Descriptive Statistics

Exploring Data with Graphs and Numerical Summaries

Data sets contain information about some group of individuals or experimental units.

This information is organized into variables.

- A variable describes a characteristic measured on a subject or experimental unit in the study.

Example 2.1: A nutritionist collected information on 20 popular brands of cereals. Below is her data set.

The variables are: sodium, sugar and CODE.

- Sodium = number of mg in one serving of cereal
- Sugar = number of grams of sugar in one serving of cereal
- CODE = whether the cereal is considered an *Adult cereal* (A) or a *Children's cereal* (C)

CEREAL	SODIUM(mg)	SUGAR(g)	CODE
Frosted Mini Wheats	0	7	A
Apple Bran	260	5	A
Apple Jacks	125	14	C
Capt Crunch	220	12	C
Cheerios	290	1	C
Cinnamon Toast	210	13	C
Corn Flakes	290	2	A
Raisin Bran	210	12	A
Crackling Oat Bran	140	10	A
Crispix	220	3	A
Frosted Flakes	200	11	C
Fruit Loops	125	13	C
Grape Nuts	170	3	A
Honey Nut Cheerios	250	10	C
Honeycomb	180	11	C
Life	150	6	A
Oatmeal Raisin Crisp	170	10	A
Sugar Smacks	70	15	C
Special K	230	3	A
Wheaties	205	3	A

- What are the subjects of this study?

There are two main types of variables

Numerical Variables (quantitative)

- A variable is called **numerical** if each variable value is a number.
    - Numbers measured on a subject are usually numerical variable values.
- In example 2.1, are any of the variables numerical?

Two Sub-types of Numerical variables

- Discrete
    - Numerical variables that can only take certain fixed values, with no intermediate values possible
        - Ex:
- Continuous
    - Continuous variables can take on any real numerical value over an interval
        - Ex:

Categorical (Qualitative) variables: A variable is called categorical if the variable values represent some quality (as opposed to quantity) of the subject. We frequently use categorical values to divide subjects into groups. For example, gender is a categorical variable and we use to sort subjects by gender.

- Subtypes: Nominal and Ordinal

  Nominal variables are purely qualitative and unordered (*i.e.* eye color)

  Ordinal variables can be ranked (*i.e.* stage of cancer). Ordinal variables frequently are given numerical variable values, but these values are substitutes for a characteristic and are not intrinsically numbers. For example when the variable is severity of cancer, a value of 4 means an individual's cancer is at an advanced stage and not easily treatable.

- What variable in example 2.1 is categorical? Is it nominal or ordinal?

Describing a Single Numerical Variable

Example 2.1 revisited: Below is data on 20 popular brands of cereals. We are now interested in summarizing the sugar variable values and understanding the distribution of these values.

The Distribution of numerical data: The distribution of a variable tells us what values the variable takes and how often it takes these values.

CEREAL	SODIUM(mg)	SUGAR(g)	CODE
Frosted Mini Wheats	0	7	A
Apple Bran	260	5	A
Apple Jacks	125	14	C
Capt Crunch	220	12	C
Cheerios	290	1	C
Cinnamon Toast	210	13	C
Corn Flakes	290	2	A
Raisin Bran	210	12	A
Crackling Oat Bran	140	10	A
Crispix	220	3	A
Frosted Flakes	200	11	C
Fruit Loops	125	13	C
Grape Nuts	170	3	A
Honey Nut Cheerios	250	10	C
Honeycomb	180	11	C
Life	150	6	A
Oatmeal Raisin Crisp	170	10	A
Sugar Smacks	70	15	C
Special K	230	3	A
Wheaties	205	3	A

Stem-and-Leaf Displays: (usually for rounded values with at least two digits)

Given data set (0,10,12,23,45,13,54,5,15,34,63,24,64,23), the stem-and-leaf plot is:

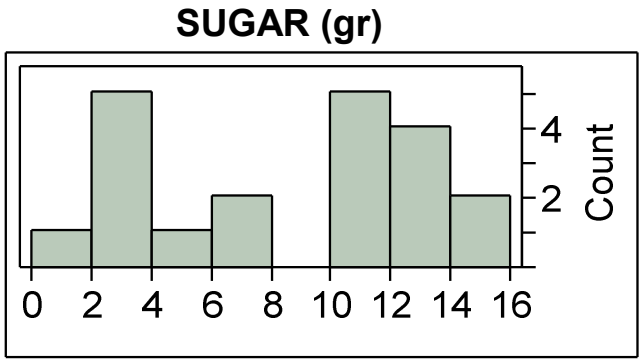
Stem(leading digit(s))	Leaf (trailing digit(s))
0	05
1	0235
2	334
3	4
4	5
5	4
6	34

R code:

```
x<-c(0,10,12,23,45,13,54,5,15,34,63,24,64,23)
stem(x); stem(x,scale=2)
```

Histograms: The most common graph used to describe the distribution of a numerical variable is the histogram.

