CHAPTER 19

Interpreting a Result That Is Not Statistically Significant

> Extraordinary claims require extraordinary proof.
>
> Carl Sagan

When you see a result that is not statistically significant, don't stop thinking. "Not statistically significant" means only that the P value is larger than a preset threshold. Thus, a difference (or correlation or association) as large as what you observed would not be unusual as a result of random sampling if the null hypothesis were true. This does not prove that the null hypothesis is true. This chapter explains how to use confidence intervals to help interpret those findings that are not statistically significant.

FIVE EXPLANATIONS FOR "NOT STATISTICALLY SIGNIFICANT" RESULTS

Let's continue the simple scenario from the last chapter. You compare cells incubated with a new drug to control cells and measure the activity of an enzyme, and you find that the P value is large enough (greater than 0.05) for you to conclude that the result is not statistically significant. The following discussion offers five explanations to explain why this happened.

Explanation 1: The drug did not affect the enzyme you are studying

Scenario: The drug did not induce or activate the enzyme you are studying, so the enzyme's activity is the same (on average) in treated and control cells.

Discussion: This is, of course, the conclusion everyone jumps to when they see the phrase "not statistically significant." However, four other explanations are possible.

Explanation 2: The drug had a trivial effect

Scenario: The drug may actually affect the enzyme but by only a small amount.

Discussion: This explanation is often forgotten.

Explanation 3: There was a Type II error

Scenario: The drug really did substantially affect enzyme expression. Random sampling just happened to give you some low values in the cells treated with the drug and some high levels in the control cells. Accordingly, the P value was large, and you conclude that the result is not statistically significant.

Discussion: How likely are you to make this kind of Type II error? It depends on how large the actual (or hypothesized) difference is, on the sample size, and on the experimental variation. This topic is covered in Chapter 26.

Explanation 4: The experimental design was poor

Scenario: In this scenario, the drug really would increase the activity of the enzyme you are measuring. However, the drug was inactivated because it was dissolved in an acid. Since the cells were never exposed to the active drug, of course the enzyme activity didn't change.

Discussion: The statistical conclusion was correct—adding the drug did not increase the enzyme activity—but the scientific conclusion was completely wrong.

Explanation 5: The results cannot be interpreted due to dynamic sample size

Scenario: In this scenario, you hypothesized that the drug would not work, and you really want the experiment to validate your prediction (maybe you have made a bet on the outcome). You first ran the experiment three times, and the result (n = 3) was statistically significant. Then you ran it three more times, and the pooled results (now with n = 6) were still statistically significant. Then you ran it four more times, and finally the results (with n = 10) were not statistically significant. This n = 10 result (not statistically significant) is the one you present.

Discussion: The P value you obtain from this approach simply cannot be interpreted.

"NOT SIGNIFICANTLY DIFFERENT" DOES NOT MEAN "NO DIFFERENCE"

A large P value means that a difference (or correlation or association) as large as what you observed would happen frequently as a result of random sampling. But this does not necessarily mean that the null hypothesis of no difference is true or that the difference you observed is definitely the result of random sampling.

Vickers (2010) tells a great story that illustrates this point:

> The other day I shot baskets with [the famous basketball player] Michael Jordan (remember that I am a statistician and never make things up). He shot 7 straight free throws; I hit 3 and missed 4 and then (being a statistician) rushed to the sideline, grabbed my laptop, and calculated a *P* value of .07 by Fisher's exact test. Now, you wouldn't take this *P* value to suggest that there is *no* difference between my basketball skills and those of Michael Jordan, you'd say that our experiment hadn't *proved* a difference.

A high P value does not prove the null hypothesis. Deciding not to reject the null hypothesis is not the same as believing that the null hypothesis is definitely true. The absence of evidence is not evidence of absence (Altman & Bland, 1995).

EXAMPLE: α_2-ADRENERGIC RECEPTORS ON PLATELETS

Epinephrine, acting through α_2-adrenergic receptors, makes blood platelets stickier and thus helps blood clot. We counted these receptors and compared people with normal and high blood pressure (Motulsky, O'Connor, & Insel, 1983). The idea was that the adrenergic signaling system might be abnormal in people with high blood pressure (hypertension). We were most interested in the effects on the heart, blood vessels, kidney, and brain but obviously couldn't access those tissues in people, so we counted receptors on platelets instead. Table 19.1 shows the results.

