High-dimensional statistics and probability

Christophe Giraud

Université Paris Saclay

M2 Maths Aléa & MathSV

False discoveries

Chapter 8

Scientific and societal concern

The Economist

OCTOBER 19TH–25TH 2013

Economist.com

Washington's lawyer surplus

How to do a nuclear deal with Iran

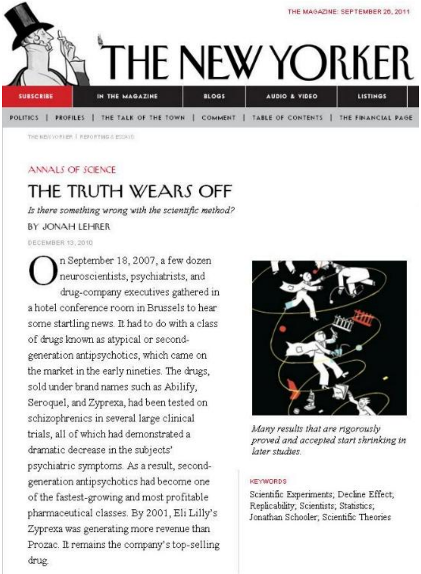
Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG

99
Es
Einsteinium



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

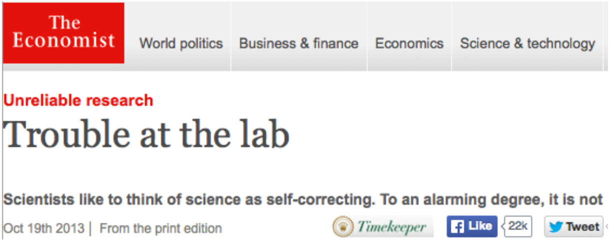
On September 18, 2007, a few dozen neuroscientists, psychiatrists, and drug-company executives gathered in a hotel conference room in Brussels to hear some startling news. It had to do with a class of drugs known as atypical or second-generation antipsychotics, which came on the market in the early nineties. 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

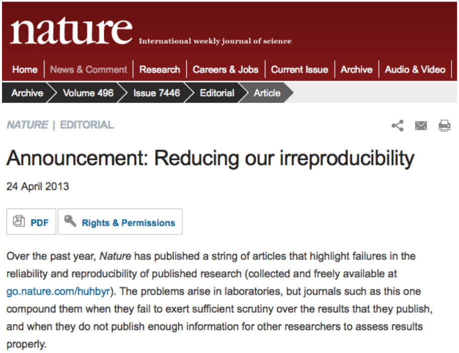
Systematic attemps to replicate widely cited priming experiments have failed

- 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

- Statistical issues
- Publication Bias
- Lack of check
- Publish or Perish
- Narcissism



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.

Special | 06 July 2018

Challenges in irreproducible research

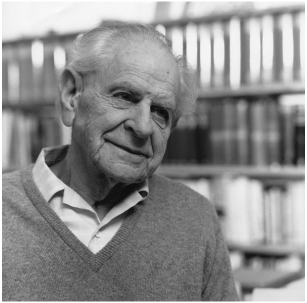
Back to the basics

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.

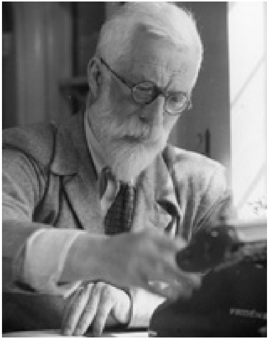
Experiment

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 $X_1, \dots, X_8$ 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 number of success

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

is larger than some threshold s_{th}.

Distribution of the test statistics

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

Choice of the threshold

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

$$\mathbb{P}\left(\mathrm{Bin}(8, 1/2) \geq s_{th}\right) \leq \alpha.$$

p-values

p-value

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

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

Remark

Since

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

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

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

Statistical issues

$$\text{Collect data first} \quad \longrightarrow \quad \text{ask (\underline{many}) questions later}$$

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

Multiple testing

Differential analysis

Question

Does the expression level of a gene vary between conditions A and B ?

