Homework 8

Rui Gao

2021-11-21

R Question

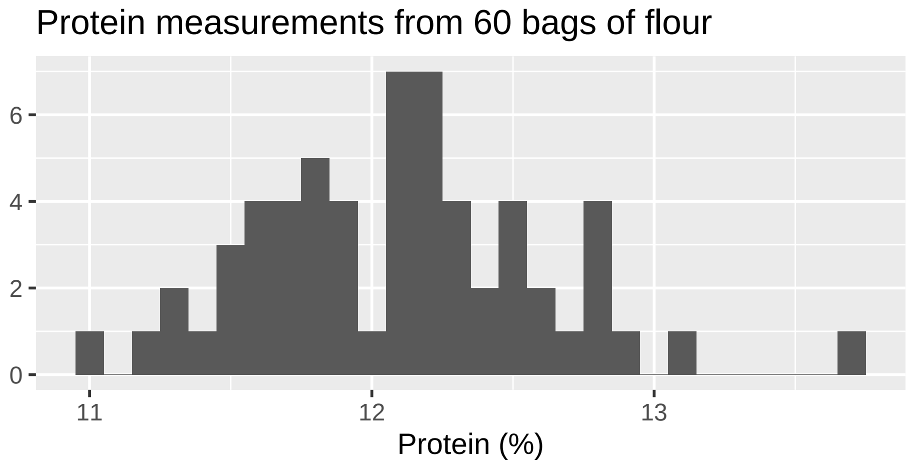
This question examines the bootstrap approach to sampling distributions, confidence intervals, and hypothesis testing.

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`:

```
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:

```
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_r <- sd(protein)
sd_r
```

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

The difference of protein containing of this brand's flour is about 0.5%, which means the quality of the product is reliable.

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.

```
boot_sd <- function(x){
  n <- length(protein)
  sample_r <- sample(x, n, replace = TRUE)
  sd(sample_r)
}
```

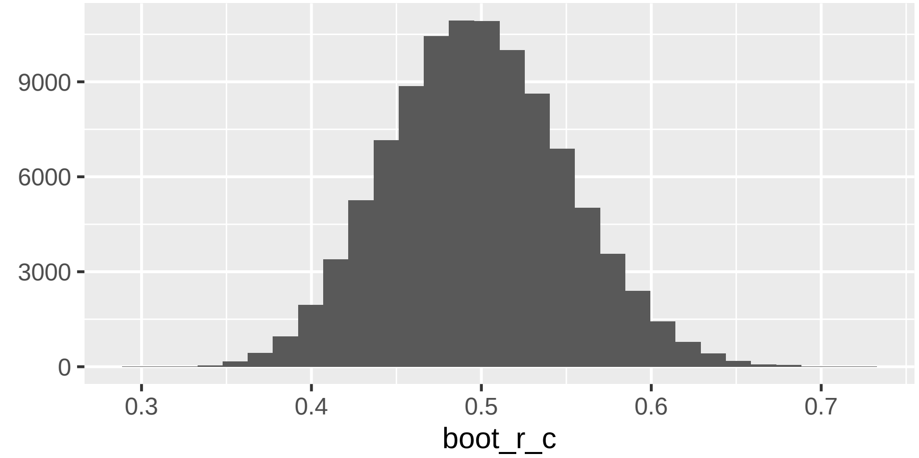
```
# Verify your function returns a single value
boot_sd(protein)
```

```
## [1] 0.4965788
```

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()`.

```
boot_r_c <- replicate(100000,boot_sd(protein))
qplot(boot_r_c)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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.

```
  #
quantile(boot_r_c, probs = seq(0,1,0.05))
```

```
##        0%        5%       10%       15%       20%       25%       30%       35% 
## 0.2992588 0.4142313 0.4308916 0.4428963 0.4525853 0.4609698 0.4687255 0.4758153 
##       40%       45%       50%       55%       60%       65%       70%       75% 
## 0.4827475 0.4895975 0.4962782 0.5029213 0.5098452 0.5169380 0.5245387 0.5327151 
##       80%       85%       90%       95%      100% 
## 0.5417775 0.5523364 0.5660154 0.5866234 0.7284496
```

I guess it's between (0.414, 0.5867)

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 would faile to reject the null hypothesis at the 0.05 level, since 0.5 is between the interval.

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.** The number of samples is small, the number are different in two groups, and the data maybe not come from normal populations

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

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

```
c_0 <- 30
d <- camping_times-c_0
d_c <- 30-d #distance form c_0
d_c
```

```
##  [1] 24 27 48 24 37  4 26 25 29 34
```

The rank's numbers that greater than 30 are 7,8,9

```
s <- 24

z <- (s-9*(9+1)/4)/(sqrt((9*10*19))/24)
z
```

```
## [1] 0.8705715
```

```
p <- 2*(1-pnorm(abs(z)))
p
```

```
## [1] 0.3839882
```

The ranger maybe is correct. But the time they stay in this place maybe influenced by each other and the distribution of the population maybe is not symmetric.

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.** They may have enough evidence to say the two bags are different by reject the null hyphothesis.But the two populations spreads of A and B maybe are not the same or in an acceptable level.