Lecture 1
Basic Statistical Machinery

Bruce Walsh. jbwalsh@u.arizona.edu. University of Arizona.

ECOL 519A, Jan 2007. University of Arizona

Probabilities, Distributions, and Expectations

Discrete and Continuous Random Variables

A **random variable** is a variable whose particular outcome (or **realization**) is not a set value but rather is drawn from some **probability distribution.** The simplest random variables are **discrete**, with the random variable x taking on values $X_1, X_2, \cdots X_k$ (a countable number of possible outcomes, so that X_∞ is allowable). The complete description of the behavior of the random variable is given by specifying

$$P_i = \Pr(x = X_i) \tag{1.1a}$$

for all the x_i values, the probability that $x = X_i$. Note that probabilities are positive and the sum of probabilities over all outcomes is one:

$$\Pr(x = X_i) \geq 0, \quad \text{and} \quad \sum_{i=1}^{k} P_i = 1 \tag{1.1b}$$

The **cumulative probability function**, $cdf(X)$ is given by $cdf(X) = \Pr(x \leq X)$. An example of a discrete random variable is the genotype at a particular locus in a randomly-drawn individual. Suppose the alleles are A and a, then (in a diploid) the possible outcomes for a randomly drawn genotype are AA, Aa, or aa.

A **continuous random variable** can take on any possible value over some interval (or sets of intervals), for example the height of a randomly-chosen individual. If x is continuously distributed, it makes no sense to specify $\Pr(x = X_i)$ since the probability that x takes on any specific value is infinitesimally small. Rather, it is more meaningful to consider the probability that x lies within a specific range of values, say x_1 and x_2. This quantity is described by the **probability density function** $p(x)$, which satisfies the integral

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x) \, dx \tag{1.2a}$$

Note that any function that satisfies the continuous analogs of Equation (1b),

$$p(x) \geq 0 \quad \text{and} \quad \int_{-\infty}^{\infty} p(x) dx = 1 \tag{1.2b}$$

is a probability density function. The cdf is given by

$$cdf(z) = \int_{-\infty}^{z} p(x) \, dx$$

Joint and Conditional Probabilities

The probability of joint occurrence of a pair of random variables (x, y) is specified by the **joint probability density function**, $p(x, y)$,

$$\mathrm{P}(\, y_1 \leq y \leq y_2,\ x_1 \leq x \leq x_2 \,) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} p(x, y)\, dx\, dy \tag{1.3a}$$

We often ask questions of the form: What is the distribution of y given that x equals some specified value? For example, we might want to know the probability that parents whose height is 68 inches have offspring with height exceeding 70 inches. To answer such questions, we use $p(y|x)$, the **conditional density** of y given x, where

$$\mathrm{P}(\, y_1 \leq y \leq y_2 \,|\, x \,) = \int_{y_1}^{y_2} p(\, y \,|\, x \,)\, dy \tag{1.3b}$$

The joint probability density functions $p(x, y)$, and conditional density function $p(y|x)$, are connected by

$$p(x, y) = p(\, y \,|\, x \,)\, p(x), \quad \text{hence} \quad p(\, y \,|\, x \,) = \frac{p(x, y)}{p(x)} \tag{1.3c}$$

where $p(x) = \int_{-\infty}^{+\infty} p(\, y, x \,)\, dy$ is the **marginal density** of x.

Two random variables, x and y, are said to be **independent** if $p(x, y)$ can be factored into the product of a function of x only and a function of y only, i.e.,

$$p(x, y) = p(x)\, p(y) \tag{1.3d}$$

If x and y are independent, knowledge of x gives no information about the value of y. From Equation 1.3c, if $p(x, y) = p(x)\, p(y)$, then $p(\, y \,|\, x \,) = p(y)$.

The relationship between conditional and joint probabilities (Equation 1.3c) is a general one, holding for probabilities as well as probability density functions. For example, suppose the genotypes AA and Aa both give green offspring, while aa gives red offspring. Further, suppose that we have crossed a pure red (aa) and green (AA) lines, so that in the F_2, $\mathrm{Pr}(AA) = 0.25, \mathrm{Pr}(Aa) = 0.5$. What is the probability that a green offspring is AA? Aa?

$$\mathrm{Pr}(AA|\text{green}) = \frac{\mathrm{Pr}(AA \text{ and green})}{\mathrm{Pr}(\text{green})} = \frac{0.25}{0.25 + 0.5} = \frac{1}{3}$$

Likewise $\mathrm{Pr}(Aa|\text{green}) = 2/3$.

