# Math 107

Confidence Intervals: Bootstrapping
(Section 3.3)

# Pulling the Sample up by its Bootstraps

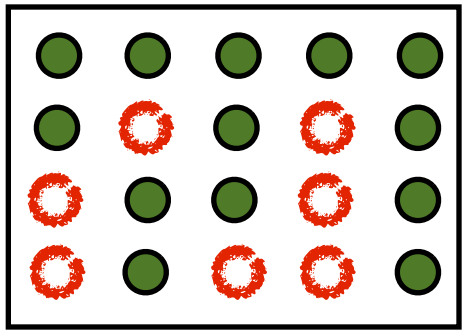
We only have one sample, but we need to understand the sampling distribution!

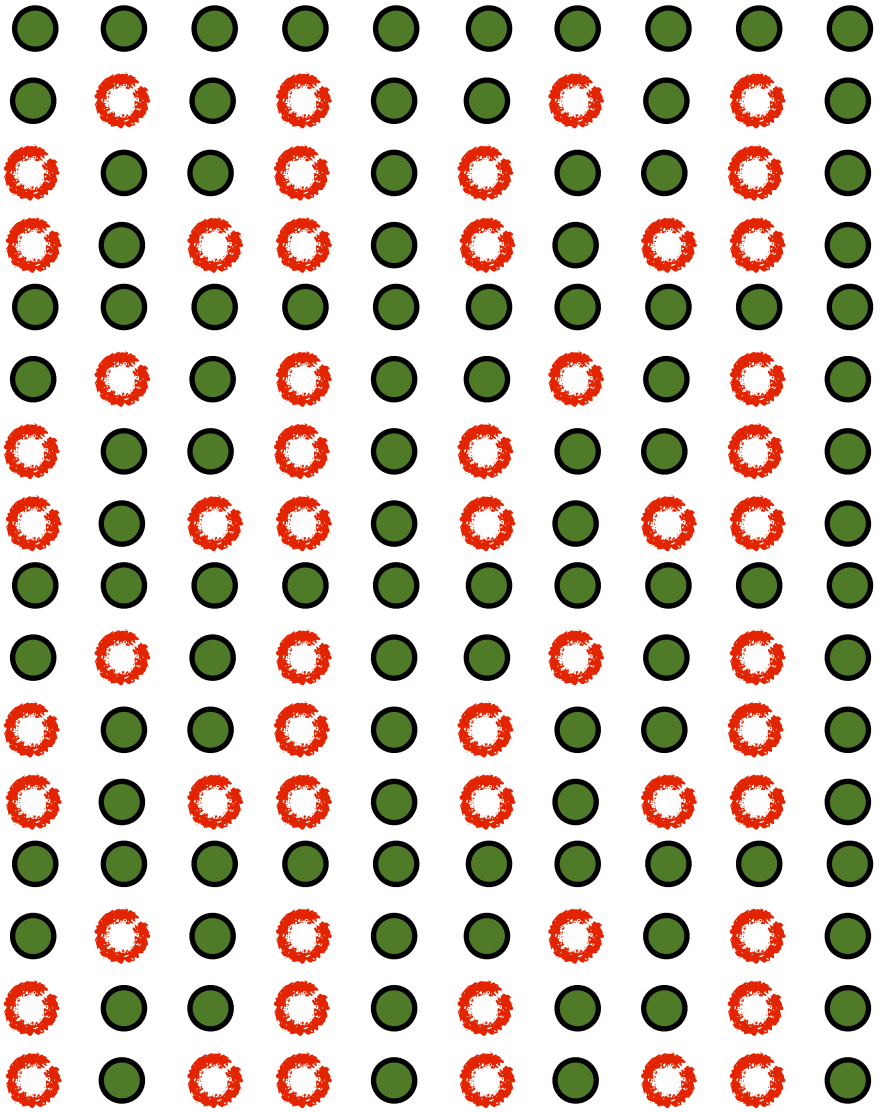
If we can assume that the sample is representative of the population, then the population can be thought of as many copies of the original sample.

We can simulate a sampling distribution by sampling with replacement from the original sample!

