Chapter 14: Analysis of Categorical Data

14.1 **a.** H_0: $p_1 = .41$, $p_2 = .10$, $p_3 = .04$, $p_4 = .45$ vs. H_a: not H_0. The observed and expected counts are:

	A	B	AB	O
observed	89	18	12	81
expected	200(.41) = 82	200(.10) = 20	200(.04) = 8	200(.45) = 90

The chi–square statistic is $X^2 = \frac{(89-82)^2}{82} + \frac{(18-20)^2}{20} + \frac{(12-8)^2}{8} + \frac{(81-90)^2}{90} = 3.696$ with 4 –1 = 3 degrees of freedom. Since $\chi^2_{.05} = 7.81473$, we fail to reject H_0; there is not enough evidence to conclude the proportions differ.

b. Using the Applet, *p*–value = $P(\chi^2 > 3.696) = .29622$.

14.2 **a.** H_0: $p_1 = .60$, $p_2 = .05$, $p_3 = .35$ vs. H_a: not H_0. The observed and expected counts are:

	admitted unconditionally	admitted conditionally	refused
observed	329	43	128
expected	500(.60) = 300	500(.05) = 25	500(.35) = 175

The chi–square test statistic is $X^2 = \frac{(329-300)^2}{300} + \frac{(43-25)^2}{25} + \frac{(128-175)^2}{175} = 28.386$ with 3 – 1 = 2 degrees of freedom. Since $\chi^2_{.05} = 7.37776$, we can reject H_0 and conclude that the current admission rates differ from the previous records.

b. Using the Applet, *p*–value = $P(\chi^2 > 28.386) = .00010$.

14.3 The null hypothesis is H_0: $p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$ vs. H_a: not H_0. The observed and expected counts are:

lane	1	2	3	4
observed	294	276	238	192
expected	250	250	250	250

The chi–square statistic is $X^2 = \frac{(294-250)^2+(276-250)^2+(238-250)^2+(192-250)^2}{250} = 24.48$ with 4 –1 = 3 degrees of freedom. Since $\chi^2_{.05} = 7.81473$, we reject H_0 and conclude that the lanes are not preferred equally. From Table 6, *p*–value < .005.

Note that R can be used by:

```
> lanes <- c(294,276,238,192)
> chisq.test(lanes,p = c(.25,.25,.25,.25))    # p is not necessary here

        Chi-squared test for given probabilities

data:  lanes
X-squared = 24.48, df = 3, p-value = 1.983e-05
```

14.4 The null hypothesis is H_0: $p_1 = p_2 = \ldots = p_7 = \frac{1}{7}$ vs. H_a: not H_0. The observed and expected counts are:

	SU	M	T	W	R	F	SA
observed	24	36	27	26	32	26	29
expected	28.571	28.571	28.571	28.571	28.571	28.571	28.571

The chi–square statistic is $X^2 = \frac{(24-28.571)^2+(36-28.571)^2+\ldots+(29-28.571)^2}{28.571} = 24.48$ with 7 –1 = 6 degrees of freedom. Since $\chi^2_{.05} = 12.5916$, we can reject the null hypothesis and conclude that there is evidence of a difference in percentages of heart attacks for the days of the week

14.5 **a.** Let p = proportion of heart attacks on Mondays. Then, H_0: $p = \frac{1}{7}$ vs. H_a: $p > \frac{1}{7}$. Then, $\hat{p} = 36/200 = .18$ and from Section 8.3, the test statistic is

$$z = \frac{.18-1/7}{\sqrt{\frac{(1/7)(6/7)}{200}}} = 1.50.$$

Since $z_{.05} = 1.645$, we fail to reject H_0.

b. The test was suggested by the data, and this is known as "data snooping" or "data dredging." We should always apply the scientific method: first form a hypothesis and then collect data to test the hypothesis.

c. Monday has often been referred to as the most stressful workday of the week: it is the day that is farthest from the weekend, and this realization gets to some people.

14.6 **a.** $E(n_i - n_j) = E(n_i) - E(n_j) = np_i - np_j$.

b. Define the sample proportions $\hat{p}_i = n_i / n$ and $\hat{p}_j = n_j / n$. Then, $\hat{p}_i - \hat{p}_j$ is unbiased for $p_i - p_j$ from part **a** above.

c. $V(n_i - n_j) = V(n_i) + V(n_j) - 2\text{Cov}(n_i, n_j) = np_i(1-p_i) + np_j(1-p_j) + 2np_ip_j$.

d. $V(\hat{p}_i - \hat{p}_j) = \frac{1}{n^2}V(n_i - n_j) = \frac{1}{n}\left(p_i(1-p_i) + p_j(1-p_j) + 2p_ip_j\right)$.

e. A consistent estimator is one that is unbiased and whose variance tends to 0 as the sample size increases. Thus, $\hat{p}_i - \hat{p}_j$ is a consistent estimator.

f. Given the information in the problem and for large n, the quantity

$$Z_n = \frac{\hat{p}_i - \hat{p}_j - (p_i - p_j)}{\sigma_{\hat{p}_i - \hat{p}_j}}$$

is approx. normally distributed, where $\sigma_{\hat{p}_i - \hat{p}_j} = \sqrt{\frac{1}{n}\left(p_i(1-p_i) + p_j(1-p_j) + 2p_ip_j\right)}$.

