> **intro —** Introduction to power and sample-size analysis

Description

Power and sample-size (PSS) analysis is essential for designing a statistical study. It investigates the optimal allocation of study resources to increase the likelihood of the successful achievement of a study objective.

Remarks and examples   **stata.com**

Remarks are presented under the following headings:

This entry describes statistical methodology for PSS analysis and terminology that will be used throughout the manual. For a list of supported PSS methods and the description of the software, see [PSS] **power**. To see an example of PSS analysis in Stata, see *An example of PSS analysis in Stata*. For more information about PSS analysis, see Lachin (1981), Cohen (1988), Cohen (1992), Wickramaratne (1995), Lenth (2001), Chow, Shao, and Wang (2008), and Julious (2010), to name a few.

Power and sample-size analysis

Power and sample-size (PSS) analysis is a key component in designing a statistical study. It investigates the optimal allocation of study resources to increase the likelihood of the successful achievement of a study objective.

How many subjects do we need in a study to achieve its research objectives? A study with too few subjects may have a low chance of detecting an important effect, and a study with too many subjects may offer very little gain and will thus waste time and resources. What are the chances of achieving the objectives of a study given available resources? Or what is the smallest effect that can be detected in a study given available resources? PSS analysis helps answer all of these questions. In what follows, when we refer to PSS analysis, we imply any of these goals.

We consider prospective PSS analysis (PSS analysis of a future study) as opposed to retrospective PSS analysis (analysis of a study that has already happened).

Statistical inference, such as hypothesis testing, is used to evaluate research objectives of a study. In this manual, we concentrate on the PSS analysis of studies that use hypothesis testing to investigate the objectives of interest. The supported methods include one-sample and two-sample tests of means, variances, proportions, correlations, and more. See [PSS] **power** for a full list of methods.

Before we discuss the components of PSS analysis, let us first revisit the basics of hypothesis testing.

Hypothesis testing

Recall that the goal of hypothesis testing is to evaluate the validity of a hypothesis, a statement about a population parameter of interest θ, a target parameter, based on a sample from the population. For simplicity, we consider a simple hypothesis test comparing a population parameter θ with 0. The two complementary hypotheses are considered: the null hypothesis $H_0: \theta = 0$, which typically corresponds to the case of "no effect", and the alternative hypothesis $H_a: \theta \neq 0$, which typically states that there is "an effect". An effect can be a decrease in blood pressure after taking a new drug, an increase in SAT scores after taking a class, an increase in crop yield after using a new fertilizer, a decrease in the proportion of defective items after the installation of new equipment, and so on.

The data are collected to obtain evidence against the postulated null hypothesis in favor of the alternative hypothesis, and hypothesis testing is used to evaluate the obtained data sample. The value of a test statistic (a function of the sample that does not depend on any unknown parameters) obtained from the collected sample is used to determine whether the null hypothesis can be rejected. If that value belongs to a rejection or critical region (a set of sample values for which the null hypothesis will be rejected) or, equivalently, falls above (or below) the critical values (the boundaries of the rejection region), then the null is rejected. If that value belongs to an acceptance region (the complement of the rejection region), then the null is not rejected. A critical region is determined by a hypothesis test.

A hypothesis test can make one of two types of errors: a type I error of incorrectly rejecting the null hypothesis and a type II error of incorrectly accepting the null hypothesis. The probability of a type I error is $\Pr(\text{reject } H_0 | H_0 \text{ is true})$, and the probability of a type II error is commonly denoted as $\beta = \Pr(\text{fail to reject } H_0 | H_0 \text{ is false})$.

A power function is a function of θ defined as the probability that the observed sample belongs to the rejection region of a test for a given parameter θ. A power function unifies the two error probabilities. A good test has a power function close to 0 when the population parameter belongs to the parameter's null space ($\theta = 0$ in our example) and close to 1 when the population parameter belongs to the alternative space ($\theta \neq 0$ in our example). In a search for a good test, it is impossible to minimize both error probabilities for a fixed sample size. Instead, the type-I-error probability is fixed at a small level, and the best test is chosen based on the smallest type-II-error probability.

