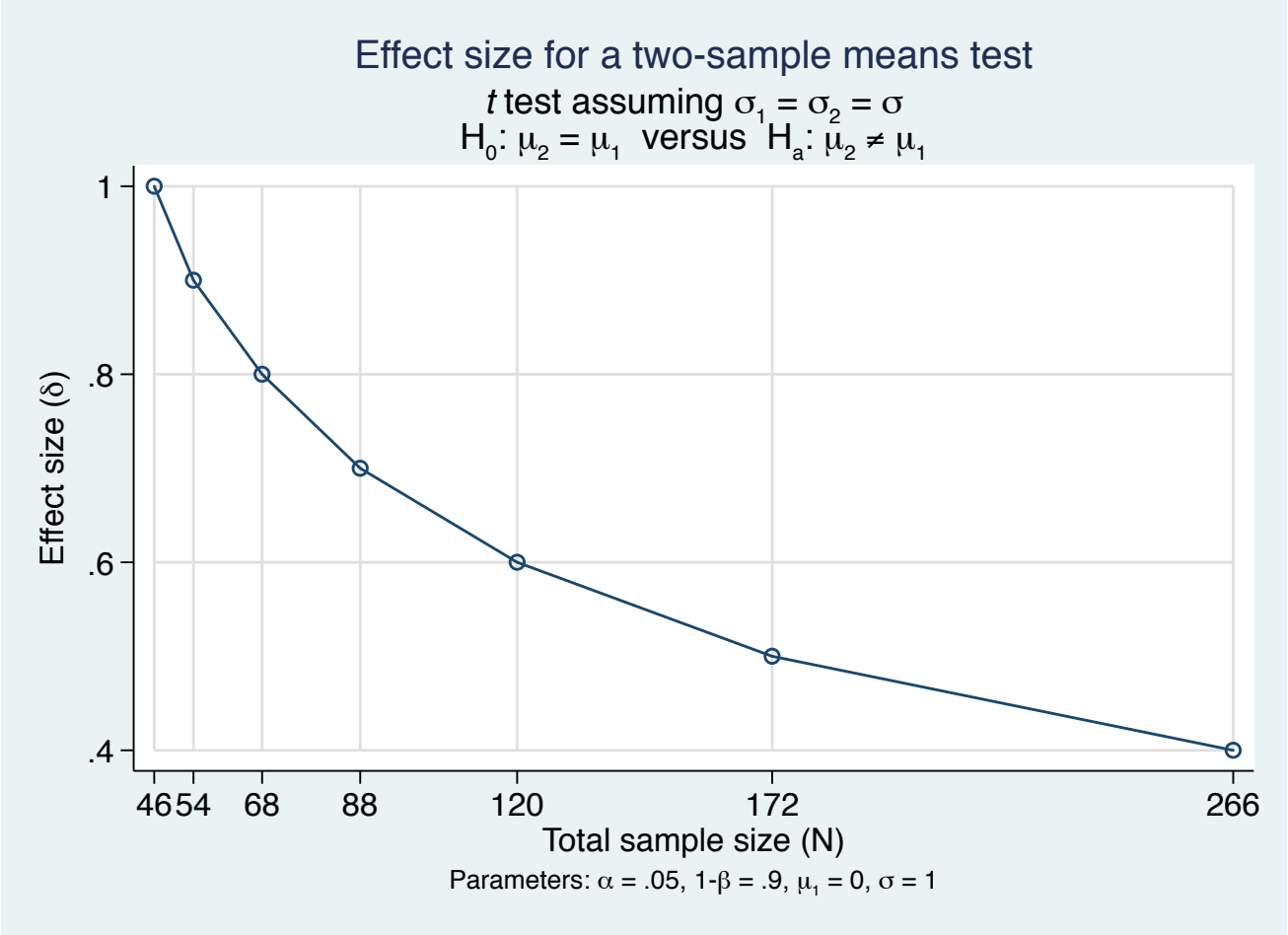
The RCSI Sample size handbook

A rough guide

May 2021 version

Ronán M Conroy

rconroy@rcsi.ie



Stata command

```
. power twomeans 0 (.4(.1)1), power(0.9) graph(ydimension(delta)
xdimension(N))
```

How to use this guide

This guide has sample size ready-reckoners for many common research designs. Each section is self-contained You need only read the section that applies to you.

If you are new to sample size calculation, read the first section first.

Examples

There are examples in each section, aimed at helping you to describe your sample size calculation in a research proposal or ethics committee submission. They are largely non-specialist. If you have useful examples, I welcome contributions.

Feedback

If you have trouble following this guide, please email me. Your comments help to improve it. If you spot an error, please let me know.

Support

This guide has slowly percolated around the internet. I'm pleased to handle queries from staff and students of RCSI and affiliated institutions. However, I cannot deal with queries from elsewhere. I'm sorry.

Warranty?

This document is provided as a guide. While every attempt has been made to ensure its accuracy, neither the author nor the Royal College of Surgeons in Ireland takes any responsibility for errors contained in it.

What's new

This version May 2021.

2020 : Updated text additional references and discussions based on current literature, updated web links.

2021 : Intraclass correlation section added.

Introduction : sample size and why it’s important

Sample size is an important issue in research. Ethics committees and funding agencies are aware that if a research project is too small, it misses failing to find what it set out to detect. Not only does this waste the input of the study participants (and frequently, in the case of animal research, their lives) but by producing a false negative result a study may do a disservice to research by discouraging further exploration of the area.

And, of course, if a study is too large it will waste resources that could have been spent on something else.

So the ideal sample size is one that collects sufficient data to have a good chance of measuring what you set out to measure.

Key issues: representativeness and precision

When choosing a sample, there are two important issues:

- will the sample be **representative** of the population, and
- will the sample be **precise** enough.

The first criterion of a good sample is sample **representativeness**. An unrepresentative sample will result in biased conclusions, and the bias cannot be eliminated by taking a larger sample. For this reason, sampling methodology is the first thing to get right.

The second criterion is sample **precision**. The larger the sample, the smaller the margin of uncertainty (confidence interval) around the results. However, there is another factor that also affects precision: the variability of the thing being measured. The more something that varies from person to person the bigger your sample needs to be to achieve the same degree of certainty about your results.

This guide deals with the issue of sample size. Remember, however, that sample size is of secondary importance to sample representativeness.

Key terms used in this sample size calculation

Precision – what it is, what determines it

Precision is the amount of potential error in a finding. Low-precision studies have wide margins of error around their findings, while high-precision studies have narrow margins of error. **The degree of precision is partly determined by the sample size**. In some sample size calculations, you will need to begin by deciding how much precision you require or, equivalently, the degree of uncertainty you are prepared to tolerate in your findings.

Precision is also determined by the variability in the thing you are studying. If something has little variation, such as body temperature, then you

will require a smaller sample than for something that varies quite a lot, like blood pressure.

You can easily imagine variability when it comes to things measured on a numeric scale. But what about things that are measured on a simple dichotomous scale – present/absent, true/false for example. Years ago, a colleague came up with an excellent example. Imagine a crowd of spectators where the supporters of one team wore white and supporters of the other team wore black. If one team had 100% support, the crowd would be all one colour – no variability. The maximum variability would occur where there was 50% support for each team. This is exactly what happens with dichotomous variables. The closer the prevalence is to 50%, the higher the variability. At 0% and 100% there's no variability at all.

Prevalence

Prevalence is how frequently a characteristic is found is in the population you are studying. Although we speak of prevalences every time we say something like “ten percent of people” or “a third of new admissions”, we rarely use the word prevalence for these fractions or percentages. This guide will use ‘prevalence’ as a general term for proportions, fractions and percentages.

Variability

The more variable is the thing we are studying, the more data we will have to gather in order to achieve a given level of precision. This makes sense intuitively when we are measuring something on a numeric scale. But it also applies to other types of measurement too, even to percentages.

Looking at the tables that show sample sizes for different prevalences, you will see that the required sample size rises as the prevalence approaches 50%. This is because when 50% of people have a characteristic and 50% do not, that characteristic has the highest person-to-person variability. As the prevalence nears zero or 100%, variability decreases, and so the required sample size will also decrease.

For continuous variables, the standard deviation is used as a measure of variability. This is sometimes known, or guessable, from previously published work, and this guide will tell you how to do this. But even if it is unknown, the guide will show you how to make an informed guess.

Effect size

Many sample size calculations require you to stipulate an **effect size**. This is the smallest effect that is **clinically significant** (as opposed to statistically significant). Clinical significance is a health research term that is used to mean “practical significance” or “real life significance”. The task of deciding on the smallest effect that would be clinically significant requires knowledge of the purpose of the research and the current state of knowledge and practice.

For example, if you are planning to compare two treatments, you have to decide how big a difference between two groups should be before it would be regarded as clinically important. You might define it as the smallest effect that would be noticed by the person being treated, or the smallest effect that would alter the management of the patient, or the smallest effect required to change the person's diagnosis.

The whole question of what constitutes a clinically significant finding is outside the scope of statistics. However, you will see from the tables that I have tried to help out by translating the rather abstract language of effect size into terms of patient benefit or differences between people.

What effect size isn't

It is important to realise what effect size **is not**. Effect size is not the effect that you think is there. We tend to have high hopes for our theories, and therefore hope that the treatment or risk factor we are interested in will have a very important effect. However, in sample size calculation, effect size is always the *smallest* effect that would be clinically significant. Not the one that you hope is there.

Importantly, too, effect size is **not** what was published by someone else. Again, this is an estimate of the actual effect size, but research must have adequate power to detect the smallest clinically significant effect size. Often the early publications in a field are biased towards larger effect sizes. This is not just because of publication bias, but also because methodologies will improve and things will always work less well when they leave the lab for the real world.

Power

Power if the chance that what you are looking for will be detected in your sample, if it actually exists. No sample, however big, is a guarantee that you will detect what you are looking for. However, it is foolish to do research without a reasonable chance that your study will detect it if it exists. And that "*if it exists*" is very important. **The power of a study is its chance of detecting an effect of a given size, if an effect of at least that size exists**.

Decades ago, studies were often run with 80% power. That is to say, there was an 80% chance that they would detect the effect if it existed but was the smallest clinically significant effect. And, therefore, there was a 20% chance – one in five – that they would fail to detect it and come to a false-negative conclusion.

A 20% probability of a false-negative conclusion is now regarded as unacceptable by ethics committees. Why waste the time (and lives) of research participants on projects that have a built-in 20% chance of failure? The sample sizes in this guide assume that you want 90% or even 95% power to detect what you are looking for.

Things that are not sample size calculations

Before going on to cover specific research scenarios, I should mention some things appear in ethics submissions and grant applications that are not acceptable as sample size calculations.

The following are the most usual offenders:

Everyone else used six animals per group

The legal advisor to RCSI's research ethics committee has advised us that there is no legal defence that runs *Well, everyone else did it too*. So the fact that someone else used this sample size does not justify it, whether or not the research was published. There are grounds for using 6 animals per group, but they are laid out below under comparing the means of two groups.

We did another study that used 10 patients and it was significant/got published

It's important, too, to emphasise here the point made above, that sample size should be set to detect the minimum effect that would be clinically significant, *not* the effect that someone else found or that the researcher thinks is there. Small studies are only likely to be published if they find something interesting. So they are likely to be misleading about the potential effect size.

We have limited funding/the student is only available for x weeks

Limited funding and limited time are not excuses for doing bad research. If you spend your resources on a research project that has no reasonable chance of being able to answer the research question because it is simply too small, then you have wasted your limited resources.

This is just a student project

Finally, student projects often lack the time and resources to recruit a sample big enough to have decent statistical power. Ethics committees understand that student research is where students learn research methods. So long as the application is accompanied by a calculation that shows the applicant is aware of the power of the proposed sample size, and the potential effect that this will have on the analysis and interpretation of the data, small sample size is not in itself an obstacle to receiving ethical approval – though it will probably be an obstacle to publication.

1. Sample size for percentages or proportions

This section give guidelines for sample sizes for
- studies that measure the proportion or percentage of people who have some characteristic,
- and for studies that compare this proportion with either a known population or with another group.

The characteristic being measured can be a disease, an opinion, a behaviour : anything that can be measured as present or absent.

Prevalence

Prevalence is the technical term for the proportion of people who have some feature. You should note that for a prevalence to be measured accurately, the study sample should be a valid sample. That is, it should not contain any significant source of bias.

1.1 Sample size for simple prevalence studies

The sample size needed for a prevalence study depends on how precisely you want to measure the prevalence. (**Precision** is the amount of error in a measurement.) The bigger your sample, the less error you are likely to make in measuring the prevalence, and therefore the better the chance that the prevalence you find in your sample will be close to the real prevalence in the population. You can calculate the margin of uncertainty around the findings of your study using **confidence intervals**. A confidence interval gives you a maximum and minimum plausible estimate for the true value you were trying to measure.

Step 1: decide on an acceptable margin of error

The larger your sample, the less uncertainty you will have about the true prevalence. However, you do not necessarily need a tiny margin of uncertainty. For an exploratory study, for example, a margin of error of ±10% might be perfectly acceptable. A 10% margin of uncertainty can be achieved with a sample of only 100. However, to get to a 5% margin of error will require a sample of 384 (four times as large).

Step 2: Is your population finite?

Are you sampling a population which has a defined number of members? Such populations might include all the physiotherapists in private practice in Ireland, or all the pharmacies in Ireland. If you have a finite population, the sample size you need can be significantly smaller.

Step 3: Simply read off your required sample size from table 1.1.

