Midterm Exam 1

Name: _____ PID: _____

This exam has 7 problems with several questions each, worth 100 total points.

You are permitted to use a calculator (without internet capabilities) and one handwritten 8.5×11 "cheat sheet". You may not use your phone, tablet, laptop, or any other electronic device with communication capabilities, note-storing capabilities, or the ability to access the internet. You must turn in your cheat sheet with the exam. Your cheat sheet must clearly have your name on it.

- **Rounding:** Unless otherwise specified, you should either represent your answers as fractions or round them to 3-4 decimal points (or significant digits, if the decimal starts with any zeros). It is strongly advised that you do not round anything until the END of a computation.

Showing your work: On questions where you use a calculator to reach the result, you are REQUIRED to show your work. Merely writing down the answer at the end, without additional support, will NOT receive full credit unless otherwise specified.

- To receive full credit, you must at least write down the equation you are using to complete the computation, plus one additional computational element to demonstrate your understanding.
- It is strongly recommended that the second written piece of evidence be either the step of plugging in all the relevant values into the equation or a list assigning values to each variable in the equation.
- If you make a mistake plugging values into your calculator and end up with the wrong answer, these additional elements will earn substantial partial credit.

Collaborating with anyone else on this exam is strictly forbidden. Using artificial intelligence of any kind on this exam is strictly forbidden. Do not post any question from this exam on Piazza, Chegg, Stack Exchange or any other help platform or tutoring service.

****IMPORTANT**** In signing your name below, you are promising to uphold the university's Honor Code while taking this exam, and certifying on your honor that all work contained herein is strictly your own and that you have received no help from anyone else in completing this exam. **Your exam will not be graded without a signature here**.

On My Honor: _____

STOP

WAIT UNTIL YOU ARE INSTRUCTED TO PROCEED

1. (8 points) A build-your-own-sandwich shop is collecting data on the types of sandwiches its customers build. The owners are particularly interested in which meat options (ham, turkey, roast beef, and salami) and which cheese options (American, Swiss, cheddar, and provolone) its customers like to pair together. Across 231 customers, the shop collects the following data:

	Ham	Turkey	Roast Beef	Salami	
American	23	18	5	8	54
Swiss	30	11	6	1	48
Cheddar	17	24	26	7	74
Provolone	12	6	18	19	55
	82	59	55	35	231

For each of the questions asking for a proportion below, please show any arithmetic steps required to reach your final answer. No additional justification is required, but it may earn partial credit if your final answer is incorrect.

1.1. (2pts) What proportion of the people observed did **not** use turkey?

$$\frac{82 + 55 + 35}{231} = \frac{172}{231}$$

1.2. (2pts) What proportion of the people observed built ham and Swiss sandwiches?

$$\frac{30}{231}$$

1.3. (2pts) What proportion of the people observed meet at least one of the following sandwich-building conditions?

- They chose to put roast beef on their sandwich.
- They chose to put provolone cheese on their sandwich.

$$\frac{55 + 55 - 18}{231} = \frac{92}{231}$$

1.4. (2pts) Does the least popular sandwich combination use the least popular meat option and/or the least popular cheese option? (Briefly) justify your answer.

Yes. The least popular sandwich is salami and swiss, with only 1 order out of 231. The least popular meat is salami, with the fewest orders (35) among all meats, and the least popular cheese is Swiss, with the fewest orders (48) among all cheeses.

2. (12 points) Researchers are interested in whether backpack weights born by college students in North Carolina on a daily basis are leading to chronic back pain. Three methods have been proposed for gathering a sample for this study. For each method, identify one good sampling technique being implemented and one possible issue (e.g. bias, skew, limitations) with the resulting data.

   2.1. (4pts) Choose 10 colleges at random in North Carolina. Visit each campus and spend a few days talking to students from all parts of the campus about their backpack weights and their level of pain.

