University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAC32 (K. Butler), Midterm Exam
October 4, 2023

Aids allowed (on paper, no computers):

- My lecture overheads (slides)
- Any notes that you have taken in this course
- Your marked assignments
- My assignment solutions
- Non-programmable, non-communicating calculator

This exam has 7 numbered pages plus this cover page.
In addition, you should have a booklet of Figures to refer to during the exam.
The maximum marks available for each part of each question are shown next to the question part.
If you need more space, use the last page of the exam. Anything written on the back of the page will not be graded.
You may assume throughout this exam that the code shown in Figure 1 of the booklet of Figures has already been run.

1. When you are found guilty of wrongdoing, does it help to smile at the person who decides what your punishment is? An experiment was conducted to investigate this. Participants in the experiment pretended to be members of a college disciplinary panel judging students accused of cheating. For each suspect, along with a description of the offence, a picture was provided with either a smile or neutral facial expression. Each participant said what they thought was a suitable punishment based on the evidence they had seen and a leniency score was calculated based on the disciplinary decisions made by the participants. (A higher leniency score means a *smaller* punishment.)

   The data file is shown in Figure 2, and is in the file `smiles.txt` in the same folder as your current R Studio project.

   (a) [3] What R code would read the data from the file into a dataframe called `smiles` and display (at least some of) that dataframe? (Here, and elsewhere in this exam, you do not know what the output from your code is going to be, so I need only the code you would run to carry out the task.)

   (b) [3] What code will make a suitable graph of the two variables in your dataframe? Justify your choice of graph briefly.

   (c) [2] A graph is shown in Figure 3. This may or may not be the same as the graph you gave code for earlier. What does this plot tell you that would be of interest to the researchers who designed this experiment? Explain briefly.

   (d) [3] What code would work out the number of observations in each group, along with the mean leniency score of each group?

2. Pew Research Center conducted a survey in 2018, asking a sample of U.S. adults to categorize five factual and five opinion statements. This dataset provides data from this survey, with information on the age group of the participant as well as the number of factual and opinion statements they classified correctly (out of 5). Some of the data are shown in Figure 4. A total of 5,035 adults were surveyed altogether. The dataframe is called `fact_opinion`.

   (a) [2] Suppose we want to make a graph that will enable us to see which age group has the most respondents in the survey. What code will make such a graph?

   (b) [2] Suppose you run the code shown in Figure 5. What will happen? Explain briefly.

   (c) [3] A plot is shown in Figure 6. What does this plot tell you about how the age groups differ? Explain briefly. (The variable on the x-axis, though quantitative, is treated as ordered categorical for this plot.)

3. The US is divided into ten "health regions" that each contain several states. For each region, for males and females separately and for urban and rural residents separately, the regional mortality (death) rates from various causes are recorded. Some of the data are shown in Figure 7. The dataframe is called `mortality`.

   In the question parts below, give code to display what is requested, unless otherwise stated.

   (a) [2] Display (only) the columns for health region, cause of death and death rate.

   (b) [2] Display the columns whose names begin with S (either uppercase or lowercase), without naming or numbering the columns in your code.

   (c) [2] Select the columns whose names have the letter A in them somewhere (uppercase or lowercase), without naming or numbering them.

   (d) [2] Select the categorical variables (that are text), again without naming or numbering them.

(e) [3] Display the rows that are for health region 04.

(f) [3] For each of the health regions, display the median death rates from heart disease (but not the death rates from any other cause).

(g) [3] Display any death rates that are either over 230 or are for Cancer (or both), along with the cause of death.

(h) [4] Display the two lowest mortality rates and their accompanying causes of death for each health region (which you should also display), but only for females.

4. Shrimp cocktail is a seafood dish consisting of shelled, cooked shrimp in a sauce, and is served in a glass. It used to be a popular starter at restaurants. Shrimp cocktail is required to contain a certain percentage (by weight) of shrimp. Samples of a certain brand of shrimp cocktail were sent to 18 different labs for analysis, with the results shown in Figure 8.

   (a) [2] What code would obtain a 90% confidence interval for the mean percentage of shrimp (by weight)?

   (b) [3] Figure 9 shows the code and output for an analysis of these data. What specifically do you conclude? Explain briefly (by which I mean that your reader should end up convinced that you have drawn an appropriate conclusion).

