Chapter 5 Notes and elaborations for STAT 141-Introduction to Statistics

Assignment:
Chapter 5 is pretty good, so I'll be following the text pretty closely. The book likes to talk about the mean of a random variable, I tend to call it the expected value of the random variable or the expectation of X, written as E(X).

I will not ask you to compute the variance or standard deviation of a random variable, so you don't have to do those parts of the homework from sections 5.2 and 5.3. Before you can do section 5.3 you'll need to understand how to do combinations and factorials as I show you below.

Do the following exercises:
5.1: 1-6 you should know except the baseball thing. 7-17 odd, 21, 23, 29
5.2: 1, 7, 11, 13, 15, 17, 19.
5.3: 1, 5, 7, 9, 11, 17, 21.

In the following sections, we discuss random variables, expectation of variables, combinations and factorials, and the infamous binomial distribution.

5.0 Random variables

Definition 5.0.0 A **random variable** is a variable whose numerical value is given by chance. There are also random elements which are what we call a random variable that assumes values that are not numerical. Do note, the proper mathematical definition for a random is complicated and unnecessary for our purposes.

Examples of random variables are assigning X to a number rolled on a die or letting Y = the blood pressure of an American. A random element might be given as Z = gender of a person.

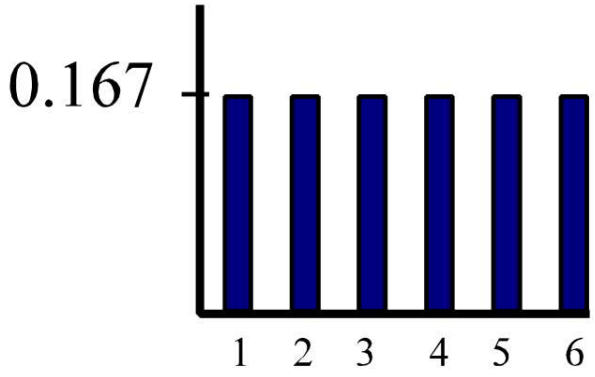
There are two types of random variables we will consider for this class, although it isn't hard to construct random variables outside these two.

A **discrete random variable** can assume only a countable number of distinct values. If a random variable can only be a finite number of distinct values, then it must be discrete. Examples of discrete random variables include the number of spoons in a kitchen, number in attendance at a basketball game, the number of times you need to shoot a basketball from the foul line until you make 3 baskets in a row, etc. A more complicated example is a person's height in centimeters, rounded to the nearest tenth.

A **continuous random variable** is one that assumes values from a range (a continuum). They require some different handling when it comes to probability. Height of a person is a continuous random variable. Consider that when we say somebody is 5' 2", that is only an approximation. It is not an exact measurement. In fact, the probability that someone is exactly 5' 2" is zero. The questions we ask about continuous random variables involve ranges or intervals. For example, what is the probability that if you choose a person at random, they will be between 4' 10" and 5' 6" tall? Or, what is the probability that if you choose a person at random, they are at least 5' 2" tall?

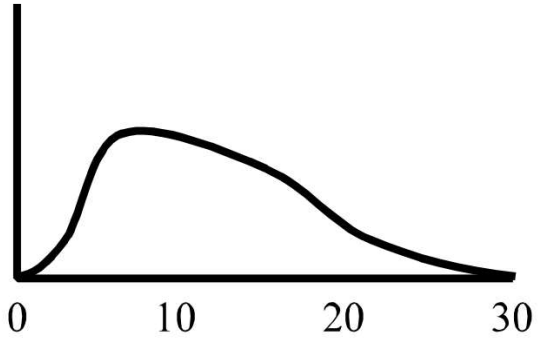
When we speak about a **distribution** or the distribution of a random variable, we are referring to the probability of a random variable being equal to a value, in the discrete case, or about it being within a given range, in the continuous case. When we make histograms of data, what we are actually trying to see is the distribution of the random variable.

