Série de Fiches Méthodologiques en Recherche et en Rédaction Scientifique
Fiche Méthodologique n°1/2021:
Comment calculer et interpréter la valeur de «p» dans une étude épidémiologique

How to calculate and interpret the p-value in an epidemiological study?

كيف تحسب وتفسّر قيمة p في دراسة وبائية؟

Joël Ladner[1,2], Ahmed Ben Abdelaziz[1,3], Réseau Maghrébin: Pédagogie-Recherche-Publication en Sciences de la Santé(RPP2S)

1. Réseau Maghrébin: Pédagogie-Recherche-Publication en Sciences de la Santé (RPP2S)

2. CHU de Rouen Normandie. Département d'Epidémiologie et de Promotion de la Santé, Rouen, France

3. Laboratoire de Recherche LR19SP01. Université de Sousse. Tunisie (Email : ahmedbenabdelaziz.prp2s@gmail.com)

Cette série...

Cette série…

Le Réseau Maghrébin PRP2S et la Rédaction de la revue «La Tunisie Médicale» ont l'honneur de continuer d'une manière régulière, à partir du numéro de février 2021, et pour la deuxième année successive, la série des fiches techniques en épidémiologie, en bio statistique et en rédaction médicale scientifique.

Cette série a eu un grand succès au cours de sa première année d'édition en 2020, comme indique le nombre de téléchargements dépassant significativement celui des articles originaux et illustrant un besoin très manifeste des jeunes chercheurs, au renforcement de leurs capacités en méthodologie de recherche scientifique en sciences de santé, selon une pédagogie centrée sur l'acquisition des compétences pratiques de recherche biomédicale.

En effet, nos fiches méthodologiques décrivent, d'une manière standardisée, les modes d'usage des concepts, des outils et des méthodes utilisés d'une part lors du continuum de la recherche biomédicale scientifique, dès la phase conceptuelle jusqu'à la phase rédactionnelle et d'autre part lors des différentes phases de la rédaction médicale scientifique, depuis l'étape de la recherche documentaire jusqu'à l'étape de la communication médicale scientifique.

Cette série est rédigée par les experts du Réseau Maghrébin PRP2S, en méthodologie de recherche, exerçant dans les universités du Grand Maghreb et les facultés sœurs au Nord de la Méditerranée. Chaque fiche répond à trois questions essentielles (Quoi ? Pourquoi ? Comment) du concept étudié, en se basant sur un article publié dans la revue «La Tunisie Médicale».

Le coordinateur de la série «Fiches Méthodologiques»

Professeur Ahmed Ben Abdelaziz (Président du Réseau Maghrébin PRP2S)

Email : ahmedbenabdelaziz.prp2s@gmail.com

Série des Fiches méthodologiques
Sommaire

Correspondance
Joël Ladner
Email: joel.ladner@univ-rouen.fr

ETUDE DE CAS

La revue «Tunis Med» a publié en 2009 un article original intitulé «Les troubles de la personnalité dans un groupe de consultants en psychiatrie : aspects généraux et caractéristiques comparatives du Cluster B » (1). L'objectif de ce travail était de décrire, au sein d'un groupe de consultants en psychiatrie au CHU Hached de Sousse (Tunisie), entre janvier 2000 et décembre 2004, les caractéristiques des patients présentant un trouble de la personnalité, et de préciser les particularités du cluster B, comparé aux deux autres clusters. Parmi les résultats de cette étude:

1. Les troubles de la personnalité ont été notés chez 6% des consultants.
2. Les personnalités du Cluster B étaient les plus fréquentes (54,7%).
3. Les troubles addictifs et somatoformes étaient plus fréquents dans le cluster B, les troubles anxieux dans le cluster C et les troubles psychotiques dans le cluster A.

Ci-dessous un extrait du tableau 2 de cet article, intitulé : «*Comparisons of scociodemographic features and medical history between cluster B and the two other clusters»*.

	Cluster A	Cluster B	Cluster C	p
Mean âge	40.2 ± 12	30 ± 9.6	33.1 ± 10.2	p1=0.001 p2=NS
Sex ratio	3.6	0.8	0.6	p1=0.018 p2=NS
High school or more	7 (50%)	39 (49.3%)	24 (75%)	p1=NS p2=0.011

p1 : cluster B vs cluster A. p2 : cluster B vs cluster C.

QUIZ

Répondre aux deux questions suivantes:

1 - Concernant la comparaison du sex-ratio, quelle(s) est (sont) la(es) réponse(s) exacte(s) ?

A. Il est significativement plus élevé dans le cluster A que dans le cluster B

B. Pour la comparaison entre les clusters B et C, p=0.45

C. Les valeurs de «p» ont été obtenues par un test du Chi^2 ou un test de Fisher exact

2 - Concernant la valeur de «p» pour la variable *«high school or more»* entre les clusters A et B, elle peut prendre la(es) valeur(s) de:

A. 0,04

B. 0,06

C. 0,86

3 - Quel(s) est (sont) le(s) test(s) statistique(s) utilisé(s) pour comparer les âges ?

