A One-Page Primer on **Statistical Power**

by Carlisle Rainey
based on Rainey (2025)

> *"Of course, investigators could not have known how underpowered their research was, as their training had not prepared them to know anything about power, let alone how to use it in research planning. One might think that after 1969, when I published my power handbook that made power analysis as easy as falling off a log, the concepts and methods of power analysis would be taken to the hearts of null hypothesis testers. So one might think. (Stay tuned.)" – Cohen (1990)*

What is statistical power?

Statistical power is the chance that your data lead you to *reject* the null hypothesis if that null hypothesis *is incorrect*. Power depends on the magnitude of the effect, so you must assume an effect for a power calculation. Claims about power have the form:

"the experiment has ____% power to detect an effect of ____."

Why does power matter to the researcher?

You are a clever theorist and hypothesize that *X* increases *Y*; then the null hypothesis is that *X* does not increase *Y*. You design an experiment and statistical test. You plan to reject the null hypothesis if the p-value is less than 0.05. Two types of errors can occur.

- *Type I*: First, you might incorrectly reject the null hypothesis. That is, the null is true, but your p-value is less than 0.05 and you reject it.
- *Type II*: Second, you might incorrectly fail to reject the null hypothesis. That is, the null is false (i.e., the research hypothesis is correct), but your p-value is greater than 0.05 and you do not reject the null.

Let's think more on this second "error." You've paid all the costs of designing and fielding an experiment. But in the end, you cannot distinguish between your research hypothesis and the null hypothesis. This is a **wasted opportunity**. Researchers care about statistical power because they do not want to waste their time, effort, and money.

Why does power matter to the reader?

When reading a paper, why should you care about power?

- *My perspective:* Power doesn't matter after running the experiment. All relevant information is in the confidence interval. Caveat: power *does* change the meaning of "not significant," but this information is in the CI.
- *Others' perspective:* Power matters greatly after running the experiment. Authors and reviewers dismiss papers without significant findings. This significance filter combined with low power causes published estimates to diverge from the truth.

How can I compute statistical power?

To compute statistical power, first ask: "for what effect?" Identify the range of substantively interesting effects and choose the smallest effect in this range. This is the smallest effect of substantive interest (SESOI). The experiment should be well-powered for the SESOI.

Fancy software can compute power for you, but I like simple rules of thumb. Experiments have:

- 80% power to detect effects that are 2.5 times the standard error, and
- 95% power to detect effects that are 3.3 times the standard error.

As a reader, you can simply multiply the reported standard errors by 2.5 and ask yourself: "is this effect substantively small?" If 2.5 times the reported standard error is a substantively large effect, then the experiment is underpowered.

As a researcher, you can predict the standard error of a simple two-group, difference-in-means comparison using $SE = \frac{2\sigma}{\sqrt{2n}}$, where σ is the standard deviation of the outcome variable and n is the sample size *per condition*. You can obtain a suitable value of σ from an existing study of the same outcome and population. Or from a pilot study, if you plan to use one. As a too-crude shortcut in a pinch, you might use range/4.

Regression adjustment can increase power; it shrinks the standard error by a factor of about $\sqrt{1 - \widetilde{R^2}}$, where $\widetilde{R^2}$ is the $\widetilde{R^2}$ of the regression of the predictors on the outcome variable. Unless predictors are *highly* predictive of the outcome, adjustment doesn't matter much. However, in pre–post designs, the researcher measures the outcome before and after the treatment and adjusts for the pre-treatment measure. This shrinks the standard error at least 30%, typically 50%, and perhaps 70%. You can obtain a suitable value of $\widetilde{R^2}$ from a pilot study, if you have one.

What should I read next?

Cohen (1988) provides the foundational introduction to power analysis; Cohen (1990) offers an engaging personal and historical perspective. DeGroot and Schervish (2010, ch. 9) and Casella and Berger (2002, ch. 8) provide a thorough technical discussion. Jones and Tukey (2000) discuss Type II errs as "wasted opportunities."

Bloom (1995) offers a practical and intuitive approach using the minimum detectable effect. Rainey (2025) updates Bloom's perspective. Greenland et al. (2016) offer an accessible discussion organized around potential misunderstandings. Cohen (1992), Lenth (2001), and Meyvis and van Osselaer (2018) also offer practical guides.

Rainey (2014), McCaskey and Rainey (2015), and Lakens, Scheel, and Isager (2018, see esp. pp. 261-263) discuss the SESOI. Lovakov and Agadullina (2021) offer empirical rules for effect sizes derived from social psychology. Leon, Davis, and Kraemer (2011) and Albers and Lakens (2018) explain why pilot data should not be used to estimate the effect for the power calculation, though Marco Perugini, Gallucci, and Costantini (2014) propose a conservative alternative.

Hoenig and Heisey (2001) explain why power should not be used for interpreting results. Gelman and Carlin (2014) explain the pernicious effects of filtering on statistical significance when power is low. Arel-Bundock et al. (2025) show that observed power in political science is very low. Ioannidis, Stanley, and Doucouliagos (2017) examine economics; Stanley, Carter, and Doucouliagos (2018) examine psychology.

Meyvis and van Osselaer (2018) discuss regression adjustment in the context of experiments. Clifford, Sheagley, and Piston (2021) show huge boosts in power of the pre–post design; Jordan, Ollerenshaw, and Trexler (2025) validate their findings.

Lakens (2022) offers a broader perspective on sample size justification beyond statistical power. Blair et al. (2019) and Blair, Coppock, and Humphreys (2023) offer a useful conceptual framework and software for comprehensive evaluation of research designs.