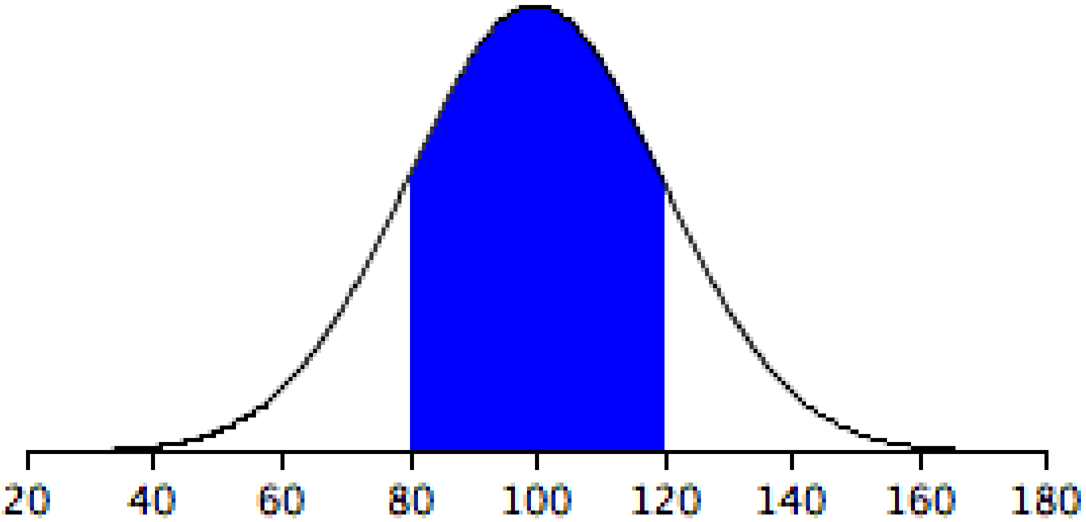
Sampling distributions



Overview

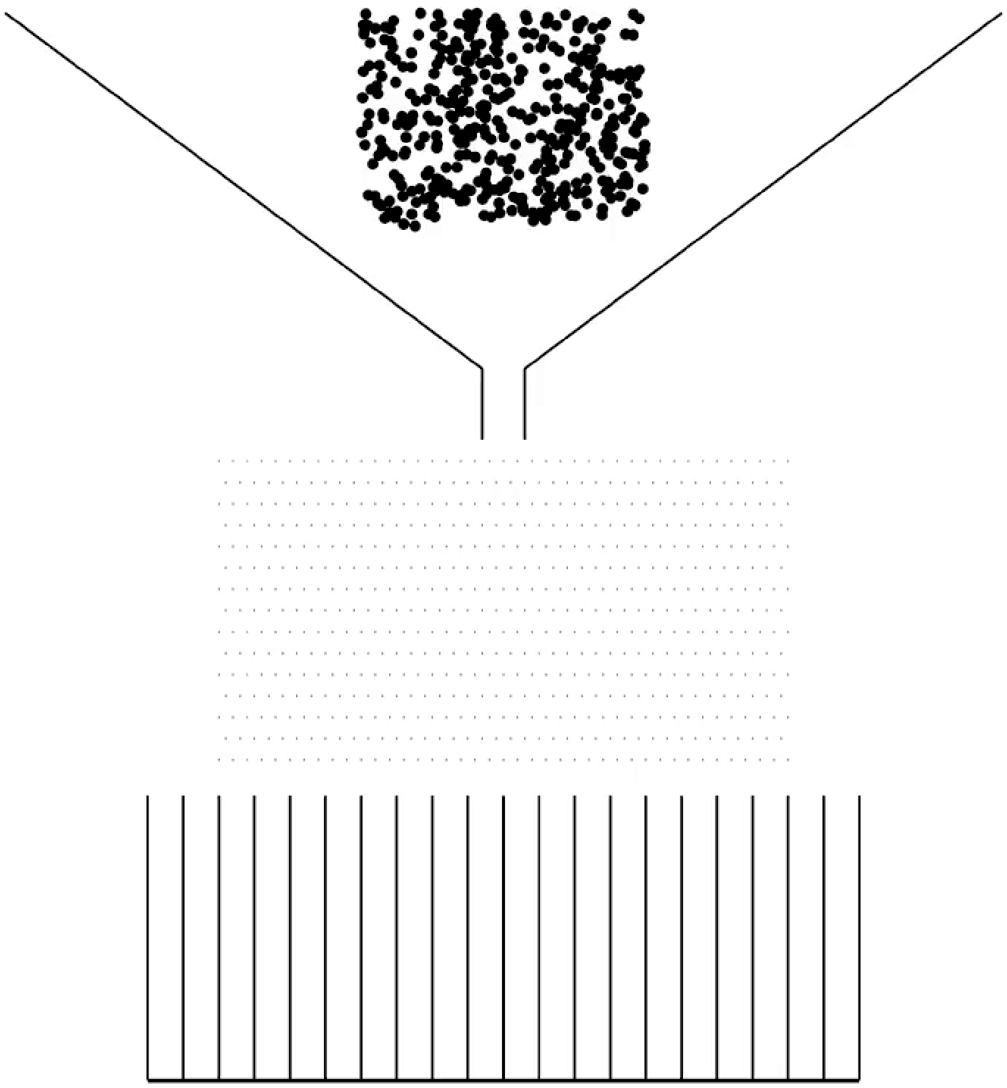
Very quick review

For loops

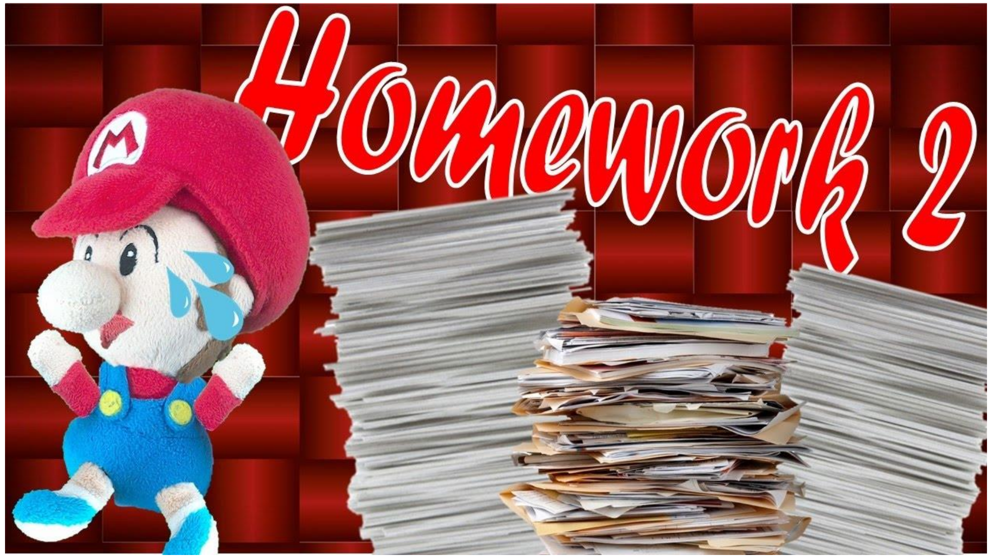
Generating random numbers and selecting random samples

Sampling distributions

If there is time: confidence intervals



Announcements



Homework 2 has been posted

- Due Sunday (9/15) at 11pm
- Start early on it!
  - You can do problems 1 and 2 after today's class
- How was homework 1?

Dean's Extension needed for extensions for undergraduates

Extensions for grad students are allowed but need to be requested a week in advance

Announcement: Office hours cancelled for today

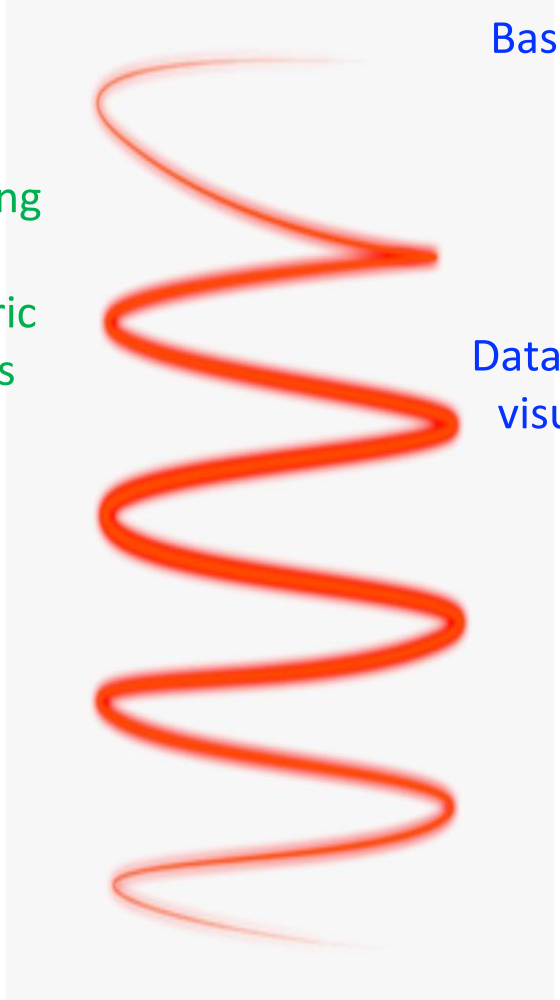
Unfortunately, I need to cancel my office hours today

Feel free to come to my office hours tomorrow (Wednesday) at 2pm in Kline Tower room 1253

Plan for the semester

1	Aug 29	Course overview, introduction to R, descriptive statistics
2	Sep 3-6	Review of central statistical concepts and exploratory analysis using R
3	Sep 10-12	Confidence Intervals and the bootstrap
4	Sep 17-19	Review of hypothesis tests and permutation tests in R
5	Sep 24-26	Parametric tests and theories of hypothesis testing

Analysis **R**



Base R

Resampling and parametric methods

Data wrangling visualization

We will be using simulations to justify and validate methods we use throughout the semester

