Exam 2

Ayush Jain

04/03/22

```
library(tidyverse)
library(tidymodels)
```

Load the data here:

```
sleep <- read.csv("sleep.csv")
```

Exercise 1

```
sleep %>%
  filter(Run != 0.0)
```

```
##    ï..Date Bedtime TTS TST TBT Alc Cal Run
## 1   29-Apr    1305  25 443 505   0   0 2.5
## 2    1-May    1230  30 393 500   0   0 1.6
## 3    3-May    1330  30 424 511   0   0 2.5
## 4    7-May    1230  55 406 540   0   0 2.9
## 5    9-May    1235  35 450 535   0   0 2.5
## 6   11-May    1255  40 430 515   0   0 3.2
## 7   25-May    1220  25 473 520   0   0 2.7
## 8   27-May    1230  30 480 540   0   1 3.2
## 9   22-Jun    1145  20 465 525   0   0 2.4
## 10  24-Jun    1212  25 508 568   0   0 3.1
## 11  26-Jun    1244  20 453 495   0   0 2.9
## 12  28-Jun    1215  20 478 510   0   0 2.9
## 13   1-Jul    1238  20 442 477   1   0 2.9
## 14   3-Jul    1250  25 459 512   0   0 3.5
## 15   5-Jul    1148  15 458 497   1   0 3.7
## 16   8-Jul    1222  15 493 528   1   0 3.1
## 17  10-Jul    1222  10 442 468   0   0 3.1
## 18  12-Jul    1105  35 448 535   0   0 0.4
## 19  13-Jul    1159  30 509 555   0   0 4.8
## 20  15-Jul    1327  25 439 478   0   0 2.9
## 21  18-Jul    1135  45 456 535   0   0 3.1
## 22  20-Jul    1206  25 464 504   0   0 3.2
## 23  22-Jul    1220  20 459 500   0   0 6.0
## 24  24-Jul    1145  25 431 495   0   0 2.9
## 25  26-Jul    1150  35 408 480   0   0 6.4
```

```
## 26   16-Aug    1126   30  490  543   0    1  6.4
## 27   18-Aug    1314   10  473  493   1    1  2.9
## 28   20-Aug    1331   20  410  445   0    0  3.2
## 29   22-Aug    1223   20  437  467   0    1  6.9
## 30   24-Aug    1204   60  422  516   0    1  2.9
## 31   26-Aug    1139   10  533  566   0    1  3.2
## 32   29-Aug    1227   15  446  486   0    0  2.9
## 33   31-Aug    1211   20  472  509   0    1  2.9
```

```
n = 72
ran = 33
pran = ran/n
sleep %>%
  filter(Alc == 1.0) %>%
  mutate(stats = ifelse(Run == 0.0,
                        0, 1)) %>%
  summarise(mean = mean(stats))
```

```
##        mean
## 1 0.6666667
```

```
sleep %>%
  filter(Alc == 0.0) %>%
  mutate(stats = ifelse(Run == 0.0,
                        0, 1)) %>%
  summarise(mean = mean(stats))
```

```
##        mean
## 1 0.4393939
```

Answer: Looking at the dataset, it is clear that alcohol consumption and running are not disjoint events, since there are observations where both occur.

P(ran) = 33/72 = 45.83% P(ran | consumed alcohol) = 66.7% P(ran | no alcohol) = 43.94%

I would be concerned comparing these probabilities because the sample size is not large enough and also because looking at the data set, there are very few (only 6) entries with alcohol consumption, compared to 66 without it

Exercise 2

```
set.seed(3)

sleep %>%
  filter(Run > 0) %>%
  summarize(mean = mean(Run))
```

```
##       mean
## 1 3.324242
```

```
null_dist0 <- sleep %>%
  specify(response = Run) %>%
  hypothesize(null = "point", mu = 3.32) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

null_dist0 %>%
  summarise(lower = quantile(stat, 0.05),
            upper = quantile(stat, 0.95))
```

```
## # A tibble: 1 x 2
##   lower upper
##   <dbl> <dbl>
## 1  2.96  3.69
```

