

Plots and statistics for categorical
and quantitative data

Overview

Review of vectors and continuation of data frames

Statistics and plots for categorical data in R

Statistics and plots for quantitative data in R

Announcement: Homework 1

Due Sunday September 8th at 11pm

- I recommend getting started early on this!

To download the homework please do the following:

> `library(SDS230)`

> `download_homework(1)`

From the file panel, open the homework and try knitting it

Review: vectors

Creating vectors

```
> s <- c("statistics", "data", "science", "fun")
```

```
> z <- 2:10
```

Accessing elements of vectors

```
> s[4]
```

```
> s[c(1, 2)]
```

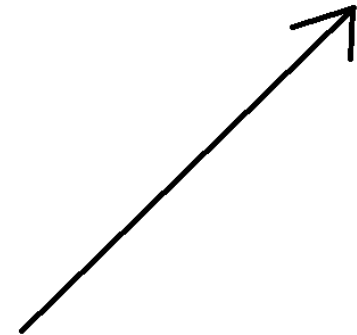
Applying functions to vectors

```
> sqrt(z)
```

```
> mean(z)
```

```
> z > 3
```

You just got



Vectored

OKCupid data

The screenshot shows the OKCupid website interface. At the top, the OKCupid logo is on the left, and a navigation bar contains links for Messages, Matches, Connections, and Treasures. A user is logged in as 'BigDaddyC_taco', with a 'Sign out' button. The profile page for 'BigDaddyC_taco' is displayed, showing a profile picture, age (21), gender (M), orientation (Straight), status (Single), and location (Chicago, Illinois). The profile is marked as 'Online Now'. Below the profile picture are tabs for 'About', 'Photos', 'Questions', and 'Personality'. The 'About' tab is selected, showing a 'My self-summary' and 'What I'm doing with my life' sections. The 'My Details' section is also visible on the right.

My self-summary

I'm a young, ambitious and outgoing individual. I love traveling, having recently been to South America and through the southern states on a road trip with friends. I'm a very caring/emotional person. I enjoy anything artistic and always up for new activities. Also, I've been told I'm too perfect.

What I'm doing with my life

- Working two marketing jobs in downtown and Lincoln Park areas of Chicago.
- Full-time student at DePaul University studying Marketing/Sales.
- Volunteer on South Side of Chicago (Pilsen, Little Village & Englewood).
- Writer for my blog, The Plaid Tie

My Details

Last Online	Online now!
Ethnicity	Hispanic / Latin
Height	6' 0" (1.83m).
Body Type	Fit
Diet	Mostly anything
Smokes	No
Drinks	Rarely
Drugs	Never

Did everyone read the article [The Big Lies People Tell in Online Dating?](#)

Thoughts?

Review: Data frames

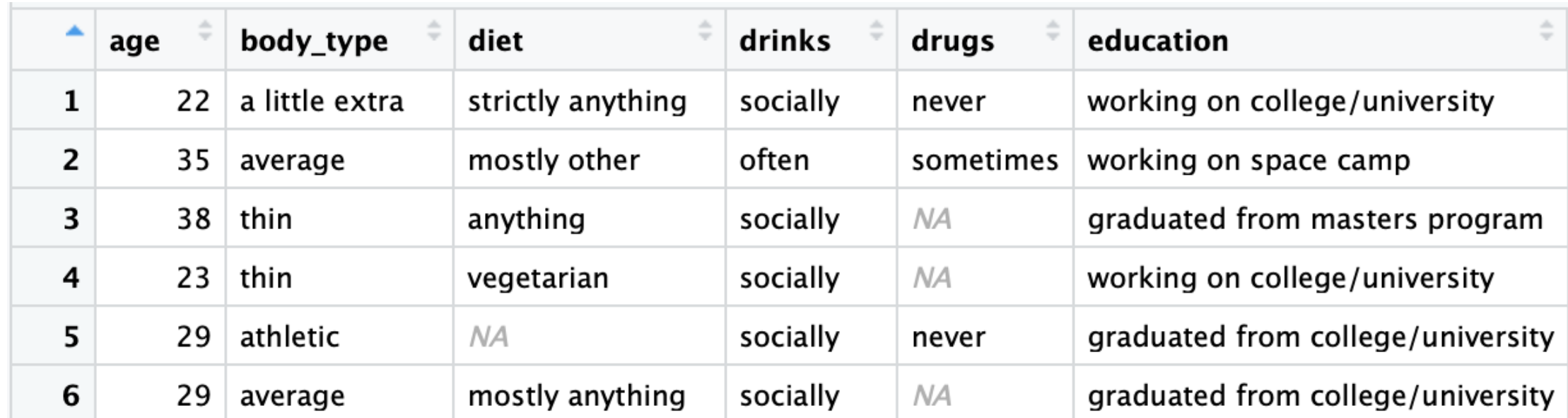
Data frames contain structured data

- > `library(SDS230)`
- > `download_data("profiles_revised.csv")` *# only needs to be run once*
- > `profiles <- read.csv("profiles_revised.csv")`
- > `View(profiles)` *# the View() function only works in R Studio!*

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Review: Data Frames

Variables



	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Cases

An Example Dataset

Quantitative Variable

Categorical Variable

Cases
(observational units)

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Review: Data frames

We can extract the columns of a data frame as vector objects using the \$ symbol

```
> the_ages <- profiles$age
```

Can you get the `mean()` age of users in this data set?

```
> mean(the_ages)
```

Review: Extracting rows from a data frame

We can extract rows from a data frame in a similar way as extracting values from a vector by using the square brackets

```
> profiles[1, ] # returns the first row of the data frame
```

```
> profiles[, 1] # returns the first column of the data
```

Note, the first column of the profiles data frame is the variable *age*, so we can also get the first column using:

```
> profiles$age # this is the same as profiles[, 1]
```

Review: Extracting rows from a data frame

We can create vectors of numbers specifying which rows we want to extract from a data frame

```
# create a vector with the numbers 1, 10, 20
```

```
> my_vec <- c(1, 10, 20)
```

```
# use my_vec to get the 1st, 10th, and 20th row in profiles
```

```
> small_profiles <- profiles[my_vec, ]
```

```
> dim(small_profiles) # number of rows and columns in the data frame
```

Review: Extracting rows from a data frame

Finally, we can also extract rows by creating a Boolean vector that is of the same length as the number of rows in the data frame

TRUE values will be extracted from the data frame, while **FALSE** values will not

```
# create a vector of booleans
```

```
> my_bools <- c(TRUE, FALSE, TRUE)
```

```
# use the Boolean vector to get the 1st and 3rd row
```

```
> small_profiles[my_bools, ]
```

Questions?



Let's try some quick warm-up exercises in R!

Categorical data

Categorical variables

What is a categorical variable?

- A: A categorical variable assigns each observation to one of k groups

What are some of the categorical variables in the profiles data frame?

- Is heights a categorical variable?

For categorical variables, we usually want to view:

- How many items are each category OR
- The proportion (or percentage) of items in each category

$$\text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$

Categorical data

```
# Get information about drinking behavior
```

```
> drinking_vec <- profiles$drinks
```

```
# Create a table showing how often people drink
```

```
> drinks_table <- table(drinking_vec)
```

```
> drinks_table
```


Relative frequency table

We can create a relative frequency table using the function:

```
> prop.table(my_table)
```

Can you create a relative frequency table for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
```

```
> prop.table(drinks_table)
```

What is the proper statistical notation for these values: \hat{p} or π ?

Bar plots

(pun intended?)

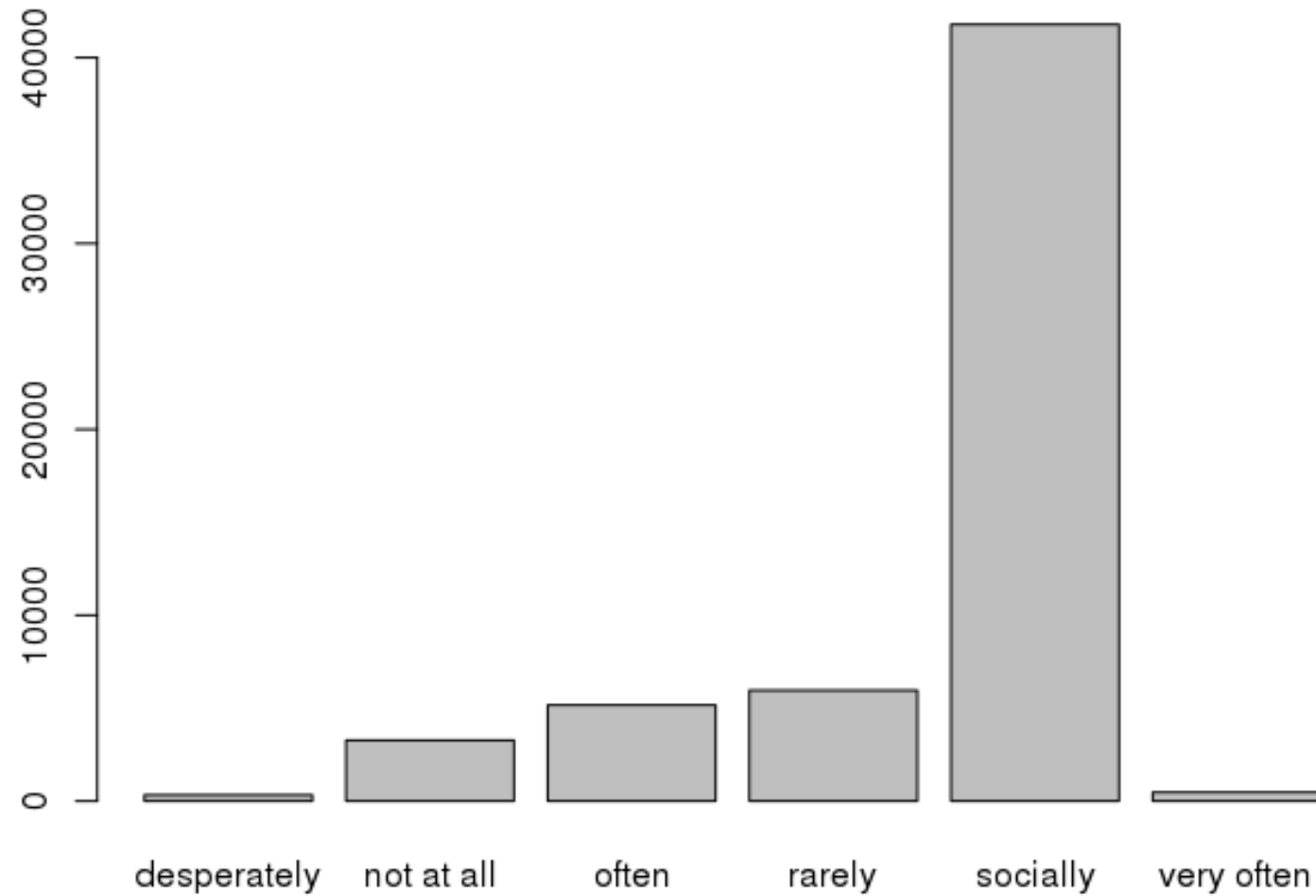
We can plot the number of items in each category using a bar plot

```
> barplot(my_table)
```

Can you create a bar plot for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
```

```
> barplot(drinks_table)
```



What is wrong with this plot?

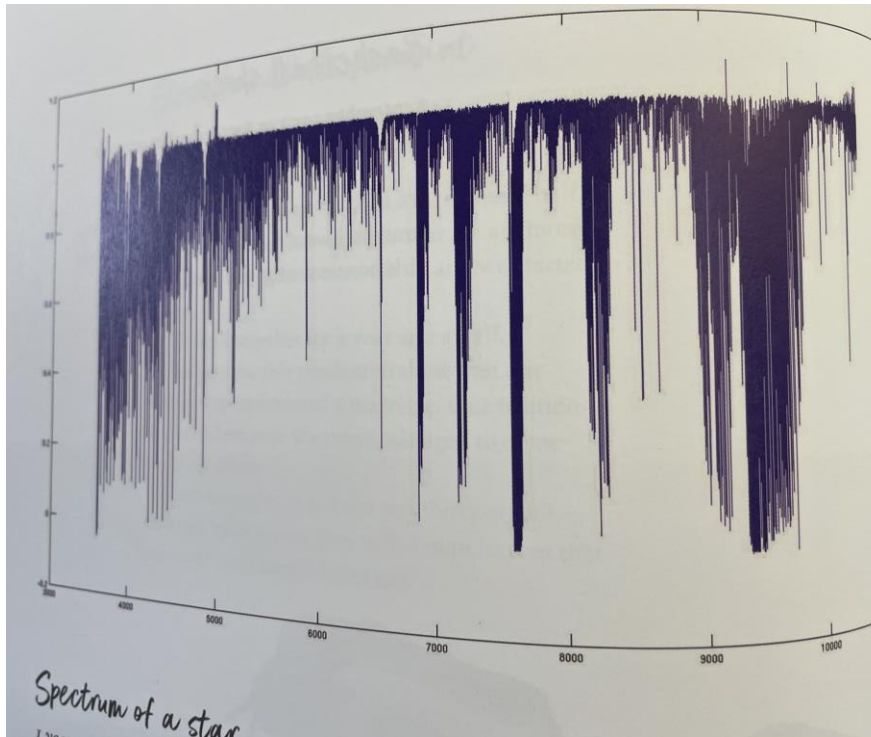
- A: the axes are not labeled!!!



If you don't want exes, label you axes!

Side note: the museum of broken relationships

Last summer I went to Croatia and visited the [Museum of Broken Relationships](#)...



A spectrum of a star

1 year

Beijing, China

We are both astronomers. On my 26th birthday he sent me a spectrum of a star in the Orion constellation as my birthday gift. This star, named π_3 , is 26 light years away from the Earth.

He said, 'Look, at the time when you were born, the light left this star, passing through the endless interstellar space, the countless dust and nebula, arriving here after a 26-light-year journey. So, have you. Here you meet your starlight, and I meet you.'

(你看,在你出生的那一刻,有一束光芒从这颗恒星出发,

它穿越漫无边际的星际空间,穿过数不尽的尘埃和星云,经过

26光年的旅程来到这里。

你从那一刻起,也经过漫长的岁月的旅程来到这里。

你和你的星光在这里相遇,

我和你在这里相遇。)

I could only hear the sound of my heart beating then.

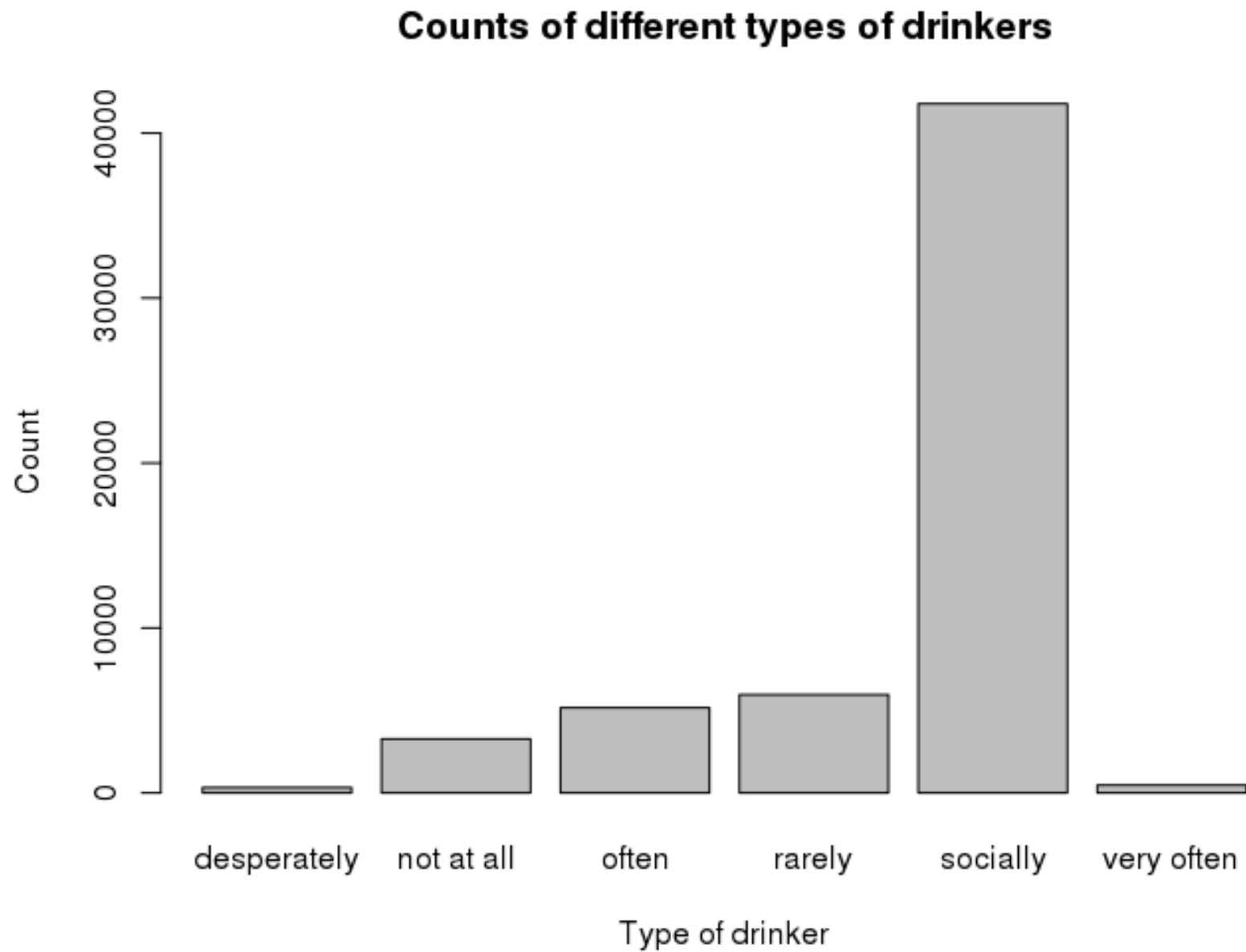
Though we have since broken up, every time I see the Orion constellation, I relive some sweet memories.

Details matter!

Can you figure out how to label the axes?

- A: ? barplot
- A: xlab and ylab!

```
> barplot(drinks_table,  
          ylab = "Count",  
          xlab = "Type of drinker",  
          main = "Counts of different types of drinkers")
```

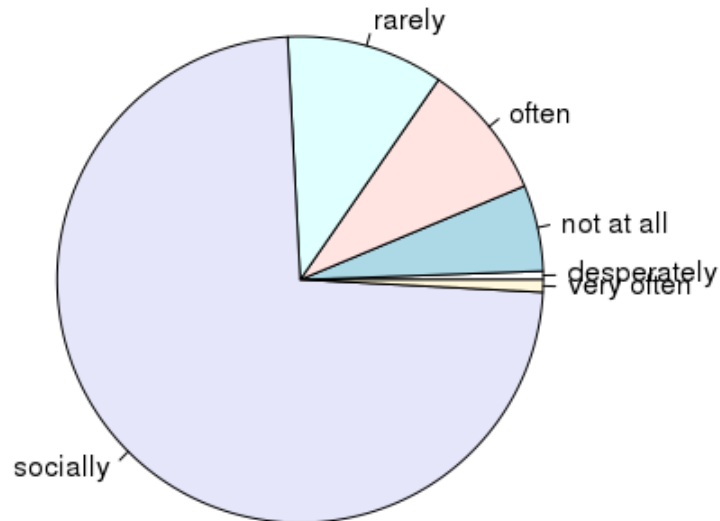


So much better!!!

Pie charts

We can also use the `pie()` function to create pie charts

> `pie(drinks_table)`



World's Most Accurate Pie Chart



Which is best: bar plots or pie charts?

```
> barplot(table(profiles$sex, useNA = "always"))
```

```
> pie(table(profiles$sex, useNA = "always"))
```

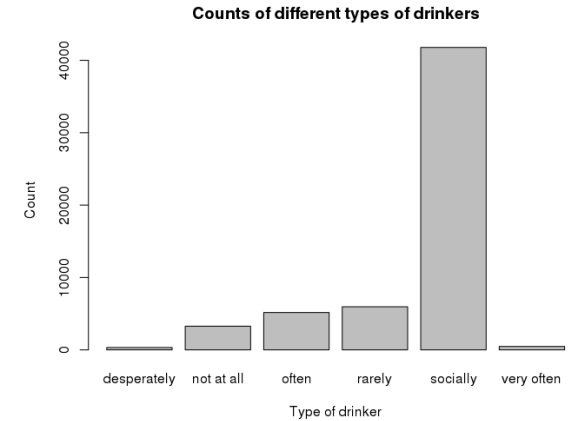
**BE
BEST**

Q1: Is one better than the other?

Q2: Can you figure out how to add colors to these plots?

Removing social drinkers

Social drinkers are dominating our plot 😞



We can get rid of social drinkers by only plotting counts less than 10,000

```
> nonsocial_inds <- drinks_table < 10000  
> nonsocial_drinks_table <- drinks_table[nonsocial_inds]  
> barplot(nonsocial_drinks_table)
```

It's a Match!



You and Booze have liked each other.

Questions?



Quantitative data

Quantitative data: statistics

There are several statistics that describe the central tendency of quantitative data?

- The mean: `mean()`
- The median: `median()`

Which of these measures is robust to outliers?

Can you calculate the mean and median of OkCupid user's heights?

What went wrong?

`mean(v, na.rm = TRUE)`

What is the proper statistical notation for the mean of OkCupid user's heights:
 \bar{x} or μ ?

Quantitative data: Visualizing heights

Q: How can we visualize the heights in the profiles data frame?

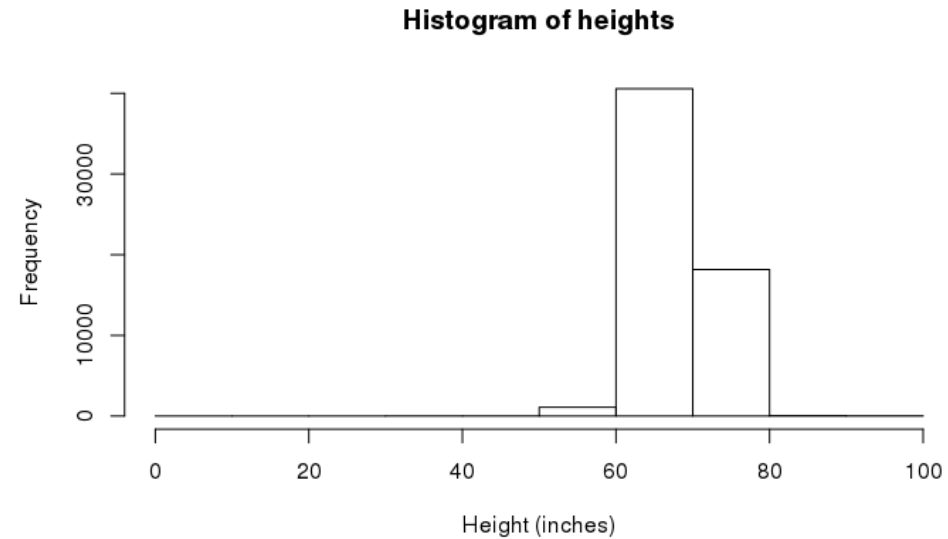
A: Histograms!

A: Boxplots

A: Many other options too

Histograms of heights

Height (inches)	Frequency Count
(0-10]	6
(10-20]	0
(20-30]	1
(30-40]	13
(40-50]	9
(50-60]	1097
(60-70]	40575
(70-80]	18164
(80-90]	50
>90	28



Visualizing heights

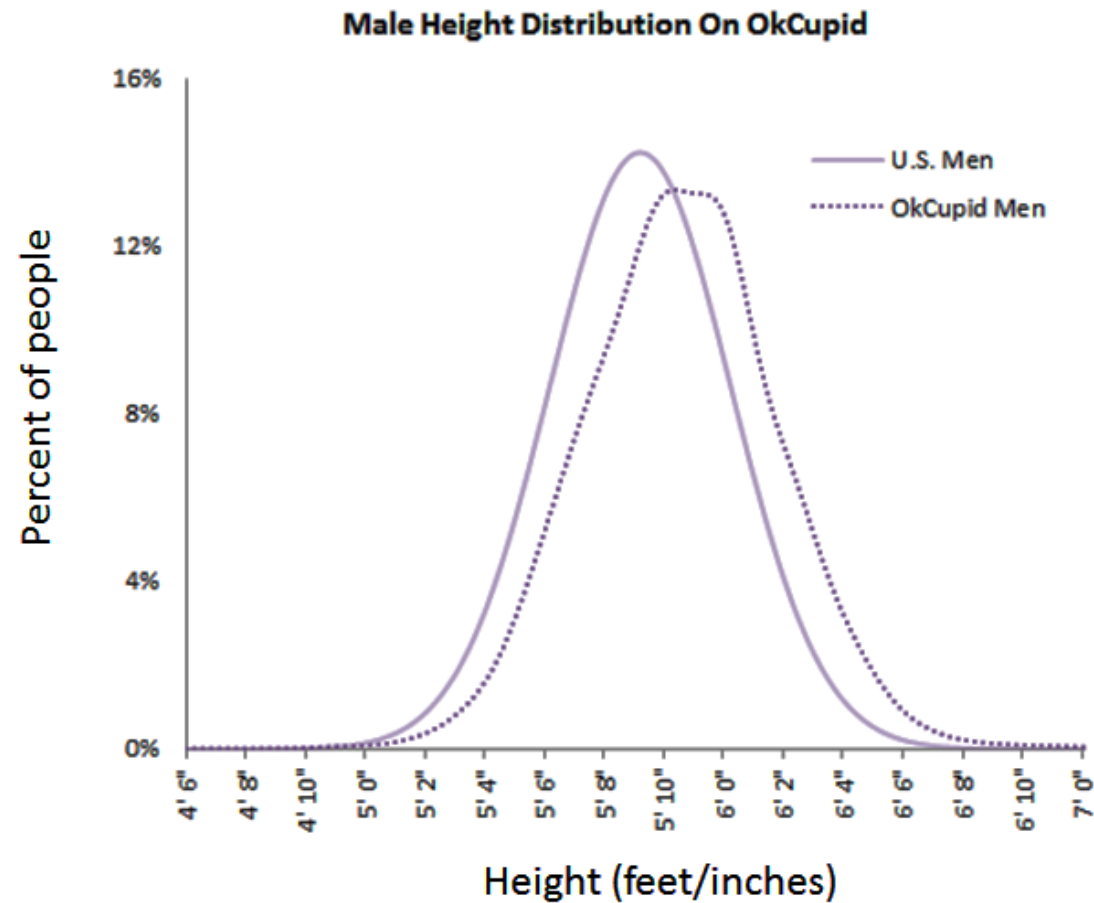
We can create histograms in R using the `hist()` function

Can you create a histogram of heights?

