c1
 Standard Deviation of x
Standard deviation of x = 2.11570
```

Some Properties of Variance

- The value of s is always greater than or equal to zero.
- The larger the value of s^2 or s, the greater the variability of the data set.
- If s^2 or s is equal to zero, all the measurements must have the same value.
- In order to measure the variability in the same units as the original observations, we compute the standard deviation s.

The Variance –Grouped Data

In calculating the variance and standard deviation from grouped data we assume that all values falling into a particular class interval are located at the midpoint of the interval. The variance of a sample, then, is given by

$$s^2 = \frac{\sum_{i=1}^{k} (m_i - \overline{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1}$$

The following computing formula for the sample variance may be preferred:

$$s^2 = \frac{n\sum_{i=1}^{k} m_i^2 f_i - \left(\sum_{i=1}^{k} m_i f_i\right)^2}{n(n-1)}$$

where

$$n = \sum_{i=1}^{k} f_i$$

Class Interval	Midpoint m_i	Frequency f_i	$m_i f_i$	$m_i^2 f_i$
10-19	14.5	5	72.5	1051.3
20-29	24.5	19	465.5	11404.8
30-39	34.5	10	345.0	11902.5
40-49	44.5	13	578.5	25743.3
50-59	54.5	4	218.0	11881.0
60-69	64.5	4	258.0	16641.0
70-79	74.5	2	149.0	11100.5
Total		57	2086.5	89724.3

$$s^2 = \frac{n\sum_{i=1}^{k} m_i^2 f_i - \left(\sum_{i=1}^{k} m_i f_i\right)^2}{n(n-1)}$$

$$= \frac{57(89724.3) - (2086.5)^2}{57(57-1)} = 238.35$$

Data Display

71.8483	66.3890	66.2304	62.6633	56.3896	72.7303	59.3119
58.2468	66.6568	81.6851	59.6392	67.1141	89.6984	66.0986
68.9907	55.3811	77.0741	56.1753	45.8898	61.6749	63.3153
56.2484	71.2606	78.2773	69.5257	97.3577	74.7657	62.2872
58.7144	65.8996	72.5456	64.6450	69.3609	65.8606	67.4463
72.8546	65.9247	83.5393	51.2912	73.1480	69.5310	55.0728
75.3135	85.7150	84.1734	62.0618	79.1964	76.6877	63.4280
68.5589	50.8236	72.1436	74.2940	87.2444	67.6740	94.4568
70.7153	76.8975	74.2013	67.1729	66.6304	77.2414	64.0028
61.4217	84.1182	88.0001	77.0757	77.8508	73.9554	88.4847
90.4772	66.2349	74.3869	61.3872	85.1908	73.9719	79.7538
68.2543	81.0005	73.6131	68.8121	73.2348	96.5489	81.8984
79.5355	56.4083	76.2816	89.1318	93.5616	68.5015	78.4543
84.6485	70.7936	77.5622	60.1408	78.2083	58.1339	68.8597
56.7745	60.8253					

Direct computation of standard deviation from original data

MTB > stan c1
Standard Deviation of C1

Standard deviation of C1 = 10.7733

MTB > GSTD
*** NOTE * The character graph commands are obsolete.**
*** NOTE * Standard Graphics are now enabled, and Professional Graphics are**
*** disabled. Use the GPRO command when you want to re-enable**
*** Professional Graphics.**
MTB > hist c1;
SUBC> incr 10.
Histogram
Histogram of C1 N = 100

```
Midpoint     Count
   50.0        3  ***
   60.0       24  ************************
   70.0       39  ***************************************
   80.0       22  **********************
   90.0       10  **********
  100.0        2  **
```

Computation from grouped data

m_i	f_i	$m_i f_i$	$m_i^2 f_i$
50	**3**	150	7500
60	**24**	1440	86400
70	**39**	2730	191100
80	**22**	1760	140800
90	**10**	900	81000
100	**2**	200	20000
Total	**100**	**7180**	**526800**

$$s^2 = \frac{n\sum_{i=1}^{k} m_i^2 f_i - \left(\sum_{i=1}^{k} m_i f_i\right)^2}{n(n-1)}$$

$$= \frac{100(526800) - (7180)^2}{100(100-1)} = 113.898$$

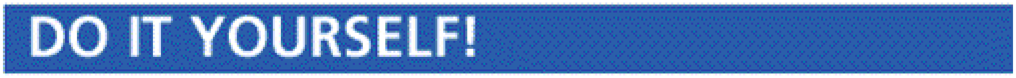
$$s = \sqrt{113.898} = 10.6723$$

Compare this with the following

Standard deviation of C1 = 10.7733

Why Divide by (n-1)?

