Distribution Overview: Probability by the Seat of the Pants

Paul Johnson

August 30, 2011

1 Why Do We Need Probability Concepts?

I'll start by challenging you.

1. Describe the range of test scores you expect when we test my students in U.S. politics?
2. Describe the number of times per month that your neighbor's dog will wiggle under the fence and escape.

Most people I know agree that answers like the following are, more or less, acceptable.

1. The test scores range from 0 to 100, the proportion of students who earn each score will be something like this:

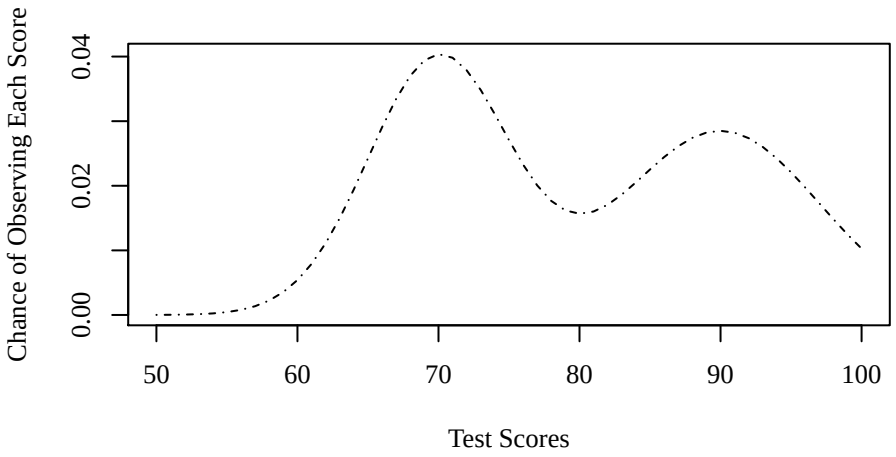


Figure 1: A Probability Distribution of Test Scores

I have to be a little bit careful in describing this figure. The curve represents probabilities, but what does that mean? At the current time, my best answer is this. Draw

one person's name from a hat (all names equally likely). Without any additional information, the curve tells me that the most likely score is approximately 70. Scores below 70 are less likely, but if we consider 80 and above, the most likely score is 90. These beliefs reflect my experience as a teacher. My class tends to have one big clump of solid C students and an smaller, well defined clump of students for whom the average is A-. When those two groups are combined, a "two humped camel" graph emerges.

2. The dog will probably not escape at all, but there is a decent chance it will escape once, and lesser still is the chance of 2 escapes, and the chance is lower for each successive number of escapes.

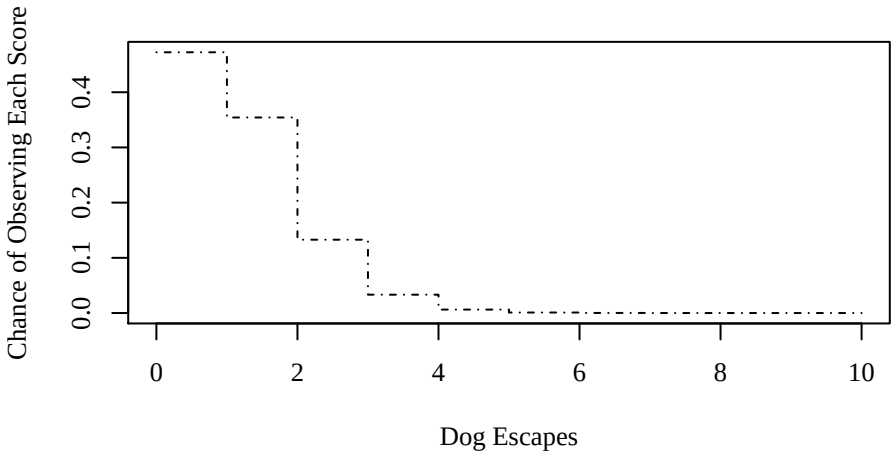


Figure 2: The Chances of a Given Number of Dog Escapes

Maybe you want to re-draw these answers. I don't mind. Beliefs are likely to vary. The key idea is that we can use models, "probability distributions," to describe our expectations. They specify the range of observations and the chances that various scores will be observed.

Describing what we expect to see in a sample is one of the important purposes of probability models.

Now let me challenge you again.

1. Describe the average score we are likely to observe on the first test in my US politics course.

   I suppose you'll nag me for more information. This semester, the average on the first test was 78.3. Last year, the average was 77.1. Before that, it was 79.2. I have tenure now, and I expect to impose myself on the students for about 100 more years, so you can wait and give your answer later, if you want to. I'll forward the data to you.

2. Suppose you keep track of your neighbor's dog escapes for months and months. What do you expect the average number of escapes per month will be?

These questions ask for us to summarize across a series of observations. These are a little bit more abstract, but not too difficult. There is no right or wrong answer, I am only asking about your opinion.

Here are my answers.

1. I believe the test averages will follow this pattern.

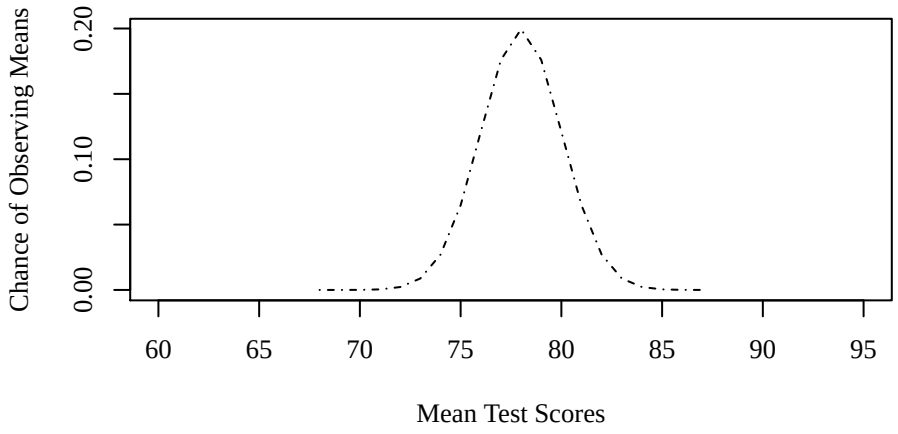


Figure 3: My Beliefs About the Class Mean

I think the average is most likely to be 78, and I'm very confident it will be in a range from 74 to 82. There's still a chance it might be more extreme, but I'm pretty doubtful. This distribution has one hump. I've not drawn it that way by mistake. It seems that when we average the scores of a class, we smooth out the bumps of the score distribution. The distribution of the means is simpler.

2. I expect the average number of dog escapes will be between 0 and 1, and there's a very small chance that it will be greater than 1. And, obviously, it cannot be negative.

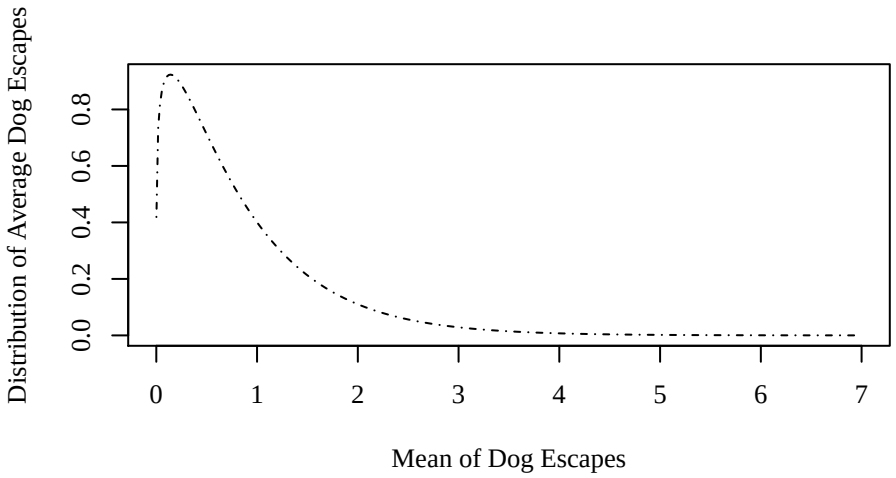


Figure 4: My Beliefs About the Number of Dog Escapes

> In my mind, the average number of escapes is 0.84. That's just what I think, on the basis of my experience, and the chance that the mean number of escapes is greater than 5 is, well, really small. I think it is more likely that monkeys will fly out of your ... ear than it is that the average number of escapes will be 7.

The first challenges asked us to describe the range of observed outcomes "right now." Somehow, it seems more "tangible" and "realistic" to summarize observed scores.

The second set of challenges is somewhat more abstract, but to me, problems of this type are more satisfying and interesting. They ask us to describe the range of outcomes we might observe across a series of experiments, *even if we do not actually conduct the experiments!* If you want to wait until I'm 150 years old, or until you have lived next to the same neighbor for 100 years, you can have an exhaustive set of data and revise your estimates. It will be pretty boring, and, to the surprise of most people, it is not necessary. This latter type of reasoning–imagining the variation of estimates that would arise–is the most important part of statistics. It is called "inferential statistics". From a set of observations, we can make statements not only about summary values like "means," but we can also summarize our uncertainty about those estimates.

2 Key Terms: population and random variable

There is a problem with the word "population." Ordinary people use that to refer to "all the people who are alive," but statisticians usually mean something different. Quite often, statisticians use the term population to mean "something from which observations are drawn" and the conclusions they draw about that process are thought to characterize the process that generates observations, rather than simply describing "all of the people." To differentiate the "population as process" from "population as finite collection" usages of the term, some will

refer to former as a "superpopulation." That term is not widely used in applied statistics, but statisticians do use it when they are trying to be very clear.

Most of the time, when we are talking about statistics, we are not talking about estimating features of "the people" (or "the fish in the river" or whatever), instead we want to characterize the process that governs the creation of data. If we were studying a finite population, all of our conclusions would be instantly out of date when research is reported, since members of that finite population will have died or otherwise exited.

When we study the population (the superpopulation), researchers usually have in mind a bigger meaning:

population - process that generates observations. This has the same meaning as "stochastic process"

We seek to "characterize" a population by a mathematical formula, an equation that depends on some important coefficients we call "parameters."

parameter: adjustable characteristic that alters the qualities of a probability process.

Our goal in probability analysis is to write down a "mathematical model" that we can understand, and then investigate its properties. We usually do that by imagining how predictable observations from that process might be.

random variable - an observation–a number–drawn from population (a score "pulled" from a random process that is thought to be governed by mathematical laws).

Once we have a pretty good understanding of the theoretical properties, then we might try to translate them into a study of observations from out there in the "real world." We wonder, "do those *things* we observe *out there* come from a process that resembles the theory that we explore in our minds?"

Example: Rolling dice. If I say each side of a die is likely to land facing up with probability 1/6, I don't mean to say I've rolled a million dice and counted. I intend to characterize the process itself, not the results on a million rolls. If I roll the die 50 times and try to re-evaluate the probabilities, I'm not trying to estimate the number of 1's I'd get in a sample of 1 million rolls. Instead, I want to know the chances of a 1 on any given roll of that die.

Example: Coin flips. Imagine a fair coin. Each outcome is equally likely: the probability of head is 0.5 and the probability of a tail is 0.5. We have a probability model for that kind of process. The "Binomial distribution" describes the chances of observing a certain number of heads and tails. Then we wonder if the referee who administers coin flips before football games is a random process of that type. If there are 10 games this week, we might collect a string of data, {H,T,H,...}. Those ten flips represent a collection of scores from a random process. From that sample, we often want to find out if the referee's coin is "fair." That is, does the data match the theory?

The term "random" is frequently misunderstood. It does not mean "equally likely". It means outcomes are generated according to a given probability process. The term "random" thus means "patterned unpredictability." A process would still be called random if there were two outcomes and one occurs 'almost all the time.'

Finite Population Interpretation

A competing interpretation of population is the "finite population" view. I think this is not usually useful in the advanced contexts of statistics, but it is frequently taught in elementary statistics courses (even as the teacher will usually say, "this is not quite right, but it might help you to get started").

In this view, the population is thought of as a fixed collection of things from which we draw examples (as if we were blindfolded pulling numbers in a BINGO parlor).

In an introductory course on probability, almost every student has suffered through silly exercises like this.

The classic example: Consider an Urn with 1000 balls, X are red, $1000 - X$ are blue. From a sample of 20 balls, 6 of which are red, estimate the fraction that is red in the population of 1000 balls.

Perhaps this view could be useful if we could convince ourselves that there is a fixed, finite set of things we want to know about. But I don't have many Urns full of balls I need to study. I'm straining myself to think of a realistic example. Consider dental cavities among prison inmates. Suppose that no new prisoners are brought in, none are released. None of them can die or escape. The "population" might actually mean "these particular prisoners." We want to know what fraction of the prisoners have a cavity in the second molar on the left lower jaw. Our prison is too large to actually check them all, so we have the dentist examine some of them. From that sample, we might estimate the number of cavities.

I often belittle this view, saying its practitioners are mainly interested in tedious projects, such as estimating the number of left-handed red heads in a Cincinnati. There are various weaknesses in the finite population interpretation. It is not exactly fatally flawed, but it does make some parts of statistics very difficult to understand (in my humble opinion).

Here is one such example. The idea of drawing "independently and identically distributed" (iid) observations into a sample is not practical within the finite population perspective. We need iid observations in order to derive almost all of the results in inferential statistics and maximum likelihood analysis. The observations must be independent so that we can multiply them to calculate their joint probability. They must be identical so we can act as if they come from the same distribution. If we take the finite population approach, it is impossible to draw a sample of identically distributed observations because taking one case out of the population alters the characteristics of the remaining cases, and the next draw will not be statistically identical to the original case.

In some contexts, one can reach the same conclusion from either perspective. Because some sampling ideas are more easily developed with an urn of colored balls, that model is still used in elementary statistics. In some practical areas of applied statistics, where it is necessary to find out how many acres are currently infected with a blight, the finite approach is prominent. But almost all of the workaday tools of modern research scientists are based on the idea that the observations we are studying are drawn from a random probability process, rather than a finite collection of things.

3 Probability Theory: The Language of Statistics

3.1 Sample Space

Generally speaking, the sample space is the set of all "outcomes" (people, opinions, animals, etc) that might be observed as a result of a random process.

Discrete Sample Space: the possible outcomes are drawn from a finite list, $X = \{1, 2, 3, \ldots\}$

Continuous Sample Space: possible outcomes are drawn from a continuum (the real numbers, $\mathbb{R}$), either bounded (such as $[0, 1]$ or infinite $(-\infty, +\infty)$ or half closed, $[0, +\infty)$).

3.2 Probability

In my opinion, it is easiest to think of probability as a property of a "region" or "subset" of possible outcomes.

$p(x \leq 4)$ the chance that one randomly drawn observation x is less than or equal to 4.
There are many different kinds of notation. If I'm worried the reader will forget, I will often write $Pr(x < 4)$ or even $Prob(x < 4)$. There is something to be said for letters from other alphabets. How would you feel about $\pi(x \leq 4)$?

$p(7 \leq x \leq 9)$ the chance that x is in $[7, 9]$ – between 7 and 9, inclusive.

I don't think it pays to invest too much effort to answer the question "what is probability?" This is a point of contention between competing schools of thought, and by the time you are qualified to answer the question, you will ready have decided which camp you prefer, and the answer will seem completely obvious to you.

In a nutshell, the competing views are as follows.

1. View 1. Probability is "long run relative frequency".

   Suppose we draw observations over and over, forever. After an infinite number of draws, the observed fraction of x's observed will match $p(x)$. If I ask you, "what is the chance that the next outcome will be x," and you answer, "wait a minute, I have to conduct an infinite series of experiments to find out," then you belong in this group of scholars.

2. View 2. Probability as "degree of belief".

   Probability models summarize a person's theory of the world, a belief about what is likely to happen in any one "flip of the coin" or "roll of the dice." How strongly do you believe that the next observation will be x? If your answer is, $p(x)$, then you belong in this group.

These two views are, essentially, the difference of opinion that keeps the "frequentist" statisticians and the "Bayesian" statisticians at war with each other.

How to avoid the philosophical disagreement over the meaning of probability.

Consider a discrete variable that can take on values $\{1, 2, 3, 4, 5\}$.

A probability model must list the possible outcomes and assign a number $p(x_i)$ to each one.

