Course Introduction

“Very true,” said the Duchess: “flamingoes and mustard both bite. And the moral of that is – “’Birds of a feather flock together. ‘ “
“Only mustard isn’t a bird,” Alice remarked.
“Right, as usual,” said the Duchess: “what a clever way you have of putting things!”

- Alice in Wonderland

This is an outline of the entire course, using a “course roadmap.” It begins with an introduction to statistical literacy is introduced using several examples. Note how we are often poor at evaluating probability! A brief overview of each unit is provided.

Table of Contents

1. Course Roadmap

Nature

Nature is full of variation. Variation might be from time to time, from person to person, or from one repeated measurement to the next. Or, it might be from one treatment to the next or from one exposure to the next. Which variation is **"real"**, representing differences that are systematic in some knowable way, and which variation is **"noise",** representing what we call natural or random variation which we think of (for the moment anyway) as "noise" ? Do we even know what we're talking about when we distinguish "real" from "random"?

Populations/ Sample

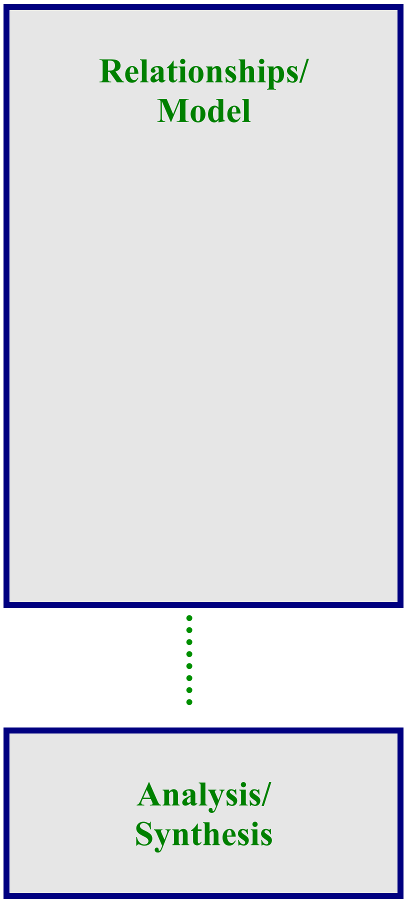
A ***population*** is a class of individuals. An example is the collection of individuals who voted in the 2012 U.S. presidential election. Numerical facts about a population are called ***parameters***. If we could study a population by examining each and every member, we would be doing a census. This course is not about censuses.

More often, what we can examine is only a part of a population; that part is called a ***sample***. Numerical facts about a sample are called ***statistics***. Statistics from a sample are used to make generalizations to the population (provided the method of sampling was appropriate). This is called inference.

Observation/ Data

Observation and **data** may not be the same. What does your mind's eye "register" when you ***observe*** a flower? You might describe the flower as red, with 5 petals, and having a strong aroma. "Red", "5 petals", "strong aroma" are your ***data***. Data are the result of selection (which attributes of the flower matter to you in the first place?) and measurement (what value scheme are you using?). Just think of the many attributes of the flower that were *not* selected as your data!

A ***variable*** is a characteristic or attritubute, something whose value can vary. ***Data*** are the values you obtain by measurement of the variable. "Color" is a variable. "Red" is a data value.



Relationships/ Model

Analysis/ Synthesis

A ***relationship*** exists between two variables if they co-vary (eg; – the relationship between excessive sun exposure and occurrence of skin cancer)

Statistical modeling is used to discover relationships. Beginning with the data, models are fit to the data and ***not*** the other way around! It is important to appreciate that there might well be several models that are a good description of the available data.

A good model is one that (i) explains a good amount of the variability in the data (***adequacy***); and is then (ii) minimally adequate (***parsimony***), meaning: it represents your best understanding of the factors that are related to your response variable while, at the same time, being as simple as possible.

The existence of a relationship does ***not*** mean there is causality.

2. A Feel for Things

A variety of illustrations provide a feel for things.

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 1 – 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 2 – 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, that is: is the excess beyond what we might have anticipated by chance?

Example 3 – 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 (association $\neq$ causation).

Example 4 – 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
(Consider this. We all view the world through tinted lenses)

Example 5 – 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.

