**The final is due Thursday, Dec. 14th by 12:00 PM (PST) on Canvas.**

Instructions:

* This exam includes 7 questions and 2 extra credit questions and is worth 120 (out of 100) points.
* For essay and short answer questions, answers should be typed. You are encouraged to use bullet points to organize your ideas as you answer questions (it also makes it easier to grade!).
* All answers should be reported below the question. **To make it easier for the TAs to grade, please type all answers in blue.**
* For questions requiring computer work, answers should either be typed or be copy/pasted from the **edited, relevant output** (i.e., include only what is needed and highlight your answer; please do not include unedited or irrelevant output). You may include your code at **the very end** of the final if you would like, but it is not required.
* You may include a photo of any work done by hand, but please try to make your answers legible. Please make any answers/solutions easy for the TA’s to find (if we have to guess what your answer is from a sheet of hand-written work, we may not find the answer you intended to submit).
* This exam gives you the opportunity to work independently. You may consult with the TA’s or Emilio concerning the final, but not with anyone else. You can also consult your book, class notes, and previous labs and homework.
* You will not be able to submit corrections for the final.

**1. (30 points) Imagine that you are interested in examining the relationship between mood and weather. You ask 4 people to fill out a questionnaire about their mood for 70 consecutive days and also record the maximum temperature for each day. The data for weather and mood from 4 individuals are in the “tempmood.csv” data set.**

(a) Plot the data in a way that you find meaningful to illustrate the relationship between mood and weather. Comment on your plot. You may consider changing your plot after you complete the rest of the questions, to best illustrate your research findings. [Note: the plot deemed “the best” (e.g., most informative, most interesting, most understandable, etc.) by your teaching team will be awarded bonus points! Have fun! [5 pts]

This plot shows the regression line of mood (y-axis) predicted by temperature (x-axis) for each participant, along with the individual data points. It shows that for some participants, the slope was close to a constant while for others it showed a positive or negative trend.

A graph showing different colored lines

Description automatically generated

(b) Compute means and standard deviations for each of the variables in the data set. [ 4 pts]

|  |  |  |
| --- | --- | --- |
|  | Mean | SD |
| Temp | 80.0 | 3.0 |
| Part1 | 3.0 | 0.8 |
| Part2 | 4.3 | 1.2 |
| Part3 | 5.1 | 1.8 |
| Part4 | 5.9 | 2.3 |

(c) Compute the sum of cross-products and covariance between weather and mood across all participants and report them below. Then explain what each of these indices indicate about the relation between weather and mood? [4 pts]

Sum of cross-products = -395.37.

Covariation = -1.41.

These indices indicate that there is an inverse or negative relationship between these two variables.

(d) What is the correlation between weatherand mood across all participants? What does the correlation indicate about the relationship between weather and mood? [2 pts]

*r* = -0.244. That means that there is a weak negative relationship between mood and weather.

(e) Estimate the regression equation of the line that best represents the relationship between weather and mood for all individuals as a single group. [2 pts]

(f) Estimate and report the regression equations for the lines best representing the relationship between weather and mood for each individual participant. [4 pts]

(g) Calculate and report the standardized beta-weight for the association between weather and mood (rounded to two decimal places) for each participant. Identify which participant had the strongest association between weather and mood, and which participant had the weakest association between weather and mood. [4 pts]

Participant 2 had the weakest association between mood and weather, while participant 4 had the strongest association between these two variables.

(h) What can you say about differences in the relationship between weather and mood across individual participants? [2 pts]

For Participant 1, weather and mood have a positive relationship, meaning that their mood moves in the same direction as the temperature increases or decreases. For Participant 2, the association was not statistically different from zero, implying that there is no association between the two variables for this participant. For Participants 3 and 4, the association is negative. As the temperature increases, the mood decreases and vice-versa. Participant 4 is also the most affected by temperature since they have the strongest regression coefficient among all subjects.

(h) You submit the result of all these analyses for publication but the editor rejects the manuscript on the basis of: (i) a lack of power to examine your research questions, and (ii) the fact that there are only 4 individuals in your data set and, thus – they claim – you cannot generalize to the population. Nevertheless, you are convinced – or just have a hunch – that there might be something valuable here and write back arguing that the data and analyses are worth disseminating. What would you say to support your argument? [3 pts]

I acknowledge the editor’s concern about the lack of power since a sample of four may not be sufficient to detect most effects in psychological research. However, it can be argued that because we found a statistically significant effect of temperature on mood, we suppose that the effect size of this relationship is very strong, making it detectable even with a sample as low as four.

