Intuitive Biostatistics: Interpreting Nonsignificant P values

THE TERM SIGNIFICANT

You've already learned that the term *statistically significant* has a simple meaning: The P value is less than a preset threshold value alpha. That's it! In plain language, a result is statistically "significant" when the result would be surprising if there were really no differences between the overall populations.

It's easy to read far too much into the word significant because the statistical use of the word has a meaning entirely distinct from its usual meaning. Just because a difference is statistically significant does not mean that it is important or interesting. A statistically significant result may not be scientifically significant or clinically significant. And a difference that is not significant (in the first experiment) may turn out to be very important. This is important, so I'll repeat it: **Statistically significant results are not necessarily important or even interesting.**

EXTREMELY SIGNIFICANT RESULTS

Intuitively, you'd think that P = 0.004 is more significant than P = 0.04. Using the strict definitions of the terms, this is not correct. Once you have set a value for alpha, a result either is statistically significant or is not statistically significant. It doesn't matter whether the P value is very close to alpha or far away. Many statisticians feel strongly about this, and think that the word significant should never be prefaced by an adjective. Most scientists are less rigid, and refer to *very significant* or *extremely significant* results when the P value is tiny.

When showing P values on graphs, investigators commonly use a "Michelin Guide" scale. *: $P < 0.05$ (significant), **: $P < 0.01$ (highly significant); ***: $P < 0.001$ (extremely significant). When you read this kind of graph, make sure that you look at the key that defines the symbols, as different investigators use different threshold values.

BORDERLINE P VALUES

If you follow the strict paradigm of statistical hypothesis testing and set alpha to its conventional value of 0.05, then a P value of 0.049 denotes a statistically significant difference and a P value of 0.051 denotes a not significant difference. This arbitrary distinction is unavoidable since the whole point of using the term *statistically significant* is to reach a crisp conclusion from every experiment without exception.

Rather than just looking at whether the result is significant or not, it is better to look at the actual P value. That way you'll know whether the P value is near a or far from it. When a P value is just slightly greater than alpha, some scientists refer to the result as *marginally significant* or *almost significant.*

When the two-tailed P value is between 0.05 and 0.10, it is tempting to switch to a one-tailed P value. The

one-tailed P value is half the two-tailed P value and is less than 0.05, and the results become "significant" as if by magic. Obviously, this is not an appropriate reason to choose a one-tailed P value! The choice should be made before the data are collected.

One way to deal with borderline P values would be to choose between three decisions rather than two. Rather than decide whether a difference is significant or not significant, add a middle category of *inconclusive.* This approach is not commonly used.

THE TERM NOT SICNIFICANT

If the P value is greater than a preset value of alpha, the difference is said to be *not significant.* This means that the data are not strong enough to persuade you to reject the null hypothesis. People often mistakenly interpret a high P value as proof that the null hypothesis is true. That is incorrect. A high P value does not prove the null hypothesis. This is an important point worth repeating: A *high P value does not prove the null hypothesis.* As you've already learned in Chapter 11, concluding that a difference is not statistically significant when the null hypothesis is, in fact, false is called a Type II error.

When you read that a result is not significant, don't stop thinking. There are two approaches you can use to evaluate the study. First, look at the confidence interval (CI). Second, ask about the power of the study to find a significant difference if it were there.

INTERPRETING NOT SIGNIFICANT RESULTS WITH CONFIDENCE INTERVALS

Example 12.1

Ewigman et al. investigated whether routine use of prenatal ultrasound would improve perinatal outcome (reference 1). They randomly divided a large group of pregnant women into two groups. One group received routine ultrasound sonogram exams twice during the pregnancy. The other group received monograms only if there was a clinical reason to do so. The physicians caring for the women knew the results of the monograms and cared for the patients accordingly. The investigators looked at several outcomes. Table 12.1 shows the total number of adverse events, defined as fetal or neonatal deaths of moderate to severe morbidity.

Table 12.1. Results of Example 12.1

	Adverse Outcome	**Healthy Baby**	**Total**
Routine monograms	383	7,302	7,685
Sonograms only when indicated	373	7,223	7,596
Total	756	14,525	15,281

The null hypothesis is that the rate of adverse outcomes is identical in the two groups. In other words, the null hypothesis is that routine use of ultrasound neither prevents nor causes perinatal mortality or morbidity. The two-tailed P value is 0.86. The data provide no reason to reject the null hypothesis.

Before you can interpret the results, you need to know more. You need to know the 95% CI for the relative risk. For this study, the relative risk is 1.02, with the 95% CI ranging from 0.88 to 1. 17. The null hypothesis can be restated as follows: In the entire population, the relative risk is 1.00. The data are consistent with the null hypothesis. This does not mean that the null hypothesis is true. Our CI tells us that the data are also consistent (within 95% confidence) with relative risks ranging from 0.88 to 1.17.

