# CHAPTER ONE

# BASIC STATISTICAL THEORY

> "Statistical methods are objective methods by which group trends are abstracted from observations on many separate individuals."[1]

Medicine is founded on the principles of statistical theory. As physicians, we employ **scientific reasoning** each day to diagnose the conditions affecting our patients and treat them appropriately. We apply our knowledge of and previous experience with various disease processes to create a differential diagnosis which we then use to determine which disease our patient is most likely to have. In essence, we propose a **hypothesis** (our diagnosis) and attempt to prove or disprove it through laboratory tests and diagnostic procedures. If the tests disprove our hypothesis, we must abandon our proposed diagnosis and pursue another.

## DEDUCTIVE AND INDUCTIVE REASONING

In proposing a diagnosis, we can either propose a general theory (such as "all patients with chest pain are having an acute myocardial infarction") and make a specific diagnosis based upon our theory, or we can make specific observations (such as a CK-MB band of 10%, S-T segment changes on EKG, tachycardia, etc.) and arrive at a general diagnosis.

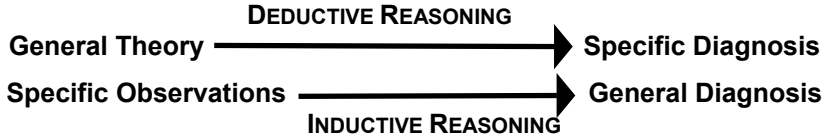


**Figure 1-1: Deductive and Inductive Reasoning**

There are thus two forms of scientific reasoning. **Deductive reasoning** involves proposing a general theory and using it to predict a specific conclusion. **Inductive reasoning** involves identifying specific observations which are used to propose a general theory. Physicians use **deductive reasoning** every day. If we are presented with a patient who has right lower quadrant abdominal pain and the last four patients we have seen with this finding have had acute appendicitis, we will likely use deductive reasoning to predict that this patient will also have acute appendicitis. We use a general theory (i.e., "all patients with right lower quadrant abdominal pain have acute appendicitis") to arrive at a specific conclusion or diagnosis. We could just as easily have used **inductive reasoning** to arrive at the same diagnosis. Upon completing our history and physical examination, we might have referred to the medical literature on acute appendicitis and found that 55% of patients have the classic right lower quadrant abdominal pain, 90% have nausea, and up to 33% may have a normal white blood cell count. Based upon our history, physical examination, laboratory findings, and this data from the medical literature (all specific observations), we could then have used inductive reasoning to make a general diagnosis of acute appendicitis.

Deductive and inductive reasoning are used throughout medical research to answer questions about specific groups of patients. A hypothesis or theory regarding the question of interest is proposed and and a study is performed in which attempts are made to disprove the study hypothesis using various tests and diagnostic procedures. From the observations collected on these study patients, the original hypothesis is either accepted or modified and the study conclusions are applied to the treatment of patients in the general population.

## HYPOTHESIS TESTING

The **hypothesis** is the foundation of any research study or clinical trial. To answer any question, we must first propose two hypotheses. The first, the **null hypothesis**, is the cornerstone of statistical inference. The null hypothesis states that "there is no difference between the groups or treatments being evaluated". The null hypothesis is tested against the second of our two hypotheses, the **alternate hypothesis**, which states that "there is a difference".

**Null Hypothesis - there is no difference**
**Alternate Hypothesis - there is a difference**

If we reject the null hypothesis of no difference, we must accept the alternate hypothesis and conclude that there is a difference. It is important to remember that we can never prove that the null hypothesis is true; there is always a possibility that a difference exists, but that we have not found it yet. We have only but to show that a difference exists, however, to prove the null hypothesis is false. Therefore, we propose a null hypothesis and attempt to disprove it using statistical analysis. We use statistics because we can never be absolutely certain that a difference does or does not exist. We can therefore never be absolutely certain that we are correct in either accepting or rejecting any hypothesis. Statistical tests are used to either reject the null hypothesis or fail to reject it based on the statistical probabilities associated with the outcome of interest.

## TYPE I ERRORS, TYPE II ERRORS, AND SIGNIFICANCE LEVELS

There are two types of errors associated with using the null hypothesis. A **Type I error** occurs if the null hypothesis is rejected incorrectly (i.e., we conclude there is a difference when in fact there is not). The probability of making a Type I error is known as the **significance level** of the test. It is also commonly referred to as "**alpha**." The significance level is a measure of how willing we are to make a Type I error. More formally stated, it is the probability of obtaining a result at least as unlikely as that which we have observed if the null hypothesis is really true. The familiar "**p-value**," upon which such emphasis is placed in the medical literature, originates from the significance level. A p-value of 0.02 means that there is a 2% chance that we are wrong when we conclude that a difference exists. Traditionally, most researchers use a significance level of 0.05 to define statistical significance; that is, they are willing to accept a 5% chance of falsely concluding that a difference exists when it does not (a Type I error). Another way to look at this is that there is a 95% probability of correctly accepting (or rejecting) the null hypothesis. If we are very concerned about committing a Type I error, we might decrease our chosen significance level to 0.01 such that we are now willing to accept only a 1% chance of incorrectly rejecting the null hypothesis. The second potential error in using the null hypothesis is a **Type II error,** the probability of which is known as "**beta**." A Type II error occurs when we fail to reject the null hypothesis when a difference does exist. We thus miss a potentially important effect by falsely concluding that there is no difference.

