Lab 3: Loops and central limit theorem

Julieth Santamaria

February 8, 2019

Summary

Key Terms

- **Central limit theorem**: The mean of a sufficiently large number of independent draws from any distribution will be normally distributed

Application

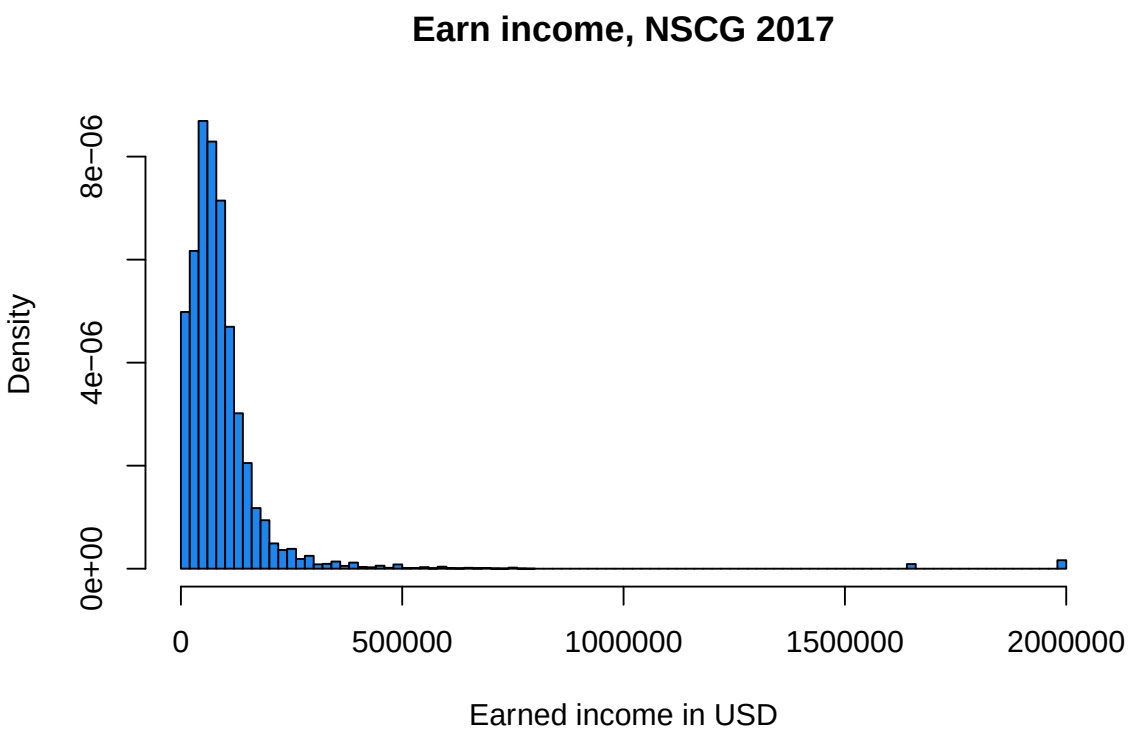
1. Load nscg17 (the same dataset we used last week). Set the working directory to the labs folder in your computer. Then, run the following lines. We run these lines last week but just to check, what are these lines doing?

```
setwd("G:/My Drive/U of M/TA/TA APEC3003/APEC 3003 - 2019/APEC 3003 R work/labs/")
load("../data/nscg17.rdata")
nscg17 <- within(nscg17, {
      earn[earn == 9999998] <- NA
    })
```

   Notice that earnings is not the same as salary (the variable you'll have to clean in assignment 1).

2. Do a histogram of the variable earnings

```
  hist(nscg17$earn,
    breaks=100, # Number of cells for the histogram
    freq=FALSE, # If TRUE ->frequencies, if FALSE -> probability densities
    col="dodgerblue2", # Sets the color of your bars
    xlab="Earned income in USD", # Sets the label of your x-axis
    main="Earn income, NSCG 2017")  # Title of the graph
```



3. Run a simulation.
   3.1. Get familiar with loops

   - Think of something you want to do this weekend. Store a small description in an object called `weekend.plans`. Then, make R repeat your plans 5 times

```
weekend.plans="I'll swim"
for (j in 1:5) {
  print(weekend.plans)
}
```

```
## [1] "I'll swim"
## [1] "I'll swim"
## [1] "I'll swim"
## [1] "I'll swim"
## [1] "I'll swim"
```