Different people might interpret these data in two ways:

- The confidence interval is very narrow, and is centered close to1.0. These data convince me that routine use of ultrasound is neither helpful nor harmful. To reduce costs, I'll use ultrasound only when there is an identified problem.
- The confidence interval is narrow, but not all that narrow. There have been plenty of studies showing that ultrasound doesn't hurt the fetus, so I'll ignore the part of the confidence interval above 1.00. But ultrasound gives the obstetrician extra information to manage the pregnancy, and it makes sense that using this extra information will decrease the chance of a major problem. The confidence interval goes down to 0.88, a reduction of risk of 12%. If I were pregnant, I'd certainly want to use a risk-free technique that reduces the risk of a sick or dead baby by as much as 12%! The data certainly don't prove that routine ultrasound is beneficial, but the study leaves open the possibility that routine use of ultrasound might reduce the rate of truly awful events by as much as 12%. 1 think the study is inconclusive. Since ultrasound is not prohibitively expensive and appears to have no risk, I will keep using it routinely.

To interpret a not significant P value, you must look at the CI. If the entire span of the CI contains differences (or relative risks) that you consider to be trivial, then you can make a strong negative conclusion. If the CI is wide enough to include values you consider to be clinically or scientifically important, then the study is inconclusive. Different people will appropriately have different opinions about how large a difference (or relative risk) is scientifically or clinically important and may interpret the same not significant study differently.

In interpreting the results of this example, you also need to think about benefits and risks that don't show up as a reduction of adverse outcomes. The ultrasound picture helps reassure parents that their baby is developing normally and gives them a picture to bond with and to show relatives. This can be valuable regardless of whether it reduce-, the chance of adverse Outcomes. Although statistical analyses focus on one outcome at a time, you must consider all the outcomes when evaluating the results.

INTERPRETING *NOT SIGNIFICANT* P VALUES USING POWER ANALYSES

A not significant P value does not mean that the null hypothesis is true. It simply means that your data are not strong enough to convince you that the null hypothesis is not true. As you have already learned, obtaining a not significant result when the null hypothesis is really false is called a Type 11 error. When you obtain (or read about) a not significant P value, you should ask yourself this question: "What is the chance that this is a Type 11 error?" That question can only be answered if you specify an alternative hypothesis -a difference delta (or relative risk R) that you hypothesize exists in the overall population. Then you can ask, "If the true difference is delta (or the true relative risk is R), what is the chance of obtaining a statistically significant result in an experiment of this size? The answer is termed the power of the study,

Table 12.2 shows the power of example 12.1 for various hypothetical relative risks. This table uses the sample size of the example and sets (alpha = 0.05 and the risk in the control subjects to 5.0%. Calculating the power exactly is

quite difficult, but there are several ways to calculate power approximately. You'll learn about the approximation used to create this table in Chapter 27. In Chapter 23 you'll learn how to calculate a similar table for studies that compare two means (rather than two proportions). GraphPad StatMate performs these calculations.

Table 12.2. A Power Analysis of Example 12.1

Relative Risk	Power
0.95	11%
0.90	30%
0.85	60%
0.80	84%
0.75	97%

If the experimental treatment (routine ultrasound) reduced the risk by 25% (so the relative risk = 0.75), the study had a 97% power to detect a significant difference. If the treatment was really that good, then 97% of studies this size would wind up with a statistically significant result, while 4% would come up with a **not** significant result. In contrast, the power of this study to detect a risk reduction of 5% (relative risk = 0.95) is only 11%. If the treatment truly reduced risk by 5%, only 11% of -studies this size would come up with a statistically significant result.

The numbers are particular to this study. The principle is universal. All studies have very little power to detect tiny differences and enormous power to detect large differences. If you increase the number of subjects, you will increase the power.

As you can see, calculations of power can aid interpretation of results that are not statistically significant. When reading biomedical research, however, you'll rarely encounter power calculations in papers that present not significant results. This is partly a matter of tradition, and it is partly because it is difficult to define the smallest difference or relative risk that you think it is important.

SUMMARY

A result is statistically significant when the P value is less than the preset value of alpha. This means that the results would be surprising if the null hypothesis were really true. The statistical use of the term *significant* is quite different than the usual use of the word. Statistically significant results may or may not be scientifically or clinically interesting or important.

When results are statistically not statistically significant, it means that the results are not inconsistent with the null hypothesis. This does *not* mean that the null hypothesis is true. When interpreting results that are not significant, it helps to look at the extent of the Cl and to calculate the power that the study would have found a statistically significant result if the populations really were different (with a difference of a defined size).

Reference

BG Ewigman, IP Crane, FD Frigoletto et al. Effect of prenatal ultrasound screening on perinatal outcome. *N Engl J Med* 329:821-827, 1993.

GraphPad Home