Homework 8

Nora Quick

2021-11-20

R Question

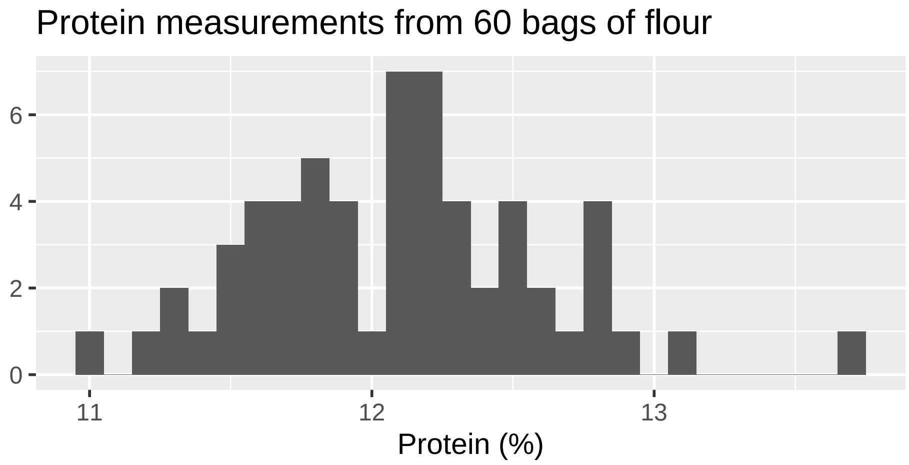
This question examines the bootstrap approach to sampling distributions, confidence intervals, and hypothesis testing.

Suppose you are embarking on a baking business. A key ingredient for you is flour, and of most importance is that the protein content has minimal variation. That is, for consistent baked products you need the protein levels to be consistent across bags. You sample and test 60 bags from your current favorite flour brand for their protein content (measured in %). The measurements are stored below in an object called `protein`:

```r
protein <- c(12.06, 11.16, 11.35, 11.89, 12.49, 12.19, 11.89, 12.47, 12.42,
11.57, 12.2, 11.04, 12.17, 12.82, 11.81, 11.86, 11.75, 11.82,
12.17, 11.63, 11.54, 12.76, 12.2, 12.13, 12.08, 12.56, 12.77,
13.12, 12.15, 12.07, 11.48, 11.61, 12.28, 12.38, 11.67, 11.67,
11.55, 12.16, 12.92, 11.85, 12.53, 12.29, 12.06, 12.06, 12.01,
12.81, 11.78, 11.66, 11.4, 12.33, 12.21, 11.93, 12.71, 11.65,
12.32, 12.52, 11.84, 12.56, 13.72, 11.29)
```

A histogram of these measurements is shown below:

```r
qplot(protein, binwidth = 0.1) +
  labs(
    title = "Protein measurements from 60 bags of flour",
    x = "Protein (%)"
  )
```



a) Calculate a point estimate for the **standard deviation of protein content** for the population of bags of this brand of flour, and give a one sentence summary of your result in context of the data.

```
sd <- sd(protein, na.rm=FALSE)
sd
```

```
## [1] 0.504584
```

There is about a half a percent (0.5%) difference in the variance between all 60 bags of flour.

b) Fill in the body of the function `boot_sd()` that takes a bootstrap sample of size 60 from the the input `x`; and calculates and returns the standard deviation of the bootstrap sample.

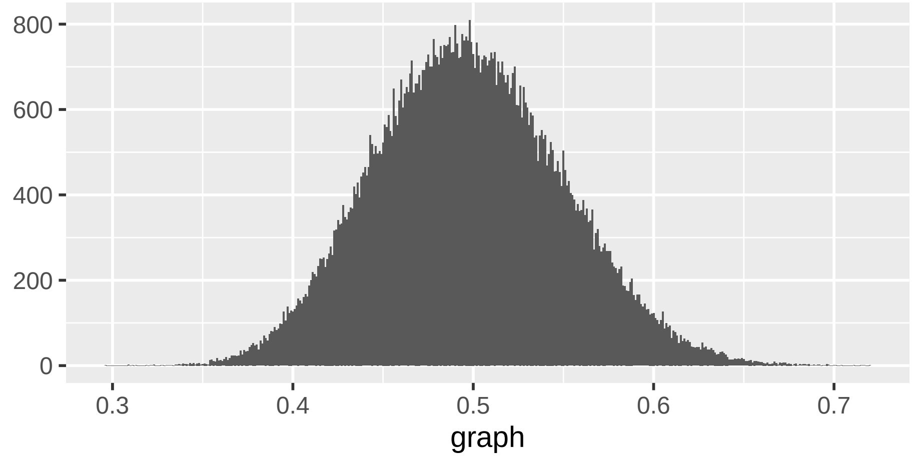
```
#Legnth of protein should be 60

boot_sd <- function(x){
  l <- length(x)
  s <- sample(x, l, replace = TRUE)
  sd(s)
}
```

```
# Verify your function returns a single value
boot_sd(protein)
```

```
## [1] 0.4873682
```

