Exercise 1

PLSCI 7201

9/1/2020

Summary

The purpose of these exercises is to impliment some of the approaches we discussed in the lecture and to get an idea of your proficiency in R. I've provided a few examples of code that can be modified to answer each question. Feel free to come up with your own code to answer the questions. Make sure your analyses are reproducible.

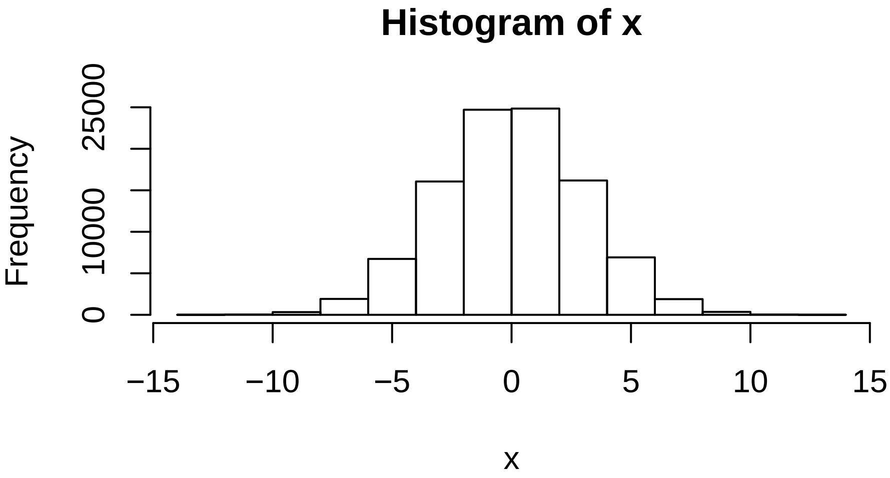
Pt 1. Statistical inference

A large group of researchers were interested in studying plant height. They randomly sampled 160,000 observations for a given species. Since this is a very large sample size, we can assume that this data is the "population" discussed in class. In theory, this might not be a great assumption, but for the purposes of the following exercises it is likely fine.

- **Q1:** Plot histogram of the population height. Does it appear to be normally distributed? Why or why not?

Example: A histogram can be plotted using the `hist()` function in R. The function accepts a numeric vector. For example:

```
set.seed(991837)
# rnorm will generate a random set of 100000 numbers from a Normal distribution
# with mean = 0 and standard deviation = 3
x <- rnorm(1e+05, 0, 3)
hist(x)
```

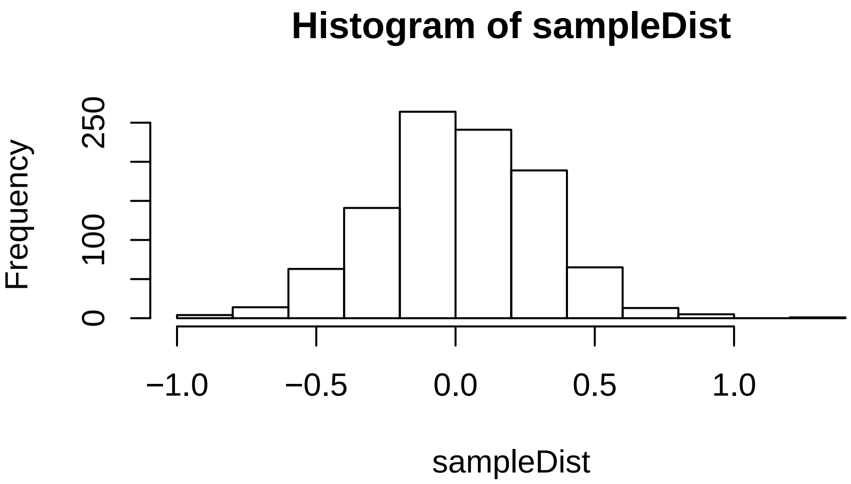


Generate a sampling distribution for the mean with a sample size $n = 100$

The distribution should be composed of 1000 draws of size n. Compute the mean of the sampling distribution. Be sure to use `set.seed()` to ensure that your work is reproducible.