D. ANOVA

E. Test t de Student

F. Test de Man et Whitney

INTRODUCTION

Les tests statistiques sont les compagnons de fortune (ou d'infortune) des articles scientifiques. Pour certains chercheurs, il est pratiquement impossible de se passer de tests statistiques qui font très souvent une condition de crédibilité à toute recherche; une marque de qualité scientifique leur est associée. Des tests statistiques simples, tels que le *Chi-carré* ou le test *t de Student* seront très fréquemment utilisés. Au final, toutes ces techniques statistiques pour se lancer à la quête du Graal, le (fameux) «*petit p*» significatif demeure fortement ancré dans la démarche de recherche. En médecine, comme dans d'autres disciplines scientifiques, un consensus international s'est établi pour considérer une différence significative, si la valeur de «p» est <0,05, c'est-à-dire si le hasard a moins de 5 chances sur 100 d'expliquer les différences observées. Ce seuil est totalement arbitraire, il est devenu mythique car la majorité des chercheurs attendent (plus ou moins fortement) des résultats significatifs; nous savons tous qu'il est plus aisé de publier un papier avec des «p» significatifs que non. En 1925, Fisher suggéra de fixer le seuil de significativité à 5%. L'histoire raconte que ce seuil à 5% a eu sa préférence, car il percevait 5% de royalties sur ses publications.

LE PETIT «p»: POURQUOI ?

Des connaissances statistiques simples permettent de pratiquer des tests statistiques usuels, sans solliciter l'aide du professionnel de l'art, le biostatisticien ou l'épidémiologiste. Mais des connaissances élémentaires sont indispensables dans une démarche de Lecture Critique d'Article et d'Evidence Based Medecine (EBM).

LE PETIT «p»: C'EST QUOI ?

La «p value» est la pierre angulaire des tests statistiques. La règle est simple, car la valeur du «p» a un gold standard: le seuil de significativité. Il est très souvent choisi pour une valeur de «p» inférieure ou égale à 0,05 (2). Quelle lecture de la valeur de «p» issue d'un test statistique ?:

- Soit p>0,05: la différence n'est pas significative, on ne peut pas conclure à une différence.
- Soit p≤0,05: la différence est significative, le risque pris est précisé, sa valeur est appelée degré de signification.

Il n'existe pas, en effet, de différence fondamentale entre p=0,06 (non significatif) et p=0,04 (significatif), même si ces deux valeurs apportent une information scientifique. C'est pour cela, qu'il est recommandé dans un article d'indiquer la valeur de «p» et d'éviter de mettre NS (Non Significatif). La valeur de «p» ne doit pas être jugée ex abstracto, mais replacée dans son contexte, en se rappelant bien que la signification statistique est différente de la signification clinique.

LE PETIT «p»: COMMENT ?

Le principe général des tests statistiques (tests d'hypothèse) et le «p» obtenu par ces tests reposent sur la formulation d'une hypothèse nulle (H_0), que l'on cherche à rejeter au profit d'une hypothèse alternative (H_1). L'hypothèse nulle sous-tend l'absence de différence entre les deux (ou plus) échantillons. Rejeter H_0 c'est accepter, avec un risque d'erreur, qu'il existe une différence significative entre les échantillons.

QU'APPELLE-T-ON RISQUE DE PREMIÈRE ET DE DEUXIÈME ESPÈCE ?

L'erreur α (ou risque de première espèce) consiste à conclure à tort qu'il y a une différence (par exemple : traitement meilleur, facteur de risque d'une pathologie), alors que le hasard est responsable des différences observées (fluctuation d'échantillonnage).

L'erreur β (ou risque de deuxième espèce) consiste à conclure à tort qu'il n'y a pas de différence, alors qu'en réalité, il en existe une. En définitive, l'hypothèse nulle est celle que l'on ne veut pas, car elle exprime souvent l'absence de différence. Par contre, l'hypothèse alternative est celle que l'on aimerait accepter.

La puissance d'un test statistique est égale à 1- β. C'est la probabilité de rejeter H_0 quand elle est fausse (et H_1 est vraie). Il se peut, en fait, qu'une vraie différence existe, mais elle n'est pas retrouvée car la puissance de l'étude est insuffisante, en rapport avec un échantillon (nombre de personnes incluses) trop petit. Il arrive aussi que des personnes (patients) soient perdues de vue (ou ayant refusé l'inclusion), entrainant alors une perte de puissance. Il est donc recommandé d'anticiper cette perte de suivi et d'augmenter la taille de l'échantillon calculée, en incluant n% de patients en plus.

Quels sont les facteurs qui influencent la puissance d'un test statistique ?

