# PH142 Midterm I

September 17, 2018

First and last name (print clearly): _____

Student number (print clearly): _____

Week day and time of lab section you attend: _____

| Question | points |
|---|---|
| 1 | 10 |
| 2 | 9 |
| 3 | 13 |
| 4 | 16 |
| Total | 48 |

Notes:

- You can use the back of each page as scratch paper.
- No points will be given for answers on the back of each page.
- Cellphones and computers must be stored and on silent.
- You must show your student ID when you submit your test.

**Question 1 (10 points total)**

Fill in the blank of the following statements:

1.a) [1 point] Correlation measures the _____ association between two variables. [linear]

1.b) [1 point] `geom_histogram()` is used to visualize _____ variables. [quantitative. 1/2 if said either continuous or discrete]

1.c) [1 point] The five number summary consists of the five following values: minimum, first quartile, _____, third quartile and the maximum. [median]

Circle the correct answer(s):

1.d) [1 point] True or false: Association equates to causation. [false]

1.e) [1 point] True or false: The `cor()` function requires two categorical variables to run. [false]

1.f) [1 point] Out of the four following options, circle those that are resistant measures: [median, quartiles]

- mean
- median
- standard deviation
- quartiles

1.g) [1 point] If a particular distribution is skewed left, the mean is most likely _____ the median. [less than]

- less than
- greater than
- equal to

1.h) [1 point] The correlation of a graph _____ affected by changing the unit of measurement of the x and/or y variable. [is not]

- is
- is not

Provide a short answer:

1.j) [1 point] You are tidying a data frame called `PH142_data` and need to add a new variable to use in a regression model in the following step. Compare the two pieces of code below. Both chunks will run, but one is preferred. Which one is preferred and why?

Chunk 1: `PH142_data <- PH142_data %>% mutate(...)`

vs.

Chunk 2: `PH142_data %>% mutate(...)`

Brief explanation:

Solution: Chunk 1 is preferred because it saved the change (i.e., adding the new variable) for later use whereas chunk 2 only prints it to the screen.

1.k) [1 point] Suppose you run the following code in your R console:

```
x <- 4
y <- 2
z == x+y
```

Does `z` show up in your environment? If so, what does `z` contain? If not, what could change in the code to add `z`?

Solution: No. Change `z == x + y` to `z <- x + y` (preferred) or `z = x + y` (acceptable, still full marks)

**Question 2 (9 points total)**

A data set that you have read into R is called `california_data`. This data set contains information on the number of cases of Pertussis, and the resulting hospitalizations and deaths among kindergartners for the forty largest counties in California.

The variables in this data set include:

| Variable | Description |
|---|---|
| COUNTY | String label for the county |
| Cases | Number of reported Pertussis cases |
| Hospitalizations | Number of cases leading to hospitalization |
| Deaths | Number of cases resulting in death |
| Case_Rate | Case rate per 1,000 members of the population |

The original dataset (`california_data`) and a tidied dataset (`tidy_ca_data`) are shown to you on two separate loose-leaf papers. They contain the same number of rows.

2.a) Circle the functions that were applied to go from the initial data frame (`california_data`) to the tidied data frame (`tidy_ca_data`):

- `rename()`
- `select()`
- `arrange()`
- `mutate()`
- `filter()`
- `group_by()`
- `summarize()`

[circle: `select()`, `arrange()`, and `mutate()`]

2.b) Write the code using the circled `dplyr` functions to apply functions to the initial data frame `california_data` and result in `tidy_ca_data`.

Solution:

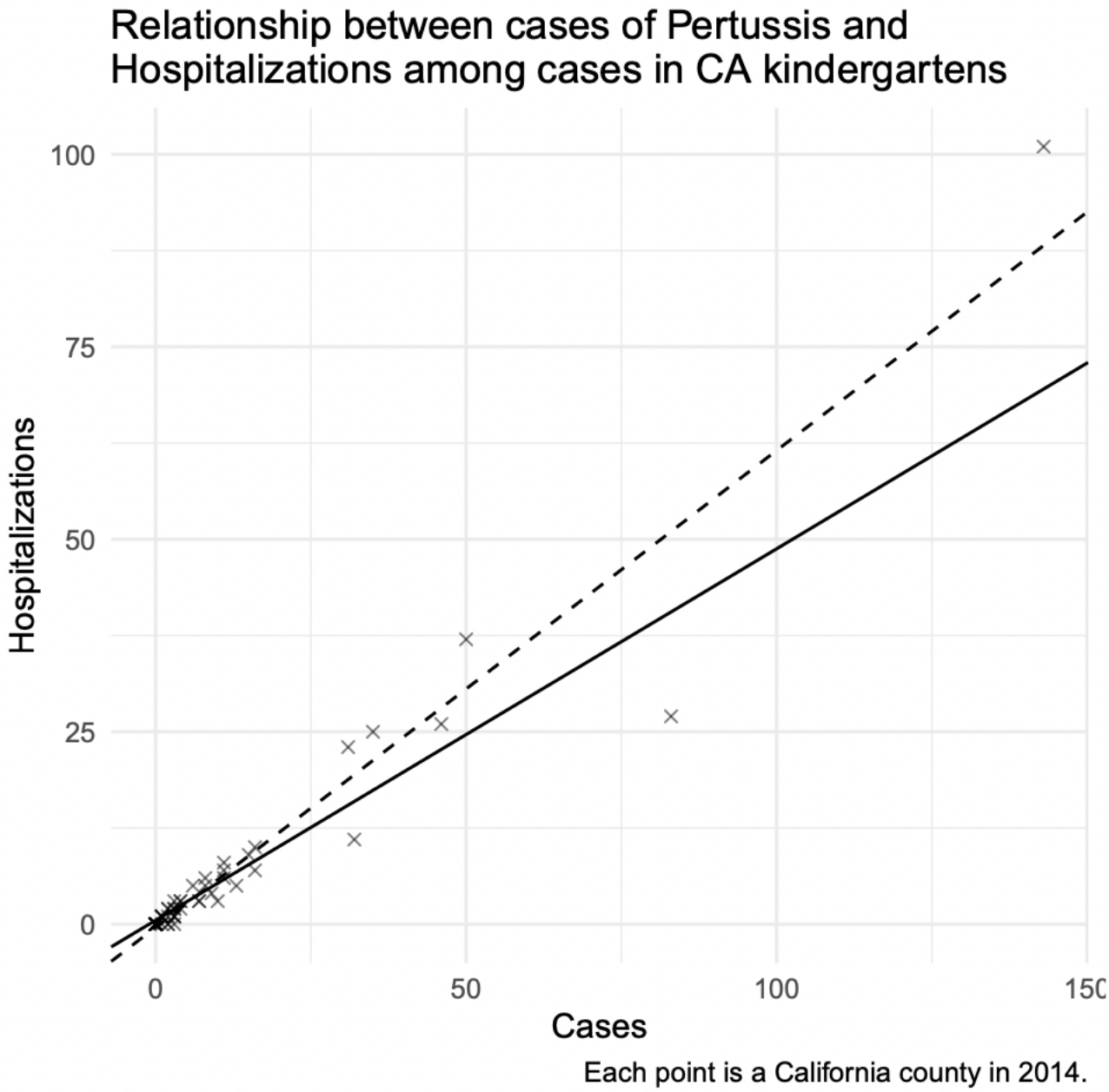
```
tidy_ca_data <- california_data %>%
  mutate(Hosp_Rate = Hospitalizations/Cases) %>%
  select(-Deaths) %>%
  arrange(-Case_Rate)
```

Students can split the above solution into multiple parts if they assign each intermediate step to a object that is then referenced in the following step.

Remove some marks for:

- Beginning a statement with the pipe (%>%) since that code would not compile
- Variable names not specified correctly

The graph below displays the number of Pertussis cases vs. Pertussis hospitalizations in 2014 for **all** California counties (not just the 40 included in `california_data`). The dashed line is the line of best fit including the point at x = 140, y = 101 and the solid line is the line of best fit excluding that point.




2.c) [1 point] What type of outlier is the point at x = 140, y = 101 considered to be? [circle c]

a. An outlier
b. An influential point
c. Both (a) and (b)
d. None of the above

2.d) [1 point] Circle the inequality that best describes the relationship between Hospitalizations and Cases: [circle a]

a. $r > 0.8$
b. $0.35 < r < 0.5$
c. $-0.32 < r < 0$
d. $-0.7 < r < -0.99$

2.e) [1 point] You execute the function `dim(all_california_data)` and get the following output:

```
[1] 58 5
```

Circle the correct interpretation of the output: [circle d]

a. The dataset has 58 unique observations and 5 repeated observations, for a total of 63 observations
b. The dataset has 3 rows and 58 columns
c. The dataset has 1 row with the following values: 58 and 5, in two separate columns
d. The dataset has 58 rows and 5 columns

**Question 3 (13 points total)**

Consider the following data set named HELP_study. The dataset contains information on adults admitted to hospital for treatment for alcohol, cocaine, or heroin. Selected variables in the dataset include:

| Variable | Description |
|---|---|
| age | age of patient |
| gender | gender of patient |
| racegrp | race/ethnicity of patient |
| homeless | housing status |
| drink_count | average number of drinks consumed per day, in the past 30 days |
| substance | primary substance of abuse |

