Methods for Describing Data Sets

2.1 Describing Data Graphically

In this section, we will work on organizing data into a special table called a frequency table. First, we will classify the data into categories. Then, we will create a table consisting of two columns. The first column will have the label for the categories. We will call those categories the "classes". The second column will tell the viewer how many data values belong in each category. We will call those counts the "frequencies."

A **<u>class</u>** is one of the categories into which data can be classified.

Class ~ Category

<u>Class frequency</u> is the number of observations belonging to the class.

Suppose for a second that after work one day, I said to a friend, "Wow, today 20 people in my class had green eyes!" My friend might respond, "Out of how many people?" Twenty people having green eyes in the class is only interesting if there were say 22 students in the room. Then 20 out of 22 is impressive, but if there were 200 students…not so special. This points to the need for the more commonly used concept of relative frequency. Think: The number of people with some trait relative to the whole group.

Relative Frequency = $\frac{Frequency}{n}$

Where n = total number of data values

The above number will always be a decimal between 0 and 1 inclusive, but if we prefer a percentage then we can convert relative frequency to percent by multiplying by 100.

Class percentage = (Class relative frequency) X 100

Let's practice with some real data from a recent study conducted by the Pew Research Center.

Example 6 Twenty-eight college graduates rated their job satisfaction in a survey for the Pew Research Center. Organize the data below into a frequency table:

Subject	Job Satisfaction	Subject	Job Satisfaction
1	Completely Satisfied	15	Somewhat Satisfied
2	Dissatisfied	16	Somewhat Satisfied
3	Somewhat Satisfied	17	Dissatisfied
4	Completely Satisfied	18	Somewhat Satisfied
5	Completely Satisfied	19	Dissatisfied
6	Somewhat Satisfied	20	Completely Satisfied
7	Somewhat Satisfied	21	Completely Satisfied
8	Completely Satisfied	22	Somewhat Satisfied
9	Somewhat Satisfied	23	Somewhat Satisfied
10	Somewhat Satisfied	24	Completely Satisfied
11	Dissatisfied	25	Somewhat Satisfied
12	Completely Satisfied	26	Completely Satisfied
13	Somewhat Satisfied	27	Somewhat Satisfied
14	Somewhat Satisfied	28	Completely Satisfied

Solution: We will use the job satisfaction levels as our classes. Then we will count the number of subjects that have reported each level of satisfaction.

Job Satisfaction	**Frequency**
Completely	10
Somewhat	14
Dissatisfied	4
Total	28

Example 7 Convert the frequency table above into a relative frequency table.

Solution:

Job Satisfaction	**Relative Frequency**	**Class Percentage**
Completely	10/28 = 0.357	35.7%
Somewhat	14/28 = 0.500	50.0%
Dissatisfied	4/28 = 0.143	14.3%
Total	28/28 = 1.00	100%

*Notice something about this relative frequency table: the relative frequencies must add up to 1.00.

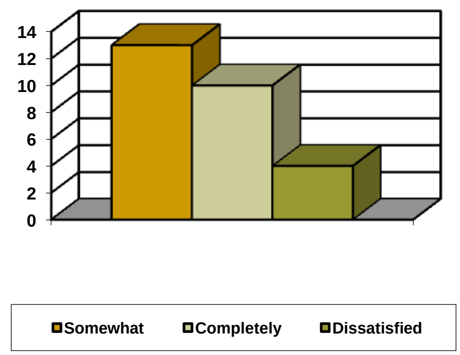
Finally, a lot of the ways to organize qualitative data are familiar to us all (e.g., pie charts, bar graphs,…), but one graph might be new to you.

A **Pareto Diagram** is a bar graph that arranges the categories by height from tallest (left) to smallest (right).

Here is a side by side comparison between a bar graph and a Pareto diagram:

Bar

Pareto



Graphical Methods for Describing Quantitative Data

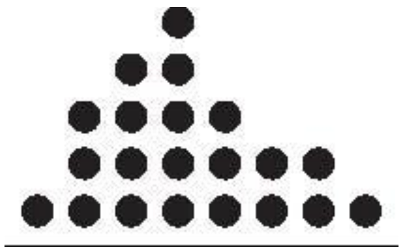
In this section, we are dealing with numerical data, so we will discuss three kinds of graphs:

- Dot plots
- Stem-and-leaf diagrams
- Histograms

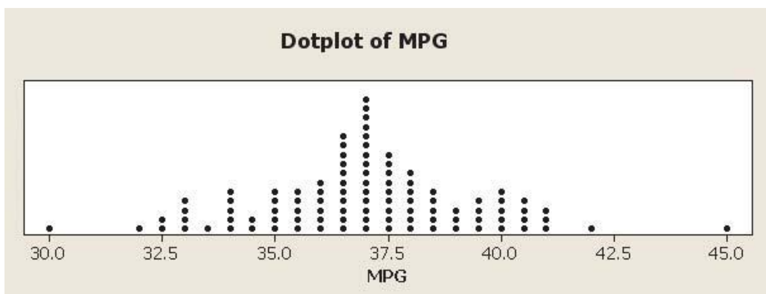
Dot plots display a dot for each observation along a horizontal number line

--Duplicate values are piled on top of each other

--The dots reflect the shape of the distribution



Here is an example where we plotted MPG ratings for autos on a number line. Each dot represents a car's MPG rating from the study:



I love this kind of graph. It is so simple and yet so helpful.

The next kind of graph is called a stem-and-leaf display.

- A **stem-and-leaf display** shows the number of observations that share a common value (the stem) and the precise value of each observation (the leaf).

1	3
2	2489
3	126678
4	37
5	2

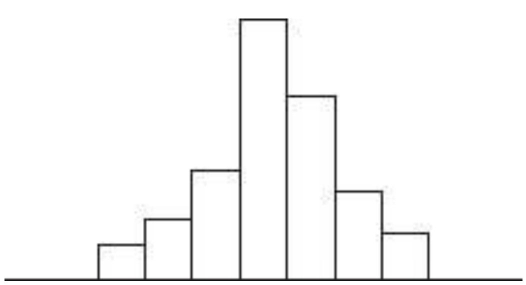
The numbers can be seen clearly in this display. Can you tell what the original numbers were assuming they were 2 digit whole numbers? Answer: 13, 22, 24, 28, 29, 31, 32, 36, 36, 37, 38, 43, 47, 52. If you turned this graph onto its side it would form a shape like the dot plot! The advantage is that you can easily see the original numbers from the data in the graph.

- Below is another example, which list the number of wins by teams at the MLB 2007 All-Star Break:

	stem unit = 10	**leaf unit =1**
Frequency	**Stem**	**Leaf**
9	**3**	**4 6 6 8 8 8 8 9 9**
17	**4**	**0 0 2 2 3 4 4 4 4 5 7 7 8 9 9 9 9**
4	**5**	**2 2 3 3**
n = 30		

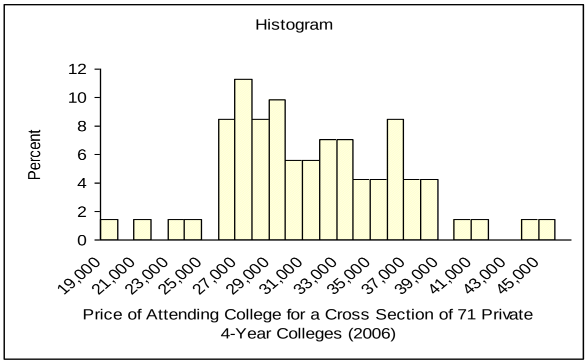
The final kind of graph, perhaps the most important of the three, we will discuss is the histogram.