# Simulating the Population

Original Sample

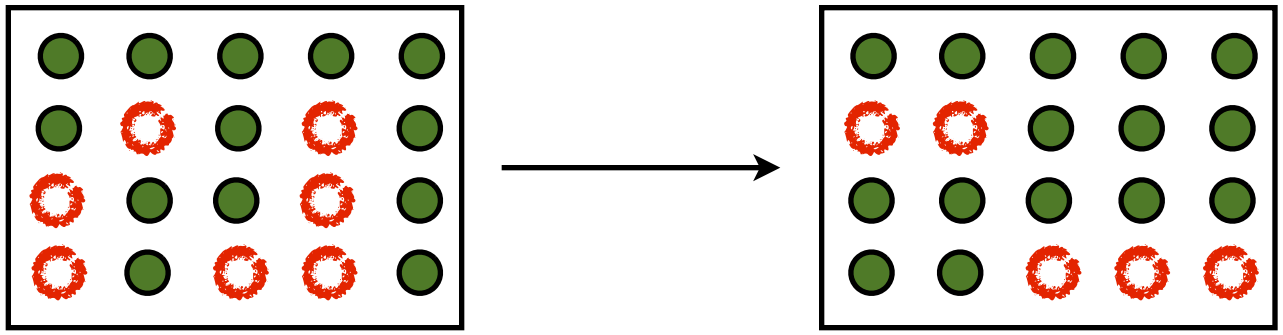




Simulated Population

# Bootstrap Sample

A **bootstrap sample** is created by sampling with replacement from the original sample, using the same sample size.



Original Sample

Bootstrap Sample

# Your Turn

The original sample data contains values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18,19, 20, 21, 22

# Your Turn

The original sample data contains values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18,19, 20, 21

# Your Turn

The original sample data contains values

18, 19, 19, 20, 21

Is the following a possible bootstrap sample?

18, 18,19, 20, 21

# Bootstrapping

Let n be the sample size.

1. Draw a **bootstrap sample** of size n with replacement from the sample
2. Compute the statistic of interest (this is called the **bootstrap statistic**)
3. Repeat steps 1 and 2 many times—say 1,000 or 10,000
4. Create the **bootstrap distribution**

# American Community Survey

- Each year since 2005, the US Census Bureau surveys about 3.5 million households (sampled randomly)
- Data collected from the ACS have been crucial in government and policy decisions, helping to determine the allocation of more than $400 billion in federal and state funds each year.
- Let's take a look at a random sample of 1,000 respondents.

# American Community Survey

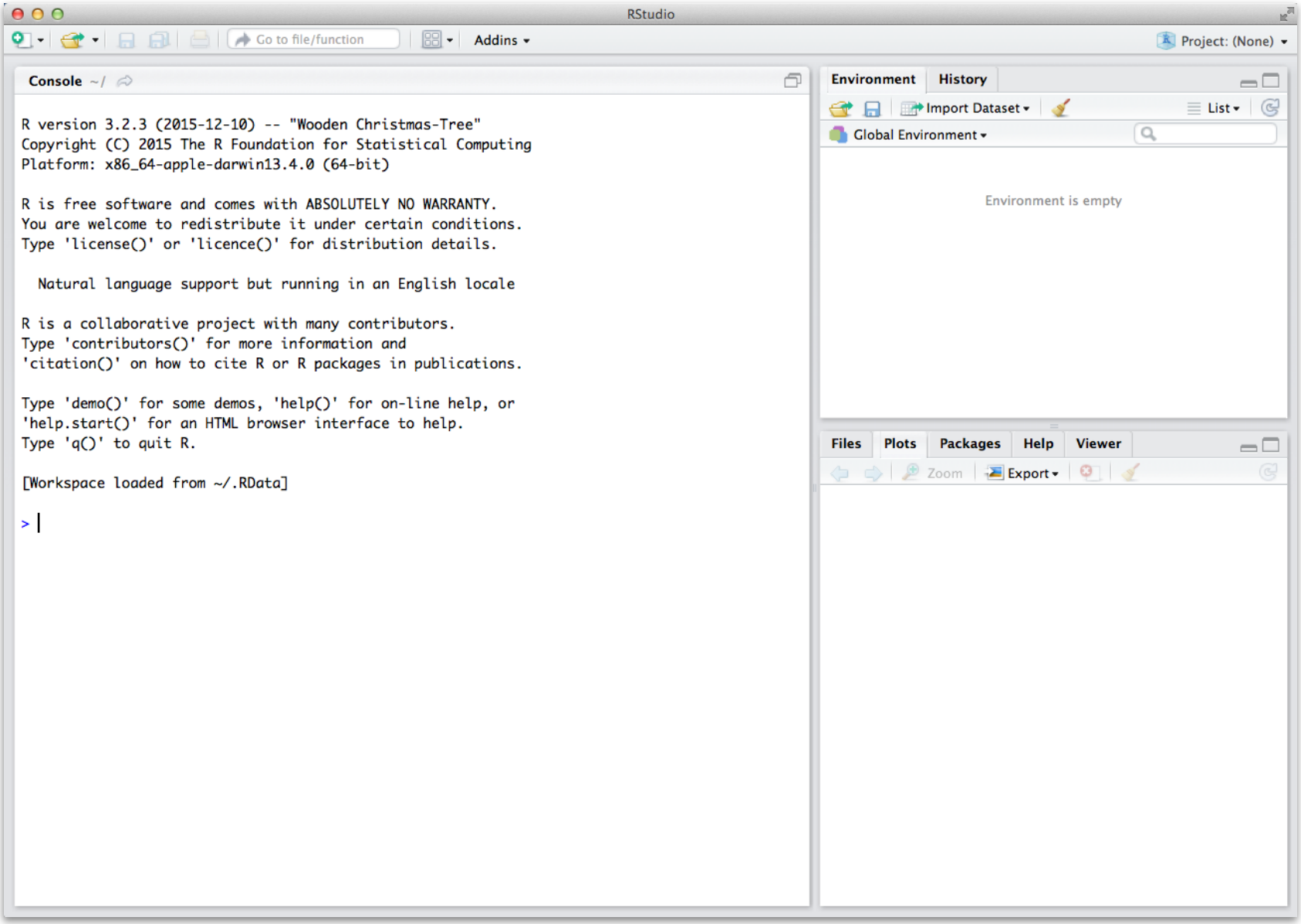
| Sex | Age | Married | Income | HoursWk | Race | USCitizen | HealthInsurance | Language |
|---|---|---|---|---|---|---|---|---|
| F | 31 | not married | 60 | 40 | white | citizen | yes | other |
| M | 31 | not married | 0.36 | 12 | black | citizen | yes | english |
| M | 75 | not married | 0 | | white | citizen | yes | english |
| F | 80 | not married | 0 | | white | citizen | yes | english |
| M | 64 | married | 0 | | white | citizen | yes | english |
| M | 14 | not married | NA | | white | citizen | yes | english |
| M | 78 | married | 0 | | white | citizen | yes | english |
| M | 35 | not married | 87 | 40 | white | citizen | yes | other |
| F | 70 | married | 0 | 1 | white | citizen | yes | english |

