

## DS 1000

### Assignment 2 – due October 20, 2021, at 23:55 EST

#### Chapters 3, 4 and 5

- Questions with the computer symbol  must be answered using Python. All codes must be provided.
- Submission must be done via Gradescope.





#### Question 1 (10 pts)

A high-profile consulting company chooses its new entry-level employees from a pool of recent college graduates using a five-step interview process. Unfortunately, there are usually more candidates who complete the interview process than the number of newly available positions. As a result, cumulative GPA is used as a tie-breaker. GPAs for the successful interviewees are Normally distributed, with a mean of 3.33 and a standard deviation of 0.20.

- (3 pts) What is the proportion of candidates with a GPA between 3.13 and 3.53? Use the 68-95-99.7 rule for Normal distributions.
- (3 pts) What is the proportion of candidates with a GPA above 2.93? Use the 68-95-99.7 and draw a picture that illustrates your rationale.
- (4 pts) How high must be a candidate GPA be to be placed in the top 5% of all successful interviewees? Use Table A from the textbook. Show all your work.





#### Question 2 (12 points)

The file **RBC>Returns.csv** contains the % daily change in Royal Bank Stock price from 1995 until October 1<sup>st</sup>, 2021.

- (3 pts)  Make a histogram of the % daily changes with the density curve estimate on the same plot. Comment on the shape of the distribution.
- (3 pts)  Calculate the mean ( $\bar{x}$ ) and standard deviation ( $s$ ) of the % daily changes.
- (3 pts)  Using the results of part b), find the number of data points with % daily change between  $\bar{x} - 3s$  and  $\bar{x} + 3s$ . Divide this number by the total number of data points in the dataset obtaining a proportion. Compare this proportion to the proportion between  $\mu - 3\sigma$  and  $\mu + 3\sigma$  given by the Normal density curve. Based on your results, does there appear to be a departure from the Normal distribution? Explain.
- (3 pts)  Make a boxplot of these % daily changes. Does the boxplot support your findings in part b)? Explain.

### Question 3 (18 pts)




The common fruit fly *Drosophila melanogaster* is the most studied organism in genetic research because it is small, is easy to grow, and reproduces rapidly. The length of the thorax (where the wings and legs attach) in a population of male fruit flies is approximately Normal, with mean 0.800 millimeter (mm) and standard deviation 0.078 mm.

- a) (3 pts)  What proportion of flies have thorax length less than 0.7 mm?
- b) (3 pts) How would you calculate the proportion in part a) if you only had access to Table A in your textbook?
- c) (3 pts)  What proportion of flies have thorax length greater than 1.0 mm?
- d) (3 pts) Draw a picture illustrating how you can calculate the proportion of flies with thorax length between 0.7 mm and 1.0 mm.
- e) (3 pts)  Calculate the proportion illustrated in part c).
- f) (3 pts)  What value of thorax length gives a 25% proportion of flies above it?

### Question 4 (10 pts)

The deadly Ebola virus is a threat to both people and gorillas in Central Africa. An outbreak in 2002 and 2003 killed 91 of the 95 gorillas in seven home ranges in the Congo. To study the spread of the virus, scientists measured the time in number of days until deaths began across different distances (numbers of home ranges) separating a group of gorillas from the first group infected. Here are the data (also available at the file [gorillas.csv](#)):




|          |   |    |    |    |    |    |
|----------|---|----|----|----|----|----|
| Distance | 1 | 3  | 4  | 4  | 4  | 5  |
| Time     | 4 | 21 | 33 | 41 | 43 | 46 |

- a) (3 pts)  Make a scatterplot. Which variable is the explanatory variable? What kind of pattern does your plot show?
- b) (3 pts)  Find the Pearson correlation coefficient  $r$  between distance and time.
- c) (4 pts)  Recalculate the correlation  $r$  from part c) with time in days replaced by time in number of weeks until deaths began (for example, 4 days becomes 4/7 weeks; 21 days becomes 3 weeks; 33 days becomes 33/7 weeks, and so on). Did the correlation between distance and time change in comparison to the one found in part c)? Why or why not?

### Question 5 (18 pts)




Do poorer people tend to have shorter lives than richer people? Two researchers ranked all counties in the United States by their poverty level and then divided them into 20 groups, each representing approximately 5% of the overall U.S. population. The bottom 5% were the least impoverished (wealthiest), and the top 5% (the 100th percentile) were the most impoverished. Life expectancies at birth for each of the 20 groups were calculated based on life tables. The file

**lifeexp.csv** contains the poverty percentiles (with higher percentiles corresponding to greater poverty) and life expectancies at birth (in years) in 2010 for males and females in the 20 groups.

- a) (4 pts)  Make a scatterplot of the life expectancy versus poverty level percentile rank, using separate colors for men and women.
- b) (3 pts) What does your plot show about the pattern of life expectancy? What does it reveal about the effect of sex on life expectancy?
- c) (4 pts)  Calculate the Pearson correlation coefficient  $r$  between life expectancies and poverty level percentile ranks for men and women altogether. What does the result show about the relationship between these two variables?
- d) (4 pts)  Calculate the Pearson correlation coefficient  $r$  between life expectancies and poverty level percentile ranks for men and women separately. How do these results compare with the correlation obtained in part c)?
- e) (3 pts) Based on your previous answers, is it important to consider the effect of sex when analyzing the relationship between life expectancy and poverty? Explain your answer.

#### Question 6 (18 pts)

The file **beer.csv** contains data on the price of a hotdog against the price of beer (per ounce) at 30 major league ballparks in 2019 in the United States.




- a) (3 pts)  Make a scatterplot of the data points along with the regression line using the `sns.lmplot()` function in Python. What does the plot say about the relationship between price of a hotdog and price of beer?
- b) (4 pts)  Obtain the least-squares regression line for predicting the price of a hotdog in dollars from the price of beer (per ounce) in dollars.
- c) (3 pts) What does the slope say about the relationship between price of a hotdog and price of beer?
- d) (4 pts) What is predicted price of a hotdog when one ounce of beer costs 0.35 dollars? Compare this price with the predicted price of a hotdog when one ounce of beer costs 0.65 dollars.
- e) (4 pts)  How much of the variation in price of a hotdog is explained by the fitted regression line? What does this say about the strength of the relationship between price of a hotdog and price of beer?

#### Question 7 (4 pts)

Researchers analyzed data from more than 5000 adults and found that the more diet sodas a person drank, the greater the person's weight gain. Does this mean that drinking diet soda causes weight gain? Give a more plausible explanation for this association.

### Question 8 (10 pts)

The file [homicide.csv](#) contains data from 2015 for the 11 counties in Ohio, United States, with homicide and suicide rates (rates are per 100,000 people).

- a) (3 pts)  Make a scatterplot that shows how suicide rate can be predicted from homicide rate. Comment on the pattern your plot shows.
- b) (3 pts)  Make two new scatterplots by adding to the original data new data points. In the first scatterplot, add Point A corresponding to a homicide rate of 21.8 and suicide rate of 27.6. In the second scatterplot, add Point B corresponding to a homicide rate of 20.2 and a suicide rate of 14.0. In which direction is each of these points an outlier?
- c) (4 pts)  Add the least-squares regression lines to each scatterplot above: for the original 11 counties, for the original 11 counties plus Point A, and for the original 11 counties plus Point B. Which new point is more influential for the regression line? Explain in simple language why each new point moves the line in the way your graphs show.