

# **The Labyrinth of Multiple Testing: How to avoid the pitfall of false positives**

## **Introduction to Hypothesis testing**

**12th SISMEC National Congress 2023**

Livio Finos and Angela Andreella



- Full professor in statistics at the University of Padova
- **E-mail:**  
livio.finos@unipd.it



- Researcher in social statistics at the University Ca' Foscari Venezia
- **E-mail:**  
angela.andreella@unive.it

# Outline

## 1. Introduction

## 2. Hypothesis testing

2.1 Individual hypothesis testing

2.2 Multiple hypothesis testing

# American Statistical Association's Ethical Guidelines for Statistical Practice

Recognize that any frequentist statistical test has a random chance of indicating **significance** when it is **not really present**.

Selecting the one “significant” result from a multiplicity of parallel tests poses a grave risk of an **incorrect conclusion**.

Failure to disclose the full extent of tests and their results in such a case would be highly misleading.

# American Statistical Association's Ethical Guidelines for Statistical Practice

Recognize that any frequentist statistical test has a random chance of indicating **significance** when it is **not really present**.

Selecting the one “significant” result from a multiplicity of parallel tests poses a grave risk of an **incorrect conclusion**.



Failure to disclose the full extent of tests and their results in such a case would be highly misleading.

e.g. VaxGen's AIDSVAX trial ...

# VaxGen's AIDSVAX trial

VaxGen announced the results of the **first-ever efficacy trial** of an AIDS vaccine on 24 February 2003:

## The vaccine prevent HIV infection?









	Total	Infected			
All subjects	1679	96		5.8%	PLACEBO
	3330	191		5.7%	VACCINE

*"We saw absolutely no difference between the vaccine and placebo groups. Everyone was pretty depressed."*

but the next day...

# VaxGen's AIDSVAX trial

...by **broking the data down into racial groups** – which they say was part of the original design – the vaccine appeared to have worked in blacks:

	Total	Infected			Fisher's exact test
White	1508	81		5.4%	$p_W = 0.898$
	3003	179		6.0%	
Black	111	9		8.1%	$p_B = 0.015$
	203	4		2.2%	
Asian	20	2		10.0%	$p_A = 0.301$
	53	4		3.8%	
Other	40	6		15.0%	$p_O = 0.345$
	71	6		8.5%	

*"The numbers were small, which concerned us, but the result was highly statistically significant. They were pretty incredible results."*

# Criticisms

## 1. Failure to account for multiplicity

*"The  $p$ -values were not adjusted."*

## 2. Selective reporting (data snooping)

*"It's all murky because it's all post hoc analysis. They might as well do a subgroup analysis based on signs of the zodiac."*

**If you torture your data long enough, they will confess to you whatever you want to hear!**



# Revived interest in multiple testing

“-omics”

e.g., genomics experiments with microarray data: which genes are differentially expressed?

model selection

e.g., multiple regression: which coefficients matter?

...

## Clinical trials

### sources of multiplicity

- multiple endpoints
- several treatments
- multiple time points
- subgroup analysis
- interim analysis
- ...

### regulatory guidelines

- statistical principles for clinical trials (ICH E9)
- points to consider on multiplicity issues in clinical trials (EMA)
- ...

# Outline

## 1. Introduction

## **2. Hypothesis testing**

2.1 Individual hypothesis testing

2.2 Multiple hypothesis testing

# Hypothesis Testing: One Single Test

## Two Hypotheses under comparison

- $H_0$ : two groups are **Equal**, no relationship between  $X$  and  $Y$ .
- $H_1$ : two groups are **Different**, there is a relationship between  $X$  and  $Y$ .

Each test produces a p-value  $p$ :

if  $p \leq .05$  ( $\alpha = .05$ ), we **reject**  $H_0$  (and lean towards  $H_1$ ).

