WIKIPEDIA

Power of a test

The statistical **power** of a binary hypothesis test is the probability that the test correctly rejects the null hypothesis (H_0) when a specific alternative hypothesis (H_1) is true. It is commonly denoted by $1-\beta$, and represents the chances of a "true positive" detection conditional on the actual existence of an effect to detect. Statistical power ranges from 0 to 1, and as the power of a test increases, the probability β of making a type II error by wrongly failing to reject the null hypothesis decreases.

Contents

Notation

This article uses the following notation

- β = probability of a Type II error, known as a "false negative"
- 1-β = probability of a "true positive", which is correctly rejecting the null hypothesis. **"1-β" is also known as the power of the test.**
- α = probability of a Type I error, known as a "false positive"
- 1-α = probability of a "true negative", which is correctly accepting the null hypothesis

Description

For a type II error probability of β, the corresponding statistical power is 1 – β. For example, if experiment E has a statistical power of 0.7, and experiment F has a statistical power of 0.95, then there is a stronger probability that experiment E had a type II error than experiment F. This reduces experiment E's sensitivity to detect significant effects. However, experiment E is consequently more reliable than experiment F due to its lower probability of a type I error. It can be equivalently thought of as the probability of accepting the alternative hypothesis (H_1) when it is true — that is, the ability of a test to detect a specific effect, if that specific effect actually exists. Thus,

$$\text{power} = \Pr\left(\text{reject } H_0 \mid H_1 \text{ is true}\right).$$

If H_1 is not an equality but rather simply the negation of H_0 (so for example with $H_0 : \mu = 0$ for some unobserved population parameter μ, we have simply $H_1 : \mu \neq 0$) then power cannot be calculated unless probabilities are known for all possible values of the parameter that violate the null hypothesis. Thus one generally refers to a test's power *against a specific alternative hypothesis*.

As the power increases, there is a decreasing probability of a type II error, also called the *false negative rate* (β) since the power is equal to 1 – β. A similar concept is the type I error probability, also referred to as the *false positive rate* or the level of a test under the null hypothesis.

In the context of binary classification, the power of a test is called its *statistical sensitivity*, its *true positive rate*, or its *probability of detection*.

Power analysis

A related concept is "power analysis". Power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. For example: "How many times do I need to toss a coin to conclude it is rigged by a certain amount?"[1] Power analysis can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size. In addition, the concept of power is used to make comparisons between different statistical testing procedures: for example, between a parametric test and a nonparametric test of the same hypothesis.

Background

Statistical tests use data from samples to assess, or make inferences about, a statistical population. In the concrete setting of a two-sample comparison, the goal is to assess whether the mean values of some attribute obtained for individuals in two sub-populations differ. For example, to test the null hypothesis that the mean scores of men and women on a test do not differ, samples of men and women are drawn, the test is administered to them, and the mean score of one group is compared to that of the other group using a statistical test such as the two-sample *z*-test. The power of the test is the probability that the test will find a statistically significant difference between men and women, as a function of the size of the true difference between those two populations.

Factors influencing power

Statistical power may depend on a number of factors. Some factors may be particular to a specific testing situation, but at a minimum, power nearly always depends on the following three factors:

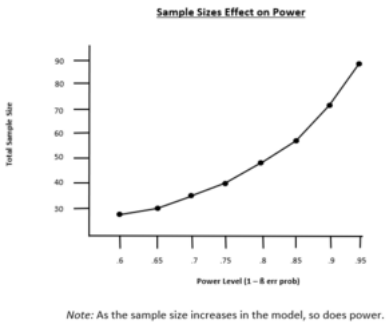
- the statistical significance criterion used in the test

- the magnitude of the effect of interest in the population
- the sample size used to detect the effect

A **significance criterion** is a statement of how unlikely a positive result must be, if the null hypothesis of no effect is true, for the null hypothesis to be rejected. The most commonly used criteria are probabilities of 0.05 (5%, 1 in 20), 0.01 (1%, 1 in 100), and 0.001 (0.1%, 1 in 1000). If the criterion is 0.05, the probability of the data implying an effect at least as large as the observed effect when the null hypothesis is true must be less than 0.05, for the null hypothesis of no effect to be rejected. One easy way to increase the power of a test is to carry out a less conservative test by using a larger significance criterion, for example 0.10 instead of 0.05 . This increases the chance of rejecting the null hypothesis (obtaining a statistically significant result) when the null hypothesis is false; that is, it reduces the risk of a type II error (false negative regarding whether an effect exists). But it also increases the risk of obtaining a statistically significant result (rejecting the null hypothesis) when the null hypothesis is not false; that is, it increases the risk of a type I error (false positive).

