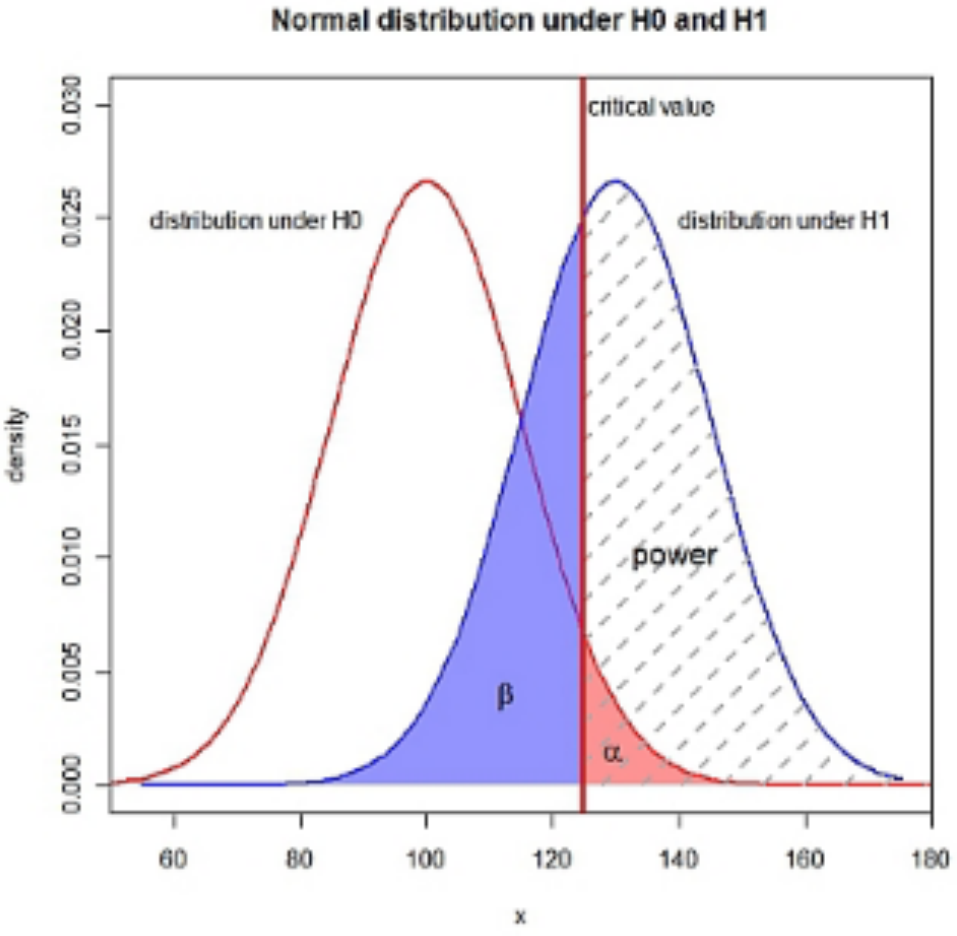
		Null hypothesis (H_0) is	
		True	False
Decision about null hypothesis (H_0)	Don't reject	True Negative Confidence: $P\left(S^{obs} \not\subset D^{crit} \mid H_0\right) = 1 - \alpha$	False Negative Type II error: $P\left(S^{obs} \not\subset D^{crit} \mid H_1\right) = \beta$
	Reject	False Positive Type I error: $P\left(S^{obs} \subset D^{crit} \mid H_0\right) = \alpha$	True Positive Power: $P\left(S^{obs} \subset D^{crit} \mid H_1\right) = 1 - \beta$

Key trade-off

- Type I error and Type II errors are inversely related
- The more demanding in terms of minimizing Type I error (false positives), the less likely to find any results (but might end up with false negatives (higher Type II error)



Statistical power is one piece in a puzzle that has four related parts; they are:

- **Effect Size**. The quantified magnitude of a result present in the population. Effect size is calculated using a specific statistical measure, such as Pearson's correlation coefficient for the relationship between variables or Cohen's d for the difference between groups.
- **Sample Size**. The number of observations in the sample.
- **Significance**. The significance level used in the statistical test, e.g. alpha. Often set to 5% or 0.05.
- **Statistical Power**. The probability of accepting the alternative hypothesis if it is true.

