Hypothesis Testing

Errors, Power, Likelihood Ratios, and Multiple Comparisons

Hypothesis Testing Framework

Definition — Statistical Hypothesis Test

A **statistical hypothesis test** is a procedure for deciding between two competing claims about a population parameter θ.

Null Hypothesis

$$H_0 : \theta \in \Theta_0$$

The hypothesis to be tested — the status quo assumption.

Alternative Hypothesis

$$H_1 : \theta \in \Theta_1$$

The research hypothesis — what we want to detect.

Mathematical Requirements

- **Mutually exclusive:** $\Theta_0 \cap \Theta_1 = \emptyset$
- **Exhaustive:** $\Theta_0 \cup \Theta_1 = \Theta$ (the complete parameter space)

Types of Hypotheses

Simple hypothesis — specifies a single parameter value

$$H_0 : \theta = \theta_0$$

Composite hypothesis — specifies a range of values

$$H_0 : \theta \geq \theta_0$$

Two-sided

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \neq \theta_0$$

Deviations in **either** direction.

Right-tailed

$$H_0 : \theta \leq \theta_0$$
$$H_1 : \theta > \theta_0$$

Increases above θ_0.

Left-tailed

$$H_0 : \theta \geq \theta_0$$
$$H_1 : \theta < \theta_0$$

Decreases below θ_0.

Rejection Regions, Live

The alternative decides **where** the rejection region goes; α decides **how big** it is.

