Commentary

The significance of non-significance

E.M. BERRY, C. COUSTÈRE-YAKIR and N.B. GROVER

From the Department of Human Nutrition & Metabolism and Department of Experimental Medicine and Cancer Research, Hebrew University, Jerusalem, Israel

Summary

We discuss the implications of empirical results that are statistically non-significant. Figures illustrate the interrelations among effect size, sample sizes and their dispersion, and the power of the experiment. All calculations (detailed in Appendix) are based on actual noncentral t-distributions, with no simplifying mathematical or statistical assumptions, and the contribution of each tail is determined separately. We emphasize the importance of reporting, wherever possible, the a priori power of a study so that the reader can see what the chances were of rejecting a null hypothesis that was false. As a practical alternative, we propose that non-significant inference be qualified by an estimate of the sample size that would be required in a subsequent experiment in order to attain an acceptable level of power under the assumption that the observed effect size in the sample is the same as the true effect size in the population; appropriate plots are provided for a power of 0.8. We also point out that successive outcomes of independent experiments each of which may not be statistically significant on its own, can be easily combined to give an overall p value that often turns out to be significant. And finally, in the event that the p value is high and the power sufficient, a non-significant result may stand and be published as such.

Introduction

‘By their very nature, witch hunts are not concerned with the truth. A favorite means of detection of medieval witches was to throw a suspect, bound hand and foot, into the river. If she floated, she was possessed of unnatural powers. If she drowned, she was innocent. Either way, she was dispatched from this world, which was presumably the result intended in the first place ... it is impossible to prove a negative.’

The purpose of this article is to discuss the nature of non-significance. In order to do so, it is necessary to define terms and concepts used in diagnostic and statistical testing. Suppose we believe that the blood level of albumin is higher in diabetics. Let us assume that in the healthy population, albumin is a normally-distributed variable with a mean $\mu_1$ and a standard deviation $\sigma_1$. Then if we draw a random sample of size $n_1$ from this population and measure the albumin levels $x_1$, we would expect that the mean of that sample, $\bar{x}_1$, is distributed about $\mu_1$ with a standard error given by $\sigma_1/\sqrt{n_1}$; this is represented by the heavy bell-shaped curve in Figure 1. Consider now a random sample of size $n_2$ drawn from the diabetic population. If the two populations differ as regards albumin level, then $\bar{x}_2$ would be distributed about another mean, $\mu_2$, say, with a standard error $\sigma_2/\sqrt{n_2}$, where $\sigma_2$ is the standard deviation of the diabetic population, and may or may not differ from $\sigma_1$. This is shown by the light bell-shaped curve in Figure 1. If they do not differ, then $\bar{x}_2$ will be distributed like $\bar{x}_1$, about $\mu_1$.

The true state of nature is, of course, unknown to...
likely it is that the latter is distributed about \( m \), we could formulate a null hypothesis, found. Such an outcome could have been caused by chance (or false-positive rate) and is usually set at 1% or at 5%. The 5% figure has a long history based on such subjective notions of likelihood and was formalized by Sir Ronald A. Fisher in 1925; it has since acquired ‘a magical life of its own’.2

The thick vertical line in Figure 1 has been drawn such that the area to its right under the heavy bell-shaped curve (horizontal hatch) is exactly \( \alpha \). In other words, if \( H_0 \) is true and \( X_2 \) is distributed about \( \mu_2 \), then in a proportion \( \alpha \) of the cases we shall wrongly infer that \( X_2 \) is distributed about \( \mu_2 \). All this is well-known and can be found in any introductory text on statistical inference. The purpose of the present article is to consider in more detail than is usual the complementary error, that which occurs when a value of \( p > \alpha \) is obtained, and so the null hypothesis is not rejected even though \( x_2 \) is actually distributed about \( \mu_2 \) if it is true. This is called the \( \beta \) or Type II error4 and is illustrated in Figure 1 by the area to the left of the vertical line under the light bell-shaped curve (vertical hatch). We shall find it convenient (see Appendix) to introduce the parameter \( \Delta \) as the so-called effect size \( \mu_2 - \mu_1 \) expressed in units of \( \sigma_1 \): \( \Delta = (\mu_2 - \mu_1)/\sigma_1 \); since means and standard deviations have the same units, \( \Delta \) so defined is dimensionless.