All four variables are related. For example, a larger sample size can make an effect easier to detect, and the statistical power can be increased in a test by increasing the significance level

A power analysis involves estimating one of these four parameters given values for three other parameters. This is a powerful tool in both the design and in the analysis of experiments that we wish to interpret using statistical hypothesis tests.

For example, the statistical power can be estimated given an effect size, sample size and significance level. Alternately, the sample size can be estimated given different desired levels of significance.

> *Power analysis answers questions like "how much statistical power does my study have?" and "how big a sample size do I need?".*

— Page 56, The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results, 2010.

Perhaps the most common use of a power analysis is in the estimation of the minimum sample size required for an experiment.

> *Power analyses are normally run before a study is conducted. A prospective or a priori power analysis can be used to estimate any one of the four power parameters but is most often used to estimate required sample sizes.*

— Page 57, The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results, 2010.

As a practitioner, we can start with sensible defaults for some parameters, such as a significance level of 0.05 and a power level of 0.80. We can then estimate a desirable minimum effect size, specific to the experiment being performed. A power analysis can then be used to estimate the minimum sample size required.

Confidence Intervals for the Mean - Large Sample

We wish to construct interval estimates for the population mean μ. To begin with, we assume that:

(i) the population standard deviation σ is known;

(ii) the size of our random sample n is greater than or equal to 30.

Under the above assumptions, the *Central Limit Theorem* tells us that the random variable

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

is approximately a standard normal random variable.

This allows us to use the table of standard normal probabilities to make inferences about how "close" values of $\bar{X}$ "typically" are to the true mean μ.

95% Confidence Intervals for the Mean

Let $z_{0.025}$ denote the value such that 0.025 or 2.5% of the Z-distribution lies in the tail to the right of $z_{0.025}$.

Then

$$P(-z_{0.025} \leq Z \leq z_{0.025}) = 0.95.$$

It can be checked from the table of standard normal probabilities that $z_{0.025} = 1.96$.

Thus

$$P(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96) \quad = \quad 0.95$$

Rearranging the above inequalities, we find that

$$P(\bar{X} - 1.96(\frac{\sigma}{\sqrt{n}}) \leq \mu \leq \bar{X} + 1.96(\frac{\sigma}{\sqrt{n}})) = 0.95.$$

95% Confidence Intervals for the Mean

Error Magnitude and Sample Size

In fact, we can be 95% confident that the error $|\bar{x}-\mu| \leq E$ provided that

$$1.96\frac{\sigma}{\sqrt{n}} \quad \leq \quad E$$

which after rearrangement becomes

$$n \geq \left(\frac{1.96\sigma}{E}\right)^2.$$

In general, if we wish to be $100(1-\alpha)\%$ confident that the error in estimating the population mean with a value of the sample mean $\bar{x}$ is at most E then we need to choose a sample size

$$n \geq \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2.$$

Example

The light bulb manufacturer in our earlier example wants to choose a random sample and to be 95% confident that the error in estimating the mean lifetime of a bulb with the sample mean is less than 5 hours. How large a sample is required?

Solution:

$E = 5$, $\sigma = 20$, so we need to choose

$$n \geq \left(\frac{1.96(20)}{5}\right)^2 = 61.47.$$

So we need a sample of size 62 or larger.

Confidence Intervals for Means and Prop

Error Magnitude and Sample Size

Large Sample Confidence Intervals for Proportions

- We use the sample proportion $\hat{P}$ to estimate the population proportion p.
- For instance, if we wish to estimate the proportion of people in the electorate who will vote for a particular candidate, we could choose a sample of 400 people and count the number x of people in our sample who intend to vote for the candidate.
- Then the sample proportion takes the value $\hat{p} = \frac{x}{n}$.

