Bootstrap Confidence Intervals

Robert Edwards

22 February 2019

1 Introduction

In previous weeks we have seen many examples of calculating sample statistics such as means, percentiles, standard deviations and regression coefficients. These sample statistics are used as point estimates of population parameters which describe the population from which the sample of data was taken. That last sentence assumes you're familiar with concepts and terminology about sampling (e.g. from the Statistical Inference course in 1st Semester) so here is a summary of some key terms:

1. **Population**: The population is a set of N observations of interest.
2. **Population parameter**: A population parameter is a numerical summary value about the population. In most settings, this is a value that's unknown and you wish you knew it.
3. **Census**: An exhaustive enumeration/counting of all observations in the population in order to compute the population parameter's numerical value exactly. When N is small, a census is feasible. However, when N is large, a census can get very expensive, either in terms of time, energy, or money.
4. **Sampling**: Collecting a sample of size n of observations from the population. Typically the sample size n is much smaller than the population size N, thereby making sampling a much cheaper procedure than a census. It is important to remember that the lowercase n corresponds to the sample size and uppercase N corresponds to the population size, thus n???N. Point estimates/sample statistics: A summary statistic based on the sample of size n that estimates the unknown population parameter.
5. **Representative sampling**: A sample is said be a representative sample if it "looks like the population". In other words, the sample's characteristics are a good representation of the population's characteristics.
6. **Generalizability**: We say a sample is generalizable if any results based on the sample can generalize to the population.
7. **Bias**: In a statistical sense, we say bias occurs if certain observations in a population have a higher chance of being sampled than others. We say a sampling procedure is unbiased if every observation in a population had an equal chance of being sampled.
8. **Random sampling**: We say a sampling procedure is random if we sample randomly from the population in an unbiased fashion.

1.1 Inference via Sampling

The logic of inference via sampling is:

- If the sampling of a sample of size n is done at **random**, then
- The sample is **unbiased** and **representative** of the population, thus
- Any result based on the sample can **generalize** to the population, thus
- The **point estimate/sample statistic** is an estimate of the unknown population parameter of interest

and thus we have **inferred** something about the population based on our sample.

1.2 Task 1

In 2013 National Public Radio in the USA reported a poll of President Obama's approval rating among young Americans aged 18-29 in an article Poll: Support For Obama Among Young Americans Eroding. Here is a quote from the article:

"After voting for him in large numbers in 2008 and 2012, young Americans are souring on President Obama.

According to a new Harvard University Institute of Politics poll, just 41 percent of millennials (adults ages 18-29) approve of Obama's job performance, his lowest-ever standing among the group and an 11-point drop from April."

Identify each of the following terms in this context. (NB. Do not enter any R code below, but you can access the solution by clicking "Hint".)

- Population: Millennials (Americans aged 18-29)
- Population parameter: The true population proportion p of young Americans who approve of Obama's job performeance.
- Census: Young Americans and asking them if they approve of Obama's job performance.

- Sampling: One way is to get phone records from a dtabase and pick out n phone numbers.
- Point estimates/sample statistics: Th sample proportion $\hat{p}$ of young Americans in the sample that approve of Obama's job performance.

- Representative sampling: Does the sampe of $n = 2089$ young Americans accurately represent the population of all young Americans age 18-29?
- Generalizability: is $\hat{p} = 0.41$ a good estimate of p?
- Bias: Are there any sources of bias in the study? Sampling bias, self-selection bias...
- Random sampling: Was the sampling randomly done?

2 Inference Using Sample Statistics

Scenario	Population parameter	Population Notation	Point estimate/sample statistic	Sample Notation
1	Population proportion	p	Sample proportion	$\hat{p}$
2	Population mean	μ	Sample mean	$\bar{x}$
3	Difference in population proportions	$p_1 - p_2$	Difference in sample proporti	ons $\hat{p}_1 - \hat{p}_2$
4	Difference in population means	$\mu_1 - \mu_2$	Difference in sample means	$\bar{x}_1 - \bar{x}_2$
5	Population regression intercept	β_0	Sample regression intercept	$\hat{\beta}_0$ *or* b_0
6	Population regression slope	β_1	Sample regression slope	$\hat{\beta}_1$ *or* b_1