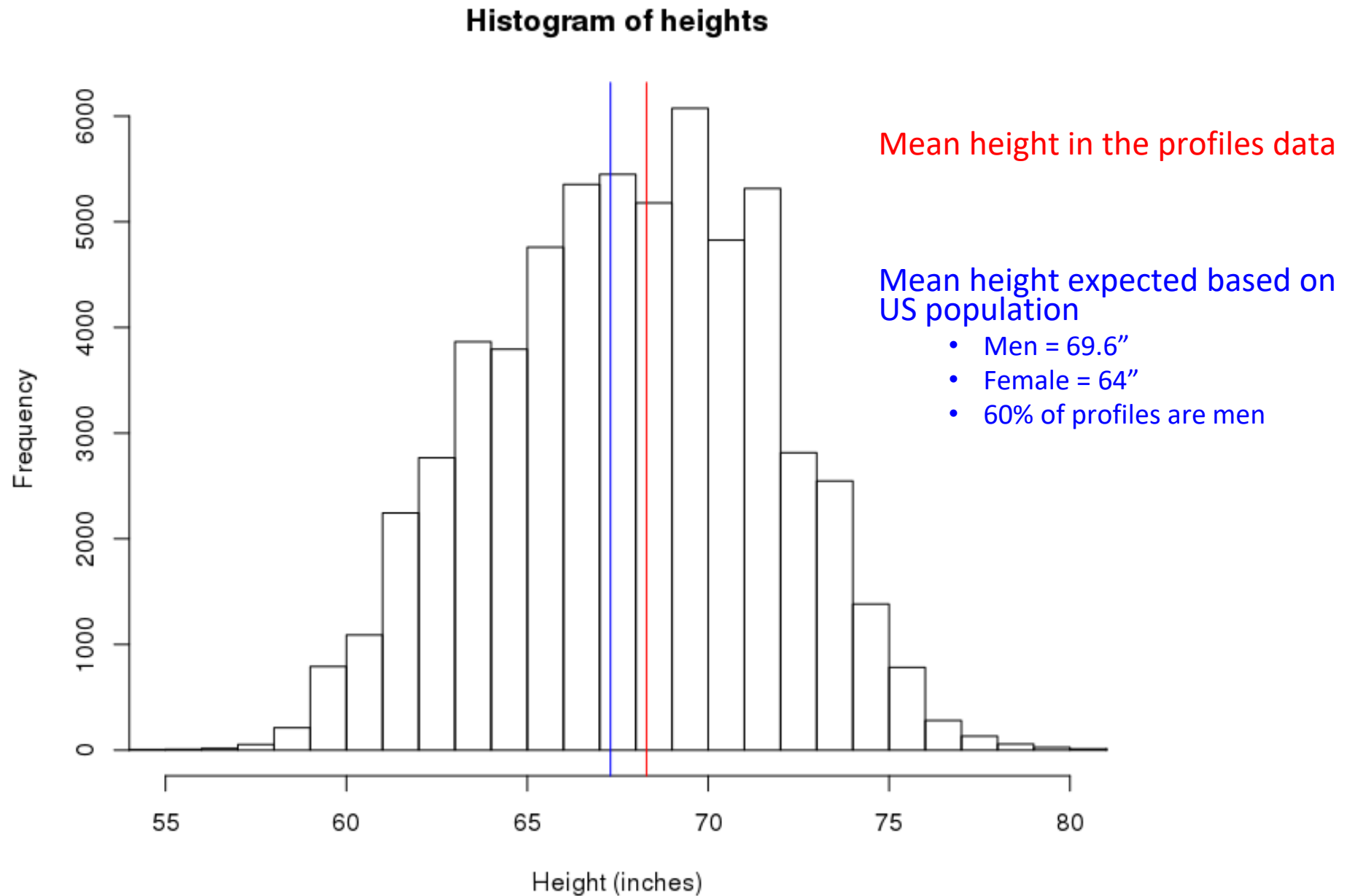
```
> hist(profiles$height)
```

```
> hist(profiles$height, breaks = 50)
```

OkCupid users are taller than the average person



Can we see this in the profiles data?



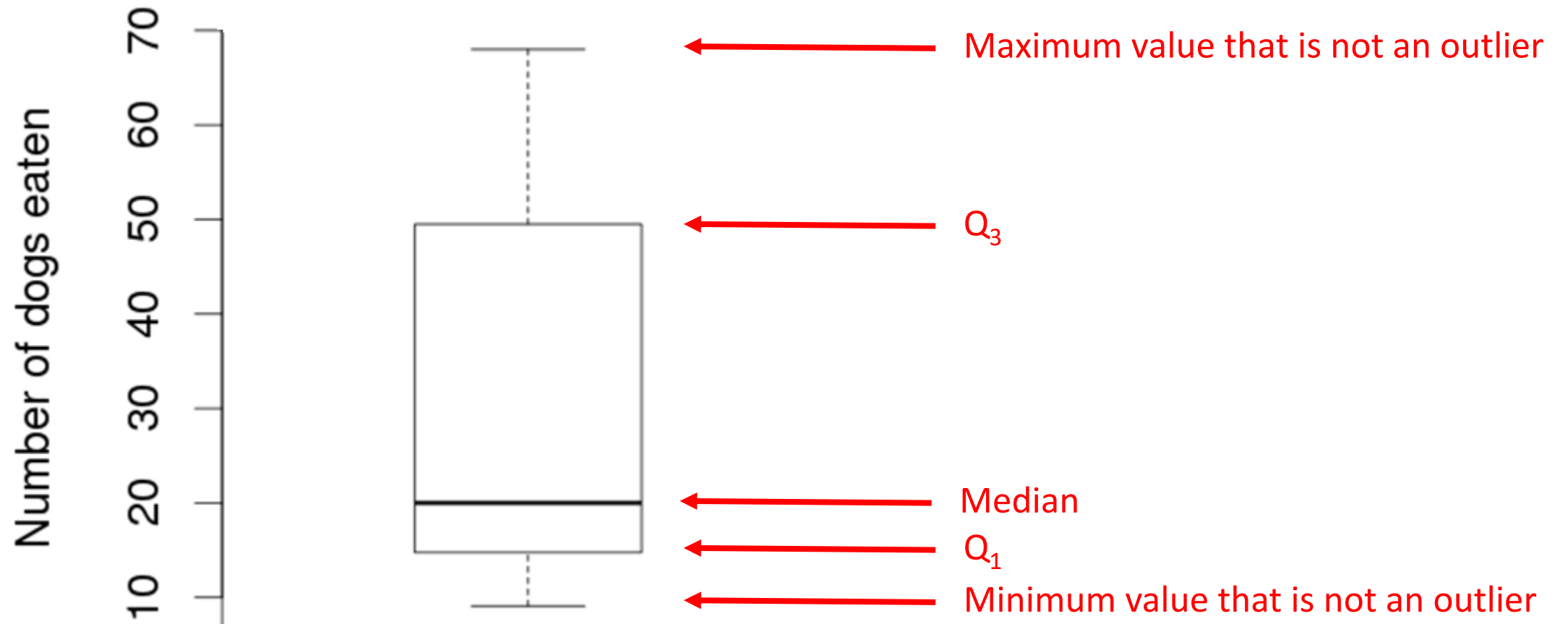
`abline()` adds lines to plots

Has Categorical Data



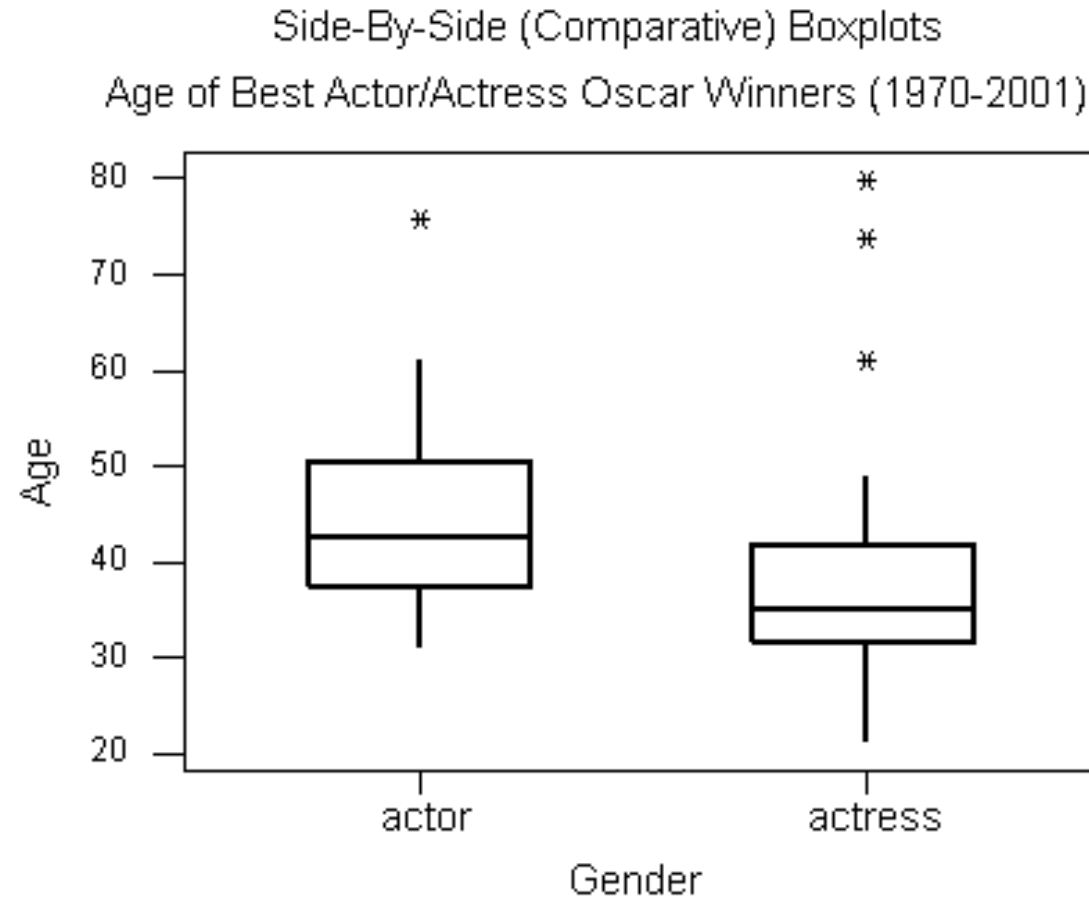
Uses Histogram

Box plots can also visualize quantitative data



R: `boxplot(v)`

Side-by-side boxplots



Useful for comparing distributions!

- What does the figure above show?

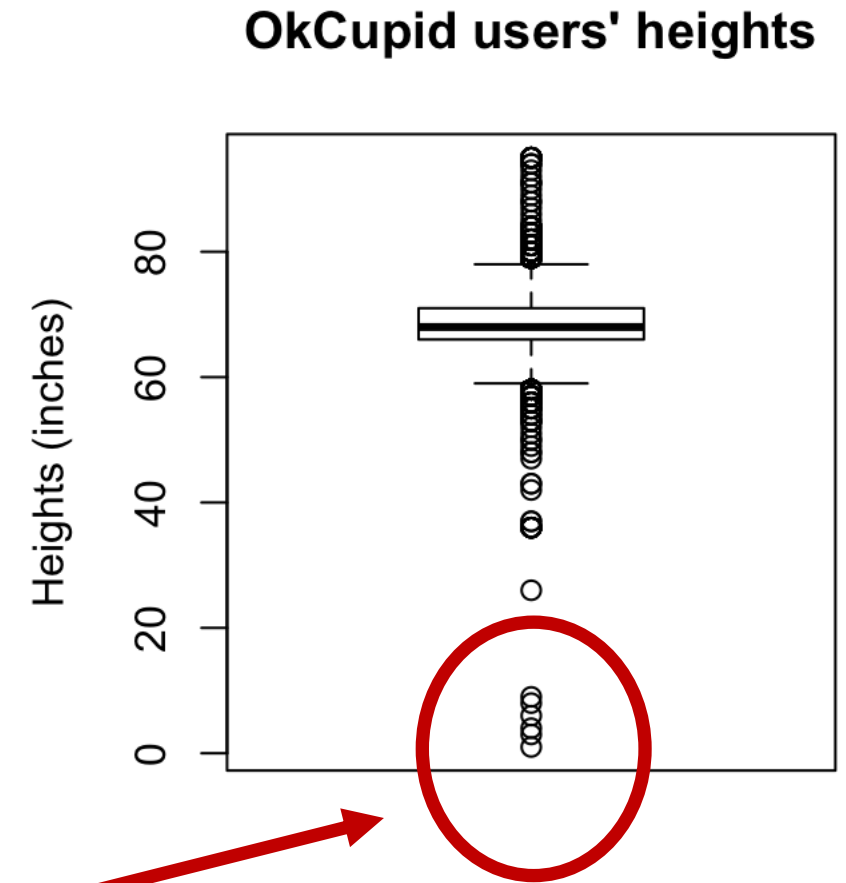
Outliers

Outliers on boxplots are values that are more than $1.5 * IQR$

What should we do if we have outliers?

Investigate!

- If there are due to an error, remove them



People under 20" tall?

Outliers

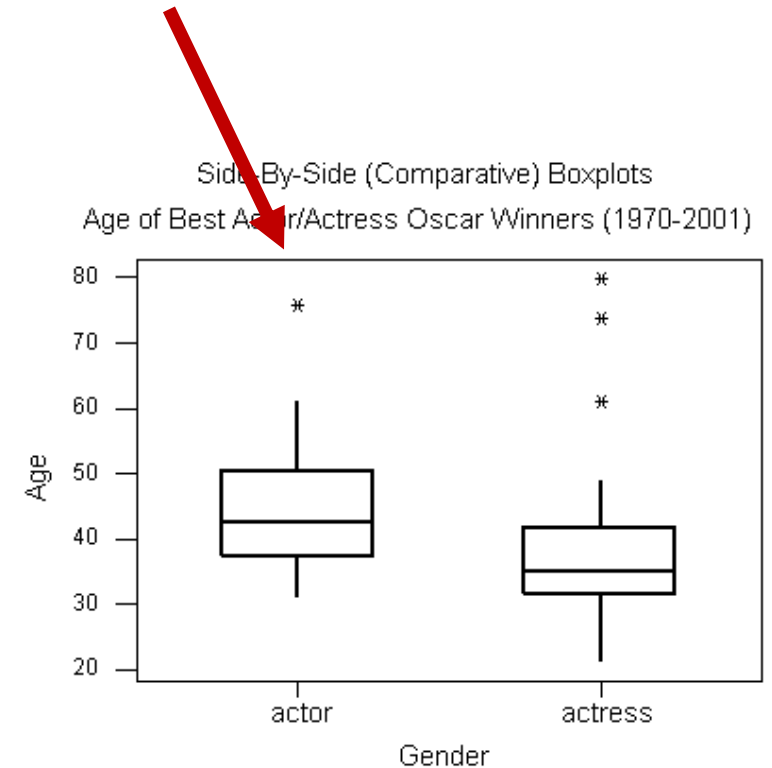
Outliers on boxplots are values that are more than $1.5 * IQR$

What should we do if we have outliers?

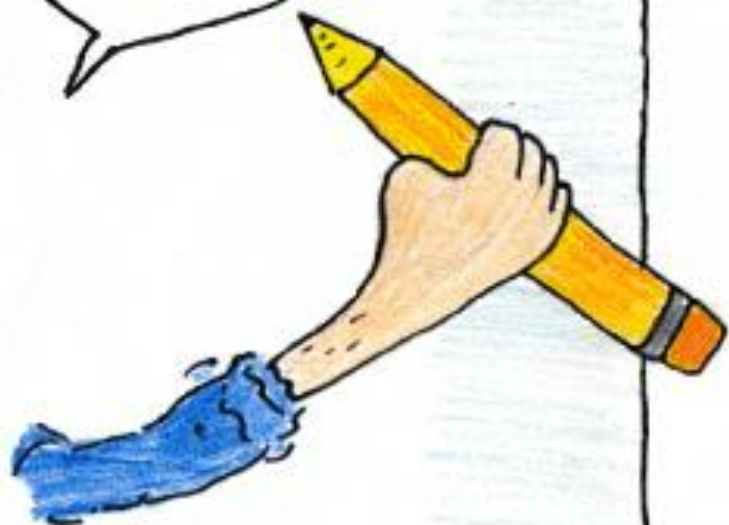
Investigate:

- If there are due to an error, remove them
- **If not, need to account for them**

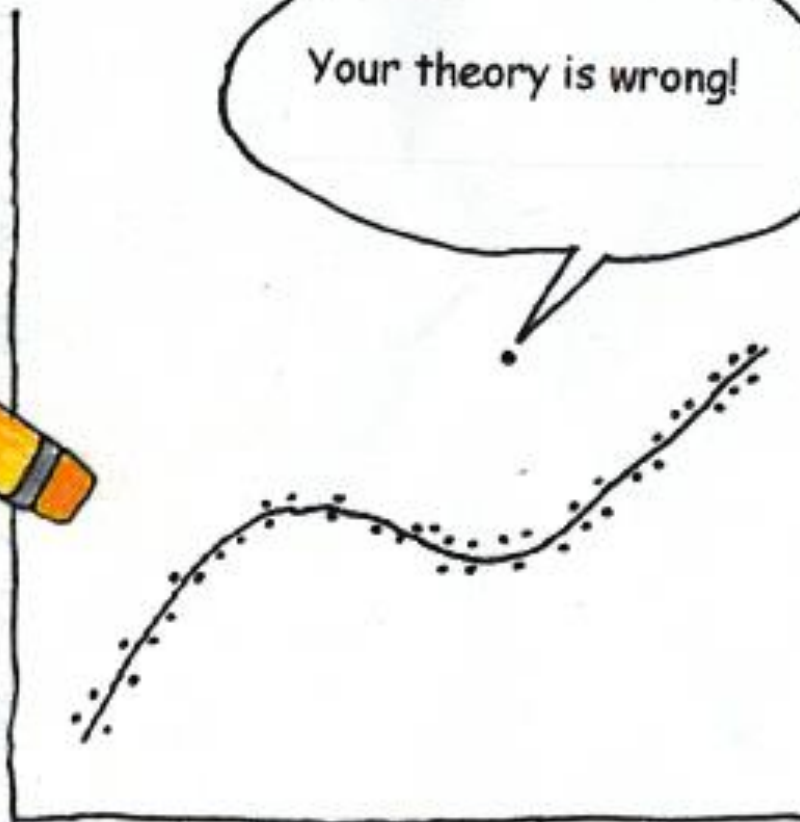
Who is this actor?



Out, liar!



Your theory is wrong!



Ben Shabat

Questions?



CitiBike data

Let's look at the bike share data from NYC

```
> load('daily_bike_totals.rda')
```



[CitiBike analysis](#)

What does each case correspond to?

We can use the `dim()` function to get how many cases and variables there are

- How many are there?

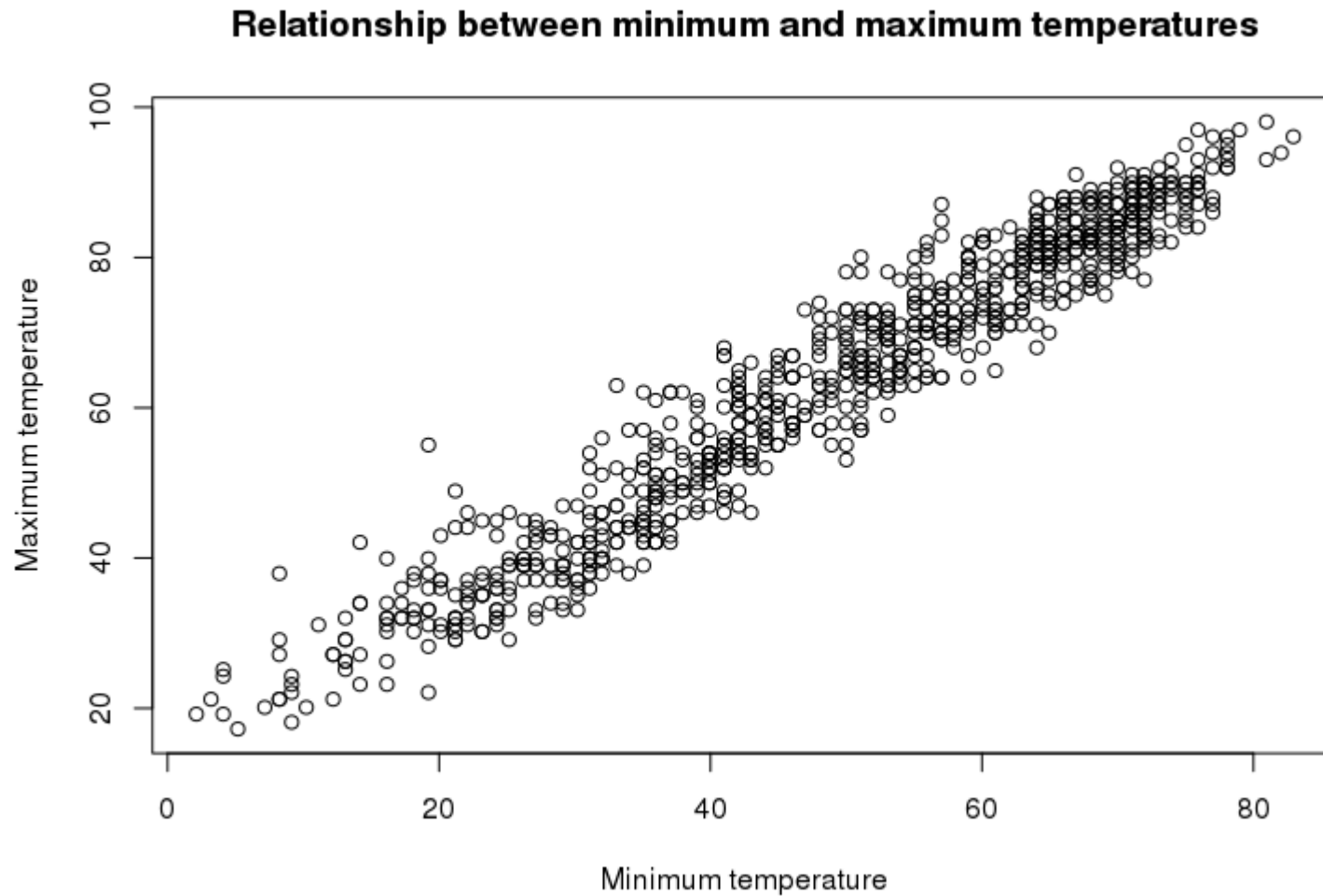
Scatter plots

We can use the `plot(x, y)` function to create scatter plots

Can you create a scatter plot of the relationship between the minimum and maximum temperatures?

```
> plot(bike_daily_data$min_temperature,  
       bike_daily_data$max_temperature,  
       xlab = "Minimum temperature",  
       ylab = "Maximum temperature",  
       main = "Relationship between min and temp")
```

Scatter plots



Plotting time series

We can use the `plot(x, y)` function to plot time series

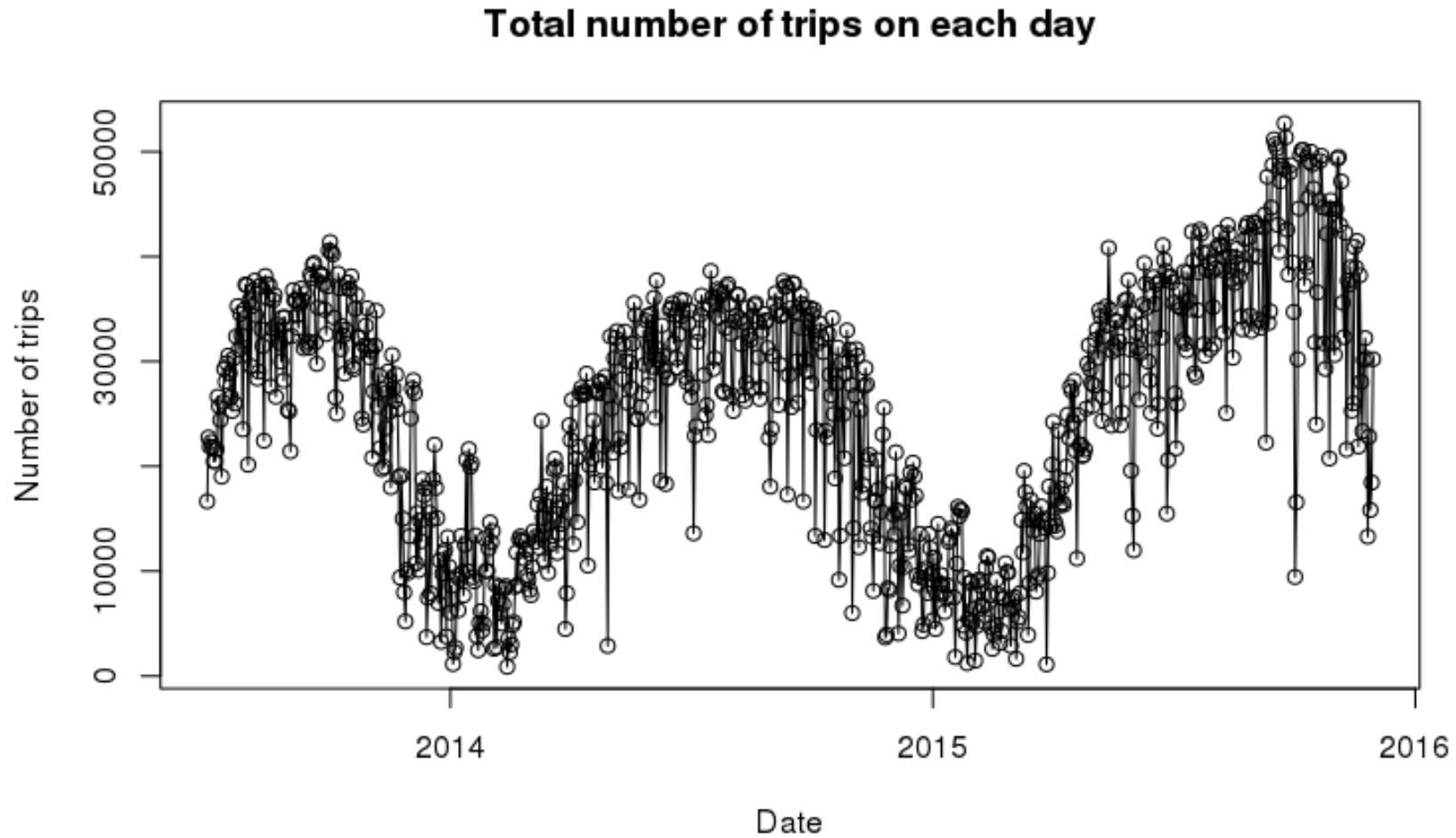
we can connect the points in a plot using

> `plot(x, y, type = 'l')` # connected points

> `plot(x, y, type = 'o')` # both points and dots

```
> plot(bike_daily_data$date, bike_daily_data$trips,  
       type = 'o',  
       xlab = "Date",  
       ylab = "Number of trips",  
       main = "Total number of trips on each day")
```

Plotting time series



Announcement: Homework 1

Due Sunday September 8th at 11pm

- I recommend getting started early on this!