Example: For example, we can set up a loop to sample from the vector `x` created above.

```
sampleDist <- NULL  # creates an empty object to store the results

# a for loop will iteratively run the same code over some index i
for (i in 1:1000) {

    # The sample function will sample (at random) a set of values from a vector
    # specified by the argument x in the function. In this case I am using it to
    # create a random index from 1:100000 to sample from the vector of observations.
    # The set.seed arguement removes the randomness of the sampling process. So, if
    # we use the same value for set.seed then we will always sample the same numbers.

    set.seed(i)  # i is the current iteration of the loop
    tmp.x <- x[sample(x = 1:1e+05, size = 100, replace = F)]  # sample from the population

    # There are two potentially new functions used here. mean() computes the mean.
    # c() is the combine function which is used to combine two vectors. So, here
    # we're basically appending the vector sampleDist at each iteration
    sampleDist <- c(sampleDist, mean(tmp.x))
}

hist(sampleDist)
```



```
mean(sampleDist)
```

```
[1] 0.01509066
```

- **Q2:** Modify the code above (or write your own code) to generate a sampling distribution for the mean. We have 160000 observations recorded for the population. Report the mean computed from the sampling distribution. How does this value compare with the population mean?

Hypothesis testing - Z-test

We can assume that the mean of the sampling distribution created above should be very close to the population mean. Here we will randomly select a sample of 50 observations ($n = 50$) from the sample, compute the mean, and perform a Z-test to determine if the population mean is different from our estimate. Use $\sigma = 23.91$.

Example: The code below shows how we can compute the Z-statistic and the p-values associated with the statistic. I will continue with our population defined by the vector `x`. In this case I know $\mu = 0$ and $\sigma = 3$.

```r
set.seed(1834)
tmp.x <- x[sample(1:1e+05, size = 50, replace = F)]
xbar = mean(tmp.x)
xbar
```

```
[1] -0.2039231
```

```r
xbar = 0.06043461

pop.sd = 3
pop.mu = 0

Zstat = (xbar - pop.mu)/(pop.sd/sqrt(50))
Zstat
```

```
[1] 0.1424457
```

```r
# We can compute the p values for the Z-stat using the pnorm function. Since the
# distribution is symmetrical and were doing a two tailed test the p value is
# twice the area under the curve.
pnorm(q = Zstat, mean = 0, sd = 1, lower.tail = F) * 2
```

```
[1] 0.8867279
```

- **Q3:** What is the p-value from the Z-test? What does this value mean? Do you think there is a difference?

Pt 2. Bootstrap resampling

Continuing with the sample $n = 50$ that we used above, lets compute the standard error of the sample mean using bootstrap resampling. We know that the SE of the sample mean is $\frac{s}{\sqrt{n}}$, so we'll compare our bootstrap estimates to this value. Perform 1000 bootstrap iterations and sample the data with replacement (we won't select a subset since the data is not very large).

Example: Again, we can do this pretty simply using a for loop. Using the toy example `x`.

```
set.seed(1834)
tmp.x <- x[sample(1:1e+05, size = 50, replace = F)]
xbar = mean(tmp.x)
xbar
```

```
[1] -0.2039231
```

```
# Create a matrix to store the boot strap samples. Each column will be one
# iteration.  length() returns the length of a vector
boot.samples <- matrix(NA, ncol = 1000, nrow = length(tmp.x))

for (i in 1:1000) {
    set.seed(12 + i)
    # Sample and replace each column of the matrix witht the bootstrap sample.
    boot.samples[, i] <- tmp.x[sample(1:length(tmp.x), size = length(tmp.x), replace = T)]
}

# Get the means of each bootstrap sample. colmeans will compute the means of each
# column in the matrix
boot.means.i <- colMeans(boot.samples)

# The bootstrap standard error is just the standard deviation of the bootstrap
# means
boot.se <- sd(boot.means.i)
boot.se
```

```
[1] 0.4851917
```

```
sd(tmp.x)/sqrt(length(tmp.x))
```

```
[1] 0.491944
```

- **Q4:** Report the values for the bootstrap standard error and the SE obtained using $\frac{s}{\sqrt{n}}$? Is the bootstrap standard error reliable?

Pt 3. Hypothesis testing - one sample t-test.

As a new researcher you are interested in determing whether the height of a specific group of plants is different from the population mean. In this example, you have no information on the population. However a recent study evaluated plant height for a very large population ($n = 160000$) and reported an average height of 60 cm ($\mu = 60$). You recorded height for 18 plants. Perform a one sample t-test using the data provided and report the p-value for a two tailed test. Your data for these 18 plants is provided as Q3_sampleHt.Rds.

- Report the p-value and the t-statistic. What do each of these mean? Do you think the average height of the group is different from the the population mean?