Chapter 16
Inference with Categorical Data

16.1 Use integration by parts $\int udv = uv - \int vdu$ with $u = x^{\frac{a}{2}}$ and $dv = e^{-\frac{x}{2}}dx$ to establish $\int_0^\infty x^{\frac{a}{2}} e^{-\frac{x}{2}} dx = a\int_0^\infty x^{\frac{a}{2}-1} e^{-\frac{x}{2}} dx$ for $a \geq 1$. Now suppose T has a chi-square distribution with ν degrees of freedom. Then use this to show

$$E(T) = \int x f_{\chi^2}(x)dx = \frac{\int_0^\infty x^{\frac{\nu+2}{2}} e^{-\frac{x}{2}} dx}{\int_0^\infty x^{\frac{\nu}{2}} e^{-\frac{x}{2}} dx} = \frac{\nu\int_0^\infty x^{\frac{\nu}{2}} e^{-\frac{x}{2}} dx}{\int_0^\infty x^{\frac{\nu}{2}} e^{-\frac{x}{2}} dx} = \nu$$

16.3 Refer to Section 10.1. To show the marginal density function of K_1 is binomial, factor the joint density of $K_1,\ldots K_c$, into two parts, and then sum over all possible values of $k_2,\ldots k_c$.

$$f_{K_1}(k_1) = \sum_{k_2,\ldots k_c} f(k_1, k_2, \ldots k_c) = A \cdot B,$$

where the first factor A is $A = \frac{n!}{k_1!(n-k_1)!}\pi_1^{k_1}(1-\pi_1)^{n-k_1}$

and the second factor B is $B = \sum_{k_2,\ldots k_c} \frac{(n-k_1)!}{k_2!\ldots k_c!} \cdot \left(\frac{\pi_2}{1-\pi_1}\right)^{k_2} \cdots \left(\frac{\pi_c}{1-\pi_1}\right)^{k_c}$

Notice that the sum B is that of a multinomial probability mass function of c – 1 categories and sample size n – k_1. Since the sum is over all possible values, it must sum to one.

16.5 Many mutually exclusive different values of K_{c-1}, and K_c, lead to $K = K_{c-1}, + K_c$,. Sum the probability mass function for $K_1,\ldots K_c$ over all these values to find the probability mass function of K_j, j = 1,…c – 2 and K.

$$f(k_1, k_2, \ldots k_{c-2}, k) = \sum_{k_{c-1}+k_c=k} \frac{n!}{k_1!\ldots k_{c-1}!k_c!}\pi_1^{k_1}\pi_2^{k_2}\cdots\pi_{c-1}^{k_{c-1}}\pi_c^{k_c} =$$

$$\frac{n!}{k_1!\ldots k_{c-2}!k!}\pi_1^{k_1}\pi_2^{k_2}\cdots\pi_{c-2}^{k_{c-2}} \sum_{k_{c-1}+k_c=k} \frac{k!}{k_{c-1}!k_c!}\pi_{c-1}^{k_{c-1}}\pi_c^{k_c} = \frac{n!}{k_1!\ldots k_{c-2}!k!}\pi_1^{k_1}\pi_2^{k_2}\cdots\pi_{c-2}^{k_{c-2}}\pi^k$$

Use the binomial theorem to see that the sum on the second line above equals $(\pi_{c-1} + \pi_c)^k$.

16.7 This exercise is computationally intensive. One approach is to use induction on n. Here we use the notation χ_n^2 to denote the chi-square statistic based on a sample size of n. For n = 1,

$\chi_1^2 = \sum_{j=1}^{c} \frac{(X_j - \pi_j)^2}{\pi_j}$ where exactly one of the X_j is 1 and the other c – 1 are all 0.

Also $\Pr(X_j = 1) = \pi_j$. This is the special case of Formula 16.1 with n = 1. Note the X_j are Bernoulli random variables and, if $X_j = 1$, then $\chi_1^2 = \frac{(1-\pi_j)^2}{\pi_j} + (1-\pi_j)$. So

$$Var(\chi_1^2) = E(\chi_1^2 - (c-1)) = \sum_{j=1}^{c} \left(\frac{(1-\pi_j)^2}{\pi_j} + (1-\pi_j) - (c-1) \right)^2 \pi_j$$

$$= \sum_{j=1}^{c} \left(\frac{1}{\pi_j} - c) \right)^2 \pi_j = \sum_{j=1}^{c} \frac{1}{\pi_j} - c^2$$