	Reported Enthusiasm for Shunt		
	Marked	**Moderate**	**None**
Design			
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 6 – 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 that, for these same data, we also have information about **asbestos exposure**. In particular, suppose we can "partition" the entire data into two subsets, one where exposure to asbestos has occurred and one where it has not occurred. Thus, within each subset, all persons have "similar" levels of exposure (we have "controlled" for asbestos 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%

We see that “controlling for asbestos exposure” eliminates the apparent relationship, since 22% is similar to 21% and 0.6% is similar to 0.4%). Is exposure to asbestos associated with cancer?

So perhaps asbestos exposure is the culprit, rather than proximity to electricity transmission equipment. Now let’s “partition” the available data “the other way around” namely: one subset where proximity to electricity transmission equipment is near and the other subset where proximity is not near.

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%

Now it appears that asbestos exposure is associated with cancer, regardless of location of residence, since 22% is very different from 0.6%and 21% is very different from 0.4%.

So 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.

Putting it all together

The information available 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. Again, consider this. We all have our agendas. Thus, we bring to decision-making settings our priorities of judgment. Some of these are in our awareness and, possibly, are desirable. Others are not. Examples of decision-making settings where there might exist biases that are *known and desirable* are the following:

- Judicial system (bias: "innocent until proven guilty")
- Diagnostic testing (bias: "when in doubt, continue to suspect disease")
- Type I, II error (stay tuned … we'll get to this in unit 7, hypothesis testing)

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 the values of statistics from a sample to make estimates about unknown population parameter values.

- **Inference making** – through the fitting and comparison of competing models of the data, we obtain a comparison (hypothesis test) of competing explanations (hypotheses) of the phenomena we have observed.

Example 7 -

In 1969, the average number of serious accidents per 1000 workers per year in a large factory was 10. In 2015, 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 helpful:

Signal - Treatment effect, Exposure effect, Secular trend

Noise - Natural variation, Random error

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

Description	**Inference Making**
Example: From a data set consisting of 573 cholesterol values obtained from a simple random sample of a specified population, calculate the sample mean and use this to obtain an estimate of the unknown population mean cholesterol value**.**	**Example**: Is excessive occupational exposure to video display terminals (computer monitors) during pregnancy associated with a greater likelihood of spontaneous abortion?
Solution: Confidence interval for the unknown population mean value. We will learn how to do this in Unit 6, *Estimation***.**	***Solution***: Two sample test of equality of occurrence of spontaneous abortion. We will learn how to do this in Unit 7, *Hypothesis Testing*.

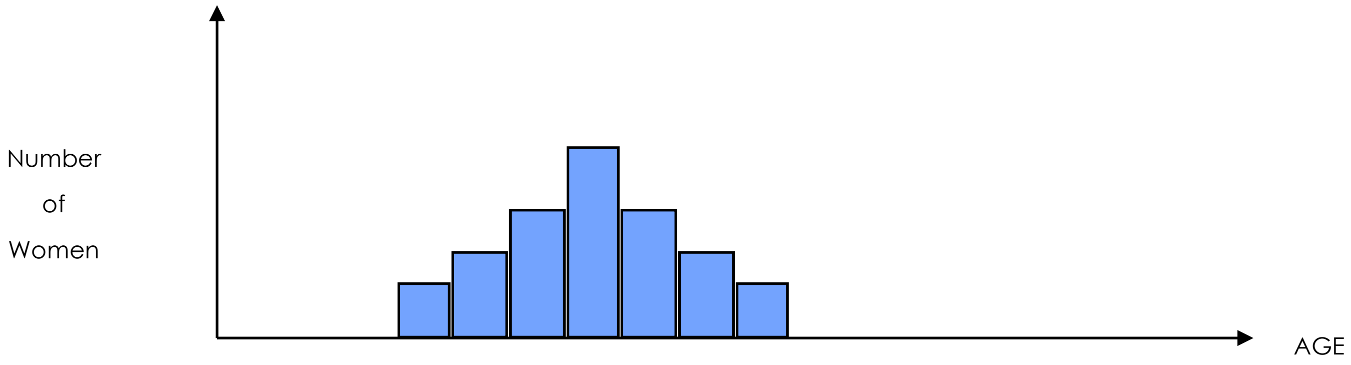
3. Overview, Unit by Unit

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. This course will emphasize the importance of looking at data.