Table 1.1 Sample sizes for prevalence studies

Acceptable margin of error	Size of population					
	Large	**5000**	**2500**	**1000**	**500**	**200**
±20%	24	24	24	23	23	22
±15%	43	42	42	41	39	35
±10%	96	94	93	88	81	65
±7.5%	171	165	160	146	127	92
±5%	384	357	333	278	217	132
±3%	1067	880	748	516	341	169

Example 1: Sample size for a study of the prevalence of burnout in students at a large university

A researcher is interested in carrying out a prevalence study using simple random sampling from a population of over 11,000 university students. She would like to estimate the prevalence to within 5% of its true value.

Since the population is large (more than 5,000) she should use the first column in the table. A sample size of 384 students will allow the study to determine the prevalence of anxiety disorders with a confidence interval of ±5%. Note that if she wants increase precision so that her margin of error is just ±3%, she will have to sample over 1,000 participants. Sample sizes increase rapidly when very high precision is needed.

Example 2: Sample size for a study of a finite population

A researcher wants to study the prevalence of bullying in registrars and senior registrars working in Ireland. There are roughly 500 doctors in her population. She is willing to accept a margin of uncertainty of ±7.5%.

Here, the population is finite, with roughly 500 registrars and senior registrars, so the sample size will be smaller than she would need for a study of a large population. A representative sample of 127 will give the study a margin of error (confidence interval) of ±7.5% in determining the prevalence of bullying in the workplace, and 341 will narrow that margin of error to ±3%.

Frequently asked questions

Suppose my study involves analysing subgroups, how do I calculate sample size?

In some cases, you may be interested in percentages or prevalences within subgroups of your sample. In this case, you should check that they sample size will have enough power to give you an acceptable margin of error within the **smallest subgroup of interest**.

For example, you may be interested in the percentage of mobile phone users who are worried about the effects of radiation. A sample of 384 will allow you to measure this percentage with a margin of error of no more than ±5% of its true value.

However, you are also interested in subgroups, such as men and women, older and younger people, people with different levels of education etc. You reckon that the smallest subgroup will be older men, who will probably make up only 10% of the sample. This would give you about 38 men, slightly fewer than you need for a margin of error of ±20%. If this is not acceptable, you might increase the overall sample size, use stratified sampling (where a fixed number of each subgroup is recruited) or decide not to analyse rarer subgroups.

If you want to compare subgroups, however, go to section 1.3

What if I can only survey a fixed number of people?

You can use the table to find the approximate margin of error of your study. You will then have to ask yourself if this margin of error is acceptable. You may decide not to go ahead with the study because it will not give precise enough results to be useful.

How can I calculate sample size for a different margin of error?

All these calculations were done on a simple web page at

http://www.surveysystem.com/sscalc.htm

1.2 Sample sizes for studies comparing a prevalence with a hypothesised value

This section give guidelines for sample sizes for studies that measure the proportion or percentage of people who have some characteristic with the intention of comparing it with a percentage that is already known from research or hypothesised.

This characteristic can be a disease, and opinion, a behaviour, anything that can be measured as present or absent. You may want to demonstrate that the population you are studying has a higher (or lower) prevalence than some other population that you already know about. For example, you might want to see if medical students have a lower prevalence of smoking than other third level students, whose prevalence is already known from previous work.

Effect size

To begin with, you need to ask what is the smallest difference between the prevalence in the population you are studying and the prevalence in the reference population that would be considered meaningful in real life terms? This difference is often called a **clinically significant difference** in medicine, to draw attention to the fact that it is the smallest difference that would be important enough to have practical implications.

The bigger your study, the greater the chance that you will detect such a difference. And, of course, the smaller the difference that you consider to be clinically significant, the bigger the study you need to detect it.

Step 1: Effect size: Decide on the smallest difference the study should be capable of detecting

You will have to decide what is the smallest difference between the group that you are studying and the general population that would constitute a 'clinically significant difference' – that is, a difference that would have real-life implications. If you found a difference of 5%, would that have real-life implications? If not, would 10%? There is a certain amount of guesswork involved, and you might do well to see what the norm was in the literature.

For instance, if you were studying burnout in medical students and discovered that the rate was 5% higher than the rate for the general student population, would that have important real-life implications? How about if it was 10% lower? 10% higher? At what point would we decide that burnout in medical students was a problem that needed to be tackled?

Step 2: Prevalence: How common is the feature that you are studying in the population?

Sample sizes are bigger when the feature has a prevalence of 50% in the population. As the prevalence in the population group goes towards 0% or 100%, the sample size requirement falls. If you do not know how common the feature is, you should use the sample size for a 50% prevalence as being the worst-case estimate. The required sample size will be no larger than this, no matter what the prevalence turns out to be.

Step 3: what power do you want to detect a difference between the study group and the population?

A study with 90% power is 90% likely to discover the difference between the groups *if such a difference exists*. And 95% power increases this likelihood to 95%. So if a study with 95% power fails to detect a difference, the difference is unlikely to exist. You should aim for 95% power, and certainly accept nothing less than 90% power. Why run a study that has more than a 10% chance of failing to detect the very thing it is looking for?

Step 4: Use table 1.2 to get an idea of sample size

	Population prevalence 50%		Population prevalence 25%		Population prevalence 10%	
Difference between prevalences	Power		Power		Power	
	90%	95%	90%	95%	90%	95%*
+5%	1041	1287	883	1092	536	663
+10%	253	312	240	296	169	208
+15%	107	132	113	139	88	109
+20%	56	69	66	81	56	69
+25%	32	39	43	52	39	48
+30%	19	24	29	36	29	35
-5%	1041	1287	673	832	13	16
-10%	253	312	134	166		
-15%	107	132	43	52		
-20%	56	69	13	16		
–25%	32	39				
–30%	19	24				

Table 1.2 Comparing a sample with a known population

The table gives sample sizes for 90% and 95% power in three situations: when the population prevalence is 50%, 25% and 10%.

If in doubt about the prevalence, err on the high side.

*Sample Stata code for column

```
. power oneproportion .1 (.15(.05).4), test(wald) power(.95)
```

The bit that says `(.15(.05).4)` is a neat way of passing Stata a list of values. This one says “start at .15, increment by .05 and finish at .4”.

Example: Study investigating whether depression is more common in elderly people in nursing homes than in the general elderly population, using a limited number of available patients.

Depression has a prevalence of roughly 10% in the general elderly population. There are approximately 70 persons two nursing homes who will all be invited to participate in the research. A sample size of 69 would give the study 95% power to detect a 20% higher prevalence of depression in these participants compared with the general population.

Example: Study recruiting patients with low HDL cholesterol levels to see if there is a higher frequency of an allele suspected of being involved in low HDL. The population frequency of the allele is known to be 25%

The researchers decide that to be clinically significant, the prevalence of the allele would have to be twice as high in patients with low HDL cholesterol. A sample of 36 patients will give them a 90% chance of detecting a difference this big or bigger, while 45 patients will give them a 95% chance of detecting it.

Reference

These calculations were carried out using Stata Version 13 using the `power` command.

1.3 Sample sizes for studies comparing proportions between two groups

This section give guidelines for sample sizes for studies that measure the proportion or percentage of people who have some characteristic with the intention of comparing two groups sampled separately, or two subgroups within the same sample.

This is a common study design in which two groups are compared. In some cases, the two groups will be got by taking samples from two populations. However, in many cases the two groups may actually be subgroups of the same sample. If you plan on comparing two groups within the same sample, the sample size will have to be increased. Instructions for doing this are at the end of the section.

Step 1: Effect size: Decide on the difference the study should be capable of detecting

You will have to decide what is the smallest difference between the two groups that you are studying that would constitute a 'clinically significant difference' – that is, a difference that would have real-life implications. If you found a difference of 5%, would that have real-life implications? If not, would 10%? There is a certain amount of guesswork involved, and you might do well to see what the norm is in the literature.

Step 2: Prevalence: How common is the feature that you are studying in the comparison group?

Sample sizes are bigger when the feature has a prevalence of 50% in one of the groups. As the prevalence in one group goes towards 0% or 100%, the sample size requirement falls. If you do not know how common the feature is, you should use the sample size for a 50% prevalence as being the worst-case estimate. The required sample size will be no larger than this no matter what the prevalence turns out to be.

Step 3: Power: what power do you want to detect a difference between the two groups?

A study with 90% power is 90% likely to discover the difference between the groups if such a difference exists. And 95% power increases this likelihood to 95%. So if a study with 95% power fails to detect a difference, the difference is unlikely to exist. You should aim for 95% power, and certainly accept nothing less than 90% power. Why run a study that has more than a 10% chance of failing to detect the very thing it is looking for?

Step 4: Use table 1.3 to get an idea of sample size

The table gives sample sizes for 90% and 95% power in three situations: when the prevalence in the comparison group is 50%, 25% and 10%. If in doubt, err on the high side. The table shows the number in **each** group, so the total number is **twice** the figure in the table!

	Prevalence in one group 50%		Prevalence in one group 25%		Prevalence in one group 10%	
Difference between the groups	Power		Power		Power	
	90%*	95%	90%	95%	90%	95%
5%	2095	2590	1674	2070	918	1135
10%	519	641	440	543	266	329
15%	227	280	203	251	133	164
20%	124	153	118	145	82	101
25%	77	95	77	95	57	70
30%	52	63	54	67	42	52

Table 1.3 Numbers needed in each group

*Sample Stata command that generated the figures in this column

```
. power twoproportion .5 (.45(-.05).2),  power(.9)
```

The notation `.5 (.45(-.05).2)` is a way of telling Stata to generate a list of values starting with 0·5, decreasing in units of 0·05 and ending with 0·2)

Example: Study investigating the effect of a support programme on smoking quit rates

The investigator is planning a study of the effect of a telephone support line in improving smoking quit rates in patients post-stroke. She knows that about 25% of smokers will have quit at the end of the first year after discharge. She feels that the support line would make a clinically important contribution to management if it improved this this to 35%. The programme would not be justifiable from the cost point of view if the reduction were smaller than this. So a 10% increase is the smallest effect that would be clinically significant.

From the table she can see that two groups of 440 patients would be needed to have a 90% power of detecting a difference of at least 10%, and two groups of 543 patients would be needed for 95% power. She writes in her ethics submission:

Previous studies in the area suggest that as few as 25% of smokers are still not smoking a year after discharge. The proposed sample size of 500 patients in each group (intervention and control) will give the study a power to detect a 10% increase in smoking cessation rate that is between 90% and 95%.

Example: Study comparing risk of hypertension in women who continue to work and those who stop working during a first pregnancy.

Women in their first pregnancy have roughly a 10% risk of developing hypertension. The investigator wishes to compare risk in women who stop working and women who continue. She decides to give the study sufficient power to have a 90% chance of detecting a doubling of risk associated with continued working. The sample size, from the table, is two groups of 266 women. She decides to increase this to 300 in each group to account for drop-outs. She writes in her ethics submission:

Women in their first pregnancy have roughly a 10% risk of developing hypertension. We propose to recruit 300 women in each group (work cessation and working). The proposed sample size has a 90% power to detect a twofold increase in risk, from 10% to 20%.

Comparing subgroups within the same sample

This often happens when the two groups being compared are subgroups of a larger sample. For example, if you are comparing men and women coronary patients and you know that two thirds of patients are men.

A detailed answer is beyond the scope of a ready-reckoner table, because the final sample size will depend on the relative sizes of the groups being compared. Roughly, if one group is twice as big as the other, the total sample size will be 20% higher, if one is three times as big as the other, 30% higher. In the case of the coronary patients, if two thirds of patients are men, one group will be twice the size of the other. In this case, you would calculate a total sample size based on the table and then increase it by 20%.

Stata code

Suppose you are comparing two groups from the same sample. You are expecting the two groups to have a 20% and 80% prevalence. In this case, the ratio of the two groups is 80:20 which is 4:1. The Stata code for 90% power that gives the first column in the table above now reads

```
power twoproportions .5 (.45(-.05).2), test(chi2) power(0.9)
nratio(4)
```

You can see that you simply have to specify `nratio()` to get the appropriate calculation.

Frequently-asked questions

What is 90% or 95% power?

Just because a difference really exists in the population you are studying does not mean it will appear in every sample you take. Your sample may not show the difference, even though it is there. To be ethical and value for money, a research study should have a reasonable chance of detecting the smallest difference that would be of clinical significance (if this difference actually exists, of course). If you do a study and fail to find a difference, even though it exists, you may discourage further research, or delay the discovery of something useful. For this reason, you study should have a reasonable chance of finding a difference, if such a difference exists.

A study with 90% power is 90% likely to discover the smallest clinically significant difference between the groups if such a difference exists. And 95% power increases this likelihood to 95%. So if a study with 95% power fails to detect a difference, the difference is unlikely to exist. You should aim for 95% power, and certainly accept nothing less than 90% power. Why run a study that has more than a 10% chance of failing to detect the very thing it is looking for?

What if I can only study a certain number of people?

You can use the table to get a rough idea of the sort of difference you study might be able to detect. Look up the number of people you have available.

Reference

These calculations were carried out using Stata release 13 `power` command

1.4a Sample sizes for population case-control studies

This section give guidelines for sample sizes for studies that measure the effect of a risk factor by comparing a sample of people with the disease with a control sample of disease-free individuals **drawn from the same population**. The effect of the risk factor is measured using the odds ratio.
Population case-control studies have the disadvantage that the controls and cases may differ on variables that will have an effect on disease risk (confounding variables), so a multivariable analysis will have to be carried out to adjust for these variables. The sample sizes shown here are inflated by 25% to allow for the loss of statistical power that will typically result from adjusting for confounding variables.
If you are controlling for confounding variables by carrying out a **matched case-control study, see section 1.4b**.

A case-control study looks for risk factors for a disease or disorder by recruiting two groups of participants: cases of the disease or disorder, and controls, who are drawn from the same population as the cases but who did not develop the disease.

