Spring 2021 Midterm I SOLUTIONS

1. [1 point] What functions are necessary to visualize the distribution of a continuous variable? Choose all that apply.

a) `geom_histogram()`
b) `ggplot()`
c) `geom_point()`
d) `geom_bar()`
e) `aes()`
f) `geom_smooth()`

SOLUTION: a) `geom_histogram()`, b) `ggplot()`, e) `aes()`

2. [3 points total] Below is text taken from the abstract of "A Randomized Trial Comparing Acupuncture, Simulated Acupuncture, and Usual Care for Chronic Low Back Pain" (Arch Intern Med. 2009 May 11; 169(9): 858–866. doi:10.1001/archinternmed.2009.65.)

> Background: Acupuncture is a popular complementary and alternative treatment for chronic back pain. Recent European trials suggest similar short-term benefits from real and sham acupuncture needling. This trial addresses the importance of needle placement and skin penetration in eliciting acupuncture effects for patients with chronic low back pain. Methods: 638 adults with chronic mechanical low back pain were randomized to: individualized acupuncture, standardized acupuncture, simulated acupuncture, or usual care. Ten treatments were provided over 7 weeks by experienced acupuncturists. The primary outcomes were back-related dysfunction (Roland Disability score, range: 0 to 23) and symptom bothersomeness (0 to 10 scale). Outcomes were assessed at baseline and after 8, 26 and 52 weeks.

i. [1 point] In our introductory lecture on the PPDAC method we discussed types of problems. What type of a problem is this study addressing?

a) Observational
b) Causative/Etiologic
c) Predictive
d) Descriptive

SOLUTION: b) Causative/Etiologic

ii. [1 point] What type of variable is the exposure variable in this study?

a) Nominal
b) Discrete
c) Continuous
d) Ordinal

SOLUTION: a) Nominal

iii. [1 point] What type of variable are the primary outcomes in this study?

a) Nominal
b) Discrete
c) Continuous
d) Ordinal

SOLUTION: c) Continuous

3. [1 point] The overall rate of asthma in city A is less than in city B. Therefore, the rate of asthma for each age group in city A must be less than the rate of asthma in the corresponding age group in city B. Is this statement true or false?

a) True
b) False

SOLUTION: b) False

4. [6 points total] You are given a dataset, `covid_data` which has 4 columns (`county`, `state`, `num_deaths`, and `population`).

i. [1 point] Select the line(s) of code you could run so that you only have data from California and Washington. Select all that apply.

a) `covid_data %>% select(state == "California", state == "Washington")`
b) `covid_data %>% filter(state %in% c("California", "Washington"))`
c) `covid_data %>% filter(state == "California" & state == "Washington")`
d) `covid_data %>% filter(state == "California" | state == "Washington")`

SOLUTION:

a) `covid_data %>% select(state == "California", state == "Washington")`

d) `covid_data %>% filter(state == "California" | state == "Washington")`

ii. [1 point] With `covid_data %>% arrange(county, -num_deaths)`, how will this line of code sort the data?

a) Sort `county` in descending order first, then `num_deaths` in ascending order
b) Sort `county` in ascending order first, then `num_deaths` in descending order
c) Sort `county` in ascending order first, then `num_deaths` in ascending order
d) Sort `num_deaths` in ascending order first, then `county` in descending order

SOLUTION: b) Sort `county` in ascending order first, then `num_deaths` in descending order

iii. [1 point] Say you want to use this data to run a linear regression to predict `num_deaths` based on `population`. Write one line of `R` code to perform this regression.

SOLUTION: `lm(num_deaths ~ population, data = covid_data)`

iv. [1 point] Say that you assign the results of your previous `R` command to the variable name `covid_model`. You then run `tidy(covid_model)` and see the following output. Report and interpret the slope of the regression line. Don't forget units! (Note: This is purely an example and does not necessarily reflect actual COVID data)

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic     p.value
##   <chr>          <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)    8.1       0.242    121.   0.0000142
## 2 population     0.035     0.135      7.40 0.000000246
```

SOLUTION: Slope is 0.035 deaths per one unit population. That is, as population increases by one person, the expected number of deaths increases by 0.035.

v. [1 point] Report the y-intercept of the regression line (with units!). Does this intercept make sense in context? Explain.

SOLUTION: y-intercept is 8.1 deaths. No it does not make sense, because you cannot have 8.1 deaths in a population of zero people.

vi. [1 point] Use the `tidy(covid_model)` output to predict the number of COVID deaths in a county with a population of 50,000 people. Show all of your calculations.

SOLUTION: 1758.1 deaths

5. [3 points total] The dataset `insure_data` includes data on medical costs. Below is a data dictionary with information on each of the variables.

Column	Description
`age`	age of primary beneficiary
`sex`	male, female
`bmi`	Body Mass Index (kg/mˆ22)
`children`	Number of children covered by health insurance/Number of dependents
`smoking`	smoker (yes/no)
`region`	the beneficiary's residential area in the US, northeast,southeast, southwest, northwest.
`charges`	individual medical costs billed by health insurance

