APPENDIX A

Distributional catalog

This appendix provides a summary of the statistical distributions used in this book. For each distribution, we list its abbreviation, its support, constraints on its parameters, its density function, and its mean, variance, and other moments as appropriate. We also give additional information helpful in implementing a Bayes or EB analysis (e.g., relation to a particular sampling model or conjugate prior). We denote density functions generically by the letter p, and the random variable for the distribution in question by a Roman letter: x for scalars, x for vectors, and V for matrices. Parameter values are assigned Greek letters. The reader should keep in mind that many of these distributions are most commonly used as prior distributions, so that the random variable would itself be denoted as a Greek character.

In modern applied Bayesian work, a key issue is the generation of random draws from the distributions in this catalog. Since whole textbooks (Devroye, 1986; Ripley, 1987; Gentle, 1998) have been devoted to this subject, we make no attempt at a comprehensive treatment here, but only provide a few remarks and pointers. First, many modern Bayesian software packages (such as BUGS and others described in Appendix Section C.3) essentially "hide" the requisite sampling from the user altogether. High-level (but still general purpose) programming environments (such as S-Plus, Gauss, or MOW) feature functions for generating samples from most of the standard distributions below. Lower-level languages (such as Fortran or C) often include only a $Uniform(0, 1)$ generator, since samples from all other distributions can be built from these. In what follows, we provide a limited number of generation hints, and refer the reader to the aforementioned textbooks for full detail.

A.1 Discrete

A.1.1 Univariate

Binomial: $X \sim Bin(n, \theta)$, $x = 0, 1, 2, \ldots, n$, $0 \leq \theta \leq 1$, n any positive

© 2000 by CRC Press LLC

integer, and

$$p(x|n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} ,$$

where

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

$E(X) = n\theta$ and $Var(X) = n\theta(1-\theta)$.

The binomial is commonly assumed in problems involving a series of independent, success/failure trials all having probability of success θ *(Bernoulli trials): X is* the number of successes in n such trials. The binomial is also the appropriate model for the number of "good" units obtained in n independent draws with replacement from a finite population of size N having G good units, and where $\theta = G/N$, the proportion of good units.

Poisson: $X \sim Po(\theta)$, $x = 0, 1, 2, \ldots, \theta > 0$, and

$$p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}$$

$E(X) = \theta$ and $Var(X) = \theta$.

The Poisson is commonly assumed in problems involving counts of "rare events" occurring in a given time period, and for other discrete data settings having countable support. It is also the limiting distribution for the binomial when n goes to infinity and the expected number of events converges to θ.

- Negative Binomial: $X \sim NegBin(r,\theta)$, $x = 0, 1, 2, \ldots$, $0 \leq \theta \leq 1$, r any positive integer, and

$$p(x|r,\theta) = \binom{x+r-1}{x}\theta^r(1-\theta)^x .$$

$E(X) = r(1-\theta)/\theta$ and $Var(X) = r(1-\theta)/\theta^2$.

The name "negative binomial" is somewhat misleading; "inverse binomial" might be a better name. This is because it arises from a series of Bernoulli trials where the number of *successes,* rather than the total number of trials, is fixed at the outset. In this case, X is the number of failures preceding the r^{th} success. This distribution is also sometimes used to model countably infinite random variables having variance substantially larger than the mean (so that the Poisson model would be inappropriate). In fact, it is the marginal distribution of a Poisson random variable whose rate θ follows a gamma distribution (see Section A.2 below).

- Geometric: $X \sim Geom(\theta) \equiv NegBin(1,\theta)$, $x = 0, 1, 2, \ldots$, $0 \leq \theta \leq 1$, and

$$p(x|\theta) = \theta(1-\theta)^x .$$

© 2000 by CRC Press LLC

$E(X) = (1-\theta)/\theta$ and $Var(X) = (1-\theta)/\theta^2$.

The geometric is the special case of the negative binomial having $r = 1$; it is the number of failures preceding the first success in a series of Bernoulli trials.

