AM 166 Problem Set 2 2/14/03

10.100 In Ω, $\widehat{\mu}_1 = \overline{Y}_1$, $\widehat{\mu}_2 = \overline{Y}_2$, and

$$\widehat{\sigma}^2 = \frac{\sum (Y_{1j}-\overline{Y}_1)^2 + \sum (Y_{2j}-\overline{Y}_2)^2}{n_1+n_2}$$

Then

$$L(\widehat{\Omega}) = \frac{1}{(2\pi\widehat{\sigma})^{(n_1+n_2)/2}} e^{-(n_1+n_2)/2}$$

In Ω_0, $\mu_1 = \mu_2$ and σ^2 is unknown. Hence

$$L(\Omega_0) = \frac{1}{(2\pi)^{(n_1+n_2)/2}\sigma^{n_1+n_2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum (y_{1j}-\mu)^2 + \sum (y_{2j}-\mu)^2\right]\right\}$$

and

$$\ln L = -\left(\frac{n_1+n_2}{2}\right)\ln 1\pi - \left(\frac{n_1+n_2}{2}\right)\ln\sigma^2 - \frac{1}{2\sigma^2}\left[\sum (y_{1j}-\mu)^2 + \sum (y_{2j}-\mu)^2\right]$$

Taking derivatives with respect to μ and σ^2, we have

$$\frac{d\ln L}{d\sigma^2} = \frac{-(n_1+n_2)}{2\sigma^2} + \frac{\left[\sum (y_{1j}-\mu)^2 + \sum (y_{2j}-\mu)^2\right]}{2\sigma^4} = 0$$

$$\frac{d\ln L}{d\mu} = \frac{\sum (y_{1j}-\mu) + \sum (y_{2j}-\mu)}{\sigma^2} = 0$$

Hence

$$\sum y_{1j} - n_1\mu + \sum y_{2j} - n_2\mu = 0$$

or

$$\widehat{\mu} = \frac{\sum y_{1j} + \sum y_{2j}}{n_1+n_2} = \frac{n_1\overline{y}_1 + n_2\overline{y}_2}{n_1+n_2} \qquad \text{and} \qquad \widetilde{\sigma}^2 = \frac{\sum_i\sum_j (y_{ij}-\widehat{\mu})^2}{n_1+n_2}$$

Finally,

$$L(\widehat{\Omega}_0) = \frac{1}{(2n\widetilde{\sigma}^2)^{(n_1+n_2)/2}} \qquad \text{and} \qquad \lambda = \left(\frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2}\right)^{(n_1+n_2)/2} \le k$$

is the likelihood ratio test.

In order to show that this reduces to the two-sample t test of Section 10.8, define

$$SS_{Y_1} = \sum (Y_{1j} - \overline{Y}_1)^2 \qquad \text{and} \qquad SS_{Y_2} = \sum (Y_{2j} - \overline{Y}_2)^2$$

so that

$$\widehat{\sigma}^2 = \frac{SS_{Y_1} + SS_{Y_2}}{n_1+n_2}$$

Consider

$$\begin{aligned}(n_1+n_2)\widetilde{\sigma}^2 &= \sum\sum (Y_{ij}-\widehat{\mu})^2 = \sum (Y_{1j}-\overline{Y}_1)^2 + n_1(\overline{Y}_1-\widehat{\mu})^2 + \sum (Y_{2j}-\overline{Y}_2)^2 \\ &\quad + n_2(\overline{Y}_2-\widehat{\mu})^2 \\ &= SS_{Y_1} + SS_{Y_2} + n_1\left(\overline{Y}_1 - \frac{n_1\overline{Y}_1+n_2\overline{Y}_2}{n_1+n_2}\right)^2 + n_2\left(\overline{Y}_2 - \frac{n_1\overline{Y}_1+n_2\overline{Y}_2}{n_1+n_2}\right)^2 \\ &= SS_{Y_1} + SS_{Y_2} + \frac{n_1n_2(\overline{Y}_1-\overline{Y}_2)^2}{n_1+n_2}\end{aligned}$$

