Focus on Data

Focus on Data: Statistical Design of Experiments and Sample Size Selection Using Power Analysis

Johannes Ledolter[1,3] and Randy H. Kardon[2,3]

[1]Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, Iowa, United States
[2]Department of Ophthalmology and Visual Sciences, University of Iowa, Iowa City, Iowa, United States
[3]Center for the Prevention and Treatment of Visual Loss, Iowa City VA Health Care System, Iowa City, Iowa, United States

Correspondence: Johannes Ledolter, Department of Business Analytics, Tippie College of Business, University of Iowa, Iowa City, IA 52242, USA; johannes-ledolter@uiowa.edu.

Received: May 6, 2020
Accepted: May 26, 2020
Published: July 9, 2020

Citation: Ledolter J, Kardon RH. Focus on data: Statistical design of experiments and sample size selection using power analysis. *Invest Ophthalmol Vis Sci.* 2020;61(8):11. https://doi.org/10.1167/iovs.61.8.11

Purpose. To provide information to visual scientists on how to optimally design experiments and how to select an appropriate sample size, which is often referred to as a power analysis.

Methods. Statistical guidelines are provided outlining good principles of experimental design, including replication, randomization, blocking or grouping of subjects, multifactorial design, and sequential approach to experimentation. In addition, principles of power analysis for calculating required sample size are outlined for different experimental designs and examples are given for calculating power and factors influencing it.

Results. The interaction between power, sample size and standardized effect size are shown. The following results are also provided: sample size increases with power, sample size increases with decreasing detectable difference, sample size increases proportionally to the variance, and two-sided tests, without preference as to whether the mean increases or decreases, require a larger sample size than one-sided tests.

Conclusions. This review outlines principles for good experimental design and methods for power analysis for typical sample size calculations that visual scientists encounter when designing experiments of normal and non-Gaussian sample distributions.

Keywords: design of experiments, statistical power, sample size, randomization, repeated measures, cluster designs

This tutorial in the "Focus on Data" series provides information on how to optimally design experiments and how to select an appropriate sample size, which is often referred to as a power analysis. A power analysis is of great importance when planning an experiment that has a reasonably good chance of detecting treatment effects if they exist. In addition, the size of the effect anticipated should be of practical importance and the experimental design should ensure reproducibility of results. Funding agencies and journals consider rigor and reproducibility as major criteria for funding and publication. Sample size determination is an important tool for learning about what can be achieved in a study. There is much more to a power analysis than to justify grant proposals. It gives the investigator perspective on whether a well-designed experiment is feasible and likely to accept or refute a hypothesis based on an estimated measurement variability and the effect size that is anticipated. A well-considered power analysis to justify the sample size is an essential requirement for all studies.

Designing a good experiment is not easy, as in the beginning there is much uncertainty. Although one may have a certain response in mind, it is uncertain how to best measure it. Factors that can affect the response may be anticipated, but there is usually uncertainty about the relevant level of each factor that should be tested and the optimal sampling interval for detecting a response. Also, some of the factors may interact in how they affect the response (for example, treatment effects may depend on age or sex), and the experimental design should make it possible to assess such interactions. Moreover, not every factor can be anticipated and incorporated into the experiment. A good experimental design with proper randomization can ensure that unanticipated, omitted factors will equally impact the treatments to be studied.

When one designs an experiment, one usually knows very little about the measurement variability and the effect size. It would be so much easier if one already knew the results. Although it is certainly true that the best time to design an experiment is after the results of the experiment have come in, knowing this does not help design the experiment. What helps is understanding sound experimental design principles, access to pilot data, and knowing the relevant prior literature as it relates to the questions that are being studied. Fortunately, one usually does not start from the very beginning but can build on the research that has been carried out before.

Good Principles of Experimental Design

Detailed discussions of important design principles can be found in books on the statistical design of experiments, such Box et al.,[1] Ledolter and Swersey,[2] and Montgomery.[3] The seminal contributions of Fisher[4] have shaped this field. The following are important statistical design principles:

