

Statistical Machine Learning

p -values and Multiple Testing

Horacio Gómez-Acevedo
Department of Biomedical Informatics
University of Arkansas for Medical Sciences

March 17, 2022



Hypothesis test

- ▶ A *statistical hypothesis* H is a conjecture about the probability distribution of a population.
- ▶ A hypothesis H is said to be *simple* if the distribution of the population is completely specified by H . If not, then H is called a *composite* hypothesis.
- ▶ H_0 is called the *null hypothesis* and H_1 the *alternative hypothesis*.
- ▶ We say that we commit a *type I error* if we decide to accept H_1 , whereas in reality H_0 is true.
- ▶ The probability of committing a type I error will be denoted by α .
- ▶ Acceptance of H_0 whereas H_1 is true is called a *type II error*.
- ▶ The probability of committing a type II error will be denoted by β .

Example

A factory has packets of coffee with and adjusted weight of 500 grams. We assume that the weight of the packages is $N(500, 50)$ -distributed. A container with coffee packages contains packages wrongly processed with weight $N(490, 50)$ -distributed. To determine the container that has incorrectly weight the packages, we draw a sample (package) of each of two containers X_1, X_2 . Based on the outcome of the 2-vector (X_1, X_2) we will make a conjecture about the distribution of the population. Namely

H_0 : the population is $N(500, 50)$ – distributed

H_1 : the population is $N(490, 50)$ – distributed

Let's suppose our decision as

If both X_1 and $X_2 \leq 496$ then we accept H_1 , otherwise we accept H_0 .

Example (cont)

Let's define our *critical region* G as

$$G = \{(x_1, x_2) \in \mathbb{R}^2 : x_1, x_2 \leq 496\}$$

Then, we can formulate our decision as

$$= \begin{cases} \text{if } (X_1, X_2) \in G & \text{then we choose } H_1 \text{ as our conjecture.} \\ \text{if } (X_1, X_2) \notin G & \text{then we choose } H_0 \text{ as our conjecture.} \end{cases}$$

We can calculate α and β as follows

$$\begin{aligned} \alpha &= P(\text{acceptance of } H_1 | H_0 \text{ is true}) \\ &= P((X_1, X_2) \in G | \mu = 500, \sigma^2 = 50) \\ &= P(X_1 \leq 496 | \mu = 500, \sigma^2 = 50) \cdot P(X_2 \leq 496 | \mu = 500, \sigma^2 = 50) \\ &= 0.081 \end{aligned}$$

Example (cont)

$$\begin{aligned}\beta &= P(\text{acceptance of } H_0 | H_1 \text{ is true}) \\&= P((X_1, X_2) \notin G | \mu = 490, \sigma^2 = 50) \\&= 1 - P((X_1, X_2) \in G | \mu = 490, \sigma^2 = 50) \\&= 1 - P(X_1 \leq 496 | \mu = 490, \sigma^2 = 50) \cdot P(X_2 \leq 496 | \mu = 490, \sigma^2 = 50) \\&= 0.356\end{aligned}$$

Hypothesis test

A *Hypothesis test* is a collection

$$(X_1, \dots, X_n; H_0; H_1 : G)$$

where X_1, \dots, X_n is a sample, H_0 and H_1 hypotheses concerning the probability distribution of the population and $G \subset \mathbb{R}^n$ a Borel set (meaning a collection of open sets)

If H_0 is a simple statistical hypothesis. The *level of significance* of the hypothesis test $(X_1, \dots, X_n; H_0; H_1; G)$ is understood to be the number

$$\alpha = P_{X_1, \dots, X_n}^{H_0}(G)$$

Thus, we say that α represents the probability of committing a type I error.

The power function

With our previous setup the β could not be used for composite hypothesis. Thus, we have to define something more general.

- ▶ Let $f(\cdot, \theta)_{\theta \in \Theta}$ be a family of probability densities
- ▶ Let's assume that the population X_1, \dots, X_n has a probability density $f(\cdot, \theta)$ where $\theta \in \Theta$.
- ▶ Let's assume that H_0 and H_1 are statements of the type

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

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

For a fixed $\theta \in \Theta$, the probability distribution of the population is completely specified. Then, we define for every $\theta \in \Theta_1$

$$\beta(\theta) = P_{X_1, \dots, X_n}^{\theta}(G^c)$$

The expression $1 - \beta(\theta)$ is called the *power function* for $\theta \in \Theta_1$.

Example (cont)

From our previous example, given is a $N(\mu, 50)$ -distributed population, where $\mu \leq 500$. The family of probability densities $f(\cdot, \mu)$ where $\mu \leq 500$ and

$$f(x, \mu) = \frac{1}{\sqrt{100\pi}} \exp\left(-\frac{(x - \mu)^2}{100}\right)$$

The parameter space is defined by $\Theta = (-\infty, 500]$.

