# How to determine the correct sample size of a research

**Dr Aseel Mugahed**

**M.B.B.S, PGD Public health,**
**Msc Global health**


**Training course in research methodology, research protocol development and scientific writing 2023, Geneva 2023**

# Contents

- Sample size calculation
- Factors that must be estimated to calculate sample size
- Steps to follow in sample size determination
- How to calculate sample size in cross-sectional studies
- How to calculate sample size in case control studies
- How to calculate sample size in clinical trials
- Further examples
- References for more reading
- Useful websites for sample size calculation

# Sample size calculation

- Sample size calculation or estimation has **no one single formula that can be universally** applied to all situations and circumstances.
- Sample size estimation can be done either by using;
  - Manual calculation,
  - Sample size software,
  - Sample size tables from scientific published articles,
  - Adopting various acceptable rules-of-thumb.

Bujang MA, 2021

# Factors that must be estimated to calculate sample size

- There are four factors that must be known or estimated to calculate sample size:

1. The effect size;
2. The population standard deviation;
3. The power of the experiment;
4. The significance level.

Dell et al., 2002

# 1.The effect size

# The effect size

- What do we mean by the *effect size*?
  - Simply put, the effect size is the minimal difference between the studied groups that the investigator wishes to detect or the difference between estimation and unknown parameter that the investigator wants to estimate.

Pourhoseingholi MA et al., 2013

# Key points about the effect size

- The effect size is independent of the sample size, only data is used to calculate effect sizes.
- A large effect size means that a research finding has practical significance, while a small effect size indicates limited practical applications.
- The practical significance shows that the effect is large enough to be meaningful in the real world.
- The most common effect sizes are **Cohen's (d)** and **Pearson's (r)** .

Bujang MA, 2021; Bhandari P, 2022

# Cohen's (d) & Pearson's (r) effect sizes

- **Cohen's (d)** measures the size of the difference between two groups.

**Cohen's *d* formula**

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

where,

- $\bar{x}_1$ = mean of Group 1
- $\bar{x}_2$ = mean of Group 2
- $s$ = standard deviation

- **Pearson's (r)** measures the strength of the relationship between two variables. It is better to use a statistical software to calculate Pearson's r accurately from the raw data.

**Pearson's *r* formula**

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

where,

- $r_{xy}$ = strength of the correlation between variables x and y
- $n$ = sample size
- $\sum$ = sum of what follows
- $X$ = every x-variable value
- $Y$ = every y-variable value
- $XY$ = the product of each x-variable score times the corresponding y-variable score

# How to estimate the effect size

The effect size can be estimated from:

- **Literature review**
  - It is better to obtain the needed information from recent articles (within 5 years) that used an almost similar design, same treatment and similar patient characteristics

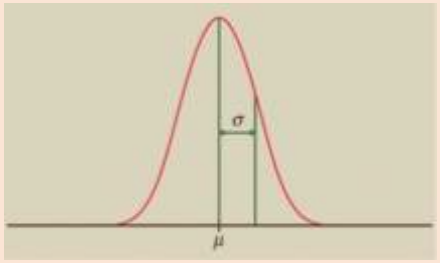
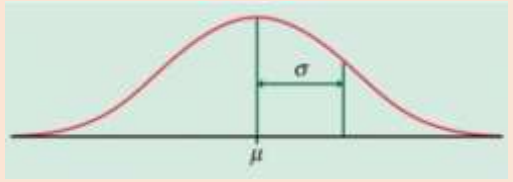
- **From historical data or secondary data**
  - Provided that the researcher has access to all data of different groups
  - Not always feasible since a new intervention may not have been assessed yet

- **Educated guess or expert opinion**
  - Researchers/experts can use their experience and knowledge to set up an effect size that is scientifically or clinically meaningful

# 2. The population standard deviation

# The population standard deviation

- What do we mean by *standard deviation*?
  - The standard deviation (or σ) is defined as a measure of how dispersed the data is in relation to the mean. In other words, it measures the typical distance between each data point and the mean.
- What do we understand by low and high standard deviation?
  - Low standard deviation means data are clustered around the mean (μ), and high standard deviation indicates data are more spread out.

| Low standard deviation | High standard deviation |
| --- | --- |
|  |  |

  - The standard deviation depends on whether the data is being considered a population of its own, or a sample representing a larger population.

