SYMSYS 130: Research Methods in the Cognitive and Information Sciences (Spring 2013)

Homework #2 – Instructor's Responses
May 19, 2013

Please respond to the following questions with short essays (300-500 words, not more). Answers will be scored out of 25 points total, based on the following criteria (5 pts each):

- informativeness (interesting, nonobvious),
- correctness (sound, accurate),
- thoroughness (convincing, rigorous),
- coherence (consistent, well constructed), and
- conciseness (clear, succinct).

1. Imagine that you have been involved in an intense debate over an email list, which includes people with whom you have strong disagreements. Toward the end of the debate, which lasts for a few days, your email account is suddenly overridden with spam email messages in an apparent email spoofing attack. You think someone on the list may be responsible for this attack, due to the timing of it and the fact that you have not been involved in such a debate before on this list, nor have you been a victim of this type of attack. Is this theory falsifiable, in Popper's terms? What predictions could be derived from it that could be tested? How might confirmation bias be avoided as you explore the evidence?

Popper says that if a theory is testable, then it is falsifiable, although “there are degrees of testability” (from: “Science as Falsification”). Testing a theory, for Popper, is an attempt “to refute it”. This is done by deriving “risky predictions” from the theory. A prediction is risky “if, unenlightened by the theory in question, we should have expected an event which was incompatible with the theory – an event which would have refuted the theory.” In this case, my theory is an attempt to explain why I have been email spammed: that it was done by someone on the email list who is angry with my comments on the listserv, who wants to hurt me as a result. The theory posits an isolated attack, done by one other person with me as the target, because of a recent event involving both of us. If the theory is correct, then people who have not been involved in debates involving this person should not be spammed.

Consider a different theory, namely that this is a random attack which has befallen many other people and is unrelated to the list debate. Assuming I have no other candidate reasons why someone would specificallly target me, aside from the “list debate” theory, this “random attack” theory is the most plausible alternative. A prediction of the random attack theory is that if I look for others who have been attacked in a similar fashion, I will find people who have been, and who have not been involved in debates involving anyone on the email list. If the prediction is borne out, it would evidentially confirm the “random attack” theory and refute the “list debate” theory. I could test this prediction by asking friends whom I know not to be on the list if they have experienced a similar attack (describing it in as much detail as I can). I could post the question on the Internet and see whether strangers from afar have been similarly victimized. If I find that others have experienced similar attacks, and cannot attribute these to anything they may have done that made someone angry, I have fairly strong evidence for the “random attack” theory. So the “list debate” theory is falsifiable, in Popper's terms. It generates a *negative* prediction, namely that I will not find evidence for the “random attack” theory, which, because it is an alternative to the “list debate” theory, would refute the “list debate” theory. Confirmation bias is avoided here by attempting to confirm the alternative theory and to refute the theory I start with. If I do find evidence that others have been similarly attacked in a way that cannot be attributed to another person's anger, it is still possible that I have been attacked out of anger, even as others have been attacked randomly. So Popper might say the “list debate” theory does not forbid the evidenceI am looking for, and this bears on the quality and degree of testability of my theory according to him.

2. A researcher is interested in the effect of discussion with people of different viewpoints on the post-discussion viewpoints of participants. Pre- and post-discussion attitudes are measured on a 0-10 scale, with average responses of 5.1. The researcher finds on average that for participants who start out with extreme positions (≤3 or ≥7), responses on post-discussion questions are less extreme (closer to 5) than the response on the same question pre-discussion. Explain how this result could arise without discussion actually having any systematic effect on attitudes. What is a better way to measure whether viewpoints are more extreme pre- or post-discussion?

The researcher is looking at the post-discussion attitudes of two groups whose pre-discussion attitudes are relatively far away from both the midpoint (5) and the mean (5.1) of the response scale. Suppose that participants answer the questions randomly (equal likelihood across the scale from 0 to 10). [This is known as a “uniform distribution”.] This could result in a mean response for all participants of 5.1 assuming the group is not large enough for the sample mean to converge to the distribution mean of 5. Alternatively, suppose just that responses are relatively random, with a slight bias upward from 5, resulting in the average of 5.1. In either of these scenarios, we would expect the average post-discussion response for each of our two groups (those answering ≤3 and ≥7 on the pre-discussion questions, respectively) to be around 5 or 5.1. Yet this would not reflect discussion having any systematic effect on attitudes, since, by hypothesis, respondents are answering randomly or randomly with a slight bias that is shared across both groups. This is an extreme instance of regression to the mean, in which the regression is 100% toward the mean due to the complete lack of correlation between pre- and post-discussion responses.

A better way to measure whether viewpoints are more extreme pre- or post-discussion would be to look at the average deviation of responses from the midpoint of the scale, comparing the pre- and post-discussion responses. This can be done by computing the average of $|\text{pre}_i-5|$ and $|\text{post}_i-5|$ across all n participants, where pre_i is the response score for the *i*th participant pre-discussion, and post_i is the response score for the ith participant post-discussion. We would then compare

$$\text{AvgDev}_{\text{Pre}} = \sum_{\text{i=1 to n}} |\text{pre}_i-5| /n$$

to

$$\text{AvgDev}_{\text{Post}} = \sum_{\text{i=1 to n}} |\text{post}_i-5| /n$$

in order to see if viewpoints are more extreme pre- or post-discussion. Another way to do this would be to compare the sample standard deviations of pre_i and post_i, which would measure how far an average response was from the sample mean of 5.1, pre- and post-discussion. These approaches avoid regression to the mean as a confounding explanation for responses becoming less extreme after discussion for groups that are chosen on the basis of being on one side of the scale or the other before discussion, because we are not conditioning on pre-discussion responses in order to predict post-discussion responses.

