# Statistiques en grande dimension

Christophe Giraud[1,2]

(1) Université Paris-Sud
(2) Ecole Polytechnique

M2 Data Science & ISG

## Objectifs du cours

1. expliquer le cadre conceptuel de la stats en grande dimension (dans un contexte simple)
2. expliciter comment jouent diverses quantités sur la difficulté stat d'un problème
3. donner les clés pour comprendre la littérature scientifique en stats
4. acquérir les outils mathématiques basiques du domaine

⟶ cours sur les "fondements mathématiques" mais axé sur la compréhension plutôt que sur les maths:

- pas de preuves longues
- des petits calculs pour comprendre les choses intuitivement
- pratique par des exos

## Thématiques du cours

1. Contrôle des fausses découvertes: tests multiples
2. Fléau de la dimension: recherche de structures
3. Concept: sélection de modèle
4. Compromis algorithmique: convexification
5. Tirer partie de corrélations: faible rang

# Documents

## Documents

- Le poly
- Le site web du cours (avec les slides en ligne la veille)

http://datascience-x-master-paris-saclay.fr/catalogue/statistique-en-grande-dimension/

- Correction de certains exos (en cours d'écriture):

http://high-dimensional-statistics.wikidot.com

⚠

Cours de M2. . .

- il faut travailler le cours
- il faut s'entraider

# Evaluation

## Etape 1

1. **s'enroller sur le site web** (sinon pas de note et pas d'infos sur le cours)
2. Pour ISG: m'écrire d'ici ce soir pour qu'on vous crée des comptes

## Etape 2

Examen écrit de 3h fin janvier

## Pour ISG

Mini-projet sur la classification supervisée (février)

# Organisation du cours

1. Fausses découvertes et tests multiples
2. Fléau de la dimension et adaptation aux structures
3. Sélection de modèles
4. Convexification
5. Structures complexes et sélection d'estimateurs
6. Faible rang
7. Faible rang et sparsité
8. Classification supervisée (ISG)
9. Classification supervisée (ISG)
10. Classification supervisée (ISG)

# False discoveries

# Scientific and societal concern

The Economist

OCTOBER 19TH–25TH 2013 Economist.com

Washington's lawyer surplus

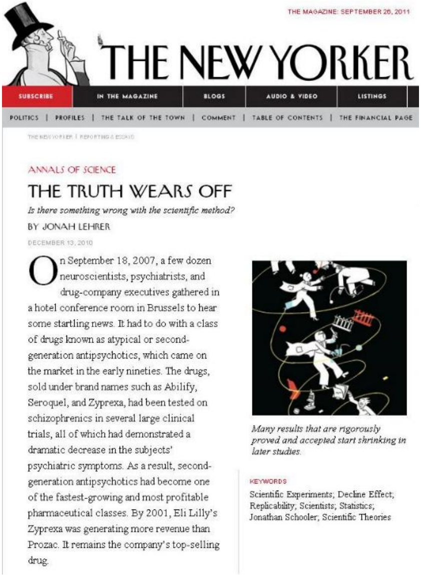
How to do a nuclear deal with Iran

Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG



THE MAGAZINE: SEPTEMBER 26, 2011

THE NEW YORKER

SUBSCRIBE
IN THE MAGAZINE
BLOGS
AUDIO & VIDEO
LISTINGS
POLITICS | PROFILES | THE TALK OF THE TOWN | COMMENT | TABLE OF CONTENTS | THE FINANCIAL PAGE
THE NEW YORKER | REPORTING & ESSAYS

