Version: September 3, 2010

Notes on the Multivariate Normal and Related Topics

Let me refresh your memory about the distinctions between *population* and *sample*; *parameters* and *statistics*; *population distributions* and *sampling distributions*. One might say that anyone worth knowing, knows these distinctions. We start with the simple univariate framework and then move on to the multivariate context.

A) Begin by positing a *population* of interest that is operationalized by some random variable, say X. In this Theory World framework, X is characterized by *parameters*, such as the expectation of X, $\mu = \mathrm{E}(X)$, or its variance, $\sigma^2 = \mathrm{V}(X)$. The random variable X has a *(population) distribution*, which for us will typically be assumed normal.

B) A *sample* is generated by taking observations on X, say, $X_1, \ldots, X_n$, considered independent and identically distributed as X, i.e., they are exact copies of X. In this Data World context, statistics are functions of the sample, and therefore, characterize the sample: the sample mean, $\hat{\mu} = \frac{1}{n}\Sigma_{i=1}^{n} X_i$; the sample variance, $\hat{\sigma}^2 = \frac{1}{n}\Sigma_{i=1}^{n}(X_i - \hat{\mu})^2$, with some possible variation in dividing by $n-1$ to generate an unbiased estimator of σ^2. The statistics, $\hat{\mu}$ and $\hat{\sigma}^2$, are *point estimators* of μ and σ^2. They are random variables by themselves, so they have distributions that are called *sampling distributions*.

The general problem of statistical inference is to ask what the sample statistics, $\hat{\mu}$ and $\hat{\sigma}^2$, tell us about their population counterparts,

such as μ and σ^2. Can we obtain some notion of accuracy from the sampling distribution, e.g., confidence intervals?

The multivariate problem is generally the same but with more notation:

A) The population (Theory World) is characterized by a collection of p random variables, $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$, with parameters: $\mu_i = \mathrm{E}(X_i)$; $\sigma_i^2 = \mathrm{V}(X_i)$; $\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathrm{E}[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))]$; or the correlation between X_i and X_j, $\rho_{ij} = \sigma_{ij}/\sigma_i\sigma_j$.

B) The sample (Data World) is defined by n independent observations on the random vector $\mathbf{X}$, with the observations placed into an $n \times p$ data matrix (e.g., subject by variable) that we also (with a slight abuse of notation) denote by a bold-face capital letter, $\mathbf{X}_{n\times p}$:

$$\mathbf{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} .$$

The statistics corresponding to the parameters of the population are: $\hat{\mu}_i = \frac{1}{n}\Sigma_{k=1}^n x_{ki}$; $\hat{\sigma}_i{}^2 = \frac{1}{n}\Sigma_{k=1}^n (x_{ki} - \hat{\mu}_i)^2$, with again some possible variation to a division by $n-1$ for an unbiased estimate; $\hat{\sigma}_{ij} = \frac{1}{n}\Sigma_{k=1}^n (x_{ki} - \hat{\mu}_i)(x_{kj} - \hat{\mu}_j)$; and $\hat{\rho}_{ij} = \hat{\sigma}_{ij}/\hat{\sigma}_i\hat{\sigma}_j$.

To obtain a good sense of what the estimators tell us about the population parameters, we will have to make some assumption about the population, e.g., $[X_1, \ldots, X_p]$ has a multivariate normal distribution. As we will see, this assumption leads to some very nice results.

