36-720 Homework 1 Solutions
Fall 2007

Problem 1 (10 points):

For a single multinomial distribution, under the null, $p_{ij} = p_i p_j$, so the d.f. is I-1+j-1 as the number of free row parameters is I-1 ($\sum_i p_i = 1$) and the number of free column parameters is J-1 ($\sum_j p_j = 1$). Under the alternative (unrestricted model), the number of free parameters is I*J-1 (one lost for $\sum_{i,j} P_{ij} = 1$). So the d.f. for the single multinomial sampling is d.f. = IJ-1 - (I-1+J-1) = (I-1)(J-1).

Problem 2 (30 points):

a) Let $(N_1, ..., N_J)$ have a Multinomial distribution with parameters $(n, \mathbf{p})$, where $\mathbf{p} = (p_1, ..., p_J)$. The probability mass function is

$$P(N_1 = n_1, \cdots, n_J = n_J) = \frac{n!}{n_1! \cdots n_J!} p_1^{n_1} ... p_J^{n_J}$$

where $\sum_{i=1}^t n_i = n$ and $\sum_{i=1}^t p_i = 1$. Multiply and divide this probability function by $P(N_1 = n_1)$, where N_1 is Binomial(n, p_1). We get,

$$\begin{aligned} P(N_1 = n_1, \cdots, N_t = n_J) &= \frac{n!}{n_1!(n-n_1)!} p_1^{n_1}(1-p_1)^{n-n_1} \frac{n!}{n_1! \cdots n_J!} p_1^{n_1} \cdots p_J^{n_J} \\ & \quad \frac{n_1!(n-n_1)!}{n!} \frac{1}{p_1^{n_1}} \frac{1}{(1-p_1)^{n-n_1}} \\ &= \frac{n!}{n_1!(n-n_1)!} p_1^{n_1}(1-p_1)^{n-n_1} \frac{(n-n_1)!}{n_2! \cdots n_J!} \frac{p_2^{n_2} \cdots p_J^{n_J}}{(1-p_1)^{n-n_1}} \\ &= \frac{n!}{n_1!(n-n_1)!} p_1^{n_1}(1-p_1)^{n-n_1} \frac{(n-n_1)!}{n_2! \cdots n_J!} \left(\frac{p_2}{1-p_1}\right)^{n_2} \cdots \left(\frac{p_J}{1-p_1}\right)^{n_J} \end{aligned}$$

which is the product of a Binomial(n, p_1) times a Multinomial$(n - n_1, \mathbf{p}^*)$, with $\mathbf{p}^* = \left(\frac{p_2}{1-p_1}, \cdots, \frac{p_J}{1-p_1}\right)$.

If we apply the same decomposition to the Multinomial$(n-n_1, \mathbf{p}^*)$, and then to Multinomial$(n-n_1-n_2, \mathbf{p}^{**})$, and so on, we get

$$P(N_1 = n_1, \cdots, N_t = n_J) = P(N_1 = n_1) P(N_2 = n_2 | N_1 = n_1) \cdots P(N_{J-1} = n_{J-1} | N_i = n_i, 1 \le i \le J-2)$$

where N_1 is Bin(n, p_1) and N_k given $N_1, \cdots$, and N_{k-1} is Bin$\left(n - \sum_{i=1}^{k-1} n_i, \frac{p_k}{1-\sum_{i=1}^{k-1} p_i}\right)$, for $k = 2, , J-1$.

b) Say the first sample has distribution $Multi_1(n_1, P_{11}, P_{12}, ..., P_{1J})$ and the second sample has distribution of $Multi_2(n_2, P_{21}, P_{22}, ..., P_{2J})$. Then with the product-multinomial sampling, the likelihood function is $L(p) = Multi_1(n_1, P_{11}, P_{12}, ..., P_{1J}) * Multi_2(n_2, P_{21}, P_{22}, ..., P_{2J})$, which gets maximized at $\hat{p}_{ij} = \frac{n_{ij}}{n}$.

Under the null $H_0 : p_{1j} = p_{2j} = \pi_j$, j=1,...,J, the likelihood ratio function becomes $L(p) = Multi_1^0(n_1, \pi_1, \pi_2, ..., \pi_t) * Multi_2^0(n_2, \pi_1, \pi_2, ..., \pi_J)$, which get maximized at $\hat{p}_{ij}^0 = \hat{\pi}_j = \frac{n_{.j}}{n_{..}}$.

The likelihood ratio test statistics is

$$\lambda = \frac{L(\hat{p}^0)}{L(\hat{p})} = \frac{Multi_1^0 * Multi_2^0}{Multi_1 * Multi_2}$$

then $G^2 = -log\lambda$, each of the multinomials can be factored to J-1 binomials, as we proved in part a. So $G^2 = -\sum_1^{J-1} log(\frac{Bin_{1j}^0 * Bin_{2j}^0}{Bin_{1j} * Bin_{2j}}) = \sum_{J-1} G_1^2$ as for a 2*2 table, $G_1^2 = -log\lambda = -log(\frac{Bin_1^0 * Bin_2^0}{Bin_1 * Bin_2})$.