To download the homework please do the following:

> `library(SDS230)`

> `download_homework(1)`

From the file panel, open the homework and try knitting it

For loops

For loops are useful when you want to repeat a piece of code many times under similar conditions

The syntax for a for loop is:

```
for (i in 1:100) {
```

```
    # do something
```

```
}
```



This is repeated 100 times
i is incremented by 1 each time

Homework 1

Homework 1:

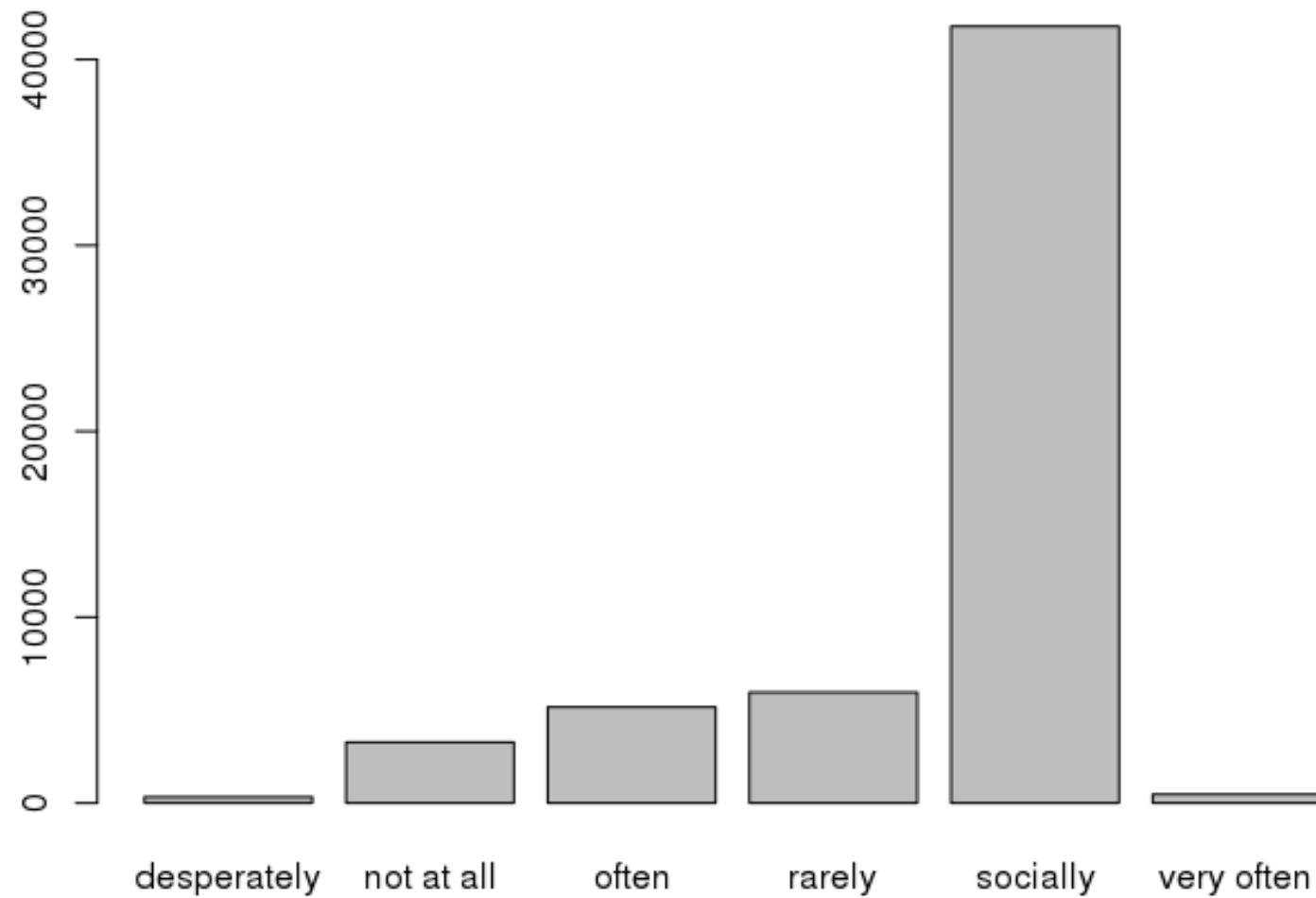
> [SDS230::download_homework\(1\)](#)

Due on Gradescope by 11pm on Sunday September 11th

- Instructions for how to submit homework on Gradescope are on Canvas

Review of categorical data in R

Let's quickly review plots and statistics in R



What is missing on this plot?



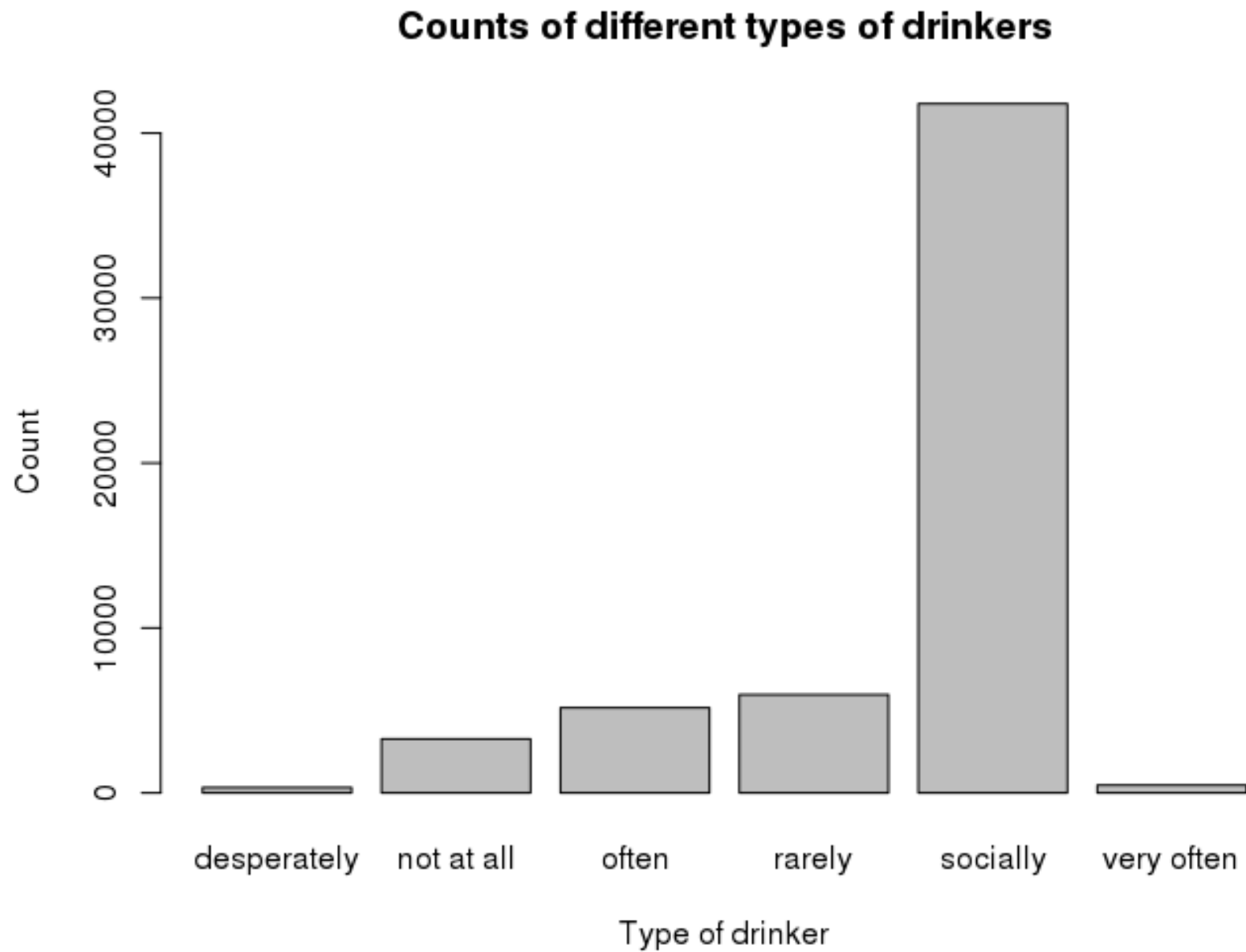
If you don't want exes, label you axes!

Details matter!

Can you figure out how to label the axes?

- A: ? barplot
- A: xlab and ylab!

```
> barplot(drinks_table,  
          ylab = "Count",  
          xlab = "Type of drinker",  
          main = "Counts of different types of drinkers")
```



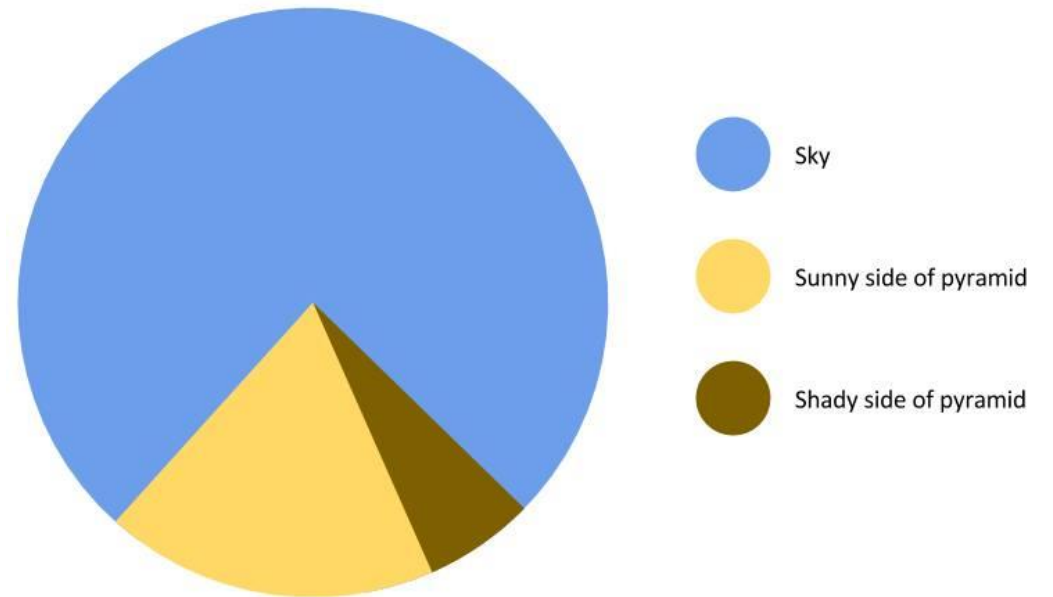
So much better!!!

Review of R from last class

Plotting categorical data

```
> barplot(drinks_table,  
          ylab = "Count",  
          xlab = "Type of drinker",  
          main = "Counts of drinkers")
```

```
> pie(drinks_table)
```



Questions about the logistics?



Review and continuation of...

Data frames and structured data

Statistics and plotting of categorical and quantitative data

OK Cupid data

The screenshot shows the OkCupid website interface. At the top, the OkCupid logo is on the left, and navigation links for Messages, Matches, Connections, and Treasures are in the center. A user is logged in as 'BigDaddyC_taco'. The profile for 'BigDaddyC_taco' is displayed, showing a photo of a man, his age (21), gender (M), orientation (Straight), status (Single), and location (Chicago, Illinois). He is marked as 'Online Now'. Below the photo are tabs for About, Photos, Questions, and Personality. The 'About' tab is selected, showing a 'My self-summary' and 'What I'm doing with my life' sections. To the right of the summary is a 'My Details' table.

BigDaddyC_taco
21 / M / Straight / Single
Chicago, Illinois
Online Now

My self-summary

I'm a young, ambitious and outgoing individual. I love traveling, having recently been to South America and through the southern states on a road trip with friends. I'm a very caring/emotional person. I enjoy anything artistic and always up for new activities. Also, I've been told I'm too perfect.

What I'm doing with my life

- Working two marketing jobs in downtown and Lincoln Park areas of Chicago.
- Full-time student at DePaul University studying Marketing/Sales.
- Volunteer on South Side of Chicago (Pilsen, Little Village & Englewood).
- Writer for my blog, The Plaid Tie

My Details

Last Online	Online now!
Ethnicity	Hispanic / Latin
Height	6' 0" (1.83m).
Body Type	Fit
Diet	Mostly anything
Smokes	No
Drinks	Rarely
Drugs	Never

Survey question 1: Are you familiar with the website OkCupid?

Are you familiar with website OkCupid?

Yes, I know the site	11 respondents	17%	<div></div> ✓
I am vaguely familiar with it	36 respondents	55%	<div></div>
I have never heard of it before	19 respondents	29%	<div></div>
		0%	<div></div>

Data Frames

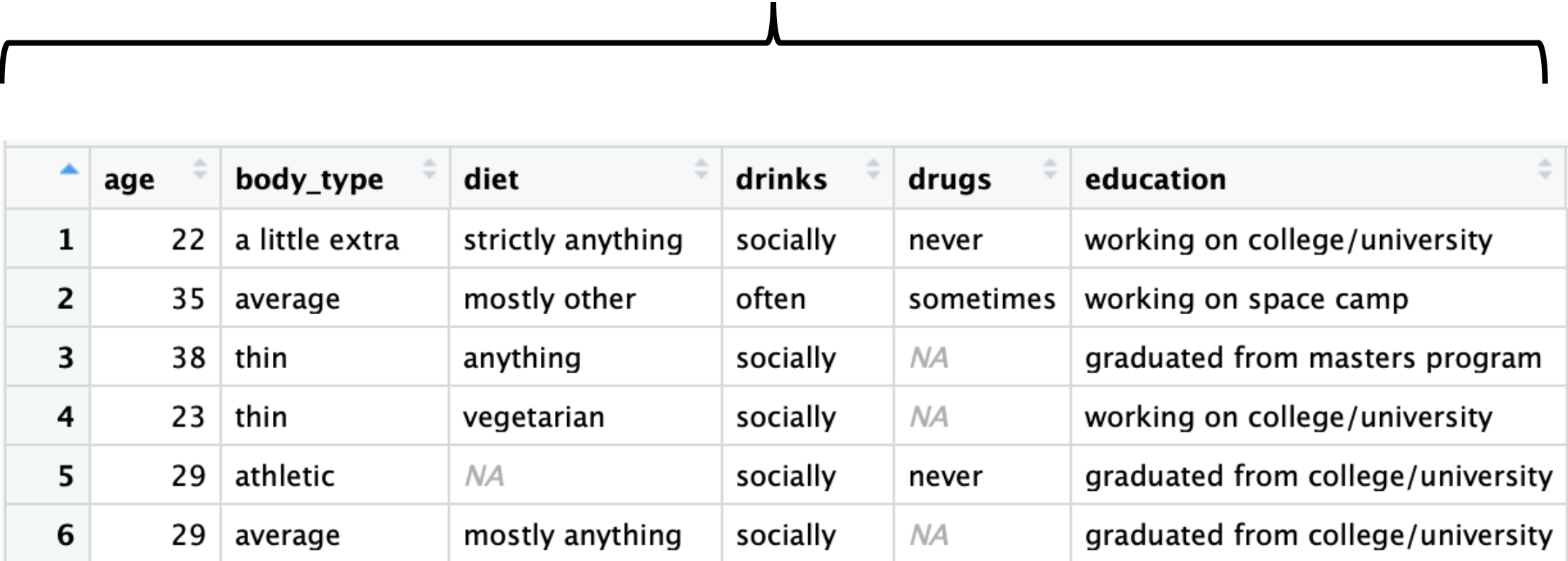
Data frames contain structured data

▲	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Data Frames

Variables

Cases



	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Data Frames

Quantitative Variable

Categorical Variable

Cases
(observational units)

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

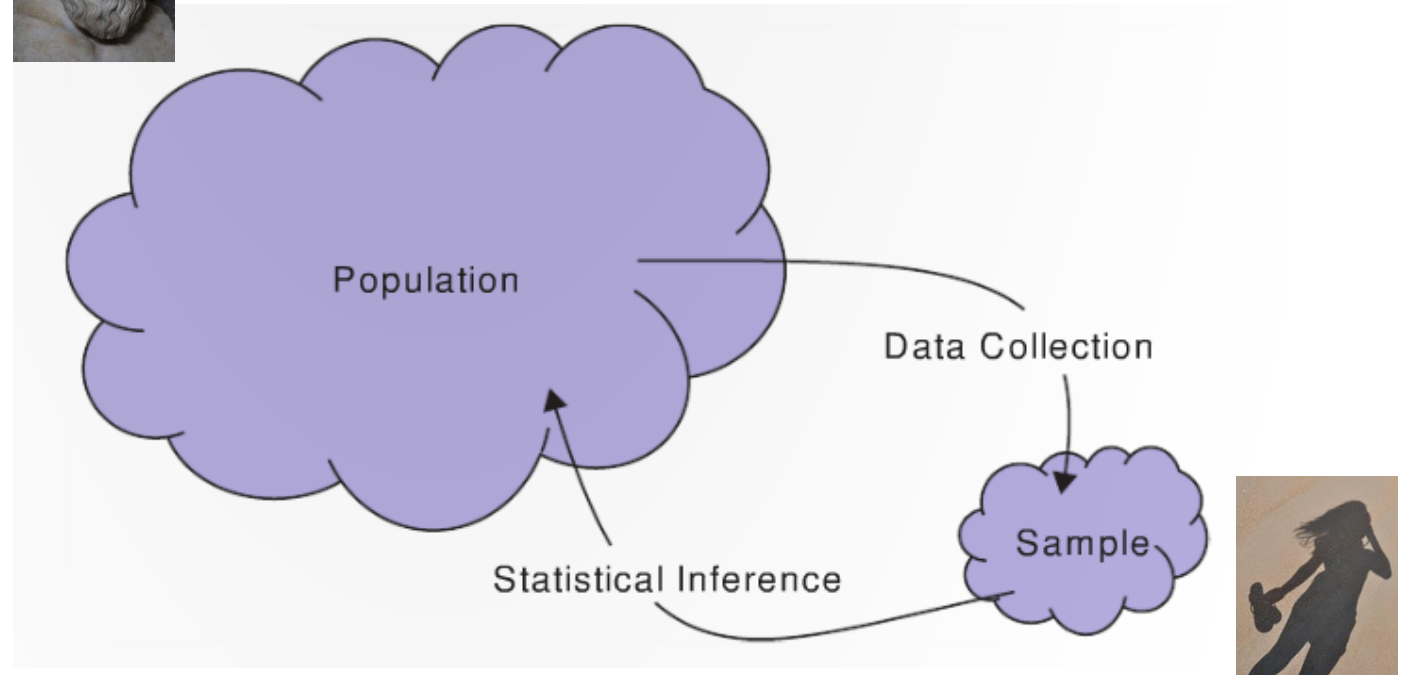
Where does data come from?



DATA SCIENCE!!!



Population: all individuals/objects of interest



Sample: A subset of the population

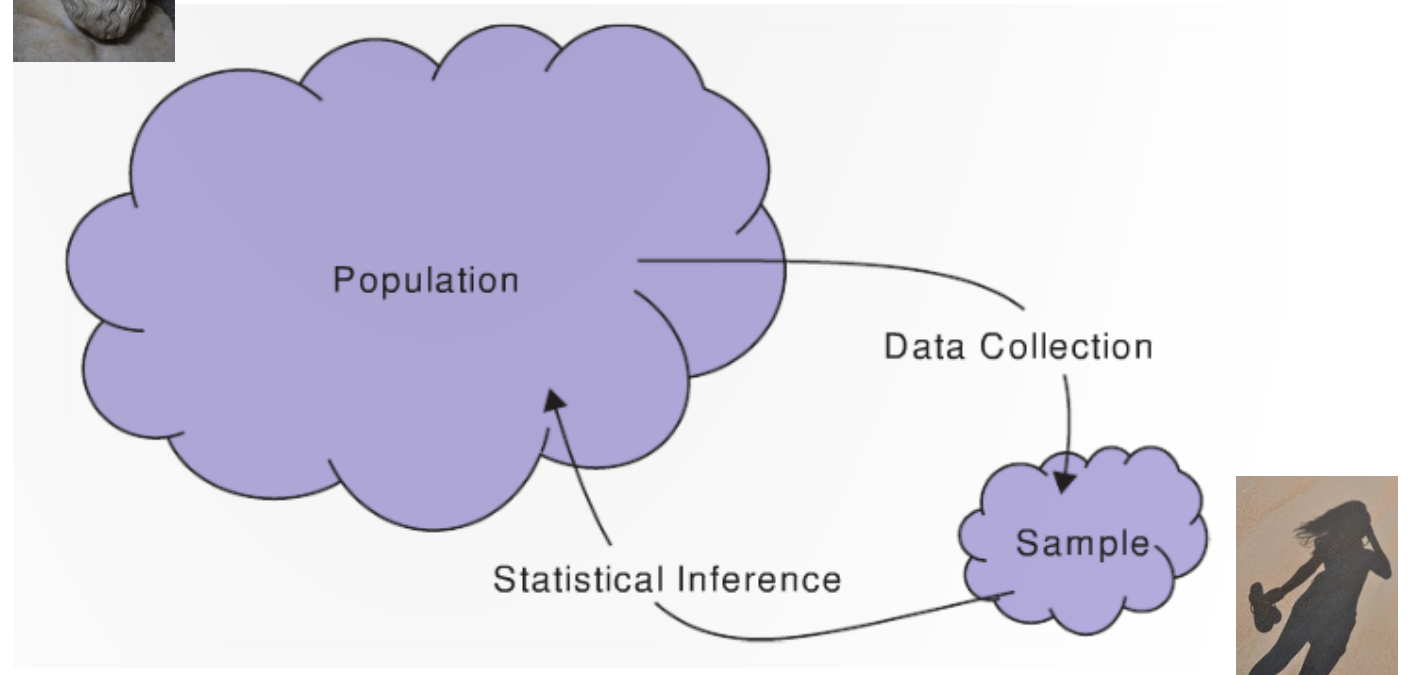
Where does data come from?