# Errors

		Null hypothesis	
		False (two groups are different)	True (two groups are equal)
Test	Rejected	True discovery	Type I error
	Not rejected	Type II error	True negative

- **Type I** (false positive): **Reject**  $H_0$  when it is **True**

$$\mathbb{P}(\text{Type I Error}) = \mathbb{P}(p \leq .05 | H_0) = .05$$

- **Type II** (false negative): **Fail** to reject  $H_0$  when it is **False**

$$\mathbb{P}(\text{Type II Error}) = \mathbb{P}(p > .05 | H_1)$$

$$\text{Power: } \mathbb{P}(p \leq .05 | H_1) = 1 - \mathbb{P}(p > .05 | H_1) = 1 - \mathbb{P}(\text{Type II Error})$$

# Asymmetric Importance of Errors

Control  $\mathbb{P}(\text{Type I Error})$  (e.g.,  $\leq 0.05$ )

and

find the test with the maximum **Power** (minimum **Type II Error**)

*It's important to remember that:*

- **A significant p-value** ( $p \leq \alpha$ ) allows us to think that  $H_1$  is true, while
- **A non-significant p-value** ( $p > \alpha$ ) does NOT allow us to think that  $H_0$  is true; we simply don't have enough evidence to reject it.

# Type I Error

Suppose  $H_0 : \mu_1 - \mu_2 = 0$  and  $H_1 : \mu_1 - \mu_2 < 0$

test statistic  $T = \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}}$  ( $\hat{\sigma}$  estimate of the std dev of  $\bar{x}_1 - \bar{x}_2$ )

under  $H_0$ :  $T \sim t_{n_1+n_2-2}$ , then

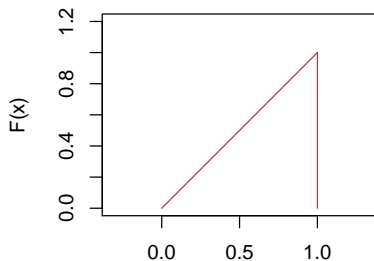
$$\mathbb{P}(T \leq t_\alpha | H_0) = \alpha \quad \forall \alpha$$

$$\mathbb{P}(F(T) \leq F(t_\alpha) | H_0) = \alpha \quad \forall \alpha$$

$$\mathbb{P}(P \leq \alpha | H_0) = \alpha \quad \forall \alpha$$

**Se vera H0**

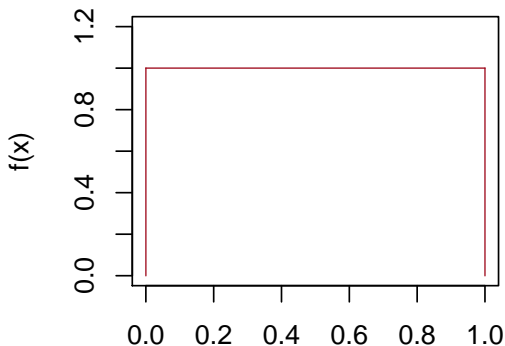
consequently,  $P \sim U(0,1)$



# Type I Error

Under  $H_0$ , the p-value is a **uniform random variable**  $U(0, 1)$

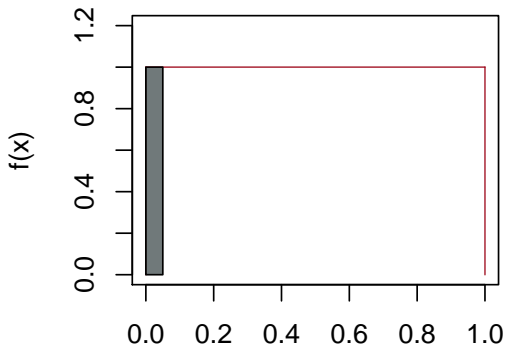
**Se vera  $H_0$**



# Type I Error

Type I Error:  $\mathbb{P}(p \leq .05 | H_0) = .05$

**Se vera H0**

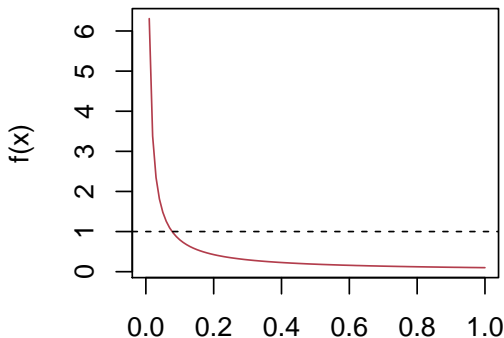




# Power

Under  $H_1$ , the p-value is **stochastically smaller** than a uniform random variable  $U(0, 1)$  (No test distortion)

