MATH 5510: Topics in Statistics

Sample Exam Questions

Notes:

1. The actual exam will be much shorter than this.
2. Portions of the first three questions are adapted from Pearl, Judea; Glymour, Madelyn; Jewell, Nicholas P. (2016) *Causal Inference in Statistics: A Primer*. Wiley. Kindle Edition.

Question 1

What is wrong with the following claims?

(a) "Data show that income and marriage have a high positive correlation. Therefore, your earnings will increase if you get married."

(b) "Data show that as the number of fires increase, so does the number of fire fighters. Therefore, to cut down on fires, you should reduce the number of fire fighters."

(c) "Data show that people who hurry tend to be late to their meetings. Don't hurry, or you'll be late."

Question 2

A baseball batter Tim has a better batting average than his teammate Frank. However, someone notices that Frank has a better batting average than Tim against both right-handed and left-handed pitchers. How can this happen? (Present your answer in a Paik diagram or a table or both)

Question 3

Determine, for each of the following causal stories, whether you should use the aggregate (marginal) or the segregated (conditional) data to attempt to determine the true effect.

(a) There are two treatments used on kidney stones: Treatment A and Treatment B. Doctors are more likely to use Treatment A on large (and therefore, more severe) stones and more likely to use Treatment B on small stones. Should a patient who doesn't know the size of his or her stone examine the general population data, or the stone size-specific data when determining which treatment will be more effective?

(b) There are two doctors in a small town. Each has performed 100 surgeries in his career, which are of two types: one very difficult surgery and one very easy surgery. The first doctor performs the easy surgery much more often than the difficult surgery and the second doctor performs the difficult surgery more often than the easy surgery. You need surgery, but you do not know whether your case is easy or difficult. Should you consult the success rate of each doctor over all cases, or should you consult their success rates for the easy and difficult cases separately, to maximize the chance of a successful surgery?

(c) In a study of a group of male 50 to 55-year-old long-time smokers, researchers compared a group of heavy smokers with matched group (same age range, sex and similar socioeconomic and environmental backgrounds) of light smokers. It was found that lung function was worse in the group of heavy smokers than in the group of light smokers. The researchers also measured the amount of tar deposit in the lungs of the subjects and classified subjects as having heavy or light tar deposits. Would you get a better

indication of the effect of smoking by comparing the aggregated data for the two groups or by comparing the tar level specific data?

Question 4

When interpreting a study that purports to show a relationship between two variables, what are the three questions you should ask? Discuss as succinctly as you can the consequences of the answers to those questions.

Question 5

R. A. Fisher insisted that causal inference was impossible in the absence of an experiment with random assignment to a 'treatment' variable. Discuss why Fisher's position could be considered correct but why it may be considered impractical?

Question 6

Give an example of a situation in which we would be interested in predictive inference and an example in which we would be interested in causal inference.

Question 7

Here are some fictitious data on the rate of complications for appendectomies performed at University Hospital, a large urban teaching and research hospital, and in County Hospital, a small-town hospital.

	Complications	No complications	Total cases
University Hospital	64 (8%)	736	800
County Hospital	10 (5%)	190	200

Suppose that appendectomies can be classified as high risk or low risk and that the high risk cases tend to be directed disproportionately to University Hospital instead of County Hospital. Construct two tables, one for each level of risk, and draw a Paik diagram that shows how it is possible for both high- and low-risk patients to have a lower probability of complications at University Hospital than at County Hospital, although, overall the probability of complications is higher at University Hospital than at County Hospital.

Question 8

Suppose a test for glaucoma has a sensitivity of .95 and a specificity of .90. You receive the test as a routine test on a regular visit to your optometrist. The prevalence of glaucoma in your age, ethnic and gender group among people who have not been previously diagnosed is 1 per 100. The test, alas, is positive. Use a natural frequency table to find the probability that you have glaucoma given the positive test result.

