Case C→C: The Idea of the Chi-Square Test

Learning Objective: In a given context, carry out the appropriate inferential method for comparing relationships and draw the appropriate conclusions.

Learning Objective: Specify the null and alternative hypotheses for comparing relationships.

The Chi-Square Test for Independence

The chi-square test for independence examines our observed data and tells us whether we have enough evidence to conclude beyond a reasonable doubt that two categorical variables are related. Much like the previous part on the ANOVA F-test, we are going to introduce the hypotheses (step 1), and then discuss the idea behind the test, which will naturally lead to the test statistic (step 2). Let's start.

Step 1: Stating the hypotheses

Unlike all the previous tests that we presented, the null and alternative hypotheses in the chi-square test are stated in words rather than in terms of population parameters. They are:

H_o: There is no relationship between the two categorical variables. (They are independent.)

H_a: There is a relationship between the two categorical variables. (They are not independent.)

Example

In our example, the null and alternative hypotheses would then state:

H_o: There is no relationship between gender and drunk driving.

H_a: There is a relationship between gender and drunk driving.

Or equivalently,

H_o: Drunk driving and gender are independent

H_a: Drunk driving and gender are not independent

and hence the name "chi-square test for independence."

Comment

Algebraically, independence between gender and driving drunk is equivalent to having equal proportions who drank (or did not drink) for males vs. females. In fact, the null and alternative hypotheses could have been re-formulated as

H_o: proportion of male drunk drivers = proportion of female drunk drivers

H_a: proportion of male drunk drivers ≠ proportion of female drunk drivers

However, expressing the hypotheses in terms of proportions works well and is quite intuitive for two-by-two tables, but the formulation becomes very cumbersome when at least one of the variables has several possible values, not just two. We are therefore going to always stick with the "wordy" form of the hypotheses presented in step 1 above.

The Idea of the Chi-Square Test

The idea behind the chi-square test, much like previous tests that we've introduced, is to measure how far the data are from what is claimed in the null hypothesis. The further the data are from the null hypothesis, the more evidence the data presents against it. We'll use our data to develop this idea. Our data are represented by the observed counts:

	Drank Alcohol in Last 2 Hours?		
Gender	Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Observed counts

How will we represent the null hypothesis?

In the previous tests we introduced, the null hypothesis was represented by the null value. Here there is not really a null value, but rather a claim that the two categorical variables (drunk driving and gender, in this case) are independent.

To represent the null hypothesis, we will calculate another set of counts — the counts that we would expect to see (instead of the observed ones) if drunk driving and gender were really independent (i.e., if H_o were true). For example, we actually observed 77 males who drove drunk; if drunk driving and gender were indeed independent (if H_o were true), how many male drunk drivers would we expect to see instead of 77? Similarly, we can ask the same kind of question about (and calculate) the other three cells in our table.

In other words, we will have two sets of counts:

- the observed counts (the data)
- the expected counts (if H_o were true)

We will measure how far the observed counts are from the expected ones. Ultimately, we will base our decision on the size of the discrepancy between what we observed and what we would expect to observe if H_o were true.

How are the expected counts calculated? Once again, we are in need of probability results. Recall from the probability section that if events A and B are independent, then P(A and B) = P(A) * P(B). We use this rule for calculating expected counts, one cell at a time.

Here again are the observed counts:

Gender ↓	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Applying the rule to the first (top left) cell, if driving drunk and gender were independent then:

P(drunk and male) = P(drunk) * P(male)

By dividing the counts in our table, we see that:

P(Drunk) = 93 / 619 and

P(Male) = 481 / 619,

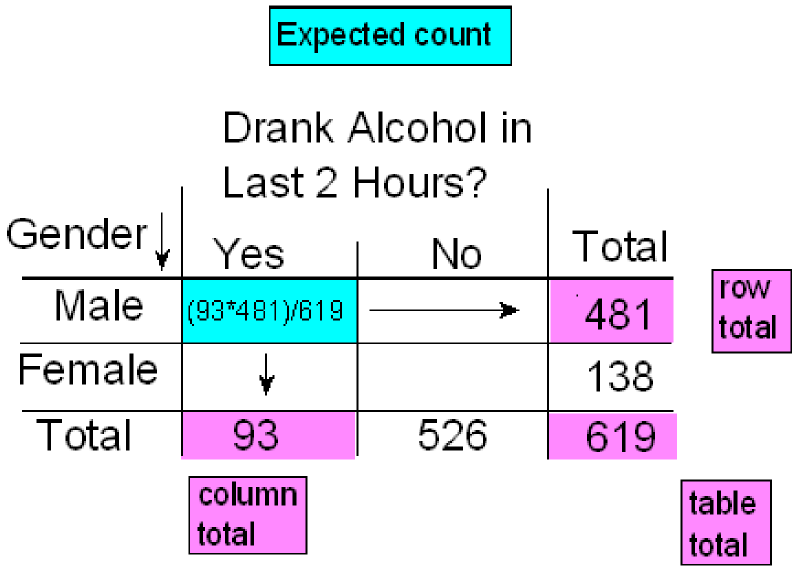
and so,

P(Drunk and Male) = (93 / 619) (481 / 619)

Therefore, since there are total of 619 drivers, **if drunk driving and gender were independent**, the **count** of drunk male drivers that I would **expect** to see is:

$$619 * P\left(Drunk \;\; and \;\; Male\right) = 619\left(\frac{93}{619}\right)\left(\frac{481}{619}\right) = \frac{93*481}{619}$$

Notice that this expression is the product of the column and row totals for that particular cell, divided by the overall table total.



