And yet...looking at the confidence interval above, we see that the difference in HS12 between cover types 1 and 2 is tiny when compared to the general size of HS12, in the 200s. Thus HS12 is not going to help us guess which cover type exists at a given location. In this sense, the difference is not "significant" at all. And this is why the American Statistical Association released their historic position paper, warning that p-values were overused and often misinterpreted.

10.15 Problems with Significance Testing

Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path — Paul Meehl, professor of psychology and the philosophy of science, referring to Fisher, one of the major founders of statistical methodology

Significance testing is a time-honored approach, used by tens of thousands of people every day. But although significance testing is mathematically correct, many consider it to be at best noninformative and at worst seriously misleading.

10.15.1 History of Significance Testing, and Where We Are Today

When the concept of significance testing, especially the 5% value for α, was developed in the 1920s by Sir Ronald Fisher, many prominent statisticians opposed the idea — for good reason, as we'll see below. But Fisher was so influential that he prevailed, and thus significance testing became the core operation of statistics.

So, significance testing became entrenched in the field, in spite of being widely recognized as potentially problematic to this day. Most modern statisticians understand this, even if many continue to engage in the practice.[6] For instance, there is an entire chapter devoted to this issue in one of the best-selling elementary statistics textbooks in the US [16].

One of the authors of that book, Professor David Freedman of UC Berkeley, was commissioned to write a guide to statistics for judges [15]. The

[6] Many are forced to do so, e.g. to comply with government standards in pharmaceutical testing. My own approach in such situations is to quote the test results but then point out the problems, and present confidence intervals as well.

discussion there of potential problems with significance testing is similar to that of our next section here. These views are held by most statisticians, and led to the ASA statement cited above.

10.15.2 The Basic Issues

To begin with, it's questionable to test H_0 in the first place, because we almost always know *a priori* that H_0 is false.

Consider the coin example, for instance. No coin is absolutely perfectly balanced — e.g. the *bas relief* bust of Abraham Lincoln on the "heads" side of the US penny would seem to make that side heavier — and yet that is the question that significance testing is asking:

$$H_0 : p = 0.5000000000000000000000000000... \tag{10.30}$$

We know before even collecting any data that the hypothesis we are testing is false, and thus it's nonsense to test it.

But much worse is this word "significant." Say our coin actually has $p = 0.502$. From anyone's point of view, that's a fair coin! But look what happens in (10.24) as the sample size n grows. If we have a large enough sample, eventually the denominator in (10.24) will be small enough, and $\widehat{p}$ will be close enough to 0.502, that Z will be larger than 1.96 and we will declare that p is "significantly" different from 0.5. But it isn't! Yes, 0.502 is different from 0.5, but NOT in any significant sense in terms of our deciding whether to use this coin in the Super Bowl.

The same is true for government testing of new pharmaceuticals. We might be comparing a new drug to an old drug. Suppose the new drug works only, say, 0.4% (i.e. 0.004) better than the old one. Do we want to say that the new one is "signficantly" better? This wouldn't be right, especially if the new drug has much worse side effects and costs a lot more (a given, for a new drug).

Note that in our analysis above, in which we considered what would happen in (10.24) as the sample size increases, we found that eventually *everything* becomes "signficiant"—even if there is no practical difference. This is especially a problem in computer applications of statistics, because they often use very large data sets.

That is what we saw in the forest cover example above. The p-value was essentially 0, yet the difference in population means was so small that it

was negligible in terms of our goal of predicting cover type.

In all of these examples, the standard use of significance testing can result in our pouncing on very small differences that are quite insignificant to us, yet will be declared "significant" by the test.

Conversely, if our sample is too small, we can miss a difference that actually *is* significant — i.e. important to us — and we would declare that p is NOT significantly different from 0.5. In the example of the new drug, this would mean that it would be declared as "not significantly better" than the old drug, even if the new one is much better but our sample size wasn't large enough to show it.

In summary, the basic problems with significance testing are

- H_0 is improperly specified. What we are really interested in here is whether p is *near* 0.5, not whether it is *exactly* 0.5 (which we know is not the case anyway).
- Use of the word *significant* is grossly improper (or, if you wish, grossly misinterpreted).

10.15.3 Alternative Approach

I was in search of a one-armed economist, so that the guy could never make a statement and then say: "on the other hand" — President Harry S Truman

If all economists were laid end to end, they would not reach a conclusion— Irish writer George Bernard Shaw

Note carefully that this is not to say that we should not make a decision. We *do* have to decide, e.g. decide whether a new hypertension drug is safe or in this case decide whether this coin is "fair" enough for practical purposes, say for determining which team gets the kickoff in the Super Bowl. But it should be an informed decision.

In fact, the real problem with significance tests is that they **take the decision out of our hands**. They make our decision mechanically for us, not allowing us to interject issues of importance to us, such possible side effects in the drug case.

Forming a confidence interval is the more informative approach. In the coin example, for instance:

- The width of the interval shows us whether n is large enough for $\widehat{p}$ to be reasonably accurate.

- The location of the interval tells us whether the coin is fair enough for our purposes.

Note that in making such a decision, we do NOT simply check whether 0.5 is in the interval. That would make the confidence interval reduce to a significance test, which is what we are trying to avoid. If for example the interval is (0.502,0.505), we would probably be quite satisfied that the coin is fair enough for our purposes, even though 0.5 is not in the interval.

On the other hand, say the interval comparing the new drug to the old one is quite wide and more or less equal positive and negative territory. Then the interval is telling us that the sample size just isn't large enough to say much at all.

In the movies, you see stories of murder trials in which the accused must be "proven guilty beyond the shadow of a doubt." But in most noncriminal trials, the standard of proof is considerably lighter, *preponderance of evidence*. This is the standard you must use when making decisions based on statistical data. Such data cannot "prove" anything in a mathematical sense. Instead, it should be taken merely as evidence. The width of the confidence interval tells us the likely accuracy of that evidence. We must then weigh that evidence against other information we have about the subject being studied, and then ultimately make a decision on the basis of the preponderance of all the evidence.

Yes, juries must make a decision. But they don't base their verdict on some formula. Similarly, you the data analyst should not base your decision on the blind application of a method that is usually of little relevance to the problem at hand—significance testing.

10.16 The Problem of "P-hacking"

The (rather recent) term *p-hacking* refers to the following abuse of statistics.[7]

[7]The term *abuse* here will not necessarily connote intent. It may occur out of ignorance of the problem.