Name: *Caleb McWhorter — Solutions*
MAT 121
Summer 2019
Homework 3

"Leon, of course it's wrong! The odds against you being right are. . . staggering! You have a great advantage, you know the outcome. You will be wrong. Don't fear it; embrace your wrongness!"
–Dick Solomon, 3rd Rock from the Sun

Problem 1: Construct a dataset with 6 different positive numbers such that the mean of the dataset is less than the median of the dataset. Then construct a dataset with 6 different positive numbers such that the mean of the dataset is greater than the median of the dataset.

Solution. *There are infinitely many answers. Possibilities include:*

$$1, \quad 20, \quad 30, \quad 40, \quad 50, \quad 60$$

and

$$10, \quad 20, \quad 30, \quad 40, \quad 50, \quad 100$$

Can you see how they were constructed?

Problem 2: Explain what is wrong with the following statement using course concepts: schools with vending machines, especially ones with fatty foods, tend to have higher rates of student obesity. To prevent student obesity, these vending machines should be removed from the schools.

Solution. *This is an example of correlation does not imply causation. It could be that the obese students are 'causing' there to be (more) vending machines or 'causing' (influencing) vending machines to have fattier foods. It could be a coincidence (they do happen). More likely, there is a confounding variable involved. Finally, there could be indeed be a causality relationship, but this can not be determined at face value.*

Problem 3: Which measures of center are resistant? Which measures of center are not?

Solution. *Median is the only resistant measure of center we discussed. The mean and midrange are not resistant measures of center.*

Problem 4: Explain what a p-value is.

Solution. *A p-value is the probability that, if you repeated an experiment, you would see a measurement at least as 'extreme', i.e. large, as the one currently observed.*

Problem 5: The total number of ski rentals at a local ski resort x number of weeks after the start of the season appears to follow a linear pattern. The owner creates a linear regression for the data and finds $\hat{y} = 127.1x + 49.3$ with $r = 0.914$. According to the model, how many ski rentals appear to be made each week during the season? What percent of the variability in the data seems to be explained by the model?

Solution. *The number of ski rentals each week would be approximated by the slope of the model. So there seem to be approximately 127 ski rentals each week during the season. The percent of the variability in the data is explained by the model is the r^2 value. We know $r = 0.914$ so that $r^2 = 0.835$. Therefore, 83.5% of the variability in the data explained by the model.*

Problem 6: During their first semester of college, a student receives the following course grades. What is the students GPA? Be sure to show all your computation.

Course	Credits	Letter Grade
Vector Calculus	4	A
Quantum Mechanics I	3	A−
Women in Music	3	B
Tree Climbing	1	A
Marketing & Media	3	B+
Russian I	4	C−

Solution.

$$\textit{GPA} = \frac{\sum \textit{weight} \cdot \textit{credits}}{\textit{total credits}} = \frac{4(4) + 3.66(3) + 3(3) + 4(1) + 3.33(3) + 1.66(3)}{18} = 3.052$$

Problem 7: Consider the following dataset:

$$14, \quad 15, \quad 16, \quad 17, \quad 23, \quad 26, \quad 29, \quad 33, \quad 46, \quad 67, \quad 71$$

What percentile does the data values 17 represent? What is the value of P_{75} for this dataset?

Solution. *There are 11 values and 3 of them less than 17. Therefore, 17 represents the* $\frac{3}{11} \cdot 100 = 27.3$*rd percentile. Now we know* P_{75} *is the* $\frac{75}{100} \cdot 11 = 8.25 \rightsquigarrow 9$*th data value, which is 46. Therefore,* $P_{75} = 46$.

Problem 8: Suppose the scores of an exam are normally distributed with mean 81 and standard deviation 3. What proportion of students received between a 78 and a 90? Suppose John received a 67 on this exam. Susan took a similar exam, which also had scores that were normally distributed with mean 68 and standard deviation 6. Susan received a 45. Relative to their exams, who did worse? Explain.

Solution. *The proportion of students that received between a 78 and a 90 was...*

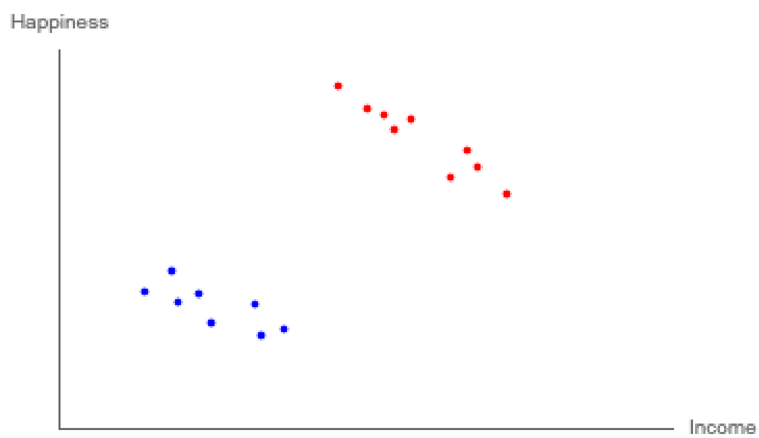
$$0.341 + 0.341 + 0.136 + 0.021 = 0.839.$$

The z*-score measure the 'unusualness' of a score in a distribution. In this case, the lower the* z*-score, the worse a student did. We know that...*

$$z_{\textit{John}} = \frac{67 - 81}{3} = \frac{-14}{3} = -4.67$$
$$z_{\textit{Susan}} = \frac{45 - 68}{10} = \frac{-23}{6} = -3.83$$

Therefore, John did worse on his exam.

Problem 9: In his 1951 paper "The Interpretation of Interaction in Contingency Tables", Edward Simpson described what is now known as Simpson's Paradox. Simpson's Paradox, in essence, is when a trend appears in groups of data but when the data is combined, the trend reverses. Consider the following graph of income and 'happiness.'



The red dots (the upper collection of dots) were females surveyed while the blue dots (the lower collection of dots) were males surveyed. What trend between income and happiness does there seem to be for females and males, respectively and individually? What does the overall trend between income and happiness appear to be for all people? How does this serve as an example of Simpson's paradox?

Solution. *Observe that for both males and females, having a greater income seems to result in a decrease in happiness because there is a downward trend for both groups. However ignoring the groups and looking at the data as a whole, a person having a greater income seems to have more happiness. Therefore, the trend reverses when changing from looking at groups to looking at the overall data.*

Problem 10: Continuing the discussion of Simpson's paradox begun in Problem 9, consider the following scenario: a company gives out bonuses to the best sales person. Two people seem to be in contention for the top sales person: Krystina and Ashton. They sell two sizes of cars. Their sales are summarized below:

Salesperson	Small Cars	Large Cars	Total Cars
Ashton	90 out of 100	16 out of 20	106 out of 120
Krystina	20 out of 20	85 out of 100	105 out of 120

Who had the best overall sales rate? Who had best overall sales rate in each of the categories 'small' and 'large'? Why is this an example of Simpson's paradox?

Solution. *The following chart sums up the percentages of sales:*

Salesperson	Small Cars	Large Cars	Total Cars
Ashton	90%	80%	88.3%
Krystina	100%	85%	87.5%

Notice Krystina was the better salesperson in each category. So one would assume that she would be the better salesperson overall. However, Ashton had the better overall sales percentage. So by this metric, he would be the better salesperson. Therefore, the trend versus depending on whether one looks at each category or the overall data. This is quite the decision for the company!