Lecture 1

Introduction and Overview

Lecture 1

- Welcome to the class !
- Samples and Populations
- An Overview of Study Designs
- Types of Data

Section A: So, Why Do I Need Biostatistics in My Life?

3

Some May Say…

- "The world is in the midst of a data craze….."
- "This is the era of BIG data"..
  - Genomics
  - Medical Informatics
  - Medical Imaging
  - Internet Usage
- "Data has never been more relevant"..

Statistics!

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

Hal Varian , Google Chief Economist, August 2009

Statistics!

Headline from *Harvard Business Review*:

Data Scientist: The Sexiest Job of the 21st Century[1]

[1] Davenport T, Patil D. Harvard Business Review (October 2012)

Statistics!

Headline from *New York Times*:

For Today's Graduate, Just One Word: Statistics[2]

[2] Lohr S. New York Times (August 2009)

Data Are Everywhere!

- Research results and data are certainly utilized and summarized in the popular media

From *The Baltimore Sun*, 8/23/12:

Elmo makes apples more appealing to kids

"Kids took nearly twice as many apples when they had Elmo stickers on them as when they didn't, researchers from Cornell University said in a letter in the August issue of the Archives of Pediatrics and Adolescent Medicine.

8

Data Are Everywhere!

- From *The NY Times*, 8/23/12:

The Widespread Problem of Doctor Burnout

"Analyzing questionnaires sent to more than 7,000 doctors, researchers found that almost half complained of being emotionally exhausted, feeling detached from their patients and work or suffering from a low sense of accomplishment. The researchers then compared the doctors' responses with those of nearly 3,500 people working in other fields and found that even after adjusting for variables like gender, age, number of hours worked and amount of education, the doctors were still more likely to suffer from burnout.

9

Data Are Everywhere!

- From *The Washington Post*, August 5, 2009:

DC to offer STD Tests in Every High School

"The program conducted last year at eight high schools found that 13 percent of about 3,000 students tested positive for an STD, mostly gonorrhea or chlamydia, according to the D.C. Department of Health. "

10

Data Provides Information

- Good Data Can Be Analyzed and Summarized to Provide Useful Information
- Bad Data Can Be Analyzed and Summarized to Provide Incorrect/Harmful/Non-informative Information

11

Steps in a Research Project

- Planning/Design of Study
- Data collection
- Data analysis
- Presentation
- Interpretation

Biostatistics CAN play a role in each of these steps! (but sometimes is only called upon for the data analysis part)

12

Biostatistics Issues

- Planning/Design of studies
  - Primary Question(s) of Interest:
    - Quantifying information about a single group?
    - Comparing multiple groups?
  - Sample size
    - How many subjects needed total?
    - How many in each of the groups to be compared?
  - Selecting Study Participants
    - Randomly chosen from "master list"?
    - Selected from a pool of interested persons?
    - Take whoever shows up?
  - If group comparison of interest, how to assign to groups?

13

Biostatistics Issues

- Data Collection
- Data Analysis
  - How best to summarize the information coming from the raw data
  - Dealing with variability (both natural and sampling related):
    - Important patterns in data are obscured by variability
    - Distinguish real patterns from random variation
  - Inference: using information from the single study coupled with information about variability to make statement about the larger population/process of interest: What statistical methods are appropriate given the data collected?

14

Biostatistics Issues

- Presentation
  - What summary measures will best convey the "main messages" in the data about the primary (and secondary) research questions of interest
  - How to convey/ rectify uncertainty in estimates based on the data
- Interpretation
  - What do the results mean in terms of practice, the program, the population etc..?

15

Statistical Reasoning 1: Goals

- Summarization
- Measurement of Associations
- Interval Estimation and Statistical Inference
- Sample Size Considerations when Designing A Study

16

Statistical Reasoning 2: Goals

- Adjustment
- Assessing Effect Modification (Statistical Interactions)
- Prediction Using Potentially Multiple Inputs
- Linear, logistic and time-to-event regression

17

Universal Goals

- Throughout all of our endeavors the focus will be on
  - interpreting the results of statistical procedures correctly
  - summarizes the results from published studies in an understandable fashion
  - assessing the strengths and weaknesses of published research results including:
    - ▸ study design
    - ▸ clarity of the research question(s)
    - ▸ appropriateness of the statistical methods
    - ▸ clarity of the reported results
    - ▸ appropriateness of the overall scientific/substantive conclusions