- Replication. Observing a certain result just once or twice does not make it reliable. Natural variability is present everywhere; results from repeated trials on the same subject vary, and results from trials on different subjects vary even more. Replicating the experiment increases the reliability and rigor of the results.
- Randomization. Allocation of treatments randomly to experimental subjects ensures the validity of an inference in the presence of unspecified disturbances by making certain that the risk of such disturbances is spread evenly among the treatment groups. For example, in visual science, this might entail assigning different treatments to groups of patients at random, or randomizing the order of an assigned treatment to a group of subjects or eyes. Without randomization, treatment differences may be confounded with other variables that are not controlled by the experimenter. Anticipating confounding variables in advance can help tremendously in randomizing subjects based on these variables, such as age, sex, animals from the same litter or cage, severity of disease, or level at baseline.
- Blocking. Randomizing the assignment of treatments to subjects or eyes is important as it spreads the existing variability among subjects equitably across all treatments. However, the experimenter can do considerably better if the experimental subjects can be grouped into blocks, such that units are homogenous within the same block, but different across blocks. In the visual sciences, eyes can be blocked by subjects, or in the case of mice, by cage or litter. Responses on eyes from different subjects vary considerably, whereas the responses on eyes from the same subject are usually related with much smaller variability. When studying effects, one frequently treats only one eye, whereas keeping the other eye as a within-subject control. This approach assumes that the treatment will only affect the eye that receives it, which may not be the case in every situation. If the effect is restricted to the treated eye, then the large subject effect that affects both eyes in a similar way can be removed, resulting in an increase of the precision of the comparison, potentially making it more sensitive to detecting an effect, if one exists. Also, it is more efficient to design the experiment such that each treatment is applied to the same subject (or eye) at different time points in which the effect of the first treatment is no longer present and will not affect the second treatment effect. The consecutive arrangement of the treatments can always be randomized to make sure that treatment effects are not compromised by the order. A within-subject comparison of the effectiveness of a treatment or a drug is subject to fewer interfering variables than a comparison across different subjects.
- A multifactor design should be considered, instead of a one factor at-a-time experimental approach. A common, but inefficient, approach to studying the effects of several factors is to carry out successive experiments in which the levels of each factor are changed one at-a-time. Fisher[4] showed that a better approach is to vary the factors simultaneously and to study the response at each possible factor-level combination. Such approach makes it possible to learn about interaction effects (e.g., whether the effect on the response when changing one factor depends on the level of another factor).
- Sequential approach to experimentation. Each experiment contributes to one's understanding. The results of one experiment are critical to determine the next experimental steps. Hence only a portion of the overall research plan and budget should be spent on the initial experiment.

In medical research, investigators run experiments all the time, and evidence-based medicine relies on randomized experiments to confirm which of several treatments are the most effective.

The search for effective ways to design experiments and issues of sample size and statistical power are commonplace in scientific experimentation. If experiments are executed poorly, little, or even nothing, will be learned from the resulting data. Although it is true that most experiments increase knowledge (one usually learns "something" through experimentation), the experimenter wants to learn as efficiently as possible. Relatively few experimental runs (observations) are needed in efficient experimental designs to get precise estimates of the factor effects. Sir Ronald Fisher, the eminent statistician and scientist who developed this area, said that "a well-designed experiment may improve the precision of the results tenfold, for the same cost in time and labor" (Fisher,[4] page 217).

Power Analysis: Importance of Calculating the Required Sample Size

Prior to running an experiment, one needs to determine the sample size required to identify scientifically meaningful effects. In other words, one must address the question whether a certain sample size is sufficient to detect a specified response effect if it really exists. If the sample size is too small, observed effects may not be statistically significant and meaningful effects may not be uncovered, even if they do exist. Conversely, including too many samples when not necessary cannot only be a waste of resources, but can also expose subjects to unnecessary risks, and may reveal statistically significant, but clinically irrelevant, results.

It is very important to know whether the data that are collected from an experiment have a realistic chance of detecting meaningful effects. Consider, for example, an experiment on mice that studies the effect of a new drug or genetic treatment on improving visual function. Typically one knows how large an effect will be considered clinically meaningful (e.g., a 30%–40% effect size for a promising new drug being tested in preclinical studies). Research studies are expensive, and costs increase with the number of subjects that need to be recruited into the study. Prior to running the experiment, one must calculate the statistical power of detecting (practically) meaningful effects. For some planned experiments, detection may not be feasible; many more subjects may be needed to learn about clinically important effects. If one cannot afford the required sample sizes, one must restructure or abandon the problem in favor of problems that can be solved with the budget at hand. If there is little chance that meaningful effects can be detected, resources are better expended elsewhere. Although medical grant proposals typically require a section on sample size and power, these sections are usually written in a defensive manner to justify the experimental plan the investigator has settled on a long time ago. Often these sections are written to defend a prior assumption the investigator has, and they rarely assess critically whether the planned research is worth its cost or whether the effect size is appropriate. Many times they represent an intricate "song and dance" to justify why limited funds can be used to study something experimenters want to study anyway. Experimenters need to understand that sample size studies are there to help them determine if a question can reasonably be answered; sample size studies are not there to game the system to achieve funding.

