EIGHT

RESPONSE MODEL METHODS

1. UNBIASEDNESS: BY TAKING AVERAGES

The theory of estimation is concerned with finding a function on the sample space whose values will cluster in some reasonable manner about the value of a parameter of interest. In an application there would then be substantial hope that the calculated value of such a function was a good approximation to the true value of the parameter.

The very wealth of functions on a sample space leads to major problems in choosing a function that is in some sense best. The concept that has been most fruitful theoretically and for applications is that of unbiasedness - to be examined in the first three sections of this chapter. There are other approaches such as the decision theoretic which provide extensions and alternatives but none with the immediate relevance and fruitfulness of the unbiasedness approach.

For notation let y be the response with values in a sample space $\mathcal{S}$ and θ be the parameter with values in a parameter space Ω. And let $\{P_\theta : \theta \in \Omega\}$ be the class of measures for the response y.

Now suppose we are interested in some particular real valued parameter $\beta(\theta)$ and $t(y)$ is a real valued function that we are considering as an estimator for $\beta(\theta)$. Then we define

$\mathbb{D}$ t *is an unbiased estimator for* β *if*

$$E\{t(y)|\theta\} = \beta(\theta) \quad \textit{for all} \quad \theta \quad \textit{in} \quad \Omega \, .$$

Note that t is a function from S into $\mathbb{R}$ and that β is a function from Ω into $\mathbb{R}$. This definition extends easily to cover a vector estimator $\underset{\sim}{t}(y) = (t_1(y), \ldots, t_r(y))'$ of a vector parameter $\underset{\sim}{\beta}(\theta) = (\beta_1(\theta), \ldots, \beta_r(\theta))'$:

$\mathbb{D}$ $\underset{\sim}{t}$ *is an unbiased estimator for* $\underset{\sim}{\beta}$ *if*

$$E\{t_s(y)|\theta\} = \beta_s(\theta) \quad \textit{for all} \quad \theta \quad \textit{in} \quad \Omega \quad \textit{and for} \quad s = 1,\ldots,r \, .$$

Example 1. Let $(y_1, \ldots, y_n)$ be a sample from the normal distribution (μ, σ^2) with $\theta = (\mu, \sigma^2)$ in $\Omega = \mathbb{R} \times \mathbb{R}^+$. For the parameter μ (first coordinate projection) we can consider, say, the sample average $\bar{y}$ and the sample median $\tilde{y}$:

$$E(\bar{y}|\theta) = \mu \quad , \quad E(\tilde{y}|\theta) = \mu \, .$$

(We have not discussed the distribution of the sample median but we can argue generally that the distribution of $\tilde{y} - \mu$ is symmetric about zero, and then $E\tilde{y} = \mu$) . For the parameter $\theta = (\mu, \sigma^2)$ we can consider say the sample average and sample variance $(\bar{y}, s_y^2)$:

$$E\left[(\bar{y}, s_y^2) : \theta\right] = (\mu, \sigma^2) .$$

Of course we can construct many other estimators for these parameters and our concern in the first three sections of this chapter will center largely on finding unbiased estimators that are best in some sense. To the degree however that we accept our model, we will limit our attention to estimators based on the likelihood statistic which is $(\bar{y}, s_y^2)$. We need consider then only those functions that are effectively defined on the space $\mathbb{R} \times \mathbb{R}^+$ of $(\bar{y}, s_y^2)$.

(a) Comparison of estimators.

Consider further the example involving a sample $(y_1, \ldots, y_n)$ from the normal (μ, σ^2) . The unbiased estimator $\bar{y}$ has variance σ^2/n and normal distribution form. The unbiased estimator $\tilde{y}$ can be shown to have variance $\pi\sigma^2/2n$ (approximately) and normal distribution form (approximately). It is reasonable then to compare the estimators on the basis of variance and to prefer $\bar{y}$ which has smaller variance. Indeed reciprocal variance provides a somewhat reasonable indicator of the precision of an unbiased estimator. And we can then numerically compare estimators of the basis of reciprocal variance. This gives one definition for the comparitive efficiency of two estimators:

D *The efficiency of an unbiased estimator* t_1 (of $\beta(\theta)$) *with respect to an unbiased estimator* t_2 (of $\beta(\theta)$) *is*

$$\mathrm{Eff}(t_1, t_2) = \frac{1/\mathrm{Var}(t_1)}{1/\mathrm{Var}(t_2)} .$$

For the sample average and sample median (normal case) we have

$$\mathrm{Eff}(\tilde{x}, \bar{x}) = \frac{2n/\pi\sigma^2}{n/\sigma^2} = \frac{2}{\pi} = \cdot 64 .$$

The definition of efficiency has an interesting interpretation for the common cases where variance is proportion to reciprocal sample size. For example, if $\tilde{y}$ is based on a sample of 100 and $\bar{y}$ is based on a sample of 64 ; then

$$\mathrm{Var}(\tilde{y}) = \frac{\pi}{2}\,\frac{\sigma^2}{100} = \frac{\sigma^2}{64} = \mathrm{Var}(\bar{y}) ;$$

and we can say that $\tilde{y}$ uses 64% of a sample in comparison with $\bar{y}$.

For the case of vector unbiased estimators the comparison by means of variance becomes more complicated. Each coordinate has a variance and in addition there are covariances. Consider two unbiased estimators $\underset{\sim}{t}_1$ and $\underset{\sim}{t}_2$ of a parameter $\underset{\sim}{\beta}$; let $\Sigma_{11}(\theta)$ and $\Sigma_{22}(\theta)$ be the variance matrices of the two estimators. Then we will say that $\underset{\sim}{t}_1$ *is better than* $\underset{\sim}{t}_2$ *if*

$$\Sigma_{22}(\theta) - \Sigma_{11}(\theta)$$

is positive semi definite; and we then say that $\Sigma_{11}(\theta)$ *is smaller than* $\Sigma_{22}(\theta)$ writing $\Sigma_{11}(\theta) \prec \Sigma_{22}(\theta)$. Note that a symmetric matrix M is *positive semi definite* if *the quadratic expression*

$$\underset{\sim}{\ell}'M\underset{\sim}{\ell} \geq 0 \qquad \text{all } \underset{\sim}{\ell} \text{ in } \mathbb{R}^r ,$$

or if M is *the covariance matrix*

$$M = \mathrm{VAR}(\underset{\sim}{w})$$

of a distribution say indicated by $\underset{\sim}{w}$ or if M is *the inner product matrix*

$$M = BB'$$

of the row vectors in an $r \times r$ *matrix* B . These equivalences are available from matrix theory. We can then say that $\underset{\sim}{t}$ *is better than* $\underset{\sim}{t}_2$ *if the ellipsoid of concentration of* $\underset{\sim}{t}_1$ *is contained in the ellipsoid of concentration of* $\underset{\sim}{t}_1$ (ellipsoid of concentration defined by Four: Problems 56, 57.) or if

$$\Sigma_{11}^{-1}(\theta) - \Sigma_{22}^{-1}(\theta)$$

is positive semi definite. Note that we obtain a partial ordering $\prec$ on the positive semi definite (symmetric) matrices; the reverse ordering holds on the inverse matrices.

(b) Combining estimates by weighted averages

Suppose we have two unbiased estimates t_1 , t_2 of $\beta(\theta)$ and wish to consider linear combinations as possible estimates of $\beta(\theta)$

Theorem 1. *If* t_1 , t_2 *are unbiased for* $\beta(\theta)$, *then*

(i) *Any linear combinations of* t_1 *and* t_2 *that is unbiased for* $\beta(\theta)$ *has the form* $at_1 + (1-a)t_2$ *(the sum of the weights is* 1*)* .

(ii) *Among such linear unbiased estimate the one with smallest variance at the parameter value* θ_0 *has*

$$\frac{a}{1-a} = \frac{1\Big/\left[\mathrm{Var}(t_1|\theta_0) - \mathrm{cov}(t_1,\ t_2|\theta_0)\right]}{1\Big/\left[\mathrm{Var}(t_2|\theta_0) - \mathrm{cov}(t_1,\ t_2|\theta_0)\right]}$$

(Weight by reciprocal excess of variance over covariance).

Proof: (i) Consider a linear combination $at_1 + bt_2$. If this is unbiased for $\beta(\theta)$ then

$$E\left(at_1(y) + bt_2(y)\,|\,\theta\right)$$

$$= a\beta(\theta) + b\beta(\theta) = \beta(\theta)$$

and hence $b = 1 - a$, provided of course we exclude the ultra trivial $\beta(\theta) \equiv 0$.

(ii) The variance of $at_1 + (1-a)t_2$ at θ_0 is

$$a^2\mathrm{Var}\left(t_1|\theta_0\right) + 2a(1-a)\mathrm{Cov}\left(t_1,\ t_2|\theta_0\right) + (1-a)^2\mathrm{Var}\left(t_2|\theta_0\right) .$$

Setting the derivative with respect to a equal to zero gives

$$2a\left\{\mathrm{Var}\left(t_1|\theta_0\right) - \mathrm{Cov}\left(t_1,\ t_2|\theta_0\right)\right\}$$

$$= 2(1-a)\left\{\mathrm{Var}\left(t_2|\theta_0\right) - \mathrm{Cov}\left(t_1,\ t_2|\theta_0\right)\right\}$$

which is equivalent to the quoted condition. Another method of proof involves completing-the-square in a to form a quadratic expression plus a constant.

Note that with uncorrelated estimates, minimum variance is obtained by *weighting by reciprocal variance.*

Example 2. Suppose that t_1 and t_2 are unbiased estimates of θ and that

$$\operatorname{Var}(t_1) = \cdot 126\sigma^2 \ , \qquad \operatorname{Var}(t_2) = \cdot 276\sigma^2 \ ,$$

$$\operatorname{Cov}(t_1, t_2) = -\cdot 100\sigma^2 \ .$$

Then the best linear unbiased estimate is

$$\left(\frac{1}{\cdot 226} + \frac{1}{\cdot 376}\right)^{-1}\left(\frac{t_1}{\cdot 226} + \frac{t_2}{\cdot 376}\right) \ .$$

There is an analogous theorem for vector estimates. For simplicity here we give the special version for uncorrelated estimates:

Theorem 2. *If* $\underset{\sim}{t}_1$, $\underset{\sim}{t}_2$ *are uncorrelated unbiased estimates for* $\underset{\sim}{\beta}(\theta)$ *(with* $\beta_1(\theta)$*, ...,* $\beta_r(\theta)$ *linearly independent), then*

(i) *any linear combination of* $\underset{\sim}{t}_1$ *and* $\underset{\sim}{t}_2$ *that is unbiased for* $\underset{\sim}{\beta}$ *has the form*

$$A\underset{\sim}{t}_1 + (I-A)\underset{\sim}{t}_2$$

where A *is* r × r *and* I *is the* r × r *identity.*

(ii) *Among such linear unbiased estimates the estimate*

$$\left(\Sigma_{11}^{-1}(\theta_0) + \Sigma_{22}^{-1}(\theta_0)\right)^{-1}\left(\Sigma_{11}^{-1}(\theta_0)\underset{\sim}{t}_1 + \Sigma_{22}^{-1}(\theta_0)\underset{\sim}{t}_2\right)$$

has smallest covariance matrix at θ_0 .

Proof: (i). Consider a linear combination $A\underset{\sim}{t}_1 + B\underset{\sim}{t}_2$ where A and B are r × r ; unbiasedness gives

$$E\left(At_1 + Bt_2\right) = A\underset{\sim}{\beta} + B\underset{\sim}{\beta} = (A+B)\underset{\sim}{\beta}$$

and hence $A + B = I$.

(ii) The differentiation approach does not adapt easily to the vector case. The alternative approach by completing a quadratic expression is relatively straightforward but some manipulation is required to then obtain the quoted expressions; it is a special case ~~that for~~ of the gauss-markov theorem later in this section and can readily be omitted. The variance matrix of

$$A\underset{\sim}{t}_1 + (I-A)t_2 = (A,\ I - A)\begin{pmatrix}\underset{\sim}{t}_1\\ \underset{\sim}{t}_2\end{pmatrix}$$

is

$$\mathrm{VAR}\left(At_1 + (I - A)t_2\right)$$

$$= (A,\ I - A) \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} (A,\ I - A)'$$

$$= A\left(\Sigma_{11} + \Sigma_{22}\right)A' - A\Sigma_{22} - \Sigma_{22}A' + \Sigma_{22}$$

$$= \left(A - \cdots\ \right)\left(\Sigma_{11} + \Sigma_{22}\right)\left(A - \cdots\ \right)' + \cdots .$$

The cross terms $A\Sigma_{22}$, $\Sigma_{22}A'$ determine the term in the brackets and thus determine the remainder term:

$$\mathrm{VAR}\left(A\underset{\sim}{t}_1 + (I - A)\underset{\sim}{t}\right)$$

$$= \left[A - \Sigma_{22}\left(\Sigma_{11} + \Sigma_{22}\right)^{-1}\right]\left(\Sigma_{11} + \Sigma_{22}\right)\left[A - \Sigma_{22}\left(\Sigma_{11} + \Sigma_{22}\right)^{-1}\right]'$$

$$+ \Sigma_{22} - \Sigma_{22}\left(\Sigma_{11} + \Sigma_{22}\right)^{-1}\Sigma_{22}$$

$$= \left[A - \left(\Sigma_{11}^{-1} + \Sigma_{22}^{-1}\right)^{-1}\Sigma_{11}^{-1}\right]\left(\Sigma_{11} + \Sigma_{22}\right)\left[A - \left(\Sigma_{11}^{-1} + \Sigma_{22}^{-1}\right)\Sigma_{11}^{-1}\right]'$$

$$+ \left(\Sigma_{11}^{-1} + \Sigma_{22}^{-1}\right)^{-1}$$

where the algebra of the final step is justified in the next paragraph. Each term in the final expression is an inner product matrix and the first term can be eliminated by choosing

$$A = \left(\Sigma_{11}^{-1} + \Sigma_{22}^{-1}\right)^{-1}\Sigma_{11}^{-1}$$

with the result that the minimized variance matrix is $(\Sigma_{11}^{-1} + \Sigma_{22}^{-1})^{-1}$.

The missing algebra is given by

$$\Sigma_{22}(\Sigma_{11} + \Sigma_{22})^{-1} = \Sigma_{22}\Sigma_{22}^{-1}(\Sigma_{11}\Sigma_{22}^{-1} + I)^{-1}$$

$$= (\Sigma_{22}^{-1} + \Sigma_{11}^{-1})^{-1}\Sigma_{11}^{-1} ,$$

and

$$\Sigma_{22} - \Sigma_{22}(\Sigma_{11} + \Sigma_{22})^{-1}\Sigma_{22} = \Sigma_{22} - (\Sigma_{22}^{-1} + \Sigma_{11}^{-1})^{-1}\Sigma_{11}^{-1}\Sigma_{22}$$

$$= (\Sigma_{22}^{-1} + \Sigma_{11}^{-1})^{-1}(\Sigma_{22}^{-1} + \Sigma_{11}^{-1} - \Sigma_{11}^{-1})\Sigma_{22}$$

$$= (\Sigma_{22}^{-1} + \Sigma_{11}^{-1}) .$$

(c) Combining potential estimates by weighted averages

Suppose that we know that the mean of a response $\underset{\sim}{y} = (y_1, \ldots, y_n)'$ lies somewhere in an r dimensional subspace formed by vectors $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_r$ and that the variation about the mean is scaled by a common standard deviation $\sigma(\theta)$ and has zero correlation between coordinates:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1r} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nr} \end{pmatrix} \begin{pmatrix} \beta_1(\theta) \\ \vdots \\ \beta_r(\theta) \end{pmatrix} + \sigma(\theta) \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

where $(\underset{\sim}{\beta}, \sigma)$ is in $\mathbb{R}^r \times \mathbb{R}^+$, and the $u_1, \ldots, u_n$ have zero mean, unit variance, and zero correlation. This is called the *linear model* with homoscedastic uncorrelated error. It can be presented in vector matrix notation as

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \sigma\underset{\sim}{u}$$

and X is called the *design matrix*.

In an application the vectors $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ would represent the patterns that one might expect in the general level for the response vector: the model presents the mean or general level of the response vector as an element of $\mathcal{L}(\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)$. The response vector $\underset{\sim}{y}$ deviates from the general level $X\underset{\sim}{\beta}$ due to a response scaling σ of the vector $\underset{\sim}{u}$ describing the internal variation in the system. Sometimes the space for the mean response is not linear but can be approximated over a reasonable range by the linear space represented by an $X\underset{\sim}{\beta}$.

Each of the coordinates y_i can be thought of as an ingredient of an estimate for the β's ; certainly the mean of each y_i is linear in the β's . The following gauss markov theorem is concerned with using linear combinations of the y's to estimate the β's .

Theorem 3. (i) *A linear construct* $A\underset{\sim}{y}$ *is an unbiased estimate of* $\underset{\sim}{\beta}$ *if* $AX = I$. (ii) *Among such linear unbiased estimates, the estimate*

$$\hat{\beta} = (X'X)^{-1}X'\underset{\sim}{y}$$

has minimum covariance matrix; the minimum covariance matrix is

$$(X'X)^{-1}\sigma^2 \ .$$

Proof: (i) If $A\underset{\sim}{y}$ is unbiased for $\underset{\sim}{\beta}$ then $EA\underset{\sim}{y} = AX\underset{\sim}{\beta} = \underset{\sim}{\beta}$; hence $AX = I$. (ii) Consider the minimization of the variance matrix of $A\underset{\sim}{y}$,

$$\mathrm{VAR}(A\underset{\sim}{y}) = A\ \mathrm{VAR}(\underset{\sim}{y})A' = AI\sigma^2A' = \sigma^2AA' \ ,$$

or equivalently the minimization of the matrix

$$AA' = (A - \cdots)(A - \cdots)' + \cdots$$

where we seek entries on the right side that utilize the relation $AX = I$; this suggests

$$AA' = (A - CX')(A - CX')' + K$$

$$= AA' - C - C' + CX'XC' + K \; .$$

The right side simplifies if $C = (X'X)^{-1}$ and $K = C$:

$$AA' = \left(A - (X'X)^{-1}X'\right)\left(A - (X'X)^{-1}X'\right) + (X'X)^{-1} \; .$$

Each term on the right side is an inner product matrix and the first term can be elimated by choosing

$$A = (X'X)^{-1}X'$$

with the result that the minimized variance matrix is $(X'X)^{-1}\sigma^2$.

The estimate obtained by the gauss markov theorem is the least squares estimate due to Gauss. The least squares estimate is obtained as the point $X\underset{\sim}{b}$ closest to $\underset{\sim}{y}$ in the sense of least squares (the euclidean distance). The least squares estimate can be derived by the same method of completing a quadratic expression: the squared distance from $\underset{\sim}{y}$ to $X\underset{\sim}{b}$ is

$$(\underset{\sim}{y} - X\underset{\sim}{b})'(y - Xb)$$

$$= b'X'Xb - b'X'\underset{\sim}{y} - \underset{\sim}{y}'X\underset{\sim}{b} + \underset{\sim}{y}'\underset{\sim}{y}$$

$$= (b - \cdots)'X'X(b - \cdots) + \cdots$$

$$= \big(b - (X'X)^{-1}X'\underset{\sim}{y}\big)'X'X\big(b - (X'X)^{-1}X'\underset{\sim}{y}\big)$$

$$+ \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}'X(X'X)^{-1}X'y \;;$$

thus the closest point is obtained with

$$\underset{\sim}{b} = \underset{\sim}{\hat{b}} = (X'X)^{-1}X'\underset{\sim}{y}$$

and the minimized distance is

$$s^2(\underset{\sim}{y}) = (\underset{\sim}{y} - X\underset{\sim}{\hat{b}})'(\underset{\sim}{y} - X\underset{\sim}{\hat{b}})$$

$$= \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}'X(X'X)^{-1}X'\underset{\sim}{y}$$

$$= \underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}'X\underset{\sim}{b} \;.$$

See Figure 1.

Now consider the generalized linear model

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{u}$$

where X has rank r and $\underset{\sim}{u}$ has zero mean and variance matrix $M\sigma^2$. Then the generalized gauss markov theorem is

Theorem 4. (i) *A linear construct* $A\underset{\sim}{y}$ *is an unbiased estimate of* $\underset{\sim}{\beta}$ *if* $AX = I$. (ii) *Among such linear unbiased estimates,*

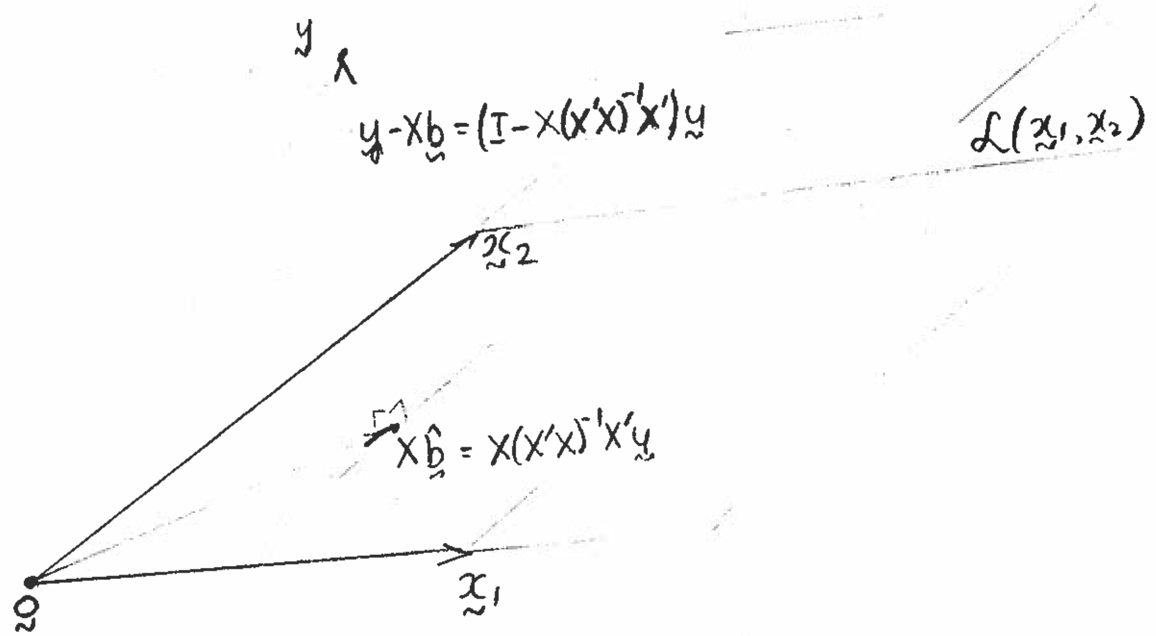
$\underset{\sim}{y}$
$\underset{\sim}{y} - X\underset{\sim}{b} = (I - X(X'X)^{-1}X')\underset{\sim}{y}$
$\mathcal{L}(\underset{\sim}{x}_1, \underset{\sim}{x}_2)$
$\underset{\sim}{x}_2$
$X\hat{\underset{\sim}{b}} = X(X'X)^{-1}X'\underset{\sim}{y}$
$\underset{\sim}{x}_1$
$\underset{\sim}{0}$

Figure 1: The point Xb in $L(X)$ that is closest to $\underset{\sim}{y}$ has $\hat{b} = (X'X)^{-1}X'\underset{\sim}{y}$. The projection $X\hat{b}$; the orthogonal complement or deviation vector $\underset{\sim}{y} - X\underset{\sim}{b}$.

the estimate

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}\underset{\sim}{y}$$

has minimum covariance matrix; the minimum covariance matrix is

$$(X'V^{-1}X)^{-1}\sigma^2 \ .$$

Problems

1. Consider the sample average $\bar{y}$ and median $\tilde{y}$ for a sample $(y_1, \ldots, y_n)$ from the normal (θ, σ_0^2) . The covariance of $\bar{y}$ and $\tilde{y}$ is σ^2/n . Formally derive the best linear combination of $\bar{y}$ and $\tilde{y}$ as an estimate of θ .

2. Let $\bar{y}_1$ be an estimate of μ based on a first sample of m and let $\bar{y}_2$ be an estimate of μ based on a second sample of n from the same distribution. Derive the linear compound of $\bar{y}_1$ and $\bar{y}_2$ that has minimum variance (assume the variance of the underlying distribution exists).

3. Let $t_1, \ldots, t_k$ be independent unbiased estimates of θ with variances $\sigma_1^2, \ldots, \sigma_k^2$. Show that $(\sigma_1^{-2} + \cdots + \sigma_k^{-2})^{-1}(\sigma_1^{-2}t_1 + \cdots + \sigma_k^{-2}t_k)$ is the minimum variance linear compound that is unbiased for θ .

4. Let $\underset{\sim}{t}$ and $\underset{\sim}{u}$ be independent estimates of $\underset{\sim}{\beta}$ with variance matrices

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\sigma^2 \quad \text{and} \quad \begin{pmatrix} 1 & .95 \\ .95 & 1 \end{pmatrix}\sigma^2$$

respectively. Determine the unbiased linear compound of $\underset{\sim}{t}$ and $\underset{\sim}{u}$ which has minimum variance matrix.

5. Use Theorem 3 to prove the special case given by Theorem 2.

6. Let $\underset{\sim}{t}_1, \ldots, \underset{\sim}{t}_k$ be uncorrelated unbiased estimates of $\underset{\sim}{\beta}$ with variance matrices $\Sigma_{11}(\theta), \ldots, \Sigma_{kk}(\theta)$. Show that $(\Sigma_{11}^{-1} + \cdots + \Sigma_{kk}^{-1})^{-1}(\Sigma_{11}^{-1}\underset{\sim}{t}_1 + \cdots + \Sigma_{kk}^{-1}\underset{\sim}{t})$ with matrices evaluated at θ_0 gives the best linear compound having minimum variance matrix at $\underset{\sim}{\theta}_0$. Use Theorem ~~3~~.4

7. Let $\underset{\sim}{y} = \beta\underset{\sim}{x} + \sigma\underset{\sim}{e}$ where the e's have unit variances and are uncorrelated. Present the model in general linear model form and show the best linear unbiased estimate of β is $\hat{\beta} = \Sigma x_i y_i / \Sigma x_i^2$.

8. Let $\underset{\sim}{y}_1 = (y_{11}, \ldots, y_{n1})' = \mu_1 \underset{\sim}{1} + \underset{\sim}{u}_1$ and $\underset{\sim}{y}_2 = (y_{11}, \ldots, y_{n1})' = \mu_2 \underset{\sim}{1} + \underset{\sim}{v}_2$ where the u's and v's have zero means, common variances σ^2 and zero covariances. Present the combined model in general linear model form and determine the unbiased estimate of $(\mu_1, \mu_2)'$ which is linear in $\underset{\sim}{y}_1$ and $\underset{\sim}{y}_2$ and has minimum variance matrix.

9. Let $\underset{\sim}{y} = \alpha\underset{\sim}{1} + \beta\underset{\sim}{x} + \sigma\underset{\sim}{e}$ where the e's have unit variances and are uncorrelated. Present the model in general linear model form and show that the best linear unbiased estimate of (α, β) has $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and $\hat{\beta} = \Sigma(x_i - \bar{x})y_i / \Sigma(x_i - \bar{x})^2$. Determine the covariance matrix of $(\hat{\alpha}, \hat{\beta})$.

10. Consider the method of least square for finding $X\underset{\sim}{b}$ in $L(X)$ closest to $\underset{\sim}{y}$. Show that the projection of $\underset{\sim}{y}$ into the subspace $L(X)$ is obtained from the transformation matrix $X(X'X)^{-1}X'$ and that the projection of $\underset{\sim}{y}$ into $L^{\perp}(X)$ is obtained from the transformation matrix $(I - X(X'X)X')$.

11. Continuation. A projection matrix P has the idempotent property $P = PP$. Check that the projection matrices in Problem 10 are idempotent.

12. Another proof of Theorem 2. Note that $\underset{\sim}{d} = \underset{\sim}{t}_1 - \underset{\sim}{t}_2$ is an unbiased estimate of zero. Show that $\underset{\sim}{u} = \Sigma_{11}^{-1}(\theta_0)\underset{\sim}{t}_1 + \Sigma_{22}^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{t}_2$ has zero covariance with $\underset{\sim}{d}$. Show that $A\underset{\sim}{u} + B\underset{\sim}{d}$ generates all linear compounds of $\underset{\sim}{t}_1$ and $\underset{\sim}{t}_2$. Determine A, B to obtain the minimum variance matrix (at $\underset{\sim}{\theta}_0$) unbiased estimate of $\underset{\sim}{\beta}$.

13. Extension of Theorem 2 for the case of possibly correlated estimates. Show that the proof as given can be modified in detail and the best estimate uses $S_{11} = \Sigma_{11} - \Sigma_{12}$ and $S_{22} = \Sigma_{22} - \Sigma_{21}$ in place of Σ_{11} and Σ_{22}. Note that S_{11} and S_{22} are no longer symmetric. Rule: Weight by the inverse of the excess of variance over covariance.

14. Prove the general gauss-markov Theorem 4.

15. Show that the best linear unbiased estimate $\hat{\beta}$ in the generalized gauss-markov theorem is The weighted least squares solution obtained by minimizing $(\underset{\sim}{u} - X\underset{\sim}{b})'V^{-1}(\underset{\sim}{y} - X\underset{\sim}{b})$.

16. Consider the normal linear model where $\underset{\sim}{y} = X\underset{\sim}{\beta} + \sigma\underset{\sim}{e}$ and $\underset{\sim}{e}$ is a sample from the standard normal and X is the design matrix with rank r . Show that the maximum likelihood estimate of β is $\hat{\beta} = (X'X)^{-1}X'\underset{\sim}{y}$ which also is the minimum variance linear estimate (gauss markov) and is the least squares estimate. And show that the least squares estimate of σ^2 is $\hat{\sigma}^2 = (\underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}'X\underset{\sim}{b})/n$ which is n^{-1} time the minimized sum of squares.

2. UNBIASEDNESS: BY ANALYZING LOCALLY

Unbiasedness was discussed in Section 1 from the point of view of obtaining best averages among given estimates or potential estimates. In this section we discuss unbiasedness from the point of view of a small neighbourhood on the parameter space. We discuss it as if we knew that the true parameter value was in the small neighbourhood; we search for the best estimate in the restricted problem that has this very small parameter space. This approach *does* produce solutions for the original model with a full parameter space.

For notation let y be the response with values in a sample space S, and let θ be a *real valued* parameter with values in $\Omega = \mathbb{R}$. We assume that the class of measures can be represented by a density function satisfying the general assumptions in Seven: Section 1 and the differentiation assumptions in Seven: Section 5(i), (ii).

Now consider unbiasedness for a parameter space that consists of a small neighbour of some value θ_0. The regular definition of unbiasedness for a function $t(y)$ can be expressed as:

$$E(t(y)|\theta) \equiv \theta_0 + (\theta - \theta_0) \qquad \text{all } \theta \text{ in } \Omega .$$

We now define *local unbiasedness* (θ_0) as satisfying the preceding up to a first derivative at θ_0 :

ID t *is locally unbiased* (θ_0) *if*

(i) $E\left(t(y)|\theta_0\right) = \theta_0$,

(ii) $\frac{\partial}{\partial\theta} E\left(t(y)|\theta\right)\Big|_{\theta=\theta_0} = 1$.

Obviously a (globally) unbiased estimate is locally unbiased at any parameter point. On the other hand a locally unbiased (θ_0) estimate need not be unbiased for other parts of the parameter space.

(a) A variance bound for locally unbiased estimates

In Seven: Section 5 we differentiated

$$\int f(y|\theta)dy = 1$$

and obtained

$$E\left\{S(y|\theta)|\theta\right\} = 0$$

$$Var\left\{S(y|\theta)|\theta\right\} = I(\theta)$$

$$= -E\left\{S'(y|\theta)|\theta\right\} .$$

For the present context we view y as the full response under consideration and correspondingly $S(y|\theta)$ and $I(\theta)$ as the full score and information functions; if there is a sample then it is represented here by the response y .

Now let $t(y)$ be a locally unbiased (θ_0) estimate:

$$\int t(y)f(y|\theta)dy = \theta_0 + (\theta - \theta_0) + o(\theta - \theta_0) \ ;$$

and assume that its variance exists for θ near θ_0 . In order to apply our differentiation routine on the preceding integral we assume

(iii) $|\partial \ell n\ f(y|\theta)/\partial\theta|^2 f(y|\theta) < M(y)$ *in a neighbourhood of* θ_0 *and* $M(y)$ *is integrable,*

and use Four: Problem 4 and Section 1 (9). Now differentiating the integral at θ_0 and using

$$\frac{\partial f(y|\theta)}{\partial\theta} = S(y|\theta)f(y|\theta)$$

we obtain

$$\int t(y)S(y|\theta_0)f(y|\theta_0)dy = 1$$

or

$$\text{cov}\left[t(y),\ S(y|\theta_0)|\theta_0\right] = 1 \ .$$

The covariance inequality in Four: Section 2 can now be used to obtain

$$\mathrm{Var}(t(y)|\theta_0)\mathrm{Var}\left(S(y|\theta_0)|\theta_0\right) \geq 1$$

or equivalently:

Theorem: *If* t *is a locally unbiased* (θ_0) *estimate for a model satisfying* (i), (ii) *in Seven, Section 5, and* (iii) *in this section, then*

$$\mathrm{Var}(t(y)|\theta_0) \geq \frac{1}{I(\theta_0)}$$

with equality if and only if t(y) *and* $S(y|\theta_0)$ *are affinely related.*

This *information inequality* gives a lower bound to the variance of a locally unbiased (θ_0) estimate.

Corollary: *If* t *is an unbiased estimate of* θ *, then*

$$\mathrm{Var}(t(y)|\theta) \geq \frac{1}{I(\theta)}$$

with equality at some θ *values if and only if* t(y) *and* $S(y|\theta)$ *are affinely related at those* θ *values.*

Example 3. Let $(y_1, \ldots, y_n)$ be a sample from the normal distribution (θ, σ_0^2) with θ in $\Omega = \mathbb{R}$. We have

$$S(\underset{\sim}{y}|\theta) = \frac{\Sigma(y_i - \theta)}{\sigma_0^2} ,$$

$$I(\theta) = \frac{n}{\sigma_0^2} ;$$

and hence

$$\mathrm{Var}(t(\underset{\sim}{y})|\theta_0) \geq \frac{\sigma_0^2}{n}$$

for any locally unbiased (θ_0) estimate. But we know that $\bar{y}$ is a (globally) unbiased estimate with variance σ_0^2/n at the information lower bound. Thus $\bar{y}$ has *uniformly* (re θ) *minimum variance* among unbiased estimates of θ ; it is UMV unbiased.

(b) The best locally unbiased estimate

Consider the locally unbiased (θ_0) estimation of the parameter θ . The theorem on the information lower bound contains a clue as to how to construct the best locally unbiased estimate. We consider affine functions

$$a + bS(y|\theta_0)$$

of the score and try to determine a, b so that the function is locally unbiased (θ_0) ; if we are successful then we have a locally unbiased (θ_0) estimate that has minimum variance.

The first property of a locally unbiased (θ_0) estimate gives

$$E\left(a + bS(y|\theta_0)|\theta_0\right) = \theta_0 \ .$$

But the mean of S is zero; hence $a = \theta_0$. The second property of a locally unbiased (θ_0) estimate gives

$$\frac{\partial}{\partial\theta} E\left(a + bS(y|\theta_0)|\theta\right)\Big|_{\theta=\theta_0} = 1 \ .$$

Again applying our differentiation routine through the sign of integration we obtain

$$\int\left(a + bS(y|\theta_0)\right)S(y|\theta_0)f(y|\theta_0)dy = 1$$

or

$$b\int S^2(y|\theta_0)f(y|\theta_0)dy = 1 \ ;$$

hence $b = I^{-1}(\theta_0)$. Thus the estimate

$$\theta_0 + I^{-1}(\theta_0)S(y|\theta_0)$$

is locally unbiased (θ_0) and it has minimum variance $I^{-1}(\theta_0)$ among locally unbiased (θ_0) estimates.

Note the very interesting property that this best locally unbiased estimate uses only the score $S(y|\theta_0)$; the score is all there is to the likelihood function if we restrict our parameter space to first derivative change at θ_0 . This is consistent with our earlier argument that only the observed likelihood function should be used with the model for purposes of inference.