Example -

Suppose we have the ages of 573 women visiting a prenatal care clinic. The listing of these ages, all 573 of them, is hard to appreciate. Instead, it would be nicer to have a summary that communicates something useful (e.g. – what is typical, how much do the ages vary, etc). Possible summaries include: the average age, range of ages, or a graph of some sort. The following is an example of a histogram summary.



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, exhibited a more favorable response if in fact the treatment and control therapies are equivalent? (*clinical trials*)

Example of diagnostic testing - 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 that an individual will ever develop breast cancer are 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 in Unit 2, the answer in this example is a 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.

Some Commonly Used Terms and Notation:

Population: Entire class of individuals. N = # in population (if finite)	**Sample:** A part of the population (subset). n = # in sample
Parameter: A numerical fact about the population. Parameter values are represented using Greek letters. For example, the average value of a variable, taken over all the individuals in the population is represented using the Greek letter μ **Tip** – In general, we do ***not*** get to see population parameter values.	**Statistic** A number - A numerical fact about the sample. Values of statistics are represented using Roman letters. For example, the average value of a variable X, taken over the individuals in the sample is represented using the notation $\overline{X}$. **Tip** – We do get to see values of statistics

In this unit, you will be introduced to the idea of drawing a simple random sample from a population . 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 population of N=293 at a particular school. The 20 children in the sample were each given the test. Based on a summarization of their test values, an estimate is made concerning the blood levels of all 293 children in the school.

Units 4 and 5 -
Bernoulli and Binomial Distribution
Normal Distribution

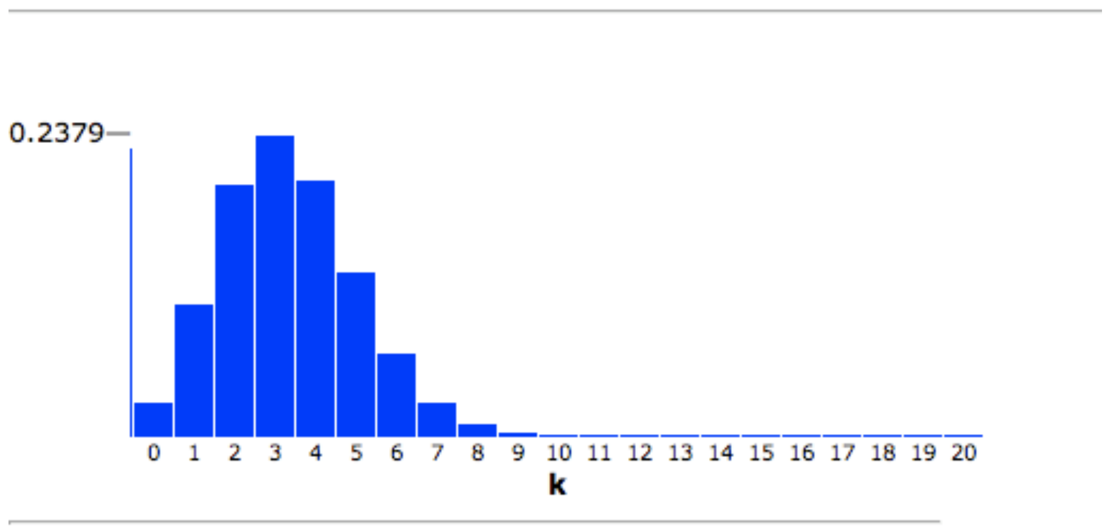
The patterns of occurrence of many phenomena can be described well using some known probability models. In units 4 and 5, you will be introduced to three probability models: Bernoulli, Binomial, and Normal.

The **Bernoulli (Bernoulli trial)** probability model is useful for modeling the pattern of discrete outcomes in one instance where there are only two possible outcomes (eg – "success" or "failure").

> **Example -** The outcome of tossing a fair coin one time is modeled using the Bernoulli probability model. It says that "heads" occurs with probability 50% and tails occurs with probability 50%

The **Binomial** probability model is useful for modeling the net result of a multiple number of Bernoulli trials. (eg – "what are the chances of 7 sixes in 20 rolls of a single die?").