You may wonder why you need to divide by (n-1) rather than n when computing the sample variance. Just as we used the sample mean



***Do It Yourself!:* Why Divide by *n - 1*?**

http://metalab.uniten.edu.my/~abdrahim/matb344/beaver/sampleStDev.html

The reason for dividing by n-1 rather than n is the theoretical consideration referred to as *degrees of freedom*. In computing the variance, we say that we have n-1 *degrees of freedom*.

We reason as follows:

The sum of the deviations of the values from their mean is equal to zero. If, then, we know the values of n-1 of the deviations from the mean, we know the nth one, since it is automatically determined because of the necessity for all n values to add to zero.

Expected Value of S^2

The following is a proof that the formula for the sample variance, S^2, is unbiased. Recall that it seemed like we should divide by n, but instead we divide by n-1. Here's why.

First, recall the formula for the sample variance:

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

Now, we want to compute the expected value of this:

$$\begin{aligned} E[S^2] &= E\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}\right] \\ &= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \end{aligned}$$

Now, let's multiply both sides of the equation by n-1, just so we don't have to keep carrying that around, and square out the right side, just like we did with that shortcut formula for SSX, above.

$$\begin{aligned} (n-1)E[S^2] &= E\left[\sum_{i=1}^{n}x_i^2 - 2\bar{x}x_i + \bar{x}^2\right] \\ &= E\left[\sum x_i^2\right] - E\left[\sum 2\bar{x}x_i\right] + E\left[\sum \bar{x}^2\right] \\ &= E\left[\sum x_i^2\right] - E\left[2\bar{x}\sum x_i\right] + E\left[\bar{x}^2\sum 1\right] \end{aligned}$$

Now, if you think about it, it's clear that $\sum x_i = n\bar{x}$, so we can rewrite the middle term on the RHS in terms of $\bar{x}$:

$$\begin{aligned}
(n-1)E[S^2] &= E\left[\sum x_i^2\right] - E\left[2\bar{x}(n\bar{x})\right] + E\left[\bar{x}^2 \sum 1\right] \\
&= \sum E\left[x_i^2\right] - nE\left[\bar{x}^2\right] \\
&= nE\left[x_i^2\right] - nE\left[\bar{x}^2\right] \\
\frac{n-1}{n}E[S^2] &= E\left[x_i^2\right] - E\left[\bar{x}^2\right]
\end{aligned}$$

Let's write that again as a numbered equation:

$$\frac{n-1}{n}E[S^2] = E\left[x_i^2\right] - E\left[\bar{x}^2\right] \qquad (1)$$

Unfortunately, the expected value of the square of something is not equal to the square of the expected value, so we seem to have hit an impasse with both terms on the RHS. But, we're not out of tricks yet. Each of those terms is an expected value of something squared: a second moment. Let's use the trick about moments that we saw above.

First, let *Y* be the random variable defined by the sample mean, $\bar{x}$. We're trying to figure out the expected value of its square.

$$
\begin{aligned}
E[Y^2] = E[\bar{x}^2] &= \mathrm{var}[Y] + E[Y]^2 \\
&= \mathrm{var}\left[\frac{1}{n}\sum x_i\right] + \mu^2 \\
&= \frac{1}{n^2}\mathrm{var}\left[\sum x_i\right] + \mu^2 \\
&= \frac{1}{n^2}\sum \mathrm{var}[x_i] + \mu^2 \\
&= \frac{1}{n^2}\sum \sigma^2 + \mu^2 \\
&= \frac{1}{n^2}(n\sigma^2) + \mu^2 \\
&= \frac{1}{n}\sigma^2 + \mu^2
\end{aligned}
$$

We can substitute this stuff for the second term on the RHS of equation 1. Also, note that the first term on the RHS of equation 1 is the second moment of X, so that can also be re-written. Doing both substitutions gives us:

$$
\begin{aligned}
\frac{n-1}{n}E[S^2] &= [\sigma^2 + \mu^2] - [\frac{1}{n}\sigma^2 + \mu^2] \\
&= \sigma^2 - \frac{1}{n}\sigma^2 \\
E[S^2] &= \sigma^2
\end{aligned}
$$

This is why S^2 with the n-1 denominator is an unbiased estimator.

Tchebysheff Theorem

Given a number k greater than 1 and a set of n measurements, at least $[1-(1/k^2)]$ of the measurements will lie within k standard deviations of their mean.

Example

The mean and variance of a sample of n= 25 measurements are 75 and 100, respectively. Use Tchebysheff's Theorem to describe the distribution of measurements.

$\bar{x} = 75$, $s^2 = 100$ and $s = 10$.

The distribution of measurements is centered about $\bar{x} = 75$, and **Tchebysheff's Theorem** states**:**

- At least $[1-(1/2^2)] = 3/4$ of the 25 measurements lie in the interval $\bar{x} \pm 2s = 75 \mp 2(10)$ - that is, 55 to 95.