Reading Histograms:

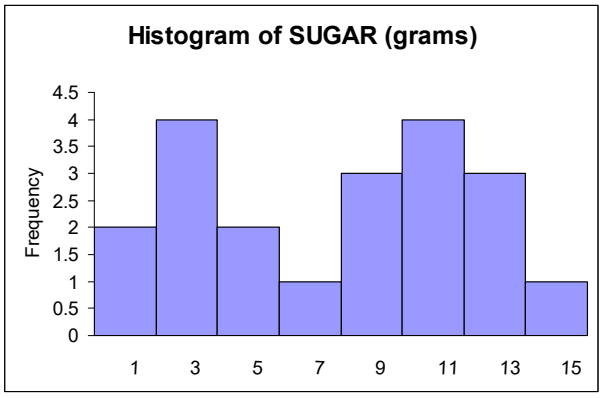


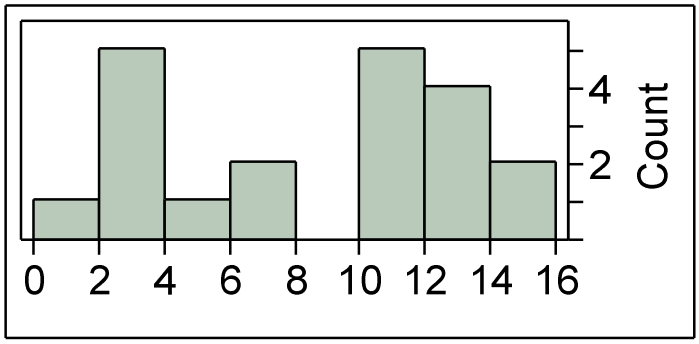
- Variable values are plotted on the horizontal axis. The range of values for the left bar is 0 up to 2 but not including 2. The range of values for the 2nd bar are from 2 up to 4 but not including 4, etc.

  How many cereals take a value of 4 up to 6 (but not 6) grams of sugar?

- The height of each bar is its frequency or the proportion of subjects taking values in that range.

- NOTE: The shape of a histogram *of the same data set* can vary depending on how wide the bars are. Here, the bars are 2 wide but on page 8 a histogram of sugar has bars of width 2.5. The 2 histograms of sugar look different.





Same data set but the range of values for each box is slightly different

- The stem-leaf plot can be considered as a rotated histogram.

- What does this histogram tell us about the distribution of sugar in these 20 cereals?

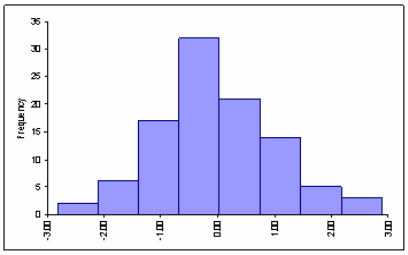
R code:

```
x<-c(7,5,14,12,1,13, 2, 12,10, 3, 11, 13, 3, 10, 11, 6, 10, 15, 3, 3)
hist(x); hist(x,breaks=c(0,4,8,12,16))
```

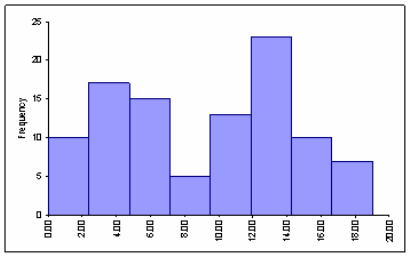
Describing the Shape of the distribution of numerical data

Modality (number of modes)

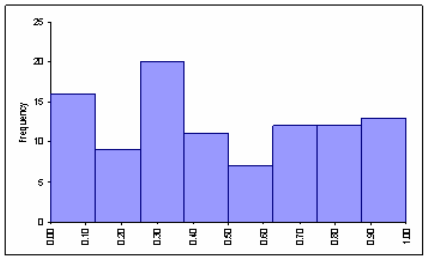
Unimodal



Bimodal

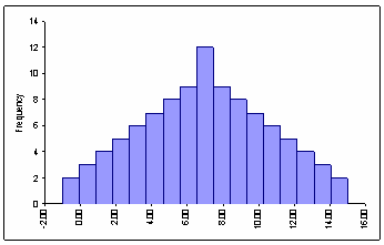


Uniform

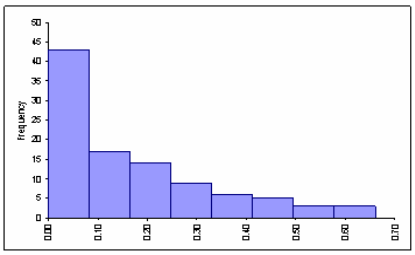


Skewness (of unimodal distributions)

