STU22005 Laboratory Session 2

Professor Caroline Brophy

Week beginning 15th February 2021

Instructions

- In this lab sheet, there will be a series of notes for you to work through. For each section, you should run the R code yourself on your own computer and verify the output provided. Ask your lab demonstrator any questions you have as you work through the material.
- In the last section of the sheet, there are several questions where you will write your own code. Your lab demonstrator is there to help if you get stuck in this section.
- Create a script for all the code that you write in this lab session. Save the script into the folder you created last week for the module. Save your code regularly throughout the class. (Do not write your code into the "Console" window as this will not save in an R script.)

Sampling distributions

From the lecture notes Section 3.3:

"The 'sampling distribution' of a sample statistic is NOT the same as the 'sample distribution' which is the distribution of the raw data in the sample."

In this lab session, we are going to explore this statement for ourselves.

Samples that are normally distributed.

Simulate 1000 datasets, each of size 10, from a N(100,16) distribution. Note that we always write N(μ, σ^2), so 16 is the variance, not the standard deviation.

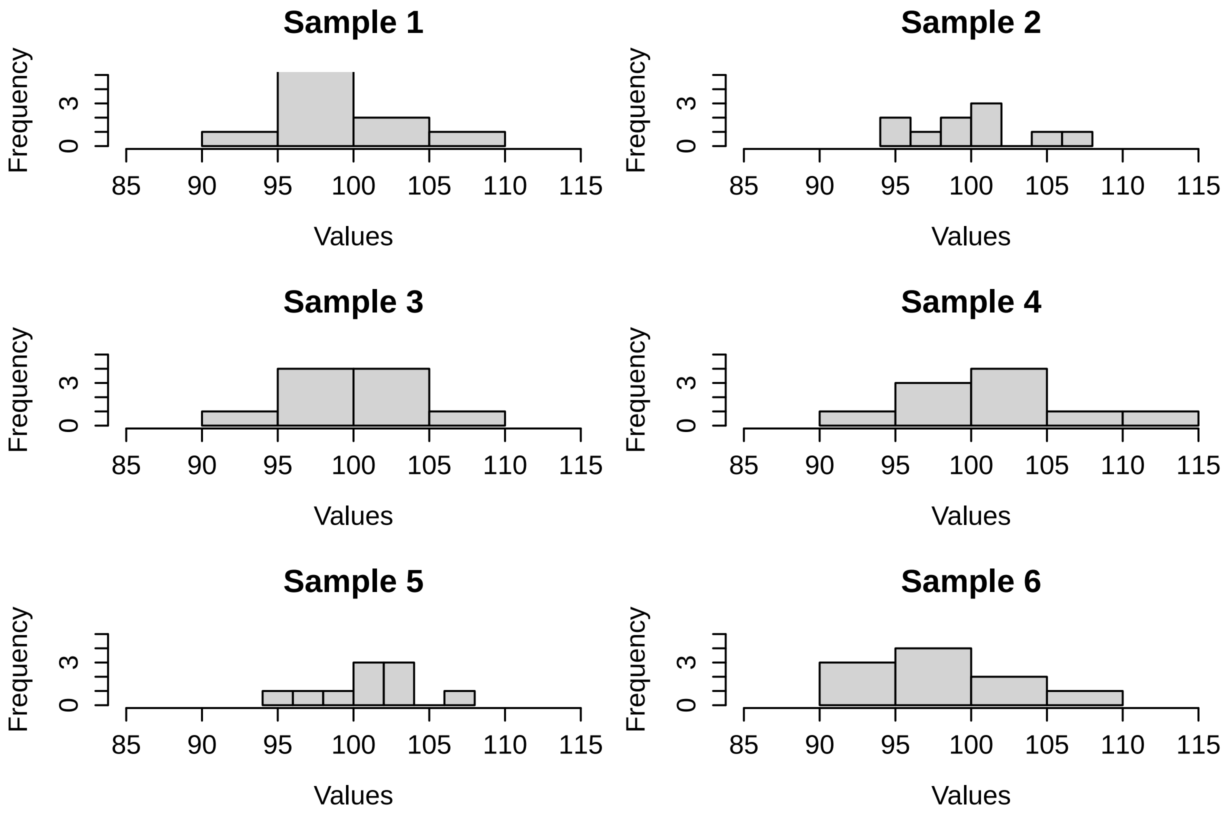
- Compute the mean for each sample and store it in a new vector. Display the first six means. What is the `matrix()` function doing? You can run the code `?matrix` to find out more.

```r
# Set the seed value
set.seed(8423)

# Simulate values from the normal distribution
simdata <- rnorm(n = 10000, mean = 100, sd = 4)

# Create a matrix with 1000 samples in each row of 10 columns
matrixdata <-  matrix(simdata, nrow = 1000, ncol = 10)

# Compute the mean of each sample (each row) and print the first six means
means <- apply(matrixdata, 1, mean)
means[1:6]
```

```
## [1]  99.63755  99.94784 100.19990 102.21240 101.05745  97.88532
```

