IntroComp_Midterm_WeixuanChen

Weixuan Chen

2/26/2023

```
library(ggplot2)
```

1

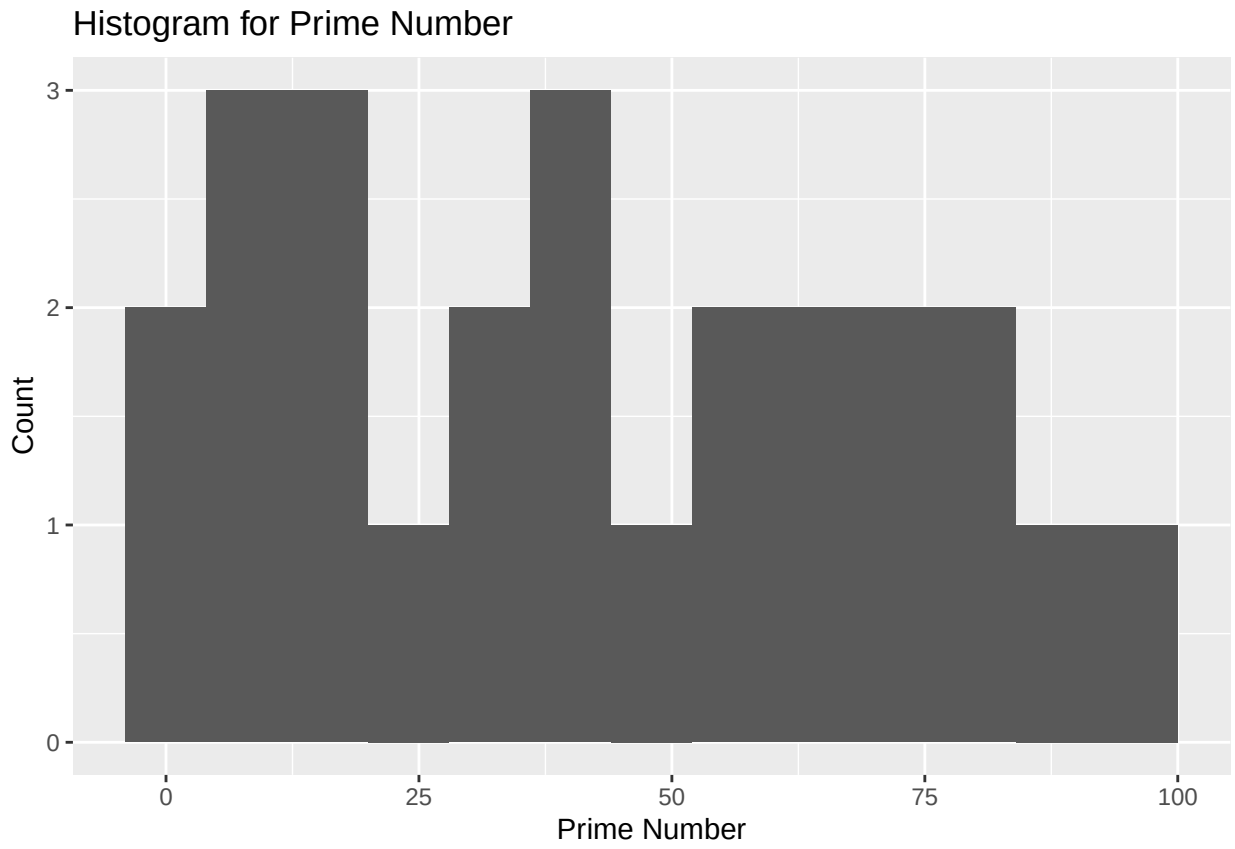
```
prime <- c()
for(i in 2:100){
  if(i == 2){
    prime <- append(prime, i)
  }
  else if(any(i %% 2:(i-1) == 0)){
    next
  }
  else{
    prime <- append(prime, i)
  }
}

prime
```

```
##  [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

