To look for an association between two categorical variables, we first put the data into a table (called a two-way table). We then looked at the marginal distributions, and finally examined the conditional distributions. The conditional distributions are key to determining if there is a significant association between two categorical variables, and what this association looks like.

The conditional distributions, recall, are the distributions of one variable for each level of the other variable. If these distributions are the same for each level of the second variable, we would conclude there was no association. If these distributions differ for the different levels of the second variable, we might conclude that there is an association.

When the conditional distribution suggests there may be an association, we want to see if that association is *statistically significant* (greater than what we would expect to see just by chance—i.e. sampling variation). To do this, we will do some hypothesis testing.

The hypothesis test that we use in this situation is called the Pearson chi-square test. And it follows the same outline as our other hypothesis tests, but uses a new test statistic (that has a new distribution called a chi-square distribution).

1. State the null and alternative hypotheses. Generally speaking, they are

   H_0: There is no association between rows and columns; they are independent.
   H_a: The row and column variables are not independent.

2. Compute the test statistic, χ^2
3. Find the p-value, based on the appropriate χ^2-distribution. This is *always* a single-tailed value.
4. Draw a conclusion.

Chi-square tests are generally used to answer 1 of 2 questions (or, the same question phrased 2 ways):

1. Does the percent in a certain category change from population to population?

   Example: Does the percent of deaths from heart disease change from country to country? Or we could say, is there an association between the percent of deaths from heart disease and which country a person is from? (data from 1965, males aged 35-64 in selected countries)

2. Are these two categorical variables independent.?

   Example: Is alcohol consumption associated with oesophageal cancer? (data from Tuyns et all 1977)

Let's back up. Where does this Pearson chi-square come from?

1 Looking at one (categorical) variable with k values

The test statistic is:

$$\chi^2 = \sum \frac{[(\text{observed count}) - (\text{expected count})]^2}{\text{expected count}}, \qquad \text{with } df = k - 1,$$

where the *observed count* is an actual cell count from the sample, and the *expected count* is the (average) count a cell would have if there were no association between the variables.

Decision rule: At the α level of significance, reject H_0 if $\underline{P \leq \alpha}$, and do not reject H_0 if $\underline{P > \alpha}$.

Example: The following data give for each day of a one-week period the number of crimes reported to a certain police station. In general, is crime distributed uniformly across days of the week?

Day	Su	Mo	Tu	We	Th	Fr	Sa	Total
# of crimes	48	38	31	40	39	40	44	280
Expected # if each $p_i = 1/7$	40	40	40	40	40	40	40	

Following our solution scheme:

1. **State null and alternative hypotheses**:

   H_0: The # of crimes per day is the same for each day of the week.
   H_a: The # of crimes is not the same for each day of the week.

2. **Compute the test statistic**:
   First we need to know the expected counts for each day of the week. These expected counts are what would suppose the counts, on average, to be, if the null hypothesis were true. Guess what they ought to be.

   Since the denominators in the formula for the test statistic are all equal, we may write it as

$$\chi^2 = \frac{1}{40}\left(8^2 + 2^2 + 9^2 + 0^2 + 1^2 + 0^2 + 4^2\right) = \frac{166}{40} = 4.15.$$

3. **Compute the P-value**:

   Using Table F with $df = 7 - 1 = 6$, we see that $P(\chi^2 > 4.15) > 0.25$.

4. **Draw a conclusion**:

   At any reasonable significance level we find this sample data insufficiently different (as measured by the test statistic) from what we would have expected under H_0 to reject it. In other words, the sample data is consistent with the belief that crimes are committed with the same frequency regardless of the day of the week. $(\chi^2 = 4.15, df = 6, P > 0.25)$

2 Looking for a relationship between 2 (categorical) variables

Example: Are estrogen use and endometrial cancer related?

1. **State null and alternative hypotheses**.

   To do this, we must discern what variables are of interest to us. For the above question, the two variables are duration of estrogen use and status of endometrial cancer. These may be thought of as categorical variables, with estrogen use serving as the explanatory variable.

   H_0: Whether or not a woman gets endometrial cancer is independent of how long she has used estrogen (there is no association).

   H_a: There is an association between endometrial cancer and duration of estrogen use.

   At this point an appropriate sample (involving some sort of random selection process) would be taken to help decide between these two options. It is usually helpful to arrange the sample data in a two-way (contingency) table. Generally speaking, the values of the explanatory variable are placed along rows, and the values of the response variable are placed along columns.