Bayes' Theorem

An extremely useful formula is Bayes' theorem. Suppose there are n possible outcomes $(b_1, \cdots, b_n)$ of a random variable that we cannot observe. Given the observed outcome of a correlated variable A, what is the probability of b_j? From the definition of a conditional probability, $\mathrm{Pr}(b_j \,|\, A) = \mathrm{Pr}(b_j, A) / \mathrm{Pr}(\, A \,)$. We can decompose this further, by noting that $\mathrm{Pr}(b_j, A) = \mathrm{Pr}(b_j) \mathrm{Pr}(A \,|\, b_j)$ and $\mathrm{Pr}(A) = \sum_i^n \mathrm{Pr}(b_i)\ \mathrm{Pr}(A \,|\, b_i)$. Putting these together gives Bayes' theorem,

$$\mathrm{Pr}(b_j \,|\, A) = \frac{\mathrm{Pr}(b_j) \mathrm{Pr}(A \,|\, b_j)}{\mathrm{Pr}(A)} = \frac{\mathrm{Pr}(b_j) \mathrm{Pr}(A \,|\, b_j)}{\displaystyle\sum_{i=1}^{n} \mathrm{Pr}(b_i)\ \mathrm{Pr}(A \,|\, b_i)} \tag{1.4}$$

As a application of Bayes' theorem, consider a locus that contributes to height, and suppose we observe an individual over 70 inches. Suppose the population frequencies of the genotypes and the corresponding conditional probability they have offspring over 70 inches are as follows:

Genotype:	QQ	Qq	qq
Freq(Genotypes)	0.5	0.3	0.2
Prob(height > 70 | genotype)	0.3	0.6	0.9

Hence, the probability that a random individual exceeds a height of 70 is

$$0.5 \cdot 0.3 + 0.3 \cdot 0.6 + 0.2 \cdot 0.9 = 0.51$$

and

$$\Pr(\text{individual over 70 is } QQ) = \Pr(QQ \,|\, \text{over 70 }) = \frac{\Pr(QQ)\Pr(\text{over 70}\,|\,QQ)}{\Pr(\text{over 70})} = \frac{0.5 \cdot 0.3}{0.51} = 0.294$$

Expectations of Random Variables

The **expected value** of a function $f(x)$ of a random variable (the **expectation of f**) is just its average value,

$$E[f(x)] = \int_{-\infty}^{+\infty} f(x)p(x)dx, \quad \text{or } (x \text{ discrete}) \quad E[f(x)] = \sum_i \Pr(x = X_i)f(X_i) \qquad (1.5)$$

In particular, the **arithmetic mean**, μ, also known as the **first moment about the origin**, is

$$\mu = \int_{-\infty}^{+\infty} z\,p(z)\,dz = E(z) \qquad (1.6)$$

For a character denoted by z, the sample estimate of the mean is generally denoted by $\bar{z}$, and estimated as the average of the n measures,

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i$$

A second important expectation is the spread of values about their mean, the **variance** (a term introduced in Fisher's 1918 paper). Also known as the **second moment about the mean**, the variance is the expected squared deviation of an observation from its mean,

$$\sigma^2 = \int_{-\infty}^{+\infty} (z - \mu)^2\, p(z)\,dz = E\left[(z - \mu)^2\right] \qquad (1.7)$$

Because $\mu = E(z)$, this quantity can be expressed more simply by expanding $(z - \mu)^2$ to obtain

$$\sigma^2 = E(z^2 - 2z\mu + \mu^2) = E(z^2) - 2\mu E(z) + \mu^2 = E(z^2) - \mu^2 \qquad (1.8)$$

where we have used two useful properties of expectations,

$$E(x + y) = E(x) + E(y)$$
$$E(c\,x) = c\,E(x)$$

for a constant c. The standard estimator for the variance is

$$Var(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right)$$

More generally, the rth moment about the mean is given by

$$\mu_r = E\left[(x - \mu)^r\right] = \int_{-\infty}^{+\infty} (z - \mu)^r p(z)\,dz \qquad (1.9)$$

The Normal Distribution

The most common distribution used in quantitative genetics for a continuous character is the **normal** or **Gaussian distribution**. If z is a normally distributed variable, its density function is given by

$$p(z) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right] \tag{1.10}$$

Note that the normal distribution is completed specified by its mean μ and variance σ^2.

