

## Introduction

|               |                            |   |
|---------------|----------------------------|---|
| <b>Topics</b> | 1. Why Biostatistics ..... | 3 |
|               | 2. Course Overview .....   | 9 |

## 1. Why Biostatistics

A variety of settings illustrate the need for and use of biostatistics.

### Example – Genetic Counseling

A couple has a baby with a genetic defect.

They are considering having another baby.

What is the likelihood that the second child will have a genetic defect also?

### Example – Prognosis

A physician is considering several therapies for the treatment of a patient.

Which therapy should be used?

Each therapy produces a result that is somewhere between success and failure.

The final choice is “weighed” against the others.

**Probabilities are a tool in decision making.**

### Example – Federal Drug Testing

Is a food additive carcinogenic?

An investigator explores this in an experiment that compares two groups.

Only some of the controls develop cancer.

Only some of the treated individuals develop cancer.

Is the excess number of cancers among treated individuals meaningful?

### **Example – Smoking and Cancer**

Lung cancer occurs only sometimes.

It is not an invariable consequence of smoking.

Interest is identifying the factors related to a variable outcome.

**Biostatistical inference about associations is not equivalent to the understanding of deterministic phenomena.**

### **Example – Justice versus Medicine**

In the judicial system, we say “innocent until proven guilty”

- We err in the direction of “letting go free” a guilty person.

In the practice of medicine, we say it is “better to order another test”

- We err in the direction of suspecting disease.

**Accepted and known biases influence decision making**

### Example – Investigation of the Portacaval Shunt

*Source:* Grace, Muench, Chalmers (1966) summarized the findings in over 50 studies. These were then classified according to study design.

| <u>Design</u>            | <u>Reported Enthusiasm for Shunt</u> |          |      |
|--------------------------|--------------------------------------|----------|------|
|                          | Marked                               | Moderate | None |
| No controls              | 24 (75%)                             | 7        | 1    |
| Observational Controlled | 10 (67%)                             | 3        | 2    |
| Randomized Trial         | 0 (0%)                               | 1        | 4    |

Since 1966, we have seen the increasing use of randomization designs.

**Unknown biases influence decision making**

### Example – Is living near electricity transmission equipment associated with occurrence of cancer?

|      | Cancer | Not  |     |
|------|--------|------|-----|
| Near | 200    | 1646 | 11% |
| Not  | 50     | 7289 | 1%  |

Among those living near electricity equipment, 11% have cancer.

Among those living elsewhere, only 1% have cancer.

Is this a meaningful difference?

Suppose we control for asbestos exposure. Within each group, all persons have “similar” levels of exposure.

#### Exposed to Asbestos

|      | Cancer | Not |     |
|------|--------|-----|-----|
| Near | 194    | 706 | 22% |
| Not  | 21     | 79  | 21% |

#### Not exposed to Asbestos

|      | Cancer | Not  |      |
|------|--------|------|------|
| Near | 6      | 940  | 0.6% |
| Not  | 29     | 7210 | 0.4% |

Controlling for asbestos exposure eliminates the apparent relationship.

Is exposure to asbestos associated with cancer?

Let’s look at this, controlling for proximity to transmission equipment.

#### Residence Near Transmission Equipment

|          | Cancer | Not |      |
|----------|--------|-----|------|
| Asbestos | 194    | 706 | 22%  |
| Not      | 6      | 940 | 0.6% |

#### Residence Not Near Transmission Equipment

|          | Cancer | Not  |      |
|----------|--------|------|------|
| Asbestos | 21     | 79   | 21%  |
| Not      | 29     | 7210 | 0.4% |

Asbestos exposure is associated with cancer, regardless of location of residence.

What happened?

Persons living near transmission equipment and who were exposed to asbestos were more likely to be sampled than were people living near transmission equipment who were not exposed to asbestos.

**Biased sampling can lead to spurious findings.**

## Biostatistics is a Tool

The information available to us is often incomplete. Decision making then requires some kind of evaluation of probability.

- ◆ Statistical methodologies are tools for managing these issues

The goal is to inform decision making

- Family planning
- Patient care
- ◆ Tobacco and lung cancer (Experiment)
- ◆ Tobacco and lung cancer (Observation)

Uncertainty is not necessarily approached objectively. Norms of decision making influence our decision making. We want a fair decision. Or, we want a decision that is faithful to agreed upon priorities of judgment.

- Judicial system
- ◆ Diagnostic testing
- ◆ Type I, II error

As well, we want a fair decision when we are not knowledgeable of our biases.

- Portacaval shunt
- We don't do science in a vacuum

Investigators must consider as fully as possible all of the factors which might be related to the observed outcomes.