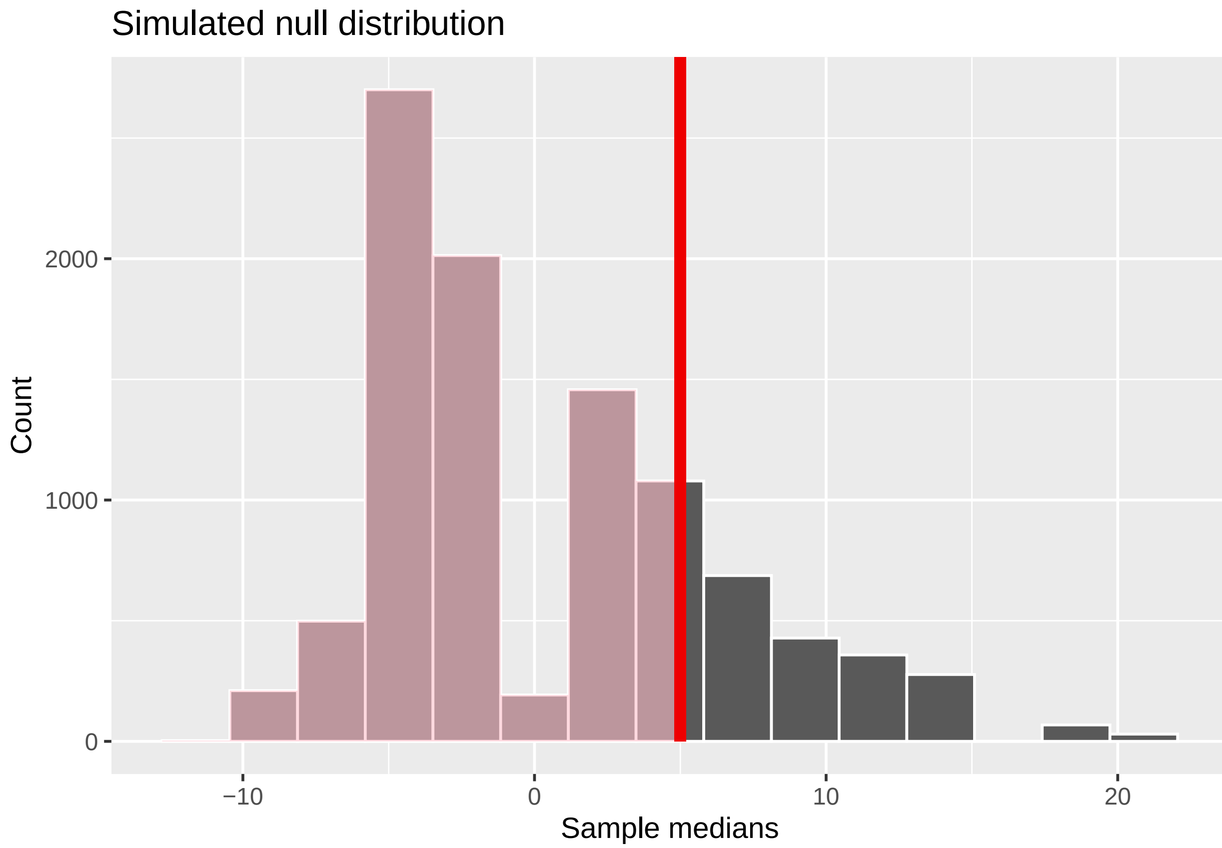
Answer: The mean number of miles ran is 3.32 miles.

No, CLT cannot be used to construct a 90% confidence interval around this mean. The sample size is barely large enough (33) and the observations are not randomly sampled or independent. Also, the sample size is not smaller than 10% of the population size

Generating a 90% confidence interval with bootstrapping gives us a lower value of 2.96 and upper value of 3.69. This means that we are 90% confident that the true mean running distance is between 2.96 and 3.69 miles (given that the individual ran on that day).

Exercise 3

```
set.seed(3)
sleepmaybe <- sleep %>%
  mutate(CalOrNa = ifelse(Cal == 1, "Yes", "No"))
CalYesMedian <- sleep %>%
  filter(Cal == 1) %>%
  summarise(median(TTS)) %>%
  pull()
CalNoMedian <- sleep %>%
  filter(Cal == 0) %>%
  summarise(median(TTS)) %>%
  pull()
DiffMed = CalYesMedian - CalNoMedian
null_dist <- sleepmaybe %>%
  specify(response = TTS, explanatory = CalOrNa) %>%
  hypothesize(null = "independence") %>%
  generate(10000, type = "permute") %>%
  calculate(stat = "diff in medians",
            order = c("Yes", "No"))
visualise(null_dist) +
  shade_p_value(obs_stat = DiffMed, direction = "less") +
  labs(x = "Sample medians",
       y = "Count",
       title = "Simulated null distribution")
```




```
null_dist %>%
  get_p_value(obs_stat = DiffMed, direction = "less")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.815
```

