Probability Review

Obviously, we cannot review all of an introductory probability course in this short time period. Instead, we will focus on the concepts with which students most often struggle and which are frequently needed in a biostatistics class. These include conditional probability and the basics of hypothesis testing.

Recall, that for any two events $A$ and $B$ with $P(B) > 0$, the conditional probability of $A$ given $B$ is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

This definition leads directly to the Multiplication Rule:

$$P(A \cap B) = P(A|B)P(B)$$

Similarly, for three events:

$$P(A \cap B \cap C) = P(A|B \cap C)P(B|C)P(C)$$

An extremely useful tool which you will encounter over and over again in this course is the Law of Total Probability.

**Theorem:** Law of Total Probability

Let $B_1, \ldots, B_k$ be a partition of the sample space. Then for any event $A$

$$P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$$

Another useful theorem is Bayes’ Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**Example:** According to the Arizona Chapter of the American Lung Association, 7% of the population has lung disease. Of those people having lung disease, 90% are smokers; and of those not having lung disease, 74.7% are non-smokers. What are the chances that a smoker has lung disease?
Another important concept from probability theory is that of the independence of events. This means that if we learn whether or not one event has actually happened, we gain no information about whether the other event will or will not happen.

**Definition:** Two events $A$ and $B$ with $P(A) > 0$ and $P(B) > 0$ are independent if and only if

$$P(A \cap B) = P(A)P(B)$$

Independence should not be confused with mutually exclusive.

**Example:** If two events are mutually exclusive, can they be independent?

**Definition:** Two events $A$ and $B$ are called conditionally independent given a third event $C$ (with $P(C) > 0$) if

$$P(A \cap B|C) = P(A|C)P(B|C)$$

**Example:** Consider the events $A, B, C$ shown in the Venn diagram below.

(a) Find $P(A|B)$.

(b) Are $A$ and $B$ independent?

(c) Are $A$ and $B$ conditionally independent given $C$?
The DNA Molecule

The deoxyribonucleic acid molecule stores genetic instructions for all living organisms. It is arranged in the famous double-helix structure (see Origami). It consists of two complementary strands of molecules. The strands consist of alternating phosphate groups and deoxyribose sugars, which in turn are linked to one of four nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The two strands are connected in the sense that an adenine molecule on strand one is always accompanied by a thymine molecule in the corresponding position on strand two. In the same sense, guanine is complementary to cytosine. Think of DNA as a twisted ladder where the phosphate sugar backbone is the rails and the base pairs are the rungs.

The information in the DNA strands is encoded in the order of base pairs. Combinations of three base pairs called a codon (such as ATA, GGT etc.) encode for an amino acid. Proteins are strands of amino acids and hence a string of codons can be used as “assembly instructions” for a protein. There are twenty amino acids encoded by this genetic code.

**Question:** How many different codons exist?

Not all bases in the DNA molecule are protein coding. There are long stretches of “junk”-DNA (whose function we do not know, yet) interspersed with protein coding regions. The protein coding regions together with certain regulatory sequences are referred to as genes.

How is the information translated from the DNA molecule to form a protein? The chromosomes in the cell’s nucleus contain the wound up DNA molecule. The proteins are mostly found in a cell’s cytoplasm. How do they get there? The DNA molecule in the nucleus has the ability to partially unwind and separate its two strands temporarily. Then another molecule is formed, complementary to a strand of DNA. This temporary molecule, called messenger RNA (mRNA) is single stranded. It has the ability to leave the nucleus and travel to the areas of the cell where proteins are constructed. Ribosomes “read” the information in the mRNA molecules and assemble the appropriate chain of amino acids. Certain start and stop sequences tell the ribosomes when to begin and end this assembly process.

The Central Dogma of Molecular Biology states that DNA is transcribed into RNA which is then translated into protein.
DNA Replication

The DNA is stored wound up tightly in chromosomes. In a diploid organism (such as humans) every cell contains two copies of every chromosome with exception of the sex chromosomes. Males have a copy of the \(x\)-chromosome and a copy of the \(y\)-chromosome, while females have two copies of the \(x\)-chromosome.

