PLSC 502: "Statistical Methods for Political Research"

Discrete Probability Distributions
September 27, 2016

Overview

We'll spend the next couple days discussing probability distributions. There are a lot of probability distributions, most of which aren't used (or even seen) very often in the social sciences. We'll focus on the ones that are, and that make up (maybe) 90 percent of all commonly-used ones. We'll start today with several related to the binomial, all of which are *discrete*. Next time we'll move on to some distributions used for continuous random variables.

Bernoulli

Imagine a binary $X \in \{0, 1\}$ with:

$$\begin{aligned} X &= 0 \text{ with probability } 1 - \pi \\ &= 1 \text{ with probability } \pi. \end{aligned}$$

That is, X's PDF is:

$$f(x) = \begin{cases} 1 - \pi & \text{for } X = 0 \\ \pi & \text{for } X = 1 \end{cases}$$

which we can usefully rewrite as

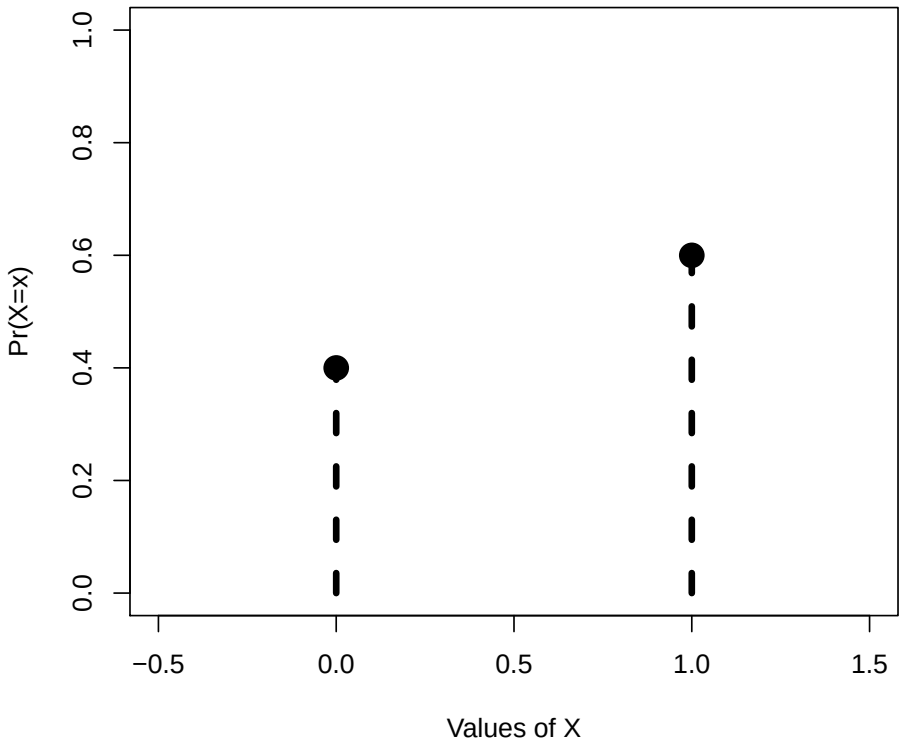
$$f(x) = \pi^x (1 - \pi)^{1-x}, \; x \in \{0, 1\}. \tag{1}$$

This is a *Bernoulli* variable; we say "X is distributed Bernoulli," and write:

$$X \sim \text{ Bernoulli}(\pi)$$

The Bernoulli is a one-parameter distribution for a discrete variable that can take on only two values; it's therefore the natural choice for binary / dichotomous variables. Because X can only take on two values, we say that X has "support" only in $\{0, 1\}$; that means that the only possible values X can take on are those two. It's pretty simple, when you look at it:

Figure 1: Bernoulli PDF, $\pi = 0.6$



What can we say about X? Well, the CDF is:

$$\begin{aligned} F(x) &= \sum_x f(x) \\ &= \begin{cases} 1-\pi & \text{for } X = 0 \\ 1 & \text{for } X = 1 \end{cases} \end{aligned}$$

