Live Exercise: Sampling births from US territories

Corinne Riddell

September 16, 2020

What we've covered so far

- Data manipulation
- Data visualization
- Measures of central tendency, measures of variation
- Distributions of a single variable
- Relationships between 2 quantitative or 2 categorical variables
- Dr. Kang Dufour's lecture on experimental and observational study design

What is next?

- Probability (Part II)
- Statistical Inference (Part III)

Today

- Learn how to take a random sample in R
- Introduce the idea of probability
- See how our estimate of probability is related to sample size

Description of the data

These data are births by place of occurrence for U.S. territories (American Samoa, Guam, N. Mariana Islands, Puerto Rico, and US Virgin Islands) from the year 2015.

This is a subset of the data downloaded from here. You can find more information about the data set here.

Data dictionary

Here is the data dictionary for this dataset:

Variable	Description
babyID	Unique identifier: row number
dbwt	Birth weight in Grams: 227-8165 grams
combgest	Combined gestation, in weeks: 17th to 47th week of gestation
sex	Assigned sex at birth: M (Male) or F (Female)
dob_mm	Birth month
cig_rec	If the mother reports smoking in any of the three trimesters of pregnancy she is classified as a smoker: (Y) Yes, (N) No, or (U) Unknown

Import the data into R

```r
library(tidyverse)
birth_data <- read_csv(file = "./data/L03_US-territories-births.csv") %>% select(-X1)
```

```r
head(birth_data)
```

```
## # A tibble: 6 x 6
##   babyID  dbwt combgest sex   dob_mm cig_rec
##    <dbl> <dbl>    <dbl> <chr>  <dbl> <chr>  
## 1      1  2977       37 M          1 N      
## 2      2  3191       41 M          1 Y      
## 3      3  1786       32 F          1 N      
## 4      4  4489       39 M          1 N      
## 5      5  3203       38 M          1 N      
## 6      6  3203       39 F          1 N
```

Objective

- Today's objective is to take a **sample** from the **population** of births.
- We will first calculate the **true** probability of an infant being born less that 5 lbs 8 ounces, or 2500 grams, which is the traditional cutoff used to classify an infant as low birthweight.
- In real life settings, we would not know the **true** probability, we would have to estimate it based on a sample. Ideally, we have a random sample from the population to make this estimate. Why?
- We will see how well we can estimate the true probability based on random samples of varying sizes.

Step 1: Add a variable to the dataset for low birthweight (LBW)

```r
birth_data <- birth_data %>% mutate(lbw = dbwt < 2500)
```

- What does the variable `lbw` store?
- `lbw` stores "logical" values, which means it is either equal to `TRUE` or `FALSE`
- Variables that take only two values are called **binary** variables. They are most commonly stored as logical data (TRUE/FALSE), numeric (0/1) or categorical ("Yes"/"No").

Step 2: Calculate the probability in overall population of the US territories

```r
lbw_population <- birth_data %>% summarize(true_prob_lbw = mean(lbw))
lbw_population
```

```
## # A tibble: 1 x 1
##   true_prob_lbw
##           <dbl>
## 1         0.102
```

- How did we take a mean of values equal to `TRUE` or `FALSE`? R treats `TRUE` as equivalent to `1` and `FALSE` as equivalent to `0`.
- The mean of a variable coded as 0 or 1 is the **proportion** of individuals with the low birth weight.
- Remember: we do not usually know the true value because we rarely have data on every individual in a sample.

Step 3: Take a random sample of size n=10

```r
random_sample_n10 <- birth_data %>%
  sample_n(size = 10) %>%
  mutate(sample_size = n()) #this adds the sample size to every row of the new data frame
```

Step 4: Progressively increase the sample size and store those samples

Sample the rows of data without replacement using the following sample sizes. Assign your samples each to a different object.

```
1. 10
2. 50
3. 100
4. 200
5. 500
6. 1000
7. 5000
8. 10000
9. 36724 (i.e, the entire target population)
```

