Paper 99-2019

Statistical Power Calculations Using SAS Software

Russ Lavery Contractor

ABSTRACT

The term “statistical power” is a shorthand for “the power of a statistical test to detect a change in Y” but many books only use the phrase “statistical power” without any explanation/expansion of the term. Only using the phrase “statistical power” makes learning the topic more difficult. Calculation of “the power of a test to detect a change in Y” is tremendously important in research effectiveness and is directly tied to the cost of doing research. The topic is considered to be difficult and learning how to handle the many types of statistical power calculations is, to some extent, requires an overview of all of the statistics courses you’ve ever taken. SAS has done much to make the calculation of statistical power fast and easy.

The first section of this paper will be a review of the concepts, and theory, of statistical power. In order to illustrate principles of statistical power we will use only one statistical test – a test to detect a change in the mean of a variable. Teaching statistical tests graphically requires creating a different kind of graphic for each test. A test of means has an easy to create, and understand, graphic. The second part of the paper will be an overview of how to use SAS to easily perform statistical power calculations.

INTRODUCTION

A great number of the statistical power calculations involve determining the number of subjects required to have “reasonable” (often 80%) statistical power. Since the number of subjects is usually linked to the cost of doing the research project, understanding statistical power is a very important and practical topic.

The goal of a researcher is to get sufficient dollars to afford to be able to recruit, and manage, enough subjects so that her research project has a “reasonable” chance of detecting the effect of the research manipulation (detecting the effect of the researcher changing a level of an X variable). The consequences of not doing having enough subjects (or too many) are very severe, but often not well understood. An overview of the consequences, and a few examples, are worth reviewing.

When the sample size and power for a research project are too low the following list bad things happen. The researcher often fails to answer the research question. The researcher fails to detect associations between Y and X variables. The researcher inconveniences subjects for no reason. The false results can misdirect future research. An underpowered (too few subjects) study wastes resources that could be better applied to another research project. Finally; if the research hypothesis is true and the study is underpowered, conducting underpowered research delays the discovery of a true relationship and the delivery of the results of that research to society (e.g. delayed medical treatments).

When the sample size and power for a research project are too large the following bad things happen. The study is more difficult to administer. The study is more costly than it needs to be. The study denies budgets to other useful studies. The study inconveniences more subjects than are required. Finally, because big studies generally take longer, an overpowered study delays the delivery of the results of that research to society (e.g. delayed medical treatments).

Here are some examples that illustrate the prevalence and importance of this issue.

A recently published meta-review of literature discovered no relationship between 18 genes that, over twenty years, have been thought to influence depression. The initial discovery, and the seminal article on this issue, involved one gene of the 18 and was published in 1996. That first, underpowered study involved gene research on *fewer than 1,000 patients*. That one gene has been the subject of over 450 research papers.

The 18 genes, linked to depression by the early, small-N studies, have been the subject of over 1,000 papers. Millions of dollars, and 20 years of research effort, have likely been wasted because the initial studies did not have a preliminary power analysis.

A review article of papers that had been published in prestigious medical journals between 1975 and 1990 suggests that more than 80% of the trials, that reported negative results, did not have enough sample size (not enough subjects) to detect a 25% change in the Y variable. Nearly 2/3 of the studies did not have enough statistical power to detect a 50% change in the Y variable. A 50% change in the Y variable is a huge change and few drugs will be expected to have an effect that large.

The book "Bad Pharma" shows a chart of over 30 studies, conducted between 1959 and 1988 on one drug. Through this almost 30 year period, the results of the drug investigation were unclear. A recently done meta-analysis, combining the sample sizes of all the studies, concluded that that the individual studies were lacking in statistical power and that the research issue could have been resolved in 1973 – 15 years earlier.

STATISTICAL POWER ANALYSIS IS CONSIDERED DIFFICULT

Statistical power analysis is considered to be a difficult topic. Each type of statistical analysis (t-test, paired t-test, ANOVA, logistic regression, survival analysis, comparison of percentages, etc.) has its own logic, picture and graphic. While the logic and formulas are different, the underlying principles are the same and this paper will endeavor to use pictures to illustrate the underlying principles.

Statistical power is an area of ongoing research and it might the worth warning the reader that some of their research questions have no "cookbook formula for getting the answer". In those cases, a simulation is the only way to determine the power of your particular research proposal/plan.

THE CONDITIONS THAT AFFECT STATISTICAL POWER

When a researcher is doing a study she typically changes the level of some X variable and expects this change in X to produce a change in Y. We are investigating the power of the statistical test, under certain conditions, to detect that change in Y.

There is no such thing as "the statistical power of a test"! There is only "*the statistical power of a test under certain conditions*". If you change the conditions, you change the statistical power of the test.

THERE ARE FIVE CONDITIONS THAT AFFECT STATISTICAL POWER.

One of Five: The type of statistical test being used in the analysis affects the ability of the test to detect a change in Y. Some tests, typically tests that assume normality as opposed to nonparametric tests, are more powerful than others. In this paper, we will not investigate the different power associated with different statistical tests. There are just too many tests.

Two of Five: The level of alpha (the researchers willingness to make a type I error) has an effect on the power of the test.

Three of Five: The "degree of shift" or "the effect size" affects the ability of the test to detect a change in Y. Big changes in the Y variable are easier to detect.

Four of Five: The variability in the data affects the ability of the test to detect a change in Y. It is harder to detect the effect of changing an X variable when there's a lot of variability in the X and Y variables.

Five of Five: The sample size affects the ability of the test to detect a change in Y.

THE TWO TYPES OF MISTAKES ONE CAN MAKE IN DOING A STATISTICAL ANALYSIS.

A researcher can make two kinds of logical mistakes. She can: 1) reject a true Ho and 2) fail to reject a false Ho. Please see the figures below to see a graphical representation of the two kinds of mistakes that can be made in doing statistical analysis.

Start off by assuming that we administered a questionnaire measuring antisocial behavior to a subject group (see gold curve).

We know that our subject group has a distribution shown by the gold curve. We set up a reject region, shown by the red triangles, where 95% of the area under the yellow curve is between the gold triangles. The researcher has a drug to administer and hopes it will decrease antisocial scores. The decrease might be represented by the black distribution (our drug has little effect) or with the blue distribution (the drug was effective).

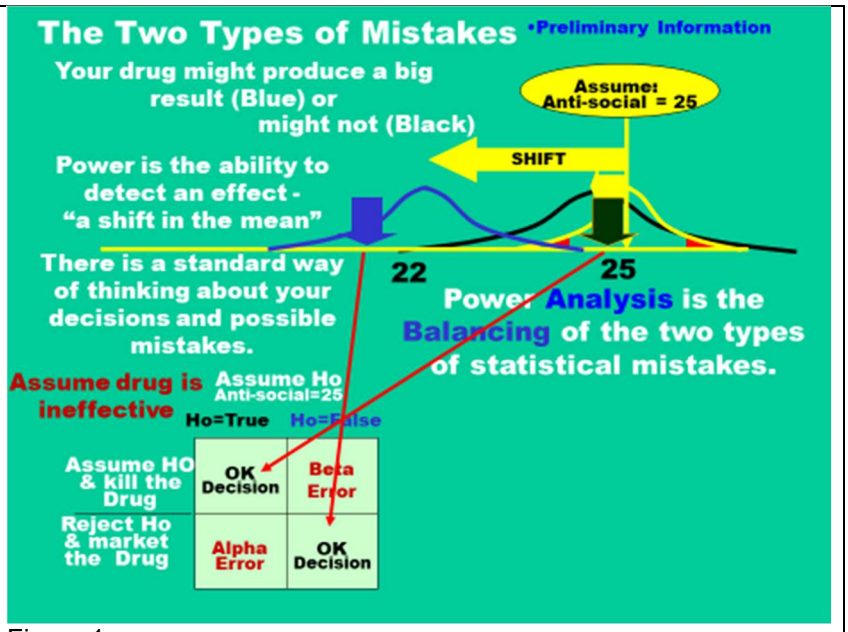


Figure 1

A researcher can make two kinds of mistakes but this paper will start off logic by showing, in Figure 1, correct decisions – the two kinds of correct decisions.