The impact of Type I and Type II errors is dependent upon the nature of the hypothesis being tested. If we are evaluating a drug to prevent sepsis and commit a Type I error (i.e., we conclude that the drug is better than placebo when it is not), we may promote a drug that in reality has no effect. Conversely, if we are testing a vaccine to prevent AIDS and commit a Type II error (i.e., we falsely conclude that the vaccine does not improve survival), a lifesaving vaccine may be ignored. As we have seen, one way to decrease the probability of a Type I error is to lower the significance level or alpha. By doing so, however, we make it harder to accept the alternate hypothesis and more likely that we will commit a Type II error and miss a significant difference. Thus, as the risk of committing a Type I error is decreased (such as by lowering the significance level of the test to 0.01), the risk of making a Type II error must necessarily increase. In order to lower the probability of making either a Type I or Type II error, we must increase the number of observations. As the number of observations increases, we will be more likely to detect a difference if one exists thus decreasing the chance of making an error.

## POWER

The ability to detect a difference is known as the **power** of a test and is defined as 1- beta (where "beta" represents the probability of making a Type II error). The higher the power of a test, the more likely a difference will be detected (if one exists). If finding a large difference between two prospective treatments (known as the **effect size**) is likely, only a small number of observations will be required to prove that the difference is real. As the effect size one wishes to detect becomes smaller, however, the number of observations necessary to detect a significant difference becomes larger. A clinical trial can therefore prove that no difference exists only if it has sufficient power to detect a significant difference to begin with. A study may likewise fail to detect a clinically significant difference due to a lack of sufficient power. A "non-significant" result only means that the existing data was not strong enough to reject the null hypothesis of no difference; a larger **sample size** might have provided the power necessary to reject the null hypothesis. Power is thus proportional to the number of observations. The sample size necessary to either accept or reject the null hypothesis is calculated from the significance level (the probability of making a Type I error), the power (the probability of finding a difference when there really is one), and the effect size (the clinically significant

difference we wish to detect) and should be calculated before a clinical trial is begun to ensure that the study has the ability to detect any differences which may be present. Calculation of the sample size necessary for a clinical study is known as "**power analysis**" and will be discussed in Chapter Nine.

## ONE-TAILED VERSUS TWO-TAILED TESTS

Careful definition of study hypotheses is the foundation of any research study. If the study hypotheses are not clearly stated at the beginning, conclusions based on the study observations will be difficult to make. In any study comparing two patient groups, tests, or outcomes (designated here as A and B), there are only three possible hypotheses:

**1) There is no difference between A and B.**

**2) There is a difference between A and B.**

**3) A is superior to B (or vice versa).**

The first hypothesis is the null hypothesis of "no difference" which we attempt to disprove in any study. The second and third hypotheses are both forms of the alternate hypothesis. The second hypothesis states that there is a difference between A and B, but does not indicate the direction of the difference. This is the alternate hypothesis which we use if we have no *a priori* impression as to whether A or B is superior to the other. This is known as a "**two-tailed**" alternate hypothesis as we must test to see whether A > B and B > A. The third hypothesis clearly states A is superior to B. In order to prove that this is true we need only look in one direction (A > B). This is an example of a "**one-tailed**" alternate hypothesis and is the one we would choose if we are only interested in proving that A is superior to B. Because of this directionality, it is easier to prove a one-tailed hypothesis than a two-tailed hypothesis (which must look in both directions) and the statistical method which we use to test our hypothesis must take this into account.

## VARIABILITY: THE REASON FOR STATISTICAL ANALYSIS

We can never be absolutely certain that we are 100 percent correct in either accepting or rejecting any hypothesis. Such is the case with our clinical diagnoses as well. Not all patients with acute appendicitis present with the same signs and symptoms. There is always a certain **variability** present which imparts a degree of uncertainty to each of our diagnoses. The best that we can do is to make the most likely diagnosis given the data and the potential for error present. Variation can be inherent (due to normal biologic differences between patients) or introduced (such as an erroneous lab value, misinterpreted test, or missed physical exam finding). It is the presence of this inherent uncertainty that requires us to use statistical analysis throughout the practice of medicine. Because of variability, it is impossible to know the absolute chance of an event occurring or diagnosis being correct (such as acute appendicitis). There is always a small possibility that our diagnosis may be wrong. The concept of variability is central to basic statistical theory and is used in the calculation of virtually all statistical tests. It is not just the presence of variability that is important, but the amount. If we could measure the absolute variability present in our clinical observations, our diagnoses could be fairly accurate because we would always know how likely it was that we were wrong. It is impossible, however, to measure every source of variability and we use statistics to predict how much variability is likely to be present and how likely our diagnosis is wrong due to variability.

