3.2 MEASURES OF CENTER

DEFINITION

A **measure of center** is a value at the center or middle of a data set.

DEFINITION

The **arithmetic mean (aka mean)** of a set of data is the measure of center found by adding the data values and dividing the total by the number of data values.

$$\text{mean} = \frac{\sum x}{n} = \frac{\text{sum of all data values}}{\text{number of data values}}$$

**One advantage of the mean is that it is relatively reliable, so that when samples are selected from the same population, sample means tend to be more consistent than other measures of center. Another advantage of the mean is that it takes every data value into account. However, because the mean is sensitive to every value, just one extreme value can affect it dramatically. Because of this fact, we say the mean is not a resistant measure of center.

NOTATION

Σ : sum

x : variable used to represent individual data values

n : number of data values in a sample

$\bar{x} = \frac{\Sigma x}{n}$: mean of a set of sample values

$\tilde{x}$: median, N : number of data values in a population

(mu) $\mu = \frac{\Sigma x}{N}$: mean of a set of population values

Example 1: Find the mean of the following numbers:

17 23 17 22 21 34 27

$$\bar{x} = \frac{\Sigma x}{n}$$

$$\bar{x} = \frac{17+23+17+22+21+34+27}{7}$$

$$\bar{x} = \frac{161}{7} = \boxed{23}$$

DEFINITION

The **median** of a data set is the measure of center that is the ___middle___ value when the original data values are arranged in ___order___ of increasing (or decreasing) magnitude. The median is often denoted ___$\tilde{x}$___ (pronounced "x-tilde"). To find the median, first ___sort___ the values, then follow one of these two procedures:

1. If the number of data values is ___odd___, the median is the number located in the exact ___middle___ of the list.
2. If the number of data values is ___even___, the median is the ___average___ of the ___middle___ two numbers.

**The median is a ___resistant___ measure of center, because it does not change by ___large___ amounts due to the presence of just a few ___extreme___ values.

Example 2:

a. Find the median of the following numbers:

17 23 17 22 21 34 27

17 17 21 22 23 27 34

$$\boxed{\tilde{x} = 22}$$

b. Find the median of the following numbers

17 23 17 22 34 27

17 17 22 23 27 34

$\tilde{x} = \frac{22+23}{2} = \boxed{22.5}$

DEFINITION

The **mode** of a data set is the value that occurs with the greatest __frequency__.

A data set can have more than one mode, or no mode.

- When two data values occur with the same greatest frequency, each one is a __mode__ and the data set is __bimodal__.
- When more than two data values occur with the same greatest frequency, each is a __mode__ and the data set is said to be __multimodal__.
- When no data value is repeated, we say there is no __mode__.

**The mode is the only measure of center that can be used with data at the __nominal__ level of measurement.

Example 3:

a. Find the mode of the following numbers:

17 23 17 22 21 34 27

mode is 17

b. Find the mode of the following numbers

17 23 17 22 21 34 27 22

data set is bimodal → the modes are 17 and 22

DEFINITION

The **midrange** of a data set is the measure of center that is the value ___midway___ between the ___minimum___ and ___maximum___ values in the original data set. It is found by adding the maximum data value to the minimum data value and then dividing the sum by two.

$$\text{midrange} = \frac{\text{min. data value} + \text{max. data value}}{2}$$

**The midrange is rarely used because it is too ___sensitive___ to extremes since it uses only the minimum and maximum data values.

Example 4: Find the midrange of the following numbers:

17 23 17 22 21 34 27

minimum: 17, maximum: 34

$$\text{midrange} = \frac{17+34}{2} = \boxed{25.5}$$

ROUND-OFF RULE FOR THE MEAN, MEDIAN, AND MIDRANGE

Carry ___one___ more decimal place than is present in the original data set. Because values of the mode are the same as some of the original data values, they can be left without any rounding.

