Looking at *p* Values: Did the Study Results Happen by Chance Alone?

By
M. Gail Woodbury
PhD BScPT
and
Janet Kuhnke
BSN MS ET

INTRODUCTION

This article outlines the role of *p* values in research. It presents a streamlined discussion of a complex topic, with the aim of supporting wound care practitioners in reading research articles effectively. We recommend two easy-to-read books, *Medical Stats Made Easy*[1] and *PDQ Statistics*,[2] to help you increase your ability to read and use research effectively. A relevant research text might also be of use.

Wound care practitioners and specialists read research articles – the background, methods used, sample size, analyses, results, discussion and interpretation – looking for implications for practice. Scientific studies in the literature often contain a statement such as the following, quoted from the Canadian leg-ulcer care community study, referring to before and after implementation of an evidence-based service: "The proportion of daily visits…dropped from 38% to 6% (Pearson χ^2 test 60.1, $p<0.001$)."[3] Even if you know that the *p* value is **significant**, what does that actually mean? If the *p* value had not been significant (i.e. greater than 0.05), what would that mean?

The *p* value refers to **probability**. Probability ranges from 0 to 1.00. By convention the *p* value is often set at 0.05, which means that a result or statistic as large as that found in the sample would have occurred by chance alone five times out of 100.[4]

Samples and hypotheses

Whenever a research study is done, a sample is used to represent the target population the researcher wants to understand; it is rarely possible to test the whole population. The sample provides an **estimate**, such as a mean or proportion, of the true value of the mean or proportion for the target population.

The investigator starts a study by stating the **null hypothesis**, or H_0, as well as the alternative or **research hypothesis**, H_A or H_1. Let us consider the example of comparing two products or therapeutic approaches, A and B, and determining a mean (i.e. average) value for each. The null hypothesis (H_0) states there is no difference between the means of the products, i.e.:

$A = B$

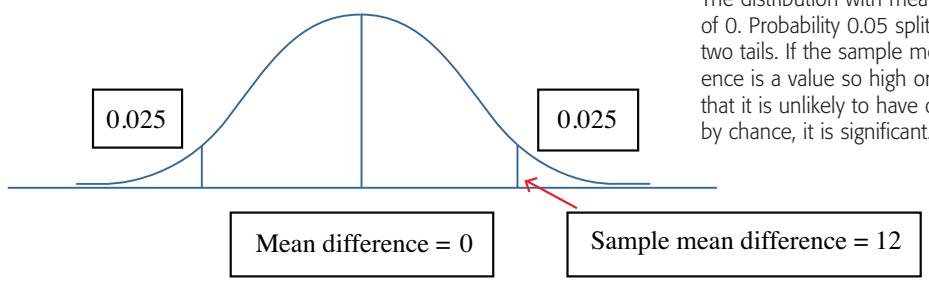
Or $mean_A = mean_B$

The alternative hypothesis (H_A) states there is a difference, i.e.:

$A \neq B$

Or $mean_A \neq mean_B$

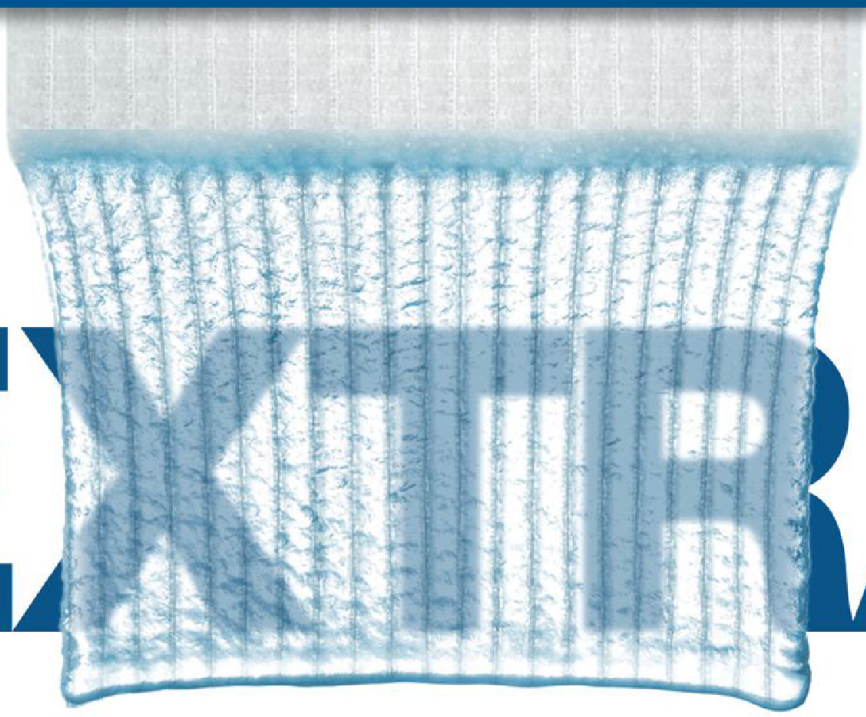
FIGURE 1



The distribution with mean difference of 0. Probability 0.05 split between two tails. If the sample mean difference is a value so high or so low that it is unlikely to have occurred by chance, it is significant.

