Spring 2023 STAT 240 In-Person (Part 1) Midterm

1st Letter of Last/Family Name | Last/Family Name as in Canvas | First/Given Name as in Canvas | Student ID

Instructor (Circle) **Bret Larget** **Bi Cheng Wu** **Hamna Hannan**

Lecture Time (Circle) **MWF 8:50 - 9:40** **MWF 1:20 - 2:10** **MWF 2:25 - 3:15** **TH 8:00 - 9:15**

Discussion (Circle Name and Time)

TA	time 1	time 2	time 3
Shane Huang	T 7:45 am	T 8:50 am	T 9:55 am
Christian Varner	M 2:25 pm	M 3:30 pm	M 4:35 pm
Cameron Jones	M 2:25 pm	M 3:30 pm	M 4:35 pm
Ryan Yee	M 2:25 pm	M 3:30 pm	M 4:35 pm
Congwei Yang	T 7:45 am	T 12:05 pm	Tue 1:20 pm
Jingyang Lyu	T 7:45 am	T 12:05 pm	W 7:45 am
Nathaniel Pritchard	W 7:45 am	W 8:50 am	W 4:35 pm
Haoran Xiong	T 4:35 pm	W 7:45 am	W 4:35 pm

Instructions:

1. You may use one regular sheet of paper with self-prepared notes. You may use both sides of the paper.
2. You may not consult other resources, your phone, a computer, online info, nor your neighbor's exam.
3. Do all of your work in the space provided. Use the backs of pages if necessary, indicating clearly that you have done so (so the grader can easily find your complete answer).

Sections

- Name, Lecture, and Discussion (2 points)
- Multiple Choice (24 points, eight questions worth 3 points each).
- Short Answer (12 points, three questions worth 4 points each).
- Data Analysis (12 points, three questions worth 4 points each).

Scoring

Problem	Name	1 - 5	6 - 8	9 - 11	12	13	14	Total
Possible	2	15	9	12	4	4	4	50
Points								

Multiple Choice (24 points). Each problem is worth 3 points.

Circle the correct answer or answers as indicated.

Problem 1. Which type of graph is typically most effective at displaying the relationship between two quantitative variables. **Circle one answer.**

(a) bar chart (b) density plot (c) histogram (d) scatter plot

Problem 2. Which of the following commands calculate the proportion of missing values in a vector X? **Circle all correct answers.**

(a) `mean(is.na(X))`

(b) `mean(X == NA)`

(c) `sum(X == NA) / count(X)`

(d) `sum(is.na(X)) / length(X)`

Problem 3. Which are valid names for an R object? **Circle all correct answers.**

(a) `bucky_badger` (b) `1_bucky_badger` (c) `bucky_badger_1` (d) `bucky=badger`

Problem 4. Which command keeps all rows of the a data set `obesity` where the variable `n` is not missing, the variable `sex` is "female", and the variable `age` is not "05-17"? **Circle one answer.**

(a) `obesity %>% filter(n != NA, sex = "female", age != "05-17")`

(b) `obesity %>% filter(n != NA, sex == "female", age == !"adult")`

(c) `obesity %>% filter(is.na(n), sex == "female", age == !"05-17")`

(d) `obesity %>% filter(is.na(n), sex == "female", age != "05-17")`

Problem 5. The plot of a histogram of a variable `X` from a data set `df` appears to be symmetric, bell-shaped, and centered near 0. The value of `df %>% summarize(sd = sd(X))` is `15.0`. Which numerical value will be closest to the result of `df %>% summarize(p = mean( between(X, -30, 30)))`? **Circle one answer.**

(a) 0 (b) 0.68 (c) 0.95 (d) 99

Problem 6. What is the output of the following command, (suppressing variable names), where March 8, 2023 is a Wednesday and August 3, 2023 is a Thursday? **Circle one answer.**

```
tibble(date = mdy("3/8/2028")) %>%
  summarize(day = day(date),
            wday = wday(date, label = TRUE),
            yday = yday(date),
            month = month(date, label = TRUE))
```

(a) `8  Wed  67 Mar`
(b) `8  67  Wed Mar`
(c) `3  Thu  215 Aug`
(d) `Thu 3  215 Aug`

Problem 7. A data set `grocery_items` has variables named `item`, `type`, and `price`. A data set named `grocery_list` has variables named `item` and `n`. The values in the columns `item` match if the same item is part of both data sets. Some items in `grocery_items` may not by in `grocery_list` and some items in `grocery_list` may not by in `grocery_items`. Which description matches the contents of `df` after executing the following code? No items are repeated within either data set. **Circle one answer.**

```
df = grocery_list %>%
  full_join(grocery_items, by = "item")
```

(a) A data frame with one row for each item in both data sets and columns `item` and `n` only.

(b) A data frame with one row for each item in both data sets and columns `item`, `n`, `type` and `price`.

(c) A data frame with one row for each item in `grocery_list` and columns `item`, `n`, `type` and `price`.

(d) A data frame with one row for each item in either data set, columns `item`, `n`, `type` and `price`, and the value `NA` in columns `n`, `type`, and `price` in rows where this information was missing.

Problem 8

A data frame `sfo` has no missing data, $7 \times 24 = 168$ rows, and columns named `Day`, `Hour`, and `n` where `Day` is an abbreviated day of the week (one of `Sun`, ..., `Sat`), `Hour` is one of `0000-0100` to `2300-0000`, and `n` is an integer count. Each possible combination of `Day` and `Hour` is included in the data frame.

