11.220-Quantitative Reasoning and Statistical Methods for Planning I
Midterm exam: April 12, 2000

Name: ___________________________ Academic advisor’s name: ______________________

TA: □ Saleem Ali □ Raja Shankar □ Sumeeta Srinivasan

Instructions (please read carefully!):

1. You are allowed to use two textbooks of your choice and two pages (4 sides) of your notes during the exam. You may not share these materials with other students.

2. You may use a calculator during the exam; however, you may not use any statistical functions on the calculator. You are only permitted to use the calculator’s simple arithmetic operations (addition, subtraction, multiplication, and division). Please read and sign the statement at the bottom of this page regarding these restrictions.

3. For each problem, show your work in the space provided. We encourage you to draw pictures where appropriate. Partial credit will be given if we have evidence that you framed the problem correctly and/or were headed in the right direction to obtain the correct answer. Points will be deducted if you are asked to show your work for a problem and fail to do so. You may use the back of any page to complete your work. If you need additional scratch paper during the exam, please let us know and we will provide it for you. In all cases, indicate clearly what your final answer is for each question.

4. Remember to keep your writing clear, tight, and to the point. You do not need more than a few well-written sentences to answer any of the questions on the exam.

5. Point values are included in [brackets] next to each question on the exam.

6. During the exam, please let us know if you find unclear printing, typing, or errors on the exam. We will not give hints about how to answer a question. If you think a question is unclear, clearly state your assumptions on the exam and complete the question to the best of your ability.

7. Note that one student will be taking the exam at a later date. Please do not discuss the exam with anyone else until the exams have been returned (or your grades have been reported to you).

Please read this statement carefully and sign your name below.

I certify that I have neither given to nor received assistance from another person on this exam, and that I have used only simple arithmetic functions on my calculator.

Signature: ___________________________
Multiple choice questions: Each is worth 3 points.

1. The 58th percentile of a distribution is...
   ___a. ...58% of the average (mean) value of the dataset.
   ___b. ...the value with 58% of the data values below it.
   ___c. ...the value with 58% of the data values above it.
   ___d. ...58% of the sample size.
   ___e. ...none of the above.

2. Given the regression equation \( \hat{Y} = -4.3 + 5.9X \), which of the following statements is incorrect?
   ___a. The \( r^2 \) value could be less than 0.
   ___b. The correlation between \( x \) and \( y \) is positive.
   ___c. The slope of the line is 5.9.
   ___d. Given an \( X \) value of 2, the predicted value of \( Y \) is 7.5.
   ___e. The intercept value is less than 0.

3. Pearson’s correlation coefficient (\( r \)) should not be used to analyze variables that...
   ___a. ...are not normally distributed.
   ___b. ...have values less than 0.
   ___c. ...are perfectly correlated.
   ___d. ...have a curvilinear relationship.
   ___e. ...none of the above.

4. Jennifer is researching creativity among Americans. She chooses the variable “skull size,” measured in centimeters diameter, as an indicator for the construct “creativity.” Given your knowledge of measurement principles, what would you conclude about Jennifer’s indicator?
   ___a. It is a fairly valid and reliable measure of creativity.
   ___b. It is a fairly valid, but not a very reliable, measure of creativity.
   ___c. It is a fairly reliable, but not a very valid, measure of creativity.
   ___d. It is neither a very reliable nor a very valid measure of creativity.

6. An \( r^2 \) value of 0.75 in a simple ordinary least squares regression model... Check all that apply.
   ___a. ...is probably not high enough for a model to be considered “good.”
   ___b. ...signifies that the dependent and independent variable are positively correlated.
   ___c. ...has nothing to do with the correlation of the dependent and independent variable.
   ___d. ...means that changes in the independent variable explain 75% of the variation in the dependent variable.
   ___e. ...none of the above.
7. The graph below (reprinted from the Economist) presents transportation fatality data from the UK for 1995.

[Graph showing transportation fatality data for 1995 with modes of transport including Motorcycle, Foot, Pedal cycle, Car, Bus, Rail, Air, and Water.]

[8] (a) Which of the following statements about the graph, or about the data it presents, is (are) correct? Check all that apply.