MEAN FROM A FREQUENCY DISTRIBUTION

When working with data summarized in a frequency distribution, we don't know the ___exact___ values falling in a particular ___class___. To make calculations possible, we assume that all sample values in each class are equal to the class ___midpoint___. We can then add the ___products___ from each ___class___ to find the total of all sample values, which we can then ___divide___ by the sum of the frequencies, $\sum f$

frequency class midpoint

$$\bar{x} = \frac{\sum (f \cdot x)}{\sum f}$$

Example 5: Find the mean of the data summarized in the given frequency distribution.

Tar (mg) in nonfiltered cigarettes	Frequency
10-13	1
14-17	0
18-21	15
22-25	7
26-29	2

Tar (mg) in nonfiltered cigs	Frequency (f)	Midpoint (x)	f · x
10-13	1	11.5	11.5
14-17	0	15.5	0
18-21	15	19.5	292.5
22-25	7	23.5	164.5
26-29	2	27.5	55

$$\bar{x} = \frac{\Sigma f \cdot x}{\Sigma f}$$

$$\bar{x} = \frac{523.5}{25}$$

$$\boxed{\bar{x} \doteq 20.9\,mg}$$

WEIGHTED MEAN

When data values are assigned different weights, we can compute a weighted mean.

$$\bar{x} = \frac{\sum(w \cdot x)}{\sum w}$$

Example 6: A student earned grades of 92, 83, 77, 84, and 82 on her regular tests. She earned grades of 88 on the final and 95 on her class project. Her combined homework grade was 77. The five regular tests count for 60% of the final grade, the final exam counts for 10%, the project counts for 15%, and homework counts for 15%. What is her weighted mean grade? What letter grade did she earn?

Average exam score is: $\frac{92+83+77+84+82}{5} = 83.6$

x	w	$x \cdot w$
83.6	0.60	50.16
88	0.10	8.8
95	0.15	14.25
77	0.15	11.55
Sum	1.00	84.76

$$\bar{x} = \frac{\Sigma x \cdot w}{\Sigma w}$$

$$\bar{x} = \frac{84.76}{1.00}$$

$$\bar{x} \doteq 84.8$$

Her weighted mean grade is approximately 84.8%. She earned a B in the class.

SKEWNESS

A comparison of the mean, median, and mode can reveal information about the characteristic of **skewness**. A distribution of data is said to be skewed if it is not symmetric and extends more to one side than the other.

A Comparison of the Mean, Median, and Mode

The mean, median, and mode are affected by what is called skewness (i.e., lack of symmetry) in the data.

- Here is Figure 15.6, which showed a normal curve, a negatively skewed curve, and a positively skewed curve:

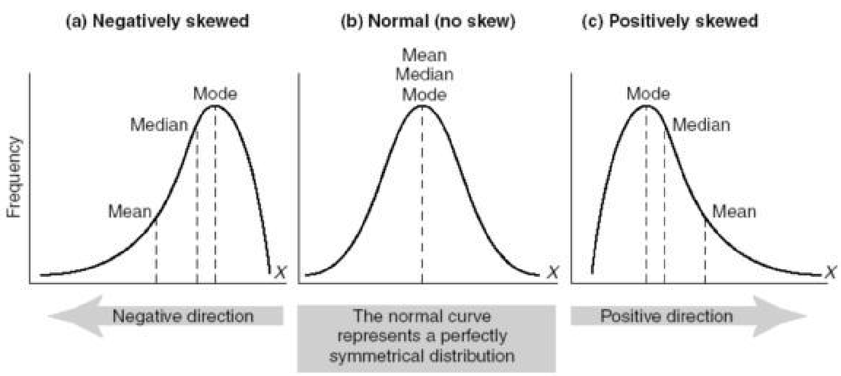


FIGURE 15.6 Examples of normal and skewed distributions

- Look at the above figure and note that when a variable is normally distributed, the mean, median, and mode are the same number.
- When the variable is skewed to the left (i.e., negatively skewed), the mean shifts to the left the most, the median shifts to the left the second most, and the mode the least affected by the presence of skew in the data.
- Therefore, when the data are negatively skewed, this happens:
  mean < median < mode.
- When the variable is skewed to the right (i.e., positively skewed), the mean is shifted to the right the most, the median is shifted to the right the second most, and the mode the least affected.
- Therefore, when the data are positively skewed, this happens:
  mean > median > mode.
- If you go to the end of the curve, to where it is pulled out the most, you will see that the order goes mean, median, and mode as you "walk up the curve" for negatively and positively skewed curves.