Finding and Using Health Statistics [cited 2023 Aug 23]

# How to calculate standard deviation

## The population standard deviation ($\sigma$)

The population standard deviation is used when;

a. The researchers have the entire population
b. The researchers have a sample of a larger population, but they are ONLY interested in this sample and DO NOT wish to generalize the findings to the population.

If the data is being considered a population on its own, we divide $\sum (X - \mu)^2$ by the number of data points (scores in sample), n.

The **population standard deviation**

$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{n}}$$

where,

$\sigma$ = population standard deviation
$\sum$ = sum of...
$\mu$ = population mean
n = number of scores in sample.

***Steps to reach the standard deviation of a data***

1. *Subtract the mean from each data point (X)*
2. *Square the value*
3. *Add them all together*
4. *Divide by the total number of data points (scores)*
5. *Take the square root*

## The sample standard deviation (S)

In statistics, researchers usually use a sample from which they wish to GENERALIZE to a population. In this case researchers will need to use the sample standard deviation.

If the data is a sample from a larger population, we divide $\sum (X-\bar{X})^2$ by one fewer than the number of data points (scores) in the sample, n-1.

The **sample standard deviation formula** is:

$$s = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$$

where,

s = sample standard deviation
$\sum$ = sum of...
$\bar{X}$ = sample mean
n = number of scores in sample.


Finding and Using Health Statistics [cited 2023 Aug 23]; Laerd Statistics [cited 2022 Sep 9]

# Key points about standard deviation

- Note that confusion may arise when the name "sample" standard deviation is used, as it can be incorrectly interpreted as the standard deviation of the sample itself and not the estimate of the population standard deviation based on the sample.
- The standard deviation is used with continuous data, BUT NOT with categorical data.
- The standard deviation is appropriate when the continuous data IS NOT SIGNIFICANTLY SKEWED OR HAS OUTLIERS.
- The standard deviation can be derived from a literature review, pilot study or from any reliable source.


Laerd Statistics [cited 2022 Sep 9]

Finding and Using Health Statistics [cited 2023 Aug 23]

# 3. The power

# The power of a study

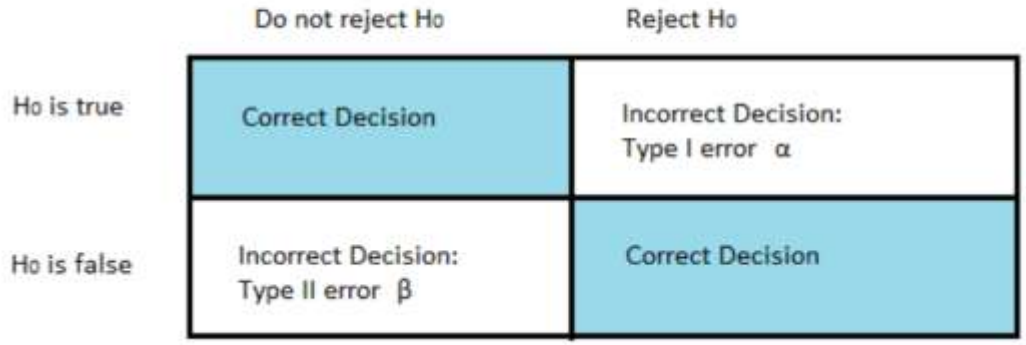
- What do we mean by the *power of a study*?
  - The power of a study can be defined as the pre-study probability that the effect will be detected, and that the test will reject the null hypothesis. In other words, if the power is too low, there is little chance of detecting a significant difference.
  - The power is arbitrarily set to 80% or 90% which means that the study has an 80% or 90% chance of having statistically significant results.