We start off, always, by assuming our training has no effect – that the distribution of scores after training is the same as the distribution of scores before training (gold distribution).

If our training only shifts the scores as much as is shown by the black distribution, and we happen to get an observation shown by the black arrow, we would continue to assume that our experimental treatment was ineffective. We would stop drug development and that would be a good decision.

If the training was effective and shifted the scores as much as is shown by the blue distribution, and we had an observation indicated by the blue arrow, we would conclude that the distribution of scores has decreased. We would reject our assumption of no change in scores and that would be an okay decision. We would go on to market the drug.

Figure 2 shows the two kinds of mistakes. Let's assume that the drug did not shift the scores at all and that the distribution after training is still the gold distribution.

When post-drug measurements are taken, the researcher might get very unlucky and pull a score indicated by the yellow arrow.

That arrow is in the reject Ho region. The researcher would reject Ho and conclude that the drug was effective.

A company would market the ineffective drug and likely lose money. Losing money is what made the author call the sample "unlucky".

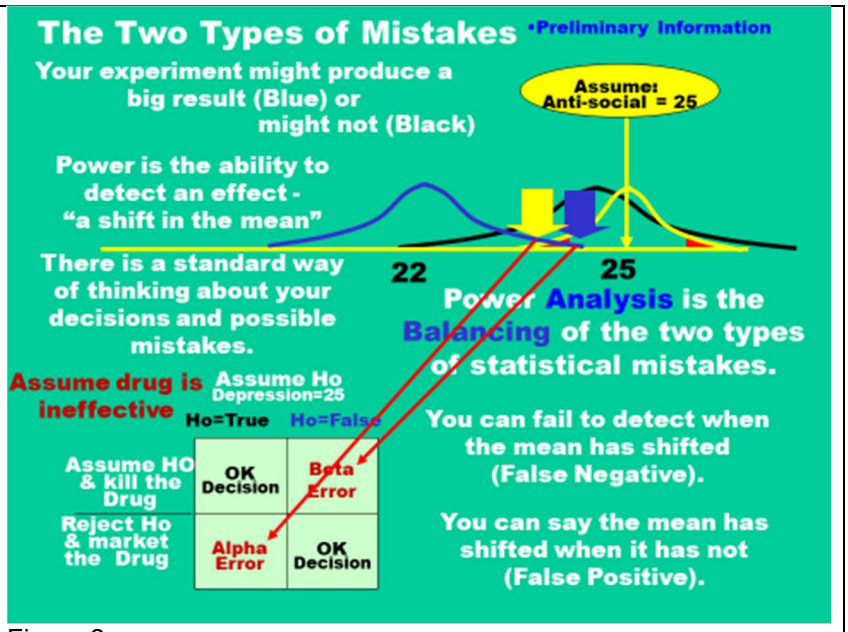


Figure 2

If the drug was effective in reducing antisocial behavior it might shift the scores as far as the blue distribution.

On any particular day the researcher, sampling from the blue drug, might be very unlucky in taking her sample. Assume that she pulled a sample with a score indicated by the blue arrow. This is a very unusual score for the "effective drug" – very high.

The sample score is in the "do not reject region" and the researcher would continue to assume that the drug is ineffective. Drug development would stop – even though the drug really did reduce anti-social behavior. A company would miss the chance to sell a good drug.

Let's think a bit about the business prospects for the black distribution. This distribution shows a drug that seems not very effective in reducing the average score for antisocial behavior. It has produced a small shift in the average. The important consideration should be "is the observed reduction in average large enough to have some practical, or business, usefulness". The degree of statistical shift (Z score) is not of primary importance in business statistics. Statistics, in research, is important in so far as it helps solve business or research problems.

Statistical power analysis is the balancing of the two types of mistakes that researchers can make.

FIVE THINGS THAT AFFECT STATISTICAL POWER

1 OF 5) DIFFERENT STATISTICAL TESTS HAVE DIFFERENT POWERS

Some tests are, due to their nature, more powerful than other tests (or are just considered "not very powerful" for a certain task). As an example, the chi-square goodness of fit test is often said to have a weak ability (low-power) to reject Ho even when the distributions, by eye, seem obviously different.

2 OF 5) THE EFFECT OF ALPHA ON POWER

Before starting any sort of statistical analysis you have to have some idea about how your data behaves.

Here the researcher used a questionnaire to assign a bunch of people an antisocial score. The average antisocial score was 25 and she knows how the subjects are spread out (variability).

The gold distribution summarizes what she knows about the research subjects. They have an average of 25 and some variability around that average. She can say that 95% of the research subjects are between 25 +/- some number (see red triangles).

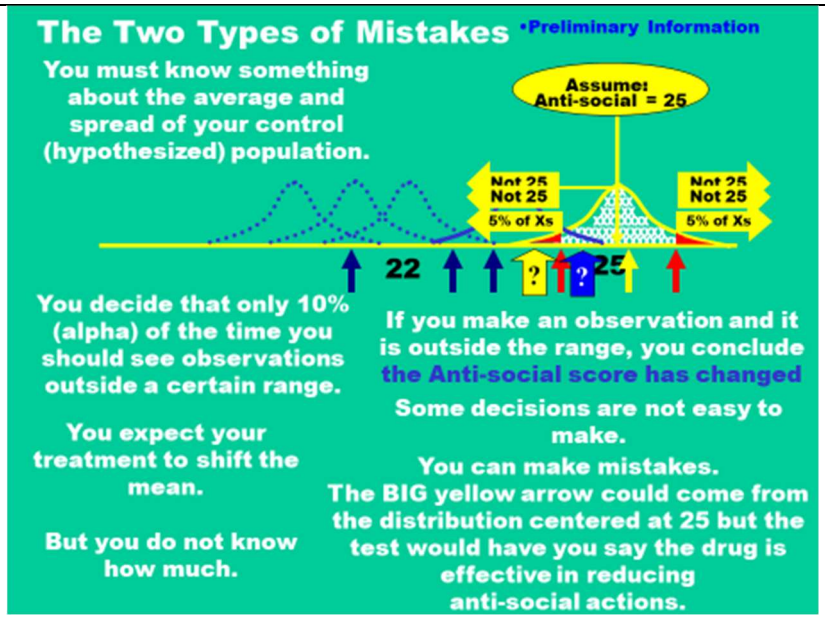


Figure 3

If 95% of the research subjects are within a certain range and she decides that a subject is outside of that range is an indication of a change in the average antisocial score, then she is willing to make a mistake 5% of the time – alpha is .05. An alpha of .05 is very common.

The blue dotted distributions are “shifted distributions”. The researcher hopes that the drug reduces antisocial scores but, before doing the experiment, she is not sure by how much which blue distribution). This section discusses the implications of not knowing how much of an effect the drug (or any other experimental manipulation) will have.

Note that the solid blue distribution represents the result of a treatment that is not very effective – it doesn’t change antisocial scores by very much.

Think of the change in Y as the distance between the center of the gold distribution in the center of the solid blue distribution. That distance, the change in the centers, is often referred to as the “effect size” of the treatment. Y is the score on the antisocial measuring questionnaire and X is the treatment given to the subjects.

Before a researcher starts on a study she never really knows how much of an “effect size” will be produced by her changing the level of X (administering a drug, or maybe some social training for these antisocial people).

It might be that she only gets an effect as large as what is indicated by the solid blue distribution. It might be that her treatment is very effective and the antisocial distribution shifts way to the left – to one of the dashed distributions.

If a researcher increases alpha (willingness to make a type I mistake) that pulls the reject Ho region towards the center of the gold distribution. The original Ho region is indicated by the red triangles under the gold curve.

The region of increased alpha is indicated by the pink area under the gold curve.

Increasing the alpha increases the percentage of the blue distribution in the "reject Ho region" and increases the "power of the test to detect a change in the mean" for a particular "effect size" or "shift in the center of the distribution".

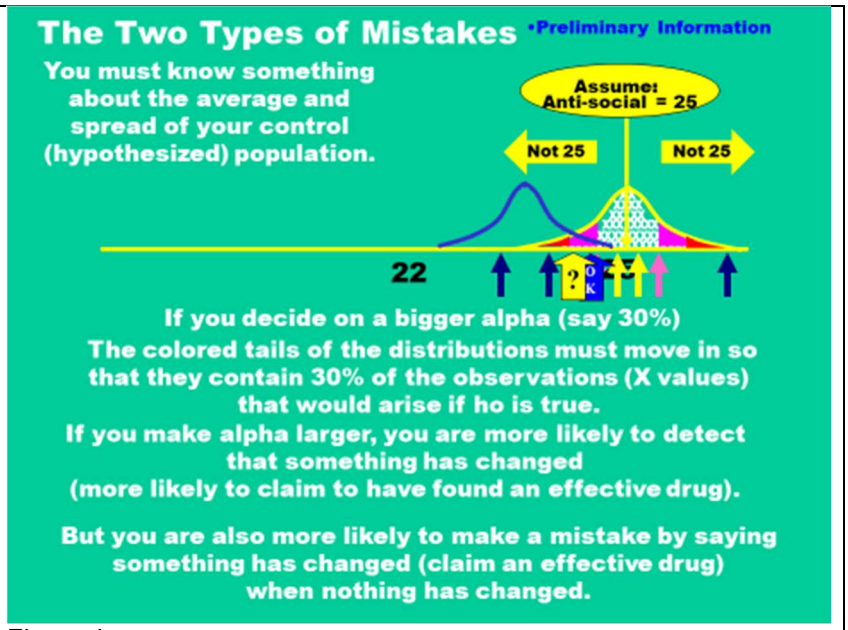


Figure 4

We indicate alpha by the colored triangles under the gold curve. A researcher is more likely to reject Ho (and detect a shift in the mean) if she makes alpha larger.

In summary, increasing alpha makes the test more powerful. Increasing alpha increases the chance that she will reject Ho - that the test will detect a shift in the mean.

However; increasing alpha also increases the chance that the test will reject Ho due to chance (more of the gold distribution is in the reject Ho region and more of a distribution that has a small degree of shift, (see black distribution in Figure 2 for an example) is in the reject region.

3 OF 5) STATISTICAL POWER DEPENDS ON THE "EFFECT SIZE" - IN OUR CASE DEGREE OF SHIFT OF THE MEAN

We reject Ho, and detect a change in the mean, when we get a sample in the reject Ho region. In Figure 5 the black distribution represents our belief about our subjects. The dotted blue distribution represents the distribution scores after subjects have been treated.

The ability to detect a change in the mean score is shown by the percentage of the blue curve in the reject Ho region – the shaded blue area.

The stronger the effect of our treatment the more the blue distribution shifts and the greater the percentage of the blue distribution in the reject Ho region.

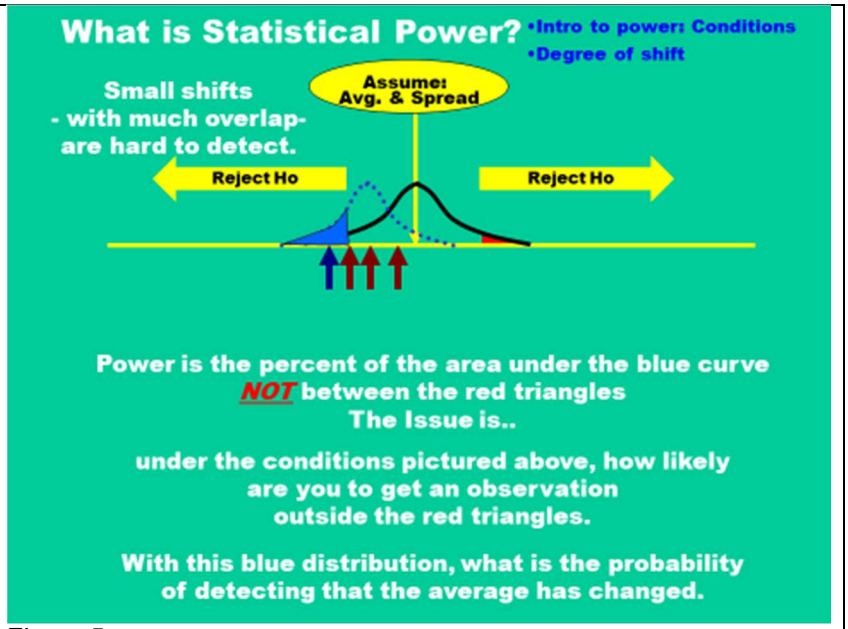


Figure 5

4 OF 5) AN INCREASE IN VARIABILITY IN THE DISTRIBUTIONS DECREASES THE ABILITY TO DETECT A CHANGE

In Figure 6 we illustrate how changing the variability, the spread of *only* the Ho distribution, affects the percentage of the blue curve outside the reject Ho region.

We have three examples stacked on top of each other. The effect of our treatment, the degree of shift, does not changed in the three examples.

If the subjects in our study are more consistent the distributions become narrower and there is less overlap between the Ho distribution and the shifted distribution. This means more power.

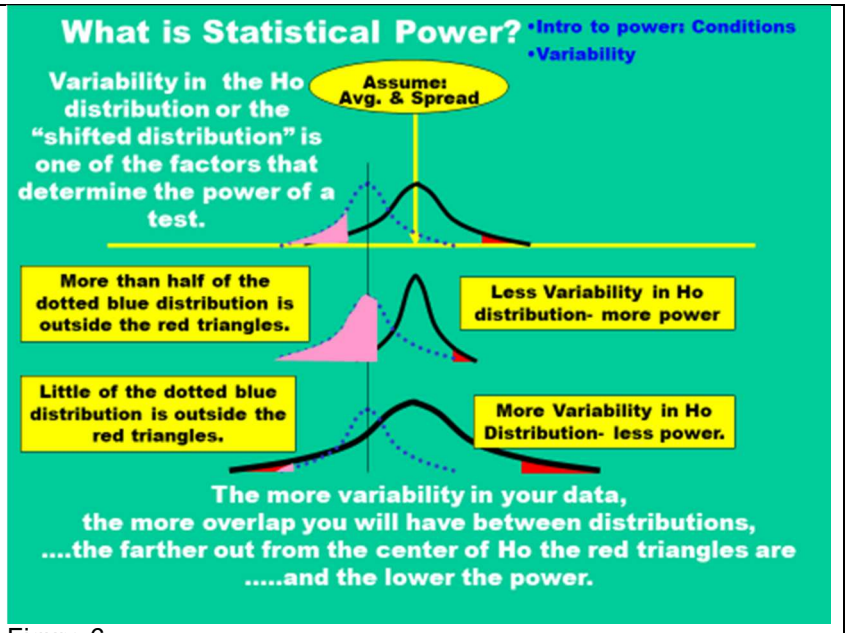


Figure 6

You can see that where there is less variability in the Ho distribution we have more power to detect a change in the mean. When the spread of the Ho distribution increases, when we have more variability in the data, it becomes more difficult to detect a change in the mean because less of the blue distribution is in the reject Ho region.

Figure 7 illustrates the same phenomenon that we saw in Figure 6.

The difference between Figure 6 and Figure 7 is that, in Figure 7, both the Ho distribution and the distribution of scores after treatment have shrunk (gotten skinny) or gotten wider by the same amount.

As was seen above, as the variability in the data increases it becomes harder to reject Ho- harder to detect a shift in the average.