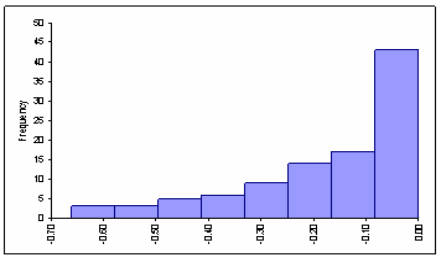
Symmetric



Skewed Right

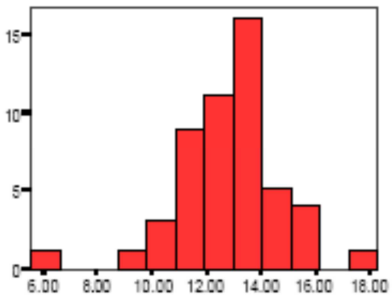


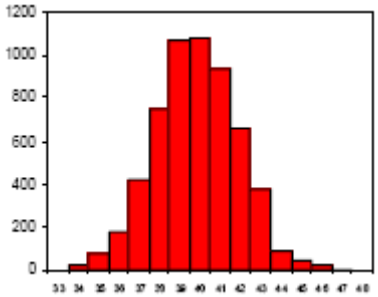
Skewed Left

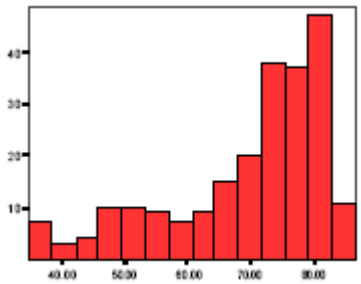


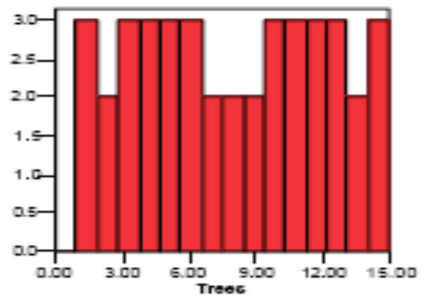
Example 2.3: How would you describe the shape of the distribution of these data sets?

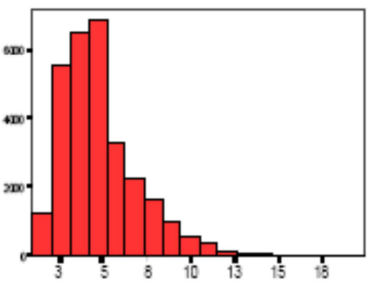
- Identify both the modality and skewness (for data that are unimodal).

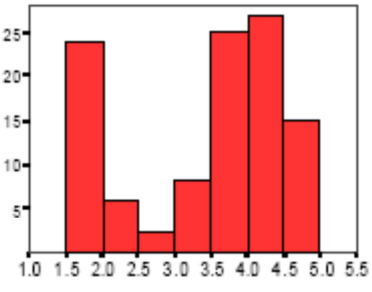












We use four different properties to describe numerical variables:

- Shape
- Center
- Variability
- Unusual observations (data points whose value differs noticeably from the other values)

Three statistics used to measure the center of a data set

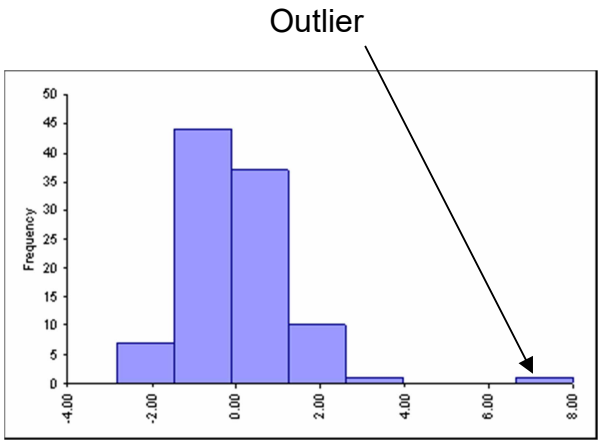
	Mean	Median
Population Parameters	μ	M
Sample Statistics	$\overline{x}$	$\hat{M}$

- **Mode** (least important statistic)
  - The value taken most frequently by the subjects
- **Mean**
  - The "average" value of a variable
  - The sample mean of a data set is defined as
  - The population mean of a census data is defined as
- **Median**
  - The middle value of a data set
  - The sample median of a data set is calculated as
  - The population median is defined as
- **Quartile and Percentile**

Outliers

Outliers are extreme observations taking values far away from the bulk of the data values.

