EDS 222: Midterm Exam

October 27, 2022

Professor: Tamma Carleton

- You have until 10:45am 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. Vacation mode

For his final project, Lewis needs an estimate of indoor air pollution in homes throughout Mexico City. While visiting the hip La Roma neighborhood of the city on vacation, he surveys households by knocking on front doors throughout La Roma. For each household where someone answers the door and allows him to enter, Lewis uses a handheld air monitor to record particulate matter inside. He gets tired after 3 hours and heads to a cafe.

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

Answer: Population: all indoor spaces across all buildings in Mexico City. Sample: indoor spaces in the La Roma neighborhood where residents allowed Lewis and friends to enter the building.

b. Lewis computes the mean indoor air pollution in his sample. Is this statistic likely to be an unbiased estimate of the population parameter? Why or why not?

Answer: No, this sample mean is likely to be a very biased estimate of the population mean. This is because Lewis and friends will be obtaining a convenience sample. The La Roma neighborhood does not represent the whole city, and people who allow Lewis and friends into their homes may be systematically different from people who do not let them in.

c. Lewis recalls his lecture notes from EDS 222 and thinks maybe a stratified sampling approach would help him. If he were to redo his survey by stratifying all households across the city, would he want the households within each stratum to be (choose one): (i) very similar in their indoor air pollution levels; or (ii) very different in their indoor air pollution levels.

Answer: (i). Stratification helps obtain unbiased estimates when only a small sample can be collected, but only when observations within each stratum look very similar. The idea is that stratification ensures you sample from the full distribution (i.e., low and high indoor air pollution levels) and then you only need a handful of observations from each stratum to get an accurate estimate with a small overall sample size.

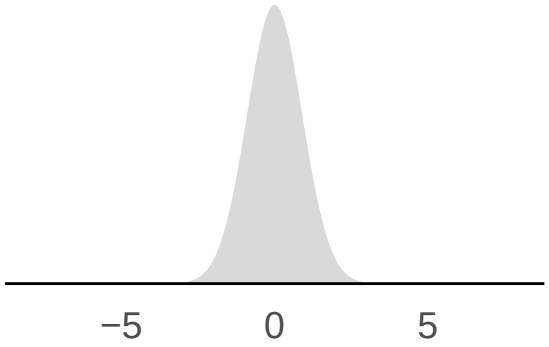
2. Feeling normal?

Consider a normal distribution with mean of $\mu = 0$ and standard deviation of $\sigma = 0.9$. The probability that a random variable distributed following this distribution takes on a value > 2 is 1.3%.

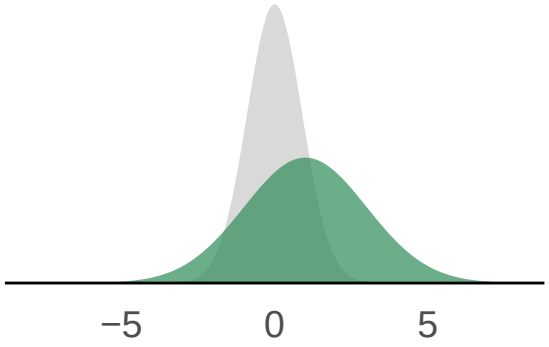
a. What is this distribution's variance?

Answer: variance is the square of the standard deviation. Therefore, the variance is 0.9^2=0.81.

b. The following graph displays this probability density function. Draw on top of the graph a new distribution with (approximately) $\mu = 1$ and $\sigma = 2$.



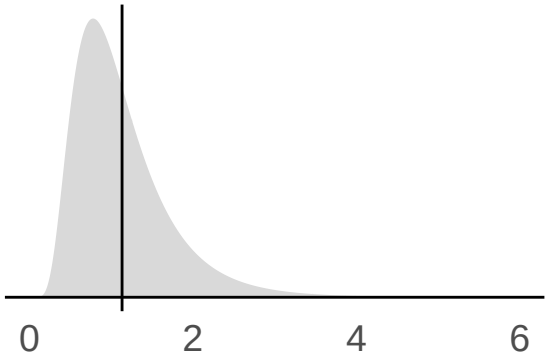
Answer:



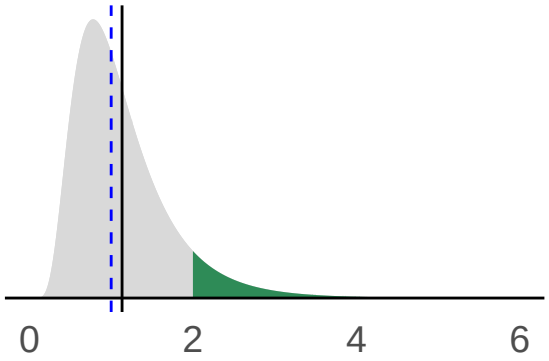
c. Is the probability that a random variable takes on a value > 2 under the new distribution (with $\mu = 1$ and $\sigma = 2$) smaller or larger than 1.3%?

Answer: Higher, both because of mean shift and increase in the variance.

d. Plotted below is the distribution of annual rainfall in meters for a village in the Amazon rainforest. 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.



Answer: The median will be below the black line, as the mean is pulled to the right by the long right tail. The median is shown in the blue dashed line and the shading indicates the probability that annual rainfall exceeds 2:



3. No more trash!

The vector `trash` contains 12 elements, each recording the number of trash items Colleen collected on 12 different beach cleanup days she spent at Campus Point beach.

```
trash = c(5, 17, 0, 12, 14, 14, 2, 15, 7, 22, 37, 6)
```

a. What is the median number of trash items in Colleen's sample?

Answer: Ordering the observations from smallest to largest, we obtain the ordered vector:

```
trash_ordered = sort(trash)
trash_ordered
```

```
##  [1]  0  2  5  6  7 12 14 14 15 17 22 37
```

The median is the mean of the 6th and 7th observations, which is (12+14)/2 = 13.

b. A "quartile" is the word used to describe the the 4-quantile. What is the first quartile of Colleen's sample distribution? What is the third quartile?

Answer: We divide the vector into 4 equally sized groups. Each group has three observations. The first quartile is the edge of the first group, which is (5+6)/2=5.5. The third quartile is the edge of the third group, which is between 15 and 17, so is (15+17)/2 = 16. I will accept 5, 5.5, and 6 as correct answers for the first quartile, and 15, 16, and 17 as correct answers for the third quartile, as these quartiles are not uniquely defined in this small sample.

c. Colleen made a data entry error and one of her entries of "14" should actually have been a "44". Will the first quartile be affected by this error? Will the third quartile be affected by this error?

Answer: The first quartile is unaffected by this data entry error. The third quartile will change from 16 to (17+22)/2 = 19.5.

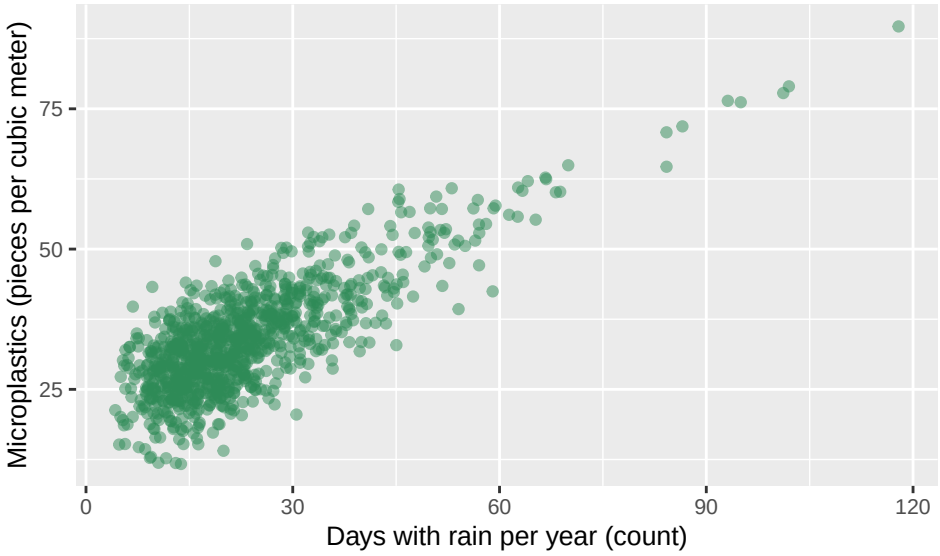
Long answer questions (35 points each)