An upper bound for a type-I-error probability is a significance level, commonly denoted as α, a value between 0 and 1 inclusively. Many tests achieve their significance level—that is, their type-I-error probability equals α, $\Pr(\text{reject } H_0 | H_0 \text{ is true}) = \alpha$—for any parameter in the null space. For other tests, α is only an upper bound; see example 6 in [PSS] **power oneproportion** for an example of a test for which the nominal significance level is not achieved. In what follows, we will use the terms "significance level" and "type-I-error probability" interchangeably, making the distinction between them only when necessary.

Typically, researchers control the type I error by setting the significance level to a small value such as 0.01 or 0.05. This is done to ensure that the chances of making a more serious error are very small. With this in mind, the null hypothesis is usually formulated in a way to guard against what a researcher considers to be the most costly or undesirable outcome. For example, if we were to use hypothesis testing to determine whether a person is guilty of a crime, we would choose the

null hypothesis to correspond to the person being not guilty to minimize the chances of sending an innocent person to prison.

The power of a test is the probability of correctly rejecting the null hypothesis when the null hypothesis is false. Power is inversely related to the probability of a type II error as $\pi = 1 - \beta = \Pr(\text{reject } H_0 | H_0 \text{ is false})$. Minimizing the type-II-error probability is equivalent to maximizing power. The notion of power is more commonly used in PSS analysis than is the notion of a type-II-error probability. Typical values for power in PSS analysis are 0.8, 0.9, or higher depending on the study objective.

Hypothesis tests are subdivided into one sided and two sided. A one-sided or directional test asserts that the target parameter is large (an upper one-sided test $H\colon \theta > \theta_0$) or small ($H\colon \theta \leq \theta_0$), whereas a two-sided or nondirectional test asserts that the target parameter is either large or small ($H\colon \theta \neq \theta_0$). One-sided tests have higher power than two-sided tests. They should be used in place of a two-sided test only if the effect in the direction opposite to the tested direction is irrelevant; see *One-sided test versus two-sided test* below for details.

Another concept important for hypothesis testing is that of a *p*-value or observed level of significance. *P*-value is a probability of obtaining a test statistic as extreme or more extreme as the one observed in a sample assuming the null hypothesis is true. It can also be viewed as the smallest level of α that leads to the rejection of the null hypothesis. For example, if the *p*-value is less than 0.05, a test is considered to reject the null hypothesis at the 5% significance level.

For more information about hypothesis testing, see, for example, Casella and Berger (2002).

Next we review concepts specific to PSS analysis.

Components of PSS analysis

The general goal of PSS analysis is to help plan a study such that the chosen statistical method has high power to detect an effect of interest if the effect exists. For example, PSS analysis is commonly used to determine the size of the sample needed for the chosen statistical test to have adequate power to detect an effect of a specified magnitude at a prespecified significance level given fixed values of other study parameters. We will use the phrase "detect an effect" to generally mean that the collected data will support the alternative hypothesis. For example, detecting an effect may be detecting that the means of two groups differ, or that there is an association between the probability of a disease and an exposure factor, or that there is a nonzero correlation between two measurements.

The general goal of PSS analysis can be achieved in several ways. You can

- compute sample size directly given specified significance level, power, effect size, and other study parameters;
- evaluate the power of a study for a range of sample sizes or effect sizes for a given significance level and fixed values of other study parameters;
- evaluate the magnitudes of an effect that can be detected with reasonable power for specific sample sizes given a significance level and other study parameters;
- evaluate the sensitivity of the power or sample-size requirements to various study parameters.

The main components of PSS analysis are

- study design;
- statistical method;
- significance level, α;
- power, $1 - \beta$;

- a magnitude of an effect of interest or clinically meaningful difference, often expressed as an effect size, δ;

- sample size, N.

Below we describe each of the main components of PSS analysis in more detail.