Very often selection is practiced by **truncation**, wherein all individuals above a threshold value (T) are saved, the others culled. If the trait distribution before selection is normal, the distribution following selection is a **truncated normal distribution.**

Recalling Equation 1.3c, we can write the conditional distribution for the density of phenotype z after selection as

$$\frac{p(z)}{\Pr(z > T)} = \frac{p(z)}{\int_T^\infty p(z)\,dz} \tag{1.11a}$$

Since the denominator, $\int_T^\infty p(z)\,dz$, is the sum of frequencies for phenotypes greater than T (i.e., the fraction of individuals allowed to reproduce), it is a measure of the intensity of selection, and we denote this fraction saved by π_T, and the conditional density is given by

$$p(z|z > T) = \frac{p(z)}{\pi_T} \quad \text{for} \quad z > T \tag{1.11b}$$

The mean trait value after selection is thus

$$E[\,z\,|\,z > T\,] = \int_t^\infty \frac{z\,p(z)}{\pi_T}\,dz = \mu_s = \mu + \frac{\sigma \cdot p_T}{\pi_T} \tag{1.12a}$$

where p_T is the height of the standard normal curve at the truncation point,

$$p_T = (2\pi)^{-1/2} \exp\left[-\frac{(T-\mu)^2}{2\,\sigma^2}\right]$$

In a similar fashion, the variance of the selected population can be shown to be

$$\sigma_s^2 = \left[1 + \frac{p_T \cdot (z-\mu)/\sigma}{\pi_T} - \left(\frac{p_T}{\pi_T}\right)^2\right]\sigma^2 \tag{1.12b}$$

The quantity within brackets gives the fraction of phenotypic variance remaining after selection.

Covariances

One of the most useful measures in quantitative genetics is the **covariance** between two variables, which is a (linear) measure of association. Formally, the covariance, $Cov(x, y)$, of two random variables x and y is defined by

$$\begin{aligned}
Cov(x, y) &= E[\,(x - \mu_x) * (y - \mu_y)\,] \\
&= E(\,xy\,) - \mu_x\mu_y \\
&= \text{mean of the product} - \text{product of the means}
\end{aligned} \tag{1.13}$$

As the figure (below) shows, if x and y are positively associated, then $Cov(x, y) > 0$, while if they are negatively associated, then $Cov(x, y) < 0$. Note that the covariance is a measure of the *linear* association between two variables — even though x perfectly predicts y is the far right panel, there is no *linear* trend, so that $Cov(x, y) = 0$. While $Cov(x, y) = 0$ when x and y are independent, the converse is NOT true, as $Cov(x, y) = 0$ does not necessarily imply that x and y are independent (again, as evidenced by the last panel).
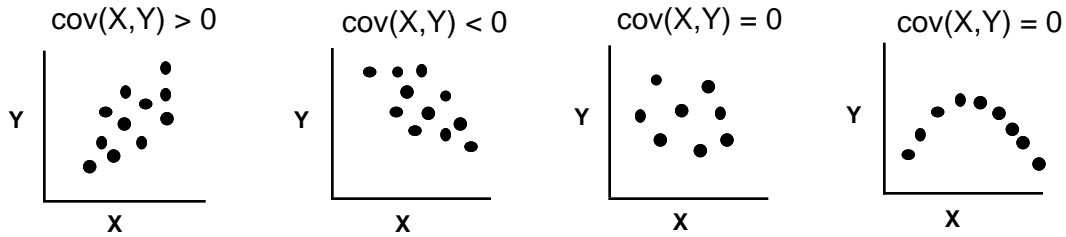
The covariance is estimated for a sample of n paired observations (x_i, y_i) by

$$Cov(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i y_i - n\overline{x} \cdot \overline{y} \right) \tag{1.14}$$

