Appendix B

More probability distributions and related mathematical constructs

This chapter covers probability distributions and related mathematical constructs that are referenced elsewhere in the book but aren't covered in detail. One of the best places for more detailed information about these and many other important probability distributions is Wikipedia.

B.1 The gamma and beta functions

The GAMMA FUNCTION $\Gamma(x)$, defined for $x > 0$, can be thought of as a generalization of the factorial $x!$. It is defined as

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$$

and is available as a function in most statistical software packages such as `R`. The behavior of the gamma function is simpler than its form may suggest: $\Gamma(1) = 1$, and if $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$. This means that if x is a positive integer, then $\Gamma(x) = (x-1)!$.

The BETA FUNCTION $B(\alpha_1, \alpha_2)$ is defined as a combination of gamma functions:

$$B(\alpha_1, \alpha_2) \stackrel{\text{def}}{=} \frac{\Gamma(\alpha 1)\Gamma(\alpha 2)}{\Gamma(\alpha 1 + \alpha 2)}$$

The beta function comes up as a normalizing constant for beta distributions (Section 4.4.2). It's often useful to recognize the following identity:

$$B(\alpha_1, \alpha_2) = \int_0^1 x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}\, dx$$

B.2 The Poisson distribution

The POISSON DISTRIBUTION is a generalization of the binomial distribution in which the number of trials n grows arbitrarily large while the mean number of successes πn is held constant. It is traditional to write the mean number of successes as λ; the Poisson probability density function is

$$P(y;\lambda) = e^{\lambda}\frac{\lambda^y}{y!} \quad (y = 0, 1, \dots) \tag{B.1}$$

The Gamma distribution is conjugate for the Poisson parameter λ, hence it is common to use a Gamma prior on λ in Bayesian inference.

B.3 The hypergeometric distribution

One way of thinking of the binomial distribution is as n repeated draws from a bag with M marbles, πM of which are black and the rest of which are white; each outcome is recorded and the drawn marble is replaced in the bag, and at the end the total number of black marbles is the outcome k. This picture is often called SAMPLING WITH REPLACEMENT. The HYPERGEOMETRIC distribution is similar to this conception of the binomial distribution except that the marbles are not replaced after drawn—this is SAMPLING WITHOUT REPLACEMENT. The hypergeometric distribution has three parameters: the number of marbles M, the number of black marbles m, and the number of draws n; the probability mass function on the number of "successes" X (black marbles drawn) is

$$P(X = r) = \frac{\binom{m}{r}\binom{M-m}{n-r}}{\binom{M}{m}}$$

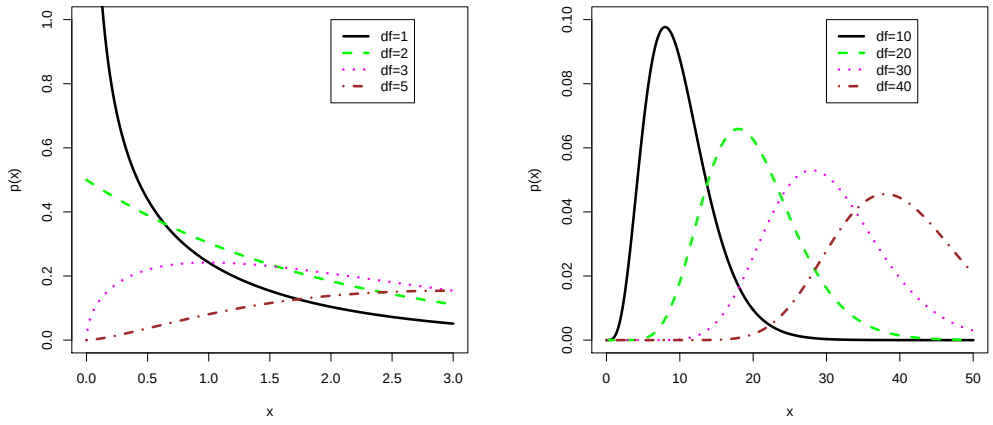
In this book, the hypergeometric distribution comes up in discussion of Fisher's exact test (Section 5.4.3).