Good technique: multistage sampling, clusters are colleges but not all students surveyed

Possible issue(s): survey wording / convenience bias, depending on execution; also, could inquire into whether cluster sampling is representative in this case

   2.2. (4pts) Select 50,000 students at random from across all North Carolina colleges to receive a digital survey about the weight of their backpack and their level of pain.

Good technique: simple random sampling across NC

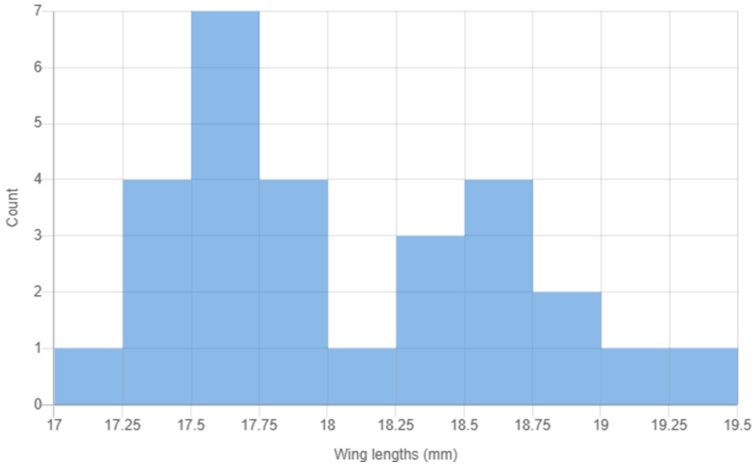
Possible issue(s): nonresponse bias and/or volunteer bias

   2.3. (4pts) Divide up student records at UNC-Chapel Hill according to gender and field of study (humanities, social sciences, engineering, etc). For each group, survey 100 random UNC-Chapel Hill students about their backpack weights and their level of pain.

Good technique: stratified sampling according to key characteristics

Possible issue(s): sampling from only 1 campus, may not be generalizable

3. (14 points) Ecologists in Greenland investigated the effects of rising summer temperatures on the wing lengths of a local butterfly species by recording summer temperatures each year as well as the wing lengths of butterflies born in those summers. The following histogram summarizes the data on butterfly wing lengths (in mm) collected by the ecologists.



3.1. (2pts) Is this data from an experiment or an observational study?

Observational study

3.2. (3pts) Describe the shape of this histogram in terms of both symmetry and modality.

Bimodal, right-skew

3.3. (3pts) Based on the shape of the histogram, do you expect the mean of this data to be higher or lower than the median? Explain your choice.

The mean is expected to be higher than the median, because it is sensitive to the extreme values in the tail on the right of the histogram.

3.4. (4pts) The ecologists observed a strong negative relationship between temperature and wing length, and concluded that higher summer temperatures cause the butterflies to exhibit shorter wings. Is this a reasonable conclusion? Why or why not?

No. This is an observational study, and therefore we cannot make causative conclusions.

3.5. (2pts) Suppose you want to conduct a follow-up study to verify the ecologists' results. Name one additional variable that could be a confounder.
(Note: the possible existence of confounding variables does NOT automatically mean the ecologists' conclusion was unreasonable. We don't have information on what other variables the ecologists were considering.)

Many possible (largely interconnected) answers. Air quality / pollution levels, other climate factors such as humidity, population levels of local predators, availability of food sources

4. (16 points) A golfer keeps careful track of how many putts he makes when putting from various distances away from the hole. The data below shows his number of successful putts Y for each distance X in meters. The golfer is interested in predicting how many putts he will make from various distances.

X	Y
3	85
4	88
5	60
6	61
7	43

4.1. (4pts) What is the mean of the variable Y?

Mean of Y: (1/5) * (85 + 88 + 60 + 61 + 43)
= 337/5 = 67.4 made putts

4.2. (4pts) What is the standard deviation of the variable Y?

Standard deviation of Y:
(1/4) * ((85 - 67.4)^2 + (88 - 67.4)^2 + (60 - 67.4)^2 + (61 - 67.4)^2 + (43 - 67.4)^2)
= (1/4) * (309.76 + 424.36 + 54.76 + 40.96 + 595.36)
= (1/4) * 1425.2 = 356.3 made putts

4.3. (4pts) The least squares regression line for this data is $\hat{y} = 122.9 - 11.1x$. Interpret the slope and y-intercept of this line in the context of the data.