The **correlation**, $r(x, y)$ is a scaled measure of the covariance, where

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x) Var(y)}} \tag{1.15a}$$

Since the range of correlation is restricted between to -1 and $+1$, it provides a standard metric for comparing the amount of association between pairs of variables that show different levels of variation. For example, a covariance of 10 implies a relatively small association if both variables have a variance of 100 ($r = 10/100 = 0.1$), but complete association if both variables have a variance of 10 ($r = 10/10 = 1$). Rearranging Equation 1.15a,

$$Cov(x, y) = r(x, y)\sqrt{Var(x) Var(y)} \tag{1.15b}$$

showing that a weak correlation can still give a large covariance if the variances of x and y are large,

Useful Properties of Variances and Covariances

- The covariance function is symmetric, $Cov(x, y) = Cov(y, x)$

- The covariance of a variable with itself is the variance, e.g., $Cov(x, x) = Var(x)$

- If a is a constant, then $Cov(ax, y) = a \cdot Cov(x, y)$

- $Var(ax) = a^2 Var(x)$. This follows since $Var(ax) = Cov(ax, ax) = a^2 Cov(x, x) = a^2 Var(x)$

- $Cov(x + y, z) = Cov(x, z) + Cov(y, z)$, i.e., the covariance of a sum is the sum of covariances. More generally,

$$Cov \left(\sum_{i=1}^{n} x_i, \sum_{j=1}^{m} y_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(x_i, y_j) \tag{1.16}$$

- $Var(x + y) = Var(x) + Var(y) + 2Cov(x, y)$. Hence, the variance of a sum, $Var(x + y)$, equals the sum of the variances, $Var(x) + Var(y)$, only when the variables have a covariance of zero.

Regressions

Suppose we have a scatterplot of x and y values and we wish to find the best linear relationship for predicting y given an observed x value. Thus, we wish to solve for a and b in the **linear regression**

$$y = a + bx + e \tag{1.17}$$

where $\hat{y} = a + bx$ is the predicted value for y given x and the **residual** $e = y - \hat{y}$ is the difference between the true and predicted values of y. When information on x is used to predict y, x is referred to as the **predictor** or **independent variable** and y as the **response** or **dependent variable**.

   The standard solution for a and b is to use **least-squares**, finding the values of a and b the minimize the sum of squared residuals,

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2 \tag{1.18}$$

The solutions are

$$a = \overline{y} - b\,\overline{x} \tag{1.19a}$$

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)} \tag{1.19b}$$

The slope b (the **regression coefficient**) can also be written as $b_{y|x}$ to signify that the slope is for the regression of y on x, i.e., the demoniator in b is the variance of x. Equation 1.19 implies that the predicted value $\hat{y}$ for y given we know x is

$$\hat{y} = \overline{y} + b_{y\,|\,x}(\,x - \overline{x}\,) \tag{1.20}$$

Thus, the least-squares estimators for the intercept and slope of a linear regression are simple functions of the observed means, variances, and covariances. From the standpoint of quantitative genetics, this property is exceedingly useful, since such statistics are readily obtainable from phenotypic data.

Properties of Least-squares Regressions

1. *The regression line passes through the means of both x and y.* Hence $\overline{y} = a + b\,\overline{x}$.

2. *The average value of the residual is zero, $\overline{e} = 0$.*

3. *For any set of paired data, the least-squares regression parameters, a and b, define the straight line that maximizes the amount of variation in y that can be explained by a linear regression on x.* Since $\overline{e} = 0$, it follows that the variance of residual errors about the regression is simply $\overline{e^2}$. This variance is the quantity minimized by the least-squares procedure.

4. *The residual errors around the least-squares regression are uncorrelated with the predictor variable x.* This statement follows since

$$\mathrm{Cov}(x, e) = \mathrm{Cov}[\,x, (y - a - b\,x)\,] = \mathrm{Cov}(x, y) - \mathrm{Cov}(x, a) - b\,\mathrm{Cov}(x, x)$$
$$= \mathrm{Cov}(x, y) - 0 - b\,\mathrm{Var}(x)$$
$$= \mathrm{Cov}(x, y) - \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}\,\mathrm{Var}(x) = 0$$

   Note, however, that $\mathrm{Cov}(x, e) = 0$ does not guarantee that e and x are independent. If the true regression is nonlinear, then $E(\,e\,|\,x\,) \neq 0$ for some x values, and the predictive power

of the linear model is compromised. Even if the true regression is linear, the variance of the residual errors may vary with x, in which case the regression is said to display **heteroscedasticity**. If the conditional variance of the residual errors given any specified x value, $\sigma^2(\,e\,|\,x\,)$, is a constant (i.e., independent of the value of x), then the regression is said to be **homoscedastic**.

