Lecture 3: Some Special Distributions

MATH E-156: Mathematical Foundations of Statistical Software

February 5, 2018

1 Introduction

In this lecture, we will look at three very special probability distributions (in your homework you'll meet two more important distributions).

2 Distributions and Parameters

In our review of basic probability theory, sometimes we worked with a specific concrete random variable, such as our example of the colored balls in the urn. But there were also situations that were more general, in which we had a whole family of random variables. For instance, in our final example, we worked out all the theory for uniform random variables on the range $[0, b]$, and this is actually a whole collection of uniform random variables, one for each value of b. Similarly, in your homework you developed the theory of exponential random variables, geometric random variables, and Pareto random variables. We call these sets of related random variables "probability distributions", and this implies that we are thinking about not just a single particular random variable, but a whole family of random variables that somehow share common characteristics. So, rather than talk about a certain exponential random variable X or another exponential random variable Y, we will think about **all** exponential random variables, and we will be interested in making general statements that apply to every memeber in the group.

Typically, the members of a probability distribution will all have densities that have a standard algebraic form, except that certain values have to be

specified. For instance, if X is an exponential random variable, then it has the density function

$$f_X(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, \ x > 0$$

This describes a family of density functions, which are indexed by λ. That is, if we select the value $\lambda = 2$, then we have the density function

$$f_X(x) = 2e^{-2x}, \quad x > 0$$

Now we have a density function with all positive real numbers as its support. On the other hand, if we select the value $\lambda = 3$, then we have the density function

$$f_X(x) = 3e^{-3x}, \quad x > 0$$

So for each value of λ, we get a different density function. In this case, we call λ a parameter, and we say that the exponential distribution has one parameter λ. More generally, a probability distribution can have multiple parameters, and distributions that can be indexed using parameters are called *parametric* distributions. In this course, we will be very interested in such parametric distributions, and a large amount of work will be devoted to trying to determine the values of unknown parameters using observed data.

3 The Binomial Distribution

The binomial distribution is our first special distribution. It's actually one of the hardest distributions to understand, because it involves some tricky counting arguments. We'll start out with a concrete example, and then develop the theory from there.

3.1 A Concrete Example: Part 1

Let's consider Tom Gravy, star quarterback of the New England Clam Chowder. Tom completes 60% of his passes, so we can say that the probability of Tom completing a single pass is 60%. Now we can ask: what is the probability that if Tom throws 5 passes, he will complete exactly 3 of them? In order to answer this question, we have to make some assumptions, the most important being that each pass is independent of all the others. We also

will assume that each pass is completed with 60% probability. These assumptions might not be entirely realistic, but they will serve as a good first approximation.

We can start by considering one possible sequence of passes, such that out of 5 attempts 3 are completed. I will use C to denote a completed pass, and I to denote an incomplete pass.

$$[C, I, C, C, I]$$

This notation means that Tom Gravy first threw a complete pass, then an incomplete pass, then two complete passes, and finally an incomplete pass. Note that I am using single brackets here; later on we'll have to consider another type of sequence, for which I will use double brackets.

What is the probability of this sequence? Since we are assuming that each pass is independent of the others, we can decompose the probability of this sequence into the product of the individual pass probabilities:

$$\begin{aligned} \Pr([C, I, C, C, I]) &= \Pr(C) \cdot \Pr(I) \cdot \Pr(C) \cdot \Pr(C) \cdot \Pr(I) \\ &= (0.60) \times (0.40) \times (0.60) \times (0.60) \times (0.40) \\ &= 0.01944 \end{aligned}$$

So far, so good. But note that this is not the only sequence of 5 passes with 3 completions. For instance, the sequence $[I, C, C, I, C]$ is also a sequence of 5 passes with 3 completions. The probability of this sequence is:

$$\begin{aligned} \Pr([I, C, C, I, C]) &= \Pr(I) \cdot \Pr(C) \cdot \Pr(C) \cdot \Pr(I) \cdot \Pr(C) \\ &= (0.40) \times (0.60) \times (0.60) \times (0.40) \times (0.60) \\ &= 0.01944 \end{aligned}$$

So we have the same probability as before. But wait – there's more! How about the sequence $[C, C, I, C, I]$? Now we have:

$$\begin{aligned} \Pr([C, C, I, C, I]) &= \Pr(C) \cdot \Pr(C) \cdot \Pr(I) \cdot \Pr(C) \cdot \Pr(I) \\ &= (0.60) \times (0.60) \times (0.40) \times (0.60) \times (0.40) \\ &= 0.01944 \end{aligned}$$

Once again, we get the same probability. This shouldn't be surprising: if we have 5 passes, and 3 of them are completions, then we will have 3 factors of

$\Pr(C) = 0.6$ and 2 factors of $\Pr(I) = 0.4$, and since the order in which we multiply these factors doesn't influence the final result, we will always end up with the same result.

There seem to be many different sequences of 5 passes with 3 completions. How many? In this example, the numbers are small enough that it is possible to explicitly list them all:

$$\begin{array}{ll} [C, C, C, I, I] & [C, C, I, C, I] \\ [C, C, I, I, C] & [C, I, C, C, I] \\ [C, I, C, I, C] & [C, I, I, C, C] \\ [I, C, C, C, I] & [I, C, C, I, C] \\ [I, C, I, C, C] & [I, I, C, C, C] \end{array}$$

So there are 10 possible such sequences. We know that each individual sequence has probability 0.01944. Thus, the probability of 3 completions out of 5 pass attempts is

$$\Pr(X = 3) = 10 \times 0.01944 = 0.1944$$

And thus we have the answer to our question.

We've solved our original problem, but there are still many other problems. For instance, what is the probability that Tom Gravy attempts 11 passes, and 8 of them are completions? And how can we adapt these calculations for another quarterback who has a different completion probability? Really, what we want is a general theory that will show us how to calculate the probability given three input values:

- The number of pass attempts, denoted n.
- The number of pass completions, denoted k.
- The probability of an individual pass completion, denoted p.

Using this notation, we found that for $n = 5$, $k = 3$, and $p = 0.6$, the probability was 0.1944. But ultimately we want to be able to do this calculation for any values of n, k, and p.

Despite the fact that we were working with a very specific example, we should be able to draw 2 conclusions at this point:

- First, every sequence of n passes, with k completions, will result in the same probability (remember, this depends on our assumptions of independence and identical distribution).
- Second, there are potentially a lot of these sequences, and it would be good if we could come up with some way to count them systematically.

So we need to a way to systematically count all the sequences of 5 passes with 3 completions.

3.2 Binomial Coefficients

Let's return to the problem of counting the number of sequences of 5 passes with 3 completions. We can think of this as starting with an empty sequence $[-,-,-,-,-]$ with 5 locations, and selecting three of these locations in which to write a C. This suggests a more general approach: our goal should be to develop a method for counting the number of ways of choosing k objects from a total of n objects. This number is called a *binomial coefficient*, and is denoted:

$$\binom{n}{k} = \left(\begin{array}{l} \text{Number of ways to choose } k \text{ objects} \\ \text{from a total of } n \text{ objects} \end{array} \right)$$

The symbol on the left is pronounced "n choose k", and in fact in LaTeXis typeset as `{n \choose k}`.

In order to develop a formula for the binomial coefficient, we need to engage in some careful counting arguments, and these can sometimes be a little tricky. It might take a little while to fully digest them, so if you don't fully understand everything the first time you see it, don't get upset!

Counting permutations

Let's start out with a fairly simple counting argument. We have a collection of objects, say $\{1, 2, 3\}$, and we want to count all the different ways that we can order these objects. That is, we want to count all the possible *ordered sequences* of the objects. For instance, we could put the three numbers in order so that 2 is the first number, 3 is the second number, and 1 is the third number. I'll denote this ordered sequence using double brackets, so we can

express this particular ordering as $[[2, 3, 1]]$. It's easy to list all the possible orderings of the three objects, and it turns out that there are 6 of them:

$$\begin{array}{ll} [[1, 2, 3]] & [[1, 3, 2]] \\ [[2, 1, 3]] & [[2, 3, 1]] \\ [[3, 1, 2]] & [[3, 2, 1]] \end{array}$$

How about if we have 4 objects, say $\{1, 2, 3, 4\}$. Then there are 24 such ordered sequences:

$$\begin{array}{lll} [[1, 2, 3, 4]] & [[1, 2, 4, 3]] & [[1, 3, 2, 4]] \\ [[1, 3, 4, 2]] & [[1, 4, 2, 3]] & [[1, 4, 3, 2]] \\ [[2, 1, 3, 4]] & [[2, 1, 4, 3]] & [[2, 3, 1, 4]] \\ [[2, 3, 4, 1]] & [[2, 4, 1, 3]] & [[2, 4, 3, 1]] \\ [[3, 1, 2, 4]] & [[3, 1, 4, 2]] & [[3, 2, 1, 4]] \\ [[3, 2, 4, 1]] & [[3, 4, 1, 2]] & [[3, 4, 2, 1]] \\ [[4, 1, 2, 3]] & [[4, 1, 3, 2]] & [[4, 2, 1, 3]] \\ [[4, 2, 3, 1]] & [[4, 3, 1, 2]] & [[4, 3, 2, 1]] \end{array}$$

Let's think about how we can count these ordered sequences. Consider the ordered sequences of 3 objects: how could we systematically generate these? Suppose we start out with an empty sequence $[[-, -, -]]$, and we choose a position for the object 1. There are three possible choices:

$$[[1, -, -]] \qquad [[-, 1, -]] \qquad [[-, -, 1]]$$

Then, for each of these, there are two choices for the position of the object 2. For instance, if we just focus on the sequence $[[-, 1, -]]$, then the two possibilities are:

$$[[2, 1, -]] \qquad [[-, 1, 2]]$$

For the object 3, there is now only 1 open position remaining, so we can fill this automatically, and we have:

$$[[2, 1, 3]] \qquad [[3, 1, 2]]$$

Thus, there were 3 choices for the position of the first object, and for each of these choices there were 2 choices for the position of the second object, and for each of these there was only one choice for the position of the third object. So the total number of possible sequences is $3 \times 2 \times 1 = 6$.

How about if we have 4 objects? Then there are 4 possible positions for the first object:

$$[[1,-,-,-]] \qquad [[-,1,-,-]]$$
$$[[-,-,1,-]] \qquad [[-,-,-,1]]$$

For each of these, there are three possible positions for the second object. For instance, if we focus on the sequence $[[-,1,-,-]]$, then the three possibilities are:

$$[[2,1,-,-]] \qquad [[-,1,2,-]] \qquad [[-,1,-,2]]$$

For each of these, there are 2 positions for the third object. If we think about the sequence $[[-,1,2,-]]$, then we have:

$$[[3,1,2,-]] \qquad [[-,1,2,3]]$$

At this point, there is only one object left, and for each sequence there is only one empty position left, so the position of the fourth object is completely determined, and we can just fill in the remaining location:

$$[[3,1,2,4]] \qquad [[4,1,2,3]]$$

Thus, there were 4 choices for the position of the first object, and for each of these choices there were 3 choices for the position of the second object, and for each of these there were 2 choices for the third object, and finally there was only one choice for the position of the forth object. So the total number of possible sequences is $4 \times 3 \times 2 \times 1 = 24$, which is the same number that we obtained by explicitly tabulating all possibilities.

Now let's generalize this to k objects. There will be k positions available for the first object, and for each of these choices there will be $k-1$ choices for the second object, and for each of these sequences there will be $k-2$ choices for the third object, and so on. Thus, the total number of sequences will be:

$$k \times (k-1) \times (k-2) \times \ldots \times 3 \times 2 \times 1 = k!$$

When we take k objects and list them in a specific order, this is called a *permutation* of the objects. Thus, $[[3,1,2]]$ and $[[2,3,1]]$ are permutations of the set of objects $\{1,2,3\}$. Notice here that order matters: the permutations $[[3,1,2]]$ and $[[2,3,1]]$ are considered distinct because the objects are listed

in different orders. So, we can summarize the results of this subsection by saying:

$$k! = \left(\begin{array}{c} \text{Total number of permutations} \\ \text{of } k \text{ objects} \end{array} \right)$$

When we were thinking about the sequence of Tom Gravy's passes, I mentioned that a specific sequence of passes was denoted with single brackets like this: $[C, I, C, C, I]$, but that later on we would encounter another set of sequences that would use double brackets. These are the permutations that we've been working with in this section. There are important differences between permutations and Tom Gravy passing sequences:

- The permutations consist of re-orderings of k objects, hence they must have length k and every item in the list must be distinct.
- The Tom Gravy passing sequences consist of C and I symbols, and these can be repeated. So there is no restriction on the length of a Tom Gravy passing sequence; in our example we are focused on sequences of length 5, but we could have chosen sequences of length 3 or 7 or 22. Also, the items in the list will not necessarily be distinct.

Counting passing sequences

Let's go back to Tom Gravy, and recall that for 5 passes with 3 completions there were 10 possible sequences:

$$\begin{array}{ll} [C, C, C, I, I] & [C, C, I, C, I] \\ [C, C, I, I, C] & [C, I, C, C, I] \\ [C, I, C, I, C] & [C, I, I, C, C] \\ [I, C, C, C, I] & [I, C, C, I, C] \\ [I, C, I, C, C] & [I, I, C, C, C] \end{array}$$

Notice that there are 5 slots, and we can choose 3 of them to hold a C, so this enumeration is the same as asking for how many ways there are to choose 3 items from a total set of 5 items.

We can think about this systematically. We can think of an empty sequence $[-, -, -, -, -]$ by keeping track of all the choices that we can make

as we insert three instances of C into this sequence. When we put the first C into the empty sequence, there are 5 possible choices:

$$[C, -, -, -, -]$$
$$[-, C, -, -, -]$$
$$[-, -, C, -, -]$$
$$[-, -, -, C, -]$$
$$[-, -, -, -, C]$$

Now when we insert the second C, for each of these sequences there will be 4 available positions. For instance, let's consider the sequence $[-, -, C, -, -]$:

$$[C, -, C, -, -]$$
$$[-, C, C, -, -]$$
$$[-, -, C, C, -]$$
$$[-, -, C, -, C]$$

It's similar for the third C: for each of these sequences, there will be three available positions. Let's consider the sequence $[-, -, C, C, -]$

$$[C, -, C, C, -]$$
$$[-, C, C, C, -]$$
$$[-, -, C, C, C]$$

Of course, once we've selected the position for the third C, we're done, because we're only concerned with the situation where Tom Gravy throws 3 completions out of 5 passes. So we can fill in the remaining empty slots with I, giving us the final sequences:

$$[C, I, C, C, I]$$
$$[I, C, C, C, I]$$
$$[I, I, C, C, C]$$

How many ways are there to populate the empty sequence $[-, -, -, -, -]$? There were 5 choices for the first C, and for each of the resulting sequences there were 4 choices for the second C, and for each of these sequences there were 3 remaining choices for the last C. So this comes out to $5 \times 4 \times 3 = 60$.

At first glance, this seems wrong. After all, we explicitly enumerated all the possible sequences with 3 completions out of 5 passes, and we found

that there were only 10 of them. But now it seems that there are 60 ways to populate the empty sequence $[-,-,-,-,-]$ with 3 copies of C. What's going on here?

The resolution of this seeming paradox is that the value of 60 is keeping track of all the possible different sequences of choices that we can populate the empty sequence. But many of these will result in the same final passing sequence. For instance, consider the sequence $[C,-,C,C,-]$:

- We could first select the first position, resulting in $[C,-,-,-,-]$. Then we could select the the third position for the second C, resulting in $[C,-,C,-,-]$. Finally, we could select the fourth position for the third C, giving $[C,-,C,C,-]$. Thus we would fill out the empty sequence like this:

$$\begin{aligned}[-,-,-,-,-] &\rightarrow [C,-,-,-,-] \\ &\rightarrow [C,-,C,-,-] \\ &\rightarrow [C,-,C,C,-]\end{aligned}$$

- On the other hand, for the first C we could have selected the fourth position, resulting in $[-,-,-,C,-]$. For the second C, we could have selected the first position, giving $[C,-,-,C,-]$. Then for the final C we could have selected the third position, resulting in $[C,-,C,C,-]$. Now we would have this process:

$$\begin{aligned}[-,-,-,-,-] &\rightarrow [-,-,-,C,-] \\ &\rightarrow [C,-,-,C,-] \\ &\rightarrow [C,-,C,C,-]\end{aligned}$$

And there are other possible processes as well. So there can be multiple sequences of *choices* for filling in the slots that give us the same final sequence of passes $[C,-,C,C,-]$.

How many such sequences of choices? Let's use the notation $[[1,3,4]]$ to denote the process of choosing the first position first, the third position next, and the fourth position last. Just to be clear: the sequence $[C,-,C,C,-]$, with single brackets, represents a sequence of Tom Gravy passes, while the sequence $[[1,3,4]]$, with double brackets, represents a sequence of choices that

we make to populate the empty sequence $[-,-,-,-,-]$. Then there are six possible sequences of choices that will result in the sequence $[C,-,C,C,-]$:

$$\begin{array}{ll} [[1,3,4]] & [[1,4,3]] \\ [[3,1,4]] & [[3,4,1]] \\ [[4,1,3]] & [[4,3,1]] \end{array}$$

Do you see what this is? It's just the set of all possible permutations of the set of objects $\{1,3,4\}$, and we know from the previous subsection that the total number of such permutations must be $3! = 6$. In fact, this is true for *every* sequence of Tom Gravy passes. For instance, consider the sequence $[I,C,C,I,C]$. There are 6 different sequences of choices that we could have made that will give us this sequence of passes:

$$\begin{array}{ll} [[2,3,5]] & [[2,5,3]] \\ [[3,2,5]] & [[3,5,2]] \\ [[5,2,3]] & [[5,3,2]] \end{array}$$

Now we have everything we need to count Tom Gravy passing sequences. First, count the number of sequences of *choices* for filling out the five positions with 3 Cs. Then, divide by the number of permutations of 3 objects, because we know that any such permutation will give us the same final passing sequence. We know that there are 60 possible sequences of choices that will place 3 copies of C into a sequence with 5 positions, and there are 6 permutations for every set of three indices, so we have

$$\text{Number of passing sequences} = \frac{60}{6} = 10$$

And that's what we found originally when we enumerated all the possible sequences of 5 passes with 3 completions. In words, we can write this calculation as:

$$\text{Number of passing sequences} = \frac{\text{Number of sequences of choices}}{\text{Number of permutations}}$$

There's a nice way to think about this. We can think of a Tom Gravy passing sequence such as $[C,I,C,C,I]$ as the set $\{1,3,4\}$, where the numbers in the set represent the positions in the sequence containing a C. But we

don't care about the order of these numbers, because the set $\{4, 1, 3\}$ would give us the exact same passing sequence $[C, I, C, C, I]$. So what we have is a set of 5 positions, and we want to count all the *unordered* subsets of size 3. In order to do this, we first count all the *ordered* subsets of size 3, because that's easy to do, and then we divide by all the permutations.

Our real goal is to develop a general formula, beyond the immediate example of 5 passes with three completions. So let's now ask: how many passing sequences of length n with k completions are there? We start with an empty sequence of n positions, and we have to populate k positions with a C. The number of choices for the first C is n, the number of choices for the second C is $n-1$, and so on, down to $n-k+1$ choices for the position of the kth C. Thus, the number of sequences of choices for selecting k positions in the empty sequence of n positions in which to place a C is:

$$\underbrace{n \times (n-1) \times \ldots \times (n-k+1)}_{k \text{ choices}}$$

Now we need to divide by the number of permutations of k objects. But we know that that will just be $k!$. So we have the formula

$$\binom{n}{k} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!}$$

If you try this out with our example of $n = 5$ and $k = 3$, you should obtain

$$\frac{5 \times 4 \times 3}{3!} = \frac{60}{6} = 10$$

Many people employ a slightly different notation for binomial coefficients. Note that

$$n! = n \cdot (n-1) \cdot \ldots \cdot (n-k+1) \cdot (n-k)!$$

Then we have:

$$\begin{aligned} \binom{n}{k} &= \frac{n \cdot (n-1) \cdot \ldots \cdot (n-k+1)}{k!} \cdot \frac{(n-k)!}{(n-k)!} \\ &= \frac{n!}{k! \cdot (n-k)!} \end{aligned}$$

This alternative version is the more common form of the binomial coefficient, because it's a little more compact than the version we derived, and tends to be a little more friendly to algebraic manipulation. Unfortunately, it obscures the logic behind the derivation of the formula, and it tends to be computationally inefficient, because you end up computing the $(n-k)!$ factor twice, once in the numerator and once in the denominator, and these ultimately cancel out. Of course both forms are equivalent, but often one form will be more suitable for a given purpose.

Binomial coefficients are everywhere

One reason why I've spent so much time on binomial coefficients is because they arise in many differrent situations beyond statistical inference, and it's useful to understand them. For instance, suppose we want to sum the numbers from 1 to n: is there a nice formula for that? It turns out that there is just such a formula, and it involves a binomial coefficient:

$$1 + 2 + \ldots + n = \binom{n+1}{2}$$

There's a clever argument for proving this identity. The right-hand side is counting the number of pairs of objects that can be drawn from a collection of $n+1$ objects. The left-side is also counting all these pairs, but in a very specific way. Suppose we label the objects $1, 2, \ldots, n, n+1$. Then there are n pairs where the lowest element has the label 1: $\{1,2\}, \{1,3\}, \ldots, \{1,n\}, \{1,n+1\}$. Next, there will be $n-1$ pairs where the lowest element has the label 2: $\{2,3\}, \{2,4\}, \ldots, \{2,n\}, \{2,n+1\}$. We continue this way until we get to the element with label n, and there will only be 1 such pair: $\{n, n+1\}$. At this point, we've enumerated all the pairs, and if we add them all up we get $n + (n-1) + \ldots + 3 + 2 + 1$. So both sides of the equation are counting the same thing, the collection of pairs of $n+1$ objects, and thus they must be equal. (You can also prove the identity using induction, but for my taste that's not as pretty as this counting argument.)

The moral of the story is that binomial coefficients are everywhere, and it's good to understand them.

3.3 Moments of the Binomial Distribution

Let's calculate the expected value of a binomial distribution. Let X denote a binomial random variable with parameters n and p. Then we have:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{x \in \Omega_X} x \cdot \Pr(X = x) \\
&= \sum_{x=0}^{n} x \cdot \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x}
\end{aligned}$$

Notice what happened in the last line: when x is zero, the entire expression in the summation is 0, so really the sum starts at $x = 1$. Then, because x starts at 1, it will never take on the value 0, so we are justified in cancelling an x in the denominator of the binomial coefficient. (Notice here that we are using the factorial version of the binomial coefficient, because this makes it easier to perform algebraic manipulations). Now let's pull out the n from the numerator of the binomial coefficient, and a p from the second factor (note how we are again justified in doing this because x starts at 1 and not 0):

$$\sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} = np \cdot \sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}$$

Now we will re-index the sum; this is analogous to performing a substitution in an integral. Let's use s as our new index, and we will set $s = x - 1$, so that $x = s + 1$. When we sum over x values, the sum ranges from $x = 1$ to $x = n$, but with our re-indexing the sum will now range from $s = 0$ to

$s = n - 1$. The binomial coefficient now becomes:

$$\begin{aligned}
\frac{(n-1)!}{(x-1)! \cdot (n-x)!} &\rightarrow \frac{(n-1)!}{s! \cdot (n-(s+1))!} \\
&= \frac{(n-1)!}{s! \cdot ((n-1)-s)!} \\
&= \binom{n-1}{s}
\end{aligned}$$

For the factors with p, we have:

$$\begin{aligned}
p^{x-1} \cdot (1-p)^{n-x} &\rightarrow p^s \cdot (1-p)^{n-(s+1)} \\
&= p^s \cdot (1-p)^{(n-1)-s}
\end{aligned}$$

Let's put all of this together:

$$\sum_{x=1}^{n} \frac{(n-1)!}{(x-1)!(n-x)!} p^{x-1}(1-p)^{n-x} \rightarrow \sum_{s=0}^{n-1} \frac{(n-1)!}{s! \cdot ((n-1)-s)!} p^s (1-p)^{(n-1)-s}$$

Now do you see what the right-hand side is? The algebraic expression in the sum is just the probability mass function for a binomial distribution with $n-1$ trials and s successes, and since we are summing from 0 to $n-1$ this must be 1. So:

$$\left(\sum_{s=0}^{n-1} \frac{(n-1)!}{s! \cdot ((n-1)-s)!} p^s (1-p)^{(n-1)-s} \right) = 1$$

Amazing! Now let's do the whole derivation:

$$\begin{aligned}
\mathrm{E}[X] &= \sum_{x=0}^{n} x \cdot \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} \frac{n!}{(x-1)!(n-x)!} p^x (1-p)^{n-x} \\
&= np \cdot \sum_{s=0}^{n-1} \frac{(n-1)!}{s! \cdot ((n-1)-s)!} p^s (1-p)^{(n-1)-s} \\
&= np \cdot 1 \\
&= np
\end{aligned}$$

So if X is a binomial random variable with parameters n and p, the expected value (or first moment) of X is np.

To calculate the second moment, there is a little algebraic trick that will make life much easier: instead of directly calculating the second moment $\mathrm{E}[X^2]$, we instead calculate the expression $\mathrm{E}[X \cdot (X-1)]$. Do you see why? It's because the expression $x \cdot (x-1)$ will cancel nicely with the $x!$ term in the denominator of the binomial coefficient. Once again, it's nice to work with the factorial form of the binomial coefficient when doing algebraic manipulations, and this is why it's the form that you see most often in textbooks. This time, we'll make the substitution $s = x - 2$, hence $x = s + 2$, but otherwise

everything is the same as before:

$$\begin{aligned}
\mathrm{E}[X \cdot (X-1)] &= \sum_{x \in \Omega_X} x \cdot (x-1) \cdot \Pr(X = x) \\
&= \sum_{x=0}^{n} x \cdot (x-1) \cdot \frac{n!}{x! \cdot (n-x)!} \cdot p^x (1-p)^{n-x} \\
&= \sum_{x=2}^{n} \frac{n!}{(x-2)! \cdot (n-x)!} \cdot p^x (1-p)^{n-x} \\
&= n(n-1)p^2 \cdot \sum_{x=2}^{n} \frac{(n-2)!}{(x-2)! \cdot (n-x)!} \cdot p^{x-2} (1-p)^{n-x} \\
&= n(n-1)p^2 \cdot \sum_{s=0}^{n-2} \frac{(n-2)!}{s! \cdot ((n-2)-s)!} \cdot p^s (1-p)^{(n-2)-s} \\
&= n(n-1)p^2 \cdot 1 \\
&= n^2 p^2 - np^2
\end{aligned}$$

At this point, we've calculated $\mathrm{E}[X \cdot (X-1)]$, but what we really want is the second moment $\mathrm{E}[X^2]$. Here we just need to use the linearity of the expectation operator:

$$\begin{aligned}
\mathrm{E}[X^2] &= \mathrm{E}[X^2] - \mathrm{E}[X] + \mathrm{E}[X] \\
&= \mathrm{E}[X^2 - X] + \mathrm{E}[X] \\
&= \mathrm{E}[X \cdot (X-1)] + \mathrm{E}[X] \\
&= (n^2 p^2 - np^2) + np \\
&= n^2 p^2 + np - np^2
\end{aligned}$$

Thus, the variance is:

$$\begin{aligned}
\text{Var}[X] &= \text{E}[X^2] - (\text{E}[X])^2 \\
&= (n^2p^2 + np - np^2) - (np)^2 \\
&= np - np^2 \\
&= np(1-p)
\end{aligned}$$

3.4 What About the Cumulative Probability Function?

We've derived the probability mass function for the binomial distribution, along with the mean and variance. Now surely it's time for the cumulative probability function, right? Unfortunately, there is no good answer here, and no simple closed form expression exists for the cumulative probability function of the binomial distribution. Sure, we could just go back to the definition:

$$\begin{aligned}
F_X(x) &= \sum_{k=0}^{x} \text{Pr}(X = k) \\
&= \sum_{k=0}^{x} \binom{n}{k} \cdot p^k (1-p)^{n-k}
\end{aligned}$$

Of course this is true, but it's also not much help, because what it's really saying is, "To calculate the cumulative probability function for the value x, calculate every individual probability up to and including x." So this is not saving us any work. What we really want is some sort of clever shortcut that enables us to avoid doing this brute-force calculation, and the answer is that there is no clever shortcut (at least for an exact answer). On the other hand, in practice this isn't really a problem, because we can use software to obtain this value, and both R and Excel have built-in functions for this calculation.

3.5 Section Summary

Here's what we've developed in this section. Let X be a binomial random variable with parameters n and p. Then:

$$\begin{aligned} \Pr(X = k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ \mathrm{E}[X] &= np \\ \mathrm{Var}[X] &= np(1-p) \end{aligned}$$

4 Transformations of a Random Variable

Before we encounter the normal and chi-squared distributions, we're going to pause for a moment and develop some more probability theory. This is generally not taught in beginning probability and statistics classes, but it will be very important in developing the theory that we need in order to do statistics.

Suppose we have a continuous random variable X, and we want to apply some sort of function to it, which we will call $g(x)$. If we know the density function for X, what can we say about the density function for the new random variable $Y = g(X)$? For instance, suppose X is an exponential random variable with parameter $\lambda = 1$, so that the density function for X is:

$$f_X(x) = e^{-x}, \quad x > 0$$

Now suppose that $g(x) = x^2$, so that $Y = X^2$. What is the density function $f_Y(y)$? It's tempting to make some intuitive guesses; for instance, maybe we should just square the input value, so that we end up with this:

$$f_X^*(y) = e^{-y^2}. \quad y > 0$$

But this is wrong, because this function won't integrate to 1:

$$\int_0^\infty f_X^*(y) \cdot dy \;=\; \int_0^\infty e^{-y^2} \cdot dy \;=\; \frac{\sqrt{\pi}}{2}$$

(By the way, I'm using the asterisk notation to indicate that this isn't really a density function). OK, that won't work. Again, since $g(x)$ is the square function, maybe we should just square the density:

$$f_X^{**}(x) = \left(e^{-x}\right)^2 = e^{-2x}$$

Again, this won't integrate to 1:

$$\int_0^\infty f_X^{**}(x) \cdot dx \;=\; \int_0^\infty e^{-2x} \cdot dx \;=\; \frac{1}{2}$$

Hmmm . . . perhaps we should stop guessing and approach this in a systematic manner.

To start with, we should note a very basic and important property of the cumulative probability function. Here's the definition of the cumulative probability function for a continuous random variable:

$$F_X(x) = \int_{-\infty}^{x} f_X(s) \cdot ds$$

Now, what is the derivative of $F_X(x)$ with respect to x? By the Fundamental Theorem of Calculus, we have:

$$\begin{aligned} \frac{dF_X(x)}{dx} &= \frac{d}{dx}\int_{-\infty}^{x} f_X(s) \cdot ds \\ &= f_X(x) \end{aligned}$$

In other words, the derivative of the cumulative probability function $F_X(x)$ at the point x is just the density function $f_X(x)$ at that point.

This suggests a strategy for finding the density function $f_Y(y)$ of the transformed random variable $Y = X^2$: if we could somehow obtain an expression for the cumulative probability function of Y, then we could differentiate this to obtain the density function. However, the only cumulative probability function that we know about is $F_X(x)$, the cumulative probability function for X. Thus, we have to somehow express $F_Y(y)$ in terms of $F_X(x)$. How

can we do this? Here's an example for $F_Y(9)$:

$$\begin{aligned} F_Y(9) &= \Pr(Y \leq 9) \\ &= \Pr(X^2 \leq 9) \\ &= \Pr(X \leq 3) \\ &= F_X(3) \end{aligned}$$

Let's stop and think about this for a moment. We said that $Y = X^2$. So for instance what is $F_Y(9)$, the probability that Y is less than or equal to 9? In order for this to occur, it must be the case that X is less than or equal to 3, and this is just $F_X(3)$. More generally, we have:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(X^2 \leq y) \\ &= \Pr(X \leq \sqrt{y}) \\ &= F_X(\sqrt{y}) \end{aligned}$$

So far, so good – we've found a way to express $F_Y(y)$, the cumulative probability function for Y, in terms of $F_X(x)$, the cumulative probability function for X.

Our next step to find the density function $f_Y(y)$ is to take the derivative of $F_Y(y)$:

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d}{dy}\left[F_X(\sqrt{y})\right]$$

Notice the expression in the second factor on the right-hand side: $F_X(\sqrt{y})$. This is a composite function, because we first apply the square-root function to y, and then we apply the function $F_X(x)$ to the output $\sqrt{y}$. How do we take the derivative of the composite function $F_X(\sqrt{y})$? We use the chain rule

from elementary calculus:

$$\frac{d}{dy}\left[F_X(\sqrt{y})\right] \;=\; \left.\frac{dF_X(x)}{dx}\right|_{x=\sqrt{y}} \cdot \frac{d}{dy}\left[\sqrt{y}\right]$$

Notice that, by what we saw a few minutes ago,

$$\begin{aligned}\left.\frac{dF_X(x)}{dx}\right|_{x=\sqrt{y}} &= f_X(\sqrt{y}) \\ &= e^{-\sqrt{y}}\end{aligned}$$

And by standard elementary calculus, we have:

$$\frac{d}{dy}\left[\sqrt{y}\right] \;=\; \frac{1}{2\cdot\sqrt{y}}$$

Putting this all together, we have:

$$\begin{aligned}\frac{d}{dy}\left[F_X(\sqrt{y})\right] &= \left.\frac{dF_X(x)}{dx}\right|_{x=\sqrt{y}} \cdot \frac{d}{dy}\left[\sqrt{y}\right] \\ &= \left(e^{-\sqrt{y}}\right)\cdot\left(\frac{1}{2\cdot\sqrt{y}}\right) \\ &= \frac{e^{-\sqrt{y}}}{2\cdot\sqrt{y}}\end{aligned}$$

That was a concrete example – what's the general formulation of this method? We start with a random variable X and a function $g(x)$, and we want to find the density function of the transformed random variable $Y = g(X)$. We have to do two things:

- First, we have to express $F_Y(y)$ in terms of $F_X(x)$, the cumulative probability function for X.
- Next, we take the derivative with respect to y of $F_Y(y)$ to obtain the density function $f_Y(y)$; this will involve the chain rule.

For the first step, we have:

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(g(X) \leq y) \\ &= \Pr(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)) \end{aligned}$$

For the second step, we take the derivative with respect to y, for which we will have to use the chain rule:

$$\begin{aligned} \frac{dF_Y(y)}{dy} &= \frac{dF_X(g^{-1}(y))}{dy} \\ &= \left.\frac{dF_X(x)}{dx}\right|_{x=g^{-1}(y)} \cdot \frac{d}{dy}\left[g^{-1}(y)\right] \\ &= f_X(g^{-1}(y)) \cdot \frac{d}{dy}\left[g^{-1}(y)\right] \end{aligned}$$

There was a subtle technical issue in this derivation that I hid from you: $g(x)$ can't be an arbitrary function, because it has to have an inverse so that can write an expression such as $g^{-1}(y)$, and it has to preserve order, so that we can go from the statement $g(X) \leq y$ to the statement $X \leq g^{-1}(y)$. This means that $g(x)$ is strictly increasing i.e. if $a < b$, then $g(a) < g(b)$. If $g(x)$ is strictly decreasing (if $a < b$, then $g(a) > g(b)$), then the formula becomes:

$$\frac{dF_Y(y)}{dy} = -f_X(g^{-1}(y)) \cdot \frac{d}{dy}\left[g^{-1}(y)\right]$$

You might be a little nervous about that minus sign out in front; after all, we're trying to compute a density, and those are always positive. But if $g(x)$ is strictly decreasing, then the derivative will be negative, and this will cancel out the minus sign. Thus, there are two formulas for the density of the transformed random variable, depending on whether $g(x)$ is strictly increasing or strictly decreasing:

- If $g(x)$ is strictly increasing, so that if $a < b$, then $g(a) < g(b)$, then the density function for the transformed random variable $Y = g(X)$ is:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \frac{d}{dy}\left[g^{-1}(y)\right]$$

- If $g(x)$ is strictly decreasing, so that if $a < b$, then $g(a) > g(b)$, then the density function for the transformed random variable $Y = g(X)$ is:

$$f_Y(y) = -f_X(g^{-1}(y)) \cdot \frac{d}{dy}\left[g^{-1}(y)\right]$$

Actually, we can combine these formulas into one expression, by using an absolute value:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{d}{dy}\left[g^{-1}(y)\right]\right|$$

At first, this method might seem tricky and complicated. But with a little practice, you should get the comfortable with it. We'll see two applications of this method in the rest of today's lecture, at which point I know you will be eager to try your hand at this, so you'll get to do a homework problem using this approach. As always, try to understand the concrete example first, and then use that as a guide for the formal theory.

5 The Standard Normal Random Variable

Now we will discuss the normal distribution, which is unquestionably the most important distribution in all of statistics and probability. I could just give you the density function for the general case, but I know you would find that too easy. Instead, we will first develop the concept of the standard normal random variable, and then apply our concept of a transformation of a random variable, which will give us some valuable geometric insight about this distribution.

A *standard normal* random variable Z has the density function:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{z^2}{2}\right\}$$

Notice that I used the letter Z to denote the random variable, and this is the universal convention. In other words:

- If you have a standard normal random variable, and you don't use Z to denote it, that's weird.
- If you have normal random variable that doesn't have parameters $\mu = 0$ and $\sigma^2 = 1$ and use Z to denote it, that's weird.

Don't be weird.

What is the expected value of Z? We have to be a little careful here. By the definition of the expected value for a continuous random variable, we have:

$$\begin{aligned} \mathrm{E}[Z] &= \int_{\Omega_Z} z \cdot f_Z(z) \cdot dz \\ &= \int_{-\infty}^{+\infty} z \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \end{aligned}$$

Up to this point, everything is fine.

Now we need to evaluate that integral. There is an argument from "folk" or "street" calculus that is incorrect, but gives the right answer. This incorrect argument is based on the concept of an *odd* function: a function $g(z)$ is *odd* if $g(-z) = -g(z)$ for all z. If we then integrate this odd function over a symmetric range $(-a, a)$, it must be 0, because any contribution to the integral from the value $g(x)$ will be cancelled out by the negative contribution from $g(-x)$. Thus, we don't have to know anything about $g(x)$ or do any special tricks to see that:

$$\int_{-a}^{+a} g(x) \cdot dx = 0$$

All of this is true. But now the "street" calculus non-proof extends this argument to this integral:

$$\int_{-\infty}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz$$

The integrand here is odd, and we seem to be integrating over the symmetric range $(-\infty, +\infty)$. So, using this concept about integrating an odd function over a symmetric range, we must have:

$$\int_{-\infty}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \;=\; 0$$

What could be wrong about such an argument?

The problem with this "proof" is that we aren't really integrating over a symmetric interval of the form $(-a, +a)$. Instead, we are attempting to evaluate what's called a double improper Riemann integral, that is, the range of integration is not between two numbers, but rather is the result of a limiting process:

$$\int_{-\infty}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \;=\; \lim_{\substack{b \to \infty \\ a \to -\infty}} \int_a^b z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz$$

The problem with integrals of this form is that they can be sensitive to the rates at which a and b go out to infinity. According to the theory of Riemann integration, in order to evaluate this integral we must break it up into the two tails and evaulate each integral individually:

$$\int_{-\infty}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \;=\; \int_{-\infty}^{0} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \;+\; \int_{0}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz$$

The two integrals on the right-hand side are easy to evaluate, although they are also improper Riemann integrals, so we will be very precise:

$$\begin{aligned}
\int_0^{\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz &= \lim_{b \to \infty} \int_0^b z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \\
&= \lim_{b \to \infty} -\exp\left\{-\frac{z^2}{2}\right\} \Bigg|_0^b \\
&= \lim_{b \to \infty} \left[1 - \exp\left\{-\frac{b^2}{2}\right\}\right] \\
&= 1
\end{aligned}$$

By the same argument, we have:

$$\begin{aligned}\int_{-\infty}^{0} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz &= \lim_{a \to \infty} \int_{a}^{0} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \\ &= -1\end{aligned}$$

Then:

$$\begin{aligned}\int_{-\infty}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz &= \int_{-\infty}^{0} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz + \int_{0}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \\ &= 1 - 1 \\ &= 0\end{aligned}$$

And finally, after all this, we have:

$$\begin{aligned}\mathrm{E}[Z] &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} z \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \\ &= \frac{1}{\sqrt{2\pi}} \cdot 0 \\ &= 0\end{aligned}$$

For the variance of Z, we need to calculate the second moment – in fact, because the expected value of Z is 0, the variance is exactly equal to the second moment. Using the definition of the second moment, we have:

$$\begin{aligned}\mathrm{E}[Z^2] &= \int_{\Omega_Z} z^2 \cdot f_Z(z) \cdot dz \\ &= \int_{-\infty}^{+\infty} z^2 \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz \\ &= \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} z^2 \cdot \exp\left\{-\frac{z^2}{2}\right\} \cdot dz\end{aligned}$$

I know it's been a while since you've had a chance to do any fun integrals, so I'm going to let you do this for one of your homework problems. It looks

scary, but in fact it's actually not all that bad, and with a nice substitution it becomes easy.

So let's summarize what we've done in this subsection. Let Z be a random variable with the density function:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{z^2}{2}\right\}$$

Then Z is called a *standard normal* random variable, and we have:

$$\begin{aligned} \mathrm{E}[Z] &= 0 \\ \\ \mathrm{Var}[Z] &= 1 \end{aligned}$$

One last point before we move on: what about the cumulative probability function, and the quantile function? The short answer is that there is no closed form algebraic formula for these function – we can't "solve" for them algebraically. (This is similar to the situation with the binomial distribution.) Instead, we have to use numerical software (i.e. Excel, R, Matlab, etc.) to obtain particular values. However, there are four values for the quantile function that are very nice to know:

q	$Q_Z(q)$
0.025	-1.96
0.05	-1.645
0.95	+1.645
0.975	+1.96

Remember how to interpret the quantile function: the first row of this table means that, for a standard normal random variable, we have:

$$F_Z(-1.96) = 0.025$$

That is, the probability that a standard normal random variable will have a realized value less than or equal to -1.96 is 2.5%.

The quantile values in the table are useful, but even more useful are two results about intervals. The first is:

$$\Pr(-1.645 \leq Z \leq 1.645) = 0.90$$

The second is:

$$\Pr(-1.96 \leq Z \leq 1.96) \;=\; 0.95$$

We will return to these two simple facts throughout this course.

6 The (General) Normal Distribution

Now we're going to apply some of our transformation theory. Let Z denote a standard normal random variable (as usual), let μ and σ be constants, with $\sigma > 0$, and define a new, transformed random variable X by:

$$X = \sigma \cdot Z + \mu$$

We can think of this as the function $g(Z) = \sigma Z + \mu$; notice that $g(z)$ is a strictly increasing function, because $\sigma > 0$. Then the inverse of this function is:

$$g^{-1}(x) = \frac{x - \mu}{\sigma}$$

Now recall our general formula for transforming a random variable:

$$f_X(x) = f_Z(g^{-1}(x)) \cdot \frac{d}{dx}\left[g^{-1}(x)\right]$$

Let's take each component on the right-hand side individually. For the first part, we know that

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{z^2}{2}\right\}$$

Then

$$\begin{aligned} f_Z(g^{-1}(x)) &= f_Z\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right\} \end{aligned}$$

For the second component, we have:

$$\begin{aligned} \frac{d}{dx}\left[g^{-1}(x)\right] &= \frac{d}{dx}\left[\frac{x-\mu}{\sigma}\right] \\ &= \frac{1}{\sigma} \end{aligned}$$

Putting this all together, we have:

$$\begin{aligned} f_X(x) &= f_Z(g^{-1}(x)) \cdot \frac{d}{dx}\left[g^{-1}(x)\right] \\ &= \left(\frac{1}{\sqrt{2\pi}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right\}\right) \cdot \left(\frac{1}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right\} \end{aligned}$$

Now we can state the formal definition. Let Z be a standard normal random variable, and let μ and $\sigma > 0$ be real constants. Then the random variable $X = \sigma Z + \mu$ is called a *general normal* random variable, and has the density function:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left\{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right\}$$

Here we go again: the expression "general normal random variable" is my own, and non-standard, so nobody outside of this course will know what you mean. I am introducing this terminology because I want to be able to distinguish normal distributions with any legal value for mu and σ^2 from the standard normal distribution.

6.1 Moments of the (General) Normal Distribution

Now that we have the density function for a (general) normal distribution, in principle it's straightforward to calculate the moments. According to our official definition, the first moment is:

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right\} \cdot dx$$

Hmmm ... this looks as though there might be some algebra involved. How about the second moment:

$$\mathrm{E}[X^2] = \int_{-\infty}^{+\infty} x^2 \cdot \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{1}{2} \cdot \frac{(x-\mu)^2}{\sigma^2}\right\} \cdot dx$$

OK, this is *definitely* going to involve a bunch of algebra. Is there any shortcut here?

Well, you will be delighted to learn that there *is* a shortcut, and a very nice one at that. Remember how we derived the (general) normal random variable: we transformed the standard normal random variable using the transformation $X = \sigma Z + \mu$. Also, recall that $\mathrm{E}[Z] = 0$. So, using the linearity of the expectation operator, we have:

$$\begin{aligned} \mathrm{E}[X] &= \mathrm{E}[\sigma Z + \mu] \\ &= \sigma \cdot \mathrm{E}[Z] + \mu \\ &= \sigma \cdot 0 + \mu \\ &= \mu \end{aligned}$$

So we didn't have to do that first big ugly integral after all. How about the variance? Here we use the standard properties of the variance:

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{Var}[\sigma \cdot Z + \mu] \\ &= \mathrm{Var}[\sigma \cdot Z] \\ &= \sigma^2 \cdot \mathrm{Var}[Z] \\ &= \sigma^2 \cdot 1 \\ &= \sigma^2 \end{aligned}$$

And so we didn't have to do the second big really ugly integral, either.

Here's what you CANNOT do. Don't be lazy and think, "Well, the expected value is always denoted by μ, and X has a parameter denoted by μ, so the expected value of X is μ." No no no! It's certainly true that we conventionally use μ to denote the expected value of a random variable. But this is not a proof, or even a very good argument! The point is that we first have to **show** that the expected value is equal to μ, and then we are justified in using the notation μ. Similarly for the variance – we have to **show** that the variance equals σ^2, and then that justifies us in using that notation.

6.2 The Cumulative Probability and Quantile Functions for the General Normal Distribution

What about the cumulative probability and the quantile function for the general normal distribution? In the case of the standard normal distribution,

I've said that there are no closed formulas for these functions, and so it must be the case that there are no closed form expressions for the general normal distribution as well. You might think there is nothing more to be said on this topic, but in fact there is one interesting insight to be had.

We will start with the cumulative probability function. Let X be a general normal random variable with parameters μ and σ^2; thus $X = \sigma Z + \mu$, where Z is a standard normal random variable. Then we have:

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) \\ &= \Pr(\sigma Z + \mu \leq x) \\ &= \Pr\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= F_Z\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

Notice what we've accomplished here: we started with a cumulative probability for the general normal random variable X, and we've shown that this is the same as a cumulative probability for the standard normal random variable Z. In other words, we've transformed a probability problem for a general normal random variable into a probability problem about the standard normal random variable. I like to say: secretly, every normal random variable X is really just the standard normal random variable Z. That means that whenever we make a probability statement about a general normal random variable, there is a corresponding statement about a standard normal random variable, and we can solve the former by solving the latter. In fact, this was traditionally the usual approach: given a probability statement about a particular general normal random variable, the problem would first be transformed into a question about the standard normal distribution, and then the answer could be looked up in a table of the cumulative probability function for this distribution. Thus, only one table was required, rather than a separate table for every general normal random variable.

There is a very nice interpretation for this expression:

$$\frac{x-\mu}{\sigma}$$

The denominator $x-\mu$ measures the distance of x from the population mean, and by dividing this distance by σ, we are effectively converting this distance into units of standard deviations. Thus, suppose X is a general normal distribution with parameters $\mu = 3.8$ and $\sigma^2 = 2.25$, and suppose $x = 6.5$. If the variance is $\sigma^2 = 2.25$, then the standard deviation is $\sigma = 1.5$, and

$$\begin{aligned} \frac{x-\mu}{\sigma} &= \frac{6.5-3.8}{1.5} \\ &= 1.8 \end{aligned}$$

So X is 1.8 standard deviations above the population mean.

Now let's consider the quantile function for a general normal distribution with parameters mu and σ^2: if $Q_X(q) = x$, then by the definition of the quantile function we must have:

$$F_X(x) = q$$

But this means that

$$F_Z\left(\frac{x-\mu}{\sigma}\right) = q$$

Thus, just as with the cumulative probability function, questions about the quantile function for the general normal distribution can be transformed into questions about the quantile function for the standard normal distribution. In particular, let's suppose that $q = 0.95$, so that we want to find the particular value of x so that:

$$F_X(x) = 0.95$$

In terms of the standard normal distribution, the corresponding equation is:

$$F_Z\left(\frac{x-\mu}{\sigma}\right) = 0.95$$

But we've also seen that

$$F_Z(1.645) = 0.95$$

Thus,

$$\frac{x-\mu}{\sigma} = 1.645$$

Solving for x, we obtain

$$x = \mu + 1.645\sigma$$

This is a very useful result: for **every** general normal random variable, the 95% quantile is $\mu + 1.645\sigma$. In fact, we can use this idea to derive a nice table:

q	$Q_x(q)$
0.025	$\mu - 1.96\sigma$
0.05	$\mu - 1.645\sigma$
0.95	$\mu + 1.645\sigma$
0.975	$\mu + 1.96\sigma$

A nice way to interpret this table is that for every general normal random variable, the 95th percentile is 1.645 standard deviations above the population mean.

6.3 Summary

Let's summarize the results of this section.

Let Z be a standard normal random variable, and let μ and $\sigma > 0$ be real constants. Then the random variable $X = \sigma \cdot Z + \mu$ is a general normal random variable, and has expected value, variance, and cumulative probabiity function:

$$\begin{aligned} \mathrm{E}[X] &= \mu \\ \mathrm{Var}[X] &= \sigma^2 \\ F_X(x) &= F_Z\left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

7 The Chi-Squared Distribution

The chi-squared distribution plays a critical role in statistical inference, and it's reasonable to say that only the normal distribution is more important. We'll start out by looking at a special case of this distribution, and then state the general form.

