Sample Size and Power

General remarks

One of the most transparent reasons why statistical analyses are based on the means of samples rather than just single values is that they are in some sense an improvement on using single values. In less vague terms, the sample mean, $\bar{x}$, becomes a more and more precise estimate of the population mean, μ, as the sample size increases. A quantitative measure of this precision is the standard error, $\sigma/\sqrt{n}$, which decreases as the precision increases. The larger n becomes the smaller is the standard error.

The dependence of the standard error on sample size can be exploited when a study is being planned. The investigators need to decide how much precision is needed for their purposes and design the study accordingly. It is wasteful, and possibly unethical, to recruit many more patients than you need, while on the other hand a study that recruits too few may well be pointless.

There are several methods of determining sample size and each has several variants, depending on the kind of outcome being measured. Sample sizes could be based directly on the measure of precision, so that the width of a confidence interval, or equivalently the size of a standard error, is required to be less than a prescribed value. An alternative method is to set the sample size so that a hypothesis test has a given power. The latter is probably more widely used, although the former is, perhaps, rather underused.

In either case the sample size will be given by a formula or a table which will require the values of certain unknown parameters to be specified. It would not be unreasonable to think of the study as a means of obtaining such estimates, and consequently their prior specification in a sample size calculation can often seem rather unhelpful. Nevertheless, this is problem is inevitable. For some parameters logically satisfactory ways round this impasse can be found, but for others there is no option but to attempt to find some estimate, perhaps from the literature or from a pilot study. For this reason it is important to realise that sample size calculations do not give precise values: they depend on parameters whose values are unknown and they will vary, sometimes alarmingly, as the values used for the parameters vary.

It is a wise precaution when performing a sample size calculation to do a *sensitivity analysis*, where the range of sample sizes obtained from a range of parameter values is considered. Of course, this will not help in deciding what parameter values are correct, nor should it be used to accord special status to parameter values corresponding to convenient samples sizes. However the exercise is useful in telling the investigators how much reliance it is prudent to place on the computed sample size.

It should also be appreciated that in some pieces of research, the area is insufficiently understood to allow a sensible sample size calculation to be performed. For example, if it is impossible to find adequate values for the parameters needed to provide the sample size that ensures a hypothesis test has a certain power, then perhaps it should be admitted that it is too early in the process of research for a hypothesis to be tested. At an early stage it is inevitable that sample sizes will be determined almost arbitrarily: in these circumstances the results of the study may need to be used with caution.

Methods based on estimation

The confidence interval for a mean is $\bar{x} \pm ts/\sqrt{n}$, where *s* is the sample standard deviation and *t* is the appropriate point from a *t*-distribution. Strictly speaking the value of *t* will change with *n* but in fact, once $n>30$, the value changes little as *n* changes. Consequently it is much simpler to base sample size calculations on the approximate confidence interval $\bar{x} \pm zs/\sqrt{n}$ where *z* is the appropriate point of a standard Normal distribution. So, for example, for a 95% interval $z=1.96$. If the calculation results in a value of *n* below 30 then it might be prudent to increase the value slightly to allow for this approximation, but this is seldom needed in practice. The width of the confidence interval is $2zs/\sqrt{n}$, which is approximately $4s/\sqrt{n}$ for a 95% interval. Consequently sample size calculations which prescribe a limit for the width of a confidence interval are essentially the same as those which put a limit on the size of the standard error, they merely differ by a known factor 2*z*

If the standard error is required to be less than *L* then this means that *n* must exceed s^2/L^2. The value of *s* will clearly be unknown, but it can be replaced by σ, but this too is unknown. It is necessary to find some prior estimate of this parameter and this is substituted in σ^2/L^2. It is not really material whether you think of the process as estimating *s* or σ.

