Chapter 23

Use and Abuse of Statistical Inference

Lecture Slides

Case Study: Use and Abuse of Statistical Inference 1

Suppose we take a look at the more than 10,000 mutual funds on sale to the investing public.

Any Internet investment site worth clicking on will tell you which funds produced the highest returns over the past (say) 3 years.

Ocean/Corbis

Case Study: Use and Abuse of Statistical Inference 2

In 2011, one site claimed that the ProFunds Internet Inv. Fund was among the top 1% over the 3 preceding years.

If we had bought this fund in 2008, we would have gained 50.2% per year.

Comparing this return with the average over this period for all funds, we find that ProFunds Internet Inv. Fund return is significantly higher.

Doesn't statistical significance suggest that the ProFunds Internet Inv. Fund is a good investment?

Case Study: Use and Abuse of Statistical Inference 3

In this chapter, we will take a careful look at what statistical confidence and statistical significance do and do not mean.

We will discuss some abuses of statistical inference.

By the end of this chapter, you will be able to answer the question of whether there was strong evidence that the ProFunds Internet Inv. Fund was a sound investment in 2011.

Using Inference Wisely 1

We have met the two major types of statistical inference: confidence intervals and significance tests.

We have, however, seen only two inference methods of each type, one designed for inference about a population proportion *p* and the other designed for inference about a population mean μ.

There are many methods for inference about various parameters in various settings. The reasoning of confidence intervals and significance tests remains the same, regardless of the method.

Using Inference Wisely 2

The first step in using inference wisely is to understand your data and the questions you want to answer and fit the method to its setting.

The design of the data production matters. "Where do the data come from?" remains the first question to ask in any statistical study.

Using Inference Wisely 3

For our confidence interval and test for a proportion *p*:

- The data must be a simple random sample (SRS) from the population of interest. When you use these methods, you are acting as if the data are an SRS. In practice, it is often not possible to actually choose an SRS from the population. Your conclusions may then be open to challenge.

- These methods are not correct for sample designs more complex than an SRS, such as stratified samples. There are other methods that fit these settings.

Using Inference Wisely 4

For our confidence interval and test for a proportion *p*:

- There is no correct method for inference from data haphazardly collected with bias of unknown size. Fancy formulas cannot rescue badly produced data.

- Other sources of error, such as dropouts and nonresponse, are important. Remember that confidence intervals and tests use the data you collect and ignore these errors.

Using Inference Wisely 5

Know how confidence intervals behave. All confidence intervals share these behaviors:

- The confidence level says how often the method catches the true parameter when sampling many times. We never know whether this specific data set gives us an interval that contains the true value of the parameter. All we can say is that "we got this result from a method that works 95% of the time." This data set might be one of the 5% that produce an interval that misses the true value of the parameter. If that risk is too high for you, use a 99% confidence interval.

Using Inference Wisely 6

Know how confidence intervals behave. All confidence intervals share these behaviors:

- High confidence is not free. A 99% confidence interval will be wider than a 95% confidence interval based on the same data. To be more confident, we must have more values to be confident about. There is a trade-off between how closely we can pin down the true value of the parameter (the precision of the confidence interval) and how confident we are that we have captured its true value.

Using Inference Wisely 7

Know how confidence intervals behave. All confidence intervals share these behaviors:

- Larger samples give narrower intervals. If we want high confidence and a narrow interval, we must take a larger sample. The width of our confidence interval for *p* goes down by a factor of the square root of the sample size. To cut the interval in half, we must take four times as many observations. This is typical of many types of confidence intervals.

Using Inference Wisely 8

Know what statistical significance says.

Many statistical studies hope to show that some claim is true.

The purpose of significance tests is to weigh the evidence that the data give in favor of such claims.

That is, a test helps us know if we found what we were looking for.

Using Inference Wisely 9

Know what statistical significance says.

To do this, we ask what would happen if the claim were not true.

That's the null hypothesis—no difference between the two drugs, no difference between women and men.

A significance test answers only one question: "How strong is the evidence that the null hypothesis is not true?"