The Type II error is the ‘other side’ of statistical significance3 and can be quantified; it is equivalent to the false-negative error. The quantity \( 1 - \beta \) is called the power of the statistical test, and expresses the probability of detecting a true effect, a true positive.6 In cases where the power is low, interpretation of a negative result becomes ambiguous.7

**When to accept the null hypothesis**

There is a distinction between a negative conclusion (no effect, no difference) and no conclusion (that the study was inconclusive). Once the statistical test has shown that chance alone could have produced the observed results with a probability greater than \( \alpha \) (the formal meaning of a non-significant finding), then the researcher should be wary of other explanations. To observe that nothing happened does not mean that nothing happened.8 All it means is that in this particular experiment, no significant effect was found. Such an outcome could have been caused by a true but small difference, or by chance alone, and there is good reason to be suspicious of declarative conclusions drawn from a null hypothesis that is not rejected.

Thus, what one ought to do in designing an experiment in addition to putting a limit on the risk of being wrong in concluding that the blood concentration of albumin is higher in diabetics than in the general population (a Type I error), is to put a limit also on the risk of being wrong in saying that the blood concentrations in the two groups are the same when in fact they are not (a Type II error). Without incorporating this concept into the design it is difficult, if not impossible, to interpret a result that is statistically non-significant. Concern for the probability of missing an important therapeutic improvement, because of insufficient power (small sample size), deserves more attention in the planning of clinical trials. In an analysis of \( \beta \) error, the conclusion was reached that many of the therapies discarded as ineffective after ‘inconclusive, negative’ trials could still have a clinically meaningful effect.9

**Figure 1.** Geometric representation of entities discussed in article. Heavy curve, distribution of means of random samples of size \( n_1 \), drawn from a normally-distributed population with mean \( \mu_1 \) and standard deviation \( \sigma_1 \); light curve, distribution of means of random samples of size \( n_2 \), drawn from a normally-distributed population with mean \( \mu_2 \) and standard deviation \( \sigma_2 \). Horizontal hatch, probability \( \alpha \) of rejecting null hypothesis when it is in fact true; vertical hatch, probability \( \beta \) of accepting null hypothesis when it is in fact false. Specific effect size \( \Delta \) is defined as \((\mu_2 - \mu_1)/\sigma_1 \), Figure drawn to scale for \( \sigma_1 = \sigma_2 \), \( n_1 = n_2 = 20 \), \( \alpha = 0.05 \) 1-tail, \( \beta = 0.20 \). Abscissa in units of \( \sigma_1 \), origin at \( \mu_1 \).
Reason given for non-significant result: small sample size

It is very often said that the results did not attain statistical significance because of inadequate sample size. The reductio ad absurdum of this argument is that if the sample size is made sufficiently large then the result will always be significant. Correlation data afford an instructive example of this phenomenon. For a sample size $n$ of 10, an observed correlation coefficient must be at least 0.55 to be statistically significant at the 5% level; for an $n$ of 25 this falls to 0.34, and for an $n$ of 100 it need be only 0.165. This leads to the confusion between statistical significance and medical relevance, and the error of equating a highly significant result obtained on a large sample with a very strong correlation. The square of the correlation coefficient estimates the proportion of variance in one of the variables that can be attributed to (or accounted for by) the other. Thus, a correlation of 0.5 (which is considered high) implies a 25% commonality, the remaining 75% being attributable to other factors. A correlation of 0.3 or even 0.2 might be highly significant statistically given a large enough $n$, but only means 9% or 4% of shared variance. This can be a source of misinterpretation in large-scale epidemiological studies.\(^\text{10}\)

Setting Type II error

It has been proposed\(^\text{6}\) as a convention that, when the investigator has no other basis for choosing the desired power (1-$\beta$), a value of 0.8 be used, so that $\beta$ is 0.2. Together with the common practice of fixing $\alpha$ at 0.05, this implies that the relative seriousness of these two kinds of errors is around 0.20:0.05—that is, that Type I errors are considered to be four times as serious. (In fact, statistical inference may be viewed as an exercise in balancing these two errors.) A ratio of 0.20:0.05 is not always clinically acceptable, as in the case of suspected appendicitis, where $\alpha$ is commonly about 0.30 (in order to make $\beta$ very small and not miss the diagnosis). Such a high false-positive rate (30%) under these circumstances is considered proper, even good medicine. Sir William Blackstone (1723–80) opined that ‘it is better that 10 guilty persons escape [false negative] than one innocent suffer [false positive], an example from a different discipline where a Type I error is deemed to be very serious indeed.