And:

$$\begin{aligned} \mathrm{E}(X) &= \sum_x x f(x) \\ &= (0)(1-\pi) + (1)(\pi) \\ &= \pi \end{aligned}$$

And so:

$$\begin{aligned}
\text{Var}(X) &= \sum_x [X - \text{E}(X)]^2 f(x) \\
&= \sum_x [X - \pi]^2 f(x) \\
&= (0-\pi)^2(1-\pi) + (1-\pi)^2\pi \\
&= \pi^2 - \pi^3 + \pi - 2\pi^2 + \pi^3 \\
&= \pi^2 - \pi \\
&= \pi(1-\pi)
\end{aligned}$$

This makes sense: This value will be at its largest when $\pi = 0.5$, which is when we'd see the greatest amount of variation between "0s" and "1s."

I won't derive it, but the skewness of a Bernoulli variate is:

$$\text{Skewness} = \frac{(1-\pi) - \pi}{\sqrt{(1-\pi)\pi}}.$$

One thing that is immediately apparent is that the skewness is 0 when $\pi = 0.5$ (as it should be).

More generally, the *moment-generating function* for a Bernoulli variate is:

$$\begin{aligned}
\psi(t) &= \int_{-\infty}^{\infty} \exp(tx)\, dF(x) \\
&= \sum_{n=0}^{1} \exp(tn)\pi^n(1-\pi)^{1-n} \\
&= \exp(0)(1-\pi) + \exp(t)\pi \\
&= (1-\pi) + \pi\exp(t) \qquad (2)
\end{aligned}$$

Note that the kth derivative of (2) with respect to t is just

$$\frac{\partial^k \psi(t)}{\partial^k t} = \pi \exp(t) \;\forall\; k.$$

This means that X's kth moment *around zero* (that is, the kth "raw moment") is just:

$$\text{E}(X^k) = \pi \;\forall\; k > 0.$$

This is because $X \in \{0, 1\}$, so that the long run expectation of X^k is just the expectation of kth power of either 0 or 1, where the latter appears with probability π. More important (and useful), this leads to expressions for the *central moments* (those around the mean). For example, the first moment around the mean is just

$$M_1 = \pi,$$

the second central moment is

$$M_2 = \pi(1-\pi),$$

and so forth.

Binomial

The Bernoulli is the simplest discrete probability distribution; moreover, it is the "building block" for a large set of important discrete distributions. Probably the most important is the *binomial*, which is most easily thought of as the number of "1s" ("successes") in n independent Bernoulli trials, each with identical probability π:

$$f(x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} \tag{3}$$

where $\pi \in [0,1]$ is the probability of "success," $n \in \{0,1,2,...\}$ is the number of trials, and

$$\binom{n}{x} \equiv \frac{n!}{x!(n-x)!}.$$

We refer to such a variable as being "binomial-distributed," and write

$$X \sim \text{binomial}(n,\pi).$$

The binomial is called that after the *binomial theorem* for the expansion of powers of sums in mathematics, which states that

$$(a+b)^n = \sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}. \tag{4}$$

For example,

$$\begin{aligned}(a+b)^2 &= a^2+2ab+b^2,\\ (a+b)^3 &= a^3+3a^2b+3ab^2+y^3,\\ (a+b)^4 &= a^4+4a^3b+6a^2b^2+4ab^3+b^4,\end{aligned}$$

and so forth.

