Lecture 21. The Multivariate Normal Distribution

21.1 Definitions and Comments

The *joint moment-generating function* of $X_1, \dots, X_n$ [also called the moment-generating function of the random vector $(X_1, \dots, X_n)$] is defined by

$$M(t_1, \dots, t_n) = E[\exp(t_1 X_1 + \cdots + t_n X_n)].$$

Just as in the one-dimensional case, the moment-generating function determines the density uniquely. The random variables $X_1, \dots, X_n$ are said to have the *multivariate normal distribution* or to be *jointly Gaussian* (we also say that the random vector $(X_1, \dots, X_n)$ is *Gaussian*) if

$$M(t_1, \dots, t_n) = \exp(t_1\mu_1 + \cdots + t_n\mu_n) \exp\left(\frac{1}{2}\sum_{i,j=1}^{n} t_i a_{ij} t_j\right)$$

where the t_i and μ_j are arbitrary real numbers, and the matrix A is symmetric and positive definite.

Before we do anything else, let us indicate the notational scheme we will be using. Vectors will be written with an underbar, and are assumed to be column vectors unless otherwise specified. If $\underline{t}$ is a column vector with components $t_1, \dots, t_n$, then to save space we write $\underline{t} = (t_1, \dots, t_n)'$. The row vector with these components is the transpose of $\underline{t}$, written $\underline{t}'$. The moment-generating function of jointly Gaussian random variables has the form

$$M(t_1, \dots, t_n) = \exp(\underline{t}'\underline{\mu}) \exp\left(\frac{1}{2}\underline{t}'A\underline{t}\right).$$

We can describe Gaussian random vectors much more concretely.

21.2 Theorem

Joint Gaussian random variables arise from nonsingular linear transformations on independent normal random variables.

Proof. Let $X_1, \dots, X_n$ be independent, with X_i normal $(0,\lambda_i)$, and let $\underline{X} = (X_1, \dots, X_n)'$. Let $\underline{Y} = B\underline{X} + \underline{\mu}$ where B is nonsingular. Then $\underline{Y}$ is Gaussian, as can be seen by computing the moment-generating function of $\underline{Y}$:

$$M_{\underline{Y}}(\underline{t}) = E[\exp(\underline{t}'\underline{Y})] = E[\exp(\underline{t}'B\underline{X})] \exp(\underline{t}'\underline{\mu}).$$

But

$$E[\exp(\underline{u}'\underline{X})] = \prod_{i=1}^{n} E[\exp(u_i X_i)] = \exp\big(\sum_{i=1}^{n} \lambda_i u_i^2/2\big) = \exp\big(\frac{1}{2}\underline{u}'D\underline{u}\big)$$

where D is a diagonal matrix with λ_i's down the main diagonal. Set $\underline{u} = B'\underline{t}, \underline{u}' = \underline{t}'B$; then

$$M_{\underline{Y}}(t) = \exp(\underline{t}'\underline{\mu}) \exp(\frac{1}{2}\underline{t}'BDB'\underline{t})$$

and BDB' is symmetric since D is symmetric. Since $\underline{t}'BDB'\underline{t} = \underline{u}'D\underline{u}$, which is greater than 0 except when $\underline{u} = \underline{0}$ (equivalently when $\underline{t} = \underline{0}$ because B is nonsingular), BDB' is positive definite, and consequently $\underline{Y}$ is Gaussian.

Conversely, suppose that the moment-generating function of $\underline{Y}$ is $\exp(\underline{t}'\underline{\mu}) \exp[(1/2)\underline{t}'A\underline{t})]$ where A is symmetric and positive definite. Let L be an orthogonal matrix such that $L'AL = D$, where D is the diagonal matrix of eigenvalues of A. Set $\underline{X} = L'(\underline{Y} - \underline{\mu})$, so that $\underline{Y} = \underline{\mu} + L\underline{X}$. The moment-generating function of $\underline{X}$ is

$$E[\exp(\underline{t}'\underline{X})] = \exp(-\underline{t}'L'\underline{\mu})E[\exp(\underline{t}'L'\underline{Y})].$$

The last term is the moment-generating function of $\underline{Y}$ with $\underline{t}'$ replaced by $\underline{t}'L'$, or equivalently, $\underline{t}$ replaced by $L\underline{t}$. Thus the moment-generating function of $\underline{X}$ becomes

$$\exp(-\underline{t}'L'\underline{\mu}) \exp(\underline{t}'L'\underline{\mu}) \exp\big(\frac{1}{2}\underline{t}'L'AL\underline{t}\big)$$

This reduces to

$$\exp\big(\frac{1}{2}\underline{t}'D\underline{t}\big) = \exp\big(\frac{1}{2}\sum_{i=1}^{n}\lambda_i t_i^2\big).$$

Therefore the X_i are independent, with X_i normal $(0, \lambda_i)$. ♣

21.3 A Geometric Interpretation

Assume for simplicity that the random variables X_i have zero mean. If $E(U) = E(V) = 0$ then the covariance of U and V is $E(UV)$, which can be regarded as an inner product. Then $Y_1 - \mu_1, \dots, Y_n - \mu_n$ span an n-dimensional space, and $X_1, \dots, X_n$ is an orthogonal basis for that space. We will see later in the lecture that orthogonality is equivalent to independence. (Orthogonality means that the X_i are uncorrelated, i.e., $E(X_iX_j) = 0$ for $i \neq j$.)

