# Lecture 3

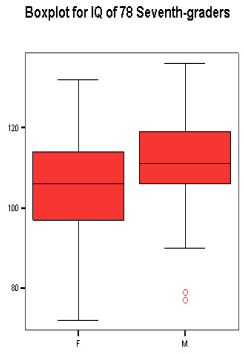
## Last time

Measure of center: mean and median

Which one is a resistant measure?

Standard deviation measures the observations average distance away from the mean.

What information does this box plot give us?



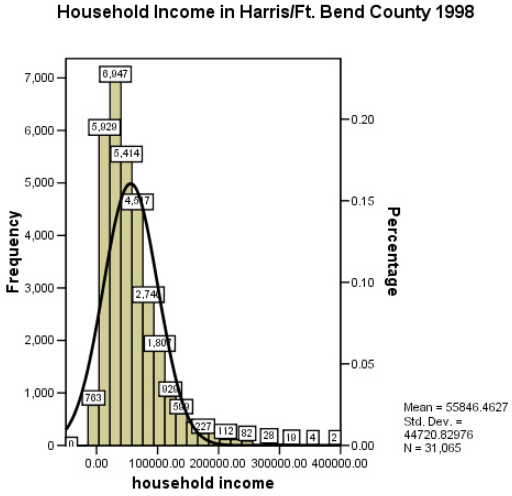

# 1 Density Curves

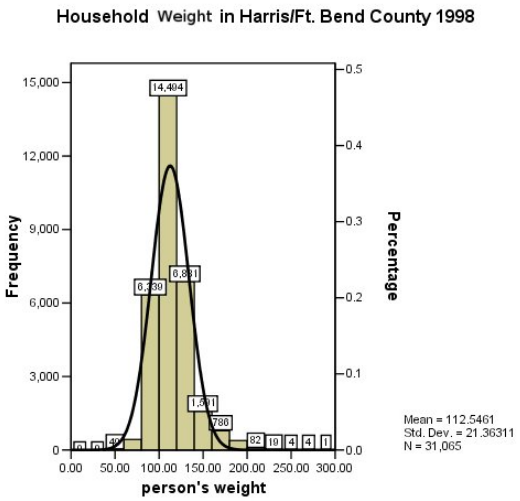
We have already gone through three steps to analyze data; now we will add a fourth:
(1) Plot the data - in a histogram or stemplot
(2) Look at the shape, center, spread and outliers
(3) Calculate the center and spread numerically
(4) Summarize a regular overall pattern consisting of many observations in one smooth curve

Given a histogram of values, a smooth curve called a **density curve**, can be drawn through the tops of the bars, making a good approximate description of the distribution of the data. The density curve is a mathematical model of the data which ignores minor irregularities and outliers but gives a compact picture of the overall pattern. They are easier to work with than a histogram or stemplot.

## Properties of a Density Curve

- It is always at or above the horizontal axis (ie, it never has negative values)
- It has an area of 1 (or 100%) underneath the entire curve




The density curve gives us a good estimation of the frequency or percentage of people at a certain income or weight. We trust that this is a random sample of 31,065 people that accurately represent the entire population of Harris and Ft. Bend County in 1998.

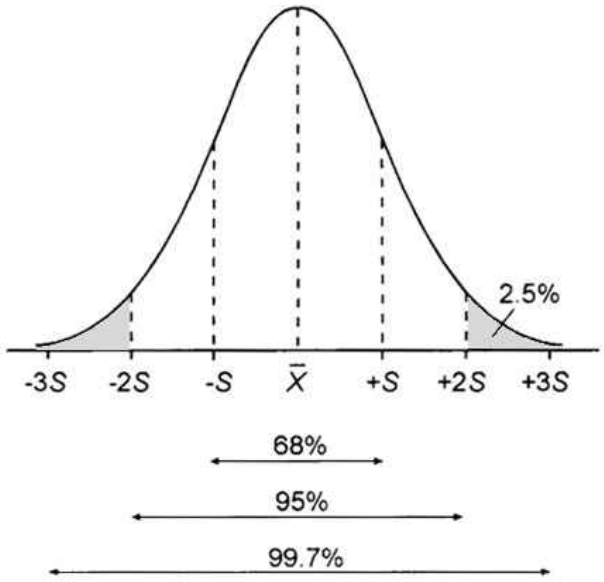
Note that these histograms and graphs represent probabilities.