Many statistics packages have modules for determining the appropriate sample sizes (see Appendix for some examples and links). Some programs are dedicated to this task exclusively, such as the sample size/power programs by Lenth.[5] Lenth's sample size applets (they are free, good, and easy to use) cover many different situations, including continuous outcome variables (with an emphasis on means and variances), categorical outcome variables (with an emphasis on proportions), and correlations. G*Power, developed by Faul et al.,[6,7] is another free software program available for both Macintosh and PC platforms.

Statistical power is defined as the probability that a data-based test will correctly reject a false null hypothesis (e.g., the means of two distributions are equal). The higher the statistical power, the smaller the type II error of not rejecting a false null hypothesis (false-negative result). Incorporated into the power analysis is also the specified type I error (false-positive result) of rejecting the null hypothesis when the difference was really only due to chance alone. Statistical power can be thought of as the probability of finding a difference in population characteristics when such differences actually exist. Of course, power increases with the magnitude of the differences one wants to detect; it is easier for data-based tests to detect larger differences than smaller differences. Experiments with low statistical power may not uncover meaningful effects, even if they do exist. A minimum level of statistical power must be sought, at least 80% or greater, to detect a specified practically relevant difference. The choice of power (e.g., 80%, 90%) is related to how certain one wants to be that the experimental design (e.g., sample size) is sufficient to detect a meaningful difference if one does exist. In designing clinical trials, the Consolidated Standards of Reporting Trials (CONSORT) has an agreed on CONSORT Statement, which is an evidence-based, minimum set of recommendations for reporting randomized trials (http://www.consort-statement.org). It offers a standard way for authors to prepare reports of trial findings, facilitating their complete and transparent reporting, and aiding their critical appraisal and interpretation. Sample size calculation is one of the key requirements.

Sample size selection for some typical problems that visual scientists encounter in their research are discussed next.[8]

Power Analysis: Detecting a Difference from a Known Mean of a Single Normal Distribution (One-Sample Situation)

We test the null hypothesis $H_0 : \mu = \mu_0$ against the one-sided (lower-tailed) alternative hypothesis $H_1 : \mu < \mu_0$. We test the research hypothesis whether or not an intervention reduces the mean from its current known value μ_0. When determining the appropriate sample size, we need to specify values for the four following items:

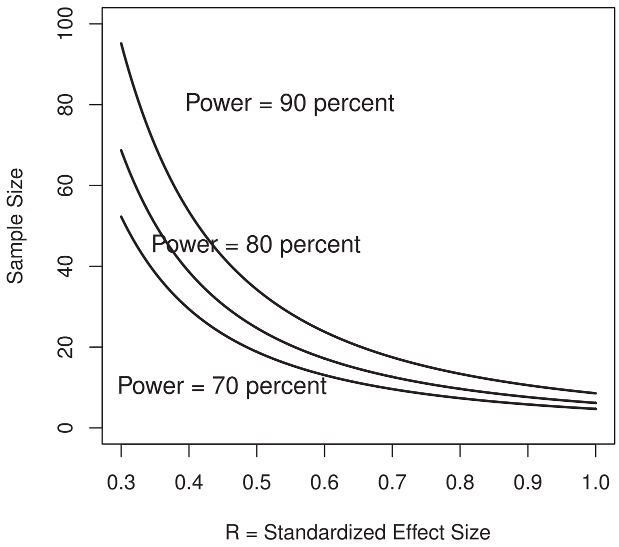

Figure 1. Plotting the required sample size against the standardized effect size (e.g., effect size divided by the SD of the measurements) for 5% significance level and 70%/80%/90% power. For anticipated larger effects, fewer samples are required, but if one is trying to achieve greater power (e.g., top line of 90% power), the required sample size increases.

- $\sigma = \sqrt{Var(Y)}$, the **SD** (standard deviation = square root of the variance) of the normally distributed measurement variable Y. Prior data in the literature or pilot data provide a planning value for the SD.
- The **significance level** (that is, the probability of falsely rejecting a true null hypothesis); usually $\alpha = 0.05$.
- The **power** (usually 0.80, or 80%) to detect a specified **meaningful change** (commonly referred to as **effect size**) $\delta = \mu_1 - \mu_0 < 0$. $\beta = 1 - Power$ (here $\beta = 1 - 0.8 = 0.2$) is the probability of a type II error of accepting the null hypothesis H_0 if the mean has indeed shifted to $\mu_1 = \mu_0 + \delta < \mu_0$.

Result: The required sample size for detecting a change δ with power $1 - \beta$ is

$$n = (z_\alpha + z_\beta)^2(\sigma/\delta)^2;$$

z_α and z_β are percentiles (z-scores) of the standard normal distribution; they can be looked up in normal probability tables. For 5% significance level, $z_{\alpha=0.05} = -1.645$; for 80% power and type II error of 0.20, $z_{\beta=0.20} = -0.8416$.