Dell et al., 2002; Jones SR et al., 2003

# The type I and type II errors

- What do we mean by the *type I and type II errors*?
  - The type I error Alpha (α): is the rejection of a true null hypothesis. In other words, it corresponds to the level of confidence in sample size calculation, which is the degree of uncertainty or probability that a sample value lies outside a stated limits. **False positive**
  - The type II error Beta (β): is the failure to reject a false null hypothesis. In other words, it corresponds to power, which means the ability of a statistical test to reject the false null hypothesis. **False negative**

Statistical Power and Beta




Pourhoseingholi MA et al., 2013
StatisticsHowTo.com [cited 2022 Sep 5]

# 4. The significance level

# The significance level

- What do we mean by the *significance level*?
  - The significance level, also known as **alpha (α)** is the probability that a positive finding is due to chance alone. It is a measure of the strength of the evidence that must be present in the sample before rejecting the null hypothesis and concluding that the effect is statistically significant.
- What does it mean when the researcher determines a significance level of 0.05 or 0.01 before conducting the experiment?
  - It simply means that the investigator wishes the chance of mistakenly designating a difference "significant" (when in fact there is no difference) to be no more than 5% or 10%. The significance level is correlated with power: increasing the significance level (from 5% to 10%) increases power

Dell et al., 2002

# Steps to follow in sample size determination

**1. Understand the objective of the study**

The objective of a study has to be measurable, meaning that it can be determined by using statistical analysis.

**2. Select the appropriate statistical analysis**

Researchers have to decide the appropriate analysis or statistical test to be used to answer the study objective. The formula that will be used to estimate or calculate the sample size will be the same formula for performing the statistical test that will be used to answer the objective of study.

**3. Calculate or estimate the sample size**

Estimating or calculating the sample size can be done either by using manual calculation, sample size software, sample size tables from scientific published articles, or by adopting various acceptable rules-of-thumb.

**4. Provide an additional allowance during subject recruitment to cater for a certain proportion of non-response**

A minimum required sample size means the minimum number of subjects a study must have after recruitment is completed. Therefore, the researchers must ideally be able to recruit subjects at least beyond the minimum required sample size. It is advisable to add 20-30% more. If the chance of non response is high then it can be increased up to 40-50%.

**5. Write a sample size statement by outlining steps from 1 to 4**

The sample size statement is usually included in the protocol or manuscript. There are various styles to write the statement.

Bujang MA, 2021

# Examples of free sample size calculator tools

- **StatCalc: https://www.cdc.gov/epiinfo/user-guide/statcalc/statcalcintro.html**
  - A utility tool in Epi Info™ and statistical calculator that produces summary epidemiologic information.
  - Six types of calculations are available including Sample Size and Power calculations for Population Survey, Cohort or Cross-Sectional, and Unmatched Case-Control.
  - Epi Info™ is a free software that can be downloaded from the Centers for Disease Control and Prevention (CDC) website at https://www.cdc.gov/epiinfo.
  - Watch the Epi Info™ 7 Tutorial Videos.

- **OpenEpi.com: https://www.openepi.com**
  - An open-source web tool that provides additional epidemiologic statistics not included in StatCalc.

- **ClinCalc LLC. Sample Size Calculator: https://clincalc.com/stats/SampleSize.aspx**
  - A free online sample size calculator

# How to calculate sample size in cross-sectional studies

- In cross-sectional studies or surveys the researchers usually aim to study the prevalence of some disease in a community or finding the average value of some quantitative variable in a population.
- In cross-sectional study, the sample size formula is as follows,

Sample size formula for
**Cross-sectional studies with qualitative variables**

$$\text{Sample size} = \frac{Z_{1-\alpha/2}^2 p(1-p)}{d^2}$$

Here

$Z_{1-\alpha/2}$ = Is standard normal variate (at 5% type 1 error ($P<0.05$) it is 1.96 and at 1% type 1 error ($P<0.01$) it is 2.58). As in majority of studies $P$ values are considered significant below 0.05 hence 1.96 is used in formula.
p = Expected proportion in population based on previous studies or pilot studies.
d = Absolute error or precision – Has to be decided by researcher.

