Sample size calculation
a quick guide

Thursday 28 October 2004

Ronán Conroy
rconroy@rcsi.ie

How to use this guide
This guide has sample size ready-reckoners for a number of common research designs. Each section is self-contained You need only read the section that applies to you.

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.

Contents
1. Sample size for percentages or proportions
   1. Sample sizes for simple prevalence studies 2
   2. Sample sizes for studies comparing a prevalence with a hypothesised value 4
   3. Sample sizes for studies comparing proportions between two groups 6

2. Sample sizes and powers for comparing two groups where the variable is measured on a continuous scale that is (more or less) normally distributed 9

3. Online resources including free software and web pages that calculate sample size for you 12

Feedback
If you have trouble following this guide, please email me. Your comments help to improve it.

Ronán Conroy
rconroy@rcsi.ie
1. Sample size for percentages or proportions

This paper gives guidelines for sample sizes for studies which measure the proportion or percentage of people who have some characteristic, and for studies which compare this proportion with either a known population or with another group. This characteristic can be a disease, and opinion, a behaviour, anything that can be measured as present or absent. 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 the table.

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.

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
Sample sizes for different margins of error for populations of different sizes

<table>
<thead>
<tr>
<th>Margin of error</th>
<th>Size of population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Large 5000 2500 1000 500 200</td>
</tr>
<tr>
<td>±20%</td>
<td>24 24 24 23 23 22</td>
</tr>
<tr>
<td>±15%</td>
<td>43 42 42 41 39 35</td>
</tr>
<tr>
<td>±10%</td>
<td>96 94 93 88 81 65</td>
</tr>
<tr>
<td>±7.5%</td>
<td>171 165 160 146 127 92</td>
</tr>
<tr>
<td>±5%</td>
<td>384 357 333 278 217 132</td>
</tr>
<tr>
<td>±3%</td>
<td>1067 880 748 516 341 169</td>
</tr>
</tbody>
</table>

**Example 1: Sample size for a study of the prevalence of anxiety disorders in medical patients**

The sample size of 400 patients will allow the study to determine the prevalence of anxiety disorders with a confidence interval of ±5%.

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

There are just over 500 registrars and senior registrars in Ireland. A representative sample of 127 of these will give the study a margin of error (confidence interval) of ±7.5% in determining the prevalence of bullying in the workplace.

**Analysing subgroups**

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, or decide not to analyse rarer subgroups.

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

*Sample sizes and powers: percentages and proportions* 3
1.2 Sample sizes for studies comparing a prevalence with a hypothesised value

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. You might want to demonstrate that medical students have a lower prevalence of smoking than other third level students, whose prevalence is already known from previous work.

In this case, you need to ask what is the smallest difference between the prevalence in the population I am 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: 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 smoking in medical students and discovered that the rate was 5% lower than the rate for the general population, would that have important clinical implications? How about if it was 10% higher? 20% higher?

Step 2: 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 the table to get an idea of sample size

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, err on the high side.

<table>
<thead>
<tr>
<th>Difference between the prevalences</th>
<th>50% prevalence in population</th>
<th>25% prevalence in population</th>
<th>10% prevalence in population</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power</td>
<td>Power</td>
<td>Power</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>95%</td>
<td>90%</td>
</tr>
<tr>
<td>5 %</td>
<td>1047</td>
<td>1294</td>
<td>825</td>
</tr>
<tr>
<td>10 %</td>
<td>259</td>
<td>319</td>
<td>214</td>
</tr>
<tr>
<td>15 %</td>
<td>113</td>
<td>139</td>
<td>97</td>
</tr>
<tr>
<td>20 %</td>
<td>62</td>
<td>76</td>
<td>56</td>
</tr>
<tr>
<td>25 %</td>
<td>38</td>
<td>46</td>
<td>36</td>
</tr>
<tr>
<td>30 %</td>
<td>25</td>
<td>30</td>
<td>25</td>
</tr>
</tbody>
</table>

Sample sizes and powers: percentages and proportions
Example: Study investigating whether photosensitivity is more common in patients with Asperger's syndrome than in the general population, using a limited number of available patients.

