Ma 494 — Theoretical Statistics

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. Consider the heat-loss data in Problem 9.3.5 in the text (p574), which has one column of data for a control group of ten individuals that do not have Reynaud's syndrome and a second column of ten individuals with Reynaud's syndrome.

(By the way, modern medical and statistical practice does not call a control group "Normal Subjects", as the text does, since this suggests that the test group is "abnormal". "Unaffected" is OK. Biologists often use the term "wild type" for unaffected individuals, but this is usually not applied to humans.)

An expert in Reynaud's syndrome has theoretical arguments that suggest that the heat-loss times for the Reynaud's-syndrome subjects should have a smaller VARIANCE than unaffected subjects. Assume that the data for the unaffected subjects are independent normal $N(\mu_1, \sigma_1^2)$ and that for the Reynaud's-syndrome subjects are $N(\mu_2, \sigma_2^2)$. (The expert is not interested in μ_1 or μ_2.)

Let $\tau = \sigma_2^2/\sigma_1^2$ be the relative variance of the Reynaud's-syndrome subjects, so that $\sigma_2^2 = \tau\sigma_1^2$. Use the F-statistic S_Y^2/S_X^2 to test

$$H_0 : \tau = 1 \qquad \text{against} \qquad H_1 : \tau < 1$$

Find or bracket the P-value of the test in this case. (That is, either find the exact P-value to two or three decimal places or else bracket the P-value by saying something like $0.05 < P < 0.10$.)

Is the expert correct that the heat-loss values for the Reynaud's-syndrome subjects have a significantly smaller variance, using the level of significance $\alpha = 0.05$?

2. (Like Problem 10.3.12 in the text, p614–615.) Past experience has been that the time Y served in prison for defendants convicted of grand theft has been $f_Y(y) = (1/9)y^2$ for $0 \le y \le 3$. A review of 50 individuals convicted of this crime in recent years shows that 8 served less than one year, 16 between one and two years, and 26 between two and three years. It was suggested that recent judicial reforms may have altered the lengths of sentences.

(i) Is the recent data consistent with the previous distribution for Y? Find or bracket the P-value. (Hint: Break up the interval [0,3] into three cells.)

(ii) If you used a chi-square cell test in part (i), how many degrees of freedom did you use?

(iii) If the result in part (i) is significant, which time interval contributed the most to the Pearson chi-square statistic? What does this say about how sentencing has changed?

3. At a particular genetic locus (or gene), individuals can be of one of three genotypes, aa, Aa, or AA. A genetic locus is said to be in *Hardy-Weinberg (HW) equilibrium* at this locus if the population proportions of the three genotypes (aa, Aa, and AA) are in the ratios $p^2 : 2pq : q^2$ where $0 < p < 1$ and $q = 1-p$. Mendel's laws say that controlled experimental samples should follow HW proportions. Possible causes of deviation from HW equilibrium in a natural population are migration, inbreeding, or population heterogeneity, all of which can cause a relative deficit of Aa individuals, or certain types of selection that cause an excess of Aa individuals. (Animal breeders like the term "hybrid vigor" for an excess of Aa individuals.)

A sample of $n = 150$ individuals were typed at this locus. The results were

Table 1. Distribution of a Natural Sample

Genotypes	aa	Aa	AA	Sum
Counts	18	86	46	150

(i) Test whether or not the locus is in HW equilibrium. What is the P-value (or else bracket the P-value as in Problem 1)?

(ii) If you used a chi-square cell test in part (i), how many degrees of freedom did you use?

(iii) If the locus is not in HW equilibrium, which of the three genotypes makes the largest contribution to the chi-square statistic? Is there an excess or a deficit in the number of individuals with this genotype in comparison with HW proportions?

4. Let (Y_i, X_i) be observations for $1 \le i \le n$ with $Y_i, X_i > 0$ such that the X_i are considered deterministic. Assume

$$Y_i = \beta X_i + \sigma Z_i \qquad \text{where} \;\; Z_i \;\; \text{are independent } N(0,1)$$

This describes a linear regression line that is forced to go through $(Y, X) = (0, 0)$, which is sometimes used to calibrate measurements by two different methods.

(i) Find the MLEs $\widehat{\beta}$ and $\widehat{\sigma}$.

(ii) Show that $\widehat{\beta}$ and $\widehat{\sigma}^2$ are independent.

(iii) Show that $n\widehat{\sigma}^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom. (That is, $(n/(n-1))\widehat{\sigma}^2 \approx \sigma^2\chi^2_{n-1}/(n-1)$.)

(*Hint*: Follow the proof for the independence for $\overline{X}$ and S_X^2 for a single normal sample or else for $\overline{Y}$, $\widehat{\beta}$, and SSE for the regression $Y_i = \mu + \beta X_i + \sigma Z_i$.)

5. Two thousand (2000) individuals were classified in two different ordinal classifications, property U with values 1,2,3 and property V with levels 1,2,3,4. (Think of height and weight ranges for 2000 marmots, or else book values and market capitalization for 2000 companies.)

An expert wants to test the hypothesis that U and V are independent.

The basic data in this case is $Y_i = (U_i, V_i)$ $(1 \leq i \leq 2000)$ for 2000 marmots or 2000 companies, where $1 \leq U_i \leq 3$ and $1 \leq V_i \leq 4$. Since there are $3 \times 4 = 12$ possible paired values (U_i, V_i) and 2000 observations, it is convenient to store the data as a 3×4 contingency table:

Table 2. Contingeny table for 2000 (U, V) values

	V:	1	2	3	4	Sums:
	1	66	98	127	180	471
U:	2	111	136	170	228	645
	3	168	193	240	283	884
Sums:		345	427	537	691	2000

An associate of the expert carries out Pearson's standard chi-square test for independence, and obtains $X = 8.3105$ for $(r-1)(c-1) = 6$ degrees of freedom. This leads to the non-significant P-value $P = 0.2162$.

However, the expert has theoretical reasons for suspecting that the variables U_i and V_i are correlated and decides to use Mantel's chi-square trend test instead. Mantel's test has the implicit alternative H_1 that U_i and V_i are correlated, as opposed to the Pearson chi-square test for independence, which, being the GLRT test, has an omnibus alternative, or alternatively the implicit alternative H_1 that the cell frequencies are exactly the cell proportions $p_{uv} = \widehat{p}_{uv} = X_{uv}/n$ where X_{uv} are the cell counts in Table 2. Both Pearson's and Mantel's test have the same null hypothesis H_0 that the U_i and V_i are independent multivariate Bernoulli random variables.

Use Mantel's trend test to find out whether the data in Table 2 comes from a population for which (U, V) are independent. What is the P-value (or bracket the P-value)? If the test distribution given H_0 has a chi-square distribution, what is the number of degrees of freedom?

(*Hints*: (1) See class notes or Section 9 in the Math 494 notes for a discussion of Mantel's trend test.

(2) It may be easier to express sample moments like $\overline{U} = (1/n)\sum_{i=1}^{n} U_i$ and $\overline{UV} = (1/n)\sum_{i=1}^{n} U_i V_i$ in terms of the tabled values in Table 2.

(3) Try coding $U = 0, 1, 2$ and $V = 0, 1, 2, 3$ instead of $1, 2, 3$ and $1, 2, 3, 4$. The value of r will be the same but sample means and variances may be easier to calculate. Be careful!)