Midterm Exam

Summer 2024

[YOUR NAME]

Due Date: 2024-07-11

Overview

This is your midterm exam. It consists of five questions plus one additional extra credit question.

Survey EC

The extra credit is awarded if you respond to a short survey about this course. The survey is not part of SNU's official teaching evaluations. I use it to help me improve the course in the second half of the month, and respond to your specific needs. To receive the extra credit, take the survey and then copy the secret completion code in the space provided at the bottom of the survey. The survey is **anonymous**, meaning that the completion code is the same for everyone (so please don't share it!). I won't know what you wrote, so please be as honest as possible.

Grading

The exam is worth 15 total points. The point values are indicated next to each question in brackets. The extra credit is worth 3 points, meaning that the maximum possible score is 120%.

Resources

You are permitted to rely on course resources from the first part of the summer intensive session. These include all lecture slides, problem sets, answer keys, homeworks, and lecture notes, as well as any and all posts to Campuswire. You can use ChatGPT as long as you copy the questions you asked into your answers to each question.

You are **not** permitted to review recordings during the midterm.

Codebook

The midterm uses the `fn_cleaned_final.rds` dataset, the codebook for which is reproduced below:

Name	Description
placed	The position between 1 and 100 that the player finished the game (lower is better)
mental_state	Whether the player was drunk or sober when they played
eliminations	How many times they killed an enemy player
assists	How many times they helped a teammate kill an enemy player
accuracy	The proportion of total shots that hit an enemy player

Name	Description
hits	The number of shots that hit an enemy player
head_shots	The number of shots that hit an enemy player in the head
damage_taken	The amount of damage the player received from enemy players

Logistics

You have three hours to complete this exam, although it is designed to only take one hour. You must upload the PDF of the knitted output to the eTL website under the assignment labelled “midterm”. If you need help converting to a PDF, see this tutorial (https://github.com/jbisbee1/ISP_Data_Science_2024/blob/main/Psets/ISP_pset_0_HELPER.pdf).

Good luck!

*Copy the link to ChatGPT you used here: _____

Question 0

Require `tidyverse` and load the `fn_cleaned_final.rds` data to an object called `fn`.

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

```
fn <- read_rds("https://github.com/jbisbee1/ISP_Data_Science_2024/raw/main/data/fn_clean
ed_final.rds")
```

Question 1 [1 point]

How many observations are in this dataset? How many variables? What is the unit of analysis?

```
fn
```

```
## # A tibble: 957 × 8
##    placed mental_state eliminations assists accuracy  hits head_shots
##     <dbl> <chr>               <dbl>   <dbl>    <dbl> <dbl>      <dbl>
##  1     17 sober                   2       0   0.194     10          1
##  2     41 sober                   0       2   0.324     17          0
##  3     36 drunk                   3       0   0.337     38          0
##  4     28 drunk                   1       4   0.105     22          3
##  5      3 drunk                   3       2   0.622     49         18
##  6     15 drunk                   0       1   0.0582     4          3
##  7      9 drunk                   2       2   0.265     43          2
##  8     29 drunk                   3       2   0.272     14          3
##  9     11 drunk                   4       0   0.383     53         13
## 10      1 drunk                   1       2   0.328     27          0
## # ℹ 947 more rows
## # ℹ 1 more variable: damage_taken <dbl>
```

```
summary(fn)
```

```
##      placed      mental_state        eliminations      assists
##  Min.   : 1.00   Length:957         Min.   :0.000   Min.   :0.00
##  1st Qu.: 1.00   Class :character   1st Qu.:1.000   1st Qu.:0.00
##  Median :11.00   Mode  :character   Median :2.000   Median :1.00
##  Mean   :14.58                      Mean   :2.526   Mean   :1.51
##  3rd Qu.:24.00                      3rd Qu.:4.000   3rd Qu.:2.00
##  Max.   :73.00                      Max.   :9.000   Max.   :7.00
##     accuracy           hits          head_shots      damage_taken
##  Min.   :0.0000   Min.   :  0.00   Min.   : 0.000   Min.   :  0
##  1st Qu.:0.1736   1st Qu.: 13.00   1st Qu.: 1.000   1st Qu.:153
##  Median :0.2469   Median : 25.00   Median : 3.000   Median :217
##  Mean   :0.2605   Mean   : 29.95   Mean   : 4.829   Mean   :245
##  3rd Qu.:0.3256   3rd Qu.: 39.00   3rd Qu.: 6.000   3rd Qu.:322
##  Max.   :0.9472   Max.   :118.00   Max.   :36.000   Max.   :749
```

```
glimpse(fn)
```

```
## Rows: 957
## Columns: 8
## $ placed       <dbl> 17, 41, 36, 28, 3, 15, 9, 29, 11, 1, 17, 1, 18, 9, 1, 1, …
## $ mental_state <chr> "sober", "sober", "drunk", "drunk", "drunk", "drunk", "dr…
## $ eliminations <dbl> 2, 0, 3, 1, 3, 0, 2, 3, 4, 1, 1, 3, 2, 4, 4, 5, 3, 1, 3, …
## $ assists      <dbl> 0, 2, 0, 4, 2, 1, 2, 2, 0, 2, 2, 0, 0, 1, 0, 4, 6, 6, 1, …
## $ accuracy     <dbl> 0.19371429, 0.32400265, 0.33653340, 0.10506617, 0.6216160…
## $ hits         <dbl> 10, 17, 38, 22, 49, 4, 43, 14, 53, 27, 11, 33, 22, 42, 40…
## $ head_shots   <dbl> 1, 0, 0, 3, 18, 3, 2, 3, 13, 0, 2, 1, 3, 2, 11, 2, 12, 2,…
## $ damage_taken <dbl> 282, 203, 206, 262, 437, 151, 176, 222, 198, 92, 336, 103…
```

There are 957 observations in the dataset, and 8 variables. Based on looking at the data, it seems that the unit of analysis is a single Fortnite game.