NEW AQUACEL® EXTRA™

Hydrofiber® dressing with strengthening fiber



EXTRA

STRENGTH · ABSORBENCY · CONFIDENCE

More to love about AQUACEL® dressings

NEW AQUACEL® EXTRA™ dressing:

- 9x stronger[1a]
- 39% more absorbent[1a]
- Manages a wide range of exudate levels

ConvaTec CRC Customer Relations Center Centre des relations avec la clientèle

For more information, please call our Customer Relations Center (Registered Nurses on staff) at 1-800-465-6302, Monday through Friday, 8:00 AM to 6:00 PM (EST), or visit our Web Site at www.convatec.ca

[a]As compared to original AQUACEL® dressing.

Reference: 1. Preliminary assessment of the physical properties of AQUACEL® EXTRA vs AQUACEL® & DURAFIBER™. *Scientific Background Report*. WHRI3461 TA214. 2011, Data on File, ConvaTec Inc.

Hydrofiber® TECHNOLOGY

AQUACEL® Dressings
TRIED. TRUE. TRUSTED.™

ConvaTec

After the study data have been collected and the mean values calculated, the means are compared by hypothesis testing (in this instance, the hypothesis test would be a Student's t-test to compare 2 mean values). Based on the *p* value associated with the hypothesis test, a decision has to be made to either reject the null hypothesis or not reject the null hypothesis. (Statisticians do not talk about accepting the null hypothesis, only **not rejecting** it, because a non-significant hypothesis test does not provide proof of the validity of the null hypothesis.)

Error and significance

When the sample data indicate a big difference such that the *p* value is small (i.e. a value less than 0.05) then the null hypothesis is rejected and there is a risk of type I error. The extent of risk of **type I error** depends upon the significance level of the test, which is set by convention at 0.05. The **significance level** is also called the alpha level and represented by α. When the data indicate there is no difference between groups and the null hypothesis is not rejected, then there is a risk of **type II error**, the amount of which is set by the beta level (represented by β). The statistical power of the hypothesis test, denoted $1 - \beta$, is the probability of finding a significant difference when there is a true difference in the population. These concepts are illustrated in Table 1, which is found in many statistical textbooks.[5–7]

TABLE 1

Type I and type II error

Decision based on sample data and hypothesis test	Reality	
	H_0 is true	H_0 is not true
Not statistically significant (do not reject H_0)	Correct conclusion $1 - \alpha$	Type II error β
Statistically significant (reject H_0)	Type I error α	Correct conclusion $1 - \beta$ (power)

Level of significance indicates the level of risk the investigator is prepared to take in falsely finding significance (i.e. type I error). The researcher must consider how serious the consequences would be if a difference were found by chance that was not true in the population. When the alpha level is set at 0.05, as it often is, the researcher is willing to accept that 5 times out of 100, significance will be found by chance alone. When the alpha level is set at 0.1, the researcher is prepared to accept that risk 10 times out of 100 (i.e. 10% of the time).

Type I and II error

- Type I error refers to the error or risk of rejecting a true null hypothesis, i.e. finding significance where there is not a true difference.
- Type II error refers to the error or risk of not rejecting a false null hypothesis, i.e. not finding significance where there is a true difference.

Let us consider the distribution in Figure 1, with the difference between the means being zero. Think of it as the distribution of all possible samples of the same size from the target population. Our study with the same sample size would be just one example. Let us pretend the data indicate a mean difference of 12 and the hypothesis test has a probability of p=0.015. This sample mean would be unlikely under the null hypothesis and would fall at the extreme end of the distribution, as shown. This result is unlikely to have occurred by chance, so we conclude that we can reject the null hypothesis and say the result is significant.

Conclusion

In summary, *p* values indicate the probability or risk of error – the likelihood that the difference occurred by chance alone. Therefore, when we read the results of a study and they are significant, we know there is still a risk that the result occurred by chance. This illustrates the importance of interpreting evidence from a body of literature as opposed to from just a single study. It also highlights the importance of systematic reviews with meta-analysis for making important clinical decisions.