- **Histograms** are graphs of the frequency or relative frequency of a variable.
  - Class intervals make up the horizontal axis (x-axis).
  - The frequencies or relative frequencies are displayed on the vertical axis (y-axis).



Histograms are like bar charts for numerical data, but they never have gaps between the bars (unless the frequency for the class is zero).

Here is an example where there are categories with no frequency (that is why some of the bars have spaces between them):



We are going to spend some time on learning to create histograms by hand, but understand that these graphs are often better left to computer software programs that can rapidly create them to perfection. Ok, before we begin creating a histogram, let's recap and expand what we have already learned about the frequency table.

Recall: A **relative frequency distribution (or **<u>frequency table</u>**) lists data values (usually in groups), along with their corresponding <u>relative frequencies</u>. ***"Relative"*** here means relative to our **sample size (n).** Also, ***"Frequencies"*** is just another way to say **counts.**

$$\textbf{Relative Frequency} = \frac{Frequency}{n}$$

In a relative frequency distribution, each data value belongs to an interval of numbers called a **<u>class</u>**. Each class has a lower and upper class limit that define the interval.

<u>Lower class limits</u> are the smallest numbers that can belong to the different classes.

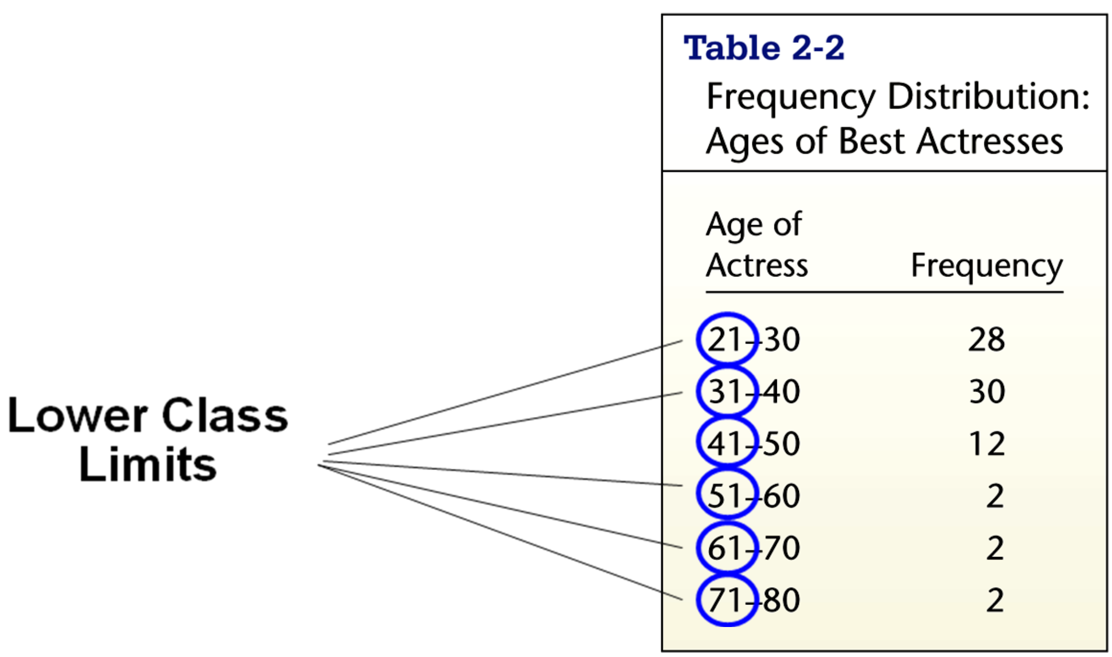


Table 2-2
Frequency Distribution: Ages of Best Actresses

Age of Actress	Frequency
21-30	28
31-40	30
41-50	12
51-60	2
61-70	2
71-80	2

Upper class limits are the largest numbers that can belong to the different classes.

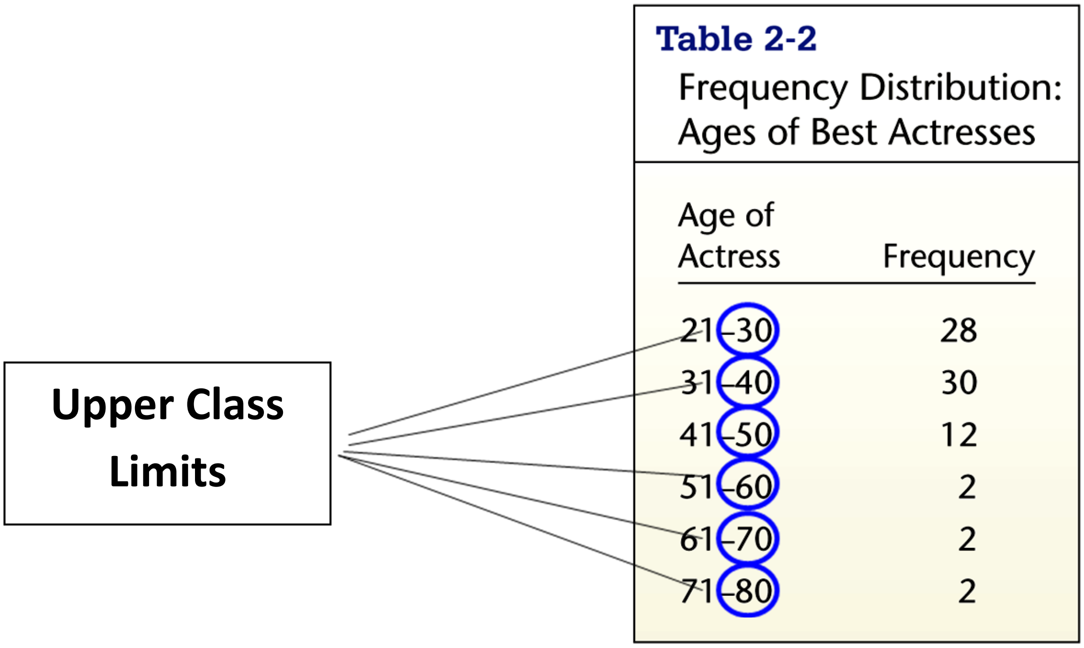
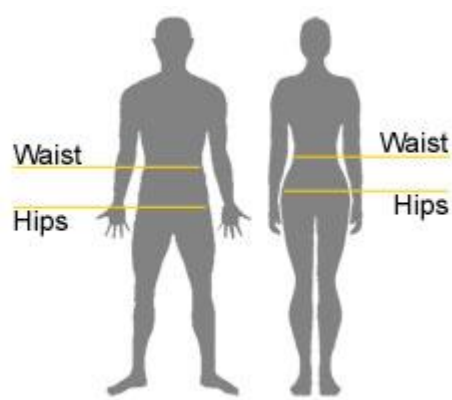


Table 2-2
Frequency Distribution: Ages of Best Actresses

Age of Actress	Frequency
21–30	28
31–40	30
41–50	12
51–60	2
61–70	2
71–80	2

Sample Data – The data below is from a study of waist-to-hip ratios (waist circumference divided by hip circumference) for Playboy centerfold models (1953 – 2001). The **lower class limits** are 0.52, 0.56, 0.60, 0.64, 0.68, 0.72, and 0.76. The **upper class limits** are 0.55, 0.59, 0.63, 0.67, 0.71, 0.75, and 0.79.

