Inference for Two-way Tables

- Two-way table for categorical dataset;
- Chi-square test for two-way table;
- Chi-square goodness of fit test;

Example: Background music and consumer behavior

In a study conducted in a Northern Ireland supermarket, researchers counted the number of bottles of French, Italian, and other wine purchased while shoppers were subject to one of three "treatments": no music, French accordion music, and Italian string music.

The following **two-way table** summarizes the data:

	Music			
Wine	None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Example (cont.)

From the table of counts, we can ascertain the (empirical) joint distribution, marginal distributions, and conditional distributions of wine type and music type:

	Music			
Wine	None	French	Italian	Total
French	0.123	0.160	0.123	0.407
Italian	0.045	0.004	0.078	0.128
Other	0.177	0.144	0.144	0.465
Total	0.346	0.309	0.346	1.000

We are interested in determining whether there is relationship between the row variable (wine type) and the column variable (music type).

If this were the *true distribution*, then the answer would be clear: music and wine are not independent, so there *is* a relationship.

However, this table is *random*, and we want to know whether or not music and wine are independent *under the true distribution*. This requires a statistical test.

The χ^2-test for a $r \times c$ Table

Hypotheses

- H_0: there is *no association* between the row variable and the column variable.
- H_a: there is an association between the two variables.

The alternative hypothesis H_a does not specify any particular direction of the association because. It includes all of the many kinds of association that are possible and as a result, we cannot describe H_a as either one-sided or two-sided.

Intuition for the Test

Suppose H_0 is true, and the two variables are independent. What counts would we expect to observe?

Recall that under the independence assumption,

$$P(A \text{ and } B) = P(A)P(B)$$

Wine	Music None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

Let $n = 243$ be the total number of bottles of wine that is sold. Assuming independence, the expected number of French wine that is sold while French Music is being played is estimated to be:

$$n \times P(\text{Type of Wine is French and Music is French})$$

$$= nP(\text{Type of Wine is French})P(\text{Music is French})$$

$$= 243\left(\frac{99}{243}\right)\left(\frac{75}{243}\right) = \frac{(99)(75)}{243} = 30.555$$

What is the observed number?

What is the observed number and the expected number of Italian wine that is sold while French Music is being played (under H_0)?

$$n \times P(\text{Type of Wine is Italian and Music is French})$$

$$= nP(\text{Type of Wine is Italian})P(\text{Music is French})$$

$$= 243\left(\frac{31}{243}\right)\left(\frac{75}{243}\right) = \frac{(31)(75)}{243} = 9.567$$

Thus, for each cell, we have

$$\text{Expected Cell Count} = \frac{\text{row total} \times \text{col total}}{\text{total count}}$$

Our test will be based on a measure of *how far the observed table is from the expected table.*

Example (cont.)

For the supermarket example, the expected counts are:

Wine	Music None	French	Italian	Total
French	34.22	30.56	34.22	99
Italian	10.72	9.57	10.72	31
Other	39.06	34.88	39.06	113
Total	84	75	84	243

The observed counts are:

Wine	Music None	French	Italian	Total
French	30	39	30	99
Italian	11	1	19	31
Other	43	35	35	113
Total	84	75	84	243

The X^2 (Chi-Squared) Statistic

To measure how far this *expected* table is from the *observed* table, we will use the following test statistic:

$$X^2 = \sum_{\text{all cells}} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

The χ^2 Distribution

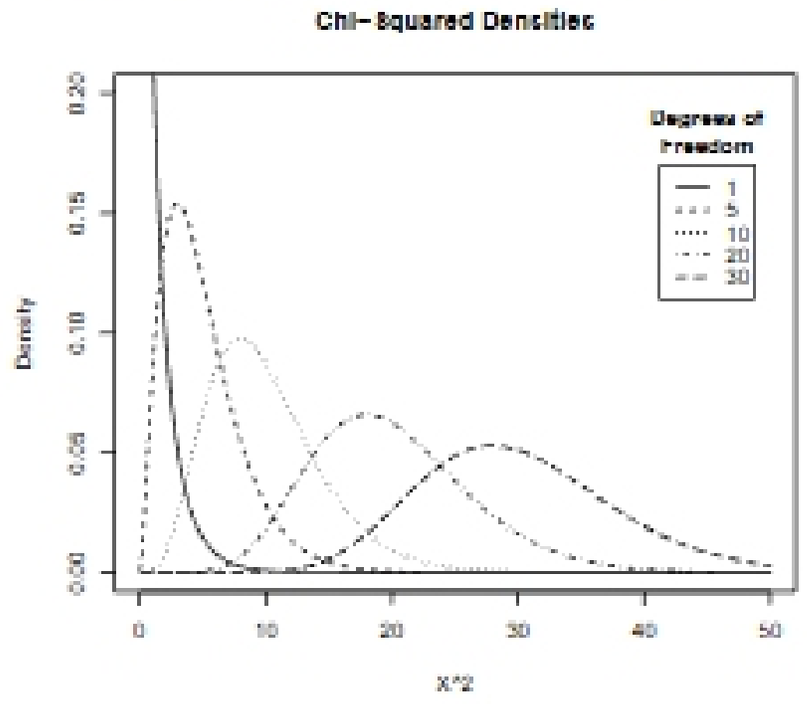
Under H_0, the X^2 test statistic has an approximate χ^2 distribution with $(r-1)(c-1)$ degrees of freedom, denoted $\chi^2_{(r-1)(c-1)}$.

Why $(r-1)(c-1)$?

Recall that our "expected" table is based on some quantities estimated from the data: namely the row and column totals.

Once these totals are known, filling in any $(r-1)(c-1)$ undetermined table entries actually gives us the whole table. Thus, there are only $(r-1)(c-1)$ freely varying quantities in the table.

What does the χ^2 distribution look like?



- Unlike the Normal or t distributions, the χ^2 distribution takes values in $(0, \infty)$.
- As with the t distribution, the exact shape of the χ^2 distribution depends on its degrees of freedom.

p-Value for the χ^2-Test

If the observed and expected counts are very different, X^2 will be large, indicating evidence against H_0. Thus, the p-value is always based on the right-hand tail of the distribution.

There is no notion of a two-tailed test in this context.

The p-value is therefore

$$P(\chi^2_{(r-1)(c-1)} \geq X^2)$$

Recall that X^2 has an *approximate* $\chi^2_{(r-1)(c-1)}$ distribution. When is the approximation valid?

For any two-way table larger than 2×2, we require that the average expected cell count is at least 5 and each expected count is at least one.

For 2×2 tables, we require that each expected count be at least 5.

Example (cont.)

Let's get back to our example...

Recall the observed and expected counts:

	Observed			Expected			
	Music			Music			
Wine	None	Fr.	It.	None	Fr.	It.	Tot.
French	30	39	30	34.22	30.56	34.22	99
Italian	11	1	19	10.72	9.57	10.72	31
Other	43	35	35	39.06	34.88	39.06	113
Total	84	75	84	84	75	84	243

$$X^2 = \frac{(30-34.22)^2}{34.22} + \frac{(39-30.56)^2}{30.56} + \frac{(30-34.22)^2}{34.22} + \cdots + \frac{(35-34.88)^2}{34.88} + \frac{(35-39.06)^2}{39.06} = 18.28$$

The table is 3×3, so there are $(r-1)(c-1) = 2 \times 2 = 4$ degrees of freedom.

Finally, the p-value is found from the χ^2 distribution table with 4 degrees of freedom:

$$0.001 \leq P(\chi^2_4 \geq 18.28) \leq 0.0025$$