Chp 22 Two Categorical Variables: The Chi-Square Test

☆ a new test that we can use to determine if there is a relationship between two categorical variables!

Ex 1

Attitudes toward recycled products. Some people think recycled products are lower in quality than other products, a fact that makes recycling less practical. Here are data on attitudes toward coffee filters made of recycled paper.[5]

	Think the Quality of the Recycled Product Is		
	Higher	**Same**	**Lower**
Buyers	20	7	9
Nonbuyers	29	25	43

☆ note: Two way table for categorical data.

(a) It appears that people who have bought the recycled filters have more positive opinions than those who have not. Give percents to back up this claim. Make a bar graph that compares your percents for buyers and nonbuyers.

(b) Association does not prove causation. Explain how buying recycled filters might improve a person's opinion of their quality. Then explain how the opinion a person holds might influence his or her decision to buy or not. You see that the cause-and-effect relationship might go in either direction.

Chp 22 (cont)

★ use chi-square to do multiple comparisons at once (do not use separate CIs for each of 2 variables)

w/ in general
Ho: there is no relationship between 2 variables
Ha: Ho not true

To test Ho, we compare observed counts from our sample w/ expected counts, if Ho were true. If observed counts are far from expected counts, that is evidence against Ho.

Expected Counts in any cell of a two-way table, when Ho is true, is

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

why? If p = probability of "success" + we have n tries, then we expect np "successes".
and in general $p = \frac{\text{row total}}{\text{table total}}$ + n = column total
$\Rightarrow np$ gives expected count (above)

Ex 2

Attitudes toward recycled products. Exercise 22.2 describes a comparison of the attitudes of people who do and don't buy coffee filters made of recycled paper. The null hypothesis "no relationship" says that in the population of all consumers, the proportions who hold each attitude are the same for buyers and nonbuyers.

(a) Find the expected cell counts if this hypothesis is true and display them in a two-way table. Add the row and column totals to your table and check that they agree with the totals for the observed counts.

(b) Are there any large deviations between the observed counts and the expected counts? What kind of relationship between the two variables do these deviations point to?

Chp 22 (cont)

Ex 2 (cont)

Chi-Square Statistic

χ^2 = measure of how far the observed counts in a 2-way table are from expected counts.

$$\chi^2 = \sum_{\text{all cells in table}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

note: ① $\chi^2 \geq 0$ + $\chi^2 = 0$ only when observed counts = expected counts

② large values of χ^2 = evidence against H_0 (because observed counts are far from what we expect if H_0 true)

③ Although H_a is many-sided, χ^2 test is one-sided because any violation of H_0 tends to produce large χ^2-value.
small χ^2-value not evidence against H_0.

Chp 22 (cont).

Cell Counts required for χ^2 test

Use χ^2 test when no more than 20% of expected counts are less than 5, and when all individual expected counts ≥ 1.

Ex 3

Attitudes toward recycled products. Your data analysis in Exercise 22.2 found that people who have bought recycled coffee filters tend to have more positive opinions about the filters than those who have not. Figure 22.4 gives Minitab output for the two-way table in Exercise 22.2. (Ex 1)

(a) Verify from the output that the data meet the cell count requirement for use of chi-square.

(b) What are the chi-square statistic and its P-value? Explain in simple language what it means to reject H_0 in this setting.

(c) Give an overall conclusion that refers to row percents to describe the nature of the relationship between attitude and decision to buy or not.

Session

	Higher	Same	Lower	All
Buyer	20	7	9	36
	55.56	19.44	25.00	100.00
	13.26	8.66	14.08	36.00
Nonbuyer	29	25	43	97
	29.90	25.77	44.33	100.00
	35.74	23.34	37.92	97.00
All	49	32	52	133
	36.84	24.06	39.10	100.00
	49.00	32.00	52.00	133.00

Cell Contents: Count
% of Row
Expected count

Pearson Chi-Square = 7.638, DF = 2, P-Value = 0.022

The key for cell entries in this table is at the bottom.

FIGURE 22.4
Minitab output for the study of consumer attitudes toward recycled products.

M1070

Chp 22 (cont)

Uses of χ^2 test

Use χ^2 test for

H_0: no relationship between 2 categorical variables

when you have 2-way table from one of these situations

① Independent SRSs from 2 or more populations, w/ each individual classified according to one categorical variable.

② A single SRS with each individual classified according to both of two categorical variables.

