Probability Lecture III (August, 2006)

1 Some Properties of Random Vectors and Matrices

We generalize univariate notions in this section.

Definition 1 *Let* $\mathbf{U} = ||U_{ij}||_{k\times l}$*, a matrix of random variables. Suppose* $E|U_{ij}| < \infty$ *for all* i, j*. Define the expectation of* $\mathbf{U}$ *by*

$$E(\mathbf{U}) = (E(U_{ij}))_{k\times l}.$$

The following are some properties of random vectors:
Let $\mathbf{U}$, respectively $\mathbf{V}$, denote a random k, respectively l, vector.

1. If $\mathbf{A}_{m\times k}, \mathbf{B}_{m\times l}$ are nonrandom and $E\mathbf{U}$, $E\mathbf{V}$ are defined, then

$$E(\mathbf{AU} + \mathbf{BV}) = \mathbf{A}E(\mathbf{U}) + \mathbf{B}E(\mathbf{V}).$$

Definition 2 *For a random vector* $\mathbf{U}$*, suppose* $EU_i^2 < \infty$ *for* $i = 1, \cdots, k$ *or equivalently* $E(|\mathbf{U}|^2) < \infty$*, where* $|\cdot|$ *denotes Euclidean distance. Define the variance of* $\mathbf{U}$*, often called the variance-covariance matrix, by*

$$\begin{aligned}\mathrm{Var}(\mathbf{U}) &= E(\mathbf{U} - E(\mathbf{U}))(\mathbf{U} - E(\mathbf{U}))^T \\ &= (\mathrm{Cov}(U_i, U_j))_{k\times k}\end{aligned}$$

a symmetric matrix.

2. If $\mathbf{A}$ is $m \times k$ as before,

$$\mathrm{Var}(\mathbf{AU}) = \mathbf{A}\ \mathrm{Var}(\mathbf{U})\mathbf{A}^T.$$

Note that $\mathrm{Var}(\mathbf{U})$ is $k \times k$, $\mathrm{Var}(\mathbf{AU})$ is $m \times m$.

3. Let $\mathbf{c}_{k\times 1}$ denote a constant vector. Then

$$\mathrm{Var}(\mathbf{U} + \mathbf{c}) = \mathrm{Var}(\mathbf{U}).$$

$$\mathrm{Var}(\mathbf{c}) = (0)_{k\times k}.$$

4. The variance of any random vector is *nonnegative definite symmetric matrix.*

To see this, note that if $\mathbf{a}_{k\times 1}$ is constant we can apply $\mathrm{Var}(\mathbf{AU}) = \mathbf{A}\ \mathrm{Var}(\mathbf{U})\mathbf{A}^T$ to obtain

$$\begin{aligned}\mathrm{Var}(\mathbf{a}^T\mathbf{U}) &= \mathrm{Var}(\Sigma_{j=1}^k a_j U_j) \\ &= \mathbf{a}^T \mathrm{Var}(\mathbf{U})\mathbf{a} = \Sigma_{i,j} a_i a_j \mathrm{Cov}(U_i, U_j).\end{aligned}$$

Because the variance of any random variable is nonnegative and a is arbitrary, we conclude from the above equalities that $\mathrm{Var}(\mathbf{U})$ is a *nonnegative definite symmetric matrix.*

Definition 3 *Define the moment generating function (m.g.f.) of* $\mathbf{U}_{k\times 1}$ *for* $\mathbf{t} \in R^k$ *by*

$$M(\mathbf{t}) = M_{\mathbf{U}}(\mathbf{t}) = E(e^{\mathbf{t}^T\mathbf{U}}) = E(e^{\Sigma_{j=1}^k t_j U_j}).$$

5. If $\mathbf{U}_{k\times 1}, \mathbf{V}_{k\times 1}$ are independent then

$$M_{\mathbf{U}+\mathbf{V}}(\mathbf{t}) = M_{\mathbf{U}}(\mathbf{t})M_{\mathbf{V}}(\mathbf{t}).$$

2 The Bivariate Normal Distribution

The family of k-variate normal distributions arises on theoretical grounds when we consider the limiting behavior of sums of independent k-vectors of random variables. In this section we focus on the case $k = 2$ where all properties can be derived relative easily.

A planar vector (X, Y) has a *bivariate normal distribution* if, and only if, there exist constants $a_{ij}, 1 \le i, j \le 2, \mu_1, \mu_2$, and independent standard normal random variables Z_1, Z_2 such that

$$X = \mu_1 + a_{11}Z_1 + a_{12}Z_2$$
$$Y = \mu_2 + a_{21}Z_1 + a_{22}Z_2.$$

In matrix notation, if $\mathbf{A} = (a_{ij}), \mu = (\mu_1, \mu_2)^T, \mathbf{X} = (X, Y)^T, \mathbf{Z} = (Z_1, Z_2)^T$, the definition is equivalent to

$$\mathbf{X} = \mathbf{AZ} + \mu. \tag{1}$$

Two important properties follow from the definition.