Now assume that the formula $Var(\chi_n^2) = 2(c-1) \frac{\sum_{j=1}^{c} \frac{1}{\pi_j} - (c+1)^2 + 3}{n}$ is correct for some n and demonstrate it must also be correct for n + 1. Let K_j, j = 1,..,c be the counts for the first n observations and X_j the counts for the new, $(n+1)^{st}$ observation. So the K_j ,j = 1,..,c follow a multinomial distribution with parameters n and $\pi_1,\ldots,\pi_c$, the X_j follow a multinomial distribution with parameters 1 and $\pi_1,\ldots,\pi_c$, and the K_j are independent of the X_j. Now the counts for all n+1 observations are $K_j + X_j$, j = 1,..,c, and the new chi-square statistic is

$$\chi_{n+1}^2 = \sum_{j=1}^{c} \frac{(K_j + X_j - (n+1)\pi_j)^2}{(n+1)\pi_j}$$

$$= \frac{n}{n+1}\chi_n^2 + \frac{2n}{n+1} \sum_{j=1}^{c} \frac{(K_j - n\pi_j)(X_j - \pi_j)}{n\pi_j} + \frac{1}{n+1} \sum_{j=1}^{c} \frac{(X_j - \pi_j)^2}{\pi_j}$$

Then use the formulas in Section 10.4 and compute the variance of χ_{n+1}^2 as the sum of the variance and covariances of the terms on the right hand side. We know that the variance of χ_n^2 is $Var(\chi_n^2) = 2(c-1) + \frac{\sum_{j=1}^{c} \frac{1}{\pi_j} - (c+1)^2 + 3}{n}$ and that the variance of $\chi_1^2 = \sum_{j=1}^{c} \frac{(X_j - \pi_j)^2}{\pi_j}$ is $Var(\chi_1^2) = \sum_{j=1}^{c} \frac{1}{\pi_j} - c^2$. Now show that the variance of $\sum_{j=1}^{c} \frac{(K_j - n\pi_j)(X_j - \pi_j)}{n\pi_j}$ is $\frac{(c-1)}{n}$ and all three covariance terms are 0.

Substituting these results into the equation

$$Var(\chi^2_{n+1}) = \left(\frac{n}{n+1}\right)^2 Var(\chi^2_n) + \left(\frac{2n}{n+1}\right)^2 Var\left(\sum_{j=1}^{c} \frac{(K_j - n\pi_j)(X_j - \pi_j)}{n\pi_j}\right) + \left(\frac{1}{n+1}\right)^2 Var\left(\sum_{j=1}^{c} \frac{(X_j - \pi_j)^2}{\pi_j}\right)$$

completes the proof.

16.9 The null hypothesis is H_0: π_1=9/16, π_2=3/16, π_3=3/16, π_4=1/16. The alternative hypothesis is that the π_j are not all equal to these values. There are 80 observations. Under the null hypothesis $80\pi_j$ are expected in the j^{th} class. The chi-square statistic is

$$\chi^2 = \frac{(47-45)^2}{45} + \frac{(12-15)^2}{15} + \frac{(18-15)^2}{15} + \frac{(3-5)^2}{5} = \frac{94}{45} = 2.0\overline{8}.$$

The *P*-value is $\Pr(X_3^2 > 2.0\overline{8}) \approx 0.55$. The *P*-value is large, so the null hypothesis would be retained at the usual significance levels. The evidence does not contradict the claim that the data are a random sample from a population with a 9:3:3:1 ratio.

16.11 The null hypothesis is H_0: π_1=0.08, π_2=0.92. The alternative hypothesis is that the π_j are not all equal to these values. There are 100 observations, $k_1 = 15$ and $k_2 = 85$. Under the null hypothesis $100\pi_j$ are expected in the j^{th} class. The chi-square statistic is

$\chi^2 = \frac{(15-8)^2}{8} + \frac{(85-92)^2}{92} = 6.6576$. The *P*-value is

$\Pr(X_1^2 > 6.6576) \approx 0.0099 < 0.05 = \alpha$. The *P*-value is small, so the null hypothesis would be rejected at the 5% significance level. The evidence contradicts the failure rate is 8%. The failure rate is out of control.

16.13 Table for computing the chi-square statistic. $\chi^2 = \sum_j \frac{(O_j - E_j)^2}{E_j} = 3.4$

Face J	Observed count O_j	Expected Count E_j	$(O_j - E_j)$	χ^2 terms $(O_j - E_j)^2/E_j$
1	1	5/6	1/6	1/30
2	1	5/6	1/6	1/30
3	0	5/6	–5/6	25/30
4	1	5/6	1/6	1/30
5	2	5/6	7/6	49/30
6	0	5/6	–5/6	25/30
Totals	5	5	0	3.4

.Split	χ^2	Probability
5–0	25.0	$6/6^5$
4–1	15.4	$150/6^5$
3–2	10.6	$300/6^5$
3–1–1	8.2	$1200/6^5$
2–2–1	5.8	$1800/6^5$
2–1–1–1	3.4	$3600/6^5$
Total		P-value = $7056/6^5$ = 0.907

The P-value is large. There is insufficient evidence to support the claim that the die is not fair.