Sample size formula for
**Cross-sectional studies with quantitative variables**

$$\text{Sample size} = \frac{Z_{1-\alpha/2}^2 SD^2}{d^2}$$

$Z_{1-\alpha/2}$ = Is standard normal variate as mentioned in previous section.
SD = Standard deviation of variable. Value of standard deviation can be taken from previously done study or through pilot study.
d = Absolute error or precision as mentioned in previous section

Charan J et al., 2013

# Example 1: sample size for cross-sectional studies (qualitative variables)

- Suppose a researcher wants to estimate the proportion of women suffering from *obstetric fistula* in a population of a certain village, and according to previous studies the actual number of cases is not more than 15%. The researcher wants to find the sample size with an absolute error/precision of 5% and type 1 error of 5%.

Using the following formula for qualitative variables;

$$\text{Sample size} = \frac{Z_{1-\alpha/2}{}^2 p(1-p)}{d^2}$$

Sample size $=(1.96)^2 \times 0.15(1-0.15)/0.05^2 = 196$

So the researcher will need at least 196 women for the study.

Formula taken from:
Charan J et al., 2013

# Example 2: sample size for cross-sectional studies (quantitative variables)

- Suppose a researcher wants to know the average age of patients with *obstetric fistula* in the same village as the previous example at 5% of type of 1 error and a precision of plus/minus 5 years. The standard deviation, based on previous studies, is 25 years.

Using the following formula for quantitative variables;

$$\text{Sample size} = \frac{Z_{1-\alpha/2}{}^2 SD^2}{d^2}$$

* $Z_{1-\alpha/2}$ = Is the standard normal variate (at 5% type 1 error (P<0.05) it is 1.96)

Sample size = $(25)^2(1.96)^2/(5)^2 = 96$

So roughly, the sample size will be around 96 women who have *obstetric fistula*.

Formula taken from:
Charan J et al., 2013

# How to calculate sample size in case control studies

- In case control studies the researchers are interested in identifying the risk factors that may be associated to a disease in a group of people compared to another group of people who do not have the disease. The former will be the cases in the study while the latter will be the controls.
- In case control study, the sample size formula is as follows,

Sample size formula for
**Case control studies with qualitative variables**

$$\text{Sample size} = \frac{r+1}{r}\frac{(p^*)(1-p^*)(Z_\beta + Z_{\alpha/2})^2}{(p_1 - p_2)^2}$$

$r$ = Ratio of control to cases, 1 for equal number of case and control
$p^*$ = Average proportion exposed = proportion of exposed cases + proportion of control exposed/2
$Z_\beta$ = Standard normal variate for power = for 80% power it is 0.84 and for 90% value is 1.28. Researcher has to select power for the study.
$Z_{\alpha/2}$ = Standard normal variate for level of significance as mentioned in previous section.
$p_1 - p_2$ = Effect size or different in proportion expected based on previous studies. $p_1$ is proportion in cases and $p_2$ is proportion in control.

Sample size formula for
**Case control studies with quantitative variables**

$$\text{Sample size} = \frac{r+1}{r}\frac{SD^2(Z_\beta + Z_{\alpha/2})^2}{d^2}$$

SD = Standard deviation = researcher can take value from previously published studies
d = Expected mean difference between case and control (may be based on previously published studies.)
$r$, $Z_\beta$, $Z_{\alpha/2}$ are already explained in previous sections.