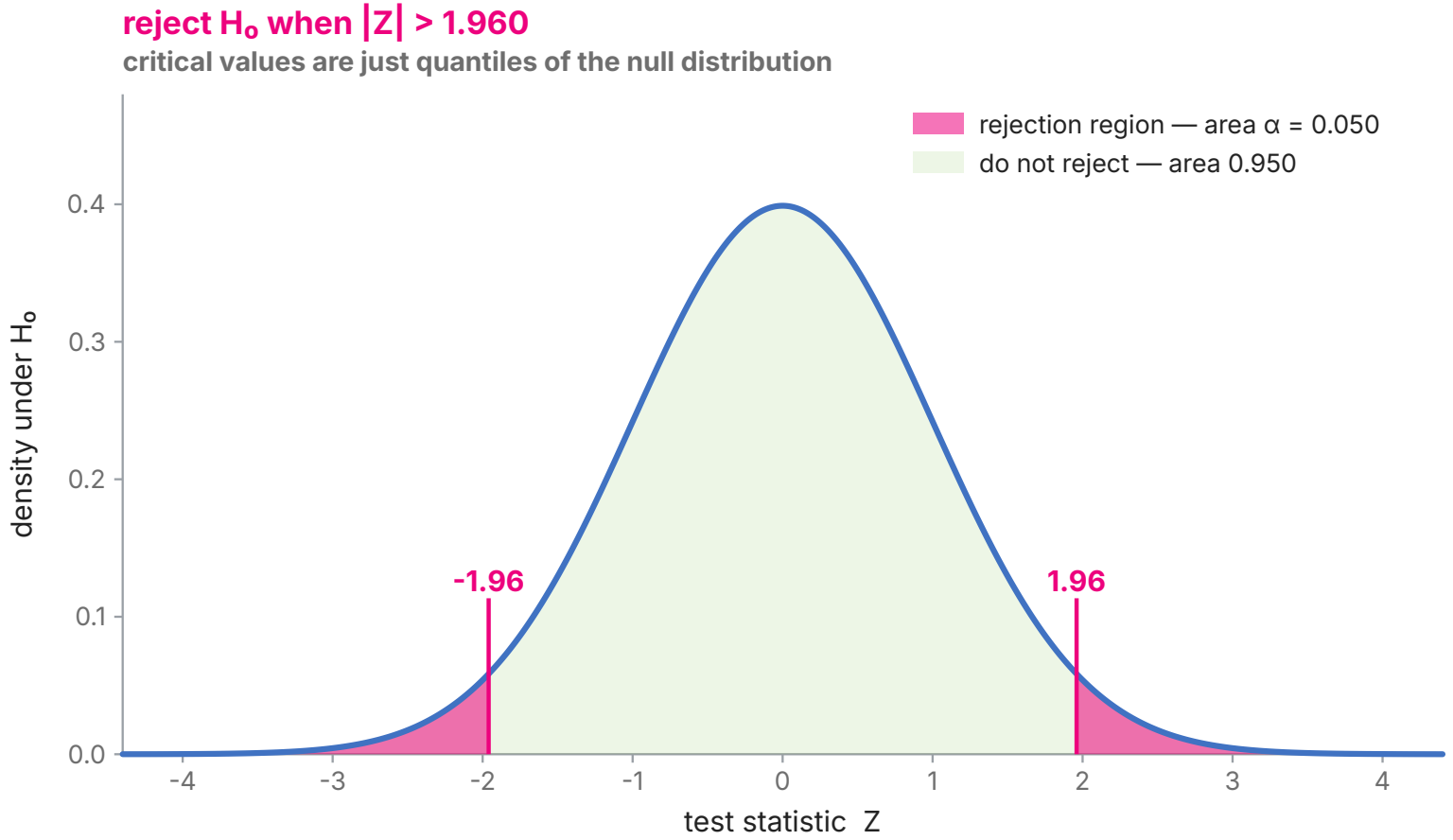
Alternative

Two-sided H_1: $\theta \neq \theta_0$

Significance α

0.05

The curve is the distribution of the test statistic **under** H_0. The shaded area *is* α — the probability of rejecting a true null.



Test Statistics and Decision Framework

Definition — Test Statistic

A **test statistic** $T(X)$ is a function of the sample data used to decide between H_0 and H_1.

Decision Rule

A test is defined by its **critical region** (rejection region) C:

$$\text{Reject } H_0 \text{ if } T(X) \in C$$

Critical Value

The boundary value c such that

$$P(T(X) > c \mid H_0) = \alpha$$

Significance Level

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true})$$

Common choices: 0.01, 0.05, 0.10.

Errors and Power

Type I and Type II Errors

Decision	H_0 true	H_0 false
Reject H_0	Type I error — probability α	Correct decision — power $1 - \beta$
Fail to reject H_0	Correct decision — probability $1 - \alpha$	Type II error — probability β

Error Definitions

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ true})$$

$$\beta = P(\text{Fail to reject } H_0 \mid H_1 \text{ true})$$

The Trade-off

- Decreasing α **increases** β — at fixed sample size
- Only a larger n reduces both at once

Two Errors, One Threshold, Live

One threshold splits both curves at once: you cannot shrink one shaded area without growing the other.