5. *When x are y are bivariate normally distribute, the true regression, the value of $E(\,y\,|\,x\,)$, is both linear and homoscedastic.*

6. *The regression of y on x is different from the regression of x on y unless the means and variances of the two variables are equal.* This distinction is made by denoting the regression coefficient by $b(y, x)$ or $b_{y,x}$ when x is the predictor and y the response variable.

Regressions and Correlations

Correlations and regression slopes are related as follows:

$$r(x, y) = \frac{Cov(x, y)}{\sqrt{Var(x)Var(y)}} = \frac{Cov(x, y)}{Var(x)}\sqrt{\frac{Var(x)}{Var(y)}} = b_{y|x}\sqrt{\frac{Var(x)}{Var(y)}} \tag{1.21}$$

Thus, if the variances of x and y are the same, then $r(x, y) = b_{y|x} = b_{x|y}$.

The correlation coefficient has two other useful properties:

1. *r is a standardized regression coefficient (the regression coefficient resulting from rescaling x and y such that each has unit variance).* Letting $x' = x/\sqrt{Var(x)}$ and $y' = y/\sqrt{Var(y)}$ gives $Var(x') = Var(y') = 1$, implying

$$b(y', x') = b(x', y') = \text{Cov}(x', y') = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\,\text{Var}(y)}} = r \tag{1.22}$$

Thus, when variables are standardized, the regression coefficient is equal to the correlation coefficient regardless of whether x' or y' is chosen as the predictor variable.

2. *The squared correlation coefficient measures the proportion of the variance in y that is explained by assuming that $E(y|x)$ is linear.* The variance of the response variable y has two components:

$$Var(y|x) = r^2\, Var(y) \tag{1.23a}$$

the amount of variance accounted for by the linear model (the **regression variance**), and

$$Var(y - \hat{y}|x) = Var(e|x) = (1 - r^2)\, Var(y) \tag{1.23b}$$

the remaining variance not accountable by the regression (the **residual variance**).

Maximum Likelihood

The method of maximum likelihood (ML) for estimating and hypothesis testing is very widely used in genetics and quantitative genetics. The basic idea underlying ML is quite simple. Usually, when specifying a probability density function (say, a normal with unknown mean μ and unit variance), we treat the pdf as a function of z (the value of the random variable) with the distribution parameters Θ assumed to be known. With maximum likelihood estimation, we reverse the roles of the observed value and the distribution parameters by asking: Given a vector of observations $\mathbf{z}$, what can we

say about Θ? To specify this alternative interpretation, the density function is denoted as $\ell(\Theta \mid \mathbf{z})$, the **likelihood** of Θ given the observed vector of data $\mathbf{z}$. This defines a **likelihood surface**, as $\ell(\Theta \mid \mathbf{z})$ assigns a value to each possible point in the Θ-parameter space given the observed data $\mathbf{z}$. The **maximum likelihood estimate** (MLE) of the unknown parameters, $\widehat{\Theta}$, is the value of Θ corresponding to the maximum of $\ell(\Theta \mid \mathbf{z})$, i.e., the MLE is the value of Θ that is "most likely" to have produced the data $\mathbf{z}$. It is usually easier to find the maximum of a likelihood function by first taking its log and working with the resulting **log-likelihood**

$$L(\Theta \mid \mathbf{z}) = \ln\left[\,\ell(\Theta \mid \mathbf{z})\,\right] \tag{1.24}$$

L is also referred to as the **support**. Since the natural log is a monotonic function, $\ell(\Theta)$ has the same maxima as $\ln\left[\ell(\Theta)\right]$, so that the maximum of L also corresponds to the maximum of the likelihood function. The **score** S of a likelihood function is the first derivative of L with respect to the likelihood parameters, $S(\theta) = \partial L(\theta)/\partial\theta$ s. From elementary calculus it follows that the score evaluated at the MLE is zero, $\mathbf{S}(\widehat{\Theta}) = \mathbf{0}$. This provides one approach for obtaining MLEs.

