## 0.1 STAT220: BASIC STATISTICS, WIN 2008

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

| | | |
|---|---|---|
| **Teaching assistants:** | **Aneesh Hariharan,** | **AA, AB** |
| | **Jennifer Chunn,** | **AE, AD** |
| | **Kevin See,** | **AC, AF** |

- **The web page:**
**http://www.stat.washington.edu/courses/stat220/winter08**
**or go to http://www.stat.washington.edu and click on "220" under Winter Quarter Courses**

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

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

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

**NOTE: Schedule is not all updated yet– it will be updated through the quarter.**



## 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; Crime statistics, Traffic statistics**
- **Weather reports; Record highs and lows; precipitation.**
- **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**
- **Note the news EVENT is not a statistical study, – although the event may become part of the statistical record.**

## 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 controls which subjects will receive the treatment. These subjects are 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. FPP P.5-6

- 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
- Compare the rates not the numbers!
- 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.



## SALK VACCINE TRIAL RESULTS: FPP P.6

| The NFIP Study | | | Randomized controlled double-blind experiment | | |
|---|---|---|---|---|---|
| | Size in 1,000s | Rate per 100,000 | | Size in 1,000s | Rate per 100,000 |
| Grade 2 (vaccine) | 225 | 25 | Treatment (vaccine) | 200 | 28 |
| Grades 1,3 (control) | 725 | 54 | Control (placebo) | 200 | 71 |
| Grade 2 (no consent) | 125 | 44 | No consent | 350 | 46 |

- Grades 1,3 may not be good control for Grade 2.
  age, contagion.
- "No consent" is NOT a good control:
  BOTH consent and disease risk are associated with income level.
- In NFIP study, Grades 1,3 contain both consent & no-consent children.
- Note the two non-consent groups have very similar rates
- Note the two treatment groups have very similar rates.
- Assuming groups are "similar", there are ways to figure out whether differences could occur just "by chance'.
- The control groups are quite different:
  NFIP would underestimate effect of the vaccine.

## 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

• Must randomize within eligible group

For example, cannot assign/exclude certain combinations due to severity.

Or consent, as in the NFIP polio trial vs the randomized trial.



Blank page for your notes

## 1.4 BREAST CANCER SCREENING TRIAL: FPP Pp 21-23

First large-scale study of the effectiveness of breast-cancer screening, run by the Health Insurance Plan (HIP) of Greater New York starting in 1963. The subjects were 62,000 women aged 40 to 64. The women were randomly divided into two groups. The control group of 31,000 were offered usual health care, while the treatment group were offered extra screening. Of these, 20,200 came in for the screening tests, but the other 10,800 refused. The numbers of deaths in the first 5 years of the study are shown in the table. The rates in the table are deaths per 1,000 women.

| | | Cause of death | | | |
|---|---|---|---|---|---|
| | | Breast cancer | | All other causes | |
| | | Number | Rate | Number | Rate |
| Treatment group | | | | | |
| Examined | 20,200 | 23 | 1.1 | 428 | 21 |
| Refused | 10,800 | 16 | 1.5 | 409 | 38 |
| Total | 31,000 | 39 | 1.3 | 837 | 27 |
| Control group | 31,000 | 63 | 2.0 | 879 | 28 |

The epidemiologists who worked on the study found:

(i) screening has little impact of diseases other than cancer,

(ii) poorer women were less likely to accept screening than richer ones,

(iii) most diseases affect the poor more heavily than the rich.



(a) Screening saves lives. Which numbers in the table show this?

(b) Why is the death rate from other causes overall in the treatment group about the same as the rate in the control group?

(c) Why is the death rate from other causes higher in the "refused" group than in the "examined" group?

(d) Breast cancer affects the rich more than the poor. Which numbers show this association between breast cancer and income?

(e) To show that screening reduces the risk of death from breast cancer, someone wants to compare the rates 1.1 and 1.5 in the table. Is this a good comparison? Is it biased against screening? For screening?

(f) In the first year of HIP, 67 breast cancers were detected in the examined group, 12 in the refused group and 58 in the control group. True or false and explain: screening causes breast cancer.

## 1.5 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?
- Pretend example – boys/girls in polio trial



Blank page for your notes

## 1.6 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.
- Example: Bias in graduate admissions at Berkeley (1973) FPP 17-19: Numbers of applicants (#) and percent admitted (%).

| Major | Men | | Women | | Combined | |
|---|---|---|---|---|---|---|
| | # | % | # | % | % adm | % women |
| A | 825 | 62 | 108 | 82 | 64.3 | 11.6 |
| B | 560 | 63 | 25 | 68 | 63.2 | 4.3 |
| C | 325 | 37 | 593 | 34 | 35.1 | 64.6 |
| D | 417 | 33 | 375 | 35 | 33.9 | 47.3 |
| E | 191 | 28 | 393 | 24 | 25.3 | 67.3 |
| F | 373 | 6 | 341 | 7 | 6.5 | 47.8 |
| plus other majors | | | | | | |
| Total | 8,442 | 44 | 4,321 | 35 | 40.9 | 33.8 |

- Majors A and B have high acceptance rates, but lower proportions of women applying.



## CONFOUNDING FACTORS AND SIMPSON'S PARADOX

- Choice of major is confounded with gender.
- This is an example of Simpson's paradox.
- Control for known confounding factors – stratify!
  that is, analyze in smaller more homogeneous groups
- In above example, when we analyze the data for each major separately, there is no evidence of bias against women.
- Confounding factors must be associated with both disease or outcome and with exposure
  —with both lung cancer and smoking
  — Berkeley major: with acceptance rate, and with gender.
- Other considerations in observational studies:
  Can we observe? outcomes, behaviors,
  but not beliefs or attitudes (contrast with survey).
  Cost? – time is money.
  Stratification can be expensive: need larger samples.
  Observer presence may affect outcome?

## 1.7 SAMPLING FROM A POPULATION

- 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.

Example: The Gallup and Literary Digest Polls: FPP Pp 334-336

- 1936 Landon vs Roosevelt presidential election:
  population = US voting population.
  parameter = % voting for Roosevelt. (In fact, 62%).
- Literary Digest (LD) sample:
  2.4 million responses to mailing of 10 million postcards.
  selected from phone books and club membership lists.
  Prediction: 43%
- George Gallup's sample:
  Sample of 50,000 people, according to his methods.
  Prediction: 56%. (Predicted correct Roosevelt victory).
  Sample of 3,000 people chosen according to LD method.
  Prediction of LD's prediction: 44%
- Biases in LD's sample:
  Selection bias: Phones (in 1936), Club memberships??
  Income level was a confounding factor – associated with vote and LD's sampling.
  Response bias: – probably not in this case.
  Non-response bias: only 24% response: who responds?



Blank page for your notes

## 1.8 SELECTING A SAMPLE

• A simple random sample (FPP. P.339): everyone has the same chance of being in the sample independently of everyone else.
Taking a simple random sample from a large population is IMPOSSIBLE.

• 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. –both present in 1936 Literary Digest poll.

• 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



Blank page for your notes