Question: Is the okcupid profiles data frame a population or a sample?

Question: If the OkCupid profiles data frame is a sample, what is the population?



Parameters: $\pi, \mu, \sigma, \rho, \beta$



Statistics: $\hat{p}, \bar{x}, s, r, b$

How do we get sample of data?

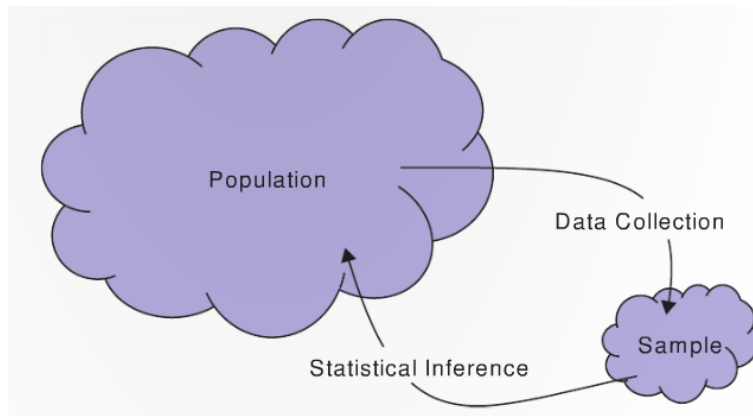
Simple random sample: each member in the population is equally likely to be in the sample

“Random selection”

Soup analogy!

Q: Why is this good?

A: Allows for generalizations to the population! (no bias)



Question: Is the OkCupid profiles data a simple random sample?

Observational and experimental studies

An **experiment** is a study in which the researcher actively controls one or more of the explanatory variables.

An **observational study** is a study in which the researcher does not actively control the value of any variable but simply observes the values as they naturally exist

Question: Is the OkCupid profiles data from an experiment or observational study?



Observational and experimental studies

Question: Would it be ethical to run an experiment on OkCupid users?

- [OkCupid experiments on humans](#)

OkCupid deleted all the text on everyone's profiles and measured how star ratings changed

- i.e., only the picture was available

The screenshot shows an OkCupid profile for a user named 'H<3artbreaker', who is 19 years old and male. The profile includes a photo of a young man, a match percentage of 5% (Match), 3% (Friend), and 98% (Enemy). There are five star ratings for the profile. The profile is divided into sections: 'About', 'Photos', and 'The Two of Us'. The 'About' section contains three paragraphs of text. The 'His Details' section is a table with the following information:

His Details	
Height	5' 7"
Diet	Mostly anything
Drinks	--
Drugs	--
Sign	Pisces
Job	Artist/musical/writer
Income	More than \$1,000,000
Pets	Likes dogs

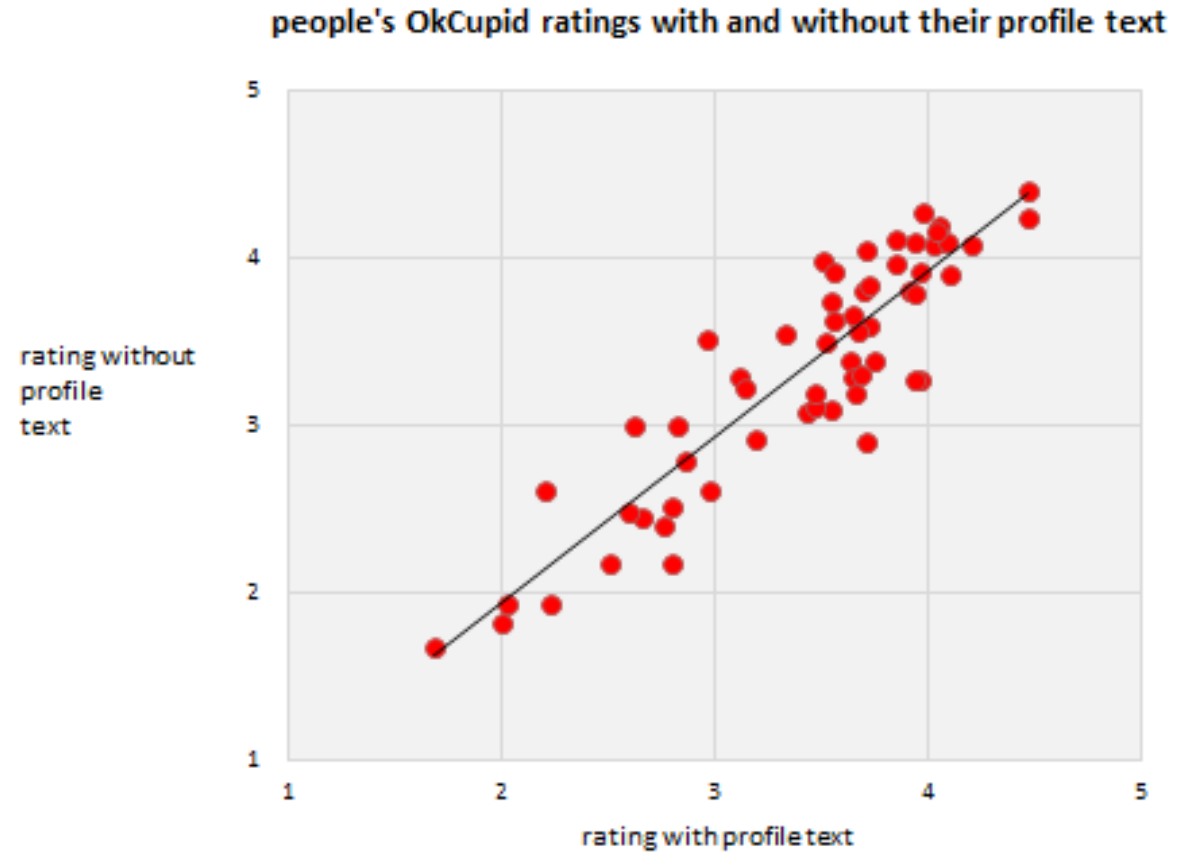
Observational and experimental studies

Question: Would it be ethical to run an experiment on OkCupid users?

- [OkCupid experiments on humans](#)

OkCupid deleted all the text on everyone's profiles and measured how star ratings changed

- i.e., only the picture was available







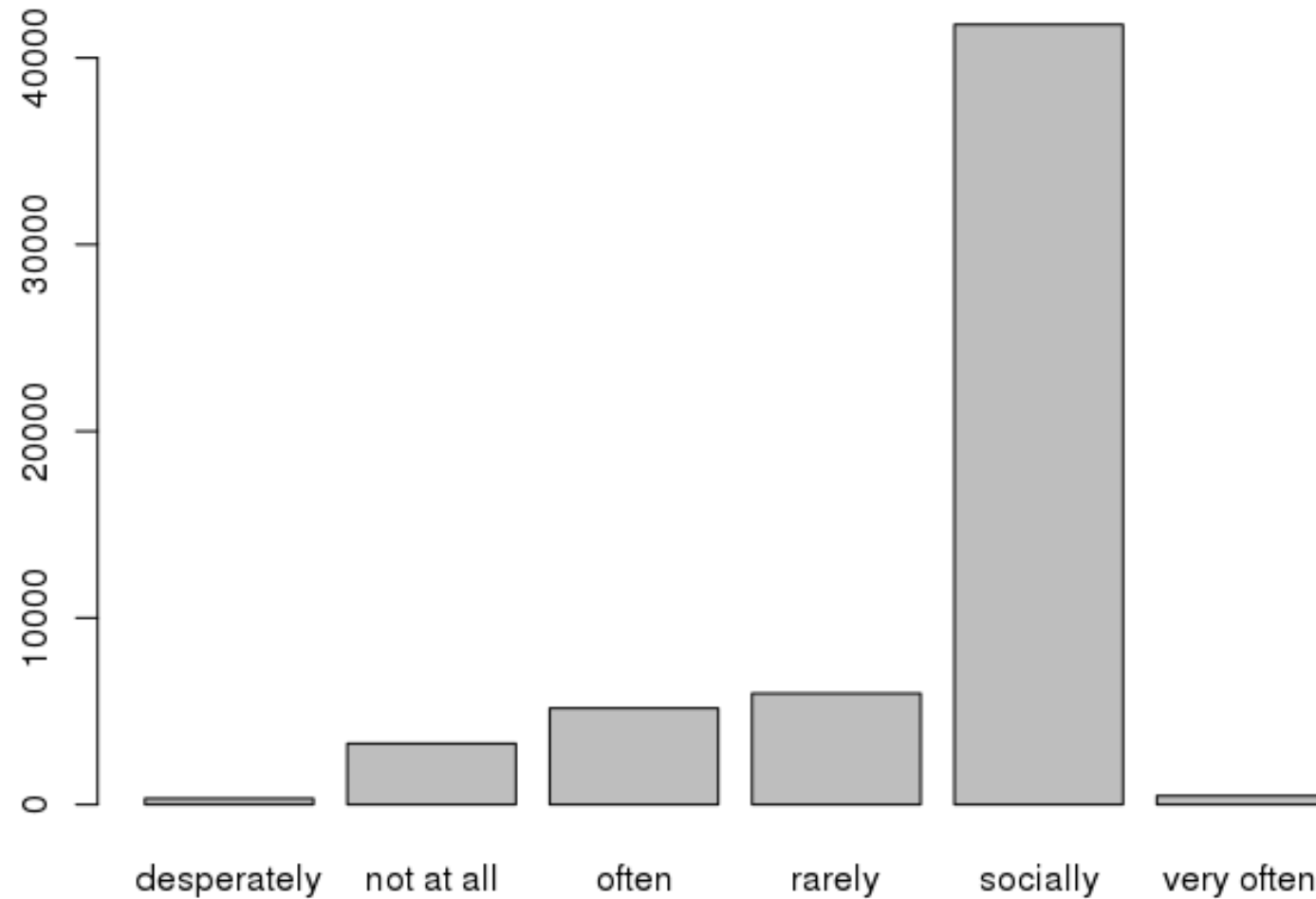
Questions?



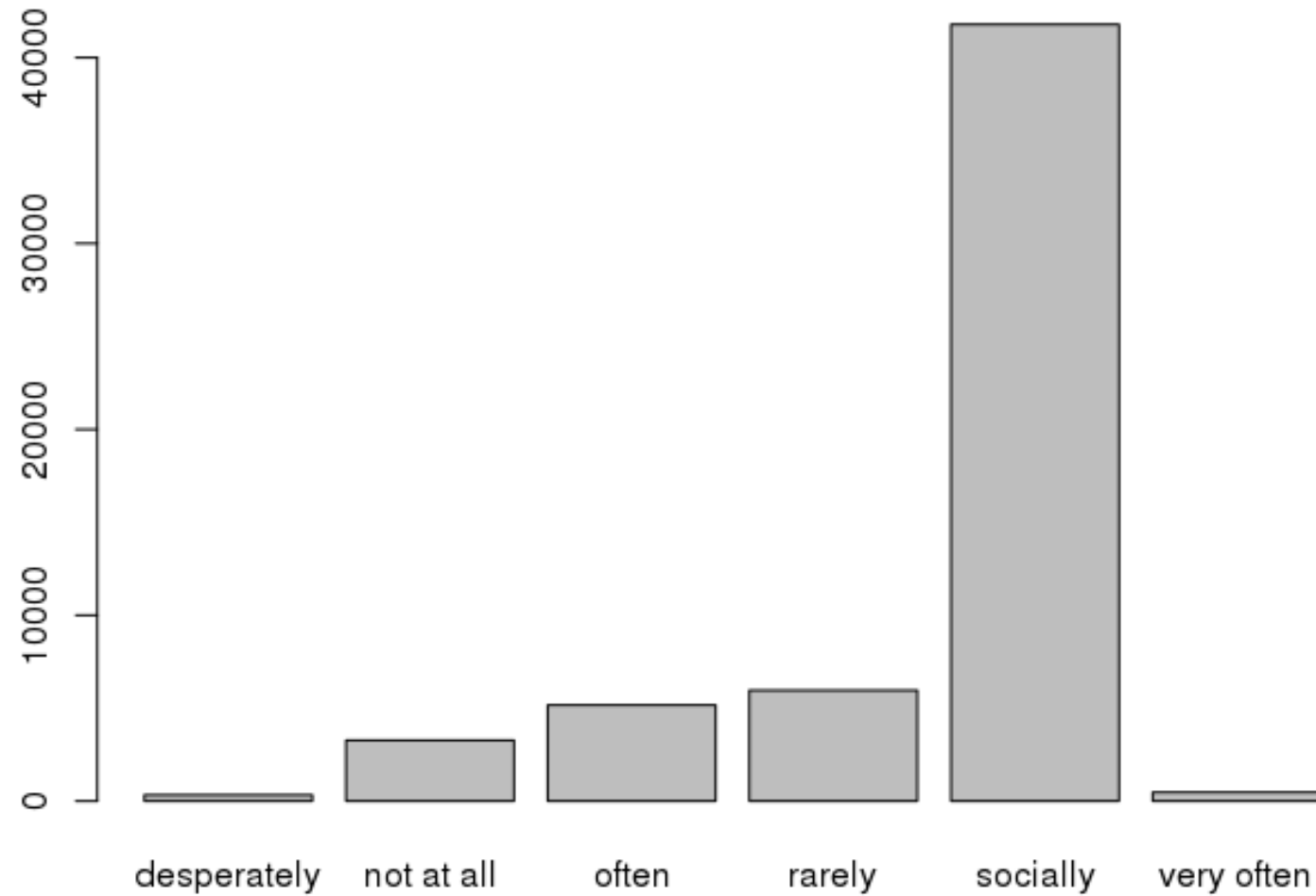
Let's quickly review plots and statistics in R

Is height a categorical variable in the OkCupid data frame?

Yes	8 respondents	9%	 ✓
No	60 respondents	67%	
Yes and No depending on the analysis	22 respondents	24%	
I'm not sure		0%	



Survey question 5: What is missing on this plot and can you figure out how to add it?



A: the axes are not labeled!!!



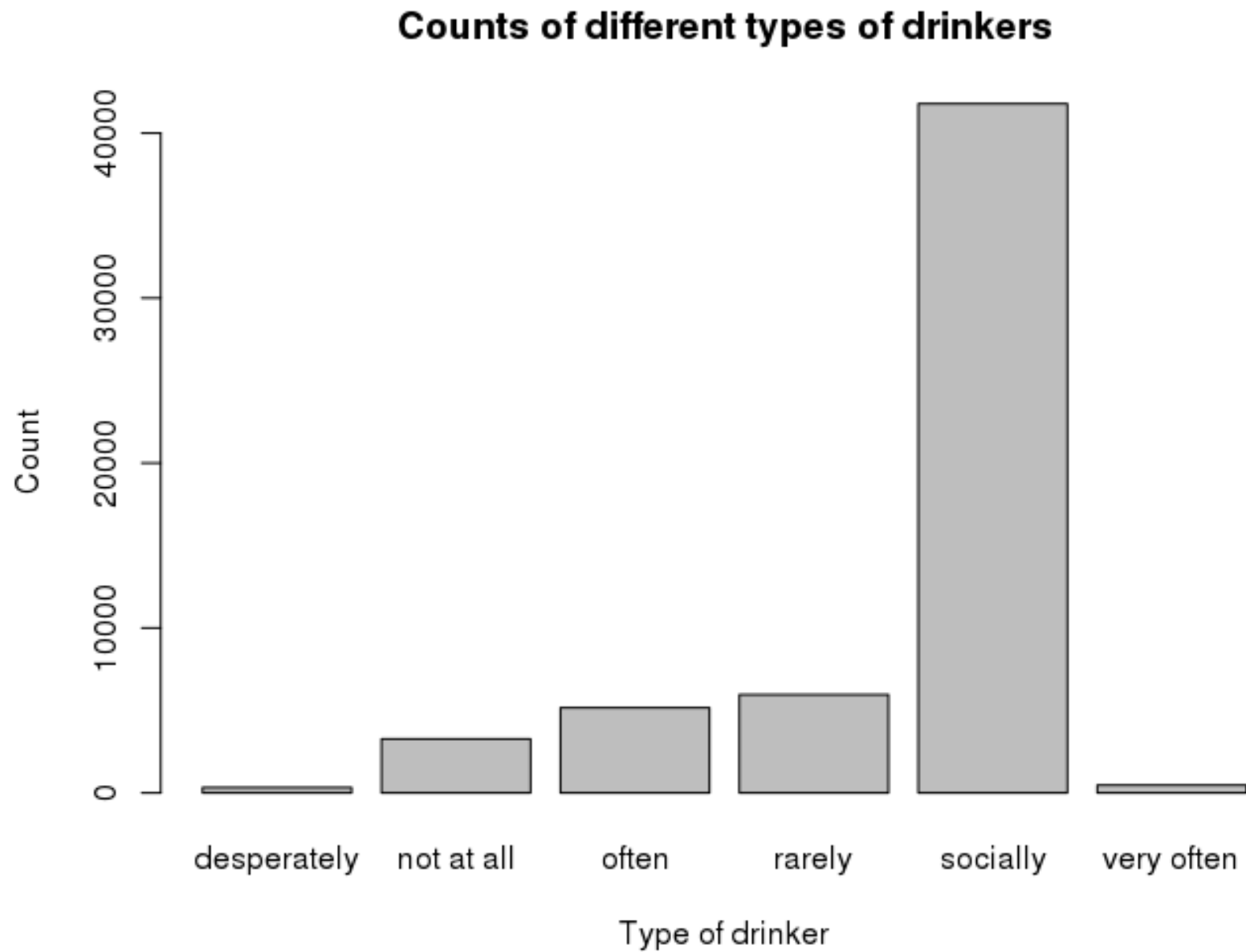
If you don't want exes, label you axes!

Details matter!

Can you figure out how to label the axes?

- A: ? barplot
- A: xlab and ylab!

```
> barplot(drinks_table,  
          ylab = "Count",  
          xlab = "Type of drinker",  
          main = "Counts of different types of drinkers")
```



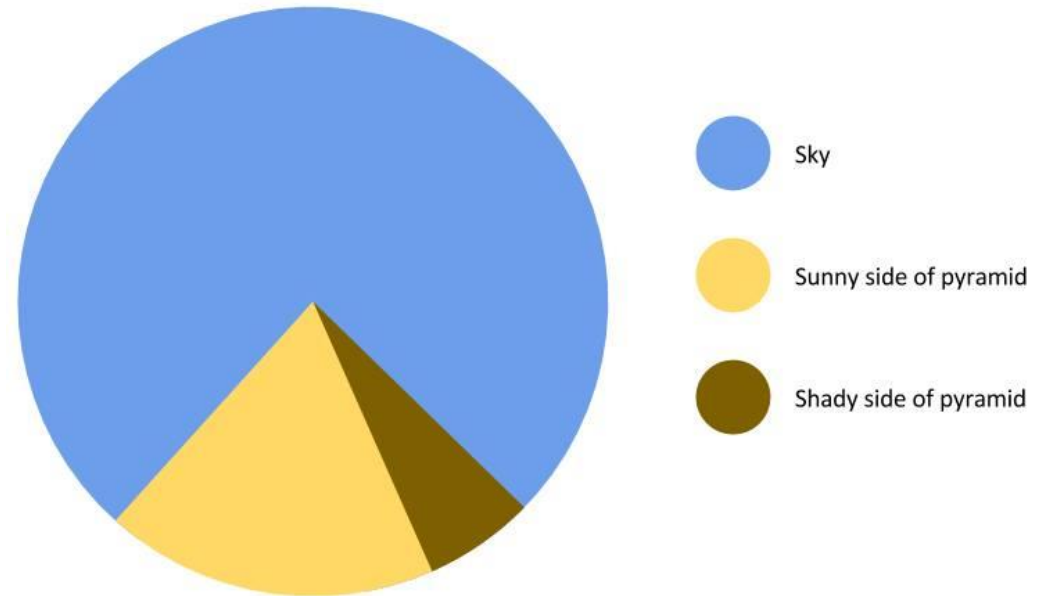
So much better!!!

Review of R from last class

Plotting categorical data

```
> barplot(drinks_table,  
          ylab = "Count",  
          xlab = "Type of drinker",  
          main = "Counts of drinkers")
```

```
> pie(drinks_table)
```



Homework 1

> `library(SDS230)`

> `download_homework(1)`

Due on Gradescope by 11:59pm on Sunday September 13th

- Instructions for how to submit homework on Gradescope are on Canvas

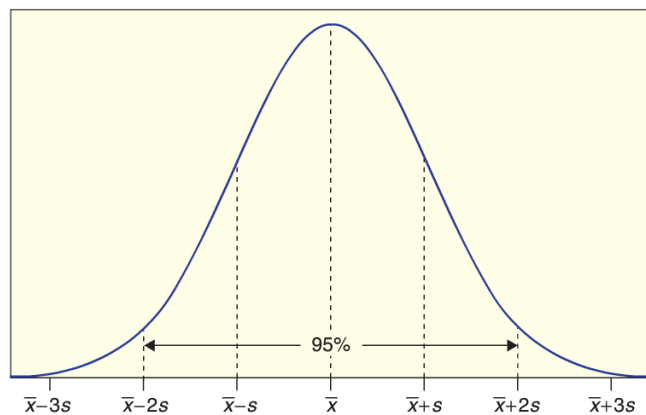
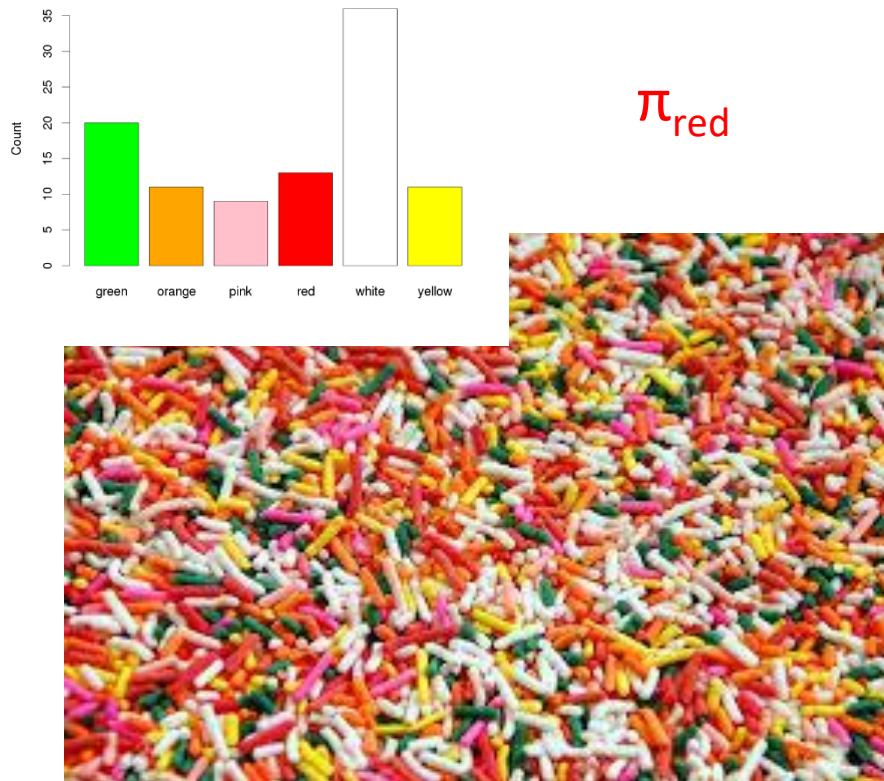
Sampling distributions

A distribution of ***statistics*** is called a ***sampling distribution***

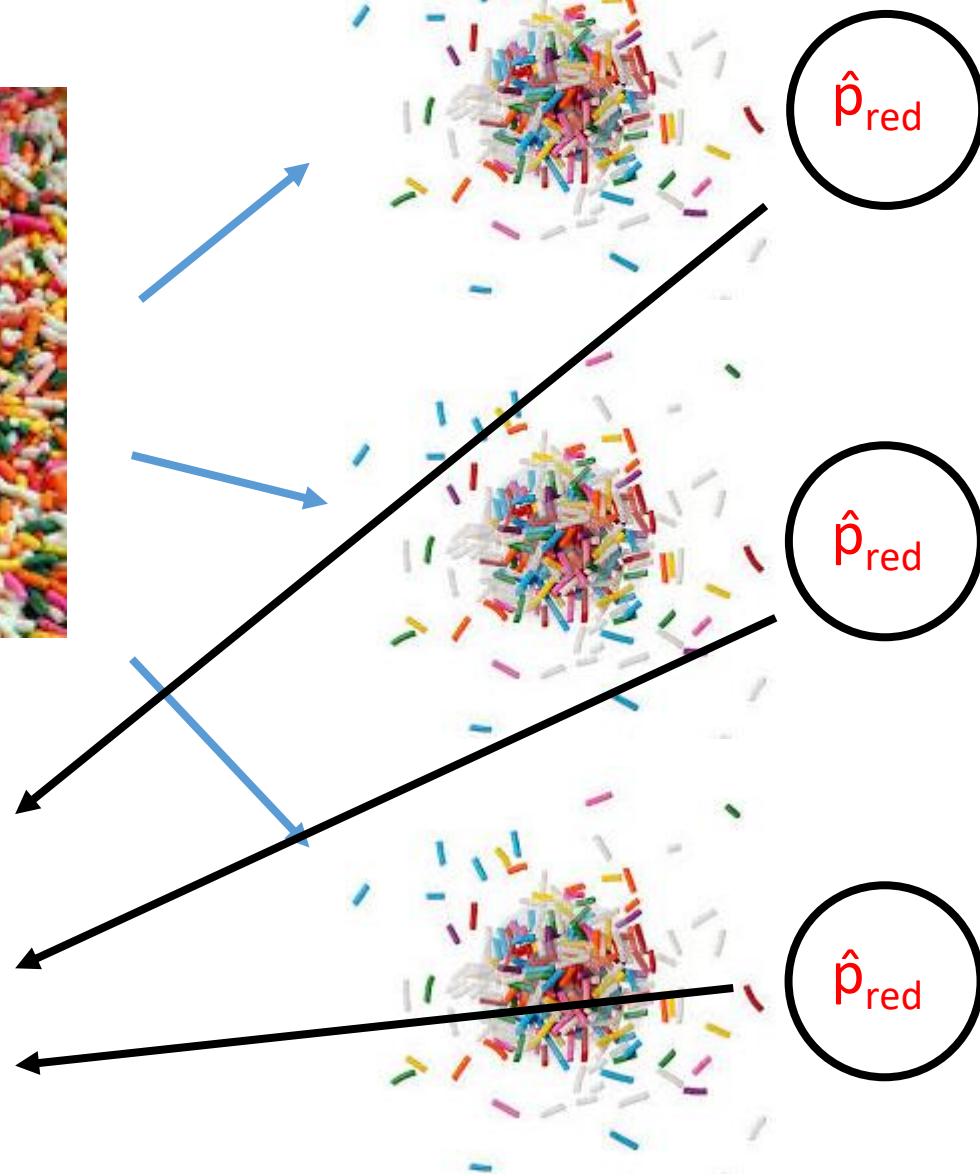
Reminder: For a ***single categorical variable***, the main statistic of interest is the ***proportion*** (\hat{p}) in each category

- (shadow of the parameter π)

$$\hat{p} = \text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$



Sampling distribution!