Waist-to-Hip Ratio	Frequency
0.52 – 0.55	6
0.56 – 0.59	17
0.60 – 0.63	86
0.64 – 0.67	230
0.68 – 0.71	208
0.72 – 0.75	23
0.76 – 0.79	6



Class boundaries are like the class limits, in that they use a range of values to define the class, but they do so without the gaps that are sometimes created by class limits.

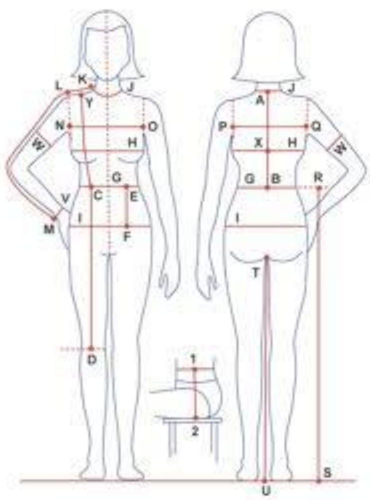
Class boundaries are obtained as follows:

Step 1: Subtract the upper class limit of the first class from the lower class limit of the second class.

Step 2: Divide the number we found in step one in half.

Step 3: Subtract the result from step two from the first lower class limit to find the first class boundary; add the result from step two to each upper class limit to find the rest of the upper class boundaries.

Example 8- Find the class boundaries for the frequency table for waist-to-hip ratios for centerfold models.

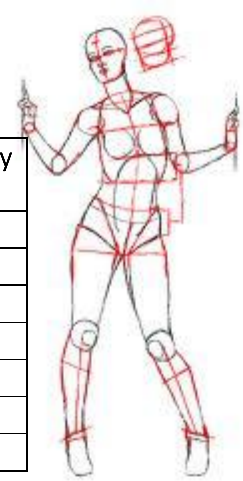


Waist-to-Hip Ratio	Class Boundaries	Frequency
0.52 – 0.55		6
0.56 – 0.59		17
0.60 – 0.63		86
0.64 – 0.67		230
0.68 – 0.71		208
0.72 – 0.75		23
0.76 – 0.79		6

Class midpoints are the midpoints of the classes. Each class midpoint can be found by adding the lower class limit to the upper class limit and dividing the sum by 2.

Example 9 - Find the class midpoints for the waist-to-hip ratio example.

Waist-to-Hip Ratio	Class Boundaries	Class Midpoints	Frequency
0.52 – 0.55	0.515 – 0.555		6
0.56 – 0.59	0.555 – 0.595		17
0.60 – 0.63	0.595 – 0.635		86
0.64 – 0.67	0.635 – 0.675		230
0.68 – 0.71	0.675 – 0.715		208
0.72 – 0.75	0.715 – 0.755		23
0.76 – 0.79	0.755 – 0.795		6



The **class width** is the difference between two consecutive lower class limits or two consecutive lower class boundaries.

Example 10 - Find the class width for the waist-to-hip ratio example.

Solution 0.56 – 0.52 = 0.04 or you could use 0.79 – 0.75 = 0.04 as long as you take the difference between two consecutive lower (or upper) class limits you get the width for this table which is 0.04.

We should learn how to create these frequency tables when we are presented with a set of numerical data. Here are some guidelines:

Guidelines for creating a relative frequency table:

1. **Determine the number of classes to use:**
   Using 5 to 20 classes is best, but the recommendation to use 5 to 20 classes is vague. Here is a little more detail:

<25 observations:	5-6 classes
25-50 observations	7-14 classes
>50 observations	15-20 classes

   (If we assume the population data has a bell shaped distribution, it is possible to use Sturges' formula to determine the number of classes: $K = 1 + 3.3219 * \log n$, where K is the number of classes, and n is the number of values in the data set.)

2. **Calculate the Range:** Range = Highest value – Lowest value
3. **Determine the class width:**

   Class width $\geq \left(\frac{Range}{number of classes} \right)$ *Round up if necessary

4. **Select the lower limit of the first class:** This lower limit should be either the lowest data value or a convenient number that is a little bit smaller. This value is referred to as the starting point.
5. **Use the class width to obtain the other lower class limits:** For example, add the class width to the starting point to get the second lower class limit.
6. **Determine the upper class limits:** Place the upper and lower class limits in a column
7. **Determine the frequencies:** Count the number of values belonging to each class
8. **Calculate the relative frequencies:** Divide each of the values in the frequency column by the total number of data values.

Example 11 - A medical research team studied the ages of patients who had strokes caused by stress. The ages of 34 patients who suffered stress strokes are below. Construct a frequency distribution for these ages. Use 8 classes beginning with a lower class limit of 25.

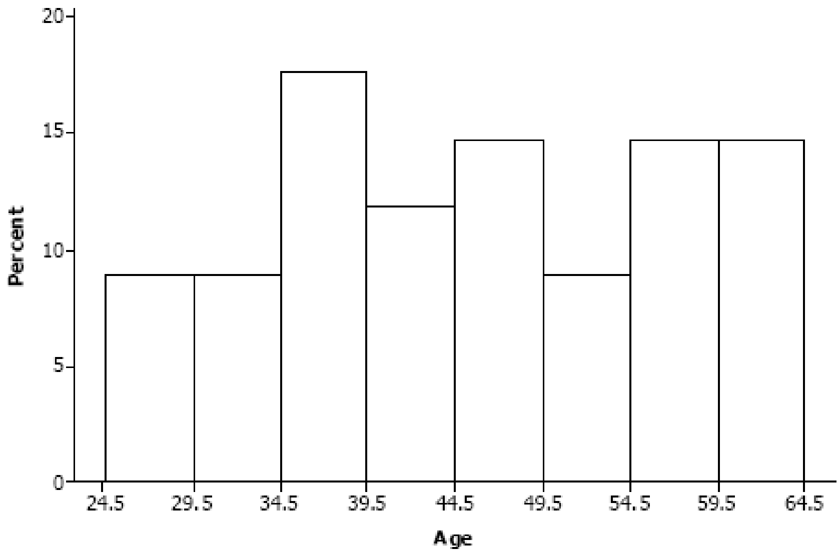
27 28 29 30 32 32 36 36 36 36 38 38 40 40 40 41 45 46 46 46 47 50 50 50 55 56 57 58 58 60 61 61 61 62

A **relative frequency histogram** is a bar graph in which the heights of the bars represent the proportion of occurrence for particular classes. The **classes** or categories/subintervals are plotted along the horizontal (x) axis, and the vertical scale is based on the relative frequencies.