Now, since $\hat{p}_i$ and $\hat{p}_j$ are consistent estimators,

$$W_n = \frac{\sigma_{\hat{p}_i - \hat{p}_j}}{\hat{\sigma}_{\hat{p}_i - \hat{p}_j}} = \frac{\sqrt{\frac{1}{n}\left(p_i(1-p_i) + p_j(1-p_j) + 2p_ip_j\right)}}{\sqrt{\frac{1}{n}\left(\hat{p}_i(1-\hat{p}_i) + \hat{p}_j(1-\hat{p}_j) + 2\hat{p}_i\hat{p}_j\right)}}$$

tends to 1 (see Chapter 9). Therefore, the quantity

$$Z_n W_n = \frac{\hat{p}_i - \hat{p}_j - (p_i - p_j)}{\sigma_{\hat{p}_i - \hat{p}_j}} \left(\frac{\sigma_{\hat{p}_i - \hat{p}_j}}{\hat{\sigma}_{\hat{p}_i - \hat{p}_j}} \right) = \frac{\hat{p}_i - \hat{p}_j - (p_i - p_j)}{\sqrt{\frac{1}{n}\left(\hat{p}_i(1-\hat{p}_i) + \hat{p}_j(1-\hat{p}_j) + 2\hat{p}_i\hat{p}_j\right)}}$$

has a limiting standard normal distribution by Slutsky's Theorem. The expression for the confidence interval follows directly from the above.

14.7 From Ex. 14.3, $\hat{p}_1 = .294$ and $\hat{p}_4 = .192$. A 95% (large sample) CI for $p_1 - p_4$ is

$$.294 - .192 \pm 1.96\sqrt{\frac{.294(.706) + .192(.808) + 2(.294)(.192)}{1000}} = .102 \pm .043 \text{ or } (.059, .145).$$

There is evidence that a greater proportion use the "slow" lane since the CI does not contain 0.

14.8 The hypotheses are H_0: ratio is 9:3:3:1 vs. H_a: not H_0. The observed and expected counts are:

category	1 (RY)	2 (WY)	3 (RG)	4 (WG)
observed	56	19	17	8
expected	56.25	18.75	18.75	6.25

The chi–square statistic is $X^2 = \frac{(56-56.25)^2}{56.25} + \frac{(19-18.75)^2}{18.75} + \frac{(17-18.75)^2}{18.75} + \frac{(8-6.25)^2}{6.25} = .658$ with 3 degrees of freedom. Since $\chi^2_{.05} = 7.81473$, we fail to reject H_0: there is not enough evidence to conclude the ratio is not 9:3:3:1.

14.9 **a.** From Ex. 14.8, $\hat{p}_1 = .56$ and $\hat{p}_3 = .17$. A 95% (large sample) CI for $p_1 - p_3$ is

$$.56 - .17 \pm 1.96\sqrt{\frac{.56(.44) + .17(.83) + 2(.56)(.17)}{100}} = .39 \pm .149 \text{ or } (.241, .539).$$

b. There are three intervals to construct: $p_1 - p_2$, $p_1 - p_3$, and $p_1 - p_4$. So that the simultaneous confidence coefficient is at least 95%, each interval should have confidence coefficient 1 – (.05/3) = .98333. Thus, we require the critical value $z_{.00833} = 2.39$. The three intervals are

$$.56 - .19 \pm 2.39\sqrt{\frac{.56(.44) + .19(.81) + 2(.56)(.19)}{100}} = .37 \pm .187$$

$$.56 - .17 \pm 2.39\sqrt{\frac{.56(.44) + .17(.83) + 2(.56)(.17)}{100}} = .39 \pm .182$$

$$.56 - .08 \pm 2.39\sqrt{\frac{.56(.44) + .08(.92) + 2(.56)(.08)}{100}} = .48 \pm .153.$$

14.10 The hypotheses of interest are H_0: $p_1 = .5$, $p_2 = .2$, $p_3 = .2$, $p_4 = .1$ vs. H_a: not H_0. The observed and expected counts are:

defect	1	2	3	4
observed	48	18	21	13
expected	50	20	20	10

It is found that $X^2 = 1.23$ with 3 degrees of freedom. Since $\chi^2_{.05} = 7.81473$, we fail to reject H_0; there is not enough evidence to conclude the proportions differ.

14.11 This is similar to Example 14.2. The hypotheses are H_0: Y is Poisson(λ) vs. H_a: not H_0. Using $\bar{y}$ to estimate λ, calculate $\bar{y} = \frac{1}{400}\Sigma_i y_i f_i = 2.44$. The expected cell counts are estimated as $\hat{E}(n_i) = n\hat{p}_i = 400\frac{(2.44)^{y_i}\exp(-2.44)}{y_i!}$. However, after $Y = 7$, the expected cell count drops below 5. So, the final group will be compiled as $\{Y \geq 7\}$. The observed and (estimated) expected cell counts are below:

# of colonies	n_i	$\hat{p}_i$	$\hat{E}(n_i)$
0	56	.087	34.86
1	104	.2127	85.07
2	80	.2595	103.73
3	62	.2110	84.41
4	42	.1287	51.49
5	27	.0628	25.13
6	9	.0255	10.22
7 or more	20		400 – 394.96 = 5.04

The chi–square statistic is $X^2 = \frac{(56-34.86)^2}{34.86} + \ldots + \frac{(20-5.04)^2}{5.04} = 69.42$ with 8 – 2 = 6 degrees of freedom. Since $\chi^2_{.05} = 12.59$, we can reject H_0 and conclude that the observations do not follow a Poisson distribution.

