STA 291
Fall 2009

LECTURE 3

THURSDAY, 3 September

Administrative Notes

Suggested Problems:

- 3.1, 3.2, 3.7, 3.12

HW 1 (sic) is on MyStatLab!

- Graded [10 points]

  See online when you log onto your MyStatLab account

  Up to three attempts to take: *best* try counts

  You get to see number right, not which problems

- Add period is over ☹

Review: Qualitative Variables

(=Categorical Variables)

Nominal or Ordinal

- Nominal variables have a **scale of unordered categories**
- Ordinal variables have a **scale of ordered categories**

If not Qualitative, then what?

- Then they're **Quantitative**

- Quantitative variables are measured numerically, that is, for each subject, a number is observed
- The scale for quantitative variables is called **interval scale**

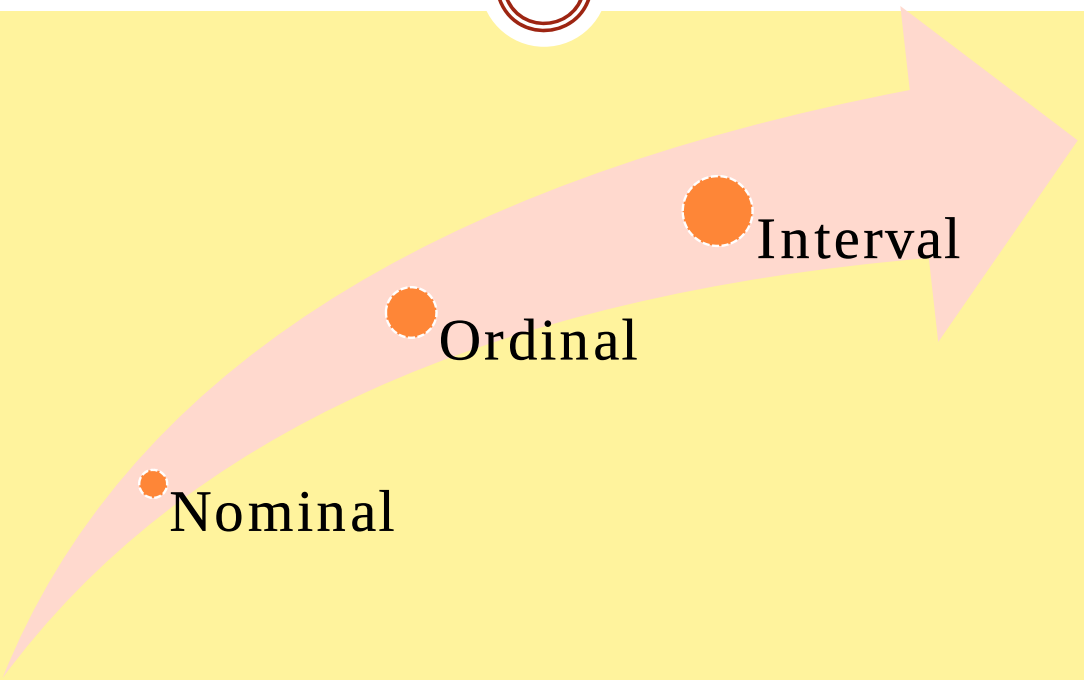
Scale of Measurement Example

The following data are collected on newborns as part of a birth registry database:

- Ethnic background: African-American, Hispanic, Native American, Caucasian, Other
- Infant's Condition: Excellent, Good, Fair, Poor
- Birthweight: in grams
- Number of prenatal visits

What are the appropriate scales?

Why is it important to distinguish between different types of data?



Some statistical methods only work for quantitative variables, others are designed for qualitative variables. The higher the level, the more information and the better statistical methods we may use.

Discrete versus Continuous

- A variable is **discrete** if it has a finite number of possible values
- ***All*** qualitative (categorical) variables are discrete.
- *Some* quantitative (numeric) variables are discrete—which are not?
- A variable is **continuous** if it can take all the values in a continuum of real values.

Discrete versus Continuous, Which?

- Discrete versus Continuous for quantitative variables:

  - discrete quantitative variables are (almost) always counts

  - continuous quantitative variables are everything else, but are usually physical measures such as time, distance, volume, speed, etc.

Simple Random Sample

- One drawn so that every possible sample has the same probability of being selected.
- The sample size is usually denoted by n.