   (c) [2] A graph of the shrimp percentages is shown in Figure 10. On the basis of this graph and the information given in the question, what are *two* reason why the analysis above is trustworthy? Explain briefly.

5. A manufacturer is concerned about the environmental impact of the smokestack emissions of its factory. In particular, the manufacturer measures the amount of carbon monoxide emitted from the smokestacks of its factory, and from a factory of a competitor, and wants to show that the manufacturer has less of an environmental impact than the competitor. A smaller carbon monoxide emission is better. The data are shown in Figure 11. There are nine observations from the manufacturer's smokestack and ten from the competitor's smokestack (each measured at different times). The dataframe is called `monoxide`.

   (a) [3] A plot is shown in Figure 12. What code was used to make this plot? Why is each part of your code necessary?

   (b) [3] What code would run a suitable t-test for these data? Justify your choice briefly.

   (c) [3] The output from an appropriate test is shown in Figure 13. (Note that this may or may not be the same test as you gave code for in the previous part.) What do you conclude from this output, in the context of the data? (Note that some of the text in the Figure has run off the side of the page and is not visible. Use what you can see.) Explain briefly.

   (d) [2] Why might you have some doubts about running a t-test here?

Use this page if you need more space. Be sure to label any answers here with the question and part they belong to.

Numbered Figures begin here:

```
library(tidyverse)
library(readxl)
```

Figure 1: Packages

Group	Leniency
neutral	6
smile	3.5
smile	4.5
smile	6
smile	4
neutral	2.5
smile	7.5
smile	2.5
smile	3.5
neutral	4
neutral	2.5
neutral	4.5
smile	3.5
smile	9
neutral	3
smile	3
smile	5
neutral	4.5
smile	5.5
smile	5