Let X = the single roll of a fair die. Then the distribution of X is the whole numbers 1 to 6, each with probability 1/6 which is about equal to 16.7% Graphically, this can be seen below.



Distribution for a roll of a die.

Let the continuous random variable Y be equal to the lifespan of a car. Asking questions about a specific value of a continuous random variable doesn't actually make much sense. Questions about continuous variables come in the form of ranges of values. In the histogram of continuous random variable, probability is represented by area. The total area under the distribution curve is 100%. So in the example to the below, I would find the probability of a car lasting between 10 and 20 years by finding the area under the curve between 10 and 20. (Looks something like about 60% to my eyes.)



Lifespan of a car
(I made this up.)

Definition 5.0.1 A **distribution table** for a finite random variable consists of a list of all possible values of the variable together with their probabilities.

Example 5.0.0

Your experiment is rolling a die. Assign X to the number rolled. Find the distribution table for X.

X can take on any of the values 1, 2, 3, 4, 5, or 6. The probability of any of these values is 1/6.

X = x	**P(X = x)**
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

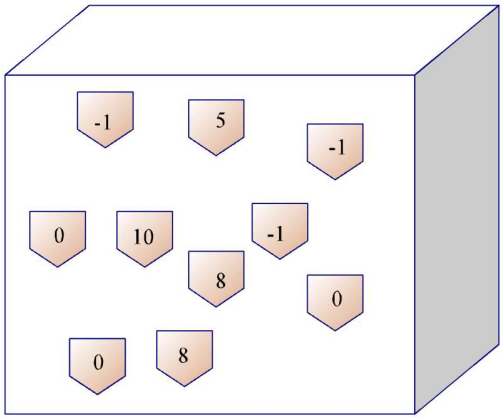
Things to note:

- **All** possible values of X are shown in the first column. They are the possible values of x.
- All their probabilities are there. Notice they sum to 1. Always a good idea to make sure this is true.
- Remember, probabilities must be between 0 and 1, inclusive.

Example 5.0.2

A box contains a total of 10 tickets. One ticket has 10 written on it, two tickets have 8 written on them, one ticket has 5 written on it, three tickets have -1 written on them, and the rest have 0 written on them. One ticket is drawn from the box and the random variable X is assigned the value written on the ticket. Find the distribution table for X.

First, you might draw this picture:



Why do my tickets look like baseball plates? Because it seemed like a good idea. Provided all draws are equally likely, we see that X can be -1, 0, 5, 8, or 10 and each value's probability is the total number of tickets of that type, divided by the total number of tickets, 10. So:

X = x	**P(X = x)**
-1	3/10
0	3/10
5	1/10
8	2/10
10	1/10

Do notice that random variables can have negative values. Zero even! The probabilities never can be negative, though.

Example 5.0.3

Frog pays $5 to play a "Crazy-Fives" with a fair pair of dice. If he rolls doubles, he gets his money back. If the roll sums to five, he wins $20. Let X be the net earnings that Frog gets in one play of Crazy-Fives. Find the distribution of X.

I will write (3,4) to mean that you roll a 3 on the first die and a 4 on the second die.

First, note that you can't roll doubles and sum to five at the same time. (Otherwise the rules aren't clear.) Second, what can happen here?

There are 36 equally likely outcomes for the roll of a pair of dice. But only 3 possible values of X.

- Pay $5 win $5 means no net gain or loss. So that totals to $0.
- Pay $5 win $20 means a total of $15 earned.
- Pay $5 win $0 means a net gain of -$5.

How can these happen?

- Totals to $0. This happens for the 6 possible ways to get doubles.
- Totals to $15. This happens with any of these rolls: (1,4), (4,1), (2,3), and (3,2). That's 4 ways.
- Totals to -$5. This happens for all the other rolls. 36 – 6 – 4 = 26 ways.

