

# Section 7: Hypothesis Testing

Ricky Truong (rickytruong@college.harvard.edu),  
Emily Xing (exing@college.harvard.edu)

## 1 Introduction

### 1.1 Logistics

Welcome! The goal of our weekly section is to review content from lecture, with emphasis on big-picture intuition, mathematical proof, and hands-on practice.

Specifically, we aim for our sections to be interactive, well-paced, inclusive, and fun! Please do not hesitate to reach out if you have any questions/feedback!

### 1.2 Office Hours

- Mondays, 7:30–9:30 PM in Adams D-Hall (Ricky).
- Fridays, 11 AM–12 PM in Maxwell-Dworkin 2nd Floor (Emily).
- Saturdays, 10:30–11:30 AM in Cabot D-Hall (Emily).

## 2 Big Picture

This week we will explore *frequentist hypothesis testing*, one of the most widely used—and sadly abused—tools in applied statistics.

The core idea is that we want to make a *binary decision* about the world based on data: frequentist hypothesis testing asks, *is the observed data consistent with a pre-specified null hypothesis?* We specify the null hypothesis  $H_0$  and the alternative  $H_1$  *before* looking at the data, collect data, and then either **reject** or **retain**  $H_0$ . For instance, is a drug effective or not?

Three key quantities organize the whole framework:

- The *power function*  $\beta(\theta)$ , which gives the probability of rejecting  $H_0$  as a function of the true parameter. Ideally,  $\beta(\theta)$  is small when  $\theta \in \Theta_0$  and large when  $\theta \in \Theta_1$ .
- The *size*  $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$ , which is the worst-case Type I error probability (or false positive rate) we will tolerate. We calibrate our test to achieve a pre-specified size (commonly 0.05, though this is convention and not a law of nature).
- The *p-value*, which is the smallest  $\alpha$  at which we would have rejected  $H_0$  given the observed data. It is a summary of the evidence against  $H_0$ , not a probability that  $H_0$  is true.

We will also see that hypothesis tests and confidence intervals are *dual* to each other. That is, every confidence interval induces a family of tests, and every test induces a confidence interval. Finally, for parametric models, we will examine the Wald, score, and likelihood ratio tests give three related—but distinct—ways to test the same null hypothesis.

### 3 The Hypothesis Testing Framework

**Definition 1** (Statistical hypothesis). Partition the parameter space  $\Theta$  into two disjoint pieces:  $\Theta = \Theta_0 \cup \Theta_1$ . We test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1.$$

$H_0$  is called the **null hypothesis** and  $H_1$  is called the **alternative hypothesis**. The null is **simple** if  $\Theta_0 = \{\theta_0\}$  and **composite** otherwise (that is, it tests more than one parameter value). Tests of the form  $H_0 : \theta \leq \theta_0$  vs.  $H_1 : \theta > \theta_0$  are called **one-sided**; tests of the form  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$  are called **two-sided**.

**Definition 2** (Retention and rejection regions). The **retention region**  $A$  is the set of data values for which we retain  $H_0$ . Its complement  $A^c$  is the **rejection (critical) region**. We reject  $H_0$  if and only if  $y \in A^c$ .

If the test is based on a scalar test statistic  $T(y)$ , typical rejection regions are:

- One-sided:  $\{y : T(y) > c\}$  with **critical value**  $c$ .
- Two-sided:  $\{y : T(y) < c_L \text{ or } T(y) > c_U\}$  with critical values  $c_L < c_U$ .

**Definition 3** (Power function and size). The **power function** of a test is

$$\beta(\theta) = P_{Y;\theta}(Y \in A^c) = \int_{A^c} f_{Y;\theta}(y) dy,$$

the probability of rejecting  $H_0$  as a function of the true  $\theta$ . The **size (or level)** of the test is

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta),$$

the maximum Type I error probability over all null values.

- Ideally:  $\beta(\theta)$  is small for  $\theta \in \Theta_0$  (few false positives) and large for  $\theta \in \Theta_1$  (few false negatives).
- ~~⊗~~: Always compare tests at *equal size*. “Always reject” has power 1 everywhere but size 1—it is useless. High power is only meaningful relative to a fixed size constraint.

