Hypothesis Testing — Swiftie

September 20, 2025

Big Picture

Think of hypothesis testing like this: you make a **claim** about Taylor Swift–world (streams, ticket sales, merch choices), then you check a **sample** of data to see if the claim holds up or if what you saw could just be luck. The same logic powers day-to-day decisions in **machine learning** (credit risk modeling, product A/B tests, model monitoring, feature selection).

1 Core Concepts (with Swift & ML intuition)

Null hypothesis (H_0)

- **Meaning:** The "nothing new here" claim; a concrete baseline.
- **Swift example:** After *The Eras Tour* film, the average daily streams stayed the same as before. For example, $H_0 : \mu = 100$ streams/user/day.
- **ML / credit risk example:** A new scorecard does not change the default rate. $H_0 : p_{\text{default,new}} = p_{\text{default,old}}$.

Alternative hypothesis (H_1 or H_a)

- **Meaning:** The competing claim you want evidence for.
- **Swift example:** Average daily streams increased ($\mu > 100$) or changed ($\mu \neq 100$).
- **ML / credit risk example:** The new model reduces default rate ($p_{\text{new}} < p_{\text{old}}$).

Significance level (α)

- **Meaning:** False-alarm tolerance; risk of claiming a change when there isn't one (Type I error).
- **Common choice:** $\alpha = 0.05$ (5%).
- **Swift example:** Willing to be wrong 5% of the time when declaring streams went up.
- **ML / credit risk example:** Require $p \leq 0.01$ before shipping a new underwriting rule because errors are costly.

p-value

- **Meaning (plain):** If the **null were true**, how surprising is the data you observed? Smaller p = more surprising (i.e., stronger evidence against H_0).
- **Decision rule:** If $p \le \alpha \Rightarrow$ reject H_0 (evidence favors H_1). If $p > \alpha \Rightarrow$ do not reject H_0.
- **Swift example:** $p = 0.003$ for "streams increased" $\Rightarrow$ strong evidence of a boost.
- **ML / credit risk example:** $p = 0.18$ for "default rate dropped" $\Rightarrow$ not enough evidence to change policy yet.

Misconception watch: p is *not* "the probability H_0 is true." It's about the data's surprise level *given* H_0.

Type I & Type II errors; Power

- **Type I (false positive):** Say "something changed" when it didn't. Chance $\approx \alpha$.
- **Type II (false negative):** Miss a real change (fail to reject H_0 when H_1 is true).
- **Power** $(1 - \beta)$: Chance to detect a true effect. Bigger samples / bigger effects $\Rightarrow$ higher power.
- **Swift example:** Too-small fan sample may miss a real lift in streams.
- **ML / credit risk example:** Under-powered test may miss a real drop in default rate after a new score threshold.

2 Z-Tests (means & proportions; large n or known σ)

2.1 One-sample z-test for a mean

Use when: Testing a mean with known population σ (or large n so the standard error is reliable).
Test statistic:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}.$$

Swift example (worked): Claim: "Average streams didn't change." $H_0 : \mu = 100$. After the film you sampled $n = 64$ users: sample mean $\bar{x} = 105$. Assume historical $\sigma = 10$.

$$z = \frac{105 - 100}{10/\sqrt{64}} = \frac{5}{1.25} = 4.0.$$

Two-sided $p \approx 0.000063 \Rightarrow$ reject H_0 at $\alpha = 0.05$. Strong evidence streams increased.
ML / credit risk uses:

- Did a new feature pipeline reduce the mean loss or mean processing time?
- Did a new model configuration change the average margin per approved customer?
- For small n or unknown $\sigma \Rightarrow$ use a **t-test** (same intuition; different reference distribution).[1]

[1] In practice, analysts often use a t-test for means because σ is rarely known.