Large-sample Properties of MLEs

MLEs have several important features when the sample size is large:

1. **Consistency**: As the sample size increases, the MLE converges to the true parameter value, e.g., $\widehat{\Theta} \to \Theta$.

2. **Invariance**: If $f(\Theta)$ is a function of the unknown parameters of the distribution, then the MLE of $f(\Theta)$ is $f(\widehat{\Theta})$, i.e., the MLE of a function of the parameters is simply that function evaluated at the MLE. For example, the MLE of $\sqrt{\theta} = (\,\widehat{\theta}\,)^{1/2}$.

3. **Asymptotic normality and efficiency**: As the sample size increases, the sampling distribution of the MLE converges to a normal and (generally) no other estimation procedure has a smaller variance. Hence, for sufficiently large sample sizes, estimates obtained via maximum likelihood typically have the smallest confidence intervals.

4. **Variance**: For large sample sizes, the variance of an MLE (assuming a single unknown parameter) is approximately the negative of the reciprocal of the second derivative of the log-likelihood function evaluated at the MLE $\widehat{\theta}$,

$$\sigma^2\big(\,\widehat{\theta}\,\big) \simeq -\left(\left.\frac{\partial^2\ L(\theta \mid \mathbf{z})}{\partial\,\theta^2}\,\right|_{\theta=\widehat{\theta}}\right)^{-1} \tag{1.25}$$

   This is just the reciprocal of the curvature of the log-likelihood surface at the MLE. The flatter the likelihood surface around its maximum value (the MLE), the larger the variance; the steeper the surface, the smaller the variance. The minus sign appears because the second derivative is negative (downward curvature) at the maximum of the likelihood function.

Likelihood Ratio Tests

Maximum likelihood provides for extremely convenient tests of hypotheses in the form of **likelihood-ratio**, or LR, tests that examine whether a reduced model provides the same fit as a full model. The likelihood-ratio test statistic is given by

$$LR = 2\ln\left(\frac{\ell(\widehat{\Theta} \mid \mathbf{z})}{\ell(\widehat{\Theta}_r \mid \mathbf{z})}\right) = -2\ln\left(\frac{\ell(\widehat{\Theta}_r \mid \mathbf{z})}{\ell(\widehat{\Theta} \mid \mathbf{z})}\right) = -2\left[L(\widehat{\Theta}_r \mid \mathbf{z}) - L(\widehat{\Theta} \mid \mathbf{z})\right] \tag{1.26}$$

$\ell(\widehat{\Theta} \mid \mathbf{z})$ is the likelihood evaluated at the MLE and $\ell(\widehat{\Theta}_r \mid \mathbf{z})$ is the maximum of the likelihood function, subject to the restriction that r parameters unconstrained in the full likelihood analysis are assigned fixed values. For example, a likelihood ratio test that two loci are unlinked (the recombination fraction $\Theta = 1/2$ is

$$LR = 2\ln\left(\frac{\ell(\widehat{\Theta} \mid \mathbf{z})}{\ell(0.5 \mid \mathbf{z})}\right)$$

For sufficiently large sample size, *the LR test statistic is approximately χ_r^2-distributed*, a χ^2 with r degrees of freedom, although there can be exceptions when the null hypthesis can take on a point-mass value (i.e., have a nonzero probability of taking a discrete value such as 0).

Bayesian Statistics

Bayesian statistical analysis is becoming very popular in quantitative genetics. Informally, Bayesian analysis is a natural extension of maximum likelihood. Indeed, many of the desirable features of maximum likelihood are large sample properties, while Bayesian analysis is exact for small sample. The basic idea of Bayesian statistics follows, not surprisingly, from Bayes theorem. We are ultimately interested in the distribution $p(\theta|\mathbf{x})$ – the conditional distribution of the unknown parameter(s) given the data. This is also known as the **posterior distribution**. Note immediately that unlike maximum likelihood where we consider a **point estimator** (the MLE of some parameter) a Bayesian analysis is concerned with the entire *distribution* of the unknown parameter as the object of interest.

