Chapter 10: Categorical Data

10.1 a. Yes, because $n\hat{\pi} = 30 > 5$ and $n(1 - \hat{\pi}) = 120 > 5$. Samples with $n < 25$ would be suspect.

b. $.2 \pm 1.645\sqrt{(.2)(.8)/150} \Rightarrow (0.15, 0.25)$ is a 90% C.I. for π.

10.2 When $n\pi > 5$ and $n(1 - \pi) > 5$.

10.3 a. $\hat{\pi} = 1202/1504 = 0.8 \Rightarrow$ 95% C.I. for $\pi : 0.8 \pm 1.96\sqrt{(.8)(.2)/1504} \Rightarrow (0.780, 0.820)$

b. 90% C.I. for $\pi : 0.8 \pm 1.645\sqrt{(.8)(.2)/1500} \Rightarrow (0.783, 0.817)$

10.4 a. Yes, the binomial assumptions hold. The samples are independent, the trials are identical, the probability of success remains constant, and there are two possible outcomes.

b. Yes, $\pi = \frac{1}{3}, n = 50 \Rightarrow n\pi = \frac{50}{3} > 5$ and $n(1 - \pi) = \frac{100}{3} > 5$

c. $\hat{\pi} = \frac{21}{54} = 0.389 \Rightarrow$ 95% C.I. for $\pi : 0.389 \pm 1.96\sqrt{(.389)(.611)/54} \Rightarrow (0.259, 0.519)$
The C.I. is too wide to be very informative since as an estimate of π it provides values from 26% to over 50% for π.

In order to decrease the width, the sample size would need to be increased.

10.5 The 95% C.I.'s are summarized here: Note that $\hat{\pi}$ remains essentially unchanged from % Responding because n=1230 is very large.

Condition	95% C.I. on proportion having condition
Sore Throat	$0.30 \pm 1.96\sqrt{(.30)(.70)/1234} \Rightarrow (0.274, 0.326)$
Burns	$0.28 \pm 1.96\sqrt{(.28)(.72)/1234} \Rightarrow (0.255, 0.305)$
Alcohol	$0.25 \pm 1.96\sqrt{(.25)(.75)/1234} \Rightarrow (0.226, 0.274)$
Overweight	$0.22 \pm 1.96\sqrt{(.22)(.78)/1234} \Rightarrow (0.197, 0.243)$
Pain	$0.21 \pm 1.96\sqrt{(.21)(.79)/1234} \Rightarrow (0.187, 0.233)$

10.6 a. By grouping the classes into similar type, it might be possible to summarize the data more concisely. Percentages are helpful but would not add to 100% because one adult might use more than one of the remedies. The numerator of the percentage would refer to users of an OTC remedy and the denominator to the number of patients.

b. A 95% C.I. using the normal approximation requires that both $n\hat{\pi}$ and $n(1 - \hat{\pi})$ exceed 5. This condition would hold in every OTC category except Sprays/Inhalers, Anesthetic throat lozenges, Room vaporizers and Other products.

10.7 $\hat{\pi} = 88/254 = 0.346 \Rightarrow$ 90% C.I. for $\pi : 0.346 \pm 1.645\sqrt{(0.346)(0.654)/254} \Rightarrow (0.297, 0.395)$

10.8 The 95% C.I.'s are given here:

Statement	95% C.I. on Proportion
Others Don't Report	$.56 \pm 1.96\sqrt{(.56)(.44)/504} \Rightarrow (0.516, 0.604)$
Government is Careless	$.50 \pm 1.96\sqrt{(.50)(.50)/504} \Rightarrow (0.456, 0.544)$
Cheating can be Overlooked	$.46 \pm 1.96\sqrt{(.46)(.54)/504} \Rightarrow (0.416, 0.504)$

10.9 a. A bar chart with the responses along the horizontal axis and the percentages along the vertical axis would allow comparison of the responses.

b. Yes, since the C.I.'s would reflect the sampling errors of the point estimators and hence be more informative of the size of the true proportions.

c. The report lists only a select few responses in the United States, ignoring the most and least popular ones as well as almost all of the foreign figures. For those percentages reported, it does not include the sample size and so the reader gets no idea of the accuracy of the reported sample proportions as estimates of the population proportions.

