EDS 222: Midterm Exam Answer Key

Fall 2023

Professor: Tamma Carleton

- You have 1 hour and 15 minutes to complete the exam.
- Clearly state your answers to the questions.
- Be concise!
- If you are unsure about a question, please raise your hand and ask for clarification. If necessary, clearly state any assumptions you are making.
- Please answer all questions and if you draw a figure, make sure to label your axes.

Short answer questions (10 points each)

1. Birds aren't real

For his final project, Charlie is using bird diversity data collected from a citizen science project in which people upload sightings of individual species with associated latitude and longitude coordinates. Charlie is using these data to construct an estimate of bird diversity (i.e., total number of unique bird species) in census tracts deemed to be "Disadvantaged Communities" (DAC) by the state of California under State Bill 32 (SB-32).

a. What is the population of interest, and what is the sample?

Population: All birds in DAC communities. (2)

Sample: Sightings of individual species by people uploaded to the citizen science project. (2)

b. Charlie computes the total count of unique species in the dataset that are geolocated in DAC census tracts in California. Is this statistic likely to be an unbiased estimate of the population parameter? Why or why not?

It is likely to be biased (1) because untrained people are probably unable to spot, identify, and report all the bird species within a tract. People might also not frequent all parts of the tract, especially places where birds reside, causing the true number of bird species to be underreported. (2)

c. Charlie repeats this exercise to compute total unique species geolocated in all non-DAC communities, which tend to be wealthier and contain more birding enthusiasts. He estimates a much larger number of unique species in non-DAC communities. Should Charlie conclude that non-DAC communities have higher bird diversity than DAC communities? Briefly explain why or why not.

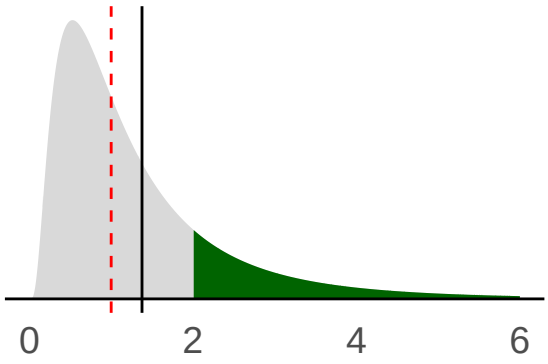
No, he should not (1) because people in NDAC tracts are wealthier and better educated; and hence more of them might be birding enthusiasts. Therefore, unique species of birds in NDAC tracts are more likely to be sighted, identified, and reported to the portal. Bird species in DAC tracts will be relatively under-reported compared to NDAC tracts. Charlie's conclusion might stem from bias in the sample rather than true differences in bird diversity. (2)

2. Rainy days

Consider the probability distribution of annual rainfall in meters for a village in the Amazon rainforest.

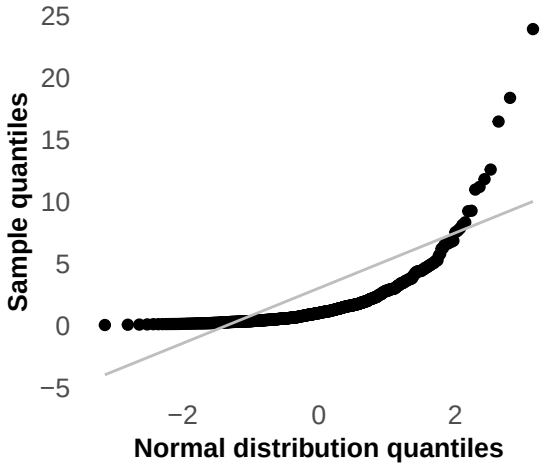
a. This probability distribution is plotted below. Mean annual rainfall is shown as a vertical black line. Is median annual rainfall above or below the mean? Show on this graph the probability that annual rainfall is greater than 2 meters.

Median is below the mean (the red line). (2) The probability that annual rainfall is greater than 2 meters is shaded in green below. (2)



b. Draw on the blank quantile-quantile (Q-Q) plot below roughly what you expect the Q-Q plot for this variable to look like.

(2) for each 1/3 of the distribution drawn.



3. Microplastics

The vector `micro` contains 10 elements, each recording the number of microplastic particles detected in fresh seafood sold at Whole Foods.

```
micro = c(6, 8, 12, 0, 6, 0, 0, 88, 2, 22)
```

a. What is the median number of microplastic particles in this sample?

[1] 0 0 0 2 6 6 8 12 22 88

median (between 5th and 6th positions) $= (6+6)/2 = 6$ **(2)**

b. A "quintile" is the word used to describe the the 5-quantile. What is the first quintile of this sample distribution? What is the fourth quintile?

