Capture – Recapture Method of Sampling

Capture – Recapture is a common method used to estimate the size of a population by sampling. Biologists and ecologists use this method extensively to estimate wild animal populations.

- **Step 1**: **Capture** a sample of the animal you want to count in the area you want to know about. (your book calls the number captured n_1)
  - **Tag** all the captured animals (given them each an identifying mark)
  - **Release** them back into the wild
- **Step 2**: **Recapture** after enough time for the released individuals to re-mix with the whole population, capture a new sample of individuals and count the number of tagged and the number of untagged individuals in this second sample.

The logic of capture-recapture computations: The computation is the proportion formed by two correctly stacked ratios.

<u>IF</u> we can assume that the recaptured sample is representative of the whole population, then

$$\frac{\#\ tagged\ in\ total\ \ population}{Total\ \ population} = \frac{\#tagged\ in\ the\ recapture\ sample}{Total\ \#in\ the\ recapture\ sample}$$

I suggest you memorize this rather than the formula on page 530 of your text

Notice that we KNOW the # tagged in the total population – that is the number we tagged in the first sample.

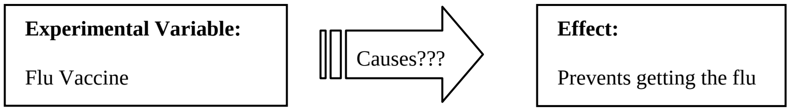
Example 13.4 (reworded – This is the way such questions will appear on the exam)
A large pond is stocked with catfish. You capture 200 catfish, tag and release them. You wait enough time for the tagged fish to spread out more with the general population. Then you capture another sample. This sample has 250 catfish. Of the 250 catfish in this second sample, 35 have tags.

If the second sample is representative of the catfish population in the pond, estimate the number of catfish in the pond.

Section 13.5 Clinical Studies (Clinical Trials)

Clinical Studies do not collect data for the same purposes as surveys and censuses.

Instead, Clinical Studies attempt to determine whether a single variable can cause a certain effect.



Experimental Variable:

Flu Vaccine

Causes???

Effect:

Prevents getting the flu

New vaccines and drug treatments are put through clinical studies before being officially approved for public use.

Things that are “unhealthy” like cigarettes and caffeine are officially identified as “unhealthy” after clinical studies show that people who include significant amounts of them in their lifestyle have more health problems than people who do not include them.

Controlled Clinical Study Methodology

A **controlled clinical study** uses two groups:

- **treatment group** (receives the actual treatment)
- **control group** (sometimes called the comparison group) should only differ from the treatment group in that they do not receive the treatment.

Confounding Variable: a characteristic (not the one being studied) in which the control and treatment groups differ. Then you can’t tell whether the effect was due to the characteristic being studied or due to this other characteristic or a combination of both.

Randomized Controlled Study: subjects are randomly assigned to either the treatment or control group

We can only deduce that the treatment CAUSES the effect if the treatment group experiences the effect and the control group does not experience the effect.

Placebo Effect: just the idea that one is getting treatment can produce positive results. People receiving a **placebo** (a harmless, inactive substance like a “sugar pill”) often report experiencing improvement.

Blind Study: The placebo effect cannot be eliminated, but it can be controlled by giving a placebo to the control group and conducting a **blind study,** in which neither the treatment nor the control group know whether they are getting the real treatment or the placebo.

Double Blind Study: The scientists conducting the study are also not aware of whether the participant is getting the real treatment or the placebo. (participants and researchers are both “blind”).

Even clinical studies that are properly designed can lead to conflicting conclusions. But when clinical studies of the same variable, done in different labs by different groups, **consistently find the same conclusion**, the clinical study method is persuasive.

It is important to use control groups because:

Association is NOT Causation: Just because 2 conditions occur together does not mean one condition causes the other. They may both be caused by some 3rd condition, or they may just coincide by chance. **ALSO** a single effect can have many possible and actual causes.

Example: The school district that receives the most federal money and pays the highest teachers' salaries has the lowest national test scores. Does higher teacher pay cause low test scores?

Example: A black cat crosses your path in the morning and by afternoon you have lost your job. Did the black cat crossing your path cause you to lose your job?

Class Practice – Clinical Studies

Study #1: In order to determine the effectiveness of a new drug for HIV treatment, the researchers conducted a study at the Park HIV Clinic in Philadelphia. The clinic first asked all 8,000 of their HIV patients who were between the ages of 20 and 40 years of age if they would be willing to participate.