21.4 Theorem

Let $\underline{Y} = \underline{\mu} + L\underline{X}$ as in the proof of (21.2), and let A be the symmetric, positive definite matrix appearing in the moment-generating function of the Gaussian random vector $\underline{Y}$. Then $E(Y_i) = \mu_i$ for all i, and furthermore, A is the *covariance matrix* of the Y_i, in other words, $a_{ij} = \operatorname{Cov}(Y_i, Y_j)$ (and $a_{ii} = \operatorname{Cov}(Y_i, Y_i) = \operatorname{Var} Y_i$).

It follows that the means of the Y_i and their covariance matrix determine the moment-generating function, and therefore the density.

Proof. Since the X_i have zero mean, we have $E(Y_i) = \mu_i$. Let K be the covariance matrix of the Y_i. Then K can be written in the following peculiar way:

$$K = E\left\{ \begin{bmatrix} Y_1 - \mu_1 \\ \vdots \\ Y_n - \mu_n \end{bmatrix} (Y_1 - \mu_1, \dots, Y_n - \mu_n) \right\}.$$

Note that if a matrix M is n by 1 and a matrix N is 1 by n, then MN is n by n. In this case, the ij entry is $E[(Y_i - \mu_i)(Y_j - \mu_j)] = \text{Cov}(Y_i, Y_j)$. Thus

$$K = E[(\underline{Y} - \underline{\mu})(\underline{Y} - \underline{\mu})'] = E(L\underline{X}\underline{X}'L') = LE(\underline{X}\underline{X}')L'$$

since expectation is linear. [For example, $E(M\underline{X}) = ME(\underline{X})$ because $E(\sum_j m_{ij}X_j) = \sum_j m_{ij}E(X_j)$.] But $E(\underline{X}\underline{X}')$ is the covariance matrix of the X_i, which is D. Therefore $K = LDL' = A$ (because $L'AL = D$). ♣

21.5 Finding the Density

From $\underline{Y} = \underline{\mu} + L\underline{X}$ we can calculate the density of $\underline{Y}$. The Jacobian of the transformation from $\underline{X}$ to $\underline{Y}$ is $\det L = \pm 1$, and

$$f_{\underline{X}}(x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi})^n} \frac{1}{\sqrt{\lambda_1 \cdots \lambda_n}} \exp\big(-\sum_{i=1}^n x_i^2/2\lambda_i\big).$$

We have $\lambda_1 \cdots \lambda_n = \det D = \det K$ because $\det L = \det L' = \pm 1$. Thus

$$f_{\underline{X}}(x_1, \dots, x_n) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det K}} \exp\big(-\frac{1}{2}\underline{x}'D^{-1}\underline{x}\big).$$

But $\underline{y} = \underline{\mu} + L\underline{x}$, $\quad \underline{x} = L'(\underline{y} - \underline{\mu})$, $\quad \underline{x}'D^{-1}\underline{x} = (\underline{y} - \underline{\mu})'LD^{-1}L'(\underline{y} - \underline{\mu})$, and [see the end of (21.4)] $K = LDL'$, $\quad K^{-1} = LD^{-1}L'$. The density of $\underline{Y}$ is

$$f_{\underline{Y}}(y_1, \dots, y_n) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det K}} \exp\big[-\frac{1}{2}(\underline{y} - \underline{\mu})'K^{-1}(\underline{y} - \underline{\mu})\big].$$

21.6 Individually Gaussian Versus Jointly Gaussian

If $X_1, \dots, X_n$ are jointly Gaussian, then each X_i is normally distributed (see Problem 4), but *not conversely*. For example, let X be normal (0,1) and flip an unbiased coin. If the coin shows heads, set $Y = X$, and if tails, set $Y = -X$. Then Y is also normal (0,1) since

$$P\{Y \le y\} = \frac{1}{2}P\{X \le y\} + \frac{1}{2}P\{-X \le y\} = P\{X \le y\}$$

because $-X$ is also normal (0,1). Thus $F_X = F_Y$. But with probability 1/2, $X + Y = 2X$, and with probability 1/2, $X + Y = 0$. Therefore $P\{X + Y = 0\} = 1/2$. If X and Y were jointly Gaussian, then $X + Y$ would be normal (Problem 4). We conclude that X and Y are individually Gaussian but not jointly Gaussian.