The data frame `df` created with the following code will have how many rows, how many columns, and what will the column names be? **Circle one answer.**

```
df = sfo %>%
  pivot_wider(names_from = Day, values_from = n)
```

(a) 168 rows and 3 columns named `Hour`, `Day`, and `n`

(b) 24 rows and 8 columns named `Hour`, `Sun`, ..., `Sat`

(c) 24 rows and 7 columns named `Sun`, ..., `Sat`

(d) 7 rows and 25 columns named `Day`, `0000-0100`, ..., `2300-0000`

Short Answer (12 points). Each problem is worth 4 points

Problems 9-11 are based on a small data set `df` below which has numerical variables `x` and `y` and a categorical variable named `color`.

```
##   x  y color
## 1 1 -1   red
## 2 2 -5  blue
## 3 3  0  blue
## 4 4  5   red
## 5 5  1  blue
## 6 6  5  blue
```

Problem 9. Write the result of the following code.

```
df %>%
  slice_max(y, n = 1)
```

Problem 10 Write the result of the following code.

```
df %>%
  mutate(s = x+y) %>%
  group_by(color) %>%
  summarize(max = max(s)) %>%
  arrange(desc(color))
```

Problem 11 Write the result of the following code.

```
df %>%
  select(-x) %>%
  filter(y > 0) %>%
  group_by(color) %>%
  mutate(n = n())
```

Data Analysis (12 points). Three problems worth 4 points each.

Each problem asks you to interpret the output from the following data analysis of the `obesity` and `education` data sets we studied in class. The obesity data set has undergone some transformation from its raw form.

The `obesity` data set includes one row for each `zip/sex/age` combination, a total of 7740 rows, and variables:

- `zip` is a zip code (a 5-digit a categorical variable);
- `sex` is either "female" or "male";
- `age` is an age range from "05-17", "18-34", "35-54", "55-74", and "75+";
- `obese` is the number of sampled individuals who are obese;
- `n` is the number of sampled individuals;
- `pop` is the population of individuals in the zip code/sex/age range combination;
- `obese_pop` is an estimate of the number of individuals in `pop` who are obese: `obese_pop` contains some missing data; and

The `education` data set has one row for each zip code and variables:

- `zip` as above;
- `pct_f_bach` which is the percentage of women aged 25 and older with a bachelors degree; and
- `pct_m_bach` which is the percentage of men aged 25 and older with a bachelors degree

Read the questions before reading the code. Only read parts of the code needed to answer each question.

```
summary_1 = obesity %>%
  drop_na(obese) %>%
  group_by(sex, age) %>%
  summarize(v = 100 * sum(n) / sum(pop)) %>%
  ungroup() %>%
  mutate(sex = recode(sex,
                      "female" = "A",
                      "male" = "B")) %>%
  pivot_wider(names_from = sex, values_from = v)

summary_1
```

Summary for Problem 12

```
## # A tibble: 5 x 3
##   age       A     B
##   <chr> <dbl> <dbl>
## 1 05-17  43.4  42.8
## 2 18-34  38.8  29.3
## 3 35-54  42.0  35.5
## 4 55-74  50.8  47.1
## 5 75+    56.3  59.6
```

Problem 12 What do the values in column A and column B represent for each age group in `summary_1`?

```
obesity_2 = obesity %>%
  drop_na() %>%
  filter(age != "05-17") %>%
  group_by(zip) %>%
  mutate(x = n()) %>%
  filter(x == 8) %>%
  arrange(zip, sex, age) %>%
  group_by(zip, sex) %>%
  summarize(pop = sum(pop),
            obese_pop = sum(obese_pop),
            z = 100*obese_pop / pop)

education = education %>%
  rename(female = pct_f_bach,
         male = pct_m_bach) %>%
  pivot_longer(female:male, names_to = "sex", values_to = "w")

education_2 = education %>%
  inner_join(obesity_2, by = c("zip", "sex")) %>%
  mutate(y = 100*obese_pop / pop,
         z = pop * w / 100) %>%
  relocate(y, .after = w)

education_2 %>%
  print(n = 4)
```

Summary for Problems 13 and 14

```
## # A tibble: 612 x 7
##   zip   sex        w     y   pop obese_pop     z
##   <chr> <chr>  <dbl> <dbl> <dbl>     <dbl> <dbl>
## 1 53002 female  25.4  37.7   977      369.  248.
## 2 53002 male    16.2  38.4  1052      404.  170.
## 3 53005 female  53.9  26.8  8039     2153. 4333.
## 4 53005 male    55.6  34.7  7150     2479. 3975.
## # ... with 608 more rows
```

```
summary_2 = education_2 %>%
  group_by(sex) %>%
  summarize(U = 100*sum(obese_pop) / sum(pop),
            V = 100*sum(z) / sum(pop))
```

```
summary_2
```

```
## # A tibble: 2 x 3
##   sex        U     V
##   <chr>  <dbl> <dbl>
## 1 female  39.8  31.8
## 2 male    41.3  30.0
```

Problem 13 What do the values in columns U and V represent for each sex in `summary_2`?

Problem 14. Add meaningful axis labels and a title to the following plot.

```
g = ggplot(education_2, aes(x = w/100, y = y/100)) +
  geom_point(alpha = 0.5) +
  geom_smooth(se = FALSE, color = "gray") +
  xlab("") +
  ylab("") +
  scale_x_continuous(labels = scales::percent) +
  scale_y_continuous(labels = scales::percent) +
  facet_grid(cols = vars(sex)) +
  theme_bw()
```