Chapter 4 Lab: Statistical Inference

Sampling from Millbrae, California

In this lab, we'll investigate the ways in which the estimates that we make based on a random sample of data can inform us about what the population might look like. We're interested in formulating a *sampling distribution* of our estimate in order to get a sense of how good of an estimate it might be.

The Data

The dataset that we'll be considering comes from the town of Millbrae, California, near San Francisco. The U.S Census Bureau has recorded information on all 20,718 residents of Millbrae, including age and household income. All residents of Millbrae represent our statistical population. In this lab we would like to learn as much as we can about the residents by taking smaller samples from the full population. Let's load the data.

```
set.seed(341)
data(millbrae)
```

We see that two vectors are loaded into the workspace: `age`, which contains the ages of all 20,718 residents of Millbrae and `income`, which contains the incomes for all 500 households to which those residents belong. For now, we'll focus on `age`. Let's look at the distribution of ages in Millbrae by calculating some summary statistics and making a histogram.

```
summary(ages)
hist(ages)
```

QUESTION 1: How would you describe this population distribution?

The Unknown Sampling Distribution

In this lab, we have access to the entire population, but this is rarely the case in real life. Gathering information on an entire population is often extremely costly or even impossible. Because of this, we often take a smaller sample survey of the population and use that to make educated guesses about the properties of the population.

If we were interested in estimating the mean age in Millbrae based on a sample, we can use the following command to survey the population.

```
samp1 <- sample(ages,75)
```

This command allows us to create a new vector called `samp1` that is a simple random sample of size 75 from the population vector `ages`. At a conceptual level, you can imagine randomly choosing 75 names from the Millbrae phonebook, calling them up, and recording their ages. You would be correct in objecting that the phonebook probably doesn't contain all of the residents and that there will almost certainly be people that don't pick up the phone or refuse to give their age. These are issues that can make gathering data very difficult and are a strong incentive to collect a high quality sample.

QUESTION 2: How would you describe the distribution of this sample? How does it compare to the distribution of the population?

If we're interested in estimating the average age of all the residents in Millbrae, our best guess is going to be the sample mean from this simple random sample.

```
mean(samp1)
```

Our estimate of the mean is 40.17, which is just a bit below the true population mean of 42.29. So our sample mean turns out to be a pretty good estimate of the average age, and we were able to get it by sampling less than 1% of the population.

QUESTION 3: Take a second sample, also of size 75, and call it `samp2`. How does the mean of `samp2` compare with the mean of `samp1`? If we took a third sample of size 150, intuitively would you expect the sample mean to be a better or worse estimate of the population mean?

Not surprisingly, every time we take another random sample, we get a different sample mean. It's useful to get a sense of just how much variability we should expect when estimating the population mean this way. This is what is captured by the *sampling distribution*. Because we have access to the population, we can build up the sampling distribution for the sample mean by repeating the above steps 5000 times.

```
sample.means <- rep(0,5000)

for(i in 1:5000){
    samp <- sample(ages,75)
    sample.means[i] <- mean(samp)
    }

hist(sample.means, freq = FALSE)
```

Here we rely on the computational ability of R to quickly take 5000 samples of size 75 from the population, compute each of those sample means, and store them in a vector called `sample.means`.

QUESTION 4: How many elements are there in `sample.means`? How would you describe this sampling distribution? On what value is it centered? Would you expect the distribution to change if we instead collected 50,000 sample means?

Approximating the Sampling Distribution

The sampling distribution that we just computed tells us everything that we would hope for about the average age of the residents of Millbrae. Because the sample mean is an unbiased estimator, the sampling distribution is centered at the true average age of the the population and the spread of the distribution indicates how much variability is induced by sampling only 75 of the residents.

We computed the sampling distribution for mean age by drawing 5000 samples from the population and calculating 5000 sample means. This was only possible because we had access to the population. In most cases you don't (if you did, there would be no need to estimate!). Therefore, you have only your single sample to rely upon ... that, and the Central Limit Theorem.

The Central Limit Theorem states that, under certain conditions, the sample mean follows a normal distribution. This allows us to make the inferential leap from our single sample to the full sampling distribution that describes every possibly sample mean you might come across. But we need to look before we leap.

QUESTION 5: Does `samp1` meet the three conditions for the sample mean to be nearly normal, as described in section 4.4?

