Useful Probability Distributions

- Standard Normal Distribution
- Binomial
- Multinomial
- Hypergeometric
- Poisson
- Beta Binomial
- Student's t
- Beta
- Gamma
- Dirichlet
- Multivariate Normal and Correlation

Standard Normal Distribution

$$X \sim N(\theta, \sigma^2)$$

Standardization

$$Z = (X - \theta)/\sigma$$

- A general normally distributed random variable is transformed into one which has a standard normal distribution
- E(Z)=0 and Var(Z)=1
- \bullet Division by σ ensures the resulting statistic is dimensionless
- Pr(Z < z) is denoted $\Phi(z)$: cumulative distribution function
- Tabulated, e.g., z=1.6449, $\Phi(z)=0.950$, z=2.5758, $\Phi(z)=0.995$
- $\Phi(-z)=1-\Phi(z)$

Student's t-distribution

- Similar to Standard Normal Distribution
- \bullet Standard deviation σ of a Normal Distribution is rarely known
- \bullet T-statistic takes into account uncertainty associated with estimating σ

If
$$X_i \sim N(\theta, \sigma^2)$$
,

$$\overline{\mathbf{X}} = \sum_{i=1}^{n} X_{i}$$

has a normal distribution with $\overline{X} \sim N(\theta, \sigma^2/n)$

If S is an estimate of std dev

then $(\overline{X} - \theta)/(S/\sqrt{n})$ has a t-distribution

which has a greater dispersion than standard normal

Binomial

- Models a sequence of independent trials in which there are only two possible mutually exclusive outcomes
- *n* is the number of trials
- X is the number of successes
- p is the probability of success in any individual trial, and q=1-p
- P(X=x), x=0,1...,n is denoted p_i

$$Pr(X = x) = \binom{n}{x} p^{x} (1-p)^{n-x}$$
$$= \binom{n}{x} p^{x} q^{n-x}$$

- Distribution of *X* is denoted in shorthand as $X \sim Bin(n,p)$
- If X is number of sixes in 10 throws of a fair die, then $X \sim Bin(10, 1/6)$
- E(X)=np, Var(X)=npq

Normal Approximation to the Binomial

- Binomial is a discrete distribution where it is tedious to evaluate exact probabilities for large number of events, e.g., probability of 530 or fewer heads in 1000 tosses of a fair coin
- *X~N(np,npq)*

$$\Pr(X \le 530 \mid n = 1000, p = 0.5) = \sum_{x=0}^{530} \binom{n}{x} 0.5^{1000}$$

$$= \Phi((530.5 - 500) / \sqrt{250})$$

$$= \Phi(1.929) = 0.9731$$
If $X \sim N(\mu, \sigma^2)$

$$Z = (X - \mu) / \sigma \text{ is } N(0,1)$$

$$\Pr(Z < z) \text{ is denoted } \Phi(z)$$

Z has a standard normal distribution which is tabulated

Multinomial Distribution

- Generalization of Binomial
- Models a sequence of independent trials where there are *k* possible mutually exclusive outcomes

$$Pr(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1! x_2! ... x_k!} p_1^{x_1} ... p_k^{x_k}$$

$$\sum_{i=1}^{k} p_i = 1$$

Hypergeometric Distribution

- Binomial is with replacement
 - Models a sequence of independent trials where there are 2 possible mutually exclusive outcomes
- Hypergeometric is without replacement
 - E.g., Probability of the number *X* of illicit tablets in a sample of size *m* from a consignment of size *N* in which *R* are illicit and *N-R* are licit is

$$\Pr(X = x) = \frac{\binom{R}{x} \binom{N - R}{m - x}}{\binom{N}{m}}$$

If N=20, R=10, m=6 then
$$Pr(X=3)={}^{10}C_3$$
 ${}^{10}C_3$ ${}^{20}C_6$ = 0.37

Beta-Binomial Distribution

- Consignment of tablets, a proportion of which are suspected drugs. For large consignments, probability distribution of the proportion t which are drugs can be modeled with a beta distribution, which treats the proportion t as a variable which is continuous over the interval (0,1)
- For small consignments, say N<50, a more accurate distribution, which recognizes the discrete nature of possible values of the proportions is used

