

## Course Introduction

<b>Topics</b>		
	1. Why Biostatistics .....	2
	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

One goal is to **inform decision making**, as in the examples described in previously:

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

Uncertainty is not necessarily approached objectively. We bring to decision making settings priorities of judgment. Some of these are in our awareness. Others are not. Some of the examples described previously where priorities of judgment *are known* are the following:

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

An example where the influences are *not* necessarily in our awareness is the following:

- Portacaval shunt

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 2009, 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.*

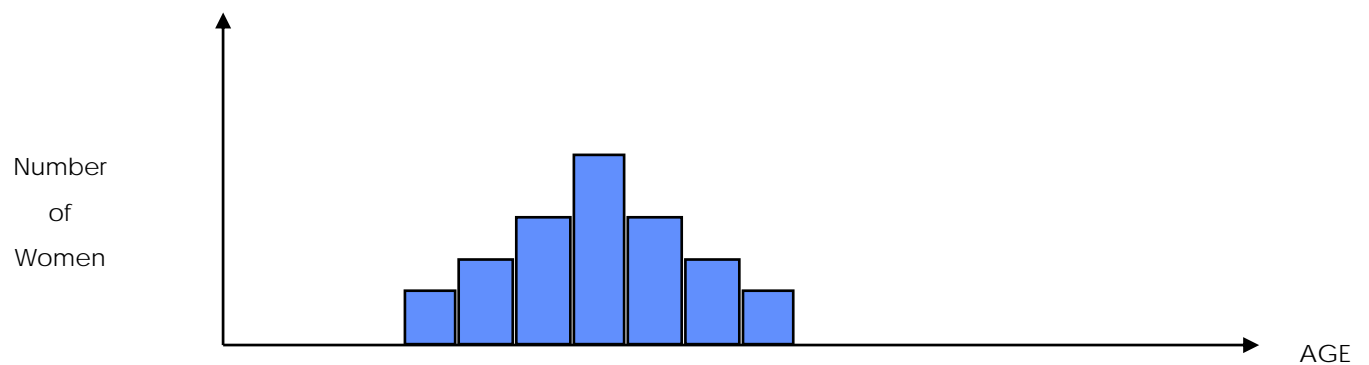
<b>Descriptive Statistics</b>	<b>Inferential Statistics</b>
<b>Example:</b> Among 573 cholesterol values, what is a typical value?	<b>Example:</b> Is exposure to video display terminals (computer monitors) associated with adverse health outcomes?
<b>Solution:</b> Confidence interval for the population mean, obtained using data in a random sample.	<b>Solution:</b> Two sample test of equality of rates of adverse outcome, obtained using data from two random samples.

## 2. Course Overview

### Unit 1 - Summarizing Data

In this unit, you will learn methods for graphical and numerical summarization of data. These techniques enable us to condense a great amount of data into an easily digested format.

**Example -** 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.



## Unit 2 - Introduction to Probability

In this unit, you will gain an appreciation of some ideas of chance (eg – the chances of a fair coin landing “heads” is 0.50) and the basics of calculating probabilities. This understanding is useful when asking questions such as

- What are the chances that a person with a positive test result is truly diseased? (*diagnostic testing*)
- What were the chances that the treatment group, relative to the control group, experienced the better response that was observed under the assumption that the treatment and control therapies are equivalent? (*clinical trials*)

**Example -** Suppose it is known that the probability of a positive mammogram is 80% for a woman with breast cancer and is 9.6% for a woman without breast cancer. Suppose further that, in the general population, the chances of breast cancer is 1%.

If we are told that an individual patient is known to have a positive mammogram, we can use an approach known as **Bayes Rule** to solve for the probability that she is truly diseased. As we shall see, the answer in this example is 7.8% likelihood.

### Unit 3 - Populations and Samples

In this unit, we will discuss the principles, and conditions, under which we can generalize conclusions about a sample to inferences about a population.

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

#### Some Commonly Used Terms and Notation:

**Population:**

Entire Group of Interest

$N = \#$  in population (if finite)

**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 =  $\bar{X}$ )

In this unit, you will be introduced to the idea of drawing a **simple random sample** of size  $n$  from a finite population of size  $N$ . You will also learn that if a sample is *not* obtained in an appropriate manner (based on a probability model), then it may not be possible to generalize findings from analysis of the sample to inferences about the population.

**Example** - Since blood tests are costly to administer, a simple random 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.

