GraphPad StatMate 2

Copyright (C) 2004, GraphPad Software, Inc.

Table of Contents

© 2011 GraphPad Software, Inc.

© 2011 GraphPad Software, Inc.

1 Introducing GraphPad StatMate

GraphPad StatMate helps you with power and sample size analyses. StatMate is used for two purposes:

- To help you decide how many subjects to use in your study (or to justify the size you've already chosen!).
- To evaluate the power of a completed experiment that resulted in a "not significant" result.

Note: StatMate does not perform any statistical analyses. If you want to perform a t test, a chi-square test, linear regression, ANOVA, etc., then you have launched the wrong program.

If you prefer to follow detailed examples (with explanations), we provide a step-by-step example of how to determine sample size for an experiment that you are planning and another on how to compute the power of a completed experiment.

I welcome comments about StatMate and these help screens.

Harvey Motulsky
CEO and founder, GraphPad Software
hmotulsky@graphpad.com

2 Sample size for a contemplated experiment

2.1 Sample size tutorial. Unpaired t test

2.1.1 Introduction to sample size calculations

Many experiments and clinical trials are run with too few subjects. An underpowered study is a wasted effort because even substantial treatment effects are likely to go undetected. Even if the treatment substantially changed the outcome, the study would have only a small chance of finding a "statistically significant" effect.

When planning a study, therefore, you need to choose an appropriate sample size. The required sample size depends on your answers to these questions (which will be discussed in depth later in this example):

© 2011 GraphPad Software, Inc.

- How much scatter do you expect?
- How willing are you to risk mistakenly finding a difference by chance?
- How big a difference you are looking for?
- How sure do you need to be that your study will detect a difference, if it exists? In other words, how much statistical power do you need?

The first question requires that you estimate the standard deviation you expect to see. If you can't estimate the standard deviation, you can't compute how many subjects you will need. If you expect lots of scatter, it is harder to discriminate real effects from random noise, so you'll need lots of subjects.

The second question is answered with your definition of statistical significance. Almost all investigators choose the 5% significance level, meaning that P values less than 0.05 are considered to be "statistically significant". If you choose a smaller significance level (say 1%), then you'll need more subjects.

The third and fourth questions are trickier. Everyone would prefer to plan a study that can detect very small differences, but this requires a large sample size. And everyone wants to design a study with lots of power, so it is quite certain to return a "statistically significant" result if the treatment actually works, but this too requires lots of subjects. Rather than asking you to answer those last two questions, StatMate presents results in a table so you see the tradeoffs between sample size, power, and the effect size you can detect. You can look at this table, consider the time, expense and risk of your experiment, and decide on an appropriate sample size. Note that StatMate does not directly answer the question "how many subjects do I need?" but rather answers the related question "if I use N subjects, what information can I learn?". This approach to sample size calculations is recommended by R. A. Parker and N. G. Berman (Am. Statistician 57:166-170, 2003).

In some cases, StatMate's calculations may convince you that it is impossible to find what you want to know with the number of subjects you are able to use. This can be very helpful. It is far better to cancel such an experiment in the planning stage, than to waste time and money on a futile experiment that won't have sufficient power. If the experiment involves any clinical risk or expenditure of public money, planning such a study can even be considered unethical.

Next: Begin the example.

2.1.2 1. Introduction to the example

Using StatMate is entirely self-explanatory, and this example (and this entire help system) emphasizes the logic behind sample size calculations rather than the mechanics of using StatMate.

First a bit of background. Platelets (the small cells in blood that help with clotting) have $alpha_2$-adrenergic receptors. Epinephrine (adrenaline) in the blood binds to these receptors, which make the platelets more sticky, so they aggregate, helping blood clot.

© 2011 GraphPad Software, Inc.

Hypertension (high blood pressure) is a complicated disease (or group of diseases) but there is lots of evidence that the adrenergic signalling system might be abnormal in hypertension.

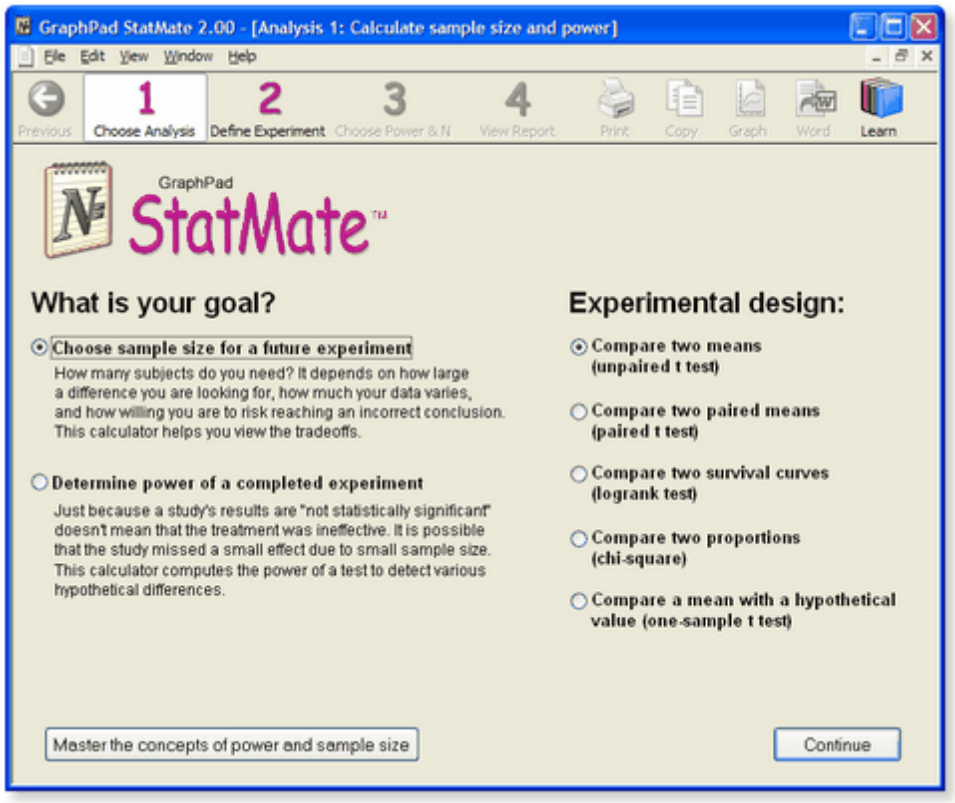
We were most interested in the heart, blood vessels, kidney and brain but obviously couldn't access those tissues in people. Instead, we decided to count the number of alpha$_2$-adrenergic receptors on platelets, and compare people with and without hypertension. How many subjects should we use?

Next: Start StatMate.

2.1.3 2. Start StatMate

In step 1, you choose the kind of analysis you want StatMate to perform by answering two questions.

- What is your goal? For this example, we want to compute sample size for a new study. Later, we'll go through an example of determining the power of a completed experiment.
- What is your experimental design? In this example, we plan to compare the mean of two groups using an unpaired t test.



Next: Estimate the SD.

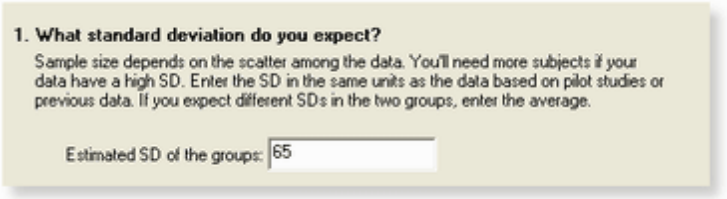
© 2011 GraphPad Software, Inc.

2.1.4 3. Estimate the SD

On step 2, StatMate asks us to enter the standard deviation (SD) we expect to see.

If you expect to see a lot of scatter among values (high SD), then you'll need a large sample size. On the other hand, if you expect to see very little scatter among the values (low SD), you won't need so many subjects. There simply is no way to estimate the sample size you need unless you can estimate the amount of scatter you expect to see.

For this example, we use data from other studies of platelet alpha$_2$-adrenergic receptors done for different reasons. These studies show that the average number of receptors per platelet is about 250 and the standard deviation is about 65. Why so high? It probably is a combination of biological variation, experimental error in counting the receptors, and experimental error in counting the platelets. Using a prior study to obtain an SD value is usually better than using a pilot study.



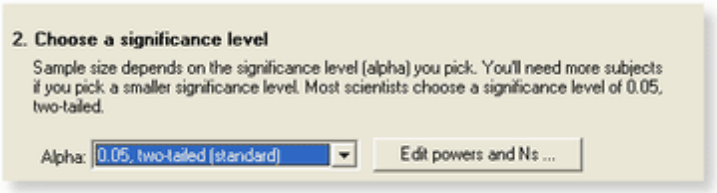
Next: Choose a significance level.

2.1.5 4. Choose a significance level

StatMate next asks you to choose a significance level, α (alpha). This is the P value below which you deem results "statistically significant" and thus reject the null hypothesis.

Ideally, you should set a value for α based on the consequence of making a Type I error (concluding that a difference is statistically significant when in fact the treatment is ineffective and the difference was due to random variation). If the consequences of making a Type I error are serious, set α to a smaller value, and you'll need more subjects. If the consequences of a Type I error are minor, set α to a higher value so you can get by with fewer subjects.

Most investigators always set alpha to 0.05, two-tailed, and we'll do the same. Learn more about the meaning of alpha, about one- and two-tailed tests, and about how to choose an appropriate value for alpha.



Note: The button "Edit powers

© 2011 GraphPad Software, Inc.

and Ns..." gives you a chance to revise the list of powers and sample sizes used on the next screen. In almost all cases, the defaults will be just fine and we'll keep the defaults for this example.

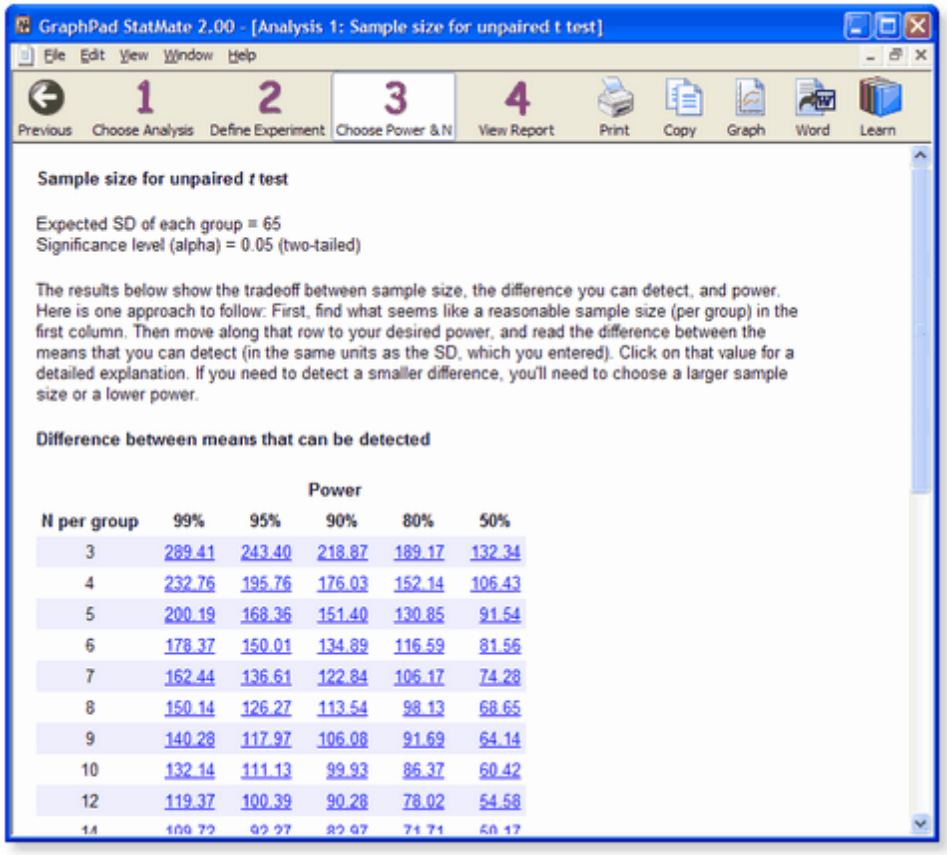
Next: View tradeoffs of sample size, power, and detectable difference.

2.1.6 5. View tradeoff of sample size and power

Some programs would ask you at this point how much statistical power you desire and how large an effect size you are looking for. The program would then tell you what sample size you need. The problem with this approach is that you often can't say how much power you want, or how large an effect size you are looking for. You want to design a study with very high power to detect very small effects, with a very strict definition of statistical significance. But doing so requires lots of subjects, more than you can afford. What you need to do is review the possibilities and understand the tradeoffs.

StatMate presents a table showing the tradeoff between sample size, power, and the effect size that you will be able to detect as statistically significant.

© 2011 GraphPad Software, Inc.



The table presents lots of information.