Study design

A well-designed statistical study has a carefully chosen study design and a clearly specified research objective that can be formulated as a statistical hypothesis. A study can be observational, where subjects are followed in time, such as a cross-sectional study, or it can be experimental, where subjects are assigned a certain procedure or treatment, such as a randomized, controlled clinical trial. A study can involve one, two, or more samples. A study can be prospective, where the outcomes are observed given the exposures, such as a cohort study, or it can be retrospective, where the exposures are observed given the outcomes, such as a case–control study. A study can also use matching, where subjects are grouped based on selected characteristics such as age or race. A common example of matching is a paired study, consisting of pairs of observations that share selected characteristics.

Statistical method

A well-designed statistical study also has well-defined methods of analysis to be used to evaluate the objective of interest. For example, a comparison of two independent populations may involve an independent two-sample t test of means or a two-sample χ^2 test of variances, and so on. PSS computations are specific to the chosen statistical method and design. For example, the power of a balanced or equal-allocation design is typically higher than the power of the corresponding unbalanced design.

Significance level

A significance level α is an upper bound for the probability of a type I error. With a slight abuse of terminology and notation, we will use the terms "significance level" and "type-I-error probability" interchangeably, and we will also use α to denote the probability of a type I error. When the two are different, such as for tests with discrete sampling distributions of test statistics, we will make a distinction between them. In other words, unless stated otherwise, we will assume a size-α test, for which $\Pr(\text{reject} H_0 | H_0 \text{ is true}) = \alpha$ for any θ in the null space, as opposed to a level-α test, for which $\Pr(\text{reject } H_0 | H_0 \text{ is true}) \leq \alpha$ for any θ in the null space.

As we mentioned earlier, researchers typically set the significance level to a small value such as 0.01 or 0.05 to protect the null hypothesis, which usually represents a state for which an incorrect decision is more costly.

Power is an increasing function of the significance level.

Power

The power of a test is the probability of correctly rejecting the null hypothesis when the null hypothesis is false. That is, $\pi = 1 - \beta = \Pr(\text{reject } H_0 | H_0 \text{ is false})$. Increasing the power of a test decreases the probability of a type II error, so a test with high power is preferred. Common choices for power are 90% and 80%, depending on the study objective.

We consider prospective power, which is the power of a future study.

Clinically meaningful difference and effect size

Clinically meaningful difference and effect size represent the magnitude of an effect of interest. In the context of PSS analysis, they represent the magnitude of the effect of interest to be detected by a test with a specified power. They can be viewed as a measure of how far the alternative hypothesis is from the null hypothesis. Their values typically represent the smallest effect that is of clinical significance or the hypothesized population effect size.

The interpretation of "clinically meaningful" is determined by the researcher and will usually vary from study to study. For example, in clinical trials, if no prior knowledge is available about the performance of the considered clinical procedure, then a standardized effect size (adjusted for standard deviation) between 0.25 and 0.5 may be considered clinically meaningful.

The definition of effect size is specific to the study design, analysis endpoint, and employed statistical model and test. For example, for a comparison of two independent proportions, an effect size may be defined as the difference between two proportions, the ratio of the two proportions, or the odds ratio. Effect sizes also vary in magnitude across studies: a treatment effect of 1% corresponding to an increase in mortality may be clinically meaningful, whereas a treatment effect of 10% corresponding to a decrease in a circumference of an ankle affected by edema may be of little importance. Effect size is usually defined in such a way that power is an increasing function of it (or its absolute value).

More generally, in PSS analysis, effect size summarizes the disparity between the alternative and null sampling distributions (sampling distributions under the alternative hypothesis and the null hypothesis, respectively) of a test statistic. The larger the overlap between the two distributions, the smaller the effect size and the more difficult it is to reject the null hypothesis, and thus there is less power to detect an effect.

For example, consider a z test for a comparison of a mean μ with 0 from a population with a known standard deviation σ. The null hypothesis is $H_0\colon \mu = 0$, and the alternative hypothesis is $H_a\colon \mu \neq 0$. The test statistic is a sample mean or sample average. It has a normal distribution with mean 0 and standard deviation σ as its null sampling distribution, and it has a normal distribution with mean μ different from 0 and standard deviation σ as its alternative sampling distribution. The overlap between these distributions is determined by the mean difference $\mu - 0 = \mu$ and standard deviation σ. The larger μ or, more precisely, the larger its absolute value, the larger the difference between the two populations, and thus the smaller the overlap and the higher the power to detect the differences μ. The larger the standard deviation σ, the more overlap between the two distributions and the lower the power to detect the difference. Instead of being viewed as a function of μ and σ, power can be viewed as a function of their combination expressed as the standardized difference $\delta = (\mu - 0)/\sigma$. Then, the larger $|\delta|$, the larger the power; the smaller $|\delta|$, the smaller the power. The effect size is then the standardized difference δ.

