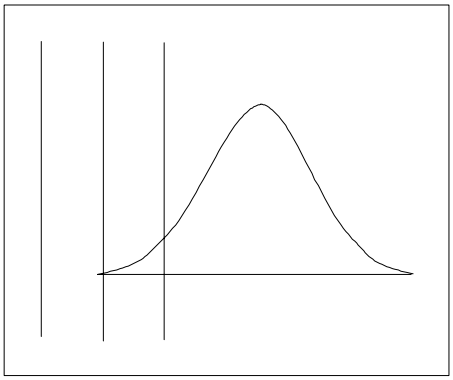
The ECHIP Sample Size Calculator*

Version 1.0

Robert E. Wheeler



Contents

1 Introduction

In todays newspaper (7 April 2000) an AP story reports on a paper appearing in the Lancet from a university hospital in Oslo showing that aspirin is as useful as the method which has been used for years in preventing secondary strokes. The story reports the observed percentages of secondary strokes in the two groups as 8.5 and 7.5, and elaborates with amazement that a simple, cheap remedy can replace a costly alternative. The enthusiastic reporter also tells us that the study was "extensive," involving 449 men and women with irregular heartbeats. Oh, the wonders of modern science, but wait, let me get a pencil. Hum, with 449 subjects I find that it is not possible to discriminate finer than about 10 percentage points with any assurance[1]! This means that the two rates could really differ by quite a bit, and that the closeness of the observed result is very likely due to chance. Now I have not read the paper, and the enthusiastic reporter may have got it all wrong, but the story is common enough, as attested by quite a few surveys of the literature in several scientific fields.

Cohen (1962) called attention to the phenomenon in the psychological literature. In his book, Cohen (1988) he elaborated further and discussed it at length. Others have reported the same thing: Bailar (1992) has a chapter on it with respect to medicine, and Sedlmeier, and Gigerenzer(1989) and Rossi (1995) who surveyed the literature some twenty years after Cohen's book tell a story of too little change. The tale is much the same if not worse in the physical sciences – thousands of experimental designs are run every year by engineers and scientists without adequate evaluation. Indeed most courses which teach industrial experimentation never mention the topic.

Cohen's survey concentrated on the power of statistical tests assessed after the fact, which can indicate too small sample sizes to detect effects of practical interest. The underlying problem is a failure to match resources to the task. Not everything one desires can be done. Resources such as subjects and time are always limited, and some projects are impossible of fulfillment. In engineering and the physical sciences where I have been a consultant for a time, it is always possible to work around a problem. To find another way. When I tell an engineer that an idea is infeasible with the resources at hand, invariably a new idea is developed, and often what was at first thought to be the goal is discarded and a new one substituted. The new idea may not come at once, but if the need is there, it will come, and it will be one capable of execution. I have consulted enough in psychology, medicine, and social science to know that this can happen there too. The consulting statistician well knows of what I speak, for too often a bedraggled experimenter appears on the doorstep lugging a mass of data, about which the statistician can only provide assurances that it is well and truly dead[2].

A sample size calculation is the resource evaluation methodology for controlled experiments. This methodology has been available since Neyman-Pearson (1933), and although widely acknowledged in academic circles, it has only re-

[1] The width of a 95% confidence interval on the difference of two proportions in the neighborhood of 0.08 is approximately 10 percentage points.

[2] Cribbed from R.A. Fisher.

cently begun to be used in actual practice. The methodology, using power for sample size determination, was laid out very carefully in Scheffé (1959), but it seemed to attract little attention in spite of Scheffé's considerable reputation and the adoption of his book as a teaching standard. To a large extent the recent popularity is due in the social and behavioral sciences to the efforts of Jacob Cohen, whose massive book, Cohen (1969,77,88), laid out the methodology in exhaustive detail. His work was largely ignored from the first publication until the early 1990's, but his persistence and many papers seem to have had some effect; but clearly not as much as had been hoped for according to the surveys cited above.

Part of the difficulty seems to lie in the awkward nature of the power paradigm, which insists on precision about alternative hypotheses. In many cases, an alternative hypothesis must be manufactured in order to use the theory. Consider the qualification of a generic drug, where the null hypothesis is not the straw man, but in fact something to be proved! Many studies in the social an biological sciences are indeed intended to establish the validity null hypothesis.

A better procedure in such situations is to select a sample size such that one may claim the null with reasonable assurance. For this reason an alternate calculation is offered, wherein sample sizes are calculated such that the resulting confidence intervals will be suitably small. The idea has received considerably study. It is of particular concern to those who deal with bioequivalence problems such as the qualification of genetic drugs: see Westlake (1979) for example. Jason Hsu (1996) has written extensively on the subject, and has made the calculations and sample size software freely available over the Internet[3]. Perhaps the availability of such a calculation on a Palm device will increase its visibility, even though it is tucked away and must be consciously accessed so as not disturb those who are satisfied with the usual methods based on power.

In Wheeler (1974) I indicated practical methods for linear models. My methodology differs substantially from Cohen's, in that I focus on the response scale and differences in its values as the most immediately understandable quantities, and phrase everything in terms of this scale. Cohen references values and combinations of the parameters, which is convenient from a computational point of view, but by and large parameters are not things about which most investigators can make informed judgments. I say, for example, "is a 10% improvement in quality of economic value to you," whereas, Cohen prefers to ask, "how small will you allow the sum of squares of the main effect parameters to become." There are of course other differences, and they are discussed later. It should be noted, that I am far from alone in this viewpoint. Fleiss (1981) for example, goes to considerable trouble to ensure differences of interest are expressible in a meaningful way.

My emphasis is on experiments in which a variable or factor is under the control of the experimenter; that is in experiments where one can make a change and observe a response, hopefully. For such, the nature of the response should be well enough understood so that the experimenter can make meaningful judg-

[3]http://www.stat.ohio-state.edu/ jch/

ments about differences on the response scale. This contrasts with "sampling" situations, where one samples from a population while observing two or more variates and judges the merit based on some measure of dependence, such as a correlation. The response in this case is the measure of dependence about which it is usually difficult to make meaningful judgments with respect to differences on the scale, which is often dimensionless. For such sampling situations, there is always a question about the degree to which the statistical assumptions are satisfied, which adds to the difficulties in justifying a sample size calculation. In my experience, too many such situations arise because of inadequate thought.

The sample size calculations provided deal mainly with linear models under various distributions. There is a module devoted to response surface experiments, since that is the business of ECHIP. In addition, there is a multiway module, which deals with t-tests, main effects, interactions, for normal, binomial, Poisson, and chi-squared data. The binomial includes logistic regression and contingency tables are treated. There is a general sample size calculation module which enables sample size to be calculated for any sort of linear model with data from the distributions cited – part of this is an accurate calculation of the noncentrality parameters of the noncentral F and chi-squared distributions.

The Exposition section contains an introduction to the ideas at a very elementary level. It discusses important practical matters that are not always given their proper weight; and of course all the remainder is phrased in these terms. In particular, I define three terms of importance: (1) *least difference of interest* (LDI); (2) *resolution bounds*; and (3) *contrast*. These terms are key to understanding and using the software.

For the convenience of those who are familiar with Cohen's methods, the software supports his effect size protocols, but the emphasis is on linear functions and contrasts, which are more immediate to the experimenter's need, and which I trust will be found, by those whose patience I try, to be more useful in the end. I even allow the calculations to be run in reverse, so that one can find the power or resolution as a function of the sample size: a practice subject to misuse, as discussed in Appendix B.

2 Exposition

2.1 What is a sample size calculation, and why is it needed?

This section is for those who know nothing about the subject.

2.1.1 All you need to know

No matter how carefully data is collected, it is common to obtain different values on repetition. Often the investigator will refine the technique in an attempt to reduce the variation, but there always comes a time when nothing more in this line is practicable, and other means must be sought. This is where statistics comes into it, and the variation is reduced by averaging. For most distributions,

the variation in data averages decreases as the square root of the number of observations increases. This is not a particularly efficient way to reduce variation, and should be resorted to only after other methods of refinement have reached their limit, but that is what statistics offers.

Sample size calculations are all about finding the number of replications to achieve a desired degree of precision. The precision of a measurement is commonly indicated by plus an minus bounds. Thus one might say of a measurement, that it is 5 inches plus or minus 1/10 th inches. This usually is just a rough indicator of the precision, and one is often a bit vague about the 1/10 th part. In statistics, such roughness is replaced by exact statements, and one says "5 inches plus or minus 1/10 th inches with 95% confidence," where the *confidence* is a probability that can be calculated according to a formula and which has an exact meaning in terms of repeated frequency. We will not delve into this precise meaning too closely, since it will take us out of our way. For the curious, any elementary statistics book will provide details. If you are really a novice, try Gonick and Smith (1993) or Freedman (1991).

The width of the plus or minus bounds shrink as the square root of the number of observations increase, and sample size calculations are nothing more than calculations which find the number of observations that make the bounds equal to the least difference of interest, LDI. That is it, and now you understand what it is all about. If you use either the Scheffé or Tukey kernel, the sample size will be calculated in this fashion, and the Power kernel will produce sample sizes that do the same thing.

Unfortunately, there is much confusion about sample size calculations. Most of the confusion comes from the fact that most common sample size calculations involve the power of a statistical test, which is a hard concept. It involves two probabilities, not just the one called "confidence" above. You can stop reading right here and never have to worry further about the details if you adopt the rule to always set the probability called "alpha" to 0.05, and to always interpret the probability called "power" as meaning "confidence." If however, you need more, read on.

2.1.2 More details

Figure (1) shows the scatter of a measurement with a 95% region marked in curve A. The proportion outside this region, above the "Critical value," is referred to as "alpha," and called "the significance level," or sometimes "the size of the test." The width of the curve depends on the sample size. The larger the sample size, the narrower the width. There are two curves, A and B, marked in this figure. As the sample size increases, both become more concentrated about their centers, and the overlap decreases. A sample size calculation finds the sample size so that the overlap becomes small to the degree specified by the investigator. In a power calculation, one always supposes that there is a type B curve, and that the problem is to say whether or not an observation comes from one or the other.

For example, one might measure the weight of test animals, and curve A

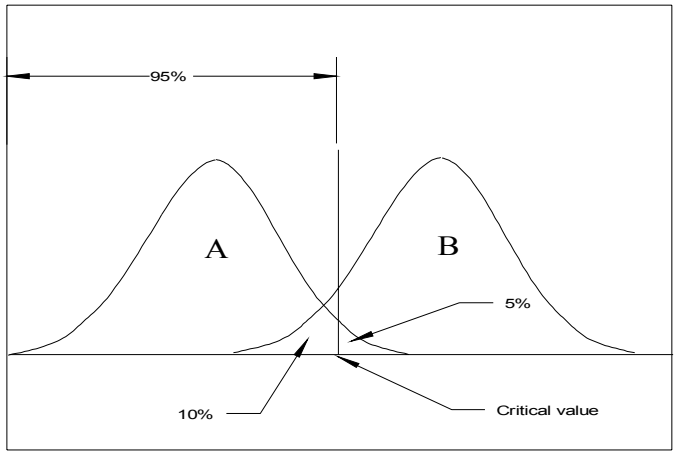


Figure 1: Confidence limits and power

might represent the normal scatter in the measurements. Measurements in the alpha region would be unusual, and if this unusual behavior could be induced through some treatment, then there would be justification for a claim of a significant effect. The claim that would be made is that the treatment has changed the weight so that in fact distribution A is no longer correct, but in fact the B distribution represents the scatter about a different weight center. You have thus tested the hypothesis that the treatment had no effect against the alternative that the treatment had an effect represented by B.

Before running the test, the sample size is chosen so that the scatter about both centers will be sufficiently concentrated so that the result will be unambiguous. This is the power sample size calculation. Power is the probability that a B curve will be identified when the treatment actually has an effect – it is the 90% of the B curve area on the right hand side above the "Critical value."

In the sample size calculation, the LDI is set equal to the distance between the centers of the A and B distributions, and a sample size is found such that the overlap will be whatever the user wants, as controlled by choosing alpha and power.

It will turn out after the data has been collected, that the width of the plus and minus confidence bonds about the observed measurements are approximately equal to the LDI when the confidence coefficient is set equal to the power that was used. Thus the power calculation has achieved the same end as before, the control of the width of the confidence interval. It has however, required the investigator to be specific about B, something that is not always easy to do.

2.1.3 A final thought

Models are at best approximations. This is especially true in statistics where one assumes a distribution for the data. It follows that if enough data is taken, there will be a discrepancy between the model and the data, and a statistical test will show a significant result. Indeed with enough data, the level of significance can be made as small as one desires! Berkson (1938) with his tongue only partially in his cheek asked what is the point in applying a test to a moderate or small sample if we already know that a large sample will show a highly significant result?

The problem arises of course from a focus on "statistical significance," which can be made arbitrarily small with enough data. If one focuses instead on the estimate and the errors in its measurement, then all that large quantities of data will do is increase our confidence about the location of the parameter. Even when the model is in error, as it almost always is, the estimate will usually still be of value and the plus or minus bounds will shrink appropriately with the sample size. It makes sense, therefore, to concentrate on the distance between these bounds rather on the statistical significance, and this should be the goal in choosing a sample size – to adjust this distance to an appropriate size, say the LDI. Either power or a direct calculation based on the width of the confidence interval may be used to achieve this goal.

2.2 Least difference of interest, LDI

It is obvious that the larger the data set the finer the discrimination. It is also obvious that the greater the underlying difference between things the easier it will be to discriminate between them. Detecting a difference in, say, burglary rates between two affluent suburbs will surely take more data than will detecting a difference between an affluent suburb and an inner city area. These are fundamental ideas about sample size and differences of interest, and are the essence of the power problem, which is to decide on the sample size needed to detect a difference of a given magnitude.

In research, resources are always limited, and it is important to adjust the quantity to the need. Taking too much data is undesirable for a variety of reasons: cost being one, but perhaps as important is the fact that excess data can lead to discriminations which are unimportant in a practical sense, which being present must be reported and explained, sometimes to the embarrassment of the researcher and elation of the critics. The task is therefore to find the "right amount" to take.

Let us start with a very simple situation. Suppose we have two treatments (or categories, or types, or what you will), such that the data taken in both may be assumed to come from the same probability distribution, and the only difference between the two will be with respect to their means. Let μ_1 and μ_2 be the population means (or parameters) and $\hat{\mu_1}$ and $\hat{\mu_2}$ be the sample estimates of these means when a sample of n observations is taken for each treatment; that is $2n = N$ observations in all.

The goal is to make a decision about the magnitude of $\delta = \mu_1 - \mu_2$. Statistical tests usually suppose δ to be zero, and make their calculations on this assumption. The reasoning being that if one obtains data which should occur infrequently on the assumption of $\delta = 0$, then one may reasonably argue against $\delta = 0$ and conclude that an effect is present.

The magnitude of such an effect is important, because clearly, some magnitudes may be too small to be of practical interest. In the burglary example above, a difference of one burglary per hundred years is unlikely to be of practical interest, while one per week may well be. The idea involved here is central. In all practical situations, there will be some magnitudes too small to matter and there will be some that are large enough so that they must be considered. Finding that your watch looses one second a week will hardly matter to anything you care to do. Finding that it looses a minute a day may cause concern, and five minutes will surely bother you.

Somewhere between the unimportant practically and the clearly important lies a demarcation value, called the "least difference of interest," or LDI for short. In most situations it is likely to be a region rather than a point, a band of uncertainty; but to get on with things let us suppose it to be a point with a definite value. This value is central to making statistical decisions, and in deciding on sample size. Different things will be concluded from the experiment depending whether or not one decides that δ is less than or greater than the LDI, and greater sample sizes will be required to detect small δ's than to detect larger ones.

The LDI needs to be decided upon before an experiment can get underway. Too often it is not so decided, and in some situations it is quite difficult.

In many cases, it is a mutable quantity that must be arrived at by adjusting one's expectations. For example, marketing may decree that the product will meet certain specifications. Say, they decide to advertise a failure rate of one per thousand. To achieve this a research effort involving the testing of 20000 specimens over a six month period will be required, at a cost of $200,000. When confronted with this fact, it becomes apparent that the benefits are inadequate, and so a compromise failure rate of one per hundred is agreed upon which results in a testing effort of $10,000. The LDI was changed. This frequently happens. There are usually many roads to a destination, and careful considerations may show that the first route is not always the practical choice.