It is important to ensure that the correct standard error is used. The above uses the standard error of a sample mean. If the aim is to set a limit on the standard error of the difference between the means of two groups, then that standard error should be used. If the responses in the two groups have a common standard deviation, then standard error of $\bar{x}_1 - \bar{x}_2$ is:

$$s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (1)$$

where *s* estimates the common standard deviation and n_1 and n_2 are the sizes of the two groups. Placing a limit on (1) is not sufficient to determine two samples sizes. However, it is usual to assume that the two groups have the same size (it is very easy to adapt the calculation to allow any other pre-specified ratio between the two sample sizes). In this case (1) become $s\sqrt{2}/\sqrt{n}$ where *n* is the size of each group. Consequently putting a limit of *L* on the standard error of the difference between two groups requires

$$n \geq \frac{2s^2}{L^2}.$$

Methods based on hypothesis tests

The ideas behind this approach are the same regardless of the type of outcome variable that is the basis of the calculation. However, the details do differ substantially and therefore two cases are dealt with here, namely a Normally distributed variable and a binary variable. The main ideas are introduced in this context and the necessary modifications are dealt with when binary data are considered.

Normally distributed outcome

Suppose interest is focussed on the comparison of two groups with respect to a variable that has a Normal distribution. In particular the null hypothesis of interest is that the means in the two groups are equal. It is assumed that the responses in the two groups share a common population standard deviation.

The usual method of testing this hypothesis is to use an unpaired *t*-test which is based on the test statistic

$$z = \frac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

The *P*-value is found by referring this to a *t*-distribution on the appropriate degrees of freedom. Provided the combined sample size exceeds about 30, a simpler approximation is to refer the statistic to a standard Normal distribution.

Hitherto it has been recommended that the size of the P-value be used as a measure of the evidence against the null hypothesis. A similar approach, but which has a slightly different emphasis, is to reject the null hypothesis if the P-value is below some critical value, *C*. Two kinds of error could be made:

Type I - *the null hypothesis is rejected when it is true*

Type II - *The null hypothesis is not rejected when it is false.*

The probability of the first is determined by the critical *C* and this does not depend on sample size. For example, a Type I error probability or rate of 5% corresponds to *C*=1.96.

The probability of a Type II error depends on several things, including the sample size, and *n* can be set by specifying, amongst other things, the Type II error rate.

The Type II error is 1 minus the power, because the power of the test is the probability that a false null hypothesis is rejected.

The formula for the sample size will be given below and its components described. The mathematical justification of the formula can be found in the appendix. However, although it is rather imprecise, there is some value in a more heuristic explanation of how power can be used to set sample size can be given and this is now given.

Heuristic explanation

When testing the difference in means between two groups we are trying to decide if $\mu_1 - \mu_2$ is zero or not. We have to try to do this on the basis of the difference in sample means, $\bar{x}_1 - \bar{x}_2$. While we know that $\bar{x}_1 - \bar{x}_2$ contains information on $\mu_1 - \mu_2$, we also know that $\bar{x}_1 - \bar{x}_2$ is an imprecise estimate of $\mu_1 - \mu_2$. The standard error of this difference, *se* measures this imprecision.

Consider two cases: one in which $\mu_1 - \mu_2$ has a given (non-zero) value and another in which $\mu_1 - \mu_2$ is twice this value. Both cases have the same *se*. The distributions of the observed values of $\bar{x}_1 - \bar{x}_2$ in the two cases is shown in figure 1. It is clear that you will have a much better chance of being able to conclude that $\mu_1 - \mu_2$ is not zero if $\mu_1 - \mu_2$ is larger.

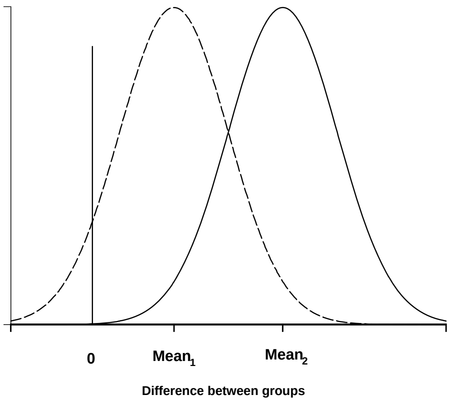

Figure 1: distribution of sample means: solid case has twice the population mean of the dashed case.