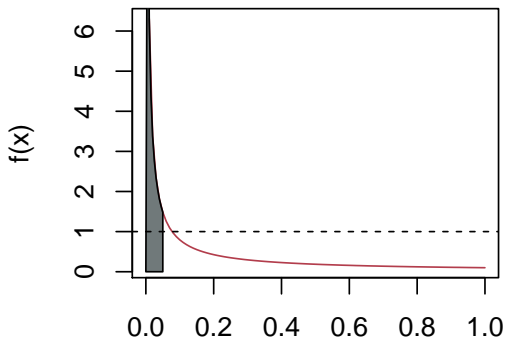
**Se vera H1 (esempio)**



# Power

Under  $H_1$ :  $\mathbb{P}(p \leq .05 | H_1) > .05$ , in our case = .74

**Se vera H1 (esempio)**



# Hypothesis testing: Multiple Tests

The goal is to test  $m \geq 2$  hypotheses simultaneously from the same data.

Each test carries the risk of making a **Type I error**  $\rightarrow$  the risk of having **AT LEAST one** may become unmanageable.

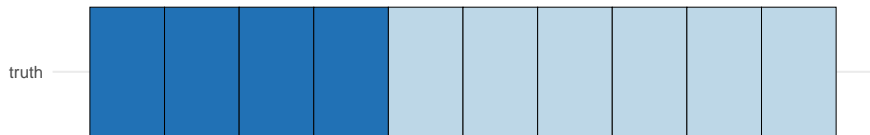
## Two Tests (Independent) Case

### EXAMPLE: independent tests

Probability of **AT LEAST one** (false) rejection?

$$\begin{aligned}\mathbb{P}(p_1 \leq .05 \cup p_2 \leq .05 | H_0) &= .05 + .05 - (.05 \cdot .05) = 1 - (1 - .05)^2 \\ &= .0975 = 1 - (1 - \alpha)^2 > \alpha\end{aligned}$$

# Multiple Tests



# Multiple Tests

truth									
rej.									

# Multiple Tests

truth									
rej.									
err.				type 2			type 1	type 1	

# Error control

		Null hypothesis		
		False	True	Tot
Test	Rejected	$S$	$V$ (false discoveries)	$R$
	Not rejected	$T$	$U$	$m - R$
Tot		$m_1$	$m_0$	$m$

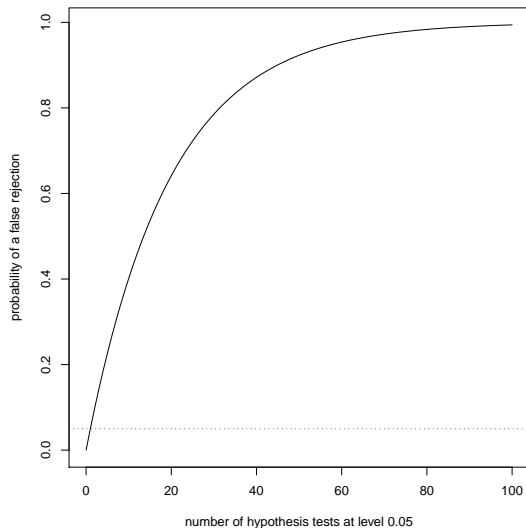
Probability of AT LEAST one false rejection (independent tests):

$$\mathbb{P}(V > 0) = 1 - (1 - \alpha)^m$$

This quickly becomes a problem if  $m$  becomes large ...



# Error control



# Type I error

- How to **define** the Type I error when there are many hypotheses?
- Which procedures **control** this error?