0.1 Developing the Multivariate Normal Distribution

Suppose $X_1, \ldots, X_p$ are p continuous random variables with density functions, $f_1(x_1), \ldots, f_p(x_p)$, and distribution functions, $F_1(x_1), \ldots, F_p(x_p)$, where

$$P(X_i \leq x_i) = F_i(x_i) = \int_{-\infty}^{x_i} f_i(x_i)dx \ .$$

We define a p-dimensional random variable (or random vector) as the vector, $\mathbf{X}' = [X_1, \ldots, X_p]$; $\mathbf{X}$ has the joint cumulative distribution function

$$F(x_1, \ldots, x_p) = \int_{-\infty}^{x_p} \cdots \int_{-\infty}^{x_1} f(x_1, \ldots, x_p)dx_1 \cdots dx_p \ .$$

If the random variables, $X_1, \ldots, X_p$ are independent, then the joint density and cumulative distribution functions factor: $f(x_1, \ldots, x_p) = f_1(x_1) \cdots f_p(x_p)$ and $F(x_1, \ldots, x_p) = F_1(x_1) \cdots F_p(x_p)$. The independence property will not be assumed in the following; in fact, the whole idea is to investigate what type of dependency exists in the random vector $\mathbf{X}$.

When $\mathbf{X}' = [X_1, X_2, \ldots, X_p] \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the joint density function has the form

$$\phi(x_1, \ldots, x_p) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\} \ .$$

In the univariate case, the density provides the usual bell-shaped curve: $\phi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2}[(x - \mu)/\sigma)^2]\}$.

Given the MVN assumption on the generation of the data matrix,

$\mathbf{X}_{n\times p}$, we know the sampling distribution of the vector of means:

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_p \end{pmatrix} \sim \text{MVN}(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$$

In the univariate case, we have that $\hat{\mu} \sim \text{N}(\mu, (1/n)\sigma^2)$. As might be expected, a Central Limit Theorem (CLT) states that these results also hold asymptotically when the MVN assumption is relaxed. Also, if we estimate the variance-covariance matrix, $\boldsymbol{\Sigma}$, with divisions by $n-1$ rather than n, and denote the result as $\hat{\boldsymbol{\Sigma}}$, then $\text{E}(\hat{\boldsymbol{\Sigma}}) = \boldsymbol{\Sigma}$.

Linear combinations of random variables form the backbone of all of multivariate statistics. For example, suppose $\mathbf{X}' = [X_1, X_2, \ldots, X_p] \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and consider the following q linear combinations:

$$\begin{aligned} Z_1 &= c_{11}X_1 + \cdots + c_{1p}X_p \\ &\vdots \\ Z_q &= c_{q1}X_1 + \cdots + c_{qp}X_p \end{aligned}$$

Then the vector $\mathbf{Z}' = [Z_1, Z_2, \ldots, Z_q] \sim \text{MVN}(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}')$, where

$$\mathbf{C}_{q\times p} = \begin{pmatrix} c_{11} & \cdots & c_{1p} \\ \vdots & & \vdots \\ c_{q1} & \cdots & c_{qp} \end{pmatrix} .$$

These same results hold in the sample if we observe the q linear combinations, $[Z_1, Z_2, \ldots, Z_q]$: i.e., the sample mean vector is $\mathbf{C}\hat{\boldsymbol{\mu}}$ and the sample variance-covariance matrix is $\mathbf{C}\hat{\boldsymbol{\Sigma}}\mathbf{C}'$.

0.2 Multiple Regression and Partial Correlation (in the Population)

Suppose we partition our vector of p random variables as follows:

$$\mathbf{X}' = [X_1, X_2, \ldots, X_p] =$$

$$[[X_1, X_2, \ldots, X_k], [X_{k+1}, X_{k+2} \ldots, X_p]] \equiv [\mathbf{X}_1', \mathbf{X}_2']$$

Partitioning the mean vector and variance-covariance matrix in the same way,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \; ; \; \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{12}'} & \boldsymbol{\Sigma_{22}} \end{pmatrix},$$

we have

$$\mathbf{X}_1' \sim \text{MVN}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_{11}}) \; ; \; \mathbf{X}_2' \sim \text{MVN}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_{22}}) \; .$$

$\mathbf{X}_1'$ and $\mathbf{X}_2'$ are statistically independent vectors if and only if $\boldsymbol{\Sigma_{12}} = \mathbf{0_{k \times p}}$, the $k \times p$ zero matrix.

