

**STP 226
ELEMENTARY STATISTICS**

NOTES

PART 2 - DESCRIPTIVE STATISTICS

CHAPTER 3

DESCRIPTIVE MEASURES

Chapter 2 covered organizing data into tables, and summarizing data with graphical displays.

We will now use numbers called **descriptive measures** to describe data sets.

3.1 Measures of Center

Measures of central tendency / measures of center/averages.

- where center or most typical value of a data set lies

Mean – sum of the observations divided by the number of observations.

Example: 4, 5, 7, 8, 10

Mean = $(4 + 5 + 7 + 8 + 10)/5 = 34/5 = 6.8$

Median – divides the bottom 50% of the data from the top 50%

Arrange the data in increasing order

1. If the # of observations is odd, the median is the observation exactly in the middle.
2. If the # of observations is even, the median is the mean of the two middle observations.

For n observations, the position of the median is the $(n + 1)/2$ th position in the ordered distribution.

Example: 4, 5, 7, 8, 10 (in order) **Median** = 7

Example: 4, 5, 7, 8, 10, 12 **Median** = $(7 + 8)/2 = 15/2 = 7.5$

Mode – the value that occurs most frequently in the data set

Obtain the frequency of each value

1. If the greatest frequency is 1, then there is no mode.
2. If the greatest frequency is 2 or greater, then any value with that greatest frequency is the mode of the data set.

Example: 2, 3, 3, 3, 4, 4, 5 **Mode** = 3

Example: 2, 3, 3, 3, 4, 4, 4, 5 **Mode** = 3 and 4

Example: 1, 3, 5, 4, 7, 9, 0 **Mode** = none

- Mode is the only appropriate measure of center for qualitative data.

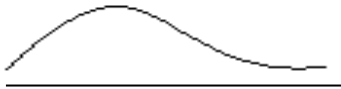
Example: A sample of 100 ASU students had 25% Democrats, 60% Republicans and 15% Libertarians.

Mode = Republicans

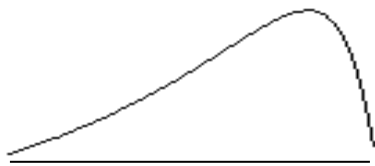
Mean is affected by extreme observations - **nonresistant**

Median is not sensitive to extreme observations - **resistant**

In a **right skewed** distribution the mean is pulled more towards the right, no effect on the median (Median < Mean)



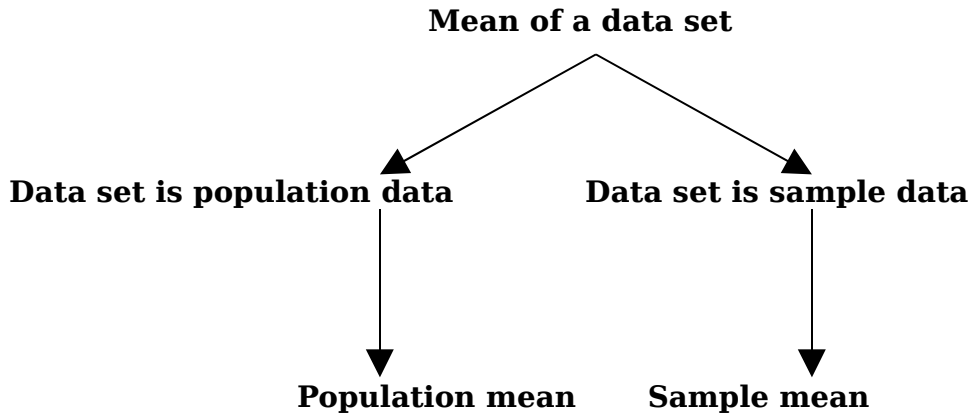
In a **left skewed** distribution, the mean is pulled more towards the left, no effect on the median (Median > Mean)



In a **symmetric** distribution, mean and median are at the same position (same value)



Trimmed means – cut off a small percentage of the observations on both ends of the distribution and then compute the mean. this makes the mean a more resistant measure since the extreme observations may be deleted from the data set



This is the same for the **median, the mode, or any other descriptive measure**.