Outcome	$x_i =$	1	2	3	4	5
probability	$p(x_i)$	1/5	1/5	1/5	1/5	1/5

Here we suppose the 5 outcomes are equally likely.

It is not necessary to take that view, however. Fiddle the $p's$ however you like:

Outcome	$x_i =$	1	2	3	4	5
probability	$p(x_i)$	0.05	0.1	0.2	0.3	0.35

And your friend who hates the number 3 might as well have a turn writing down his favorite model:

Outcome	$x_i =$	1	2	3	4	5
probability	$p(x_i)$	0.25	0.25	0.0	0.25	0.25

These are all probability models because:

1. No outcome is less likely than impossible (that is, $p(x_i) \geq 0.0$).

2. The sum of the probabilities of all outcomes is 1.0.

The reader can re-assign the probabilities in any way he or she wants to, as long as the result meets those requirements. The result will be a probability model, in my opinion.

I suggest you stop worrying about subjective meaning of $p(x)$ at this stage. Just try to develop an intuition for the analysis that we perform with these numbers. The mathematics of probability analysis will "flow," no matter how you interpret the probability values. That's why I don't think it is worthwhile to spend too much time worrying about what a probability number "really *is*." If you really want to get metaphysical about it, please wait.

3.3 Multiplication and Addition Principles

Multiplication Principle. The chance that m separate things will happen equals the product of their individual probabilities.

Example 1: Roll a die. Let's suppose it is a fair die. The chance of it landing on 1 is 1/6. Roll again. The chance of 1 again is 1/6. Thus, the chance of rolling 1 twice in a row is $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$. The chance of rolling another 1 is 1/6, but the chance of rolling three 1's is $\frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} = \frac{1}{216} = 0.00462963$. And so forth.

Now, suppose I pull a die from my pocket and roll three 1's in a row. If the die is fair, you've observed something that is fairly unusual. The probability of three 1's is smaller than 5 out of 1,000 sets of rolls. Now suppose I roll another 1. Wow, that's crazy! The multiplication principle says the probability of a fair die generating four 1's is 0.000771605. Then I roll another 1! The probability is 0.0001286. A fair die would give me five 1's in a row only about 1 time in 10,000.

At some point, you have to interrupt me with this story and ask to inspect my die. It is a special one I had made up just for this purpose. There is only one spot on all six faces.

And that is the lesson to be learned. If you work out a probability model, and then observations seem to fly in the face of the probability model, then you have to at least consider the possibility that your probability model is wrong.

Example 2: Survey 1000 people, ask them if the death penalty should be enforced against convicted first degree murders. The chance that respondent 1 will say yes is $p(x_1 = Yes)$. The chance that the second person says yes is $p(x_2 = Yes)$, and so forth. Then the chance of observing a particular $(Yes, Yes, No, Yes, \ldots, No)$ pattern is the product of all respondent probabilities:

$$\begin{aligned} & p(x_1 = Yes, x_2 = Yes, x_3 = No, x_4 = Yes, \ldots, x_{1000} = No) \\ = \quad & p(x_1 = Yes)p(x_2 = Yes)p(x_3 = No)p(x_4 = Yes)\ldots p(x_{1000} = No). \quad (1) \end{aligned}$$

The probability of observing a whole sample is equal to the products of the individual events.

This is only true, obviously, if each person's response is independent of each other person's response, but that is commonly assumed in public opinion surveys. The probability of observing a sample of responses is thus calculated by combining the probabilities of individual responses.

We can use a mathematical trick to make analysis of this overall probability more manageable. Remember the following mathematical law for logarithms:

$$log(x \cdot y) = log(x) + log(y). \quad (2)$$

The mnemonic for that is, "The log of a product is the sum of the logs."

That means we can convert the long quantity in expression (1) into a sum of logs,

$$\begin{aligned} & log(p(x_1 = Yes, x_2 = Yes, x_3 = No, x_4 = yes, \ldots, x_{1000} = No) = \\ & log(p(x_1 = Yes)) + log(p(x_2 = Yes)) + \ldots + log(p(x_{1000} = No)). \quad (3) \end{aligned}$$

In a more advanced statistical context, this kind of reasoning is known as "maximum likelihood analysis," a procedure through which we study the probability of observing a set of responses as a product of individual probabilities. This last expression would be known as the "log of the likelihood function." It is a very important foundation in probability analysis.

Note: These calculations presuppose the two events are "independent." The chance that one will occur is not affected by the chance that the other will occur.

Addition_Principle. To determine the probability that one event among a list of possibilities might occur, add their probabilities together.

Example 1: Roll a die with 6 sides. How likely am I to roll a number smaller than 3?

$$p(x = 1) + p(x = 2) = \sum_{i<3} p(x = i)$$

Example 2: People can be right-handed, left-handed, or ambidextrous. What is the probability that a randomly selected person will not be left handed?

$$p(x = "right - handed") + p(x = "ambidexterous")$$

Example 3: The chance that a randomly chosen American will say she is a "strong Democrat" is 0.15. The chance that a person is a "weak Democrat" is 0.20. That means the chance that the person is a Democrat, either strong or weak, is

$$p(x = "Strong\, Democrat") + p(x = "weak\, Democrat") = 0.15 + 0.20$$

The chance that one among a collection of possibilities will occur is equal to the sum of their probabilities.

4 Terminology for Discrete Distributions

A discrete distribution is one for which we have a list of possible outcomes, $\{x_1, x_2, \ldots, x_m\}$. Outcomes can be placed into one-to-one correspondence with the integer 'counting numbers'.

If an observation is drawn at random, the chance that it will fall into the $j'th$ category, meaning the observed value will be x_j, is $p(x_j)$.

4.1 Probability Mass Function

Any formula $p(x)$ can be a **probability mass function** (PMF) if

$$1 \geq p(x) \geq 0$$

and

$$\sum_{x_j \in X} p(x_j) = \sum_{j=1}^{m} p(x_j) = 1$$

Sometimes I use notation $P(X = k)$ instead of $p(x_k)$. Either way should be clear enough.

Discrete distributions are needed when the process we are considering offers up outcomes that are "grainy," in the sense that they cannot be averaged together. In the case of a die, for example we can observe a 3 or a 4, but nothing in between. We never roll 3.5. An animal is either a cat or a dog; except in cartoons, there is no such thing as a "cat-dog".

5 Continuous Random Variables

Until now, I made this easy on myself by discussing coins, dice, and Democrats. Most of the important distributions we need to work on are not discrete. Rather, they are continuous: the outcomes correspond to segments of the real number line or the Cartesian plane.

With continuous numbers, we might have a temperature of 80, or 90, or any number between them. Variables like time, the proportion of citizens who call themselves Democrats,

blood pressure, and so forth, lend themselves much more readily to treatment on a mathematical continuum.

How are "continuous numbers" different from "discrete numbers"? I hate to bring up bad memories for readers, but do you remember high school geometry? When the teacher discussed points and lines, did she say "a point has no width"? That seemed wrong. It bothered me. It seemed obvious that if I drew a dot on the paper, my point did have some width. With a magnifying glass and a ruler, I could measure it. With a microscope, I could even see texture inside the dot. That bothered me so much I couldn't pay attention for a whole semester.

Here's the part I did not appreciate. When I measured the width of the pencil mark, I was simply mistaken in thinking I was measuring a point. I was measuring the distance from the outer left edge of the pencil dot to the outer right edge of the dot. I was not measuring "a point," I was measuring the diameter of a region. I was measuring the "distance between two points." Points play the role of "edge markers" and two points mark off a region.

In the same sense that a "point has no width," a single point has no probability. A single point does have "probability density," however. And, after we understand density, we can make the next step to measure the "probability" between two points (no magnifying glass required).

5.1 PDF: probability density function:

The probability density function (PDF) describes the probability density of each possible outcome. The term "probability density" is somewhat elusive, but I expect it will reveal itself to you before this section is finished.

For the sake of clarity–to keep the continuous density separate from the discrete probability model separate in our minds–it is customary to use a different letter for the PDF. I'll use the letter f, most of the time.

A pdf f for x must meet two requirements.

1. The value of $f(x)$ cannot be negative

$$f(x) \geq 0 \tag{4}$$

2. The area under that function must be 1.0.

$$\int f(x)dx = 1.0 \tag{5}$$

The domain of a continuous variable might be the whole range from $(-\infty, +\infty)$. Any "chunk" of the real line will also do, such as $[0, 1]$ or $(0, \infty)$.

A continuous density function must be interpreted differently than a discrete distribution.

Where do probability density functions come from?

Any function that has non-negative values and a finite value for its integral can be converted into a PDF.

How can this ugly thing be a PDF?

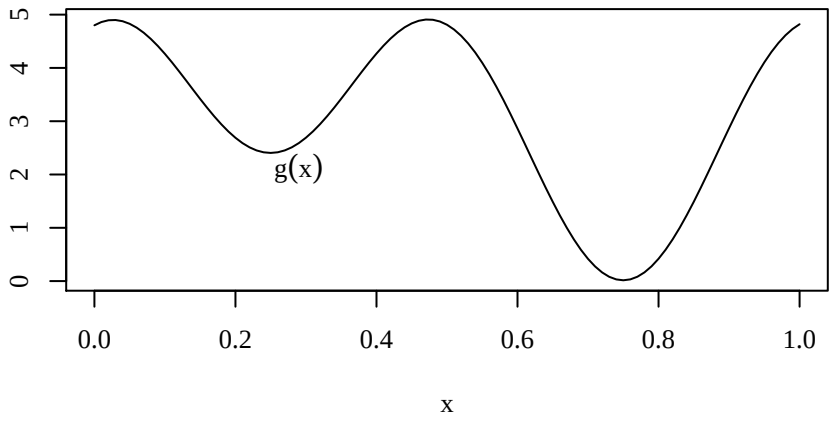


Figure 5: A Function That Might Become a PDF

We need to normalize it, so that the area under the curve is 1.0. Calculate the integral (the area under the curve between 0 and 1),

$$\int_0^1 g(x)dx = 17.$$

The magic number is 17. We convert $g(x)$ into a PDF by division, which normalizes $g(x)$.

$$f(x) = \frac{g(x)}{17}.$$

That is a valid PDF because

$$\int_0^1 \frac{g(x)}{17}dx = 1.0$$

Note that this assumes that $g(x) \geq 0$, for all x, but that's a pretty minor restriction, since you can add and multiply anything you want.

The function $g(x)$ in this example is called the "kernel" of the pdf, because it describes the substantively important variation that we are studying. The value 1/17 is a "normalizing constant." It is not substantively important, it is only needed to bring the area under the curve down from 17 to 1.0.

Claim: Any positive function with a finite integral can be the kernel of a density function.

5.2 Cumulative Distribution Function

A point has no width, no probability. But the chance of an outcome between 0.001 and 0.002 can be calculated. That region might be what we have in mind when we say "point," but it is a region. When we talk about hitting the target in darts or golf, we don't actually intend to hit a particular point, we mean a region between two points. Finding the chances of an outcome in that region is the main purpose of the cumulative distribution function (CDF). It helps us to answer questions like "what is the chance that the outcome will be greater than 9?" or "what is the chance that the outcome will be between points A and B?"

The cumulative distribution function is commonly referred to by the capital letter of the density function. $F(x^u)$ represents the probability that a randomly drawn value from the distribution $f(x)$ will be smaller than a target value, x^u. It is the area under f from the "left edge" of the possible outcomes (which may be $-\infty$) up to x^u.

$$F(x^u) = \int_{-\infty}^{x^u} f(x)dx. \tag{6}$$

We often refer to the CDF simply as $F(x)$, keeping in mind that the parenthesized value is the upper limit of an integral. If we use the letter x for the input variable in $F(x)$, then the math teachers will want us to use some other letter for outcomes. For example,

$$F(x) = \int_{-\infty}^{x} f(y)dy. \tag{7}$$

I call the upper limit x^u in order to avoid choosing a new letter.

A Brief Example: The Uniform Distribution

The simplest PDF is the Uniform probability model, $U(a, b)$, which holds that all of the outcomes between a and b are equally likely. The PDF is $f(x) = \frac{1}{b-a}$ and the probability that the outcome will fall between any two values, say x^l and x^u, is easy to calculate:

$$Prob(x^l < x \leq x^u) = \int_{x^l}^{x^u} \frac{1}{b-a}dx = \frac{x^u - x^l}{b-a}$$

This is particularly easy if we take the Uniform on $[0, 1]$, because the PDF just $f(x) = 1.0$.

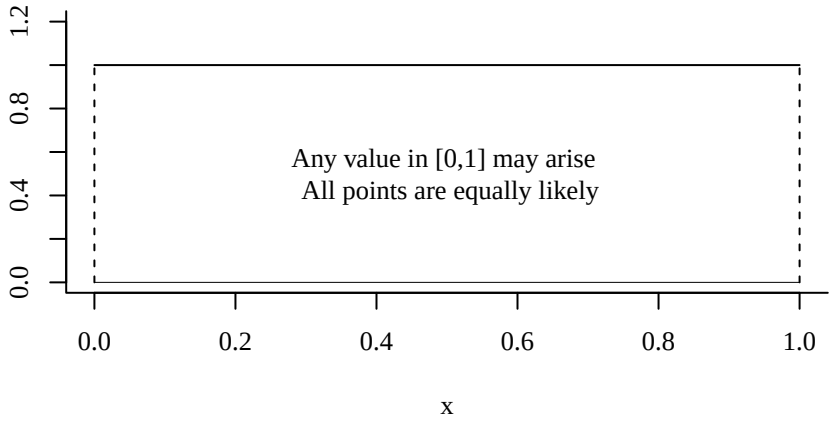


Figure 6: Uniform Distribution

The fact that $f(x) = 1.0$ does not mean that a point x is "certain" to occur. This is one of the little mathematical wrinkles that make PDFs different from PMFs. The value of 1.0 is a density value, and its main substantive meaning is that we are able to group together sets of outcomes and then the probability of an outcome in a set is given by the integral of the PDF. For example, the chance that an outcome will lie between 0.2 and 0.5 is represented by the shaded region below:

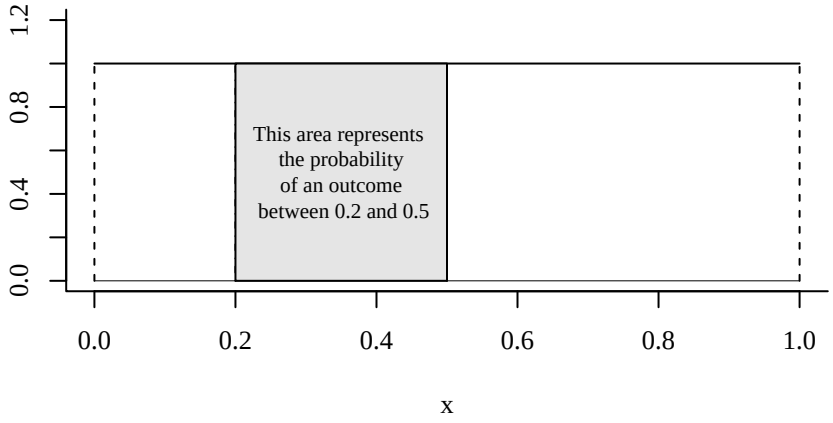


Figure 7: Probability that an Outcome will Lie Between Two Values

Because we are considering the simple case, where $a = 0$ and $b = 1$, the CDF is easily

seen to be

$$F(x^u) = \int_0^1 f(x)dx = x^u$$

The linkage between the PDF and the CDF is illustrated next. Consider the PDF of the Uniform, with special attention to the chances of outcomes smaller than 0.3, 0.5, and 0.7.

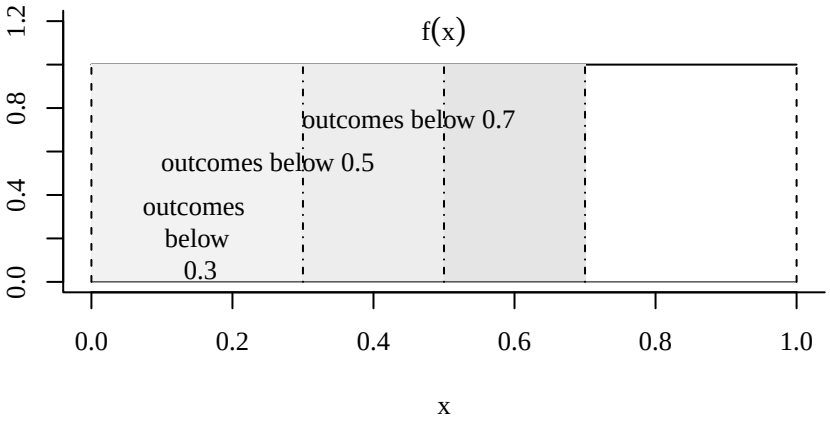


Figure 8: A Uniform Probability Density Function

Those three values of x are marked on this curve, which represents the cumulative distribution.

