

M2 Mycologie

Rappels biostatistiques

1. Panorama des techniques statistiques

- Bases de l'estimation et des tests d'hypothèse
- Lois de probabilité utiles pour la génétique
- Analyse probabilistique des séquences génomiques
- Classification d'espèces par approches multivariées

2. Applications bioinformatiques

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 [3]

L'inférence en 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.

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

Quelques mots-clés


- test paramétrique, non paramétrique et test exact
- risque d'erreur et puissance statistique
- estimateur consistant et efficace
- compromis biais/variance
- significativité statistique et pratique

Puissance statistique et erreur de mesure [2]


PLOS MEDICINE

[Browse](#) | [For Authors](#) | [About Us](#)

[plos.org](#) [create account](#) [sign in](#)

Search 

[advanced search](#)

 OPEN ACCESS

1,107,226 VIEWS

1,399 CITATIONS

12,900 SAVES

10,224 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

 CrossMark

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](#)

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

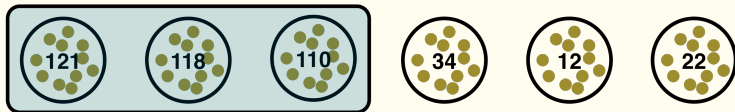
On dispose de 6 lots contenant des cellules en culture (pendant 24h), dont 3 ont reçu un supplément de vitamine E (groupe expérimental). Après 10 jours, on examine les auto-radiographies pour dénombrer le nombre total de cellules dans chaque lot.

Le technicien qui apporte les résultats rapporte au chercheur que les étiquettes permettant d'identifier quels lots ont été traités ont été égarées [1].



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 ?

Définition d'un cadre décisionnel

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.

Démarche du test d'hypothèse

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.

Construction d'un estimateur

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 ?

	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

Interprétation du test

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" ?

Opérationnalisation

Si l'on suppose une pièce bien équilibrée et des lancers indépendants, le nombre attendu de "Pile" est $10 \times 0.5 = 5$. La fréquence observée de "Pile" 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 :

$$\begin{aligned}\Pr(k \geq 4) &= 0.828125 && \text{(one-sided test)} \\ \Pr(k \leq 4) &= 0.376953 && \text{(one-sided test)} \\ \Pr(k \leq 4 \text{ or } k \geq 6) &= 0.753906 && \text{(two-sided test)}\end{aligned}$$

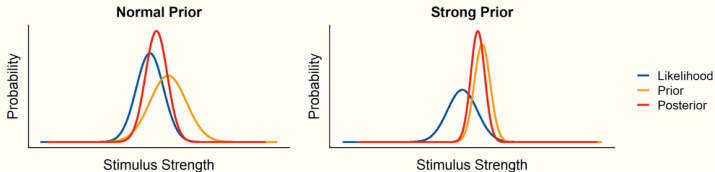
Le résultat suggère que cette séquence de Pile/Face n'est pas incompatible avec l'hypothèse d'équi-distribution des deux côtés de la pièce.

Différents cadres de raisonnement pour l'inférence

- approche fréquentiste : ce qui a été discuté jusqu'à présent
- approche par vraisemblance :
- approche bayésienne

Approche bayésienne

Graphical illustration of likelihood, prior and posterior in a Bayesian framework, for both a normal, relatively shallow prior, and a strong, extremely precise prior.



doi: <https://doi.org/10.1371/journal.pone.0236732.g001>

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

Exemple de la loi binomiale

TODO : CDF et PDF

Variables continues

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

TODO : CDF et PDF

Tests exacts, approchés, paramétriques et non paramétriques

- Les tests paramétriques constituent de bonnes approximations aux tests exacts (permutation), en général.
- Les tests non-paramétriques ont, pour certains, une efficacité relative $> 80\%$ par rapport aux tests paramétriques.

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 dans une fenêtre de taille fixée à l'avance ?

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 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. Onensemence 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 ?

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

Le modèle de Jukes-Cantor en phylogénie

On souhaite comparer deux espèces (eucaryotes) ayant un ancêtre commun. Des mutations aléatoires intervenant sur le génome au cours de leur évolution, un alignement des deux séquences ne donnera pas 100 % d'identité.

Matrice de transition :

$$\begin{pmatrix} -0.886 & 0.190 & 0.633 & 0.063 \\ 0.253 & -0.696 & 0.127 & 0.316 \\ 1.266 & 0.190 & -1.519 & 0.063 \\ 0.253 & 0.949 & 0.127 & -1.329 \end{pmatrix}$$

Si l'on se trouve dans l'état A, on y restera un temps exponentiel de paramètre $-q_{ii} = 0.886$. La probabilité d'observer la transition $A \rightarrow C$ est donnée par $-q_{ij}/q_{ii} = \frac{0.190}{0.886}$.

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 $N_0 m^n$.

Références

- [1] Phillip Good. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*. New York: Springer-Verlag, 2005.
- [2] J. P. A. Ioannidis. “Why Most Published Research Findings Are False”. In: *PLoS Medicine* 2 (2005), e124.
- [3] Stephen Senn. *Dicing with Death*. Cambridge University Press, 2003.