21.7 Theorem

If $X_1, \dots, X_n$ are jointly Gaussian and uncorrelated ($\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$), then the X_i are independent.

Proof. The moment-generating function of $\underline{X} = (X_1, \dots, X_n)$ is

$$M_{\underline{X}}(\underline{t}) = \exp(\underline{t}'\underline{\mu}) \exp\big(\frac{1}{2}\underline{t}'K\underline{t}\big)$$

where K is a diagonal matrix with entries $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ down the main diagonal, and 0's elsewhere. Thus

$$M_{\underline{X}}(\underline{t}) = \prod_{i=1}^{n} \exp(t_i\mu_i) \exp\big(\frac{1}{2}\sigma_i^2 t_i^2\big)$$

which is the joint moment-generating function of independent random variables $X_1, \dots, X_n$, whee X_i is normal (μ_i, σ_i^2). ♣

21.8 A Conditional Density

Let $X_1, \dots, X_n$ be jointly Gaussian. We find the conditional density of X_n given $X_1, \dots, X_{n-1}$:

$$f(x_n|x_1, \dots, x_{n-1}) = \frac{f(x_1, \dots, x_n)}{f(x_1, \dots, x_{n-1})}$$

with

$$f(x_1, \dots, x_n) = (2\pi)^{-n/2}(\det K)^{-1/2} \exp\big[-\frac{1}{2}\sum_{i,j=1}^{n} y_i q_{ij} y_j\big]$$

where $Q = K^{-1} = [q_{ij}], y_i = x_i - \mu_i$. Also,

$$f(x_1, \dots, x_{n-1}) = \int_{-\infty}^{\infty} f(x_1, \dots, x_{n-1}, x_n)\, dx_n = B(y_1, \dots, y_{n-1}).$$

Now

$$\sum_{i,j=1}^{n} y_i q_{ij} y_j = \sum_{i,j=1}^{n-1} y_i q_{ij} y_j + y_n \sum_{j=1}^{n-1} q_{nj} y_j + y_n \sum_{i=1}^{n-1} q_{in} y_i + q_{nn} y_n^2.$$

Thus the conditional density has the form

$$\frac{A(y_1, \dots, y_{n-1})}{B(y_1, \dots, y_{n-1})} \exp[-(Cy_n^2 + D(y_1, \dots, y_{n-1})y_n]$$

with $C = (1/2)q_{nn}$, $\quad D = \sum_{j=1}^{n-1} q_{nj} y_j = \sum_{i=1}^{n-1} q_{in} y_i$ since $Q = K^{-1}$ is symmetric. The conditional density may now be expressed as

$$\frac{A}{B} \exp\big(\frac{D^2}{4C}\big) \exp\Big[-C(y_n + \frac{D}{2C})^2\Big].$$

We conclude that

$$\boxed{\text{given} \quad X_1, \dots, X_{n-1}, \quad X_n \text{ is normal.}}$$

The conditional variance of X_n (the same as the conditional variance of $Y_n = X_n - \mu_n$) is

$$\frac{1}{2C} = \frac{1}{q_{nn}} \quad \text{because} \quad \frac{1}{2\sigma^2} = C, \sigma^2 = \frac{1}{2C}.$$

Thus

$$\boxed{\operatorname{Var}(X_n | X_1, \dots, X_{n-1}) = \frac{1}{q_{nn}}}$$

and the conditional mean of Y_n is

$$-\frac{D}{2C} = -\frac{1}{q_{nn}} \sum_{j=1}^{n-1} q_{nj} Y_j$$

so the conditional mean of X_n is

$$\boxed{E(X_n | X_1, \dots, X_{n-1}) = \mu_n - \frac{1}{q_{nn}} \sum_{j=1}^{n-1} q_{nj}(X_j - \mu_j).}$$

Recall from Lecture 18 that $E(Y|X)$ is the best estimate of Y based on X, in the sense that the mean square error is minimized. In the joint Gaussian case, the best estimate of X_n based on $X_1, \dots, X_{n-1}$ is linear, and it follows that the best linear estimate is in fact the best overall estimate. This has important practical applications, since linear systems are usually much easier than nonlinear systems to implement and analyze.

Problems

1. Let K be the covariance matrix of *arbitrary* random variables $X_1, \dots, X_n$. Assume that K is nonsingular to avoid degenerate cases. Show that K is symmetric and positive definite. What can you conclude if K is singular?
2. If $\underline{X}$ is a Gaussian n-vector and $\underline{Y} = A\underline{X}$ with A nonsingular, show that $\underline{Y}$ is Gaussian.
3. If $X_1, \dots, X_n$ are jointly Gaussian, show that $X_1, \dots, X_m$ are jointly Gaussian for $m \leq n$.
4. If $X_1, \dots, X_n$ are jointly Gaussian, show that $c_1X_1 + \cdots + c_nX_n$ is a normal random variable (assuming it is nondegenerate, i.e., not identically constant).