To read more about effect sizes in Stata, see [R] **esize**, although PSS analysis may sometimes use different definitions of an effect size.

Sample size

Sample size is usually the main component of interest in PSS analysis. The sample size required to successfully achieve the objective of a study is determined given a specified significance level, power, effect size, and other study parameters. The larger the significance level, the smaller the sample size, with everything else being equal. The higher the power, the larger the sample size. The larger the effect size, the smaller the sample size.

When you compute sample size, the actual power (power corresponding to the obtained sample size) will most likely be different from the power you requested, because sample size is an integer. In the computation, the resulting fractional sample size that corresponds to the requested power is

usually rounded to the nearest integer. To be conservative, to ensure that the actual power is at least as large as the requested power, the sample size is rounded up. For multiple-sample designs, fractional sample sizes may arise when you specify sample size to compute power or effect size. For example, to accommodate an odd total sample size of, say, 51 in a balanced two-sample design, each individual sample size must be 25.5. To be conservative, sample sizes are rounded down on input. The actual sample sizes in our example would be 25, 25, and 50. See *Fractional sample sizes* in [PSS] **unbalanced designs** for details about sample-size rounding.

For multiple samples, the allocation of subjects between groups also affects power. A balanced or equal-allocation design, a design with equal numbers of subjects in each sample or group, generally has higher power than the corresponding unbalanced or unequal-allocation design, a design with different numbers of subjects in each sample or group.

One-sided test versus two-sided test

Among other things that affect power is whether the employed test is directional (upper or lower one sided) or nondirectional (two sided). One-sided or one-tailed tests are more powerful than the corresponding two-sided or two-tailed tests. It may be tempting to choose a one-sided test over a two-sided test based on this fact. Despite having higher power, one-sided tests are generally not as common as two-sided tests. The direction of the effect, whether the effect is larger or smaller than a hypothesized value, is unknown in many applications, which requires the use of a two-sided test. The use of a one-sided test in applications in which the direction of the effect may be known is still controversial. The use of a one-sided test precludes the possibility of detecting an effect in the opposite direction, which may be undesirable in some studies. You should exercise caution when you decide to use a one-sided test, because you will not be able to rule out the effect in the opposite direction, if one were to happen. The results from a two-sided test have stronger justification.

Another consideration: Dropout

During the planning stage of a study, another important consideration is whether the data collection effort may result in missing observations. In clinical studies, the common term for this is dropout, when subjects fail to complete the study for reasons unrelated to study objectives.

If dropout is anticipated, its rate must be taken into consideration when determining the required sample size or computing other parameters. For example, if subjects are anticipated to drop out from a study with a rate of R_d, an ad hoc way to inflate the estimated sample size n is as follows: $n_d = n/(1 - R_d)^2$. Similarly, the input sample size must be adjusted as $n = n_d(1 - R_d)^2$, where n_d is the anticipated sample size.

Sensitivity analysis

Due to limited resources, it may not always be feasible to conduct a study under the original ideal specification. In this case, you may vary study parameters to find an appropriate balance between the desired detectable effect, sample size, available resources, and an objective of the study. For example, a researcher may decide to increase the detectable effect size to decrease the required sample size, or, rarely, to lower the desired power of the test. In some situations, it may not be possible to reduce the required sample size, in which case more resources must be acquired before the study can be conducted.

Power is a complicated function of all the components we described in the previous section—none of the components can be viewed in isolation. For this reason, it is important to perform sensitivity analysis, which investigates power for various specifications of study parameters, and refine the

sample-size requirements based on the findings prior to conducting a study. Tables of power values (see [PSS] **power, table**) and graphs of power curves (see [PSS] **power, graph**) may be useful for this purpose.