Case-control studies are **observational studies**. In experimental studies, we can hold conditions constant so that the only difference between the two groups we are comparing is that one group was exposed to the risk factor and the other was not. In observational studies, however, there can be other differences between those exposed to the risk factor and those not exposed. For example, if you are looking at the relationship between diarrhoeal disease in children and household water supply, households with high quality water will differ in other ways from households with low quality water. They are more likely to be higher social class, wealthier, and more likely to have better sanitation. These factors, which are associated with both the disease and the risk factor, are called **confounding factors**.

Understanding confounding factors is important in designing and analysing case-control studies. Confounding factors can distort the apparent relationship between a risk factor and a disease, so their effects have to be adjusted for statistically during the analysis. In the diarrhoeal disease example, you might need to adjust your estimate of the effect of good water quality in the household for the association between good water quality and presence of a toilet. Any case-control study must identify and measure potential confounding factors.

Sample size and adjustment for confounding factors

Allowing for confounding factors in the analysis of case-control studies increases the required sample size, because the statistical adjustment will

increase the margin of uncertainty around the estimate of the risk factor's odds ratio. If you don't understand the last bit, don't worry. The important thing is that you have to gather extra data in a case control study to allow you sufficient statistical power to adjust for confounding variables. How much extra data depends on how strongly the confounding factor is associated with the risk factor and the disease. Cousens and colleagues (see references) recommend increasing the sample size by 25%, based on simulation studies. **The sample sizes in the tables in this section are inflated by 25% in line with this recommendation**.

Step 1: Prevalence: What is the probable prevalence of the risk factor in your population?

The prevalence of the risk factor will affect your ability to detect its effect. If most of the population is exposed to the risk factor, it will be common in your control group, making it hard to detect its effect, for example. If you are unsure about the prevalence of the risk factor in the population, err on the extreme side – that is, if it is rare, use the lowest estimate you have as the basis for calculations, and if it is common use the highest estimate.

Step 2: Effect Size: What is the smallest odds ratio that would be regarded as clinically significant?

The odds ratio expresses the impact of the factor on the risk of the disease or disorder. Usually we are only interested in risk factors that have a sizeable impact on risk – and odds ratio of 2, for example – but if you are studying a common, serious condition you might be interested in detecting an odds ratio as low as 1.5, because even a 50% increase in risk of something common or serious will be important at the public health level.

Step 3: Power: What statistical power do you want?

With 90% power, you have a 90% chance of being able to detect a clinically significant odds ratio. That is, though, a 10% chance of doing the study and failing to detect it. With 95% power, you have only a 5% chance of failing to detect a clinically significant odds ratio, if it exists.

Step 4: Look up the number of cases from table 1.4

	Smallest odds ratio that would be clinically significant					
	1.5	2	2.5	3	4	5
Prevalence of the risk factor	90% Power to detect the odds ratio					
10%	1581	493	264	175	103	73
20%	929	300	165	113	69	50
30%	739	246	140	98	61	46
40%	674	231	134	95	63	49
50%	674	239	141	103	69	55
60%	730	265	161	118	81	65
70%	869	324	200	149	105	85
80%	1184	453	284	215	154	128
90%	2186	855	546	416	304	254
	95% Power to detect the odds ratio					
10%	1988	619	331	220	129	91
20%	1168	376	208	141	86	64
30%	929	309	175	121	78	59
40%	848	291	169	120	79	61
50%	848	300	178	129	86	69
60%	919	334	203	149	103	83
70%	1091	408	251	188	131	108
80%	1489	569	358	270	194	160
90%	2749	1075	686	524	383	320

Table 1.4a Number of cases required for a case control study

Note 1: This assumes a study that recruits an equal number of controls.

Note 2: The table has an allowance of 25% extra participants to adjust for confounding.

Example: A study to detect the effect of smoking on insomnia in elderly.

Step 1 is to estimate how common smoking is in the elderly. The current population estimate is that about 27% of the elderly smoke.

Step 2 is to specify the minimum odds ratio that would be clinically significant. In this case, we might decide that an odds ratio of 2.5 would be the smallest one that would be of real importance.

The table gives a sample size of 140 cases and 140 controls for 90% power to detect an odds ratio of at least 2.5 with a smoking prevalence of 30%. This is probably close enough to 27% to be taken as it is.

When analysing the data, the effect of smoking may be confounded by the fact that smoking is more common in men, and insomnia is also more common in men. So the apparent relationship between insomnia and smoking could be partly due to the fact that both are associated with male sex. We can adjust the odds ratio for sex, and for other confounding factors during the analysis. Although this will reduce the study power, the sample size table has a built-in allowance of 25% extra to deal with the loss of power due to confounding.

In an ethics submission, you would write

The sample size was calculated in order to have 90% power to detect an odds ratio of 2.5 or greater associated with smoking, given that the prevalence of smoking is approximately 30% in the target population. The sample size was inflated by 25% to allow for the calculation of an odds ratio adjusted for confounding variables such as gender, giving a planned sample size of 140 cases and 140 controls.

Frequently-asked questions

I only have 30 cases available to me – what can I do?

Looking at the table, it is clear that you cannot do a lot. You have a 90% chance of detecting a ten-fold increase in risk associated with a risk factor that is present in at least 20% of the population and at most 40%. Sample sizes for case-control studies are generally larger than people think, so it’s a good idea to look at the table and consider whether you have enough cases to go ahead.

Is there any way I can increase the power of my study by recruiting more controls?

Yes. If you have a limited number of cases, you can increase the power of your study by recruiting more controls.

Step 1 : Look up the number of cases you need from table 1.4

Step 2: Use table 1.5 to look up an adjustment factor based on the number of controls per case that you plan on recruiting. Multiply the original number of cases by the adjustment factor.

Step 3: the number of controls you require is based on this adjusted number.

Example: An obstetrician is interested in the relationship between manual work during pregnancy and risk of pre-eclampsia. She does some preliminary research and finds that about 20% of her patients do manual work during their pregnancy. She is interested in being able to detect an odds ratio of 3 or more associated with manual work. Since pre-eclampsia is comparatively rare, she plans to recruit three controls for each case.

Number of controls per case	Multiply the number of cases by
2	0.75
3	0.67
4	0.63
5	0.60

Table 1.4a1 Effect of multiple controls per case on sample size

From table 1.4, she needs 113 patients with pre-eclampsia for 90% power. Recruiting three controls per case, she can reduce this by a third (0.67), giving 113 x 0.67 = 75.7 cases (76 in round figures). However, she will have to recruit three controls per case, giving 228 controls (76 x 3). Although this is pretty close to the size of study she would have had to do with a 1:1 case-control ratio, it will be quicker to carry out, because recruiting the cases will be the slowest part of the study.

Reference

The calculations in this section were carried out with Stata, using formulas in

Cousens SN, Feachem RG, Kirkwood B, Mertens TE and Smith PG. Case-control studies of childhood diarrhoea: II Sample size.World Health Organization. CDD/EDP/88.3 Undated.

A scanned version may be downloaded here: http://www.ircwash.org/resources/case-control-studies-childhood-diarrhoea

1.4b Sample sizes for matched case-control studies

This section gives sample sizes for studies that compare cases of a disease or disorder with matched controls drawn from the same population.

Introduction

Case-control studies are widely used to establish the strength of the relationship between a risk factor and a health outcome. Case-control studies are **observational studies**. In experimental studies, we can hold conditions constant so that the only difference between the two groups we are comparing is that one was exposed to the risk factor and one was not. In observational studies, however, there can be other differences between those exposed to the risk factor and those not exposed. For example, if you are looking at the effect of diet on mild cognitive impairment, you would be aware that the main risk factor for cognitive impairment is age. Diet also varies with age. Age, then, is a factor which is associated with both the disease and the risk factor. These factors are called **confounding factors**. Confounding factors can distort the true relationship between a risk factor and a disease unless we take them into account in the design or the analysis of our study.

We can deal with the presence of confounding variables in the design of our study by matching the cases and controls on key confounders. In matched case-control designs, healthy controls are matched to cases using one or more variables. **In practice, the most efficient matching strategy is to match on at most two variables.** Matching on many variables makes it very difficult to locate and recruit controls. And although matching on many variables is intuitively attractive, it doesn't actually increase statistical efficiency – in fact, matching on more than three variables actually reduces the power of your study to detect risk factor relationships. Altman recommends that "in a large study with many variables it is easier to take an unmatched control group and adjust in the analysis for the variables on which we would have matched, using ordinary regression methods. Matching is particularly useful in small studies, where we might not have sufficient subjects to adjust for several variables at once." (Bland & Altman, 1994).

Matching cases and controls will produce a correlation between the probability of exposure within each case-control pair. This increases the statistical power of the study. The sample size will depend on the degree of correlation between the cases and controls. This is rarely possible to estimate, so these calculations are based on a case-control correlation of phi=0·2. This is the recommended action where the correlation is unknown (Dupont 1980).

It is important to note that when you analyse a matched case-control study, you must incorporate the matching into the analysis using procedures like conditional logistic regression. Analysing it as an unmatched case-control study

biases the estimates of the risk factor effects in the directly of 1. In other words, calculated risk factor effects will be smaller than they really are.

Sample size calculation

1. What is the prevalence of the risk factor in the controls? The tables give possibilities of 10%, 20% 25%, 50% and 75%. If in doubt, select the estimate furthest from 50%. For example, if you think that the prevalence is somewhere between 10% and 20%, estimate sample size based on a 10% prevalence.

2. What is the smallest odds ratio that would be of real life importance (clinically significant)?

3. Look up the sample size for 90% and 95% power in the table.

	Smallest odds ratio that would be clinically significant					
	1.5	2	2.5	3	4	5
Prevalence of the risk factor	90% Power to detect the odds ratio					
10%	1501	454	236	152	86	59
20%	885	279	150	100	59	43
30%	705	230	128	87	54	40
40%	644	217	124	86	55	41
50%	644	223	130	92	59	45
60%	697	248	147	105	69	54
70%	827	301	181	131	88	68
80%	1126	418	254	186	126	99
90%	2072	784	482	355	243	192
	95% Power to detect the odds ratio					
10%	1851	557	289	185	103	70
20%	1091	342	184	122	71	51
30%	869	283	156	106	65	47

40%	794	266	151	104	66	49
50%	794	274	158	111	71	54
60%	860	304	179	127	83	64
70%	1020	369	221	159	105	81
80%	1389	513	311	226	151	118
90%	2557	963	589	432	292	229

Table 1.4b Number of cases required for a matched case control study

Multiple controls per case

Where there are multiple controls per case, you can get greater statistical power. If you don't have enough cases, you could consider this strategy. Recruiting two controls per case will reduce your case sample size by roughly 25% for the same statistical power, and recruiting three controls per case will reduce it by roughly a third. However, the total size of your study will increase because of the extra controls.

Number of controls per case	Multiply the number of cases by
2	0.75
3	0.67
4	0.63
5	0.60

Table 1.4b1 Effect of multiple controls per case on sample size

Example

A researcher wishes to conduct a matched case-control study of the effect of regular alcohol consumption on risk of falls in older people. She estimates that 20% to 30% of the elderly population consume alcohol regularly. She decides that an odds ratio of 2.5 would be regarded as clinically significant.

She uses the lower estimate of prevalence – 20% – for sample size calculation. She will require 150 case-control pairs to achieve 90% power. This is a very large number of falls patients, and she will only have a maximum of 60 patients available to her, so she realises that she will only reach 90% power to detect an odds ratio of 4.

Recruiting 60 patients would take a long time, so she considers recruiting two controls for each patient, which would reduce the number of patients from 60 to 45, though increasing the number of controls to 90.

Should you use a matched design?

Matched designs seem to offer advantages in being able to control for confounding variables. However, there are two points to be considered. The first is that a matched design will under-estimate the strength of the risk factor effect if it is analysed without taking the matching into account, so it is important to use an appropriate statistical technique (such as conditional logistic regression).

More importantly, matching can make it hard to find controls. In many situations it is probably better to adjust for confounding variables statistically and use an unmatched case-control design.

However, there are two cases where matching can be beneficial:

1. There are strong, known risk factors that are not of interest. Variables like age, smoking, diabetes are well-studied and have strong effects on risk of many diseases. Matching on these variables can greatly increase study power (Stürmer 2000). However, one-to-one matching may be less efficient than frequency matching, and the paper by Stürmer *et al* is a useful read before you decide on a matching strategy.

2. Matching may be used to control for background variables that are hard to measure or are unknown. For example, in hospital studies time of admission may have a considerable effect on patient outcomes – patients admitted when the hospital is very busy may receive different treatment to those admitted when it is quiet. Matching cases and controls by time of admission can be used to control for these contextual variables.

References

Bland J M, Altman D G. Statistics notes: Matching BMJ 1994; 309 :1128

Stürmer, T., and H. Brenner. "Potential gain in precision and power by matching on strong risk factors in case-control studies: the example of laryngeal cancer." Journal of epidemiology and biostatistics 5.2 (2000): 125-131.

These calculations were carried out using the Stata command `sampsi_mcc`, written by Adrian Mander, of the MRC Human Nutrition Research, Cambridge, UK.

A typical command is

```
sampsi_mcc , p0(.1) power(.9) solve(n) alt(4) phi(.2) m(1)
```

which sets the prevalence at .1, the power at 90%, the hypothesised odds ratio (alternative odds ratio) at 4 and asks Stata to solve the problem for N, the sample size. The command also includes two options that are not actually needed, since they are the defaults: phi, the correlation between case-control pairs, is set at 0.2 and the matching (m) is set to one control per case.

The formulas are drawn from

Dupont W.D. (1988) Power calculations for matched case-control studies. Biometrics 44: 1157-1168.