1st quintile (between 2nd and 3rd positions) $= (0+0)/2 = 0$ **(2)**

4th quintile (between 7th and 8th positions) $= (12+22)/2 = 17$ **(2)**

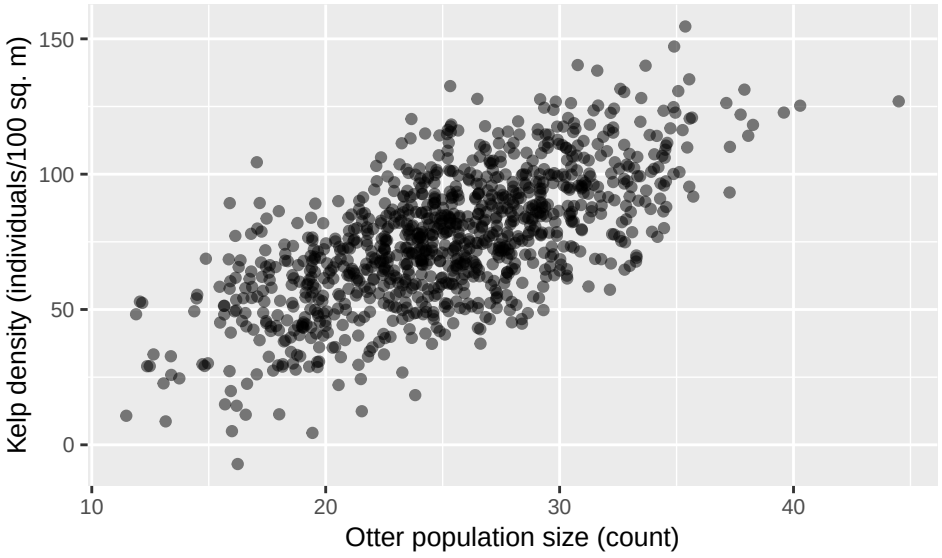
c. It is discovered that the value `88` is erroneous, arising from faulty microplastics detection technology. Which of the following summary statistics are "robust" (i.e., do not change) when this value is corrected to its true value of `17`? Circle all that apply.

- **median(2)**
- **mean (1)**
- **variance (1)**

Long answer questions (35 points each)

1. Keystone species

Sea otters play a critical ecological role as a "keystone species" of the kelp ecosystem. Otters protect kelp forests by eating their top predator, the sea urchin: a healthy otter population keeps urchins in check, enabling kelp to thrive. Oksana is interested in quantifying the relationship between otter populations and kelp forest density across the California coast. She has obtained data from the non-profit group Reef Check containing measurements of the kelp density (individuals per 100 m^2) and otter population counts (number of individuals) for hundreds of monitoring sites along the coast. Her data are plotted below (note units on the axes labels).