The required sample size decreases inversely with $R^2 = (\delta/\sigma)^2$. The ratio $R = |\delta|/\sigma$ expresses the size of the detectable meaningful change as a fraction of the SD; we refer to it as the standardized effect size. Figure 1 plots the sample size against R, for 5% significance and three different values of power (70%, 80%, and 90%). For given R, one can find graphically the sample size that is required to detect that change. Approximately 25 observations are needed to detect a change of half an SD with 80% power; fewer (19) observations are needed for 70% power, and more (35) observations are needed for 90% power.

The (result) equation involves five quantities, and the relationship between them can be displayed graphically in other ways. For example, for fixed sample size (and fixed significance level), the power can be graphed against the

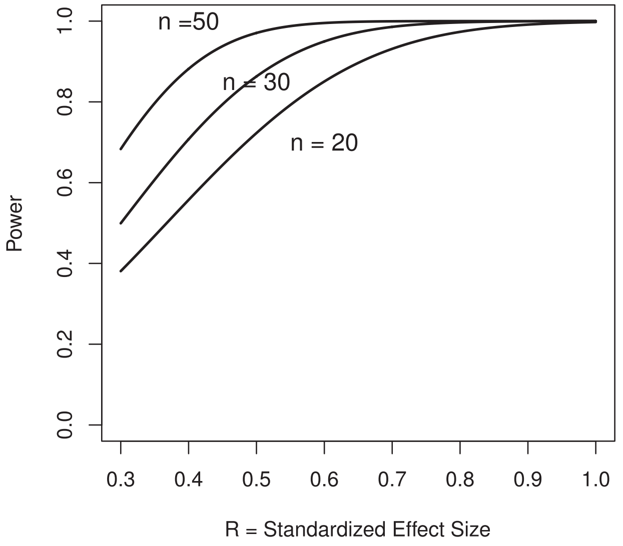

Figure 2. Plotting the power against the standardized effect size R (e.g., effect size divided by the SD of the measurements) for 5% significance level and three different sample sizes (20, 30, 50). Power decreases for anticipated smaller effects, and power increases for larger sample sizes. Fixing power at 80%, for example, one can detect a change of 0.35 (SD) with 50 samples, a change of 0.45 (SD) with 30 samples, and a change of 0.56 (SD) with 20 samples.

standardized effect size R. The graph in Figure 2 shows how power decreases for decreasing standardized effect size.

Facts to Remember.

- Sample size increases with power. The more power you want, the larger the sample size.
- Sample size increases with decreasing detectable difference. The smaller the difference or effect size you expect, the larger the sample size that will be required.
- Sample size increases proportionally to the variance. The larger the uncertainty of the outcome measurement (variability of a result), the larger the sample size must be. The sample size quadruples with a doubling of the SD.
- Tests are typically one-sided as one expects increases (or decreases) in the mean. Two-sided tests, without preference whether the mean increases or decreases, require a larger sample size than one-sided tests. For a two-sided test, the term z_α in the earlier noted result is replaced by $z_{\alpha/2}$. For $\alpha = 0.05$, $z_{\alpha/2} = -1.96$.

Example: For the general population, mean thickness of the inner retina is known to be 100 µm, based on prior research publications. The subject variability is large, with SD of approximately 20 µm. We are interested in whether individuals from a certain ethnic group have a thinner (smaller) mean retinal thickness. How many subjects from this ethnic group need to be studied to confirm with 80% power a reduced mean thickness of 5 µm? In this case, $\mu_0 = 100$ and $\sigma = 20$. For 5% significance, and 80% power to detect a reduction of $\delta = -5$ units, we need $n = (z_\alpha + z_\beta)^2(\sigma/\delta)^2 = (-1.645 - 0.8416)^2(20/-5)^2 \approx 100$ individuals.

Example: A new glaucoma strain of mice has been developed through breeding on a Black 6 background strain (C57BL/6). The investigators are interested in a power analysis to determine how many mice of the new strain are needed to test for a significant increase in intraocular pressure (IOP). From the literature, it is known that C57BL/6 mice at 7 months of age have IOP with mean 13.3 mm Hg and SD 1.25 mm Hg. An increase in IOP in the new strain would be considered significant if it were increased by 0.5 to 13.8 mm Hg. For 5% significance, and 80% power to detect an increase of $\delta = 0.5$mm Hg, we need $n = (z_\alpha + z_\beta)^2(\sigma/\delta)^2 = (-1.645 - 0.8416)^2(1.25/0.5)^2 = 39$ mice.