When a cell divides into two (mitosis), the information contained in the DNA needs to be duplicated exactly. In this process, called replication, the two parent strands of DNA split and each individual strand serves as a template for the synthesis of a new daughter strand. DNA strands have a direction and the new nucleotides will be attached by an enzyme called DNA polymerase in sequential order. The ends of a DNA strand are labeled as the 3’ and the 5’ end. Replication always proceeds in the direction of 5’ to 3’. The separation of parent strands and the synthesis of the daughter strands takes place simultaneously in different parts of the DNA molecule. Once replication is complete, the chromosome pairs chromatids line up and are pulled apart in the dividing cell. This way, each new cell ends up with the same set of chromosomes.

**Question:** What is the difference between DNA transcription and DNA replication?
Reproduction

Meiosis is the process by which one diploid cell divides twice to create four haploid cells (containing only one version of each gene). The resulting gametes (egg and sperm in humans) combine during fertilization to create a zygote cell which is diploid with genetic material from both parents. The zygote cell divides and grows into a new organism (offspring). During the meiosis process the chromosomes of each parent undergo recombination - shuffling of the genes on the chromosomes. Thus each zygote has a unique genetic DNA content.

Example: If the genes for two different traits are located on different chromosomes, then what is the probability that the offspring inherits both genes from mom (dad)?

The closer two genes are to each other on the same chromosome, the more likely it is for them to be inherited together (without a cross-over event between them). Knowing the probabilities of cross-over events allows us to build a genetic map, a diagram of each chromosome showing the relative position of each gene.

Example: QTL analysis is the study of quantitative trait loci. The goal is to find genes that are linked to a quantitative trait (such as yield in maize, or obesity in humans). This is done by comparing the target trait with a number of genetic markers. These are traits that can be easily observed for whom the position of the gene that controls them is known. Using linkage probabilities it is then possible to estimate the position of the gene(s) that control the target trait.
DNA Sequencing

The goal of DNA sequencing is to discover the exact order of nucleotides in a biological DNA sample. Since the discovery of the double helix structure of DNA in 1953 by Watson and Crick (and Franklin) the first systematic attempts at sequencing were made in the 1970s. At first, these methods were very tedious and labor intensive (Sanger sequencing). Since then, many advances in sequencing have been made: pyrosequencing in 1996 and parallelizing the process in the early 2000s yielding next generation sequencing.

In the early days, for instance in the original human genome project (1990 - 2003), the genome to be sequenced was replicated several times and then cut into many small random pieces (50 to a couple hundred nucleotides long) which were sequenced individually and then puzzled back together to yield the full sequence. This approach is called shotgun sequencing. These days, technology is able to read sequences that are several thousand nucleotides long, however, the error rate in those reads is comparatively high.

Sequencing costs have also decreased dramatically over time. You can have your whole genome sequenced for about $1,000 in just a few days now, whereas the same task for the human genome project took 13 years and cost $2.7 billion dollars.

**SANGER SEQUENCING:** The predominant method of sequencing for about 25 years in the late 1990s relied on replicating the (short) strand of DNA to be sequenced many times and dividing the many identical copies of the single strand into four test tubes. Chemical reactions lead to many partial copies of the DNA strand that were of different length, all started in the same place and all ended with the same letter in each test tube. That is, in one tube all strands ended with “A”, in another all strands ended with “G” etc. Then, the different strands from all four tubes were sorted by size on a gel. From this gel one could read off the original sequence.

**Example:** In the gel on the right, the heaviest DNA fragments are the ones on the top. Read off the base pair sequence of the DNA fragment analyzed in this chain termination experiment:
Pyrosequencing: When a complementary nucleotide attaches itself to a single template strand of DNA a small molecule called a phyrophosphate is released. This phyrophosphate can be converted into another molecule and the conversion will emit a small amount of light. The single strand that is to be sequenced is now exposed to solutions containing just one of the four possible nucleotides. When light is detected, a nucleotide is incorporated. If there are several identical letters in the sequence that are all incorporated at the same time, the amount of light emitted increases proportionally.

Dye-Terminator Sequencing uses terminator dideoxynucleotides that are labeled with four different fluorescent dyes. This method does not require the separation of DNA fragments into four different portions and is thus more expedient and much quicker. The DNA template strands are again mixed with normal base pairs and all four terminating color coded termination bases in an “alphabet soup”. DNA polymerase randomly picks normal or terminating base pairs for incorporation into the new strands. The next phase sorts new DNA fragments by size and uses lights of different wavelengths to detect the “ending” base for each strand. The results are read by a computer (no longer a human as in the Sanger method).