   - Create an object 'J' that contains your favorite number. Then, print your favorite number 10 times

```
J<-8
for (j in 1:10) {
    print(J)
  }
```

```
## [1] 8
## [1] 8
## [1] 8
## [1] 8
## [1] 8
## [1] 8
```

```
## [1] 8
## [1] 8
## [1] 8
## [1] 8
```

- Use a loop to print numbers from 1-10.

```
K<-10              # This sets the number at which the loop should stop
for(j in 1:K) {    # Loop starts: j starts in 1 and ends in J
  print(j)         # prints j
}                  # The loop will continue until j equates 10
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
```

3.2. Write a loop that samples 3000 people, calculates the mean for those people, and stores the sample mean in a vector "'salarymeans"'. Run the loop 80 times.

```
N <- 3000 # Number of observations in the sample
R <- 80 # Number of repetitions
earn.means <- rep(NA, R) # Creates an empty vector with R elements
```

- First, let's do the first repetition (when R=1).

```
# 1. Create a sample of size 3000 (or N) from the variable earnings
    earn.sample<- sample(nscg17$earn, N)

# 2. Clculate the mean earnings of the sample and store it as the first element of the vector earn
    earn.means[1] <- mean(earn.sample, na.rm=TRUE)
```

- Now, insert those commands in a loop. Remember that you'll repeat the same procedure 800 (or R) times.

```
for(j in 1:R) {
   earn.sample<- sample(nscg17$earn, N)
   earn.means[j] <- mean(earn.sample, na.rm=TRUE) # Assigns the mean of the sample to row j
}
```

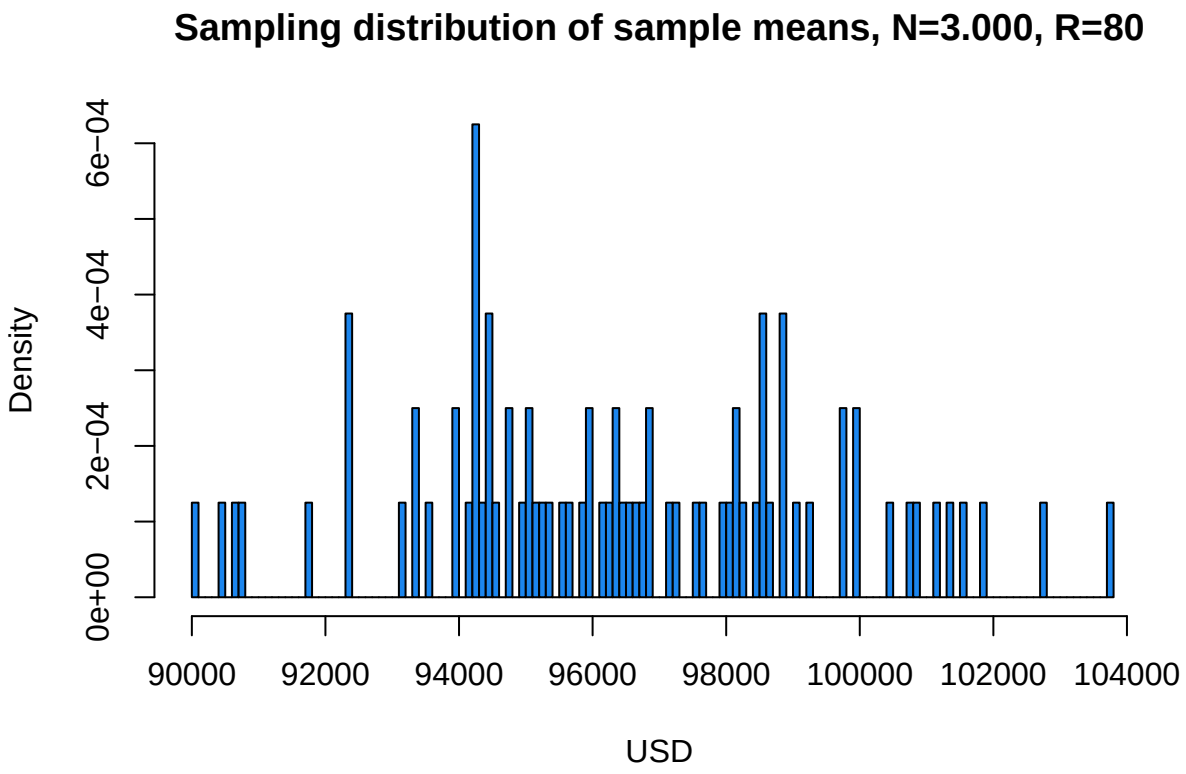
- This calculates the difference between the two means

```
mean(nscg17$earn, na.rm=TRUE) - mean(earn.means, na.rm=TRUE)
```

```
## [1] 109.834
```

