

Common Probability Distributions: The Data Scientist's Crib Sheet

Data scientists have hundreds of probability distributions from which to choose. Where to start?

Data science, whatever it may be, remains a big deal. “A data scientist is better at statistics than any software engineer,” you may overhear a [pundit](https://twitter.com/josh_wills/status/198093512149958656) say, at your local tech get-togethers and hackathons. The applied mathematicians have their revenge, because statistics hasn't been this [talked-about since the roaring 20s](https://en.wikipedia.org/wiki/Ronald_Fisher#Statistical_Methods_for_Research_Workers) (https://en.wikipedia.org/wiki/Ronald_Fisher#Statistical_Methods_for_Research_Workers). They have their own legitimizing Venn diagram (<https://pbs.twimg.com/media/BdoZ6NjlcAAGv3w.png>) of which people don't [make fun](http://41.media.tumblr.com/tumblr_lqfu7kJIFN1qz6f4bo1_500.jpg) (http://41.media.tumblr.com/tumblr_lqfu7kJIFN1qz6f4bo1_500.jpg). Suddenly it's you, the engineer, left out of the chat about [confidence intervals](https://en.wikipedia.org/wiki/Confidence_interval) instead of tutting at the analysts who have never heard of the Apache Bikeshed project for distributed comment formatting. To fit in, to be the life and soul of that party again, you need a crash course in stats. Not enough to get it right, but enough to sound like you could, by making basic observations.

Probability distributions are fundamental to statistics, just like data structures are to computer science. They're the place to start studying if you mean to talk like a data scientist. You can sometimes get away with simple analysis using [R](https://www.r-project.org/) or [scikit-learn](http://scikit-learn.org/) without quite understanding distributions, just like you can manage a Java program without understanding hash functions. But it would soon end in tears, bugs, bogus results, or worse: sighs and eye-rolling from stats majors.

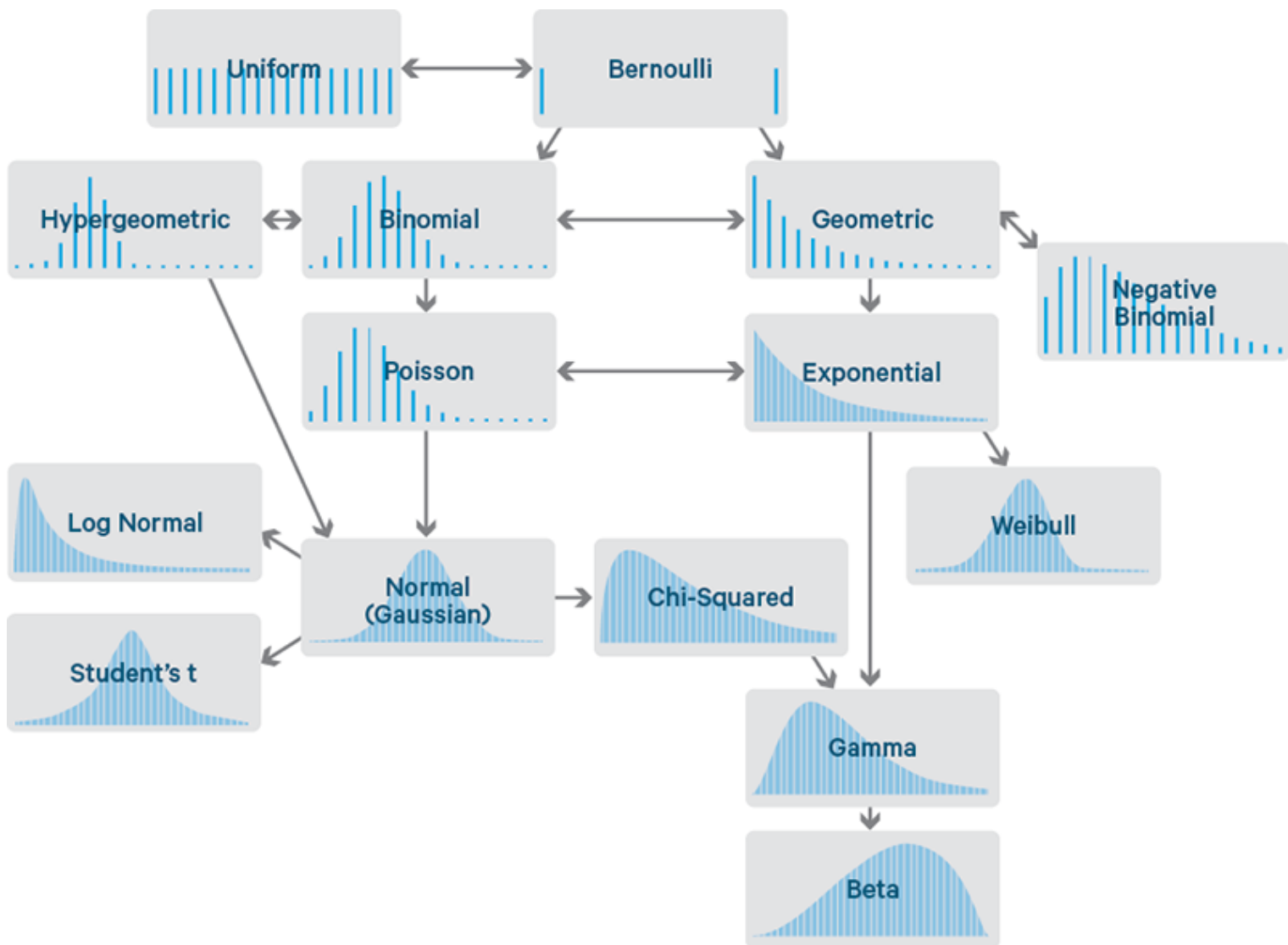
There are hundreds of probability distributions, some sounding like monsters from medieval legend like the Muth (<http://www.math.wm.edu/~leemis/chart/UDR/PDFs/Muth.pdf>) or Lomax (https://en.wikipedia.org/wiki/Lomax_distribution). Only about 15 distributions turn up consistently in practice though. What are they, and what clever insights about each of them should you memorize?

