

Lab 4: Sampling distribution

Not graded, just practice

Katie Schuler

2024-09-19

Practice your new stats skills with these practice exam questions! Best to open a fresh Google Colab notebook and test things out! Refer to the study guide to find answers as well.

Tip

More than one answer may be correct!

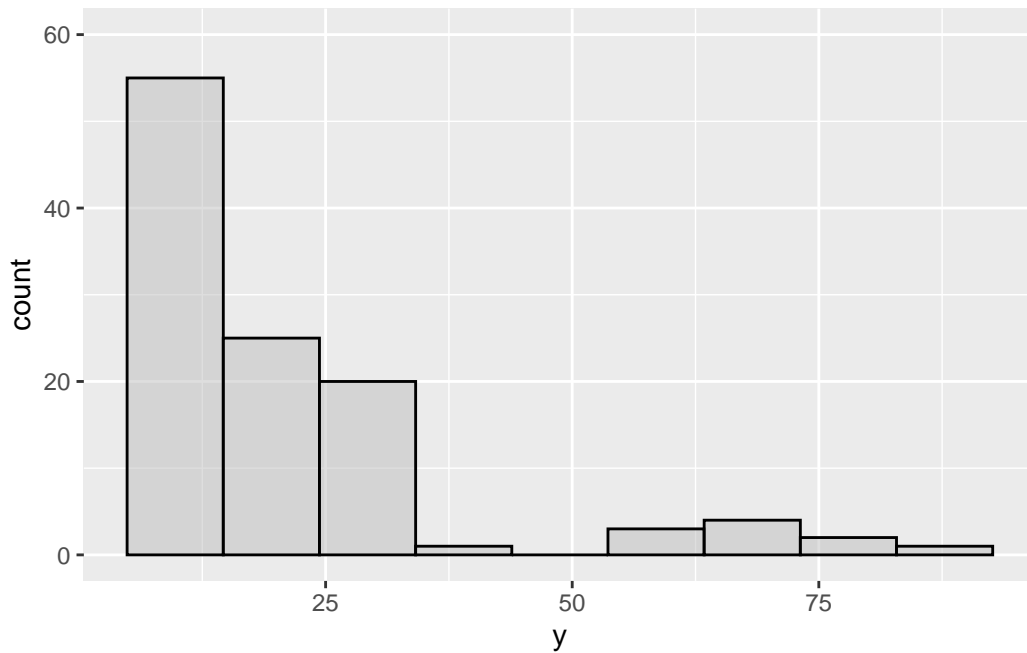
If you would like to practice with a set of data, you can import the following dataset with `read_csv`:

```
# brain volumes simulated from Ritchie et al  
"http://kathrynschuler.com/datasets/brain_volume.csv"
```

0.1 Descriptive statistics

- (a) Which of the following is the best choice to visualize the frequency distribution of a set of data? Choose one.
- (A) `geom_rug()`
 - (B) `geom_histogram()`
 - (C) `geom_point()`
 - (D) `geom_smooth()`
- (b) Which of the following would summarize the central tendency of a set of data? Choose all that apply.

- (A) mean
 - (B) median
 - (C) standard deviation
 - (D) inter quartile range (IQR)
- (c) Which of the following would summarize the spread of a set of data? Choose all that apply
- (A) mean
 - (B) median
 - (C) standard deviation
 - (D) inter quartile range (IQR)
- (d) Which of the following are paramteric statistics?
- (A) mean
 - (B) median
 - (C) standard deviation
 - (D) inter quartile range (IQR)
- (e) Given the following figure, which summary statistics would best describe these data?



- (A) mean
- (B) median
- (C) standard deviation
- (D) inter quartile range (IQR)

(f) Given the following code, which of the following would fill in the blank to return the value below which 20% of the data fall.

```
data %>% summarise(
  lower = quantile(y, _____)
)
```

- (A) 20
- (B) -20
- (C) below
- (D) 0.2

0.2 Probability distributions

- (a) Write code to generate 200 data points, sampled from a gaussian distribution with a mean of 0 and a standard deviation of 1.

```
rmnorm(200, mean = 0, sd = 1)
```

- (b) Suppose you sampled 500 data points from a uniform distribution and stored the result in `data`. Then, you use the following code to compute the summary stats. What is the height of the probability density function at a value of 5?

```
data %>% summarise(  
  n = n(),  
  mean = mean(y),  
  sd = sd(y),  
  lower = quantile(y, 0),  
  upper = quantile(y, 1)  
)  
  
# A tibble: 1 x 5  
      n mean    sd lower upper  
  <int> <dbl> <dbl> <dbl> <dbl>  
1   500  7.57  1.41  5.02  9.98
```

-
- (c) Suppose your data is normally distributed and has a mean of 25 and a standard deviation of 5. What is the probability a random value drawn from your dataset will be less than 20? Select the closest value.

- (A) 0.0483
- (B) 0.1589
- (C) 1
- (D) 0

0.3 Sampling variability

- (a) True or false, the `parameter` is the mean of the population and the `parameter estimate` is the mean of your sample?
- (A) TRUE
 - (B) FALSE
- (b) What do we call the probability distribution of the values our parameter estimate can take on?

-
- (c) Suppose we want to quantify the spread of the sampling distribution. What method could we choose? Choose all that apply.
- (A) mean
 - (B) median
 - (C) standard error
 - (D) confidence interval
- (d) For a typical experiment, how many samples from the population is practical for us to take? Enter a number.

0.4 Bootstrapping

- (a) True or false, when we generate the bootstrap sampling distribution, we sample our original sample *with replacement*.
- (A) TRUE
 - (B) FALSE
- (b) Suppose we want to generate the bootstrap sampling distribution for the mean of set of data, `data`, with one variable: `reaction_time`. Write code that uses the `infer` package to accomplish this, generating 1000 samples.

```
data %>%
  specify(response = reaction_time) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "mean")
```

- (c) Suppose we store our bootstrap sampling distribution from part b in a variable called `bootstrap_distribution`. Which two arguments should we add to the code below to compute the 68% confidence interval and assign it to the value `ci`?

```
ci <- bootstrap_distribution %>%
  get_confidence_interval(_____, _____)
```

- (A) `type="se", level = 68`
- (B) `type="se", level = 0.68`
- (C) `type="percentage", level = 0.68`
- (D) `type="percentage", level = 68`

- (d) Suppose we store our bootstrap sampling distribution in `bootstrap_distribution` and we want to visualize the confidence interval we just computed in c. Which of the following could we add to the code below? Choose all that apply.

```
bootstrap_distribution %>%
  visualize() +
  _____
```

- (A) `get_confidence_interval(endpoints = ci)`
- (B) `shade_ci(endpoints = ci)`
- (C) `shade_confidence_interval(endpoints = ci)`
- (D) `get_ci(endpoints = ci)`