Charan J et al., 2013

# Example 1: sample size for case-control studies (qualitative variables)

- Suppose a researcher wants to know the relation between female genital mutilation in childhood and psychiatric disorders in adulthood. He will take two groups of women: one group with psychiatric disorders and the other with no psychiatric disorder. The researcher will then go retrospectively to study the FGM history in both groups. What will be the sample size of the study if the researcher decides to have an equal number of cases and controls and to fix the power of this study at 80%? (assuming the expected proportions in the case group and control group are 0.35 and 0.20 respectively)

Using the following formula for qualitative variables;

$$\text{Sample size} = \frac{r+1}{r}\frac{(p^*)(1-p^*)(Z_\beta + Z_{\alpha/2})^2}{(p_1 - p_2)^2}$$

p* = Average proportion exposed = (proportion of exposed cases + proportion of control exposed)/2 = (0.35 + 0.20)/2 = 0.275

Sample size = $2(0.275)(1\text{-}0.275)(0.84+1.96)^2/1\ (0.35+0.20)^2 = 138.9$

So the researcher will need to have 139 cases and 139 controls for the study.

Formula taken from:
Charan J et al., 2013

# Example 2: sample size for case-control studies (quantitative variables)

- Suppose that a researcher wants to study the link between birth weight and obesity in adulthood. He will take two groups of adults: one group will be obese (cases) and the other will be non-obese (control) and then trace back the weight of birth. Suppose that the **SD** in previous studies was 1kg and the mean difference between the case and control groups was 250gm.

Using the following formula for quantitative variables;

$$\text{Sample size} = \frac{r+1}{r}\frac{SD^2(Z_\beta + Z_{\alpha/2})^2}{d^2}$$

*$Z_{\alpha/2}$ = Is the standard normal variate (at 5% type 1 error (P<0.05) it is 1.96)

Sample size = $2\times1^2(0.84+1.96)^2/1\ (0.25)^2 = 250.8$

So roughly the sample size will be around 251 subjects in each group

Formula taken from:
Charan J et al., 2013

# How to calculate sample size in clinical trials

- In clinical trials the researchers are interested in knowing the effect of an intervention.
- In clinical trials, the sample size formula is as follows,

**Qualitative end point** (live/death, male/female..)

**Sample size formula**

$$\text{Sample size} = \frac{2(Z_{\alpha/2} + Z_{\beta})^2 P(1-P)}{(p_1 - p_2)^2}$$

$Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$ (From Z table) at type 1 error of 5%

$Z_{\beta} = Z_{0.20} = 0.842$ (From Z table) at 80% power

$p_1 - p_2$ = Difference in proportion of events in two groups

P = Pooled prevalence = [prevalence in case group ($p_1$) + prevalence in control group ($p_2$)]/2

**Quantitative end point** (age, weight, height…)

**Sample size formula**

$$\text{Sample size} = \frac{2SD^2(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$$

SD – Standard deviation = From previous studies or pilot study

$Z_{\alpha/2} = Z_{0.05/2} = Z_{0.025} = 1.96$ (From Z table) at type 1 error of 5%

$Z_{\beta} = Z_{0.20} = 0.842$ (From Z table) at 80% power

d = effect size = difference between mean values

So now formula will be

$$\text{Sample size} = \frac{2SD^2(1.96 + 0.84)^2}{d^2}$$

Charan J et al., 2013

# Example 1: sample size for clinical trials (qualitative end points)

- Suppose a researcher wants to know the effect of a drug on the mortality of breast cancer (BC) among women. He selected two groups: one to take the drug and the other to take placebo for a certain period of time, during which both groups will be under medical supervision. Assume that previous studies state 20% of BC patients die in a specific time. The researcher thinks if the drug increases the survival to 30% then the findings can be considered clinically significant. The researcher chooses the significance level to be 5% and the power of the study at 80%.

Using the following qualitative formula; $$\text{Sample size} = \frac{2(Z_{\alpha/2} + Z_{\beta})^2 P(1-P)}{(p_1 - p_2)^2}$$

* effect size=difference of the proportions. This equals to: 0.2-0.3=-0.1

Pooled prevalence = (0.20 + 0.30)/2 = 0.25

Sample size= $2(0.84+1.96)^2\ 0.25(1-0.25)/\ (-0.1)^2 = 294$

So the researcher will need to have 294 subjects per group.