Example 12: Construct a relative frequency histogram for the 34 ages of patients who suffered stress strokes using the table below:

Age	Relative frequency
25 - 29	$\frac{3}{34} \approx 8.82\%$
30 - 34	$\frac{3}{34} \approx 8.82\%$
35 - 39	$\frac{6}{34} \approx 17.65\%$
40 - 44	$\frac{4}{34} \approx 11.76\%$
45 - 49	$\frac{5}{34} \approx 14.71\%$
50 - 54	$\frac{3}{34} \approx 8.82\%$
55 - 59	$\frac{5}{34} \approx 14.71\%$
60 - 64	$\frac{5}{34} \approx 14.71\%$

Solution: The class boundaries are 24.5, 29.5, 34.5, 39.5, 44.5, 49.5, 54.5, 59.5, and 64.5.



*Note: if the class widths are not all the same, we will use % / (x axis variable) for the y axis.

Histograms when the Class Width is not Uniform

Histogram Steps (for unequal class widths):

1. Use the steps from above to create the classes.

2. Fill in the frequencies.

3. Convert the frequencies into percents.

4. Find the height of each bar (rectangle).

   *Recall from geometry that the area of a rectangle is area = height X width, so H = A/W. Also, remember that area here is the same as percent. Width is the width of the specific class you are working with.

5. Draw and label the x and y axis.

   *The y-Axis label is called the Density Scale, and it is equal to percent per X unit.

6. Draw the bars (rectangles).

Left End Point Convention for Continuous Data: the bars of the histogram contain the left end point but not the right. This is also the convention used when a frequency table has no gaps between the classes.

Example 12.5 Create a histogram for the frequency table below:

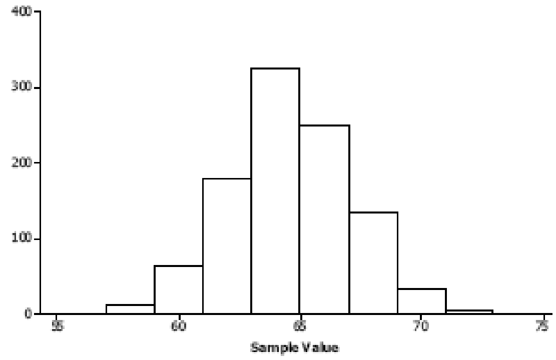
Time Spent Studying /Week	Frequency
0 – 2hrs	28
2 – 4	71
4 – 6	43
6 – 8	21
8 – 15	12

Histograms are often constructed to help analyze data. The shape of the histogram can be used to determine whether the data have a distribution that is approximately normal. Normal distributions correspond to histograms that are roughly bell-shaped (see figure below). Histograms can also be used to approximate the center of the data.

Example 13: Based on the histogram above from example 12, do the data appear to have a distribution that is approximately normal?

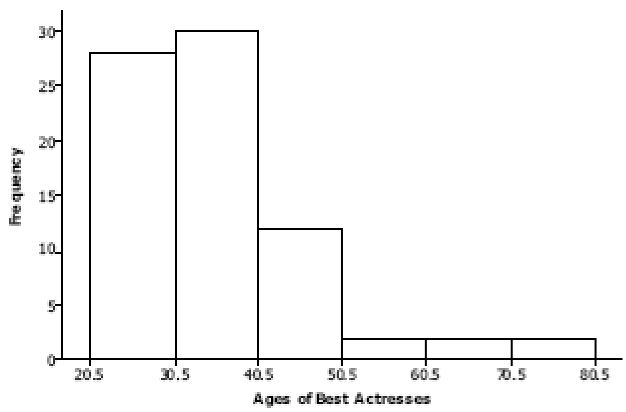
Solution: No, not based on the above drawing.

Example 14: Based on the histogram shown below, do the data appear to have a distribution that is approximately normal?



Solution: Yes, this is very close to a bell shaped distribution, so perhaps the data is normally distributed.

Example 15: Based on the histogram shown below, estimate the center of the data.



Solution: This is subjective, but I might say it is 39 or 38 yrs old. Imagine if you could lift the histogram off of the paper and hold it, the mean would be the location where you would need to place your finger so that half the weight (mass) is on the left and the other half of the weight (mass) is on the right.

2.2 Summation notation

Some people say math is a foreign language; well, this section definitely has parallels to learning the alphabet of a foreign language. In fact, some of the letters are indeed Greek. Below we break down a commonly used system of shorthand from mathematics:

- **Individual observations in a data set are denoted**
  $x_1, x_2, x_3, x_4, \ldots x_n$.

We use a Greek letter, capital Sigma, to denote summation. We use this symbol often, so it deserves special mention:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

The above notation tells us to add all the values of variable x from the first (x_1) to the last (x_n) .

Example, if If $x_1 = 1$, $x_2 = 2$, $x_3 = 3$ and $x_4 = 4$,

$$\sum_{i=1}^{n} x_i = 1 + 2 + 3 + 4 = 10$$

- Sometimes we will have to square the values before we add them:

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \ldots + x_n^2$$

- Other times we will add them and then square the sum:

$$\left(\sum_{i=1}^{n} x_i\right)^2 = \left(x_1 + x_2 + x_3 + \ldots + x_n\right)^2$$

Example 16: Using {2, 4, -3, 7, 1} find the following:

$$\sum_{i=1}^{5} X_i$$

$$\left(\sum_{i=1}^{5} X_i\right)^2$$

$$\sum_{i=1}^{5} X_i^2$$

Example 16.5 Using {-1, 5, 9, 6, 0}, find $\sum_{i=1}^{5}(x_i - 3)^2$

2.3 Numerical Measures of Central Tendency

There are two ideas that come up when trying to capture the information obtained in a sample of data. Those ideas are: what is the typical data value like from this population or sample, and how similar or clustered are the data values (or members of the population or sample)?

- When summarizing data sets numerically two ideas arise
  - Are there certain values that seem more typical for the data?
  - How typical are they?

Central tendency is the tendency of the data to cluster, or center, about certain numerical values.

Variability is the same as the spread or clustering of the data. Measures of v**ariability** are designed to show how strongly the data cluster around the center of the distribution.

Below we will discuss three common measures of *central tendency:*

The **Mean** is found by summing up all the measurements and then dividing by the number of measurements.

$$\overline{x} = \frac{\sum_{i=1}^{n} X_i}{n}$$

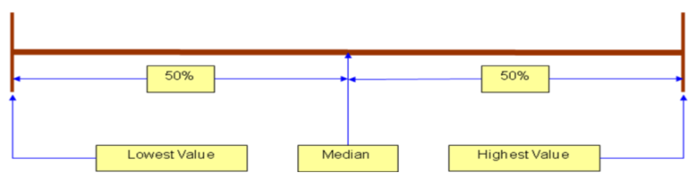
- The mean of a *sample* is typically denoted by x-bar, but the *population mean* is denoted by the Greek symbol μ *(pronounced mew)*.

$$\mu = \text{population mean}\,,\ \overline{x} = \text{sample mean}$$

Example 17: Find the sample mean for the numbers: 21, 2, 1, 3, 24, 120, 36, 1, 1, and 1.

