Distributions arising from categorical trials

Jacob Fine

April 9th, 2025

Introduction

The simplest concepts can be the basis of the most important and fundamental paradigms in science. Here, we explore the simple concept of an experiment with two or more possible categorical outcomes, starting with the classic Bernoulli distribution, and use it to derive more complex distributions, including the Poisson-Binomial and the Dirichlet-Multinomial.

Bernoulli distribution

Consider an experiment with two possible outcomes, success or failure. Let $X \in \{0, 1\}$ be a random variable that models the outcome of this experiment, i.e., a coin toss, where we can say $X = 1$ for heads and $X = 0$ for tails. Therefore, $X \sim \text{Bern}(p)$ and the probability mass function (PMF) is given by

$$P(X = x) = p^x(1 - p)^{1-x}$$

Binomial distribution

Let us now consider a sequence of n independent and identically distributed (iid) Bernoulli experiments. Suppose we are interested in the distribution of the sum of these iid Bernoullis, so we now have $X \sim \text{Binom}(n, k)$ where k here represents the number of successes. There are a total of $\binom{n}{k}$ ways to choose k successes from the n Bernoulli trials. And since the trials are iid, we have that

$$P(X = k) = \binom{n}{k} p^k(1 - p)^{n-k}$$

Here, both n and p were fixed. But what distributions would we get if they were not fixed? This will be explored in the next sections.

Negative binomial distribution

Let us first consider the case where we are interested in the probability of the r^{th} success after the first k failures. Here, the meaning of k is different than the meaning in the previous example (Binomial) to keep with conventional notation. We are therefore want to have $r + k = n$ trials and of them, are choosing k failures until the r^{th} success. The distribution of the number of k failures until the r^{th} success follows a negative binomial, which has the PMF (where p now is the probability of success):

$$P(X = k) = \binom{r + k - 1}{k}(1 - p)^k p^r$$

We can think of the coefficient $\binom{r+k-1}{k}$ as resulting from the fact that we are conducting $r + k$, and stop on the r^{th} trial (hence the minus one). Of these, we are choosing k failures with probability $1 - p$. So the binomial coefficient just counts the total number of these combinations.

Geometric distribution

A special case of the negative binomial distribution is when we care about the number of failures k until the first success. So we can substitute $r = 1$ into the negative binomial PMF to obtain the PMF for the geometric distribution:

$$P(X = k) = (1 - p)^k p$$

In both cases (Geometric and Negative Binomial) the number of trials was not fixed, but we were still dealing with the sum of iid Bernoulis with probability of success p. But what if p was different for each Bernoulli in the sum? This brings us to the interesting case of the 'Poisson-Binomial' distribution, where we have independent but not idendically distributed Bernoullis.

Poisson-binomial distribution

This Poisson-Binomial is a generalization of the Binomial experiment where the probability of success in each independent trial is different, and given by p_i. The total number of trials is still fixed at n, and let us now consider k as the number of desired successes (not to be confused with our definition of k from before). We are using the same definitions of variables as in the Binomial case.

Consider the set of all indexes of the trials $\{1, 2...n\}$. Now consider the set of all subsets of the trial indexes F_k such that $|F_k| = \binom{n}{k}$ and every subset in F_k contains k elements. That is, $F_k = \{A \subseteq \{1, 2...n\} \mid |A| = k\}$. To develop this, let us consider the probability that all k events in $A \in F_k$ of length k succeed, which, since each event has a probability of success p_i and is independent, is given by:

$$P(X=k,A)=\prod_{i\in A}p_i\prod_{j\notin A}(1-p_j)$$

We just take the probabilities of each event in the subset A and multiply them, and then continue multiplying by all events that are not in the subset A, but still in F_k. This value is essentially probability that we have $X=k$ successes AND that the successes occured at the indices in A. So to get the total probability, we simply just sum over all possible subsets $A\in F_k$:

$$P(X=k)=\sum_{A\in F_k}\prod_{i\in A}p_i\prod_{j\notin A}(1-p_j)$$

We can now see clearly why in the case where the Bernoulli random variables are iid such that $p_1=p_2=...p_n$, since there are $|F_k|=\binom{n}{k}$ subsets, this would give us the binomial distribution.