Assume we have planned on only 20 mice. What would the power be to detect an increase of 0.5 mm Hg? Solving the equation in the earlier noted result for β, we obtain $\beta \approx 0.44$ and a power of 56% (considerably less than the planned 80%). With 20 mice, what change could we detect with 80% power? Solving the equation in the earlier noted result for δ, we obtain $\delta \approx 0.70$mm Hg. There is a good chance that a change of 0.50 mm Hg stays undetected.

Applying this Result to Paired Comparisons. The result can be used in the paired (blocked) test with response $D = Y_2 - Y_1$, where Y_2 is the response under treatment 2 and Y_1 is the response under treatment 1. The two groups may reflect treatment and control, or after-treatment and baseline. An important aspect in the paired comparison is that both treatments are applied on the same subject, allowing us to express the treatment effect with the difference of the two measurements. After taking differences, the problem reduces to a one-sample comparison, and the previous result can be applied. All the researcher needs to provide is a planning value of the SD of the differences, $\sigma = \sqrt{Var(D)}$, the mean of the differences under the null hypothesis μ_0, and a meaningful detectable difference.

Example: Assume an experiment in which eyes of Black 6 mouse strain (C57BL/6) are treated with a pressure lowering eye drop. Drops are administered to one randomly selected eye of each mouse. The change in the IOP after and before treatment (D = treatment IOP – baseline IOP) reflects the effectiveness of the medication. Fortunately, a number of publications assess the variability of the difference in IOP measurements from the same eye at two different time points, and a planning value for the SD of such difference can be obtained from the literature. In the case of mice, $\sigma = \sqrt{Var(D)} \approx 1$ mm Hg. We wish to test whether the treatment is effective and whether the mean of treatment/baseline differences is less than 0. A reduction of 0.5 mm Hg is considered clinically significant. With this information, the number of mice should be $n = (z_\alpha + z_\beta)^2(\sigma/\delta)^2 = (-1.645 - 0.8416)^2(1/-0.5)^2 \approx 25$.

Power Analysis: Comparing Means of Two Independent Normal Distributions

We test the null hypothesis $H_0 : \mu_2 - \mu_1 = 0$ against the one-sided (lower-tailed) alternative $H_1 : \mu_2 - \mu_1 < 0$. Both group means, μ_1 and μ_2, are unknown and must be estimated from the sampled data. This makes the problem different from the one-sample situation discussed previously, in which one mean is known with certainty. For a two-sample comparison, we need to specify values for the following five quantities:

- σ_1 and σ_2 : **two SDs** that need not be equal
- **significance level**; usually $\alpha = 0.05$
- **power** (usually 0.80) to detect a **specified meaningful difference** (effect size) $\delta = \mu_2 - \mu_1 < 0$

Result 1: (Ledolter[8]). The required total sample size (for groups 1 and 2 together) is

$$N = (z_\alpha + z_\beta)^2[(\sigma_1 + \sigma_2)/\delta]^2.$$

The sample sizes of the two groups, n_1 and n_2, must be selected proportional to the SDs: $n_1 = \frac{\sigma_1}{\sigma_1+\sigma_2}N$ and $n_2 = \frac{\sigma_2}{\sigma_1+\sigma_2}N$.

Result 2: When the SDs are the same ($\sigma_1 = \sigma_2 = \sigma$), the sample size for either of the two groups is $n = 2(z_\alpha + z_\beta)^2(\sigma/\delta)^2$, for a combined sample size of $N = 2n = 4(z_\alpha + z_\beta)^2(\sigma/\delta)^2$.

For equal SDs, Figure 1 can be used to determine the required sample sizes, but the value on the y-ordinate must be doubled when obtaining the sample sizes for each of the two groups. In the one-sample case, the reference level is given with certainty. In the two-sample case, larger sample sizes are required as two means need to be estimated.

Example: Ledolter and Kardon[9] studied the average retinal nerve fiber layer (RNFL) thickness from an optic disc scan for both normal subjects and glaucoma patients on optimal treatment. As expected, the average RNFL thickness of normal subjects was considerably larger than that of glaucoma subjects. They also found that the variability in RNFL thickness among glaucoma patients (SD = 10 µm) was larger than that of normal subjects (SD = 8.5 µm). The larger SD of the glaucoma group is expected because there is a large range of disease severity and response to treatment affecting the thickness of the RNFL.

Let us assume that one wants to study the RNFL difference between normal subjects and glaucoma patients from a certain subgroup (such as age, sex, ethnic origin) for whom we have incomplete information on its mean RNFL thickness. Although group means are unknown, we have good reason to assume that group variabilities are similar to the ones from our earlier normal/glaucoma study. Suppose one plans for 80% power to detect a mean reduction of 5 µm. How many subjects should one sample?

