Common probability distributions

1 Discrete distributions

1.1 Uniform

The uniform distribution is the simplest discrete probability distribution. It assigns equal probability to N different outcomes, usually represented with numbers $1, 2, \ldots, N$.

$$\begin{aligned} X &\sim \text{Uniform}(N) \\ \text{P}(X = x|N) &= 1/N \quad \text{for } x = 1, 2, \ldots, N \\ \text{E}[X] &= \frac{N+1}{2} \\ \text{Var}[X] &= \frac{N^2-1}{12} \end{aligned}$$

One common example is the outcome of throwing a fair six-sided die ($N = 6$).

1.2 Bernoulli

The Bernoulli distribution is used for binary outcomes, coded as 0 and 1. It has one parameter p, which is the probability of "success" or 1.

$$\begin{aligned} X &\sim \text{Bern}(p) \\ \text{P}(X = x|p) &= p^x(1-p)^{1-x} \quad \text{for } x = 0, 1 \\ \text{E}[X] &= p \\ \text{Var}[X] &= p(1-p) \end{aligned}$$

One common example is the outcome of flipping a fair coin ($p = 0.5$).

1.3 Binomial

The binomial distribution counts the number of "successes" in n independent Bernoulli trials (each with the same probability of success). Thus if $X_1, \ldots, X_n$ are independent Bernoulli(p)

random variables, then $Y = \sum_{i=1}^{n} X_i$ is binomial distributed.

$$\begin{aligned} Y &\sim \text{Binom}(n, p) \\ \text{P}(Y = y|n, p) &= \binom{n}{y} p^y (1-p)^{n-y} \quad \text{for } y = 0, 1, \ldots, n \\ \text{E}[Y] &= np \\ \text{Var}[Y] &= np(1-p) \end{aligned}$$

where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$.

1.4 Poisson

The Poisson distribution is used for counts, and arises in a variety of situations. The parameter $\lambda > 0$ is the rate at which we expect to observe the thing we are counting.

$$\begin{aligned} X &\sim \text{Pois}(\lambda) \\ P(X = x|\lambda) &= \frac{\lambda^x \exp(-\lambda)}{x!} \text{ for } x = 0, 1, 2, \ldots \\ \text{E}[X] &= \lambda \\ \text{Var}[X] &= \lambda \end{aligned}$$

A Poisson process is a process wherein events occur on average at rate λ, events occur one at a time, and events occur independently of each other.

Example: Significant earthquakes occur in the Western United States approximately following a Poisson process with rate of two earthquakes per week. What is the probability there will be at least 3 earthquakes in the next two weeks? Answer: the rate per two weeks is $2 \times 2 = 4$, so let $X \sim \text{Pois}(4)$ and we want to know $P(X \geq 3) = 1 - P(X \leq 2) = 1 - P(X = 0) - P(X = 1) - P(X = 2) = 1 - e^{-4} - 4e^{-4} - \frac{4^2 e^{-4}}{2} = 1 - 13e^{-4} = 0.762$. Note that $0! = 1$ by definition.

1.5 Geometric

The geometric distribution is the number of failures before obtaining the first success, i.e., the number of Bernoulli failures until a success is observed, such as the first head when

flipping a coin. It takes values on the positive integers starting with 0 (alternatively, we could count total trials until first success, in which case we would start with 1).

$$\begin{aligned} X &\sim \text{Geo}(p) \\ P(X = x|p) &= p(1-p)^x \text{ for } x = 0, 1, 2, \ldots \\ \text{E}[X] &= \frac{1-p}{p} \end{aligned}$$

If the probability of getting a success is p, then the expected number of trials until the first success is $1/p$ and the expected number of failures until the first success is $(1-p)/p$.

Example: What is the probability that we flip a fair coin four times and don't see any heads? This is the same as asking what is $P(X > 3)$ where $X \sim Geo(1/2)$. $P(X > 3) = 1 - \text{P}(X = 0) - P(X = 1) - \text{P}(X = 2) - \text{P}(X = 3) = 1 - (1/2) \cdot 1 - (1/2)(1/2) - (1/2)(1/2)^2 - (1/2)(1/2)^3 = 1/16$.

1.6 Negative Binomial

The negative binomial distribution extends the geometric distribution to model the number of failures before achieving the rth success. It takes values on the positive integers starting with 0.

$$\begin{aligned} Y &\sim \text{NegBinom}(r, p) \\ P(Y = y|r, p) &= \binom{r+y-1}{y} p^r (1-p)^y \quad \text{for } y = 0, 1, 2, \ldots \\ \text{E}[Y] &= \frac{r(1-p)}{p} \\ \text{Var}[Y] &= \frac{r(1-p)}{p^2} \end{aligned}$$

Note that the geometric distribution is a special case of the negative binomial distribution where $r = 1$.

Because $0 < p < 1$, we have $\text{E}[Y] < \text{Var}[Y]$. This makes the negative binomial a popular alternative to the Poisson when modeling counts with high variance (recall, that the mean equals the variance for Poisson distributed variables).

