Lab 6 Solution

Introduction to basic probability models Part II

July 9th, 2020

Contents

Watch the video for Session 6

Watch the video lecture associated to Session 6 (find the link on the website) – it will walk you through concepts on a) normal distribution, b) populations, samples, and estimates, and c) central limit theorem (CLT).

Populations, Samples, and Estimates Exercises

We will be using the following dataset:

```
library(downloader)
url <- "https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/mice_pheno.csv"
filename <- basename(url)
download(url, destfile=filename)
dat <- read.csv(filename)
```

Remove missing values:

```
dat <- na.omit(dat)
```

1. Use `dplyr` to create a vector x with the body weight of all males on the control (`chow`) diet. What is this population's average?

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# check each column and frequency
table(dat$Sex)
```

```
##
##   F   M
```

```
## 425 416
```

```
table(dat$Diet)
```

```
##
## chow   hf
##  448  393
```

```
x <- dat %>% filter(Sex == "M", Diet == "chow") %>% select(Bodyweight) %>% unlist
pop_avg_control <- mean(x)
pop_avg_control
```

```
## [1] 30.96381
```

2. Now use the `rafalib` package and use the `popsd` function to compute the population standard deviation.

```
# install.packages("rafalib")
library(rafalib)
popsd(x)
```

```
## [1] 4.420501
```

3. Set the seed at 1. Take a random sample X of size 25 from x. What is the sample average?

```
set.seed(1)
sample_avg_control <- mean(sample(x, size = 25, replace = FALSE))
sample_avg_control
```

```
## [1] 30.5196
```

4. Use `dplyr` to create a vector y with the body weight of all males on the high fat (`hf`) diet. What is this population's average?

```
y <- dat %>% filter(Sex == "M", Diet == "hf") %>% select(Bodyweight) %>% unlist
pop_avg_hf <- mean(y)
pop_avg_hf
```

```
## [1] 34.84793
```

5. Now use the `rafalib` package and use the `popsd` function to compute the population standard deviation.

```
popsd(y)
```

```
## [1] 5.574609
```

6. Set the seed at 1. Take a random sample Y of size 25 from y. What is the sample average?

```
set.seed(1)
sample_avg_hf <- mean(sample(y, size = 25, replace = FALSE))
sample_avg_hf
```

```
## [1] 35.8036
```

7. What is the difference in absolute value between $\bar{y} - \bar{x}$ and $\bar{X} - \bar{Y}$? Note: X and Y are the sampled vectors defined before, x and y represent the population vectors.

```
abs(sample_avg_hf - sample_avg_control)
```

```
## [1] 5.284
```

```
abs(pop_avg_hf - pop_avg_control)
```

```
## [1] 3.884116
```

8. Repeat the above for females. Make sure to set the seed to 1 before each sample call. What is the difference in absolute value between $\bar{y} - \bar{x}$ and $\bar{X} - \bar{Y}$? Note: X and Y are the sampled vectors defined before, x and y represent the population vectors.

```
x_female <- dat %>% filter(Sex == "F", Diet == "chow") %>% select(Bodyweight) %>% unlist
pop_avg_control_female <- mean(x_female)
pop_avg_control_female
```

```
## [1] 23.89338
```

```
popsd(x_female)
```

```
## [1] 3.416438
```

```
set.seed(1)
sample_avg_control_female <- mean(sample(x_female, size = 25, replace = FALSE))
sample_avg_control_female
```

```
## [1] 24.2528
```

```
y_female <- dat %>% filter(Sex == "F", Diet == "hf") %>% select(Bodyweight) %>% unlist
pop_avg_hf_female <- mean(y_female)
pop_avg_hf_female
```

```
## [1] 26.2689
```

```
popsd(y_female)
```

```
## [1] 5.06987
```

```
set.seed(1)
sample_avg_hf_female <- mean(sample(y_female, size = 25, replace = FALSE))
sample_avg_hf_female
```

```
## [1] 28.3828
```

```
abs(sample_avg_hf_female - sample_avg_control_female)
```

```
## [1] 4.13
```

```
abs(pop_avg_hf_female - pop_avg_control_female)
```

```
## [1] 2.375517
```

9. For the females, our sample estimates were closer to the population difference than with males. What is a possible explanation for this?
   a. The population variance of the females is smaller than that of the males; thus, the sample variable has less variability.
   b. Statistical estimates are more precise for females.
   c. The sample size was larger for females.
   d. The sample size was smaller for females.

Answer a

CLT Exercises

We will be using the same dataset `dat` as above.

1. (Conceptual) If a list of numbers has a distribution that is well approximated by the normal distribution, what proportion of these numbers are within one standard deviation away from the list's average?

Answer 68% (to be precise 68.27%)