A test answers this question by giving a *P*-value.

Using Inference Wisely 10

Know what statistical significance says.

The *P*-value tells us how likely data as or more extreme than ours would be if the null hypothesis were true.

Data that are very unlikely and have a small *P*-value are good evidence that the null hypothesis is not true.

This kind of indirect evidence against the null hypothesis (and for the effect we hope to find) is less straightforward than a confidence interval.

Using Inference Wisely 11

Know what your methods require.

Our significance test and confidence interval for a population proportion p require that the population size be much larger than the sample size.

They also require that the sample size itself be reasonably large so that the sampling distribution of the sample proportion $\hat{p}$ is close to Normal.

Using Inference Wisely 12

Know what your methods require.

If you plan to use statistical inference in practice, you will need help from a statistician (or need to learn lots more statistics) to manage the details.

Most of us read about statistical studies more often than we actually work with data ourselves.

Concentrate on the big issues, not on the details of whether the authors used exactly the right inference methods.

Using Inference Wisely 13

Know what your methods require.

Does the study ask the right questions?

Where did the data come from?

Do the results make sense?

Does the study report confidence intervals so you can see both the estimated values of important parameters and how uncertain the estimates are?

Does it report *P*-values to help convince you that findings are not just good luck?

The Woes of Significance Tests 1

The purpose of a significance test is usually to give evidence for the presence of some effect in the population.

The effect might be a probability of heads different from one-half for a coin or a longer mean survival time for patients given a new cancer treatment.

If the effect is large, it will show up in most samples: the proportion of heads among our tosses will be far from one-half, or the patients who get the new treatment will live much longer than those in the control group.

The Woes of Significance Tests 2

Small effects, such as a probability of heads only slightly different from one-half, will often be hidden behind the chance variation in a sample.

This is as it should be: big effects are easier to detect.

That is, the *P*-value will usually be small when the population truth is far from the null hypothesis.

A test measures only the strength of evidence against the null hypothesis. It says nothing about how big or how important the effect we seek in the population really is.

The Woes of Significance Tests 3

Pay particular attention to the size of the sample when you read the result of a significance test. Here's why:

- Larger samples make tests of significance more sensitive. If we toss a coin hundreds of thousands of times, a test of H_0: $p = 0.5$ will often give a very low *P*-value when the truth for this coin is $p = 0.502$. The test is right—it found good evidence that p really is not exactly equal to 0.5—but it has picked up a difference so small that it is of no practical interest. **A finding can be statistically significant without being practically important.**

The Woes of Significance Tests 4

- On the other hand, tests of significance based on small samples are often not sensitive. If you toss a coin only 10 times, a test of H_0: $p = 0.5$ will often give a large *P*-value even if the truth for this coin is $p = 0.7$. Again, the test is right—10 tosses are not enough to give good evidence against the null hypothesis. Lack of significance does not mean that there is no effect, only that we do not have good evidence for an effect. Small samples often miss important effects that are really present in the population. As the cosmologist Martin Rees said, "Absence of evidence is not evidence of absence."

Example: Antidepressants versus a Placebo

Through a Freedom of Information Act request, two psychologists obtained 47 studies used by the Food and Drug Administration for approval of the six antidepressants prescribed most widely between 1987 and 1999. Overall, the psychologists found that there was a statistically significant difference in the effects of antidepressants compared with a placebo, with antidepressants being more effective. However, the psychologists went on to report that antidepressant pills worked 18% better than placebos, a statistically significant difference, “but not meaningful for people in clinical settings.”

The Woes of Significance Tests 5

Whatever the truth about the population, whether p = 0.7 or p = 0.502, more observations allow us to estimate p more closely.

If p is not 0.5, more observations will give more evidence of this, that is, a smaller *P*-value.

Because statistical significance depends strongly on the sample size as well as on the truth about the population, statistical significance tells us nothing about how large or how practically important an effect is.

The Woes of Significance Tests 6

Large effects (such as $p = 0.7$ when the null hypothesis is $p = 0.5$) often give data that are insignificant if we take only a small sample.

