

Lecture 14 – Introduction to hypothesis testing

Last Modified: October 21, 2026

Outline:

14.1 Null and alternative hypotheses	1
14.2 Hypothesis tests	2
14.3 Error types	3
14.4 Power Function	4
14.5 Worked example	5

Hypothesis testing answers binary questions about a parameter, such as...

- Our first example from class: do people prefer to turn their head to the right when kissing?
- Is a drug better than placebo?
- Is the average waiting time in the ER > 30 minutes?
- Is my data Gaussian?

How do we formulate such a question using statistics?

Consider the ER question. Suppose we collect i.i.d. data $X_1, \dots, X_n$, the ER waiting times of different patients. Our parameter of interest is $\mu = \mathbb{E}[X_1]$. The binary question about the parameter is: is $\mu \leq 30$ or is $\mu > 30$?

How do we go about answering this question?

The first strategy that comes to mind is to compute $\bar{X}_n$, and check if $\bar{X}_n > 30$. But $\bar{X}_n$ is not a perfect proxy for μ . We could have $\bar{X}_n > 30$ even if $\mu < 30$. So we should only decide $\mu > 30$ if $\bar{X}_n > 30 + c$ for a large enough c . For example if $\bar{X}_n > 60$ minutes, then $\mu < 30$ is very unlikely (but not impossible). With incomplete data we have to accept the possibility of error. The language of hypothesis testing will help us reason about the types of errors and probability of making them, and how we should decide where to set the cut off.

14.1 Null and alternative hypotheses

Definition 1 (Hypotheses). In a statistical model $\{f_\theta : \theta \in \Theta\}$, a **hypothesis** is a statement that θ belongs to a specified subset of Θ . We consider

$$\underbrace{H_0 : \theta \in \Theta_0}_{\text{null hypothesis}} \quad \text{vs} \quad \underbrace{H_1 : \theta \in \Theta_1}_{\text{alternative hypothesis}},$$

where $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 \subseteq \Theta$.

Example 1. In the ER waiting time example, we might test

$$H_0 : \mu \leq 30 \quad \text{vs} \quad H_1 : \mu > 30.$$

For the effectiveness of a drug regulating blood pressure, let θ be the average change in a patient's blood pressure after taking the drug. We might test

$$\underbrace{H_0 : \theta = 0}_{\text{No effect}} \quad \text{vs} \quad \underbrace{H_1 : \theta \neq 0}_{\text{Presence of effect}}$$

How do we choose which hypothesis to call the null (H_0) and which to call the alternative (H_1)? The operating principles are

innocent (H_0) until proven guilty (H_1),

or

status quo (H_0) vs discovery (H_1).

Example 2.

- Suppose someone is suing Duke hospital for falsely claiming that their waiting times are below 30 minutes. The default presumption is $H_0 : \mu \leq 30$ (hospital is innocent). To prove the alternative hypothesis $H_1 : \mu > 30$ (guilty), the person suing the hospital would need to bring data as evidence to reject the default assumption of innocence.
- A pharma company petitions the FDA to approve the drug they developed. Then H_0 : the placebo outperforms the drug (status quo), H_1 : the drug outperforms the placebo (a scientific discovery).
- A scientist claims she's discovered a gene that cures all cancer. Then H_0 : cancer-curing gene doesn't exist (status quo), H_1 : cancer-curing gene really does exist (scientific discovery) .

The alternative is the claim for which we require convincing evidence.

Remark 1 (Interpreting a decision). We either **reject** H_0 or **fail to reject** H_0 (rather than “accept H_1 ”). Failing to reject means only that there is insufficient evidence against the null.