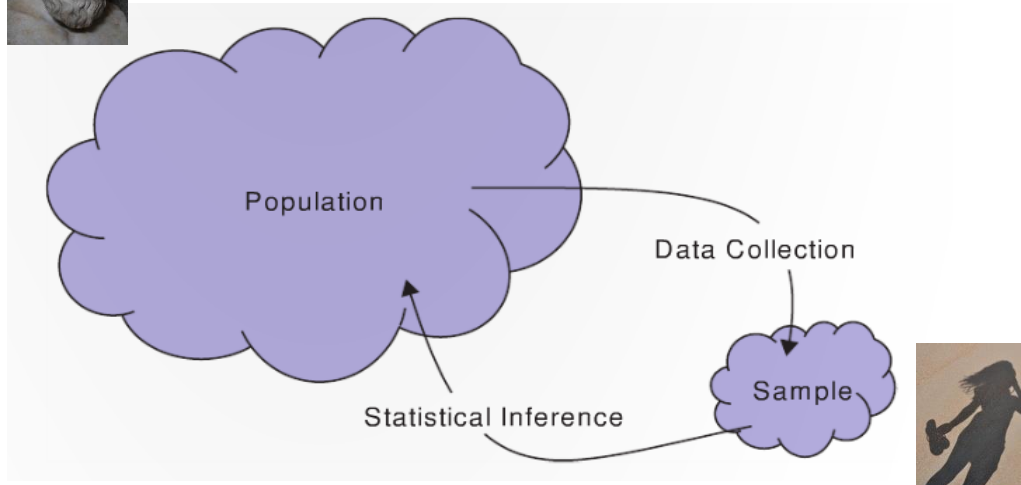
Sampling distribution

Why would we be interested in the sampling distribution?

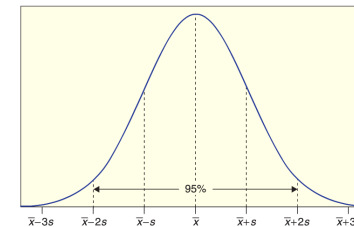
- If we knew what the sampling distribution was, then we could evaluate how much we should trust individual statistics



Parameters: π , μ , σ , ρ , β



Sampling distribution



Statistics: \hat{p} , \bar{x} , s , r , b

Sampling distributions

```
sampling_dist <- NULL
for (i in 1:1000) {
  rand_data <- runif(100)  # generate n = 100 points from U(0, 1)
  sampling_dist[i] <- mean(rand_data)  # save the mean
}

hist(sampling_dist)
```

Sampling distributions

Distribution of OkCupid user's heights $n = 100$

```
heights <- profiles$height
```

```
# get one random sample of heights from 100 people
```

```
height_sample <- sample(heights, 100)
```

```
# get the mean of this sample
```

```
mean(height_sample)
```

Sampling distributions

Distribution of OkCupid user's heights $n = 100$

```
sampling_dist <- NULL
for (i in 1:1000) {
    height_sample <- sample(heights, 100)  # sample 100 random heights
    sampling_dist[i] <- mean(height_sample) # save the mean
}

hist(sampling_dist)
```


For loops

For loops are useful when you want to repeat a piece of code many times under similar conditions

The syntax for a for loop is:

```
for (i in 1:100) {
```

```
    # do something
```

```
}
```



This is repeated 100 times
i is incremented by 1 each time

For loops

For loops are useful when you want to repeat a piece of code many times under similar conditions

The syntax for a for loop is:

```
for (i in 1:100) {  
    print(i)  
}
```



This is repeated 100 times
i is incremented by 1 each time

For loops

For loops are particular useful in combination with vectors that can store the results

```
my_results <- NULL    # create an empty vector to store the results
for (i in 1:100) {
    my_results[i] <- i^2
}
```

Sometimes there are more efficient ways to do the same thing without for loops

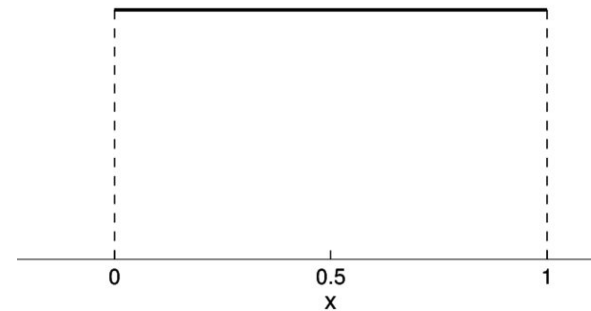
```
> (1:100)^2
```


Generating random data

R has built in functions to generate data from different distributions

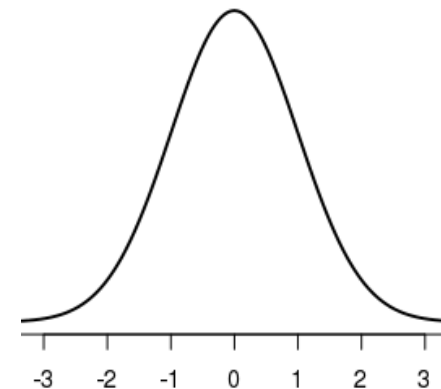
The uniform distribution:

```
# generate n = 100 points from U(0, 1)  
rand_data <- runif(100)  
hist(rand_data)
```



The normal distribution

```
# generate n = 100 points from N(0, 1)  
rand_data <- rnorm(100)  
hist(rand_data)
```



Generating random data

If we want the same sequence of random numbers we can set the random number generating seed

```
> set.seed(123)
```

```
> runif(100)
```

Q: Why would we want the same sequence of random number?

A: Reproducibility!

Sample statistics

Q: What is a statistic?

A: A statistic is number computed from a function on a sample of data

The sample mean \bar{x} (shadow of the parameter μ)

```
rand_data <- runif(100) # generate n = 100 points from U(0, 1)
mean(rand_data)
```

Q: If we repeat the code above will we get the same statistic?

- A: unlikely

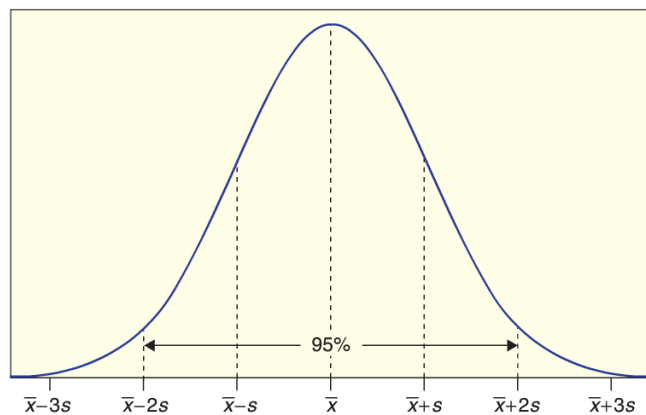
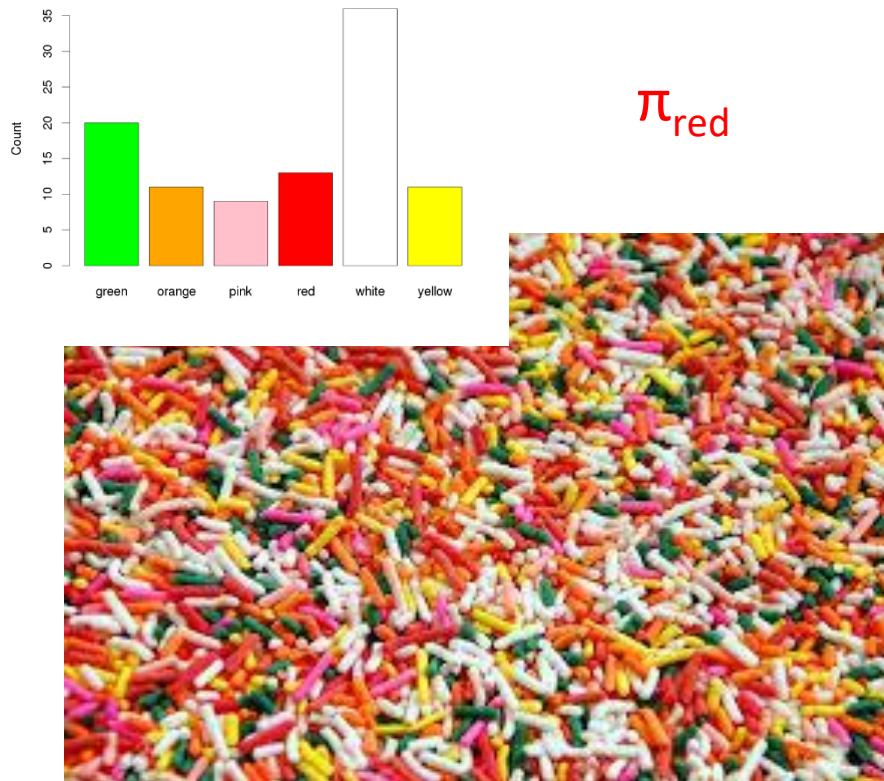
Sampling distributions

A distribution of ***statistics*** is called a ***sampling distribution***

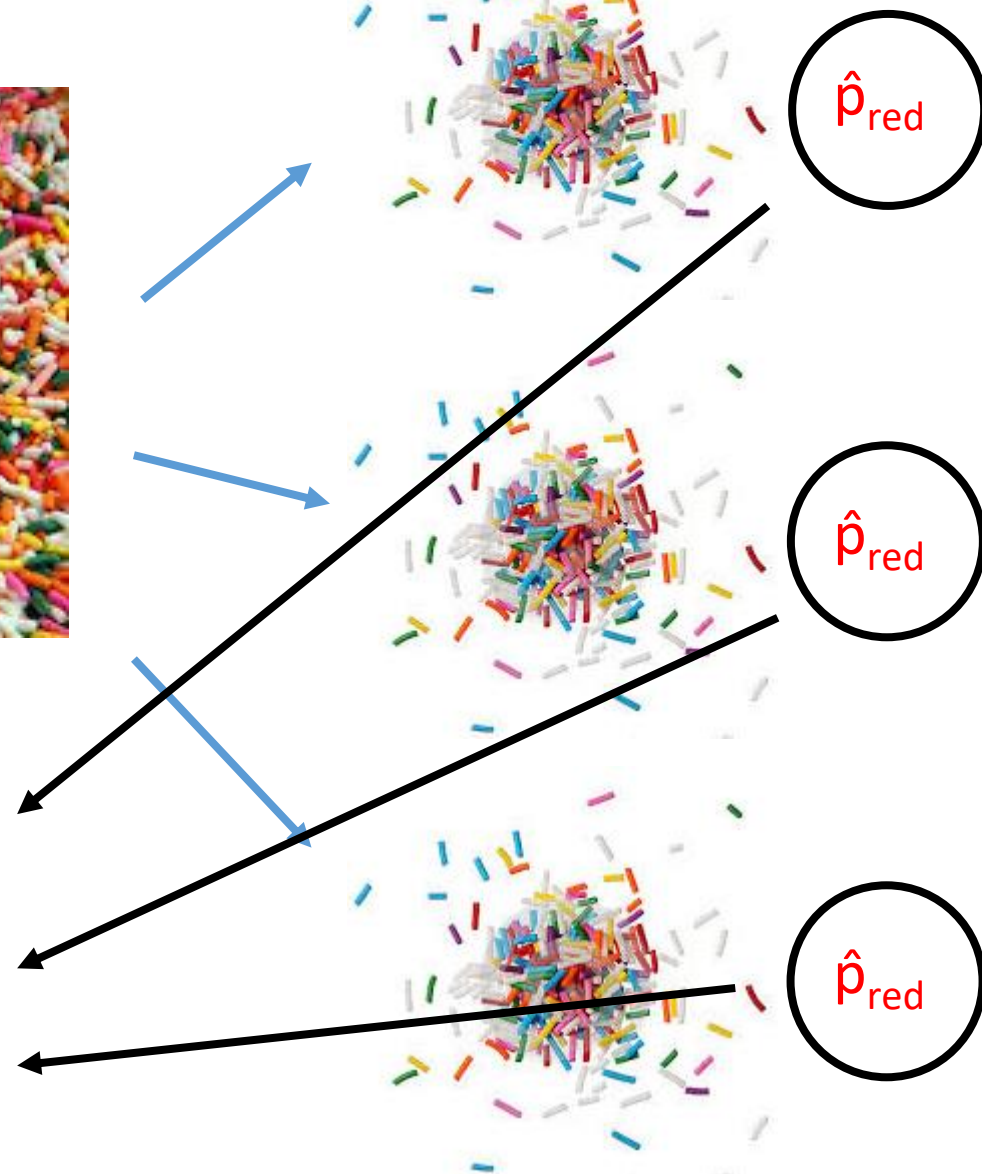
Reminder: For a ***single categorical variable***, the main statistic of interest is the ***proportion*** (\hat{p}) in each category

- (shadow of the parameter π)

$$\hat{p} = \text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$



Sampling distribution!



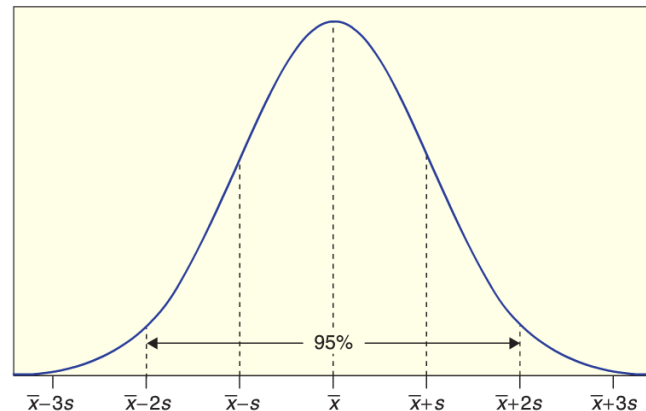
Sampling distribution

Q1: Would we ever calculate the sampling distribution in practice?

- A: No, we would never repeat an experiment that many times

Q2: Why would we be interested in the sampling distribution?

- If we knew what the sampling distribution was, then we could evaluate how much we should trust individual statistics



Sampling distribution

Sampling distributions

```
sampling_dist <- NULL
for (i in 1:1000) {
  rand_data <- runif(100)  # generate n = 100 points from U(0, 1)
  sampling_dist[i] <- mean(rand_data)  # save the mean
}

hist(sampling_dist)
```

Sampling distributions

Distribution of OkCupid user's heights $n = 100$

```
heights <- profiles$height
```

```
# get one random sample of heights from 100 people
```

```
height_sample <- sample(heights, 100)
```

```
# get the mean of this sample
```

```
mean(height_sample)
```

Sampling distributions

Distribution of OkCupid user's heights $n = 100$

```
sampling_dist <- NULL
for (i in 1:1000) {
    height_sample <- sample(heights, 100)  # sample 100 random heights
    sampling_dist[i] <- mean(height_sample) # save the mean
}

hist(sampling_dist)
```

Homework 1

Homework 1 due on Sunday September 8th at 11:59pm

TA office hours are on a [google calendar on Canvas](#)

Data frames

Let's get a data frame with some OkCupid data!

- > `install.packages("okcupiddata")` # only needs to be run once
- > `library(okcupiddata)`
- > `View(profiles)` # the `View()` function only works in R Studio!

	age	body_type	diet	drinks	drugs	education
1	22	a little extra	strictly anything	socially	never	working on college/university
2	35	average	mostly other	often	sometimes	working on space camp
3	38	thin	anything	socially	NA	graduated from masters program
4	23	thin	vegetarian	socially	NA	working on college/university
5	29	athletic	NA	socially	never	graduated from college/university
6	29	average	mostly anything	socially	NA	graduated from college/university

Data frames

When data is loaded from a package it isn't visible in the environment pane. We can make it visible using the `data()` function.

```
> library(okcupiddata)
```

```
> data(profiles)
```

We can extract the columns of a data frame as vector objects using the `$` symbol

```
> the_ages <- profiles$age
```

Data frames

Survey question 2: What is the mean age of OkCupid users in the profiles data frame?

```
> mean(the_ages)
```

Extracting rows from a data frame

We can extract rows from a data frame in a similar way as extracting values from a vector by using the square brackets

```
> profiles[1, ] # returns the first row of the data frame
```

```
> profiles[, 1] # returns the first column of the data
```

Note, the first column of the profiles data frame is the variable *age*, so we can also get the first column using:

```
> profiles$age # this is the same as profiles[, 1]
```

Extracting rows from a data frame

We can also create vectors of numbers or booleans specifying which rows we want to extract from a data frame

```
# create a vector with the numbers 1, 10, 20
```

```
> my_vec <- c(1, 10, 20)
```

```
# use my_vec to get the 1st, 10th, and 20th row in profiles
```

```
> small_profiles <- profiles[my_vec, ]
```

```
> dim(small_profiles) # number of rows and columns in the data frame
```

Extracting rows from a data frame

Finally, we can also extract rows by creating a boolean vector that is of the same length as the number of rows in the data frame

TRUE values will be extracted from the data frame while FALSE values will not

```
# create a vector of booleans
```

```
> my_bools <- c(TRUE, FALSE, TRUE)
```

```
# use the Boolean vector to get the 1st and 3rd row
```

```
> small_profiles[my_bools, ]
```


Questions will be answered in class or on Piazza!



Categorical data

Categorical variables

What is a categorical variable?

- A: A categorical variable assigns each observation to one of k groups

Which variables in the profiles data frame are categorical?

- **Survey question 3**: Is height a categorical variable in the OkCupid data frame?

For categorical variables, we usually want to view:

- How many items are each category OR
- The proportion (or percentage) of items in each category

$$\text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$

Categorical data

```
# Get information about drinking behavior
```

```
> drinking_vec <- profiles$drinks
```

```
# Create a table showing how often people drink
```

```
> drinks_table <- table(drinking_vec)
```

```
> drinks_table
```

Relative frequency table

We can create a relative frequency table using the function:

```
> prop.table(my_table)
```

Can you create a relative frequency table for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
```

```
> prop.table(drinks_table)
```

Survey question 4: What is the proper statistical notation for these values: \hat{p} or π ?

Bar plots

(pun intended?)

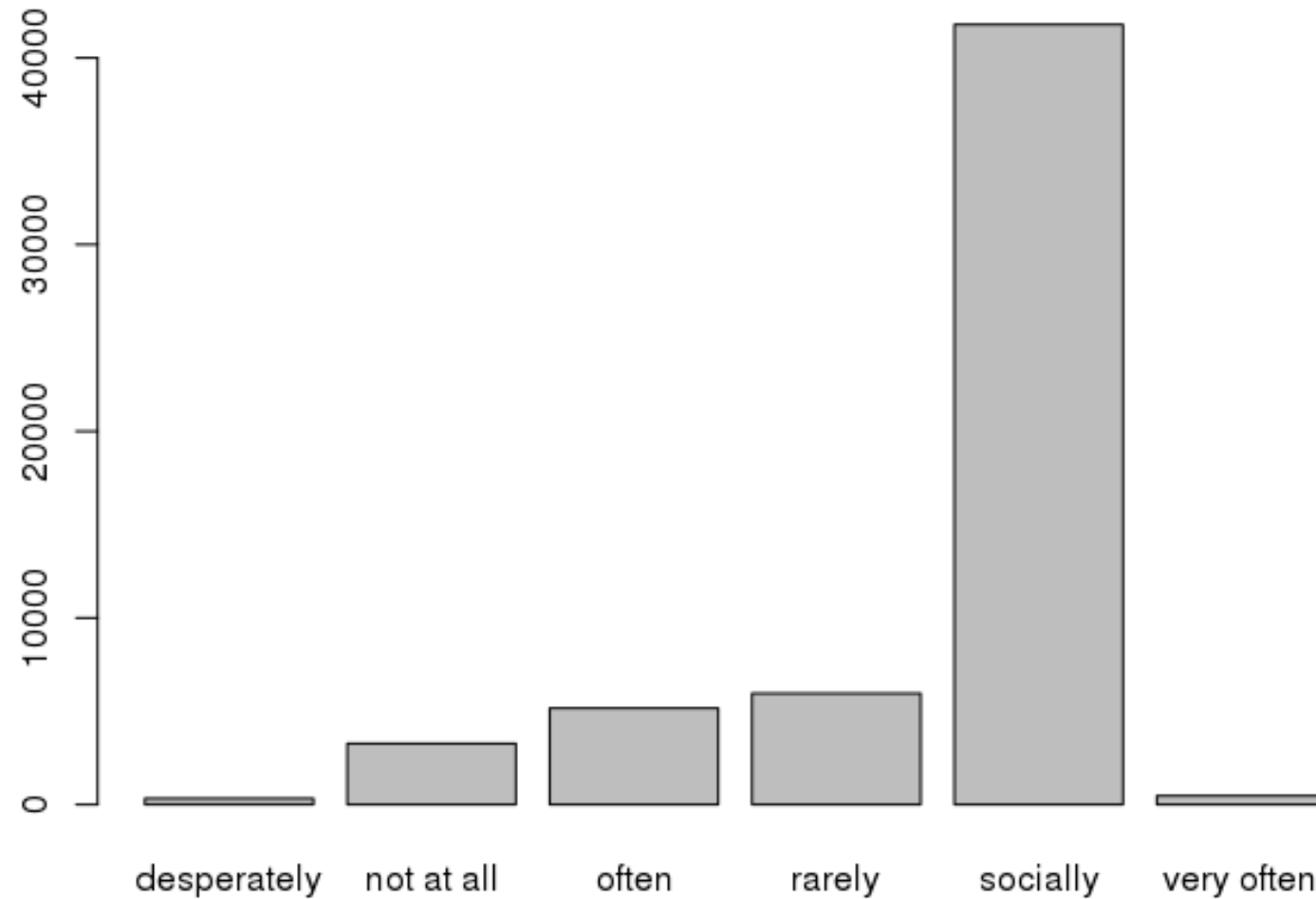
We can plot the number of items in each category using a bar plot

```
> barplot(my_table)
```

Can you create a bar plot for the drinking behavior of the people in the okcupid data set?

```
> drinks_table <- table(profiles$drinks)
```

```
> barplot(drinks_table)
```

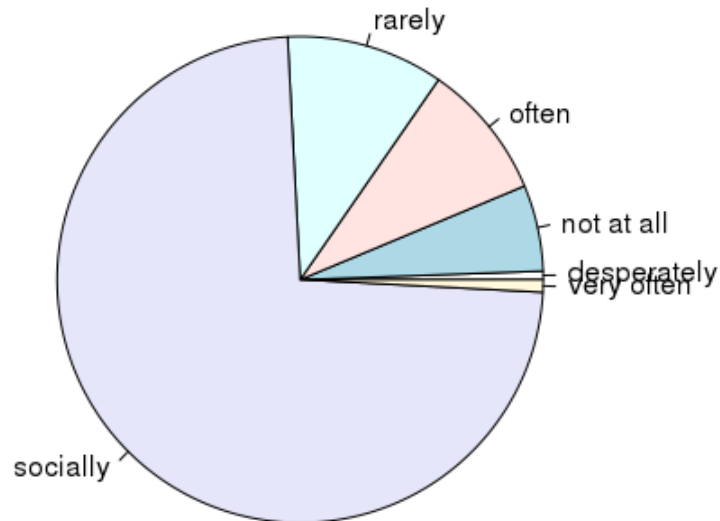


Survey question 5: What is missing on this plot and can you figure out how to add it?

Pie charts

We can also use the `pie()` function to create pie charts

> `pie(drinks_table)`



World's Most Accurate Pie Chart



Survey question 6: Which is best: bar plots or pie charts?