3 Bootstrapping

The moderndive package contains a sample of 40 pennies collected and minted in the United States. Let's explore this sample data first: The pennies_sample data frame has rows corresponding to a single penny with two variables:

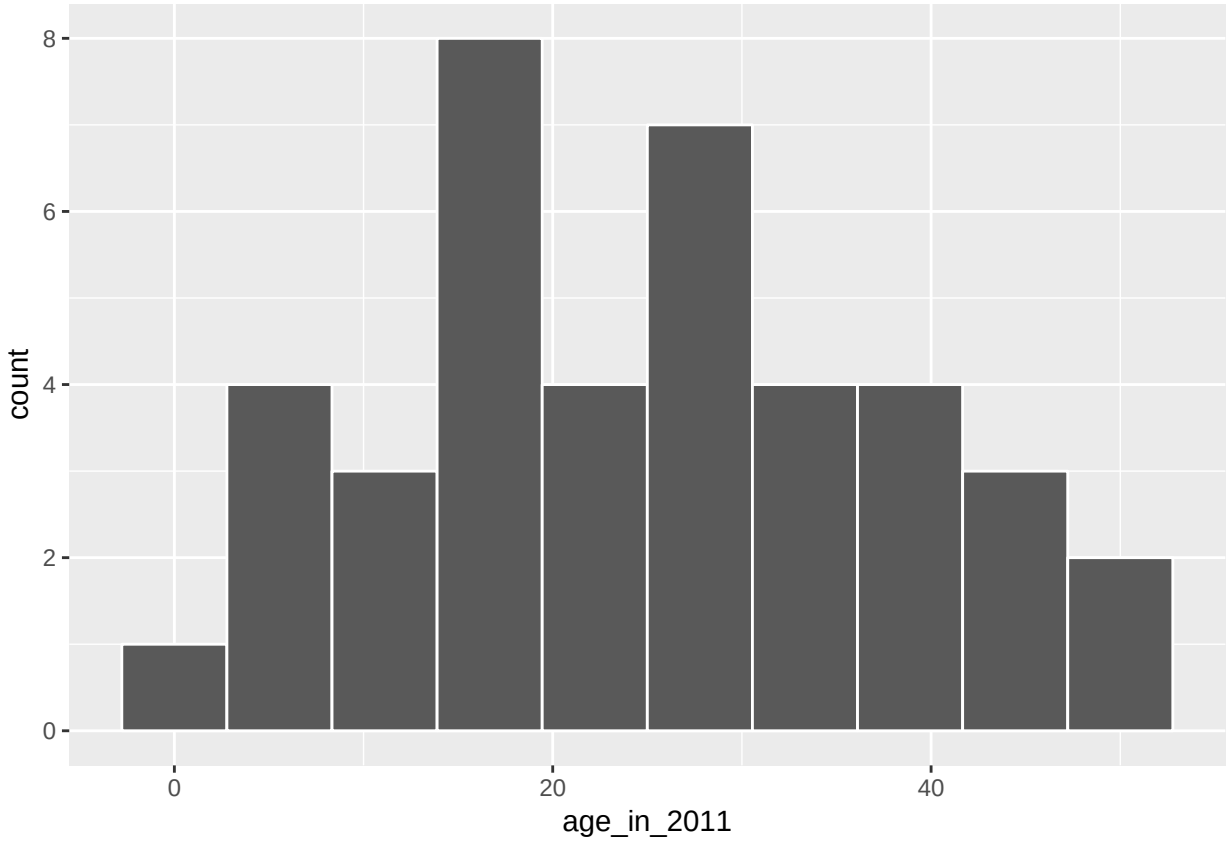
- year of minting as shown on the penny and
- age_in_2011 giving the years the penny had been in circulation in 2011 as an integer, e.g. 15, 2, etc.

Suppose we are interested in understanding some properties of the mean age of all US pennies from this data collected in 2011. How might we go about that? Let's begin by understanding some of the properties of pennies_sample using data wrangling from Week 2 and data visualization from Week 1.

3.1 EDA

First, let's visualize the values in this sampl as a histogram:

```
ggplot(pennies_sample, aes(x = age_in_2011)) +
  geom_histogram(bins = 10, color = "white")
```



We see a roughly symmetric distribution here that has quite a few values near 20 years in age with only a few larger than 40 years or smaller than 5 years. If pennies_sample is a representative sample from the population, we'd expect the age of all US pennies collected in 2011 to have a similar shape, a similar spread, and similar measures of central tendency like the mean.