- Construct a histogram of the first six samples. Note, in the code that follows, we are using a loop to construct multiple histograms, instead of writing out the `hist` code several times. Examine the code, and think about what `par`, `matrixdata[i,]` and `paste0` are doing. You can run the code `?par` to find out more about it.

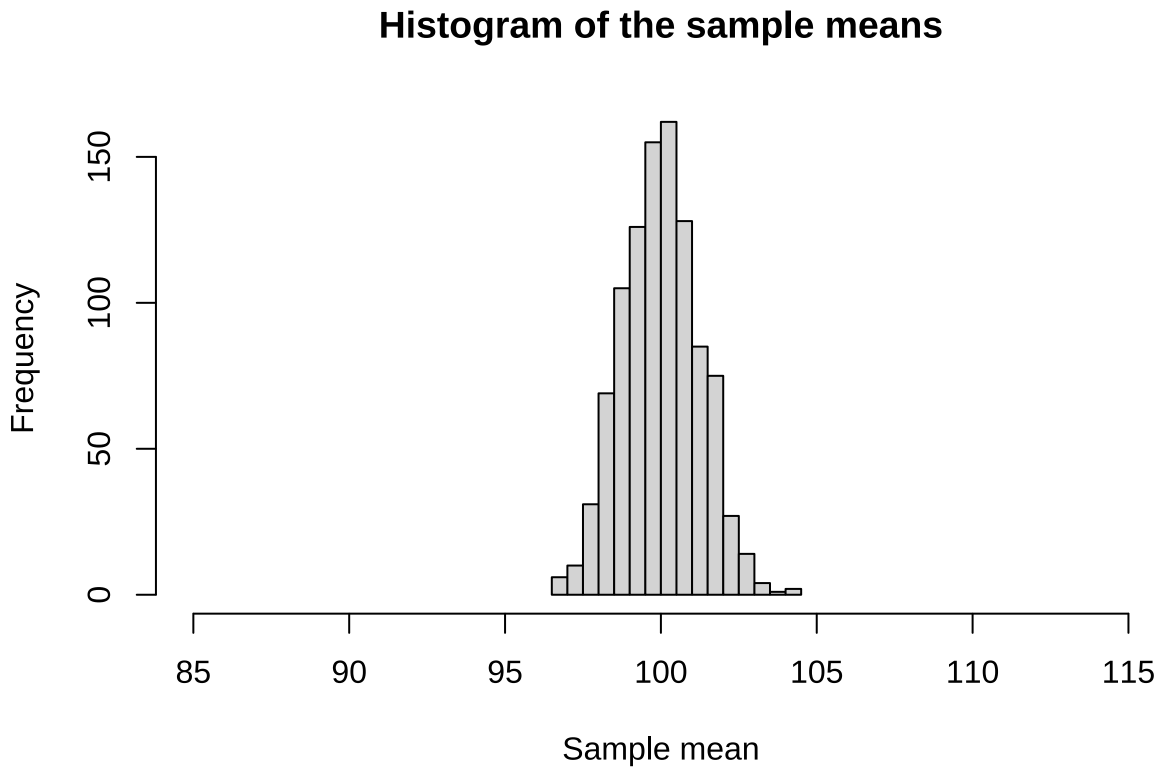
```
# Plot the first six samples
par(mfrow=c(3,2), mar=c(4,4,4,1), oma=c(0.5,0.5,0.5,0))
for (i in c(1:6)){
   hist(matrixdata[i,], main = paste0("Sample ", i), xlab = "Values",
        cex.main = 1.5, cex.lab = 1.2, cex.axis = 1.2,
        ylim = c(0,5), xlim = c(85,115))
}
```



```
# Reset the number of plots that appear in the window at a time
par(mfrow=c(1,1))
```

- Now construct a histogram of the sample means from each of the 1000 samples.

```
# Generate histogram (keeping the y scale the same as the earlier graphs for comparison)
hist(means, main = "Histogram of the sample means", xlab = "Sample mean", xlim = c(85,115))
```



- How would you describe this sampling distribution? What do you notice when you compare it to the histograms of the samples? (Note that the x-axis is on the same scale as the earlier graphs to aid comparison.)