1.7 Multinomial

Another generalization of the Bernoulli and the binomial is the multinomial distribution, which is like a binomial when there are more than two possible outcomes. Suppose we have n trials and there are k different possible outcomes which occur with probabilities $p_1, \ldots, p_k$. For example, we are rolling a six-sided die that might be loaded so that the sides are not equally likely, then n is the total number of rolls, $k = 6$, p_1 is the probability of rolling a one, and we denote by $x_1, \ldots, x_6$ a possible outcome for the number of times we observe rolls of each of one through six, where $\sum_{i=1}^{6} x_i = n$ and $\sum_{i=1}^{6} p_i = 1$.

$$f(x_1, \ldots, x_k | p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} .$$

Recall that $n!$ stands for n factorial, which is the product of n times $n-1$ times $\ldots 1$, e.g., $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. The expected number of observations in category i is np_i.

2 Continuous distributions

2.1 Uniform

The uniform distribution is used for random variables whose possible values are equally likely over an interval. If the interval is (a, b), then the uniform probability density function (PDF) $f(x)$ is flat for all values in that interval and 0 everywhere else.

$$\begin{aligned} X &\sim \text{Uniform}(a, b) \\ f(x|a, b) &= \frac{1}{b-a} I_{\{a \leq x \leq b\}}(x) \\ \text{E}[X] &= \frac{a+b}{2} \\ \text{Var}[X] &= \frac{(b-a)^2}{12} \end{aligned}$$

The standard uniform distribution is obtained when $a = 0$ and $b = 1$.

2.2 Beta

The beta distribution is used for random variables which take on values between 0 and 1. For this reason (and other reasons we will see later in the course), the beta distribution is

commonly used to model probabilities.

$$\begin{aligned} X &\sim \text{Beta}(\alpha, \beta) \\ f(x|\alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{\{0<x<1\}}(x) \\ \text{E}[X] &= \frac{\alpha}{\alpha+\beta} \\ \text{Var}[X] &= \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function introduced with the gamma distribution. Note also that $\alpha > 0$ and $\beta > 0$. The standard Uniform$(0, 1)$ distribution is a special case of the beta distribution with $\alpha = \beta = 1$.

2.3 Exponential

The exponential distribution is often used to model the waiting time between random events. Indeed, if the waiting times between successive events are independent from an Exp(λ) distribution, then for any fixed time window of length t, the number of events occurring in that window will follow a Poisson distribution with mean $t\lambda$.

$$\begin{aligned} X &\sim \text{Exp}(\lambda) \\ f(x|\lambda) &= \lambda e^{-\lambda x} I_{\{x\geq 0\}}(x) \\ \text{E}[X] &= \frac{1}{\lambda} \\ \text{Var}[X] &= \frac{1}{\lambda^2} \end{aligned}$$

Similar to the Poisson distribution, the parameter λ is interpreted as the rate at which the events occur.

2.4 Double Exponential

The double exponential (or Laplace) distribution generalizes the exponential distribution for a random variable that can be positive or negative. The PDF looks like a pair of back-to-back

exponential PDFs, with a peak at 0.

$$\begin{aligned} X &\sim \text{DExp}(\lambda) \\ f(x|\mu,\tau) &= \frac{\tau}{2} e^{-\tau|x-\mu|} \\ \text{E}[X] &= \mu \\ \text{Var}[X] &= \frac{1}{2\tau^2} \end{aligned}$$

Most of the probability mass for the double exponential distribution is near 0. For this reason, it is commonly used as a "shrinkage prior" for situations where we have a collection of parameters and believe that many of them should be 0, but we don't know which ones are 0. This might arise, for example, with the coefficients in multiple regression.

2.5 Gamma

If $X_1, X_2, \ldots, X_n$ are independent (and identically distributed $\text{Exp}(\lambda)$) waiting times between successive events, then the total waiting time for all n events to occur $Y = \sum_{i=1}^{n} X_i$ will follow a gamma distribution with shape parameter $\alpha = n$ and rate parameter $\beta = \lambda$.

$$\begin{aligned} Y &\sim \text{Gamma}(\alpha, \beta) \\ f(y|\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} I_{\{y \geq 0\}}(y) \\ \text{E}[Y] &= \frac{\alpha}{\beta} \\ \text{Var}[Y] &= \frac{\alpha}{\beta^2} \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function, a generalization of the factorial function which can accept non-integer arguments. If n is a positive integer, then $\Gamma(n) = (n-1)!$. Note also that $\alpha > 0$ and $\beta > 0$.

The exponential distribution is a special case of the gamma distribution with $\alpha = 1$. The gamma distribution commonly appears in statistical problems, as we will see in this course. It is used to model positive-valued, continuous quantities whose distribution is right-skewed. As α increases, the gamma distribution more closely resembles the normal distribution.

2.6 Inverse-Gamma

The inverse-gamma distribution is the conjugate prior for σ^2 in the normal likelihood with known mean. It is also the marginal prior/posterior for σ^2 in the model of Lesson 10.2.

