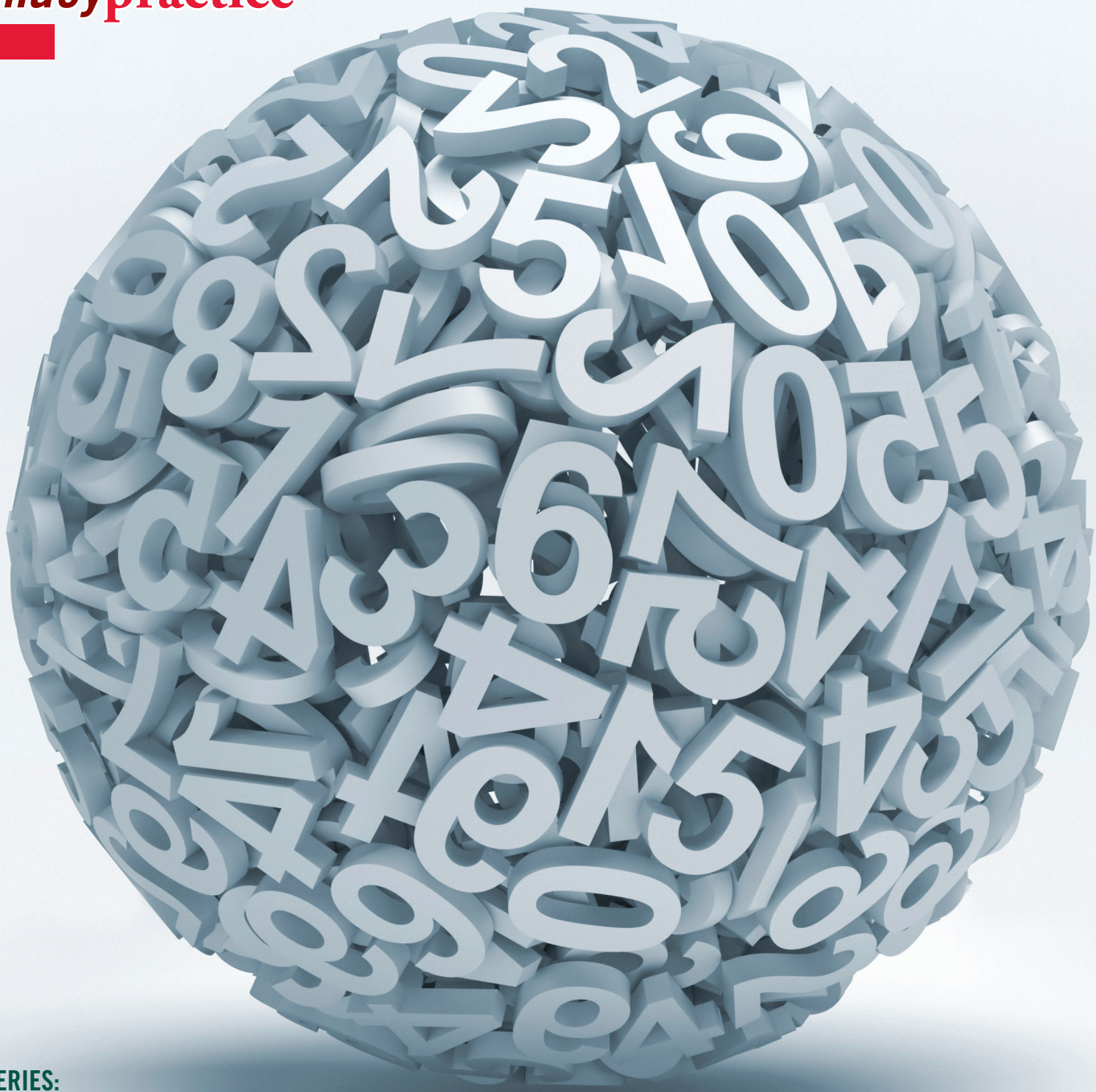

JOURNAL SERIES:

Statistics Review Part 4: Statistical Significance, Power, and Error

by Kevin Look, PharmD, PhD and Amanda Margolis, PharmD, MS, BCACP

This is the fourth article in a series of articles designed to review the basic, fundamental concepts of biostatistics. This article describes statistical significance, power, and error.

In order to fully grasp the importance of a study's findings and limitations, it is important to understand some key concepts related to statistical significance, power, and error. In this article, we will discuss how to interpret research findings by discussing p-values and statistical significance. We will then discuss common types of errors encountered in pharmacy research (Type I and Type II errors), and the importance of statistical power to research. Finally, we briefly discuss the difference between statistical and clinical significance.

