Master’s Comprehensive/Doctoral Preliminary Examination
Measurement, Statistics and Evaluation Program
Department of Human Development and Quantitative Methodology
University of Maryland, College Park

Spring 2016, February 12

Declarative Knowledge Component: 9a.m. – 11a.m.
Question 1

a. Draw a completely labeled interaction diagram (plot) for samples in a 2×3 factorial design (A has 2 levels; B has 3 levels) with between-group variability due to A mean differences and an A×B interaction, but none due to B differences. (There are many possible correct answers to this.) Please be as precise as possible in your drawing.

![Diagram for question 1a](image)

b. Draw a completely labeled interaction diagram (plot) for samples in a 2×3 factorial design (A has 2 levels; B has 3 levels) with between-group variability due to A mean differences and B mean differences, but none due to any A×B interaction. (There are many possible correct answers to this.) Please be as precise as possible in your drawing.

![Diagram for question 1b](image)

c. Explain how you would code and test this interaction in the multiple regression / general linear model

There are many answers to this. One could dummy code the A main effect as 0 (A1) and 1 (A2). The B main effect could in turn be coded with two dummy variables – B1 B2, where B1 membership is 1 0, respectively; B2 membership is 0 1, respectively; and B3 membership is 0 0, respectively. The interaction would take two dummy variables, the product A*B1 and the product A*B2.

To assess the significance of the interaction, enter the three main effect variables (A, B1, B2) as a block in a multiple regression model; then enter the two interaction products as a block. The significance of the change in R² will provide the test of the interaction.
Question 2

Edited text-type output for a logistic regression (with two predictors entered simultaneously) appears below. The dependent variable Y has two categories, each with 24 observed cases (i.e., 48 cases total).

<table>
<thead>
<tr>
<th>Variable(s) Entered on Step Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1..</td>
</tr>
<tr>
<td>X1</td>
</tr>
<tr>
<td>X2</td>
</tr>
</tbody>
</table>

| -2 Log Likelihood | 52.669 |
| Goodness of Fit   | 46.838 |
| Cox & Snell - R^2 | .251   |
| Nagelkerke - R^2  | .335   |

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>df</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>13.873</td>
<td>2</td>
</tr>
</tbody>
</table>

------------------- Variables in the Equation ---------------------

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>-.9959</td>
<td>.3449</td>
<td>8.3357</td>
<td>1</td>
<td>.0039</td>
<td>.3694</td>
</tr>
<tr>
<td>X2</td>
<td>.1093</td>
<td>.0611</td>
<td>3.1973</td>
<td>1</td>
<td>.0738</td>
<td>1.1155</td>
</tr>
<tr>
<td>Constant</td>
<td>2.6248</td>
<td>1.7306</td>
<td>2.3005</td>
<td>1</td>
<td>.1293</td>
<td></td>
</tr>
</tbody>
</table>

a. Write three full fitted equations – one for predicting the log(odds), one for predicting the odds, and one for predicting the probability of the outcome of interest.

\[
\text{logit}(p)' = b_0 + b_1X_1 + b_2X_2
\]

\[
\text{odds}' = \exp(b_0 + b_1X_1 + b_2X_2)
\]

\[
p' = \frac{1}{1 + \exp[-(b_0 + b_1X_1 + b_2X_2)]}
\]

b. Provide a complete interpretation of the regression weight (both B and Exp(B)) for X1.

B: For a one unit increase in X1 we expect a .9959 decrease in the logit(p), holding all else constant.

