Counterintuitive Results for Statistical Tests with High Power

Jim Kennedy
Version of April 20, 2016

Counterintuitive, contradictory results can occur with a classical statistical hypothesis test with extremely high power (typically above .95). If the power of a test is so high that beta (probability of type II error or false negative) is smaller than alpha (probability of type I error or false positive), outcomes with p-values just below the alpha level are more likely to occur under the null hypothesis than under the alternative hypothesis. Even though the study outcome is more consistent with the null hypothesis than with the alternative hypothesis, the null hypothesis is rejected. This inversion between likelihood and p-values is explained and shown graphically in Figures 1 to 3 below.

The relative size of alpha and beta indicate the relative importance of Type I and Type II errors. The traditional recommendation is alpha = .05 and beta = .20 (power = 1 – beta = .80). This rule of thumb indicates that Type I errors are more important than Type II errors—which is a controversial assumption for most scientific research. Higher power is obviously preferable when possible.

When beta is smaller than alpha, the statistical analysis gives priority to avoiding Type II errors. The inversion noted above reflects the fact that some Type I errors are treated as acceptable in order to avoid Type II errors. This situation may be planned by the researchers, or it may be an unintended consequence when large amounts of data are available.

Statistical tests with sufficient power to make beta less than alpha are unlikely for most individual research studies. However, research syntheses can be expected to increasingly have this property as adequately powered individual studies are conducted.

The optimal option when beta is smaller than alpha (typically when power is above .95) usually is to set alpha and beta to be equal. However, this requires an accurate estimate of effect size and may be difficult to implement if the range of possible effect sizes is wide. Alpha and beta must be set prospectively.

Another option is to ignore the inversion. The probability is small that the outcome of the analysis will be in the area with the inversion. However, the potential controversy may be large when that outcome does occur.

A more general option is to recalibrate the interpretation of p-values. This option has been suggested by various others for various methodological reasons. The potential inversion described here adds another reason for the recalibration. The interpretation of p-values could be:

$p \leq .05$ and $> .01$ – tentative statistical evidence for an effect; confirmation is needed

$p \leq .01$ and $> .001$ – moderate statistical evidence for an effect

$p \leq .001$ and $> .0000001$ – strong statistical evidence for an effect

$p \leq .0000001$ – compelling statistical evidence for an effect

These interpretations would be particularly appropriate for meta-analyses or other situations with very high power. They also promote the recognition that multiple well-powered studies are typically needed for strong evidence.

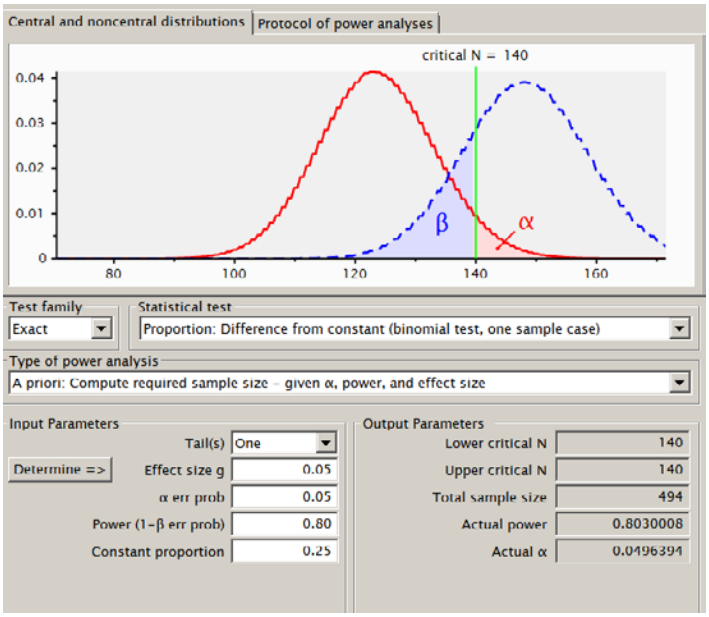

Figure 1. Plot of typical power analysis from the GPower program. The red curve is the null hypothesis and the blue curve is the alternative hypothesis for a binomial test with alpha = .05 and power = .80 (beta = .20). The green line is the number of events or successful trials for a significant outcome. If the study outcome is to the right of the green line, the null hypothesis is rejected. The height of the curve at any point represents the probability (likelihood) that the number of events on the horizontal axis occurred under the null hypothesis (red curve) or under the alternative hypothesis (blue curve). For this case, all outcomes in the rejection (α) area are more likely under the alternative hypothesis than under the null hypothesis—i.e., the blue line is above the red line for the area to the right of the green line.