X = x	**P(X = x)**
-5	26 / 36
0	6 / 36
15	4 / 36

Definition 5.0.2 The **expected value** of a random variable is a theoretical number that is the long-term average of many independent trials of a random variable. ("In the limit" if you've had calculus. If not, don't worry about it.)

Imagine a box with only two numbers in it: 0 and 1. Assuming an equally-likely experiment of drawing from this box, we "expect" to see, in 100 draws, exactly 50 0's and 50 1's. If we take the average of these 100 draws you get 0.5. This is the expected value for this experiment. Modifying it just a little, imagine that the box instead contains three 0's and only one 1. From 100 draws you would "expect" to see 75 0's and 25 1's. The expected value for this experiment is 0.25. This isn't a truly useful way to compute expectation, but it might help in understanding what it means.

In the case of a finite discrete random variable, the expected value of a random variable is the weighted mean of all the values the variable can take where the weight for each value is the probability that that value occurs.

That is, suppose a random variable X can take values of x_1, x_2, x_3, and so on up to x_n. Assume that, for each one of these values that $p_i = P(X = x_i)$. Because summing all the probabilities is 1, we don't have to divide in the formula for weighted voting. And the expected value of X, written as E(X) is given by:

$$E(X) = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n$$

If the random variable is continuous or otherwise can attain an infinite number of different variables, we need calculus to compute the expectation. (Integrals and limits.) So now I will start to teach you calculus. . . just kidding. I won't ask you to compute expectations of these types of distributions. But the interpretation of the expected value is this same no matter which type of random variable we are discussing, including types I haven't even mentioned.

Again, you can think of expected value like this: If you repeat an experiment that gives you the random variable many, many times and take the average of all the X's, you are approximating the expected value of the random variable associated with the experiment. To get the actual expected value, we repeat the experiment an infinite number of times. So, the expected value of a random variable is a theoretical object, since we can't really repeat the experiment an infinite number of times.

Let's do a simple example.

Example 5.0.4
Your experiment is rolling a die. Assign X to the number rolled. Find the expected value of X.

We found this distribution table in example 5.0.1.

X = x	P(X = x)
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

So, using the formula for the expected value of X, we have:

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{21}{6} = 3.5$$

Notice that, if we have a distribution table, finding the expected value is easy. Multiply the two columns together. Add these new numbers.

So, when the experiment is rolling a die, and X is the random variable assigned to the number rolled, then we have E(X) = 3.5. Notice that the expected value of the random variable is not a value in the sample space. That's alright. *On average*, we expect to get a 3.5 when we roll a die. Not on a specific roll.

Example 5.0.5

A box contains a total of 10 tickets. One ticket has 10 written on it, two tickets have 8 written on them, one ticket has 5 written on it, three tickets have -1 written on them, and the rest have 0 written on them. One ticket is drawn from the box and the random variable X is assigned the value written on the ticket. Find the distribution table for X.

We found the distribution table in example 5.0.2. I have copied it here.

X = x	**P(X = x)**
-1	3/10
0	3/10
5	1/10
8	2/10
10	1/10

To find the expected value, the formula tells us to multiply the two columns together and add the numbers.

$$E(X) = -1\left(\frac{3}{10}\right) + 0\left(\frac{3}{10}\right) + 5\left(\frac{1}{10}\right) + 8\left(\frac{2}{10}\right) + 10\left(\frac{1}{10}\right) = \frac{28}{10} = 2.8$$

Example 5.0.6

Frog pays $5 to play a "Crazy-Fives" with a fair pair of dice. If he rolls doubles, he gets his money back. If the roll sums to five, he wins $20. Let X be the net earnings that Frog gets in one play of Crazy-Fives. Find the expected value of X.

We found the distribution table in example 5.0.3. I have copied it here.

