STAT117 Spring 2026 Exam #2

- Ask if you have any questions about the exam.
- Be concise in your answers. You will be penalized for adding irrelevant or incorrect information to a response.
- You don't need to round decimals, but if you choose to, keep at least two decimals unless otherwise noted.
- You are allowed a reference sheet. Use both sides of one 8.5 by 11 inch piece of paper and put whatever you want on it. You will turn this sheet in with your exam. Be sure to put your name on it so that it can be returned to you later.
- Calculator use is permitted. You may NOT use your phone as a calculator. You will need to use your laptop, but the only programs you are allowed to have open are Jamovi and the standard Windows or Mac calculator. You need to close (not just minimize) other programs (including your web browser and all messaging apps). You are not allowed to consult any papers (other than your reference sheet), websites, books, or any person except the instructor during the exam.
- You need to download the dataset before the exam. You are not allowed to open a web browser during the exam. You must turn off WIFI on your computer.
- Cases of suspected cheating will be referred to the dean. We take cheating very seriously because it benefits student learning when students get grades that reflect what they know. For this reason, you are allowed to have at your seat only your test, your reference sheet, your computer, a calculator, and writing implements (pen or pencils). You will need to put your jacket, backpack, cell phone, smart watch and anything else you carry to class at the front of the room during the exam.
- I might tell you where to sit, or move your seat, during the exam.

The dataset used for the exam is **GSS_SP2026.sav** (same as exam 1). This dataset is gathered from the General Social Survey 2024, a large scale survey conducted by NORC at the University of Chicago . The levels of measurement in Jamovi are not set correctly, so you need to judge for yourself. Sometimes, the level of measurement is subjective or controversial, and the relevant question will tell you what to do.

Note 1: the data is set up such that any value between -100 and -10 is automatically treated as a missing value. You don't need to do anything about these as Jamovi knows how to handle them (for instance, they are excluded from calculations of any statistics).

Note 2: these three variables are similar: `dem10fut`, `dem10pst`, `demtoday`. Make sure that you are using the correct one as specified in relevant questions.

1. (20 points) In the following problems, you are asked about the relationship between pairs of variables in the data set.

(a) The variables `dem10fut` and `dem10pst` measure people's assessment to "How well will democracy work in America in ten years" and "How well did democracy work in America ten years ago," on a scale of 0 (very poorly) to 10 (very well). Treat both variables as Numeric variables. The relationship between `dem10fut` and `dem10pst` can best be described by the Pearson / Spearman (circle one) correlation coefficient. The value of the correlation coefficient is _____. Is there a relationship between the two variables? If so, state the direction and strength of the relationship and provide a one-sentence interpretation of the relationship in the context of the problem.

Pearson

0.420

Positive, moderate

People who think that democracy in America in ten years will work very well also tend to think that democracy in America ten years ago worked very well, in general.

(b) The relationship between `income` (total family income) and `degree` (respondent's highest degree) can best be described by the Pearson / Spearman (circle one) correlation coefficient. The value of the correlation coefficient is _____. Is there a relationship between the two variables? If so, state the direction and strength of the relationship and provide a one-sentence interpretation of the relationship in the context of the problem.

Spearman

0.337

Positive, moderate

People who have a higher degree tend to have a higher family income, in general.

For the following open-ended problems, report the name and value of at least one appropriate statistic to support your conclusion. When describing the relationship, interpret the relationship in context using terms that someone who has not taken statistics could understand (for example, "there is a positive relationship" is not sufficient) and avoid causal language. If you are using a correlation coefficient, be sure to indicate the direction and strength of the relationship.

(c) Is there a relationship between `othlang` (can respondent speak language other than English/Spanish) and `dem10fut` (how well will democracy work in America in ten years, treat as a Numeric variable)?

Pearson r = -0.119, negative, weak. Respondents who can speak another language (1=Yes) tend to think that democracy in America in ten years will work very well, i.e., more optimistic, compared to respondents who cannot speak another language (2=No).