_X_ a. Based on the Economist’s index, water is the safest mode of transport among those presented. Yes.

___b. One rail passenger died in the UK in 1995. No; the expression “rail=1” simply means that rail has been chosen as a reference category with which the other transport forms are being compared. (Moreover, the index is based on distance traveled, not number of fatalities per transport mode.)

___c. The rate of death among persons on motorcycles was 90% higher than that among pedestrians (travelers on foot). No; remember that the index values compare each mode to rail. Thus, we can say that the rate of death on motorcycles in 1995 was (220-130)/130=69.2% higher than that among pedestrians.

___d. For every one person killed on a train in 1995, 99 cyclists were killed. No; the index is based on distance traveled, not on number of deaths.

_X_ e. The rate of death among pedestrians (travelers on foot) was 130% higher than that among rail passengers. Yes.

8. In their 1999 study of homelessness in Los Angeles County, Cousineau and Simabukuro found that roughly 5% of all LA county adults were without a place to sleep or live at some point during the period 1994-1999. Suppose you would like to conduct a study of homelessness in the LA area and want to interview only individuals who have experienced homelessness. You draw a random sample of 20 adults in LA county, and you want to find the probability that 3 or fewer of them have been homeless in the past five years.
[4] a. To save some time, you would like to use the normal approximation for the sampling
distribution of a sample proportion to answer this question. Is it appropriate to do so in
this case? Why or why not?

No, we should not use the normal approximation in this case. The criteria for using the
approximation are:

\[ np > 10 \quad \text{AND} \quad n(1-p) > 10 \]

In our case, we have:

\[ np = (20)(.06)=1.2 \quad \text{and} \quad n(1-p) = (20)(.94) = 18.8 \]

[8] b. Using the correct approach, show the strategy you would use to determine the probability
that 3 or fewer of the people in your sample are not registered to vote. Write out the
equation(s) you would use to find this value and make it clear how you would arrive at
your answer; however, you do not need to do the computations and arrive at a final
probability value.

We thus work this out as a binomial problem with \( n=20 \) and \( p=.05 \). It requires 4 values:
The probability of finding 3, 2, 1, or no homeless persons in our sample. The equations
are:

\[ P(X = r) = \binom{n}{r} p^r q^{n-r} \]

\[ P(X = r) = \binom{20}{3} (.05)^3 (.95)^{17} = 0.0596 \]
\[ P(X = r) = \binom{20}{2} (.05)^2 (.95)^{18} = 0.1887 \]
\[ P(X = r) = \binom{20}{1} (.05)^1 (.95)^{19} = 0.3774 \]
\[ P(X = r) = \binom{20}{0} (.05)^0 (.95)^{20} = 0.3585 \]

To find the probability that \( X \leq 3 \), we simply add these four probabilities together:

\[ 0.0596 + 0.1887 + 0.3774 + 0.3585 = 0.9841 \]

[4] c. What is the expected value (i.e., the mean) of \( x \) in this example?

For a binomial random variable, the expected value is given as \( E(x) = \mu = np \). In our
case, \( np = (20)(.05) = 1 \).

[6] d. Based on your answer to part (c), what do you learn about planning your study,
particularly regarding the feasibility of finding individuals who fit your criteria using a
simple random sampling procedure? (Two or three sentences should suffice.)

The probability that your sample will have very few individuals who have experienced
homelessness—the population you are interested in—is very high. This is because you
are looking for people who have a characteristic that is quite rare in the population
(consider that a simple random sample of 100 is expected to yield only 6 individuals that
you want to interview!). Evidence suggests that it will be quite expensive and time
consuming to draw an SRS large enough to yield a “sufficient” number of cases
(individuals who have been homeless) for your study. If you are constrained for time or
money, perhaps you should consider using a different sampling approach for your study.
9. A stratified random sample of students from eight high schools in one city were given a written survey which asked about their smoking behavior and the smoking behavior of their parents.

<table>
<thead>
<tr>
<th>Student smokes</th>
<th>Student does not smoke</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both parents smoke</td>
<td>400</td>
</tr>
<tr>
<td>One parent smokes</td>
<td>416</td>
</tr>
<tr>
<td>Neither parent smokes</td>
<td>188</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1004</strong></td>
</tr>
</tbody>
</table>

[6] a. What is the probability that a randomly selected student is a smoker if neither of his/her parents smoke? (Show your work.)