Lecture 22. The Bivariate Normal Distribution

22.1 Formulas

The general formula for the n-dimensional normal density is

$$f_{\underline{X}}(x_1,\dots,x_n)=\frac{1}{(\sqrt{2\pi})^n}\frac{1}{\sqrt{\det K}}\exp\big[-\frac{1}{2}(\underline{x}-\underline{\mu})'K^{-1}(\underline{x}-\underline{\mu})\big]$$

where $E(\underline{X})=\underline{\mu}$ and K is the covariance matrix of $\underline{X}$. We specialize to the case $n=2$:

$$K=\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}=\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix},\qquad \sigma_{12}=\operatorname{Cov}(X_1,X_2);$$

$$K^{-1}=\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2\\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix}=\frac{1}{1-\rho^2}\begin{bmatrix}1/\sigma_1^2 & -\rho/\sigma_1\sigma_2\\ -\rho/\sigma_1\sigma_2 & 1/\sigma_2^2\end{bmatrix}.$$

Thus the joint density of X_1 and X_2 is

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

The moment-generating function of $\underline{X}$ is

$$M_{\underline{X}}(t_1,t_2)=\exp(\underline{t}'\underline{\mu})\exp\left(\frac{1}{2}\underline{t}'K\underline{t}\right)$$

$$=\exp\left[t_1\mu_1+t_2\mu_2+\frac{1}{2}\left(\sigma_1^2t_1^2+2\rho\sigma_1\sigma_2t_1t_2+\sigma_2^2t_2^2\right)\right].$$

If X_1 and X_2 are jointly Gaussian and uncorrelated, then $\rho=0$, so that $f(x_1,x_2)$ is the product of a function $g(x_1)$ of x_1 alone and a function $h(x_2)$ of x_2 alone. It follows that X_1 and X_2 are independent. (We proved independence in the general n-dimensional case in Lecture 21.)

From the results at the end of Lecture 21, the conditional distribution of X_2 given X_1 is normal, with

$$E(X_2|X_1=x_1)=\mu_2-\frac{q_{21}}{q_{22}}(x_1-\mu_1)$$

where

$$\frac{q_{21}}{q_{22}}=-\frac{\rho/\sigma_1\sigma_2}{1/\sigma_2^2}=-\frac{\rho\sigma_2}{\sigma_1}.$$

Thus

$$E(X_2|X_1=x_1)=\mu_2+\frac{\rho\sigma_2}{\sigma_1}(x_1-\mu_1)$$

and

$$\operatorname{Var}(X_2|X_1=x_1)=\frac{1}{q_{22}}=\sigma_2^2(1-\rho^2).$$

For $E(X_1|X_2=x_2)$ and $\operatorname{Var}(X_1|X_2=x_2)$, interchange μ_1 and μ_2, and interchange σ_1 and σ_2.

22.2 Example

Let X be the height of the father, Y the height of the son, in a sample of father-son pairs. Assume X and Y bivariate normal, as found by Karl Pearson around 1900. Assume $E(X) = 68$ (inches), $E(Y) = 69$, $\sigma_X = \sigma_Y = 2, \rho = .5$. (We expect ρ to be positive because on the average, the taller the father, the taller the son.)

Given $X = 80$ (6 feet 8 inches), Y is normal with mean

$$\mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(x - \mu_X) = 69 + .5(80 - 68) = 75$$

which is 6 feet 3 inches. The variance of Y given $X = 80$ is

$$\sigma_Y^2(1 - \rho^2) = 4(3/4) = 3.$$

Thus the son will tend to be of above average height, but not as tall as the father. This phenomenon is often called *regression*, and the line $y = \mu_Y + (\rho\sigma_Y/\sigma_X)(x - \mu_X)$ is called the *line of regression* or the *regression line.*

Problems

1. Let X and Y have the bivariate normal distribution. The following facts are known: $\mu_X = -1, \sigma_X = 2$, and the best estimate of Y based on X, i.e., the estimate that minimizes the mean square error, is given by $3X + 7$. The minimum mean square error is 28. Find μ_X, σ_Y and the correlation coefficient ρ between X and Y.
2. Show that the bivariate normal density belongs to the exponential class, and find the corresponding complete sufficient statistic.

Lecture 23. Cramér-Rao Inequality

23.1 A Strange Random Variable

Given a density $f_\theta(x), -\infty < x < \infty, \quad a < \theta < b$. We have found maximum likelihood estimates by computing $\frac{\partial}{\partial\theta} \ln f_\theta(x)$. If we replace x by X, we have a random variable. To see what is going on, let's look at a discrete example. If X takes on values x_1, x_2, x_3, x_4 with $p(x_1) = .5, p(x_2) = p(x_3) = .2, p(x_4) = .1$, then $p(X)$ is a random variable with the following distribution:

$$P\{p(X) = .5\} = .5, \quad P\{p(X) = .2\} = .4, \quad P\{p(X) = .1\} = .1$$

For example, if $X = x_2$ then $p(X) = p(x_2) = .2$, and if $X = x_3$ then $p(X) = p(x_3) = .2$. The total probability that $p(X) = .2$ is .4.

