Ma 494 — Theoretical Statistics

Solutions for Problem Set #6 — Due April 30, 2010

Prof. Sawyer — Washington University

NOTE: 5 problems on 3 pages. Different parts of problems may not be equally weighted.

1. From the text (page 574), $S_Y = 0.20$ and $S_X = 0.37$. Thus the observed statistic $T = (S_Y/S_X)^2 = (0.20/0.37)^2 = 0.29218$. Since the sample sizes $n_X = n_Y = 10$, T has an F-distribution $T \approx F_{9,9}$ given $H_0 : \tau = 1$. Since we are testing for $\tau = \sigma_Y^2/\sigma_X^2 < 1$, the P-value is $P(F_{9,9} \le T_{\rm obs}) = P(F_{9,9} \le 0.29218) = 0.0405$. Thus the second group has a significantly smaller variance at the level of significance $\alpha = 0.05$.

2. Break up the range $0 \le Y \le 3$ into three cells $0 \le Y \le 1$, $1 \le Y \le 2$, and $2 \le Y \le 3$. Then, given H_0, the three cell probabilities are $p_1 = \int_0^1 (1/9)y^2\,dy = 1/27$, $p_2 = \int_1^2 (1/9)y^2\,dy = (2^3-1)/27 = 7/27$, and $p_3 = \int_2^3 (1/9)y^2\,dy = (3^3-2^3)/27 = 19/27$. For $n = 50$ individuals, the expected numbers are $E_i = (50)p_i$ and the three summands for the Pearson chi-square statistic are $Y_i = (\text{Obs}_i - E_i)^2/E_i$ so that the chi-square statistic is $X = Y_1 + Y_2 + Y_3$. This leads to the table

Counts :	8	16	26	Total : 50
p_i :	0.0370	0.2593	0.7037	Total : 1.00
E_i :	1.852	12.963	35.185	Total : 50.00
Y_i :	20.412	0.712	2.398	Total : 23.531

If we don't combine the first two cells, then the Pearson chi-square P-value is $P = P(\chi_2^2 \ge 23.531) = 0.000008$. Even though a simulation test using the Pearson test statistic as a score gave $P = 0.000125$, which is statistically valid, we can't use Pearson's chi-square approximation for the value of his statistic to reject H_0 because of the small value of $E_1 = np_1 = 1.852$. (In fact, the simulation shows that the chi-square approximate P-value is too small.)

However, the table with three cells does give qualitative information about what is going on: The chi-square values Y_i suggest that the problem is the observed excess in Cell 1, which suggests a relatively larger number of defendants with sentences of one year or less. This could be due to either (i) prisoners who would have served for more than one year are receiving a shorter sentence, (ii) defendants who would not have been sent to prison earlier are now being given sentences of up to one year, or of course (iii) both.

If we combine cells 1 and 2, we get

Counts :	24	26	Total : 50
p_i :	0.2963	0.7037	Total : 1.00
E_i :	14.815	35.185	Total : 50.00
Y_i :	5.695	2.398	Total : 8.093

This leads to P-value $P = P(\chi_1^2 \geq 8.093) = 0.0044$, which is highly significant. (The same simulation gave $P = 0.005388$ with 1,000,000 simulations. This is significantly higher than $P = 0.0044$, but is in the correct ball park.)

This is also due to the first cell, now for sentences up to two years. Again, this could also be due either to defendants with a sentence longer than two years being given a shorter sentence, defendants who earlier would not have been sent to prison being given a sentence of two years or less, or some of each.

3. Here the alternative H_1 is that the three cell probabilities $p_1 = p_{AA}$, $p_2 = p_{Aa}$, and $p_3 = p_{aa}$ are arbitrary subject to $p_1 + p_2 + p_3 = 1$, and H_0 is that they are of the form $p_1 = p_{AA} = p^2$, $p_2 = p_{Aa} = 2pq$, and $p_3 = p_{aa} = q^2$ for some unknown value of p with $q = 1 - p$. The Pearson chi-square statistic is

$$X = \sum_{i=1}^{3} \frac{(X_i - n\widehat{p}_i)^2}{n\widehat{p}_i}$$

where X_i are the observed counts and $\widehat{p}$ is the maximum-likelihood estimator of p given H_0. Given H_0 (for some unknown p), X has an asymptotic χ^2 distribution with $df = 3 - 1 - 1 = 1$ degree of freedom, where we subtract one degree of freedom for the estimated parameter.

