Scientific Inquiry

Hypothesis Testing and Power in Research

Carol P. Vojir
Column Editor: Diane Hudson-Barr

Scientific Inquiry *provides a forum to facilitate the ongoing process of questioning and evaluating practice, presents informed practice based on available data, and innovates new practices through research and experimental learning.*

Diane Hudson-Barr, PhD, RN, is a Clinical Nurse Specialist in the Newborn Critical Care Center at the University of North Carolina Hospitals. Dr. Hudson-Barr brings her expertise on providing family-centered and developmentally-appropriate care and experience as a researcher in neonatal and pediatric pain to editing the Scientific Inquiry column.

Many of the important advances in nursing research are done through hypothesis testing. Like shoes, hypotheses come in pairs. We have a *null* hypothesis, jargon for a hypothesis that proposes that our findings simply are the result of chance processes in the sample and represent no systematic effect in the population, and an *alternative* or research hypothesis, which proposes a plausible and generalizable explanation for the results we expect to find through the research. One example of a null hypothesis might be "There will be no difference in reported pain between subjects who receive an injection through a pain diffusion shield and subjects who receive an injection by the standard technique." The paired alternative hypothesis might be "Subjects who receive an injection through a pain diffusion shield will report less pain than subjects who receive an injection by the standard technique."

This "twosies" thinking process is not unique to research; each of us works from multiple hypothesis perspectives routinely, albeit much more informally. If you come on shift at 6 AM in an inpatient setting, you expect to find your patients in their beds; that's the routine (null) situation. If one patient is not in his/her bed, the most likely explanation is that he/she is in the restroom (alternative), although other explanations may be equally likely. The point is that nurses think this way routinely and informally.

Developing Hypotheses

Because hypotheses are developed in advance of the research, the pairs of hypotheses (at least one pair for each specific aim of the research) help shape the research plan. If we actually wanted to test whether use of a pain diffusion shield results in less pain when patients receive injections, we would first develop a pair of hypotheses like those previously given. It is reasonable to expect that use of a pain diffusion shield will result in less pain for the subjects. Where there is a qualifying adjective (less) on our outcome variable (pain), the alternative hypothesis is called "directional". If experience and/or the literature do not support a qualified alternative hypothesis, one is obliged to use an unqualified or nondirectional hypothesis. In our example, the nondirectional hypothesis would be "There will be a difference in reported pain between subjects who receive an injection through a pain diffusion shield and subjects who receive an injection by the standard technique." Such a hypothesis requires the researcher to look for a difference of more pain (if that were plausible) as well as a difference of less pain.

Specification of null and alternative hypotheses is important because they contain part of the "blueprint" for the statistical analysis. If the null hypothesis is "There will be no difference in reported pain between subjects who receive an injection through a pain diffusion shield and subjects who receive an injection by the standard technique," we know from use of the word "difference" that a statistical procedure that tests differences will be needed. The null hypothesis also tells us that (a) the outcome or dependent variable is reported pain, (b) the manipulated or independent variable is use of the pain diffusion shield during an injection, and (c) we will need two groups of subjects, one group to receive injections through the pain

diffusion shield and one group to receive injections by the standard technique. The alternative hypothesis, "Subjects who receive an injection through a pain diffusion shield will report less pain than subjects who receive an injection by the standard technique," directs us to look for less reported pain from the subjects who received injections through the pain diffusion shield in comparison with pain reported by subjects receiving injections by the standard technique.

Which Hypothesis Explains the Findings?

Unfortunately, we can not just look at the difference in average reported pain between the two groups and know intuitively whether that difference represents chance findings (the situation specified by the null hypothesis) or a real difference in pain that is attributable to use of the pain diffusion shield (the situation specified by the alternative hypothesis). We have to use a statistical procedure (here a *t*-test) along with a statistical "growth chart" to determine which hypothesis is the more likely explanation for our findings. These statistical "growth charts" (significance tables) are very much like the reference charts used to document children's growth. Just as a child's height and weight can be compared with a table of heights and weights for children of similar ages and genders to determine whether the child is growing normally, the difference in average reported pain from a statistical procedure can be referenced to a table of all the outcomes that could occur if the null hypothesis (chance differences) was the underlying cause for those outcomes. Pediatric growth charts tell the nurse in percentiles how a child compares in height and weight with the normal population. Significance tables tell the researcher the probability, called a *p*-value, that the null hypothesis is the better explanation for the study results.

Usually nurses will give advice designed to help children grow normally to the parents of children whose heights and weights are extremely low or extremely high. A somewhat similar process occurs with significance tables. It is typical to decide in research that if the obtained *p*-value is lower than a cutoff probability, called alpha, which the researcher chooses while setting up the research study, the alternative hypothesis is a better explanation than the null hypothesis for the research outcome. Alpha is customarily set at 0.05, that is, only one time in twenty would a result so extreme belong to the set of values associated with the null hypothesis as the correct explanation for the research results.

Once alpha is chosen, the decision about which hypothesis is the better explanation for the results is straightforward; if the *p*-value is less than or equal to alpha, the alternative hypothesis is judged to better explain the results. If the *p*-value is greater than alpha, the null hypothesis is judged to better explain the results. So if we test the hypothesis about the effect on reported pain with use of the pain diffusion shield and the *p*-value that we get from analyzing the data is 0.13 (larger than 0.05), we conclude that the better explanation for these results is chance findings (the null hypothesis); the difference between the groups, even if there is less reported pain in the pain diffusion shield group, is not statistically significant. On the other hand, if our obtained *p*-value is 0.03 (less than 0.05), we conclude that the better explanation for less reported pain is use of the pain diffusion shield; the difference in reported pain between the groups is statistically significant.