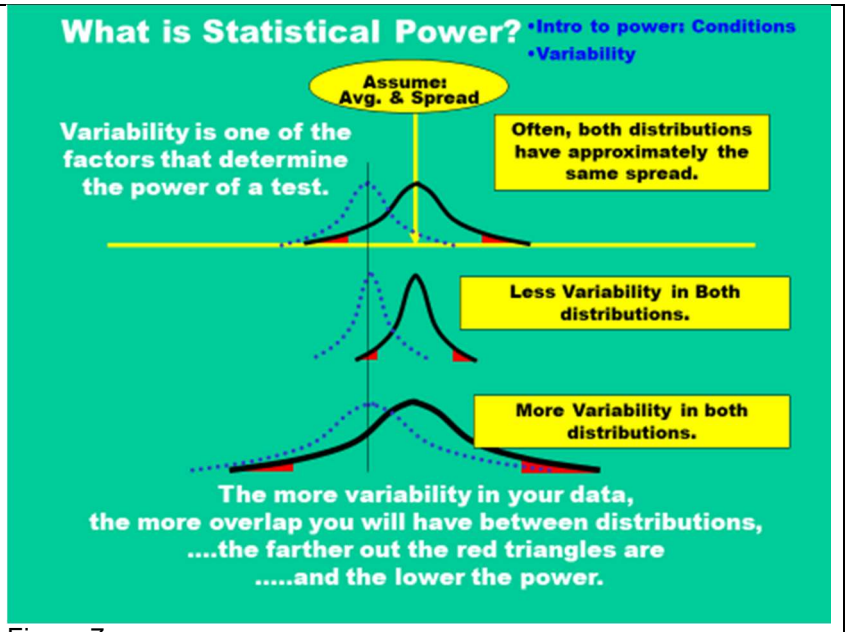


Figure 7

Figure 8 brings the two previous concepts together and makes explicit an important and subtle point. It is not the degree of shift (the effect size) or the variability in the data that is important.

What is important is the degree of shift *compared* to the variability in the data.

As the slide says, if we could reduce the overlap we could have a more powerful test. A larger shift would reduce the overlap but a larger shift involves inventing a new treatment. It's unlikely that a researcher can invent a more powerful treatment in any short amount of time.

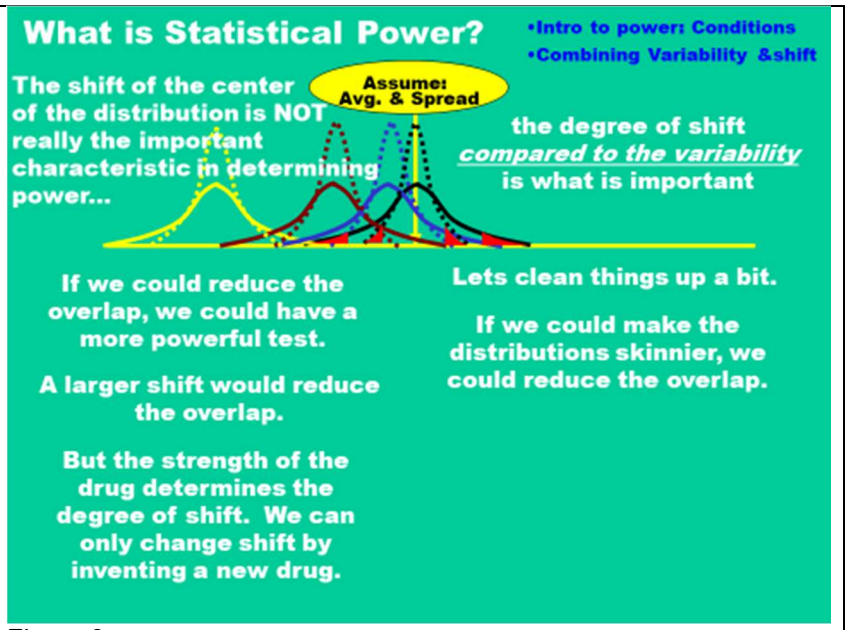


Figure 8

If we could make the distributions skinnier, see dotted lines above, without changing their means we could have a more powerful test.

This brings us to the next topic. We ***can*** make the distributions skinnier if we stop thinking of individual people in our sample and think of doing our analysis using the "distributions of an average" of N people.

5 OF 5) MAKING THE DISTRIBUTIONS SKINNIER BY INCREASING THE NUMBER OF OBSERVATIONS - THE CENTRAL LIMIT THEOREM IN ACTION

The central limit theorem

Before we start talking about increasing sample size we should review the central limit theorem and get a little bit of understanding about why increasing the sample size will be so useful.

When you make a histogram by plotting single (n=1) observations from a distribution, you are making a histogram of what is called the "parent distribution".

If you pull, and plot, single observations from the parent population you are plotting the average of a sample – but the number of observations in your samples is equal to one. This is the top row of Figure 9.

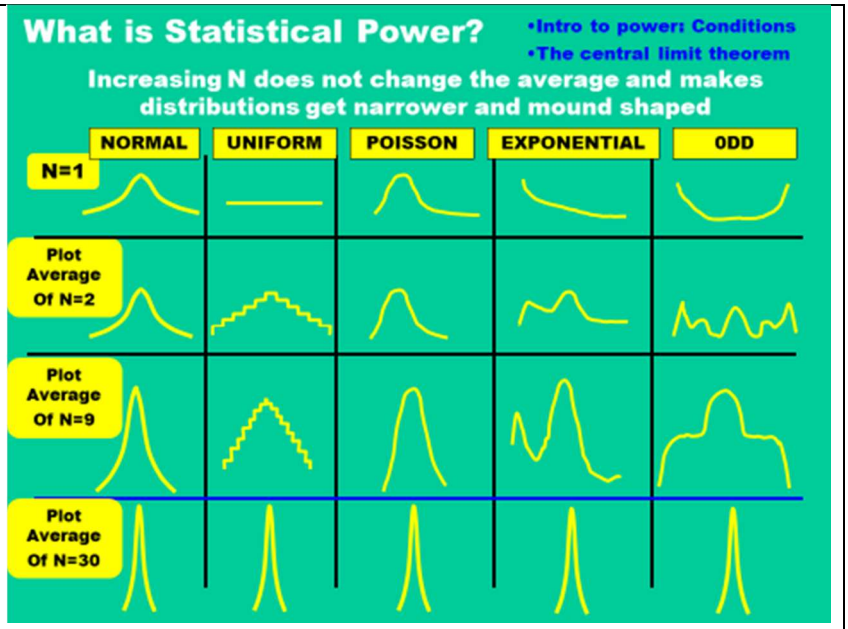


Figure 9

If you repeatedly take 30 observations (bottom row of Figure 9), from whatever distribution you like - and calculate the average - and then plot the distribution of the sample averages- that plot will look like a normal curve and get thinner.

The average of the "plots of averages", the center of your "plot of averages" will be the same as the average of the parent distribution. Mathematicians find this exciting (because they are math people) and business people find this exciting because they can use this characteristic to make a money. Many business decisions can be made knowing the average for something (average sales of the store versus average sales of another store, average number of people who drive by this location versus some other location etc.)

Another useful feature of dealing with averages of samples, where the number of observations in a sample is greater than 30, is that the distribution of sample averages is so close to the theoretical normal curve that you can use just one statistical table (the normal) when you want to do calculations. This is still important today but was hugely important just a few decades ago. At that time, only having to use one statistical table simplified problem analysis by a huge degree.

A slight digression is in order. It will illustrate some of the reasons why we deal with the distribution of sample averages rather than the parent population.

Imagine rolling one die. The distribution of outcomes is shown in the upper right of Figure 10. Possible outcomes are one to six and each outcome has a probability of .166. The average value for a role of one die is 3.5.

The distribution of averages, when the sample has a sample size equal to one, is dead flat.