In general, if X denotes the number of items in a sample of size n that have the attribute of interest, then the sample proportion is the statistic:

$$\hat{P} = \frac{X}{n}.$$

Confidence Intervals for Proportions

We have already seen that for large samples ($n \geq 30$) the distribution of

$$\frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

is approximately a standard normal distribution.

If $z_{\alpha/2}$ is the value of Z for which

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

then a $100(1-\alpha)\%$ confidence interval for p is given by

$$\hat{p} - z_{\alpha/2}\sqrt{p(1-p)/n} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{p(1-p)/n}.$$

Confidence Intervals for Proportions

- In the above formula, we require the population proportio[illegible] to compute the confidence interval.

Error Bounds and Sample Size

- As with estimating the mean, it is possible to make assertions with prescribed levels of confidence about the size of the error in estimating the population proportion p with a value $\hat{p}$ of the sample proportion.
- We can assert with $100(1-\alpha)\%$ confidence that the difference between $\hat{p}$ and p is at most

$$z_{\alpha/2}\sqrt{p(1-p)/n}.$$

Error Bounds and Sample Size

If we want to be $100(1-\alpha)\%$ confident that the estimation error is less than or equal to some value E, we require

$$z_{\alpha/2}\sqrt{p(1-p)/n} \leq E$$

or

$$n \geq \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p).$$

This formula involves the population proportion p, which is unknown.

Error Bounds and Sample Size

А/Б-тестирование

Что такое А/Б тесты?

Minimal sample size to estimate a population proportion

$$n = \frac{\hat{p}(1-\hat{p})z^2}{m^2},$$

where $\hat{p}$ is the estimated proportion, m the margin of error and z the z-score corresponding to the selected confidence level (for example $z = 1.96$ for a confidence level of 95%).

Minimal sample size to estimate a population mean

$$n = \frac{\sigma^2 z^2}{m^2},$$

where σ is the (expected) standard deviation in the population, m the margin of error and z the z-score corresponding to the selected confidence level (for example $z = 1.96$ for a confidence level of 95%).

It can be seen from the formula for sample size that sample size increases with an increase in the population variability, the degree of confidence and the precision level required of the estimate. Because the sample size is directly proportional to σ^2, the larger the population variability, the larger the sample size. Likewise, a higher degree of confidence implies a larger value of z, and thus a larger sample size. Both σ^2 and z appear in the numerator. Greater precision means a smaller value of D, and thus a larger sample size because D appears in the denominator.

6 *If the resulting sample size represents 10% or more of the population, the finite population correction (fpc) should be applied.*[6] The required sample size should then be calculated from the formula:

$$n_c = \frac{nN}{(N+n-1)}$$

where n = sample size without fpc
n_c = sample size with fpc.

7 *If the population standard deviation, σ^2, is unknown and an estimate is used, it should be re-estimated once the sample has been drawn*. The sample standard deviation, s, is used as an estimate of σ. A revised confidence interval should then be calculated to determine the precision level actually obtained.

Suppose that the value of 55.00 used for σ was an estimate because the true value was unknown. A sample of $n = 465$ is drawn, and these observations generate a mean

If we assume the simple random sampling is **with replacement**, then the sample values are independent, so the covariance between any two different sample values is zero. This fact is used to derive these formulas for the standard deviation of the estimator and the estimated standard deviation of the estimator. The first two columns are the parameter and the statistic which is the unbiased estimator of that parameter.

		standard deviation of the estimator	usual estimator of the standard deviation of the estimator
μ	$\overline{X}$	$\sqrt{\frac{\sigma^2}{n}}$	$\sqrt{\frac{s^2}{n}}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
p	$\hat{p}$	$\sqrt{\left(\frac{p(1-p)}{n}\right)}$	$\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}}$

Return to the top.

