EDMS Master’s Comprehensive Examination / Doctoral Preliminary Examination
Fall 2015, September 18
Morning Session: 9a.m. – 12p.m.
Statistics Question 1

The term “robust” is often used in conjunction with statistical tests. Answer the following questions related to robustness.

a. What does it mean to say that a statistical test is “robust” or that it is “not robust”? Provide an explanation of robustness in the context of statistical tests in general.

Typically, “robust” means that the technique still functions properly in its primary intent. If we are talking about a statistical test, robustness usually refers to maintaining Type I error rate to a target nominal $\alpha$ level and not unduly reducing (or increasing) power. For studies whose focus is on the estimation of a parameter, such as a regression coefficient, robustness might refer to maintaining a low level of bias in the estimation of that coefficient and possible maintaining a desired level of precision in the estimation (i.e., not too much variability of the parameter estimates).

b. Provide an example of a scenario/conditions in which a specific statistical test (of your choosing) is behaving in a robust manner. Explain thoroughly the context for your example and what it is about that statistical test that you believe is causing it to be robust in the scenario/context described.

Answers may vary. One simple example is the t-test’s robustness to nonnormality with larger sample sizes. The reason for this comes from the central limit theorem, whereby the sampling distribution of means is increasingly normal with increased sample size even if the population distribution is not.

c. Provide an example of a scenario/conditions in which a specific statistical test (of your choosing) is not behaving in a robust manner. Explain thoroughly the context for your example and what it is about that statistical test that you believe is causing it to be nonrobust in the scenario/context described.

Answers may vary. One simple example is the pooled sample t-test’s nonrobustness to heterogeneity of variance when the sample sizes are increasingly unequal. If the population with the larger variance is the one from which the larger sample is drawn, the standard error of the difference between means used in the denominator of the t-test will be too large, leading to unduly small t-values (i.e., that tend to fail to reject $H_0$). On the other hand, if the population with the larger variance is the one from which the smaller sample is drawn, the standard error of the difference between means used in the denominator of the t-test will be too small, leading to unduly large t-values (i.e., that tend to over-reject $H_0$).
Statistics Question 2

Imagine that a researcher wanted to try to establish that a new treatment was just as effective as an old treatment (i.e., had the same population mean).

a. Discuss the problems with using traditional hypothesis testing (e.g., a two-sample t-test) in designing/conducting such an investigation.

In short, these tests are designed to reject a null hypothesis of no effect rather than to reject a null hypothesis of some degree of effect. Practically speaking, this means that having insufficient power leads to an apparent conclusion of no treatment difference when, in fact, it may merely be driven by a relatively small sample size. Thus, it operates backwards and has as its default the claim we are trying to test.

b. Suggest a strategy, using hypothesis testing methods or otherwise, that the researcher might use to assist in conducting the desired investigation.

One option might be to set a threshold for practical equivalence of two treatments. For example, if a treatment doesn’t raise test scores more than 5 points, one might conclude that the treatments are equivalent for all practical purposes. Thus, one could set a null hypothesis that the population means are 5 or more points apart, and design the test to have to have sufficient power to reject this in favor of being closer than 5 points apart. In this way the default is that the population means are not equivalent, and one must get sufficient power to reject that in favor of practical equivalence.
**Section 1: Statistics**

**Statistics Question 3**

a. What is an interaction between two variables, A and B? What does it mean for them to interact?

For a given outcome Y, an interaction between A and B means that the effect of A on Y differs for the different levels of B (or vice versa). As an example, perhaps kids’ achievement (Y) is influenced by their motivation (A), but the extent to which their motivation has an influence on their achievement dependent on their simultaneous level of effort (B). Motivation may not be effective except in the presence of higher levels of effort.

b. Give an example, graphically or numerically, where an interaction between A and B impacts the inference one draws about the effect of A and the effect of B on some outcome variable Y.

Consider an ANOVA with two levels of A and two levels of B factors. A means plot is shown below.

![Means Plot](image)

In this fairly typical example, the effect of A depends entirely on B. A1 is much more effective if B1 is present, but much less effective if B2 is present.

c. Give an example, graphically or numerically, where an interaction between A and B does not affect the inference one draws about the effect of A or the effect of B on some outcome variable Y.