In the social sciences, the units of the scales are not always well understood, being in many cases almost artifacts of the procedure. Let me cite the example discussed in the G*Power tutorial[4]. Cognitive psychologists studying amnesic memory observed a score difference of $16 - 14.5 = 1.5$ between amnesic and normal in a stem completion test[5]. On recognition tests, the difference had been observed to be $13 - 8 = 5$. The difficulty is that these are simply scores and one has little criteria to use in deciding what the LDI should be. I will

[4]Currently available at http://www.psychologie.uni-trier.de:8000/projects/gpower.html This is a very nice free program, and the documentation contains an informative "state of affairs" discussion as of 1997.

[5]e.g. give a word starting with sci.

discuss this example in more detail later, but for not let it suffice that such scales occur frequently and must be dealt with. It is not a problem that can be shifted to other derived quantities as is often attempted, but must be dealt with directly, and when compromise is necessary, it should be along the lines of the previous paragraph: that is changing one's expectations rather than bulling ahead and taking data that cannot achieve a goal.

2.3 Resolution bounds

Now that the idea of a LDI has been presented, we move on to discuss ways in which data may be acquired for its evaluation. If there were no statistical scatter in data, then one would simply compare the LDI with δ and make a judgment about importance. There is scatter, however, and so it must be accounted for in some way. The simplest way is to construct an interval to show the band of statistical scatter of $\hat{\delta}$ from sample to sample. Suppose the interval is $\hat{\delta} \pm k$. Now if zero is outside this interval and if the interval really represents the range of fluctuations of $\hat{\delta}$, then it should be clear that the parameter δ is unlikely to be zero. With suitable assumptions, this is equivalent to a statement that $\hat{\delta}$ is statistically significant[6].

The interesting case, from our viewpoint, is the other one, when $\hat{\delta}$ is not statistically significant. That is when zero falls inside the interval $\hat{\delta} \pm k$. This is the case of a Scotch verdict – not proven. There are two possibilities: (1) δ is truly zero: (2) δ is non-zero, but too small to be detected by our procedure in view of the statistical scatter. It is like an election where there is a suspicion of miscount. Clearly if one candidate has won substantially, there will be no problem, but if the election is close the possibility of miscount can obscure the results. It all depends on how large the miscount is. If it can be shown to be small with respect to the vote difference then matters can proceed, but if not, it is a Scotch verdict.

The diagram in Figure (2) will help. It shows an observation bounded by a $\pm k$ interval which includes zero. Since this interval represents the band of statistical scatter, it is clearly possible for the parameter δ to be zero, and hence one could not claim a non zero value with any justification. However, note that the upper limit of this band, marked as "Resolution Bound" in the diagram is a limit beyond which one would be surprised to find δ. This resolution bound, is a reasonable limit on how large δ might actually be, and as such it can be compared with the LDI to make informed statements about the ability of the data to resolve the LDI.

The resolution bound is a measure after the fact[7] of the quantify, or in

[6] In case this seems abstract, let me point out that this is the usual situation when making a t-test. One supposes that the common distribution of the observations is the normal distribution with a variance σ^2, and thus the constant k is $t_{df}(\alpha) \times \hat{\sigma}/\sqrt{df}$, where $t_{df}(\alpha)$ is the 100α percentage point of a t-distribution with df degrees of freedom and $\hat{\sigma}$ is the sample estimate of σ. The statement that the confidence interval excludes zero is logically equivalent to the statement that the t-test is significant at the α level.

[7] Or post hoc as some are fond of saying :)

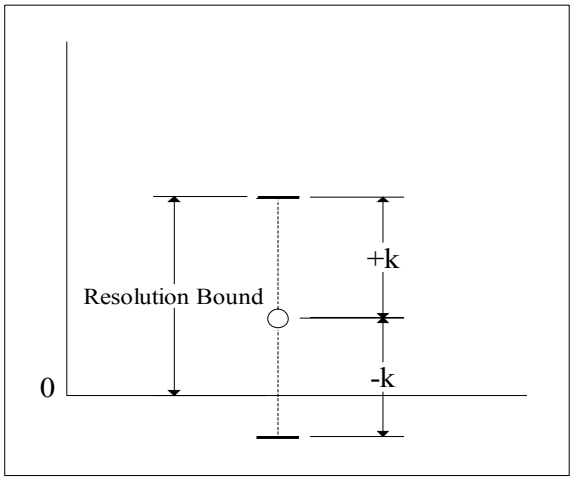


Figure 2: Resolution

terminology to be made precise, of the revolving power of the data. If in the election miscount, it could be established that the resolution bound of the count is $\pm 1/8\%$ and that the two candidates differed by $1/2\%$, then a recount will not be needed, since it will not change the results.

Earlier I mentioned a cognitive study cited in the G*Power documentation. In this study the $\hat{\delta}$ for the stem completion test was 1.5. The sample sizes were 4 and 8 subjects each, and the $\hat{\sigma}$ was about 3. This gives us, approximately, an interval of 1.5 ± 3.8[8], and a resolution bound of about 5.3. Since the observed $\hat{\delta}$ on a recognition test was 5, and since this was considered an important difference in practice, it is quite clear that the stem completion test with its resolution bound of 5.4 was inadequate to detect differences of interest. I will discuss this further in a bit, because this sort of problem is sometimes treated with what I believe to be a questionable methodology.

A resolution bound, or its generalization as appropriate, may be defined for any statistical quantity. For the usual estimates it is quite simply the maximum absolute confidence limit of the estimate[9].

In all cases, a comparison between the LDI and the resolution bound after the completion of an experiment is a complete expression of the ability of the experiment to resolve differences of interest. If the LDI is approximately the same size as the resolution bound, then the experiment was adequately designed, and one may conclude with reasonable safety that non-significant quantities are of no practical interest. They still remain unknown, but whatever they may be they are unlikely to be larger than the resolution bound which matches the LDI and hence of no practical interest. The Scotch verdict has been set aside.

[8] t-test 95% confidence limits assuming normality.

[9] It is important to note that the resolution bound is not equivalent to the bound that may be obtained by solving for effect size as a function of power: Hoenig and Heisey(2001)

In most experiments one has several parameters of interest, and it is quite possible to adjust the experiment so that the resolutions of all estimates are consistent with the LDI's. Figure (3) shows the output of the analysis of a near optimal experiment that was designed to detect a LDI of 4.5 psi. I will not explain the output in detail. It is a standard ECHIP effects output table, which is similar to a regression analysis output. The model terms are listed on the right. I want at this time to call attention to the column labeled "RESLTN." This column gives the resolution bound for the non-significant terms in the model, and it may be seen that they are all in the ball park of the design goal 4.5, and in this case the experimenter was able to dismiss those terms and thus the pressure variable from consideration. Whatever the values of the parameters, they are unlikely to be larger than the LDI and thus to have no practical effect on the process.

Although a side issue, it is important to note that the LDI and indeed the effects estimates in this table are in the units of the response, which was psi, and thus something about which the experimenter could form judgments based on experience. To appreciate this point, one only has to think of an equivalent analysis using standardized regression coefficients, where the basic variables have been scaled to a -1 to +1 range. What, pray, does a 0.1 change in one of the coefficients mean, and how does it relate to LDI?

2.4 Contrasts

The statistics of most interest in an analysis are linear functions of the observations. The grand mean of the data is such a linear function, since it is simply the addition of the data values. Estimates of main effects are linear functions of the observations, as are estimates of interactions and regression coefficients. The estimated difference, $\hat{\delta}$ between the sample means of the two treatments in our simple example above is a linear function of the observations. All such statistics are estimates of parameters, just as $\hat{\delta}$ is an estimate of δ, and it is in terms of these parameters that we formulate hypotheses.

One may perform a sample size calculation for any function of the parameters, but linear functions are the most common, and the most common linear functions are the contrasts. A contrast is just what its name implies, a contrast between parameters. the linear function $\delta = \mu_1 - \mu_2$ is a contrast, since it contrasts μ_1 with μ_2. Such contrasts are commonly referenced by a vector [1,-1], meaning $\delta = 1 \times \mu_1 - 1 \times \mu_2$. A contrast for three parameters might be [1,-2,1], which contrasts the extreme parameters with the middle one. The defining characteristic of a contrast is that its elements sum to zero. Another contrast for three parameters is [1,0,-1] which contrasts only the two extremes.

A common hypothesis is that a parameter is zero, but such a hypothesis could be verified only with infinite data, so we agree to a compromise and settle for the ability to say something like, "if is not zero, it is unlikely to be larger in magnitude than Δ." We are guided in our choice of Δ by the LDI.

One can nominate Δ's for either of the three parameter contrasts above.

```
EFFECTS  RESLTN SIG  TERM

  89.935               0 CONSTANT
  16.118         ***   1 temperature
  -1.271   4.168        2 pressure
  14.977         ***   3 duration
   0.283   3.523        4 temperature*pressure
 -63.471         ***   5 temperature*duration
  -0.809   3.813        6 pressure*duration
 -17.792         ***   7 temperature^2
   0.838   3.266        8 pressure^2
 -14.543         ***   9 duration^2



 Residual SD    = 2.088168
 Replicate SD   = 2.522499

 N terms            = 10
 N unique trials    = 15
 N replicates       = 5
 N total trials     = 20
 Cross val  RMS = 1.906303
```

Figure 3: Effects output

The meaning attached to the Δ[10] is different in the two cases. For [1,0,-1], the Δ corresponds to the minimum absolute value of a contrast between two parameters, which relates directly to LDI. For [1,-2,1] it corresponds to a contrast between the middle and end values. In general when one has three parameters, both contrasts are likely to be of interest, and it may well be that the sample sizes required to detect the two Δ's are different, since the sample size depends on both the value of Δ and the elements in the vector.

A quandary? Not much of one, since it is easily seen that the larger sample size will do for both. It turns out that [1,0,-1] requires the larger sample size, so this is the one that will usually be considered. As far as the sample size calculation, it does not matter how the parameters are assigned, since the sample size depends only on the elements in the contrast: i.e. [1,0,-1],[-1,0,1],[1,-1,0],[0,1,-1],[-1,1,0], and [0,1,-1] require the same sample size.

Popular methodologies currently in use follow Cohen and do not focus on single contrasts, but rather on an omnibus combination of them. Such calculations apply to all possible contrasts, the majority of which are of little practical interest. It would, for example, be rare for the contrast [1/3,1/2,-5/6] to be of interest, but this and all of an infinity of others are included when the popular methodologies are used. Now this may not seem to be a terribly important point, since by using an omnibus combination one will surely capture those contrasts of practical interest, and the resulting sample sizes are often not much different. But there is something more important involved. That is the ability of the experimenter to relate the omnibus value to practical differences in the response.

Any single contrast can be directly related to the LDI. In the case of a two-level contrast, like [-1,0,1], the relation is obvious; the Δ corresponds directly to the LDI. For a contrast like [1,-2,1] it turns out that one half the Δ is equal to the LDI[11]. For the contrast, [3,-1,-2] one has LDI equal to 5/14 of the Δ. etc. Thus for any single contrast, one may find the equivalent LDI, and thus judge it in terms wholly familiar to the experimenter, assuming of course that the experimenter understands the response scale. The omnibus combination is nearly impenetrable to interpretation; so much so that Cohen devised a three point rating scale, "small," "medium," and "large" to allow judgments to be made. It is very nearly impossible to relate the omnibus combination to LDI's or anything of practical interest in terms of the response scale. The concerned reader might like to read section 3.1 of Fleiss (1981) for illustrations of several ways to be specific about this matter.

The calculations focus on contrasts. The Response Surface and Multiway Effects modules ask the user only for the LDI, and produce by default appro-

[10]The value Δ is called a *detectable value*, and is the smallest absolute value of a function of the parameters that may be detected. To be precise, we mean that a *detectable value* is it is the smallest value of a function of the parameters that will produce a significant result with at least the power designated. Such detectable values when combined with an assumed probability distribution for the data and with the additional parameters α and *power* may be used to calculate a sample size. The ECHIP power calculation program does this calculation.

[11]Details may be found in Appendix A.

priate sample sizes for two-level contrasts. The Multiway Effects module allows the user to specify particular contrasts of interest.

2.5 Effect size, σ and the signal to noise ratio

A common population standard deviation, σ, is assumed for normally distributed data. Since it is a population parameter, its actual value is unknown, and to make progress with sample size calculations, some value must be assumed. The resulting calculations can be only as accurate as the assumed value, which in practice means that all sample size calculations are approximate. In general, a sample size calculation for normally distributed observations should not be taken as a precise calculation. Doubling or halving the calculated sample size is usually quite acceptable. The calculation should be taken as an "in the ball park" value, and no extensive effort should be made to achieve it exactly. The greatest value to be found in calculating sample size is to exclude from consideration those proposed projects whose goals can never be attained with the resources at hand.

In the equations used to calculate sample size, σ appears only in the ratio Δ/σ, and this leads some to treat this as a "signal to noise" ratio and to ignore its separate components. This is poor practice, since each parameter in the ratio has a definite meaning and deserves separate consideration. The Δ is important because it is referenced to the LDI, which is a value of immense importance and which requires careful consideration as has been discussed. The σ represents the variability in the data, and needs to be determined as precisely as possible from empirical studies or from a deep understanding of the data generating mechanism.

To be clear, σ represents the statistical error in the observations. It may be estimated by replicate tests. If for example, the observations are viscosity measurements of paint batches, then an estimate of σ may be obtained by repeating the setup and production of such batches and calculating the sample standard deviation of such replicates. It is important not to confuse such full scale set up repeats with the variation that might be observed by sampling repeatedly from the same setup, since this second variation will usually be smaller, and not represent the setup to setup error. As another example, consider a cognitive test to be applied to individuals. The σ should be an estimate obtained from replications of this test on individuals, and not from replicate tests on the same individual from time to time. In short σ should be estimated in exactly the same way that it will be in the actual testing.

The failure to keep these two parameters separate will result in a vagueness in the results that may well obviate their utility. The current practice of ignoring this advice by using a signal to noise combination such as Cohen's f, leads to the need to choose vague "small," "medium," and "large" criteria. Such choices place no restriction on the LDI, and their use conceals the fact that small, medium, and large sample sizes are being specified. Consider an experiment in which the quantity of interest may be measured with different instruments – say a micrometer and a desktop ruler. The use of a "medium" criterion will

produce the same sample size for the school ruler and the micrometer even though a specified LDI will be more difficult to detect with the desktop ruler. The idea is pernicious. Lenth (2001) makes some interesting comments about this.

Cohen was quite proud of his systematic treatment of this three element classification, since he says, referring to it, in the preface to Cohen(1969), "Whatever originality this work contains falls primarily in this area." I am sure others feel as he did, but I am also sure that there are those who see it as I do as an obfuscation, and the discarding of important information.

3 The number pad

Figure 4: Number Pad

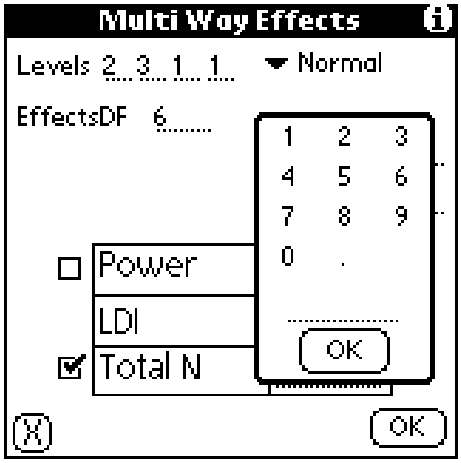


Figure (4) shows the number pad which appears when any field is clicked. It may be used to either select digits, or write in values using Graffiti. It may be turned on and off by clicking on the "NumberPad" menu item. When it is turned off, values may be entered into fields using Graffiti. As these values are entered the other fields automatically change.

4 The Response Surface Module

This module is intended to be straightforward to use, and is structured to deliver good answers for those doing response surface experimentation.

The initial screen is shown in Figure (6). Many users will never need to do more than is available on this screen. This screen shows that the required sample size is about 18 trials for a linear model when both the σ and the LDI are equal.

Figure 5: Initial settings

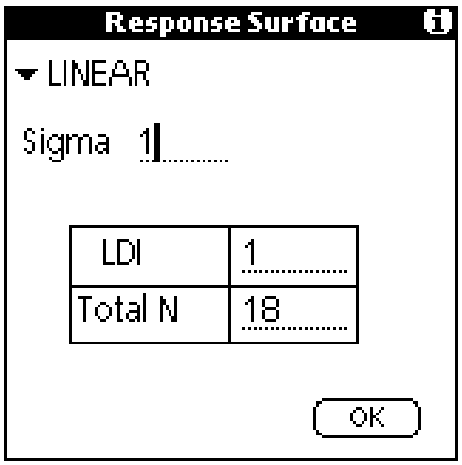


Figure 6: Initial settings

When calculating sample size, attention need be paid only to a single variable pair of variables. The full model may involve many variables, but its nature is irrelevant for the calculation. There are a couple of reasons for this, but the most important ones are that the experimental design is assumed to be highly efficient and that only about 5 or 6 extra trials are taken to estimate the experimental error. This is typically the case with an ECHIP design.