The results were analyzed using an unpaired t test (see Chapter 30). The average number of receptors per platelet was almost the same in both groups, so of course the P value was high, 0.81. If the two populations were Gaussian with identical means, you'd expect to see a difference as large or larger than that observed in this study in 81% of studies of this size.

Clearly, these data provide no evidence that the mean receptor number differs in the two groups. When my colleagues and I published this study 30 years ago, we stated that the results were not statistically significant and stopped there, implying that the high P value proves that the null hypothesis is true. But that was not a complete way to present the data. We should have interpreted the CI.

The 95% CI for the difference between group means extends from –45 to 57 receptors per platelet. To put this in perspective, you need to know that the average number of α_2-adrenergic receptors/platelet is about 260. We can therefore rewrite the CI as extending from –45/260 to 57/260, which is from –17.3% to 21.9%, or approximately plus or minus 20%.

It is only possible to properly interpret the CI in a scientific context. Here are two alternative ways to think about these results:

- A 20% change in receptor number could have a huge physiological impact. With such a wide CI, the data are inconclusive, because they are consistent with no difference, substantially more receptors on platelets from people with hypertension, or substantially fewer receptors on platelets of people with hypertension.

	CONTROLS	HYPERTENSION
Number of subjects	17	18
Mean receptor number (receptors/platelet)	263	257
SD	87	59

Table 19.1. Number of α_2-adrenergic receptors on the platelets of controls and people with hypertension.

- The CI convincingly shows that the true difference is unlikely to be more than 20% in either direction. This experiment counts receptors on a convenient tissue (blood cells) as a marker for other organs, and we know the number of receptors per platelet varies a lot from individual to individual. For these reasons, we'd only be intrigued by the results (and want to pursue this line of research) if the receptor number in the two groups differed by at least 50%. Here, the 95% CI extended about 20% in each direction. Therefore, we can reach a solid negative conclusion that either there is no change in receptor number in individuals with hypertension or that any such change is physiologically trivial and not worth pursuing.

The difference between these two perspectives is a matter of scientific judgment. Would a difference of 20% in receptor number be scientifically relevant? The answer depends on scientific (physiological) thinking. Statistical calculations have nothing to do with it. Statistical calculations are only a small part of interpreting data.

EXAMPLE: FETAL ULTRASOUNDS

Ewigman et al. (1993) investigated whether the routine use of prenatal ultrasounds would improve perinatal outcomes. The researchers randomly divided a large pool of pregnant women into two groups. One group received routine ultrasound exams (sonograms) twice during their pregnancy. The other group was administered sonograms only if there was a clinical reason to do so. The physicians caring for the women knew the results of the sonograms and cared for the women accordingly. The investigators looked at several outcomes. Table 19.2 shows the total number of adverse events, defined as fetal or neonatal deaths (mortality) or moderate to severe morbidity.

The null hypothesis is that the risk of adverse outcomes is identical in the two groups. In other words, the null hypothesis is that routine use of ultrasounds neither prevents nor causes perinatal mortality or morbidity, so the relative risk equals 1.00. Chapter 27 explains the concept of relative risk in more detail.

Table 19.2 shows that the relative risk is 1.02. That isn't far from the null hypothesis value of 1.00. The two-tailed P value is 0.86.

	ADVERSE OUTCOME	TOTAL	RISK (%)	RELATIVE RISK
Routine ultrasound	383	7,685	5.0	1.020
Only when indicated	373	7,596	4.9	
Total	756	15,281		

Table 19.2. Relationship between fetal ultrasounds and outcome.

The risks in Column 4 are computed by dividing the number of adverse outcomes by the total number of pregnancies. The relative risk is computed by dividing one risk by the other (see Chapter 27 for more details).

Interpreting the results requires knowing the 95% CI for the relative risk, which a computer program can calculate. For this example, the 95% CI ranges from 0.88 to 1.17.

Our data are certainly consistent with the null hypothesis, because the CI includes 1.0. 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.

