

The significance of significance

By Cathy Carter-Snell, RN MN

Abstract

In this article, the concepts of statistical and clinical significance are reviewed. Implications of significance levels for error and power are discussed as well as issues in interpreting significance.

You find a research article in your nursing journal which deals with a treatment you would like to implement in your unit. After reading the article, however, you become discouraged. In the conclusions section, the authors describe the difference between the regular and experimental treatment as not significant. Should you be discouraged? What might explain this lack of significance?

In order to answer this question, we need to first briefly discuss probability theory. Health care research has traditionally relied on probability theory, in which a study is designed with a preset (apriori) significance level, also known as an alpha level. Once the research is conducted, a probability is calculated and, if it is less than the preset level, the study results are considered statistically significant. This then raises the issue of whether the results are clinically significant.

We began this research series in a previous article by exploring the concept of evidence-based practice as an overall clinical goal. Understanding significance is important to interpreting research evidence. The purpose of this article is to explore the concepts of apriori alpha levels and their implications for error in research studies, followed by a discussion of clinical and statistical significance.

Understanding significance levels

When a research study is designed, the researcher must first decide at what point their potential findings will be considered statistically significant. Statistical significance is achieved when the probability of obtaining a result is smaller than what would be anticipated by chance. Statistical testing examines the variability in a sample and the probability that the results obtained in the sample population are more than expected from chance or random variation. Consider the results if you flip a coin. Although we anticipate a 50% chance of heads or tails, we know that we never get exactly that with 100 tosses. If we got 42 heads and 58 tails, for instance, measuring statistical significance would calculate if the difference between the two

is more than we would expect with chance. If it is less likely or probable, then it is considered statistically significant. A typical significance (alpha) level chosen as a cut-off for a study would be 0.05, or chance of finding a difference by error five times out of 100. If an obtained probability is smaller than 0.05, such as $p=0.023$, it would be considered statistically significant.

It should be noted in probability theory that if something is statistically significant, it does not “prove” there is a difference. We only know that there is probably a difference. Unfortunately, if huge samples of research subjects are used it may also result in falsely positive significant results. This is why multicentre trials or nationwide surveys only use a portion of the population, rather than trying to use as many as possible. If multiple tests are performed on the same data, eventually it is possible to also have falsely significant results. For this reason, if more than one statistical test is used, some may choose more stringent significance levels or use other means to ensure it will be possible to detect a difference through their research design. One simple method to control for this risk is the “Bonferonni split”. In this method, the desired significance level is divided by the number of tests. If two tests were to be performed and a significance level of 0.05 was desired, the actual cut-off would be a probability of less than 0.025 for each test before it would be considered significant. This is controversial, however, as some researchers believe you should just publish the obtained probability and let the reader determine if it is significant.

Lack of significance does also not prove there is no difference. It may mean that there is probably not a difference. It could also mean that there were problems with the research study. Examples include research tools or methods which were not sensitive enough to detect a difference, too small a sample size, or less sensitive statistical tests in relation to the type of data obtained. In a later article, we will discuss the issue of types of tests and levels of data. At its simplest, the more precisely an item is measured, the more sensitive the statistical tests to noting a difference. If we measure pain control in terms of “poor”, “adequate”, and “excellent”, this is less sensitive than a pain scale measured in millimetres or relying on the amount of sedation equivalents used. Sample size, the variability or sensitivity of tests to a difference, and the type of test used all affect the ability of the study to detect a true difference. This is also known as “power”.

A lack of significance could also be due to random differences in group composition. If there is something abnormally distributed between groups which is unexpected, it could alter the ability to see a difference. Consider a study in which groups of abdominal surgery patients were compared for pain relief and post-operative outcomes, and were randomly assigned to either patient-controlled analgesia (PCA) or intramuscular (IM) analgesia groups. It was anticipated there would be a significant difference in pain relief with the PCA group compared to the IM group, yet there was not. In comparing the groups after random assignment, there were more patients who had received hysterectomies in the PCA group than the IM group. Most of these were not the newer laparoscopy type, but the invasive abdominal type. It was noted that the PCA group had a longer length of post-operative stay. The invasive nature of the surgery combined with the psychologic implications of the surgery for the women could contribute to this difference.