10.10 a. A table summarizing the results is given here:

Statement	No	Yes
Understand Radiation	70%	30%
Misconceptions About Space-Rockets	40%	60%
Understand How Telephone Works	80%	20%
Understand Computer Software	75%	25%
Understand Gross National Product	72%	28%

b. Unmentioned details include a complete list of questions asked and the manner in which they were stated. The article also does not report how the survey was conducted. Thus, the results may be biased if the sample was not selected in a random fashion. For example, if the questionaire were given by mail, the responses would come from only those individuals who were able to read and write. This would bias the results because illiterate people probably understand less about technology than literate ones. Other demographic characteristics of the sample might also bias the results.

10.11 a. $n\pi_o = (800)(0.096) = 76.8 > 5$ and $n(1 - \pi_o) = (800)(1 - 0.096) = 723.2 > 5$ thus the normal approximation would be valid.

b. $H_o : \pi \geq 0.096$ versus $H_a : \pi < 0.096$
$\hat{\pi} = 35/800 = 0.04375, //z = \frac{0.04375-0.096}{\sqrt{(0.096)(0.904)/800}} = -5.02 \Rightarrow$
p-value $= P(z < -5.02) < 0.0001 \Rightarrow$ Reject H_o and
conclude there is significant evidence that $\pi < 0.096$.

10.12 a. $\hat{\pi} = 562/1504 = 0.374 \Rightarrow$ 95% C.I. on $\pi : (0.350, 0.398)$
Half width of C.I. is 0.024

b. Using $\hat{\pi} = 0.374$, $n = \frac{(1.96)^2(0.374)(.626)}{(0.01)^2} = 8984.1 \Rightarrow n = 8985$

10.13 $\hat{\pi} = 10/24 = 0.417 \Rightarrow 95\%$ C.I. on $\pi : (0.220, 0.614)$

10.14 a. $\hat{\pi}_{Adj.} = \frac{\frac{3}{8}}{100+\frac{3}{4}} = 0.00372$

b. 99% C.I. on $\pi : (0, 1 - (.005)^{\frac{1}{100}}) = (0, 0.0516)$

c. $H_o : \pi \geq 0.01$ versus $H_a : \pi < 0.01$

Because 0.01 falls in the C.I., fail to reject H_o and conclude that the data fails to support the company's claim. The level of significance of the test would be $\alpha = \frac{1-.99}{2} = 0.005$. The problem is that with such a small value for π_o, the sample size must be much larger in order for the company to be able to support its claim.

10.15 $\hat{\pi}_1 = 109/200 = 0.545$ (Republicans) and $\hat{\pi}_2 = 86/200 = 0.43$ (Democrats)

$z = \frac{0.545-0.43}{\sqrt{\frac{0.545(1-0.545)}{200}+\frac{0.43(1-0.43)}{200}}} = 2.32 \Rightarrow$ p-value $= 0.0102$

Reject H_o and conclude that a large proportion of Republicans are in favor of the incentives.

10.16 Because p-value $= 0.0218 < 0.05$, reject H_o and conclude that the data supports the hypothesis that the rates of satisfied customers served by the two methods are different.

10.17 95% C.I. on $\pi_1 - \pi_2 : (0.013, 0.167)$

Because 0 is not contained within the C.I., H_o is rejected and hence the conclusion is the same as was in Exercise 10.16.

10.18 a. 95% C.I. on $\pi_1 - \pi_2 : 0.478 - 0.376 \pm 1.96\sqrt{\frac{0.478(1-0.478)}{473} + \frac{0.376(1-0.376)}{439}} \Rightarrow (0.038, 0.166)$

b. Yes, $n\hat{\pi}$ and $n(1-\hat{\pi})$ are greater than 5 for both samples.

c. Yes, because 0 is not contained within the C.I., H_o is rejected.

10.19 $H_o : \pi_1 = \pi_2$ versus $H_a : \pi_1 \neq \pi_2$

p-value $= 0.0018 < 0.05 \Rightarrow$ Reject H_o and conclude there is significant evidence that the population proportions are different.

10.20 a. $H_o : \pi_1 = \pi_2$ versus $H_a : \pi_1 \neq \pi_2$

$\hat{\pi}_1 = 0.32, \hat{\pi}_2 = 0.20 \Rightarrow$

$z = \frac{0.32-0.20}{\sqrt{\frac{0.32(1-0.32)}{310}+\frac{0.2(1-0.2)}{309}}} = 3.44 \Rightarrow$ p-value $= 0.0006$

Reject H_o and conclude there is significant evidence in the proportion of males with new hair growth.

b. What was the amount of hair growth? What side effects were observed? What characteristics distinguished males who demonstrated hair growth from those who did not?