```
> barplot(table(profiles$sex, useNA = "always"))
```

```
> pie(table(profiles$sex, useNA = "always"))
```

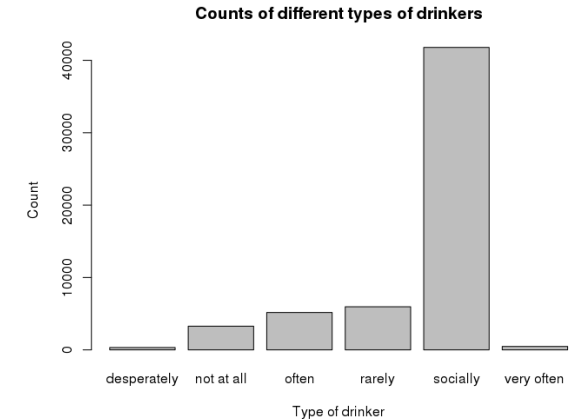
**BE
BEST**

Is one better than the other?

Can you figure out how to add colors to these plots?

Removing social drinkers

Social drinkers are dominating our plot 😞



We can get rid of social drinkers by only plotting counts less than 10,000

```
> nonsocial_inds <- drinks_table < 10000  
> nonsocial_drinks_table <- drinks_table[nonsocial_inds]  
> barplot(nonsocial_drinks_table)
```

It's a Match!



You and Booze have liked each other.

Questions will be answered in class or on Piazza!



Quantitative data: statistics

There are several statistics that describe the central tendency of quantitative data

- The mean: `mean()`
- The median: `median()`

Survey question 7 : Which of these measures is robust to outliers?

Can you calculate the mean and median of OkCupid user's heights?

What went wrong?

`mean(v, na.rm = TRUE)`

Survey question 8 What is the proper statistical notation for the mean of OkCupid user's heights: \bar{x} or μ ?

Quantitative data: Visualizing heights

Q: How can we visualize the heights in the profiles data frame?

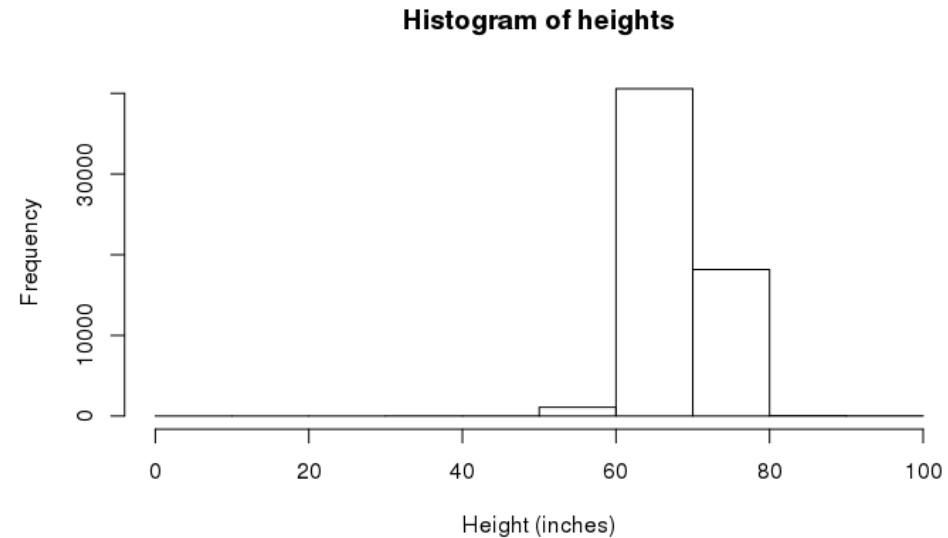
A: Histograms!

A: Box plots

A: Many other options too

Histograms of heights

Height (inches)	Frequency Count
(0-10]	6
(10-20]	0
(20-30]	1
(30-40]	13
(40-50]	9
(50-60]	1097
(60-70]	40575
(70-80]	18164
(80-90]	50
>90	28



Visualizing heights

We can create histograms in R using the `hist()` function

Can you create a histogram of heights?

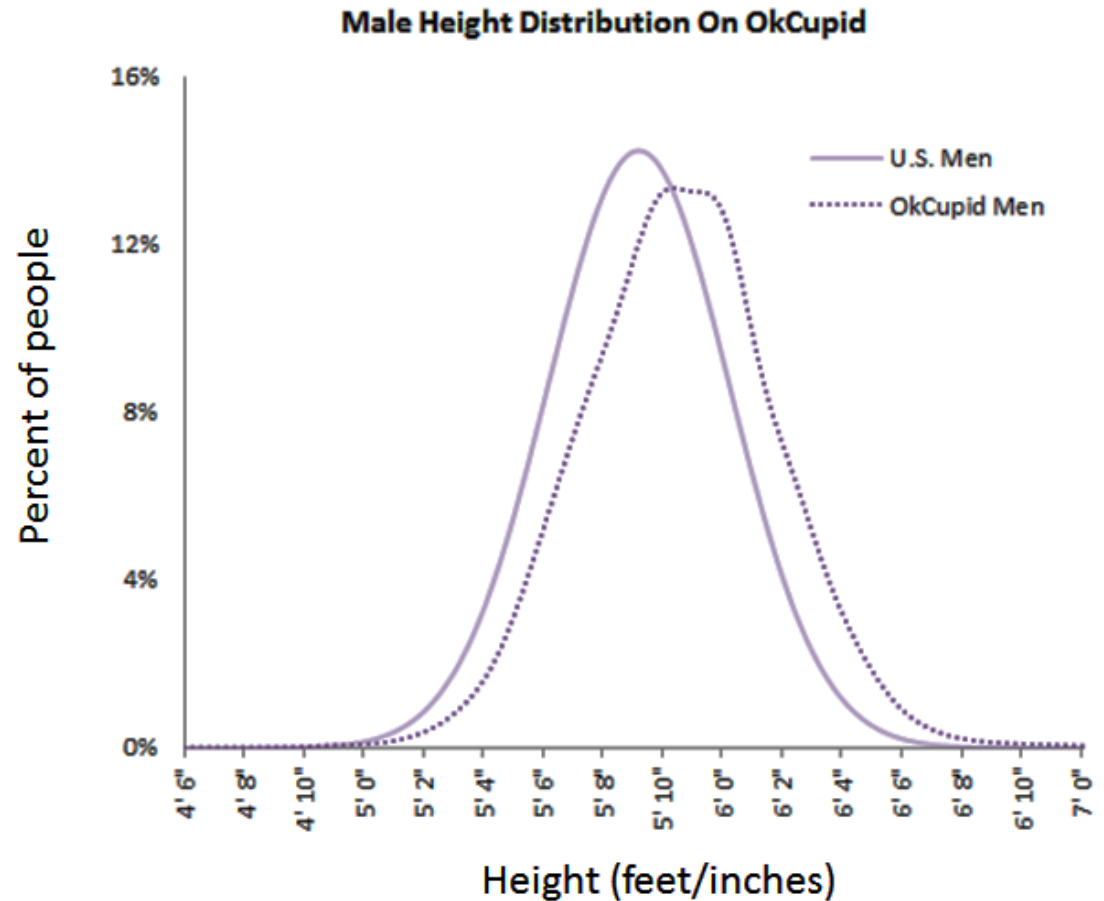
```
> hist(profiles$height)
```

```
> hist(profiles$height, nclass = 50)
```

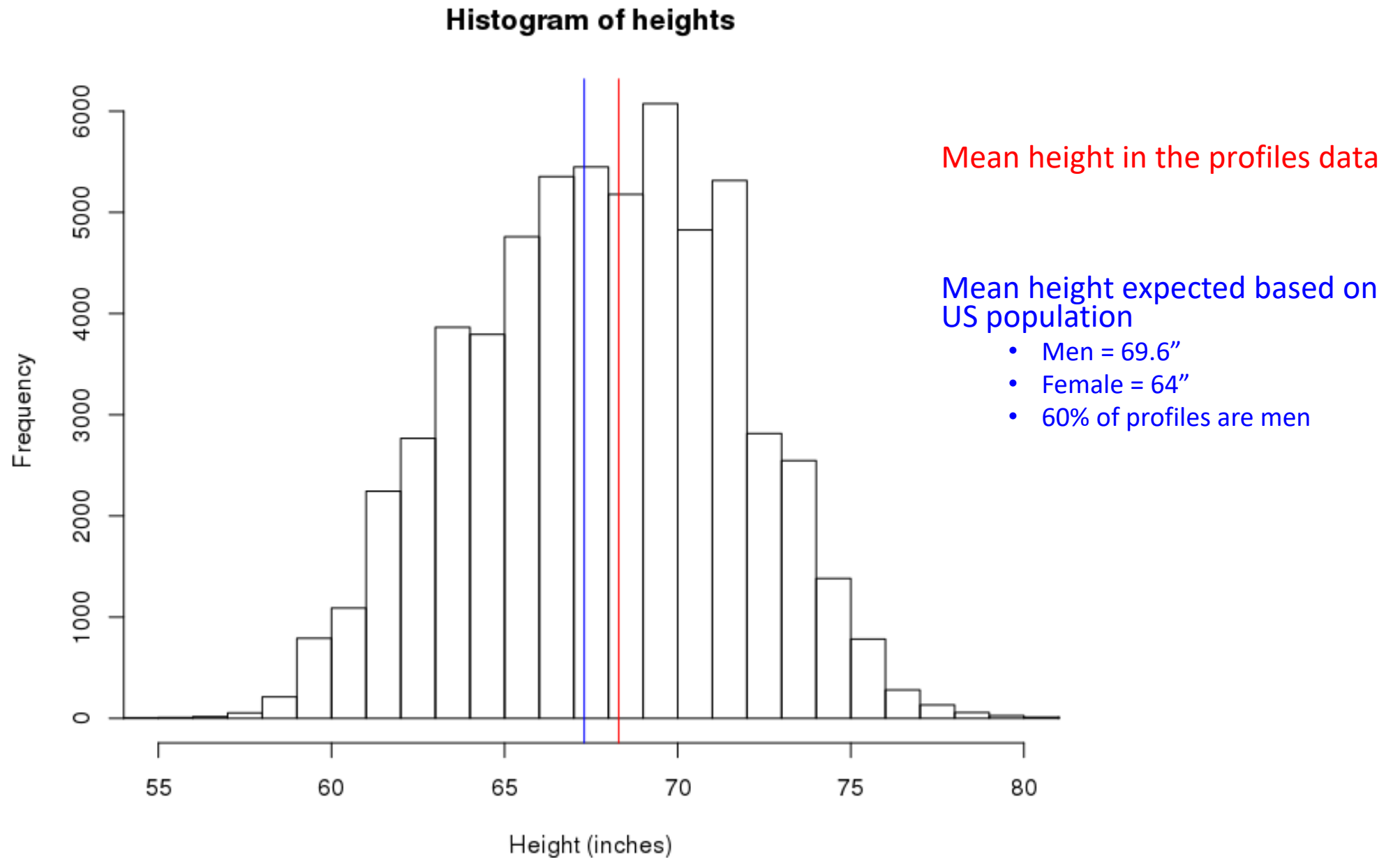
OkCupid users are taller than the average person

Survey question 9 Please read the article: "The Big Lies People Tell In Online Dating"

Describe one thing you found interesting in the article and be prepared to discuss it in class on Tuesday

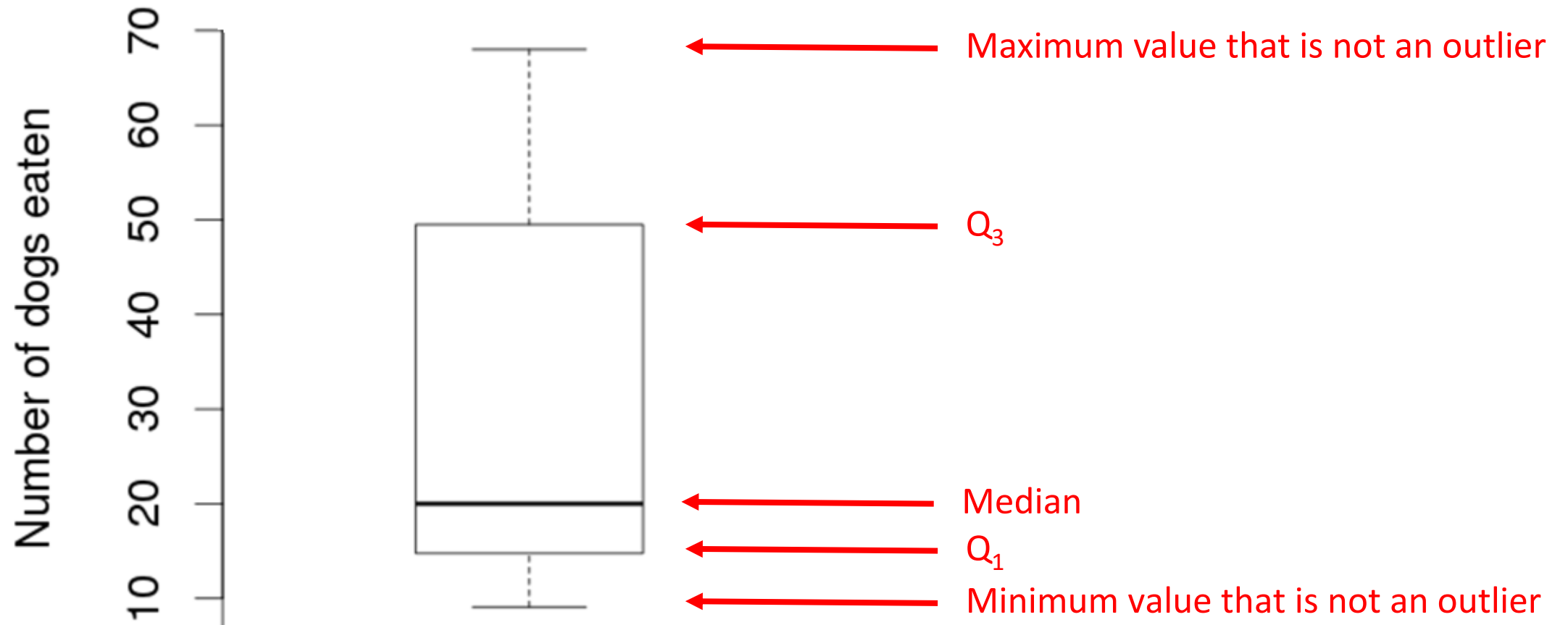


Can we see this in the profiles data?



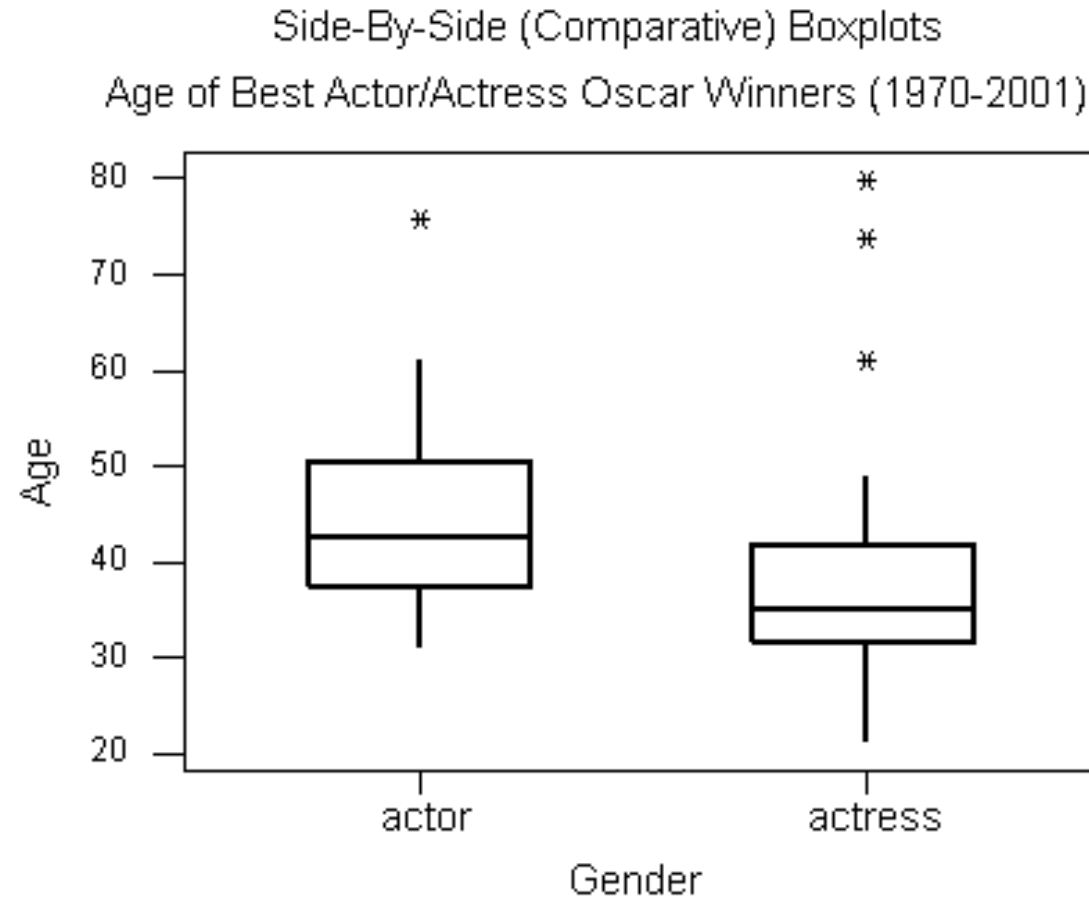
`abline(v = 65)` adds a vertical line to a plot at the value of 65

Box plots can also visualize quantitative data



R: `boxplot(v)`

Side-by-side boxplots



Useful for comparing distributions!

- What does the figure above show?

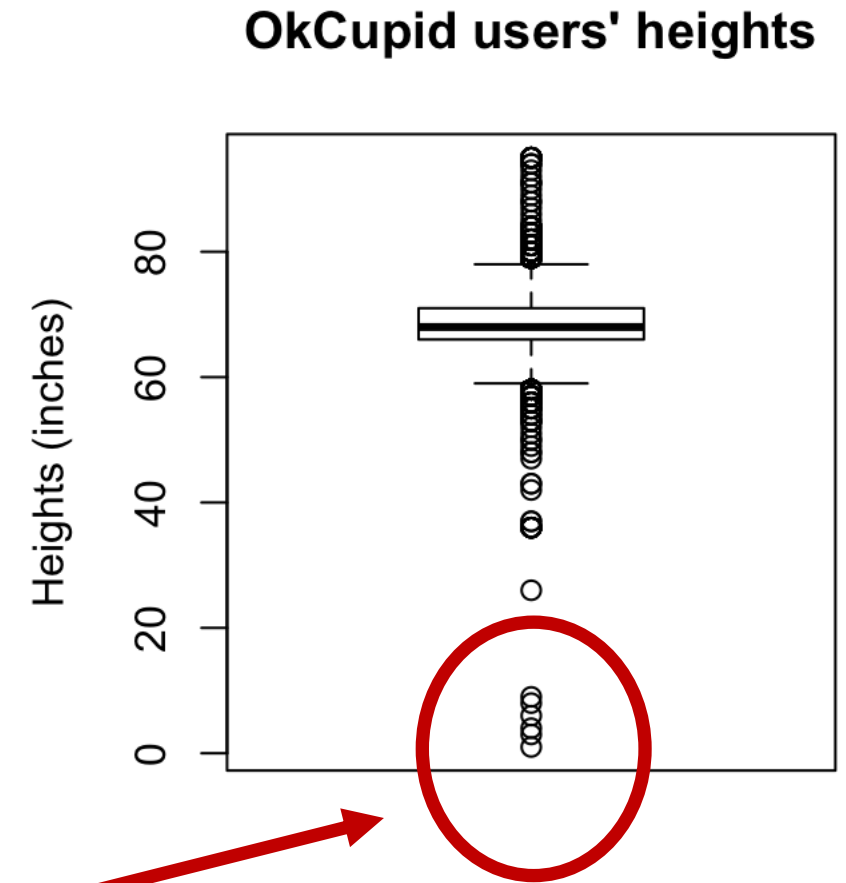
Outliers

Outliers on boxplots are values that are more than $1.5 * IQR$

What should we do if we have outliers?

Investigate!

- If there are due to an error, remove them



People under 20" tall?

Outliers

Outliers on boxplots are values that are more than $1.5 * IQR$

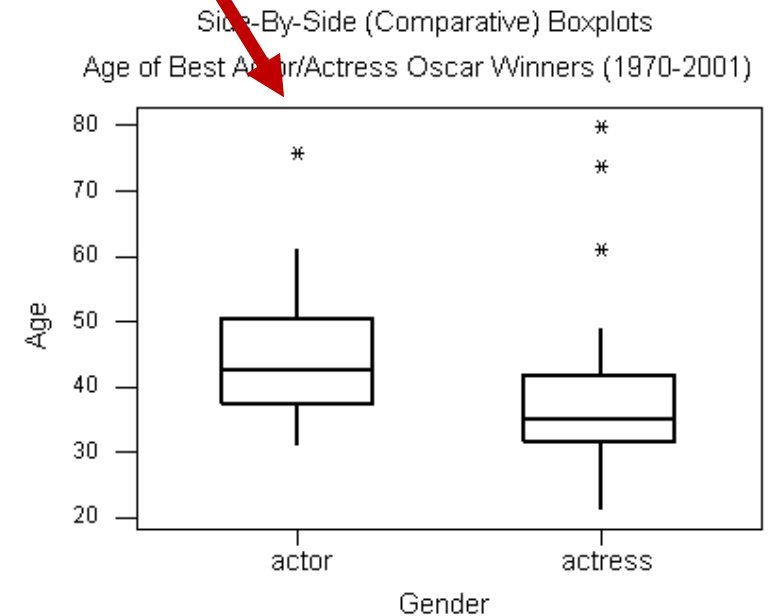
What should we do if we have outliers?

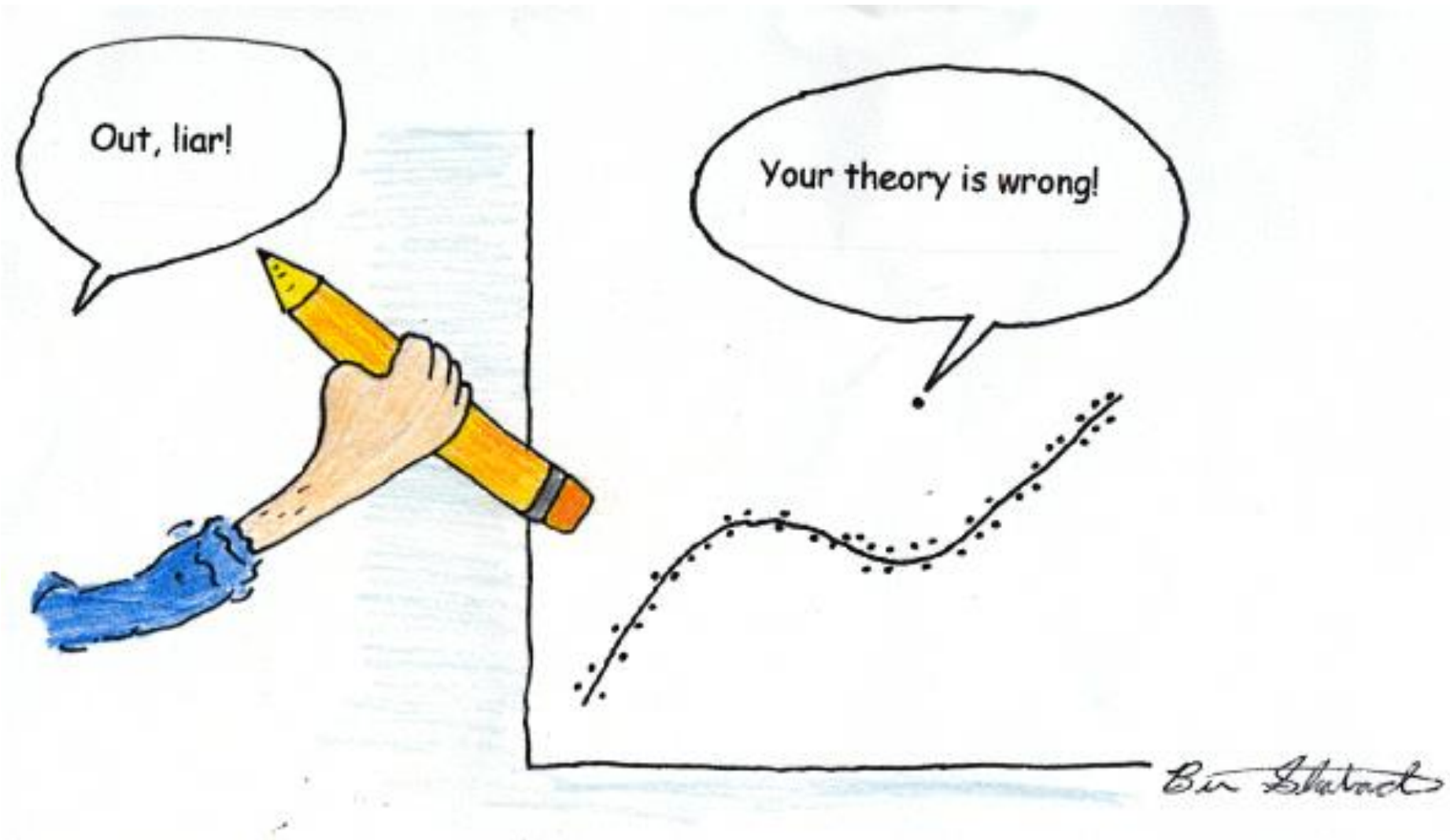
Investigate:

- If there are due to an error, remove them
- **If not, need to account for them**

Survey question 10

Who is this actor?





Questions will be answered in class or on Piazza!



CitiBike data

Let's look at the bike share data from NYC