The continuous case is, at first sight, easier to handle. If X has density f and $X = x$, then $f(X) = f(x)$. But what is the density of $f(X)$? We will not need the result, but the question is interesting and is considered in Problem 1.

The following two lemmas will be needed to prove the Cramér-Rao inequality, which can be used to compute uniformly minimum variance unbiased estimates. In the calculations to follow, we are going to assume that all differentiations under the integral sign are legal.

23.2 Lemma

$$E_\theta\big[\frac{\partial}{\partial\theta} \ln f_\theta(X)\big] = 0.$$

Proof. The expectation is

$$\int_{-\infty}^{\infty} \big[\frac{\partial}{\partial\theta} \ln f_\theta(x)\big] f_\theta(x)\, dx = \int_{-\infty}^{\infty} \frac{1}{f_\theta(x)} \frac{\partial f_\theta(x)}{\partial\theta} f_\theta(x)\, dx$$

which reduces to

$$\frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} f_\theta(x)\, dx = \frac{\partial}{\partial\theta}(1) = 0. \quad \clubsuit$$

23.3 Lemma

Let $Y = g(X)$ and assume $E_\theta(Y) = k(\theta)$. If $k'(\theta) = dk(\theta)/d\theta$, then

$$k'(\theta) = E_\theta\big[Y \frac{\partial}{\partial\theta} \ln f_\theta(X)\big].$$

Proof. We have

$$k'(\theta) = \frac{\partial}{\partial\theta} E_\theta[g(X)] = \frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} g(x) f_\theta(x)\, dx = \int_{-\infty}^{\infty} g(x) \frac{\partial f_\theta(x)}{\partial\theta}\, dx$$

$$= \int_{-\infty}^{\infty} g(x) \frac{\partial f_\theta(x)}{\partial \theta} \frac{1}{f_\theta(x)} f_\theta(x)\, dx = \int_{-\infty}^{\infty} g(x) \big[\frac{\partial}{\partial \theta} \ln f_\theta(x)\big] f_\theta(x)\, dx$$

$$= E_\theta[g(X) \frac{\partial}{\partial \theta} \ln f_\theta(X)] = E_\theta[Y \frac{\partial}{\partial \theta} \ln f_\theta(X)]. \ \clubsuit$$

23.4 Cramér-Rao Inequality

Under the assumptions of (23.3), we have

$$\operatorname{Var}_\theta Y \geq \frac{[k'(\theta)]^2}{E_\theta\big[\big(\frac{\partial}{\partial \theta} \ln f_\theta(X)\big)^2\big]}.$$

Proof. By the Cauchy-Schwarz inequality,

$$[\operatorname{Cov}(V, W)]^2 = (E[(V - \mu_V)(W - \mu_W)])^2 \leq \operatorname{Var} V \operatorname{Var} W$$

hence

$$[\operatorname{Cov}_\theta(Y, \frac{\partial}{\partial \theta} \ln f_\theta(X))]^2 \leq \operatorname{Var}_\theta Y \operatorname{Var}_\theta \frac{\partial}{\partial \theta} \ln f_\theta(X).$$

Since $E_\theta[(\partial/\partial\theta) \ln f_\theta(X)] = 0$ by (23.2), this becomes

$$(E_\theta[Y \frac{\partial}{\partial \theta} \ln f_\theta(X)])^2 \leq \operatorname{Var}_\theta Y\, E_\theta[(\frac{\partial}{\partial \theta} \ln f_\theta(X))^2].$$

By (23.3), the left side is $[k'(\theta)]^2$, and the result follows. $\clubsuit$

23.5 A Special Case

Let $X_1, \dots, X_n$ be iid, each with density $f_\theta(x)$, and take $X = (X_1, \dots, X_n)$. Then $f_\theta(x_1, \dots, x_n) = \prod_{i=1}^n f_\theta(x_i)$ and by (23.2),

$$E_\theta\big[\big(\frac{\partial}{\partial \theta} \ln f_\theta(X)\big)^2\big] = \operatorname{Var}_\theta \frac{\partial}{\partial \theta} \ln f_\theta(X) = \operatorname{Var}_\theta \sum_{i=1}^n \frac{\partial}{\partial \theta} \ln f_\theta(X_i)$$

$$= n \operatorname{Var}_\theta \frac{\partial}{\partial \theta} \ln f_\theta(X_i) = n E_\theta\big[\big(\frac{\partial}{\partial \theta} \ln f_\theta(X_i)\big)^2\big].$$