If we assume the simple random sampling is **without replacement**, then the sample values are **not** independent, so the covariance between any two different sample values is **not** zero. In fact, one can show that

Covariance between two different sample values: $\operatorname{cov}\left(X_i, X_j\right) = -\frac{\sigma^2}{N-1}$ for $i \neq j$

This fact is used to derive these formulas for the standard deviation of the estimator and the estimated standard deviation of the estimator. The first two columns are the parameter and the statistic which is the unbiased estimator of that parameter.

		standard deviation of the estimator	estimator of the standard deviation of the estimator
μ	$\overline{X}$	$\sqrt{\frac{\sigma^2}{n}\left(1-\frac{n-1}{N-1}\right)}$	$\sqrt{\frac{s^2}{n}\left(1-\frac{n}{N}\right)}$ where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
p	$\hat{p}$	$\sqrt{\left(\frac{p(1-p)}{n}\right)\left(1-\frac{n-1}{N-1}\right)}$	$\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}\left(\frac{n}{n-1}\right)\left(1-\frac{n}{N}\right)}$

In sampling without replacement, the formula for the standard deviation of all sample means for samples of size n must be modified by including a finite population correction. The formula becomes:

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ where N is the population size, N=6 in this example, and n is the sample size, n=4 in this case. The finite population correction is the the second square root in this formula. Using this formula, you get the correct standard deviation for the the population of 360 sample means, namely, 0.540062.

Most of the time sampling is done without replacement. However, when n, the sample size, is less than 0.05 times the population size, N, the finite population correction can be dropped. For example, if N=1000, 0.05N = 0.05 1000=50, so if the sample size is 50 or less, the finite population correction can be dropped.

8 In some cases, precision is specified in relative rather than absolute terms. In other words, it may be specified that the estimate be within plus or minus R percentage points of the mean. Symbolically:

$$D = R\mu$$

In these cases, the sample size may be determined by:

$$n = \frac{\sigma^2 z^2}{2}$$

$$= \frac{C^2 z^2}{R^2}$$

where the coefficient of variation $C = \sigma/\mu$ would have to be estimated.

Table 15.2 Summary of sample size determination for means and proportions

Steps	Means	Proportions
Specify the level of precision	$D = \pm$ €5.00	$D = p - \pi = \pm 0.05$
Specify the confidence level (CL)	CL = 95%	CL = 95%
Determine the z value associated with the CL	z value is 1.96	z value is 1.96
Determine the standard deviation of the population	Estimate σ. $\sigma = 55$	Estimate π:π = 0.64
Determine the sample size using the formula for the standard error	$n = \frac{\sigma^2 z^2}{D^2}$ $= \frac{55^2(1.96)^2}{5^2}$ $= 465$	$n = \frac{\pi(1-\pi)z^2}{D^2}$ $= \frac{0.64(1-0.64)(1.96)^2}{(0.05)^2}$ $= 355$
If the sample size represents ≥10% of the population, apply the finite factor popular correction (fpc) If necessary, re-estimate the confidence interval by employing s to estimate σ	$n_c = \frac{nN}{N+n-1}$ $= \bar{X} \pm z s_{\bar{x}}$	$n_c = \frac{nN}{N+n-1}$ $= p \pm z s_p$
If precision is specified in relative rather than absolute terms, determine the sample size by substituting for D	$D = R\mu$ $n = \frac{C^2 z^2}{R^2}$	$D = R\pi$ $n = \frac{z^2(1-\pi)}{R^2\pi}$

2.0.1.1 Evan Miller formula

Formula to calculate number of sample is:

$$n = \frac{(z_{\alpha/2} * \sqrt{2 * \sigma^2_{control}}) + z_{\beta} * \sqrt{\sigma^2_{control} + \sigma^2_{experiment}})^2}{\delta^2} \tag{3}$$

where,

$$\sigma^2 = p(1-p) \tag{4}$$

with,
n: number of sample needed in each group
α: desired statistical significance level
β: 1 - power
δ: absolute practical significance

