Midterm

Christina Lee

10/18/2021

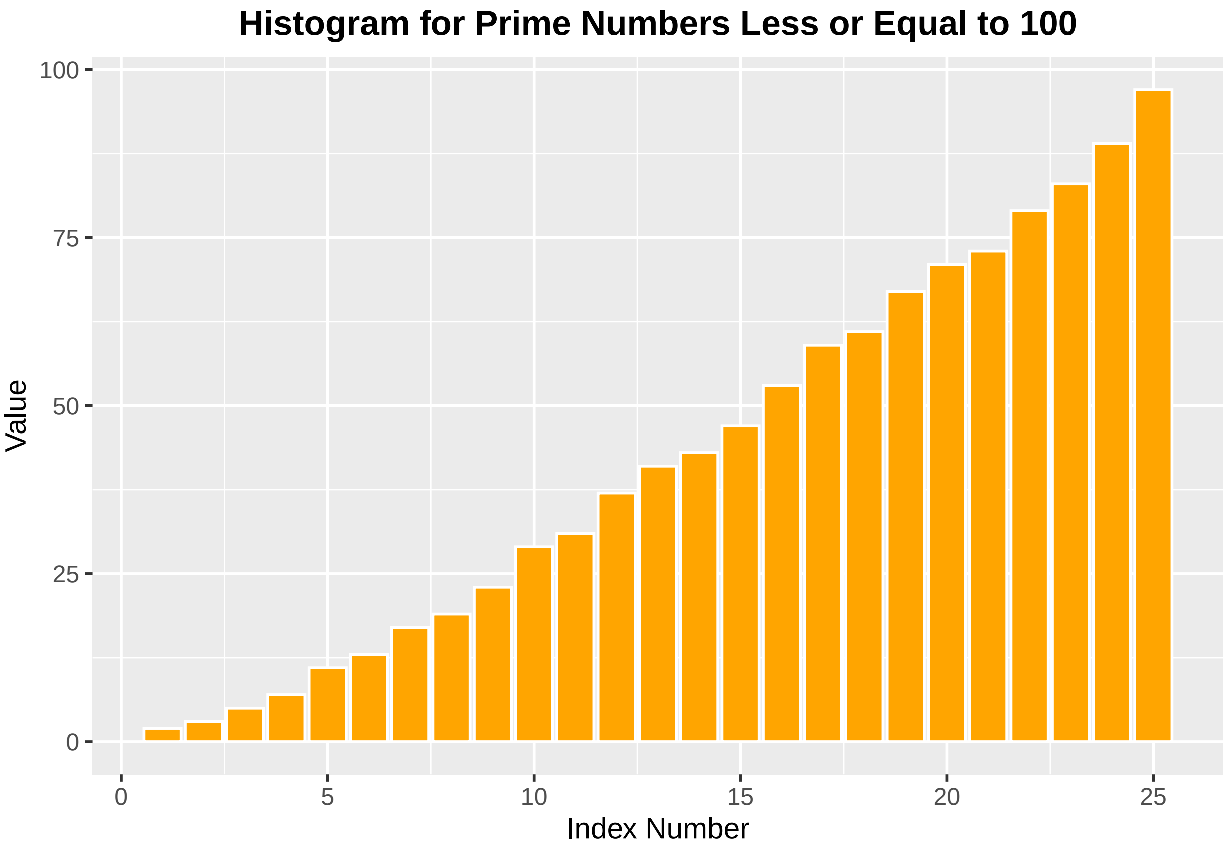
1. Using R, write a script that calculates all the prime numbers less than or equal to 100. A prime number is a positive integer greater than 1 that is only divisible (without remainder) by 1 and itself. Hint: write a loop that tests each number i from 2 to 100 against all integers less than i using the modulo function %%; for any given i, if no integer less than i divides i evenly then i is prime and should be added to your vector of primes. (15 pt)

```
prime_numbers <- function(x)
{
# numbers that are less than 2 are not prime numbers
  if(x < 2)
    return(FALSE)
# create empty vector to store results.
  primes <- vector()
# run the for loop to test for prime numbers beginning from 2
  for(i in 2:x)
    if(!any(i %% primes == 0))
      primes <- c(primes, i)
  return(primes)
}
# assign the results in a new variable for the histogram
prime100 <- prime_numbers(100)

# print all prime numbers from 2 up to 100
prime_numbers(100)
```

```
##  [1]  2  3  5  7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71 73 79 83 89 97
```