Beta-binomial distribution

In the regualr Binomial distribution, we are assuming that p is a fixed value that we already know. But what if it isn't fixed, and all we know is the probability distribution of p? We often choose to model the prior distribution $p\sim\text{Beta}(\alpha,\beta)$, since when combined with the likelihood $P(X|p)$ which is a binomial distirbution, gives a posterior $P(p|X)$ that is also a Beta distribution. For this reason we call the Beta distribution a conjugate prior, since when combined with a particular likelihood function, gives a posterior that is in the same family of distributions as the prior. Then, we can use Bayes' theorem and integrate over p to get the distribution of $P(X)$, which is a Beta-Binomial distirbution, i.e., $X\sim\text{BetaBin}(n,\alpha,\beta)$. We therefore want to integrate $P(X|p)P(p)$ over $p\in[0,1]$.

$$P(X)=\int_p P(X|p)P(p)dp$$

Conisder that

$$P(X\mid p)=\binom{n}{k}p^k(1-p)^{n-k}$$

and (the Beta prior for p)

$$P(p)=\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}$$

We want to solve:

$$P(X)=\int_0^1\binom{n}{k}p^k(1-p)^{n-k}\cdot\frac{1}{B(\alpha,\beta)}p^{\alpha-1}(1-p)^{\beta-1}\,dp$$

So we first take out the constants, and simplify to obtain

$$P(X)=\binom{n}{k}\frac{1}{B(\alpha,\beta)}\int_0^1 p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\,dp$$

Fascinatingly, we can see that the integral

$$\int_0^1 p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\,dp$$

is a case of the Beta function ($\int_0^1 p^{x-1}(1-p)^{y-1}dp = B(x,y)$) where $x = k+\alpha$ and $y = n-k+\beta$. So we have obtained the PMF of a Beta-Binomial, $X \sim \text{BetaBin}(n,\alpha,\beta)$ as:

$$P(X) = \binom{n}{k}\frac{B(k+\alpha, n-k+\beta)}{B(\alpha,\beta)}$$

This, like a regular binomial, represents the probability of k successes but uses a Beta prior distribution of p. We can interpret α and β as the 'initial' number of observed successes and failures, respectively. Since the mean of the Beta-Binomial is $\mu = \frac{n\alpha}{\alpha+\beta}$, we are effectively estimating the probability of success with $\hat{p} = \frac{\alpha}{\alpha+\beta}$. Since the variance of the Beta prior is $\text{Var}(p) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, we can see that as $\alpha+\beta \to \infty$, the variance in the prior goes to zero, and the Beta-Binomal converges in distribution to a Binomial distribution. Before convergence, the Beta-Binomial incorporates a greater variance for the increased uncertainty of the probability of success p.

Multinomial distribution

Until now, we have considered distributions related to the Binomial 'built' on top of Bernoulli distirbutions, i.e., having two outcomes. But what if we considered the case where each trial had c possible outcomes? This gives rise to the Multinomial distribution. Here, let a random variable X_i represent each of the c outcomes, such that each X_i occurs with a probability p_i, such that $\sum_{i=1}^{c} p_i = 1$. We no longer speak about 'success' or 'failure' to emphasize the fact of > 2 outcomes. Also, note that this is different than the Poisson-Binomial, since in that case, we were considering n Bernoulii distributions but each distribution still had two outcomes. Though the distribution of each Bernoulli was different. Here, are now dealing with n iid 'c-noulis': c outcomes instead of just $c = 2$. Technically we call this a 'Categorical distribution' where each category occurs with probability p_i. A Bernoulli distribution is a Categorical distribution with $c = 2$. Therefore, we have n iid categorical distributions $X \sim \text{Categorical}(p_1, \ldots, p_c)$. So, the Bernoulli is to the Binomial what the Categorical is to the Multinomial. The PMF of the multinomial is just given by the product of how many times each category was observed multiplied by the total number of combinations of that number of each category from all n trials. We use the multinomial coefficient, $\frac{n!}{x_1!\cdots x_c!}$, which gives the total number of ways to choose each category a specific number of times (assuming the counts of each category we choose adds up to the total number of experiments n, that is, $x_1 + x_2 + ... + x_c = n$). Let $\mathbf{X}$ be the random vector $\mathbf{X} = (X_1, X_2, ...X_c)$. Therefore,

