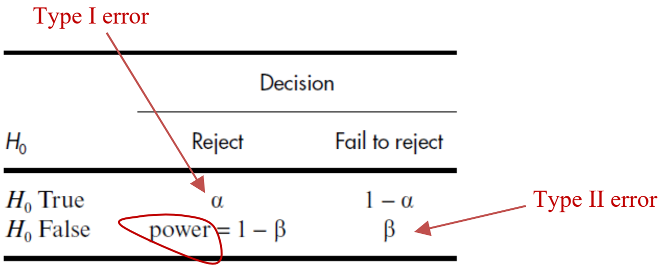
Power

Power Concept Review

Power is the probability of making a correct statistical decision (rejecting null, H_0) when the null hypothesis is false in the population. In other words, based on a test of our sample data, we correctly conclude the alternative hypothesis when it is actually true in the population. The reason for tortured language of "failing to reject" is that we never want to conclude that the null hypothesis is true for certain. We want to retain it if we do not find evidence against it. The idea is rooted in ideas of Hume (1798/1963) and Popper (1980) that we cannot prove a theory to be true. We can only find evidence against it. Over time, we develop theories that organize and explain a wealth of evidence consistent with specific hypotheses and have not been falsified.

Consider two options for a true state of affairs in the population, say that $H_0: \mu_1 = \mu_2$ or $H_1: \mu_1 \neq \mu_2$, the hypothesis addressed with a t test. Our statistical decision can be incorrect in two ways— by rejecting H_0 when we should retain it, a Type I error, and by failing to reject H_0 when we should reject it, a Type II error. A correct decision, rejecting H_0 when it is false in the population (H_1 is true) is statistical power.



Type I error

	Decision	
H_0	Reject	Fail to reject
H_0 True	α	$1 - \alpha$
H_0 False	power = $1 - \beta$	β

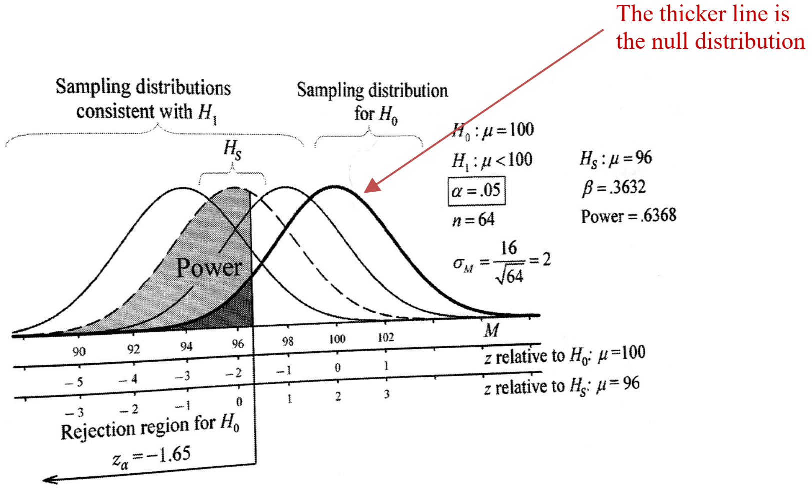
Type II error

Myers, J.L., & Well, A.D., & Lorch, R.F.,Jr. (2010). Research design and statistical analysis (3rd Edition). Mahwah, NJ: Erlbaum. (p. 78)

The null hypothesis is rejected (i.e., we decide the test is statistically significant) when we find that the sample value is far from the null hypothesis value (here, $H_0: \mu_1 - \mu_2 = 0$), in which we have estimated that there is a low probability that the sample value belongs to the sampling distribution that would be constructed if the null hypothesis was true in the population. The null rejection implies that we believe the statistical value belongs to the alternative sampling distribution, one that we could imagine constructing with random samples drawn from a population in which the alternative hypothesis is true, $H_1: \mu_1 - \mu_2 \neq 0$. If the null hypothesis is true, we have a good idea about the shape of the sampling distribution because of the central limit theorem. We know less about the location (central point or expected value, $E(\bar{Y})$) and shape of the sampling distribution under the alternative hypothesis. The true size of difference between the population means is always unknown, although we can estimate it with an effect size measure based on our sample. That sample value, however, is subject to sampling variability.

Noncentrality Parameter

The noncentrality parameter describes the degree of difference between the H_1 and H_0 values. The usual sampling distribution we discuss that is related to hypothesis testing is the distribution centered around the null hypothesis value, which, in this context, is referred to as the *central t distribution*. Thus, the *noncentral distribution* is the sampling distribution that would be created when the alternative hypothesis is true. But since the alternative hypothesis can be true in many different ways, depending on the size of (or even direction of) the group difference, there are many possible noncentral distributions. The degree of the difference of the means in this case is described by the noncentrality parameter.