Samples that are exponentially distributed.

Simulate 1000 datasets, each of size 10, from a exponential distribution with $\lambda = 0.5$. From properties of exponential distributions, what is the mean and variance of this distribution?

```
# Set the seed value
set.seed(96358285)

# Simulate values from the exponential distribution
simdata <- rexp(n = 10000, rate = 0.5)

# Create a matrix with 1000 samples in each row of 10 columns
matrixdata <-  matrix(simdata, nrow = 1000, ncol = 10)
```

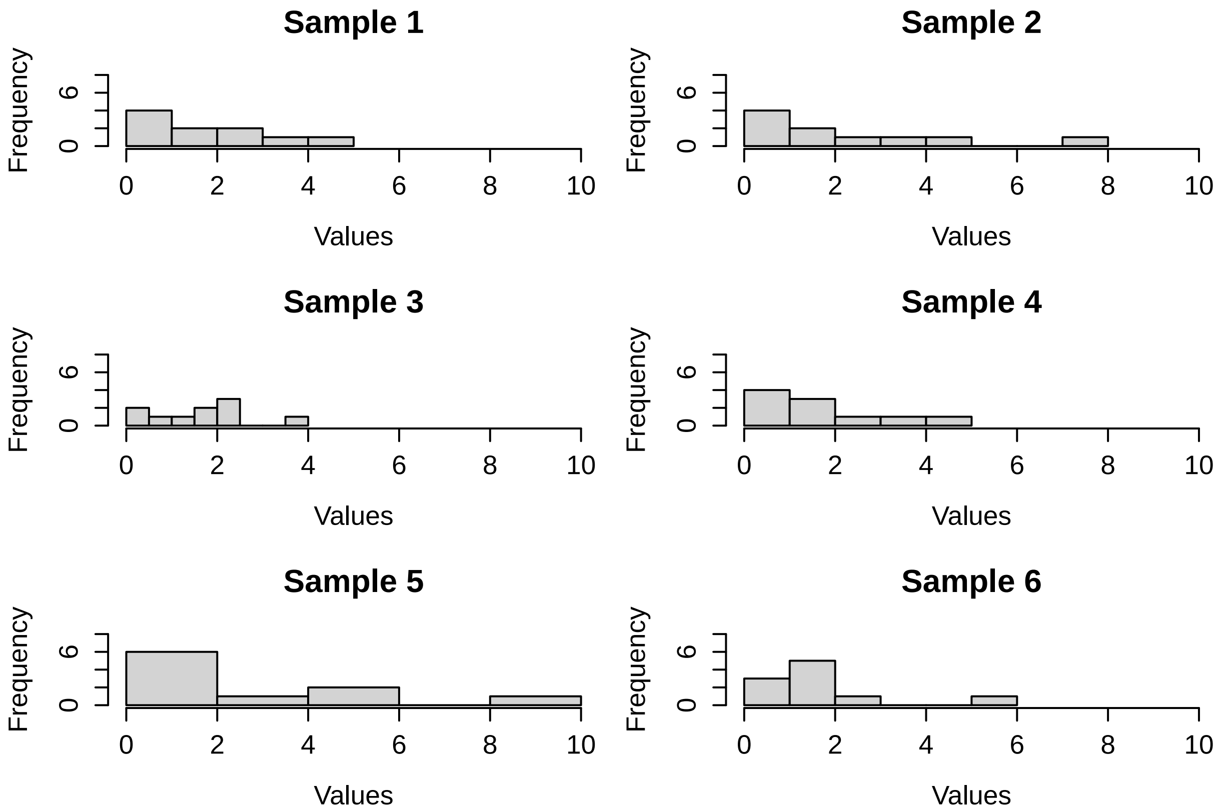
- Compute the mean for each sample and store it in a new vector. Display the first six means.

```
#Compute the mean of each sample (each row)
means <- apply(matrixdata, 1, mean)
means[1:6]
```

```
## [1] 1.684177 2.168568 1.604019 1.704819 2.696661 1.570264
```

- Construct a histogram of the first six samples.