```
pnorm(1) - (1 - pnorm(1))
```

```
## [1] 0.6826895
```

```
## OR
pnorm(1) - pnorm(-1)
```

```
## [1] 0.6826895
```

2. (Conceptual) What proportion of these numbers are within two standard deviations away from the list's average?

Answer 95% (to be precise 95.45%)

```
pnorm(2) - (1 - pnorm(2))
```

```
## [1] 0.9544997
```

```
## OR
pnorm(2) - pnorm(-2)
```

```
## [1] 0.9544997
```

3. (Conceptual) What proportion of these numbers are within three standard deviations away from the list's average?

Answer 99.7% (to be precise 99.73%)

```
pnorm(3) - (1 - pnorm(3))
```

```
## [1] 0.9973002
```

```
## OR
pnorm(3) - pnorm(-3)
```

```
## [1] 0.9973002
```

4. Define y to be the weights of males on the control diet. What proportion of the mice are within one standard deviation away from the average weight (remember to use popsd for the population sd)?

```
y <- dat %>% filter(Sex == "M", Diet == "hf") %>% select(Bodyweight) %>% unlist
upper_bound1 <- mean(y) + popsd(y)
lower_bound1 <- mean(y) - popsd(y)

mean(y <= upper_bound1 & y >= lower_bound1)
```

```
## [1] 0.7098446
```

```
## MORE ELEGANT WAY (Credit to Breakout Room 1)
prop <- (y-mean(y))/popsd(y)
mean(abs(prop) <= 1)
```

```
## [1] 0.7098446
```

5. What proportion of these numbers are within two standard deviations away from the list's average?

```
upper_bound2 <- mean(y) + 2* popsd(y)
lower_bound2 <- mean(y) - 2* popsd(y)

mean(y <= upper_bound2 & y >= lower_bound2)
```

```
## [1] 0.9378238
```

```
## MORE ELEGANT WAY (Credit to Breakout Room 1)
prop <- (y-mean(y))/popsd(y)
mean(abs(prop) <= 2)
```

```
## [1] 0.9378238
```

6. What proportion of these numbers are within three standard deviations away from the list's average?

```
upper_bound3 <- mean(y) + 3* popsd(y)
lower_bound3 <- mean(y) - 3* popsd(y)

mean(y <= upper_bound3 & y >= lower_bound3)
```

```
## [1] 0.9948187
```

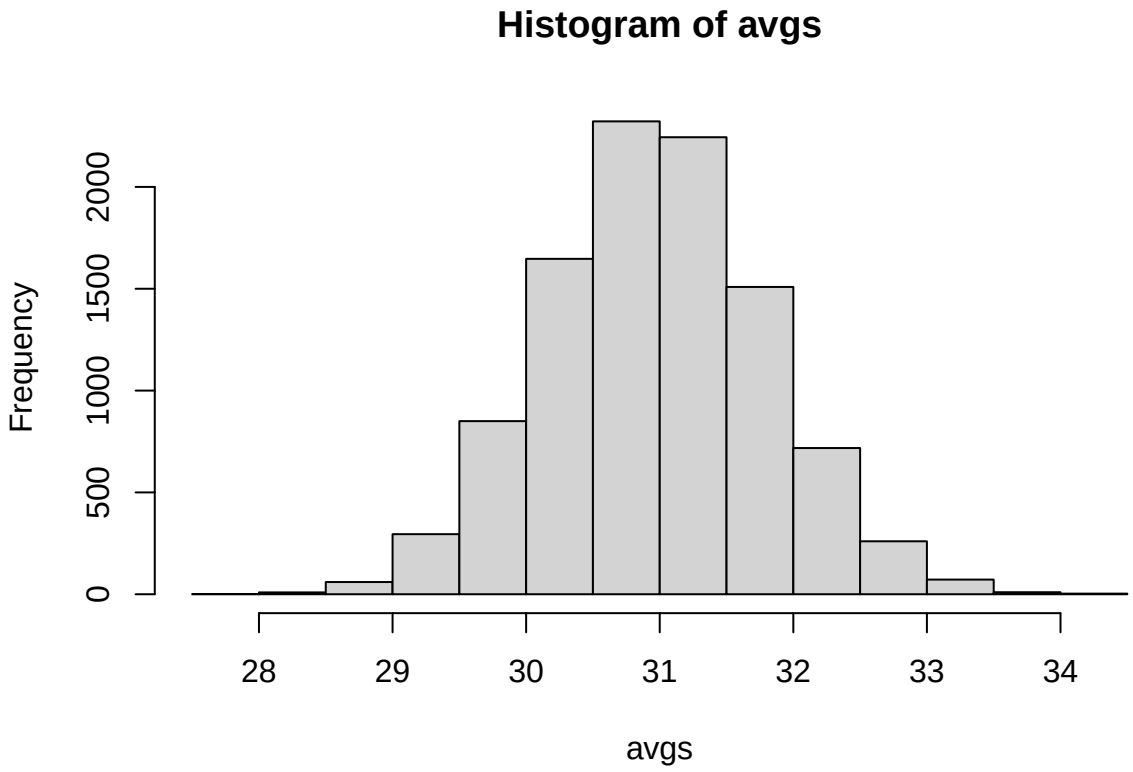
```
## MORE ELEGANT WAY (Credit to Breakout Room 1)
prop <- (y-mean(y))/popsd(y)
mean(abs(prop) <= 3)
```