1. Rain, rain, go away

Microplastics are tiny plastic particles that cause environmental degradation. In particular, ocean microplastics are consumed by marine animals, causing injury and/or death. Many factors influence the quantity of microplastics we see in the ocean, and Javier is interested in investigating whether storm activity in the

California Central Coast contributes to microplastics here. His hypothesis is that the rainfall transports land-based plastic pollution into marine environments through runoff, where it breaks down into microplastics.

Javier's data contain measurements of the pieces of microplastic per cubic meter of water from a variety of NOAA monitoring stations off the Central Coast, which he has merged to rainfall data from nearby rain gauges. His data are plotted below (note units on the axes labels).



a. Below is a chunk of Javier's code showing a simple linear regression relating microplastics to the number of days per year with rainfall. How many pieces of microplastic per cubic meter does Javier predict will be present in a location with 45 days of rain per year (feel free to round up to the nearest integer)?

```
summary(lm(micro ~ rainydays))
```

```
##
## Call:
## lm(formula = micro ~ rainydays)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.9484  -3.7799  -0.0535   3.8864  17.3402
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.24433    0.37803   53.55   <2e-16 ***
## rainydays    0.58942    0.01421   41.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.897 on 998 degrees of freedom
## Multiple R-squared:  0.633,  Adjusted R-squared:  0.6326
## F-statistic:  1721 on 1 and 998 DF,  p-value: < 2.2e-16
```

Answer: $20.2 + 0.59 \times 45$ equals approximately 46.75 pieces per cubic meter.

b. Javier presents a report showing the regression analysis above, and a NOAA official says they are skeptical of the estimated relationship because locations in the Central Coast with more rainy days also tend to be more populated, and plastic littering is a bigger problem in populated areas. Which assumption of Ordinary Least Squares is this critique challenging?

Answer: Exogeneity. The assumption that there is nothing in the error term that is correlated with the independent variable.

c. Which property of OLS would be threatened by this critique?

Answer: Unbiasedness. OLS suffers from omitted variable bias if this critique is true.

d. Suppose Javier could collected data on the population density for each catchment area around each microplastics monitoring station. If Javier adds population density as a second independent variable to his regression, do you expect his estimated slope coefficient on `rainydays` to increase or decrease relative to the simple linear regression he ran above? Briefly explain your reasoning.

Answer: We expect the slope coefficient on `rainydays` to fall in this new regression, because the positive correlation between rainy days and population will inflate the slope coefficient in the regression where population is omitted.

2. Polluting planes

Ruth is interested in the effect of noise pollution from airport traffic on mental health. For a sample of residents in the greater Los Angeles area, she has data on self-reported mental health information, socioeconomic status (SES), and distance between residence and the closest major airport.

a. First, Ruth estimates the following simple linear regression, where *days sad*$_i$ is the number of days in the last month that resident i felt depressed, sad, or anxious. *Low SES*$_i$ is a binary variable indicating whether the resident is of low socioeconomic status or not. When Ruth runs this regression, she recovers estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. Interpret these two coefficients in words.