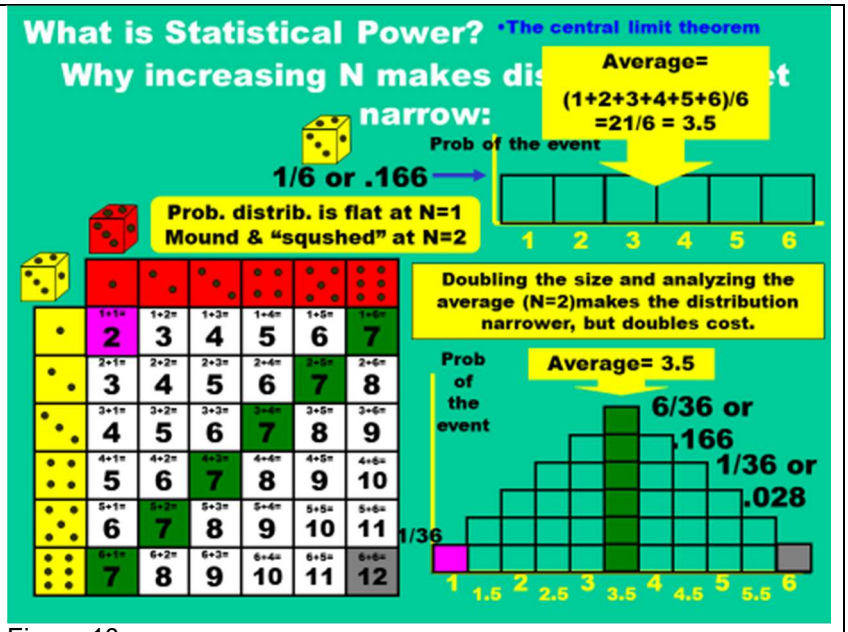


Figure 10

The rest of the slide illustrates the results of rolling two die. There are 36 possible outcomes and all outcomes are equally likely. By an outcome, I mean something like a one on the *first* die and a one on the *second* die or a five on the *first* die and a *two* on the second die. All the combinations of the two dice are equally likely.

The way to calculate probability, if you are a frequentist, is to count up the number of ways a certain event can occur and divide that by the total number of ways that anything can occur. The probability of a seven is calculated as: six different ways a seven can occur divided by 36 possible outcomes. A seven (average is 3.5) is the most likely outcome when rolling two dice and the probability is 6/36 or .166.

There is only one combination of two dice that will produce a two or a twelve. These are shown in Figure 10. The chance of rolling a two (average is 1) or a 12 (average is 6) is 1/36 or .028.

You can use Figure 10 to calculate the probability of rolling either a total of six or an eight (5/36) - or of rolling a total of five or a nine (4/36) - or a rolling a total of four or a 10 (3/36) - or of rolling a total of three or eleven (2/36).

The X axis for the histogram in the lower right-hand corner of Figure 10 shows the *averages* for the role of two die. That histogram is a plot of sample averages where the sample size is equal to two.

Notice the huge effect of taking a sample of size two rather than a sample of size one. The distribution has changed from dead flat to very mound shaped. Plotting “sample averages from samples taken from the parent distribution” is very powerful.

The average of the “distribution of sample averages” is the same as the average of the parent distribution and that can be very useful in business. That one number (the average) can be used to make decisions.

Additionally; the distribution of sample averages is mound shaped and “clustered around the average of the parent distribution”. This clustering increases the *precision* of your estimate of the average.

Imagine if one wanted to know the average of the role of a die, and rolled just one die, hoping that *one number* would be representative of the average of parent distribution. One would have to be pretty lucky to get useful information (a 3 or a 4 is close to the average) from just one roll. The average of the distribution is a 3.5 but you would be equally likely to get a six, or a one, from your one role of the die. If you roll two dice you are much more likely to get an average close to the parent population average.

Here’s a small summary.

Taking a sample and working with the sample average means:
1) the center of the parent and “average of samples” distributions are the same
2) The distributions of sample averages get “skinnier” and more precise
3) The distributions of the sample averages assume a “normal” shape and you can do analysis using just one statistical table

Figure 11 shows two different situations. The upper graph is where the shift is large compared to the variability in the data and it would be easy to detect the shift.

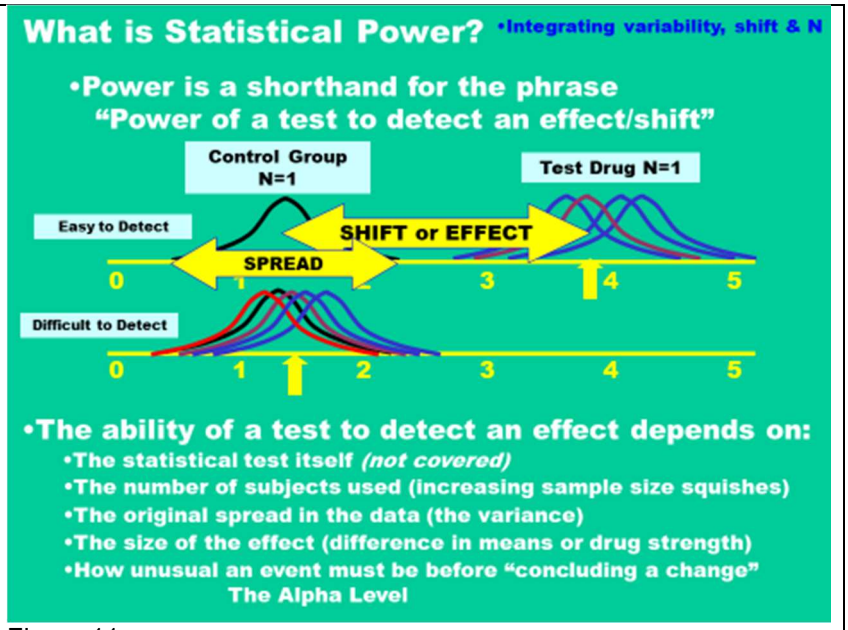


Figure 11

The lower graphs shows lots of overlap. Increasing the sample size makes distributions skinnier, reducing overlap and making tests have more “statistical power”. Increasing N decreases overlap.

THE BIG PROBLEM WITH TAKING AVERAGES: $$$

The big problem with taking averages is that increasing the sample size increases the cost of doing the study, increases the time to do the study and increases the complexity of the study.

Increasing the sample size might not increase the cost of the study in a linear manner. If a researcher is studying antisocial behavior it might be that she can get only 50 appropriate subjects from one special school (like a Juvenile Justice Alternative Education Program). If the study needs 200 students, to have sufficient power, the researcher might have to contact schools in several different states. Not being able to be constantly present at the study site greatly increases the chances of mistakes. Bigger, more complex, studies need different management styles and more resources. That means more dollars.

This brings us back to the topic of the paper. We want to use SAS procedures to calculate the appropriate number of subjects. The calculation of the number of subjects should be done in the research proposal stage – before any time is spent *doing* the study. Some academic journals now require a power analysis as part of submission to the Journal. Journals have finally recognized a historic problem with poorly powered studies. Calculating statistical power before starting work is critical.

If a researcher does not have enough money to fund a study with enough subjects to have a reasonable alpha and power – the researcher should not do an underpowered study- not even start one. The researcher should come up with another research proposal. Additionally; a researcher should not overpower a study. Inconveniencing subjects with no research benefit is unethical.

PICTURES OF THE EFFECT OF INCREASING SAMPLE SIZE (INCREASING N)

Figure 12 shows the effect of increasing sample size and is approximately to scale.

The black distribution is our hypothesized distribution (Ho) and we are trying to determine if three different drugs can produce a shift in the mean.

The area between the red triangles, under the black distribution, contains 95% of the area – so alpha is .05.

The gold drug is very effective and the purple drug is largely ineffective. The red drug is moderately effective.

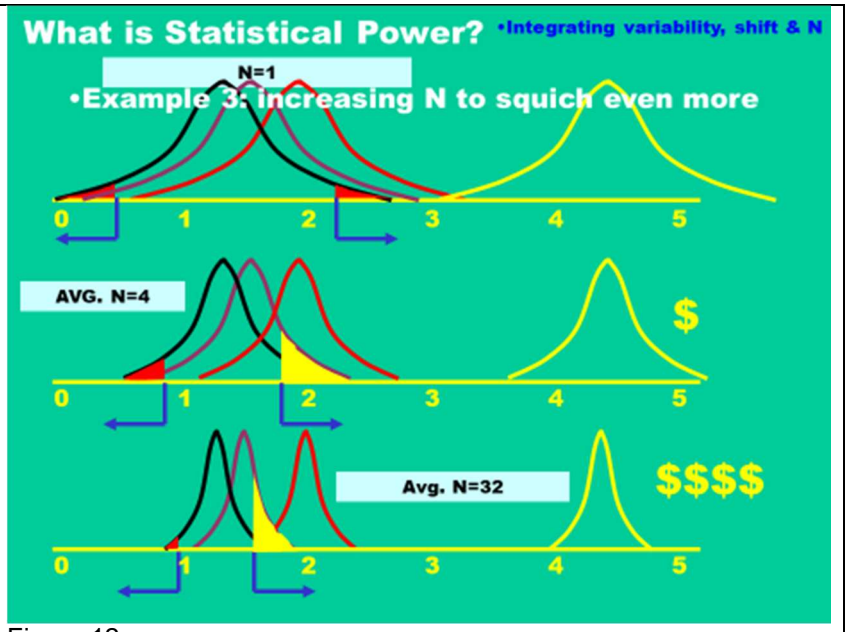


Figure 12

Even with a sample size of one, we would be able to determine that the drug that is described by the gold distribution is effective. It is very unlikely that, even with a sample size of one, the gold distribution would produce an observation in the "do not reject" region under the black curve. With a sample size of one, the purple and red drugs are underpowered. The purple drug has a small effect and that small effect will be very difficult to detect.