Question 2 [3 points]

Research Question: What is the relationship between mental state and accuracy? Why do you think so? Based on your theory, what is the X and the Y variable?

I expect players who are sober to have a higher accuracy. This is based on the well-documented finding that alcohol reduces reaction time and fine motor skills, both of which are helpful for accuracy. According to this logic, I assume that mental state causes accuracy, which means mental state is the X variable and accuracy is the Y variable.

Question 3 [4 points]

Following the 3-step process discussed in class, provide univariate visualizations of the X and the Y variable (steps 1 and 2). Why did you choose the `geom_...()` that you did for each variable? Are there alternative `geom_...()` choices that would also work?

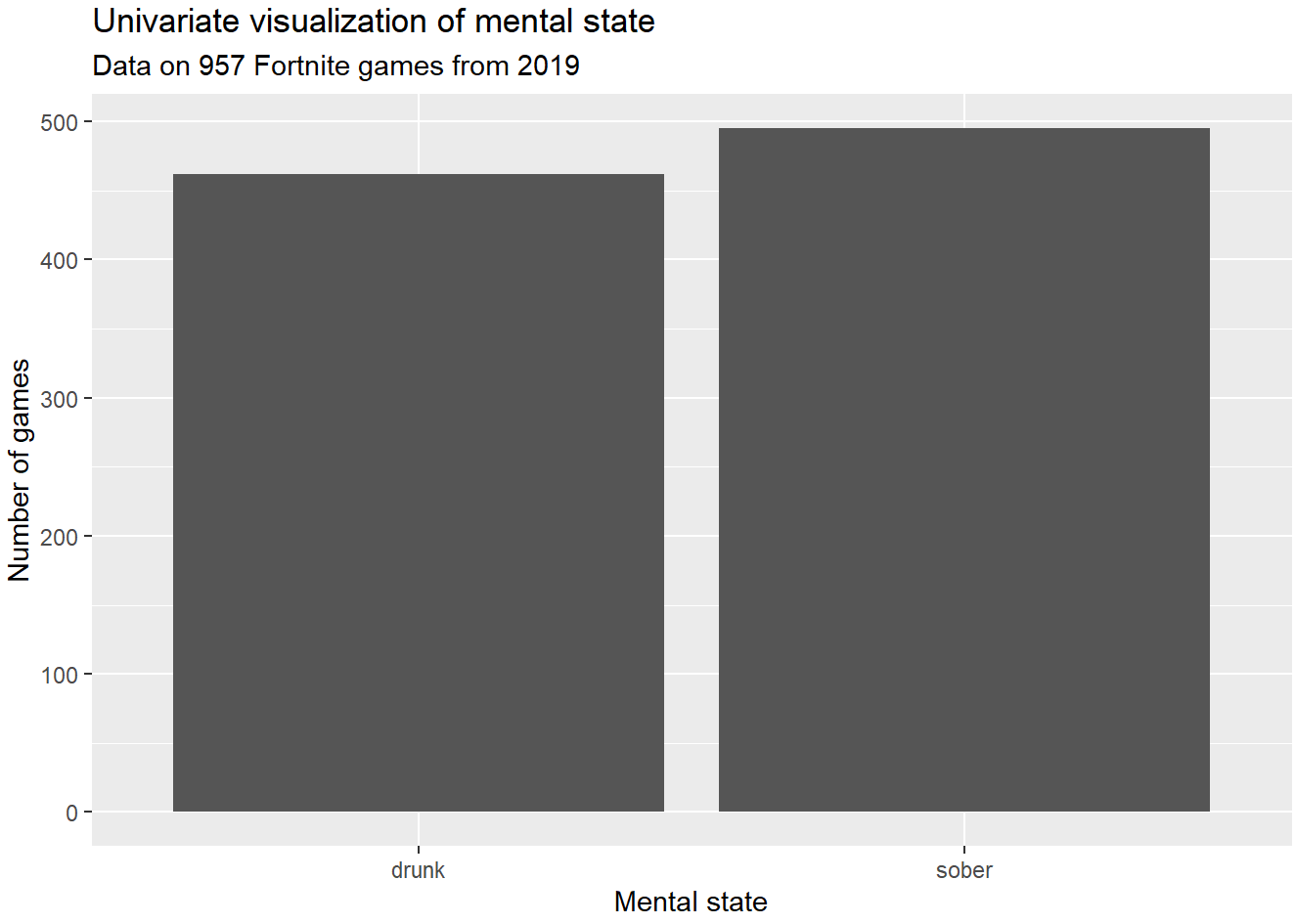
```
# Step 1: Look
fn %>%
  select(mental_state,
         accuracy)
```

```
## # A tibble: 957 × 2
##    mental_state accuracy
##    <chr>           <dbl>
##  1 sober          0.194
##  2 sober          0.324
##  3 drunk          0.337
##  4 drunk          0.105
##  5 drunk          0.622
##  6 drunk          0.0582
##  7 drunk          0.265
##  8 drunk          0.272
##  9 drunk          0.383
## 10 drunk          0.328
## # ℹ 947 more rows
```

```
glimpse(fn %>%
          select(mental_state,
                 accuracy))
```