2. Using R, create a histogram of the result from 1 using ggplot. Be sure to nicely label your axes and title the graph. (5pt)

```
library(ggplot2)
indexnum <- seq(1, 25)
prime100 <- as.data.frame(cbind(indexnum, prime100))

ggplot(data = data.frame(prime100), mapping = aes(x = indexnum, y = prime100)) +
  geom_bar(stat = "identity",col="white",fill="orange") + xlab("Index Number") +
  ylab("Value") + ggtitle("Histogram for Prime Numbers Less or Equal to 100") +
    theme(plot.title = element_text(hjust = 0.5, face="bold"),
          plot.subtitle = element_text(hjust = 0.5))
```



3. You flip a coin five times.

a. What's the chance of getting three or more heads in a row? (5 pt)

The total observations = 2ˆ5 = 32
There are 8 different possible outcomes where you get 3 consecutive heads or more:

$$(HHHHH, HHHHT, THHHH, HHHTH, HTHHH, HHHTT, TTHHH, THHHT) = 8$$

$$P(3\ or\ more\ H) = \frac{8}{32} = 0.25$$

b. What's the chance of getting three or more heads in a row conditional on knowing the first flip was a heads? (5 pt) (Hint: the best approach here is to count up events in sample space.)

Based on the fact that we know that first flip was a head, the total observations becomes = 2ˆ4 = 16
There are then 5 different possible outcomes where you get 3 consecutive heads or more:

$$(HHHHH, HHHHT, HTHHH, HHHTH, HHHTT) = 5$$

$$P(3\ or\ more\ H) = \frac{5}{16} = 0.3125$$

4. NASA has declared that the Earth is likely to be hit by an asteroid this year based on an astronomical observation it has made. These things are hard to judge for certain, but it is known that the test NASA used is pretty good – it has a sensitivity of 99% and a false positive rate of only 1%. It is further known that the general probability of an asteroid hitting earth in any given year is 1 in 100,000. What is the probability we will actually be hit by an asteroid this year given NASA's test? (10 pt)

To answer this, plug what we know into Bayes' equation:
P(Hit) = 0.00001
P(+|Hit) = 0.99

P(+|No Hit) = 0.01
P(No Hit) = 1 - 0.00001 = 0.99999

$$P(Hit|+) = \frac{P(Hit)\ P(+|Hit)}{P(Hit)\ P(+|Hit)\ +\ P(No\,Hit)\ P(+|No\,Hit)}$$

$$P(Hit|+) = \frac{(0.00001)(0.99)}{(0.00001)(0.99) + (0.99999)(0.01)} = \frac{0.0000099}{0.0100098} = 0.098\%$$

5. The average number of snow days in Boston in a winter month is 1. Assuming these events follow a poisson distribution, calculate (using R) the probability of getting 5 or more snow days in a month. (5 pt)

To answer this, I used ppois() function to calculate the probability of getting snow 5 or more times a month by first calculating the chance of getting snow 4 or fewer times (which is my first argument =4) and with a mean of 1 (my second argument). Then I took 1 minus the ppois() function and named it variable "snowdays" to output the answer.

```r
snowdays <- 1- ppois(4,1)
snowdays
```

```
## [1] 0.003659847
```

6. You want to know how many hours of sleep the average college student gets. You start out with a preliminary survey of 10 people, and get the following data (in hours): 7,6,5,8,6,6,4,5,8,7. You hypothesize that despite what doctors recommend, the average college student does not get 7 hours of sleep a night. What does your survey say about your hypothesis? State your null hypothesis, research hypothesis (two tailed), and calculate your threshold value, test statistic, and p value (be sure to show your work). Do you reject the null or not? (10 pt)

$$H_a : \mu \neq 7$$
$$H_0 : \mu = 7$$

Let's first calculate the sample mean.

$$7 + 6 + 5 + 8 + 6 + 6 + 4 + 5 + 8 + 7 = 62$$
$$\frac{62}{10} = 6.2$$
$$\bar{x} = 6.2$$

Then, let's calculate the standard deviation.

$$(7 - 6.2)^2 = 0.64$$

$$(6 - 6.2)^2 = 0.04$$

$$(5 - 6.2)^2 = 1.44$$

$$(8 - 6.2)^2 = 3.24$$

$$(6-6.2)^2 = 0.04$$
$$(6-6.2)^2 = 0.04$$

$$(4-6.2)^2 = 4.84$$

$$(5-6.2)^2 = 1.44$$

$$(8-6.2)^2 = 3.24$$

$$(7-6.2)^2 = 0.64$$
$$0.64 + 0.04 + 1.44 + 3.24 + 0.04 + 0.04 + 4.84 + 1.44 + 3.24 + 0.64 = 15.6$$

$$(\frac{1}{10}) * 15.6 = 1.56$$
$$SD = \sqrt{1.56}$$
$$SD = 1.25$$

Now, we can calculate the SE.