Duration of use (yrs)	Endometrial cancer cases	Healthy	Total
None	274 (41.3%)	390 (58.7%)	664
under 1	11 (61.1%)	7 (38.9%)	18
1–5	17 (68%)	8 (32%)	25
above 5	36 (92.3%)	3 (7.7%)	39
Total	338 (45.3%)	408 (54.7%)	746

   Here, the values available for the two variables of interest are:

   Explanatory variable: no estrogen use, use for under a year, use for a period of 1–5 yrs, and use for longer than 5 yrs

   Response variable: endometrial cancer did occur, endometrial cancer did not occur

   The percents included in parentheses do not need to be included in the table. They are helpful here in demonstrating the sample differences in proportions of occurences of endometrial cancer between different levels of estrogen use. We have computed such percentages when talking about two-way tables before (Section 2.6). They give the conditional distribution of endometrial cancer status given estrogen use. You can tell this because percentages add to 100% along rows.

2a. **Compute the expected counts**.

For this handout, we will use the notation:

o_{ij} = observed (sample) count in the ith row and jth column.

R_i = total count for row i.

C_j = total count for row j.

n = total count for the whole table.

e_{ij} = the expected count for the ijth cell.

As with all tests of hypothesis, we will compute a test statistic based on the assumption that H_0 is true. Our expected counts are based upon this assumption as well.

$$\text{If } H_0 \text{ is true, then } \; e_{ij} = \frac{R_i C_j}{n}.$$

Assuming independence of the two variables (H_0), the expected count in the 3rd row, 1st column would be

$$e_{31} \;=\; \frac{(25)(338)}{746} \;=\; 11.32708.$$

The original two-way table is given below, this time with the expected counts appearing in parentheses.

Duration of use (yrs)	Endometrial cancer cases	Healthy	Total
None	274 (300.85)	390 (363.15)	664
under 1	11 (8.16)	7 (9.84)	18
1–5	17 (11.33)	8 (13.67)	25
above 5	36 (17.67)	3 (21.33)	39
Total	338	408	746

2b. **Compute the test statistic**.

Once again, the test statistic has the formula

$$\chi^2 \;=\; \sum \frac{[\text{(observed count) - (expected count)}]^2}{\text{expected count}} \;=\; \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}.$$

What is different in the two-variable case is the number of degrees of freedom, which is $df = (\#\text{rows} - 1)(\#\text{columns} - 1)$.

For our endometrial cancer data, the value of the test statistic is

$$\begin{aligned} \chi^2 \;&=\; \frac{(274-300.85)^2}{300.85} + \frac{(390-363.15)^2}{363.15} + \frac{(11-8.16)^2}{8.16} + \frac{(7-9.84)^2}{9.84} \\ &\quad + \frac{(17-11.33)^2}{11.33} + \frac{(8-13.67)^2}{13.67} + \frac{(36-17.67)^2}{17.67} + \frac{(3-21.33)^2}{21.33} \\ &=\; 46.1455. \end{aligned}$$

3. **Compute the P-value**:
   Here there are degrees of freedom $df = (4-1)(2-1) = 3$. We consult Table F to find that the probability of getting sample results at least as extreme as this when H_0 is true is $P(\chi^2 \geq 46.15) < 0.0005$.

4. **Draw a conclusion**:
   We reject H_0 (that is, reject at any of the three significance levels usually used), concluding that duration of estrogen use does play a role in the appearance of endometrial cancer.

Several notes:

A. Some two-way tables are 2×2—that is, they have just two rows and two columns. In such cases, the χ^2-test is not the only test we might use to test the null hypothesis of *independence of variables*. The 2-proportion test of significance suits the same purpose. (See the 2nd paragraph on p. 632 of your textbook.)

B. Like the other inference procedures we have learned, results from a χ^2 test are valid only under certain assumptions. One such assumption, of course, is that the sample is a random sample. But another assumption is that the **expected cell counts must be large enough**. Specifically, for 2×2 tables, all four expected values need to be 5 or more. For larger tables, the average of the expected counts needs to be 5 or more, and the smallest expected count needs to be at least 1.