Only 2000 volunteered to participate in the study. All 2000 of those volunteers were given a battery of medical assessments to determine the severity of symptoms they were experiencing and prognosis.

The researchers looked at the results of these medical assessments and found there were 150 of these volunteers who were in the beginning stages of HIV infection and were showing only minimal symptoms. These 150 patients became the participants in the study.

By random assignment, 75 were assigned to "Group A" and the other 75 were assigned to "Group B". Group A received injections from "Drug A" vials while Group B received injections from "Drug B" vials.

One vial was the experimental drug and the other vial was a placebo treatment. Neither the patients nor the researchers knew whether "Drug A" or "Drug B" was the actual treatment drug.

Participants received the injections once a week for 6 months. At the end of the 6 months of treatment, the patients were again given the same battery of medical assessment to determine the severity of symptoms they were experiencing and prognosis. The average level of health was found to be significantly better for Group B.

Group B turned out to be the group that had received the real drug treatment.

1. What is the sampling frame in this study?

2. What is the target population of this study?

3. Does it matter to the results of the study that the participants were volunteers? Why or why not?

4. What purpose did the initial medical screening to select the 150 actual participants serve in the methodology of this survey?

5. What makes this study a **controlled** study?

6. What makes this study a **randomized controlled** study?

7. Is this study **blind** or **double blind**? How can you tell?

Study #2: In order to determine the effectiveness of a new vaccine that is alleged to cure "math anxiety", a clinical study was conducted. One thousand college students enrolled in math courses across the U.S. were chosen to participate in the study. The 1,000 students were broken up into two groups. Those enrolled in calculus courses or higher were given the real vaccine. The students in remedial and basic math courses were given a fake vaccine consisting of sugared water. None of the students knew whether they were being given the real or the fake vaccine, but the researcher conducting the experiment knew. At the end of the semester the students were given a test that measured their level of math anxiety. The students in the treatment group showed significantly lower levels of math anxiety than those in the control group. On the basis of this experiment the vaccine was advertised as being highly effective in fighting math anxiety.

1. The **sampling frame** in this study consists of
   A. the treatment and control groups
   B. all U.S. college students
   C. all U.S college students enrolled in math classes
   D. all students that suffer from "math anxiety"
   E. None of the above
2. The **target population** in this study consists of
   A. treatment and control groups
   B. all U.S. college students
   C. all U.S. college students enrolled in math classes
   D. all students that suffer from "math anxiety"
   E. None of the above
3. The **control group** in this experiment consists of
   A. the 1,000 volunteer college students used for the study
   B. the students given the real vaccine
   C. the students given the fake vaccine
   D. This experiment has no control group because it used volunteers.
   E. None of the above
4. This experiment can best be described as a
   A. double blind randomized controlled experiment
   B. double blind controlled placebo experiment
   C. blind randomized controlled experiment
   D. blind controlled placebo experiment
   E. All of the above
5. The results of this experiment should be considered unreliable because
   F. only college students were used
   G. the treatment and control groups were not the same size
   H. the sample was too small
   I. the treatment and control groups represented two very different segments of the population
   J. None of the above
6. Which of the following is most likely confounding variable for this experiment?
   K. the student's background in mathematics
   L. the student's grade level (freshman, sophomore, junior, senior)
   M. the type of college attended (two year, four year, university)
   N. the student's sex (male, female)
   O. None of the above

Chapter 14: Descriptive Statistics

Descriptive Statistics: statistics that summarize or otherwise describe large amounts of numerical data.

- Present data visually as pictures or graphs
- Numerical summaries like **measures of center** and **measures of spread**

Data set: a collection of data values

Data points: the individual data values in the data set

Raw data: data as it was first gathered before any summarizing or computational manipulation

N is the size of the data set (the population of data)

Frequency: how often a particular data value occurs

Outliers: Extreme values in the data that do not fit the overall pattern of the data.

Create a Frequency Tally

Make a frequency tally for this **data set** of exam scores. This is a first step in **organizing data.**

95, 90, 85, 90, 70, 15, 70, 50, 55, 80, 70, 80, 60, 45, 70, 75, 75, 75, 60, 65

100	
95	
90	
85	
80	
75	
70	

65	
60	
55	
50	
45	
40	
35	

30	
25	
20	
15	
10	
5	
0	

Create a Frequency Table

For the frequency tally above, make a frequency table.

Important: In a Frequency Table, you only include the scores that actually happened.