$$SE = \frac{SD}{\sqrt{n}}$$

$$SE = \frac{1.25}{\sqrt{10}} = 0.40$$

Our Test statistics will then be:

$$Test\ statistic = \frac{\bar{x} - \mu_0}{SE}$$
$$Test\ statistic = \frac{6.2 - 7}{0.40} = -2$$

To verify calculated results, here are the summary data calculated by R:

```
avgsleep <- c(7,6,5,8,6,6,4,5,8,7)

n <- length(avgsleep)
sample_mean <-mean(avgsleep)
Sd <- sd(avgsleep)
se <- Sd / sqrt(n)

# print summary data
n
```

```
## [1] 10
```

```
sample_mean
```

```
## [1] 6.2
```

```
Sd
```

```
## [1] 1.316561
```

```
se
```

```
## [1] 0.4163332
```

The Threshold Value:

```
qt(0.975,9) # upper threshold
```

```
## [1] 2.262157
```

```
qt(0.025,9) # lower threshold
```

```
## [1] -2.262157
```

The P-value:

```
2*(1-pt(2,9))# p-value for two-tailed test
```

```
## [1] 0.07655282
```

Results:
Since p-value 0.08 is $>$ 0.05 we fail to reject the null hypothesis. We can also look at the t statistic value. Since our t statistic value = -2 is not greater than 2.26, nor less than -2.26 the threshold value we also fail to reject the null hypothesis.Thus, perhaps we can say that college student do get an average of 7 hours of sleep a night.

7. Despite the disappointing results in 6, you are confident in your hypothesis. Assuming your sample standard deviation and mean do not change and you want to survey as few people as possible, how many additional people would you have to survey to reject the null at the 0.05 level? (5 pt)

To solve this problem, in order to reject the null at the 0.05 level we can increase n in the t-statistics equation to change our current t-statistics (t-stat= -2) for it be to less than or greater than the threshold value +/-2.26. Therefore, we can solve it with simple algebra:

$$Test\ statistic = \frac{\bar{x} - \mu_0}{\frac{sd}{\sqrt{n}}}$$

$$Test\ statistic = \frac{6.2 - 7}{\frac{1.25}{\sqrt{n}}}$$

$$-2.26 = \frac{6.2 - 7}{\frac{1.25}{\sqrt{n}}}$$

$$-2.26\frac{1.25}{\sqrt{n}} = -0.8$$

$$\frac{1.25}{\sqrt{n}} = \frac{0.8}{2.26}$$

$$\frac{1.25}{\sqrt{n}} = 0.354$$

$$\sqrt{n} = 3.531$$

$$n = (3.531)^2 = 12.47$$
$$n = 13$$

Here, we have calculated n = 13. To test if to see if our t-statstics is now greater or less than the threshold value +/-2.26 if we increase our n to 13:

$$Test\ statistic = \frac{6.2 - 7}{\frac{1.25}{\sqrt{13}}}$$

$$Test\ statistic = -2.30$$

```r
2*(1-pt(2.30,12)) # p-value
```

```
## [1] 0.04019757
```

Our t-statistic -2.30 is now less and greater than the threshold value +/-2.26, which means we can reject the null hypothesis. In addition, the p-value is now 0.04 < 0.05, which we can again reject the null. This indicates that we need to sample at least 13 people in total or 3 additional people in order to reject the null (if you used the sd calculated by R, which is 1.316561, then you would need at least 14 people in total or 4 additional people to reject null).

8. You survey the same 10 individuals in the same order during finals period, and this time get the the following responses: 5,4,5,7,5,4,5,4,6,5. Do your data show that college students get significantly less sleep than usual during finals? You may answer this one using any combination of hand calculations and R you prefer. (10 pt)

To see if my data show that college students get significantly less sleep than usual during finals, I will conduct a paired t-test since I am surveying the same 10 individuals in Q6, which means I am testing differences in means with dependent samples.

$$H_a : \mu_{finals} < \mu_{usual}$$
$$H_0 : \mu_{finals} = \mu_{usual}$$

```r
Finals_Hours = c( 5,4,5,7,5,4,5,4,6,5)
Usual_Hours = c(7,6,5,8,6,6,4,5,8,7)


# Create a data frame
avgsleep.hours <- data.frame (group = rep(c("Finals_Hours", "Usual_Hours"), each=10),
                           sleep_hrs = c(Finals_Hours, Usual_Hours)
                           )
# Compute paired t-test
t.test(sleep_hrs ~ group, data= avgsleep.hours, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sleep_hrs by group
## t = -3.6742, df = 9, p-value = 0.005121
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.9388174 -0.4611826
## sample estimates:
## mean of the differences
##                    -1.2
```

```
qt(0.975,9) # upper threshold for two-tailed test.
```

```
## [1] 2.262157
```

```
qt(0.025,9) # lower threshold for two-tailed test.
```

```
## [1] -2.262157
```

```
2 * pt(3.67,9,lower.tail = FALSE) # P-value
```

```
## [1] 0.005154842
```

Results: Since p-value 0.005 is < 0.05 we reject the null hypothesis. We can also look at the t statistic value. Since our t statistic value is greater than 2.26 and less than -2.26 the threshold value we again reject the null hypothesis. Thus, perhaps we can say that college student do get significantly less sleep during finals.

