RVLS home > HyperStat Online > **Chapter 11 Power**

Contents

Other Sites

Analysis Tools

Power analysis for numerous designs and statistics
John Pezzullo

Power analysis for ANOVA designs
by Michael Friendly

Statistical considerations for clinical trials and scientific experiments
by David Schoenfeld

Power and sample size page
by Russ Lenth

Text

Statistical power
by William Trochim

Power analysis
by StatSoft

Sample size, power and precision (PDF)
by B. Gerstman

Difference between means: Type I and Type II errors and power
by T. D. V. Swinscow; revised by M. J. Campbell

Two sample size practices that I don't recommend (pdf file)
by Russ Lenth

The ideas behind statistical power
by N. Scott Urquhart

Important Factors in Designing Statistical Power Analysis Studies
by Robin High

Power Analysis for Regression and Stepwise Regression
by David C. Howell

Identification of Misconceptions in Learning Statistical Power with Dynamic Graphics as a Remedial Tool
by Chong Ho Yu and John T. Behrens

Books

Statistical Power Analysis : A Simple and General Model for Traditional and Modern Hypothesis Tests
by Kevin R. Murphy and Brett Myors (Editor)

Statistical Power Analysis for the Behavioral Sciences
by Jacob Cohen

How Many Subjects : Statistical Power and Analysis in Research
by Helena Chmura Kraemer and Sue Thiemann

Power (1 of 3)

Next

Power is the probability of correctly rejecting a false null hypothesis. Power is therefore defined as: 1 - b where b is the Type II error probability. If the power of an experiment is low, then there is a good chance that the experiment will be inconclusive. That is why it is so important to consider power in the design of experiments. There are methods for estimating the power of an experiment before the experiment is conducted. If the power is too low, then the experiment can be redesigned by changing one of the factors that determine power.

Consider a hypothetical experiment designed to test whether rats brought up in an enriched environment can learn mazes faster than rats brought up in the typical laboratory environment (the control condition). Two groups of 12 rats each are tested. Although the experimenter does not know it, the population mean number of trials it takes to learn the maze is 20 for the enriched condition and 32 for the control condition. The null hypothesis that the enriched environment makes no difference is therefore false.

Next

Null hypothesis (1 of 4)

Next

The null hypothesis is an hypothesis about a population parameter. The purpose of hypothesis testing is to test the viability of the null hypothesis in the light of experimental data. Depending on the data, the null hypothesis either will or will not be rejected as a viable possibility.

Consider a researcher interested in whether the time to respond to a tone is affected by the consumption of alcohol. The null hypothesis is that $\mu_1 - \mu_2 = 0$ where μ_1 is the mean time to respond after consuming alcohol and μ_2 is the mean time to respond otherwise. Thus, the null hypothesis concerns the parameter $\mu_1 - \mu_2$ and the null hypothesis is that the parameter equals zero.

The null hypothesis is often the reverse of what the experimenter actually believes; it is put forward to allow the data to contradict it. In the experiment on the effect of alcohol, the experimenter probably expects alcohol to have a harmful effect. If the experimental data show a sufficiently large effect of alcohol, then the null hypothesis that alcohol has no effect can be rejected.

Next

Null hypothesis (2 of 4)

Prev Next

It should be stressed that researchers very frequently put forward a null hypothesis in the hope that they can discredit it. For a second example, consider an educational researcher who designed a new way to teach a particular concept in science, and wanted to test experimentally whether this new method worked better than the existing method. The researcher would design an experiment comparing the two methods. Since the null hypothesis would be that there is no difference between the two methods, the researcher would be hoping to reject the null hypothesis and conclude that the method he or she developed is the better of the two.

The symbol H_0 is used to indicate the null hypothesis. For the example just given, the null hypothesis would be designated by the following symbols:
H_0: $\mu_1 - \mu_2 = 0$
or by
H_0: $\mu_1 = \mu_2$.
The null hypothesis is typically a hypothesis of no difference as in this example where it is the hypothesis of no difference between population means. That is why the word "null" in "null hypothesis" is used -- it is the hypothesis of no difference.

Prev Next

Null hypothesis (3 of 4)

Prev Next

Despite the "null" in "null hypothesis," there are occasions when the parameter is not hypothesized to be 0. For instance, it is possible for the null hypothesis to be that the difference between population means is a particular value. Or, the null hypothesis could be that the mean SAT score in some population is 600. The null hypothesis would then be stated as: H_0: m = 600. Although the null hypotheses discussed so far have all involved the testing of hypotheses about one or more population means, null hypotheses can involve any parameter. An experiment investigating the correlation between job satisfaction and performance on the job would test the null hypothesis that the population correlation (r) is 0. Symbolically, H_0: r = 0.

Some possible null hypotheses are given below:

H_0: m=0

H_0: m=10

H_0: m_1 - m_2 = 0

H_0: p = .5

H_0: p_1 - p_2 = 0

H_0: m_1 = m_2 = m_3

H_0: r_1- r_2= 0

Prev Next

Null hypothesis (4 of 4)

Prev

Next section: Steps in hypothesis testing

When a one-tailed test is conducted, the null hypothesis includes the direction of the effect. A one-tailed test of the differences between means might test the null hypothesis that $m_1 - m_2$ is greater than 0. If $M_1 - M_2$ were much less than 0 then the null hypothesis would be rejected in favor of the alternative hypothesis: $m_1 - m_2 < 0$.