References

1. Harris M, Taylor G. *Medical Stats Made Easy*. London, UK: Matrin & Dunitz; 2003.
2. Norman GR, Streiner DL. *PDQ Statistics*, 3rd ed. Hamilton: BC Decker Inc.; 2003.
3. Harrison MB, Graham ID, Lorimer K, Friedberg E, Pierscianowski T, Brandys T. Leg-ulcer care in the community, before and after implementation of an evidence-based service. *CMAJ*. 2005; 172(11):1447-1452.
4. Last JM. *A Dictionary of Epidemiology*, 2nd ed. Toronto: Oxford University Press; 1988: p.1450.
5. Lobiondo-Wood G, Haber J. *Nursing Research in Canada: Methods and Critical Appraisal for Evidenced-Based Practice*, 2nd ed. Toronto: Mosby Elsevier; 2009.
6. Colton T. *Statistics in Medicine*. Boston: Little Brown and Company; 1974.
7. Elzey F. *An Introduction to Statistical Methods in the Behavioral Sciences*. Monterey, CA: Brooks/Cole Publishing Company; 1976.

Interprétation des valeurs *p* : les résultats de l'étude sont-ils uniquement le fruit du hasard?

PAR
M. GAIL WOODBURY
PHD BSCPT
ET
JANET KUHNKE
BSN MS ET

INTRODUCTION

Le présent article porte sur le rôle des valeurs *p* dans la recherche. Il explique un sujet complexe en termes simples pour aider les praticiens du soin des plaies à bien comprendre les comptes rendus de recherches. Les auteurs recommandent au lecteur deux livres faciles à lire, intitulés *Medical Stats Made Easy*[1] et *PDQ Statistics*[2], qui lui permettront de mieux comprendre les comptes rendus et d'utiliser plus efficacement les renseignements qu'ils contiennent. Un livre pertinent sur la recherche pourrait aussi être utile.

Les praticiens et spécialistes du soin des plaies lisent les comptes rendus de recherches – contexte, méthodes employées, taille des échantillons, analyses, résultats, discussion et interprétation – pour comprendre comment ils modifient la pratique. Les études scientifiques publiées contiennent souvent un énoncé comme le suivant, tiré de l'étude canadienne menée en milieu communautaire sur le soin des ulcères de jambe et portant sur la différence entre avant et après la mise en place d'un service fondé sur des données probantes : « La proportion des visites quotidiennes [...] est passée de 38 % à 6 % (test du chi carré de Pearson : 60,1; $p < 0{,}001$). »[3] On sait que la valeur *p* est **significative**, mais qu'est-ce que ça veut dire exactement? Et si la valeur *p* n'avait pas été significative (soit supérieure à 0,05)?

La valeur p désigne la **probabilité**. La probabilité va de 0 à 1,00. Par convention, la valeur p est souvent fixée à 0,05, ce qui veut dire qu'un résultat ou une statistique aussi important que celui retrouvé dans l'échantillon aurait pu être uniquement le fruit du hasard cinq fois sur 100[4].

Échantillons et hypothèses

Les études de recherche portent toujours sur des échantillons qui représentent la population cible que les chercheurs désirent comprendre, car il est rarement possible d'évaluer l'ensemble de la population. L'échantillon donne une **estimation**, par exemple une moyenne ou une proportion, de la valeur réelle de la moyenne ou de la proportion dans la population cible.

Au début d'une étude, le chercheur formule **l'hypothèse nulle**, soit H_0, et **l'hypothèse de recherche** alternative, soit H_A ou H_1. Prenons l'exemple de la comparaison entre deux médicaments ou démarches thérapeutiques, A et B, et de la détermination d'une valeur moyenne pour chacune. Selon l'hypothèse nulle (H_0), il n'y a pas de différence entre les moyennes des produits :

$A = B$

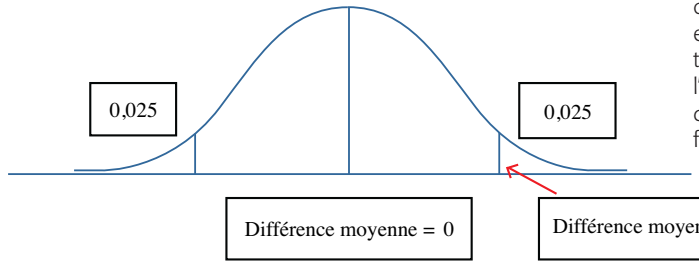
ou $\text{moyenne}_A = \text{moyenne}_B$

Selon l'hypothèse alternative (H_A), il y a une différence :

$A \neq B$

ou $\text{moyenne}_A \neq \text{moyenne}_B$

FIGURE 1



Distribution avec différence moyenne de 0. Probabilité de 0,05 partagée entre les deux queues de la distribution. Si la différence moyenne dans l'échantillon est si grande ou si petite qu'elle est peu susceptible d'être le fruit du hasard, elle est significative.

