Lab Assignment IV

Saba Dzmanashvili

11/16/2021

## Sampling distribution and Inferential Statistis

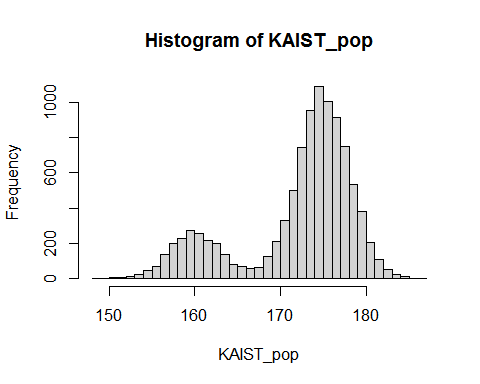
In this lab, we will practice once more to understand how sampling distribution works step-by-step. Let’s say our population size is 10k and we measure the height of KAIST students. The proportion of man is very high in KAIST, which will make the height distribution bimodal.

Let’s create a hypothetical KAIST students height distribution.

KAIST\_pop <- c(rnorm(8000, 175, 3), rnorm(2000, 160, 3))   
set.seed(1001)

**Question 1:** Check its distribution using a histogram and describe its shape.

hist(KAIST\_pop, 40)



*Shape of* KAIST\_pop *looks like two normal distributions merged together, one flat and another tight. It has bimodal shape*

**Question 2:** Check the population mean and standard deviation.

**Answer:**

population\_mean = mean(KAIST\_pop) # Mean of KAIST population  
population\_sd = sd(KAIST\_pop) # Standard deviation of KAIST population  
  
print(population\_mean)

## [1] 172.0239

print(population\_sd)

## [1] 6.713577

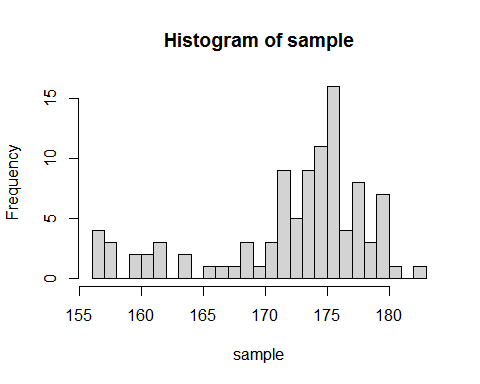
Now, sample 100 students randomly from the population.

n = 100  
sample <- sample(KAIST\_pop, n, replace = FALSE)

**Question 3:** Again, check its distribution using a histogram, and compute mean and sd, then compare them to your population distribution. If you’re using RMarkdown and the composition of sample keeps changing whenever you run the code due to its randomness, you can set the random seed by using “set.seed(ANY NUMBERS)”

**Answer:**

hist(sample,20)



mean(sample) # Mean of sample population

## [1] 172.1447

sd(sample) # Standard deviation of sample population

## [1] 6.387715

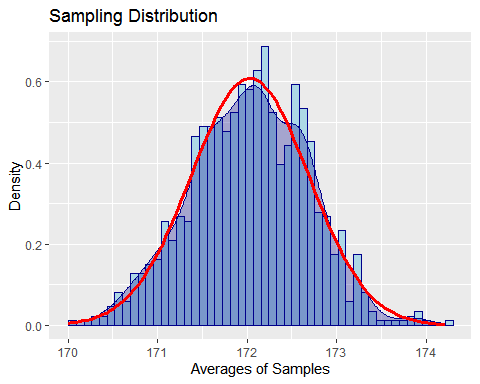
Now, let’s repeat the same procedure for 1,000 times, and save their average.

sampling\_distribution\_n100 <- NULL # Create a vacant vector where we will save sampled means.  
  
for (i in 1:1000) {  
 sample\_mean <- mean(sample(KAIST\_pop, n, replace = FALSE))  
 sampling\_distribution\_n100 <- c(sampling\_distribution\_n100, sample\_mean)  
}

**Question 4:** Check the distribution of sampling distribution by using a histogram, and compute a mean and standard deviation.

**Answer:**

library(ggplot2)  
sampling\_mean = mean(sampling\_distribution\_n100) # Mean of sampling distribution  
sampling\_sd = sd(sampling\_distribution\_n100) # Standard Deviation of Sampling Distribution  
  
ggplot(data.frame(x = sampling\_distribution\_n100), aes(x)) +  
 geom\_histogram(bins=50,aes(y=..density..), color="darkblue", fill="lightblue") +   
 geom\_density(alpha=.3, fill="darkblue",col='darkblue',size=0.7) +   
 stat\_function(fun = dnorm, args = list(mean = sampling\_mean,   
 sd = sampling\_sd), colour ='red', size =1.2) +   
 labs(x = "Averages of Samples", y = "Density", title="Sampling Distribution")



print(sampling\_mean)

## [1] 172.0317

print(sampling\_sd)

## [1] 0.6560506

*Red line is the normal distribution with the mean and standard deviation of* sampling\_distribution\_n100

(continued) Describe the shape of distribution and compare it to the distribution from population. Do they look similar or not, and why?

*The sampling distribution looks similar to the normal distribution and it differs lot from original KAIST population distribution. That’s because KAIST population distribution was bimodal, and sampling distribution is unimodal. They differ, because from CLT (Central Limit Theorem) sampling distribution will be approximated to the normal distribution as long as the sample size is large.*

(continued) Describe the mean and standard deviation. Is it similar to population mean and standard deviation?

