18.650: Fundamentals of Statistics October 16, 2023

Lecture 14— Hypothesis Testing

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

Hypothesis testing answers binary questions, such as...

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

How do we formulate such a question using statistics?
Consider the ER question. Suppose we collect i.i.d. data $X_1, \ldots, X_n$, the ER waiting times of different patients. Our parameter of interest is $\mu = \mathbb{E}[X_1]$. Is $\mu > 30$? Equivalently, is $\mu - 30 > 0$? (It helps to make 0 be the standard reference for comparison).

To answer this question, the first strategy that comes to mind is to compute $\bar{X}_n$, and check if $\bar{X}_n - 30$ is large. But how large is large? We will learn to quantify "large" in a precise way.

1 Terminology

Definition 1.1: Test & Rejection region

A *test* is a function Ψ : data $\to \{0, 1\}$. In particular, a test is an estimator. (Recall that an estimator is any function of the data).
The *rejection region* of a test is $R = \{\text{datasets for which } \Psi(\text{data}) = 1\}$.

Since Ψ only takes two values, zero or one, this means that the rejection region R fully characterizes Ψ. In fact, we can write Ψ in the equivalent form

$$\Psi(\text{data}) = \mathbb{1}\{\text{data} \in R\}.$$

Example.

Going back to the ER example, a *test* might be $\Psi(X_1, \ldots, X_n) = \mathbb{1}(\bar{X}_n > 31)$. The corresponding *rejection region* is $R = \left\{(X_1, \ldots, X_n) \mid \bar{X}_n > 31\right\}$. In other words, R is given by all the datasets $X_1, \ldots, X_n$ whose sample mean is larger than 31.

Definition 1.2: Test statistic

A test *statistic* is a function that summarizes the data and is sufficient to compute a test Ψ.

Example.

$\bar{X}_n$ is a test statistic for the test $\Psi(\text{data}) = \mathbb{1}(\bar{X}_n > 30)$. Note that $\bar{X}_n^3$ is also sufficient to compute Ψ, but we typically define the test statistic to be the most natural one.

2 The hypothesis testing problem

Let Θ be the full parameter space, and let Θ_0, Θ_1 split Θ into two disjoint subsets. A hypothesis test takes the form

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

Example.

In the waiting room example,

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

How do we choose which hypothesis to call H_0 and which to call H_1? The operating principles are

$$\textit{innocent } (H_0) \textit{ until proven guilty } (H_1),$$

or

$$\textit{status quo } (H_0) \textit{ vs discovery } (H_1).$$

Example.

> - Suppose someone is suing a 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 of 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 at the Broad claims she's discovered the gene for perfect GPA. Then H_0 : GPA gene doesn't work, H_1 : GPA gene does actually work (scientific discovery) .

2.1 Error types

We use a test to accept or reject the null hypothesis based on the value of the test statistic. There are two types of errors the test could make.

	test concludes H_0 ($\Psi = 0$)	test concludes H_1 ($\Psi = 1$)
H_0 true	✓	Type I
H_1 true	Type II	✓

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

Example.

> 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.

The probability of a test committing an error depends on the true value of the parameter. For example, $\mathbb{P}_\theta(\Psi = 1)$ is the probability that the test commits a type I error when $\theta \in \Theta_0$ is the ground truth. Similarly, $\mathbb{P}_\theta(\Psi = 0)$ is the probability that the test commits a type II error when $\theta \in \Theta_1$ is the ground truth.

Example.

> In the ER example, suppose the ground truth is some $\mu \leq 30$, and the test we are using has rejection region $\{\bar{X}_n \geq 31\}$. Then
>
> $$\mathbb{P}_\mu(\Psi = 1) = \mathbb{P}_\mu(\bar{X}_n \geq 31)$$
>
> is the probability of a type I error.