10.21 a. $z = \frac{0.90-0.36}{\sqrt{\frac{0.9(1-0.9)}{100}+\frac{0.36(1-0.36)}{100}}} = 9.54 \Rightarrow$ p-value $= P(z > 9.54) < 0.0001$

Reject H_o and conclude there is significant evidence that the death rate after 30 days is greater for Cocaine group than for the Heroin group.

b. If the physical response to the two drugs is the same for humans, cocaine is a very dangerous drug, even more so than heroin.

10.22 $H_o : \pi_1 = 0.40, \pi_2 = 0.20, \pi_3 = 0.25, \pi_4 = 0.15$

H_a : at least on of the $\pi_i s$ differs from its hypothesized value

$E_i = n\pi_{io} \Rightarrow \quad E_1 = 1000(.4) = 400, \quad E_2 = 1000(.2) = 200,$

$E_3 = 1000(.25) = 250, \quad E_4 = 1000(.15) = 150$

$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - E_i)^2}{E_i} = 49.067$ with $df = 4 - 1 = 3 \Rightarrow$ p-value $< 0.0001 \Rightarrow$

Reject H_o. There is substantial evidence that the whole life policies have decreased in popularity and the universal life have increased in popularity.

10.23 $H_o : \pi_1 = \frac{1}{3}, \pi_2 = \frac{1}{3}, \pi_3 = \frac{1}{3}$

H_a : at least on of the groups had probability of interning different from $\frac{1}{3}$

$E_i = n\pi_{io} = 63\pi_{io} \Rightarrow \quad E_1 = 21, \quad E_2 = 21 \quad E_3 = 21$

$\chi^2 = \sum_{i=1}^{3} \frac{(n_i - E_i)^2}{E_i} = 6.952$ with $df = 3 - 1 = 2 \Rightarrow$ p-value $= 0.0309 > 0.01 \Rightarrow$

Fail to reject H_o. The data does not appear to contradict the claim that students finishing an internship are equally distributed from the three industries.

10.24 $H_o : \pi_1 = 0.25, \pi_2 = 0.48, \pi_3 = 0.20, \pi_4 = 0.07$

H_a : at least on of the $\pi_i s$ differs from its hypothesized value

$E_i = n\pi_{io} \Rightarrow \quad E_1 = 400(.25) = 100, \quad E_2 = 400(.48) = 192,$

$E_3 = 400(.20) = 80, \quad E_4 = 400(.07) = 28$

$\chi^2 = \sum_{i=1}^{4} \frac{(n_i - E_i)^2}{E_i} = 181.65$ with $df = 4 - 1 = 3 \Rightarrow$ p-value $< 0.0001 \Rightarrow$

Reject H_o. There is substantial evidence that the proportion of mentally ill patients of four social classes housed in a county health facility differ from the proportion residing in the county in general.

10.25 $H_o : \pi_1 = 0.50, \pi_2 = 0.40, \pi_3 = 0.10$

H_a : at least on of the $\pi_i s$ differs from its hypothesized value

$E_i = n\pi_{io} \Rightarrow \quad E_1 = 200(.5) = 100, \quad E_2 = 200(.4) = 80, \quad E_3 = 200(.1) = 20$

$\chi^2 = \sum_{i=1}^{3} \frac{(n_i - E_i)^2}{E_i} = 6.0$ with $df = 3 - 1 = 2 \Rightarrow$ p-value $= 0.0498 \Rightarrow$

Reject H_o at the $\alpha = 0.05$ level. There is substantial evidence that the distribution of registered voters is different from previous elections.

10.26 $H_o : \pi_1 = 0.6, \pi_2 = 0.3, \pi_3 = 0.1$

H_a : at least on of the $\pi_i s$ differs from its hypothesized value

$E_i = n\pi_{io} \Rightarrow \quad E_1 = 85(.6) = 51, \quad E_2 = 85(.3) = 25.5, \quad E_3 = 85(.1) = 8.5$

$\chi^2 = \sum_{i=1}^{3} \frac{(n_i - E_i)^2}{E_i} = 27.745$ with $df = 3 - 1 = 2 \Rightarrow$ p-value $< 0.001 \Rightarrow$

Reject H_o at the $\alpha = 0.05$ level. There is substantial evidence that the distribution of responses for depressed adults differs from the responses of nondepressed adults.