Score												
Frequency												

Relative Frequency

Relative Frequency: the percent of the total population that had that value rather than the actual number that had that value.

Relative frequency is used most commonly when the actual frequencies are very large numbers. This makes them easier to compare.

Important: If you are graphing relative frequencies, be careful to:

- Include the N-value in the title of the so that the actual data values can be reconstituted, if desired
- Be sure to label the frequency scale as "relative frequency" or "percent of total".

Example: Find the relative frequencies for these raw data

What do we need to find first????

Score	**Frequency**	**Relative Frequency**
5	3,500	
4	2,000	
3	1,250	
2	1,000	
1	250	

Example: In a high stakes exam used for academic scholarship awards, N=200,000 and the relative frequency of the of a perfect score is 0.04%. How many students made a perfect score on the exam?

Bar Graphs of Frequencies

Bar graphs are often used to show frequencies. The higher the bar, the more frequent that data value .

Pictograph: uses pictures or icons to create the length of the bars

Characteristics of Bar Graphs:

- Bars are separated from each other, not right up against each other
- Height of bar indicates the frequencies of each score (or length of bar in horizontal)
- Bar graphs are usually limited to 12 or fewer bars. More than that is difficult to read.

Steps for creating a proper bar graph:

Step 1: **Organize the data values**

Step 2: Make the vertical scale (usually) the frequencies. Use equal intervals and be sure to label the scale “frequency”

Step 3: Make the horizontal scale the possible values. Make it an equal interval scale, including values that did not occur. Include a word-label that tells what those values represent.

Step 4. Draw a bar above each value that did occur, making the bar as long as the frequency for that value. Keep the bars more narrow than the space between values so that the bars do not touch one another.

Example:

Create a bar graph to display this data set as <u>relative frequencies.</u> The letters represent the letters of the correct answers on a multiple choice test.

A B A C B C C B C A A A C C B B B C B B C A B A B
C B B B C B A A B A B C C A A A C C B B A A A B C

Step 1:
Organize the data values
↓

Steps 2-3 →
Graph the values

Misleading Graphical Representations

Example from p. 553 of text: “Cheating” on the choice of starting value on vertical axis and stretching the scale on the vertical axis to make it look like there is more change than there actually is.

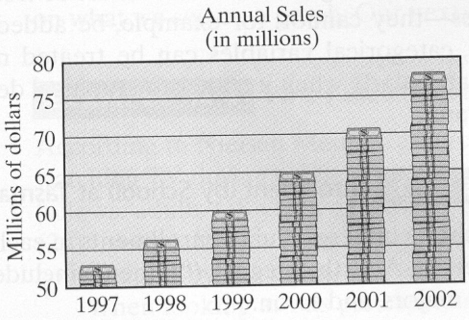


FIGURE 14-4
XYZ Corp. annual sales.

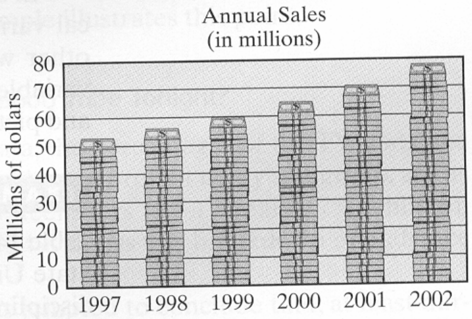


FIGURE 14-5
XYZ Corp. annual sales.

Example from P. 555 of text: When comparing characteristics of a population that is broken up into categories, it is essential to take into account the relative sizes of the various categories.

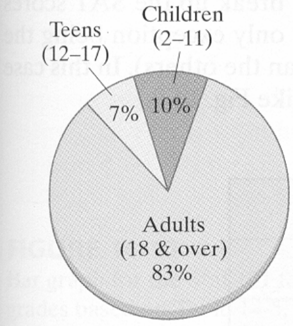


FIGURE 14-8
Audience composition for prime-time TV viewership by age group.
(*Source:* Nielsen Media Research.)

In the General Population:

Children (2-11 years) comprise 15% of the U.S. population

Teens (12-17) comprise 8% of the U. S. population

Adults comprise approximately 75% of the U. S. population

Creating Circle Graphs (Pie Charts)

Circle Graphs (Pie Charts) are good for showing the respective sizes of categories within a whole population. The circle represents the whole and the size of the sectors (the "slices") are proportional to the relative frequency of each category.

Remember 25% may be thought of as the ratio $\frac{25}{100}$.

Also remember that a circle contains 360˚.