Probability and Chance

Biomedical research usually starts with a question such as "How does Drug X affect A1c in patients with Type II diabetes?" Next, researchers identify the population of interest and select an appropriate sample from that population. A population (often

denoted with a capital "N") is the entire group that the data collected is intended to apply to (e.g., all patients with Type II diabetes). A sample (often denoted with a lowercase "n") is a representative subset of the population from which the data is collected (e.g., data from a small number of Type II diabetic patients). It is impractical or impossible to measure the entire population due to time, resource, and cost constraints; imagine trying to measure the A1c of every person with Type II diabetes in the United States! Therefore, research is typically performed on a representative sample taken from this population. However, what we are actually interested in knowing is the behavior of the entire population of interest (all diabetics).

Statistics help with generalizing the characteristics of subjects in the sample to a population of interest. This is called inferential statistics, which uses sample evidence to draw inferences or conclusions about a whole population. It is important to note that the results of a study can be generalized only to a population that is similar to the study sample; the results may not be generalizable to other populations. (For example, a study of the effectiveness of a new type of insulin performed in Type I diabetics using an insulin pump may not be generalized to Type I diabetics not using an insulin pump, or to Type II diabetics.) Inferential statistics are based on probabilities, or the likelihood of events. Because of the natural variability among individuals, it is important to determine whether an observed difference in the sample is a true difference or merely a chance occurrence. That is, did the patient's A1c decrease because of the medication, or was it just due to random chance?

Probability is an estimation of how likely it is that something will happen or that a statement is true. The probability of an event occurring ranges from 0 (never occurs) to 1 (always occurs). Suppose, for example, we wanted to predict the outcome of flipping a fair coin. Since a coin has only two sides, we know there are only two possible outcomes: heads or tails. Therefore, the probability of the coin landing heads up is 1/2 (or 0.5), and the probability of the coin landing tails up is 1/2 (or 0.5). Similarly, if we were to roll a 6-sided die, the probability of rolling a 2 is 1/6, the probability of rolling an even number is 3/6 (or 1/2), and so on.

TABLE 1. Error Summary

	H_0 True (Truly no treatment effect)	*H_A True (Truly a treatment effect)*
Reject H_0 (Treatment effect found)	**Type I Error** (Find a treatment effect when there is truly no effect) Probability = α (significance level)	**Correct** (Find a treatment effect when there is truly a treatment effect) Probability = 1 - β (Power)
Reject H_A (No treatment effect found)	**Correct** (Do not find a treatment effect when there is truly no effect) Probability = 1 - α	**Type II** (Do not find a treatment effect when there is truly a treatment effect) Probability = β (1 - power)

Now suppose we flipped a coin 50 times and got 50 heads in a row. Since we know the probability of getting a heads is 1/2, the probability of getting 50 consecutive heads is very low. According to this low probability value, most reasonable people would conclude that heads were favored; that is, that getting a heads was more likely than getting a tails. This illustrates a very important principle: **that if, under a given assumption** (e.g., the probability of a heads is 1/2), **the probability of a particular sample's results is exceptionally small** (e.g., 50 straight heads), **one would conclude that the assumption is incorrect.**

Hypothesis Testing and P-Values

A hypothesis is a claim or statement about a property of a population. For example, "The manufacturer of Drug X claims it reduces A1c by 1.0% in patients with diabetes." Research involves two types of hypotheses: a null hypothesis and an alternative hypothesis. In biomedical research, the null hypothesis, typically denoted H_0, is the claim that the intervention has no effect. The alternative hypothesis, typically denoted H_A, is the claim that the intervention actually does have an effect. In this case: H_0: Drug X has no effect on A1c. H_A: Drug X has an effect on A1c.

To investigate a hypothesis, we analyze a sample through a clinical study to determine if the results are reasonably consistent or conflicting with the hypothesis. This is commonly done using a p-value. The p-value is the probability of obtaining results as extreme or more extreme than the results observed given H_0 is true. For example, assuming our H_0 is true that there is no effect of the study medication, the p-value is the probability of obtaining the observed (or larger) difference in A1c in our intervention group compared to the control group. (This can also be done using confidence intervals, which was discussed in part 3 of this series. The statistical tests used to determine the p-value will be covered in part 5 of this series.) Because there is inherent variability or error in the way things are measured, we define a pre-specified "acceptable" error level, typically designated using α. The biomedical standard for an acceptable error level is 5% (or α=0.05), although in some instances a lower (1% or α=0.01) or higher (10% or α=0.1) level of error may be acceptable. An α level of 0.05 means there is a less than 1:20 (5%) chance that the results occurred due to random error.