23.6 Theorem

Let $X_1, \dots, X_n$ be iid, each with density $f_\theta(x)$. If $Y = g(X_1, \dots, X_n)$ is an unbiased estimate of θ, then

$$\operatorname{Var}_\theta Y \geq \frac{1}{n E_\theta\big[\big(\frac{\partial}{\partial \theta} \ln f_\theta(X_i)\big)^2\big]}.$$

Proof. Applying (23.5), we have a special case of the Cramér-Rao inequality (23.4) with $k(\theta) = \theta, k'(\theta) = 1$. ♣

The lower bound in (23.6) is $1/nI(\theta)$, where

$$I(\theta) = E_\theta\big[\big(\frac{\partial}{\partial\theta}\ln f_\theta(X_i)\big)^2\big]$$

is called the *Fisher information*.

It follows from (23.6) that if Y is an unbiased estimate that meets the Cramér-Rao inequality for all θ (an *efficient estimate*), then Y must be a UMVUE of θ.

23.7 A Computational Simplification

From (23.2) we have

$$\int_{-\infty}^{\infty}\big(\frac{\partial}{\partial\theta}\ln f_\theta(x)\big)f_\theta(x)\,dx = 0.$$

Differentiate again to obtain

$$\int_{-\infty}^{\infty}\frac{\partial^2 \ln f_\theta(x)}{\partial\theta^2}f_\theta(x)\,dx + \int_{-\infty}^{\infty}\frac{\partial \ln f_\theta(x)}{\partial\theta}\frac{\partial f_\theta(x)}{\partial\theta}\,dx = 0.$$

Thus

$$\int_{-\infty}^{\infty}\frac{\partial^2 \ln f_\theta(x)}{\partial\theta^2}f_\theta(x)\,dx + \int_{-\infty}^{\infty}\frac{\partial \ln f_\theta(x)}{\partial\theta}\big[\frac{\partial f_\theta(x)}{\partial\theta}\frac{1}{f_\theta(x)}\big]f_\theta(x)\,dx = 0.$$

But the term in brackets on the right is $\partial \ln f_\theta(x)/\partial\theta$, so we have

$$\int_{-\infty}^{\infty}\frac{\partial^2 \ln f_\theta(x)}{\partial\theta^2}f_\theta(x)\,dx + \int_{-\infty}^{\infty}\big(\frac{\partial}{\partial\theta}\ln f_\theta(x)\big)^2 f_\theta(x)\,dx = 0.$$

Therefore

$$I(\theta) = E_\theta\big[\big(\frac{\partial}{\partial\theta}\ln f_\theta(X_i)\big)^2\big] = -E_\theta\big[\frac{\partial^2 \ln f_\theta(X_i)}{\partial\theta^2}\big].$$

Problems

1. If X is a random variable with density $f(x)$, explain how to find the distribution of the random variable $f(X)$.
2. Use the Cramér-Rao inequality to show that the sample mean is a UMVUE of the true mean in the Bernoulli, normal (with σ^2 known) and Poisson cases.

Lecture 24. Nonparametric Statistics

We wish to make a statistical inference about a random variable X even though we know nothing at all about its underlying distribution.

24.1 Percentiles

Assume F continuous and strictly increasing. If $0 < p < 1$, then the equation $F(x) = p$ has a unique solution ξ_p, so that $P\{X \le \xi_p\} = p$. When $p = 1/2$, ξ_p is the median; when $p = .3$, ξ_p is the 30-th percentile, and so on.

Let $X_1, \dots, X_n$ be iid, each with distribution function F, and let $Y_1, \dots, Y_n$ be the order statistics. We will consider the problem of estimating ξ_p.

24.2 Point Estimates

On the average, np of the observations will be less than ξ_p. (We have n Bernoulli trials, with probability of success $P\{X_i < \xi_p\} = F(\xi_p) = p$.) It seems reasonable to use Y_k as an estimate of ξ_p, where k is approximately np. We can be a bit more precise. The random variables $F(X_1), \dots, F(X_n)$ are iid, uniform on (0,1) [see (8.5)]. Thus $F(Y_1), \dots, F(Y_n)$ are the order statistics from a uniform (0,1) sample. We know from Lecture 6 that the density of $F(Y_k)$ is

$$\frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, \quad 0 < x < 1.$$

Therefore

$$E[F(Y_k)] = \int_0^1 \frac{n!}{(k-1)!(n-k)!} x^k (1-x)^{n-k}\, dx = \frac{n!}{(k-1)!(n-k)!} \beta(k+1, n-k+1).$$

Now $\beta(k+1, n-k+1) = \Gamma(k+1)\Gamma(n-k+1)/\Gamma(n+2) = k!(n-k)!/(n+1)!$, and consequently

$$E[F(Y_k)] = \frac{k}{n+1}, \quad 1 \le k \le n.$$

Define $Y_0 = -\infty$ and $Y_{n+1} = \infty$, so that

$$E[F(Y_{k+1}) - F(Y_k)] = \frac{1}{n+1}, \quad 0 \le k \le n.$$

(Note that when $k = n$, the expectation is $1 - [n/(n+1)] = 1/(n+1)$, as asserted.)