*Mean of sampling distribution and the KAIST population are very close to each other. The difference is:*

print(population\_mean) # Mean of KAIST population

## [1] 172.0239

print(sampling\_mean) # Mean of sampling distribution

## [1] 172.0317

*Standard deviation of each distribution differ and they should differ by factor of sqrt(n), where n is the sample size:*

sd(KAIST\_pop) / sqrt(n) # Standard Deviation of KAIST population

## [1] 0.6713577

sd(sampling\_distribution\_n100) # Standard Deviation of Sampling Distribution

## [1] 0.6560506

## Confidence Intervals

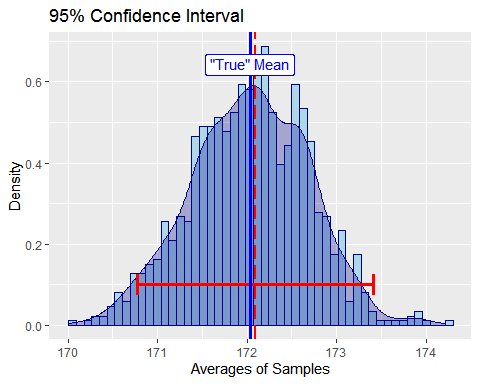
Let’s again sample with size 100.

sample\_n100 <- sample(KAIST\_pop, n, replace = FALSE)

**Question 5:** Compute the mean and standard ‘error’. Then, compute the 95% confidence interval. Explain the meaning of the confidence interval in English sentences.

**Answer:**

sample\_mean = mean(sample\_n100)  
  
standard\_err = sd(KAIST\_pop) / sqrt(n)  
margin\_of\_error = 1.96\*standard\_err  
  
confidence\_interfal\_left\_bound = sample\_mean - margin\_of\_error  
confidence\_interval\_right\_bound = sample\_mean + margin\_of\_error  
data\_sample = data.frame(y=0.1, x=c(confidence\_interfal\_left\_bound, confidence\_interval\_right\_bound))  
  
ggplot(data.frame(x = sampling\_distribution\_n100), aes(x)) +  
 geom\_histogram(bins=50,aes(y=..density..), color="darkblue", fill="lightblue") +   
 geom\_density(alpha=.3, fill="darkblue",col="darkblue", size=0.7) +   
 geom\_vline(xintercept = sample\_mean, color = "red", linetype = "longdash", size=1) +   
 geom\_errorbarh(aes(y=y, xmin=x[1], xmax=x[2],height = .05), data\_sample, col='red', size=1.2) +   
 geom\_vline(xintercept = sampling\_mean, color = "blue", size=1.2) +  
 annotate(x=sampling\_mean,y=+Inf,label="\"True\" Mean",vjust=2,geom="label", col='blue') +   
 labs(x = "Averages of Samples", y = "Density", title="95% Confidence Interval")



print(sample\_mean) # Mean of sample

## [1] 172.0898

print(standard\_err) # Standard Error

## [1] 0.6713577

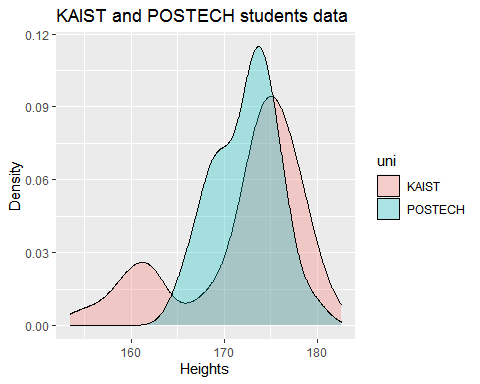
*The 95% confidence interval of sample is the the interval for which everytime we are going to draw a sample from the distribution and calculate its mean, 95% of the time 95% of the time the calculate mean will be in that exact interval. That’s why we are calling it confidence interval and give the estamitation as well.*

**Question 6:** Does your confidence interval cover the ‘true’ mean? What does that mean?

**Answer:** *It covers the “true” mean. As explained above, “true” mean is the original unknown mean of the distribution from which the sample was drawn. We can’t find out was is EXACTLY that mean, but from central limit theorem we can always get the value close to original one as close as possible. And this approximation level depends on how many times we want do traw the samples. The more samples, the better we know the “true” value.*

Postech students randomly sample 30 students and find out that their average is 172.8..

set.seed(1005)  
sample\_postech <- rnorm(30, 173, 3)  
  
KAIST\_heights <- data.frame(height = sample\_n100)  
POSTECH\_heights <- data.frame(height = sample\_postech)  
  
KAIST\_heights$uni <- "KAIST"  
POSTECH\_heights$uni <- "POSTECH"  
  
heights <- rbind(KAIST\_heights,POSTECH\_heights)  
  
ggplot(heights, aes(height, fill = uni)) +   
 geom\_density(alpha=.3) +   
 labs(x = "Heights", y = "Density", title="KAIST and POSTECH students data")



**Question 7:** Does this sample support the argument that Postech students are statistically significantly taller than KAIST students? Use a t-test and examine this hypothesis and interpret the results.

ttest = t.test(sample\_postech, sample\_n100)  
print(ttest)

##   
## Welch Two Sample t-test  
##   
## data: sample\_postech and sample\_n100  
## t = 0.23264, df = 97.358, p-value = 0.8165  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.595220 2.018875  
## sample estimates:  
## mean of x mean of y   
## 172.3017 172.0898

*Since p-value of T-test is*

## [1] 0.8165256

*and is greater than 0.05, therefore, we can NOT reject the hypothesis that Postech students are statistically significantly taller than KAIST students.*