Calculations of power and sample size

The four major parameters of statistical inference (see Appendix; for didactic reasons, we regard $\alpha$ here as major) are the combined sample size $n$ (= $n_1 + n_2$), the specific effect size $\Delta$, the power $1-\beta$, and the level of significance $\alpha$. They are so related that when any three of them are known, the fourth is completely determined (provided the various other minor parameters are fixed).

We consider the situation in which a medical researcher is comparing the effect of two treatments and has to decide on the size of the groups or, given the $n$, on the power of the experiment; effect size should always be based on clinical rather than statistical considerations.

Figure 2 illustrates the relationship between power and sample size for various values of $\Delta$ with $\alpha=0.05$ 2-tail, $n_2=n_1$ and $\sigma_2=\sigma_1$. It can be applied as follows. 1. To calculate the power for a specific effect size $\Delta$ and combined sample size $n$. Thus, a sample size of 100 and a $\Delta$ of 0.4 would yield a power of 0.5, provided $\alpha=0.05$ 2 tail, $n_2=n_1$, and $\sigma_2=\sigma_1$. 2. To calculate the sample size $n$ for a specific effect size $\Delta$ and power, under the same assumptions concerning the minor parameters.

When the sample sizes required to achieve a particular power are not feasible, something has to give and a compromise made: either admitting a higher $\alpha$, or $\beta$ (or both), or settling for a larger $\Delta$.

Figure 3 shows the effects of partitioning the total sample size unequally, and of an increased variance in the diabetic population: the total sample size $n$ is plotted as a function of specific variance $\Delta$ for various ratios of $n_2 : n_1$ for $\sigma_2 = \sigma_1$ (left-hand panel) and for $\sigma_2 - \sigma_1$ (right-hand panel), for $\alpha=0.05$ 2-tail and a power of 0.80.

When the standard deviations are the same in the two populations, a $\Delta$ of 0.75 requires a total sample size of 58 to reach a power of 0.80 with $n_2 = n_1$, for instance, but an $n$ of 76 with $n_2 : n_1$ at a distribution of 25:75. Thus, in a situation in which it is difficult (or costly) to recruit subjects with disease, acceptable power levels may still be maintained by raising the

![Figure 2](image-url)
Figure 3. Total sample size $n$ as a function of specific effect size $\Delta$ for various values of the relative sample sizes $n_2:n_1$, for $\alpha=0.05$ 2-tail, power $(1-\beta)=0.08$. Left-hand panel: $\sigma_2=\sigma_1$, right-hand panel: $\sigma_2=2\sigma_1$.

total sample size $n$ appropriately: in the present example, an increase in $n$ from 58 to 76 permitted a decrease in the number of ill subjects from 29 to 19. When the standard deviation of the sick population is twice that of the healthy, on the other hand, the corresponding figures are an increase in total $n$ from 142 to 249 in order to allow a decrease in the size of the disease group from 71 to 63, not a very efficient trade-off.

Estimating effect size

The researcher should specify beforehand what magnitude of treatment effect would be regarded as clinically meaningful. Sometimes this is difficult to assess, in which case sample size determinations may be made for different effect sizes that give similar power. Power determinations should be performed a priori, before the experiment is carried out. However, if the observed difference between the two groups is regarded as being clinically relevant, then the power for such an effect size may be determined post hoc from Figures 2 or 3.

Non-significance when power insufficient

Figure 4 shows the relationship between the observed $p$ value (one- and two-tailed) and the total sample size required to yield a power of 0.8 in a subsequent experiment, assuming that the measured effect size in the sample is the same as the true effect size in the population. Thus, if for an initial sample size $n_0$ of 20, a two-tailed $p$ value of 0.30 was observed (clearly a non-significant result), then in the next experiment a sample size of about 110 would be required in order to reach a power of 0.8. If the outcome of this second experiment is again statistically non-significant, then it is reasonable to accept the null hypothesis and to conclude that the two treatments (say) were indeed equivalent, the error incurred in doing so being about 20%.

In this manner, it is possible to qualify a non-significant result by giving the sample size that would be necessary in future experiments to attain a power of 0.8; the smaller the initial $p$ value, the less $n$ has to be increased. It can be shown that the above assumption concerning the measured effect size implies that all non-significant experiments have a post hoc power less than 0.5; as $p$ decreases, the power increases until at $p=\alpha$ it is very slightly above 0.5.