Although obvious, this is an important point: the power of a test is the probability that the test rejects the null hypothesis when the null hypothesis is false. However, unlike the null hypothesis being true, which defines a value for $\mu_1 - \mu_2$ (it is zero), the null hypothesis being false only entails $\mu_1 - \mu_2 \neq 0$, so the null hypothesis can be false in an infinite number of ways, and the power will not be the same for all these values. In other words, the power of a test is actually a function of the true difference, $\mu_1 - \mu_2$, it is not a single number. In order to specify the power as a single number the value of $\mu_1 - \mu_2$ must also be specified. Of course, $\mu_1 - \mu_2$ is always unknown and the way this is handled is explained below, once the formula for sample size has been given.

A second circumstance to consider is when there are two cases where, the difference in the population means is the same in both cases but the standard errors are different. This is shown in figure 2.

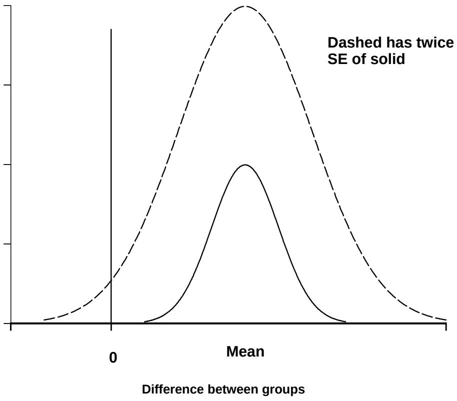

Figure 2: distribution of sample means: solid case has half the standard error of the

dashed case, both cases have the same mean.

It is clear that we have a much better chance of inferring that $\mu_1 - \mu_2$ is non-zero in the case with the smaller standard error. The standard error depends on the sample size and it can be made as small as we like by making the sample size sufficiently large. If the standard error is sufficiently small, then the distribution of $\bar{x}_1 - \bar{x}_2$ will be clustered sufficiently tightly about $\mu_1 - \mu_2$ that (provided $\mu_1 - \mu_2$ really is not zero) we will be very likely to be able to infer that $\mu_1 - \mu_2 \neq 0$. This is the basis of using this approach to set sample sizes.

Sample size formula

Suppose we are comparing two groups, with the responses in group 1 having a Normal distribution with mean μ_1 and standard deviation σ and group two being the same but that the mean is μ_2. A test of the null hypothesis that these means are equal will have Type II error β (so power 1-β) if the size of *each* group is:

$$n = \frac{2\sigma^2 (z_\beta + z_{\frac{1}{2}\alpha})^2}{(\mu_1 - \mu_2)^2}. \qquad (2)$$

The term z_β is simply the value that is exceeded by a proportion β of a standard normal population. The term in α is related to the significance level used to reject the null hypothesis. Other aspects of the formula are in accord with the heuristic explanation given above. If the outcome measure is more variable, i.e. σ is larger, then a larger value of *n* is required. If the difference between the groups is smaller, then again *n* will need to be larger if the same power is to be obtained.

The values of the *z* terms in (2) can be found from tables or statistical packages. However, when determining sample sizes only a few values of α and β are used, so the most commonly encountered values of $2(z_\beta + z_{\frac{1}{2}\alpha})^2$ = *A* say, can easily be tabulated. If the null hypothesis is considered to be rejected if a two-sided *P*-value is less than 0.05 then α=0.05 and $z_{\frac{1}{2}\alpha} = 1.96$. It is common to find that powers of 80%, 90% and 95% are considered, and this then gives the following.

Power	80%	90%	95%
A	15.7	21.0	26.0

The formula for sample size calculation becomes

$$n = A \times \frac{\sigma^2}{(\mu_1 - \mu_2)^2} = A \times \left(\frac{\sigma}{(\mu_1 - \mu_2)} \right)^2 \qquad (3)$$

with *A* taken from the above table.