Small effects (such as $p = 0.502$) often give data that are highly significant if we take a large enough sample.

The Woes of Significance Tests 7

Beware the naked *P*-value

The *P*-value of a significance test depends strongly on the size of the sample, as well as on the truth about the population.

It is bad practice to report a naked *P*-value (a *P*-value by itself) without also giving the sample size and a statistic or statistics that describe the sample outcome.

The Advantages of Confidence Intervals 1

Give a confidence interval

Confidence intervals are more informative than significance tests because they actually estimate a population parameter.

They are also easier to interpret.

It is good practice to give confidence intervals whenever possible.

Significance at the 5% Level Isn't Magical 1

The *P*-value of a test of significance describes the degree of evidence provided by the sample against the null hypothesis.

How small a *P*-value is convincing evidence against the null hypothesis?

Significance at the 5% Level Isn't Magical 2

This depends mainly on two circumstances:

- How plausible is H_0? If H_0 represents an assumption that the people you must convince have believed for years, strong evidence (small *P*) will be needed to persuade them.
- What are the consequences of rejecting H_0? If rejecting H_0 in favor of H_a means making an expensive changeover from one type of product packaging to another, you need strong evidence that the new packaging will boost sales.

Significance at the 5% Level Isn't Magical 3

Different people will often insist on different levels of significance.

Giving the *P*-value allows each of us to decide individually if the evidence is sufficiently strong.

But the level of significance that will satisfy us should be decided before calculating the *P*-value.

Computing the *P*-value and then deciding that we are satisfied with a level of significance that is just slightly larger than this *P*-value is an abuse of significance testing.

Significance at the 5% Level Isn't Magical 4

Users of statistics have often emphasized standard levels of significance such as 10%, 5%, and 1%.

Courts have tended to accept 5% as a standard in discrimination cases.

This emphasis reflects the time when tables of critical values rather than computer software dominated statistical practice.

Significance at the 5% Level Isn't Magical 5

The 5% level ($\alpha = 0.05$) is particularly common.

There is no sharp border between "significant" and "insignificant," only increasingly strong evidence as the *P*-value decreases.

There is no practical distinction between the *P*-values 0.049 and 0.051. It makes no sense to treat $P \leq 0.05$ as a universal rule for what is significant.

Beware of Searching for Significance 1

Statistical significance ought to mean that you have found an effect that you were looking for.

The reasoning behind statistical significance works well if you decide what effect you are seeking, design a study to search for it, and use a test of significance to weigh the evidence you get.

In other settings, significance may have little meaning. In an exploratory data analysis, you may find patterns with significant *P*-values.

Beware of Searching for Significance 2

Searching data for suggestive patterns is certainly legitimate.

Exploratory data analysis is an important part of statistics. But the reasoning of formal inference does not apply when your search for a striking effect in the data is successful.

The remedy is clear. Once you have a hypothesis, design a study to search specifically for the effect you now think is there. If the result of this study is statistically significant, you have real evidence.

Inference as Decision 1

Tests of significance were presented in Chapter 22 as methods for assessing the strength of evidence against the null hypothesis.

The assessment is made by the *P*-value, which is the probability computed under the assumption that the null hypothesis is true.

The alternative hypothesis (the statement we seek evidence for) enters the test only to help us see what outcomes count against the null hypothesis.

Inference as Decision 2

We have also seen signs of another way of thinking in Chapter 22.

A level of significance chosen in advance points to the outcome of the test as a decision.

If the *P*-value is less than ≤ α, we reject H_0 in favor of H_a. Otherwise, we fail to reject H_0.

The transition from measuring the strength of evidence to making a decision is not a small step.

Inference as Decision 3

It can be argued (and is argued by followers of Fisher) that making decisions is too grand a goal, especially in scientific inference.

A decision is reached only after the evidence of many experiments is weighed, and, indeed, the goal of research is not “decision” but a gradually evolving understanding.

It is rare (outside textbooks) to set up a level α in advance as a rule for decision making in a scientific problem.

