Power and Sample Size, Part 1

Guidance for the Right Size and Right Approach to Sample Size

By Carol Parendo

Q How many samples should I test?

A Selecting the correct sample size requires consideration of several variables. Test too few and it loses statistical power and the ability to detect a *meaningful* difference. Test too many and risk wasting resources and impacting the project schedule. The key is detecting a meaningful difference, given reasonable error probabilities. To accomplish this, the sample must sufficiently represent the population for hypothesis testing. Hypothesis tests determine if something is statistically the same (null hypothesis) or different (alternative hypothesis). Sample sizes must be able to detect a practical minimum difference for a given power. This article discusses the inputs for these statistical calculations, followed by an example to solidify the concepts.

The drivers for determining sample size are intuitive. Larger sample sizes will detect a smaller difference. There are diminishing returns as the sample size increases. Data type also impacts sample size. Continuous variables require a much smaller sample size than pass/fail (attribute) samples. Estimates for the population such as standard deviation or proportion defective are also necessary. If variability can be reduced, sample size can also be reduced.

This intuitive understanding can also be expressed through a series of equations. Starting with Equation 26 from the practice for calculating and using basic statistics (E2586):

$$\hat{\theta} \pm z_{1-\alpha/2} \times se(\hat{\theta})$$

where:

$\hat{\theta}$ is the test statistic;

$Z_{1-\alpha/2}$ is the value from the standard normal table at a given α; and

$se(\hat{\theta})$ is the standard error of the test statistic.

Next, rewrite this equation for a specific case of continuous data where the test statistic is the average of the sample ($\bar{X}$) of size *n* and the standard deviation (σ) is known.

$$\bar{X} \pm \boxed{z_{1-\alpha/2} \times \sigma/\sqrt{n}} \quad \textbf{Margin of Error (ME)}$$

The margin of error is the minimum detectable difference for the hypothesis test. Then, solving for sample size (*n*) yields:

$$n = \left[\frac{z_{1-\alpha/2}\,\sigma}{ME}\right]^2$$

This equation for sample size (*n*) aligns with the intuition mentioned earlier. Sample size (*n*) and the margin of error (*ME*) are inversely proportional, so a tighter margin of error will drive a higher sample size. Conversely, sample size (*n*) and population standard deviation (σ) are directly proportional, and a higher standard deviation will drive a higher sample size.

In addition to margin of error, the maximum chance of obtaining an incorrect conclusion from a hypothesis test needs to be specified. There are two possibilities for making an incorrect conclusion. A false positive is called a type I error and a false negative is called a type II error. The probability of a type I error is designated as α, and the probability of a type II error is designated as β. Power is defined as 1 – β. **Table 1** depicts these errors when using a sample to estimate the truth for the population.

An appropriate α (type I error) and β (type II error) need to be determined for the situation. A common default value for α is .05. For β, a common value is .10. As stated earlier, power is 1 - β, so specifying a power of .90 is common.

		The Truth (Based on the Entire Population)	
		Same	Different
Your Conclusion (Based on Your Sample)	Same (Non-Significant)	Correct	Incorrect Type II Error
	Different (Significant)	Incorrect Type I Error	Correct

