In complex dynamic systems such as biological organisms, how is it possible to distinguish genuine – or “statistically significant” – sources of variation, from purely “random chance” effects? Why is it important to do so?

Consider the following three experimental scenarios…

- In a clinical trial designed to test the efficacy of a new drug, participants are randomized to either a control arm (e.g., a standard drug or placebo) or a treatment arm, and carefully monitored over time. After the study ends, the two groups are then compared to determine if the differences between them are “statistically significant” or not.

- In a longitudinal study of a cohort of individuals, the strength of association between a disease such as COPD (Chronic Obstructive Pulmonary Disease) or lung cancer, and exposure to a potential risk factor such as smoking, is estimated and determined to be “statistically significant.”

- By formulating an explicit mathematical model, an investigator wishes to describe how much variation in a response variable, such as mean survival time after disease diagnosis in a group of individuals, can be deterministically explained in terms of one or more “statistically significant” predictor variables with which it is correlated.

This first course is an introduction to the basic but powerful techniques of statistical analysis – techniques which formally implement the fundamental principles of the classical scientific method – in the general context of biomedical applications. How to:

1. formulate a hypothesis about some characteristic of a variable quantity measured on a population (e.g., mean cholesterol level, proportion of treated patients who improve),

2. classify different designs of experiment that generate appropriate sample data (e.g., randomized clinical trials, cohort studies, case-control studies),

3. investigate ways to explore, describe and summarize the resulting empirical observations (e.g., visual displays, numerical statistics),

4. conduct a rigorous statistical analysis (e.g., by comparing the empirical results with a known reference obtained from Probability Theory), and finally,

5. infer a conclusion (i.e., whether or not the original hypothesis is rejected) and corresponding interpretation (e.g., whether or not there exists a genuine “treatment effect”).

These important biostatistical techniques form a major component in much of the currently active research that is conducted in the health sciences, such as the design of safe and effective pharmaceuticals and medical devices, epidemiological studies, patient surveys, and many other applications. Lecture topics and exams will include material on:

- Exploratory Data Analysis of Random Samples
- Probability Theory and Classical Population Distributions
- Statistical Inference and Hypothesis Testing
- Regression Models
- Survival Analysis
A Brief Overview of Statistics

Statistics is a quantitative discipline that allows objective general statements to be made about a population of units (e.g., people from Wisconsin), from specific data, either numerical (e.g., weight in pounds) or categorical (e.g., overweight / normal weight / underweight), taken from a random sample. It parallels and implements the fundamental steps of the classical scientific method: (1) the formulation of a testable null hypothesis for the population, (2) the design of an experiment specifically designed to test this hypothesis, (3) the performance of which results in empirical observations, (4) subsequent analysis and interpretation of the generated data set, and finally, (5) conclusion about the hypothesis.

Specifically, a reproducible scientific study requires an explicit measurable quantity, known as a random variable (e.g., IQ, annual income, cholesterol level, etc.), for the population. This variable has some ideal probability distribution of values in the population, for example, a bell curve (see figure), which in turn has certain population characteristics, a.k.a. parameters, such as a numerical “center” and “spread.” A null hypothesis typically conjectures a fixed numerical value (or sometimes, just a largest or smallest numerical bound) for a specific parameter of that distribution. (In this example, its “center” – as measured by the population mean IQ – is hypothesized to be 100.) After being visually displayed by any of several methods (e.g., a histogram; see figure), empirical data can then be numerically “summarized” via sample characteristics, a.k.a. statistics, that estimate these parameters without bias. (Here, the sample mean IQ is calculated to be 117.) Finally, in a process known as statistical inference, the original null hypothesis is either rejected or retained, based on whether or not the difference between these two values (117 – 100 = 17) is statistically significant at some pre-specified significance level (say, a 5% Type I error rate). If this difference is “not significant” – i.e., is due to random chance variation alone – then the data tend to support the null hypothesis. However, if the difference is “significant” – i.e., genuine, not due to random chance variation alone – then the data tend to refute the null hypothesis, and it is rejected in favor of a complementary alternative hypothesis.

Formally, this decision is reached via the computation of any or all of three closely related quantities:

1) **Confidence Interval** = the observed sample statistic (117), plus or minus a margin of error. This interval is so constructed as to contain the hypothesized parameter value (100) with a pre-specified high probability (say, 95%), the confidence level. If it does not, then the null is rejected.

2) **Acceptance Region** = the hypothesized parameter value (100), plus or minus a margin of error. This is constructed to contain the sample statistic (117), again at a pre-specified confidence level (say, 95%). If it does not, then the null hypothesis is rejected.

3) **p-value** = a measure of how probable it is to obtain the observed sample statistic (117) or worse, assuming that the null hypothesis is true, i.e., that the conjectured value (100) is really the true value of the parameter. (Thus, the smaller the p-value, the less probable that the sample data support the null hypothesis.) This “tail probability” (0%-100%) is formally calculated using a test statistic, and compared with the significance level (see above) to arrive at a decision about the null hypothesis.

Moreover, an attempt is sometimes made to formulate a mathematical model of a desired population response variable (e.g., lung cancer) in terms of one or more predictor (or explanatory) variables (e.g., smoking) with which it has some nonzero correlation, using sample data. Regression techniques can be used to calculate such a model, as well as to test its validity.

This course will introduce the fundamental statistical methods that are used in all quantitative fields. Material will include the different types of variable data and their descriptions, working the appropriate statistical tests for a given hypothesis, and how to interpret the results accordingly in order to formulate a valid conclusion for the population of interest. This will provide sufficient background to conduct basic statistical analyses, understand the basic statistical content of published journal articles and other scientific literature, and investigate more specialized statistical techniques if necessary.
**POPULATION**

Random Variable: $X$ = IQ score, having an ideal distribution of values

Null Hypothesis: Mean $\mu = 100$

(about a parameter)

**Experiment** to test hypothesis

Statistical Inference

Conclusion: Does the experimental evidence tend to support or refute the null hypothesis?

**RANDOM SAMPLE**

Observations

Analysis of empirically-generated data (e.g., via a histogram):

Statistic: Mean $\bar{x} = 117$

(estimate of parameter)