Figure 2: Smiles leniency data

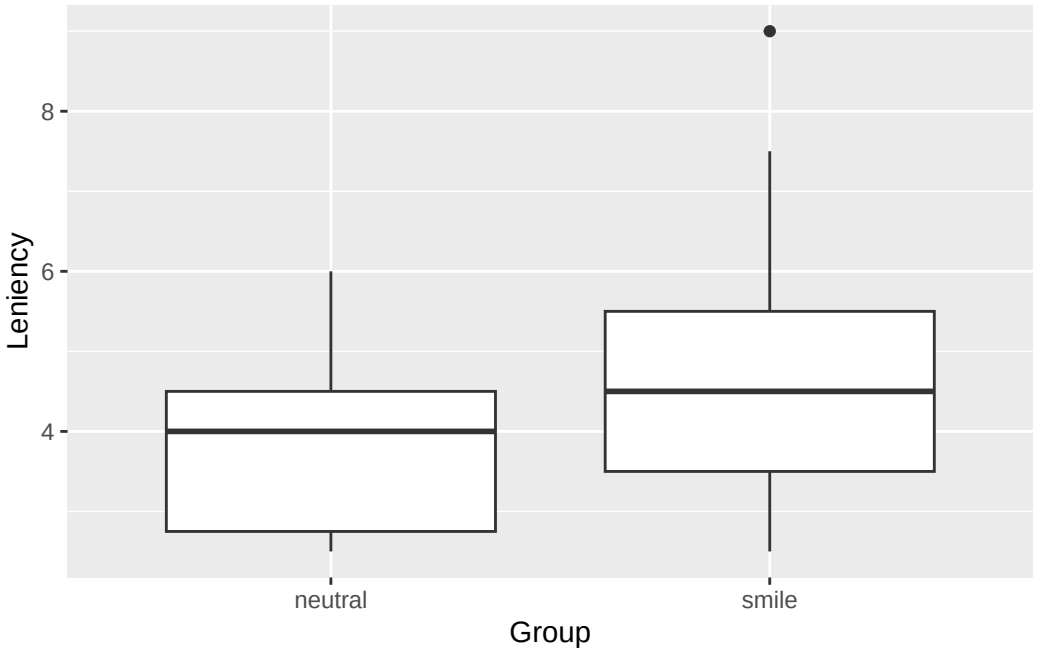

Figure 3: Smiles leniency plot

```
# A tibble: 30 x 3
   age_group fact_correct opinion_correct
   <chr>            <dbl>           <dbl>
 1 18-49                3               5
 2 18-49                5               5
 3 18-49                5               5
 4 50+                  4               1
 5 18-49                2               4
 6 50+                  5               5
 7 18-49                5               5
 8 50+                  4               2
 9 18-49                2               5
10 50+                  4               3
11 50+                  2               5
12 18-49                3               5
13 50+                  1               4
14 18-49                3               3
15 50+                  3               3
16 50+                  3               2
17 18-49                5               5
18 50+                  3               3
19 50+                  2               5
20 18-49                5               5
21 50+                  5               1
22 18-49                2               5
23 50+                  4               3
24 18-49                3               1
25 50+                  5               5
26 18-49                1               5
27 50+                  3               5
28 50+                  4               3
29 50+                  1               4
30 18-49                5               5
```

Figure 4: Fact and opinion survey data (30 randomly chosen rows)

```
fact_opinion %>% count(age_group) -> counted
ggplot(counted, aes(x = age_group)) + geom_bar()
```

Figure 5: Some code

```
ggplot(fact_opinion, aes(x = fact_correct, fill = age_group)) +
  geom_bar(position = "dodge")
```