Use of the formula

The choice of α and β, the Type I and II error rates, is a reflection of the investigator's views on the acceptability, or otherwise, of either type of error. Once chosen then A is determined and the formula in (3) needs to be applied, but neither σ nor $\mu_1 - \mu_2$ is known, so how do you proceed? Of course, to apply (3) you only need to know the ratio of these quantities and this can sometimes be exploited to simplify matters. However, it is generally a good idea to avoid this simplification unless you are very familiar with the variables you are dealing with.

Some value for σ needs to be obtained from somewhere, whether from the literature or existing data or a specially designed pilot study. Two notes of caution should be sounded here:

1. the standard deviation should measure the *same* variation as is present in the data values that go to make up each of $\bar{x}_1, \bar{x}_2$. If these measure the change in serum cholesterol from baseline to the end of a trial, it is no use finding a value for the standard deviation of a cholesterol level: it is the standard deviation of the *change* that needs to be entered in (3).
2. If the value of σ is based on an estimate s from a small pilot study, remember that s is going to be a highly variable estimate of σ. In these circumstances it is particularly important to perform a sensitivity analysis. Constructing a confidence interval for σ from s is likely to be a worthwhile if somewhat salutary exercise.

It is important to be clear how to think of the value of $\mu_1 - \mu_2$ that is entered in (3). It was remarked above that the power of a test was a function of $\mu_1 - \mu_2$ and this is made explicit in figure 3.

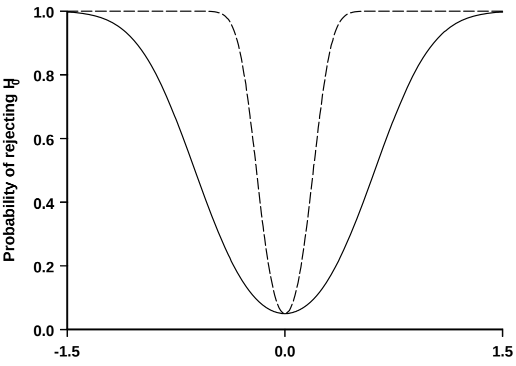

Figure 3: power curves for two tests: sample sizes larger in dashed case than solid.

As $\mu_1 - \mu_2$ gets larger, either positively or negatively, the probability of rejecting the null hypothesis approaches 1. However, if $\mu_1 - \mu_2 = 0$ then the probability of rejecting the null hypothesis (which is now the Type I error rate) is fixed at 0.05 (or, more generally, α), so the curves for all tests, whatever sample size they use, must pass through the point (0,0.05). At a given value of $\mu_1 - \mu_2$ (strictly $(\mu_1 - \mu_2)/\sigma$), the higher curve in figure 3 corresponds to the larger sample sizes. The idea behind the sample size calculation is to focus on a value of $\mu_1 - \mu_2$ and then choose *n* so that the height of the power curve at that value of $\mu_1 - \mu_2$ is the desired power, such as 0.8, or 0.9 or 0.95.

How is the value of $\mu_1 - \mu_2$ chosen? It is clear that the power for values of the abscissa between $\mu_1 - \mu_2$ and 0 is less then the value set for $\mu_1 - \mu_2$. This is inevitable: if you want to have good power for a very small value of $\mu_1 - \mu_2$ then a very large sample size will be needed. Consequently, investigators have to accept that it will be difficult for their study to detect reliably very small values of $\mu_1 - \mu_2$. However, very small values of $\mu_1 - \mu_2$ are unlikely to be of interest: no physician is going to want to know if treatment X reduced mean blood pressure by 1 mmHg more than treatment Y. On the other hand, if the study is such that there is poor power to detect a clinically important value of $\mu_1 - \mu_2$, then the study may well end up giving a non-significant result (say P=0.5) when, in fact, there is an important difference between the groups.

The approach that is used is to decide on what value of $\mu_1 - \mu_2$ corresponds to the minimum clinically important difference between the groups. The study is then designed to have good power at that value of $\mu_1 - \mu_2$. If the true value of $\mu_1 - \mu_2$ is larger than the minimum clinically important difference then the study will have even higher power. If the true value of $\mu_1 - \mu_2$ is smaller than the minimum clinically important difference, then the study may be too small to be able to detect this difference reliably but in this case the study will miss a difference that has been deemed to be clinically unimportant.

