Lab 5 Solution

Zelin (James) Li

7/6/2020

Contents

Watch the video for Session 5

Probability and random variables are an essential component of data analysis, as they allow us to model noise, error and uncertainty in the data. The next two lectures will thus focus on giving you the basics of probability theory.

1. Watch the video lecture associated to Session 5 (find the links on the website) – it will walk you through concepts on random variable (RV), null distribution, and probability distributions.
2. You can complement this set of video lectures by reading the following chapter on random variables.

Random Variable Exercises

We will be using the following dataset:

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

`x` represents the weights for the entire population

Run the previous chunk of code to load the data and answer the following questions:

1. What is the average of these weights?

```
avg_all <- mean(x)
```

2. After setting the seed at 1, `set.seed(1)` take a random sample of size 5. What is the absolute value (use abs) of the difference between the average of the sample and the average of all the values?

```
set.seed(1)
sample5 <- sample(x, size = 5, replace = FALSE)
avg_sample5 <- mean(sample5)
abs(avg_all - avg_sample5)
```

```
## [1] 0.3293778
```

3. After setting the seed at 5, `set.seed(5)` take a random sample of size 5. What is the absolute value of the difference between the average of the sample and the average of all the values?

```r
set.seed(5)
sample5 <- sample(x, size = 5, replace = FALSE)
avg_sample5 <- mean(sample5)
abs(avg_all - avg_sample5)
```

```
## [1] 0.3813778
```

4. Why are the answers from 2 and 3 different?
    a. Because we made a coding mistake.
    b. Because the average of the x is random.
    c. Because the average of the samples is a random variable.
    d. All of the above.

Answer c. Because the average of the samples is a random variable.

5. Set the seed at 1, then using a for-loop take a random sample of 5 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of x ?

```r
set.seed(1)
avgs <- c()
for (i in 1:1000){
    avgs[i] <- mean(sample(x, size = 5, replace = FALSE))
}

# simplest way
mean(abs(avgs - avg_all) > 1)
```

```
## [1] 0.503
```

```r
# alternative way
sum(abs(avgs - avg_all) > 1) / length(avgs)
```

```
## [1] 0.503
```

6. We are now going to increase the number of times we redo the sample from 1,000 to 10,000. Set the seed at 1, then using a for-loop take a random sample of 5 mice 10,000 times. Save these averages. What percent of these 10,000 averages are more than 1 ounce away from the average of x ?

```r
set.seed(1)
avgs_iter10000 <- c()
for (i in 1:10000){
    avgs_iter10000[i] <- mean(sample(x, size = 5, replace = FALSE))
}

# simplest way
mean(abs(avgs_iter10000 - avg_all) > 1)
```

```
## [1] 0.5084
```

```r
# alternative way
sum(abs(avgs_iter10000 - avg_all) > 1) / length(avgs_iter10000)
```

```
## [1] 0.5084
```

7. Note that the answers to 5 and 6 barely changed. This is expected. The way we think about the random value distributions is as the distribution of the list of values obtained if we repeated the experiment an infinite number of times. On a computer, we can't perform an infinite number of iterations so instead, for our examples, we consider 1,000 to be large enough, thus 10,000 is as well. Now if instead we change the sample size, then we change the random variable and thus its distribution. Set the seed at 1, then

using a for-loop take a random sample of 50 mice 1,000 times. Save these averages. What percent of these 1,000 averages are more than 1 ounce away from the average of x ?

```
set.seed(1)
avgs_size50 <- c()
for (i in 1:1000){
    avgs_size50[i] <- mean(sample(x, size = 50, replace = FALSE))
}

# simplest way
mean(abs(avgs_size50 - avg_all) > 1)
```

```
## [1] 0.014
```