Let $P(A) = \text{probability that student is a smoker}$
and $P(B) = \text{probability that neither parent smokes}$

We would not assume that the smoking behavior of students and parents is independent.

This is a conditional probability question. We have to find $P(A/B)$. One can directly compute this from the table as $188/1356 = 0.139$. 1356 is the number of cases in which neither parent smokes. In 188 out of these 1356 cases the student smokes. Hence the answer as $188/1356$.

Alternatively, $P(A/B)=P(A \text{ and } B)/P(B)$

$P(A \text{ and } B) = 188/5375$, $P(B)=1356/5375$

Therefore $P(A/B) = (188/5375)*(5375/1356) = 188/1356 = 0.139$.

[6] b. What is the probability that a randomly selected student is a smoker, given that either one or both of his/her parents smoke? (Show your work.)

Let $P(A) = \text{probability that student is a smoker}$
and $P(B) = \text{probability that one or both parents smoke}$

This is also a conditional probability question. We have to find $P(A/B)$. One can directly compute this from the table as $(400+416)/(1780+2239) = 816/4019 = 0.203$. 4019 is the number of cases in which one or both parents smokes. In 816 out of these 4019 cases the student smokes. Hence the answer as $816/4019$.

Alternatively, $P(A/B)=P(A \text{ and } B)/P(B)$
\[ P(A \text{ and } B) = \frac{816}{5375}, \quad P(B) = \frac{4019}{5375} \]

Therefore \( P(A/B) = (\frac{816}{5375}) \times (\frac{5375}{4019}) = \frac{816}{4019} = 0.203. \)

[4] c. Based on your answers to parts (a) and (b), what do you conclude regarding the relationship between a student’s decision to smoke and his/her parents’ smoking behavior? (Just one or two clear statements.)

It appears that students who have at least one smoking parent have a substantially higher probability of being smokers themselves (probably not a very surprising result).

10. The weights of newborn children in the United States vary according to the normal distribution with mean 7.5 pounds and standard deviation 1.25 pounds. The government classifies a newborn as having low birth weight (LBW) if his/her weight is less than 5.5 pounds. A newborn is classified as high birth weight (HBW) if his/her weight is more than 10.0 pounds. (National Center for Health Statistics)

[8] (a) What is the probability that a baby chosen at random is a low birth weight (LBW) baby, i.e., that s/he weighs less than 5.5 pounds at birth? (Show your work.)

This is a straight-forward normal probability problem. We form a z-score with the LBW value and find the associated p-value from the standard normal table:

\[ z = \frac{X - \mu}{\sigma} = \frac{5.5 - 7.5}{1.25} = -1.60 \]

From the standard normal table, we find a value of 0.4452 associated with a z-score of 1.60 (remember that the negative does not matter when finding the value in the table, only when thinking about what part of the distribution we care about and thus how to compute the final probability value).

Because the value 0.4452 represents the area between the mean value, 7.5, and the z-score of –1.60, and we are interested in the area in the tail of the distribution (i.e., beyond the score of –1.60), we subtract the table value from 0.5:

\[ P = 0.50 - 0.4452 = 0.0548 \]

[8] (b) What percentage of babies born each year in the US have “normal” weights (i.e., birth weights between 5.5 and 10 pounds)? (Show your work.)

We want to find the area bounded by two z-scores: -1.60 (from part a) and one computed for 10 pounds. This second z-score is:

\[ z = \frac{X - \mu}{\sigma} = \frac{10 - 7.5}{1.25} = 2.0 \]

The p-value for a z-score of 2.0 is 0.4772. Remember that the standard normal table only deals with one half of the distribution at a time. The z-score for 5.5 pounds is in the side of the distribution below the mean while the score of 10 pounds is above the mean. Thus,
If the add the values for the two z-scores, we find the area between them, which is the percentage of all birthweights greater than 5.5 pounds and less than 10 pounds.

\[ P = 0.4552 + 0.4772 = 0.9324 \]  
\[ 93.25\% \] of all births

10. The University of Michigan’s General Social Survey (GSS) is a nationwide survey of adults conducted each year using a simple random sampling procedure. In 1986, the GSS included a question about the length of commute (one way, in minutes) that the 1,330 respondents had each day, on average, to their place of work. The responses from this question are provided below.