See also: significance test and significance level

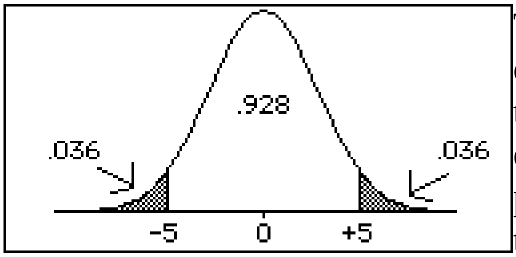
Next section: Steps in hypothesis testing

Prev

One- and two-tailed tests (1 of 4)

Next

In the section on "Steps in hypothesis testing" the fourth step involves calculating the probability that a statistic would differ as much or more from parameter specified in the null hypothesis as does the statistic obtained in the experiment. This statement implies that a difference in either direction would be counted. That is, if the null hypothesis were: H_0: m- m = 0 and the value of the statistic M_1- M_2 were +5, then the probability of M_1- M_2 differing from zero by five or more (in either direction) would be computed. In other words, probability value would be the probability that either M_1- $M_2 \geq 5$ or M_1- $M_2 \leq$ -5. Assume that the figure shown below is the sampling distribution of M_1- M_2.



The figure shows that the probability of a value of +5 or more is 0.036 and that the probability of a value of -5 or less is .036. Therefore the probability of a value either greater than or equal to +5 or less than or equal to -5 is 0.036 + 0.036 = 0.072.

Next

One- and two-tailed tests (2 of 4)

Prev Next

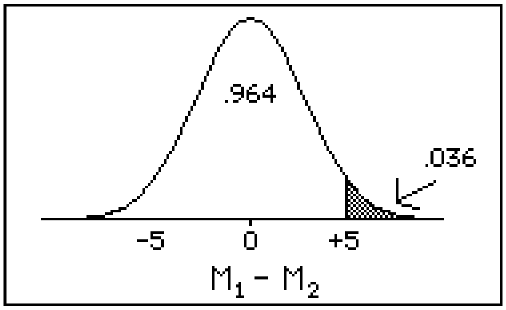
A probability computed considering differences in both directions is called a "two-tailed" probability. The name makes sense since both tails of the sampling distribution are considered. There are situations in which an experimenter is concerned only with differences in one direction. For example, an experimenter may be concerned with whether or not m- m is greater than zero. However, if m- m is not greater than zero, the experimenter may not care whether it equals zero or is less than zero. For instance, if a new drug treatment is developed, the main issue is whether or not it is better than a placebo. If the treatment is not better than a placebo, then it will not be used. It does not really matter whether or not it is worse than the placebo. When only one direction is of concern to an experimenter, then a "one-tailed" test can be performed. If an experimenter were only concerned with whether or not m- m is greater than zero, then the one-tailed test would involve calculating the probability of obtaining a statistic as great or greater than the one obtained in the experiment.

Prev Next

One- and two-tailed tests (3 of 4)

Prev Next

In the example, the one-tailed probability would be the probability of obtaining a value of M_1- M_2 greater than or equal to five given that the difference between population means is zero. The shaded area in the figure is greater than five. The figure shows that the one-tailed probability is .036. It is easier to reject the null hypothesis with a one-tailed than with a two-tailed test as long as the effect is in the specified direction. Therefore, one-tailed tests have lower Type II error rates and more power than do two-tailed tests. In this example, the one-tailed probability (.036) is below the conventional significance level of .05 whereas the two-tailed probability (.072) is not. Probability values for one-tailed tests are always one half the value for two-tailed tests as long as the effect is in the specified direction.



Prev Next

One- and two-tailed tests (4 of 4)

Prev

Next section: Confidence intervals and hypothesis testing

One-tailed and two-tailed tests have the same Type I error rate.

One-tailed tests are sometimes used when the experimenter predicts the direction of the effect in advance. This use of one-tailed tests is questionable because the experimenter can only reject the null hypothesis if the effect is in the predicted direction. If the effect is in the other direction, then the null hypothesis cannot be rejected no matter how strong the effect is. A skeptic might question whether the experimenter would really fail to reject the null hypothesis if the effect were strong enough in the wrong direction. Frequently the most interesting aspect of an effect is that it runs counter to expectations. Therefore, an experimenter who committed him or herself to ignoring effects in one direction may be forced to choose between ignoring a potentially important finding and using the techniques of statistical inference dishonestly. One-tailed tests are not used frequently. Unless otherwise indicated, a test should be assumed to be two-tailed.

Next section: Confidence intervals and hypothesis testing

Prev

Hypothesis testing

Hypothesis testing is a method of inferential statistics. An experimenter starts with a hypothesis about a population parameter called the null hypothesis. Data are then collected and the viability of the null hypothesis is determined in light of the data. If the data are very different from what would be expected under the assumption that the null hypothesis is true, then the null hypothesis is rejected. If the data are not greatly at variance with what would be expected under the assumption that the null hypothesis is true, then the null hypothesis is not rejected. Failure to reject the null hypothesis is not the same thing as accepting the null hypothesis.

More information

Inferential statistics

Next Section: Variables

Inferential statistics are used to draw inferences about a population from a sample. Consider an experiment in which 10 subjects who performed a task after 24 hours of sleep deprivation scored 12 points lower than 10 subjects who performed after a normal night's sleep. Is the difference real or could it be due to chance? How much larger could the real difference be than the 12 points found in the sample? These are the types of questions answered by inferential statistics.

There are two main methods used in inferential statistics: estimation and hypothesis testing. In estimation, the sample is used to estimate a parameter and a confidence interval about the estimate is constructed.

In the most common use of hypothesis testing, a "straw man" null hypothesis is put forward and it is determined whether the data are strong enough to reject it. For the sleep deprivation study, the null hypothesis would be that sleep deprivation has no effect on performance.

Next Section: Variables

Why the null hypothesis is not accepted (1 of 5)

Next

A null hypothesis is not accepted just because it is not rejected. Data not sufficient to show convincingly that a difference between means is not zero do not prove that the difference is zero. Such data may even suggest that the null hypothesis is false but not be strong enough to make a convincing case that the null hypothesis is false. For example, if the probability value were .15, then one would not be ready to present one's case that the null hypothesis is false to the (properly) skeptical scientific community. More convincing data would be needed to do that. However, there would be no basis to conclude that the null hypothesis is true. It may or may not be true, there just is not strong enough evidence to reject it. Not even in cases where there is no evidence that the null hypothesis is false is it valid to conclude the null hypothesis is true. If the null hypothesis is that μ_1 - μ_2 is zero then the hypothesis is that the difference is exactly zero. No experiment can distinguish between the case of no difference between means and an extremely small difference between means. If data are consistent with the null hypothesis, they are also consistent with other similar hypotheses.