Quick review

Basics of R

```
> my_vec <- c(5, 28, 19)
> my_vec[3]
> my_vec[3]  <-  7
```

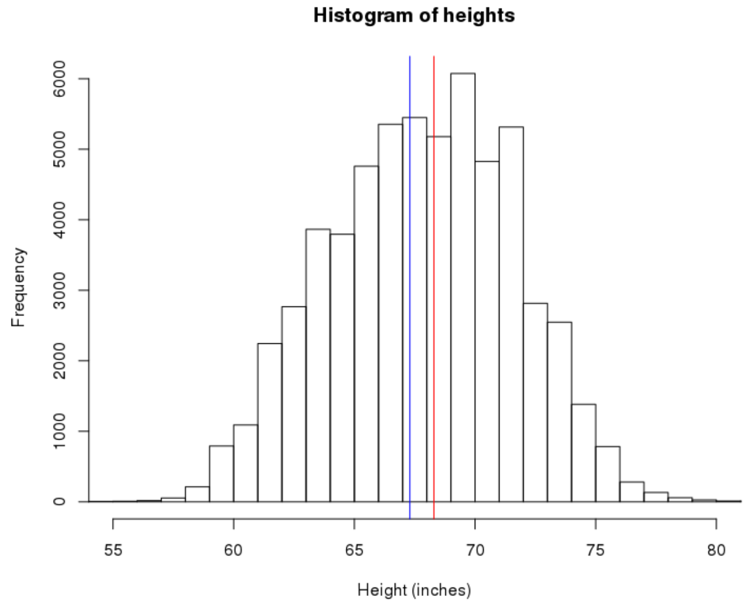
How to plot categorical data

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

Quick review

How to plot quantitative data:

```
> hist(profiles$height)
> abline(v = 67)
```



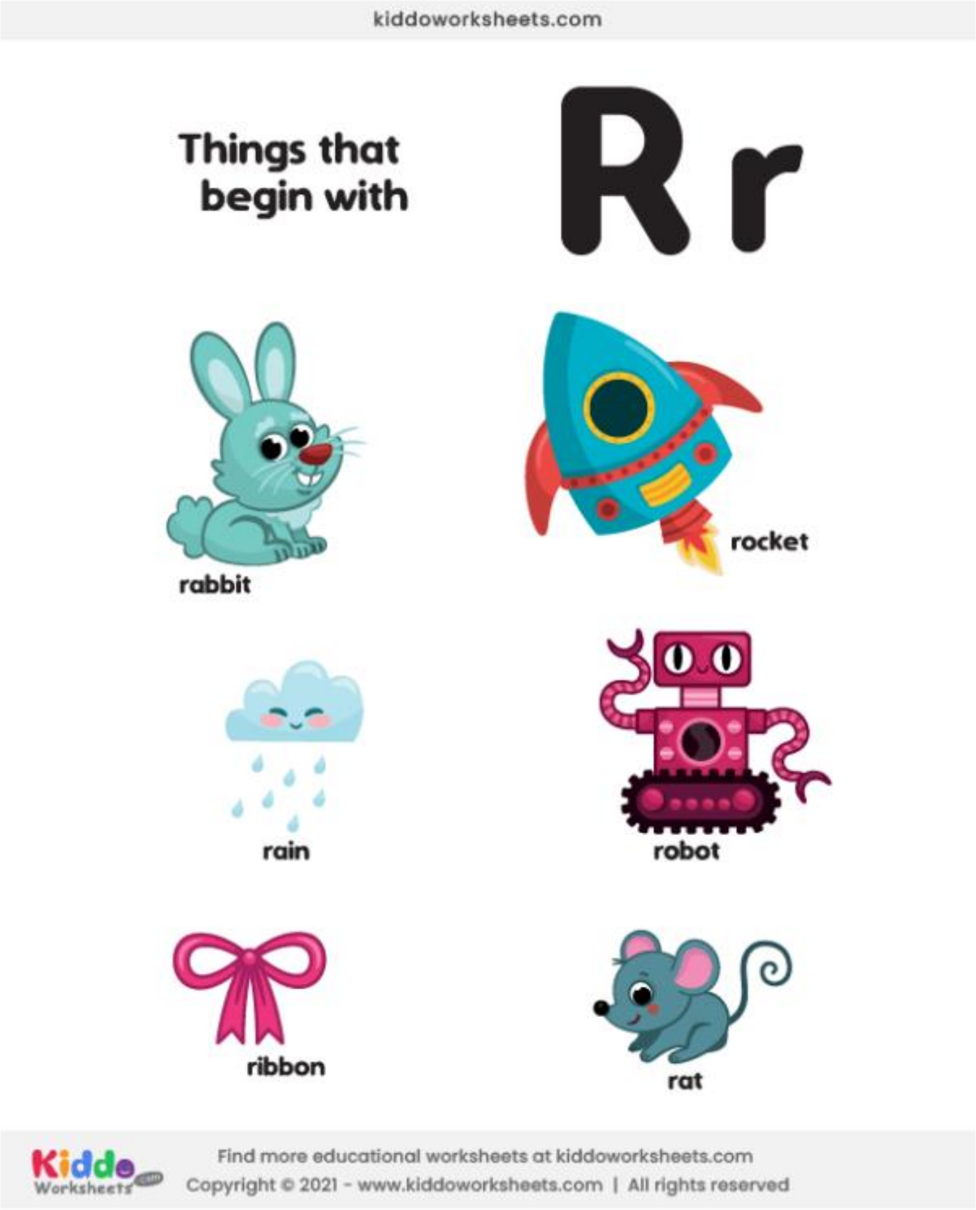
Staying organized

It is useful to create separate folders for different homework and even for the difference pieces of class code.