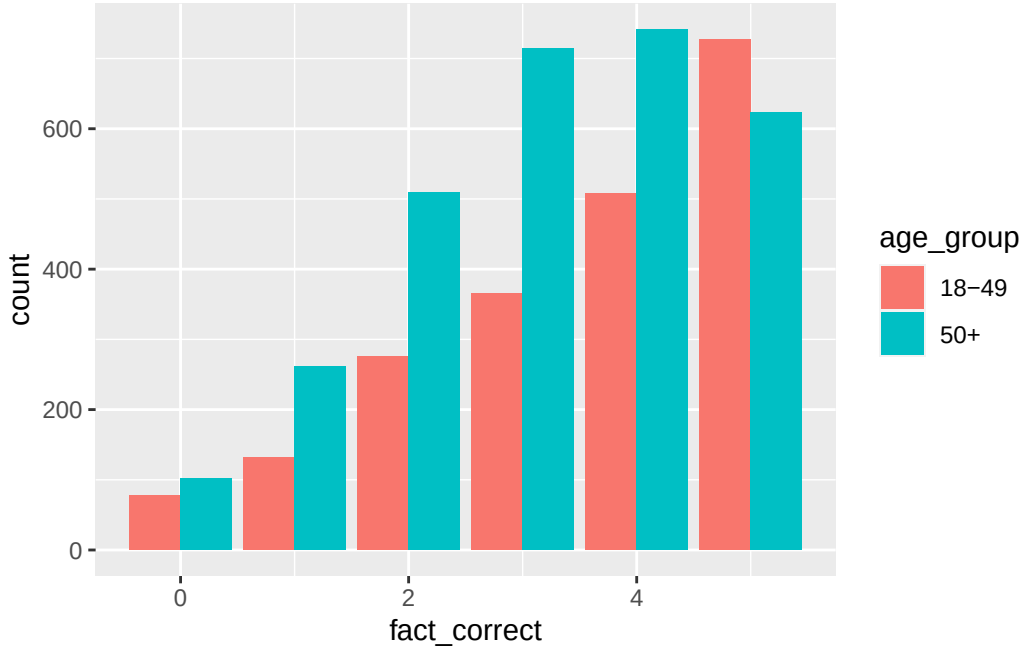

Figure 6: Fact and opinion survey plot

```
# A tibble: 30 x 6
   Region        Status Sex    Cause                     Rate    SE
   <chr>         <chr>  <chr>  <chr>                    <dbl> <dbl>
 1 HHS Region 08 Urban  Male   Unintentional injuries    55.3   0.7
 2 HHS Region 10 Urban  Male   Cancer                   191.    1.1
 3 HHS Region 10 Urban  Female Cerebrovascular diseases  35.2   0.4
 4 HHS Region 06 Urban  Male   Alzheimers                20.7   0.3
 5 HHS Region 10 Urban  Male   Unintentional injuries    49.8   0.6
 6 HHS Region 03 Rural  Female Cancer                   157.    1.4
 7 HHS Region 10 Rural  Male   Cerebrovascular diseases  37.1   1
 8 HHS Region 02 Rural  Male   Flu and pneumonia         19.6   0.9
 9 HHS Region 09 Rural  Male   Heart disease            206.    2.7
10 HHS Region 02 Urban  Female Flu and pneumonia         14.5   0.2
11 HHS Region 06 Urban  Male   Cancer                   202.    0.7
12 HHS Region 07 Rural  Male   Lower respiratory         65.9   0.9
13 HHS Region 01 Rural  Female Diabetes                  15     0.6
14 HHS Region 07 Urban  Female Diabetes                  16     0.3
15 HHS Region 09 Rural  Male   Unintentional injuries    79.1   1.8
16 HHS Region 04 Rural  Male   Unintentional injuries    79.1   0.7
17 HHS Region 01 Urban  Female Cancer                   140.    0.8
18 HHS Region 08 Urban  Female Nephritis                  8.3   0.3
19 HHS Region 07 Rural  Female Cancer                   150.    1.3
20 HHS Region 07 Rural  Male   Unintentional injuries    68.1   1
21 HHS Region 01 Rural  Male   Lower respiratory         51.7   1.3
22 HHS Region 09 Urban  Female Suicide                    5.2   0.1
23 HHS Region 05 Urban  Female Nephritis                 12.9   0.1
24 HHS Region 06 Rural  Male   Unintentional injuries    77.2   0.9
25 HHS Region 08 Rural  Male   Unintentional injuries    71     1.3
26 HHS Region 08 Urban  Female Alzheimers                30.9   0.5
27 HHS Region 10 Rural  Male   Alzheimers                22.9   0.8
28 HHS Region 06 Urban  Male   Heart disease            220.    0.8
29 HHS Region 05 Rural  Female Unintentional injuries    32.3   0.4
30 HHS Region 03 Rural  Male   Lower respiratory         62.1   1
```

Figure 7: US regional mortality rates data (randomly chosen rows)

```
  my_url <- "http://ritsokiguess.site/datafiles/shrimp.csv"
  shrimp <- read_csv(my_url)
```

```
Rows: 18 Columns: 1
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (1): percent

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
  shrimp
```

```
# A tibble: 18 x 1
   percent
     <dbl>
 1    32.2
 2    33
 3    30.8
 4    33.8
 5    32.2
 6    33.3
 7    31.7
 8    35.7
 9    32.4
10    31.2
11    26.6
12    30.7
13    32.5
14    30.7
15    31.2
16    30.3
17    32.3
18    31.7
```

Figure 8: Shrimp cocktail data

```r
with(shrimp, t.test(percent, mu = 34, alternative = "less"))
```

```
    One Sample t-test

data:  percent
t = -5.0761, df = 17, p-value = 4.674e-05
alternative hypothesis: true mean is less than 34
95 percent confidence interval:
    -Inf 32.5503
sample estimates:
mean of x
 31.79444
```