X = x	**P(X = x)**
-5	26 / 36
0	6 / 36
15	4 / 36

To find the expected value, the formula tells us to multiply the two columns together and add the numbers.

$$E(X) = -5\left(\frac{26}{36}\right) + 0\left(\frac{6}{36}\right) + 15\left(\frac{4}{36}\right) = \frac{-74}{36} \approx -2.055$$

What does this mean? (I can never say that without thinking about double-rainbow man. Need a pick-me-up? Go look up the video. Skip to the 2:30 part or so if you're super Tic-Toc impatient.)

Let me start again. The expected value of the earnings for one play of the game is about -$2.06. That is, on average, you should expect to pay about that much each time you play. Of course, you never know what's going to happen if you play a handful of times. But in a place like a casino, you can just about set your watch by watching the amount of profit. If you have hundreds of people playing, and know the expectation per game (and they know it for ever single game) the **law of large numbers** kicks in. Go ahead and read the Wikipedia entry on "the law of large numbers" if you have time.

No, they don't have to load the dice or mark the cards at all. The game is already rigged for you to lose right at the start, and they don't hide the mathematics of it. And that's why this probabilist sees gambling on game machines at a casino (or scratch-off tickets, etc.) as a stupid tax. As in more stupider, more taxier.

5.1 Factorials and Combinations

Factorials are really not too bad. They are defined only for whole numbers starting with zero: 0, 1, 2, 3, 4, . . . We'll start with some examples, and I'm sure you'll get it.

5! = 5*4*3*2*1 = 120 4! = 4*3*2*1 = 24

3! = 3*2*1 = 6 2! = 2*1 = 2

In general, we write

n! = n*(n-1)*(n-2)* . . .* 3 * 2 *1.

There are some special factorials to remember, namely 1! and 0!.

1! = 1 and

0! = 1.

A couple of more properties of factorials before we get to combinations:

Factorials are not distributive operators. That is, (5-3)! must be computed as (5-3)! = (2)! = 2! = 2. It is crucial that you know (5-3)! ≠5! - 3!. Just do the work inside the parenthesis first.

Also, know that 4! * 3! ≠ 12!. You must first compute each factorial, and then multiply the resulting numbers (or just expand them all out and multiply).

Last but not least, we don't define factorials for negative numbers.

Now on to combinations. There are many uses and places in mathematics for combinations, but I will not burden you with them here. We need to merely be able to compute them in order to study the binomial distribution.

There are two commonly used symbols for combinations. We will follow the book (this material is near the end of Chapter 4) and use nCr , which is read "the combination of n items taken r at a time." Thus ${}_5C_3$ is read "the combination of 5 items taken 3 at a time" or "five choose three". The other commonly used symbol is $\binom{n}{r}$

Here is the formula:

$$\text{nCr} = \binom{n}{r} = \frac{n!}{(n-r)!\,r!}$$

Computing a combination. Notice that much of a combination will cancel after expanding the factorials.

Find ${}_7C_2$.

$$ {}_7C_2 = \frac{7!}{(7-2)!2!} = \frac{7!}{5!2!} = \frac{7*6*5*4*3*2*1}{5*4*3*2*1*(2*1)} = \frac{7*6}{2} = 21 $$

One more.

Find ${}_9C_6$.

$$ {}_9C_6 = \frac{9!}{(9-6)!6!} = \frac{9!}{3!6!} = \frac{9*8*7*6*5*4*3*2*1}{(3*2*1)6*5*4*3*2*1} = \frac{9*8*7}{3*2*1} = 3*4*7 = 84 $$

5.2 **Binomial Probability**

There are many, many probability distributions that we will not study (most of them). We study the binomial probability distribution at this level simply because it is easy to understand and it has many practical applications. It's one of the more famous discrete probability distributions.

Definition 5.2.0 A **binomial experiment** is an experiment with a fixed number of independent trials, where each outcome falls into exactly one of two categories.