Alternative

Right-tailed

Effect size δ = $(\mu_1 - \mu_0)/\sigma$

0.5

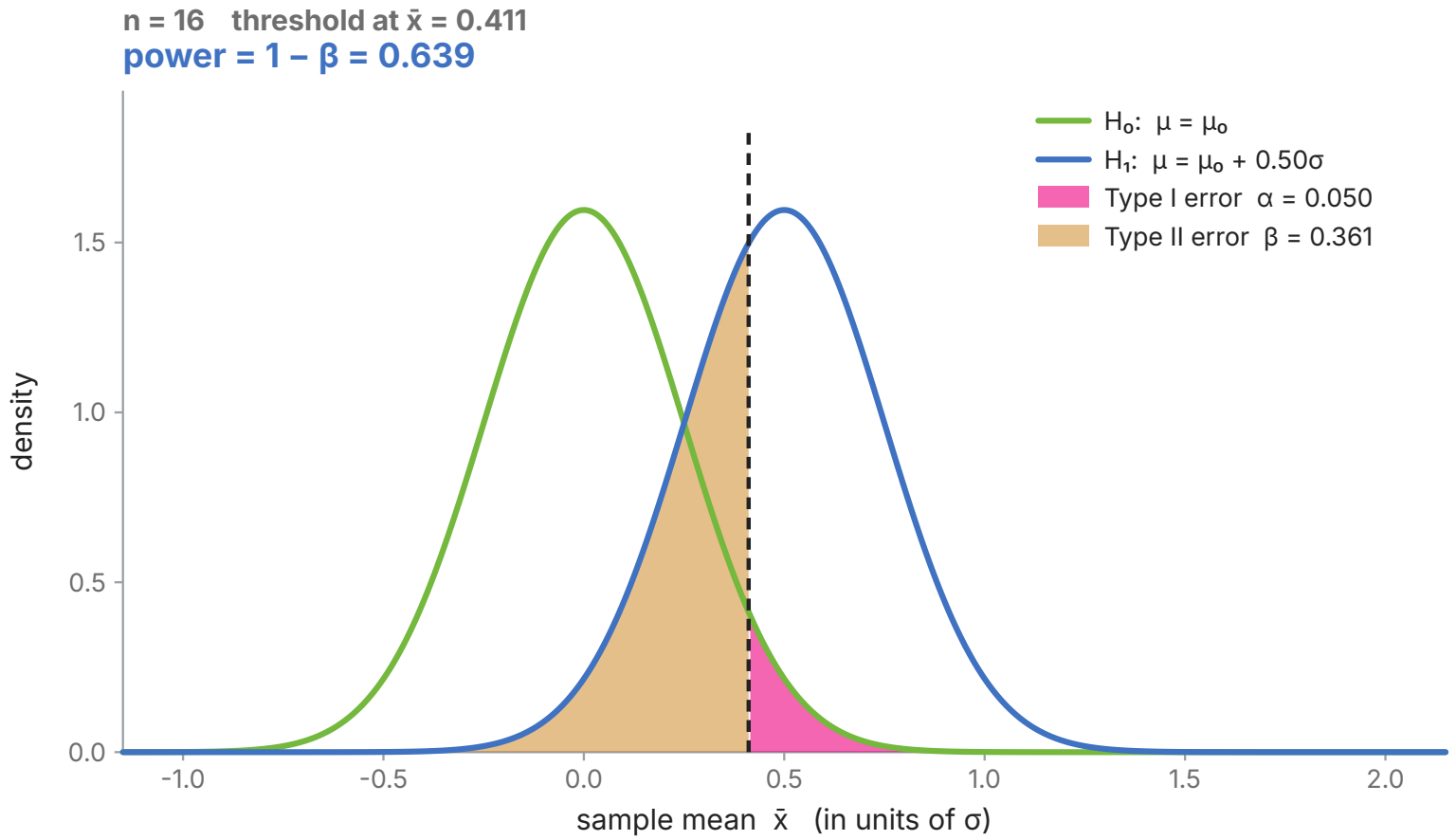
Sample size n

16

Significance α

0.05

Both curves are the sampling distribution of $\bar{X}$, under H_0 and under H_1. Raising n narrows **both** — the only move that shrinks α and β together.



Statistical Power

Definition — Power

The **power** of a test is the probability of correctly rejecting H_0 when it is false:

$$\text{Power} = 1 - \beta = P(\text{Reject } H_0 \mid H_1 \text{ true})$$

Power Function

$$\pi(\theta) = P_\theta(\text{Reject } H_0)$$

How power varies with the **true** parameter value.

Properties

- $\pi(\theta) \leq \alpha$ for $\theta \in \Theta_0$
- π grows with $|\theta - \theta_0|$
- π grows with n

Factors

- Sample size $n \uparrow$
- Significance level $\alpha \uparrow$
- Effect size $\uparrow$
- Population variance $\downarrow$

The Power Curve, Live

Each curve is one sample size. Where they cross θ_0, every curve is pinned at α.