So where does the mean value fall for this sample? This point will be known as our **point estimate** and provides us with a single number that could serve as the guess to what the true population mean age might be. Recall how to find this using the dplyr package:

```
x_bar <- pennies_sample %>%
  summarize(stat = mean(age_in_2011))
```

We've denoted this sample mean as $\hat{x}$, which is the standard symbol for denoting the mean of a sample. Our point estimate is, thus, $\hat{x} = 25.1$. Note that this is just one sample providing just one sample mean to estimate the population mean. To construct a **confidence interval** (and to do any sort of *statistical inference* for that matter) we need to know about the **sampling distribution** of this sample mean, i.e. how would its values vary if many samples of the same size were drawn from the same population.

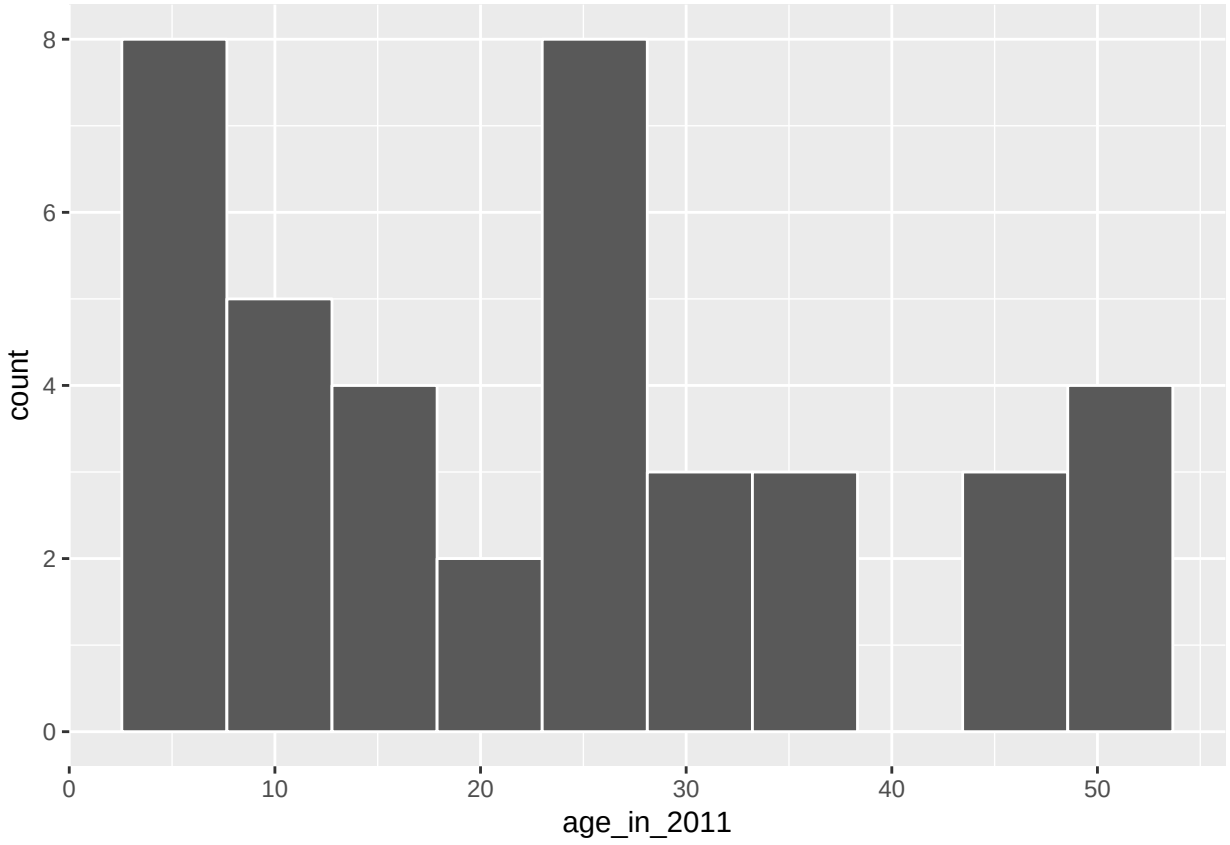
The process of **bootstrapping** allows us to use a single sample to generate many different samples that will act as our way of approximating a sampling distribution using a created **bootstrap distribution** instead. We will "pull ourselves up by our bootstraps" (as the saying goes in English, see here) using a single sample (pennies_sample) to get an idea of the **sampling distribution** of the sample mean.

