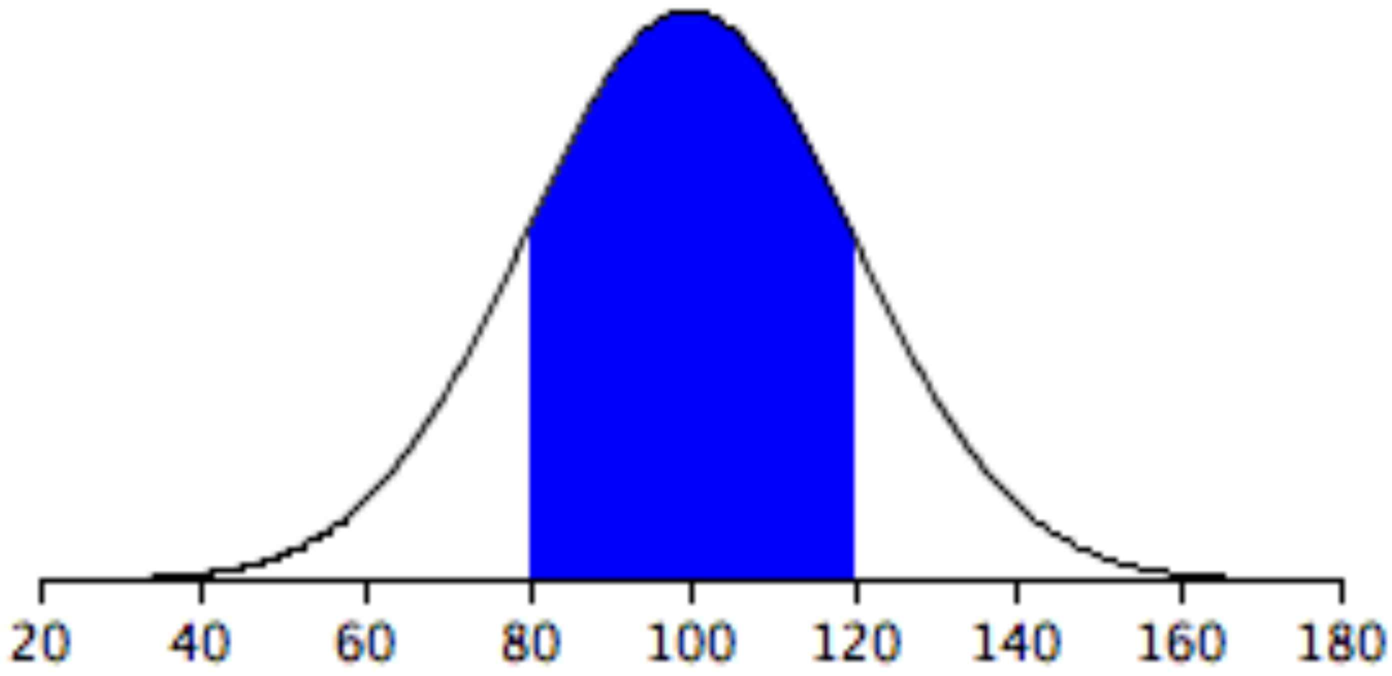
Distributions of data and statistics



Overview

R odds and ends

for loops

Probability density functions and random numbers

Sampling distributions

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

For loops and R odds and ends…

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?