14.12 This is similar to Ex. 14.11. First, $\bar{y} = \frac{1}{414}\Sigma_i y_i f_i = 0.48309$. The observed and (estimated) expected cell counts are below; here, we collapsed cells into $\{Y \geq 3\}$:

# of accidents	n_i	$\hat{p}_i$	$\hat{E}(n_i)$
0	296	.6169	255.38
1	74	.298	123.38
2	26	.072	29.80
3	18	.0131	5.44

Then, $X^2 = \frac{(296-255.38)^2}{255.38} + \ldots + \frac{(18-5.44)^2}{5.44} = 55.71$ with 4 – 2 = 2 degrees of freedom. Since $\chi^2_{.05} = 5.99$, we can reject the claim that this is a sample from a Poisson distribution.

14.13 The contingency table with observed and expected counts is below.

	All facts known	Some facts withheld	Not sure	Total
Democrat	42	309	31	382
	(53.48)	(284.378)	(44.142)	
Republican	64	246	46	356
	(49.84)	(265.022)	(41.138)	
Other	20	115	27	162
	(22.68)	(120.60)	(18.72)	
Total	126	670	104	900

a. The chi–square statistic is $X^2 = \frac{(42-53.48)^2}{53.48} + \frac{(309-284.378)^2}{284.378} + \ldots + \frac{(27-18.72)^2}{18.72} = 18.711$ with degrees of freedom (3–1)(3–1) = 4. Since $\chi^2_{.05} = 9.48773$, we can reject H_0 and conclude that there is a dependence between part affiliation and opinion about a possible cover up.

b. From Table 6, *p*–value < .005.

c. Using the Applet, *p*–value = $P(\chi^2 > 18.711) = .00090$.

d. The *p*–value is approximate since the distribution of the test statistic is only approximately distributed as chi–square.

14.14 R will be used to answer this problem:

```
> p14.14 <- matrix(c(24,35,5,11,10,8),byrow=T,nrow=2)
> chisq.test(p14.14)

        Pearson's Chi-squared test

data:  p14.14
X-squared = 7.267, df = 2, p-value = 0.02642
```

a. In the above, $X^2 = 7.267$ with a *p*–value = .02642. Thus with $\alpha = .05$, we can conclude that there is evidence of a dependence between attachment patterns and hours spent in child care.

b. See part **a** above.

14.15 **a.** $$X^2 = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{\left[n_{ij} - E(\hat{n}_{ij})\right]^2}{E(\hat{n}_{ij})} = \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{\left[n_{ij} - \frac{r_i c_j}{n}\right]^2}{\frac{r_i c_j}{n}} = n\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2 - \frac{2n_{ij}r_i c_j}{n} + \left(\frac{r_i c_j}{n}\right)^2}{r_i c_j}$$

$$= n\left[\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_i c_j} - 2\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}}{n} + \sum_{j=1}^{c}\sum_{i=1}^{r}\frac{r_i c_j}{n^2}\right]$$

$$= n\left[\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_i c_j} - 2 + \frac{\left(\sum_{i=1}^{r} r_i\right)\left(\sum_{j=1}^{c} c_j\right)}{n^2}\right] = n\left[\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_i c_j} - 2 + \frac{n \cdot n}{n^2}\right]$$

$$= n\left[\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_i c_j} - 1\right].$$

b. When every entry is multiplied by the same constant k, then

$$X^2 = kn\left[\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{(kn_{ij})^2}{kr_i kc_j} - 1\right] = kn\left[\sum_{j=1}^{c}\sum_{i=1}^{r}\frac{n_{ij}^2}{r_i c_j} - 1\right].$$

Thus, X^2 will be increased by a factor of k.

14.16 The contingency table with observed and expected counts is below.

Church attendance	Bush	Democrat	Total
More than …	89	53	142
	(73.636)	(68.364)	
Once / week	87	68	155
	(80.378)	(74.622)	
Once / month	93	85	178
	(92.306)	(85.695)	
Once / year	114	134	248
	(128.604)	(119.400)	
Seldom / never	22	36	58
	(30.077)	(27.923)	
Total	405	376	781

The chi–square statistic is $X^2 = \frac{(89-73.636)^2}{73.636} + \ldots + \frac{(36-27.923)^2}{27.923} = 15.7525$ with (5 – 1)(2 – 1) = 4 degrees of freedom. Since $\chi^2_{.05} = 9.48773$, we can conclude that there is evidence of a dependence between frequency of church attendance and choice of presidential candidate.

b. Let p = proportion of individuals who report attending church at least once a week. To estimate this parameter, we use $\hat{p} = \frac{89+53+87+68}{781} = .3803$. A 95% CI for p is

$$.3803 \pm 1.96\sqrt{\frac{.3803(.6197)}{781}} = .3803 \pm .0340.$$

14.17 R will be used to solve this problem:

Part a:

```
> p14.17a <- matrix(c(4,0,0,15,12,3,2,7,7,2,3,5),byrow=T,nrow=4)
> chisq.test(p14.17a)

        Pearson's Chi-squared test

data:  p14.17a
X-squared = 19.0434, df = 6, p-value = 0.004091

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(p14.17a)
```

Part b:

```
> p14.17b <- matrix(c(19,6,2,19,41,27,3,7,31,0,3,3),byrow=T,nrow=4)
> chisq.test(p14.17b)
```

```
        Pearson's Chi-squared test

data:  p14.17b
X-squared = 60.139, df = 6, p-value = 4.218e-11

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(p14.17b)
```

a. Using the first output, $X^2 = 19.0434$ with a *p*–value of .004091. Thus we can conclude at $\alpha = .01$ that the variables are dependent.
b. Using the second output, $X^2 = 60.139$ with a *p*–value of approximately 0. Thus we can conclude at $\alpha = .01$ that the variables are dependent.
c. Some of the expected cell counts are less than 5, so the chi–square approximation may be invalid (note the warning message in both outputs).