*Example:*

Calculating the value with the histogram, we see that 26,316 people out of the 31,065 sample have incomes less than \$100,000. What percent of people is this?

The area under the smoothed density curve to the left of \$100,000 also represents an approximation of the percentage of people with incomes less than \$100,000. We will learn how to make this approximation, but for now if it is enough to know that it is 85%.

We can use the **empirical rule** on distributions that are symmetric and mound-shaped:
68% of the distribution lies within one standard deviation of the mean
95% of the distribution lies within two standard deviations of the mean
99.7% of the distribution lies within three standard deviations of the mean




*Example:*
Can we use the empirical rule on the *weight* distribution for people in Harris and Ft. Bend County?

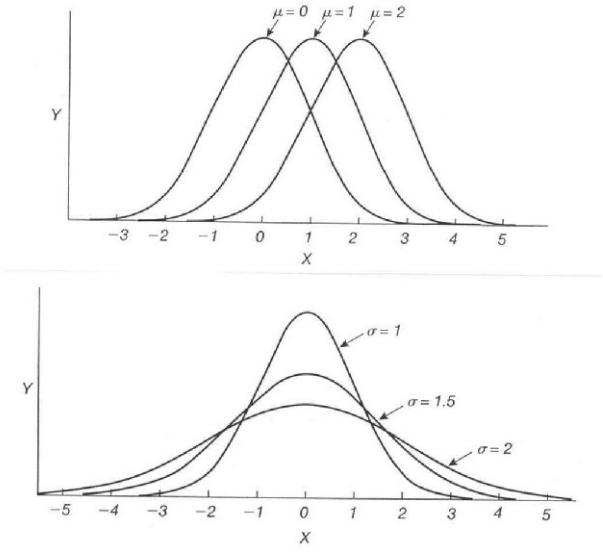
NOTE: Remember that $\bar{X}$ and $s$ represent the mean and standard deviation calculated directly from the (sample) data. We collect data from a sample in order to make inferences about the entire population. When we talk about the population we use $\mu$ (muoo) for the population mean and $\sigma$ (sigma) for the population standard deviation.

# 2 The Normal Distribution

There are several classes of density curves (or distributions) that have the same general shape. One of the most commonly used is the **normal distribution**.

## Important properties of the normal distribution

(1) They are all symmetric, uni-modal, and bell-shaped
(2) They are centered at their mean $\mu$
(3) Their spread is determined by $\sigma$
(4) Their shape is described completely by $\mu$ and $\sigma$




We will abbreviate a variable $X$ that has a normal distribution with mean $\mu$ and standard deviation $\sigma$ as:

$$X \sim N(\mu, \sigma) \tag{1}$$

## Standard normal distribution

Any variable, call it $X$, can be standardized by subtracting its mean ($\mu$) and dividing by its standard deviation $\sigma$. The resulting value is commonly called a **Z-score**.

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

The value of $Z$ tells us how many standard deviations the original observation ($X$) falls away from the mean, and in which direction (bigger or smaller)

*Example:*

The lengths of an adult South American rain forest beetle species are distributed normally with $\mu$=5.6 cm and $\sigma$=.32 cm.

$Z = \frac{length - 5.6}{.32}$

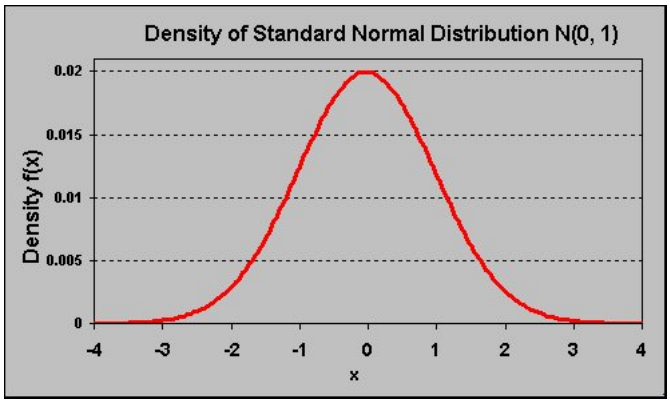
What is the Z-score for a beetle of length 5.1 cm.?

