Hypothesis Testing

Robert L. Wolpert
Department of Statistical Science
Duke University, Durham, NC, USA

1 An Example

Mardia et al. (1979, p. 121) reprint data from Frets (1921) giving the length and breadth (in millimeters) of the heads of the first and second son in a sample of \( n = 25 \) families, from a study of heredity in humans. If we assume a multivariate normal model then the following statistics are sufficient:

\[
\bar{x} = \begin{bmatrix} \bar{x}_1 = 185.72 \\ \bar{x}_2 = 151.12 \\ \bar{x}_3 = 183.84 \\ \bar{x}_4 = 149.24 \end{bmatrix}
\]

\[
\frac{1}{n} S = \begin{bmatrix} 91.481 & 50.753 & 66.875 & 44.267 \\ - & 52.186 & 49.259 & 33.651 \\ - & - & 96.775 & 54.278 \\ - & - & - & 44.222 \end{bmatrix}
\]

the sample mean \( \mu = \bar{x} = \frac{1}{n} \sum X_\alpha \) and the sample covariance \( \Sigma = \frac{1}{n} S \) where \( S := \sum (X_\alpha - \bar{x})(X_\alpha - \bar{x})' \).

If we model \( \{X_\alpha\} \sim_{\text{iid}} \text{No}(\mu, \Sigma) \) for \( 1 \leq \alpha \leq 25 \), the log likelihood function for \( \mu \) and \( \Lambda := \Sigma^{-1} \) is

\[
\ell(\mu, \Lambda) = \frac{n}{2} \log |\Lambda/2\pi| - \frac{1}{2} \text{tr} \Lambda S - \frac{n}{2} (\bar{x} - \mu)' \Lambda (\bar{x} - \mu)
\]

In this section we’ll consider only the “length” measurements of the two sons, \( X_1 \) and \( X_3 \). We will test each of the null hypotheses

\[
H_0^1 : \mu_1 = 180 \\
H_0^2 : \mu_3 = 180 \\
H_0^3 : \mu_1 = \mu_3 = 180
\]

against the omnibus alternative—first for known \( \Lambda \), then for unknown. For now we’ll follow the sampling-theory paradigm and find \( P \)-values for these
hypotheses on the basis of the \( n = 25 \) observations of the \( p = 2 \)-dimensional data \( \{x_1, x_3\} \), with summary statistics

\[
\bar{x} = \begin{bmatrix} \bar{x}_1 = 185.72 \\ \bar{x}_3 = 183.84 \end{bmatrix} \quad \frac{1}{2} \Sigma = \begin{bmatrix} 91.481 & 66.875 \\ 66.875 & 96.775 \end{bmatrix}.
\]

1.1 Likelihood Ratio Tests

Each of our hypotheses will be of the form \( H_j : \theta \in \Theta_j \) for some set \( \Theta_j \subset \Theta \) of possible parameters \( \theta \) governing the distribution of the observables through their joint pdf \( f(x \mid \theta) \). The traditional sampling-theory approach to testing a hypothesis \( H_0 \) of this form against an alternative \( H_1 \) is to construct the (generalized) likelihood ratio against the null

\[
B(x) := \frac{\sup_{\theta \in \Theta_1} f(x \mid \theta)}{\sup_{\theta \in \Theta_0} f(x \mid \theta)}
\]

or, equivalently, twice its logarithm, the deviance

\[
\delta(x) = 2[\ell_j^*(x) - \ell_0^*(x)]
\]

where

\[
\ell_j(x) = \log \sup_{\theta \in \Theta_j} f(x \mid \theta)
\]

for \( j = 0, 1 \), and “reject” \( H_0 \) for sufficiently large values of \( B(x) \) (or of \( \delta(x) \))—say, for \( \delta(x) \geq c \). The significance level of the test is the maximum rejection probability \( P[\ell(X) \geq c \mid \theta] \) if the hypothesis is true (i.e. for \( \theta \in \Theta_0 \)), while the “P-value” is \( P(x) = \sup_{\theta \in \Theta_0} P[\delta(X) \geq \delta(x) \mid \theta] \) for the observed data value \( x \), the maximum probability of observing \( B(x) \) (or \( \delta(x) \)) at least this large if \( H_0 \) is true.