Une fois les données recueillies et les valeurs moyennes calculées, on compare les moyennes au moyen d'un test d'hypothèse (dans le cas qui nous intéresse, le test d'hypothèse serait un test de Student pour comparer deux valeurs moyennes). Selon la valeur *p* associée au test d'hypothèse, on décide de rejeter ou de ne pas rejeter l'hypothèse nulle. (Les statisticiens ne parlent pas d'acceptation, mais plutôt de non-rejet de l'hypothèse nulle, parce qu'un test d'hypothèse non significatif ne prouve pas la validité de l'hypothèse nulle.)

Erreur et signification

Quand les données d'échantillon indiquent qu'il y a une grande différence, donc que la valeur *p* est faible (c'est-à-dire inférieure à 0,05), l'hypothèse nulle est rejetée et il y a un risque **d'erreur de type I**. L'importance du risque d'erreur de type I dépend du **seuil de signification** du test, qui est fixé par convention à 0,05. Le seuil de signification est aussi appelé seuil alpha (représenté par α). Quand les données indiquent qu'il n'y a pas de différence entre les groupes et que l'hypothèse nulle n'est pas rejetée, il y a un **risque d'erreur de type II**, dont l'importance dépend du seuil bêta (représenté par β). La puissance statistique du test d'hypothèse (soit 1 - β) est la probabilité d'observer une différence significative quand il y a une différence réelle dans la population. Ces concepts sont illustrés au tableau 1, qu'on retrouve dans de nombreux manuels de statistique[4–7].

Le seuil de signification indique l'importance du risque que la signification observée soit fausse (soit une erreur de type I) que le chercheur est prêt à accepter. Le chercheur doit se demander si les conséquences seraient graves si une différence observée était le fruit du hasard, et donc n'était pas réelle. Un seuil alpha fixé à 0,05, ce qui est souvent le cas, indique qu'un chercheur est prêt à accepter que cinq fois sur 100, la signification observée soit uniquement le fruit du hasard. Un seuil alpha de 0,1 indique que le chercheur est prêt à accepter que la signification soit uniquement le fruit du hasard dix fois sur 100 (soit dans 10 % des cas).

Examinons la distribution présentée dans la figure 1, selon laquelle la différence entre les moyennes est de zéro. Cette distribution est comme celle de tous les échantillons possibles de la même taille dans la population cible. Notre étude, qui porte sur un échantillon de la même taille, ne serait qu'un exemple. Supposons que les données indiquent que la différence est de 12 et que le test d'hypothèse donne une probabilité (valeur *p*) de 0,015. Cette moyenne d'échantillon est peu probable selon l'hypothèse nulle et se situe à la limite extrême de la distribution, comme le montre la figure. Comme un tel résultat est peu susceptible d'être le fruit du hasard, on peut conclure qu'il convient de rejeter l'hypothèse nulle et que le résultat est significatif.

TABLEAU 1

Erreur de type I et erreur de type II

Décision fondée sur les données d'échantillon et le test d'hypothèse	Réalité	
	H_0 est vraie	H_0 n'est pas vraie
Pas de différence statistiquement significative (non-rejet de H_0)	Bonne conclusion 1 - α	Erreur de type II β
Différence statistiquement significative (rejet de H_0)	Erreur de type I α	Bonne conclusion 1 - β (puissance)

Erreur de type I et erreur de type II

- L'erreur de type I désigne le risque de rejeter une hypothèse nulle vraie, c'est-à-dire d'observer une différence significative quand il n'y en a pas en réalité.
- L'erreur de type II désigne le risque de ne pas rejeter une hypothèse nulle fausse, c'est-à-dire de ne pas observer de différence significative quand il y en a une en réalité.

Conclusion

En résumé, la valeur *p* indique la probabilité ou le risque d'erreur, soit la probabilité que la différence soit uniquement le fruit du hasard. Par conséquent, quand on lit dans un compte rendu que les résultats d'une étude sont significatifs, on sait qu'il est quand même possible que ces résultats soient le fruit du hasard. C'est pourquoi il est important d'examiner les données probantes d'un ensemble d'études, et pas seulement d'une seule. C'est aussi pourquoi les importantes décisions cliniques doivent être fondées sur des examens méthodiques et des méta-analyses.

Références (voir page 8)