ANNALS OF SCIENCE

THE TRUTH WEARS OFF

*Is there something wrong with the scientific method?*

BY JONAH LEHRER

DECEMBER 13, 2010

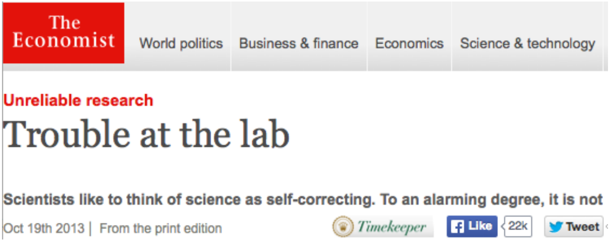
On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. The drugs, sold under brand names such as Abilify, Seroquel, and Zyprexa, had been tested on schizophrenics in several large clinical trials, all of which had demonstrated a dramatic decrease in the subjects' psychiatric symptoms. As a result, second-generation antipsychotics had become one of the fastest-growing and most profitable pharmaceutical classes. By 2001, Eli Lilly's Zyprexa was generating more revenue than Prozac. It remains the company's top-selling drug.

*Many results that are rigorously proved and accepted start shrinking in later studies.*

KEYWORDS

Scientific Experiments; Decline Effect; Replicability; Scientists; Statistics; Jonathan Schooler; Scientific Theories

# Lack of reproducibility



Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

Oct 19th 2013 | From the print edition

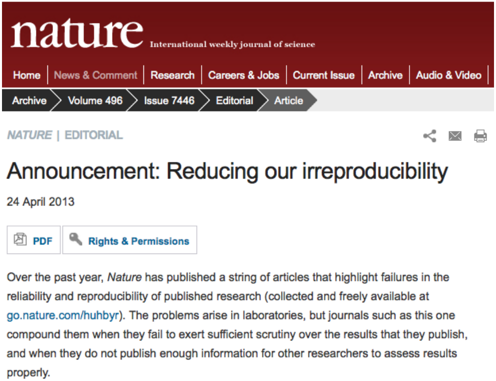
### Systematic attemps to replicate widely cited priming experiments have failed

- Amgen could only replicate 6 of 53 studies they considered landmarks in basic cancer science
- HealthCare could only replicate about 25% of 67 seminal studies
- etc

# What has gone wrong?

## Main Flaws

- Misusage of Statistics
- Publication Bias
- Publish or Perish


nature
International weekly journal of science
Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video
Archive > Volume 496 > Issue 7446 > Editorial > Article
PDF
Rights & Permissions

*NATURE* | EDITORIAL

### Announcement: Reducing our irreproducibility

24 April 2013

Over the past year, *Nature* has published a string of articles that highlight failures in the reliability and reproducibility of published research (collected and freely available at go.nature.com/huhbyr). The problems arise in laboratories, but journals such as this one compound them when they fail to exert sufficient scrutiny over the results that they publish, and when they do not publish enough information for other researchers to assess results properly.

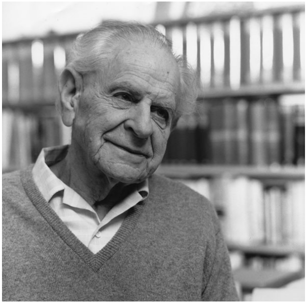
# Back to the basics

## Status of science

An hypothesis or theory can only be empirically tested.

Predictions are deduced from the theory and compared with the outcomes of experiments.

An hypothesis can be falsified or corroborated.



Karl Popper (1902-1994)

# An historical example (1935)

## The lady testing tea

A lady claims that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.

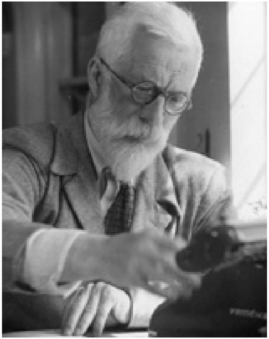
## Experiment

8 cups are brought to the lady and she has to determine whether the milk or the tea was added first.

## Test

Modeling: the success $X_1, \ldots, X_8$ are i.i.d. with $\mathcal{B}(\theta)$ distribution.