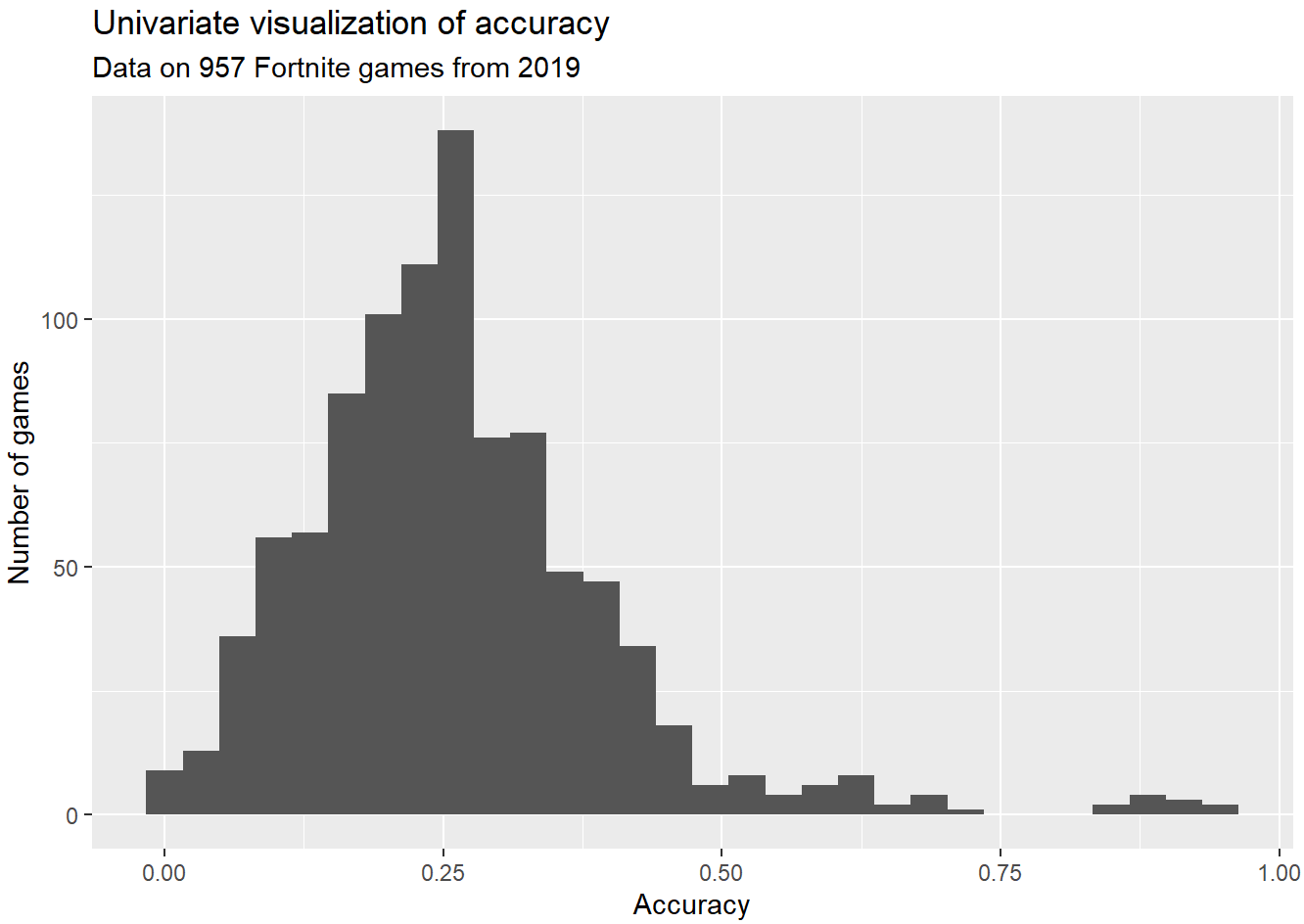
```
## Rows: 957
## Columns: 2
## $ mental_state <chr> "sober", "sober", "drunk", "drunk", "drunk", "drunk", "dr…
## $ accuracy     <dbl> 0.19371429, 0.32400265, 0.33653340, 0.10506617, 0.6216160…
```

```
# Step 2: Visualize
fn %>%
  ggplot(aes(x = mental_state)) +
  geom_bar() +
  labs(x = 'Mental state',
       y = 'Number of games',
       title = 'Univariate visualization of mental state',
       subtitle = 'Data on 957 Fortnite games from 2019')
```




```
fn %>%
  ggplot(aes(x = accuracy)) +
  geom_histogram() +
  labs(x = 'Accuracy',
       y = 'Number of games',
       title = 'Univariate visualization of accuracy',
       subtitle = 'Data on 957 Fortnite games from 2019')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



I chose to use `geom_bar()` for `mental_state` because this variable looks like a binary categorical variable. I chose to use `geom_histogram()` for `accuracy` because this variable looks like a numeric continuous variable. I could have also used `geom_density()` for `accuracy()` , since either works well for visualizing continuous data.

Then analyze the data in two ways. First, calculate the average accuracy for both mental states. Based on this comparison, what do you conclude? Does the data support your hypothesis?

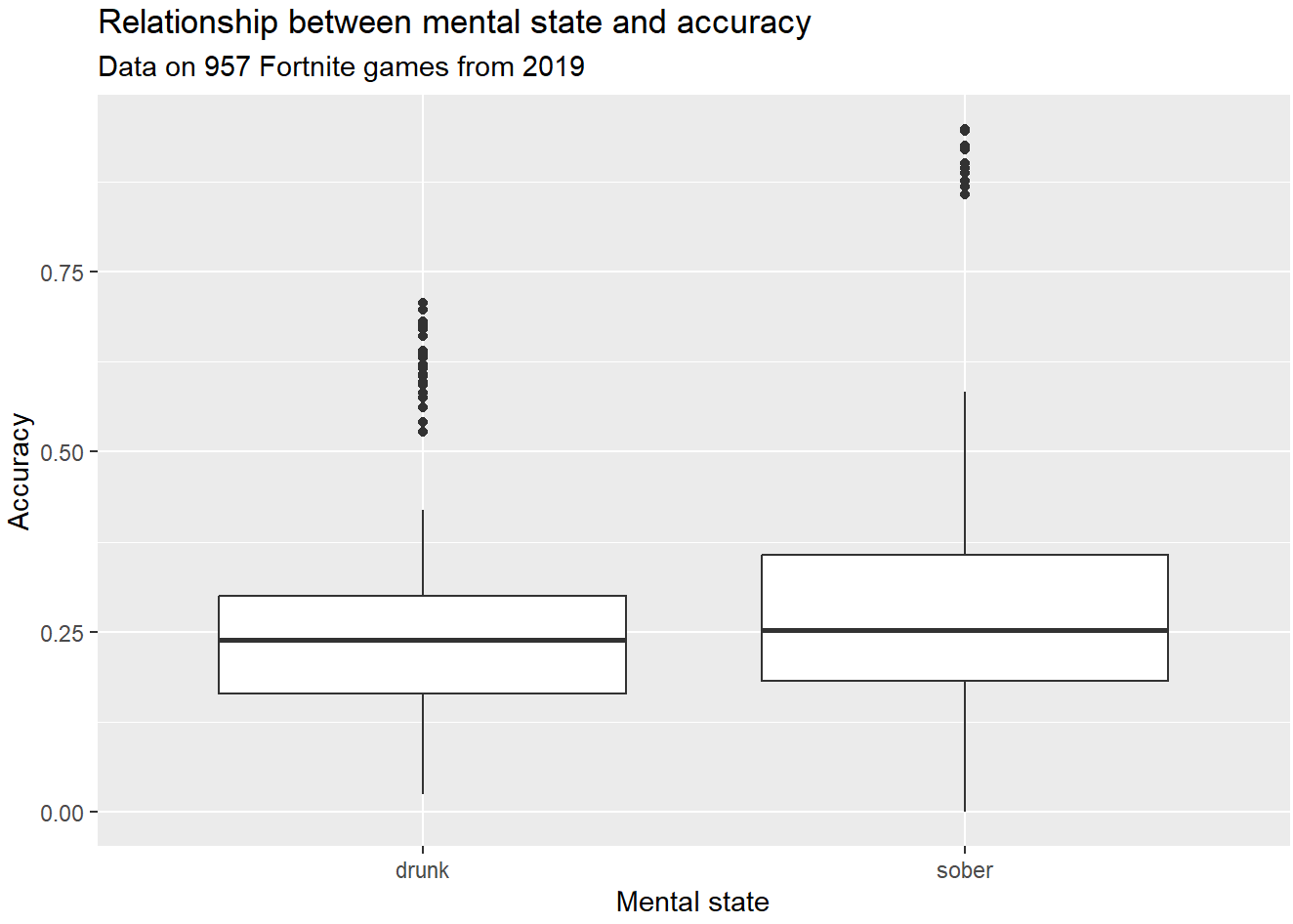
```
fn %>%
  group_by(mental_state) %>%
  summarise(avg_acc = mean(accuracy))
```