- EX: height of a professional basketball player



Comparing the Mean & Median

In general, if the shape of the distribution is:

- Symmetric
- Skewed Right

- Skewed Left

A statistic is called **robust** or **resistant** if it is not strongly influenced by extreme values like outliers.

- The _____ is robust to outliers.
- The _____ is *not* robust to outliers.
    - The data set 0,1,2,2,3,3,4,4,5,6 has mean 3.0 and median 3.0. If the number 60 is substituted for 6, the median stays the same but the mean increases to 5.7.

A data value is called **influential** if it noticeably shifts the value of the sample mean towards it.

Measures of variability of Numerical Data

- **Variance** The variance is the average of the square of the deviation from the sample mean
    - $x_i - \overline{x}$ is the deviation of the i^{th} data value from the sample mean
    - The variance is not resistant to outliers
- **Standard Deviation** The standard deviation is the square root of the variance.

	Variance	Standard Deviation
Population Parameter	σ^2	σ
Sample Statistic	s^2	s

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

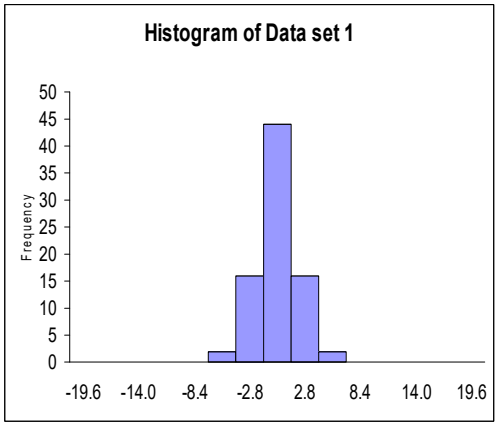
The standard deviation measures how tightly the data values are clustered around the mean
The standard deviation is very close in value to the average distance of the data from $\overline{x}$.

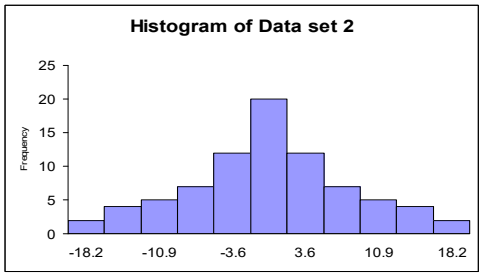
- The standard deviation is NOT resistant to outliers
- The population variance is defined as

- **NOTE:** The denominator in sample variance and population variance are different.
- **NOTE:** Since the standard deviation is just the square root of the variance, they both give us the same information about how much the data values vary from the mean.

The mean and standard deviation are usually paired together. Both use all the data values in their calculation – this is good – but both have the disadvantage that they are not resistant to outliers.

- Below, each data set consists of 80 numbers and the $\overline{x}$ = 0 for both data sets.
- The horizontal (x-axis) scale is the same in both plots.
- In data set 1, the histogram is narrow and tall with a standard deviation *s* = 2.2.
- In data set 2, the histogram is broad and flat with a standard deviation *s* = 9.0.





Why we care about the standard deviation:

If the standard deviation is large, then many data values won't be close to the mean and this may be important:

Car's MPG: When buying a car, the MPG for the make of car is given. But this value is only the average MPG. The actual MPG you'll get on your car could be very different. MPG will of course depend on how are car is driven but it could also depend on factors at the manufacturing plant. What you'd like to know if how much the MGP varies between many cars when the same driver tests many cars. We'll call this variability the manufacturer MPG standard deviation.

So, suppose you purchase a car whose stated MPG = 28. If the manufacture MPG standard deviation in MPG is 4, then the car you purchase could easily only get 24 MPG or less.

Blood pressure: My mother-in-law's average blood pressure is about 130. This isn't bad for an 85 year old woman. But, the SD ≅ 15 and so some days her blood pressure is down around 105 – 110 and she feels crummy. Other days her blood pressure is up to 150 which is dangerous. Her blood pressure can be very high in the morning and then low in the evening.