Here is the code and output to help you understand the contents of this data frame:

```
dim(HELP_study)
```

```
## [1] 453   7
```

```
head(HELP_study)
```

```
##   age gender homeless drink_count racegrp substance hospitalizations
## 1  37   male   housed          13   black   cocaine                3
## 2  37   male homeless          56   white   alcohol               22
## 3  26   male   housed           0   black    heroin                0
## 4  39 female   housed           5   white    heroin                2
## 5  32   male homeless          10   black   cocaine               12
## 6  47 female   housed           4   black   cocaine                1
```

The following two-way table was created using the variables `homeless` and `substance` from `HELP_study`:

| Group | Alcohol | Cocaine | Heroin |
|---|---|---|---|
| Housed | 74 | 93 | 77 |
| Homeless | 103 | 59 | 47 |

3.a) [1 point] What type of variable is `homeless`? Circle one: [nominal]

- continuous
- discrete
- nominal
- ordinal

3.b) [1 point] What type of variable is `substance`? Circle one: [nominal]

- continuous
- discrete
- nominal
- ordinal

3.c) [2 points] Given the (little) information you know about this study, is there a clear explanatory and response variable between `homeless` and `substance`? If so, which variable is which? If not, why not?

Solution: No because it could be that becoming homeless affects substance abuse, or that substance use affects housing status and it is not clear which hypothesis the investigators want to test.

3.d) [3 points] Based on the two-way table, calculate the conditional distribution of substance abuse among homeless individuals. Round to two decimal places and show your work.

Solution:

Number of homeless individuals = 103 + 59 + 47 = 209

% alcohol use among homeless: 103/209 = 49.28%

% cocaine use among homeless: 59/209 = 28.23%

% herion among homeless: 47/209 = 22.49%

Using the data from the two-way table and your conditional distribution calculations, suppose you create a data frame called `two-way` with six rows that encodes the proportion of individuals using each type of substance among homeless and housed individuals. The data frame looks like this (but with the `percent`) column filled in:

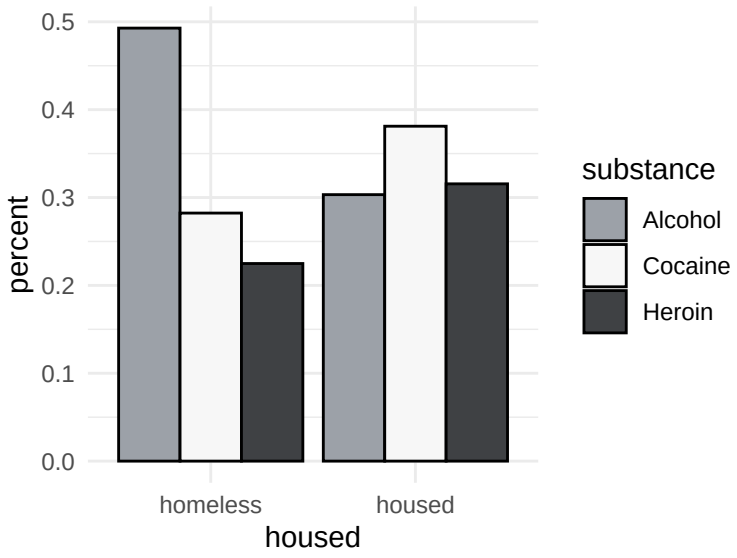
```
two_way <- tribble(    ~ housed,    ~ substance, ~ percent,
                       "housed",    "Alcohol",             ,
                       "housed",    "Cocaine",             ,
                       "housed",    "Heroin",              ,
                       "homeless", "Alcohol",              ,
                       "homeless", "Cocaine",              ,
                       "homeless", "Heroin",                 )
```

3.e) [6 points] Fill in the blanks to create the chart shown below using the `two-way` data frame. To help you, choose from the eight options listed below (two options will be left over).

```
ggplot(two_way, aes(x = _____________ , y = _____________)) +
  geom_bar(aes( _____________ = _____________ ), _____________,
           position = _____________, col = "black" ) +
  scale_fill_manual(values = c("#9ca1a8", "#f7f7f7", "#3f4144")) +
  theme_minimal()
```

Solution: Blanks in order: housed, percent, fill, substance, stat="identity", "stack"

- `stat = "identity"`
- `"dodge"`
- `percent`
- `col`
- `housed`
- `substance`
- `fill`
- `"stack"`