Test: $\mathcal{H}_0 : \theta = 1/2$ versus $\mathcal{H}_1 : \theta > 1/2$



R.A. Fisher (1890-1962)

# Hypothesis testing

## Testing statistics

We reject the hypothesis $\mathcal{H}_0$ : "the lady cannot discriminate" if the outcome of the variable

$$\widehat{S} = X_1 + \ldots + X_8$$

is larger than some threshold $s_{th}$.

## Choice of the threshold

We choose the threshold $s_{th} = s_{th}(\alpha)$ such that the probability to reject wrongly $\mathcal{H}_0$ is smaller than $\alpha$ (e.g. 5%)

$$\mathbb{P}_{1/2}(\widehat{S} \geq s_{th}(\alpha)) = \alpha.$$

**Reminder:** under $\mathcal{H}_0$ the distribution of $\widehat{S}$ is $\mathrm{Bin}(8, 1/2)$.

# $p$-values

## $p$-value

The $p$-value of the observation $\widehat{S}(\omega_{obs})$, is the probability to observe $\widehat{S}$ larger than $\widehat{S}(\omega_{obs})$ when $\mathcal{H}_0$ is true

$$\hat{p}(\omega_{obs}) = T_{1/2}\left(\widehat{S}(\omega_{obs})\right), \quad \text{where } T_{1/2}(s) = \mathbb{P}\left(\mathrm{Bin}(8,1/2) \geq s\right).$$

## Remark

Since

$$\widehat{S}(\omega_{obs}) \geq s_{th}(\alpha) \iff \hat{p}(\omega_{obs}) \leq \alpha$$

we reject $\mathcal{H}_0$ if the $p$-value is smaller than $\alpha$.

## Foundations of science

Science is largely based on $p$-values. An hypothesis/theory is falsified or corroborated depending on the size of the $p$-value of the outcome of some experiment(s)/observation(s).

# Where does-it go wrong?

## Publications issues

- Publication bias
- Publishing pressure
- Lack of check: replication is not ”recognized” and exponential growth of the number of scientific publications

## Small sample size

Cost of adding individuals in experiments

## New paradigm: era of Big Data

Collect data first ⟶ ask (many) questions later

## Curse of dimensionality

Issue of multiple testing (one aspect of the curse of dimensionality)

# Multiple testing

# Analyse différentielle

## Question

Est-ce que le niveau d'expression d'un gène diffère entre une condition A et une condition B ?

## Données issues d'une expérience

| Conditions | Mesures |
|---|---|
| A | $X_{A1}, \ldots, X_{Ar}$ |
| B | $X_{B1}, \ldots, X_{Br}$ |

## Objectif

Différentier entre les 2 hypothèses

$\mathcal{H}_0$ : "la moyenne des $X_{Ai}$ et des $X_{Bi}$ sont les mêmes"

$\mathcal{H}_1$ : "la moyenne des $X_{Ai}$ et des $X_{Bi}$ sont différentes"

# Exemple de test

$Y_i = X_{Ai} - X_{Bi}$ pour $i = 1, \dots, r$.

**Rejet** de $\mathcal{H}_0$ si

$$\widehat{S} := \frac{|\overline{Y}|}{\sqrt{\widehat{\sigma}^2/r}} \ \geq \ s = \text{seuil à fixer}$$

avec $\widehat{\sigma}^2 = \overline{\mathrm{var}}(Y)$

**Choix du seuil** pour contrôler le risque de rejeter $\mathcal{H}_0$ à tort

$$\mathbb{P}_{\mathcal{H}_0}(\widehat{S} \geq s_\alpha) \leq \alpha$$

**Test :** $T = \mathbf{1}_{\widehat{S} \geq s_\alpha}$

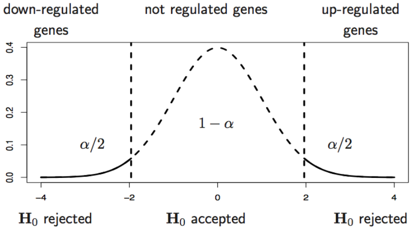
## Modèle statistique

$$X_{Ai} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_A, \sigma_A^2) \quad \text{and} \quad X_{Bi} \overset{i.i.d.}{\sim} \mathcal{N}(\mu_B, \sigma_B^2)$$

On a alors $\mathcal{H}_0 = \text{“}\mu_A = \mu_B\text{”}$.