Alternative

Two-sided

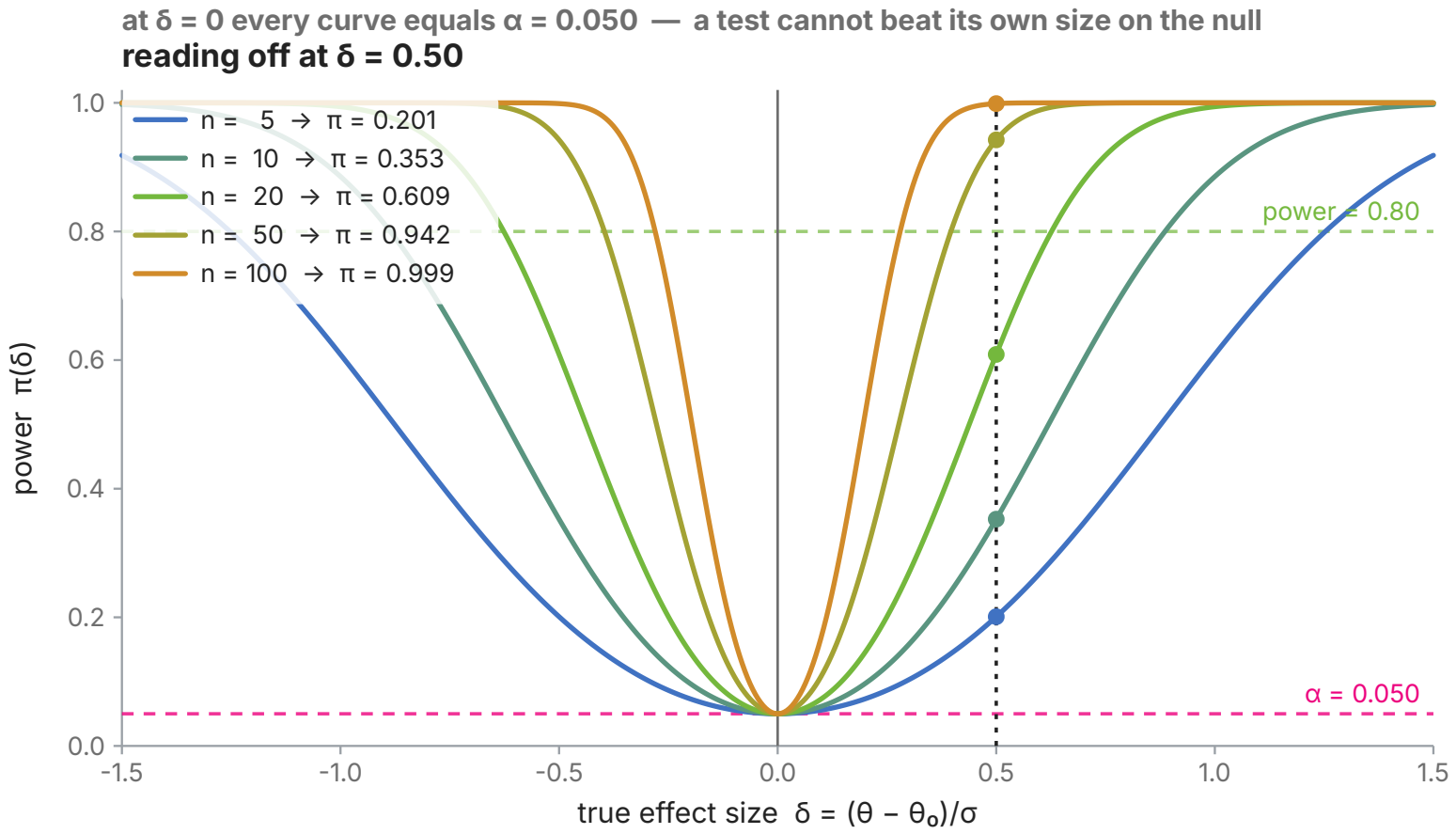
Significance α

0.05

Read off at δ =

0.5

$\delta = (\theta - \theta_0)/\sigma$ is the true effect in standard deviations. The legend reports $\pi(\delta)$ for each n at the marked value.



Likelihood Ratio Tests

Likelihood Ratio Test (LRT)

Definition — Likelihood Ratio Statistic

$$\Lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta; x)}$$

LRT Decision Rule

Reject H_0 if $\Lambda(x) < c$, with c chosen so that

$$P(\Lambda(X) < c \mid H_0) = \alpha$$

Theorem — Wilks

Under regularity conditions, as $n \to \infty$,

$$-2 \log \Lambda(X) \xrightarrow{d} \chi^2_k$$

where $k = \dim(\Theta) - \dim(\Theta_0)$.

Why it matters

- One recipe for a huge family of tests
- Asymptotically optimal
- p-values straight from a χ^2 table

The Likelihood Ratio, Live

$-2 \log \Lambda$ is exactly **twice the drop** in log-likelihood from its peak down to the best point allowed by H_0.