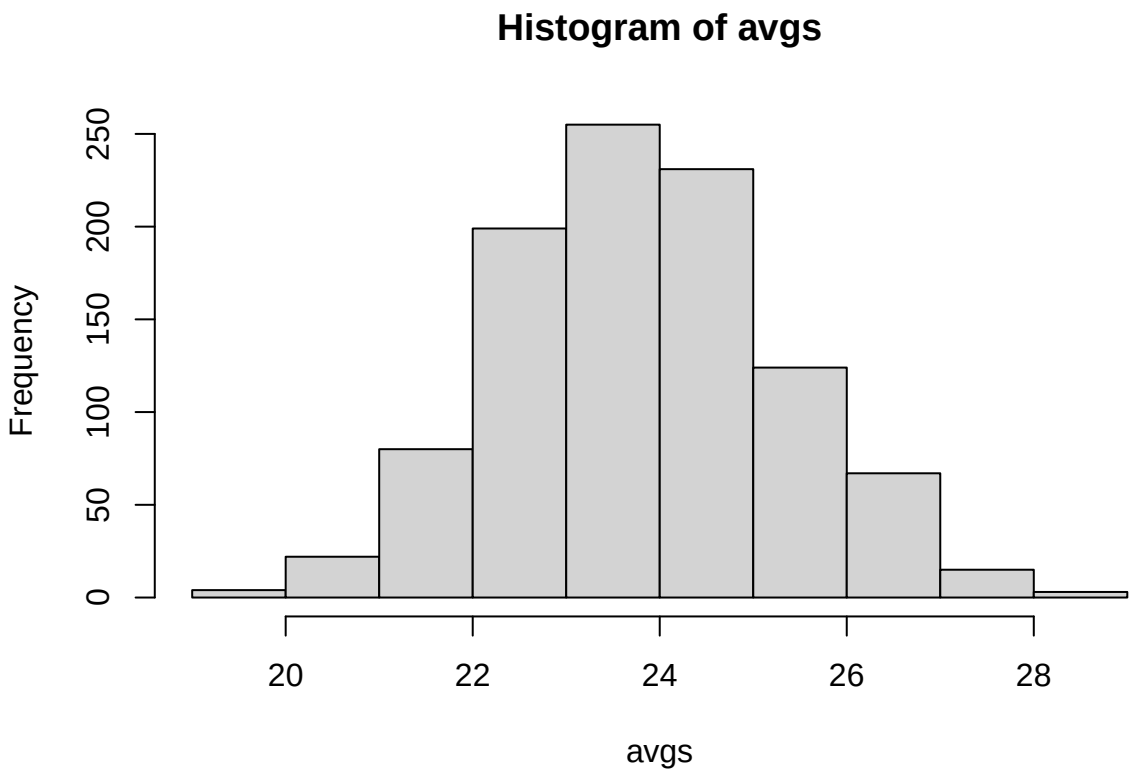
```
# alternative way
sum(abs(avgs_size50 - avg_all) > 1) / length(avgs_size50)
```

```
## [1] 0.014
```

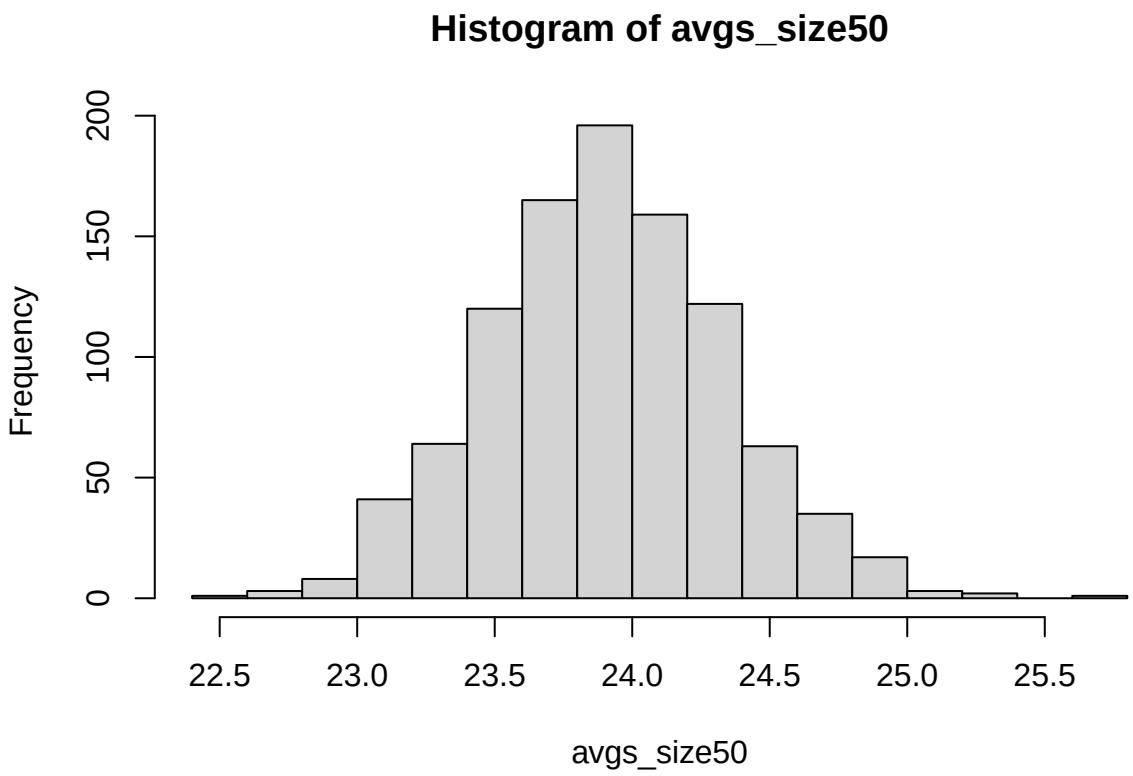
8. Use a histogram `hist()` to "look" at the distribution of averages we get with a sample size of 5 and a sample size of 50. How would you say they differ?
   a. They are actually the same.
   b. They both look roughly normal, but with a sample size of 50 the spread is smaller.
   c. They both look roughly normal, but with a sample size of 50 the spread is larger.
   d. The second distribution does not look normal at all.