The standard deviation of her blood pressure matters because it is hard to control her blood pressure with medicine since she needs one drug for high blood pressure but another for low blood pressure. If the standard deviation were small, then she could just take the same drug every day and control her blood pressure.

Alternative measures of the Variability

Range = maximum – minimum

- The range measures how spread out the data values are.
- The range is not resistant to outliers

Interquartile Range (abbr, IQR):

- The IQR is the range of the middle 50% of the data values: **IQR = Q3 – Q1**
- The IQR is resistant to outliers

First quartile (Q1) is the 25th percentile. NOTE: Q1 is not a measure of variability

- Q1 is median of the observations whose position in the ordered list of variable values is to the left of the location of the overall median

Third quartile (Q3) is the 75th percentile. NOTE: Q3 is not a measure of variability

- Q3 is the median of the observations greater than the overall median
- Both Q1 and Q3 are measures of location of numbers but not measures of variability
- You are not expected to calculate Q1 or Q3.

EXAMPLE: Find median, Q1 and Q3 and the IQR of the 20 sodium variable values in the cereal data.

0	70	125	125	140	150	170	170	180	200	205	210	210	220	220	230	250	260	290	290

- The IQR is resistant to outliers – which is good – but gives no information about the spread of data values less than Q1 or greater than Q3
- The median and IQR are usually paired together because both are resistant to outliers.

Five Number Summary & Outliers

Five Number Summary

- Minimum
- Q1
- Median

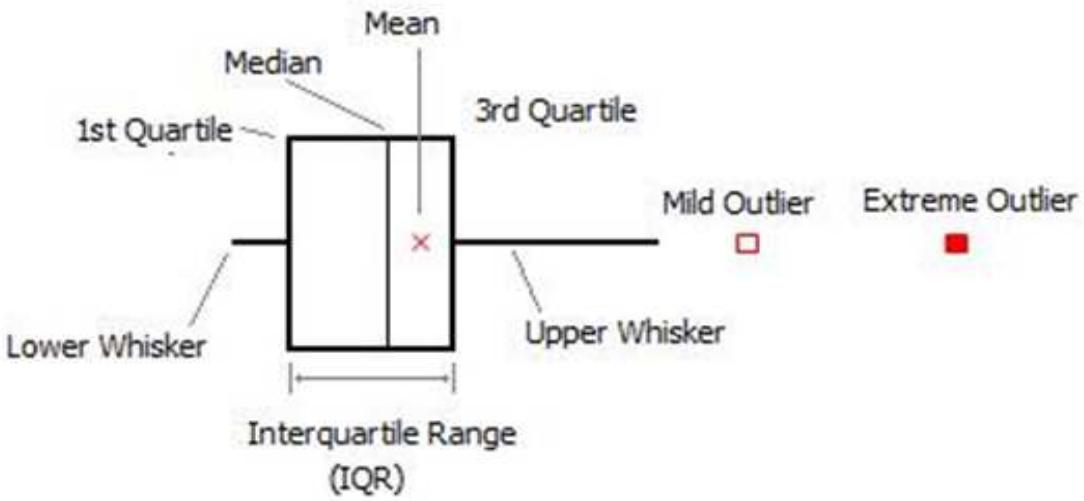
- Q3
- Maximum

Outliers: The formal definition of an outlier is any observation that meets one of the following criteria:

- An outlier is any data value that is less than Q1 - 1.5*IQR or greater than Q3 + 1.5*IQR
- An **extreme outlier** is any data value that is smaller than Q1 – 3*IQR or greater than Q1 + 3*IQR
    - Extreme outliers are frequently influential points whereas an outlier that isn't also an extreme outlier may not be an influential point.

Boxplots: Graphical Versions of the Five Number Summary

Box plots can be used to identify _____ in the data.

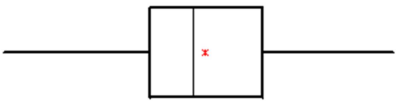


Whiskers extend to the furthest observations that are no more than 1.5 IRQ from the edges of the box. Mild outliers are observations between 1.5 IRQ and 3 IRQ from the edges of the box. Extreme outliers are greater than 3 IRQ from the edges of the box.

You can also use box plots to identify the _____ of a distribution.


Skewed Left


Skewed Right

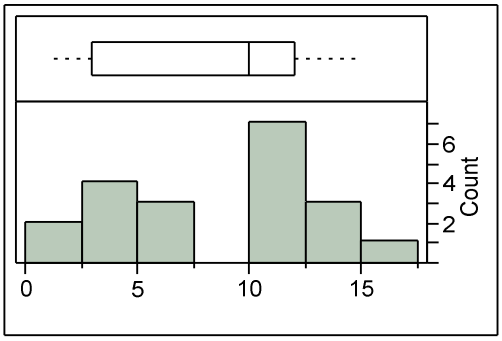

Symmetric

Symmetric: The whiskers will be approximately equal in length.

