ST102 Outline solutions to Exercise 18 (2013–14)

1. (a) The total number of observations is $n = 4 + 8 + 17 + 37 + 62 + 45 + 17 + 10 = 200$. The expected frequencies are calculated according to the fitted Normal distribution $N(54.9,\ 223.11)$. Hence the expected frequency in the interval $(a, b]$ is

$$n \times P\left(a < N(54.9,\ 223.11) \leq b\right) = 200 \times P\left(\frac{a - 54.9}{\sqrt{223.11}} < N(0,\ 1) \leq \frac{b - 54.9}{\sqrt{223.11}}\right).$$

   From this, we obtain $E_1 = 1.95$ and $E_8 = 9.29$.

   (b) To ensure $E_i \geq 5$ for all i, we combine the first two intervals together.

Interval	$(-\infty, 30]$	$(30, 40]$	$(40, 50]$	$(50, 60]$	$(60, 70]$	$(70, 80]$	$(80, \infty)$	Total
O_i	12	17	37	62	45	17	10	200
E_i	9.55	22.30	42.44	52.43	42.07	21.92	9.29	200
$O_i - E_i$	2.45	−5.30	−5.44	9.57	2.93	−4.92	0.71	0
$(O_i - E_i)^2/E_i$	0.63	1.26	0.70	1.75	0.20	1.10	0.05	5.69

   Under H_0, $T = \sum_i (O_i - E_i)^2/E_i \sim \chi^2_{7-1-2} = \chi^2_4$, as we have estimated 2 parameters in the fitted Normal distribution. Since $t = 5.69 < 7.78 = \chi^2_{0.1,\,4}$, we cannot reject the null hypothesis and hence there is no significant evidence to reject normality of the data.

2. We compute the marginals first, obtain $n = O_{\cdot\cdot} = 2223$, then compute the expected frequencies $E_{ij} = O_{i\cdot}O_{\cdot j}/n$ and the differences $O_{ij} - E_{ij}$ etc. in the tables below.

E_{ij}	Men	Women	Boys	Girls	Total
Survived	537.360	134.022	20.326	14.291	706
Died	1154.640	287.978	43.674	30.709	1517
Total	1692	422	64	45	2223

$O_{ij} - E_{ij}$	Men	Women	Boys	Girls	Total
Survived	−205.360	183.978	8.674	12.709	0
Died	205.360	−183.978	−8.674	−12.709	0
	0	0	0	0	0

$(O_{ij} - E_{ij})^2/E_{ij}$	Men	Women	Boys	Girls	Total
Survived	78.482	252.553	3.702	11.301	346.037
Died	36.525	117.536	1.723	5.259	161.043
					507.080

   Under the null hypothesis of independence, $T \sim \chi^2_{(2-1)(4-1)} = \chi^2_3$. Since $t = 507.084 > 11.345 = \chi^2_{0.01,\,3}$, we reject the independence hypothesis.

   The above statistical analysis reveals that there is highly significant evidence indicating that whether or not someone survived depends on the gender/age category. Looking

at the differences table, one can clearly see that the number of adult male deaths is much larger than the expected number under the null hypothesis of independence. In contrast, the numbers of deaths for women, boys and girls are all smaller than the expected numbers. (Of course, we know why this happened!)

3. Under $\mathrm{H}_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$ for $i, j = 1, 2$, and $E_{ij} = O_{i\cdot}O_{\cdot j}/n$. Note $n = a + b + c + d$. Hence

$$\begin{array}{c|cc} E_{ij} & & \\ \hline & (a+b)(a+c)/n & (a+b)(b+d)/n \\ & (c+d)(a+c)/n & (c+d)(b+d)/n \end{array}$$

$$\begin{array}{c|cc} O_{ij}-E_{ij} & & \\ \hline & a-(a+b)(a+c)/n & b-(a+b)(b+d)/n \\ & c-(c+d)(a+c)/n & d-(c+d)(b+d)/n \end{array}$$

$$= \begin{array}{c|cc} O_{ij}-E_{ij} & & \\ \hline & (ad-bc)/n & (bc-ad)/n \\ & (bc-ad)/n & (ad-bc)/n \end{array}$$