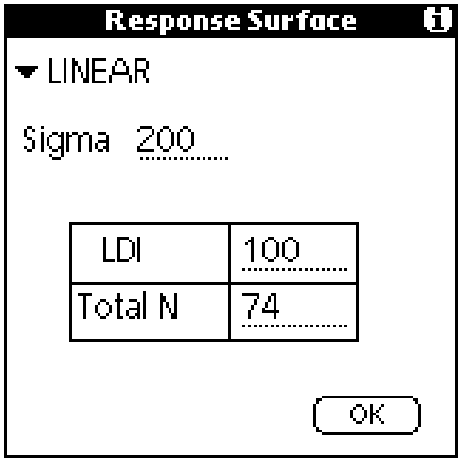
In general, a best guess at the actual experimental error should be input in the Sigma field. The better the guess, the more accurate will be the sample size (Total N) prediction. The sample size that is produced by the calculation should not be taken as absolute. Half or twice the value output will usually be acceptable – this is because of the uncertainty about σ. See Section (4.3.1) for situations in which this rule does not apply.

The experimental design will require a certain number of trials due to its combinatorial nature; but this is only part of the story because the goal is to detect effects of certain magnitudes which is determined by a sample size calculation. If the design is not large enough to detect effects of interest, then it will have to be at least partially replicated to attain the number indicated by the sample size calculation.

The decision about sample size is made by comparing the detection ability of a design with the LDI, the least difference of interest, which is in the units of the response. It is a value that can be discussed and agreed upon as a worthwhile minimum, or changed if that seems wise. After the fact, it will be found that statistically insignificant effects are unlikely to have magnitudes larger than the LDI. The factors associated with such effects can thus be put aside as not worthy of investigation. They may be revisited later if a decision is made to lower the

LDI.

Figure 7: Abrasion example



Example: As an example to firm up ideas, consider an experiment investigating abrasion loss in grams per hour (gph) of a material. A loss of 10 gph is negligible, but a loss of 500 gph is substantial. Someplace between the two one finds a value on the borderline between negligible and interesting. This value is determined by the practical implications of abrasion loss such as cost or appearance factors. This value is the LDI. Say it is 100 gph, and say also that σ seems to be in the neighborhood of 200 gph. Figure (7) shows the calculation, which indicates that a total of 74 trials are needed.

This is a substantial sample size in industry, and may well represent more effort than is possible. If so, there are two choices. (1) Redefine the LDI to a larger value. (2) Think very hard about the test, the measurement and the whole proposal, and try to decrease the measurement error or find a different solution. The second choice is always interesting, and sometimes on consideration, the problem shifts to quite a different one, as for example the realization that abrasion is not the response of most interest, but rather some measure of hardness.

4.1 Model choice

Figure (8) shows the models available from the drop down list. They correspond to those available in ECHIP. The LINEAR and CATEGORICAL models involve a single variable, while the others involve two variables.

The required sample size increases with the complexity of the model, because a more complex model describes a more complex surface. Figure (9) shows the calculation for a quadratic model, which indicates that 91 trials are required to detect a LDI equal to σ. The linear model in Figure (6) required only 18 trials.

Figure 8: Model choices

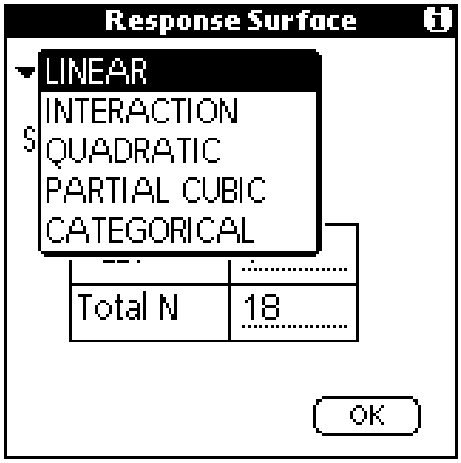
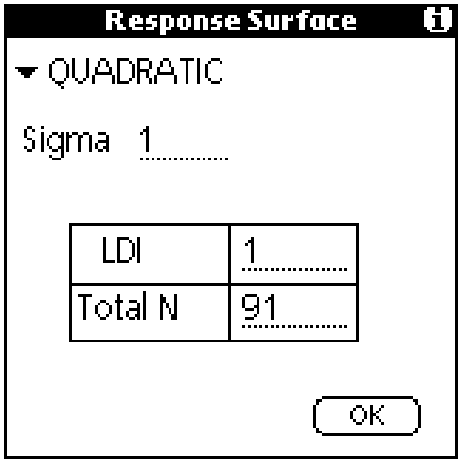
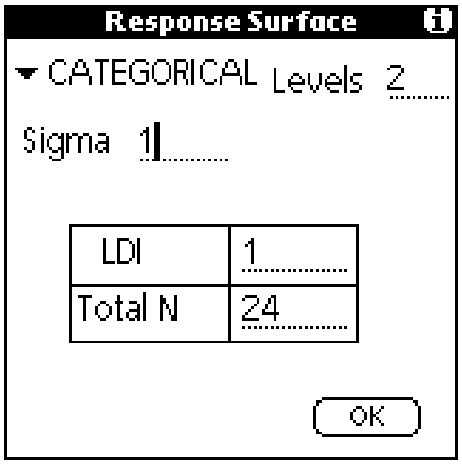


Figure 9: A quadratic model



Because sample size increases with model complexity, the shrewd experimenter will prefer to run experiments sequentially, and build up to the complex models using design augmentation. Of course, this is not always possible, and worrisome problems such as random shifts in the environment between experiments must be considered, but when possible it should be attempted.

Figure 10: A Categorical Variable



The only model that needs special comment is the categorical model, and this is because when it is chosen a Levels field appears, as is shown in Figure (10). Most of the time, a sample size calculation involving a single categorical variable is all that is necessary, and this is all that is provided in this module. If it is important to plan for more, the Multiway Effects Module may be used.

4.2 Guessing σ

Perhaps the greatest good that comes from a sample size calculation, lies in the rejection of bad experiments. If any reasonable guess at σ produces an impractical sample size, then the proposal deserves more thought, and only foolish researchers will proceed to an almost certain failure. There do seem to be a number of such fools around nowadays, as the references cited in the introduction show. People in groups or in committees seem to be more foolish than individual investigators. I seldom encounter a researcher so stubborn as to proceed in the face of sample size evidence, but I have often seen task forces plod on to certain doom.

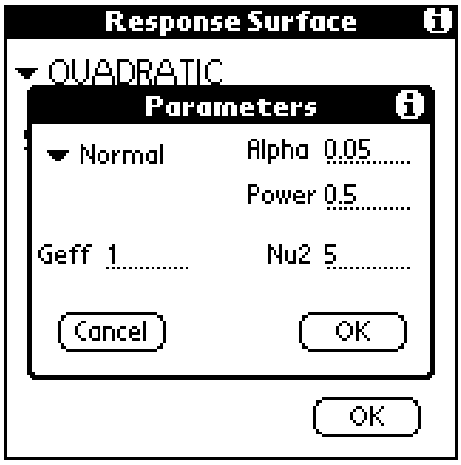
Steps should be taken to estimate σ as precisely as possible. Often pertinent data is at hand, and sometimes replicate trials can be run to firm up a guess; but except in unusual situations, it is not usually worthwhile to expend a large amount of effort for this purpose. Experience indicates that the likelihood of success is large if it is possible to run an experiment with the number of trials

ranging from half to double the value produced by the sample size calculation. There are a number of reasons for this, but basically things are too uncertain before an experiment to justify great precision in this matter. The best strategy seems to be to "get in the ballpark."

There is one point that causes confusion for those who do not often work with statistics, and this is about the sources of variation, and what σ is supposed to represent. The parameter σ is the population standard deviation of response variable. It is the value that describes the variation of this variable upon repetition. It does not represent the error in measuring this variable alone, but the error in the setup of the process as well. Too often engineers mistake the measurement part for the whole. For example, suppose the response variable is the viscosity in a container, and that each experimental unit is a container. The error in the viscosity measurement is only part of the error. The whole of the error is a collection of errors due to the set up and production of the container. The simple rule is to look at the units that will form the data for the final statistical analysis. It is the total error in these units that should be estimated and used as σ.

4.3 Parameters

Figure 11: The parameters dialog



This section is more technical, and should seldom be of concern or interest to most users.

The parameters dialog may be accessed by clicking the menu button at the bottom left of the green pad. Figure (11) shows the parameters dialog that will appear.

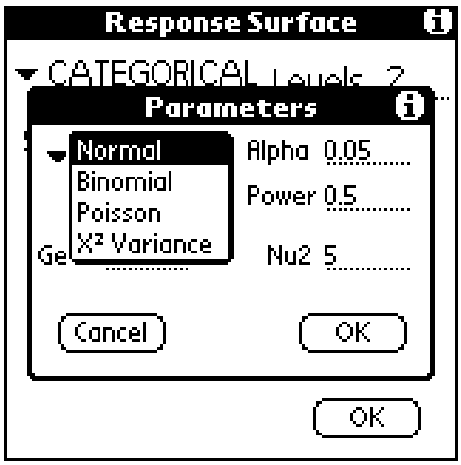
Sample size is calculated using the power of a test. For this the Alpha and Power fields containing the test size and power. The power is set to 0.5 by

default which experience shows is a good value for response surface experiments. The numerator df of the F-distribution is set by the complexity of the model. The denominator df in the Nu2 field is set to 5, since this is the usual value for an ECHIP design. It should be noted that it is also an appropriate value for an estimate of the residual error variance: the gain in power for larger values diminishes rapidly.

The efficiency of the design may be accounted for by adjusting the value of the G-efficiency in the Geff field, although in general there are too many other uncertainties in sample size calculation to make this of much value.

4.3.1 Non-normal distributions

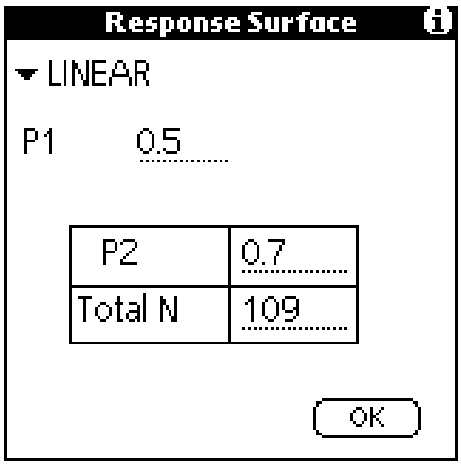
Figure 12: Distributions



The distributions that are available are shown on the drop down list in Figure (12). Switching to any non-normal distribution changes the fields on the main screen. For example, choosing binomial results in the display shown in Figure (13). Non-normal distributions are treated by transforming the response using a normalizing transformation. For such, it is necessary to specify the LDI with respect to two values. For more details the interested reader may consult Section (5.9) in the Multiway Effects documentation.

One final comment is appropriate here. It is necessary to guess the unknown parameter σ when a normal distribution is assumed. This is unnecessary for non-normal distributions, and thus their sample size calculations turn out to be more reliable than the normal ones, even though an approximate normalizing transformation needs to be used in order to make the calculations!

Figure 13: Calculation with a binomial distribution



5 The Multiway Effects Module

This section and the module it describes, assumes some statistical knowledge. Terms like "degrees of freedom," "ANOVA," "regression," "test size," "power," and "probability distribution" are assumed to be familiar and used without explanation. It is unlikely that anyone would attempt a sample size calculation without this knowledge, but for those who are interested an do not feel comfortable, the two references previously given may be helpful: Gonick and Smith (1993) or Freedman (1991), and in addition Moore and McCabe(1993) is a standard elementary textbook containing a great many examples.

5.1 The starting screen

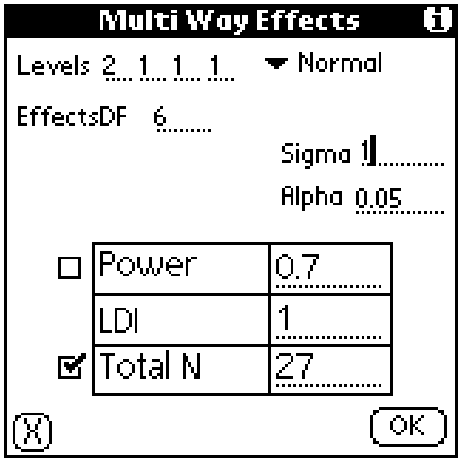
Although this module deals with a variety of distributions and calculations, the central theme is linear models which includes both regression and ANOVA. Even though the calculations are made for particular variables or sets of variables, a more complex model is assumed. This is input to the program through the EffectsDF field. This field should contain the sum of all effects in the model. For regression this is the number of terms in the model, excluding the constant. For ANOVA, the degrees of freedom allocated to effects is illustrated for the three-way case in Table (1). Here there are three factors at levels I,J,and K. The sum of the effects is inserted in the EffectsDF field in the program.

Figure (14) shows the startup screen for the MultiWay module. The top left of the screen shows that there is one factor at 2 levels, and that there are 6 df for effects; that is all factors involved in the model account for a total of 6 df. The "Total N" which appears in the box in the center of the screen indicates that there are 27 observations in all (i.e. the sample size is 27). From Table (1), it may be seen that this means that the "Error" is 27-1-6 = 20 df.

Table 1: ANOVA df

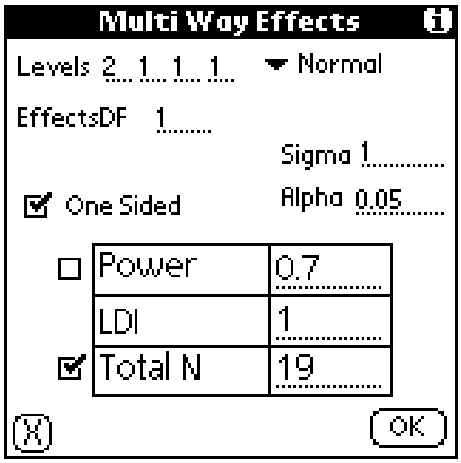
Source	df
A main effects	I-1
B main effects	J-1
C main effects	K-1
AB interactions	(I-1)(J-1)
BC interactions	(J-1)(K-1)
AC interactions	(I-1)(K-1)
ABC interactions	(I-1)(J-1)(K-1)
Effects total (= EffectsDF)	IJK-1
Error (= ν_2)	IJK(M-1)
Total	IJKM-1
Total N	IJKM

Figure 14: Startup screen



The calculation that was made to find his Total N of 27, started with the information that there was one factor at 2 levels and that the EffectsDF was 6. It made use of the other entries shown on the screen and iterated until everything balanced. The result was a sample size of 27. In this case, it means that with a total sample size of 27, a LDI of 1 unit between the levels of a two level factor will be detected with a power of 0.7 for a 0.05 level test. It also means that the width of a confidence interval, or resolution bound, on the observed difference between the two levels will be approximately 1 unit wide.

Figure 15: One sided test

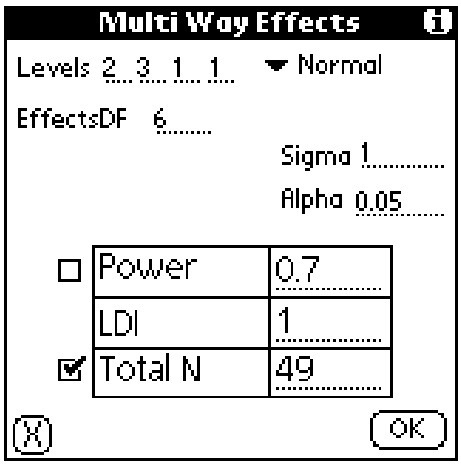


Example: As an illustration, suppose an experimenter were measuring the time in seconds for a chemical reaction to respond to a catalyst, and suppose the population standard deviation of the measurement error was 1 second ($\sigma = 1$), and that this was also the least difference of interest, LDI: then if catalyst A and catalyst B were each used in 13 trials, a difference in their reaction times would likely be detected if the population difference exceeded 1 second.