Question 9

A group of major medical journals are now requiring that all authors who intend to publish in their journal must preregister their experimental designs and their intended analyses for all the clinical endpoints (responses) they intend to report before obtaining data if they intend to publish their results in their journal. Authors must also agree to publish their findings whether the results achieve statistical significance or not. In what ways does this policy contribute to mitigating lack of reproducibility?

Question 10

1. A study investigated whether there was a higher risk of complications when women gave birth at home with the assistance of a midwife instead of giving birth in a maternity ward in a hospital. 400 women who chose to give birth at home and 2,000 women who gave birth in a hospital were studied. The following table summarizes the number of 'complications' in each group:

	Complication	No complications	Total
Home Births	20	380	400
Hospital Births	200	1800	2000
Total	220	2180	2400

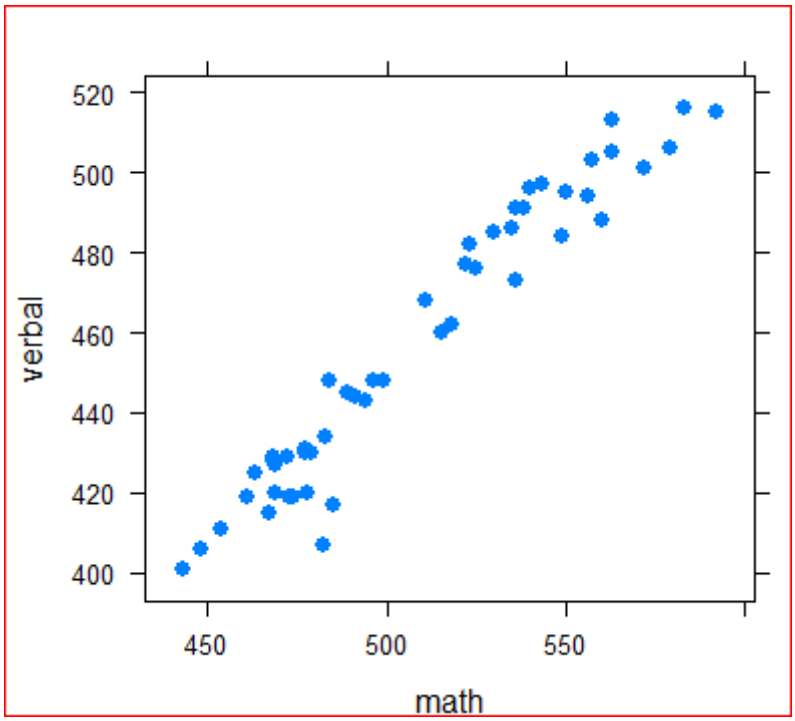
a) Do you think that this is an experimental or an observational study? Justify briefly.

b) The data suggest that it is safer (in the sense of a lower rate of complications) to give birth at home than to give birth in the hospital. Discuss whether this implies that a woman should consider giving birth at home in order to reduce her risk of complications. Identify at least one plausible confounding factor and one plausible mediating factor that could partly explain the results of the study (continue overleaf if you need more space).

c) Choose a possible confounding factor and use an Agresti diagram to show how controlling for this confounding factor could reverse the direction of association between the rate of complications and the location of birth: home or hospital.

Question 11

2. This graph shows the mean verbal and mean math scores on the SAT test in each of the 50 states of the US.



Make an intelligent guess of the correlation between these two variables. Show the basis for your guess– perhaps by drawing an appropriate geometric figure on the graph above.

Question 12

Suppose you were to read about a study based on a random survey of Ontario medical records that shows that smokers have twice as high a risk of kidney disease as non-smokers. Is it reasonable to conclude that smoking causes a higher risk of kidney disease?

A. No, because the result was clearly based on an observational study.

B. Yes, because the result was clearly based on a randomized experiment.

C. The answer depends on whether the research was based on an observational study or a randomized experiment, and it isn't obvious which was used.

D. The answer depends on whether the research was based on a random sample, which it was, so it is safe to conclude that there is a causal relationship.