What I call the Master Theorem refers to the conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$; it is multivariate normal with mean vector of order $k \times 1$:

$$\boldsymbol{\mu_1} + \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_{22}^{-1}} (\mathbf{x}_2 - \boldsymbol{\mu_2}) \; ,$$

and variance-covariance matrix of order $k \times k$:

$$\boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}} \boldsymbol{\Sigma_{22}^{-1}} \boldsymbol{\Sigma_{12}'} \; .$$

If we denote the $(i,j)^{th}$ partial covariance element in this latter matrix ('holding' all variables in the second set 'constant') as $\sigma_{ij \cdot (k+1) \cdots p}$, then the partial correlation of variables i and j, 'holding' all variables in the second set 'constant', is

$$\rho_{ij \cdot (k+1) \cdots p} = \sigma_{ij \cdot (k+1) \cdots p} \Big/ \sqrt{\sigma_{ii \cdot (k+1) \cdots p} \sigma_{jj \cdot (k+1) \cdots p}} \; .$$

Notice that in the formulas just given, the inverse of $\boldsymbol{\Sigma_{22}}$ must exist; otherwise, we have what is called "multicollinearity," and solutions don't exist (or are very unstable numerically if the inverse 'almost' doesn't exist). So, as one moral: don't use both total and subtest scores in the same set of "independent" variables.

If the set $\mathbf{X}_1$ contains just one random variable, X_1, then the mean vector of the conditional distribution can be given as

$$\mathrm{E}(X_1) + \boldsymbol{\sigma}'_{\mathbf{12}}\boldsymbol{\Sigma}^{-\mathbf{1}}_{\mathbf{22}}(\mathbf{x}_2 - \boldsymbol{\mu_2}) \ ,$$

where $\boldsymbol{\sigma}'_{\mathbf{12}}$ is $1\times(p-1)$ and of the form $[\mathrm{Cov}(X_1, X_2), \ldots, \mathrm{Cov}(X_1, X_p)]$. This is nothing but our old regression equation written out in matrix notation. If we let $\boldsymbol{\beta}' = \boldsymbol{\sigma}'_{\mathbf{12}}\boldsymbol{\Sigma}^{-\mathbf{1}}_{\mathbf{22}}$, then the conditional mean vector (predicted X_1) is equal to $\mathrm{E}(X_1) + \boldsymbol{\beta}'(\mathbf{x}_2 - \boldsymbol{\mu_2}) = \mathrm{E}(X_1) + \beta_1(x_2 - \mu_2) + \cdots + \beta_{p-1}(x_p - \mu_p)$. The covariance matrix, $\boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma}^{-\mathbf{1}}_{\mathbf{22}}\boldsymbol{\Sigma}'_{\mathbf{12}}$, takes on the form $\mathrm{Var}(X_1) - \boldsymbol{\sigma_{12}}\boldsymbol{\Sigma}^{-\mathbf{1}}_{\mathbf{22}}\boldsymbol{\sigma}'_{\mathbf{12}}$, i.e., the variation in X_1 that is *not explained*. If we take explained variation, $\boldsymbol{\sigma_{12}}\boldsymbol{\Sigma}^{-\mathbf{1}}_{\mathbf{22}}\boldsymbol{\sigma}'_{\mathbf{12}}$, and consider the proportion to the total variance, $\mathrm{Var}(X_1)$, we define the squared multiple correlation coefficient:

$$\rho^2_{1\cdot 2\cdots p} = \frac{\boldsymbol{\sigma_{12}}\boldsymbol{\Sigma}^{-\mathbf{1}}_{\mathbf{22}}\boldsymbol{\sigma}'_{\mathbf{12}}}{\mathrm{Var}(X_1)} \ .$$

In fact, the linear combination, $\boldsymbol{\beta}'\mathbf{X}_2$, has the highest correlation of any linear combination with X_1; and this correlation is the positive square root of the squared multiple correlation coefficient.