An example of PSS analysis in Stata

Consider a study of math scores from the SAT exam. Investigators would like to test whether a new coaching program increases the average SAT math score by 20 points compared with the national average in a given year of 514. They do not anticipate the standard deviation of the scores to be larger than the national value of 117. Investigators are planning to test the differences between scores by using a one-sample t test. Prior to conducting the study, investigators would like to estimate the sample size required to detect the anticipated difference by using a 5%-level two-sided test with 90% power. We can use the `power onemean` command to estimate the sample size for this study; see [PSS] **power onemean** for more examples.

Below we demonstrate PSS analysis of this example interactively, by typing the commands; see [PSS] **GUI** for point-and-click analysis of this example.

We specify the reference or null mean value of 514 and the comparison or alternative value of 534 as command arguments following the command name. The values of standard deviation and power are specified in the corresponding `sd()` and `power()` options. `power onemean` assumes a 5%-level two-sided test, so we do not need to specify any additional options.

```
. power onemean 514 534, sd(117) power(0.9)
Performing iteration ...
Estimated sample size for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m != m0
Study parameters:
        alpha =    0.0500
        power =    0.9000
        delta =    0.1709
           m0 =  514.0000
           ma =  534.0000
           sd =  117.0000
Estimated sample size:
            N =       362
```

The estimated required sample size is 362.

Investigators do not have enough resources to enroll that many subjects. They would like to estimate the power corresponding to a smaller sample of 300 subjects. To compute power, we replace the `power(0.9)` option with the `n(300)` option in the above command.

```
. power onemean 514 534, sd(117) n(300)
Estimated power for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m != m0
Study parameters:
        alpha =    0.0500
            N =       300
        delta =    0.1709
           m0 =  514.0000
           ma =  534.0000
           sd =  117.0000
Estimated power:
        power =    0.8392
```

For a smaller sample of 300 subjects, the power decreases to 84%.

Investigators would also like to estimate the minimum detectable difference between the scores given a sample of 300 subjects and a power of 90%. To compute the standardized difference between the scores, or effect size, we specify both the power in the `power()` option and the sample size in the `n()` option.

```
. power onemean 514, sd(117) power(0.9) n(300)
Performing iteration ...
Estimated target mean for a one-sample mean test
t test
Ho: m = m0   versus   Ha: m != m0; ma > m0
Study parameters:
        alpha =     0.0500
        power =     0.9000
            N =        300
           m0 =   514.0000
           sd =   117.0000
Estimated effect size and target mean:
        delta =     0.1878
           ma =   535.9671
```

The minimum detectable standardized difference given the requested power and sample size is 0.19, which corresponds to an average math score of roughly 536 and a difference between the scores of 22.

Continuing their analysis, investigators want to assess the impact of different sample sizes and score differences on power. They wish to estimate power for a range of alternative mean scores between 530 and 550 with an increment of 5 and a range of sample sizes between 200 and 300 with an increment of 10. They would like to see results on a graph.

We specify the range of alternative means as a *numlist* (see [U] **11.1.8 numlist**) in parentheses as the second command argument. We specify the range of sample sizes as a *numlist* in the `n()` option. We request a graph by specifying the `graph` option.

```
. power onemean 514 (535(5)550), sd(117) n(200(10)300) graph
```
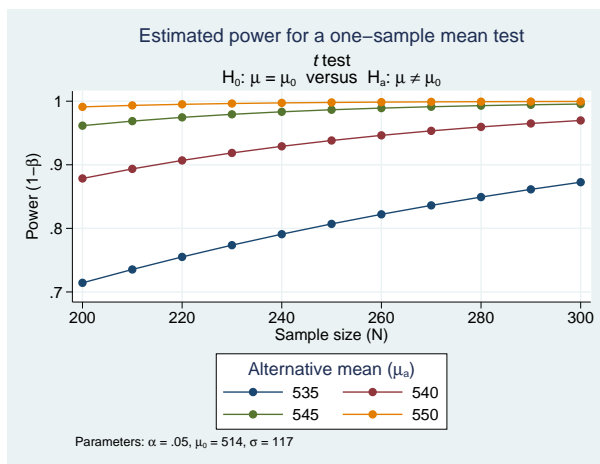