Now

$$\begin{aligned}\lambda^{2/(n_1+n_2)} = \frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2} &= \frac{SS_{Y_1}+SS_{Y_2}}{SS_{Y_1}+SS_{Y_2}+\left[\frac{n_1n_2(\overline{Y}_1-\overline{Y}_2)^2}{n_1+n_2}\right]} \\ &= \frac{1}{1+\frac{(\overline{Y}_1-\overline{Y}_2)^2}{\left\{\left[\left(\frac{1}{n_1}\right)+\left(\frac{1}{n_2}\right)\right]S^2(n_1+n_2-2)\right\}}} = \frac{1}{1+\frac{t^2}{n_1+n_2-2}}\end{aligned}$$

where $S^2 = \frac{SS_{Y_1}+SS_{Y_2}}{n_1+n_2-2}$. Since we are considering only $\mu_1 > \mu_2$, or equivalently $\overline{Y}_1 - \overline{Y}_2 > 0$, $t = \sqrt{t^2}$ will be positive. Hence, small values of λ imply large positive values of t, and a one-tailed t test is implied.

14.1 One thousand cars were each classified according to the lane that they occupied (1 through 4). The objective is to determine whether or not some lanes were preferred over others. This is a multinomial experiment with $k = 4$ cells. If no lane is preferred over another, the probability that a car will be driven in lane i, $i = 1, 2, 3, 4$, is $\frac{1}{4}$. The null hypothesis is then

$$H_0\text{: } p_1 = p_2 = p_3 = p_4 = \tfrac{1}{4}$$

Notice that the hypothesis to be tested is a test of specified numerical values for the probabilities rather than a test of their relationship to one another. Hence no degrees of freedom are lost for estimating cell probabilities. The test statistic is

$$X^2 = \sum_{i=1}^{k} \frac{[n_i - E(n_i)]^2}{E(n_i)}$$

which, when n is large, will possess an approximate chi-square distribution in repeated sampling. The values of n_i are the actual counts observed in the experiment, and

$$E(n_i) = np_i = 1000\left(\tfrac{1}{4}\right) = 250$$

A table of observed and expected cell counts is shown at the right. Then

	Lane 1	Lane 2	Lane 3	Lane 4
n_i	294	276	238	192
$E(n_i)$	250	250	250	250

$$X^2 = \frac{(294-250)^2}{250} + \frac{(276-250)^2}{250} + \frac{(238-250)^2}{250} + \frac{(192-250)^2}{250} = \frac{6120}{250} = 24.48$$

To obtain the rejection region for this test, the degrees of freedom associated with X^2 must be determined. The number of degrees of freedom is equal to the number of cells, k, less 1 degree of freedom for each linearly independent restriction placed on $n_1, n_2, \ldots, n_k$. For this example, $k = 4$ and one degree of freedom is lost because of the restriction that $\sum_i n_i = n$. Hence X^2 has $(k-1) = (4-1) = 3$ degrees of freedom and the appropriate upper-tailed rejection region is

$$X^2 \geq X^2_{3,.005} = 7.81$$

Thus the conclusion is to reject the null hypothesis, with a probability of error equal to $\alpha = .05$. Remember that a one-tailed test is employed, using the upper-tail values of X^2, because large deviations of the observed cell counts will tend to contradict H_0. Hence we will reject the null hypothesis when X^2 is large. Since $24.48 > 12.8381 = X^2_{3(.005)}$, the p-value is less than .005.

14.2 If the frequency of occurrence of a heart attack is the same for each day of the week, then when a heart attack occurs, the probability that it falls in one cell (day) is the same as for any other cell (day). Hence,

$$H_0\colon\ p_1 = p_2 = \ldots = p_7 = \tfrac{1}{7}$$

vs.

$$H_a\colon \text{ at least one } p_i \text{ is different from the others}$$

or equivalently,

$$H_a\colon\ p_i \neq p_j \text{ for some pair } i \neq j.$$

Since $n = 200$,

$$E(n_i) = np_i = 200\left(\tfrac{1}{7}\right) = 28.571429$$

and the test statistic is

$$X^2 = \tfrac{(24-28.571429)^2}{28.571429} + \ldots + \tfrac{(29-28.571429)^2}{28.571429} = \tfrac{103.71429}{28.571429} = 3.63.$$

The degrees of freedom for this test of specified cell probabilities are $k - 1 = 7 - 1 = 6$, and the upper-tailed rejection region is

$$X^2 > \chi^2_{6,.05} = 12.59.$$

and H_0 is not rejected. There is insufficient evidence to indicate a difference in frequency of occurrence from day to day.