Example 3 continued. Consider the sample $(y_1, \ldots, y_n)$ from the normal (θ, σ_0^2) with θ in $\mathbb{R}$. The best locally unbiased (θ_0) estimate is

$$t(\underset{\sim}{y}) = \theta_0 + \frac{1}{n/\sigma_0^2} \frac{\Sigma(y_i - \theta_0)}{\sigma_0^2}$$

$$= \bar{y} .$$

Note that it does not depend on θ_0 . Thus we have obtained an estimate that is the best locally unbiased (θ_0) for all θ_0 ; accordingly, it is the globally UMV unbiased estimate of θ .

As a simple corollary to the preceding analysis we obtain: *If* t *is an unbiased estimate of* θ *with variance at the information lower bound then*

$$t(y) = \theta + I^{-1}(\theta)S(y|\theta)$$

and the right side is independent of θ .

(c) What models have information-bound estimates?

Under the assumptions (i), (ii) and (iii) we now consider what models admit unbiased estimates at the information lower bound. If t is an unbiased estimate of θ with variance at the information lower bound then t has the following form

$$t(y) = \theta + I^{-1}(\theta)S(y|\theta)$$

or

$$S(y|\theta) = I(\theta)t(y) - \theta I(\theta) .$$

Integration with respect to θ then gives

$$\ell n\ f(y|\theta) = \psi(\theta)t(y) + \phi(\theta) + k(y)$$

where $k(y)$ is the constant of integration and ψ and ϕ are the obvious indefinite integrals. Thus

$$f(y|\theta) = \gamma(\theta)\exp\{t(y)\psi(\theta)\}h(y)$$

and the model is an exponential model with one ψ function.

Thus only the exponential model admits unbiased estimates at the information lower bound.

Now consider an exponential model

$$f(y|\theta) = \exp\{\phi(\theta) + t(y)\psi(\theta)\}h(y)$$

with θ in $\Omega = \mathbb{R}$. The parameter θ may not have an unbiased estimate at the information lower bound. In fact only the parameter $\phi'(\theta)/\psi'(\theta)$ has an unbiased estimate at its information bound.

(d) With a vector parameter

Now consider the case of a vector parameter $\underset{\sim}{\theta} = (\theta_1, \ldots, \theta_r)'$ in $\Omega = \mathbb{R}^r$ or in a connected open set of $\mathbb{R}^r$. And suppose that the vector analogues of the assumptions (i), (ii), (iii) hold.

An estimate $\underset{\sim}{t}$ of $\underset{\sim}{\theta}$ *is locally unbiased* (θ_0) *if*

(i) $E(\underset{\sim}{t}(y)|\underset{\sim}{\theta}_0) = \theta_0$

(ii) $\dfrac{\partial}{\partial\theta_s} E(t_{s'}(y)|\underset{\sim}{\theta})\Big|_{\underset{\sim}{\theta}=\underset{\sim}{\theta}_0} = \delta_{ss'}$

where $\delta_{ss'} = 1$ if $s = s'$ and $= 0$ otherwise.

From Seven: Section 5 we have

$$E\left[\underset{\sim}{S}(y|\underset{\sim}{\theta})\,|\,\underset{\sim}{\theta}\right] = 0$$

$$\mathrm{var}\left[\underset{\sim}{S}(y|\underset{\sim}{\theta})\,|\,\underset{\sim}{\theta}\right] = I(\underset{\sim}{\theta})$$

$$= -E\left\{\frac{\partial \underset{\sim}{S}(y|\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}'}\,\middle|\,\underset{\sim}{\theta}\right\}$$

where $I(\underset{\sim}{\theta})$ is now the information matrix for the full sample. And by the differentiation routine the local unbiasedness gives

$$\mathrm{COV}\left[\underset{\sim}{t}(y), \underset{\sim}{S}(y|\underset{\sim}{\theta})\right] = I \ ,$$

the $r \times r$ identity matrix.

The information inequality for a locally unbiased $(\underset{\sim}{\theta}_0)$ estimate is

$$\mathrm{VAR}(\underset{\sim}{t}(y)\,|\,\theta_0) \succeq I^{-1}(\theta_0) \ .$$

The best locally unbiased $(\underset{\sim}{\theta}_0)$ estimate is

$$\underset{\sim}{\theta} + I^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{S}(y|\underset{\sim}{\theta}_0) \ .$$

An unbiased estimate at the information lower bound has the form

$$\underset{\sim}{t}(\underset{\sim}{y}) = \underset{\sim}{\theta} + I^{-1}(\underset{\sim}{\theta})\underset{\sim}{S}(\underset{\sim}{y}|\underset{\sim}{\theta}) \ ;$$

and this occurs only for an exponential model with r ψ functions.

Problems

17. Let $(x_1, \ldots, x_n)$ be a sample from the bernoulli distribution with p in $\Omega = (0, 1)$. Derive the minimum variance locally unbiased (p_0) estimate of p. What about a UMV unbiased estimate?

18. Let y have a poisson distribution (θ) with θ in $\mathbb{R}^+$. Derive the minimum variance locally unbiased (θ_0) estimate of θ. What about a UMV unbiased estimate?

19. Let $(y_1, \ldots, y_n)$ be a sample from the exponential distribution $f(y|\theta) = \theta^{-1}\exp\{-y/\theta\}$ on $\mathbb{R}^+$ with θ in $\Omega = \mathbb{R}^+$. Derive the minimum variance locally unbiased (θ_0) estimate of θ. What about a UMV unbiased estimate?

20. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ_0, σ^2) with σ^2 in $\mathbb{R}^+$. Derive the minimum variance locally unbiased (σ_0^2) estimate of σ^2. What about a UMV unbiased estimate?

21. Consider a statistical model $f(y|\theta)$ satisfying the assumptions in Section 2 and with a real parameter θ in $\Omega = \mathbb{R}$.

Now let $\beta(\theta)$ be a continuously differentiable monotone increasing function mapping Ω into $\Omega^* \subset \mathbb{R}$, and let $b(\theta) = d\theta(\beta)/d\beta$. Then show that the score function and the information function (relativ to β) are $S(y|\theta)b(\theta)$ and $I(\theta)b^2(\theta)$.

22. Let $(y_1, \ldots, y_n)$ be a sample from $f(y|\theta) = \theta y^{\theta-1}$ on $(0, 1)$ and $= 0$ otherwise with θ in $\Omega = \mathbb{R}^+$. Derive the minimum variance locally unbiased (θ_0) estimate of θ. What about a UMV unbiased estimate?

23. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ, σ^2) with $\theta = (\mu, \sigma^2)$ in $\mathbb{R} \times \mathbb{R}^+$. Calculate $\underset{\sim}{S}(\underset{\sim}{y}|\underset{\sim}{\theta})$ and $I(\theta)$.

24. Continuation. Derive the minimum variance-matrix locally unbiased $(\underset{\sim}{\theta}_0)$ estimate of $\underset{\sim}{\theta}$. What about an unbiased estimate with variance matrix at the information lower bound?

25. Continuation. Calculate the variance matrix of $(\bar{y}, s_y^2)$ and compare with the information lower bound.

26. Continuation. Show that the only parameter $\underset{\sim}{\beta}$ with an unbiased estimate at the information lower bound is $(\mu, \mu^2 + \sigma^2)$ (or an affine equivalent).

27. Consider independent models: $f_1(y|\theta)$ with $S_1(y|\theta)$ and $I_1(\theta)$; and $f_2(y|\theta)$ with $S_2(y|\theta)$ and $I_2(\theta)$. For

local unbiased (θ_0) estimation we have $\theta_0 + I^{-1}(\theta_0)S_1(y|\theta)$ from the first model and $\theta_0 + I_2^{-1}(\theta_0)S_2(y|\theta_0)$ from the second model. Combine these local estimates by the methods in Section 1 and show that the combined estimate is the best locally unbiased (θ_0) estimate from the combined model.

28. Continuation. The vector version: $f_1(y_1|\underset{\sim}{\theta})$ with $\underset{\sim}{S}_1(y|\underset{\sim}{\theta})$ and $I(\underset{\sim}{\theta})$; and $f_2(y|\theta)$ with $\underset{\sim}{S}_2(y|\theta)$ and information matrix $J(\theta)$. Combine the best locally unbiased (θ_0) estimates by the methods in Section 1 and show that the combined estimate is the best locally unbiased $(\underset{\sim}{\theta}_0)$ estimate from the oombined model.

29. Let Σ be the variance matrix of $\begin{pmatrix} \underset{\sim}{x} \\ \underset{\sim}{y} \end{pmatrix}$ and let $\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ be the corresponding matrix partition. If Σ_{11} is non-singular, then prove the generalized covariance inequality: $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \succ 0$ (zero matrix) with equality if and only if $\underset{\sim}{y} = \underset{\sim}{a} + B\underset{\sim}{x}$. Compare with Four: Problem 105.

30. Continuation. Prove the vector form of the information inequality for the variance matrix of a locally unbiased (θ_0) estimate of $\underset{\sim}{\theta}$.

31. Continuation. Show that the best locally unbiased $(\underset{\sim}{\theta}_0)$ estimate is $\underset{\sim}{\theta} = \underset{\sim}{\theta}_0 + I^{-1}(\underset{\sim}{\theta}_0)\underset{\sim}{S}(y|\underset{\sim}{\theta}_0)$.

32. Continuation. Show that the information lower bound is attained for the unbiased estimate of the vector parameter only if the model is exponential with r ψ functions.

33. Continuation of Subsection (c). Show that the only parameters with unbiased estimates at the information lower bound (exponential model with one ψ function) are these with the form $a + c\phi'(\theta)/\psi'(\theta)$.

34. Continuation. Consider the reduced exponential model $f(y|\theta) = \exp\{\phi(\theta) + \underset{\sim}{\psi}'(\theta)\underset{\sim}{a}(y)]h(y)$ with r ψ-functions and $\underset{\sim}{\theta}$ in $\mathbb{R}^r$. Show that the only vector (r dim) parameters with unbiased estimates with variance matrix at the information lower bound are those with the form $\underset{\sim}{a} + C(\partial\underset{\sim}{\psi}'(\theta)/\partial\underset{\sim}{\theta})^{-1}\partial\phi(\theta)/\partial\underset{\sim}{\theta}$ where the prime here denotes transpose and C is $r \times r$.

3. UNBIASEDNESS: BY TAKING MEANS

We now discuss some general questions connected with unbiasedness. Does a particular parameter have unbiased estimates? Can we improve a given unbiased estimate? Is there a best unbiased estimate in the sense of say minimum variance?

For notation, let y be the response in a sample space S and let θ be the general parameter with values in Ω. And suppose we are interested in some real valued parameter $\beta(\theta)$ or some vector valued parameter $\underset{\sim}{\beta}(\theta)$.

In SEVEN we have seen that the likelihood function is all that the model presents concerning a response value and that for purposes of inference it suffices to have the value of the likelihood statistic. Now for unbiased estimation we should expect some clear preferences to show for the unbiased estimates that depend on the likelihood statistic. The following rao-blackwell theorem not only gives technical preferences but also tells how an arbitrary unbiased estimate can be converted into one dependent on the likelihood statistic.

(a) By taking means

Variance is reduced by taking means. For an unbiased estimate the variance can be reduced by taking means over component parts of the distribution that do not involve the parameter.

Theorem. *If* $\underset{\sim}{t}(y)$ *is an unbiased estimate of* $\underset{\sim}{\beta}(\theta)$ *and if* $s(y)$ *is the likelihood statistic, then*

(i) $E\left(\underset{\sim}{t}(y) : s(y)\right) = \underset{\sim}{r}\left(s(y)\right)$ *is an unbiased estimate of* $\underset{\sim}{\beta}(\theta)$ *based on the likelihood statistic* $s(y)$.

(ii) $\text{VAR}\left(\underset{\sim}{t}(y)|\theta\right) \leq \text{VAR}\left[\underset{\sim}{r}\left(s(y)\right)|\theta\right]$ *with equality if and only if* $\underset{\sim}{t}(y) = \underset{\sim}{r}\left(s(y)\right)$ *is already a function of* $s(y)$ *with probability one.*

Proof. The conditional distribution given the likelihood statistic does not depend on the parameter θ . Accordingly, the conditional mean

$$E\left(\underset{\sim}{t}(y) : s(y)\right) = \underset{\sim}{r}\left(s(y)\right)$$

does not depend on θ and is thus a function on the sample space. Then by Seven, Section 5, the mean of the conditional mean is the marginal mean

$$E\left[E\left(\underset{\sim}{t}(y) : s(y)\right)|\theta\right] = E\left(\underset{\sim}{t}(y)|\theta\right) = \underset{\sim}{\beta}(\theta) \;;$$

this establishes the unbiasedness of $\underset{\sim}{r}\left(s(y)\right)$.

We now apply the formula for mean variance about regression (Four: Section 5):

$$E\left[\text{VAR}\left(\underset{\sim}{t}(y) : s(y)\right)|\theta\right]$$
$$= \text{VAR}\left(\underset{\sim}{t}(y)|\theta\right) - \text{VAR}\left[\underset{\sim}{r}\left(s(y)\right)|\theta\right] .$$

Thus the variance matrix for $\underset{\sim}{r}(s(y))$ is smaller than the variance matrix for $\underset{\sim}{t}(y)$, with equality if and only if the mean square of the deviation of $\underset{\sim}{t}(y)$ from its conditional mean is zero; that is, if and only if $\underset{\sim}{t}(y)$ is equal to its conditional mean with probability one.

Note that the theorem as stated remains true if $s(y)$ is any sufficient statistic. The likelihood statistic provides the largest reduction and would thus be the preferred sufficient statistic whenever it is available.

The theorem shows that an unbiased estimate not based on the likelihood statistic can be improved by taking its mean value given the likelihood statistic. This reduces the variance matrix; it also reduces the variance of any coordinate that is not already a function of the likelihood statistic; and it reduces the variance of any linear combination of coordinates (as an unbiased estimate of the corresponding linear combination of parameter coordinates) that is not already a function of the likelihood statistic.

Example 4. Consider a sample $(x_1, \ldots, x_n)$ from the bernoulli distribution (p) with p in $[0, 1]$. Suppose we are interested in estimating p and have been given the very naive estimate x_1 for p. Of course x_1 can take only the values $0, 1$ whereas p can be any value in $[0, 1]$; nevertheless p

is technically unbiased:

$$E(x_1|p) = 1\cdot p + 0\cdot(1-p)$$

$$= p \ .$$

The likelihood statistic for the sample $(x_1, \ldots, x_n)$ is the binomial variable $y = \Sigma_1^n x_i$. We apply the theorem and calculate the mean value of x_1 given y :

$$E(x_1 : y) = 1\,\frac{y}{n} + 0\,\frac{n-y}{n}$$

$$= \frac{\Sigma x_i}{n} = \hat{p} \ .$$

The conditional distribution is available from symmetry: y 1's and $n-y$ 0's and equal probability to each possible sequence. Note that

$$\mathrm{Var}(x_1|p) = pq \ , \quad \mathrm{Var}(\hat{p}|p) = \frac{pq}{n} \ .$$

Thus the theorem takes us from the trivial estimate x_1 to the estimate $\hat{p}$ that we have considered several times before.

Example 3 continued. Consider a sample $(y_1, \ldots, y_n)$ from the normal distribution (θ, σ_0^2) with θ in $\Omega = \mathbb{R}$. Suppose we are interested in θ and that we have taken the first estimate that has been presented to us; say y_1 . Of course

$$E(y_1) = \theta .$$

The likelihood statistic is Σy_i or equivalently $v_1 = \sqrt{n}\,\bar{y}$. The notation in Seven: Section 2 allows us to express

$$y_1 = \frac{1}{\sqrt{n}} v_1 + a_{21}v_2 + \cdots + a_{n1}v_n$$

in terms of independent v's that are normally distributed; the means of $v_2, \ldots, v_n$ are equal to zero. We can then calculate the conditional mean of y_1 immediately

$$\begin{aligned} E(y_1|v_1) &= \frac{1}{\sqrt{n}} v_1 + a_{21}\cdot 0 + \cdots + a_{n1}\cdot 0 \\ &= \bar{y} . \end{aligned}$$

Note that

$$\mathrm{Var}(y_1) = \sigma_0^2 , \quad \mathrm{Var}(\bar{y}) = \frac{\sigma_0^2}{n} .$$

Thus the theorem takes us from y_1 to the now familiar $\bar{y}$ as a better unbiased estimate of θ .

The rao-blackwell theorem is a powerful constructive theorem. But it does need the conditional distribution given the likelihood statistic. In the two examples the conditional distribution was derived easily. The examples however are somewhat exceptional; conditional distributions often are tedious and difficult to derive. We now consider some ways of coming to grips directly with the unbiased estimates that are based on the likelihood statistic.

(b) From likelihood expansions

Let $s(y)$ be the likelihood statistic and let $g(s|\theta)$ be its density function. Of course by Seven, Section 3, the likelihood function obtained from y with $f(y|\theta)$ is the same as that obtained from $s(y)$ with $g(s|\theta)$.

Now consider some *estimable* parameter $\beta(\theta)$, *a parameter that has an unbiased estimate.* By the preceding theorem there is an unbiased estimate based on the likelihood statistic $s(y)$ and it has at least as small variance. Thus there is a function $r(s)$ such that

$$\beta(\theta) = \int r(s)g(s|\theta)ds \ .$$

But we can write this

$$\beta(\theta) = \int r(s)k(s)L(s|\theta)ds$$

where $k(s)$ is a factor that scales the representative likelihood $L(s|\theta)$ to give $g(s|\theta)$ as a function of θ . Thus the estimable parameters are those that can be expressed as a linear compound of the likelihood functions (relative to the lebesgue or counting measure).

We can now examine two rather different questions by a common approach.

First. Suppose there is a second model $g^*(s|\theta)$ that has the same family of likelihood functions. Then

$$1 = \int g(s|\theta)ds \qquad \text{all} \quad \theta$$

$$1 = \int g^*(s|\theta)ds \qquad \text{all} \quad \theta \;;$$

hence

$$0 = \int d(s)g(s|\theta)ds \qquad \text{all} \quad \theta$$

where $d(s) = \left[g^*(s|\theta) - g(s|\theta)\right] \big/ g(s|\theta)$ is independent of θ by the common likelihood property. If g^* is different from g then there must be a 'nontrivial unbiased estimate $d(s)$ of zero'.

Second. Suppose that a parameter $\beta(\theta)$ has two unbiased estimates based on the likelihood statistic. Then

$$\beta(\theta) = \int r(s)g(s|\theta)ds \qquad \text{all} \quad \theta$$

$$\beta(\theta) = \int r^*(s)g(s|\theta)ds \qquad \text{all} \quad \theta \, ;$$

hence

$$0 = \int d(s)g(s|\theta)ds \qquad \text{all} \quad \theta$$

where $d(s) = r^*(s) - r(s)$ is the difference of the estimates. If there are two different unbiased estimates of a parameter, then there must be a 'nontrivial unbiased estimate of zero'.

The two questions lead us to the following definition

D *The statistical model* $\{g(s|\theta) : \theta \in \Omega\}$ *is said to be complete, if*

$$\int d(s)g(s|\theta)ds = 0 \qquad \text{all} \quad \theta$$

implies that $d(s) = 0$ *with probability one.*

A statistical model is complete if there are no nontrivial unbiased estimates of zero. In practice we will say that s is complete provided the context makes clear the statistical model for s.

Example 5. The poisson family is complete. Consider s with a poisson distribution (λ) with λ in $\mathbb{R}^+$. Suppose that $d(s)$ is an unbiased estimate of zero:

$$\sum_{s=0}^{\infty} d(s) \frac{\lambda^s e^{-\lambda}}{s!} = 0 , \qquad \text{all} \quad \lambda \in \mathbb{R}^+ ;$$

$$\sum_{s=0}^{\infty} \frac{d(s)}{s!} \lambda^s = 0 \qquad \text{all} \quad \lambda \in \mathbb{R}^+ .$$

Thus we have a power series in λ that is convergent to zero for all λ on the positive axis. The coefficients of such a power series are unique; hence $d(s) = 0$ for $s = 0, 1, \ldots$. Thus the poisson family is complete.

Example 3 continued. Consider s with a location-normal distribution (θ, τ_0^2) with θ in $\mathbb{R}^+$. Suppose that $d(s)$ is an unbiased estimate of zero:

$$\int_{-\infty}^{\infty} d(s) \frac{1}{\sqrt{2\pi}\, \tau_0} \exp\left\{-\frac{1}{2\tau_0^2} (s-\theta)^2\right\} ds = 0 ,$$

$$\int_{-\infty}^{\infty} \left[d(s) \exp\left\{-\frac{s^2}{2\tau_0^2}\right\}\right] \exp\left\{s \frac{\theta}{\tau_0^2}\right\} ds = 0 ,$$

$$\int_{-\infty}^{\infty} d^+(s) \exp\left\{-\frac{s^2}{2\tau_0^2}\right\} \exp\{st\} ds$$

$$= \int_{-\infty}^{\infty} d^-(s) \exp\left\{-\frac{s^2}{2\tau_0^2}\right\} \exp\{st\} ds$$

for all θ in $\mathbb{R}$ or for all $t = \theta/\tau_0^2$ in $\mathbb{R}$. Let D be the value of the last two integrals when $t = 0$; then

$$\int_{-\infty}^{\infty} p_1(s)\exp\{st\}ds = \int_{-\infty}^{\infty} p_2(s)\exp\{st\}ds$$

for all t in $\mathbb{R}$ where

$$p_1(s) = \frac{d^+(s)}{D} \exp\left\{-\frac{s^2}{2\tau_0^2}\right\}$$

$$p_2(s) = \frac{d^-(s)}{D} \exp\left\{-\frac{s^2}{2\tau_0^2}\right\}$$

are probability density functions. But the equation says that $p_1(s)$ and $p_2(s)$ have the same moment generating functions and hence are identical. Thus $d(s) = 0$ almost everywhere and it follows that the location-normal model is complete.

The analysis preceding the definition of completeness now allows us to state the following theorems.

Theorem. *If the model for the likelihood statistic is complete, then it is the only model that has the given family of likelihood functions.*

The following lehmann-scheffé theorem when used with the rao-blackwell theorem shows that any unbiased estimate based on a complete likelihood statistic is automatically the unique *uniformly minimum variance* (UMV) unbiased estimate of the parameter given by its mean value: the argument being that we always get smaller variance in going to the likelihood statistic and if there is at most one unbiased estimate based on the likelihood statistic then it must be UMV.

Theorem. *There is at most one unbiased estimate of a parameter based on a complete likelihood statistic.*

Example 5 continued. Let $(y_1, \ldots, y_n)$ be a sample from the poisson distribution (θ) with θ in $\Omega = \mathbb{R}^+$. The likelihood statistic is $s = \Sigma y_i$ and we now know that it is complete. It follows for example that $\bar{y}$ is UMV unbiased for θ. And it follows that any function of $\bar{y}$ whose mean exists is UMV unbiased for the parameter represented by its mean value.

Suppose now that we are interested in estimating $e^{-\theta}$, the probability of a zero count for the typical application. We could start with a trivial estimate such as

$$\begin{aligned} h(y_1) &= 1 \qquad y_1 = 0 \,, \\ &= 0 \qquad \text{otherwise;} \end{aligned}$$

its mean is $e^{-\theta}$. And we could then improve it to get the UMV estimate by taking the mean given Σy_i . Or, more analytically, we could go directly to the best estimate r :

$$e^{-\theta} = \sum_{s=0}^{\infty} r(s) \frac{(n\theta)^s \; e^{-n\theta}}{s!} ,$$

$$e^{(n-1)\theta} = \sum_{0=0}^{\infty} r(s) n^s \frac{\theta^s}{s!} ,$$

$$(n-1)^s = r(s) n^s ,$$

$$r(s) = \left(1 - \frac{1}{n}\right)^s .$$

This is the UMV unbiased estimate of $e^{-\theta}$.

Example 3 continued. Let $(y_1, \ldots, y_n)$ be a sample from the normal distribution (θ, σ_0^2) with θ in $\Omega = \mathbb{R}$. The likelihood statistic is $s = \bar{y}$ and we know now that it is complete. It follows that $\bar{y}$ is UMV unbiased for θ . And that $\bar{y}^2$ is UMV unbiased for $E\bar{y}^2 = \theta^2 + \sigma_0^2/n$. And then that $\bar{y}^2 + \frac{n-1}{n} \sigma_0^2$ is UMV unbiased for $\mu_2 = \mu^2 + \sigma^2$. One can generate UMV unbiased estimate of parameters almost as fast as one can write down functions of $\bar{y}$. Of course there is really only *one* parameter θ - however one labels it, and if s is the estimate for θ it is reasonable to think of $\beta(s)$ as the estimate of $\beta(\theta)$. The property of unbiasedness is typically spoiled by a nonlinear function β . The fact that we need to find a new estimate (UMV unbiased) is really a measure of the arbitrariness of the property of unbiasedness.

In some sense θ is the natural parameter here and there is perhaps some consolation in using the best (UMV) unbiased estimate of the natural parameter and appropriately transforming it to get natural estimates of perhaps not quite so natural parameters.

(c) UMV unbiased: uniqueness

We have obtained the UMV unbiased estimate of a parameter from the uniqueness of the unbiased estimate based on a complete likelihood statistic. We can show quite generally however that a UMV unbiased estimate of a parameter is necessarily unique:

Theorem. *If* $t_1(y)$ *and* $t_2(y)$ *are* UMV *unbiased for* $\beta(\theta)$, *then* $t_1 = t_2$ *with probability one.*

Proof. Let $\sigma^2(\theta)$ be the minimum variance attained by t_1 and t_2 , and consider the estimate $a\,t_1(y) + (1-a)t_2(y)$ with $0 < a < 1$.

$$\mathrm{Var}\left(a\,t_1(y) + (1-a)t_2(y)\,|\,\theta\right)$$

$$= E\left[\left(a\left(t_1(y) - \beta(\theta)\right) + (1-a)\left(t_2(y) - \beta(\theta)\right)\right)^2\middle|\,\theta\right] .$$

The square function from $\mathbb{R}$ into $\mathbb{R}$ is strictly convex (Six: Section 2). Thus

$$\left(a\,x_1 + (1-a)x_2\right)^2 \leq a\,x_1^2 + (1-a)x_2^2$$

with equality if and only if $x_1 = x_2$. Thus

$$\mathrm{Var}\left(a\,t_1(y) + (1-a)t_2(y)\,|\,\theta\right)$$

$$\leq a\,\sigma^2(\theta) + (1-a)\sigma^2(\theta) = \sigma^2(\theta)$$

with equality if and only if $t_1(y) = t_2(y)$ with probability one. But $\sigma^2(\theta)$ is the lower bound for the variance; hence the equality must hold and $t_1 = t_2$ with probability one.

(d) Overview

We have obtained the UMV unbiased estimate for the case of a complete likelihood statistic. The problems will demonstrate the types of model that have completeness: the exponential model with the number of ψ function equal to the dimension of the parameter; the variable-carrier models (monotone boundaries) where the number of boundary types is equal to the number of parameters; fairly straightforward mixtures of these. These also are the cases where the dimension of the likelihood statistic does not increase with sample size.

In broad terms the local methods of Section 2 work easily for the natural parameters of the exponential model; the completeness methods work for estimable parameters for the exponential model and for the monotone carrier cases. We will see later just how general this range of models is.

Problems

35*. Let $\underset{\sim}{s} = (s_1, \ldots, s_r)'$ have an exponential model $f(\underset{\sim}{s}|\theta) = \gamma(\theta)\exp\{\Sigma_1^r s_u\theta_u\}h(\underset{\sim}{s})$ with respect to lebesgue measure or a counting measure or some other σ-finite measure on $\mathbb{R}^r$ with $\underset{\sim}{\theta}$ in $\Omega = \mathbb{R}^r$ or = a subset of $\mathbb{R}^r$ that contains a rectangle. Then $\underset{\sim}{s}$ is complete. Method: If the rectangle does not contain the origin, then set $\underset{\sim}{\theta} = \underset{\sim}{\phi} + \underset{\sim}{a}$ and redistribute terms so that the new parameter has a rectangle containing the origin. Follow Example 3 and use the uniqueness theorem for multivariate moment generating functions (Four: Section 4).

36. The scale normal. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ_0, σ^2) with σ^2 in $\Omega = \mathbb{R}^+$. Show that the likelihood statistic $\Sigma(y_i - \mu_0)^2$ is complete. Use Problem 35.

37. The location-scale normal. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ, σ^2) with (μ, σ^2) in $\Omega = \mathbb{R} \times \mathbb{R}^+$. Show that the likelihood statistic $(\Sigma y_i, \Sigma y_i^2)$ is complete. Use Problem 35.

38. The scale exponential. Let $(y_1, \ldots, y_n)$ be a sample from the exponential distribution $f(y|\theta) = \theta^{-1}\exp\{-y/\theta\}$ on $\mathbb{R}^+$ with θ in $\Omega = \mathbb{R}^+$. Show the likelihood statistic Σy_i is complete.

39. Consider the normal linear model $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{u}$ where $\underset{\sim}{u}$ is a sample from the normal $(0, \sigma_0^2)$ with $\underset{\sim}{\beta}$ in $\Omega = \mathbb{R}^r$. Show that the likelihood statistic $X'\underset{\sim}{y}$ is complete. Use Problem **35**. For future reference note that $\underset{\sim}{b}(\underset{\sim}{y}) = (X'X)^{-1}X'\underset{\sim}{y}$ is an equivalent function.

40. Consider the normal linear model $\underset{\sim}{y} = X\underset{\sim}{\beta} + \sigma\underset{\sim}{u}$ where $\underset{\sim}{u}$ is a sample from the normal $(0, 1)$ with $\theta = (\underset{\sim}{\beta}, \sigma^2)$ in $\Omega = \mathbb{R}^r \times \mathbb{R}^+$. Show that the likelihood statistic $(X'y, \underset{\sim}{y}'\underset{\sim}{y})$ is complete. Use Problem 1. For future reference note that $(\underset{\sim}{b}(\underset{\sim}{y}), s^2(\underset{\sim}{y}))$ is an equivalent function.

41. Let $(x_1, \ldots, x_n)$ be a sample from the bernoulli distribution (p) with p in $[0, 1]$. Show that the likelihood statistic $y = \Sigma x_i$ is complete.

42. Let y have the hypergeometric distribution (N, n, D) with D in $\{0, 1, \ldots, N\}$. Show that y is complete.

43. Let $(y_1, \ldots, y_r)$ be multinomial $(n; p_1, \ldots, p_r)$ with $p_i \geq 0$, $\Sigma p_i = 1$. Show that $(y_1, \ldots, y_r)$ is complete.

44. Let $(y_1, \ldots, y_r)$ be multihypergeometric $(N, n, D_1, \ldots, D_r)$ with $D_i = 0, 1, \ldots$ and $\Sigma D_i = N$. Show that $(y_1, \ldots, y_r)$ is complete.

45. Let $(y_1, \ldots, y_n)$ be a sample from the uniform $(0, \theta)$ with θ in $\mathbb{R}^+$. Show that the likelihood statistic $\max y_i$ is complete.

46. Let $(y_1, \ldots, y_n)$ be a sample from the model $f(y|\theta) = k(\theta)\phi(y/\theta)h(y)$ where ϕ is the indicator function for the interval $(0, 1)$. Show that the likelihood statistic $\max y_i$ is complete.

47. Let $(y_1, \ldots, y_n)$ be a sample from the uniform (θ_1, θ_2) with $\theta_1 < \theta_2$ and θ_i in $\mathbb{R}$. Show that the likelihood statistic $(\min y_i, \max y_i)$ is complete.

48. Let (y_1, y_2) be a sample from the uniform $(\theta, \theta+1)$. Show that the likelihood statistic $(y_{(1)}, y_{(2)})$ is *not* complete.

49. The location-scale exponential. Let $(y_1, \ldots, y_n)$ be a sample from $f(y|\theta,\tau)=\tau^{-1}c(y-\theta)\exp\{-(y-\theta)/\tau\}$. Show that the likelihood statistic is $(\min y_i, \Sigma y_i)$. Show that the likelihood statistic is complete.

50. Let $(y_1, \ldots, y_n)$ be a sample from *some* density function on $\mathbb{R}$ (Ω would be an index set for the class of density functions or the class of piecewise continuous density functions). The likelihood statistic is $(y_{(1)}, \ldots, y_{(n)})$:

See Three: Problem 37.

(a) Show that an alternative form of the likelihood statistic is $(\Sigma y_i, \Sigma y_i^2, \ldots, \Sigma y_i^n)$.

(b) The density functions as described above include those of the exponential model $f(y) = \phi(\underset{\sim}{\theta})\exp\{-y^{2n} + \theta_1 y^1 + \cdots + \theta_n y^n\}$. Deduce that the likelihood statistic is complete.

51. Let $(x_1, \ldots, x_n)$ be a sample from the bernoulli (p) with p in $[0, 1]$. Show that the only estimable parameters are the polynomials of degree less than or equal to n in p .

52. Let $(y_1, \ldots, y_k)$ be a sample from the binomial distribution (n, p) with p in $[0, 1]$. Determine the UMV unbiased estimate of q^n , the probability of a zero count.

53. Let $(y_1, \ldots, y_n)$ be a sample from the uniform $(0, \theta)$. Show $2\bar{y}$ is an unbiased estimate of θ . Determine the UMV unbiased estimate of θ . How do the variances compare for n large?

54. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ, σ^2) with (μ, σ^2) in $\Omega = \mathbb{R} \times \mathbb{R}^+$. Show that $s_y(n-1)^{1/2}\Gamma((n-1)/2)\big/2^{1/2}\Gamma(n/2) \doteq s_y(1+1/4(n-1))$ is the UMV unbiased estimate of σ .

55. Measurements are made to determine a period of oscillation

β . Let y_0 be measured time at which the system is in a certain phase, and let $y_1, \ldots, y_n$ be the measured times of the next n recurrences to that phase. If the measurement error is normally distributed without bias and with common variance then the linear model $\underset{\sim}{y} = \alpha\underset{\sim}{1} + \beta\underset{\sim}{x} + \sigma\underset{\sim}{e}$ is appropriate where $\underset{\sim}{e}$ is a sample from the standard normal and $\underset{\sim}{x} = (0, 1, 2, \ldots, n)$. Find the UMV unbiased estimate b of β in a convenient form for computation (n even; n odd).

56. (Continuation). Consider the following estimates

$$b_1 = \frac{y_n - y_0}{n}$$

$$b_2 = \frac{\sum_{n/2+1}^{n} y_i - \sum_{0}^{n/2-1} y_i}{\left(n^2 + 2n\right)/4} \text{ (n even)}, = \frac{\sum_{(n+1)/2}^{n} y_i - \sum_{0}^{(n-1)/2} y_i}{(n+1)^2/4} \text{ (n odd)} .$$

(a) Show that b_1 , b_2 are unbiased estimates of β . Calculate the variances of b , b_1 , b_2 .

(b) Suppose b_1 is available for $n = 300$. Find values of n for which b and b_2 would have the same precision.

57. For the normal linear model $\underset{\sim}{y} = X\underset{\sim}{\beta} + \sigma\underset{\sim}{e}$ where $\underset{\sim}{e}$ is a sample from the standard normal, the least squares estimates $b_1, \ldots, b_r$ are UMV unbiased for $\beta_1, \ldots, \beta_r$. Show that $s^2(\underset{\sim}{y})/(n-r)$ is UMV unbiased: this uses the definition of

$s^2(\underset{\sim}{y})$ as the minimized sum of squares of deviations; see Section 1(c).

58. Continuation of Seven: Problem 53. Consider a sample of n from the multivariate normal $(\underset{\sim}{\mu}, \Sigma)$ with $\underset{\sim}{\mu}$ in $\mathbb{R}^k$ and Σ is the space of positive definite symmetric matrices $\left(\text{an open set of } \mathbb{R}^{k+k(k+1)/2}\right)$; see Seven; Problem 53. Let

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{k1} \\ \vdots & & \vdots \\ y_{n1} & \cdots & y_{nk} \end{pmatrix}$$

where the n rows record the n elements in the sample. Show that

$$(\bar{y}_1,\ldots,\bar{y}_n) = n^{-1}\underset{\sim}{1}'Y, \quad \begin{pmatrix} s_{11} & \cdots & s_{1k} \\ \vdots & & \vdots \\ s_{k1} & \cdots & s_{kk} \end{pmatrix} = (Y - \underset{\sim}{1}\, n^{-1}\underset{\sim}{1}'Y)'(Y - \underset{\sim}{1}\, n^{-1}\underset{\sim}{1}'Y)/(n-1)$$

where $\underset{\sim}{1}$ is the one vector with n coordinate and the s_{ij} are sample covariances (Four: 2(c)). Show that the sample means and covariances are UMV unbiased estimates of the means and Co-variances of the distribution.