Be sure to **set your working directory** properly so that R can find the relevant files.

A little more R…

For loops



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 particularly useful in conjunction with vectors…

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

Try this at home!: Use a for loop to create a vector that holds the values at multiples of 3 from 3 to 300

- i.e., 3, 6, 9, …, 300

Let's try it in R!

Generating random data

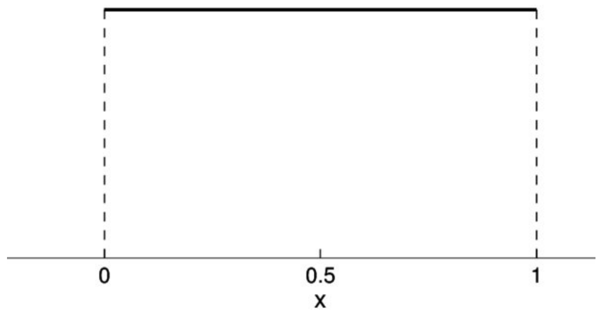
Generating random data

R has built in functions to generate data from different distributions

- All these functions start with the letter ***r***

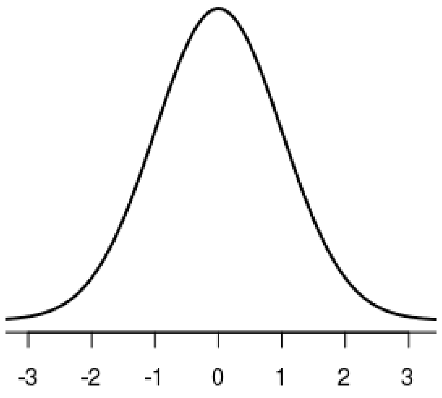
The uniform distribution

```
# 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?

Sampling data

The sample(v, n) function samples **n** random points from a vector **v**

For example, suppose we had a vector with the ages of all US citizens in a vector called pop_ages

We could sample the ages of 100 random people using:

- rand_sample <- sample(pop_ages, 100)

We can sample with replacement using the replace = TRUE argument:

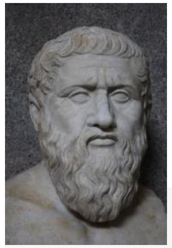
- rand_sample_replace <- sample(pop_ages, 100, replace = TRUE)