If the data set is **population data** then the measures are called **population measures**.

If the data set is **sample data** then the measures are called **sample measures**.

- **The Sample Mean (\bar{x})**

Observations labeled with **subscripts**, x_1, x_2 , etc

Summation notation $\Sigma x = x_1 + x_2 + \dots$

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

Sample mean – mean of observations of a sample.

$$\bar{x} = \Sigma x_i / n$$

Example:

106	122	120	123
118	114	138	131
128	124	119	130

Compute \bar{x} , Σx^2 , $\Sigma (x - \bar{x})$, $\Sigma (x - \bar{x})^2$, median, mode

x	x - \bar{x}	(x - \bar{x})²
106	-16.75	280.5625
114	-8.75	76.5625
118	-4.75	22.5625
119	-3.75	14.0625
120	-2.75	7.5625
122	-0.75	0.5625
123	0.25	0.0625
124	1.25	1.5625
128	5.25	27.5625
130	7.25	52.5625
131	8.25	68.0625
138	15.25	232.5625
1473	0	784.25
Σx	$\Sigma (x - \bar{x})$	$\Sigma (x - \bar{x})^2$
12	n	
122.7		
5	\bar{x}	
122.5	median	
none	mode	

Computing Sample Mean for grouped data:

$$\bar{x} = \Sigma (x*f)/n \quad \text{where } n = \Sigma f$$

<i>Number of school children=x</i>	<i>Frequency=f</i>	<i>x*f</i>
0	2	0
1	5	5
2	4	8
3	3	9
4	4	16
5	2	10
	20	48

$$\bar{x} = 48/20 = 2.4$$

If the data is grouped using classes that are not single value, we can estimate the mean by using **Midpoints** of each class:

Days to maturity	Frequency =f	Midpoint =x	x*f
30<40	3	35	105
40<50	1	45	45
50<60	8	55	440
60<70	10	65	650
70<80	7	75	525
80<90	7	85	595
90<100	4	95	380
	40		2740

$\bar{x} = 2740/40 = 68.5$ (estimate of the sample mean)

Using raw data (below) true sample mean is $\bar{x} = 2731/40 = 68.275$, so we did OK

70 64 99 55 89 87 65 62 38 67 70 60 69 78 39 75 56 71 51
99
68 95 86 57 53 47 50 55 81 80 98 51 36 63 66 85 79 83 70
64

3.2 Measures of Variation; The Sample Standard Deviation

Measures of spread (measures of variation)

Range = Max. value - Min. value

- the difference between the largest and smallest observations.

Two data sets may have the same mean but different ranges.

The Sample Standard Deviation - measures variation

- tells how far, on average, the observations are from the mean

- like the mean, it is not a resistant measure

Deviations from the mean: $(x_i - \bar{x})$ - how far each observation is from the mean

Sum of squared deviations $\sum (x_i - \bar{x})^2$

For a variable x, the standard deviation of the observations for a sample is called a **sample standard deviation**. It is denoted by s_x or, when no confusion will arise simply by s. We have

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \quad \text{where } n \text{ is the sample size.}$$

Steps to calculate standard deviation

1. Calculate the sample mean, \bar{x}
2. Construct a table to obtain the sum of squared deviations, $\sum (x_i - \bar{x})^2$
3. Substitute into the formula for s above.

The more variation there is in a data set, the larger its standard deviation.