Next

Why the null hypothesis is not accepted (2 of 5)

Prev Next

Thus, if the data do not provide a basis for rejecting the null hypothesis that $\mu_1 - \mu_2 = 0$ then they almost certainly will not provide a basis for rejecting the hypothesis that $\mu_1 - \mu_2 = 0.001$. The data are consistent with both hypotheses. When the null hypothesis is not rejected then it is legitimate to conclude that the data are consistent with the null hypothesis. It is not legitimate to conclude that the data support the acceptance of the null hypothesis since the data are consistent with other hypotheses as well. In some respects, rejecting the null hypothesis is comparable to a jury finding a defendant guilty. In both cases, the evidence is convincing beyond a reasonable doubt. Failing to reject the null hypothesis is comparable to a finding of not guilty. The defendant is not declared innocent. There is just not enough evidence to be convincing beyond a reasonable doubt. In the judicial system, a decision has to be made and the defendant is set free. In science, no decision has to be made immediately. More experiments are conducted.

Prev Next

Why the null hypothesis is not accepted (3 of 5)

Prev Next

One experiment might provide data sufficient to reject the null hypothesis, although no experiment can demonstrate that the null hypothesis is true. Where does this leave the researcher who wishes to argue that a variable does not have an effect? If the null hypothesis cannot be accepted, even in principle, then what type of statistical evidence can be used to support the hypothesis that a variable does not have an effect. The answer lies in relaxing the claim a little and arguing not that a variable has no effect whatsoever but that it has, at most, a negligible effect. This can be done by constructing a confidence interval around the parameter value. Consider a researcher interested in the possible effectiveness of a new psychotherapeutic drug. The researcher conducted an experiment comparing a drug-treatment group to a control group and found no significant difference between them. Although the experimenter cannot claim the drug has no effect, he or she can estimate the size of the effect using a confidence interval. If μ_1 were the population mean for the drug group and μ_2 were the population mean for the control group, then the confidence interval would be on the parameter μ_1 - μ_2.

Prev Next

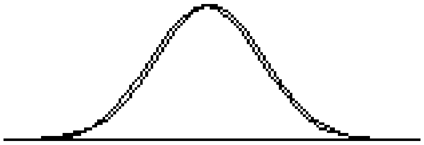
Why the null hypothesis is not accepted (4 of 5)

Prev Next

Assume the experiment measured "well being" on a 50 point scale (with higher scores representing more well being) that has a standard deviation of 10. Further assume the 99% confidence interval computed from the experimental data was:

$-.5 \le \mu_1 - \mu_2 \le 1$

This says that one can be 99% sure that the mean "true" drug treatment effect is somewhere between -.5 and 1. If it were -.5 then the drug would, on average, be slightly detrimental; if it were 1 then the drug would, on average, be slightly beneficial. But, how much benefit is an average improvement of 1? Naturally that is a question that involves characteristics of the measurement scale. But, since 1 is only .10 standard deviations, it can be presumed to be a small effect. The overlap between two distributions whose means differ by .10 standard deviations is shown below. Although one distribution is slightly to the right of the other, the overlap is almost complete.



Prev Next

Why the null hypothesis is not accepted (5 of 5)

Prev

Next section: The precise meaning of the p value

So, the finding that the maximum difference that can be expected (based on a 99% confidence interval) is itself a very small difference would allow the experimenter to conclude that the drug is not effective. The claim would not be that it is totally ineffective, but, at most, its effectiveness is very limited.

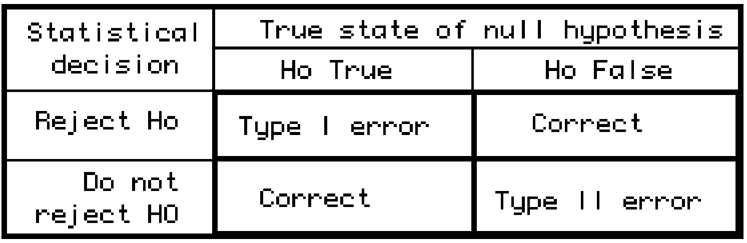
See also: pages 2-3 of "Confidence intervals & hypothesis testing"

Next section: The precise meaning of the p value

Prev

Type I and II errors (1 of 2)

Next



Statistical decision	True state of null hypothesis	
	Ho True	Ho False
Reject Ho	Type I error	Correct
Do not reject HO	Correct	Type II error

There are two kinds of errors that can be made in significance testing: (1) a true null hypothesis can be incorrectly rejected and (2) a false null hypothesis can fail to be rejected. The former error is called a Type I error and the latter error is called a Type II error. These two types of errors are defined in the table. The probability of a Type I error is designated by the Greek letter alpha (α) and is called the Type I error rate; the probability of a Type II error (the Type II error rate) is designated by the Greek letter beta (ß) . A Type II error is only an error in the sense that an opportunity to reject the null hypothesis correctly was lost. It is not an error in the sense that an incorrect conclusion was drawn since no conclusion is drawn when the null hypothesis is not rejected.

Next

Type I and II errors (2 of 2)

Prev

Next section: One- and two-tailed tests

A Type I error, on the other hand, is an error in every sense of the word. A conclusion is drawn that the null hypothesis is false when, in fact, it is true. Therefore, Type I errors are generally considered more serious than Type II errors. The probability of a Type I error (α) is called the significance level and is set by the experimenter. There is a tradeoff between Type I and Type II errors. The more an experimenter protects him or herself against Type I errors by choosing a low level, the greater the chance of a Type II error. Requiring very strong evidence to reject the null hypothesis makes it very unlikely that a true null hypothesis will be rejected. However, it increases the chance that a false null hypothesis will not be rejected, thus lowering power. The Type I error rate is almost always set at .05 or at .01, the latter being more conservative since it requires stronger evidence to reject the null hypothesis at the .01 level then at the .05 level.

