This project uses data collected by the National Government as part of its "National Health and Nutrition Examination Survey (NHANES)." There have been three such large surveys and have proven extremely useful. Elizabeth Kolbert in the *New Yorker* writes this about the NHANES project

> One of the most comprehensive data sets available about Americans – how tall they are, when they last visited a dentist, what sort of cereal they eat for breakfast, whether they have to pee during the night, and, if so, how often – comes from a series of studies conducted by the federal Centers for Disease Control and Prevention. Participants are chosen at random, interviewed at length, and subjected to a battery of tests in special trailers that the C.D.C. hauls around the country. The studies, known as the National Health and Nutrition Examination Surveys, began during the Eisenhower Administration and have been carried out periodically ever since.

Her article is on how one scientist's look back in time through the NHANES data showed Americans getting more and more overweight as a group. Here we look at some of the data on college age kids.

Load the data

First we need to load the data. Not to hard, simply enter this command:

As in life, the details matter – those are upper and lower case letters.

```
> source("http://www.math.csi.cuny.edu/verzani/classes/MTH214/R/college.txt")
```

Afterwards you should have a big data set sitting on your computer stored as `df`. To see, you can describe the data with:

The `str` function succinctly describe each variable. Here we have factors (categorical variables) and numeric variables. The numbers by the factors are how the values get stored internally.

```
> str(df)

'data.frame':        1755 obs. of  11 variables:
 $ race      : Factor w/ 4 levels "white","black",..: 1 1 2 2 1 2 1 2 1 2 ...
```

```
 $ gender    : Factor w/ 2 levels "male","female": 1 1 2 2 2 1 2 1 1 1 ...
 $ age       : num  21 19 19 21 18 22 22 22 21 19 ...
 $ ageunit   : Factor w/ 2 levels "months","years": 1 1 1 1 1 1 1 1 1 1 ...
 $ ageMonths : num  261 234 235 254 220 275 271 271 258 236 ...
 $ familysize: num  4 1 3 3 4 6 3 6 8 5 ...
 $ metro     : Factor w/ 2 levels "metro","non-metro": 1 1 1 2 2 2 2 1 2 1 ...
 $ region    : Factor w/ 4 levels "NE","MidWest",..: 4 2 2 3 1 3 3 4 4 2 ...
 $ weight    : num  180 165 163 103 100 ...
 $ BMI       : num  25.5 24.5 24.1 17.9 18.5 23.6 26.2 19.1 24.9 26 ...
 $ height    : num  70.4 68.9 69 63.7 61.9 71.7 62.2 68.4 70.6 66.3 ...
```

Our goal here is to explore the data set graphically – there are 1761 cases – using our graphics: dotplots, stem-and-leaf plots, histograms, densityplots and quantile-normal plots.

We will use R's `ggplot2` package to do so, as the graphs are somewhat prettier than the standard ones.

To load this package, you issue the command:

```
> library(ggplot2)
```

(If that doesn't work, you can try downloading and installing the package with the command:

```
> install.packages("ggplot2")
```

You will be asked to select a spot to download the package from. Pick a close by one.)

Heights

We mentioned in class that heights are generally thought to be normally distributed. Do we see that in the data?

A histogram of heights can be made with the command `qplot`:

```
> qplot(df$height)
```

The use of `df$height` is to get at the `height` variable stored in the data frame `df`. The `qplot` function actually has a convenient alternate form which we use, as it proves helpful when we consider more than one variable:

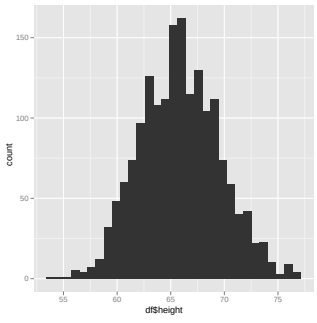

Figure 1: Histogram: qplot default

```
> qplot(height, data = df)
```

Question 0.1. Describe the shape of the distribution of heights.

Question 0.2. Describe the center and spread of the distribution of heights

Question 0.3. Use the functions `mean` and `sd` (to find s) to compute the center and spread.

Question 0.4. Use the functions `median` and `IQR` to compute the center and spread.

Question 0.5. Explain why you might expect the mean and median to be similar.

Question 0.6. Did you expect the standard deviation and the IQR to be identical?

Question 0.7. The percentiles look at the data by answering which value has a given percent of the data (or chance) less than it. To answer the question: for a given value, what percent of the sample is less can be done easily using the following idiom. If we want to know what percent of all heights are 5' 5" or shorter we need to find the ratio of those shorter than the amount and the total number. Here are the commands:

Percentiles are identical to quantiles, the 25th percentile being the 0.25 quantile.

```
> sum(df$height <= 5 * 12 + 5)/length(df$height)

[1] 0.4131054
```

What percent are less than 6 feet tall? More than 5 feet tall?

