

M2 Mycologie

Rappels biostatistiques

Hasard, probabilités et statistiques

Estimation et test d'hypothèse

Quelques lois de probabilités utiles

Applications en biologie et bioinformatique

Statisticians are applied philosophers

Philosophers argue how many angels can dance on the head of a needle; statisticians count them. Or rather, count how many can probably dance. (...) We can predict nothing with certainty but we can predict how uncertain our predictions will be, on average that is. Statistics is the science that tells us how. – Stephen Senn [8]

Hasard, probabilités et statistiques

Deux questions récurrentes :

- Quelle est la meilleure manière de caractériser l'objet d'intérêt dans une étude scientifique ?
- Peut-on généraliser les effets observés au cours d'une expérience à une population plus vaste ?

La première question soulève le choix d'un estimateur efficace et consistant, la seconde celle de définir une statistique de test en lien avec une distribution de probabilité.


$$\underbrace{\text{variable mesurée}}_{\text{échantillon aléatoire}} = \underbrace{\text{vraie valeur}}_{\text{population théorique}} + \underbrace{\text{erreur de mesure}}_{\text{aléatoire et/ou systématique}}$$

Puissance statistique et erreur de mesure [6]


plos.org create account sign in

PLOS MEDICINE

Browse For Authors About Us

Search 

advanced search

 OPEN ACCESS	1,107,226	1,399	12,900	10,224
ESSAY	VIEWS	CITATIONS	SAVES	SHARES

Why Most Published Research Findings Are False

John P. A. Ioannidis

Published: August 30, 2005 • DOI: 10.1371/journal.pmed.0020124

Article

About the Authors

Metrics

Comments

Related Content

Download PDF

Print

Share

Abstract

Modeling the Framework for False Positive Findings

Bias

Testing by Several Independent Teams

Corollaries

Most Research Findings Are False for Most Research Designs and

Abstract

Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is

Related PLOS Articles

When Should Potentially False Research Findings Be Considered Acceptable?

Most Published Research Findings Are False—But a Little Replication Goes a Long Way

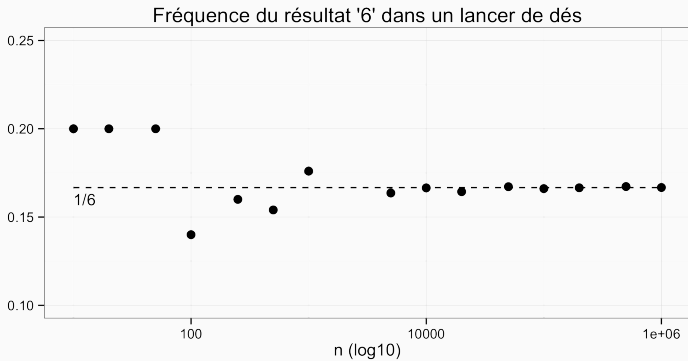
4

1. Peng, R. Reproducible research and Biostatistics. *Biostatistics* (2009), 10(3):405.
2. Prinz, F et al. Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery* (2011), 10:712.
3. Simmons, JP et al. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science* (2011), 22(11): 1359.

Lorsque l'on connaît exactement la loi qui gouverne la survenue d'événements, l'analyse combinatoire suffit dans la plupart des cas.