Step 4 code

```r
random_sample_n50 <- birth_data %>% sample_n(size = 50) %>% mutate(sample_size = n())

random_sample_n100 <- birth_data %>% sample_n(size=100) %>% mutate(sample_size = n())

random_sample_n200 <- birth_data %>% sample_n(size=200) %>% mutate(sample_size = n())

random_sample_n500 <- birth_data %>% sample_n(size=500) %>% mutate(sample_size = n())

random_sample_n1000 <- birth_data %>% sample_n(size=1000) %>% mutate(sample_size = n())

random_sample_n5000 <- birth_data %>% sample_n(size=5000) %>% mutate(sample_size = n())

random_sample_n10000 <- birth_data %>% sample_n(size=10000) %>% mutate(sample_size = n())

whole_pop <- birth_data %>% sample_n(size = nrow(birth_data)) %>% mutate(sample_size = n())
```

Step 5: calculate the estimate of the proportion of LBW for each random sample

- For each sample, we want to estimate the proportion of `LBW` babies to see how much it differs from the true proportion in the entire population.
- We code do this using `summarize` for each sample and by writing that code ten times, but there is an easier way.

Step 5: code to estimate the proportions more efficiently

The function `bind_rows(df1, df2, df3, ...)` can be used to stack multiple data frames (e.g. df1, df2, df3, . . .) on top of wach other when they have each contain the same variables. Bind together the 9 data frames created in the previous code chunk using `bind_rows()`and assign the stacked data frame the name `stacked_samples`:

```r
stacked_samples <- bind_rows(random_sample_n10, random_sample_n50,
                             random_sample_n100, random_sample_n200,
                             random_sample_n500, random_sample_n1000,
                             random_sample_n5000, random_sample_n10000,
                             whole_pop)
```

Step 5: code to estimate the proportions more efficiently

Estimate the proportion of babies with low birthweight using each of your samples in `stacked_samples`. Hint: `group_by()` and `summarize()` will come in handy! Assign the output to a data frame called `sample_estimates`

```
sample_estimates <- stacked_samples %>%
  group_by(sample_size) %>%
  summarize(estimated_proportion_lbw = mean(lbw))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

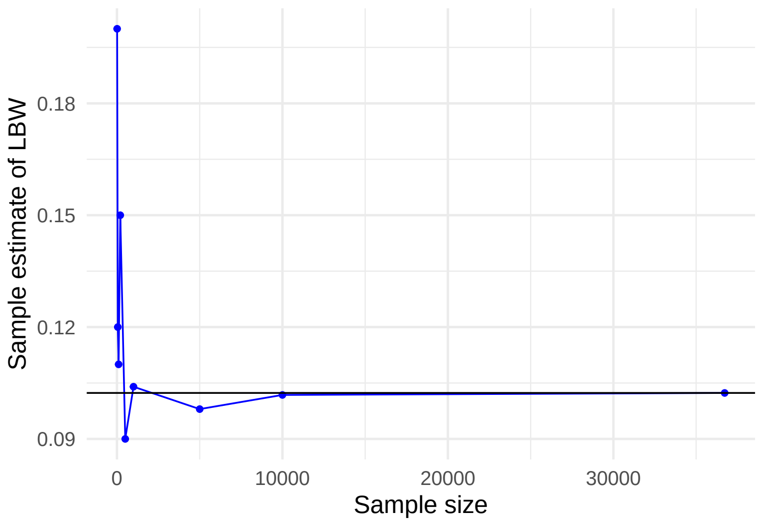
```
sample_estimates
```

