BST 675 – Fall 2011 – Dr. Charnigo

Unit II: Discrete Random Variables

a. Motivating Case Study #1: Assessing the potential of a new pharmaceutical

This motivating case study is adapted from pp. 132-133 of Larsen and Marx.

A pharmaceutical company is experimenting with a new affordable AIDS medication that may strengthen a victim's immune system. Thirty monkeys infected with HIV have been given the drug. Researchers will wait six weeks and then count the number of monkeys whose immunological responses show marked improvement. Any inexpensive drug that works on at least 60% of those consuming it would be considered a major breakthrough. Any medication that works on less than 50% of those consuming it would not have much commerical potential.

Already a few objections can be anticipated. First, for a disease as lethal as AIDS, an inexpensive medication that works on even a small fraction of those consuming it may be quite marketable, provided that the side effects are not overwhelming. Second, immunological responses are naturally continuous phenomena, so some subjectivity is introduced in proposing a threshold that constitutes marked improvement. Third, the medication may turn out to be somewhat less or somewhat more effective in humans than in monkeys.

Notwithstanding these objections, this scenario provides a motivation for defining a random variable.

Letting I be shorthand for "[Marked] Improvement" and N be shorthand for "No [Marked] Improvement", we can define a sample space S consisting of all possible strings of length 30 constructed from the symbols I and N. For instance, $IIIIIIIIIIIIIIINNNNNNNNNNNNNNN$ is one element of the sample space, which corresponds to the first fifteen monkeys responding well to the medication and the last fifteen monkeys responding poorly.

Assuming that $IIIIIIIIIIIIIIINNNNNNNNNNNNNNN$ is not somehow indicative of an experimental error (e.g., the last fifteen monkeys getting the wrong medication), the pharmaceutical company should not really regard $IIIIIIIIIIIIIIINNNNNNNNNNNNNNN$ as importantly different from, say, $ININININININININININININININ$. The pharmaceutical company should care only about the number of I's appearing in the string. Indeed, an inferential statistician would call the number of I's a sufficient statistic, in the sense that the number of I's captures all the information that the experiment provides regarding the fraction of HIV-infected monkeys for which the medication is effective. However, we are getting ahead of ourselves, so let's step back and introduce the concept of a random variable.

Let X be a function that takes elements of the sample space S as inputs and produces as outputs real numbers. For instance, we can take X to be the function whose output is the number of I's appearing in a string such as $IIIIIIIIIIIIIIINNNNNNNNNNNNNNN$. The string $IIIIIIIIIIIIIIINNNNNNNNNNNNNNN$ is an input, while the corresponding output in that case is the real number 15. We then refer to X as a random variable. A common abuse of semantics is to identify the random variable with its output by saying, for instance, that X is the number of I's appearing in a string (rather than that X is a function whose output is the number of I's appearing in a string). However, we will live with the abuse of semantics because this is how we have been acclimated to random variables in our introductory statistical methods courses. In any event, the key idea to take home about X is that we are collapsing a 30-dimensional piece of information into a single real number.

Now, let p denote the fraction of HIV-infected monkeys for which the medication is effective. We may be inclined to say that $p = X/30$, but this is not quite right. The pharmaceutical company could be somewhat "lucky" or somewhat "unlucky" with this particular group of thirty monkeys. So, indeed, $X/30$ only estimates p. Nonetheless, the pharmaceutical company will have X available and not p, so the pharmaceutical company's decision about whether to proceed further with this medication will have to be made based on X.

Suppose, then, that the pharmaceutical company decides to proceed if and only if $X \geq 16$. Two questions of interest, which we will address at the end of Unit II, are as follows:

(a) If p equals (or exceeds) 0.60, then what is the probability that $X \leq 15$? This is the probability of (mistakenly) failing to pursue a marketable medication.

(b) If p equals (or falls short of) 0.50, then what is the probability that $X \geq 16$? This is the probability of (mistakenly) pursuing an unmarketable medication.

b. Motivating Case Study #2: Evaluating the adequacy of hospital resources

This motivating case study is adapted from pp. 277-278 of Larsen and Marx.

A hospital is located in a rural area that has twelve thousand elderly residents. The probability that any specific one of these twelve thousand elderly residents will have a myocardial infarction and will need to be connected to a special cardiac monitoring machine on any specific day is estimated to be one in eight thousand. Currently, the hospital has three such machines. Are the hospital's resources adequate?

At first glance, the answer may appear to be yes. With the aforementioned probability of one in eight thousand, we anticipate that on a given day one or two elderly residents will need the special cardiac monitoring machine. (I use the word "anticipate" rather than "expect" because, as we will see later in Unit II, "expect" has a technical meaning in probability.) There will be no problem on such a day because the hospital has three machines.

Yet, there is some possibility that the given day may be a "bad" day and that three elderly residents will need the machine. Or four, in which case there will be a problem because the hospital only has three machines. So, some more effort is required to address the question of resource adequacy.

Formally, we can define a sample space S consisting of all possible strings of length 12,000 constructed from the symbols M ("Myocardial Infarction") and N ("No Myocardial Infarction"). Yet, the hospital should care only about the number of M's in the string. Therefore, let us define the random variable X to be the number of M's in the string. The adequacy of the hospital's resources can then be assessed by computing $P(X > 3)$ and determining whether this probability exceeds a threshold set by the hospital administration. Such a threshold will reflect how often the hospital administration is willing to tolerate the denial of machine access to an elderly resident with a myocardial infarction.

For instance, suppose that the hospital administration is willing to tolerate an average of one such denial per 30 days. In this case, the threshold is $1/30$ and a finding that $P(X > 3)$ exceeds $1/30$ indicates that the hospital's resources are not adequate. If the hospital's resources are not adequate, then the hospital administration may consider acquiring a fourth special cardiac monitoring machine.

Now, $P(X > 4)$ will be less than $P(X > 3)$. (How do we know this?) Yet, $P(X > 4)$ will still be a positive number, not zero. So, with a small enough threshold, we may still find that the hospital's resources are not adequate even after acquisition of a fourth machine.