14.2 Hypothesis tests

Definition 2. A **hypothesis test** is a rule that specifies

1. For which sample values the decision is to reject H_0
2. For which sample values the decision is to not reject H_0 .

A hypothesis test can be specified by a function $\psi: \mathcal{X}^n \rightarrow \{0, 1\}$ where

$$\begin{aligned} \psi(x) = 1 &\implies H_0 \text{ is rejected} \\ \psi(x) = 0 &\implies H_0 \text{ is not rejected} \end{aligned}$$

This function partitions the sample space $\mathcal{X}^n$ into two disjoint sets:

$$\begin{aligned} \mathcal{R} &:= \{x : \psi(x) = 1\} && \text{(Rejection region)} \\ \mathcal{R}^c &:= \{x : \psi(x) = 0\} && \text{(Region of failing to reject)} \end{aligned}$$

In many cases, the function ψ is defined in terms of some statistic T .

Example 3. Continuing with the ER waiting times question, recall the hypotheses

$$H_0 : \mu \leq 30 \quad \text{versus} \quad H_1 : \mu > 30.$$

As we discussed, a natural approach is to reject based on the value of $T(\underline{x}) = \frac{1}{n} \sum_{i=1}^n x_i$. For example, we might choose

$$\psi(\underline{x}) = \mathbb{1}_{\{\frac{1}{n} \sum_{i=1}^n x_i > 31\}} \quad \Longleftrightarrow \quad \mathcal{R} = \left\{ \underline{x} : \frac{1}{n} \sum_{i=1}^n x_i > 31 \right\}$$

Is this test working well? To evaluate this, we need to talk about error types.

14.3 Error types

If we are performing a test, then we do not know whether H_0 or H_1 is the true one, i.e. whether the true parameter is in Θ_0 or Θ_1 ! So the types of errors we can make depend on the truth and our decision:

Decision of Test		
		Decision
Truth	H_0	Correct
	H_1	Type II Error
Decision	Correct	Type I Error
	Type II Error	Correct

By the “innocent until proven guilty” principle, Type I is the more serious error.

Example 4. A test commits a Type I error if it concludes a drug works better than the placebo when it actually doesn't. This is considered a more serious error than a Type II error, in which the test concludes the drug does not work better than placebo, when it actually does.

Error probabilities:

$$\begin{aligned} \text{Probability of Type I error :} & \quad \mathbb{P}_\theta(\psi(\underline{X}) = 1), \quad \theta \in H_0, \\ \text{Probability of Type II error :} & \quad \mathbb{P}_\theta(\psi(\underline{X}) = 0), \quad \theta \in H_1. \end{aligned}$$

Note that the probability of the two error types depends not just on whether H_0 or H_1 is true, but on the specific θ value within H_0 or H_1 .

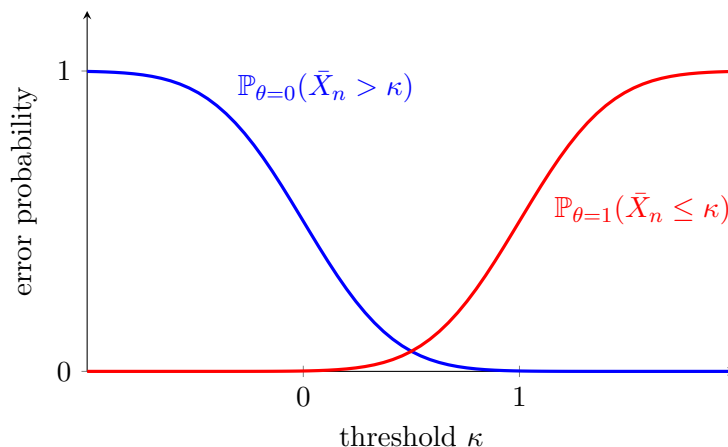
Example 5 (Tradeoff between error types). Suppose $X_i \stackrel{\text{iid}}{\sim} \mathcal{N}(\theta, 1)$. Consider the hypotheses

$$H_0 : \theta = 0 \quad \text{versus} \quad H_1 : \theta = 1.$$

We'll use the test $\psi(\underline{x}) = \mathbb{1}_{\{\bar{x}_n > \kappa\}}$ for some κ . How should we choose κ ? Consider what happens to the Type I and Type II errors. They are:

$$\begin{aligned} \mathbb{P}(\text{Type I}) &= \mathbb{P}_{\theta=0}(\bar{X}_n > \kappa), \quad (\text{decreasing in } \kappa) \\ \mathbb{P}(\text{Type II}) &= \mathbb{P}_{\theta=1}(\bar{X}_n \leq \kappa), \quad (\text{increasing in } \kappa) \end{aligned}$$

Thus, there is a tradeoff between the probabilities of Type I error and Type II errors. It is impossible to make both error probabilities arbitrarily small.