```
## # A tibble: 9 x 2
##   sample_size estimated_proportion_lbw
##         <int>                    <dbl>
## 1          10                    0.2
## 2          50                    0.12
## 3         100                    0.11
## 4         200                    0.15
## 5         500                    0.09
## 6        1000                    0.104
## 7        5000                    0.098
## 8       10000                    0.102
## 9       36724                    0.102
```

Step 6: Visualize the results

- Make a line plot of the estimates of the probabilities versus the sample size.
- Add a horizontal line to the line plot at the true value that you are striving to estimate.
- You might also want to add points on top of the line to see exactly where the estimates are.

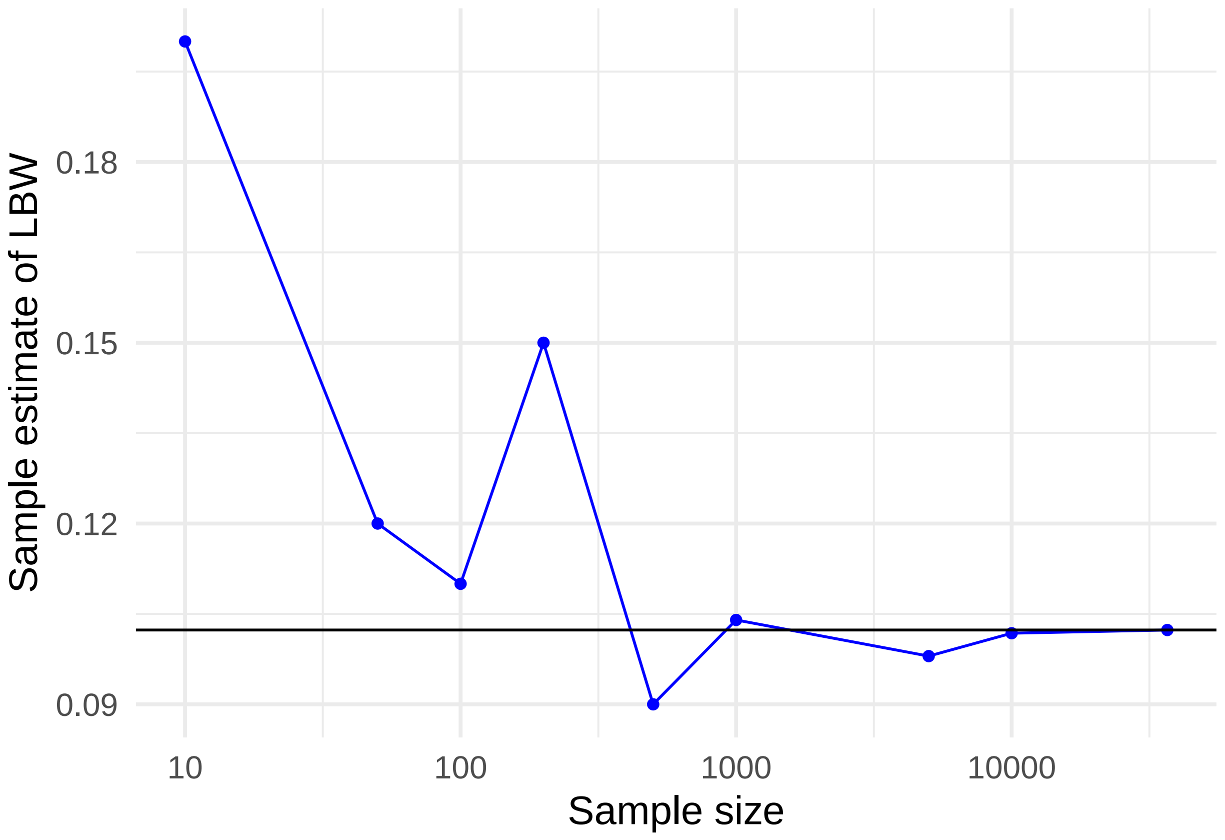
```
ggplot(sample_estimates, aes(x = sample_size, y = estimated_proportion_lbw)) +
  geom_line(col = "blue") +
  geom_point(col = "blue") +
  geom_hline(yintercept = lbw_population %>% pull(true_prob_lbw)) +
  #scale_x_log10() +
  labs(y = "Sample estimate of LBW", x = "Sample size") +
  theme_minimal(base_size = 15)
```



Step 6: Visualize the results

- Because the scale of the x axis is so large, try using `scale_x_log10` to convert the x scale to a logarithm.

```r
ggplot(sample_estimates, aes(x = sample_size, y = estimated_proportion_lbw)) +
  geom_line(col = "blue") +
  geom_point(col = "blue") +
  geom_hline(yintercept = lbw_population %>% pull(true_prob_lbw)) +
  scale_x_log10() +
  labs(y = "Sample estimate of LBW", x = "Sample size") +
  theme_minimal(base_size = 15)
```



Check your understanding!

1) What happens as sample size increases?
2) Will it **always** be the case that a higher sample size produces an estimate closer to the true value than a lower sample size?