**2. (10 points) The following matrices are the covariance and correlation matrix, respectively, of variables *X*1, *X*2, *X*3, *X*4, and *X*5.Using the information provided in the matrices, fill in the gray boxes with the appropriate values.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **X1** | **X2** | **X3** | **X4** | **X5** |
|  | **X1** |  | - | - | - | - |
|  | **X2** | -0.5 | 0.25 | - | - | - |
| **Covariance** | **X3** | 1.8 |  |  | - | - |
|  | **X4** | -1.08 | -0.135 | -2.43 | 7.29 | - |
|  | **X5** | 6.48 |  | 6.48 | 4.374 |  |
|  |  |  |  |  |  |  |
|  |  | **X1** | **X2** | **X3** | **X4** | **X5** |
|  | **X1** | 1.0 | - | - | - | - |
| **Correlation** | **X2** | -0.5 | 1.0 | - | - | - |
|  | **X3** |  | 0.25 | 1.0 | - | - |
|  | **X4** |  |  | -0.3 | 1.0 | - |
|  | **X5** |  | 0.5 |  | 0.3 | 1.0 |

**3. (20 points) Researchers were interested in the role of extracurricular activities (sports: 0 = other extracurricular activities; 1 = participation in sports) and biological sex (female: 0 = male; 1 = female) on standard normal score of adolescent perceptions of social acceptance (PSA).**

**The data can be found in the “socialacceptance.csv” file. Determine whether factors of extracurricular activity type and biological sex are associated with adolescent PSA.**

a) State the type of design of the study. [2 pts]

The study is a 2x2 factorial design.

b) Test whether there are group differences in adolescent PSA based on extracurricular activity, biological sex, and their interaction (use Type II SS). If there is a significant interaction effect, be sure to conduct appropriate follow-up analyses and report their outcomes. Write a report (no longer than a page) in which you report your findings as you would in a journal article (i.e., include text, table, and figure). [18 pts]

A two-way factorial ANOVA with Social Acceptance as the dependent variable showed a significant interaction between biological sex and extracurricular activities, *F*(1, 196) = 6.15, *p* = 0.014. Simple effects analyses showed that among males, those that did other extracurricular activities had statistically significant lower social acceptance than those that participated in sports, *F*(1, 100) = 5.41, *p* = 0.02. A similar effect was observed among females. The females who did sports had statistically significant higher social acceptance than the females who did other extracurricular activities, *F*(1, 96) = 28.5, *p* < 0.001.

Table 1. ANOVA table of main effects and interaction between sex and extracurricular activity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SS | *df* | *F* | Sig. |
| Sports = Yes | 25.42 | 1 | 30.33 | <0.001 |
| Sex = Female | 3.58 | 1 | 4.28 | 0.040 |
| Sports\*Sex | 5.15 | 1 | 6.15 | 0.014 |
| Within | 164.24 | 196 |  |  |
| Total | 198.40 | 199 |  |  |

Table 2. Simple effects of sex and extracurricular activity on social acceptance.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | SS | *df* | *F* | Sig. |
| Group: Female |  |  |  |  |
| Sports = yes | 26.52 | 1 | 28.50 | <0.001 |
| Within | 89.37 | 96 |  |  |
| Group: Male |  |  |  |  |
| Sports = yes | 4.05 | 1 | 5.41 | 0.022 |
| Within | 74.87 | 100 |  |  |

A graph with different colored lines

Description automatically generated

**4. (10 points) Explain why *r* must be between -1 and +1. Please, do not use more than 1 or 2 paragraphs. You can append calculations, if you need them.**

The r coefficient is bounded between -1 and +1 because it is the scaled covariance between two variables in terms of their standard deviations. Therefore, both variances will be on the same scale. Given that , when |x| = |y| (i.e., a perfect relationship), , and , resulting in .

**5. (20 points) Say you run a simple regression with predictor variable and outcome variable .** **You fit the following model:**

a) What is the interpretation of ? What is the interpretation of ? **[4 points]**

represents the intercept of the regression line, that is, the predicted value of Y when is equal to zero.

is the slope of the regression line. It represents the change in Y for a one-unit increase in .

b) When will be equal to the correlation between and ? Why? **[4 points]**

The coefficientwill be equal to the correlation between Y and when these two variables are standardized. When that happens both variables will have a mean of 0 and a standard deviation of 1, implying that . Therefore,

**After running the analysis, you remember a covariate that you believe is related to but is not substantively of interest.**

c) What are the benefits of including the covariate in the model? Include two benefits and explain in them in detail. **[4 points]**

Including a covariate can help control for confounding variables that might otherwise distort the estimation of the effect of interest. If the covariate is related to both X and Y, by including it in the model we can statistically control for its influence on both variables, isolating the effect of X on Y. Another advantage of including a covariate is obtaining more precise estimates for the predictors. When the covariate is associated with the outcome variable, including it in the model helps to explain some of the variation in the outcome. It reduces the residual variance, leading to more precise estimates of the effect of X on Y.

