Intermediate Statistics Final Assignment

- Due **DECEMBER 6** by end of day.
- Marked out of 60 points.

- Late submissions ***cannot*** be accepted. Final grades need to be submitted shortly after your deadline. Exceptions are only possible for severe and **documented** scenarios.

1. Comparing Multiple Groups (10 points)

a. Use the Titanic data set to investigate whether a passenger's country of residence affected the cost of their ticket. First, run your *usual* significance test without checking assumptions. Briefly report your overall findings, including post-hoc comparisons and what they found.

b. Then, examine the data for normality. If deciding this assumption has been violated, explain how you know this and then run the appropriate non-parametric significance test. Post-hoc analysis is **not** required here.

c. Formally summarize your findings (statistical and research), indicating which non-parametric test you used to obtain your results. Refer to your Evans text for examples of non-parametric reporting.

2. Bivariate Relationships (10 points)

a. Use Draw my Data (http://www.robertgrantstats.co.uk/drawmydata.html) to create a near perfect, negative linear relationship with a minimum sample size of 20. Copy-and-paste or import your data into SPSS. Report an appropriate visualization and a correlation significance test. Briefly summarize your findings, inventing research variables of your choice (and that make sense with the min and max values you provided).

b. Now, use Draw my Data to create a non-linear relationship that has a medium to strong correlation coefficient (*r* **>** .3, positive or negative) with a minimum sample size of 20. Import this data to SPSS and produce a correlation significance test **without** any accompanying visualization. Briefly summarize your findings, again inventing research variables of your choice.

c. Besides the fact that your data are artificial, can you explain what else is misleading about your results in part b) of this question?

3. Model Selection (15 points)

The “substanceAbuse” SPSS file contains data from 120 participants who were in a substance abuse rehabilitation program at least five years ago. Upon participants’ initial admission into the program, they were psychologically evaluated, including their depression rates, daily stress levels, amount and perception of social support, and propensity for substance abuse. We also have information about the number of relapses participants experienced since their initial admission.

- We want to investigate how well these variables predict the number of relapses.

a. Conduct a regression model using all the available predictors. Report your Model Summary, ANOVA, and Coefficients tables. What is the strongest predictor of relapses? Which is the weakest predictor variable?

b. Now, start over the process but this time conduct a stepwise selection method for choosing the best multiple regression model. Report your Model Summary table that includes R^2 Change, as well as your ANOVA and Coefficients tables. By how much did R^2 change from the initial model (i.e., Model 1) to the final model selected?

c. In your final model in **part b)**, which predictors were retained? Comparing your model from **part a)** of this question to your final model of the stepwise selection, is your strongest predictor still the same variable? If so, has its relationship with the outcome changed at all?

4. Model Assumptions (10 points)

a. Using the final regression model recovered from the stepwise selection procedure from the previous question (**Question 3, b.**), run a new regression model with just those predictors. (Remember to set Method back to "Enter."). But this time investigate your model for the following assumptions:

   i. linearity,

   ii. homoscedasticity,

   iii. normality of the residuals,

   iv. and multicollinearity.

b. If these assumptions have been satisfied, formally report the findings of your multiple regression model, indicating both overall model statistics and specific predictor statistics. If you think any of the specified assumptions have been violated, also include in your summary of your results a warning that your data may violate those specific assumptions.

5. Measurement Reliability (10 points)

Use the LAMP data set and codebook to investigate the reliability of the Brink Depression Scale (BDS).

a. First determine the BDS' split-half reliability. Report your Reliability Statistics table, taking note of your split-half reliability estimate.

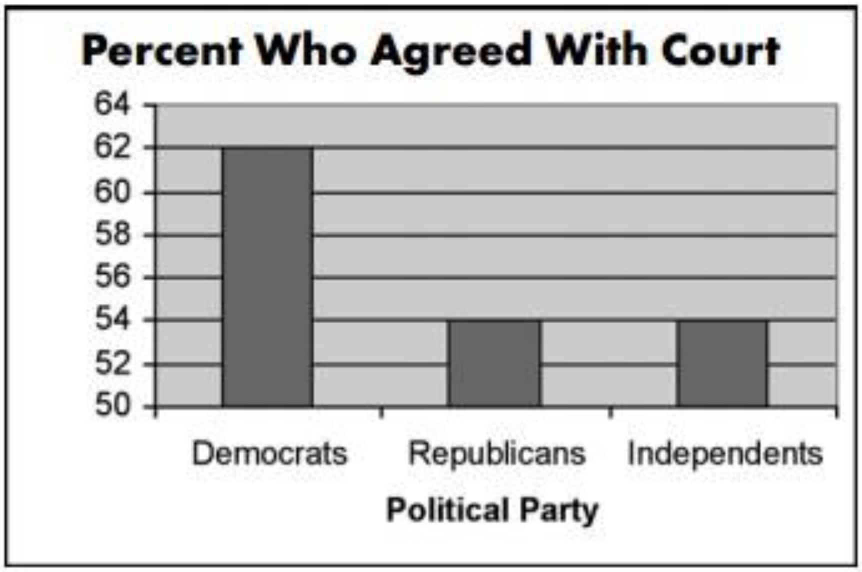
b. Now, investigate the internal consistency of the BDS using coefficient alpha. Report Reliability Statistics and Item-Total Statistics tables.

c. Formally summarize your findings, including both reliability estimates (split-half and coefficient alpha) and how many items are on the scale.

d. Do you think any items should be dropped to further improve reliability? If so, explain why. Re-run your reliability analysis without the item(s) you chose to exclude.

6. Visualization (5 points)

Below is a bar graph from an American news outlet explaining people's differences of opinion on the ultimate decision in a controversial court case. People have been sorted into political party ideology, indicating within each party how many agree with the court decision.



a. Is there anything potentially misleading about the bar graph as shown? Fully explain what you mean.

b. What change(s) would you make to this graph? Explain if you would use an alternative type of graph to represent the same information. (You do *not* have to limit yourself to only visualizations discussed in this course.)