It is often helpful to work with the reduced fraction for the percent rather than the number of degrees.

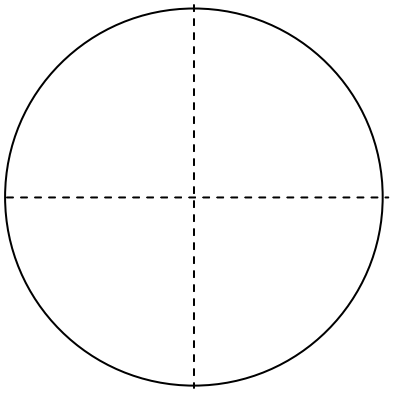
So the basic proportion relationship when figuring out the size of each sector is:

$$\frac{\%}{100} = \frac{\text{size of sector in degrees}}{360 \text{ degrees}}$$

Make a Circle Graph to Represent the following data:

N = 100,000 marbles that are either red, blue, green, yellow, or purple.

Type	Percent	Degrees
Red	25%	
Blue	50%	
Green	10%	
Yellow	10%	
Purple	5%	



Section 14.2 Variables

Variable: in statistics a variable is any characteristic that varies within the population. Examples: what color, what size, what kind, how many of . . .

Categorical Variable: (qualitative variable) represents a quality that is not normally measured numerically. For instance, gender.

- Categorical variables can be counted and quantified but we need to be very careful how we use those values and how we interpret them. We should not be giving a numerical average for variables like gender or hair color that are not actually numerical values to start with.

Numerical Variable: (Quantitative variable) represents a measurable quality.

- **Discrete numerical** variables are characteristics that cannot be measured in "infinitely" small fractions of a value (counting actual people, test scores, shoe sizes)

- **Continuous numerical** variables are characteristics that can be measured in "infinitely" small fractions of a value (time, distance traveled, volume of liquid)

- In the real world this distinction is blurred by
  - **Rounding off** values that are actually continuous so they seem discrete (like measurements of length to the nearest quarter inch)
  - **Calculations** or subdividing that create as many decimal places as necessary
    (often number of people, like 2.5 children per family is an average)

Practice: For each situation below, indicate whether the variable should be considered Categorical or Numerical. If Numerical, is it discrete or continuous?

	(a) Hair color: blond, brown, black, red, grey
	(b) Gender: male, female
	(c) Shoe size: $4, 4\frac{1}{2}, 5, 5\frac{1}{2}, 6, 6\frac{1}{2}, 7, 7\frac{1}{2}, 8, \ldots$
	(d) Ethnicity: Black, Native American, Hispanic, . .
	(e) Height in inches
	(f) Gender when asked to record it as a "1" if female and a "2" if male

Histograms and Class Intervals (text pp. 555-558)

Histogram: a variation of a bar graph showing relative frequencies.

Remember relative frequencies are the percent of the total population that had that value. In really large data sets where the raw values are very close together, we frequently group the raw data values into **equal-sized classes** and counting how many raw data values fall in each of these **class intervals**.

Important: When creating histograms, it is mathematically correct to draw bars for adjacent categories touching one another (different from regular bar graphs). This is because the class intervals are continuous, where as the original categories discrete. (letter grades like A, B, C, D, F are discrete, but class intervals like 89-100 are continuous).

Example:
GPA's for all students at MSUM would be an example of a large data set where the values are not well-separated. The values only run from 0 to 4 and are computed to 3 decimal places so you get values like 3.725, 3.724, 3.726, 3.725 all of which are very close together and are pretty much the same GPA.

Teachers frequently make histograms of grades grouped by A, B, C, D, and F to look at the type of grade distribution in their classes. This means that A+, A and A- are all recorded in the same bar. This is another example of data where the values run together and can be categorized together.

Class Practice: Make a histogram of the following quiz averages. Use the final grade chart below to group the quiz averages into letter grades.

72 85.5 93.5 68 73.5 82.5 80 79.5
56.5 87.5 89.5 71 79.5 86 75 76.5
83 86.5 78 67

Grading Scale:

A	91-100
B	81-90
C	71-80
D	61-70
F	60 -

Remember that the frequencies are not the scores, but ________________________________.