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
```

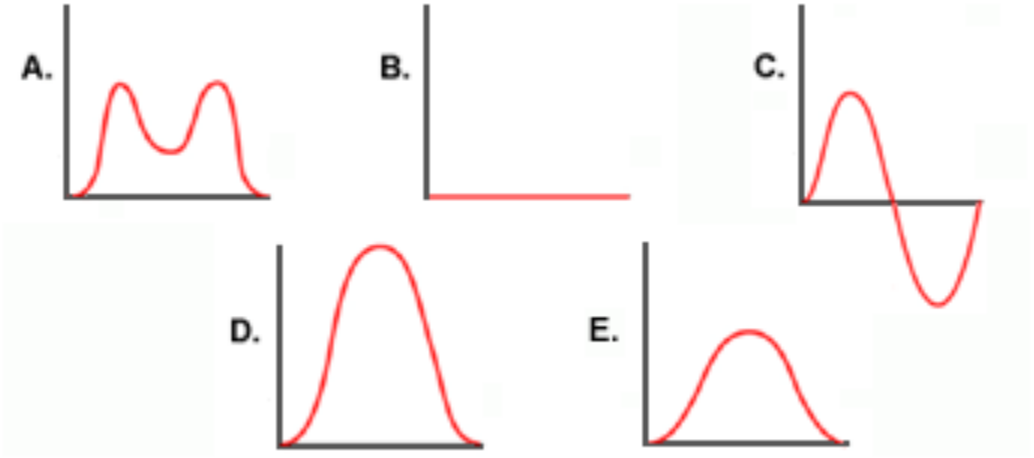
Probability density functions and generating random data

Density Curves

A **density curve** is a mathematical function f(x) that has two important properties:

1. The total area under the curve f(x) is equal to 1
2. The curve is always ≥ 0

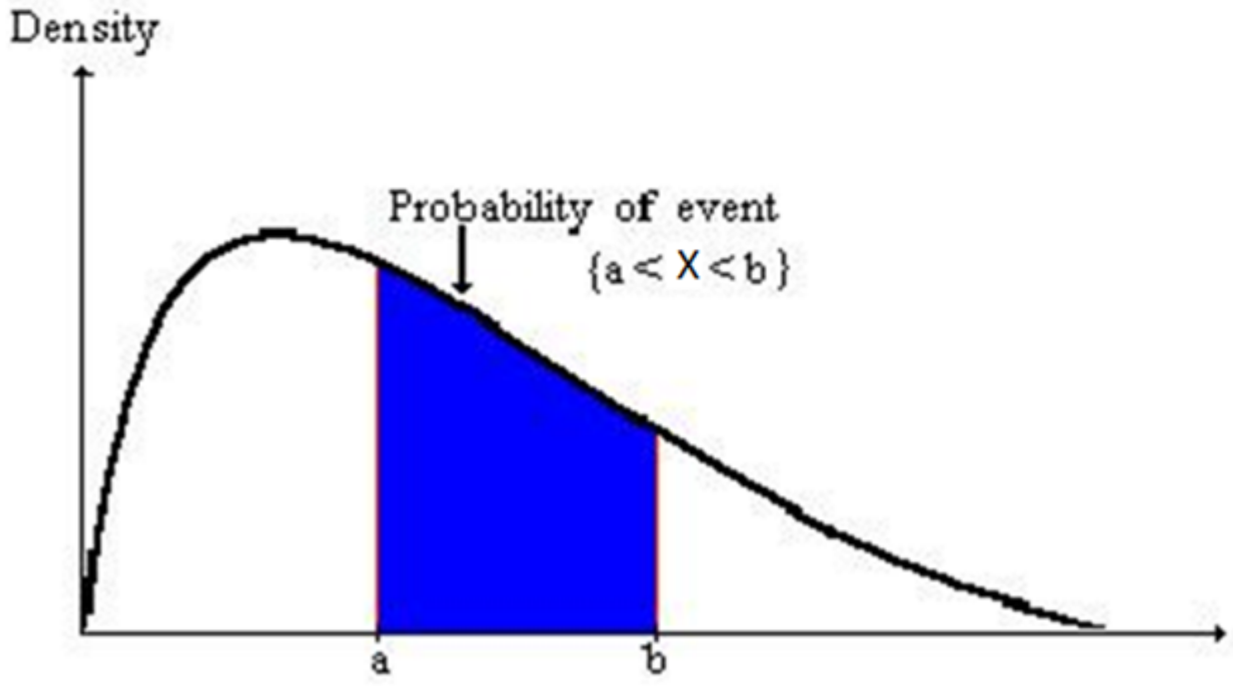
Which of these could ***not*** be a density curve?



Density Curves

The <u>area under the curve</u> in an interval [a, b] models the probability that a random number X will be in the interval