If we draw a sample X_1, X_2 of size 2 from this population. We can define $\Theta_0 = \{500\}$ and $\Theta_1 = (-\infty, 500)$. This corresponds to the following hypotheses

$$H_0: \mu = 500 \quad \text{against} \quad H_1: \mu < 500$$

If we choose the following critical region

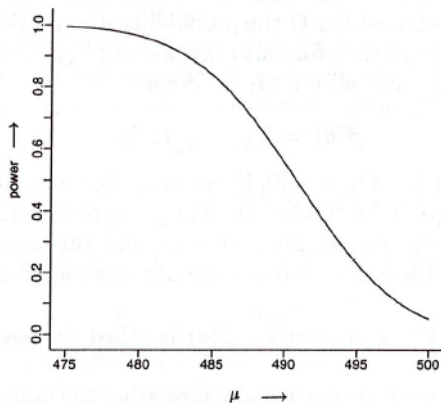
$$G = \{(x_1, x_2) \in \mathbb{R}^2 : x_1, x_2 \leq 494.63\}$$

Example (cont)

Then, our 5-tuple $(X_1, X_2; H_0; H_1; G)$ constitutes a hypothesis test. The size α of the critical region G is $\alpha = 0.05$, and the power function

$$\begin{aligned} 1 - \beta(\mu) &= 1 - P_{X_1, X_2}^{\mu}(G^c) = P_{X_1, X_2}^{\mu}(G) = P((X_1, X_2) \in G) \\ &= P(X_1 \leq 494.63 \text{ and } X_2 \leq 494.63 | \mu = \mu, \sigma^2 = 50) \\ &= P(X_1 \leq 494.63 | \mu = \mu, \sigma^2 = 50) \\ &\quad \cdot P(X_2 \leq 494.63 | \mu = \mu, \sigma^2 = 50) \end{aligned}$$

Power function example



Normally Distributed Case

Suppose we are dealing with a $N(\mu, \sigma^2)$ -distributed population where both μ and σ are unknown. If we wish to test the hypothesis

$$H_0: \mu = \mu_0 \quad \text{against} \quad H_1: \mu \neq \mu_0$$

Critical regions based on the likelihood ratio are of type

$$G = \{(x_1, \dots, x_n) \in \mathbb{R}^n : \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| \geq c\}$$

where $s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$

The outcome of the variable

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

is decisive in the choice to reject H_0 or not. This is called the *Test statistic* for this hypothesis test. Under the H_0 the test statistics is t -distributed with $n - 1$ degrees of freedom.

Normally distributed case

The decision becomes

if $|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}| \geq c$, then assume H_1

if $|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}| < c$, then assume H_0

And we have the following equivalence

$$P(|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}| \geq u) \leq \alpha \iff u \geq c$$

If the outcome u of $|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}|$ satisfies

$$P(|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}| \geq |u|) \leq \alpha$$

then, we accept H_1 .

p -values

The expression

$$P\left(\left|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}\right| \geq |u|\right)$$

is called the p -value associated with the outcome u of the test statistic.

Therefore, we normally represent the decision rule

if $P\text{-value} \leq \alpha$, then assume H_1

if $P\text{-value} > \alpha$, then assume H_0

Multiple Testing

Consider the problem of simultaneously testing m null hypotheses H_j , $j = 1, \dots, m$ and denote by R the number of rejected hypothesis. In the frequentist setting, the situation can be summarized as

	No. not rejected	No. rejected	
No. true null hypotheses	U	V	m_0
No. nontrue null hypotheses	T	S	m_1
	$m - R$	R	m

Multiple Testing (cont)

The m hypotheses are assumed to be known in advance, while the number m_0 and m_1 of true and false null hypothesis are unknown parameter, R is an observable random variable, and S , T , U , and V are unobservable random variables.

In general, we want to minimize the number V of false positives, or type I errors, and the number T of false negative or type II errors. The standard approach is to prespecify an acceptable type I error rate α that seek tests that minimize the type II error rate, within the class of tests with type I error rate α .

Test for multiple comparisons

In terms of these random variables, we can define the main rates used in the present context. When we are facing testing hypotheses of possibly thousands of significance tests, there are a number of alternatives dealing

- ▶ The per-comparison error rate (PCER). The expected value of the number of type I error over the number of hypotheses

$$PCER = \mathbb{E}(V)/m$$

- ▶ The family-wise error rate (FWER). The probability of at least one type I error

$$P(V \geq 1)$$

- ▶ The false discovery rate (FDR) is the expected proportion of type I error among rejected hypotheses

$$FDR = \mathbb{E}(V/R; R > 0) = \mathbb{E}(V/R | R > 0)P(R > 0)$$

Adjusted p -values

To account for multiple hypothesis testing, one may calculate the adjusted p -values. Given a test procedure, the adjusted p -value corresponding to the test of a single hypothesis H_j can be defined as the level of the entire test procedure at which H_j would just be rejected, given the values of all test statistics involved.

Control of the FWER at level α , the Bonferroni procedure rejects any hypothesis H_j with p -value less than or equal to α/m .

FDR control


The adjusted p -value goes according to this formula


$$\tilde{p}_{r_i} = \min_{k=i, \dots, m} \{ \min(m p_{r_k}, 1) \}$$

This adjustment leads to strong control of the FDR under the additional assumption of independence of the test statistics.

References

Materials and some of the pictures are from (Pestman, 1998), and (Speed, 2000).

 [Pestman, W. R. \(1998\). *Mathematical Statistics. An Introduction*. De Gruyter. ISBN: 3-11.015356-4.](#)

 [Speed, T., ed. \(2000\). *Statistical Analysis of Gene Expression Microarray Data*. CRC Press. ISBN: 978-1584883272.](#)

I have used some of the graphs by hacking TiKz code from StakExchange, Inkscape for more aesthetic plots and other old tricks of T_EX