14.18 The contingency table with observed and expected counts is below.

	16–34	35–54	55+	Total
Low violence	8	12	21	41
	(13.16)	(13.67)	(14.17)	
High violence	18	15	7	40
	(12.84)	(13.33)	(13.83)	
Total	26	27	28	81

The chi–square statistic is $X^2 = \frac{(8-13.16)^2}{13.16} + \ldots + \frac{(7-13.83)^2}{13.83} = 11.18$ with 2 degrees of freedom. Since $\chi^2_{.05} = 5.99$, we can conclude that there is evidence that the two classifications are dependent.

14.19 The contingency table with the observed and expected counts is below.

	No	Yes	Total
Negative	166	1	167
	(151.689)	(15.311)	
Positive	260	42	302
	(274.311)	(27.689)	
Total	426	43	469

a. Here, $X^2 = \frac{(166-151.689)^2}{151.689} + \ldots + \frac{(42-26.689)^2}{26.689} = 22.8705$ with 1 degree of freedom. Since $\chi^2_{.05} = 3.84$, H_0 is rejected and we can conclude that the complications are dependent on the outcome of the initial ECG.
b. From Table 6, *p*–value < .005.

14.20 We can rearrange the data into a 2×2 contingency table by just considering the type *A* and *B* defects:

	B	$\overline{B}$	Total
A	48	18	66
	(45.54)	(20.46)	
$\overline{A}$	21	13	34
	(23.46)	(10.54)	
Total	69	31	100

Then, $X^2 = 1.26$ with 1 degree of freedom. Since $\chi^2_{.05} = 3.84$, we fail to reject H_0: there is not enough evidence to prove dependence of the defects.

14.21 Note that all the three examples have $n = 50$. The tests proceed as in previous exercises. For all cases, the critical value is $\chi^2_{.05} = 3.84$

a. 20 (13.44) 4 (10.56) $X^2 = 13.99$, reject H_0: species segregate
8 (14.56) 18 (11.44)

b. 4 (10.56) 20 (13.44) $X^2 = 13.99$, reject H_0: species overly mixed
18 (11.44) 18 (14.56)

c. 20 (18.24) 4 (5.76) $X^2 = 1.36$, fail to reject H_0
18 (19.76) 8 (6.24)

14.22 **a.** The contingency table with the observed and expected counts is:

	Treated	Untreated	Total
Improved	117	74	191
	(95.5)	(95.5)	
Not Improved	83	126	209
	(104.5)	(104.5)	
Total	200	200	400

$X^2 = \frac{(117-95.5)^2}{95.5} + \ldots + \frac{(126-104.5)^2}{104.5} = 18.53$ with 1 degree of freedom. Since $\chi^2_{.05} = 3.84$, we reject H_0; there is evidence that the serum is effective.

b. Let p_1 = probability that a treated patient improves and let p_2 = probability that an untreated patient improves. The hypotheses are H_0: $p_1 - p_2 = 0$ vs. H_a: $p_1 - p_2 \neq 0$. Using the procedure from Section 10.3 (derived in Ex. 10.27), we have $\hat{p}_1 = 117/200 = .585$, $\hat{p}_2 = 74/200 = .37$, and the "pooled" estimator $\hat{p} = \frac{117+74}{400} = .4775$, the test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{.585 - .37}{\sqrt{.4775(.5225)\left(\frac{2}{200}\right)}} = 4.3.$$

Since the rejection region is $|z| > 1.96$, we soundly reject H_0. Note that $z^2 = X^2$.

c. From Table 6, p–value < .005.

14.23 To test H_0: $p_1 - p_2 = 0$ vs. H_a: $p_1 - p_2 \neq 0$, the test statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

from Section 10.3. This is equivalent to

$$Z^2 = \frac{(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \frac{n_1 n_2 (\hat{p}_1 - \hat{p}_2)^2}{(n_1 + n_2)\hat{p}\hat{q}}.$$

However, note that

$$\hat{p} = \frac{Y_1 + Y_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

Now, consider the X^2 test from Ex. 14.22. The hypotheses were H_0: independence of classification vs. H_a: dependence of classification. If H_0 is true, then $p_1 = p_2$ (serum has no affect). Denote the contingency table as

	Treated	Untreated	Total
Improved	$n_{11} = n_1\hat{p}_1$	$n_{12} = n_2\hat{p}_2$	$n_{11} + n_{12}$
Not Improved	$n_{21} = n_1\hat{q}_1$	$n_{22} = n_2\hat{q}_2$	$n_{21} + n_{22}$
Total	$n_{11} + n_{21} = n_1$	$n_{12} + n_{22} = n_2$	$n_1 + n_2 = n$

The expected counts are found as follows. $\hat{E}(n_{11}) = \frac{(n_{11}+n_{12})(n_{11}+n_{21})}{n_1+n_2} = \frac{(y_1+y_2)(n_{11}+n_{21})}{n_1+n_2} = n_1\hat{p}$.