- At least $[1-(1/3^2)] = 8/9$ of the 25 measurements lie in the interval $\bar{x} \pm 3s = 75 \mp 3(10)$ -that is, 45 to 105.

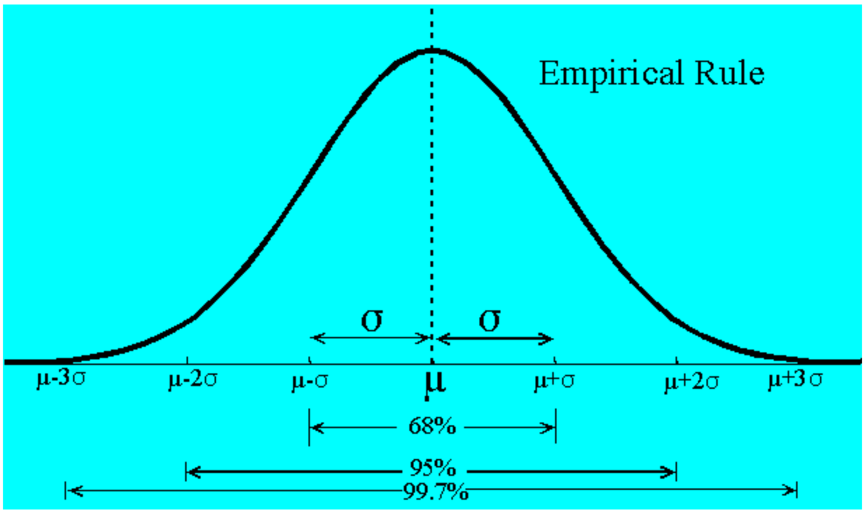
Since Tchebysheff's Theorem applies to any distribution, it is very conservative. This is why we emphasize "at least $[1-(1/k^2)]$ **in this Theorem.**

Empirical Rule

Given a distribution of measurements that is approximately <u>mound-shaped</u>.

- ***The interval*** $(\mu \pm \sigma)$ ***contains approximately 68% of the measurements.***
- ***The interval*** $(\mu \pm 2\sigma)$ ***contains approximately 95% of the measurements.***
- ***The interval*** $(\mu \pm 3\sigma)$ ***contains approximately 99.7% of the measurements.***

The mound-shaped distribution shown in the following figure is commonly known as the **normal distribution** and will be discussed in detail in later sections.



Example

The mean and standard deviation are found to be 12.8 and 1.7 respectively. Describe the sample data using the Empirical Rule.

$\bar{x} = 12.8$, $s = 1.7$.

To describe the data

$(\bar{x} \pm s) = 12.8 \pm 1.7$ **11.1 to 14.5**

Approximately 68% of the measurements fall into the interval 11.1 to 14.5.

$(\bar{x} \pm 2s) = 12.8 \pm 2(1.7)$ **9.4 to 16.2**

Approximately 95% of the measurements fall into the interval 9.4 to 16.2.

$(\bar{x} \pm 3s) = 12.8 \pm 3(1.7)$ **7.7 to 17.9**

Approximately 99.7% of the measurements fall into the interval 7.7 to 17.9.

Exercises:

Compare TCHEBYSHEFF'S THEOREM and EMPIRICAL RULE on the LAB.

When you use these two tools, for describing a set of measurements, Tchebysheff's Theorem will always be satisfied, but it is a very conservative estimate of the fraction of measurements that fall into a particular interval. If it is appropriate to use the Empirical Rule (mound-shaped data), this rule will give you a more accurate estimate of the fraction of measurements that fall into the interval.

A check on the calculation of s

Tchebysheff Theorem and the Empirical rule can be used to detect gross errors in the calculation of s. Very rough approximation can be useful in checking for large errors in calculation of s. If the range R, is about four standard deviations, or 4s, we can write

$$R \approx 4s \quad \text{or} \quad s \approx R/4$$

Data Display

C1
34 48 43 19 28 8 11 7 4 33 40 6 40 9 1
30 16 14 18 25

Standard deviation of C1 = 14.4990

MTB > range c1

Range of C1

Range of C1 = 47

Measures of Relative Standing

Sometimes we need to know <u>the position of one observation to others in a set of data</u>. The mean and standard deviation of the scores can be used to calculate a z-score, which measures the relative standing of a measurement in a data set.

Definition

The sample z-score is a measure of relative standing defined by

$$z_{score} = \frac{x - \bar{x}}{s} \qquad z_i = \frac{x_i - \bar{x}}{s}$$

A z-score measures the distance between an observation and the mean, measured in nits of standard deviation. The z-score is a valuable tool for determining whether a particular observation is <u>likely</u> to occur quite frequently or whether it is <u>unlikely</u> and might be considered an <u>outlier</u>.

