Homework 4

Data Science Team

Due Tuesday, July 26, 2022 at 10am

1. The sampling distribution of the sample mean

In this exercise we will explore the two main ways for approximating the sampling distribution of a statistic: the central limit theorem (CLT) and the bootstrap. The central limit theorem tells us that the sample mean $\bar{x}$ based on n observations from a population with mean μ and standard deviation σ has approximately a $N(\mu, \sigma/\sqrt{n})$ distribution.

a. Load the flight data from Homework 2.

b. Suppose we are going to take a small sample of n of these flights to try to estimate the mean of x=`arr_delay`. Since you have the entire population, you can compute μ and σ. Find μ and σ, ignoring the missing values by using the option `na.rm=TRUE`. Now state what the CLT says the sampling distribution of $\bar{x}$ is for a sample of size $n = 100$.

c. Now take **one** sample of size 100. What are your estimates of μ and σ based on this single sample?

d. The CLT employs the standard error $\sigma/\sqrt{n}$. But if we have only a sample (and not the whole population) then σ is unknown and needs to be estimated with $\hat{\sigma}$ from c. What is the resulting standard error?

e. Use the bootstrap with 10,000 replicates to obtain an estimate of the standard error of $\bar{x}$, based on the one sample from c. How does this estimate compare to the standard error employed by the CLT, i.e. the answer in d?

f. Since in this case we have the whole population (and not just a sample), we can simulate the sampling distribution of $\bar{x}$. Draw 1,000 samples of size 100 and compute the resulting 1,000 sample means. Does a histogram of these suggest that the sampling distribution is approximately normal as stated by the CLT? Find out what the function `qqnorm` does and use it to assess normality.

g. Now use the bootstrap to approximate the sampling distribution: Draw 1,000 bootstrap samples based on your one sample from c. Make a histogram of the resulting 1,000 bootstrap sample means. Comparing the shape of the histograms in f and g, does it appear that the bootstrap provides a better approximation to the shape of the sampling distribution than the CLT?

2. Bootstrapping when the sampling distribution isn't normal

This exercise will have you work with the Claridge data:

```
library(boot)
data("claridge")
```

This exercise concerns the **sample correlation** of the variables in this dataset. The sample correlation is a measure of how **dependent** two variables are. You can calculate it using the `cor` function. We will take the perspective that the claridge data is a sample from a much larger population that has **population correlation** ρ, and we will use the sample correlation as an estimate of ρ.

a. Call the two variables in the claridge dataset x and y. Compute their sample correlation.

b. Now we want to form an interval that gives us a sense of how much variability there is for this value in repeated sampling. Take 10,000 bootstrap resamples of the claridge data and recompute the sample correlation each time. (Note: when you do this, resample with replacement **entire rows** of the data. . . if you resample x and y independently, then the correlation calculation will be totally wrong. Why?)

c. The central limit theorem doesn't always apply, and when it does, the sample size required for it to be a good approximation can vary a lot depending on the statistic being computed. Suppose for a moment that the CLT does apply here and is a good approximation for the sampling distribution of the sample correlation. Use the bootstrap standard error to form a CLT 95 percent interval for the population correlation.

d. You know another way to find a confidence interval using the bootstrap **quantiles**. Get a 95 percent confidence interval this way. How does it compare to your answer in c?

e. Make a histogram of the bootstrap samples. Does it look normal? If not, which of the two intervals do you trust more and why?

Bootstrap: True or False

(a) Let $X_1, ..., X_{50}$ be independent draws from $N(\mu, \sigma)$ (i.e. Normal distribution with mean mu and standard deviation σ) and let the sample mean be $\bar{X}$. True or false: $\bar{X}$ is likely to be off μ by something like $\sigma/\sqrt{50}$, just due to random error.

For (b) - (d): Now let $X_{i(k)}$ be independent draws from $N(\mu, \sigma)$, for $i = 1, ..., 50$ and $k = 1, ..., 100$. Let $\bar{X}_{(k)} = \frac{1}{50}\sum_{i=1}^{50} X_{i(k)}$, $s^2_{(k)} = \frac{1}{50}\sum_{i=1}^{50}[X_{i(k)} - \bar{X}_{(k)}]^2$, $\bar{X}_{\text{ave}} = \frac{1}{100}\sum_{k=1}^{100}\bar{X}_{(k)}$, $V = \frac{1}{100}\sum_{k=1}^{100}[\bar{X}_{(k)} - \bar{X}_{\text{ave}}]^2$.

For each of the following, indicate whether it is true and false and briefly explain.

(b) $\{\bar{X}_{(k)} : k = 1, ..., 100\}$ is a sample of size 100 from $N(\mu, \sigma/\sqrt{50})$.

(c) $|\bar{X}_{(k)} - \bar{X}_{\text{ave}}| < 2\sqrt{V}$ for about 95 of the k's.

(d) $\bar{X}_{\text{ave}}$ is $N(\mu, \sigma/\sqrt{5000})$.