Beta Distribution

- Consignment of N tablets, No of illicit is R
- Proportion of illicit is R/N which has a finite no of values ranging from 0/N to N/N in steps of 1/N
- As N increases proportion becomes closer to a continuous measurement over interval (0,1)
- Modeled by a beta distribution
- Denote true proportion by random variable θ (0 < θ <1)

$$p(\theta / \alpha, \beta) = \frac{\theta^{\alpha - 1} (1 - \theta)^{\beta - 1}}{B(\alpha, \beta)}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$
 known as the beta function

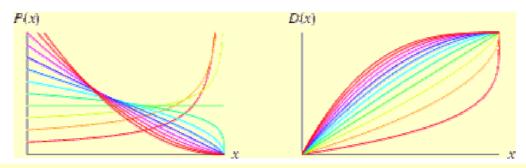
 Γ () is the gamma function defined as

$$\Gamma(x+1) = x!$$

$$\Gamma(1/2) = \sqrt{\pi}$$

- Values of α and β reflect prior beliefs before inspection (Bayesian philosophy)
- Large value of α relative to β would imply a belief that θ was high
- Neutral belief would have $\alpha = \beta = 1$

Beta Distribution



It is a general type of statistical distribution which is related to the gamma distribution. Beta distributions have two free parameters, which are labeled α and β

The domain is [0,1], and the probability function P(x) and distribution function D(x) are given by

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

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

$$D(x) = I(x; a, b),$$

where B(a,b) is the beta function, I(x;a,b), is the regularized beta function, and $\alpha,\beta>0$

Gamma Distribution

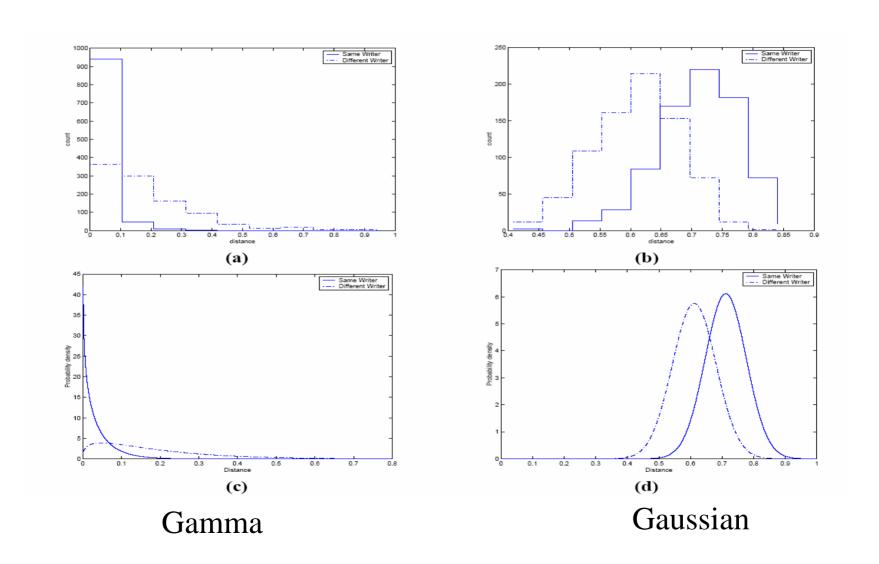
 Probability Density function expressed in terms of te Gamma function

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

Alternatively expressed as

$$g(x; \alpha, \beta) = x^{\alpha - 1} \frac{\beta^{\alpha} e^{-\beta x}}{\Gamma(\alpha)}$$
 for $x > 0$

Modeling distributions of distances in writer verification



Dirichlet Distribution

- Generalization of Beta distribution to k categories analogous to generalization of binomial distribution to multinomial distribution
- Example: proportion of illicit drugs when there are k types of drugs
 - Given a consignment of size N
 - No of tablets of each type is R_i , i=1...,k
 - Proportions are R_i/N
 - As N increases the proportions are continuous over (0,1)

Dirichlet Distribution

• Characterized by k parameters $\{\alpha_1,...\alpha_k\}$ chosen to represent prior beliefs about proportions

$$f(\theta_1, ... \theta_k) = \frac{\theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}}{B(\alpha_1, ... \alpha_k)} \quad 0 < \theta_i < 1, \sum_{i=1}^k \theta_i = 1$$

where

$$B(\alpha_1,..\alpha_k) = \frac{\Gamma(\alpha_1)..\Gamma(\alpha_k)}{\Gamma(\alpha_1 + ..\alpha_k)}$$