Proposition 4 *The marginal distributions of the components of a bivariate normal random vector are (univariate) normal or degenerate (concentrate on one point).*

Note that the converse of the proposition is not true (See problem B.4.10 in Bickel and Doksum [2001]). Also, note that

$$E(X) = \mu_1 + a_{11}E(Z_1) + a_{12}E(Z_2) = \mu_1,\ E(Y) = \mu_2$$

and define

$$\sigma_1 = \sqrt{\text{Var } X},\ \sigma_2 = \sqrt{\text{Var } Y}.$$

Then X has $\mathcal{N}(\mu_1, \sigma_1^2)$ and Y a $\mathcal{N}(\mu_2, \sigma_2^2)$ distribution.

Proposition 5 *If we apply an affine transformation* $\mathbf{g}(x) = \mathbf{C}x + \mathbf{d}$ *to a vector* $\mathbf{X}$*, which has a bivariate normal distribution, then* $\mathbf{g}(\mathbf{X})$ *also has such a distribution.*

This is clear because

$$\mathbf{CX} + \mathbf{d} = \mathbf{C}(\mathbf{AZ} + \mu) + \mathbf{d} = (\mathbf{CA})\mathbf{Z} + (\mathbf{C}\mu + \mathbf{d}). \tag{2}$$

We define the *variance-covariance matrix* of (X, Y) (or of the distribution of (X, Y)) as the matrix of central second moments

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \tag{3}$$

where

$$\rho = \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_1\sigma_2}.$$

This symmetric matrix is in many ways the right generalization of the variance to two dimensions.

Theorem 6 *Suppose that* $\sigma_1\sigma_2 \neq 0$ *and* $|\rho| < 1$*. Then the density of* $\mathbf{X}$ *is*

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi\sqrt{\det \mathbf{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)\right). \tag{4}$$

Remark 7 *From (3) we see that* $\mathbf{\Sigma}$ *is nonsigular iff* $\sigma_1\sigma_2 \neq 0$ *and* $|\rho| < 1$*. Bivariate normal distribution with* $\sigma_1\sigma_2 \neq 0$ *and* $|\rho| < 1$ *are referred to as nondegenerate, whereas others are degenerate.*

Remark 8 *When* $\rho = 0$, $p_{\mathbf{X}}(\mathbf{x})$ *becomes the joint density of two independent normal variables. Thus, in the bivariate normal case, correlation zero is equivalent to independence.*

Exercise 9 *Given nonnegative constants* σ_1, σ_2, *a number* ρ *such that* $|\rho| < 1$ *and numbers* μ_1, μ_2, *construct a random vector* $(X, Y)^T$, *where*

$$X = \mu_1 + \sigma_1 Z_1, \; Y = \mu_2 + \sigma_2(\rho Z_1 + \sqrt{1-\rho^2} Z_2)$$

Check that $(X, Y)^T$ has a bivariate normal distribution with vector of means $(\mu_1, \mu_2)^T$ and variance-covariance matrix $\mathbf{\Sigma}$ as given in (3).

3 Convergence in Probability v.s. in Distribution

3.0.1 Chebyshev's inequality

If X is any random variable and a is a constant, then

$$P(|X| \geq a) \leq \frac{E(X^2)}{a^2}.$$

3.1 Converagence in Probability

Definition 10 *If a sequence of random variables,* $\{Z_n\}$, *is such that* $P(|Z_n - \alpha| > \varepsilon)$ *approaches zero as* n *approaches infinity, for any* $\varepsilon > 0$ *and where* α *is some scalar, then* Z_n *is said to converge in probability to* α.

Definition 11 *A sequence of random vectors* $\mathbf{Z}_n \equiv (Z_{n1}, Z_{n2}, ..., Z_{nd})^T$ *converages in probability to* $\mathbf{Z} \equiv (Z_1, Z_2, ..., Z_d)^T$ *iff*

$$|\mathbf{Z}_n - \mathbf{Z}| \xrightarrow{\mathcal{P}} 0$$

or equivalently $Z_{nj} \xrightarrow{\mathcal{P}} Z_j$ *for* $1 \leq j \leq d$.

3.1.1 A Law of Large Numbers

Theorem 12 ***Example 13*** *Let* $X_1, X_2, \cdots, X_i, \cdots$ *be a sequence of independent random variables with* $E(X_i) = \mu$ *and* $Var(X_i) = \sigma^2$. *Let* $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. *Then, for any* $\varepsilon > 0$,

$$P(|\bar{X}_n - \mu| > \varepsilon) \to 0 \quad \text{as } n \to \infty$$

Proof. We first find $E(\bar{X}_n)$ and $Var(\bar{X}_n)$:

$$E(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^n E(X_i) = \mu$$

Since the X_i are independent,

$$Var(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{\sigma^2}{n}$$

The desired result now follows immediately from Chebyshev's inequality, which states that

$$P(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{Var(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0, \quad \text{as } n \to \infty$$

■

3.2 Convergence in Distribution

Definition 14 *Let $X_1, X_2, \cdots$ be a sequence of random variables with cumulative distribution functions $F_1, F_2, \cdots$, and let X be a random variable with distribution function F. We say that X_n converges in distribution to X if*

$$\lim_{n\to\infty} F_n(x) = F(x)$$

at every point at which F is continuous.