**Definition 4** (Type I and Type II errors).

|          |              | Truth                         |                                |
|----------|--------------|-------------------------------|--------------------------------|
|          |              | $H_0$ true                    | $H_0$ false                    |
| Decision | Reject $H_0$ | Type I error (false positive) | Correct                        |
|          | Retain $H_0$ | Correct                       | Type II error (false negative) |

$$P(\text{Type I error} \mid \theta \in \Theta_0) = \beta(\theta). \quad P(\text{Type II error} \mid \theta \in \Theta_1) = 1 - \beta(\theta).$$

## 4 Calibrating the Size

### 4.1 Two-sided test

**Example 1** (Normal,  $\sigma^2$  known). Let  $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\theta, \sigma^2)$  with  $\sigma^2$  known. For  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$ , use test statistic

$$T(Y) = \frac{\sqrt{n}\bar{Y}}{\sigma} \sim \mathcal{N}\left(\frac{\sqrt{n}\theta}{\sigma}, 1\right).$$

The symmetric rejection region  $\{T(Y) < c_L\} \cup \{T(Y) > c_U\}$  with  $c_U = -c_L$  achieves size  $\alpha$  when

$$c_U = Q_{\mathcal{N}(0,1)}(1 - \alpha/2), \quad c_L = Q_{\mathcal{N}(0,1)}(\alpha/2).$$

### 4.2 One-sided test

**Example 2** (Normal, one-sided). For  $H_0 : \theta \leq \theta_0$  vs.  $H_1 : \theta > \theta_0$  using  $T(Y) = \sqrt{n}(\bar{Y} - \theta_0)/\sigma$ , the power function is

$$\beta(\theta) = 1 - F_{\mathcal{N}(0,1)}\left(c - \frac{\sqrt{n}(\theta - \theta_0)}{\sigma}\right),$$

which is strictly *increasing* in  $\theta$ . Therefore  $\sup_{\theta \leq \theta_0} \beta(\theta) = \beta(\theta_0)$ , so controlling size only requires calibrating at the boundary:  $c = Q_{\mathcal{N}(0,1)}(1 - \alpha)$ .

- One-sided tests with monotone power, size control at the boundary  $\theta_0$  automatically controls size over the entire null space  $\Theta_0$ .

**Definition 5** (z-test). For any consistent estimator  $\hat{\theta}$  and consistent estimator  $\hat{\sigma}$  for  $\text{SD}(\sqrt{n}\hat{\theta})$ , the z-test statistic is

$$z = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

under  $H_0$ , by the CLT and Slutsky's theorem. This gives a test with nominal size  $\alpha$ .

### 4.3 T-test

**Theorem 1** (t-statistic has a t-distribution). Let  $Y_1, \dots, Y_n \stackrel{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\mu$  and  $\sigma^2$  both unknown. Then

$$T(Y) = \frac{\sqrt{n}(\bar{Y} - \mu)}{\hat{\sigma}} \sim t_{n-1},$$

where  $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{j=1}^n (Y_j - \bar{Y})^2$ . The t-test is exact (finite-sample), not asymptotic. Can you figure out what distribution it follows asymptotically?

- **One-sided**  $H_0 : \mu \leq \mu_0$  vs.  $H_1 : \mu > \mu_0$ : reject if  $\sqrt{n}(\bar{Y} - \mu_0)/\hat{\sigma} > Q_{t_{n-1}}(1 - \alpha)$ .
- **Two-sided**  $H_0 : \mu = \mu_0$  vs.  $H_1 : \mu \neq \mu_0$ : reject if  $|\sqrt{n}(\bar{Y} - \mu_0)/\hat{\sigma}| > Q_{t_{n-1}}(1 - \alpha/2)$ .

- $\otimes$ : The t-test requires (1)  $\hat{\theta}$  exactly Normal under  $H_0$ , (2)  $\hat{\sigma}^2 \sim \sigma^2 \chi^2(m)/m$  exactly, and (3)  $\hat{\theta} \perp\!\!\!\perp \hat{\sigma}^2$ . For example, the t-test is *not* valid for Poisson data:  $\hat{\lambda} = \bar{X}$  is not exactly Normal for finite  $n$ , and  $\widehat{\text{Var}}(\hat{\lambda}) = \hat{\lambda}/n$  is not independent of  $\hat{\lambda}$ . But, even if our data is incorrectly specified, estimates can still be unbiased; CIs will just be off.