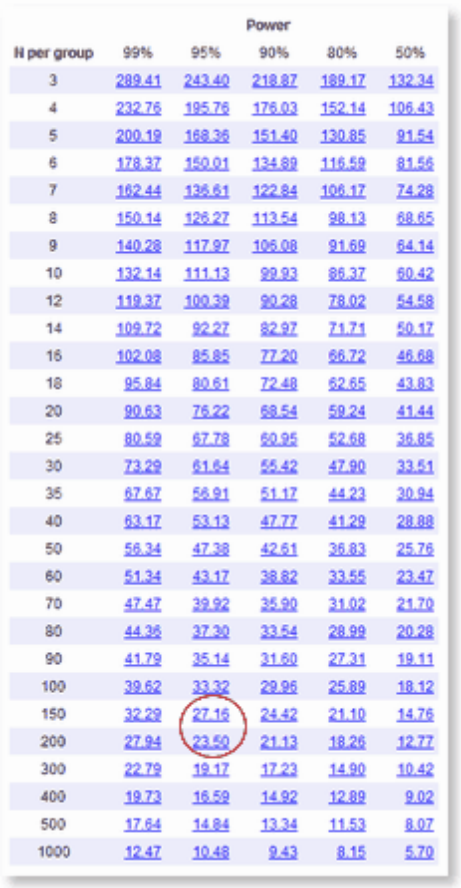
- Each row in the table represents a potential sample size you could choose. The numbers refer to the sample size in each group. Learn how to plan for unequal sample sizes.
- Each column represents a different power. The power of a study is the answer to this question: If the true difference between means equals the tabulated value, what is the chance that an experiment of the specified sample size would result in a P value less than α, and thus be deemed "statistically significant". Learn about choosing an appropriate power. You can change the list of powers used by clicking "Edit Powers and Ns..." in step 2.
- Since this example is for a unpaired t test, each value in the table is a difference between the means of the two groups, expressed in the same units as the SD you entered on step 2. In this example, the data are expressed as number of receptors per platelet.

Now comes the hard part. You need to look over this table and find a satisfactory combination of sample size, power, and a difference you can detect. We'll show several approaches.

Next: Approach A. Design a definitive study, even if it takes a lot of subjects.

© 2011 GraphPad Software, Inc.

2.1.7 6. Tradeoffs of sample size and power. Approach A



	Power				
N per group	99%	95%	90%	80%	50%
3	289.41	243.40	218.87	189.17	132.34
4	232.76	195.76	176.03	152.14	106.43
5	200.19	168.36	151.40	130.85	91.54
6	178.37	150.01	134.89	116.59	81.56
7	162.44	136.61	122.84	106.17	74.28
8	150.14	126.27	113.54	98.13	68.65
9	140.28	117.97	106.08	91.69	64.14
10	132.14	111.13	99.93	86.37	60.42
12	119.37	100.39	90.28	78.02	54.58
14	109.72	92.27	82.97	71.71	50.17
16	102.08	85.85	77.20	66.72	46.68
18	95.84	80.61	72.48	62.65	43.83
20	90.63	76.22	68.54	59.24	41.44
25	80.59	67.78	60.95	52.68	36.85
30	73.29	61.64	55.42	47.90	33.51
35	67.67	56.91	51.17	44.23	30.94
40	63.17	53.13	47.77	41.29	28.88
50	56.34	47.38	42.61	36.83	25.76
60	51.34	43.17	38.82	33.55	23.47
70	47.47	39.92	35.90	31.02	21.70
80	44.36	37.30	33.54	28.99	20.28
90	41.79	35.14	31.60	27.31	19.11
100	39.62	33.32	29.96	25.89	18.12
150	32.29	27.16	24.42	21.10	14.76
200	27.94	23.50	21.13	18.26	12.77
300	22.79	19.17	17.23	14.90	10.42
400	19.73	16.59	14.92	12.89	9.02
500	17.64	14.84	13.34	11.53	8.07
1000	12.47	10.48	9.43	8.15	5.70

In this approach, we want to plan a fairly definitive study and have plenty of time and funding.

What power should we use? We chose the traditional significance level of 5%. That means that if there truly is no difference in mean receptor number between the two groups, there still is a 5% probability that we'll happen to get such a large difference between the two groups that we'll end up calling the difference statistically significant. We also want a 5% probability of missing a true difference. So we'll set the power equal to 100%-5%, or 95%.

What size difference are we looking for? While we haven't yet studied people with hypertension, we know that other studies have found that the average number of receptors per platelet is about 250. How large a difference would we care about? Let's say we want to find a 10% difference, so a difference between means of 25 receptors per cell.

Look down the 95% power column, to find values near 25. This value is about half way between N=150 and N=200, so we need about 175 subjects in each group.

That is a lot of subjects. The next page, shows an approach that justifies fewer subjects.

Next: Use an approach that justifies fewer subjects.

© 2011 GraphPad Software, Inc.

2.1.8 7. Tradeoffs of sample size and power. Approach B

	Power				
N per group	99%	95%	90%	80%	50%
3	289.41	243.40	218.87	189.17	132.34
4	232.76	195.76	176.03	152.14	106.43
5	200.19	168.36	151.40	130.85	91.54
6	178.37	150.01	134.89	116.59	81.56
7	162.44	136.61	122.84	106.17	74.28
8	150.14	126.27	113.54	98.13	68.65
9	140.28	117.97	106.08	91.69	64.14
10	132.14	111.13	99.93	86.37	60.42
12	119.37	100.39	90.28	78.02	54.58
14	109.72	92.27	82.97	71.71	50.17
16	102.08	85.85	77.20	66.72	46.68
18	95.84	80.61	72.48	62.65	43.83
20	90.63	76.22	68.54	59.24	41.44
25	80.59	67.78	60.95	52.68	36.85
30	73.29	61.64	55.42	47.90	33.51
35	67.67	56.91	51.17	44.23	30.94
40	63.17	53.13	47.77	41.29	28.88
50	56.34	47.38	42.61	36.83	25.76
60	51.34	43.17	38.82	33.55	23.47
70	47.47	39.92	35.90	31.02	21.70
80	44.36	37.30	33.54	28.99	20.28
90	41.79	35.14	31.60	27.31	19.11
100	39.62	33.32	29.96	25.89	18.12
150	32.29	27.16	24.42	21.10	14.76
200	27.94	23.50	21.13	18.26	12.77
300	22.79	19.17	17.23	14.90	10.42
400	19.73	16.59	14.92	12.89	9.02
500	17.64	14.84	13.34	11.53	8.07
1000	12.47	10.48	9.43	8.15	5.70

In this approach, we want a smaller sample size, and are willing to make compromises for it.

What power should we use? It is pretty conventional to use 80% power. This means that if there really is a difference of the tabulated size, there is a 80% chance that we'll obtain a "statistically significant" result ($P<0.05$) when we run the study, leaving a 20% chance of missing a real difference of that size.

What size difference are we looking for? While we haven't yet studied people with hypertension, we know that other studies have found that the average number of receptors per platelet is about 250. How large a difference would we care about? In approach A, we looked for a 10% difference. Let's look instead for a 20% difference, so a difference between means of 50 receptors per cell.

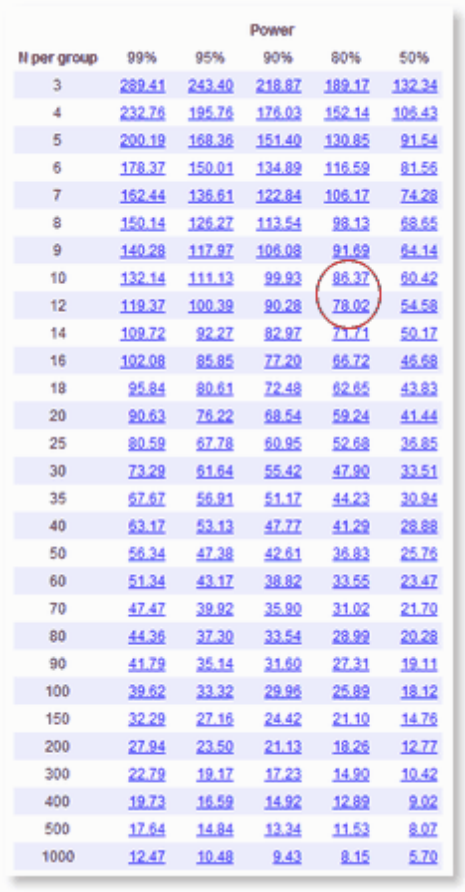
Look down the 80% power column, to find values near 50. This value is about half way between N=25and N=30, so we need about 28 subjects in each group.

That still seems like a lot. Can we justify even fewer?

Next: Justify even fewer subjects.

© 2011 GraphPad Software, Inc.

2.1.9 8. Tradeoffs of sample size and power. Approach C



	Power				
N per group	99%	95%	90%	80%	50%
3	289.41	243.40	218.87	189.17	132.34
4	232.76	195.76	176.03	152.14	106.43
5	200.19	168.36	151.40	130.85	91.54
6	178.37	150.01	134.89	116.59	81.56
7	162.44	136.61	122.84	106.17	74.28
8	150.14	126.27	113.54	98.13	68.65
9	140.28	117.97	106.08	91.69	64.14
10	132.14	111.13	99.93	86.37	60.42
12	119.37	100.39	90.28	78.02	54.58
14	109.72	92.27	82.97	71.71	50.17
16	102.08	85.85	77.20	66.72	46.68
18	95.84	80.61	72.48	62.65	43.83
20	90.63	76.22	68.54	59.24	41.44
25	80.59	67.78	60.95	52.68	36.85
30	73.29	61.64	55.42	47.90	33.51
35	67.67	56.91	51.17	44.23	30.94
40	63.17	53.13	47.77	41.29	28.88
50	56.34	47.38	42.61	36.83	25.76
60	51.34	43.17	38.82	33.55	23.47
70	47.47	39.92	35.90	31.02	21.70
80	44.36	37.30	33.54	28.99	20.28
90	41.79	35.14	31.60	27.31	19.11
100	39.62	33.32	29.96	25.89	18.12
150	32.29	27.16	24.42	21.10	14.76
200	27.94	23.50	21.13	18.26	12.77
300	22.79	19.17	17.23	14.90	10.42
400	19.73	16.59	14.92	12.89	9.02
500	17.64	14.84	13.34	11.53	8.07
1000	12.47	10.48	9.43	8.15	5.70

Let's say that our budget (or patience) only lets us do a study with 11 subjects in each group. How much information can we obtain? Is such a study worth doing?

With a small study, we know we are going to have to make do with a moderate amount of power. But the rightmost column is for a power of only 50%. That means that even if the true effect is what we hypothesize, there is only a 50% chance of getting a "statistically significant" result. In that case, what's the point of doing the experiment? We want more power than that, but know we can't have a huge amount of power without a large sample size. So let's pick 80% power, which is pretty conventional. This means that if there really is a difference of the tabulated size, there is a 80% chance that we'll obtain a "statistically significant" result ($P<0.05$) when we run the study, leaving a 20% chance of missing a real difference.

If we look down the 80% power column, in the N=11 row, we find that we can detect a difference of 86.4. We already know that the mean number of alpha2-adrenergic receptors is about 250, so a sample size of 12 in each group has 80% power to detect a 35% (86.4/250) change in receptor number.

This sample size analysis has helped us figure out what we can hope to learn given the sample size we already chose. Now we can decide whether the experiment is even worth doing. Different people would decide this differently. But some would conclude much smaller differences might be biologically important, and that if we can only detect a huge change of 35%, and even that with only 80% power, it simply isn't even worth doing the experiment.

Next: How can all three of these approaches be correct?

© 2011 GraphPad Software, Inc.

2.1.10 9. How can all three approaches be correct?

If you specify exactly what power you want, and how large an effect you want to detect, StatMate can tell you exactly how many subjects you need.

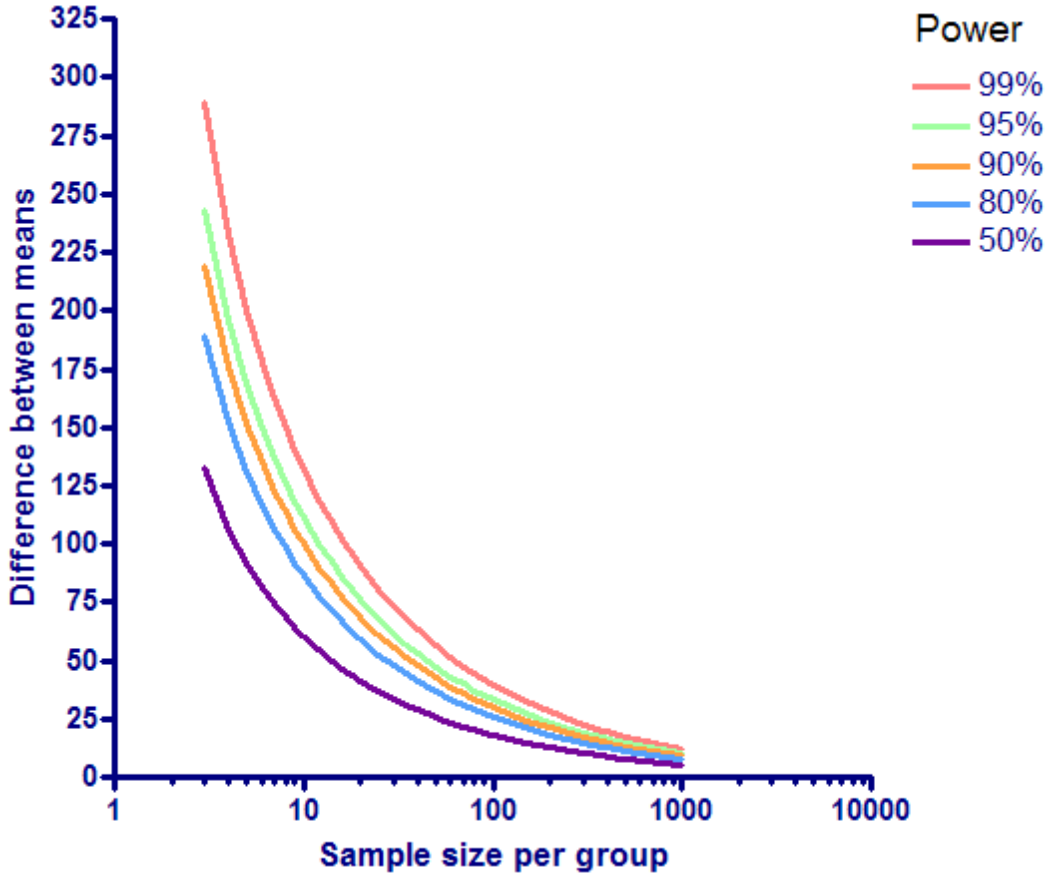
But generally, you won't be sure about what power you want (or are willing to accept) or how large an effect you want to detect. Therefore, you can justify almost any sample size. It depends on how large a effect you want to find, how sure you want to be to find it (power), and how willing you are to mistakenly find a significant difference (alpha). So there is no one right answer. It depends on why you are looking for a difference and on the cost, hassle and risk of doing the experiment.