```
> load('daily_bike_totals.rda')
```



CitiBike analysis

What does each case correspond to?

We can use the `dim()` function to get how many cases and variables there are

- How many are there?

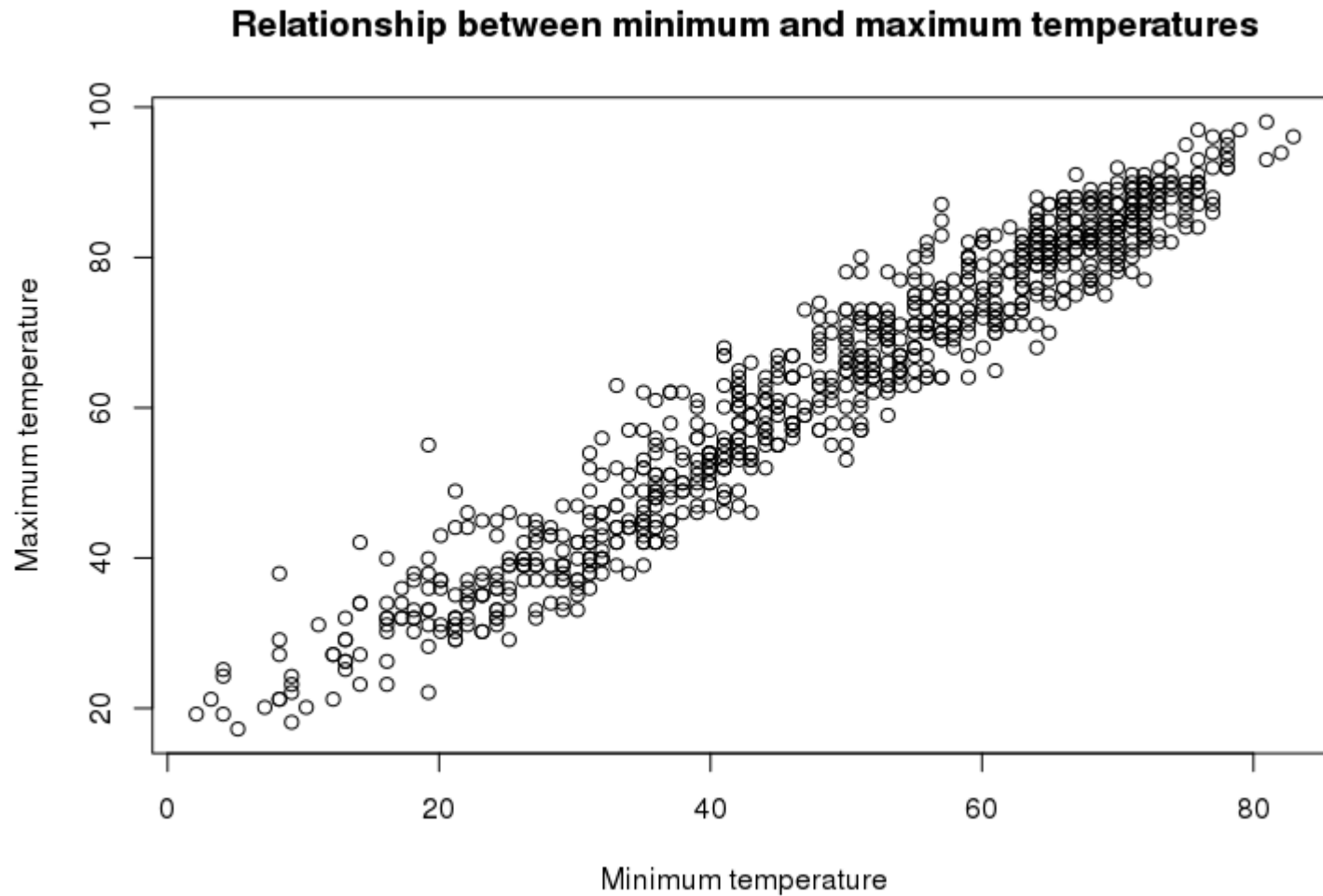
Scatter plots

We can use the `plot(x, y)` function to create scatter plots

Can you create a scatter plot of the relationship between the minimum and maximum temperatures?

```
> plot(bike_daily_data$min_temperature,  
      bike_daily_data$max_temperature,  
      xlab = "Minimum temperature",  
      ylab = "Maximum temperature",  
      main = "Relationship between min and temp")
```

Scatter plots



Plotting time series

We can use the `plot(x, y)` function to plot time series

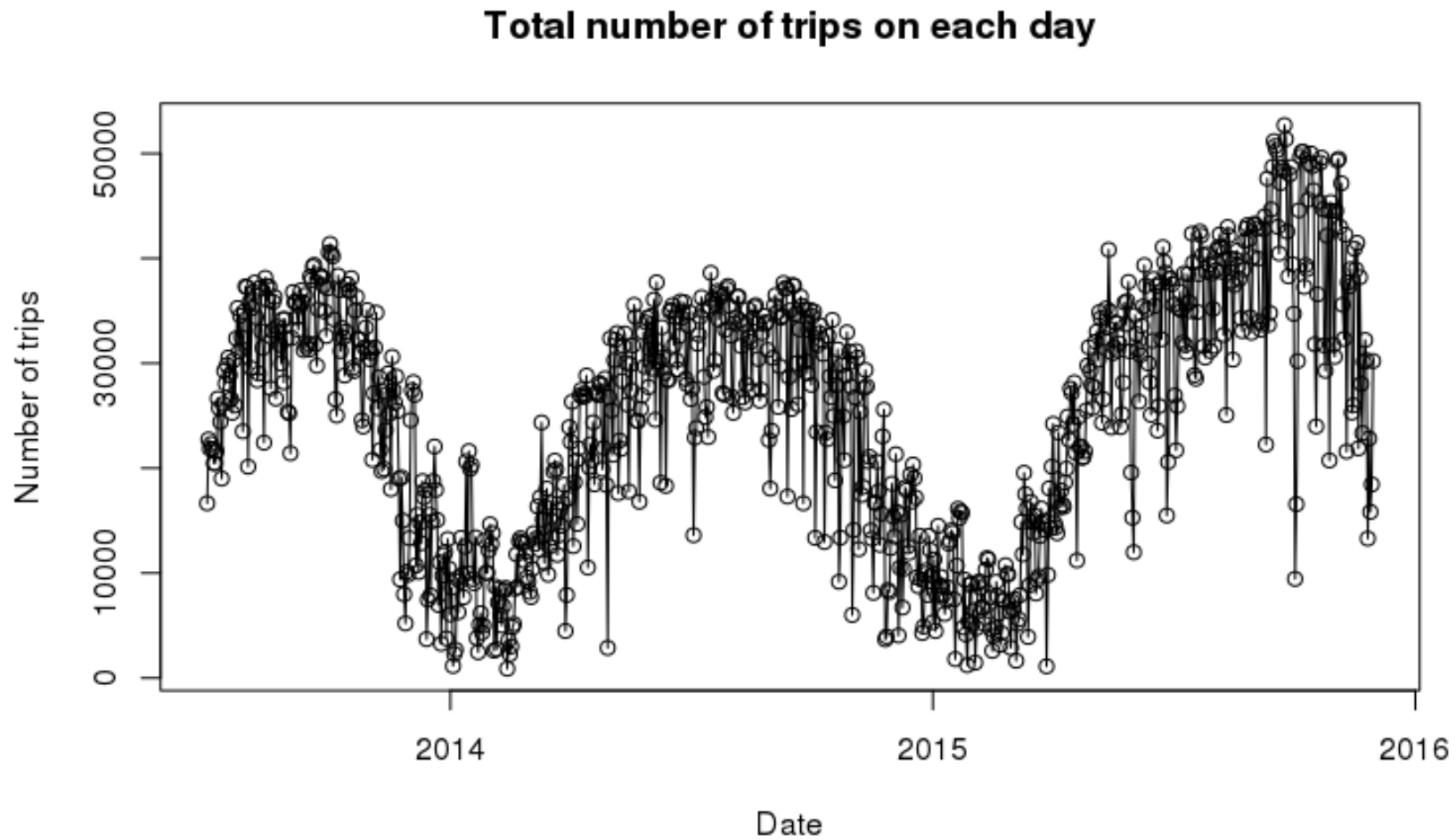
we can connect the points in a plot using

```
> plot(x, y, type = 'l') # connected points
```

```
> plot(x, y, type = 'o') # both points and dots
```

```
> plot(bike_daily_data$date, bike_daily_data$trips,  
       type = 'o',  
       xlab = "Date",  
       ylab = "Number of trips",  
       main = "Total number of trips on each day")
```


Plotting time series



Homework 1

Homework 1 has been posted

> `library(SDS230)`

> `download_homework(1)`

Due on Gradescope by 11:59pm on Sunday September 13th

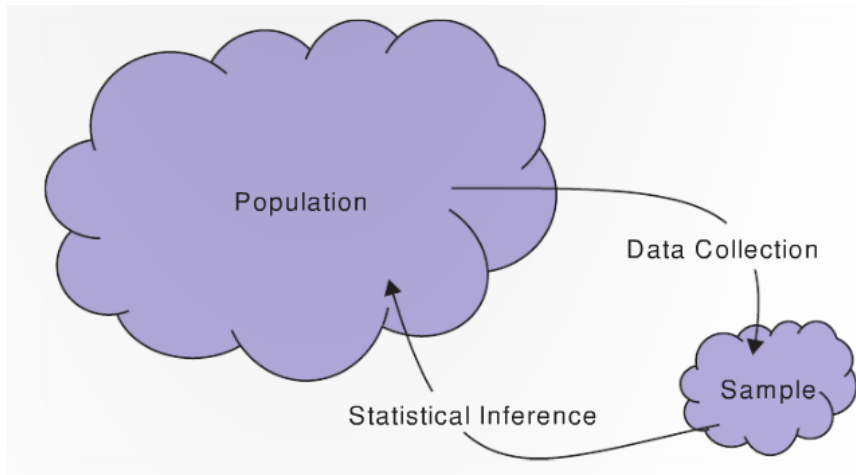
- Instructions for how to submit homework on Gradescope are on Canvas

Quick Review of central concepts in Intro Statistics

Concepts from Intro Stats

The Truth[®] is out there

- If we have infinite data we could compute parameters
- We can estimate parameters with statistics
 - statistics are functions of our data



	Sample Statistic	Population Parameter
Mean	\bar{x}	μ
Standard deviation	s	σ
Proportion	\hat{p}	π
Correlation	r	ρ
regression slope	b	β

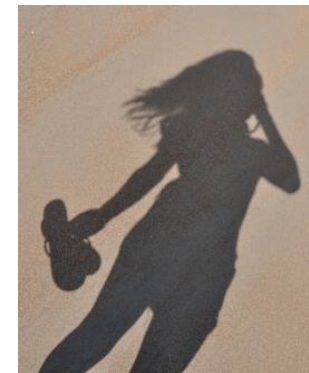
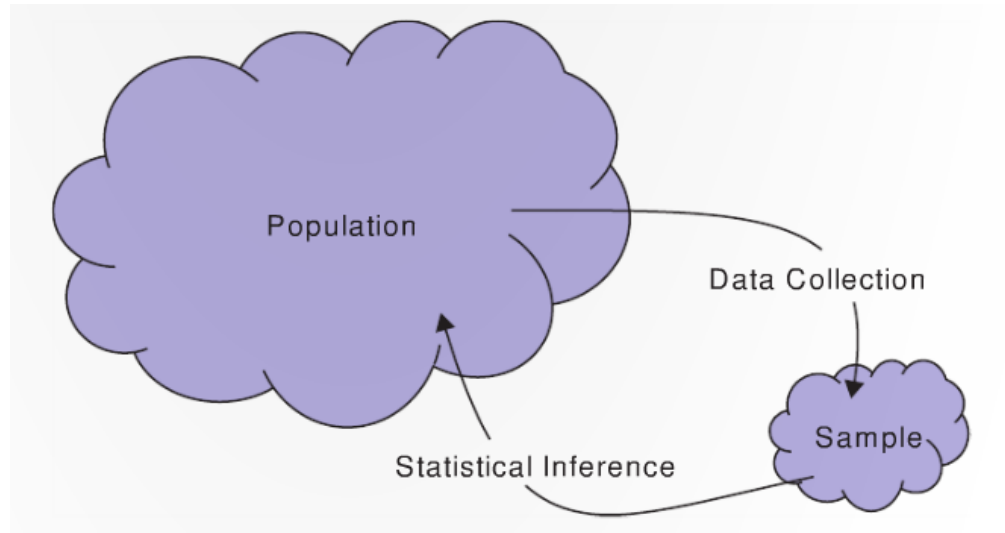
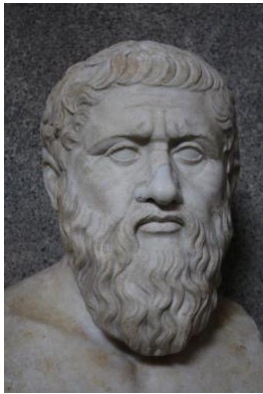
Descriptive and inferential statistics

Descriptive Statistics: describe the sample of data we have

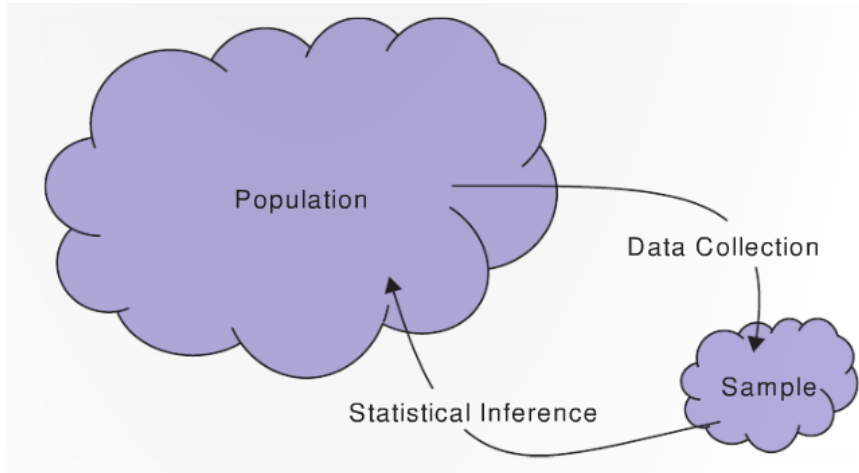
- i.e., describe the shadows

Inferential Statistics: use the sample to make claims about properties of the population/process

- i.e., try to use the data to get at the Truth



Sampling



Simple random sample: each member in the population is equally likely to be in the sample

- This is called *random selection*

Soup analogy!

Q: Why is this good?

A: Allows for generalizations to the population! (no bias)

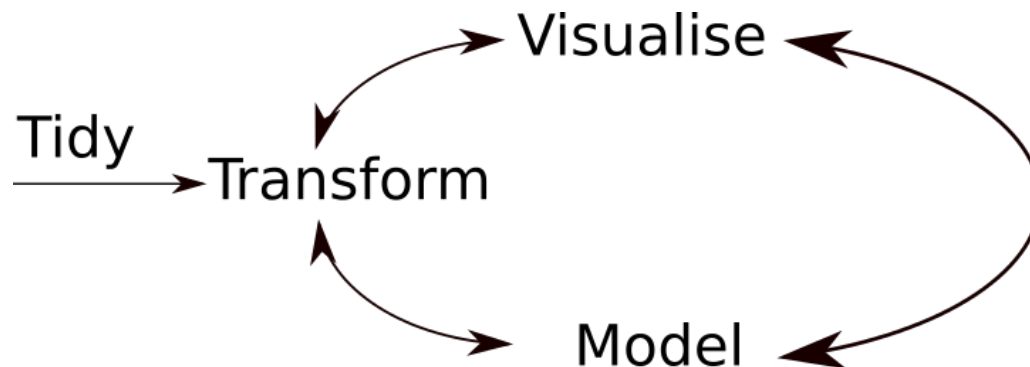


Our goal

Our goal is to try to uncover the Truth and convey the results to others

- More than one way to get at the Truth
 - no one 'right way' to analyze data
- Many ways to make mistakes too

Getting at the Truth can be an iterative process



“All models are wrong
some models are useful”
- George Box

Questions?



Review of R from last class

Assignment:

```
> a <- 5
```

Data types:

```
> s <- "s is a terrible name for an object" # string
```

```
> b <- TRUE # boolean
```

Functions:

```
> sqrt(49)
```

```
> ? sqrt # help
```

Review of R from last class

Vectors:

```
> v <- c(TRUE, TRUE, FALSE)    # a vector of booleans
```

Accessing elements of a vector:

```
> v[3]
```

```
> v[c(2, 3)]
```


For loops

- Then we have all the material for both days, can split it up between days or not
- Can do sampling distribution stuff as backup if need more material
- Ok, not there yet for next week, but hopefully can wrap things up (lots of cleaning) tomorrow

- Sampling
- Sampling distributions
- Sampling and bias via simulations

Question



Q: What was the movie, 'Pirates of the Caribbean' rated?

A: PG-13

Q: Worst joke of the semester?

A: We are just getting started!

A warning about terminology



“Boy, those French: They have a different word for everything!”

- Steve Martin

Boy, those Statisticians: They use common words to mean something different!



Bias,
confidence,
significance

My thoughts

Statistics is a way to use data to answer questions:

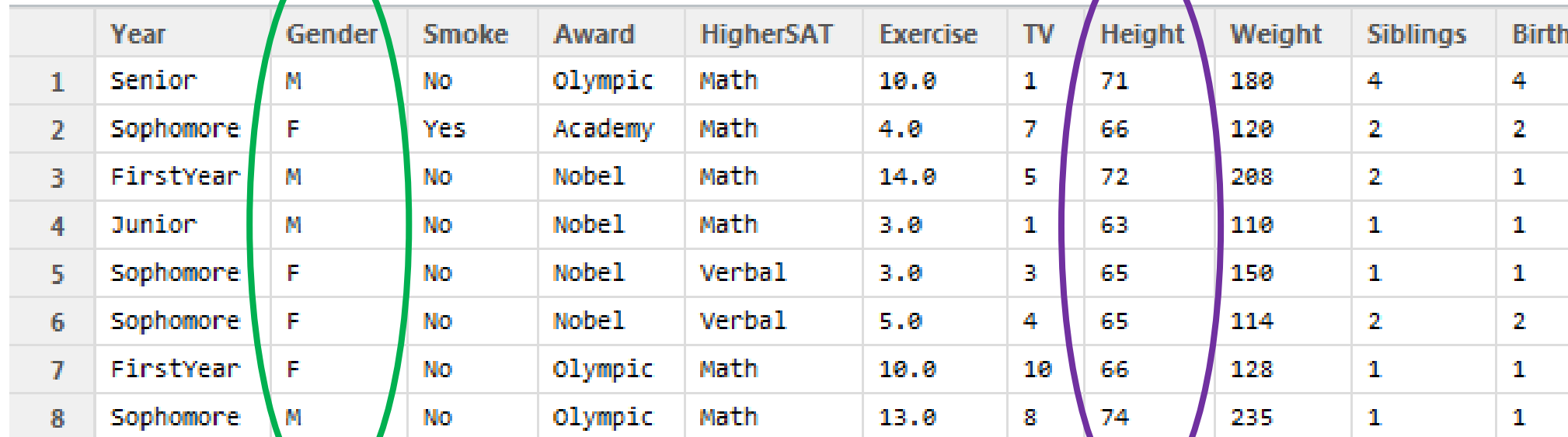
- Often we use a small amount of data to answer questions about a larger underlying phenomenon
- We want to know the truth, and not be fooled by randomness
 - Quantify uncertainty and randomness

It's part of an argument

- Don't blindly trust statistical tests, think about the results!
 - Do you really believe them?
- Be your own worst critic and try to prove yourself wrong

Explanatory and Response Variables

Sometimes we use one variable (**the explanatory variable**) to understand/predict another variable (**the response variable**)



	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings	Birth
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2	1
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1	1

Your turn to examine datasets

What are the observational units (cases)?

Which variables are: quantitative or categorical?

Find examples of possible explanatory and response variables

Student Survey

What are the observational units (cases)?

Which variables are: quantitative or categorical?

Find examples of possible explanatory and response variables

	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings	BirthOrder
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2	1
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1	1

FloridaLakes

This dataset describes characteristics of water and fish samples from 53 Florida lakes. Some variables (e.g. Alkalinity, pH, and Calcium) reflect the chemistry of the water samples. Mercury levels were recorded for a sample of large mouth bass selected at each lake.

Source: Lange, Royals, and Connor, Transactions of the American Fisheries Society (1993)

<i>ID</i>	An identifying number for each lake
<i>Lake</i>	Name of the lake
<i>Alkalinity</i>	Concentration of calcium carbonate (in mg/L)
<i>pH</i>	Acidity
<i>Calcium</i>	Amount of calcium in water
<i>Chlorophyll</i>	Amount of chlorophyll in water
<i>AvgMercury</i>	Average mercury level for a sample of fish (large mouth bass) from each lake
<i>NumSamples</i>	Number of fish sampled at each lake
<i>MinMercury</i>	Minimum mercury level in a sampled fish
<i>MaxMercury</i>	Maximum mercury level in a sampled fish
<i>ThreeYrStdMercury</i>	Adjusted mercury level to account for the age of the fish
<i>AgeData</i>	Mean age of fish in each sample

ICUAdmissions

Data from a sample of 200 patients following admission to an adult intensive care unit (ICU).

Source: DASL dataset downloaded from <http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html>

<i>ID</i>	Patient ID number
<i>Status</i>	Patient status: 0 = lived or 1 = died
<i>Age</i>	Patient's age (in years)
<i>Sex</i>	0 = male or 1 = female
<i>Race</i>	Patient's race: 1 = white, 2 = black, or 3 = other
<i>Service</i>	Type of service: 0 = medical or 1 = surgical
<i>Cancer</i>	Is cancer involved? 0 = no or 1 = yes
<i>Renal</i>	Is chronic renal failure involved? 0 = no or 1 = yes
<i>Infection</i>	Is infection involved? 0 = no or 1 = yes
<i>CPR</i>	Patient gets CPR prior to admission? 0 = no or 1 = yes
<i>Systolic</i>	Systolic blood pressure (in mm Hg)
<i>HeartRate</i>	Pulse rate (beats per minute)
<i>Previous</i>	Previous admission to ICU within 6 months? 0 = no or 1 = yes
<i>Type</i>	Admission type: 0 = elective or 1 = emergency
<i>Fracture</i>	Fractured bone involved? 0 = no or 1 = yes
<i>PO2</i>	Partial oxygen level from blood gases under 60? 0 = no or 1 = yes
<i>PH</i>	pH from blood gas under 7.25? 0 = no or 1 = yes
<i>PCO2</i>	Partial carbon dioxide level from blood gas over 45? 0 = no or 1 = yes
<i>Bicarbonate</i>	Bicarbonate from blood gas under 18? 0 = no or 1 = yes
<i>Creatinine</i>	Creatinine from blood gas over 2.0? 0 = no or 1 = yes
<i>Consciousness</i>	Level: 0 = conscious, 1 = deep stupor, or 2 = coma

HappyPlanetIndex

Data for 143 countries from the Happy Planet Index Project, <http://www.happyplanetindex.org>, that works to quantify indicators of happiness, well-being, and ecological footprint at a country level. Region of the world is coded as: 1 = Latin America, 2 = Western nations, 3 = Middle East, 4 = Sub-Saharan Africa, 5 = South Asia, 6 = East Asia, 7 = former Communist countries.

Source: Downloaded from <http://www.happyplanetindex.org/data/>

<i>Country</i>	Name of country
<i>Region</i>	Code for region of the world, with code given in the description above.
<i>Happiness</i>	Score on a 0 to 10 scale for average level of happiness (10 is happiest)
<i>LifeExpectancy</i>	Average life expectancy (in years)
<i>Footprint</i>	Ecological footprint—a measure of the (per capita) ecological impact
<i>HLY</i>	Happy Life Years—combines life expectancy with well-being
<i>HPI</i>	Happy Planet Index (0–100 scale)
<i>HPIRank</i>	HPI rank for the country
<i>GDPperCapita</i>	Gross Domestic Product (per capita)
<i>HDI</i>	Human Development Index
<i>Population</i>	Population (in millions)

HollywoodMovies2011

Information for all 136 movies released from Hollywood in 2011.

Source: McCandless, D., “Most Profitable Hollywood Movies,” from “Information is Beautiful,” davidmccandless.com, accessed January 2012. The data were compiled late in 2011 so they reflect results as of December 2011.

<i>Movie</i>	Title of movie
<i>LeadStudio</i>	Studio that released the movie
<i>RottenTomatoes</i>	Rotten Tomatoes rating (reviewers)
<i>AudienceScore</i>	Audience rating (via Rotten Tomatoes)
<i>Story</i>	General theme—one of 21 themes
<i>Genre</i>	Action, Adventure, Animation, Comedy, Drama, Fantasy, Horror, Romance, or Thriller
<i>TheatersOpenWeek</i>	Number of screens for opening weekend
<i>BOAverageOpenWeek</i>	Average box office income per theater—opening weekend
<i>DomesticGross</i>	Gross income for domestic viewers (in millions)
<i>ForeignGross</i>	Gross income for foreign viewers (in millions)
<i>WorldGross</i>	Gross income for all viewers (in millions)
<i>Budget</i>	Production budget (in millions)
<i>Profitability</i>	WorldGross divided by Budget
<i>OpeningWeekend</i>	Opening weekend gross (in millions)

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains an R script with the following code:


```
1
2
3 # install.packages("Lahman")
4
5 require("Lahman")
6
7
8 # calculate average salary as a function of the year
9
10 player.salaries <- Salaries
11
12 mean(player.salaries[player.salaries$yearID == 2000, "salary"])
13
14 #mean(player.salaries[player.salaries$yearID %in% 1990:2000, "salary"]) # takes average over all years in range
15
16 mean(player.salaries[player.salaries$yearID == 2000, "salary"])
17
18 compute.ave.year.salary <- function(year){
19   d <- subset(player.salaries, yearID == year)
20   mean(d$salary, na.rm = TRUE)
21 }
22
23
24 ave.yearly.salary <- sapply(1985:2013, compute.ave.year.salary)
25 plot(1985:2013, ave.yearly.salary)
26
27
28 # cool, can use a function as an argument (weird syntax for second argument, but ok)
29 compute.stat.year.salary <- function(year, func){
30   d <- subset(player.salaries, yearID == year)
31   func(d$salary, na.rm = TRUE)
32 }
33
```
- Environment:** Shows the 'Global Environment' with a data object 'team.data.2014' containing 30 observations and 29 variables.
- Files:** A file explorer showing the project structure: 'Home > Baseball_Statistics > Class 01'. It lists files: 'class02.R', 'leagues_MLB_2001_teams_standard_batting.csv', and 'leagues_MLB_2014_teams_standard_batting.csv'.
- Console:** Displays the output of the R script execution, showing a data frame with columns for player ID, year, team, league, and various statistics. The output is truncated, showing only the first few rows.