Table 1 — Type I and Type II Errors

Continued on page 54

Continued from page 52

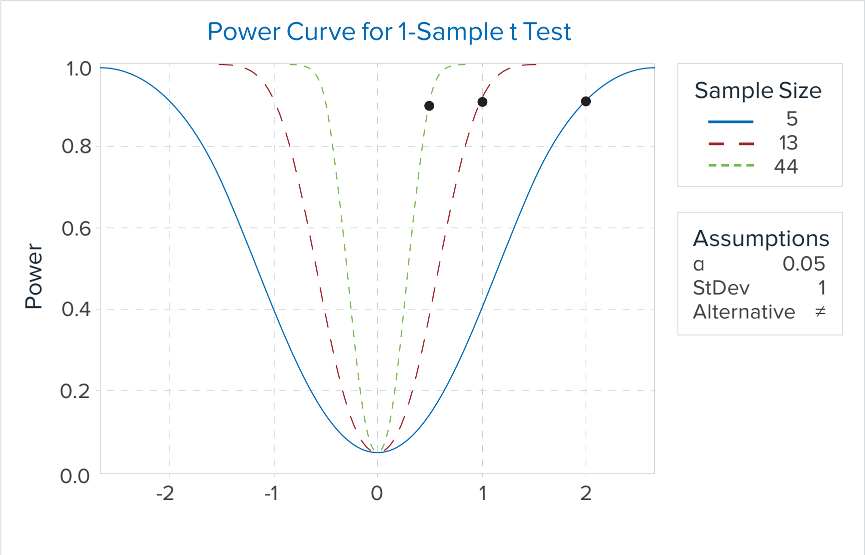

Figure 1 — Power Curve Example for a 1-Sample t Test

Adjust these values according to the specific risks associated with making an error. When incorporating the β risk, the equation is modified further, thus increasing the sample size.

$$n = \left[\frac{(z_{1-\alpha/2} + z_{1\text{-}\beta})\sigma}{ME}\right]^2$$

where:

n is the sample size;

$z_{1-\alpha/2}$, $z_{1\text{-}\beta}$ are the values from the standard normal table at a given α or β;

σ is the population standard deviation or an estimate the population standard deviation; and

ME is the margin of error or the minimum detectable difference.

The *Z* assumes a known σ (population standard deviation). For smaller sample sizes, it is important to replace the *Z* with a *t*. This calculation increases in complexity since the *t* value is dependent on sample size, so this method typically employs statistical software such as Minitab®.

A power curve can express this calculation visually. **Figure 1** shows a power curve for a 1-sample t-test using common inputs (two-tailed, α = .05, power = .90, standard deviation = 1.0, *ME* = .5, 1.0, and 2.0).

The points in **Figure 1** show the different intersections that meet the required .9 power condition. Power changes as we depart from 0 difference (or the null hypothesis). **Figure 1** specifically shows that the sample size increases from 5 to 13 to 44 when the minimum detectable difference decreases from 2 to 1 to .5. This illustrates the trade-off between sample size and detecting a smaller difference.

An example of using power and sample size in practice follows.

Problem: A company advertises that their boxes of cereal contain 13 oz of product. Engineering recently completed a repair on the filling equipment. You must verify that this did not affect the (mean) fill weight and it remains on target at 13 oz. Engineering states that a mean weight within ± .1 oz is of no practical difference. Engineering does not know the standard deviation of the weights and states that the range of weights is 12.5–13.5 oz. It is determined to use a difference of .1 and that α = .10 and power = .80 is reasonable given the risk.

Sampling Solution: You will need to know the if mean weight is at the target weight of 13 oz using a 1-sample t test. The standard deviation is estimated from the range of 1 oz. For measures that are typically normally distributed, a reasonable assumption is to use 1/6 of the range to yield an estimated standard deviation of .167 oz. (Reference E122-17 Figure 1 for this rule of thumb to estimate.) The software calculates a minimum of 19 cereal boxes to be weighed for the hypothesis test. Engineering states they will weigh 20. Engineering provides the data in the same order it comes off the filling equipment in case of an unlikely scenario such as shifting or drifting of data.

Summary

The calculations and approach (through an example) to obtain proper data for hypothesis testing was discussed. Statistical software is readily available for precise calculations involving sample size for the hypothesis test of interest. Having a conceptual understanding of the drivers for power and sample size calculations is key. Statistical software provides prompts for entering a difference to detect or sample size. The software will calculate the parameter not entered. The allowable type I and type II errors are entered in as α and as power (the complement of β or 1 − β). Then it will ask for a parameter that is used to estimate the population. In the case of continuous data, it is the estimate of the population standard deviation. In the case of proportion data (pass/fail data), it is the estimate of the population proportion.

When creating new data, follow best practices such as recording run order. In the case of two samples or more, randomization or semi-randomization should be implemented if feasible. This is not possible when working with historical data or data that has already been provided.

The next Data Points column (part 2) will discuss the next step of hypothesis testing. ■

Carol Parendo is a technical fellow at Collins Aerospace with over 30 years of experience as a mechanical engineer and statistician in both the aerospace and medical fields. Parendo is a member at large on the executive subcommittee of the committee on quality and statistics (E11).

John Carson, Ph.D., senior statistician for Neptune and Co., is the Data Points column coordinator. He is a member of the committee on quality and statistics (E11) as well as the committees on petroleum products, liquid fuels, and lubricants (D02), air quality (D22), and environmental assessment, risk management, and corrective action (E50).