Next: Graph the relationship between N and power.

2.1.11 10. Graph the relationship between N and power

StatMate does not create graphs itself. But if you own a copy of GraphPad Prism version 4.01 (Windows) or 4.0b (Mac) or later, just click the graph button to make an instant graph in Prism. Each curve is for a different power, and shows the relationship between the sample size you could choose for each group (X) and the difference you would then detect as "significant" (Y).

© 2011 GraphPad Software, Inc.



As you go from left to right, the curves go down. This makes sense -- if you use more subjects (collect more data), then you'll be able to reliably detect smaller differences. Each curve is for a different power. If you choose a higher power, the curve shifts to the right. This also makes sense -- if you want more power (to have less chance of missing a real difference), then you'll need more subjects.

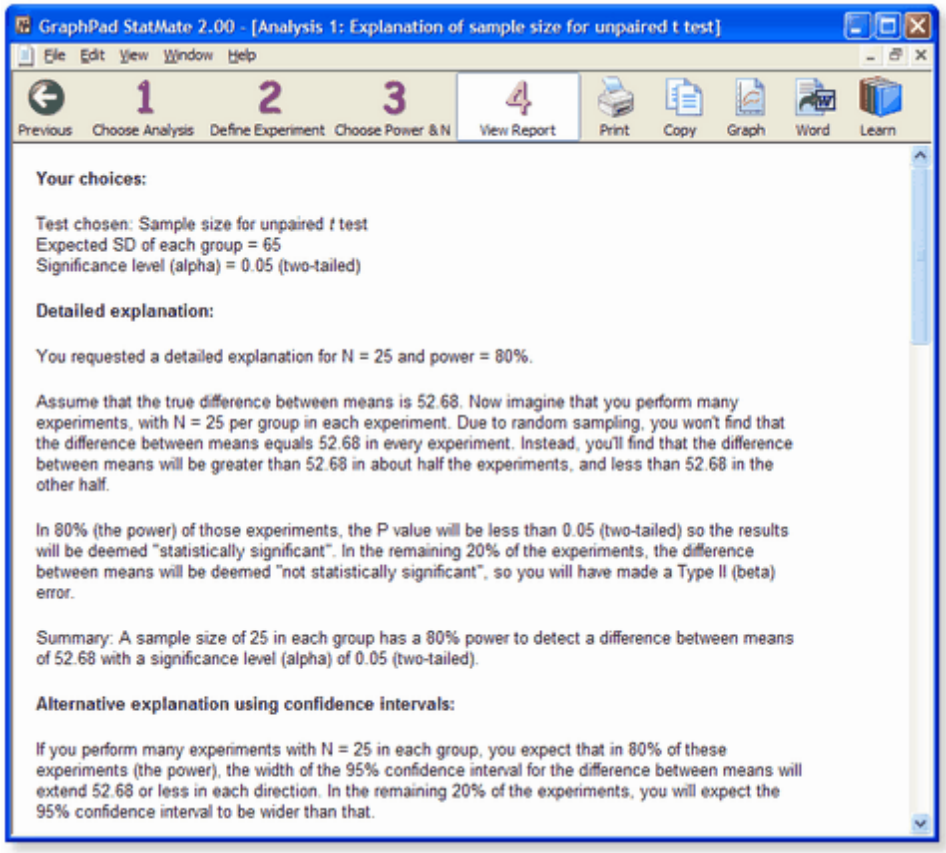
If you don't own Prism, you can copy and paste the table into another program for graphing.

Next: StatMate puts all the results into a single report.

2.1.12 11. StatMate's report

We'll choose the sample size chosen in approach B. In step 3 of StatMate, each value is a link. We click the link in the 80% power column and the N=25 row (52.68), and StatMate presents a complete report. You can send the entire report to Word with a click of a button (Windows only), or via copy and paste.

© 2011 GraphPad Software, Inc.



The screen shot above shows the first two of four sections of the report: a reiteration of your choices, and a detailed interpretation. The report then shows the entire table of tradeoffs (which you have already seen) and a discussion of when it makes sense to design studies with unequal sample sizes (which is explained here).

Note: This ends the example. You should now be ready to use StatMate on your own, or browse the rest of the help screens.

2.2 Defining experiments for sample size calculations

2.2.1 Unpaired t test

An unpaired t test compares measurements in two groups of subjects. Sample size calculations are based on your estimate of the standard deviation you expect to see. A t test compares the difference between means with the standard deviation between groups (accounting for sample size), so there simply is no way to estimate the sample size you need unless you can estimate the amount of scatter you expect to see.

Estimate the standard deviation by looking at previous studies. For example, if your study

© 2011 GraphPad Software, Inc.

is testing how a drug changes blood pressure you can estimate the standard deviation of blood pressure from studies that test other drugs. If you've done a pilot study, you can use the SD from it but pilot studies are usually not needed and the SD you obtain from a pilot study may not be very accurate.

You also need to choose a significance level (alpha). Most scientists choose a significance level of 5%, meaning that a P value less than 0.05 is deemed "statistically significant".

2.2.2 Paired t test

A paired t test is used when you have before and after measurements in each subject, or measurements in paired subjects. Therefore you must estimate both the amount of scatter and the degree to which the before and after measurements are correlated. StatMate offers two ways to enter this information.

- Enter estimates for both the SD you expect to see in each group and the correlation coefficient (r) you expect to see.
- Enter an estimate for the SD of the differences between pairs. Imagine you had already collected data. For each pair, compute the difference. Then compute the SD of this list of differences.

Estimate the standard deviation by looking at previous studies. For example, if your study is testing how a drug changes blood pressure you can estimate the standard deviation of blood pressure from studies that test other drugs. If you've done a pilot study, you can use the SD from it but pilot studies are usually not needed and the SD you obtain from a pilot study may not be very accurate.

You also need to choose a significance level (alpha). Most scientists choose a significance level of 5%, meaning that a P value less than 0.05 is deemed "statistically significant".

2.2.3 Comparing two proportions (chi-square)

You'll often do studies where you compare two groups and the outcome is either/or (binary). The results for each group are expressed as a proportion, and you want to compare the proportions between groups.

The necessary sample size depends, in part, on the proportion "success" (an arbitrary label for one of the two alternative outcomes) in the control group. You'll need the fewest subjects when the control group has a proportion success near 0.50. StatMate asks you to enter an estimated value for the control group.

You may be surprised that you have to tell StatMate whether you expect to see an increase or a decrease. This is necessary because the power to detect a difference between two

© 2011 GraphPad Software, Inc.

proportions depends not only on the difference (or ratio) of the two proportions, but also on their actual values. The power to find a difference between 0.2 and 0.4 is not the same as the power to find a difference between 0.4 and 0.6 (same difference) or between 0.4 and 0.8 (same ratio). So it is not enough to tell StatMate that you expect a control proportion of 0.4 – you also need to tell it whether you expect to see an increase or decrease.

Note: The sample size required to find an increase from 0.2 to 0.4 is exactly the same as the sample size needed to find a decrease from 0.4 to 0.2. In one case you enter the control proportion as 0.2 and tabulate an increase. In the other case you enter the control proportion as 0.4 and tabulate a decrease.

You can express the results as a difference between the two proportions or as their ratio. The results are equivalent either way, so choose the format that you find easier to interpret.

You also need to choose a significance level (alpha). Most scientists choose a significance level of 5%, meaning that a P value less than 0.05 is deemed "statistically significant".

2.2.4 One-sample t test

The one-sample t test is used to test whether the mean of a group differs significantly from some hypothetical value (often 0.0, 1.0 or 100). The test depends on the standard deviation of the values. If the data are very scattered (high SD) you'll need a larger sample size (or suffer from lower power) than if the values have a low SD. There simply is no way to estimate the sample size you need unless you can estimate the amount of scatter you expect to see.

If possible, estimate the standard deviation by looking at previous studies. If you've done a pilot study, you can use the SD from it but pilot studies are usually not needed and the SD you obtain from a pilot study may not be very accurate.

You also need to choose a significance level (alpha). Most scientists choose a significance level of 5%, meaning that a P value less than 0.05 is deemed "statistically significant".

Sample size calculations for a one-sample t test are based on the same principles as those for an unpaired (two-sample) t test. Review the example which explains sample size determination for the unpaired t test in detail.

© 2011 GraphPad Software, Inc.

2.2.5 Suvival analysis

Survival analysis is used to analyze studies that compare the time until a one-time event occurs. This event is often death, which explains the name "survival analysis". But survival analysis can be used in any situation where the outcome you are tracking is all-or-none and can only happen once. So it could be time until a tumor metastasizes, or time until first heart attack. The event doesn't have to be dire. You could track time until a student graduates, or time until a patient recovers from a coma.

StatMate computes sample size based on these simplifying assumptions:

- The hazard ratio is the same at all times. For example, treated patients have half the death rate at all times as control subjects.
- All subjects are followed the same length of time.
- Few subjects drop out along the way.

Sample size depends on the proportion of subjects who will be event-free at the end of the study. If the event only occurs to a small fraction of subjects, of course you'll need to start with a larger sample size.

For the group (usually the untreated control group) for which you think the event will occur more rapidly, estimate the proportion of subjects who will be event-free (will still be alive) at the end of the study.

Also choose how you want to tabulate the results. StatMate gives you three choices:

- Increase in proportion. This the difference in the proportion of event-free (surviving) subjects. You entered your estimate of the proportion of event-free subjects in one group. The other group will have a proportion of event-free (surviving) subjects equal to that proportion plus the difference that StatMate calculates.
- Hazard ratio. The hazard ratio is the ratio of the slopes of the survival curves. A hazard ratio of 2 means the control group dies at twice the rate of the treated group. Another way to look at the hazard ratio is the median survival time in one group divided by the median survival curve in the other group. So a hazard ratio of 2 means that the median survival of the treated group is twice that of the control group.
- Hazard ratio (larger is worse). Use this alternative choice if you want a hazard ratio of 2 to mean the treated group dies at twice the rate as the control group.

You also need to choose a significance level (alpha). Most scientists choose a significance level of 5%, meaning that a P value less than 0.05 is deemed "statistically significant".

2.3 Questions about sample size calculations

2.3.1 Do I need to decide sample size in advance?

Yes, you need to choose a sample size before starting your study, and StatMate can help you figure out what size is most appropriate.

© 2011 GraphPad Software, Inc.

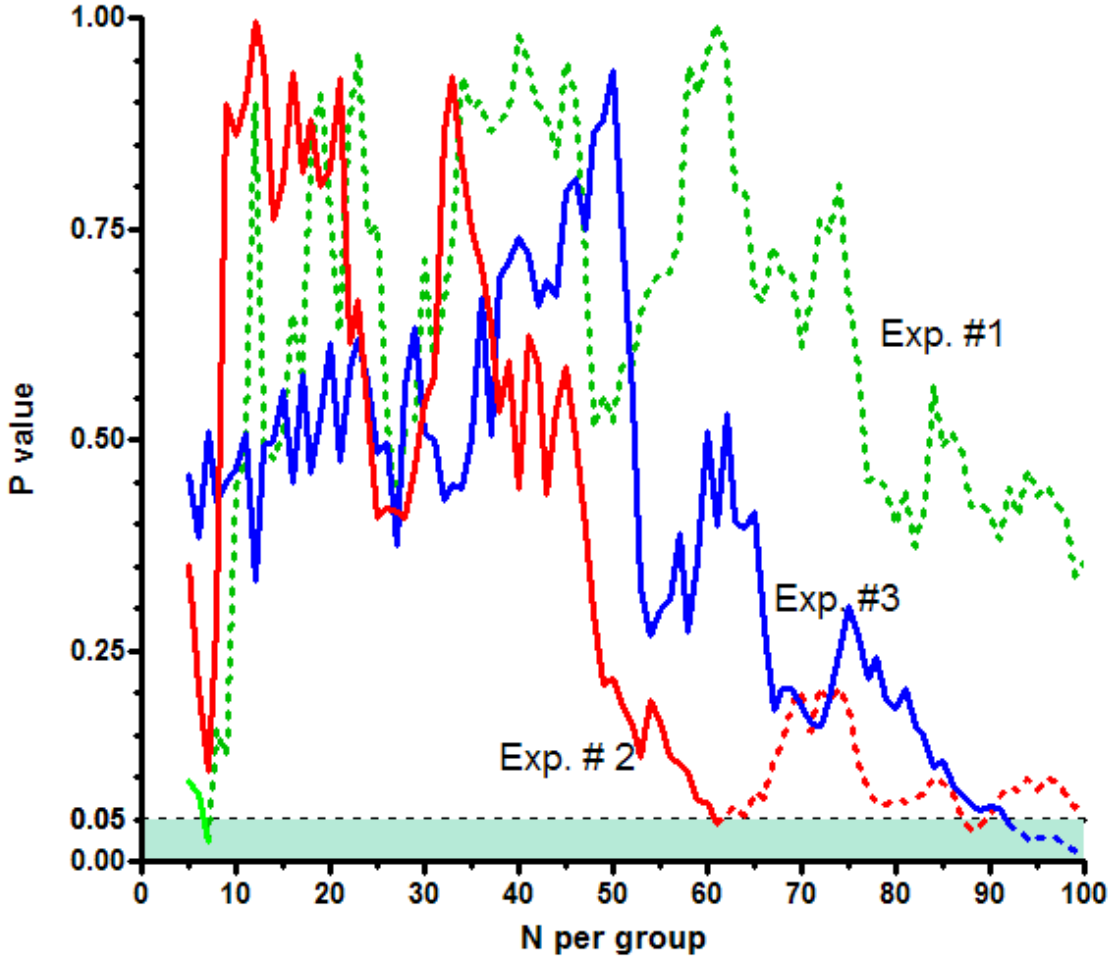
You might ask: Why bother with calculating sample size before the study starts? Why not do the analyses as you collect data? If you don't have a significant result, then collect some more data, and reanalyze. If you do obtain a statistically significant result, stop the study.