$$\text{days sad}_i = \beta_0 + \beta_1 \text{Low SES}_i + \varepsilon_i$$

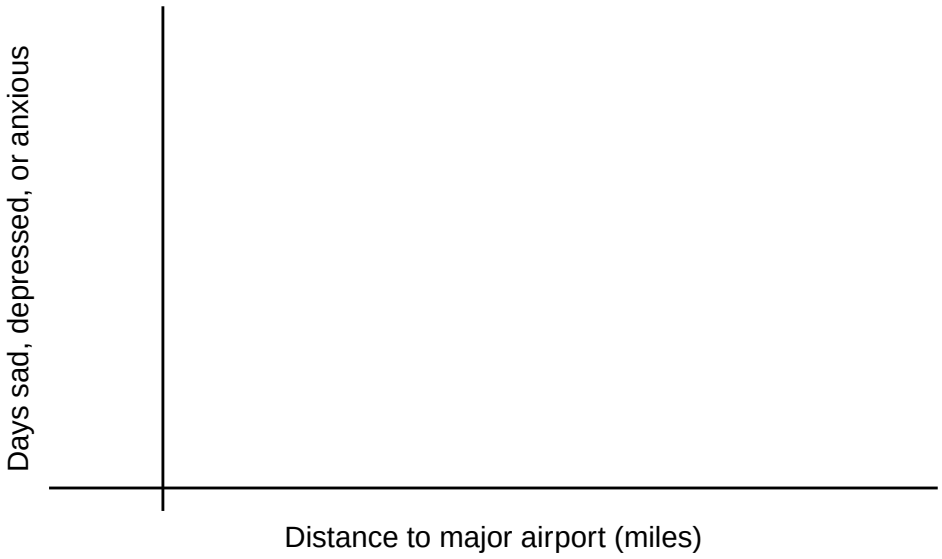
Answer: $\hat{\beta}_0$ tells us the average number of days that a resident who is not of low socioeconomic status felt depressed, sad, or anxious in the last month. $\hat{\beta}_1$ tells us the average *difference* between days reported as depressed, sad, or anxious in the last month by low SES residents versus all other residents.

b. Ruth now adds distance to major airport to the model, estimating the following regression, where *distance*$_i$ is a continuous variable measuring the distance between resident i's home and the closest major airport, measured in miles. Interpret $\hat{\beta}_2$ in words.

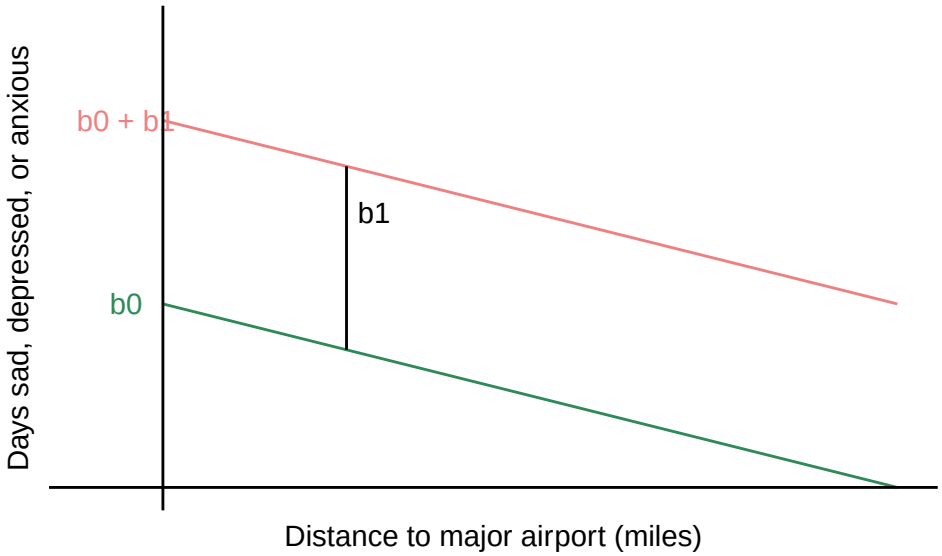
$$\text{days sad}_i = \beta_0 + \beta_1 \text{Low SES}_i + \beta_2 \text{distance}_i + \varepsilon_i$$

Answer: $\hat{\beta}_2$ is the change in days reported as sad, depressed, or anxious in the last month due to a one mile increase in distance between residence and major airport.

c. Suppose Ruth finds that $\hat{\beta}_1 > 0$ and $\hat{\beta}_2 < 0$. Draw (approximately) on the following figure the predicted relationship between *days sad* and *distance* for both low SES and non-low SES residents Label $\hat{\beta}_0$ and $\hat{\beta}_1$.