When the model involves a single two level variable, it is sometimes possible to specify in advance a one sided hypothesis. Thus when there is a single two level variable and when the EffectsDF field is set to 1, a one sided checkbox appears as in Figure (15), where it may be seen that for a one sided test the sample size is slightly smaller.

The calculation was made for only one factor, but there are other factors in the model, and calculations may be made for any of them. For example, if in this same experiment, the interest were focused on the interaction between a two and three level factor, then the calculation would appear as in Figure (16), where it may be seen that now a sample size of 49 is required. It is assumed in this that the experiment giving rise to the interaction is balanced so that the

Figure 16: A 2x3 Effect



49 observations are divided among the 6 cells so that each cell receives about 8 observations.

In practice, the number of trials required by the combinations of the factors places a lower bound on the sample size. In the case illustrated, the design involved at least two factors; one at 2 levels and one at 3 levels. This means that the design requires at least 6 trials for its structure. Had there been more factors, then there would have been more trials required by the nature of the design. It can happen that a sample size calculation indicates fewer trials than are required by the design. It can also happen, as it did in this illustration, that the sample size is larger than that of the design combinations; in which case, some of the design combinations must be replicated.

Replicated trials are used in the analysis to improve the estimate of error, but often there are many more replications than are really needed for this purpose[12]. In the illustration, 49 trials are required, most of which will be allocated in the analysis to the error term. The trials are of course needed to detect the differences of interest, but allowing them to be used only for improving the error term is wasteful. A strategy used by the best experimenters is to include additional factors in the experiment.

5.2 Menus and drop down lists

There are two menus and one drop down list. Figure (17) shows the menus and Figure (18) the list. The menus are accessed by clicking on the menu button at the lower left of the green keypad.

The first menu provides the following options:

1. **NumberPad:** Starts and stops the number pad.

[12]In general the point of diminishing returns sets in at about 5 or 6 df for error!

Figure 17: Menus

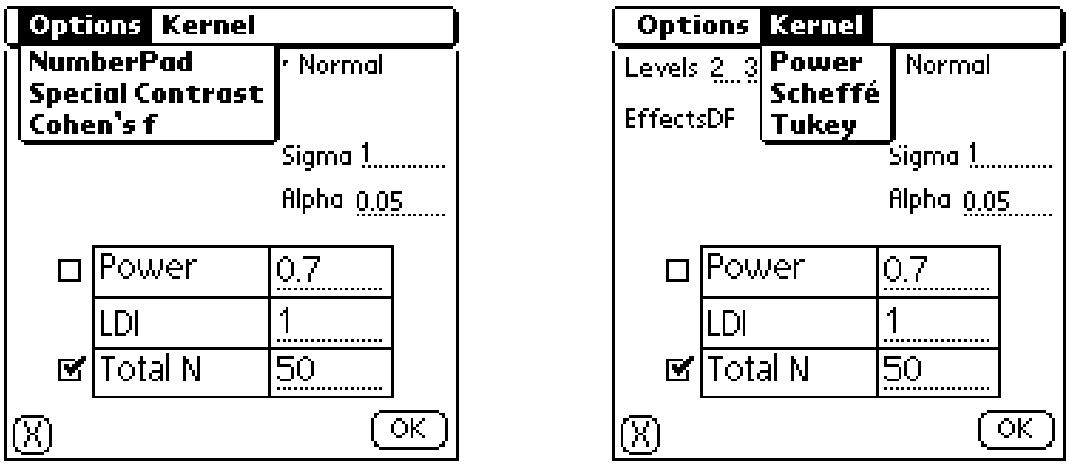
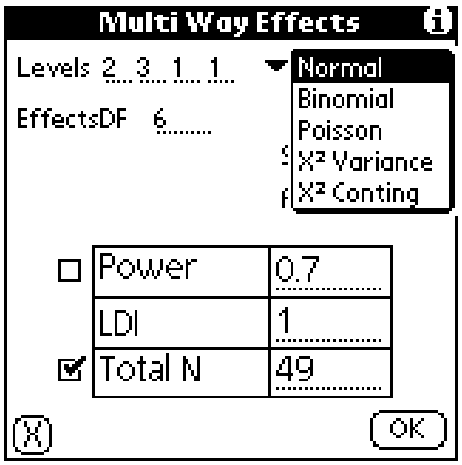


Figure 18: The distribution drop down list.



2. **Special Contrast:** Allows the input of theta values for special contrasts.
3. **Cohen's f:** Switches back and forth from LDI to Cohen's f

The second menu allows the choice between various kernels for the calculations:

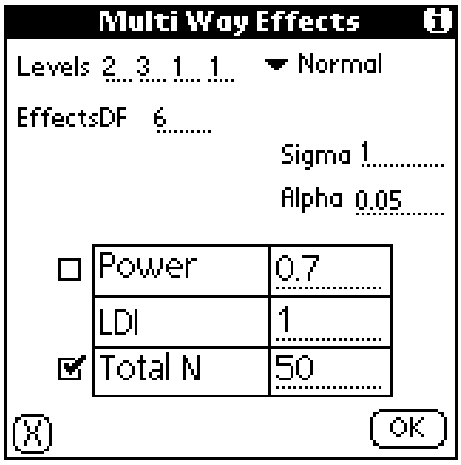
1. **Power:** Finds sample size using power. This is the default.
2. **Scheffé:** Finds sample size using Scheffé's multiple comparison intervals.
3. **Tukey:** Finds sample size using Tukey's multiple comparison intervals.

The drop down lists allows the choice of a distribution:

1. **Normal:** The distribution is normal. This is the default.
2. **Binomial:** The distribution is binomial. This enables logistic analysis.
3. **Poisson:** The distribution is Poisson.
4. χ^2 Variance: The distribution is χ^2, and may be used to analyze variances.
5. χ^2 Conting: The distribution is χ^2, and may be used for contingency tables and goodness of fit.

5.3 Exact Calculation Button

Figure 19: A 2x3 Effect after an exact calculation.



Palm devices are surprisingly capable, but still they are not full size computers, and cannot be expected to perform complex numerical calculations instantly. For this reason, most calculations are performed in approximate mode. It is usually quite accurate. An alternate mode is available for calculations which may possibly be improved upon. When this is possible an "exact calculation button"

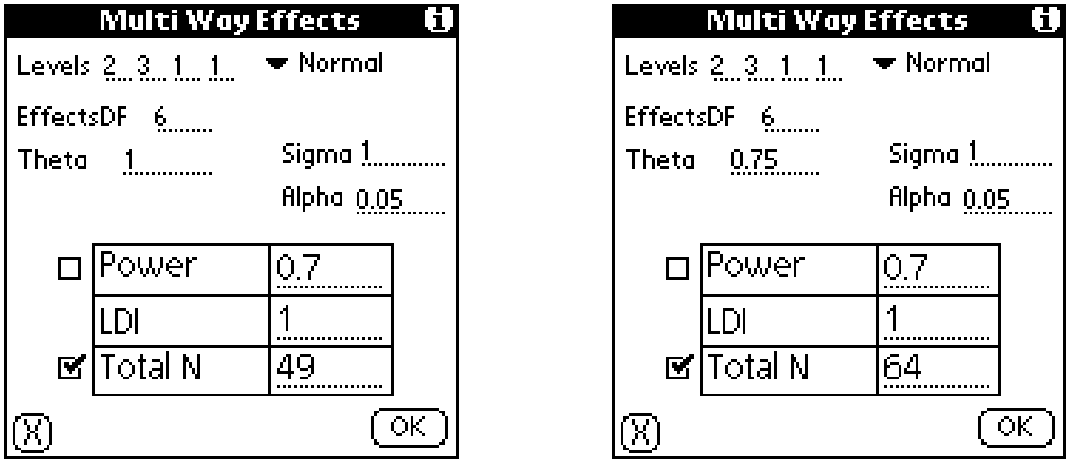
will appear at the lower left of the screen. Clicking on this button will engage an exact calculation. An hourglass icon will appear on the screen and remain while the calculation is being performed. Have patience, let it run its course.

Pressing the exact button for the example in the previous section produces the result shown in Figure (19). It may be seen that the result is only slightly different: a change from 49 to 50 in sample size.

In those rare cases where numerical difficulties prevent an exact calculation, an asterisk will appear beside the exact button.

5.4 Special Contrasts

Figure 20: MultiWayTheta



When the "Special Contrast" menu item is selected. An entry field labeled "Theta" appears. Into this may be entered the specification for a special contrast, as shown in Figure (20). The default contrasts are two-level contrasts, having elements -1 and 1. These are usually the most interesting contrasts. When the special contrast menu item is first selected, the theta for the appropriate two-level contrast is shown, as it is in Figure (20). For this 2x3 table, the two-level contrast elements are as shown in the left of Table (2).

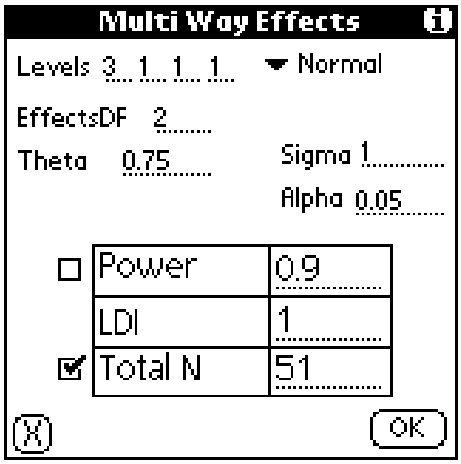
Table 2: Contrast elements for 2x3 tables

1	0	-1
-1	0	1

1	-2	1
-1	2	-1

Theta is defined as the sum of squares of the elements of a contrast divided by its range. For the two-level contrast, this gives $4/2^2 = 1$, as shown at the left of Figure (20). For the contrast shown at the right of Table (2), one has $theta = 12/4^2 = 0.75$ as shown at the right of Figure (20). The sample size has increased to 64. Changing the contrast has therefore made a difference, but please note that this difference is unlikely to be important because σ is unknown. In general σ will not be known within a factor of two, which means that one might reasonably choose to use any sample size from 25 to 100.

Figure 21: Carnauba Wax example



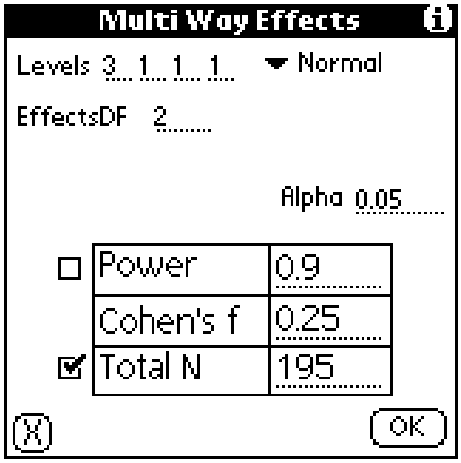
Example: A manufacturer of fiber reinforced plastic products uses Carnauba Wax in a process. There are three suppliers for this wax, and the question of the equivalence of the products has arisen. It has been decided to compare them with a test. If A, B and C denote the three suppliers, and if y is an average value for a process characteristic that will be affected by the wax, then pairwise contrasts like $y_A - y_C$ will be interest, but since supplier C seems to give better service, the comparison $y_A + y_B - y_C$ needs to be examined too. For this last comparison θ is 3/4, and just for illustration suppose $\sigma = 1$ and $LDI = 1$. Figure (21) shows the calculation, where it may be seen that 51 observations will be needed, or about 17 per wax. As usual, however, the pairwise contrasts, require a larger sample size, some 75 observations or 25 per wax, and thus the prudent experimenter will use this sample size.

5.5 Cohen's f

Cohen's f is defined as the standard deviation of a set of parameters divided by the population standard deviation. This is a vague dimensionless unit, that can be related to actual effects only by special arguments. Such arguments are

given in Cohen (1988). The problem is so awkward that Cohen assigned three values to be used universally for all effects (main or interaction). These are 0.10 for "small," 0.25 for "medium," and 0.40 for "large."

Figure 22: Using Cohen's f with a "medium" effect size.



Example: Figure (22) shows a calculation for the Carnauba wax example in Figure (21) using Cohen's "medium" value for f. The sample size is 195. A "large" value would have to be used in obtain sample sizes on the order of those produced for the contrasts in Figure (21). This hardly seems reasonable for the problem, and points up the vagueness of these effect sizes.

It is difficult to compare calculations involving more than one factor with the tables from Cohen (1988). He tabled things in an odd way and made compromises that lead to erroneous results. A better way is to use G*Power whose URL is given on page (10).

5.6 Calculating power instead of sample size

If the checkbox beside Power is checked, power will be calculated instead of sample size. Figure (23) shows the result of such a calculation. The resulting power is 0.699, which differs from the 0.7 that was input originally because the sample size of 49 is an integer and rounding has occurred.

This sort of calculation is occasionally interesting, but it should not be used as a substitute for resolution bounds. The post hoc calculation which is discussed in Appendix B is statistically unsound.

Example: The Chapin Social Insight Test is a psychological test designed to measure how accurately the subject appraises other people. The scores

Figure 23: Power calculated as a function of sample size.

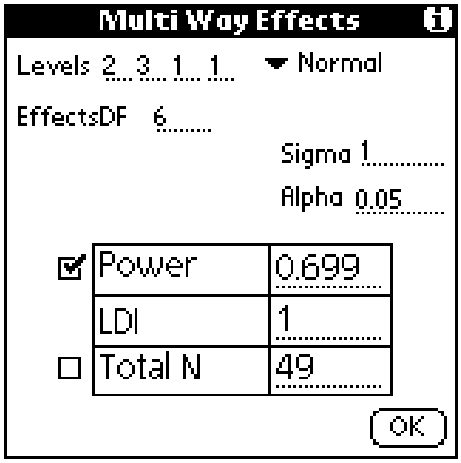
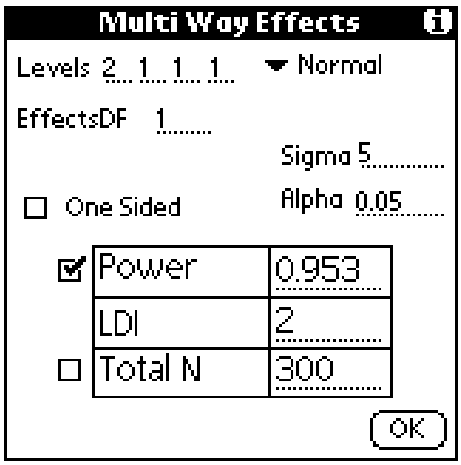


Figure 24: Power for Chapin Social Insight Test



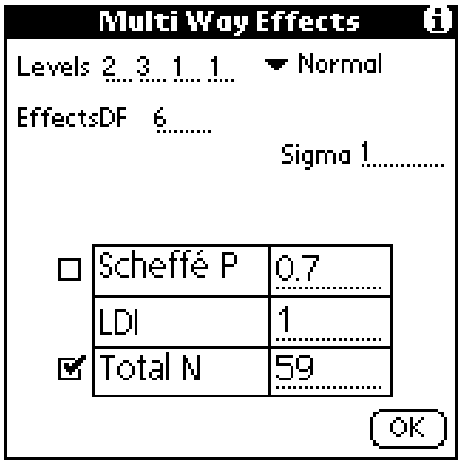
range from 0 to 41. A sample of about 300 students evenly divided between males and females was available, and it was proposed to administer this test as part of a battery. The sample standard deviations usually run about 5. Figure (24) shows that for an LDI of 2, the power is about 0.95. It is of course possible that the test measures noting at all, but 2 seems a small number in view of the range of scores. The experiment was run, and the difference between the averages of the scores for the two sexes was 1 which was not statistically significant. The resulting claim of no difference between the sexes was supported by the fact that the confidence interval on the difference in averages ran from -0.25 to 2.25. Clearly if there are differences, they must not be greater than 2.25 or so.

Alert: It should be noted that the power will sometimes change substantially. This is because power is usually calculated precisely, while by default, sample size is calculated approximately. The exact calculation button should be cliked before using power checkbox in order to avoid this behavior.

5.7 Scheffé's kernel

Selecting this kernel changes the sample size calculation from a power calculation to one in which the widths of confidence intervals are set equal to the LDI. Figure (25) shows the calculation corresponding to Figure (19), which indicates that 59 observations are required instead of the 49 given by power. This is a negligible change.