Slope: for each increase of 1 yard in distance from the hole, we predict the golfer will make 11.1 fewer putts.

Y-intercept: at a distance of 0 yards from the hole, we predict the golfer will make 122.9 putts.

4.4. (4pts) The least squares regression line above predicts that the golfer will make about 73 putts (rounded to the nearest integer) from a distance of 4.5 meters. Is this prediction trustworthy? Why or why not? Justify your answer.

Yes, this prediction is trustworthy. Our range of x-data has a minimum of 3 and a maximum of 7. Since 4.5 is within this range, this prediction is an interpolation and is therefore trustworthy.

5. (12 points) Some students were interested in how an acidic environment, such as one caused by acid rain, might affect the growth of plants. They planted alfalfa seeds in 15 cups and randomly chose five to get plain water, five to get a moderate amount of acid, and five to get a stronger acid solution. The plants were grown in an indoor room, so the students assumed that the distance from the windows might have an effect on growth rates. For this reason, they arranged the cups in five rows of three with one cup from each acid level in each row.

   5.1. (2pts) Is this an experiment or an observational study?

   Experiment

   5.2. (4pts) Identify the explanatory and response variables in this study, and identify any variables used for blocking or stratification (depending on your answer to 2.1).

   Explanatory: acid levels

   Response: alfalfa growth

   Blocking: distance from the windows

   5.3. (3pts) Can the results of this study be used to establish a causal relationship between the explanatory and response variables you identified? Explain your response.

   Yes. This is an experiment, so we will be able to draw conclusions about how acid levels in water cause changes in alfalfa growth.

   5.4. (3pts) Can the results of this study be generalized for plant growth and acidic conditions? Explain your response.

   No. This is an experiment, taking place in heavily controlled conditions, and so it cannot be generalized. The easiest example is to point out that different plants are likely to respond differently to acid (we cannot generalize alfalfa results to other plants) but there are many possible responses.

The 2005-06 Pittsburgh Steelers were Super Bowl victors for the first time in 25 years, a feat achieved by winning a wild-card game to get into the playoffs and pushing through several tough challenges along the way. The next few questions focus on data from this season.

6. (20 points) In American football, a team makes progress by reaching a certain checkpoint on the field to earn a "first down", which then gives them the opportunity to earn another "first down" by reaching the next checkpoint, until they get far enough to score. The following data is the number of first downs the Steelers achieved in each of the 20 total games in the 2005-06 season, ordered from smallest to largest:

10	13	14	14	14	16	17	18
18	19	19	20	20	20	20	20
21	25	25	28				

6.1. (4pts) Which type of numerical data is this - discrete or continuous? Explain your choice.

Discrete. Since a "first down" is measured by achieving a checkpoint, the number of first downs must always be an integer, so the possible values for this variable will be the numbers 0, 1, 2, 3, ... and are discrete.

6.2. (4pts) What is the 5-number summary for this dataset? Make sure to show all computations for values that require them.