E. No, because the baseline risk of kidney disease is not given

Question 13

1. Simpson's Paradox occurs when

    A. No baseline risk is given, so it is not known whether or not a high relative risk has practical importance.

    B. A confounding variable rather than the explanatory variable is responsible for a change in the response variable.

    C. The direction of the relationship between two variables changes when the categories of a third variable are taken into account.

    D. The results of a test are statistically significant but are really due to chance.

2. A test to detect HIV had a sensitivity of 95%. This means that

    A. 95% of people who test positive will actually have HIV.

    B. 95% of people with HIV will test positive.

    C. 95% of people who do not have HIV will test negative.

    D. 95% of people who test negative will actually not have HIV.

3. A test to detect HIV had a specificity of 90%. This means that

    A. 90% of people who test positive will actually have HIV.

    B. 90% of people with HIV will test positive.

    C. 90% of people who do not have HIV will test negative.

    D. 90% of people who test negative will actually not have HIV.

Question 14

In a screening procedure in which all employees of a large firm are required to take an HIV test, your friend tests positive. Suppose that the prevalence of HIV for people similar to your friend is 2%.

a) What is the probability that your friend has HIV in view of the positive test result?

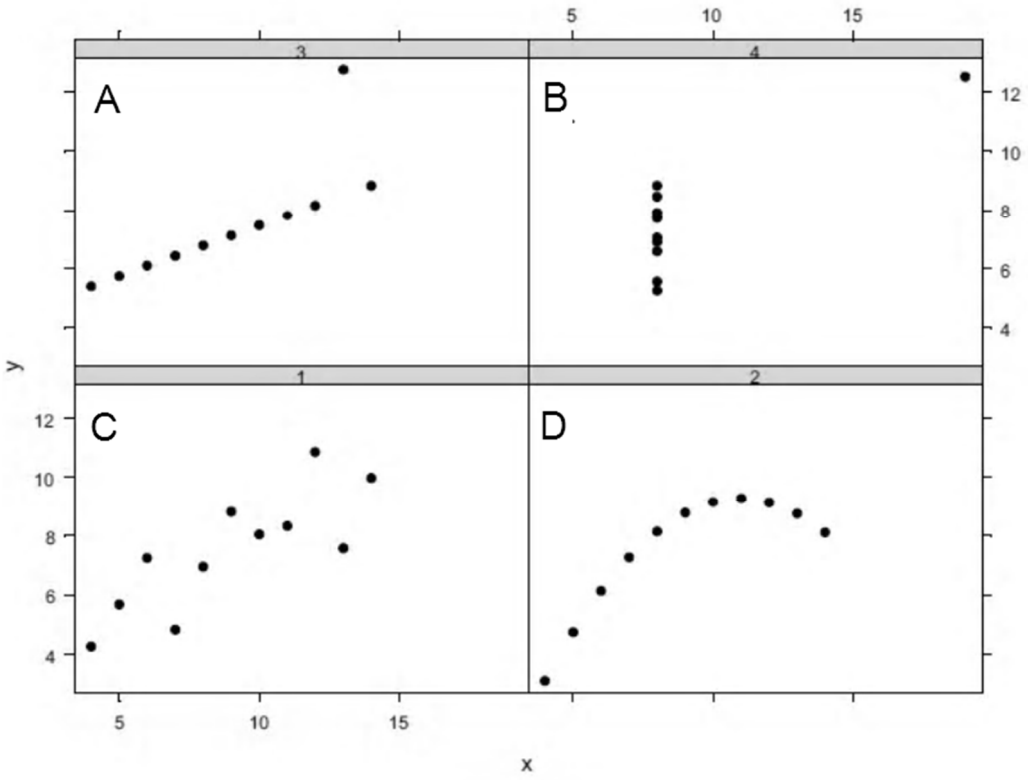
b) What would the probability be for someone who is at high risk and belongs to a group with a prevalence of 30%?