Solution: $\bar{x} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{21+2+1+3+24+120+36+1+1+1}{10} = 21$

The **Median** is intuitively the middle number when the measurements are arranged in numerical order. The median is defined as a value such that at least half of the data values are less than or equal to it and at least half of the values are greater than or equal to it. The median is also called the 50^{th} percentile since 50% of the data is below the median and 50% is above.



Notice the median only looks at one or two numbers in the center of the data set. Doesn't that seem like a waste of information? It does to me, but on the other hand the median, because it ignores all the other values, isn't unduly affected by really big or small numbers in the data set. For example, what would happen if you averaged a list of numbers which were the personal net worths of 46 randomly selected Americans and Bill Gates net worth was part of the list? The average obtained would be artificially high, and would not truly capture the typical American's personal net worth. In this sort of situation, you may wish to use the median instead of the mean.

Guidelines for calculating the sample Median:

Arrange the sample data from smallest to largest

- If n is odd, M is the middle number.
- If n is even, M is the mean of the two middle numbers.

Notation for the median: Sample median is denoted by $\tilde{x}$ (x-tilda) and the population median is denoted by η (eta).

The **Mode** is the data value that occurs most frequently.

Example 18: Find the mean, median, and mode for the following data set: 75, 75, 70, 80, 97, 53, 60, and 90.

Solution: mean = 75, median =75, and the mode = 75.

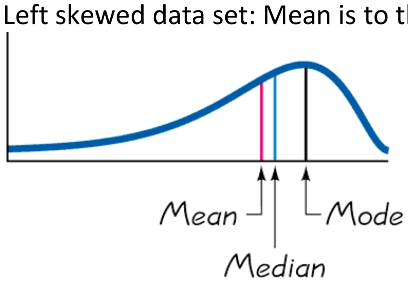
So what is the best measure of the center? Well, that depends on the data set you are dealing with. In general, the sample mean is preferred because it varies less as an estimate of the true center, and it uses input from every data value. However, when extreme values are present (very big compared to the rest of the data or very small) the median is generally preferred. Yet, if the data is qualitative how can you use either mean or median?

Remember, these measures are supposed to help us state what the typical member of the population is like. For example, what is the typical eye color? The answer is probably brown—that is the modal eye color in the USA, correct? You can't add and divide eye colors to find an average, nor could you put them in order to find the middle one, so we use the mode. For qualitative data, the mode is a good choice.

2.4 Skewed Distributions

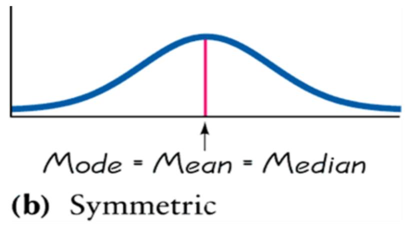
A distribution is **Skewed** when one side of the distribution has more extreme values than the other. If the population mean is greater than or less than the population median, the distribution is skewed.

- Left skewed data set: Mean is to the left of the median

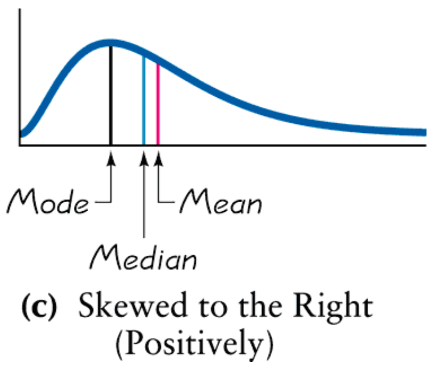


(a) Skewed to the Left (Negatively)

■ Perfectly symmetric data set:
Mean = Median = Mode

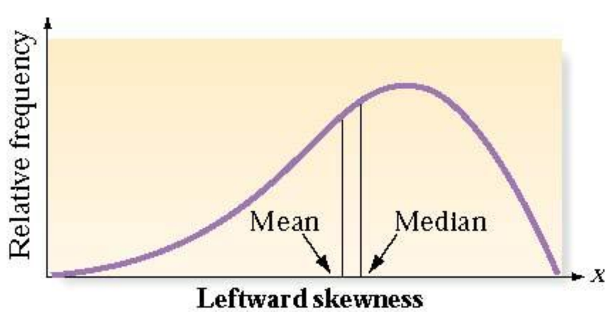


■ Right skewed data set: The mean is on the right of the median.



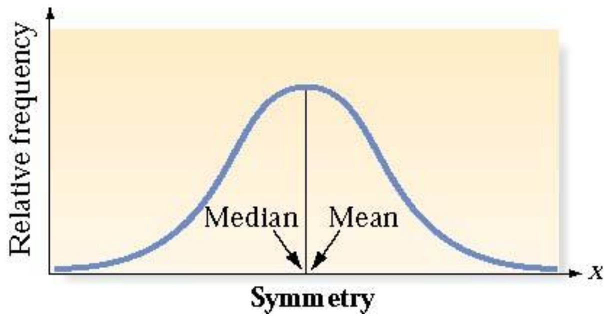
Example 19: If the population mean for a group is 37 and the population median is 45, is the population most likely left-skewed, right-skewed, or symmetric?

Solution: Left-skewed, the mean < median.



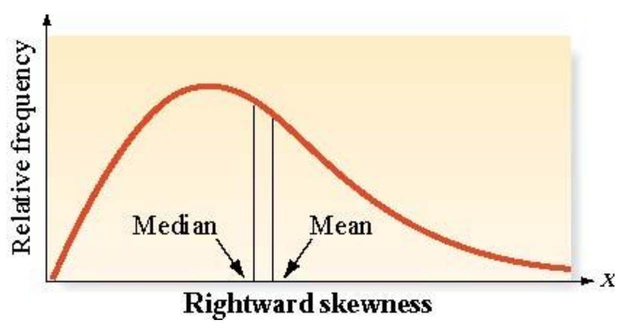
Example 20: If the population mean for a group is 37 and the population median is 37, is the population most likely left-skewed, right-skewed, or symmetric?

Solution: Symmetric, the mean = median.



Example 21: If the population mean for a group is 103 and the population median is 95, is the population most likely left-skewed, right-skewed, or symmetric?

Solution: Right-skewed, the median < mean.



2.5 Measures of Variability

Remember, in the last section we spoke about the idea of central tendency (literally, the tendency for data to cluster around some central point). Think of this tendency like a magnetic force and data points on the number line as little balls of iron. A strong magnet would draw all of the points to it leaving no points far away from the center of its magnetic pull. A very weak magnet may not be able to create as much clustering in the center, leaving lots of little points spread out around it at varying distances.

We want to be able to measure the strength of central tendency in a population. That way we know how clustered together the data values are. This is very important because if the values are all clustered tightly around some central point we know a great deal about the population just by knowing its center. For example, in very clustered data sets, we might be able to safely stereotype, because nearly all of the population will exhibit traits similar to the center. Prediction becomes easy when this is the case. We will call these measures of the strength of the central tendency the **measures of variability**.