A.1.2 Multivariate

9 Multinomial: $\mathbf{X} \sim Mult(n, \boldsymbol{\theta})$ for $\mathbf{X} = (x_1, \ldots, x_k)'$, where $x_i \in \{0, 1, 2, \ldots, n\}$ and $\sum_{i=1}^{k} x_i = n$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ where $0 \leq \theta_i \leq 1$ and $\sum_{i=1}^{k} \theta_i = 1$, and

$$p(\mathbf{x}|n, \boldsymbol{\theta}) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} \theta_i^{x_i} .$$

$E(X_i) = n\theta_i$, $Var(X_i) = n\theta_i(1-\theta_i)$, and $Cov(X_i, X_j) = -n\theta_i\theta_j$.

The multinomial is the multivariate generalization of the binomial: note that for $k = 2$, $Mult(n, \boldsymbol{\theta} = Bin(n, \theta$ with $\theta = \theta_1 = 1 - \theta_2$. For a population having $k > 2$ mutually exclusive and exhaustive classes, X_i is the number of elements of class i that would be obtained in n independent draws with replacement from the population. Note well that this is really only (k-1)-dimensional distribution, since $x_k = n - \sum_{i=1}^{k-1} x_i$ and $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$.

A.2 Continuous

A.2.1 Univariate

- Beta: $X \sim Beta(\alpha, \beta)$, $x \in [0, 1]$, $\alpha > 0$, $\beta > 0$, and

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} ,$$

where $\Gamma(\cdot)$ denotes the *gamma function,* defined by the integral equation

$$\Gamma(\alpha) \equiv \int_0^\infty y^{\alpha-1} e^{-y} dy , \quad \alpha > 0 .$$

Using integration by parts, we have that $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, so since $\Gamma(1) = 1$, we have that for integer α,

$$\Gamma(\alpha) = (\alpha-1)! .$$

One can also show that $\Gamma(1/2) = \sqrt{\pi}$.

For brevity, the pdf is sometimes written as

$$\frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} ,$$

© 2000 by CRC Press LLC

where $B(\cdot, \cdot)$ denotes the *beta function,*

$$B(\alpha, \beta) \equiv \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} .$$

Returning to the distribution itself, we have $E(X) = \alpha/(\alpha+\beta)$ and $Var(X) = \alpha\beta/[(\alpha+\beta)^2(\alpha+\beta+1)]$; it is also true that $E(X^2) = \alpha(\alpha+1)/[(\alpha+\beta)(\alpha+\beta+1)]$. The beta is a common and flexible distribution for continuous random variables defined on the unit interval (so that rescaled and/or recentered versions may apply to any continuous variable having finite range). For $\alpha < 1$, the distribution has an infinite peak at 0; for $\beta <$ l, the distribution has an infinite peak at l. For $\alpha = \beta = 1$, the distribution is flat (see the next bullet point). For α and β both greater than 1, the distribution is concave down, and increasingly concentrated around its mean as $\alpha + \beta$ increases. The beta is the conjugate prior for the binomial likelihood; for an example, see Subsection 2.3.4.

Uniform: $X \sim Unif(\theta_L, \theta_U)$, $x \in [\theta_L, \theta_U]$, $-\infty < \theta_L < \theta_U < \infty$, and

$$p(x|\theta_L, \theta_U) = \frac{1}{\theta_U - \theta_L} .$$

$E(X) = (\theta_L + \theta_U)/2$ and $Var(X) = (\theta_U - \theta_L)^2/12$. Clearly $Unif(0, 1) \equiv Beta(1, 1)$.

- Normal (or Gaussian): $X \sim N(\mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$E(X) = \mu$ and $Var(X) = \sigma^2$. The normal is the single most common distribution in statistics, due in large measure to the Central Limit Theorem. It is symmetric about p, which is not only its mean, but its median and mode as well.

A $N(0, 1)$ variate Z can be generated as $\sqrt{-2\log(U_1)}cos(2\pi U_2)$, where U_1 and U_2 are independent $Unif(0, 1)$ random variates; in fact, the quantity $\sqrt{-2\log(U_1)}sin(2\pi U_2)$ produces a second, independent $N(0,1)$ variate. This is the celebrated *Box-Muller* (1958) method. A $N(\mu, \sigma^2)$ variate can then be generated as $\mu + \sigma Z$.

- Double Exponential (or *Laplace*):$X \sim DE(\mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right)$$

$E(X) = \mu$ and $Var(X) = 2\sigma^2$.