10.27 $H_o : \pi_1 = 0.0625, \pi_2 = 0.25, \pi_3 = 0.375, \pi_4 = 0.25, \pi_5 = 0.0625$

H_a : at least on of the $\pi_i s$ differs from its hypothesized value

$E_i = n\pi_{io} \Rightarrow \quad E_1 = 125(.0625) = 7.8125, \quad E_2 = 125(.25) = 31.25,$

$E_3 = 125(.375) = 46.875, \quad E_4 = 125(.25) = 31.25, \quad E_5 = 125(.0625) = 7.8125$

$\chi^2 = \sum_{i=1}^{5} \frac{(n_i - E_i)^2}{E_i} = 7.608$ with $df = 5 - 1 = 4 \Rightarrow$ p-value $= 0.1070 \Rightarrow$

Fail to reject H_o. The data appear to fit the hypothesized theory that the securities analysts perform no better than chance, however, we have no indication of the probability of a Type II error.

10.28 a. $\chi^2 = 15.97$

b. p-value = 0.0139

c. At the $\alpha = 0.05$ level, there is significant evidence of a relationship.

d. No. There are two $E_{ij}s$ that are less than 5. However, the guideling for the applying the χ^2-statistic is met because only $2/12 = 17\%$ of the $E_{ij}s$ are less than 5 and all the $E_{ij}s$ are greater than 1.

10.29 Yes, since the row percentages differ considerably for the four categories of schools.

10.30 a. The expected counts are $E_{ij} = n_{i.}n_{.j}/250$ and are displayed in the following table:

	Column			
Row	1	2	3	4
1	16.0	22.4	25.6	16.0
2	34.0	47.6	54.4	34.0

b. df = (2-1)(4-1) = 3

c. Using the chi-square approximation with df = 3 and $\chi^2 = 13.025, \Rightarrow$
p-value = 0.0046. Thus, there is significant evidence of a relationship.

10.31 p-value = 0.0046

10.32 a. Using the chi-square approximation with df = 1 and $\chi^2 = 0.012, \Rightarrow$ p-value = 0.9128. Thus, there is not significant evidence to reject the hypothesis of independence.

b. There was a considerable loss of important information. Combining the age categories masks the differences found when examining the data with a greater number of categories. In Exercise 10.30, the hypothesis of independence was rejected.

10.33 a. The 25%, 40%, and 35% claims concerning opinions on union membership was made for industrial workers as a whole without regard to membership status. The relevant data are the column totals of those favoring, those indifferent, and those opposed industrial workers, i.e., 210, 240, and 150, respectively.

b. The following table summarizes the information needed for the goodness-of-fit test:

Preference	Theoretical	Expected	Observed
	Proportions	Frequencies	Frequencies
	π_i	$E_i = 600\pi_i$	n_i
Favor	0.25	150	210
Indifferent	0.40	240	240
Oppose	0.35	210	150

$H_o : \pi_1 = 0.25, \pi_2 = 0.40, \pi_3 = 0.35$ versus H_a : Specified proportions are not correct
$\chi^2 = \sum_{i=1}^{3} \frac{(n_i - E_i)^2}{E_i} = 41.143$ with $df = 3 - 1 = 2 \Rightarrow$ p-value $< 0.0001 \Rightarrow$
Reject H_o. There is significant evidence that the speaker's claim is not supported by the data.

10.34 With $df = 2$, p-value $= P(\chi^2 > 41.143) < 0.0001$

10.35 H_o : Membership Status and Opinion are independent Versus H_a : Membership Status and Opinion are related

The expected values in each cell and the cell chi-square values are given in the following table with the expected given above the cell chi-square values:

Status	Favor	Indifferent	Opposed
Members	70	80	50
	70.00	18.05	20.48
Nonmembers	140	160	100
	35.00	9.03	10.24

$\chi^2 = \sum_{i,j} \frac{(n_{ij} - E_{ij})^2}{E_{ij}} = 162.80$ with $df = (2-1)(3-1) = 2 \Rightarrow$ p-value $< 0.0001 \Rightarrow$
Reject H_o. There is significant evidence that the Membership Status and Opinion are related.

10.36 The percentages are given in the following table:

Status	Favor	Indifferent	Opposed
Members	70%	21%	9%
Nonmembers	17.5%	49.5%	33.0%

The percentages are very different for the two groups, indicating a strong relation between membership status and opinion.