Consider an ANOVA with two levels of A and two levels of B factors. A means plot is shown below.
In this fairly typical example, although interaction variability is present, A1 is always better than A2 and B2 is always better than B1. So in this sense our conclusions are uncompromised by interaction. More completely, however, our assessment of the magnitude of superiority of A1 over A2, or of B2 over B1, is affected by the interaction.

Statistics Question 4

After conducting a multiple regression with two predictors (X1, X2), a researcher had a highly significant F-value for the whole regression model, a very high multiple correlation (R), and yet the p-values associated with each individual X1 and X2 variable in the regression model output were both nonsignificant. Explain what situation(s) might give rise to this phenomenon and why that situation(s) would give rise to that phenomenon in the regression output.

The theme of the answer should be high collinearity between X1 and X2, both of which also correlate highly with Y. So you get a high R value by virtue of their relation with Y, but neither variable contributes much above and beyond the other because of their own high relation. In situations of high collinearity you can also get greatly inflated standard errors for the slopes. For these reasons the marginal contribution would not necessarily be statistically significant, especially with smaller sample sizes.
The following is an example of course grades scheme for one of the EDMS courses that reflects university level guidance.

**Course Grades**

Your assignments and exams will be combined using a weighted average grading scheme with the corresponding weights given below.

**Assessment Weight:**
- Total homework points converted to a percentage 50%
- Total midterm exam points converted to a percentage 25%
- Total final exam points converted to a percentage 25%
- Final letter grades will then be assigned based on the given scale.

**Overall Course Percent Grade:**
- 98.00% ~ 100.00%: A+
- 92.00% ~ 97.99%: A
- 88.00% ~ 91.99%: A-
- 85.00% ~ 87.99%: B+
- 82.00% ~ 84.99%: B
- 78.00% ~ 81.99%: B-
- 75.00% ~ 77.99%: C+
- 72.00% ~ 74.99%: C
- 68.00% ~ 71.99%: C-
- 65.00% ~ 67.99%: D+
- 62.00% ~ 64.99%: D
- 58.00% ~ 61.99%: D-
- ≤ 57.99%: F

**Assessment Question 1**

A naïve student asks if the final grade is norm-referenced or criterion-referenced. As an assessment expert, please provide your rationale for thinking the final grade is norm-referenced or criterion-referenced, and if there is any room for some other interpretation of the final grades based on the given information.

Given that the final grade is not determined with respect to a norm (e.g. a distribution of scores) but decided with respect to some fixed criterion, it is more reasonable to say that the final grade is criterion-referenced. However, the description itself does not provide enough information to make the final exam grade as clearly established criterion-referenced score. For example, neither specification of objectives or learning outcome to be measured nor mechanisms to assign points for each component (e.g., homework) is not provided. If any of the components is associated with the norm-referenced scoring, there is ambiguity in interpreting the final grade as fully criterion-referenced.
Assessment Question 2

Now let’s think about the validity of the final grade for this course. Define what we mean by “validity” and write a short research proposal to collect and analyze data properly to support construct, content, predictive, and consequential validity of the final grade with empirical evidence. You can assume this course is one of the EDMS core courses to facilitate your discussion (e.g., EDMS 646 or EDMS 651).

The answers can vary. However, the key is that the responses should address each aspect of validity – construct, content, predictive, and consequential validity and exhibits understanding of concepts and evidence. For construct validity related evidence, examination of underlying factor structure can be one way. For the content validity, to develop blue print for homework and exam items and get feedback from expert (e.g., face validity) can be an option. As of predictive validity, the correlation between final grades in given course and midterm/final grades in the next sequence of course can be examined. For consequential validity related evidence, one needs to collect information how the final grades of this course are used (e.g., fellowship criteria, or prerequisite for upper level courses) to make decisions if the interpretation of the final grades is well aligned with the consequential purpose of using the grades.
EDMS Master’s Comprehensive Examination / Doctoral Preliminary Examination
Fall 2015, September 18
Afternoon Session: 1p.m. – 4p.m.
Section 3: Measurement

Measurement Question 1

The Spearman-Brown “prophecy” formula is one of the methods to estimate the “internal consistency” reliability of a test.

a. Explain how the Spearman-Brown formula can be used to estimate reliability given data from a single test administration.

Divide the test into two halves (e.g., even- vs. odd-numbered items), compute scores for each “half-test,” and compute the correlation between the two sets of scores. Assuming the two half-tests are parallel, this correlation is the reliability of either test. Then use the Spearman-Brown formula to “upward-correct” this reliability estimate to obtain the reliability of the full-length test.

b. Describe two potential drawbacks to this procedure.