If the conditions are met, then we can find the approximate sampling distribution by plugging in our best estimate for the population mean and standard error: $\bar{x}$ and $s/\sqrt{n}$.

```
xbar <- mean(samp1)
se <- sd(samp1)/sqrt(75)
```

We can add a curve representing this approximation to our existing histogram using the command `lines`. The x-coordinates cover the range of the x in the histogram and the y-coordinates are the values that correspond to this particular normal distribution.

```
lines(x = seq(33,52,.1), y = dnorm(seq(33,52,.1),mean = xbar,sd = se))
```

We can see that the line does a decent job of tracing the histogram that we derived from having access to the population. In this case, our approximation based on the CLT is a good one.

Confidence Intervals

Return for a moment to the question that first motivated this lab: what is our best estimate of the mean age of the residents of Millbrae? If we were to take only one random sample, `sample1`, our best estimate would be $\bar{x}$. That serves as a good *point estimate* but it would be useful to also communicate how uncertain we are in that estimate. That is exactly the information that is captured by the sampling distribution that we just approximated. If we provided a histogram of the sampling distribution itself, all but the stasticians would throw up their hands in exhasperation (and rightfully so). Instead, we summarize it with two numbers that make up the *confidence interval*.

According to section 4.2.3, we can calculate a 95% confidence interval for a sample mean by adding and subtracting 1.96 standard errors to our point estimate.

```
hi <- xbar + 1.96*se
lo <- xbar - 1.96*se
abline(v = c(lo,hi), col = "red")
```

The last command adds to the plot two red lines representing the confidence interval. Our aim was to create an interval that would capture the mean age of the population, 42.29.

Though we can see on the plot that we were successfull, we can verify this using the following custom function.

```
oi.contains(lo, hi, mean(ages))
```

This function checks to see if the mean is between the `lo` and `hi` and returns either TRUE or FALSE. This will be a useful tool in the following section.

Interpreting a Confidence Interval

What exactly does it mean to be 95% confident and what exactly are we confident of? The goal is to get an accurate estimate of the average age of everyone in Millbrae, the population parameter, and after constructing a 95% confidence interval we say we are 95% confident that it contains this parameter. To understand the meaning of "95% confident" we need to consider how well this procedure works when conducted on many different samples.

In the following simulation, we return to the population, from which we will draw 100 different random samples. For each of the samples, we will calculate the sample mean and standard error and use those to form a 95% confidence interval. Then, we check to see if the interval contains the parameter and store that result in a vector called `results`. Keep in mind that because we are resampling from the population, this is a hypothetical experiment that usually wouldn't be possible in the real world.

```
hi <- rep(0,100)
lo <- rep(0,100)
results <- rep(0,100)

for(i in 1:100){
    samp <- sample(ages,75)
    xbar <- mean(samp)
    se <- sd(samp)/75
    hi[i] <- xbar + 1.96*se
    lo[i] <- xbar - 1.96*se
    results[i] <- oi.contains(lo[i], hi[i], mean(ages))
    }
```

Note that along the way, we are saving three pieces of information: the 100 results, the 100 upper bounds on the intervals (`hi`), and the 100 lower bounds on the intervals (`lo`). To visualize this output, we can plot the first 50 intervals that we have constructed using the following custom function.

```
oi.plot95ci(lo, hi, mean(ages))
```

Each of the 50 confidence intervals are represented by 50 horizontal lines centered on their 50 sample means (the black dots). The vertical line represents the population parameter. It's apparent that our sample means sometimes overestimate and sometimes underestimate the population mean, but that most interval estimate do manage to capture it. Any interval that fails to contain the population parameter is highlighted in red.

We can tabulate the full results from all 100 samples, keeping in mind that R stores TRUE as 1 and FALSE as 0.

```
results
sum(results)
```

QUESTION 6: How many of the 100 confidence intervals that you produced contain the population parameter? Repeat the simulation several more times and note the results. Using this experiment as a reference, come up with a definition of what we mean by "95% confident".

On Your Own

So far we have only focused on estimating the mean age of the residents of Millbrae. Now we'll try to estimate the mean household income. Use `set.seed(211)` then take a random sample of size 75 from `incomes`.

1. Using the sample, what is your best point estimate of the population mean?
2. Check the conditions for the sampling distribution of $\bar{x}_{income}$ to be nearly normal.
3. Regardless of whether or not the conditions were met, form an interval estimate - a 95% confidence interval - for the mean household income.
4. Since you have access to the population, compute the sampling distribution for $\bar{x}_{income}$ by taking 5000 samples from the population of size 75 and computing 5000 sample means. Describe this sampling distribution.
5. Explore the interpretation of a 95% confidence interval by resampling from the population to form a total of 10,000 intervals, each based on a sample of size 75. What proportion of the intervals capture the true parameter?
6. What is your conclusion about forming confidence intervals based on the normal model in this circumstance?