Photosensitivity has a prevalence of roughly 10% in the general population. There are approximately 60 persons with Asperger's syndrome in the two study centres who will all be invited to participate in the research. A sample size of 60 would give the study approximately 90% power to detect a 15% higher prevalence of photosensitivity in Asperger's syndrome compared with the general population. This sample size would have roughly a 95% power to detect a 20% higher prevalence.

1.3 Sample sizes for studies comparing proportions between two groups

Step 1: 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 was in the literature.

Step 2: 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: 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 the table 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!

**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.

**What if the groups are not the same size?**
This is beyond the scope of a ready-reckoner table. 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.

**Example: Study investigating the effect of a pre-discharge treatment programme on rate of readmission**

Previous studies in the area suggest that as many as 25% of patients are readmitted within a year of discharge. The proposed sample size of 500 patients in each group (intervention and control) will give the study a power to detect a 10% reduction in readmission 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. We propose to recruit 350 women in each group (work cessation and working). The proposed sample size has a 95% power to detect a twofold increase in risk, from 10% to 20%.
### Sample sizes for comparing the prevalence in two groups

The table shows the number needed in **each** group

**Sample sizes and powers: percentages and proportions**

---

<table>
<thead>
<tr>
<th>Difference between the groups</th>
<th>Power</th>
<th>Power</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>90%</td>
<td>95%</td>
</tr>
<tr>
<td>5%</td>
<td>2134</td>
<td>2630</td>
</tr>
<tr>
<td>10%</td>
<td>538</td>
<td>661</td>
</tr>
<tr>
<td>15%</td>
<td>240</td>
<td>293</td>
</tr>
<tr>
<td>20%</td>
<td>134</td>
<td>163</td>
</tr>
<tr>
<td>25%</td>
<td>85</td>
<td>103</td>
</tr>
<tr>
<td>30%</td>
<td>58</td>
<td>70</td>
</tr>
</tbody>
</table>

### 25% prevalence in one group

<table>
<thead>
<tr>
<th>Power</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>1714</td>
<td>2110</td>
</tr>
<tr>
<td>10%</td>
<td>460</td>
<td>563</td>
</tr>
<tr>
<td>15%</td>
<td>216</td>
<td>264</td>
</tr>
<tr>
<td>20%</td>
<td>128</td>
<td>155</td>
</tr>
<tr>
<td>25%</td>
<td>85</td>
<td>103</td>
</tr>
<tr>
<td>30%</td>
<td>61</td>
<td>73</td>
</tr>
</tbody>
</table>

### 10% prevalence in one group

<table>
<thead>
<tr>
<th>Power</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>957</td>
<td>1174</td>
</tr>
<tr>
<td>10%</td>
<td>286</td>
<td>349</td>
</tr>
<tr>
<td>15%</td>
<td>146</td>
<td>177</td>
</tr>
<tr>
<td>20%</td>
<td>92</td>
<td>111</td>
</tr>
<tr>
<td>25%</td>
<td>65</td>
<td>78</td>
</tr>
<tr>
<td>30%</td>
<td>49</td>
<td>58</td>
</tr>
</tbody>
</table>
2: Sample sizes and powers for comparing two groups where the variable is measured on a continuous scale that is (more or less) normally distributed.

Step 1: decide on the difference that you want to be able to detect and express it in standard deviation units.

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 might not be.