Example

In a trial to compare a new treatment for influenza, zanamivir with placebo the primary outcome variable was the number of days to the alleviation of symptoms (MIST Study Group, Lancet, 352, 1877-1881). A previous study suggested a value for σ of 2.75 days. It was decided that a change of one day in the mean number of days to alleviation of symptoms was of interest, but that any smaller improvement would not be of clinical value.

The aim, therefore, is to determine how many patients should be allocated to each treatment group. If it is decided that it is important not to miss a difference of 1 day, then a power of 90% may be selected. Assuming the significance level is 5%, then *A* = 21.0 and as σ = 2.75 days and $\mu_1 - \mu_2$ = 1 day, formula (3) gives the number of patients in each group as:

$$21.0 \times \frac{2.75^2}{1^2} = 158.92 \approx 159$$

so the trial should aim to recruit 318 patients and allocate them equally between the groups. If σ had been taken as 2 or 3.5 days the number per group would have been 84 or 258, showing how uncertainty in the prior estimate of variability translates into noticeable uncertainty in the calculated sample size. (to put this kind of variation in context, note that if 2.75 days were an estimate based on a sample of size 25 then a 95% confidence interval for σ would be 2.15 to 3.83 days).

Binary outcome

The principles in the case when the outcome is binary are unchanged. The test that the population proportions in the two groups are the same needs to be performed at a given significance level, which is usually taken as 5%. The investigator needs to specify the power that is required. Thus α and β need to be decided just as in the formula above. The minimum clinically important difference between the two proportions, $\pi_1 - \pi_2$ must also be specified.

The main difference between the Normal and binary cases arises in specifying the remaining parameter, which in the Normal case would be σ, the measure of variability. There is no direct analogue of this parameter in the binary case. The standard error of a proportion from a sample of size *n* is $\sqrt{\pi(1-\pi)/n}$, i.e. the variability is determined by a function of the population proportion. The way this is resolved is by specifying both π_1 and π_2 rather than just their difference $\pi_1 - \pi_2$. This is because the dependence of the standard error of a proportion on the value of that proportion means that different numbers of patients are needed to detect $\pi_1 - \pi_2$=0.2, when this is from, e.g., 0.1 to 0.3 or from 0.4 to 0.6. Providing both π_1 and π_2 rather than just their difference allows the technique to take proper account of this. In some ways this makes it easier to set a sample size when the outcome is binary because the calculations depend on parameters that are more likely to be known than in the normal case.

Various formulae are available. In the same way that the formula (2) is closely related to the *t*-test, the formulae for the binary case are generally related to the χ^2 test. An alternative is to use an approach related to Fisher's Exact test. Mathematically this is rather complicated but an easily used table is available which actually makes it one of the easiest methods to use. It is given in table 3B reproduced below from Casagrande, Pike and Smith, *Applied Statistics* 1978, 27, 176-180.

In order to estimate a sample size the investigator must specify values for π_1 and π_2: the change $\pi_1 - \pi_2$ can be determined in the same way as $\mu_1 - \mu_2$ in the normal case, i.e. as the minimum clinically important difference. In addition one of π_1 and π_2 must also be given (of course any other independent combination, such as the mean of the proportions, would suffice but this is seldom convenient). It is often the case that one of the proportions will measure the success rate of the *status quo,* so is likely to be known.

Table 3B assumes that π_1 and π_2 have been labelled so that $\pi_2 < \pi_1$: the values used in the table are π_2 (defining the rows) and π_1 - π_2 (defining the columns). Only values of π_2 up to 0.50 are given in the table, so does this mean it cannot be used if the smaller proportion is greater than this value? It does not and this can be illustrated by considering an example. Suppose the sample size needed to detect a change from 0.6 to 0.8 was sought. This could be thought of as the size of study needed to detect an increase in success rate from 60% to 80%. This is, of course equivalent to the failure rate decreasing from 40% to 20%. Table 3B does not specify whether the binary outcome is success or failure and, in purely mathematical terms, they are equivalent,