Conclusion Errors

p-values are probabilities, though, and the researcher is taking a calculated risk by using the alpha cutoff probability. Mistakes, called conclusion errors, can occur. For example, if our *p*-value was 0.03 and we conclude that subjects who received injections through the pain diffusion shield reported significantly less pain than subjects who received injections by the standard technique, there is a small chance that we have drawn an incorrect conclusion. We would make what is called a type-I error, finding significance in

the sample when it is not present in the population. Conversely, if our *p*-value is 0.13 and we conclude that subjects who received injections through the pain diffusion shield did not report significantly less pain than subjects who received injections by the standard technique, there is a finite chance that we have made a second type of mistake called a type-II error (not finding a difference in the sample when it is present in the population). Typically, conclusion errors are not discovered until someone else repeats the research or questions the findings because they are inconsistent with the body of existing literature.

Estimating and Enhancing Power

No one wants to make intentional mistakes. How can a researcher "stack the deck" in favor of error-free results? Good research design is crucial. Part of that research design process involves consideration of the statistical concept called power and the elements of research design that enhance power.

Power is defined as the probability of finding a statistically significant result in the sample when that result exists in the population. The concepts that underlie the definition and estimation of power, while abstract and circuitous, are straightforward. Although we reference the significance table associated with the null hypothesis being true in the population, theoretically there is a second significance table, one associated with the alternative hypothesis being true in the population. Significance tables associated with the null hypothesis are used in research because they can be worked out easily. Alternative hypothesis significance tables are hypothetical because of the infinite number of significance tables that would be needed to encompass all the possible outcomes associated with the alternative hypotheses from *just one* research study.

Even though an alternative hypothesis significance table is impossible to obtain, we can use data from published articles or pilot studies to estimate the effect size, which is an *inverse* measure of the overlap between the null hypothesis significance table and a plausible alternative hypothesis significance table. The amount of overlap, which is the situation where results from the statistical procedure used have different *p*-values in both null hypothesis and alternative hypothesis significance tables, is important. If overlap is estimated to be large, our chances of finding a significant result in the sample when it exists in the population are severely diminished. It is a little like the interference you experience when you want to talk with one of your two children and the other child becomes jealous and interrupts frequently. Using the null significance table is easiest if there is no "interference" from an alternative significance table that is too similar or overlapping.

We estimate the overlap inversely through the effect size. The form of the effect size depends on the statistical procedure used, but basically the larger the effect size, the smaller the overlap between the significance tables. Researchers can use the effect size, along with other aspects of the study such as alpha (the criterion probability for significance), and the sample size to determine the maximally useful estimate, power. Power is the degree to which the null hypothesis and alternative hypothesis significance tables do not overlap expressed as a probability. The current standard for power in nursing is 0.8.

Power is affected by at least six factors: (1) the alpha cutoff, (2) the nature of the statistical procedure, (3) the reliability of the measures used to quantify the variables, (4) the directionality of the alternative hypothesis, (5) the sample size, and (6) the research or statistical design considerations. We know that if the alpha cutoff is relaxed (jargon for made larger, for example set at 0.10 instead of at 0.05), smaller values of the statistical procedure will be necessary to achieve statistical significance. In this way, power is increased.

Parametric statistical procedures, which have very restrictive assumptions about the data, are more powerful than nonparametric statistical procedures, which do not have the same restrictive assumptions. Reliable measures, which are more precise in assessment, are more powerful than unreliable measures. Directional

(qualified) alternative hypotheses with their focused outcomes are more powerful than nondirectional (unqualified) hypotheses. Larger sample sizes are associated with more power than smaller sample sizes. Finally, certain research and statistical considerations are more powerful than others; for example, in our pain study, sampling subjects who are known to experience more pain with injections might be a more sensitive and powerful test of the effect of the pain diffusion shield than a "regular" sample. We also could choose a more powerful statistical technique for analysis than the *t*-test. These factors are all considered as a part of the research design process.

Suppose we estimate power as a part of planning our study to investigate whether use of a pain diffusion shield results in less pain when patients receive injections. We choose an alpha cutoff probability of 0.05, a directional (qualified) alternative hypothesis, and a sample size of 20 subjects in each injection group — pain diffusion shield and standard technique. From preliminary work we have done, we estimate the effect size to be 0.5 (moderate). From power tables for our parametric *t*-test procedure, power is estimated to be 0.5, which is not very close to 0.8 (but not a disaster either).

What could we do to increase our estimated power? One of the easiest factors to consider is choosing more subjects. To achieve an estimated power of 0.8, we would need 50 subjects in each of the two groups rather than the 20 per group we had originally planned. What would happen if we relaxed (increased) alpha? With 20 subjects per group and an estimated effect size of 0.5, power is estimated to be 0.6. If we both increased the sample size and alpha, a power of 0.8 could be achieved with 36 subjects in each group. Each of these factors has an effect, and both together have more of an effect. The trade-off for these enhancements to power would be in time and resources to recruit more subjects and in the potential consequences of increasing the risk of a type-I error.

Conclusion

In summary, we have discussed hypotheses and power. Points to remember include: (1) hypotheses come in pairs, (2) careful specification of both hypotheses helps in the research design process, (3) the thinking process is simply a more formal version of thinking processes used daily by nurses, (4) deciding about statistical significance is similar to relating information to pediatric growth charts, and (5) the likelihood of mistakes from research findings is decreased by good research design, particularly attention to power estimation.

Carol P. Vojir, PhD
Associate Research Professor
Director, Center for Nursing Research
School of Nursing
University of Colorado at Denver and Health Sciences Center
Denver, CO

Author contact: carol.vojir@uchsc.edu, with a copy to the Editor: roxie.foster@uchsc.edu