For the group 1 of normal patients, $\sigma_1 = 8.5$ µm; for group 2 of glaucoma patients, $\sigma_1 = 10$ µm. For a detectable difference of interest $\delta = -5$ µm, 80% power ($\beta = 0.20$), and significance level $\alpha = 0.05$, the combined sample size from Result 1 is $N = (z_\alpha + z_\beta)^2[(\sigma_1 + \sigma_2)/\delta]^2 = (-1.645 - 0.8416)^2[(8.5 + 10)/-5]^2 \approx 85$. We should sample (8.5/18.5)85 = 39 healthy subjects and (10/18.5)85 = 46 glaucoma patients. We should sample more glaucoma patients as their variability is larger.

The two SDs are not that different. For further illustration, we take the larger SD as the planning value for the common SD. From Result 2, the sample size in each of the two groups is $n = 2(z_\alpha + z_\beta)^2(\sigma/\delta)^2 = 2(-1.645 - 0.8416)^2(10/-5)^2 \approx 50$, for a combined sample size of 100.

Example: A second example from the visual science considers an experiment that investigates whether a topical medication can reduce the IOP. The experiment compares a group of mice receiving the medication with another group of mice of the same strain receiving a placebo drop. The two groups are matched on similar levels of IOPs; effectiveness is measured by changes in IOP from baseline prior to receiving the treatment or placebo. We compare two groups: group 1 consisting of mice receiving the placebo, and group 2 receiving the treatment. The SD of differences in IOP taken on the same subject at different times is 1.16 mm Hg, and there are good reasons to assume that the SDs in the treatment and the placebo groups are about the same. If we want 80% power to detect a mean change of IOP of 0.5 mm Hg, $n = 2(z_\alpha + z_\beta)^2(\sigma/\delta)^2 = 2(-1.645 - 0.8416)^2(1.16/0.5)^2 \approx 34$, for a combined sample size of 68.

Power Analysis: Comparing Means of Two Independent Log-Normal Distributions

It is often easier for investigators to specify treatment effects as percentage changes in the means and variability in terms of coefficients of variation. The detectable effect of interest is then the proportionate change in the means, $f = \frac{E(Y_2)}{E(Y_1)} - 1$; for example, a 20% increase in the mean when $f = 0.2$, or a 30% decrease in the mean when $f = -0.3$. The measurements Y for each of the two groups are often log-normally distributed (the logarithm of the data sample transforms it to a normal distribution), with different means but equal coefficients of variation $c = \frac{\sqrt{Var(Y)}}{E(Y)}$.

Prior data can be used to check the distributional assumptions. A plot of the two (group-specific) histograms should show skewed log-normal distributions with long right tails. Histograms of the log-transformed data should be normal; they can have different means, but their SDs should be approximately the same. The SD of the untransformed measurements should be proportional to the mean, with the coefficients of variation in the two groups approximately the same. Their average can be taken as a planning value for c.

Result: (Van Belle and Martin[10]). The objective is to detect a 100f percent proportionate change in the means, and to do so with power $1 - \beta$. For two log-normal distributions with equal coefficients of variation c, the number of observations needed in each group is

$$n = 2(z_\alpha + z_\beta)^2\left[\frac{\sqrt{\log(1 + c^2)}}{\log(1 + f)}\right]^2.$$

Example: Activation of neurons by sensory stimuli follow a proportional law (referred to as the Weber-Fechner law[11,12]), and measures of sensitivity to stimuli tend to follow log-normal distributions. We are comparing two groups of mice: a normal group and one with a new, genetically engineered form of retinitis pigmentosa with damage to the rods and cones. The mice in each of the two groups are exposed to a series of different stimuli differing in light intensity, and the amplitude of the electroretinogram (ERG) is recorded in response to a flash of light at each intensity. Amplitudes at each intensity follow log-normal distributions with coefficient of variation $c = 0.30$.[13] We expect that the ERG response in the normal group will be larger than that of the retinitis pigmentosa group. We want 80% power to detect a 20% greater ERG response in the mean of the normal group. For $c = 0.30$, $\sqrt{\log(1 + (0.30)^2)} = 0.2936$; for $f = 0.20$, $\log(1+ 0.2) = 0.1823$. We need $n = 2(-1.645 - 0.8416)^2(0.2936/0.1823)^2 = 32$ mice in each group, for a total of 64 mice.

Power Analysis: Cluster Designs for Comparing Two Means

For a two-group comparison with equal variances the sample size for each of the two groups is $n = 2(z_\alpha + z_\beta)^2(\sigma/\delta)^2$; see our earlier Result 2. This result assumes that the two treatments are assigned to the experimental units (e.g., subjects, mice, and others) at random.