**Concept Checker 1.** For  $H_0 : \lambda \geq 1$  vs.  $H_1 : \lambda < 1$  with  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$ :

1. Why can't we use a t-test here? Why is a z-test valid instead?
2. Write down the z-test statistic and rejection region for size  $\alpha = 0.05$ .
3. What is the power function  $\beta_z(\lambda)$ ?

Solution

## 5 Three Likelihood-Based Tests

Throughout, we focus on  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$  using the MLE  $\hat{\theta}$ .

**Idea.** All three tests measure how far the observed data is from the null value  $\theta_0$ —just differently on the log-likelihood curve. the Wald test uses *horizontal distance*  $\hat{\theta} - \theta_0$ , the score test uses the *slope* at  $\theta_0$ , and the likelihood ratio test uses the *vertical drop*  $\log L(\hat{\theta}) - \log L(\theta_0)$ .

**Definition 6** (Wald test). Using the asymptotic pivot  $\sqrt{I_Y(\theta_0)}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, 1)$  under  $H_0$ :

$$W(Y) = \sqrt{I_Y(\theta_0)}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, 1).$$

Reject if  $|W(Y)| > Q_{\mathcal{N}(0,1)}(1 - \alpha/2)$ , or equivalently if  $W(Y)^2 = I_Y(\theta_0)(\hat{\theta} - \theta_0)^2 > Q_{\chi_1^2}(1 - \alpha)$ .

**Definition 7** (Score test). Using the score function  $s(\theta_0; y) = \partial \log L(\theta_0; y) / \partial \theta$ , which has mean 0 and variance  $I_Y(\theta_0)$  under  $H_0$ :

$$T(Y) = \frac{s(\theta_0; Y)}{\sqrt{I_Y(\theta_0)}} \sim \mathcal{N}(0, 1).$$

Reject if  $|T(Y)| > Q_{\mathcal{N}(0,1)}(1 - \alpha/2)$ .

**Definition 8** (Likelihood Ratio (LR) test). Let  $\Lambda(y) = 2[\log L(\hat{\theta}; y) - \log L(\theta_0; y)]$ . Under  $H_0$ :

$$\Lambda(Y) \sim \chi_1^2.$$

Reject if  $\Lambda(Y) > Q_{\chi_1^2}(1 - \alpha)$ . The LR statistic is always  $\geq 0$ , since  $\hat{\theta}$  maximizes the likelihood.

- All three tests have nominal size  $\alpha$ . For a *quadratic* log-likelihood (exact NEF), they are numerically identical.
- **Which is best?** The score test requires only evaluating quantities at  $\theta_0$  (not the MLE). The Wald test has the worst finite-sample behavior. The LR test is generally preferred in practice: it does not require estimating  $I_Y(\theta)$  and enjoys invariance under reparameterization.

**Concept Checker 2.** For a two-sided z-test of  $H_0 : \mu = 0$  vs  $H_1 : \mu \neq 0$  with  $\sigma = 1$ :

1. If  $n$  increases, what happens to  $\beta(\mu)$  for  $\mu \neq 0$ ? For  $\mu = 0$ ?
2. If you increase  $\alpha$  from 0.05 to 0.10, what happens to power? To Type I error?
3. A colleague says “my test has higher power, so it’s better.” What’s missing from this claim?

Solution

**Theorem 2** (Neyman-Pearson Lemma). For a simple null  $H_0 : \theta = \theta_0$  against a simple alternative  $H_1 : \theta = \theta_1$ , the most powerful test at size  $\alpha$  rejects when

$$\text{LR}(y) = \frac{L(\theta_1; y)}{L(\theta_0; y)} > c,$$

where  $c$  is chosen so that  $P_{Y;\theta_0}(\text{LR}(Y) > c) = \alpha$ . No other size- $\alpha$  test can achieve higher power against  $\theta_1$ .

- The Neyman-Pearson lemma extends to one-sided tests for exponential families (delivering *uniformly most powerful* tests), but generally does not extend to two-sided tests.

**Example 3** (All three tests for Exponential data). Let  $Y_1, \dots, Y_n \stackrel{\text{i.i.d.}}{\sim} \text{Expo}(\theta)$ , test  $H_0 : \theta = \theta_0$  vs.  $H_1 : \theta \neq \theta_0$ . Using  $\hat{\theta} = 1/\bar{Y}$  and  $I_Y(\theta) = n/\theta^2$ :

$$W = n \left( \frac{1}{\theta_0 \bar{Y}} - 1 \right)^2, \quad S = n(1 - \theta_0 \bar{Y})^2, \quad \Lambda = 2n \{ -\log(\theta_0 \bar{Y}) + (\theta_0 \bar{Y} - 1) \},$$

all asymptotically  $\chi_1^2$  under  $H_0$ .

