CHAPTER 19

# Interpreting a Result That Is Not Statistically Significant

> I was gratified to be able to answer promptly, and I did. I said I didn't know.
>
> MARK TWAIN

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

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

A large P value means that a difference (correlation, 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 (2006a) told 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).

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

**Table 19.1. Number of $\alpha_2$-adrenergic receptors on the platelets of controls and people with hypertension.**

## EXAMPLE: $\alpha_2$-ADRENERGIC RECEPTORS ON PLATELETS

Epinephrine, acting through $\alpha_2$-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 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 with 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 had 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 I published this study 25 years ago, I 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/platelet. To put this in perspective, you need to know that the average number of receptors per platelet is about 260. Therefore, the 95% confidence interval extends approximately plus or minus 20%.

It is only possible to properly interpret the confidence in a scientific context. Here are two alternative, contradictory, 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.
- 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

| | ADVERSE OUTCOME | TOTAL | RISK | RELATIVE RISK |
|---|---|---|---|---|
| Routine ultrasound | 383 | 7,685 | 0.050% | 1.020 |
| Only when indicated | 373 | 7,596 | 0.049% | |
| 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).

> 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 any such change is physiologically trivial and not worth pursuing.

Those two conclusions contradict each other. The difference 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 ultrasound would improve perinatal outcome. They randomly divided a large group of pregnant women into two groups. One group received routine ultrasound exams (or, *sonograms*) twice during the pregnancy. The other group received 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 ultrasound neither prevents nor causes perinatal mortality or morbidity, so the relative risk equals 1.00. Chapter 27 will explain 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-tail P value is 0.86.

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