18

Section B: Samples and Populations

19

Learning Objectives

- Upon completing this lecture section you should be able to:
  - Explain the difference between a population and sample (so far as the terms are used in research)
  - Give examples of populations, and of a corresponding sample from a population
  - Explain that characteristics of a randomly selected data sample should imperfectly mimic the characteristics of the population from which the sample was taken
  - Explain how non-random samples may differ systematically from the populations from which they were taken

20

Population Versus Sample

- *Sample :* A subset (part) of a larger group (population) from which information is collected to learn about the larger group
  - For example, twenty-five 18-year-old male college students in the United States
- *Population:* The entire group for which information is wanted
  - For example, all 18-year-old male college students in the United States

21

Random Sampling

- For studies it is optimal if the sample which provides the data is representative of the population under study
  - Certainly not always possible!
- For this term, we will make this assumption unless otherwise specified
- One way of getting a representative sample: simple random sampling
  - A sampling scheme in which every possible subsample of size *n* from a population is equally likely to be selected

22

Random Sampling

- If a sample is randomly selected from a population, the characteristics of the sample should (imperfectly) mimic those of the population
- How can a random sample be obtained?
  - First , a "master list" of the population must be enumerated
  - Using a computer, a random subset of any size can be drawn from the population

23

Population Versus Random Sample

- Generally speaking, with research we want to learn truths in a population, but can only estimate these from an imperfect sample of observations from the population

Population

20% M < 30 years
15% M ≥ 30 years
26% F< 30 years
39% F ≥ 30 years

Random Sample

24

Population Versus Sample: Example 1

- Researchers wanted to learn about the pulmonary health in clinical population of men. There were able to sample 113 men from this population, and measure the systolic blood pressure of each male in the sample.

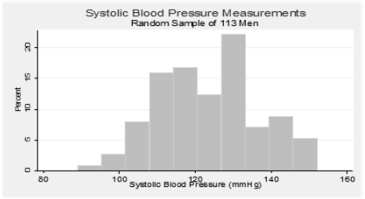
25

Population Versus (Random) Sample: Example 1

- Researchers wanted to learn about the pulmonary health in clinical population of men. There were able to sample 113 men from this population, and measure the systolic blood pressure of each male in the sample.

Sample mean= 123.6 mmHg

Sample sd = 12.9 mmHg



26

Population Versus (Random) Sample: Example 2[2]

- Researchers wanted to characterize the risk of mother to infant HIV transmission (within 18 months of birth). The researchers studied a 183 births to 183 randomly sampled HIV+ women .

27

Population Versus Sample: Example 2[2]

- Researchers wanted to characterize the risk of mother to infant HIV transmission (within 18 months of birth). The researchers studied 183 births to HIV+ women and found that 40 of the children tested positive for HIV within 18 months, for a transmission percentage of 22%

[2]**Connor** E, et al. Reduction of Maternal-Infant Transmission of Human Immunodeficiency Virus Type 1 with Zidovudine Treatment. *New England Journal of Medicine* (1994). 331(18); 1173-1180

28

Other Types of (Non-Random) Samples

- Other types of sampling may be necessary, but may also result in samples whose elements do not reflect he makeup of the populations of interest (bias)
  - Voters (not registered, but those who will actually vote) in the US Presidential Election
  - Intravenous Drug users in Chennai
  - Patients with a Certain Disease
  - Homeless persons in Baltimore
  - Men who have sex with men (MSM) in Malawi

29

Other Types of (Non-Random) Samples

- Other types of sampling may be necessary, but may also result in samples whose elements do not reflect he makeup of the populations of interest

Population

20% M < 30 years
15% M ≥ 30 years
26% F< 30 years
39% F ≥ 30 years

30

Other Types of (Non-Random) Samples

- What kinds of sampling strategies can be employed that may/may not result in a random sample?
  - Voters (not registered, but those who will actually vote) in the US Presidential Election

    Random digit dialing

31

Other Types of (Non-Random) Samples

- What kinds of other sampling strategies can be employed?
  - Intravenous Drug users in Chennai
  - Homeless persons in Baltimore
  - Men who have sex with men (MSM) in Malawi

    Convenience Sampling

    Respondent Driven Sampling

32

Other Types of (Non-Random) Samples

- What kinds of other sampling strategies can be employed?
  - Patients with a Certain Disease

