

BIOS 600: Principles of Statistical Inference

Introduction, Reproducibility, and Types of Data

Fall 2012

Reading (on Sakai)

- ▶ Perry, 'What is the role of biostatistics in modern medicine?',
Discovery Health
- ▶ AmStat News Article on Reproducible Research
- ▶ Why all the fuss about reproducibility? (NY Times Article)
- ▶ Pagano and Gauvreau, Chapters 1 and 2.1

What is Biostatistics?

Biostatistics is the science of obtaining, analyzing and interpreting data in order to understand and improve human health.

Biostatisticians forge advances in science that benefit human health through innovations in biostatistical methodology and theory as well as the thoughtful implementation of biostatistical methods in practice.

Welcome to BIOS 600!

What have I gotten into?

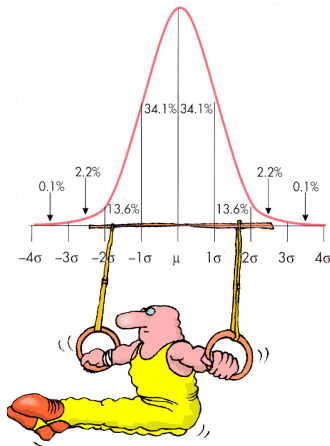
BIOS 600

- ▶ is an introductory course in probability theory and statistical inference
- ▶ provides a tour of basic statistical methods commonly encountered in public health and biomedical research
- ▶ places emphasis on understanding of basic statistical methods, use of the methods to evaluate evidence from studies, and communication of statistical results to non-statisticians
- ▶ requires use of standard statistical software (e.g., Stata)

Learning Objectives

By December, students successfully completing the course will

- ▶ have a basic working knowledge of important statistical topics including descriptive statistics and probability, inference on means and proportions, regression methods, and nonparametrics
- ▶ understand how to evaluate which methods are appropriate in answering a research question for a given study design
- ▶ be able to evaluate data using modern statistical software and interpret analysis results



Learning Objectives

By December, students successfully completing the course will

- ▶ be able to evaluate straightforward statistical usage in public health and medicine, with a focus on relevant research publications
- ▶ have the tools to interact knowledgeably with biostatisticians in planning, conducting, analyzing, and reporting public health and medical research (and know how to determine when a biostatistician should be consulted)



Strategies for Success

- ▶ Keep up with the readings (they were assigned for a reason!)
- ▶ Embrace statistical computing (computers rarely bite!)
- ▶ Attend all lab sessions, which will reinforce concepts and computing skills
- ▶ DO NOT violate the honor code
- ▶ ASK EARLY if you feel you may fall behind!
- ▶ Read syllabus carefully for important information

Reproducibility

Reproducibility is one of the primary components of the scientific method. It is the ability of an entire experiment or study to be reproduced by an individual investigator. In biostatistics, it is critical to keep careful records so that an entire analysis, from defining variables to analytical results, can be independently reproduced. Safe storage of datasets and retention of clearly-documented computer code are essential components of reproducibility.

Reproducibility

Duke University and other research units were rocked by the recent scandal involving a series of cancer clinical trials that determined genomic signatures of tumors to select the most effective chemotherapy.

A number of errors were found with the research, and many of these were very basic – such as switching the label on a 0/1 variable (so that if 0 was treatment A and 1 was treatment B, they interpreted 0 as treatment B and 1 as treatment A, so that each patient was assigned the *least* effective chemotherapeutic agent) and reading data incorrectly by attaching the wrong column labels (they were off by one column).

The researcher had a number of opportunities to fix these problems but failed to retract or correct all the papers.

Reproducibility

Eventually (after another publication determined the researcher lied about having obtained a Rhodes Scholarship), the studies were stopped and the researcher resigned from Duke.

Don't let this be you!

What are strategies you can use to enhance reproducibility?

Big Picture

Statistics is the process by which we convert data into useful information. As part of this process, we

- ▶ collect data
- ▶ summarize data
- ▶ interpret the results

What is the Population?

First we identify the group we would like to learn about. This group is called the *population*. This population could be, for example, all babies born in sub-Saharan Africa, all breast cancer patients in North Carolina, or all adults in the United States.



Graphic from the CMU Open Learning Initiative.