OR

mean 5.48 vs 4.699 OR median 6 vs 5; Respondents who can speak another language typically have a higher value for `dem10fut` (more optimistic).

(d) Is there a relationship between `othlang` (can respondent speak language other than English/Spanish) and `income` (total family income)?

Rho = -0.033, no relationship.

OR

median 12 vs 12; No relationship; regardless of whether the respondent can speak another language, people typically reported "25,000 or more" for total family income.

2. (13 points) Create a contingency table for the variables `sex` and `gendtech` ("Are women or men benefiting more from digital tech").

(a) Use the table to fill out the blanks.

- i. The count (NOT percentage) of male and female respondents, respectively: _____
- ii. Among male respondents, what percent believe "men are benefitting more than women"? _____%
- iii. Among male respondents, what percent believe "neither men nor women are benefitting"? _____%
- iv. Among female respondents, what percent believe "men are benefitting more than women"? _____%
- v. Among female respondents, what percent believe "neither men nor women are benefitting"? _____%

577, 755, 9.7, 19.4, 16.2, 21.9

(b) Is there a relationship between `sex` and `gendtech`? You must select two numbers from (a) to draw a conclusion. If there is a relationship, describe the relationship in the context of the problem.

Yes, by comparing 9.7% vs 16.2%.

Count as correct if said yes by comparing 19.4% vs 21.9%

(-1) if said there is no relationship since 19.4% and 21.9% are similar

(c) Can we use a correlation coefficient to draw a conclusion between `sex` and `gendtech`? Why or why not? Briefly explain.

No, since `gendtech` is a nominal variable.

3. (10 points) The following questions are based on the variable `childs` (number of children, treat it as a Numeric variable).

(a) If a respondent has no children (i.e., `childs` = 0), what would be the corresponding z-score? Perform this calculation by hand (not using Jamovi). You must show your work to get full credit.

Mean = 1.75; SD = 1.664

(0 – 1.75)/1.664 = - 1.05

(b) What would be the value for `childs`, if the corresponding z-score is 1.35? Perform this calculation by hand (not using Jamovi). You must show your work to get full credit.

1.35*1.664 + 1.75 = 3.996 or 4

4. (25 points) As used in Question 1, the variables `dem10fut` and `dem10pst` measure people's assessment to "How well will democracy work in America in ten years" and "How well did democracy work in America ten years ago" (0=very poorly to 10=very well; both treated as Numeric variables). Perform a regression analysis that uses `dem10pst` to predict `dem10fut`. For interpretations, you can use the variable names directly (e.g.: "when `dem10pst` increases by 1…")

(a) Write down the regression equation.

Yhat = 0.506X + 1.723

(b) Report the value of the intercept. Interpret the value of the intercept in the context of this problem using language that someone who hasn't taken statistics would understand or indicate why it would not be meaningful to do so.

1.723

For respondents with a `dem10pst` value of 0, the predicted value of `dem10fut` is 1.723.

(c) Report the value of the slope and interpret it in the context of this problem using language that someone who hasn't taken statistics would understand or indicate why it would not be meaningful to do so.

0.506

When `dem10pst` increases by 1 point, the predicted value of `dem10fut` increases by 0.506.

(d) Find the person with ID 2 (should be in the second row if you have not rearranged the data). Calculate the predicted value of `dem10fut` for this person.