Inference as Decision 4

More commonly, users think of significance at level 0.05 as a description of good evidence.

This is made clearer by talking about *P*-values, and this newer language is spreading.

Inference as Decision 5

Yet there are circumstances in which a decision or action is called for as the end result of inference.

Acceptance sampling is one such circumstance.

The supplier of a product (for example, potatoes to be used to make potato chips) and the consumer of the product agree that each truckload of the product shall meet certain quality standards.

When a truckload arrives, the consumer chooses a sample of the product to be inspected.

Inference as Decision 6

On the basis of the sample outcome, the consumer will either accept or reject the truckload.

Inference as decision changes the ways of reasoning used in tests of significance.

Inference as Decision 7

Tests of significance fasten attention on H_0, the null hypothesis. If a decision is called for, however, there is no reason to single out H_0.

There are simply two alternatives, and we must accept one and reject the other.

It is convenient to call the two alternatives H_0 and H_a, but H_0 no longer has the special status (the statement we try to find evidence against) that it had in tests of significance.

Inference as Decision 8

In the acceptance sampling problem, we must decide between

H_0: the truckload of product meets standards
H_a: the truckload does not meet standards

on the basis of a sample of the product. There is no reason to put the burden of proof on the consumer by accepting H_0 unless we have strong evidence against it.

Inference as Decision 9

It is equally sensible to put the burden of proof on the producer by accepting H_a unless we have strong evidence that the truckload meets standards.

Producer and consumer must agree on where to place the burden of proof, but neither H_0 nor H_a has any special status.

Inference as Decision 10

In a decision problem, we must give a decision rule: a recipe based on the sample that tells us what decision to make.

Decision rules are expressed in terms of sample statistics, usually the same statistics we would use in a test of significance.

In fact, we have already seen that a test of significance becomes a decision rule if we reject H_0 (accept H_a) when the sample statistics is statistically significant at level α and otherwise accept H_0 (reject H_a).

Inference as Decision 11

Suppose, then, that we use statistical significance at level α as our criterion for decision. And suppose that the null hypothesis H_0 is really true.

Then, sample outcomes significant at level α will occur with probability α.

But now we make a wrong decision in all such outcomes by rejecting H_0 when it is really true.

That is, significance level α now can be understood as the probability of a certain type of wrong decision.

Inference as Decision 12

Now H_a requires equal attention.

Just as rejecting H_0 (accepting H_a) when H_0 is really true is an error, so is accepting H_0 (rejecting H_a) when H_a is really true.

We can make two kinds of errors.

If we reject H_0 (accept H_a) when in fact H_0 is true, this is a Type I error.

If we accept H_0 (reject H_a) when in fact H_a is true, this is a Type II error.

Inference as Decision 13

		TRUTH ABOUT THE POPULATION	
		H_0 true	H_a true
DECISION BASED ON SAMPLE	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Inference as Decision 14

		TRUTH ABOUT THE TRUCKLOAD: Does meet standards	TRUTH ABOUT THE TRUCKLOAD: Does not meet standards
DECISION BASED ON SAMPLE	Reject the truckload	Type I error	Correct decision
	Accept the truckload	Correct decision	Type II error

Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

Inference as Decision 15

So the significance level α is the probability of a Type I error.

In acceptance sampling, this is the probability that a good truckload will be rejected.

The probability of a Type II error is the probability that a bad truckload will be accepted.

A Type I error hurts the producer, while a Type II error hurts the consumer.

Inference as Decision 16

Any decision rule is assessed in terms of the probabilities of the two types of error.

This is in keeping with the idea that statistical inference is based on probability.

We cannot (short of inspecting the whole truckload) guarantee that good lots will never be rejected and bad lots never accepted.

Inference as Decision 17

But by random sampling and the laws of probability, we can say what are the probabilities of both kinds of errors.

Because we can find out the monetary cost of accepting bad truckloads and rejecting good ones, we can determine how much loss the producer and consumer each will suffer in the long run from wrong decisions.

