Statistically Speaking

Statistical Power

Regina L. Nuzzo, PhD

Terminology:

- **Sensitivity**: The “true positive” rate for a test; the proportion of “positives” that the test correctly flags as being positive.
- **Specificity**: The “true negative” rate for a test; the proportion of “negatives” that the test correctly excludes as being negative.
- **Type I error rate**: The “false-positive” rate for a statistical test; the proportion of “negatives” that will be incorrectly flagged as being positive. Also known as \( \alpha \), or the significance threshold.
- **Type II error rate**: The “false-negative” rate for a statistical test; the proportion of “positives” that will be incorrectly excluded as being negative.
- **Positive predictive power**: Of the cases flagged by the test as being positives, the proportion that are truly “positives.”
- **Negative predictive power**: Of the cases excluded by the test as being negatives, the proportion that truly are “negatives.”
- **Effect size**: A general term for the quantitative magnitude or strength of the effect being studied. Specific examples include the difference in outcome measure between 2 groups; the correlation between 2 measures; the change in an outcome measure over time.
- **Power**: The prestudy probability that a statistical test will correctly return a result of “statistically significant” for true effects that are no smaller than a certain size.

Statistical power is a prevalent but widely misunderstood concept used in planning studies and interpreting results. This article will discuss the concept of power, its relationship to sensitivity and specificity, the difficulties in achieving high power, implications of low power, and common misconceptions in interpreting power.

Concept of Power

Statistical power often is thought of solely in the context of determining sample size for a planned study, but power is in fact more relevant and useful beyond this context, and thus it’s useful to understand it from a statistician’s perspective. It’s important to remember that power is a theoretical property of a statistical test in a given study situation, in much the same way that sensitivity and specificity are properties of medical tests, as will be discussed herein. When we speak of “power calculations,” we usually are either using our knowledge of a specific study situation to estimate how strong its power will be, or we have a target power level in mind for a situation and are estimating how large the sample size we will need to achieve our goal.

There are 4 main components that affect the power of a study: sample size, significance threshold, population variance of the effect, and effect size. Knowing these 4 values will allow a researcher to calculate the theoretical power of a given study design. Of these 4 factors, sample size is the component most under a researcher’s control; the latter 3 factors typically are dictated by the situation. Thus, researchers usually use power calculations to work backwards: for a given significance threshold, population variance, and effect size, they calculate what sample size is necessary to design a study with 80% power. These 4 factors will be discussed in more detail herein.

Conceptually, the power of a statistical test is similar to the power of a telescope: it is the ability to separate a true pattern from its background. The smaller the object to be detected, or the more distracting the environment is, the more magnification power is needed to detect a true signal. A statistically high-powered, placebo-controlled randomized study, for instance, is more likely to spot even subtle differences between the treatment and control groups than a low-powered study would, even when there is high person-to-person variability (“noise”) in the results. Thus, designing a study to have high power reduces the chance of overlooking true findings (ie, reducing false negatives, or Type II errors), and it also increases the chances that a significant finding in fact represents a true effect (a true positive).
Statistical Power

A

1000 Hypotheses

Power = 80%

500 True Effects

400 True Positives

100 False Negatives

475 True Negatives

25 False Positives

575 Negatives: 82% Correct 18% Wrong Negative Predictive Value = 82%

425 Positives: 94% Correct 6% Wrong Positive Predictive Value = 94%

B

1000 Hypotheses

Power = 30%

500 True Effects

150 True Positives

350 False Negatives

475 True Negatives

25 False Positives

825 Negatives: 58% Correct 42% Wrong Negative Predictive Value = 58%

175 Positives: 86% Correct 14% Wrong Positive Predictive Value = 86%

C

1000 Hypotheses

Power = 80%

250 True Effects

200 True Positives

50 False Negatives

712.5 True Negatives

37.5 False Positives

763 Negatives: 93% Correct 7% Wrong Negative Predictive Value = 93%

238 Positives: 84% Correct 16% Wrong Positive Predictive Value = 84%

D

1000 Hypotheses

Power = 30%

250 True Effects

275 True Positives

175 False Negatives

712.5 True Negatives

37.5 False Positives

888 Negatives: 80% Correct 20% Wrong Negative Predictive Value = 80%

113 Positives: 67% Correct 33% Wrong Positive Predictive Value = 67%
More formally, the power of a statistical test for a given effect size \( d \) is the probability before the study that the test will return a result of "statistically significant" (eg, \( P < .05 \)) when the true effect size is \( d \) or larger. This is the probability of getting a "true positive" when the population effect is no smaller than \( d \). The formula for power can depend on the individual study design and require a number of crucial assumptions, so researchers often consult a statistician for help or use available software, such as G*Power [1,2].

It is important to note that it is incomplete to simply discuss the power of a study or a test without discussing a specific effect size. A calculated power is always dependent on the hypothesized effect size in the population that is being sought; different effect size values will yield different estimates of power. Thus, instead of writing simply about the "power of a study," we should write, for example, about the "power to detect a true mean difference of at least 5 pounds," or the "power against a mean change of 2 points."