2.2 One-sample z-test for a proportion

Use when: Testing a proportion with large enough n.
Test statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}.$$

Swift example (worked): H_0: "50% of fans click the pre-save link." You observe 560 of 1,000 did (56%).

$$z \approx \frac{0.56 - 0.50}{\sqrt{0.5 \cdot 0.5/1000}} \approx 3.795.$$

Two-sided $p \approx 0.00015 \Rightarrow$ reject H_0. Pre-save rate is higher than 50%.
ML / credit risk uses:

- Did the new model cut the approval default rate proportion?
- Did a new fraud rule change the flag rate?
- In online A/B tests, compare conversion or acceptance rates between variants.

Two-sample variants: Compare two means (A vs. B) or two proportions (control vs. treatment) to decide whether to ship a model or policy change.

3 Chi-Square (χ^2) Tests (counts & categories)

3.1 Goodness-of-fit (does a categorical variable match an expected pattern?)

Statistic:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \qquad \text{df} = \text{categories} - 1.$$

Swift example: Expect equal preference (25% each) for four eras in a poll: *Folklore, Midnights, 1989, Lover*. Out of 1,000 votes you saw: 260, 240, 270, 230. Expected each = 250.

$$\chi^2 = \tfrac{(260-250)^2}{250} + \tfrac{(240-250)^2}{250} + \tfrac{(270-250)^2}{250} + \tfrac{(230-250)^2}{250} = 0.4 + 0.4 + 1.6 + 1.6 = 4.0.$$

With df = 3, $p \approx 0.26 \Rightarrow$ do not reject H_0 (looks close to even).
ML / credit risk uses:

- Does the class distribution match expectations after re-sampling? (e.g., churn positives per bucket)
- Do PD deciles (expected defaults) align with observed defaults? (Basic calibration check; see also Hosmer–Lemeshow below.)

3.2 Independence test (are two categorical variables associated?)

Statistic: Same χ^2 formula using a contingency table; df = $(r - 1)(c - 1)$.

Swift example: Are *age group* (Under-25 vs 25+) and *song type* (Up-tempo vs Ballad) independent?

	Up-tempo	**Ballad**	**Row total**
Under-25	200	100	300
25+	150	150	300
Column total	**350**	**250**	**600**

Expected counts: $\begin{bmatrix} 175 & 125 \\ 175 & 125 \end{bmatrix}$. Then

$$\chi^2 \approx 17.14, \quad \text{df} = 1, \quad p < 0.0001.$$

Conclusion: Preference is associated with age group.

ML / credit risk uses:

- **Feature selection:** χ^2 test between categorical feature and target (default vs not) to keep features with signal (often summarized via Cramér's V).
- **Bias/fairness diagnostics:** Check if approval depends on protected group more than expected.
- **Monitoring:** Are bucketed feature distributions shifting by month? (time $\times$ category association)

For small counts or expected cells < 5, consider **Fisher's exact test** instead of χ^2.

4 The 5-Step Recipe (from question to decision)

1. **State H_0 and H_1.** "Average streams didn't change" vs "they increased."
2. **Choose α and a test.** z/t for means or proportions; χ^2 for counts.
3. **Compute the test statistic** (z or χ^2) from the sample.
4. **Get the p-value** and compare to α. $p \leq \alpha \Rightarrow$ reject H_0; $p > \alpha \Rightarrow$ don't reject.
5. **Interpret in plain English** and check **practical significance** and **assumptions**.

5 Assumptions & Practical Tips

- **Sampling & independence:** Random or representative samples; observations roughly independent.
- **Large-sample rules:** z-tests and χ^2 need sufficient counts. For means with small n and unknown σ, use **t-tests**.
- **Effect size vs significance:** A tiny p can hide a trivial effect. Always check the **magnitude** (e.g., +0.3% streams may be operationally moot).
- **Multiple testing:** If you try many songs/markets/hyper-parameters, adjust (e.g., Bonferroni or FDR) to control false discoveries.
- **Power planning:** Decide up front what effect size matters and compute required n.

6 Deeper ML / Credit-Risk Examples

Feature selection with χ^2 (classification)

Target Y = default (1/0). Candidate categorical feature X = employment sector (binned). Test independence of X vs Y. Keep features with small p; optionally rank strength with Cramér's V.