Next-generation Sequencing or high-throughput sequencing methods parallelize the sequencing process. Analyzing many thousands of sequences at once makes the process both much faster and cheaper. The trick lies in making the fluorescent termination bases removable. Many short one-stranded DNA sequences are attached to primers on a glass slide and amplified to make many identical copies of each sequence in a spot. One type of terminating nucleotide is added at a time. The terminating nucleotides are indicated by four different fluorescent dyes. A picture is taken, the termination end and the dye of the nucleotide are chemically removed and the process is repeated.

In any sequencing method, the short sequenced pieces of DNA need to be assembled into much longer chromosomes. This is done based on the overlap of the randomly generated sequences, sometimes with the help of a reference genome.
Hypothesis Testing

Hypothesis testing is the process of using sample data to decide which of two contradicting statements about a population parameter is true. Since we do not have observations on every individual in the population, we do not see the "whole picture" but we have some insight through observing the data from the sample. This limited amount of information can still be used to make an informed choice.

**Definition:** A statistical hypothesis is a claim about a population parameter (such as the population mean $\mu$ or a population proportion $p$). The **null hypothesis**, denoted by $H_0$ is the claim that is initially assumed to be true. The **alternative hypothesis**, denoted by $H_a$ is an assertion that is contradictory to $H_0$.

If the observed data is plausible (has high probability) under the null hypothesis assumption, we will accept this claim as true. If the observed data has very low probability under the null hypothesis, we will reject the null hypothesis in favor of the alternative.

**Testing procedure:** A statistical hypothesis test consists of several components.

1. The null hypothesis statement and the alternative statement. The null hypothesis is usually phrased as an equality (e.g., $H_0 : \mu = 0$ or $H_0 : p = 0.5$). The alternative can be phrased as an equality (e.g., $H_a : \mu = 3$) or an inequality (e.g., $H_a : \mu \neq 0$ or $H_a : p > 0.5$).

2. A **test statistic**. This is a function whose value can be computed from the sample data. We have to know the (theoretical) distribution of the test statistic function if the null hypothesis $H_0$ is true. The decision whether to accept or reject $H_0$ is based on the value of the test statistic computed from the data.

3. A **rejection region** - the set of all test statistic values for which the null hypothesis $H_0$ will be rejected.

**Errors:** There are two possible errors that can be made in hypothesis testing:

- Rejecting the null hypothesis $H_0$ when it is true (type I).
- Accepting the null hypothesis $H_0$ when it is false (type II).

Ideally, one would want to keep the probabilities of both these errors as small as possible. However, the error probabilities are related and if one error probability is made smaller the other one usually will increase. The choice of rejection region determines the probabilities of both a type I and type II error.

**Definition:** The probability of a type I error $\alpha$ is called the (significance) level of the test. The probability of a type II error is usually denoted by $\beta$. The quantity $1 - \beta$ represents the test’s ability to correctly reject a false null hypothesis and is called the **power** of the test.
Example: Two bags contain 2 white and 2 black (bag 1) and 1 white and 3 black marbles (bag 2), respectively. A bag is chosen at random and from that bag two marbles are selected at random.

(a) Let $p$ denote the proportion of white marbles in the selected bag. Formulate a null hypothesis and alternative hypothesis for this example.

(b) Based on the colors of the two marbles drawn, formulate a test statistic function. What are the possible values of this test statistic function?

(c) What are the values of the test statistic function that make it unlikely that the selected bag is bag 1? This is the rejection region for this hypothesis test.

(d) Compute the probability that the null hypothesis is rejected when it is in fact true ($\alpha = P(\text{type I error})$).

(e) Compute the probability that the null hypothesis is not rejected when it is in fact false ($\beta = P(\text{type II error})$).
P-Values

One way to report the results of a hypothesis test is to say whether or not the test statistic value fell into the rejection region and subsequently, whether or not the null hypothesis was rejected at a specified level of significance \( \alpha \). This yes/no decision does not convey any information about how soundly the null hypothesis was rejected. Where in the rejection region did the observed test statistic value fall?