Model

Normal mean, σ = 1 H_0: μ = 0

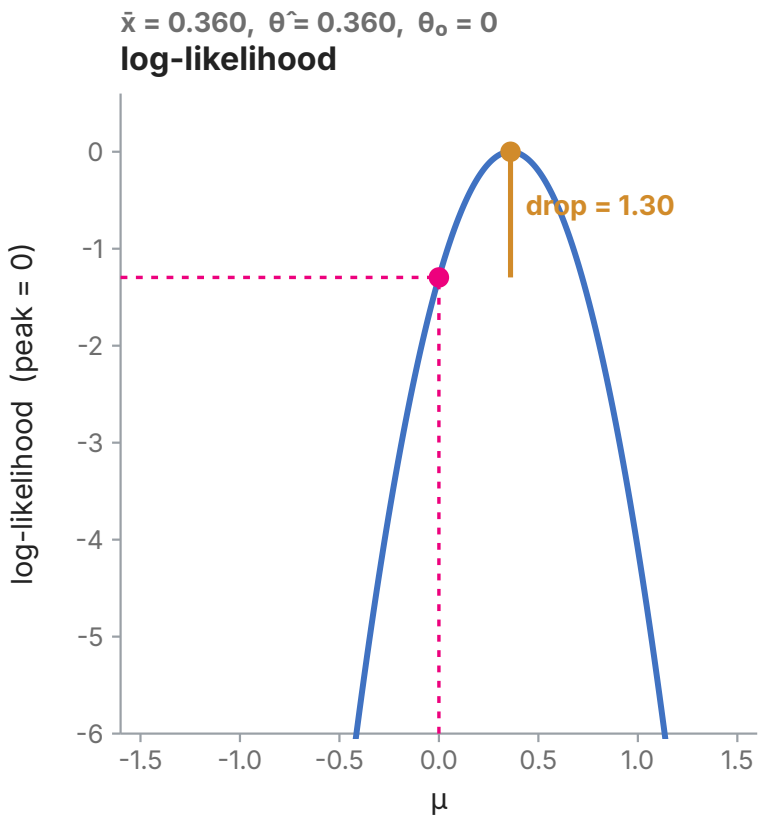
Sample size n

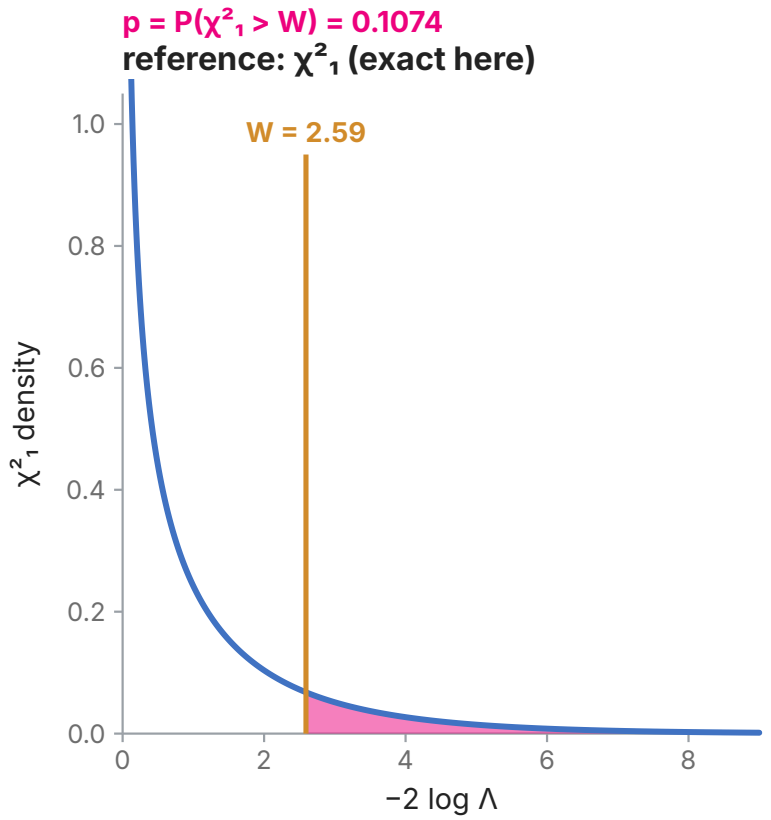
20

Observed statistic

0.62

Left: the log-likelihood, shifted so its peak is 0. Right: the χ^2_1 reference. For the normal mean the χ^2_1 law is **exact**; otherwise it is the large-n approximation.





P-values

P-values: Definition and Interpretation

Definition — p-value

The **p-value** is the probability of observing a test statistic as extreme or more extreme than the observed value, **assuming** H_0 is true:

$$p\text{-value} = P\big(T(X) \geq T(x_{\text{obs}}) \mid H_0\big)$$

Correct Interpretation

- Small $p \Longrightarrow$ evidence **against** H_0
- $p < \alpha \Longrightarrow$ reject H_0 at level α
- p measures **compatibility** of the data with H_0

A p-value is NOT

- $P(H_0 \text{ true} \mid \text{data})$
- The probability of making an error
- A measure of effect size

A large p-value is **not** evidence that H_0 is true — only that the data do not contradict it.

What a p-value Actually Does, Live

Under H_0 the p-value is **exactly uniform**. Every property people expect of it follows from that one fact.