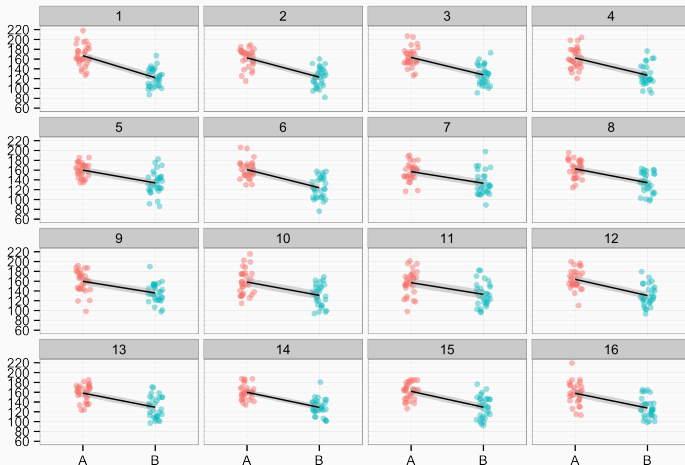
- Probabiliste : chaque face d'un dé a une probabilité $1/6$ d'apparaître, donc je peux savoir quelles sont mes chances de succès à chaque tirage.
- Statisticien : regardons combien de fois chaque face apparaît après un certain nombre de tirage, et ensuite je déciderai si mes observations sont compatibles avec l'hypothèse d'un dé équilibré.

Hasard, probabilité et statistique (2)

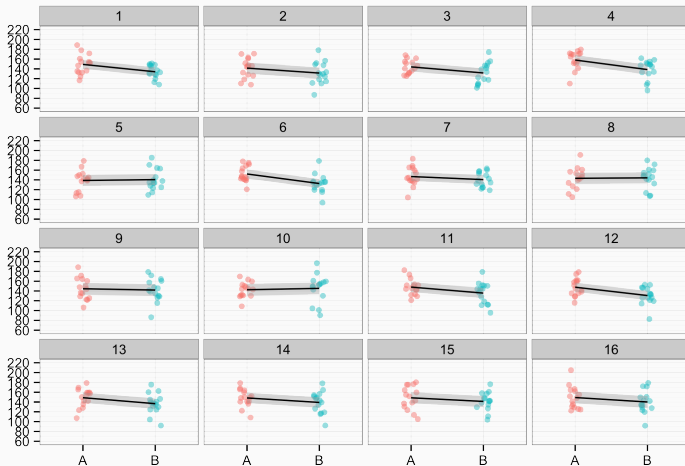


Considérons une expérience imaginaire où on simule des données de pression systolique dans deux échantillons tirés dans des populations où les moyennes sont de 160 et 130 mmHg. L'écart-type est le même dans les deux populations (20). La différence théorique est donc de 30 mmHg.

Fluctuations d'échantillonnage (2)



Fluctuations d'échantillonnage (3)



Taille d'effet et nombre de sujets nécessaires [3]

TABLE 1—Sample sizes required per group at the two sided 5% significance level for different values of d and power (d =expected mean difference/standard deviation)

d	Power ($1-\beta$)				
	99	95	90	80	50
0.10	3676	2600	2103	1571	770
0.20	920	651	527	394	194
0.30	410	290	235	176	87
0.40	231	164	133	100	49
0.50	148	105	86	64	32
0.60	104	74	60	45	23
0.70	76	54	44	33	17
0.80	59	42	34	26	13
0.90	47	34	27	21	11
1.00	38	27	22	17	9
1.10	32	23	19	14	8
1.20	27	20	16	12	7
1.30	23	17	14	11	6
1.40	20	15	12	9	5
1.50	18	13	11	8	5

Estimation et test d'hypothèse

Construction d'un estimateur et d'un test d'hypothèse

On dispose de 6 lots contenant des cellules en culture (pendant 24h), dont 3 ont reçu un supplément de vitamine E (groupe expérimental). Après 10 jours, on examine les auto-radiographies pour dénombrer le nombre total de cellules dans chaque lot.

Le technicien qui apporte les résultats rapporte au chercheur que les étiquettes permettant d'identifier quels lots ont été traités ont été égarées [5].



Formulation d'une hypothèse

Si les trois premiers lots correspondent au groupe traité à la vitamine E, alors *a priori* l'expérience semble concluante : quel que soit le lot, le nombre de cellules apparaît largement supérieur à n'importe lequel des trois derniers lots.



Est-il possible d'évaluer la plausibilité d'un tel résultat ?