Let's try it in R!

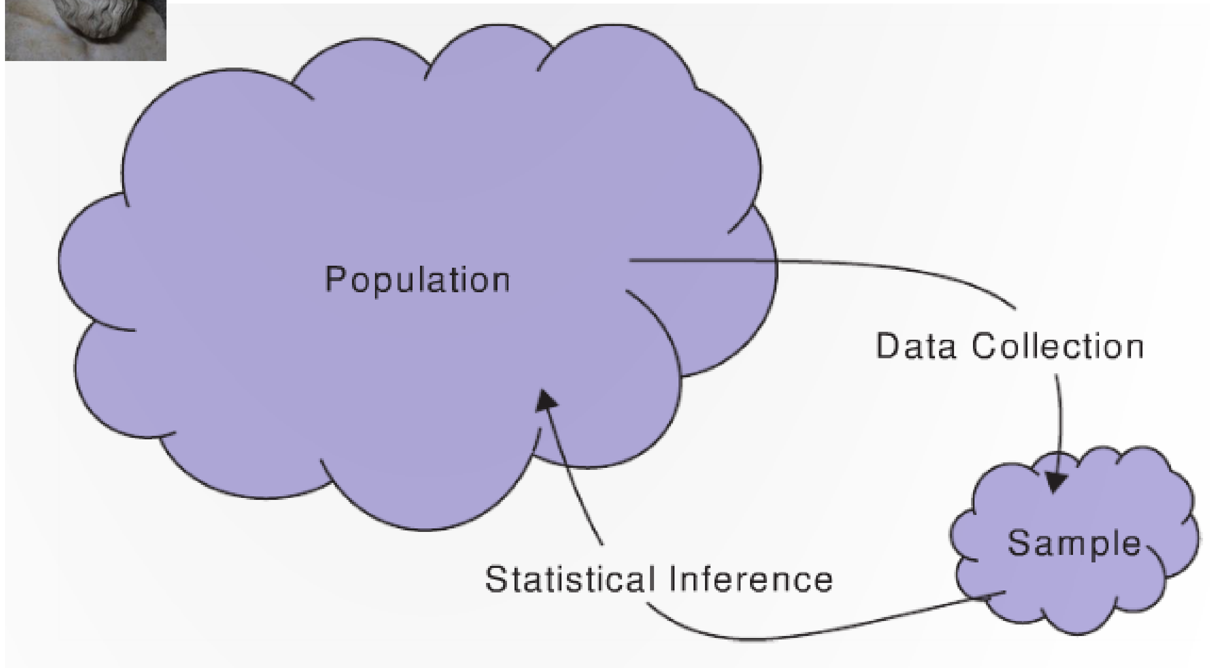
Questions?

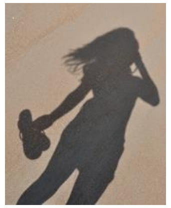
Review and extension of statistical concepts

Where does data come from?



Population: all individuals/objects of interest

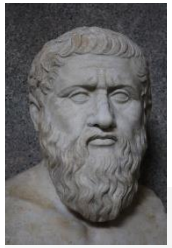




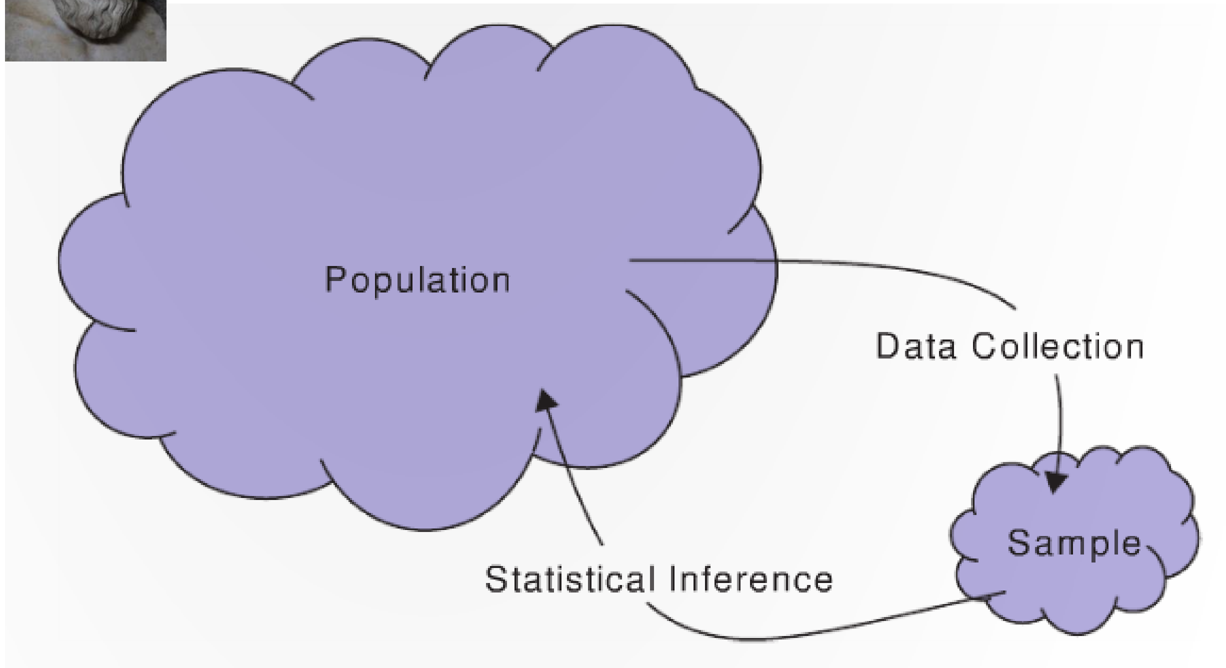
Sample: A subset of the population

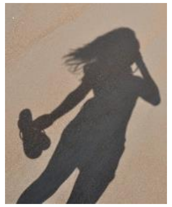
Where does data come from?

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



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





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

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

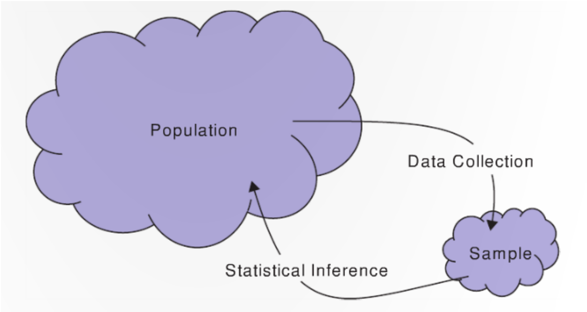
"Random selection"

Soup analogy!

Q: Why is this good?

A: Allows for generalizations to the population!

- No sampling bias
- Statistic (on average) equal parameter
  - E.g., $E[\bar{x}] = \mu$



Questions:

- Is the OkCupid profiles data a simple random sample?
- Would we expect sampling bias from statistics computed from the OkCupid profiles?

Big picture for the week

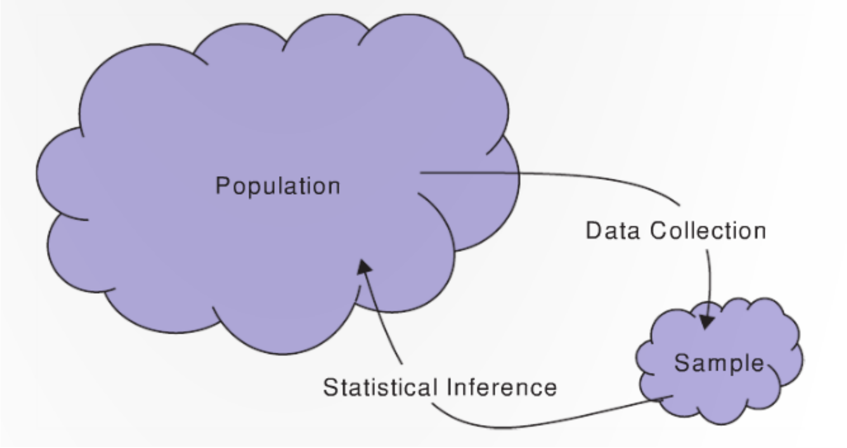
Statistics are point estimates of parameters