<u>Variability</u> is the same as the spread or clustering of the data. Measures of v**ariability** are designed to show how strongly the data cluster around the center of the distribution.

You know what the word 'vary' means. If a population's data values do not vary much, then the population has strong central tendency. The group is then more homogenous. Think of a 4 inch dry wall screw, the little black screws you can find at Home Depot. How many of those do you need to see to know what a 4-inch, dry-wall screw looks like? Just one, since they are all so similar. In other words, they do not vary much. If you know one, you know them all. When variation is high in a group, it is difficult to capture the essence of something with just one example from the population. Think of humans. Are all 20 year olds alike? If you know one, do you know them all? I would say no.

The question is how do you measure this variation? Here are some ideas:

The **<u>Range</u>** is the largest measurement minus the smallest measurement.

$$Range = Max - Min$$

(Benefits: easy to calculate, and easy to interpret; downsides: it is insensitive when the data set is large.)

The sample **<u>Variance</u>** for a sample of n measurements is equal to the sum of the squared distances from the mean divided by n – 1.

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

or the formula can be written $s^2 = \frac{n\sum x^2 - \left(\sum x\right)^2}{n(n-1)}$. This formula is the one I recommend.

Example 22: Calculate the sample variance for the following numbers: 2, 7, 3, 9, and 12.

The sample **Standard Deviation** is the square root of the sample variance.

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

Example 23: Calculate the sample standard deviation for the following numbers: 2, 9, 5, 7, 4, and 2.

- As before, Greek letters are used for populations and Roman letters for samples

The symbols:

Population Variance σ^2, sample variance s^2

Population Standard Deviation σ, sample standard deviation s

Notice something very important: the unit of measurement (i.e., ft., inches, miles, yrs, dollars,…) for standard deviation is the same unit the original data was measured in, and it is the same as the mean. Variance does not have the same units. Its unit will be the square of the units for the original data.

It is possible for us to estimate the standard deviation for a sample by using something called the **range rule of thumb**. We will see that at the end of the next section, but first let's find a way to interpret the standard deviation.

In the section, we will learn something about *where the data will lie relative to the mean*. In other words, we will learn how to determine how much data will be within certain distances from the mean.

2.6 Chebyshev's Theorem

Chebyshev's Rule: For any number greater than 1, the proportion of measurements that will fall within k standard deviations of the mean is at least $1-\frac{1}{k^2}$.

The k in the above formula is the number of standard deviations away from the mean. If we have a symmetric interval of the form: $[\mu - k\sigma, \mu + k\sigma]$, we know at least $(1-\frac{1}{k^2})$100% of the data will lie inside the interval.

Chebyshev's Rule

- Valid for *any* data set
- For any number *k* >1, at least $(1-\frac{1}{k^2})$100% of the observations will lie within *k* standard deviations of the mean

k	k^2	1/ k^2	(1- 1/ k^2)%
2	4	.25	75%
3	9	.11	89%
4	16	.0625	93.75%

Example 24: The average weight of 19 year old women in America is 148.6 pounds with a standard deviation of 23.9 pounds. What is the minimum percentage of 19 year old women who weigh between 96.0 and 201.2 pounds? What is the maximum percentage of women who weigh more than 201.2 pounds?

Example 24.5: Imagine you wanted to avoid having to pay for car repairs before your car is paid off, so

you are considering purchasing a new Toyota Corolla. What is the maximum percentage of Corollas that will require a major repair before you pay off the five year car loan? To answer this, assume the following: You will put 20,000 miles on the car per year for each of the first five years you drive it. This means, the car will reach 100,000 miles by the time you pay it off. Also, assume the average mileage for Toyotas undergoing their **first** major repair is 150,000 with a standard deviation of 10,416 miles.

Chebyshev's rule is great, because it does not assume anything about the distribution of the data. This is good because we often do not know the distribution of the data. However, if we can assume the distribution of the data is bell shaped, we have a more precise rule to turn to.

2.7 The Empirical Rule

- **The Empirical Rule**
  - **Useful for mound-shaped, symmetrical distributions**
  - **~68% will be within the range** $(\bar{x}-s, \bar{x}+s)$
  - **~95% will be within the range** $(\bar{x}-2s, \bar{x}+2s)$
  - **~99.7% will be within the range** $(\bar{x}-3s, \bar{x}+3s)$

Another way to write the same rule is as follows (notice how we express the intervals above in words below):

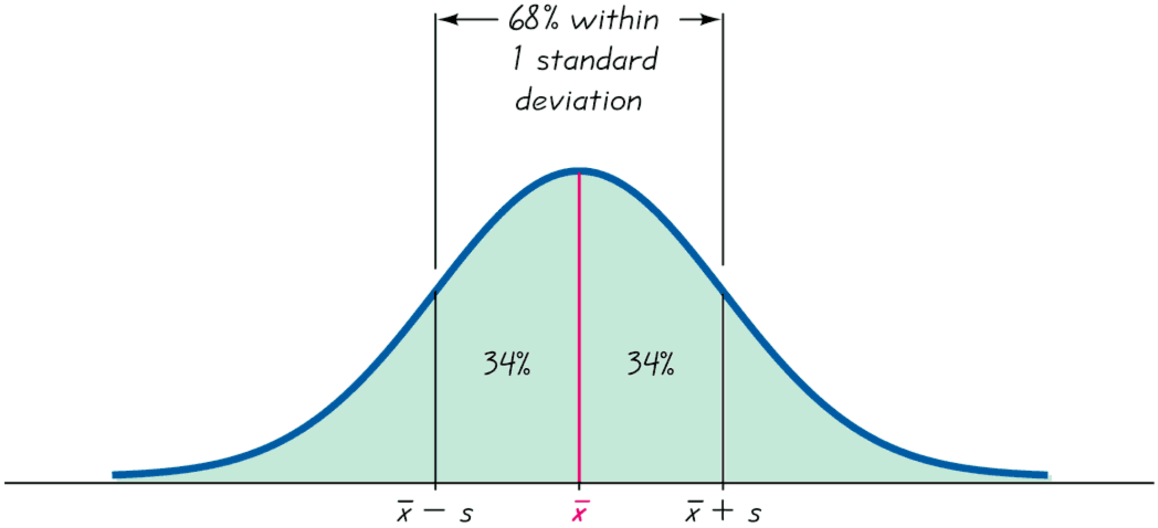
Empirical Rule:

68% of the measurements will fall within 1 σ of the mean

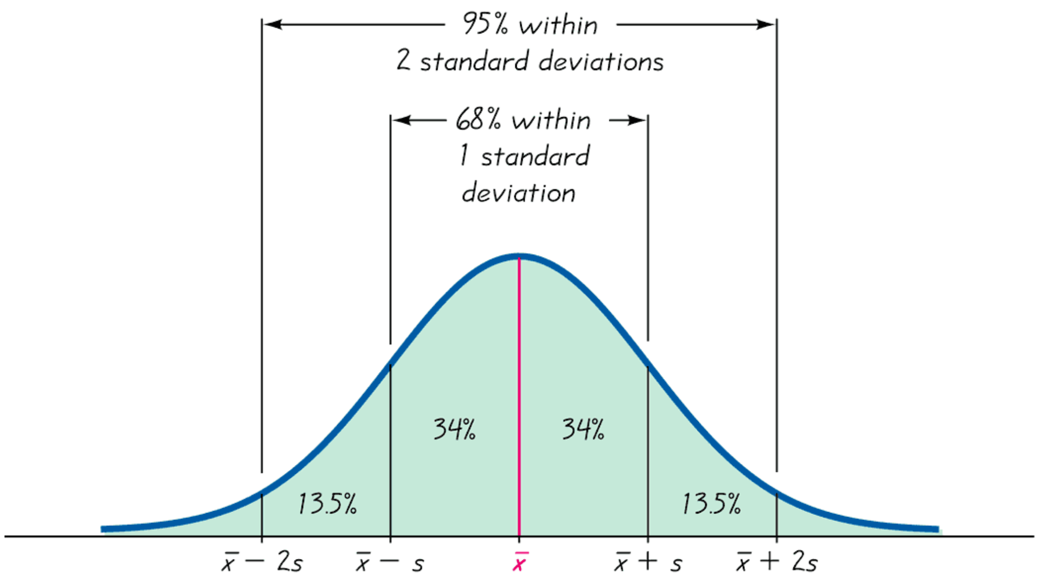
95% of the measurements will fall within 2 σ's of the mean

99.7% of the measurements will fall within 3 σ's of the mean

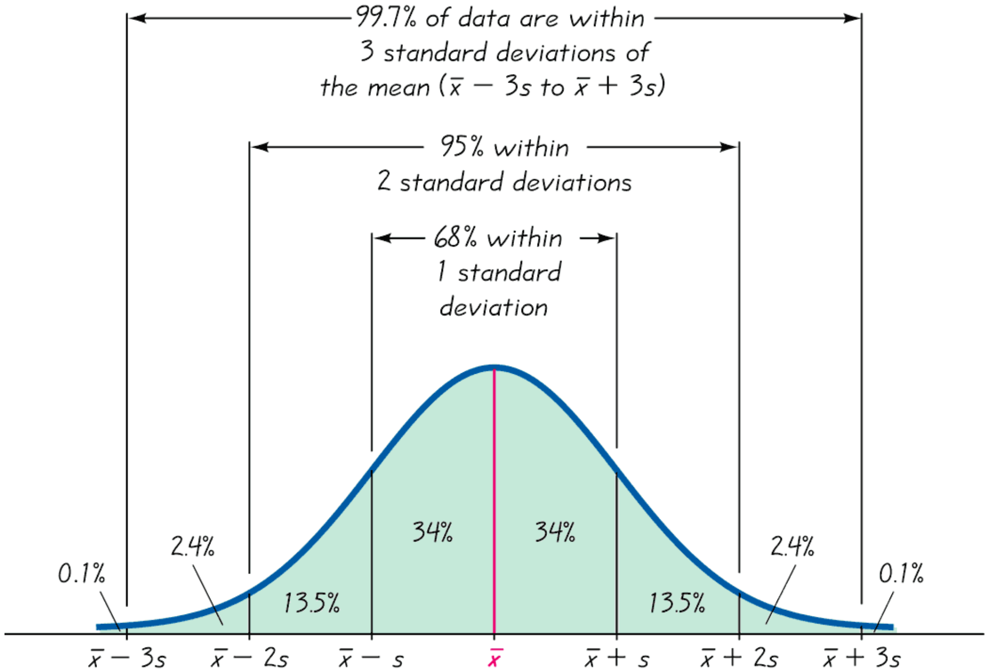
Finally, the pictures below will clear up any confusion you have about the empirical rule. The first drawing is of the interval of one standard deviation away from the mean.



The drawing below now considers two standard deviations away from the mean:



Finally, this drawing shows three standard deviations about the mean.



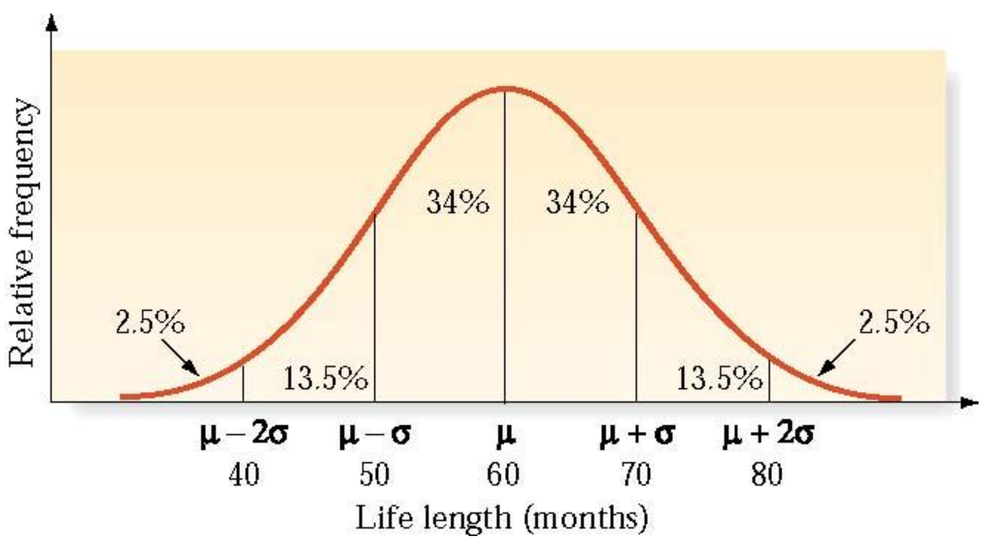
Example 25: The average weight of 19 year old males is 176.8 pounds. The standard deviation is 26.4 pounds. If there is reason to believe that the distribution of weights for 19 year old males is normally distributed (or bell shaped), what is the approximate percentage of 19 year old males that will weigh between 150.4 pounds and 203.2 pounds?

Example 26: Hummingbirds beat their wings in flight an average of 55 times per second. Assume the standard deviation is 10, and that the distribution is symmetric and mounded.

- Approximately what percentage of hummingbirds beat their wings between 45 and 65 times per second?
- Between 55 and 65?
- Less than 45?

Range Rule of Thumb

Now that we have these two rules to help us interpret the standard deviation, we can derive a rule of thumb for approximating the standard deviation. To estimate the value of the standard deviation, *s*, create the interval: $\left[\frac{R}{6},\frac{R}{4}\right]$ where R is the range = (maximum value) – (minimum value). The value for s should be between these two numbers. If you have a lot of numbers in your sample, use a value nearer to $\frac{R}{6}$. If you want to know why standard deviation can be approximated by $\frac{R}{6}$ or $\frac{R}{4}$, we should look at the drawing above of the bell curve. Since almost all of the values are between $\bar{x}-3s$ and $\bar{x}+3s$, it makes sense to think that the smallest value we are likely to encounter in a sample of data is $\bar{x}-3s$ and the largest would be $\bar{x}+3s$. Then, the range for that sample would be R = (maximum value) – (minimum value) = $(\bar{x}+3s)-(\bar{x}-3s)$ = $(\bar{x}+3s-\bar{x}+3s)=6s$. Now, if R = 6s, we can solve for s to get $s\approx\frac{R}{6}$. For small samples we are not likely to get data values as small or as large as $\bar{x}-3s$ and $\bar{x}+3s$, so we can use $\frac{R}{4}$ as an approximation for s.