Il faut définir un cadre décisionnel comprenant une hypothèse à tester et un outil permettant de prendre une décision :

- Il nous faut un moyen de comparer l'effet de l'adjonction de vitamine E par rapport à la situation où les lots ne sont pas traités.
- Un test statistique judicieusement choisi nous permettra de tester l'invraisemblance d'une hypothèse, appelée hypothèse nulle et formulée dans un cadre hypothético-déductif.

Définition d'un cadre décisionnel (2)

Si la différence observée est suffisamment grande, et on considérera que c'est le cas s'il y a moins de 5 % de chance d'observer un résultat aussi extrême, alors on conclue que celle-ci ne peut vraisemblablement pas être expliquée par de simples fluctuations d'échantillonnage et que les données observées ne sont pas compatibles avec l'hypothèse nulle d'absence d'effet, appelée H_0 .

On rejettera donc H_0 si la probabilité d'observer, du seul fait du hasard, une différence au moins aussi grande que celle observée entre les effets de A et B est inférieure à 5 %. Cette probabilité est appelée *degré de signification*. Ce seuil de signification est arbitraire, mais largement admis dans la communauté biomédicale. En somme, on accepte de se tromper dans 5 % des cas en rejetant l'hypothèse d'absence de différence.

Des risques d'erreur asymétriques

		True diagnosis		
		Positive	Negative	
Screening test	Positive	TP	FP (α)	↔ PPV
	Negative	FN (β)	TN	↔ NPV
		↕ Se	↕ Sp	

1. Définir une hypothèse nulle (H_0), une hypothèse alternative, et les risques associés à la prise d'une décision concernant le résultat observé à partir d'un échantillon.
2. Choisir une statistique de test, S .
3. Calculer la valeur de S .
4. Définir la distribution d'échantillonnage de S sous H_0 .
5. Conclure à partir de cette distribution.

Soit H_0 "la vitamine E ne modifie pas la croissance des cultures" ; en d'autres termes, les étiquettes "traité" ou "non traité" n'apportent aucune information du point de vue de la mesure considérée (tous les lots sont "échangeables"). Il y a $\binom{6}{3} = 20$ manières de définir un groupe composé de 3 éléments pris parmi 6. Considérons la somme de l'ensemble des cellules développées dans les 3 lots définissant un même groupe. Appelons la s . Ici, $s_{\text{obs}} = 121 + 118 + 110 = 349$.

Quelles sont les valeurs possibles de s lorsque l'on recombine les lots pour former deux groupes indépendants ?

Construction d'un estimateur (2)

	L1	L2	L3	s
1	121	118	110	349
2	121	118	34	273
3	121	118	12	251
-	-	-	-	-
18	110	34	22	166
19	110	12	22	144
20	34	12	22	68

Parmi les 20 résultats possibles, le résultat $s_{\text{obs}} = 349$ est le plus extrême et il y a exactement $1/20 = 5\%$ de chances d'observer un résultat aussi extrême.

Il est donc peu probable que les résultats observés (les trois premiers lots sont ceux qui ont été traités) puissent s'expliquer simplement par les fluctuations d'échantillonnage.

Un jeu de pile ou face

On lance une pièce 10 fois et on observe la séquence de résultats suivants :

P P P P F F F P F P

- Question générale : la pièce est-elle truquée ? (à reformuler sous forme d'hypothèse nulle)
- Question subsidiaire : combien de temps doit-on attendre, en moyenne, avant d'observer le premier événement "face" ?

Si l'on suppose une pièce bien équilibrée et des lancers indépendants, le nombre attendu de "Face" est $10 \times 0.5 = 5$. La fréquence observée de "Face" dans l'expérience est de $4/10 = 0.4$.

