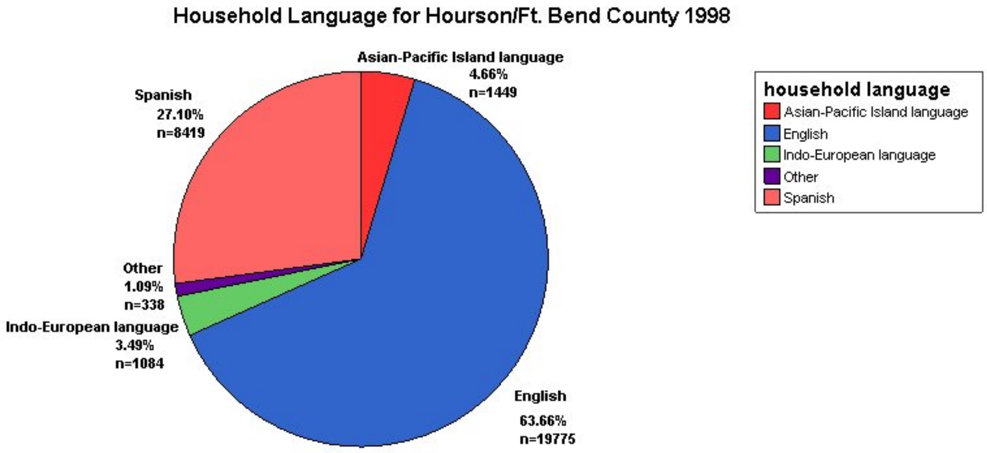
# Lecture 2

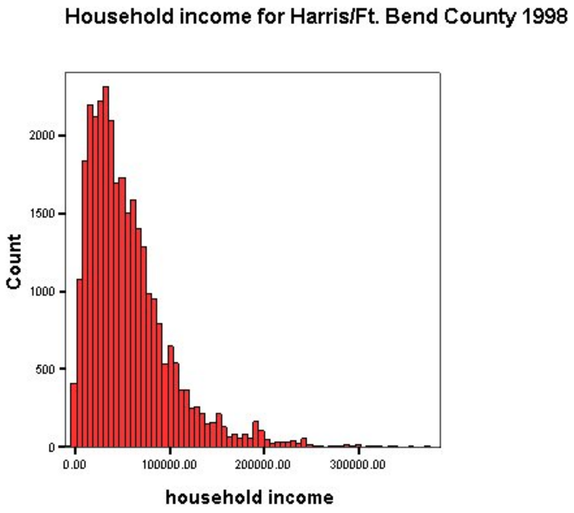
## Last time

### Why are we learning statistics?

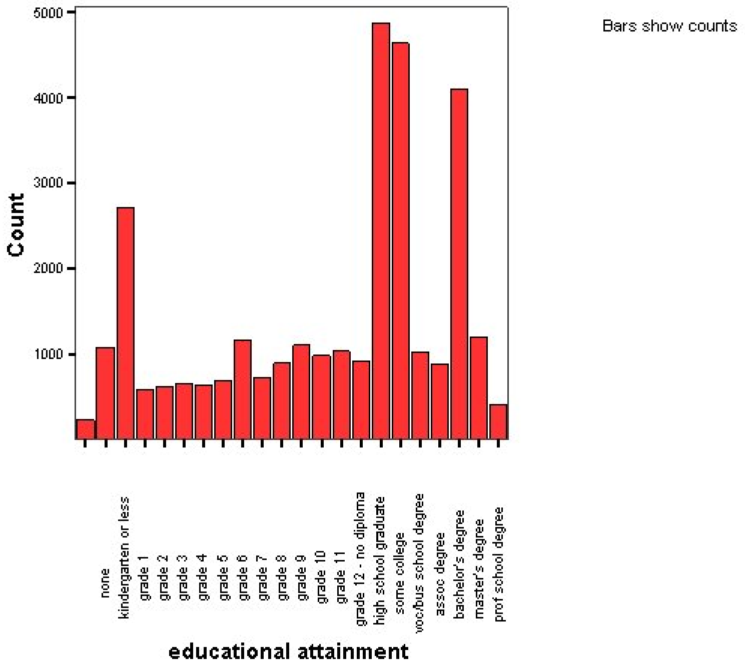
We see statistic everyday on the news, in sports, in papers, all over the place. A new, promising field in statistics is bioinformatics that studies new treatment effectiveness as well as genetic ties to physiological characteristics. What other areas use statistics?