9. You are a very bad gardener, and hypothesize that feeding houseplants vodka might help them relax and grow better. You perform an experiment to test your hypothesis, giving 15 houseplants water spiked with vodka, and 15 houseplants water alone. These are your results:

condition	*live*	*die*
treatment	4	11
control	8	7

This looks pretty bad for the treatment, but being better at statistics than you are at gardening, you test it using the chi-square test. What are your results? Please do this by hand and show your work, though you may confirm your results using R. (15 pt)

Percent treatment = 15/30 = 0.5
Percent control = 15/30 = 0.5
Percent live = 12/30 = 0.4
Percent die= 18/30= 0.6

P(live&treatment) = 0.4 * 0.5 * 30 = 6, thus the total number of live houseplants we would expect to see in treatment group is 6
P(live&control) = 0.4 * 0.5 *30 = 6, thus the total number of live houseplants we would expect to see in control group is 6*
P(die&treatment) = 0.6 0.5 * 30 = 9, thus the total number of dead houseplants we would expect to see in treatment group is 9
P(die&control) = 0.6 * 0.5 * 30 = 9, thus the total number of dead houseplants we would expect to see in control group is 9

Next, let's calculate the test statistic:

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$X^2 = \frac{(4-6)^2}{6} + \frac{(11-9)^2}{9} + \frac{(8-6)^2}{6} + \frac{(7-9)^2}{9} = 2$$

Our test statistics is 2

$$X^2 = 2$$

Our degrees of freedom:

$$df = (r-1)(c-1)$$

$$df = (2-1)(2-1) = 1$$

```
qchisq(0.95,df=1) # threshold value
```

```
## [1] 3.841459
```

```
1-pchisq(2,df=1) # p-value
```

```
## [1] 0.1572992
```

1.
$$H_a = Variables\ are\ not\ independent$$
$$H_o = Variables\ are\ independent$$
2. Since Xˆ2= 2 < 3.8 threshold value, we fail to reject Null.
3. Since p-value 0.157 > 0.05, we again fail to reject Null.

Therefore, we fail to reject the null hypothesis and accept our null hypothesis, which indicates that the treatment and control groups are independent. This suggests that whether a houseplant lives or dies is not dependent on what they are watered with. This may be due to the small sample size.

10. Perhaps you got things backwards, and plants need more stimulation to thrive. So you adjust your experiment into three treatment groups: water, vodka, and coffee. These are your results:

condition	*mean days alive*	*sd*	*n*
water	50	10	20
vodka	45	7	10
coffee	55	4	10

The overall mean is 50 days. Use an F test to determine if there is any significant difference among these three groups. Please do this by hand and show your work. (15 pt)

$$H_a = The\ average\ days\ alive\ is\ different\ across\ the\ treatment\ groups$$
$$H_o = The\ average\ days\ alive\ is\ the\ same\ across\ the\ treatment\ groups$$

$$f_{stat} = \frac{average\ variance\ between\ groups}{average\ variance\ within\ groups}$$

$$Between\ Groups = \frac{n_1(y_1 - y)^2 + ...n_G(y_G - y)^2}{df = G - 1}$$

$$Within\,Groups = \frac{(n_1 - 1)s_1^2 + ...(n_G - 1)s_G^2}{df = N - G}$$

$$Between\,Groups = \frac{20(50-50)^2 + 10(45-50)^2 + 10(55-50)^2}{3-1} = 250$$

$$Within\,Groups = \frac{(20-1)(10)^2 + (10-1)(7)^2 + (10-1)(4)^2}{40-3} = 67.162$$

$$F\,Statistic = \frac{250}{67.162} = 3.7$$

$$df_1 = 3 - 1 = 2$$

$$df_2 = 40 - 3 = 37$$

```r
qf(0.95,2,37)  #threshold value for F test.
```

```
## [1] 3.251924
```

```r
1-pf(3.7,2,37) #p-value
```

```
## [1] 0.03428838
```

Results:
1. Since F-stat= 3.7 is greater than 3.25 threshold value, we reject the null.
2. Since p-value= 0.03 $<$ 0.05, we again reject the null.

The results suggests that there is a significant difference in mean days alive among these three groups, such that, the mean days alive of my houseplants were at least affected by one of these treatment groups.