## 6 P-values

**Definition 9** (p-value, simple null). For a simple null  $H_0$  with test statistic  $T(Y)$  that rejects for large  $T$ , the **p-value** is

$$p(y) = P(T \geq t \mid H_0), \quad t = T(y).$$

It is the probability of observing a result at least as extreme as the data, *assuming  $H_0$  is true*.

**Definition 10** (p-value, general null). For any null  $H_0$ , with retention region  $A_\alpha$  for each size  $\alpha$ , the p-value is

$$p(y) = \min\{\alpha : y \in A_\alpha^c\},$$

the smallest size at which the test would have rejected  $H_0$ .

**Theorem 3** (p-values are Uniform under  $H_0$ ). *Let  $T(Y)$  be a continuous test statistic. Then  $p(Y) \sim \text{Unif}(0, 1)$  under  $H_0$ .*

*Proof.* The p-value is  $p = 1 - F_T(T)$ , where  $F_T$  is the CDF of  $T$  under  $H_0$ . By universality of the Uniform,  $F_T(T) \sim \text{Unif}(0, 1)$ . Since  $1 - U \sim \text{Unif}(0, 1)$  for  $U \sim \text{Unif}(0, 1)$ , we have  $p(Y) \sim \text{Unif}(0, 1)$ .  $\square$

- **Consequence:** A histogram of p-values from many independent replications of an experiment where  $H_0$  is true should look *flat*. A spike near zero suggests a true effect; a spike near 1 may suggest a conservative test.
- ~~⊗~~: The p-value is *not* the probability that  $H_0$  is true or the probability the result is due to chance. The threshold 0.05 is a convention, not a mathematical truth.
- **p-hacking:** Running many experiments and selectively reporting only those with  $p < 0.05$  (while pretending to have run only one) is FRAUD! Under  $H_0$ , we expect 5% of tests to produce  $p < 0.05$  by chance alone.

**Concept Checker 3.** Suppose  $T(Y) \sim \mathcal{N}(0, 1)$  under  $H_0$  and we observe  $T(y) = 1.44$ .

1. Compute the p-value for the one-sided test  $H_1 : \theta > \theta_0$ .
2. At what significance levels  $\alpha$  would we reject  $H_0$ ?
3. Compute the p-value for the two-sided test  $H_1 : \theta \neq \theta_0$ .

Solution

## 7 Duality: Tests $\leftrightarrow$ Confidence Intervals

**Idea.** Hypothesis tests and confidence intervals are two sides of the same coin! Any confidence interval gives a family of tests; any family of tests gives a confidence interval. The duality makes the information content of each transparent.

**Theorem 4** (Inverting a confidence interval). *If  $C(Y)$  is a  $1 - \alpha$  confidence interval for  $\theta$ , then retaining  $H_0 : \theta = \theta_0$  whenever  $\theta_0 \in C(Y)$  defines a valid  $\alpha$ -sized test.*

*Proof.*  $P_{Y;\theta_0}(\theta_0 \notin C(Y)) = 1 - P_{Y;\theta_0}(\theta_0 \in C(Y)) = 1 - (1 - \alpha) = \alpha.$  □ □

**Theorem 5** (Inverting a test). *For each  $\theta_0$ , let  $A(\theta_0)$  be the retention region of a size- $\alpha$  test of  $H_0 : \theta = \theta_0$ . Then*

$$C(y) = \{\theta : y \in A(\theta)\}$$

*is a  $1 - \alpha$  confidence interval for  $\theta$ .*

*Proof.*  $P_{Y;\theta}(\theta \in C(Y)) = P_{Y;\theta}(Y \in A(\theta)) = 1 - \alpha$ , since  $A(\theta)$  is the retention region of a size- $\alpha$  test. □ □

• **Practical takeaway:** A  $1 - \alpha$  CI conveys the result of *infinitely many* hypothesis tests simultaneously—it tells you the set of all null values that would be retained. This is why CIs are generally more informative than a single reject/retain decision.