    Random sample from a clinical/hospital population

33

Summary

- Generally speaking, with regards to pubic health and medical research, not all elements of a population can be studied. As such, a sample is taken from the population of interest.
  - Random sampling is the best strategy for getting a sample whose characteristics imperfectly mimic the population
  - However.. random sampling is not always feasible: other approaches can be used, and the sampling procedure needs to be considered when applying the results from the sample to the population

34

Section C: Study Design Types

35

Learning Objectives

- At the end of this lecture section you will be able to:
  - Describe the similarities and differences between the randomized cohort, observational cohort, case-control and observational cross-sectional studies
  - Explain the major analytical challenge that comes from comparing outcomes across groups where the group membership has not been randomized
  - Start to become aware of some of the major issues to consider when making conclusions based on study results (ie: mapping the statistics to the scientific/clinical/substantive)

36

Common Study Design Types

- Prospective Cohort Studies
  - Randomized/controlled study design
  - Observational (Cohort) Studies

  Subjects are classified as to their exposure(s) status at study start, and followed over time to see who develops outcome(s)

- Case/Control Studies

  Subjects are chosen based on their outcome status, and the exposure(s) that occurred prior to outcome are assessed

37

Common Study Design Types

- (Observational) Cross-Sectional Studies

  Everything assessed at the same point in time

38

The Research Process

- The goal of much research is to characterize differences in (sub) populations
  - Does the live polio vaccine reduce the risk of contracting polio?
  - How does the weights of Nepalese children less than a year old differ by sex?
  - Does calorie labeling on restaurant menus result in a reduction in amount of calories consumed?
  - Does AZT reduce the risk of HIV transmission from mother to child?

39

The Prospective Research Process

- The first step is to either:
  - Get a representative sample from the general population under study (A)
  - Get representative samples from the populations under study (ie: the groups to be compared) (B)
- The second step (possibly) is to either assign the sample members to the groups of interest (randomization) or to classify them based on their "self-selected" membership (ex: current smokers vs non-smoker)
- How the second step is done (or if first step A is employed) determines the type of study being done

40

Randomized Trials (Prospective Cohort)

- Important for accounting for many kinds of biases
- Randomization, done correctly on a large number of subjects nearly ensures that the only systematic difference in the groups being compared is the exposure(s) of interest

41

Example 1: Salk Polio Vaccine Trials[1]

- A very famous randomized trial

200, 745 Vaccinated for Polio

≈ 400,000 School Children Randomized

201,229 Given a Placebo

1 Meier, P. The biggest public health experiment ever: The 1954 field trial of the Salk poliomyelitis vaccine. In J. M. Tanur, F. Mosteller, W. H. Kruskal, R. F. Link, R. S. Pieters & G. R. Rising. Statistics: A Guide To the Unknown (1972).

42

1954 Salk Polio Vaccine Trial

- At the end of the follow-up period there were 82 cases in the vaccine group and 162 in the placebo group
- Subsequently analyses report slightly different numbers because some false positives were discovered in each of the two groups

43

Benefits of Randomization

- Randomization helps protect against self selection biases
  - Examples:
    Males more likely to volunteer for placebo than females
    Smokers less likely to be in exposed group
    Healthier persons sign up for the intervention
- The goal of randomization is to eliminate any systematic differences in characteristics of subjects in each of the exposure groups under study, save for the exposure itself

44

Example 2: Hormone Replacement Therapy[2]

- Another very famous randomized trial

8,508 Given Therapy

16,608 Women

8,102 Given Placebo

2 The Women's Health Initiative Study Group. Risks and Benefits of Estrogen Plus Progestin in Health Postmenopausal Women: Principle Results from the Women's Health Initiative Randomized Controlled Trial. (2002) *Journal of The American Medical Association* . 288 (3) 321-333.

45

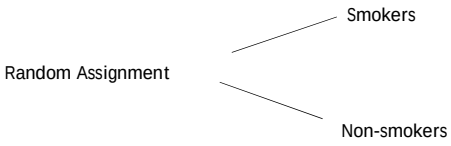
Hormone Replacement Therapy

- At the end of the follow-up period there were 163 CHD cases in the therapy group and 122 in the placebo group

46

Randomization Is Not Always Possible!

- Unfortunately (at least for scientific purposes), you cannot always perform randomized trials!!



