

Designing Your Experiment: How Large The Sample Do You Need?

Introduction

In planning and designing a statistical analysis one of the key issues is the sample size necessary to produce statistically sound inference. As many paradigms in statistics, the sample size determination is a trade-off problem, the designer would like to maximize sample size, but there are obvious concerns about economy, logistic, ethics, plausibility, and other issues of desired extensive sampling. Sometimes the retrospective power or observed power can be found by a so called post-hoc analysis to see “if the test was powerful enough or not.” This in principle should never be done. Using power to determine a sample size is PROSPECTIVE in nature.

The sample size depends on the statistical model, statistical procedure, significance/confidence level adopted, and on precision/power desired. For Bayesian statisticians the sample size depends on posterior measures of precisions.

In this terse guide, we describe how to select a sample size when (i) constructing a confidence interval for the parameter of interest, (ii) in testing for the means in one and two sample problems, and (iii) in designing a more complex experiment such as factorial or block design. The approach could be rather theoretical involving noncentrality parameters of distribution statistics under the alternatives, or practical involving the preset effect size, championed by Cohen (1988), or both.

Since this handout is aimed to readers who are not experts in statistics, we will use the effect size approach combined with software implementations. An effect size helps to determine whether a statistically significant difference is a difference of practical concern. Often, the effect size is a deviation from null hypothesis in units of population standard deviation that we would like to detect. For example, in testing that the means of two populations μ_1 and μ_2 are the same, we would like that chosen test discovers deviation $|\mu_1 - \mu_2| \geq \sigma$, 95% of the time. The parameter σ is common standard deviation for the two populations.

Some call the effect size approach the *T-shirt effect sizes* since the effects come in three sizes: “small”, “medium”, and “large” and sometimes even as X-large. For example, if one chooses standard definitions of significance ($\alpha = 0.05$), power ($1 - \beta = 80\%$), and an effect size, then there is no need for any pilot experiments. If T-shirt definitions fit all your studies, then all studies need a sample size of 26 in each group to detect a large effect, 65 in each group to detect a medium effect, 400 in each group to detect a small effect. Choosing standard effect sizes is really the same as picking standard sample sizes - robbing Peter to pay Paul. For more criticism of canned effect sizes see Lenth (2001).

Clearly more precise analyses are possible and involve considerations concerning sample/population variability, variability among subjects, correlations among the regression predictors, and other modeling and practical issues. These methods are beyond the scope of this overview and interested reader can consult extensive statistical literature on the subject.

Confidence Intervals

The only power analysis standardly taught in introductory statistics courses involves confidence intervals for unknown normal location parameter, μ .

When observations X_1, X_2, \dots, X_n are coming from a normal distribution with unknown mean μ and known variance σ^2 , the $(1 - \alpha) \times 100\%$ confidence interval for the population mean is

$$\left[\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right],$$

where

- α is a small number (for 95% confidence, α is 0.05.),
- $z_{1-\alpha/2}$ is $1 - \alpha/2$ quantile of standard normal distribution, i.e., the point that divides the unit area below the standard normal density to $(1 - \alpha/2) \times 100\%$ to the left and $\alpha/2 \times 100\%$ to the right, as in Figure for $z_{0.995}$.
- n is the sample size
- \bar{X} is the sample mean $\bar{X} = \frac{X_1+X_2+\dots+X_n}{n}$, and
- σ is population standard deviation, a square root of variance.

Figure 1: Standard normal density and 0.995 quantile $z_{0.995}$ (Figure notation is $z_{0.005}$ since it uses upper tail area, not quantile notation). It is used for finding 99% confidence interval. Some important normal quantiles are $z_{0.9} = 1.28$, $z_{0.95} = 1.645$, $z_{0.975} = 1.96$, $z_{0.99} = 2.33$ and $z_{0.995} = 2.58$. Other standard normal quantiles are provided in appendix.

If the variance is not known, then the $(1 - \alpha)$ 100% confidence interval is

$$\left[\bar{X} - t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}} \right],$$

where $s = \sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}$ is the sample standard deviation and $t_{n-1, 1-\alpha/2}$ is $(1 - \alpha/2)$ quantile of Student's t distribution with $n - 1$ degrees of freedom.

For the population proportion p the approximate $(1 - \alpha)$ 100% confidence interval is

$$\left[\hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right],$$

where \hat{p} is observed proportion in a sample of size n .

The equations for sample size are derived by constraining the length w of a confidence interval.

(i) Equation for sample size for estimating the mean: σ^2 known.

$$n \geq \frac{4z_{1-\alpha/2}^2 \sigma^2}{w^2}$$

where w is the width of the interval ($w = \text{Upper bound} - \text{Lower bound}$). This follows from the fact that the width of the interval is $2z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$. If σ is not known, it is replaced by s ,

$$n \geq \frac{4t_{n-1, 1-\alpha/2}^2 s^2}{w^2} \approx \frac{4z_{1-\alpha/2}^2 s^2}{w^2},$$

as long as n exceeds 30.

(ii) Equation for sample size for estimating the proportion

$$n \geq \frac{4z_{1-\alpha/2}^2 \hat{p}(1-\hat{p})}{w^2}.$$

Designing the sample size usually precedes the sampling. In absence of data, \hat{p} is our best guess for the population proportion. In absence of any prior information about p , the most conservative choice for \hat{p} is 0.5.

Cholesterol Level. Suppose you are designing a cholesterol study experiment and would like to estimate the mean cholesterol level of all Tech’s students. You will take a random sample of n students and measure their cholesterol level. Previous studies have shown that the standard deviation is 25 and you will use this value in planning your study.

If you wanted a 99% confidence interval with a total length not exceeding 12, how many students should you include in your sample?

$$[n \geq \frac{4 \cdot 2.58^2 \cdot 25^2}{12^2} = 115.5625]$$

Right to Die. The Gallup Poll estimated the support among Americans for “right to die” laws. For the survey, 1528 adults were asked whether they favor voluntary withholding of life-support systems from the terminally ill. The results: 1238 Said yes.

- (a) Find a 99% confidence interval for the percentage of all adult Americans who are in favor of “right to die” laws.
- (b) If the margin error is to be smaller than 0.02, what sample size is needed to achieve that requirement. Assume $\hat{p} = 0.8$.

[The margin of error is usually assumed as $w/2$, thus $n \geq \frac{4 \cdot 2.5758^2 \cdot 0.8 \cdot 0.2}{0.04^2} = 2653.9$].

Sample Sizes in Testing Hypotheses

The very first task in testing problems is a formulation of hypotheses. Clearly, there will be at least (and in most of the cases exactly) two competing hypotheses. The hypothesis which is believed to be true state of nature is standardly denoted by H_0 (null hypothesis) and the competing (alternative, research) hypothesis is denoted by H_1

It is important to specify which of the two hypothesis is H_0 since the subsequent testing procedure depends on this assignment.

Rule: When our goal is to establish an assertion with substantive support obtained from the sample, the negation of the assertion is taken to be the *null* hypothesis H_0 , and the assertion itself is taken to be the research hypothesis H_1 .

If the null hypothesis is rejected when in fact it is true, then the *error of I kind* is committed. If, on the other hand, a wrong null hypothesis is not rejected, then the *error of II kind* is committed (Table 1).

	decide H_0	decide H_1
true H_0	OK	error of I kind
true H_1	error of II kind	OK

It is customary to denote by α the probability of error of the first kind and by β the probability of error of second kind.

A good testing procedure keeps these two probabilities minimal. However, minimizing both errors simultaneously, for the fixed sample size, is often impossible. Usually by minimizing α , the probability of the error of second kind, β , increases.

Sometimes there is no clear dichotomy in the hypotheses such as ‘established truth’ vs ‘research hypotheses’ and both hypotheses look like research hypotheses. For instance, the hypotheses “The new drug is safe” and “The new drug is not safe” are both research hypotheses. In such cases one chooses H_0 in such a way that the error of first kind is more serious than the error of second kind.

If “The new drug is not safe” is chosen as H_0 then the error of first kind is (reject true H_0 , i.e. use unsafe drug) is more serious (at least for the patient) than the error of second (accept false H_0 , i.e. do not use a safe drug).

This is the reason why α is a fixed small number. We want to control the probability of a more serious mistake, the type I error. The second benefit of fixing α at a few values have been brief statistical tables. Standardly used value for α is 5%, with 1% or 10% used less frequently. ¹

¹This choice starts with Fisher (1926, page 504): “... it is convenient to draw the line at about the level at which we can say: *Either there is something*

Power of the α -level test for $H_0 : \mu = \mu_0$, vs. $H_1 : \mu = \mu_1$.

- One sided test:

$$1 - \beta = \Phi \left(\frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}} - z_{1-\alpha} \right);$$

- Two sided test:

$$1 - \beta \approx \Phi \left(\frac{|\mu_0 - \mu_1|}{\sigma/\sqrt{n}} - z_{1-\alpha/2} \right);$$

Often we are in a situation to decide about sample size n (design the experiment). For example, it may be up to us to decide how many respondents to interview in a pool. We can design a sample size in interval estimation to achieve the preassigned precision and confidence level.

As a framework of designing a sample size in the testing setup, consider a problem of testing $H_0 : \mu = \mu_0$ using statistic \bar{X} based on a sample of size n . Let the alternative has fixed value for μ , i.e. $H_1 : \mu = \mu_1 (\neq \mu_0)$. Assume $\alpha = 0.05$. How large n should be so that the power $1 - \beta$ is 0.90? The power of test is a probability of rejecting wrong H_0 , and in our case rejecting $H_0 : \mu = \mu_0$ when $H_1 : \mu = \mu_1$ is true. Naturally, we require that the test has good power, often 0.8, 0.9, 0.95, or similar. Achieving desired power, given α and H_1 fixed requires manipulating the sample size, and this link provides the basis for the sample size design.

The sample size for fixed α, β ,

$$n \geq \frac{1}{d^2} (z_{1-\alpha} + z_{1-\beta})^2,$$

where $d = |\mu_1 - \mu_0|/\sigma$. If the alternative is two sided $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$. In that case the sample size is approximate.

If σ^2 is not known, substitution of an estimate will give an approximate sample size.

Since σ^2 is usually unknown (prior to experiment) the sample size is often determined by expressing the desired precision in units of σ .

An example is Cohen's ² standardized effect size $d = \frac{|\mu_1 - \mu_0|}{\sigma}$. Cohen recommends $d = 0.25$ (small effects), $d = 0.5$ (medium effects), and $d = 0.8$ (large effects). Informally, Cohen's standardized effects are called *T-shirt* effects.

Serum Sodium Levels. ³ These data give the results of analysis of 20 samples of serum measured for their sodium content. The average value for the method of analysis used is 140 ppm. Use $\alpha = 0.05$.

140	143	141	137	132	157	143	149	118	145
138	144	144	139	133	159	141	124	145	139

in the treatment, or a coincidence has occurred such as does not occur more than once in twenty trials. If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."

²Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd edition). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155-159.

³Source: National Quality Control Scheme, Queen Elizabeth Hospital, Birmingham, referenced in *Data* by D. F. Andrews and A. M. Herzberg, Springer, 1985.

(i) Is there evidence that the mean level of sodium in this serum is different from 140 ppm? [Note: The wording: *different from*, not *larger than* suggests the alternative].

[$t = \frac{145.55-140}{9.455/\sqrt{20}} = 2.625$. The alternative is two-sided. Critical value is $t_{20-1,1-0.025} = 2.086$ thus reject H_0 .]

(ii) What sample size is needed to achieve the 90%-power against the alternative $H_1 : \mu = 145$? Use $\alpha = 0.05$.

Since σ is not known, from the preliminary study we find it is approximately 10. This corresponds to the standardized effect $d = (145 - 140)/10 = 0.5$ which is a *medium* effect, according to Cohen's scale. If we did not have any information on σ and no data, we would use d directly.

The formula involving $z_{1-\alpha/2}$ and $z_{1-\beta}$ is giving an approximate value, more accurate formula should use $t_{n-1,1-\alpha/2}$ and $t_{n-1,1-\beta}$, but in this case the exact solution for n is hard to obtain (the algorithm has to be iterative). The approximation with z 's in place of t 's is good if the resulting n is not very small, say not less than 15.

$$n \approx \frac{1}{d^2} (z_{1-\alpha} + z_{1-\beta})^2 = \frac{1}{0.5^2} (1.96 + 1.28155)^2 = 42.0306$$

Thus, take $n = 43$.

The MINITAB⁴ output for this sample size is $n = 44$

```
Testing mean = null (versus not = null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 10
```

Difference	Sample Size	Target Power	Actual Power
5	44	0.9	0.900031

In α -level testing the binomial proportion,

$$H_0 : p = p_0, \text{ v.s. } H_1 : p \neq p_0,$$

the power of the test against specific alternative $p = p_1$ is

$$1 - \beta = \Phi \left[\sqrt{\frac{p_0 q_0}{p_1 q_1}} \left(\frac{\sqrt{n} |p_0 - p_1|}{\sqrt{p_0 q_0}} - z_{1-\alpha/2} \right) \right] = \Phi \left[\sqrt{\frac{p_0 q_0}{p_1 q_1}} \left(z_{\alpha/2} + \frac{\sqrt{n} |p_0 - p_1|}{\sqrt{p_0 q_0}} \right) \right].$$

If the alternative is one sided, then $z_{1-\alpha/2}$ is replaced by $z_{1-\alpha}$. The sample size needed to find the effect $|p_0 - p_1|$ significant $(1 - \beta) \times 100\%$ of the time (test would have power $1 - \beta$) is

$$n = \frac{p_0 q_0 \left(z_{1-\alpha/2} + z_{1-\beta} \sqrt{\frac{p_1 q_1}{p_0 q_0}} \right)^2}{(p_0 - p_1)^2}.$$

Is the cloning of humans moral? Recent Gallup Poll conducted May 2-4, 2004, estimates that 88% Americans believe that cloning humans is morally unacceptable. Results are based on telephone interviews with a randomly selected national sample of $n = 1000$ adults, aged 18 and older.

(i) Test the hypothesis that the true proportion is 0.9, versus the two sided alternative, based on the Gallup's data. Use $\alpha = 0.05$.

⁴Minitab is a computer program designed to perform basic and some advanced statistical tasks. It combines the user-friendliness of Microsoft Excel with the ability to perform complex statistical analyses. It is available on BME public cluster machines.

```

>> clear all
>> n=1000; phat = 0.88;
>> p0=0.9; q0=1-p0;
>> Z = (phat - p0)/sqrt(p0 * q0/n)

Z = -2.1082
>> crit = norminv(0.975)

crit = 1.9600
% The rejection region is (-infinity, -1.96) U (1.96, infinity) and Z=-2.1082
% falls in. Reject H_0

>> pval = 2*normcdf(-2.1028)

pval = 0.0355
% pval < 0.05=alpha => Reject H_0

```

(ii) What is the power of this test against the specific alternative $p = 0.85$.

Recall that the power is $1 - \beta = \Phi \left[\sqrt{\frac{p_0 q_0}{p_1 q_1}} \left(z_{\alpha/2} + \frac{|p_0 - p_1| \sqrt{n}}{\sqrt{p_0 q_0}} \right) \right]$.

```

>> normcdf( sqrt( 0.9*0.1/(0.85*0.15)) * (norminv(0.025) +...
abs(0.9-0.85)*sqrt(1000)/sqrt(0.9*0.1)) )

ans = 0.9973

```

(iii) Suppose that the difference $|p_1 - p_0|$ exceeding 0.03 is to be found significant at least 80% of the time. What sample size ensures this. In our problem the required precision is achieved for $p_1 = 0.87$ or $p_0 = 0.93$ since the alternative is two sided.

Exact power of t-test.

The formulas for power of t test that involve Φ and normal quantiles are only approximations. These approximations are good for larger sample sizes, but for small sample sizes the exact formulas are available. The reason why approximations are used at the first place is more historic, exact formulas need calculation of non-central t that was not readily available – in contrast to standard normal tables.

Assume that we test $H_0 : \mu = 10$ vs $H_1 : \mu > 10$ at level $\alpha = 0.05$. Sample of size $n = 20$ gave $\bar{X} = 12$ and $s = 5$. What is the (EXACT) power of the test against alternative $H_1 : \mu = 13$.

Well, the exact power is $Pr(t \in RR | t \sim nct(df = n - 1, ncp = (\mu_1 - \mu_0)\sqrt{n}/\sigma))$, since under H_1 , t has noncentral t -distribution with $n - 1$ degrees of freedom and noncentrality parameter $\frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}$.

```

>> n=20; Xb=12; mu0 = 10; s=5; mu1= 13;
>> alpha=0.05;
>> % normal approximation
>> pow = 1-normcdf( norminv(1-alpha) - (mu1-mu0)*sqrt(n)/s )

pow = 0.8505

>> % exact power

```

```
>> pow1 = 1- nctcdf( tinvc(1-alpha, n-1), n-1, (mu1-mu0)*sqrt(n)/s)
pow1 = 0.8266
```

When the noncentral t is not available, a good approximation for power is

$$1 - \Phi \left(\frac{t_{n-1,\alpha} - |\mu_0 - \mu_1| \sqrt{n}/s}{\sqrt{1 + \frac{t_{n-1,1-\alpha}^2}{2(n-1)}}} \right).$$

```
>> %approximation with correction
>> 1-normcdf((tinvc(1-alpha,n-1)-
(mu1-mu0)/s * sqrt(n))/sqrt(1 + (tinvc(1-alpha,n-1))^2/(2*n-2)))
ans = 0.8209
```

Two Sample Problem

The group sample size for fixed α, β , in testing equality of means in two normal populations using independent samples, $H_0 : \mu_1 = \mu_2$, versus one sided alternative, is

$$n \geq \frac{2}{d^2} (z_{1-\alpha} + z_{1-\beta})^2,$$

where $d = |\mu_1 - \mu_2|/\sigma$, and σ is common population standard deviation. If the alternative is two sided $z_{1-\alpha}$ is replaced by $z_{1-\alpha/2}$. In that case the sample size is approximate. It is assumed that the group sample sizes are equal, i.e., that the total sample size is $N = 2n$. If the variances are not the same,

$$n \geq \frac{\sigma_1^2 + \sigma_2^2}{|\mu_0 - \mu_2|^2} (z_{1-\alpha} + z_{1-\beta})^2,$$

Mushrooms. Making spore prints is an enormous help in identifying genera and species of mushrooms. To make a spore print, mushroom fans take a fresh, mature cap and lay it on a clean piece of glass. Left overnight or possibly longer the cap should give you a good print. Family of Amanitas is one that has the most poisonous (*Amanita Phalloides*, *Amanita verna*, *Amanita virosa*, *Amanita Pantherina*, etc.), and the most delicious species (*Amanita Cesarea*, *Amanita Rubescens*). Two independent samples of $m = 12$ and $n = 15$ spores of *A. Pantherina* ("Panther") and *A. Rubescens* ("Blusher"), respectively, are analyzed. In both species of mushrooms the spores are smooth and elliptical and the largest possible measurement is taken (great axis of the ellipse). It was found that the means are $\bar{X}_1 = 6.3$ microns and $\bar{X}_2 = 7.5$ microns with standard deviations of $s_1 = 2.12$ microns and $s_2 = 1.94$ microns.

(i) The researcher is interested in testing the hypothesis that population mean sizes of spores for these two mushrooms, μ_1 and μ_2 are the same, against the two sided alternative. Use $\alpha = 0.05$.

Sol. $s_p^2 = (s_1^2(m-1) + s_2^2(n-1))/(m+n-2) = (2.12^2 * 11 + 1.94^2 * 14)/(11+14) = 4.0852$ is pooled sample variance, and $s_p = 2.0212$ is pooled sample standard deviation.

Also, $\sqrt{1/12 + 1/15} = 0.3873$.

The test statistics $t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{1/n + 1/m}} = -1.2/2.02/0.3873 = -1.5338$. In the two sided test, the critical value at $\alpha = 0.05$ is $t_{m+n-2, 1-\alpha/2} = t_{25, 0.975} = 2.060$, and the hypothesis H_0 is not rejected, since $|t| < 2.060$.

(ii) What sample sizes are needed so that researcher should be able to reject the null hypothesis of no effect (no difference between the population means) with power $1 - \beta = 0.9$ against the “medium” standardized effect size of $d = 0.5$. Significance α is as in (i), 0.05.

Sol. $z_{0.90} = 1.2816$ and $z_{0.975} = 1.96$ and the group sample size should be $2/0.5^2 \cdot (1.96 + 1.2816)^2 = 84.0638 \leq 85$. To achieve desired power and detect the deviation of $d = 0.5$, independent samples of $m = 85$ and $n = 85$ spores of *A. Pantherina* (“Panther”) and *A. Rubescens* (“Blusher”) should be taken.

Note that when the alternative is two sided, as in this case, the formula $2/0.5^2 \cdot (1.96 + 1.2816)^2$ is only an approximation and more precise sample sizes (calculated numerically by G*Power and MINITAB) are 86 each, rather than 85.

2-Sample t Test

Testing mean 1 = mean 2 (versus not =)
Alpha = 0.05 Assumed standard deviation = 1

Difference	Sample Size	Target Power	Actual Power
0.5	86	0.9	0.903230

The sample size is for each group.

More complex two sample designs

Suppose that we are interested in testing the equality of normal population means when the underlying variances in the two populations are σ_1^2 and σ_2^2 , and not necessarily equal. Also assume that desired proportion of sample sizes to be determined is $k = n_2/n_1$, that is $n_2 = k \times n_1$. This proportion may be dictated by cost of sampling or by abundance of the populations. When equal group samples are desired, $k = 1$, of course. Then,

$$n_1 = \frac{(\sigma_1^2 + \sigma_2^2/k)(z_{1-\alpha/2} + z_{1-\beta})^2}{|\mu_1 - \mu_2|^2}, \quad n_2 = k \times n_1. \quad (1)$$

As before, μ_1, μ_2, σ_1^2 , and σ_2^2 are unknown, and in absence of any data, one can express $|\mu_1 - \mu_2|^2$ in units of $\sigma_1^2 + \sigma_2^2/k$, to elicit d^2 .

However, if a preliminary samples are available, then μ_1, μ_2, σ_1^2 , and σ_2^2 can be estimated by $\bar{X}_1, \bar{X}_2, s_1^2$ and s_2^2 , respectively, and plugged in the formula for sample size (2).

Mushrooms continued. Suppose that two independent samples of $m = 12$ and $n = 15$ spores of *A. Pantherina* (“Panther”) and *A. Rubescens* (“Blusher”), respectively, are only a pilot study. It was found that the means are $\bar{X}_1 = 6.3$ microns and $\bar{X}_2 = 7.5$ microns with standard deviations of $s_1 = 2.12$ microns and $s_2 = 1.94$ microns. Suppose that Blushers are twice more common than Panthers.

Determine the sample sizes for future the study that will find the difference obtained in the preliminary samples significant at the level $\alpha = 0.05$ with the power of $1 - \beta = 0.90$.

Solution: Here, based on abundance of mushrooms, $2n_1 = n_2$, and $k = 2$. Substituting $\bar{X}_1, \bar{X}_2, s_1^2$ and s_2^2 into (2), one gets,

$$n_1 = \frac{(2.12^2 + 1.94^2/2)(z_{0.975} + z_{0.9})^2}{|6.3 - 7.5|^2} = 46.5260 \approx 47.$$

Here, the effect size was $|6.3 - 7.5|/\sqrt{2.12^2 + 1.94^2/2} = 0.4752$, which is corresponds to $d = 0.4752\sqrt{2}$.

The above “plug-in” principle is controversial. Proponents argue that in the absence of any information on μ_1, μ_2, σ_1^2 , and σ_2^2 the most “natural” procedure is to use their Maximum Likelihood estimators, $\bar{X}_1, \bar{X}_2, s_1^2$ and s_2^2 . Opponents say that one is looking for a sample size that will find the pilot difference significant at the level α with the preassigned power. They argue that, due to routinely small sample sizes in the pilot studies, the estimators for population means and variances can be quite unreliable. This unreliability is compounded by taking the ratios and powers to express the sample size.

Sample Sizes in ANOVA

The one-way analysis of variance (ANOVA), fixed effect model, generalizes the two sample t -test to more than two populations/samples. In terms of ANOVA, the j th observation within the i th treatment (group), y_{ij} , is modeled as

$$y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

where number of treatments is k , ($1 \leq i \leq k$, $k \geq 2$) and number of replicates within the i th treatment is n_i , ($1 \leq j \leq n_i$, $n_i \geq 1$). The total sample size is $N = n_1 + \dots + n_k$. If the design is *balanced*, all n_i are equal (to n) and $N = n \times k$.

The errors ϵ_{ij} are assumed to be independent normal random variables with mean 0 and common variance σ^2 .

The population group means μ_i s can be split to μ (grand mean) and α_i s (the treatment effects), where $\mu = 1/k \sum_i \mu_i$ and $\sum_i \alpha_i = 0$, and the null hypothesis of equality of all population means, $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$, can be equivalently rewritten in terms of population treatment effects α_i as, $H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$. Under H_0 , the test statistics has F distribution with $k - 1$ and $k(n - 1)$ degrees of freedom. However, if H_0 is not true, the test statistics has noncentral F distribution with $k - 1$ and $k(n - 1)$ degrees of freedom and noncentrality parameter $\lambda = \frac{\sum_i n_i \alpha_i^2}{\sigma^2}$ which is $\lambda = n \frac{\sum_i \alpha_i^2}{\sigma^2}$ for balanced designs. In 'Before Computers' times so called Hartley-Pearson charts have been used, nowadays they are as obsolete like the Table of Logarithms Book. It is important that software you are using can calculate percentiles of non-central distributions (t , χ^2 and F) since for important designs noncentral distributions model deviations from H_0 we want to control.

Example: Suppose $k = 4$ treatment means are to be compared at the significance level $\alpha = 0.05$. The experimenter is to decide how many replicates n to run at each level, so that the hypothesis H_0 is rejected with the probability of 0.9 if $f^2 = \frac{1/k \sum_i \alpha_i^2}{\sigma^2} = 0.25$, or if $\sum_i \alpha_i^2$ is equal to overall σ^2 .

For example, if you have access to MATLAB, then for $n = 16$, i.e., $N = 64$ the power is calculated in a one-line command:

```
>> 1-ncfcdf( finv(1-0.05, 3, 60), 3, 60, 16)
ans =
    0.9167
```

Here we used the precision $\sum_i \alpha_i^2 \approx \sigma^2$, which corresponds to $\lambda = n = 16$.

Here we used the fact that the power in ANOVA is, by definition,

$$1 - \beta = P(F^{nc}(k - 1, N - k, \lambda) > F^{-1}(1 - \alpha, k - 1, N - k)), \quad (2)$$

where $F^{nc}(k - 1, N - k, \lambda)$ is a random variable with noncentral F distribution with $k - 1$ and $N - k$ degrees of freedom and noncentrality parameter $\lambda = \frac{n \sum_i \alpha_i^2}{\sigma^2}$. The quantity $F^{-1}(1 - \alpha, k - 1, N - k)$ is the $1 - \alpha$ percentile (a cut point for which the upper tail has the probability α) of standard F distribution with $k - 1$ and $N - k$ degrees of freedom. Recall, α is prescribed significance level of the test, standardly taken as 5%.

One can now try different values of n (treatment sample sizes) to achieve the desired power; this change affects only two arguments in equation (2): $N - k = nk - k = (n - 1)k$ and $\lambda = n \frac{\sum_i \alpha_i^2}{\sigma^2}$.

If we wanted to plot the power for different sample sizes, the following simple matlab session will do it, see Figure . The inputs are k -number of treatments and the significance level α .

```

>> k=4; %number of treatments
>> alpha = 0.05; % significance level
>> x = linspace(1,30,100); %define x-grid for plot
>> y=[]; %set values of power for n=x (no need to be integer)
>> for i=1:100
    y =[y 1-ncfcdf( finv(1-alpha, k-1, k*(x(i)-1)), k-1, k*(x(i)-1), x(i))];
end
>> plot(x, y)
>> xlabel('n'); ylabel('power')

```

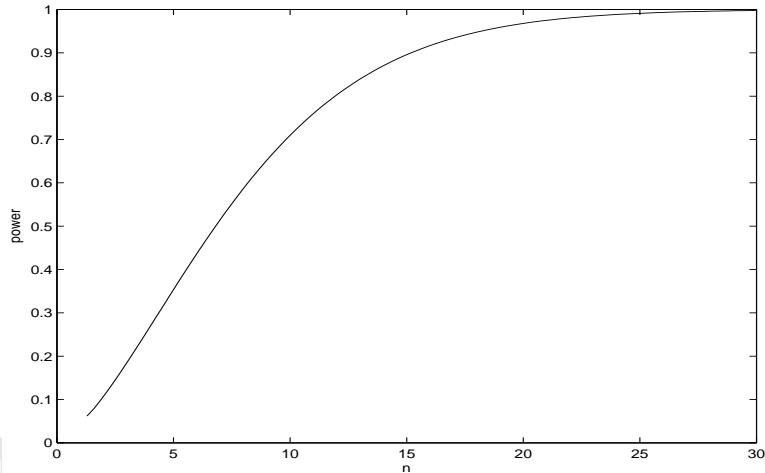


Figure 2: Power for $n < 30$ in a fixed effect ANOVA with $k = 4$ treatments and $\alpha = 0.05$.

In the above analysis, n is treatment sample size and the total sample size is $k \times n = N$. However, if we have a *block* design, i.e., the k treatments are applied to every subject, then number of subjects is equal to the number of blocks, b , but the total number of measurements is $N = nb$.

Sample Size in Regression

The evaluation of power in a regression with $p - 1$ variables, and p parameters (intercept added) requires specification of significance level and precision. Let's say that we want power so that total sample size of $n = 40$ will make $R^2 = 0.2$ significant for $\alpha = 0.05$ and number of predictors $p - 1 = 3$. Cohen's effect size here is $f^2 = R^2/(1 - R^2)$ and $f^2 \approx 0.04$ is small, $f^2 \approx 0.16$ is medium, and $f^2 \approx 0.36$ is a large effect.

Note that from $f^2 = R^2/(1 - R^2)$ one gets $R^2 = f^2/(1 + f^2)$.

The power would be found similarly as in ANOVA,

$$1 - \beta = P(F^{nc}(p - 1, N - p, \lambda) > F^{-1}(1 - \alpha, p - 1, n - p)), \quad (3)$$

where $\lambda = Nf^2$ is the noncentrality parameter.

For $R^2 = 0.2$ that is, $f^2 = 0.25$ and sample size of $N = 40$ and $p = 4$ one gets, $\lambda = 10$, and the power of about 71%.

```

>> p=4; N=40; lam=10;
>> 1-ncfcdf( finv(1-0.05,p-1, N-p), p-1, N-p, lam)

ans =    0.7097

```

G*Power

There are many specialized computer packages for power analysis both commercial and free. We describe a free package G*Power for Windows. G*Power performs high-precision statistical power analyzer for the most common statistical tests such as

- t -tests (independent samples, correlations, and any other t -test),
- F -tests (ANOVAS, multiple correlation and regression, and any other F -test), and
- χ^2 -tests (goodness of fit and contingency tables).

G*Power computes (i) power values for given sample sizes, Cohen's effect sizes, and alpha levels (post-hoc power analyses), (ii) sample sizes for given effect sizes, alpha levels, and power values (a priori power analyses), and (iii) alpha and beta values for given sample sizes, effect sizes, and beta/alpha ratios (compromise power analyses).

The program can be downloaded from

<http://www.psych.uni-duesseldorf.de/aap/projects/gpower/>.

G*Power also computes the effect size measures from the specified alternative hypothesis. A description of the program and the relevant background is published in Erdfelder et al (1996).

Figure is a screen capture of the G*Power analysis of the ANOVA example. The Cohen's effect size f^2 is connected to noncentrality parameter λ of the noncentral F distribution defined by the alternative H_1 , and the relationship is

$$\lambda = N \times f^2,$$

where N is the total sample size and $f^2 = \frac{\sum \alpha_i^2}{k\sigma^2}$.