So similarly, $\hat{E}(n_{21}) = n_1\hat{q}$, $\hat{E}(n_{12}) = n_2\hat{p}$, and $\hat{E}(n_{22}) = n_2\hat{q}$. Then, the X^2 statistic can be expressed as

$$\begin{aligned} X^2 &= \frac{n_1^2(\hat{p}_1 - \hat{p})^2}{n_1\hat{p}} + \frac{n_1^2(\hat{q}_1 - \hat{q})^2}{n_1\hat{q}} + \frac{n_2^2(\hat{p}_2 - \hat{p})^2}{n_2\hat{p}} + \frac{n_2^2(\hat{q}_2 - \hat{q})^2}{n_2\hat{q}} \\ &= \frac{n_1(\hat{p}_1 - \hat{p})^2}{\hat{p}} + \frac{n_1[(1-\hat{p}_1) - (1-\hat{p})]^2}{\hat{q}} + \frac{n_2(\hat{p}_2 - \hat{p})^2}{\hat{p}} + \frac{n_2[(1-\hat{p}_2) - (1-\hat{p})]^2}{\hat{q}} \end{aligned}$$

However, by combining terms, this is equal to $X^2 = \frac{n_1(\hat{p}_1 - \hat{p})^2}{\hat{p}\hat{q}} + \frac{n_2(\hat{p}_2 - \hat{p})^2}{\hat{p}\hat{q}}$. By substituting the expression for $\hat{p}$ above in the numerator, this simplifies to

$$\begin{aligned} X^2 &= \frac{n_1}{\hat{p}\hat{q}}\left(\frac{n_1\hat{p}_1 + n_2\hat{p}_1 - n_1\hat{p}_1 - n_2\hat{p}_2}{n_1 + n_2}\right)^2 + \frac{n_2}{\hat{p}\hat{q}}\left(\frac{n_1\hat{p}_2 + n_2\hat{p}_2 - n_1\hat{p}_1 - n_2\hat{p}_2}{n_1 + n_2}\right)^2 \\ &= \frac{n_1 n_2(\hat{p}_1 - \hat{p}_2)^2}{\hat{p}\hat{q}(n_1 + n_2)} = Z^2 \text{ from above. Thus, the tests are equivalent.} \end{aligned}$$

14.24 **a.** R output follows.

```
> p14.24 <- matrix(c(40,56,68,84,160,144,132,116),byrow=T,nrow=2)
> chisq.test(p14.24)

        Pearson's Chi-squared test

data:  p14.24
X-squared = 24.3104, df = 3, p-value = 2.152e-05     <-- reject H0
```

b. Denote the samples as 1, 2, 3, and 4. Then, the sample proportions that provide parental support for the four groups are $\hat{p}_1 = 40/200 = .20$, $\hat{p}_2 = 56/200 = .28$, $\hat{p}_3 = 68/200 = .34$, $\hat{p}_4 = 84/800 = .42$.

i. A 95% CI for $p_1 - p_4$ is $.20 - .42 \pm 1.96\sqrt{\frac{.20(.80)}{200} + \frac{.42(.58)}{200}} = -.22 \pm .088$.

ii. With 6 confidence intervals, each interval should have confidence coefficient $1 - (.05/6) = .991667$. Thus, we require the critical value $z_{.004167} = 2.638$. The six intervals are:

$p_1 - p_2$:	$-.08 \pm .112$	
$p_1 - p_3$:	$-.14 \pm .116$	(*)
$p_1 - p_4$:	$-.22 \pm .119$	(*)
$p_2 - p_3$:	$-.06 \pm .122$	
$p_2 - p_4$:	$-.14 \pm .124$	(*)
$p_3 - p_4$:	$-.08 \pm .128$	

iii. By considering the intervals that do not contain 0, these are noted by (*).

14.25 **a.** Three populations (income categories) are under investigation. In each population, members are classified as one out of the four education levels, thus creating the multinomial.

b. $X^2 = 19.1723$ with 6 degrees of freedom, and p-value = 0.003882 so reject H_0.

c. The sample proportions are:

- at least an undergraduate degree and marginally rich: 55/100 = .55
- at least an undergraduate degree and super rich: 66/100.66

The 95% CI is

$$.55 - .66 \pm 1.96\sqrt{\frac{.55(.45)}{100} + \frac{.66(.34)}{100}} = -.11 \pm .135.$$

14.26 **a.** Constructing the data using a contingency table, we have

Machine Number	Defectives	Nondefectives
1	16	384
2	24	376
3	9	391

In the chi–square test, $X^2 = 7.19$ with 2 degrees of freedom. Since $\chi^2_{.05} = 5.99$, we can reject the claim that the machines produce the same proportion of defectives.

b. The hypothesis of interest is H_0: $p_1 = p_2 = p_3 = p$ against an alternative that at least one equality is not correct. The likelihood function is

$$L(\boldsymbol{p}) = \prod_{i=1}^{3} \binom{400}{n_i} p_i^{n_i} (1 - p_i)^{400 - n_i}.$$

In Ω, the MLE of p_i is $\hat{p}_i = n_i / 400$, $i = 1, 2, 3$. In Ω_0, the MLE of p is $\hat{p} = \Sigma n_i / 1200$. Then,

$$\lambda = \frac{\left(\frac{\Sigma n_i}{1200}\right)^{\Sigma n_i}\left(1-\frac{\Sigma n_i}{1200}\right)^{1200-\Sigma y_i}}{\prod_{i=1}^{3}\left(\frac{n_i}{400}\right)^{y_i}\left(1-\frac{n_i}{400}\right)^{400-n_i}}.$$

Using the large sample properties, $-2\ln\lambda = -2(-3.689) = 7.378$ with 2 degrees of freedom. Again, since $\chi^2_{.05} = 5.99$, we can reject the claim that the machines produce the same proportion of defectives.

14.27 This exercise is similar to the others. Here, $X^2 = 38.429$ with 6 degrees of freedom. Since $\chi^2_{.05} = 12.59$, we can conclude that age and probability of finding nodules are dependent.