Exp(B): For a one unit increase in X1 we expect the odds of an outcome of 1 to become .3694 times what they were, holding all else constant. That is about a 63% reduction in odds with each increase of 1 unit in X.
c. Assuming \( Y=0 \) represents a failure and \( Y=1 \) represents a success, would you predict success or failure for someone with \( X_1=3.7 \) and \( X_2=14.1 \)? Provide predicted log(odds), odds, and probability to explain your answer.

\[
\text{logit}(p) = 2.6248 - .9959X_1 + .1093X_2 \\
= 2.6248 - .9959(3.7) + .1093(14.1) \\
= .4811 \quad \text{predict failure since > 0}
\]

\[
\text{odds} = \exp(.4811) \\
= 1.618 \quad \text{predict success since > 1}
\]

\[
p = \frac{1}{1+ \exp[-(.4811)]]} \\
= .618 \quad \text{predict success since > .5}
\]
Question 3

a. What is standard error of measurement? Define the term both conceptually and mathematically using key concepts and notations in classical test theory (CTT).

The standard error of the measurement (SEM) is the standard deviation of an examinee’s observed scores to parallel test forms around his/her true score. It is used to gauge the precision of reported observed scores, and as such is integral to constructing confidence intervals and test statistics for inferences about each examinee’s true score.

Mathematically SEM is calculated as follows.

\[ S_E = S_X \sqrt{1 - r_{XX}} \]

where S stands for sample standard deviation and X and E are observed scores and error scores, respectively. \( r_{XX} \) is the reliability of scores that is defined as the correlation between two scores X and X’ obtained from parallel tests.

b. Explain how standard error of measurement is related to reliability and validity of scores.

As it can be seen from the equation above, lower reliability of test scores indicates larger magnitude of standard error of measurement and vice versa. Validity of test scores cannot be assured without achieving a certain level of reliability. Accordingly, larger standard error of measurement will harm validity of test scores too.

c. What is standard error of measurement in item response theory (IRT)? Explain the commons and differences in defining and calculating standard error of measurement between CTT and IRT.

In IRT, the standard error of measurement is defined as \( \frac{1}{\sqrt{I}} \), where I stands for information function. Therefore, standard error of measurement in IRT is also a function rather than a single number. The standard error of measurement of an examinee’s score in both theories describes the precision of scores. While the standard error of measurement of an examinee’s score in CTT is a constant across levels of observed scores, the standard errors of measurement in IRT differ across true latent trait levels.
Question 4

Shadish, Cook, and Campbell (2002) introduced four general types of validity regarding inference from studies in the social sciences.

a. Name and define these four validity types.

*Statistical Conclusion Validity*: The validity of inferences about the correlation between treatment and outcome.

*Internal Validity*: The validity of inferences about whether observed covariation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured.

*Construct Validity*: The validity of inferences about the higher order constructs that represent sampling particulars.

*External Validity*: The validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables.

b. Briefly describe (in 2-3 sentences) an example for each of the four where the validity has been threatened.

Each example can be one of the lists for each category as follows.

*Threats to Statistical Conclusion Validity*: Low statistical power, Violated assumptions of statistical tests, Fishing and the error rate problem, Unreliability of measures, Restriction of range, Unreliability of treatment implementation, Extraneous variance in the experimental setting, Heterogeneity of units, inaccurate effect size estimation

*Threats to Internal Validity*: Ambiguous temporal precedence, Selection, History, Maturation, Regression, Attrition, Testing, Instrumentation, Additive and interactive effects of threats to internal validity

*Threats to Construct Validity*: Inadequate explication of constructs, Construct confounding, Mono-Operation Bias, Mono-Method Bias, Confounding constructs with levels of constructs, Treatment sensitive factorial structure, Reactive self-report changes, Reactivity to the experimental situation, Experimenter expectancies, Novelty and disruption effects, Compensatory equalization, Compensatory rivalry, Resentful demoralization, Treatment diffusion
Threats to External Validity: Interaction of the causal relationship with unit, Interaction of the causal relationship over treatment variations, Interaction of the causal relationship with outcomes, Interactions of the causal relationships with settings, Context-dependent mediation

c. Rubin’s Causal Model is directly relevant to one of the types of validity. Identify this validity and briefly explain the connection.

Internal Validity. Rubin’s Causal Model is a theoretical and statistical framework that tries to draw causal inferences from not only experimental studies but also quasi-experimental studies or observed studies by using the concept of potential outcome.