Nous pouvons formuler une hypothèse nulle selon laquelle $p = 0.5$, et l'hypothèse alternative est $p \neq 0.5$. En utilisant un test binomial, il est possible de vérifier si la proportion observée diffère de celle attendue théoriquement, en considérant un risque de 5 % de prendre une mauvaise décision en rejetant l'hypothèse nulle.

Voici les résultats calculés à l'aide d'un logiciel statistique (test binomial) :

$$\Pr(k \geq 4) = 0.828125 \quad (\text{one-sided test})$$

$$\Pr(k \leq 4) = 0.376953 \quad (\text{one-sided test})$$

$$\Pr(k \leq 4 \text{ or } k \geq 6) = 0.753906 \quad (\text{two-sided test})$$

Le résultat suggère que cette séquence de Pile/Face n'est pas incompatible avec l'hypothèse d'équi-distribution des deux côtés de la pièce.

Analyse de séquences ADN

Au lieu d'un alphabet à 2 lettres, {P, F}, considérons un alphabet à 4 lettres, {A, C, T, G}. On peut toujours se ramener au cas binomial discuté précédemment.

The DNA sequence from different regions of two closely linked human fetal globin genes were analyzed with regard to base composition: regions flanking the genes, introns, exons, and the region between the genes. [9]

Region	Length	A	C	G	T
5' Flanking	1000	0.33	0.23	0.22	0.22
3' Flanking	1000	0.29	0.15	0.26	0.30
Introns	1996	0.27	0.17	0.27	0.29
Exons	882	0.24	0.25	0.28	0.22
Between Genes	2487	0.32	0.19	0.18	0.31

Analyse de séquences ADN (2)

Les régions bordantes (en particulier 5' qui contient les promoteurs) sont-elles plus riches en A par rapport aux exons ?

Considérons les deux variables aléatoires $X_{A,5'} \sim \text{Bin}(n_1, p_1)$ et $X_{A,\text{exon}} \sim \text{Bin}(n_2, p_2)$, et $H_0 : p_1 = p_2$, ou en d'autres termes $H_0 : p_1 - p_2 = 0$ (versus $H_A : p_1 - p_2 > 0$). Il s'agit d'un test unilatéral pour deux échantillons.

On montre que la statistique de test

$$z_{\text{obs}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_1(1 - \hat{p}_1)/n_1 + \hat{p}_2(1 - \hat{p}_2)/n_2}}$$

suit une loi normale centrée-réduite sous H_0 .

Ici, $\hat{p}_1 = 0.33$, $\hat{p}_2 = 0.24$, $n_1 = 1000$ et $n_2 = 882$, soit $z_{\text{obs}} = 4.35$. La probabilité d'observer une statistique de test aussi élevée est $< 10^{-4}$.

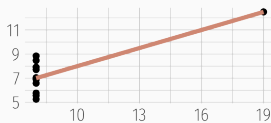
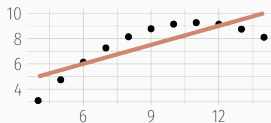
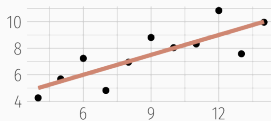
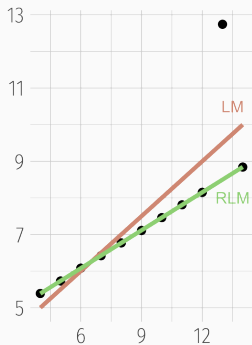
- surexpression d'un gène dans des données de RNA-Seq
- co-occurrence de deux espèces dans des données de méta-génomique
- similitude de deux arbres phylogénétiques
- correspondance entre deux motifs de séquences ADN ou protéique

One of the principal uses of statistical models is to attempt to explain variation in measurements. This variation may be due to a variety of factors, including variation from the measurement system, variation due to environmental conditions which change over the course of a study, variation from individual to individual (or experimental unit to experimental unit), etc. Factors which are not controlled from observation to observation can introduce variation in measured values. In designed experiments, the experimenter deliberately changes the levels of experimental factors to induce variation in the measured quantities, to lead to a better understanding of the relationship between experimental factors and the response. – Armitage and Colton [1]