Pr(a < X < b) is the area under the curve from a to b



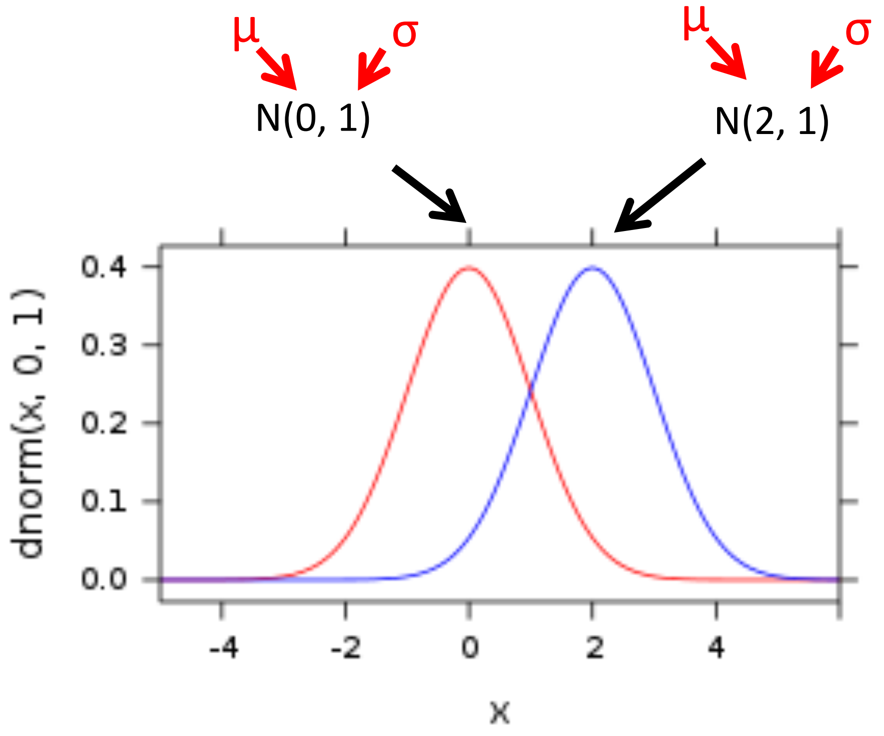
Example: Normal Density Curve

Normal distributions are a family of bell-shaped curves

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

There are two parameters that characterize normal curves, which are:

- The mean: μ
- The standard deviation: σ



Notation: X ~ N(μ, σ)

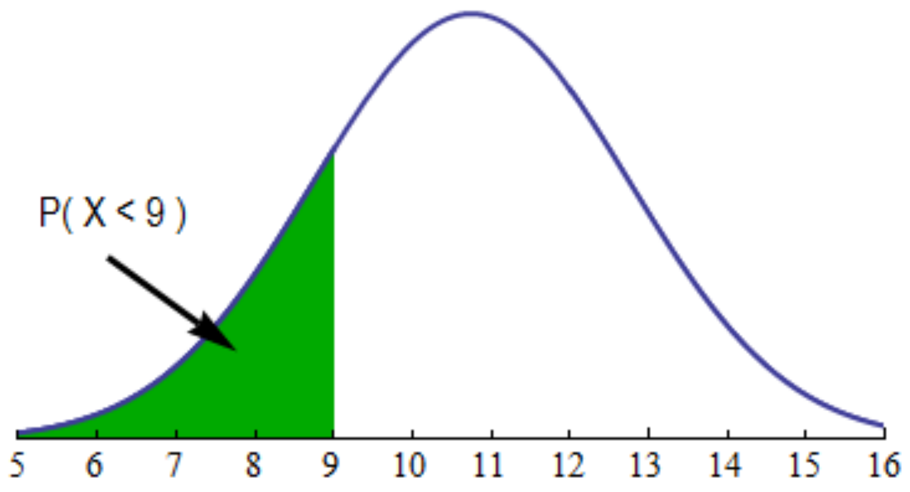
Densities, probabilities and quantiles for **normal distributions**