Definition 15 *A sequence $\{\mathbf{Z}_n\}$ of random vvectors converges in law (in distribution) to $\mathbf{Z}$, written $\mathbf{Z}_n \longrightarrow \mathbf{Z}$, iff*

$$h(\mathbf{Z}_n) \xrightarrow{\mathcal{L}} h(\mathbf{Z})$$

for all functions $h : \mathbf{R}^d \longrightarrow \mathbf{R}$, h continuous.

3.2.1 Central Limit Theorem

Theorem 16 ***The Multivariate Central Limit Theorem****. Let $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ be independent and identically distributed random k vectors with $E|X_1|^2 < \infty$. Let $E(\mathbf{X}_1) = \mu$, $Var(\mathbf{X}_1) = \mathbf{\Sigma}$, and let $\mathbf{S}_n = \Sigma_{i=1}^n \mathbf{X}_i$. Then, for every continuous functiong: $g : R^k \to R$,*

$$g\left(\frac{\mathbf{S}_n - n\mu}{\sqrt{n}}\right) \xrightarrow{\mathcal{L}} g(\mathbf{Z})$$

where $\mathbf{Z} \sim \mathcal{N}_k(\mathbf{0}, \mathbf{\Sigma})$.

3.2.2 The $O_P, \asymp p$, and o_P Notation

The following asymptotic order in probability notation is useful.

$$\mathbf{U}_n = o_P(1) \text{ iff } \mathbf{U}_n \xrightarrow{P} 0$$
$$\mathbf{U}_n = O_P(1) \text{ iff } \forall\epsilon > 0,\ \exists M < \infty \text{ such that } \forall n \ \ P\left[|\mathbf{U}_n| \geq M\right] \leq \epsilon$$
$$\mathbf{U}_n = o_P(\mathbf{V}_n) \text{ iff } \frac{|\mathbf{U}_n|}{|\mathbf{V}_n|} = o_P(1)$$
$$\mathbf{U}_n = O_P(\mathbf{V}_n) \text{ iff } \frac{|\mathbf{U}_n|}{|\mathbf{V}_n|} = O_P(1)$$
$$\mathbf{U}_n \asymp p\ \mathbf{V}_n \text{ iff } \mathbf{U}_n = O_P(\mathbf{V}_n) \text{ and } \mathbf{V}_n = O_P(\mathbf{U}_n).$$

Note that

$$O_P(1)o_P(1) = o_P(1),\ O_P(1) + o_P(1) = O_P(1),$$

and $\mathbf{U}_n \xrightarrow{\mathcal{L}} \mathbf{U} \implies \mathbf{U}_n = O_P(1)$.

Example 17 *Suppose $\mathbf{Z}_1, \mathbf{Z}_2, \cdots, \mathbf{Z}_n$ are iid as $\mathbf{Z}_1$ with $E|\mathbf{Z}_1|^2 < \infty$. Set $\mu = E|\mathbf{Z}_1|$, then $\overline{\mathbf{Z}}_n = \mu + O_p(n^{-\frac{1}{2}})$ by the central limit theorem.*

Exercise 18 *Let X_i be the last digit of D_i^2, where D_i is a random digit between 0 and 9. For instance, if $D_i = 7$ then $D_i^2 = 49$ and $X_i = 9$. Let $\bar{X}_n = (X_1 + \cdots + X_n)/n$ be the average of a large number n of such last digits, obtained from independent random digits $D_1, \cdots, D_n$.*

a) Predict the value of $\bar{X}_n$ for large n.

b) Find a number ϵ such that for $n = 10,000$ the chance that your prediction is off by more than ϵ is about 1 in 200.

c) Find approximately the least value of n such that your prediction of $\bar{X}_n$ is correct to within 0.01 with probability at least 0.99.

d) Which can be predicted more accurately for large n: the value of $\bar{X}_n$, or the value of $\bar{D}_n = (D_1 + \cdots + D_n)/n$?

e) If you just had to predict the first digit of $\bar{X}_{100}$, what digit should you choose to maximize your chance of being correct, and what is that chance?

References

[1] Peter J. Bickel and Kjell A. Doksum (2001) Mathematical Statistics: Basic Ideas and Selected Topics, Vol. I, 2nd Edition. *Prentice Hall*

[2] Geoffrey Grimmett and David Stirzaker (2002) Probability and Random Processes, 3rd Edition, *Oxford University Press.*

[3] Jim Pitman (1993) Probability, *Springer-Verlag New York, Inc.*

[4] John A. Rice (1995) Mathematical Statistics and Data Analysis, 2nd Edition, *Duxbury Press.*