The key point is that on the average, each $[Y_k, Y_{k+1}]$ produces area $1/(n+1)$ under the density f of the X_i. This is true because

$$\int_{Y_k}^{Y_{k+1}} f(x)\, dx = F(Y_{k+1}) - F(Y_k)$$

and we have just seen that the expectation of this quantity is $1/(n+1)$, $\quad k = 0, 1, \dots, n$. If we want to accumulate area p, set $k/(n+1) = p$, that is, $k = (n+1)p$.

Conclusion: If $(n+1)p$ is an integer, estimate ξ_p by $Y_{(n+1)p}$.
If $(n+1)p$ is not an integer, we can use a weighted average. For example, if $p = .6$ and $n = 13$ then $(n+1)p = 14 \times .6 = 8.4$. Now if $(n+1)p$ were 8, we would use Y_8, and if $(n+1)p$ were 9 we would use Y_9. If $(n+1)p = 8 + \lambda$, we use $(1-\lambda)Y_8 + \lambda Y_9$. In the present case, $\lambda = .4$, so we use $.6Y_8 + .4Y_9 = Y_8 + .4(Y_9 - Y_8)$.

24.3 Confidence Intervals

Select order statistics Y_i and Y_j, where i and j are (approximately) symmetrical about $(n+1)p$. Then $P\{Y_i < \xi_p < Y_j\}$ is the probability that the number of observations less than ξ_p is at least i but less than j, i.e., between i and $j-1$, inclusive. The probability that exactly k observations will be less than ξ_p is $\binom{n}{k}p^k(1-p)^{n-k}$, hence

$$P\{Y_i < \xi_p < Y_j\} = \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}.$$

Thus (Y_i, Y_j) is a confidence interval for ξ_p, and we can find the confidence level by evaluating the above sum, possibly with the aid of the normal approximation to the binomial.

24.4 Hypothesis Testing

First let's look at a numerical example. The 30-th percentile $\xi_{.3}$ will be less than 68 precisely when $F(\xi_{.3}) < F(68)$, because F is continuous and strictly increasing. Therefore $\xi_{.3} < 68$ iff $F(68) > .3$. Similarly, $\xi_{.3} > 68$ iff $F(68) < .3$, and $\xi_{.3} = 68$ iff $F(68) = .3$. In general,

$$\xi_{p_0} < \xi \iff F(\xi) > p_0, \quad \xi_{p_0} > \xi \iff F(\xi) < p_0$$

and

$$\xi_{p_0} = \xi \iff F(\xi) = p_0.$$

In our numerical example, if $F(68)$ were actually .4, then on the average, 40 percent of the observations will be 68 or less, as opposed to 30 percent if $F(68) = .3$. Thus a larger than expected number of observations less than or equal to 68 will tend to make us reject the hypothesis that the 30-th percentile is exactly 68. In general, our problem will be

$$H_0 : \xi_{p_0} = \xi \quad (\iff F(\xi) = p_0)$$

$$H_1 : \xi_{p_0} < \xi \quad (\iff F(\xi) > p_0)$$

where p_0 and ξ are specified. If Y is the number of observations less than or equal to ξ, we propose to reject H_0 if $Y \geq c$. (If H_1 is $\xi_{p_0} > \xi$, i.e., $F(\xi) < p_0$, we reject if $Y \leq c$.) Note that Y is the number of nonpositive signs in the sequence $X_1 - \xi, \dots, X_n - \xi$, and for this reason, the terminology *sign test* is used.

Since we are trying to determine whether $F(\xi)$ is equal to p_0 or greater than p_0, we may regard $\theta = F(\xi)$ as the unknown state of nature. The power function of the test is

$$K(\theta) = P_\theta\{Y \geq c\} = \sum_{k=c}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

and in particular, the significance level (probability of a type 1 error) is $\alpha = K(p_0)$.

The above confidence interval estimates and the sign test are *distribution free*, that is, independent of the underlying distribution function F.

Problems are deferred to Lecture 25.

Lecture 25. The Wilcoxon Test

We will need two formulas:

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}, \quad \sum_{k=1}^{n} k^3 = \left[\frac{n(n+1)}{2}\right]^2.$$

For a derivation via the calculus of finite differences, see my on-line text "A Course in Commutative Algebra", Section 5.1.

The hypothesis testing problem addressed by the Wilcoxon test is the same as that considered by the sign test, except that:

(1) We are restricted to testing the *median* $\xi_{.5}$.