`dem10fut` = 2

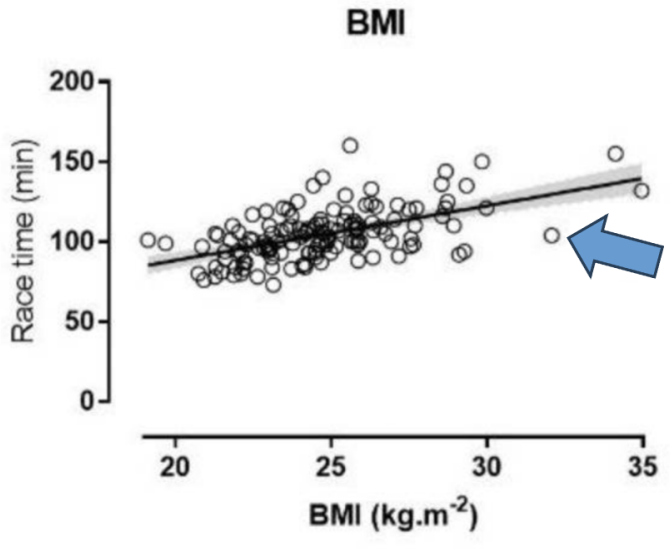
`dem10pst` = 5

Predicted `dem10fut` = 0.506(5) + 1.723 = 4.253

(e) Does the model overpredict or underpredict the value of `dem10fut` in (d)? Calculate the residual.

2 - 4.253 = -2.253; overpredict

5. (20 points) You do not need Jamovi for this question. The following plot is from an article "Predictors of half-marathon performance in male recreational athletes." It shows the correlation between half-marathon race time (in minutes) and body mass index (`BMI`). The regression line is shown on the graph.



(a) For a male recreational runner with a `BMI` of 25, what is the predicted `race time`? Give a rough estimate from the graph and indicate on the graph how this estimate is obtained.

Around 100

(b) There is a data point indicated by an arrow. Does the regression model over- or under-predict for this data point? Briefly explain.

Overpredict; any reasonable explanation is fine.

(c) The slope of the regression line is 3.4 (given in the article). Provide an interpretation of this value in the context.

When a person's `BMI` increases by 1, the predicted race time increases by 3.4 minutes.

(d) Which one of the following is true regarding the intercept of the regression line? Choose one.

- i. Not interpretable since `race time` cannot be 0.
- ii. Not interpretable since `BMI` cannot be 0.
- iii. It is a value between 50 and 100.
- iv. It can be interpreted as the predicted `race time`.

ii. Not interpretable since `BMI` cannot be 0.

6. (12 points) You do not need Jamovi for this question. The following information are from the same article quoted in Question 5.

Few research has been conducted on predictors of recreational runners' performance, especially in half-marathon running. The purpose of our study was (a) to investigate the relationship of `half-marathon race time` with training, anthropometry and physiological characteristics, and (b) to develop a formula to predict `half-marathon race time` in male recreational runners.

Half-marathon race time correlated with physiological, anthropometric and training characteristics, with the faster the runner, the better the score in these characteristics (e.g., `VO2max`, r = - 0.59; `BMI`, r = 0.55; `weekly running distance`, r = -0.53, $p<0.001$). `Half-marathon race time` was related to and could be predicted by `BMI`, `VO2max` and `weekly running distance`. Based on these relationships, a prediction formula for `race time` was developed providing a practical tool for recreational runners and professionals working with them.

(a) Which type of runners tend to be faster (i.e., lower `half-marathon race time`)? Choose one from each of the blanks:

Runners with (higher / lower) `VO2max`; runners with (higher / lower) `BMI`; runners with (higher / lower) `weekly running distance`.

Briefly explain how you can tell.

Higher, lower, higher; based on the signs of the correlation coefficient.

(b) The article says "`Half-marathon race time` was related to and could be predicted by `BMI`, `VO2max` and `weekly running distance`." Which of the following statement is true? Choose one.

- i. `BMI`, `VO2max`, and `weekly running distance` are the dependent variables.
- ii. `Half-marathon race time` is the dependent variable.
- iii. The prediction formula must be based on a simple linear regression.
- iv. When predicting `half-marathon race time`, residual values can only be positive.

ii. `Half-marathon race time` is the dependent variable.

(c) Based on these findings, can we conclude that a change in `BMI` will directly cause a change in running performance? Briefly explain.

No, correlation does not imply causation; any reasonable explanation in the context is fine. E.g., there could be other confounding variables that affect both `BMI` and running performance, such as age or gender; maybe younger runners tend to have a lower BMI and also better running performance, but this doesn't mean that a lower BMI directly causes better performance.