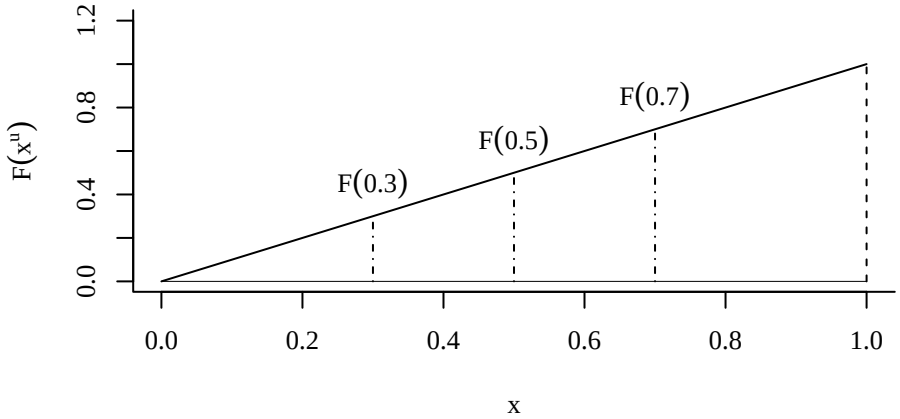


Figure 9: Cumulative Distribution

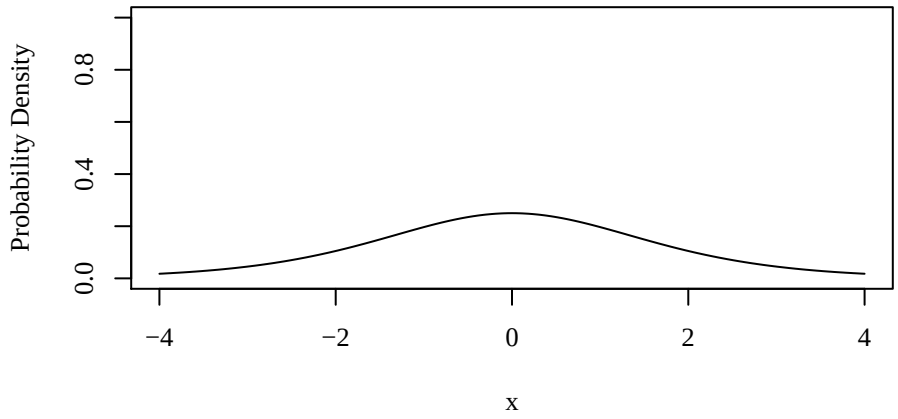
All CDF's are strictly increasing functions of x^u. As x^u moves to the right, $F(x^u)$ always grows. For PDFs that are unimodal, the CDF is an "S-shaped curve." In Figure 10, the

PDF of a unimodal, symmetric PDF is presented, along with the S-shaped CDF that it leads to. The figure is based on the "logistic distribution," a distribution that is often used in categorical models of choice because is has a comparatively simple CDF. It is one of the very few distributions for which the CDF appears to be simpler than the PDF. The PDF is

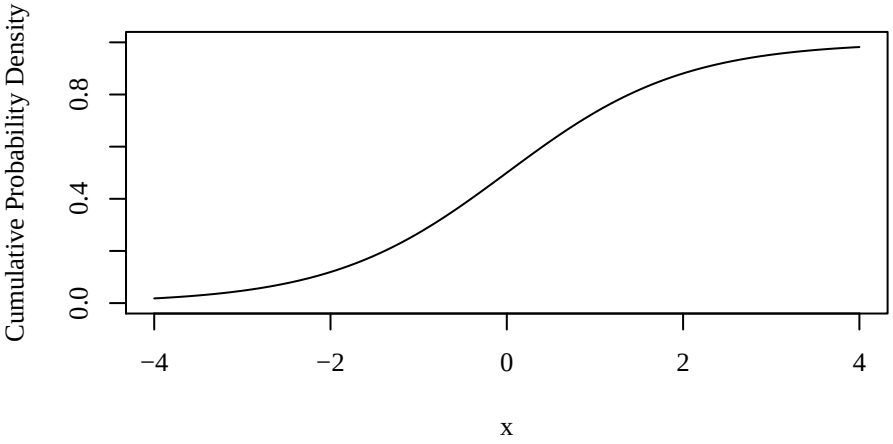
$$f(x) = \frac{1}{\sigma} \frac{exp(\frac{x-\mu}{\sigma})}{\left(1 + exp(\left(\frac{x-\mu}{\sigma}\right))\right)^2} \tag{8}$$

but the CDF is simply:

$$F(x^u) = \frac{1}{1 + e^{-\frac{x^u-\mu}{\sigma}}} \tag{9}$$



(a) The PDF of a Logistic Distribution



(b) The CDF of a Logistic Distribution

Figure 10: The CDF is S-shaped

6 Moments

The **expected value** of a distribution is defined as the "probability weighted sum" of outcomes.

For a continuous distribution, with density $f(x)$:

$$E[x] = \int_{-\infty}^{+\infty} f(x) \cdot x \, dx. \tag{10}$$

For a discrete distribution, with probability mass function $p(x_j)$:

$$E[x] = \sum_{j=1}^{m} p(x_j)x_j, \quad where\, p(x_j) = prob.\, of\, outcome\, x_j. \tag{11}$$

As a result, the expected value of x is often referred to as the "population mean" or "theoretical mean" of a distribution that generates x.

The expected value is the first moment of the distribution. The expected value of x^2, $E[x^2]$, is the second moment, $E[x^3]$ is the third moment, and so forth. In a course on mathematical statistics, one learns that the moments characterize a distribution and allow us to make many important calculations, including the variance, as we see next.

The **population variance** of a distribution (or just "variance") is expected value of a squared deviation from the expected value. That is to say, it is the "probability weighted sum" of the squared differences between outcomes and their expected values. Variance is formally defined as a weighted sum of squared deviations

$$Var[x] = \int_{-\infty}^{+\infty} f(x) \cdot (x - E[x])^2\, dx, \tag{12}$$

which can be rearranged as

$$Var[x] = \int_{-\infty}^{+\infty} f(x) \cdot x^2\, dx - E[x]^2 = E[x^2] - E[x]^2. \tag{13}$$

Repeat out loud: "The variance of x equals the expected value of x squared minus the Expected value of x, squared." The variance can be calculated from the first two moments.

For a discrete distribution, the variance is defined similarly

$$Var[x] = \sum_{j=1}^{m} p(x_j)(x_j - E[x])^2, \tag{14}$$

and like (12), it is easily rearranged as

$$E[x^2] - E[x]^2 \tag{15}$$

6.1 How are Expected Value and Population Variance different from the "average" and the "variance"?

Here is a Very Important Point: Expected value and population variance are "theoretical quantities." They are characterizations of a probability model.

One can estimate the expected value and the variance with a sample, but should never (never) lose sight of the fact that $E[x]$ and $Var[x]$ are defined by a probability model.

The relationship between a sample mean and the expected value is demonstrated most easily with a discrete variable. Suppose a sample of scores is collected. The observed frequencies, the counts for outcome x_j, would be

$$freq(x_j) = \frac{\#\, of\, observations\, of\, x_j}{\#\, of\, observations} \tag{16}$$

Sample average, which is often referred to as $\bar{x}$, (pronounced "x bar"), is familiar to us as the sum of observations divided by N. With discrete data (m possible outcomes), the sample average, can also be calculated as

$$\bar{x} = \sum_{j=1}^{m} freq(x_j)x_j \tag{17}$$

The formula for the expected value (10) is almost the same, except that the observed $freq(x_j)$ is replaced by the *true probability* $p(x_j)$.

Similarly, the sample variance can be calculated as

$$\sum_{j=1}^{m} freq(x_j) \cdot (x_j - \bar{x}) \tag{18}$$

which is almost the same as the population variance in (14).

Now, I would like to unburden myself by expressing frustration about statistical notation. In many advanced statistical models, a parameter is given a Greek letter, say μ or λ, and an estimate calculated from data is differentiated by the addition of a hat, as in $\widehat{\mu}$ or $\widehat{\lambda}$. That practice is clear and consistent, but it has not trickled down to introductory statistics. In introductory statistics, the sample mean is commonly called $\bar{x}$, even though it would be more meaningfully written as

$$\widehat{E[x]}. \tag{19}$$

Some models will have a parameter, say μ_x or λ_x, that we find is equal to $E[x]$. In those cases, it is only natural to refer to the sample average as $\widehat{\mu_x}$ or $\widehat{\lambda_x}$. Nevertheless, it is quite common to refer to a sample average as $\bar{x}$.

Notation about estimates of variance is even more troublesome. The most direct notation for a sample estimate of variance would be $\widehat{Var[x]}$, and yet that is almost never done. Instead, most authors seize on the fact that in one particular distribution, there is a parameter called σ^2 that determines the distribution's variance. If $Var[x] = \sigma_x^2$, one will find all manner of notations to refer to the sample estimate, such as s_x^2. Before computerized type setting made it convenient to insert symbols above characters, s_x^2 was expedient. Nevertheless, it is much more clear to refer to an estimate as $\widehat{Var[x]}$ or $\widehat{\sigma_x^2}$.

7 The Forces of Nature May Not Know Our Formula (or agree with us about the parameters).

There's a lot of discussion in philosophy of science about whether or not our models exist "out there" in the world. Do leaves optimally rotate to consume sunlight? What does it mean to say the coin behaves "as if" it follows a probability model? I'm not sure if coins, trees, or social events can do math; I'm inclined to think they do not. Nevertheless, I also think that the patterns observed in nature can be described by mathematical patterns, patterns that

we socially construct and exchange among ourselves in mathematical language. And as we interact with nature, we try to improve our mental models by adjusting them.

A parameter is a numerical coefficient in a formula. It exists in our minds. We may project that parameter onto nature, saying things like "the data is produced by a process with parameters α and β." That kind of statement is common. It is certainly more fun to say "God generated this data with parameters γ and ϕ." I'm afraid this way of talking does confuse my students, but I still do it because it is fun, and somehow it exposes our mission. We believe parameters determine the generation of data; we want to know if our current understanding of those parameters is accurate.

The "parameter estimates" that are gathered from observation are not thought of as "exact matches" for the parameters in the models in our minds. If they were exact, science would be simpler, and possibly less interesting. Any experiment or set of observations will lead to a statement about the parameters that are most likely to correspond with the data.

8 Some Distributions

If any function–or just about any function–can serve as the basis for the creation of a probability density function, how can we bring our research problem under control? Are we supposed to study just any old function that pops into our heads?

I'm sorry to say the answer appears to be "yes," or, perhaps "maybe." On one hand, statisticians have already built a catalog of useful distributions. We have pretty good reason to believe that there are 10 or 15 "really important" distributions, functions that active researchers remember by name. On the other hand, there is an almost indefinite variety of possible probability density functions. In the late 1980s, I recall seeing a list of about 90 distributions that were more or less understandably different. In the early 2000s, the list had grown to 130 or so. Virtually any distribution can be generalized, distorted, truncated, or otherwise varied.

In the following list, I have collected the distributions that are most important in the foundations for most students. These are not comprehensive treatments, of course. Those can be found in many other places. My emphasis here is on understanding the "shape" of the various distributions, recognizing their parameters and the moments of the distributions.

A checklist of the features that are worth mentioning for each distribution should include

- domain. For what values is the formula defined? Does it take values in the whole real number line, or just for positive values, or perhaps just an interval like $[0, 1]$?
- location. Where is the "center" of the distribution (and is that center point substantively meaningful?).
- shape. Is it unimodal? Is it symmetric about its center point?
- scale. How widely "spread out" are the scores?

8.1 Exponential Distribution

The exponential distribution is shaped like a "ski slope," as illustrated in Figure 11. It represents the time that one must wait before an "event" occurs if the chance of an event depends only on the amount of time that passes. "Delta t" is the amount of time that passes, $\Delta t = t_2 - t_1$. If the probability of an "event" is $\lambda \cdot \Delta t$ (for Δt shrinking to 0), then the time waited before an event is exponentially distributed.

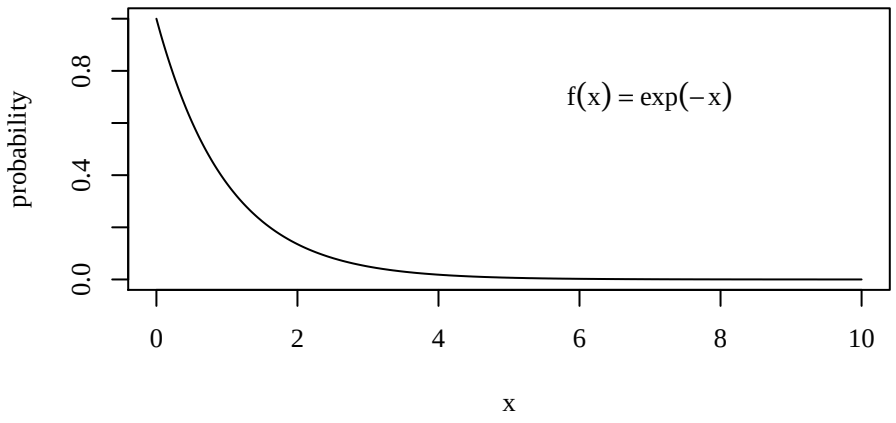


Figure 11: Exponential Density

8.1.1 Probability Density Function

$$f(x;\lambda) = \lambda e^{(-\lambda x)},\ where\, x \geq 0 \tag{20}$$

This is a very simple formula, with only one parameter, λ, which is called the "rate" parameter. In some texts, the parameter is expressed as the reciprocal of λ, so the density would be

$$f(x;\mu) = \frac{1}{\mu} e^{-x/\mu} \tag{21}$$

The letter e stands for Euler's (pronounced "oiler's") constant, roughly equal to 2.7182818 .... The value e^x is often represented as $exp(x)$.

In some texts, the density function will be written with the parameter as a subscript, as in $f_\lambda(x)$. That works well, except when there are several parameters. Sometimes it is simply written as $f(x)$ and the parameters are implicit. I prefer to include the parameters in parentheses after a semi-colon, mainly because they are printed in a more readable size.

Consider the exponential density when $\lambda = 1$,

$$f(x;\lambda = 1) = e^{-x}. \tag{22}$$

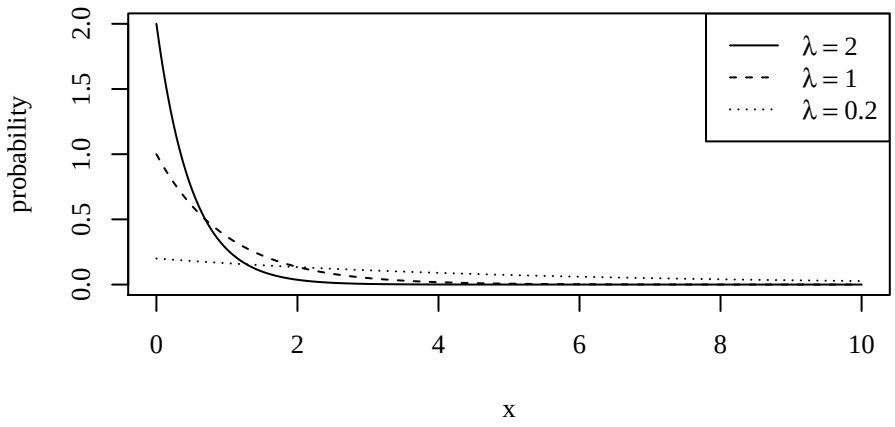
All of these notations represent that same function:

$$exp(-x) = \frac{1}{exp(x)} = \frac{1}{e^x}$$

The value of $exp(-x)$ shrinks 0 smoothly as x grows to infinity (see Figure 11).