Now, What's a Probability Distribution?

Things happen all the time: dice are rolled, it rains, buses arrive. After the fact, the specific outcomes are certain: the dice came up 3 and 4, there was half an inch of rain today, the bus took 3 minutes to arrive. Before, we can only talk about how likely the outcomes are. Probability distributions describe what we think the probability of each outcome is, which is sometimes more interesting to know than simply which single outcome is most likely. They come in many shapes, but in only one size: probabilities in a distribution always add up to 1.

For example, flipping a fair coin has two outcomes: it lands heads or tails. (Assume it can't land on edge or be stolen by a seagull in mid-air.) Before the flip, we believe there's a 1 in 2 chance, or 0.5 probability, of heads. The same is true for tails. That's a probability distribution over the two outcomes of the flip, and if you can follow that sentence, you've already mastered the [Bernoulli distribution](https://en.wikipedia.org/wiki/Bernoulli_distribution) (https://en.wikipedia.org/wiki/Bernoulli_distribution).

Despite exotic names, the common distributions relate to each other in intuitive and interesting ways that make them easy to recall, and remark on with an air of authority. Several follow naturally from the Bernoulli distribution, for example. It's time to reveal a map of the relationships.



<http://blog.cloudera.com/wp-content/uploads/2015/12/distribution.png>

Each distribution is illustrated by an example of its probability density function (https://en.wikipedia.org/wiki/Probability_density_function) (PDF). This post deals only with distributions of outcomes that are single numbers. So, the horizontal axis in each box is the set of possible numeric outcomes. The vertical axis describes the probability of outcomes. Some distributions are discrete, over outcomes that must be integers like 0 or 5. These appear as sparse lines, one for each outcome, where line height is the probability of that outcome. Some are continuous, for outcomes that can take on any real numeric value like -1.32 or 0.005. These appear as dense curves, where it's areas under sections of the curve that give probabilities. The sums of the heights of lines, and areas under the curves, are always 1.

Print, cut along the dotted line, and take it with you in your wallet or purse. This is your field guide to spotting distributions and their relatives.