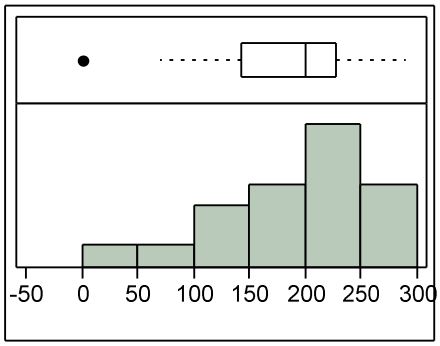
Skewed: One whisker will be much longer than the other.

Comparing box plots and histograms

Cereal data - Sugar (see page 1)



Cereal data – Sodium (see page 1)



What box plots do well:

- Show all outliers
- Give the 5 number summary

What histograms do well:

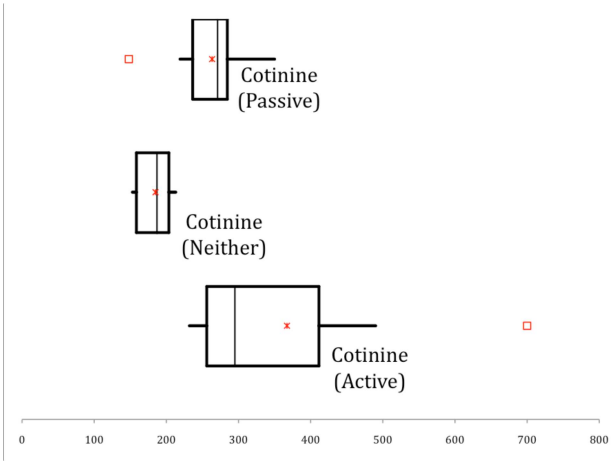
- Indicate the shape of the data's distribution.
- Only with histograms can we tell if a data set has bimodal distribution

Side-by-Side Boxplots for data comparison

Example:

Active smokers	490	418	405	328	700	292	295	272	240	232
Passive smokers	254	219	287	257	271	282	148	273	350	293
Nonsmokers	158	163	153	207	211	159	199	187	200	213

Sources: Sherif, NA et al (2004) Detection of cotinine in neonate meconium as a marker for nicotine exposure in utero. *Eastern Mediterranean Health Journal*, 10, 96-105.



Example for you to look over on your own:

Set 1: 1 1 1 1 1 1 1 1 1 $\bar{x}$ = 1 s = 0 $\hat{M}$ = 1 Q1 = 1 Q3 = 1 IQR = 0

Set 2: 1 2 3 4 5 6 7 8 9 $\bar{x}$ = 5 s = 2.7 $\hat{M}$ = 5 Q1 = 2.5 Q3 = 7.5 IQR = 5

Set 3: -9 -8 -7 -6 -5 -4 -3 -2 -1 $\bar{x}$ = –5 s = 2.7 $\hat{M}$ = –5 Q1= –7.5 Q3= –2.5 IQR = 5

Below is given the calculation for the IQR of data set 3:

- IQR = Q3 – Q1 = -2.5 – (-7.5) = 5. The IQR is positive but all data values are negative!

Set 4: 5 10 15 20 25 30 35 40 45 $\bar{x}$ = 25 s= 13.7 $\hat{M}$ = 25 Q1=12.5 Q3= 37.5 IQR= 25

Set 5: 5 10 15 20 25 30 35 40 150 $\bar{x}$ =37.7 s= 44.0 $\hat{M}$ = 25 Q1=12.5 Q3= 37.5 IQR= 25

- Is the IQR ever negative? (No because IQR = Q3 – Q1 and Q3 is always a larger #.)
- Which of $\bar{x}$, s, $\hat{M}$, Q1, Q3, IQR are measures of variability? (Only s and IQR)
- Are ANY measures of variability negative? (No)

Choosing the best Measure of Center and Variability

- When dealing with strongly skewed distributions, it is somewhat customary to report the median ("midpoint") rather than the mean ("arithmetic average"). However, a health organization or a government agency may need to include all survival times, and thus calculate the mean, to estimate the cost of medical care for a given disease and plan medical staffing appropriately. Relying only on the median would result in underestimating the medical and financial needs. The mean and median measure center in different ways, and both are useful.
- As a rule of thumb, when there are extreme outliers or the data is very skewed, the median is a preferred measure of center because the mean and standard deviation are affected by extreme observations.
- If the data is only slightly skewed and there are no extreme outliers, then the mean is usually the preferred measure of center.