As the name implies, the inverse-gamma distribution is related to the gamma distribution. If $X \sim \text{Gamma}(\alpha, \beta)$, then the random variable $Y = 1/X \sim \text{Inverse-Gamma}(\alpha, \beta)$ where

$$f(y) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{-(\alpha+1)} \exp\left(-\frac{\beta}{y}\right) I_{\{y>0\}}$$
$$\text{E}(Y) = \frac{\beta}{\alpha - 1} \quad \text{for } \alpha > 1\,.$$

The relationship between gamma and inverse-gamma suggest a simple method for simulating draws from the inverse-gamma distribution. First draw X from the $\text{Gamma}(\alpha, \beta)$ distribution and take $Y = 1/X$, which corresponds to a draw from the $\text{Inverse-Gamma}(\alpha, \beta)$.

2.7 Normal

The normal, or Gaussian distribution is one of the most important distributions in statistics. It arises as the limiting distribution of sums (and averages) of random variables. This is due to the central limit theorem. Because of this property, the normal distribution is often used to model the "errors," or unexplained variation of individual observations in regression models.

The standard normal distribution is given by

$$Z \sim \text{N}(0, 1)$$
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$$
$$\text{E}[Z] = 0$$
$$\text{Var}[Z] = 1$$

Now consider $X = \sigma Z + \mu$ where $\sigma > 0$ and μ is any real constant. Then $\text{E}(X) = \text{E}(\sigma Z + \mu) = \sigma\text{E}(Z) + \mu = \sigma \cdot 0 + \mu = \mu$ and $\text{Var}(X) = \text{Var}(\sigma Z + \mu) = \sigma^2\text{Var}(Z) + 0 = \sigma^2 \cdot 1 = \sigma^2$. Then, X follows a normal distribution with mean μ and variance σ^2 (standard deviation σ) denoted as

$$X \sim \text{N}(\mu, \sigma^2)$$
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

The normal distribution is symmetric about the mean μ, and is often described as a "bell-shaped" curve. Although X can take on any real value (positive or negative), more than 99% of the probability mass is concentrated within three standard deviations of the mean.

The normal distribution has several desirable properties. One is that if $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ are independent, then $X_1 + X_2 \sim \mathrm{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Consequently, if we take the average of n independent and identically distributed (iid) normal random variables,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \,,$$

where $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ for $i = 1, 2, \ldots, n$, then

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) . \tag{1}$$

2.8 t

If we have normal data, we can use Equation 1 to help us estimate the mean μ. Reversing the transformation from the previous section, we get

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0, 1) \,. \tag{2}$$

However, we may not know the value of σ. If we estimate it from data, we can replace it with $S = \sqrt{\sum_i (X_i - \bar{X})^2/(n-1)}$, the sample standard deviation. This causes the expression (2) to no longer be distributed as standard normal, but as a standard t distribution with $\nu = n-1$ degrees of freedom.

$$\begin{aligned}
Y &\sim \mathrm{t}_\nu \\
f(y) &= \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\nu\pi}} \left(1 + \frac{y^2}{\nu}\right)^{-(\frac{\nu+1}{2})} \\
\mathrm{E}[Y] &= 0 \ \text{ if } \nu > 1 \\
\mathrm{Var}[Y] &= \frac{\nu}{\nu - 2} \ \text{ if } \nu > 2
\end{aligned}$$

The t distribution is symmetric and resembles the normal distribution, but with thicker tails. As the degrees of freedom increase, the t distribution looks more and more like the standard normal distribution.

2.9 Dirichlet

Just as the beta distribution is the conjugate prior for a binomial likelihood, the Dirichlet distribution is the conjugate prior for the multinomial likelihood. It can be thought of as a multivariate beta distribution for a collection of probabilities (that must sum to 1).

The probability density function for the random variables $Y_1, \ldots, Y_K$ with $Y_k > 0$ and $\sum_{k=1}^{K} Y_k = 1$ is given by

$$f(y_1, y_2, \ldots, y_K | \alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \cdot p_1^{\alpha_1 - 1} \cdot p_2^{\alpha_2 - 1} \cdot \ldots \cdot p_K^{\alpha_K - 1}$$

where $\alpha_k > 0$ for $k = 1, \ldots, K$. The expected value for Y_i is $\alpha_i / \sum_{k=1}^{K} \alpha_k$.

Example: The example for the multinomial distribution describes an experiment to estimate the probabilities associated with a loaded die. Suppose roll the die many times and model the data as multinomial, where x_1 is the number of 1's observed, x_2 is the number of 2's observed, etc., and we use a Dirichlet prior for the probabilities $p_1, \ldots, p_6$ associated with each face of the die. Similar to the binomial-beta model, the posterior distribution for the probabilities is Dirichlet with updated parameters: $\alpha_1 + x_1$, $\alpha_2 + x_2$, ..., $\alpha_6 + x_6$.