Bernoulli and Uniform

You met the Bernoulli distribution above, over two discrete outcomes—tails or heads. Think of it, however, as a distribution over 0 and 1, over 0 heads (i.e. tails) or 1 heads. Above, both outcomes were equally likely, and that’s what’s illustrated in the diagram. The Bernoulli PDF has two lines of equal height, representing the two equally-probable outcomes of 0 and 1 at either end.

The Bernoulli distribution could represent outcomes that aren’t equally likely, like the result of an unfair coin toss. Then, the probability of heads is not 0.5, but some other value p , and the probability of tails is $1-p$. Like many distributions, it’s actually a family of distributions defined by parameters, like p here. When you think “Bernoulli (https://en.wikipedia.org/wiki/Jacob_Bernoulli)” just think “(possibly unfair) coin toss.”

It’s a short jump to imagine a distribution over many equally-likely outcomes: the uniform distribution ([https://en.wikipedia.org/wiki/Uniform_distribution_\(discrete\)](https://en.wikipedia.org/wiki/Uniform_distribution_(discrete))), characterized by its flat PDF. Imagine rolling a fair die. The outcomes 1 to 6 are equally likely. It can be defined for any number of outcomes n or even as a continuous distribution.

Associate the uniform distribution with “rolling a fair die.”

Binomial and Hypergeometric

The binomial distribution (https://en.wikipedia.org/wiki/Binomial_distribution) may be thought of as the sum of outcomes of things that follow a Bernoulli distribution. Toss a fair coin 20 times; how many times does it come up heads? This count is an outcome that follows the binomial distribution. Its parameters are n , the number of trials, and p , the probability of a “success” (here: heads, or 1). Each flip is a Bernoulli-distributed outcome, or trial (https://en.wikipedia.org/wiki/Bernoulli_trial). Reach for the binomial distribution when counting the number of successes in things that act like a coin flip, where each flip is independent and has the same probability of success.

Or, imagine an urn with equal numbers of white and black balls. Close your eyes and draw a ball and note whether it is black, then put it back. Repeat. How many times did you draw a black ball? This count also follows a binomial distribution.

Imagining this odd situation has a point, because makes it simple to explain the hypergeometric distribution (https://en.wikipedia.org/wiki/Hypergeometric_distribution). This is the distribution of that same count if the balls were drawn *without replacement* instead. Undeniably it’s a cousin to the binomial distribution, but not the same, because the probability of success changes as balls are removed. If the number of balls is large relative to the number of draws, the distributions are similar because the chance of success changes less with each draw.

When people talk about picking balls from urns without replacement, it’s almost always safe to interject, “the hypergeometric distribution, yes,” because I have never met anyone who actually filled urns with balls and then picked them out, and replaced them or otherwise, in real life. (I don’t even know anyone who owns an urn.) More broadly, it should come to mind when picking out a significant subset of a population as a sample.

Poisson

What about the count of customers calling a support hotline each minute? That’s an outcome whose distribution sounds binomial, if you think of each second as a Bernoulli trial in which a customer doesn’t call (0) or does (1). However, as the power company knows, when the power goes out, *2 or even hundreds of people*

can call in the same second. Viewing it as 60,000 millisecond-sized trials still doesn't get around the problem—many more trials, much smaller probability of 1 call, let alone 2 or more, but, still not technically a Bernoulli trial. However, taking this to its infinite, logical conclusion works. Let n go to infinity and let p go to 0 to match so that np stays the same. This is like heading towards infinitely many infinitesimally small time slices in which the probability of a call is infinitesimal. The limiting result is the Poisson distribution (https://en.wikipedia.org/wiki/Poisson_distribution).

Like the binomial distribution, the Poisson distribution is the distribution of a count—the count of times something happened. It's parameterized not by a probability p and number of trials n but by an average rate λ , which in this analogy is simply the constant value of np . The Poisson distribution is what you *must* think of when trying to count events over a time given the continuous rate of events occurring.

When things like packets arrive at routers, or customers arrive at a store, or things wait in some kind of queue, think “Poisson (https://en.wikipedia.org/wiki/Simon_Denis_Poisson).”

Geometric and Negative Binomial

From simple Bernoulli trials arises another distribution. How many times does a flipped coin come up tails before it first comes up heads? This count of tails follows a geometric distribution (https://en.wikipedia.org/wiki/Geometric_distribution). Like the Bernoulli distribution, it's parameterized by p , the probability of that final success. It's not parameterized by n , a number of trials or flips, because the number of failure trials is the outcome itself.