However, sometimes the randomization is carried out on **clusters** that consist of groupings of the experimental units. Clusters may be cages of animals, and experimental units could be mice. Clusters may be patients, and experimental units could be eyes. The randomization is at the cluster level: the treatment groups (experimental and control) are assigned to clusters at random, and each of m experimental units in a cluster is assigned to the same treatment. Although the data of interest comes from the experimental units in the two experimental groups, the randomization is carried out on the clusters. In the example with patient eyes, we may assign $n = 10$ patients each to one of the two treatments, for a total of 20 patients. For cluster size $m = 2$, this generates a total of 40 eyes, with 20 eyes for each treatment.

Usually subjects from the same cluster tend to be alike. Because observations from the same cluster are most likely correlated, with intracluster correlation coefficient $\rho > 0$, the m observations in a cluster do not carry the same weight as m independent observations. For the retinal thickness example in Ledolter et al.,[14] the intracluster correlation is approximately 0.8.

Result: The required number of observations n (number of clusters, k, times number of observations in each cluster, m) in each treatment group is

$$n = km = 2(z_\alpha + z_\beta)^2(\sigma/\delta)^2\,[1 + (m - 1)\rho]$$

Discussion: The intracluster correlation inflates the sample size that we obtain under complete random sampling, $2(z_\alpha + z_\beta)^2(\sigma/\delta)^2$, by the factor $[1 + (m - 1)\rho]$. For $\rho = 0$, we are back to our earlier Result 2. For $\rho = 1$, we must multiply the sample size that we obtain under complete random sampling by the number of experimental units in the cluster (m). Here experimental units in a cluster are carbon-copies of each other. The m experimental units in a cluster basically count as one unit (and not as m).

In the presence of large intracluster correlation it is important to randomize over many clusters so as to maximize the efficiency of the experiment. Taking more and more replicates within the cluster does not increase the power of the experiment by much, but taking more clusters does. With perfect correlation you may as well leave off one eye from each cluster and save yourself the work collecting measurements on the second eye. For $\rho = 0.80$ and say $n = km = 2(z_\alpha + z_\beta)^2(\sigma/\delta)^2[1 + (m - 1)\rho] = 100$ eyes, you take 50 subjects and analyze both of their eyes. It would be wrong to ignore the intracluster correlation and calculate the number of eyes from $n = km = 100/[1 + (m - 1)\rho] = 56.55$, taking only 28 subjects with their 56 eyes.

Comments

Our review covers parametric tests that assume normality, and the derived sample sizes represent a best-case "scenario." When non-Gaussian distributions are sampled (that cannot be transformed to a Gaussian distribution, such as through a log transformation), nonparametric tests are often used instead because the normal distribution assumption may not be valid for parametric statistical tests. Nonparametric tests (such as the Wilcoxon signed-rank test) are less efficient than their parametric equivalents (two-sample t-test) if the underlying distributions are in fact normal, and being less efficient implies that the sample size must be increased to achieve the same power. For the two-sample comparison, Lehman[15] shows that in most situations the sample sizes derived for the parametric test should be increased by approximately 15%. If one plans to use a nonparametric test, a good rule of thumb adds approximately 15% to the sample size that is required for the parametric test.

Also, experimenters usually anticipate losing experimental units before all data can be recorded and analyzed; patients withdraw form studies and animals die during the course of the experiment. Usually, experimenters know the proportion of units they expect to lose. This proportion needs to be added to the sample size.

This tutorial has focused on the application of power analysis before a study is carried out for understanding whether the analysis being proposed is likely to be meaningful based on the number of samples needed, the variability of the sample, the effect size anticipated, and the level of significance and power desired, while at the same time considering the resources that would be needed. Another application of a power analysis is "post hoc" or after the study has been completed, when meaningful statistical significance has not been obtained. In some publications, it is often stated that a larger sample size would be needed to bolster the hypothesis, if it was not supported by the limited samples collected in their study; however, is this always the case? Applying a power analysis after a "negative result" study may be useful for answering this question.

Conclusions

This review has covered typical sample size calculations that you may encounter when designing your experiments. We have omitted the comparison of proportions, the assessment of correlations, comparisons that involve more than two groups of one factor in the 1-way ANOVA setting, and studies involving two factors in the 2-way ANOVA setting. Guidelines have also been provided for power analyses of non-Gaussian sample distributions (e.g., nonparametric testing). Although technical details can get complicated quickly (for example, power calculations for the correlation coefficient make use of the Fisher's z-transformation to normalize the distribution of the Pearson correlation coefficient; sample size in ANOVA models with more than two groups are usually powered for the maximum difference between means[3]), power analysis software is readily available (see Appendix). The book by Cohen[16] is another source for formulas, tables, and much useful practical discussion.