### Power and Significance, as Related to Sensitivity and Specificity

Power is akin to the sensitivity of a medical test. A screening test with high sensitivity for a certain disease has a high probability of detecting a patient’s disease if it is present and will therefore correctly flag a large fraction of diseased patients. Likewise, a statistical test with 80% power for effect size \( d \) is sensitive enough to pick up true effects as small as \( d \) about 80% of the time over many repetitions. In other words, a highly sensitive medical test will overlook very few diseased patients; a high-powered statistical test will miss very few true effects of a given size.

Specificity also has an analog in statistical testing. A highly specific medical test has a better chance of correctly giving a healthy person a clean bill of health and therefore not incorrectly flagging too many healthy patients with a false positive. In statistical testing, specificity is controlled by the significance threshold, or \( \alpha \), which is established before a study is done. A statistical test with a significance threshold of .01 is specific enough to ignore truly nonsignificant results 99% of the time over many repetitions. In other words, a highly specific test will produce very few false positives.

Just as is the aim in medical testing, the crucial hypothesis test in a study will ideally have both high sensitivity and specificity. Yet, in general, increasing a test’s specificity will decrease its sensitivity. Thus, tradeoffs must be balanced, and costs of errors in specific situations must be weighed. In a research setting, high sensitivity helps prevent truly effective interventions or results from being overlooked, whereas high specificity helps prevent ineffectual interventions from being promoted and spurious associations from being published.

### Difficulties in Increasing Power

Although great attention has been paid to the problems of false positives [3], there is an increasing awareness that the problems stemming from low power also deserve serious consideration, especially true in light of increasing evidence of low power in many studies [4]. In challenging research environments, however, adequate power in a study can be difficult to achieve. It is standard in much of biomedical research to design a study so that the primary hypothesis test will have 80% power for the smallest plausible effect that will have clinical relevance; in other words, it is desired that a true effect of this given size would have an 80% chance of being flagged as statistically significant. When designing a study, researchers should keep in mind the 4 study components that will affect its power:

1. **Sample size**: Increasing a study’s sample size will increase how precisely we can estimate the true effect size, which will in turn increase the study’s power.
2. **Significance threshold:** All other things being equal, power and significance thresholds work in opposition to one another. Keeping the sample size identical but setting a less stringent threshold for significance (e.g., increasing alpha from .05 to .10) will increase a study’s power and thus decrease its false negatives, but it will do so at the cost of increasing the number of false positives. On the other hand, a stricter threshold for significance (decreasing alpha from .05 to .01, for example) will have the advantage of producing fewer false positives, but it will also decrease power and thus increase false negatives. Designing a study with 80% power and a significance threshold of .05 essentially places 4 times the priority on minimizing false positives compared with minimizing false negatives. In such a study, a null result will have a 5% chance of incorrectly coming up significant, whereas a true result will have a 20% chance of incorrectly coming up nonsignificant.

3. **Population variance:** For research situations fortunate enough to have in the study population a low variability in the effect being measured, a study with a particular sample size will have greater power to detect the effect than a study with an identical size but greater population variance. In other words, when researchers expect a high population variance in the results, their study will need a greater sample size to achieve the standard 80% power than would a study in which low variance is expected.

4. **Effect size:** All other factors being equal, the power to detect a large effect is greater than the power to detect a small effect; that is, it is more difficult to detect subtle effects. Thus, if researchers want the ability to detect small as well as large effects at 80% power, they need to collect a greater sample size than if they were only looking for large effects.

### Problems Stemming From Low Power

In ideal situations, researchers could conduct a study with a sample size large enough to achieve very high power. In reality, however, it can often be difficult or impossible to collect a large number of observations, as the result of cost, limited time, patient scarcity or dropouts, or other constraints, so it’s often not practically possible to achieve the desired level of power, a fact that is leading some researchers to explore alternatives methods to determining an appropriate sample size for a study [5]. Although changing the significance level is an easy way to increase power, it is difficult to do in practice; an alpha of .05 is standard in much of research, and in fact some researchers are calling for even more stringent thresholds such as .005 or .001 [6].

In light of the difficulties in achieving high power, it’s important to look at how low power affects study outcomes. The first important impact is the most widely known: More true effects are overlooked with a low-powered study than with a high-powered one. For example, for a set of hypotheses all tested at 80% power for a given effect size, only 20% of true effects of that size will be incorrectly classified as nonsignificant. If the hypotheses were tested at 25% power, however, three-quarters of true effects will be overlooked.

A second effect is less widely appreciated: In low-powered studies that test many comparisons and conditions, a greater fraction of statistically significant results may in fact be false positives. This problem becomes even worse in research situations in which more hypotheses are “exploratory” and fewer correspond to true effects. This situation is analogous to using a medical test with low sensitivity to screen the general population versus screening a targeted high-risk group: the general population includes more asymptomatic individuals unlikely to have the disease, so a higher proportion of patients flagged by the test as having the disease will in fact be healthy. In this situation, the positive predictive value of the test will be low, meaning that a greater fraction of positives will be false positives [7]. The relationship among exploratory hypotheses, power, positive predictive value, and negative predictive value is illustrated in Figure 1.