4. HYPOTHESIS TESTING

Consider a response y that is normally distributed (θ, σ_0^2) with θ in $\mathbb{R}$. And suppose an investigator is concerned with a hypothesis $H_0 : \theta \leq \theta_0$ suggested by some theoretical or normative considerations; the alternative then is given by the hypothesis $H_1 : \theta > \theta_0$. With a sample $(y_1, \ldots, y_n)$ of n, our earlier discussions would suggest calculating the observed $\bar{y}$ and comparing it with the normal distribution $(\theta_0, \sigma_0^2/n)$; the significance can be assessed reasonably by the probability (θ_0) of values as large or larger than the observed $\bar{y}$.

A common mathematical formulation of this testing problem envisages the making of a decision on the basis of an observed sample $(y_1, \ldots, y_n)$. As an example such a decision or formal test procedure consider: Accept H_0 if $\bar{y} < \mu_0 + 1{\cdot}64\sigma_0/\sqrt{n}$; Reject H_0 (and accept H_1) if $\bar{y} \geq \mu_0 + 1{\cdot}64\sigma_0/\sqrt{n}$.

Some earlier comments on the use of such decision procedures may be found at the end of Section 2 in Chapter Six. Certainly there are industrial and developmental situations where entities are produced repetitively and where batches are sampled and then accepted or rejected on the basis of a test applied to the sample. In such contexts it may often be possible to determine the loss that would result from a certain decision in a certain situation; then for any decision procedure, the risk or mean loss can be calculated and used to assess the decision procedure. In the scientific applications however, there are many potential users for the results from an investigation, and an appropriate decision may well vary from user to user. The proper role of the investigator is to assemble the information and the evidence that is available and to present conclusions that are supported by the evidence. Certainly the occasion may arise where the evidence is overwhelming and a reasonably firm judgment is appropriate. But often the evidence will not be overwhelming, and it should be presented in as accessible a form as possible; the various users would then be able to make judgments appropriate to their situations and where necessary make what ever decisions are warranted by those judgments.

The mathematical formulation of test and decision procedures, however, does involve some attractive mathematics and it can provide some basis for the choice of a function to use for tests of significance. In this section and several succeeding sections we investigate the formulation and methods for some two-decision procedures: Accept the hypothesis; or Reject the hypothesis.

Now consider a response $\underset{\sim}{y}$ with values in a sample space S and with a probability measure P_θ with a parameter θ taking values in Ω . Suppose that an investigator has a hypothesis concerning the system he is investigating. And suppose the hypothesis can be identified in terms of the parameter θ ; specifically that the points in Ω can be checked and those for which the hypothesis holds can be identified as a set say H_0 and the remaining points for which the hypothesis does not hold can be identified as the complementary set H_1. Thus the hypothesis induces a partition of Ω into two sets called the null hypothesis H_0 and the alternative hypothesis H_1. The term null hypothesis derives from a common situation where an experimenter adopts the initial position or hypothesis that a treatment does not affect (is null for) a response (or that two treatments do not differ in their effects on the response). If the experimenter has randomized his treatments to the experimental units and if he has controlled factors that are accessible and has randomized against possible other factors, then the hypothesis identifies a set H_0 such that the *only* alternative to explain apparent affects in the sample is that the treatment (or the difference in treatments) *causes* response effects.

In the abstract minimum we now have a partition of Ω into sets H_0, H_1; see Figure 2.

If an investigator has committed himself to making a decision, Accept H_0, say d_0, or Reject H_0 $(\text{Accept } H_1)$, say d_1, then he must have a decision function δ which maps S into $\{d_0, d_1\}$. To use the decision function δ he inserts his observed response $\underset{\sim}{y}$ and obtains the decision $\delta(\underset{\sim}{y})$. The statistical problem then becomes one of choosing a good, reasonable, or best decision function δ. Alternatively he can think in terms of a *critical region* C in the sample space; the critical region C consists of the points that produce the decision d_1 (the remaining points of course are those that produce the decision d_0). The statistical problem is then one of choosing a good, reasonable, or best critical region C . See Figure 2.

Consider the example at the beginning of this section with $\sigma_0 = .5$ and $n = 25$. The decision function δ recorded there is given by

$$\begin{aligned} \delta(\underset{\sim}{y}) &= d_0 \quad \text{if} \quad \underset{\sim}{y} \in \{\underset{\sim}{y} : \bar{y} < \theta_0 + .164\} \\ &= d_1 \qquad\qquad \in \{\underset{\sim}{y} : \bar{y} \geq \theta + .164\} \end{aligned}$$

or in terms of the image space of the function $\bar{y}$ by

$$\begin{aligned} \delta(\underset{\sim}{y}) &= d_0 \quad \text{if} \quad \bar{y} < \theta_0 + .164 \\ &= d_1 \qquad\qquad \geq \theta_0 + .164 \,. \end{aligned}$$

The corresponding critical or rejection region C is

$$C = \{\bar{y} \geq \theta_0 + .164\} \;;$$

in this form the event can be interpreted as a set on the sample space S or as a set on the image space (re $\bar{y}$) .

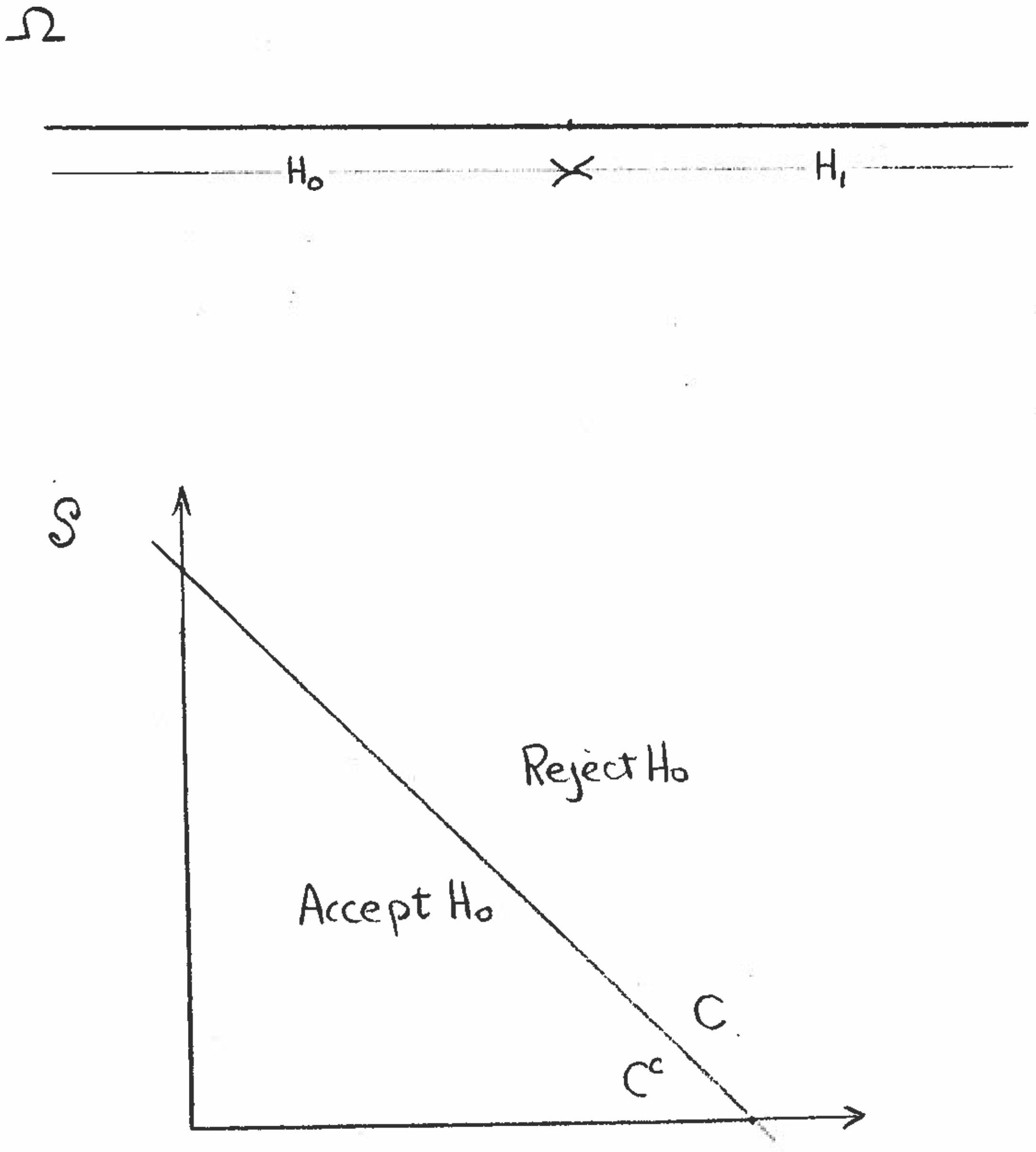

Figure 2: The parameter space Ω as partitioned into hypothesis H_0 and alternative H_1 . The sample space S as partitioned according to a particular decision procedure.

The decision procedure δ just described can be assessed in part on the basis of Figure 3. The space for $\bar{y}$ is divided into the acceptance and rejection regions. And over the space of $\bar{y}$ is pictured the typical distribution for $\bar{y}$; it is located at θ and scaled by $\cdot 1$. The probabilities for acceptance and rejection can be calculated for any value of θ .

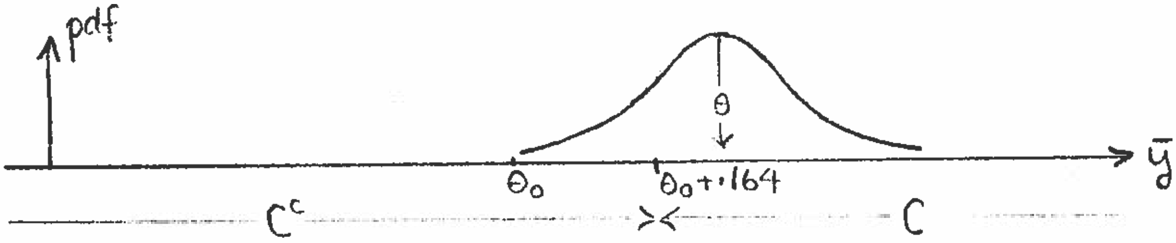

Figure 3: The sample space (in terms of $\bar{y}$) with C and C^c .
The distribution of $\bar{y}$ for a typical θ .

The commitment to using a decision function leaves us in the position of comparing decision functions on the basis of their performance characteristics. For this we define the *power function* of a δ or of a C to be

$$\begin{aligned} P_\delta(\theta) &= P\left(\delta(\underset{\sim}{y}) = d_1 \middle| \theta\right) = P\left(C \middle| \theta\right) \\ &= \mathrm{Prob}\left(\text{Rejecting } H_0 \middle| \theta\right) \\ &(=)\ P\left(\bar{y} \geq \theta_0 + \cdot 164 \middle| \theta\right) \\ &(=)\ P\left[\underset{\sim}{z} \geq 1\cdot 64 + \frac{\theta_0 - \theta}{\cdot 1}\right] \\ &(=)\ 1 - G\left[1\cdot 64 + 10(\theta_0 - \theta)\right] \end{aligned}$$

where the last three expressions refer to the example with G as the standard normal distribution function. Alternatively we define the *operating characteristic function* (the complement of the power function) of a δ or C to be

$$\begin{aligned} OC_\delta(\theta) &= P\left(\delta(\underset{\sim}{y}) = d_0 \middle| \theta\right) = P\left(C^c \middle| \theta\right) \\ &= \mathrm{Prob}\left(\text{Accepting } H_0 \middle| \theta\right) \\ &(=)\ G\left[1\cdot 64 + 10(\theta_0 - \theta)\right] . \end{aligned}$$

The power function for the δ in the example is plotted in Figure 4. Other decision functions would generally have other power functions.

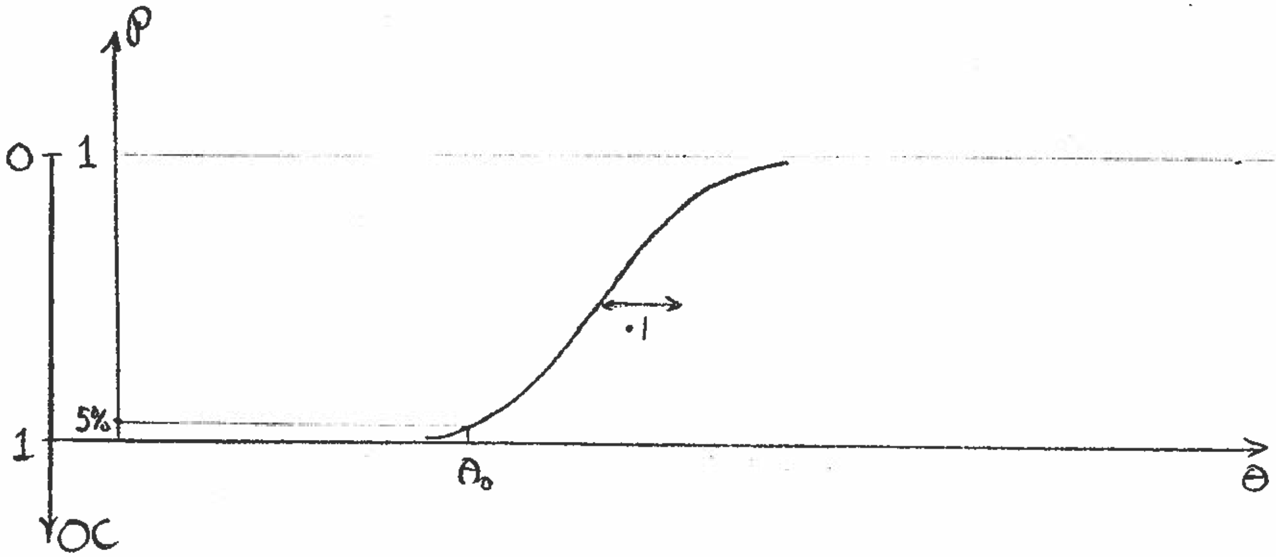

Figure 4: The power function $P(\theta)$ for the test: Reject if $\bar{y} \geq \theta_0 + \cdot 164$. It has the shape of a normal distribution function with standard deviation ·1 .

If H_0 is true, then the decision d_1 is an error; it is called a Type I error. The probability of a Type I error with a decision function δ is

$$\begin{aligned} \alpha_\delta(\theta) &= P_\delta(\theta) && \theta \in H_0 \\ &= 0 && \in H_1 \;. \end{aligned}$$

If H_1 is true, then the decision d_0 is an error; it is called a Type II error. The probability of a Type II error with a decision function δ is

$$\begin{aligned} \beta_\delta(\theta) &= 0 && \theta \in H_0 \\ &= 1 - P_\delta(\theta) && \in H_1 \;. \end{aligned}$$

The error functions α_δ and β_δ are plotted in Figure 5 for the decision function δ of the example.

In certain industrial applications it may be possible to determine the financial loss $\ell(d_i, \theta)$ that would be entailed if the decision d_i is made when the parameter has the value θ . A loss function might have the special form

$$\begin{aligned} \ell(d_0, \theta) &= 0 && \theta \in H_0 \\ &= b(\theta) && \in H_1 \end{aligned}$$

$$\begin{aligned} \ell(d_1, \theta) &= a(\theta) && \theta \in H_0 \\ &= 0 && \in H_1 \end{aligned}$$

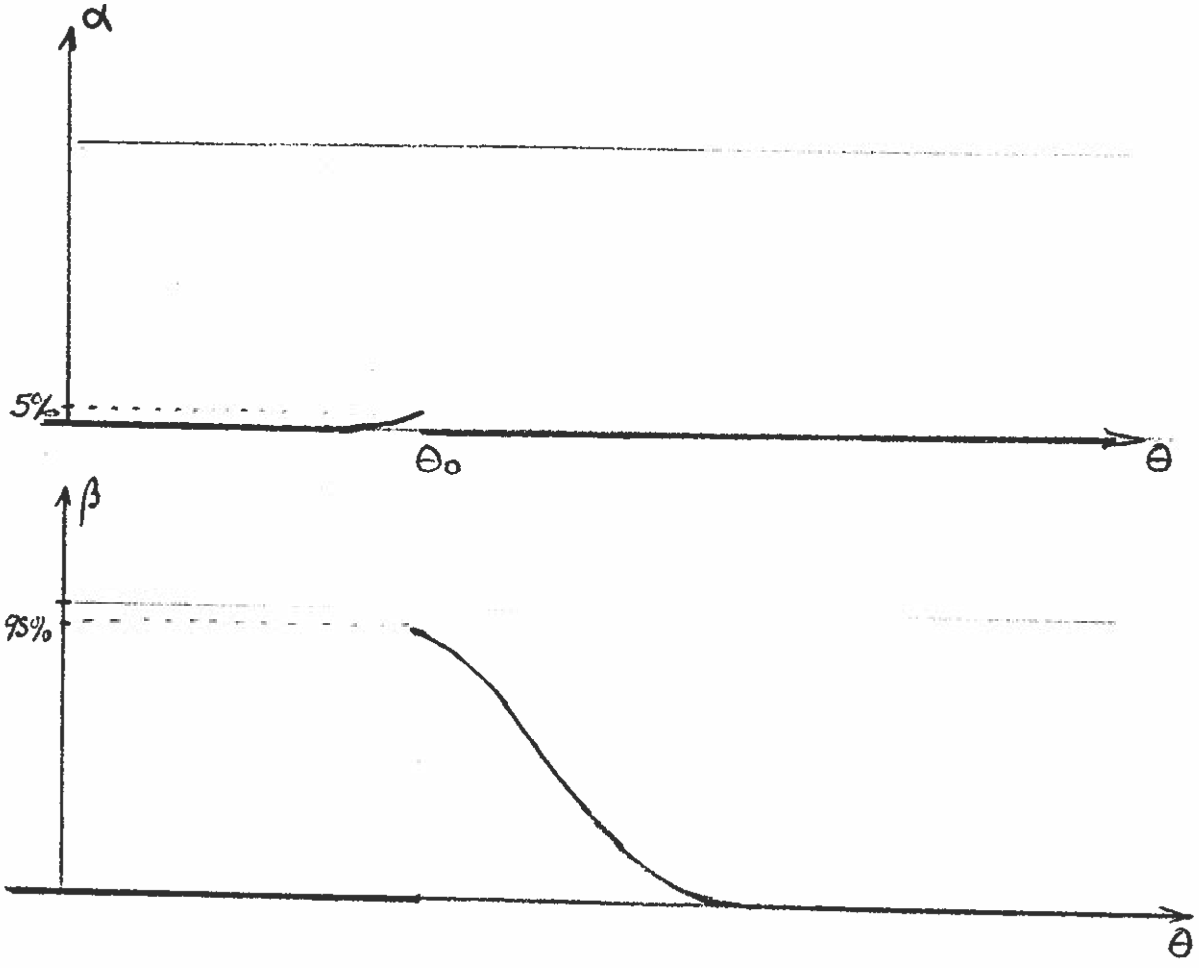

Figure 5: For the particular δ (Reject if $\bar{y} \geq \theta_0 + \cdot 164$) the probability $\alpha_\delta(\theta)$ of a Type I error and the probability $\beta_\delta(\theta)$ of a Type II error.

giving no loss if the decision is a correct decision. The mean value of the loss using a δ can then be calculated; it is called the *risk function*

$$\begin{aligned} R_\delta(\theta) &= E\left[\ell(\delta(\underset{\sim}{y}),\ \theta)\,|\,\theta\right] \\ &= \ell(d_0,\ \theta)\left(1 - P_\delta(\theta)\right) + \ell(d_1,\ \theta)P_\delta(\theta) \\ &(=)\ b(\theta)\beta_\delta(\theta) + a(\theta)\alpha_\delta(\theta) \end{aligned}$$

where the final expression refers to the special form for the loss function. We would now choose a decision function so that the risk is small - ideally so that the risk is small for each θ , or as a compromise so that the maximum risk is small - or small for certain θ values of special concern.

Now consider the choice of a decision function δ (or critical region C) on the basis of performance characteristics. The usual approach involves restricting our attention to those decision functions that have a certain bound α (say 5% or 1%) on the probability of a Type I error; the rationale being that deciding a treatment effect exists when in fact it doesn't is serious and should be guarded against. Thus we restrict our attention now to those tests that satisfy

$$(1) \qquad P(C|\theta) \le \alpha \qquad \theta \text{ in } H_0 .$$

A test C that satisfies (i) is called a *size* α *test.* The test in the normal example was a 5% test (also say a 10% test but a poor 10% test).

Then among tests of size α we look for one for which the power

$$(2) \qquad P(C|\theta) = \text{large} \qquad \theta \in H_1 \ .$$

Of course the test that maximizes the power $P(C|\theta)$ for one θ may not be the test that maximizes it for another. Some sort of a compromise is then needed. See Figure 6.

A hypothesis is called *simple* if it consists of one parameter point and *composite* if it consists of more than one parameter point. For the case of a simple hypothesis against a simple alternative the problem of choosing the size α test that has maximum power is resolved by the following neyman-pearson lemma.

For the statistical model $f(y|\theta)$ with θ in $\Omega = \{\theta_0, \theta_1\}$, consider the problem of testing the hypothesis $H_0 = \{\theta_0\}$ against the alternative $H_1 = \{\theta_1\}$:

Lemma. *A test having*

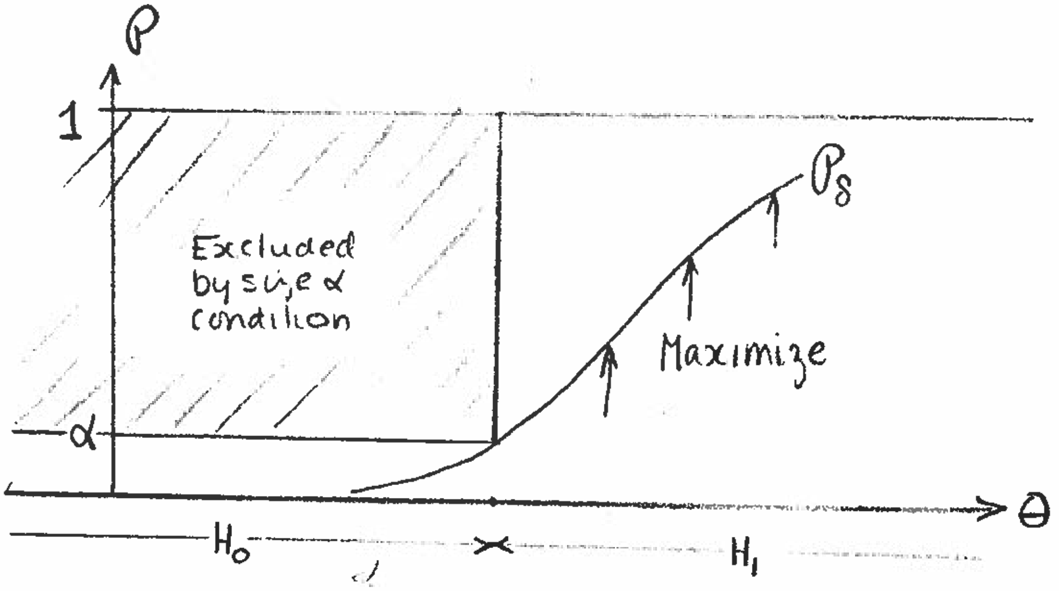

Figure 6: The size α tests δ have a power function P_δ that does not intersect the excluded region. Maximizing power is suggested by the arrows above the set H_1 .

(iii) $$\frac{L(y|\theta_1)}{L(y|\theta_0)} \geq k \qquad y \in C$$

$$\leq k \qquad \in S - C$$

is a most powerful test at some size α *for testing* H_0 *against* H_1 . *If the value of* k *and then* C *can be chosen to satisfy* (*exact* size α)

(iv) $$\int_C f(y|\theta_0)dy = \alpha ,$$

then C *is most powerful among tests having size* α .

Proof: The test is to reject for large value of the likelihood ratio

$$\frac{L(y|\theta_1)}{L(y|\theta_0)} = \frac{f(y|\theta_1)}{f(y|\theta_0)} = L_3(y) .$$

The critical value (beyond which rejection occurs) is determined so that the test has exact size α .

The problem is one of finding a set C containing α of the $f(y|\theta_0)$ probability and yet as much as possible of the $f(y|\theta_1)$ probability; points with large values of $f(y|\theta_1)/f(y|\theta_0)$ should then go in C first.

Consider a set C satisfying (iii) and (iv) for some α. Let D be any other set having size α; then

$$\int_D f(y|\theta_0)dy \leq \alpha = \int_C f(y|\theta_0)dy .$$

Let $C_0 = C \cap D$ and $D^* = D - C_0$, $C^* = C - C_0$. Subtracting the integral over the intersection set C_0 gives

$$\int_{D^*} f(y|\theta_0)dy \leq \int_{C^*} f(y|\theta_0)dy .$$

On D^* which is outside C we have

$$\frac{f(y|\theta_1)}{k} \leq f(y|\theta_0)$$

and on C^* which is inside C we have

$$f(y|\theta_0) \leq \frac{f(y|\theta_1)}{k} .$$

Hence

$$\int_{D^*} \frac{f(y|\theta_1)}{k} dy \leq \int_{D^*} \frac{f(y|\theta_1)}{k} dy .$$

Then cancelling the k and adding on the integral over C_0

gives

$$\int_D f(y|\theta_1)dy \le \int_C f(y|\theta_1)dy$$

which states that the power of D is less than or equal to the power of C. Hence C has maximum power among tests of size α.

Example 3 continued. Consider a sample from the normal distribution (θ, σ_0^2) with θ in $\mathbb{R}$. To illustrate the lemma we take $\Omega = \{\theta_0, \theta_1\}$ and consider testing $H_0 : \theta_0$ as $H_1 = \theta_1$ where $\theta_1 > \theta_0$. By the lemma the most powerful test is to reject the hypothesis H_0 for large values of $L(\theta|y_1)/L(\theta|y_0) = L_3(y)$, or equivalently to reject for large values of

$$\begin{aligned}\ell(\theta|y_1) - \ell(\theta|y_0) &= \ell_3(y)\\ &= -\frac{1}{2\sigma_0^2}\Sigma(y_i - \theta_1)^2 + \frac{1}{2\sigma_0^2}\Sigma(y_i - \theta_0)^2\\ &= \frac{\theta_1 - \theta_0}{2\sigma_0^2}\Sigma\left\{y_i - \frac{\mu_1 + \mu_0}{2}\right\},\end{aligned}$$

or equivalently to reject for large values of the likelihood statistic Σy_i or $\bar{y}$.

The most powerful test at exact size α can be then determined by calculating the critical value exceeded with probability α under the θ_0 distribution. For $\bar{y}$ the critical value is $\theta_0 + z_\alpha \sigma_0 / \sqrt{n}$ where z_α is the point exceeded with probability α for the standard normal. Thus the critical region C is

$$C = \{\underset{\sim}{y} : \bar{y} \geq \theta_0 + z_\alpha \sigma_0 / \sqrt{n}\} .$$

Note that the test does not depend on μ_1 and hence is most powerful for all θ values in the alternative $H_1 : \theta > \theta_0$.

The power of the test is

$$\begin{aligned} &P\left(\bar{y} \geq \theta_0 + z_\alpha \sigma_0 / \sqrt{n} \,|\, \theta_1\right) \\ &= P\left(z \geq z_\alpha + \frac{\theta_0 - \theta_1}{\sigma_0 / \sqrt{n}}\right) \\ &= 1 - G\left(z_\alpha - \frac{\theta_1 - \theta_0}{\sigma_0 / \sqrt{n}}\right) \end{aligned}$$

where G is the standard normal distribution function; see Figure 4.

In applying the lemma to discrete distributions we can reasonably expect difficulties in satisfying the exact size α condition (iv). For suppose we are to reject for large values of a discrete variable y and that under H_0 the discrete variable has say a poisson distribution with mean 1 . Then under H_0 $P(y \geq 4) = \cdot 019$ and $P(y \geq 3) = \cdot 080$. if we wanted an $\alpha = 5\%$ test then we would be wanting to split the probability at the point $y = 3$. In effect this can be accomplished by using a *randomized* test function:

ID *A real valued function* ϕ *on the sample space* S *is a test function if* $0 \leq \phi(y) \leq 1$ *for all* y *in* S *.*

The test function for a nonrandomized test with critical region C is the indicator function for that region C . More generally a test function ϕ is used as follows: if $\phi(y) = 1$ then the hypothesis is rejected; if $\phi(y) = 0$, then the hypothesis is accepted; if $\phi(y) = a$ then dice or random numbers are used and the hypothesis is rejected with probability a and accepted with probability $1 - a$.

For test functions, the lemma can now be presented in the following form.

Lemma. *For testing* H_0 *against* H_1 *the test*

$$
\begin{array}{llr}
\phi(y) & = 1 & \dfrac{L(y|\theta_1)}{L(y|\theta_0)} > k \\[1em]
 & = a & = k \\[1em]
 & = 0 & < k
\end{array}
$$

where a *and* k *are chosen to satisfy* (exact size α)

(v) $$\int \phi(y) f(y|\theta_0) dy = \alpha$$

is most powerful among tests having size α .

Proof: The earlier proof can be modified to cover the corresponding part of the present lemma: for any alternative test ψ at size α consider

$$P_\phi(\theta_0) - P_\psi(\theta_0)$$

$$= \int (\phi(y) - \psi(y)) f(y|\theta_0) dy$$

$$\geq \int (\phi(y) - \psi(y)) k f(y|\theta_1) dy$$

and use the likelihood ratio inequalities together with the sign of $\phi(y) - \psi(y)$.

The essential part of the present lemma is to show that a , k *can* be chosen to satisfy the exact size α condition (v). For this the poisson example preceding the lemma suggests the pattern:

$$\begin{aligned} \phi(y) &= 1 && y \geq 4 \\ &= \frac{.050 - .019}{.080 - .019} && y = 3 \\ &= 0 && y \leq 2 ; \end{aligned}$$

this appropriately splits the probability at $y = 3$ and gives $E\phi(y) = .05$ for the poisson distribution with mean equal to 1 . For the general case let H be the distribution function (see Figure 7) for the likelihood ratio

$$L_3(y) = \frac{L(y|\theta_1)}{L(y|\theta_0)}$$

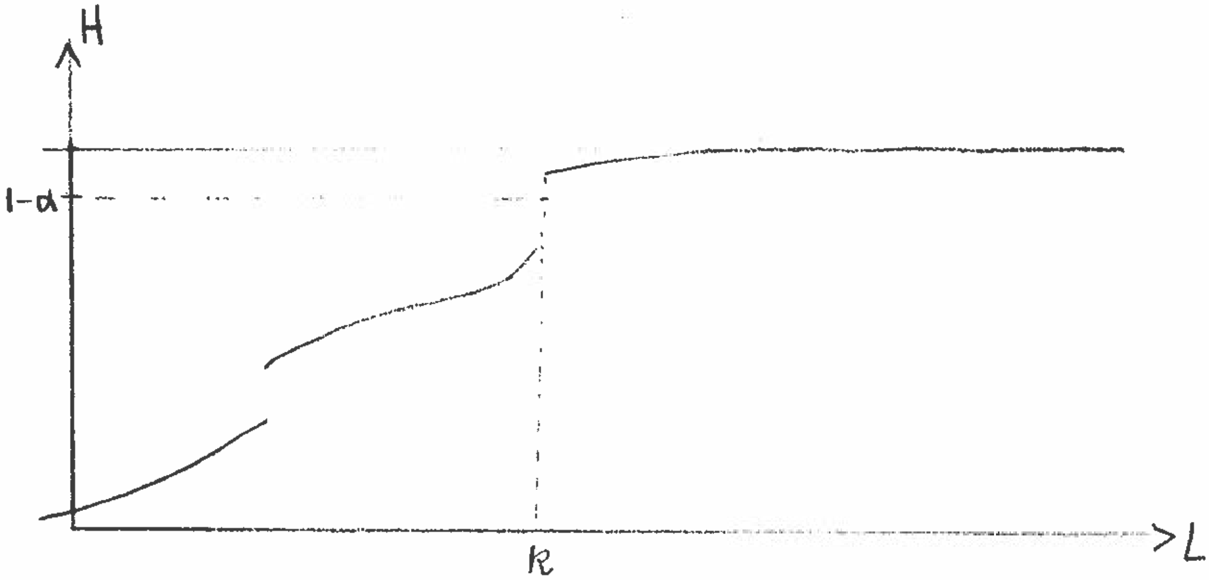

Figure 7: The distribution function H of the likelihood ratio under the (θ_0) distribution on S .

as derived from the θ_0 distribution on the sample space (note that the probability of a zero denominator is zero under the θ_0 distribution). Let k be the value such that $H(k) \geq 1-\alpha$ and $H(k-0) \leq 1-\alpha$. Then the required test function is

$$\begin{aligned} \phi(y) &= 1 & L_3(y) &> k \\ &= \frac{H(k) - (1-\alpha)}{H(k) - H(k-0)} & &= k \\ &= 0 & &< k . \end{aligned}$$

Problems

59. An investigator knows that a response is approximately normal with variance 1·44 . Previously the mean had been 75 but a new treatment may have increased it. For testing the hypothesis that the mean is 75 against the alternative that it is larger, determine the most powerful (1%) test for a sample size of 10 . Plot the power function. If the experimenter wants to be at least 95% certain of detecting a mean equal to 76 , find the minimum sample size.

60. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ_0, σ^2) .

(a) Find the most powerful size α test for $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1 < \sigma_0$.

(b) Does the test depend on σ_1 ?

(c) Give an expression for the power function of the test in terms of the distribution function of a chi-square variable.

61. Continuation. Find the most powerful size α test for $H_0 : \sigma = \sigma_0$ against $H_1 : \sigma = \sigma_1 > \sigma_0$. Does the test depend on σ_1?

62. Let $(x_1, \ldots, x_n)$ be a sample from the bernoulli distribution (p) . Find the form of a most powerful size α test for

(a) $H_0 : p = p_0$ against $H_1 : p = p_1 > p_0$. Does the test depend on p_1 ?

(b) $H_1 : p = p_0$ against $H_1 : p = p_1 < p_0$. Does the test depend on p_1 ?

63. Let y have the poisson distribution (θ) . Find the form of a most powerful size α test of

(a) $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 > \theta_0$. Does the test depend on θ_1 ?

(b) $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 < \theta_0$. Does the test depend on θ_1 ?

64. Let $(y_1, \ldots, y_n)$ be a sample from the exponential distribution $f(y|\theta) = \theta \exp\{-\theta y\}$ on $\mathbb{R}^+$ (say, the interarrival times of THREE Section 6. Find the form of a most

powerful test of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 > \theta_0$ (the critical value can be presented in terms of a percentage point of a familiar distribution!). Does the test depend on θ_1 ?

65. Let $(y_1, \ldots, y_n)$ be a sample from the uniform distribution $(0, \theta)$. Find the form of a most powerful size α test for

(a) $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 > \theta_0$. Does the test depend on θ_1 ?

(b) $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1 < \theta_0$. Does the test depend on θ_1 ?

66. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ, σ^2) . For testing the hypothesis $H_0 : \mu = \mu_0$ against $H_1 : \mu > \mu_0$ a reasonable test of size α is to reject if $t = \sqrt{n}(\bar{y} - \mu_0)/s_y > t_\alpha$ where t_α is the point exceeded with probability α by a t-variable on $n-1$ degrees of freedom. The power function of this test is based on the general distribution of t derived from the (μ, σ^2) distribution on $\mathbb{R}^n$. Show that the probability differential for this noncentral t distribution is

$$\sqrt{\pi}\ \exp\left\{-\frac{\delta^2}{2}\right\} \sum_{r=0}^{\infty} 2^{r/2}\,\frac{\delta^r}{r!}\,\frac{\Gamma\left(\frac{f+1+r}{2}\right)}{\Gamma\left(\frac{f}{2}\right)}\,\frac{(t/\sqrt{f})^{r/2}}{(1+t^2/f)^{(f+1+r)/2}}\,\frac{dt}{\sqrt{f}}$$

where $f = n - 1$ is the degrees of freedom and $\delta = \sqrt{n}(\mu_1 - \mu_0)/\sigma$ is the noncentrality parameter.

Method: (a) Express $t/\sqrt{f}$ as w/χ where w is normal $(\delta, 1)$ and χ is independent with a chi distribution on f degrees of freedom. (b) For the density of w expand the quadratic exponential putting the cross term into an exponential series. (c) Then use the method of TWO Problem 35 and term by term integration.

