# Mathematics and Biostatistics

An Introduction to Biostatistics for the Health Sciences

1st Semester 2021

Lecture 1

# Textbook

- The text for this course is:
- ***Introduction to the Practice of Statistics: 6th Edition***
- Authors: David Moore, George McCabe and Bruce Craig
- Published by W.H. Freeman and Company, New York, 2009

- ***Or any other textbooks you prefer***

# Introduction

- Definition: Statistics is the branch of mathematical science that is focused on the collection, organization, and interpretation of numerical information (Data).

- Note that within this definition the term interpretation includes making inferences from the observed data.

- In this class we will focus on the general concepts of statistical theory, methods, and operations, and data Analysis

# Introduction

- The term data is a synonym for numerical information.
- Another term that implies numerical information is measurement.
- In general, data refers to a collection of measurements that often have a common relational component (being made on the same person(s), or object(s) for some specific health reason, research project etc.).

# Statistics and Uncertainty

- Typically one has a set of observations that we believe somehow represent a larger group of observations. The goal being to infer from the observed data to the unobserved larger group.
- Since we do not view all of the possible observations there is some uncertainty in the inference.
- The uncertainty is measured using the ideas of probability and variability.

# Biostatistics

- Biostatistics is a specialized discipline of statistics that deals with statistical applications in the biological and health sciences.

- The design of health surveys, clinical trials, vital statistics, cancer survivorship studies and biological field studies are some specific biostatistical applications.

# Collection of Information

- Collection deals with exactly what you would think—how the data are obtained.

- Methods of Sampling and Research Designs are aspects of collection. These often involve randomization as in ‘randomized clinical trials’, or Random sampling (aka scientific sampling)

- Various designs for sampling and research are available.

# Organization and Presentation of Data

- Organization-This aspect may refer to how the data are stored in a database. However it mostly refers to presentation of the data, so as to efficiently and effectively bring out the information content.

- This area of statistics is often referred to as descriptive statistics. It also includes aspects of data management and coding.

# Inference from Data

- Inference is making a generalization from a few specific measurements (sample) to a larger set of measurements (population).

- It is a statement that goes beyond the given data. Hence uncertainty

EX, You observe 10 diabetic patients and infer all patients are diabetic or at least the majority of them are diabetic. This goes beyond the 10 patients of the sample to talking about all patients.

# Inference

- It is the nature of most problems that not every item or person in the world can be measured for a study.
- The solution is typically to observe a sample or subset of all possible persons or items.
- From this subset an inference will be made to the entire collection or population of persons or items.

# Inference and Probability

- In Statistics, inference takes the form of probability statements which allow a way to measure the uncertainty in our statements.

- These probability statements are usually related to hypothesis tests and confidence intervals that we will cover.

# Probability

- Probabilities are numbers in the range [0, 1] that are used to indicate the chance of some event occurring.

- The value 0 indicates no chance at all, while the value 1 indicates the event is certain to happen.

- Probabilities can never be negative or larger than 1.

# Probability

- In many applications probabilities are expressed as a percentage ranging from 0% to 100%. We will treat probability and percentage as synonyms in most cases.

- We commonly hear of such usage in terms of the chance of disease infection caused by specific type of bacteria being 40% or some other value.
- Again note 0% implies no infection of disease and 100% indicates disease is certain if not already occurring.

# Schools of Statistical Inference

1) Classical or Frequentist School (which is what we will study).
2) Bayesian School
3) Likelihood School

# Frequentist School

- Frequentist use the notion of frequency as the basis for probability. For example, tossing a coin say 1000 times and recording 551 heads.
- The frequentist would say the best estimate of the probability of heads is the number of observed heads divided by the number of tosses. Here that would be
- Prob(Heads) = 551/1000 = 0.551, or 55.1%

# Other Schools

- Though the freqeuntist school has some intuitive appeal and has been the dominant school in science at this time, both the likelihood and Bayesian approaches have strengths that the frequentist aprroach does not have.
- However, both the likelihood and Bayesian approaches require a much greater knowledge of mathematical statistics. An in depth discussion is beyond this course.

# End of Lecture 1