**Concept Checker 4.** Suppose a 95% CI for an odds ratio  $\theta$  is  $(0.85, 1.20)$ .

1. Can we reject  $H_0 : \theta = 1$  at  $\alpha = 0.05$ ? Why?
2. Can we reject  $H_0 : \theta = 1.5$  at  $\alpha = 0.05$ ?
3. Does “failing to reject  $\theta = 1$ ” mean the treatment has no effect?

Solution

## 8 Bayesian posteriors?

A bit more intuitively, rather than computing p-values, Bayesians find *posterior probabilities* of the null. Simulation makes this easy.

**Example 4** (Using posterior draws). A researcher specifies a prior  $\pi(\theta)$  and likelihood  $f(y | \theta)$ , then simulates  $B = 10^4$  draws  $\theta^{[1]}, \dots, \theta^{[B]}$  from the posterior  $\theta | y$  and posts them online.

- (a) **Approximate 95% credible interval for  $\theta$ :** Use sample quantiles of the posterior draws:

$$\left[ \hat{Q}_{\theta^{[1]}, \dots, \theta^{[B]}}(0.025), \hat{Q}_{\theta^{[1]}, \dots, \theta^{[B]}}(0.975) \right].$$

This is *approximate* because  $B$  is finite—as  $B \rightarrow \infty$ , sample quantiles converge to the true posterior quantiles (error is  $O(B^{-1/2})$ ).

- (b) **Posterior median of a transformation  $\psi = g(\theta)$ :** Compute transformed draws  $\psi^{[b]} = g(\theta^{[b]})$  for each  $b$ , then take the sample median  $\hat{Q}_{\psi^{[1]}, \dots, \psi^{[B]}}(0.5)$ . No analytic derivation of the distribution of  $g(\theta)$  is needed.
- (c) **Posterior probability of  $H_0 : \psi \leq 0$ :** We want  $P(\psi \leq 0 | y) = \mathbb{E}[\mathbf{1}(\psi \leq 0) | y]$ . Approximate via the sample proportion:

$$\hat{P}(H_0 | y) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(\psi^{[b]} \leq 0).$$

This is an actual posterior probability of the hypothesis, not a p-value.

• **Frequentist vs. Bayesian:** A frequentist p-value is  $P(T \geq t | H_0)$ —a probability about the data given the null. A Bayesian posterior probability  $P(H_0 | y)$  is a probability about the hypothesis given the data. They are fundamentally different answers to different questions, with frequentist testing requiring a null, and posterior probabilities requiring a prior.

## 9 Practice Problems

**Problem 1** (Tests on Poisson). Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$  for  $n = 20$ . Consider  $H_0 : \lambda \geq 1$  vs.  $H_1 : \lambda < 1$ .

- (a) Construct a z-test with size  $\alpha = 0.05$ . Give the test statistic and rejection region. Is a t-test applicable? Why or why not?
- (b) Given observations  $x_1, \dots, x_n$ , how do you compute the p-value?
- (c) Find the power function  $\beta_z(\lambda)$ .
- (d) Construct an exact test using the distribution of  $n\bar{X}$ . Can you achieve exact size  $\alpha = 0.05$ ? Why or why not?

## Solution

**Problem 2** (Likelihood Ratio Test for Poisson). Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Pois}(\lambda)$ . Consider  $H_0 : \lambda \geq 1$  vs.  $H_1 : \lambda < 1$ .

- Show the log-likelihood  $\ell(\lambda)$  is strictly concave. What is  $\hat{\lambda}_{\text{MLE}}$ ?
- Find the constrained MLE  $\hat{\lambda}_0$  over the null space  $\Theta_0 = [1, \infty)$ .
- Compute the LRT statistic  $\Lambda(x)$  when  $\bar{x} < 1$ ; when  $\bar{x} \geq 1$ .
- Why do we never reject  $H_0$  when  $\bar{x} \geq 1$ ?

## Solution

**Problem 3** (Tests on Variance). Let  $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$  with  $\mu = 0$  known,  $\sigma^2$  unknown. Test  $H_0 : \sigma^2 \leq 1$  vs.  $H_1 : \sigma^2 > 1$  at level  $\alpha = 0.05$ .