The problem with this approach is that you'll keep going if you don't like the result, but stop if you do like the result. The consequence is that the chance of obtaining a "significant" result if the null hypothesis were true is a lot higher than 5%.

The graph below illustrates this point via simulation. We simulated data by drawing values from a Gaussian distribution (mean=40, SD=15, but these values are arbitrary). Both groups were simulated using exactly the same distribution. We picked N=5 in each group and computed an unpaired t test and recorded the P value. Then we added one subject to each group (so N=6) and recomputed the t test and P value. We repeated this until N=100 in each group. Then we repeated the entire simulation three times. These simulations were done comparing two groups with identical population means. So any "statistically significant" result we obtain must be a coincidence -- a Type I error.

The graph plots P value on the Y axis vs. sample size (per group) on the X axis. The blue shaded area at the bottom of the graph shows P values less than 0.05, so deemed "statistically significant".

© 2011 GraphPad Software, Inc.



The green curve shows the results of the first simulated set of experiments. It reached a P value less than 0.05 when N=7, but the P value is higher than 0.05 for all other sample sizes. The red curve shows the second simulated experiment. It reached a P value less than 0.05 when N=61 and also when N=88 or 89. The blue curve is the third experiment. It has a P value less than 0.05 when N=92 to N=100.

If we followed the sequential approach, we would have declared the results in all three experiments to be "statistically significant". We would have stopped when N=7 in the green experiment, so would never have seen the dotted parts of its curve. We would have stopped the red experiment when N=6, and the blue experiment when N=92. In all three cases, we would have declared the results to be "statistically significant".

Since these simulations were created for values where the true mean in both groups was identical, any declaration of "statistical significance" is a Type I error. If the null hypothesis is true (the two population means are identical) we expect to see this kind of Type I error in 5% of experiments (if we use the traditional definition of alpha=0.05 so P values less than 0.05 are declared to be significant). But with this sequential approach, all three of our experiments resulted in a Type I error. If you extended the experiment long enough

© 2011 GraphPad Software, Inc.

(infinite N) all experiments would eventually reach statistical significance. Of course, in some cases you would eventually give up even without "statistical significance". But this sequential approach will produce "significant" results in far more than 5% of experiments, even if the null hypothesis were true, and so this approach is invalid.

Conclusion: It is important that you choose a sample size and stick with it. You'll fool yourself if you stop when you like the results, but keep going when you don't.

Note: There are some special statistical techniques for analyzing data sequentially, adding more subjects if the results are ambiguous and stopping if the results are clear. Look up 'sequential medical trials' in advanced statistics books to learn more.

2.3.2 Do I need to run a pilot study before computing sample size?

It is not usually necessary to do a pilot study first.

Pilot studies are often done to get an estimate of the standard deviation (for data to be analyzed by t test). But a small pilot study won't give you a very accurate value for the standard deviation. If possible, use data from other studies. Perhaps the same variable was measured for other reasons.

If you are comparing proportions or survival curves, you can't do a sample size analysis without first estimating the percentage "success" and "failure" (or survival) in the two groups. A small pilot study won't give you very accurate values. For example, say you do a pilot study with ten subjects, and 3 have one outcome and 7 the other. So the proportion "success" in that sample is 30%. The 95% confidence interval for this proportion ranges from 7% to 65%. You need a much larger study -- hardly a pilot study -- to get narrow confidence intervals.

2.3.3 How accurate is a SD from a pilot experiment?

The standard deviation you determine from a pilot experiment may not be as accurate as you think. By chance, the values you measure in your subjects may be more (or less) scattered than the population. This uncertainty can be expressed as the 95% confidence interval of the SD, which is easy to compute.

To compute the confidence interval, we assume that the subjects you sample are randomly selected from (or at least representative of) a larger population whose values distributed according to a Gaussian distribution. You can compute the confidence interval of a standard deviation using the equation shown at the bottom of this page, or (much easier) using one of the free web-based GraphPad QuickCalcs (link to the web site).

© 2011 GraphPad Software, Inc.

The confidence interval depends on sample size (N), as shown in this table which shows the 95% confidence interval of a standard deviation (SD) as a function of sample size (N). If you assume that your values were randomly sampled from a Gaussian population, you can be 95% sure that this interval contains the true population standard deviation.

N	95% CI of SD
3	0.52*SD to 6.29*SD
5	0.60*SD to 2.87*SD
10	0.69*SD to 1.83*SD
25	0.78*SD to 1.39*SD
50	0.84*SD to 1.25*SD
100	0.88*SD to 1.16*SD
500	0.94*SD to 1.07*SD
1000	0.96*SD to 1.05*SD

Most people are surprised at how wide the intervals are, especially for small sample sizes. The standard deviation you determine from a handful of subjects may be quite far from the true population standard deviation. If you rely upon this SD to do sample size computations, your estimate of sample size may be inaccurate. To bypass this problem, make sure that your estimate of SD is based on more than a few values.

Why are the confidence intervals asymmetrical? Since the SD is always a positive number, the lower confidence limit can't be less than zero. This means that the upper confidence interval usually extends further above the sample SD than the lower limit extends below the sample SD. With small samples, this asymmetry is quite noticeable.

If you want to compute these confidence intervals yourself, use these Excel equations:

Lower limit: =SD*SQRT((N-1)/CHINV((alpha/2),N-1))
Upper limit: =SD*SQRT((N-1)/CHINV(1-(alpha/2),N-1))

Alpha is 0.05 for 95% confidence, 0.01 for 99% confidence, and N is the number of values used to compute the standard deviation.

2.3.4 What values of alpha and power should I pick?

The values you pick for alpha and power depend on the experimental setting, and on the consequences of making a Type I or Type II error (review the difference).

Let's consider three somewhat contrived examples. All three examples involve a screening test to detect compounds that are active in your system. In this context, a Type I error occurs when you conclude that a drug is effective, when it really is not. A Type II error occurs when you conclude that a drug is ineffective, when it fact it is effective. But the consequences of making a Type I or Type II error depend on the context of the experiment. Let's consider four situations.

- Situation A. You are screening drugs from a library of compounds, chosen with care.

© 2011 GraphPad Software, Inc.

You know that some of the "hits" will be false-positives (Type I error) so plan to test all those "hits" in another assay. Therefore, the consequence of a Type I error is that you need to retest that compound. You don't want to retest too many compounds, so can't make alpha huge. But it might make sense to set it to a fairly high value, perhaps 0.10. A Type II error occurs when you conclude that a drug has no statistically significant effect, when in fact the drug is effective. The consequences of a Type II error are more serious than the consequences of a Type I error. Once the drug is deemed to have a "not significant" effect, it won't be tested again. You've picked the library of compounds with some care, and don't want to mistakenly overlook a great drug. Summary: high power, high alpha.

- Situation B. You are in a hurry, so want to avoid retesting the "hits". Instead, you'll use the results of this experiment -- the list of hits and misses -- to do a structure-activity relationship, which will then be used to come up with a new list of compounds for the chemists to synthesize. This will be a expensive and time-consuming task, so a lot is riding on this experiment, which won't be repeated. In this case, the consequences of both a Type I and Type II error are pretty bad, so you set alpha to a small value (say 0.01) and power to a large value (perhaps 99%). Choosing these values means you'll need a larger sample size, but the cost is worth it here. Summary: high power, low alpha.

- Situation C. You are testing a screening assays, rather than screening for drugs. Before beginning a big screening project, you want to have some idea what fraction of compounds from your library will be "hits". There are no bad consequences to either Type I or Type II errors, since you just want to get an idea about how well your assay system works. This means you can choose a small sample size that results in low power and high alpha. Summary: low power, high alpha.

Note: In all three situations, sample size calculations are used to determine the number of times (replicates) you should test each drug. It is not trying to help you figure out how many drugs to test.

These examples are a bit contrived, and certainly should not be used to help anyone design drug screening assays. But these examples make the point that you should choose values for alpha and power after carefully considering the consequences of making a Type I and Type II error. These consequences depend on the scientific context of your experiment.

© 2011 GraphPad Software, Inc.

2.3.5 How do I choose an effect size?

A final reason to think about a pilot study is to figure out what effect you expect to see. A power analysis always determines the power to detect a certain effect size. But this effect size does not need to come from a pilot study, and probably shouldn't. You want to determine the power to detect some effect size that you would consider to be scientifically or clinically of interest. This can be determined from the context of the experiment. The sample size calculation does not have to be for the effect size you expect to see. In fact, if you really knew the effect size you expect to see, then there really wouldn't be any need to run the study at all. Rather it should be for the smallest effect size you would care about. For this, you don't need a pilot study.

2.3.6 Does it help to use standard definitions of small and large effect sizes?

Jacob Cohen in Statistical Power Analysis for the Behavioral Sciences (second edition) makes some recommendations for what to do when you don't know what effect size you are looking for. He limits these recommendations to the behavioral sciences (his area of expertise), and warns that all general recommendations are more useful in some circumstances than others. Here are his guidelines for an unpaired t test:

- A "small" difference between means is equal to one fifth the standard deviation.
- A "medium" effect size is equal to one half the standard deviation.
- A "large" effect is equal to 0.8 times the standard deviation.

So if you are having trouble deciding what effect size you are looking for (and therefore are stuck and can't determine a sample size), Cohen would recommend you choose whether you are looking for a "small", "medium", or "large" effect, and then use the standard definitions.

Russell Lenth (1) argues that you should avoid these "canned" effect sizes, and I agree. You must decide how large a difference you care to detect based on understanding the experimental system you are using and the scientific questions you are asking. Cohen's recommendations seem a way to avoid thinking about the point of the experiment.

Consider the example discussed in these help screens. We anticipated, based on prior studies, that the standard deviation of $alpha_2$ receptor numbers on platelets will be about 65 receptors/platelet. If we were to use Cohen's suggestions, our effect size (from which we compute sample size) would be computed just from this estimated standard deviation, without regard to the expected mean number of receptors per platelet. If we were somehow to improve the methodology to make the experiments more consistent, and thus reduce the standard deviation, we'd also lower the effect size we are looking for. This really doesn't make any sense. It makes sense to look for an effect as a fraction of the mean. It doesn't make sense to define the effect you care about scientifically to be a fraction of the standard deviation.

Note: If you do choose standard definitions of alpha (0.05), power (80%), and effect size

© 2011 GraphPad Software, Inc.

(see above), then there is no need for any calculations. If you accept those standard definitions for all your studies, then all studies need a sample size of 26 in each group to detect a large effect, 65 in each group to detect a medium effect, 400 in each group to detect a small effect. Choosing standard effect sizes is really the same as picking standard sample sizes.

1. Lenth, R. V. (2001), "Some Practical Guidelines for Effective Sample Size Determination," The American Statistician, 55, 187-193. (see an earlier draft posted online)

2.3.7 Does it make sense to use standard definitions of alpha and power?

Almost all researchers set alpha to 0.05. This means that if the null hypothesis were true, and you did lots of experiments, then in 5% of the experiments where there really is no difference at all, you'd incorrectly conclude that the difference is "statistically significant".

There isn't such a strong tradition for a particular value of power, but many researchers choose a sample size to achieve a power of 80%. That means that if the effect you want to see really exists, there is an 80% chance that your experiment will result in a statistically significant result, leaving a 20% chance of missing the real difference (beta is 0.20).

Using these standard definitions, you are four times more likely to make a Type II error (miss a real effect) than a Type I error (falsely conclude that an effect is "significant"). This makes sense if the consequences of making a Type I error are four times as bad (as costly) as making a Type II error. But the relative consequences of Type I and II errors depends on the context of the experiment. In some situations, making a Type I error has more serious consequences than making a Type II error. In other situations, the consequences of a Type II error might be worse.

While it is tempting to simply choose standard values for alpha and power, it is better to choose values based on the relative consequences of Type I and Type II errors.

2.3.8 How can simulations verify the sample size results?

A good way to verify and understand a sample size calculation is to simulate many experiments.