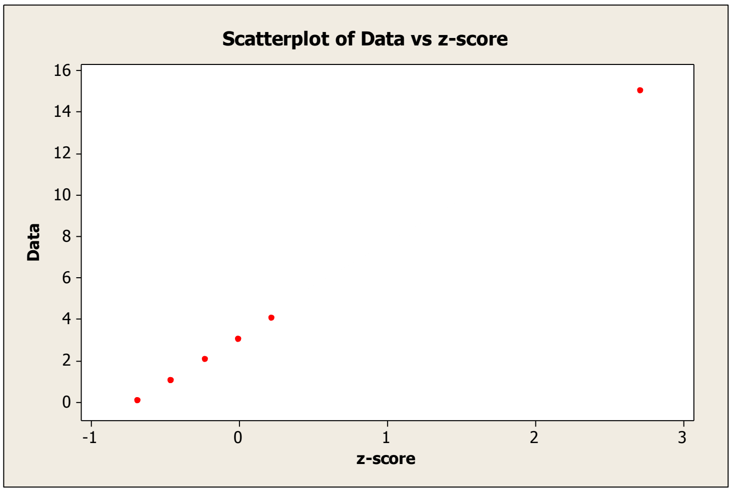
According to Tchebysheff's Theorem and the Empirical Rule

- **At least 75% and more likely 95% of the observations lie within two standard deviations of their mean: their z-scores are between -2 and +2.** *Observations with z-scores exceeding* ***2*** *in absolute value happen less than* ***5%*** *of the time and are considered* ***somewhat unlikely*****.**

- **At least 89% and more likely 99.7% of the observations lie within three standard deviations of their mean: their z-scores are between -3 and +3.** *Observations with z-scores exceeding* ***3*** *in absolute value happen less than* ***1%*** *of the time and are considered* ***very unlikely*****.**

Example:

Data	z-score
1	-0.45227
1	-0.45227
0	-0.67840
15	2.71360
2	-0.22613
3	0.00000
4	0.22613
0	-0.67840
1	-0.45227
3	0.00000



Standard deviation of Data = 4.42217

Mean of Data = 3

The z-score of the suspected outlier x=15, is calculated as

$$z_{score} = \frac{x - \bar{x}}{s} = \frac{15 - 3}{4.42217} = 2.71360$$

Hence, the measurement x=15 lies 2.71 standard deviations above the sample mean 3. Although the z-score does not exceed 3, it is close enough so that you might suspect that x=15 is an outlier.

The pth percentile

A percentile is another measure of relative standing and is most often used for **large data set** (University Entrance Exam). Percentiles are **not very useful for small data sets**.

Definition

A set of n measurements on the variable x has been arranged in order of magnitude. The pth-percentile is the value of x that is greater than p% of the measurements and is less than the remaining (100-p) %.

Example: Suppose you have been notified that your score of 470 on the verbal Graduate Record Examination placed you at the 70^{th} percentile in the distribution of scores. Where does your score of 470 stand in relation to the scores of others who took the examination?

Scoring the 70^{th} percentile means that 70% of all the examination scores were lower than your score and 30% were higher.

Calculating Sample Quartiles

Definition ***A set of n measurements on the variable x has been arranged in order of magnitude. The lower quartile (first quartile),*** Q_1***, is the value of x that is greater than one-fourth of the measurements and is less than the remaining three-fourths. The second quartile is median. The upper quartile (third quartile),*** Q_3***, is the value of x that is greater than three-fourths of the measurements and is less than the remaining one-fourth.***

- ***When the measurements are arranged in order of magnitude, the lower quartile,*** Q_1***, is the value of x in position 0.25(n+1), and the upper quartile ,*** Q_3***, is the value of x in position 0.75(n+1).***
- ***When 0.25(n+1) and 0.75(n+1) are not integers, the quartiles are found by interpolation, using the values in two adjacent positions.***

Definition ***The interquartile range (IQR) for a set of measurements is the difference between the upper and lower quartiles; that is***

$$IQR=Q_3-Q_1$$

The five number summary consist of the smallest number, the lower quartile, the median, the upper quartile, and the largest number, presented

Example: Find the lower and upper quartiles for this set of measurements:

16, 25, 4, 8, 11, 13, 20, 8, 11 ,9

Rank the n=10 measurements from smallest to largest

4, 8, 9, 11, 11, 13 16, 18, 20, 25

Calculate

Position of Q_1=.25(n+1)=.25(10+1)=2.75

Position of Q_3=.75(n+1)=.75(10+1)=8.25

$$Q_1=8+0.75(9-8) = 8.75$$

$$Q_3=18+0.25(20-18) = 18.5$$

$$IQR=Q_3-Q_1=18.50-8.75= 9.75$$

The five number summary

4, 8.75 ,12.0 ,18.5 ,25.0

```
MTB > Describe 'data';
SUBC>   QOne;
SUBC>   Median;
SUBC>   QThree;
SUBC>   IQRange;
SUBC>   Minimum;
SUBC>   Maximum;
SUBC>   NMissing.