Significance, power and confidence

The choice of a significance level influences the chance for random errors in a study. If a significance level of 0.05 is chosen, you are saying that you are willing to accept that five times out of 100 a positive finding will actually be only due to chance and not actually exist. This is called Type I error. It perhaps makes sense to choose a more stringent significance level such as 0.01 or 0.001. This brings further problems,

however. As you decrease the chance of false positives or Type I error, you also decrease the probability of being able to detect a true difference if it exists (the power). This is also known as "power". Power of a statistical test is defined as "the probability that it will yield statistically significant results". Essentially, it is likely that a study will be able to detect a true difference when in fact it exists. It relies on three factors - the size of the sample, the alpha level set, and the type of statistical test being used. A more stringent significance level may greatly decrease power and yield a non-significant result. One way this is counteracted is to increase the number of subjects in the study. Many researchers estimate their sample size for a study based on power for given tests. Unfortunately, if subjects are lost from studies or the estimated effects are not as large as anticipated, the obtained power of a test at the end of the study is often very low.

In one study of nursing research articles published in the late 1980s and early 1990, it was shown that most nursing studies had low power particularly due to small sample sizes. This is a reality in clinical studies with only small groups of patients available to us. The lack of adequate samples is changing slowly over time as nurse researchers begin to share or access large datasets such as Statistics Canada data or other researchers' data, to participate in multicentre trials, and to conduct studies of data across



THE SCARBOROUGH HOSPITAL


The Scarborough Hospital (TSH) is a multi-site urban community hospital that delivers innovative, high quality patient care, advocates for our community's health and wellness issues, and is a leader in research, teaching and learning. We are currently offering full and part time opportunities for:

EMERGENCY REGISTERED NURSES

We believe **The Scarborough Hospital** is the best place to practice nursing and build your career. In addition to our excellent services, the hospital is in the process of creating a new state-of-the-art Emergency and Critical Care Wing, which will house a greatly expanded emergency department, intensive and coronary care units and diagnostic imaging facilities. We offer: In-house degree and specialty certificate programs, fitness centres offering a variety of programs with an adjacent parkland; an extensive Nursing Orientation program; Sponsorship to a 12 week critical care course at Seneca college; dedicated Nurse Educators, Managers and Preceptors; relocation assistance; continuing educational assistance and flexible scheduling options.

Contact Lucy Sangregorio at lsangregorio@tsh.to or (416) 431-8200 ext.6137 to find out more.

**The Evidence is Clear:
Expand Your Forensic Knowledge ONLINE!**




Become a professional in the growing field of forensics and meet new challenges where the worlds of health, science and law meet.


Earn your Certificate of Achievement in Forensic Studies from a leading Canadian college through web-based learning. Complete ANY four of these courses (12 credits*) which best fit your learning needs:

- FORE 4401 - Forensic History, Risk Populations and Issues
- FORE 4403 - Forensic Psychiatric and Correctional Populations
- FORE 4405 - Victims of Violence
- FORE 4407 - Forensic Science
- FORE 4409 - Expert Witness Testimony
- FORE 4411 - Crime Scene Investigation & Evidence
- FORE 4413 - Sexual Assault Examination and Intervention Theory
- FORE 4415 - Sexual Assault Examination and Intervention Practicum

(* courses may be transferable)



Call toll-free in North America:
1-888-240-7201
E-mail: fore@mtroyal.ca
www.mtroyal.ca/forensic



MOUNT ROYAL COLLEGE
Faculty of Continuing Education & Extension
CALGARY, CANADA

studies. This aggregation of data is known by a few terms, most commonly meta-analyses for quantitative data and meta-synthesis for qualitative data. These concepts will also be explored in a later paper.

Given the impact of statistical significance on power, we need to consider the impact of the results and the risks involved in order to choose a significance level. If there is a high risk if there is a false positive, such as a potentially toxic medication, a more stringent significance level would generally be desired. In order to have a reasonable chance of finding a difference, a power level of at least 0.80 is desired. This is an 80% chance of finding a difference if it does really exist. The more stringent level will increase the sample size greatly, which helps explain why so many drug companies have moved to multicentre trials. On the other hand, in a study of pain relief with one well-accepted treatment compared to another common treatment, a less stringent level may be chosen and fewer patients will be required to conduct the study.