Under suitable regularity conditions (asymptotic normality and a bit more), if \( \Theta_0 \subset \Theta_1 \subset \mathbb{R}^q \) with \( \dim(\Theta_0) = r < q \), the asymptotic distribution of \( \delta(x) \) for large sample-size \( n \) is

\[
\delta(x) \Rightarrow \chi^2_{q-r}.
\]

1.2 One-dimensional Hypotheses, known \( \Lambda \)

First consider only the first son’s head width, \( X_1 \), and hypothesis \( H_1^1 \) that its mean is \( \mu_1 = 180 \). If we are given the precision—say, \( \sigma_1^{-2} = 1/100 \)—
then the maximum log likelihoods under $H_0^1$ : $\mu_1 = 180$ and its alternative $H_1$ : $\mu_1 \in \mathbb{R}$ are $\log f(x | \theta_j)$ where $\theta_j$ is the MLE under the restriction $\theta \in \Theta_j$.

$$
\ell_0^\theta = \frac{n}{2} \log(\chi^2 / 2\pi) - \frac{1}{2} \chi^2 S - \frac{n}{2}(\bar{x}_1 - 180)^2 / \chi^2 S
$$

$$
= \frac{n}{2} \log \frac{0.01}{2\pi} - \frac{1}{2} 0.01 S - \frac{25}{2} (185.72 - 180)^4 0.01 (185.72 - 180)
$$

$$
\ell_1^\theta = \frac{n}{2} \log \frac{0.01}{2\pi} - \frac{1}{2} 0.01 S
$$

and hence

$$
\delta = 2[\ell_1^\theta - \ell_0^\theta]
$$

$$
= n \Lambda (\bar{x} - 180)^2 = 0.25 \times 5.72^2 = 8.1796
$$

Since $\Theta_0$ is $r = 0$-dimensional and $\Theta_1$ is $q = 1$-dimensional, $\delta(x)$ has approximately a $\chi^2_1$ distribution under the null hypothesis and so the $P$-value would be approximately $P[\chi^2_1 > 8.1796] = 2\Phi(-\sqrt{8.1796}) = 0.004236$, so the hypothesis would be rejected at level $\alpha = 0.01$. The critical values of $\delta(x)$ for rejecting at levels $\alpha = 0.01$ and $\alpha = 0.05$ would be $2.58^2 = 6.635$ and $1.96^2 = 3.841$, respectively.

Similarly, the hypothesis $H_0^2$ : $\mu_3 = 180$ would have

$$
\delta(x) = 2[\ell_1^\theta - \ell_0^\theta] = n \Lambda (\bar{x}_3 - 180)^2 = 0.25 \times 3.84^2 = 3.6864,
$$

leading to $P$-value $P(x) = 2\Phi(-\sqrt{3.6864}) = 0.0549$, so $H_0^2$ cannot be rejected at level $\alpha = 0.05$ (or at any lower level like $\alpha = 0.01$, of course).

1.2.1 Composite Hypothesis $H_0^3$

How can we test the $p = 2$-dimensional hypothesis $H_0^3$ : $\mu_1 = \mu_3 = 180$?

Simply noting that one of the two one-dimensional hypotheses was rejected at level $\alpha = 0.01$ is not enough to reject $H_0^3$ at that level because of the “multiple comparisons” issue— the probability of rejecting at least one of $k$ hypotheses at level $\alpha$ may have probability greater than $\alpha$ if $H_0$ is true. By subadditivity it can’t have probability more than $k \times \alpha$, though, so the naïve Bonferroni multiple-comparison correction is valid— reject $H_0^3$ at level $\alpha$ if either $H_0^1$ or $H_0^2$ can be rejected at level $\alpha / 2$. Somewhat better are any of:

1. If $x_1$ and $x_3$ are independent, the probability of rejecting either at level $\gamma$ is $[1 - (1 - \gamma)^2]$ if $H_0^3$ is true, which will be no more than
\[ \alpha \text{ if we take } \gamma = 1 - \sqrt{1 - \alpha}; \text{ thus we can reject at levels } \alpha = 0.01 \text{ or } \alpha = 0.05 \text{ if either individual hypothesis may be rejected at level } \gamma = 1 - \sqrt{1 - \alpha} = 0.00501 \text{ or } 0.0253, \text{ respectively (slightly higher than Bonferroni).} \]

2. Under \( H_0^3 \), each of \( z_i := \sqrt{n} \bar{x}_i - 180 \) has a standard normal \( \text{No}(0, 1) \) distribution, hence so too does \( (z_1 + z_2)/\sqrt{2} \); a valid test of \( H_0^3 \) could be based on \( P\text{-value } 2\Phi(-|z_1 + z_2|/\sqrt{2}) \). For these data \( z_1 = 2.86 \) and \( z_2 = 1.92 \), and hence \( z^* = (z_1 + z_2)/\sqrt{2} = 3.380 \) would lead to \( P(x) = 7.25 \times 10^{-4} \) and rejection of \( H_0^3 \).