Directions

Students will have 24 hours to take the procedural knowledge exam component. This part of the exam begins at 3pm on Friday, February 12, 2016 and student solutions are due by email to Jannitta Graham (jgraham7@umd.edu), no later than 3pm on Saturday, February 13, 2016; late submissions may count as a failure. For this component, students are allowed to use any resource at their disposal (e.g., books, internet) with the exception of discussing the exam with any other individual including, but not limited to, other graduate students or faculty members at any institute. All work on this component should be completely your own.

Please follow the writing guidelines outlined on pp. 2-3 of the Master’s Comprehensive/Doctoral Preliminary Examination Policies and Procedures document found at:


We wish you all the best on this component of the exam and are looking forward to receiving your responses. If you require any clarification, on matters of procedure or on the questions themselves, please contact Dr. Yang immediately (jyang@umd.edu). Clarification questions and responses will be shared with other examinees for fairness.

The 2015-2016 Examination Committee
A randomized pretest-posttest design was implemented and an experiment was conducted that examined three different interventions to decrease depression. Specifically, 30 individuals diagnosed with depression were randomly assigned to one of three conditions: (1) selective serotonin reuptake inhibitor (SSRI) antidepressant medication, (2) placebo, and (3) wait list control. The Beck Depression Inventory (BDI) was administered to each individual prior to the study, and then later, was administered a second time at the end of the study. The data are displayed in Table 1 below. A lower BDI score indicates a lower level of depression.

Table 1

<table>
<thead>
<tr>
<th>SSRI</th>
<th>Placebo</th>
<th>Wait List</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>Post</td>
<td>Pre</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>11</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>25</td>
<td>15</td>
<td>11</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>22</td>
</tr>
</tbody>
</table>

Your tasks:

a. Analyze the presented data using three analytic methods: (i) ANOVA on the posttest scores only, (ii) ANOVA on gain scores (gain scores = posttest scores – pretest scores), (iii) ANCOVA using the pretest score as a covariate. Present the necessary output, and provide syntax or code as an appendix.

b. What do the results of comparing the three different approaches suggest about the relative merits of ANCOVA, ANOVA on the posttest scores only, and ANOVA on gain scores for analyzing data from a randomized pretest-posttest design?

c. Move forward using the ANCOVA approach. For the statistical tests to be valid, a minimal number of assumptions must be met concerning elements of the ANCOVA model. State and test these assumptions. If they are approximately satisfied then provide evidence supporting this claim. If one or more assumptions are not approximately satisfied, then implement a remedy and rerun the analysis. What assumptions must be made in order to assume estimates from a natural experiment are unbiased causal estimates? Do these assumptions apply to the design implemented here?
d. Write up the results of the ANCOVA analysis based on part (c). Do not forget the computation of effect sizes and confidence intervals of parameter estimates of interest in your write-up.

e. A colleague would like to apply an ANCOVA to a similar pretest-posttest experimental design as the one above but would like to have some idea of the number of individuals (sample size) to randomly assign to groups and has turned to you for help. Your colleague provides you with the following information about the design: (1) would like to have power of 0.90, (2) there are four experimental conditions, (3) the correlation between the pretest and posttest scores is $r = 0.55$, and (4) only a small effect is expected. Perform a power/sample size analysis for your colleague and write up the results.
Question 2. EDMS 623

The following question uses data found in a space-delimited file, item.dat. The data file contains responses from $N = 200$ examinees on 10 items. Each item was scored 0 (incorrect) or 1 (correct).

a. Using ideas from Classical Test Theory (CTT), perform an item analysis. Compute and comment on three components: (i) item difficulty, (ii) item discrimination, and (iii) reliability. Write up your results in a report using tables and graphs if warranted.

b. Provide a rationale for your choice of reliability coefficient. Discuss other reliability coefficients that could have been used instead of the one you chose for answering part (a).

c. Discuss in detail at least three shortcomings of using CTT. Pick one of three shortcomings you previously discussed and demonstrate/explain the impact of this shortcoming to the item analysis scenario in part (a).