```

Descriptive Statistics: data

Variable	N*	Minimum	Q1	Median	Q3	Maximum	IQR
data	0	4.00	8.75	12.00	18.50	25.00	9.75

The Coefficient of Variation

The standard deviation is useful as a measure of variation within a given set of data. When one desires to compare the dispersion in two sets of data, however, comparing the two standard deviations may lead to fallacious result. It may be that the two variables involved are measured in different units.

What is the needed in situations like these is a measure or relative variation rather than absolute variation. Such a measure is found in the ***coefficient of variation***, which ***expresses the standard deviation as a percentage of the mean.***

$$C.V. = \frac{s}{\bar{x}}(100)$$

	Sample 1	Sample 2
Age	25 years	11 years
Mean weight	72 Kg	40 Kg
Standard Deviation	10 Kg	10 Kg

We have for the 25 years old **C.V.=10/72(100)=13.8889**

And for the 11- years olds **C.V.=10/40(100)=25.**

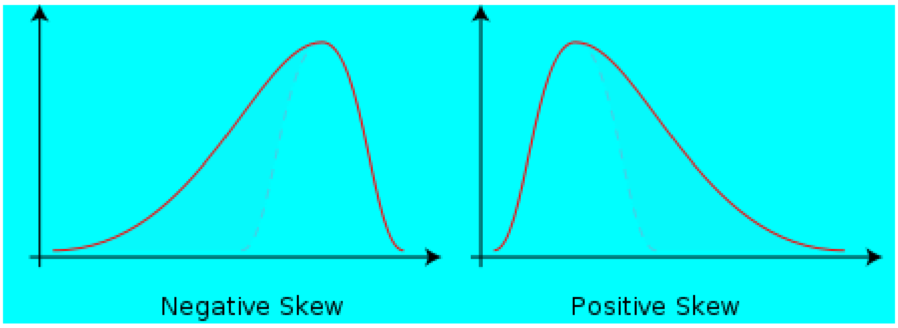
If we compare these results we get quite a different impression.

Skewness

Definition

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

The degree to which a data set is not symmetrical. Like many other basic statistics, skewness can help you establish an **initial understanding of your data**. You can evaluate skewness via a graph (like a histogram) or through the skewness statistic.

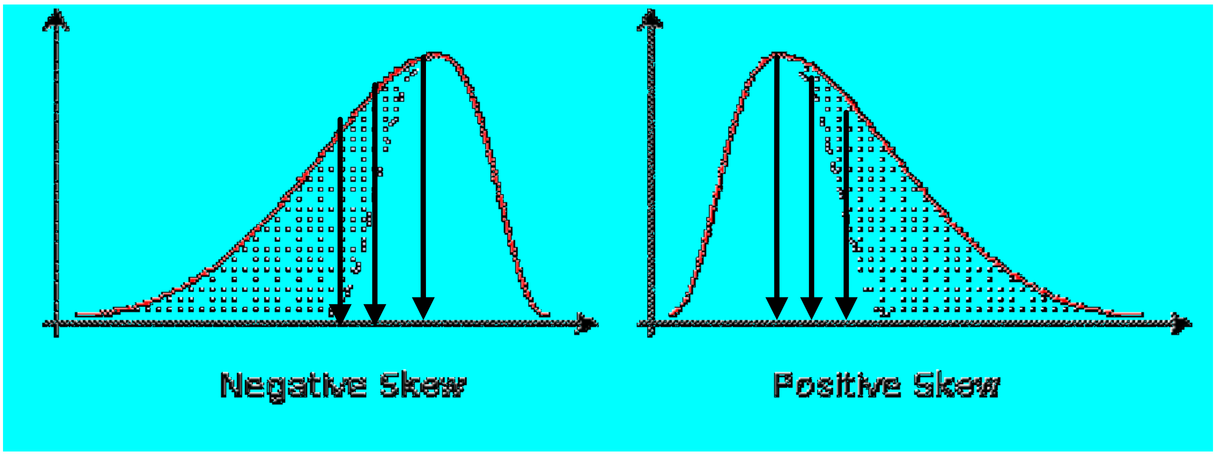


Skewness is a measure of asymmetry. A **negative value indicates skewness to the left**, and a **positive value indicates skewness to the right**. A **zero value does not necessarily indicate symmetry**.