14.3 **a.** Let p denote the true proportion of heart attacks occurring on Monday. The hypothesis to be tested is

$$H_0\colon\ p = \tfrac{1}{7} \qquad \text{vs.} \qquad H_a\colon\ p > \tfrac{1}{7}$$

From Section 8.3 and 10.3, the observed value of the test statistic is

$$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.18-\frac{1}{7}}{\sqrt{\frac{\left(\frac{1}{7}\right)\left(\frac{6}{7}\right)}{200}}} = 1.50.$$

From Table IV, Appendix 3, we see that $P(z > 1.645) = .05$. Hence, we take $z > 1.645$ as the rejection region. Since the observed value of z is not in the rejection region, there is not sufficient evidence of heart attacks being more likely to occur on Monday than any other day of the week.

b. The test is suggested by the data. Our hypothesis should test a question of interest formulated before the experiment is conducted. Data is then gathered to support or refute the given hypothesis. That is, we should apply the scientific method.

c. Monday is popularly known as the most stressful workday of the week; one has a long five days until the weekend. One might wish to investigate if this extra stress results in a disproportionate amount of heart attacks.

14.5 Similar to previous exercises. The null hypothesis to be tested is

$$H_0\colon\ p_1 = .69;\ p_2 = .21;\ p_3 = .07;\ p_4 = .03$$

against the alternative that at least one of these probabilities is incorrect. The observed and expected cell counts are shown below.

User	1	2	3	4
n_i	102	32	12	4
$E(n_i)$	103.5	31.5	10.5	4.5

The test statistic is

$$X^2 = \tfrac{(102-103.5)^2}{103.5} + \ldots + \tfrac{(4-4.5)^2}{4.5} = .2995$$

and the p-value with $k - 1 = 3$ d.f. is p-value $> .95$. The null hypothesis is not rejected and we cannot conclude that the figures given are inaccurate.

14.8 This is similar to Exercise 14.7. Calculate

$$\widehat{\lambda} = \overline{y} = \frac{0(296)+1(74)+\ldots+8(1)}{414} = 483.09$$

The observed and estimated cell counts are given in the table at the right.

y	n_i	$\widehat{p}_i$	$\widehat{E}(n_i)$
0	296	.6169	255.38
1	74	.298	123.38
2	26	.072	29.80
3 or more	18	.0131	5.44

Then

$$X^2 = \frac{(296-255.38)^2}{255.38} + \ldots + \frac{(18-5.44)^2}{5.44} = 55.71$$

The rejection region with $k-2=2$ degrees of freedom is $X^2 \geq 5.99$, and the null hypothesis is rejected. The data do not come from a Poisson distribution.

14.9 a. The table of estimated expected cell counts is

	JAS Score		
3-year follow-up	Less than −5	−5 to 5	Greater than 5
Died	17.09	16.24	15.67
Alive	162.91	154.76	149.33

The test statistic is

$$X^2 = \frac{(21-17.09)^2}{17.09} + \frac{(17-16.24)^2}{16.24} + \ldots + \frac{(154-149.33)^2}{149.33} = 2.56$$

with $(r-1)(c-1) = 2$ d.f. Since

$$X^2 = 2.56 < 5.99 = \chi^2_{.05},$$

we do not reject H_0 at the $\alpha = .05$ level. There is insufficient evidence to indicate a dependence between mortality rate and level of Type A behavior.

b. Since $X^2 = 2.56 < 4.61 = \chi^2_{.10}$, $p > .10$. Thus the results are not significant.

14.10 a. The hypothesis of independence between attachment pattern and child care time is tested using the chi-square statistic. The contingency table, including column and row totals and the estimated expected cell counts in parentheses, follows.

Pattern	0–3 hours	4–19 hours	20–54 hours	Total
Secure	24 (24.09)	35 (30.97)	5 (8.95)	64
Anxious	11 (10.91)	10 (14.03)	8 (4.05)	29
Total	35	45	13	93

The test statistic is

$$X^2 = \frac{(24-24.09)^2}{24.09} + \frac{(35-30.97)^2}{30.97} + \ldots + \frac{(8-4.05)^2}{4.05} = 7.267$$

and the rejection region is $X^2 > \chi^2_{2,.05} = 5.99$ with 2 d.f., H_0 is rejected. There is evidence of a dependence between attachment pattern and child care time.

b. The value $X^2 = 7.267$ is between $\chi^2_{.05}$ and $\chi^2_{.025}$ so that $.025 < p\text{-value} < .05$. The results are significant.