We can plot the density curve using:

dnorm(x_vec, mu, sigma)

We can get the probability that we would get a random value less than x using:

pnorm(x_vec, mu, sigma)



We can get the quantile values using:

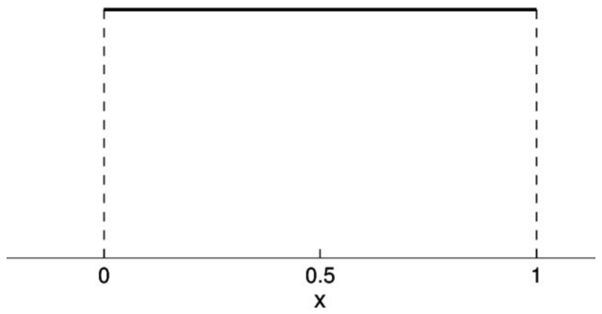
qnorm(area, mu, sigma)

Generating random data

R has built in functions to generate data from different distributions

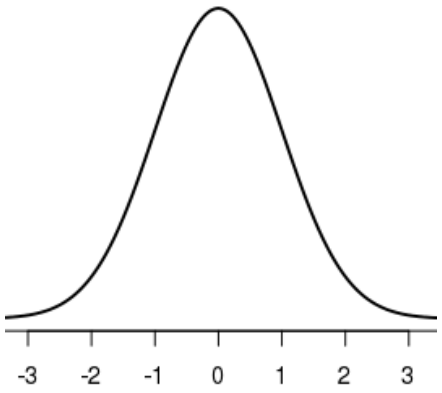
The uniform distribution:

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



The normal distribution