TABLE 3B

One-tailed tests

Upper figure: Test of significance at 2·5% level, probability 80%
Middle figure: Test of significance at 2·5% level, probability 90%
Lower figure: Test of significance at 0·5% level, probability 95%

	$\delta = P_1 - P_2$													
P_2 = *smaller probability of success*	0·05	0·10	0·15	0·20	0·25	0·30	0·35	0·40	0·45	0·50	0·55	0·60	0·65	0·70
0·05	466	151	82	55	39	31	24	20	17	14	13	11	10	9
	606	198	106	69	51	38	31	24	21	18	16	13	12	11
	999	321	170	110	80	61	48	40	33	28	24	21	19	16
0·10	721	214	110	69	49	36	29	23	19	17	13	12	11	9
	952	281	143	89	62	47	36	29	24	20	18	16	13	11
	1587	462	232	144	100	74	58	47	38	32	27	23	20	18
0·15	942	267	131	81	56	41	32	25	22	17	16	12	11	9
	1247	351	172	104	72	53	41	32	26	22	18	16	13	11
	†	582	282	171	116	85	65	52	42	35	29	24	21	18
0·20	1133	311	150	90	61	44	33	27	22	18	16	12	11	9
	1505	410	197	117	79	58	43	33	28	23	18	16	13	11
	†	683	323	193	129	93	70	55	43	36	29	24	21	18
0·25	1290	347	164	98	64	48	37	29	23	18	16	12	11	9
	1713	459	216	127	85	61	47	37	29	23	18	16	13	11
	†	763	357	209	138	98	74	57	46	36	29	24	20	16
0·30	1416	375	174	102	69	48	37	29	23	18	16	12	10	—
	1877	496	231	133	87	61	47	37	29	23	18	16	12	—
	†	825	381	220	144	103	75	57	46	36	29	23	19	—
0·35	1509	393	183	102	70	49	37	29	22	17	13	11	—	—
	2008	524	242	140	87	61	47	37	28	22	18	13	—	—
	†	872	395	229	145	103	75	57	43	35	27	21	—	—
0·40	1587	404	183	108	70	48	37	27	22	17	13	—	—	—
	†	533	244	141	87	61	47	33	26	20	16	—	—	—
	†	903	405	229	145	103	74	55	42	32	24	—	—	—
0·45	1600	416	183	102	69	48	33	25	19	14	—	—	—	—
	†	546	244	140	87	61	43	32	24	18	—	—	—	—
	†	898	405	229	144	98	70	52	38	28	—	—	—	—
0·50	1600	404	183	102	64	44	32	23	17	—	—	—	—	—
	†	533	242	133	85	58	41	29	21	—	—	—	—	—
	†	903	395	220	138	93	65	47	33	—	—	—	—	—

† These values could not be computed since they were too large for the maximum job size (56K bytes) of the computer used.

so the problem of determining the size of study needed to detect a change from 60% to 80% is solved by using $\pi_2 = 0.20$ and $\pi_1 - \pi_2 = 0.20$.

For each combination of π_2 and $\pi_1 - \pi_2$ there are three entries in the table. They are the sample size *for each group* for three different combinations of significance level and power. The top two figures are for a two-sided significance level of 5% (the table describes them as one-sided levels of 2.5% but this amounts to the same thing) and the last figure has a significance level of 1%. The top figure gives a power of 80%, the middle corresponds to 90% and the bottom figure is for 95% power.

So, for example, if the success rate of the standard treatment is 25% and it is important to be able to detect a change of 15%, at the 5% significance level, when using the new treatment (i.e. that the success rate on the new treatment is 40%) then two groups each of 216 patients would give a 90% chance of detecting this difference. If the groups only contained 164 patients the power would be 80%.

General Remarks

It should be clear that a sample size calculation is far from a precise and objective exercise. If uncertainty about key parameters lead to sample size estimates varying between, e.g. 400 and 800, it may be asked why they are helpful at all? It must be conceded that it is often unwise to take a sample size calculation as an immutable target for the study. However, even if a sample size can plausibly range between 400 and 800, it certainly excludes 100, and if this is the largest number of patients that can be recruited in the foreseeable future, then the calculation has been useful in stopping a woefully inadequate design from proceeding.