Actually, the same can be said about $P(X > 5)$ and acquisition of a fifth machine, or about $P(X > 6)$ and acquisition of a sixth machine. Really, the only way to ensure that no elderly resident is ever denied access is to acquire twelve thousand machines. However, that is unreasonable. Putting aside the massive acquisition (and maintenance) costs, there is also the problem of where the twelve thousand machines will be stored.

Thus, the hospital administrators have to ask themselves how often they are willing to tolerate a denial (once per 30 days on average? once per 90 days? once per year?) and then acquire a number of machines sufficient to ensure that denials do not occur intolerably often.

c. Probability mass functions and cumulative distribution functions (Cf. pp. 148-161 of Larsen and Marx)

Recall that a random variable X is a function from the sample space S to the set of real numbers $\mathbb{R}$. (There are also some technical requirements involving sigma fields that we will eschew.)

A question of interest is how we should define a probability involving a random variable X, say the probability that $X \in D$, where D is a subset of $\mathbb{R}$. For instance, if $D = [0, 1]$, then $X \in D$ means that $0 \leq X \leq 1$. Or, if $D = \{0\}$, then $X \in D$ means that $X = 0$.

To motivate an appropriate definition, suppose that we conduct a taste test in which each of 10 grocery shoppers is asked to try "Brand A" and "Brand B" of peanut butter and then state which one he or she prefers. Assuming that a response of "no preference" is not allowed, the sample space S may be taken to consist of strings of length 10 containing A's and B's. Two such strings are $A, A, A, B, B, B, B, B, B, B$ and $B, B, B, B, B, B, B, A, A, A$. Let X denote the number of shoppers who prefer Brand A. Suppose that all elements of S are equally likely. Implicit in this supposition is a belief that Brand A and Brand B are of identical quality (or, at least, that the typical grocery shopper is not capable of discriminating between them). Then we should have

$P(X = 0) =$

$P(X = 1) =$

and

$P(X = 2) =$

As the preceding example suggests, $X \in D$ should inherit its probability from the underlying sample space S. More specifically, we define

$$P(X \in D) := P(\{\omega \in S : X(\omega) \in D\}).$$

Regarding the notation, ω represents a generic element of the sample space S. In the preceding example, ω would be a string of length 10 containing A's and B's such as $A, A, A, B, B, B, B, B, B, B$ or $B, B, B, B, B, B, B, A, A, A$. Also, the colon $:$ is read out loud as "such that". Hence, the probability that X belongs to D is defined as the probability that was attached to the subset of S whose members get mapped to D by X. To illustrate, let us take $D := \{0\}$, so that $X \in D$ means $X = 0$. Then, since the only element of S that gets mapped to 0 is $B, B, B, B, B, B, B, B, B, B$, we define the probability that $X = 0$ to be the probability that was attached to $B, B, B, B, B, B, B, B, B, B$.

Since probabilities involving a random variable X are inherited from the underlying sample space S, all of the axioms and subsequent computational formulas from Unit I apply when we are evaluating probabilities involving a random variable X. In particular, for any subsets $D, D_1, D_2, \ldots \subset \mathbb{R}$, we have:

1. $P(X \in D) \geq 0$.
2. $P(X \in \mathbb{R}) = 1$.
3. $P(X \in \cup_{i=1}^{\infty} D_i) = \sum_{i=1}^{\infty} P(X \in D_i)$ when $D_i \cap D_j = \emptyset$ for $i \neq j$.

An interesting practical point is that, since the axioms and computational formulas from Unit I do apply when we are evaluating probabilities involving a random variable X, sometimes we don't actually need to define the sample space S explicitly when we are solving problems. For instance, suppose that we take systolic blood pressure measurements on 10 subjects and let X denote the average of the 10 measurements. If I tell you that $P(X \geq 140) = 0.40$ and $P(X > 150) = 0.20$, then you can determine $P(140 \leq X \leq 150)$ even though we said nothing about the underlying sample space S.

Now that we have discussed how to define probabilities involving a random variable X, we are in a position to define probability mass functions and cumulative distribution functions. Defining the latter first will be easier.

Let X be a random variable. The cumulative distribution function of X is denoted by $F(x)$ and defined as $P(X \leq x)$. Note that (lower case) x is a placeholder for 2 or 8 or any other real number. That is, $F(2) = P(X \leq 2)$ and $F(8) = P(X \leq 8)$.

For example, suppose that there are two children in a family, not twins, and each child has a 50% chance of inheriting a disease. Since the children are not twins, we may assume that whether the first child inherits the disease is independent of whether the second child inherits the disease. In this case, the probability that neither child inherits the disease is $0.5 \times 0.5 = 0.25$, the probability that both children inherit the disease is also $0.5 \times 0.5 = 0.25$, and the probability that exactly one child inherits the disease is $0.5 = 1 - 0.25 - 0.25$. Let X denote the number of children who inherit the disease. What is $F(x)$?

A cumulative distribution function $F(x)$ must satisfy three conditions:

1. $F(x)$ is nondecreasing in x.
2. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
3. $F(x)$ is right continuous in x.

By right continuous we mean that, for any x, $\lim_{\delta \searrow 0} F(x + \delta) = F(x)$. Note that any continuous function is right continuous but not vice versa.

To illustrate why the first condition must hold, suppose that $F(1) = 0.5$ and $F(2) = 0.3$. Then

$$0.3 = P(X \leq 2) = P(1 < X \leq 2) + P(X \leq 1) = P(1 < X \leq 2) + 0.5,$$

which forces $P(1 < X \leq 2) = -0.2$. That violates the first axiom of probability.