```
## [1] 0.9948187
```

Note that the numbers for the normal distribution and our weights are relatively close. Also, notice that we are indirectly comparing quantiles of the normal distribution to quantiles of the mouse weight distribution.

7. Here we are going to use the function replicate to learn about the distribution of random variables. All the above exercises relate to the normal distribution as an approximation of the distribution of a fixed list of numbers or a population. We have not yet discussed probability in these exercises. If the distribution of a list of numbers is approximately normal, then if we pick a number at random from this distribution, it will follow a normal distribution. However, it is important to remember that stating that some quantity has a distribution does not necessarily imply this quantity is random. Also, keep in mind that this is not related to the central limit theorem. The central limit applies to averages of random variables. Let's explore this concept. We will now take a sample of size 25 from the population of males on the chow diet. The average of this sample is our random variable. We will use the replicate to observe 10,000 realizations of this random variable. Set the seed at 1, generate these 10,000 averages. Make a histogram of these 10,000 numbers against the normal distribution using `hist()`. We can see that, as predicted by the CLT, the distribution of the random variable is very well approximated by the normal distribution. What is the average of the distribution of the sample average?

```
library(dplyr)
y <- filter(dat, Sex=="M" & Diet=="chow") %>% select(Bodyweight) %>% unlist
avgs <- replicate(10000, mean( sample(y, 25)))
hist(avgs)
```




```
mean(avgs)
```

```
## [1] 30.96865
```

8. What is the standard deviation of the distribution of sample averages?

```
sd(avgs)
```

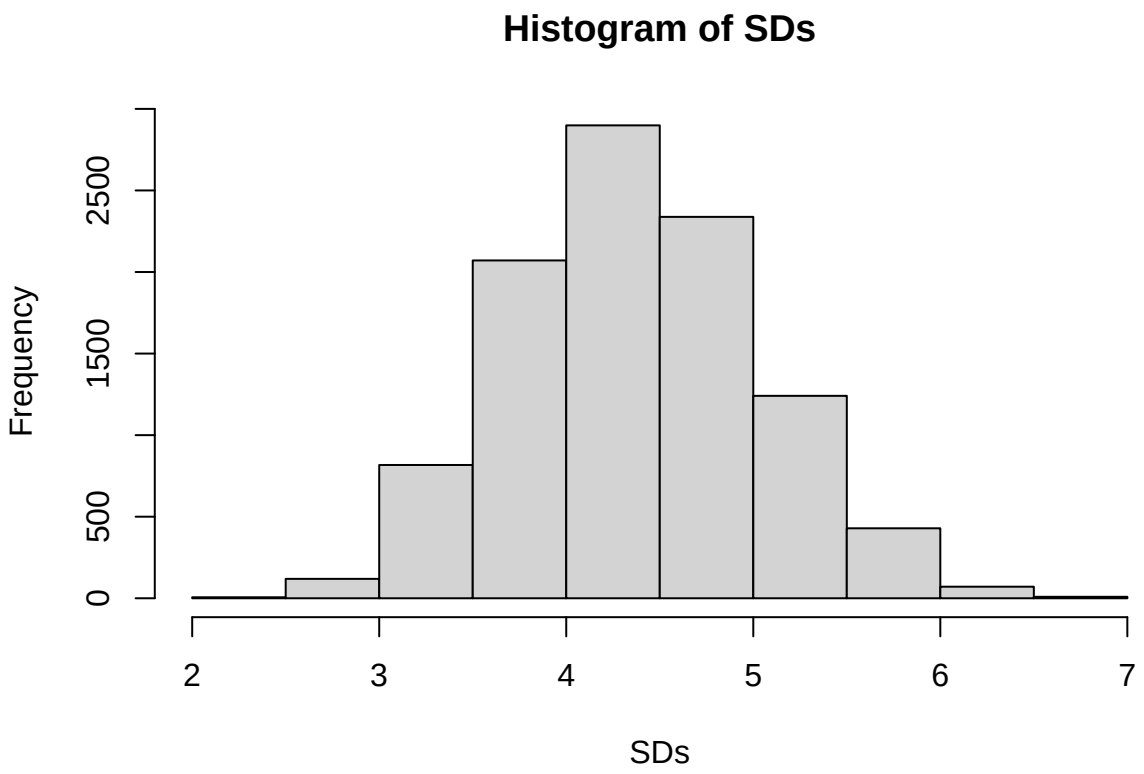
```
## [1] 0.8267168
```

9. According to the CLT, the answer to exercise 7 should be the same as `mean(y)`. You should be able to confirm that these two numbers are very close. Which of the following does the CLT tell us should be close to your answer to exercise 8?
   a. `popsd(y)`
   b. `popsd(avgs)/sqrt(25)`
   c. `sqrt(25) / popsd(y)`
   d. `popsd(y)/sqrt(25)`

Answer: d

10. In practice we do not know $\sigma($`popsd(y)`$)$ which is why we can't use the CLT directly. This is because we see a sample and not the entire distribution. We also can't use `popsd(avgs)` because to construct averages, we have to take 10,000 samples and this is never practical. We usually just get one sample. Instead we have to estimate `popsd(y)`. As described, what we use is the sample standard deviation. Set the seed at 1, using the `replicate` function, create 10,000 samples of 25 and now, instead of the sample average, keep the standard deviation. Look at the distribution of the sample standard deviations. It is a random variable. The real population SD is about 4.5. What proportion of the sample SDs are below 3.5?