```r
# 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?

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?

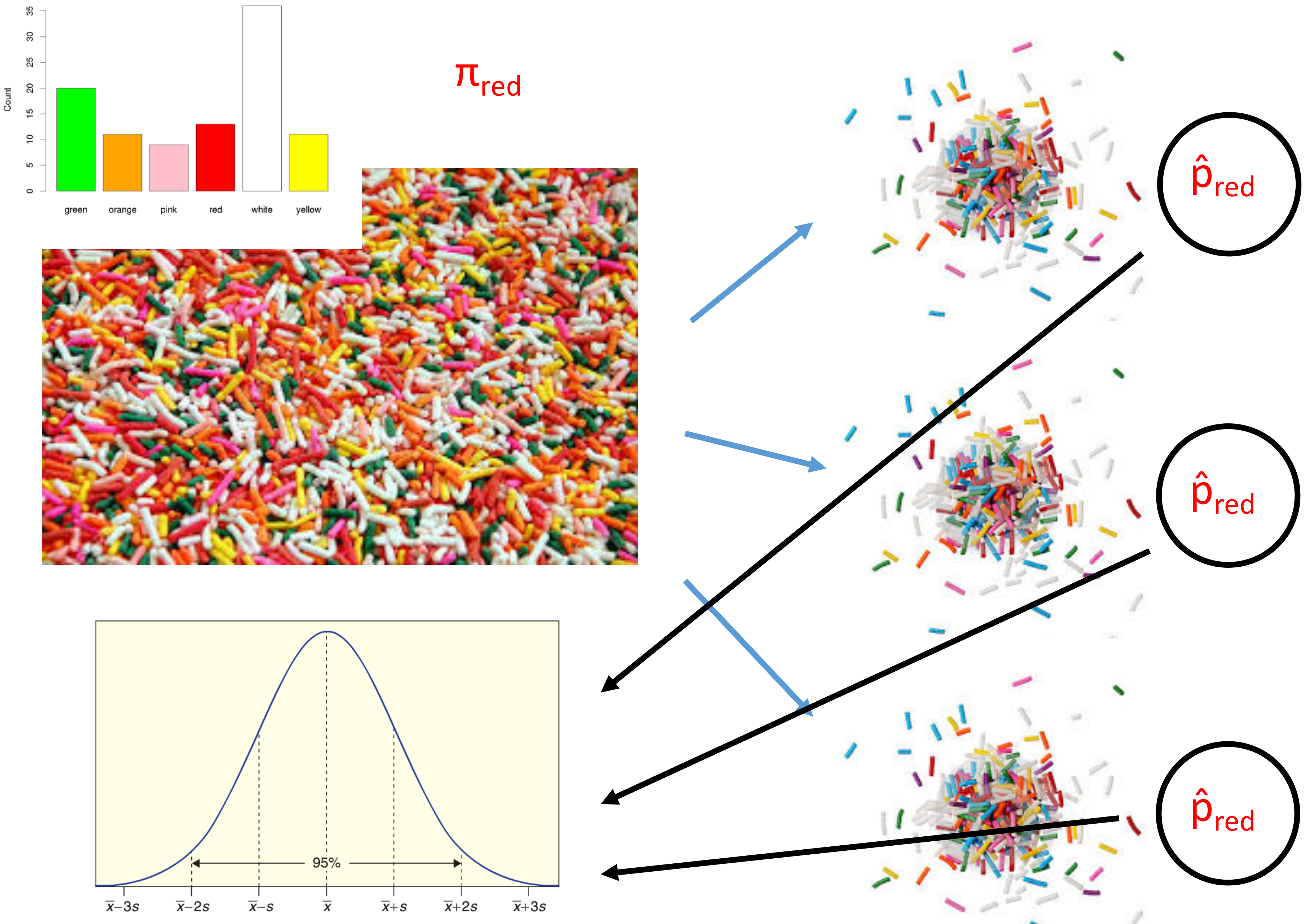
Sampling distributions

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

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

- (shadow of the parameter π)

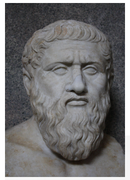
$$\hat{p} = \text{Proportion in a category} = \frac{\text{number in that category}}{\text{total number}}$$



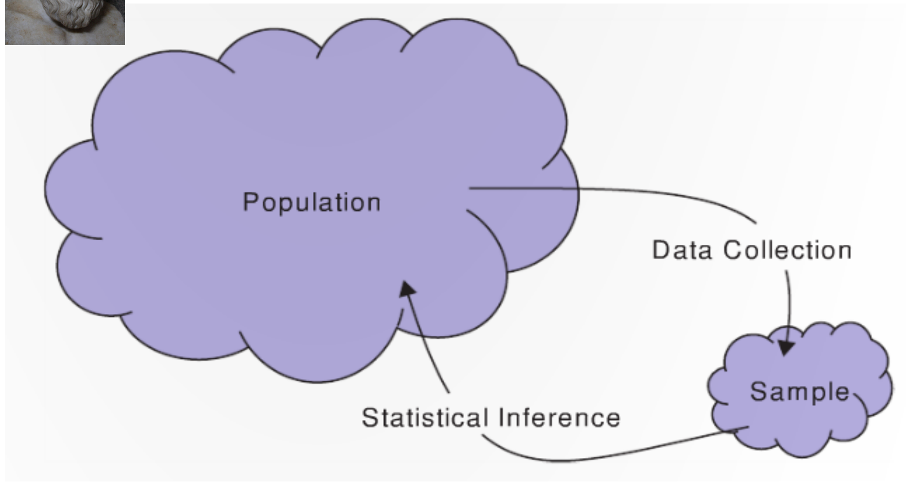
Sampling distribution!

Sampling distribution

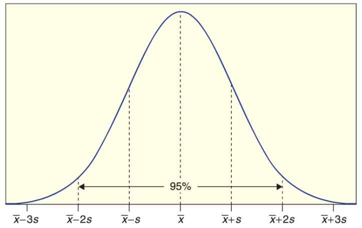
Why would we be interested in the sampling distribution?



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)
```

Next week: confidence intervals and the bootstrap

Videos for the next class will be posted online on (or before) Monday

Tuesday’s class is optional office hours

- Ask questions about the videos, homework, etc.
- Fill out survey by 3pm on Wednesday

We will meet again synchronously on Thursday (9/17)