2.0.1.2 Pooled SE Stanford formula

Formula to calculate number of sample is:

$$n = \frac{2\sigma^2(Z_{\beta} + Z_{\alpha/2})^2}{\delta^2} \tag{5}$$

where,

$$\sigma^2 = p_{pool}(1 - p_{pool}) \tag{6}$$

with,
n: number of sample needed in each group
α: desired statistical significance level
β: 1 - power
Z_i: respective z-score under the standard normal distribution (0.84 for 80% power and 1.96 for 5% alpha)
δ: absolute practical significance

An effect size refers to the size or magnitude of an effect or result as it would be expected to occur in a population.

The effect size is estimated from samples of data.

The effect size does not replace the results of a statistical hypothesis test. Instead, the effect size complements the test. Ideally, the results of both the hypothesis test and the effect size calculation would be presented side-by-side.

- **Hypothesis Test**: Quantify the likelihood of observing the data given an assumption (null hypothesis).
- **Effect Size**: Quantify the size of the effect assuming that the effect is present.

Examples of effect sizes include the correlation between two variables,[2] the regression coefficient in a regression, the mean difference, or the risk of a particular event (such as a heart attack) happening.

Relationship to test statistics

Sample-based effect sizes are distinguished from test statistics used in hypothesis testing, in that they estimate the strength (magnitude) of, for example, an apparent relationship, rather than assigning a significance level reflecting whether the magnitude of the relationship observed could be due to chance. The effect size does not directly determine the significance level, or vice versa. Given a sufficiently large sample size, a non-null statistical comparison will always show a statistically significant result unless the population effect size is exactly zero (and even there it will show statistical significance at the rate of the Type I error used). For example, a sample Pearson correlation coefficient of 0.01 is statistically significant if the sample size is 1000. Reporting only the significant *p*-value from this analysis could be misleading if a correlation of 0.01 is too small to be of interest in a particular application.

Difference family: Effect sizes based on differences between means [edit]

The raw effect size pertaining to a comparison of two groups is inherently calculated as the differences between the two means. However, to facilitate interpretation it is common to standardise the effect size; various conventions for statistical standardisation are presented below.

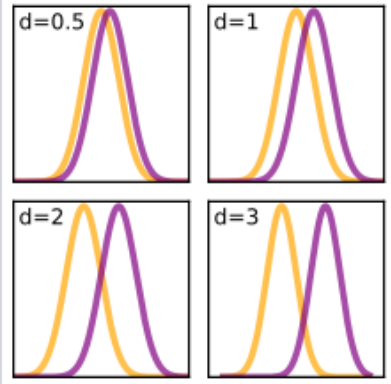
Standardized mean difference [edit]

A (population) effect size θ based on means usually considers the standardized mean difference (SMD) between two populations[21]:78

$$\theta = \frac{\mu_1 - \mu_2}{\sigma},$$

where μ_1 is the mean for one population, μ_2 is the mean for the other population, and σ is a standard deviation based on either or both populations.

In the practical setting the population values are typically not known and must be estimated from sample statistics. The several versions of effect sizes based on means differ with respect to which statistics are used.



Plots of Gaussian densities illustrating various values of Cohen's d.

This form for the effect size resembles the computation for a *t*-test statistic, with the critical difference that the *t*-test statistic includes a factor of $\sqrt{n}$. This means that for a given effect size, the significance level increases with the sample size. Unlike the *t*-test statistic, the effect size aims to estimate a population parameter and is not affected by the sample size.

SMD values of 0.2 to 0.5 are considered small, 0.5 to 0.8 are considered medium, and greater than 0.8 are considered large.[22]

Cohen's *d* [edit]

Cohen's *d* is defined as the difference between two means divided by a standard deviation for the data, *i.e.*

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}.$$

Jacob Cohen defined *s*, the pooled standard deviation, as (for two independent samples):[9]:67

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

where the variance for one of the groups is defined as

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1,i} - \bar{x}_1)^2,$$

and similarly for the other group.