3. With \( z_j \) as above, under \( H_0^3 \) the test statistic \( Y = (z_1)^2 + (z_3)^2 \) has a \( \chi^2 \) distribution, leading to \( P(x) = \exp(-Y/2) = e^{-5.933} = 0.00265 \), and rejection again at level \( \alpha = 0.01 \).

### 1.2.2 LLR for Composite Hypothesis \( H_0^3 \)

A more principled approach is to compute the log likelihood ratio for the \( r = 0 \)-dimensional hypothesis \( H_0^3 \) and its \( q = 2 \)-dimensional alternative:

\[
\ell_0 = \frac{n}{2} \log |\Lambda/2\pi| - \frac{1}{2} \text{tr } \Lambda S - \frac{n}{2}(\bar{x} - \mu_0)^\top \Lambda (\bar{x} - \mu_0)
\]

\[
= \frac{n}{2} \log \begin{bmatrix} \frac{0.01}{2\pi} & 0 \\ 0 & \frac{0.01}{2\pi} \end{bmatrix} - \frac{1}{2} \text{tr } \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} 91.481 & 66.875 \\ 66.875 & 54.278 \end{bmatrix} - \frac{25}{2} \begin{bmatrix} 5.72 & 3.84 \\ 3.84 & 3.84 \end{bmatrix} \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix} \]

\[
\ell_1 = \frac{n}{2} \log \begin{bmatrix} \frac{0.01}{2\pi} & 0 \\ 0 & \frac{0.01}{2\pi} \end{bmatrix} - \frac{1}{2} \text{tr } \begin{bmatrix} 0.01 & 0 \\ 0 & 0.01 \end{bmatrix} \begin{bmatrix} 91.481 & 66.875 \\ 66.875 & 54.278 \end{bmatrix} - 0
\]

and hence

\[ \delta(x) = 0.25(5.72^2 + 3.84^2) = 11.866, \]

leading (as in 3. above) to \( P(x) = \exp(-11.866/2) = 0.00265 \).

### 1.2.3 Confidence Ellipses

The same calculations lead to confidence ellipses of the form

\[ C_{1-\alpha}(x) = \{ \mu : n(\bar{x} - \mu)^\top \Lambda (\bar{x} - \mu) \leq c_\alpha \} \]
for \( c_\alpha \) chosen such that \( \Pr[\delta(x) > c_\alpha \mid H_0] = \alpha \); in this problem \( c_\alpha = -2 \log \alpha \), so for example the 95% ellipse is

\[
C_{0.05} = \{ \mu : 25[(\mu_1 - 185.72)^2/100 + (\mu_2 - 183.84)^2/100] \leq 5.99 \}
\]

\[
= \{ \mu : (\mu_1 - 185.72)^2 + (\mu_2 - 183.84)^2 \leq 23.966 \},
\]

the circle of radius 4.8955 centered at \([\bar{x}_1, \bar{x}_3]'\).

### 1.3 Unknown Precision

Now consider the same problem with \( \Lambda \) unknown.

**Lemma 1.** If \( D \in S^p_+ \) and \( n > 0 \) then the function of \( G \in S^p_+ \)

\[
f(G) = -n \log |G| - \text{tr} G^{-1}D
\]

attains its maximum value at \( \hat{G} = \frac{1}{n} D \), and there takes the value \( f(\hat{G}) = np \log n - n \log |D| - np \).

**Proof.** Let \( D = EE' \) and set \( H := E'G^{-1}E \); then \( G = EH^{-1}E' \), so

\[
|G| = |E||H^{-1}||E'| = |D|/|H|,
\]

and

\[
\text{tr} G^{-1}D = \text{tr} G^{-1}EE' = \text{tr} E'G^{-1}E = \text{tr} H,
\]

so we can rewrite \( f(G) = g(H) \) with

\[
g(H) = -n \log |D| + n \log |H| - \text{tr} |H|.
\]