If λ is very small, the decline in the value of $f(x;\lambda)$ is very gradual. The rates of decline are displayed in Figure 12

Figure 12: Three Exponential Densities



8.1.2 Cumulative Distribution Function

The cumulative distribution, the probability that a randomly drawn value will be smaller than k, is a very workable problem for a student who has completed elementary calculus.

$$F(k;\lambda) = \int_0^k \lambda e^{-\lambda x}$$

$$= -e^{-\lambda x} \Big|_0^k$$

$$= 1 - e^{-\lambda k} \tag{23}$$

8.1.3 Moments

First, begin with the result. The expected value of x for an exponential distribution is

$$E[x] = \frac{1}{\lambda} \tag{24}$$

This is not difficult to derive. Begin with the definition in expression (10) and insert the exponential:

$$E[x] = \int_0^\infty f(x) \cdot x \, dx = \int_0^\infty \lambda e^{-\lambda x} x \, dx. \tag{25}$$

This can be calculated with integration by parts.

$$\begin{aligned} &= -xe^{-\lambda x} \, |_0^\infty + \int e^{-\lambda x} \, dx \\ &= 0 - 0 + \left(-\frac{1}{\lambda} e^{-\lambda x} \right) |_0^\infty \\ &= lim_{x \to \infty} - \frac{1}{\lambda} e^{-\lambda x} + \frac{1}{\lambda} e^{-\lambda 0} \\ &= 0 + \frac{1}{\lambda} \end{aligned}$$

The variance of the exponential distribution is

$$Var[x] = \frac{1}{\lambda^2} \tag{26}$$

The easiest way to demonstrate that with elementary tools is by remembering that

$$Var[x] = E[x^2] - E[x]^2 \tag{27}$$

We have already derived $E[x]$, so we just need to solve for

$$E[x^2] = \int_0^\infty \lambda e^{-\lambda x} x^2 \, dx. \tag{28}$$

The work requires integration by parts, twice. When that is finished, we find

$$E[x^2] = \frac{2}{\lambda^2}$$

and so

$$Var[x] = -\frac{2}{\lambda^2} - \left(\frac{1}{\lambda} \right)^2 = \frac{1}{\lambda^2}$$

I've written this out because it is important for prospective researchers to understand that results about distributions must be derived *by someone* before they can be put to use. The characterization of a distribution can sometimes be a difficult problem that will require tools from mathematical statistics.

8.1.4 Comments

This is sometimes used to describe waiting times for events that are likely to happen quickly. For example, if we ask, "how long will we wait to hear a dial tone if we pick up a telephone," the answer (as of 2011, at least) is usually "no time at all." However, sometimes there is a period of silence before the dial tone appears.

This distribution is simple and very workable. Most applied research will be based on one of its more complicated relatives that is describe in the following sections, but one should not move past the exponential too quickly. It is the cornerstone of the "exponential family" of distributions, the family upon which the "generalized linear model" is based.

8.2 Normal Distribution

This is the single most important distribution in statistics. It is uni-modal and symmetric on the real number line. It may represent observed variables like IQ scores, while it also plays a vital role in the study of sampling distributions. When we draw samples over and over and calculate estimates from them, those estimates are likely to be normally distributed (this result is known as the Central Limit Theorem).

The starting point of the normal is the apparently simple function, $exp(-x^2)$, which is illustrated in Figure 13. It appears that we might simply have wondered out loud, "what happens if we square the input in an exponential distribution?" As we shall see below, the distribution may be moved to the left and right, or it may be stretched or squeezed, but the essence of it is simply $exp(-x^2)$.

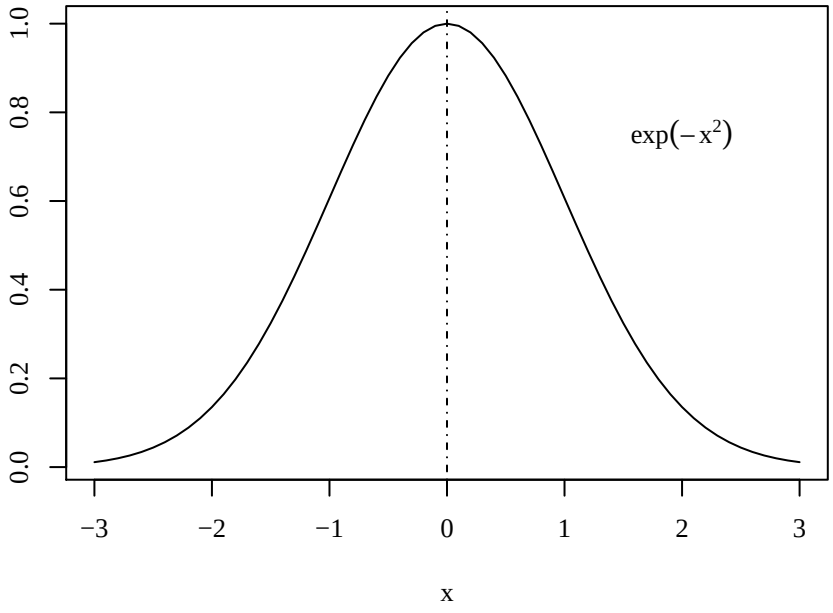


Figure 13: The Simplified Kernel of a Normal Distribution

Changes in μ and σ^2 have rather superficial effects on the distribution. They shift and scale it, nothing more interesting. As a result, it is often common to standardize a random variable by calculating it as a Z statistic.

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{29}$$

This new variable, Z_i, follows the standardized normal distribution, $N(0, 1)$. It is important to note that we do not lose any information by converting x_i into Z_i. The original variable x_i can be recovered from Z_i by multiplication, as in

$$x_i = \mu + Z_i\sigma. \tag{30}$$

This offers hints about a way with which to generate random samples from $N(\mu, \sigma^2)$. Many computer programs have built-in generators for standardized random normal variables. We can draw observations from $N(0, 1)$ and then re-scale, as in expression (30), to obtain draws from $N(\mu, \sigma^2)$.

8.2.1 Probability Density Function

The normal distribution, which I refer to as $N(\mu, \sigma^2)$, describes a continuous variable that takes on values in the real number line. The two parameters, μ and σ^2, determine the location and scale of the distribution. The probability density function is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)},\ -\infty < x < \infty. \tag{31}$$

The normalizing constant, $1/\sqrt{2\pi\sigma^2}$, is required to assure that the area under the pdf will be 1.0. The symbol π is, of course, "pi", roughly equal to 3.14159..., and e is Euler's constant.

The Greek letter μ, pronounced "mu", is a "location" parameter and σ^2, pronounced "sigma squared", is a "scale" parameter. In the literature, one will find the formula re-arranged in various ways, for example:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{ or}$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

To make sense out of the pdf, concentrate on the kernel, the part that depends on x.

$$e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{32}$$

To simplify, set $\mu = 0$ and $\sigma = 1$ (making this a standard normal distribution, $N(0, 1)$). That leaves us with

$$e^{-\frac{1}{2}x^2}$$

Throw away the fraction from the exponent, since it is a simple scaling factor. That means the relationship between x and its probability density is simple. We only need to understand

$$e^{-x^2} \tag{33}$$

The parameters μ and σ^2 move and stretch the shape, but they do not fundamentally change it. As illustrated above in Figure 13, that function is unimodal and symmetric.

The parameter μ is thus a "location parameter." The peak of the normal distribution's density, its mode (most likely value), will always correspond with μ. The shape of the distribution is unaffected by a change in μ. It simply shifts to the left or right, as illustrated in Figure 14.

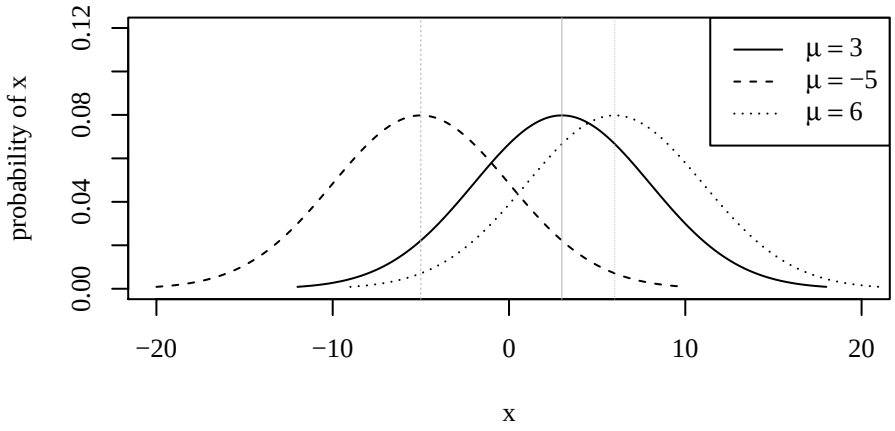


Figure 14: Three Normal Distributions

On the other hand, adjusting the σ^2 parameter changes the scale of the distribution. If σ^2 is very small, then points are tightly clustered around μ. Of course, it is possible to adjust both the location (μ) and scale (σ^2) at the same time. That is illustrated in Figure 15.

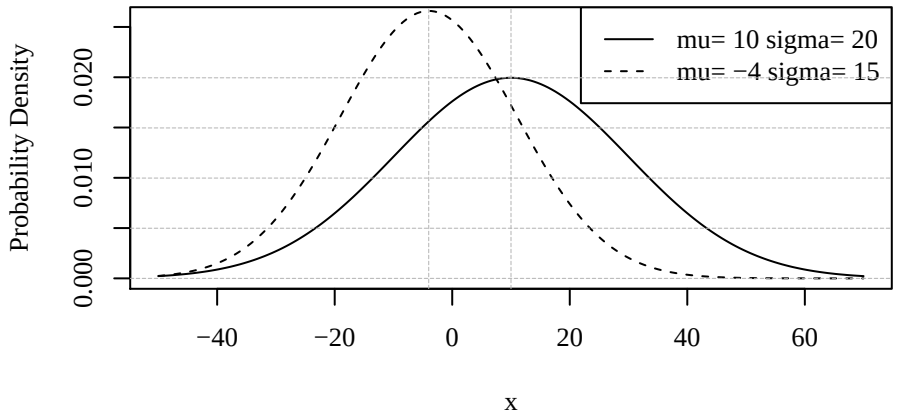


Figure 15: Compare 2 Normal Distributions

8.2.2 Cumulative Distribution Function

One of the truly frustrating facts in statistics is that the CDF of the normal cannot be simplified into an easily calculated formula. The chance of an outcome smaller than k cannot be written down more simply than the CDF itself,

$$F(k;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_0^k e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{34}$$

Hence, when we need to make calculations of $F(k;\mu,\sigma)$, it is necessarily to perform numerical integration. That is a difficult prospect. It was done in the days before computers by teams of calculating assistants who prepared large tables of solutions that were published in the appendices of most statistics texts. I was stunned to notice recently that the statistics text with which I have been teaching no longer includes those tables, presumably because they are "in the computer."

8.2.3 Moments

Since the normal is unimodal and symmetric, it should come as no surprise that the expected value of x,

$$\begin{aligned} E[x] &= \int_{-\infty}^{\infty} f(x;\mu,\sigma^2) \cdot x \, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} \cdot x \, dx \end{aligned} \tag{35}$$

is simply the center point of the distribution, μ. That is,

$$E[x] = \mu \tag{36}$$

The **variance** of a distribution is the "probability weighted sum" of the squared differences between outcomes and their expected values. It is a little bit harder to believe that this complicated expression

$$\begin{aligned} Var[x] &= \int_{-\infty}^{\infty} f(x;\mu,\sigma^2) \cdot (x - E[x])^2 \, dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} \cdot (x - E[x])^2 \, dx \end{aligned} \tag{37}$$

simplifies to this:

$$Var[x] = \sigma^2 \tag{38}$$

This claim can be derived with a tool called a "moment generating function" that is presented in the first part of a course on mathematical statistics.

The main point here is that the normal distribution's expected value and variance are extremely simple results. The parameter μ happens to be the expected value. It is also the mode and the median. The variance happens to be the parameter σ^2. This is not usually true that a distribution's parameters end up being equal to its expected value and variance. In some ways, we are spoiled by the normal distribution.

8.2.4 Comments

The normal distribution has many interesting qualities. Although in some ways it is a complicated function, in some ways it is very easy to work with. No doubt, on the difficult side of the ledger, we have the problem that the cumulative distribution function has no workable analytic solution. The only way calculate the chances of an outcome between two points is by numerical approximation. That makes computer programs run more slowly and, unless we are very careful, less accurately than they should.

On the easy side of the ledger, however, it is not too hard to calculate joint probabilities. Consider, for example, the chance that 2 independent observations from $N(\mu, \sigma^2)$ will be equal to particular values x_1 and x_2. That will be the product of the two densities,

$$f(x_1;\mu,\sigma^2) \cdot f(x_2;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{1}{2\sigma^2}(x_1-\mu)^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{1}{2\sigma^2}(x_2-\mu)^2}. \tag{39}$$

We can group like terms and use the laws of exponents to write this as

$$\frac{1}{\sqrt{2\pi\sigma^2}} \, e^{-\frac{1}{2\sigma^2}\{(x_1-\mu)^2+(x_2-\mu)^2\}}. \tag{40}$$

This generalizes easily if we need to know the chance that N separate observations will be $x_1, x_2, \ldots, x_N$,

$$\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2\sigma^2}\{\sum_{i=1}^{N}(x_i-\mu)^2\}} \tag{41}$$

Why is this useful? In maximum likelihood analysis, we frequently need to calculate optimal estimates for μ and σ^2. To my eye, it is quite obvious that maximizing this expression is the same as minimizing the sum of squares in the numerator. If that is not obvious to you, remember that

- a parameter estimate that maximizes a function also maximizes a monotone transformation of that function, and
- maximizing a function is the same as minimizing its negative.

The logarithm is a monotone transform. Take the natural log of (41),

$$\begin{aligned} & ln(\frac{1}{\sqrt{2\pi\sigma^2}}) + ln\left[\, e^{-\frac{1}{2\sigma^2}\{\sum_{i=1}^{N}(x_i-\mu)^2\}}\right] \\ = \;& ln(\frac{1}{\sqrt{2\pi\sigma^2}}) - \frac{1}{2\sigma^2}\{\sum_{i=1}^{N}(x_i-\mu)^2\}. \end{aligned} \tag{42}$$

It turns out that the maximum likelihood estimate of μ is the sample average:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{43}$$

The maximum likelihood estimate of σ^2 is

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{N}(x_i-\hat{\mu})^2}{N}. \tag{44}$$

These are clean, workable formulas. We did not run into any complicated "numerical approximations" or "iterative algorithms." Many ML estimators require more difficult calculation, but it is worth remembering that at least one of them is easy.

When calculating the variance from a sample, many students in introductory statistics courses have been terrorized by the question, "should we divide by N or $N-1$?" Professors usually respond by dissembling and ambiguating, or reciting a poem about "used degrees of freedom."

But I'm here to give a straight answer. If the student wants the ML estimator, the answer is "divide by N."

On the other hand, the ML estimator may not be the one the professor wants. The expected value of the ML estimator of the variance is not equal to the true $Var[x] = \sigma^2$. That is, it can be shown that

$$E[\widehat{\sigma^2}] = \sigma^2 - \frac{\sigma^2}{N} = \frac{N-1}{N}\sigma^2 \tag{45}$$

The ML estimator is just a bit too small. It equals the "true variance" σ^2 minus a fraction that depends on N. An unbiased formula for estimating the variance from a sample is

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{N}(x_i - \hat{\mu})^2}{N-1}. \tag{46}$$

So, the complete answer to the student's question is a question that should be phrased back to the professor. "Would you like an ML estimate or would you like an unbiased estimate?"

8.3 Uniform Distribution

The uniform distribution has already been discussed in Section 5.2 and illustrated in Figures 6 through 8. It is presented here mainly for completeness.

The uniform is a continuous distribution that is defined between two points, $[a, b]$. All points between a and b are equally likely.

8.3.1 Probability Density Function

The probability density of the uniform is "flat". It does not have parameters in the usual sense. Its height is simply determined by its width. If the range is $a = 0$ to $b = 1$, then the height of the density has to be 1.0 in order to guarantee that the total area under the curve is equal to 1.0. On the other hand, if the uniform has to stretch from $a = -10$ to $b = +10$, then the height of the curve must be $1/20 = 0.05$.