0.3 Moving to the Sample

Up to this point, the concepts of regression and kindred ideas, have been discussed only in terms of population parameters. We now have

the task of obtaining estimates of these various quantities based on a sample; also, associated inference procedures need to be developed. Generally, we will rely on maximum-likelihood estimation and the related likelihood-ratio tests.

To define how maximum-likelihood estimation proceeds, we will first give a series of general steps, and then operationalize for a univariate normal distribution example.

A) Let $X_1, \ldots, X_n$ be univariate observations on X (independent and continuous, and depending on some parameters, $\theta_1, \ldots, \theta_k$. The density function of X_i is denoted as $f(x_i; \theta_1, \ldots, \theta_k)$.

B) The likelihood of observing $x_1, \ldots, x_n$ for values of $X_1, \ldots, X_n$ is the joint density of the n observations, and because the observations are independent,

$$\mathrm{L}(\theta_1, \ldots, \theta_k) \equiv \prod_{i=1}^{n} f(x_i; \theta_1, \ldots, \theta_k) \ ,$$

i.e., we assume $x_1, \ldots, x_n$ are already observed, and that L is a function of the parameters only.

Now, choose parameter values such that $\mathrm{L}(\theta_1, \ldots, \theta_k)$ is at a maximum — i.e., the probability of observing that particular sample is maximized by the choice of $\theta_1, \ldots, \theta_k$. Generally, it is easier and equivalent to maximize the log-likelihood using $\ell(\theta_1, \ldots, \theta_k) \equiv \log(\mathrm{L}(\theta_1, \ldots, \theta_k))$.

C) Generally, the maximum values are found through differentiation, and by setting the partial derivatives equal to zero. Explicitly,

we find $\theta_1, \ldots, \theta_k$ such that

$$\frac{\partial}{\partial \theta_j} \ell(\theta_1, \ldots, \theta_k) = 0 \ ,$$

for $j = 1, \ldots, k$. (We should probably check second derivatives to see if we have a maximum, but this is almost always true, anyways.)

Example:

Suppose $X_i \sim \mathrm{N}(\mu, \sigma^2)$, with density

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_i - \mu)^2/2\sigma^2\} \ .$$

The likelihood has the form

$$\mathrm{L}(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_i - \mu)^2/2\sigma^2\} =$$

$$(1/\sqrt{2\pi})^n (1/\sigma^2)^{n/2} \exp\{-(1/2\sigma^2) \sum_{i=1}^{n} (x_i - \mu)^2\} \ .$$

The log-likelihood reduces:

$$\ell(\mu, \sigma^2) = \log \mathrm{L}(\mu, \sigma^2) =$$

$$\log(1/\sqrt{2\pi})^n + \log(1/\sigma^2)^{n/2} + (-(1/2\sigma^2) \sum_{i=1}^{n} (x_i - \mu)^2) =$$

$$\text{constant} - (n/2) \log \sigma^2 - (1/2\sigma^2) \sum_{i=1}^{n} (x_i - \mu)^2 \ .$$

The partial derivatives have the form:

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = -(1/2\sigma^2) \sum_{i=1}^{n} 2(x_i - \mu)(-1) \ ,$$

and

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = -(n/2)(1/\sigma^2) - (1/2(\sigma^2)^2)(-1) \sum_{i=1}^{n} (x_i - \mu)^2 \ .$$

Setting these two expressions to zero, gives

$$\hat{\mu} = \sum_{i=1}^{n} x_i/n \ ; \ \hat{\sigma}^2 = (1/n) \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \ .$$

Maximum likelihood (ML) estimates are generally consistent, asymptotically normal (for large n), and efficient; a function of the sufficient statistics; and have an invariance to operations performed on them – e.g., the ML estimate for σ is just $\sqrt{\hat{\sigma}^2}$. As the ML estimator for $\hat{\sigma}^2$ shows, ML estimators are not necessarily unbiased.