According to the "T-shirt" convention, the effect of $f^2 = 0.25$ found in the previous example falls between Medium and Large. This effect corresponds to noncentrality parameter $\lambda = kn/4$, since by definition $\lambda = Nf^2$, and $N = kn$ (balanced design). For one-way ANOVA the effects are classified as $f^2 = 0.04$ small, $f^2 = 0.16$ medium, and $f^2 = 0.36$ large.

If, in the previous example, we wanted to control for the small effect $f = 0.2$, the required sample size achieving the power of 0.9 would be $N = 4 \times 98 = 392$.

Minitab

Minitab calculates sample size by measuring deviation from H_0 that does not involve $\sum_i \alpha_i^2$, but maximal difference between two means, $\max |\mu_i - \mu_j|$. When H_0 is true, the maximum difference between the treatment means is 0, i.e., $\max |\mu_i - \mu_j| = \max |\alpha_i - \alpha_j| = 0$. The detection of differences in treatment means that are comparable to σ is considered reasonable, that is,

$$\frac{\max |\mu_i - \mu_j|}{\sigma} = \frac{\max |\alpha_i - \alpha_j|}{\sigma} \approx 1.$$

Note that in this case for finding sample size no estimator of variance (standard deviation) is needed since we elicit precision in units of standard deviation, $\frac{\max |\alpha_i - \alpha_j|}{\sigma}$.

In the analysis below we assumed quite arbitrarily that $\sigma = 0.3$ and that $\max |\mu_i - \mu_j| = 0.3$ is the required precision. The same result would be obtained for $\sigma = 30$ and $\max |\mu_i - \mu_j| = 30$, for example.

Power and Sample Size
One-way ANOVA

Alpha = 0.05 Assumed standard deviation = 0.3 Number of Levels = 4

SS Means	Sample Size	Target Power	Actual Power	Maximum Difference
----------	-------------	--------------	--------------	--------------------

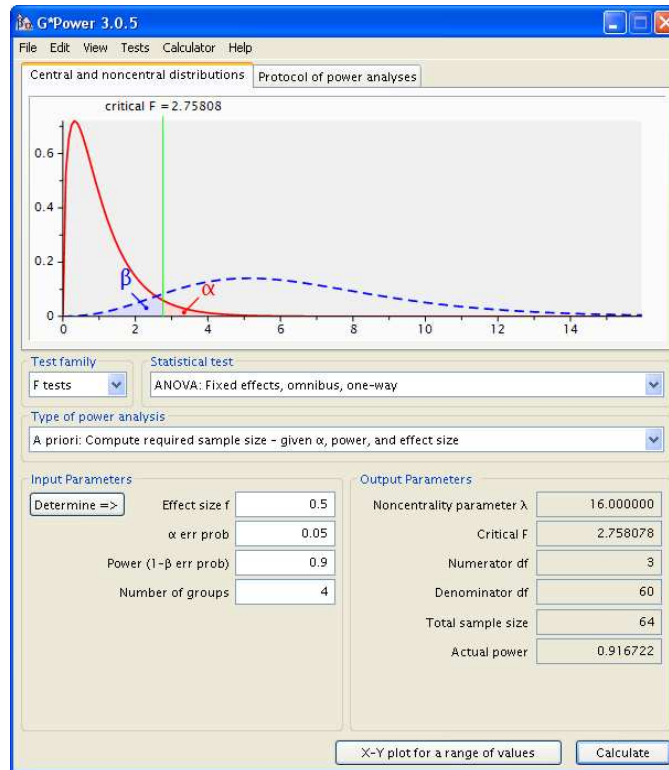


Figure 3: G*Power screen capture. Notice that the effect size $f = 0.5$ (or $f^2 = 0.25$) corresponds to $\sum_i \alpha_i^2 \approx \sigma^2$.

0.045 30 0.9 0.907115 0.3

The sample size is for each level.

A Case Study: Precision of the Pedometer.

The project assigned to students in BME1300 class several semesters ago was to test the precision of a pedometer (the counter of steps).

One group organized the experiment as follows. The subjects would walk on a treadmill for some time under different conditions and the pedometer count was contrasted to the human count obtained by several designated student-counters who simultaneously counted the steps. The human count was considered exact.

There were 4 treatments defined by two treadmill speeds Normal at 2.7 mi/h and Slow at 2 mi/h. Under the both speeds the subject will walk with unconstrained gait and constrained gait (Steps constrained by stepping marks on the treadmill)

-
- 1 = NU (Normal – Unconstrained)
 - 2 = NC (Normal – Constrained)
 - 3 = SU (Slow – Unconstrained)
 - 4 = SC (Slow – Constrained)
-

The student researchers wanted to test that the difference between the pedometer count and the human count was insignificant under all 4 regimes of walking thus statistically confirming the precision of the pedometer under this experimental design.

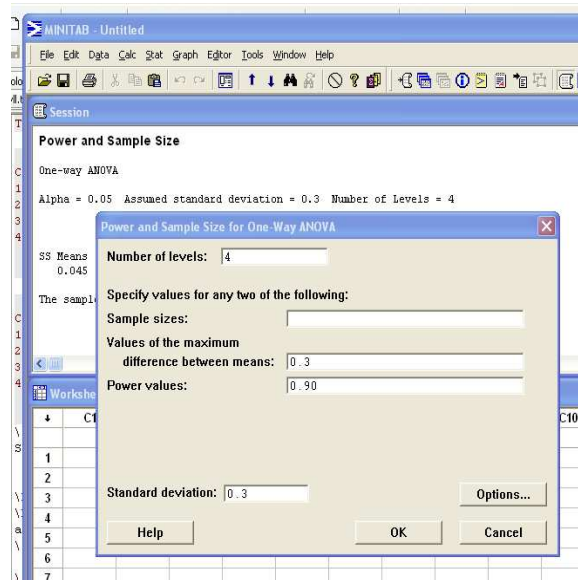


Figure 4: Minitab sample size

To account for difference between the subjects, each subject went through all 4 treatments. In total, $n = 24$ subjects produced $N = 4 \cdot 24 = 96$ data points (pedometer count - student count). This repeated measurements on a single subject leads to **block design**.