Question 15

How could you distinguish whether data are 'experimental' or 'observational'? Explain, using as little technical language as you can, why causal conclusions are reasonable with one kind of data but more problematic with the other. [Continue on the back if you need more space]

Question 16

The figure below shows four scatterplots: A, B, C and D.



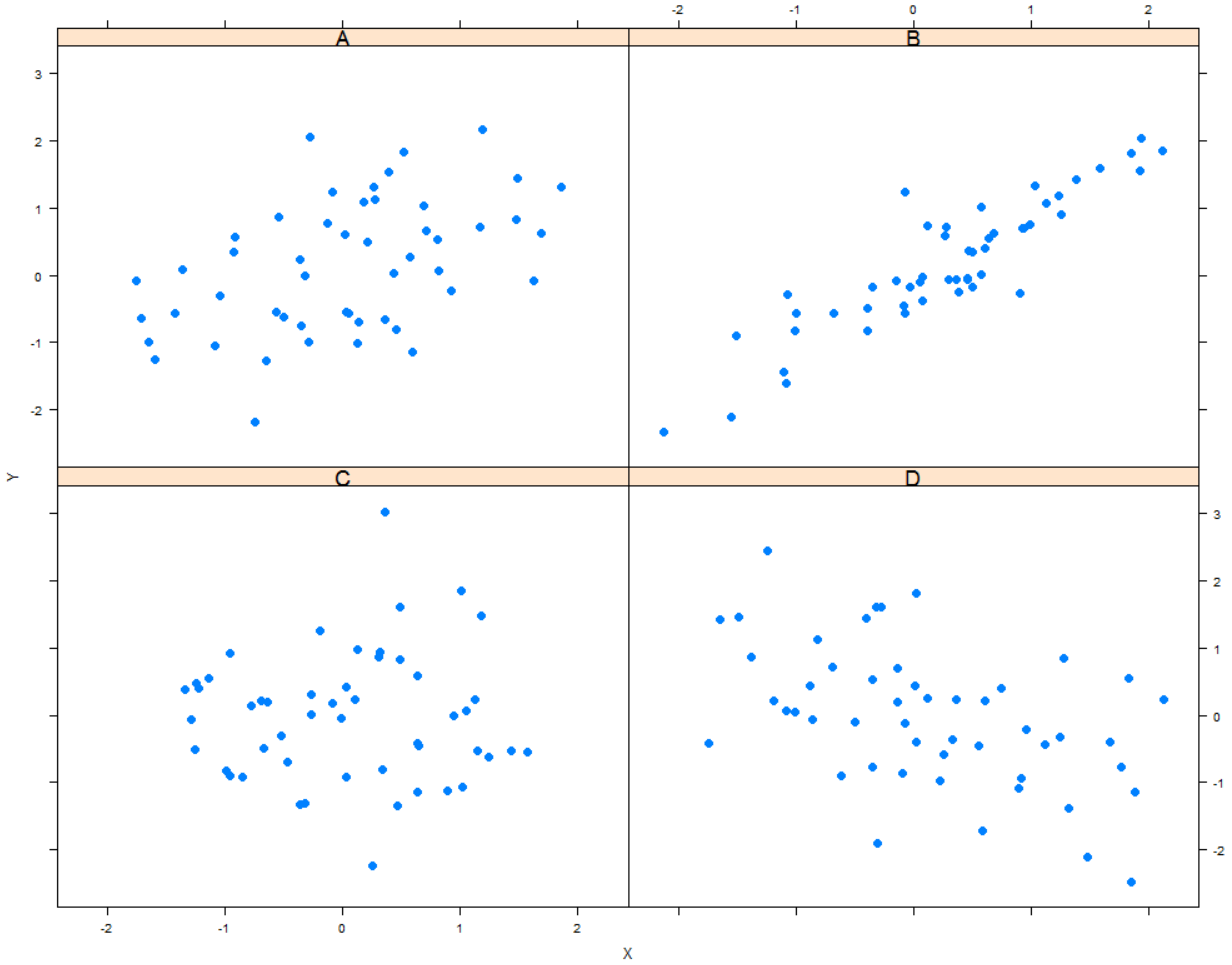
For each of the four scatterplots, identify whether a simple linear regression would be appropriate and briefly support your answer.