The **magnitude of the effect** of interest in the population can be quantified in terms of an effect size, where there is greater power to detect larger effects. An effect size can be a direct value of the quantity of interest, or it can be a standardized measure that also accounts for the variability in the population. For example, in an analysis comparing outcomes in a treated and control population, the difference of outcome means $\bar{Y} - \bar{X}$ would be a direct estimate of the effect size, whereas $(\bar{Y} - \bar{X})/\sigma$ would be an estimated standardized effect size, where σ is the common standard deviation of the outcomes in the treated and control groups. If constructed appropriately, a standardized effect size, along with the sample size, will completely determine the power. An unstandardized (direct) effect size is rarely sufficient to determine the power, as it does not contain information about the variability in the measurements.

The **sample size** determines the amount of sampling error inherent in a test result. Other things being equal, effects are harder to detect in smaller samples. Increasing sample size is often the easiest way to boost the statistical power of a test. How increased sample size translates to higher power is a measure of the efficiency of the test — for example, the sample size required for a given power.[2]



An example of how sample size affects power levels

The precision with which the data are measured also influences statistical power. Consequently, power can often be improved by reducing the measurement error in the data. A related concept is to improve the "reliability" of the measure being assessed (as in psychometric reliability).

The design of an experiment or observational study often influences the power. For example, in a two-sample testing situation with a given total sample size n, it is optimal to have equal numbers of observations from the two populations being compared (as long as the variances in the two populations are the same). In regression analysis and analysis of variance, there are extensive theories and practical strategies for improving the power based on optimally setting the values of the independent variables in the model.

Interpretation

Although there are no formal standards for power (sometimes referred to as π), most researchers assess the power of their tests using π = 0.80 as a standard for adequacy. This convention implies a four-to-one trade off between β-risk and α-risk. (β is the probability of a type II error, and α is the probability of a type I error; 0.2 and 0.05 are conventional values for β and α). However, there

will be times when this 4-to-1 weighting is inappropriate. In medicine, for example, tests are often designed in such a way that no false negatives (type II errors) will be produced. But this inevitably raises the risk of obtaining a false positive (a type I error). The rationale is that it is better to tell a healthy patient "we may have found something—let's test further," than to tell a diseased patient "all is well."[3]

Power analysis is appropriate when the concern is with the correct rejection of a false null hypothesis. In many contexts, the issue is less about determining if there is or is not a difference but rather with getting a more refined estimate of the population effect size. For example, if we were expecting a population correlation between intelligence and job performance of around 0.50, a sample size of 20 will give us approximately 80% power (α = 0.05, two-tail) to reject the null hypothesis of zero correlation. However, in doing this study we are probably more interested in knowing whether the correlation is 0.30 or 0.60 or 0.50 . In this context we would need a much larger sample size in order to reduce the confidence interval of our estimate to a range that is acceptable for our purposes. Techniques similar to those employed in a traditional power analysis can be used to determine the sample size required for the width of a confidence interval to be less than a given value.

Many statistical analyses involve the estimation of several unknown quantities. In simple cases, all but one of these quantities are nuisance parameters. In this setting, the only relevant power pertains to the single quantity that will undergo formal statistical inference. In some settings, particularly if the goals are more "exploratory", there may be a number of quantities of interest in the analysis. For example, in a multiple regression analysis we may include several covariates of potential interest. In situations such as this where several hypotheses are under consideration, it is common that the powers associated with the different hypotheses differ. For instance, in multiple regression analysis, the power for detecting an effect of a given size is related to the variance of the covariate. Since different covariates will have different variances, their powers will differ as well.

Any statistical analysis involving multiple hypotheses is subject to inflation of the type I error rate if appropriate measures are not taken. Such measures typically involve applying a higher threshold of stringency to reject a hypothesis in order to compensate for the multiple comparisons being made (*e.g.* as in the Bonferroni method). In this situation, the power analysis should reflect the multiple testing approach to be used. Thus, for example, a given study may be well powered to detect a certain effect size when only one test is to be made, but the same effect size may have much lower power if several tests are to be performed.