1: There are many different ways to divide the test in half, thus there is no unique reliability estimate.
2: The assumption that the two half-tests are parallel is very unlikely to be true in practice. In this case, Spearman-Brown may over- or underestimate the true reliability.

c. What is an alternative measure of internal consistency, and why might this be an attractive alternative?

An attractive alternative is coefficient alpha. First, alpha provides a unique estimate of reliability. Second, alpha assumes that the test items are true score equivalent (i.e., essentially tau-equivalent), which is a less strict assumption than assuming parallel items (or parallel half-tests). And even if this less strict assumption is not met, we know that alpha is an underestimate of true reliability (as long as item errors are uncorrelated). Whereas with Spearman-Brown, when the assumption of parallelism is not met, it may over- or underestimate true reliability.

d. Will a longer test always yield higher reliability? Why or why not?

No. If all items on a test are parallel, Spearman-Brown shows us that adding items to a test will always increase reliability. However, this may not be true if the scores of the new item are not highly correlated with the scores of the existing test items. In this case, the new item makes a small contribution to the true score variance (of the total test scores) and a relatively larger contribution to the error variance (of the total test scores), and we are essentially adding “noise” to the test scores. Referring to the definitional formula for reliability:

\[
\text{rel} = \frac{\text{var}(T)}{[\text{var}(T) + \text{var}(E)]}
\]

we can see that if the new item’s contribution to var(E) is much larger than its contribution to var(T), the reliability of the test may actually decrease.
Measurement Question 2

While analyzing some test data, a psychometrician computed $p$-values for each item (i.e., the proportion of correct responses) as well as separate $p$-values for women and men. She found that the difference between women’s and men’s $p$-values was statistically significant for Item 1.

a. What conclusion can the psychometrician make about this item? In other words, how can this result be interpreted?

The psychometrician can state that women and men performed differently on this item (on average), but she cannot say why. It may be because one gender is more proficient than the other, and the item is not biased. Alternatively, the item might actually be biased against men or against women. Formally, all we can say is that this item exhibits item impact. We cannot say that the item functions differently for the two groups because we have only compared the average performance of the two groups (rather than subgroups that are matched in terms of proficiency.)

b. Suppose the psychometrician wanted to compute the standardized $p$-difference (comparing women and men) for Item 1. (As a reminder, the standardized $p$-difference is a popular method to detect differential item functioning.) What additional information would she need in order to compute this statistic?

She must know how many women and how many men are in each possible score category (based on the total test score). And in each category, she must know the proportion of correct responses for each gender.

So she must compute the difference between the $p$-values for each score category, and then she must compute a weighted sum of these differences (where the weights are determined by the proportion of women in each category, the proportion of men in each category, or the proportion of the total sample in each category).

c. Suppose that the standardized $p$-difference is significantly different from zero. How should the psychometrician interpret this result?

She can conclude that the item functions differently for the two groups. Specifically, among examinees in a given score category (assuming the item exhibits uniform DIF), one gender performs better than the other. This implies that the item measures some construct (other than the primary construct of interest), and the two groups differ in terms of their distribution on this additional construct. However, the item is not necessarily biased. It must be determined whether this additional construct is relevant to the purposes of the test.
Measurement Question 3

The table below contains information about three test items that have been calibrated with the two-parameter logistic model (2PLM). The left half displays the discrimination ($a$) and difficulty ($b$) parameters for each item. The right half displays the probability of a correct response at each of several ability ($\theta$) values.

<table>
<thead>
<tr>
<th>Item</th>
<th>2PLM parameters</th>
<th>Probabilities at several $\theta$ values</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>1</td>
<td>1.2</td>
<td>-0.6</td>
</tr>
<tr>
<td>2</td>
<td>1.5</td>
<td>-0.1</td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
<td>0.7</td>
</tr>
</tbody>
</table>

a. An examinee responded correctly to the first two items, but responded incorrectly to Item 3. Compute the likelihood of this response pattern at each of three ability levels: $\theta = 0$, 0.5, and 1.0. Which of these three values is the best estimate of the examinee’s true ability?