Question 17

P-values are almost universally used in scientific research to evaluate evidence against null hypotheses. Criminal courts that assess evidence under the assumption that the accused is innocent follow a procedure that appears formally very similar to testing a null hypothesis. Explain how the use of p-values in court testimony can lead to unfair convictions.

Question 18

The figure below shows four scatterplots.



[8] Of the following possible values for correlation:

a) +1, b) –1, c) 0, d) +0.9, e) –0.9, f) +0.5, g) –0.5

the correct correlation for each scatterplot is:

A)_____ B) _____ C) _____ D) _____

Question 19

4. A study shows that heavy users of sunscreen lotion have a higher chance of developing skin cancer.
   a) Does this imply that you should avoid using sunscreen lotion in order to reduce your chances of developing skin cancer? Why?

   b) List at lease one potential *confounding factor* and one potential *mediating factor*. Justify briefly.

Question 20

Suppose a statistics teacher wants to know whether the number of hours students spent studying in a group affects the final course grade. For each of the following explain whether the research method described is a randomized experiment or an observational study:

a) Each student keeps a log of the hours he or she spends studying in a group and reports the total after the course is completed.
b) Students are randomly assigned to study groups. The teacher tells each group how often to meet. This varies from one hour the day before each test to two hours per week.
c) Students voluntarily join groups based on how often the groups will meet. The groups are designated as meeting weekly, meeting only before exams, or meeting whenever members feel it is necessary.

Question 21

Simpson's Paradox occurs when

A. No baseline risk is given, so it is not known whether or not a high relative risk has practical importance.

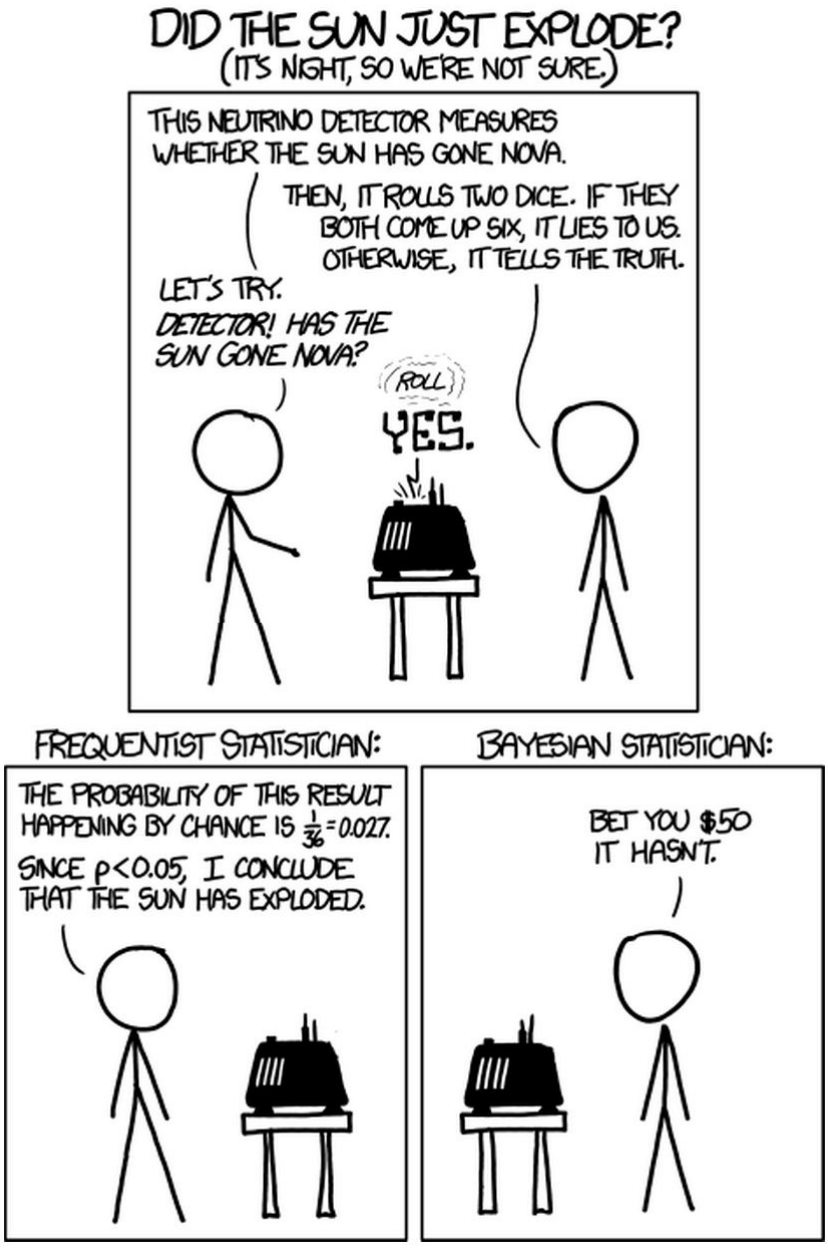
B. A confounding variable rather than the explanatory variable is responsible for a change in the response variable.