1.5 Sample size for logistic regression with a continuous predictor variable

This section give guidelines for sample sizes for studies that measure the effect of a continuous predictor (for example, body mass index) on the risk of an endpoint (for example ankle injury). The data may come from a cross-sectional, case-control or cohort study.

Introduction

Logistic regression allows you to calculate the effect that a predictor variable has on the occurrence of an outcome. It can be used with cross-sectional data, case-control data and longitudinal (cohort) data. The effect of the predictor variable is measured by the odds ratio. A researcher may be interested, for example, on the effect that body weight has on the probability of a patient not having a complete clinical response to a standard 70mg dose of enteric aspirin, or the effect that depression scores have on the probability that the patient will not adhere to prescribed treatment.

Step 1: Variability : Estimate the mean and standard deviation of the predictor variable

You will probably be able to estimate the mean value quite easily. If you cannot find an estimate for the standard deviation, you can use the **rule of thumb** that the typical range of the variable is four standard deviations. By asking yourself what an unusually low and an unusually high value would be, you can work out the typical range. Dividing by four gives a rough standard deviation.

For example, adult weight averages at about 70 kilos, and weights under 50 or over 100 would be unusual, so the ‘typical range’ is about 50 kilos. This gives us a ‘guesstimate’ standard deviation of 12.5 kilos (50÷4).

Step 2: Baseline: What is the probability of the outcome at the average value of the predictor?

A good rule of thumb is that the probability of the outcome at the average value of the predictor is the same as the probability of the outcome in the whole sample. So if about 20% of patients have poor adherence to prescribed treatment, this will do as an estimate of the probability of poor adherence at the average value of the predictor.

Step 3: Effect size: what is the smallest increase in the probability of the outcome associated with an increase of one standard deviation of the predictor that would be clinically significant?

Clinical significance, or real-life significance, means that an effect is important enough to have real-life consequences. In the case of treatment failure with aspirin, if the probability of treatment failure increased from 10% at the

average weight to 25% one standard deviation higher, it would certainly be of clinical importance. Would an increase from 10% to 20% be clinically important? Probably. But any smaller increase probably would not. So in this case, we would set 10% and 20% as the prevalence at the mean and the smallest increase the be detected one standard deviation higher.

Step 4. Read off the required sample size from the table.

Table 1.5 Sample size for logistic regression

Prevalence at mean value	Prevalence 1 SD higher	Odds ratio	N for 90% power
5%	10%	2.1	333
10%	15%	1.6	484
10%	20%	2.3	172
20%	25%	1.3	734
20%	30%	1.7	220
20%	40%	2.7	98
20%	50%	4.0	143
25%	30%	1.3	825
25%	35%	1.6	238
25%	40%	2.0	128
25%	50%	3.0	93
30%	35%	1.3	889
30%	40%	1.6	249
30%	50%	2.3	93
30%	60%	3.5	106
40%	45%	1.2	933
40%	50%	1.5	250
40%	60%	2.3	87
40%	80%	6.0	499
50%	55%	1.2	865
50%	60%	1.5	225
50%	75%	3.0	81
50%	80%	4.0	133

Example

A researcher wishes to look at the effect of stigma on the risk of depression in medical patients. Previous research suggests that the prevalence of depression is about 20%. We can take this as the prevalence at the mean stigma score. The researcher wishes to be able to detect an increase in prevalence of 10% at one standard deviation above the mean value. She will need 172 patients to have a 90% chance of detecting a relationship this strong.

Reference:

These calculations were carried out using the `powerlog` command written for Stata by Philip B. Ender, UCLA Institut for Digital Research and Education.

The command is supported by an online tutorial at the IDRE website: http://www.ats.ucla.edu/stat/stata/dae/logit_power.htm

1.6 Sample sizes for logistic or Cox regression with multiple predictors

This section reviews guidelines on the number of cases required for studies in which logistic regression or Cox regression are used to measure the effects of risk factors on the occurrence of an endpoint. Earlier recommendations stated that you needed ten events (endpoints) per predictor variable. Subsequent work suggested that this isn't strictly true, and that 5–9 events per predictor may be yield estimates that are just as good. However, the jury is still out. The section includes guidelines on designing studies with multiple predictors. There isn't a table because the number of potential scenarios is impossibly big.

Introduction

Logistic regression builds a model the estimate the probability of an event occurring. We can use logistic regression where we have data in which each participant's status is known: the event of interest has either occurred or has not occurred. For example, we might be analysing a case-control study of stress fractures in athletes. Stress fractures are either present (in the cases) or absent (in the controls). We can use logistic regression to analyse the data. Or we might be analysing data on whether or not a patient is prescribed an antibiotic for symptoms of the common cold. In this case, we know whether each patient was or was not prescribed an antibiotic.

However, in follow-up studies, we often have data on people who *might* experience the event but they have not experienced it *yet*. For example, in a cancer follow-up study, some patients have experienced a recurrence of the disease, while others are still being followed up and are disease free. We cannot say that those who are disease free will not recur, but we know that their time to recurrence must be greater than their current follow-up time. This kind of data is called **censored data**. Censoring is where we do not know the exact value we are trying to measure, but we know that it has to be larger than (or smaller than) a particular value, or that it lies between two values (such as knowing that the person's age is between 45 and 54).

Where we want to predict the risk of the occurrence of an event, but we have some censored observations, we can use Cox regression (sometimes called a proportional hazards general linear model, which is what Cox himself called it. You can see why people refer to it as Cox regression!).

The ten events per predictor rule

There was a very influential paper published in the 1990s by Peduzzi et al (1996) based on simulation studies which concluded that for logistic regression you needed ten *events* (not patients) per predictor variable if you were calculating a multivariate model.

Example: a researcher wants to look at factors affecting the development of hypertension in first-time pregnancies. If the researcher has 5 explanatory variables, they will need to recruit a sample big enough to yield 50 cases of hypertension. Around 20% of first-time mothers will develop hypertension, so these 50 cases will be 20% of the required sample. So a total sample of 250 will be required so that there will be the required 50 cases

More recent research has cast doubt on this rule

More recently, bigger and more comprehensive simulation studies have cast doubt on this hard-and-fast rule. Vittinghoff and McCulloch (2007), in a very widely-cited paper, concluded that "*problems are fairly frequent with 2–4 events per predictor variable, uncommon with 5–9 events per predictor variable, and still observed with 10–16 events per predictor variable. Cox models appear to be slightly more susceptible than logistic. The worst instances of each problem were not severe with 5–9 events per predictor variable and usually comparable to those with 10–16 events per predictor variable*."

In other words, with between 5 and 9 events per predictor variable, their models performed more or less as well as models with 10-16 events per variable. As a safe minimum, then, it appears that there should be **at least 5 events per predictor variable**.

Since then, further simulation studies where prediction models are validated against new datasets tend to confirm that 10 events per variable is a minimum requirement (see Wynants 2015) for logistic regression. These studies are important because they are concerned with the **generalisability** of findings.

The importance of the number and type of predictor variables

The second factor that will influence sample size is the nature of the study. Where the predictor variables have low prevalence and you intend running a multivariable model with several predictors, then the number of events per variable required for Cox regression is of the order of 20. As you might imagine, increasing the number of predictor variables and decreasing their prevalence both require increases in the number of events per variable.

Sample size requirements

Based on current research, the sample should have at least 5 events per predictor variable ideally 10. Sample sizes will need to be larger than this if you are performing a multivariate analysis with predictor variables that have low prevalences. In this case, you may require up to 20 events per variable, and should probably read the paper by Ogundimu *et al*.

Correlated predictors – a potential source of problems

One consideration needs to be mentioned: correlations between your predictor variables. If your predictor variables are uncorrelated, the required sample size will be smaller than if they are correlated. And the stronger the correlation, the larger the required sample size. Courvoisier (2011) points out that the size of the effect associated with the predictor and the correlations between the predictors all affect the statistical power of a study. And Kocak and colleagues, using simulation studies, report that the problem is especially significant in small samples.

One solution to the problem is to design the analysis carefully.

1. Choose predictor variables based on theory, not availability. It is better to use a small set of predictors that test an interesting hypothesis than to have a large number of predictors that were chosen simply because the data were there.

2. Make sure that predictors don't overlap. If you put education and social class into a prediction model, they measure overlapping constructs. The well-educated tend to have higher social class. Does your hypothesis really state that the two constructs have different effects? Choose one good measure of each construct rather than having multiple overlapping measures.

Frequently asked questions

That's all very well but I have only 30 patients

That's health research. I worked on what was, at the time, one of the world's largest studies of a rare endocrine disorder. It has 16 patients. We are often faced with a lack of participants because we are dealing with rare problems or rare events. In such a case, we do what we can. What this section is warning is is that with rare conditions our statistical power is low. The only strategy in this case is the one outlined above: keep to a small, theoretically-justified set of predictors that have as little overlap as possible. And try and collaborate with other centres to pool data.

References

Courvoisier, D.S. et al., 2011. Performance of logistic regression modeling: beyond the number of events per variable, the role of data structure. Journal of Clinical Epidemiology, 64(9), pp.993–1000.

Kocak M, Onar-Thomas A. A Simulation-Based Evaluation of the Asymptotic Power Formulas for Cox Models in Small Sample Cases. The American Statistician. 2012 Aug 1;66(3):173-9.

Ogundimu EO, Altman DG, Collins GS. Adequate sample size for developing prediction models is not simply related to events per variable. Journal of Clinical Epidemiology. Elsevier Inc; 2016 Aug 1;76(C):175–82.

Peduzzi, P. et al., 1996. A simulation study of the number of events per variable in logistic regression analysis. Journal of Clinical Epidemiology, 49(12), pp.1373–1379.

Vittinghoff, E. & McCulloch, C.E., 2007. Relaxing the rule of ten events per variable in logistic and Cox regression. American Journal of Epidemiology, 165(6), pp.710–718.

Wynants L, Bouwmeester W, Moons KGM, Moerbeek M, Timmerman D, Van Huffel S, et al. A simulation study of sample size demonstrated the importance of the number of events per variable to develop prediction models in clustered data. Journal of Clinical Epidemiology. Elsevier Inc; 2015 Dec 1;68(12):1406–14.

2: Sample sizes and powers for comparing two means where the variable is measured on a continuous scale that is (more or less) normally distributed.

This section give guidelines for sample sizes for studies that measure the difference between the means of two groups, or that compare the means of the same group measured under two different conditions (often before and after an intervention).

2.1 Comparing the means of two groups

Studies frequently compare a group of interest with a control group or comparison group. If your study involved measuring something on the same people twice, once under each of two conditions, you need the next section.

Step 1: Effect size: decide on the difference that you want to be able to detect

The first step in calculating a sample size is to decide on the smallest difference between the two groups that would be 'clinically significant' or 'scientifically significant'. For example, a difference in birth weight of 250 grammes between babies whose mothers smoked and babies whose mothers did not smoke would be certainly regarded as clinically significant, as it represents the weight gain of a whole week of gestation. However, a smaller difference – say 75 grammes – probably would not be.

It is hard to define the smallest difference that would be clinically significant. An element of guesswork in involved. What is the smallest reduction in cholesterol that would be regarded as clinically worthwhile? It may be useful to search the literature and see what other investigators have done. And bear in mind that an expensive intervention will need to be associated with quite a large difference before it would be considered worthwhile.

NB: Effect size should not be based on your hopes or expectations!

Note, however, that the sample size depends on the smallest clinically significant difference, not on the size of the difference you expect to find. You may have high hopes, but your obligation as a researcher is to give your study enough power to detect the smallest difference that would be clinically significant.

Step 2: Convert the smallest clinically significant difference to standard deviation units.

Step 2.1. What is the expected mean value for the control or comparison group?

Step 2.2. What is the standard deviation of the control or comparison group?

How to get an approximate standard deviation

If you do not know this exactly, you can get a reasonable guess by identifying the highest and lowest values that would *typically* occur. Since most values will be within ±2 standard deviations of the average, then the highest typical value (2 standard deviations above average) and lowest typical value (2 below) will span a range of four standard deviations.

An approximate standard deviation is therefore

$$\text{Approximate SD} = \frac{\text{Highest typical value} - \text{Lowest typical value}}{4}$$

For example: a researcher is measuring fœtal heart rate, to see if mothers who smoke have babies with slower heart rates. A typical rate is 160 beats per minute, and normally the rate would not be below 135 or above 175. The variation in 'typical' heart rates is 175–135 = 30 beats. This is about 4 standard deviations, so the standard deviation is about 7.5 beats per minute. (This example is real, and the approximate standard deviation is pretty close to the real one!)

How to get an approximate standard deviation from a published confidence interval

Another potential source of standard deviation information is from published research. Although the paper may not include a standard deviation, it may include a confidence interval. The Cochrane Handbook has a useful formula for converting this to a standard deviation:

$$\text{Standard deviation} = \sqrt{N}\,\frac{\text{Upper CI limit} - \text{Lower CI limit}}{3{\cdot}92}$$

where N is the number of cases.

Step 3. What is the smallest difference between the two groups in the study that would be considered of scientific or clinical importance?

This is the minimum difference which should be detectable by the study. You will have to decide what is the smallest difference between the two groups that you are studying that would constitute a 'clinically significant difference' – that is, a difference that would have real-life implications.

In the case of the foetal heart rate example, a researcher might decide that a difference of 5 beats per minute would be clinically significant.

Note again that the study should be designed to have a reasonable chance of detecting the minimum clinically significant difference, and not the difference that you think is actually there.

Step 4. Convert the minimum difference to be detected to standard deviation units by dividing it by the standard deviation

$$\frac{\text{Minimum difference to be detected}}{\text{Standard deviation}}$$

Following our example, the minimum difference is 5 beats, and the standard deviation is 7.5 beats. The difference to be detected is therefore two thirds of a standard deviation (0.67)

Step 5: Use table 2.1 to get an idea of the number of participants you need in each group to detect a difference of this size.