The p-value obtained in a study can be compared to the pre-specified error level to determine whether or not the findings are consistent with the null hypothesis. Returning to our example, based upon our findings, we can do one of two things: if we find results that are inconsistent with the null hypothesis (p-value < α, typically a p-value less than 0.05), we reject the null hypothesis that Drug X has no effect and accept the alternative hypothesis that Drug X does have an effect on A1c. Alternatively, if our results are reasonably consistent with the hypothesis (p-value ≥ α, typically a p-value greater than 0.05), we would fail to reject the null hypothesis and conclude that Drug X has no effect on A1c. Note that we cannot accept the null hypothesis; we can only say that we do not have enough evidence to reject it.

Type I and Type II Errors and Power

Despite our best efforts, errors in statistical measurement can occur. Errors are typically classified as either a Type I error or a Type

II error. A Type I error occurs when H_0 is rejected when it is true (a "false positive"). This occurs when a researcher claims a finding was statistically significant when in fact it was not. This commonly occurs when small sample sizes are used that allow random variation to have a large effect. This can also occur when a large number of statistical tests are performed (called "repeated testing"); positive results can be found simply by undertaking so many comparisons that significant results will eventually be found by chance.[1] A Type II error occurs when we fail to reject H_0 when it is false (a "false negative"). This occurs when a researcher claims there was no significant difference when there actually was. This commonly occurs when the sample size is too small for a difference to be detected or when our α is set too small.

Conventionally, we allow a 20% risk of a Type II error (typically designated using β=0.2); however, the actual value of β varies with the size of α, the size of the effect, the size of the sample, and the variance of the original distribution. Statistical power is the ability of a study to detect a statistically significant difference, and is defined as 1-β. Since we typically set β to be 20%, the statistical power is typically set at 80% (1 - 0.2 = 0.8), although it can be set higher. Many studies do not have adequate sample sizes to be definitive in their conclusions and are underpowered; there is actually a significant difference but the sample size is not large enough to detect it.[1] Table 1. summarizes the different types of error.

Statistical versus Clinical Significance

A final important concept is the difference between statistical significance and clinical significance. A statistically significant finding may not always be clinically meaningful in practice. For example, a study may find that a new diabetes drug significantly reduces patient A1c by 0.1% compared to placebo ($p<0.05$). Although this finding may be statistically significant, it may not be a clinically important finding for many patients. However, the clinical significance of a finding varies on the finding and the situation. Although for many patients a 0.1% reduction in A1c may not be clinically useful, it may be useful for a patient who has contraindications or adverse reactions to other diabetes medications.

Summary

This article reviewed several key concepts related to statistical significance, power, and error. This information is useful to understand the importance of a study's findings and limitations. The next article in this series will cover the use of statistical tests to determine a p-value.

Kevin Look is an Assistant Professor in the Social & Administrative Sciences Division at the University of Wisconsin School of Pharmacy, Madison, WI. Amanda Margolis is a Lecturer at the UW-Madison School of Pharmacy and a Clinical Pharmacist at the William S. Middleton Memorial Veterans Hospital, Madison, WI.

Practice questions

1. Which of these statements is the correct interpretation of a p-value? An exceptionally small p-value indicates the difference between two groups is:
   a. likely due to chance
   b. unlikely to be due to chance
   c. is clinically significant
   d. is not clinically significant

2. A recent study found that Drug X reduces cardiovascular mortality by 5% compared to placebo (p=0.08). Assuming α=0.05 and the following hypotheses, what conclusion would we make?
   H_0: Drug X is not significantly different from placebo
   H_A: Drug X is significantly different placebo
   a. We reject H_0 and accept H_A
   b. We reject H_0 and fail to reject H_A
   c. We accept H_0
   d. We fail to reject H_0

3. A study of a new medication for diabetes finds a 0.8% lowered A1c in the treatment group compared to the placebo group with a p-value of 0.057. Which of the following statements regarding error is correct?
   a. Type I error could be possible
   b. Type II error could be possible
   c. Type I and II errors could be possible
   d. There is unlikely to be an error

Answers:

1. **b.** The p-value is the probability of obtaining the observed (or larger) difference in A1c in our intervention group compared to the control group. When the p-value is very small we can reject the null hypothesis that there is no difference between the two groups.

2. **d.** Since our p-value is greater than α, we would fail to reject H_0 (no effect). We cannot accept H_0, only fail to reject it.

3. **b.** Using the conventional cutoff of 0.05 for the p-value, we find the results of this study are not statistically significant. A Type II error is possible in that there may truly be a difference but the study was underpowered to detect a difference.

References and suggestions for further review:

1. Redmond AC, Keenan AM. Understanding Statistics. Putting p-values into perspective. J Am Podiatr Med Assoc. 2002;92(5):297-305.
2. Munro BH. Statistical Methods for Health Care Research. Fifth Edition. Philadelphia, PA: Lippencott Williams and Wilkins; 2004.