Next section: One- and two-tailed tests

Prev

Significance level

In hypothesis testing, the significance level is the criterion used for rejecting the null hypothesis. The significance level is used in hypothesis testing as follows: First, the difference between the results of the experiment and the null hypothesis is determined. Then, assuming the null hypothesis is true, the probability of a difference that large or larger is computed . Finally, this probability is compared to the significance level. If the probability is less than or equal to the significance level, then the null hypothesis is rejected and the outcome is said to be statistically significant. Traditionally, experimenters have used either the .05 level (sometimes called the 5% level) or the .01 level (1% level), although the choice of levels is largely subjective. The lower the significance level, the more the data must diverge from the null hypothesis to be significant. Therefore, the .01 level is more conservative than the .05 level. The Greek letter alpha is sometimes used to indicate the significance level. See also: Type I error and significance test

Significance test (1 of 2)

Next

A significance test is performed to determine if an observed value of a statistic differs enough from a hypothesized value of a parameter to draw the inference that the hypothesized value of the parameter is not the true value. The hypothesized value of the parameter is called the "null hypothesis." A significance test consists of calculating the probability of obtaining a statistic as different or more different from the null hypothesis (given that the null hypothesis is correct) than the statistic obtained in the sample. If this probability is sufficiently low, then the difference between the parameter and the statistic is said to be "statistically significant."

Just how low is sufficiently low? The choice is somewhat arbitrary but by convention levels of .05 and .01 are most commonly used.

For instance, an experimenter may hypothesize that the size of a food reward does not affect the speed a rat runs down an alley. One group of rats receives a large reward and another receives a small reward for running the alley. Suppose the mean running time for the large reward were 1.5 seconds and the mean running time for the small reward were 2.1 seconds.

Next

Significance test (2 of 2)

Prev

The difference between means is thus 2.1 - 1.5 = .6 seconds. The test of whether this difference is significant consists of determining the probability of obtaining a difference as large or larger than .6 seconds given there is really no effect of magnitude of reward. If the probability is low (below the significance level) then the null hypothesis that magnitude of reward makes no difference is rejected in favor of the alternate hypothesis that it does make a difference. The null hypothesis is not accepted just because it is not rejected.

Prev

Estimating power (1 of 4)

Next

The primary purpose of power analysis is to guide in the choice of sample size. First, the experimenter specifies the power that he or she wishes to achieve. Then, the sample size needed for that level of power can be estimated. The calculations for power depend on the size of the effect in the population. Therefore, the first and most difficult step in choosing a sample size is to estimate the size of the effect. If there are published experiments similar to the one to be conducted, then the effects obtained in these published studies can be used as a guide to the size of the effect. There is a need for caution, however, since there is a tendency for published studies to contain overestimates of effect sizes. Often previous studies are not sufficiently similar to a new study to provide a valid basis for estimating the effect size. In this case, it is possible to specify the minimum effect size that is considered important. For example, an experimenter interested in the effectiveness of a course that prepares students for the quantitative portion of the SAT

Next

Estimating power (2 of 4)

Prev Next

might consider a difference of 30 points between the treatment group and the control group to be the minimum effect size worth detecting. Estimating the effect size includes estimating the population variance as well as the population means. In the above example, the experimenter would have to estimate the variance of the SAT (it is about 10000) before the calculations could be done. Frequently it is easiest to specify the effect size in terms of the number of standard deviations separating the population means. Thus, one might find it easier to estimate that the population mean for the experimental group is .5 standard deviations above the population mean for the control group than to estimate the two population means and the population variance. Fortunately, the power for any experiment in which the difference between population means is the same number of population standard deviations apart is the same. The power for an experiment in which $\mu_1 = 10$, $\mu_2 = 20$, and s = 20 is the same as the power for an experiment in which $\mu_1 = 5$, $\mu_2 = 7$, and s = 4. In both cases, the means are 0.5 standard deviations apart.

Prev Next

Estimating power (3 of 4)

Prev Next

How much power is enough power? Naturally, the more power the better. However, in some experiments it is very time consuming and expensive to run each subject. In these experiments, the experimenter usually must accept less power than is typically found in experiments in which subjects can be run cheaply and easily. In any case, power below .25 would almost always be considered too low and power above .80 would be considered satisfactory. Keep in mind that a Type II error is not necessarily so bad since a failure to reject the null hypothesis does not mean that the research hypothesis should be abandoned. If the results are suggestive, further experiments should be conducted to see if the existence of the effect can be confirmed. The power of the two experiments taken together will be greater than the power of either one.

Prev Next

Estimating power (4 of 4)

Prev

Next chapter: Introduction to between-subjects ANOVA

The computation of power when the experimenter is going to use a t test rather than a z test (as used in the section on factors affecting power for computational convenience) is much more complicated. These formulas for power are not given here. Instead, you should use a computer program to compute power such as Russ Lenth's power calculation page.

Next chapter: Introduction to between-subjects ANOVA

Prev

Sample size

The sample size is very simply the size of the sample. If there is only one sample, the letter "N" is used to designate the sample size. If samples are taken from each of "a" populations, then the small letter "n" is used to designate size of the sample from each population. When there are samples from more than one population, N is used to indicate the total number of subjects sampled and is equal to (a)(n). If the sample sizes from the various populations are different, then n_1 would indicate the sample size from the first population, n_2 from the second, etc. The total number of subjects sampled would still be indicated by N.

When correlations are computed, the sample size (N) refers to the number of subjects and thus the number of pairs of scores rather than to the total number of scores.

The symbol N also refers to the number of subjects in the formulas for testing differences between dependent means. Again, it is the number of subjects, not the number of scores.