```
df_prime <- as.data.frame(prime)
ggplot(df_prime,aes(x=prime)) +
  geom_histogram(binwidth = 8) +
  xlab('Prime Number') +
  ylab('Count') +
  ggtitle('Histogram for Prime Number')
```



3

a) We have 5 tosses, so the total sample space has $2^5 = 32$ combinations. For the event of 3 or more heads in a row, we can consider p(3 or more heads in a row) = p(3 heads in a row) + p(4 heads in a row) + p(5 heads in a row). For p(3 heads in a row), we have {HHHTT, THHHT, TTHHH, HHHTH, HTHHH} 5 outcomes. For p(4 heads in a row), we have {HHHHT, THHHH} 2 outcomes, and for p(5 heads in a row), we have {HHHHH} 1 outcome. Therefore, we have $\frac{(5+2+1)}{32} = \frac{8}{32} = \frac{1}{4}$.

b) If we add the condition to our probabilities, we have p(3 or more heads in a row | first toss is head) = p(3 heads in a row | first toss is head) + p(4 heads in a row | first toss is head) + p(5 heads in a row | first toss is head). So based on our previous analysis, we have to narrow down the possibilities, because now we know the first toss is head. So the remaining outcomes are{HHHTT, HHHTH, HTHHH, HHHHT, HHHHH} 5 outcomes. The probability is $\frac{5}{32}$

4

p(hit by asteriod | NASA test positive) = p(NASA test positive| hit by asteriod)*p(hit by asteriod)/p(NASA test positive), where p(NASA test positive) = p(NASA test positive| hit by asteriod)*p(hit by asteriod) + p(NASA test positive| not hit by asteriod)*p(not hit by asteriod)

```
pNASA.H = 0.99
pH = 1/100000
pNASA.NH = 0.01
p.NH = 1 - pH
```

```r
pNASA = pNASA.H*pH + pNASA.NH*p.NH

pH.NASA = pNASA.H*pH/pNASA
pH.NASA
```

```
## [1] 0.0009890307
```

5

p(5 or more snow days in a month) = 1 - p(less than 5 snow days in a month)

```r
## 4 or less snow days in a month
ppois(4, lambda = 1)
```

```
## [1] 0.9963402
```

```r
## 5 or more snow days in a month
1 - ppois(4, lambda = 1)
```

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

```r
ppois(4, lambda = 1, lower=FALSE)
```

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

6

H0: Average sleep $\mu_{sleep} = 7$ H1: Average sleep $\mu_{sleep} \neq 7$

And we have few samples so use t-test.

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

sleep_mean <- mean(sleep_time)
sleep_sd <- sd(sleep_time)
sleep_size <- length(sleep_time)

(sleep_mean - 7)/(sleep_sd/sqrt(sleep_size))
```

```
## [1] -1.921538
```

```r
#Our test
t.test(sleep_time, mu = 7)
```

```
##
##  One Sample t-test
##
## data:  sleep_time
```

```
## t = -1.9215, df = 9, p-value = 0.08684
## alternative hypothesis: true mean is not equal to 7
## 95 percent confidence interval:
##  5.258189 7.141811
## sample estimates:
## mean of x
##       6.2
```

```
#Lower tail
qt(0.025, 9)
```

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

We can see the critical value for $t_{crit,9} = -2.262$, and our test statistic is - 1.92, which is larger than critical value. So we do not reject the null hypothesis.

We can see our calculated t-statistic agree with the compute results.

7

Since we know the critical value is -2.262, we can set $\frac{mean-diff}{sd/\sqrt{n}} = -2.262$ and solve for n. $\frac{mean-diff}{sd/\sqrt{n}} = -2.262$

```
new_size <- (-2.262*sleep_sd/(sleep_mean - 7))^2
new_size
```

```
## [1] 13.85758
```

We need at least 14 people to reject the null. Now we have 10 people, so we need 4 extra people to reject the null hypothesis.