Definition 2.1: Size and level of a set

The size of a test Ψ is

$$\text{size}(\Psi) = \max_{\theta \in \Theta_0} \mathbb{P}_\theta(\Psi = 1),$$

i.e. the maximum possible probability of a type I error. The test Ψ is said to have *level* α (a number between 0 and 1) if $\text{size}(\Psi) \leq \alpha$. The typical levels used are $\alpha = 5\%$ and $\alpha = 1\%$

Remark.

> The maximum type I error probability $\mathbb{P}_\theta(\Psi = 1)$ is always achieved for θ on the boundary between Θ_0 and Θ_1; see Figure 2 (the function $\beta(\theta)$ is introduced in Definition 2.2 below). Heuristically this makes sense since for θ's on the boundary between Θ_0 and Θ_1, it is hardest to correctly decide whether to accept or reject the null hypothesis.

Note that a test which always accepts the null hypothesis has size 0 — 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. Therefore, given a certain allowable level α, we try to max out the Type I error over all $\theta \in \Theta_0$, which means that for θ on the boundary, we want $\mathbb{P}_\theta(\Psi = 1)$ to equal α exactly. This will help keep the Type II error low.

The *power* function helps us reason about Type I and Type II errors.

Definition 2.2: Power

The *power* function is defined as

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

For a perfect test Ψ (see Figure 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 2 and 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.

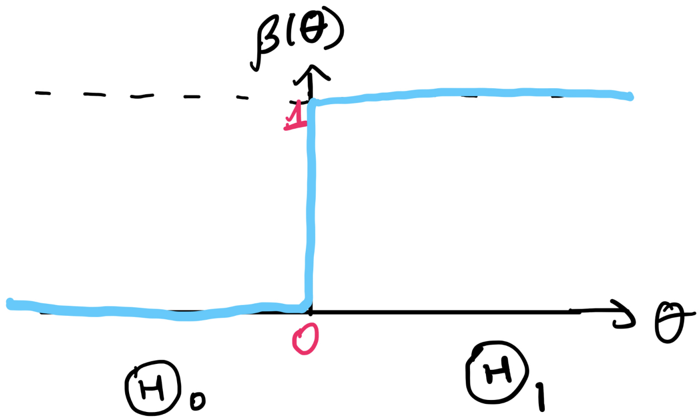

Figure 1: The power function for a perfect test, with zero type I error and zero type II error.

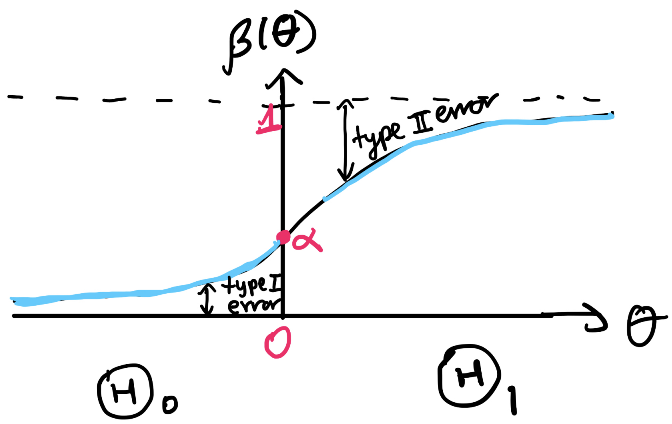

Figure 2: The power function for a test of size α and level α. Note the largest type I error is achieved at the boundary between Θ_0 and Θ_1.

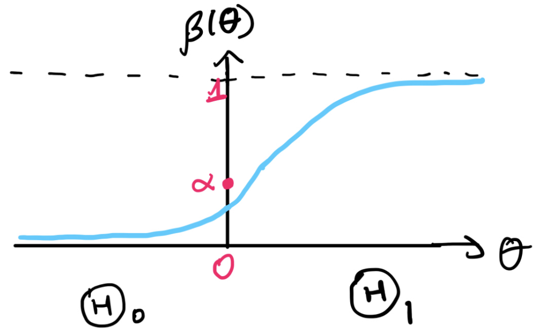

Figure 3: The power function for a test at level α, but whose size is less than α.