The default graph plots the estimated power on the y axis and the requested sample size on the x axis. A separate curve is plotted for each of the specified alternative means. Power increases as the

sample size increases or as the alternative mean increases. For example, for a sample of 220 subjects and an alternative mean of 535, the power is approximately 75%; and for an alternative mean of 550, the power is nearly 1. For a sample of 300 and an alternative mean of 535, the power increases to 87%. Investigators may now determine a combination of an alternative mean and a sample size that would satisfy their study objective and available resources.

If desired, we can also display the estimated power values in a table by additionally specifying the `table` option:

```
. power onemean 514 (530(5)550), sd(117) n(200(10)300) graph table
(output omitted )
```

The `power` command performs PSS analysis for a number of hypothesis tests for continuous and binary outcomes; see [PSS] **power** and method-specific entries for more examples. See the `stpower` command in [ST] **stpower** for PSS analysis of survival outcomes. Also, in the absence of readily available PSS methods, consider performing PSS analysis by simulation; see, for example, Feiveson (2002) and Hooper (2013) for examples of how you can do this in Stata.

References

Agresti, A. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: Wiley.

ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. 2002. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). *Journal of the American Medical Association* 288: 2981–2997.

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. 3rd ed. New York: Wiley.

Armitage, P., G. Berry, and J. N. S. Matthews. 2002. *Statistical Methods in Medical Research*. 4th ed. Oxford: Blackwell.

Barthel, F. M.-S., P. Royston, and M. K. B. Parmar. 2009. A menu-driven facility for sample-size calculation in novel multiarm, multistage randomized controlled trials with a time-to-event outcome. *Stata Journal* 9: 505–523.

Casagrande, J. T., M. C. Pike, and P. G. Smith. 1978a. The power function of the "exact" test for comparing two binomial distributions. *Journal of the Royal Statistical Society, Series C* 27: 176–180.

——. 1978b. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* 34: 483–486.

Casella, G., and R. L. Berger. 2002. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury.

Chernick, M. R., and C. Y. Liu. 2002. The saw-toothed behavior of power versus sample size and software solutions: Single binomial proportion using exact methods. *American Statistician* 56: 149–155.

Chow, S.-C., J. Shao, and H. Wang. 2008. *Sample Size Calculations in Clinical Research*. 2nd ed. New York: Marcel Dekker.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Erlbaum.

——. 1992. A power primer. *Psychological Bulletin* 112: 155–159.

Connor, R. J. 1987. Sample size for testing differences in proportions for the paired-sample design. *Biometrics* 43: 207–211.

Davis, B. R., J. A. Cutler, D. J. Gordon, C. D. Furberg, J. T. Wright, Jr, W. C. Cushman, R. H. Grimm, J. LaRosa, P. K. Whelton, H. M. Perry, M. H. Alderman, C. E. Ford, S. Oparil, C. Francis, M. Proschan, S. Pressel, H. R. Black, and C. M. Hawkins, for the ALLHAT Research Group. 1996. Rationale and design for the antihypertensive and lipid lowering treatment to prevent heart attack trial (ALLHAT). *American Journal of Hypertension* 9: 342–360.

Dixon, W. J., and F. J. Massey, Jr. 1983. *Introduction to Statistical Analysis*. 4th ed. New York: McGraw–Hill.

Feiveson, A. H. 2002. Power by simulation. *Stata Journal* 2: 107–124.

Fisher, R. A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* 10: 507–521.

Fleiss, J. L., B. Levin, and M. C. Paik. 2003. *Statistical Methods for Rates and Proportions*. 3rd ed. New York: Wiley.

Graybill, F. A. 1961. *An Introduction to Linear Statistical Models, Vol. 1*. New York: McGraw–Hill.

Guenther, W. C. 1977. Desk calculation of probabilities for the distribution of the sample correlation coefficient. *American Statistician* 31: 45–48.