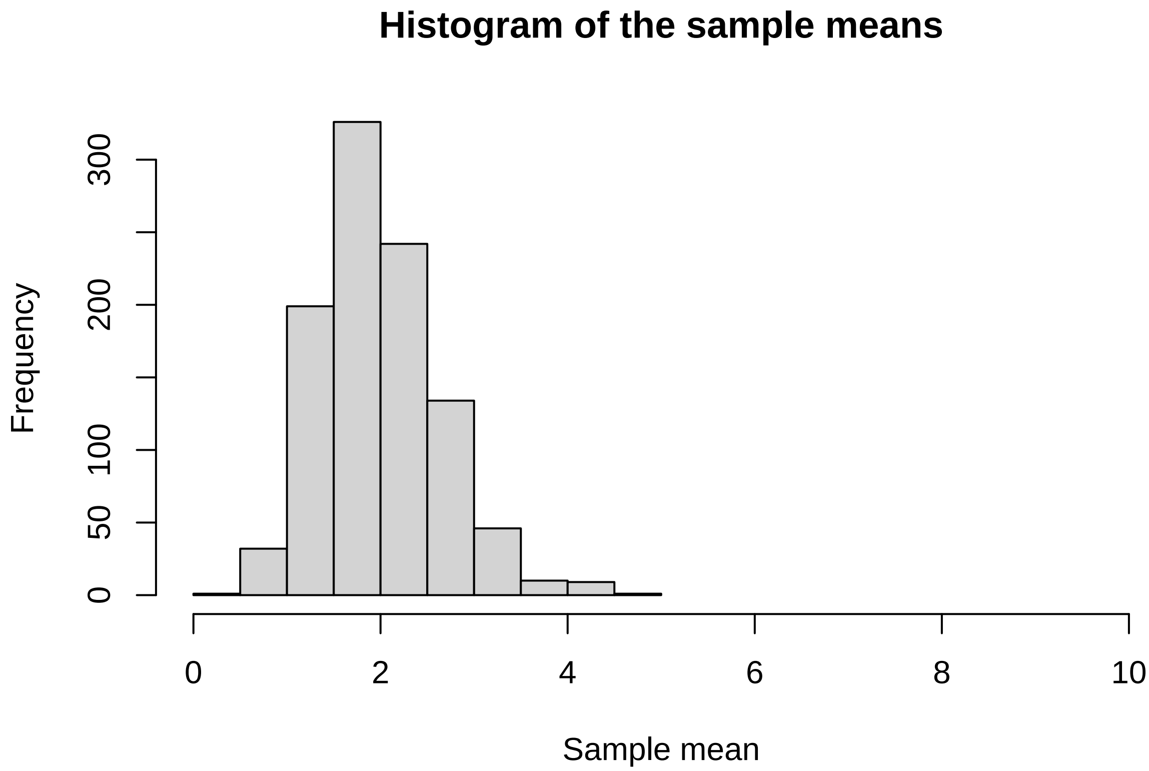
```r
# Plot the first six samples
par(mfrow = c(3,2), mar = c(4,4,4,1), oma = c(0.5,0.5,0.5,0))
for (i in c(1:6)){
   hist(matrixdata[i,], main = paste0("Sample ", i), xlab = "Values",
        cex.main = 1.5, cex.lab = 1.2, cex.axis = 1.2,
        ylim = c(0,8), xlim = c(0,10))
}
```



```r
# Reset the number of plots that appear in the window at a time
par(mfrow = c(1,1))
```

- Now construct a histogram of the sample means from each of the 1000 samples.

```r
# Generate histogram (keeping the y scale the same as the earlier graphs for comparison)
hist(means, main = "Histogram of the sample means", xlab = "Sample mean", xlim = c(0, 10))
```



- How would you describe this sampling distribution? Do you think it is a little bit skewed or is it symmetric?

Putting your code into practice

Save the code you have written so far in this lab session. Generally, it is a good idea to save your code every five minutes!

For the remainder of this laboratory session: DO NOT USE COPY / PASTE. Write out your own code. Try to remember the code yourself, but you can look back at the earlier code whenever you need to. Typing out your own code gives you a better chance of understanding what each part of the code does, and will help your R programming skills to develop.