Acknowledgments

Supported by the Center for the Prevention and Treatment of Visual Loss, Iowa City VA Health Care Center (RR&D C9251-C; RX003002), and an endowment from the Pomerantz Family Chair in Ophthalmology (RK).

Disclosure: **J. Ledolter**, None; **R.H. Kardon**, None

References

1. Box GEP, Hunter JS, Hunter WG. *Statistics for Experimenters: Design, Innovation, and Discovery*. 2nd ed. New York: Wiley & Sons; 2005.

2. Ledolter J, Swersey A. *Testing 1-2-3: Experimental Design with Applications in Marketing and Service Operations*. Stanford, CA: Stanford University Press; 2007.
3. Montgomery D. *Design and Analysis of Experiments*. 8th ed. New York: Wiley & Sons; 2012.
4. Fisher RA. *The Design of Experiments. Edinburgh: Oliver and Boyd, 1935 (various later editions, such as 9th ed)*. New York: Macmillan Publishing Company; 1971.
5. Lenth RV. Java applets for power and sample size [Computer software]. 2006. Available at: http://www.stat.uiowa.edu/∼rlenth/Power. Accessed July 1, 2020.
6. Faul F, Erdfelder E, Lang AG, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods*. 2007;39:175–191.
7. Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav Res Methods*. 2009;41:1149–1160.
8. Ledolter J. Economic field experiments: comments on design efficiency, sample size and statistical power. *J Econ Manag*. 2013;9:271–290.
9. Ledolter J, Kardon RH. Assessing trends in functional and structural characteristics: a survey of statistical methods with an example from ophthalmology. *Trans Vis Sci Tech*. 2018;7(5):34, https://doi.org/10.1167/tvst.7.5.34.
10. Van Belle G, Martin DC. Sample size as a function of coefficient of variation and ratio of means. *Am Stat*. 1993;47:165–167.
11. Brus J, Heng JA, Polanía R. Weber's law: a mechanistic foundation after two centuries. *Trends Cogn Sci*. 2019;23:906–908.
12. Pardo-Vazquez JL, Castiñeiras-de Saa JR, Valente M, et al. The mechanistic foundation of Weber's law. *Nat Neurosci*. 2019;22:1493–1502.
13. Vollrath D, Yasumura D, Benchorin G, et al. Tyro3 modulates Mertk-associated retinal degeneration. *PLoS Genet*. 2015;11:e1005723.
14. Ledolter J, Gramlich OW, Kardon RH. Focus on data: display of data. *Invest Ophthalmol Vis Sci*. 2020;61(6):25, https://doi.org/10.1167/iovs.61.6.25.
15. Lehmann EL. *Nonparametrics: Statistical Methods Based on Ranks*. Revised 1st ed. Upper Saddle River, NJ: Prentice Hall; 1998:76–81.
16. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.

Appendix: Useful Software for Power Analysis and Sample Size Determination

Here we list useful software. This list is by no means exhaustive.

Minitab (State College, PA, USA); https://www.sas.com/

Sample size determination for a wide variety of situations, including: 1-sample mean, 2-sample means, paired mean comparison, 1-sample proportion, 2-sample proportions, 1-sample variance, 2-sample variances, 1-way ANOVA (powered for the maximum difference between means).

The software determines the required sample size, and draws a power curve that shows how, for the given sample size, the power changes as a function of the detectable change.

One can specify the sample size and calculate the power of detecting a certain specified change (effect size), or one can specify the sample size and the power and calculate the effect size that can be detected. All software packages listed here help carry out the tedious calculations that we have done from first principles (see the second example of the power calculations in the one-sample setting).

The R Project for Statistical Computing; https://www.r-project.org/

R Statistical Software provides extensive coverage through

- built-in R Functions in library(stats)
- library(pwr), which implements power analysis procedures as outlined in Cohen[16]
- several other specialized power analysis packages

SAS (Cary, NC, USA); https://www.sas.com/

Extensive coverage through the procedures: PROC POWER and PROC GLIMPOWER

Lenth RV: Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA, USA; http://www.stat.uiowa.edu/∼rlenth/Power

Lenth RV. Java Applets (piface.jar) for Power and Sample Size; http://www.stat.uiowa.edu/∼rlenth/Power

Lenth RV. Some practical guidelines for effective sample size determination. *Am Stat*. 2001;55:187–193.

G*Power: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

G*Power is a freeware software tool for either Mac or PC operating systems used to compute statistical power analyses for a number of statistical applications. G*Power can also be used to compute effect sizes and to graphically display the results of power analyses.