If we increase the sample size to four, all of the distributions of sample averages get skinnier. A significant, but still small, percentage of the purple distribution is in the reject Ho region. With N equals four, the test for the purple drug is still underpowered (it appears to be ~ 20%). The power of the purple distribution is represented by the yellow triangle under the purple distribution that is in the reject Ho region.

With N=4, the red distribution is still underpowered though the chance of detecting that shift seems to be about .7. We know that, since the center of the red distribution is in the reject Ho region, the power of this test must be greater than .5 - but it is hard to estimate by eye.

If we increase the sample size to 32 observations from each drug. The red drug has almost 100% power. The purple drug might have ~ 30% power. The gold drug is 100% powered.

If a researcher were only trying to compare the black distribution to the red and the gold distributions having a sample size of 32 is a waste of money. A sample size of 32 overpowers the study.

SETTING SAMPLE SIZES BASED ON THE COST OF BEING WRONG

In real life, people try and assign a cost to being wrong when they're doing the statistical analysis. The different ways a researcher can be wrong can have different cost structures.

Pharmaceutical research is often done in multiple stages.

Stage I: Ho is that the drug is not effective. The task is to screen many drugs using laboratory tests and use relatively inexpensive experiments to see if a new molecule *might have promise*. The cost of passing the drug onto the next level of screening is relatively low.

In this stage, the big cost is to test a drug that *is effective* and mistakenly *say that it's not* – the penalty is lost sales revenue. Some other company will discover this drug and bring it to market. In Stage 1 a researcher will set alpha to be very large in order to get high statistical power (a high ability to detect an effect). The drug company wants to send all *remotely promising drugs* on to the next stage of testing. At this stage, tests are constructed to have very high power and a large alpha is tolerated because the cost of mistakenly calling a drug "effective" is low.

Stage 2: Ho is that the drug is not effective. The cost of mistakenly calling a drug "effective" is higher. This stage tests drugs for effectiveness using animals and more expensive chemical tests. The cost of passing a drug onto the next level of screening is relatively high.

In this stage, there is a significant cost involved in sending an ineffective drug on to the next level of screening but the major worry is still that the company will not investigate, and market, a powerful drug.

In Stage 2 a researcher will set alpha to be smaller (than stage 1) in order to "remove very non-performing drugs" but still wants high statistical power (an ability to detect a small effect). The drug company wants to send moderately promising to highly promising drugs on to the next stage of testing.

Stage 3: Ho is that the drug is not effective. The cost of mistakenly calling a drug "effective" is higher. This stage tests drugs for effectiveness using small scale tests involving people. The cost of passing a drug onto the next level of screening is high.

In this case there is a significant cost involved in sending an ineffective drug on to the next level of screening. In Stage 3 a researcher will set alpha to be small in order to avoid false positives and increase the sample size to get statistical power (an ability to detect an effect). The drug company wants to only send very promising drugs on to the next stage of testing.

Stage 4: Ho is that the drug is not effective. The FDA only wants drugs with proven effectiveness on the market. This stage tests drugs for effectiveness using large scale tests involving people and companies submit results to the FDA. In Stage 4 a researcher will recruit thousands of subjects in order to get statistical power (an ability to detect an effect) and set alpha to be small.

THE NEED FOR STATISTICAL AID IN DOING CALCULATIONS

In the sections above we saw characteristics of a study and how they affect statistical power. In the real world there is one more complication – subjects leaving the study.

If a research project is going to take several years to complete you can expect that some people will move out of the area and be lost to the researcher. Some people may die. If the treatment involved (either a drug or some social training) makes the subjects uncomfortable (emotionally or physically) some subjects will leave the study to avoid this discomfort. It is very difficult to predict a dropout rate for a study.

In addition to the unpredictability in the dropout rate there are other unpredictability's in getting accurate numbers to plug into the formulas that calculate statistical power.

The researcher usually does not know the effect size of the treatment.

Additionally, the variance inside the different treatment groups is usually not known.

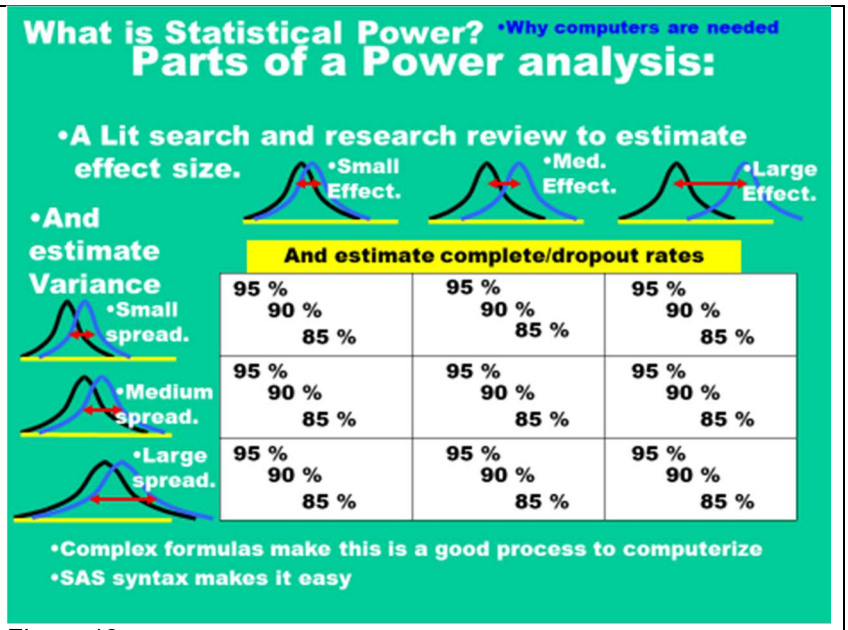


Figure 13

There are two main ways in which a researcher can cope with this situation.

If the researcher has enough money, the researcher can do what is called a "pilot study". This is a small N study with the same process as the study to be used to answer the major research issue. The goal of the pilot study is to get reasonable estimates of variance, effect size and dropout rates.

Whether the researcher has enough money to do a pilot study or not, the researcher will always do a literature search. In a literature search the researcher looks for studies of similar drugs/treatments applied to similar subjects using a similar study process. From the literature search a researcher can get rough estimates of the numbers that must be put into the formulas to calculate statistical power.

A very common process is to estimate, for each of the numbers in the power formula, three numbers: worst-case, most likely case and best case.

At this stage, a researcher can, given some time, calculate predicted statistical power for all of the different combinations of effect size, variance and dropout rate- but this is not really enough. This large amount of effort will only give a limited number (27 in figure 13) of point estimates and not provide the researcher much of an idea of the sensitivity of statistical power to small variations in the important parameters of the formula.

THIS IS WHERE SAS COMES IN.

USING SAS TO CALCULATE STATISTICAL POWER

SAS has a four options for calculating statistical power and they do an amazing job of reducing the tedium of repeated complex calculations – and the graphics are amazing.

First Option: there is a power and sample size application that can be run from a drop-down menu. The author is an old-school "type the code" sort of guy and this option will not be explored in this paper.

The author's problem with drop-down menu calculations is that they don't leave a record of what was done. The author would rather have a little bit of code (often with the title power_calculations.sas) in the project directory as a record of what was done.