To implement this design, the two way additive ANOVA in MINITAB is used. When subjects are a factor, then this two-way additive ANOVA is in fact the *block design*. The differences y_{ij} (pedometer count minus student count in treatment i on subject j) are modeled as

$$y_{ij} = \mu_{ij} + \epsilon_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad i = 1, \dots, a, \quad j = 1, \dots, b,$$

where in this example $a = 4$ and $b = 24$. The mean μ_{ij} is represented as a sum of common mean μ , treatment effect α_i , subject or block effect β_j , and random error ϵ_{ij} . The errors are assumed normal with 0 mean and variance σ^2 . To make this decomposition of μ_{ij} unique, it is assumed that $\sum_i \alpha_i = 0$ and $\sum_j \beta_j = 0$.

We are interested in testing $H_0 : \alpha_1 = \dots = \alpha_a = 0$ (contributions to the mean by treatments are all equal to zero). The other null hypothesis of equality of contributions by subjects $H'_0 : \beta_1 = \dots = \beta_b = 0$ is not of interest.

In general, blocking is used to eliminate the difference between subjects entering the experiment from the analysis. Operationally, this is done by separating total sum of squares (total variability in data) as $SST = SSA + SSB + SSE$. Now, the inference on treatments would depend on SSA and SSE only but not on the SSB . The degrees of freedom for SSA , SSB , SSE and SST are $a - 1$, $b - 1$, $(a - 1)(b - 1)$ and $ab - 1$.

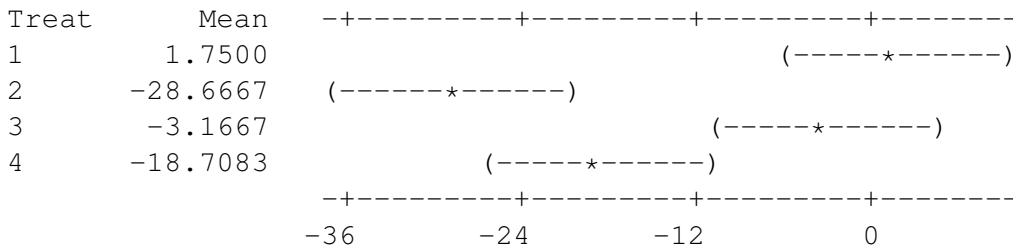
The ANOVA table produced in MINITAB is

Two-way ANOVA: Diff versus Treat, Subject

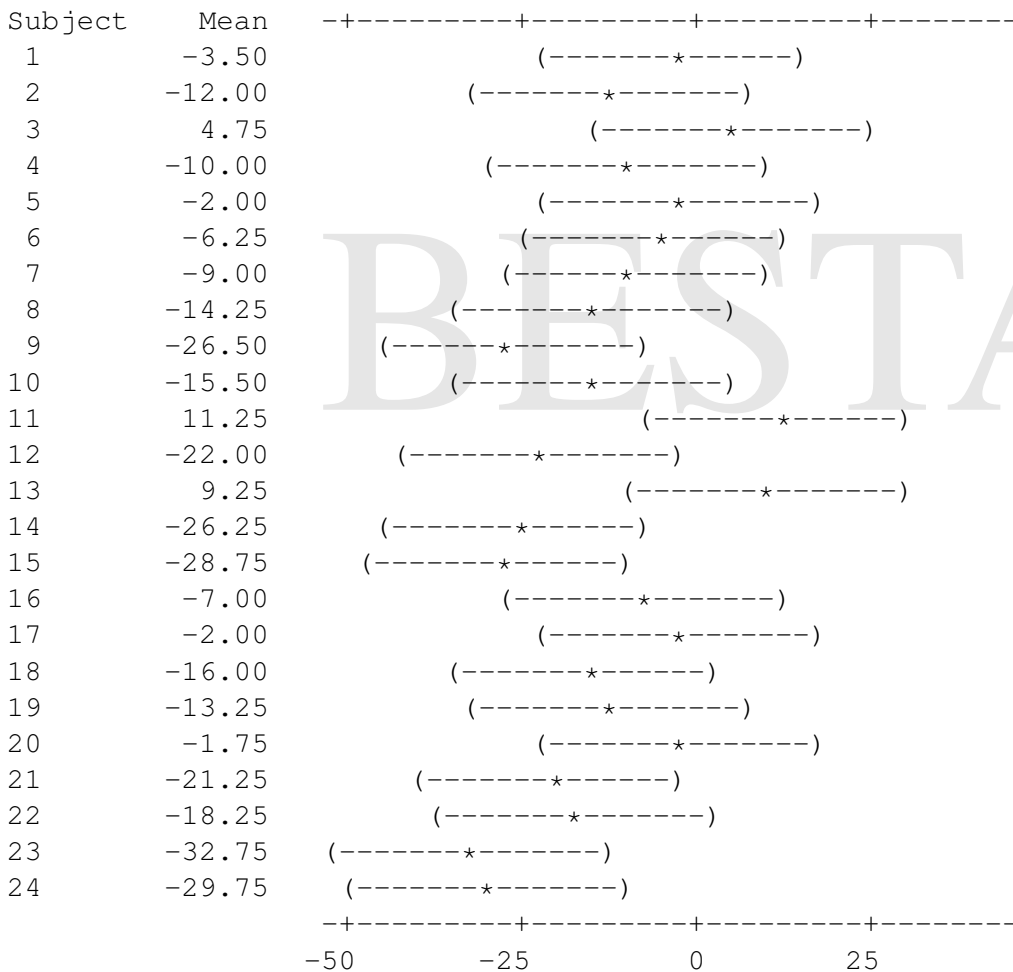
Source	DF	SS	MS	F	P
Treat	3	14153.1	4717.70	12.13	0.000
Subject	23	13676.5	594.63	1.53	0.091
Error	69	26843.6	389.04		
Total	95	54673.2			

S = 19.72 R-Sq = 50.90% R-Sq(adj) = 32.40%

Individual 95% CIs For Mean Based on Pooled StDev



Individual 95% CIs For Mean Based on Pooled StDev



Conclusion is that contribution to error by treatments is significant, with p -value essentially 0, or exactly

```
>> 1-fcdf(12.13, 3, 23)
```

```
ans = 5.7940e-005
```

The difference between subjects was not found significant at $\alpha = 5\%$ (p-value=0.091), so blocking did not contribute to a more precise analysis of treatments.

Power and Sample Size

For the block design, the noncentrality parameter $\lambda = \frac{b \sum_i \alpha_i^2}{\sigma^2}$. Assume that $\sum_i \alpha_i^2 \approx 2\sigma^2$ is required precision. This correspond to $f^2 = 0.5$ which is Cohen's Extra-Large effect.