Sampling from the Population

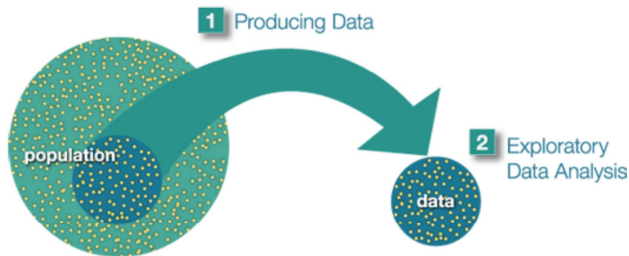
Suppose we wish to know what adults in the U.S. think about universal health care. It would be virtually impossible to ask every adult in the U.S. this question (though the Census still tries to do this, despite good advice!). Usually we have to compromise by taking a *sample* of people from the population for further study. We have to be careful that our *sample* is a representative one – for example, we would have biased results if our sample consisted only of Tea Party members.



Graphic from the CMU Open Learning Initiative.

Collecting Data

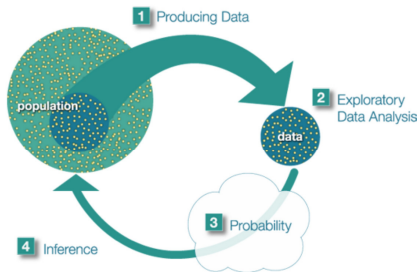
Suppose our sample consists of 5000 U.S. adults, and we ask each adult whether or not they support universal health care. We'll then need to take the 5000 answers and summarize them in some way. For example, we can calculate the % of those in our sample who support universal health care – say it's 65%.



Graphic from the CMU Open Learning Initiative.

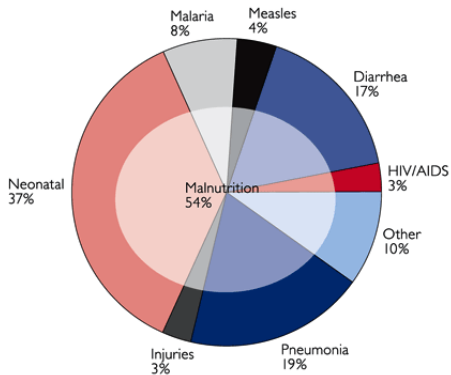
Drawing Conclusions

We then use *probability* and *inference* theory to help us determine whether there is significant support of universal health care and to characterize the uncertainty in our estimate (65%). We then draw conclusions about our original population (all U.S. adults)



Graphic from the CMU Open Learning Initiative.

Types of Data



Nominal Data

Classify into named categories without numeric meaning, e.g.

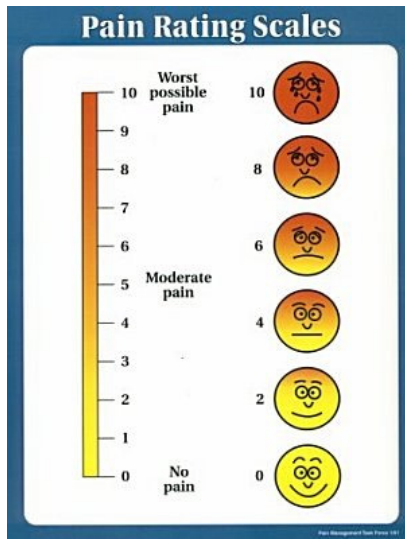
- ▶ biological sex (0=male, 1=female; 'M'=male, 'F'=female) – this variable is binary or dichotomous (two possible values)
- ▶ health insurance provider (0=uninsured, 1=medicaid, 2=medicare, 3=private)
- ▶ blood type (A, B, AB, O)
- ▶ whether or not you have colon cancer
- ▶ also called categorical data

Ordinal Data

Categories are ordered, but differences between values not easily measured; only relative comparisons are made about differences between levels

- ▶ Colon cancer stage 0, I, IIA, IIB, IIC, IIIA, IIIB, IIIC, IVA, IVB
- ▶ Likert scale: 5=strongly agree, 4=agree, 3=neutral, 2=disagree, 1=strongly disagree