While the details are beyond the scope of this note, formula (2) can be used in other ways. In particular it is possible to use it to calculate the power of a study with a given number of patients. So, if an investigator knows that in a year he will be able to recruit 200 patients, then (2) can be used to find the value of z_β and hence the power, $1-\beta$ to detect a given $\mu_1 - \mu_2$. The approach should be used cautiously: it is all too easy to inappropriately convince oneself that a power of, say, 65%, is adequate, if the alternative is to abandon the study altogether.

In some sample size calculations the investigators anticipate a certain drop-out rate from the study and inflate their sample size estimates accordingly. This can be quite sensible and can guard against an unfortunate loss power. However, it should be borne in mind that in these circumstances loss of power may not be the principal problem. In the example of a clinical trial the groups would be randomized and therefore comparable. There is no guarantee that the patients dropping out of one group are comparable with the group dropping out of the other, so the loss of comparability may be much more serious than the loss of power, and additional recruitment cannot compensate for this.

A final issue is that of so-called *post hoc* power. Once a study has been completed an investigator may try to work out what power was actually achieved. This may be especially tempting when the study is non-significant, inasmuch as a 'non-significant' P-value is widely acknowledged to be uninformative if the study had inadequate power, so ruling out this possibility is particularly important. This practice is, however, generally misguided. As figure 3 shows, the power is a function of $\mu_1 - \mu_2$ which is unknown both before and after the study.. Of course, after the study the investigator has some knowledge of $\mu_1 - \mu_2$ through the observed value of $\bar{x}_1 - \bar{x}_2$. There are, however, two comments that are relevant here.

i) The investigators should be interested in their study having adequate power to detect a clinically meaningful difference, and this is determined by thinking about the clinical problem and exercising clinical judgment: the observed difference in sample means is not relevant and therefore neither is evaluating the power at $\bar{x}_1 - \bar{x}_2$ from figure 3.

ii) Attempts to determine the power at the true difference by using the fact that $\bar{x}_1 - \bar{x}_2$ is an estimator of $\mu_1 - \mu_2$ are not sensible. The former is only an estimator of the latter and will vary from it: if the study has resulted in a non-significant P-value then the variation of $\bar{x}_1 - \bar{x}_2$ about $\mu_1 - \mu_2$ can be substantial. Given that parts of the curve in figure 3 are very steep, even small variations about $\mu_1 - \mu_2$ will result in widely varying estimates of power. The range of possible powers could be found by constructing a confidence interval for $\mu_1 - \mu_2$ and then finding the powers corresponding to the end points of this interval. This is bizarre: the interest in the study is, or should be, in the value of $\mu_1 - \mu_2$ and a confidence interval is an appropriate measure of this: translating this to a power is unnecessary and misleading. Once a study has been completed, interest should focus on *confidence*, not *power*.

Appendix deriving formula (2) (Not Assessed)

The derivation of formula (2) starts be considering the test statistic $D = (\bar{x}_1 - \bar{x}_2)/se$. Later in the derivation *se* will be replaced by its expression in terms of sample size, namely:

$$\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}, \qquad \text{(A1)}$$

but for the moment *se* will suffice.

A test of the null hypothesis $\mu_1 - \mu_2$=0 is performed by seeing if *D* is between -$z_{½\alpha}$ and $z_{½\alpha}$. where α is the significance level (so $z_{½\alpha}$ is the value which is exceeded by a proportion ½α of a standard Normal distribution). For example, if the test is two-sided at the 5% level then the null hypothesis is not rejected if *D* is between -1.96 and 1.96. This region is shown by the shaded box in figure A1.

Under the null hypothesis, *D* has a standard Normal distribution (i.e. a Normal distribution with mean 0 and standard deviation 1), so the choice of -$z_{½\alpha}$ and $z_{½\alpha}$. ensures that the chance of rejecting the null hypothesis is indeed α, the Type I error. This is depicted by the dashed Normal curve in figure A1.