Now write \( H = TT' \) with \( T \) lower-triangular; then the maximum of

\[
g(H) = -n \log |D| + n \log |T|^2 - \text{tr} TT'
\]

\[
= -n \log |D| + \sum_{i=1}^p (n \log t_{ii}^2 - t_{ii}^2) - \sum_{i > j} t_{ij}^2
\]

occurs at \( t_{ii}^2 = n \) and \( t_{ij} = 0, i \neq j \), or \( H = nI \). Then \( G = \frac{1}{n} EE' = \frac{1}{n} D \).
As functions of $\Sigma = \Lambda^{-1}$, twice the log likelihood $2\ell(\mu, \Lambda)$ is of the form considered in Lemma (1) under both $H_0$ and $H_1$; thus

$$
\ell(\mu, \Lambda) = -\frac{np}{2} \log 2\pi - \frac{n}{2} \log |\Lambda| - \frac{1}{2} \text{tr} \Lambda [S + n(\bar{x} - \mu)(\bar{x} - \mu)' - n \mu S^{-1} \mu']
$$

$$
\ell^*_0 = \sup_{\Lambda \in \mathbb{F}_2^+} \ell(\mu_0, \Lambda) = \ell(\mu_0, n(S + n dd')^{-1}) \quad \text{where} \quad d := (\bar{x} - \mu)
$$

$$
\ell^*_1 = \max_{\mu \in \mathbb{R}^2, \Lambda \in \mathbb{F}_2^+} \ell(\mu, \Lambda) = \ell(\bar{x}, nS^{-1})
$$

$$
\delta(x) = 2[\ell(\bar{x}, nS^{-1}) - \ell(\mu_0, n(S + n dd')^{-1})] = n \log |S + n dd'| - n \log |S|,
$$

a monotone increasing function $\delta(x) = n \log R$ of

$$
R = \frac{|S + n(\bar{x} - \mu)(\bar{x} - \mu)'|}{|S|} = \frac{|I + n S^{-1/2}(\bar{x} - \mu)(\bar{x} - \mu)'S^{-1/2}|}{1 + n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu)}
$$

(Footnote 1)

$$
T^2 := n(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu) \quad \text{with} \quad \nu := n - 1
$$

has Hotelling’s $^aT_p^2(\nu)$ distribution, while $\frac{n-1}{p}(\bar{x} - \mu)'S^{-1}(\bar{x} - \mu)$ has Snedecor’s $F_{n-p}^p$. For these data,

$$
F = \frac{n - p}{p(n - 1)} T^2 = \frac{23}{2} [5.72 \ 3.84] \begin{bmatrix} 0.02030 & -0.01403 \\ -0.01403 & 0.02003 \end{bmatrix} \begin{bmatrix} 5.72 \\ 3.84 \end{bmatrix} = 3.947
$$

leading to an exact $P$-value of $P(x) = \text{Pr}[F_{23}^2 > 3.947] = 0.0336$, with rejection at $\alpha = 0.05$ but not at $\alpha = 0.01$.

\footnote{All eigenvalues are one for this matrix except that for eigenvector $S^{-1/2}(\bar{x} - \mu)$}
The deviance here was \( \delta(x) = n \log \left(1 + n (\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu) \right) = 7.3768 \), leading to an approximate \( P \)-value of \( P(x) \approx \exp(-7.3768/2) = 0.025 \), which would lead to the same conclusions. Confidence ellipses are again available: for example, since \( P[F_{23}^2 > 3.422] = 0.05 \) and \((2/23) \times 3.422 = 0.2975767\), a 95% confidence set can be constructed as

\[
C_{0.95}(x) = \left\{ \mu : (\bar{x} - \mu)^T S^{-1} (\bar{x} - \mu) \leq \frac{p c_0}{n - p} \right\}
\]

\[
= \left\{ \mu : \begin{bmatrix} \mu_1 - 185.72 \\ \mu_3 - 183.84 \end{bmatrix}^T \begin{bmatrix} 0.02030 & -0.01403 \\ -0.01403 & 0.02003 \end{bmatrix} \begin{bmatrix} \mu_1 - 185.72 \\ \mu_3 - 183.84 \end{bmatrix} \leq 0.2976 \right\}
\]

where \( c_0 = 0.3422 \) is the appropriate critical value of the \( F_{n-p}^p \) distribution. This also leads to simultaneous 95% confidence intervals for all possible linear combinations \( \alpha_1 \mu_1 + \alpha_2 \mu_2 \) (for example, for \( [\mu_2 - \mu_1] \) and \( [\frac{\mu_1 + \mu_2}{2}] \)).

References