The two categories are usually called 'success' and 'failure'. Some examples are experiments where the outcomes are right or wrong, yes or no, defective or not defective, win or lose, boy or girl, eyes are blue or eyes are not blue, etc. Whatever we are counting is the 'success.' For example, if you are running an experiment where the outcomes are right or wrong, like a multiple choice test, and you are finding, say, the probability that you get 10 out of 20 questions right if you randomly guess, then the 'success' is right and the 'failure' is wrong. If instead you were finding the probability that you get 5 out of 20 wrong if you randomly guess, the 'success' is wrong and the 'failure' is right. So, be conscientious of the categories when doing calculations.

By a fixed number of independent trials, we mean that our experiment actually consists of an experiment, called a trial, repeated a fixed number of times, and each trial is independent of every other. For example, rolling a die 10 times (each roll is a trial), taking a 20 question multiple choice test via random guessing (each question is a trial), recording whether or not the students in your online class have blue eyes (each student is a trial), etc.

Definition 5.2.1 A **binomial random variable** counts the number of successes in a binomial experiment.

Let X = the number of successes in an experiment. Then X is a binomial random variable if the following 3 requirements are met

1. <u>The experiment has a finite number of trials, known in advance</u>.

2. <u>Each trial is independent of every other trial.</u> The previous statement holds deep mathematical meaning, but much of it is beyond the scope of this course. I suggest you review the concept of independence from the probability section. Basically none of the trials have any bearing on any of the others.

3. <u>The probability of success is the same for all trials</u> and P(success) + P(failure) = 1.

Definition 5.2.2 A binomial probability distribution is the probability distribution of a binomial random variable.

Before we go any further, we need some notation.

- We use *n* to denote the number of trials (remember this is fixed).
- S and F denote success and failure, the two possible categories of all outcomes.

- The probability of success in one of n trials is p, i.e., P(S) = p.
- The probability of failure in one of n trials is q, i.e., P(F) = q.
- We use x to denote the specific number of successes in n trials.
- Finally, P(X = x) = the probability of getting exactly x successes among the n trials.

There are a couple of things I'd like to mention with regards to the notation. First, it would benefit you to think of n, the number of trials, as synonymous with sample size. Secondly, x, the specific number of successes in n trials, is what we call the realization of X, our binomial random variable. Your book is just simply wrong in writing P(X) in this context.

Now I think we are ready for the binomial (probability) formula:

Equation 5.2.0. If X is a binomial with n trials where each trial has probability p of success then the probability of exactly x successes is:

$$P(X = x) = \binom{n}{x} p^x q^{(n-x)}$$

- The combination is the number of ways to get x successes which means $(n - x)$ failures.
- p^x is the probability of getting x successes in x independent trials.
- q^{n-x} is the probability of getting n-x successes in n-x independent trials.

Let's disassemble this formula via an example.

Example 5.2.0

Consider a 4 question multiple choice quiz. Each question has 5 options. It's written in ancient Greek, which you forgot to study, so you answer each question by randomly guessing.

What's the probability of getting exactly 3 correct?

First, let's digest this problem. We'll call getting a problem correct a "success" in this context.

Let p = probability of guessing correctly in a single (trial) question.
Let q = probability of guessing incorrectly.
P(S) = p = 1/5 = 0.2 and P(F) = q = 4/5 = 0.8.
There are n = 4 trials.
We're interested in exactly x = 3 successes and n-x = 1 failures.

Now we can substitute into the formula:

$$P(X = 3) = \binom{n}{x} p^x q^{(n-x)} = \binom{4}{3} 0.2^3 0.8^{(1)}$$

$$= \frac{4*3*2*1}{(3*2*1)*(1)} * 0.008 * 0.8 = 4 * 0.008 * 0.8 = 0.0256$$

What are all the values that X could be?

Considering that X counts the number of successes in 4 trials, it could be 0, 1, 2, 3, or 4. We just calculated the probability that X = 3. Hopefully you can now find any of these probabilities with the use of the binomial formula given in equation 5.2.0. You should check that you can find one or more of these other values. I found them all and arranged them in a distribution table as given below.