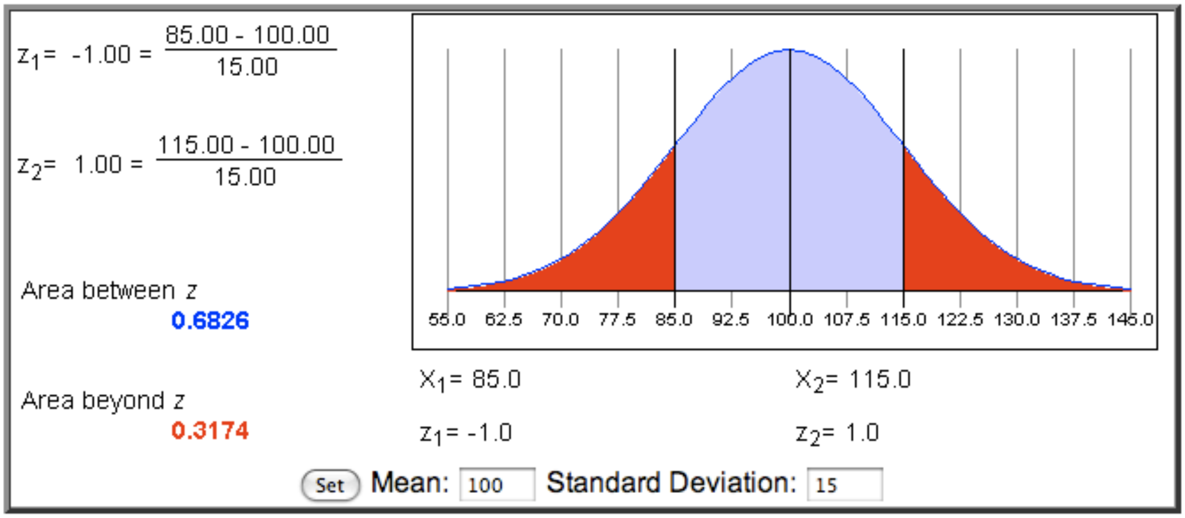
> **Example -** The probability of obtaining a six in one rolling of a single die is 16.67%. Suppose you roll the single die 20 times. The probabilities of obtaining 0 sixes, 1 six, 2 sixes, etc, is an example of the binomial probability distribution. A graph of this probability distribution is shown below. On the horizontal axis, "k" refers to the number of sixes obtained; thus, k might be 0, 1, 2, … , 20. On the vertical axis is the probability of getting that many ("k") sixes in 20 rolls. For example, you can see in this graph that the probability of getting k=3 sixes in 20 rolls is .2379 or about a 24% chance.



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

The **Normal** (also called **Gaussian**) probability model is one model that 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. A graph of this probability distribution is shown below. On the horizontal axis, "x_1 and x_2" refer to two different IQ test scores – 85 and 115, respectively. The values "z_1 and z_2" below are standardizations of "x_1 and x_2" and are called standardized z-scores. The smooth bell-shaped curve is called the probability density function. Probabilities here are calculated as areas under this curve. This graph says that the probability is .6826 (representing a 68% chance, approximately) that a randomly sampled individual has an IQ that is between 85 and 115. Much more on this in Unit 5.



$$z_1 = -1.00 = \frac{85.00 - 100.00}{15.00}$$

$$z_2 = 1.00 = \frac{115.00 - 100.00}{15.00}$$

Area between z
0.6826

Area beyond z
0.3174

X_1= 85.0 X_2= 115.0

z_1= -1.0 z_2= 1.0

Mean: 100 Standard Deviation: 15

Source: http://strader.cehd.tamu.edu/Mathematics/Statistics/NormalCurve/

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 – A particular school has N_1= 293 children. On the basis of a simple random sample of size "n_1=50" and the measurement of low density cholesterol (LDL) on each child, it is of interest to estimate the average LDL of all of the 293 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 n_2 =25 students is taken from the N_2 =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?

	Blood Pressure		
Subject	**Before**	**After**	**Difference**
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.

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 “c” represent counts. Thus, in this example of n deaths, we observe “a” deaths due to heart attack in smokers.