Following the example, the nearest value in the table to 0.67 is 0.7. The researcher will need two groups of 43 babies each to have a 90% chance of detecting a difference of 5 beats per minute between smoking and non-smoking mothers' babies. To have a 95% chance of detecting this difference, the researcher will need 54 babies in each group.

Table 2.1 Sample size for comparing the means of two groups

Difference to be detected (SD units)	N in each group 90% power*	N in each group 95% power	Chance that someone in group 1 will score higher than someone in group 2
2	7	8	92%
1.5	11	13	86%
1.4	12	15	84%
1.3	14	17	82%
1.25	15	18	81%
1.2	16	20	80%
1.1	19	23	78%
1	23	27	76%
0.9	27	34	74%
0.8	34	42	71%
0.75	39	48	70%
0.7	44	55	69%
0.6	60	74	66%
0.5	86	105	64%
0.4	133	164	61%
0.3	235	290	58%
0.25	338	417	57%
0.2	527	651	55%

Sample Stata code for the first entries in this column:

```
power twomeans 0 (2 1.5 1.4 1.3 1.25 1.2 1.1 1), power(0.9)
```

If you intend using the Wilcoxon Mann-Whitney test,
multiply the sample size by 1.16

Subgroup analysis : the effect of prevalence

Researchers often have to compare a subgroup of the sample with the remainder. For example, they may be interested in comparing folate intake in vegetarians and non-vegetarians using data from a population survey. Ideally, they would use a stratified random sample and recruit equal numbers of vegetarians and non-vegetarians. However is they are analysing a population survey, they will be faced with a situation in which the vegetarians are in a minority.

Table 2.11 shows the effect on sample size of decreasing prevalence of a subgroup.

Prevalence of the smaller subgroup	multiply total sample size by
33%	1.125
25%	1.33
20%	1.5
15%	1.95
10%	2.7

As you can see from the table, the effect of reducing prevalence only really begins to bite as the prevalence drops below 20%. Effectively, your sample size needs to be an eighth bigger to compare a subgroup of 33% with the rest, a third bigger if the prevalence is 25%, half as big again if the prevalence is 20%, twice as big if the prevalence is 15% and three times as big if the prevalence is 10%. The multipliers were taken by averaging the effect over a range of effect sizes between 0·2 and 1 standard deviations.

Stata code that generated the results in that table

```
power twomeans 0 (.2(.2)1), power(0.9) nratio(1(1) 4 5.6666667 9 19)
```

The pieces of code that have a list of numbers in brackets cause Stata to make one calculation for each of the numbers. You can write a list of numbers as a list (for example `4 5.6666667 9 19`) or as a start number(an increment)an end number, as in `(.2(.2)1)`, which means 'start with .2, increment by .2 until you reach 1.

Frequently-asked questions

What is 90% or 95% power?

Just because a difference really exists in the population you are studying does not mean it will appear in every sample you take. Your sample may not show the difference, even though it is there. To be ethical and value for money, a research study should have a reasonable chance of detecting the smallest difference that would be of clinical significance (if this difference actually exists, of course). If you do a study and fail to find a difference, even though it exists, you may discourage further research, or delay the discovery of

something useful. For this reason, you study should have a reasonable chance of finding a difference, if such a difference exists.

A study with 90% power is 90% likely to discover the difference between the groups if such a difference exists. And 95% power increases this likelihood to 95%. So if a study with 95% power fails to detect a difference, the difference is unlikely to exist. You should aim for 95% power, and certainly accept nothing less than 90% power. Why run a study that has more than a 10% chance of failing to detect the very thing it is looking for?

How do I interpret the column that shows the chance that a person in one group will have a higher score than a person in another group?

Some scales have measuring units that are hard to imagine. We can imagine fœtal heart rate, which is in beats per minute, but how do you imagine scores on a depression scale? What constitutes a 'clinically significant' change in depression score?

One way of thinking of differences between groups is to ask what proportion of the people in one group have scores that are higher than average for the other group. For example we could ask what proportion of smoking mothers have babies with heart rates that are below the average for non-smoking mothers? Continuing the example, if we decide that a difference of 5 beats per minute is clinically significant (which corresponds to just about 0.7 SD), this means that there is a 69% chance that a non-smoking mother's baby will have a higher heart rate than a smoking mother's baby. (Of course, if there is no effect of smoking on heart rate, then the chances are 50% – a smoking mothers' baby is just as likely to have higher heart rate as a lower heart rate).

This information is useful for planning clinical trials. We might decide that a new treatment would be superior if 75% of the people would do better on it. (If it was just the same, then 50% of people would do better and 50% worse.) This means that the study needs to detect a difference of about 1 standard deviation (from the table). And the required size is two groups of 26 people for 95% power.

The technical name for this percentage, incidentally, is the Mann-Whitney statistic. You will also encounter it as the c statistic, Harrell’s c, and even as the area under the ROC curve.

I have a limited number of potential participants. How can I find out power for a particular sample size?

You may be limited to a particular sample size because of the limitations of your data. There may only be 20 patients available, or your project time scale only allows for collecting data on a certain number of participants. You can use the table to get a rough idea of the power of your study. For example, with only 20 participants in each group, you have more than 95% power to detect a difference of 1.25 standard deviations (which only needs two groups of 17) and

slightly less than 90% power to detect a difference of 1 standard deviation (you would really need 2 groups of 22).

But what if the difference between the groups is bigger than I think?

Sample sizes are calculated to detect the smallest clinically significant difference. If the difference is greater than this, the study's power to detect it is higher. For instance, a study of two groups of 43 babies has a 90% power to detect a difference of 0.7 standard deviations, which corresponded (roughly) to 5 beats per minute, the smallest clinically significant difference. If the real difference were bigger – say, 7.5 beats per minute (1 standard deviation) then the power of the study would actually be 99.6%. (This is just an example, and I had to calculate this power specifically; it's not in the table.) So if your study has adequate power to detect the smallest clinically significant difference, it has more than adequate power to detect bigger differences.

I intend using a Wilcoxon (Mann Whitney) test because I don't think my data will be normally distributed

The first important point is that the idea that the data should be normally distributed before using a t-test, or linear regression, *is a myth*. It is the *measurement errors* that need to be normally distributed. But even more important, studies with non-normal data have shown that the t-test is extremely robust to departures from normality (Fagerland, 2012; Fagerland, Sandvik, & Mowinckel, 2011; Rasch & Teuscher, 2007).

A second persistent misconception is that you cannot use the t-test on small samples (when pressed, people mutter something about “less than 30” but aren’t sure). Actually, you can. And the t-test performs well in samples as small as N=2! (J. de Winter, 2013) Indeed, with very small samples indeed, the Wilcoxon-Mann Whitney test is unable to detect a significant difference, while the t-test is (Altman & Bland, 2009).

Relative to a t-test or regression, the Wilcoxon test (also called the Wilcoxon Mann-Whitney U test) can be less efficient if your data are close to normally distributed. However, a statistician called Pitman showed that the test was never less than 86.4% as efficient. So inflating your sample by 1.16 should give you at least the same power that you would have using a t-test with normally distributed data. With data with skewed distributions, or data in which the distributions are different in the two groups, the Wilcoxon Mann-Whitney test can be more powerful than a t-test, so

My data are on 5-point Likert scales and my supervisor says I cannot use a t-test because my data are ordinal

Simulation studies comparing the t-test and the Wilcoxon Mann-Whitney test on items scored on 5-point scales have given heartening results. In most scenarios, the two tests had a similar power to detect differences between groups. The false-positive error rate for both tests was near to 5% for most situations, and

never higher than 8% in even the most extreme situations. However, when the samples differed markedly in the shape of their score distribution, the Wilcoxon Mann-Whitney test outperformed the t-test (J. C. de Winter & Dodou, 2010).

Methods in Stata and R

The sample sizes were calculated using Stata Release 14, using the `power` command. The Mann-Whitney statistic was calculated using the mwstati command for Stata written by Rich Goldstein, and based on formulas in Colditz et al (1988) above.

You can also use the package `pwr` in R. The R code for the fœtal heart rate example, where we want to detect a difference of 0.67 standard deviations is

```
> pwr.t.test(n=NULL, d=.67,power=.9,type="two.sample")

     Two-sample t test power calculation

              n = 47.79517
              d = 0.67
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number in *each* group
```

References and useful reading

These calculations were carried out using Stata release 12

Altman, D. G., & Bland, J. M. (2009). Parametric v non-parametric methods for data analysis. *Bmj*, *338*(apr02 1), a3167–a3167. doi:10.1136/bmj.a3167

Conroy, R. M. (2012). What hypotheses do “nonparametric” two-group tests actually test? *The Stata Journal*, *12*(2), 1–9.

Higgins JPT. Cochrane Handbook for Systematic Reviews of Interventions. The Cochrane Collaboration; 2011. Available from: www.handbook.cochrane.org.

de Winter, J. (2013). Using the Student’s t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation*, *18*(10), 1–12.

de Winter, J. C., & Dodou, D. (2010). Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, *15*(11), 1–12.

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies--a paradox of statistical practice? *BMC Medical Research Methodology*, *12*, 78. doi:10.1186/1471-2288-12-78

Fagerland, M. W., Sandvik, L., & Mowinckel, P. (2011). Parametric methods outperformed non-parametric methods in comparisons of discrete

numerical variables. *BMC Medical Research Methodology*, *11*(1), 44. doi:10.1186/1471-2288-11-44

Colditz, G. A., J. N. Miller, and F. Mosteller. (1988). Measuring Gain in the Evaluation of Medical Technology. International Journal of TechnologyAssessment. 4, 637-42.

Rasch, D., & TEUSCHER, F. (2007). How robust are tests for two independent samples? *Journal of Statistical Planning and Inference*, *137*(8), 2706–2720.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83.

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *The British Journal of Mathematical and Statistical Psychology*, *57*(Pt 1), 173–181. doi:10.1348/000711004849222

2.2 Sample sizes for comparing means in the same people under two conditions

One common experimental design is to measure the same thing twice, once under each of two conditions. This sort of data are often analysed with the *paired t-test*. However, the paired t-test doesn't actually use the two values you measured; it subtracts one from the other and gets the average difference. The null hypothesis is that this average difference is zero.

So the sample size for paired measurements doesn't involve specifying the means for each condition but specifying the mean difference.

Step 1: decide on the difference that you want to be able to detect.

The first step in calculating a sample size is to decide on the smallest difference between the two measurements that would be 'clinically significant' or 'scientifically significant'. For example, if you wanted to see how effective an exercise programme was in reducing weight in people who were overweight, you might decide that losing two kilos over the one-month trial period would be the minimum weight loss that would count as a 'significant' weight loss..

It is often hard to define the smallest difference that would be clinically significant. An element of guesswork in involved. What is the smallest reduction in cholesterol that would be regarded as clinically worthwhile? It may be useful to search the literature and see what other investigators have done.

Effect size should not be based on your expectations!

Note, however, that the sample size depends on the smallest clinically significant difference, not, on the size of the difference you expect to find.

Step 2: Convert the smallest clinically significant difference to standard deviation units.

Step 2.1. What is the standard deviation of the differences?

This is often very hard to ascertain. You may find some published data. Even if you cannot you can get a reasonable guess by identifying the biggest positive and biggest negative differences that would *typically* occur. The biggest positive difference is the biggest difference in the expected direction that would typically occur. The biggest negative difference is the biggest difference in the opposite direction that would be expected to occur. Since most values will be within ±2 standard deviations of the average, then the biggest positive difference (2 standard deviations above average) and biggest negative (2 below) will span a range of four standard deviations. An approximate standard deviation is therefore

$$\text{Approximate SD of differences} = \frac{\text{Biggest typical positive difference} - \text{Biggest typical negative difference}}{4}$$

For example: though we are hoping for at least a two kilo weight loss following exercise, some people may lose up to five kilos. However, others might actually gain as much as a kilo, perhaps because of the effect of exercise on appetite. So the change in weight can vary from plus five kilos to minus one, a range of six kilos. The standard deviation is a quarter of that range: one and a half kilos.

Step 2.2. Convert the minimum difference to be detected to standard deviation units by dividing it by the standard deviation

$$\frac{\text{Minimum difference to be detected}}{\text{Standard deviation of the difference}}$$

Following our example, the minimum difference is 2 kilos, and the standard deviation is 1.5 kilos. The difference to be detected is therefore one and a third standard deviations (1.33).

Step 3: Use table 2.2 to get an idea of the number of participants you need in each group to detect a difference of this size.

Following the example, the nearest value in the table to 1.33 is 1.3. The researcher will need to study seven people to have a 90% chance of detecting a weight loss of 2 kilos following the exercise programme. To have a 95% chance of detecting this difference, the researcher will need 8 people.

Table 2.2 Sample sizes for comparing means in the same people under two conditions

Difference to be detected (SD units)	N required for 90% power*	N required for 95% power	Percentage of people who will change in the hypothesised direction
2	5	6	98%
1.5	7	8	93%
1.4	8	9	92%
1.3	9	10	90%
1.25	9	11	89%
1.2	10	12	88%
1.1	11	13	86%
1	13	16	84%
0.9	16	19	82%
0.8	19	23	79%
0.75	21	26	77%
0.7	24	29	76%
0.6	32	39	73%
0.5	44	54	69%
0.4	68	84	66%
0.3	119	147	62%
0.25	171	210	60%
0.2	265	327	58%

Sample sizes for studies which compare mean values on the same people measured under two different conditions

*Stata code for this column:

```
power pairedmeans, sddiff(1) altdiff( 2 1.5 1.4 1.3 1.25 1.2 1.1 1
      0.9 0.8 0.75 0.7 0.6 0.5 0.4 0.3 0.25 0.2) power(0.9)
```

Note that the Stata code includes a list of values for the alternative hypothesis difference. Note also that you can run this command from the Stata menus and dialogues.