To illustrate why the second condition must hold, suppose that $\lim_{x \to \infty} F(x) = 2$. By the definition of limit, we can find a large x such that $F(x)$ is as close to 2 as we like. So, maybe $F(800) = 1.9$. However, by monotonicity of probability, $P(X \leq 800) \leq P(X \in \mathbb{R})$. So, if the former probability is 1.9, then the latter probability cannot be 1. That violates the second axiom of probability. An almost identical argument can be made if 2 is replaced by any other number greater than 1, and, although not as obviously, we can also force a violation if 2 is replaced by any number less than 1.

A good probability exercise is to determine whether a proposed cumulative distribution function is valid. For instance, does $1_{\{0 \leq x \leq 10\}} Cx + 1_{\{x > 10\}}$ define a valid cumulative distribution function? (The answer may depend on C.) What about $1_{\{0 \le x \le 10\}} C(x^2 - x) + 1_{\{x > 10\}}$?

If $F(x)$ is a step function (i.e., piecewise constant except at values x assumed by X with positive probability), then we say that X is a discrete random variable. Some authors define a discrete random variable as one that can assume a finite or countably infinite number of values. This definition is acceptable because a nondecreasing function can have at most a countably infinite number of steps. However, the proof of that assertion is delicate, so I omit it.

If $F(x)$ is continuous (i.e., both right continuous and left continuous), then we say that X is a continuous random variable. Some textbook authors (incorrectly) assert that a continuous random variable is one that can take on any value in a continuum. Although a continuous random variable can take on any value in a continuum, not all random variables that can take on any value in a continuum are continuous. Consider, for example, a random variable whose cumulative distribution function is $1_{\{0 \le x < 2\}} 0.25x + 1_{\{x \ge 2\}}$. Such a random variable may describe the length of time that a student spends on a two-hour in class examination. There is a 50% chance that the student will have his or her examination snatched away by the proctor at the end of two hours!

Next we define the probability mass function of a discrete random variable X as $f(x) := P(X = x)$ for any $x \in \mathbb{R}$. Note that (lower case) x is a placeholder for 2 or 8 or some other number, so that $f(2) = P(X = 2)$ and $f(8) = P(X = 8)$. A probability mass function must satisfy $f(x) \ge 0$ for all $x \in \mathbb{R}$ and $\sum_{x \in \mathbb{R}: f(x) > 0} f(x) = 1$. Conversely, any function $f(x)$ with these properties may be interpreted as a probability mass function.

For example, let λ be a positive number and put $f(x) := 1_{\{x \in \{0,1,2,\ldots\}\}} C(\lambda) \lambda^x / x!$. How can $C(\lambda)$ be chosen so that $f(x)$ is a probability mass function? If X is a random variable with this probability mass function, which we write in shorthand as $X \sim f(x)$, then what is $P(X > 1)$?

d. Expected values, means, and variances (Cf. pp. 173-203 of Larsen and Marx)

Let X be a discrete random variable with probability mass function $f(x)$, let $g(x)$ be any function (satisfying technical conditions that we will take for granted in this course), and let $\mathcal{X}$ denote the set of real numbers x on which $f(x) > 0$ (called the "support set" or simply the "support" of X).

We define the expected value of $g(X)$, also called the mean of $g(X)$, as

$$E[g(X)] := \sum_{x \in \mathbb{R}: f(x)>0} g(x)f(x) = \sum_{x \in \mathcal{X}} g(x)f(x) = \sum_{x \in \mathcal{X}} g(x)P(X = x),$$

provided that the sum is absolutely convergent. If the sum is not absolutely convergent, then we say that $E[g(X)]$ does not exist as a finite number. If the sum is not absolutely convergent and $g(x) \geq 0$ for all but finitely many $x \in \mathcal{X}$, then we may also say that $E[g(X)] = \infty$. Note that the expected value of $g(X)$, when it does exist as a finite number, is just a weighted average of all values of $g(X)$ that occur with nonzero probabilities, the weights being the probabilities themselves.

For example, let X be the number of flips required to get your first tails on a fair coin. Then X has probability mass function $f(x) := (1/2)^x$ for $x \in \{1, 2, \ldots\}$. Putting $g(x) := x$, we have

$$E[X] = \sum_{x=1}^{\infty} x(1/2)^x = \sum_{x=1}^{\infty} (1/2)^x + \sum_{x=2}^{\infty} (1/2)^x + \sum_{x=3}^{\infty} (1/2)^x + \cdots$$
$$=$$

This computation suggests an answer if you were asked, before you flipped the coin, what you expected X to be. However, despite what we saw in this example, we are not generally guaranteed that $E[X] \in \mathcal{X}$, even when $E[X]$ exists as a finite number. (Can you construct another example in which $E[X]$ exists as a finite number but does not belong to $\mathcal{X}$?)

Again, let X be the number of flips required to get your first tails on a fair coin. Putting $g(x) := 2^x - C$ for some positive constant C, we have

$$E[g(X)] = \sum_{x=1}^{\infty} \{(2^x - C)(1/2)^x\} = \sum_{x=1}^{\infty} \{1 - C(1/2)^x\} = \infty.$$

To see the last equality above, let x^* be the smallest positive integer greater than $-\log C/\log(1/2)+1$. Then, for every positive integer $x \geq x^*$, we have $C(1/2)^x < (1/2)$. Hence,

$$\sum_{x=x^*}^{\infty}\{1-C(1/2)^x\} = \lim_{n\to\infty}\sum_{x=x^*}^{n}\{1-C(1/2)^x\} \geq \lim_{n\to\infty}\sum_{x=x^*}^{n}(1/2)$$
$$= \lim_{n\to\infty}(n-x^*+1)(1/2) = \infty.$$

Since $\sum_{x=1}^{x^*-1}\{1-C(1/2)^x\}$ is finite, we also have $\sum_{x=1}^{\infty}\{1-C(1/2)^x\} = \infty$.

Here is an interpretation of the above result. I sell you a fair coin for C dollars and promise to pay you 2^X dollars if you get your first tails on flip X. Your net winnings are $g(X)$. If $C < 2$, then your net winnings are positive with probability 1 and you will surely choose to play (assuming you are a rational human being). However, the calculations above show that your expected net winnings are positive infinity regardless of the price C. This is called the St. Petersburg Paradox and nicely demonstrates that, in real life situations that can be modeled probabilistically, our behaviors are not necessarily governed by expected values. (If they were, then nobody would play the lottery.)