$$k_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} [(x_i - \bar{x})/s]^3$$

n is the number of nonmissing observations

s is the standard deviation

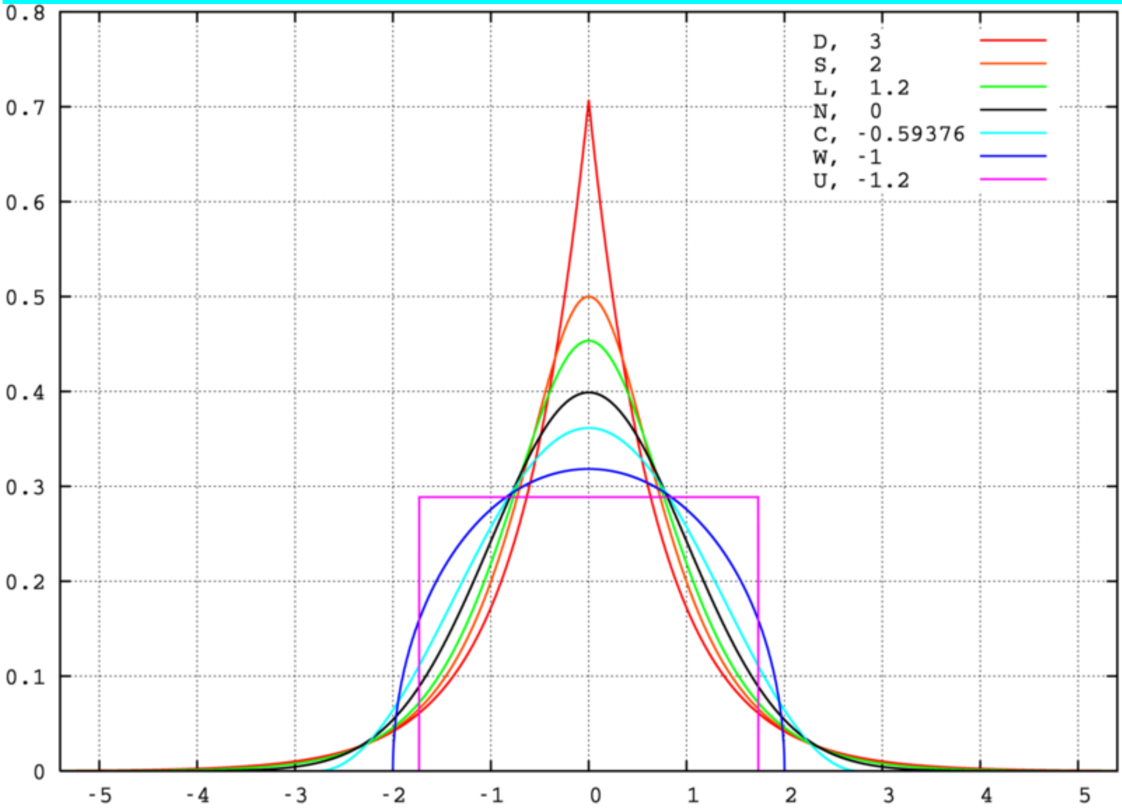


Negative Skew:
Mean< Median < Mode

Positive Skew:
Mode<Median<Mean

Kurtosis

Kurtosis is one measure of how different a distribution is from the ***normal distribution***. ***A positive value typically indicates that the distribution has a sharper peak than the normal distribution. A negative value indicates that the distribution has a flatter peak than the normal distribution.***



$$k_2 = \frac{1}{(n-1)s^4} \sum_{i=1}^{n} [(x_i - \bar{x})]^4 - 3$$

n is the number of nonmissing observations

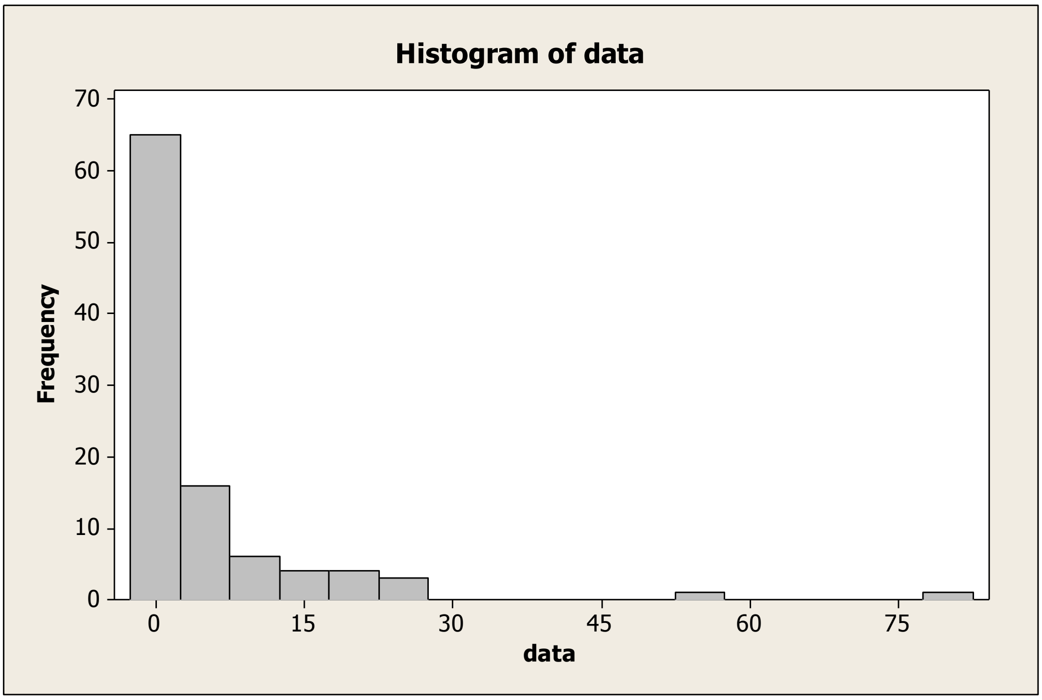
s is the standard deviation

Example:

```
MTB > random 100 c1;
SUBC> weibull 0.5 2.5 .