A binomial distribution is a two-parameter distribution, n and π; if X is a binomial variate, it necessarily has support only in the set $\{0,1,2,...n\}$. The "choose" term in (3) reflects all possible orderings of "0s" and "1s" on the individual trials. Two independent Bernoulli trials illustrate:

$$\begin{aligned}
\Pr(X = 0) &= \Pr(X_1 = 0, X_2 = 0) \\
&= \Pr(X_1 = 0) \times \Pr(X_2 = 0) \\
&= (1 - \pi)^2
\end{aligned}$$

$$\begin{aligned}
\Pr(X = 1) &= \Pr(X_1 = 1, X_2 = 0 \textit{ or } X_1 = 0, X_2 = 1) \\
&= \Pr(X_1 = 1) \times \Pr(X_2 = 0) + \Pr(X_1 = 0) \times \Pr(X_2 = 1) \\
&= \pi(1 - \pi) + (1 - \pi)\pi
\end{aligned}$$

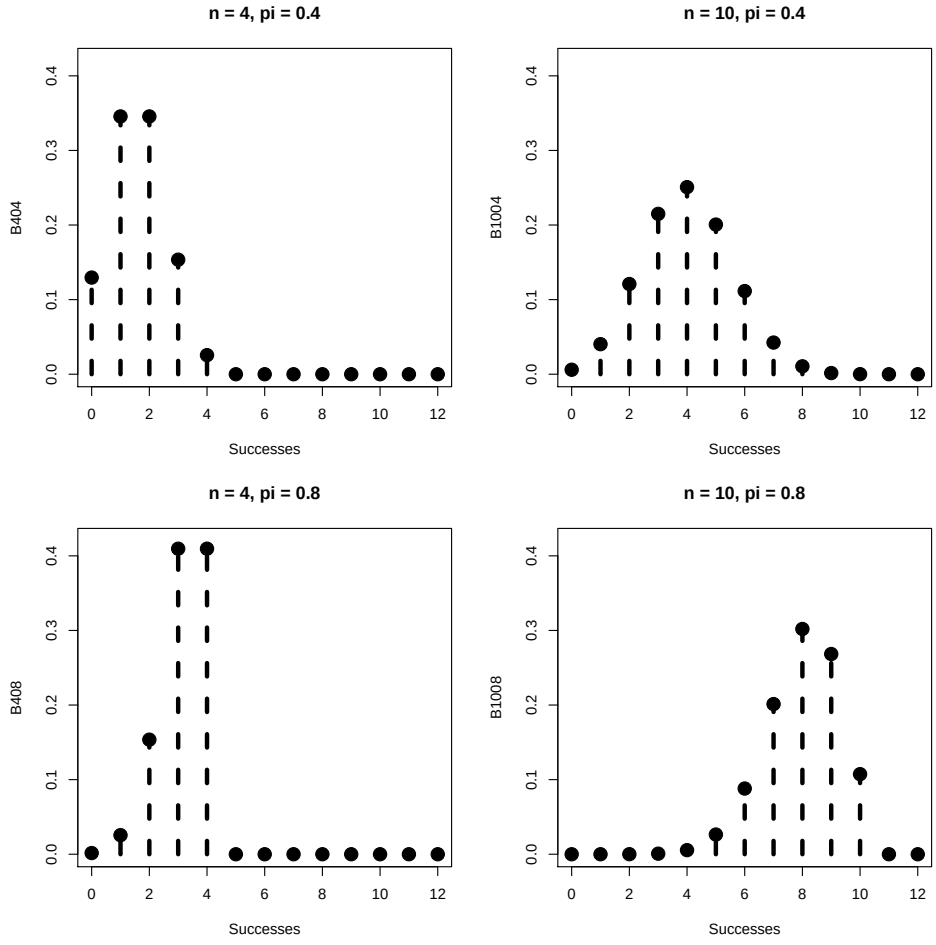
$$\begin{aligned}
\Pr(X = 2) &= \Pr(X_1 = 1, X_2 = 1) \\
&= \Pr(X_1 = 1) \times \Pr(X_2 = 1) \\
&= \pi^2
\end{aligned}$$

The binomial CDF – which we can think of intuitively as the probability of observing x or fewer "successes" in n Bernoulli trials with probability of success π – is then just

$$\begin{aligned}
F(x) &= \sum_x f(x) \\
&= \sum_{j=0}^{x} \binom{n}{j} \pi^j (1 - \pi)^{n-j}
\end{aligned}$$