47

Non-Randomized Design: Observational Cohort Studies

- Studies in which subjects "self-select" to be in exposure groups: i.e. subjects are not randomized
- Sometimes this is the only type of study that can be done
- Outcome/exposure relationships are of interest
  - Sometimes difficult to directly assess because of selection bias issues which may lead to systematic differences between the exposure groups other then the exposure of interest
  - Examples:
    Smokers more likely to drink alcohol.
    Vegetarians more likely to exercise.

48

Example 3: Needle Exchange and HIV Infection[3]

- New York City: relative risk of HIV infection for intravenous drug users (IVDUS) by needle exchange program participation

As per the authors:

" Interpretation :We observed an individual-level protective effect against HIV infection associated with participation in a syringe-exchange programme. "

[3] Des Jarlais et al. HIV incidence among injecting drug users in New York City syringe-exchange programmes. (1996) *The Lancet. 348. 987-991.*

49

Needle Exchange

- Adjusted for the following . . .
  - Age, gender, race, frequency of injection

50

Example 4: HPV Vaccine and Sexual Activity in Teens[4]

- From the article abstract:

"**RESULTS:** The cohort included 1398 girls (493 HPV vaccine–exposed;905 HPV vaccine–unexposed). ... "

"CONCLUSIONS: HPV vaccination in the recommended ages was not Associated with increased sexual activity–related outcome rates."

[4] Bednarczyk et al. Sexual Activity-Related Outcomes After Human Papillomavirus Vaccination of 11- to 12-Year-Olds. (2012) *Pediatrics.* 130 (5). 798-805.

51

HPV Vaccine and Sexual Activity in Teens

- Results were adjusted for other characteristics of the teens including "health care–seeking behavior and demographic characteristics.".

52

Observational Cohort Studies

- Issues to consider in the analyses of observational studies:

53

Observational Studies

- Sometimes are performed to study results that will then be studied with a follow-up randomized trial

54

Example 5: Beta Carotene and Cancer[5]

- **"Abstract/Background.** Observational studies suggest that people who consume more fruits and vegetables containing beta carotene have somewhat lower risks of cancer and cardiovascular disease, and earlier basic research suggested plausible mechanisms. Because large randomized trials of long duration were necessary to test this hypothesis directly, we conducted a trial of beta carotene supplementation."
- **"Conclusions.** In this *(randomized)* trial among healthy men, 12 years of supplementation with beta carotene produced neither benefit nor harm in terms of the incidence of malignant neoplasms, cardiovascular disease, or death from all causes.

[5] Hennekens C, et al. Lack of Effect of Long-Term Supplementation with Beta Carotene on the Incidence of Malignant Neoplasms and Cardiovascular Disease (1996). *New England Journal of Medicine*.

55

Case-Control Studies: Retrospective Research Process

- In the previously discussed prospective cohort-studies (randomized and observational), the subjects had their exposure status assigned to them, or were selected and then the exposure status was classified: the outcome of interest was assessed over time, after the exposure had occurred
- In situations in which researchers wish to study exposures associated with rare outcomes, it is not necessarily feasible to do a prospective cohort study. Such an approach would require a very large number of enrollees in order to see any outcomes in the samples being compared

56

Case-Control Studies

- A useful alternate approach to a cohort study in this scenario is a case-control study
- In this design, enrollees are selected on whether they have the outcome or not (usually a rare disease), and then exposure(s) is assessed

57

Example 6: Smoking and Lung Cancer[6]

- Another landmark study in public heath/medicine

  The selection of cases (lung cancer patients) and controls (persons without lung cancer) is described in detail in the article.

[6] Doll R and Hill A. Smoking and Carcinoma of the Lung: Preliminary Report, (1950). *British Medical Journal*. pps 739-748.

58

Example 6: Smoking and Lung Cancer

- Another landmark study in public heath/medicine

  " The method of the investigation was as follows: Twenty London hospitals were asked to co-operate by notifying all patients admitted to them with carcinoma of the lung, .." (and several other cancers)

  ".....for each lung-carcinoma patient visited at a hospital the almoners were instructed to interview a patient of the same sex, within the same five-year age group."

59

Example 6: Smoking and Lung Cancer

- Summary of findings, as per the authors

  "Consideration has been given to the possibility that the results could have been produced by the selection of an unsuitable group of control patients, by patients with respiratory disease exaggerating their smoking habits, or by bias on the part of the interviewers. Reasons are given for excluding all these possibilities, and it is concluded that *smoking is an important factor in the cause of carcinoma of the lung*."