If one grasps that simple fact, then it is easy to see the PDF is

$$f(x) = \frac{1}{b-a} \tag{47}$$

8.3.2 Cumulative Distribution Function

The graph of the linkage between the PDF and the CDF has already been presented in Figure 8. The formal representation of the chance of an outcome smaller than k is

$$\begin{aligned} F(k) &= \int_a^k \frac{1}{b-a} dx \\ &= \frac{1}{b-a}(k-a) \end{aligned} \tag{48}$$

8.3.3 Moments

The expected value is exactly in the center of the domain, half way between a and b.

$$E[x] = \frac{1}{2}(a+b)$$

The variance depends on the width of the domain, of course.

$$Var[x] = \frac{1}{12}(b-a)^2$$

8.3.4 Comments

The uniform represents the idea that any outcome is equally likely. It is mainly useful as a theoretical representation of the idea that there is no basis for predicting one value over another. To a Bayesian statistician, a "uniform prior" is used to mean that a person is completely unsure about what to expect. In many game theory models, the uniform distribution is used because it is very easy to work with.

8.4 Gamma Distribution

The gamma distribution is continuous and defined for positive real numbers, $[0, \infty)$. Depending on the values of its parameters, it may be either "ski-slope" shaped or it may be single-peaked, with a more-or-less exaggerated tail on the right. It can be used to represent the density of any variable that is restricted to non-negative values, and is frequently used in models of waiting times and survival.

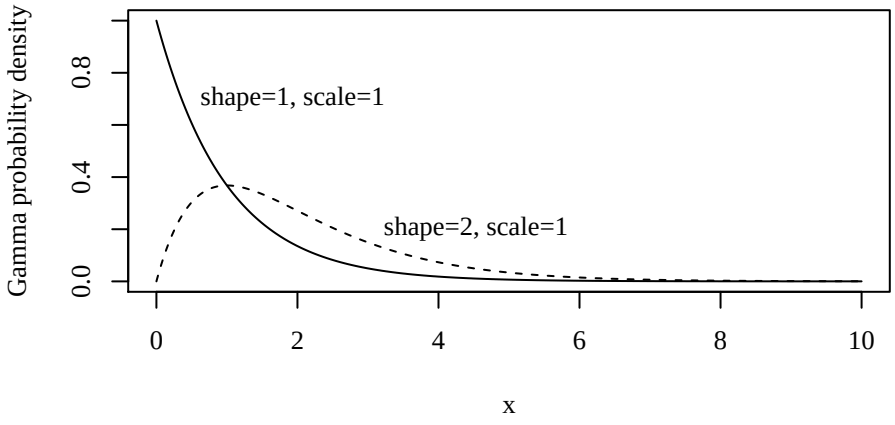


Figure 16: Gamma Density

8.4.1 Probability Density Function

Like the beta and the normal, the gamma distribution is a two parameter distribution. The parameters are often called shape (α) and scale (β). I will refer to it as $Gamma(\alpha, \beta)$. The PDF is

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta}, where\, x \geq 0, \alpha{>}0, \beta{>}0. \tag{49}$$

The symbol $\Gamma(\alpha)$ is a normalizing constant. It is known as the gamma function. It can be thought of as an extension of the factorial function to the real number line. For integers, $\Gamma(\alpha) = (\alpha - 1)!$

How would someone ever think of a horrible formula like that? If I were just making this up, I would reason as follows. Start with the exponential distribution's PDF, in which the essential shape is determined by e^{-x}. That is a bit boring and inflexible; it is always smoothly declining from left to right. To spice that up a bit, multiply by $x^{\alpha-1}$.

$$x^{\alpha-1}e^{-x} \tag{50}$$

If $\alpha = 1$, this just reproduces the exponential (since $x^0 = 1$). However, if $\alpha > 1$, the shape changes. We have a single-peaked function with a mode in the interior of the domain, as illustrated in Figure 17. That is why α is called a "shape" parameter.

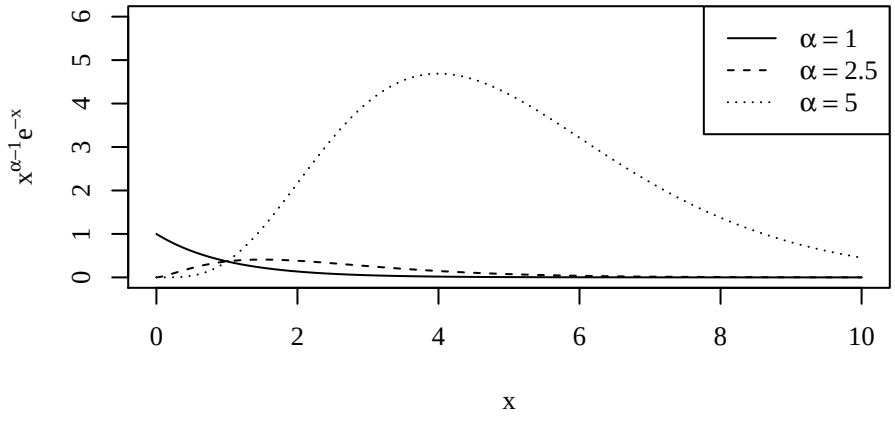


Figure 17: Gamma Kernel (Unscaled)

Expression (50) is the basis for my new probability model, but the area under the curve is not 1.0. A normalizing constant is required. Clearly, it must be

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}. \tag{51}$$

That function, $\Gamma(\alpha)$, is called the gamma function. If we divide (50) by (51), we have a valid pdf. We have almost finished the derivation of (49), except that we have not yet introduced β. However, that is easily remedied. Replace the variable x by a new variable equal to x/β (and employ a change of variables), and we find the result is exactly right.

The gamma distribution seems to "pop up" in many contexts. Recall that the exponential distribution describes the time we have to wait before an event occurs. The gamma can be derived as a model for the amount of time we have to wait that event to repeat itself several times. It makes sense, then, that the gamma distribution with $\alpha = 1$, that is,

$Gamma(1, \beta)$, is identical to the exponential distribution (because we only waited for the event one time). $Gamma(2, \beta)$ would represent the time we wait for two events, and so forth. This interpretation can be used to derive the gamma's pdf, but we have to be somewhat cautious about the interpretation. The shape parameter α can take on any real values greater than 0, it is not limited to integers like 1, 2, and so forth. The interpretation of those non-integer α values is not facilitated by the "waiting for events" interpretation.

8.4.2 Cumulative Distribution Function

The CDF represents the chance of a score lower than k. As one might expect from the functional form, this integral is difficult to solve:

$$F(k; \alpha, \beta) = \int_0^k \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx. \tag{52}$$

We can take note of the fact that $\Gamma(\alpha)\beta^\alpha$ does not depend on x to rewrite this as

$$F(k; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^k x^{\alpha-1} e^{-x/\beta} dx. \tag{53}$$

As far as I know, it cannot be further simplified, which means that in order to calculate $F(k; \alpha, \beta)$, it is necessarily to numerically approximate the area under the curve described by the integral.

8.4.3 Moments

The expected value of the gamma is the product of its two parameters.

$$E[x_i] = \alpha \cdot \beta \tag{54}$$

This indicates that one can shift the distribution to the right by increasing either the shape or the scale parameter.

The variance of the gamma responds to both parameters as well, but it is more sensitive to the scale parameter.

$$Var[x_i] = \alpha \cdot \beta^2 \tag{55}$$

As one should expect from the graph of the gamma distribution, the expected value does not coincide with the mode. If $\alpha > 1$, then the distribution has a single-peaked appearance in which the most likely outcome, the mode, is

$$mode = \beta(\alpha - 1) \tag{56}$$

8.4.4 Comments

The gamma distribution is important partly because it is flexible enough to describe a variety of possible beliefs about the chances of outcomes on the positive real numbers. The gamma distribution was used to summarize my beliefs about the likely number of break-outs by my neighbor's dog in Figure 4.

Another important point is that the gamma distribution is centrally located in the family of distributions. Other distributions can seen as special cases of the gamma. For example, the $\chi^2(\nu)$ distribution (which is discussed below) has the same PDF as $Gamma(\frac{\nu}{2}, 2)$. If $\alpha = 1$, the gamma simplifies into an exponential distribution (20).

The gamma distribution has special properties that are inherited by all of its special cases. One of the most intriguing facts about the gamma distribution is the additivity property. The sum of observations from gamma distributions with various α_i, but the same scale (β), is distributed as $Gamma(\alpha_1 + \ldots + \alpha_n, \beta)$.

The gamma's probability density function is easy to "re-parameterize" for particular projects. In particular, we can fiddle around with the scale parameter to achieve various desired effects. In regression models of "count" data, sometimes we need to add "noise" in the form of a positive random variable that has a fixed expected value but an adjustable amount of variance. For that purpose, set the gamma scale parameter to be $1/\alpha$. The expected value will be

$$E[x] = \alpha \cdot \frac{1}{\alpha} = 1. \tag{57}$$

As we adjust α, the expected value stays fixed at 1, but the variance is still sensitive.

$$Var[x] = \alpha(\frac{1}{\alpha})^2 = \frac{1}{\alpha}. \tag{58}$$

Three probability density curves are plotted for the special case in which $\beta = 1/\alpha$ in Figure 18.

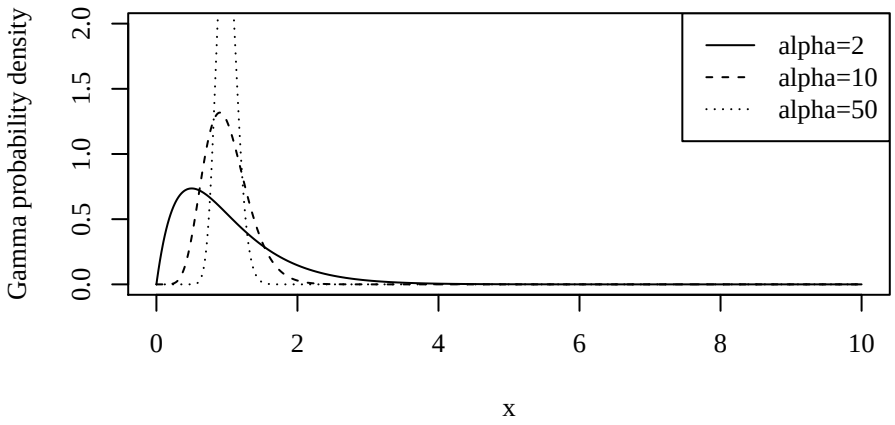


Figure 18: Gamma Density when $\beta = 1/\alpha$

From sample data, the gamma's parameters can be estimated in a variety of ways. The maximum likelihood estimate can be derived by iterative approximation, but there is no "closed form" solution from which to calculate estimates of α and β. The "method of moments" can be used to get a quick "first take" on parameter estimates. The method proceeds as follows. First, calculate the sample mean and variance. Let's call these $\widehat{E[x]}$ and $\widehat{Var[x]}$. Second, use those values in place of $E[x]$ and $Var[x]$ in expressions (54) and (55).

$$\widehat{E[x]} = \hat{\alpha} \cdot \hat{\beta} \tag{59}$$

and

$$\widehat{Var[x]} = \hat{\alpha} \cdot \hat{\beta}^2 \tag{60}$$

This is called the "method of moments" because we have proceeded as though the sample mean and variance are actually equal to the theoretical moments. Those expressions can be rearranged so that

$$\hat{\alpha} = \frac{\widehat{E[x]}^2}{\widehat{Var[x]}} \tag{61}$$

$$\hat{\beta} = \frac{\widehat{Var[x]}}{\widehat{E[x]}} \tag{62}$$

8.5 Beta Distribution

The beta distribution, $Beta(\alpha, \beta)$, has two parameters which jointly define its shape. Like the uniform distribution, the beta distribution is defined on a closed interval. For simplicity,

we consider only the version that is defined on the domain $[0, 1]$.

The beta's parameters can be adjusted to dramatically change its appearance. It can be single peaked, skewed, or two peaked. Three example beta distributions are presented in Figure 19.

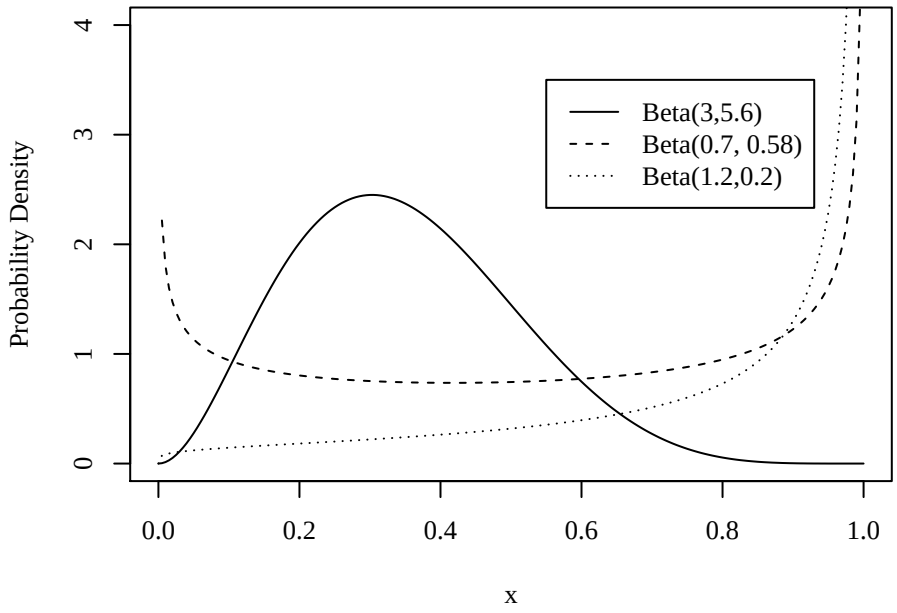


Figure 19: Beta Density Functions

8.5.1 Probability Density Function

The standard *Beta*'s pdf is defined on $[0, 1]$:

$$f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \; where \, x \in [0, 1] \tag{63}$$

and the normalizing constant is called the beta function

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt \tag{64}$$

Incidentally, the beta function is equal to a ratio of gamma functions,

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{65}$$

and the fraction formed by two gamma variables that have the same scale parameter, $x_1/(x_1 + x_2)$, is distributed as a beta variable.

8.5.2 Cumulative Distribution Function

The chance that a draw from a beta density is less than k is

$$F(k;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \int_0^k x^{\alpha-1}(1-x)^{\beta-1} dx \tag{66}$$

Sometimes the part on the right (the integral) is called the incomplete beta function, but as far as I can see, no simplifying analytical benefit is had by re-labeling it. Generally, it can only be calculated by numerical integration.

8.5.3 Moments

The expected value of a variable that is beta distributed is:

$$E[x] = \mu = \frac{\alpha}{\alpha+\beta} \tag{67}$$

and the variance is

$$Var[x] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{68}$$

If $\alpha > 1$and $\beta > 1$, the peak of the density is in the interior of [0,1]. In that case, the mode of the *Beta* distribution is

$$mode = \gamma = \frac{\alpha-1}{\alpha+\beta-2} \tag{69}$$

If α or $\beta < 1$, the mode may be at an edge.
If $\alpha = \beta = 1$, then the beta is identical to a uniform distribution.

8.5.4 Comments

The two parameters that determine the shape and nature of a beta distribution are not so intuitively meaningful as are the parameters for the normal.

The beta distribution fits into a larger mosaic of probability models, but in my experience it has two especially important uses. First, it can summarize our beliefs about how likely something is to "be true" or "to occur." Since the beta's formula is so flexible, it can describe virtually any shape that we might subjectively impose in a model.

Second, the beta can be used as an output variable in a regression modeling exercise. If a dependent variable is a proportion, then it may naturally be interpreted as a draw from a beta distribution. Predictors are used in an effort to account for the fact that the observed proportion is high for some units and low for others.

8.6 Chi-Squared

The Chi-Squared distribution depends on only one parameter, which is I will refer to by the Greek letter ν (pronounced "nu"). The Chi-Squared may be referred to as $\chi^2(\nu)$ or χ^2_ν . The χ^2 represents the probability of a variable that is defined on the interval $[0,\infty)$.

The Chi-Squared distribution is used to describe the "sum of squared mistakes" or "mismatches" between expectation and observation. If that sum is very small–close to 0–it means the cumulative mismatch between expectations inspired by a model is small. In all kinds of regression modeling, we often need to decide if one model is "closer" to the data than another. The difference between the models usually boils down to a Chi-Squared statistic.

