Statistical Power: An intuitive guide

Prepared by Prof. Soudeep Deb

August 2026

The basic idea

Imagine that a company launches a pilot study to evaluate a new product, process, or policy. There are two possible realities:

- **Nothing meaningful has changed.** Any apparent improvement in the sample is simply random fluctuation.
- **A meaningful improvement really exists.** The sample should ideally provide enough evidence for us to recognize it.

A hypothesis test draws a line between results that are treated as ordinary random variation and results considered sufficiently unusual under the "nothing has changed" explanation. This line is called the **critical value** or **decision threshold**.

We deliberately set the threshold so that false positives (i.e., the type I error) are uncommon. If the significance level is 5%, then a result is allowed to cross the threshold only about 5% of the time when the null hypothesis is true. This false-alarm probability is commonly denoted as α.

However, a cautious threshold creates another risk: a real improvement might exist and yet the sample might not cross the threshold. That is a **Type II error**. Its probability is usually denoted by β.

Power is the probability that the study crosses the decision threshold when a specified, meaningful effect really exists. Thus,

$$\boxed{\text{Power} = 1 - \beta.}$$

A study with 80% power has an 80% chance of detecting the effect it was designed to detect, if that effect truly exists.

Power is best considered *before* collecting data. Managers must specify what size of effect would matter in practice – for example, a five-point increase in satisfaction, a 25% reduction in process variability, a ten-percentage-point increase in conversion etc. Power is then used to choose a sample size capable of detecting that effect. Generally, there are four aspects that determine power:

1. **Effect size:** Larger improvements are easier to notice.
2. **Sample size:** More observations reduce random noise.
3. **Variability:** Consistent data make a signal easier to distinguish.
4. **Decision threshold:** A stricter false-alarm rule reduces power.

It is important to recognize that power is not the probability that the alternative hypothesis is true. Rather, it describes how a testing procedure behaves over many repetitions under an assumed alternative effect. This is illustrated with two case studies below: one related to a test of mean and another for variance.

Example 1: Power when evaluating an average

Three Broomsticks' new beverage

The Three Broomsticks Inn is testing a new butterbeer-inspired beverage. The manager, Madam Rosmerta, will launch it only if its population average satisfaction score is above the current benchmark of 70 points. Past research suggests that individual ratings have a standard deviation of approximately 10 points. The team plans to survey 25 independently selected customers.

Naturally, the business decision can be expressed by the following hypotheses:

$$H_0 : \mu \leqslant 70 \qquad \text{against} \qquad H_1 : \mu > 70.$$

This is an **upper-tailed** test: only an unusually high sample average supports launch. At the 5% significance level, we know that the decision threshold should be set as the z-score exceeding the upper 5% quantile, i.e., 1.65. Equivalently, the rejection region is

$$\bar{X} > 70 + 1.65 \left(\frac{10}{\sqrt{25}} \right) = 73.3.$$

In words, the sample average must exceed 73.3 before Madam Rosmerta treats the improvement as more than ordinary sampling fluctuation.

One must however realize that the sample average can stay below 73.3 even when the true mean is more than 70. This will be called the type II error, as the alternative hypothesis is not being detected correctly. Hence, the concept of *power* asks the question: "what is the chance of the rejection region detecting such a deviation correctly?"

To understand this better, turn attention to Figure 1 where the rejection region of the test is the right side of the decision threshold (dotted vertical line). The sampling distribution under the null hypothesis (taking $\mu = 70$) is shown using navy-colored curve, while the sampling distribution under an alternative scenario (taking $\mu = 75$) is presented using an orange-colored curve.