14.28 **a.** The chi–square statistic is $X^2 = 10.2716$ with 1 degree of freedom. Since $\chi^2_{.05} = 3.84$, we can conclude that the proportions in the two plants are different.

b. The 95% lower bound is

$$.73 - .51 - 1.645\sqrt{\tfrac{.73(.27)}{100} + \tfrac{.51(.49)}{100}} = .22 - .11 = .11.$$

Since the lower bound is greater than 0, this gives evidence that the proportion at the plant with active worker participation is greater.

c. No. The chi–square test in (a) only detects a difference in proportions (equivalent to a two–tailed alternative).

14.29 The contingency table with observed and expected counts is below.

	City A	City B	Nonurban 1	Nonurban 2	Total
w/ lung disease	34	42	21	18	115
	(28.75)	(28.75)	(28.75)	(28.75)	
w/o lung disease	366	358	379	382	1485
	(371.25)	(371.25)	(371.25)	(371.25)	
Total	400	400	400	400	1600

a. Using the above, it is found that $X^2 = 14.19$ with 3 degrees of freedom and since $\chi^2_{.05} = 7.81$, we can conclude that there is a difference in the proportions of lung disease for the four locations.

b. It is known that cigarette smoking contributes to lung disease. If more smokers live in urban areas (which is possibly true), this could confound our results. Thus, smokers should probably be excluded from the study.

14.30 The CI is $.085 - .105 \pm 1.96\sqrt{\frac{.085(.915)}{400} + \frac{.105(.895)}{400}} = -.02 \pm .041$.

14.31 The contingency table with observed and expected counts is below.

	RI	CO	CA	FL	Total
Participate	46	63	108	121	338
	(63.62)	(78.63)	(97.88)	(97.88)	
Don't participate	149	178	192	179	698
	(131.38)	(162.37)	(202.12)	(202.12)	
Total	195	241	300	300	1036

Here, $X^2 = 21.51$ with 3 degrees of freedom. Since $\chi^2_{.01} = 11.3449$, we can conclude that there is a difference in participation rates for the states.

14.32 See Section 5.9 of the text.

14.33 This is similar to the previous exercises. Here, $X^2 = 6.18$ with 2 degrees of freedom. From Table 6, we find that $.025 < p$–value $< .05$, so there is sufficient evidence that the attitudes are not independent of status.

14.34 R will be used here.

```
> p14.34a <- matrix(c(43,48,9,44,53,3),byrow=T,nrow=2)
> chisq.test(p14.34a)

        Pearson's Chi-squared test

data:  p14.34a
X-squared = 3.259, df = 2, p-value = 0.1960
>

> p14.34b <- matrix(c(4,42,41,13,3,48,35,14),byrow=T,nrow=2)
> chisq.test(p14.34b)

        Pearson's Chi-squared test

data:  p14.34b
X-squared = 1.0536, df = 3, p-value = 0.7883

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(p14.34b)
```

a. For those drivers who rate themselves, the *p*–value for the test is .1960, so there is not enough evidence to conclude a dependence on gender and driver ratings.

b. For those drivers who rate others, the *p*–value for the test is .7883, so there is not enough evidence to conclude a dependence on gender and driver ratings.

c. Note in part **b**, the software is warning that two cells have expected counts that are less than 5, so the chi–square approximation may not be valid.

14.35 R:

```
> p14.35 <- matrix(c(49,43,34,31,57,62),byrow=T,nrow=2)
> p14.35
     [,1] [,2] [,3]
[1,]   49   43   34
[2,]   31   57   62
> chisq.test(p14.35)

        Pearson's Chi-squared test

data:  p14.35
X-squared = 12.1818, df = 2, p-value = 0.002263
```

In the above, the test statistic is significant at the .05 significance level, so we can conclude that the susceptibility to colds is affected by the number of relationships that people have.

14.36 R:

```
> p14.36 <- matrix(c(13,14,7,4,12,9,14,3),byrow=T,nrow=2)
> chisq.test(p14.36)

        Pearson's Chi-squared test

data:  p14.36
X-squared = 3.6031, df = 3, p-value = 0.3076

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(p14.36)
```

a. From the above, we fail to reject the hypothesis that position played and knee injury type are independent.
b. From the above, *p*–value = .3076.
c. From the above, *p*–value = .3076.

14.37 The hypotheses are H_0: Y is binomial(4, p) vs. H_a: Y isn't binomial(4, p). The probability mass function is

$$p(y) = P(Y = y) = \binom{4}{y} p^y (1-p)^{4-y}, \; y = 0, 1, 2, 3, 4.$$

Similar to Example 14.2, we can estimate p by using the MLE (see Chapter 10; think of this as an experiment with 400 trials):

$$\hat{p} = \frac{\text{number of successes}}{\text{number of trials}} = \frac{0(11)+1(17)+2(42)+3(21)+4(9)}{400} = .5$$

So, the expected counts are $\hat{E}(n_i) = 100\,\hat{p}(i) = \binom{4}{i}(.5)^i(.5)^{4-i} = \binom{4}{i}(.5)^4$, $i = 0, \ldots, 4$. The observed and expected cell counts are below.

	0	1	2	3	4
n_i	11	17	42	21	9
$\hat{E}(n_i)$	6.25	25	37.5	21	6.25

Thus, $X^2 = 8.56$ with 5 – 1 – 1 = 3 degrees of freedom and the critical value is $\chi^2_{.05} = 7.81$. Thus, we can reject H_0 and conclude the data does not follow as binomial.