Faceting

The `qplot` function has some neat features. For example, you might think that the data is perhaps multimodal – it includes both genders and we know that on average females are shorter. To split data up by genders is referred to as "faceting" and is done with this command:

Make note of the extra "tilde," as you will need to include this when you use this command.

```
> qplot(height, facets = ~gender, data = df)
```

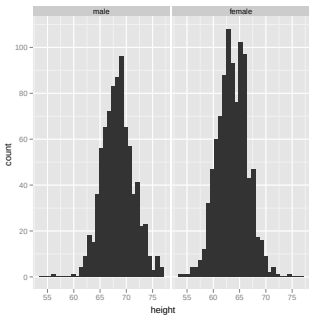

Figure 2: Histograms split up by gender. This is done with the `facets` argument. Notice the "tilde." Make note of the double equals sign (to test equality) and the use of `"male"` which is the code for the males. For testing, one has several types: <, <=, == >= and >.

Question 0.8. Produce the faceted diagram, then estimate the median value for males and for females.

Question 0.9. We haven't seen how to split up a variable by a factor, or categorical variable. There are many ways. One easy way is to use the `subset` command to extract values. For instance, we have

```
> justMales <- subset(df, subset = gender == "male")
> mean(justMales$height)
```

```
[1] 68.41786
```

Repeat for the females.

Question 0.10. The centers of the two distribution are clearly different. Are the spreads? The shapes?

Densityplots

A density plot is not too hard to make using `qplot`. It just requires our asking. How this is done is by adding a request to use the "density statistic":

```
> qplot(height, stat = "density", data = df)
```

Question 0.11. Make the density plot by entering that command. Do you see evidence of bimodality?

Question 0.12. The Census bureau uses a breakdown of 4 major regions: West, Midwest, South and NE. Make four comparative densityplots by faceting with the region variable. Is there a noticeable difference in the distribution of heights from region to region?

Boxplots

The boxplot shows center, spread and shape with a graphic that lends itself to side-by-side comparisons. Single boxplots are not as useful, as say a histogram to see a shape, so for `qplot` one specifies two variables – a categorical one to split the data up with and numeric one to make boxplots of. To illustrate, we have

```
> qplot(gender, height, geom = "boxplot", data = df)
```

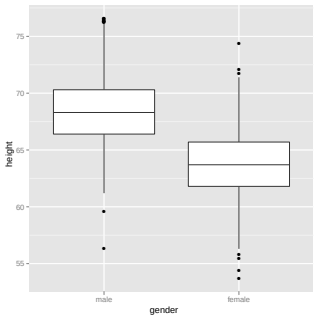

Figure 3: Boxplots of height by gender

The use of `geom` here is short for use the geometry of a boxplot to illustrate the data. Sometimes this is `geom` – to describe a display, and sometimes `stat` to describe how to summarize the data.

Quantile plots

Producing quantile plots is not too much different, We use the "qq statistic." However, we also need to inidate that our data is the `sample`. (One can use such graphs to compare two samples or other theoretical distribution besides the normal.) Here is how to get a plot for all the height data:

```
> qplot(sample = height, stat = "qq", data = df)
```

To use `gender` to make different graphics for males and females can be done by using the `colour` argument specifying which factor to color by:

```
> qplot(sample = height, stat = "qq", colour = gender, data = df)
```

Question 0.13. Based on the quantile plot for both genders, do both seem to be normally distributed?

Weights

Although height data is mostly normal, or can be well approximated by the normal anyways, one wouldn't expect weight data to show this, as it is much easier to be 50 pounds overweight than 50 pounds underweight, say. That is, we don't expect symmetry – a key feature of the shape of normally distributed data.

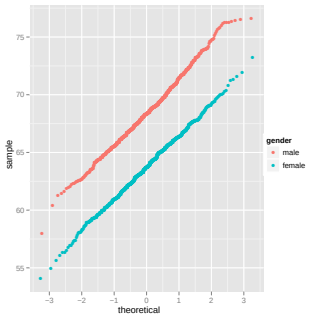

Figure 4: Quantile plot of all heights with color to group males and females. Normal data falls along a more or less straight line.

Question 0.14. Make a histogram and estimate the mean weight for the entire sample. Confirm with the `mean` function. Write down both values and comment if you are way off.

Question 0.15. From your histogram, estimate the median, then confirm with the `median` function. Again, write down both and comment if you are way off.

Question 0.16. Did you guess that the mean and median are similar or different? Comment why and whether you were right. In your comments did the shape of the distribution come into play?

Question 0.17. Find estimates for the mean and median for just the males and just the females by making densityplots of each.

Question 0.18. Make qqplots of weight coloring by gender. Do you see the graphic producing a "straight line?"

You will never get a perfectly straight line – this graph is produced from a sample and sampling introduces variability.

BMI

Dating back to Quetelet (over 200 years ago) people knew that looking at weight alone can be misleading, as taller people are going to be heavier. One should take a ratio of the two (in some manner) to offset this. Quetelet suggested a power relationship, which has become the BMI – a ratio of weight to height squared.

This is a discussion for chapter 2, which we will get back to later in the term.

Weight is proportional to volume (multiply by the density). The volume of a cube of side x is x^3, the area of square of side x is x^2. So the right ratio of volume to dimension is a power of 3 and the right ratio of area to dimension is a power of 2. The BMI uses a power of 2.

Question 0.19. Make a histogram of the BMI variable. Describe the center, spread and shape.

Question 0.20. Make a plot comparing the distribution of BMI for males and females. Does there appear to be a difference due to gender? If not, explain. If so, comment.

Question 0.21. A BMI over 25 is considered overweight and a bmi of over 30 obese. these seem a little strict to me, personally, but the cdc has these guidelines. what percent of the sample has as bmi more than 25? more than 30

Question 0.22. is bmi stable over all age ranges? it definitely isn't for ages 1 through 18, but may be for the ages 18 - 22 represented in this data. to check, we can produce boxplots for each age as follows, specifying that `age` is a factor (a categorical variable):

```
> qplot(factor(age), BMI, data = df, geom = "boxplot")
```

Make the graphic and discuss if there seems to be any change.

Answers:

0.1 Basically unimodal and symmetric

0.2 Roughly it is 65 with a apread aroudn 5.

0.3

```
> mean(df$height)

[1] 65.87903

> sd(df$height)

[1] 3.706184
```

0.4

```
> median(df$height)

[1] 65.8

> sd(df$height)

[1] 3.706184
```

0.5 The mean and median are identical for perfectly symmetric data.

0.6 They both measure spread, but do so differently. For the normal distribution we actually can compare, as the value of $\sigma = 1$ (σ is the equivalent of s) but we say $Q_3 = -Q_1 = 0.67$, so we get IQR = 1.35 (after rounding). Checking, we get

```
> c(1.35 * sd(df$height), IQR(df$height))

[1] 5.003348 5.200000
```

are similar.

0.7

```
> sum(df$height <= 6 * 12)/length(df$height)
```

```
[1] 0.9470085
```

```
> 1 - sum(df$height <= 5 * 12)/length(df$height)
```

```
[1] 0.9470085
```

0.8 Roughly 68 and 64

0.9

```
> justFemales <- subset(df, subset = gender == "female")
> mean(justFemales$height)
```

```
[1] 63.6929
```

0.10 The shapes are both basically unimodal and symmetric, so by that description they are the same – of course, they are different too!

The spreads look similar, perhaps a bit more for males. To get a numeric difference, we have:

```
> c(males = sd(justMales$height), females = sd(justFemales$height))
```

```
   males  females
2.941076 2.790807
```

0.11 No.

0.12 The y axis is a count – this hasn't been scaled to have area 1 – so they look different, but aren't so much. Although perhaps NE is more spread out.

0.13 Yes. Although the very tails for the males show some possible difference.

0.14

```
> plot(weight, data = df)
> mean(df$weight)
```

```
[1] 151.5766
```

0.15 Right skewed data has a median less than the mean, and here we guess that it will be the case. We guess between 140 and 150. The actual value is

```
> median(df$weight)

[1] 144.7
```

0.16 Should guess that right skewed data has mean greater than the median.

0.17

```
> qplot(weight, facets = ~gender, stat = "density", data = df)
```

A guess from these is around 160 for males, and 140 for females.

0.18 To produce the graphic, one can do:

```
> qplot(sample = weight, colour = gender, stat = "qq", data = df)
```

From this we see a pronounced upward curve. This data is non normal. (To read the graph, on the left a curve up is a short tail a curve down a long tail. On the right, a curve up is a long tail, a curve down a short tail – when the "theoretical" is on the x axis.)

0.19 The data is unimodal and skewed right. The median is around 23 and the IQR around 28 - 20. To check we have

```
> summary(df$BMI)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.40   20.90   23.30   24.55   26.90   50.40
```

0.20 A boxplot is a good device for this – both data sets are unimodal. This is produced with

```
> qplot(gender, BMI, geom = "boxplot", data = df)
```

Although we see similar shapes (skewed right) and similar centers (around 23), there is perhaps more spread for females. The ratio weight and height is supposed to remove the dependency on height, so this difference comes from something else.

0.21

```
> sum(df$bmi > 25)/length(df$bmi)

[1] NaN

> sum(df$bmi > 30)/length(df$bmi)

[1] NaN
```

0.22 The shape and spread is similar, but the median seems to be creeping upward.