```
head(insure_data)
```

```
## # A tibble: 6 x 7
##     age sex      bmi children smoker region    charges
##   <dbl> <chr>  <dbl>    <dbl> <chr>  <chr>       <dbl>
## 1    19 female  27.9        0 yes    southwest  16885.
## 2    18 male    33.8        1 no     southeast   1726.
## 3    28 male    33          3 no     southeast   4449.
## 4    33 male    22.7        0 no     northwest  21984.
## 5    32 male    28.9        0 no     northwest   3867.
## 6    31 female  25.7        0 no     southeast   3757.
```

i. [1 point] Which variables in `insure_data` are continuous? Select all that apply.

a) `age`
b) `sex`
c) `bmi`
d) `children`
e) `smoker`
f) `region`
g) `charges`

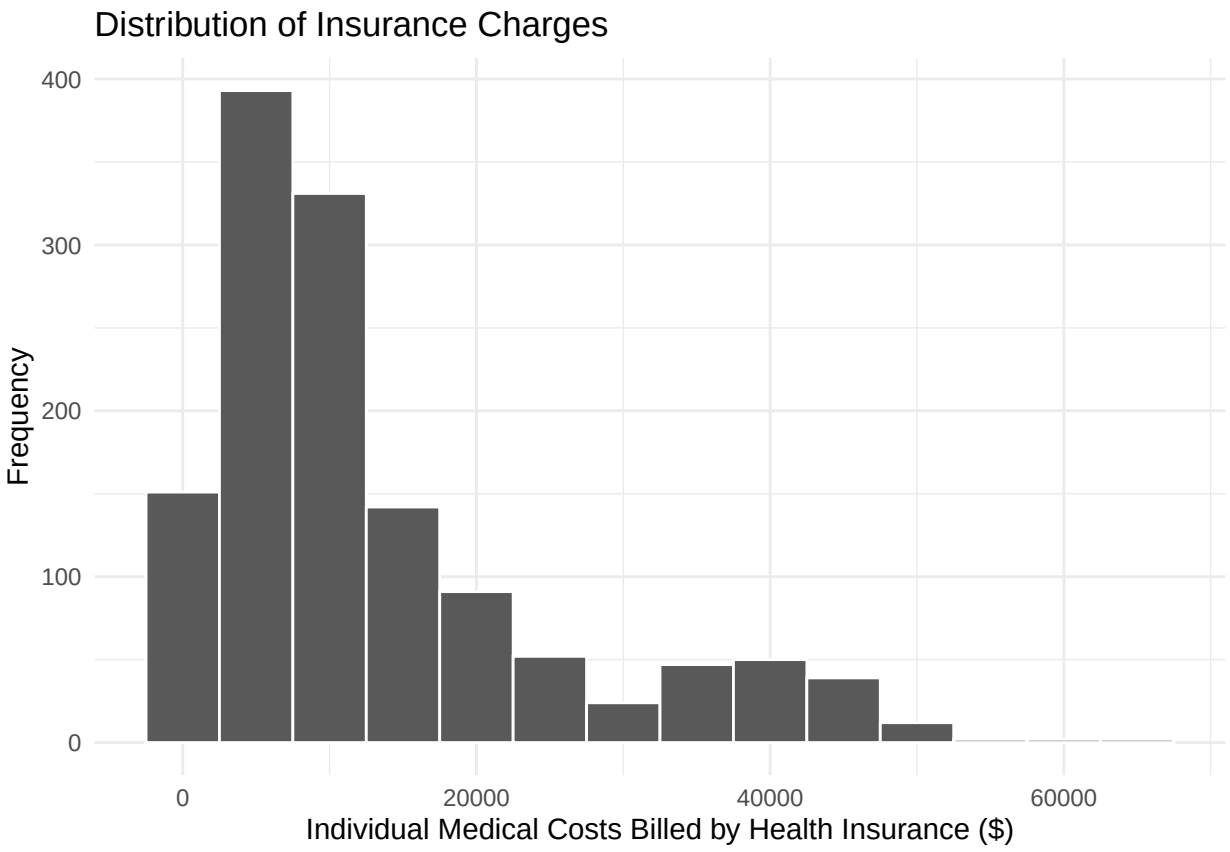
SOLUTION:

a) `age`

c) `bmi`

g) `charges`

Use this histogram to answer 5.2 through 5.4



ii. [1 point] Describe the distribution in no more than one sentence.

SOLUTION: Shape: bimodal, skewed to the right

iii. [0.5 point] Based on the histogram, what can you say about the mean and median of the distribution?

a) mean = median
b) mean > median
c) mean < median

SOLUTION: b) mean > median

iv. [0.5 point] Pick the sentence that is most correct.

a) The mean is approximately equal to $5000
b) The mean is smaller than $5000
c) The mean is larger than $5000
d) Not enough information to choose

SOLUTION: c) The mean is larger than $5000

6. [6 points total] You want to test two drugs, Drug 1 and Drug 2. You give each drug to a group of people and then count the number of successes (improvements) and failures (no change) for each group.