Let X count the number of correct answers on our 4 question multiple choice quiz. Then the probability distribution table of X is as follows:

X	0	1	2	3	4
P(X)	0.4096	0.4096	0.1536	0.0256	0.0016.

What is the expected value of X?

Again, we multiply the values of X by it's probabilities and add.

E(X) = 0*0.4096+1*0.4096+2*0.1536+3*0.0256+4*0.0016 = 0.8.

Now, let's think about this. 4 trials. Each one is independent. Each has 20% of success. How many successes would you expect, intuitively? 4 * 0.2 = 0.8. As it turns out, this intuition can be proven. (We won't do that here.)

Equation 5.2.1 If X is a binomial with *n* trials where each trial has probability *p* of success then the expectation of X is given by:

$$E(X) = np.$$

Example 5.2.2 Roll a die until you get a four. Let Y be count of the number of rolls it takes. Is this a binomial random variable? No. The number of trials is random. A binomial random variable will always have a fixed number of trials.

Example 5.2.3 Draw from a box of 20 marbles randomly and without replacement exactly 4 times. Let Z be the number of green marbles drawn. Suppose there are 5 green marbles and 15 other color marbles in the box. Is Z binomial? No. The probability of drawing a green marble will change on the second draw depending on the results of the first draw. In a binomial experiment, the probability is the same in every trial.

Example 5.2.4

The experiment is to roll a die 5 times. Let X = number of threes that are rolled. Answer the following questions.

1. Is X a binomial random variable?
2. Find the probability of tossing 4 threes.
3. Find the probability of tossing at least 2 threes.
4. What is E(X)?

1. There are going to be exactly 5 trials and each trial is independent of each other. We'll consider "roll a 3" a success and "do not roll a 3" as a failure. Those probabilities won't change between trials. Yes, X is a binomial random variable.

2. We must find the probability of tossing 4 threes. Let's get everything we need:

n = 5, p = 1/6, q = 5/6, and x = 4.

So, we have:

$$P(X=4) = \binom{n}{x} p^x q^{(n-x)} = \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^1 = \binom{5}{4} * \left(\frac{1^4}{6^4}\right)\left(\frac{5^1}{6^1}\right) = \binom{5}{4} * \left(\frac{1^4 * 5^1}{6^4 * 6^1}\right)$$

$$= \frac{5*4*3*2*1}{(4*3*2*1)*(1)} * \left(\frac{5}{6^5}\right) = \left(\frac{5^2}{6^5}\right) = \left(\frac{25}{7776}\right)$$

$$\approx 0.0032.$$

3. We must find the probability of tossing at least 2 threes. What works? x = 2, 3, 4, and 5 all work. We could repeat what we did in problem 2 for all four values of x, (well, we just did x = 4 so we wouldn't do that again) and then add these 4 numbers. But I have a lazier way of doing this.

How would we find the probability of not tossing at least 2 threes? That would be x = 0 or x = 1. Using the binomial formula with n = 5, p = 1/6, q = 5/6, and x = 0 simplifies to:

$$P(X = 0) = 3125/7776 \approx 0.402$$

And again with n = 5, p = 1/6, q = 5/6, and x = 0 yields:

$$P(X = 1) = 5*(625/7776) \approx 0.402$$

Adding these, we see that P(tossing fewer than 2 threes) ≈ 0.804

$$P(\text{tossing at least 2 threes}) = 1 - P(\text{tossing fewer than 2 threes}) \approx 1 - 0.804 = 0.196$$

4. Now we calculate the expected value of X. We know from Equation 5.2.1 that

$$E(X) = 5*(1/6) = 5/6.$$

Let’s just write out the probability distribution table and check this the old-fashioned way.

X	0	1	2	3	4	5
P(X)	3125/7776	3125/7776	1250/7776	250/7776	25/7776	1/7776

Using the probability distribution table, we get

$$E(X) = 5/6$$