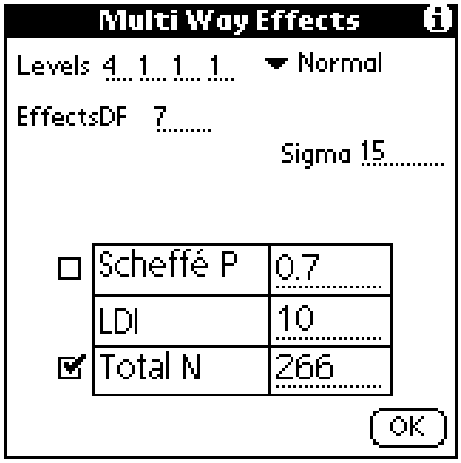
Figure 25: Sample size using the Scheffé kernel.



The Scheffé kernel controls for all possible contrasts even though the sample size is for the particular contrast specified. This means that as the number of cells under consideration grows, so will the sample size. The interested reader

may like to refer to Scheffé (1959) for more details about the S method of multiple comparison which is used.

Figure 26: Cardiovascular study using the Scheffé kernel.



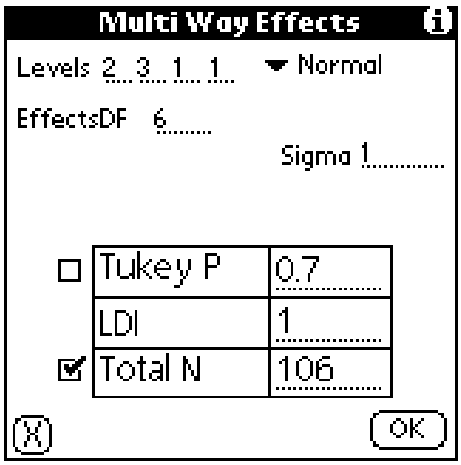
Example: As part of a study of cardiovascular function, individuals of both sexes from four occupation groups are to be studied. The design is simply a replicated 2x4 with 7 degrees for all effects. Heart rates after 6 minutes of a specified exercise are to be measured. The average heart rates of males may be taken to be about 130 and that of females about 150. Various sets of data indicate that the standard deviation for individuals is in the 12 to 17 range. The goal is to detect differences larger than 10 between the occupations if it exists.

A power calculation with $\alpha = 0.01$, $power = 0.70$ and $\sigma = 15$ gives a sample size of 240, or about 60 individuals per group. If instead $\alpha = 0.05$ had been assumed, then the sample size would have been only about 160, or 40 per group. Figure (26) shows the calculation using the Scheffé kernel which indicates a need for some 266 subjects, or about 65 per group. These differences are not of great practical importance, because of the uncertainty about σ, and the prudent experimenter will choose something in the neighborhood of 60 per group.

5.8 Tukey's kernel

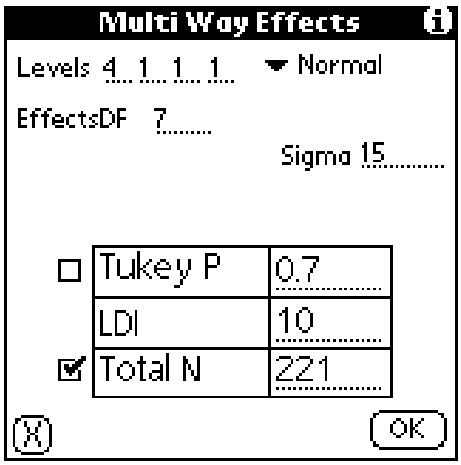
This kernel is most appropriate for two-level contrasts. It differs from Scheffé's kernel in that it does not control for all possible contrasts, but for a subset of them. Figure (27), shows the calculation corresponding to Figure (19), which in this case gives a sample size of 106. Clicking the exact button does not change the result. This sample size is larger than that given for the Scheffé kernel, but

Figure 27: Sample size using the Tukey kernel.



in general, the Tukey sample sizes will be the smaller. The interested reader my refer to Scheffé for details about the T method which is used.

Figure 28: Cardiovascular study using the Tukey kernel.



Example: Figure (28) shows the calculation for the cardiovascular example from Section (5.7). It may be seen that in this case, the use of the Tukey kernel produces a sample size of 221 which is smaller than that from either the Scheffé kernel or the power kernel.

5.9 Binomial, Poisson, and χ^2 Variance

An important characteristic of the normal distribution is that it is unbounded in both directions. Non-normal distributions are bounded on one or two sides. This causes a compression of values near the boundary. With normality the LDI remains constant.

For non-normal distributions the LDI changes as it nears a boundary. For example, the proportions of a binomial variable are bounded between zero and one. Thus the difference between 1% and 2% is often meaningful and important, while that between 50% and 51% may be negligible. For example, it might be a cause to celebrate if a defect rate could be reduced from 2% to 1%.

The Poisson and χ^2 distributions are bounded by zero, since all values must be positive, and as values approach zero they bunch up, and the LDI must change to accommodate this. For them, as for the binomial, the LDI changes across the scale.

This dependence on scale makes things difficult, and a common approach is to transform the response to an unbounded scale, and then to use existing theory on the transformed values. Such transformations frequently have the effect of making the resulting data appear to be normally distributed with constant variance, and such transformations are sometimes referred to as "normalizing transformations," or "variance stabilizing" transformations. I use such transformations in this program.

The rub is of course, that the transformed data isn't really normal, and any calculations made on that assumption must be considered approximate. The proper way to deal with the problem is by use of a generalized linear model, winch not only transforms but adjusts the calculations to the distribution. See McCullagh and Nelder (1989) for details. Unfortunately, the calculational complexities of this method make sample size calculations for complex linear models rather difficult since, among many difficulties, the coefficients are obtained by an iterative calculation.

Asymptotic[13]results can be used to partially overcome these difficulties, but they apply only to large samples.

The most practical way seems to be to use a variance stabilizing transformation which I have done. A simulation study with respect to this would be useful. In addition, for those common situations where the comparison is between two groups, I use accurate calculations from the literature.

In spite of the use of a transformation, sample size calculations for non-normal distributions are more reliable than those made assuming normality, because for them there is no need to guess at σ. Whereas the normal results are only "in the ballpark" the non-normal results are "on the money."

[13]Asymptotic calculations have been made for the binomial and Poisson cases. Whittemore (1981) treated the binomial and Signorini (1991) the Poisson. Both assume the independent variables to be random, which presents a problem to those who would use their results in a fixed effects case which is what almost all standard analysis methods assume.

5.9.1 Binomial and logistic

5.9.1.1 Binomial Since the binomial scale is not uniform, two proportions are required to locate the LDI. These proportions $P1$ and $P2$ are transformed and their transformed values are used to set the LDI on the transformed scale, which is hopefully a uniform scale. The transform used is the arcsine transformation: that is a proportion p is transformed to $(2 \arcsin \sqrt{p})$. The result is a variable with constant variance and an approximately normal distribution.

Figure 29: Binomial (logistic) calculation.

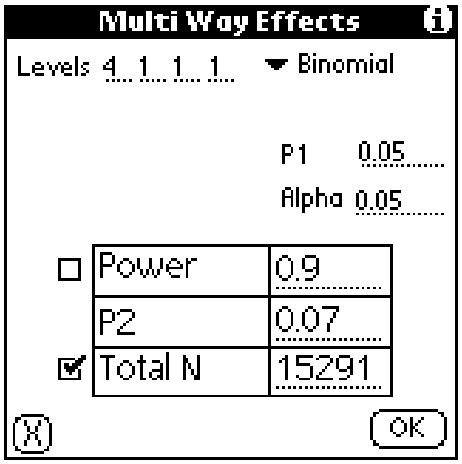


Figure (29) shows fields for two values, P1, and P2. In this figure P1 is 0.05, and P2 is 0.07, which is about 40% larger. The actual calculation is done on the transformed scale using an LDI that corresponds to the difference between the P's on the proportion scale. The result in this case is enormous, 15291 observations are required to detect a difference between two levels. A change of 40% between values near the center of the scale would result in a much smaller sample size. For example if P1 were 0.50, and P2 were 0.70, then the sample size would be only 645.

One may switch to the binomial to perform any of the available sample size calculations, and may analyze the results using ANOVA, regression, or any other appropriate technique, after first transforming the observations to a normal scale. One may also use a generalized linear model procedure such as logistic regression – this procedure is available in ECHIP.

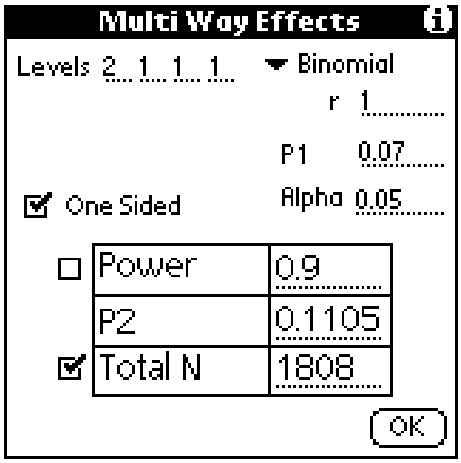
5.9.1.2 Logistic There are a number of research areas in which linear models are appropriate with binomial response variables. Cohort studies, for example, in which subjects are followed over time and a binary response, such as death, are modeled as functions of several variables, or planned experiments in which subjects are subjected to combinations of treatments. For most of these

the preferred transformation during analysis is the logistic transformation, and the analysis is referred to as logistic regression or ANOVA. The transformation is $logit(p) = \log(p/(1-p))$ and where p is a proportion. The logits are then modeled with a linear model in the independent variables of the problem, which might look like $logit(p) = \mu_0 + \mu_1 x_1 + \mu_2 x_2 \ldots$ The μ_0 coefficient represents the model intercept, and if the variables are properly coded, the center of the data. See Hosmer and Lemeshow (1989) for details about logistic regression. The LDI's on the logit scale are those appropriate for values near μ_0.

The logit transformation produces unbounded values, but the tails of the distribution are not a good match for the normal; hence I use the arcsine transformation rather than the logit to calculate the sample sizes.

Whittemore (1981) gives sample size calculations for logistic regression using asymptotic approximations which in general apply only for small probabilities. She also assumes the independent variables to be random, which requires the researcher to be specific about their distribution. Her results depend very heavily on this specification.

Figure 30: Fisher's exact calculation.

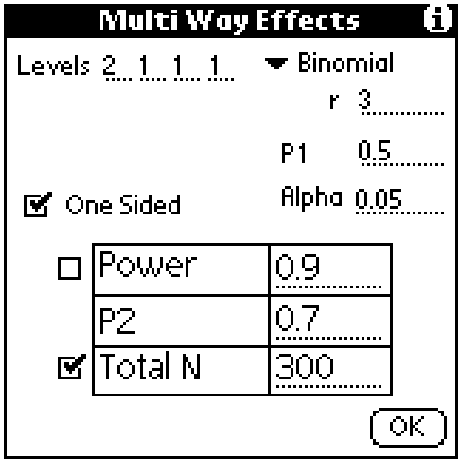


5.9.1.3 Calculations for two samples In many situations, there is only one variable under investigation, and that one has only two levels, which is to say that is there are two proportions to be compared. The problem has a considerable literature since it is very common and very important. It may be modeled as a 2x2 contingency table, and tested using the χ^2 approximation, but an exact test is available, called Fisher's exact test, Fisher (1935), and the literature contains methods for calculating exact sample sizes. It seems reasonable to suppose that users are likely to be interested in the exact calculation, so when there is a single two level factor, a sample size is calculated for Fisher's exact test. The procedure implemented here is due to Fleiss (1980), which is an approximation,

but a very good one. Casagrande and Pike (1978) give some exact values that may be compared with the program's calculations. For this calculation, one may choose either a one or two sided test by checking the checkbox.

It is interesting to compare this exact result with that of Whittemore. If one chooses a small proportion, the sample size will be large enough to justify her calculation. She gives an example involving the effect of serum cholesterol levels on coronary heart disease, CHD, in which the base probability that an individual will develop CHD during a period is 0.07, and indicates that the total sample size should be 582 to detect a change by an odds ratio of 1.65 when the test is a one sided test at 0.05 with 0.90 power. Such an odds ratio implies that the probability under the alternate hypothesis is 0.1105, and for this the sample size for Fisher's exact test is 1808, as shown in Figure (30). There is a considerable discrepancy between these results which is likely due to her assumption of a normally distributed independent variable. If one assumes a binomially distributed independent variable with binomial probability 0.5, then her results indicate a sample size of 2215.

Figure 31: Unequal sample sizes.



5.9.1.4 Unequal sample sizes Occasionally it makes sense to use unequal sample sizes. If for example, one treatment is more difficult or expensive than another, it might make sense to increase the precision of the test by taking a greater number of the least difficult treatment. Figure (31) shows a calculation for such a case, where the larger sample is to be three times the smaller. As may be seen, 3 has been inserted in the r field.

The result is a total sample size of 300, and $300/(r+1) = 75$ should be allocated to one treatment and 225 to the other.

5.9.2 LDI for proportions

Fleiss (1981) offers some guidance in choosing an LDI when comparing proportions, which I take as a model to illustrate the sort of thinking that may be used in making the choice. Fleiss discusses two situations: (1) comparison with a control, and (2) replication of a result.

5.9.2.1 Comparison with a control Suppose that P_1 is the rate of success for a standard treatment, and a new treatment is to be tested. What should be the goal for the new treatment's proportion, P_2? Surely the new treatment must do as well as the standard, plus it should impact the proportion of failures of the standard treatment in order to be considered an improvement. Suppose it is determined that it is economically or otherwise advantageous to succeed in a fraction F of these failures, then since $(1-P_1)$ is the failure rate of the current treatment, it follows that P_2 should be set to $P_1 + F(1-P_1)$, and the sample size calculated for these two proportions, P_1 and P_2.

Example: consider the situation where mothers who attend a hospital clinic experience $100P_1 = 25\%$ premature births. A visiting nurse program is proposed, but its expense is such that it can be justified only if it will make a 20% improvement. Thus $P_2 = P_1 + 0.20 \times (1-P_1) = 0.35$. The sample size for a one sided test at 5% with power of 0.80 is 557, which is likely too large to be practical. Changing from 20% to 40% brings the sample size down to 77, but it is unlikely that this great an improvement due to visiting nurses is possible. Clearly the project as proposed is not feasible, and something else should be attempted. Perhaps attention could be switched from a visiting nurse program to community education with the aim of attracting mothers to the clinic.

5.9.2.2 Replication of a result An odds ratio is the ratio of the odds for two treatments[14], and it may be useful to verify this ratio under different circumstances. The actual rates for the two treatments may be different, but it often happens that the relative odds (i.e. the odds ratio) remain the same. For example the chance that a pedestrian is hit by an auto may be higher in a city than in a village, but the relative odds between jaywalking and non-jaywalking situations may be the same. Thus a study done in one population may produce the same treatment vs non-treatment odds ratio as that done in a different population with different base rates.

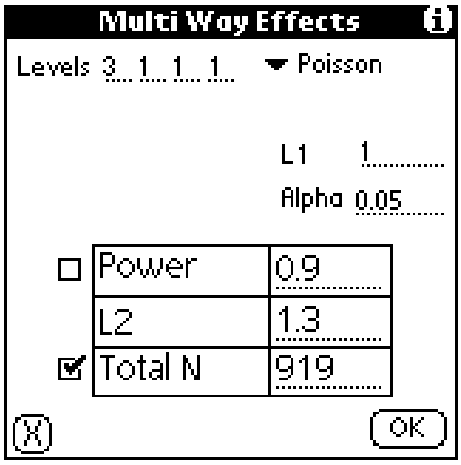
Odds are calculated by dividing a proportion by its compliment. Thus if something occurs 1/3 of the time the odds are $\frac{1/3}{2/3}$ or 1 to 2. It follows that if the ratio of two odds is R and if the control proportion in the new population is P_1 with odds $O = P_1/(1-P_1)$, then the treatment proportion P_2 in the new population may be found by solving $R = O/[P_2/(1-P_2)]$, which gives $P_2 = O/(R+O)$.

[14]The log of the odds ratio is the logit.

Example: Suppose that a survey of data in New England shows that the odds ratio between children who complete high school and those from one-parent families who complete high school to be $R = 2$. A researcher in the Midwest finds that the proportion of students who compel high school in the local community is $P_1 = 0.8$, and from this calculates the odds as $O = 4$ and then finds $P_2 = 4/(2 + 4) = 0.67$. The research may be replicated with a sample size in the neighborhood of 400.

5.9.3 Calculations assuming the Poisson distribution

Figure 32: Poisson for a linear model.