Frequently-asked questions

What is 90% or 95% power?

Just because a difference really exists in the population you are studying does not mean it will appear in every sample you take. Your sample may not show the difference, even though it is there. To be ethical and value for money, a research study should have a reasonable chance of detecting the smallest difference that would be of clinical significance (if this difference actually exists, of course). If you do a study and fail to find a difference, even though it exists, you may discourage further research, or delay the discovery of something useful. For this reason, you study should have a reasonable chance of finding a difference, if such a difference exists.

A study with 90% power is 90% likely to discover the difference between the two measurement conditions if such a difference exists. And 95% power increases this likelihood to 95%. So if a study with 95% power fails to detect a difference, the difference is unlikely to exist. You should aim for 95% power, and certainly accept nothing less than 90% power. Why run a study that has more than a 10% chance of failing to detect the very thing it is looking for?

How do I interpret the column that shows the percentage of people who will change in the hypothesised direction?

Some scales have measuring units that are hard to imagine. We can imagine foetal heart rate, which is in beats per minute, but how do you imagine scores on a depression scale? What constitutes a 'clinically significant' change in depression score?

One way of thinking of differences between groups is to ask what proportion of the people will change in the hypothesised direction. For example we could ask what proportion of depressed patients on an exercise programme would have to show improved mood scores before we would consider making the programme a regular feature of the management of depression. If we decide that a we would like to see improvements in at least 75% of patients, then depression scores have to fall by 0.7 standard deviation units. The sample size we need is 22 patients for 90% power, 27 for 95% power (the table doesn't give 75%, I've used the column for 76%, which is close enough).

The technical name for this percentage, incidentally, is the Mann-Whitney statistic.

I have a limited number of potential participants. How can I find out power for a particular sample size?

You may be limited to a particular sample size because of the limitations of your data. There may only be 20 patients available, or your project time scale only allows for collecting data on a certain number of participants. You can use the table to get a rough idea of the power of your study. For example, with only 20 participants, you have more than 90% power to detect a difference of 0.75

standard deviations (which only needs two groups of 17) and slightly less than 95% power to detect a difference of 0.8 standard deviations (you would really need 21 participants).

But what if the difference is bigger than I think?

Sample sizes are calculated to detect the smallest clinically significant difference. If the actual difference is greater than this, the study's power to detect it is higher.

Reference and methods

These calculations were carried out using Stata release 15 with the `power` command

You can also use the `pwr` package in R. Here is the calculation for a difference of 0.5 standard deviations with 90% power.

```
 pwr.t.test(n=NULL, d=.5,power=.9,type="paired")

     Paired t test power calculation

              n = 43.99548
              d = 0.5
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: n is number of *pairs*
```

2.3 Calculating sample sizes for comparing two means: a rule of thumb

Sample size for comparing two groups

Gerald van Belle gives a good rule of thumb for calculating sample size for comparing two groups. You do it like this:

1. Calculate the smallest difference between the two groups that would be of scientific interest.

2. Divide this by the standard deviation to convert it to standard deviation units (this is the same two steps as before)

3. Square the difference

4. For 90% power to detect this difference in studies comparing two groups, the number you need in each group will be

$$\frac{21}{(\text{Difference})^2}$$

Round up the answer to the nearest whole number.

5. For 95% power, change the number above the line to 26.

Despite being an approximation, this formula is very accurate.

Studies comparing one mean with a known value

If you are only collecting one sample and comparing their mean to a known population value, you may also use the formula above. In this case, the formula for 90% power is

$$\frac{11}{(\text{Difference})^2}$$

Round up the answer to the nearest whole number.

For 95% power, replace the number 11 above the line by 13.

See the links page at the end of this guide for the source of these rules of thumb.

3.1 Sample size for correlations or regressions between two variables measured on a numeric scale

This section give guidelines for sample sizes for studies that measure the relationship between two numeric variables. Although these sample sizes are often based on correlations, they can also be applied to linear regression, and both types of measure are shown in the table.

Introduction : correlation and regression

Correlations are not widely used in medicine, because they are hard to interpret. One interpretation of a Pearson correlation (r) can be got by squaring it: this gives the proportion of variation in one variable that is linked to variation in another variable. For example, there is a correlation of 0.7 between illness-related stigma and depression, which means that just about half the variation in depression (0.49, which is 0.7^2) is linked to variation in illness-related stigma.

Despite the fact that correlations are measured on a continuous scale, researchers often try to make them more interpretable by converting them to what have been called, unkindly, tee-shirt sizes : small, medium and large. They do this referencing the work of Cohen (1992) , who recommended Pearson r values of 0.10, 0.30, and 0.50 to demarcate small, medium, and large effects, respectively. However, recent research has cast doubt on this classification. A major review of studies by Gignac and Szodorai (2020) based on 708 meta-analytically derived correlations, reported that the 25th, 50th, and 75th percentiles corresponded to correlations of 0.11, 0.19, and 0.29, respectively. Fewer than 3% of correlations met Cohen's definition of "large". They suggest that in real life, the terms small, medium and large more closely correspond to correlations of 0.10, 0.20 and 0.30.

Regressions are much more widely used, since they allow us to express the relationship between two variables in natural units – for example, the effect of a one-year increase in age on blood pressure. Because regressions are calculated in natural units, people often cite the proportion of variation shared between the two variables.

In fact, correlation is just an alternative form of reporting the results of a regression, so the p-value for a regression will be the same as the p-value for a Pearson correlation.

Steps in calculating sample size for correlation or regression

Step 1: How much variation in one variable should be linked to variation in the other variable for the relationship to be clinically important?

This is hard to decide, but it is hard to imagine a correlation being of 'real life' importance if less than 20% of the variation in one variable is linked to variation in the other variable.

Step 2: Use the table to look up the corresponding correlation and sample size

% Shared variation	Correlation	Sample size 90% power*	Sample size 95% power
10%	0.32	99	122
15%	0.39	65	80
20%	0.45	48	59
25%	0.5	38	47
30%	0.55	31	37
35%	0.59	26	32
40%	0.63	23	27
45%	0.67	19	23
50%	0.71	17	20

Table 3.1

Sample sizes for Pearson's r correlation and for simple regression based on percentage of shared variance

*Stata command for this column:

```
power onecorrelation 0 (0.32 0.39 0.45 0.5 0.55 0.59 0.63 0.67
0.71), power(0.9)
```

Reference

These calculations were carried out in Stata 15 with the command `power`

Cohen J. A power primer. Psychological bulletin. 1992 Jul;112(1):155.

Gignac, G., Szodorai, E. (2016). Effect size guidelines for individual differences researchers Personality and Individual Differences 102(Personality and Individual Differences5472013), 74-78. https://dx.doi.org/10.1016/j.paid.2016.06.069

3.2 Sample size for correlations based on the desired confidence interval

Introduction

Another approach to sample size for correlations is to decide on the degree of precision you want to measure the correlation with. Do you want to measure it to a margin of error of ±0·2? ±0·05? The table gives required sample sizes. It is also probably useful for determining the approximate power of your study if you know the likely sample size.

You can use this method to set a maximum confidence interval for your study, but you can also use it as a test of whether the correlation is bigger (or smaller) than a specified value.

Steps in calculating sample size for correlation based on confidence interval width

Step 1 : Decide what size correlation you want to be able to detect

You need to decide on the size of correlation that your study should be capable of detecting. This should be the smallest correlation that has real-life or clinical significance. You can base this on the percentage of variance that the variables share, in which case you should use table 3.2.1 for your sample size estiate, or you can base it on the size of the correlation, in which case you can use table 3.2.2.

Step 2a : Decide on the maximum width of the confidence interval

This is the widest margin of uncertainty that you would be prepared to accept. You can do this based on the absolute margin of error you would like – you might decide you would like correlations measured with a confidence interval of at most ±0·1.

Step 3b : Alternatively, decide on what is an unacceptably low correlation

There is a second way you might set the width of the confidence interval is by deciding what would be an unacceptably low correlation. An unacceptably low correlation is the largest correlation that is still too small to have any real-life or clinical importance.

Subtract this correlation from the correlation you want to be able to detect. This gives you the maximum width of the confidence interval.

Step 3 : Use the tables to corresponding sample size

% Shared variation	Correlation	Margin of error ±0·2	Margin of error ±0·15	Margin of error ±0·1
10%	0.32	79	139	311
15%	0.39	71	125	278
20%	0.45	64	111	247
25%	0.5	57	99	219
30%	0.55	50	86	190
35%	0.59	44	76	167
40%	0.63	39	66	144
45%	0.67	33	56	121
50%	0.71	28	47	99

Table 3.2.1

Sample sizes for correlations based on % shared variation

% Shared variation	Correlation	Margin of error ±0·2	Margin of error ±0·15	Margin of error ±0·1
9%	0.3	81	143	320
16%	0.4	70	123	273
25%	0.5	57	99	219
36%	0.6	43	74	161
49%	0.7	30	49	105
64%	0.8	18	28	56
81%	0.9	8	10	16

Table 3.2.2

Sample sizes for correlations based on absolute size of correlation

Examples

Sample size based on confidence interval

A researcher wants to correlate oral health literacy with oral health quality of life. However, they are both measured by short questionnaires which will have a limited reliability (in other words, a significant percentage of the variation in scores will be error variation rather than being true variation in health literacy

or quality of life). For this reason, even a correlation of 0·4 (meaning that 16% of the variation in oral health quality of life is linked with variation in oral health literacy) would be of real life importance. She would ideally like to measure this with a precision of ±0·1.

From table 3.2.2, she will need 87 participants to achieve a 95% confidence interval of ±0·1.

Methods

These correlations were calculated in R, using the library `userfriendlyscience`.

Table 3.2.3 was produced from a single R command, after loading the package

```
library("userfriendlyscience")
pwr.confIntR(seq(.3, .9, by = .1), w=seq(.2, .05,b=-.05))
```

Reference

Moinester, M., & Gottfried, R. (2014). Sample size estimation for correlations with pre-specified confidence interval. The Quantitative Methods of Psychology, 10(2), 124-130.

3.3 Sample sizes for Kendall's tau-b correlation

Spearman's rho is often used as a correlation for data measured on ordinal scales. However, it has a major disadvantage : it's pretty much impossible to explain what it measures. You cannot translate a rho of, say, 0·7 into some interpretable measure of the size of the effect.

On the other hand, Kendall's tau-b has an actual interpretation – that is, you can convert it into a measure of effect size. The idea behind tau-b is simple and ingenious. If you were to take a pair of observations, what is the probability that they would show a positive correlation. That is, what is the probability that the observation with the higher value on one variable would also have the higher value on the other?

The correlation is calculated by first getting the proportion of pairs of observations showing a positive correlation – this will fall between zero and one. Multiplying this by two gives it a range of 0 to 2, and subtracting 1 changes the range so that it runs from —1 to +1.

Step 1 : decide on the effect size you want to measure

You can base a sample size for tau on this effect size. What proportion of pairs of observations should show a positive correlation?

Step 2 : decide on the precision you want

The precision of the correlation is measured by its 95% confidence interval. The table presents confidence intervals of ±0·2, ±0·15 and ±0·1. Bear in mind that a confidence interval of ±0·1 is 0·2 units wide. That's probably as imprecise as you would want to go.

Step 3 : locate the effect size and value of tau from the table

% pairs of observations positively correlated	Kau-b correlation	N for confidence interval width **0·2**	N for confidence interval width **0·15**	N for confidence interval width **0·1**
65%	0.30	143	251	560
68%	0.35	133	234	521
70%	0.40	122	214	478
73%	0.45	111	194	431
75%	0.50	99	172	382
78%	0.55	86	150	331
80%	0.60	73	127	280
83%	0.65	61	104	229
85%	0.70	49	83	180
88%	0.75	37	62	134

Table 3.3

Sample size required for a confidence interval of a given size around a measured Kendall's tau-b correlation

Methods

The calculations were made with the presize package for R. The reference is

Armando Lenz and Alan G. Haynes and Andreas Limacher. presize: Precision Based Sample Size Calculation. R package version 0.2.3, 2021. https://CRAN.R-project.org/package=presize

The right hand column of the table can be calculated with these R commands:

```
library(presize)
prec_cor(seq(.3, .75, by=.05),conf.width = .1, method="kendall")
```

4. Sample size for reliability studies

This section give guidelines for sample sizes for studies that measure Cronbach's alpha, an index of the reliability – strictly speaking the internal consistency – of a set of items designed to measure a trait. The topic of scale development is a complex one, so the section gives guidance on the methodology of analysis and the interpretation of alpha.

Introduction : An apology

I wish there were a simple answer to this problem, and there isn't. Please read the following carefully.

Cronbach's alpha

The reliability of a measurement scale is the degree to which all the items measure the same thing. Reliability is specific: it describes the performance of a scale in a specific population tested under specific conditions. So it is important to make sure that scales are reliable when used in realistic conditions with realistic participants.

In developing a new measurement scale, or showing that a measurement scale works in a new setting, it is useful to measure its reliability. Reliability is usually measured using Cronbach's alpha coefficient, which is scaled between zero and one, with zero meaning that the items in the scale have nothing in common and one meaning that they are all perfectly correlated. In practice, it is wildly unlikely that anyone would develop a scale in which all the items were unrelated, so there is no point in testing whether your reliability is greater than zero. Instead, you have to specify a minimum value for the reliability coefficient.