It is also important to consider the statistical power of a hypothesis test when interpreting its results. A test's power is the probability of correctly rejecting the null hypothesis when it is false; a test's power is influenced by the choice of significance level for the test, the size of the effect being measured, and the amount of data available. A hypothesis test may fail to reject the null, for example, if a true difference exists between two populations being compared by a t-test but the effect is small and the sample size is too small to distinguish the effect from random chance.[4] Many clinical trials, for instance, have low statistical power to detect differences in adverse effects of treatments, since such effects may be rare and the number of affected patients small.[5]

A priori vs. *post hoc* analysis

Power analysis can either be done before (*a priori* or prospective power analysis) or after (*post hoc* or retrospective power analysis) data are collected. *A priori* power analysis is conducted prior to the research study, and is typically used in estimating sufficient sample sizes to achieve adequate power. *Post-hoc* analysis of "observed power" is conducted after a study has been completed, and uses the obtained sample size and effect size to determine what the power was in the study, assuming the effect size in the sample is equal to the effect size in the population. Whereas the utility of prospective power analysis in experimental design is universally accepted, post hoc power

analysis is fundamentally flawed.[6][7] Falling for the temptation to use the statistical analysis of the collected data to estimate the power will result in uninformative and misleading values. In particular, it has been shown that *post-hoc* "observed power" is a one-to-one function of the *p*-value attained.[6] This has been extended to show that all *post-hoc* power analyses suffer from what is called the "power approach paradox" (PAP), in which a study with a null result is thought to show *more* evidence that the null hypothesis is actually true when the *p*-value is smaller, since the apparent power to detect an actual effect would be higher.[6] In fact, a smaller *p*-value is properly understood to make the null hypothesis *relatively* less likely to be true.

Application

Funding agencies, ethics boards and research review panels frequently request that a researcher perform a power analysis, for example to determine the minimum number of animal test subjects needed for an experiment to be informative. In frequentist statistics, an underpowered study is unlikely to allow one to choose between hypotheses at the desired significance level. In Bayesian statistics, hypothesis testing of the type used in classical power analysis is not done. In the Bayesian framework, one updates his or her prior beliefs using the data obtained in a given study. In principle, a study that would be deemed underpowered from the perspective of hypothesis testing could still be used in such an updating process. However, power remains a useful measure of how much a given experiment size can be expected to refine one's beliefs. A study with low power is unlikely to lead to a large change in beliefs.

Example

The following is an example that shows how to compute power for a randomized experiment: Suppose the goal of an experiment is to study the effect of a treatment on some quantity, and compare research subjects by measuring the quantity before and after the treatment, analyzing the data using a paired t-test. Let A_i and B_i denote the pre-treatment and post-treatment measures on subject i, respectively. The possible effect of the treatment should be visible in the differences $D_i = B_i - A_i$, which are assumed to be independently distributed, all with the same expected mean value and variance.

The effect of the treatment can be analyzed using a one-sided t-test. The null hypothesis of no effect will be that the mean difference will be zero, i.e. $H_0 : \mu_D = 0$. In this case, the alternative hypothesis states a positive effect, corresponding to $H_1 : \mu_D > 0$. The test statistic is:

$$T_n = \frac{\bar{D}_n - 0}{\hat{\sigma}_D/\sqrt{n}},$$

where

$$\bar{D}_n = \frac{1}{n}\sum_{i=1}^{n} D_i,$$

n is the sample size and $\hat{\sigma}_D/\sqrt{n}$ is the standard error. The test statistic under the null hypothesis follows a Student t-distribution with the additional assumption that the data is identically distributed $N(\mu_D, \sigma_D^2)$. Furthermore, assume that the null hypothesis will be rejected at the significance level of $\alpha = 0.05$. Since n is large, one can approximate the t-distribution by a normal distribution and calculate the critical value using the quantile function Φ^{-1}, the inverse of the cumulative distribution function of the normal distribution. It turns out that the null hypothesis will be rejected if

$$T_n > 1.64\,.$$

Now suppose that the alternative hypothesis is true and $\mu_D = \theta$. Then, the power is

$$\begin{aligned} B(\theta) &= \Pr\left(T_n > 1.64 \,\middle|\, \mu_D = \theta\right) \\ &= \Pr\left(\frac{\bar{D}_n - 0}{\hat{\sigma}_D/\sqrt{n}} > 1.64 \,\middle|\, \mu_D = \theta\right) \\ &= \Pr\left(\frac{\bar{D}_n - \theta + \theta}{\hat{\sigma}_D/\sqrt{n}} > 1.64 \,\middle|\, \mu_D = \theta\right) \\ &= \Pr\left(\frac{\bar{D}_n - \theta}{\hat{\sigma}_D/\sqrt{n}} > 1.64 - \frac{\theta}{\hat{\sigma}_D/\sqrt{n}} \,\middle|\, \mu_D = \theta\right) \\ &= 1 - \Pr\left(\frac{\bar{D}_n - \theta}{\hat{\sigma}_D/\sqrt{n}} < 1.64 - \frac{\theta}{\hat{\sigma}_D/\sqrt{n}} \,\middle|\, \mu_D = \theta\right) \end{aligned}$$

For large n, $\dfrac{\bar{D}_n - \theta}{\hat{\sigma}_D/\sqrt{n}}$ approximately follows a standard normal distribution when the alternative hypothesis is true, the approximate power can be calculated as

$$B(\theta) \approx 1 - \Phi\left(1.64 - \frac{\theta}{\hat{\sigma}_D/\sqrt{n}}\right).$$

According to this formula, the power increases with the values of the parameter θ. For a specific value of θ a higher power may be obtained by increasing the sample size n.