```
## # A tibble: 2 × 2
##   mental_state avg_acc
##   <chr>          <dbl>
## 1 drunk          0.245
## 2 sober          0.275
```

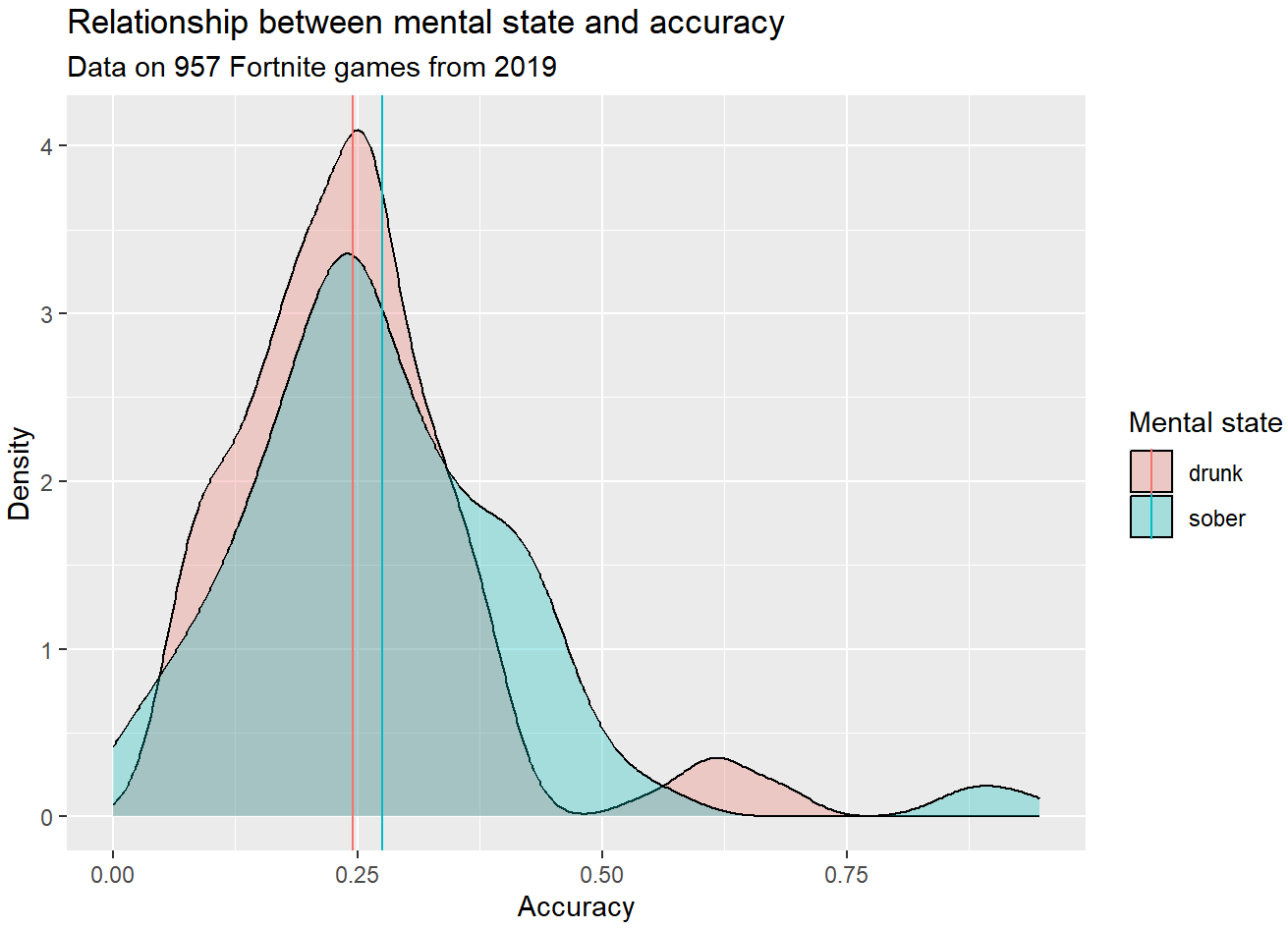
Based on this analysis, I would conclude that sober players have a better accuracy than drunk players. Their accuracy is 27.5% while drunk players is only 24.5%, or 3 percentage points lower.

Second, create a multivariate visualization of the relationship between mental state and accuracy. As before, explain why you chose the `geom_...()` *you did, and discuss whether another approach would also work.*