## Units 4 and 5 - Bernoulli and Binomial Distribution Normal Distribution

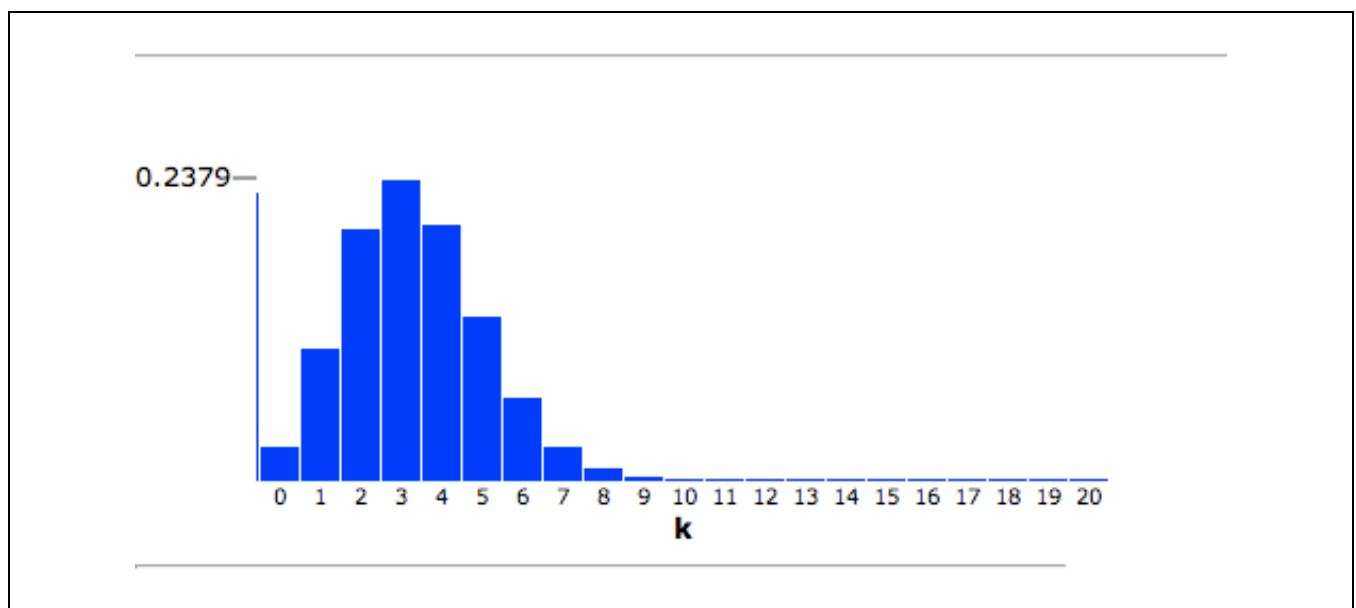
The pattern of occurrence of many phenomena in nature can be described well using some known probability models. In units 4 and 5, you will be introduced to three probability models: Bernoulli, binomial, and normal.

The **Bernoulli** probability model is useful for describing the pattern of discrete outcomes of a single trial in which there only two possible outcomes (eg – “success” or “failure”).

**Example** - The pattern of outcomes of tossing a fair coin one time is “heads” which occurs with probability 50% and tails which occurs with probability 50%

The **Binomial** probability model is useful for describing the pattern of discrete outcomes of a multiple number of trials where, for each, there only two possible outcomes (eg – “success” or “failure”).