Second Option: PROC Power handles many of the simple and moderately complicated power calculations and produces great output. Power calculations handled by PROC Power are:

PROC POWER< just print options>;
LOGISTIC<options>; **MULTREG<options>;** **ONECORR<options>;**
ONESAMPLEFREQ<options>; **ONESAMPLEMEANS<options>;** **ONEWAYANOVA <options>;**
PAIREDFREQ<options>; **PAIREDMEANS<options>;**
TWOSAMPLEFREQ<options>;
TWOSAMPLEMEANS<options>;
TWOSAMPLESURVIVAL<options>;
TWOSAMPLEWILCOXON<options>;
PLOT<plot-options> </ graph-options>;
Run;

Third Option: PROC GLMPower performs power analysis for complex general linear models. PROC GLMPower supports type III sums of squared tests as well as contrast of fixed class affects. It handles continuous and categorical covariates, unbalanced design and produces customized graphics.

PROC GLMPower uses parameters inside the PROC GLMPower syntax itself, but also uses what is called an "exemplary data set" that passes population response means and design weights to PROC GLM power.

Fourth Option: When a researcher's statistical analysis does not fit in any of the wide variety of analysis that SAS has created -SAS will facilitate a researcher doing a simulation to calculate power.

EXAMPLES OF PLOT OPTIONS

It is incredible how much of the tedium of this process has been eliminated by SAS. Since the plot options can be applied to many different types of statistical test, the sections of examples will start off by showing some of the power of the plotting in PROC Power.

It should be noted that the missing parameter, the parameter name followed by an equals and a period, is the signal to product power that this is the "unknown thing" to be calculated. You can ask SAS to calculate power, N or effect size.

In the top plot power is the "unknown thing". In the middle and bottom plots the number of pairs are the unknown things.

The difference between the middle and the bottom plots is the request for the type of plot.

Note that we can easily request multiple calculations.

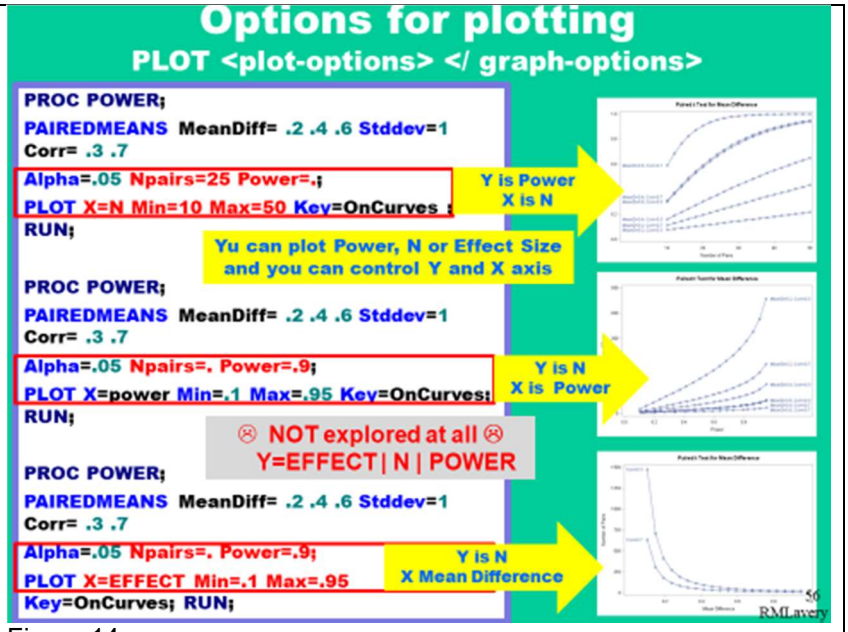


Figure 14

Above, we simply tell SAS we are interested in mean differences of .2, .4 and .6 and correlations of .3 and .7. This makes multiple plots and allows the investigation of sensitivity to changes in the parameters.

Figure 15 shows some of the additional flexibility in the types of charts produced.

Charting is a very important and useful feature of the way SAS has implemented power calculations.

You can see that power is not linear and there are many plateaus or asymptotes in these plots.

Plots are required to understand the sensitivity of the effect of other parameters and SAS creates excellent plots.

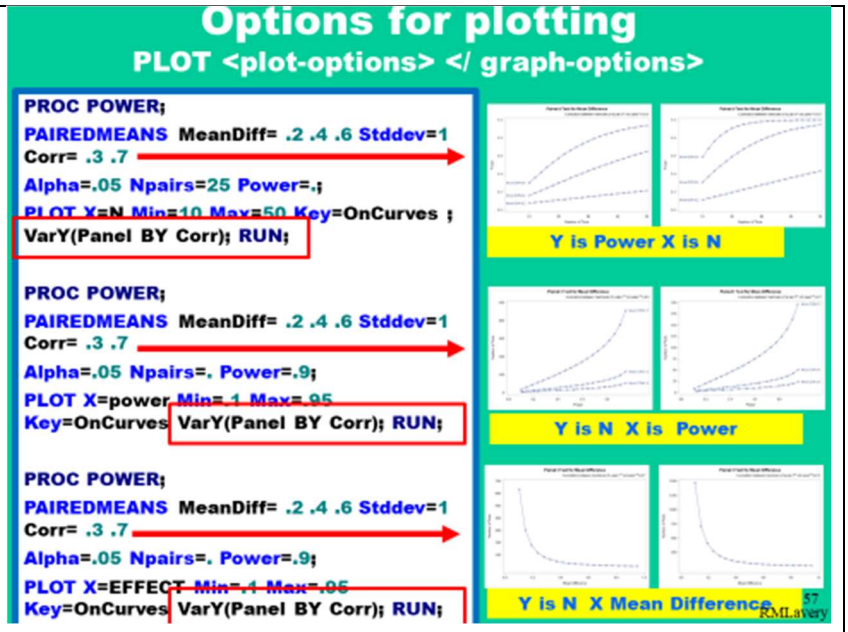


Figure 15

EXAMPLES OF LOGISTIC OUTPUT FROM PROC POWER

One of the reasons that power analysis is so difficult is that there are so many different ways to a bit of statistical analysis and so many possible different internal structures in the data.

The left side of Figure 16 shows some of the options that you can set for logistic regression.

The right side of Figure 16 shows one particular call – one request for a power analysis.

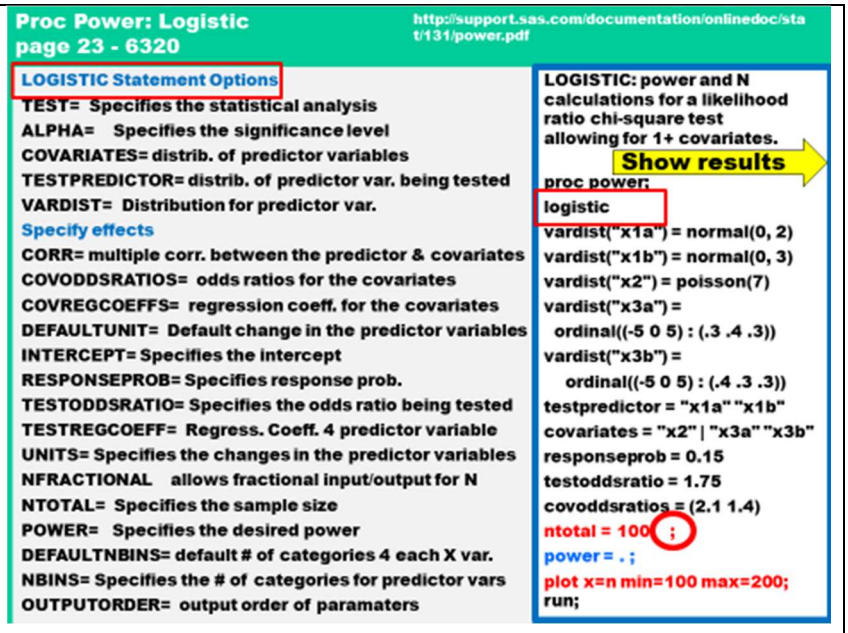


Figure 16

In order to understand the output of this power analysis for a PROC Logistic, a reader must be very familiar with all of the parameters that were set and with the terms used in table headings.