Computational formula for s is

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

Example: (a) Compute standard deviation using Definition:

10 15 11 15 10 14 13 16 12 14 10 16

x	x - \bar{x}	(x - \bar{x})²
10	-3	9
15	2	4
11	-2	4
15	2	4
10	-3	9
14	1	1
13	0	0
16	3	9
12	-1	1
14	1	1
10	-3	9
16	3	9
$\Sigma x =$	$\Sigma (x - \bar{x}) =$	$\Sigma (x - \bar{x})^2 = 60$
156	0	

n=12, $\bar{x}=13$ $S^2 = 60/11=5.45$

S=2.34

(b) Using Computational formula for s:

x	x²
10	100
15	225
11	121
15	225
10	100
14	196
13	169
16	256
12	144
14	196
10	100
16	256
Σ x=156	Σ x²=2088

$$n=12, S^2 = [2088-(156)^2/12]/(11)=60/11=5.45 \quad \text{and} \quad S=2.34$$

If data is grouped, we compute S as follows:

$$s = \sqrt{\frac{\sum (x^2 * f) - (\sum (x * f))^2 / n}{n - 1}}$$

where $n = \sum f$

Number of school children=x	Frequency=f	x*f	x ²	x ² *f
0	2	0	0	0
1	5	5	1	5
2	4	8	4	16
3	3	9	9	27
4	4	16	16	64
5	2	10	25	50
	n=20	Σ x*f =48		Σx²*f = 162

$$n=20, S^2 = [162-(48)^2/20]/19=2.46, S=1.57$$

Three Standard Deviations Rule:

Almost all of the observations in any data set lie within three (3) standard deviations to either side of the mean.

More Precise Rules for any data set:

Chebychev's rule : ~ 89% of the observations in any data set lie within three standard deviations to either side of the mean.

Chebychev's rule (more precisely): For any data set and any number $k > 1$, at least $100(1 - 1/k^2)\%$ of the observations lie within k standard deviations to either side of the mean.

If the distribution is ~ bell-shaped, the **Empirical Rule** implies that ~ 99.7% of the observations lie within three standard deviations to either side of the mean.

3.3 The five-number summary; Boxplots

Median, Percentiles, Deciles, Quartiles, Interquartile Range are all resistant measures.

- **Percentiles, Deciles, and Quartiles**

Percentiles – divide the distribution into 100 equal parts (P_1, P_2, \dots, P_{99})

P_1 divides the bottom 1% of the data from the top 99%

P_2 divides the bottom 2% of the data from the top 98%

Etc,

Thus, the median is the 50th percentile

Deciles – divide the distribution into 10 equal parts (D_1, D_2, \dots, D_9)

D_1 divides the bottom 10% of the data from the top 90%

D_2 divides the bottom 20% of the data from the top 80%

Etc,

Thus, the median is D_5

Quartiles – divide the distribution into 4 equal parts (Q_1, Q_2, Q_3)

Q_1 divides the bottom 25% of the data from the top 75%

Q_2 divides the bottom 50% of the data from the top 50%

Q_3 divides the bottom 75% of the data from the top 25%

Thus the median is Q_2

To find the Quartiles

Arrange the data in increasing order.

1. Q_1 is the median of the data set that lies at or below the median of the entire data set.
2. Q_2 is the median of the entire data set.
3. Q_3 is the median of the data set that lies at or above the median of the entire data set.

Examples:

- n=7 Data: 3, 4, 5, 6, 12, 13, 14
 $Q_1 = (4+5)/2 = 4.5$
 $Q_2 = 6$
 $Q_3 = (12+13)/2 = 12.5$

- n=10 Data: 1, 3, 4, 5, 6, 12, 13, 14, 15, 18
 $Q_1 = 4$
 $Q_2 = (6+12)/2 = 9$
 $Q_3 = 14$

Interquartile Range (IQR) – difference between the first and third quartiles.

$$\mathbf{IQR = Q_3 - Q_1}$$

IQR gives the range of the middle 50% of the observations (approximately)

The **five-number summary** of a data set consists of the minimum, maximum, and the quartiles in increasing order. Min., Q_1 , Q_2 , Q_3 , Max.