To obtain an expression for the posterior, recall Bayes' theorem (Equation 1.4):

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\,p(\theta)}{\int p(\mathbf{x}|\theta)p(\theta)d\theta} = C \cdot p(\mathbf{x}|\theta)\,p(\theta) \tag{1.27}$$

$p(\mathbf{x}|\theta)$ is just the likelihood, while $p(\theta)$ is the **prior** and represents any prior knowledge we may have about the unknown parameter. Commonly used are **flat** or **uninformative priors** in which the prior is set equal to a constant over some interval and zero elsewhere. The integral in the denominator of Equation 1.27 is to ensure that the distribution integrates to one.

One reason that Bayesian methods have recently become very popular is that the very difficult issues of analytically obtaining the full posterior distribution for interesting problems has been complete circumvented by **Markov Chain Monte Carlo** (**MCMC**) methods (such as the **Gibbs sampler**). These approaches allow us to use very intensive computer iterations to simulate draws from any distribution of interest.

One of the nicest features about a bayesian analysis is the **marginal posterior distribution** — suppose we are really only interested in estimating one particular parameter, but to do so we have to estimate not only that parameter but also a number of so-called **nuisance parameters**. To obtain a marginal posterior distribution for the parameter of interest, one simply integrates the posterior over all of the nuisance parameters. The resulting distribution is now only a function of our parameter of interest. By integrating over the posterior, we have completely and fully accounted for how all of the uncertainly in estimating the nuisance parameters translates into uncertainty in the parameter of interest.

Lecture 1 Problems

1. Suppose $x_1, \cdots x_n$ are n uncorrelated random variables, all with the same variance σ^2. Compute $Cov(x_i, \bar{x})$ — the covariance between a sample point and the mean.

2. There are two types of families in your population – normal families with an equal chance of a male or female offspring and sex-bias families that have only females. Suppose the population frequencies of normal families is 0.99 and of sex-bias families is 0.01. What is the probability that a family with seven girls (and no boys) is a sex-bias family? Hint: Think Bayes.

3. Suppose $Cov(x, y) = 20$, $Var(x) = 20$, $Var(y) = 40$, $\mu_x = 10$, $\mu_y = 30$. Compute:

    a. The regression of y on x

    b. The regression of x on y

    c. The correlation between y on x

    d. What fraction of the total variance in x is accounted for by knowing the value of y?

Solutions for Lecture 1 Problems

1.
$$Cov(x_i, \overline{x}) = Cov\left(x_i, \frac{1}{n}\sum_{j=1}^{n} x_j\right) = \frac{1}{n}\sum_{j=1}^{n} Cov(x_i, x_j) = \frac{Cov(x_i, x_i)}{n} = \frac{\sigma^2}{n}$$

2. From Bayes' thereom,

$$\Pr(\text{sex bias}|7 \text{ girls}) = \frac{Pr(7 \text{ girls}|\text{sex bias}) \cdot \Pr(\text{sex bias})}{\Pr(7 \text{ girls}|\text{sex bias}) \cdot \Pr(\text{sex bias}) + \Pr(7 \text{ girls}|\text{normal}) \cdot \Pr(\text{normal})}$$

Since
$$\Pr(7 \text{ girls}|\text{normal}) = (1/2)^7, \qquad \Pr(7 \text{ girls}|\text{sex bias}) = 1$$

$$\Pr(\text{sex bias}|7 \text{ girls}) = \frac{1 \cdot 0.01}{1 \cdot 0.01 + (1/2)^7 \cdot 0.99} = 0.56$$

3. a. The regression of y on x:

$$b_{y|x} = \frac{Cov(x,y)}{Varx} = \frac{20}{20} = 1, \quad y = \mu_y + b_{y|x}(x - \mu_x) = 30 + (x - 10) = x + 20$$

b. The regression of x on y:

$$b_{y|x} = \frac{Cov(x,y)}{Vary} = \frac{20}{40} = 0.5, \quad x = \mu_x + b_{x|y}(y - \mu_y) = 10 + 0.5(y - 30) = 0.5y - 5$$

c. The correlation between y on x

$$r(x,y) = \frac{20}{\sqrt{20 \cdot 40}} = 0.707$$

d. What fraction of the total variance in x is accounted for by knowing the value of y?

$$r^2(x,y) = 0.5$$