Comparing models (A/B) with two-sample proportion z-test

Goal: See if the default rate differs for accounts scored/approved by Model A vs Model B. $H_0 : p_A = p_B$, $H_1 : p_A \neq p_B$ (or one-sided if you only care about "B lower than A"). Use to decide whether to ship Model B. Also compare approval rates, fraud flag rates, conversion rates.

Calibration check (Hosmer–Lemeshow, χ^2-based)

Bin predicted PDs into deciles, compare expected vs observed defaults in each decile with a χ^2-type statistic. Use to detect mis-calibration (e.g., model predicts PD=10% in a bin but observed is 15%). *Swift analogy:* bin fans by "propensity to stream" and compare predicted vs observed streamers in each bin.

Monitoring drift post-deployment

Use χ^2 tests on categorical/binned features across time (Month 1 vs Month 6). A related companion metric is **Population Stability Index (PSI)** for tracking shifts across bins.

Policy/Risk threshold tuning

Z-test for proportions: When you raise the approval threshold, test if bad-rate (defaults/approvals) dropped without tanking approval rate. **Two-sample mean/t tests:** Check business KPIs (mean margin, loss given default) before vs after.

Fairness checks

χ^2 independence: Is approval independent of protected group after controlling for risk? Proportion z-tests: Compare error rates (e.g., false positives) across groups.

7 Worked Numbers (copy-paste ready)

Z for a mean (Swift streams)

$n = 64,\ \bar{x} = 105,\ \mu_0 = 100,\ \sigma = 10.$ $\quad z = \dfrac{105 - 100}{10/\sqrt{64}} = 4.0 \Rightarrow$ two-sided $p \approx 0.000063 \Rightarrow$ reject H_0.

Z for a proportion (Swift pre-save rate)

$n = 1000,\ \hat{p} = 0.56,\ p_0 = 0.50.$ $\quad z \approx 3.795 \Rightarrow$ two-sided $p \approx 0.00015 \Rightarrow$ reject H_0.

χ^2 independence (Age $\times$ Song type)

Observed $\begin{bmatrix} 200 & 100 \\ 150 & 150 \end{bmatrix}$; Expected $\begin{bmatrix} 175 & 125 \\ 175 & 125 \end{bmatrix}$. $\quad \chi^2 \approx 17.14$, df $= 1$, $p < 0.0001 \Rightarrow$ associated, not independent.

8 Quick Formula Sheet

$$\text{Z for a mean:} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

$$\text{Z for a proportion:} \quad z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

$$\text{Two-sample proportion z-test:} \quad z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\chi^2 \text{ (GOF/Independence):} \quad \chi^2 = \sum \frac{(O - E)^2}{E}, \quad \text{df} = \text{categories} - 1 \text{ (GOF) or } (r - 1)(c - 1) \text{ (independe}$$

Power ideas: Bigger n reduces standard error $\Rightarrow$ higher chance to detect real effects.

9 Picking the Right Test (cheat sheet)

- **Mean (big n or known σ) $\Rightarrow$ Z-test** (practically: often **t-test**).
  Swift: Did average listens per fan change?
- **Proportion $\Rightarrow$ Z-test for proportions**.
  Swift: Did pre-save rate exceed 50%? *Credit risk:* Did bad-rate drop with new model?
- **One categorical vs target pattern $\Rightarrow$ χ^2 goodness-of-fit**.
  Swift: Are era preferences evenly split?
- **Two categorical variables $\Rightarrow$ χ^2 independence** (or Fisher exact if small counts).
  Credit risk: Is default associated with employment band?

TL;DR

Use hypothesis tests to turn noisy data into decisions. Frame a clear H_0/H_1, pick the right test (z/χ^2), compute a p-value, and always read the result alongside **effect size**, **power**, and **assumptions**—whether you're checking if Swift streams jumped or if your new credit model truly cut default risk.