From Counts to Conclusions:

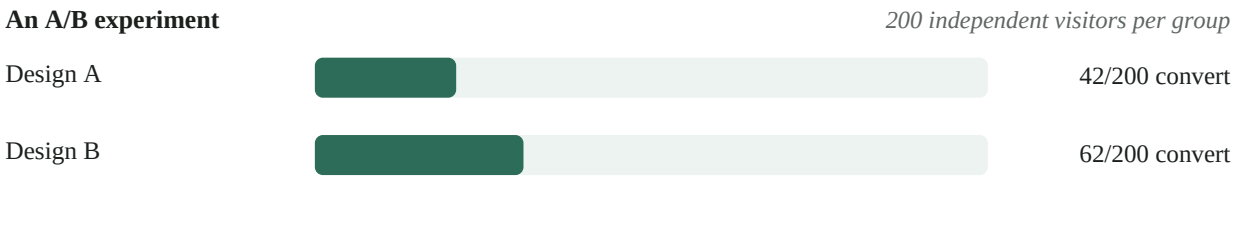
Hypothesis Tests in Practice

BAMS course insight - Mathematical Statistics

Abstract.

An online platform compares two versions of its sign-up page. A survey asks whether preferences follow a proposed distribution. A bank observes different default frequencies across borrower groups. In each case, we see differences in data: how do we decide which are compatible with random sampling, and what exactly does a statistical test tell us?

This project develops **hypothesis testing for categorical and numerical data** through concrete problems. We begin with binomial and multinomial models, derive Pearson's chi-square tests for goodness-of-fit and independence, and establish their connection to tests for two proportions. We then examine tests for means, assess their assumptions and compare the conclusions obtained from significance tests, confidence intervals and effect sizes.



A binary-outcome experiment: two observed rates can differ even if the underlying conversion probabilities are equal.

Prerequisites / Difficulty.

Completed Mathematical Statistics: point estimation, confidence intervals, central limit theorem and basic hypothesis testing. Elementary probability and Python or R for numerical examples.

Accessible to intermediate. The common project uses familiar distributions and simple contingency tables. Exact tests, resampling and multiple comparisons are optional extensions.

Starting materials.

A. Agresti (2018), *An Introduction to Categorical Data Analysis*, 3rd ed. Begin with binomial inference and two-way contingency tables.

M. Çetinkaya-Rundel and J. Hardin, *Introduction to Modern Statistics*, 2nd ed. Use the chapters on inference and categorical data for worked applications.

R. L. Wasserstein and N. A. Lazar (2016), *The ASA Statement on p-Values: Context, Process, and Purpose*. For interpreting statistical evidence.

What to do.

The common goal is to develop a **small family of hypothesis tests from their sampling models**, then apply them to realistic data questions. Each test is derived, used and evaluated rather than treated as a formula to memorize.

1. Start with the experiment. Describe an A/B conversion experiment, a one-variable categorical survey and a comparison of a numerical outcome across groups. For each case, specify the observations, sampling assumptions, parameter of interest, and null and alternative hypotheses. Distinguish independent samples from paired measurements.

2. Derive Pearson's chi-square statistic. For a categorical distribution, calculate the expected cell counts under the null and study $X^2 = \Sigma\,(O - E)^2/E$. Apply it first to goodness-of-fit and then to independence in a contingency table. Derive the expected counts from row and column totals, identify the degrees of freedom, and explain when the large-sample approximation may be unreliable.

3. Connect chi-square tests and proportions. In a 2×2 table with independent binomial samples, derive the pooled two-proportion z statistic for equality of probabilities. Show that its square equals the **uncorrected Pearson chi-square statistic**. Explain the connection between the resulting two-sided tests and why directional alternatives need separate treatment.

4. Compare numerical outcomes. Use a two-sample Welch t test for independent numerical measurements and a paired t test for repeated observations. Explain why these designs require different standard errors and why a test for proportions is more appropriate for binary data than blindly substituting a t test.

5. Investigate what the tests report. Simulate experiments under null and alternative models. Examine rejection rates, power, effects of sample size and the consequences of sparse categorical counts. Report estimates, confidence intervals and effect sizes alongside p-values. Discuss what statistical significance does not establish about practical importance or causation.

Further directions.

- **Small samples.** Compare a chi-square approximation with Fisher's exact test in a 2×2 table, under its sampling conditions.
- **Randomization tests.** Use label permutations to test an effect in a randomized experiment.
- **Many hypotheses.** Investigate false positives across repeated tests and compare simple multiple-testing corrections.

Expected outcome.

An illustrated guide with worked examples from A/B testing, surveys and observational comparisons; derivations linking the principal test statistics; and a reproducible set of simulations showing the role of sampling design, sample size and assumptions. Different case studies can be developed in parallel by group members.