5.9.3.1 Poisson A Poisson process is one of the most fascinating processes. The Poisson distribution describes one part of it and represents the count of events per unit exposure, such as the number of radioactive counts per unit time or the number of particles in a unit volume, or, interestingly enough, the number of wars engaged in by a country in a given time, Richardson (1942). Such counts may occur at different rates, λ, and it is usually this parameter on which interest focuses. For example, "did Great Britain have a greater war rate in its heyday than the U.S. does now?"

Wars might be modeled in terms of various factors, say exchange rate, industrial capacity, length of orations in congress or parliament, etc. The model might look like $\lambda = t \exp(\mu_0 + \mu_1 x_1 + \ldots)$, where t is in units of exposure such as time, and the x's are independent variables.

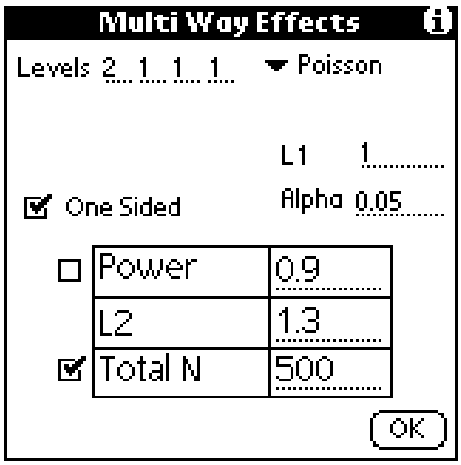
LDI's for the Poisson, like other non-normal distributions, vary across the scale, and require two values to define the LDI on the transformed scale. The square root transformation is used: that is the value x is transformed to $\sqrt{x}$. Figure (32) shows an example, where the LDI is defined by two Poisson λ's

with $L1 = 1$ and $L2 = 1.3$. Means may be input instead of λ's, since a Poisson mean is simply λt, where t is in units of exposure. In actuality only the ratio of the two values is of importance: the same sample size will be obtained for $L1 = 5$, $L1 = 6.5$ and for $L1 = 1$, $L1 = 1.3$. Note that there is no EffectsDF field available to specify the number of terms in the model[15].

One may switch to the Poisson to perform any of the available sample size calculations, and may analyze the results using ANOVA, regression, or any other appropriate technique, after first transforming the observations to a normal scale. One may also use a generalized linear model procedure with the log link which is appropriate for a Poisson – this procedure is available in ECHIP.

Signorini (1991) provides asymptotic results which apply only to large samples. He also assumes the independent variables to be random, which requires the researcher to be specific about their distribution. His results depend very heavily on this specification.

Figure 33: Poisson for two samples.



5.9.3.2 Calculations for two samples: In many situations, there is only one variable under investigation, and that one has only two levels, which is to say that is there are two λ's to be compared. The sample size may be found in the context of a linear model simply by setting up a single factor at two levels. Like the binomial, this problem has a considerable literature since it is very common and very important.

There are several possible tests, since a Poisson process has several facets. These are described by Birnbaum (1954), and several are implemented by Gail (1974). The one most consistent with the linear model above involves observing a Poisson process until a fixed count total is reached.

[15]The number of terms impacts a linear model primarily through the number of degrees of freedom available to estimate σ, but there is no σ to estimate for a non-normal model

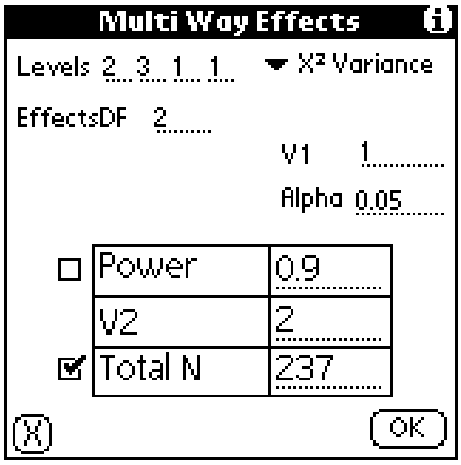
For example, if one desires to detect a ratio of λ's of 1.3, one might calculate as in Figure (33), where it is seen that a total of 500 counts are needed. The experiment would consist of observing two processes until a total of 500 counts were obtained. Nelson (1987) describes the appropriate analysis of the data.

The sample size is calculated using the approximate formula given by Gail (1974). It differs from the result of an exact calculation only due the rounding of the size and power probabilities, and thus for large ratios of λ's may differ by a few counts from the exact values.

Example: Signorini (1991) gives an example in which the ratio of the λ's is 1.3. For $\alpha = 0.05$ and $power = 0.90$, his calculation results in a sample size of 555 when he assumes that the independent variable is binomial distributed with binomial probability 0.5. This is in reasonable agreement with Figure (33).

5.9.4 Calculations for the χ^2 variance

Figure 34: Variance calculation.

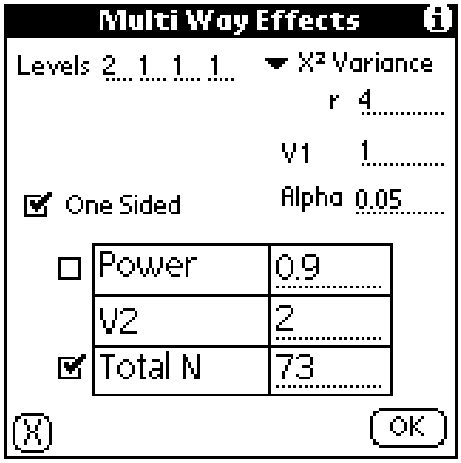


5.9.4.1 Variances Sample variances from normally distributed data are distributed like $\sigma^2\chi^2/df$ with df equal to one minus the number of observations on which the sample variance is calculated. Like other non-normal distributions the LDI varies across the scale, and two values are required to define the LDI on the transformed scale. The transformation used is due to Wilson and Hilferty (1931): for a value x, with ν degrees of freedom, the transformation is $\left(\sqrt{\frac{9\nu}{2}}\left[(\frac{x}{\nu})^{1/3} - 1 + \frac{2}{9\nu}\right]\right)$. Population variances should be input in the V1 and V2 fields to define the LDI.

Example: Suppose, for example, a manufacturer has two plants, and three production lines in each, and suppose that it is important to ensure that the variability of the product from both plants and all lines does not differ by more than 2 to 1. Figure (34) illustrates the calculation. It may be seen that estimates of variance based on about 40 df from each of the six lines will be able to detect a LDI corresponding to a 2 to 1 ratio. The estimates may be obtained in any way such that their total df adds up to 40 for each of the 6 lines: that is 10 estimates with 4 df each may be used or one with 40 df.

One may switch to the χ^2 variance to perform any of the available sample size calculations, and may analyze the results using ANOVA, regression, or any other appropriate technique, after first transforming the observations to a normal scale. One may also use a generalized linear model – this procedure is available in ECHIP.

Figure 35: Two-hour vs. once-a-day



5.9.4.2 Calculations for two samples As with the binomial and Poisson, the two sample situation is of most interest. In addition, such two sample situations often involve unequal sample sizes because the two samples frequently come from different strata. A split-plot experiment, for example, has an error associated with the variation among the whole plots, and one within the plots. Repeated measurements experiments are similar in that the variation from measurement to measurement within an individual is different than that between individuals. If the variances for different strata in such designs are to be compared, sampling will naturally produce more data from the second strata than the first.

Note: The calculation assumes that the larger sample size is always associated with the smaller variance.

Note: The power calculation is always exact. Hence the power may change when the power checkbox is clicked if the sample size was calculated approximately.

Example: As an example, suppose that the variance in systolic blood pressure measured at two hour intervals were be compared with the same measurement taken once-a-day. In this case there would be more two-hour measurements than once-a-day measurements. Suppose 5 two hour measurements are made producing 4 df each day. There will thus be a 4 times as many two-hour as once-a-day degrees of freedom, and the sample sizes will be in a ratio of 4 to 1. The value 4 should be input in the r field of the program.

It is usually argued in such situations that the once-a-day measurements should be more variable than the two-hour measurements, because they are subject to all the sources of variation of the two-hour measurements plus others that occur between days. Suppose that it is desired to detect a 2 to 1 ratio between the two variances.

Figure (35) shows an appropriate sample size calculation for detecting a 2 to 1 ratio of population variances. The approximate calculation indicates that some 73 df are required. An exact calculation shows 91 instead of 73. This is the total degrees of freedom required, and it should be split in a 4 to 1 ratio. The smaller sample size is associated with the once-a-day measurements and will require about 19 days (18 df). The total experiment will need $5 \times 19 = 95$ measurements over 19 days, with 5 two hour measurements per day.

5.10 χ^2 contingency tables and goodness of fit

This topic has an extensive literature, and about all that can be done in this brief writeup is to give a few equations for evaluating the parameter τ needed by the program.

A contingency table with r rows and c columns can arise in three experimental situations depending on whether the row or column sums are fixed or not – see Section (5.10.6). The analysis does not depend on these situations, since it is done conditionally on the row and column sums, but the power and sample size do depend on the situation. For the limiting χ^2 distribution, each situation has an appropriate formulation.

5.10.1 Comparative trials

5.10.1.1 2x2 tables The most common situation is called a comparative trial and has either the rows or the column sums fixed. The Fisher exact test for a 2x2 contingency table may be viewed as a comparative trial. In this there are two populations, and the null hypothesis is that both have a common occurrence probability, α. The alternate hypothesis is that the two populations each have a different occurrence probabilities, π_1 and π_2. A 2x2 table is constructed

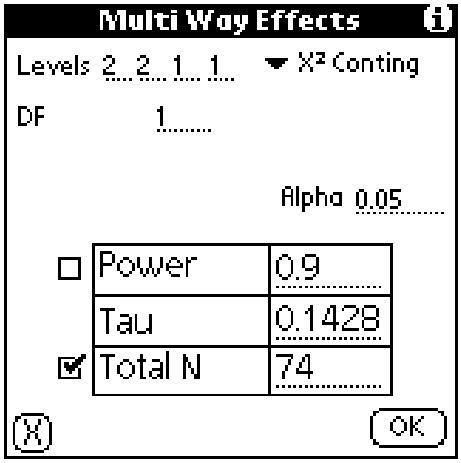
Table 3: Probabilities for a 2x2 table

Null			Alternative		
1/3	2/3	Q_1	1/2	1/2	Q_1
1/3	2/3	Q_2	1/7	6/7	Q_2

containing these probabilities. Assuming that the row sums are fixed, the null table is illustrated on the left of Table (3), where $\alpha = 1/3$. The alternate table is as on the right of Table (3) with $\pi_1 = 1/2$ and $\pi_2 = 1/7$. The proportions of the total sample that will be assigned to the two rows are indicated by Q_1 and Q_2 with of course $Q_1 + Q_2 = 1$.

Figure (36) illustrates a sample size calculation for this case which results in a total sample size of 73. The binomial calculation for Fisher's exact test produces a sample size of 79 for a two-sided test and thus the two calculations are in reasonable agreement. The total sample would of course be divided between the two rows according to Q_1 and Q_2. The degrees of freedom is 1.

Figure 36: Contingency Table calculation.



The calculation is made by inserting a value in the Tau field. The appropriate equation for a 2x2 table with alternate hypotheses probabilities π_1 and π_2 is:

$$\tau = Q_1 Q_2 \frac{(\pi_1 - \pi_2)^2}{\alpha(1-\alpha)}, \tag{1}$$

It should be noted that α in this case is not equal to $Q_1\pi_1 + Q_2\pi_2$, although there is nothing to prevent this choice if desired. Such a choice reduces the alternate specification to a single parameter, since $\pi_i = \alpha \pm \delta$.

For the illustration in Figure (36), the calculation is

$$\tau = \frac{1}{2}\frac{1}{2}\left[\frac{(1/2 - 1/7)^2}{1/3 \times 2/3}\right] = 0.1435,$$

with $Q_1 = Q_2 = 1/2$.

Sample size calculations for 2x2 tables are relatively straightforward, in that the alternates are fairly clear cut. Things become complicated when there are more rows or columns since the choice of alternate hypotheses is vast.

5.10.1.2 2xc tables For the general 2xc table, τ is given by

$$\tau = Q_1 Q_2 \left[\sum_j (\pi_{1j} - \pi_{2j})^2/\alpha_j\right], \tag{2}$$

where the π_{1j} and π_{2j} are the parameters for the jth column and $\alpha_j = Q_1\pi_{1j} + Q_2\pi_{2j}$. It is important to note that the parameters in each row sum to unity; that is, it is assumed that $\sum_j \pi_{ij} = 1$. The degrees of freedom is (c-1).

Example: Consider the 2x3 table given by Cohen (1988) on page 219. He gave the alternative population probabilities as shown on the left of Table (4) as an independence trail, where all the probabilities sum to unity, and the null probabilities are completely specified for each entry as the product of the corresponding marginals: e.g. the null for cell (1,1) is $0.6 \times 0.45 = 0.27$.

In the comparative trial on the right of this table, where the rows have been rescaled to sum to unity, it is only assumed that the column probabilities have some common value to be estimated from the data using the row proportions 0.60 and 0.40. The value of τ is 0.1197 from Equation (2), and the total sample size is 106. It is worth noting that Equation (2) is algebraically identical to the calculation used by Cohen.

Table 4: Two parameter forms

	Independence trail					Comparative trial			
	Dem.	Rep.	Ind.	marginal		Dem.	Rep.	Ind.	sum
Men	.22	.35	.03	.60		.367	.583	.05	1.00
Women	.23	.10	.07	.40		.575	.25	.175	1.00
marginal	.45	.45	.10	1.00	α	.45	.45	.10	1.00

5.10.1.3 rxc tables A general expression for rxc tables may be obtained by representing the π_{ij} in terms of deviations from the null hypothesis column[16] parameters α_j. Define $\pi_{ij} = \alpha_j + \delta_{ij}$, then

[16]Columns and rows may be interchanged if desired.

$$\tau = \sum_j \frac{1}{\alpha_j} \left[\sum_i Q_i \delta_{ij}^2 - \left(\sum_i Q_i \delta_{ij}) \right)^2 \right], \tag{3}$$

where the degrees of freedom is (r-1)(c-1).

Since the π_{ij} in each row sum to unity, the δ_{ij} in each row must sum zero, but the δij need not sum to zero in columns; however, if they do, then the second term in equation (3) vanishes which considerably simplifies the expression:

$$\tau = \sum_j \frac{1}{\alpha_j} \left[\sum_i Q_i \delta_{ij}^2 \right]. \tag{4}$$

This last expression is very convenient, and should serve most needs. I find it useful to choose the $\{\delta_{ij}\}$ as elements of a suitably scaled contrast, which brings the specification into line with contrasts for ANOVA as discussed in previous sections.

The theoretical basis for these equations is due to Mitra (1958). Other papers of interest are Meng (1966), and Lachin (1977). The equations above are in the form given by Lachin.

Example: Example 8.14 in Moore and McCabe (1993) discusses a 3x3 table relating smoking habits to socioeconomic status, SES. The rows classify individuals according to their smoking experience as "current smokers," "former smokers," and "never smoked." The hypothesis of interest was whether or not the proportions falling into these classes differed according to SES. There were about twice as many "High" SES individuals available as there were for either "Middle" or "Low" SES. The null hypothesis was that the proportions for the three smoking classifications would be the same. One possible form for the alternate hypotheses $\{\delta_{ij}\}$ is shown in Table (5).

Table 5: Alternate hypotheses δ_{ij} for smoking habits study

	Current	Former	Never	Q_i
High	-0.2	0.2	0	1/2
Middle	-0.1	0.1	0	1/4
Low	0	0	0	1/4
α	1/3	1/3	1/3	

Using equation (3) gives $\tau = 0.04125$, and from this a sample size of 374 for $\alpha = 0.05$ and $power = 0.9$. The actual study used 356.

5.10.2 Double dichotomy

A second experimental situation of interest is the double dichotomy, where none of the marginal totals is fixed. For the comparative trial $\{\alpha_i\}$ represented the

common column null hypothesis probability values. In a double dichotomy, one also supposes that the rows have common probability values $\{\beta_i\}$ under the null hypothesis. Defining deviations in terms of these by $\pi_{ij} = \alpha_i\beta_j + \delta_{ij}$ gives the following expression for τ:

$$\tau = \sum_{ij} \frac{\delta_{ij}^2}{\alpha_i\beta_j} - \sum_i \frac{\delta_{i.}^2}{\alpha_i} - \sum_j \frac{\delta_{.j}^2}{\beta_j}, \tag{5}$$

where $\sum_i \delta_{ij} = \delta_{.j}$ $\sum_j \delta_{ij} = \delta_{i.}$and it is only assumed that $\sum_{ij} \delta_{ij} = 0$.