**Finally, you include the covariate within the analysis and fit the following model:**

d) What is the interpretation of each of the coefficients in this model? **[4 points]**

is intercept of the regression line and represents the predicted value of Y when both and are equal to zero.

now represents the change in Y for a one-unit increase in , holding constant. It informs us the expected change in Y when increases by one unit, regardless of the level of .

can be interpreted similarly as . It is the expected change in Y when increases by one unit, regardless of the level of . That is, the effect of after controlling for the effect of .

e) When will be equal to the correlation between and ? When will be equal to the correlation between and ? Why? **[4 points]**

Similar to the answer in (b), one of the conditions is that all variables are in standardized form. However, given that now these are semi-partial regression coefficients, a second condition is that and are orthogonal. Orthogonality ensures that the effects of each predictor are independent and do not overlap. The terms of the equation representing the correlation between the two predictors will be zero. Then, the equation will be reduced to the correlation between the response variable and a single predictor. For instance,

**6. (10 points) Suppose you are hired to serve as a statistical consultant. In each of the following cases, what advice would you give to your client concerning the procedures and/or conclusions he or she has drawn, or about the kind of statistical techniques most suitable? Be sure to briefly explain the reasoning underlying your advice.**

(a) A researcher studies the effects of education (HS or less, Some College, 4 Year College Degree, Graduate/Professional Degree) on income by randomly calling 5,000 participants in the United States. At a presentation of his results, several colleagues suggest that effects of education on income may not be robust when considering other predictors such as work experience, time with their current employer, age, personal investments, etc. What sort of analysis did the researcher most likely conduct, and how can the researcher address these criticisms of his research? **[2]**

The researcher likely used a one-way ANOVA to evaluate how a categorical variable (education level) and a continuous variable (income) are related. However, the colleagues have pointed out a potential issue with confounding variables. Factors such as age, work experience, time with current employer, and personal investments are all possible factors that may affect both education level and income. The researcher should consider including these variables as covariates in a multiple regression analysis/ANCOVA to address this issue. This will help isolate the effect of education on income while controlling for the influence of other variables.

(b) A researcher is interested in predicting the mental health (mentally stable versus mentally unstable) of college students based on their reported level of stress. What kind of sample should she collect, and what statistical technique would be best to address her research question? **[2 ]**

Ideally, a random sample of college students to ensure generalizability, and to reflect the proportion of the two conditions in the population of interest. Since the outcome variable is binary, the researcher should consider a logistic regression analysis.

(c) A researcher collected data from undergraduate and graduate students at universities across the country in a study of the relation between age (range: 18 – 46 years; *M* = 23.5) and openness to experience. The researcher found a significant, negative relation between age and openness to experience (*r*(2,998) = -0.13, *p* < .05). She used this finding to argue that as people age, their openness to new experiences decreases, and that this explains why elderly individuals (aged 60 years and above) have difficulty in learning about novel technology and ideological shifts. Is this a reasonable conclusion? Why or why not? **[2 points]**

The researcher’s conclusion cannot be defended for several reasons. Firstly, the study was conducted only on undergraduate and graduate students, representing a narrow age range of 18-46 years. This narrows the applicability of the findings to the overall population, especially the elderly above 60 years old, as the conclusions suggest. Secondly, the researcher found a relatively small correlation coefficient of -0.13, indicating a weak negative relationship. This doesn't necessarily imply a linear decrease in openness to experience with increasing age. Thirdly, the study didn't account for other factors that might influence age and openness to experience.

(d) A researcher studied a group of 100 students by having them complete a survey once a quarter, every quarter, for two years via an online survey form. The survey consisted of several items meant to measure anxiety, self-competence, and academic performance. What method(s) of analysis would be applicable to this type of data? Justify your recommendations. **[2 points]**

Given that time points are categorized in quarters, one option is analyzing it with a Repeated-measures ANOVA. It is an appropriate choice if the outcome is measured on a continuous scale. Another option is using Latent growth curve analysis. This option becomes particularly relevant because the variables assessed are latent and indirectly measured from the observed items. Therefore, the researcher could model the individual growth trajectories of the latent variables over time.

(e) A researcher received a small grant to conduct a study and is debating on how to spend the money. Her options are to (1) give a test to 300 individuals on one occasion; (2) give a test to one individual on 300 occasions; (3) give a test to 30 individuals on 10 occasions; (4) give a test to 10 individuals on 30 occasions; or (5) any combination of the above. What factors would you need to consider in consider in your recommendation of which data collection methods should she should use and why? **[2 points]**

A first consideration is what research question she is trying to understand. These designs can refer, for instance, to description, testing individual differences, group-level trends, or changes within individuals over time. Another important factor is considering the nature of the analysis, whether longitudinal or cross-sectional, favoring some options over others. I would also have to ask the researcher what effect size she expects to find since this may imply different requirements for sample size and frequency of measurement. Considering the expected generalizability of her findings is also important, given that it is also relevant to deciding on a sample size.