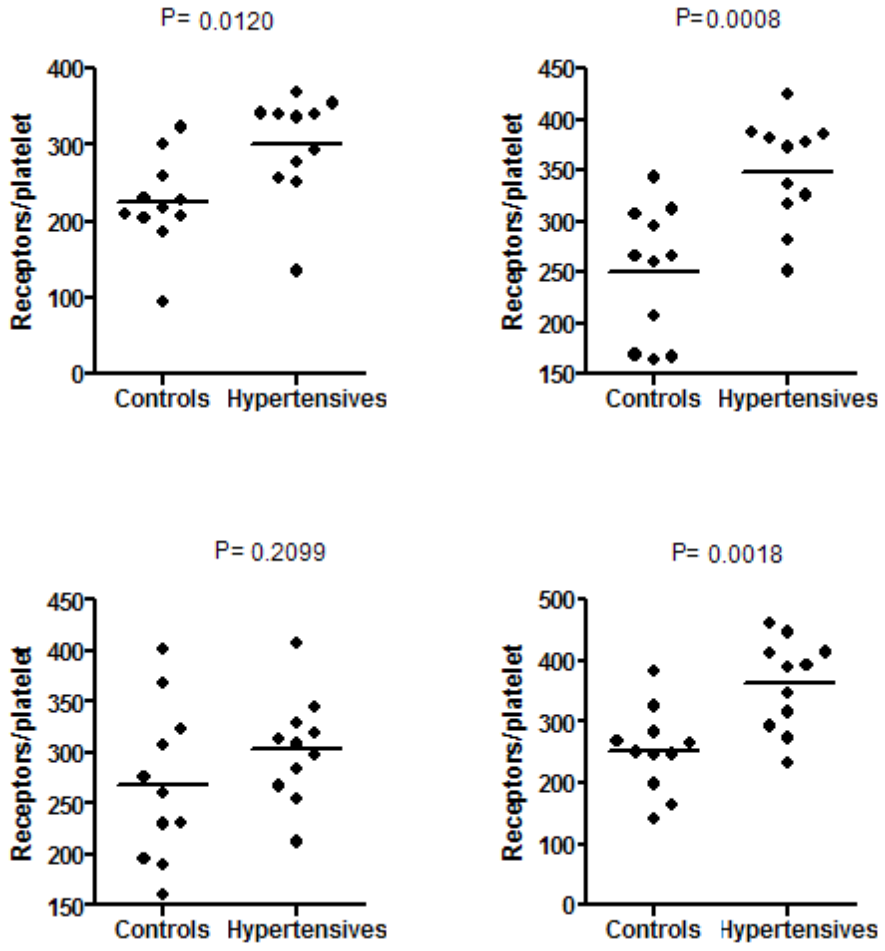
We'll base our simulations on the example shown in the example, using Approach C. In this example, we are comparing $alpha_2$-adrenergic receptor number in two groups, and have assumed, based on prior data, that the SD within both groups is 65 receptors/platelet.

© 2011 GraphPad Software, Inc.

We decided here that the smallest difference we would be interested in is a difference of 80 receptors/platelet and StatMate showed us that we can detect that effect size with 80% power if we choose 11 subjects for each group.

Let's simulate this situation. We'll assume 250 receptors/cell in one group (previous data for control subjects) and 330 receptors/cell in the other (so the difference is 80 receptors/cell as we specified). We'll set the standard deviation in each group to 65, and choose a sample size of 11 per group.

We can perform the simulations using GraphPad Prism. Prism's simulation analysis is set up to simulate curves, so here we simulated the model Y=slope*X + intercept, setting the slope equal to zero (so the X values are ignored) and setting the intercept equal to 250 (column A, control) or 330 (column B, hypertensive), and adding Gaussian error with SD=65. The graph below shows four such simulations, with the P value (from unpaired t test) superimposed.



In three of the four simulated data sets, the P value was less than 0.05, so the difference would be deemed statistically significant. In the other experiment, the P value was higher

© 2011 GraphPad Software, Inc.

than 0.05, so the difference would be deemed not statistically significant. The data were simulated from the same simulated populations as the others, so there really is a difference between the means of the two groups, But in this particular experiment, we happened to sample somewhat higher values in the controls, lower values in the hypertensive, resulting in a "not significant" difference.

We simulated this kind of experiment 1000 times (using a Prism script). The two-tailed P value was less than 0.05 in 759 of these experiments (75.9%), and was higher than 0.05 in the remaining 241 experiments (sometimes much higher, up to a P value of 0.88). In this set of one thousand simulations, therefore, the power was 75.9%. This is quite close to the predicted power of 80%.

For this example, the simulations are used simply as a way to understand the concepts of power and sample size. It is easier to use StatMate to do the calculations for you. But with some other experimental designs it might be very hard to compute sample size from equations. In such situations, simulations might be the best way to determine sample size.

2.3.9 Should I report my sample size calculations?

It is conventional to report sample size calculations in publications of clinical research. You often see in such papers statements like "We chose a sample size of 25 in each group so we would have a 80% power of detecting a mean difference of 53 receptors/platelet with a significance level of 5% (two sided), assuming a standard deviation of 65 receptors/cell".

SA Goodman and JA Berlin (Annals of Internal Medicine 121:200-206, 1994) argue that this does not help a reader interpret or evaluate your results. The results stand on their own, and the rationale for the sample size used doesn't have any impact on the interpretation of the data. The only real purpose in reporting the sample size calculations is to demonstrate that you designed the study carefully and are statistically sophisticated.

2.3.10 What if I plan to use a nonparametric test?

Nonparametric tests are used when you are not willing to assume that your data come from a Gaussian distribution. Commonly used nonparametric tests are based on ranking values from low to high, and then looking at the distribution of sum-of-ranks between groups. This is the basis of the Wilcoxon rank-sum (test one group against a hypothetical median), Mann-Whitney (compare two unpaired groups), Wilcoxon matched pairs (compare two matched groups), Kruskal-Wallis (three or more unpaired groups) and Friedman (three or more matched groups).

When calculating a nonparametric test, you don't have to make any assumption about the distribution of the values. That is why it is called nonparametric. But if you want to calculate necessary sample size for a study to be analyzed by a nonparametric test, you must make an assumption about the distribution of the values. It is not enough to say the distribution is *not* Gaussian, you have to say what kind of distribution it *is*. If you are willing to make such an assumption (say, assume an exponential distribution of values, or a uniform distribution) you should consult an advanced text or use a more advanced program to compute sample size.

© 2011 GraphPad Software, Inc.

But most people choose a nonparametric test when they don't know the shape of the underlying distribution. Without making an explicit assumption about the distribution, detailed sample size calculations are impossible. But all is not lost! Depending on the nature of the distribution, the nonparametric tests might require either more or fewer subjects. But they never require more than 15% additional subjects if the following two assumptions are true:

- You are looking at reasonably high numbers of subjects (how high depends on the nature of the distribution and test, but figure at least a few dozen).
- The distribution of values is not really unusual (doesn't have infinite tails, in which case its standard deviation would be infinitely large).

So a general rule of thumb is this: If you plan to use a nonparametric test, compute the sample size required for a t test and add 15%.

(Information from pages 76-81 of Nonparametrics : Statistical Methods Based on Ranks, by Erich L. Lehmann, Prentice Hall.)

2.3.11 What is beta?

If you read about sample size calculations, you may encounter the term "beta". Beta is simply defined to equal 100% minus power.

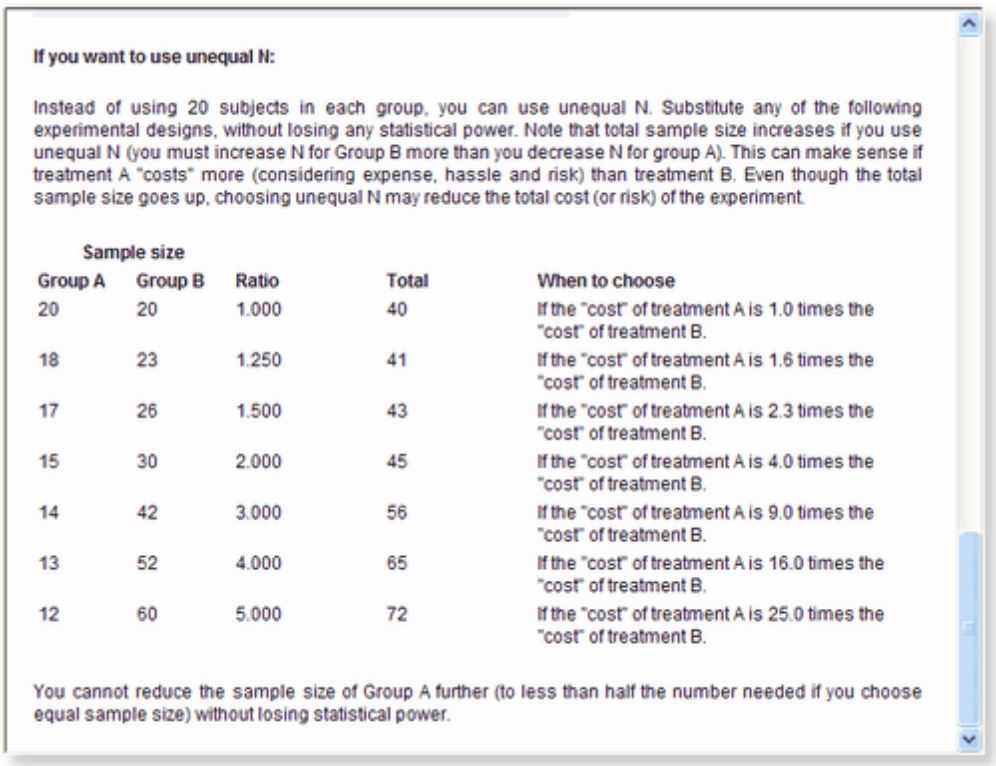
A power of 80% means that even if the effect size is really what you are looking for, your experiment has an 80% chance of obtaining a "statistically significant" result, and thus a 20% chance of obtaining a "not significant" result. In other words, there is a 20% chance that you'll make a Type II error (missing a true effect of a specified size). Beta, therefore, equals 20% or 0.20.

2.3.12 When should I plan to use unequal sample sizes?

In most cases you should plan to use the same sample size in each group. But you don't have to. Consider using unequal sample sizes when one treatment is much more expensive, difficult or risky, or when it is hard to find appropriate subjects for one group (for example when comparing patients with a rare disease to controls).

At the bottom of the detailed report, StatMate shows you alternative experimental designs that all have the same power to detect the effect size you chose..

© 2011 GraphPad Software, Inc.



If you want to use unequal N:

Instead of using 20 subjects in each group, you can use unequal N. Substitute any of the following experimental designs, without losing any statistical power. Note that total sample size increases if you use unequal N (you must increase N for Group B more than you decrease N for group A). This can make sense if treatment A "costs" more (considering expense, hassle and risk) than treatment B. Even though the total sample size goes up, choosing unequal N may reduce the total cost (or risk) of the experiment.

Sample size				
Group A	Group B	Ratio	Total	When to choose
20	20	1.000	40	If the "cost" of treatment A is 1.0 times the "cost" of treatment B.
18	23	1.250	41	If the "cost" of treatment A is 1.6 times the "cost" of treatment B.
17	26	1.500	43	If the "cost" of treatment A is 2.3 times the "cost" of treatment B.
15	30	2.000	45	If the "cost" of treatment A is 4.0 times the "cost" of treatment B.
14	42	3.000	56	If the "cost" of treatment A is 9.0 times the "cost" of treatment B.
13	52	4.000	65	If the "cost" of treatment A is 16.0 times the "cost" of treatment B.
12	60	5.000	72	If the "cost" of treatment A is 25.0 times the "cost" of treatment B.

You cannot reduce the sample size of Group A further (to less than half the number needed if you choose equal sample size) without losing statistical power.

You can reduce the number of subjects in one group (but never to less than half of what the size would have been if the sample sizes were equal), but must increase the number of subjects in the other group even more. If you choose unequal sample size, the total number of subjects will increase. But if one treatment is more costly (in expense, risk and effort) than the other, using unequal sample size will reduce the cost of the experiment.

If you are curious, the math of these calculations is very simple. The number that enters into sample size and power calculations is the harmonic mean of the two sample sizes. The harmonic mean is the reciprocal of the mean of the reciprocals of the two sample sizes, which is equivalent to the equation below. All the experimental designs proposed by StatMate have the same harmonic mean of sample sizes.

$$\mathrm{N} = \text{Harmonic Mean} = \frac{2 \cdot \mathrm{N}_1 \cdot N_2}{N_1 + N_2}$$

2.3.13 When should I use a one-tail P value?

The computation of sample size depends on what value you pick for alpha -- the P value below which you deem a result to be statistically significant. If you pick a larger value for alpha, you can get by with fewer subjects. When choosing alpha, you also need to choose if you plan to compute a one-tailed (also called one-sided) or two-tailed (two-sided) P value.

© 2011 GraphPad Software, Inc.

Review the difference.

If you choose alpha=0.05 as your threshold (as is conventional) but choose a one-tailed, rather than a two-tailed, significance level, StatMate will compute smaller sample sizes. This is appropriate only when:

- You can predict (while designing the study) whether any change will be an increase or a decrease. The issue is not whether you can predict a difference (that is the point of doing the study) but on whether you interpret an increase and decrease the same way. You should only choose a one-tail P value when previous data, physical limitations or common sense tell you that a difference, if any, can only go in one direction.
- If your data surprise you and end up showing a difference in opposite direction to what you predicted, you must attribute this change to chance and thus conclude the difference is "not statistically significant". For example, if you predicted that a treatment would increase the variable you are measuring, then you must deem any decrease -- no matter how large -- to be a coincidence and conclude the effect is "not statistically significant" .

The first condition is easy -- everyone thinks they can predict the results of an experiment. The second condition is harder. When the results come in opposite to what you expect, the natural temptation is to be intrigued and want to understand what happened. It is very hard, when you see results that are opposite of what you expected, to conclude that the results are due to chance and are not "statistically significant". For this reason, few scientists use one-tailed P values.

Suggestion: Use two-tailed P values unless you have a very strong reason to use a one-tail P value.

2.3.14 Why do different programs compute different sample sizes?

Sample size calculations are based on some simplifications, and different texts give different equations which give slightly different results. Sample size calculations are based on estimated values (the value of SD, or the control proportion), and arbitrary choices (for alpha, power, and effect size). The inputs are not precise, so the results can't be. So a modest difference between programs is just not something to worry about.

2.3.15 How are the calculations done?

With the exception of survival analyses, all calculations were done using equations obtained from Jacob Cohen, Statistical Power Analysis for the Behavioral Sciences (second edition), Lawrence Erlbaum Assoc, 1988.