67. Let ψ be a most powerful size α test of $f(y|\theta_0)$ against $f(y|\theta_1)$. Then for some k

$$\begin{aligned} \psi(y) &= 1 & \qquad \frac{f(y|\theta_1)}{f(y|\theta_0)} &> k \\ &= 0 & &< k \end{aligned}$$

almost everywhere with respect to the lebesgue or the counting measure that carries the density. Thus all most powerful tests have essential likelihood ratio form and differ only in how the size is established for points where the ratio is equal to k .

Method: Let ϕ be as given in the lemma. From size and power show that

$$\int(\phi(y) - \psi(y))\big(f(y|\theta_1) - kf(y|\theta_0)dy \leq 0 \ ;$$

and from properties of ϕ show that the integrand ≥ 0 .

68. Continuation. If $E(\psi(y)|\theta_0) < \alpha$, then $k = 0$ and the maximum power at θ_1 is unity.

69. Generalization of the hypothesis testing lemma. Let f_0 and f_1 be real valued integrable functions relative to a measure μ . Among test functions ϕ for which

$$* \qquad \int \phi(y) f_0(y) dy = c$$

a necessary and sufficient condition that ϕ maximize

$$\int \phi(y) f_1(y) dy$$

is that ϕ satisfy $*$ and have the form

$$\begin{aligned} \phi(y) &= 1 \qquad & f_1(y) &> kf_0(y) \\ &= 0 & &< kf_0(y) \end{aligned}$$

for some k .

5. UNIFORMLY MOST POWERFUL TESTS

The most powerful test of a simple hypothesis against a simple alternative is based on the likelihood function -- reject if the likelihood function has a θ_1 to θ_0 ratio that is large. It is of course reasonable for a good test to be based on the likelihood statistic; see Seven, Sections 1,3. Now with a general parameter space Ω we can in fact show that any test function ϕ has a corresponding test function ψ with the same power where ψ is based on the likelihood statistic s . The power function of a test ϕ is

$$\begin{aligned} P_\phi(\theta) &= P(\text{"Reject"}|\theta) \\ &= E(\phi(y)|\theta) \quad . \end{aligned}$$

Let

$$\psi(s(y)) = E(\phi(y):s(y))$$

be the mean value of ϕ from the conditional distribution (independent of θ) given the likelihood statistic. Then

$$E(\psi(s(y))|\theta) = E(\phi(y)|\theta)$$

and hence

$$P_\phi(\theta) = P_\psi(\theta)$$

for all θ in Ω . Thus averaging over the conditional distribution given the likelihood statistic produces a test function dependent on the likelihood statistic and it does not alter the power function. In terms of performance characteristics the two tests are equivalent. And thus *in anu hypothesis testing problem we can clearlu restrict our attention to the test functions that are based on the likelihood statistic.*

(a) Simple hypothesis, composite alternative

Now consider a hypothesis testing problem with a simple hypothesis $H_0:\Theta_0$ and a composite alternative H_1. Among *tests of size* α

(i) $$E(\phi(y)|\Theta_0) \leq \alpha$$

we look for one with *maximum power*

(ii) $$E(\phi(y)|\Theta)$$

for Θ in the alternative H_1. Of course the test with maximum power at Θ_1 may not be the test with maximum power at another value Θ_1'. Some sort of compromise would be needed -- perhaps choosing the test with maximum power at some important or distinct value in the alternative. For certain special problems however we are successful and we do obtain a *uniformlu most powerful* (UMP) test. Consider an example:

Example 3 continued: Consider a sample $(y_1, \ldots, y_n)$ from the normal (Θ,σ_0^2) and suppose we are interested in

$$H_0:\Theta_0$$

$$H_1:\Theta > \Theta_0 \quad .$$

By the example in Section 4 the test of size α having maximum power at Θ_1 is

$$\begin{aligned}\phi(\underset{\sim}{y}) &= 1 \quad \text{if } \bar{y} \geq \Theta_0 + z_\alpha\sigma_0/\sqrt{n}\\ &= 0 \quad\quad\; < \Theta_0 + z_\alpha\sigma_0/\sqrt{n} \;.\end{aligned}$$

But this test does not depend on Θ_1 ; it has maximum power for each $\Theta_1 > \Theta_0$ and hence is a UMP size α test.

It is natural to ponder what kind of problems will yield a UMP test. The lemma in Section 4 gives the test having maximum power at a particular alternative value; it is to reject for large values of the likelihood ratio. We will obtain a UMP test if large values of the likelihood ratio

$$\frac{L(y|\Theta_1)}{L(y|\Theta_0)}$$

are the same for the various possible Θ_1 values. Towards this we define a monotone likelihood ratio model:

ID *The model* $f(y|\Theta)$ *with* Θ *in* $\Omega = \mathbb{R}$ *is a monotone likelihood ratio* (MLR) *model if there is a real valued function* $t(y)$ *such that*

$$\frac{f(y|\Theta_2)}{f(y|\Theta_1)}$$

is a nondecreasing function of $t(y)$ *for each* $\Theta_1 < \Theta_2$.

In particular it follows that the likelihood statistic can be expressed as a function of $t(y)$.

As a prime example consider the exponential model

$$f(y|\Theta) = \gamma(\Theta)\exp\{t(y)\psi(\Theta)\}h(y)$$

where ψ is a nondecreasing function. As special examples we have the location parameter normal, the binomial, and the poisson.

For a monotone likelihood ratio model with hypothesis testing problem

$$H_0 : \Theta_0$$

$$H_1 : \Theta > \Theta_0$$

we then have (proof as in the example) the following form of the hypothesis testing lemma.

Lemma. *The test*

$$\begin{aligned} \phi(y) &= 1 & \text{if} \quad t(y) &> k \\ &= a & &= k \\ &= 0 & &< k \end{aligned}$$

where k,a *satisfy* (exact size α)

$$E(\phi(y)|\Theta_0) = \alpha$$

is a uniformly most powerful test at level α .

(b) Composite hypothesis

Now consider a hypothesis testing problem with a composite hypothesis H_0 and suppose we want the size α test

(v) $$E(\phi(y)|\Theta) \leq \alpha \qquad \Theta \in H_0$$.

haivng maximum power for the alternative Θ_1 . Typically we would be interested in a composite alternative but in line with our general approach we first examine a simple alternative and in effect use a parameter space $\Omega = H_0 \cup \{\Theta_1\}$

The mathematical problems of closing in on a test that satisfies each of the inequalities

(iii) $$E(\phi(y)|\Theta) \leq \alpha \qquad \Theta \text{ in } H_0$$

are rather severe. Sometimes we can find a value Θ_0 in H_0 with density $f(y|\Theta_0)$ that is toughest to distinguish from the alternative $f(y|\Theta_1)$. In that case we would expect that the best test ϕ of Θ_0 against Θ_1 would also have

$$E(\phi(y)|\Theta) \leq \alpha$$

for the other Θ's that are more easily distinguished from Θ_1 . In effect we examine the tests (a larger class) that satisfy the single inequality

(iv) $$E(\phi(y)|\Theta_0) \leq \alpha \quad ,$$

and find the test having maximum power

$$E(\phi(y)|\Theta_1) \;;$$

if the resulting test is one satisfying all the inequalities (iii) then clearly it is the best test in the smaller class satisfying the inequalities (iii).

In other cases we can find a probability average of the densities of H_0 that is toughest to distinguish from the alternative $f(y|\Theta_1)$. We look for a probability measure λ on H_0 such that

$$f_\lambda(y) = \int_{H_0} f(y|\Theta)d\lambda(\Theta)$$

is toughest to distinguish from $f(y|\Theta_1)$. Any test satisfying all the inequalities (iii) will satisfy

(v) $$E(\phi(y)|\lambda) \leq \alpha$$

since

$$\begin{aligned} E(\phi(y)|\lambda) &= \int \phi(y) f_\lambda(v)dv \\ &= \int_S \phi(y) \int_{H_0} f(y|\Theta)d\lambda(\Theta)dv \\ &= \int_{H_0} \left[\int_S \phi(y)f(y|\Theta)dv\right]d\lambda(\Theta) \\ &\leq \alpha \end{aligned}$$

using fubini in Four Section 1(10).

Our method then is to examine the larger class of tests that satisfy (v) and find the test having maximum power at Θ_1 ; if the resulting test is one in the smaller class of tests satisfying all the inequalities (iii) then clearly it is the best test in that smaller class satisfying (iii).

Example 3 continued. Consider a sample $(y_1,\ldots,y_n)$ from the normal (Θ,σ_0^2) and suppose we are interested in

$$H_0:\Theta \leq \Theta_0$$

$$H_1:\Theta > \Theta_0 \quad .$$

Consider a value Θ_1 in H_1 and suppose we look for a value in H_0 that is "closest" to H_1; this suggest Θ_0 . The most powerful size α test of Θ_0 against Θ_1 is

$$\begin{aligned} \phi(\underset{\sim}{y}) &= 1 \qquad \text{if } \bar{y} \geq \Theta_0 + z_\alpha\sigma_0/\sqrt{n} \\ &= 0 \qquad\qquad\;\; < \Theta_0 + z_\alpha\sigma_0/\sqrt{n} \quad . \end{aligned}$$

Is this test in the smaller class of size α tests for H_0 ? The power function of ϕ is

$$\int_{\mathbb{R}^n} \phi(\underset{\sim}{y})f(\underset{\sim}{y}|\Theta)d\underset{\sim}{y} = P(\bar{y} \geq \Theta_0 + z_\alpha\sigma_0/\sqrt{n}|\Theta)$$

$$= P\left(\frac{\bar{y}-\Theta}{\sigma_0/\sqrt{n}} \geq \frac{\Theta_0-\Theta}{\sigma_0/\sqrt{n}} + z_\alpha\right)$$

and with $\Theta < \Theta_0$ the power is $\leq \alpha$. The test *is* in the smaller class and hence is the most powerful test of H_0 against Θ_1. But the test does not depend on Θ_1 ; hence it is a UMP size α test for

$$H_0 : \Theta \leq \Theta_0$$
$$H_1 : \Theta > \Theta_0 \quad .$$

Now consider a monotone likelihood ratio model (with real valued function $t(y)$) and suppose we are interested in the hypothesis testing problem

$$H_0 : \Theta \leq \Theta_0$$
$$H_1 : \Theta > \Theta_0 \quad .$$

We then have the following form of the hypothesis testing lemma:

Lemma. *The test*

$$\begin{aligned} \phi(y) &= 1 & \text{if} \quad t(y) &> k \\ &= 0 & &= k \\ &= 0 & &< k \end{aligned}$$

where k,a *satisfy*

$$E(\phi(y) | \Theta_0) = \alpha$$

is a uniformly most powerful size α *test of* H_0 *against* H_1 .

Proof: The earlier form of the lemma gives the UMP size test ϕ of Θ_0 against Θ_1. Is the test in the smaller class of tests of size α for H_0 ? Consider the power at some value $\Theta' < \Theta_0$:

$$\alpha(\Theta') = E(\phi(y)|\Theta') \quad .$$

By the earlier form of the lemma the test ϕ is most powerful at level $\alpha(\Theta')$ for testing Θ' against Θ_0. But the test $\phi^*(y) = \alpha(\Theta')$, which stupidly rejects with probability $\alpha(\Theta')$ regardless of the data, is an $\alpha(\Theta')$ test of Θ' against Θ_0; hence $\alpha(\Theta') \leq \alpha$. The test ϕ is then of size α for H_0 and hence is most powerful for testing H_0 against Θ_1 . But the test doesn't depend on Θ_1 . Hence it is UMP size α for H_0 against H_1 . Note that the proof uses monotone likelihood ratio only for parameter pairs where one of the values is Θ_0 .

(c) Locally most powerful tests

Consider a hypothesis testing problem $H_0:\Theta_0$ against $H_1:\Theta > \Theta_0$ and suppose there does not exist a uniformly most powerful test of size α. One possibility mentioned earlier is to seek an important or distinctive value in the alternative and choose the most powerful test of Θ_0 against Θ_1 :

$$\phi(y) = 1 \qquad \text{if} \quad \frac{f(y|\Theta_1)}{f(y|\Theta_0)} > k$$
$$= a \qquad\qquad = k$$
$$= 0 \qquad\qquad < k$$

where k and a are chosen to give exact size α under Θ_0. With such a Θ_1 value the hope would be that the test had reasonable power for other Θ values in H_1.

Another possibility is to determine the test that has maximum power for Θ close to Θ_0, that is for the Θ values that typically are difficult to distinguish from Θ_0. For this suppose that the log-likelihood is continuously differentiable for Θ near Θ_0 and that Assumption (i) in Seven, Section 9 holds; then

$$\ln \frac{L(y:\Theta_0+\delta)}{L(y:\Theta_0)} = S(y|\Theta_0)\delta + 0(\delta^2)$$

and for any test ϕ

$$\frac{d}{d\Theta} E(\phi(y)|\Theta)\Big|_{\Theta=\Theta_0} = \int \phi(y)S(y|\Theta_0)f(y|\Theta_0)dy$$

where $S(y|\Theta)$ is the score function from Seven, Section 1. Thus the size α test that maximizes the slope of the power function at Θ_0 is given by

$$\phi(y) = 1 \qquad \text{if} \quad S(y|\Theta_0) > k$$
$$= a \qquad\qquad = k$$
$$= 0 \qquad\qquad < k$$

where k and a are chosen to give the test the exact size α under Θ_0 . With such a test the hope would be that good power for Θ close to Θ_0 would mean reasonable power for other presumably more easily detected alternative values.

Example 3 continued. Consider a sample $(y_1,\ldots,y_n)$ from the normal (Θ,σ_0^2) and suppose we are interested in $H_0:\Theta_0$ against $H_1:\Theta > \Theta_0$. The locally most powerful test is to reject for large values of

$$\mathbf{S}(\underset{\sim}{y}|\Theta_0) = \frac{\bar{y}-\Theta_0}{\sigma_0^2/n}$$

or equivalently for larger values of $\bar{y}$. Then, as we would expect, this gives the UMP test described earlier in this section.

(d) Large sample methods

For testing Θ_0 against Θ_1 the likelihood ratio is typically easy to calculate. The critical value of k however requires the Θ_0 distribution of the likelihood ratio and may well be difficult to calculate.

And similarly for testing Θ_0 against $\Theta_0 + \delta$ (small positive δ) the score function $S(y|\Theta_0)$ is typically easy to calculate. The critical value however requires the Θ_0 distribution of $S(y|\Theta_0)$ and may well be difficult to calculate. Fortunately the large sample distributions are commonly available.

Let $(y_1,\ldots,y_n)$ be a sample from $f(y|\Theta)$. And suppose the mean μ_0 and variance $\sigma_0^2(>\ )$ of

$$\ln f(y|\Theta_1) - \ln f(y|\Theta_0)$$

exist under the Θ_0 distribution:

$$\mu_0 = E(\ln f(y|\Theta_1) - \ln f(y|\Theta_0)|\Theta_0)$$

$$= \int \ln \frac{f(y|\Theta_1)}{f(y|\Theta_0)} f(y|\Theta_0)dy$$

$$\mu_0^2 + \sigma_0^2 = E((\ln f(y|\Theta_1) - \ln f(y|\Theta_0))^2|\Theta_0)$$

$$= \int \ln^2 \frac{f(y|\Theta_1)}{f(y|\Theta_0)} f(y|\Theta_0)dy \quad .$$

Then by the central limit theorem (Five, Section 2), the likelihood ratio

$$\ln \frac{f(\underset{\sim}{y}|\Theta_1)}{f(\underset{\sim}{y}|\Theta_0)} = \sum_1^n \ln \frac{f(y_i|\Theta_1)}{f(y_i|\Theta_0)}$$

has a (Θ_0) distribution with limiting normal form located at $n\mu_1$ and scale by $\sqrt{n}\,\sigma_1$. The approximate size α test of Θ_0 against Θ_1 is then

$$\begin{aligned} \phi(\underset{\sim}{y}) &= 1 \qquad \text{if} \quad \ln \frac{f(\underset{\sim}{y}|\Theta_1)}{f(\underset{\sim}{y}|\Theta_0)} \geq n\mu_0 + z_\alpha \sqrt{n}\,\sigma_0 \\ &= 0 \qquad\qquad\qquad\qquad\quad < n\mu_0 + z_\alpha \sqrt{n}\,\sigma_0 \quad . \end{aligned}$$

The approximate power can be obtained from the mean μ_1 and variance σ_1^2 of $\ln f(y|\Theta_1) - \ln f(y|\Theta_0)$ under the Θ_1

distribution

$$P_\phi(\Theta_1) \approx 1 - G\left(z_\alpha \frac{\sigma_0}{\sigma_1} - \frac{\mu_1-\mu_0}{\sigma_1/\sqrt{n}}\right)$$

where G is the standard normal distribution function.

For the locally most powerful test suppose that $f(y|\Theta)$ satisfies the regularity conditions (i) , (ii) of Seven, Section 5. Then

$$E(S(y|\Theta_0)|\Theta_0) = 0$$

$$\mathrm{Var}(S(y|\Theta_0)|\Theta_0) = I(\Theta_0) \quad .$$

And for the sample of n ,

$$E(S(\underset{\sim}{y}|\Theta_0)|\Theta_0) = 0$$

$$\mathrm{Var}(S(\underset{\sim}{y}|\Theta_0)|\Theta_0) = nI(\Theta_0) \quad .$$

And then by the central limit theorem (Five, Section 2) the score function

$$S(\underset{\sim}{y}|\Theta_0)$$

has a (Θ_0) distribution with limiting normal form located at 0 and scaled. by $\sqrt{n}\ I^{\frac{1}{2}}(\Theta_0)$. The approximate size test is then given by

$$\begin{aligned} \phi(\underset{\sim}{y}) &= 1 \qquad \text{if} \quad S(\underset{\sim}{y}|\Theta_0) \geq Z_\alpha\sqrt{n}\ I^{\frac{1}{2}}(\Theta_0) \\ &= 0 \qquad\qquad\qquad\qquad\quad < Z_\alpha\sqrt{n}\ I^{\frac{1}{2}}(\Theta_0) \quad . \end{aligned}$$

Now consider the hypothesis testing problem

$$H_0 \ : \Theta = \Theta_0$$

$$H_1 \ : \Theta \neq \Theta_0 \quad .$$

A natural extension of the likelihood ratio test is to use

$$L(y) = \frac{\sup\limits_{\Theta\in\mathbb{R}} f(y|\Theta)}{f(y|\Theta_0)} = \frac{f(y|\hat{\Theta}(y))}{f(y|\Theta_0)} \quad ,$$

the maximum likelihood relative to the hypothesized value Θ_0 and to reject the hypothesis for large values of $L(y)$. Fortunately the large sample distribution is readily available.

By the assumptions and results in Seven, Section 5 the approximate size α test of Θ_0 is given by

$$\begin{array}{rll} \phi(y) = 1 & \quad 2\ln \dfrac{f(\underset{\sim}{y}|\hat{\Theta}(y))}{f(\underset{\sim}{y}|\Theta_0)} & \geq \chi^2_\alpha \\ = 0 & & < \chi^2_\alpha \end{array}$$

where χ^2_α is the value exceeded with probability α by a chi-square variable on 1 degree of freedom.

For the case on an r dimensional parameter the multivariable results in Seven, Section 5 give a test as in the preceding paragraph but based on a chi-square variable on r degrees of freedom.

Problems

70. Continuation of Problem 60; scale normal. Find the uniformly most powerful (UMP) size α test of $H_0 : \sigma \geq \sigma_0$ against $H_1 : \sigma < \sigma_0$.

71. Continuation of Problem 61; scale normal. Find the UMP size α test of $H_0 : \sigma \leq \sigma_0$ against $H_1 : \sigma > \sigma_0$.

72. Continuation of Problem 62; the binomial. Find the UMP size α test of

(a) $H_0 : p \leq p_0$ against $H_1 : p > p_0$.

(b) $H_0 ; p \geq p_0$ against $p < p_0$.

73. Continuation of Problem 63; the poisson. Find the UMP size α test of

(a) $H_0 : \Theta \leq \Theta_0$ against $H_1 : \Theta > \Theta_0$.

(b) $H_0 : \Theta \geq \Theta_0$ against $H_1 : \Theta < \Theta_0$.

74. Continuation of Problem 64; the exponential. Find the UMP size α test of

(a) $H_0 : \Theta \leq \Theta_0$ against $H_1 : \Theta > \Theta$.

(b) $H_0 : \Theta \geq \Theta_0$ against $H_1 : \Theta < \Theta_0$.

75. Continuation of Problem 65; the uniform $(0,\Theta)$. Find the UMP size α test for the two-sided problem $H_0 : \Theta = \Theta_0$ against $H_1 : \Theta \neq \Theta_0$; there is such a test in contrast to the preceding problems.

76. Let $(y_1,\ldots,y_n)$ be normal (μ,σ^2) and consider the composite hypothesis $H : \sigma \geq \sigma_0$ against the simple alternative $H_1 : \sigma = \sigma_1 \ (<\sigma_0)$, $\mu = \mu_1$. As hypothesis distribution "closest" to the alternative consider

$\sigma = \sigma_0$, $\mu = \mu_1$.

(a) Show that the most powerful size α test of (μ_1, σ_0) against (μ_1, σ_1) is to reject if $\sum(y_i - \mu_1)^2 \le \sigma_0^2 \chi^2_{1-\alpha}$ where $\chi^2_{1-\alpha}$ is the point exceeded with probability $1 - \alpha$ by a chi-square variable on n degrees of freedom.

(b) Argue that the preceding test has correct size α for the full hypothesis $H_0 : \sigma_0 \ge \sigma_0$, $\mu \in \mathbb{R}$.

(c) Conclude that a UMP size α test does not exist for $H : \sigma \ge \sigma_0$ against $H : \sigma < \sigma_0$.

77. Continuation. Let $n \ge 2$ and consider the composite hypothesis $H : \sigma \le \sigma_0$ against the simple alternative $H_1 : \sigma = \sigma_1$ $(> \sigma_0)$, $\mu = \mu_1$. By the initial arguments in Section 5 it suffices to examine tests based on $\bar{y}$ (normal $(\mu, \sigma^2/n)$) and the independent $\sum(y_i - \bar{y})^2$ (σ^2 chi-square on $n-1$ degrees of freedom). As hypothesis *combination* 'closest' to the alternative consider $\sigma^2 = \sigma_0^2$ and give μ a normal probability distribution located at μ_1 and scaled by $(\sigma_1^2 - \sigma_0^2)/n$. (This describes a probability measure λ)

(a) Show that the most powerful test of $f_\lambda(\underset{\sim}{y})$ against $f(\underset{\sim}{y} | \mu_1, \sigma_1^2)$ is to reject if $\sum(y_i - \bar{y})^2 \ge \sigma_0^2 \chi^2_\alpha$ where χ^2_α is the α point of the chi-square distribution on n degrees of freedom.

(b) Argue that the test is UMP size α for $H_0 : \sigma \le \sigma_0$ against $H_1 : \sigma > \sigma_0$. (Compare with Problem 76 (c)).

78. Let $(y_1,\ldots,y_n)$ be a sample from *some* distribution having a density function relative to Lebesgue on $\mathbb{R}$. Let h designate the median of the density f (if a range of medians take h to be the inf) and consider the $H_0 : h = h_0$ against $H_1 : h > h_1$.

(a) Find the most powerful test of H_0 against the alternative f_1 . Method: Write $f_1 = p_1 f^- + q_1 f^+$ where f^- and f^+ are the conditional densities given $y < h_0$ and given $y > h_0$ and consider the hypothesis density $f_0 = \frac{1}{2}f^- + \frac{1}{2}f^+$.

(b) Show that the sign test is UMP size α for the original problem.

79. Let $(y_1,\ldots,y_n)$ be a sample from the normal (μ,σ^2) and consider the problem $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$. For this consider first $H_0 : \mu \leq 0$ against the simple $H_1 : \mu = \mu_1$ (>0) , $\sigma = \sigma_1$.

(a) Show that the most powerful test of $(0,\sigma_0^2)$ against (μ_1,σ_1^2) with $\sigma_0^2 > \sigma_1^2$ is to reject if

$$\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_0^2}\right) \Sigma y_i^2 - 2\frac{\mu_1}{\sigma_1^2}\Sigma y_i < k \quad .$$

(b) Consider a rejection region of the form $C = \{\underset{\sim}{y} : \Sigma(y_i-a)^2 < ka^2\}$ with $k < 1$. Consider $P(C|(\mu,\sigma^2))$. For $\mu = 0$ argue that there is a positive value of σ for which the probability content of C is maximum. And for σ given argue that $P(C|\mu,\sigma^2)$ decreases as μ decreases (for $\mu \leq 0$) .

(c) For given σ_0, show that σ_1 and μ_1 can be found so that C is the likelihood ratio test in (a) . Hence show that C is the most powerful test at some size α of $H_0 : \mu \leq 0$ against $H_1 : \mu = \mu_1$, $\sigma = \sigma_1$. It follows that there does not exist a UMP size α of $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$ for some α values. Some further analysis shows that the range is $0 < \alpha < \frac{1}{2}$.

80. Let $y_{11}, \ldots, y_{1m})$ be a sample from the normal $(0,\sigma_1^2)$ and $(y_{21}, \ldots, y_{2n})$ be a sample from the normal $(0,\sigma_2^2)$. (The case of known means but responses relocated to zero for convenience of calculations). Find the form of the most powerful test of $H_0 : \sigma_1 = \sigma_2 = \sigma_0$ against $H_1 : \sigma_1 = \sigma_a , \sigma_2 = \sigma_b$ where $\sigma_a > \sigma_b$. Show that σ_0 can be chosen so that the test is to reject for large values of

$$F = \frac{\Sigma y_{1j}^2/m}{\Sigma y_{2j}^2/n} ,$$

that is, for values of F that exceed the α point of the F distribution on m over n degrees of freedom. (Canonical F in ONE Problem 70, and TWO Problems 36, 49. For independent normal variables z_i , the distribution of

$$G = \frac{Z_1^2 + \ldots \; Z_m^2}{Z_{m+1}^2 + \ldots + Z_{m+n}^2} , \quad F = \frac{\left(Z_1^2 + \ldots + Z_m^2\right)/m}{\left(Z_{m+1}^2 + \ldots + Z_{m+n}^2\right)/n} = G\,\frac{1/m}{1/n}$$

is respectively canonical $F\left(p = \frac{m}{2}\ ,\ \ q = \frac{n}{2}\right)$ and ordinary F with m over n degrees of freedom. The probability element for ordinary F is

$$\frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \quad \frac{\left(\frac{m}{n}\ F\right)^{\frac{m}{2}-1}}{\left(1 + \frac{m}{n}\ F\right)^{\frac{m+n}{2}-1}}\ d\ \frac{m}{n}\ F \Bigg) \quad .$$

Conclude that the test: reject if $F > F_\alpha$, is a UMP size test for $H_0 : \sigma_1 = \sigma_2$ against $H_1 : \sigma_1 > \sigma_2$.

81. Continuation. Show that the power function of the preceding test is given by

$$1 - H\left(\frac{\sigma_2^2}{\sigma_1^2}\ F_\alpha\right)$$

where H designates the distribution function for ordinary F on m over n degrees of freedom.

82. Let $(y_{11}\ ,\ \ldots,\ y_{1m})$ be a sample from the normal (μ,σ_0^2) and $(y_{21}\ ,\ \ldots,\ y_{2n})$ be a sample from the normal (μ_2,σ_0^2) with σ_0^2 know.

(a) Find the most powerful size α test of $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 = \mu_a,\ \ \mu_2 = \mu_b$ with $\mu_a > \mu_b$. As tough value under H_0 try $\mu_0 = (m\mu_a + n\mu_b)/(m+n)$. The test is to reject for large values of $\bar{y}_1 - \bar{y}_2$.

(b) Find the UMP size α test of $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 > \mu_2$.

6. UNBIASED TESTS

Some very simple hypothesis testing problems do not have uniformly most powerful tests. Let $(y_1,\ldots,y_n)$ be a sample from the normal (μ,σ^2) and consider the hypothesis testing problem, $H_0:\mu_0\ ,\sigma_0^2$ against $H_1:\mu \neq \mu_0,\sigma_0^2$ (the variance known case). The most powerful test for μ values greater than μ_0 is to reject for large values of $\bar{y}$. Correspondingly the most powerful test for μ values less than μ_0 is to reject for small values of $\bar{y}$. There is no uniformly most powerful test although a reasonable test is to reject for large deviations $|\bar{y} - \mu_0|$.

Or consider the hypothesis testing problem, $H_0:\mu_0,\sigma^2 \in \mathbb{R}^+$ against $H_1:\mu > \mu_0,\sigma^2 \in \mathbb{R}^+$. For values of $\alpha < \frac{1}{2}$, even this one-sided problem does not have a uniformly most powerful test (Problem 79). But certainly a reasonable test from all our earlier considerations is to reject for large values of

$$t = \frac{\bar{y}-\mu_0}{s_y/\sqrt{n}} \quad .$$

Or consider the hypothesis testing problem, $H_0:\mu_0,\sigma^2 \in \mathbb{R}^+$ against $H_1:\mu \neq \mu_0,\sigma^2 \in \mathbb{R}^+$. This differs from the preceding in having an enlarged alternative and again there can be no uniformly most powerful test. Certainly a reasonable test is to reject for large values of $|t|$.

Some very simple hypothesis testing problems for the mathematically very nice normal distribution do not have solutions in terms of the theory we have developed so far. Our approach now is to restrict attention to tests satisfying some attractive property -- unbiasedness in this section and invariance in the next section -- with the hope that a uniformly most powerful test may be found in the restricted class of nice tests.

Consider a hypothesis testing problem H_0 against H_1

ID *A test* ϕ *is an unbiased size* α *test of* H_0 *against* H_1 *if*

(i)
$$\begin{aligned} E(\phi(y)|\Theta) &\leq \alpha \quad \text{if} \quad \Theta \varepsilon H_0 \\ &\geq \alpha \qquad\quad\ \ \varepsilon H_1 \end{aligned}$$

The first part of the condition is that ϕ be of size α. The second part -- the essential part of unbiasedness -- is that the probability of rejection when rejection should occur be greater than the probability allowed when rejection should not occur.

In many standard problems power functions are continuous, even continuously differentiable functions of Θ. Often then we can relax from the unbiased tests to a larger class of tests that satisfy a weaker condition involving *equalities*. If the best test in the larger class happens to be in the smaller class of unbiased tests then we have obtained the best *unbiased* test.

(a) Local Unbiasedness

Consider the hypothesis testing problem $H_0:\Theta = \Theta_0$ against $H_1:\Theta \neq \Theta_0$, and suppose that any power function is continuously differentiable with respect to Θ .

An unbiased size α test ϕ satisfies

$$\begin{aligned} &E(\phi(y)|\Theta_0) \leq \alpha \\ &E(\phi(y)|\Theta) \geq \alpha \quad \Theta \neq \Theta_0 \quad . \end{aligned}$$

By the continuity an equivalent set of conditions is

$$\begin{aligned} &E(\phi(y)|\Theta_0) = \alpha \\ &E(\phi(y)|\Theta) \geq \alpha \quad \Theta \neq \Theta_0 \quad . \end{aligned}$$

By the continuous differentiability a weaker set of conditions is

(ii)
$$\begin{aligned} &E(\phi(y)|\Theta_0) = \alpha \\ &\frac{d}{d\Theta} E(\phi(y)|\Theta)|_{\Theta=\Theta_0} = 0 \quad ; \end{aligned}$$

a test ϕ satisfying these conditions is called a *locally unbiased size* α *test.* See Figure 8.

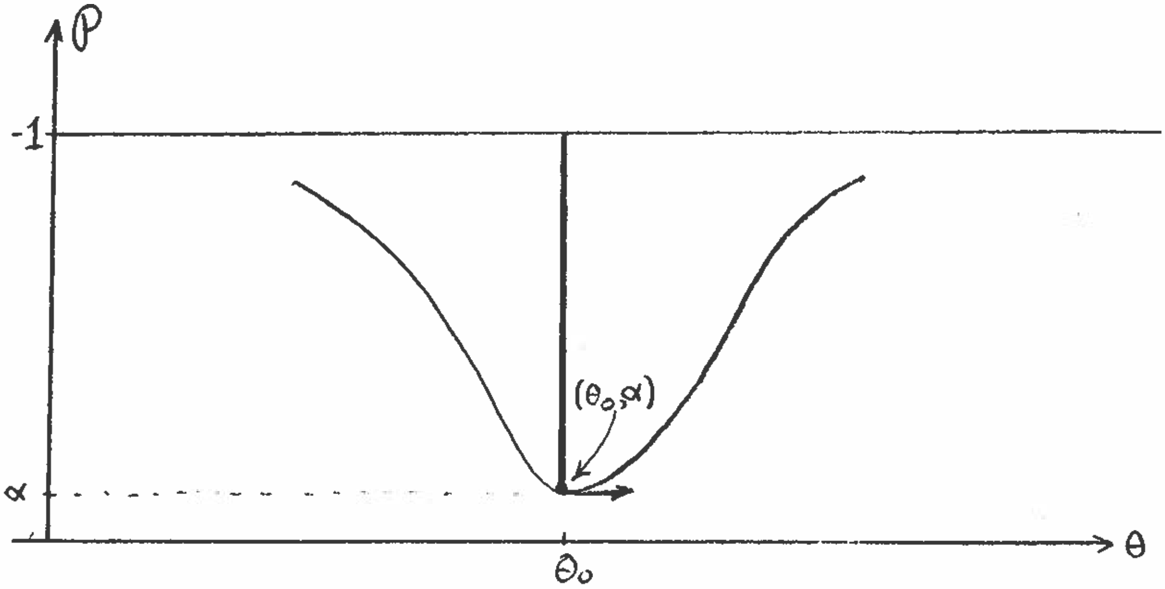

Figure 8: The power function of an unbiased size α test necessarily has a zero derivative at the point (Θ_0,α) .

Now suppose the conditions in Section 2(a) are fulfilled. We can then differentiate within the integration sign and a *locally unbiased size* α *test satisfies*

$$\int \phi(y) f(y|\Theta_0) dy = \alpha$$

$$\int \phi(y) S(y|\Theta_0) f(y|\Theta_0) dy = 0 \ .$$

Our mathematical approach is to look among tests satisfying

(iii) $$\int \phi(y)(a_1 + a_2 S(y|\Theta_0)) f(y)|\Theta_0) dy = c$$

for one that maximizes

$$\int \phi(y) f(y|\theta_1) dy \ .$$

If such a test satisfies (ii) then it is the most powerful (at θ_1) locally unbiased size α test. And if it satisfies (i) then it is the most powerful (at θ_1) unbiased size α test.

Example 3 continued. Let $(y_1,\ldots,y_n)$ be a sample from the normal (θ,σ_0^2) and suppose we seek the most powerful unbiased size α test for the problem $H_0:\theta = \theta_0$ against $H_1:\theta \neq \theta_0$.

The notation becomes somewhat simpler if we relocate the response relative to θ_0 and thus in effect examine the problem $H_0:\theta = 0$ against $H_1:\theta \neq 0$.

The score function at $\theta = 0$ is

$$S(y,0) = \frac{\bar{y}}{\sigma_0^2/n}$$

and

$$a_1 + a_2 S(y,0) = b_1 + b_2\bar{y} \ .$$

Then from the generalized lemma (Problem 69) we obtain the most powerful (at θ_1) test satisfying the condition (iii):

$$\begin{aligned} \phi(y) &= 1 \qquad f(\underset{\sim}{y}|\theta_1) > k(b_1+b_2\bar{y}) f(\underset{\sim}{y}|\theta_0) \\ &= 0 \qquad\qquad\qquad\ < k(b_1+b_2\bar{y}) f(\underset{\sim}{y}|\theta_0) \end{aligned}$$

or

$$\begin{array}{ll} \phi(\underset{\sim}{y}) = 1 & \exp\left\{\dfrac{(\theta_1 - 0)n\bar{y}}{\sigma_0^2}\right\} > C_1 + C_2\bar{y} \\ \quad = 0 & \qquad\qquad < C_1 + C_2\bar{y}_0 \end{array} \quad .$$

But this is an exponential function in comparison with a linear function; hence

$$\begin{array}{ll} \phi(\underset{\sim}{y}) = 1 & \bar{y} < d_1 \quad \text{or} \quad d_2 < \bar{y} \\ \quad = 0 & d_1 < \bar{y} < d_2 \end{array}$$

The following symmetric choice of limits gives an unbiased size α test for the original H_0 against H_1 :

$$\begin{array}{ll} \phi(\underset{\sim}{y}) = 1 & \dfrac{|\bar{y}|}{\sigma_0/\sqrt{n}} \geq Z_{\alpha/2} \\ \quad = 0 & \qquad < Z_{\alpha/2} \end{array}$$

where a standard normal variable exceeds $Z_{\alpha/2}$ with probability $\alpha/2$. This test is the most powerful (at θ_1) unbiased size α test of $H_0 : \theta_0 = 0$; it does not depend on θ_1 and hence is the UMP unbiased size α test for the original problem. The test can be presented in the following alternative form

$$\begin{array}{ll} \phi(\underset{\sim}{y}) = 1 & \text{if} \quad \dfrac{n\bar{y}^2}{\sigma_0^2} \geq \chi^2_\alpha \\ \quad = 0 & \qquad\qquad < \chi^2_\alpha \end{array}$$

where a chi-square variable on one degree of freedom exceeds

χ^2_α with probability α .

