Agenda 10/9/2006

- Feedback on thought papers
- Reminder: Midterm a week from Friday
- General announcement
- New material
  - Types of observational research
  - Variables and their measurement
  - Example study
  - Distributions, central tendencies, and standard deviation

Thought papers

- Next time:
  - Please write your TA’s name on your work.
  - Entire packet should be stapled together and turned in to the original author’s TA.
  - If you’re missing a staple for some reason, make sure the headers of each piece identify the original author’s name and TA.
- Fantastic job on this round of thought papers!!!
- Remember that formal essays will have to be more tightly argued than these (we’ll go over writing mechanics in lecture and/or section), but for what the thought papers were intended to be, they were great.

Noise in lecture

- I’ve received some complaints from students about the amount of noise in class, especially towards the end of lecture.
- Please note that this kind of noise is not ok, because it prevents other students from being able to hear and focus on what’s going on during scheduled class time.
- This includes zipping bags, packing up books, etc.
- Please do not start packing up until 11:50. If I’m still talking at 11:50, go ahead and move.
- If you absolutely must get up during class for any reason, please do so very discreetly and as quietly as possible.

New material…

- Why do researchers code their data?
- Types of sampling - continuous vs. time, event, situation sampling
- Why does developing a coding system require a lot of planning, good judgment, and thought?
- Variables - what are they, what types are there

Key ideas 10/9/2006

- Be familiar with the examples of observational research projects under “a tour of observational research” module
- Difference between observational and experimental research
- Types of observational research
- Observer bias, reactivity, how to minimize them
Key ideas 10/9/2006

- What is a distribution? How do distributions differ for "score" versus nominal or categorical variables?
- Normal distribution
- mean, median, mode, standard deviation - how to calculate them and intuitively what they mean, what they are telling you

Observational research

- Observational researchers:
  - Always observe some type of phenomena
  - Make a record of their observations and analyze data from the record
  - Do not manipulate variables (e.g., actively do something to influence the phenomena, in order to test it under different conditions - that's an "experiment")

Why conduct observational research?

- Description ( atheoretical) - get a general characterization of phenomena in a certain domain or situation
- Exploratory (pretheoretical) - get a characterization of phenomena with the goal of later developing hypotheses and/or theories
- Confirmation or falsification (theoretical) - test a hypothesis.
  - If the hypothesis is a causal claim, observation usually won't be a good enough test, an experiment is usually needed

Types of observational research

- Naturalistic observation
  - Passive, unobtrusive, natural environment.
  - Minimize observer bias (e.g. by blinding observers to the goals of the study)
- Participant observation
  - Observer interacts with phenomena, usually in a natural context.
  - Minimize reactivity (reactivity is the influence of observer on what/who is being observed), e.g. avoid significant role in the situation, record observations when out of sight.

Types of observational research

- Structured observation
  - Natural environment, no active manipulation, but focused on phenomena of interest, e.g., arranging to be present at the time the phenomena of interest occurs
- Indirect observation:
  - interviews, surveys, self-reports. nonexperimental.
- "Field experiments"
  - the situation is actively manipulated by the researcher, are not pure observational research, but they occur in natural environments.

Types of records

- Narrative records
  - Involves recording as much information as possible, often in regular prose style
  - May be difficult to use in answering focused questions
  - Useful for telling particular stories in full context
  - In order to perform higher level analyses, however, one often needs to more systematically categorize what is observed
- Coding systems - used to categorize observations
- Both types of records have their uses
Types of records - coding systems and timing

- Continuous coding - constantly recording data. Often overkill, need to select when to record.
- Time sampling - record at predetermined interval
- Event sampling - wait until target phenomena occurs
- Situation sampling - collecting data from different situations
  - Sometimes the initial data capture is in coded form, sometimes "coding" step comes after the initial recording of observations.
  - Sometimes more than one person does one or both steps, and then the judgments are compared/discussed to ensure agreement - this helps with standardization and reliability.

Risks in observational research

- Observer bias - when expectations influence perception
  - Can moderate this by blinding observers about the goals of the research.
- Reactivity - when awareness of being observed changes the behavior from what it otherwise would have been
  - Can try to be unobtrusive, or use a measure that is unobtrusive.