A third effect of low power is perhaps even less well known but is gaining wider attention: Effect sizes of statistically significant results in low-powered studies can be exaggerated, so that significant effects appear larger than they really are [4,8]. This problem has obvious implications for follow-up studies and reproducibility of research; it is sometimes referred to as a “Type M (magnitude) error” or “the winner’s curse,” the latter referencing the common phenomenon of promising early findings not panning out in later studies. When initial studies are underpowered to detect the true effect of an intervention, it is often only those studies “lucky” enough to have drawn a sample with an unusually large effect that will produce statistically significant results and be published. Thus, the effects in these studies that pass the “statistical significance filter” are often inflated in publication. This is a problem for follow-up studies because they are usually designed to replicate these too-large effects, and since more power is needed to detect smaller effects, these follow-up studies are often underpowered and thus often will fail to find a significant effect.

### Misconceptions About Prestudy and Poststudy Power

Power should be used to guide the design of a study. When performing power calculations before data collection to aid in study design, researchers should choose the smallest effect size that is both plausible and clinically relevant. Although it may be difficult to estimate the true effect size, researchers should
consider the size below which an effect would be considered negligible. This is often a difficult task, but as statistician Stephen Senn points out, “An astronomer does not know the magnitude of new stars until he has found them, but the magnitude of star he is looking for determines how much he has to spend on a telescope” [9].

Power provides very little information after a study is completed. Power is the probability before the study that the statistical test will correctly pick up on a hypothesized effect size if it is present. After a study, therefore, prestudy power calculations cannot reveal how well the hypothesized effects fit with data actually observed in the study.

After a study, some researchers instead use the observed data to calculate power, in what is known as “observed power” or “post hoc power.” This post hoc power, however, is simply a restatement of the P value and therefore contains no new information [10]. For example, after finding a nonsignificant effect, sometimes researchers will point to a low post hoc power to make the argument that the study was simply underpowered to detect the observed effect, and the results likely to be a false negative. Yet for any nonsignificant result, it can be shown that the power (with very few exceptions) will necessarily be <.50%. So using post hoc power does not provide more help in interpreting nonsignificant results.

Power Calculation:

**Example Calculations for Power and Sample Size**

Software such as G*Power will easily perform the following calculations; they are provided here for readers seeking a slightly deeper conceptual understanding. Note that the Z test statistic rather than the more accurate t test statistic is used in the interest of providing a simple illustration.

Suppose researchers want to use the Functional Independence Measure to investigate the difference between treatment group and a control group of patients. They would like to have enough power in their study to detect a difference $|\mu_1 - \mu_2|$ of at least 22 points, which is the Minimum Clinically Important Difference. From previous studies they believe the observations are roughly normally distribution and the population standard deviation $\sigma$ for the Functional Independence Measure score is about 25 points for each group. They plan to do a standard 2-sided hypothesis test of no difference between groups, with a significance threshold $\alpha$ of .05 and equal sample sizes for each group.

What sample size do the researchers need to have a study with 80% power to detect a difference of 22 points?

Solution:

$$n = \text{minimum sample size for each group} = \frac{2\sigma^2 (z_{\alpha/2} + z_\beta)^2}{(\mu_1 - \mu_2)^2}$$

where

$\sigma =$ population standard deviation (assumed to be equal for both groups)

$z_{\alpha/2} =$ standard normal value for which 100 ($\alpha/2$)% of the values fall in the upper tail (for the typical significance threshold of $\alpha = .05$, $z_{\alpha/2} = 1.96$)

$z_\beta =$ standard normal value at which 100 $\beta$% of the values fall in the upper tail (for the typical power of 80%, $\beta = 1 - .020$ and $z_\beta = 0.841$)

$\mu_1 - \mu_2 =$ the difference that the study will be powered to detect

So in this example, $\sigma = 25$, $\alpha = .05$, $z_{\alpha/2} = 1.96$, $\beta = .20$, $z_\beta = 0.841$, and $|\mu_1 - \mu_2| = 22$

Therefore, $n = \frac{2\cdot25^2(1.96+0.841)^2}{22^2} = 20.26$. So a minimum of 21 patients in each group would be necessary for the study to have 80% power to detect a difference of 22 points, given the assumptions.

Note: Closer inspection of the formula for the sample size shows the relationship between power and the 4 elements: sample size, population standard deviation, significance threshold, and expected effect size. A larger population standard deviation $\sigma$ will increase the sample size necessary to achieve the desired power; a larger $\alpha$ (less strict threshold) will decrease $z_{\alpha/2}$ and therefore decrease the necessary sample size; accepting a lower power will increase $\beta$, decrease $z_\beta$, and therefore decrease the necessary sample size; searching for a larger mean difference $|\mu_1 - \mu_2|$ will decrease the necessary sample size.
References


Disclosure

R.L.N. Department of Science, Technology, and Mathematics, Gallaudet University, 800 Florida Avenue, NE, Washington, DC 20002-3695. Address correspondence to: R.L.N.; e-mail: regina.nuzzo@gallaudet.edu
Disclosure: nothing to disclose
Accepted August 11, 2016.