Example: Suppose that the probabilities are as on the left of Table (4), and the row α's as before, but that $\beta_1 = \beta_2 = 0.50$, then it will be found that $\tau = 0.1551$ instead of 0.1197, and 82 trials will suffice instead of the 106 cited previously.

5.10.3 McNemar's test

Table 6: 2x2 Table for McNemar's test

	Observations			Probabilities	
	0	1	total		
0	n_{00}	n_{01}	$n_{0.}$	π_{00}^1	π_{01}^1
1	n_{10}	n_{11}	$n_{1.}$	π_{10}^1	π_{11}^1
total	$n_{.0}$	$n_{.1}$	N		

Mitra (1958) has shown how to calculate the power of McNemar's test, which is a binary crossover design. The formulation that maximizes the power assumes that N indivuals are chosen and subjected to two treatments, one after the other. If these are classified in a 2x2 table such as Table (6), then

$$\tau = \frac{(\pi_{10}^1 - \pi_{01}^1)^2}{2\pi^0}, \tag{6}$$

where $\sum \pi_{ij}^0 = 1$, and the null hypothesis is $\pi_{01}^0 = \pi_{10}^0 = \pi^0$, with the superscript 0 denoting the null and the superscript 1 the alternative. Note π^0 is not necessesarily equal the average of π_{01}^1 and π_{10}^1

5.10.4 Multiway tables

There is little theory for tables more complex than rxc; however, since the limiting distribution is χ^2, I would be surprised if Rule 1 from Scheffé (1959) would not apply. This rule states that the noncentrality parameter may be obtained by replacing each observation by its expectation. Hence the following should apply

$$\tau = \sum_{\omega} \frac{(\pi_{\omega}^1 - \pi_{\omega}^0)^2}{\pi_{\omega}^0}, \tag{7}$$

where ω is a set of indices, and the exponents 0 and 1 denote the null and alternate hypotheses.

5.10.5 Goodness of fit

Given a set of counts, $x_i, i = 1, \ldots, N$, and a corresponding set of probabilities, $\pi_i^0, i = 1, \ldots, N$, the statistic

$$X^2 = N \sum_{i} \frac{(x_i/N - \pi_i^0)^2}{\pi_i^0},$$

is distributed under certain conditions as a χ^2 variate. When these conditions are met, the limiting sample size may be calculated from a noncentral χ^2 with noncentrality parameter $N\tau$, where

$$\tau = \sum_{i} \frac{(\pi_i^1 - \pi_i^0)^2}{\pi_i^0}, \tag{8}$$

with the $\{\pi_i^1\}$ being a set of alternative probabilities. The conditions and results may be found in many places, Kendall (1967) chapter 30 for example. The probabilities $\{\pi_i^0\}$ may be estimated, and the appropriate degrees of freedom reduced by the number of estimates.

It should be obvious that Equation (8) is the same as Equation (4) used with a single multilevel factor.

The most common use is when the $\{x_i\}$ represent counts falling into cells specified by the values of a probability distribution with the probabilities representing the portions of the distribution assigned to the cells. For example, when the distribution is the normal, and its mean and standard deviation are estimated from the data, then the appropriate degrees of freedom is $N - 2$.

5.10.6 Further comments

It is supposed that N observations $\{x_{ij}\}$ following a multinomial distribution with probabilities $\{\pi_{ij}\}$ are arranged in a table with r rows and c columns. The statistic

$$X^2 = \sum_{ij} \frac{(x_{ij} - N\pi_{ij})^2}{N\pi_{ij}} \tag{9}$$

is calculated and used as a test statistic. As $N \to \infty$ this statistic is distributed like χ^2 with (r-1)(c-1) degrees of freedom. If there are s restrictions on the probabilities, the degrees of freedom is reduced by s.

There are three distinct experimental situations which give rise to such an rxc table: (1) the double dichotomy, DD, in which both the row and column

sums are free; (2) the homogeneity or comparative trial, CT, in which either the row or the column sums are fixed; and (3) the independence trial, IT, in which both the row and column sums are fixed. This classification describes the way in which the data occurs, not the analysis. The analysis is usually done conditionally on fixed marginals, and the limiting distribution of X^2 under the null is χ^2.

Situations in which an IT might actually arise are rare, and after the fact, it is not easy to identify the experimental situation even when attention is restricted to DD and CT. Kendall (1967) in example 33.4 illustrates the problem by referring to a table given in example 33.1, reproduced here as Table (7).

Table 7: Effect of cholera inoculation

	Not-attacked	Attacked	Totals
Inoculated	276	3	279
Not-inoculated	473	66	539
Totals	749	69	818

The comments about this are:

> "The table in Example 33.1 above is certainly not of our last type, with both sets of marginal frequencies fixed, but it is not clear, without further information, which of the other types it belongs to. Possibly 818 persons were examined and then classified into the 2x2 table. Alternatively, two samples of 279 inoculated and 539 not-inoculated persons were separately examined and each classified into "attacked" and "not-attacked." It is also possible that two samples of 69 attacked and 749 not-attacked persons were classified into "inoculated" and "not-inoculated." There are thus three ways in which the table might have been formed, one of the double-dichotomy type and two of the homogeneity type. Reference to the actual process by which the observations were collected would be necessary to resolve the choice."

Although the null distribution of X^2 remains the same for these three experimental situations, the power curves for exact calculation do not because each of the three gives rise to a different class of alternative hypotheses with a different τ. See Harkness (1964) for the 2x2 exact case. For the limiting χ^2, the expectations are different, but sample size may be calculated following Mitra (1958).

6 The General Calculation Module

To avoid needless repetition, I will assume that the Multiway Effects section has been read. I will also assume a greater sophistication on the part of the reader.

Figure 37: General Calculation Startup Screen.

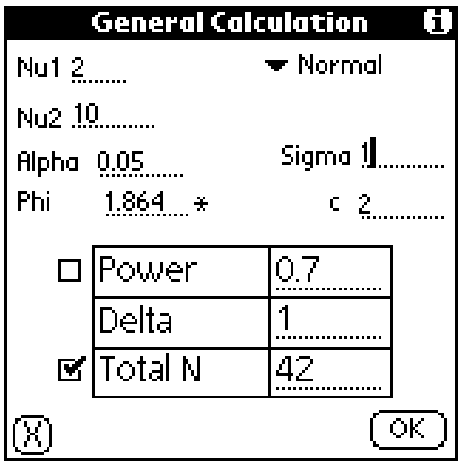


Figure (37) shows the start up screen for this module. It differs in several ways from the Multiway Effects start up screen, most notably in the presence of the Nu1, Nu2 and Phi fields. The first two fields refer to the numerator and denominator df (ν_1 and ν_2) of the F-distribution, and the last to the usually tabled function of the noncentrality parameter, λ^2 of the noncentral F-distribution: i.e. $\lambda^2 = \phi^2(\nu_1 + 1)$.

In the Multiway Effects Module, ν_1 was calculated from the levels fields, and ν_2 by iteration from $N - df$, where N is the total sample size and df the model degrees of freedom (EffectsDF + 1).

The Power, Scheff, and Tukey kernels are available from the menu, as they were in the Multiway Effects Module; however, there are no menu item for Cohen's f nor for special contrasts. These are both available by direct input as will be described. The same distributions are present from the drop down list, except for χ^2 contingency tables, which are completely treated in the Multiway Effects Module.

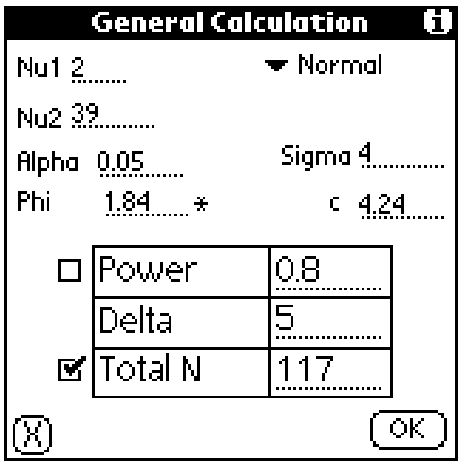
To adjust for one sided tests, insert 2α in the Alpha field.

6.1 General calculation

The calculation is controlled by the value in the c field. The total sample size N is given by $N = ((c\lambda\sigma)/\Delta)^2$, where Δ is the parametric function value to be detected. If for example, Δ is the standard deviation of a set of parameters (Cohan's f), then $c = 1$ and $\sigma = 1$. If Δ is the detectable value for a linear functional ψ then $(c\sigma)^2/N$ is the variance of $\hat{\psi}$. For example, suppose $\psi = w\mu$ is a linear functional of the π dimensional parameter vector μ, then $c^2 = w'w\pi$.

Example 1: Consider a $3x4^2$ factorial experiment where there are three fertilizer compounds to be tested on in a 4x4 plot area, and suppose one wants

Figure 38: Example1.



to detect a difference in yield of 5 bushels between the middle fertilizer and the others. The contrast of interest is thus w=[-1,2,-1], from which one has $c = \sqrt{6 \cdot 3} = \sqrt{18} \approx 4.24$. If interactions are not thought important, then the model has 8 df, and thus $\nu_1 = 2$ and ν_2=39, which gives the calculation shown in Figure (38) when σ is guessed to be 4 bushels.

This result of 117 trials differs from that which would have been obtained from the Multiway Effects calculation because of the definition of Δ. In general Δ is equal to the LDI only when the contrast is a difference of two values, such as [-1,0,1]. For this contrast, $c = 2.45$ which produces $N = 39$ in agreement with the Multiway Effects calculation. Appendix A explains the translation from LDI to Δ. Some care is needed in setting up a calculation. If for example, the contrast w=[-1/2,1,-1/2] had been chosen the sample size would have been 29; yet both contrast the middle fertilizer with the others. Whereas the Multiway Effects Module scales everything to a common LDI, the General Calculation leaves this up to the user; and in this case, changing the contrast changed the meaning of Δ by a factor of two!

Example 2: Consider the same example, but using Cohen's f with a "medium" value of 0.25. Figure (39) shows the calculation. The result is 162 trials, but frankly, I do not know that this means, nor what is really being calculated here in terms of the problem.

6.2 The Phi field

For the power kernel, this field shows ϕ, where $\lambda^2 = \phi^2(\nu_1 + 1)$ is the noncentrality parameter of the F-distribution. The asterisk that appears beside the field

Figure 39: Example 2 using Cohen's f.

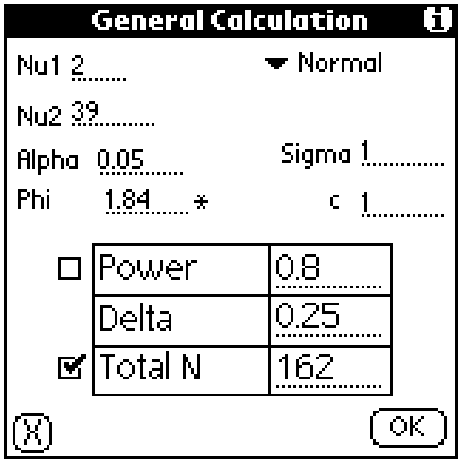
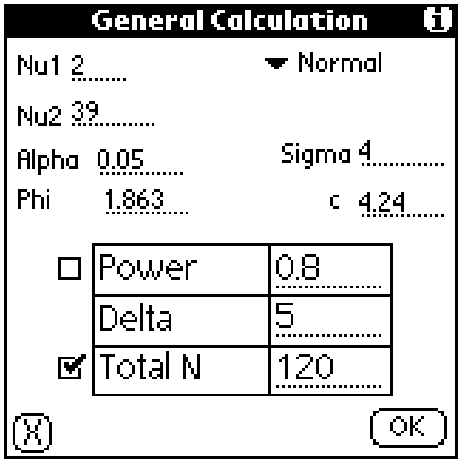


Figure 40: Exact calculation



indicates that the value is an approximation. The approximation is generally good, and is used in order to make it possible to change on-screen values easily; however, an exact calculation may be obtained by clicking the *Exact* button in the lower left hand corner of the screen. Figure (40) shows the result. A fairly difficult numerical series is evaluated to obtain the exact result, which may on occasion converge slowly. When this happens, the asterisk will reappear after the calculation. There will be no practical effect on the resulting sample size when this happens. Checks on the accuracy of the calculation may be made by comparing the results with graphs in Odeh and Fox (1975).

For non-normal distributions, $\lambda^2 = \phi^2(\nu_1 + 1)$ is the noncentrality parameter of the noncentral χ^2 distribution, and its accuracy may be checked by comparing the results with the tables in Pearson and Hartley (1972).

For the Scheffé and Tukey kernels, ϕ is defined as a function of a percentage point of a distribution. See Appendix C for details.

6.3 Non-normal distributions

The calculations remain the same as for the normal, except that Δ is now defined by two values. These are transformed by a normalizing transformation and the difference of the transformed values used to define Δ on the transformed scale. The meaning of the c field is unchanged.

The transformations used are:

1. **binomial:** $(2 \arcsin \sqrt{p})$, for proportion p.
2. **Poisson:** $\sqrt{x}$, for value x.
3. χ^2 **variance:** $\left(\sqrt{\frac{9\nu}{2}}\left[(\frac{x}{\nu})^{1/3} - 1 + \frac{2}{9\nu}\right]\right)$, for a value x with ν the degrees of freedom: due to Wilson and Hilferty (1931).

A LDI for arbitrary linear functions

Let t be a vector of parameters, and c't a linear function with vector c, then c(c't/c'c) is that part of t ascribed to the function. If ρ is the range of the elements of c, then $\rho(|c't|/c'c)$ is the range of the contribution of t made by the function, and if Δ is the detectable value for the function, then LDI is given by $\rho\Delta/c'c$. For c'=[1,-2,1], one has ρ=3, and c'c=6, hence LDI=$\Delta/2$. For c'=[1/3,1/2,-5/6], one has ρ=8/6, and c'c=19/18, hence LDI=$24\Delta/19$.

B Post Hoc, observed power, and other misuses.

Such procedures attempt to assess the quality of a study *a posteriori*[17]. As an academic exercise, this may be of interest, but there seems to be considerable

[17]Which is, I would have thought better Latin than *post hoc*

confusion about it, with some attempting to use it to weight the evidence about the null hypothesis. There is no logical basis for this.

There is nothing to prevent running the calculations in reverse – that is by starting with the sample size and solving for power. It can be interesting, and the ECHIP power calculator makes it possible. However, it should not be substituted for an examination of the resolution bounds involved. The statistical difficulty with reversing the calculations is that one cannot assess the statistical errors involved, nor indeed easily state what is being estimated, if anything is. Contrast this with a resolution bound which is a clearly defined entity from a statistical and probabilistic viewpoint. It is a confidence limit on an estimate.

The *post hoc* procedure has been used in survey papers such as Cohen(1962) where it can provide an assessment of the adequacy of sample sizes as used in a particular area of study. This can be legitimate.

There are two large problems with *a posteriori* power procedures. First of all, it is an improper calculation of a probability. Power is of course, the probability that the alternate hypothesis will be correctly judged to occur when it is true. The value of this probability changes upon collection of the data, and the probabilities before and after the data are not necessarily equal. Zumbo and Bruno (1998) illustrate this for a simple case in which the probabilities before and after the data are 0.483 and 0.935 respectively. If, as most *a posteriori* users do, the usual power formulas had been used after taking the data and, upon finding a significant result, judging the alternate to be true, the user would mistakenly have assessed the probability of the alternate hypothesis to be 0.483, rather than the correct value of 0.935.

Secondly, it can not be used to add something to the interpretation of the results, such as an assessment of the likelihood of the null. This is in fact what is attempted by those who calculate *observed power*, where the observed values are fed into a reversed power calculation, and the size of the resulting power is used as evidence of the adequacy of the study. In point of fact, the *observed power* is completely determined by the observed "p value:" Hoenig and Heisey (2001). It follows that those who perform such *observed power* calculations will inevitably find that the power is low.

It is true that some scales in the social sciences are difficult, but it is on these scales that results must be judged. I suspect the difficulty with these scales is confounded by apparently scale free quality of Cohen's f and by the use of Δ/σ without attempting to separate the two parts.

C Alternates to Power for sample size selection.

I assume a statistical background in this section.