Coding systems

- You can adopt a conventional coding system and/or a system that has already been in use by other researchers in a given field
  - Would help with relating your data to other research
- You can create your own coding system if there isn’t already one that suits your needs - requires a great deal of planning, foresight, and good judgment
- The way you capture and code data constrains how you can analyze it - need to consider what you need to analyze, what you may want to look at "post-hoc", as well as resource limitations, among other things.

Variables - some vocabulary

- Variables - "things that vary," the general categories that are to be measured.
- Units - the standard of measurement used in giving the value (e.g. pounds vs. kilograms, years vs. months)
- Variables - "things that vary," the general categories that are to be measured.
- Values - the particular measurements or "answers" with respect to a variable, e.g., the measurement or result for an individual subject, or a range of possible results (e.g., 0 to 110 years if the variable were "Age")
- Units - the standard of measurement used in giving the value (e.g. pounds vs. kilograms, years vs. months)
- Data field - the spot or "place" (e.g. in a database, or on a paper form, or wherever data is to be recorded) where the value for a given variable can be recorded or captured.

Variables - some vocabulary

Examples

- If you're measuring age, and entering the data on a form and your subject is 20 years old, the variable is "Age," the value is "20," the units is "years," and the data field is the space on the form where you write down (capture) the value of "20."
- If you're recording height in an excel spreadsheet, and your subject is 5’6”, the variable is "Height," the value could be "66" and the units would be "inches" (or, the value could be "5.5" and the units would be "feet"). The cell in the Excel spreadsheet where you capture the value is the data field.
- If you're measuring political party, the variable would be "political party," and the values could be "republican," "democratic," "libertarian," etc. and in this case, there wouldn't be any relevant units because of the sort of variable it is. And of course, wherever you entered the value would be the data field.

Types of variables

- Nominal ("named," sometimes called "categorical")
- Ordinal ("ordered")
- Interval (where the intervals between possible values are equal)
- Ratio (numerical, with a meaningful zero point)
Types of variables: Nominal (categorical)

- no obvious ranking or ordering of the possible values, values usually labeled with a finite number of names (sometimes with a catch-all “other” category for items that don’t easily fit into the given options), mostly not numerically represented
  - E.g. gender, race, political party, disease type, country, language spoken, etc etc
  - Capturing data in nominal/categorical form limits how it can be analyzed and graphically represented

Types of variables: Ordinal (ordered, or rank)

- Values can be ordered with respect to one another (e.g. “Strongly agree, agree, neutral, disagree, strongly disagree” or “Outstanding, very good, good, fair, poor, failure.”)
- Numbers associated with the values need not convey any information about precise quantities.
- Can do more sorts of analysis with ordinal variables than with nominal/categorical ones, but not as much as with interval and ratio variables, because the rankings aren’t necessarily very precise in terms of quantity and “distance” from each other.

Types of variables: Interval

- A type of “score” variable, sometimes called “numerical.”
- Intervals (or “distances”) between possible successive values are constant, or equal
- Continuous measure, numbers specify differences between terms. E.g., temperature
- Units must be specified
- Don’t permit us to infer multiplicative relationships (e.g. 10 degrees isn’t necessarily twice as warm as 5).

Types of variables: Ratio

- A type of “score” variable, sometimes called “numerical.”
- Sometimes the values are called “scores.”
- Continuous measure, meaningful zero-point on scale
- Units must be specified
- E.g. Age, height, time, date, duration, speed, frequency, weight, mass, concentration (i.e. of a substance in a liquid), volume, calories, etc etc.

Discussion - 1

Say we needed to do an exploratory observational research study concerning the dietary habits of the campus population.
For now, we want to generally see whether different “groups” tend to have “healthier” versus “unhealthier” diets, but we have to decide for ourselves how to interpret that. Later, we may want to do further analyses of our data set to answer sub-questions, e.g. whether time spent on campus, gender, race, activities, relate in interesting ways to diets.

Discussion - 2

- We have to design our own study and construct our own coding system
- However, our resources are limited.
  - We can only afford to capture 15 different variables, due to the financial cost of data capture and analysis.
  - We only have 1 quarter to plan, 1 quarter to conduct the study.
Discussion - 3

• Any suggestions as to how we might interpret the general question?
• Any suggestions as to how we might collect our data?
• What sorts of variables might we want to collect?
• How do the resource limitations constrain what we can do?