<table>
<thead>
<tr>
<th>Difference to be detected (SD units)</th>
<th>N in each group 90% power</th>
<th>N in each group 95% power</th>
<th>Chance that someone in group 1 will score higher than someone in group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>6</td>
<td>7</td>
<td>92%</td>
</tr>
<tr>
<td>1.5</td>
<td>10</td>
<td>12</td>
<td>86%</td>
</tr>
<tr>
<td>1.4</td>
<td>11</td>
<td>14</td>
<td>84%</td>
</tr>
<tr>
<td>1.3</td>
<td>13</td>
<td>16</td>
<td>82%</td>
</tr>
<tr>
<td>1.25</td>
<td>14</td>
<td>17</td>
<td>81%</td>
</tr>
<tr>
<td>1.2</td>
<td>15</td>
<td>19</td>
<td>80%</td>
</tr>
<tr>
<td>1.1</td>
<td>18</td>
<td>22</td>
<td>78%</td>
</tr>
<tr>
<td>1</td>
<td>22</td>
<td>26</td>
<td>76%</td>
</tr>
<tr>
<td>0.9</td>
<td>26</td>
<td>33</td>
<td>74%</td>
</tr>
<tr>
<td>0.8</td>
<td>33</td>
<td>41</td>
<td>71%</td>
</tr>
<tr>
<td>0.75</td>
<td>38</td>
<td>47</td>
<td>70%</td>
</tr>
<tr>
<td>0.7</td>
<td>43</td>
<td>54</td>
<td>69%</td>
</tr>
<tr>
<td>0.6</td>
<td>59</td>
<td>73</td>
<td>66%</td>
</tr>
<tr>
<td>0.5</td>
<td>85</td>
<td>104</td>
<td>64%</td>
</tr>
<tr>
<td>0.4</td>
<td>132</td>
<td>163</td>
<td>61%</td>
</tr>
<tr>
<td>0.3</td>
<td>234</td>
<td>289</td>
<td>58%</td>
</tr>
<tr>
<td>0.25</td>
<td>337</td>
<td>416</td>
<td>57%</td>
</tr>
<tr>
<td>0.2</td>
<td>526</td>
<td>650</td>
<td>56%</td>
</tr>
</tbody>
</table>

It is hard to define the smallest difference that would be clinically significant. An element of guesswork is involved. What is the smallest reduction in cholesterol that would be regarded as clinically worthwhile? It

*Sample sizes and powers: comparing two means*
may be useful to search the literature and see what other investigators have done.

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 different to standard deviation units.

1. What is the expected mean value for the control group?
2. What is the standard deviation of the control group?
   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 foetal 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.

3. What is the smallest difference between the treated and control group (or between any 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.

   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.

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

   \[ \text{Minimum difference to be detected} \div \text{standard deviation} \]

   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 3: Use the table 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.

Interpretation 1: 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. However, a research study should have a reasonable chance of detecting a difference (if it is there, 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?

Interpretation 2: 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 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 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.

Sample sizes and powers: comparing two means
Finding 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.

Example: a study examining the effect of smoking on cardiac output in pregnancy.

Mean cardiac output rises 0.7 litres/min between booking visit and the 36th week of pregnancy. The sample size was calculated on the basis of detecting a halving of this rise (0.35 litres vs 0.7 litres). The standard deviation of the rise, based on previous research, is 1.2 litres/min. The study was therefore designed to detect a difference of 0.3 litres with 95% power, giving a sample size of 289 in each group. The proposed sample size of 350 in each group is conservative, and allows for a participant drop-out rate of 20%.

Example: a study examining differences in pain scores between two different injection sites.

150 patients will be randomised into two groups of 75, to receive the injection either at site A or site B. This sample size gives a 95% power to detect a difference of 0.75 standard deviations in pain score between site A and site B. Since pain scores are measured on an arbitrary 100 millimetre visual analogue scale, the clinical significance of a 0.75-standard deviation difference between the sites can be more readily appreciated by saying that such a difference implies that 70% of patients would experience less pain if injected at site A than they would if injected at site B.
Calculating sample sizes for comparing two means: a rule of thumb

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. Online resources

Try the online sample size calculators at
http://home.clara.net/sisa/sampshlp.htm
http://members.aol.com/johnp71/javastat.html#Power
http://hedwig.mgh.harvard.edu/sample_size/size.html

You can also look for sample size software to download at
http://members.aol.com/johnp71/javasta2.html

Good help pages at http://www.cmh.edu/stats/size.asp
and both help and online calculation at
http://www.stat.uiowa.edu/~rlenth/Power/

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
http://graphpad.com/StatMate/statmate.htm

There is a free calculator for the Palm at
http://www.bobwheeler.com/stat/SSize/ssize.html
which includes a very extensive manual. The interface is a bit sparse but the
manual makes it all clear.

Gerard van Belle's chapter on rules of thumb for sample size calculation can
be downloaded from
http://www.seattlecrc.org/vbelle/chapters/webchapter2.pdf

Finally, *Ask Professor Mean* is an excellent resource for all your statistical
queries. The page relevant to sample size is at
http://www.cmh.edu/stats/size.asp