Introduction

There are three type of lies: lies, damned lies, and statistics.

-Benjamin Disraeli

- Are statistics lies? Can statistics lie?
  (MY answer is NO!?)

- Why use statistics?

  It is easy to lie with statistics,
  but it is easier to lie without them.

  -Frederick Mosteller

Statistics: 1 Numerical data assembled and classified so as to present significant information; 2 The science of compiling such data.

Literally,

"Statistic" means numerical summarization

More broadly it has come to mean the discipline of drawing conclusions (*making inferences*) from data.

Biostatistics is a growing sub-discipline of statistics that focuses on the Health sciences (medicine, biology, etc.)

- Public Health
- Environmental science
- Clinical medicine
- Health policy
- Laboratory (life sciences)
- Epidemiology
- Genetics

In general, the field of Biostatistics tends to be less theoretical than the field of Statistics, and more focused on methods for making statistical inferences from data sets.

What does drawing statistical conclusions or making statistical inferences mean?

For example: Suppose I collect a data set "**D**". Then I may ask:

1) What should I believe, now that I have observed "**D**"?

2) What should I do, now that I have observed "**D**"?

3) What do these data ("**D**") tell me about one hypothesis versus another?

(#3 restated - How should I interpret these data as statistical evidence regarding one hypothesis over another?)

All three questions are perfectly valid and important in their own right.
(We'll learn to how to answer each question.)

The answer to each question is unique, and that answer does not provide an answer to any other question.

In fact each question requires different preliminary information to formulate an answer:

1) Prior beliefs; Probability model
2) Cost and benefits of each action; Prior beliefs; Probability model
3) Probability model

In addition to 'making inferences' from data, statistical methods are useful for:

1) Estimating unknown quantities
2) Making predictions about a process
3) Quantifying the uncertainty in 1 & 2

"If there is one concept underlying the subject of statistics, it is that of variability."

-Douglas Altman

Two types of variability:

1) Patterned, ordered variability
2) Chaotic, random, unpredictable variability

Statistics (and probability) is responsible for dealing with the chaotic, random, variability.

How to:

- Describe and measure it (both mathematically and verbally)
- Control it (and recognize when you can't)
- Allow for it
- Detect and measure patterns in the presence of it.

The secret:

Underlying the chaotic, unpredictable variability of individuals yields a quite orderly, consistent, and predictable pattern of variability in large groups of individuals.

Example: Effect of AZT on maternal-infant transmission of HIV virus.

363 HIV-infected pregnant women

- 180 were given AZT treatment
- 183 were given placebo

	Baby's HIV status: Positive	Baby's HIV status: Negative	
Placebo	40	143	183
AZT	13	167	180

Does AZT treatment prevent transmission?

Does AZT treatment reduce the chance (i.e., probability) of transmission?

Do these observations prove that AZT treatment reduces the probability that the virus will be transmitted to the baby?

Do you believe that AZT treatment reduces the probability of transmission?

Should you act as if AZT treatment reduces the probability of transmission?

These observations do **not** tell us what the two transmission rates are.

They represent statistical evidence about the rates.

They enable us to estimate the rates
(e.g., by 40/183=0.22 and 13/180=0.07)

If we were to repeat this study with another group of pregnant women, we would almost certainly not come up with these exact same numbers.
(Because of the chaotic, random variability of Individuals.)

But we would probably observe a similar result.
(We will learn why.)

KEY DISTINCTION

Sample rates	vs.	Population Rates
(40/183, 13/180)	vs.	(**p**, **a**)

- Ratio of sample rates (0.22 / 0.07=3) estimates the ratio of pop. rates (**p**/**a**=?)

- We do **not** know **p** or **a**, or that the placebo rate is three times the AZT rate (**p**/**a**=3).

- We do have evidence about these quantities. We have evidence that the placebo rate is three times the AZT rate.

See Connor et.al. (NJEM 1994; 331: 1173-80).

How strong is our evidence?

How accurate are our estimates?

Statistics is supposed to provide tools for answering such questions. That is tools for:

- Measuring uncertainty
- Measuring the strength of evidence in the data
- Measuring and controlling the risk of errors (e.g., the risk of estimating **p**/**a** to be about 3 when it is really only 1).

Final thought

Three essential elements to any statistical problem/method/inference:

1. Data
2. Probability model (assumptions)
3. Context (generalizability)

Statistics do not lie; they merely reflect the three elements listed above.