- The transmission equipment, asbestos, cancer example
- Experimental design

The tools of biostatistics are of two types:

- **Description** – we use summaries to understand a population
- **Inference making** – we wish to compare competing hypotheses

**Example**

In 1969, the average number of serious accidents per 1000 workers per year in a large factory was 10. In 2004, the average number of serious accidents per 1000 workers per year in the same factory was 7. Is the downward trend from 10 to 7 real or a reflection of natural variation?

**Example**

The spaceship Voyager 2 is circling the planet Uranus. What is the “blip” on our radio receiver here on earth? Is it a true signal? Or, is it random noise such as cosmic rays, magnetic fields, or whatever?

The “signal-to-noise ratio” concept is useful in epidemiology:

**Signal - Treatment effect, Exposure effect, Secular trend**

**Noise - Natural variation, Random error**

*Random error is the “noise” in the “signal-to-noise ratio” analogy.*

---

**Descriptive Statistics****Inferential Statistics**

---

**Example:** Among 573 cholesterol values, what is a typical value?

**Example:** Is exposure to VDT during pregnancy associated with adverse outcomes?

**Solution:** Confidence interval for the mean.

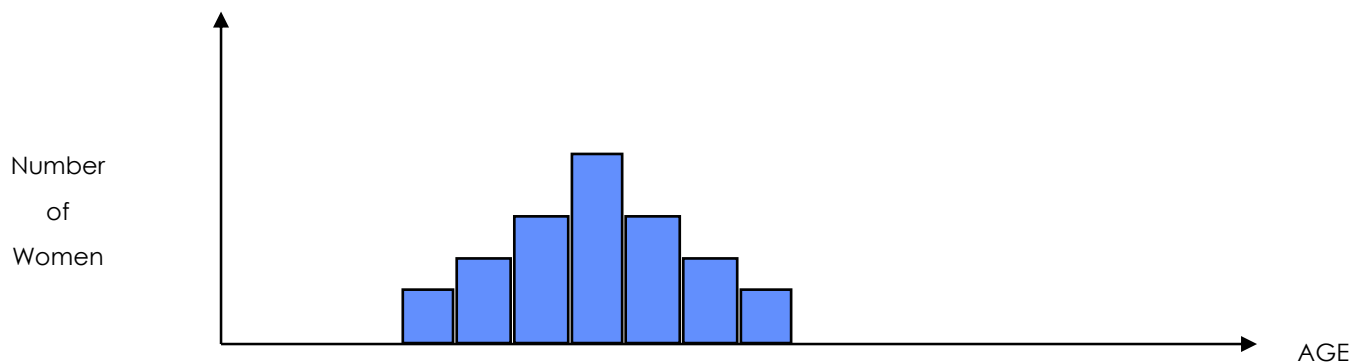
**Solution:** Two sample test For equality of rates.

## 2. Course Overview

### Topic 1 - Summarizing Data: Descriptive Statistics and Graphical Presentation

These techniques enable us to condense a great amount of data into an easily digested format.

**Eg:** Suppose we had the ages of 573 women visiting a prenatal care clinic. If someone were interested in this information he/she wouldn't be overjoyed to have a statistician hand him/her a list of 573 numbers. Instead, computing the average age, range of ages, or drawing a picture gives an easily understood summary of the ages of these women.



### Topic 3: Populations, Samples and Probability Distributions

Statistical Inference is the theory and methodology for generalizing from a sample to a population.

We will discuss the principles that allow us to use information from a small subgroup to reach conclusions about a larger group of interest.

#### Some Commonly Used Terms and Notation:

Population:

Entire Group of Interest  
 $N = \#$  in population

Sample:

Small subset of population  
 $n = \#$  in sample

parameter: summary  
measure of population,  
often denoted by Greek  
letter (i.e., mean =  $\mu$ )

statistic: summary  
measure of sample values,  
(i.e., sample mean =  $\bar{X}$ )

Very often a sample of size  $n$  is drawn from the  $N$  individuals in a population. On the basis of this sample we make inferences concerning the population.

**eg:** Since blood tests are costly to administer, a sample of  $n=20$  children were selected from the  $N=293$  of a particular school. These 20 were given the test and, based on their results, a statement is made concerning the blood levels of all 293 children in the school.

**Note:** If a sample is NOT drawn in an appropriate manner from a population, it may not be representative of that population. In that case, results from the sample may not be generalizable to the population.

We will discuss appropriate methods of sampling as well as the basics of probability that allow us to make statements about the likelihood of selecting a particular sample.

Topics 6 and 7:  
Estimation and Hypothesis Testing

We will apply the principles of statistical inference in a variety of common settings:

I. The One Sample Problem

Eg: Interest lies in making a statement about one population (the 293 children in the school) based upon a single sample. Interest may lie in estimating the average level of the blood test, or the amount of variation among the children. Or interest may lie in deciding whether or not the average level is above some specific value (hypothesis testing).