Clinical significance

Tests relying on statistical significance have been criticized in many areas, as they may miss clinical significance. Clinical significance is present if the findings are meaningful clinically. Consider the results of the PCA and IM study again. The IM patients remained an average of 5.32 days in hospital compared to an average of 5.9 days, or 11 hours longer, in the PCA group. While not statistically significant with a two-tailed alpha level of 0.05, it was potentially clinically significant. An additional 11 hours could mean another full day's stay in hospital, incurring more costs. This is supported by an increased delay in the time to first ambulation in the PCA group which again was not statistically significant, but could have contributed to the longer length of stay.


Interpreting significance

Reports of research studies generally include statements of probability and may or may not include a discussion of the "cut-off" alpha level selected by the researcher. Study conclusions and tables usually focus on the presence or absence of statistical significance. This is also a source of controversy in the research world. Some argue that, rather than choosing an arbitrary "cut-off" level, we should be seeking the probable range in which the true value lies by reporting a range of values in which we are confident that the true value likely exists. This is known as a "confidence interval". The standard used is a 95% confidence interval, which means a range of values in which we are 95% confident that the true value lies. The width of this interval varies with the amount of measurement error in the study. Yet another approach used by researchers is to simply report the probability obtained and let readers determine for themselves whether it is significant in their eyes.

We generally discourage implementing findings from one study with significant results. It is recommended you look for other supporting studies. There is a note of caution, however. Finding a number of studies with similar results

may not always increase certainty about the findings, or give a true picture of the issue. There have been publishing biases noted in some instances. Many research journals tend not to publish studies with non-significant findings, or only publish those which the peer reviewers favour. This does not give a fair representation of the variability of results. An additional problem in publishing is that some researchers publish the results of the same studies in two or more journals, giving the impression that there are more articles favouring the results. You will have to read carefully and look closely at the authorship to detect this problem.

Conclusion

Now let's go back to that non-significant study you were looking at in the beginning. You will have to ask yourself if there were factors in your experience which would have interfered with their ability to detect a difference such as inappropriate methods or a poor question. Was their sample size reasonably large, such as at least 20 to 40 subjects per study group? Is there a power level reported with the completed data which is acceptable, or a 95% confidence interval used? Was the cut-off level unreasonable given the risk with the study so that power was unreasonably affected? You may also want to see if there were other similar studies in which results were also non-significant. You will also want to consider the clinical significance of the findings. If there is truly no statistically significant difference in treatments, can you then choose the one which is most comfortable for patients or least expensive? If it was statistically significant, will it make an important impact on your practice, or is it only a numerical exercise? While numbers may help, only you can determine the significance of significance in your setting. 

References

- Altman, D.G., Fore, S.M., Gardner, M.J., & Pocock, S.J. (2000). **Chapter 14: Statistical guidelines for contributors to medical journals**. In D.G. Altman, D. Machin, T.N. Bryant, & M.J. Gardner (2nd ed., pp. 171-190). Bristol: BMJ Books.
- Cohen, J. (1977). **Statistical power analysis for the behavioural sciences**. Orlando: Academic Press.
- Gardner, M.J., & Altman, D.G. (2000). Chapter 3: Confidence intervals rather than p values. In B (Eds), **Statistics with confidence** (2nd ed., pp. 15-27). Bristol: BMJ Books.
- Hojat, M., Gonnella, J.S., & Calleigh, A.S. (2003). Impartial judgment by the 'gatekeepers' of science: fallibility and accountability in the peer review process. **Advances in Health Sciences Education: Theory and Practice**, 8(1), 75-96.
- Melander, H., Ahlqvist-Rastad, J., Meiger, G., & Beerman, B. Evidence based medicine - selective reporting from studies sponsored by pharmaceutical industry: Review of studies in new drug applications. **BMJ: British Medical Journal**, 326(7400), 1171-1174.
- Polit, D.F., & Sherman, R.E. (1990). Statistical power in nursing research. **Nursing Research**, 39(6), 365-369.
- Snell, C.C., Fothergill-Bourbonnais, F., & Durocher-Hendriks, S. (1997). Patient controlled analgesia and intramuscular injections: A comparison of patient pain experiences and postoperative outcomes. **Journal of Advanced Nursing**, 25(4), 681-690.