## Displaying data with graphs

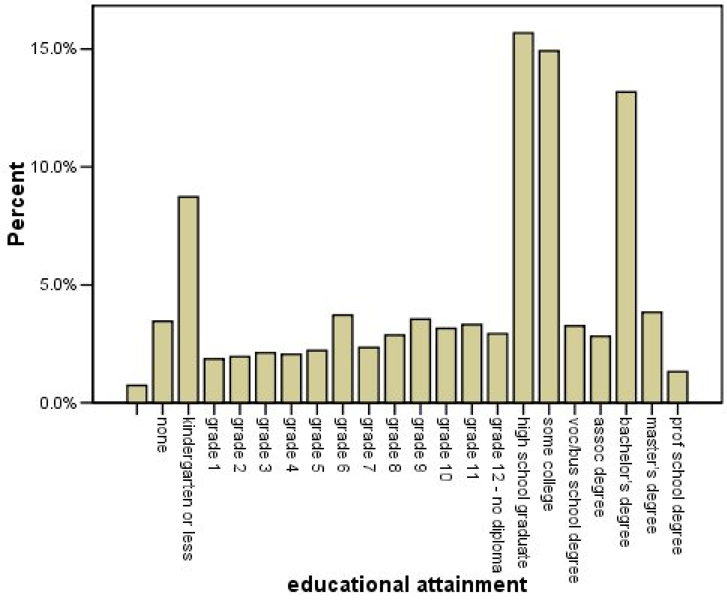




Educational Attainment for Harris/Ft. Bend County 1998



Education for Harris/Ft. Bend County 1998



# 1 Describing Data with Numbers

### Measures of Central Tendency

The **mean** is useful for estimation of the center when the data is distributed symmetrically and is free of outliers. The mean is sometimes called the average. Here we are discussing the sample mean which is denoted by $\bar{X}$. This is different than the population mean denoted by $\mu$, we will discuss this later.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1}$$

NOTE: $X_i$ represents one observation/individual of the variable $X$. $X$ can represent any variable and $X_i$ is one value of $X$ seen for a specific observation/individual. *For example:* Part of the BRFSS data

```
:
   state genhlth physhlth exerany hlthplan smoke100 height weight wtdesire age
1     22      good    0        0        1        0        70    175    175    77
2     25      good    30       0        1        1        64    125    115    33
3     6       good    2        1        1        1        60    105    105    49
4     6       good    0        1        1        0        66    132    124    42
5     39 very good    0        0        1        0        61    150    130    55
6     42 very good    0        1        1        0        64    114    114    55
7     6  very good    0        1        1        0        71    194    185    31
8     48 very good    1        0        1        0        67    170    160    45
9     6       good    2        0        1        1        65    150    130    27
10    48      good    3        1        1        0        70    180    170    44
```

Let $X$ be *weight* and we have information about $X$ for 10 people so $X_1 = 175, X_2 = 125, X_3 = 105, X_4 = 132$,...

NOTE: $n$ denotes your sample size. What's your sample size for the variables above?

*Example:*
Let's calculate the average weight of these first 10 individuals in the BRFSS study:

The **median** is the midpoint of the distribution of observations, where half the observations lie above the median and half the observations lie below the median.

How to find the median:
(1) Order the observations from smallest to largest
(2) If $n$ is odd, the median $M$ is the $(\frac{n+1}{2})$ observation so

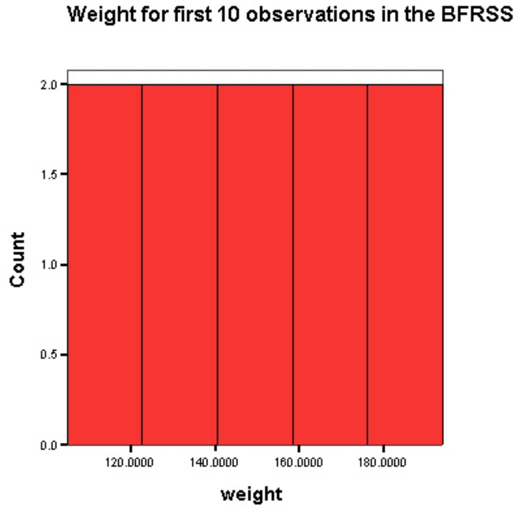
$$M = X_{(\frac{n+1}{2})} \tag{2}$$

(3) If $n$ is even, the median $M$ is the average of the observation $(\frac{n}{2})$ and the observation above $(\frac{n}{2})$ so

$$M = \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2} \tag{3}$$

*Example*:
Now, let's find the median weight of the first 10 observations from the BRFSS.