Alternatively, we can do the calculation directly without invoking part (a), but this involves considerable algebraic manipulation that is equivalent to reproving part (a).

Problem 3. EX. 2.7.2 (15 points):

Both the χ^2 and likelihood ratio test give a test-statistic of about 14.99. Comparing to a χ^2 distribution with $(2-1)(12-1) = 11$ degrees of freedom, this gives a p-value of 0.18. Therefore, we fail to reject the null hypothesis of independence of gender and birth month. Normally we would stop here, but because the problem asked us to use all the tools developed in the first two lectures, we proceed to examine the nature of the residuals under the null model. The Pearson residuals are given in the following table.

	Female	Male
Jan	0.41	−0.40
Feb	0.87	−0.84
Mar	1.01	−0.98
Apr	−1.51	1.46
May	−0.53	0.51
Jun	−0.10	0.10
Jul	−0.64	0.62
Aug	0.48	−0.47
Sep	0.26	−0.26
Oct	1.05	−1.01
Nov	−0.02	0.02
Dec	−1.20	1.15

There appears to be a temporal pattern in the residuals, with more than the expected number of males being born between April and July as well as in November and December. However, compared to the degrees of freedom, the residuals are not very large. Perhaps with a larger sample size we would be able to detect lack of independence, in which case we could test for such a pattern using an odds ratio collapsing over the relevant months.

As an observational study, we would like to think that n_{++} is fixed and it's better to think of this as a single multinomial sampling.

Problem 4 EX. 2.7.5 (15 points):

1.

$$\begin{aligned} Pr(y_1 = r_1, t = t_0) &= Pr(y_1 = r_1, y_2 = t_0 - r_1) \\ &= \binom{N_1}{r_1} p_1^{r_1} (1-p_1)^{N_1 - r_1} \binom{N_2}{t_0 - r_1} p_2^{t_0 - r_1} (1-p_2)^{N_2 - (t_0 - r_1)} \end{aligned}$$

2. Assuming $p_1 = p_2, t = y_1 + y_2 \sim \text{Binomial}(N_1 + N_2, p_1)$, so

$$\begin{aligned} Pr(y_1 = r_1 | t = t_0) &= Pr(y_1 = r_1, y_2 = t_0 - r_1)/Pr(t = t_0) \\ &= \frac{\binom{N_1}{r_1} p_1^{r_1} (1-p_1)^{N_1 - r_1} \binom{N_2}{t_0 - r_1} p_2^{t_0 - r_1} (1-p_2)^{N_2 - (t_0 - r_1)}}{\binom{N_1 + N_2}{t_0} p_1^{t_0} (1-p_1)^{N_1 + N_2 - t_0}} \\ &= \binom{N_1}{r_1} \binom{N_2}{t_0 - r_1} / \binom{N_1 + N_2}{t_0} \end{aligned}$$

   This is hypergeometric distribution. Both columns and rows are fixed. Each row considered as an independent Binomial, the row totals are fixed; as the condition, t, and thus the column totals are fixed.

3. $Pr(y_1 = 3 | t = 10) = \binom{5}{3}\binom{8}{7}/\binom{5+8}{10} = .2797$

   The condition $Pr(y_1 = r_1 | t = 10) \leq Pr(y_1 = 3 | t = 10)$ is satisfied for $r_1 = 2, 3$, and 5, which gives a p-value of 0.51.

Problem 5 (30 points):

a) No matter what the value of K, the expected values for the 4 cells are:

	A	$\bar{A}$	Total
B	225	75	300
$\bar{B}$	75	25	100
	300	100	400

We can now write the Pearson chi-square statistic as

$$\begin{aligned} \chi^2 &= \frac{K^2}{225} + \frac{K^2}{75} + \frac{K^2}{75} + \frac{K^2}{25} \\ &= K^2 (\frac{1}{225} + \frac{2}{75} + \frac{1}{25}) \\ &= K^2 \frac{16}{225} \end{aligned}$$

Since the critical value of χ^2 with 1 degree of freedom at 1% level is 6.635, the test rejects if $\chi^2 \geq 6.635$. Thus, we need to have

$$\begin{aligned} K^2 \frac{16}{225} &\geq 6.635 \\ K^2 &\geq \frac{225}{16} 6.635 \\ &= 93.3 \\ |K| &\geq 9.659. \end{aligned}$$

If we consider only integer values of K, then the result is that $K \geq 10$ or $K \leq -10$. Thus χ^2 is useful for detecting departures from independence in both positive and negative directions.

b) Using G^2 in place of χ^2, we get

$$\begin{aligned} G^2 &= 2 \sum O \log \frac{O}{E} \\ &= 2[(225+K)\log\frac{225+K}{225} + 2(75-K)\log\frac{75-K}{75} + (25+K)\log\frac{25+K}{25}] \\ &\geq 6.635. \end{aligned}$$

since, as before, we reject H_0 at the 1% level if $G^2 \geq 6.65$. Solving this inequality in Maple yield $|K| \geq 9.887$. Alternatively you may just have tried integer values numerically in which case you would have discovered that $|K| \geq 10$.

c) In this instance the integer values in parts (a) and (b) are the same and we appear to have equivalent results. This will not happen in general.