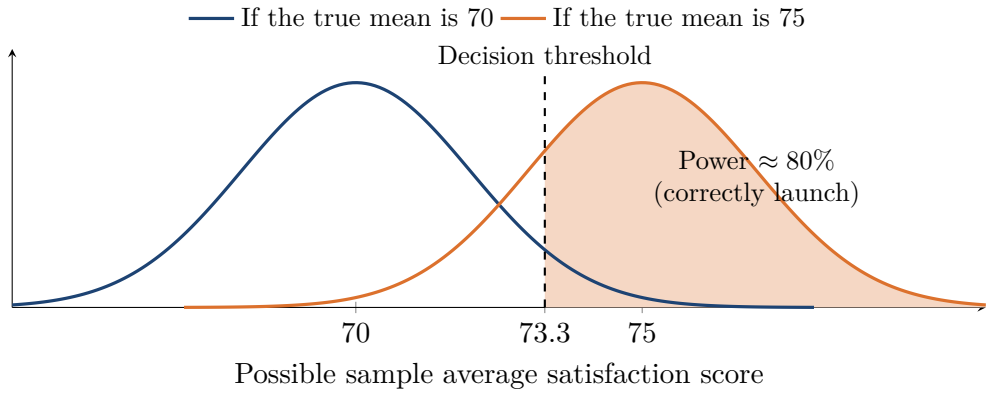

Figure 1: Power is the orange area to the right of the decision threshold. Curves are only presented as illustrative sampling distributions.

Based on the above discussion, it is clear that the area of the rejection region under the null distribution is 5%. The area of the same region under the alternative distribution is

$$P\left(\bar{X} > 73.3 \mid \mu = 75\right) = P\left(Z > \frac{73.3 - 75}{10/\sqrt{25}}\right) = P\left(Z > -0.85\right) \approx 80.4\%$$

Hence, if the true mean really is 75, the probability that the test will be rejected is approximately 80.4%; which indicates that a sample of 25 is well suited (with power of nearly 80.4%) of detecting

a deviation of 5 units. Of course, it is not equally capable of detecting smaller deviations. For example, if the true mean were 72, power would be only about 20.8% (confirm by similar calculations as above).

An associated problem is that of sample size determination for pre-specified targets of type I error and type II error. For a normally distributed outcome with a known standard deviation, if one wants to ensure a type I error α along with a type II error β corresponding to a deviation of d units from the null value, then the required sample size is

$$n \approx \left[\frac{(z_{1-\alpha} + z_{1-\beta})\ \sigma}{\text{deviation}} \right]^2.$$

As an illustration, suppose the objective is to ensure the test is developed at 1% level of significance; and should not have more than 2% type II error when there is a 3-point deviation in the satisfaction score. Thus, we may use

$$z_{1-\alpha} = z_{0.99} = -2.58, \quad z_{1-\beta} = z_{0.98} = -2.05,$$

which indicates that the required sample size is

$$n \geqslant \left[\frac{(-2.58 - 2.05)\ 10}{3} \right]^2 \approx 238.2.$$

That is, a minimum of 239 sample observations should be taken.

Example 2: Power when evaluating consistency

A butterbeer-bottling process

In the Three Broomsticks Inn, Madam Rosmerta is planning to buy a premium butterbeer-bottling machine. She is less concerned with the average fill (which is known to be adhering to the required accuracy) than with consistency (which is yet untested). The existing process has a standard deviation of 4 ml, and she is planning to adopt the new machine only if it reduces variability.

It is straightforward to see that it is a **lower-tailed** test because an unusually small sample variance provides evidence of better consistency. Thus, the hypotheses are

$$H_0 : \sigma^2 = 16 \qquad \text{against} \qquad H_1 : \sigma^2 < 16.$$

Suppose that Madam Rosmerta decides to take a sample of 30 independently filled bottles and subsequently use the data to conduct the required test at a 5% significance level. Then, using the concept of chi-squared test, the test should be rejected if the sample variance S^2 is such as that

$$\frac{(30-1)S^2}{16} < \chi^2_{29;0.95} = 17.71, \quad \text{i.e.,} \quad S^2 < 9.77.$$

Hence, the sample variance must be below approximately 9.77 ml^2 before Madam Rosmerta concludes that consistency has improved.

Now, suppose that the new machine truly reduces the standard deviation from 4 ml to 3 ml, so that the true variance falls from 16 to 9 ml^2. With 30 bottles, the probability of obtaining a sample variance below 9.77 is approximately 65.7%. This can be obtained as

$$P\left(S^2 > 9.77 \mid \sigma = 3\right) = P\left(\frac{(30-1)S^2}{9} < \frac{29 \times 9.77}{9} \right) = P\left(\chi^2_{29} < 31.48\right) \approx 65.7\%$$

Therefore, the study has only moderate power: even if the improvement is a magnitude of 1 unit in standard deviation, it would fail to detect it about one-third of the time. This can be visualized from the plots presented in Figure 2, where the blue curve represents the sampling distribution under the null hypothesis, green curve is used for the same under the alternative; and the colored area is the case where the test rejects under the alternative scenario (to reflect power).