Like the normal, this distribution is symmetric and unimodal, but has heavier tails and a somewhat different shape, being strictly concave up on both sides of μ.

© 2000 by CRC Press LLC

- Logistic: $X \sim Logistic(\mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left[1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right]^2} .$$

  $E(X) = \mu$ and $Var(X) = (\pi^2/3)\sigma^2$.

  The logistic is another symmetric and unimodal distribution, more similar to the normal in appearance than the double exponential, but with even heavier tails.

- **t** (or *Student's t*): $X \sim t(\nu, \mu, \sigma^2)$, $x \in \Re$, $\nu > 0$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2, \nu) = \frac{\Gamma[(\nu+1)/2]}{\sigma\sqrt{\nu\pi}\Gamma(\nu/2)}\left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\nu+1)/2}$$

  $E(X) = \mu$ (if $\nu > 1$) and $Var(X) = \nu\sigma^2/(\nu - 2)$ (if $\nu > 2$). The parameter v is referred to as the *degrees of freedom* and is usually taken to be a positive integer, though the distribution is proper for any positive real number ν. The t is a common heavy-tailed (but still symmetric and unimodal) alternative to the normal distribution. The leading term in the pdf can be rewritten

$$\frac{\Gamma[(\nu+1)/2]}{\sigma\sqrt{\nu\pi}\Gamma(\nu/2)} = \frac{1}{\sigma\sqrt{\nu}B(\frac{1}{2}, \frac{\nu}{2})} ,$$

  where $B(.,.)$ again denotes the beta function.

- **Cauchy:** $X \sim Cau(\mu, \sigma^2) \equiv t(1, \mu, \sigma^2)$, $x \in \Re$, $-\infty < \mu < \infty$, $\sigma^2 > 0$, and

$$p(x|\mu, \sigma^2) = \frac{1}{\sigma\pi\left[1 + \left(\frac{x-\mu}{\sigma}\right)^2\right]} .$$

  $E(X)$ and $Var(X)$ do not exist, though μ is the median of this distribution. This special case of the t distribution has the heaviest possible tails and provides a wealth of counterexamples in probability theory.

- **Gamma:** $X \sim G(\alpha, \beta)$, $x > 0$, $\alpha > 0$, $\beta > 0$, and

$$p(x|\alpha, \beta) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} .$$

  $E(X) = \alpha\beta$ and $Var(X) = \alpha\beta^2$. Note also that if $X \sim G(\alpha, \beta)$, then $Y = cX \sim G(\alpha, c\beta)$.

  The gamma is a flexible family for continuous random variables defined on the positive real line. For $\alpha < 1$, the distribution has an infinite peak at 0 and is strictly decreasing. For $\alpha = 1$, the distribution intersects the vertical axis at $1/\beta$, and again is strictly decreasing (see the Exponential distribution below). For $\alpha > 1$, the distribution starts at the origin,

© 2000 by CRC Press LLC

increases for a time and then decreases; its appearance is roughly normal for large a. The gamma is the conjugate prior distribution for a Poisson rate parameter θ.

- **Exponential:** $X \sim Expo(\beta) \equiv G(1, \beta)$, $x > 0$, $\beta > 0$, and

$$p(x|\beta) = \frac{1}{\beta} e^{-x/\beta} .$$

$E(X) = \beta$ and $Var(X) = \beta^2$.

- **Chi-square:** $X \sim \chi^2(\nu) \equiv G(\nu/2, 2)$, $x > 0$, $\nu > 0$, and

$$p(x|\nu) = \frac{x^{\nu/2-1} e^{-x/2}}{\Gamma(\nu/2) 2^{\nu/2}} .$$

$E(X) = \nu$ and $Var(X) = 2\nu$. As with the t distribution, the parameter ν is referred to as the *degrees of freedom* and is usually taken to be a positive integer, though a proper distribution results for any positive real number ν.