Answer: Null hypothesis: Calcium-Magnesium supplements do not reduce the median time to fall asleep Alternative hypothesis: Calcium-Magnesium supplements reduce the median time to fall asleep

$$H_0 : M_0 - M_1 = 0$$

$$H_1 : M_0 - M_1 < 0$$

where H0 is the null hypothesis, H1 is the alternative hypothesis, M0 is median time with supplements, M1 is median time without supplements

The p-value of 0.8152 is higher than our alpha cut-off (0.05). This means we fail to reject the null hypothesis i.e. calcium-magnesium supplements do not reduce the median time taken to fall asleep

Exercise 4

```
sleepnew <- sleep %>%
  mutate(timeAwake = TBT - TST)
m_main <- linear_reg() %>%
  set_engine("lm") %>%
  fit(timeAwake ~ TTS + Alc + Cal + Run, data = sleepnew)
m_main %>%
  tidy()
```

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    31.7      6.37      4.98  4.80e- 6
## 2 TTS             1.20     0.152     7.93  3.25e-11
## 3 Alc            -9.05     8.55     -1.06  2.94e- 1
## 4 Cal            -3.22     5.73     -0.563 5.76e- 1
## 5 Run            -1.99     1.26     -1.57  1.20e- 1
```

Answer: Our linear model is:

$$\hat{y} = b_0 + b_1\ x_1 + b_2\ x_2 + b_3\ x_3 + b_4\ x_4$$

where y_hat = timeAwake, b0 is the intercept (time awake with no alcohol, no running, no calcium supplements, and fell asleep instantly), b1 is the slope of x1 (time to sleep), b2 is the slope of x2 (alcohol consumption), b3 is the slope of x3 (calcium magnesium supplement consumption), b4 is the slope of x4 (miles ran)

Fitted with values, the model is:

$$\hat{y} = 31.7 + 1.2\ TimeToSleep - 9.0\ Alcohol - 3.2\ Calcium - 2.0\ MilesRan$$

For the Run slope, all other factors held constant, every mile ran reduces the time spent awake by 1.98 (or 2) minutes

Exercise 5

Answer: The null hypothesis for any given predictor is that all else held constant, the given factor has no impact on the time spent awake. The alternative hypothesis is that all else held constant, the given factor does have an impact on the time spent awake.

The only statistically significant factor here is the time to sleep. The extremely low p-value (<0.05) means that we can reject the null hypothesis.

However, the other variables could still be correlated - the p-value is only used to make a conclusion about whether the null hypothesis can be rejected or not. A failure to reject the null hypothesis (as is the case here) does not imply acceptance of the null hypothesis. Further tests could be used to test the correlation. Also, the statistical significance is only determined for a given factor with all others held constant. It is possible that with the addition or removal of a factor, this changes.

Exercise 6

```r
m_main2 <- linear_reg() %>%
  set_engine("lm") %>%
  fit(timeAwake ~ Alc + Cal + Run, data = sleepnew)
m_main2 %>%
  tidy()
```

```
## # A tibble: 4 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   76.0        4.23    18.0   1.58e-27
## 2 Alc          -29.5       11.3     -2.62  1.09e- 2
## 3 Cal           -0.950      7.90    -0.120 9.05e- 1
## 4 Run           -4.93       1.67    -2.96  4.24e- 3
```

```r
glance(m_main)$adj.r.squared
```

```
## [1] 0.5649382
```

```r
glance(m_main2)$adj.r.squared
```

```
## [1] 0.169204
```

Answer: In this model, alcohol and running time also become statistically significant (p-value < 0.05). Glancing at the adjusted r-squared values for both models, one reason for this could be that since they barely explain the model's variability (only about 16.9%), they become statistically significant as compared to the previous model where with the addition of time to sleep, the factors explained about 56.5% of the model's variability. This means that it is likely that the time to sleep was the most important factor. Also, it is possible that without time to sleep being held constant, the significance of the other variables change.

The change in adjusted r-squared values is also why I like the previous model more, because the omission of time to sleep as a factor makes the model less significant

Exercise 7

```r
calculation = 75.9891484 - 29.4727764 - (4.9324321*3.7)
calculation
```

```
## [1] 28.26637
```

Answer: Our new linear model is:

$$\hat{y} = 76.0 - 29.5\ Alcohol - 0.9\ Calcium - 4.9\ MilesRan$$

where y_hat is the outcome (time spent awake) and TimeToSleep, Alcohol, Calcium and MilesRan are the predictors.

For the following values, there is one observation that fits the values. However, the difference between the predicted and actual value gives us a positive residual of 39-28.27 = 10.73