It is not possible to guarantee a sufficient large power for all values of θ, as θ may be very close to 0. The minimum (infimum) value of the power is equal to the confidence level of the test, α, in this example 0.05 . However, it is of no importance to distinguish between $\theta = 0$ and small positive values. If it is desirable to have enough power, say at least 0.90, to detect values of $\theta > 1$, the required sample size can be calculated approximately:

$$B(1) \approx 1 - \Phi\left(1.64 - \frac{\sqrt{n}}{\hat{\sigma}_D}\right) > 0.90,$$

from which it follows that

$$\Phi\left(1.64 - \frac{\sqrt{n}}{\hat{\sigma}_D}\right) < 0.10\,.$$

Hence, using the quantile function

$$\frac{\sqrt{n}}{\hat{\sigma}_D} > 1.64 - z_{0.10} = 1.64 + 1.28 \approx 2.92 \qquad \text{or} \qquad n > 8.56\hat{\sigma}_D^2,$$

where $z_{0.10}$ is a standard normal quantile; refer to the Probit article for an explanation of the relationship between Φ and z-values.

Extension

Bayesian power

In the frequentist setting, parameters are assumed to have a specific value which is unlikely to be true. This issue can be addressed by assuming the parameter has a distribution. The resulting power is sometimes referred to as Bayesian power which is commonly used in clinical trial design.

Predictive probability of success

Both frequentist power and Bayesian power use statistical significance as the success criterion. However, statistical significance is often not enough to define success. To address this issue, the power concept can be extended to the concept of predictive probability of success (PPOS). The success criterion for PPOS is not restricted to statistical significance and is commonly used in clinical trial designs.

Software for power and sample size calculations

Numerous free and/or open source programs are available for performing power and sample size calculations. These include

- G*Power (https://www.gpower.hhu.de/)
- WebPower Free online statistical power analysis (https://webpower.psychstat.org)
- Free and open source online calculators (https://powerandsamplesize.com)
- PowerUp! provides convenient excel-based functions to determine minimum detectable effect size and minimum required sample size for various experimental and quasi-experimental designs.
- PowerUpR is R package version of PowerUp! and additionally includes functions to determine sample size for various multilevel randomized experiments with or without budgetary constraints.
- R package pwr
- R package WebPower
- Python package statsmodels (https://www.statsmodels.org/)

See also

- Cohen's h
- Effect size
- Efficiency
- Neyman–Pearson lemma
- Sample size
- Uniformly most powerful test

References

1. "Statistical power and underpowered statistics — Statistics Done Wrong" (https://www.statisticsdonewrong.com/power.html). *www.statisticsdonewrong.com*. Retrieved 30 September 2019.
2. Everitt, Brian S. (2002). *The Cambridge Dictionary of Statistics*. Cambridge University Press. p. 321. ISBN 0-521-81099-X.
3. Ellis, Paul D. (2010). *The Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results*. United Kingdom: Cambridge University Press.

4. Ellis, Paul (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge University Press. p. 52. ISBN 978-0521142465.
5. Tsang, R.; Colley, L.; Lynd, L.D. (2009). "Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials". *Journal of Clinical Epidemiology*. **62** (6): 609–616. doi:10.1016/j.jclinepi.2008.08.005 (https://doi.org/10.1016%2Fj.jclinepi.2008.08.005). PMID 19013761 (https://pubmed.ncbi.nlm.nih.gov/19013761).
6. Hoenig; Heisey (2001). "The Abuse of Power". *The American Statistician*. **55** (1): 19–24. doi:10.1198/000313001300339897 (https://doi.org/10.1198%2F000313001300339897).
7. Thomas, L. (1997). "Retrospective power analysis" (http://eprints.st-andrews.ac.uk/archive/00000417/01/ThomasCB1997.pdf) (PDF). *Conservation Biology*. **11** (1): 276–280.

Sources

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). ISBN 0-8058-0283-5.
- Aberson, C.L. (2010). *Applied Power Analysis for the Behavioral Science*. ISBN 1-84872-835-2.

External links

- StatQuest: P-value pitfalls and power calculations (https://www.youtube.com/watch?v=UFhJefdVCjE) on YouTube