Section 14.3 Numerical Summaries of Data (text p. 558)

Another way to summarize data and make large data sets more comprehensible, is to numerically summarize them numerically. There are two main ingredients in such a summary:

- **Measures of Location** (how the data "line-up" in an ordered list of the values)
  - **Mean** (Tells the **center** of the data)
  - **Median** (Tells the **center** of the data)
  - **Percentile** (Tells the **percentile-rank** the data value)
  - **Quartile** (Tells the **center AND the quarter-marks**)
  - **"The Five-Number Summary"**
    (Minimum, Quartile 1, Median, Quartile 3, Max)
  - **Box-Plots** (Graphic representation of the 5-score summary)
- **Measures of Spread** (how the data "bunch-up")
  - **Range** (Max – Min)
  - **Interquartile Range** (Q3 –Q1)
  - **Variance** (an intermediate step to get to the standard deviation)
  - **Standard Deviation** (average distance from the mean)

MEAN

Mean: The mean is the arithmetic average.

- It tells you where the center of the data is in terms of "weight" (balance point) or "volume" (equally full point -- If you think of the bars in the bar graph as being tubes filled with liquid, the average would be how full each tube would be if you used all the liquid, but completely evened-out the liquid so that each tube had the same amount in it.

- **To find the mean**: Add up the scores and divide by the number of scores in the list.

- There are other types of averages, but the mean is the most commonly used. It has one drawback as a measure of center – it can be strongly influenced by extreme outliers.

- **Be careful when you compute the mean for a frequency table of values. Working from a frequency table is a weighted average.**

(see the next example)

Computing a Mean from a Frequency Table

Example: For the frequency chart below, you cannot get the average by adding just the numbers in the first row to get the total of the scores BECAUSE, for instance, the score of 9 did not happen just once, it happened 10 times.

If you need to compute the mean from a frequency table, it is easiest to add another row to the table that gives the “weighted” score. To get the weighted score, you multiply the exam score by the number of times it happened (the frequency). For instance, the score of 9 happened 10 times, so its weighted score is 9×10 = 90 because that is what you would get if your wrote out all 10 of them and added them up.

Point Scores on a 24-Point Exam													
Exam Score	1	6	7	8	9	10	11	12	13	14	15	16	24
Frequency	1	1	2	6	10	16	13	9	8	5	2	1	1

Weighted Scores													

Then the mean of ALL OF THE SCORES = $\dfrac{\text{Total of all of the Weighted Scores}}{\text{Total of all of the Frequencies}}$

Now finish computing the mean for this Exam.

Median

Median: The median the other measure of center that you are responsible for.

- It is the value which occurs in the middle of the data when the data are put in numerical order. So it is the center of the data, physically. Half of the data values are above the median and half are below it in this ordered list.

- Computing the Mean.
  - If there are an odd number of values in the ordered list, it is the center value.
  - If there are an even number of values in the ordered list, it is the average of the center two values.

- The Median is useful because it is a measure of center that is not influenced by extreme outlier values.

Class Practice:

1. Data Set: 5, 10, 8, 7, 4, 10, 9, 5, 7, 8, 2, 1, 7, 6, 10

(a) what is the mean of this data set?

(b) what is the median of this data set?

2. Data Set: 100, 90, 70, 90, 20, 80

(a) what is the mean of this data set?

(b) what is the median of this data set?

3. Data Set:

Score	0	4	6	7	8	9	10
Frequency	2	1	2	1	5	8	6

(a) what is the mean of this data set?

(b) What is the median of this data set?

Percentiles

Percentile: Percent and Percentile are NOT the same thing.

- On a test, a score of **<u>90 percent</u>** means that you got, proportionally speaking, 90 out of 100 correct. It compares the amount you scored <u>to the total possible score</u>.

- A percentile-rank compares how you did compared to everyone else. A **"90th percentile**" score means that, proportionally speaking, you did as well or better than 90% of the people who took the test. It compares your score <u>to all the other scores.</u>

Compute a Percentile:

Step 1: List the data in numerical order, from least to greatest.

Remember Percentile tells the relative position in the ordered list of data. So think of the data as being an ordered list of values like

$d_1,\ d_2,\ d_3,\ d_4,\ \ldots, d_N$ where each of these have a numerical value, but they also have a position value (the subscript)

Step 2: Compute the **locator** for the p^{th}–percentile using the formula:

$$\text{L} = \left(\frac{p}{100}\right)N\ .$$

Step 3:

When L is a whole number: the $p^{th} - percentile = \dfrac{d_L + d_{L+1}}{2}$

When L is not a whole number: the $p^{th} - percentile =$ L rounded up

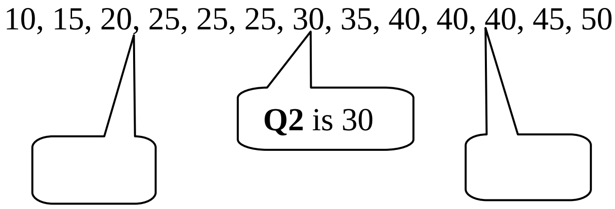
Try It Yourself: Find the 80th percentile value for the following GPA's:

3.4, 3.9, 3.3, 3.6, 3.5, 3.4, 4.0, 3.7, 3.3, 3.8, 3.6, 3.9, 3.7, 3.4, 3.6

Quartiles

The Quartiles divide the data set into four quarters. The data are first ordered.