Another Example:

Suppose $X_1, \ldots, X_n$ are observations on the Poisson discrete random variable X (having outcomes $0, 1, \ldots$):

$$P(X_i = x_i) = \frac{\exp(-\lambda)\lambda^{x_i}}{x_i!} \ .$$

The likelihood is

$$\mathrm{L}(\lambda) = \prod_{i=1}^{n} P(X_i = x_i) = \frac{\exp(-n\lambda)\lambda^{\sum_i x_i}}{\Pi_i \, x_i!} \ ,$$

and the log-likelihood

$$\ell(\lambda) = \log \mathrm{L}(\lambda) = -n\lambda + \log(\lambda^{\sum_i x_i}) - \log \prod_i x_i! \ .$$

The partial derivative

$$\frac{\partial \ell(\lambda)}{\partial \lambda} = -n + (1/\lambda) \sum_{i=1}^{n} x_i \ ,$$

and when set to zero, gives

$$\hat{\lambda} = \sum_{i=1}^{n} x_i/n \ .$$

0.3.1 Likelihood Ratio Tests

Besides using the likelihood concept to find good point estimates, the likelihood can also be used to find good tests of hypotheses. We will develop this idea in terms of a simple example: Suppose $X_1 = x_1, \ldots, X_n = x_n$ denote values obtained on n independent observations on a $\mathrm{N}(\mu, \sigma^2)$ random variable, and where σ^2 is assumed known. The standard test of $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$, would compare $(\bar{x}. - \mu_0)^2/(\sigma^2/n)$ to a χ^2 random variable with one-degree of freedom. Or, if we chose the significance level to be .05, we could reject H_0 if $|\bar{x}. - \mu_0|/(\sigma/\sqrt{n}) \geq 1.96$. We now develop this same test using likelihood ideas:

The likelihood of observing $x_1, \ldots, x_n$, $\mathrm{L}(\mu)$, is only a function of μ (because σ^2 is assumed to be known):

$$\mathrm{L}(\mu) = (1/\sigma\sqrt{2\pi})^n \exp\{-(1/2\sigma^2) \sum_{i=1}^{n} (x_i - \mu)^2\} \ .$$

Under H_0, $\mathrm{L}(\mu)$ is at a maximum for $\mu = \mu_0$; under H_1, $\mathrm{L}(\mu)$ is at a maximum for $\mu = \hat{\mu}$, the maximum likelihood estimator. If H_0 is true, $\mathrm{L}(\hat{\mu})$ and $\mathrm{L}(\mu_0)$ should be close to each other; If H_0 is false, $\mathrm{L}(\hat{\mu})$ should be much larger than $\mathrm{L}(\mu_0)$. Thus, the decision rule would be to reject H_0 if $\mathrm{L}(\mu_0)/\mathrm{L}(\hat{\mu}) \leq \lambda$, where λ is some number less than 1.00 and chosen to obtain a particular α level. The ratio, $(\mathrm{L}(\mu_0)/\mathrm{L}(\hat{\mu})) = \exp\{-(1/2)(\hat{\mu} - \mu_0)^2\}/(\sigma^2/n)$. Thus, we could rephrase the decision rule: reject H_0 if $(\hat{\mu} - \mu_0)^2\}/(\sigma^2/n) \geq$

$-2\log(\lambda)$, or if $|\bar{x}. - \mu_0|/(\sigma/\sqrt{n}) \geq \sqrt{-2\log\lambda}$. Thus, for an .05 level of significance, choose λ so $1.96 = \sqrt{-2\log\lambda}$. Generally, we can phrase likelihood ratio tests as:

$$-2\log(\text{likelihood ratio}) \sim \chi^2_{\nu-\nu_0} \ ,$$

where ν is the dimension of the parameter space generally, and ν_0 is the dimension under H_0.

0.3.2 Estimation

To obtain estimates of the various quantities we need, merely replace the variances and covariances by their maximum likelihood estimates. This process generates sample partial correlations or covariances; sample multiple squared correlations; sample regression parameters; and so on.