$Z = \frac{5.1 - 5.6}{.32} = -1.6$

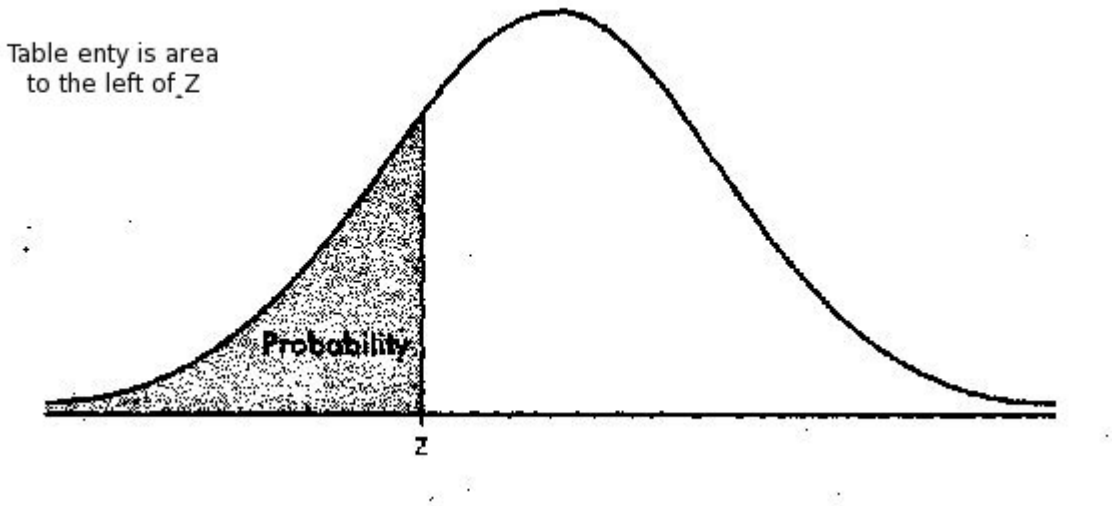
Therefore, this beetle's length is 1.6 standard deviations smaller than the mean.

If the variable that is standardized has a normal distribution, then standardizing it produces the **standard normal distribution**

- It is denoted by N(0,1) because a standardized variable has mean = 0 and standard deviation = 1
- In more concise notation, if $X \sim N(\mu, \sigma)$ then $Z \sim N(0, 1)$



We use the area under distribution curves to find the percentage of a population that falls within a certain range of values. Because we can transform any normal distribution into a N(0,1) distribution, we only need to use one table of values to find the area under a N(0,1) curve, it's Table A in the appendix and also inside the front cover of the textbook. The table entry for each value $Z$ is the area under the curve and to the left of $Z$ as shown below




*Example*:

Using the Weight data from Harris/Ft Bend County we will assume that our population mean $= \mu = 112.55$ and our population standard deviation $= \sigma = 21.36$.

What percentage of people weighed less than 100 pounds?

$Z = \frac{100-112.55}{21.36} = -0.59$

Look up -.59 on Table A and you get .2776. Therefore, 27.76% of people weighed less than 100 pounds.

What percent of people weighed more than 120 pounds?

$Z = \frac{120-112.55}{21.36} = 0.35$

The table value for 0.35 is .6368. Is this the answer we are looking for?

What percentage of people weighed between 100 and 150 pounds?

Let $X$ be weight then we are looking for $100 \leq X \leq 150$.

$\frac{100-112.55}{21.36} \leq \frac{X-112.55}{21.36} \leq \frac{150-112.55}{21.36}$

$-.59 \leq Z \leq 1.75$

Area between -.59 and 1.75 = area left of 1.75 - area left of -.59 = .9599-.2776 = .6823 Therefor, 68.23% of the people weighed between 100 and 150 pounds.

**Finding a value given a proportion (backwards standardization)**

You may want to know an observed value for a given proportion. To do this, we find the Z-score for the proportion and then "unstandardize" it by multiplying it by the standard deviation and then adding the mean to that value.

*Example:*
How much did the top 25% of people weigh in Harris/Ft Bend county in 1998?
Find the closest tabled value to .75 and then the Z-score associated with it.

Solve the equation $\frac{X-112.55}{21.36} = .67$ for $X$

In general:

$$X = \mu + \sigma(Z) \tag{3}$$