Interpretations will not be covered in this paper.

It was mentioned, earlier in the paper, that making PROC Power and PROC GLM Power demonstrate all they can do involves almost all of every statistics course a reader is likely to have had.

Proc Power: Logistic
page 23 - 6320
http://support.sas.com/documentation/onlinedoc/stat/131/power.pdf

Fixed Scenario Elements	
Method	Shieh-O'Brien approximation
Response Probability	0.15
Odds Ratio for Test Predictor	1.75
Covariate Odds Ratios	2.1 1.4
Total Sample Size	100
Alpha	0.05

Computed Power

Index	Test Pred	Covariates		Test Pred Unit	Cov Units		Total N Bins	Power
1	x1a	x2	x3a	1	1	1	300	0.931
2	x1a	x2	x3b	1	1	1	300	0.927
3	x1b	x2	x3a	1	1	1	300	0.999
4	x1b	x2	x3b	1	1	1	300	0.999

Figure 17

Again, a plot is more useful than a table.

Here is the plot of power vs total sample size for the logistic in Figure 16.

Proc Power: Logistic
page 23 - 6320
http://support.sas.com/documentation/onlinedoc/stat/131/power.pdf

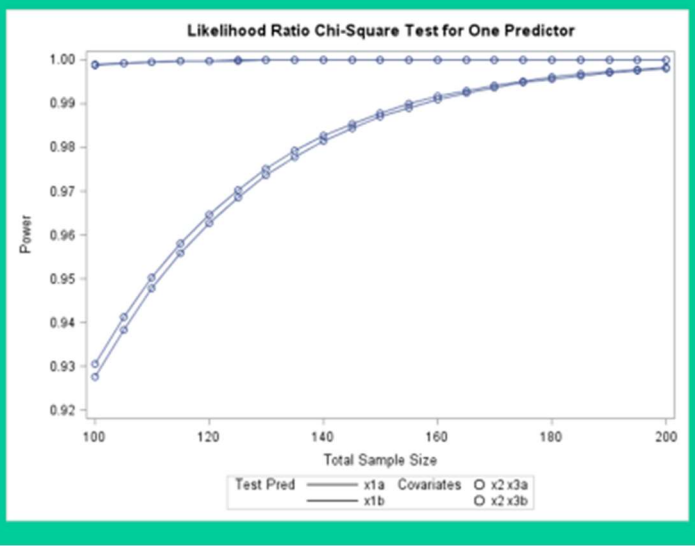


Figure 18

EXAMPLES OF REPEATED MEASURE ANALYSIS OUTPUT FROM PROC GLM POWER

Using PROC GLMPower is more complicated than using PROC Power because the research issues addressed are more complicated.

Figure 19 shows the code for a repeated measures problem.

PROC GLMPower requires the creation of an extra file that is used to pass parameters to the procedure.

Example 47.3 Repeated Measures ANOVA http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_glmpower_examples03.htm

```
Data Pain;
input Treatment $
    PainMem0     PainMem1Wk
    PainMem6Mo PainMem12Mo;
datalines;
SensoryFocus    2.40  2.38  2.05   1.90
StandardOfCare 2.40  2.39  2.36   2.30;
Run;

ods graphics on;
PROC GLMPower data=Pain;
CLASS Treatment;
MODEL PainMem0 PainMem1Wk PainMem6Mo
      PainMem12Mo  = Treatment;
repeated Time contrast;
POWER mtest = hlt alpha = 0.01
  power = .9     ntotal = .
  StdDev = 0.92 1.04
  Matrix ("PainCorr") =
     lear(0.6, 0.8, 4, 0 1 26 52)
  CorrMat = "PainCorr";
PLOT y=power   min=0.05 max=0.99
      yopts=(ref=0.9) vary (linestyle by stddev,
      symbol by dependent source); run;
ods graphics off;
```

Repeated Measures

Logan, Baron, and Kohout (1995) and Guo et al. (2013) study a dental intervention on the memory of pain.

The study compare sensory focus to standard care over a period of a year, asking patients to self-report immediately after the procedure and then again at 1 week, 6 , and 12 months.

The correlation is thought to be linear exponent autoregressive (LEAR), with a correlation of about 0.6 and a decay rate of about 0.8 over one-week intervals. X has 2 levels (sensory focus versus standard care), and you allocate to a balanced design. The within-subject factor is time, with four levels (0, 1, 26, and 52 weeks).

Show results

Figure 19

Interpreting the output requires an understanding of repeated measures and mixed models.

That will not be covered here.

Example 47.3 Repeated Measures ANOVA http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.htm#statug_glmpower_examples03.htm

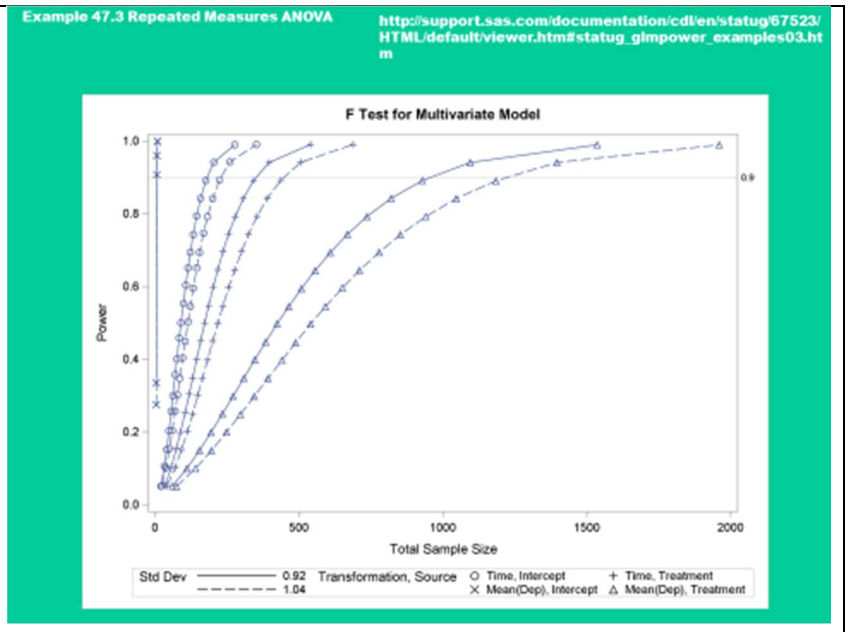
Conjectured Correlation Matrix

Time (week)	0	1	26	52
0	1	0.6	0.491	0.399
1	0.6	1	0.495	0.402
26	0.491	0.495	1	0.491
52	0.399	0.402	0.491	1

Computed N Total

Index	Transformation	Source	Std Dev	Effect	Num DF	Den DF	Actual Power	N Total
1	Time	Intercept	0.92	Time	3	176	0.90	180
2	Time	Intercept	1.04	Time	3	226	0.90	230
3	Time	Treatment	0.92	Time*Treatment	3	346	0.90	350
4	Time	Treatment	1.04	Time*Treatment	3	442	0.90	446
5	Mean(Dep)	Intercept	0.92	Intercept	1	4	0.96	6
6	Mean(Dep)	Intercept	1.04	Intercept	1	4	0.90	6
7	Mean(Dep)	Treatment	0.92	Treatment	1	950	0.90	952
8	Mean(Dep)	Treatment	1.04	Treatment	1	1214	0.90	1216

Figure 20

As we have so often seen, a plot is much more informative than a series of tables. Here is a plot of power vs total sample size for a mixed model.	 Figure 21

CONCLUSION

Power analysis is a critical step in any research project. The penalties that come from under-powering, or over-powering, a research project can be very severe. Millions of dollars and years of research have been wasted because studies were conducted that were under-powered and had almost no chance of detecting that the Y variable changed when the experimenter changed his X variables.

SAS has several different ways for calculating statistical power. It's accuracy and charts make it an excellent choice for this critical part of every research study.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Russ Lavery
Contractor – Bryn Mawr, PA
russ.lavery@verizon.net