```
fn %>%
  ggplot(aes(x = mental_state,
             y = accuracy)) +
  geom_boxplot() +
  labs(x = 'Mental state',
       y = 'Accuracy',
       title = 'Relationship between mental state and accuracy',
       subtitle = 'Data on 957 Fortnite games from 2019')
```



```
fn %>%
  ggplot(aes(x = accuracy,
             fill = mental_state)) +
  geom_density(alpha = .3) +
  geom_vline(data = fn %>%
               group_by(mental_state) %>%
               summarise(accuracy = mean(accuracy)),
             aes(xintercept = accuracy,color = mental_state)) +
  labs(x = 'Accuracy',
       y = 'Density',
       fill = 'Mental state',
       color = 'Mental state',
       title = 'Relationship between mental state and accuracy',
       subtitle = 'Data on 957 Fortnite games from 2019')
```



I chose to use `geom_boxplot()` because the X variable is categorical and the Y variable is continuous. I could have also used `geom_density()` and set the fill color to be the mental state. I prefer the boxplot approach because it is easier to compare the two distributions side-by-side instead of overlapping.

Finally, based on all the preceding results, what do you conclude about the research question? Is there a relationship between a player's mental state and their accuracy in the video game Fortnite?

> Based on this analysis, I would conclude that sober players have a better accuracy than drunk players. However, the multivariate visualization suggests that this difference is small.

Question 4 [4 points]

How confident are you in your answer? Use bootstrapped simulations to express your uncertainty as a percentage. Use 100 simulations and set the `size` of each sample equal to the total number of observations in the dataset.

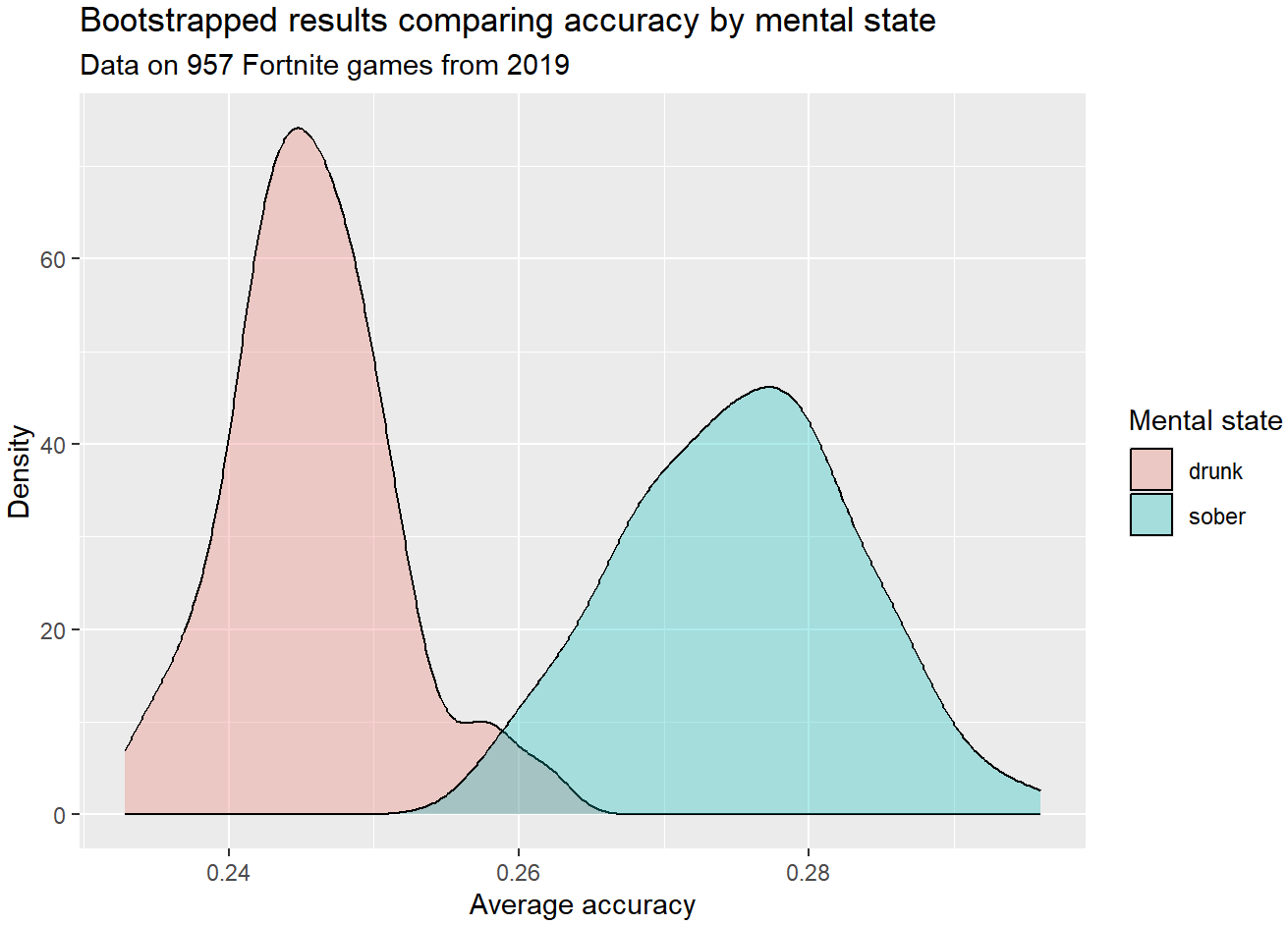
```
set.seed(123)
bsRes <- NULL
for(i in 1:100) {
  simulation <- fn %>%
    sample_n(size = nrow(.),
             replace = T)