The likelihood of p given H_0 is

$$\begin{aligned} L(p) &= \prod_{i=1}^{3} p_i^{X_i} = (p^2)^{X_1}\big(2p(1-p)\big)^{X_2}\big((1-p)^2\big)^{X_3} \\ &= p^{2X_1+X_2}(1-p)^{X_2+2X_3}\, 2^{X_2} \end{aligned}$$

for $0 < p < 1$. Solving

$$\frac{d}{dp}\log L(p) = \frac{2X_1 + X_2}{p} - \frac{X_2 + 2X_3}{1-p} = 0$$

leads to $\widehat{p} = (2X_1 + X_2)/(2(X_1 + X_2 + X_3)) = (36 + 86)/300 = 122/300 = 0.4067$. In turn, this leads to the table

	AA	Aa	aa	
Counts :	18	86	46	Total : 150
$\widehat{p}_i$:	0.1654	0.4826	0.3520	Total : 1.00
E_i :	24.807	72.387	52.807	Total : 50.00
Y_i :	1.868	2.560	0.877	Total : 5.305

Given H_0, the Pearson statistic $X \approx \chi_1^2$ as indicated above. The P-value is

$$P = P(\chi_1^2 \geq 5.305) = P(|Z| \geq 2.303) = 2P(Z \geq 2.303) = 0.02126$$

using the fact $\chi_1^2 \approx Z^2$ for a standard normal random variable Z to make it easier to find P. This is significant at $\alpha = 0.05$ and we reject H_0. The high value of X appears to be due to an excess of Aa individuals after $p = \widehat{p}$ is estimated, and may be an example of "hybrid vigor".

4. Given $Y_i = \beta X_i + \sigma Z_i$ for constant X_i and standard normal Z_i, we have $Y_i \approx N(\beta X_i, \sigma^2)$ and the likelihood of (β, σ) given $Y = (Y_1, Y_2, \dots, Y_n)$ is

$$\begin{aligned} L(\beta,\sigma,Y) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{\frac{1}{2}\sigma^2}(Y_i-\beta X_i)^2\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i-\beta X_i)^2\right) \end{aligned}$$

Thus

$$\log L(\beta,\sigma,Y) \;=\; C \;-\; n\log\sigma \;-\; \frac{1}{2\sigma^2}\sum_{i=1}^n (Y_i-\beta X_i)^2$$

For fixed σ, this is maximized when the sum is minimized, or at the solution of

$$\begin{aligned} \frac{\partial}{\partial\beta}\sum_{i=1}^n (Y_i-\beta X_i)^2 \;&=\; \sum_{i=1}^n \frac{\partial}{\partial\beta}(Y_i-\beta X_i)^2 \;=\; \sum_{i=1}^n (-2X_i)(Y_i-\beta X_i) \\ =\; -2\left(\sum_{i=1}^n X_iY_i \;-\; \beta\sum_{i=1}^n X_i^2\right) \;&=\; 0 \end{aligned}$$

Thus the MLE of β is $\widehat{\beta} = \left(\sum_{i=1}^n X_iY_i\right) \big/ \left(\sum_{i=1}^n X_i^2\right)$. Similarly

$$\frac{\partial}{\partial\sigma}\log L(\widehat{\beta},\sigma,Y) \;=\; -\frac{n}{\sigma} \;+\; \frac{1}{\sigma^3}\sum_{i=1}^n (Y_i-\widehat{\beta}X_i)^2 \;=\; 0$$

and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (Y_i-\widehat{\beta}X_i)^2$. Since $Y_i = \beta X_i + \sigma Z_i$,

$$\widehat{\beta} \;=\; \frac{\sum_{i=1}^n X_iY_i}{\sum_{i=1}^n X_i^2} \;=\; \frac{\sum_{i=1}^n X_i(\beta X_i+\sigma Z_i)}{\sum_{i=1}^n X_i^2} \;=\; \beta \;+\; \sigma\,\frac{\sum_{i=1}^n X_iZ_i}{\sum_{i=1}^n X_i^2}$$

Set

$$W_1 \;=\; C(\widehat{\beta}-\beta)/\sigma \;=\; \frac{\sum_{i=1}^n X_iZ_i}{C}, \qquad C = \sqrt{\sum_{i=1}^n X_i^2}$$

Then $\mathrm{Var}(W_1) = \sum_{i=1}^n X_k^2/C^2 = 1$, W_1 is standard normal, and $\widehat{\beta}-\beta = \sigma W_1/C$.

Use Gram-Schmidt orthogonalization or something similar to construct an $n\times n$ matrix

$$A = \begin{pmatrix} X_j/C \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix} \qquad (1 \le j \le n,\ \text{rows } 1 \le i \le n)$$

where the row vectors $a_1 = (X_j/C)$ and $a_i = (a_{ij})$ for $2 \le i \le n$ are an orthonormal basis for R^n. Then A is an orthogonal or rotation matrix. Set

$$W \;=\; \begin{pmatrix} W_1 \\ \vdots \\ W_n \end{pmatrix} \;=\; A \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix}$$