True effect δ

0

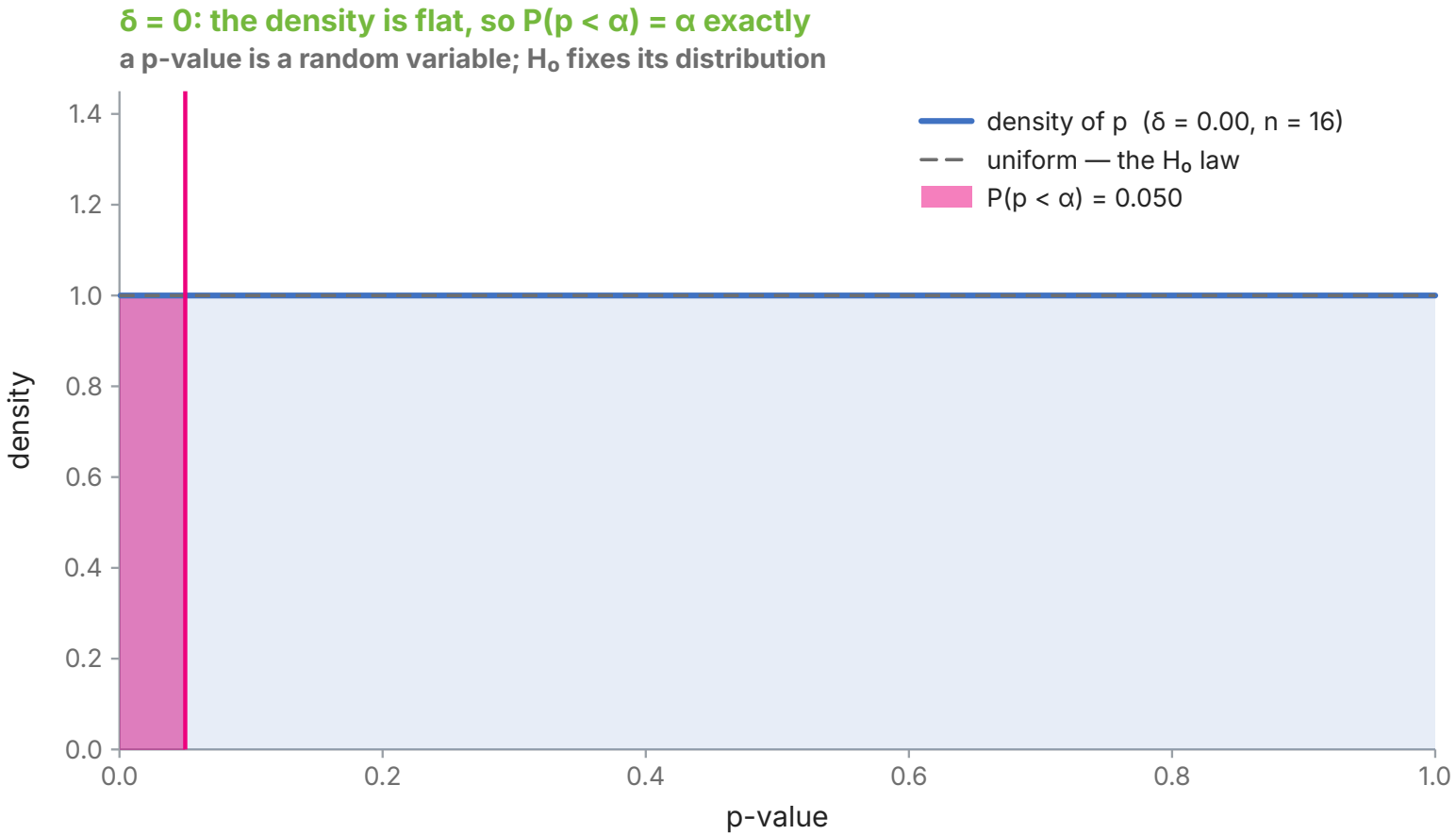
Sample size n

16

Significance α

0.05

Exact density of the two-sided p-value for a z-test. At $\delta = 0$ it is flat at 1 — so $P(p < \alpha) = \alpha$, which is what "level α" means. Raising δ or n piles the mass up against 0.



Worked Examples

Example 1: Normal Mean Test

Problem Setup

Test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ for $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

Case 1 — Known Variance

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \text{ under } H_0$$

Reject if $|Z| > z_{\alpha/2}$;

$$p = 2P(Z > |z_{\text{obs}}|)$$

Case 2 — Unknown Variance

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \text{ under } H_0$$

with

$$S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2.$$

Reject if $|T| > t_{n-1,\alpha/2}$;

$$p = 2P(t_{n-1} > |t_{\text{obs}}|)$$

Why t, not z, Live

Estimating σ costs you something. The price is paid in the tails.