a. Below is a chunk of Oksana's code showing a simple linear regression relating kelp density to otter population size. What kelp density does Oksana predict will be present in a monitoring site with an otter population of 25 (feel free to round up to the nearest integer)?

```r
summary(lm(kelp ~ otters))
```

```
##
## Call:
## lm(formula = kelp ~ otters)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.862 -12.047   0.683  12.400  56.135
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.4499     2.9216  -1.181    0.238
## otters        3.1538     0.1136  27.756   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.17 on 998 degrees of freedom
## Multiple R-squared:  0.4357, Adjusted R-squared:  0.4351
## F-statistic: 770.4 on 1 and 998 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\hat{y} &= -3.45 + 3.15(25) \\ &= 75.3 \approx 75\end{aligned} \tag{5}$$

b. Oksana learns through her research on kelp forests that oil spills damage kelp forests by impacting kelp spores. She is worried about omitted variables bias. What second condition must to be true for oil spills to pose an omitted variables bias threat to Oksana's regression?

Oil spills must also affect otter population counts. (10)

c. Suppose Oksana could collect data on quantity of oil spills near every monitoring site in her data. If she adds oil spills as a second independent variable to her regression, do you expect her estimated slope coefficient on `otters` to increase or decrease relative to the simple linear regression she ran above? Briefly explain.

Her estimated slope should decrease. (5)

Oil spill is likely to negatively affect both otter counts and kelp density (3), exaggerating/biasing upward the association between them. In the model not accounting for oil spill, higher otter counts were capturing some of the effects (positive on kelp) of less oil spills (2), therefore, accounting for oil spill will yield a lower estimate for otter.

d. When Oksana runs the regression `lm(kelp ~ otters + oil_spills)`, she recovers an R^2 value of 0.52, and an adjusted-R^2 value of 0.39. The data provider Reef Check asks Oksana to provide them with a model that *best predicts* kelp density for the entire population. Should Oksana show Reef Check the model that includes `oil_spills` as an independent variable, or not? Briefly explain.

Oksana should be using adjusted R^2 as a metric since there are multiple variables. (5) Since $0.39 < 0.435$, she should use the model without oil spills. (5)

2. Valley Fever on the rise

Valley fever is an infectious disease caused by a fungus that lives in the soils in Central California and parts of Arizona. When dust from soil containing this fungus is inhaled, valley fever can infect lungs and cause moderate to severe respiratory conditions. Sofia is interested in studying the relationship between valley fever, agricultural workers, and drought in California. She has census tract level data from 2021 indicating valley fever cases, agricultural employment, and drought for all census tracts in California.

a. First, Sofia estimates the following multiple linear regression, where $valley\ fever_i$ is the valley fever case rate (cases per 100,000 people) in census tract i, $ag\ percent_i$ is the percent of tract i's population that is employed in agriculture, and $drought_i$ is a binary variable indicating whether the tract was in drought conditions or not when data were collected. When Sofia runs this regression, she recovers estimated coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. Interpret these coefficients in words.

$$\text{valley fever}_i = \beta_0 + \beta_1 \text{ag percent}_i + \beta_2 \text{drought}_i + \varepsilon_i$$

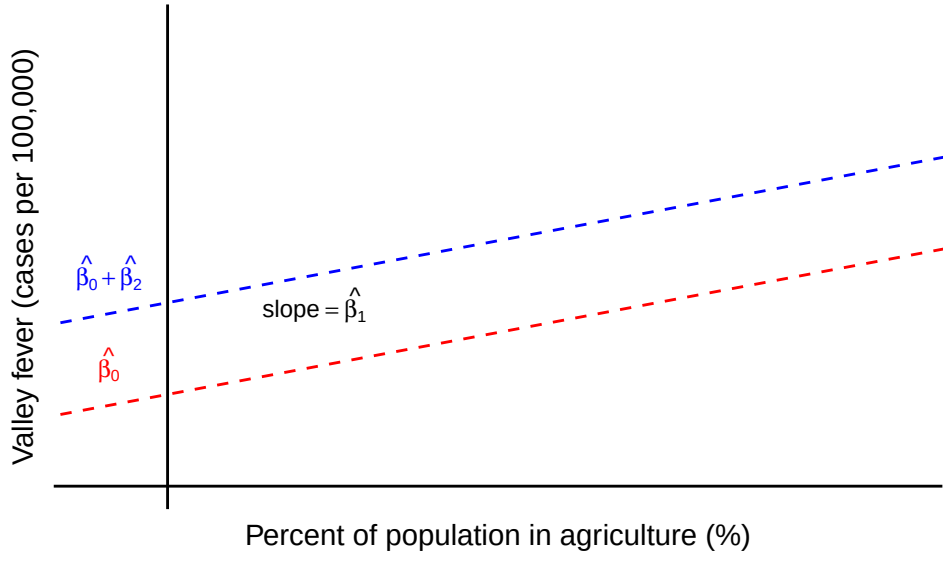
$\widehat{\beta}_0$: Average valley fever rates for tracts with no agricultural employees and in a non-drought conditions. (5)

$\widehat{\beta}_1$: Average increase in valley fever rates when there is a 1 unit / 1 percentage point increase in percent of population employed in agriculture (5), holding drought constant.

$\widehat{\beta}_2$: Average increase in valley fever rates between for tracts in drought compared to non-drought conditions (5), holding percent of population employed in agriculture constant.

- **Bonus point (+1) for stating "holding other variables constant"!**

b. Suppose Sofia finds that all three estimated coefficients are positive. Draw (approximately) on the following figure the predicted relationship between *valley fever* and *ag percent* for tracts that were in drought in 2021 and for those that were not. Label $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$.



c. Sofia reads some anecdotes about valley fever that suggest the effects of drought on cases likely differs across agricultural and non-agricultural populations, since drought increases dust exposure particularly for people working on farms. To investigate this, she estimates the following interaction model. Interpret $\hat{\beta}_3$ in words.

$$\text{valley fever}_i = \beta_0 + \beta_1 \text{ag percent}_i + \beta_2 \text{drought}_i + \beta_3 \text{ag percent}_i \times \text{drought}_i + \varepsilon_i$$

$\hat{\beta}_3$ is the average difference in the association between percent employed in agriculture and valley fever case rates for a tract in drought conditions, compared to a tract in non drought conditions. (5)

d. Suppose Sofia estimates that $\hat{\beta}_3 > 0$. Draw (approximately) the predicted relationship between *valley fever* and *ag percent* for both drought and non-drought census tracts under this new model. Label $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_1 + \hat{\beta}_3$.