Comparing two means (unpaired t test)
Computed using equation 12.2.1 of Cohen. In that equation, d equals the difference between means divided by the SD. We know the SD and all other variables in the equation

© 2011 GraphPad Software, Inc.

so solve for the difference between means.

The power of a t test with unequal sample sizes N1 and N2 is equal to the power of a t test with equal sample sizes of size N, where

$$N = \text{Harmonic Mean} = \frac{2 \cdot N_1 \cdot N_2}{N_1 + N_2}$$

One-sample t test
The distances are computed as for the unpaired t test and then divided by the square root of 2 according to equation 2.3.4 of Cohen.

Comparing paired groups (paired t test)
If you enter the anticipated SD of the differences, results are calculated as for the one-sample t test. If you enter the anticipated value of the correlation coefficient between pairs, the results for the unpaired t test are corrected according to Cohen's equation 2.3.9.

Comparing two proportions (chi-square)
We use equation 12.6.3 of Cohen, rearranged to solve for h.

Comparing survival curves (logrank test)
We used equations 10.7 and 10.10 from Parmar and Machi. There is no way to rearrange these equations to solve for the effect size, so StatMate does so iteratively.

3 Power of a completed experiment

3.1 Power analysis tutorial (unpaired t test)

3.1.1 Introduction to power analyses

If your analysis results in a "statistically significant" conclusion, it is pretty easy to interpret.

But interpreting "not statistically significant" results is more difficult. It is never possible to prove that there is no difference -- that the treatment had zero effect on the outcome. A tiny difference will go undetected. So to interpret not significant results, you have to put the results in a scientific context. How large a difference would you care about? If that difference existed, would the experiment or study have detected it?

It is never enough to just conclude "not significant". You should always look further. One approach is to interpret the confidence interval. The other approach is to look at the power the study or experiment had to detect a hypothetical difference. That is what StatMate helps you do. The two approaches really tell you the same thing. Some statisticians argue that power analyses on completed experiments are not helpful, and suggest you only look at confidence intervals. I think that power analyses can help put the results in perspective, as well as help you design the next experiment better.

© 2011 GraphPad Software, Inc.

3.1.2 1. Introduction to the example

Using StatMate is entirely self-explanatory, and this example discusses the logic behind power analysis more than the mechanics of using StatMate.

We will continue analyzing the experiment discussed in the sample size example (Clinical Science 64:265-272, 1983). Now we'll use power analysis to interpret the results.

We determined the number of alpha$_2$-adrenergic receptors on platelets of people with and without hypertension. Here are the results:

	Controls	Hypertensives
Number of subjects	17	18
Mean receptor number (receptors/platelet)	263	257
Standard Deviation	87	59

Next: The results were analyzed with a t test.

3.1.3 2. Results of a t test

The data were analyzed with an unpaired t test. Here are the results from Prism:

Table Analyzed	Data 2
Column A	Control
vs	vs
Column B	Hypertensive
Unpaired t test	
P value	0.8116
P value summary	ns
Are means signif. different? (P < 0.05)	No
One- or two-tailed P value?	Two-tailed
t, df	t=0.2402 df=33
How big is the difference?	
Mean ± SEM of column A	263.0 ± 21.00 N=17
Mean ± SEM of column B	257.0 ± 14.00 N=18
Difference between means	6.000 ± 24.98
95% confidence interval	-44.84 to 56.84
R squared	0.001746
F test to compare variances	
F,DFn, Dfd	2.126, 16, 17
P value	0.1333
P value summary	ns
Are variances significantly different?	No

Because the mean receptor number was almost the same in the two groups, the P value is very high. These data provide no evidence that the mean receptor number differs in the two groups.

© 2011 GraphPad Software, Inc.

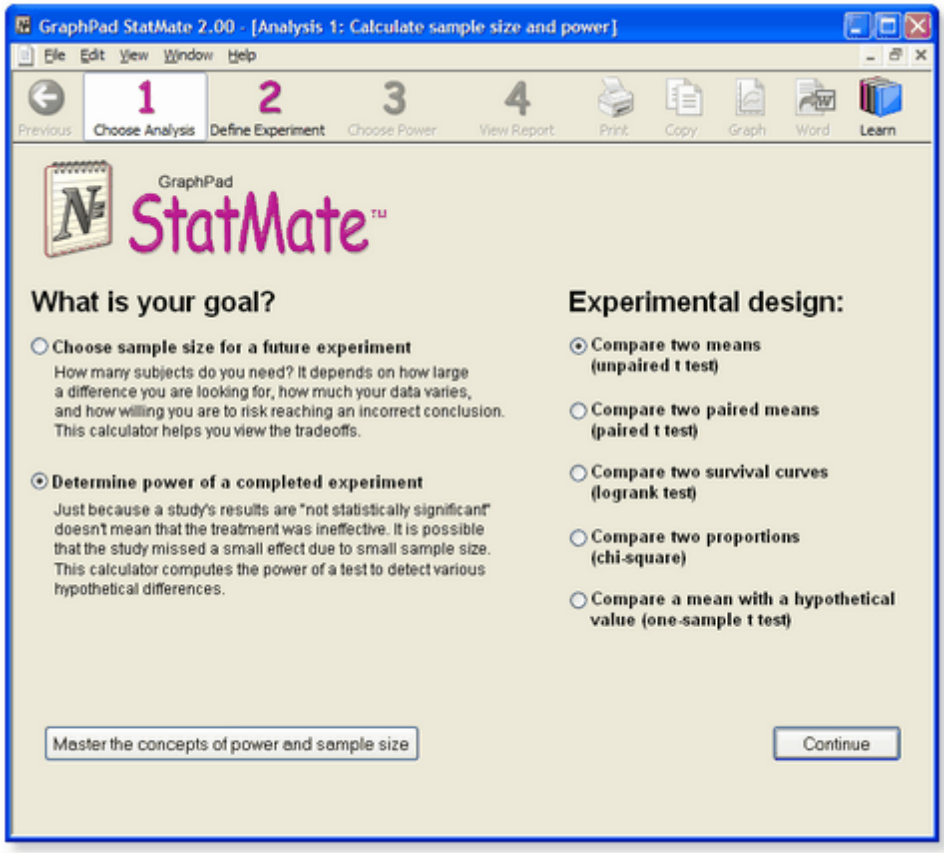
While it is tempting to just stop with the conclusion that the results are "not statistically significant" (as we did in this study published 20 years ago), there are two ways to go further. One approach, which we'll discuss later, is to interpret the confidence interval. But here we'll use power analysis to evaluate the experiment.

Next: Start StatMate.

3.1.4 3. Tell StatMate you want to peform power calculations

On step 1, you choose the kind of analysis you want StatMate to perform by answering two questions:

- What is your goal? For this example, we want to determine the power of a completed experiment.
- What is your experimental design? In this example, we plan to compare the mean of two groups using an unpaired t test.

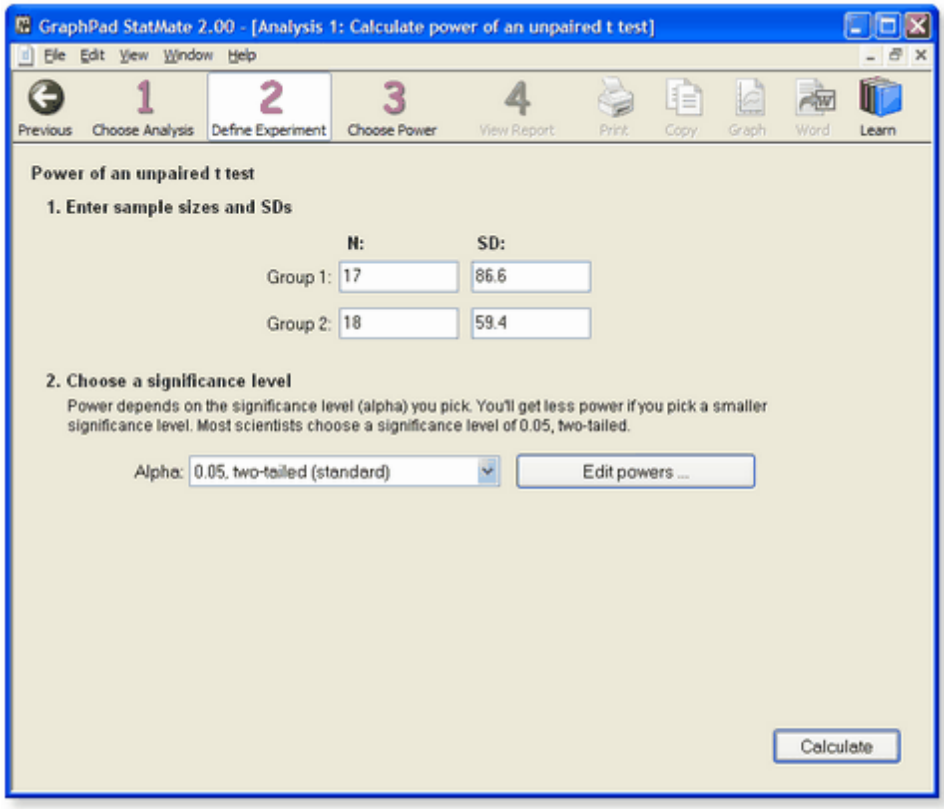


Next: Enter the data.

3.1.5 4. Enter SD and N for each group

Next enter the results of the study.

© 2011 GraphPad Software, Inc.



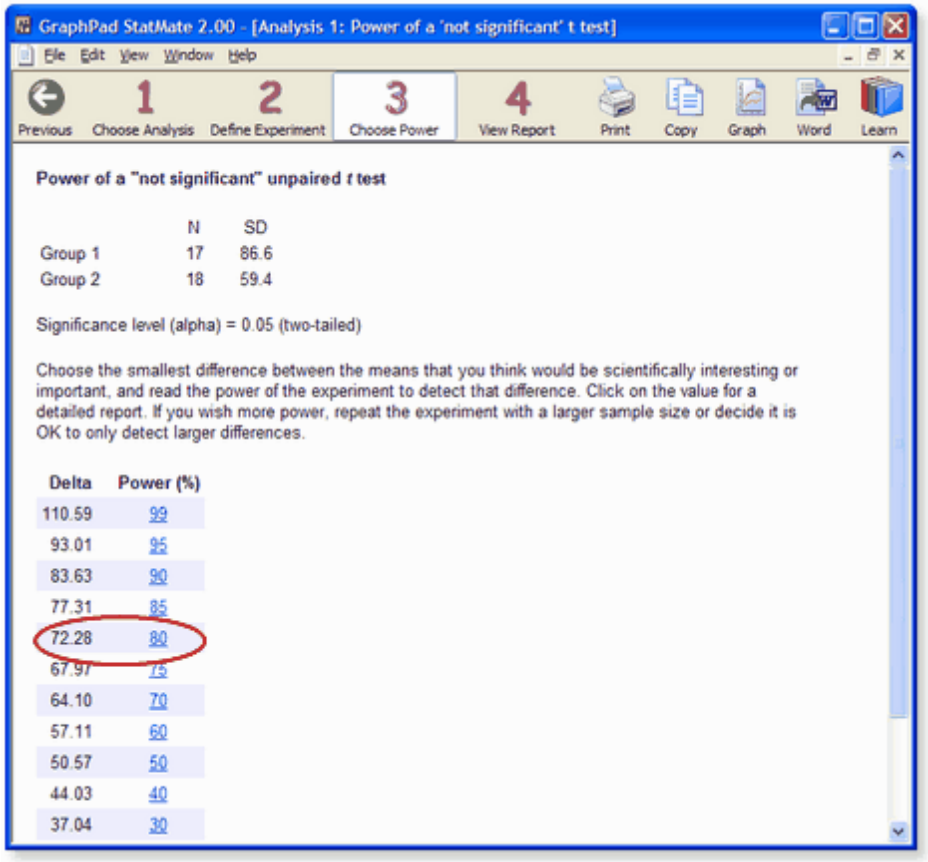
Note that you do not need to enter the mean of the two groups. Mean values are not needed for power calculations. You need only enter the size and variability of the groups.

Next: View the tradeoff of effect size and power.

3.1.6 5. View tradeoff of effect size and power

StatMate shows us the power of the study (given the sample sizes and standard deviations you entered) to detect various hypothetical differences (delta).

© 2011 GraphPad Software, Inc.



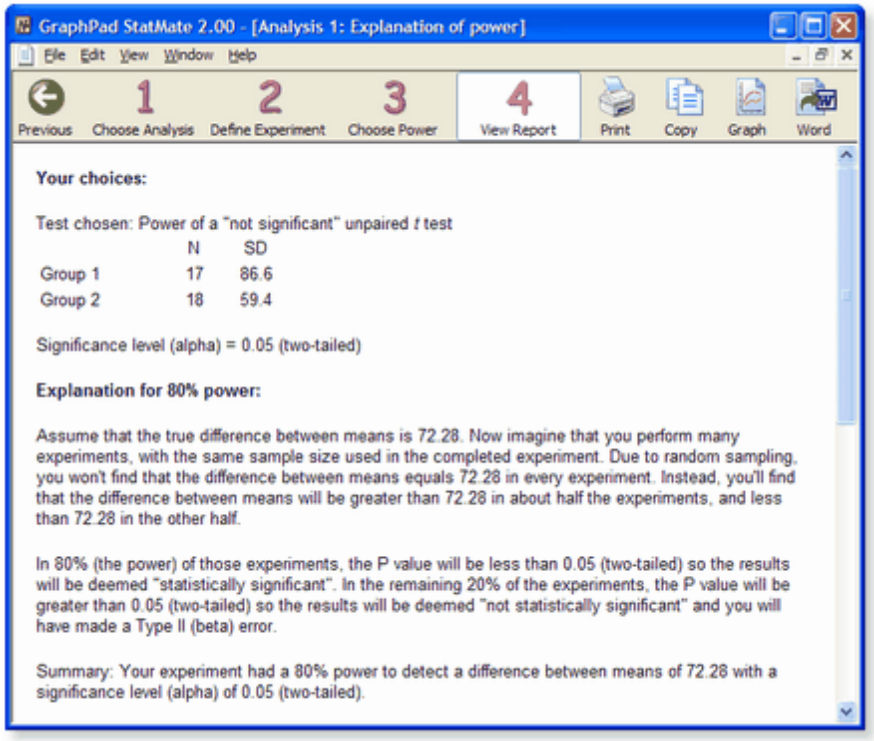
If you click on any result, StatMate will give a detailed report. Let's choose a power of 80%, which is commonly used.