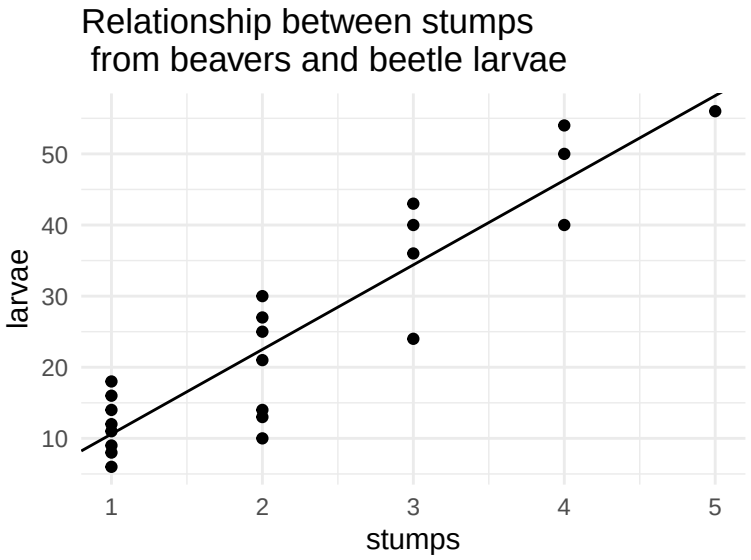


**Question 4 (16 points total)**

Ecologists sometimes find rather strange relationships in the environment. For example, do beavers benefit beetles? Researchers laid out 23 circular plots, each 4 meters in diameter, in an area where beavers were cutting down cottonwood trees. In each plot, they counted the number of stumps from trees cut by beavers and the number of clusters of beetle larvae. Ecologists think that the new sprouts from stumps are more tender than other cottonwood growth, so the beetles prefer them. If so, more stumps would produce more beetle larvae. Here are the first six lines of the data frame, and a scatter plot of the data along with a line of best fit:

```
head(beaver_data)
```

```
##   stumps larvae
## 1      2     10
## 2      2     30
## 3      1     12
## 4      3     24
## 5      3     36
## 6      4     40
```




4.a) [1 point] Is there a clear response and explanatory variable? If so, which one is which?

Solution: Yes. The x variable is stumps and the y variable is larvae.

4.b) [4 points] Suppose that the data frame has already been read into R. Write a few lines of code to run a linear model and produce the following output. (Hint: don't forget that `head()` is printed on the previous page and shows some useful information.):

Solution:

```
library(broom)
mod <- lm(larvae ~ stumps, data = beaver_data)
tidy(mod)
glance(mod)
```

Output 1:

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    -1.29      2.85    -0.451 6.57e- 1
## 2 stumps         11.9       1.14    10.5   8.67e-10
```

Output 2:

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.839         0.831  6.42      110. 8.67e-10     1  -74.4  155.  158.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

4.c) [1 point] Write the equation for the line of best fit for these data.

Solution: $larvae = -1.29 + 11.9 \times stumps$

4.d) [2 points] Interpret the slope parameter.

Solution: Increasing the number of stumps by 1 is associated with a 11.9 unit increase in the number of clusters of beetle larvae.

4.e) [2 points] Use the model to predict y when x is equal to 2.17. Round to one decimal place. Is it sensible to make a prediction at this x value? Why or why not?

Solution: larvae = -1.29 + 11.9*2.17 = -1.29 + 25.823 = 24.5

No, it isn't sensible because stumps is a discrete variable; for a discrete variable it only makes sense to make predictions at the values the variable can take, like at 2 stumps or 3 stumps.

4.f) [1 point] What is the correlation coefficient for the relationship between `stumps` and `larvae`? Round to two decimal places.

Solution: $r = \sqrt{0.839} = 0.92$ (from the regression output)

4.g) [5 points] Here is the sorted data for the 23 larvae measures. Calculate the median and the IQR. Would the corresponding box plot contain any outliers? Show calculations to justify why or why not.

```
##  [1]  6  8  9 10 11 12 13 14 14 16 18 21 24 25 27 30 36 40 40 43 50 54 56
```

Solution: Median: There are 23 measures, so the median is at the 12th ordered measure and is equal to 21.

1st quartile: There are 11 measures before the median so the first quartile is at the 6th ordered measure which is equal to 12.

3rd quartile: There are 11 measures after the median so the third quartile is at the 6th ordered measure *after* the median and is equal to 40.

The IQR is Q3-Q1 = 40-12 = 28

To check for outliers, we check if there are any measures outside of $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$:

$Q1 - 1.5 \times IQR$ = 12 - 1.5*28 = -30 There are no measures below -30.

$Q3 + 1.5 \times IQR$ = 40 + 1.5*28 = 82 There are no measures larger than 82.

Thus, there are no outliers on the boxplot.