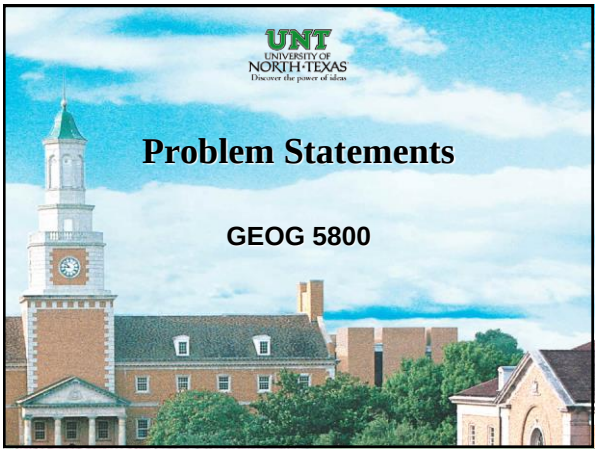
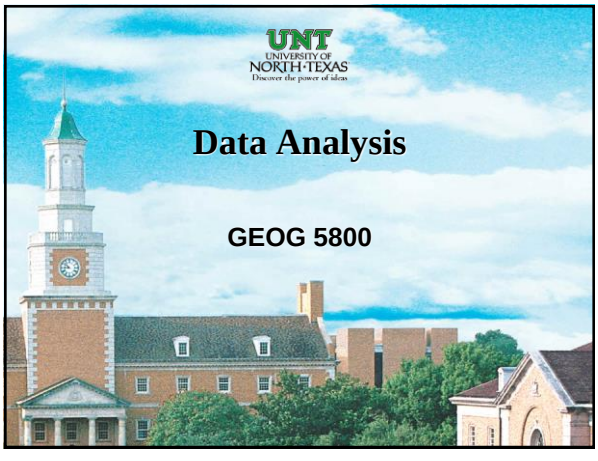


Problem Statements

- A couple of general observations/ requests
  - 1. Continue to pay attention to how you communicate your ideas
    - Build a logical case for what you do
    - If you make an assumption, state it
    - If you use specific terminology, define it (even if you think your reader knows, make sure they understand that you know the correct definition too)

Problem Statements

- A couple of general observations/ requests
  - 2. Continue to consider how you might better define your study
    - Key idea: making sure your study is one you can do in a few months, not years/decades
    - How will you know when you are done?
    - How much data will you need to collect?
    - How will you tell when your question is answered?



Data Analysis

- We focus today on what to do once you have your data
- Much of what we might discuss here ought to have been covered in other courses you have taken
- Idea here: remind you of a couple of things, highlight a few important issues

Multivariate Statistics

- Multivariate stats have increased in popularity recently in virtually all fields of physical and social sciences
- Why?
  - 1. difficult to address complicated research problem without multivariate analysis
  - 2. good software is so much more available today than 10 or 20 years ago

Multivariate Statistics

- Challenge is often not in doing the analysis, but in selecting which analysis to use
- The handout "Major Statistical Methods" is a (hopefully) useful summary of some key methods you might consider in your own research and analysis

Multivariate Statistics

- Q: What statistical methods do you have experience in using? Any not on the handout?

Considering Your Data

- Comments from here on today will focus on the data we analyze
- Key issue: number and nature of variables to include in your analysis
- In general: the fewer, the better
  - Fewer variables: more general explanation, but less explanation
  - More variables: more explanation, but also more specific to a given situation

Data Screening

- Consider six keys to screening your data file for possible problems

Data Screening

- 1. Accuracy of data file
  - If analyzing a small file, proofreading is helpful
  - However, if you have a million entries in your file, this is not so practical!
  - With large files:
    - Apply simple descriptive statistics to see if things look OK
    - Are all variables within the range they should be in?
    - Are means and standard deviations plausible?
    - If you graph the data, is the distribution reasonable?

Data Screening

- 2. Honest Correlations
  - Inflated Correlation: it is possible for high correlations to be meaningless if the variables overlap (e.g. soil moisture and rainfall)
  - Deflated Correlation: falsely low correlations may occur when the range of one variable is artificially constrained (e.g. study education K-12 vs. income)

Data Screening

- 3. Missing Data
  - Missing data might come from several sources (equipment breakdowns, insensitive equipment, incomplete survey responses)
  - How to deal with this problem?
    - Deleting cases or variables: delete if only a few cases or variables are seriously affected
    - Estimating missing data: use when missing values fall in a known range, and the impact of excluding the case would be greater than a possible inaccuracy from estimation

Data Screening

- 3. Missing Data
  - Missing data might come from several sources (equipment breakdowns, insensitive equipment, incomplete survey responses)
  - How to deal with this problem?
    - Complete the analysis with and without the missing data: quantify what the impact of the data would actually be

Data Screening

- 4. Outliers
  - Cases with such extreme values that they distort the overall statistics
  - Simplest example: linear regression
  - Important to check for outliers

Data Screening

- 4. Outliers
  - Four reasons for the presence of outliers
  - 1. Incorrect data entry
  - 2. Programming error ("missing value codes" like "9999" treated as real data)
  - 3. Outlier is a member of a population other than the one intended for sampling
  - 4. Population being sampled is not a normal distribution (expect many "outliers")

Data Screening

- 4. Outliers
  - How to detect outliers?
  - Univariate: extreme value in a single variable (relatively easy to check for this)
  - Multivariate: may be an extreme value in one variable, or it may be an extreme combination (variables individually fall within usual range)
    - For example: 14 years old, $80,000 income

Data Screening

- 4. Outliers
  - Once detected, multivariate outliers should be described so you can further deal with them
  - If only a few: examine individually and deal with each one (exclude, correct, or try something else)
  - If many: examine them as a group or several groups (maybe the outliers are similar among themselves)

Data Screening

- 4. Outliers
  - Last step with outliers is to do a couple of final checks
    - 1. Check outliers for accurate recording
    - 2. If recording is accurate, check if one variable is responsible for most outliers
  - If these checks don't help, you need to decide whether the outliers should be in the study
  - Truly multivariate outliers are the hardest to deal with (may need to leave in the study)

Data Screening

- 5. Normality and Linearity
  - Normality: important because so many statistical tests (but not all) require normality – check for skewness (lop-sidedness) and kurtosis (peakedness)
  - Linearity: again, methods like regression (univariate and multivariate) require linear relationships

Data Screening

- 6. Multicollinearity and Singularity
  - Multicollinearity: check for explanatory variables that are very highly correlated (e.g. 0.90+) and discard one or more
  - Singularity: one variable is actually made up of a combination of two or more of the other variables (again, can discard)

Data Screening

- Conclusion: have a data screening checklist and use it (see checklist handout)