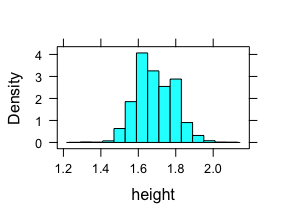
What’s your status -- Single?

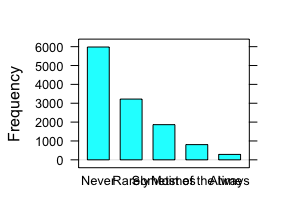
Unit 1 - Lab 2

Directions: Follow along with the slides and answer the questions in **red** font in your journal.

# Where'd we leave off ...

* The first lab ended with these two plots:

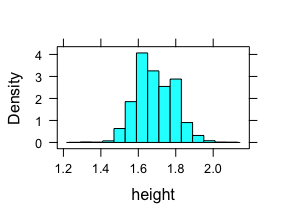


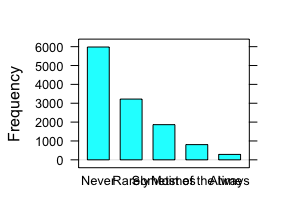


* Look at the type of values on the x-axis?

# Where'd we leave off ...

* This plot has an x-axis with numbers



* This plot has an x-axis with categories 

# Interesting...

* We know then that data comes in *at least* two different varieties.
* **Numerical variables** have values that are typically measured in units
  + Ex: Feet, inches, pounds, hours, sodas per day, etc.
* **Categorical vairables** have values that describe what category the observation belongs to.
  + Ex: Heads or tails, red, green or blue.

# Before we start!

* Load the CDC data from before by using the following command in the console.
  + Remember how to load this data set for future labs.

load("~/Dropbox/IDS Curriculum Design/Labs/Unit 1 Labs/Data/cdc.rda")

# Categorical Variables

* Have *values* that describe the *category* an observation belongs to
* For example, view your data by typing View(cdc) into the console.
  + Find the **gender** of the first person in the data.
  + Which *category* of **gender** does the person belong to?

# Categorical Variables

* Now type the following command into the console to view the names of the variables.

names(cdc)

* **Write down 3 variables that you think are *categorical* variables and why you think that they're categories**
* **View your data and write down the different *values* (or categories) for each of the 3 variables you chose**

# Bargraphs, a.k.a. Barplots

* **Bargraphs** are one of the best ways to *visualize* categorical variables.
  + One axis (x or y) will have the different categories.
  + The other axis will have the number of observations (or *Frequency*) that fall into each category.
* Type the following into the console to create a bargraph (Hit the *Zoom* button in the plot pane to make it larger).

bargraph(~helmet, data = cdc)

* **Explain what the values on the x and y axis mean. Which categories occured the most & the least often**

# More on Bargraphs

* Bargraphs are sometimes easier to read when the bars are horizontal.
  + Run the following command (Make sure to spell *TRUE* in all capital letters):

bargraph(~helmet, data = cdc, horizontal = TRUE)

* **In your opinion, are the vertical bars or horizontal bars easier to read for visualizing the *helmet* data. Why do you think that?**

# More on Bargraphs

* We also sometimes want to *split* each bar in our plot by **grouping** them into seperate categories.
  + Run the following command to **group** the bars for each category based on each person's **gender**

bargraph(~helmet, data = cdc, groups = gender, horizontal = TRUE)

* **Recreate this exact bargraph but make the bars horizontal**

# Your turn

* **Practice making AND interpretting bargraphs using the categorical variables you chose earlier in the lab**
* **Interpret each graph by explaining**
  + **What the categories are**
  + **Which categories occured more or less often**
* **Be sure to point out any interesting discoveries you make**

# Numerical Variables

* Have *values* that are **measured** in **units**.
* For example, view your data by typing View(cdc) into the console.
  + Find the **height** of the first person in the data.
  + How tall is this person? What do you think are the units?

# Watch out!

* Sometimes variables that you think would be *numerical* are actually *categorical*.
  + You might think that **age** is *numerical* because it can be measured in *years*.
* View the values of people's ages in the CDC data by running View(cdc) again.
  + Since the values contain the words "years old", R inteprets these values to be categories!

# Numerical Variables

* List the *names* of the variables again (You can go back to look up the command if you've forgotten)
  + **Write down 3 variables that you think are *numerical* variables**
  + **List some possible units of measurements for each variable.**
* View your data
  + **Write down any variables you thought were numeric but were actually categorical (See the warning on the previous slide for help)**

# Histograms!

* Just like how we used **bargraphs** to visualize *categorical variables*, histograms are useful for visualizing *numerical variables*
* Type the following to make a histogram for people's **height**

histogram(~height, data = cdc)

* **What do the values of the x-axis mean in terms of people's heights?**
* **What do the *widths* of the bins mean?**
* **What does the *height* of each bar represent?**

# More on Histograms

* By changing the width of the *bins* of a histogram, we can change the amount of detail it shows.
  + Wide bins give us a very broad view of the data.
  + Narrow bins give us a very detailed view.
* A good histogram should strike a balance between being both wide & narrow

# More on Histograms

* Run the following commands:

histogram(~height, data = cdc, nint = 3)

histogram(~height, data = cdc, nint = 41)

* **Describe what happens to the number of *bins* as the value of nint changes**
* **Find a value of nint that you think strikes the best balance between broad and detailed views of the data**
  + **Describe this histograms distribution in terms of it's *center*, *shape* and *spread*.**