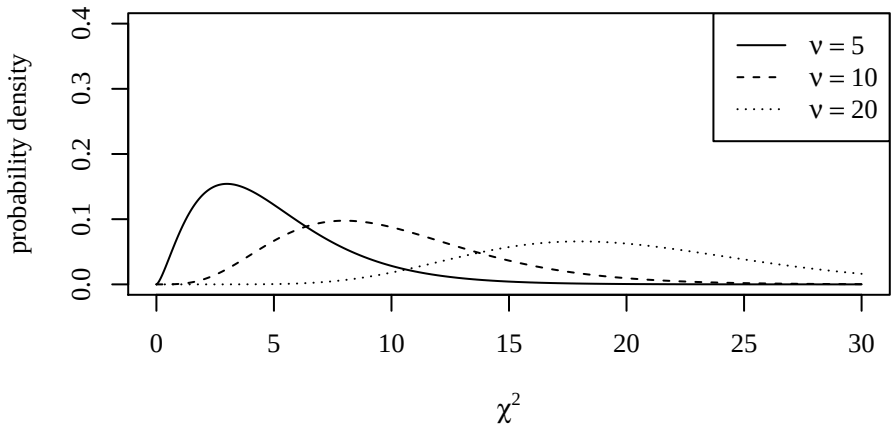


Figure 20: Density Function of χ^2

As one can see, as ν grows larger, the concentration of density shifts to the right and becomes more symmetric.

8.6.1 Probability Density Function

In the discussion of the gamma distribution, it was already mentioned that the pdf of a Chi-square distribution is identical to a gamma distribution with shape parameter $\nu/2$ and scale 2. Filling in the blanks in expression (49), the Chi-square's probability density function is

$$f(x) = \frac{1}{\Gamma(\frac{\nu}{2})(2)^{\frac{\nu}{2}}} x^{\frac{\nu}{2}-1} e^{-x/2},\ x \geq 0,\ \nu > 0. \tag{70}$$

This seems somewhat anti-climactic, but the story does not end there. The best is yet to come. Here it is in a nutshell: sum of squared random variables follows a Chi-square distribution.

Here is the big idea. Draw a collection of ν observations from a standard normal distribution,

$$Z_i \sim N(0,1),\ for\, i = 1, 2, \ldots, \nu. \tag{71}$$

Square each one, and add them together. The result is distributed as a $\chi^2(\nu)$. That is to say

$$Z_1^2 + \ldots + Z_\nu^2 \sim \chi^2(v). \tag{72}$$

When it is used in this context, the parameter that represents sample size, ν, is often called "degrees of freedom."

8.6.2 Cumulative Distribution Function

The cumulative distribution of a $\chi^2(\nu)$ variable does not reduce to a simple formula, in the same way that the cdf of $Gamma(\alpha, \beta)$ is not simple. Nevertheless, it is very important that statisticians have found numerical methods to approximate the cumulative densities of the χ^2.

8.6.3 Moments

Since the $\chi^2(\nu)$ is the same as $Gamma(\frac{\nu}{2}, 2)$, so we can use the formulas (54) and (55).

$$E[x] = \nu \tag{73}$$

$$Var[x] = 2\nu \tag{74}$$

8.6.4 Comments

Many statistical procedures can result in a estimate that is distributed as $\chi^2(\nu)$. The mismatch between the saturated model and the fitted generalized linear model, for example, is distributed as a χ^2. The squared mismatch between the observed and predicted counts in a cross tabulation table is also distributed as a χ^2.

When we calculate some estimate from a sample, say we call it $\hat{k}$, it is important for us to find out if that estimate is "in the middle of the usual range" or if it is in an extreme tail of the possibilities. If we can calculate the proportion of cases smaller than $\hat{k}$, $F(\hat{k}; \nu)$, then it is obvious we can calculate the proportion of cases that is greater than $\hat{k}$, $1 - F(\hat{k}; \nu)$. That area is represented in the following figure, in which the pdf of $\chi^2(50)$ is drawn. The shaded area on the right–values greater than 67.50–represents the top 5% of possible draws from $\chi^2(50)$.

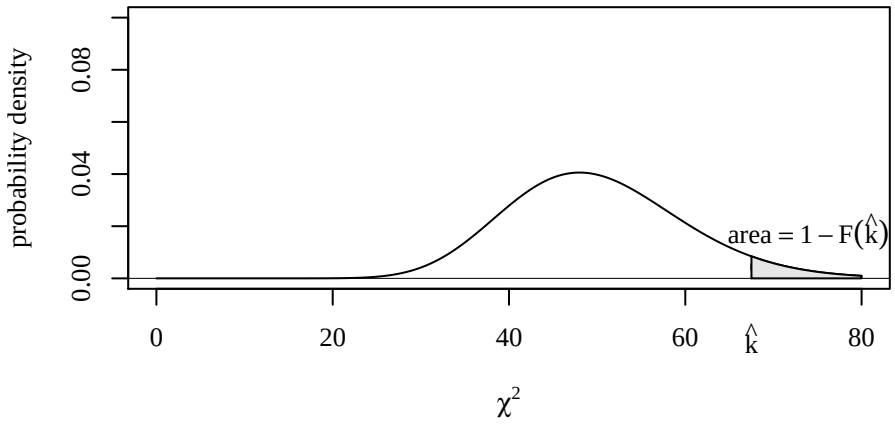


Figure 21: Extreme Values of $\chi^2(50)$

8.7 Student's t distribution

One of the most interesting stories in the folklore of statistics is that an employee of the Guinness beer company discovered this distribution, but his employer would not allow him to publish it under his own name. When William S. Gosset published "The Probable Error of A Mean" in 1908, he elected to use the pen name "Student." The finding was not immediately recognized for its value, but the famous statistician R.A. Fisher popularized Student's t distribution and made it a cornerstone in his system of hypothesis testing.

The t distribution is symmetric and uni-modal. It has one parameter, ν. Its center point–mean,median, and mode–is always at $x = 0$. It is similar to the normal distribution. Extreme outcomes are more likely in the t distribution. Statisticians say that t has "fatter tails" the normal.

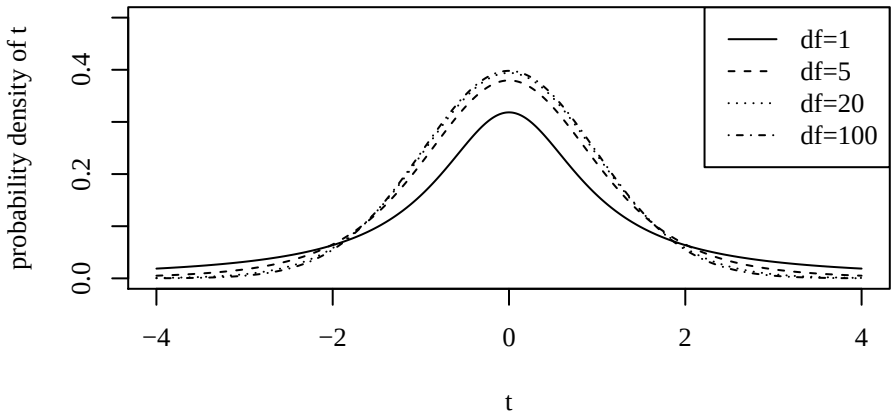


Figure 22: t Densities

8.7.1 Probability Density Function

The probability density of the t distribution is

$$f(x;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-(\frac{v+1}{2})}. \tag{75}$$

Even if we ignore the normalizing constant at the front, we are still left with a formidable expression. Does it help to write this as:

$$f(x;\nu) \propto \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\frac{v+1}{2})}}? \tag{76}$$

As x grows larger, the denominator grows larger and so the density gets smaller.

The t distribution was developed to help deal with the following problem. Suppose we collect a sample of data and from it we calculate estimates of the mean and the variance (and its square root, the standard deviation). We want to know if the observed mean is in the "middle range" of what we expect (close to the expected value) or if it is extreme. If we think of this as if it were a Z statistic (see expression (29)), it seems as though we ought to be able to make a comparison, something like

$$\frac{estimated\,mean - null\,hypothesis}{standard\,deviation\,of\,mean}. \tag{77}$$

What distribution would that estimate have? Gosset suggested that the result is distributed according to his t distribution.

At this point, I need to explain how to calculate the standard deviation of the mean. Please remember I'm using the symbol $\widehat{E[x]}$ to refer to a sample mean (not the usual symbol

$\bar{x}$) and the sample variance is $\widehat{Var[x]}$. The variance of the estimated mean across samples is much smaller than the variance itself. In fact,

$$Var[\widehat{E[x]}] = \frac{1}{N} Var(x) \tag{78}$$

Here is how that is derived. Collect a sample and calculate the average,

$$\widehat{E[x]} = \frac{x_1 + x_2 + \ldots + x_N}{N}. \tag{79}$$

Apply the variance operator to both sides.

$$\begin{aligned} Var[\widehat{E[x]}] &= Var\left(\frac{x_1 + x_2 + \ldots + x_N}{N}\right) \\ &= \frac{1}{N^2}\left(Var(x_1) + Var(x_2) + \ldots + Var(x_N)\right) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} Var(x_i) = \frac{1}{N} Var(x) \end{aligned} \tag{80}$$

Thus, the true variance (and its square root, the standard deviation) of an estimated mean are known, as long as the true variance of x itself is known.

In practice, the true variance of x is not known, and thus we are wrestling with the fact that both the mean and the variance must be estimated from the same sample. What if we could proceed *as if* the estimated standard deviation of the mean were actually correct? Gosset charted out a plan to do just that. In his own words, he sought to find the "standard deviation of the standard deviation" across samples so as to appreciate the ratio of the estimate to its estimated standard deviation.

Today, we think of the problem like this. A standard normal variable could be created if we knew the true variance, as in

$$\frac{\widehat{E[x]} - E[x]}{\sqrt{Var[x]/N}} \sim N(0,1) \tag{81}$$

Don't worry that $Var[x]$ is unknown, we will find a way to cancel it out. The ratio of the estimated variance to true variance is proportional to a χ^2 variable with $\nu = N$.

$$\frac{\widehat{Var[x]}}{Var[x]} \sim \frac{1}{N}\chi^2(N) \tag{82}$$

Divide (81) by the square root of (82),

$$\frac{\widehat{E[x]} - E[x]}{\sqrt{N}\sqrt{Var[x]}} \div \sqrt{\frac{\widehat{Var[x]}}{Var[x]}} = \frac{\widehat{E[x]} - E[x]}{\sqrt{N}\sqrt{\widehat{Var[x]}}} = \frac{\widehat{E[x]} - E[x]}{\sqrt{N}\widehat{StdDev}[x]}. \tag{83}$$

The unknown $Var[x]$ disappears, and we are left with exactly the result we were looking for. It looks like a Z statistic, but we can use an estimate of the variance. We call the denominator, $\sqrt{N}StdDev[x]$, the "standard error of the mean" because it is an estimate of the standard deviation of the mean (not the true standard deviation of the mean).

When a sample is large, then the t ratio described in expression (85) and the standard normal (29) are not noticeably different. As illustrated in Figure 23, as ν is increased, the $t(\nu)$ converges to the standard normal distribution.

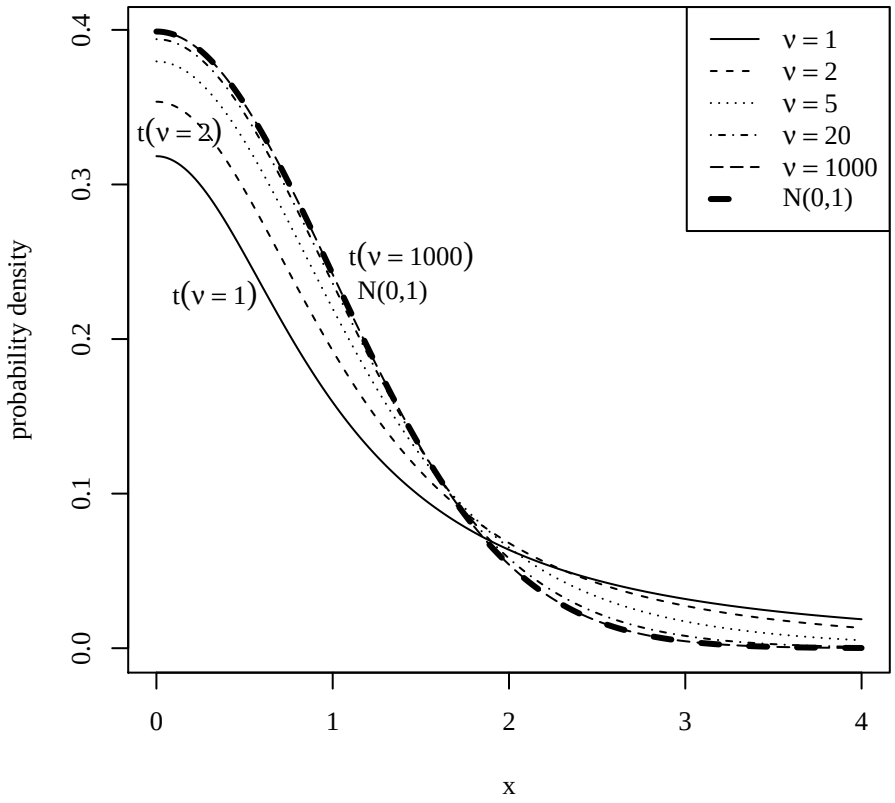


Figure 23: Normal(0,1) and $t(1000)$ Coincide

8.7.2 Cumulative Distribution Function

Like most of the other distributions that have been discussed here, there is no simple closed form with which to calculate the cdf of a t statistic. Because the cumulative probability of a t is difficult to calculate, statistics books have historically included a table against which test values can be compared.

One complication worth mentioning about the cdf is that the t distribution is usually thought of as a two-tailed distribution. That is, the sample estimate $\widehat{E[x]}$ may be grossly wrong on the low side, or on the high side. Unlike the χ^2 distribution, pictured in Figure 21,

where we look only on the right tail of the distribution for evidence of unusual cases, the t distribution has critical regions both tails. Consider Figure 24.

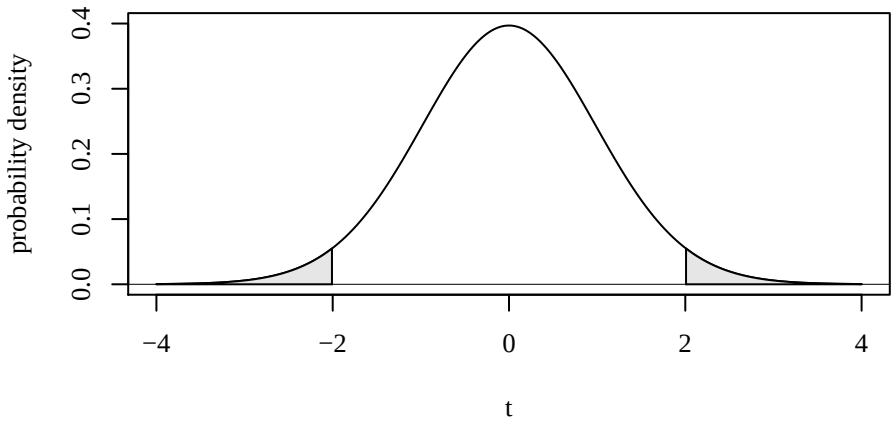


Figure 24: Extreme Values of $t(50)$

8.7.3 Moments

Supposing $\nu \geq 1$, the expected value, median, and mode of a t distribution are all 0. The variance of a t distribution is

$$Var[x] = \frac{\nu}{\nu - 2} \tag{84}$$

It is worth noting that as $\nu \to \infty$, $Var[x] \to 1.0$, consistent with the claim that the t density converges to $N(0, 1)$.

8.7.4 Comments

The t distribution is thus a handy way to find out if the average from a sample is out of line with expectations. That's important, but not so hugely important as the t distribution would become. When he popularized Student's t distribution, R.A. Fisher proposed the t as a distribution for analysis of a much larger class of problems. Basically, any problem in which the sample-based estimate is normally distributed may be compared against the t distribution, as long as we can find a standard error to use in the denominator. The term "t ratio" refers generally to the comparison of any estimator, $\hat{\theta}$, for a parameter θ, against its standard error.

$$\frac{\hat{\theta} - E[\theta]}{standard\,error(\hat{\theta})} \sim t(\nu). \tag{85}$$

There is usually some work to do when deciding what the ν parameter should be, but it is almost always $N - something$, and in most common situations, it is considered a solved problem.