16.17 Contingency Table

Type of Store	No Overcharges	One or Two	More
food	104	50	11
drug	50	37	12
other	148	73	14

Expected Count Table

Store	No Overcharge	One or Two	More	Total
food	99.86	52.91	12.23	165.00
drug	59.92	31.74	7.34	99.00
other	142.23	75.35	17.42	235.00
Total	302	160	37	499

$\chi^2 = 0.172 + 0.160 + 0.125 + 1.641 + 0.870 + 2.957 + 0.235 + 0.073 + 0.673 = 6.906$
$df = 4$, P-value $= 0.141$. Since P-value $> \alpha = .05$, the null hypothesis is maintained. The data do not support the hypothesis that overcharges occurred at different rates in different types of stores in California, 1998.

16.19 Contingency Table

Year	No Overcharges	One or Two	More	Total
1998	302	160	37	499
1999	165	107	28	300
Total	467	267	65	799

Expected Count Table

Year	No Overcharges	One or Two	More	Total
1998	291.66	166.75	40.59	499
1999	175.34	100.25	24.41	300
Total	467.00	267.00	65.00	799

$\chi^2 = 0.367 + 0.273 + 0.318 + 0.610 + 0.454 + 0.529 = 2.552$. $df = 2$, P-value $= 0.279$.

Since *P*-value $> \alpha = .05$, the null hypothesis is maintained. The data do not support the hypothesis that overcharges in California occurred at different rates in 1998 and 1999.

16.21 Contingency Table

Therapy	Success	Failure	Row Total
Acupuncture	7	6	13
Placebo	4	13	17
Relaxation	2	20	22
Column Total	13	39	52

Expected Count Table

Therapy	Success	Failure	Row Total
Acupuncture	3.25	9.75	13
Placebo	4.25	12.75	17
Relaxation	5.50	16.50	22
Column Total	13.00	39.00	52

$\chi^2 = 4.327 + 1.442 + 0.015 + 0.005 + 2.227 + 0.742 = 8.759$, $df = 2$, *P*-value $= 0.013 < \alpha = 0.05$. Since the *P*-value is less than the significance level, the null hypothesis is rejected at the 5% significance level. Because two of the expected counts are less than 5, the use of chi-square tables to compute the P-value may be inappropriate.

Under the null hypothesis of independence, and given the row and column totals, the probability of success is $\pi = 13/52$ and the distribution of the number of successes for the acupuncture therapy group follows a binomial distribution with n = 13, and $\pi = 13/52$. Also the distribution of the number of successes for the placebo group, given k successes in the acupuncture therapy group follows a binomial distribution with n = 17 – k, and $\pi = 13/52$. We could use this information to compute the exact sampling distribution of the chi-square distribution, but this would be long and complicated.

16.23 Contingency Table

Therapy	Breast cancer	No breast cancer	Row Total
Raloxifen	22	5107	5129
Placebo	39	2537	2576
Column Total	61	7664	7705

Expected Count Table

Therapy	Breast cancer	No breast cancer	Row Total
Raloxifen	40.61	5088.39	5129.00
Placebo	5088.39	2555.61	2576.00
Column Total	61.00	7664.00	7705.00

$\chi^2 = 8.525 + 0.068 + 16.975 + 0.135 = 25.704$, $df = 1$, P-value $= 3.98 \times 10^{-7}/2 = 1.99 \times 10^{-7}$. The P-value is very small, so the null hypothesis is rejected at the usual significance levels. Raloxifen reduced the rate of breast cancer compared to the placebo.

16.25 Contingency Table

Treatment	Success	Failure	Row Total
Acupuncture	7	6	13
Placebo	4	13	17
Column total	11	19	30

Expected Count Table

Treatment	Success	Failure	Row Total
Acupuncture	4.77	8.23	13.00
Placebo	6.23	10.77	17.00
Column total	11.00	19.00	30.00

$\chi^2 = 1.046 + 0.606 + 0.800 + 0.463 = 2.916$, $df = 1$, P-value $= 0.088/2 = 0.044$.

Because one of the expected counts is less than 5, the use of Chi-square tables to compute the P-value may be inappropriate. We recommend using Fisher's exact test, which has a P-value of 0.0927. The one-sided chi-square test rejects the null hypothesis that the success rate for acupuncture treatment is the same as that of the placebo in favor of the alternative hypothesis that the acupuncture treatment outperforms the placebo at the 5% significance level. Fisher's exact test does not.