At $\theta = 0$: $L = .67 \times .54 \times (1 - .27) = .264$
At $\theta = 0.5$: $L = .79 \times .71 \times (1 - .43) = .320$
At $\theta = 1.0$: $L = .87 \times .84 \times (1 - .60) = .292$
So, $\theta = 0.5$ is the best estimate of the examinee’s true ability.

b. Is the ability estimate that you found in Question 1 the examinee’s maximum likelihood ability estimate? Why or why not?

No. Conceptually, the maximum likelihood estimate would be found by computing the likelihood of the examinee’s response pattern at every possible value of $\theta$ (i.e., all real numbers), and choosing the $\theta$ value with the highest likelihood. But in Question 1, the likelihood was computed at only three possible $\theta$ values. Practically, the maximum likelihood estimate would be found by using some optimization algorithm (e.g., Newton-Raphson).

c. What is the standard error of the ability estimate that you found in Question 1? As a reminder, item information (for item $j$) under the 2PLM can be computed as follows:

$$I_j(\theta) = a_j^2 P_j(\theta)[1 - P_j(\theta)]$$

Information for item 1: $I_1(\theta = 0.5) = 1.2^2 \times .79 \times (1 - .79) = .239$
Information for item 2: $I_2(\theta = 0.5) = 1.5^2 \times .71 \times (1 - .71) = .463$
Information for item 3: $I_3(\theta = 0.5) = 1.4^2 \times .43 \times (1 - .43) = .480$

Test information: $I(\theta = 0.5) = \sum I_j(\theta = 0.5) = 1.182$

$$SE(\theta) = \sqrt{\frac{1}{I(\theta=0.5)}} = \sqrt{\frac{1}{1.182}} = .920$$
Measurement Question 4

a. Conceptually, what is item discrimination? Is it desirable for an item to be highly discriminating? Why or why not?

Conceptually, item discrimination refers to an item’s ability to distinguish between examinees who are high on the construct of interest and examinees who are low on the construct of interest. For example, if an algebra item is highly discriminating, examinees who are highly proficient in algebra are very likely to provide the correct answer, whereas examinees who are not very proficient in algebra are unlikely to provide the correct answer.

Another way to state this is that if an item is highly discriminating, the expected item response is highly sensitive to variation on the construct of interest.

Yes, it is desirable for an item to be highly discriminating. If an item does not discriminate at all, this means that all examinees, regardless of proficiency, will perform equally well on the item. In this case, an examinee’s item response tells us nothing about his or her true proficiency.

b. Describe one way to measure item discrimination using classical item analysis.

One option is the item-total correlation. This is the correlation between the scores on an item and examinees’ total scores (which should be computed based on all items other than the item of interest). In the case of a dichotomously-scored item, this is known as the point-biserial correlation (another option is the biserial correlation). This is a good measure of discrimination because a large (positive) correlation indicates that examinees with higher test scores (i.e., those who are highly proficient) should also have a high chance of responding correctly to an item. Conversely, examinees with lower test scores (i.e., those who are not so proficient) should have a low chance of responding correctly.

An alternative measure is the index of discrimination. This is computed by separating a sample of examinees into “high” and “low” proficient groups. This is often done by separating the top and bottom 30% or so of examinees on the basis of the total test score. Then (for a dichotomously-scored item), you compute the item $p$-value (i.e., proportion of correct responses) within each group, and then compute the difference between the two $p$-values. The more discriminating the item, the greater the difference between the two $p$-values.

c. How is item discrimination defined in the context of item response theory?

In the context of IRT, discrimination is defined as the steepness (i.e., the slope) of an item response curve, which is a function describing the relationship between the expected item response and an examinee’s latent trait value. At its most extreme, the curve becomes a step function so that examinees above a certain level always respond correctly and examinees below that level always respond incorrectly. At the other extreme, the curve becomes a flat (horizontal) line, and all examinees perform equally well, regardless of latent trait value.
Research Design Question 1

There can be multiple ways to construct a research design and collect data to answer a research question. Please read the following description of two studies that were actually conducted to examine the effect of vitamin C on cancer patients and answer the questions.

Cameron and Pauling (1976) presented observational data concerning the use of vitamin C as a treatment for advanced cancer. They gave vitamin C to 100 patients believed to be terminally ill from advanced cancer and studied subsequent survival. For each patient, 10 historical controls were selected of the same age and gender, the same site of primary cancer, and the same histological tumor type. Then patients receiving vitamin C were compared to the “control” patients in terms of time from “untreatability by standard therapies” to death. Cameron and Pauling found that patients receiving vitamin C survived about four times longer than the controls, and the difference was statistically significant, \( p < .0001 \). Accordingly, they concluded that “there is strong evidence that treatment …[with vitamin C]… increases the survival time.”