7.1 The Chi-Squared Distribution with One Degree of Freedom

In the previous section, we derived the general normal distribution by applying a linear transformation to a standard normal random variable:

$$X = \sigma \cdot Z + \mu$$

Now we'll apply another transformation the standard normal distribution, this time using the square function $g(z) = z^2$. Thus,

$$X = Z^2$$

Because we are working over all the real numbers, the square function does not have a unique inverse, and we have to careful. Let's start out the usual way:

$$\begin{aligned} F_X(x) &= \Pr(X \leq x) \\ &= \Pr(Z^2 \leq x) \\ &= \Pr(-\sqrt{x} \leq Z \leq \sqrt{x}) \end{aligned}$$

Do you see what happened? We had the event $Z^2 \leq x$, and now that means that Z must be less than $\sqrt{x}$ or greater than $-\sqrt{x}$, and so we end up with the event $-\sqrt{x} \leq Z \leq \sqrt{x}$. We could write this in terms of the cumulative probability function of Z:

$$\Pr(-\sqrt{x} \leq Z \leq \sqrt{x}) \;=\; F_Z(\sqrt{x}) - F_Z(-\sqrt{x})$$

Now we can take the derivative of this with respect to x to obtain the density function $f_X(x)$. Note that this will involve using the chain rule for both $F_Z(\sqrt{x})$ and $F_Z(-\sqrt{x})$, so that might be a little computationally laborious. But instead of this direct approach, we can exploit the symmetry of the standard normal distribution to get a simpler formula:

$$\begin{aligned} F_Z(\sqrt{x}) - F_Z(-\sqrt{x}) &= 2 \cdot \Pr(0 \leq Z \leq \sqrt{x}) \\ &= 2 \cdot (\Pr(Z \leq \sqrt{x}) - \Pr(Z \leq 0)) \\ &= 2 \cdot (\Pr(Z \leq \sqrt{x}) - 1/2) \\ &= 2F_Z(\sqrt{x}) - 1 \end{aligned}$$

Thus, we have shown:

$$F_X(x) = 2F_Z(\sqrt{x}) - 1$$

This derivation took a numer of steps, and you might not get the whole thing the first time you see it. However, each individual step is straightforward, so if you put in some time you should be able to grasp this.

At this point, we've achieved the first step in our process: we've re-expressed the probability statement of the cumulative distribution function for the transformed variable X in terms of the original variable Z. Next, we can differentiate this with respect to x in order to obtain the density function, and because we did that clever derivation we now only have to use the chain rule once:

$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) \\ &= \frac{d}{dx}(2 \cdot F_Z(\sqrt{x}) - 1) \\ &= 2 \cdot \left.\frac{dF(z)}{dz}\right|_{z=\sqrt{x}} \cdot \frac{d}{dx}(\sqrt{x}) \end{aligned}$$

Now the derivative of the first term in this expression is just the density function of Z, which was a standard normal random variable, evaulated at $z = \sqrt{x}$:

$$\begin{aligned} \left.\frac{dF(z)}{dz}\right|_{z=\sqrt{x}} &= \left.\frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}}\right|_{z=\sqrt{x}} \\ &= \frac{1}{\sqrt{2\pi}} \cdot e^{-x/2} \end{aligned}$$

The second term is straightforward:

$$\frac{d}{dx}(\sqrt{x}) = \frac{1}{2\sqrt{x}}$$

Putting this all together, we have:

$$\begin{aligned} f_X(x) &= 2 \cdot \frac{dF(z)}{dz}\bigg|_{z=\sqrt{x}} \cdot \frac{d}{dx}(\sqrt{x}) \\ &= 2 \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-x/2} \cdot \frac{1}{2\sqrt{x}} \\ &= \frac{e^{-x/2}}{\sqrt{2\pi} \cdot \sqrt{x}} \end{aligned}$$

And we have obtained the density function for $X = Z^2$. The random variable X is called a *chi-squared distribution with 1 degree of freedom.*

7.2 The Chi-Squared Distribution with ν Degrees of Freedom

You might be wondering about the somewhat mysterious name of the random variable whose density we have just derived. What is this "degree of freedom" thingamabob? And, since we explicitly stated that the density had 1 such "degree of freedom", that suggests that maybe there are other chi-squared distributions that have more degrees of freedom. What's going on here?

It will be a while before we can address the concept of "degrees of freedom"; for now you'll just have to accept that this is a parameter for the distribution, and leave it at that. But it's certainly the case that chi-squared distrbutions can have degrees of freedom greater than 1. So let's see the general definition.

Definition A *chi-squared distribution with ν degrees of freedom* is a probability distribution with one parameter, denoted ν and called the "degrees of freedom", with density function:

$$f_X(x) = \frac{x^{\nu/2-1} \cdot e^{-x/2}}{2^{\nu/2} \cdot \Gamma(\nu/2)}$$

If a random variable X follows a chi-squared distribution wth ν degrees of freedom, we write $X \sim \chi^2(\nu)$.

I know that this must be a disappointment for you, just being given the density function without seeing it derived as the result of some cool

transformation. The problem is that the chi-squared distribution with ν degrees of freedom is in fact **not** derived as a transformation of a random variable; instead, it's obtained by another technique, which we'll see next week. So you'll have to wait until then, but I promise you it will be very cool indeed!

Notice that if we set ν equal to 1, the density becomes:

$$\begin{aligned} f_X(x) &= \frac{x^{\nu/2-1}e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)} \\ &= \frac{x^{1/2-1}e^{-x/2}}{2^{1/2}\Gamma(1/2)} \\ &= \frac{x^{-1/2}e^{-x/2}}{2^{1/2}\Gamma(1/2)} \\ &= \frac{e^{-x/2}}{\sqrt{2\pi}\cdot\sqrt{x}} \end{aligned}$$

And this is just the density that we derived previously and called the "chi-squared distribution with 1 degree of freedom". So everything works out the way it should.

Two quick notes on terminology: the parameter ν that I'm using is **not** the letter "v" from the Roman alphabet; instead, it's the letter "nu" from the Greek alphabet. If you're typesetting in LaTeX, the symbol is `\nu`. Also, I am calling this distribution the "chi-squared" distribution, but many people just say "chi-square"; either one is fine.

7.3 Moments

Now that we have the density function for the chi-squared distribution, we can calculate the moments. Let's start with the first moment i.e. the expected

value:

$$\begin{aligned}
\mathrm{E}[X] &= \int_{\Omega_X} x f_X(x) \cdot dx \\
&= \int_0^\infty x \cdot \frac{x^{\nu/2-1} e^{-x/2}}{2^{\nu/2}\Gamma(\nu/2)} \cdot dx \\
&= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty x^{\nu/2} e^{-x/2} \cdot dx
\end{aligned}$$

Do you see what the integral is? It's a "kinda sorta" gamma function i.e. it would be a gamma function if the exponent for the exponential function were $x/2$ instead of x. We can use the formula we derived in the first lecture to evaluate this:

$$\begin{aligned}
\int_0^\infty x^{\nu/2} e^{-x/2} \cdot dx &= \frac{\Gamma(\nu/2+1)}{\left(\frac{1}{2}\right)^{\nu/2+1}} \\
&= 2^{\nu/2+1} \cdot \Gamma(\nu/2+1)
\end{aligned}$$

Now we have:

$$\begin{aligned}
\mathrm{E}[X] &= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty x^{\nu/2} e^{-x/2} \cdot dx \\
&= \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \cdot \left(2^{\nu/2+1} \cdot \Gamma(\nu/2+1)\right) \\
&= \frac{2^{\nu/2+1}}{2^{\nu/2}} \cdot \frac{\Gamma(\nu/2+1)}{\Gamma(\nu/2)} \\
&= 2 \cdot \frac{\nu}{2} \\
&= \nu
\end{aligned}$$

To obtain the variance of a chi-squared random variable with ν degrees of freedom, we could calculate the second moment $\mathrm{E}[X^2]$ and the use the standard variance formula. But I don't want to deny you the enjoyment of doing this for yourself, so this will be a homework problem. But I'll give you a hint: when you're all done, you should end up with $\mathrm{Var}[X] = 2\nu$.

So, to summarize, if X is a chi-squared random variable with ν degrees of freedom, then it has the density function:

$$f_X(x) = \frac{x^{\nu/2-1} \cdot e^{-x/2}}{2^{\nu/2} \cdot \Gamma(\nu/2)}$$

The expected value and variance of X are:

$$\begin{aligned} \mathrm{E}[X] &= \nu \\ \\ \mathrm{Var}[X] &= 2\nu \end{aligned}$$