Baseball Salaries:

In 2011, the total payroll for the New York Yankees was $202,689,028.

The average salary was $6,756,300.

The median salary was $2,100,000.

The standard deviation was $8,468,058!

- What does this tell us about the shape of the distribution of NYY salaries?

- Is the average a very good measure of what a “typical” NYY baseball player makes?

R code:

```
x<-c(7,5,14,12,1,13, 2, 12,10, 3, 11, 13, 3, 10, 11, 6, 10, 15, 3, 3)
mean(x);var(x);sd(x);
summary(x);
quantile(x, prob=0.25);
boxplot(x)
boxplot(x, x+1.5)
```

Describing a Single Categorical Variable

State	Frequency	%
Florida	289	79
Hawaii	44	12
California	34	9
Total	367	100

Ex 2.2: An oceanographer studying sharks wanted to know "*Which state has the highest percentage of shark attack?*" The researcher randomly selected 367 shark attacks from a list of all shark attacks that occurred in the US between 2000 and 2005. For each individual shark attack in the sample, the state where the attack occurred was identified. There were 289 attacks in Florida, 44 in Hawaii and 34 in California.

What is a subject in this sample?

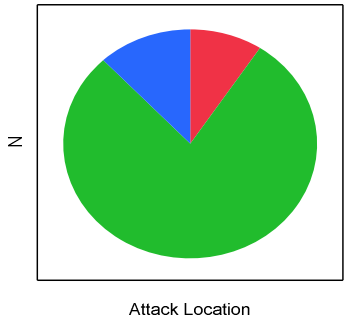
What is the variable? What type of variable is it?

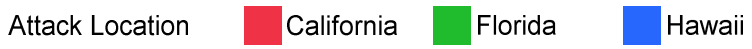
Frequency tables

- Frequency tables list all the variable values along with the number of subjects and the percent of subjects taking each variable value. Usually, the statistic of interest is %.

Pie Charts

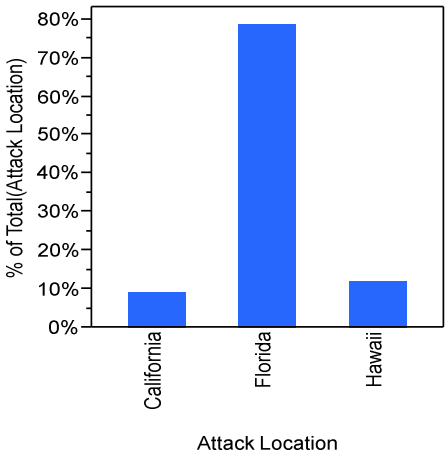
- The size of each slice corresponds to the percentage of observations in that category.





Bar Charts

- For each variable value there is a bar. The height of a bar represents either the percentage of subjects taking that variable value or a count of the number of observations in that category.



The statistics of interest

Counts

- 289 shark attacks occurred in Florida

Percentages

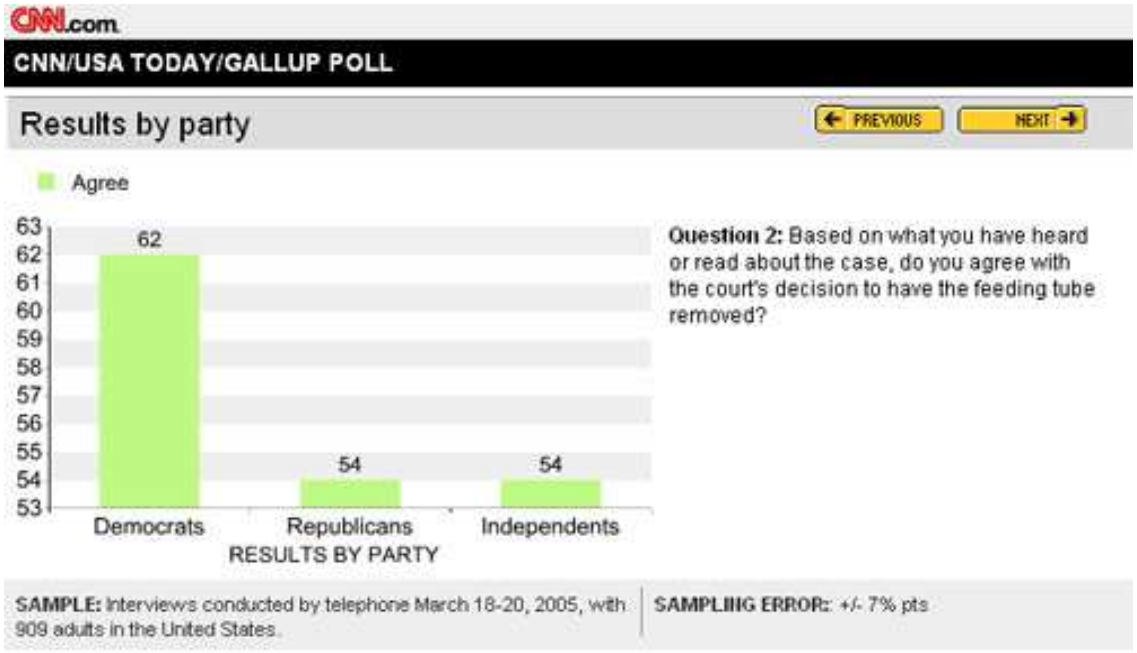
- 9% of all shark attacks in the sample occurred in California