Différents cadres de raisonnement pour l'inférence

- approche fréquentiste : ce qui a été discuté jusqu'à présent (confronter une hypothèse unique, dans une expérience contrôlée, via un principe de falsification ; Fisher, puis Neyman & Pearson). Aucune information sur $P(H_0 \mid \text{data})$. [4]
- approche par vraisemblance : utilisation des données observées pour arbitrer entre deux modèles en compétition.
- approche bayésienne : utilisation d'information externe pour évaluer *a priori* quel modèle est le plus vraisemblable (mise à jour d'une probabilité *a priori* par les données pour former une probabilité *a posteriori*) [7]

La description avant l'inférence



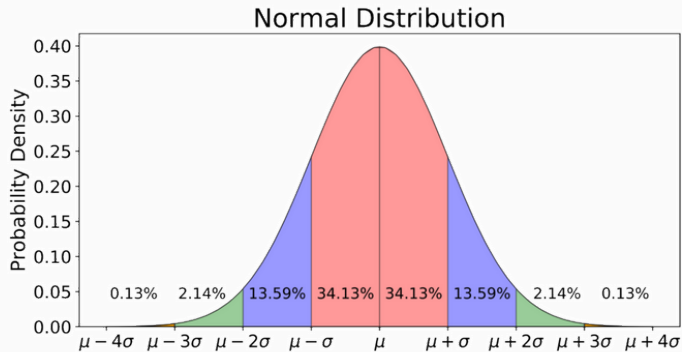
Quelques lois de probabilités utiles

Variables discrètes

loi	espérance	variance	application
binomiale	np	npq	succession d'événements 0/1
Poisson	λ	λ	comptage
binomiale négative	n/p	nq/p^2	temps d'attente avant n succès
géométrique	$1/p$	q/p^2	temps d'attente avant 1 succès

loi	espérance	variance	application
uniforme	$(b + a)/2$	$(b - a)^2/12$	distribution p-valeurs H_0
gaussienne	μ	σ^2	cumul d'erreurs indépendantes
χ^2 (Pearson)	n	$2n$	tableau de contingence
Gamma	$k\theta$	$k\theta^2$	processus temps réel

Exemple de la loi normale



- Les tests paramétriques constituent de bonnes approximations aux tests exacts (permutation), en général.
- Les tests non-paramétriques ont, pour certains, une puissance relative $\geq 80\%$ par rapport aux tests paramétriques (c'est le cas du test de Mann-Whitney-Wilcoxon pour comparer deux échantillons).

Panorama des tests statistiques usuels

non paramétrique	prédicteur	réponse	paramétrique
Spearman (ρ)	quantitative	quantitative	Pearson (r)
Fisher	qualitative	qualitative	Pearson (χ^2)
Signe	qualitative	quantitative	Student 1 éch. (t)
Kruskal-Wallis (H)	qualitative	quantitative	ANOVA 1 grp. (F)
ANOSIM	qualitative	quantitative+	MANOVA
Mann-Whitney	qualitative	quantitative	Student grp. ind. (t)
Wilcoxon	qualitative	quantitative	Student grp. app. (t)

Données omiques [2]

Omics	Biomarker data	Platforms	Features
Genome	Single nucleotide polymorphism (SNP)	Microarray	Categorical data
	Copy number variation (CNV)	DNA sequencing	Distance-driven correlation
	Loss of heterozygosity (LOH)		Extremely stable over time
	Rare variant		
Transcriptome	Gene expression	Microarray	Continuous data
	Alternative splicing	RNA sequencing	Affected by time and exposures
	Long non-coding RNA		Strong measurement noise
	Small RNA		
Proteome	Protein expression	Microarray	Continuous data
		Mass spectrometry	Affected by time and exposures
Epigenome	DNA methylation	Microarray	Continuous data
	Histone modification	Bisulfite sequencing	Affected by time and exposures
	miRNA		
Metabolome	Metabolite profiling	Mass spectrometry	Continuous data
		Nuclear magnetic resonance (NMR) spectroscopy	Affected by time and exposures
			Structured correlation
			Strongly affected by exposures