3.3 MEASURES OF VARIATION

DEFINITION

The **range** of a set of data values is the difference between the maximum and the minimum data value.

DEFINITION

The **standard deviation** of a set of sample values, denoted by s, is a measure of __variation__ of values about the __mean__. It is a type of __average__ deviation of values from the mean that is calculated by using either of the following formulas:

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

or

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

- The standard deviation is a measure of __variation__ of all values from the __mean__.
- The value of the standard deviation is usually __positive__.
  - It is zero only when all of the data values are the same __number__.
  - It is never __negative__.
- Larger values of the standard deviation indicate __greater__ amounts of __variation__.
- The value of the standard deviation can increase dramatically with the inclusion of one or more __outliers__.
- The units of the standard deviation are the same units as the original __data__ values.

General Procedure for Finding Standard Deviation (1st formula)	Specific Example Using the Following Numbers: 2, 4, 5, 16
Step 1: Compute the mean $\bar{x}$	$\bar{x} \doteq 6.75$
Step 2: Subtract the mean from each individual sample value	$2-6.75=-4.75 \rightarrow x_1-\bar{x}$ $4-6.75=-2.75 \rightarrow x_2-\bar{x}$ $5-6.75=-1.75 \rightarrow x_3-\bar{x}$ $16-6.75=9.25 \rightarrow x_4-\bar{x}$
Step 3: Square each of the deviations obtained from Step 2.	$(-4.75)^2=22.56 \rightarrow (x_1-\bar{x})^2$ $(-2.75)^2=7.56 \rightarrow (x_2-\bar{x})^2$ $(-1.75)^2=3.06 \rightarrow (x_3-\bar{x})^2$ $(9.25)^2=85.56 \quad (x_4-\bar{x})^2$
Step 4: Add all of the squares obtained from Step 3.	$\Sigma(x-\bar{x})^2=118.74$
Step 5: Divide the total from Step 4 by the number $n-1$, which is one less than the total number of sample values present.	$\frac{\Sigma(x-\bar{x})}{n-1}=\frac{118.74}{4-1}=39.58$
Step 6: Find the square root of the result from Step 5. The result is the standard deviation.	$s=\sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$

$\boxed{s \approx 6.3}$

STANDARD DEVIATION OF A POPULATION

The definition of standard deviation and the previous formulas apply to the standard deviation of ___population___ data. A slightly different formula is used to calculate the standard deviation σ of a ___population___: instead of dividing by $n-1$, we divide by the population size N.

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

DEFINITION

The **<u>variance (aka dispersion aka spread)</u>** of a set of values is a measure of ___variation___ equal to the ___square___ of the ___standard deviation___.

Sample variance: s^2

Population variance: σ^2

**The sample variance is an unbiased estimator of the ___population___ variance, which means that values of s^2 tend to target the value σ^2 of instead of systematically tending to ___overestimate___ or underestimate σ^2.

USING AND UNDERSTANDING STANDARD DEVIATION

One simple tool for understanding standard deviation is the range rule of thumb, which is based on the principle that for many data sets, the vast majority (such as 95%) lie within 2 standard deviations of the mean.

RANGE RULE OF THUMB

Interpreting a known value of the standard deviation: We informally defined usual values in a data set to be those that are typical and not too extreme. If the standard deviation of a collection of data is known, use it to find rough estimates of the minimum and maximum usual values as follows:

minimum "usual " value = (mean) – 2 x (standard deviation)

maximum "usual " value = (mean) + 2 x (standard deviation)

Estimating a value of the standard deviation *s*: To roughly estimate the standard deviation from a collection of known sample data, use

$$s \approx \frac{\text{range}}{4}$$

Example 1: Use the range rule of thumb to estimate the ages of all instructors at MiraCosta if the ages of instructors are between 24 and 60.