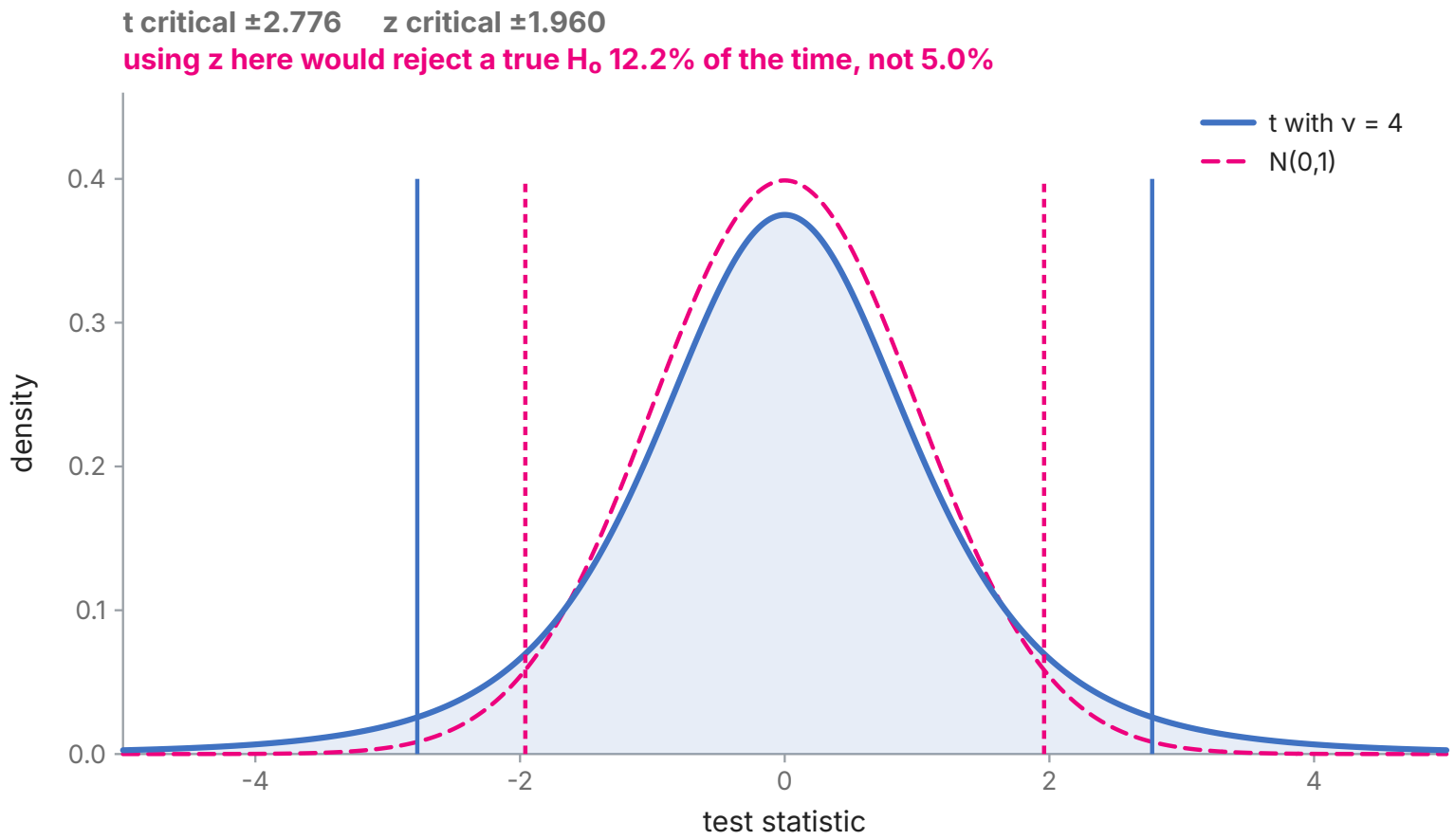
Sample size n

5

Significance α

0.05

Using $z_{\alpha/2}$ when σ is estimated does not give you level α — it gives you the larger error rate in the readout. The gap closes as $n \to \infty$.



Example 2: Variance Test

Problem Setup

Test $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$ for $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$.

Chi-Square Test

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1} \text{ under } H_0$$

Reject if $\chi^2 < \chi^2_{n-1,1-\alpha/2}$ **or** $\chi^2 > \chi^2_{n-1,\alpha/2}$.

Health Warning

- **Very sensitive** to the normality assumption
- Non-normal data can give badly wrong conclusions
- Prefer robust alternatives when in doubt

Note the null distribution is **skewed**, so the two critical values are not symmetric about $n-1$ — unlike the z and t tests.

Multiple Testing

The Multiple Testing Problem

The Problem

Running m independent tests at level α, the chance of at least one Type I error is

$$P(\text{at least one Type I error}) = 1 - (1 - \alpha)^m$$

With $\alpha = 0.05$ and $m = 20$: $P \approx 0.64$.

Family-wise Error Rate

$$\mathrm{FWER} = P(\text{reject at least one true } H_0)$$

Bonferroni: test each at α/m, which guarantees $\mathrm{FWER} \leq \alpha$.

False Discovery Rate

$$\mathrm{FDR} = E\left[\frac{\#\,\text{false discoveries}}{\#\,\text{total discoveries}}\right]$$

Less conservative than FWER control — Benjamini–Hochberg.

Multiple Testing, Live

Every square is one test. The nulls are genuinely true — every red square is a mistake the procedure made.

