# Inference for Categorical Data: Chi-Square

AP Statistics

## Are My Results Unexpected?

When data are **counts** spread across several categories, we test whether the observed counts differ from what a claim predicts. The tool is the **chi-square** 卡方 ($\chi^2$) statistic, which adds up the standardized gaps between observed and expected counts:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

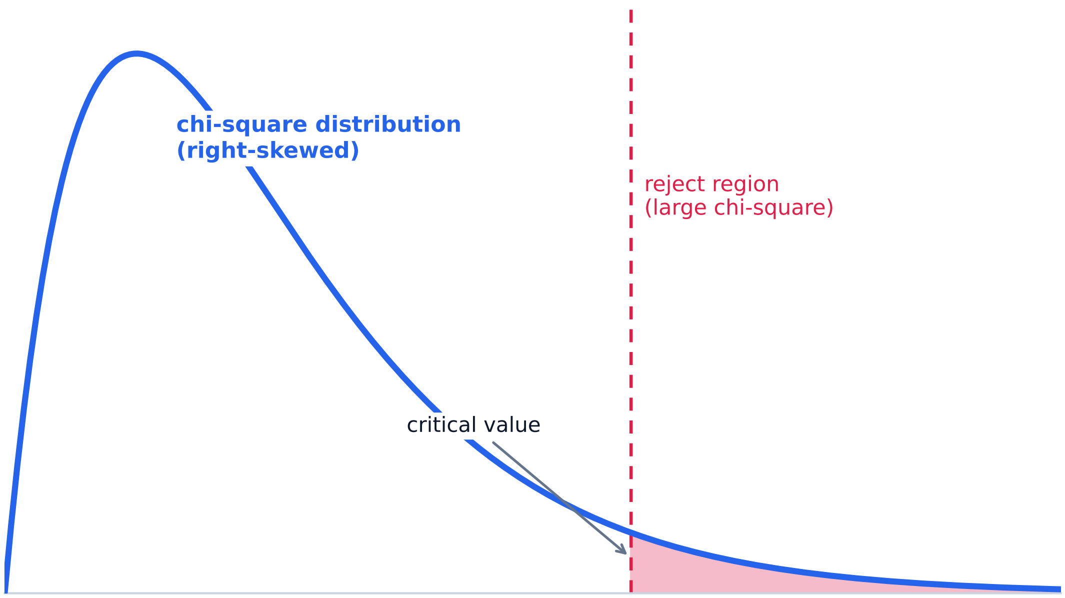
A large $\chi^2$ means the observed counts are far from expected –evidence against the claim. The **chi-square distribution** is right-skewed and depends on its **degrees of freedom** 自由度.

## Setting Up a Goodness-of-Fit Test

A **goodness-of-fit (GOF)** 拟合优度 test checks whether one categorical variable follows a **claimed distribution** (e.g. ”the die is fair”). Hypotheses:

$$H_0 : \text{the distribution is as claimed} \qquad H_a : \text{at least one proportion differs.}$$