- **Inverse Gamma:** $X \sim IG(\alpha, \beta)$, $x > 0$, $\alpha > 0$, $\beta > 0$, and

$$p(x|\alpha, \beta) = \frac{e^{-1/(\beta x)}}{\Gamma(\alpha) \beta^{\alpha} x^{\alpha+1}} .$$

$E(X) = 1/[\beta(\alpha - 1)]$ (provided a > 1) and $Var(X) = 1/[\beta^2(\alpha - 1)^2(\alpha - 2)]$ (provided $\alpha > 2$). Note also that if $X \sim IG(\alpha, \beta)$, then $Y = cX \sim IG(\alpha, \beta/c)$.

A better name for the inverse gamma might be the *reciprocal gamma,* since $X = 1/Y$ where $Y \sim G(\alpha, \beta)$ (or in distributional shorthand, $IG(\alpha, \beta) \equiv 1/G(\alpha, \beta)$).

Despite its poorly behaved moments and somewhat odd, heavy-tailed appearance, the inverse gamma is very commonly used in Bayesian statistics as the conjugate prior for a variance parameter σ^2 arising in a normal likelihood function. Choosing α and β appropriately for such a prior can be aided by solving the above equations for $\mu \equiv E(X)$ and $\tau^2 \equiv Var(X)$ for α and β. This results in

$$\alpha = (\mu/\tau)^2 + 2 \quad \text{and} \quad \beta = \frac{1}{\mu \left[(\mu/\tau)^2 + 1 \right]} .$$

Setting the prior mean and standard deviation both equal to μ (a reasonably vague specification) thus produces $\alpha = 3$ and $\beta = 1/(2\mu)$.

- Inverse Gaussian (or *Wald*): $X \sim InvGau(\mu, \lambda)$, $x > 0$, $\mu > 0$, $\lambda > 0$, and

$$p(x|\mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left[-\frac{\lambda(x - \mu)^2}{2\mu^2 x} \right]$$

$E(X) = \mu$ and $Var(X) = \mu^3/\lambda$.

© 2000 by CRC Press LLC

The inverse Gaussian has densities that resemble those of the gamma distribution. Despite its somewhat formidable form, all its positive and negative moments exist.

A.2.2 Multivariate

- **Dirichlet:** $\mathbf{X} \sim D(\boldsymbol{\alpha})$, $\mathbf{X} = (x_1, \ldots, x_k)'$ where $0 \leq x_i \leq 1$ and $\sum_{i=1}^{k} x_i = 1$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ where $\alpha_i \geq 0$, and

$$p(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}$$

  $E(X_i) = \alpha_i/\alpha_0$, $Var(X_i) = [\alpha_i(\alpha_0 - \alpha_i)]/[\alpha_0^2(\alpha_0 + 1)]$, and finally $Cov(X_i, X_j) = -(\alpha_i\alpha_j)/[\alpha_0^2(\alpha_0 + 1)]$, where $\alpha_0 \equiv \sum_{i=1}^{k} \alpha_i$.

  The Dirichlet is the multivariate generalization of the beta: note that for $k = 2$, $D(\boldsymbol{\alpha}) = Beta(\alpha_1, \alpha_2)$. Note that this is really only $(k - 1)$-dimensional distribution, since $x_k = 1 - \sum_{i=1}^{k-1} x_i$. The Dirichlet is the conjugate prior for the multinomial likelihood. It also forms the foundation for the *Dirichlet process prior,* the basis of most nonparametric Bayesian inference (see Section 2.5).

- **Multivariate Normal** (or *Multinormal*, or *Multivariate Gaussian*): $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{x} \in \Re^k$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ where $-\infty < \mu_i < \infty$, $\boldsymbol{\Sigma}$ a $k \times k$ positive definite matrix, and

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}}{(2\pi)^{k/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right],$$

  where $|\boldsymbol{\Sigma}^{-1}|$ denotes the determinant of $\boldsymbol{\Sigma}^{-1}$. $E(\mathbf{X}) = \boldsymbol{\mu}$ (i.e., $E(X_i) = \mu_i$ for all i), and $Var(X) = \boldsymbol{\Sigma}$ (i.e., $Var(X_i) = \sigma_{ii}$ and $Cov(X_i, X_j) = \sigma_{ij}$ where $\boldsymbol{\Sigma} = (\sigma_{ij})$). The multivariate normal forms the basis of the likelihood in most common linear models, and also serves as the conjugate prior for mean and regression parameters in such likelihoods.