3.3. Make a histogram of the resulting variable using 100 breaks, with your favorite color (http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf), and entitled "Sample distribution of sample means".

```
hist(earn.means,
     breaks=100,freq=FALSE, col="dodgerblue2",
     xlab="USD", main="Sampling distribution of sample means, N=3.000, R=80",
     xlim = c(90000,104000))
```

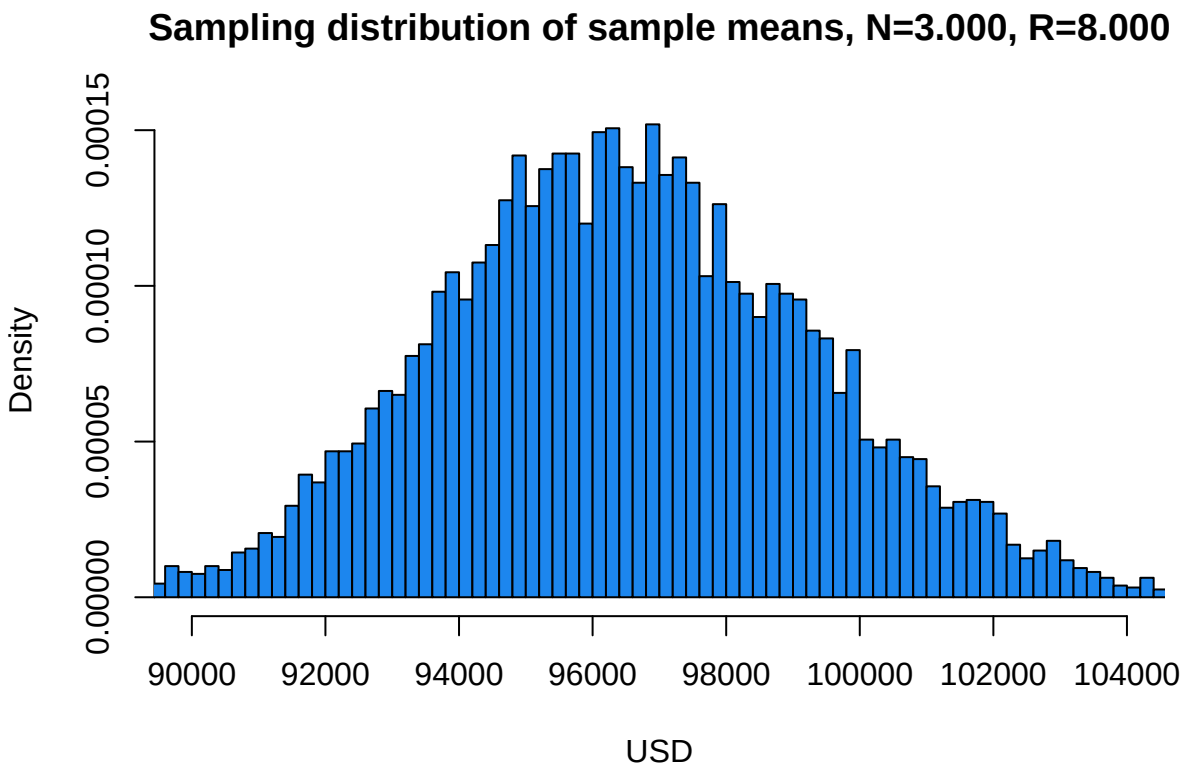


4. What happens when you increase the number of repetitions to 8.000?

```
N <- 3000 # Number of observations in the sample
R <- 8000 # Number of repetitions
earn.means <- rep(NA, R)

 for(j in 1:R) {
       earn.sample<- sample(nscg17$earn, N)
       earn.means[j] <- mean(earn.sample, na.rm=TRUE) # Assigns the mean of the sample to row j
 }

hist(earn.means,
         breaks=100,freq=FALSE, col="dodgerblue2",
         xlab="USD", main="Sampling distribution of sample means, N=3.000, R=8.000",
         xlim = c(90000,104000))
```