In general now consider the exponential model

$$f(y|\Theta) = \gamma(\Theta) \exp\{\psi(\Theta)t(y)\}h(y)$$

where $\psi(\Theta)$ is a differentiable increasing function of Θ . As examples we have the location normal, the scale normal, the binomial and the poisson. Then for the hypothesis testing problem

$$H_0 : \Theta_0$$
$$H_1 : \Theta \neq \Theta_0$$

we obtain the following lemma by the same proof as for the example.

Lemma. *The test*

$$\begin{aligned} \phi(y) &= 1 && \text{if } t(y) < d_1 \text{ or } d_2 < t(y) \\ &= a_i && t(y) = d_i \\ &= 0 && d_1 < t(y) < d_2 \end{aligned}$$

where d_1 , d_2 , a_1 , a_2 are chosen to *satisfy*

$$E(\phi(y)|\Theta_0) = \alpha$$

$$\frac{d}{d\Theta} E(\phi(y)|\Theta)\Big|_{\Theta=\Theta_0} = 0$$

is a uniformly most powerful unbiased size α *test.*

Example 6. Consider the linear model.

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{u}$$

presented in Section 1(c). Suppose that X consists of one column vector and that $\underset{\sim}{u}$ is a sample from the normal $(0,\sigma_0^2)$; then

$$\underset{\sim}{y} = \beta\underset{\sim}{x} + \underset{\sim}{u} \quad . \qquad \underset{\sim}{u} = \underset{\sim}{y} - \beta\underset{\sim}{x}$$

And suppose we seek the most powerful unbiased size α test for the problem $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$. (To test the value β_0 we would use the adjusted response $\underset{\sim}{u} - \beta_0\underset{\sim}{x}$). Note that Example 3 and the present Example 6 are equivalent in essential details: the variation has the rotationally symmetric distribution of a sample from the normal $(0,\sigma_0^2)$ and the mean is somewhare on the line $L(\underset{\sim}{1})$ in Example 3 and on the line $L(\underset{\sim}{x})$ in the present case.

The statistical model is

$$f(y|\beta) = (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_0^2}\,\Sigma\left(y_i-\beta x_i\right)^2\right\} \quad .$$

Note by Section 1(c) and Problem 7 that the sum of squares in the exponent can be rewritten

$$(\underset{\sim}{y}-\beta\underset{\sim}{x})'(\underset{\sim}{y}-\beta\underset{\sim}{x}) = \Sigma\left(y_i-bx_i\right)^2 + \left(\Sigma x_i^2\right)(b-\beta)^2$$

where

$$b = b(\underset{\sim}{y}) = \frac{\Sigma x_i y_i}{\Sigma x_i}$$

is the coefficient of the projection $b\underset{\sim}{x}$ of $\underset{\sim}{y}$ onto the line $L(x)$; see Figure 9 . The statistical model

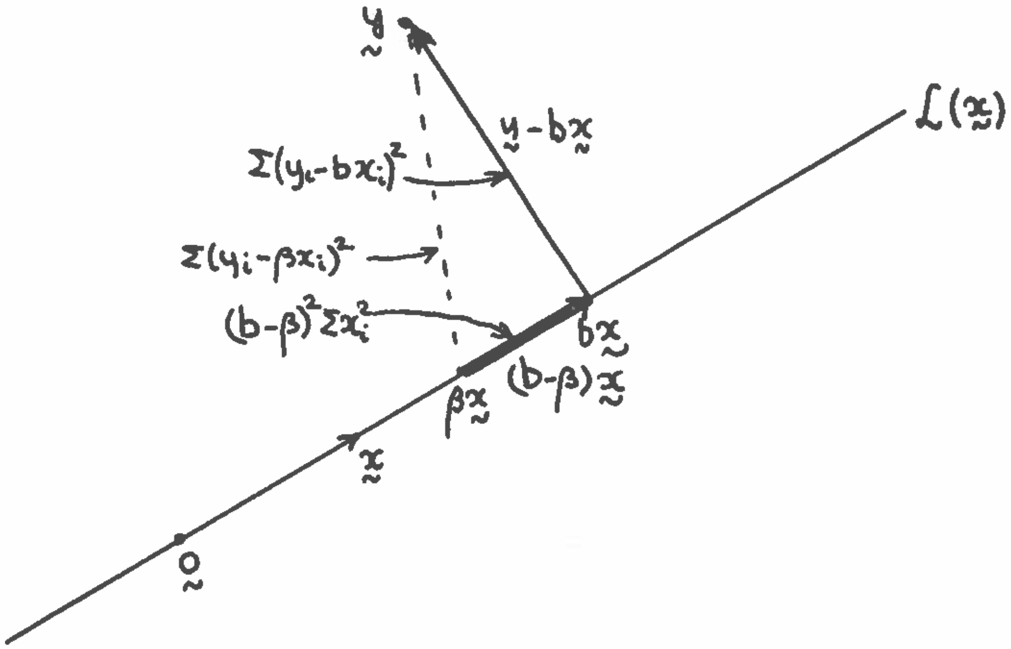
$\underset{\sim}{y}$
$\underset{\sim}{y} - b\underset{\sim}{x}$
$L(\underset{\sim}{x})$
$\Sigma(y_i - bx_i)^2$
$\Sigma(y_i - \beta x_i)^2$
$(b-\beta)^2 \Sigma x_i^2$
$b\underset{\sim}{x}$
$(b-\beta)\underset{\sim}{x}$
$\beta\underset{\sim}{x}$
$\underset{\sim}{x}$
$\underset{\sim}{0}$

Figure 9. An observed $\underset{\sim}{y}$ from a distribution centered at $\beta\underset{\sim}{x}$. The squared distance $\Sigma(y_i - \beta\underset{\sim}{x})^2$ can be separated into squared distance $(b-\beta)^2\Sigma x_i^2$ in the direction $L(\underset{\sim}{x})$ and squared distance $\Sigma(y_i - bx_i)^2$ in the orthogonal complement $L^{\perp}(\underset{\sim}{x})$.

can be written

$$f(\underset{\sim}{y}|\beta) = (2\pi\sigma_0^2)^{-\frac{n}{2}}\exp\left\{-\frac{\beta^2\Sigma x_i^2}{2\sigma_0^2}\right\}\cdot\exp\left\{\frac{\beta\Sigma x_i^2\cdot b}{\sigma_0^2}\right\}\exp\left\{-\frac{\Sigma y_i^2}{2\sigma_0^2}\right\}$$

and it has the form of the exponential model preceding the lemma. By the lemma the uniformly most powerful unbiased test is to reject for extreme values of $b = \Sigma x_i y_i/\Sigma x_i^2$. But under the hypothesis $H_0 : \beta = 0$, the function b is normally distributed with mean 0 and variance $\sigma_0^2/\Sigma x_i^2$. Hence the UMP unbiased size α test is the symmetric test

$$\phi(\underset{\sim}{y}) = 1 \qquad \text{if} \quad \frac{|b|}{\sigma_0/(\Sigma x_i^2)^{\frac{1}{2}}} \geq Z_{\alpha/2}$$

$$= 0 \qquad\qquad < Z_{\alpha/2}$$

or equivalently is

$$\phi(\underset{\sim}{y}) = 1 \qquad \text{if} \quad \frac{\Sigma x_i^2 b^2}{\sigma_0^2} \geq \chi_\alpha^2$$

$$= 0 \qquad\qquad < \chi_\alpha^2$$

where a chi-square variable on one degree of freedom exceeds χ_α^2 with probability α . The second form can be interpreted as follows: the hypothesis model is $\underset{\sim}{y} = 0\underset{\sim}{x} + \underset{\sim}{u}$; the sum of squares (SS) of deviations from the model is Σy_i^2 ; the more general model is $\underset{\sim}{y} = \beta\underset{\sim}{x} + \underset{\sim}{u}$; the sum of squares (SS) of deviations from the model is

$$\Sigma(y_i - bx_i)^2 = \Sigma y_i^2 - b^2 \Sigma x_i^2$$

(by Section 1(c)); the reduction in SS in going from the special to the more general model is

$$\Sigma \chi_i^2 \, b^2$$

which as a proportion of σ_0^2 is compared with χ_α^2 to obtain the test .

(b) Similar tests

Now more generally consider the testing of a composite hypothesis H_0 against a composite H_1 and suppose that any power function is continuous in Θ . See Figure 10.

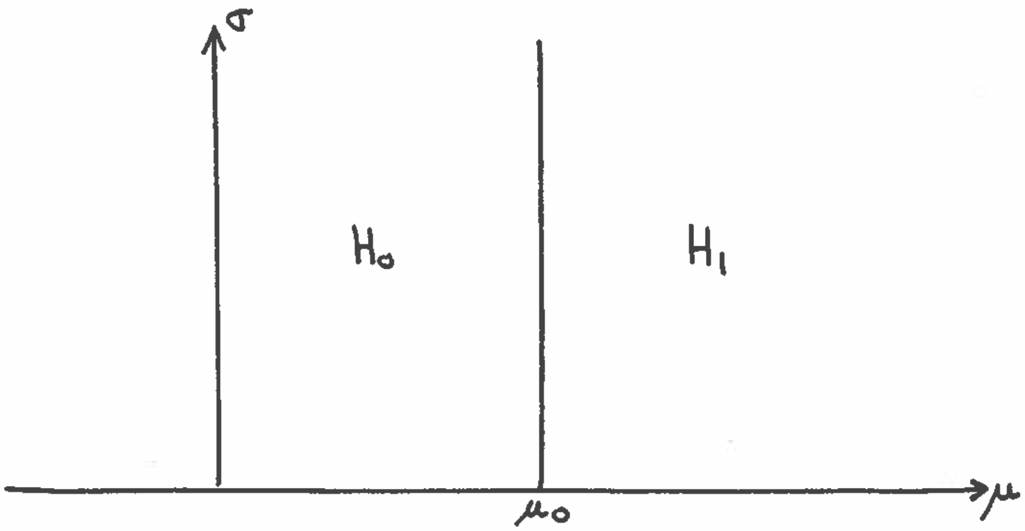

Figure 10. The hypothesis $H_0 : \mu \leq \mu_0$, $\sigma^2 \in \mathbb{R}^+$ against the alternative $H_1 : \mu > \mu_0$, $\sigma^2 \in \mathbb{R}^+$. An unbiased size α test ϕ has $E(\phi(y)|\Theta) \equiv \alpha$ for all Θ in $\omega = \bar{H}_0 \cap \bar{H}_1$ (the set of common boundary points).

An unbiased size α test ϕ satisfies

$$\begin{aligned} E(\phi(y)|\Theta) &\leq \alpha \qquad \Theta \text{ in } H_0 \\ &\geq \alpha \qquad \Theta \text{ in } H_1 \end{aligned}$$

By the continuity, any such test satisfies the condition

(iv) $$E(\phi(y)|\Theta) = \alpha \qquad \Theta \text{ in } \omega$$

where $\omega = \bar{H}_0 \cap \bar{H}_1$ is the set of common limit points in Ω ; a test ϕ satisfying (iv) is called a *similar size* α *test on* ω.

The larger class of similar tests has a condition involving equalities rather than inequalities. Moreover the equation (iv) says that $\phi(y)$ is an unbiased estimate of the constant α for the model with parameter space ω . Thus in accord with Section 3 suppose that $s(y)$ is a complete likelihood statistic (re ω). Then the unbiased estimate of α based on s is unique; hence

(v) $$E(\phi(y):s|\omega) = \alpha ,$$

and ϕ thus has *exact* size α with respect to the conditional distribution given s (the ω is included in the expression (v) since the conditional distribution typically will depend on Θ outside ω) .

Thus an unbiased size α test of H_0 against H_1 is an exact size α test of ω conditionally for each value of s where s is the complete likelihood statistic for ω. If we find the most powerful (Θ_1) test of ω conditionally for each value of s, and if the resulting test is an unbiased test of H_0 against H_1 then it is the most powerful (Θ_1) unbiased size α test.

Example 1 continued. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ,σ^2) and consider the hypothesis testing problem

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \quad \sigma^2 \in \mathbb{R}^+ \\ H_1 &: \mu > \mu_0 \quad \sigma^2 \in \mathbb{R}^+ \quad . \end{aligned}$$

Again the notation becomes simpler if we relocate the response relative to μ_0 and thus in effect examine the case $\mu_0 = 0$.

By the arguments preceding the example, an unbiased size α test of H_0 against H_1 is a similar size α test of

$$\omega : \mu = 0 \qquad \sigma^2 \in \mathbb{R}^+ \quad .$$

And since $S(\underset{\sim}{y}) = (\Sigma y_i^2)^{\frac{1}{2}}$ is a complete likelihood statistic (Problem 36), then such a test is an exact size α test (of ω) *conditionally* given any value for S; of course the conditional distribution does not depend on Θ as long as Θ is in ω.

We now consider the conditional distribution given S and find the exact size α test that has maximum power at

(μ_1, σ_1^2) with $\mu_1 > 0$. By the lemma in Section 4 the test has the form

$$\begin{aligned} \phi(\underset{\sim}{y}) &= 1 \qquad \text{if} \quad \frac{f(\underset{\sim}{y}:S|\mu_1,\sigma_1^2)}{f(\underset{\sim}{y}:S|0,\sigma_1^2)} > k \\ &= 0 \qquad\qquad\qquad\qquad\quad < k \end{aligned}$$

where f gives density for suitable coordinates conditionally given S. By THREE Section 2e the conditional density is the overall density combined with a jacobian and normalizing constant. Such extra factors will cancel in the density ratio and the test thus has the form

$$\begin{aligned} \phi(\underset{\sim}{y}) &= 1 \qquad \text{if} \quad \frac{f(\underset{\sim}{y}|\mu_1,\sigma_1^2)}{f(\underset{\sim}{y}|0,\sigma_1^2)} > k \\ &= 0 \qquad\qquad\qquad\qquad\quad < k \end{aligned}$$

where k is chosen for each contour of S to obtain exact size α conditionally. By Example 3 in Section 4 the test is to reject for large values of $\bar{y}$ or equivalently for large values of

$$t = \frac{\bar{y}}{s_{\underset{\sim}{y}}/\sqrt{n}} \;;$$

see Figure 11 or the argument at the end of SIX, Section 2b. The preceding likelihood ratio is in fact monotone in t and *conditionally* given S the test is uniformly most powerful size α for $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$ (lemma in Section 5b).

The ω ~~distribution~~ density (with $\mu = 0$) is rotationally symmetric on $\mathbb{R}^n$ and thus constant valued on the contours of S ($S^2 = \Sigma y_i^2$ = constant is a sphere centered at the origin). The function $t = \sqrt{n}\,\bar{y}/s_y$ is a function of an angle as indicated in Figure 12. It follows

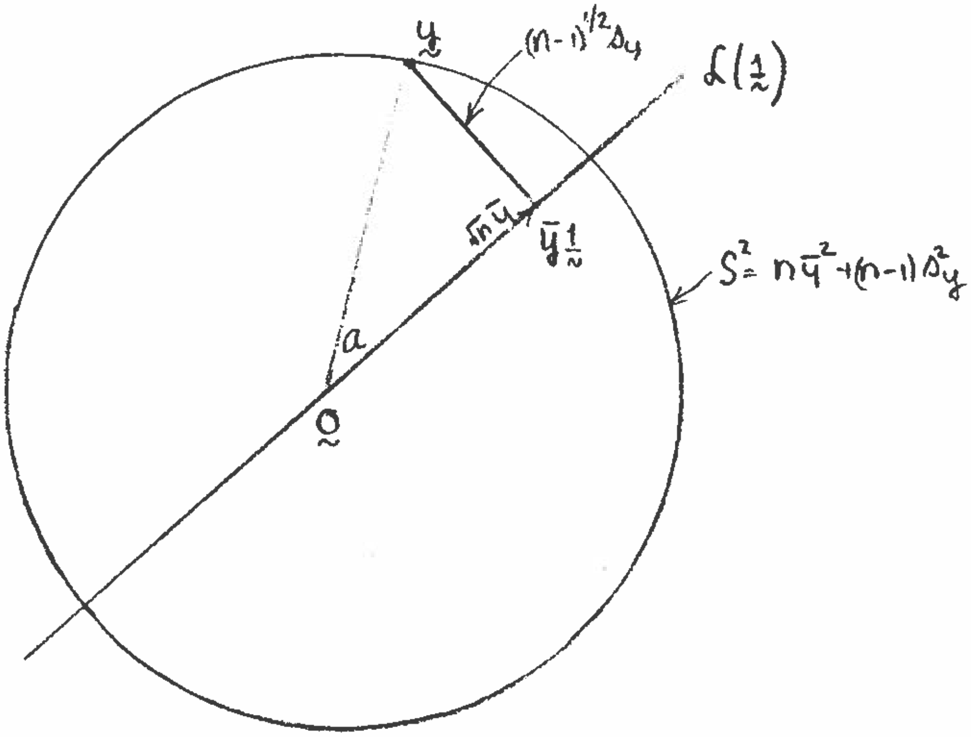
$\underset{\sim}{y}$
$(n-1)^{1/2}s_y$
$\mathcal{L}(\underset{\sim}{1})$
$\sqrt{n}\,\bar{y}$
$\bar{y}\underset{\sim}{1}$
$S^2 = n\bar{y}^2 + (n-1)s_y^2$
a
$\underset{\sim}{0}$

Figure 12. $t/\sqrt{n}-1 = \sqrt{n}\,\bar{y}/\sqrt{n-1}\ s_y$ is the cotangent of the angle between the $\underset{\sim}{1}$ vector and the $\underset{\sim}{y}$ vector.

that the conditional distribution of t given S is independent of S and thus is the same as the marginal distribution (by THREE, Section 2(b)). Hence the conditional exact size α test that has maximum power at (μ_1,σ_1^2) is

$$\begin{aligned}\phi(\underset{\sim}{y}) &= 1 \qquad t \geq t_\alpha \\ &= 0 \qquad\ \ < t_\alpha\end{aligned}$$

where a t-variable on $n-1$ degrees of freedom exceeds t_α with probability α. The unbiasedness for the original H_0 follows from the conditional unbiasedness noted at the end of the preceding paragraph. Or it can be checked directly:

$$\begin{aligned}P_\phi(\mu,\sigma^2) &= P\left\{\frac{\sqrt{n}\bar{y}}{s_y} \geq t_\alpha \middle| \mu,\sigma^2\right\} \\ &= P\left\{\frac{\sqrt{n}(\bar{y}-\mu)}{s_y} \geq t_\alpha - \frac{\sqrt{n}\mu}{s_y} \middle| \mu,\sigma^2\right\}\end{aligned}$$

which is greater or less than

$$P\left\{\frac{Z}{\chi/\sqrt{n-1}} \geq t_\alpha\right\} = \alpha$$

according as μ is greater or less than zero.

The test ϕ is a most powerful (μ_1,σ_1^2) unbiased size α test of H_0 against H_1. But the test does not depend on (μ_1,σ_1^2) provided $\mu_1 > 0$; hence the t test is UMP unbiased size α for H_0 against H_1.

And in addition by using the methods in part (a) we find that the test

$$\phi(\underset{\sim}{y}) = 1 \qquad \text{if} \quad |t| \geq t_{\alpha/2}$$

$$= 0 \qquad\qquad\qquad t_{\alpha/2}$$

is UMP unbiased size α for

$$H_0 : \mu = 0 \qquad \sigma^2 \in \mathbb{R}^+$$

$$H_1 : \mu \neq 0 \qquad \sigma^2 \in \mathbb{R}^+ \quad .$$

Now in general consider the exponential model

$$f(y|\Theta,\underset{\sim}{\psi}) = \gamma(\Theta,\underset{\sim}{\psi})\exp\left\{\Theta T(y) + \sum_1^r \psi_i s_i(y)\right\}h(y)$$

with $(\Theta,\underset{\sim}{\psi})$ in Ω, an open set in $\mathbb{R}^{r+1}$. Then for the hypothesis testing problem

$$H_0 : \Theta \leq \Theta_0$$

$$H_1 : \Theta > \Theta_0$$

we have the following lemma.

Lemma. *The test*

$$\begin{aligned}\phi(\underset{\sim}{y}) &= 1 && \text{if } T(\underset{\sim}{y}) > k(\underset{\sim}{s}) \\ &= a(\underset{\sim}{s}) && \qquad\quad\; = k(\underset{\sim}{s}) \\ &= 0 && \qquad\quad\; < k(\underset{\sim}{s})\end{aligned}$$

where $k(\underset{\sim}{s})$ *and* $a(\underset{\sim}{s})$ *are chosen to satisfy*

$$E(\phi(\underset{\sim}{y}):\underset{\sim}{s}|\Theta_0) = \alpha$$

is uniformly most powerful unbiased size α *of* H_0 *against* H_1 .

Proof. The proof parallels that for the example preceding the lemma. The set ω is the Θ_0 section of Ω; the function $\underset{\sim}{s}(y)$ is a complete likelihood statistic for ω (Problem 35). Conditionally given $\underset{\sim}{s}$ the likelihood ratio test is to reject for large values of t with critical value determined by the Θ_0 conditional distribution; the critical value will typically depend on the contour of $s(\underset{\sim}{y})$. Conditionally given $\underset{\sim}{s}$ any likelihood ratio is monotone in t whenever one of the Θ values to zero; hence conditionally given $\underset{\sim}{s}$ the test is UMP unbiased size α (by lemma, Section 5(b)). It follows that the overall test is unbiased and hence is UMP unbiased size α .

Consider further the exponential model

$$f(y|\Theta,\underset{\sim}{\psi}) = \gamma(\Theta,\underset{\sim}{\psi})\exp\left\{\Theta T(y) + \sum_1^r \psi_i s_i(y)\right\}h(y)$$

with $(\Theta, \underset{\sim}{\psi})$ in an open set of $\mathbb{R}^{r+1}$. Then for the two-sided problem

$$H_0 : \Theta = \Theta_0$$
$$H_1 : \Theta \neq \Theta_0$$

we have the following lemma.

Lemma. *The test*

$$\begin{aligned} \phi(y) &= 1 & &\text{if} \quad T(y) < d_1(\underset{\sim}{s}) \quad \text{or} \quad d_2(\underset{\sim}{s}) < T(y) \\ &= a_i(\underset{\sim}{s}) & &\qquad T(y) = d_i(\underset{\sim}{s}) \\ &= 0 & &\qquad d_1(\underset{\sim}{s}) < T(y) < d_2(\underset{\sim}{s}) \end{aligned}$$

where $d_1(\underset{\sim}{s})$, $d_2(\underset{\sim}{s})$, $a_1(\underset{\sim}{s})$, $a_2(\underset{\sim}{s})$ *are chosen to satisfy*

$$E(\phi(y):s|\Theta_0) = \alpha$$
$$\frac{d}{d\Theta} E(\phi(y):s|\Theta)|_{\Theta_0} = 0$$

is uniformly most powerful unbiased size α *for testing* H_0 *against* H_1 .

Proof. Combine the analysis for the preceding lemma with that for the lemma in part (a) .

Example 1 continued. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ, σ^2) and then consider the one and two sided hypothesis relative to $\mu = 0$. We check that the statistical model has the exponential form needed for the two lemmas

$$f(\underset{\sim}{y}|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\,\Sigma(y_i-\mu)^2\right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\}\exp\left\{\bar{y}\,\frac{n\mu}{\sigma^2} + \Sigma y_i^2\,\frac{-1}{2\sigma^2}\right\} .$$

The UMP unbiased tests are to reject for large values of $\bar{y}$ or $|\bar{y}|$ conditionally given Σy_i^2 . But these are the t-tests as derived preceding the lemmas.

Example 6 continued. Consider the linear model

$$\underset{\sim}{y} = X_r\underset{\sim}{\beta} + \underset{\sim}{u}$$

presented in Section 1(c) . Suppose that X_r consists of r linearly independent column vectors and that $\underset{\sim}{u}$ is a sample from the normal $(0,\sigma_0^2)$. And suppose we want the uniformly most powerful unbiased size α test for the problem $H_0 : \beta_r = 0$ against $H_1 : \beta_r \neq 0$ (the one-sided test can be treated with obvious modifications).

The linear model as given allows the mean to be any point in the r dimensional space $L(X_r)$ formed from the r column vectors; the hypothesis however restricts the mean to the $r-1$ dimensional space $L(X_{r-1})$ formed from the first $r-1$ column vectors of X_r .

Now consider a new set of axes (orthonormal) whose first $r-1$ axes lie in $L(X_{r-1})$ and whose r th axis lies then in $L(X_r)$. (See for example, SEVEN, Section 2(b)); the new coordinates are obtained by an orthogonal transformation. The linear model then becomes

$$
\begin{aligned}
y_1 &= \nu_1 + u_1 \\
&\vdots \\
y_r &= \nu_r + u_r \\
y_{r+1} &= \phantom{\nu_r + {}} u_{r+1} \\
&\vdots \\
y_n &= \phantom{\nu_r + {}} u_n \quad .
\end{aligned}
$$

with the testing problem $H_0 : \nu_r = 0$ against $H_1 : \nu_r \neq 0$.

The statistical model is

$$
\begin{aligned}
f(\underset{\sim}{y}|\underset{\sim}{\nu}) &= (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_0^2}\left[\sum_1^r (y_i-\nu_i)^2 + \sum_{r+1}^n y_j^2\right]\right\} \\
&= (2\pi\sigma_0^2)^{-\frac{n}{2}} \exp\left\{-\frac{\Sigma\nu_i^2}{2\sigma_0^2}\right\} \exp\left\{\frac{\nu_r}{\sigma_0^2} y_r + \sum_1^{r-1} \frac{\nu_j}{\sigma_0^2} y_j\right\} \exp\left\{-\frac{1}{2\sigma_0^2}\Sigma y_j^2\right\} .
\end{aligned}
$$

The UMP unbiased test is calculated as a conditional test given $y_1, \ldots, y_{r-1}$, But by SEVEN, Section 2(a) this distribution is given by

$$
\begin{aligned}
y_r &= \nu_r + u_r \\
y_{r+1} &= \phantom{\nu_r + {}} u_{r+1} \\
&\vdots \\
y_n &= \phantom{\nu_r + {}} u_n
\end{aligned}
$$

where the u's form a sample of $n-r+1$ from the normal $(0,\sigma_0^2)$; note that this is a simple version of the example as given in part (a) . The UMP unbiased size α test is

$$\phi(\underset{\sim}{y}) = 1 \qquad \text{if } \frac{(v_{r+1})}{\sigma_0} \geq Z_{\alpha/2}$$
$$= 0 \qquad\qquad < Z_{\alpha/2}$$

or equivalently is

$$\phi(\underset{\sim}{y}) = 1 \qquad \text{if } \frac{S^2}{\sigma_0^2} \geq \chi_\alpha^2$$
$$= 0 \qquad\qquad < \chi_\alpha^2$$

where $S^2 = v_r^2$ and χ_α^2 is the α point of chi-square on one degree of freedom.

The second form of the test can be interpreted as follows: the hypothesis model is

$$\begin{aligned} y_1 &= \nu_1 + u_1 \\ &\vdots \\ y_{r-1} &= \nu_{r-1} + u_{r-1} \\ y_r &= u_r \\ &\vdots \\ y_n &= u_n \end{aligned}$$

and the SS of deviations from the fitted model is $\sum_r^n y^2$, (note that $\hat{\nu}_1 = y_1, \ldots, \hat{\nu}_{r-1} = y_{r-1}$) ; the more general model allows a mean ν_r for y_r ; the SS of deviations from the fitted model is $\sum_{r+1}^{n} y_j^2$; the reduction in SS in going from the special to the more general model is y_r^2 which as a proportion of σ_0^2 is compared with χ_α^2 to obtain the test.

The preceding interpretation allows us to record the test for the problem in its original form at the beginning of the example: the test is as given above but with S^2 now defined by

$$S^2 = \underset{\sim}{y}\, X_r(X_r'X_r)^{-1}X_r'\, \underset{\sim}{y} - \underset{\sim}{y}'X_{r-1}(X_{r-1}'X_{r-1})^{-1}X_{r-1}'\, \underset{\sim}{y}$$

using the formula of Section 1(c) .

Example 7. Now consider the linear model

$$\underset{\sim}{y} = X_r\underset{\sim}{\beta} + \sigma\underset{\sim}{u}$$

where $\underset{\sim}{u}$ is a sample from the standard normal. And suppose that we want the UMP unbiased size α test for $H_0 : \beta_r = 0$ against $H_1 : \beta_r \neq 0$.

Consider a new set of axes as described in the preceding Example 6. The linear model then becomes

$$\begin{aligned}
y_1 &= \nu_1 + \sigma u_1 \\
&\vdots \\
y_r &= \nu_r + \sigma u_r \\
y_{r+1} &= \qquad \sigma u_{r+1} \\
&\vdots \\
y_n &= \qquad \sigma u_n
\end{aligned}$$

with the testing problem $H_0 : \nu_r = 0$ against $H_1 : \nu_r \neq 0$. The statistical model is

$$f(y|\nu) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_1^r (y_i - \nu_i)^2 + \sum_{r+1}^n y_j^2 \right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{\Sigma\nu_i^2}{2\sigma^2} \right\} \exp\left\{ y_r \frac{\nu_r}{\sigma^2} + \sum_1^n y_i^2 \frac{-1}{2\sigma^2} + \sum_1^{r-1} y_i \frac{\nu_i}{\sigma^2} \right\}$$

and the UMP unbiased test is to reject for extreme values of y_r conditionally given $y_1, \ldots, y_{r-1}, \sum_1^n y_i^2$, or given $y_1, \ldots, y_{r-1}, \sum_r^n y_i^2$. But by SEVEN, Section 2(a) this distribution is given by

$$\begin{aligned}
y_r &= \nu_r + \sigma u_r \\
y_{r+1} &= \qquad \sigma u_{r+1} \\
&\vdots \\
y_n &= \qquad \sigma u_n
\end{aligned}$$

conditionally given $\sum_r^n y_i^2$. But this is a canonical form of

Example 1 before and after the lemmas and the UMP unbiased size α test is to reject for large values of $|t|$ where

$$t = \frac{y_r}{\left[\sum_{r+1}^{n} y_j^2/(n-r)\right]^{\frac{1}{2}}}$$

or for large values of

$$F = \frac{y_r^2/1}{\sum_{r+1}^{n} y_j^2/(n-r)} \quad .$$

The test is

$$\begin{aligned} \phi(\underset{\sim}{y}) &= 1 \qquad & F &\geq F_\alpha \\ &= 0 & &< F_\alpha \end{aligned}$$

where F_α is the α point on the tail of the F distribution on 1 over n-r degrees of freedom.

The components of the F function can be interpreted directly as in the preceding version of this Example 6. It follows that the UMP unbiased size α test for the original problem is as given above but with F defined by

$$F = \frac{\underset{\sim}{y}'X_r(X_r'X_r)^{-1}X_r'\underset{\sim}{y} - \underset{\sim}{y}'X_{r-1}(X_{r-1}'X_{r-1})^{-1}X_r'\,\underset{\sim}{y}}{(\underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}'X_r(X_r'X_r)^{-1}\,X_r'\,\underset{\sim}{y})/\,(n-r)} \,.$$

Problems

83. For the exponential model $F(y|\Theta) = \gamma(\Theta)\exp\{\Theta t(y)\}h(y)$ show that a locally unbiased size α test of Θ_0 satisfies

$$E(\phi(y)|\Theta_0) = \alpha, \quad E(t(y)\phi(y)|\Theta_0) = \alpha E(t(y)|\Theta_0) \quad .$$

84. Scale normal. Let $(y_1, \ldots, y_n)$ be a sample from the normal $(0,\sigma^2)$ and let $S^2 = \Sigma y_i^2$. Show that the UMP unbiased size α test of $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$ is to reject if $S^2/\sigma_0^2 \leq d_1$ or $d_2 \leq S^2/\sigma_0^2$ where d_1 and d_2 are determined so that (d_1, d_2) is a $1-\alpha$ probability interval for chi-square on n degrees of freedom *and* for chi-square on $n+2$ degrees of freedom. Method: Use Problem 83 and use $tf_n(t) = nf_{n+2}(t)$ where f_n designate the χ^2 density on n degrees of freedom.

85. Scale exponential. Let $(y_1,\ldots,y_n)$ be a sample from the exponential $f(y|\Theta) = \Theta^{-1}\exp\{-y/\Theta\}$ on $(0,\infty)$. Determine the form of the UMP unbiased size α test of $H_0 : \Theta = \Theta_0$ against $H_1 : \Theta \neq \Theta_0$. Use the results from Problem 84.

86. Binomial. Let $(x_1, \ldots, x_n)$ be a sample from the bernoulli (p) and let $y = \Sigma x_i$. Show that the UMP unbiased size α test of $H_0 : p = p_0$ against

$H_1 : p \neq p_0$ is to reject if y is outside the interval $[d_1, d_2]$, to accept inside the interval (d_1, d_2) and to randomly a_1, a_2 reject at d_1, d_2 respectively, where the interval $[d_1, d_2]$ and the randomization at its and points are determined so as to have $1-\alpha$ probability according to the binomial (n, p_0) distribution for y *and* the binomial $(n-1, p_0)$ distribution for $y-1$. Method: Use Problem 83 and $y p_n(y|p) = np\, p_{n-1}(y-1|p)$ where $p_n(y|p)$ is the binomial probability function.

87. Poisson. Let y be poisson θ. Show that the UMP unbiased size α test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is to reject or accept on the basis of an interval $[d_1, d_2]$ with randomization at the end points (as in Problem 86) where the interval is determined so as to have $1-\alpha$ probability according to the poisson θ_0 for y *and* the poisson θ_0 for $y-1$.

88. Let $(y_{11}, \ldots, y_{1m})$ be a sample from the normal (μ_1, σ^2) and $(y_{21}, \ldots, y_{2n})$ be a sample from the normal (μ_2, σ^2). Deduce that the UMP unbiased size α test of $H_0 : \mu_1 = \mu_2$, $\sigma^2 = \sigma_0^2$ against $H_1 : \mu_1 \neq \mu_2$, $\sigma^2 = \sigma_0^2$ is to reject if

$$\frac{(\bar{y}_1 - \bar{y}_2)^2 \Big/ \left(\frac{1}{m} + \frac{1}{n}\right)}{\sigma_0^2}$$

exceeds χ^2_α, the α point on the tail of the chi square distribution on 1 degree of freedom.

89. Continuation. Deduce the UMP unbiased size α test of $H_0: \mu_1 = \mu_2, \sigma^2 \varepsilon \mathbb{R}^+$ against $H_1: \mu_1 \neq \mu_2, \sigma^2 \varepsilon \mathbb{R}^+$ is to reject if

$$\frac{(\bar{y}_1 - \bar{y}_2)^2 / \left(\frac{1}{m} + \frac{1}{n}\right)}{\left(\Sigma(y_{1i} - \bar{y}_1)^2 + \Sigma(y_{2j} - \bar{y}_2)^2\right)/(m+n-2)}$$

exceeds the α point for F on 1 over m+n-2 degrees of freedom.

7. INVARIANT TESTS

Consider a sample $(y_1, \ldots, y_n)$ from the normal distribution (μ,σ^2) with (μ,σ^2) in $\Omega = \mathbb{R} \times \mathbb{R}^+$, and , suppose we are interested in the problem $H_0 : \mu \leq 0$, $\sigma^2 \in \mathbb{R}^+$ against $H_1 : \mu > 0$, $\sigma^2 \in \mathbb{R}^+$. We have noted in the preceding section that this problem does not have a uniformly most powerful test $(\alpha < \frac{1}{2})$.

In the preceding section we introduced unbiased tests and obtained the t-test (reject for large values) -- as the uniformly most powerful test among the unbiased size α tests. In this section we introduce invariant tests and again obtain the t test -- now as the uniformly most powerful test among the invariant size α tests.