What happens if the null hypothesis is false, and $\mu_1 - \mu_2$ takes some value other than 0? No generality is lost if is it is assumed that $\mu_1 - \mu_2$>0. The null hypothesis is still tested in the same way (this must be the case, after all we never know for certain if $\mu_1 - \mu_2$=0 or not), but the distribution of *D* is no longer standard Normal - it is Normal, still with standard deviation 1 but now the mean is $(\mu_1 - \mu_2)/se$. This is shown as the solid curve in figure A1.

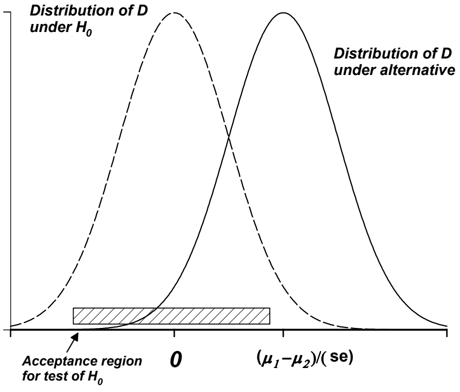

Figure A1.

As $(\mu_1 - \mu_2)/se$ gets larger the solid distribution in figure A1 moves to the right, so the chance of not rejecting the null hypothesis, i.e. the chance that *D* falls within the shaded box, gets smaller. The quantity $(\mu_1 - \mu_2)/se$ can get larger because $\mu_1 - \mu_2$ gets larger or because *se* gets smaller, i.e. the sample size gets larger.

Formula (2) arises by working out the probability that a variable with the solid distribution in figure A1 falls within the shaded region, i.e. the probability that the null hypothesis is accepted even though it is false. This is the Type II error β (remember that the power is 1-β).

The probability of a Type II error is, therefore, the probability that a Normal variable with mean $(\mu_1 - \mu_2)/se$ and standard deviation 1 falls between $-z_{½\alpha}$ and $z_{½\alpha}$. It is convenient to do this calculation in terms of a standard Normal variable (SNV), and this can be done if you slide the values $-z_{½\alpha}$ and $z_{½\alpha}$. and the solid distribution in the figure down until the solid distribution is centred about 0. The probability of a Type II error is then

$$\beta = \text{Prob(SNV} < z_{\frac{1}{2}\alpha} - (\mu_1 - \mu_2)/se\text{) - Prob (SNV} < -z_{\frac{1}{2}\alpha} - (\mu_1 - \mu_2)/se\text{)}$$

The second term above is the chance that a value from the solid distribution falls below $-z_{½\alpha}$ and this is clearly very small, so this term can be ignored. Also, by definition z_β is the value that an SNV exceeds with probability β, so by the symmetry of the standard Normal distribution the probability that a SNV is less than $-z_\beta$ is β. So the above can be written as

$$\beta = \text{Prob (SNV} < -z_\beta) \cong \text{Prob(SNV} < z_{\frac{1}{2}\alpha} - (\mu_1 - \mu_2)/se\text{)}$$

so from the second equation in this line

$$-z_\beta = z_{\frac{1}{2}\alpha} - (\mu_1 - \mu_2)/se\,.$$

And rearranging this gives:

$$\frac{1}{se^2} = \frac{(z_{\frac{1}{2}\alpha} + z_\beta)^2}{(\mu_1 - \mu_2)^2}$$

and using equation (A1) this gives:

$$\frac{n_1 n_2}{n_1 + n_2} = \frac{\sigma^2 (z_{\frac{1}{2}\alpha} + z_\beta)^2}{(\mu_1 - \mu_2)^2}$$

This is as far as we can go unless we decide on the ratio between the sizes of the two groups. If we use equal group sizes, $n_1 = n_2 = n$ then the left hand side above is simply ½*n* and (2) follows immediately. Any pre-specified value for n_1/n_2 can be chosen and modest departures from one can be practically useful without entailing serious loss of power. However, large imbalances will result in a loss of power relative the use of equal groups (with the same number of patients).