# American Community Survey

Possible Questions:

- What proportion of Americans are native English speakers?
- What is the average age of Americans?

# Bootstrapping in R



```
# Load the mosaic and ggplot2 packages
library(mosaic)
library(ggplot2)

# Load the data
acs <- read.csv("data/ACS.csv")

# Calculate the observed average age
xbar <- mean(~Age, data = acs); xbar

# To get a bootstrap sample, use the resample function
mean(~Age, data = resample(acs))
```

```
# Creating a bootstrap distribution of 1000 sample
means
age_boot <- do(1000) * mean(~Age, data =
resample(acs))

# Plot the bootstrap distribution
ggplot(data = age_boot) +
  geom_histogram(mapping = aes(x = mean)) +
  xlab("Means")

# Calculate the standard error
SE <- sd(~mean, data = age_boot); SE
```

# The Golden Rule of Bootstrapping

The bootstrap statistics are to the original sample statistic as the original sample statistic is to the population parameter.

# Center

The sampling distribution is centered around the population parameter, so the bootstrap distribution is centered around the

1. population parameter
2. sample statistic
3. bootstrap parameter
4. bootstrap statistic

We don’t care (much) about the center, just about the variability!

# Standard Error

The variability of the bootstrap statistics is similar to the variability of the sample statistics.

The standard error can be estimated by standard deviation of the bootstrap distribution!

# Plug-in CIs

We can build a **95% bootstrap CI** by "plugging in" to the **bootstrap standard deviation** into the following CI formula:

$$\text{statistic} \pm 2(\text{SE})$$

This will work for **most** parameters!

```
# Load the moasic and ggplot2 packages
library(mosaic)
library(ggplot2)

# Load the data
acs <- read.csv("data/ACS.csv")

# Calculate the observed prop of native English speakers
phat <- prop(~Language == "english", data = acs); phat

# To get a bootstrap sample, use the resample function
prop(~Language == "english", data = resample(acs))
```

```
# Creating a bootstrap distribution of 1000 sample
props.
english_boot <- do(1000) * prop(~Language == "english",
data = resample(acs))

# Plot the bootstrap distribution
ggplot(data = english_boot) +
  geom_histogram(mapping = aes(x = TRUE.)) +
  xlab("Proportions")

# Calculate the standard error
SE <- sd(english_boot$TRUE.); SE

# Calculate a 95% CI
phat - 2 * SE # Lower bound
phat + 2 * SE # upper bound
```