Definition 3. Let $\psi(\underline{x})$ be a hypothesis test.

- ψ is a test of **size** α if

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\psi(\underline{X}) = 1) = \alpha.$$

- ψ is a test of **level** α if

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\psi(\underline{X}) = 1) \leq \alpha.$$

A test having level α means that the probability of Type I error is no greater than α .

Strategy for designing tests: Typically, when we design a hypothesis test, we specify in advance that its level should be no more than α ; for example, $\alpha = 1\%$ and $\alpha = 5\%$ are standard. This is the max type I error we are willing to tolerate. Of these level α tests, which should we pick? For example, why not just choose the test ψ which always sets $\psi(\underline{x}) = 0$? It has size zero and is a valid level $\alpha = 1\%$ test. It never commits a type I error! Such a test says “everyone is innocent no matter what”, or “no drug works better than placebo”. Clearly, such a test is not very useful, it doesn’t lead to any discoveries.

Therefore, given a certain allowable level α , we try to max out the Type I error over all $\theta \in \Theta_0$. As we will see, doing so helps keep the Type II error low.

14.4 Power Function

Definition 4. Given a test for $H_0: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta_1$, the **power function** $\beta: \Theta \rightarrow [0, 1]$ is defined by

$$\beta(\theta) = \mathbb{P}_\theta(\psi = 1).$$

This is the probability of rejecting H_0 when θ is the true parameter.

If the hypotheses are *simple*, i.e.,

$$H_0: \theta = \theta_0 \quad \text{versus} \quad H_1: \theta = \theta_1,$$

then

$$\begin{aligned}\beta(\theta_0) &= \text{Probability of Type I error,} \\ \beta(\theta_1) &= 1 - \text{Probability of Type II error.}\end{aligned}$$

In general, we want $\beta(\theta)$ to be small when $\theta \in \Theta_0$ and large when $\theta \in \Theta_1$.

For a perfect test ψ (see Figure 14.1) the power function β is a step function: it is exactly zero when $\theta \in \Theta_0$, and exactly one when $\theta \in \Theta_1$.

The power functions in Figures 14.2 and 14.3 both correspond to tests with level α . But the former is more efficient, because it maxes out the Type I error at exactly α when crossing over from Θ_0 to Θ_1 .

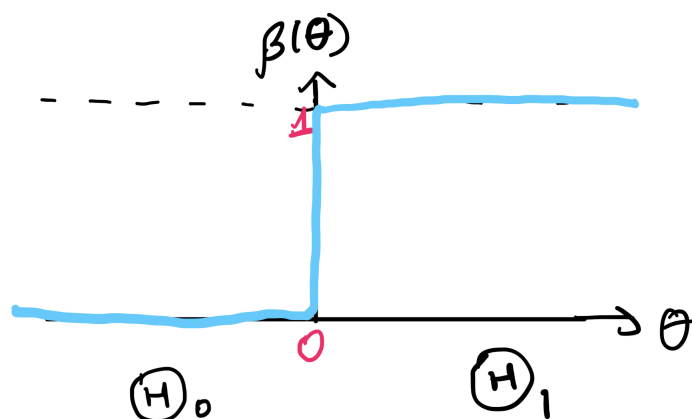


Figure 14.1: The power function for a perfect test, with zero type I error and zero type II error.