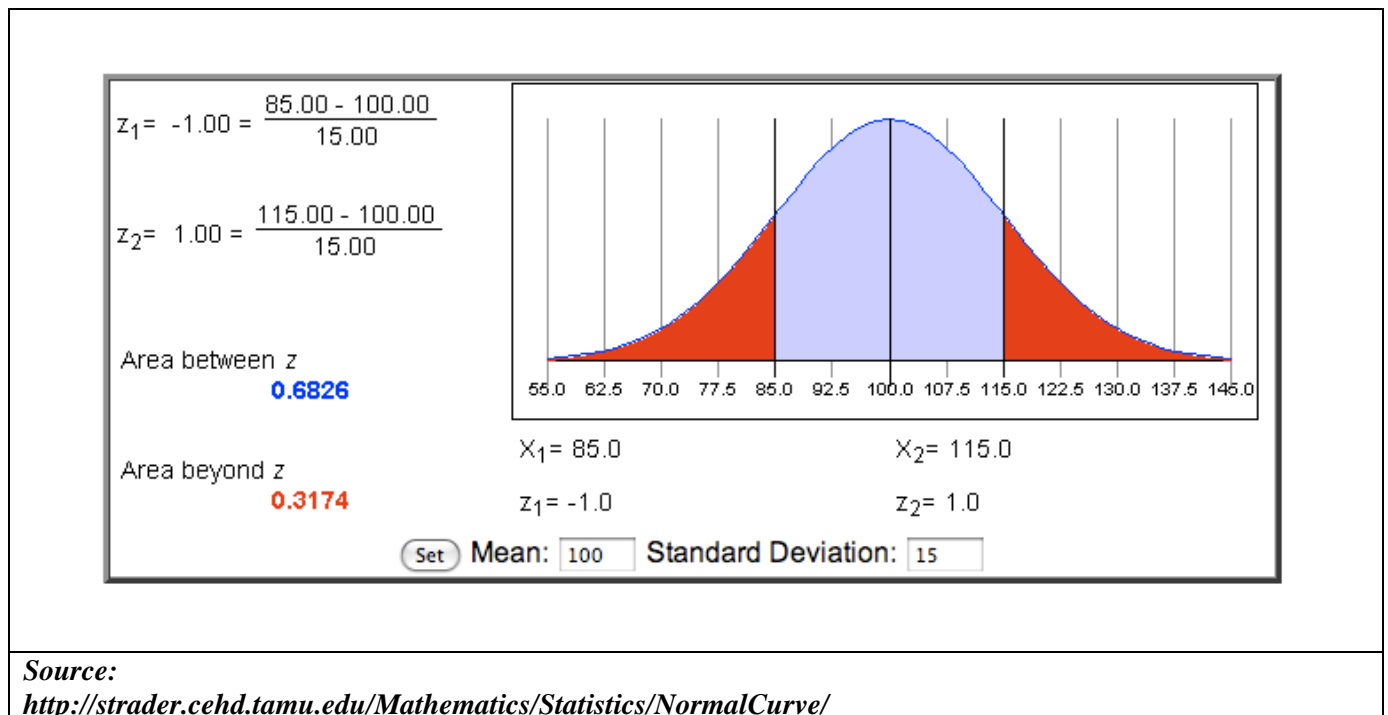
**Example** - The probability of rolling a six with a single die is 16.67%. Suppose you roll the single die 20 times. The probabilities of obtaining 0 sixes, 1 six, 2 sixes, etc, is an example of the binomial probability distribution. A graph of this probability distribution is the following. On the horizontal axis, “k” indexes the number of sixes obtained and ranges from 0 to 20. Plotted on the vertical axis is the probability that 20 rolls of a single die yields “k” sixes.



Source:  
<http://faculty.vassar.edu/lowry/binomial.html>

The **Normal** (also called **Gaussian**) probability model is useful for describing the pattern of outcomes that have values on a continuum (eg – cholesterol measurements have values that lie on a continuum)

**Example** - The pattern of scores on a standard IQ test is well described by a normal distribution. The population average IQ score is 100 and this locates the center of the normal distribution. The standard deviation (more on this later is 15). Thus we might say, “A typical IQ score is 100 give or take 15”.



**Units 6 and 7 -  
Estimation  
Hypothesis Testing**

In units 6 and 7, you will learn how to apply the principles of biostatistics (description and inference) in a variety of selected (and very common) settings. You will learn when to conclude that an observed difference is “statistically significant”. You will also learn the distinction between “statistical significance” versus “biological significance”.

**One Sample Setting**

**Example** – It is of interest to understand the nature of a single population of 293 children in a particular school based upon a single sample of size “n”. We might be interested in estimating the average level of the blood test, or the amount of variation among the children. Or, we might be interested in assessing (hypothesis testing) whether or not we can reasonably infer that the average level is above some specific value.

**Two Sample (Independent Groups) Setting**

Suppose a simple random sample of size  $n_1$  is drawn from one population and a simple random sample of size  $n_2$  is drawn from a second, independent, population. On the basis of the information in these two samples, we seek to make some inferences concerning the comparability of the two populations.

**Example, continued** -A simple random sample of 25 students is taken from the 220 students at a second, independent, school. These latter 25 were given the blood test as above. Using techniques of statistical hypothesis testing, a conclusion is drawn regarding the similarity of the blood levels at the two schools.

**Two Sample (Paired Data) Setting**

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

Subject	Blood Pressure		Difference
	Before	After	
1	$x_1$	$y_1$	$x_1 - y_1 = d_1$
2	$x_2$	$y_2$	$x_2 - y_2 = d_2$
...			
n	$x_n$	$y_n$	$x_n - y_n = d_n$

**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.**

## Unit 8 - Chi Square Tests

In unit 8, you will extend the ideas of statistical hypothesis testing to the setting of outcomes that are discrete.