Figure 9: Code and output for an analysis on the shrimp data

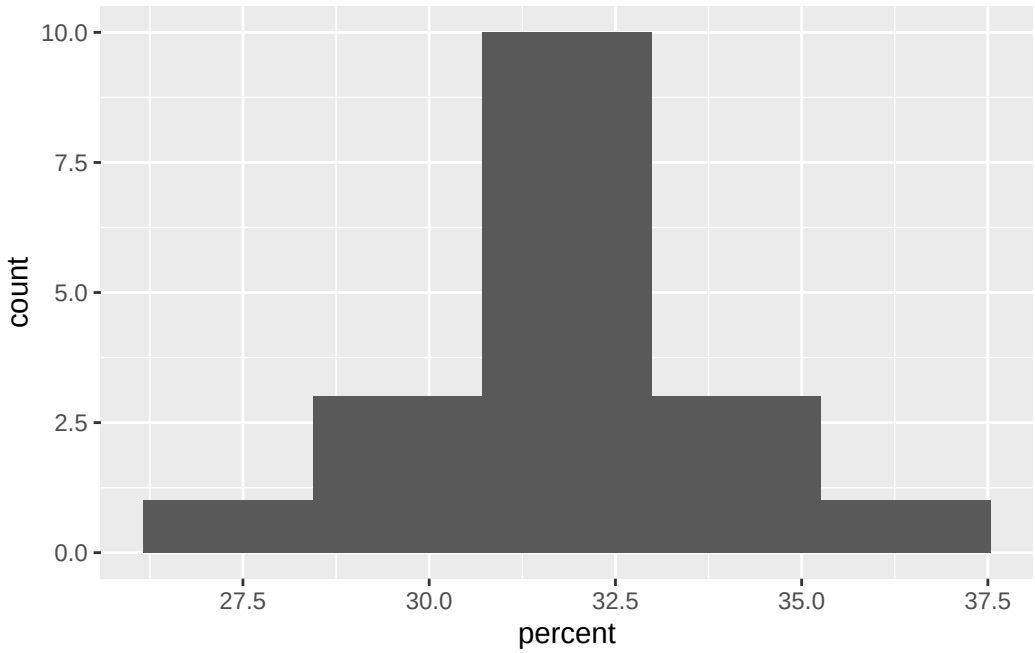

Figure 10: Histogram of shrimp data

```
# A tibble: 19 x 2
   company      emission
   <chr>           <dbl>
 1 manufacturer      2.7
 2 manufacturer      3.1
 3 manufacturer      3.1
 4 manufacturer      2.9
 5 manufacturer      2.5
 6 manufacturer      3.4
 7 manufacturer      3.4
 8 manufacturer      3.4
 9 manufacturer      2.4
10 competitor        3.7
11 competitor        3
12 competitor        3.5
13 competitor        3.8
14 competitor        2.8
15 competitor        3.5
16 competitor        3.4
17 competitor        3.6
18 competitor        2.7
19 competitor        3.7
```

Figure 11: Carbon monoxide data

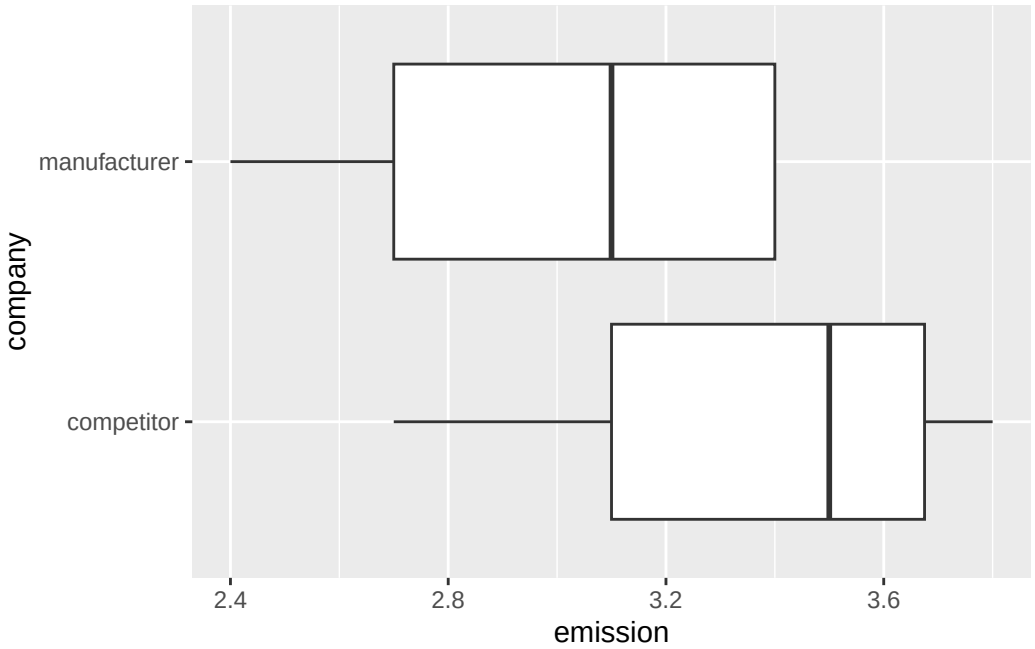

Figure 12: Plot for carbon monoxide data. Note that one of the whiskers for "manufacturer" is very short.

```
    Welch Two Sample t-test

data:  emission by company
t = 2.1187, df = 16.842, p-value = 0.02465
alternative hypothesis: true difference in means between group competitor and group manufacturer is grea
95 percent confidence interval:
 0.06802198        Inf
sample estimates:
  mean in group competitor mean in group manufacturer
                  3.370000                   2.988889
```

Figure 13: Test output for carbon monoxide data