14.38 **a.** The likelihood function is

$$L(\theta) = (-1)^n \left[\ln(1-\theta)\right]^{-n} \frac{\theta^{\Sigma y_i}}{\Pi y_i}.$$

So, $\ln L(\theta) = k - n\ln[\ln(1-\theta)] + (\ln\theta)\sum_{i=1}^{n} y_i$ where k is a quantity that does not depend on θ. By taking a derivative and setting this expression equal to 0, this yields

$$\left(\frac{1}{1-\theta}\right)\frac{n}{\ln(1-\theta)} + \frac{1}{\theta}\sum_{i=1}^{n} y_i = 0,$$

or equivalently

$$\bar{Y} = \frac{\hat{\theta}}{-(1-\hat{\theta})\ln(1-\hat{\theta})}.$$

b. The hypotheses are H_0: data follow as logarithmic series vs. H_a: not H_0. From the table, $\bar{y} = \frac{1(359)+2(146)+3(57)+\ldots+7(29)}{675} = 2.105$. Thus, to estimate θ, we must solve the nonlinear equation $2.105 = \frac{\hat{\theta}}{-(1-\hat{\theta})\ln(1-\hat{\theta})}$, or equivalently we must find the root of

$$2.105(1-\hat{\theta})\ln(1-\hat{\theta}) + \hat{\theta} = 0.$$

By getting some help from R,

```
> uniroot(function(x) x + 2.101*(1-x)*log(1-x),c(.0001,.9999))
$root
[1] 0.7375882
```

Thus, we will use $\hat{\theta} = .7376$. The probabilities are estimated as

$$\hat{p}(1) = -\frac{.7376}{\ln(1-.7376)} = .5513,\ \hat{p}(2) = -\frac{(.7376)^2}{2\ln(1-.7376)} = .2033,\ \hat{p}(3) = .1000,$$
$$\hat{p}(4) = .0553,\ \hat{p}(5) = .0326,\ \hat{p}(6) = .0201,\ \hat{p}(7, 8, \ldots) = .0374 \text{ (by subtraction)}$$

The expected counts are obtained by multiplying these estimated probabilities by the total sample size of 675. The expected counts are

	1	2	3	4	5	6	7+
$\hat{E}(n_i)$	372.1275	137.2275	67.5000	37.3275	22.005	13.5675	25.245

Here, $X^2 = 5.1708$ with 7 – 1 – 1 = 5 degrees of freedom. Since $\chi^2_{.05} = 11.07$, we fail to reject H_0.

14.39 Consider row i as a single cell with r_i observations falling in the cell. Then, $r_1, r_2, \ldots r_r$ follow a multinomial distribution so that the likelihood function is

$$L(\boldsymbol{p}) = \binom{n}{r_1\ r_2 \cdots r_r} p_1^{r_1} p_2^{r_2} \cdots p_r^{r_r}.$$

so that

$$\ln L(\boldsymbol{p}) = k + \sum_{j=1}^{r} r_j \ln p_j,$$

where k does not involve any parameters and this is subject to $\sum_{j=1}^{r} p_j = 1$. Because of this restriction, we can substitute $p_r = 1 - \sum_{j=1}^{r-1} p_j$ and $r_r = n - \sum_{j=1}^{r-1} r_j$. Thus,

$$\ln L(\boldsymbol{p}) = k + \sum_{j=1}^{r-1} r_j \ln p_j + \left(n - \sum_{j=1}^{r-1} r_j\right) \ln\left(1 - \sum_{j=1}^{r-1} p_j\right).$$

Thus, the $n - 1$ equations to solve are

$$\frac{\partial \ln L}{\partial p_i} = \frac{r_i}{p_i} - \frac{n - \sum_{j=1}^{r-1} r_j}{\left(1 - \sum_{j=1}^{r-1} p_j\right)} = 0,$$

or equivalently

$$r_i\left(1 - \sum_{j=1}^{r-1} p_j\right) = p_i\left(n - \sum_{j=1}^{r-1} r_j\right),\ i = 1, 2, \ldots, r-1. \qquad (*)$$

In order to solve these simultaneously, add them together to obtain

$$\sum_{i=1}^{r-1} r_i\left(1 - \sum_{j=1}^{r-1} p_j\right) = \sum_{i=1}^{r-1} p_i\left(n - \sum_{j=1}^{r-1} r_j\right)$$

Thus, $\sum_{i=1}^{r-1} r_i = n\sum_{i=1}^{r-1} p_i$ and so $\sum_{i=1}^{r-1} \hat{p}_i = \frac{1}{n}\sum_{i=1}^{r-1} r_i$. Substituting this into (*) above yields the desired result.

14.40 **a.** The model specifies a trinomial distribution with $p_1 = p^2$, $p_2 = 2p(1-p)$, $p_3 = (1-p)^2$. Hence, the likelihood function is

$$L(p) = \frac{n!}{n_1!n_2!n_3!} p^{2n_1}[2p(1-p)]^{n_2}(1-p)^{2n_3}.$$

The student should verify that the MLE for p is $\hat{p} = \frac{2n_1 + n_2}{2n}$. Using the given data, $\hat{p} = .5$ and the (estimated) expected cell counts are $\hat{E}(n_1) = 100(.5)^2 = 25$, $\hat{E}(n_2) = 50$, and $\hat{E}(n_3) = 25$. Using these, we find that $X^2 = 4$ with $3 - 1 - 1 = 1$ degree of freedom. Thus, since $\chi^2_{.05} = 3.84$ we reject H_0: there is evidence that the model is incorrect.

b. If the model specifies $p = .5$, it is not necessary to find the MLE as above. Thus, X^2 will have $3 - 1 = 2$ degrees of freedom. The computed test statistic has the same value as in part **a**, but since $\chi^2_{.05} = 5.99$, H_0 is not rejected in this case.