Proportions

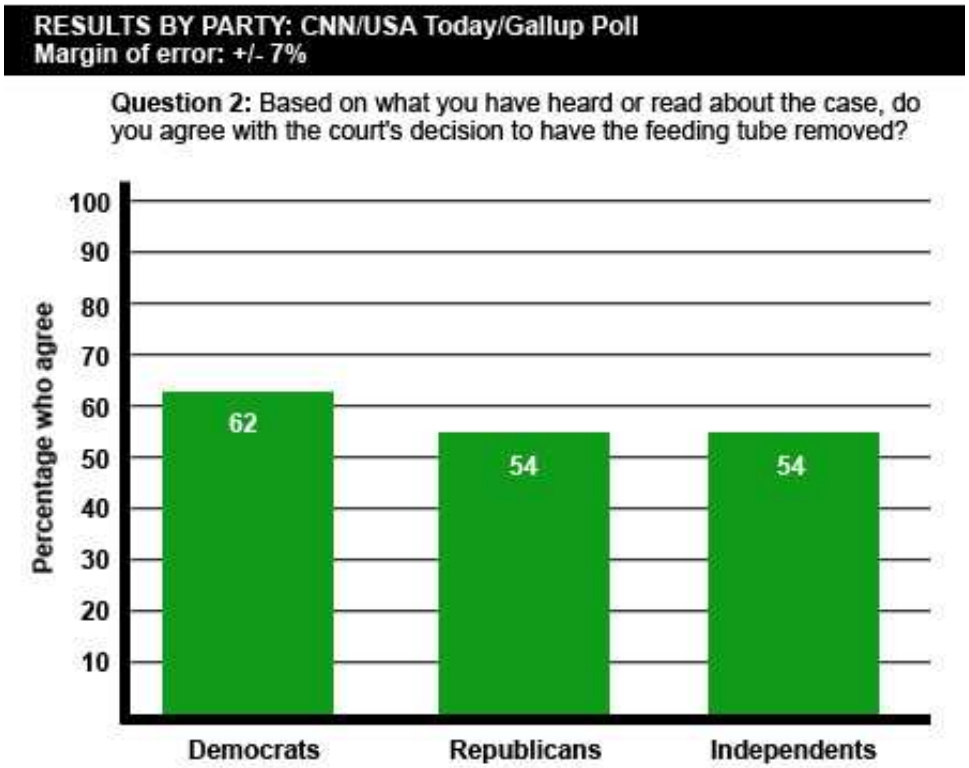
- In the sample, the proportions of shark attacks that occurred in Hawaii is 0.12

Ex 2.3: Example of how charts can be used to give false impressions

CNN.com posted the first graph on results of poll asking "Do you agree with the court's decision to the remove Terry Schiavo's feeding tube?"



A more informative graph from the same data is given below



R code:

```
freq<-c(10,20,30)
barplot(freq, names= c("Democrats","Republicans", "independents"))
pie(freq, labels=c("Democrats","Republicans", "independents"))
```

Summary of the descriptive statistics of one variables

Numerical variables

- Graphical summary
  - Use histograms and box plots

- Description
  - Distribution: the shape and any unusual observations such as outliers.
  - Measures of center: mean $\overline{x}$ and median $\hat{M}$
  - Measures of variability: standard deviation *s* and inter-quartile range IQR

Categorical Variables

- Graphical summary
  - Use frequency tables, pie charts and bar charts

- Statistics
  - Discuss counts, percentages and proportions