**Example** – Suppose smoking history is measured using an instrument with possible values of “yes” or “no”. Suppose we have information on cause of deaths and, in particular, whether or not the cause of death was a heart attack. A chi square test would be used to address the question - *Is there any relationship between smoking and death from heart attack?*

The data available to us would be in the form of a 2x2 table that has the following standard format and layout. The “a”, “b”, “c” and “d” represent counts. Thus, in this example of n deaths, we observe “a” deaths due to heart attack in smokers.

	Died of Heart Attack	Died of Other Cause
Smoker	a	b
Non-smoker	c	d

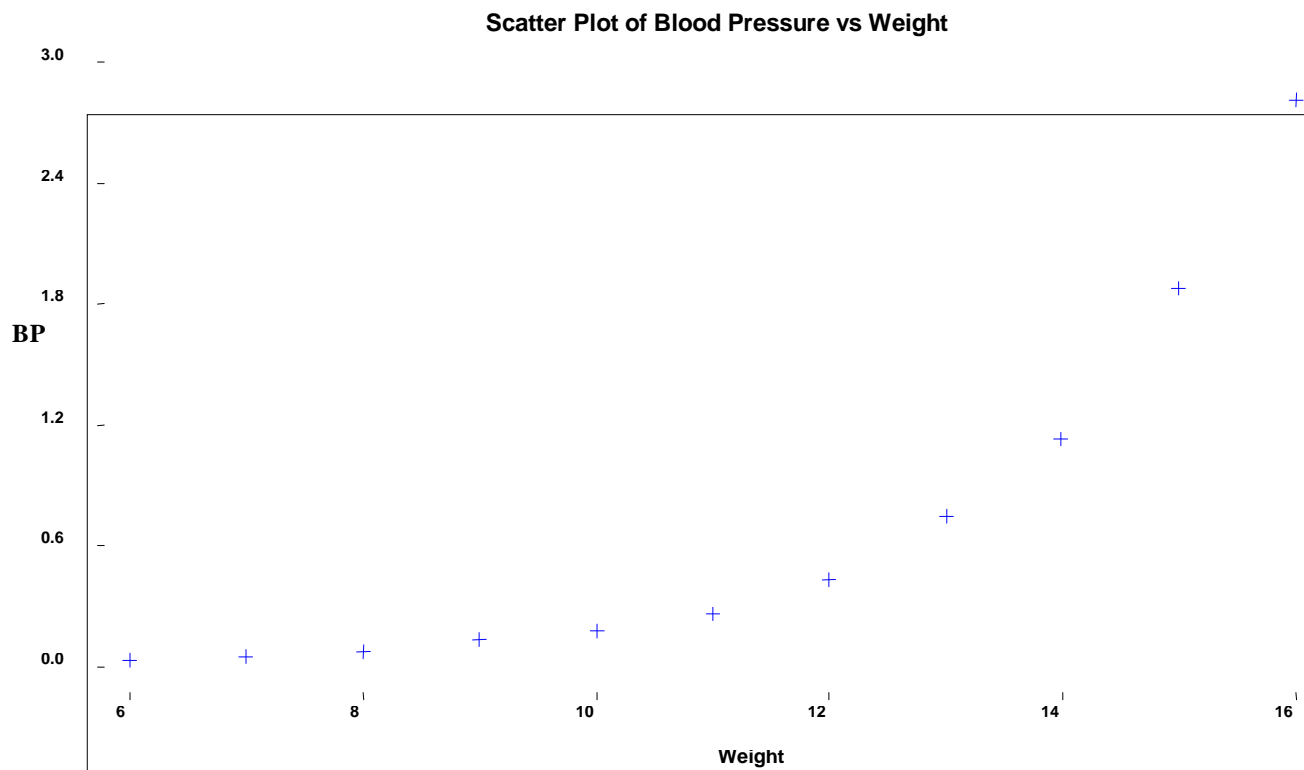
n

## Unit 9 - Regression and Correlation

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

In unit 9, you will be introduced to the ideas of simple linear regression and correlation in the setting of a single predictor variable measured on a continuum and a single outcome variable that is also measured on a continuum. In this setting, we will also assume that the pattern of values of the outcome variable is distributed normal.

**Example** - Is there a relationship between weight and blood pressure?



## 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.**