Sampling distributions consistent with H_1
Sampling distribution for H_0
H_S
Power
$H_0: \mu = 100$
$H_1: \mu < 100$
$\boxed{\alpha = .05}$
$n = 64$
$\sigma_M = \frac{16}{\sqrt{64}} = 2$
$H_S: \mu = 96$
$\beta = .3632$
Power = .6368
90 92 94 96 98 100 102
M
−5 −4 −3 −2 −1 0 1
z relative to H_0: $\mu = 100$
−3 −2 −1 0 1 2 3
z relative to H_S: $\mu = 96$
Rejection region for H_0
$z_\alpha = -1.65$

The thicker line is the null distribution

Glenberg, A., & Andrzejewski, M. (2012). *Learning from data: An introduction to statistical reasoning*. Routledge. (p. 183.)

For a t test, the noncentrality parameter is represented by δ, (lowercase Greek delta).[1]

$$\delta = \frac{(\mu_1 - \mu_2) - (\mu_{0_1} - \mu_{0_2})}{\sigma\sqrt{(1/n_1) + (1/n_2)}} = \frac{\mu_1 - \mu_2}{\sigma\sqrt{(1/n_1) + (1/n_2)}}$$

As you might suspect by now, this noncentrality parameter sounds a lot like our effect size measure for the t test. Cohen's $\hat{d}$ is the sample estimate of effect size, but we can easily obtain an estimate of the noncentrality parameter from $\hat{d}$.

$$\delta = \hat{d}\sqrt{(1/n_1) + (1/n_2)}$$

The equation is further simplified to $\hat{d}/\sqrt{n}$ for equal group sizes or repeated measures.[2] The noncentral distribution depends on the degrees of freedom (sometimes give as v in discussion of power) and the noncentrality parameter. The noncentrality parameter is valuable in power analysis, because many of the equations for power analysis are stated in terms of the relevant estimate of the noncentrality parameter.

Power Analysis

Power analysis can be conducted to determine whether an analysis already completed had sufficient power to find significance (sometimes referred to as *post hoc* power analysis) or it can be conducted when planning a study (*a priori*). The latter use of power analysis is by far the most common use of power analysis, so I will focus on that here. Typically, a researcher is interested in determining whether a given sample size from an existing study will have sufficient power or is interested in determining a sample size needed to have sufficient power. In either case, one needs to specify the effect size expected, but a range of standard values can always be used to obtain a range of power estimates or sample sizes.

Power depends on N, α, and the effect size. So, power can be estimated as long as we know these three values. Because α (the chosen Type I error rate) is usually .05 two-tailed, that is easy. If we already have a data set or have already conducted the study, we know N and may know the effect size. It is simple then to estimate power for an analysis that has already been conducted, and many software packages, at least for some procedures, allow the user to request a power estimate with a particular analysis. If the goal is

[1] For the F test and χ^2, the noncentrality parameter is usually give as λ (lambda).
[2] The text (Myers, Well, & Lorch, 2010) deviates from the use of N elsewhere in the text and uses the total sample size as n.

find the sample size needed when planning a study, then we must request or assume a particular value for power that we desire. Most typically this is arbitrarily chosen to be .80 (Cohen, 1988), although occasionally, I see researchers use .90. The remaining unknown for a priori power analysis then is the effect size. The effect size can be based on prior research, although it is important to keep in mind that effect sizes have sampling variability just as any other sample value (Simonsohn, 2015). Meta analyses that combine effect sizes from multiple studies may be more reliable. Another common, and in some ways preferable, approach is to use a range of effect sizes, estimating the N needed for each effect size. Often, researchers input the widely-used conventional values originally suggested by Cohen (1962; see Cohen, 1992, for a summary). To estimate the power probability for a t test, we could use the computation given below.

$$\text{power} = 1 - p\left\{ z \le \left[\frac{t' - \delta}{\sqrt{1 + \frac{(t')^2}{2v}}} \right] \right.$$

In the equation, p is the probability associated with a particular z value, z is the normal score value, t' is the critical value for t, usually based on two-tailed test with α = .05, δ is the noncentrality parameter, and v "nu" is the degrees of freedom ($n_1 + n_2 - 2$ for the independent samples t-test).

Software

As hand computations can become a bit tedious, power or sample size is usually estimated with the aid of special software programs. There are a variety of these available, and some are free. A review by Peng, Long, and Abaci (2012) summarizes the features and evaluates several of the most prominent. G*power is perhaps the most widely used, but it may not include all of the tests one might need (t-tests, correlation, regression, and some nonparametric tests). Perhaps more versatile as well as a more user-friendly package is SPSS SamplePower, although it is expensive.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hilsdale. NJ: Lawrence Earlbaum Associates.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,65*, 145-153.

Cohen, J. (1992) A power primer. *Psychological Bulletin, 112*, 155-159.

Hume, D. (1963). *An enquiry concerning human understanding.* Oxford: Oxford University Press. [Originally published 1798].

Peng, C. Y. J., Long, H., & Abaci, S. (2012). Power analysis software for educational researchers. *The Journal of Experimental Education, 80*, 113-136.

Popper, K. (1980). *The logic of scientific discovery, 10th edition.* London: Hutchinson.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559-569.