16.27 Contingency Table

Treatment	Mild Stroke	Moderate or severe	Row Total
Aspirin	256	253	509
No Aspirin	329	437	766
Column total	585	690	1275

Expected Count Table

Treatment	Mild Stroke	Moderate or severe	Row Total
Aspirin	233.54	275.46	509.00
No Aspirin	351.46	414.54	766.00
Column total	585.00	690.00	1275.00

$\chi^2 = 2.160 + 1.831 + 1.435 + 1.217 = 6.643$, $df = 1$, P-value $= 0.00995/2 = 0.005$.

At the 1% significance level, this one-sided chi-square test rejects the null hypothesis that the rate of mild strokes is the same for the aspirin and no-aspirin groups in favor of the alternative hypothesis that the aspirin treatment has a higher rate of mild strokes. In the study 256/509 = 50% of those in the aspirin group who had strokes had mild strokes compared to 329/766 = 43% for the other group. This chi-square test demonstrates that this difference in rates cannot easily be explained by chance variation.

16.29 Table for computing the chi-square statistic

Observed	Expected	$(O - E)^2/E$
5	5.666667	0.078431
4	5.666667	0.490196
5	5.666667	0.078431
6	5.666667	0.019608
6	5.666667	0.019608
8	5.666667	0.960784
Totals 34	34	1.647059

The null hypothesis is that the 6 categories are equally likely, the alternative is that they are not equally likely. Under the null hypothesis the sampling distribution of the chi-square statistic is approximately chi-square with 5 degrees of freedom. The *P*-value is $\Pr(\chi^2_5 \geq 1.647059) \approx 0.8955$ and the null hypothesis is retained at any of the usual significance levels. Based on the chi-square test, the data are consistent with the hypothesis that it is a random sample from a uniform distribution.

16.31 Contingency Table

Score	Smoker	Nonsmoker	Total
1	14	21	35
2	30	27	57
3	17	22	39
4	11	5	16
5–6	5	0	5
Total	77	75	152

Expected Count Table

Score	Smoker	Nonsmoker	Total
1	17.73	17.27	35.00
2	28.87	28.12	57.00
3	19.76	19.24	39.00
4	8.11	7.89	16.00
5–6	2.53	2.47	5.00
Total	77.00	75.00	152.00

$$\chi^2 = 0.785 + 0.806 + 0.044 + 0.045 + 0.385 + 0.395 + 1.034 + 1.061 + 2.403 + 2.467 = 9.424.$$

$$df = 4,\ P\text{-Value} = \Pr(\chi^2_4 \geq 9.424) \approx 0.051 > \alpha = 0.05\,.$$

The *P*-value rounds to the significance level so the null hypothesis is rejected. (If $\alpha = .050$ then the null hypothesis would be retained. Note the 95th percentile of the chi-square distribution with 4 degrees of freedom is 9.4877, which exceeds the observed chi-square statistic 9.424, so if the significance level is supposed to be precisely 5% then the null hypothesis would be retained.). However two of the expected counts are less than 5 so it may be inappropriate to use the chi-

square tables to compute a *P*-value here. You may prefer to combine the 4, 5 and 6 scores into a single group. There is some evidence that the wrinkle scores occur at different rates for smokers than nonsmokers. But the evidence is not compelling.

16.33 Contingency Table

Sex	Ambidextrous	Not Ambidextrous	Total
Men	19	981	1000
Women	7	993	1000
Total	26	1974	2000

Expected Count Table

Sex	Ambidextrous	Not Ambidextrous	Total
Men	13	987	1000
Women	13	987	1000
Total	26	1974	2000

$\chi^2 = 2.769 + 0.036 + 2.769 + 0.036 = 5.611$. $df = (r - 1)(c - 1) = 1$, P-value = $\Pr(\chi_1^2 \geq 5.611) \approx 0.018/2 < 01 = \alpha$.

Reject the null hypothesis at the 1% significance level. There is sufficient evidence to claim that men have a higher ambidexterity rate than women.

16.35 Contingency Table

Treatment	Mild Stroke	Moderate Stroke	Severe Stroke	Total
Aspirin	256	204	49	509
No Aspirin	329	324	113	766
Column total	585	528	162	1275

Expected Count Table

Treatment	Mild Stroke	Moderate Stroke	Severe Stroke	Total
Aspirin	233.54	210.79	64.67	509
No Aspirin	351.46	317.21	97.33	766
Column total	585.00	528.00	162.00	1275

$\chi^2 = 2.160 + 0.218 + 3.798 + 1.435 + 0.145 + 2.524 = 10.281$, $df = 2$,
P-value = $\Pr(\chi_2^2 \geq 10.281) \approx 0.006 < 0.01 = \alpha$. Reject the null hypothesis. There is an association between severity of stroke and aspirin usage.