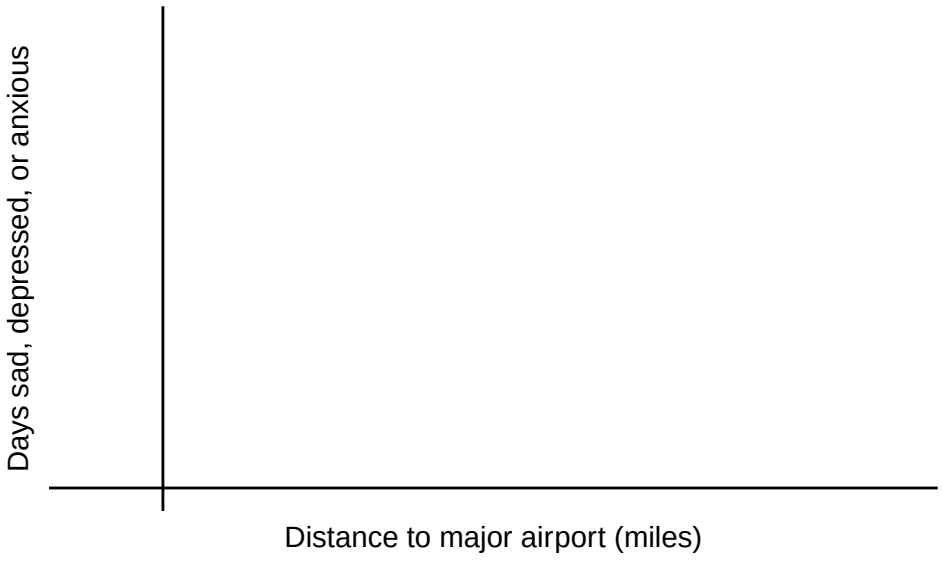


Answer:

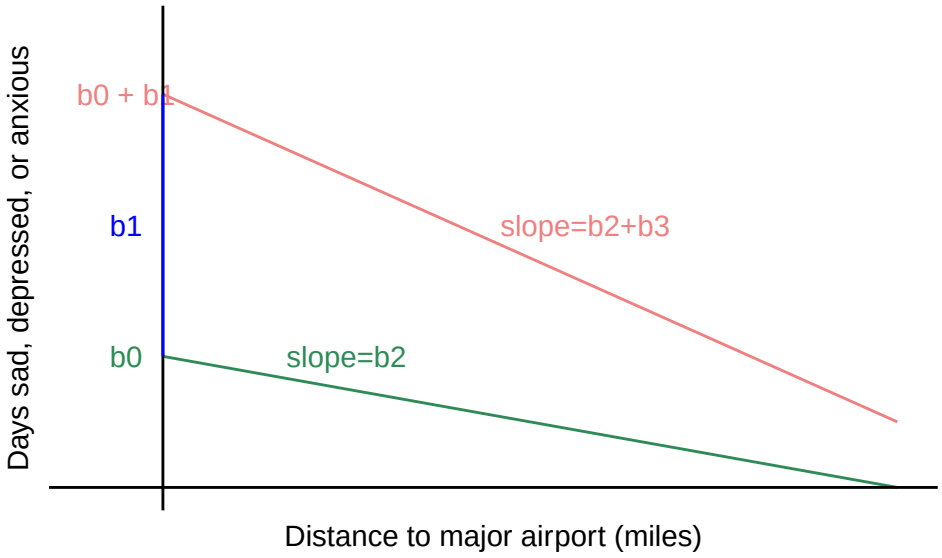


d. Prior evidence suggests people in low SES households may be more affected by noise pollution than other residents. To account for this, Ruth adds an interaction term to her model, as follows. She estimates that $\hat{\beta}_3 < 0$. Draw (approximately) the predicted relationship between *days sad* and *distance* for both low SES and non-low SES residents under this new model. Label $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_2 + \hat{\beta}_3$.

$$\text{days sad}_i = \beta_0 + \beta_1 \text{Low SES}_i + \beta_2 \text{distance}_i + \beta_3 \text{Low SES}_i \times \text{distance}_i + \varepsilon_i$$


Days sad, depressed, or anxious
Distance to major airport (miles)

Answer:


Days sad, depressed, or anxious
b0 + b1
b1
b0
slope=b2+b3
slope=b2
Distance to major airport (miles)