**7. (20 points) Imagine you were hired to help a school board analyze data. The school board created different promotional videos to get middle schoolers interested in learning to play an instrument. One of the videos promoted the school’s marching band, and a second video promoted the school’s orchestra. The school board planned to randomly assign children to view one of these two videos, but after consulting with a researcher, the school board agreed to include a control group in the study. Hence, children were randomly assigned to view the marching band video, the orchestra video, or the generic educational video. After children viewed the videos, they took a brief survey to assess their interest in playing a musical instrument. The means and SD are given in the table below.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Educational Video | Marching Band Video | Orchestra Video | Total |
| Mean | 5.2 | 22.17 | 30.8 | 19.39 |
| SD | 7.2 | 7.2 | 7.2 | 12.11 |
| n | 6 | 6 | 6 | 18 |
| *ssB = 2035.6* |  |  |  |  |

(a) If you were to use orthogonal contrasts to test for differences between the means, what is the best way that you could assign contrast weights to each group to assess whether the promotional videos were generally effective in increasing children’s interest in playing an instrument, as well as whether one video was more effective than the other? Fill in the table below with these weights, and explain why these contrast weights are more appropriate than other weights. **[6 points]**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Orchestra | Marching Band | Educational |
|  | 1/2 | 1/2 | -1 |
|  | -1 | 1 | 0 |

The contrast compares the average interest of children who saw any promotional video to the control group. The contrast examines whether there were differences in effectiveness between the orchestra and the marching band videos. These contrast weights are appropriate because they meet the requirements for orthogonality: the rows sum to zero, and the sum of the product of the columns is also zero.

(b) Using the means in the table above, perform these contrasts. Report and interpret the results of the orthogonal contrasts, including the t-value associated with each contrast (use a two-tail alpha criterion of .05). Then, summarize your conclusions as if you were writing to the school board, and make a recommendation about which video(s) to continue showing students. **[7 points]**

The significance test of the first contrast indicates that it is statistically different from zero, meaning that there is a difference in the average interest of children who saw any promotional video compared to the control group. On the other hand, the second contrast was not statistically significant. This tells us that there were no differences in interest in playing an instrument between the group assigned to the marching band video and the group assigned to the orchestra video. In conclusion, to spark interest in students to play an instrument, the school board would better show them a promotional video (either orchestra or marching band) rather than a generic educational one.

(c) Explain why contrasts weights must be chosen with care to test specific hypotheses. To illustrate your point, repeat the orthogonal contrasts using a different set of orthogonal contrast weights, and explain how your conclusions would change if you had reported the results of these contrasts to the school board. Based on these changes, make an argument for why erroneous conclusions might be drawn if inappropriate contrast weights are use. **[7 points]**

Choosing inappropriate weights can lead to testing a different hypothesis than intended, yielding misleading conclusions. For instance, had we chosen the weights (-0.5, 0.5, 0) for  we would still have orthogonal contrasts since the sum of the product of the weights sums to zero: . However, we would have obtained a t value of -4.15, which would be considered statistically significant with a two-tail alpha criterion of .05. That would have led us to wrongly conclude that there was a difference between the two groups when, in fact, there is not.

**Extra Credit**

**8. (10 points) Explain what it means to say that a correlation is a covariance expressed in *z*-scores? Derive numerically the formula for a correlation based on the formula from a covariance (and describe the steps in your own words).**

The correlation scales the covariance by the standard deviations of both variables. It is comparable to a z-standardization since it is a process to scale the data around a mean of zero and a standard deviation of one.

The covariance can be expressed as,

If we rescale x and y to z-scores, we obtain,

Then, the equation for the covariance will converge to the correlation coefficient,

**9. (10 points) Imagine that you are hired by the superintendent of a local school district to serve as the statistical consultant on a project examining children’s acquisition of mathematics skills over the course of a school year. Students were measured at 5 times throughout the school year and approximately 25% of the data are missing due to students being absent on testing days. The superintendent lets you know that children tend to start at much different levels of math ability at the start of the school year, and that some students progress more rapidly than others in increasing their ability, and some even decline. The superintendent wants you to use repeated measures ANOVA to evaluate whether the trajectories of students with mothers who completed college differ from students whose mothers did not earn a college degree. What advice would you give the superintendent (e.g., appropriateness of RM ANOVA to answer research questions, assumptions of RM ANOVA, etc.)? Explain your answer in a paragraph or two.**

While RM ANOVA can analyze repeated measures, it assumes a linear trajectory of change over time. The superintendent mentioned varying initial levels and non-uniform progress, which a simple linear model might not capture. RM ANOVA also excludes observations with missing values, meaning 25% of the students would be dropped from the analysis. Another limitation is that RM ANOVA assumes equal variances and covariances across time points. This might not hold if students progress at different rates.