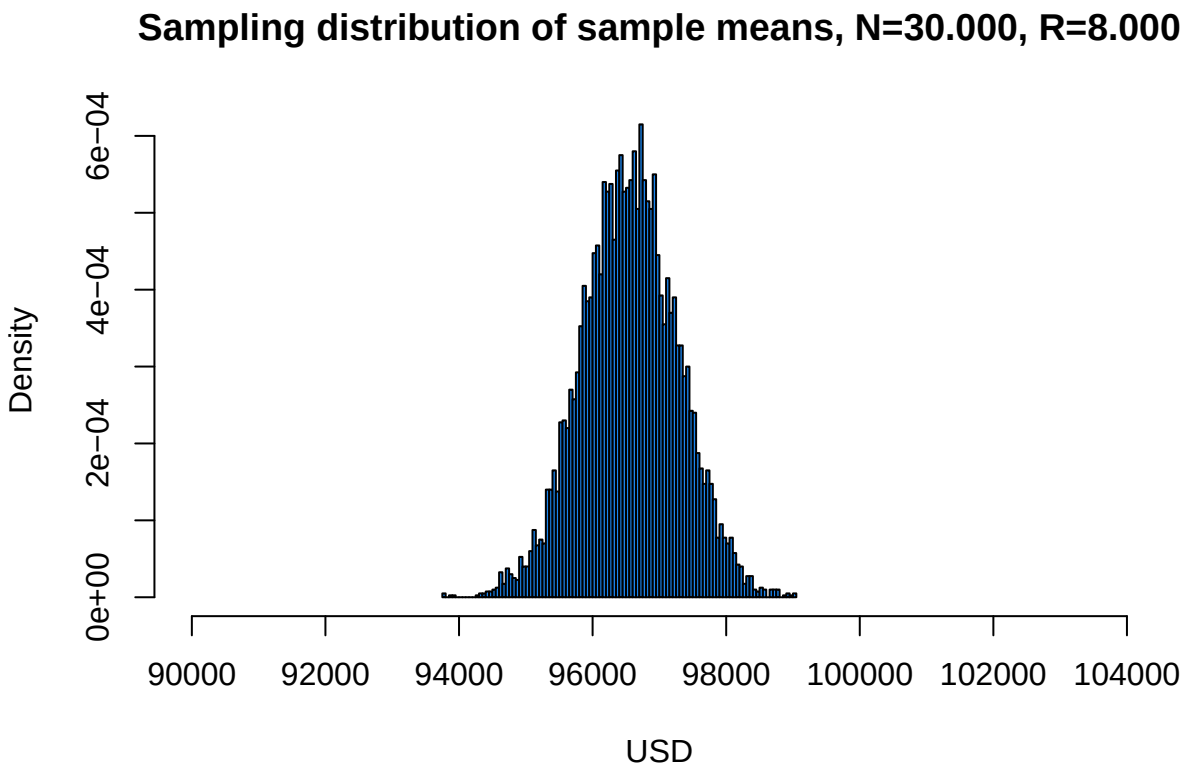


5. What happens when you increase number of observations to 30.000?

```
N <- 30000 # Number of observations in the sample
R <- 8000 # Number of repetitions
earn.means <- rep(NA, R)

 for(j in 1:R) {
       earn.sample<- sample(nscg17$earn, N)
       earn.means[j] <- mean(earn.sample, na.rm=TRUE) # Assigns the mean of the sample to row j
 }

hist(earn.means,
         breaks=100,freq=FALSE, col="dodgerblue2",
         xlab="USD", main="Sampling distribution of sample means, N=30.000, R=8.000",
         xlim = c(90000,104000))
```


Sampling distribution of sample means, N=30.000, R=8.000
Density
0e+00
2e−04
4e−04
6e−04
90000
92000
94000
96000
98000
100000
102000
104000
USD