This looks like:

Figure 2: Binomial PDF, $\pi = 0.4$ and 0.8, $n = 4$ and 10



The expected value of a binomial variate X can (easily) be shown to be

$$\mathrm{E}(X) = n\pi,$$

which is pretty intuitive since n is the number of "trials," and π is the probability of "success." Likewise,

$$\begin{aligned}
\mathrm{Var}(X) &= \sum_x [X - \mathrm{E}(X)]^2 f(x) \\
&= \sum_x (X - \pi n)^2 \binom{n}{x} \pi^x (1-\pi)^{n-x} \\
&= n\pi(1-\pi).
\end{aligned}$$

This means (again, intuitively) that the variability in a binomial variable is

- increasing in n, and
- largest when $\pi = 0.5$ for a fixed value of n.

A binomial variate is necessarily *unimodal* (except for the special case of a Bernoulli variate with $\pi = 0.5$); it can also be *skewed*, depending on the value of π.

Because the binomial is used to model the number of "successes" out of a known number of "trials," it is widely used in the social sciences. It is useful, for example, for modeling proportions or percentages where the denominator is known. So, for example, we might believe that the number of "yea" votes for bills in the U.S. House of Representatives (out of $n = 435$) follows a binomial distribution.

Geometric

Another thought experiment is to consider repeating (independent) Bernoulli trials with probability of success π until we observe the *first "success."* The number of independent Bernoulli trials needed to achieve one success is a *geometric* random variable. If X is a geometric random variable with parameter π, then

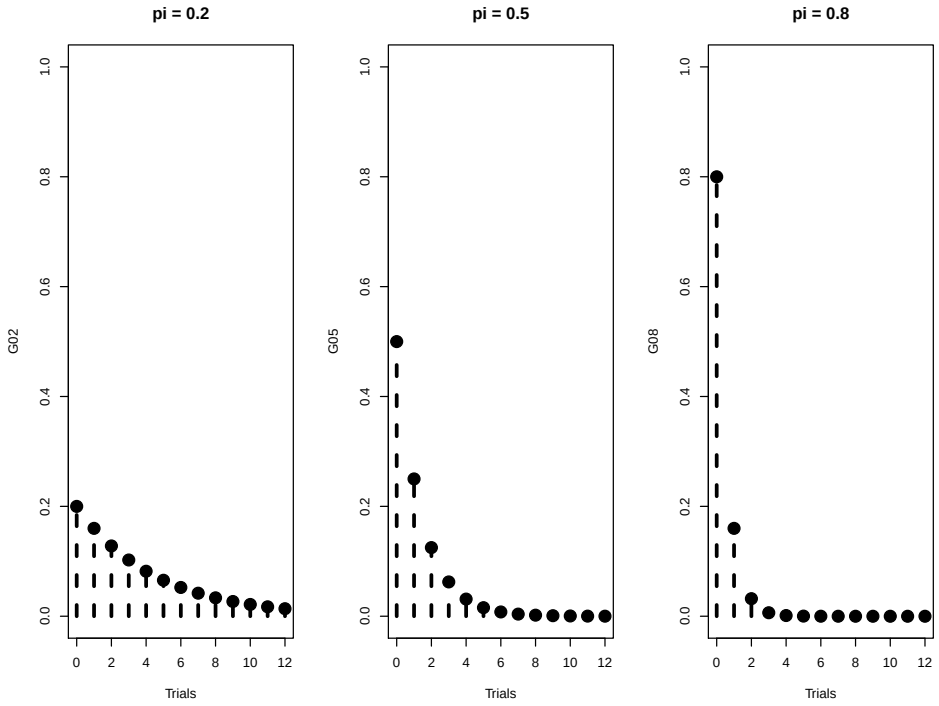
$$f(x) = \pi(1 - \pi)^{x-1}. \tag{5}$$

The geometric is thus a one-parameter distribution, with $\pi \in [0, 1]$; we write

$$X \sim \text{geometric}(\pi).$$

This looks like:

Figure 3: Geometric PDF, $\pi = 0.2$, 0.5, and 0.810



The CDF – the cumulative probability of a first success in $\{1, 2, ...\}$ trials – is then

$$\begin{aligned} F(x) &= \sum_{j=1}^{x} \pi(1-\pi)^{x-1} \\ &= 1-(1-\pi)^x. \end{aligned}$$

The expected value of the geometric distribution is (again, unsurprisingly):

$$\mathrm{E}(X) = \frac{1}{\pi},$$

suggesting (intuitively) that as the probability of success declines, we expect to undertake larger and larger numbers of trials before observing our first "success." The variance of X is similarly intuitive:

$$\mathrm{Var}(X) = \frac{1-\pi}{\pi^2}.$$

That is, the variance gets arbitrarily close to zero as the probability of success approaches 1.0, and arbitrarily large as $\pi \to 0$.

Negative Binomial

The negative binomial can be thought of as a generalization of / variant on the geometric distribution. Imagine we begin conducting independent Bernoulli trials with probability of success π, and stop the trials upon observing r successes. The distribution of the number of *failures we observe* (X) before achieving the rth success is distributed according to a *negative binomial* distribution.[1] Such a variable X has support on the nonnegative integers. A simple expression of the PDF for a negative binomial variate is:

$$f(x) = \binom{r+x-1}{r-1}\pi^r(1-\pi)^x \tag{6}$$

In this formulation,[2] the corresponding CDF – which we can think of as the probability of observing x or fewer "failures" before the rth success – is:

$$F(x) = \sum_{j=0}^{x}\binom{r+j-1}{r-1}\pi^r(1-\pi)^j.$$

Importantly, one can show via a bit of algebra that this value is equal to one minus the CDF of the binomial distribution (hence the name).

Similar to the geometric, the expected value of a negative binomial variate is:

$$\mathrm{E}(X) = \frac{(1-\pi)r}{\pi}$$

and the variance is:

$$\mathrm{Var}(X) = \frac{(1-\pi)r}{\pi^2}.$$

The skewness is:

$$\text{Skewness} = \frac{1+\pi}{\sqrt{\pi r}},$$

which makes pretty clear that a negative binomial variate will always have a positive skew (though potentially a small one, if π is small).

We can think of the negative binomial in a number of ways; two important ones are:

- As a generalization of the geometric (and, in fact, the negative binomial distribution reduces to the geometric when $r = 1$), and

[1] Formally, the negative binomial is a continuous distribution; the special case where r is an integer value is known as the *Pascal distribution* (and the real-valued case the *Polya distribution*), though in most applications the two are effectively identical.

[2] And, as Johnson, Kotz, and Kemp (2005) note, there are *lots* of different ways to formulate the negative binomial distribution.

- as a Poisson variate (see below) with "heterogeneity." We'll talk more about that later in the course, and *much* more in PLSC 504...

Poisson

Now consider a very large number of n Bernoulli trials, where the probability π of an event in any one trial is small. In such a situation, the total number of events observed will follow a Poisson distribution.

Formally, for n independent Bernoulli trials with (sufficiently small) probability of success π and where $n\pi \equiv \lambda > 0$,[3] the probability of observing exactly x total "successes" as the number of trials grows without limit is:

$$\begin{aligned} f(x) &= \lim_{n\to\infty} \left[\binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \right] \\ &= \frac{\lambda^x \exp(-\lambda)}{x!}. \end{aligned}$$

This was actually the original derivation of the Poisson distribution (by – who else? – Simeon-Denis Poisson, back in 1837). This is sometimes known as the "Law of Rare Events" motivation for the Poisson distribution.

The CDF that corresponds to (7) is:

$$F(x) = \sum_{j=0}^{x} \frac{\lambda^j \exp(-\lambda)}{x!}.$$

Alternative Motivation: Counts of Events

An alternative way to think about the Poisson distribution is with an abstract model of *event counts*. Suppose we are interested in studying events, and that those events occur over time. We might consider the *constant rate* at which events occur; call this rate λ. It's useful to think of λ as the expected number of events in any particular time "period" of length h. Imagine further that the events in question are *independent*; that is, the occurrence of one event has no bearing on the probability that another will occur.