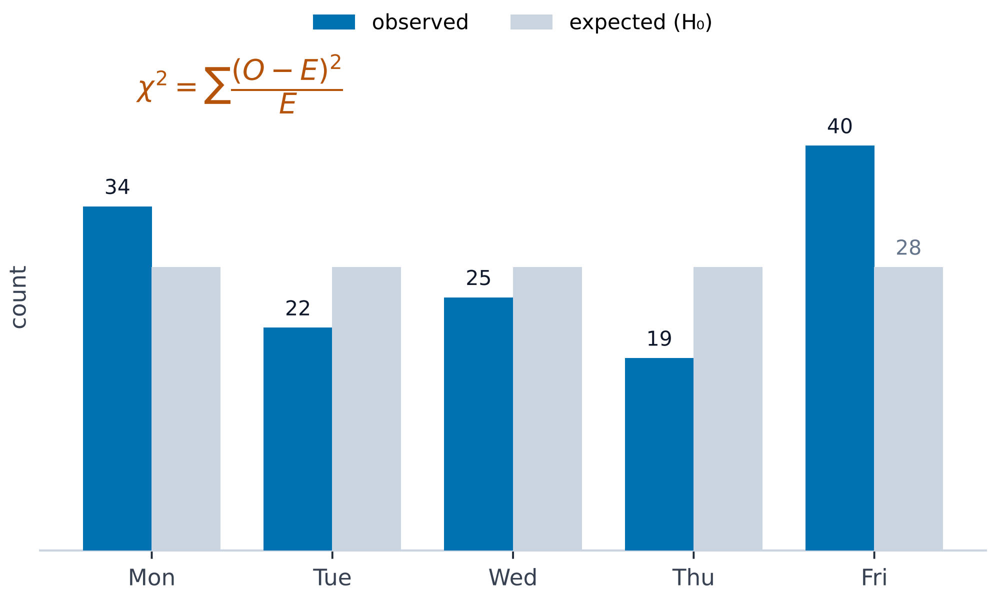
**Expected count** for each category $= n \times$ (claimed proportion). Conditions: random sample, all **expected counts** $\geq 5$, and the 10% condition.




*The chi-square distribution is right-skewed. A large statistic lands in the shaded right tail past the critical value –that is where you reject the model.*

# Carrying Out a Goodness-of-Fit Test

Compute $\chi^2 = \sum \frac{(O-E)^2}{E}$ with $df$ = (number of categories) − 1. Find the $p$-value from the chi-square distribution (upper tail), compare to $\alpha$, and conclude in context. A large component of the sum points to the category that deviates most.




*Chi-square compares observed counts with those expected under the null hypothesis*

**Worked example.** A die rolled 60 times gives counts $8, 10, 12, 9, 11, 10$. If it is fair, each expected count is $60/6 = 10$, so

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(11-10)^2}{10} = 0.4 + 0.4 + 0.1 + 0.1 = 1.0,$$

with $df = 5$. This is small (a large $p$-value), so we **fail to reject** $H_0$ –no evidence the die is unfair.

# Expected Counts in Two-Way Tables

For a two-way table, the **expected count** in a cell (under ”no association”) is

$$E = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}.$$

This is the count you would see if the row and column variables were unrelated.

**Worked example.** In a two-way table a cell's row total is 40, its column total is 50, and the grand total is 200. Its expected count is $E = \frac{40 \times 50}{200} = 10$. Repeating for every cell gives the expected table to compare against the observed one.

## Homogeneity or Independence?

Two tests use the same $\chi^2$ math but answer different questions:

- **Test for homogeneity** 同质性: are the distributions of one categorical variable the **same across several populations or groups** (separate samples/treatments)?
- **Test for independence** 独立性: are two categorical variables **associated within a single population** (one sample, two variables measured)?

The design (several samples vs one sample) decides which name and hypotheses to use.

## Carrying Out a Test for Homogeneity or Independence

Compute expected counts, then $\chi^2 = \sum \frac{(O-E)^2}{E}$ over all cells, with

$$df = (\text{rows} - 1)(\text{columns} - 1).$$

Conditions: random data, all expected counts $\geq 5$, 10% condition. Find the $p$-value, compare to $\alpha$, and conclude in context –evidence of a difference between groups (homogeneity) or of an association (independence).

## Choosing the Right Categorical Procedure

Decide by the setup: one categorical variable against a claimed distribution ⇒ **goodness-of-fit**; one sample cross-classified by two variables ⇒ **independence**; several samples/groups compared ⇒ **homogeneity**. Comparing just **two** proportions can use either a two-proportion $z$-test or a chi-square test –they agree.

## Exam tips

- Use $\chi^2 = \sum \frac{(O-E)^2}{E}$ for **categorical** data; always divide by the **expected** count.
- Pick the right test: goodness-of-fit (one variable), independence, or homogeneity (two-way table).
- Compute expected counts as $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$ and check each is $\geq 5$.
- A large $\chi^2$ (small p-value) means observed counts differ from expected by more than chance.
- State degrees of freedom correctly (categories $-1$, or $(r-1)(c-1)$).