Why the published literature contains overestimates of effect sizes

A research report containing a significant effect is much more likely to be published than is a research report containing a nonsignificant effect. This publication "bias" means that the published studies are a very select set of the studies actually conducted. Assume two investigators (A and B) are researching topics for which the difference between population means is .75 standard deviations; both experimenters use the same sample size of 15. The power for this experiment (using the .05 significance level) is .50. Assume that Investigator A's experiment produced a significant result but Investigators B's experiment did not. The effect size in Investigater A's sample was somewhat higher than the population effect size, and this result is publishable. Invesitgater B's effect size was somewhat lower than the population effect size, and this result is not publishable. In general, studies which, by chance, find effect sizes larger than the population effect size are much more likely to be published than are studies that, by chance, find effect sizes smaller than the population effect size.

The more power the better

Some researchers have argued that it is dangerous to have too much power. They claim that with too much power, it is easy to find an effect that is statistically significant but not practically significant. For instance, an experiment with 250 subjects per group would be likely to detect a trivial effect. The fallacy in this argument is that competent researchers do not confuse statistical and pracitical significance. Moreover, if the sample size is large then a confidence interval will reveal exactly how small the effect is.

Confidence Interval (1 of 2)

Next

A confidence interval is a range of values that has a specified probability of containing the parameter being estimated. The 95% and 99% confidence intervals which have .95 and .99 probabilities of containing the parameter respectively are most commonly used. If the parameter being estimated were m, the 95% confidence interval might look like the following:

$$12.5 \leq m \leq 30.2$$

What this means is that the interval between 12.5 and 30.2 has a .95 probability of containing m.

A confidence interval only has the specified probability of containing the parameter if the sample data on which it is based is the only information available about the value of the parameter. As an extreme example, consider the case in which 1000 studies estimating the value of m in a certain population all resulted in estimates between 25 and 30. If one more study were conducted and if the 95% confidence interval on m were computed (based on that one study) to be:

$$35 \leq m \leq 45$$

Next

Confidence Interval (2 of 2)

Prev

then it would be absurd to say that the probability that m is between 35 and 45 is .95. It almost certainly is not. However, if the only data you had to go on were that one study, then, from your point of view, the probability is .95.

It is important to be very precise about the sense in which a confidence interval has a specified probability of containing a parameter: If the procedure for computing a 95% confidence interval is used over and over, 95% of the time the interval will contain the parameter.

Confidence intervals can be constructed for any estimated parameter, not just m. For example, one might estimate the proportion of people who could pass a training program or the difference between the mean for subjects taking a drug and those taking a placebo. Click below for details:

Mean , s known
Mean, s estimated
Difference between means, s known
Difference between means, s estimated
Pearson's correlation
Difference between correlations
Proportion
Difference between proportions

Prev

Combining probabilities across studies

It is possible to combine the probabilities obtained from independent studies into one overall probability level. The formula is: Chi Square = -2 Σ Ln(p).

The chi square is based on 2k degrees of freedom where k is the number of probabilities being combined. This test is typically conducted using one-tailed probability values.

Consider a researcher who conducted three experiments comparing an experimental and a control condition. None of the three studies revealed significant differences (one-tailed p's = .08, .21, and .11). Combining the three p's into one,

Chi Square = -2(-2.53 -1.56 - 2.21) = 12.6.

A Table of the Chi Square distribution can be used to find that the probability value for a chi square of 12.6 with 6 df is 0.0498.

Following a nonsignificant finding (1 of 3)

Next

An experimenter wishes to test the hypothesis that sleep deprivation increases reaction time. An experiment is conducted comparing the reaction times of 10 people who have missed a night's sleep with 10 control subjects. Although the sleep-deprived subjects react more slowly, the difference is not significant, p = .10, two tailed. Should the experimenter be more or less certain that sleep deprivation increases reaction time than he or she was before the experiment was conducted. The naive approach is to argue that there was no significant difference between the sleep-deprived and the control group so the experimenter should now be less confident that sleep deprivation increases reaction time. This argument implicitly assumes that the null hypothesis should be accepted when it is not rejected. A more straightforward and more correct approach is to consider that the experimenter expected the sleep-deprived group to have slower reaction time, and they did. The experimenter's prediction was correct. It is just that the difference was not large enough to rule out chance as an explanation.

Next

Following a nonsignificant finding (2 of 3)

Prev Next

The experimenter's belief that sleep deprivation increases reaction time should be strengthened. Nonetheless, the data are not strong enough to convince a skeptic, so no attempt should be made to publish the results. Instead, the experimenter repeated the experiment. Once again, the sleep- deprived group had slower reaction times. Based on the results of the first experiment, the experimenter conducted a one-tailed test in the second experiment. However, it was not significant, p = .08, one tailed. The naive interpretation of the two experiments is that the experimenter tried twice to find a significant result and failed both times. With each failure, the strength of the experimenter's case that sleep deprivation increases reaction time is weakened. The correct interpretation is that in two out of two experiments the sleep-deprived subjects had the slower reaction times. The experimenter's case is strengthened by each experiment. Moreover, there are methods for combining the probability values across experiments. For these two experiments, the combined probability is .047.

Prev Next

Following a nonsignificant finding (3 of 3)

Prev

Next chapter: Testing hypotheses with standard errors

Taken together, these two experiments provide relatively good evidence (p<.05) that sleep deprivation increases reaction time. Naturally, the more times an outcome is replicated, the more believable the outcome. Assume the experimenter did the experiment six times and the sleep-deprived group was slower each time. If the probabiltiy values were: .10, .08, .12, .07, .19, and .13, the combined probability would be 0.009. Therefore, six nonsignificant probabilities combine to produce one highly significant probability. Compare this with the naive view that would state that the experimenter's hypothesis is almost certainly incorrect since not one of the six experiments found a significant difference between sleep-deprived and control subjects. In conclusion, a nonsignificant result means that the data are inconclusive. Collecting additional data may be all that is needed to reject the null hypothesis. If the null hypothesis is true, then additional data will make clear that the effect is at most small. The additional data can never prove that the effect is nonexistent.