  answer <- simulation %>%
    group_by(mental_state) %>%
    summarise(avg_acc = mean(accuracy)) %>%
    mutate(sim_number = i)

  bsRes <- bsRes %>%
    bind_rows(answer)
}

bsRes_wide <- bsRes %>%
  pivot_wider(names_from = 'mental_state',
              values_from = c('avg_acc'))

bsRes_wide %>%
  summarise(conf = mean(drunk < sober))
```

```
## # A tibble: 1 × 1
##    conf
##   <dbl>
## 1     1
```

> I am more than 99% confident that sober players have a higher accuracy than drunk players. This is a little surprising given how similar the two measures looked in the multivariate analysis.

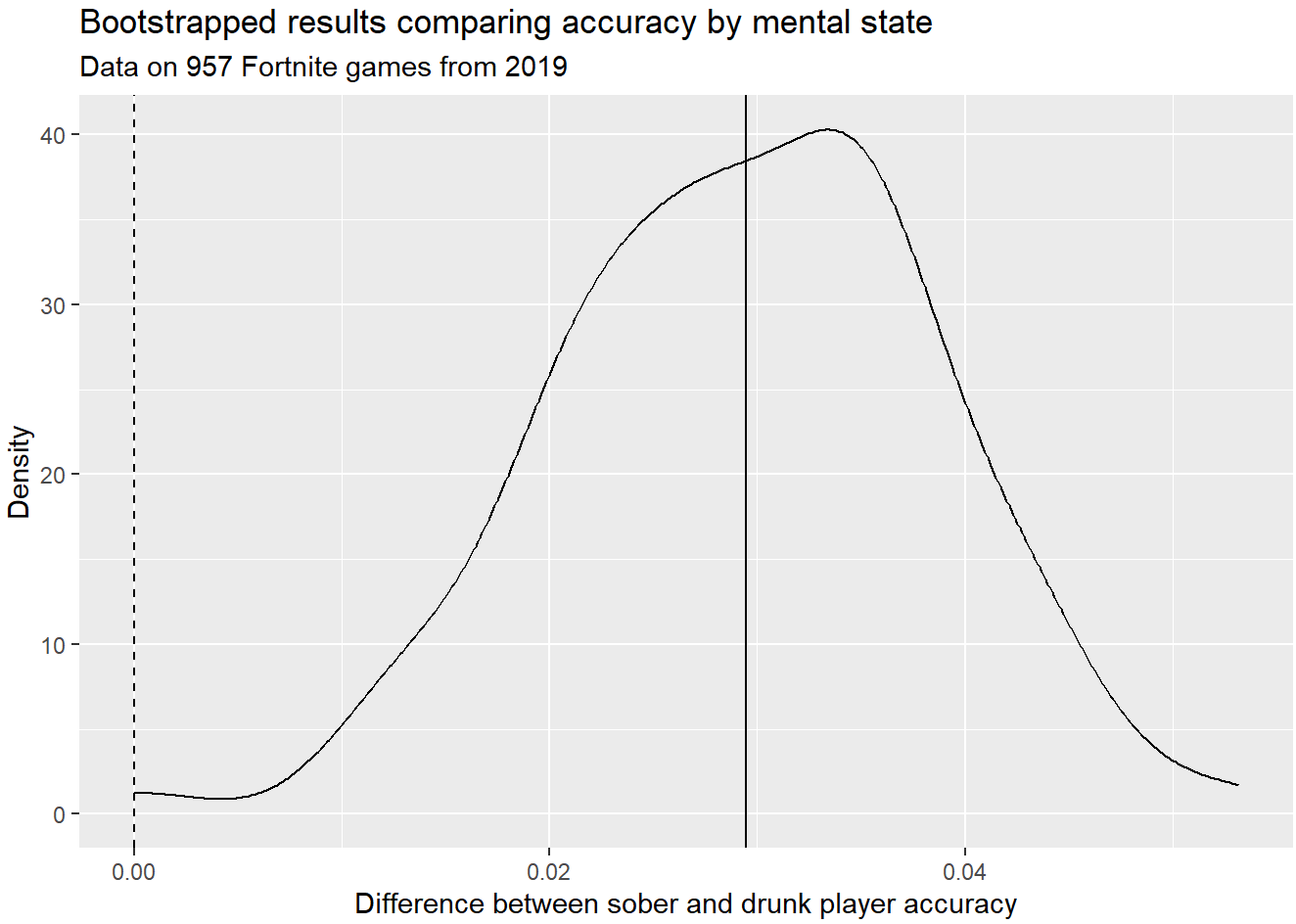
Finally, let's visualize the bootstrapped results. First, create a plot of two densities side-by-side, colored by the X variable.

```
bsRes %>%
  ggplot(aes(x = avg_acc,
             fill = mental_state)) +
  geom_density(alpha = .3) +
  labs(x = 'Average accuracy',
       y = 'Density',
       fill = 'Mental state',
       title = 'Bootstrapped results comparing accuracy by mental state',
       subtitle = 'Data on 957 Fortnite games from 2019')
```



Second, create a single density describing the difference in accuracy by mental state, with a vertical dashed line at zero, and a second vertical solid line at the average difference in accuracy. What is our best guess at the difference between drunk and sober player accuracy?

