34 2025-04-25 | Week 14 | Lecture 34

The example in this lecture is based on a very nice chi-squared tutorial by Caitlin Light, available here: `https://www.ling.upenn.edu/~clight/chisquared.htm`

34.1 Tests of Association, Revisted

We now have the things we need to test whether two things are independent or not.

Definition 97. A *contingency table* is a matrix of with rows labeled as group 1, group 2, and so forth, and columns labeled as outcome 1, outcome 2, etc. The entries of the matrix are COUNTS: the (i, j)-th entry is the number of times that group i experience outcome j.

For example, puppose we have a population of $n = 54$ students, which we divide up according two to factors

(i) whether they attended class regularly or not

(ii) whether they pass the class exam or not

Then our contingency table is

	Pass	**Fail**
Attended	25	6
Skipped	8	15

(35)

This is the *table of observed values*. We usually augment the table with "marginal totals", like so:

	Pass	**Fail**	**Total**
Attended	25	6	31
Skipped	8	15	23
Total	33	21	54

The null hypothesis is

$$H_0 = \text{ attendence and passing are independent}$$

and the alternative hypothesis is

$$H_1 = \text{ there is a relationship between attendence and passing.}$$

Let p denote the probability that a student regularly attends class, and let q denote the probability that a student passes the final exam.
Important: We do not know the true value of p and q. But we can estimate them using the marginal row and column sums from the table:

$$p \approx \frac{31}{54} \quad \text{and} \quad q \approx \frac{33}{54}$$

Question: What are the expected values of the table, assuming the null hypothesis is true?
Under the null hypothesis, we can multiply probabilities, e.g., so that the probability that a student both regularly attends class AND passes the exam is about pq. Therefore, since there are n students,

$$\begin{aligned}
&\mathbb{E}\left[\#\text{ of students who attend regularly AND pass the exam}\right] \\
&= n\mathbb{P}\left[\text{a randomly-selected student attends regularly AND pass the exam}\right] \\
&= n\mathbb{P}\left[\text{attends regularly}\right]\mathbb{P}\left[\text{passes the exam}\right] \\
&= npq
\end{aligned}$$

Similar calculations gives us the following table of expected counts:

	Pass	**Fail**
Attended	npq	$np(1-q)$
Skipped	$n(1-p)q$	$n(1-p)(1-q)$

and plugging in the values $n = 56$ and our approximations $p \approx \frac{31}{54}$ and $q \approx \frac{33}{54}$ gives

	Pass	**Fail**
Attended	18.9	12.1
Skipped	14.1	8.9

(36)

This is the *table of estimated expected values.* Using the values of Equations (35) and (36), we compute a chi-squared statistic:

$$\begin{aligned} \chi^2 &= \sum_{\text{all table entries}} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \\ &= \frac{(25-18.9)^2}{20.2} + \frac{(6-12.1)^2}{12.1} + \frac{(8-14.1)^2}{14.1} + \frac{(15-8.9)^2}{8.9} \\ &= 11.7 \end{aligned}$$

By Theorem 96, we know that χ^2 has a chi-squared distribution. The degrees of freedom thing isn't so clear from that theorem, but when working with a contingency table, the degrees of freedom is given by the equation

$$(\text{number of rows} - 1) \times (\text{number of columns} - 1)$$

which in our case, for a 2×2 table, is just 1.

Thereore, using Definition 94, the p-value is

$$\begin{aligned} p\text{-value} &= \int_{11.7}^{\infty} \frac{1}{\sqrt{2\pi}} x^{-\frac{1}{2}} e^{-x/2} dx \\ &= 0.0006 \end{aligned}$$

In other words, if attending and passing were independent, we would see results this extreme only 0.06% of the time. This is very small, so we reject the null hypothesis. Our conclusion is that attendance and passing the final exam are **not independent**, i.e., that there is a relationship between them.

At this point we are tempted to conclude that regularly attending class increases your chance of passing the final exam. This seems obvious and is certainly consistent with the data. But it's not a conclusion that follows from the test that we did. The test only told us that these two things things probably have some relationship; it doesn't give us information about the nature of that relationship.