Let us do another example. Suppose that X has probability mass function $f(x) := \exp[-\lambda]\lambda^x/x!$ for $x \in \{0,1,2,\ldots\}$, where λ is a positive real number. Putting $g(x) := x$, we have

$$E[X] = \sum_{x=0}^{\infty} x\exp[-\lambda]\lambda^x/x! = \sum_{x=1}^{\infty} x\exp[-\lambda]\lambda^x/x! = \lambda\sum_{x=1}^{\infty}\exp[-\lambda]\lambda^{x-1}/(x-1)!$$
$$=$$

Two important properties of expected value are linearity and monotonicity. Linearity says that, for any real constants c_1 and c_2,

$$E[c_1g_1(X)+c_2g_2(X)] = c_1E[g_1(X)] + c_2E[g_2(X)]$$

whenever all of the expectations exist as finite numbers. Moreover, if the last two expectations exist as finite numbers, then so does the first. Monotonicity says that if $g_1(x) \leq g_2(x) \leq g_3(x)$ for all $x \in \mathcal{X}$, then

$$E[g_1(X)] \leq E[g_2(X)] \leq E[g_3(X)].$$

In particular, if $E[g_1(X)]$ and $E[g_3(X)]$ exist as finite numbers, then so does $E[g_2(X)]$.

Consider the following problem. Suppose that X is a nonnegative discrete random variable for which $E[X^{2009}]$ exists as a finite number. Our goal is to prove that $E[X^c]$ exists as a finite number for any $c \in (0, 2009)$. At first, this seems impossible because we haven't been given enough information. In particular, neither the probability mass function nor the support set has been supplied, although we know that $\mathcal{X} \subset [0, \infty)$. Yet, we can use monotonicity to prove the claim. Our strategy will be to set $g_2(x) := x^c$ for $x \geq 0$ and then exhibit $g_1(x)$ and $g_3(x)$ with $g_1(x) \leq g_2(x) \leq g_3(x)$ for $x \geq 0$ such that $E[g_1(X)]$ and $E[g_3(X)]$ exist as finite numbers.

An obvious choice for $g_1(x)$ is   yielding $E[g_1(X)] =$
We would like to choose $g_3(x) := x^{2009}$, but unfortunately $x^{2009} < x^c$ when $x \in (0, 1)$. However, $x^c \leq 1$ when $x \in (0, 1)$, which gives us the idea to take $g_3(x) := 1 + x^{2009}$. Then by linearity we have $E[g_3(X)] = E[1] + E[X^{2009}] = 1 + E[X^{2009}]$, which is finite since $E[X^{2009}]$ was finite. Therefore, we conclude that $E[g_2(X)] = E[X^c]$ is finite.

Let X be a discrete random variable. For any integer $n \geq 1$, we define the n^{th} moment of X to be $E[X^n]$. For any integer $n \geq 2$, we define the n^{th} central moment of X to be $E[(X - \nu)^n]$, where $\nu := E[X]$ is assumed to exist as a finite number. If the n^{th} moment exists as a finite number, then so do all moments of lower order and the n^{th} central moment.

The second central moment $E[(X - \nu)^2]$ is called the variance of X. The variance describes how much X fluctuates around its expected value. The standard deviation of X is defined to be the positive square root of the variance. Unlike the variance, the standard deviation is expressed in the same units as X. For instance, if X is systolic blood pressure in mmHg, then the standard deviation of X is in mmHg while the variance is in $(\text{mmHg})^2$.

Three useful results, assuming all expectations and variances referred to exist as finite numbers, are as follows.

1. For any constants a and b, $Var[aX + b] = a^2 Var[X]$.
2. A computational formula for the variance is $E[X^2] - (E[X])^2$.
3. If $Var[X] = 0$, then $P(X = E[X]) = P(|X - E[X]| = 0) = 1$.

The first two results can be established by appealing to linearity of expectation. We prove the third result below. A useful computational tool for proof of the third result is that $E[1_{\{X\in A\}}] = P(X \in A)$ for any subset $A \subset \mathbb{R}$. To see this, put $g(X) := 1_{\{X\in A\}}$ and note that

$$E[1_{\{X\in A\}}] = \sum_{x\in\mathbb{R}:f(x)>0} 1_{\{x\in A\}} f(x) = \sum_{x\in A:f(x)>0} f(x) = P(X \in A).$$

Now, we prove the third result by contradiction. Suppose that $E[(X - E[X])^2] = 0$ but that $P(|X - E[X]| > 0) = \epsilon > 0$. Since

$$\{|X - E[X]| > 0\} = \cup_{j=1}^{\infty} \{1/j \le |X - E[X]| < 1/(j-1)\}$$

we have

$$0 < \epsilon = \sum_{j=1}^{\infty} P(1/j \le |X - E[X]| < 1/(j-1)).$$

By countable additivity, there must exist $j \in \{1, 2, \ldots\}$ such that

$$0 < \delta := P(1/j \le |X - E[X]| < 1/(j-1)) \le P(1/j \le |X - E[X]|).$$

Then, using monotonicity of expectation (twice) and the useful computational tool above, we have

$$\begin{aligned} E[(X - E[X])^2] &\ge E[(X - E[X])^2 1_{\{|X-E[X]|\ge 1/j\}}] \\ &\ge E[(1/j)^2 1_{\{|X-E[X]|\ge 1/j\}}] \\ &\ge (1/j)^2 P(|X - E[X]| \ge 1/j) \\ &\ge \delta/j^2 \\ &> 0. \end{aligned}$$

We have arrived at a contradiction. Therefore, we must conclude that $E[(X - E[X])^2] = 0$ implies $P(|X - E[X]| > 0) = 0$.

e. Moment generating functions (Cf. pp. 257-269 of Larsen and Marx)

The moment generating function of a random variable X is defined as $M_X(t) := E[\exp(tX)]$, where E denotes the expected value operator. If X is discrete, then we have $M_X(t) = \sum_{x \in \mathcal{X}} \exp(tx) f_X(x)$, where $f_X(x)$ is the probability mass function of X and $\mathcal{X} := \{x \in \mathbb{R} : f_X(x) > 0\}$ is the support set of X. I will define the expected value of a function of a continuous random variable in Unit III. Please note that, although Unit II focuses on discrete random variables, none of the three results below requires the random variables concerned to be discrete.