## PROBABILITY

We are therefore interested in the **probability** that our diagnosis is correct. Probabilities can range from zero (our diagnosis will never be correct) to one (our diagnosis will always be correct) and are calculated by dividing the number of times that an observation occurs by the number of potential observations. Suppose, for example, that upon exploratory laparotomy our patient is found to not have acute appendicitis. Our probability of accurately diagnosing acute appendicitis is now 4 cases of appendicitis out of 5 patients with abdominal pain or 0.80. We are thus 80% accurate in diagnosing acute appendicitis. Correspondingly, we have a probability of misdiagnosing acute appendicitis of 0.20 (1 missed diagnosis out of 5 patients with abdominal pain). Knowing this, we can calculate our **expected frequency** for errors in diagnosing acute appendicitis. If we see 100 patients each year with presumed acute appendicitis, we will likely misdiagnose 100 x 0.20 or 20 patients. The **odds** of an event are given by the probability that the event occurs divided by the probability that it does not occur. The odds of our correctly diagnosing acute appendicitis is therefore 0.80/0.20 or 4. We are 4 times more likely to make the correct diagnosis of acute appendicitis than the incorrect diagnosis. The **likelihood** of an event is the probability of the event occurring in different situations.

Thus, the likelihood of acute appendicitis might be different for men versus women or young people versus old.

The impact of one probability on another depends upon how they are related. If two events are mutually exclusive (i.e., the occurrence of one event excludes the occurrence of the other), the probability of either one event or the other occurring is given by the sum of their respective probabilities. If two events are independent (i.e., one event has no impact on the occurrence of the other event), the probability of both events occurring together is given by the product of their respective probabilities. For example, suppose 100 patients (60% male, 40% female) are extubated from mechanical ventilation and the incidence of reintubation is 20% (80% are successfully extubated). The probability of either remaining extubated or requiring reintubation (two mutually exclusive events) is 0.80 + 0.20 or 1.0. Similarly, the probability of being male or female (also two mutually exclusive events) is 0.60 + 0.40 or 1.0. The likelihood of being male and requiring reintubation (two independent events) is 0.60 x 0.20 or 0.12 while the likelihood of being female and requiring reintubation is 0.40 x 0.20 or 0.08. The odds of requiring reintubation and being male versus female is thus 0.12/0.08 or 1.5:1.

## SAMPLES AND POPULATIONS

The degree of variability present frequently depends on the **population** of interest. For example, an elevated white blood cell count in the presence of acute appendicitis is much more likely in a younger patient than in an elderly one. The probability of making a correct diagnosis of acute appendicitis based on an elevated white blood cell count is therefore higher in the younger patient as there is less variability in the presence of this sign. Thus, the population of interest can determine the amount of variability present.

A population is defined as all patients who have a particular characteristic of interest (i.e., all patients with acute appendicitis, all trauma patients, all Americans with high blood pressure). We can never evaluate the signs and symptoms of every patient in the world who develops acute appendicitis (or every trauma patient, every American with hypertension, etc.). We can, however, study a **sample** of the larger population of interest and make conclusions from data collected on the sample population which we then infer on the larger population (i.e., inductive reasoning). As always, the presence of variability in our sample introduces the possibility that our conclusions will be incorrect. If our inferences are to be accurate, therefore, the sample must be very similar to the larger population with regard to the characteristic of interest. The more similar the sample and the population, the more likely the conclusions we make from the sample will be accurate. One way in which to ensure that the sample is representative is to randomly choose sample patients from the population of interest. Theoretically, randomly chosen patients will have an equal probability of truly being representative of the population-at-large and will be more likely to provide accurate conclusions. Another method to obtain a representative sample is to increase the sample size. As the sample size approaches that of the population, the probability of error due to variability between the sample and the population-at-large will decrease.

It is usually impossible to test an entire population. This is another reason why we study a sample of the population of interest and make conclusions about the population based on the sample findings. There are significant advantages to studying samples. Samples can be studied more quickly, more accurately, and less expensively than entire populations. Statistical methods can then be used to estimate the probability that the sample findings truly represent those of the general patient population.

## SUGGESTED READING

1. Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals. New York: Springer-Verlag, 1990:1-7.
2. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics (2nd Ed). Norwalk, CT: Appleton and Lange, 1994:64-98.
3. Altman DG. Statistics and ethics in medical research: VII - interpreting results. Brit Med J 1980;281:1612-1614.
4. Moses LE. Statistical concepts fundamental to investigations. In: Bailar JC, Mosteller F (Eds). Medical uses of statistics (2nd Ed). Boston: NEJM Books, 1992:5-26.
5. Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: survey of 71 “negative trials”. NEJM 1978; 299:690-694.