  A $N_k(0, \mathbf{I})$ variate $\mathbf{z}$ can be generated simply as a vector of k independent, univariate $N(0, 1)$ random variates. For general covariance matrices $\boldsymbol{\Sigma}$, we first factor the matrix as $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is a lower-triangular matrix (this is often called the *Cholesky factorization;* see for example Thisted, 1988, pp.81-83). A $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variate z can then be generated as $\boldsymbol{\mu} + \mathbf{L}\mathbf{z}$.

- **Multivariate t** (or *Multi-t):* $\mathbf{X} \sim t_k(\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{x} \in \Re^k$, $\nu > 0$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ where $-\infty < \mu_i < \infty$, $\boldsymbol{\Sigma}$ a $k \times k$ positive definite matrix, and $p(\mathbf{x}|\nu, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$\frac{|\boldsymbol{\Sigma}^{-1}|^{1/2}\Gamma[(\nu + k)/2]}{(\nu\pi)^{k/2}\Gamma(\nu/2)} \left[1 + \frac{1}{\nu}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]^{-(\nu+k)/2}$$

© 2000 by CRC Press LLC

$E(\mathbf{X}) = \boldsymbol{\mu}$ (if $\nu > 1$), and $Var(\mathbf{X}) = \nu\boldsymbol{\Sigma}/(\nu - 2)$ (if $\alpha > 2$). The multivariate t provides a heavy-tailed alternative to the multivariate normal while still accounting for correlation among the elements of $\mathbf{x}$ Typically, $\boldsymbol{\Sigma}$ is called a scale matrix, and is approximately equal to the variance matrix of X for large ν.

- **Wishart**: $\mathbf{V} \sim W(\boldsymbol{\Omega}, \nu)$, $\mathbf{V}$ a $k \times k$ symmetric and positive definite matrix, $\boldsymbol{\Omega}$ a $k \times k$ symmetric and positive definite parameter matrix $\nu > 0$, and

$$p(\mathbf{V}|\nu, \boldsymbol{\Omega}) = c\frac{|\mathbf{V}|^{(\nu-k-1)/2}}{|\boldsymbol{\Omega}|^{\nu/2}} \exp\left[-\frac{1}{2}tr(\boldsymbol{\Omega}^{-1}\mathbf{V})\right],$$

provided the shape parameter (or "degrees of freedom") $\nu \geq k$. Here the proportionality constant c takes the awkward form

$$c = \left[2^{(\nu k)/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma\left(\frac{\nu + 1 - j}{2}\right)\right]^{-1}.$$

$E(V_{ij}) = \nu\Omega_{ij}$, $Var(V_{ij}) = \nu(\Omega_{ij}^2 + \Omega_{ii}\Omega_{jj})$, and finally $Cov(V_{ij}, V_{kl}) = \nu(\Omega_{ik}\Omega_{jl} + \Omega_{il}\Omega_{jk})$.

The Wishart is a multivariate generalization of the gamma, originally derived as the sampling distribution of the sum of squares and crossproducts matrix, $\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$, where $\mathbf{X}_i \overset{iid}{\sim} N_k(\boldsymbol{\mu}, \mathbf{V})$. In Bayesian analysis, just as the reciprocal of the gamma (the inverse gamma) is often used as the conjugate prior for a variance parameter σ^2 in a normal likelihood, the reciprocal of the Wishart (the *inverse Wishart*) is often used as the conjugate prior for a variance-covariance matrix $\boldsymbol{\Sigma}$ in a multivariate normal likelihood (see Example 5.6).

A random draw from a Wishart with ν an integer can be obtained using the idea in the preceding paragraph: If $\mathbf{x}_1, \ldots, \mathbf{x}_\nu$ are independent draws from a $N_k(\mathbf{0}, \boldsymbol{\Omega})$ distribution, then $\mathbf{V} = \sum_{i=1}^{\nu}\mathbf{x}_i\mathbf{x}_i'$ is a $W(\boldsymbol{\Omega}, \nu)$ random variate. Non-integral ν requires the general algorithm originally given by Odell and Feiveson (1966); see also the textbook treatment by Gentle (1998, p.107), or the more specifically Bayesian treatment by Gelfand et al. (1990).

© 2000 by CRC Press LLC