Note that $W_1 = \sum_{i=1}^n X_i Z_i/C = C(\widehat{\beta} - \beta)/\sigma$ is the same as before. Since A is an orthogonal matrix, $W_1, \ldots, W_n$ are independent standard normal random variables and

$$\|AZ\|^2 \;=\; \|Z\|^2 \;=\; \sum_{i=1}^n W_i^2 \;=\; \sum_{i=1}^n Z_i^2$$

From before, $\widehat{\sigma}^2 = (1/n)\sum_{i=1}^n (Y_i - \widehat{\beta}X_i)^2$. Since $\widehat{\beta} - \beta = \sigma(W_1/C)$,

$$\begin{aligned}
\sum_{i=1}^n (Y_i - \widehat{\beta}X_i)^2 \;&=\; \sum_{i=1}^n (\beta X_i + \sigma Z_i - \widehat{\beta}X_i)^2 \;=\; \sum_{i=1}^n (\sigma Z_i - (\widehat{\beta} - \beta)X_i)^2 \\
&= \sum_{i=1}^n (\sigma Z_i - \sigma(W_1/C)X_i)^2 \;=\; \sigma^2 \sum_{i=1}^n \Big(Z_i^2 - 2Z_i(W_1/C)X_i + (W_1/C)^2 X_i^2\Big) \\
&= \sigma^2 \left(\sum_{i=1}^n Z_i^2 \;-\; 2(W_1/C)\sum_{i=1}^n Z_i X_i \;+\; (W_1^2/C^2)\sum_{i=1}^n X_i^2 \right) \\
&= \sigma^2 \left(\sum_{i=1}^n Z_i^2 \;-\; 2(W_1/C)(CW_1) \;+\; (W_1^2/C^2)C^2 \right) \;=\; \sigma^2 \left(\sum_{i=1}^n Z_i^2 \;-\; W_1^2 \right) \\
&= \sigma^2 \left(\sum_{i=1}^n W_i^2 \;-\; W_1^2 \right) \;=\; \sigma^2 \left(\sum_{i=2}^n W_i^2 \right)
\end{aligned}$$

since $W_1 = \sum_{i=1} X_i Z_i/C$. Thus

$$\sum_{i=1}^n (Y_i - \widehat{\beta}X_i)^2/\sigma^2 \;=\; n\widehat{\sigma}^2/\sigma^2 \;=\; \sum_{i=2}^n W_i^2 \;\approx\; \chi^2_{n-1}$$

Since $W_1 = C(\widehat{\beta} - \beta)/\sigma$, $\widehat{\beta} = \beta + \sigma W_1/C$ is indepedent of $\widehat{\sigma}^2$. This completes the second part of the problem.

5. Let $1 \le a \le 3$ denote the rows of the table, $1 \le b \le 4$ the columns, and r_a $(1 \le a \le 3)$ and c_b $(1 \le n \le 4)$ the row and column sums. The sequence (X_i, Y_i)

$(1 \le i \le n)$ represents the bivariate data that is tabled. Then

$$\sum_{i=1}^{n} X_i = \sum_{a=1}^{3} a r_a = 4413, \qquad \overline{X} = (1/n)\sum_{i=1}^{n} X_i = 4413/2000 = 2.2065$$
$$\sum_{i=1}^{n} Y_i = \sum_{b=1}^{4} b c_b = 5574, \qquad \overline{Y} = (1/n)\sum_{i=1}^{n} Y_i = 5574/2000 = 2.787$$
$$\sum_{i=1}^{n} X_i^2 = \sum_{a=1}^{3} a^2 r_a = 11007, \qquad \overline{X^2} = (1/n)\sum_{i=1}^{n} X_i^2 = 11007/2000 = 5.5035$$
$$\sum_{i=1}^{n} Y_i^2 = \sum_{b=1}^{4} b^2 c_b = 17942, \qquad \overline{Y^2} = (1/n)\sum_{i=1}^{n} Y_i^2 = 14792/2000 = 7.396$$
$$\sum_{i=1}^{n} X_i Y_i = \sum_{a=1}^{3}\sum_{b=1}^{4} ab X_{ab} = 12191, \qquad \overline{XY} = (1/n)\sum_{i=1}^{n} X_i Y_i = 12191/2000 = 6.0905$$

and thus

$$\begin{aligned}
&\text{Xvar} = \overline{X^2} - \overline{X}^2 = 0.6349, \quad \text{Yvar} = \overline{Y^2} - \overline{Y}^2 = 1.2035 \\
&\text{XYcovar} = \overline{XY} - (\overline{X})(\overline{Y}) = -0.0540155 \\
&r = (-0.0540)/\sqrt{(0.6349)(1.2035)} = -0.0618 \\
&r^2 = 0.003818, \qquad X = (n-1)r^2 = 7.63273
\end{aligned}$$

(**Remark.** This uses the population means, variances, and covariance rather than the corresponding sample statistics for simplicity, but that doesn't matter for r^2 since the denominators in the variances and covariance cancel.)

Under the hypothesis H_0 that U_i and V_i are independent ("rows and columns are independent"), Mantel's $X \approx \chi_1^2$, so that the P-value for Mantel's test is

$$P = P(\chi_1^2 \ge 7.63273) = P(|Z| \ge \sqrt{7.63263}) = 2P(Z \ge 2.7672) = 0.00573$$

where we use $\chi_1^2 \approx Z^2$ for a standard normal Z to help compute the P-value. Mantel's test statistic is asymtotically χ^2 with one degree of freedom, and we reject H_0 at level of significance either $\alpha = 0.05$ or $\alpha = 0.01$, even though the same data is not significant for Pearson's chi-square test.