- Since ~95% of all the measurements will be within 2 standard deviations of the mean, only ~5% will be more than 2 standard deviations from the mean.
- About half of this 5% will be far *below* the mean, leaving only about 2.5% of the measurements at least 2 standard deviations *above* the mean.

2.8 Measures of Relative Standing

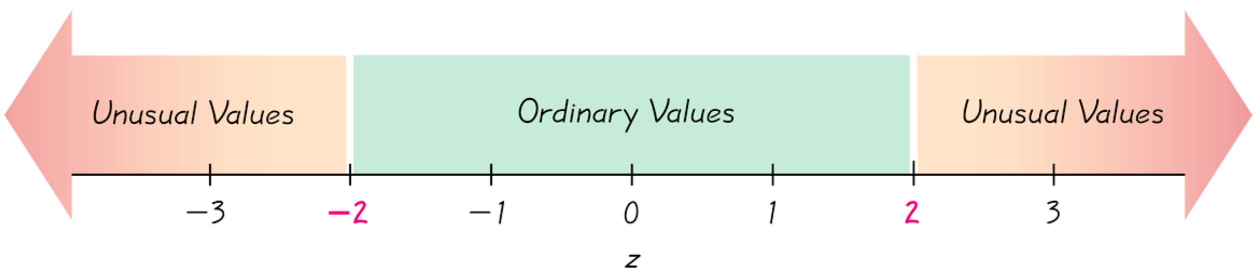
- This section introduces measures that can be used to compare values from different data sets, or to compare values within the same data set. The most important of these is the concept of the *z* score. A z - score (or standardized value) is the number of standard deviations that a given value *x* is above or below the mean.

All data values can be expressed as a z-score. It is just another scale available to us. Just like weight can be expressed in pounds or kilograms, we can express a given weight as a z-score. The formula to find a z-score for a given score (X) is as follows:

For sample data: $z = \frac{x - \overline{x}}{s}$ For population data: $z = \frac{x - \mu}{\sigma}$

Another way to view the **z-score** is the distance between a given measurement x and the mean, expressed in standard deviations.

Interpreting z-scores:



**Note: Any z-score that has an absolute value greater than 3 is considered an outlier, while |z-scores| between 2 and 3 are possible outliers. Also, whenever a value is less than the mean, its z-score is negative.

Z-scores and the Empirical rule:

Z scores are the number of standard deviations away from the mean, so using the empirical rule we can conclude:

For a perfectly symmetrical and mound-shaped distribution,

- ~68 % will have z-scores between -1 and 1
- ~95 % will have z-scores between -2 and 2
- ~99.7% will have z-scores between -3 and 3

Example 27: Compare the performance on the first exam for two different students in two different Statistics classes. The first student had a score of 72 in a class with a mean grade of 68 and a standard deviation of 5. The second student had a 70 in a class with a mean grade of 66 and a standard deviation of 4.

Example 27.5: According to a study on ethnic, gender, and acculturation influences on sexual behaviors published in 2008, the average age at the time of first intercourse for Hispanic females was 16.52 years old with a standard deviation of 2.25 years. Based on these numbers would it be unusual for a Hispanic woman to have waited until turning 21 years old before first engaging in intercourse?

Percentiles

For any set of n measurements arranged in order, the **pth percentile** is a number such that p% of the measurements fall below the pth percentile and (100 – p)% fall above it. For example, if you scored in the 94th percentile on the SAT, you did better than 94% of the exam takers and worse than 6%.

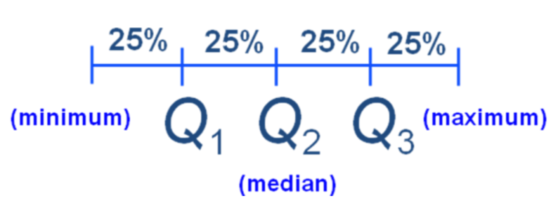
Three very important and commonly used Percentiles are:

1st Quartile = 25th percentile

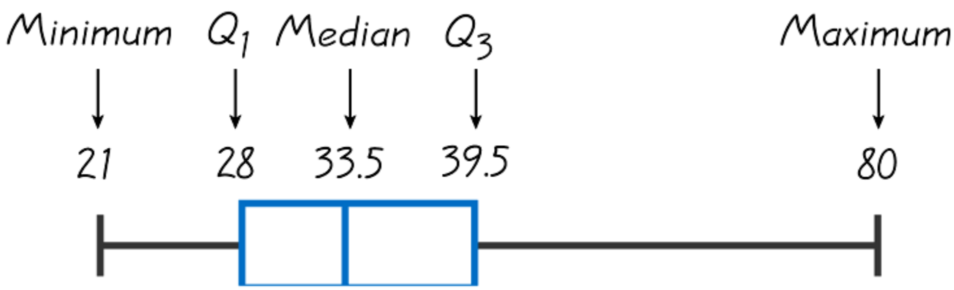
2nd Quartile = median = 50th percentile

3rd Quartile = 75th percentile

***Q_1, Q_2, Q_3* divide ranked scores into four equal parts:**



There is a useful graph that can be created using the quartiles. It is called a Boxplot. One of these is illustrated below:



(The remaining material below in 2.8 is optional)

Guidelines for finding the approximate kth – percentile:

1. Order the data.
2. Calculate $L = \left(\frac{K}{100}\right)n$ (k = percentile in question).
3. If L is an integer, add 0.50. If not, round up.
4. Find the number in the Lth position. If L is a decimal, average the two numbers sitting in the positions it lies between.

Another approach is to use the following formula:

$$L_k = (n+1)\frac{k}{100}$$

*What to do if L_k is a decimal: if the locator ended up being 14.35 you would add 0.35 (the decimal part of the locator) times the difference between the 14th and 15th value to the 14th value (the whole number part of the locator). For example, if the locator was 14.35, the 14th value was 80, and the 15th value was 83, we would perform the following calculation: 80 + (83 - 80)*0.35 = 81.05.*

Guidelines for finding the approximate percentile of a given number:

1. Count the number of values below the number
2. Add 0.5
3. Divide by n
4. Convert to a percent (multiply by 100)

Example: Use the data below to find the age corresponding to the 25th percentile, and the percentile for an actress who is 35 years old.

21	22	24	24	25	25	25	25	26	26
26	26	27	27	27	27	28	28	28	28
29	29	29	29	29	29	30	30	31	31
31	32	32	33	33	33	33	33	34	34
34	35	35	35	35	35	35	35	36	37
37	38	38	38	38	39	39	40	41	41
41	41	41	42	42	43	45	46	49	50
54	60	61	63	74	80				

Solution: 28 is the age corresponding to the 25th percentile, and 35 years old represents the 55th percentile for this data set.