10.37 a. Under the hypothesis of independence, the expected frequencies are given in the following table: $\hat{E}_{ij} = n_{i.}n_{.j}/900$

	Opinion				
Commercial	1	2	3	4	5
A	42	107	78	34	39
B	42	107	78	34	39
C	42	107	78	34	39

b. df = (3-1)(5-1) = 8

c. The cell chi-squares are given in the following table:

	Opinion				
Commercial	1	2	3	4	5
A	2.3810	3.7383	2.1667	4.2353	0.6410
B	2.8810	10.8037	0.0513	5.7647	21.5641
C	0.0238	1.8318	1.5513	0.1176	14.7692

$\chi^2 = \sum_{i,j} \frac{(n_{ij}-E_{ij})^2}{E_{ij}} = 72.521$ with $df = 8 \Rightarrow$ p-value $< 0.0001 \Rightarrow$
Reject H_o. There is significant evidence that the Commercial viewed and Opinion are related.

10.38 With $df = 8$, p-value $= P(\chi^2 > 72.521) < 0.0001$

10.39 a. The null hypothesis of statistical independence is equivalent to stating that there is no relationship between the type of order form and whether or not an order is received.

b. The p-value = 0.00007 from the chi-square test of independence. Thus, we reject the null hypothesis of independence and conclude that there is an association between the type of order form and whether or not an order was received.

10.40 a. The Row % entries suggest there is a relation between SOPHIST and PREFER. For instance, when SOPHIST=1, the PREFER values cluster towards the lower values of PREFER (suggesting that less sophisticated users have more of a preference for the current version of the program). On the extreme, the most sophisticated users, SOPHIST=3 seem to favor the new version (higher values of PREFER). If there was no relation between SOPHIST and PREFER, we would expect to see similar row percentages across the all three rows.

b. The Pearson Chi-Square has a value of 44.543 and a p-value < 0.0001. Since the p-value is less than the conventional values of α, we reject the null hypothesis of no association, and conclude that there is significant evidence of a relation between the variables SOPHIST and PREFER.

10.41 a. Control: 10%; Low Dose 14%; High Dose 19%

b. $H_o : \pi_1 = \pi_2 = \pi_3$ versus H_a : The proportions are not all equal, where π_j is probability of a rat in Group j having One or More Tumors.

$E_{ij} = 100n_{.j}/300$ and $\chi^2 = \sum_{ij} \frac{(n_{ij}-E_{ij})^2}{E_{ij}} = 3.312$ with df = (2-1)(3-1) = 2 and p-value = 0.191.

Because the p-value is fairly large, we fail to reject H_o and conclude there is not significant evidence of a difference in the probability of having One or More Tumors for the three rat groups.

c. No, since we the chi-square test failed to reject H_o.

10.42 Same results.

10.43 a. Expected cell counts are given here:

	Years of Education			
Years First Job	0-4.5	4.5-9	9-13.5	> 13.5
0-2.5	17.90	20.97	23.78	26.34
2.5-5	24.14	28.28	32.07	35.52
5-7.5	16.70	19.56	22.18	24.57
> 7.5	11.26	13.20	14.97	16.57

b. $\chi^2 = 57.830$, with df = (4-1)(4-1) = 9

c. p-value < 0.001

d. There is significant evidence that Years of Education are related to Years on First Job.

10.44 a. The results are summarized in the following table: with $\hat{\sigma}_{\hat{\pi}} = \sqrt{(\hat{\pi})(1-\hat{\pi})/504}$ and 95% C.I. $\hat{\pi} \pm 1.96\hat{\sigma}_{\hat{\pi}}$

Question	$\hat{\pi}$	$\hat{\sigma}_{\hat{\pi}}$	95% C.I.
Did Not Explain?	0.256	0.0194	(0.218, 0.294)
Might Bother?	0.913	0.0126	(0.888, 0.938)
Did Not Ask?	0.472	0.0223	(0.428, 0.516)
Drug Not Changed?	0.875	0.0147	(0.846, 0.904)

b. It would be important to know how the patients were selected, how the questions were phrased, the condition of the illness, and many other factors.

10.45 a. Building: $\chi^2 = 34.167$ with p-value < 0.0001. Thus, there is substantial evidence that the customers from the different Ownership Group categories have different distributions relative to the Building Ratings categories.

Service: $\chi^2 = 18.117$ with p-value $= 0.112$. Thus, there is not significant evidence that customers from the different Ownership Group categories have different different distributions relative to the Service Ratings categories.