This normal-sample problem has a variety of symmetries that can be expressed by groups of transformations. Perhaps the most relevant transformations are the scale transformations, the dilations and contractions about the origin:

$$G = \{[0,c] : c \in \mathbb{R}^+\}$$

where

$$[0,c]\underset{\sim}{y} = c\underset{\sim}{y} = (cy_1, \ldots, cy_n)'$$

(for notation, see THREE, Section 3). If $\underset{\sim}{y}$ has the normal distribution (μ,σ^2) , then the transformed variable $\underset{\sim}{\tilde{y}} = [0,c]\underset{\sim}{y}$ has the normal distribution $(c\mu, c^2\sigma^2)$. Note that the new distribution is in Ω and any distribution in Ω is possible depending on the mean and variance of the original variable. Also note that the new distribution is in H_0 or H_1 according as the original distribution is in H_0 or H_1. Thus the transformed variable has the same statistical model and the same hypothesis testing problem. It is reasonable then to require the test to be invariant under such transformations, that is $\phi(c\underset{\sim}{y}) = \phi(\underset{\sim}{y})$ for all $c > 0$. More informally a test of whether a mean is less or greater than zero should not depend on what unit the measurements are expressed in -- feet, inches, centimetres.

Consider the illustration further. By Section 4 it suffices to examine tests that are based on the likelihood statistic which is, say, $(\bar{y}, s(\underset{\sim}{y}))$ where

$$s(\underset{\sim}{y}) = (\Sigma(y_i-\bar{y})^2)^{\frac{1}{2}} \quad .$$

A transformation $[0,c]$ carries $\underset{\sim}{y}$ into $\underset{\sim}{\tilde{y}} = c\underset{\sim}{y}$ and correspondingly carries $(\bar{y}, s(\underset{\sim}{y}))$ into $(c\bar{y}, cs(\underset{\sim}{y}))$; see Figure 11. Any invariant function of $(\bar{y}, s(\underset{\sim}{y}))$ is necessarily constant valued on the rays

$$\{(c\bar{y},cs) : c \in \mathbb{R}^+\}$$

from the origin (for convenience ignore the probability-zero

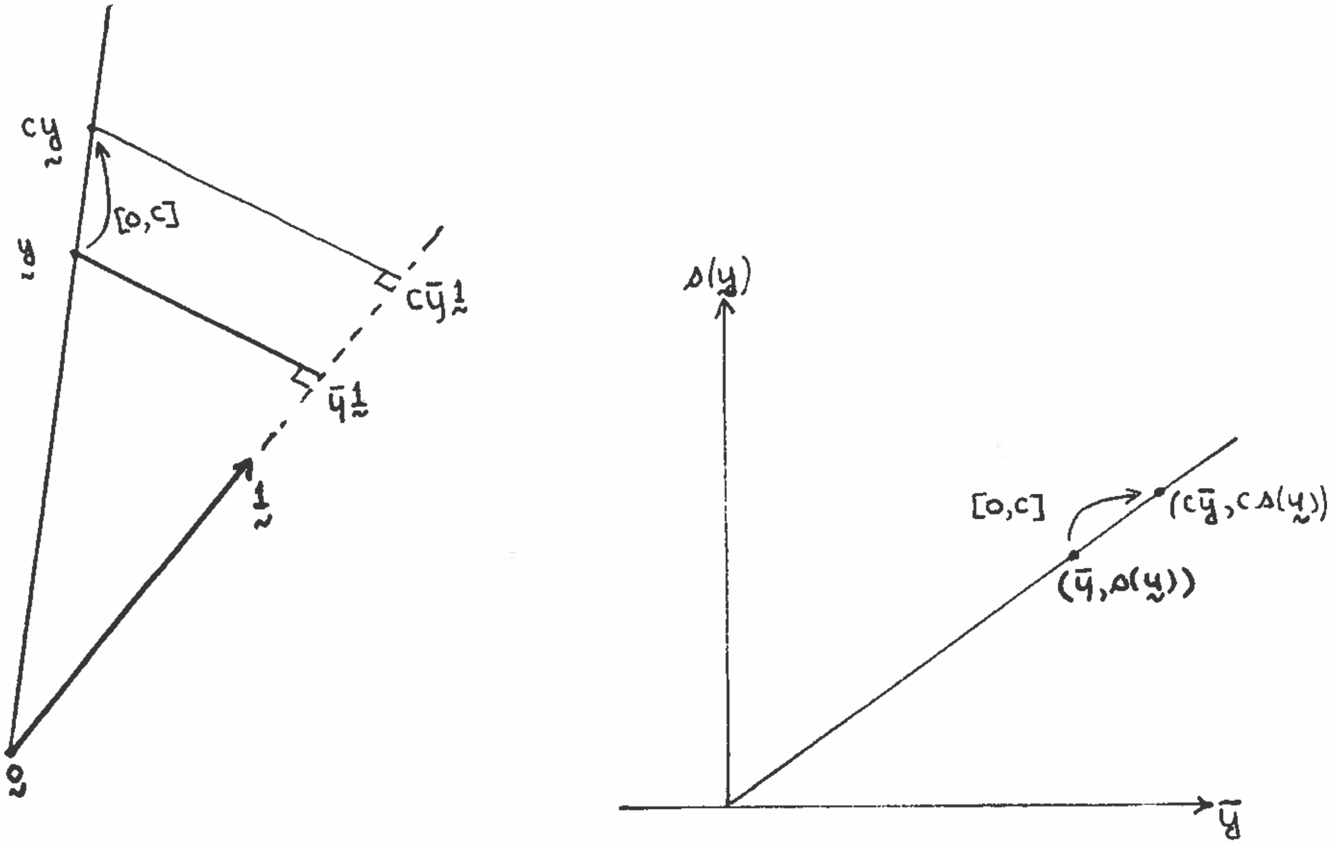
$c\underset{\sim}{y}$
$[0,c]$
$\underset{\sim}{y}$
$c\bar{y}\underset{\sim}{1}$
$\bar{y}\underset{\sim}{1}$
$\underset{\sim}{1}$
$\underset{\sim}{0}$
$s(\underset{\sim}{y})$
$[0,c]$
$(c\bar{y}, c\,s(\underset{\sim}{y}))$
$(\bar{y}, s(\underset{\sim}{y}))$
$\bar{y}$

Figure 11. The transformation $[0,c]$ on $\mathbb{R}^n$ and the induced transformation on the compatible function $(\bar{y}, s(\underset{\sim}{y}))$

set having $s(\underset{\sim}{y}) = 0)$, and hence is expressible as a function of the direction of the ray given by say

$$t = \frac{\sqrt{n}\bar{y}}{s_y} = \frac{\sqrt{n}\,\bar{y}}{s(\underset{\sim}{y})/(n-1)^{\frac{1}{2}}} \quad .$$

At the end of this section we derive the uniformly most powerful invariant test for this problem; before that, we discuss some of the concepts so far in a more general framework.

(a) Invariant models and hypotheses.

Consider a statistical model with response y in $\mathcal{S}$ and parameter θ in Ω . And let $G = \{g\}$ be a group or one-one continuously differentiable transformations of $\mathcal{S}$ into $\mathcal{S}$ (in the discrete case it suffices to have one-one transformations of the countable sample space into itself).

Consider the effect of the group G on the sample space $\mathcal{S}$; see Figure 12. From any point $\underset{\sim}{y}$ we can form

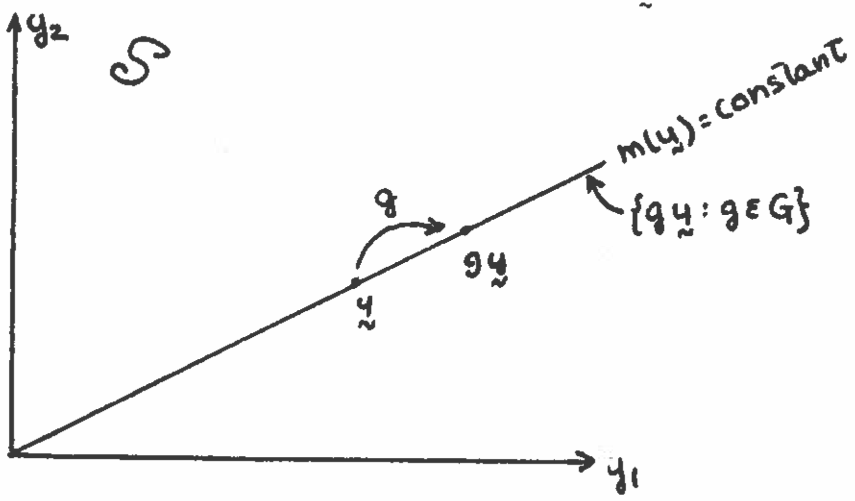
y_2
$\mathcal{S}$
$m(\underset{\sim}{y})=$constant
$\{g\underset{\sim}{y} : g \varepsilon G\}$
g
$g\underset{\sim}{y}$
$\underset{\sim}{y}$
y_1

Figure 12. The group G partitions the space S into sets $C_{\underset{\sim}{y}}$.

the set $\{gy : g \text{ in } G\}$ of all image points under the transformations. Such sets are either identical or disjoint and hence partition the space $\mathcal{S}$; see THREE, Section 3. Let $m(y)$ be a function that indexes these sets. An invariant function ϕ ($\phi(gy) = \phi(y)$ for all y and g) is constant valued on any set of the partition and hence can be expressed as a function of $m(y)$. Thus an invariant function is an arbitrary function of $m(y)$ and accordingly $m(y)$ is called the maximal invariant function.

Now suppose that the transformation group G is consonant with the model, specifically

(i) *If* y *has distribution* θ *in* Ω *, then* gy *has a distribution that is given by the parameter value* $\bar{g}\theta$ *also in* Ω .

The effect of this condition is that the possible distributions for y are the same as the possible distributions for gy . With several transformation g , h , g^{-1} we easily obtain the following distributions for variables:

Variable	y	gy	hgy	$g^{-1}gy$
Parameter	θ	$\bar{g}\theta$	$\bar{h}\bar{g}\theta$	$\overline{g^{-1}}\bar{g}\theta$
			$=\overline{hg}\theta$	$= \theta$.

It follows that the transformations $\bar{G} = \{\bar{g}\}$ on Ω are closed under product and inverse and hence form a group (and in fact the mapping $g \to \bar{g}$ preserves products and inverses, it is a homomorphism).

In a parallel way the group $\bar{G}$ partitions Ω; let $\delta(\theta)$ be a function that indexes the sets of the partition. Then an invariant function of θ is an arbitrary function of $\delta(\theta)$ and we can call $\delta(\theta)$ the maximal invariant parameter. See Figure 13.

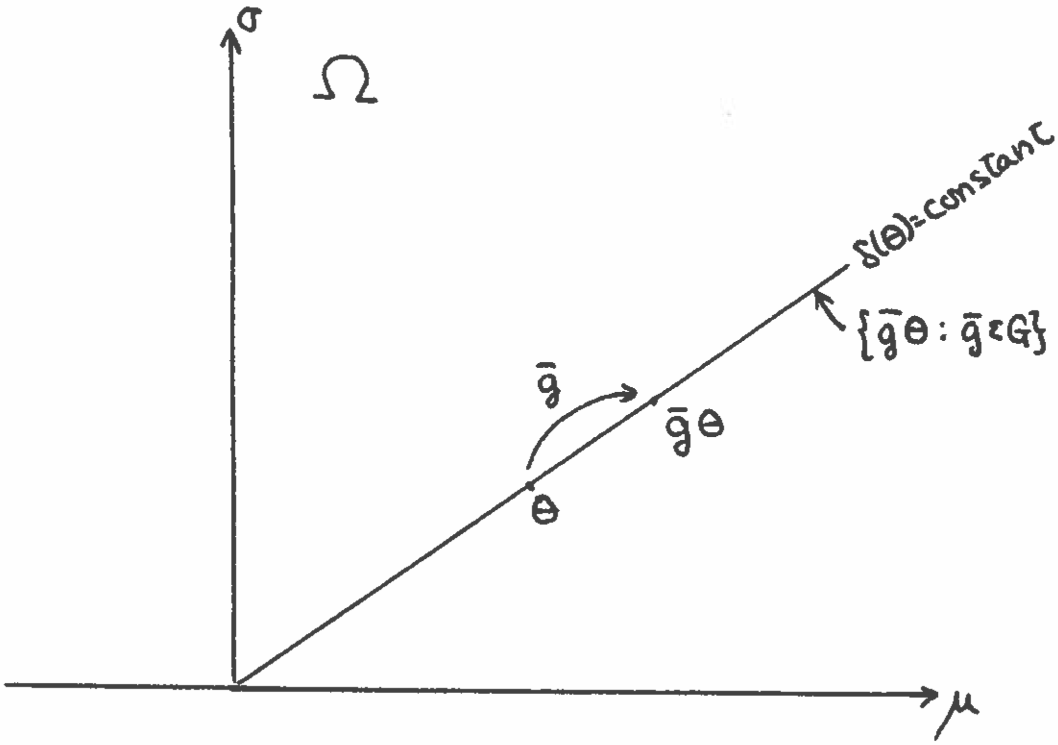
σ
Ω
$\delta(\theta)$=constant
$\{\bar{g}\theta : \bar{g}\varepsilon G\}$
$\bar{g}$
$\bar{g}\theta$
θ
μ

Figure 13. The groups $\bar{G}$ partitions the space Ω into sets $\bar{G}\theta$

We can now record a simple lemma.

Lemma. *The distribution of* $m(y)$ *depends only on* $\delta(\theta)$.

Proof. From the invariance we have

$$m(y) = m(gy) = m(\tilde{y})$$

where $\tilde{y} = gy$. If y has distribution θ, then $\tilde{y}$ has distribution $\bar{g}\theta$. From the preceding equation m has the same distribution whether derived from θ or from any other parameter value $\bar{g}\theta$ with the same value for $\delta(\theta)$. Hence the distribution of $m(y)$ depends only on $\delta(\theta)$.

Now suppose that the transformation group is consonant with the hypothesis testing problem, specifically,

(ii) *If* θ *is in* H_0 *or* H_1 , *then correspondingly* $\bar{g}\theta$ *is in* H_0 *or* H_1 .

In other words the group leaves the hypothesis testing problem invariant.

Now with conditions (i) and (ii) we find that both the model and the hypothesis testing problem are the same whether examined in terms of y or in terms of some transformed gy . It is reasonable then to restrict our attention to tests ϕ that give the same result for the various possible expressions for the response. We thus consider *invariant test functions* ϕ,

$$\phi(gy) = \phi(y),$$

or equivalently consider test functions that can be written

$$\psi(m(y))$$

as a function of the maximal invariant. From the lemma we note that the power function of an invariant test can be expressed in terms of the maximal invariant parameter $\delta(\Theta)$.

(b) An illustration.

As an illustration of the invariance arguments consider further the normal location example discussed at the beginning of this section:

Example 1 continued. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ,σ^2) and consider the hypothesis testing problem

$$H_0 : \mu \leq 0 \ , \ \sigma^2 \in \mathbb{R}^+$$
$$H_1 : \mu > 0 \ , \ \sigma^2 \in \mathbb{R}^+ \quad .$$

By the discussion at the beginning of this section it suffices to examine tests based on the likelihood statistic $(\bar{y} \ , \ s(\underline{y}))$. And then by the discussion of invariance it is reasonable to restrict our attention to tests based on the maximal invariant

$$t = \frac{\sqrt{n}\bar{y}}{s_y} = \frac{\sqrt{n}\ \bar{y}}{s(y)/(n-1)^{\frac{1}{2}}}$$

(a more convincing route would be to go to the maximal invariant on $\mathbb{R}^n$ and then to the corresponding likelihood function; it leads however to the same test based on t ; see Problem 93).

By the discussion at the beginning of this section and by Figure 12 it follows that the maximal invariant parameter is $\delta(\mu,\sigma^2) = \mu/\sigma$. Then by the lemma the distribution of t depends only on δ (see Problem 66 and be aware of the slightly different δ) . Let $f(t|\delta)$ be the density for t given δ .

We now consider the tests based on t and find the most powerful test of $\delta = 0$ against $\delta = \delta_1 > 0$:

$$\begin{aligned} \psi(t) &= 1 \qquad \text{if } \frac{f(t|\delta_1)}{f(t|0)} > k \\ &= 0 \qquad\qquad\qquad < k \quad . \end{aligned}$$

This can be calculated directly but more easily by using results from the preceding section. Let $S^2 = n\bar{y}^2 + s^2(\underset{\sim}{y}) = \Sigma y_i^2$, and $g(t:S|\mu,\sigma^2)$ and $h(S|\mu,\sigma^2)$ be the conditional density of t given S and the marginal density of S . By the example in Section 6 the likelihood ratio

$$\frac{g(t:S|\mu,\sigma^2)}{g(t:S|0,\sigma^2)}$$

with $\mu > 0$ is monotone increasing in t and the denominator

$$g(t:S|0,\sigma^2) = g_0(t)$$

is independent of S (also of σ^2) and hence is the marginal density, the ordinary t density. Thus

$$\frac{f(t|\delta_1)}{f(t|0)} = \frac{\int_0^\infty g(t{:}S|\mu,\sigma^2)h(S|\mu,\sigma^2)dS}{g_0(t)}$$

$$= \int_0^\infty \frac{g(t{:}S|\mu,\sigma^2)}{g_0(t)} h(S|\mu,\sigma^2)dS$$

is an average of increasing functions of t and hence is an increasing function of t. Hence

$$\begin{aligned} \psi(t) &= 1 \qquad && t > t_\alpha \\ &= 0 && < t_\alpha \end{aligned}$$

where the constant t_α is the value exceeded with probability α by the t variable on $n-1$ degrees of freedom.

The test ψ has size α for the enlarged hypothesis H_0; this was verified in the preceding section. Thus ψ is the most powerful (at δ_1) invariant size α test of H_0. But the test does not depend on δ_1; hence it is UMP invariant for H_0 against H_1.

In a similar manner for the problem

$$H_0 : \mu = 0 , \quad \sigma^2 \in \mathbb{R}^+$$

$$H_1 : \mu \neq 0 , \quad \sigma^2 \in \mathbb{R}^+ ,$$

we can consider the transformations

$$G = \{[0,c] : c \neq 0\} \quad .$$

The invariant tests are based on

$$|t| = \left|\frac{\sqrt{n}\bar{y}}{s_y}\right| \; ;$$

and the uniformly most powerful invariant test is to reject for large values of $|t|$.

(c) Testing several parameters

Now consider the invariance methods for testing several parameters at once.

Example 6 continued. Consider the linear model

$$\underset{\sim}{y} = X_r \underset{\sim}{\beta} + \underset{\sim}{u}$$

as examined in the preceding section. Suppose that X_r consists of r linearly independent column vectors and that $\underset{\sim}{y}$ is a sample from the normal $(0,\sigma_0^2)$. And suppose we want the uniformly most powerful invariant size α test for the problem $H_0 : \beta_r = \dots \beta_{r-s+1} = 0$ against the alternative that not all of these β's are zero.

The linear model as given allows the mean to be any point in the space $L(X_r)$. The hypothesis restricts the mean to $r-s$ dimensional space $L(X_{r-s})$ formed from the first $r-s$ column vectors X_r (by appropriate reexpression one can test any $r-s$ dimensional subspace).

Consider a new set of axes (orthonormal) whose first $r-s$ axes lie in $L(X_{r-s})$ and whose next s axes lie in $L(X_r)$; the new coordinates are obtained by orthogonal transformation. The linear model then becomes

$$\begin{aligned} y_1 &= \nu_1 + u_1 \\ &\vdots \\ y_r &= \nu_r + u_r \\ y_{r+1} &= \quad u_{r+1} \\ &\vdots \\ y_n &= \quad u_n \end{aligned}$$

with the testing problem H_0 : $\nu_r = \ldots = \nu_{r-s+1} = 0$ against the alternative that not all these ν's are zero.

As a group of transformations consider: arbitrary translations of $y_1, \ldots, y_{r-s}$; and arbitrary orthogonal transformations of $y_{r-s+1}, \ldots, y_r$. The maximal invariant is

$$\left(\sum_{r-s+1}^{r} y_j^2 \; , \; y_{r+1}, \ldots, y_n \right) \quad ;$$

and the likelihood statistic for the maximal invariant is $\sum_{r-s+1}^{r} y_j^2$. The most powerful test against any parameter point in the alternative is (Problem 95)

$$\begin{aligned}\phi(\underset{\sim}{y}) &= 1 \qquad\qquad \frac{S^2}{\sigma_0^2} \geq \chi^2_\alpha \\ &= 0 \qquad\qquad\quad\; < \chi^2_\alpha\end{aligned}$$

where $S^2 = \sum_{r-s+1}^{r} y_j^2$ and χ^2_α is the α point on the tail of the chi-square distribution on $r-s$ degrees of freedom.

The component $\sum_{r-s+1}^{r} y_j^2$ of the test function can be given direct interpretation. The hypothesis specifies a model with parameters $\nu_1,\ldots,\nu_{r-s}$; the sum of squares of deviation from the fitted model is $\sum_{r-s+1}^{n} y_j^2$. The general model has parameters $\nu_1,\ldots, \nu_r$; the sum of squares of deviations from this model general fitted model is $\sum_{r+1}^{n} y_j^2$. The reduction in sum of squares in going to the more general model is the component $\sum_{r-s+1}^{r} y_j^2$. It follows that the UMP invariant test for the original problem is as given above but with S^2 now defined by

$$S^2 = \underset{\sim}{y}'X_r(X_r'X_r)^{-1}X_r'\underset{\sim}{y} - \underset{\sim}{y}'X_{r-s}(X_{r-s}X_{r-s}')^{-1}X_{r-s}'\,\underset{\sim}{y}$$

using the formulas of Section 1(c) .

Example 7 continued. Now consider the linear model

$$\underset{\sim}{y} = X_r\underset{\sim}{\beta} + \sigma\underset{\sim}{u}$$

where $\underset{\sim}{u}$ is a sample from the standard normal. And suppose that we want the UMP invariant size α test for $H_0 : \beta_r = \dots = \beta_{r-s+1} = 0$ against the alternative that not all these β's are zero.

Consider a new set of axes as in the preceding.

Example 6. The linear model becomes

$$\begin{aligned} y_1 &= \nu_1 + \sigma u_1 \\ &\vdots \\ y_r &= \nu_r + \sigma u_r \\ y_{r+1} &= \sigma u_{r+1} \\ &\vdots \\ y_n &= \sigma u_n \end{aligned}$$

with the testing problem $H_0 : \nu_r = \dots = \nu_{r-s+1} = 0$.

As a group of transformations consider: arbitrary translations of $y_1, \dots, y_{r-s}$; arbitrary orthogonal transformations of $y_{r-s+1}, \dots, y_r$; arbitrary orthogonal transformations of $y_{r+1}, \dots, y_n$; an arbitrary scale transformation applied to all the y's . The maximal invariant is

$$G = \frac{\sum_{r-s+1}^{r} y_j^2}{\sum_{r+1}^{n} y_j^2}$$

or equivalently

$$F = \frac{\sum_{r-s+1}^{r} y_j^2 \; / \; (r-s)}{\sum_{r+1}^{n} y_j^2 \; / \; (n-r)}$$

The most powerful test against any parameter point in the alternative is (Problem 97)

$$\begin{aligned} \phi(\underset{\sim}{y}) &= 1 \qquad\qquad F \geq F_\alpha \\ &= 0 \qquad\qquad\; < F_\alpha \end{aligned}$$

where F_α is α point on the tail of the F distribution on $r-s$ over $n-r$ degrees of freedom.

The components for F can be given general interpretation and it follows that the UMP invariant size test for the original problem is as given above but with F defined by

$$F = \frac{y'X_r(X_r'X_r)^{-1}X_r'\underset{\sim}{y} - \underset{\sim}{y}'X_{r-s}(X_{r-s}X_{r-s}')^{-1}X_{r-s}'\underset{\sim}{y} \quad / \; (r-s)}{(\underset{\sim}{y}'\underset{\sim}{y} - \underset{\sim}{y}'X_r(X_r'X_r)^{-1}X_r'\underset{\sim}{y}) \; / \; n-r}$$

(d) Large sample tests

Now consider a statistical model with parameter space Ω where Ω is an open set of $\mathbb{R}^r$; and suppose that we have a hypothesis testing problem H_0 against $H_1 = \Omega - H_0$ where H_0 is an $r-s$ dimensional region, a continuously differentiable surface in Ω . A natural extension of the likelihood ratio is

$$L(y) = \frac{\sup\limits_{\Theta\varepsilon\Omega} f(y|\underset{\sim}{\Theta})}{\sup\limits_{\Theta\varepsilon H_0} f(y|\underset{\sim}{\Theta})} = \frac{f(y|\underset{\sim}{\hat{\hat{\Theta}}}(y))}{f(y|\underset{\sim}{\hat{\Theta}}(y))}$$

where $\underset{\sim}{\hat{\hat{\Theta}}}(y)$ is now the maximum likelihood estimate under the model as given and $\underset{\sim}{\hat{\Theta}}(y)$ is the maximum likelihood estimate under the restricted model specified by the hypothesis. A reasonable test then is to reject for large value of $L(y)$. Fortunately the large sample distribution of $L(y)$ is available.

Let $(y_1, \ldots, y_n)$ be a sample from a statistical model satisfying the assumptions in Seven, Section 5. Then for large samples the $(\underset{\sim}{\Theta}_0)$ distribution for the likelihood function (relative to the true value $\underset{\sim}{\Theta}_0$) is given by

$$l_n(\underset{\sim}{y}|\Theta) = -\tfrac{1}{2}(\underset{\sim}{\hat{\hat{\delta}}}-\underset{\sim}{\delta})'\, I(\underset{\sim}{\Theta}_0)(\underset{\sim}{\hat{\hat{\delta}}}-\underset{\sim}{\delta}) + \tfrac{1}{2}\hat{\hat{\delta}}'\, I(\underset{\sim}{\Theta}_0)\underset{\sim}{\hat{\hat{\delta}}}$$

where the maximum likelihood estimate $\hat{\hat{\Theta}}$ has the multivariate normal distribution with mean zero and inverse matrix $I(\Theta_0)$ and $\underset{\sim}{\Theta} = \underset{\sim}{\Theta}_0 + \underset{\sim}{\delta} n^{-\frac{1}{2}}$.

If the information matrix $I(\underset{\sim}{\Theta}_0)$ is the identity matrix then the large sample likelihood becomes

$$l_n(\underset{\sim}{y}|\underset{\sim}{\Theta}) = -\tfrac{1}{2}\sum_1^r(\hat{\hat{\delta}}_i-\delta_i)^2 + \tfrac{1}{2}\sum_1^r \hat{\hat{\delta}}_i^2$$

and

$$\sup_{\underset{\sim}{\Theta}\in\Omega} l_n(\underset{\sim}{y}|\underset{\sim}{\Theta}) = \tfrac{1}{2}\sum_{1}^{r} \hat{\hat{\delta}}_i^2 \ .$$

~~But this large sample likelihood agrees with the likelihood in Example 6 in part (c)~~ .

Now consider the hypothesis H_0 and suppose that $\underset{\sim}{\Theta}_0$ is in H_0 . For simplicity of notation suppose that the surface H_0 at $\underset{\sim}{\Theta}_0$ corresponds to $\delta_{r-s+1} = \dots = \delta_r = 0$. Then the likelihood function along the surface H_0 becomes

$$l_n(\underset{\sim}{y}|\underset{\sim}{\Theta}) = -\tfrac{1}{2}\sum_{1}^{r-s}(\hat{\delta}_i-\delta_i)^2 + \tfrac{1}{2}\sum_{1}^{r-s}\hat{\delta}_i^2$$

and

$$\sup_{\underset{\sim}{\Theta}\in H_0} l_n(\underset{\sim}{y}|\underset{\sim}{\Theta}) = \tfrac{1}{2}\sum_{1}^{r-s}\hat{\delta}_i^2 = \tfrac{1}{2}\sum_{1}^{r-s}\hat{\hat{\delta}}_i^2 \ .$$

(Of course the maximum likelihood estimates agree for the first $r-s$ coordinates).

Hence the generalized log-likelihood ratio is

$$\ln L(\underset{\sim}{y}) = \sup_{\Omega} l_n(\underset{\sim}{y}|\Theta) - \sup_{H_0} l_n(\underset{\sim}{y}|\underset{\sim}{\Theta})$$

$$= \sum_{r-s+1}^{r} \hat{\hat{\delta}}_i^2 \ .$$

Under the hypothesis H_0 , $2\ln L(\underset{\sim}{y})$ has a chi-square distribution on $r-s$ degrees of freedom.

By using the results from Example 6 it follows that ~~a~~ the UMP invariant size α test based on the large sample likelihood is to reject if $2\ln L(\underset{\sim}{y})$ exceeds the α point in the tail of the chi-square distribution on $r-s$ degrees of freedom.

This result does not depend on the information matrix being the identity matrix as assumed earlier. The standard multivariate normal becomes the general multivariate by a linear change of coordinates; and conversely an appropriate linear transformation changes the general multivariate normal to the standard multivariate normal (One: Problem 88 and Seven, Section 1). A linear change of coordinates about $\underset{\sim}{\Theta}_0$ does not change $L(\underset{\sim}{y})$ but it can be chosen to change $I(\Theta_0)$ to the identity.

Problems.

90. Let $(y_{11}, \ldots, y_{1m})$ be a sample from the normal (μ_1, σ_1^2), and $(y_{21}, \ldots, y_{2n})$ be a sample from the normal (μ_2, σ_2^2). Find the UMP invariant size α test of $H_0 : \sigma_1 = \sigma_2$ against $H_1 : \sigma_1 > \sigma_2$. Compare with Problem 80.

91. Let $(y_{11}, \ldots, y_{1m})$ be a sample from the normal (μ_1, σ^2), and $y_{21}, \ldots, y_{2n})$ be a sample from the normal (μ_2, σ^2). Show that the UMP invariant size α test of $H_0 : \mu_1 = \mu_2$, $\sigma^2 = \sigma_0^2$ against $H_1 : \mu_1 \neq \mu_2$, $\sigma^2 = \sigma_0^2$ is to reject if

$$\frac{(\bar{y}_1 - \bar{y}_2)^2 / \left(\frac{1}{m} + \frac{1}{n}\right)}{\sigma_0^2}$$

exceeds χ^2_α the α point of the tail of the chi-square distribution on 1 degree of freedom.

92. (Continuation). Show that the UMP invariant size α test of $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ is to reject if

$$\frac{(\bar{y}_1-\bar{y}_2)^2/\left(\frac{1}{m}+\frac{1}{n}\right)}{\left(\Sigma(y_{1i}-\bar{y}_1)^2+\Sigma(y_{2j}-\bar{y}_2)^2\right)/(m+n-2)}$$

exceeds the α point on the tail of the F distribution on 1 over $m+n-2$ degrees of freedom.

93. Let $f(y|\Theta)$ with y in $\mathbb{R}^n$ and Θ in Ω be a statistical model that represents probability *density*. Let G be a group of continuously differentiable transformations satisfying the assumption (i) .

(a) Show that $f(gy|\bar{g}\Theta)J(g:y)dy = f(y|\Theta)dy$ where $J(g:y) = |\partial gy/\partial y|$. Hence deduce that

$$L(gy|\cdot) = L(y|\bar{g}^{-1}\cdot) \ .$$

(b) For any function m on Ω, let $\tilde{g}m$ be the function whose value at the point Θ is $m(\bar{g}^{-1}\Theta)$. Hence show that

$$L(gy|\cdot) = \tilde{g}L(y|\cdot)$$

and deduce that $\tilde{G} = \{\tilde{g}\}$ is a group on the set of likelihood functions.

94. Noncentral chi-square distribution. Let $w_1, \ldots, w_k$ be independent normal with means $\delta_1, \ldots, \delta_k$ and unit variance. The dsstribution of $\chi^2 = w_1^2 + \ldots + w_k^2$ is called the noncentral chi-square distribution on k degrees of freedom and with noncentrality $\delta^2 = \Sigma\delta_i^2$.

(a) Let A be an orthogonal transformation with first row vector in the direction $(\delta_1, \ldots, \delta_k)$ and let $\underset{\sim}{y} = A\underset{\sim}{w}$ (see SEVEN, Section 2). Deduce that $\chi^2 = y_1^2 + \ldots + y_k^2$ where the $y_1, \ldots, y_k$ are independent normal with means $(\delta, 0, \ldots, 0)$ and unit variance.

(b) Show that the probability differential for y_1^2 is

$$\sum_{j=0}^{\infty} \exp\left\{-\frac{\delta^2}{2}\right\} \frac{(\delta^2/2)^j}{j!} \quad f_{2j+1} \; (y_1^2) \; dy_1^2$$

where f_j is the chi-square density on j degrees of freedom; compare with Problem 65.

(c) Deduce that the probability differential for χ^2 is

$$\sum_{j=0}^{\infty} \exp\left\{-\frac{\delta^2}{2}\right\} \frac{(\delta^2/2)^j}{j!} \; f_{2j+k}(\chi^2)d\chi^2 \quad .$$

95. (Continuation). Consider Example 6 in part (b).

(a) Show that the general distribution of S^2/σ_0^2 is noncentral chi-square on $r-s$ degrees of freedom with noncentrality parameter $\delta^2 = \sum_{r-s+1}^{r} \nu_j^2$.

(b) Show that the most powerful test of $\delta = 0$ against $\delta = \delta_1$ is to reject for large values of S^2/σ_0^2 .

96. The noncentral F distribution (continuation of Problem 94). Let $G = \chi_1^2/\chi_2^2$ be the quotient of independent variables where χ_1^2 is noncentral chi-square on f_1 degrees of freedom and χ_2^2 is ordinary chi-square on f_2 degrees of freedom. Show that the probability differential for G is

$$\sum_{j=0}^{\infty} \exp\left\{-\frac{\delta^2}{2}\right\} \frac{(\delta^2/2)^j}{j!} f_{f_1+2j:f_2}(G)\,dG$$

where $f_{j,k}$ is the canonical F density $(p=j/2,\ q=k/2)$ in ONE, Problem 70. Compare with Problem 80.

97. Consider Example 7 in Part (b). Show that the UMP test based on G is to reject for large values of G .

8. CONFIDENCE REGIONS

Confidence intervals have been obtained for a variety of problems in Chapters SIX and SEVEN. As an illustration consider again a sample $(y_1, \ldots, y_n)$ from the normal distribution with (μ,σ^2) in $\Omega = \mathbb{R} \times \mathbb{R}^+$.
The quantity

$$t = \frac{\bar{y}-\mu}{s_y/\sqrt{n}}$$

has a t-distribution on $n-1$ degrees of freedom as obtained from the (μ,σ^2) distribution on $\mathbb{R}^n$; let $\left(-t_{\alpha/2}\ ,\ t_{\alpha/2}\right)$ be the central interval containing $1-\alpha$ probability. Then

$$P\left(-t_{\alpha/2} < \frac{\bar{y}-\mu}{s_y/\sqrt{n}} < t_{\alpha/2} \middle| \mu,\sigma^2\right) = 1-\alpha$$

or equivalently

$$P\left(\bar{y} - t_{\alpha/2}\,\frac{s_y}{\bar{n}} < \mu < \bar{y} + t_{\alpha/2}\,\frac{s_y}{\sqrt{n}} \middle| \mu\ ,\sigma^2\right) = 1-\alpha$$

Hence the probability is $1-\alpha$ that the interval $\bar{y} \pm t_{\alpha/2}\,\dfrac{s_y}{\sqrt{n}}$ will bracket the true μ in any application; thus the interval is a $1-\alpha$ *confidence interval for* μ.

We can view the relationship

$$-t_{\alpha/2} < \frac{\bar{y}-\mu}{s_y/\sqrt{n}} < t_{\alpha/2}$$

as defining the acceptance region

$$A(\mu) = \left\{ \underset{\sim}{y} : -t_{\alpha/2} < \frac{\bar{y}-\mu}{s_y/\sqrt{n}} < t_{\alpha/2} \right\}$$

of a size α test for a hypothesized value μ; the corresponding critical region is

$$C(\mu) = \left\{ \underset{\sim}{y} : \frac{\bar{y}-\mu}{s_y/\sqrt{n}} \geq t_{\alpha/2} \right\} \quad .$$

Then for any particular response $\underset{\sim}{y}$ we see that the interval $(\bar{y} - t_\alpha \; s_y/\sqrt{n} \; , \; \bar{y} + t_\alpha \; s_y/\sqrt{n} \;)$ consists of those parameter values μ that are acceptable by the test regions just described. This connection with hypothesis testing holds generally.