Next: View StatMate's report.

3.1.7 6. StatMate's report of power analysis

Here is the report from StatMate:

© 2011 GraphPad Software, Inc.



You clicked on a row with delta=72.28, so the power calculations are based on the assumption that the true difference between means equals 72.28. You've already entered the sample sizes you used.

To understand the meaning of power, we need to think about what would happen if you performed lots of experiments, each with sample sizes of 17 and 18 like in this experiment. In each of these hypothetical experiments, the values are distributed in a Gaussian distribution with a standard deviation equal to the average of the SDs you entered[1]. You'd expect the actual difference between the means to vary from experiment to experiment. Since we have assumed that the true difference between overall means is 72.28, you'd expect about half of your simulated experiments to have a larger difference than this, and half to have a smaller difference. In what fraction of these imaginary experiments would you expect to obtain a P value less than 0.05 and thus conclude that the results are statistically significant? That's the definition of the power of an experimental design, and in this case the power is 80% (that is what you clicked on).

In summary, given the sample sizes and standard deviations you actually observed, you can conclude that this experimental design had a 80% chance (power) to detect a true difference between means of 72.28. Similarly (look at the table above), this experimental design had a 90% chance (power) to detect a difference of 83.65.

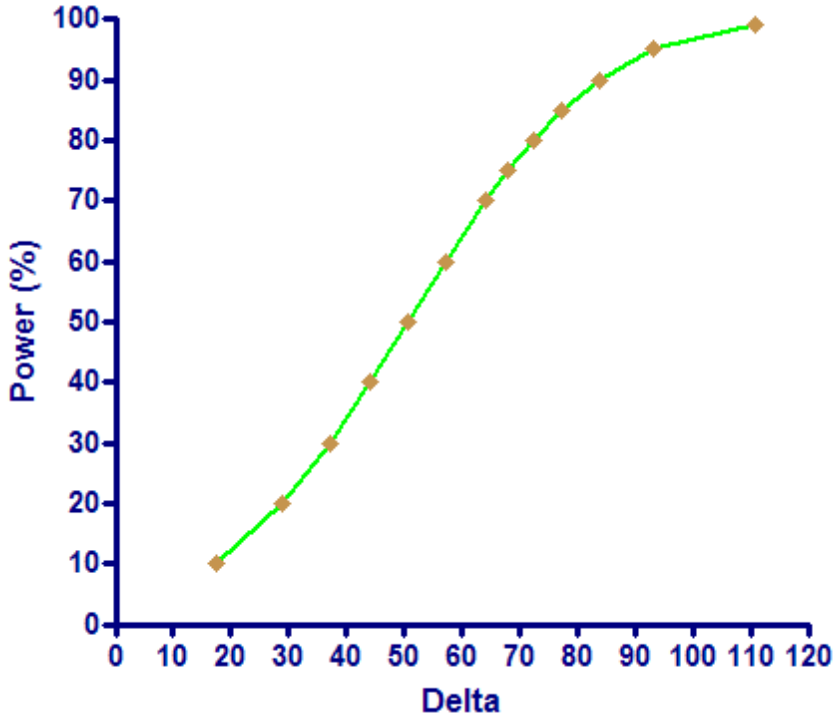
Next: Put the results in perspective.

[1] Actually it is the square root of the weighted average of the two variances, where variance is SD squared.

© 2011 GraphPad Software, Inc.

3.1.8 7. Putting the results in perspective

It is easiest to see the relationship between the difference between means (delta) and power by viewing a graph. StatMate does not create graphs itself. But if you have GraphPad Prism (Windows version 4.01, Mac 4.0b, or later), just click the graph button to make an instant graph in Prism. If you don't have Prism, you can copy and paste the table into another program for graphing.



All studies have a high power to detect "big" differences and a low power to detect "small" differences. So the graph always looks like the one above. The difference is in the numbers on the X axis. Interpreting the graph depends on putting the results into a scientific context.

When we analyzed the data with a t test, we reached the conclusion that the difference was not statistically significant. How does this power analysis help us evaluate the data?

Let's look at two alternative interpretations of the results:

- **Interpretation A.** We really care about receptors in the heart, kidney, brain and blood vessels, not the ones in the platelets (which are much more accessible). So we will only pursue these results (do more studies) if the difference was 50%. The mean number of receptors per platelet is about 260, so we would only be seriously interested in these results if the difference exceeded half of that or 130. From the graph above, you can see that this study had extremely high power to detect a difference of 130 receptors/platelet. In other words, if the difference really was that big, this study (given its sample size and

© 2011 GraphPad Software, Inc.

variability) would almost certainly have found a statistically significant difference. Therefore, this study gives convincing negative results.

- **Interpretation B.** Hey, this is hypertension. Nothing is simple. No effects are large. We've got to follow every lead we can. It would be nice to find differences of 50% (see above) but realistically, given the heterogeneity of hypertension, we can't expect to find such a large difference. Even if the difference was only 20%, we'd still want to do follow up experiments. Since the mean number of receptors per platelet is 260, this means we would want to find a difference of about 50 receptors per platelet. Reading off the graph (or the table), you can see that the power of this experiment to find a difference of 50 receptors per cell was only about 50%. This means that even if there really were a difference this large, this particular experiment (given its sample size and scatter) had only a 50% chance of finding a statistically significant result. With such low power, we really can't conclude very much from this experiment. A reviewer or editor making such an argument could correctly argue that there is no point publishing negative data with such low power to detect a biologically interesting result.

As you can see, the interpretation of power depends on how large a difference you think would be scientifically or practically important to detect. Different people may reasonably reach different conclusions.

3.2 Questions about power analysis

3.2.1 How much power do I need?

The power is the chance that an experiment will result in a "statistically significant" result given some assumptions. How much power do you need? There is no real answer to this, but these guidelines might be useful:

- If the power is less than 50% to detect some effect that you think is worth detecting, then the study is really not helpful.
- Many investigators choose sample size to obtain a 80% power.
- Ideally, your choice of acceptable power should depend on the consequence of a Type II error.

3.2.2 How can I get more power from my data?

If you are not happy with the power of your study, consider this list of approaches to increase power (abridged from Bausell and Li).

The best approach to getting more power is to collect more, or higher quality, data by:

- Increasing sample size. If you collect more data, you'll have more power.
- Increasing sample size for the group that is cheaper (or less risky). If you can't add more subjects to one group because it is too expensive, too risky, or too rare, add subjects to the other group.
- Reduce the standard deviation of the values (when comparing means) by using a more

© 2011 GraphPad Software, Inc.

homogeneous group of subjects, or by improving the laboratory techniques.

You can also increase power, by making some compromises:

- Increase your choice for alpha. Alpha is the threshold P value below which you deem the results "statistically significant". While this is traditionally set at 0.05, you can choose another value. If you raise alpha, say to 0.10, you'll increase the power of the study to find a real difference while also increasing the chance of falsely finding a "significant" difference.
- Decide you only care about a larger difference or effect size. All studies have higher power to detect a large difference than a small one.

3.2.3 How does power analysis complement confidence intervals?

All results should be accompanied by confidence intervals showing how well you have determined the differences (ratios, etc.) of interest. For our example, the 95% confidence interval for the difference between group means extends from -45 to 57 receptors/platelet. Once we accept the assumptions of the t test analysis, we can be 95% sure that this interval contains the true difference between mean receptor number in the two groups. To put this in perspective, you need to know that the average number of receptors per platelet is about 260.

The interpretation of the confidence interval (like the power analysis) must be in a scientific context. Here are two approaches to interpreting this confidence interval.

- You could say: The CI includes possibilities of a 20% change each way. A 20% change is huge. With such a wide CI, the data are inconclusive. Could be no change. Could be big decrease. Could be big increase.
- Or: The CI tells us that the true difference is unlikely to be more than 20% in each direction. Since we are only interested in changes of 50%, we can conclude that any difference is, at best, only 20% or so, which is biologically trivial. These are solid negative results.

Both statements are sensible. It all depends on how you would interpret a 20% change. Statistical calculations can only compute probabilities. It is up to you to put these in a scientific context. As with power calculations, different scientists may interpret the same results differently.

The power analysis approach is based on having an alternative hypothesis in mind. You can then ask what was the probability that an experiment with the sample size actually used would have resulted in a statistically significant result if your alternative hypothesis were true.

If your goal is simply to explain your results, the confidence interval approach is enough. If your goal is to criticize a study of others, or plan a future similar study, it might help to do a power analysis as well as interpret the confidence interval.

© 2011 GraphPad Software, Inc.

Confidence intervals and power analyses are based on the same assumptions, so the results are just different ways of looking at the same thing. You don't get additional information by performing a power analysis on a completed study, but a power analysis can help you put the results in perspective.

3.2.4 What are the assumptions when computing the power of a suvival analysis?

StatMate computes the power of a completed survival analysis using these assumptions:

- The hazard ratio is the same at all times. For example, treated patients have half the death rate at all times as control subjects.
- All subjects are followed the same length of time.
- Few subjects drop out along the way.

Some programs compute sample size assuming that the survival curves follow an exponential time course. StatMate does not make this assumption.

3.2.5 Why it is not helpful to compute the power to detect the difference actually observed?

It is never possible to just ask "what is the power of this experiment?". Rather, you must ask "what is the power of this experimental design to detect an effect of some specified size?". Which effect size should you use? How large a difference should you be looking for? It only makes sense to do a power analysis when you think about the data scientifically. It isn't purely a statistical question, but rather a scientific one.

Some programs try to take the thinking out of the process by computing only a single value for power. These programs compute the power to detect the effect size (or difference, relative risk, etc.) actually observed in that experiment. The result is sometimes called observed power, and the procedure is sometimes called a post-hoc power analysis or retrospective power analysis.

In our example, we observed a mean difference of 6 receptors/platelet. So these programs, as part of the unpaired t test results, would compute the power of this experiment (given its sample size and standard deviation) to detect a difference of 6 receptors/platelet. This power is tiny.

Let's put this in context. Remember that the mean number of receptors per platelet is about 260. So a difference between groups of 6 receptors/platelet is bit over a 2% change. That's tiny. If there really was a 2% difference between controls and hypertensives, no one would care. It would be a major waste of resources to design a study that has high power to detect such a tiny change (at least in this context; in other contexts, a 2% change might be important). So there is no point at all in computing the power of our completed study to detect such a tiny difference.

If your study reached a conclusion that the difference is not statistically significant, then by definition its power to detect the effect actually observed is very low. You learn nothing

© 2011 GraphPad Software, Inc.

new by such a calculation. What you want to know is the power of the study to detect a difference that would have been scientifically or clinically worth detecting.

These articles discuss the futility of post-hoc power analyses:

- M Levine and MHH Ensom, Post Hoc Power Analysis: An Idea Whose Time Has Passed, Pharmacotherapy 21:405-409, 2001.
- SN Goodman and JA Berlin, The Use of Predicted Confidence Intervals When Planning Experiments and the Misuse of Power When Interpreting the Results, Annals Internal Medicine 121: 200-206, 1994.
- Lenth, R. V. (2001), Some Practical Guidelines for Effective Sample Size Determination, The American Statistician, 55, 187-193

3.2.6 How are the calculations performed?

With the exception of survival analyses, all calculations were done using equations obtained from J. Cohen, Statistical Power Analysis for the Behavioral Sciences.

Comparing two means (unpaired t test)
StatMate first computes the pooled standard deviation (same calculation as used in the unpaired t test) and the harmonic mean (reciprocal of the mean of the reciprocals of the two sample sizes). Then we use equation 12.2.1 of Cohen. In that equation, d equals the difference between means divided by the SD. We know the pooled SD and all other variables in the equation so solve for the difference between means.

One-sample t test and paired t test
The distances are computed as for the unpaired t test and then divided by the square root of 2 according to equation 2.3.4 of Cohen.

Comparing two proportions (chi-square)
StatMate first computes the harmonic mean of the sample sizes of the two groups, and then uses the arcsin equation 12.6.3 of Cohen, rearranged to solve for h.

Comparing survival curves (logrank test)
StatMate uses equation 10.10 of Parmar and Machi. There is no way to rearrange this equations to solve for the effect size, so StatMate does so iteratively.

4 A review of statistical principles

4.1 A review of alpha and statistical significance

You do an experiment because you want to know whether the treatment affects the outcome. You need statistical analyses because the answer can be ambiguous.

Most statistical analyses calculate a P value that answers this question:

© 2011 GraphPad Software, Inc.

- If the treatment really does not affect the outcome, what is the chance that random variability would result in a difference as large or larger than you observed in your experiment?

The P value is a fraction between 0.0 and 1.0. If the P value is small, you are inclined to believe that the difference you observed is due to the treatment.

In many cases, you want to make a decision from the data. You want to conclude either that the difference is statistically significant, or that it is not statistically significant. This is done in a simple manner. Before running the experiment you set a threshold P value, called *alpha* or the *significance level*. By tradition, alpha is usually set equal to 0.05. After running the experiment and calculating the P value, use this logic:

- If the P value is less than or equal to alpha (usually set to 0.05), conclude that the treatment had a statistically significant affect, and you reject the null hypothesis that the treatment was ineffective.
- If the P value is greater than alpha, conclude that the treatment did not have a statistically significant effect. In other words, conclude that the data are consistent with the hypothesis that the treatment is ineffective.