```
##   drug    sex successes count success_rate
## 1    1   Male        18    30           60
## 2    1 Female         2    10           20
## 3    2   Male         7    10           70
## 4    2 Female         9    30           30
```

i. [1 point] Using the data, fill in the blanks of the following two-way table.

	Success	Failure	Total
Drug 1	20	A	40
Drug 2	B	C	40
Total	36	D	80

SOLUTION:

A: 20

B: 16

C: 24

D: 44

ii. [1 point] What is the marginal distribution of drug success?

SOLUTION: The marginal distribution of drug success is 36/80 = 45% success, 55% failure.

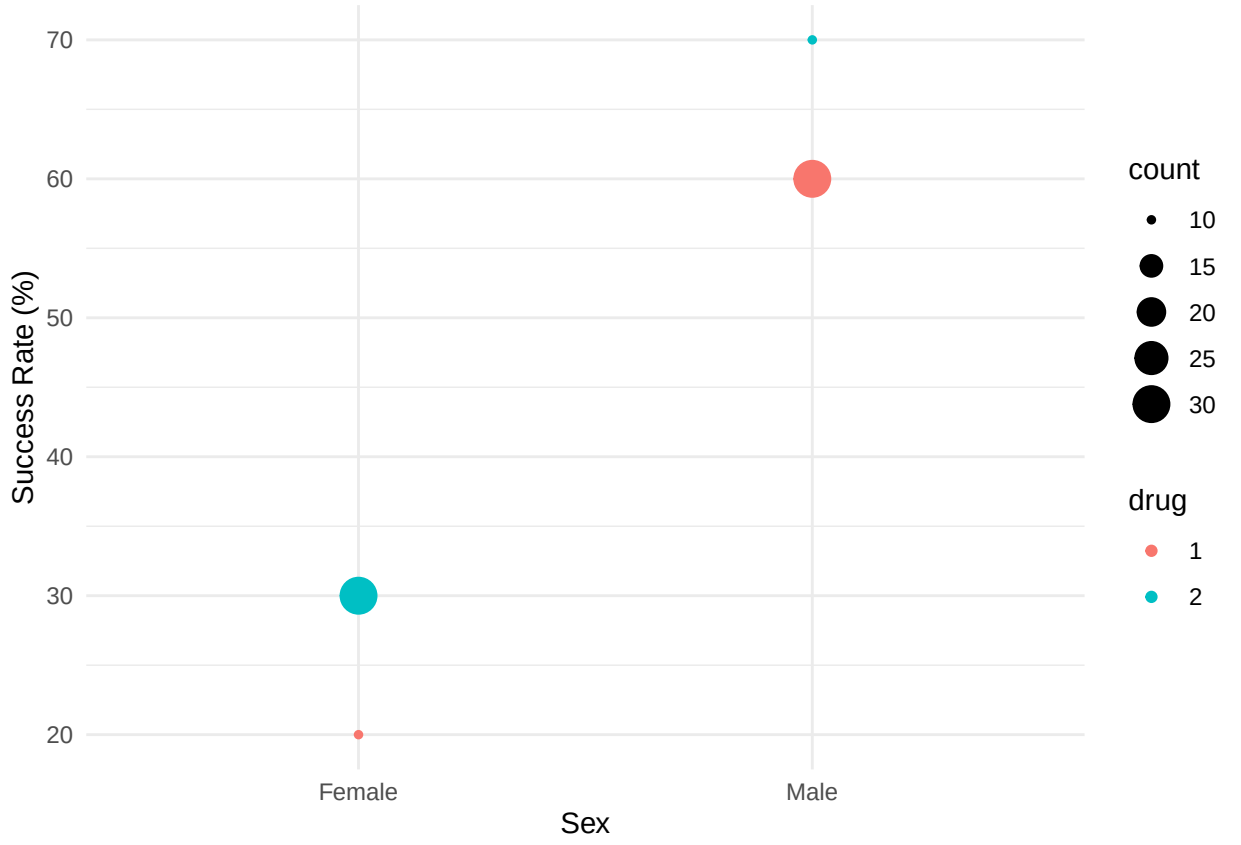
iii. [1 point] What is the conditional distribution of drug success among those who took Drug 1?

SOLUTION: The conditional distribution of drug success given Drug 1 is 20/40 = 50% success and 50% failure.

iv. [1 point] Which drug has the higher overall success rate?

a) Drug 1
b) Drug 2

SOLUTION: a) Drug 1



v. [2 points] From the visualization above, we can see that when we divide the data into groups by sex, there is a higher success rate for Drug 2. In 1-3 *brief* sentences and using your answer in 6.4, identify the cause of this phenomenon, and explain why that is the cause.

SOLUTION: Overall, Drug 1 has a higher success rate than Drug 2, but when when we stratify by sex, Drug 2 has a higher success rate. This is because sex is a confounding variable for the relationship between type of drug and drug success.

7. [1 point] Given a dataset with column names: `a`,`b`,`c`,`d`, which of the following commands is NOT equivalent to the others?

a) `select(a, b)`
b) `select(-c, -d)`
c) `select(c, d, -a, -b)`
d) `select(-c, -d, a, b)`

SOLUTION: c) `select(c, d, -a, -b)`

8. [1 point] In a study of food deserts and health outcomes in California, the age (in years), distance to a local supermarket (miles rounded to nearest two decimal points), annual income, county of residence, and presence of cardiovascular diseases of all participants were recorded. Which of these are continuous quantitative variables?

a) Age and distance to a local supermarket
b) Distance to a local supermarket only
c) Annual income and age
d) Annual income and distance to a local supermarket

SOLUTION: d) Annual income and distance to a local supermarket

9. [3 points total] You are given a dataset titled `lung_data` of patients who are enrolled in a clinical trial testing a drug to relieve symptoms of lung scarring, sarcoidosis, an illness 16 times more common among African Americans than among other races. The significance of this drug is that it is much cheaper and easier to distribute than other treatments. You are given a dataset with columns `patient_id`, `zip_code`, and `status`. `patient_id` is a randomly generated 4 digit number with no repeats, `zip_code` is the patient's zip code, and `status` is a T/F binary variable indicating whether or not they received the placebo.

i. [1 point] What type of variables are `patient_id` and `zip_code`, respectively?

a) Nominal, Nominal
b) Ordinal, Nominal
c) Discrete, Discrete
d) Discrete, Nominal

SOLUTION: a) Nominal, Nominal

ii. [1 point] Write one line of `R` code to get the number of `TRUE` values in the `status` column for each `zip_code` and assign it to the variable name `count_zip_status`.

SOLUTION:

```
count_zip_status <- lung_data %>% group_by(zip_code) %>% summarize(count = sum(status))
```

10. [1 point] Which of the following are measures of spread?

a) mean, median, mode
b) mean, median, and standard deviation
c) standard deviation and interquartile range
d) standard deviation, interquartile range, and correlation coefficient

SOLUTION: c) standard deviation and interquartile range

11. [1 point] True or False. The Pearson correlation coefficient ranges from 0 to 1.

SOLUTION: False. The correlation coefficient ranges from -1 to 1, since it describes the direction of the association between variables as well as the strength.

12. [2 points] Do outliers in a data set have more impact on the mean or median? Explain.

SOLUTION: The mean is not resistant to outliers. If a data set is larger the impact of the outliers will not be as drastic as if the data set was small. The median is typically not impacted by outliers.

13. [3 points] In 1998, GSK and the Walter Reed Institute, started cooperating to develop a Hepatitis E vaccine. Before launching phase II clinical trials, GSK had already decided the vaccine would not be commercially developed, while Walter Reed decided it would be unsuitable for US soldiers. Still, GSK and Walter Reed went ahead with phase II trials and ended up testing the candidate vaccine on 2000 Nepalese volunteers in Lalitpur, without a plan to further develop the vaccine and make it available to the local population if the trials were successful. (Source: https://www.somo.nl/wp-content/uploads/2008/02/Examples-of-unethical-trials.pdf) Explain why this study would or would not be considered ethical using at least 2 concepts of ethics from class.

SOLUTION: Not considered ethical because soldiers are easily coerced by superiors and a vulnerable population and patients were not provided informed consent on the decision to not proceed further with vaccine development.