(a) Derivation of confidence regions

Consider a statistical model with response y in a sample space S and with parameter Θ in a parameter space Ω . For any hypothesis $H_0 : \Theta = \Theta_*$ let $C(\Theta_*)$ be the critical region for a test of size α and let $A(\Theta_*)$ be the corresponding acceptance region; then

$$P(C(\Theta)|\Theta) \leq \alpha \qquad \text{for all } \Theta \text{ in } \Omega \; ,$$

$$P(A(\Theta)|\Theta) \geq 1-\alpha \qquad \text{for all } \Theta \text{ in } \Omega \; .$$

This can be presented as in Figure 14. The Θ-distribution for

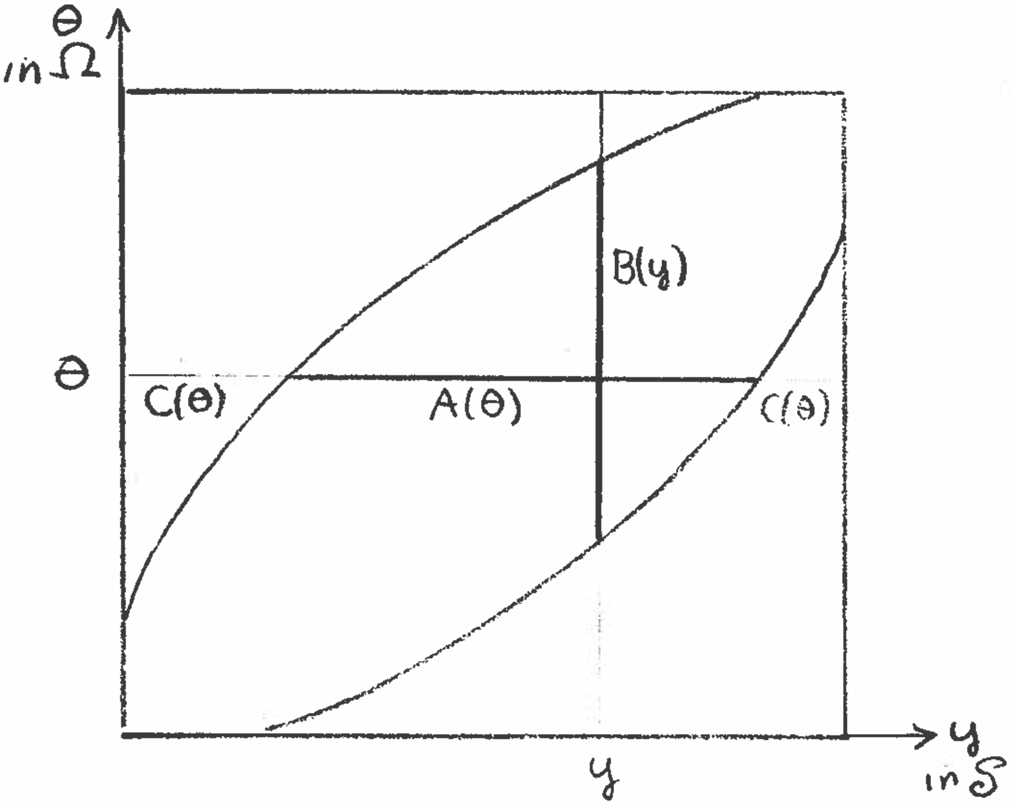

Figure 14. The θ distribution for y can be pictured as a distribution along the θ-section of 𝒮 × Ω ; let C(θ) and A(θ) be the critical region and acceptance region for a size α test for the *value* θ . The composite acceptance region A . The y-section of A provides a 1-α confidence region for θ .

for y is viewed as a distribution on the θ-section of $\mathcal{S} \times \Omega$. The critical region $C(\theta)$ and the complementary acceptance region $A(\theta)$ are presented as sets on this section; the set $A(\theta)$ collects at least $1-\alpha$ of the probability on the section.

Now let A be the composite acceptance region

$$A = \{(y,\theta) : y \in A(\theta)\} \quad ,$$

and $B(y)$ be the y-section of A.

$$B(y) = \{\theta : (y,\theta) \in A\} \quad .$$

Then the following relations are equivalent

$$y \in A(\theta) \leftrightarrow (y,\theta) \in A \leftrightarrow \theta \in B(Y) \quad .$$

Hence

$$P(A(\theta)|\theta) = P(\theta \in B(y)|\theta) \geq 1-\alpha \quad .$$

Thus the probability is at least $1-\alpha$ that the random region $B(y)$ will contain the true θ in any application, and accordingly $B(y)$ is a $1-\alpha$ confidence region for θ :

ID *A function* $B(y)$ *from the points of* $\mathcal{S}$ *to the subsets of* Ω *is a* $1-\alpha$ *confidence region for* θ *if*

$$P(\theta \in B(y)|\theta) \geq 1-\alpha$$

for all θ *in* Ω .

In an application we can think of testing each of the possible Θ values and forming a region of those that are acceptable. In particular the true value will be tested and it will be included in the region if it was acceptable by the test for it. Thus if the tests have size α the region formed will have probability $1-\alpha$ of containing the true parameter value.

A pivotal quantity provides a convenient method of constructing a confidence region.

ID *A pivotal quantity*

$$t = q(y,\Theta)$$

is a function on $S \times \Omega$ *that has a fixed distribution* (independent of Θ) *as derived from the* Θ *distribution on* S.

Let T be a set having $1-\alpha$ probability according to the fixed distribution. Then

$$A(\Theta) = \{y : q(y,\Theta) \in T\}$$

is a $1-\alpha$ acceptance region for testing the value Θ, and

$$B(y) = \{\Theta : q(y,\Theta) \in T\}$$

is a $1-\alpha$ confidence region for the parameter Θ. Note that the indicator function for the set A in Figure 14 is a pivotal quantity under certain circumstances (if the inequalities are in fact equalities). Towards obtaining a reasonable confidence interval it is natural to choose a pivotal quantity that is a function of the likelihood statistic.

Note that any $1-\alpha$ confidence region $B(y)$ in a converse way determines an acceptance region

$$A(\theta) = \{y : \theta \varepsilon B(y)\}$$

for a size α test of the value θ.

(b) Power of a confidence region.

Consider a $1-\alpha$ confidence region $B(y)$ that has been derived from a family of size α tests. Let

$$\mathcal{P}_{\theta^*}(\theta) = P(C(\theta^*)|\theta)$$

be the power function of the test for the hypothesis $H_0 : \theta = \theta_*$. Then the power

$$\begin{aligned}\mathcal{P}_{\theta^*}(\theta) &= P(S - A(\theta^*)|\theta) \\ &= P(\theta^* \notin B(y)|\theta)\end{aligned}$$

is the probability that the confidence region does not contain θ^* when the true value is θ. Thus the power function of a test becomes the probability that the confidence region does not cover a value θ^* . For all θ values save θ^* this is the probability of not covering a wrong value. For certain exceptance problems a UMP $1-\alpha$ confidence region may be found; see Problem 75.

(c) Unbiased confidence regions

A confidence region at level $1-\alpha$ has the property that the probability of not covering the true value is at most α. The interpretation of power in the preceding subsection shows that:

D A $1-\alpha$ *confidence region is unbiased if*

$$P(\Theta^* \varepsilon B(y)|\Theta) \geq \alpha \ , \ \Theta^* \neq \Theta \quad ;$$

Thus unbiasedness means that the probability of not covering wrong values is at least α .

Example 3 continued. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ,σ_0^2) . Then by Section 6

$$\bar{y} \pm Z_{\alpha/2}\, \sigma_0/\sqrt{n}$$

is a $1-\alpha$ unbiased confidence region for Θ.

(d) Invariant confidence regions

The extension of invariance to confidence regions is slightly more complicated than the preceding extension of unbiased.

Let $G(\Theta_*)$ be a group of transformations on the sample space $\mathcal{S}$ which is invariant for testing $H_0 : \Theta = \Theta_*$ against $H_1 : \Theta \neq \Theta_*$; then

$$\bar{g}\Theta_* = \Theta_*$$

for all g in $G(\theta_*)$. And let G be the smallest group of transformations that includes all the transformations in the various $G(\theta_*)$'s . And let $\bar{G}$ be the corresponding group on the parameter space Ω. Then

D *The confidence region* $B(y)$ *is invariant relative to* G *if*

$$B(gy) = \bar{g}B(y)$$

for all y *in* S *and* g *in* G.

Note the interpretation of this definition. A transformation g changes a θ distribution for y to a $\bar{g}\theta$ distribution for gy . If a response value y suggests the parameter values in $B(y)$, then it is reasonable that the response value gy should suggest the parameter values in

$$\bar{g}B(y) = \{\bar{g}\theta : \theta \text{ in } B(y)\} .$$

A family of invariant tests need not produce an invariant confidence region. But

Lemma. *An invariant confidence region* $B(y)$ *produces a family of invariant tests.*

Proof. Let g be an element of $G(\theta_*)$; thus $\bar{g}\theta_* = \theta_*$. Then the invariance of $B(y)$ gives the equivalence

$$\theta^* \text{ in } B(y) \leftrightarrow \theta^* \text{ in } B(gy) .$$

But this means

$$y \text{ in } A(\Theta^*) \leftrightarrow gy \in A(\Theta^*)$$

and hence that $A(\Theta^*)$ is the acceptance region of a test invariant under $G(\Theta_*)$.

From the lemma it follows that if a family of UMP invariant tests produces an invariant confidence region then the confidence region is UMP invariant.

Example 3 continued. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ, σ_0^2) . Then by Section 7

$$\bar{y} \pm z_\alpha \sigma_0/\sqrt{n}$$

is the uniformly most powerful invariant $1-\alpha$ confidence region; this is relative to the group

$$G = \{[a,C] : a \in \mathbb{R}, \quad C = \pm 1\}$$

of location-reversing transformations and the invariance is easily checked.

(e) Confidence regions for component parameters

Consider a statistical model with response y in a sample space $\mathcal{S}$ and with parameter Θ in a parameter space Ω and suppose that we are interested in a derived parameter $\delta(\Theta)$ with values in Δ . For any hypothesis $H_0 : \delta(\Theta) = \delta_*$ let $C(\delta_*)$ be a critical region for a test of size α and let $A(\delta_*)$ be the corresponding acceptance region. Then

$$P(C(\delta(\Theta))|\Theta) \leq \alpha \qquad \text{for all } \Theta \text{ in } \Omega ,$$
$$P(A(\delta(\Theta))|\Theta) \geq 1\text{-}\alpha \qquad \text{for all } \Theta \text{ in } \Omega .$$

Now let

$$B(y) = \{\delta : y \in \delta(\Theta)\} .$$

Then

$$y \in A(\delta) \leftrightarrow \delta \in B(y)$$

and it follows that

$$P(\delta(\Theta) \in B(y)|\Theta) \geq 1-\alpha$$

for all Θ in Ω . The interpretation of power, unbiasedness, and invariance carries over in a straightforward manner.

Example 1 continued. Let $(y_1, \ldots, y_n)$ be a sample from the normal (μ,σ^2) in $\mathbb{R} \times \mathbb{R}^+$. Then by Section 6 and 7

$$\bar{y} \pm t_\alpha \frac{s_y}{\sqrt{n}}$$

is the $1-\alpha$ confidence interval that is UMP unbiased and UMP invariant.

(f) The binomial case.

Consider a sample $(x_1, \ldots, x_n)$ from the Bernoulli distribution with p in $[0,1]$. A size α

test for the value p can be formed by finding an acceptance interval $A(p) = (a_1(p) , a_2(p))$ such that

$$\sum_{y=a_1(p)}^{a_2(p)} \binom{n}{y} p^y (1-p)^{n-y} \geq 1-\alpha \quad .$$

Note that such a test is based on the likelihood statistic $y = \Sigma x_i$. Typically we will exclude up to $\alpha/2$ of the probability on each tail of the distribution. To obtain unbiasedness we need to use a randomized test, and hence some sort of graduated or randomized confidence interval. For confidence level $1-\alpha = 95\%$ and for various samples sizes ~~yes~~, the spectrum of acceptance regions is plotted in Figure 15 (in terms of the apparent probability $\hat{p} = y/n$) . A 95% confidence region is then obtained from the appropriate vertical section: find the observed $\hat{p}$ on the horizontal scale and obtain the range of acceptable p values as the interval section between the curves labelled with the sample size.

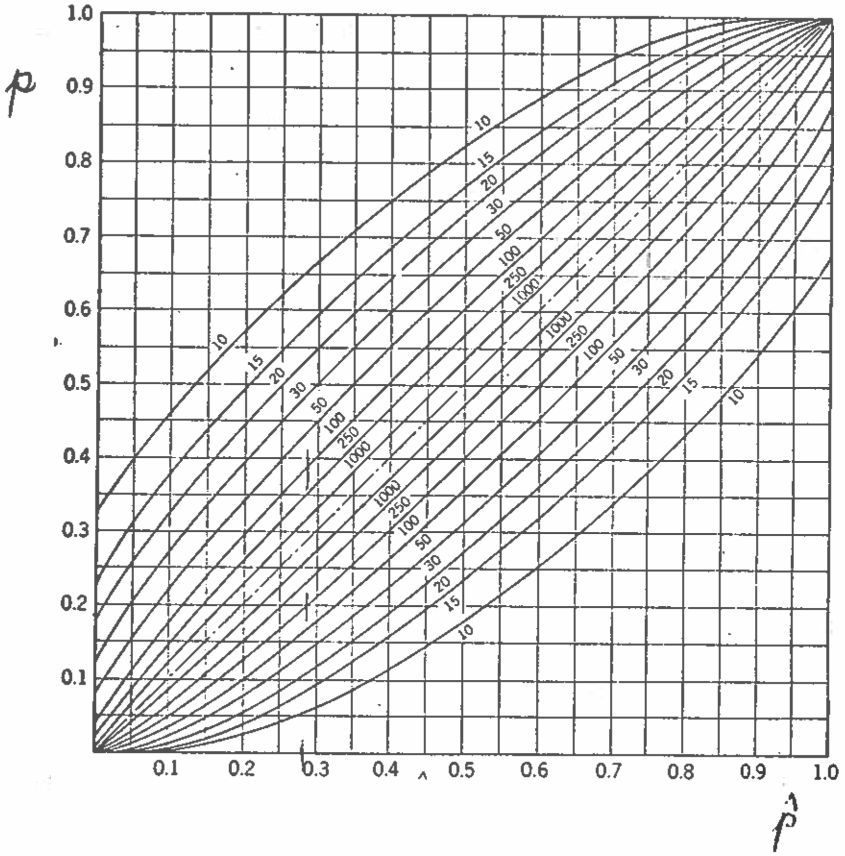

Figure 19. 95% confidence intervals for p when n = 10, 15, 20, 30, 50, 100, 250, 1000.

Problems

98. A sample of 5 was obtained from a response known to be normally distributed: 12.7, 13.3, 12.9, 13.0, 13.1 on the assumption that $\sigma = 0.06$, determine the UMP unbiased 95% confidence interval for μ .

99. (Continuation). With no information concerning σ determine the UMP unbiased 95% confidence interval for μ .

100. Two responses are known to be normally distributed with common variance. A sample of 5 from the first response gave $\bar{y}_1 = .275$ $s_1^2 = .00045$, and a sample of 6 from the second response gave $\bar{y}_2 = .295$, $s_2^2 = .00039$. Find UMP invariant 95% confidence interval for $\mu_1 - \mu_2$.

101. (Continuation). Find a central 95% confidence interval for the common variance σ^2. Use $2\frac{1}{2}\%$ on each tail; the adjustment to obtain unbiased is rarely pursued.

102. If s_1^2 is an unbiased estimate of σ_1^2 (chi-square type with df = f_1) and s_2^2 is an independent unbiased estimate of σ_2^2 (chi-square type with df = f_2) , show that

$$\left(\frac{s_1^2}{s_2^2} F^{-1}_{2\frac{1}{2}\%}(f_1, f_2) \; , \; \frac{s_1^2}{s_2^2} F_{2\frac{1}{2}\%}(f_2, f_1) \right)$$

is a 95% confidence interval for σ_1^2/σ_2^2 where $F_\alpha(f_1, f_2)$ is the α point on the tail of the F distribution on f_1 over f_2 degrees of freedom.

103. (Continuation of Problem 9 assuming normal error and Example 7). Show that the UMP unbiased 95% confidence interval for β is

$$\hat{\beta} \pm t_{2\frac{1}{2}\%} \frac{s_E}{(\Sigma(x_i-\bar{x})^2)^{\frac{1}{2}}}$$

where $s_E^2 = \Sigma(y_i - \hat{\alpha} - \hat{\beta}x_i)^2/(n-2)$.

104. (continuation). An investigater may be interested in the value of x that has mean response equal to y_0 ; this value of x is a parameter γ that can be expressed in terms of α and β : $\gamma = (x_0-\alpha)/\beta$. Find a 95% confidence interval for this parameter; use the pivotal quantity

$$\frac{\hat{\alpha} + \hat{\beta}\gamma - y_0}{C(\gamma)s_E}$$

where $C^2(\gamma)\sigma^2$ is the variance of $\hat{\alpha} + \hat{\beta}\gamma$. Note: the bounds for the interval are roots of a quadratic equation.

105. Let $(y_1, \ldots, y_n)$ be a sample from the normal distribution (μ,σ^2) . Determine the form of the 90% confidence region determined from

$$P(\Sigma(y_i-\mu)^2/\sigma^2 \leq \chi^2_{10\%}) = 90\%$$

where $\chi^2_{10\%}$ is the 10% point on the tail of the chi-square distribution on n degrees of freedom.

Some estimation methods.

106. The methods of moments. Consider a statistical model $f(y|\Theta_1, \ldots, \Theta_r)$ for a response on $\mathbb{R}^1$; let $\mu_1(\Theta_1, \ldots, \Theta_r)$, $\mu_2(\Theta_1, \ldots, \Theta_r)$, ... be the first, second, ... moments of y . Now consider a sample $(y_1, \ldots, y_n)$ and let $m_1(\underset{\sim}{y})$, $m_2(\underset{\sim}{y})$, ... be the first second, ... moments of the sample $(m_r = y_i^r /n)$. The method of moments for estimating the parameters is to find $\Theta_1^*, \ldots, \Theta_r^*$ so that the first r (typically) model moments ~~by~~ $\mu_1(\Theta_1^*, \ldots, \Theta_r^*)$, $\mu_2(\Theta_1^*, \ldots, \Theta_r^*)$, ... are equal to the corresponding sample moments. For a sample $(y_1, \ldots, y_n)$ from the normal (μ, σ^2) find the method-of-moments estimates of μ and σ^2 .

107. Consistency. Let $(y_1, \ldots, y_n)$ be a sample from the model $f(y|\Theta)$. An estimate $t_n(y_1, \ldots, y_n)$ defined for any sample size n is said to be a consistent estimate of Θ if $\operatorname{plim} t_n = \Theta$ for all Θ . For a sample from a distribution on $\mathbb{R}$ with finite variance σ^2 , show that s_y^2 and

$$\hat{\sigma}^2 = \frac{1}{n} \Sigma (y_i - \bar{y})^2$$

are consistent estimates of σ^2 .

9. SEQUENTIAL ANALYSIS

In our consideration of estimation and testing methods we have so far taken the sample size as given. Certainly for any of the methods we can examine the properties and characteristics and then choose the sample size so that the particular properties of interest are at some desired level. For example in normal sampling with known variance we could choose the sample size so that the standard test has size 1% at $\mu = 75$ and then had power 95% at $\mu = 76$; see Problem 59.

In this section we consider a sequential test of θ_0 against θ_1 such that observations are taken one by one until the θ_1 to θ_0 likelihood ratio becomes extreme -- either above an upper bound suggesting *reject* or below a lower bound suggesting *accept*. The mathematics is very attractive and there are interesting approximations that make the method surprisingly accessible. For applications the method seems particularly appropriate to industrial acceptance sampling of incoming manufactured items and developmental screening of new materials, drugs, procedures.

(a) The sequential test.

Consider a response y with statistical model $f(y:\theta)$ and typically a real parameter θ in $\Omega = \mathbb{R}$. The wald sequential test is concerned nominally with a simple hypothesis $H_0 : \Theta = \Theta_0$ against a simple alternative $H_1 : \Theta = \Theta_1$ but effectively with the more reasonable $H_0 : \Theta \leq \Theta_*$ against $H_1 : \Theta > \Theta_*$ where Θ_* is some intermediate value. After n observations the test is based on the likelihood ratio

$$L_n(y_1, \ldots, y_n) = \frac{f(y_1|\Theta_1)}{f(y_1|\Theta_0)} \cdots \frac{f(y_n|\Theta_1)}{f(y_n|\Theta_0)} .$$

and to

$$\begin{array}{lll} \text{Reject } H_0 & \text{if} & B \leq L_n \\ \text{Continue sampling} & & A < L_n < B \\ \text{Accept } H_0 & & L_n < A \end{array}$$

where *continue* means to take another sample value and repeat the procedure; see Figure 15. The values A , B are chosen to give the test the desired properties. Alternatively the test can be expressed in logarithmic form with the advantages that changes in *log*-likelihood are additive; then after n observations the test is based on the log-likelihood

$$l_n(y_1, \ldots, y_n) = \ln \frac{f(y_1|\Theta_1)}{f(y_1|\Theta_0)} + \ldots + \ln \frac{f(y_n|\Theta_1)}{f(y_n|\Theta_0)}$$

$$= Z_1 + \ldots + Z_n$$

where

$$Z = \ln f(y|\Theta_1) - \ln f(y|\Theta_0)$$

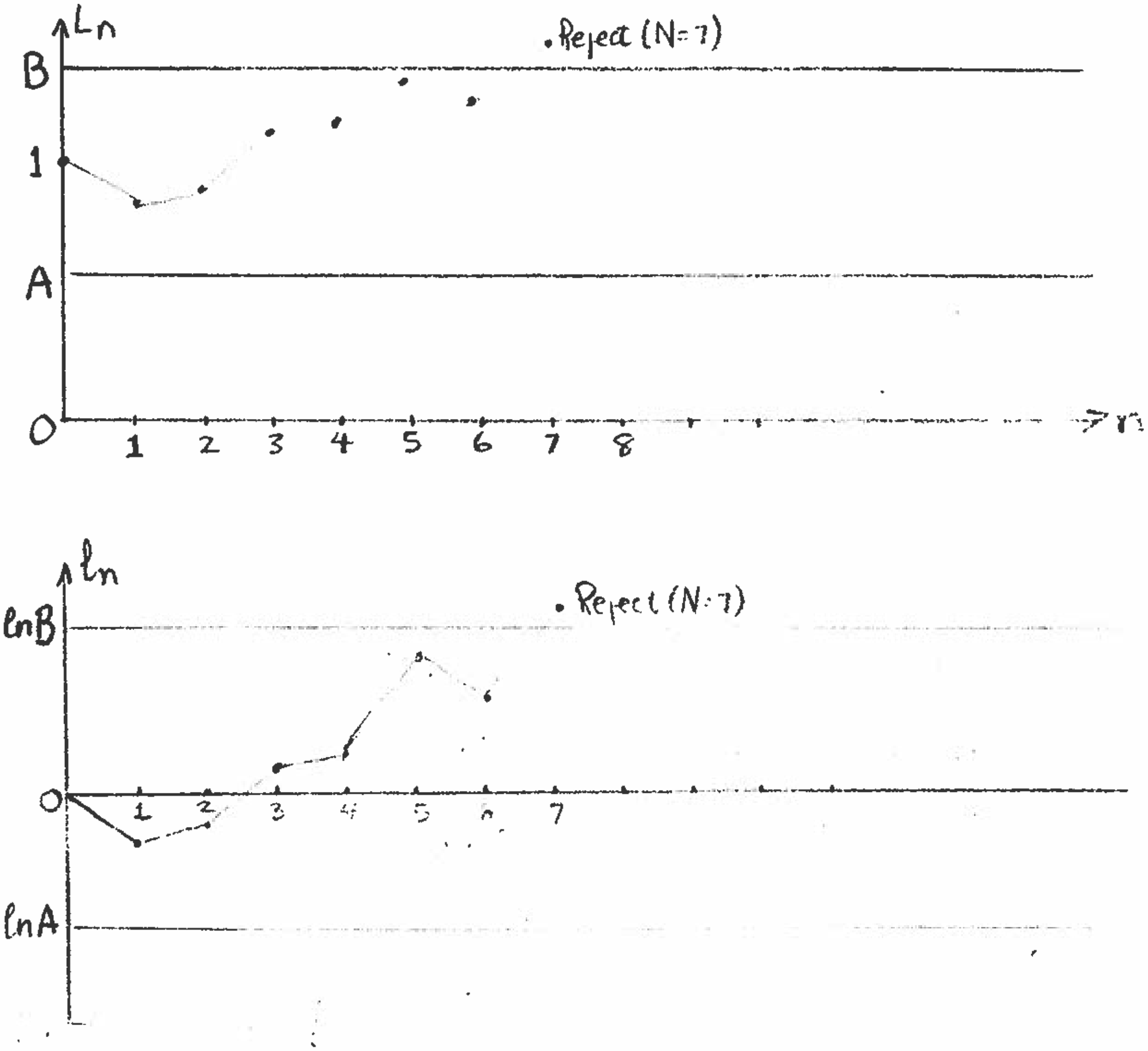

Figure 15. The likelihood or log-likelihood is calculated after each response value and sampling continues as long as it remains in a central range; when it interests the upper line the hypothesis is rejected; when it interests the lower line the hypothesis is accepted.

and to

$$\begin{array}{lll} \text{Reject } H_0 & \text{if} & \ln B \leq l_n \\ \text{Continue} & & \ln A < l_n < \ln B \\ \text{Accept } H_0 & & l_n \leq \ln A \quad . \end{array}$$

Iet N be the sample size at which a terminal decision Reject H_0 or Accept H_0 is made. N can vary in value and is a function on the countable product sample space of the problem.

The wald test has the following optimum property established by Wald and Worfowitz .

Theorem. *Among fixed and sequential two-decision procedures having*

$$P(\text{Reject } H_0 | \Theta_0) \leq \alpha \quad P(\text{Accept } H_0 | \Theta_1) \leq \beta$$

and finite $E(N|\Theta_0)$, $E(N|\Theta_1)$ *the wald sequential test with error probabilities* α *and* β *minimizes both* $E(N|\Theta_0)$ *and* $E(N|\Theta_1)$.

Consider the form of the sequential test for a simple example.

Example 3 continued. Let y be a response that is normal distributed (Θ,σ_0^2) and consider the sequential test of $H_0 : \Theta = \Theta_0$ against $H_1 : \Theta = \Theta_1$.

$$
\begin{aligned}
Z = \ln \frac{f(y|\Theta_1)}{f(y|\Theta_0)} &= \frac{1}{2\sigma_0^2}\,[(y-\mu_0)^2 - (y-\mu_1)^2] \\
&= y\,\frac{\mu_1-\mu_0}{\sigma_0^2} - \frac{\mu_1^2-\mu_0^2}{2\sigma_0^2} \\
&= \frac{\mu_1-\mu_0}{\sigma_0^2}\left(y - \frac{\mu_0+\mu_1}{2}\right)
\end{aligned}
$$

Thus the test after n observation is to

Reject if $\dfrac{\sigma_0^2 \ln B}{\mu_1-\mu_0} \leq \Sigma\left(y_i - \dfrac{\mu_0+\mu_1}{2}\right)$

Continue $\dfrac{\sigma_0^2 \ln A}{\mu_1-\mu_0} < \Sigma\left(y_i - \dfrac{\mu_0+\mu_1}{2}\right) < \dfrac{\sigma_0^2 \ln B}{\mu_1-\mu_0}$

Accept $\Sigma\left(y_i - \dfrac{\mu_0+\mu_1}{2}\right) \leq \dfrac{\sigma_0^2 \ln A}{\mu_1-\mu_0}$.

Approximate values for A and B can be obtained by a surprisingly simple analysis.

(b) Approximations for the likelihood bounds A, B .

The critical step in the proof of the hypothesis testing lemma in Section 4 can be used to derive the following simple approximations for A and B .

Lemma. *The wald sequential test with size* α *and power* $1-\beta$ *has*

$$A \approx \frac{\beta}{1-\alpha} \qquad B \approx \frac{1-\beta}{\alpha} \ ;$$

the exact values form a tighter likelihood interval

$$\frac{\beta}{1-\alpha} \leq A \quad , \qquad B \leq \frac{1-\beta}{\alpha} \qquad .$$

Proof: The calculation of the probability $P(\text{Reject } |\Theta_0)$ requires an integration over parts of $\mathbb{R}^1$, of $\mathbb{R}^2 \ldots$. Let $\bar{D}_n$ be the points $(y_1, \ldots, y_n)$ of $\mathbb{R}^n$ that give the terminal Reject H_0 :

$$\bar{D}_n = \left\{ (y_1, \ldots, y_n) : \begin{array}{ll} A < \dfrac{\Pi_1^k f(y_i|\Theta_1)}{\Pi_1^k f(y_i|\Theta_0)} < B & k=1,\ldots,n-1 \\ & \\ B \leq \dfrac{\Pi_1^n f(y_i|\Theta_1)}{\Pi_1^n f(y_i|\Theta_0)} & \end{array} \right\} .$$

Then

$$\alpha = P(\text{Reject } |\Theta_0) = \sum_{n=1}^{\infty} \int_{\bar{D}_r} \Pi_1^n f(y_i|\Theta_0) d\underset{\sim}{y}$$

$$= \sum_{n=1}^{\infty} \int_{\bar{D}_r} \frac{\Pi_1^n f(y_i|\Theta_1)}{B} d\underset{\sim}{y}$$

$$= \frac{1-\beta}{B}$$

and similarly

$$\beta = P(\text{Accept } |\Theta_1) \leq A(1-\alpha) \quad .$$

This calculation uses

$$P(\text{Reject } | \Theta_i) + P(\text{Accept } | \Theta_i) = 1$$

or equivalently

$$P(N = \infty) = 0 \; ;$$

see Problem III that the test terminates with probability one. The approximations can be solved to give the reverse expressions:

$$\alpha \approx \frac{1-A}{B-A} \qquad \beta \approx \frac{1-B^{-1}}{A^{-1}-B^{-1}} \quad .$$

Now suppose that a test is constructed using the approximate boundaries given in the lemma and let α' and β' denote the actual error probabilities for the resulting test. Then the inequalities

$$\frac{\beta'}{1-\alpha'} \le \frac{\beta}{1-\alpha} \quad \text{and} \quad \frac{1-\beta}{\alpha} \le \frac{1-\beta'}{\alpha'}$$

give

$$\alpha' \le \frac{\alpha}{1-\beta} \qquad \beta' \le \frac{\beta}{1-\alpha} \quad ,$$

and the modified inequalities

$$\beta'(1-\alpha) \le \beta(1-\alpha') \qquad \alpha'(1-\beta) \le \alpha(1-\beta')$$

when added give

$$\alpha' + \beta' \leq \alpha + \beta \qquad .$$

The use of the approximate boundaries can allow one of the error probabilities to be slightly larger than targeted. The main effect of the wider likelihood boundaries is that the sampling may continue slightly longer than needed.

(c) Approximate power function

As mentioned earlier the sequential test is concerned nominally with θ_0 against θ_1 but the typical applications are involved with a parameter that has a continuous range say $\Omega = \mathbb{R}$. The exact calculation of the power function can be excessively difficult. Fortunately there is a simple approximation.

For a parameter value θ let $h = h(\theta)$ be a power for which

$$\left(\frac{f(y|\theta_1)}{f(y|\theta_0)} \right)^h f(y|\theta)$$

is a density function (i.e., integrates to 1) . Of course $h = 0$ works but usually there is a nonzero value also. Now consider the sequential test of

$$f(y|\theta) \quad \text{against} \quad \left(\frac{f(y|\theta_1)}{f(y|\theta_0)} \right)^h f(y|\theta)$$

using the boundaries A^h , B^h . Note that this test, for example

$$A^h \leq \left(\frac{f(y_1|\Theta_1)}{f(y_1|\Theta_0)} \cdots \frac{f(y_n|\Theta_1)}{f(y_n|\Theta_0)} \right)^h \leq B^h \quad ,$$

is equivalent to the original test. The formula at the end of the proof of the lemma in section (b) then gives

$$P(\text{Reject } |\Theta) = \frac{1-A^h}{B^h-A^h} \quad .$$

If h is negative a similar analysis gives the same formula. The limiting value as h approaches zero is

$$\frac{-\ln A}{\ln B - \ln A} \quad .$$

Example 4 continued. Consider a sequential test of p_0 vs p_1 for a bernoulli variable. The equation defining $h = h(p)$ is

$$\left(\frac{p_1}{p_0}\right)^h p + \left(\frac{q_1}{q_0}\right)^h q = 1 \ ;$$

hence

$$p = \frac{1-(q_1/q_0)^h}{(p_1/p_0)^h-(q_1/q_0)^h} \qquad h \neq 0$$

$$= \frac{-\ln\ (q_1/q_0)}{\ln\ (p_1/p_0)\ -\ \ln\ (q_1/q_0)} \qquad h = 0 \quad .$$

For the sequential test with given A and B the approximate power function can be graphed against h using

$$\mathcal{P}(\Theta) = \frac{1-A^h}{B^h-A^h} \quad .$$

Then for the particular problem in this case the bernoulli, the h values can be indexed by the corresponding parameter values by the formula as in the preceding paragraph.

(d) The mean sample size.

An approximation for the mean sample size $E(N|\Theta)$ can be derived by using the same trick that produced the approximate formulas for the likelihood bounds. For this we need the wald equation

$$E(Z_1 + \ldots + Z_N|\Theta) = E(Z|\Theta)E(N|\Theta)$$

which holds for independent identically distributed Z_i with $E(Z)$ finite and for any sequential procedure with $E(N)$ finite. The left side can be reexpressed:

$$\sum_{n=1}^{\infty} \sum_{k=1}^{n} E(Z_k|N=n)P(N=n) = \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} E(Z_k|N=n)P(N=n)$$

$$= \sum_{k=1}^{\infty} E(Z_k|N \geq k)P(N \geq k) \quad .$$

The event $N \geq k$ is determined by $Z_1, \ldots, Z_{k-1}$ and thus is independent of Z_R . Hence the left side then becomes

$$= E(Z) \sum_{k=1}^{\infty} P(N \geq k)$$

$$= E(Z)E(N) \quad ;$$

using Problem 54 in FOUR.

Now consider the left side $E(Z_1 + \ldots + Z_N | \Theta)$ of the wald equation. At the termination of the sampling, $Z_1 + \ldots + Z_N$ is $\ln B$ or greater with probability $\mathcal{P}(\Theta)$ or is $\ln A$ or less with probability $1-\mathcal{P}(\Theta)$; it will exceed the bounds by at most the value of the terminal Z . Thus we have the approximation

$$E(Z_1 + \ldots Z_N | \Theta) \approx \ln B \, \mathcal{P}(\Theta) + \ln A \, (1-\mathcal{P}(\Theta)$$

The wald equation then gives

Sequential Problems

108. Let y be a response that is normally distributed (μ_0, σ^2) . Describe the sequential test of σ_0^2 against σ_1^2, using the approximate formulas to obtain the error probabilities $\alpha = .05$ and $\beta = .05$.

119. Let y be a response with the exponential distribution $f(y|\Theta) = \Theta \exp\{-\Theta y\}$ for $y > 0$. Describe the sequential test of Θ_0 against Θ_1 using the approximate formulas to obtain the error probabilities $\alpha = .05$ and $\beta = .01$.

110. Let y be a poisson response with mean Θ. Describe the sequential test of Θ_0 against Θ_1 using the approximate formula to obtain the error probabilities $\alpha = .01$ and $\beta = .01$.

111. The sequential test terminates with probability one unless $Z = 0$ with Θ probability one. Let $d = \ln B - \ln A$. If $P(Z \neq 0|\Theta) > 0$ then

(a) There is an integer r such that $P(|Z_1 + \ldots + Z_r|<d) = \rho < 1$

(b) Hence $P\left(N \geq kr+1\right) = P(|Z_1+\ldots+Z_r|<d,\ldots,|Z_{(n-1)r+1}+\ldots+Z_{nr}|<d) = \rho^k$

(c) The test terminates with probability one.

A two sample problem

112. Let $(y_1, \ldots, y_m)$ be a sample from the normal (μ,σ^2) and let $(w_1, \ldots, w_n)$ be a further sample from the same normal where n depends on the first sample variance s^2. An estimate is wanted with variance δ^2 regardless of σ^2. Given s_y^2 the estimate $a\bar{y}_1 + b\bar{w}$ of μ with $a+b = 1$ has variance $(a^2/m + b^2/n)$. Choose

a, b, n so that the estimated variance $(a^2/m + b^2/n)s_y^2 = \delta^2$. (For example choose $n \geq 1$ as small as possible so that $s^2/(m+n)$ is just less than δ^2; then argue that values of a, b can be found to give equality.