Ils sont au nombre de quatre : la taille d'échantillon, la différence à montrer, le niveau de significativité (a) et la

variabilité des observations. L'évaluation de la validité d'une hypothèse statistique se fait au moyen d'un test effectué à priori sur des données issues, d'un échantillon à priori représentatif de la population étudiée. Le test statistique est une démarche qui permet d'aboutir au rejet ou au non-rejet de l'hypothèse nulle. Le choix du test statistique, puis son utilisation correcte, nécessite une démarche adaptée et rationnelle, fondée sur la connaissance de la statistique.

Six questions à se poser avant de faire un test statistique et d'obtenir un «p»:

1. Type de variable étudiée: qualitative ou quantitative?
2. Quelle est la distribution de l'échantillon (normale, binomiale etc.)?
3. Petit ou grand échantillon?
4. Test paramétrique ou non paramétrique?
5. Données appariées ou non?
6. Bilatéralité ou unilatéralité du test?

Des logiciels de plus en plus facilement téléchargeables (certains disponibles en Open Access gratuitement) permettent un accès aisé à des outils statistiques de plus en plus complexes, même si parfois la maitrise de ces outils par des utilisateurs non avertis est souvent approximative. Il faut savoir se poser la question: quelle est la validité du «p» retrouvée ? Sachant que trop souvent, les chercheurs font confiance aveuglément à la «boîte noire» et son «fameux p», sans une connaissance suffisante des outils statistiques. Si aujourd'hui, la connaissance exacte des formules des tests n'est plus nécessaire, une connaissance appropriée des conditions et des règles de leur utilisation est obligatoire.

CONCLUSION

Une étude publiée dans le JAMA en 2016, montrait que 96% des articles publiés entre 1990 et 2015 affichaient des résultats statistiquement significatifs, avec un p≤0,05 (3). En 2018, 72 statisticiens ont proposé d'abaisser le seuil de significativité de p, en le passant à 0,005 (encadré 1), soit moins d'une chance sur deux cents que la différence retrouvée soit liée au hasard. Le très grand avantage de cette (r)évolution serait la disparition des «p» trop souvent faussement positifs (4, 5). Quand les statistiques classiques (appelées fréquentistes) donnent des résultats ambigus («p» compris entre 0,05 et 0,005), il est préférable de recourir à une analyse utilisant la méthode Bayésienne.

Encadré 1:
La valeur de «p»: vers un nouveau seuil ?

«Pour les milieux de recherche qui continuent de se fier à la vérification de l›importance des hypothèses nulles, la réduction à 0,005 du seuil de la valeur «p» pour les déclarations de nouvelles découvertes est une mesure réalisable qui améliorera immédiatement la reproductibilité. Les résultats qui n›atteignent pas le seuil de signification statistique (quelle qu›elle soit) peuvent toujours être importants et mériter d›être publiés dans des revues de premier plan s›ils abordent des questions de recherche importantes avec des méthodes rigoureuses. Cette proposition ne doit pas être utilisée pour rejeter les publications de nouvelles découvertes dont 0,005 <p< 0,05 est correctement étiqueté comme une preuve suggestive. Nous devrions récompenser la qualité et la transparence de la recherche »
(Traduction issue du blog www.redactionmedicale.fr).

L'essentiel à retenir

- **Les résultats des études sont la base de l'«Evidence-Based Medicine». Ces dernières guident les stratégies thérapeutiques et les traitements avec des conséquences essentielles pour les patients, mais aussi pour la Santé Publique (dépenses de santé par exemple). C'est pourquoi, la significativité des études est un point crucial de la recherche en santé.**
- **La détermination de la valeur de «p» est une étape importante dans l'analyse des résultats d'une étude épidémiologique.**
- **Elle est technique, car elle nécessite des connaissances et des compétences pour utiliser le test statistique approprié, qui permettra de donner la valeur de «p».**
- **Le plus souvent, le seuil de significativité est fixé à 5% (0,05), mais il peut être aussi déterminé pour un autre seuil.**
- **La valeur de ce seuil est aujourd'hui débattue, car il ne permet pas toujours une distinction claire entre significativité et non-significativité, la frontière est tenue.**

Réponses aux questions

Question 1 : A, C
Question 2 : B, C
Question 3 : B, C

RÉFÉRENCES

1. El Kissi Y, Ayachi M, Ben Nasr S, Mansour A, Ben Hadj A. Personality disorders in a group of psychiatric outpatients: general aspects and cluster B characteristics. Tunis Méd 2009; 87: 685-9.
2. Schwartz D. Méthodes statistiques à l'usage des médecins et des biologistes. 3ème edition. Paris: Flammarion; 1969
3. Chavaralias D, Xalalch JD, Ting Li, et al. Evolution of reporting p values in the Biomedical literature. JAMA 2016; 1235:1141-8.
4. Harrington D, D'Agostino RB Sr, Gatsonis C et al. New guidelines for statistical reporting in the Journal. NEJM 2019; 381: 285-6.
5. Benjamin DJ, Berger JO, Johannesson M et al. Redefine statistical significance. Nat Hum Behav 2018 ; 2 : 6–10.