Next chapter: Testing hypotheses with standard errors

Prev

Power (2 of 3)

Prev Next

The question is, "What is the probability that the experimenter is going to be able to demonstrate that the null hypothesis is false by rejecting it at the .05 level?" This is the same thing as asking "What is the power of the test?" Before the power of the test can be determined, the standard deviation (s) must be known. If s = 10 then the power of the significance test is 0.80. (Click here to see how to compute power.) This means that there is a 0.80 probability that the experimenter will be able to reject the null hypothesis. Since power = 0.80, b = 1-.80 = .20.

It is important to keep in mind that power is not about whether or not the null hypothesis is true (It is assumed to be false). It is the probability the data gathered in an experiment will be sufficient to reject the null hypothesis. The experimenter does not know that the null hypothesis is false. The experimenter asks the question: If the null hypothesis is false with specified population means and standard deviation, what is the probability that the data from the experiment will be sufficient to reject the null hypothesis?

Prev Next

Power (3 of 3)

Prev

Next section: Factors affecting power: Introduction

If the experimenter discovers that the probability of rejecting the null hypothesis is low (power is low) even if the null hypothesis is false to the degree expected (or hoped for), then it is likely that the experiment should be redesigned. Otherwise, considerable time and expense will go into a project that has a small chance of being conclusive even if the theoretical ideas behind it are correct.

Next section: Factors affecting power: Introduction

Prev

Factors affecting power : Introduction

Next section: Size of difference between means

The factors affecting power will be illustrated in terms of a simple example. Suppose an experimenter were interested in testing whether a drug affects reaction time. Subjects are tested once after taking the drug and once after taking a placebo (naturally, half of the subjects take the drug first and half take the placebo first). The difference in reaction time between the drug and placebo conditions is calculated for each subject. The null hypothesis is that the population mean difference score is zero. H_0: $\mu_{diff} = 0$ where μ_{diff} is the population mean difference score: $\mu_{diff} = \mu_{drug} - \mu_{placebo}$.

To simplify the calculations, assume that the standard deviation of the difference scores (s) is known to the experimenter. The experimenter samples N subjects at random and computes the mean difference score (M). A significance test is conducted using the formula: $z = \frac{M}{\sigma/\sqrt{N}}$.

Subsequent sections will use this example to investigate the factors affecting power.

Next section: Size of difference between means

Factors affecting power: Size of the difference between population means (1 of 3)

Next

The size of the difference between population means is an important factor in determining power. Naturally, the more the means differ from each other, the easier it is to detect the difference. In the example, the difference between means, μ_{diff} , is the population mean difference score. It represents the size of the drug effect. For instance, if there were no difference between the drug and the placebo, then μ_{diff} would be zero and there would be no effect of the drug. If the drug slows people down and, as a result, increases reaction time, μ_{diff} would be a positive number. The larger the effect of the drug, the larger the value of μ_{diff}. Assume that the standard deviation and sample size are: s = 50 and N = 25. The null hypothesis would then be rejected at the .05 level if M were larger than 19.6 or less than -19.6. (Click here for calculations.) The sampling distribution of M for four values of m_{diff} (0, 10, 20, and 30) are shown on the next page. As you will see, the farther the value of m_{diff} is from zero, the smaller the Type II error rate (b) and therefore the larger the power.

Next

Sampling Distribution (1 of 3)

Next

If you compute the mean of a sample of 10 numbers, the value you obtain will not equal the population mean exactly; by chance it will be a little bit higher or a little bit lower. If you sampled sets of 10 numbers over and over again (computing the mean for each set), you would find that some sample means come much closer to the population mean than others. Some would be higher than the population mean and some would be lower. Imagine sampling 10 numbers and computing the mean over and over again, say about 1,000 times, and then constructing a relative frequency distribution of those 1,000 means. This distribution of means is a very good approximation to the sampling distribution of the mean. The sampling distribution of the mean is a theoretical distribution that is approached as the number of samples in the relative frequency distribution increases. With 1,000 samples, the relative frequency distribution is quite close; with 10,000 it is even closer. As the number of samples approaches infinity, the relative frequency distribution approaches the sampling distribution.

Next

Sampling distribution (2 of 3)

Prev Next

The sampling distribution of the mean for a sample size of 10 was just an example; there is a different sampling distribution for other sample sizes. Also, keep in mind that the relative frequency distribution approaches a sampling distribution as the number of samples increases, not as the sample size increases since there is a different sampling distribution for each sample size.

A sampling distribution can also be defined as the relative frequency distribution that would be obtained if all possible samples of a particular sample size were taken. For example, the sampling distribution of the mean for a sample size of 10 would be constructed by computing the mean for each of the possible ways in which 10 scores could be sampled from the population and creating a relative frequency distribution of these means. Although these two definitions may seem different, they are actually the same: Both procedures produce exactly the same sampling distribution.

Prev Next

Sampling distribution (3 of 3)

Prev

Next section: Sampling distribution of the mean

Statistics other than the mean have sampling distributions too. The sampling distribution of the median is the distribution that would result if the median instead of the mean were computed in each sample.

Students often define "sampling distribution" as the sampling distribution of the mean. That is a serious mistake.

Sampling distributions are very important since almost all inferential statistics are based on sampling distributions.

Click here for interactive simulation illustrating important concepts about sampling distributions.