8.8 The F distribution

The $F(\nu_1, \nu_2)$ distribution ("F" is for Fisher) describes a variable on $[0, \infty)$. It depends on 2 parameters, ν_1 and ν_2.

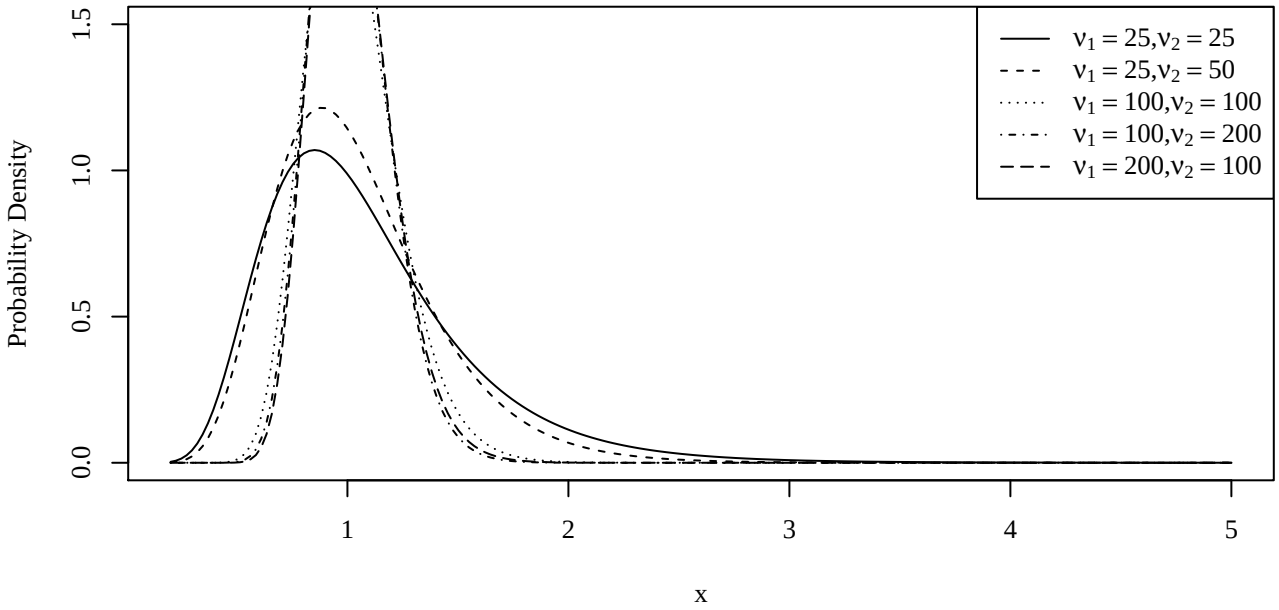


Figure 25: Density of $F(\nu_1, \nu_2)$

8.8.1 Probability Density Function

$$f(x; \nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \nu_1^{\nu_1/2} \nu_2^{\nu_2/2} x^{\frac{\nu_1}{2}-1} \left(\nu_2 + \nu_1 x\right)^{-(\nu_1+\nu_2)/2} \tag{86}$$

which may be re-arranged as

$$f(x; \nu_1, \nu_2) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{\frac{\nu_1}{2}-1} \left(1 + \frac{\nu_1}{\nu_2} x\right)^{-(\nu_1+\nu_2)/2} \tag{87}$$

I've tried to find a way to simplify that so that it would carry some intuition, but I have failed entirely. Perhaps I can make the effort worthwhile by explaining why that distribution has the shape that it does.

Here is where the F distribution comes from. Suppose we begin with the goal of comparing two samples of observations. We already know that $Z_1^2 + Z_2^2 + Z_{\nu_1}^2$ is distributed as a $\chi^2(\nu_1)$. How should we compare that against a second set of observations, one for which the sum of squares is $\chi^2(v_2)$? So far as I know, there is no known method to compare the difference of two χ^2 statistics, but it is possible to compare their ratio. If one sample size, say ν_1, is

significantly larger than the other one, then it seems obvious that its sum of squares will be larger, even if the cases are not more widely dispersed. In order to bring two sums of squares into a comparable state, we must divide the χ^2 distributed sum by the number of scores. The test statistic we want to understand is thus a ratio of "mean squares":

$$\frac{Sample\,1: \quad (Z_1^2 + Z_2^2 + \ldots + Z_{\nu_1}^2)/\nu_1}{Sample\,2: \quad (Z_1^2 + Z_2^2 + \ldots + Z_{\nu_2}^2)/\nu_2}. \tag{88}$$

The pdf of $F(\nu_1, \nu_2)$ represents the diversity we would observe if we repeatedly drew ν_1 and ν_2 observations and then formed this ratio of mean squares. If the two samples are indeed drawn from a standard normal distribution, then we expect that value will be approximately 1.0 with some variation above and below.

The density associated with example values $\nu_1 = \nu_2 = \nu$ is presented in Figure 26. If $\nu_1 = 1$, then the density of F is the same as that of squared t variable.

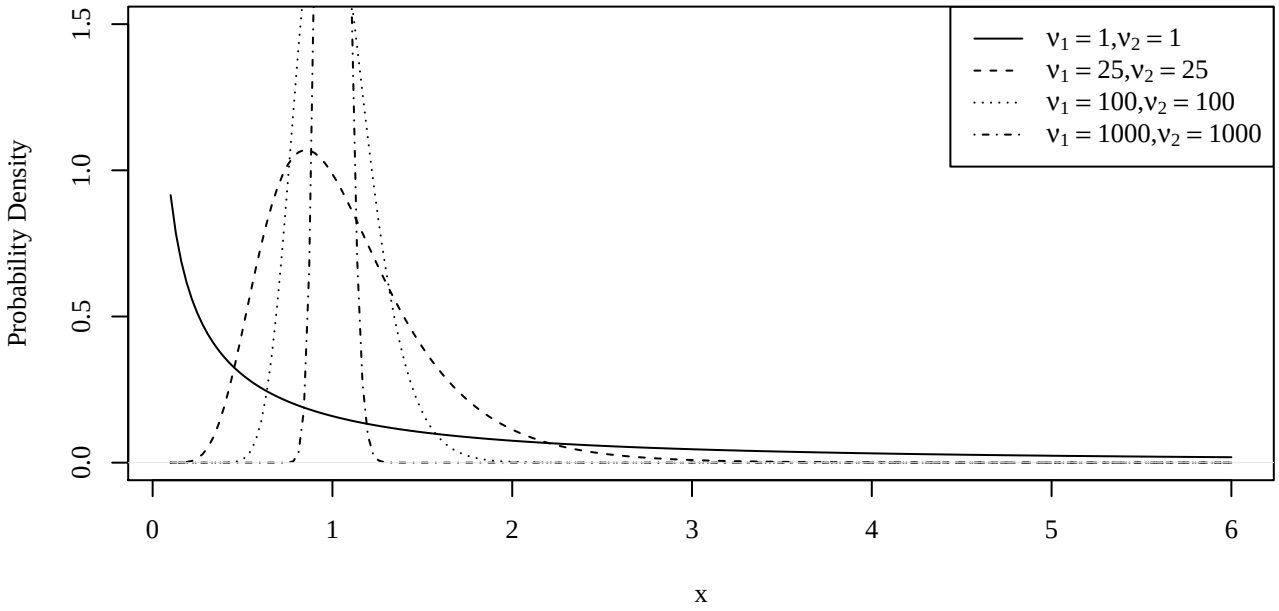


Figure 26: Density of $F(\nu_1, \nu_2)$ when $\nu_1 = \nu_2$

8.8.2 Comments

Suppose we collect samples of data from ν_1 men and ν_2 women. We'd like to know if the diversity of responses from men is greater than that of women. For each group, we calculate the "mean squares" (the estimates of the variance) and compare them. Obviously, if the ratio is 1.0, there is no question, the two are about the same. But what if the ratio is 1.5? Is a ratio so large that we would think it is inconsistent with the idea that the variances among men and women are the same? That is the sort of test for which the F works well.

The t, the χ^2, and the F are a power trio in hypothesis testing. They are, by far, the three most frequently used distributions. The t distribution represents a ratio of an estimate to its standard error. The χ^2 summarizes the distribution of a sum of squares. The F distribution can be used to analyze the *ratio* of two sums of squares. If the observed ratio is large, it means that one sum of squares is substantially larger than another. The F test compares the mismatches of 2 models and offers one way to decide if one "fits worse" than another.

8.9 Binomial Distribution

The binomial distribution, $B(N, p)$, represents the number of "events" (or "successes", or "wins", etc.) that occur when there are N "trials" (opportunities for an event, success, wins, etc.) and the chance of a success on each trial is fixed at p.

I have a special coin that returns a "head" two-thirds of the time. What is the probability that I will get any given number of heads after flipping 10 times? The binomial distribution gives the answer. Two depictions of the result are presented in Figure 27. The plot on the left highlights the fact that the outcomes are discrete steps, not real-valued outcomes. However, I fancy the plot on the right because it fits more closely together with the continuous distributions that we have studied so far.

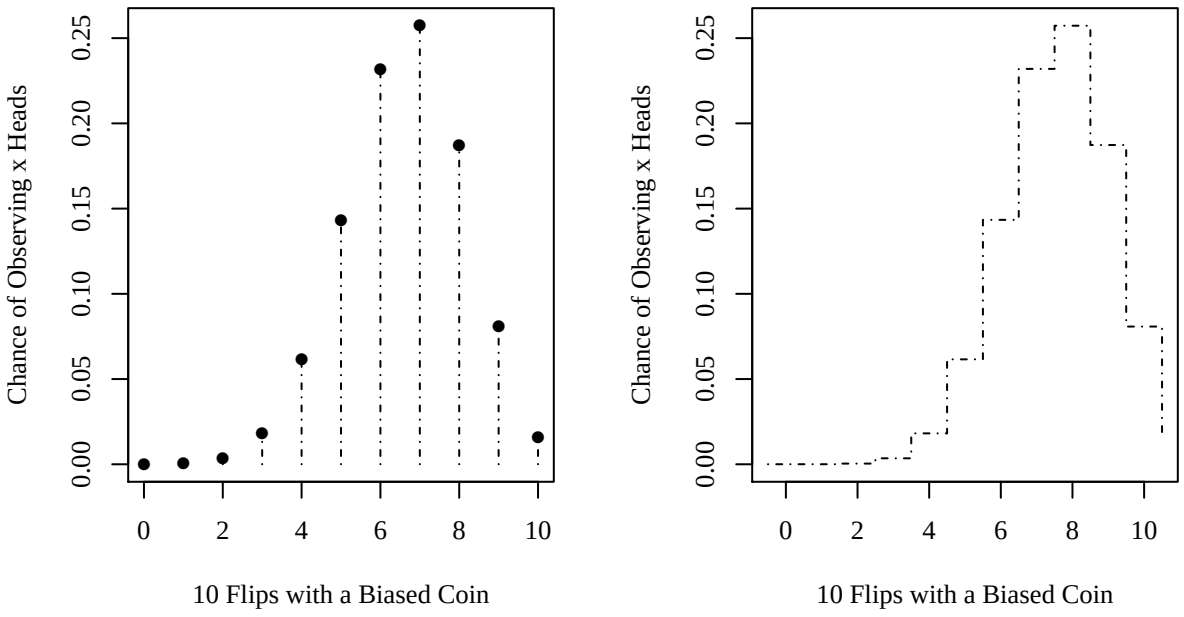


Figure 27: 50 Coin Flips

8.9.1 Probability Mass Function

The Binomial probability mass function is:

$$Prob(k|N,\pi) = \frac{N!}{(N-k)!k!}\pi^k(1-\pi)^{N-k} \tag{89}$$

It is pretty easy to derive this distribution. If there are N independent trials, and we wonder how likely we are to get k successes. The chance that the first k trials will succeed, and the rest will fail, is

$$\begin{aligned}\pi \times \pi \times \{k\,times\} \times (1-\pi) \times (1-\pi) \times \{N-k\,times\} \\ = \pi^k(1-\pi)^{N-k}\end{aligned}$$

That accounts for the second part of the binomial formula, but this is not quite done. There are many other ways to get k successes, and so we have to count all of the possible sequences. That's where the prefix comes from. It is the binomial coefficient. $\frac{N!}{(N-k)!k!}$ is the number of ways to re-arrange N things so that k are successes and $N-k$ are not.

When N is large, the binomial distribution is quite similar to a normal distribution. Lets consider an example. Suppose the chance of having a boy baby is 0.63 for all women in a community. If 437 women have babies, what is the probability that there will be 200 boys?

Inserting N and π into the previous expression, the chance of k successes is seen to be:

$$Prob(k|437,0.63) = \frac{437!}{(437-k)!k!}(0.63)^k(1-\pi)^{437-k} \tag{90}$$

If we had asked for the probability of 300 boys, we would find:

$$P(300|437,0.63) = 0.0001122501 \tag{91}$$

I've done some "hunting and pecking" with this distribution to find out which values of k are most likely. The outcomes with noticeable chances are between 240 and 310, as indicated in Figure 28. There is a mathematical proof of the fact that as N tends to infinity, the discrete probabilities of the binomial are very accurately approximated by a normal distribution. Of course, as is evident in Figure 27, that approximation will not work when N is small.

8.9.2 Cumulative Distribution

One of the big problems with analysis of continuous distributions is that the cumulative distribution function cannot be simplified. Numerical approximation is required, and there are known problems (and solutions) for that. When a distribution is discrete, no approximation is required. We simply need to calculate a sum.

8.9.3 Moments

The expected value is:

$$E[x] = \pi \cdot N \tag{92}$$

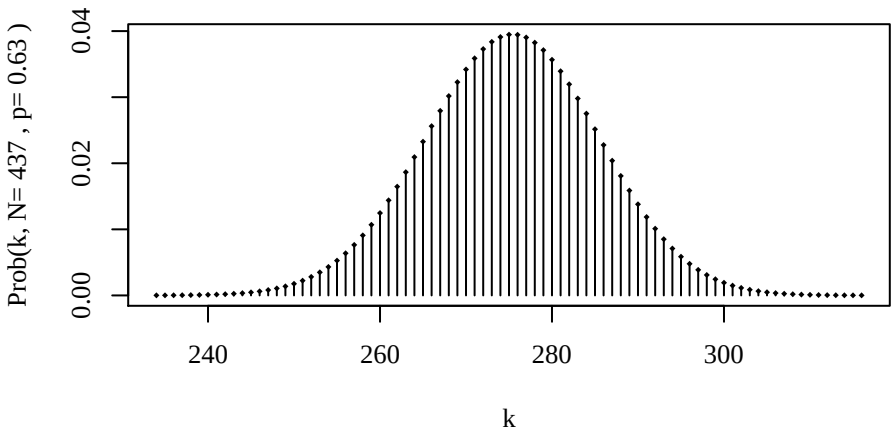


Figure 28: Binomial with N=437 and p=0.63

and the variance is

$$Var[x] = \pi(1-\pi)N \tag{93}$$

It seems obvious to me that the expected value this is correct. If we flip a coin 10 times and the chance of a "head" is π, it seems reasonable to expect $\pi \cdot 10$ heads.