If the process that gives rise to the events in question (what we'll call the *event process*) conforms to these assumptions, then it's pretty straightforward to show that as the length of the interval $h \to 0$,

- The probability of an event occurring in the interval $(t, t+h] = \lambda h$

[3] Formally, holding λ constant as $n \to \infty$ requires that $\pi \to 0$.

- The probability of no event occurring in the interval $(t, t+h] = 1 - \lambda h$

Such a variable is what is known as a *Poisson process*: events occur independently with a constant probability equal to λ times the length of the interval (that is, λh).

Next, consider our outcome variable X_t as the number of events that have occurred in the interval t of length h. For such a process, the probability that the number of events occurring in $(t, t+h]$ is equal to some value $x \in \{0, 1, 2, 3, ...\}$ is:

$$f(x) = \frac{\exp(-\lambda h)\lambda h^x}{x!} \tag{7}$$

If all the intervals are of the same length (and equal to 1), this reduces to:

$$f(x) = \frac{\exp(-\lambda)\lambda^x}{x!} \tag{8}$$

This is the way we typically see the *Poisson distribution* written. By this logic, the Poisson distribution is the limiting distribution for the number of independent (Poisson) events occurring in some fixed period of length h (for Eq. 7) or 1 (for Eq. 8).

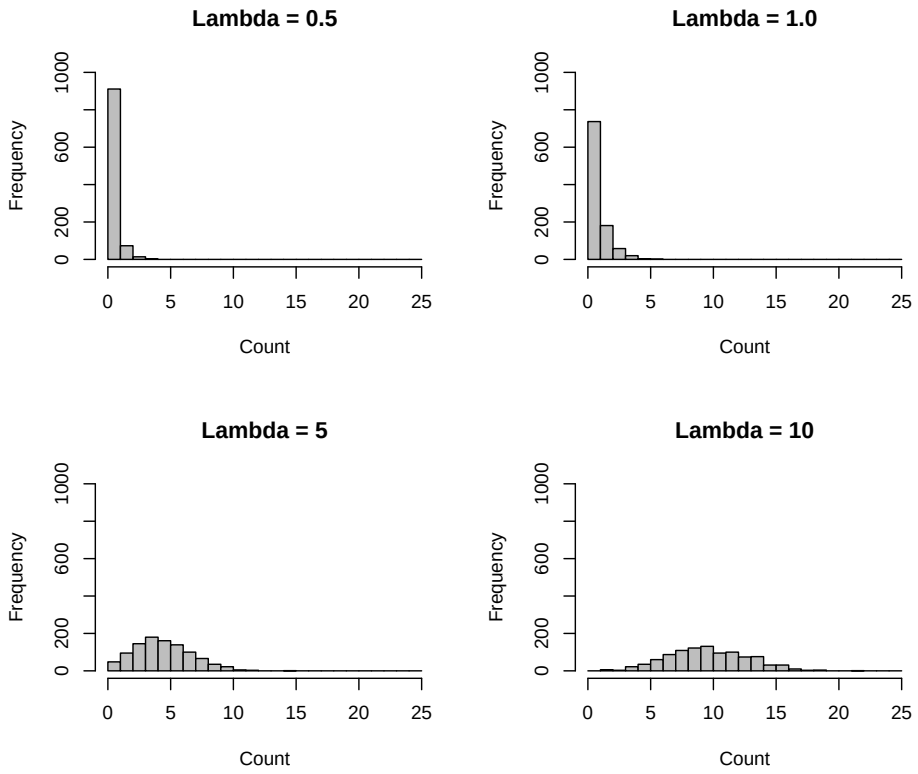
The assumptions underlying the event process – *constant arrival rates*, and *independence across events* – are key to deriving the Poisson distribution in this way. If we relax these assumptions, the resulting distribution(s) are not Poisson.