```
> nrow(Batting)
Error: could not find function "nrow"
> nrow(Batting)
[1] 97889
> unique(Batting$YearID)
```

Class policies

Important:

- Turn assignments in on time
- Attend class
- Check your Hampshire email account

You can bring a laptop to class if you have one

Contact me if:

- You have special needs
- You are worried about your performance in the class

Class surveys

Class survey:

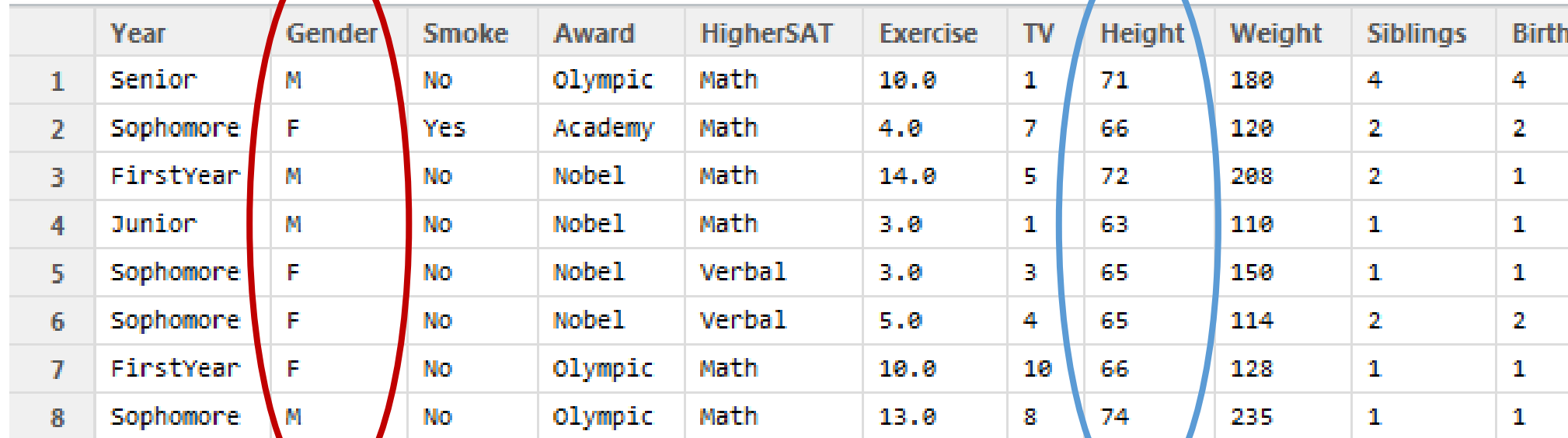
- URL: <http://goo.gl/gSwFk3>
- Will help me tailor the class to your interests

Categorical and Quantitative Variables

Cases (observational units)

Categorical Variable

Quantitative Variable



	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings	Birth
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2	1
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1	1

Practice problems for next class

1.4, 1.6, 1.10, 1.12, 1.24, 1.26

Turn your answers in to Moodle as in a pdf that was created from the RMarkdown document

```
> source("/home/shared/intro_stats_2016/cs206_functions.R")  
> get.worksheet(1)
```

Structured data – exploring the shadows



Practice problems for next class

Practice problems from Lock 5, first edition:

1.1, 1.3, 1.5, 1.11, 1.25, 1.26

Chapter is posted on Moodle

You can check the answers at the end of the book

Course objectives

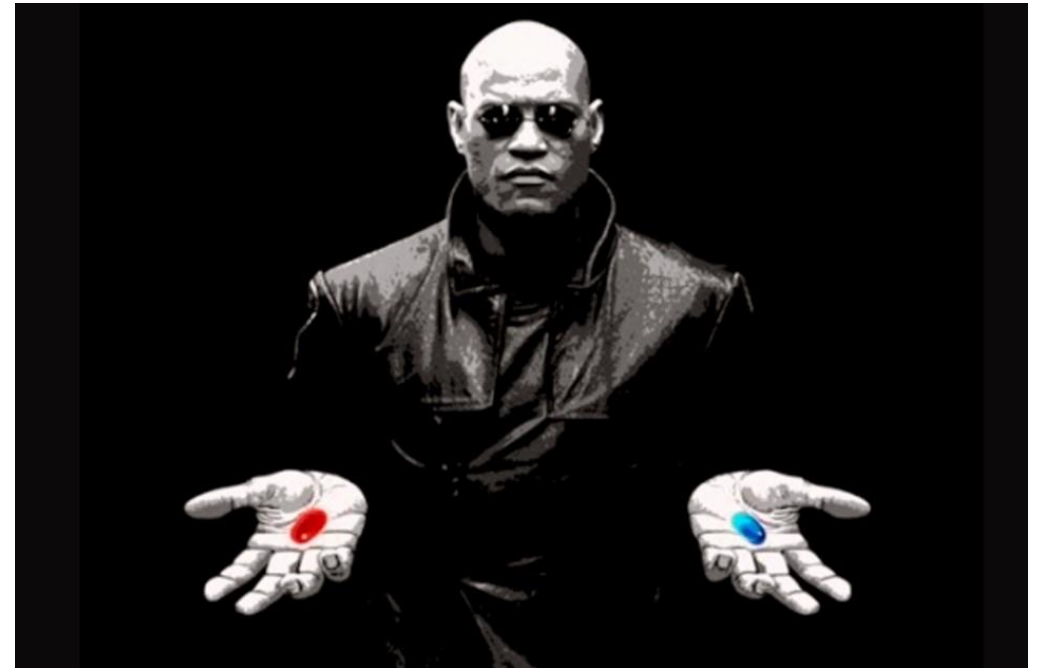
Website: <https://yale.instructure.com/courses/51220>

No required text, reading resources will be posted to canvas

Can you handle the Truth?

If not, perhaps you should not take this class

You've been warned...




Structured data – exploring the shadows



An Example Dataset (Shadows)

Variables

Cases



	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1

An Example Dataset (Shadows)

Categorical Variable

Quantitative Variable

Cases
(observational units)

	Year	Gender	Smoke	Award	HigherSAT	Exercise	TV	Height	Weight	Siblings
1	Senior	M	No	Olympic	Math	10.0	1	71	180	4
2	Sophomore	F	Yes	Academy	Math	4.0	7	66	120	2
3	FirstYear	M	No	Nobel	Math	14.0	5	72	208	2
4	Junior	M	No	Nobel	Math	3.0	1	63	110	1
5	Sophomore	F	No	Nobel	Verbal	3.0	3	65	150	1
6	Sophomore	F	No	Nobel	Verbal	5.0	4	65	114	2
7	FirstYear	F	No	Olympic	Math	10.0	10	66	128	1
8	Sophomore	M	No	Olympic	Math	13.0	8	74	235	1

Edmunds transaction data

Discuss!

- What are the observational units (cases)?
- Which variables are: quantitative or categorical?

	transactionid	date_sold	make_bought	price_bought	zip_bought	mileage_bought	color_bought
1	16966151	2014-09-27	Acura	30892.00	21043	40	BLACK
2	16914863	2014-09-27	Toyota	25566.00	15108	297	SILVER
3	15977620	2014-07-31	Nissan	34300.00	8753	0	JAVA
4	18666685	2015-01-27	Subaru	30059.00	7446	10	CRYSTAL WHITE PEARL
5	14383133	2014-04-27	Honda	32508.00	97027	21	MODERN STEEL
6	18196788	2014-12-18	Toyota	10819.66	95117	55246	WHITE
7	15722278	2014-07-24	Audi	59630.00	90401	143	GLACIER WHITE

Summary of concepts

1. Population: all individuals/objects of interest (truth)

2. Sample: A subset of the population (shadows)

3. Statistical inference: Making judgments about the population using data from the sample

4. Structured data has

- Cases/observational units: rows in a data set
- Variables: columns in a data set

5. Variables can be

- Categorical: fall into discrete categories
- Quantitative: are numbers

Course objectives

How to find the Truth® in a data set and convincingly convey the results to others

1. To extend **methods** and **concepts** learned in intro stats to more complex settings and **use computational/data science** approaches
 - Resampling methods: bootstrap CIs, and permutation tests
 - Multi-way ANOVAs
 - Multiple linear regression
 - Data visualization and wrangling
 - Statistical learning approaches: cross-validation
2. To learn how to analyze and visualize **real data sets** using **the R programming language**

- More R, plots, etc...
- Or go over R Markdown syntax???
- just in case there is more time...

Examples of questions we might look at...

Randomization tests: Are some animals psychic?



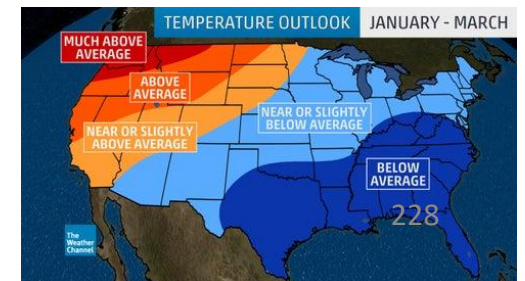
ANOVA: Are all genres of movies rated the same on average?



Data summarization: which airlines have the longest flight delays?



Data wrangling/visualization: How accurate are weather predictions?



Data frames

Data frames contain structured data

```
> car_data <- read.csv()
```

```
> View(car_data)      # the View() function only works in R Studio!
```

	transactionid	date_sold	make_bought	price_bought	zip_bought	mileage_bought	color_bought
1	16966151	2014-09-27	Acura	30892.00	21043	40	BLACK
2	16914863	2014-09-27	Toyota	25566.00	15108	297	SILVER
3	15977620	2014-07-31	Nissan	34300.00	8753	0	JAVA
4	18666685	2015-01-27	Subaru	30059.00	7446	10	CRYSTAL WHITE PEARL
5	14383133	2014-04-27	Honda	32508.00	97027	21	MODERN STEEL
6	18196788	2014-12-18	Toyota	10819.66	95117	55246	WHITE
7	15722278	2014-07-24	Audi	59630.00	90401	143	GLACIER WHITE

Data Frames

Variables

Cases



	transactionid	date_sold	make_bought	price_bought	zip_bought	mileage_bought	color_bought
1	16966151	2014-09-27	Acura	30892.00	21043	40	BLACK
2	16914863	2014-09-27	Toyota	25566.00	15108	297	SILVER
3	15977620	2014-07-31	Nissan	34300.00	8753	0	JAVA
4	18666685	2015-01-27	Subaru	30059.00	7446	10	CRYSTAL WHITE PEARL
5	14383133	2014-04-27	Honda	32508.00	97027	21	MODERN STEEL
6	18196788	2014-12-18	Toyota	10819.66	95117	55246	WHITE
7	15722278	2014-07-24	Audi	59630.00	90401	143	GLACIER WHITE

An Example Dataset (Shadows)

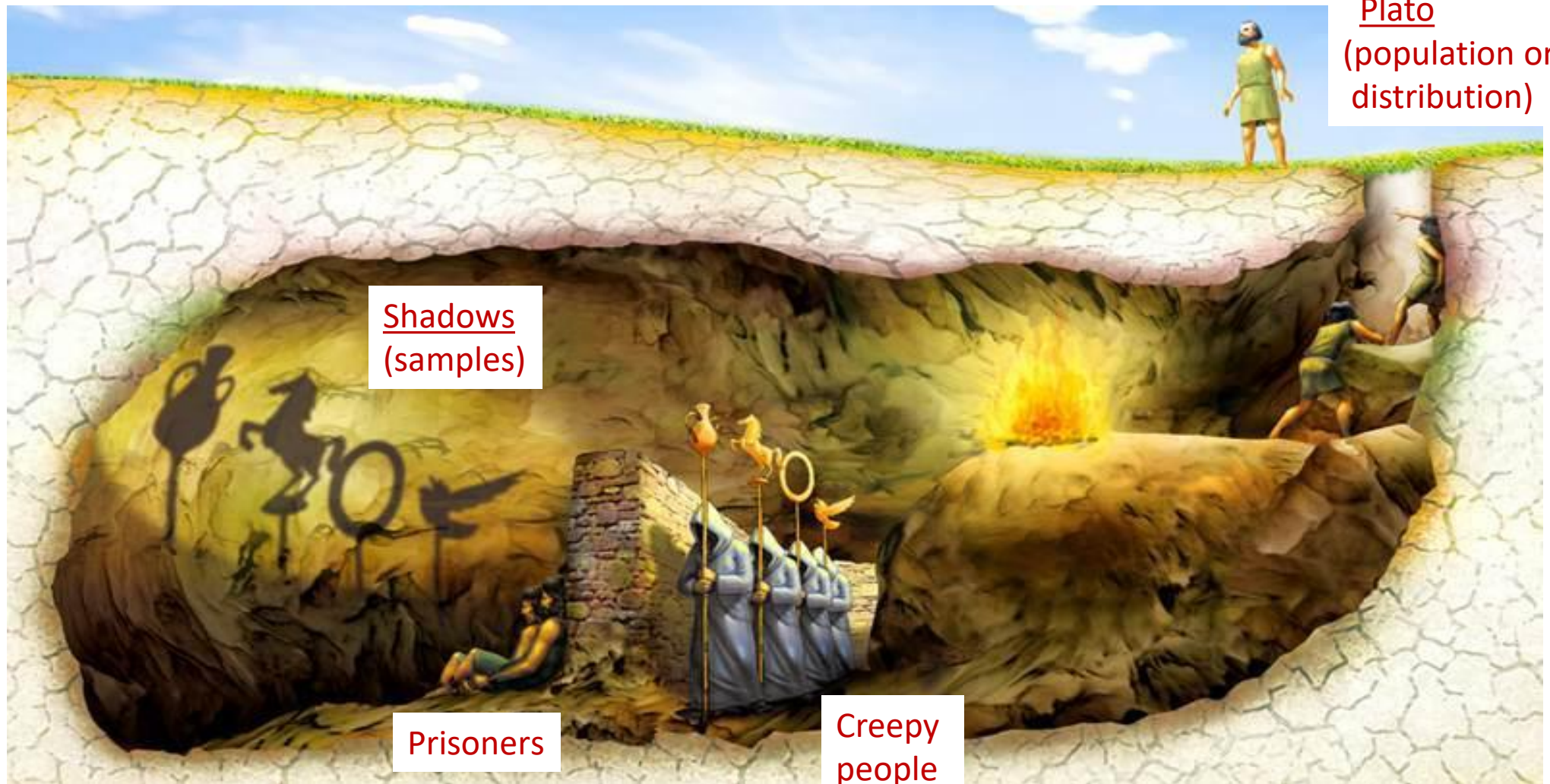
Categorical Variable

Quantitative Variable

Cases
(observational units)

	transactionid	date_sold	make_bought	price_bought	zip_bought	mileage_bought	color_bought
1	16966151	2014-09-27	Acura	30892.00	21043	40	BLACK
2	16914863	2014-09-27	Toyota	25566.00	15108	297	SILVER
3	15977620	2014-07-31	Nissan	34300.00	8753	0	JAVA
4	18666685	2015-01-27	Subaru	30059.00	7446	10	CRYSTAL WHITE PEARL
5	14383133	2014-04-27	Honda	32508.00	97027	21	MODERN STEEL
6	18196788	2014-12-18	Toyota	10819.66	95117	55246	WHITE
7	15722278	2014-07-24	Audi	59630.00	90401	143	GLACIER WHITE

Plato's cave



Course objectives

1. To extend **methods** and **concepts** from intro stats to more complex real world settings
2. To learn how to apply **computational/data science** approaches for data analysis
3. To learn how to analyze and visualize **real data sets** using **the R programming language**

How to find the Truth/trends in data sets and convincingly convey the results to others!

Emphasis in this class

Learning how to apply a range of methods that can give insights into different types of data



Gain intuition on why/how particular methods work

- No math proofs, but we will explore concepts via computational simulations



Data frames

```
> the_ages <- profiles$age
```

Can you get the mean() age of users in this data set?

```
> mean(the_ages)
```

What went wrong?

Missing values (NA) can be removed using the na.rm = TRUE argument to the mean() function

```
> mean(the_ages, na.rm = TRUE)
```

About you

Working together will be useful in this class

Let's take 3 minutes to introduce yourself to those sitting next to you



Logistics: homework 1

First homework problem set will be available after class

It is due on Sunday September 8th at 11:59pm

- 90% credit if turned by Monday September 9th at 11:59pm

The material on the problem set is based on what we will learn this week

- You should be able to answer questions 1, 3, and 5 by the end of today's class
 - Problem 5 is just reading and commenting on a blog post/Wired article

TA office hours are on a [google calendar on Canvas](#)

Logistics: homework 1

There are two ways to do the homework

1. The homework can be completed on R Studio Cloud (preferred)

<https://rstudio.cloud/spaces/25704/project/477656>

2. The homework should also work using R Studio desktop

- Download the R Markdown document from Canvas
- Run the code in the first chunk from the console

Warning: coding can be a bit frustrating at first

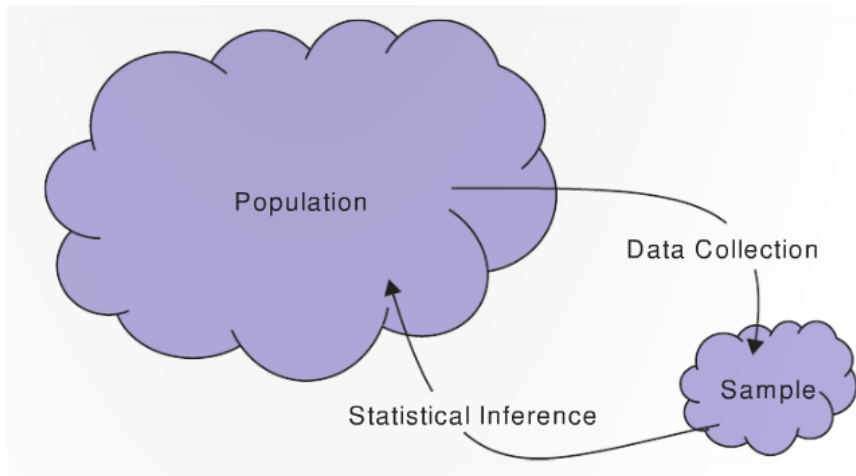
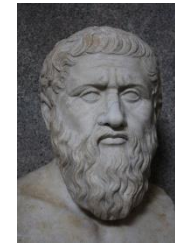
- Try it on your own, then get help as needed (Piazza, TA office hours)

Very quick review

Concepts from Intro Stats

The Truth[®] is out there

- If we have infinite data we could compute parameters
- We can estimate parameters with statistics
 - statistics are functions of our data



	Sample Statistic	Population Parameter
Mean	\bar{x}	μ
Standard deviation	s	σ
Proportion	\hat{p}	π
Correlation	r	ρ
regression slope	b	β

Logistics

R Studio Cloud outage happened during last class

- [Current R Studio Cloud status](#)
- Will keep trying for another week



[Class background survey results](#)

Homework 1

Homework 1 due on Sunday September 8th at 11:59pm

TA office hours are on a [google calendar on Canvas](#)

R packages

Packages add additional functionality to R

We will use many additional packages in this class

- `plyr`, `ggplot2`, `tidyr`, etc.

There is also a class specific package (SDS230) I wrote that you can use to download homework and other files

- All class materials are also on GitHub: <https://github.com/emeyers/SDS230>



Installing SDS230 package and LaTeX

To install the SDS230 package you first need to install the devtools package which can be done using:

```
install.packages("devtools")
```

You can then install the class SDS230 package using the function:

```
devtools::install_github("emeyers/SDS230")
```

Installing SDS230 package and LaTeX

Finally, after you have installed the SDS package, there is a function in the SDS package that installs LaTeX on your computer

- (this function uses the tinytex package)

To install LaTeX use:

```
SDS230::initial_setup()    # will install LaTeX via tinytex package
```

Test that the installation worked

```
tinytex:::is_tinytex()    # will return TRUE if it works (note: 3 colons)
```


Categorical data

Categorical variables take on one of a fixed number of possible values

For categorical variables we usually want to view:

- How many items are each category or
- The proportion (or percentage) of items in each category

```
# Get information about drinking behavior
```

```
> drinking_vec <- profiles$drinks
```

```
# Create a table showing how often people drink
```

```
> drinks_table <- table(drinking_vec)
```

```
> drinks_table
```

Follow along in R Markdown!

Please follow along by trying out these examples in the class 3 R Markdown file!

Download the file (after installing the SDS230 package):

`SDS230::download_class_code(3)`

Open the file `class_03.Rmd` in R Studio

- The `class_03.Rmd` file will appear in the lower right under the file tab
- Also see videos from class 2 on how to download class code

Follow along in R Markdown!

Run the code in the first chunk to download the required data

```
<!-- Run this once to install packages and get data needed to do
the class exercises -->
```{r setup2, eval = FALSE, include=FALSE}

install.packages("okcupiddata")

download.file("https://yale.box.com/shared/static/t3ezfphfg729x030
79aajop0d3f454wm.rda", "daily_bike_totals.rda")

```
```



Knit the file to make sure it works

- Always knit early and often!

Press the play button
to run this code once

Survey video questions

To make these videos more engaging, I have included a few survey questions on Canvas






Please fill these out these questions as you go through this video

Plots and statistics for quantitative
data and
for loops,

Logistics: class plan

Today and Thursday we will meet at the regular time (9-10:15)

Going forward, how would you prefer the content in this class is delivered?

| | | | |
|--|----------------|------|---|
| All asynchronous (i.e., all prerecorded videos) | 10 respondents | 15 % |  ✓ |
| Mostly asynchronous | 25 respondents | 38 % |  |
| Even split between asynchronous and synchronous | 25 respondents | 38 % |  |
| Mostly synchronous | 3 respondents | 5 % |  |
| All synchronous (i.e., all content delivered live) | 3 respondents | 5 % |  |

Logistics: class plan

Today and Thursday we will meet at the regular time (9-10:15)

Current plan going forward:

- Monday: asynchronous videos and homework released
- Tuesday class: office hours/live run through - **attendance optional**
- Wednesday **3pm**: Submit survey questions about videos
- Thursday class: synchronous meeting with review and activities

Logistics: class participation

Class participation grade (6%)

1. Asking and answering questions on Piazza
 2. Filling out lecture video surveys
- X** Questions during class are highly encouraged but will **not** count toward your class participation grade