$$P(\mathbf{X}) = P(X_1 = x_1, \ldots, X_c = x_c) = \frac{n!}{x_1!\cdots x_c!}\prod_{i=1}^{c} p_i^{x_i}$$

Dirichlet-multinomial distribution

We last arrive at the Dirichlet-Binomial. The Dirichlet-Multinomial is to the Multinomial what the Beta-Binomial is to the Binomial. It's just the generalization to the case where the distribution of $\mathbf{p} = (p_1, p_2, ..., p_c)$ (i.e., the probabilities of each category) follows a multivariate Beta distribution, paramaterized by $\alpha = (\alpha_1, \alpha_2, ..., \alpha_c)$ which is commonly referred to as a Dirichlet distribution. Applying similar reasoning as in the Beta-Binomial case, it can be shown that the PMF of

$$\mathbf{X} \sim \mathrm{DirMult}(n, \boldsymbol{\alpha})$$

is given by

$$P(\mathbf{x}) = \binom{n}{\mathbf{x}} \frac{B(\mathbf{x} + \boldsymbol{\alpha})}{B(\boldsymbol{\alpha})}$$

Recall that we no longer are using the notation of k successes. Each outcome of the random variable X_i occurs x_i times.

Markov-binomial distribution

Up until now, we have been assuming independence between trials. To move beyond this assumption, we can consider a Markov binomial, which is derived from a two-state Markov chain. Thus, we are now interested in modelling the probability of success conditioned on the probability of success in the previous trial. Actually, there are four possible cases:

$$P = \begin{bmatrix} P_{00} & P_{01} \\ P_{10} & P_{11} \end{bmatrix} = \begin{bmatrix} P(\text{failure} \to \text{failure}) & P(\text{failure} \to \text{success}) \\ P(\text{success} \to \text{failure}) & P(\text{success} \to \text{success}) \end{bmatrix}$$

Its PMF, for the probability of k successes after n trials, $P(X = k)$ does not have a closed-form expression. We would need to accout for the sequence of state transitions that occured, while keeping track of the total number of successes. We could simulate this, for instance, by letting our Markov chain run given some initial state probabilities π_1, π_2, a transition matrix P, over sum number of trials equal to the number of iterations n. We could count the number of successes obtained in each simulation, and then plot them.

For more complex dependence relationships in non-iid categorical distributions, we may consider using more advanced probibalistic graphical models, i.e., Bayesian networks, hidden Markov models (HMMs), etc.

Summary

All of these distributions have some underlying unity: considering the sum of random variables with a categorical number of outcomes. Sometime the random variables are iid, sometimes there are more than two categories, or we do not know the probability of each category before. But all of them are fundamentally related. The distributions are summarized in the chart below.

Table 1: Comparison of distributions arising from categorical trials

Distribution	Trials	Outcomes	Dependency	Parameters
Bernoulli	1	{0,1}	None	p
Binomial	n iid	{0,1}	Independent	n, p
Negative binomial	Variable	{0,1}	Independent	r, p
Geometric	Variable	{0,1}	Independent	p
Poisson-binomial	n non-iid	{0,1}	Independent	$p_1, \dots, p_n$
Beta-binomial	n iid	{0,1}	Depends on p	n, α, β
Multinomial	n iid	{1, …, c}	Independent	$n, p_1, \dots, p_c$
Dirichlet-multinomial	n iid	{1, …, c}	Depends on p	$n, \alpha_1, \dots, \alpha_c$
Markov-binomial	n	{0,1}	Markov property	P, π