B.4 The chi-square distribution

Suppose that we have a standard normal random variable Z—that is, $Z \sim N(0, 1)$. The distribution that the quantity Z^2 follows is called the CHI-SQUARE DISTRIBUTION with ONE DEGREE OF FREEDOM. This distribution is typically denoted as χ^2_1.

If we have k independent random variables $U_1, \dots, U_k$ such that each $U_i \sim \chi^2_1$, then the distribution of $U = U_1 + \cdots + U_k$ is the chi-squared with k degrees of freedom. This is denoted as $U \sim \chi^2_k$. The expectation of U is k and its variance is $2k$.

Figure B.1 illustrates the probability density functions for χ^2 distributions with various degrees of freedom. The χ^2_1 distribution grows asyptotically as x approaches 0, and χ^2_2 decreases monotonically, but all other χ^2_k distributions have a mode for some positive $x < k$.



(a) df≤10 (b) df≥10

Figure B.1: The χ^2 distribution with various degrees of freedom

As k grows large, more and more of the probability mass becomes located relatively close to $x = k$.

The key place where χ^2 variables arise is as the distribution of variance of a normal distribution. If we sample n points from $\mathcal{N}(\mu, \sigma^2)$ (once again: that's a normal distribution with mean μ and variance σ^2), then the quantity

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

is distributed as χ^2_{n-1}.

If U is distributed as χ^2_k, then the distribution of the quantity $1/U$ is called the INVERSE CHI-SQUARE DISTRIBUTION with k degrees of freedom. The inverse chi-square distribution is used in Bayesian inference as a conjugate prior (Section 4.4.3) for the variance of the normal distribution.

B.5 The t-distribution

Suppose once again that we have a standard normal random variable $Z \sim N(0, 1)$, and also that we have a chi-squared random variable U with k degrees of freedom. The distribution of the quantity

$$\frac{Z}{\sqrt{U/k}} \tag{B.2}$$

is called the t-DISTRIBUTION WITH k DEGREES OF FREEDOM. It has expectation 0, and as long as $k > 2$ its variance is $\frac{k}{k-2}$ (it has infinite variance if $k \leq 2$).

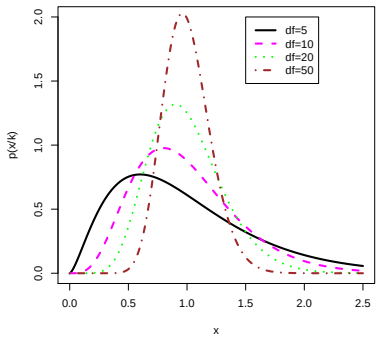


Figure B.2: The χ^2 distribution, normalized by degrees of freedom

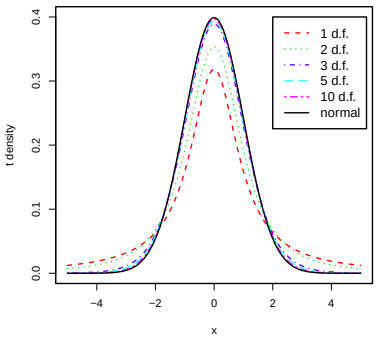


Figure B.3: The t distribution

Figure B.3 shows the probability density functions for t distributions with varying degrees of freedom, together with the standard normal distribution for reference. The t distribution is *heavier-tailed* than the normal distribution, but even with 10 degrees of freedom the t distribution is already very close to the standard normal. As the degrees of freedom grow, the t distribution converges to the standard normal; intuitively, this is because χ^2_k becomes more and more centered around k, so the quantity U/k in Equation B.2 converges to 1.