The table below contains descriptors for magnitudes of d = 0.01 to 2.0, as initially suggested by Cohen and expanded by Sawilowsky.[10]

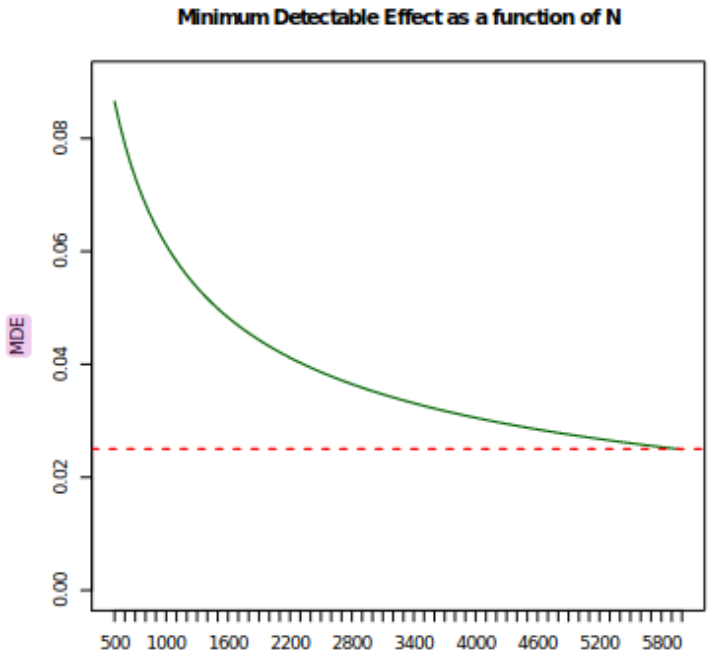
The Minimum Detectable Effect (MDE)

- Given my choice of power (usually 0.8) and my sample size (N), what's the smallest effect I can detect?

$$MDE = M_{n-2}\sqrt{\frac{\sigma^2}{Np(1-p)}}$$

- $M_{n-2} = t_{1-\frac{\alpha}{2}} + t_{1-\psi}$
- $1 - \psi = 1 - P(\text{Type II Error})$
- With $\alpha = 0.05$ and $\psi = 0.8$: $M_{n-2} = t_{0.975} + t_{0.8} = 2 + 0.8 = 2.8$
- MDE using Olken data

Minimum Detectable Effect



The MDE is estimated as,

$$MDE(k,\alpha,N,P) = (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

where $t_{(1-k)}$ is the t-statistic associated to the inverse of the power (Type II error rate) and t_{α} is the t-statistic associated with the significance level (Type I error rate). For a level of power of 80% and a significance level of 5%, these values are 0.84 and 1.96 respectively for large sample; for smaller samples, we have to use a student's t distribution with N-1 degrees of freedom instead of the normal distribution. To interpret the MDE, it is important to compare it to a given "standard" or reference value. This standard or reference value is usually derived from the theory of change and the indicator for success. For example, for a new cash transfer program the funders or implementers might decide that the anticipated treatment effect will be 100 US dollars in annual income. If the MDE for a given sample is $150, then we will not be able to statistically detect differences less than $150, and thus risk finding no statistical evidence of a positive treatment effect despite the effect's existence!

Power can be computed as,

$$t_{(1-k)}(N,\alpha,\beta_E,P) = \frac{\beta_E}{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}} - \tau_{\alpha};$$

where β_E is the effect size, or the expected change in the outcome as a consequence of the intervention. Notice that the MDE is also an effect size, but it is the minimum effect size for a given level of power and sample size. We distinguish between these two just to emphasize that the MDE is an estimate, whereas the effect size is a parameter (an assumption we make at the baseline).

The sample size can be estimated as,

$$N(k,\alpha,\beta_E,P) = \left[\frac{\sigma * (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}}{\beta_E}\right]^2;$$

where all the terms have been previously defined.