Result 1. Suppose there exists $h > 0$ such that $M_X(t) < \infty$ for all $t \in [-h, h]$. Then, for every positive integer n, $E[X^n]$ exists as a finite number and is equal to $\frac{d^n}{dt^n} M_X(t)|_{t=0}$. This is, in fact, why we refer to $M_X(t)$ as a moment generating function; we obtain moments of X by differentiating $M_X(t)$ and evaluating the derivative at $t = 0$.

Result 2. Suppose there exists $h > 0$ such that $M_X(t), M_Y(t) < \infty$ for all $t \in [-h, h]$. If $M_X(t) = M_Y(t)$ for all $t \in [-h, h]$, then X and Y have the same cumulative distribution function: $F_X(u) = F_Y(u)$ for any real u. Hence, two random variables that share the same (finite) moment generating function in a neighborhood of 0 have the same distribution. This result is sometimes useful for finding the distribution of a random variable X, especially when X has the form $c \times (X_1 + X_2 + \cdots + X_n)$ for a real constant c and random variables $X_1, X_2, \ldots, X_n$. While $f_X(x)$ may not be obvious in such a situation, $M_X(t)$ may be easy to calculate. Then, if $M_X(t)$ can be recognized as corresponding to a known cumulative distribution function, we may infer the distribution of X accordingly.

Result 3. Suppose there exists $h > 0$ such that $M_X(t), M_{X_1}(t), M_{X_2}(t), \ldots < \infty$ for all $t \in [-h, h]$. If $M_{X_i}(t) \overset{i \to \infty}{\to} M_X(t)$ for all $t \in [-h, h]$, then the cumulative distribution functions of $X_1, X_2, \ldots$ converge to the cumulative distribution function of X at all points where the latter is continuous: $F_{X_i}(u) \overset{i \to \infty}{\to} F_X(u)$ for any real u at which $F_X(u)$ is continuous. You will learn in BST 676 that this convergence of cumulative distribution functions is called "convergence in law" or, synonymously, "convergence in distribution". The Central Limit Theorem, which you learned about in your introductory statistical methods course and which you will revisit in BST 676, states that (under certain conditions, and with obvious notation) $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges in law to a standard normal random variable as $n \to \infty$. The Central Limit Theorem provides the theoretical justification for employing a Z table to make inference about μ via $\bar{X}_n$ when n is large, but the Central Limit Theorem itself is proved using this result on moment generating functions.

To see why the first result above holds, suppose that X is a discrete random variable with probability mass function $f_X(x)$ and support set $\mathcal{X}$. For $t \in (-h, h)$ we have

$$\frac{d^n}{dt^n} M_X(t) = \frac{d^n}{dt^n} \sum_{x \in \mathcal{X}} \exp[tx] f_X(x) = \sum_{x \in \mathcal{X}} \frac{\partial^n}{\partial t^n} \exp[tx] f_X(x) = \sum_{x \in \mathcal{X}} x^n \exp[tx] f_X(x).$$

Above, $\frac{\partial^n}{\partial t^n}$ denotes n consecutive partial differentiations in t. We will speak of partial derivatives in Unit IV, but for now all you need to know is that a partial derivative in t for any function of x and t is calculated by pretending that x is a constant and then taking an ordinary derivative in t. Also, note that the second equality above entails an interchange of summation and differentiation. While intuitively plausible and correct in this instance, such an interchange is not universally valid. However, discussion of the theoretical conditions justifying the interchange of summation and differentiation is beyond the scope of BST 675. Finally, from the above equalities, we obtain

$$\frac{d^n}{dt^n} M_X(t)|_{t=0} = \sum_{x \in \mathcal{X}} x^n f_X(x) = E[X^n].$$

Here is a toy example. Suppose that X has probability mass function $(1-p)1_{\{x=0\}} + p1_{\{x=1\}}$ for some constant $p \in (0, 1)$. Then

$$M_X(t) = E[\exp(tX)] = (1-p)\exp[t \times 0] + p\exp[t \times 1] = (1-p) + p\exp[t].$$

For any positive integer n, we have $\frac{d^n}{dt^n} M_X(t) = p\exp[t]$. Evaluating this expression at 0 returns p, so we conclude that $E[X^n] = p$ for any positive integer n. Of course, this is easy to confirm directly as well.

f. Binomial family (Cf. pp. 130-138 of Larsen and Marx)

A random variable X has the binomial distribution with parameters $p \in (0, 1)$ and $n \in \{1, 2, \ldots\}$ if

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} 1_{\{x \in \{0,1,\ldots,n\}\}}.$$

The usual interpretation of the binomial distribution with parameters p and n is that we conduct n independent trials, each of which results in "success" with probability p and "failure" with probability $1 - p$, and then let X denote the total number of successes. These are sometimes called Bernoulli trials.

Recall our example about interviewing grocery shoppers regarding their peanut butter preferences. Here we had $n = 10$ independent trials with success probability $p = 1/2$, where "success" is defined as preference for Brand A. This label of success is, of course, arbitrary. Actually, when the binomial distribution is used in epidemiology or in clinical trials, "success" is usually defined as the adverse outcome (such as death or disease) rather than the favorable outcome! Be that as it may, we are now in a position to revisit the grocery shopping example with a different choice of p, say $p = 3/4$ to be definite. Note that, with this different choice of p, the 2^{10} strings of length 10 are no longer equally likely and so we cannot calculate probabilities by counting.