- (a) Propose a test based on  $T_1(X) = \sqrt{n}\bar{X}$ . What is its distribution under  $\sigma^2 = 1$ ? Construct the rejection region. Comment on its power intuitively.
- (b) Improve  $T_1$  by using  $T_1^2(X)$ . What is its distribution under  $\sigma^2 = 1$ ? Find the new rejection region.
- (c) A better statistic is  $T_3 = \sum_{i=1}^n X_i^2$ . Find its distribution under  $\sigma^2 = 1$  and construct a test.
- (d) Derive the power functions  $\beta_1(\sigma^2)$ ,  $\beta_2(\sigma^2)$ ,  $\beta_3(\sigma^2)$  and compare them intuitively.

### Solution

**Problem 4** (Coins). A friend has a fair coin ( $\theta = 0.5$ ) and a biased coin ( $\theta = 0.8$ ). They select one coin and flip it 50 times, obtaining exactly 30 heads. Let  $\theta$  be the probability of heads.

- (a) Define  $\Theta$  and state appropriate hypotheses.
- (b) Perform a Wald test at level  $\alpha = 0.05$ .
- (c) Compute the p-value for the test in (b).
- (d) Compute the LRT statistic  $\Lambda(y)$ . Is the  $\chi_1^2$  approximation reasonable here?

Solution



| Formula or idea  | Description or name                                |
|--|--|
| $\Theta_0 \cup \Theta_1 = \Theta, \Theta_0 \cap \Theta_1 = \emptyset$      | Partition of parameter space                       |
| $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1$                     | Null & alternative hypotheses                      |
| $\Theta_0 = \{\theta_0\}$ , i.e. single value                              | Simple null hypothesis                             |
| Not a simple hypothesis  | Composite hypothesis                               |
| e.g. $H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$                 | One-sided hypotheses                               |
| $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$                      | Two-sided hypotheses                               |
| Reject or retain null (binary decision)                                    | Statistical hypothesis test                        |
| $y \in A$  | Retention region                                   |
| $y \in A^c$  | Rejection (critical) region                        |
| $T(y)$   | Test statistic                                     |
| Reject if $T(y) > c$   | One-sided critical value $c$                       |
| Reject if $T(y) < c_L$ or $T(y) > c_U$                                     | Two-sided critical values $c_L, c_U$               |
| $\beta(\theta) = P_{Y,\theta}(Y \notin A)$                                 | Power function                                     |
| Reject null when null is true  | Type I error (false positive, false discovery)     |
| Retain null when null is false   | Type II error (false negative)                     |
| $\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$                        | Size (level) of test                               |
| $\lim_{n \rightarrow \infty} \beta(\theta_0) = \alpha$                     | Asymptotic size control                            |
| $\beta(\theta_0)$ when asymptotically controlled                           | Nominal size                                       |
| If $\theta_0 \in C(Y)$ , retain null                                       | Duality between testing and $1 - \alpha$ CI $C(Y)$ |
| If $C(y) = \{\theta : y \in A(\theta)\}$ , then $C(Y)$ is CI               | Duality between testing and $1 - \alpha$ CI $C(Y)$ |
| $p(y) = \min\{\alpha : y \in A_\alpha^c\}$                                 | p-value: smallest $\alpha$ leading to rejection    |
| $p(Y) \sim \text{Unif}(0, 1)$ under $H_0$                                  | p-value is Uniform under the null                  |
| Report only favorite results   | p-hacking (fraud)                                  |
| $(\bar{Y} - \theta_0)/\hat{SE}(\bar{Y}) \xrightarrow{d} \mathcal{N}(0, 1)$ | t-statistic (z-test, asymptotic)                   |
| $\sqrt{n}(\bar{Y} - \mu_0)/\hat{\sigma} \sim t_{n-1}$                      | t-statistic (exact, Normal data)                   |
| $\sqrt{I_Y(\theta_0)}(\hat{\theta} - \theta_0) \sim \mathcal{N}(0, 1)$     | Wald test  |
| $s(\theta_0; y)/\sqrt{I_Y(\theta_0)} \sim \mathcal{N}(0, 1)$               | Score test   |
| $2\{\log L(\hat{\theta}) - \log L(\theta_0)\} \sim \chi_1^2$               | Likelihood ratio (LR) test                         |
| $L(\theta_1; y)/L(\theta_0; y)$ most powerful at size $\alpha$             | Neyman-Pearson Lemma (simple vs. simple)           |

Table 1: Main ideas and notation for Chapter 8 (Hypothesis Testing), from the textbook.