Hence $|O_{ij} - E_{ij}| = |ad - bc|/n$, therefore the goodness-of-test statistic is

$$\begin{aligned} T &= \frac{(ad-bc)^2}{n^2}\left(\frac{n}{(a+b)(a+c)} + \frac{n}{(a+b)(b+d)} + \frac{n}{(c+d)(a+c)} + \frac{n}{(c+d)(b+d)}\right) \\ &= \frac{(ad-bc)^2}{n}\,\frac{(c+d)(b+d)+(a+c)(c+d)+(a+b)(b+d)+(a+b)(a+c)}{(a+b)(a+c)(b+d)(c+d)} \\ &= \frac{(ad-bc)^2}{n}\,\frac{(c+d)(a+b+c+d)+(a+b)(a+b+c+d)}{(a+b)(a+c)(b+d)(c+d)} \\ &= (ad-bc)^2\frac{(c+d)+(a+b)}{(a+b)(a+c)(b+d)(c+d)} \;=\; \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}. \end{aligned}$$

4. (a) From Q3 above, we calculate

$$t = \frac{300(30\times 175 - 70\times 25)^2}{100\times 55\times 245\times 200} = 13.636.$$

Under the null hypothesis of independence, $T \sim \chi^2_1$. Since $t = 13.636 > 6.635 = \chi^2_{0.01,\,1}$, we reject the null hypothesis of independence at the 1% significance level.

(b) This requires us to test two binomial distributions, conditionally on $O_{1\cdot} = 100$ and $O_{2\cdot} = 200$. We assume

$$O_{11} \sim \mathrm{Bin}(100,\, p_1) \qquad \text{and} \qquad O_{21} \sim \mathrm{Bin}(200,\, p_2),$$

and O_{11} and O_{21} are independent. We test the hypothesis $\mathrm{H}_0 : p_1 = p_2$. First we find the MLE under H_0. The likelihood function is

$$\begin{aligned} L(p_1,\, p_2) &= \frac{100!}{O_{11}!(100-O_{11})!}p_1^{O_{11}}(1-p_1)^{100-O_{11}}\cdot\frac{200!}{O_{21}!(200-O_{21})!}p_2^{O_{21}}(1-p_2)^{200-O_{21}} \\ &\propto p_1^{O_{11}}(1-p_1)^{100-O_{11}}p_2^{O_{21}}(1-p_2)^{200-O_{21}}. \end{aligned}$$

The log-likelihood is

$$l(p_1,\, p_2) = O_{11}\log p_1 + (100 - O_{11})\log(1-p_1) + O_{21}\log p_2 + (200 - O_{21})\log(1-p_2).$$

Under H_0,

$$l(p_1) = l(p_1,\, p_1) = (O_{11}+O_{21})\log p_1 + (300 - O_{11} - O_{21})\log(1-p_1) = 55\log p_1 + 245\log(1-p_1).$$

Maximising $l(p_1)$, we obtain $\hat{p}_1 = 55/300 = 0.1833$. Then

$$E_{11} = O_{1\cdot}\hat{p}_1 = 18.33, \qquad E_{12} = O_{1\cdot}(1-\hat{p}_1) = 81.67,$$

$$E_{21} = O_{2\cdot}\hat{p}_1 = 36.67, \qquad E_{22} = O_{2\cdot}(1-\hat{p}_1) = 163.33.$$

The goodness-of-fit statistic is

$$T = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{p-d} \quad \text{under } \mathrm{H}_0.$$

We obtain the test statistic value of

$$t = 11.67^2\left(\frac{1}{18.33} + \frac{1}{81.67} + \frac{1}{36.67} + \frac{1}{163.33}\right) = 13.645.$$

Under H_0, $T \sim \chi^2_{p-d} = \chi^2_1$, where $p = 2$ is the number of free cells, and $d = 1$ is the number of estimated parameters under H_0. Since $t = 13.645 > 6.635 = \chi^2_{0.01,\,1}$, we reject the null hypothesis that the probabilities for rejection from each of the two suppliers are the same at the 1% significance level.