```
bsRes_wide <- bsRes_wide %>%
  mutate(diff = sober - drunk)

bsRes_wide %>%
  ggplot(aes(x = diff)) +
  geom_density() +
  geom_vline(xintercept = 0,linetype = 'dashed') +
  labs(x = 'Difference between sober and drunk player accuracy',
       y = 'Density',
       title = 'Bootstrapped results comparing accuracy by mental state',
       subtitle = 'Data on 957 Fortnite games from 2019') +
  geom_vline(xintercept = mean(bsRes_wide$diff))
```




```
mean(bsRes_wide$diff)
```

```
## [1] 0.02947934
```

Our best guess is that sober players are 2.9 percentage points more accurate than drunk players.

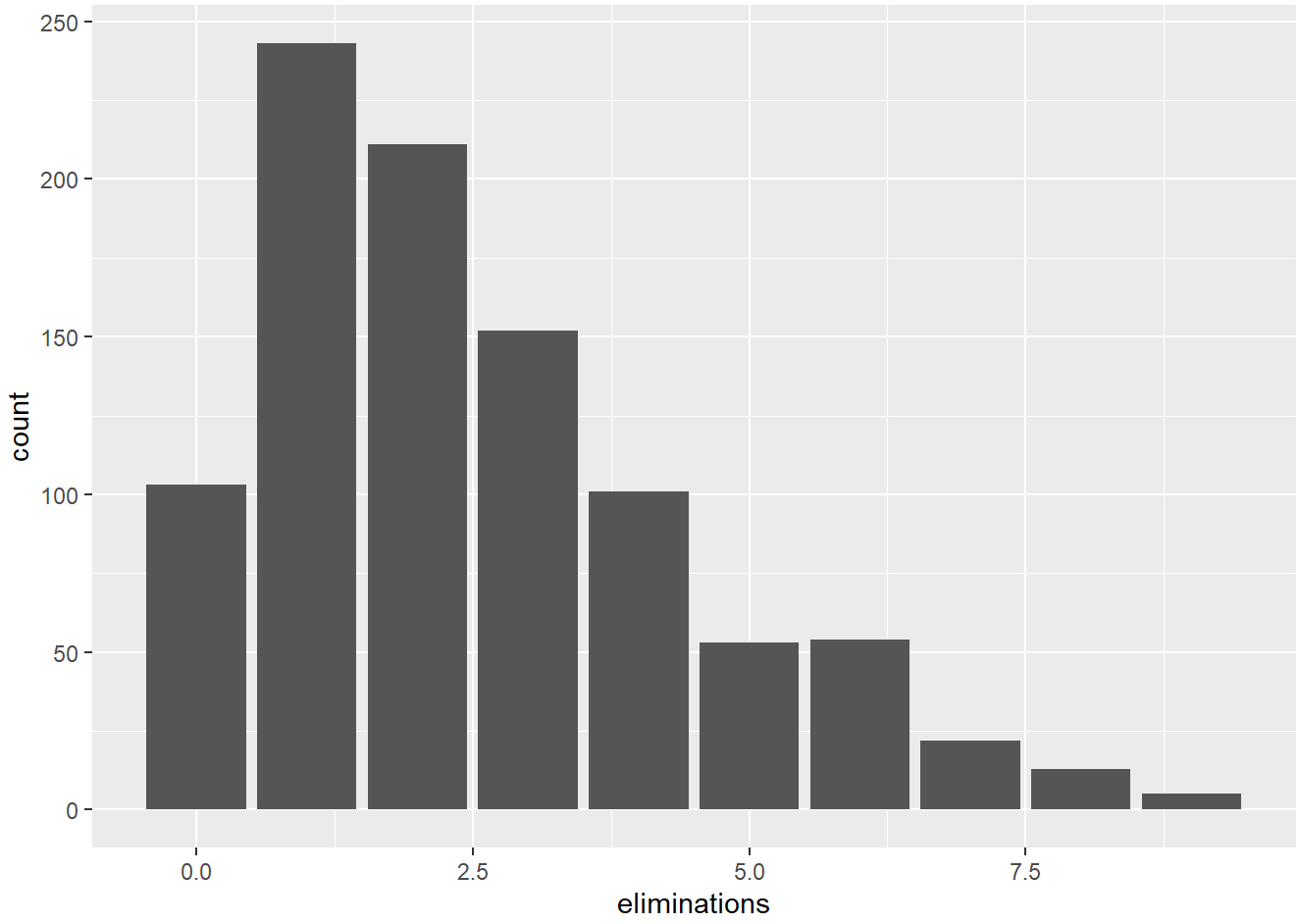
Question 5 [3 points]

Now let's look at a different variable: `eliminations`*. This captures how many times in a game the player kills another player. Does your theory change? Why or why not?*

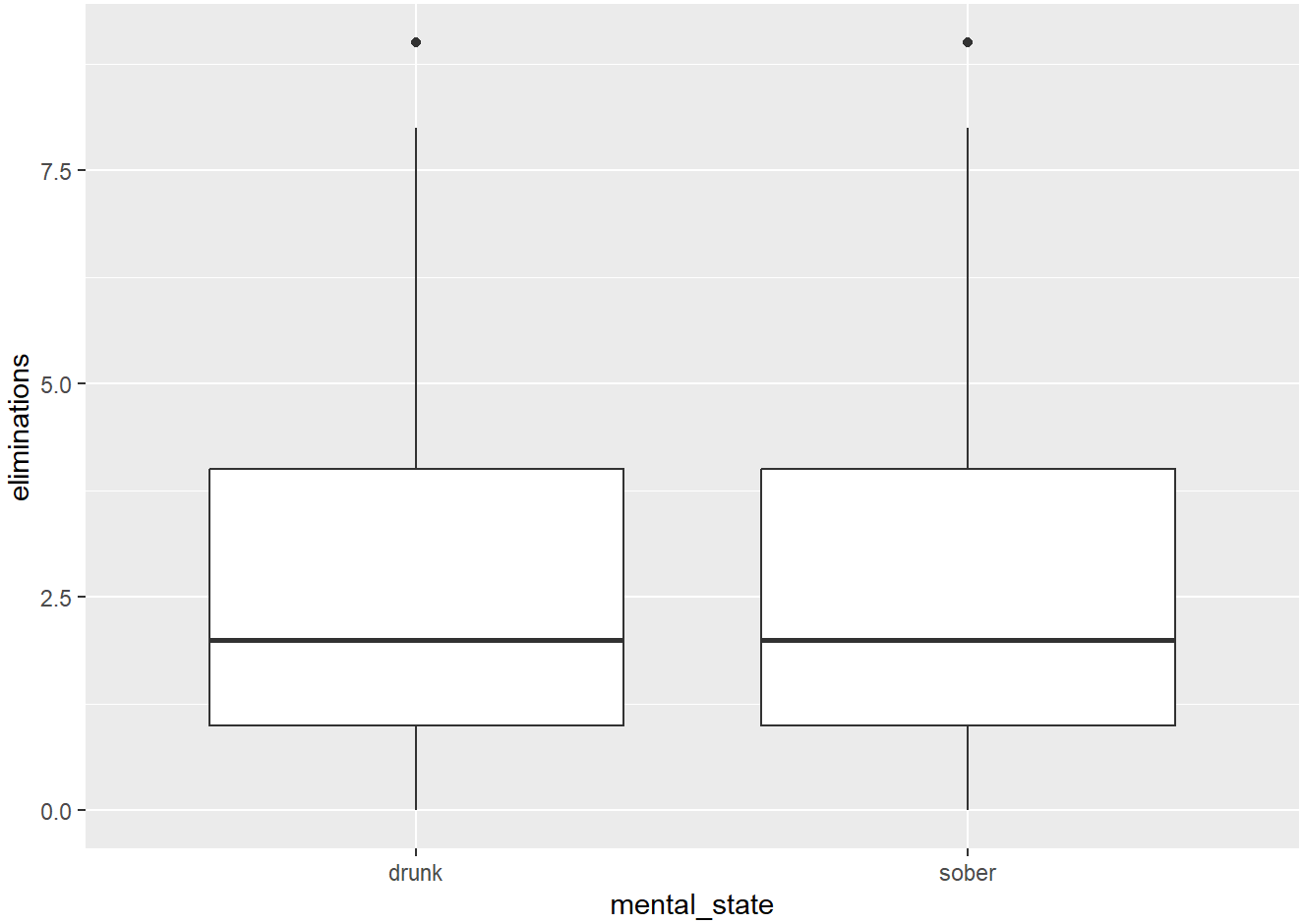
> No my theory doesn't change. Killing another player is a function of accuracy, so if sober players have higher accuracy, they should also have more eliminations.

Provide a univariate visualization for `eliminations`*, choosing the best* `geom_...()`*, and then create a multivariate visualization showing the relationship between* `eliminations` *and* `mental_state`*. In addition, calculate average eliminations by mental state. Does the data support your hypothesis?*