SRS Example

- Population of 4 students: Adam, Bob, Christina, Dana
- Select a simple random sample (SRS) of size n=2 to ask them about their smoking habits
- 6 possible samples of size n=2:

  (1) A & B, (2) A & C, (3) A & D

  (4) B & C, (5) B & D, (6) C & D

How to choose a SRS?

- Old way: use a random number table.

- A little more modern: http://www.randomizer.org
  (first lab exercise)

How **not to choose a SRS?**

- Ask Adam and Dana because they are in your office anyway
  - "convenience sample"
- Ask who wants to take part in the survey and take the first two who volunteer
  - "volunteer sampling"

Problems with Volunteer Samples

- The sample will poorly represent the population
- Misleading conclusions
- BIAS
- Examples: Mall interview, Street corner interview

Other Example

- TV, radio call-in polls
- “Should the UN headquarters continue to be located in the US?”
- ABC poll with 186,000 callers: 67% no
- Scientific random sample with 500 respondents: 28% no
- The smaller **random sample** is much more trustworthy (likely to reflect the views of the population) because it has less bias

Why are call-in polls usually biased?

People are much more likely to call in if the feel strongly about an issue:

(Israel-Palestine, Iraq, water company, mountaintop removal, equal rights for homosexuals, pedestrian safety, name of the UK mascot)

Methods of Collecting Data I

Observational Study

- An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses.
- The purpose of an observational study is to describe/ compare groups or situations.
- Example: Select a sample of men and women and ask whether he/she has taken aspirin regularly over the past 2 years, and whether he/she had suffered a heart attack over the same period

Methods of Collecting Data II

Experiment

- An experiment deliberately imposes some treatment on individuals in order to observe their responses.
- The purpose of an experiment is to study whether the treatment causes a change in the response.
- Example: Randomly select men and women, divide the sample into two groups. One group would take aspirin daily, the other would not. After 2 years, determine for each group the proportion of people who had suffered a heart attack.

Methods of Collecting Data III

Observational Study/Experiment

- **Observational Studies are passive data collection**
  - We observe, record, or measure, but don't interfere
- **Experiments are active data production**
  - Experiments actively intervene by imposing some treatment in order to see what happens
- *Experiments are preferable if they are possible*

Some Other “Good” Sampling Methods

- Stratified Sampling
- Cluster Sampling
- Systematic Sampling

Stratified Sampling

- Suppose the population can be divided into separate, non-overlapping groups (“***strata”) according to some criterion.***
- Select a simple random sample independently (and usually proportionally) from each group.
- Can be used to reduce sampling variability, but more often used when we wish to make stratum-level inferences

Cluster Sampling

- The population can be divided into a set of non-overlapping subgroups (the clusters)
- The clusters are then selected at random, and all individuals in the selected clusters are included in the sample

Systematic Sampling

- A.K.A. Systematic Random Sampling
- An initial name is selected at random
- every K^{th} name is selected after that
- *K* is computed by dividing membership list length by the desired sample size
- Not a simple random sample (why?), but often almost as good as one

Summary of Important Sampling Plans

- **Simple Random Sampling (SRS)**
  - Each possible sample has the same probability of being selected.
- **Stratified Random Sampling**
  - Non-overlapping subgroups (strata)
  - SRSs are drawn from each strata
- **Cluster Sampling**
  - Non-overlapping subgroups (clusters)
  - Clusters selected at random
  - All individuals in the selected clusters are included in the sample
- **Systematic Sampling**
  - Useful when the population consists as a list
  - A value K is specified. Then one of the first K individuals is selected at random, after which every K^{th} observation is included in the sample

Types of Bias

- **Selection Bias**

– Selection of the sample systematically excludes some part of the population of interest

- **Measurement/Response Bias**

– Method of observation tends to produce values that systematically differ from the true value

- **Nonresponse Bias**

– Occurs when responses are not actually obtained from all individuals selected for inclusion in the sample

Next Definition: Sampling Error

- Assume you take a random sample of 100 UK students and ask them about their political affiliation (Democrat, Republican, Independent)
- Now take another random sample of 100 UK students
- Will you get the same percentages?

Sampling Error (cont'd)

- No, because of sampling variability.
- Also, the result will not be exactly the same as the population percentage, unless you

  1. take a "sample" consisting of the whole population of 30,000 students (this would be called a ?)

  - or -

  2. you get very lucky

Attendance Survey Question #3

- On an index card
  - Please write down your name and section number
  - Today's Question: