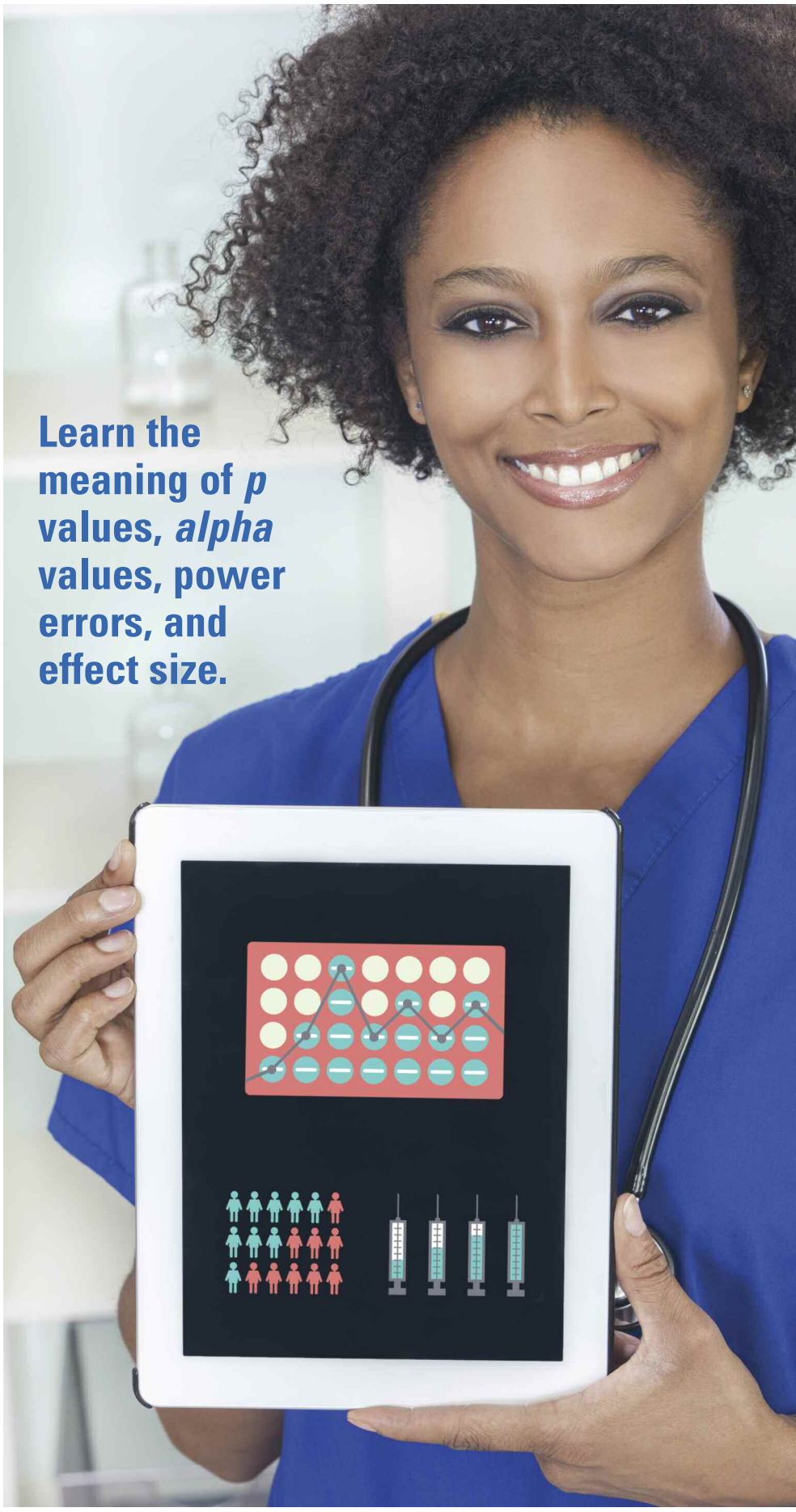
Nursing Research 101

Differentiating statistical significance and clinical significance

By Elizabeth Heavey, PhD, RN, CNM

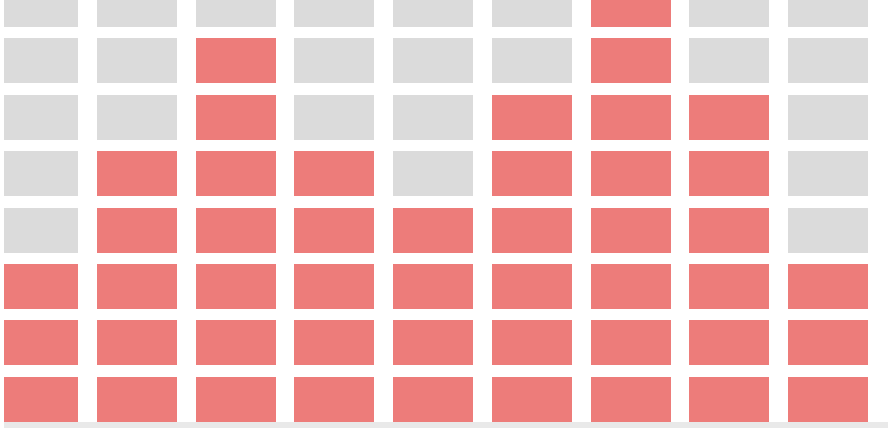


Learn the meaning of *p* values, *alpha* values, power errors, and effect size.

TO IMPLEMENT evidence-based practice, nurses must be able to comprehend and interpret research. That means you need to understand the distinction between statistical significance and clinical significance. Although the two concepts are related, they're not the same thing. To confuse matters further, many people use the terms incorrectly.

Let's look at a few examples from the nursing literature. Swartzell et al. (2013) studied the relationship between fall risk score and actual occurrence of falls in acute-care patients diagnosed with diabetes, stroke, and heart failure. For a group of patients with diabetes, they reported a *statistically significant difference* between the scores for patients who fell and for those who didn't fall.

A statistically significant difference means an association or difference exists between the variables that wasn't caused solely by normal variation or chance. The probability value (*p* value) tells you the probability or chance that the results are



just a random occurrence and not an actual difference between the variables. If the *p* value determined by the analysis is less than the certainty level (*alpha* value) required by the researcher, the results are statistically significant and not likely to be a random occurrence.

Traditionally, researchers set the *p* value (*p*) at 0.05, which means there's a 5% probability that the results weren't caused only by chance and the researcher is 95% certain a true relationship exists between the variables examined in the study.

Determining statistical significance

In the Swartzell study, the analysis resulted in a *p* value that was less than the alpha value the researchers required ($p = 0.02$, *alpha* = 0.05). This means they were confident the difference they found was real and hadn't occurred just by chance. In other words, the fall risk score was related to the actual risk of falling in diabetic patients. Patients with a high fall risk score were more likely to fall than those with a low fall risk score. Thus, the researchers established a statistically significant difference.

Determining clinical significance

To determine if a statistically significant difference is *clinically significant*, we have to go one step further. A difference is deemed clinically significant when experts in the field believe a statistically significant finding is substantial enough to be clinically important and thus should direct the course of patient care.

Swartzell et al. found a statistically significant difference—but only among diabetic patients, not among patients with heart failure or stroke ($p = 0.729$, *alpha* = 0.05). What's more, despite that difference, the instrument was unable to identify a fall risk for 44% of the patients who fell. Because of the limited ability to accurately detect a fall risk across multiple patient populations, the statistically significant result may be only minimally useful and thus isn't clinically significant. If a statistically significant result isn't clinically significant, it's not clinically helpful and shouldn't be used to guide clinical practice.

In another example, Szabo et al. (2014) reported a statistically significant increase in intracranial pressure (ICP) among intensive care unit patients during oral care ($p = 0.0031$, *alpha* = 0.05). However, these increases resolved spontaneously without intervention. Because oral care has substantial benefits, the researchers recommended such care continue despite the statistically significant difference in ICP. Thus, the statistically significant difference wasn't deemed clinically significant.

In many cases, a statistically significant difference is clinically significant.

But in many cases, a statistically significant difference *is* clinically significant. Stine et al. (2012) found that using the axillary versus rectal route to measure body temperature in children less than 1 year old leads to a statistically significant difference in recorded temperature ($p = 0.0001$, *alpha* = 0.05). Missing a fever in an infant is quite concerning. Because of the significant difference in actual temperature measurement and noted variability in the measurement depending on the route, the researchers recommended assessing body temperature by the rectal route as the standard of care in their facility. They directed a course of action to standardize the mechanism by which body tem-

Key points to remember

- A *statistically significant difference* means the researchers found an association or difference that wasn't caused solely by normal variation or chance. The probability value (*p* value) determined in the statistical test was less than the certainty level (*alpha* value) required by the researchers. Statistical significance is established by the researchers' analysis.
- Sample size can affect the ability to find a statistically significant difference. Too small a sample may lead researchers to miss a statistically significant difference (*type II* error). Too large a sample may lead them to report a difference that doesn't exist (*type I* error).
- A *clinically significant difference* means the researchers found a statistically significant difference that experts in the field believe is substantial enough to be clinically important and thus should direct the course of patient care.
- Statistical significance must always be established before clinical significance can be determined.
- Clinical significance is a subjective judgment that can't be determined by a single test.

perature is measured for children younger than age 1 because the different was significant both statistically and clinically.

Statistical significance comes first

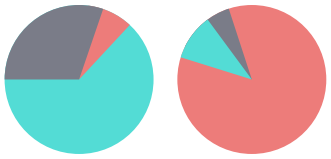
We sometimes hear a person say, "The results aren't statistically significant, but they are clinically significant." This statement is inaccurate. Statistical significance must be established before clinical significance can be determined. What that person probably meant was that he or she believes there's a statistically significant difference that wasn't detected for some reason—one of which may be an inadequate sample size. (See *Key points to remember.*)

Understanding effect size and power errors

When researchers are looking for a small difference, called the *effect size*, they need to recruit a large sample to find the small difference at a statistically significant level. If they're looking for a small difference but don't have a large enough sample to find it, a *type II* error (also called a *power error*), may occur. In this case, a statistically significant difference occurs, but researchers can't detect it because the sample size is too small.

In another example from the nursing literature, Binns-Turner et al. (2011) examined the effect of a preoperative music intervention for women with breast cancer who were about to undergo a mastectomy. They found statistically significant differences in mean arterial pressure (p = 0.003, alpha = 0.05), anxiety (p = 0.001), and pain (p = 0.007) between the women who listened to music before surgery and those who didn't. But they didn't detect a statistically significant change in heart rate (p = 0.248). They reported that seven other studies with larger samples found a statistically significant difference in heart rate with a music intervention. Because of this and the other findings they identified, they believe their study was underpowered, meaning its sample size was too small (n = 30) to detect the small difference they believe may exist in this particular variable.

If this assumption is correct, their finding relating to the impact of the music intervention on heart rate (HR) is a type II error. This means when they studied the effect of the music intervention on HR, they didn't find a statistically significant difference—yet a difference may have actually existed and they missed it. In other words, the study lacked enough power to detect an actual difference. If this study is repeated with a larger sample size and a statistically significant difference is found, experts in the field would determine if it's clinically significant. Researchers can't say the original statistically insignificant finding is clinically significant without eventually increasing the sample size and discovering if the statistically significant difference they suspected actually exists.



With a large enough sample, even minute differences occurring by chance can appear significant.

The reverse can be true, too. Researchers must be careful about using an overly large sample because it may incorrectly detect a statistically significant difference in variables arising only from chance. With a large enough sample, even minute differences occurring by chance can appear significant, even when the difference isn't meaningful. Called a *type I* error, this is another situation in which a statistically significant difference isn't clinically significant because the original conclusion was actually an error created by an overly large sample size.

To recap....

- *Statistical significance* is established by an analysis conducted by researchers.
- *Clinical significance* is established by experts in the field (including the same researchers), who decide if a statistically significant difference is clinically important.

As much as we'd like a simple answer, clinical significance is a subjective judgment and can't be determined by a single study. That's why we need nurses and other clinical experts to read and evaluate the scientific literature. ✭

Selected references

Binns-Turner PG, Wilson LL, Pryor ER, Boyd GL, Prickett CA. Perioperative music and its effects on anxiety, hemodynamics, and pain in women undergoing mastectomy. *AANA J.* 2011;79(4):S21-7.

Stine CA, Flook DM, Vincze DL. Rectal versus axillary temperatures: is there a significant difference in infants less than 1 year of age? *J Pediatr Nurs.* 2012;27(3):265-70.

Swartzell KL, Fulton JS, Friesth BM. Relationship between occurrence of falls and fall-risk scores in an acute care setting using the Hendrich II fall risk model. *Medsurg Nurs.* 2013;22(3):180-7.

Szabo CM, Grap MJ, Munro CL, Starkweather A, Merchant RE. The effect of oral care on intracranial pressure in critically ill adults. *J Neurosci Nurs.* 2014;46(6):321-9.

Elizabeth Heavey is an associate professor of nursing at the College of Brockport, State University of New York.