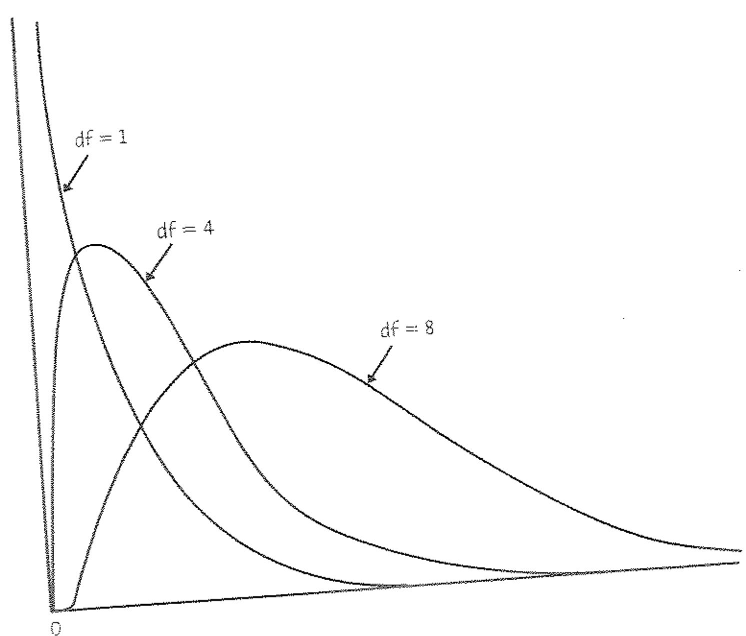
χ^2 Distribution

- a family of distributions that take only positive values & are skewed to the right.
- need to know df (degrees of freedom) in 2-way table w/ r rows and c columns

$$df = (r-1)(c-1)$$

- p-value is area under density curve to right of test statistic.

★ can use Table D in back of book.



Chp 22 (cont)

on previous page

Ex 4 **Attitudes toward recycled products.** The Minitab output in Figure 22.4 gives 2 degrees of freedom for the table in Exercise 22.2. (Ex 1)

(a) Verify that this is correct.

(b) The computer gives the value of the chi-square statistic as $\chi^2 = 7.638$. Between what two entries in Table D does this value lie? Verify that Minitab's P-value does fall between the tail probabilities p for these two entries.

(c) What is the mean value of the statistic χ^2 if the null hypothesis is true? How does the observed value of χ^2 compare with this mean?

* χ^2 can also be used to test goodness of fit of a particular distribution.

Say a categorical variable has k possible outcomes, w/ probabilities $p_1, p_2, \ldots, p_k$. Have n independent observations.

To test $H_0: p_1 = p_{10}, p_2 = p_{20}, \ldots, p_k = p_{k0}$, calculate

$$\chi^2 = \sum_{\text{all possible outcomes}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

df $= k-1$

Chp 22 (cont)

Ex 5

More on birth days. Births really are not evenly distributed across the days of the week. The data in Example 22.8 failed to reject this null hypothesis because of random variation in a quite small number of births. Here are data on 700 births in the same locale:

Day	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Births	84	110	124	104	94	112	72

(a) The null hypothesis is that all days are equally probable. What are the probabilities specified by this null hypothesis? What are the expected counts for each day in 700 births?

(b) Calculate the chi-square statistic for goodness of fit.

(c) What are the degrees of freedom for this statistic? Do these 700 births give significant evidence that births are not equally probable on all days of the week?

M1070

89

Chp 22 (cont)

EX6

Course grades. Most students in a large statistics course are taught by teaching assistants (TAs). One section is taught by the course supervisor, a senior professor. The distribution of grades for the hundreds of students taught by TAs this semester was

Grade	A	B	C	D/F
Probability	0.32	0.41	0.20	0.07

The grades assigned by the professor to students in his section were

Grade	A	B	C	D/F
Count	22	38	20	11

(These data are real. We won't say when and where, but the professor was not the author of this book.)

(a) What percents of each grade did students in the professor's section earn? In what ways does this distribution of grades differ from the TA distribution?

(b) Because the TA distribution is based on hundreds of students, we are willing to regard it as a fixed probability distribution. If the professor's grading follows this distribution, what are the expected counts of each grade in his section?

(c) Does the chi-square test for goodness of fit give good evidence that the professor's grades follow a different distribution? (State hypotheses, check the guidelines for using chi-square, give the test statistic and its P-value, and state your conclusion.)

m1070

(90)

Table entry for p is the critical value χ^* with probability p lying to its right.

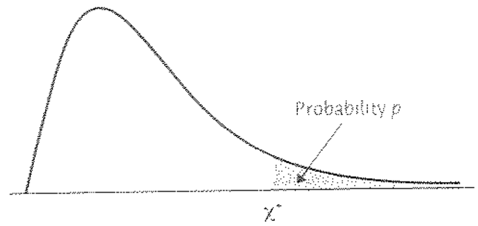


TABLE D Chi-square distribution critical values

						p						
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.7
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8	128.3
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4	153.2