Harrison, D. A., and A. R. Brady. 2004. Sample size and power calculations using the noncentral t-distribution. *Stata Journal* 4: 142–153.

Hemming, K., and J. Marsh. 2013. A menu-driven facility for sample-size calculations in cluster randomized controlled trials. *Stata Journal* 13: 114–135.

Hooper, R. 2013. Versatile sample-size calculation using simulation. *Stata Journal* 13: 21–38.

Hosmer, D. W., Jr., S. A. Lemeshow, and R. X. Sturdivant. 2013. *Applied Logistic Regression*. 3rd ed. Hoboken, NJ: Wiley.

Howell, D. C. 2002. *Statistical Methods for Psychology*. 5th ed. Belmont, CA: Wadsworth.

Irwin, J. O. 1935. Tests of significance for differences between percentages based on small numbers. *Metron* 12: 83–94.

Julious, S. A. 2010. *Sample Sizes for Clinical Trials*. Boca Raton, FL: Chapman & Hall/CRC.

Kahn, H. A., and C. T. Sempos. 1989. *Statistical Methods in Epidemiology*. New York: Oxford University Press.

Krishnamoorthy, K., and J. Peng. 2007. Some properties of the exact and score methods for binomial proportion and sample size calculation. *Communications in Statistics, Simulation and Computation* 36: 1171–1186.

Kunz, C. U., and M. Kieser. 2011. Simon's minimax and optimal and Jung's admissible two-stage designs with or without curtailment. *Stata Journal* 11: 240–254.

Lachin, J. M. 1981. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 2: 93–113.

———. 2011. *Biostatistical Methods: The Assessment of Relative Risks*. 2nd ed. Hoboken, NJ: Wiley.

Lenth, R. V. 2001. Some practical guidelines for effective sample size determination. *American Statistician* 55: 187–193.

Machin, D., M. J. Campbell, S. B. Tan, and S. H. Tan. 2009. *Sample Size Tables for Clinical Studies*. 3rd ed. Chichester, UK: Wiley–Blackwell.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 12: 153–157.

Newson, R. B. 2004. Generalized power calculations for generalized linear models and more. *Stata Journal* 4: 379–401.

Pagano, M., and K. Gauvreau. 2000. *Principles of Biostatistics*. 2nd ed. Belmont, CA: Duxbury.

Royston, P., and F. M.-S. Barthel. 2010. Projection of power and events in clinical trials with a time-to-event outcome. *Stata Journal* 10: 386–394.

Saunders, C. L., D. T. Bishop, and J. H. Barrett. 2003. Sample size calculations for main effects and interactions in case–control studies using Stata's nchi2 and npnchi2 functions. *Stata Journal* 3: 47–56.

Schork, M. A., and G. W. Williams. 1980. Number of observations required for the comparison of two correlated proportions. *Communications in Statistics, Simulation and Computation* 9: 349–357.

Snedecor, G. W., and W. G. Cochran. 1989. *Statistical Methods*. 8th ed. Ames, IA: Iowa State University Press.

Tamhane, A. C., and D. D. Dunlop. 2000. *Statistics and Data Analysis: From Elementary to Intermediate*. Upper Saddle River, NJ: Prentice Hall.

van Belle, G., L. D. Fisher, P. J. Heagerty, and T. S. Lumley. 2004. *Biostatistics: A Methodology for the Health Sciences*. 2nd ed. New York: Wiley.

Wickramaratne, P. J. 1995. Sample size determination in epidemiologic studies. *Statistical Methods in Medical Research* 4: 311–337.

Winer, B. J., D. R. Brown, and K. M. Michels. 1991. *Statistical Principles in Experimental Design*. 3rd ed. New York: McGraw–Hill.

Also see

[PSS] **GUI** — Graphical user interface for power and sample-size analysis

[PSS] **power** — Power and sample-size analysis for hypothesis tests

[PSS] **power, table** — Produce table of results from the power command

[PSS] **power, graph** — Graph results from the power command

[PSS] **unbalanced designs** — Specifications for unbalanced designs

[PSS] **Glossary**

[ST] **stpower** — Sample size, power, and effect size for survival analysis