0.1 STAT220: BASIC STATISTICS, WIN 2007

Instructor: Prof. Elizabeth Thompson
eathomp@u.washington.edu

Teaching assistants: Linqiu Wu, Sections AA, AB

Lee Barr, Sections AC, AD

Le Bao, Sections AE, AF

- **The web page:**
  www.stat.washington.edu/thompson/S220_07

Or, link from Statistics Dept home page is now working.

For course requirements,
book info (FPP), homework info,
exams and quizzes info,
and lecture notes.

In particular, check the web schedule
– it has links to the other things.

Also, office hours of TAs and Instructor (TBA).

0.2 WHAT IS STATISTICS?

- **Quantitative facts, numerical descriptions (data)**
- **Set of tools for the collection and analysis of data**
- **... in order to make decisions or draw conclusions from the data**

0.3 WHERE DO WE SEE DATA ANALYSES ?

- **News reports, Weather forcastes**
- **School records, grades, course evaluations**
- **Consumer reports, Election polls**
- **Environmental standards, air pollution, endangered species**
- **Medical and dental records, diagnostic procedures.**
- **Stock market, business plans, marketing surveys.**
- **... and many more**

0.4 WHAT DOES DOING STATISTICS INCLUDE?

- **Study design and data collection:**
  Experiments, Studies and Surveys
  FPP Chapters 1, 2, 19
- **Data description and exploration**
  Graphical and numerical summaries.
  FPP Chapters 3,4
- **Modeling Data : for example, the normal curve**
  FPP Chapter 5
- **Forecasting and prediction**
  The relationship between 2 or more variables
  Correlation and association: FPP Chapters 7,8,9
  Regression: FPP Chapters 10,11, 12
- **Understanding variation and randomness**
  Chance and randomness: FPP Chapters 16,17,18
  Accuracy and measurement error: FPP 20,23,24
- **Drawing inferences and making decisions**
  Estimates and confidence intervals: FPP Ch. 21
  Testing hypotheses: FPP Chapter 26, 27

Blank page for your notes

1.1 COLLECTION OF DATA

There are three basic study designs

1. Controlled experiments (FPP Chapter 1)
Investigator gives treatment to subjects in the treatment group.

2. Observational studies (FPP Chapter 2)
Investigator does not control who is in the treatment group.

3. Sample surveys (FPP Chapter 19)
A type of observational study. We study a sample of individuals from a population

- **Subjects: Study units, Experimental units**
- **Population: The set of individuals of interest**
- **Sample: Chosen subset of the population**
- **Variable: Characteristic or property of a subject**

Blank page for your notes

1.2 EXAMPLE: SALK VACCINE TRIALS

• From 1916, polio killed hundreds of thousands US children

• In 1954, Jonas Salk's vaccine seemed promising

• Need comparison: treatment and controls
  Vaccinated, about 500,000 children
  Unvaccinated, about 1,000,000 children
  Refused vaccination, about 500,000 children

• Randomized controlled trial:
  Investigator decides who is to be vaccinated/not.
  Use random assignment to treatment or control group— those whose parents refused vaccination are not good controls (confounding factors)

• Use of placebo, avoids placebo effect
  Subjects should not know whether they are treatment or control —use of saline solution for children in control group.

• Double blind assessment
  Neither subjects nor diagnosing physicians know who is treatment and who is control.

Blank page for your notes

1.3 TREATMENTS, RESPONSE, and FACTORS

• Treatment or control is applied to the experimental unit (or subject).

• The response is the outcome: the data we analyze.

• The treatment may involve several factors. For example, cancer treatment may involve surgery, radiation therapy and chemotherapy.

Surgery: yes or no (2 levels)

Radiation treatment: high dose, low dose, or none (3 levels)

Chemotherapy: protocol-1, protocol-2, or none (3 levels)

• Need to try (all?) combinations to assess treatments
Some combinations may not be feasible/ethical
Issues of time, cost (numbers of subjects)

• If do do all combinations, this is a complete factorial design

Blank page for your notes

1.4 RANDOMIZED CONTROLLED TRIALS

• Investigator assigns subject to treatment/control.
This avoids confounding factors How should he/she assign?

• We want treatment group to be similar to control group
but any directed attempt to make them similar may lead to bias.

• Only random assignment of eligible subjects is safe
then we can assess results objectively
Subjective confounding factors will not cause bias.

• What about obvious confounding factors.
For example, gender, in study of hormone drug reactions.
Randomization will take care of it, on average.

• But also we can stratify the subjects by gender
—Essentially do two experiments.

• Randomize or stratify ? – BOTH
Within each stratum (gender), randomize.
Issues of cost?

Blank page for your notes

1.5 OBSERVATIONAL STUDIES

- The subjects assign themselves
  Study of cancer: smokers and non-smokers
  Study of income at 40: choice of major at UW
  Study of drug: who keeps to protocol?
- WYSIWYG: Investigators just watch the outcomes!
  Association is not causation.
- How did subjects come to be in treatment/control?
  – identify likely confounding factors.
- Control for known confounding factors – stratify!
  that is, analyze in smaller more homogeneous groups
- Example: Bias in graduate admissions at Berkeley (1973)??
  FPP 17-19
  Choice of major is confounded with gender.
- This is an example of Simpson's paradox.
- Other considerations:
  Can we observe? outcomes, behaviors, but not beliefs or attitudes (contrast with survey).
  Cost? – time is money.
  Observer presence may affect outcome?

Blank page for your notes

1.6 SELECTING A SAMPLE

- We have a population of interest.
  We want to know some parameters of the population.
  But we cannot look at the whole population.

- We select a sample (subset) from the population
  We compute a statistic based on the sample, to estimate the parameter.

- We can choose individuals randomly, or judge what factors may be important.

- If we choose a random sample, we do so carefully to avoid systematic bias

- Unintended selection bias. Non-response bias.

- Large samples do not protect against bias.

- If we first consider some important factors, then
  sample randomly within categories: OK!
  This is stratified random sampling

- If we use these factors to select the sample
  this is quota sampling
  and is subject to unintended biases,
  due to confounding factors associated with selection factors