We can use sampling distributions (i.e., distributions of statistics) to tell us how much we can trust ***any one statistic*** to be a good point estimate of a parameter

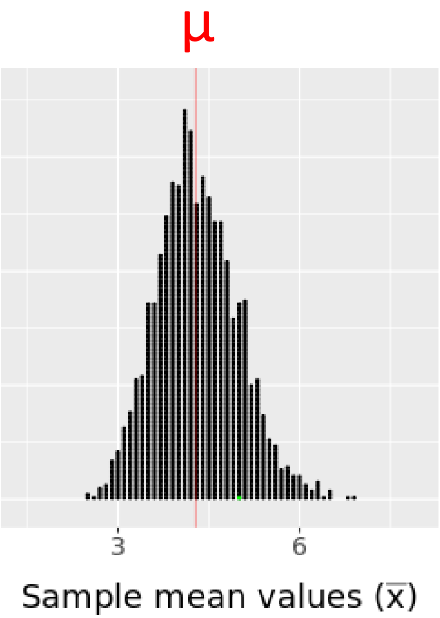
-> confidence interval

Let's starts on this now…

parameter: μ



statistic: $\bar{x}$



Sampling distribution of $\bar{x}$

Sampling distributions

Sample statistics

Q: What is a statistic?

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?

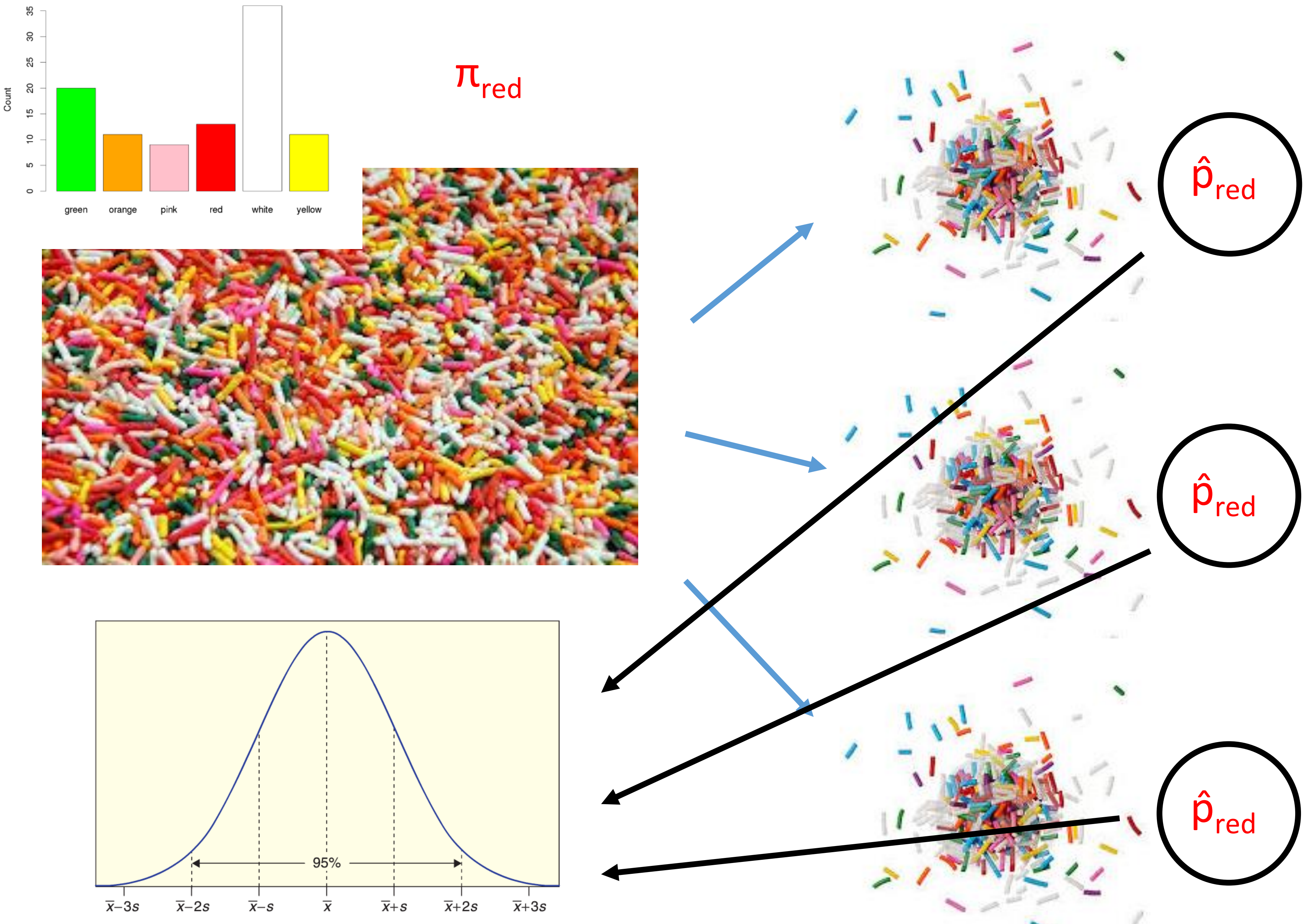
Sampling distributions

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

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

- (shadow of the parameter π)

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

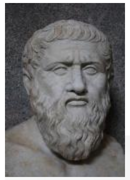


Sampling distribution!

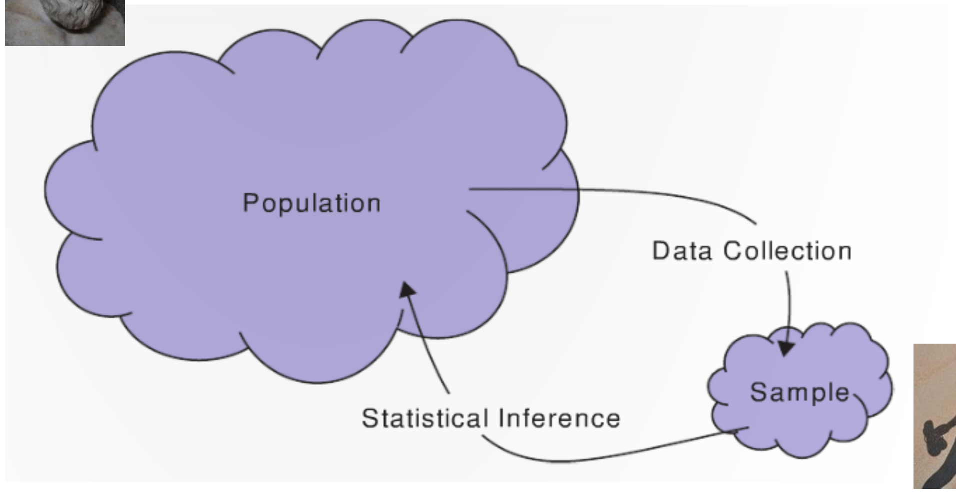
Sampling distribution

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

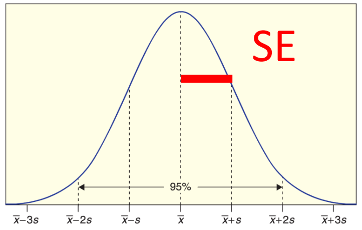


Parameters: π, μ, σ, ρ, β



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

Sampling distribution



The standard error (SE) is the standard deviation of a sampling distribution

It tells us how much statistics vary from sample to sample

Simulating 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)
```

Simulating 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)
```

Simulating sampling distributions

Distribution of OkCupid user's heights n = 100

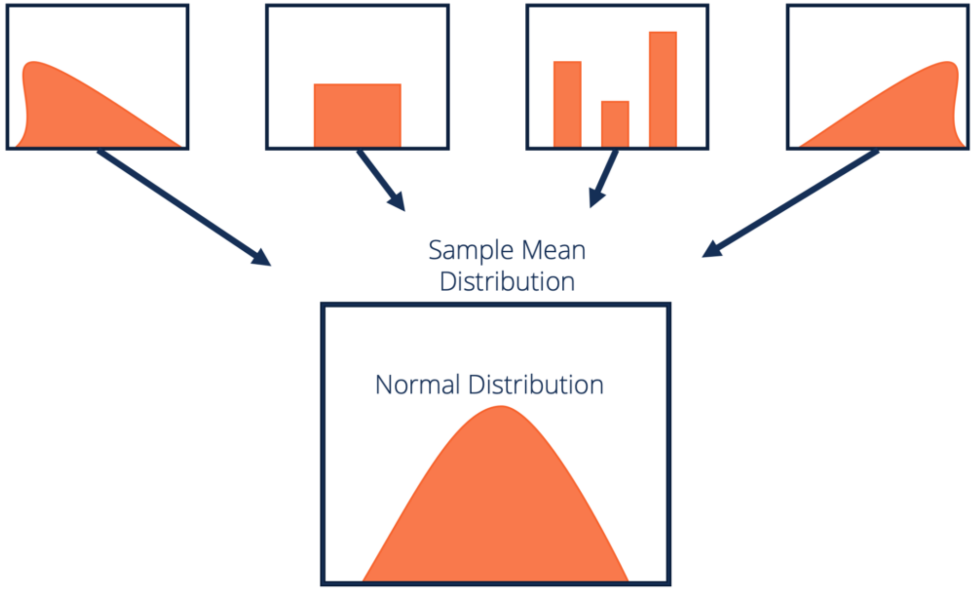
```
sampling_dist <- NULL
for (i in 1:1000) {
	height_sample <- sample(heights, 100)    # sample 100 random heights
	sampling_dist[i] <- mean(height_sample)    # save the mean
}

hist(sampling_dist)
```

The central limit theorem

The **central limit theorem** establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution.

Since many statistics we use are the sum of randomly data, many of our sampling distributions will be approximately normal

- You will explore this more on homework 2



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

Some would say this sidewalk is broken, but it’s actually normal

Confidence intervals

Point Estimate

We use the statistics from a sample as a **point estimate** for a population parameter

- $\bar{x}$ is a point estimate for…? μ

A recent New York Times/Siena College poll found that Trump's favorability rating was 46%

Symbols:

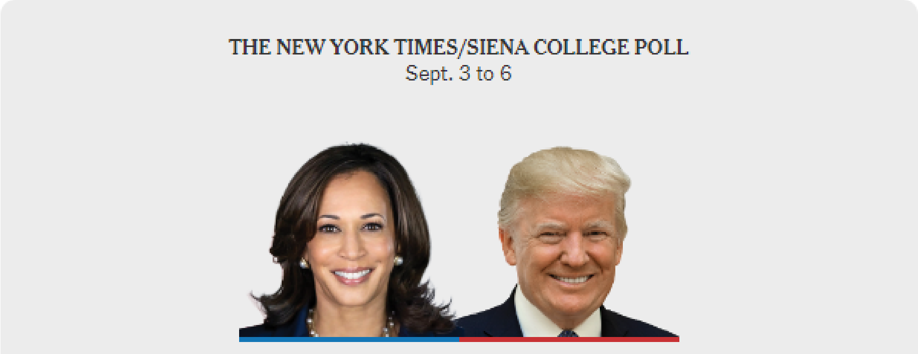
π: Trump's favorability for all voters

$\hat{p}$: Trump's favorability for those voters in our sample

Trump and Harris Neck and Neck After Summer Upheaval, Times/Siena Poll Finds

The survey finds that Donald J. Trump is retaining his support and that, on the eve of the debate, voters are unsure they know enough about where Kamala Harris stands.

Listen to this article · 10:36 min Learn more Share full article 1.9K



Interval estimate based on a margin of error

An **interval estimate** give a range of plausible values for a population parameter

One common form of an interval estimate is:

Point estimate ± margin of error

Where the **margin of error** is a number that reflects the precision of the sample statistic as a point estimate for this parameter

Example: YouGov poll

46% of American have a favorable view of Donal Trump, with a margin of error of 2.8%

- i.e., plus or minus 2.8%

How do we interpret this?

Says that the <u>population parameter</u> (π) lies somewhere between:

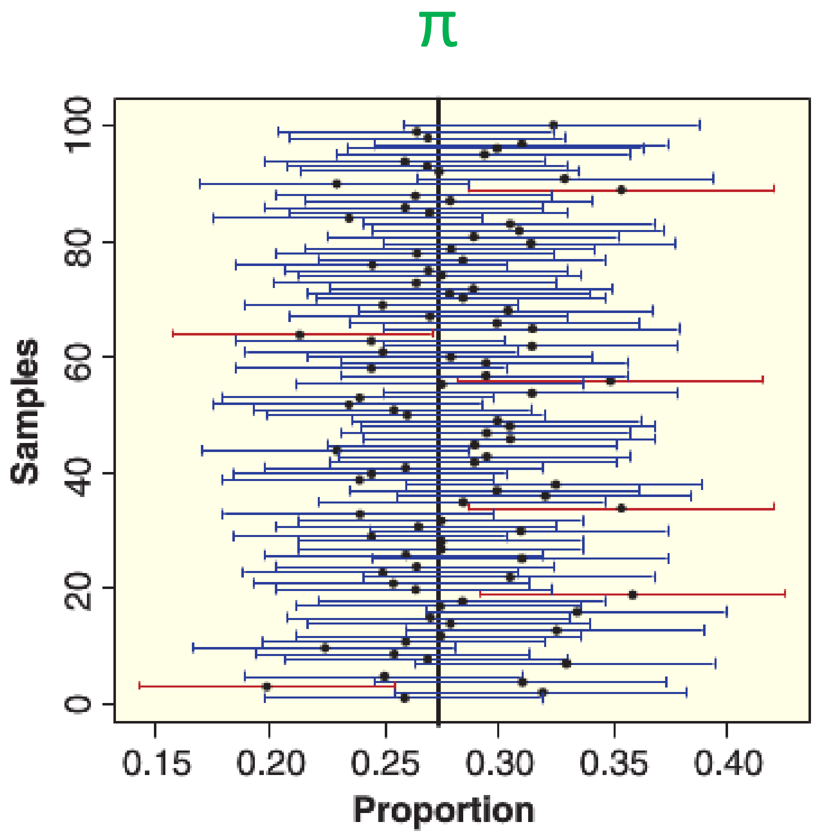
46 - 2.8 to 46 + 2.8 = 43.2 to 48.8

i.e., if they sampled all voters the true population proportion (π) would be likely be in this range

Confidence Intervals

A **confidence interval** is an interval <u>computed by a method </u>that will contain the ***parameter*** a specified percent of times

- i.e., if the interval was calculated repeatedly from many different random samples, the parameter will be in p% of these intervals



The **confidence level** is the percent of all intervals that contain the parameter

Think ring toss…

Parameter exists in the ideal world

We toss intervals at it

95% of those intervals capture the parameter