Myths about Cronbach's alpha

A mythology has grown up around the interpretation of Cronbach's alpha, based, apparently, on the published work of Nunally (1978). According to this myth, Nunally advocated an alpha of 0·7 as indicating a scale that was acceptable for use in research. In fact, it's worth quoting Nunally's paper, which offers a much more nuanced and thoughtful approach to the question:

"What a satisfactory level of reliability is depends on how a measure is being used. In the early stages of research … one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of .70 or higher will suffice… In contrast to the standards in basic research, in many applied settings *a reliability of .80 is not nearly high enough*. In basic research, the concern is with the size of correlations and with the

differences in means for different experimental treatments, for which purposes a reliability of .80 for the different measures is adequate."

"In many applied problems, a great deal hinges on the exact score made by a person on a test… In such instances it is frightening to think that any measurement error is permitted. Even with a reliability of .90, the standard error of measurement is almost one-third as large as the standard deviation of the test scores. In those applied settings where important decisions are made with respect to specific test scores, a reliability of .90 is the minimum that should be tolerated, and a reliability of .95 should be considered the desirable standard."

This extensive quotation is from Lance, C.E., Butts, M.M. & Michels, L.C., 2006. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? Organizational Research Methods, 9(2), pp.202–220.

So bear in mind that mindlessly setting a desired alpha of 0·7 and citing Nunally's original paper is wrong. He didn't say anything like that. And, second, that you need to consider carefully the context of your research in setting a minimum alpha.

Alpha only applies to unidimensional scales

One of the statistical assumptions underlying alpha is that the scale is unidimensional. That is to say, that all the items measure the same thing, and that their failure to correlate perfectly is due to measurement error. So an important part of scale development is making sure that your items are indeed unidimensional.

How many cases should a reliability study have?

The standard advice is to have at least 10 participants per item on your scale. However, this should be regarded as the bare minimum.

There are surprising differences of opinion in the literature, however, on how small your sample can be. The best current advice is based on simulation studies where authors have studied the power of samples of various sizes to detect a given alpha.

Simulation studies indicate that sample size depends on the structure of your scale. Sample sizes as small as 30 can measure alpha reliably so long as the scale items have strong inter-correlations.

First step : principal components analysis

Your analysis should begin with a principal components analysis. A principal components analysis identifies underlying 'dimensions' that account for the variation in a set of items. In the case of reliability, you should only examine the first principal component. There is a good reason for this: alpha has no

interpretation when scales combine items that measure different constructs. The first principal component measures the degree to which the items measure the same construct.

Samuels (2015), summarising the literature, makes these recommendations

1. Don’t run reliability analysis with less than 30 participants

2. If you have between 30 and 50 participants, remove items that have loadings of less than 0·4 on the first principal component. This means that that very little of the variation in the responses to that item are shared with the other scale items.

3. Rerun the principal components analysis and examine the first eigenvalue (the eigenvalue for the first principal component). If this is less than 6, do not attempt a reliability analysis; the items just don’t show enough homogeneity to yield a reliable estimate of alpha.

4. Ideally, scale items should have a loading of 0·8 or more on the first principal component. Items between 0·4 and 0·8 need to be considered carefully as candidates for inclusion.

5. If your sample size is between 50 and 100, then follow the same steps, but if your eigenvalue falls between 3 and 6, then only perform a reliability analysis if the sample size is at least 75. See Yurdugül (2008) for details of how these figures are arrived at.

References

Lance, C.E., Butts, M.M. & Michels, L.C., 2006. The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? Organizational Research Methods, 9(2), pp.202–220.

Samuels, P., 2015. Statistical Methods – Scale reliability analysis with small samples, Birmingham City University, Centre for Academic Success. DOI: 10.13140/RG.2.1.1495.5364. https://www.researchgate.net/publication/280936182_Advice_on_Reliability_Analysis_with_Small_Samples

Yurdugül, H., 2008. Minimum sample size for Cronbach's coefficient alpha : a Monte-Carlo study. Hacettepe University Journal of Education, 35, pp.397–405. http://www.efdergi.hacettepe.edu.tr/200835HALİL%20YURDUGÜL.pdf

5. Sample size calculation for agreement between two raters using a present/absent rating scale using Cohen's Kappa

This section give guidelines for sample sizes for studies that use the kappa coefficient to measure the agreement between two raters who make ratings of present/absent.

Introduction

Studies looking at the agreement between raters come in many shapes and sizes. The most basic design is where two raters are asked to rate the presence or absence of a particular feature or quality. Kappa is a statistic that measures the degree of agreement over and above the agreement you would expect by chance. You can see why just measuring percentage agreement is not enough. If you toss two coins, they will agree 50% of the time just by chance. Likewise, two raters, each of whom rates a feature as present 50% of the time will agree with each other by chance 50% of the time.

When we are studying agreement, we have to choose a null hypothesis. Normally, the null hypothesis says that the data arose by chance – that there is no actual relationship between the variables we are studying. However, this makes no sense at all when we are studying agreement. It would be ridiculous to set up a scientific study to determine whether the agreement between two pathologists was better than chance! When two raters rate the same thing, it would be unusual to find that they didn't agree any more than you would expect by chance, even in psychiatry.

So in studies of agreement, we have to set a minimum level of agreement that we want to outrule in our study. Usually we would like to outrule a level of agreement that would suggest that there was a significant problem with the reliability of the rating. So unlike other sample size methods, the researcher will have to base sample size calculation for kappa on two figures: the value of kappa to be outruled and the likely true value of kappa. In addition, the prevalence of the feature will affect sample size.

Estimating sample size for kappa

The sample size will depend on three factors:

Step 1: Prevalence of the feature

What is the approximate **prevalence** of the feature that is being rated? Sample sizes will be smallest when there is a 50% prevalence, and will get very large when the prevalence drops much below 25%.

In the calculations below, we assume that there is no systematic difference between the raters. In other words, that each rater gives more or less the same prevalence of the feature. Where you suspect that raters will give different prevalences, the sample size calculation needs to take this into account, and is well beyond the scope of this guide. However, the R package I used will perform the calculation (see below).

Step 2 : Definition of an unacceptably low level of agreement (null value)

It would be astonishing if two raters could not agree any more than you would expect by chance. So in designing the study we have to stipulate what would be an unacceptably low level of agreement. This will act as a baseline against which we can assess the actual level of agreement. Because this is the level of agreement that we wish to outrule, the value is often called the null value, or null hypothesis value.

In practice, a kappa of 0.2–0.40 is regarded as a fair level of agreement, 0.41–0.60 as moderate, 0.61–0.80 as substantial and anything above 0·8 as excellent. That said, these cutpoints have a sort of folkloric status, and the interpretation of kappa is probably best done in the context of the decision that it supports.

In the tables that follow I will tabulate sample sizes for kappa in cases where you want to demonstrate that kappa is better than 0·4 (so agreement is better than ‘fair’), better than 0·5 or 0·6 (better than ‘moderate’) and better than 0·7 and 0·8 (better than ‘substantial).

Step 3 : Effect size - what is a clinically acceptable level of agreement?

What is the level of agreement that you think should be present if the test is a reliable test? This value is often called the alternative value or alternative hypothesis value, in contrast with the null value.

For example, if the test would require substantial agreement between assessors rather than simply being moderate, then you might set up your sample size to detect a kappa of 0·75 against a null hypothesis that kappa is 0·6. This would require 199 ratings made by the two raters to achieve 90% power. However, if you hypothesised that kappa was 0·75, as before, but wanted to outrule a kappa of 0·5, the required sample size drops to a very manageable 78.

Sample sizes for kappa for two raters

Prevalence of feature	Hypothesised kappa	Kappa to be outruled (null hypothesis kappa)	90% power	95% power
0·5	0.6	0.4	156	200
	0.7	0.5	131	169
	0.8	0.6	102	133
	0.7	0.45	87	112
	0.8	0.55	68	90
	0.8	0.5	49	65
0·4 or 0·6	0.6	0.4	162	208
	0.7	0.5	137	177
	0.8	0.6	106	139
	0.7	0.45	90	117
	0.8	0.55	71	94
	0.8	0.5	51	68
0·25 or 0·75	0.6	0.4	207	265
	0.7	0.5	176	227
	0.8	0.6	137	180
	0.7	0.45	116	150
	0.8	0.55	92	121
	0.8	0.5	66	87
0·1 to 0·9	0.6	0.4	427	546
	0.7	0.5	371	479

Prevalence of feature	Hypothesised kappa	Kappa to be outruled (null hypothesis kappa)	90% power	95% power
	0.8	0.6	292	382
	0.7	0.45	242	313
	0.8	0.55	194	255
	0.8	0.5	139	183

Example

A researcher wishes to study the agreement between family doctors on whether or not to prescribe an antibiotic for uncomplicated rhinitis. The prevalence of antibiotic prescribing is about 25%. She would like to show that the kappa value for agreement is better than 0·5. She hypothesises that the true kappa might be between 0·7 and 0·8.

Looking at the table, if the true kappa is 0·7, she will need to compare the doctors' ratings for 176 patients to have a 90% power to outrule a kappa as low as 0·5. On the other hand, if the true kappa is 0·75, she would have 90% power to outrule a kappa as low as 0·45 with a sample of 116.

Limitations of these tables

There are so many potential combinations of prevalence, kappa-to-be-outruled and hypothesised kappa that these tables can only give an approximate idea of the numbers involved. And they don't cover cases where the two raters have different prevalences (which would indicate systematic disagreement!), or where there are more than two raters *etc*. To get precise calculations for a wide variety of scenarios, I recommend using the R package `irr`.

Reference

These sample sizes were calculated with the `N.cohen.kappa` command in the `irr` package in R. The command uses a formula published in

Cantor, A. B. (1996) Sample-size calculation for Cohen's kappa. Psychological Methods, 1, 150-153.

The sample sizes in the table were produced using variations on this command:

```
N.cohen.kappa(0.1, 0.1, 0.5, 0.8,power=.95)
```

Sample size for intraclass correlations

This section give guidelines for sample sizes for studies that use the intraclass correlation coefficient to measure the agreement between two or more raters who make ratings on a continuous scale.

First, you should note that to measure agreement you do not calculate a standard correlation. Think about it : if you measure distance in miles and kilometres the readings will not agree unless the distance is zero. Nevertheless they will correlate perfectly. The correlation that you need is called an intraclass correlation.

There are, confusingly, three types of intraclass correlation and each of them comes in two separate forms that measure absolute and relative agreement. A discussion of what each one measures is beyond a sample size guide. I take it that you know which one you want, and simply want a sample size.

The table presents sample sizes for intraclass correlations based on an approximate formula published by Bonett (see references)

They are calculated in R, using the presize package.

Step 1 How big do you expect the intraclass correlation to be?

The smaller the intraclass correlation you expect, the larger the sample size required to detect it. In practice, worthwhile intraclass correlations are going to be bigger than 0·6.

Step 2 How much precision do you need in measuring it?

Your study will estimate the intraclass correlation and its confidence interval. What is the maximum size of the confidence interval that is allowable? In other words, how big does the confidence interval have to be before the study is too imprecise to be of any practical use?

High-precision studies (with a confidence interval of, say, ±0·05 around the intraclass correlation) require very large numbers of participants. In the table, confidence intervals of ±0·1 and ±0·15 are shown. Confidence intervals bigger than ±1·5 are probably too imprecise.

Step 3 How many raters will you use?

This is a design question. Are you specifically interested in two particular raters, or do you want to take a random sample of raters from the population. Sample size will go down with larger numbers of raters, and study

generalisability will go up because you have a bigger sample from the pool of potential raters.

Number of raters	Hypothesised intraclass correlation	Precision ±0·1	Precision ±0·15
2	**0.65**	514	229
	0.70	405	183
	0.75	299	136
	0.80	205	94
	0.85	124	58
	0.90	61	31
4	**0.65**	275	123
	0.70	223	100
	0.75	171	77
	0.80	120	54
	0.85	74	34
	0.90	37	17
10	**0.65**	198	89
	0.70	165	74
	0.75	130	58
	0.80	93	42
	0.85	59	27
	0.90	30	14

R code used to produce this table :

```
library(presize)
rh <- c(.65,.7,.75,.8,.85,.9)
prec_icc(rh,2, conf.width = .1)
prec_icc(rh,2, conf.width = .15)
prec_icc(rh,4, conf.width = .1)
prec_icc(rh,4, conf.width = .15)
prec_icc(rh,10, conf.width = .1)
prec_icc(rh,10, conf.width = .15)
```

References

Haynes AG, Lenz A, Stalder O, Limacher A. presize: An R-package for precision-based sample size calculation in clinical research. Journal of Open Source Software. 2021 Apr 14;6(60):3118.

Bonett DG (2002). A Sample size requirements for estimating intraclass correlations with desired precision. Statistics in Medicine, 21:1331-1335.

6. Sample size for pilot studies

Introduction

The sample size methods used so far presuppose that the investigator has some kind of knowledge that can be used to make informed guesses about such things as prevalences, effect sizes *etc*. However, by their very essence pilot studies are carried out when the researcher is facing the unknown. Even so, there are some general principles which can be applied to ensure that enough data are captured by a pilot study to inform subsequent study design with the smallest use of resources.

Sample size: the law of diminishing returns

Sample size for pilot studies starts with the observation that each participant that you recruit into a study yields less information than the last one. This law of diminishing returns can be used to define a point beyond which recruiting additional participant will yield minimal improvement in estimating effects. Calculations by Julious (2005) and Van Belle (2008) both show that in studies that compare the means of two groups, if you carry on recruitment beyond a sample size of 12 per group the effect of each additional participant on the precision is minimal. If your pilot study is purely exploratory and your aim is to get a preliminary estimate of the difference between two groups, then a sample size of 12 per group can be justified on the basis of these references.