## Loi sous $\mathcal{H}_0$

$$\widehat{S} = \frac{\overline{Y}}{\sqrt{\widehat{\sigma}^2/r}} \overset{\mathcal{H}_0}{\sim} \mathcal{T}(r-1) \quad \text{(student à } r-1 \text{ degrés de liberté)}$$



## Choix du seuil $s_\alpha$

On prend $s_\alpha$ tel que $\mathbb{P}(|\mathcal{T}(r-1)| \geq s_\alpha) = \alpha$

# Exemple : analyse différentielle de 1 gène

## Data

| $i$ | $X_A$ | $X_B$ | $Y$ |
|---|---|---|---|
| 1 | 4.01 | 4.09 | -0.08 |
| 2 | 0.84 | 0.97 | -0.12 |
| 3 | 4.45 | 3.92 | -0.53 |
| 4 | 4.73 | 6.01 | 1.28 |
| 5 | 6.16 | 6.01 | 0.15 |
| 6 | 4.23 | 6.48 | -2.26 |
| 7 | 4.70 | 5.85 | -1.15 |
| 8 | 10.65 | 11.02 | -0.37 |
| 9 | 2.02 | 4.18 | -2.16 |
| 10 | 3.96 | 5.19 | -1.23 |
| mean | 4.58 | 5.37 | -0.80 |
| std | 2.60 | 2.55 | 0.96 |

## Test

| | |
|---|---|
| $r$ | 10 |
| $\overline{Y}$ | -0.80 |
| $\sqrt{\widehat{\sigma}^2}$ | 0.96 |
| $\widehat{S}$ | 2.62 |
| $p$-value | 0.03 |

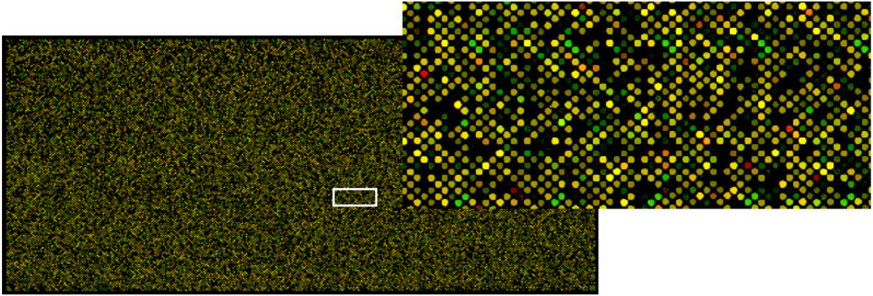
## $p$-value d'un test

Valeur de $\alpha$ pour laquelle le test change de réponse ($s_{\widehat{p}} = \widehat{S}$)

**Si $p$-value $\leq \alpha$ : $s_\alpha \leq \widehat{S}$**
le test rejette $\mathcal{H}_0$

**Si $p$-value $> \alpha$ : $s_\alpha > \widehat{S}$**
le test accepte $\mathcal{H}_0$

# Genomic data



Whole Human Genome Microarray covering over 41,000 human genes and transcripts on a standard 1" x 3" glass slide format

**High-dimensional data**

we measure 41,000 gene expression levels simultaneously!

# Blessing?

**Des nouvelles perspectives médicales**

### Objet

Personnaliser les traitements anti-cancer en combinant données cliniques et génomiques

### Moyens

RNAseq, puces CGH, etc

### Questions

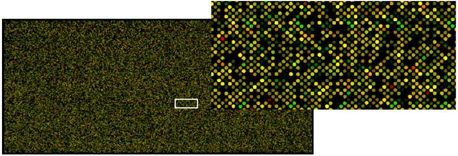
- Quelle prévision de survie?
- Quel “type” de cancer?
- Quel traitement adopter?
- etc

# Blessing?

☺ we can sense thousands of variables on each ”individual” : potentially we will be able to scan every variables that may influence the phenomenon under study.