NOTE: The mean can be easily influenced by extreme observations or outliers while the median is a **resistant measure** because it is not heavily influenced by outliers.



### Measures of Spread

The **standard deviation** is the average distance that each observation is away from the mean. The most common measure of center and spread are the mean and standard deviation. The standard deviation is just the square root of the **variance**. Here we are discussing the sample standard deviation denoted by $s$, when discuss population standard deviation we will denote it by $\sigma$ (sigma)

$$Variance \ = s^2 = \frac{(x_1 - \bar{x})^2 + (x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1} \tag{4}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{5}$$

$$Standard\ Deviation \ = s = \sqrt{s^2} \tag{6}$$

Since the standard deviation is dependent on the mean, would you guess that it is a resistant measure?

*Example*:
Let's find the standard deviation for *weight*.

NOTE: Only use $s$ to measure the spread when you are using the mean to measure the center. If $s$=0, there is no spread. What would this say about every observation?

The **Five-Number Summary** describes the distribution of the observed values of a variable by 5 numbers placed in a specific order: the minimum, first quartile, median, third quartile and the maximum.

The **first quartile**, $Q_1$, lies one-quarter of the way up the list of ordered-values (arranged increasingly).

The **third quartile**, $Q_3$, lies three-quarter of the way up the list of ordered-values (arranged increasingly).

How to find quartiles:
(1) Arrange observations in increasing order
(2) Find the median
(3) The first quartile is the median of the observations to the left of $M$
(4) The third quartile is the median of the observations to the right of $M$

Other important measures of spread are the **range** and **interquartile range (IQR)** defined as:

$$Range \;=\; Max \;-\; Min \tag{7}$$

$$IQR \;=\; Q_3 \;-\; Q_1 \tag{8}$$

*Example:*

Are you ready to find the five-number summary, range and IQR for the variable *weight*?! YEAH!

We talked about outliers being extreme observations that effect measurements like the mean and standard deviation but how do we know if an observation is an outlier or not. These are some basic rules to follow but it is still necessary to research each data set individually in order to truly know if an observation is extreme.

An observation $x$ is a **suspected outlier** if:

$$x < Q_1 - 1.5IQR \; or \; x > Q_3 + 1.5IQR \tag{9}$$

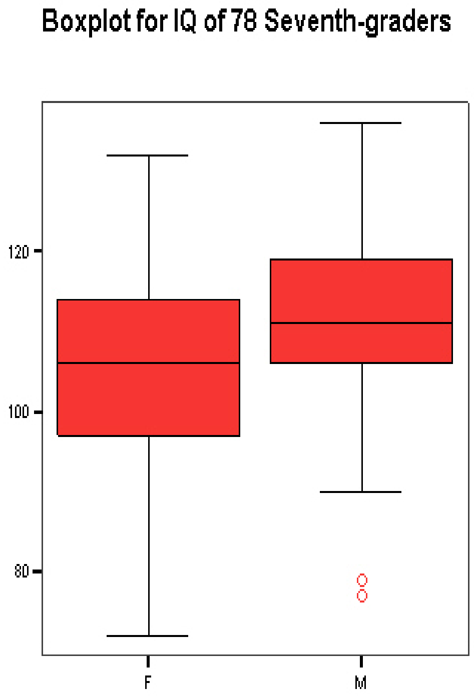
An observation $x$ is a **highly suspected (extreme) outlier** if:

$$x < Q_1 - 3IQR \; or \; x > Q_3 + 3IQR \tag{10}$$

Are there any suspected or extreme outliers in the 10 observations of *weight*?

### Graphical Displays of Numerical Summaries

The five-number summary and potential outliers can be shown on a graph called a **boxplot**. Boxplots are good ways to compare observations of a variable from different groups.


Boxplot for IQ of 78 Seventh-graders
F
M