Moertel and colleagues (1985) conducted a randomized controlled experiment comparing vitamin C to placebo for patients with advanced cancer of the colon and rectum. In the experiment, a total of 200 subjects were assignment to treatment or control on the basis of a chance mechanism such as using a random number generator. They found no indication that vitamin C prolonged survival, with the placebo group surviving slightly but not significantly longer.

a. Please comment on possible internal validity threats for the kind of research design chosen by Cameron and Pauling.

The research design used by Cameron and Pauling is an observational study in which the matched sampling is chosen to ensure comparable groups in a few important ways. However, beyond this, there is little to ensure comparability, and possible pre-existing differences cannot not be attributed to “by chance.” Particularly, the control group is formed from records of patients already dead, which means there could be other important hidden differences between the control group and the treatment group were initially alive. Furthermore, there are rooms of misinterpreting or misdiagnosing of “terminally ill” or “untreatability by standard therapies” in the process of conducting the research. If two groups differ in systematic ways (whether it is known or unknown), the variability in dependent variable cannot be attributed to the treatment variable.

b. Please comment on possible internal validity threats for the kind of research design adopted by Moertel and colleagues.

The research design used by Moertel and colleagues is a randomized control study in which the possible differences between two groups are attributed to “by chance” as long as randomization is processed properly. Diffusion, mortality, selection, or interaction with selection could be possible internal validity threats for this kind of study.
c. Moertel and colleagues (1985) used a “between-subjects design” rather than a “within-subjects design.” Please describe how these two designs are different and what the merits/limitations of each design are. In your answer, please provide your reasoning why Moertel and colleagues (1985) did not choose a within-subjects design.

A within-subject design uses a single sample of participants and exposes each individual in that sample to all conditions of the independent variable, while a between-subjects design uses different samples of participants are exposed to the conditions of the independent variable. Within-subject designs can provide greater power than between-subject designs because individual differences are controlled for (the exact same people serve in both conditions). Although within-subject designs can provide greater power, there is a great downside to these studies if they produce carryover effects. The fact that a participant experienced one condition may alter his behavior or reaction to another condition of the experiment. This is the reason why Moertel and colleagues did not choose a within-subjects design. In contrast, between-subjects designs are less powerful but can be effectively used for this kind of study when a carry-over effect is particularly concerned.

d. Assuming that statistical analyses were conducted properly, which study provides stronger evidence for researcher’s finding? In other words, which study looks “more internally valid” to you and why do you think so?

When observational studies and experiments yield different conclusions, the experiments tend to be believed in general because observational studies cannot ensure comparability between two groups even if a matching-technique is used. Even though experiments still have internal validity treats, experiments are preferred to observational studies for internal validity whenever they are feasible.
Section 4: Research Design

Research Design Question 2

Researchers are often interested in conducting an analysis while “controlling” for some key variable or variables.

a. What does it mean to “control for” other variables, conceptually speaking?

If one is interested in the effect of X on Y, but is concerned that both are affected by Z, then a failure to consider Z in any analysis of the relation between X and Y could be inaccurate. Hence we want to try to understand the relationship between X and Y above and beyond that induced by Z by “controlling for” Z in some manner.

b. Provide an example of a scenario in which a researcher might wish to control for a specific variable and why. How could one control for that variable experimentally, if at all? How could one control for that variable statistically, if at all?

I would like to know the effect of a technology training course on senior citizens’ ability to use the internet for seeking medical information. Seniors differ, however, in their prior experience with technology. Experimentally, we can randomly assign seniors to a treatment a control group, thereby balancing groups on prior technology experience before conducting the comparison. Statistically, we could examine intact treatment and control groups’ performance on outcomes of interest while treating prior experience with technology as a covariate.

c. Imagine a scenario in which two populations have equal means on a continuous control variable. Are there any circumstances under which it might be beneficial to control for that variable in analyses comparing the two populations on some outcome variable of interest? If so, explain why. If not, explain why not.

Control variables serve at least two purposes – to provide mean level adjustments and to reduce within-group (error) variability when the outcome is correlated with the control variable. Even when no mean level adjustment is needed, a control variable can greatly increase statistical power by partialing out variability on the outcome explainable by the control variable.