4.2 A review of statistical power and beta

Even if the treatment affects the outcome, you might not obtain a statistically significant difference in your experiment. Just by chance, your data may yield a P value greater than alpha.

Let's assume we are comparing two means with a t test. Assume that the two means truly differ by a particular amount, and that you perform many experiments with the same sample size. Each experiment will have different values (by chance) so a t test will yield different results. In some experiments, the P value will be less than alpha (usually set to 0.05), so you call the results statistically significant. In other experiments, the P value will be greater than alpha, so you will call the difference not statistically significant.

If there really is a difference (of a specified size) between group means, you won't find a statistically significant difference in every experiment. *Power* is the fraction of experiments that you expect to yield a "statistically significant" P value. If your experimental design has high power, then there is a high chance that your experiment will find a "statistically significant" result if the treatment really works.

The variable *beta* is defined to equal 1.0 minus power (or 100% - power%). If there really is a difference between groups, then beta is the probability that an experiment like yours will yield a "not statistically significant" result.

© 2011 GraphPad Software, Inc.

4.3 One-sided vs two-sided P values

Start with the null hypothesis that the two populations really are the same and that the observed discrepancy between sample means is due to chance.

> Note: This example is for an unpaired t test that compares the means of two groups. The same ideas can be applied to other statistical tests.

The two-tail P value answers this question: Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far (or further) apart than you observed in this experiment with either group having the larger mean?

To interpret a one-tail P value, you must first predict which group will have the larger mean before collecting any data. The one-tail P value answers this question: Assuming the null hypothesis is true, what is the chance that randomly selected samples would have means as far apart (or further) as observed in this experiment with the preselected group having the larger mean ?

A one-tail P value is appropriate only when previous data, physical limitations or common sense tell you that a difference, if any, can only go in one direction. The issue is not whether you expect a difference to exist - that is what you are trying to find out with the experiment. The issue is whether you should interpret increases and decreases the same.

You should only choose a one-tail P value when two things are true. First, you must have predicted which group will have the larger mean (or proportion) before you collected any data. That's easy, but the second criterion is harder. If the other group ends up with the larger mean - even if it is quite a bit larger - then you must attribute that difference to chance. That is a hard condition to abide by. If you see a large difference in the "wrong" direction, you inclination will be to investigate further, and not to attribute the change to chance. For this reason, it is best to avoid one-tailed P values.

4.4 Type I and Type II errors

When you reach a conclusion that a difference is "statistically significant" or "not statistically significant", you can make two kinds of mistakes:

- You conclude that there is a statistically significant effect, but the difference was actually due to random variability. The treatment had no effect, but the random variation between the groups made you conclude that the difference existed. This is called a *Type I error*.
- You conclude that there is no statistically significant effect, but the treatment was really effective. The treatment had an effect, but the random variation between the groups obscured the difference. This is called a *Type II error*.

© 2011 GraphPad Software, Inc.

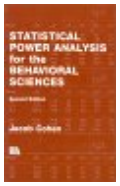
When planning an experiment and deciding on sample size, you need to decide how willing you are to make a Type I and Type II error.

5 Learn more

5.1 Books on power and sample size

Power Analysis for Experimental Research : A Practical Guide for the Biological, Medical and Social Sciences
by R. Barker Bausell and Yu-Fang Li, Cambridge University Press, 2002.
ISBN 0521809169

This book discusses t tests, ANOVA (including complicated designs) with a good conceptual explanation of the principles of power analysis. Chapter 2 is particularly interesting with a list of eleven strategies to increase the power of an experiment.

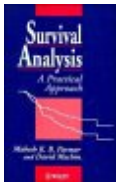


Statistical Power Analysis for the Behavioral Sciences
by Jacob Cohen, Lawrence Erlbaum Assoc, 1988.
ISBN 0805802835

A comprehensive book about computing power and sample size in many situations, including ANOVA but not survival analysis. About half the book is a discussion of principles; the other half is a set of tables. Cohen pushes the concept of "effect size". No matter what kind of data you collect, you can reduce the results down to an effect size, and can compute sample size and power based on effect size. This is a helpful concept that unifies statistical methods that are otherwise somewhat distinct. But beware of his definitions of standard effect sizes.

How Many Subjects? : Statistical Power Analysis in Research
by HC Kraemer and S Theimann. Sage publications. 1987
ICBN 0803929498

This short book clearly explains general principles. It also applies them, with a bit of math, to different kinds of analyses. It includes a good discussion of ANOVA, but not survival analysis.



Survival Analysis: A Practical Approach
by Mahesh K. B. Parmar and David Machin, John Wiley & Sons, 1995.
ISBN 0471936405

Chapter 10 explains sample size calculations for survival analysis.

© 2011 GraphPad Software, Inc.

5.2 More advanced software

StatMate is very easy to use, and to learn from, but it doesn't go as deep as some programs. If you need to compute sample size and power for situations that StatMate cannot handle, look at these programs.

- PASS. Power and Sample Size. From NCSS software.
- Power and Precision. From Biostat.
- nQuery Adviser. From Statistical Solutions.

6 Using GraphPad StatMate

6.1 How do I...?

StatMate is mostly self-explanatory, but here are answers to common questions:

How do I save my work?
StatMate is a calculator. There is no concept of document files to save. However you can export the result page as either a text file or as HTML (formatted).

How do I make a graph with StatMate?
StatMate doesn't make graphs itself. If you have Prism installed (4.01 or later for Windows, or 4.0b or later for Mac), click the graph button to create a graph within Prism. Otherwise, copy the table of numbers from StatMate and paste into another graphing program.

How do I make the report have a larger (or smaller) font?
Choices on the View menu let you use larger or smaller fonts for steps 3 and 4.

How do I make StatMate create reports with different range of N and power?
From Step 2, click "Edit powers" or "Edit powers and Ns". On the pop up dialog, edit the list of powers and sample sizes (N) that Prism uses to create the table in step 3. Check an option box on that dialog if you want your revised list to become the default.

How do I put the results into another program?
If you use Windows and want to put the results into Word, simply click the Word button (Windows only) on the StatMate toolbar. Your results will be instantly pasted into a new Word document. If you use a Mac, or wish to put the results in a program other than Word, select the results, copy to the clipboard, and then paste into the other program. StatMate will copy your report both as plain text and as HTML (web page). The program you paste into will decide which format to paste. Or use Edit... Paste Special (when available) so you can decide. You can also export the results as either a plain text file or as a HTML web page. Use the File... Export command while viewing StatMate's report.

© 2011 GraphPad Software, Inc.

6.2 How to cite StatMate

When citing analyses performed by any program, include the name of the analysis, and complete version number (including the platform and the second number after the decimal point). Use this example as a guide:

"We determined sample size using GraphPad StatMate version 2.00 for Windows, GraphPad Software, San Diego California USA, www.graphpad.com".

To find the full version number, pull down the Help menu (Windows), Apple menu (OS 8-9) or StatMate menu (OS X) and choose About StatMate.

More important than citing the name of the program is to state how you did the analysis. For example, it won't help a reader much to know that StatMate helped you choose a sample size unless you also state all the analysis choices.

6.3 License agreement

GraphPad Software, Inc. (also referred to as "we", "us" or "GraphPad) has created Windows and Macintosh versions of GraphPad Prism, GraphPad InStat, and GraphPad StatMate (herein referred to as "the Software"). The Software is licensed, not sold, and you must purchase separate licenses for each program and each platform.

USE OF AN INDIVIDUAL LICENSE. You may install one copy of the Software on any single computer. If that computer is used exclusively by you and not shared with others, then you may also install the Software on a second computer, also used exclusively by you.

USE OF A NETWORK LICENSE. You may install one copy of the Software on a network server, and ensure that the number of PC's that access the Software at any one time does not exceed the maximum number of simultaneous users specified on you license. All network users must be located at one physical site.

USE OF A DEMO VERSION. You may use the Software for a period of 30 days from the date that you first install it. Once this 30-day period has expired, you must either purchase a permanent license to use the Software, or promptly destroy all copies of the Software in your possession.

USE OF PRIOR VERSIONS IF YOU UPGRADED. If you purchased your license as an upgrade from a prior version, you may run the prior version on the same computer you installed the Software. You may not transfer your prior version license to anyone else. If you have a network license, you must ensure that the maximum number of users simultaneously accessing any version of the Software does not exceed the number of simultaneous users specified on you license.

MONEY-BACK GUARANTEE. If you are not satisfied with the Software, and you purchased it directly from GraphPad, you may return it within ninety days for a full refund (excluding shipping charges). If you purchased the software from a third party, contact that party regarding returns.

© 2011 GraphPad Software, Inc.

NO WARRANTY. The software and services offered by us are provided on an "as is" and "as available" basis without warranties of any kind, express, implied or statutory, including but not limited to, the implied warranties of title and fitness for a particular purpose. Computer programs are inherently complex, and the Software may not be free of errors. We do not warrant that the Software will be uninterrupted, timely, reliable, secure or error-free and expressly disclaim any warranties as to the materials contained therein, or the goods or services offered by us. If this disclaimer is not allowed, GraphPad's liability shall be limited to refunding the purchase price. No oral or written information or advice given by GraphPad, its dealers, distributors, agents or employees shall create a warranty or in any way increase the scope of this warranty and you may not rely on any such information or advice.

LIMITED LIABILITY. In no event will we be liable to you for any damages beyond refunding the purchase price. Neither GraphPad, its principals, shareholders, officers, employees, affiliates, contractors, subsidiaries, or parent organizations, nor anyone else who has been involved in the creation, production or delivery of any materials provided by GraphPad software, including but not limited to this software, shall be liable for any indirect, consequential, incidental or punitive damages (including damages for loss of business profits, business interruption, loss of business information, and the like) arising out of the use or inability to use such materials, even if GraphPad has been advised of the possibility of such damages.

RESTRICTIONS. You may not use the Software in medical diagnosis or treatment, or in applications or systems where the Software's failure to perform can reasonably be expected to result in significant physical injury, property damage, or loss of life. You may not translate, reverse engineer, decompile, modify, disassemble, or create derivative works derived from the Software. You may not install, use, distribute or copy the Software (or its serial number) except as expressly permitted in this License

BACKUP COPY. You may make one copy of the Software for back-up and archival purposes only.

ASSIGNMENT. You may assign your license to another person or legal entity (the "Assignee"), provided that prior to such assignment, the Assignee undertakes in writing to be bound by your obligations under this Agreement, and you transfer to the Assignee all of your copies of the Software (including electronic copies stored on computer disks or drives), including all copies of updates and prior versions of the Software. You may not rent or lease the Software to someone else.

INTELLECTUAL PROPERTY. You acknowledge that you have only the limited, non-exclusive right to use the Software as expressly stated in this license and that GraphPad retains title to the Software and all other rights not expressly granted. You agree not to remove or modify any copyright, trademark, patent or other proprietary notices that appear, on, in or with the Software. The Software is protected by United States copyright, and trademark laws and international treaty provision.

UPGRADES. GraphPad may, but is not obligated to, upgrade the software from time to time. If an upgrade is issued in less than three months from the date of purchase, you will

© 2011 GraphPad Software, Inc.

be eligible to receive the upgrade at no charge. After that period, you will have an opportunity, but not an obligation, to purchase the upgraded software.

TERMINATION. This license terminates if you fail to comply with its terms and conditions. If your license terminates, you must destroy all copies of the Software. The Limitations of Warranties and Liability set out above shall continue in force even after any termination.

ENTIRE AGREEMENT. This License is the entire agreement between you and GraphPad, and there are no other agreements, express or implied.

CHOICE OF LAW. This License shall be governed by the laws of the State of California as if between residents of said state.

SEVERABILITY. If any provision of this License is held to be invalid, illegal, or unenforceable, it will be construed to have the broadest interpretation that would make it valid and enforceable. Invalidity of one provision will not affect the other provisions.

ACKNOWLEDGEMENT. Use of the Software affirms that you have read and understood this agreement, and agree to be bound by its contents.

6.4 Technical support

Do you have the current version?
If you are having trouble with StatMate, check that you are running the current release. To do this, connect to the internet, and then drop the Help menu from StatMate and choose Check for Updates.

Updates (interim versions of GraphPad software containing bug fixes or minor improvements) are free to owners of the corresponding major releases. In contrast, *upgrades* (a new version with many new features) must be purchased.

Is the answer to your question on www.graphpad.com?
If you have a question that is not answered in this help system, the answer is very likely in the searchable Quick Answers Database at www.graphpad.com/support.

If you have questions about data analysis, also browse the library of statistical articles and links on www.graphpad.com.

Personal technical support
If you need personal help, we are happy to help. Contact us via email at support@graphpad.com, or use the form on the support page. Be sure to tell us which version and platform (Windows or Mac) you use. The full version number is not on the CD label. You have to run StatMate to find out which version you are using. Drop the Help menu (Windows), Apple menu (Mac OS 8-9) or StatMate menu (Mac OS X) and choose About StatMate. Windows versions have two digits after the decimal point (i.e. 2.00). Mac versions have a single digit and letter after the decimal (i.e. 2.0a).

© 2011 GraphPad Software, Inc.

While we reserve the right to charge for support in the future, we promise that you'll receive free support for at least one year. We can't predict how computer hardware and system software will change in the future, so cannot promise that StatMate 2, released in 2004, will work well with future versions of Windows or the Mac OS.

Note that your StatMate license does not include free statistical consulting. Since the boundary between technical support and statistical consulting is often unclear, we will usually try to answer simple questions about data analysis.

© 2011 GraphPad Software, Inc.