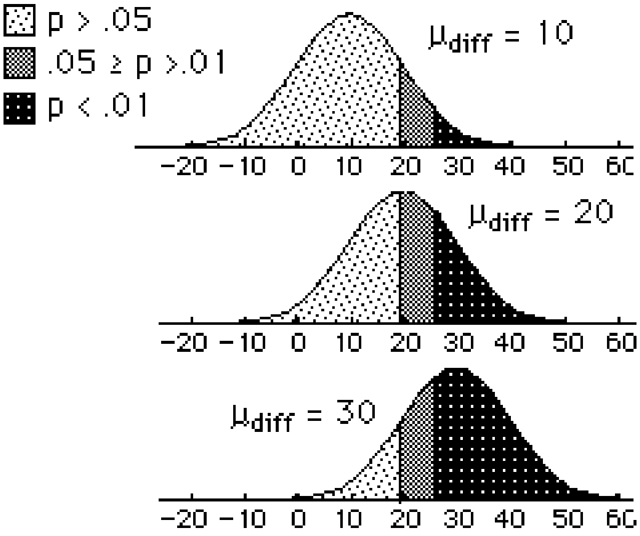
Next section: Sampling distribution of the mean

Prev

Factors affecting power: Significance level (1 of 2)

Next

A second important factor affecting power is the significance level chosen. The more conservative (lower) the significance level, the lower the power. Thus, using the .01 level will result in lower power than using the .05 level. In the example, assuming s = 50 and N = 25, the null hypothesis is rejected at the .05 level if M ≤ -19.6 or M ≥ 19.6 and is rejected at the .01 level if M ≤ -25.8 or M > 25.8. (Click here for calculations.) The figure to the right shows the power for the .05 and .01 levels for three values of μ_{diff}. The most darkly shaded area ■ is the power using the .01 level. The area in gray ▒ is the difference in power between the .05 and .01 levels.



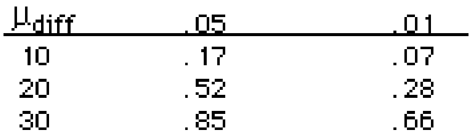
Next

Factors affecting power: Significance level (2 of 2)

Prev

Next section: Sample size

The power for the three values of m_{diff} at the two significance levels is shown below.



μ_{diff}	.05	.01
10	. 17	.07
20	.52	.28
30	.85	.66

It stands to reason that the power would be higher for the .05 level than for the .01 level. The cost of stronger protection against Type I errors is more Type II errors.

Next section: Sample size

Prev

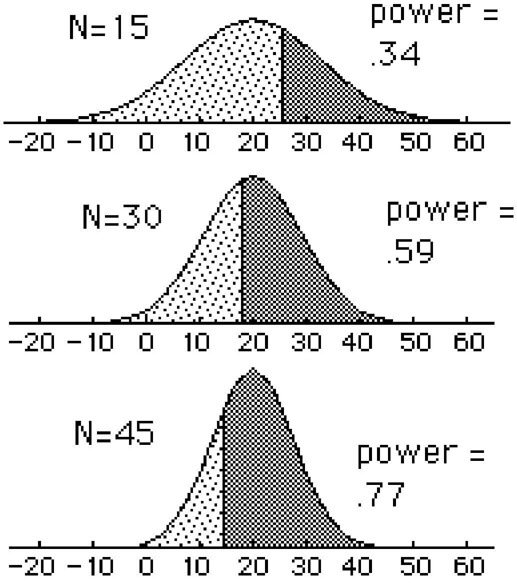
Factors affecting power: Sample size (1 of 2)

Next

Increasing N, the sample size, decreases the denominator of the equation for z: $z = \frac{M}{\sigma/\sqrt{N}}$ and therefore increases power.

The graphs on the right show the sampling distribution of M (for the example) when s = 50 and μ_{diff} = 20. The cutoff points for being significant at the .05 level were determined using the formulas: $M = (-1.96)(\sigma/\sqrt{N})$ and $M = (+1.96)(\sigma/\sqrt{N})$.

Note how the cutoff point and the variance of the distribution change as N increases. The mean of the distribution does not change.



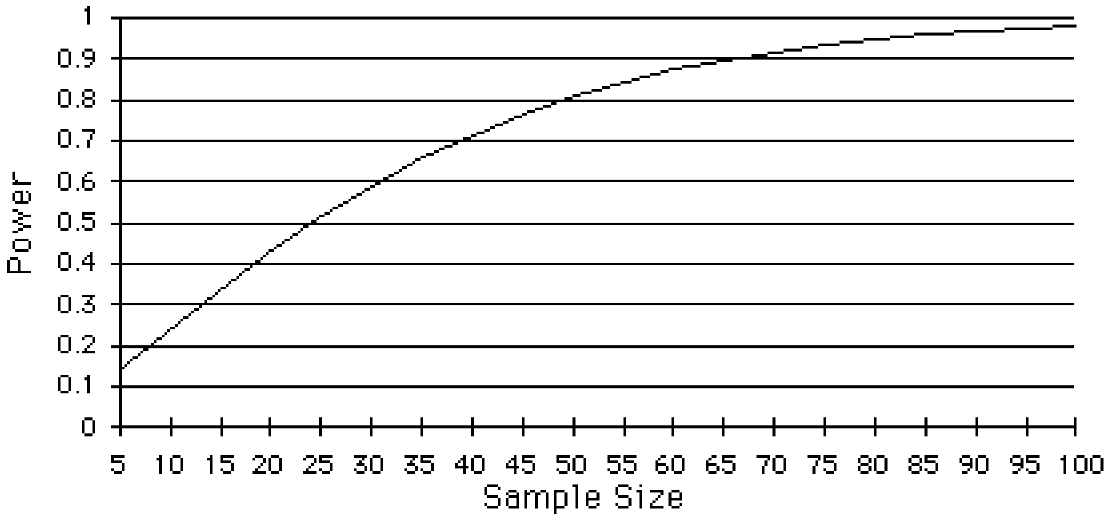
Next

Factors affecting power: Sample size (2 of 2)

Prev

Next section: Variance

Naturally, power increases as the sample size increases. The effect of sample size on power for this example is seen in more detail in the following graph. Choosing a sample size is a difficult decision. There is a tradeoff between the gain in power and the time and cost of testing a large number of subjects.