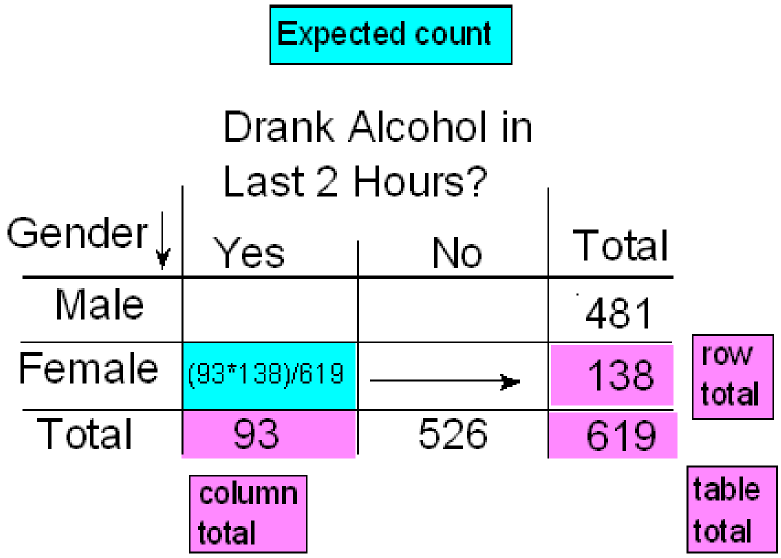
Similarly, if the variables are independent,

P(Drunk and Female) = P(Drunk) * P(Female) = (93 / 619) (138 / 619)

and the expected count of females driving drunk would be

$$\left(\frac{93}{619}\right)\left(\frac{138}{619}\right) = \frac{93*138}{619}$$

Again, the expected count equals the product of the corresponding column and row totals, divided by the overall table total:



This will always be the case, and will help streamline our calculations:

$$Expected\ \ Count = \frac{Column\ \ Total*Row\ \ Total}{Table\ \ Total}$$

Learn By Doing

1/1 point (graded)

The expected count of males not driving drunk is:

- (526 * 138) / 619
- (93 * 481) / 619
- (93 * 138) / 619
- (526 * 481) / 619 ✔

Answer

Correct: Indeed 526 is the column total for this cell, and 481 is the row total.

Submit

Learn By Doing

1/1 point (graded)

The expected count of females not driving drunk is:

- (526 * 138) / 619 ✔
- (93 * 481) / 619
- (93 * 138) / 619
- (526 * 481) / 619

Answer
Correct: Indeed, 526 is the column total for this cell, and 138 is the row total.

Submit

Here is the complete table of expected counts, followed by the table of observed counts:

Expected Counts

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	(93*481)/619=72.3	(526*481)/619=408.7	481
Female	(93*138)/619=20.7	(526*138)/619=117.3	138
Total	93	526	619

Observed Counts

Gender	Drank Alcohol in Last 2 Hours? Yes	No	Total
Male	77	404	481
Female	16	122	138
Total	93	526	619

Scenario: Gender and Ear Piercing

A study was done on the relationship between gender and piercing among high-school students. A sample of 1,000 students was chosen, then classified according to gender and according to whether or not they had any of their ears pierced. The results of the study are summarized in the following 2-by-2 table:

	Piercing?		
Gender	Yes	No	Total
Female	576	64	640
Male	72	288	360
Total	648	352	1000

Did I Get This

1/1 point (graded)

What is the expected count of non-pierced females?

- 414.72
- 126.72
- 225.28 ✔
- 3,520
- 64

Answer

Correct: Indeed, the expected count of non-pierced females is (640 * 352) / 1,000 = 225.28

Submit

Did I Get This

1/1 point (graded)

The expected count that you found in the question above is the number of non-pierced females if which of the following is true? Check all that apply.

- [ ] Gender and piercing were significantly related.
- [x] Gender and piercing were independent.
- [x] The null hypothesis were true.
- [ ] The null hypothesis were rejected.

Answer
Correct:
The expected count is the number of non-pierced females if the null hypothesis were true and the null hypothesis were true.

Submit

We see that there are differences between the observed and expected counts in the respective cells. We now have to come up with a measure that will quantify these differences. This is the chi-square test statistic.