<table>
<thead>
<tr>
<th>Commute time (minutes)</th>
<th>Frequency</th>
<th>Commute time (minutes) (cont’d.)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>54</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>34</td>
<td>25</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>23</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>30</td>
<td>155</td>
</tr>
<tr>
<td>5</td>
<td>153</td>
<td>32</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>35</td>
<td>30</td>
</tr>
<tr>
<td>7</td>
<td>13</td>
<td>40</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>45</td>
<td>63</td>
</tr>
<tr>
<td>10</td>
<td>190</td>
<td>46</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>50</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>55</td>
<td>3</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>60</td>
<td>55</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>70</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>203</td>
<td>75</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>90</td>
<td>8</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>97</td>
<td>8</td>
</tr>
<tr>
<td>20</td>
<td>180</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[4] (a) There is some evidence that the data collected for this question do not provide a very precise profile of commuting times among Americans. What is this evidence? (One or two sentences will suffice; no formulas or computations are necessary.)

The evidence is that more than 85% of respondents report a commuting time ending in a 5 or 0. In reality this is probably not true; people just “round off” to the nearest five minutes for convenience.

[10] (b) The mean one-way commute time for this sample is 20.2 minutes with a standard deviation of 17.1. Assuming a normal distribution of commuting times, compute a 99% confidence interval for the population mean one-way commute time in the US during 1986. (Show your work.) Explain in one clear sentence, to someone with no statistics training, what you can conclude about commute length in the US based on your analysis.

\[ \bar{x} - E < \mu < \bar{x} + E \]
For the GSS data: $20.2 - E < \mu < 20.2 + E$

$$20.2 - z\left(\frac{\alpha}{2}\right) \frac{s}{\sqrt{n}} < \mu < 20.2 + z\left(\frac{\alpha}{2}\right) \frac{s}{\sqrt{n}}$$

$$20.2 - (2.575) \frac{17.1}{\sqrt{1330}} < \mu < 20.2 + 2.575 \frac{17.1}{\sqrt{1330}}$$

Thus, the 99% confidence interval for mean commuting time among Americans is:

$$18.99 \text{ minutes} < \mu < 21.41 \text{ minutes}$$

which means that we are 95% confident that the mean commuting time among Americans falls between 18.99 and 21.41 minutes.

[6] (c) In the sample, 54 individuals were assigned a one-way commute time value of 0. Does this make sense? Why or why not? How would your analysis change if these individuals were excluded? (A few clear sentences.)

Individuals who do not work, or who work out of their homes, would have a commute time of 0. It makes sense to include these individuals given the population to which we want to generalize our findings (all Americans). If we were interested in knowing the mean commute time among people who do commute, we would want to use a sampling strategy that excluded individuals who have a zero value for this variable.

EXTRA CREDIT (3 points):

Suppose Saleem plans to conduct a poll of 100 randomly selected MIT students to compute an estimate the proportion of all MIT students who speak more than one language. Sumeeta argues for a sample size of 900. She knows that the standard error of the sample proportion ($\hat{p}$) who report speaking more than one language will be _____ times as large if they use the larger sample. What value belongs in this blank? (You must show your work—how you found this value—in order to get credit for this question.)

___ a. 9
___ b. 1/9
___ c. 3
___ d. 1/3
___ e. It depends on the population size (i.e., the total number of MIT students).

The value of the standard deviation decreases as the sample size $n$ increases, in particular at the rate $pn$. The standard deviation of a sample proportion is given by:
\[ \sigma_p = \sqrt{\frac{p(1-p)}{n}} \]

So a sample 9 times larger has a standard deviation one-third as large. Just plugging in an arbitrary value for \( p \), 0.5, we can test this...

\[
\sigma_p = \sqrt{\frac{.5(1-.5)}{100}} = 0.05 \quad \text{versus} \quad \sigma_{p} = \sqrt{\frac{.5(1-.5)}{900}} = 0.017
\]