When two negatives equal a positive: combining results

It is well recognized that non-significant results are not published as often as significant ones, and the power calculations presented above are meant to help evaluate which of them may be of importance. Finally, a fact of relevance which is little appreciated is that a series of experiments each yielding non-significant results may, when taken together, attain significance, even high significance. Let us say an experiment yields a value of $p$ that exceeds the preassigned level $\alpha$, but not by very much—$p=0.08$, for instance. Normally we do not reject the null hypothesis even though the probability that the observed difference between $\bar{x}_1$ and $\bar{x}_2$ arose by
Figure 4. Total sample size $n$ required to attain a power of 0.80 in a subsequent experiment as a function of the $p$ value obtained in the first experiment for various values of the initial sample size $n_0$ under the assumption that the true effect size in the population is equal to the observed effect size in the sample, for $\sigma_z = \sigma_r$, $\alpha = 0.05$. Lower abscissa, original hypothesis one-tailed; upper abscissa, original hypothesis two-tailed. Valid for all relative sample sizes $n_2/n_1$ provided they are the same in the two experiments.

chance alone, is only 8%. But suppose a subsequent experiment yields a $p$ value of 0.10. Here too we do not reject $H_0$, but consecutive probabilities of 0.08 and 0.10 imply that $H_0$ is much more unlikely than either result taken alone. Here is a method (one of several in the literature) to combine these $p$ values in order to obtain an overall probability for the entire series of experiments.

The procedure is based on two well-known properties of the $\chi^2$ distribution. Under $H_0$, $p$ is a random uniform variate and as such $-2 \ln p$ follows the $\chi^2$ distribution with two degrees of freedom (DF); in symbols, $-2 \ln p \sim \chi^2_2$. The second property proceeds directly from the definition of $\chi^2$ and states that if we have two independent random variables that follow the $\chi^2$ distribution with degrees of freedom $DF_1$ and $DF_2$, then their sum has the $\chi^2$ distribution with degrees of freedom $DF_1 + DF_2$.

These properties are applicable directly. If we label $p_1$, the result of the first experiment and $p_2$ that of the second, then $-2 \ln p_1 - 2 \ln p_2 \sim \chi^2_2 + \chi^2_2 = \chi^2_4$, and the combined $p$ is obtained by entering a $\chi^2$ table with four DF. In the above example, $p_1 = 0.08$ and $p_2 = 0.10$, giving a $\chi^2$ value of 9.657 and a $p < 0.05$.

This procedure is not limited to two experiments: for $k$ values of $p$, we have $-2 \Sigma \ln p_i \sim \chi^2_k$, where the left-hand side is a compact notation for $-2 \ln p_1 - 2 \ln p_2 - \ldots - 2 \ln p_k$. Thus, for $p_1 = p_2 = 0.10$ and $p_3 = p_4 = 0.15$, as another example, we get $\chi^2_4 = 16.799$, corresponding to a combined $p$ again $< 0.05$.

There are certain rules that must be obeyed in applying this procedure. First of all, the samples must be independent. Secondly, in one-tailed tests all the $p$ values must refer to the same tail. What this says is that if the observed difference between the means is not in the direction implied by $H_1$, then the $p$ value obtained from the usual $t$-table should be subtracted from unity (in other words, in such a case we will have $1 > p > 0.5$) before substituting it in the above expression. For two-tailed tests, there is some disagreement in the literature, but a conservative approach is to convert all two-tailed $p$ values into one-tailed $p$ values in the same tail, enter the above expression to obtain the combined one-tailed $p$ and then double it for the final overall two-tailed $p$ value. Lastly, all results should be used, no selecting allowed!

Note that the methodology does not require that all the $p$ values be derived from the same statistical test, merely that they all postulate the same $H_0$ and $H_1$. Thus one can mix results from two-sample $t$-tests, paired $t$-tests, and even more complicated designs.

A modification of this procedure, that is thought to be more powerful in the type of situations usually encountered with medical data, is to use...
Conclusions

In the reporting of experimental results, including clinical trials, the a priori power of the study should be stated, so that the reader is made aware of what the chances were of rejecting the null hypothesis given that it was false. We also suggest that, wherever possible, non-significant results be qualified by an estimate of the sample size that would be required in a subsequent experiment in order to yield a power of 0.8 under the assumption that the observed effect size in the sample is the same as the true effect size in the population. And finally, the researcher who gets a nearly statistically significant result from a well-conducted experiment should not despair, since if repetition produces similar results, then the two negatives may combine to give a positive, significant result.