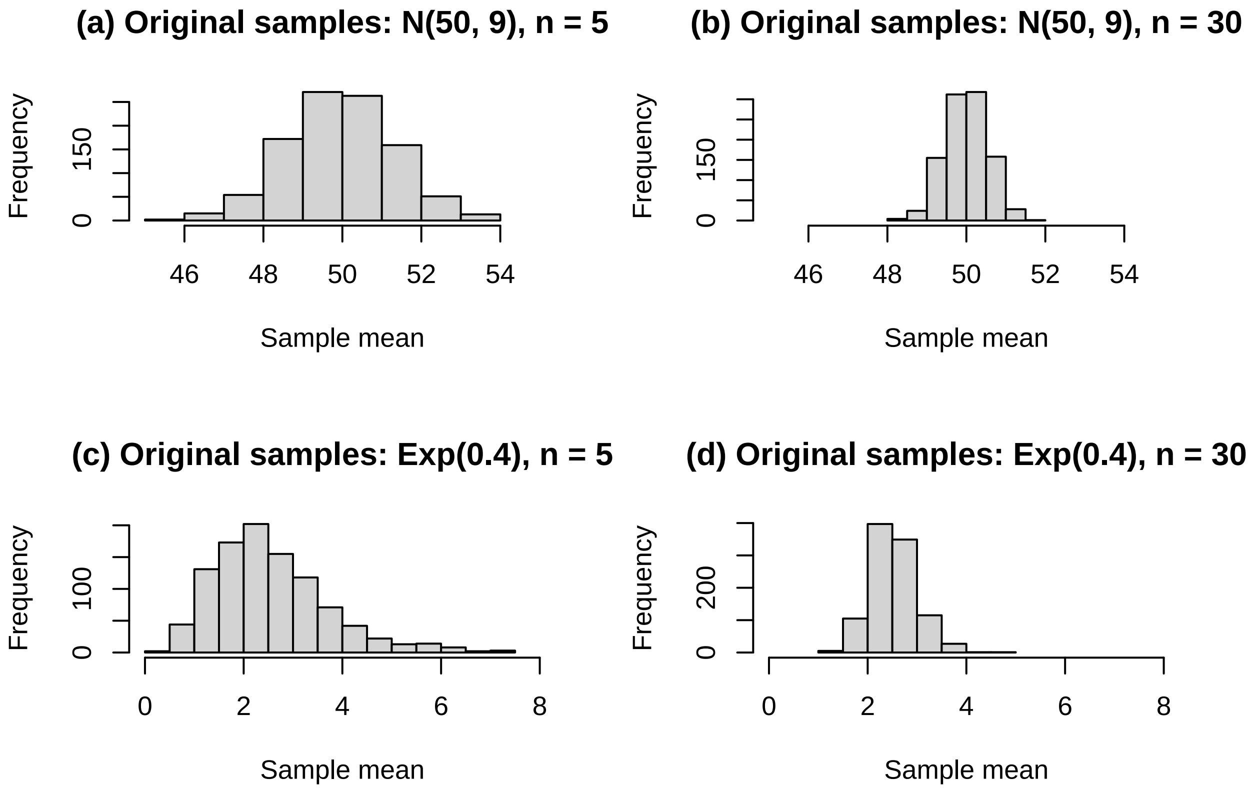
It is very good practice to include good comments for each part of your R code. Please get into the habit of this every time you write R code.

Questions

1. Simulate 1000 datasets each of size 5 from a N(50, 9) distribution. Use the seed value 6594. Find the mean for each sample and store them in a new vector (this vector will contain 1000 values). Generate a histogram of this vector of sample means.
2. Repeat the previous question, but this time use a sample size of 30 in the 1000 datasets. Use the seed value 9278. Use the same range on the histogram x-axis as in question 1.
3. Simulate 1000 datasets each of size 5 from an exponential distribution with lambda = 0.4. Use the seed value 3845. Find the mean for each sample and store them in a new vector (this vector will contain 1000 values). Generate a histogram of this vector of sample means.

4. Repeat the previous question, but this time use a sample size of 30 in the 1000 datasets. Use a seed value of 8651. Use the same range on the histogram x-axis as in question 3.
5. Create a plot with all four histograms of means in one.

Hints: To keep all plots in the same output window, use `par(mfrow = c(2,2))`. This is what your four-panel plot should look like (assuming you used the seed value as indicated in each question):



(a) Original samples: N(50, 9), n = 5

(b) Original samples: N(50, 9), n = 30

(c) Original samples: Exp(0.4), n = 5

(d) Original samples: Exp(0.4), n = 30

Comment on the four diagrams. What do you see?

Explore further:

- Why do we see the difference between plots (a) and (b)? Hint: We have seen this result already in Section 4.1 of the notes, where we proved the sampling distribution of the mean for samples of normally distributed and independent random variables.
- Why do we see the difference between plots (c) and (d)? We will see this theoretical result in Section 5! For now, describe what the comparison seems to be saying (pay attention to the original sample sizes).