60

Case-Control Studies

- Issue to handle in the analyses of results from case-control studies

61

Cross-Sectional Observational Studies

- All information (exposures, outcomes) is assessed at same time point
  - Example: current smoking status, and current flu status
- Many times cross-sectional studies are done to estimate prevalence (proportions of people with a given characteristic)

62

Example 7: Intimate Partner Violence (IPV) and SES[7]

- Analysis of a the British Crime Survey, a nationally-representative cross-sectional study from England, including men and women 16-59 years old

  Snippet from Abstract Results:
  "Lifetime IPV was reported by 23.8% of women and 11.5% of men. Physical IPV was reported by 16.8% and 7.0%, respectively; emotional-only IPV was reported by 5.8% and 4.2%, respectively."

[7] Khalifeh H, et al. Intimate Partner Violence and Socioeconomic Deprivation in England: Findings From a National Cross-Sectional Survey, (2013). *American Journal of Public Health*. 103(3): 462-472.

63

Example 7: Intimate Partner Violence (IPV) and SES

- Analysis of a the British Crime Survey, a nationally-representative cross-sectional study from England, including men and women 16-59 years old

  Conclusion from Abstract :

  "Conclusions. Physical and emotional IPV are very common among adults in England. Emotional IPV prevention policies may be appropriate across the social spectrum; those for physical IPV should be particularly accessible to disadvantaged women."

64

Cross-Sectional Studies

- Issue to handle in the analysis of results from cross-sectional studies

65

Summary

- Types of study designs
- Issue with non-randomized studies

66

Section D: Data Types

67

Learning Objectives

- In this short lecture, a brief summary is given of the types of data that frequently occur in research studies, and will be dealt with analytically in this class (both terms)
- At the end of this lecture section, you should be able to:
  - Distinguish between continuous, binary (and categorical) and time to event data
  - Give examples of each of these aforementioned data types

68

Continuous Data

- Continuous Data *(incremental measurements)*
  - Blood pressure, mmHg
  - Weight , lbs (kgs, oz etc..)
  - Height, ft (cm, in etc..)
  - Age, years (months)
  - Income level, dollars/year (Euro by year, etc..)
- A defining characteristic of continuous data is that a one unit change in the value means the same thing across the entire range of data values

69

Binary Data

- Binary (dichotomous) data: takes on only two values, "yes" or "no"
- Binary (dichotomous) data ("Yes/no" data)
  - Polio :Yes/No
  - Remission :Yes/No
  - Sex : Male/Female (or as yes/no, "is subject male?")
  - Quit Smoking: Yes/No
  - Etc..

70

Categorical Data

- Categorical data : an extension of binary data to include more than 2 possible values
- Nominal categorical data: no inherent order to categories
  - Race/ethnicity
  - Country of birth
  - Religious Affiliation
- Ordinal categorical data: order to categories
  - Income level categorized into four categories, least to greatest
  - Degree of agreement, five categories from strongly disagree to strongly agree

71

Time to Event Data

- Data that are a hybrid of continuous data and binary data
  - Whether an event occurs and time to the occurence (or time to last follow-up without occurrence)

72

Different Statistics for Different Data Types

- To compare blood pressures in a clinical trial evaluating two blood pressure-lowering medications, you could:
  - Estimate the mean difference in blood pressure change (after-before) between the two treatment groups
  - Estimate a 95% confidence interval and/or use a t-test to test for population level differences in the mean blood pressure change

73

Different Statistics for Different Data Types

- To compare the proportion of polio cases in the two treatment arms of the Salk Polio vaccine, you could:
  - Estimate the difference in proportions (risk difference) and ratio of proportions (relative risk, risk ratio)
  - Estimate 95% confidence intervals and/or use a chi-square test to test for population level differences in these quantities

74

Different Statistics for Different Data Types

- To compare differences in time to contracting HIV between HIV negative IV drug users in a needle exchange program and HIV negative IV drug users not enrolled in a needle exchange program, you could:
  - Estimate an incidence rate ratio for contracting HIV that compares these two groups
  - Construct a Kaplan-Meier curve for each group to provide a graphical description of the time to HIV profile for each group
  - Estimate a 95% confidence interval for the incidence rate ratio and/or use a log-rank to test for a population level

75