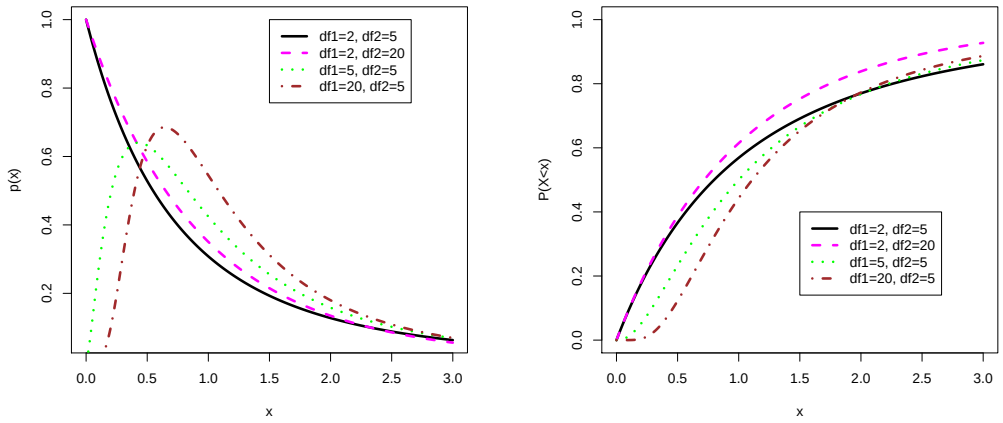
B.6 The F distribution

The F distribution, named after Ronald A. Fisher, one of the founders of the frequentist school of statistical analysis, is the distribution of the normalized ratio of two independent normalized χ^2 random variables. More formally, if $U \sim \chi^2_{k_1}$ and $V \sim \chi^2_{k_2}$, we have

$$F_{k_1,k_2} \sim \frac{U/k_1}{V/k_2} \tag{B.3}$$

Here are a few things to note about the F distribution:

- The F distribution comes up mainly in frequentist hypothesis testing for linear models (Section 6.5).
- As k_1 and k_2 grow, all the probability mass in the F distribution converges to $x = 1$. Because the variance of a sample is distributed as a χ^2 random variable, the ratio of variances in linear models (as in Figure 6.9) can be compared to the F distribution.
- Consider the case where $k_1 = 1$. Since U is then the square of a standard normal random variable, a random variable with distribution F_{1,k_2} has the same distribution as the square of a random variable with distribution t_{k_2} (compare Equation (B.2)).



(a) Density functions

(b) Cumulative distribution functions

Figure B.4: Density and cumulative distribution functions for the F distribution

It is useful to play a bit with the F distribution to see what it looks like. Figure B.4 gives sample density and cumulative distribution functions for several choices of the degrees of freedom. In general, the cumulative distribution is more interesting and pertinent than the probability density function (unless you have an anomalously low F statistic).

B.7 The Wishart distribution

Recall that the χ^2 distribution is used to place probability distributions over the inverse variance of a normal distribution (or of a sample from a normally-distributed population). The WISHART DISTRIBUTION is a multi-dimensional generalization of the χ^2 distribution; it generates inverse covariance matrices. Suppose that we have k independent observations from an $n \leq k$-dimensional multivariate normal distribution that itself has mean zero and covariance matrix Σ. Each observation $\boldsymbol{z}_i$ can be written as $\langle z_{i1}, \ldots, z_{in} \rangle$. If we write the matrix

$$\boldsymbol{Z} = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1n} \\ z_{21} & z_{22} & \cdots & z_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ z_{2n} & z_{2n} & \cdots & z_{kn} \end{bmatrix}$$

then the matrix $\boldsymbol{X} = \boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$ follows a Wishart distribution with k degrees of freedom and scale matrix Σ.

If $\boldsymbol{X}$ is Wishart-distributed, then its inverse $\boldsymbol{X}^{-1}$ is said to be INVERSE WISHART-DISTRIBUTED. The inverse Wishart distribution is used in Bayesian inference as the con-

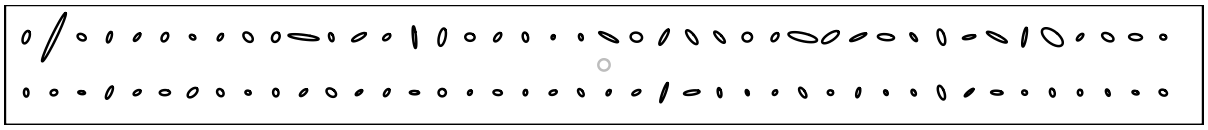


Figure B.5: Covariance-matrix samples from the two-dimensional inverse Wishart distribution with $\Sigma = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ and $k = 2$ (top row) or $k = 5$ (bottom row), represented by their characteristic ellipses. The unit circle appears in gray in the center of the figure for reference.

jugate prior (Section 4.4.3) for the covariance matrix of a multivariate normal distribution. Figure B.5 illustrates the inverse Wishart distribution for different degrees of freedom. Note that the variability in the covariance structure are more extreme when there are fewer degrees of freedom.