If the binomial distribution is “How many successes?” then the geometric distribution is “How many failures until a success?”

The negative binomial distribution (https://en.wikipedia.org/wiki/Negative_binomial_distribution) is a simple generalization. It's the number of failures until r successes have occurred, not just 1. It's therefore parameterized also by r . Sometimes it's described as the number of successes until r failures. As my life coach says, success and failure are what you define them to be, so these are equivalent, as long as you keep straight whether p is the probability of success or failure.

If you need an ice-breaker, you might point out that the binomial and hypergeometric distributions are an obvious pair, but the geometric and negative binomial distributions are also pretty similar, and then say, “I mean, who names these things, am I right?”

Exponential and Weibull

Back to customer support calls: how long until the next customer calls? The distribution of this waiting time sounds like it could be geometric, because every second that nobody calls is like a failure, until a second in which finally a customer calls. The number of failures is like the number of the seconds that nobody called, and that's *almost* the waiting time until the next call, but almost isn't close enough. The catch this time is that the sum will always be in whole seconds, but this fails to account for the wait within that second until the customer finally called.

As before, take the geometric distribution to the limit, towards infinitesimal time slices, and it works. You get the [exponential distribution](https://en.wikipedia.org/wiki/Exponential_distribution) (https://en.wikipedia.org/wiki/Exponential_distribution), which accurately describes the distribution of time until a call. It's a continuous distribution, the first encountered here, because the outcome time need not be whole seconds. Like the Poisson distribution, it is parameterized by a rate λ .

Echoing the binomial-geometric relationship, Poisson's "How many events per time?" relates to the exponential's "How long until an event?" Given events whose count per time follows a Poisson distribution, then the time between events follows an exponential distribution with the same rate parameter λ . This correspondence between the two distributions is *essential* to name-check when discussing either of them.

The exponential distribution should come to mind when thinking of "time until event", maybe "time until failure." In fact, this is so important that more general distributions exist to describe time-to-failure, like the [Weibull distribution](https://en.wikipedia.org/wiki/Weibull_distribution) (https://en.wikipedia.org/wiki/Weibull_distribution). Whereas the exponential distribution is appropriate when the rate—of wear, or failure for instance—is constant, the Weibull distribution can model increasing (or decreasing) rates of failure over time. The exponential is merely a special case.

Think of "[Weibull](https://en.wikipedia.org/wiki/Waloddi_Weibull) (https://en.wikipedia.org/wiki/Waloddi_Weibull)" when the chat turns to time-to-failure.

Normal, Log-Normal, Student's t, and Chi-squared

The [normal distribution](https://en.wikipedia.org/wiki/Normal_distribution) (https://en.wikipedia.org/wiki/Normal_distribution), or [Gaussian](https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss) (https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss) distribution, is maybe the most important of all. Its bell shape is instantly recognizable. Like [e](https://en.wikipedia.org/wiki/E_(mathematical_constant)) ([https://en.wikipedia.org/wiki/E_\(mathematical_constant\)](https://en.wikipedia.org/wiki/E_(mathematical_constant))), it's a curiously particular entity that turns up all over, from seemingly simple sources. Take a bunch of values following the same distribution—*any* distribution—and sum them. The distribution of their sum follows (approximately) the normal distribution. The more things that are summed, the more their sum's distribution matches the normal distribution. (Caveats: must be a well-behaved distribution, must be independent, only tends to the normal distribution.) The fact that this is true regardless of the underlying distribution is amazing.

This is called the [central limit theorem](https://en.wikipedia.org/wiki/Central_limit_theorem) (https://en.wikipedia.org/wiki/Central_limit_theorem), and you must know that this is what it's called and what it means, or you will be immediately heckled.

In this sense, it relates to all distributions. However it's particularly related to distributions of sums of things. The sum of Bernoulli trials follows a binomial distribution, and as the number of trials increases, that binomial distribution becomes more like the normal distribution. Its cousin the hypergeometric distribution does too. The Poisson distribution—an extreme form of binomial—also approaches the normal distribution as the rate parameter increases.