14.41 The problem describes a multinomial experiment with $k = 4$ cells. Under H_0, the four cell probabilities are $p_1 = p/2$, $p_2 = p^2/2 + pq$, $p_3 = q/2$, and $p_4 = q^2/2$, but $p = 1 - q$. To obtain an estimate of p, the likelihood function is

$$L = C(p/2)^{n_1}(p^2/2 + pq)^{n_2}(q/2)^{n_3}(q^2/2)^{n_4},$$

where C is the multinomial coefficient. By substituting $q = 1 - p$, this simplifies to

$$L = Cp^{n_1+n_2}(2-p)^{n_2}(1-p)^{n_3+2n_4}.$$

By taking logarithms, a first derivative, and setting the expression equal to 0, we obtain

$$(n_1 + 2n_2 + n_3 + 2n_4)p^2 - (3n_1 + 4n_2 + 2n_3 + 4n_4)p + 2(n_1 + n_2) = 0$$

(after some algebra). So, the MLE for p is the root of this quadratic equation. Using the supplied data and the quadratic formula, the valid solution is $\hat{p} = \frac{6960 - \sqrt{1{,}941{,}760}}{6080} = .9155$.

Now, the estimated cell probabilities and estimated expected cell counts can be found by:

$\hat{p}_i$	$\hat{E}(n_i)$	n_i
$\hat{p}/2 = .45775$	915.50	880
$\hat{p}^2/2 + \hat{p}\hat{q} = .49643$	992.86	1032
$\hat{q}/2 = .04225$	84.50	80
$\hat{q}^2/2 = .00357$	7.14	8

Then, $X^2 = 3.26$ with $4 - 1 - 1 = 2$ degrees of freedom. Since $\chi^2_{.05} = 5.99$, the hypothesized model cannot be rejected.

14.42 Recall that from the description of the problem, it is required that $\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} p_i^* = 1$. The likelihood function is given by (multiplication of two multinomial mass functions)

$$L = C\prod_{j=1}^{k} p_j^{n_j}\left(p_j^*\right)^{m_j},$$

where C are the multinomial coefficients. Now under H_0, this simplifies to

$$L_0 = C\prod_{j=1}^{k} p_j^{n_j+m_j}.$$

This is a special case of Ex. 14.39, so the MLEs are $\hat{p}_i = \frac{n_i+m_i}{n+m}$ and the estimated expected counts are $\hat{E}(n_i) = n\hat{p}_i = n\left(\frac{n_i+m_i}{n+m}\right)$ and $\hat{E}(m_i) = m\hat{p}_i = m\left(\frac{n_i+m_i}{n+m}\right)$ for $i = 1, \ldots, k$. The chi–square test statistic is given by

$$X^2 = \sum_{j=1}^{k} \frac{\left[n_i - n\left(\frac{n_i+m_i}{n+m}\right)\right]^2}{n\left(\frac{n_i+m_i}{n+m}\right)} + \sum_{j=1}^{k} \frac{\left[m_i - m\left(\frac{n_i+m_i}{n+m}\right)\right]^2}{m\left(\frac{n_i+m_i}{n+m}\right)}$$

which has a chi–square distribution with $2k - 2 - (k - 1) = k - 1$ degrees of freedom. Two degrees of freedom are lost due the two conditions first mentioned in the solution of this problem, and $k - 1$ degrees of freedom are lost in the estimation of cell probabilities. Hence, a rejection region will be based on $k - 1$ degrees of freedom in the chi–square distribution.

14.43 In this exercise there are 4 binomial experiments, one at each of the four dosage levels. So, with $i = 1, 2, 3, 4$, and p_i represents the binomial (success probability) parameter for dosage i, we have that $p_i = 1 + \beta i$. Thus, in order to estimate β, we form the likelihood function (product of four binomial mass functions):

$$L(\beta) = \prod_{i=1}^{4} \binom{1000}{n_i} (1+i\beta)^{n_i} (-i\beta)^{1000-n_i} = K \prod_{i=1}^{4} \binom{1000}{n_i} (1+i\beta)^{n_i} \beta^{1000-n_i},$$

where K is a constant that does not involve β. Then,

$$\frac{dL(\beta)}{d\beta} = \sum_{i=1}^{4} \frac{in_i}{1+i\beta} + \frac{1}{\beta} \sum_{i=1}^{4} (1000 - n_i).$$

By equating this to 0, we obtain a nonlinear function of β that must be solved numerically (to find the root). Below is the R code that does the job; note that in the association of β with probability and the dose levels, β must be contained in (–.25, 0):

```
> mle <- function(x)
+ {
+ ni <- c(820,650,310,50)
+ i <- 1:4
+ temp <- sum(1000-ni)
+ return(sum(i*ni/(1+i*x))+temp/x)
+ }
>
> uniroot(mle, c(-.2499,-.0001))  <- guessed range for the parameter
$root
[1] -0.2320990
```

Thus, we take $\hat{\beta} = -.232$ and so:

$$\hat{p}_1 = 1 - .232 = .768$$
$$\hat{p}_2 = 1 + 2(-.232) = .536,$$
$$\hat{p}_3 = 1 + 3(-.232) = .304$$
$$\hat{p}_4 = 1 + 4(-.232) = .072.$$

The observed and (estimated) expected cell counts are

Dosage	1	2	3	4
Survived	820	650	320	50
	(768)	(536)	(304)	(72)
Died	180	350	690	950
	(232)	(464)	(696)	(928)

The chi–square test statistic is $X^2 = 74.8$ with 8 – 4 – 1 = 3 degrees of freedom (see note below). Since $\chi^2_{.05} = 7.81$, we can soundly reject the claim that $p = 1 + \beta D$.

Note: there are 8 cells, but 5 restrictions:
- $p_i + q_i = 1$ for i = 1, 2, 3, 4
- estimation of β.