B.8 The Dirichlet distribution

The DIRICHLET DISTRIBUTION is a generalization of the beta distribution (Section 4.4.2). Beta distributions are probability distributions over the success parameter π of a binomial distribution; the binomial distribution has two possible outcome classes. Dirichlet distributions are probability distributions over the parameters $\pi_1, \ldots, \pi_k$ of a k-class multinomial distribution (Section 3.4.1; recall that π_k is not a true model parameter as it is fully determined by $\pi_1, \ldots, \pi_{k-1}$). The Dirichlet distribution is characterized by parameters $\alpha_1, \ldots, \alpha_k$, and $\mathcal{D}(\pi_1, \ldots, \pi_k)$ is defined as

$$\mathcal{D}(\pi_1, \ldots, \pi_k) \stackrel{\text{def}}{=} \frac{1}{Z} \pi_1^{\alpha_1 - 1} \pi_2^{\alpha_2 - 1} \ldots \pi_k^{\alpha_k - 1}$$

where the normalizing constant Z is

$$Z = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_k)}{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}$$

By comparing with the beta function and beta distribution as defined in Sections 4.4.2 and B.1, it will be apparent that the beta distribution is a Dirichlet distribution in which $k = 2$. Just as there is a beta-binomial distribution giving the probability of obtaining y successes out of N draws from a binomial distribution drawn from a beta distribution, there is a DIRICHLET-MULTINOMIAL distribution that gives the probability of obtaining $y_1, \ldots, y_k$ outcomes in each of the k response classes respectively when taking N draws from a multinomial drawn from a Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$. If we define $\alpha = \sum_{i=1}^{k} \alpha_i$, then the predictive distribution is (Leonard, 1977):

$$P(y_1, \ldots, y_k) = \int_\pi P(y_1, \ldots, y_k|\pi) P(\pi|\alpha_{1\ldots k})\, d\pi = \frac{\prod_{i=1}^{k} \binom{\alpha_i + y_i - 1}{\alpha_i}}{\binom{\alpha + N - 1}{\alpha}}$$

The special case of this predictive distribution is when we draw a multinomial distribution from the Dirichlet, and then draw one sample X from that multinomial distribution. The probability that that sample X has outcome class i is given by the value

$$P(X = i|\alpha_{1\ldots k}) = \frac{\alpha_i}{\alpha}$$

This is often convenient for using Gibbs sampling to draw samples from the posterior distribution in Bayesian models which use Dirichlet priors over multinomial distributions. An example of this usage is given in Section **??**.

The Dirichlet distribution has the following useful property. For any k-class Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$, suppose we partition the k outcome classes into a smaller, new set of $k' < k$ classes, with the j-th new class consisting of outcome classes $c_{j1}, \ldots, c_{jM_j}$. The resulting distribution over the new set of k' outcome classes is *also* Dirichlet-distributed, with parameters $\alpha_j = \sum_{i=1}^{M_j} \alpha_{ij}$. [see also Dirichlet process in Section XXX; and give example here?]

B.9 The beta-binomial distribution

We saw the beta-binomial distribution before in Section 4.4.3. If there is a binomial distribution with unknown success parameter π and we put a beta prior with parameters α_1, α_2 over π, then the marginal distribution on a sample of size n from the binomial distribution is beta-binomial, with form

$$P(m|\alpha_1, \alpha_2, m) = \binom{n}{m} \qquad\qquad = \binom{k}{r} \frac{B(\alpha_1 + m, \alpha_2 + m - n)}{B(\alpha_1, \alpha_2)}$$