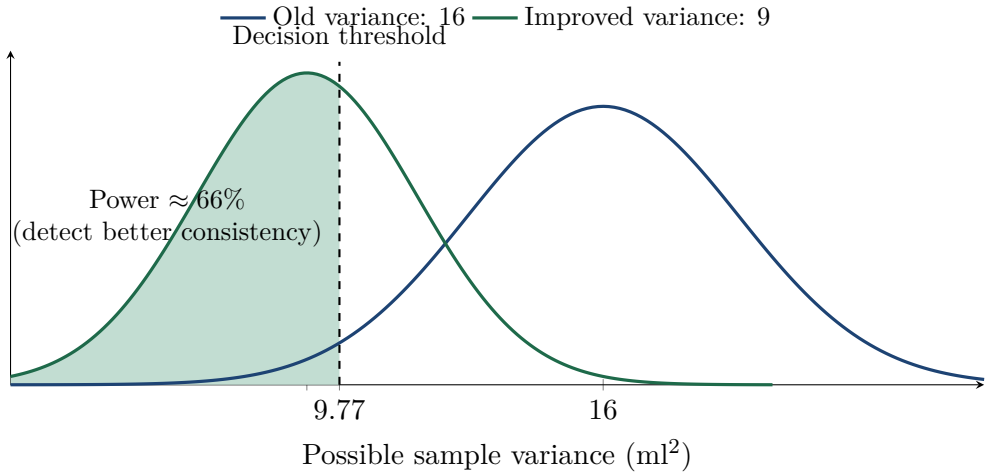

Figure 2: Power is the colored area to the left of the decision threshold. Curves are only presented as illustrative sampling distributions.

It is important to recognize that tests of variance rely strongly on the assumption that the underlying measurements are approximately normally distributed. Outliers or strongly skewed data can distort both the test and its power. One should also be able to identify that increasing the sample size would increase the power under same scenario: for example, a recalculation of the decision rule and the power would indicate that $n = 42$ bottles raises power to about 80.6% for the same situation. The managerial message is simple: reliably demonstrating improved consistency may require more observations than intuition suggests. On that note, one should observe that the sample-size calculation for a test of variance does not have a simple closed-form solution. This is because the degrees of freedom and the relevant chi-square critical value both depend on the sample size. The usual procedure is therefore to calculate power for successive values of n and select the smallest sample size that provides the desired power. To understand this, let us take a lower-tailed test with $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 < \sigma_0^2$. Consider the significance level (or type I error) α. Suppose, we want to control type II error at β (equivalently, power of $1 - \beta$) corresponding to an alternative value σ_1^2.

As we have seen above, for sample size n, the power at the alternative variance $\sigma_1^2 < \sigma_0^2$ is

$$\pi(n) = P\left[\chi^2_{n-1} < \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{n-1,1-\alpha}\right],$$

where $\chi^2_{n-1,q}$ is the upper-q-quantile of a chi-square distribution with $n - 1$ degrees of freedom. Then, the required sample size for the given conditions is the minimum value of n for which

$$P\left[\chi^2_{n-1} < \frac{\sigma_0^2}{\sigma_1^2}\chi^2_{n-1,1-\alpha}\right] \geqslant 1 - \beta.$$

This cannot be solved without numerical exploration with a software. For example, in the aforementioned butterbeer-bottling process, if we want to ensure a type I error of 5% and a type II error of 1% for wrongfully accepting the case of $\sigma = 3$, then we need to find the smallest n for which $P\left[\chi^2_{n-1} < 16\chi^2_{n-1,0.95}/9\right] \geqslant 0.99$. It can be checked in Excel that the minimum such value of n is 95.

For an upper-tailed test or a two-tailed test, similar techniques can be adopted.