3.2 The Bootstrapping Process

Bootstrapping uses a process of sampling **with replacement** from our original sample to create new **bootstrap samples** of the *same size* as our original sample. We can use the rep_sample_n() function in the infer package to explore what one such bootstrap sample would look like. Remember that we are randomly sampling from the original sample here **with replacement** and that we always use the same sample size for the bootstrap samples as the size of the original sample (pennies_sample).

```
bootstrap_sample1 <- pennies_sample %>%
  rep_sample_n(size = 40, replace = TRUE, reps = 1)

ggplot(bootstrap_sample1, aes(x = age_in_2011)) +
  geom_histogram(bins = 10, color = "white")
```



We now have another sample from what we could assume comes from the population of interest. We can similarly calculate the sample mean of this bootstrap sample, called a **bootstrap statistic**.

```
bootstrap_sample1 %>%
  summarize(stat = mean(age_in_2011))
```

```
# A tibble: 1 x 2
  replicate  stat
      <int> <dbl>
1         1  23.7
```

We'll come back to analyzing the variation in the values of different bootstrap samples' statistics shortly. But first, let's recap what was done to get to this single bootstrap sample using a tactile explanation:

1. First, pretend that each of the 40 values of age_in_2011 in pennies_sample were written on a small

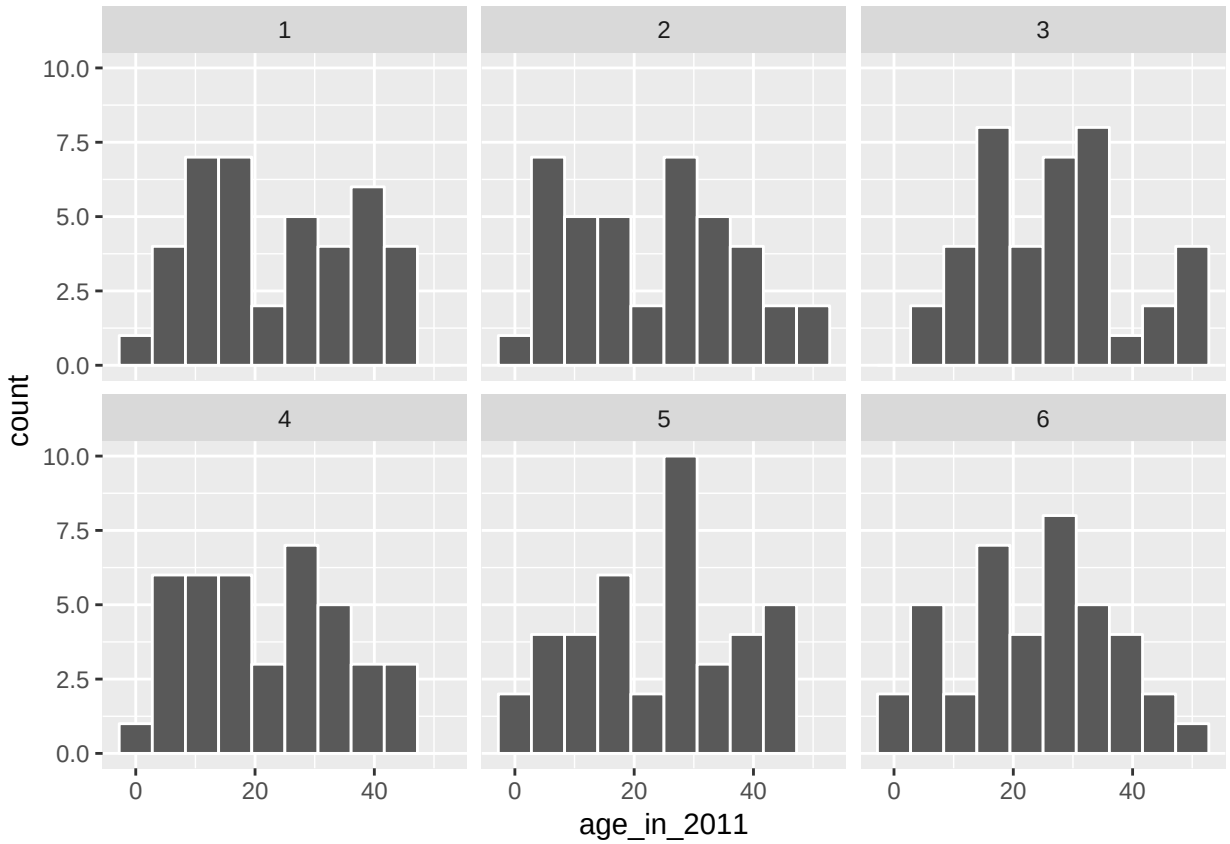
piece of paper. Recall that these values were 6, 30, 34, 19, 6, etc.
2. Now, put the 40 small pieces of paper into a receptacle such as a baseball cap.
3. Shake up the pieces of paper.
4. Draw "at random" from the cap to select one piece of paper.
5. Write down the value on this piece of paper. Say that it is 28.
6. Now, place this piece of paper containing 28 back into the cap.
7. Draw "at random" again from the cap to select a piece of paper. Note that this is the sampling with replacement part since you may draw 28 again.
8. Repeat this process until you have drawn 40 pieces of paper and written down the values on these 40 pieces of paper. Completing this repetition produces ONE bootstrap sample.