Then find the middle of the data (which is the median) and that is Quartile 2 (abbreviated **Q2**). 50% of the data are below **Q2** and 50% of the data are above **Q2**.



10, 15, 20, 25, 25, 25, 30, 35, 40, 40, 40, 45, 50

Q2 is 30

Q1 (the first quartile) is the point below which 25% of the data occur. (If there are an even number of scores, average the two middle scores)

Q3 (the third quartile) is the point below which 75% of the data occur.

Q4 (the fourth quartile mark) of course, would be the highest value in the list so that 100% of the data falls below that point.

Generally we only talk about **Q1** and **Q3**. Instead of talking about **Q2** we call it the **Median** and instead of **Q4** we call it the **Maximum.**

Example 14.14 (p. 563 of text)

During the last year, 11 homes sold in the Green Hills subdivision. The selling prices, in chronological order, were:

Find the Median and Quartiles for this situation.

$167,000
152,000
128,000
134,000
192,000
163,000
121,000
145,000
170,000
138,000
155,000

Five Number Summary (text p. 565)

Five Number Summary: consists of

• Minimum	Min
• First quartile	Q1
• Median	M
• Third quartile	Q3
• Maximum	Max

Example: Find the five-score summary for this data set:

7 4 10 8 5 6 4 6 1 3 7 5

Box Plots (sometimes called “box and whisker plots”) (text p. 566)

Box Plot: A graphical representation of the 5-number summary. Box Plots are good for comparing two similar data sets, for instance two different samples from the same population. It gives a visual way to assess whether the two samples are significantly different or not.

Step 1: Draw a scale that covers the entire range of values.

Step 2: Draw a box that has Q1 as the location of one end and Q3 as the location of the other end on this scale.

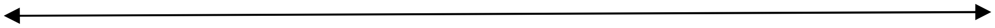
Step 3: Draw a line through the box indicating the Median, M.

Step 4. Locate the Minimum on the scale. Draw a “whisker” from the minimum to the nearest end of the box.

Step 5: Locate the Maximum data value on the scale. Draw a “whisker” from the maximum to the box.

Example: Draw a box plot for this following 5-number summary:

Min = 1, Q1 = 9, M=11, Q3=12, Max =24



Range (text p. 567)

Range: tells how spread out the data are. Take the difference (subtract) between the highest value and the lowest value of the data set. Notice that the range depends only on the most extreme values of the data: the highest (maximum) and the lowest (minimum) values.

Example: Find the range of this data set:

72 85.5 93.5 68 73.5 82.5 80 79.5 56.5 87.5

Example: Draw the box plot for the data in the previous example:

InterQuartile Range (p. 568)

Interquartile Range: the difference (subtraction) between the third quartile and the first quartile, Q3 –Q1.

Question: What percent of all the data points must lie within the interquartile range? Why?

Standard Deviation (text p. 568)

Standard Deviation: the most important and most commonly used measure of spread for a data set. It represents the average amount that the data points differ (deviate) from the **Mean**.

Step 1: Find the average (Mean) of the data set.

Step 2: Make a table with three columns and as many rows as there are data values.

Step 3. In the first column, put the data values. It makes things easier if you order them from least to greatest.

Step 4: In the second column put the answer to this computation (Data Value −Mean) for each row. Do all the subtractions in that order. Some of these values will be negative.

Step 5: In the third column, square the value in the second column.

Step 6: Total the third column. This is the sum of the squares of the deviations.

Step 7: The standard deviation, formally represented in statistics by the small greek letter "sigma"(σ) is equal

$$\sigma = \sqrt{\text{the average of (the squares of the deviations from the mean)}} = \sqrt{\frac{\text{total of column 3}}{\text{n}}}$$

Example: Find the Standard Deviation for this data set of test scores:
1, 6, 7, 8, 8, 9, 10, 11, 12, 13, 14, 15, 16, 24