113. Continuation. Prove that

$$Z = \frac{a\bar{y} + b\bar{w} - \mu}{(a^2/m + b^2/n)^{\frac{1}{2}}}$$

has a standard normal distribution given s^2 and hence that

$$t = \frac{a\bar{y} + b\bar{w} - \mu}{\delta}$$

has a t distribution on $m-1$ degrees of freedom; thus $E(t) = 0$, and $Var(t) = (m-1)\delta^2/(m-3)$ which is slightly larger than the targeted variance δ .

114. Continuation. For $m = 10$ give an expression for the minimum second sample size n to attain a 95% confidence interval for μ of prescribed length 2ℓ .

10. TEST OF THE MODEL

In this and in the preceding chapter we have been considering response models with one, two or several parameters. In a typical application to a stable system the model will have been developed on the basis of earlier experience with the same system or with similar systems. If the grounds for the model are not strong then the investigater may want to check the model by seeing if subsequent response values are in agreement with it. And even in cases where the grounds are strong, the investigater may want to check the model as a routine precaution. In this section we investigate the chi-square tests for a model. The chi-square tests cover a broad range of models and they have a readily accessible theory. There are many other tests for models but in general they are suitable or accessible only in relatively restricted or simple cases. The chi-square tests have optimum properties as based on the large sample form for the statistical model.

In Chapter ONE, Section 5 we examined the Rutherford and Geiger data involving 2608 counts on the number of α particles in a 7.5 second interval. The poisson model is a fairly standard model for such frequency counts and in fact the frequency counts from radioactive disintegration are perhaps the classic example for the poisson model. We compared the observed frequency f_i for a sample point with the expected frequency e_i by calculating a standardized deviation

$$2(\sqrt{f_i} - \sqrt{e_i}) \quad ;$$

if the model is correct then such deviations should relate in an approximate way with the standard normal. An overall test of fit for the model can be obtained by calculating the sum of squares of the standardized deviations,

$$\chi^2 = 4\Sigma(\sqrt{f_i} - \sqrt{e_i})^2 = 14.60$$

and comparing it with the chi-square distribution; we will see that the appropriate degrees of freedom is ~~less~~ $12 - 1 - 1 = 10$ where 12 is the number of cells after some grouping to avoid too small values for the e_i, less 1 for the imposed constraint $\Sigma f_i = \Sigma e_i$, less 1 for fitting the parameter $\hat{\lambda} = 3.87$ (which brings the fitted probabilities closer to the data than the true probabilities would be). The value can be compared with the 5% value 18.3 and the 1% value 23.2 . Thus the data are quite reasonable for the poisson model.

And in Chapter ONE, Section 5 we examined the Grummel and Dunningham data involving 250 observations on a continuous response. The normal model with two parameters was indicated for the particular application. To apply the chi-square test to the continuous case it is necessary to group or form cells. We compared the observed frequency f_i with the expected frequency e_i by calculating the standardized deviation

$$2(\sqrt{f_i} - \sqrt{e_i})$$

and comparing it in an approximate way with the standard normal. The overall test would be obtained by calculating

$$\chi^2 = 4\Sigma(\sqrt{f_i} - \sqrt{e_i})^2 = 3.89$$

and comparing it with the chi-square distribution on $10 - 1 - 2 = 7$ degrees of freedom: there are 10 cells after grouping to avoid too small values for the e_i, there is 1 imposed constraint $\Sigma f_i = \Sigma e_i$, and there are 2 filled parameters $\hat{\mu}, \hat{\sigma}^2$ (which bring the fitted probabilities closer to the data than the true probabilities would be). The observed value can be compared with the 5% value 14.1 and the 1% value 18.5. Thus the data conform reasonably to the model.

(a) The chi-square test

Now consider generally the problem of testing a model. In the discrete case of model with parameter θ will prescribe probabilities $p_i(\theta)$ for the various discrete sample points x_i. For a sample of n let f_i

designate the number or frequency of values x_i . By Chapter TWO, Section 6 the distribution of the frequencies $(f_1, \ldots, f_k)$ is multinomial $(n, p_1(\theta), \ldots, p_k(\theta))$ where θ is the parameter value of the distribution being sample.

In the continuous case a model with parameter θ will prescribe a density $f(y|\theta)$ on the sample space S . As part of the chi-square method we suppose that intervals or cells have been formed on the sample space, let $p_i(\theta)$ be the total (θ) probability in the i-th cell. Then for a sample of n let f_i designate the number of response values in the i-th cell. This reduces the problem to the discrete case and the distribution of the frequencies $(f_1, \ldots, f_k)$ is multinomial $(n, p_1(\theta), \ldots, p_k(\theta))$ where θ is the parameter value of the distribution being sampled.

We can derive the large sample distribution of the frequencies in several ways. By using moment generating or characteristic functions and the methods of Chapter FIVE we can show that the distribution approaches multivariate normal form given the constraint $\Sigma f_i = n$. Alternatively by using the probability function directly we will show at the end of this section that the limiting distribution of

$$t_1 = \frac{f_1 - np_1(\theta)}{(np_1(\theta))^{\frac{1}{2}}} , \ldots , t_k = \frac{f_k - np_k(\theta)}{(np_k(\theta))^{\frac{1}{2}}}$$

is that of a sample of k from the standard normal subject to the condition $\Sigma t_i\sqrt{p_i(\theta)} = 0$. And we will show that the limiting distribution of

$$Z_1 = 2(\sqrt{f_1} - \sqrt{np_1(\theta)}\,),\ \ldots,\ Z_k = 2(\sqrt{f_k} - \sqrt{np_k(\theta)}\,)$$

is that of a sample of k from the standard normal subject to the condition $\Sigma Z_i\sqrt{p_i(\theta)} = 0$. Then by Chapter FIVE (Section 2 and Problem   ) it follows that

$$\chi_1^2 = \Sigma t_i^2 = \Sigma \frac{(f_i - np_i(\theta))^2}{np_i(\theta)} = \Sigma \frac{(f_i - e_i)^2}{e_i}$$

has a limiting chi-square distribution on $k-1$ degrees of freedom and that

$$\chi_2^2 = \Sigma Z_i^2 = \Sigma 4(\sqrt{f_i} - \sqrt{np_i(\theta)})^2 = \Sigma\ 4(\sqrt{f_i} - \sqrt{e_i})^2$$

has a limiting chi-square distribution on $k-1$ degrees of freedom.

The use of the t's and hence χ_1^2 leads to the traditional chi-square test introduced by Karl Pearson in 1900. Alternatively the use of the Z's gives a somewhat better normal approximation; it corresponds to a single reexpression $2\sqrt{f_i}$ of a frequency rather than a reexpression $f_i/\sqrt{e_i}$ that depends on the parameter value being examined; and it permits the decomposition of a chi-square in correspondence with a succession of hypotheses concerning θ . We will use the second form of chi-square

$$\chi^2 = \Sigma 4\left(\sqrt{f_i} - \sqrt{e_i}\right)^2$$

for making tests of a model.

Example 8. In the breeding of a certain type of flower the offspring can have a magenta M or red R flower and can have a green g or red r stigma; the possibilities are Mg, Mr, Rg, Rr . A strong theory A specifies that they occur with relative probabilities 9, 3, 3, 1 . For 220 offspring the expected frequencies $e_i = 220p_i$ under model A are recorded in the left array and the root expected frequencies in the right array:

	g	r	
M	123.75	41.25	165
R	41.25	13.75	55
	165	55	220

	g	r
M	11.12	6.42
R	6.42	3.71

The following data were obtained, frequencies f_i in the left array and root frequencies $\sqrt{f_i}$ in the right array:

	a	r	
M	117	31	148
R	55	17	72
	172	48	220

10.82	5.57
7.42	4.12

.

The difference vector based on root frequencies is

$$\begin{pmatrix} Z_{\frac{1}{2}} & Z_{\frac{2}{2}} \\ Z_{\frac{3}{2}} & Z_{\frac{4}{2}} \end{pmatrix} = \begin{matrix} -.30 & -.85 \\ 1.00 & .41 \end{matrix} \quad .$$

The observed value of chi-square is

$$\begin{aligned} \chi^2 &= 4((-.30)^2 + (-.85)^2 + (1.00)^2 + (.41)^2) \\ &= 7.92 \end{aligned}$$

which can be compared with the chi-square distribution on 3 degrees of freedom: the 5% point is 7.81; the 1% point is 11.3. The observed chi-square gives some moderate evidence against the hypothesis and suggests that the 9, 3, 3, 1 relative probabilities may not be applicable.

Example 7. Consider how the chi-square test works for the the simple case of a binomial count y with probability p. As a multinomial we have frequencies y, $n-y$ and expected frequencies np, nq. Correspondingly we have root frequencies $\sqrt{y}$, $\sqrt{n-y}$ and root expected frequencies $\sqrt{np}$, $\sqrt{nq}$. These can be plotted as in Figure 16. The difference vector $\left(Z_1/2, Z_2/2\right)$ can be represented by the coordinate $Z/2$ on an axis tangential to the circle. Note that $Z/2$ is approximately n times the angle between the two rays from the origin:

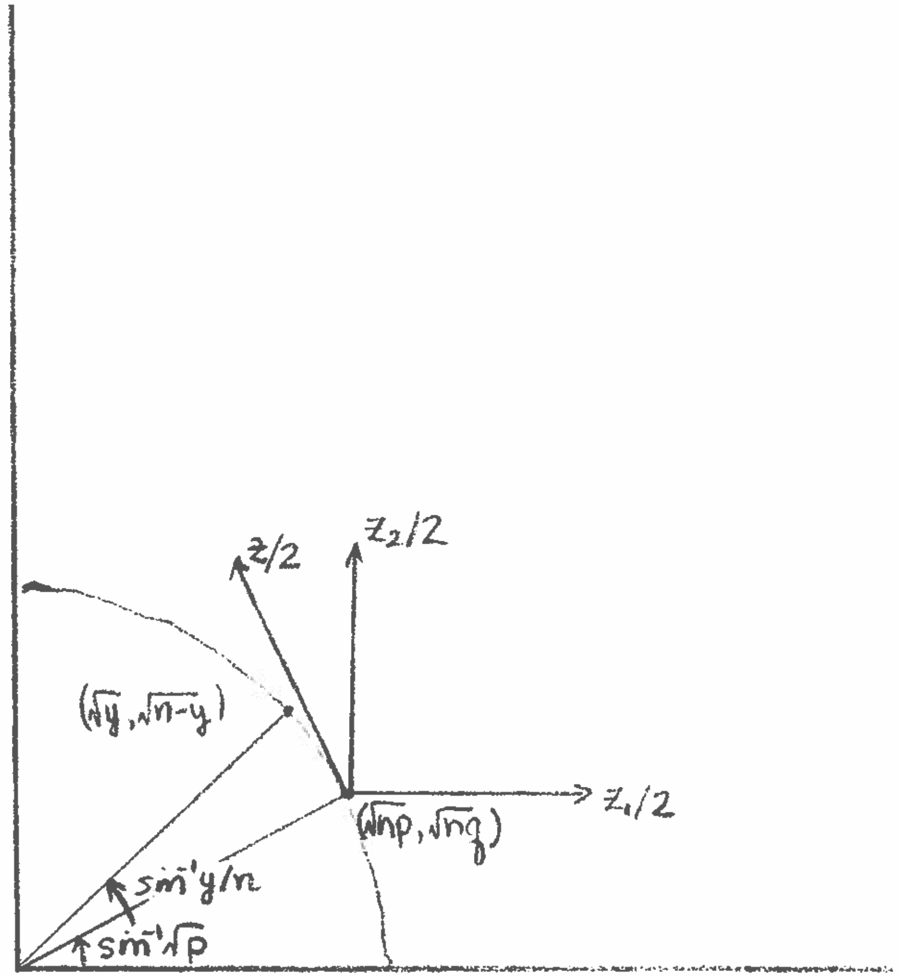

Figure 16. The root frequencies $(\sqrt{y}, \sqrt{n-y})$ and the root expected frequencies $(\sqrt{np}, \sqrt{nq})$. The difference vector $(Z_1/2, Z_2/2)$ can be represented by the coordinate $Z/2$ on an axis tangential to the circle.

$$Z = 2\sqrt{n}\left(\sin^{-1}\sqrt{\frac{Y}{n}} - \sin^{-1}\sqrt{p}\right) .$$

The *angular* transformation in the bracket is available from most statistical tables; it is used for comparing proportions with probabilities and is treated as a central normal variable with standard deviation $1/2\sqrt{n}$. This is in agreement with our use of Z as a standard normal variable.

(b) With fitted parameters.

Now suppose that the true cell probabilities are in a small neighbourhood of the point $(p_1(\theta_*) , \ldots, p_k(\theta_*))$ and that in this neighbourhood the model $(p_1(\theta), \ldots, p_k(\theta))$ is a continuously differentiable surface that is approximately linear with dimension given by the number of coordinates for θ. For the case of a single coordinate for θ we have

$$2\sqrt{np_i(\theta)} = 2\sqrt{np_i(\theta_*)} + \delta x_i$$

where

$$\delta = \sqrt{n}\ (\theta-\theta_*) \ , \qquad x_i = \frac{1}{\sqrt{p_i(\theta)}} \frac{dp_i(\theta)}{d\theta}\bigg|_{\theta_*}$$

Hence

$$\left(2\sqrt{f_1} - 2\sqrt{np_1(\theta_*)} \ , \ \ldots, \ 2\sqrt{f_k} - 2\sqrt{np_k(\theta_*)}\ \right)$$

$$= (Z_1, \ldots, Z_k) + \delta(x_1, \ldots, x_k) .$$

Thus the standardized deviations as taken from a θ_* trial value are conditioned normal variables as in part (a) but now with a mean than can lie along the vector $(x_1, \ldots, x_k)$. Then by Seven (Section 2) and Eight (Section 6, Example 6), it follows that the sum of squares of deviations from the fitted model will be chi-square but now with one degree of freedom less. Thus the minimized

$$\chi^2 = \Sigma 4\left(\sqrt{f_i} - \sqrt{np_i(\theta)}\right)^2$$

has a chi-square distribution on $k - 1 - 1 = k - 2$ degrees of freedom. In general if r independent parameter are fitted the minimized chi-square has a limiting chi-square on $k - 1 - r$ degrees of freedom.

In practice it may be easier to obtain fitted parameter values by maximum likelihood (the poisson example) or by some large sample method equivalent to maximum likelihood. The distribution for the Z's represents the limiting distribution for the multinomial, and the likelihood function from the Z's is the limiting form for the likelihood function from the multinomial. Thus the maximum likelihood estimates can provide reasonable substitutes for the minimized chi-square values. For the normal example in Chapter ONE there is a possibility of discrepancies if estimates are based on the sample mean and variance before grouping to form cells.

Example 8 continued. Now suppose that a weaker theory B allows probabilities p , q for the occurrence of flower colours M , R respectively but keeps the independence between rows and columns and keeps the relative probabilities 3 to 1 for g to r . The points Mg , Mr , Rg , Rr then have probabilities $\frac{3}{4}p_1$, $\frac{1}{4}p_1$, $\frac{3}{4}q_1$, $\frac{1}{4}q_1$. For the 220 offspring the observed proportions for M and R are $\frac{148}{220}$ and $\frac{72}{220}$. By independence these can be combined with the column probabilities $\frac{3}{4}$ and $\frac{1}{4}$ to obtain fitted cell probabilities. The expected frequencies under model B are recorded in the left array and the root expected frequencies in the right array

$$\begin{array}{cc|c} 111 & 37 & 148 \\ 54 & 18 & 72 \\ \hline 165 & 55 & 220 \end{array} \qquad \begin{array}{cc|} 10.54 & 6.08 \\ 7.35 & 4.24 \\ \hline \end{array} \; .$$

The observed frequencies and root frequencies are recorded again:

$$\begin{array}{cc|c} 117 & 31 & 148 \\ 55 & 17 & 72 \\ \hline 172 & 48 & 220 \end{array} \qquad \begin{array}{cc|} 10.82 & 5.57 \\ 7.42 & 4.12 \\ \hline \end{array} \; .$$

The difference vector is

$$\begin{array}{cc|} .28 & -.51 \\ .07 & -.12 \\ \hline \end{array} \; .$$

The observed chi-square for deviations form the fitted model is

$$\begin{aligned} \chi^2 &= 4((.28)^2 + (-.51)^2 + (.07)^2 + (-.12)^2) \ . \\ &= 1.43 \end{aligned}$$

which can be compared with the chi-square distribution on 2 degrees of freedom: the 5% point is 6.0; the 1% point is 11.3. The observed value is somewhat smaller than the mean (2) and is certainly a reasonable value; the data are in accord with the model.

Consider the difference vector for comparing the fitted model B frequencies with the expected frequencies under model A :

$$\begin{array}{cc|} -.58 & -.34 \\ .93 & .53 \\ \hline \end{array} \qquad .$$

The observed χ^2 for this difference is

$$\begin{aligned} \chi^2 &= 4((-.58)^2 + (-.34)^2 + (.93)^2 + (.53)^2) \\ &= 6.39 \ . \end{aligned}$$

If we accept theory B then theory A is equivalent to the hypothesis H_0: $p_1 = \frac{3}{4}$. By Section 6, Example 6 we can test this hypothesis by comparing the observed $\chi^2 = 6.4$ with the chi-square distribution on 1 degree of freedom: the 5% point is 3.8, the 1% point is 6.6. The observed value gives moderately strong evidence against the hypothesis.

The details for the two tests are recorded as follows (degrees of freedom (DF) and sums of squares (SS)):

Source	DF	SS
Model A after Model B	1	6.39
Deviations from Model B	2	1.43 .

The deviations from Model B are reasonably small giving some grounds for testing Model A given Model B. The details for the combined test earlier in this section are

Source	DF	SS
Deviations from Model A	3	7.92 .

The sum of squares in the first table are an approximate decomposition of the sum of squares in the second. The slight discrepancy is due to curvature of the 3 dimensional surface on which the points lie and to curvature of the line representing Model B .

(c) Chi-square tests of independence

The chi-square methods in this section can provide a test for the statistical independence of two response variables. For this suppose that the two variables are discrete or that they have been made discrete by grouping into cells as with the normal distribution example mentioned earlier in this section.

For the first variable let $A_1, \ldots, A_k$ designate the k distinct possibilities and for the second variable let $B_1, \ldots, B_\ell$ designate the ℓ distinct possibilities. Then any observation on the pair of variables corresponds to a combination A_iB_j . Then for n observations we can record frequencies f_{ij} as in the following array

$$\begin{array}{cccc|c}
\multicolumn{1}{c}{} & B_1 & \cdots & \multicolumn{1}{c}{B_\ell} & \\
A_1 & f_{11} & \cdots & f_{1k} & m_1 \\
\vdots & \vdots & & \vdots & \\
A_k & f_{k1} & \cdots & f_{k\ell} & m_k \\
\cline{2-4}
\multicolumn{1}{c}{} & n_1 & \cdots & \multicolumn{1}{c}{n_\ell} & n
\end{array}$$

Let $p_1, \ldots, p_k$ designate probabilities for the first variable $(\Sigma p_i = 1)$ and $p_1', \ldots, p_\ell'$ designate probabilities for the second variable $(\Sigma p_j' = 1)$. Then the hypothesis of independence specifies a model with cell probabilities

$$p_{ij} = p_i p_j' \quad ;$$

this model has $k-1$ free parameters for the rows and $\ell-1$ free parameters for the columns.

The row probabilities can be estimated by row proportions

$$\hat{p}_i = \frac{m_i}{n} \quad ;$$

and the column probabilities by the column proportions

$$\hat{p}_j = \frac{n_j}{n} \quad .$$

Thus the fitted model has expectations $n\hat{p}_i\hat{p}_j = m_i n_j/n$ recorded in the following array

$$\begin{array}{c|ccc|c}
 & B_1 & \cdots & B_\ell & \\
\hline
A_1 & \frac{m_1 n_1}{n} & \cdots & \frac{m_1 n_\ell}{n} & m_1 \\
\vdots & \vdots & & \vdots & \vdots \\
A_k & \frac{m_k n_1}{n} & \cdots & \frac{m_k n_\ell}{n} & m_k \\
\hline
 & n_1 & \cdots & n_\ell & n
\end{array} \quad .$$

The chi-square measure for deviations from the model is

$$\chi_1^2 = \frac{(f_{ij}-m_i n_j/n)^2}{m_i n_j/n}$$

by the pearson formula and is

$$\chi_2^2 = 4\Sigma\left(\sqrt{f_{ij}} - \sqrt{m_i n_j/n}\right)^2$$

by the standardized distance formula. An observed value would be tested against the chi-square distribution on $k\ -1-(k-1)-(\ell-1) = (k-1)(\ell-1)$ degrees of freedom.

Example 8 continued. By the earlier analysis we have no reason~~s~~ to doubt Model B. But to illustrate the independence test consider a weaker Model C that allows probabilities p_1, q_1 for the flower colour and probabilttIes p_2, q_2 for the stigma colour but specifies independence between the two categories. The observed pooportions for the rows are $\frac{148}{220}$, $\frac{72}{220}$, and for the columns are $\frac{172}{220}$, $\frac{48}{220}$. The expected frequencies under Model C are recorded in the left array and the root expected frequencies in the right array

$$\begin{array}{cc|c} 115.71 & 32.29 & 148 \\ 56.29 & 15.71 & 72 \\ \hline 172 & 48 & 220 \end{array} \qquad \begin{array}{cc|} 10.76 & 5.68 \\ 7.50 & 3.96 \\ \hline \end{array} \quad .$$

The observed frequencies and root frequencies are recorded again

$$\begin{array}{cc|c} 117 & 31 & 148 \\ 55 & 17 & 72 \\ \hline 172 & 48 & 220 \end{array} \qquad \begin{array}{cc|} 10.82 & 5.57 \\ 7.42 & 4.12 \\ \hline \end{array} \quad .$$

The difference vector is

$$\begin{array}{cc|} .06 & -.11 \\ -.08 & .16 \\ \hline \end{array} \quad .$$

The observed chi-square for deviations from Model C is

$$\chi^2 = 4((.06)^2 + (-.11)^2 + (-.08)^2 + (.16)^2)$$
$$= .19$$

which can be compared with the chi-square distribution as 1 degree of freedom. If anything, the observed value looks rather small; the data are in accord with the model.

The data and the fitted frequencies by Models C, B, A give the following tabulations of sums of squares obtained from difference vectors.

Source	DF	SS
Model A after Model B (row probabilities)	1	6.39
Model B after Model C (column probabilities)	1	1.23
Model (independence)	1	.19

This gives a three way decomposition of the original chi-square.

(d) The limiting normality.

The limiting normality of the multinomial can be derived from the limiting normality of the poisson distribution.

Let y have a poisson distribution (Θ) and condider the distribution of

$$t = \frac{y-\Theta}{\Theta^{\frac{1}{2}}}$$

as Θ approaches infinity. The limiting distribution can be obtained by using moment generating or characteristic functions (FIVE, Problem 17). We derive a stronger result concerning the probability function itself. The poisson probability at the point y is

$$\frac{\Theta^{y}e^{-\Theta}}{y!} = \frac{\Theta^{y}e^{-\Theta}}{\Gamma(y+1)} .$$

The transformation $y = \Theta + t\Theta^{\frac{1}{2}}$ gives probability

$$\frac{\Theta^{\Theta+t\Theta^{\frac{1}{2}}}e^{-\Theta}}{\Gamma(\Theta+t\Theta^{\frac{1}{2}}+1)}$$

at a possible value t, and it gives probability

$$\frac{\Theta^{\Theta+t\Theta^{\frac{1}{2}}}e^{-\Theta}}{\Gamma(\Theta+t\Theta^{\frac{1}{2}}+1)} \Theta^{\frac{1}{2}}$$

in units of the spacing between possible points for the transformed variable. The limit as $\Theta \to \infty$ can be obtained by using Sterling's formula

$$\Gamma(x+1) = \sqrt{2\pi}\, x^{x+\frac{1}{2}} \exp\{-x + 1/12x - \quad\} ;$$

the limit is

$$\frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

and it is uniform over the range of t. This establishes the standard normal limiting distribution, but it establishes more --

that the probability function in proper units approaches the standard normal density.

Now consider $(y_1, \ldots, y_k)$ with the multinomial distribution $(n; p_1, \ldots, p_k)$. This distribution can be obtained (THREE, Problem 17) as the conditional distribution of $(y_1, \ldots, y_k)$ given $\Sigma y_i = n$ where the y's are taken to be independent poisson variables with means $np_1, \ldots, np_k$. Let $(t_1, \ldots, t_k)$ be the corresponding standardized deviations:

$$t_i = \frac{y_i - np_i}{(np_i)^{\frac{1}{2}}} \qquad .$$

The sample points for $(y_1, \ldots, y_n)$ are the points on $\Sigma y_i = n$ with integer coordinates; they are uniformly spaced. The transformation to the t's is linear; correspondingly the sample points for $(t_1, \ldots, t_k)$ are uniformly spaced on $\Sigma t_i \sqrt{p_i} = 0$.

Now consider the probability function for $(t_1, \ldots, t_k)$. Except for the norming constant, this probability function is the same as the probability function obtained from the *unconditioned* distribution of the independent poisson variables. But that probability function approaches the density of a sample from the standard normal. Hence the probability function for $(t_1, \ldots, t_k)$ approaches the $\Sigma t \sqrt{p_i} = 0$ section of the normal sample density. Hence its limiting distribution is that of a sample of k from the standard normal subject to the condition $\Sigma t_i\sqrt{p_i} = 0$.

Now consider the transformation between the t's and the Z's :

$$
\begin{aligned}
Z &= 2(\sqrt{y} - \sqrt{\Theta}) \\
&= 2((\Theta + t\Theta^{\frac{1}{2}})^{\frac{1}{2}} - \Theta^{\frac{1}{2}}) \\
&= 2\Theta^{\frac{1}{2}}((1 + t\Theta^{-\frac{1}{2}})^{\frac{1}{2}} - 1) \\
&= 2\Theta^{\frac{1}{2}}(\tfrac{1}{2}t\Theta^{-\frac{1}{2}} + \tfrac{1}{2}(-\tfrac{1}{2})t^2\Theta^{-1}/2! + \ldots) \\
&= t + t^2 0(\Theta^{-\frac{1}{2}}) \quad .
\end{aligned}
$$

Thus on any bounded positive interval t approaches Z uniformly as $\Theta \to \infty$. Hence the limiting distribution of $(Z_1, \ldots, Z_k)$ is that of a sample of k from the standard normal subject to the condition $\Sigma t_i\sqrt{p_i} = 0$.

Problems.

115. A die was tossed 1600 times:

Event	1	2	3	4	5	6
Frequency	301	308	340	214	196	241

Test the fit of the model based on symmetry.

116. One hundred plants were classified according to two attributes: large L or small ℓ ; white W or coloured w . The observed frequencies are

	W	w	
L	40	20	60
ℓ	15	25	40
	55	45	100

(a) Test the model that specifies equal probabilities
($\frac{1}{2}$) in the four cells;

(b) Test the model that specifies equal column probabilities
($\frac{1}{4}$) *and* independence between the attributes;

(c) Test the model that specifies independence allowing arbitrary row and column probabilities. Record the three way decomposition of the chi-square under (a) .

11. WHEN DO THE METHODS WORK?

In this chapter we have examined the common methods available for deriving statistical procedures for the ordinary response model. For each of the methods we are able to determine the mathematical properties that are instrumental to the success of the method. In this section we summarize these properties. We then examine some common statistical models appropriate to applications and determine when the methods are successful for these models.

(a) The methods.

Local unbiasedness was used in Section 2 to derive minimum variance unbiased estimates. The success of the method required the logarithm of the density function

$$\ln f(y|\Theta) = \phi(\Theta) + \sum_{1}^{r} \psi_j(\Theta)a_j(y) + b(y)$$

to have a linear construction in which the number of essential ψ-function was equal to the dimension of the parameter.

Completeness was used in Section 3 to derive minimum variance unbiased estimates. In the fixed carrier case the principle example was the exponential model (as given in the preceding paragraph) with the number of ψ

function equal to the dimension of the parameter. In the variable carrier case the principle example involved boundaries that were monotone functions of the parameters. In any case a general survey shows that the method is essentially restricted to cases in which the likelihood statistic has the same dimension as the parameter; note Problem 48.

The likelihood ratio methods for deriving uniformly most powerful tests require effectively that the likelihood ratio to be monotone in terms of a real valued function. As noted in Section 5 this requires the likelihood statistic to be one dimensional.

The unbiasedness methods for deriving uniformly most powerful tests are closely linked to the exponential model with the number of ψ functions equal to the dimensional of the parameter. In a larger context they seem to require a combination of completeness for certain parameters and monotone likelihood ratio for others. This indicates that the method is effectively restricted to cases in which the likelihood statistic has the same dimension as the parameter.

The invariance method for obtaining uniformly most powerful tests is largely restricted to cases where the likelihood statistic has the same dimension as the parameter. And the method operates mainly to isolate component statistics for treating component parameters.

In general summary, the methods require at least that *the likelihood statistic have the dimension of the parameter* and they may require *the exponential construction with the number of ψ function equal to the dimension of the parameter.*

Now consider cases in which we have samples from a distribution. In SEVEN, Section 4 we found that fixed dimension for the likelihood statistic required an exponential construction with dimension given by the number of ψ functions (the fixed carrier case). The variable carrier case requires exponential form combined with monotone boundaries and the dimension is given by the number of ψ functions plus the number of boundaries. The case of independent coordinates seems to be at least as restricted as the more special case of samples.

In summary then, the methods are successful under sampling when *the model is exponential ~~construction~~ with one ψ function for one parameter, or two ψ functions for two parameters*, and so on; and what success there is decreases with the number of parameters. In the variable carrier case, a monotone boundary can replace a ψ function.

(a) The location model

An important and fairly common problem in applications is concerned with the general level of a response in the situation where the distribution of the variation has been identified to reasonable approximation by background experience. Let g be a density describing the variation in the response and let θ be a parameter designating the location or general level of the response distribution. The response model then has the form

$$f(y|\theta) = g(y-\theta)$$

on $\mathbb{R}$ with parameter Ω in $\Omega = \mathbb{R}$.

Now consider when such a model can be handled by the methods in this Chapter EIGHT. In the fixed carrier case the following theorem shows that g can *only* be *a normal density* or *the logarithm of a gamma variable.* A very potent restriction on the methods! These are of course the examples that are used to illustrate the methods. We now see that they are effectively the *only* examples.

Theorem. *If a location model has exponential form*

$$g(y-\theta) = \gamma(\theta)\exp\{\psi(\theta)a(y)\}h(y)$$

then $g(y)$ *is either a relocated rescaled standard normal* $(y = a + cZ)$

$$\frac{1}{\sqrt{2\pi}}\exp\{-Z^2/2\}$$

or a relocated rescaled $\underline{(\pm)}$ *log gamma* $(y = a + cw)$

$$\Gamma^{-1}(p)\left(e^w\right)^{p-1}\exp\left\{-e^w\right\}e^w \quad .$$

Proof. The exponential form shows that the region of positive density cannot vary with θ; hence $g > 0$. Suppose that g is differentiable and write $\ell = \ln g$; then

(i) $$\ell(y-\theta) = \phi(\theta) + \psi(\theta)a(y) + b(y)$$

for all y and θ. Note that neither ψ nor a can be constant; for otherwise we could rearrange and eliminate the cross term with the result that the normalizing factor $\gamma(\theta)$ would be independent of θ thus contradicting the location form $g(y-\theta)$.

The key to the proof lies in the simple joint dependence on y and θ:

$$-\frac{\partial \ell(y-\theta)}{\partial\theta} = \ell'(y-\theta) = \frac{\partial \ell(y-\theta)}{\partial y} .$$

This gives

(ii) $$\phi'(\theta) + \psi'(\theta)a(y) = -\psi(\theta)a'(y) - b'(y) .$$

Taking a difference for two values of y gives

$$\psi'(\theta) = c_1\psi(\theta) + d_1$$

and correspondingly for θ gives

$$a'(y) = c_2 a(y) + d_2 \quad ;$$

substitution back shows that $c_2 = -c_1 = c$ say. This is a common differential equation,

$$\frac{dx}{dt} + cx = d ,$$

and we obtain the solutions

$$\begin{aligned} a(y) &= k_1 e^{cy} + d_3 && c \neq 0 \\ &= k_1 y \quad + d_3 && c = 0 \\ \psi(\Theta) &= k_2 e^{-c\Theta} + d_4 && c \neq 0 \\ &= k_2 \Theta \quad + d_4 && c = 0 \quad . \end{aligned}$$

The constants d_3 and d_4 can be taken equal to zero; for otherwise we could separate terms from the cross term in the expression (i) for $\ell(y-\Theta)$ and could then redefine $\phi(y)$ and $b(\Theta)$ corresponding to the $d_3 = d_4 = 0$ case.

Consider the $c \neq 0$ case. When the expressions for $a(y)$ and $\psi(\Theta)$ are substituted in (ii) the cross terms vanish leaving

$$\phi'(\Theta) = -b'(y)$$

which has solutions

$$\begin{aligned} b(y) &= d \cdot y \ + d_5 \\ \phi(\Theta) &= -d \cdot y + d_6 \ . \end{aligned}$$

Hence

$$g(y-\Theta) = d(y-\Theta) + k_1 k_2 e^{c(y-\Theta)}$$

which gives the relocated rescaled log gamma.

Now consider the case $c = 0$. When the expressions for $a(y)$ and $\psi(\Theta)$ are substituted in (ii) we obtain

$$\phi'(\Theta) + k_1 k_2 y = -k_1 k_2 \Theta - b'(y) \ ,$$

$$\phi'(\Theta) + k_1 k_2 \Theta = K = -(b'(y) + k_1 k_2 y) \ ,$$

$$b(y) = -\frac{k_1 k_2}{2} y^2 - Ky + d_7 \ ,$$

$$\psi(\Theta) = -\frac{k_1 k_2}{2} \Theta^2 + K\Theta + d_8 \ .$$

Hence

$$g(y-\Theta) = -\frac{k_1 k_2}{2} (y-\Theta)^2 - K(y-\Theta) + d$$

which gives the relocated rescaled normal.

Difference equations can replace the differential equations. This gives analogous solutions at the rationals or at the rationals displaced by a transcendental. Continuity of f on any open interval then identifies the separate solutions. Continuity on some small interval seems like a substantial minimum for a statistical model.

In the variable carrier case some similar methods show that g can only be a relocated rescaled $(\pm)$ exponential $(\exp\{-e\}$ on $(0,\infty))$.

Thus the common location model $g(y-\Theta)$ which is concerned with the general level of a response can be handled by the standard methods only for three very special variation forms g : the normal, the exponential, and the log-gamma. These are of course the standard examples. Perhaps it should be noted that they are effectively the only examples.

(c) The location-scale model.

Another important and common problem in applications is concerned with the general level of a resonse in the situation where the form of the variation has been identified to some reasonable approximation. Let g be a density describing the standardized variation, let σ be a scaling factor that gives the actual variation and let μ be a parameter designating the general level of the response. The response model then has the form

$$f(y|\mu,\sigma) = \frac{1}{\sigma} g\left(\frac{y-\mu}{\sigma}\right)$$

on $\mathbb{R}$ with parameter (μ,σ) in $\Omega = \mathbb{R} \times \mathbb{R}^+$.

Now consider when such a model can be handled by the methods in Chapter EIGHT. It can be shown that a two dimensional likelihood statistic is available under sampling only if the distribution g is relocated rescaled $(\pm)$ (a) standard normal, (b) uniform $(0,1)$, (c) exponential $(\exp\{-x\}$ on $(0,\infty))$. These are of course the standard examples; we now see they are effectively the only examples.

(d) In summary

We have examined two important kinds of statistical model and found that the response model methods are successful only for very special distributions for the variation -- the normal, the log gamma and the exponential. Distributions in practice seem to have more probability in the tails than the normal, to be more like the t distribution with degrees of freedom around 6, 7, 8; these models are not amenable to the standard response model methods.

Problems.

117. A variable carrier model with a one dimensional likelihood statistic has a single monotone boundary. For a location model with variable lower boundary this mean $f(y-\theta) = \gamma(\theta)\, c(y-\theta-a)h(y)$ where c is the indicator function for the positive axis. Show that h must be negative exponential.