Number of tests m

100

True effects

10

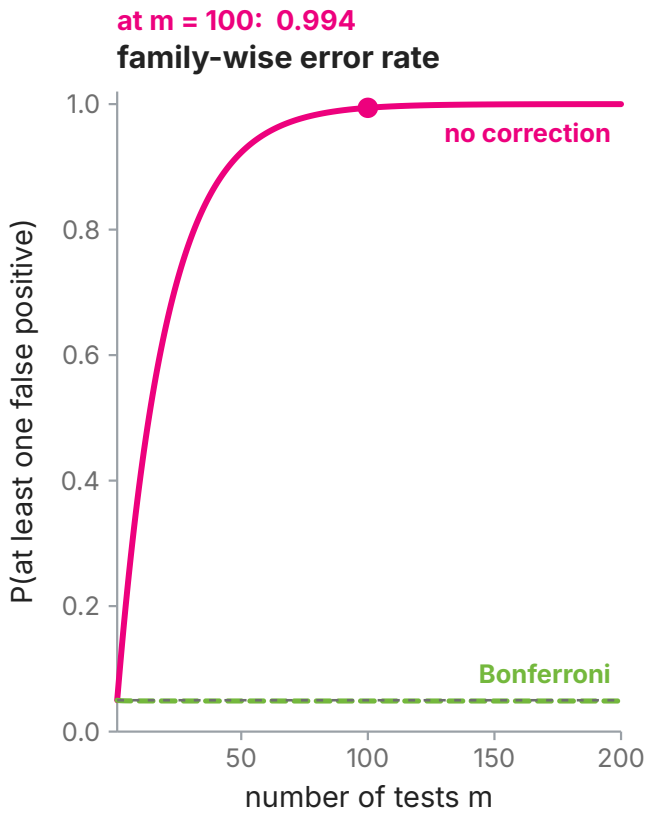
Level α

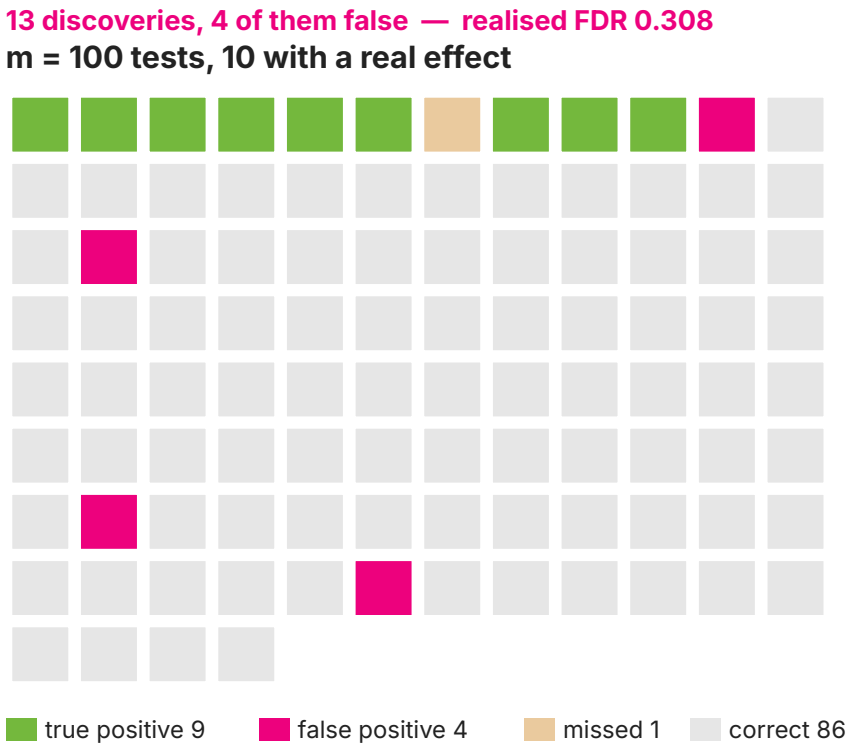
0.05

Correction

None

New draw





What the Grid Shows

- With **no correction** and 100 true nulls, you expect $100\alpha = 5$ false positives every single time — regardless of whether any effect exists.
- **Bonferroni** all but eliminates them, at the cost of missing real effects: watch the orange squares multiply.
- **Benjamini–Hochberg** sits in between — it lets a *controlled fraction* of discoveries be false, and keeps far more of the true ones.

The point

There is no correction that is simply “best”. FWER and FDR answer different questions, and you must decide which error you can live with **before** looking at the data.

Summary

Key Takeaways

(1) **Hypothesis testing** is a decision framework under uncertainty, not a proof procedure

(2) **Type I and II errors** trade off against each other at fixed n — only more data improves both

(3) **Power** is the probability of detecting a real effect, and it depends on n, α, effect size and variance

(4) **P-values** measure evidence against H_0; under H_0 they are uniform, and that is all "level α" means

(5) **Multiple testing** inflates false positives unless you correct for it

The Test Toolkit

- **LRT** — general recipe, asymptotically optimal, $-2\log\Lambda \to \chi^2_k$
- **Neyman–Pearson** — most powerful test for simple vs simple
- z-test — known variance
- t-test — unknown variance (the usual case)
- χ^2 test — variance testing

Critical Reminder

Statistical significance $\neq$ practical significance.

Best practice

Choose α and the sample size **before** seeing the data; report effect sizes alongside p-values; account for every test you ran, not just the ones that worked.