```
fn %>%
  ggplot(aes(x = eliminations)) +
  geom_bar()
```




```
fn %>%
  ggplot(aes(x = mental_state,
             y = eliminations)) +
  geom_boxplot()
```




```
fn %>%
  group_by(mental_state) %>%
  summarise(avg_elim = mean(eliminations))
```

```
## # A tibble: 2 × 2
##   mental_state avg_elim
##   <chr>           <dbl>
## 1 drunk            2.51
## 2 sober            2.54
```

Sober players have 2.54 eliminations compared to drunk players who have 2.51. However, the multivariate visualization makes it very hard to see the difference between these two groups.

Finally, use 100 bootstrapped simulations with size set to the total number of observations in the dataset to express your confidence. You don't need to visualize these results, just provide the confidence number.

```
set.seed(123)
bsRes <- NULL
for(i in 1:100) {
  simulation <- fn %>%
    sample_n(size = nrow(.),
             replace = T)

  answer <- simulation %>%
    group_by(mental_state) %>%
    summarise(avg_elim = mean(eliminations)) %>%
    mutate(sim_number = i)

  bsRes <- bsRes %>%
    bind_rows(answer)
}

bsRes_wide <- bsRes %>%
  pivot_wider(names_from = 'mental_state',
              values_from = c('avg_elim'))

bsRes_wide %>%
  summarise(conf = mean(drunk < sober))
```

```
## # A tibble: 1 × 1
##    conf
##   <dbl>
## 1  0.58
```

> I am only 58% confident that sober players have more eliminations than drunk players.

Extra Credit: Survey [3 points]

Please complete this **anonymous** course evaluation. This does not influence Professor Bisbee's career or position in the university and will only be used to improve the course. You can find the anonymous survey here (https://nyu.qualtrics.com/jfe/form/SV_b7t5vqhhbalgGZ8). Upon completing the survey, you will be given a completion code. To receive the extra credit points, please paste the completion code into the space provided below.

```
cat("D@taSc!enceForEveryone")
```

```
## D@taSc!enceForEveryone
```

NOTE: There is only one completion code to ensure that all responses are anonymized and can't be linked back to the midterm exams. To prevent students from sharing the code with their friends to get the 3 extra credit points without completing the survey, these 3 points are only provided if the number of midterms with the completion code *exactly equals the number of survey responses*. In other words, if there are 25 exams with the completion code,

but only 24 completed surveys, **all students will forfeit their extra credit points**. The purpose of this strict rule is to disincentivize the sharing of this code either by those who would fill out the survey and then share the code, or by those who would ask to be given the code without filling out the survey.