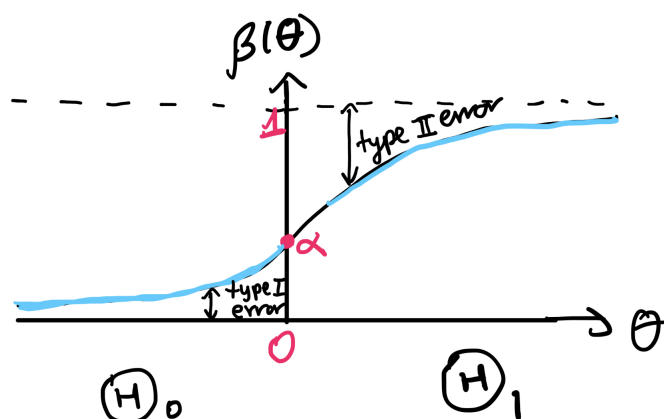


Figure 14.2: The power function for a test of size α and level α . Note the largest type I error is achieved at the boundary between Θ_0 and Θ_1 .

14.5 Worked example

Return to the ER question, $H_0 : \mu \leq 30$ versus $H_1 : \mu > 30$. Suppose

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \sigma > 0 \text{ known.}$$

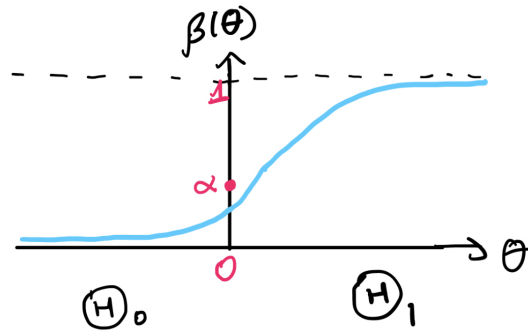


Figure 14.3: The power function for a level α test, but whose size is less than α .

Then $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$ exactly. Consider a test of the form $\psi(\underline{X}) = \mathbb{1}_{\{\bar{X}_n > 30+c\}}$. We want to choose c to make this exactly a size α test. We can express the size in terms of the power function:

$$\text{size}(\psi) = \sup_{\mu \leq 30} \beta(\mu),$$

where recall that $\beta(\mu)$ is the probability of rejecting the null when the true mean is μ . Therefore

$$\begin{aligned} \beta(\mu) &= \mathbb{P}_\mu(\bar{X}_n > 30 + c) \\ &= \mathbb{P}\left(\mu + \frac{\sigma}{\sqrt{n}}Z > 30 + c\right) \\ &= \mathbb{P}\left(Z > \frac{30 + c - \mu}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{30 + c - \mu}{\sigma/\sqrt{n}}\right), \quad Z \sim \mathcal{N}(0, 1). \end{aligned}$$

As μ increases, the cutoff inside Φ decreases, so $\beta(\mu)$ increases. This proves that the largest Type I error possible under the null is at $\mu = 30$:

$$\text{size}(\psi) = \sup_{\mu \leq 30} \beta(\mu) = \beta(30) = 1 - \Phi(c\sqrt{n}/\sigma).$$

We want to set this equal to α . Define the **upper-tail normal quantile**

$$z_\alpha = \Phi^{-1}(1 - \alpha), \quad \mathbb{P}(Z > z_\alpha) = \alpha.$$

Then we should set $c\sqrt{n}/\sigma = z_\alpha$, or $c = z_\alpha\sigma/\sqrt{n}$. Putting it all together, the size α test makes the following decision:

$$\text{Reject if } \bar{X}_n > 30 + \frac{\sigma}{\sqrt{n}}z_\alpha.$$

Example 6 (Applying the test). Suppose $n = 164$, $\sigma = 12$ minutes, and the observed mean is $\bar{x}_n = 33.4$. We have $z_{0.05} = 1.645$, so

$$30 + z_{0.05}\frac{\sigma}{\sqrt{n}} = 30 + 1.645\frac{12}{\sqrt{164}} \approx 31.54.$$

We reject when $\bar{x}_n > 31.54$. Since $33.4 > 31.54$, we reject H_0 : the data provide evidence that the population mean exceeds 30 minutes.