Example: In January 1975, the Committee on Drugs of the American Academy of Pediatrics recommended that tetracycline drugs not be used for children under the age of 8. A 2-year study was conducted in Tennessee to investigate the extent to which physicians prescribed this drug between 1973 and 1975. Of 214 physicians whose practices were in urban counties, 65 prescribed tetracycline during this period; of 330 physicians in rural counties, 172 did; and of 226 physicians in intermediate counties, 90 did. Is the proportion of physicians prescribing tetracycline the same in each type of county? Test at level $\alpha = 0.01$.

Our hypotheses:

H_0: Doctors prescribe tetracycline independent of county type.

H_a: Doctors prescribe tetracycline differently in different types of counties.

Here the type of county is the explanatory variable. A two-way table organizing the data, and giving expected cell counts (in parentheses), looks like

Type of county	# MDs prescribing tetracycline	# MDs not prescribing it	Total
Rural	172 (140.14)	158 (189.86)	330
Intermediate	90 (95.98)	136 (130.02)	226
Urban	65 (90.88)	149 (123.12)	214
Total	327	443	770

Our test statistic, then, is

$$\begin{aligned}\chi^2 &= \frac{(172-140.14)^2}{140.14} + \frac{(158-189.86)^2}{189.86} + \frac{(90-95.98)^2}{95.98} \\ &\quad + \frac{(136-130.02)^2}{130.02} + \frac{(65-90.88)^2}{90.88} + \frac{(149-123.12)^2}{123.12} \\ &= 26.04705.\end{aligned}$$

For $df = 2$, Table F indicates $P(\chi^2 > 26.05) < 0.0005$. So, we reject H_0 at the 1% level, concluding that doctors prescribe tetracycline differently based upon county type. ($\chi^2 = 26.05, df = 2, P < 0.0005$)

Example: Suppose we modify the last example to include only the sample information for urban and intermediate county doctors. Here is a two-way table which gives the data, the conditional distribution of tetracycline-using doctors across county type (the first numbers inside cell parentheses) and the expected cell counts (the latter numbers inside parentheses). (Notice that the expected cell counts are different than before because of the dropping the rural counties.)

Type of county	# MDs prescribing tetracycline	# MDs not prescribing it	Total
Intermediate	90 (39.82%, 79.61)	136 (60.18%, 146.39)	226
Urban	65 (30.37%, 75.39)	149 (69.63%, 138.61)	214
Total	155	285	440

We test the hypotheses:

H_0: Doctors prescribe tetracycline independently of county type,

H_a: Doctors prescribe tetracycline differently in different types of counties,

as before. Now, however, we have a 2×2 two-way table. So, we have two methods we can use to determine a P-value:

2-proportion test:

Let's denote quantities for intermediate counties with the subcript '1', and urban counties with '2'. If X_j gives the count of doctors who prescribe tetracycline, then, we have

$$X_1 = 90,\ n_1 = 226,\ \hat{p}_1 = \frac{X_1}{n_1} = 0.3982, \qquad \text{and} \qquad X_2 = 65,\ n_1 = 214,\ \hat{p}_1 = \frac{X_1}{n_1} = 0.3037.$$

Our hypotheses may be rewritten as

H_0: $p_1 - p_2 = 0$

H_a: $p_1 - p_2 \neq 0$

Under the null hypothesis, our pooled estimate for the common value of p_1 and p_2 is

$$\hat{p} \;=\; \frac{X_1 + X_2}{n_1 + n_2} \;=\; \frac{90 + 65}{226 + 214} \;=\; 0.3523.$$

Our z-statistic is

$$z \;=\; \frac{(0.3982 - 0.3037) - 0}{\sqrt{(0.3523)(1 - 0.3523)[(1/226) + (1/214)]}} \;=\; 2.0741.$$

Consulting Table A, we have $P(|z| > 2.07) = 2 * (0.0192) = 0.0384$.

χ-squared test:

Our test statistic is

$$\chi^2 \;=\; \frac{(90 - 79.61)^2}{79.61} + \frac{(136 - 146.39)^2}{146.39} + \frac{(65 - 75.39)^2}{75.39} + \frac{(149 - 138.61)^2}{138.61} \;=\; 4.3042.$$

Consulting Table F (with $df = 1$), we see that

$$0.025 < P(\chi^2 \geq 4.3042) < 0.05.$$

That is, the P-value is between 0.025 and 0.05. (It's actually equal to 0.0384, as we found more accurately using the other method.)