Suppose, then, that $p = 3/4$ and $n = 10$. What is the probability that $P(X = 0)$? the probability that $P(X \leq 1)$?

Let us digress for a moment. Even though a formal treatment of hypothesis testing will have to wait until BST 676, you are all familiar enough with the basic concepts that the following may be of interest to you.

Suppose that we modify our example about interviewing grocery shoppers so that p is unknown. Now the object is to estimate p from data and test, say, $H_0 : p = 3/4$ against $H_1 : p < 3/4$. If none of the 10 shoppers prefers Brand A, then we may wish to calculate the probability of such an occurrence under the null hypothesis. If that probability is sufficiently small, then we may want to reject the null hypothesis. That probability is, of course, a p-value.

Now, if 1 of the 10 shoppers prefers Brand A, then we have to decide how to define a p-value. Our initial guess may be $P(X = 1)$, but a much better idea is to define it as $P(X \leq 1)$. Thus, the p-value is the probability under the null hypothesis of getting a result at least as extreme (in its opposition to the null hypothesis) as what we actually observed.

Why use $P(X \leq 1)$ instead of $P(X = 1)$ or, more generally, $P(X \leq x)$ instead of $P(X = x)$? To understand why, suppose that n were extremely large, say $n = 10,000$. Then $P(X = x)$ would be extremely small — much less than 0.05 — no matter the value of x. This would force us to reject $H_0 : p = 3/4$ no matter what we observed. In particular, observing $X = (3/4)n$ would force rejection of $H_0 : p = 3/4$, which is unreasonable.

I have, of course, simplified matters by writing $H_0 : p = 3/4$ instead of $H_0 : p \geq 3/4$. What if I had not? Since $P(X \leq x)$ is a decreasing function of p, the most conservative approach — i.e., that which would produce the largest $P(X \leq x)$ and make most difficult rejection of the null hypothesis — would still entail calculating $P(X \leq x)$ under $p = 3/4$.

I close this section by mentioning that

$$E[X] = np \quad \text{and} \quad Var[X] = np(1-p)$$

for a binomial random variable X with parameters p and n. These calculations can be performed directly (Cf. BST 675 Midterm Examination, Fall 2010). Alternatively, we may note that

$$X = T_1 + T_2 + \cdots + T_n,$$

where $T_i := 1_{\{\text{“success” on trial i}\}}$ for $i \in \{1, 2, \ldots, n\}$. From our toy example in the last section, we know that $E[T_i] = p$ and $E[T_i^2] = p$, whence $Var[T_i] = p(1-p)$. Thus, by linearity of expectation, we have

$$E[T_1 + T_2 + \cdots + T_n] = E[T_1] + E[T_2] + \cdots + E[T_n] = np.$$

In addition, variance has a linearity property when the random variables being summed are independent (to be discussed in Unit IV), so that

$$Var[T_1 + T_2 + \cdots + T_n] = Var[T_1] + Var[T_2] + \cdots + Var[T_n] = np(1-p).$$

g. Poisson family (Cf. pp. 275-292 of Larsen and Marx)

A random variable X has the Poisson distribution with parameter $\lambda \in (0, \infty)$ if

$$P(X = x) = \exp[-\lambda]\lambda^x/x! \quad \text{for } x \in \{0, 1, \ldots\}.$$

We note that $E[X] = \lambda$ and $Var[X] = \lambda$. We have already established the former equality earlier in this Unit. A similar approach can be used to find $E[X(X-1)]$, from which $E[X^2]$ and then $Var[X]$ can be straightforwardly calculated.

The usual interpretation of the Poisson distribution with parameter λ is that events are occurring over time (space) such that: (i) the number of events occurring in one time interval (spatial region) is independent of the number of events occurring in another nonoverlapping time interval (spatial region); (ii) as the length of a time interval (area of a spatial region) shrinks to zero, the probability of there being exactly one event divided by the length of the time interval (area of the spatial region) converges to λ; and, (iii) as the length of a time interval (area of a spatial region) shrinks to zero, the probability of there being more than one event divided by the length of the time interval (area of the spatial region) converges to 0. Then X can be defined as the number of events occurring over a time interval (spatial region) of unit length (area).

Although less relevant with modern computational resources, the Poisson distribution with parameter np has historically been used to approximate the binomial distribution with parameters p and n when p is small and n is large.

To justify the Poisson approximation to the binomial distribution, put $p_n := \min\{1, \lambda/n\}$ — why do we need the minimum? — and let X_n have the binomial distribution with parameters p_n and n for $n \in \{1, 2, \ldots\}$. For large enough n, λ/n will be less than 1 and we will have $p_n = \lambda/n \in (0, 1)$. Then (Cf. BST 675 Midterm Examination, Fall 2010)

$$M_{X_n}(t) = [p_n \exp[t] + (1 - p_n)]^n = [1 + \lambda(\exp[t] - 1)/n]^n.$$

Now we need to recall a result from calculus, namely that

$$\lim_{n\to\infty}[1 + a_n/n]^n = \exp[a]$$

for any sequence $\{a_1, a_2, \ldots\}$ with limit a. Of course, if all of the a_i are equal, then a is just their common value. So, we have

$$\lim_{n\to\infty} M_{X_n}(t) = \lim_{n\to\infty}[1 + \lambda(\exp[t] - 1)/n]^n = \exp[\lambda(\exp[t] - 1)].$$

On the other hand, if X is Poisson with parameter λ, then its moment generating function is

$$\begin{aligned} M_X(t) &= E[\exp(tX)] \\ &= \sum_{x=0}^{\infty} \exp[tx]\exp[-\lambda]\lambda^x/x! \\ &= \sum_{x=0}^{\infty} (\exp[t])^x \exp[-\lambda]\lambda^x/x! \\ &= \exp[-\lambda]\sum_{x=0}^{\infty}(\lambda\exp[t])^x/x! \\ &= \exp[-\lambda]\exp[\lambda\exp(t)] \\ &= \exp[\lambda(\exp[t] - 1)]. \end{aligned}$$