Ordinal Data



Rank and Count Data

Count data: counted observations

Rank data: ranked from least to greatest or greatest to least

Causes of Death for Children Under Age 5 Globally

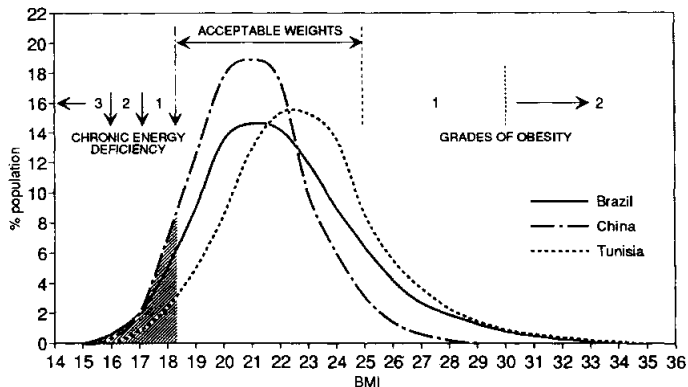
Cause	Count	Rank
Diarrhoeal diseases	1,064,000	2
HIV/AIDS	152,000	7
Injuries	228,000	6
Malaria	684,000	4
Measles	76,000	8
Neonatal death	3,040,000	1
Noncommunicable diseases	304,000	5
Pneumonia	988,000	3
Other	988,000	
Total	7.6 million	

Over 70% of these deaths are in Africa and south-east Asia.

Source: World Health Organization

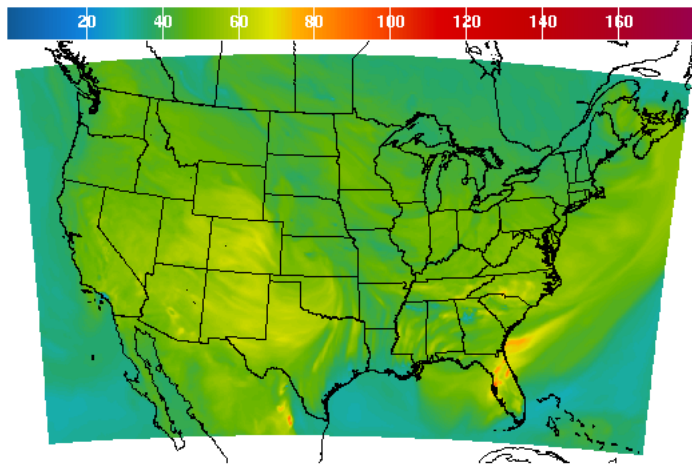
Continuous Data

Data representing measurable quantities in which the difference between any two possible data values can be arbitrarily small, e.g. birth weight, ppm ozone, BMI



* The hatched area represents the percent distribution likely to have CED (i.e. BMI < 18.5)

Continuous Data



1Hr Avg Ozone Concentration(PPB) Ending Fri Apr 27 2012 5PM EDT
(Fri Apr 27 2012 21Z)



National Digital Guidance Database

06z model run

Graphic created-Apr 27 8:19AM EDT



Example: Diet Beverage Consumption and Metabolic Syndrome

UNC Nutrition professors Kiyah Duffey & Barry Popkin recently published work linking consumption of diet beverages to metabolic syndrome (*AJCN*, 2012). What data types (nominal, ordinal, count, rank, continuous) were used?

- ▶ Metabolic syndrome (yes/no)
- ▶ Diet beverage consumption (servings per day)
- ▶ Diet beverage consumption (consumer or nonconsumer)
- ▶ Dietary pattern (prudent or Western)



Example: Diet Beverage Consumption and Metabolic Syndrome

- ▶ Race (black or non-black)
- ▶ Biological sex (male or female)
- ▶ CARDIA study center (Birmingham, Oakland, Minneapolis, or Chicago)
- ▶ Baseline age (years)
- ▶ Highest attained education (years)



Example: Diet Beverage Consumption and Metabolic Syndrome

- ▶ Smoking status (former, current, never)
- ▶ Total energy consumption (kcal/day)
- ▶ Physical activity level (exercise units per week)
- ▶ Body mass index (kg/m^2)
- ▶ Body mass index (underweight, normal weight, overweight, obese)
- ▶ Family structure (single without kids, married without kids, single with kids, married with kids)



Reading for Next Time

- ▶ Pagano and Gauvreau, remainder of Chapter 2
- ▶ Warm-up activity: spread of a particularly virulent form of the plague