```

Data Display

```
data
   10.5021     2.2097     5.4653     0.0040    21.8312    17.4945     1.2898
    2.9485     3.5788     3.3279     0.1589     2.4786     1.6979     0.4671
    4.1539     0.0543     0.0199     0.3248     0.1948    10.7923     1.6098
    0.0032     1.3626     0.0014     1.3320     1.5999     0.3397     0.5601
    0.8239     1.8025     0.0000     0.3539     1.7642    15.8646    11.0424
    0.8356     1.3491     3.5441     4.9595     0.2661     0.0697    21.6256
    7.4584     0.0020     0.1361     1.7300     2.4035     2.3521     0.0010
    0.1300    12.8130    25.3351     0.8687    53.1269     0.1662     3.3937
    0.0006     1.7983    22.4206     0.2673     0.0000     0.0006     2.8813
    6.9709     0.0420    10.3612     1.1606    27.1833     3.4946     0.9873
   20.0124     0.4510     0.3137     3.1624     5.4806     0.7834    12.2033
    2.0070     1.7640     3.2358     1.3209     0.1966     0.4443     0.3849
    0.7711     0.3343     1.1401     1.4935    13.5693     0.1381     0.0339
    0.8343     8.1633    80.2928     5.7079     0.3438     2.0176     0.0685
   27.4756     0.1405
```



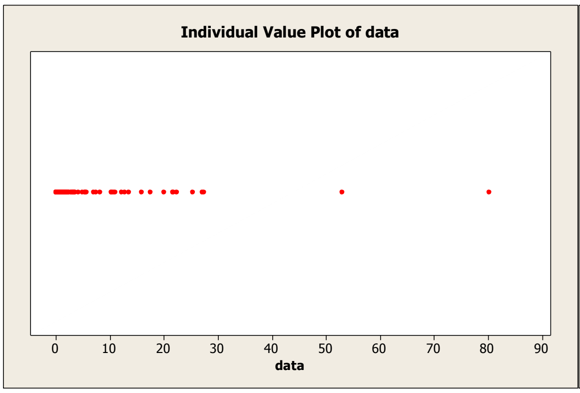
```
MTB > Describe 'data';
SUBC>   Mean;
SUBC>   SEMean;
SUBC>   StDeviation;
SUBC>   Variance;
SUBC>   CVariation;
SUBC>   QOne;
SUBC>   Median;
SUBC>   QThree;
SUBC>   IQRange;
SUBC>   Mode;
SUBC>   TRMean;
SUBC>   Sums;
SUBC>   Minimum;
SUBC>   Maximum;
SUBC>   Range;
SUBC>   SSQ;
SUBC>   Skewness;
SUBC>   Kurtosis;
SUBC>   MSSD;
SUBC>   N;
SUBC>   NMissing;
SUBC>   Count;
SUBC>   CumN;
SUBC>   Percent;
SUBC>   CumPercent;
SUBC>   GHist;
SUBC>   GNHist;
SUBC>   GIndPlot;
SUBC>   GBoxplot.