If you look at the values in bootstrap_sample1, you can see how this process plays out. We originally drew 28, then we drew 11, then 7, and so on. Of course, we didn't actually use pieces of paper and a cap here. We just had the computer perform this process for us to produce bootstrap_sample1 using rep_sample_n() with replace = TRUE set.

The process of *sampling with replacement* is how we can use the original sample to take a guess as to what other values in the population may be. Sometimes in these bootstrap samples, we will select lots of larger values from the original sample, sometimes we will select lots of smaller values, and most frequently we will select values that are near the center of the sample. Let's explore what the distribution of values of age_in_2011 for six different bootstrap samples looks like to further understand this variability.

```
six_bootstrap_samples <- pennies_sample %>%
  rep_sample_n(size = 40, replace = TRUE, reps = 6)

ggplot(six_bootstrap_samples, aes(x = age_in_2011)) +
  geom_histogram(bins = 10, color = "white") +
  facet_wrap(~ replicate)
```



We can also look at the six different means using dplyr syntax:

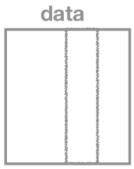
```
six_bootstrap_samples %>%
  group_by(replicate) %>%
  summarize(stat = mean(age_in_2011))
```

```
# A tibble: 6 x 2
  replicate  stat
      <int> <dbl>
1         1  23.6
2         2  23.5
3         3  26.7
4         4  22.2
5         5  24.4
6         6  23.4
```

Instead of doing this six times, we could do it 1000 times and then look at the distribution of stat across all 1000 of the replicates. This sets the stage for the infer R package (see documentation here or the "Cheat Sheet" on the DA Moodle page) that helps users perform statistical inference such as confidence intervals and hypothesis tests using verbs similar to what you've seen with dplyr. In the next section we'll walk through setting up each of the infer verbs for confidence intervals using this pennies_sample example, while also explaining the purpose of the verbs in a general framework.

4 Infer Package for Statistical Inference

The infer package makes great use of the tidyverse "pipe" %>% to create a pipeline for statistical inference. The goal of the package is to provide a way for its users to explain the computational process of confidence intervals and hypothesis tests using the code as a guide. The verbs build in order here, so you'll want to start with specify() and then continue through the others as needed.



specify()

The specify() function is used primarily to choose which variables will be the focus of the statistical inference. In addition, a setting of which variable will act as the explanatory and which acts as the response variable is done here. For proportion problems (i.e. Scenarios 1 & 3 in Table 1) we also specify which of the different levels we are calculating the proportion of (e.g. "females", "approve of Obama's job performance", etc.). To begin to create a confidence interval for the population mean age of US pennies in 2011, we start by using specify() to choose which variable in our pennies_sample data we'd like to work with. This can be done in one of two ways:

'test'