```
set.seed(1)
SDs <- replicate(10000, sd( sample(y, 25)))
hist(SDs)
```




```
# proportion of the sample SDs below 3.5
mean(SDs < 3.5)
```

```
## [1] 0.0942
```

CLT in Practice Exercises (Optional)

We will be using the following dataset for the next set of exercises:

```
url <- "https://raw.githubusercontent.com/genomicsclass/dagdata/master/inst/extdata/femaleMiceWeights.cs
filename <- "femaleMiceWeights.csv"
if(!file.exists("femaleMiceWeights.csv")) download(url,destfile=filename)
dat <- read.csv(filename)
```

1. For quantitative data, we need to estimate the population standard deviation. In several previous exercises we have illustrated statistical concepts with the unrealistic situation of having access to the entire population. In practice, we do not have access to entire populations. Instead, we obtain one random sample and need to reach conclusions analyzing that data. dat is an example of a typical simple dataset representing just one sample. We have 12 measurements for each of two populations. We think of X as a random sample from the population of all mice in the control diet and Y as a random sample from the population of all mice in the high fat diet. Define the parameter μ_x as the average of the control population. We estimate this parameter with the sample average $\bar{X}$. What is the sample average?

```
library(dplyr)
X <- filter(dat, Diet=="chow") %>% select(Bodyweight) %>% unlist
Y <- filter(dat, Diet=="hf") %>% select(Bodyweight) %>% unlist

X_bar <- mean(X)
X_bar
```

```
## [1] 23.81333
```

2. We don't know μ_X , but want to use $\bar{X}$ to understand μ_X. Which of the following uses CLT to understand how well $\bar{X}$ approximates μ_X ?
   a. $\bar{X}$ follows a normal distribution with mean 0 and standard deviation 1.
   b. μ_X follows a normal distribution with mean $\bar{X}$ and standard deviation $\frac{\sigma_x}{\sqrt{(12)}}$ where σ_x is the population standard deviation.
   c. $\bar{X}$ follows a normal distribution with mean μ_X and standard deviation σ_x where σ_x is the population standard deviation.
   d. $\bar{X}$ follows a normal distribution with mean μ_X and standard deviation $\frac{\sigma_x}{\sqrt{(12)}}$ where σ_x is the population standard deviation.

Answer: d

3. The result above tells us the distribution of the following random variable: $Z = \sqrt{12}\frac{\bar{X}-\mu_X}{\sigma_X}$. What does the CLT tell us is the mean of Z (you don't need code)?

Answer: 0

4. The result of 2 and 3 tell us that we know the distribution of the difference between our estimate and what we want to estimate, but don't know. However, the equation involves the population standard deviation σ_X, which we don't know. Given what we discussed, what is your estimate of σ_x?

```
sd(X)
```

```
## [1] 3.022541
```

5. Use the CLT to approximate the probability that our estimate $\bar{X}$ is off by more than 5.21 ounces from μ_X.

```
2 * ( 1-pnorm(5.21/sd(X) * sqrt(12) ) )
```

```
## [1] 2.356222e-09
```