Formula taken from:
Charan J et al., 2013

# Example 2: sample size for clinical trials (quantitative end points)

- Suppose that a researcher wants to study the effect of new diabetic drug against placebo. He thinks if this drug in comparison decreases the blood sugar by 10mg/dl it is considered to be clinically significant. Suppose that **SD** in previous studies was 25 mg/dl and the researcher selected the significance level to be at 5% and the power for the study at 80%. The researcher also thinks that the two tailed unpaired test is the statistically suitable test in this condition.

Using the following quantitative formula;

$$\text{Sample size} = \frac{2SD^2(Z_{\alpha/2}+Z_{\beta})^2}{d^2}$$

* effect size=10 mg/dl

$$\text{Sample size} = 2\times(25)^2(0.84+1.96)^2/(10)^2 = 98$$

So roughly the sample size will be around 98 subjects in each group

Formula taken from:
Charan J et al., 2013

# Further examples

- For more examples, go through the practical manual of sample size determination in health studies:

Lwanga SK, Lemeshow S. Sample size determination in health studies : a practical manual. World Health Organization, 1991. https://apps.who.int/iris/handle/10665/40062

# References for more reading


1. Bhandari P. What is Effect Size and Why Does It Matter? (Examples) [Internet]. Scribbr. 2020 [cited 2022 Sep 9]. Available from: https://www.scribbr.com/statistics/effect-size/

2. Bujang MA. A Step-by-Step Process on Sample Size Determination for Medical Research. Malays J Med Sci MJMS. 2021 Apr;28(2):15–27.

3. Charan J, Biswas T. How to Calculate Sample Size for Different Study Designs in Medical Research? Indian J Psychol Med. 2013;35(2):121–6.

4. Dell RB, Holleran S, Ramakrishnan R. Sample Size Determination. Ilar J. 2002;43(4):207–13.

5. Finding and Using Health Statistics [Internet]. [cited 2023 Aug 23]. Available from: https://www.nlm.nih.gov/oet/ed/stats/02-900.html

6. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31:337–50.

# References for more reading


7. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003 Sep 1;20(5):453–8.

8. Pourhoseingholi MA, Vahedi M, Rahimzadeh M. Sample size calculation in medical studies. Gastroenterol Hepatol Bed Bench. 2013;6(1):14–7.

9. Standard Deviation | How and when to use the Sample and Population Standard Deviation - A measure of spread [Internet]. Laerd Statistics [cited 2022 Sep 9]. Available from: https://statistics.laerd.com/statistical-guides/measures-of-spread-standard-deviation.php

10. Statistical Power: What it is, How to Calculate it [Internet]. Statistics How To [cited 2022 Sep 5]. Available from: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/statistical-power/

# Useful websites for sample size calculation

- ClinCalc LLC. Sample Size Calculator [website]. C2002. Available from: https://clincalc.com/stats/SampleSize.aspx
- Dean AG, Sullivan KM, Soe MM. OpenEpi: Open Source Epidemiologic Statistics for Public Health, Version. www.OpenEpi.com, updated 2013/04/06. Available from: https://www.openepi.com
- GIGAcalculator. Power & Sample Size Calculator [website]. c2017-2022. Available from: https://www.gigacalculator.com/calculators/power-sample-size-calculator.php
- Kohn MA, Senyak J. Sample Size Calculators [website]. UCSF CTSI. 20 December 2021. Available from https://www.sample-size.net/
- StatCalc in Epi Info™, Division of Health Informatics & Surveillance (DHIS), Center for Surveillance, Epidemiology & Laboratory Services (CSELS) [website]. CDC. Available from: https://www.cdc.gov/epiinfo/user-guide/statcalc/samplesize.html
- <http://www.biomath.info>: a simple website of the biomathematics division of the Department of Pediatrics at the College of Physicians & Surgeons at Columbia University, which implements the equations and conditions discussed in this article
- <http://davidmlane.com/hyperstat/power.html>: a clear and concise review of the basic principles of statistics, which includes a discussion of sample size calculations with links to sites where actual calculations can be performed
- nQuery Advisor, SPSS, MINITAB and SAS/STAT are paid statistical programs and software that can be used both for sample size calculations and statistical data analysis

# Thank you!