II. The Two Sample Problem

Suppose a sample of size  $n_1$  is drawn from one population and a sample of size  $n_2$  is drawn from another. On the basis of these two samples, statements are made concerning the comparability of the two populations.

eg: A sample of 25 students is taken from the 220 students at another school. These 25 were given the blood test as above, and, based on a comparison of the results from the two samples, a statement is made concerning the difference in blood levels of the children from the 2 schools.

## The Paired Data Problem

**eg:** Suppose a new drug is manufactured for lowering blood pressure. How do we determine if the drug does what is claimed?

| <u>Subject</u> | <u>Blood Pressure</u> |              | <u>Difference</u> |
|----------------|-----------------------|--------------|-------------------|
|                | <u>Before</u>         | <u>After</u> |                   |
| <b>1</b>       | $x_1$                 | $y_1$        | $x_1 - y_1 = d_1$ |
| <b>2</b>       | $x_2$                 | $y_2$        | $x_2 - y_2 = d_2$ |
| ...            |                       |              |                   |
| <b>n</b>       | $x_n$                 | $y_n$        | $x_n - y_n = d_n$ |

Blood pressure measurements are taken on  $n$  subjects before they start taking the new drug, and again on the same subjects after 2 weeks use of the new drug.

If the drug is successful we expect the average within-subject difference, before minus after, to be positive.

$$\text{i.e., average of } (x_i - y_i) > 0$$

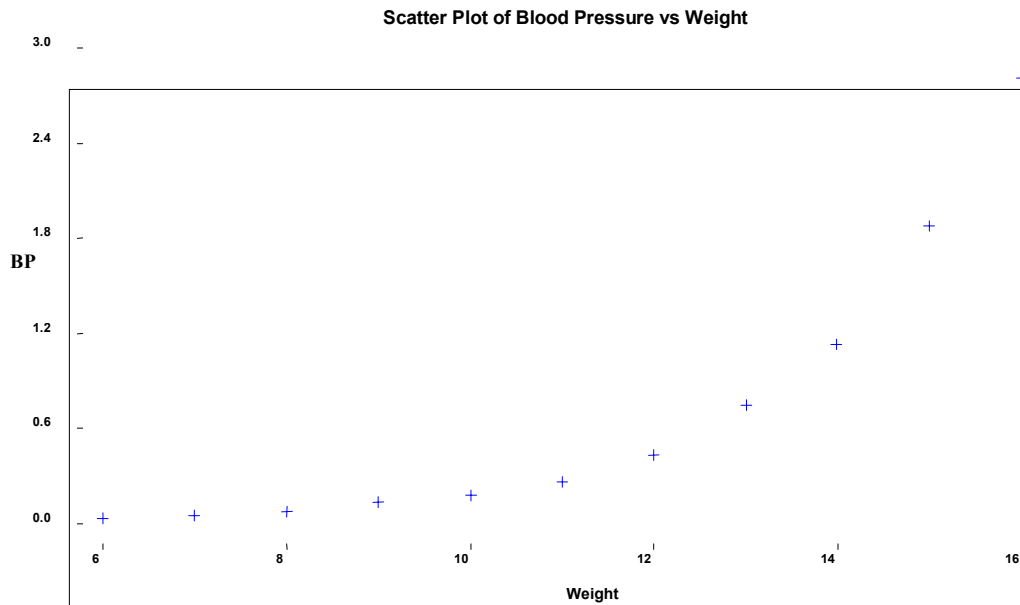
indicating that there was a drop in blood pressure.

**Note:** We'll learn when to conclude that an observed difference is "statistically significant" and we'll learn that statistical significance is not the same as biological significance.

## Topic 9: Association, Correlation and Regression

We are often interested in the relationship among several variables computed on the same individual.

eg: Is there a relationship between weight and blood pressure?



eg: Is there any relationship between smoking and death from heart attack?

|            | Died of Heart Attack | Died of Other Cause |
|------------|----------------------|---------------------|
| Smoker     | <b>a</b>             | <b>b</b>            |
| Non-smoker | <b>c</b>             | <b>d</b>            |

n

## Summary

- **Biostatistics should be informed by nature.**
  - **The signal-to-noise analogy is useful.**
  - **Statistical inference does not confer biological inference.**
  - **Meaningful inference requires the intertwining Of design and analysis.**
- **We're not certain, nor Objective, nor expert**
  - **The generic test statistic Is an expression of signal/noise**
  - **An isolated p-value is "blind" to influences of Selection, mechanism**
  - **Appropriate conclusions take into account nature.**