**Definition:** Suppose the null hypothesis \( H_0 \) is, in fact true. The \( p \)-value is the probability to observe a test statistic value at least as contradictory to \( H_0 \) as the computed value by random chance due to the selection of the sample.

**Example:** (cont.)

Suppose the two selected marbles in the previous example are black and white. Compute the \( p \)-value for the null hypothesis that the proportion of white marbles in the bag is \( p = 0.5 \).

**Definition:** If a \( p \)-value is smaller than the significance level \( \alpha \), then the corresponding value of the test statistic falls into the rejection region and the null hypothesis will be rejected.

\[
p \leq \alpha \quad \Rightarrow \quad \text{reject } H_0
\]

If the \( p \)-value is large, then it is quite likely to see data such as the observed by random chance if the null hypothesis were true and \( H_0 \) will be accepted.

\[
p > \alpha \quad \Rightarrow \quad \text{accept } H_0
\]

**Classical Hypothesis Testing**

Two very common classical testing procedures are the \( t \)-test and the \( z \)-test. The \( t \)-test is used for hypotheses involving population means and the \( z \)-test is used for null hypotheses involving population proportions.

**T-test:** Suppose a sample of size \( n \) is taken at random from a population with a Normal distribution. Our goal is to decide a statement about the population mean \( \mu \). The population standard deviation \( \sigma \) is unknown. Then the test statistic

\[
t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \sim t(df = n - 1)
\]
has a $t$-distribution with $n - 1$ degrees of freedom. Here, $\bar{x}$ is the sample mean, $\mu$ is the hypothesized mean, $s$ is the sample estimate of the standard deviation and $n$ is the sample size.

**Note:** This testing procedure assumes that the original population the sample was drawn from has a Normal distribution. That is often, but not always the case in biological applications. If this assumption is untrue, then $p$-values computed from the test statistic are uninterpretable. If the sample size is large, normality of the sample becomes less important. How large a sample has to be before the normality assumption can be dropped depends on the actual shape of the population distribution. The more "normalish" the population distribution is, the smaller the sample has to be. But for truly bizarre population distributions (e.g., bimodal) samples may have to be quite large ($n \geq 50$) for the $t$-test to be valid.

**In R:** To conduct a one-sample $t$-test in R, read in your numerical data. Suppose the data is contained in a vector called `x`. Then the $t$-test is conducted by typing `t.test(x, alternative, mu)` where `alternative` is either "two.sided", "less", or "greater" and refers to the way the alternative hypothesis is phrased. `mu` is the hypothesized value of the population mean.

**Example:** Systolic blood pressure was measured in 12 men over the age of sixty. The average measurement was 143 with a standard deviation of 11. Is this evidence enough to conclude that the mean systolic blood pressure in older men is in the stage 1 hypertension range (> 140)?

**Z-test:** Suppose that the statement of interest concerns a population proportion (such as the proportion of all women in the U.S. who will die of coronary heart disease). To decide the statement, a sample is taken from the population and a binary characteristic is observed on each individual.

The test statistic

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim \text{Normal}(0, 1)$$

has a standard normal distribution. Here, $\hat{p}$ is the estimated sample proportion, $p_0$ is the hypothesized population proportion and $n$ is the sample size.
**Note:** For this testing procedure to be valid, the sample has to be large. Usually, large is defined as containing at least ten observations of each type. That means that if the binary trait is rare, then larger samples will be required. The reason why large samples are needed is that the sample proportion which theoretically has a hypergeometric distribution can then be approximated by a binomial distribution which, in turn, can be approximated by a normal distribution.

**In R:** To conduct a $z$-test for a population proportion in R, type `prop.test(x,n,p, alternative)` where $x$ is the count of successes, $n$ is the count of trials and $p$ is the hypothesized success probability. As in the case of the $t$-test, the alternative can be specified as either “two.sided”, “less” or “greater”.

**Example:** Prevnar is a vaccine for meningitis usually given to infants. In a clinical trial, Prevnar was given to 710 children, of whom 72 experienced a loss of appetite. Competing medications cause about 13.5 percent of children to experience a loss of appetite. Can we conclude that the percentage of children experiencing a loss of appetite from Prevnar is significantly less than for other medications?