Discussion - 4

<table>
<thead>
<tr>
<th>Variables</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Living arrangement - how many in same place, on or off campus</td>
<td>Waiting at restaurants - event sampling</td>
</tr>
<tr>
<td>Socioeconomic status</td>
<td>Go to stores, apartments off campus</td>
</tr>
<tr>
<td>Work - hours per week</td>
<td>Self-report questionnaires</td>
</tr>
<tr>
<td>Mental stability</td>
<td>Grocery stores - look in bags</td>
</tr>
<tr>
<td>What you ate for dinner</td>
<td>Waiting at bus, survey</td>
</tr>
<tr>
<td>Age, year, gender, nationality</td>
<td>Don’t ask questions, just observe</td>
</tr>
<tr>
<td>Exercise - hours per week</td>
<td>Narrative or coding?</td>
</tr>
<tr>
<td>Categories from govt - residents to heavy exercise, etc.</td>
<td>Score coding?</td>
</tr>
<tr>
<td>Body weight, BMI</td>
<td>Event sampling</td>
</tr>
<tr>
<td>School grades</td>
<td>Self-reports?</td>
</tr>
<tr>
<td>Calories from different macronutrients</td>
<td>Grocery stores - look in bags</td>
</tr>
<tr>
<td>Percent of different macronutrients</td>
<td>Waiting at bus, survey</td>
</tr>
</tbody>
</table>

Moving on…

Distribution

• Representation of how many of the items (or people) measured have each of the different possible values or scores on a given variable.

Example of a Distribution for a Nominal/ Categorical Variable

- Favorite Genre of Music

Distribution for a Score Variable

- With score variables, a metric relates the different values of the variable (whereas with nominal variables, the order on the x axis doesn’t necessarily matter)
- Here, we created “bins” for different values on the continuous scale
More example distributions
score variable with a perfectly normal distribution

Descriptive Statistics
Quiz 1 Scores for Tarun’s Students

n 71
mean 32.54
median 33
mode 33
min 10
max 48
variance for sample 76.28
stdev for sample 8.73
variance for population 75.21
stdev for population 8.67

Raw Data Listing
Quiz 1 Scores for Tarun’s Students

Quiz 1 Scores for Tarun’s Sections

Different Bin Sizes

Different Bin Sizes - 2
Distributions for Score Variables

- Values of score variables can be distributed in different patterns
- "Normal curve" - scores equally distributed about the peak value, declining the further you move out from the peak
- Distributions can also be “skewed” in one direction or another
- “Bimodal distribution” - two peak values
- The size of the “bins” affects how the distribution looks

Central Tendencies

- Mean - the average score
- Median - the “middle” score, the score such that half the scores are greater, half are less
- Mode - the most frequently occurring score
- With a perfectly normal distribution, the mean, median and mode would be the same value

Central tendencies

- Take the following set of values for the variable "body weight," with the units in pounds: (98, 100, 100, 100, 125, 125, 140, 170, 210, 250)
- Here are some descriptive statistics for this set of data:
  - Number of scores: 10
  - Mean: 141.8 lbs
  - Median: 125 lbs
  - Mode: 100 lbs
  - Range: [Minimum: 98 lbs, Maximum: 250 lbs]
  - Standard Deviation: 52.6 lbs

Variance and Standard Deviation

- An important piece of information about a distribution for a numerical or score variable is how “spread out” it is
Variance and Standard Deviation

- We need a precise way to describe/characterize the variability in a sample.
- The variance:
  - for each value in the sample, take the difference of that value from the mean score (each score minus the mean score);
  - square each of the resulting values, then add the squared values together;
  - then divide that sum by the number of scores (if you're calculating the variance of a sample, you divide by the number of scores minus 1, to provide an unbiased estimate of the variance);
- Standard deviation (stdev): the square root of the variance.

Formulae for variance

- Top formula (in "summation" notation) is for variance in a population, the mu (µ) refers to the population mean.
- Bottom formula is for variance in a sample, the M refers to the sample mean, the denominator is the number of scores minus 1, to provide an unbiased estimate.
- The square roots of these provide the standard deviation.

Standard deviation in normal distributions, Z-scores

- In a normal distribution, 68% of scores are within 1 standard deviation of the mean, 95% of scores are within 2 standard deviations of the mean.
- Z-score - for an individual value, the ratio of its "distance" from the mean - an item 1 stddev from the mean has a z-score of 1.0, an item a half a stddev from the mean has a z-score of 0.5.

Just so you are aware...

(you're not responsible for remembering this yet!)