(2) We assume that $X_1, \dots, X_n$ are iid and the underlying density is symmetric about the median (so we are not quite nonparametric). There are many situations where we suspect an underlying normal distribution but are not sure. In such cases, the symmetry assumption may be reasonable.

(3) We use the magnitudes as well as the signs of the deviations $X_i - \xi_{.5}$, so the Wilcoxon test should be more accurate than the sign test.

25.1 How The Test Works

Suppose we are testing $H_0 : \xi_{.5} = m$ vs. $H_1 : \xi_{.5} > m$ based on observations $X_1, \dots, X_n$. We rank the absolute values $|X_i - m|$ from smallest to largest. For example, let $n = 5$ and $X_1 - m = 2.7, X_2 - m = -1.3, X_3 - m = -0.3, X_4 - m = -3.2, X_5 - m = 2.4$. Then

$$|X_3 - m| < |X_2 - m| < |X_5 - m| < |X_1 - m| < |X_4 - m|.$$

Let R_i be the rank of $|X_i - m|$, so that $R_3 = 1, R_2 = 2, R_5 = 3, R_1 = 4, R_4 = 5$. Let Z_i be the sign of $X_i - m$, so that $Z_i = \pm 1$. Then $Z_3 = -1, Z_2 = -1, Z_5 = 1, Z_1 = 1, Z_4 = -1$. The Wilcoxon statistic is

$$W = \sum_{i=1}^{n} Z_i R_i.$$

In this case, $W = -1 - 2 + 3 + 4 - 5 = -1$. Because the density is symmetric about the median, if R_i is given then Z_i is still equally likely to be ± 1, so $(R_1, \dots, R_n)$ and $(Z_1, \dots, Z_n)$ are independent. (Note that if R_j is given, the odds about $Z_i (i \neq j)$ are unaffected since the observations $X_1, \dots, X_n$ are independent.) Now the R_i are simply a permutation of $(1, 2, \dots, n)$, so

W is a sum of independent random variables V_i where $V_i = \pm i$ with equal probability.

25.2 Properties Of The Wilcoxon Statistic

Under $H_0, E(V_i) = 0$ and $\operatorname{Var} V_i = E(V_i^2) = i^2$, so

$$E(W) = \sum_{i=1}^{n} E(V_i) = 0, \quad \operatorname{Var} W = \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

The V_i do not have the same distribution, but the central limit theorem still applies because *Liapounov's condition* is satisfied:

$$\frac{\sum_{i=1}^{n} E[|V_i - \mu_i|^3]}{\left(\sum_{i=1}^{n} \sigma_i^2\right)^{3/2}} \to 0 \quad \text{as} \quad n \to \infty.$$

Now the V_i have mean $\mu_i = 0$, so $|V_i - \mu_i|^3 = |V_i|^3 = i^3$ and $\sigma_i^2 = \operatorname{Var} V_i = i^2$. Thus the Liapounov fraction is the sum of the first n cubes divided by the 3/2 power of the sum of the first n squares, which is

$$\frac{n^2(n+1)^2/4}{[n(n+1)(2n+1)/6]^{3/2}}.$$

For large n, the numerator is of the order of n^4 and the denominator is of the order of $(n^3)^{3/2} = n^{9/2}$. Therefore the fraction is of the order of $1/\sqrt{n} \to 0$ as $n \to \infty$. By the central limit theorem, $[W - E(W)]/\sigma(W)$ is approximately normal (0,1) for large n, with $E(W) = 0$ and $\sigma^2(W) = n(n+1)(2n+1)/6$.

If the median is larger than its value m under H_0, we expect W to have a positive bias. Thus we reject H_0 if $W \geq c$. (If H_1 were $\xi_{.5} < m$), we would reject if $W \leq c$.) The value of c is determined by our choice of the significance level α.

Problems

1. Suppose we are using a sign test with $n = 12$ observations to decide between the null hypothesis $H_0 : m = 40$ and the alternative $H_1 : m > 40$, whee m is the median. We use the statistic Y = the number of observations that are less than or equal to 40. We reject H_0 if and only if $Y \leq c$. Find the power function $K(p)$ in terms of c and $p = F(40)$, and the probability α of a type 1 error for $c = 2$.
2. Let m be the median of a random variable with density symmetric about m. Using the Wilcoxon test, we are testing $H_0 : m = 160$ vs. $H_1 : m > 160$ based on $n = 16$ observations, which are as follows: 176.9, 158.3, 152.1, 158.8, 172.4, 169.8, 159.7, 162.7, 156.6, 174.5, 184.4, 165.2, 147.8, 177.8, 160.1, 160.5. Compute the Wilcoxon statistic and determine whether H_0 is rejected at the .05 significance level, i.e., the probability of a type 1 error is .05.
3. When n is small, the distribution of W can be found explicitly. Do it for $n = 1, 2, 3$.