☹ the signal might be blurred by the noise

# Comparaison multiples : analyse différentielle de $p$ gènes



Une puce microarray permet de comparer le niveau d'expression de milliers de gènes en même temps.

**Résultat:** liste de $p$-value classées par ordre croissant

| gènes | $p$-value |
|---|---|
| 2014 | $< 10^{-16}$ |
| 1078 | $6.66\ 10^{-16}$ |
| 123 | $2.66\ 10^{-15}$ |
| 548 | $1.02\ 10^{-11}$ |
| 3645 | $3.09\ 10^{-10}$ |
| $\vdots$ | $\vdots$ |

Quels gènes sont statistiquement différentiellement exprimés?

Ceux qui ont une $p$-value $\leq 5\%$ ?

Quel contrôle du risque de déclarer à tort qu'un gène est différentiellement exprimé?

# Procedure de Benjamini et Hocheberg

**Procedure de Benjamini et Hocheberg:**

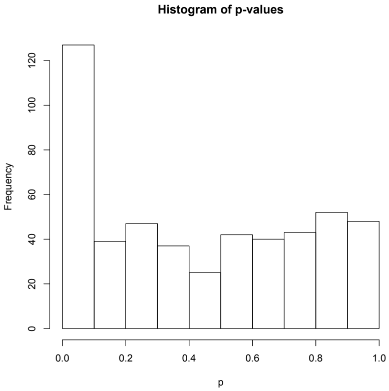
1. On ordonne les $p$-value par ordre croissant

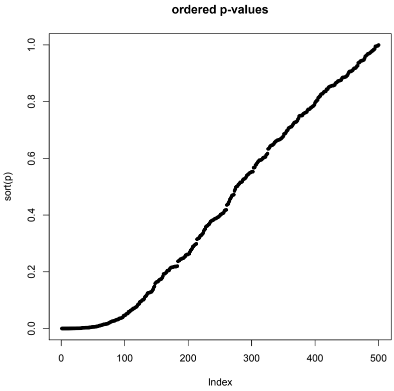
$$p(1) \leq p(2) \leq \ldots \leq p(p)$$

2. Rejet de toutes les hypothèses $\mathcal{H}_0^{(i)}$ correspondant aux $p$-values $p(1), \ldots, p(k)$ où

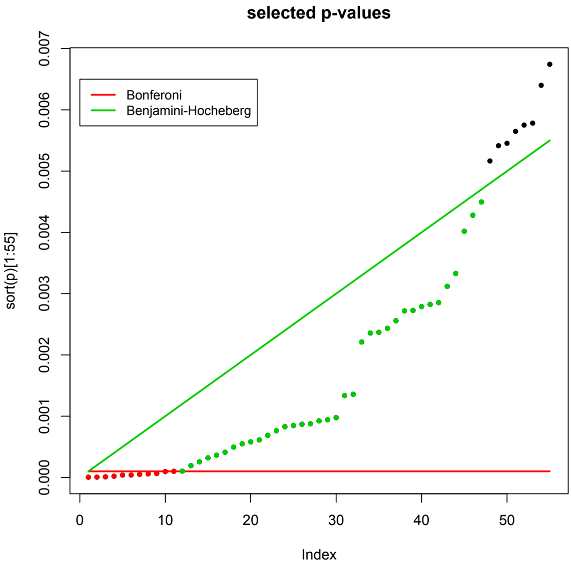
$$k = \operatorname{argmax}\{j : p(j) \leq \alpha j/p\}$$

# Exemple : *p*-values





# Exemple : comparaison avec Bonferroni ($\alpha = 5\%$)


selected p-values
Bonferoni
Benjamini-Hocheberg
sort(p)[1:55]
Index