Outliers – observations well outside of the overall pattern of the data

Lower limit = $Q_1 - 1.5 * IQR$

Upper limit = $Q_3 + 1.5 * IQR$

Potential outliers are observations outside of the Lower and Upper Limits.

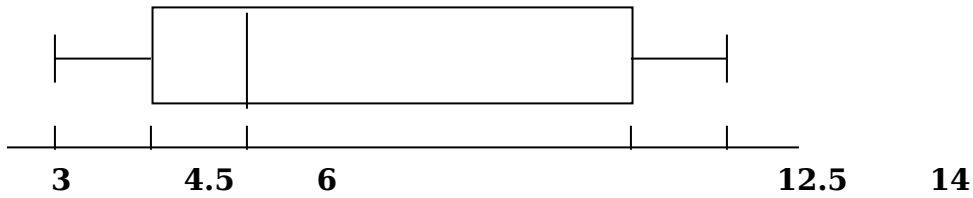
Adjacent values - most extreme obs. that are still lying within the upper and lower limits

- most extreme obs. that are not outliers.

Boxplot (box-and-whisker diagram) and the modified boxplot**To construct a boxplot**

- Determine the 5 number summary (Min, Q_1 , Q_2 , Q_3 , Max.)
- Draw a horizontal axis on which the numbers obtained in step 1 can be located. Above this axis, mark the quartiles and the minimum and maximum with vertical lines.
- Connect the quartiles to each other to make a box, and then connect the box to the minimum and maximum with lines.

The following is Boxplot for example 1 (top of previous page):



To construct a modified boxplot

1. Determine the quartiles.
2. Determine potential outliers and the adjacent values.
3. Draw a horizontal axis on which the numbers obtained in steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.
4. Connect the quartiles to each other to make a box, and then connect the box to the adjacent values with lines.
5. Plot each potential outlier with an asterisk.

The two lines stretching out on both sides are the whiskers.

3.4 Descriptive Measures for Populations; Use of Samples

Notation:

	Size	Mean
Sample	n	\bar{X}
Population	N	μ

Population Mean (Mean of a Variable) – computed in same manner as for a sample

For a variable x , the mean of all possible obs. for the entire population is called the **population mean** or **mean of the variable** x . It is denoted by μ_x or when no confusion will arise, simply by μ . For a finite population, we have

$$\mu = \frac{\sum x}{N} \quad \text{where } N \text{ is the population size.}$$

Using a Sample Mean to Estimate a Population Mean

Take a sample from the population and compute its mean. This mean is used to estimate the mean of the population (chapter 8)

Population Standard Deviation (Standard Deviation of the Variable)

For a variable x , the standard deviation of all possible obs. for the entire population is called the **population standard deviation** or **standard deviation of the variable** x . It is denoted by σ_x or, when no confusion will arise, simply by σ . For a finite population, we have

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

where N is the population size. The population standard deviation can also be obtained from the computing formula

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

Population Variance = σ^2

Using a Sample Standard Deviation to Estimate a Population Standard Deviation:

Take a sample. Compute the sample standard deviation, s and then estimate the population standard deviation, σ .

Parameter – A descriptive measure for a population.

Examples: μ , σ

Statistic – A descriptive measure for a sample.

Examples: \bar{x} , s

Standardized Variable – For a variable x , the variable

$z = \frac{x - \mu}{\sigma}$ is called the **standardized version** of x or the **standardized variable** corresponding to the variable x , z is also called the **standard score** or **z-score**

$$\mu_z = \frac{\sum z_i}{N} = 0 \qquad \sigma_z = \sqrt{\frac{\sum (z_i - \mu_z)^2}{N}} = 1$$

- The same rules for the proportions of observations being within 3 standard deviations of the mean apply.

- Standard score z is negative if x is below the mean, positive if x is above the mean and zero if x is equal to the mean.
- The value of the z score tells us how many standard deviations above or below the mean is a particular value of x .