Sample size, or in this case the block size $b = 24$ gives $\lambda = 24 \times 1/2 = 12$, and the power of about 82%:

```
>> 1-ncfcdf( finv(1-0.05, 4-1, (4-1)*(24-1)), 4-1, (4-1)*(24-1), 24/2)

ans = 0.8168
```

To achieve the power of 95% (keeping level $\alpha = 0.05$ and precision $\frac{\sum_i \alpha_i^2}{\sigma^2} = 1/2$), the block size (number of subjects) needed is $b = 36$:

```
>> b=35; 1-ncfcdf( finv(1-0.05, 4-1, (4-1)*(b-1)), 4-1, (4-1)*(b-1), b* 1/2)
ans = 0.9459
```

```
>> b=36; 1-ncfcdf( finv(1-0.05, 4-1, (4-1)*(b-1)), 4-1, (4-1)*(b-1), b* 1/2)
ans = 0.9520
```

Post-hoc analysis would involve estimating the non-centrality parameter λ from the study as $\hat{\lambda}$ and then recalculating the power with the estimator. Please, just say "No!" to post-hoc power analyses. There are many reasons, both common-sense and theoretical, why researchers should not do post-hoc power analyses. Excellent overviews can be found in Hoenig and Heisey (2001) and Levine and Ensom (2001). As Hoenig and Heisey show, power is mathematically directly related to the p -value; hence, assessing the power once you the p -value is found, adds no new information. Furthermore, as Levine and Ensom argue, the logic underlying post-hoc power analysis is fundamentally flawed. Last, but not list, a post-hoc power analyses were found to be abused; negative results have been ignored since "post-hoc power was low."

References

- [1] Chow, Shein-Chung (2003). *Design and Analysis of Clinical Trials: Concepts and Methodologies*, Wiley Series in Probability and Statistics.
- [2] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale: Erlbaum.
- [3] Erdfelder, E., Faul, F., and Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, **28**, 1–11.
- [4] isher R. A. (1926), "The Arrangement of Field Experiments," *Journal of the Ministry of Agriculture of Great Britain*, 33, 503-513.
- [5] John M. Hoenig and Dennis M. Heisey. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician* 2001: 55(1); 19-24.
- [6] Lenth, R. V. (2001). Some Practical Guidelines for Effective Sample Size Determination, *The American Statistician*, **55**, 187–193.
- [7] Levine, M., & Ensom, M. H. (2001). Post-hoc power analysis: An idea whose time has passed? *Pharmacotherapy*, 21(4), 405-409.

[8] Koele, P. (1982). Calculating power in analysis of variance. *Psychological Bulletin*, **92**, 513–516.

[9] Odeh, R. and Fox, M. (1991). *Sample Size Choice*, Statistics, a Series of Textbooks and Monographs, CRC Press.

Appendix

Standard Normal Quantiles

$1 - \alpha/2$	$\alpha/2$	Confidence $(1 - \alpha) \times 100\%$	$z_{1-\alpha/2}$
0.750	0.250	50	0.67449
0.800	0.200	60	0.84162
0.850	0.150	70	1.03643
0.900	0.100	80	1.28155
0.950	0.050	90	1.64485
0.975	0.025	95	1.95996
0.980	0.020	96	2.05375
0.990	0.010	98	2.32635
0.995	0.005	99	2.57583
0.999	0.001	99.8	3.09024

In MATLAB

```
>> lower_tail_area = [0.8 0.9 0.95 0.975 0.99 0.995];  
>> quantiles = norminv( lower_tail_area)
```

```
quantiles =
```

```
0.8416 1.2816 1.6449 1.9600 2.3263 2.5758
```

```
%for the following table
```

```
>> df=25; lower_tail_area = 0.975;  
>> tinv(lower_tail_area, df)
```

```
ans =
```

```
2.0595
```

Quantiles of Student t Distribution

Notation: $1 - \alpha/2$ is lower tail area, $\alpha/2$ upper tail area. For example, the quantiles needed for $1 - \alpha$ confidence interval are $-t_{1-\alpha/2,df} = t_{\alpha/2,df}$ and $t_{1-\alpha/2,df}$.

df	$1 - \alpha/2$						
	.80	.90	.95	.975	.99	.995	.9995
1	1.376	3.078	6.314	12.706	31.821	63.657	636.619
2	1.061	1.886	2.920	4.303	6.965	9.925	31.599
3	.978	1.638	2.353	3.182	4.541	5.841	12.924
4	.941	1.533	2.132	2.776	3.747	4.604	8.610
5	.920	1.476	2.015	2.571	3.365	4.032	6.869
6	.906	1.440	1.943	2.447	3.143	3.707	5.959
7	.896	1.415	1.895	2.365	2.998	3.499	5.408
8	.889	1.397	1.860	2.306	2.896	3.355	5.041
9	.883	1.383	1.833	2.262	2.821	3.250	4.781
10	.879	1.372	1.812	2.228	2.764	3.169	4.587
11	.876	1.363	1.796	2.201	2.718	3.106	4.437
12	.873	1.356	1.782	2.179	2.681	3.055	4.318
13	.870	1.350	1.771	2.160	2.650	3.012	4.221
14	.868	1.345	1.761	2.145	2.624	2.977	4.140
15	.866	1.341	1.753	2.131	2.602	2.947	4.073
16	.865	1.337	1.746	2.120	2.583	2.921	4.015
17	.863	1.333	1.740	2.110	2.567	2.898	3.965
18	.862	1.330	1.734	2.101	2.552	2.878	3.922
19	.861	1.328	1.729	2.093	2.539	2.861	3.883
20	.860	1.325	1.725	2.086	2.528	2.845	3.850
21	.859	1.323	1.721	2.080	2.518	2.831	3.819
22	.858	1.321	1.717	2.074	2.508	2.819	3.792
23	.858	1.319	1.714	2.069	2.500	2.807	3.768
24	.857	1.318	1.711	2.064	2.492	2.797	3.745
25	.856	1.316	1.708	2.060	2.485	2.787	3.725
26	.856	1.315	1.706	2.056	2.479	2.779	3.707
27	.855	1.314	1.703	2.052	2.473	2.771	3.690
28	.855	1.313	1.701	2.048	2.467	2.763	3.674
29	.854	1.311	1.699	2.045	2.462	2.756	3.659
30	.854	1.310	1.697	2.042	2.457	2.750	3.646
40	.851	1.303	1.648	2.021	2.423	2.704	3.551
50	.849	1.299	1.676	2.009	2.403	2.678	3.496
60	.848	1.296	1.671	2.000	2.390	2.660	3.460
100	.845	1.290	1.660	1.984	2.364	2.626	3.390
∞	0.842	1.282	1.645	1.960	2.327	2.576	3.292