C. The direction of the relationship between two variables changes when the categories of a confounding variable are taken into account.

D. The results of a test are statistically significant but are really due to chance.

Question 22

The following XKCD cartoon :



shows two statisticians interpreting the same data: one who uses a frequentist approach unquestioningly and one who uses a Bayesian approach. Make some reasonable assumptions, stating them explicitly, and calculate a reasonable value for the Bayesian statistician's posterior probability that the sun has exploded.

Discuss why there is a difference between the 'p-value' of 0.027 and the Bayesian posterior probability.

Question 23

If all scientists used 0.05 as a level of 'alpha' to decide which results are significant and worthy of publication (publish results with p-values less than 0.05 and don't publish otherwise), that would ensure that only approximately 5% of published results would be wrong and we wouldn't face the current 'crisis of reproducibility' revealing that, in some fields, a very large proportion of published results cannot be reproduced. Is this statement correct? Discuss.

Question 24

Suppose that the correlation between the grade on the midterm and the grade on the final exam in a certain Economics course is 0.7 and that the relationship is reasonably close to linear. Which of the following statements **are implied by this information?** Indicate Y (for Yes) if the statement is implied or N (for No) if it is not.

a) _____ Students who score one standard deviation below the mean on the mid-term will also, on average, score one standard deviation below the mean on the final exam.

b) _____ Students who score two standard deviations below the mean on the mid-term will tend on average to score 1.4 standard deviations below the mean on the final exam.

c) _____ 70% of students will have a lower score on the final exam than they did on the midterm and, paradoxicaly, 70% of students will have a lower score on the midterm than they did on the final exam.

d) _____ Students who do unusually well on the midterm will do better than average on the final exam but not quite as well (in terms of z-scores) as they did on the midterm.

e) _____ We can't say anything about the relationship between grades on the midterm and grades on the final exam because the relationship is essentially random.

f) _____ A student's grade on the final exam will be approximately (on average) 70% of their grade on the midterm.

g) _____ If we express grades using z-scores for each test, then the slope of the least-squares regression line of the mid-term grade on the final exam grade is 0.7.

h) _____ Although it is possible to carry out the calculations to find the least-squares line to predict the midterm grade from the final exam grade, it doesn't make sense to do so because the midterm occurs before the final exam.

i) _____ The correlation would be equal to 1 if and only if each student had the same grade on the final exam as they did on the midterm.

Question 25

Which of the following best describes the standardized (z) score for an observation? (circle one answer):

j) It is the most common score for that type of observation.
k) It is the standard deviation of the observations.
l) It is the number of standard deviations the observation falls from the mean.
m) It is one standard deviation more than the observation.
n) It is the center of the list of scores from which the observation was taken.