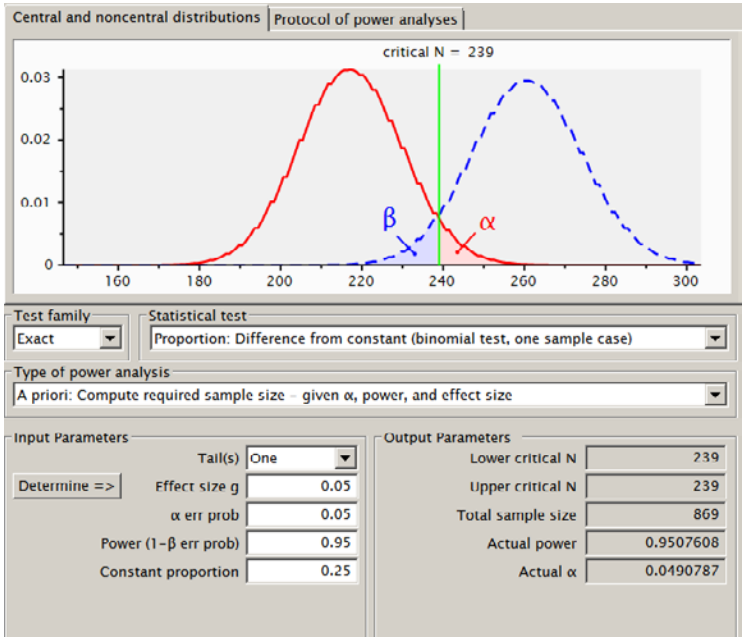

Figure 2. This shows the power analysis when alpha and beta are both equal to .05 (power = .95). The outcomes in the rejection (α) area are more likely under the alternative hypothesis than under the null hypothesis. However, the curves cross at the criterion for significance.

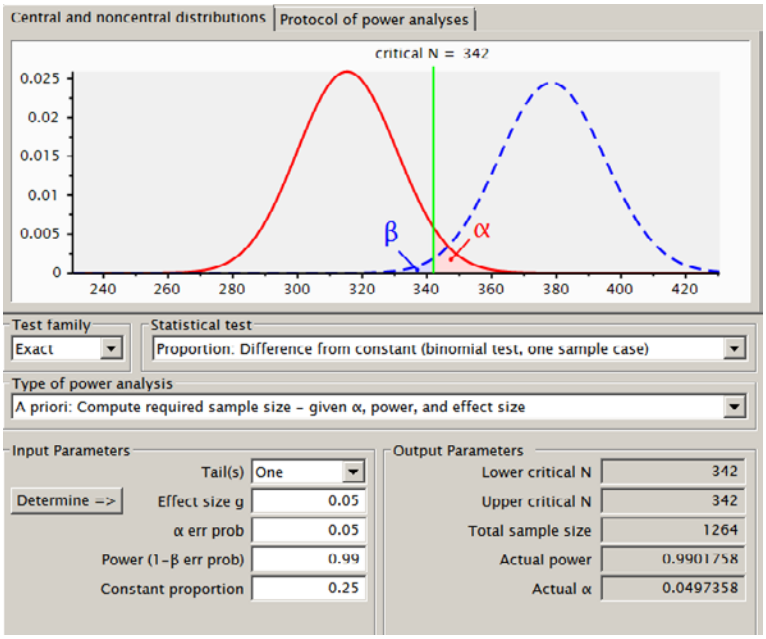

Figure 3. This shows the power analysis when beta is less than alpha. For this case beta is .01 reflecting a power of .99. Alpha remains at .05. For a segment of the rejection (α) area to the right of the green line, the likelihood for the null hypothesis (red line) is greater than for the alternative hypothesis (blue line). This is the area of inversion when a classical test will reject the null hypothesis for an outcome that is actually more likely under the null hypothesis than under the alternative hypothesis.

This situation may be desired if avoiding Type II errors is more important than avoiding Type I errors. However, it may also be an unintended consequence of large amounts of data. Note the analogous situation in Figure 1. The null hypothesis is accepted in some cases when the study outcome is more likely to have occurred under the alternative hypothesis than under the null hypothesis. This occurs because beta is larger than alpha, which implies that avoiding Type I errors is more important than avoiding Type II errors.

If large amounts of data are available, beta may be less than alpha even when alpha is small such as .01 or smaller. For example, three studies could be done for the binomial example presented in the figures here. If each study has a power of .90 with alpha of .05, a meta-analysis that combines the studies will have a power of .9968 and beta of .0032 for alpha of .01. Analyses such as weighted Stouffer's Z have power similar to the power for the combined studies. This type of analysis is often used in parapsychology to evaluate whether the studies provide overall evidence for an effect.

The inversions described here are a factor in the *Jeffreys-Lindley paradox* that states that with sufficiently large sample size a Bayesian analysis will support the null hypothesis when classical analysis supports the alternative hypothesis. Some Bayesian analysts have claimed that these differences between Bayesian and classical results indicate that classical analyses are flawed. However, this argument overlooks the roles of Type I and Type II errors. Inferential errors occur with Bayesian hypothesis tests as well as with classical hypothesis tests. The error rates can and should be quantitatively evaluated for Bayesian hypothesis tests. Similar trade-offs between Type I and Type II errors will be found for Bayesian hypothesis tests when different values for sample size, criterion for acceptable evidence, and prior probability distribution are evaluated.