Sample size to justify carrying out a full study

Sometimes there are cases when the investigator will have a preliminary estimate of the minimum difference between groups that would constitute a clinically significant difference. The purpose of the pilot study is to justify carrying out a full study. For example, before conducting a study of the effects of a physiotherapy programme on balance in the elderly, the investigators might be required to do a pilot to show that there were grounds for believing that such a programme would produce a clinically significant improvement in balance.

Cocks *et al* (2013) provide an algorithm for estimating the size of a pilot study that will give the ‘go-ahead’ to a main study. Their rule of thumb, based on calculated sample sizes for various scenarios, is to recruit 9% of of the projected final sample, or 20 participants, whichever is the greater, as a pilot. If there is no difference between the groups, then it is unlikely that the true effect size is as large as the one specified by the investigators. Note that this conclusion is based on an 80% confidence interval, not the usual 95%. If you are using this method, please read Cocks’ paper for further detail and worked examples.

Method

Calculate the sample size from section 2.1.

Use 9% of this sample size or 20 participants, whichever is the greater

If, when you analyse the pilot study, there is no significant difference between the groups, it is unlikely that the effect size reaches clinical significance.

References

Cocks, K. & Torgerson, D.J., 2013. Sample size calculations for pilot randomized trials: a confidence interval approach. Journal of Clinical Epidemiology, 66(2), pp.197–201.

Julious, S.A., 2005. Sample size of 12 per group rule of thumb for a pilot study. Pharmaceutical Statistics, 4(4), pp.287–291. Available at: http://onlinelibrary.wiley.com/doi/10.1002/pst.185/abstract.

van Belle, G., 2008. Sample Size. In Statistical Rules of Thumb. Wiley, Chichester. pp. 27–51. Download from http://vanbelle.org/chapters/webchapter2.pdf

7. Sample size for animal experiments in which not enough is known to calculate statistical power

In animal experiments, the investigator may have no prior literature to turn to. The potential means and standard deviations of the outcomes are unknown, and there is no reasonable way of guessing them. In a case like this, sample size calculations cannot be applied.

The resource equation method

The resource equation method can be used for minimising the number of animals committed to an exploratory study. It is based on the law of diminishing returns: each additional animal committed to a study tells us less than the one to reach the threshold where adding further animals will be uninformative. It should only be used for pilot studies or proof-of-concept studies.

Applying the resource equation method

1. How many treatment groups will be involved? Call this T.

2. Will the experiment be run in blocks? If so, how many blocks will be used? Call this B

A block is a batch of animals that are tested at the same time. Each block may have a different response because of the particular conditions at the time they were tested. Incorporating this information into a statistical analysis will increase statistical power by removing variability between experimental conditions on different days.

3. Will the results be adjusted for any covariates? If so, how many? Call this C

Covariates are variables that are measured on a continuous scale, such as the weight of the animal or the initial size of the tumour. Results can be adjusted for such variables, which increases statistical power.

4. Combine these three figures:

(T–1) + (B+C–1) = D

5. Add at least 10 and at most 20

The sample size should be at least (D+10) and at most (D+20).

Example of the resource equation method

An investigator wishes to examine the effect of a new delivery vehicle for an anti-inflammatory drug. The experiment will involve four treatments: a control, a group receiving a saline injection, a group receiving the vehicle alone and a group receiving the vehicle plus drug. Because of laboratory limitations, only four animals can be done on one day. The experimenter doesn't plan on adjusting the results for factors like the weight of the animal.

In this case, T (treatments) is 4 and C (covariates) is zero. So the sample size is at least 10 + (T–1) which is 10 + 3, which is 13. However, 13 animals will have to be done in at least 3 batches (assuming that the lab could manage a batch of five). This means that the experiment will probably have a minimum of 3 blocks, and more likely four. So, taking the blocks into consideration, the minimum sample size will be 10 + (T–1) + (B–1), which is 10 + 3 + 3, which is 16 animals.

The experimenter might like to aim for the maximum number of animals, to reduce the possibility that the experiment will come to a false-negative conclusion. In this case, 20 + (T–1) suggests 23 animals, which will have to be done in 6 blocks of four. 20 + (T–1) + (B–1) is 28, which means running 7 blocks of four, which requires another adjustment: an extra animal is needed because the number of blocks is now 7. The final maximum sample size is 29.

As you can see, when you are running an experiment in blocks, the sample size will depend on the number of blocks, which, in turn, may necessitate a small adjustment to the sample size.

Why do investigators use groups of 6 animals?

In early-stage research, most of the effects discovered will be dead ends. For this reason, researchers are only interested in pursuing differences between groups that are very large indeed. As can be seen from the table under “comparing the means of two groups”, two groups of 6 animals will detect a situation in which the scores of one group are almost entirely distinct from the scores of the other – there is a 92% chance that an animal in the high-scoring group will score higher than an animal in the low-scoring group.

“Everyone else used 6” is not a sample size calculation

Researchers should remember that this precludes the power to detect smaller differences, and justify their sample sizes based on the statistical power and the requirement for clinically significant effects to be very large. It’s not enough to say that everyone else used groups of 6.

8. Sample size for qualitative research

Issues

Qualitative researchers often regard sample size calculations as something that is only needed for quantitative research. However, qualitative research protocols typically contain statements like "participants will be recruited until data saturation occurs". So there is already an appreciation that a certain number of participants will be "enough participants".

Clearly, it is important when planning (and especially budgeting) a qualitative research project to know how many participants will be needed. These guidelines are partly derived from an excellent paper by Morse1

General guidance

Over-estimate your sample size when writing a proposal and budgeting it. This gives you some insurance against difficulties in recruitment, participants whose data is not very useful and other unanticipated snags.

Specific factors affecting sample size

Scope of study and nature of the topic

If the **scope of the study** is broad, then more participants will be needed to reach saturation. Indeed, broad topics are more likely to require data from multiple data sources. Doing justice to a broad topic requires a large commitment of time and resources, including large amounts of data. Broad studies should not be undertaken unless they are well-supported and have a good chance of achieving what they set out to do.

If the study addresses an obvious, clear topic, and the information will be easily obtained from the participants, then fewer participants will be needed. Topics that are harder to grasp and formulate are often more important, but require greater skill and experience from the researcher, and will require more data.

If they study topic is one about which people will have trouble talking (because it is complex, or embarrassing, or may depend on experiences which not everyone has) you will need more participants.

Quality of data and sample size

The ability of participants to devote time and thought to the interview, and to articulate their experiences and perceptions, and to reflect on them, will all affect the richness of the data. In particular, in some studies, participants may not be able to devote time to a long interview, or may not be physically or psychologically capable of taking part in a long interview, resulting in smaller

amounts of data from each interview. Where interviews are likely to be lower in information, larger sample sizes are needed.

On the other hand, when participants are being interviewed several times, this will generate more data, and sample sizes will be smaller.

Variability and sample size

The more variable the experiences, perceptions and meanings of the participants, the more participants will be needed to achieve the same degree of saturation.

Shadowed data and sample size

This is a term coined by Morse for situations in which participants talk about the experiences of others. You might call it 'secondhand data'. Collecting such data can make interviews more information rich and make better use of each participant, reducing the total sample size required. In particular, encouraging participants to compare and contrast their experiences, views and meanings with those of others can throw important light on variability in the domain you are studying. However, shadowing is no substitute for collecting first hand data, and may introduce bias.

So how many?

Morse recommends that semi-structured interviews with relatively small amounts of data per person should have 30 to 60 interviews. On the other hand, grounded theory research, with two to three unstructured interviews per person, should need 20 to 30 participants. In either case, the final choice of number should be guided by the other factors above.

A failsafe approach based on failure to detect

One question that a qualitative researcher should think about is this: if something doesn't emerge in my research (an attitude, an experience etc) then how common could it be in the population I am researching? Research, to be valid, must have a reasonable chance of detecting things that are common enough to matter. Failure to detect something important is a risk in all research, qualitative and quantitative. While you cannot guarantee that your research will absolutely detect everything important, you can at least make an estimate of the likelihood that your sample will fail to include at least one important topic/view/meaning etc.

The table shows numbers of participants and, for each number, shows how rare a theme, experience or meaning would have to be so that it was unlikely to be detected by the study.

Size of study	If you don't find something, the maximum likely prevalence is	That's roughly
60	6%	1 person in 20
40	9%	1 person in 10
30	13%	1 person in 8
20	18%	1 person in 6
15	25%	1 person in 4
10	37%	1 person in 3
8	46%	1 person in 2
5	74%	3 people in 4

Table 8.1 Sample size and likelihood of missing something important in qualitative research

As you can see, if a study of 60 people fails to identify a theme, experience or issue, that issue is probably rare – present in about one person in 20 or fewer. However, a study of 15 participants can fail to identify something which is present in one person in every four! And a study of 8 participants is quite likely to fail to find out things that affect half of the study population.

Clearly, shadowing (second hand data) can reduce these error rates by getting participants to talk about others, but this is no substitute for including the others in the research. Part of this is trying to chose a sample in such a way as to span the population, but this relies on knowing the factors that make for diversity in the population – something that may only become clear after the research is well under way.

However, both expert opinion in the area of qualitative research and the table above suggest that samples of less than 20 participants have to be justified on the grounds that they are unusually rich in data and representative.

Method

The table was calculated based on Poisson confidence intervals for zero observed frequencies at the given sample sizes, using Stata Release 14.1

References and further reading

Boddy CR. Sample size for qualitative research. Qualitative Mrkt Res: An Int J. 2016 Sep 12;19(4):426–32.

Marshall B, Cardon P, Poddar A, Fontenot R. Does sample size matter in qualitative research: a review of qualitative interviews in is research. Journal of Computer Information Systems 2013.

Morse JM. Determining Sample Size. Qual Health Res. 2000 January 1, 2000;10(1):3-5.

Morse JM. Analytic Strategies and Sample Size. Qual Health Res. SAGE Publications; 2015 Oct;25(10):1317–8.

Thomson SB. Sample Size and Grounded Theory. JOAAG. 2011 Mar 9;5(6):45–52.

van Rijnsoever FJ. (I Can’t Get No) Saturation: A simulation and guidelines for sample sizes in qualitative research. Derrick GE, editor. PLoS ONE. 2017 Jul 26;12(7):e0181689–17.

9. Resources for animal experiments

Festing, Michael FW, and Douglas G. Altman. "Guidelines for the design and statistical analysis of experiments using laboratory animals." ILAR journal 43.4 (2002): 244-258. http://ilarjournal.oxfordjournals.org/content/43/4/244.full

This paper appears as part of a collection which you can peruse here: http://ilarjournal.oxfordjournals.org/content/43/4.toc

Festing, Michael FW. "Design and statistical methods in studies using animal models of development." Ilar Journal 47.1 (2006): 5-14. http://ilarjournal.oxfordjournals.org/content/47/1/5.full?sid=6bb505df-77e8-48c3-8b9a-d67bd304deec

9. Computer and online resources

Free, highly recommended package: G*Power

 http://gpower.hhu.de/

For applications that go beyond the ones described here, including multiple regression, I can strongly recommend G*Power, which is free and multi-platform. There is an excellent manual.

Standard statistical packages

Stata also has a powerful set of sample size routines, and there are many user-written routines to calculate sample sizes for various types of study. Use the command `findit` sample size to get a listing of user-written commands that you can install.

The free professional package **R** includes sample size calculation (but requires a bit of learning). I recommend using software called **RStudio** as an interface to R. It makes R far easier to learn and use.

And no; **SPSS** will sell you a sample size package, but it isn't included with SPSS itself. If you use SPSS, my advice is to use **G*Power** and save money.

Sample size calculators and Online resources

You can look for sample size software to download at

http://statpages.org/javasta2.html

The **Graph Pad** website has a lot of helpful resources

http://graphpad.com/welcome.htm

They make an excellent sample-size calculator application called **StatMate** which gets high scores for a simple, intelligent interface and very useful explanations of the process. It has a tutorial that walks you through.

http://graphpad.com/scientific-software/statmate/

The OpenEpi website, which you can download to your computer for offline use, has some power calculations

http://www.openepi.com/Menu/OE_Menu.htm

There is a free Windows power calculation program at Vanderbilt Medical Center http://biostat.mc.vanderbilt.edu/wiki/Main/PowerSampleSize

GPower is a very comprehensive package for both Windows and Mac, available from http://gpower.hhu.de/

Online sample size calculators

WebPower

A splendid site that also offers an R package. It has a very comprehensive suite of power and sample size calculation methods. It also allows you to create a user ID so that you can save your work. There is a comprehensive manual. Recommended.

https://webpower.psychstat.org/wiki/

Manual, which has lots of useful reading, here:

https://webpower.psychstat.org/wiki/_media/grant/webpower_manual_book.pdf

Power and sample size

http://powerandsamplesize.com/

Excellent site with well-designed and validated calculators for a wide variety of study designs. Recommended.

Sealed Envelope power calculators

Calculations for clinical trials (the company provides support for clinical trials) including equivalence and non-inferiority trials

https://www.sealedenvelope.com/power/

Simple Interactive Statistical Analysis (SISA)

http://www.quantitativeskills.com/sisa/calculations/sampshlp.htm

Easy-to-use with good explanations but a smaller selection of study designs.

The survey system and Survey Monkey

http://www.surveysystem.com/sscalc.htm

https://www.surveymonkey.com/mp/sample-size-calculator/

Sample sizes for surveys. Survey Monkey has a very readable web page on sample size considerations.

Harvard sample size calculators

http://hedwig.mgh.harvard.edu/sample_size/size.html

A small selection, but clearly organised by study type.

Rules of thumb

Gerard van Belle's chapter on rules of thumb for sample size calculation can be downloaded from his website (http://www.vanbelle.org/) It's extracted from his book.