Acknowledgments

One of us (EMB) would like to thank Dr Ora Paltiel and Professor Gershom Zajicek for their constructive review of the manuscript.

References


Appendix

The parameters

In the main body of the article, we introduce quite a few parameters. The mean and standard deviation of the healthy population, μ₁ and σ₁, and its sample size n₁, the corresponding quantities for the diabetic population, μ₂, σ₂ and n₂, and the probabilities α and β, the error in rejecting the null hypothesis H₀ when it is in fact true, and the complementary error in not rejecting it when it is in fact false. In addition, there is the question of whether we are considering a one-tailed test or a two-tailed test. This is a formidable list, much too large to handle directly; fortunately, that is not necessary. Several simplifications are possible without detracting from the practical utility of the figures.

First the means, μ₁ and μ₂. We are not interested in their absolute values but only in their difference μ₂ − μ₁. This difference is known as the effect size; as such, it is not a very useful quantity because it has the dimensions of μ. We therefore define the specific effect size Δ as (μ₂ − μ₁)/σ₁; since σ₁ (and, of course, σ₂) has the dimensions of μ as well, Δ is a pure number: the difference between the means measured in units of the standard deviation of the healthy population. Similarly, we define the relative dispersion of the diabetic population as σ₂/σ₁. Considerations of homeostasis suggest that σ₂ be > σ₁; we treat two cases, σ₂ = σ₁ and σ₂ = 2σ₁. (In non-medical settings, σ₂ < σ₁ would be just as likely.)

Rather than deal with n₁ and n₂ separately, we define n the combined sample size as n₁ + n₂. The reason for this is that our calculations depend strongly on n, but are usually quite insensitive to its precise partition between the two samples. Maximum statistical efficiency requires that n₁/n₂ = (σ₂/σ₁)², but n₂ is often less than n₁ because of practical limitations; here we usually take n₂ = n₁, but present curves for n₂ = n₁ = 10:90, 25:75, 75:25, 90:10 as well.

Convention demands that x be set at 0.05 or, less commonly, at 0.01; we have conformed. There is less agreement (and much less concern, as we try to show) with β. In one case, we plot it (actually we
use power, which is $1 - \beta$ against $n$; in the others, we fix the power at 0.80.

To summarize, then, we have three major parameters $\Delta, n, \beta$, the knowledge of any two of which allows us to compute the third provided the various minor parameters are fixed: $\sigma_1, n_2, \alpha$, and whether the test is one-tailed or two-tailed.

**The computations**

All programming was done using the Professional Edition of Microsoft Fortran PowerStation. The acronyms below refer to subroutines from the IMSL Math and Stat Libraries.

For Figure 2, we used the inverse $t$ distribution (TIN) to compute $t$ from $n$ and $\alpha$, then the non-central $t$ distribution (TNDF) to obtain the power $(1 - \beta)$ as a function of $t, n$, and $\Delta$. Since the test here is two-tailed, the power consists of two components: the larger corresponds to $|t|$, the smaller (practically zero unless $n$ and $\Delta$ are both very small) to $-|t|$.

For Figure 3, because an inverse non-central $t$ distribution is not available, we solved TNDF numerically (ZBREN) to obtain $\Delta$ for a predetermined value of $\beta$ as a function of $t, n_1$, and $n_2$ where, as before, $t$ came from the inverse $t$ distribution for a given $n$ and $\alpha$. In the case of $\sigma_2 = 2 \sigma_1$, we used $t'$ in place of $t$ and corrected the DF accordingly. Since DF here is, in general, not an integer whereas TNDF requires that it be, we used harmonic interpolation in DF followed by linear interpolation in $\beta$. Again the hypothesis is two-tailed and so, at least in principle, the power $(1 - \beta)$ consists of two components.

For Figure 4, we found it more convenient to proceed backwards. For a given value of $n$ and $\alpha = 0.05$, we computed the desired $t$ of the second experiment, and from it and $n$, obtained the $\Delta$ by solving the non-central $t$ distribution numerically for fixed power. Finally, $n$ and $n_0$ were used to convert $\Delta$ to the corresponding $t$ of the initial experiment and the $t$ distribution (TDF) with $n_0$ was used to get $p$. It should be pointed out that the curves shown here apply for all values of the ratio $n_2: n_1$ provided only that it is the same in the two experiments.