[Here are some additional notes to help you understand regression:

The same potential for regression to the mean would arise if we conditioned on pre-discussion responses in groups defined as being (a) above and (b) below the midpoint or mean of the response scale. The regression phenomenon is symmetric, so it would also predict (or "post-dict") that pre-discussion responses for those who score ≤3 and ≥7 on the *post*-discussion questions, respectively, would be less extreme in each group. Regression does *not* say that dependent variable averages are less extreme than independent variable averages. It says that when two variables v and w are paired (i.e., when v_i and w_i are defined for each i in a set of n observations on both variables), the *average value* of v_i for a subset of the i's whose w_i values differ on average from the overall average (for all n observations) of w *will be closer* to the mean of v than the average w_i values for that same subset of i's are to the overall average of w.]

3. Coombs, Dawes, and Tversky (p. 17) say that "no measurement theory for intelligence is available." What justification do they have for saying this?

According to Coombs, Dawes, and Tversky, measurement is the process of assigning numbers to objects in a way that represents some property of the objects (see pages 7 and 10). They say that intelligence provides an example of "statements involving numerical values for which no measurement model exists" (p. 17). In this case, the real world property is intelligence, and the numerical scale is IQ. For a measurement model to exist, there needs to be some way of assigning numbers to objects (people, in this case) in a way that represents an empirical relation between the objects, for example: x has a higher IQ than y [formal, numerical relation] if and only if x is more intelligent than y [empirical relation] (applying the defintiion of a representation relation for measurement on p. 11). But the authors say this is an instance of "the absence of a well-defined representation relation" (p. 17). What would it mean to have such a well-defined relation? Well, IQ is defined by tests that have been constructed and a way of scoring answers on those tests across individuals in a population. So the problem must be that despite the ability to say that x scored higher on an IQ test than y, we cannot determine whether or not this implies or is impled by x being more *intelligent* than y. We may try to use IQ scores to predict what the authors loosely call "intellectual performance". The authors say (p. 18):

> "The basic dependent variable, however, remains elusive. It seems difficult to define a single measurement dependent variable whose correlation with the IQ scale should be maximized. Thus any decision concerning what to predict involves some serious theoretical considerations. Furthermore, even if the dependent variable (say intelligence) can be properly defined, one would eventually wish to predict it on the basis of some psychological theory (about intelligence) rather than on the basis of some 'blind' statistical procedure."

So the authors appear to be saying that no definition of intelligence is available that is distinct from that of the IQ scale being used to measure it. If there were such well-defined concept, it would need to representable on some type of scale (nominal, ordinal, interval, etc.). An ordinal definition requires us to agree on statements such as “x is more intelligent than y”, but we do not have an agreed upon way of doing this. People are ordered differently by different tests, depending on how they are written, and opinions typically differ about who is more intelligent than whom. So we can't use any scale that implies an ordering, which includes all but the nominal scale type. To use a nominal scale, we would just need to be able to make statements such as “x has a distinct IQ from y if and only if x is in a different intelligence category from y”, but again, we don't have an agreed upon definition of “intelligence categories” that would allow this to be tested.

4. Imagine you are designing a survey to determine which transit plan people in a given city prefer: plan A, plan B, or plan C. How might you sample the population of the city to minimize the bias in your estimates of relative support levels for these three plans? What methods of administering the survey (e.g. oral, written, or a combination) and of wording the question would minimize bias?

Reasonable ways to minimize bias in sampling from the city's population include the following:

- If we have a list of city residents, we can order the names and generate numbers at random from 1 to the number of residents to sample residents (without replacement) whose number in the ordering correspond to each generated number. This depends on having an accurate list of residents, however. Available lists will leave off some residents who have moved into the city recently, or if they are based on having a driver's license/state ID or being registered to vote, the lists will exclude many residents. Once we have the sample list, there is the problem of contacting residents. We can do so by mail or phone, but many will not respond and this may result in a bias. If we are sampling from voter lists, evidence cited by Krosnick (p. 540) suggests we will get a less biased response by not “aggressively pursuing high response rates.”
- We can try the old technique of random digit dialing., to sample from the population that has phones. If a respondent says they do not live in the city, they would not be surveyed. In the age of cell phones, many people may have more than one phone number (e.g. a land line, a work phone, and a cell phone). Some adjustment for this can be made by asking the respondent how many phones they have where they would be the most likely person to answer, and include their response with probability one over the number of phones they have.
- We can go to the waiting room for jury duty in the city, since residents are supposed to be chosen at random to be there on a given day. This would obviously restrict the sample to those old enough to serve on juries and those who are registered through the DMV as living in that city.

For administering the survey, research reviewed by Krosnick suggests the following ways to minimize bias:

- Provide a written questionnaire as the default survey method to participants, with an option for an oral administration for people who are unable to read the questionnaire. The written questionnaire leads to fewer memory limitation biases (recency) than an oral presentation (p. 550), but it can lead to a response order effect favoring choices listed earlier (the primacy effect, p. 549). To counteract this, the order or presentation of alternatives A, B, and C should be varied randomly and evenly (counter-balanced) across participants, and the letters assigned to the plans should be correspondingly varied.

- The questions should be worded as simply and succinctly as possible to reduce response order effects (p. 551), which would still produce noise even if the order is varied.
- Ask participants to rank rather than rate the three options, as this appears to lead to “higher quality data” (p. 556).