So, we see that

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t).$$

The third result for moment generating functions then tells us that, for any non-integral $x \in (0, \infty)$, we have

$$\lim_{n\to\infty} P(X_n \le x) = P(X \le x) = P(X \le \lfloor x \rfloor) = \sum_{u=0}^{\lfloor x \rfloor} \exp[-\lambda]\lambda^u/u!,$$

where $\lfloor x \rfloor$ represents the floor of x, which is the largest integer less than or equal to x.

h. Geometric and negative binomial families (Cf. pp. 317-327 of Larsen and Marx)

A random variable X has the negative binomial distribution with parameters $p \in (0,1)$ and $r \in \{1, 2, \ldots\}$ if

$$P(X = x) = \binom{r+x-1}{x} p^r (1-p)^x \quad \text{for } x \in \{0, 1, \ldots\}.$$

We have $E[X] = r(1-p)/p$ and $Var[X] = r(1-p)/p^2$.

The usual interpretation of the negative binomial distribution with parameters p and r is that we are conducting Bernoulli trials. However, instead of fixing the number of trials at n, we continue until we attain r successes. Thus, the number of trials is random. We let X denote the number of failures required to attain r successes.

An alternative definition for the negative binomial distribution entails letting Y denote the number of trials required to attain r successes. Obviously $Y = r + X$, so we have

$$P(Y = y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} \quad \text{for } y \in \{r, r+1, \ldots\},$$

$$E[Y] = r/p, \quad \text{and} \quad Var[Y] = r(1-p)/p^2.$$

Where necessary to distinguish these two definitions, I will refer to the latter as the "offset" negative binomial distribution, as the support set of Y is offset from 0.

Have another look at $E[X]$ and $Var[X]$. Can you formulate a conjecture about the circumstances under which there exists a good Poisson approximation to the negative binomial distribution?

Now, going in the other direction, suppose that you originally planned to model the number of events occurring over a time interval (spatial region) of unit length (area) as a Poisson random variable but that your data led you to believe that $E[X] = 5$ and $Var[X] = 6$. With what parameters p and r might you consider describing your data using the negative binomial distribution?

I digress here to note that a phenomenon like that described above — namely, too much variance to accommodate a Poisson distribution — is routinely encountered in ecological studies. These studies often attempt to relate a count dependent variable, such as the number of skin cancer cases occurring over a spatial region, to a variety of independent variables describing the region (such as climate or geography) but not specific people within the region. A Poisson regression model, which you will learn about in BST 760, proposes that the dependent variable follows a Poisson distribution whose mean is determined by the values of the independent variables. Yet, Poisson regression models rarely provide good fits to real data sets because of excess variance (formally called "overdispersion"). This is why epidemiologists often employ negative binomial regression models (or other more exotic alternatives such as "zero inflated" Poisson regression models).

A random variable X has the geometric distribution with parameter $p \in (0, 1)$ if

$$P(X = x) = p(1 - p)^x \quad \text{for } x \in \{0, 1, \ldots\}.$$

This is just a special case of the negative binomial distribution with parameters p and $r = 1$. As such, we have $E[X] = (1 - p)/p$ and $Var[X] = (1 - p)/p^2$.

An alternative definition for the geometric distribution entails putting $Y := 1 + X$, so we have

$$P(Y = y) = p(1 - p)^{y-1} \quad \text{for } y \in \{1, 2, \ldots\},$$

$$E[Y] = 1/p, \quad \text{and} \quad Var[Y] = (1 - p)/p^2.$$

Where necessary to distinguish these two definitions, I will refer to the latter as the "offset" geometric distribution, as the support set of Y is offset from 0.

Suppose that $X_1, X_2, \ldots, X_r$ are independent geometric random variables with parameter $p \in (0, 1)$. Can you formulate a conjecture about what will be the distribution of $X_1 + X_2 + \cdots + X_r$?

i. Resolution of motivating case studies

Our first motivating case study illustrated the definition of a random variable in the context of an experiment assessing the efficacy of a new pharmaceutical on 30 primate subjects. The sample space S consisted of all length 30 strings that could be constructed from the symbols "I" and "N", which represented improved and non-improved immunological responses respectively. We defined the random variable X to count the total number of I symbols in a string.

Now, having described binomial random variables, we are in a position to recognize that X is a binomial random variable, provided we are willing to assume that: (i) the probability of an improved immunological response is a constant $p \in (0, 1)$ for each monkey; and, (ii) the outcomes for the 30 monkeys are independent. For, indeed, the administration of the new pharmaceutical to each monkey constitutes a trial, and success on that trial can be defined as an improved immunological response.

Before we proceed to calculate probabilities, let us ponder an interesting philosophical question. Suppose I tell you that the new pharmaceutical was effective for 14 of the first 15 monkeys to which it was administered. Given this knowledge, would you think it likely or unlikely that the 16^{th} monkey would respond favorably to treatment?

This seems like a trick question, doesn't it? On the one hand, we assumed that the outcomes for the 30 monkeys were independent, so knowing what happened to the first 15 monkeys shouldn't tell you anything about what will happen to the 16^{th} monkey. On the other hand, common sense indicates that the 16^{th} monkey is likely to respond favorably.

The resolution to the trick question is that what happened to the first 15 monkeys is informative only insofar as you don't know p and are (either implicitly or explicitly) constructing an estimate of p based on the data from the first 15 monkeys. Before any experiments were conducted, you might have initially guessed a "neutral" value for p such as 0.5. Then, observing 14 successes in the first 15 trials would cause you to revise your guess upward to 0.93 or thereabouts. Hence, observing 14 successes in the first 15 trials would cause you to revise upward, from about 0.5 to 0.93, your level of belief that the 16^{th} monkey would respond favorably.

However, if I told you in the beginning that $p = 0.9$, then what happened to the first 15 monkeys should have no bearing on your level of belief that the 16^{th} monkey would respond favorably. That level of belief would be fixed at 0.9.