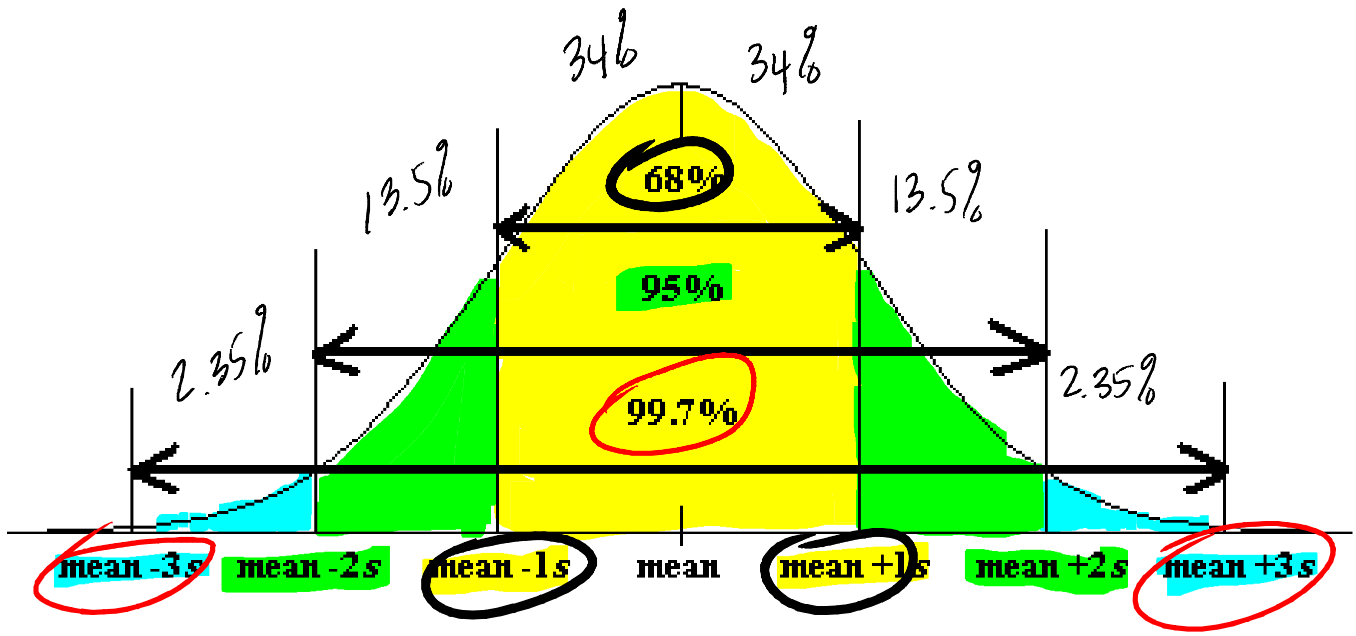
$$s \approx \frac{\text{range}}{4}$$

$$s \approx \frac{60-24}{4} = \boxed{9 \text{ yrs}}$$

EMPIRICAL (OR 68-95-99.7) RULE FOR DATA WITH A BELL-SHAPED DISTRIBUTION

Another concept that is helpful in interpreting the value of a standard deviation is the __empirical__ rule. This rule states that for data sets having a __distribution__ that is approximately __normal__, the following properties apply:

- About 68% of all values fall within 1 standard deviation of the mean
- About 95% of all values fall within 2 standard deviations of the mean
- About 99.7% of all values fall within 3 standard deviations of the mean



Example 2: The author's Generac generator produces voltage amounts with a mean of 125.0 volts and a standard deviation of 0.3 volt, and the voltages have a bell-shaped distribution. Use the empirical rule to find the approximate percentage of voltage amounts between

a. 124.4 volts and 125.6 volts

$\bar{x} = 125$

124.4 is $2 \cdot 0.3$ volts below the mean ← 2 s.d. below

and 125.6 is $2 \cdot 0.3$ volts above the mean ← 2 s.d. above

$\boxed{95\%}$

b. 124.1 volts and 125.9 volts

→ 3 s.d. below → 3 s.d. above

$\boxed{99.7\%}$