Experimental data

Conditions	Observed levels
A	$X_{A1}, \dots, X_{Ar}$
B	$X_{B1}, \dots, X_{Br}$

Goal

To differentiate between two hypotheses

$\mathcal{H}_0$: "the means of the X_{Ai} and X_{Bi} are the same"

$\mathcal{H}_1$: "the means of the X_{Ai} and X_{Bi} are differents"

Example of test

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

Reject $\mathcal{H}_0$ if

$$\widehat{S} := \frac{|\overline{Y}|}{\sqrt{\overline{\mathrm{var}}(Y)/r}} \ \geq \ s = \text{threshold to be chosen}$$

Choice of the threshold in order to avoid to wrongly reject $\mathcal{H}_0$

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

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

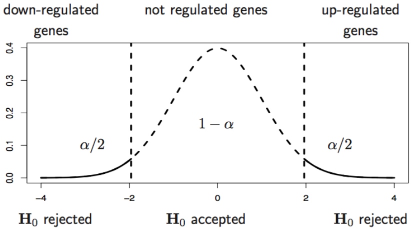
Statistical model

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

We then have $\mathcal{H}_0 \; = \; \text{“}\mu_A = \mu_B\text{”}$.

Distribution under $\mathcal{H}_0$

$$\frac{\overline{Y}}{\sqrt{\widehat{\sigma}^2/r}} \overset{\mathcal{H}_0}{\sim} \mathcal{T}(r-1) \quad \text{(student with } r-1 \text{ degrees of freedom)}$$



Choice of the threshold s_α

We choose s_α fulfilling $\mathbb{P}(|\mathcal{T}(r-1)| \geq s_\alpha) = \alpha$

Example : differential analysis of a single gene

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

$$\widehat{S} \geq s_\alpha \iff \hat{p} \leq \alpha$$

If p-value $\leq \alpha$ **:** $\widehat{S} \geq s_\alpha$

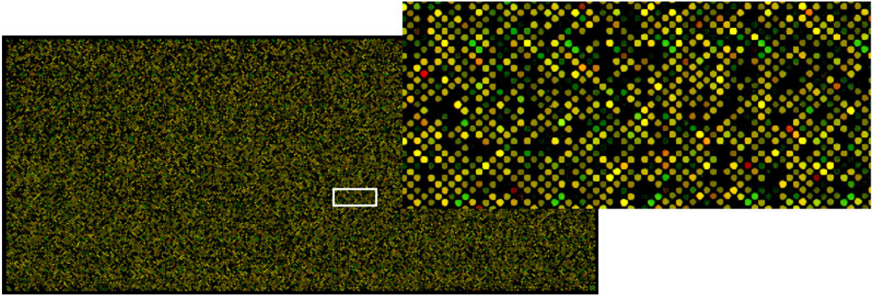
$\mathcal{H}_0$ is rejected

If p-value $> \alpha$ **:** $\widehat{S} < s_\alpha$

$\mathcal{H}_0$ is not rejected

Genomic data

We want to compare the gene expression levels for healthy/ill people.



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?

Promising medical perspectives

Object

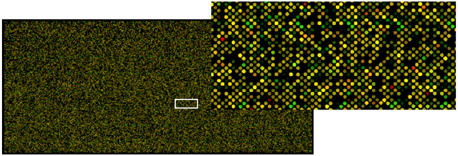
Personalized treatments against cancer by combining clinical data with genomic data

Goals

Adapt the treatment to

- the type of cancer (depending on genomic perturbations)
- the survival probability
- the personalized response to drugs
- etc

Multiple comparisons : differential analysis of p genes



A single chip allows to compare the expression levels of thousand of genes.

Ouput: an ordered list of p-values

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

Which genes have (statistically) different expression levels?

Those with a p-value $\leq 5\%$?

How many false discoveries?

An illustrative example

Assume that:

- 200 genes are differentially expressed
- you keep the p-values $\leq 5\%$

How many False Discoveries?

$$\mathbf{E}[\text{False Discoveries}] = \frac{5}{100} * (41000 - 200) = 2040$$

10 false discoveries for 1 discovery!

☹

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 curse of dimensionality : separating the signal from the noise is challenging in large multiple testing.