Q1: (14+16) / 2 = 15
Med: (19+19) / 2 = 19
Q3: (20+20) / 2 = 20

5-number summary:
10, 15, 19, 20, 28

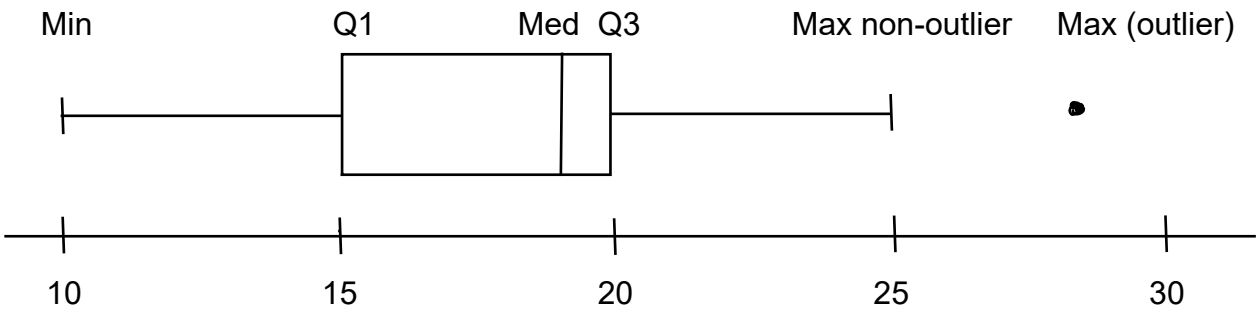
6.3. (4pts) Are there any outliers in this data, according to the IQR measure for outliers?

IQR = 20 - 15 = 5
1.5 IQR = 7.5

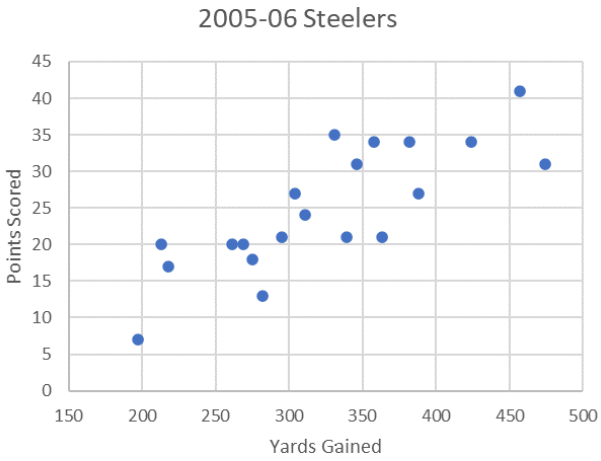
Thresholds:
Q1 - 1.5 IQR = 7.5
Q3 + 1.5 IQR = 27.5

Any point below 7.5 or above 27.5 is an outlier. Therefore there is one outlier in this data: the value 28.

6.4. (8pts) Draw a box plot for this data, including any outliers identified in the previous question.



7. (18 points) The following scatterplot shows the total yards gained and total points scored by the Pittsburgh Steelers in each of the 20 games of the 2005-06 season. The correlation coefficient is $r = 0.8056$. The team's mean points scored is 24.8 with a standard deviation of 8.53 points, and their mean yards gained is 324.35 with a standard deviation of 77.23 yards.



7.1. (3pts) Based on the structure of this scatterplot, what is the explanatory variable and what is the response variable?

Explanatory (x): yards gained
Response (y): points scored

7.2. (4pts) In the first game of the season, the Pittsburgh Steelers gained 424 yards and scored 34 points. Is this datapoint adding a positive or a negative contribution to the correlation between yards gained and points scored? Justify your answer.

424 yards is greater than the mean yards (324.35) and 34 points is greater than the mean points (24.8), putting this datapoint in "Region 2" according to our notes. The x-component and y-component for this point will both be positive, meaning their product is positive, for an overall positive contribution to the correlation.

7.3. (4pts) Compute the least squares regression line for this data, according to your decisions for explanatory and response variables from Part 1.

$b_1 = r\,(S_y / S_x) = 0.8056\,(8.53 / 77.23) = 0.08898$
$b_0 = \overline{y} - b_1\,\overline{x} = 24.8 - 0.08898\,(324.35) = -4.06$

Line: $y = -4.06 + 0.08898x$

7.4. (4pts) Does this regression line overestimate or underestimate y for the datapoint in Part 2? Justify your answer.

Predicted value: -4.06 + 0.08898 (424) = 33.67
This is less than the observed value, so the regression line is underestimating y.
(Alternative: residual is 34 - 33.67 = 0.33, which is positive, so the point has been underestimated.)

7.5. (3pts) What percentage of the variation in the response variable can be attributed to this regression line?

Percentage attributable to the regression line: $r^2 = 0.649$