8

Since we have the same group of students, where we can consider the final period as treatment, we can use paired t-test.

```
final_sleep_time <- c(5,4,5,7,5,4,5,4,6,5)

sleep_diff <- sleep_time - final_sleep_time
sleep_diff
```

```
##  [1]  2  2  0  1  1  2 -1  1  2  2
```

```
mean(sleep_diff)
```

```
## [1] 1.2
```

```
var_sleep_diff <- sum((sleep_diff-mean(sleep_diff))^2)/9
sd_sleep_diff <- sqrt(var_sleep_diff)
sd_sleep_diff
```

```
## [1] 1.032796
```

```
mean(sleep_diff)/(sd_sleep_diff/sqrt(10))
```

```
## [1] 3.674235
```

$\bar{d} = \frac{d_1+d_2+...+d_n}{n} = 1.2$ $\sigma = \sqrt{\frac{(d_1-\bar{d})^2+...+(d_n-\bar{d})^2}{n-1}} = \sqrt{\frac{(2-1.2)^2+...+(2-1.2)^2}{10-1}} = 1.032796$ $t = \frac{1.2}{1.032796/\sqrt{10}} = 3.674235$

```
t.test(sleep_time, final_sleep_time, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sleep_time and final_sleep_time
## t = 3.6742, df = 9, p-value = 0.005121
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  0.4611826 1.9388174
## sample estimates:
## mean difference
##             1.2
```

We can see the p-value is strongly less than 0.01 level. So we have enough evidence to reject the null hypothesis and conclude that college students get significantly less sleep than usual during finals.

Our result agrees with the result from computer.

9

```
live <- c(4, 8)
die <- c(11,7)

plant_watering <- as.data.frame(live)
plant_watering['die'] = die
rownames(plant_watering) <- c('treatment', 'control')
plant_watering
```

```
##           live die
## treatment    4  11
## control      8   7
```

percent treatment $= \frac{15}{30} = \frac{1}{2}$ = percent control percent live $= \frac{12}{30} = \frac{2}{5}$ percent die $= \frac{18}{30} = \frac{3}{5}$

p(live&treatment) = 0.5 * 0.4 = 0.2 p(live&control) = 0.5 * 0.4 = 0.2 p(die&treatment) = 0.5 0.6 = 0.3 p(die&control) = 0.5 * 0.6 = 0.3

```
live_exp <- c(30*0.2, 30*0.2)
die_exp <- c(30*0.3,30*0.3)

plant_watering_exp <- as.data.frame(live_exp)
plant_watering_exp['die'] = die_exp
rownames(plant_watering_exp) <- c('treatment', 'control')
plant_watering_exp
```

```
##           live_exp die
## treatment        6   9
## control          6   9
```

$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(4-6)^2}{6} + \frac{(11-9)^2}{9} + \frac{(8-6)^2}{6} + \frac{(7-9)^2}{9} = \frac{4}{6} + \frac{4}{9} + \frac{4}{6} + \frac{4}{9} = \frac{8}{6} + \frac{8}{9} = \frac{24}{18} + \frac{16}{18} = \frac{40}{18} \approx 2.22$

```
qchisq(0.95, df=1)
```

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

```
chisq.test(plant_watering, correct = F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  plant_watering
## X-squared = 2.2222, df = 1, p-value = 0.136
```

We can see our result matches the computer result. And our p-value is 0.136 (or $2.22 < 3.84$), so we do not reject the null hypothesis and conclude that there is not enough evidence to show the watering and live/die has dependence.

10

For this question, we have to calculate the Between Variance and Within Variance. For the data given, we have:

```
mean_days_alive <- c(50,45,55)
sd_alive <- c(10,7,4)
n_alive <- c(20,10,10)

plant_alive <- as.data.frame(mean_days_alive)
plant_alive['sd_alive '] <- sd_alive
plant_alive['n_alive'] <- n_alive
rownames(plant_alive) <- c('water', 'vodka', 'coffee')
plant_alive
```

```
##        mean_days_alive sd_alive  n_alive
## water               50        10      20
## vodka               45         7      10
## coffee              55         4      10
```

$\text{BV} = \frac{20(50-50)^2 + 10(45-50)^2 + 10(55-50)^2}{3-1} = \frac{250+250}{2} = 250$ $\text{WV} = \frac{(20-1)10^2 + (10-1)7^2 + (10-1)4^2}{40-3} = \frac{1900+441+144}{37} \approx 67.1621622$
$\text{F} = \frac{BV}{WV} = \frac{250}{67.1621622} \approx 3.722334$

```
qf(0.95, 2, 37)
```

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

We have our test statistic 3.72, which is greater than 3.25. So we can reject the null hypothesis and conclude that we have enough evidence to show that the mean of three groups are different.(there is significant difference between groups)