Returning to the calculation of probabilities, we can express $P(X \leq 15)$ as a function of p,

Obviously $P(X \geq 16)$ is just one minus this.

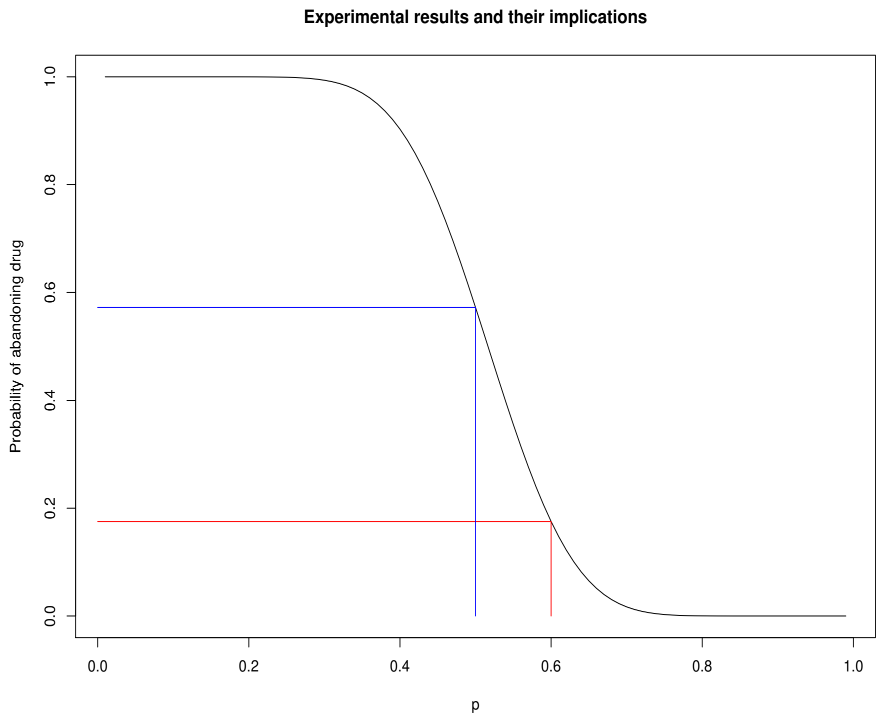
Recall that $X \leq 15$ is the event that the pharmaceutical company will abandon the drug on account of poor experimental results. This is a "correct" decision if $p \leq 0.50$ but an "incorrect" decision if $p \geq 0.60$. Likewise, $X \geq 16$ is the event that the pharmaceutical company will pursue the drug on account of satisfactory experimental results. This is a "correct" decision if $p \geq 0.60$ but an "incorrect" decision if $p \leq 0.50$.

The R code

```
postscript("C:/Documents and Settings/rjchar2/
My Documents/BST675F10/Pharma.ps")
p<-(1:99)/100
prob<-pbinom(15,30,p)
plot(p,prob,type="l",ylab="Probability of abandoning drug")
segments(0.6,0,0.6,pbinom(15,30,.6),col=2)
segments(0,pbinom(15,30,.6),0.6,pbinom(15,30,.6),col=2)
segments(0.5,0,0.5,pbinom(15,30,.5),col=4)
segments(0,pbinom(15,30,.5),0.5,pbinom(15,30,.5),col=4)
title("Experimental results and their implications")
dev.off()
```

creates the visual display on the next page.

Figure 1:



The visual display reveals that there is a 17.5% chance of incorrectly abandoning the drug if the drug works on 60% of primate subjects and that, moreover, this chance decreases as the drug becomes more effective. The visual display also reveals that there is a 57.2% chance of correctly abandoning the drug if the drug works on 50% of primate subjects and that, moreover, this chance increases as the drug becomes less effective. Hence, there is a 42.8% chance of incorrectly pursuing the drug if the drug works on 50% of primate subjects, and this chance decreases as the drug becomes less effective.

Our second motivating case study sought to assess the adequacy of hospital resources by asking whether $P(X > 3)$ was tolerably small, where X denoted the number of people who would need to use one of the three special cardiac monitoring machines available at the hospital.

Let us assume, perhaps less than realistically, that X is a binomial random variable with $n = 12,000$ trials and "success" probability $p = 1/8000$ on each trial, where "success" means that a person needs to use one of the cardiac monitoring machines. However, because n is very large and p is very small, we may proceed as if X were a Poisson random variable with mean $\lambda = np = 1.5$.

Then $P(X > 3)$ is given by the formula

which works out numerically to 0.0656. The implication is that, with three cardiac monitoring machines available, about twice a month there will be a person needing a cardiac monitoring machine who will not have access to it.

What if the hospital acquired a fourth cardiac monitoring machine? Then the relevant computation would be $P(X > 4)$, which works out numerically to 0.0186. Thus, with four cardiac monitoring machines, about every other month there will be a person needing a cardiac monitoring machine who will not have access to it.

Of course, the preceding computations are based on an approximation to an approximation. For, indeed, treating X as binomial in the first place was a sort of approximation. Is such approximation acceptable for the practical purpose of assessing the adequacy of hospital resources?

One way to answer is to calculate the frequency (over the last year, say) with which people needing to use a cardiac monitoring machine were turned away. Suppose, for example, that this happened on 25 out of the last 365 days. Then, although the assumptions underlying the binomial and Poisson probability modeling (in particular, independence of trials and constant "success" probability) may be suspect, when all is said and done, the observed fraction of days on which people are turned away, $25/365 = 0.0685$, is in reasonable concordance with the expected fraction of days as determined by the binomial and Poisson probability modeling, 0.0656. Hence, the binomial and Poisson probability modeling may be acceptable for the practical purpose of assessing the adequacy of hospital resources. On the other hand, if people were turned away on 50 out of the last 365 days, then any conclusions drawn from the binomial and Poisson probability modeling would be suspect.