c) Now use your function with `replicate()`, to produce 100,000 standard deviations based on bootstrap samples. Make a histogram of the results with `qplot()`.

```
graph <- replicate(100000, boot_sd(protein))
qplot(graph, binwidth = 0.001)
```



d) Using your 100,000 standard deviations from (c), find the 95% bootstrap confidence interval for the population standard deviation using the percentile method with `quantile()`. One of the arguments for `quantile()` is `probs`, which allows you to specify the empirical quantiles you want returned. See Lecture 6 (Module 8) to review the percentile method.

```r
quantile(graph, probs = c(0.95))
```

```
##       95%
## 0.5858208
```

e) Use bootstrapping to test the following null and alternative hypotheses at the 5% level.

$$H_0 : \sigma = 0.5$$

$$H_A : \sigma \neq 0.5$$

In two sentences at most, what do we conclude and why?

We can conclude that the null hypothesis is true. We can conclude this because we found very little variance in the protein in the flour due to the sd being around 0.5 which is the value which we wanted it to be to conclude that the null hypothesis is true.

Conceptual Questions

1. A marketer is researching the quality of suitcases. She selects two suitcase models. One of the models is extremely expensive, so she is only able to purchase three of them. The other model is fairly cheap, so she is able to purchase fifteen of them. She has a chart describing various suitcase flaws, and she looks at each of the selected suitcases and counts the number of flaws matching the chart. The results are as follows:

Flaws in expensive models: 0, 0, 0
Flaws in cheaper models: 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 2

She would like to compare the quality of the two models by comparing the the difference in mean number of flaws. **Explain why a permutation test might be more appropriate than a two sample t-test in this example.**

A permutation test would be more appropriate in this situation because of the difference in sample sizes. It is similar to the o-ring examples from lecture and lab. If we have something with a much smaller sample size we prefer a permutation test.

2. A state park ranger would like to test the hypothesis that campers spend about a typical time of 30 hours at the state park at which he works. He collects camping information from a random sample of ten campers. Specifically, he asks them how long they are planning on spending at the state park, and then converts these times to hours. He obtains the following information:

Camping times: 36, 33, 12, 36, 23, 56, 34, 35, 31, 26

NULL Hyp:

$$H_0 : \sigma = 30$$

ALT Hyp:

$$H_A : \sigma \neq 30$$

Use the signed-rank test to test the hypothesis using R and state a conclusion in the context of the problem.

```r
new_camping_times <- c(36, 33, 12, 23, 56, 34, 35, 31, 26)
wilcox.test(x = new_camping_times, y = NULL, paired = FALSE, conf.level = 0.95)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  new_camping_times
## V = 45, p-value = 0.003906
## alternative hypothesis: true location is not equal to 0
```

Given the p-value it is likely that the NULL hypothesis is true. In other words the state park ranger is correct in believing that the average time spent at the parks is about 30 hours.

(The additional 36 was gotten rid of because we can't have ties in the data. With some googling it appears that we simply want to remove ties from the data if they exist.)

3. A grass seed company is inspecting bags of grass seed for holes in the bags for two different types of bags, A and B. They would like to know if the spreads of numbers of holes for the two bag types are different. They take a random sample of 50 bags of type A and type B and count the number of holes they find in the bags. They then perform Levene's test and obtain a p-value of 0.03. Assuming they are testing at the 5% level and all necessary assumptions are met, **state a conclusion in the context of the problem.**

Due to the p-value (0.03) being smaller than 0.05 we will assume that there is enough varriance in the holes of the different bag types. In other words there is another reason for varriance other than randomization. We will reject the null hypothesis that they have the same number of holes for the alternative hypothesis that they don't have the same number of wholes.