An outcome that follows a [log-normal distribution](https://en.wikipedia.org/wiki/Log-normal_distribution) (https://en.wikipedia.org/wiki/Log-normal_distribution) takes on values whose logarithm is normally distributed. Or: the exponentiation of a normally-distributed value is log-normally distributed. If sums of things are normally distributed, then remember that products of things are log-normally distributed.

[Student's t-distribution](https://en.wikipedia.org/wiki/Student%27s_t-distribution) (https://en.wikipedia.org/wiki/Student%27s_t-distribution) is the basis of the [t-test](https://en.wikipedia.org/wiki/Student%27s_t-test) (https://en.wikipedia.org/wiki/Student%27s_t-test) that many non-statisticians learn in other sciences. It's used in reasoning about the mean of a normal distribution, and also approaches the normal distribution as its

parameter increases. The distinguishing feature of the t -distribution are its tails, which are fatter than the normal distribution's.

If the fat-tail anecdote isn't a hot enough take to wow your neighbor, go to its mildly-interesting [back-story](http://www.mlive.com/kalamabrew/index.ssf/2009/03/because_of_beer_1900s_guinness.html) (http://www.mlive.com/kalamabrew/index.ssf/2009/03/because_of_beer_1900s_guinness.html) concerning beer. Over 100 years ago, [Guinness](http://www.guinness.com/) (<http://www.guinness.com/>) was using statistics to make better stout. There, [William Sealy Gosset](https://en.wikipedia.org/wiki/William_Sealy_Gosset) (https://en.wikipedia.org/wiki/William_Sealy_Gosset) developed some whole new stats theory just to grow better barley. Gosset convinced the boss that the other brewers couldn't figure out how to use the ideas, and so got permission to publish, but only under the pen name "Student". Gosset's best-known result is this t -distribution, which is sort of named after him.

Finally, the [chi-squared distribution](https://en.wikipedia.org/wiki/Chi-squared_distribution) (https://en.wikipedia.org/wiki/Chi-squared_distribution) is the distribution of the sum of squares of normally-distributed values. It's the distribution underpinning the [chi-squared test](https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test) (https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test) which is itself based on the sum of squares of differences, which are supposed to be normally distributed.

Gamma and Beta


At this point, if you're talking about chi-squared anything, then the conversation has gotten serious. You are likely talking to actual statisticians, and you may want to excuse yourself at this point, because things like the [gamma distribution](https://en.wikipedia.org/wiki/Gamma_distribution) (https://en.wikipedia.org/wiki/Gamma_distribution) may come up. It is a generalization of *both* the exponential and chi-squared distributions. More like the exponential distribution, it is used as a sophisticated model of waiting times. For example, the gamma distribution comes up when modeling the time until the next n events occur. It appears in machine learning as the "[conjugate prior](https://en.wikipedia.org/wiki/Conjugate_prior)" (https://en.wikipedia.org/wiki/Conjugate_prior)" to a couple distributions.

Do *not* get into that conversation about conjugate priors, but if you do, be sure that you're about to talk about the [beta distribution](https://en.wikipedia.org/wiki/Beta_distribution) (https://en.wikipedia.org/wiki/Beta_distribution), because it's the conjugate prior to most every other distribution mentioned here. As far as data scientists are concerned, that's what it was built for. Mention this casually, and move toward the door.

The Beginning of Wisdom

Probability distributions are something you can't know too much about. The truly interested should check out this [incredibly detailed map of all univariate distributions](http://www.math.wm.edu/~leemis/chart/UDR/UDR.html) (<http://www.math.wm.edu/~leemis/chart/UDR/UDR.html>). Hopefully, this anecdotal guide gives you the confidence to appear knowledgeable and with-it in today's tech culture. Or at least, a way to detect, with high probability, when you should find a less nerdy cocktail party.



 analysis (<https://blog.cloudera.com/blog/tag/analysis/>) analytics (<https://blog.cloudera.com/blog/tag/analytics/>) data (<https://blog.cloudera.com/blog/tag/data/>) Data Science (<https://blog.cloudera.com/blog/tag/data-science/>) R (<https://blog.cloudera.com/blog/tag/r/>) statistics (<https://blog.cloudera.com/blog/tag/statistics/>)