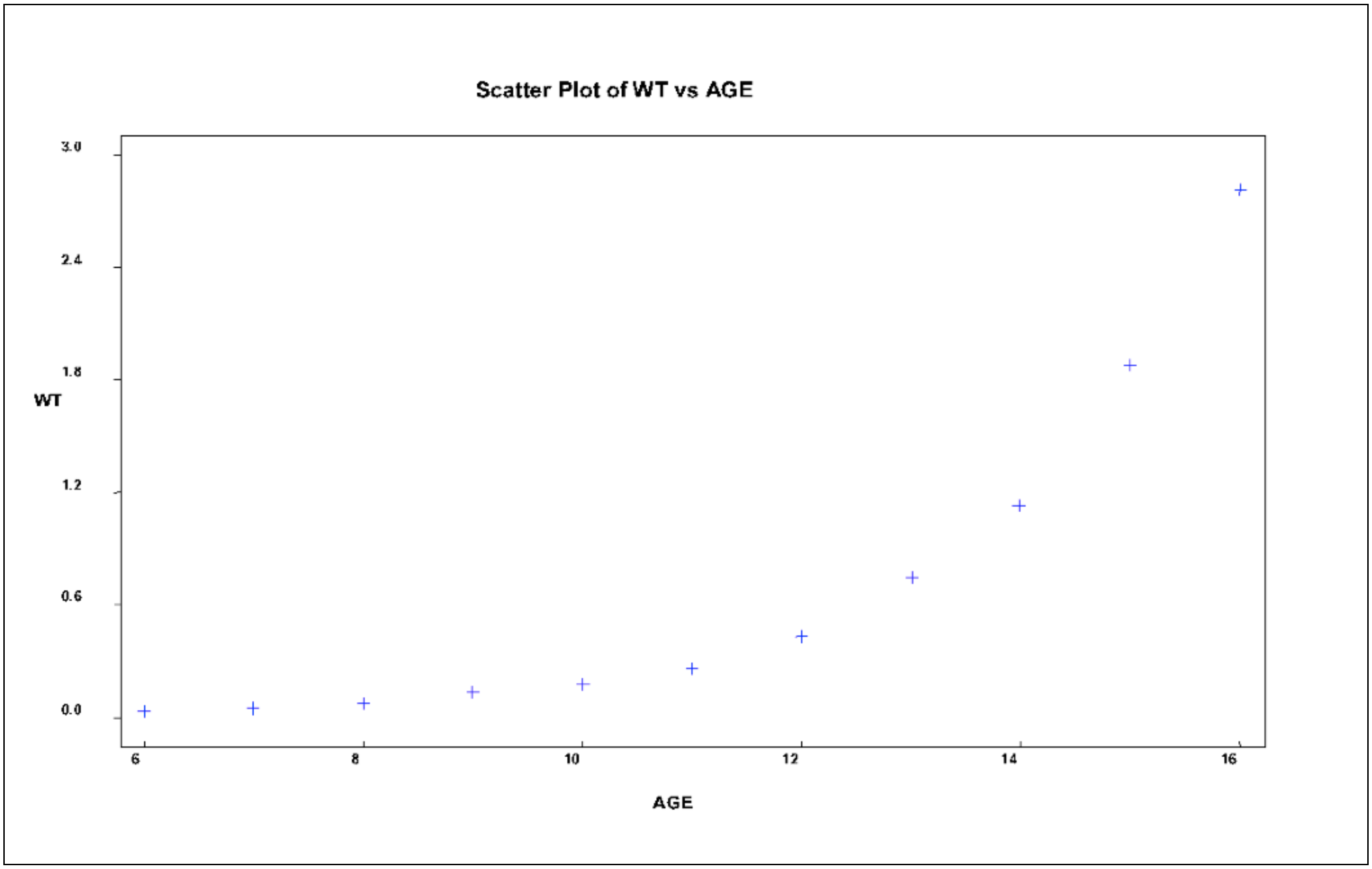
	Died of Heart Attack	**Died of Other Cause**	
Smoker	**a**	**b**	
Non-smoker	**c**	**d**	
			n

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 age?



- The plot suggests a relationship between AGE and WT
- Specifically, it suggests that older AGE is associated with higher WT
- A straight line might fit well, but another model might be better
- We have adequate ranges of values for both AGE and WT
- There are no outliers

Key Points

Biostatistics should be informed by nature.

We're not certain, nor objective, nor expert

The signal-to-noise analogy is useful.

The generic test statistic is an expression of signal/noise

Statistical inference is not the same as biological inference.

An isolated p-value is "blind" to influences of selection, mechanism

Meaningful inference requires the intertwining of design and analysis.

Appropriate conclusions take into account biological plausibility ("what makes sense") and limitations of design (eg – was sample size adequate? Was sampling representative?)