Applications en biologie et bioinformatique

En faisant l'hypothèse (erronée) que tous les nucléotides sont indépendants les uns des autres, de sorte que la probabilité d'observer n'importe lequel des nucléotides vaut $1/4$, quelle est la probabilité de trouver une séquence d'ADN donnée de longueur k dans une fenêtre de taille fixée à l'avance, n ?

$$1 - \underbrace{(1 - 1/4^k)^n}_{\text{no match}}$$

Seconde loi de Mendel

Deux organismes hétérozygotes ont pour génotype Aa et Bb. Quelle est la probabilité que leur descendant ait le génotype aa BB ?¹

	AB	Ab	aB	ab
AB	AA BB	AA Bb	Aa BB	Aa Bb
Ab	AA bB	AA bb	Aa bB	Aa bb
aB	aA BB	aA Bb	aa BB	aa Bb
ab	aA bB	aA bb	aa bB	aa bb

Puisqu'il y a indépendance, on a $P(aa) \times P(BB) = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$.

¹Rosalind bioinformatics problems

Une suspension bactérienne contient 5000 bactéries par litre. On ensemence à partir de cette suspension 50 boîtes de Pétri (1 cm^3 par boîte). Si X représente le nombre de colonies par boîte, X suit une loi de Poisson de paramètre 5, $P(\lambda = 5)$.²

Quelle est la probabilité qu'il n'y ait aucune colonie sur la boîte de Pétri ?

$$P(X = 0) = \frac{5^0 e^{-5}}{0!} = 0.0067$$

²Benjamin Jourdain, Probabilités et statistique pour l'ingénieur (2018)

Considérons N_0 brins d'ADN au début du processus. Chacun de ces brins peut être vu comme un ancêtre d'un processus de Galton-Watson, ayant pour loi de probabilité $p_1 = 1 - p$, $p_2 = p$ et $p_k = 0$ pour $k \neq 1, 2$. Ici, p représente la probabilité de succès du cycle d'amplification. L'espérance mathématique de la reproduction vaut $m = 1 + p$, et sa variance $\sigma^2 = p(1 - p) = (m - 1)(2 - m)$, avec $q = 0$ (probabilité d'extinction). Le nombre attendu de brins d'ADN après n cycles vaut alors $N_0 m^n$.

- [1] P. Armitage and E. Colton. *Encyclopedia of Biostatistics*. 2nd ed. Wiley, 2005.
- [2] Christian Baumgartner et al. *Application of Clinical Bioinformatics*. Springer, 2016.
- [3] M. J. Campbell, S. A. Julious, and D. G. Altman. “Estimating Sample Sizes for Binary, Ordered Categorical, and Continuous Outcomes in Two Group Comparisons”. In: *BMJ* 311 (1995), pp. 1145–1148.
- [4] J Cohen. “The Earth is Round ($p < .05$)”. In: *American Psychologist* 49.12 (1994), pp. 997–1003.

- [5] Phillip Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. New York: Springer-Verlag, 2005.
- [6] J. P. A. Ioannidis. “Why Most Published Research Findings Are False”. In: *PLoS Medicine* 2 (2005), e124.
- [7] Fabricia F. Nascimento, Mario dos Reis, and Ziheng Yang. “A biologist’s guide to Bayesian phylogenetic analysis”. In: *Nature Ecology & Evolution* 1 (2017), pp. 1446–1454.
- [8] Stephen Senn. *Dicing with Death*. Cambridge University Press, 2003.
- [9] Bruce S. Weir. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates. 1996.