# COMMON PROBABILITY DISTRIBUTIONS

SHANSHAN DING

We start with discrete distributions.

**1 Binomial.** The binomial distribution $\mathrm{Binom}(n,p)$ is the distribution on the number of successes in $n$ independent and identically distributed Bernoulli trials, where the probability of success in each trial is $p$. The **Bernoulli distribution** is the special case when $n=1$. For a random variable $X \sim \mathrm{Binom}(n,p)$, the probability mass function has support[1] $\{0,1,\ldots,n\}$ and is given by

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

The term $p^k(1-p)^{n-k}$ is the probability that, given a partition of $n$ trials into groups A and B of $k$ and $n-k$ trials respectively, all trials in $A$ are successes and all trials in $B$ are failures, and the binomial coefficient $\binom{n}{k} = n!/k!(n-k)!$ is the number of such partitions.

Recall that

$$\mathbb{E}\left(\sum X_i\right) = \sum \mathbb{E}(X_i) \tag{2}$$

for arbitrary random variables, and that

$$\mathrm{Var}\left(\sum X_i\right) = \sum \mathrm{Var}(X_i) \tag{3}$$

for independent random variables. Since the mean of $X \sim \mathrm{Binom}(1,p)$ is by definition $p$, the mean of $X \sim \mathrm{Binom}(n,p)$ is $np$. The variance of $X \sim \mathrm{Binom}(1,p)$ is easily seen to be $p(1-p)$, and thus the variance of $X \sim \mathrm{Binom}(n,p)$ is $np(1-p)$.

**2 Multinomial.** The multinomial distribution generalizes the binomial distribution. Here each trial has $r$ possible outcomes with probabilities $p_1,\ldots,p_r$ respectively. The pmf of the multinomial distribution is

$$\mathbb{P}(X_1=k_1,\ldots,X_r=k_r) = \frac{n!}{k_1!\cdots k_r!} p_1^{k_1}\cdots p_r^{k_r}, \tag{4}$$

where the $k_i$'s are non-negative and sum to $n$. Like with the binomial distribution, $\mathbb{E}(X_i) = np_i$ and $\mathrm{Var}(X_i) = np_i(1-p_i)$. Additionally, $\mathrm{Cov}(X_i,X_j) = -np_ip_j$, the proof of which we leave as a simple exercise.

The $n=1$ case of the multinomial distribution is called the **categorical distribution**.

[1] The support of a function is the set on which the function is not zero-valued.

**3 Poisson.** The author's personal favorite, the Poisson distribution is the distribution on the number of times an event occurs during a fixed length of time or space, where

(1) The numbers of events occurring during two disjoint intervals are independent.
(2) The probability of an event occurring during a small interval is proportional to the length of the interval. More formally, the probability of an occurrence in $[t, t+h)$ is $ch + o(h)$.[2]
(3) There are no simultaneous events. More formally, the probability of multiple events in $[t, t+h)$ is $o(h)$.

For $X \sim \mathcal{P}(\lambda)$, where $\lambda$ is the average number of events, the pmf of $X$ is given by

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{5}$$

Straightforward but tedious computations show that $\mathrm{Var}(X) = \mathbb{E}(X) = \lambda$. Moreover, using characteristic functions, one can show that the sum of two Poissons is again Poisson, namely that if $X_1 \sim \mathcal{P}(\lambda_1)$ and $X_2 \sim \mathcal{P}(\lambda_2)$, then $X_1 + X_2 \sim \mathcal{P}(\lambda_1 + \lambda_2)$.

That mean is equal to variance is an important property of the Poisson. The author has worked on problems where a family of distributions was identified as Poisson by observing this property. Also, homoscedasticity, where a vector of random variables has the same variance, should not be assumed for Poissons for this reason.

Due to criterion 3 above, the Poisson distribution is sometimes known as the "law of small numbers". In particular, we can divide an interval into $n$ subintervals small enough that there are no multiple events in the same subinterval and think of each subinterval as a Bernoulli trial with probability of success $\lambda/n$. This leads to the following:

**Theorem** (Poisson limit)**.** *For fixed* $\lambda$*,* $\lim_{n\to\infty} \mathrm{Binom}(n, \lambda/n) \to \mathcal{P}(\lambda)$.

*Proof.* Write

$$\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k}{k!}\frac{n!}{(n-k)!n^k}\left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}, \tag{6}$$

and observe that

$$\frac{n!}{(n-k)!n^k} = \frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1, \tag{7}$$

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}, \text{and} \tag{8}$$

$$\left(1 - \frac{\lambda}{n}\right)^{-k} \to 1. \tag{9}$$

Hence

$$\lim_{n\to\infty} \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}, \tag{10}$$

as was to be shown. □

[2] Here we say that $f(x)$ is $o(g(x))$ if $\lim_{x\to 0} f(x)/g(x) = 0$, i.e. $f(x)$ goes to zero faster than $g(x)$ does.

*Remark.* In fact, we can relax the assumption of fixed $\lambda$: if $n \to \infty$ and $p \to 0$ such that $np \to \lambda$, $\text{Binom}(n, p) \to \mathcal{P}(\lambda)$. The corresponding proof, however, is considerably more difficult.