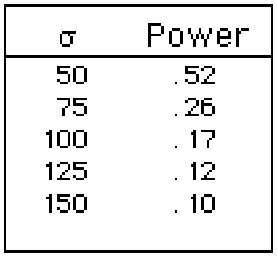


Next section: Variance

Prev

Factors affecting power: Variance (1 of 2)

Next



σ	Power
50	.52
75	.26
100	.17
125	.12
150	.10

The larger the variance (s^2), the lower the power. In the formula for z:

$$z = \frac{M}{\sigma/\sqrt{N}}$$

increasing s^2 increases the denominator and therefore lowers z and power. For the example, s is the standard deviation of the difference scores. The power of the test using the .05 significance level, for N = 25, μ_{diff} = 20, and various values of s is shown in the table on the right side of this page. The table shows that the power decreases as s increases. There are ways that an experimenter can reduce variance to increase power. One is to define a relatively homogeneous population. For instance, if one were studying reading speed, one could begin by studying the population of people in their first year at a selective college rather than the population of all English-speaking adults. The variance would be much reduced.

Next

Factors affecting power: Variance (2 of 2)

Prev

Next section: Other factors

The cost would be that the results would not be as generalizeable. A second way to reduce variance is by using a within-subjects design. In these designs, the overall level of performance of each person is subtracted out. This usually reduces the variance substantially

Next section: Other factors

Prev

Variance and standard deviation (1 of 2)

Next

The variance is a measure of how spread out a distribution is. It is computed as the average squared deviation of each number from its mean. For example, for the numbers 1, 2, and 3, the mean is 2 and the variance is:

$$s^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = .667 \, .$$

The formula (in summation notation) for the variance in a population is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where m is the mean and N is the number of scores.

When the variance is computed in a sample, the statistic

$$S^2 = \frac{\Sigma(X - M)^2}{N}$$

(where M is the mean of the sample) can be used. S^2 is a biased estimate of s^2, however. By far the most common formula for computing variance in a sample is:

$$s^2 = \frac{\Sigma(X - M)^2}{N-1}$$

which gives an unbiased estimate of s^2. Since samples are usually used to estimate parameters, s^2 is the most commonly used measure of variance.

Next

Variance and standard deviation (2 of 2)

Prev

Next section: Summary

Standard Deviation
The formula for the standard deviation is very simple: it is the square root of the variance. It is the most commonly used measure of spread.

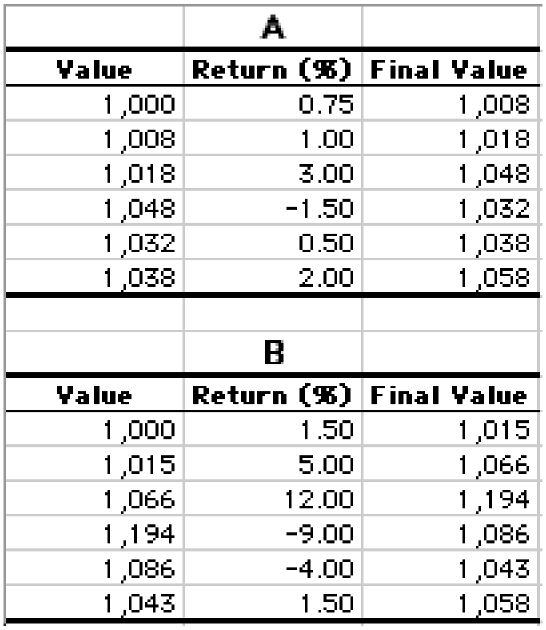
An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score. In a normal distribution, about 68% of the scores are within one standard deviation of the mean and about 95% of the scores are within two standards deviations of the mean.

The standard deviation has proven to be an extremely useful measure of spread in part because it is mathematically tractable. Many formulas in inferential statistics use the standard deviation.

Although less sensitive to extreme scores than the range, the standard deviation is more sensitive than the semi-interquartile range. Thus, the standard deviation should be supplemented by the semi-interquartile range when the possibility of extreme scores is present.

If variable Y is a linear transformation of X such that: Y = bX + A, then the variance of Y is: $b^2 \sigma^2_x$ where σ^2_x is the variance of X. The standard deviation of Y is b s_X where s_X is the standard deviation of X.

Standard Deviation as a Measure of Risk
The standard deviation is often used by investors to measure the risk of a stock or a stock portfolio. The basic idea is that the standard deviation is a measure of volatility: the more a stock's returns vary from the stock's average return, the more volatile the stock. Consider the following two stock portfolios and their respective returns (in per cent) over the last six months. Both portfolios end up increasing in value from $1,000 to $1,058. However, they clearly differ in volatility. Portfolio A's monthly returns range from -1.5% to 3% whereas Portfolio B's range from -9% to 12%. The standard deviation of the returns is a better measure of volatility than the range because it takes all the values into account. The standard deviation of the six returns for Portfolio A is 1.52; for Portfolio B it is 7.24.



	A	
Value	Return (%)	Final Value
1,000	0.75	1,008
1,008	1.00	1,018
1,018	3.00	1,048
1,048	-1.50	1,032
1,032	0.50	1,038
1,038	2.00	1,058

	B	
Value	Return (%)	Final Value
1,000	1.50	1,015
1,015	5.00	1,066
1,066	12.00	1,194
1,194	-9.00	1,086
1,086	-4.00	1,043
1,043	1.50	1,058

Further Reading:
Risk Management by Michel Crouhy et al.

The Intelligent Asset Allocator: How to Build Your Portfolio to Maximize Returns and Minimize Risk by William J. Bernstein

Personal Finance for Dummies by Eric Tyson

Next section: Summary

Prev

Other factors affecting power

Next section: Estimating power

Several other factors affect power. One factor is whether the population distribution is normal. Deviations from the assumption of normality usually lower power. A second factor affecting power is the type of statistical procedure used. Some of the distribution-free tests are less powerful than other tests when the distribution is normal distributions but more powerful when the distribution is highly-skewed. One-tailed tests are more powerful than two-tailed tests as long as the effect is in the expected direction. Otherwise, their power is zero.

The choice of an experimental design can have a profound effect on power. Within-subject designs are usual much more powerful than between-subject designs.

Next section: Estimating power