```

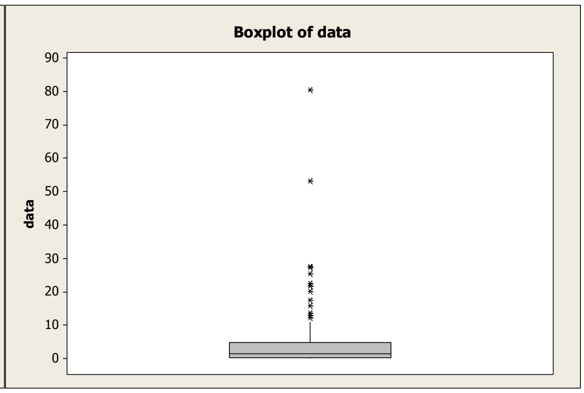
Descriptive Statistics: data

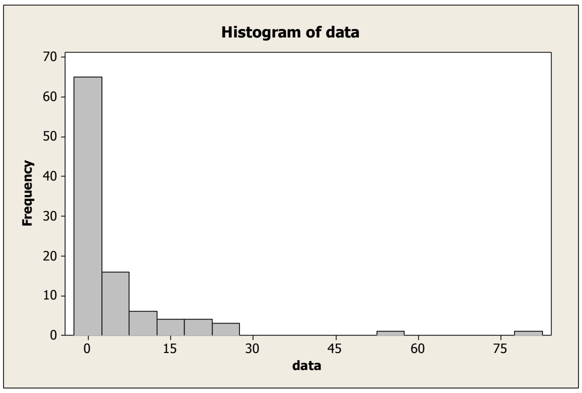
Variable	Total Count	N	N*	CumN	Percent	CumPct	Mean	SE Mean	TrMean	StDev
data	100	100	0	100	100	100	5.42	1.11	3.65	11.15

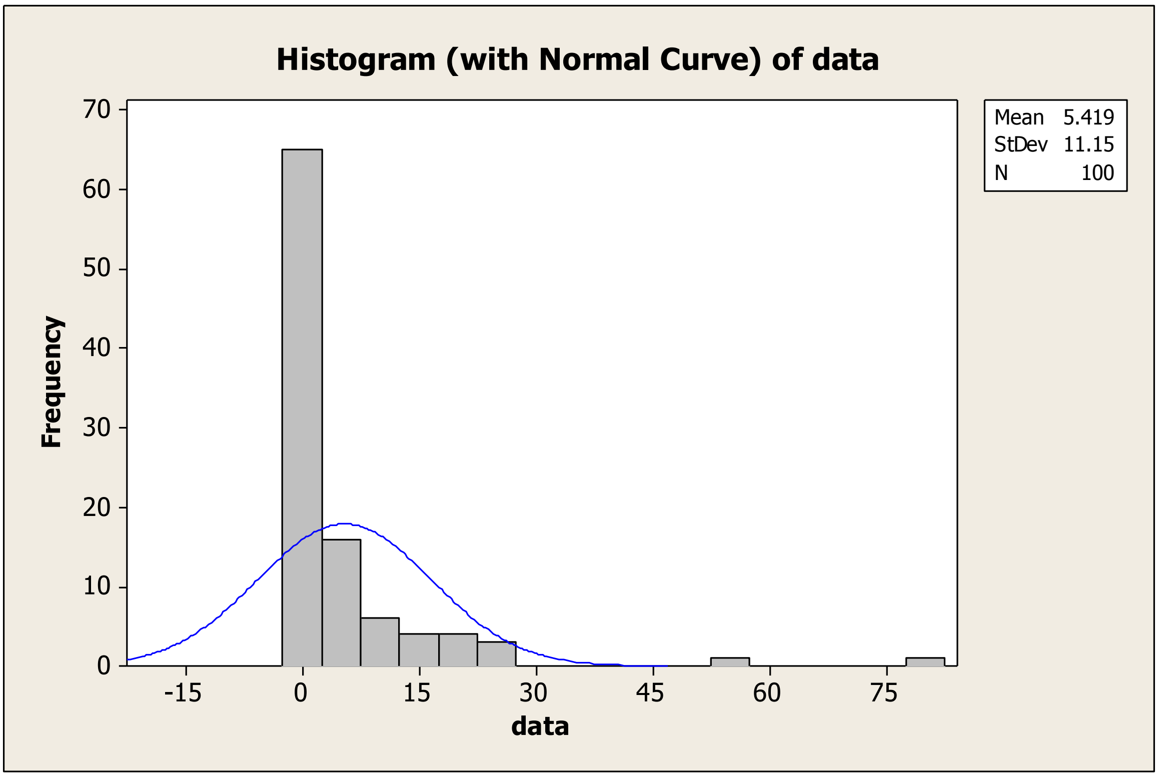
Variable	Variance	CoefVar	Sum	Sum of Squares	Minimum	Q1	Median	Q3
data	124.29	205.74	541.87	15241.44	0.00	0.28	1.43	4.76

Variable	Maximum	Range	IQR	Mode	N for Mode	Skewness	Kurtosis	MSSD
data	80.29	80.29	4.48	*	0	4.25	23.01	125.36


Individual Value Plot of data
0
10
20
30
40
50
60
70
80
90
data


Boxplot of data
90
80
70
60
50
40
30
20
10
0
data


Histogram of data
70
60
50
40
30
20
10
0
Frequency
0
15
30
45
60
75
data


Histogram (with Normal Curve) of data
Mean 5.419
StDev 11.15
N 100
70
60
50
40
30
20
10
0
Frequency
-15
0
15
30
45
60
75
data

Similar ratios

- **Relative standard deviation, $|\sigma / \mu|$**
- **Standardized moment, μ_k / σ^k**
- **Variance-to-mean ratio, σ^2 / μ**
- **Signal to noise ratio, μ / σ (in signal processing)**