Answer b. They both look roughly normal, but with a sample size of 50 the spread is smaller.

```
# sample size 5
hist(avgs)
```




```
# sample size 50
hist(avgs_size50)
```



9. For the last set of averages, the ones obtained from a sample size of 50, what percent are between 23 and 25?

```
mean((avgs_size50 >= 23) & (avgs_size50 <= 25))
```

```
## [1] 0.982
```

10. Now ask the same question of a normal distribution with average 23.9 and standard deviation 0.43.

```
mu <- 23.9
std.dev <- 0.43
lessThan25 <- pnorm(25, mean = mu, sd = std.dev, lower.tail=TRUE) # P(x < 25)
lessThan23 <- pnorm(23, mean = mu, sd = std.dev, lower.tail=TRUE) # P(x < 23)

## Note: setting lower.tail = FALSE means we are looking at the opposite, i.e.
## pnorm(23, mean = mu, sd = std.dev, lower.tail=FALSE) gives us P(x > 23).

# P(23 <= x <= 25)
lessThan25 - lessThan23
```

```
## [1] 0.9765648
```

The answer to 9 and 10 were very similar. This is because we can approximate the distribution of the sample average with a normal distribution. We will learn more about the reason for this next.

In class exercise

COVID-modeling: simplification of the Susceptible-Infected-Recovered model: - each day, X_t patients are infected. - Each of them contaminate R_0 new patients according to a Poisson distribution, and are put in quarantine in the evening. The new cases will exhibit symptoms the next day.

$$X_{t+1} = \sum_{i=1}^{X_t} Poisson(R_0) = Poisson(X_t R_0)$$

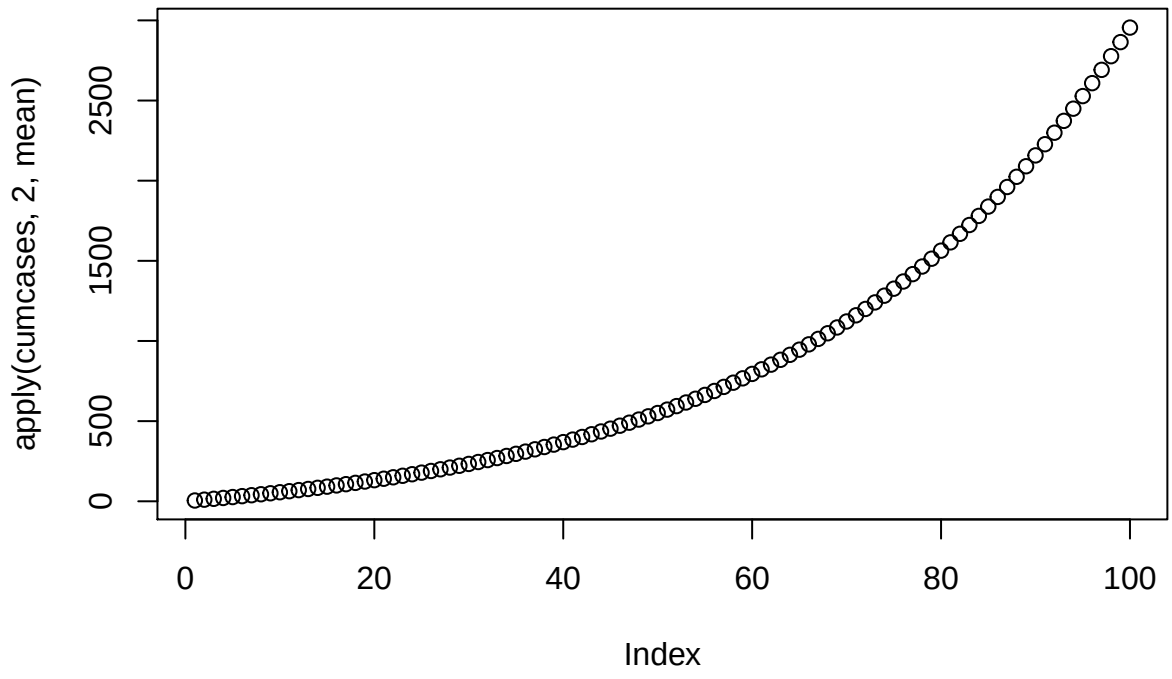
(a) Choosing $X_1 = 5$ and $R_0 = 1.03$, model 10,000 trajectories. **Note** You can first try to model 1 trajectory for (a) and (b) and think about how to expand to 10,000 trajectories.

(b) Compute for each of these epidemic trajectories the cumulative number of cases.

(c) What is the probability that the epidemic infects over 1,000 people in a month?

(d) What is the probability that the epidemic infects over 10,000 people over 100 days?

```
r0 = 1.03
cases = matrix(0, 10000, 100)
cases[,1] = 5 # X_1 = 5
for (t in 2:100){
  #cases[,t] = sapply(cases[,t-1], function(x){rpois(1, x * r0)})
  for (b in 1:10000){
    cases[b,t] = rpois(1, cases[b,t-1] * r0)
  }
}
cumcases = t(as.matrix(apply(cases,1, cumsum)))
plot(apply(cumcases,2, mean))
```




```
mean(cumcases[,30]>1000)
```

```
## [1] 0.0372
```

```
mean(cumcases[,100]>10000)
```

```
## [1] 0.1095
```