Inference as Decision 18

Advocates of decision theory argue that the kind of "economic" thinking natural in acceptance sampling applies to all inference problems.

Even a scientific researcher decides whether to announce results, or to do another experiment, or to give up research as unproductive.

Wrong decisions carry costs, though these costs are not always measured in dollars.

A scientist suffers by announcing a false effect, and also by failing to detect a true effect.

Inference as Decision 19

Decision theorists maintain that the scientist should try to give numerical weights (called utilities) to the consequences of the two types of wrong decision.

Then, the scientist can choose a decision rule with the error probabilities that reflect how serious the two kinds of error are.

This argument has won favor where utilities are easily expressed in money.

Inference as Decision 20

Decision theory is widely used by business in making capital investment decisions, for example.

But scientific researchers have been reluctant to take this approach to statistical inference.

Inference as Decision 21

In summary, in a test of significance, we focus on a single hypothesis (H_0) and single probability (the *P*-value).

The goal is to measure the strength of the sample evidence against H_0.

If the same inference problem is thought of as a decision problem, we focus on two hypotheses and give a rule for deciding between them based on sample evidence. Therefore, we must focus on two probabilities: the probabilities of the two types of error.

Inference as Decision 22

Such a clear distinction between the two types of thinking is helpful for understanding.

In practice, the two approaches often merge, to the dismay of partisans of one or the other.

We continued to call one of the hypotheses in a decision problem H_0.

In the common practice of testing hypotheses, we mix significance tests and decision rules as follows.

Inference as Decision 23

- Choose H_0 as in a test of significance.
- Think of the problem as a decision problem, so the probabilities of Type I and Type II errors are relevant.
- Type I errors are usually more serious. So choose an α (significance level), and consider only tests with probability of Type I error no greater than α.
- Among these tests, select one that makes the probability of a Type II error as small as possible. If this probability is too large, you will have to take a larger sample to reduce the chance of an error.

Inference as Decision 24

Testing hypotheses may seen to be a hybrid approach.

It was, historically, the effective beginning of decision-oriented ideas in statistics.

Hypothesis testing was developed by Jerzey Neyman and Egon S. Pearson in the years 1928–1938.

The decision theory approach came later (1940s) and grew out of the Neyman-Pearson ideas.

Inference as Decision 25

Because decision theory in its pure form leaves you with two error probabilities and no simple rule on how to balance them, it has been used less often than tests of significance.

Decision theory ideas have been applied in testing problems mainly by way of the Neyman-Pearson theory.

Fisher, who was exceedingly argumentative, violently attacked the Neyman-Pearson decision-oriented ideas, and the argument still continues.

Inference as Decision 26

The reasoning of statistical inference is subtle, and the principles at issue are complex.

If you feel that you do not fully grasp all of the ideas of this chapter and of Chapter 22, you are in excellent company.

Nonetheless, any user of statistics should make a serious effort to grasp the conflicting views on the nature of statistical inference.

Statistics in Summary 1

- Statistical inference is less widely applicable than exploratory analysis of data. Any inference method requires the right setting—in particular, the right design for a random sample or randomized experiment.
- Understanding the meaning of confidence levels and statistical significance helps prevent improper conclusions.
- Increasing the number of observations has a straightforward effect on confidence intervals: the interval gets shorter for the same level of confidence.

Statistics in Summary 2

- Taking more observations usually decreases the *P*-value of a test when the truth about the population stays the same, making significance tests harder to interpret than confidence intervals.

- A finding with a small *P*-value may not be practically interesting if the sample is large, and an important truth about the population may fail to be significant if the sample is small. Avoid depending on fixed significance levels such as 5% to make decisions.

Statistics in Summary 3

- If a test of significance is thought of as a decision problem, we focus on two hypotheses, H_0 and H_a, and give a decision rule for deciding between them based on sample evidence. We can make two types of errors. If we reject H_0 (accept H_a) when, in fact, H_0 is true, this is a Type I error. If we accept H_0 (reject H_a) when, in fact, H_a is true, this is a Type II error.