31 2026-04-22 | Week 14 | Lecture 31

31.1 Tests of independence

Based on chapters 14.2-14.4

Suppose we want to know if two things A and B are related to each other, or whether they are independent. For example, whether or not a person vapes (A) and whether they get coronary artery disease (B).

Or: whether a student attends class regularly (A) and whether they pass the class (B).

Question: How can we test this (using the hypothesis testing framework)?

Answer: We can utilize what are called "tests of independence", which we'll present in this lecture. For such tests, we require two main ingredients.

31.1.1 Ingedient 1: The chi-squared distribution

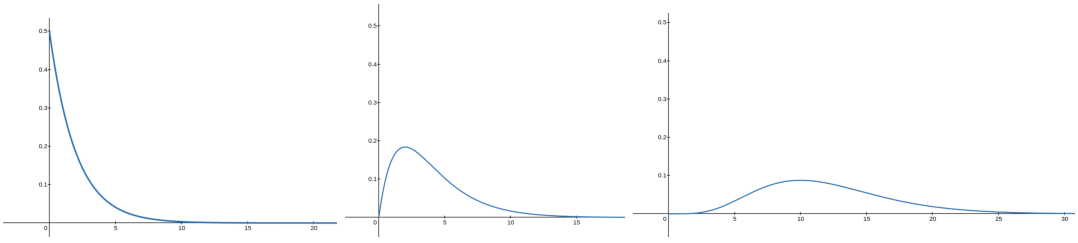
Definition 137 (Chi-Squared Distribution)**.** Let k be a positive integer. A nonnegative random variable X is said to have ***chi-squared distribution*** with d ***degrees of freedom*** if it has pdf

$$f(x) = \begin{cases} \frac{1}{2^{d/2}\Gamma(d/2)} x^{\frac{d}{2}-1} e^{-x/2} & : \quad x \geq 0 \\ 0 & : \quad x < 0 \end{cases}$$

.

Here, $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dx$. Note $\Gamma(k) = (k-1)!$ if k is a positive integer.

From left-to-right, here's what $f(x)$ looks like for $d = 2$, $d = 4$, and $d = 12$:



Proposition 138. *A χ^2 random variable X with d degrees of freedom has*

$$\mathbb{E}[X] = \alpha\beta = d \quad \textit{and} \quad \mathrm{Var}(X) = \alpha\beta^2 = 2d$$

The chi-squared distribution plays a central role in statistical inference. It is completely determined by its degrees of freedom.

31.1.2 Ingedient 2: The chi-squared statistic

Definition 139 (Chi-squared statistic)**.** Suppose we repeatedly conduct an experiment with k possible outcomes. The ***chi-squared statistic*** is defined as

$$\chi^2 := \sum_{i=1}^{k} \frac{(N_i - E_i)^2}{E_i}, \tag{21}$$

where $E_i = $ (expected # of times outcome i occurs), and $N_i = $ (# of times outcome i is observed).

Remark 140. The right-hand side of Eq. (28) has the form

$$\sum_{i=1}^{k} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Remark 141. Under general conditions, the chi-squared statistic χ^2 has approximately a chi-squared distribution with degrees of freedom determined by the setting.

31.1.3 Example of a test of independence

Definition 142. 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.

Example 143 (Test of independence)**.** For example, suppose 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

(22)

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

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

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?
If attendence and passing are indpendent, then

$$\mathbb{E}\left[\#\text{ of students who attend regularly AND pass}\right] = npq = (54)\left(\frac{31}{54}\right)\left(\frac{33}{54}\right) = 18.9$$

This is what we expect if the null hypothesis is true. Compare this with the actual observed value, which was 25. The idea of the test, which we will finish in the next lecture, is the following:

If the observed values deviate a lot from what we'd expect under H_0, we should reject H_0.

Quantifying this will require computing a chi-squared statistic, which we'll do in the next lecture.

End of Example 143. □