*Remark.* Compare this to the normal approximation of the binomial, where $p$ remains constant as $n \to \infty$. In practice, Poisson is used instead of the normal to approximate limit of the binomial when $p$ is small (e.g. $\leq 0.05$) and $np$ is bounded (e.g. $\leq 10$).

In the late 19th century the Polish-Russian scholar (proto-data scientist?) Ladislaus Bortkiewicz used the Poisson distribution to model the number of soldiers killed by horse kicks in the Prussian army. The term "law of small numbers" is attributed to him.

Next we look at a couple of continuous distributions, skipping the familiar **uniform** and **normal** distributions.

4 **Exponential.** The exponential distribution describes time elapsed between Poisson events, which is sometimes easier to record than the number of events itself. It is the continuous analogue of the **geometric distribution** (the distribution on the number of Bernoulli trials needed to get one success). For a random variable X with exponential distribution, the pdf can be derived as follows: let $c$ be the rate parameter, namely the expected number of events during a unit of time, and let $Y \sim \text{Poi}(ct)$. Then

$$\mathbb{P}(X > t) = \mathbb{P}(Y = 0) = e^{-ct}, \tag{11}$$

thus the cdf of $X$ is $1 - e^{-ct}$, and the pdf is its derivative,

$$f(t; \alpha) = ce^{-ct} \quad t \geq 0. \tag{12}$$

Furthermore, integration by parts gives that $\mathbb{E}(X) = 1/c$ and $\text{Var}(X) = 1/c^2$.

Note that by Bayes' rule,

$$\begin{aligned} \mathbb{P}(X > s + t | X > s) &= \frac{\mathbb{P}(X > s | X > s + t)\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} \\ &= \frac{e^{-c(s+t)}}{e^{-cs}} = e^{-ct} = \mathbb{P}(X > t). \end{aligned} \tag{13}$$

In words, at the present moment, the amount of time until the next event is independent of how much time has already passed since the last event. This is called the memoryless property of the exponential.

**Terminology disclaimer:** A family of exponential distributions is not the same thing as the class of exponential families of distributions, a class of distributions with convenient statistical properties that include the exponential distributions but also many other types of distributions.

5 **Beta.** The beta distribution is a family of two-parameter distributions defined by the pdf

$$f(x; \alpha, \beta) = Cx^{\alpha-1}(1 - x)^{\beta-1} \quad 0 \leq x \leq 1, \tag{14}$$

where, since the pdf must integrate to 1,

$$C = \frac{1}{\int_0^1 x^{\alpha-1}(1 - x)^{\beta-1} dx}. \tag{15}$$

The beta distribution derives its name from the integral $B(\alpha,\beta)=\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$, known as the beta function. Using integration by parts, one can show that

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{16}$$

where $\Gamma$ is the gamma function

$$\Gamma(t)=\int_0^\infty x^{t-1}e^{-x}dx, \text{ with } \Gamma(t)|_{\mathbb{Z}^+}=(t-1)!. \tag{17}$$

It is easy to compute all moments of $X\sim \mathrm{Beta}(\alpha,\beta)$:

$$\mathbb{E}(X^n)=\int_0^1 \frac{x^n x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}dx=\frac{B(\alpha+n,\beta)}{B(\alpha,\beta)}=\frac{\Gamma(\alpha+n)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+n)}, \tag{18}$$

and in particular, the identity $\Gamma(t+1)=t\Gamma(t)$ gives us that

$$\mathbb{E}(X)=\frac{\alpha}{\alpha+\beta} \text{ and } \mathrm{Var}(X)=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)^2}. \tag{19}$$

Setting $\alpha=\beta=1$ yields the uniform distribution on $[0,1]$. Setting $\alpha=\beta=1/2$ yields the **arcsine distribution** on $[0,1]$, so-named because its cdf is

$$F(x)=\frac{2}{\pi}\arcsin(\sqrt{x}). \tag{20}$$

The arcsine distribution is beyond the scope of these notes, but it is a fascinating topic with applications in finance, sports, and number theory.

The beta distribution features prominently in Bayesian statistics as the conjugate prior for the binomial distribution. Let

$$p(\theta)=\mathrm{Beta}(\alpha,\beta)=\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \tag{21}$$

be the prior on the Bernoulli parameter. After observing $k$ successes in $n$ trials, we update our belief on $\theta$:

$$\begin{aligned} p(\theta|\text{data}) &\propto p(\text{data}|\theta)p(\theta)\\ &\propto \theta^k(1-\theta)^{n-k}\theta^{\alpha-1}(1-\theta)^{\beta-1}\\ &= C\theta^{\alpha+k-1}(1-\theta)^{\beta+n-k-1}. \end{aligned} \tag{22}$$

Since this integrates to 1, we see that $C=1/B(\alpha+k,\beta+n-k)$, and thus

$$p(\theta|\text{data})=\mathrm{Beta}(\alpha+k,\beta+n-k). \tag{23}$$

That the prior and the posterior belong to the same family of distributions is highly desirable, for otherwise the posterior may not have a closed-form expression and Bayesian updating would require difficult numerical integrations.