The Poisson Distribution: Characteristics

What is this odd thing we call the Poisson distribution, anyway? The Poisson distribution has several important traits:

- It is a discrete probability distribution, with support on the non-negative integers.
- The "rate" λ can also be interpreted as the expected number of events during an observation period t. In fact, for a Poisson variate X, $\mathrm{E}(X) = \lambda$.
- As λ increases, several interesting things happen:
    1. The *mean/mode* of the distribution gets bigger (no shock there).
    2. The *variance* of the distribution gets larger as well. This also makes sense: since the variable is bounded from below, its variability will necessarily get larger with its mean. In fact, in the Poisson, the mean equals the variance (that is, $\mathrm{E}(X) = \mathrm{Var}(X) = \lambda$).
    3. The distribution becomes more Normal-looking (and, in fact, becomes more Normal, period).

Figure 4: Empirical Poisson Variates, with Varying λs



Note as well that the Poisson distribution...

- ...is not preserved under affine transformations – that is, affine transformations of Poisson variates are not themselves (necessarily) Poisson variates as well.
- ...is preserved under addition (convolution) provided that the components are independent. That is, for two Poisson variates $X_1 \sim \text{Poisson}(\lambda_{X_1})$ and $X_2 \sim \text{Poisson}(\lambda_{X_2})$, $Z = X_1 + X_2 \sim \text{Poisson}(\mu_{X_1+X_2})$ *iff* X_1 and X_2 are *independent*. (See e.g. Winkelmann 1997, Chapter 2 for proofs). However,
- ...the same is not true for differences of Poisson variates; see Johnson et al. (2005), §4.12.3 for details.

Multinomial

All the distributions we've talked about so far have their roots in the Bernoulli, which means that there have been only two potential outcomes (we've called them "success" and "failure"). But suppose that instead of just two possibilities, we instead had K possible distinct *outcomes* for each trial, where each possible outcome has some corresponding probability of

happening on each trial π_k, and (of course) $\sum_{k=1}^{K} \pi_k = 1$.

We can then think of a multi-outcome analogue to the binomial, where the variable X_k denotes the number of times we observe outcome k out of n trials. The k-vector $\mathbf{X}$ denotes these k distinct variables:

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix}$$

In this formulation, the probability of $\mathbf{X}$ – that is, the joint probability that $X_1 = x_1, X_2 = x_2, ...X_K = x_K$ follows a *multinomial distribution* with PDF:

$$f(\mathbf{x}) = \frac{n!}{x_1!x_2!...x_K!}\pi_1^{x_1}\pi_2^{x_2}...\pi_K^{x_K} \tag{9}$$

It's pretty easy to see how this is a multinomial generalization of the binomial. And, nor surprisingly, its moments look like the binomial's as well; its mean is

$$\mathrm{E}(\mathbf{X}) \equiv \mathrm{E}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix} = n\begin{pmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_K \end{pmatrix}$$

and, similarly, the variance is:

$$\mathrm{Var}(\mathbf{X}) \equiv \mathrm{Var}\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_K \end{pmatrix} = n\begin{bmatrix} \pi_1(1-\pi_1) \\ \pi_2(1-\pi_2) \\ \vdots \\ \pi_K(1-\pi_K) \end{bmatrix}$$

Note as well that, because the probabilities necessarily sum to one (and because, if a trial results in an outcome of one type, it can't also be of any of the others), the covariance between any two elements of $\mathbf{X}$ (say, X_s and X_t, $s \neq t$) is:

$$\mathrm{Cov}(X_s, X_t) = -n\pi_s\pi_t \; \forall \; s \neq t.$$

In words, this means that the covariances will always be negative, and that their magnitude will be a function of (a) the number of trials n, and (b) the relative sizes of their respective probabilities π.

Relationships Among Discrete Distributions

This is a picture of how all these different distributions relate to each other...