11.5 Commonly used distributions

In this course, we'll primarily be discussing bioinformatics where some commonly used discrete probability distribution functions are: Bernoulli, geometric, binomial, uniform and Poisson. Commonly used continuous distributions are uniform, normal (Gaussian), exponential, and gamma. Those are briefly described below. For more thorough descriptions, refer to any decent statistics text or, more simply, the relevant Wikipedia entries.

11.5.1 Bernoulli distribution

A random variable X with a *Bernoulli distribution* takes values 0 and 1. It takes the value 1 on a 'success' which occurs with probability p where $0 \leq p \leq 1$. It takes value 0 on a failure with probability $q = 1 - p$. Thus it has the single parameter p. If X is Bernoulli, $E[X] = q \cdot 0 + p \cdot 1 = p$ and $\text{Var}(X) = E[X^2] - E[X]^2 = q \cdot 0^2 + p \cdot 1^2 - p^2 = pq$.

11.5.2 Geometric distribution

X has a *geometric distribution* when it is the number of Bernoulli trials that fail before the first success. It therefore takes values in $\{0, 1, 2, 3, \ldots\}$. If the Bernoulli trials have probability p of success, the pdf for X is $P(X = x) = (1 - p)^x p$. If X is geometric,

$$E[X] = \frac{q}{p} \text{ and } \text{Var}(X) = \frac{q}{p^2}.$$

Note that the Geometric distribution can be defined instead as the total number of trials required to get a single success. This version of the geometric can only take values in $\{1, 2, 3, \ldots\}$. The pdf, mean and variance all need to be adjusted accordingly.

11.5.3 Binomial distribution

X has a *binomial distribution* when it represents the number of successes in n Bernoulli trials. There are thus two parameters required to describe a binomial random variable: n, the number of Bernoulli trials undertaken, and p, the probability of success in the Bernoulli trials. The pdf for X is

$$f(x) = P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n.$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$. For a binomial variable X,

$$E[X] = np \text{ and } Var[X] = np(1 - p) = npq.$$

We write $X \sim Bin(n, p)$ when X has a binomial distribution with parameters n and p.

11.5.4 Poisson distribution

The *Poisson distribution* is used to model the number of rare events that occur in a period of time. The events are considered to occur independently of each other. The distribution has a single parameter, λ, and probability density function

$$f(x) = \exp(-\lambda)\frac{\lambda^x}{x!} \text{ for } x = 0, 1, 2, 3, \ldots.$$

If X is Poisson,

$$E[X] = \lambda \text{ and } Var[X] = \lambda.$$

We write $X \sim Poiss(\lambda)$ when X has a Poisson distribution with parameter λ.

11.5.5 Uniform distribution (discrete or continuous)

Under the *uniform distribution*, all possible values are equally likely. So if X is discrete and takes n possible values, $P(X = x_i) = 1/n$ for all x_i.

If X is continuous and uniform over the interval $[a, b]$, the density function is $f(x) = \frac{1}{b-a}$. In this case, write $X \sim U([a, b])$.

11.5.6 Normal distribution

The *Normal*, or Gaussian, distribution, with mean μ and variance σ^2, $(\mu \in \mathbb{R}; \sigma > 0)$ has density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

We write $X \sim N(\mu, \sigma^2)$.

The normal distribution is a widely used distribution in statistical modelling for a number of reasons. A primary reason is that it arises as a consequence of the central limit theorem which says that (under a few weak assumptions) the sum of a set of identical random variables is well approximated by a normal distribution. Thus when random effects all add together, they often result in a normal distribution. Measurement error terms are typically modelled as normally distributed.

11.5.7 Exponential distribution

The *Exponential* distribution describes the time between rare events so always takes non-negative values. It has a single parameter, λ known as the rate and has density function

$$f(x) = \lambda e^{-\lambda x},$$

where $x \geq 0$. If X is exponentially distributed,

$$E[X] = \frac{1}{\lambda} \text{ and } \text{Var}(X) = \frac{1}{\lambda^2}.$$

Write $X \sim Exp(\lambda)$.

11.5.8 Gamma distribution

The *Gamma* distribution arises as the sum of a number of exponentials. It has two parameters, k and θ, called the shape and scale, respectively. These parameters can be used to specify the mean and variance of the distribution.

$$f(x) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} \exp(-x/\theta) \text{ for } x > 0,$$

where $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t}\, dt$ is the gamma function (the extension of the factorial function, $k!$, to all real numbers). The mean and variance of a gamma distributed random variable X is

$$E[X] = k\theta \text{ and } \text{Var}(X) = k\theta^2.$$

Write $X \sim Gamma(k, \theta)$.

Note that the gamma distribution has different parametrisations which result in different looking (but mathematically equivalent) expressions for the density, mean and variance — be sure to check which parametrisation is being used.