The idea of the power of a statistical test was developed by Neyman-Pearson (1933), and is the usual approach to the selection of sample size. The theory postulates two hypotheses and treats errors that may be made with respect to both. The base or null hypothesis is a straw man set up in the hope of rejection. The alternative is concluded when the null is rejected. This bifurcation imposes

a clear direction to the problem, and requires some precision in the specification of the alternative, which is sometimes a difficulty. For example, the null hypothesis that the next time my office door opens, a man will enter, suggests as an alternate, the hypothesis that a woman will enter; however, it may happen that the wind opens the door, or my dog comes in.

Fisher(1959) and others have voiced objections to such formal specifications, but seem to have offered little guidance with respect to choosing sample size. Fisher did not feel it necessary to consider alternative hypotheses in a statistical test, and argued for the scientific merits of simply rejecting hypotheses when not sufficiently supported by evidence. His influence, and the difficulty often found in specifying alternate hypotheses has led may researchers to concentrate on significance levels (p values) as the single figure of scientific merit, which has the unfortunate consequence that some researchers accept the failure to reject as proof of the null hypothesis, when it is really a Scotch verdict. Cohen(1988) rightly criticizes this attitude.

Cohen therefore adopts the Neyman-Pearson rubric, and places power in the forefront. It is not completely clear that this is the best approach, since it forces on all experimenters the need to be precise about alternate hypothesis, and even to manufacture them in order to use the theory. Consider, for example a calibration problem, or one in which the goal is to establish the reasonable equivalence of two products or treatments, such as a generic drug. In such cases, the alternative is the straw man, and the goal is in fact to try to prove the null hypothesis! Much legitimate research is directed toward establishing the reasonable validity for a null hypothesis, and this may be part of the difficulty in the studies criticized in surveys for lack of power.

In an attempt to accommodate this idea, I offer an alternative calculation based on choosing sample sizes to control resolution bounds (i.e. the size of confidence limits). This forces the focus away from a concentration on alternative hypotheses and the balance between test size and power, to that of precision in the measurement. In fact, the calculation does not depend on the test size, although the equivalent of power is present. The fact that the test size does not enter into the calculation does not mean that properties of the test are unimportant, but only that they are not paramount, and that the precision of the test must go along for the ride, so to speak, instead of doing the driving.

Jason Hsu (1966) calculates sample sizes which control both the interval and the probability that it covers the parameter. There is merit in this, as Westlake (1979) indicates. However, because of the calculational burden that would be imposed on the Palm device due to the extra integration, I chose to control only the width.

With this in mind, the calculational formulas are the same as those for the power calculation, except that the noncentrality parameter is replaced by a percentage point from a distribution, which controls the width of the confidence intervals that will be calculated a posteriori. I will now briefly outline the details.

Formula (3.1) in Wheeler(1974) gives the basic formula for power calculation. Consider a linear functional $\psi = c'\mu$, where μ is a vector of estimable

parameters. Let $\hat{\psi}$ be the least squares estimate with variance $\sigma^2_{\hat{\psi}} = c'c\pi\sigma^2/N$ with π and σ being scalars and N the sample size. Formula (3.1) is $\lambda = \Delta/\sigma_{\hat{\psi}}$ where λ^2 is the noncentrality parameter of the noncentral F distribution with ν_1 and ν_2 degrees of freedom, and Δ is the detectable value for ψ: that is the least absolute value of ψ that will be significant with at least the power specified. The formula may be rewritten as

$$N = \pi\lambda^2 c'c(\sigma/\Delta)^2. \tag{10}$$

To relate Δ to the LDI, it is useful to define an *effect* as the contribution a functional makes to the response. If one considers the vector of parameters μ to be in a space spanned by a set of orthogonal vectors, the first of which is c, then with ρ the range of the elements of c, $\rho|\psi|/c'c$ represents the magnitude of the change in μ due to ψ. In general μ will be a subset of the observations such as the cells in an interaction, hence an *effect* e_ψ will be defined as $\rho\psi/c'c$. The quantity $\gamma = \rho\Delta/c'c$ will thus be a parametric value corresponding to the LDI.

Substituting $c'c\gamma/\rho$ for Δ and taking $\theta = c'c/\rho^2$ gives the following version of equation (10) as the sample size formula based on power:

$$N = \pi\lambda^2\sigma^2/\gamma^2\theta. \tag{11}$$

For a representation involving a confidence interval based on the F distribution, let $\hat{e_\psi} = \rho\hat{\psi}/c'c$ be an estimated effect with variance $v(\hat{e}) = \pi\sigma^2/\theta N$. The width of a $1-\beta$ size S-method confidence interval for $\hat{e}$ is $2S\sqrt{v(\hat{e})}$, where $S^2 = (\nu_1 F_{\beta/2;\nu_1,\nu_2})$ – see Scheffé(1959). Equating γ to the width of the confidence interval, gives $\gamma^2 = 4S^2\pi\sigma^2/\theta N$, or

$$N = \pi 4S^2\sigma^2/\gamma^2\theta, \tag{12}$$

as the formula based on a confidence interval. Thus the noncentrality parameter λ^2 is replaced by $4S^2$.

One may do exactly the same sort of thing for any method of constructing confidence intervals. One very attractive method is Tukey's method, described in Scheffé (1959). For this one has

$$N = \pi q^2\xi^2\sigma^2/\gamma^2\theta, \tag{13}$$

where $q \equiv q_{1-\beta,\nu_1,\nu_2}$ is the $100(1-\beta)$ percentage point of the Studentized range and $\xi = \sum|c_i|/\sqrt{\rho}$.

Only the central part of these equations changes, and since it is common to table ϕ rather than λ^2, where $\lambda^2 = \phi^2(\nu_1+1)$, let us call ϕ the power *kernel*, and define the corresponding kernels for other methods with respect to it. This gives $2S/\sqrt{\nu_1+1}$ as the Scheffé kernel $q\xi/\sqrt{\nu_1+1}$ for the Tukey kernel.

Table (1) compares a few values of the power and Scheffé kernels, and Table (2) compares sample sizes for a 3 level factor and a 10 level factor using contrast $[-1\ldots1]$. Because σ, γ, and θ appear in all equations, changes in these parameters will not effect the relative sample sizes. The ratio of sample sizes

between Tukey and the others will change with the contrast however because of ξ in equation (3).

α	β	ν_1	ν_2	power	Scheffé
0.05	0.10	1	20	2.41	2.95
0.05	0.10	10	5	2.75	4.15
0.05	0.30	1	20	1.85	1.35
0.05	0.30	10	5	2.14	2.32
0.01	0.10	1	20	2.98	2.95
0.01	0.10	10	5	4.06	4.15
0.01	0.30	1	20	2.39	1.35
0.01	0.30	10	5	3.75	2.32

Table 8: A comparison of the power and Scheffé kernels.

		3 level			10 level		
α	β	power	Scheffé	Tukey	power	Scheffé	Tukey
0.05	0.10	79	112	103	405	1178	677
0.01	0.10	109	112	103	533	1178	677
0.05	0.30	50	59	45	266	854	795
0.01	0.30	74	59	45	376	854	795

Table 9: A comparison of sample sizes for $\gamma/\sigma = 1$

Although it is arguable whether or not one should equate power to the confidence coefficient, when this is done, it may be seen from Table (2) that the sample sizes for power seem to be smaller in general.

D Numerical methods

D.1 Probability distributions

The probability distributions are obtained as follows:

1. The function ϕ is evaluated by using polynomial approximations for the noncentral distributions and by using more exact series and continued fraction evaluations. The Laguerre series form of the noncentral F given by Tiku (1965) is used. The series and its derivative are used with Newton's method to evaluate the noncentrality parameter. The noncentral χ^2 is obtained from this by setting ν_2 to a large value.
2. The incomplete beta function probabilities and hence the F and t probabilities are obtained by use of Abramowitz and Stegun (1970) 26.5.8

3. The inverse of the incomplete beta function and hence the F and t percentage points are obtained by using Newton iterations with Eq 26.5.22 from Abamaowitz and Stegun (1970) as an initial approximation. The iterations are with respect to log(x) to avoid problems at the lower limit.
4. The natural logarithm of the gamma function is obtained from CACM 291 due to M.C. Pike and I.D. Hill.
5. The gamma and hence χ^2 probabilities are calculated using Eq XXVIII on page XV of Pearson (1957) when $x < 1$ or $x < \nu$ otherwise EQ 6.5.31 from Abramowitz and Stegun is used. In addition EQ 6.5.13 from Abramowitz and Stegun is used when ν is an integer.
6. The inverse of the gamma and hence χ^2 is obtained by Newtonian iterations with respect to log(x) to avoid problems at the lower limit of integration. The initial approximation is Eq 26.4.17 from Abramowitz and Stegun.
7. The inverse of the normal integral uses CACM 442 due to G.W. Hill and A.W. Davis.
8. The normal integral uses Abramowitz and Stegun (1970) 26.2.15 for $|x| < 3.1$ and otherwise uses 26.2.14.
9. The Studentized range calculation uses numerical integration for exact results; however, this seems to rarely improve the approximate results which are made using an idea from Patnaik (1950) with the help of moments of the Studentized range from Harter (1969).

D.2 Transformations

1. **binomial:** $(2 \arcsin \sqrt{p})$, for proportion p.
2. **Poisson:** $\sqrt{x}$, for value x.
3. χ^2 **variance:** $\left(\sqrt{\frac{9\nu}{2}}\left[(\frac{x}{\nu})^{1/3} - 1 + \frac{2}{9\nu}\right]\right)$, for a value x with ν the degrees of freedom: due to Wilson and Hilferty (1931).

D.3 Approximate two population sample size formulas

Because the binomial and Poisson are discrete, it is not possible to obtain critical regions of precisely the specified sizes without the use of randomization. As a consequence no attempt was made to make more exact calculations, and sample size tables will need to be consulted if this is an issue. It should be noted, as Gale (1974) points out, that these approximate formulas cannot be much improved since their major error is with respect to the precise size of the critical region. The χ^2 variance, being continuous, is of course calculated precisely when the exact key is pressed.

1. **binomial:**

$$N = \frac{m'(r+1)}{4r}\left[1+\sqrt{\left\{1+\frac{2(r+1)}{m'\delta}\right\}}\right]^2,$$

where

$$m' = \frac{[z_\alpha\sqrt{(r+1)\bar{P}\bar{Q}} + z_\beta\sqrt{rP_1Q_1+P_2Q_2}]^2}{\delta^2},$$

with $\bar{P} = (P_1 + rP_2)/(r+1)$ and $\bar{Q} = 1-\bar{P}$, z_α and z_β are the upper percentage points of the standard normal distribution, $\delta = P_2 - P_1$ with $P_2 > P_1$ for the two proportions, and N the total sample size which is $(r+1)$ times the size of the smaller sample size. This is from Fleiss (1980).

2. **Poisson:**

$$N = \frac{(z_\alpha + z_\beta)^2}{4(\sin^{-1}\sqrt{\rho/(1+\rho)} - \sin^{-1}\sqrt{1/2})^2},$$

where z_α and z_β are the upper percentage points of the standard normal distribution, and ρ is the ratio of Poisson rates. This is from Gail (1974).

3. χ^2 **variance:**

$$N = (r+1)\frac{(z_\alpha/\sqrt{r} + z_\beta)^2}{\Delta^2},$$

where N is the total sample size, r the ratio of sample sizes, z_α and z_β the upper percentage points of the standard normal distribution, and $\Delta = \sqrt{9/2}(\rho^{1/3}-1)$, with ρ the ratio of sample variances.

BIBLIOGRAPHY

1. Abramowitz, M. and Stegun, I.A. (1970). *Handbook of mathematical functions with formula, graphs and mathematical tables.* Nat. Bureau. Stds. App. Math. Series 55. Sup. of Doc.

2. Berkson. J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *JASA.* 33. 526-336.

3. Bailar, J.C. and Mosteller, F.C. ed. (1992). *Medical uses of statistics, 2nd ed* NEJM Books, Boston.

4. Birnbaum, A. (1954). Statistical methods for Poisson processes and exponential populations. *JASA.* 49. 254-266.

5. Casagrande, J.T. and Pike, M.C. (1978) An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* 34. 385-486.

6. Cohen, Jacob. (1969,77,88). *Statistical power analysis for the behavioral sciences.* Erlbaum, New Jersey.

7. Cohen, Jacob. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology,* 65. 145-153.

8. Fisher. R.A. (1959). *Statistical methods and scientific inference.* Hafner, N.Y.

9. Fisher, R.A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society, Series A.* 98. 39-54.

10. Fleiss, J.L., Tytun, A. and Ury, H.K. (1980) A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics.* 36. 343-346.

11. Fleiss, J.L. (1981). *Statistical methods for rates and proportions.* Wiley. N.Y.

12. Freedman, D., Pisani, R. Purves, R., and Adhikari, A. (1991). *Statistics.* 2d Ed. W.W. Norton, N.Y.

13. Gail M. (1974). Power computations for designing comparative Poisson trials. *Biometrics* 30. 231-237.

14. Gonick, L. and Smith, W. (1993). *The cartoon guide to statistics.* Harper Perennial, N.Y.

15. Harkness, W.L. and Katx, L. (1964). Comparison of the power functions for the test of independence in 2x2 contingency tables. *Annals of Math. Stat.* 35. 1115-1127.

16. Harter, H.L. (1969). *Order statistics and their use in testing and estimation. Vol I.* ARL document. Sup. of Doc.

17. Hoenig, J.M. and Heisey, D.M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician.* 55-1. 19-24.

18. Hosmer, D.W. and Lemeshow, S. (1989). *Applied logistic regression.* Wiley, N.Y.

19. Hsu, Jason C. (1996) *Multiple comparisons theory and methods.* Chapman Hall, London.

20. Kendall, M.G. and Stuart, A. (1967). *The advanced theory of statistics. V2, second ed.* Hafner. N.Y.

21. Lenth, Russell, V. (2001) Some practical guidelines for effective sample size determination. *The American Statistician.* 55-3. 187-193.

22. Lachin, J.M. (1977). Sample size determinations for rxc comparative trials. *Biometrics* 33. 315-324.

23. Meng, R.C. and Chapman, D.G. (1966). The power of chi square tests for contingency tables. *JASA*. 61. 965-975.

24. Moore, D.S. and McCabe, G.P. (1993). *Introduction to the practice of statistics.* W.H. Freeman. N.Y.

25. Mitra,S.K. (1958).On the limiting power function of the frequency chi-square test. *Annals of Mathematical Statistics.* 29. 1221-1233.

26. Nelson, L.S. (1987). Comparison of Poisson means: the general case. *Journal of Quality Technology.* 1, 105-109.

27. Neyman, J, and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, A*. CCXXXI. 289-337.

28. Odeh, R.E. and Fox, M. (1975). *Sample size choice, charts for experiments with linear models.* Marcel Decker, N.Y.

29. Patnaik, P.B. (1950). The use of mean range as an estimator of variance in statistical tests. *Biometrika.* 37. 78-87

30. Pearson, E.S. and Hartley, H.O. (1972). *Biometrika tables for statisticians Vol. 2.* Cambridge U. Press, Cambridge.

31. Pearson, Karl. (1957). *Tables of the incomplete* Γ*-function.* Cambridge.

32. McCullagh, P. and Neldar, J.A. (1989). *Generalized linear models.* Chapman and Hall, NY.

33. Richardson, L.F. (1942) Statistics of deadly quarrels. Reprinted in (1956). *The world of mathematics* Simon and Schuster, N.Y.

34. Rossi, J. S. (1995). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology,* 58, 646-656.

35. Scheffé, H. (1959). *The analysis of variance.* Wiley, New York.

36. Sedlmeier, P., and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin,* 105, 309-316.

37. Signorini, D. (1991). Sample size for Poisson regression. *Biometrika.* 78-2. 446-450.

38. Tiku,M.L. (1965). Laguerre series forms of the non-central χ^2 and F distributions, *Biometrika.* 52-3. 415-427.

39. Westlake, W.J. (1979). Statistical aspects of comparative bioavailabiity trials. *Biometrics.* 35. 273-280.

40. Wheeler, Robert. E. (1974). Portable power. *Technometrics.* 16-2. 193-201.

41. Wilson, E.B. and Hilferty, M.M. (1931) The distribution of chi-square. *Proceedings of the National Academy of Sciences.* 17. 684-688.

42. Whittemore, Alice S. (1981). Sample size for logistic regression with small response probability. *JASA.* 76. 27-32.

43. Zumbo, B.D. and Hubley, A.M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician.* 47-2. 385-388. A PDF copy of this paper may be obtained by e-mail from bruno.zumbo@ubc.ca.