Here are three approaches to interpreting the results:

- The CI is centered on 1.0 (no difference) and is quite narrow. These data convincingly show that routine use of ultrasound is neither helpful nor harmful.
- The CI is narrow but not all that narrow. It certainly makes clinical sense that the extra information provided by an ultrasound will help obstetricians manage the pregnancy and might decrease the chance of a major problem. The CI goes down to 0.88, a risk reduction 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% (from 5.0% to 4.4%)! The data certainly don't prove that a routine ultrasound is beneficial, but the study leaves open the possibility that routine ultrasound might reduce the rate of awful events by as much as 12%.
- The CI goes as high as 1.17. That is a 17% relative increase in problems (from 5.0% to 5.8%). Without data from a much bigger study, these data do not convince me that ultrasounds are helpful and make me worry that they might be harmful.

Statistics can't help to resolve the differences among these three mindsets. It all depends on how you interpret the relative risk of 0.88 and 1.17, how worried you are about the possible risks of an ultrasound, and how you combine the data in this study with data from other studies (I have no expertise in this field and have not looked at other studies).

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 show relatives. This can be valuable regardless of whether it reduces 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.

HOW TO GET NARROWER CIS

Both previous examples demonstrate the importance of interpreting the CI in the scientific context of the experiment. Different people will appropriately have different opinions about how large a difference (or relative risk) is scientifically or clinically important and will therefore interpret a not statistically significant result differently.

If the CI is wide enough to include values you consider clinically or scientifically important, then the study is inconclusive. In some cases, you might be able to narrow the CIs by improving the methodology and thereby reducing the SD. But in most cases, increasing the sample size is the only approach to narrowing the CI in a repeat study. This rule of thumb can help: if sample size is increased by a factor of four, the CI is expected to narrow by a factor of two. More generally, the width of a CI is inversely proportional to the square root of sample size.

WHAT IF THE P VALUE IS REALLY HIGH?

If you ran many experiments in which the null hypothesis was really true, you'd expect the P values to be uniformly distributed between 0.0 and 1.0. Half of the P values would have values greater than 0.5, and 10% would have values greater than 0.9. But what do you conclude when the P value is really high?

In 1866, Mendel published his famous paper on heredity in pea plants (Mendel, 1866). This was the first explanation of heredity and recessive traits and really founded the field of genetics. Mendel proposed a model of recessive inheritance, designed an experiment with peas to test the model, and showed that the data fit the model very well. Extremely well! Fisher (1936) reviewed these data and then pooled all of Mendel's published data to calculate a P value that answered the question: Assuming that Mendel's genetic theory is correct and every plant was classified correctly, what is the probability that the deviation between expected and observed would be as great or greater than actually observed?

The answer (the P value) is 0.99993. So there clearly is no evidence of deviation from the model. The deviation from theory is not statistically significant. That is where most people would stop. Fisher went further. If that null hypothesis were true, and you ran similar experiments many times, the P values would be uniformly distributed between zero and 1. So what is the chance of getting a P value of 0.99993 or higher? The answer is 1 – 0.99993 or 0.00007 or 0.007%. That is pretty small. It could be a rare coincidence. Fisher concluded that since the data presented match the theory so well (with very little of the expected random deviation from the theory), Mendel (or his assistants) must have massaged the data a bit. The data presented are simply too good to be true.

Q & A

If a P value is greater than 0.05, can you conclude that you have disproven the null hypothesis?

No.

If you conclude that a result is not statistically significant, it is possible that you are making a Type II error as a result of missing a real effect. What factors influence the chance of this happening?

The probability of a Type II error depends on the significance level (α) you have chosen, the sample size, and the size of the true effect.

By how much do you need to increase the sample size to make a CI half as wide?

A general rule of thumb is that increasing the sample size by a factor of 4 will cut the expected width of the CI by a factor of 2. (Note that 2 is the square root of 4.)

What if I want to make the CI one-quarter as wide as it is?

Increasing the sample size by a factor of 16 will be expected to reduce the width of the CI by a factor of 4. (Note that 4 is the square root of 16.)

Can a study result can be consistent both with an effect existing and with it not existing?

Yes! Clouds are not only consistent with rain but also with no rain. Clouds, like noisy results, are inconclusive (Simonsohn, 2016).

CHAPTER SUMMARY

- If a statistical test computes a large P value, you should conclude that the findings would not have been unusual if the null hypothesis were true.
- You should *not* conclude that the null hypothesis of no difference (or association, etc.) has been proven.
- When interpreting a high P value, the first thing to do is look at the size of the effect.
- Also look at the CI of the effect.
- If the CI includes effect sizes that you would consider to be scientifically important, then the study is inconclusive.