There is a simple way to demonstrate that. Think of the outcome, the number of successes, as a sum of 0's and 1's. For instance, the observed sample:

$$0, 1, 1, 0, 1, 1, 0, 0 \ldots, 1, 0 \tag{94}$$

is really just a realization of Bernoulli trials, and the number of successes is just the sum of those trials, as in

$$x_1 + x_2 + x_3 + \ldots + x_{N-1} + x_N \tag{95}$$

Those are "statistically independent" samples of size 1, and each one has probability of success equal to π. So, considering just one "event" in isolation, the chance is π of observing a 1 and $(1-\pi)$ chance of observing 0. So the expected value of that one draw is

$$E[x_1] = \pi \cdot 1 + (1-\pi) \cdot 0 = \pi \tag{96}$$

So of you think of the Binomial as the sum of N of those experiments,

$$\begin{aligned} E[x_1 + x_2 + \ldots x_N] &= E[x_1] + E[x_2] + \ldots + E[x_n] \\ &= \pi + \pi + \ldots + \pi \\ &= N \cdot \pi \end{aligned} \tag{97}$$

The variance can be derived similarly. Consider just one draw, x_1, in isolation. Its variance is

$$\begin{aligned} Var[x_1] &= \pi(1 - E[x_1])^2 + (1 - \pi)(0 - E[x_1])^2 \\ &= \pi(1 - \pi)^2 - (1 - \pi)(-\pi)^2 \\ &= \pi(1 - 2\pi + \pi^2) + \pi^2 - \pi^3 \\ &= \pi - 2\pi^2 + \pi^3 + \pi^2 - \pi^3 \\ &= \pi - \pi^2 = \pi(1 - \pi) \end{aligned} \tag{98}$$

The Binomial distribution is a sum of N of those variables, and they are all statistically independent of each other. Thus, the law for calculating the variance of a sum of terms applies.

$$\begin{aligned} Var[x_1 + x_2 + \ldots x_N] &= Var[x_1] + Var[x_2] + \ldots + Var[x_N] \\ &= \pi(1 - \pi) + \pi(1 - \pi) + \ldots + \pi(1 - \pi) \\ &= \pi(1 - \pi)N \end{aligned} \tag{99}$$

8.9.4 Comments

In statistical modeling research, the most common use of the binomial distribution is in regression with categorical "Yes" or "No" outcomes. These may be "logistic" or "probit" regression models. Suppose we group observations into 5 categories. For each category, we collect data on trial conditions (these will act as predictors) as well as the outcomes, Y_i successes out of N_i trials (and thus the observed success rate is Y_i/N_i). Using the predictors (or whatever other information we have handy), we calculate predicted success rates for the 5 groups, $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$. For each group, it is necessary to calculate how likely we were to observe Y_i successes, and the sum of those probabilities is the likelihood function. We would like to adjust our predictive approach so as to maximize the likelihood, of course.

8.10 Poisson Distribution

The Poisson is a discrete distribution with outcomes in the set $0, 1, \ldots, \infty$. It is commonly used to describe "event counts." A Poisson distribution was used to generate Figure 2.

8.10.1 Probability Mass Function

The Poisson has a single parameter, which is customarily known as λ. The probability that there are x occurrences in a time-interval is:

$$f(x; \lambda) = e^{-\lambda}\frac{\lambda^x}{x!}, \, where \, x \geq 0, \lambda > 0 \tag{100}$$

where e is Euler's constant and $x!$ is the factorial of x ($x! = x \cdot (x-1) \cdot \cdots 2 \cdot 1$). This probability model can be derived in several ways. The famous French mathematician Simeon Poisson proposed this model in the mid 1800s, reasoning as follows. Begin with the idea of

time passing in "small chunks," Δt. Suppose the chance of one event during that time is approximately $\lambda \cdot \Delta t$ (and, as Δt shrinks to 0, $\lambda \Delta t$ approximates the chance of an event more and more closely). Assume further that the chance of a second event in a particular chunk of time is vanishingly small. Then the analysis of some differential equations results in the formula proposed in (100). Another derivation begins with a binomial distribution, $B(N, \pi)$. Let $\lambda = \pi N$, so $\pi = N/\lambda$. Insert that into the binomial probability mass function, and let N grow arbitrary large and let p grow smaller. Several limits must be calculated, and one finds that when p is not very large, the Poisson pmf very closely approximates $B(\pi, N)$ as $N \to \infty$. Hence, when it is difficult to calculate $B(N, \pi)$ because N is large, the Poisson model can serve as a reasonable approximation.

The term $e^{-\lambda}$ (same as $1/e^{\lambda}$) is a normalizing constant. The kernel of this probability model is simply

$$\frac{\lambda^x}{x!} \tag{101}$$

The values that this implies are presented in Table 1.

Table 1: Kernel of Poisson Probability Model

x		0	1	2	3	4	5	...		∞
$\lambda^x/x!$		1	λ^1	$\lambda^2/2!$	$\lambda^3/3!$	$\lambda^4/4!$	$\lambda^5/5!$			$\lambda^\infty/\infty!$

The sum of the items in the second row is

$$exp(\lambda) = 1 + \lambda + \lambda^2/2! + \lambda^3/3! + \lambda^4/4! + \lambda^5/5! + \ldots + \lambda^\infty/\infty! \tag{102}$$

The right side is equal to $exp(\lambda)$ because that is the definition of the exp function. To verify that, differentiate the sum in expression (102) and notice the result is exactly the same sum. That's a good piece of evidence that it really is $exp(\lambda)$.

It is obvious that the values depend on whether λ is greater than 1. If λ is less than one, then the most likely outcome is always 0 and higher values are progressively less likely. On the other hand, if λ is greater than 1, then the story is quite different. There is a "race" between the numerator, which is growing rapidly, and the denominator, which will always win out in the end, but will trail in the early stages.

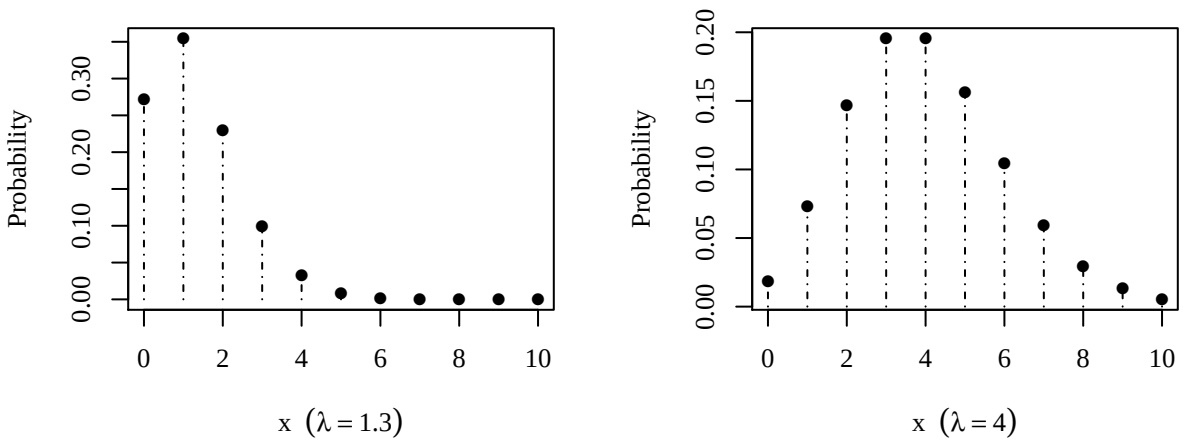


Figure 29: Poisson Mass Function with $\lambda = 1.3$ and 4.0

8.10.2 Moments

The expected value is equal to its variance, and both of them are equal to λ.

$$E(x) = \lambda$$

$$Var(x) = \lambda$$

My first inclination was to avoid proving this result because I thought it required the use of moment generating functions from mathematical statistics. However, a colleague demonstrated a more simple method of deriving the expected value. The sum of strictly positive outcomes

$$E[x] = \sum_{i=1}^{\infty} x_i \cdot e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}, \tag{103}$$

Shifting the index variable from $i = 0$ to $i = 1$ causes the value of x_i to become $(x_i + 1)$.

$$E[x] = \sum_{i=0}^{\infty} (x_i + 1) \cdot e^{-\lambda} \frac{\lambda^{(x_i+1)}}{(x_i + 1)!} \tag{104}$$

Remove the common factor λ from the summation and cancel like terms in the numerator and denominator:

$$E[x] = \lambda \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}. \tag{105}$$

Note that the sum must be 1.0, because we are adding up all of the probabilities from 0 to ∞.

$$E[x] = \lambda \times 1 = \lambda \tag{106}$$

The proof that the variance is also λ follows from the same type of argument in which we calculate $E[x^2]$.

8.10.3 Comments

Poisson regression became something of a fad in the 1990s. Many variables were counts, and a model predicting a mean for each case (λ_i) is relatively easy to construct. By far, the most commonly used predictive model is the exponential form,

$$\lambda_i = exp(\beta_0 + \beta_1 z_i) \tag{107}$$

where z_i is a predictor (independent variable). The Poisson regression falls into the family of generalized linear models, and there are standard routines for calculating estimates of the parameters β_0 and β_1 for any of the models in that class.

In many data sets, the application of the Poisson model will be somewhat wanting because the data will exhibit more dispersion than the model predicts. In that case, it is now common to revise the model to include a multiplicative random error that is drawn from a $Gamma(\alpha, 1/\alpha)$ distribution. Call that gamma variable ε_i and the new model becomes

$$\lambda_i = \varepsilon_i \times exp(\beta_o + \beta_1 z_i) \tag{108}$$

That generates a model with the same expected value (since $E[\varepsilon_i] = 1$), but a higher amount of variance. The primary reason for using that particular type of random variable is that the combined effect of mixing in ε_i and then drawing from $Poisson(\lambda_i)$ is known to be a variable that has the so-called negative binomial distribution. That is distribution has the same expected value, λ_i but a larger variance.

8.11 If I Had an Infinite Amount of Time and Space

When I started drafting this essay, I believed there were 10 distributions that most applied researchers would have in mind on a day-to-day basis. I've discussed 10 distributions, and I think I would like to revise my estimate to 14 or 15. If I had an infinite amount of time, I would write little reports on (at least) the following distributions

1. Negative binomial. The output of a Poisson process with additional randomness is negative binomial.

2. Multivariate normal. The outcome is a vector of correlated outcomes

3. Multinomial. This helps to figure out the chances of getting (142 red, 213 blue, 187 orange, 258 brown, 200 green) in a bag of 1000 *M&M* candies. In contrast, the binomial only helps to figure out the chances of getting (455 white, 545 pink) in bag of *Good & Plenty*.

4. Dirichlet. This extends the beta distribution to multiple dimensions. It allows to talk about the possibility that the *M&M Mars* company does not always use the same probability mixture for their bags of candy. Perhaps they choose $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ according to some random scheme.

5. Wishart. The output of this process is a square matrix that can be interpreted as a covariance matrix.

9 Conclusions

In some ways, I feel encouraged by the state of probability modeling. Compared to the time when I was in graduate school, when computers were inaccessible and there was little understanding of distributions except for the normal, t, and F distributions, we have an enormous amount of conceptual power at our fingertips. Any probability model one can imagine can be brought into use relatively easily. If one can offer a plausible reason for investigation of that new model, then others will probably want to consider variations on the parameters and domain.

As a case in point, I would mention the growth of interest in the so-called skewed distributions. The most prominent is the skew-normal distribution, which is obtained by taking the normal pdf, $f(x_i; \mu, \sigma^2)$ and multiplying by a skew factor, which is 2 times the cdf of the normal.

$$skew - normal\, pdf\, :\, f(x; \mu, \sigma^2) \times 2 \times \int_{-\infty}^{\alpha x} f(t; 0, 1) dt \cdot \tag{109}$$

If the skew factor $\alpha = 0.0$, then the skew disappears and this is just the same old normal. As far as I can tell, this was skew framework was originally proposed in 1985 (Azzalini, A. (1985). "A class of distributions which includes the normal ones". Scand. J. Statist. 12: 171–178) and there are now proposals for skew versions of most distributions (the ones we previously thought were symmetric, such as t).

I am encouraged, but also frightened because it appears to grow more and more difficult for part-timers like me to comprehend the magnitude of the probability modeling enterprise. In a 2008 article for the teacher's corner in the American Statistician, Leemis and McQueston assembled the single most overwhelming piece of line art I have ever seen. A snapshot is reproduced in Figure 30:

If I were a graduate student, here is where I would start.

First, I would probably take three or four more math courses than I took. I took calculus for engineering students, linear algebra, and sat in on courses in differential equations and mathematical statistics. Sitting in is never as good as actually taking the courses, and I've often regretted that I did not take the time. I never enrolled in real analysis, but I wish I had. If you are a student who has already taken these courses, and you think you know everything, go find some physicists who do asymptotic distribution theory. There is plenty more to learn.

Second, I would explore as many distributions as I could find pre-packaged for whatever statistical software is in vogue in the future. At one time, that was SAS, now it seems to be S+/R. I would try to plot the pdf's and cdf's, explore the sampling distributions of their

means. I'd try to verify the textbook claims about the way in which some function of a random variable from one distribution converges to another one. I'd write short summaries of the distributions for my use.

Third, I'd get a copy of an open-source scientific programming library, such as the GNU Scientific Library or CERN's COLT, and I would study ways to integrate those tools with my preferred statistical modeling tools. It seems certain to me that when new statistical distributions appear on the scene, they will first be offered in a low-level programming language like C, Fortran, or Java, and a familiarity with a scientific programming library will facilitate the integration of those new distributions with my collection.

Finally, I'd find a practitioner of Bayesian statistics. Even if you don't choose to be a Bayesian, it is still likely that working with one of them will help. Bayesians are, almost by definition, forced to live in a forest of statistical distributions and they need ways to make those distributions work together, more or less.

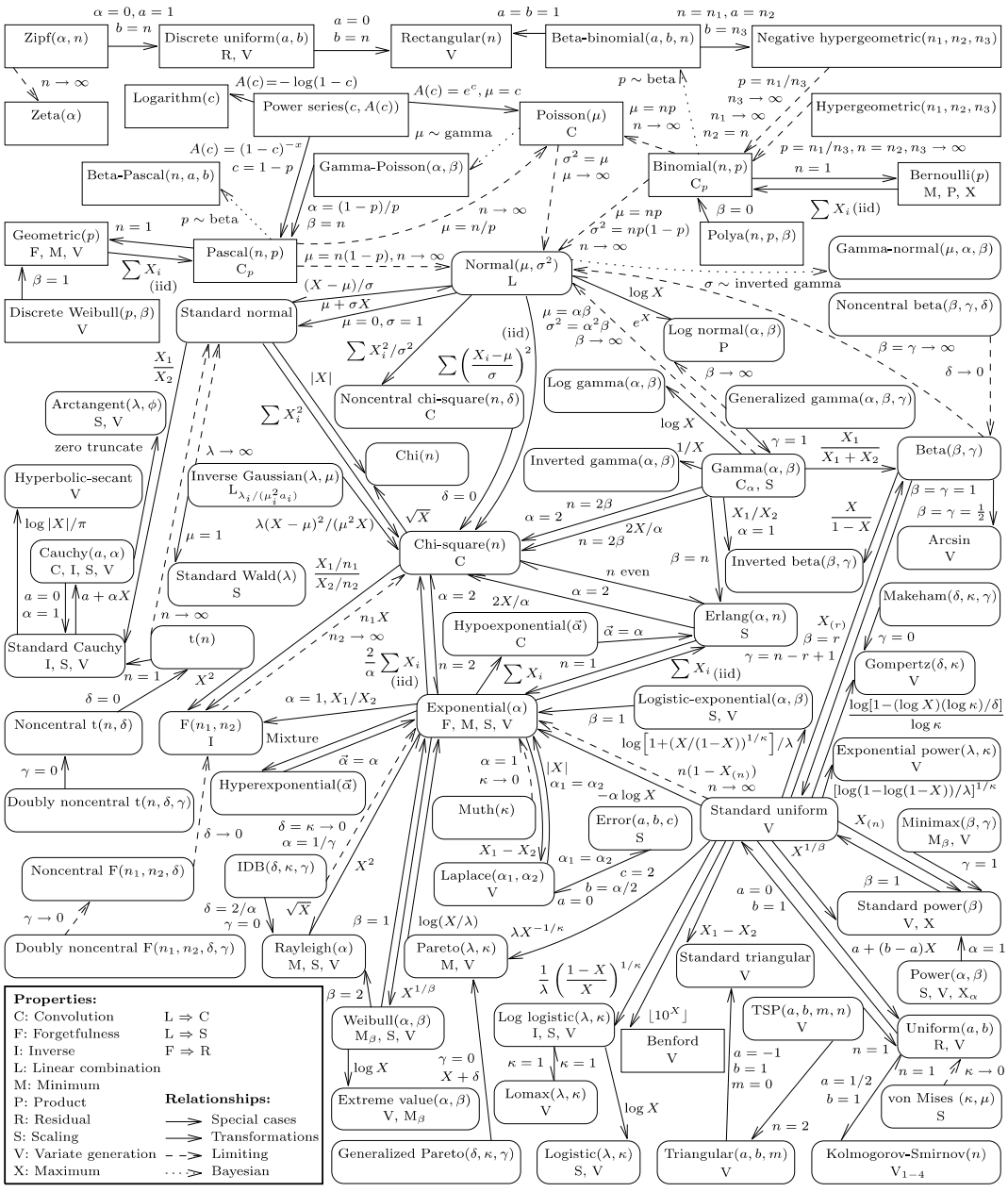


Figure 1. Univariate distribution relationships.

Figure 30: Leemis and McQueston Distribution Diagram