BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2015

3 Probability distributions and their stories

We have talked about the three levels of building a model in the biological sciences.

Model 1 A cartoon or verbal phenomenological description of the system of interest.

Model 2 A mathematization of Model 1.

Model 3 A description of what we expect from an experiment, given our model. In the Bayesian context, this is specification of the likelihood and prior.

We have talked briefly about specifying priors. We endeavor to be uninformative when we do not know much about model parameters or about the hypothesis in question. We found in the homework that while we should try to be a uninformative as possible (using maximum entropy ideas), we can be a bit sloppy with our prior definitions, provided we have many data so that the likelihood can overcome any prior sloppiness.

To specify a likelihood, we need to be careful, as this is very much at the heart of our model. Specifying the likelihood amounts to choosing a probability distribution that describes how the data will look under your model. In many cases, you need to derive the likelihood (or even numerically compute it when it cannot be written in closed form). In many practical cases, though, the choice of likelihood is among standard probability distributions. These distributions all have "stories" associated with them. If your data and model match the story of a distribution, you know that this is the distribution to choose for your likelihood.

3.1 Review on probability distributions

Before we begin talking about distributions, let's remind ourselves what probability distributions are. We cut some corners in our definitions here, but these definitions are functional for most of our data analysis purposes.

A **probability mass function** (PMF), $P(x)$, describes the probability of a discrete variable obtaining value x. The variable x takes on discrete values, so the **normalization condition** is

$$\sum_x P(x) = 1. \tag{3.1}$$

A **probability distribution function** (PDF), which we shall also call $P(x)$, is defined

such that the probability that a continuous variable x is $a \leq x \leq b$ is

$$\int_a^b \mathrm{d}x\, P(x). \tag{3.2}$$

3.1.1 Moments

PMFs and PDFs have **moments**. The way they are defined can vary, but we will define the nth moment for a PMF as

$$\langle x^n \rangle = \sum_i x_i^n P(x_i), \tag{3.3}$$

and for a PDF as

$$\langle x^n \rangle = \int \mathrm{d}x\, x^n P(x). \tag{3.4}$$

These moments are often used to compute summary statistics.

$$\text{mean} = \langle x \rangle \tag{3.5}$$

$$\text{variance} = \left\langle x^2 \right\rangle - \langle x \rangle^2 = \left\langle (x - \langle x \rangle)^2 \right\rangle \tag{3.6}$$

$$\text{skew} = \frac{\langle (x - \langle x \rangle)^3 \rangle}{\langle (x - \langle x \rangle)^2 \rangle^{\frac{3}{2}}} \tag{3.7}$$

$$\text{Pearson kurtosis} = \frac{\langle (x - \langle x \rangle)^4 \rangle}{\langle (x - \langle x \rangle)^2 \rangle^2} \tag{3.8}$$

$$\text{Fisher kurtosis} = \frac{\langle (x - \langle x \rangle)^4 \rangle}{\langle (x - \langle x \rangle)^2 \rangle^2} - 3. \tag{3.9}$$

We will present PMFs and PDFs for distributions below. We only show the univariate forms; the multivariate version are easily derived or looked up.

3.1.2 Sampling

Given that we know a probability distribution, we can take **samples** out of it. This means that we can randomly draw numbers and the probability that we draw a certain number x is proportional to the PMF or PDF, $P(x)$. As we will soon see, sampling out of a distribution is often easier than computing the distribution over a range of values because many of those values are zero.

3.2 Discrete distributions and their stories

3.2.1 Bernoulli

Story. A single trial with either a success ($x = 1$) or failure ($x = 0$) is performed. The Bernoulli distribution defines the probability of getting each outcome.

Parameter. The Bernoulli distribution is parametrized by a single value, p, the probability that the trial is successful. These trials are called **Bernoulli trials**.

Example. Check to see if a given bacteria is competent, given that it has probability p of being competent.

Probability mass function.

$$P(x; p) = \begin{cases} 1 - p & x = 0 \\ p & x = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

3.2.2 Geometric

Story. We perform a series of Bernoulli trials until we get a success. We have k failures before the success.

Parameter. The Geometric distribution is parametrized by a single value, p, the probability that the Bernoulli trial is successful.

Example. Consider actin polymerization. At each time step, an actin monomer may add to the filament ("failure"), or an actin monomer may fall off ("success") with (usually very low) probability p. The length of actin filaments are geometrically distributed.

Probability mass function.

$$P(k; p) = (1 - p)^k p. \tag{3.11}$$

The Geometric distribution is only defined for non-negative integer k.

3.2.3 Negative binomial

Story. We perform a series of Bernoulli trials until we get r successes. The number of failures, n, before we get r successes is negative binomially distributed.

Parameters. There are two parameters: the probability p of success for each Bernoulli trial, and the desired number of successes, r.

Example. Bursty gene expression can give mRNA count distributions that are negative binomially distributed. Here, "success" is that a burst in gene expression stops. So, the parameter p is related to the length of a burst in expression (lower p means a longer burst). The parameter r is related to the frequency of the bursts. If multiple bursts are possible within the lifetime of mRNA, then $r > 1$. Then, the number of "failures" is the number of mRNA transcripts that are made in the characteristic lifetime of mRNA.

Probability mass function.

$$P(n; r, p) = \binom{n + r - 1}{r - 1} p^r (1 - p)^n. \tag{3.12}$$

Here, we use a combinatorial notation;

$$\binom{n + r - 1}{r - 1} = \frac{(n + r - 1)!}{(r - 1)!\, n!}. \tag{3.13}$$

Note that if $r = 1$, this distribution reduces to the Geometric distribution.

3.2.4 Binomial

Story. We perform n Bernoulli trials. The number of successes, k, is binomially distributed.

Parameters. There are two parameters: the probability p of success for each Bernoulli trial, and the number of trials, n.

Example. Distribution of plasmids between daughter cells in cell division. Each of the n plasmids as a chance p of being in daughter cell 1 ("success"). The number of plasmids, k, in daughter cell 1 is binomially distributed.

Probability mass function.

$$P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{3.14}$$

3.2.5 Poisson

Story. Rare events occur with a rate λ per unit time. There is no "memory" of previous events; i.e., that rate is independent of time. There are k event that occur in unit time.

Parameter. The single parameter is the rate λ of the rare events occurring.

Example. The number of mutations in a strand of DNA per unit length (since mutations are rare) are Poisson distributed.

Probability mass function.

$$P(k; \lambda) = \frac{\lambda^k}{k!} \, \mathrm{e}^{-\lambda}. \tag{3.15}$$

Note. The Poisson distribution is a limit of the binomial distribution in which the number of trials goes to infinity, but the expected number of successes, np, stays fixed. Thus,

$$P_{\text{Poisson}}(k; \lambda) \approx P_{\text{Binomial}}(k; n, p), \tag{3.16}$$

with $\lambda = np$. Considering the biological example of mutations, this is binomially distributed: There are n bases, each with a probability p of mutation, so the number of mutations, k is binomially distributed. Since p is small, it is approximately Poisson distributed.

3.2.6 Hypergeometric

Story. Consider an urn with w white balls and b black balls. Draw n balls from this urn without replacement. The number white balls drawn, k is hypergeometrically distributed.

Parameters. There are three parameters: the number of draws n, the number of white balls w, and the number of black balls b.

Example. There are N finches on an island, and n_t of them are tagged. You capture n finches. The number of tagged finches k is hypergeometrically distributed, $P(k; n_t, N - n_t, n)$, as defined below.

Probability mass function.

$$P(k; w, b, n) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}. \tag{3.17}$$

Note. This distribution is analogous to the Binomial distribution, except that the Binomial distribution describes draws from an urn with replacement. In the analogy, $p = w/(w + b)$.

3.3 Continuous distributions and their stories

3.3.1 Uniform

Story. Any outcome in a given range has equal probability.

Parameters. The Uniform distribution is not defined on an infinite or semi-infinite domain, so bounds, $x_{\min}$ and $x_{\max}$ are necessary parameters.

Example. Anything in which all possibilities are equally likely.

Probability density function.

$$P(x; x_{\min}, x_{\max}) = \begin{cases} \frac{1}{x_{\max} - x_{\min}} & x_{\min} \leq x \leq x_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

3.3.2 Gaussian (a.k.a. Normal)

Story. Any quantity that emerges from a large number of subprocesses tends to be Gaussian distributed provided none of the subprocesses is very broadly distributed.

Parameters. The Gaussian distribution has two parameters, the mean μ, which determines the location of its peak, and the standard deviation σ, which is strictly positive (the $\sigma \to 0$ limit defines a Dirac delta function) and determines the width of the peak.

Example. We measure the length of many *C. elegans* eggs. The lengths are Gaussian distributed.

Probability density function.

$$P(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-(x-\mu)^2/2\sigma^2}. \tag{3.19}$$

Note. This is a limiting distribution in the sense of the central limit theorem, but also in that many distributions have a Gaussian distribution as a limit. This is seen by formally taking limits of, e.g., the Gamma, Student-t, Binomial distributions, which allows direct comparison of parameters.

3.3.3 Log-normal

Story. If $\ln x$ is Gaussian distributed, x is log-normally distributed.

Parameters. As for a Gaussian, there are two parameters, the mean logarithm, $\ln \mu$, and the variance σ^2.

Example. A measure of fold change in gene expression can be log-normally distributed.

Probability density function.

$$P(x; \mu, \sigma) = \frac{1}{x\sqrt{2\pi\sigma^2}} \, \mathrm{e}^{-(\ln x - \ln \mu)^2/2\sigma^2}. \tag{3.20}$$

3.3.4 Von Mises

Story. Gaussian, except on a periodic domain.

Parameters. As for a Gaussian, with μ being the location of the peak, and β being analogous to the variance.

Example. Repeated measurements on a periodic domain, e.g., the location of an ingression along the azimuthal angle of a developing embryo.

Probability density function.

$$P(\theta; \mu, \beta) = \frac{1}{2\pi I_0(\beta)} \, \mathrm{e}^{\beta \cos(\theta - \mu)}, \tag{3.21}$$

where $I_0(\beta)$ is a modified Bessel function of the first kind.

3.3.5 Chi-square

Story. If X_1, X_2, ..., X_n are Gaussian distributed, $X_1^2 + X_2^2 + \cdots + X_n^2$ is χ^2-distributed.

Parameters. There is only one parameter, the degrees of freedom n.

Probability density function.

$$P(x; n) \equiv \chi_n^2(x; n) = \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)} \, x^{\frac{n}{2}-1} \, \mathrm{e}^{-x/2}. \tag{3.22}$$

3.3.6 Student -t/Cauchy

Story. We get this distribution whenever we marginalize an unknown σ out of a Gaussian distribution with a Jeffreys distribution for σ.

Parameters. The Student-t distribution is peaked, and its peak is located at m. The peak's width is dictated by parameter s. Finally, we define the "degrees of freedom" as n.

Example. The story says it all!

Probability density function.

$$P(x; m, s, n) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} \frac{\sqrt{\pi n s^2}}{\left(1 + \frac{(x-m)^2}{ns^2}\right)^{\frac{n+1}{2}}}. \tag{3.23}$$

Note. For $n \to \infty$, we get a Gaussian distribution. When $n = 1$, we get the **Cauchy distribution**,

$$P(x; m, s) = \left[\pi s \left(\frac{x-m}{s}\right)^2\right]^{-1}. \tag{3.24}$$

3.3.7 Aside: Poisson processes

A **Poisson** process is a sequence of "arrivals" or events at different points on a timeline such that the number of point in a particular interval are Poisson distributed. This means that the amount of time between any two arrivals is independent of all other inter-arrival times.

3.3.8 Exponential

Story. This is the waiting time for an arrival from a Poisson process. I.e., the inter-arrival time of a Poisson process is exponentially distributed.

Parameter. The single parameter is the average arrival rate, r.

Example. The time between conformational switches in a protein is exponentially distributed (under simple mass action kinetics).

Probability density function.

$$P(x; r) = r\mathrm{e}^{-rx}. \tag{3.25}$$

Note. The Exponential distribution is the continuous analog of the Geometric distribution. The "rate" in the Exponential distribution is analogous to the probability of success of the Bernoulli trial.

3.3.9 Gamma

Story. The amount of time we have to wait for a arrivals of a Poisson process. More concretely, if we have events, X_1, X_2, ..., X_a that are exponentially distributed, $X_1 + X_2 + \cdots + X_a$ is gamma distributed.

Parameters. The number of arrivals, a, and the rate of arrivals, r.

Example. Any multistep process where each step happens at the same rate. This is common in molecular rearrangements, and we will use it in class to describe the nature of processes triggering microtubule catastrophe.

Probability density function.

$$P(x; r) = \frac{1}{\Gamma(a)} \frac{(rx)^a}{x} \mathrm{e}^{-rx}, \tag{3.26}$$

where $\Gamma(a)$ is the gamma function.

Note. The Gamma distribution is the continuous analog of the Negative Binomial distribution.

3.3.10 Weibull

Story. Distribution of $x = y^\beta$ if y is exponentially distributed. For $\beta > 1$, the longer we have waited, the more likely the event is to come, and vice versa for $\beta < 1$.

Parameters. There are two parameters, both strictly positive: the shape parameter β, which dictates the shape of the curve, and the scale parameter λ, which dictates the rate of arrivals of the event.

Example. This is a model for aging. The longer an organism lives, the more likely it is to die.

Probability density function.

$$P(x;\lambda,\beta) = \frac{\beta}{\lambda}\left(\frac{x}{\lambda}\right)^{\beta-1} \mathrm{e}^{-(x/\lambda)^{\beta}}. \tag{3.27}$$