

DAVID FREEDMAN
ROBERT PISANI
ROGER PURVES

Statistics

Third Edition

W • W • NORTON & COMPANY

NEW YORK • LONDON



WHARTON REPROGRAPHICS

29

A Closer Look at Tests of Significance

One of the misfortunes of the law [is that] ideas become encysted in phrases and thereafter for a long time cease to provoke further analysis.

—OLIVER WENDELL HOLMES, JR. (UNITED STATES, 1841–1935)¹

1. WAS THE RESULT SIGNIFICANT?

How small does P have to get before you reject the null hypothesis? As reported in section 4 of chapter 26, many statisticians draw lines at 5% and 1%. If P is less than 5%, the result is “statistically significant,” and the “null hypothesis is rejected at the 5% level”; if P is less than 1%, the result is “highly significant.” However, the question is almost like asking how cold it has to get before you are entitled to say, “It’s cold.” A temperature of 70°F is balmy, –20°F is cold indeed, and there is no sharp dividing line.

Logically, it is the same with testing. There is no sharp dividing line between probable and improbable results. A P -value of 5.1% means just about the same thing as 4.9%. However, these two P -values can be treated quite differently, because many journals will only publish results which are “statistically significant”—the 5% line. Some of the more prestigious journals will only publish results which are “highly significant”—the 1% line.² These arbitrary lines are taken so seriously that many investigators only report their results as

“statistically significant” or “highly significant.” They don’t even bother telling you the value of P , let alone what test they used.

Investigators should summarize the data, say what test was used, and report the P -value instead of just comparing P to 5% or 1%.

Historical note. Where do the 5% and 1% lines come from? To find out, we have to look at the way statistical tables are laid out. The t -table is a good example (section 6 of chapter 26). Part of it is reproduced below as table 1.

Table 1. A short t -table.

<i>Degrees of freedom</i>	10%	5%	1%
1	3.08	6.31	31.82
2	1.89	2.92	6.96
3	1.64	2.35	4.54
4	1.53	2.13	3.75
5	1.48	2.02	3.36

How is this table used in testing? Suppose investigators are making a t -test with 3 degrees of freedom. They are using the 5% line, and want to know how big the t -statistic has to be in order to achieve “statistical significance”—a P -value below 5%. The table is laid out to make this easy. They look across the row for 3 degrees of freedom and down the column for 5%, finding the entry 2.35 in the body of the table: the area to the right of 2.35 under the curve for 3 degrees of freedom is 5%. So the result is “statistically significant” as soon as t is more than 2.35. In other words, the table gives the cutoff for “statistical significance.” Similarly, it gives the cutoff for the 1% line, or for any other significance level listed across the top.

R. A. Fisher was one of the first to publish such tables, and it seems to have been his idea to lay them out that way. There is a limited amount of room on the page. Once the number of levels was limited, 5% and 1% stood out as nice round numbers, and they soon acquired a magical life of their own. With computers everywhere, this kind of table is almost obsolete. So are the 5% and 1% levels.³

2. DATA SNOOPING

The point of testing is to help distinguish between real differences and chance variation. People sometimes jump to the conclusion that a result which is statistically significant cannot be explained as chance variation. This is false. Once in a while, the average of the draws will be 2 SEs above the average of the box, just by chance. More specifically, even if the null hypothesis is right, there is a 5% chance of getting a difference which the test will call “statistically significant.” This 5% chance could happen to you—an unlikely event, but not impossible. Similarly, on the null hypothesis, there is 1% chance to get a difference which is highly significant but just a fluke.

Put another way, an investigator who makes 100 tests can expect to get five results which are “statistically significant” and one which is “highly significant” even if the null hypothesis is right in every case—so that each difference is just due to chance. (See exercise 5 on pp. 484–485.) You cannot determine, for sure, whether a difference is real or just coincidence.

To make bad enough worse, investigators often decide which hypotheses to test only after they have seen the data. Statisticians call this *data snooping*. Investigators ought to say how many tests they ran before statistically significant differences turned up. And to cut down the chance of being fooled by “statistically significant” flukes, they ought to test their conclusions on an independent batch of data—for instance by replicating the experiment.⁴ This good advice is seldom followed.

Data-snooping makes *P*-values hard to interpret.

Example 1. Clusters. Liver cancer is a rare disease, often thought to be caused by environmental agents. The chance of having 2 or more cases in a given year in a town with 10,000 inhabitants is small—perhaps 1/2 of 1%. A cluster of liver cancer cases (several cases in a small community) prompts a search for causes, like the contamination of the water supply by synthetic chemicals.⁵

Discussion. With (say) 100 towns of this size and a 10-year time period, it is likely that several clusters will turn up, just by chance. There are $100 \times 10 = 1,000$ combinations of towns and years; and $0.005 \times 1,000 = 5$. If you keep on testing null hypotheses, sooner or later you will get significant differences.

One form of data snooping is looking to see whether your sample average is too big or too small—before you make the statistical test. To guard against this kind of data snooping, many statisticians recommend using *two-tailed* rather than

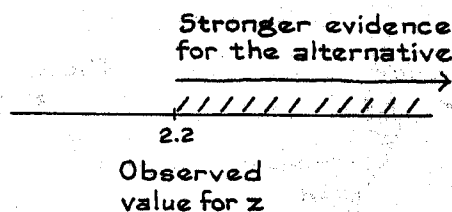
one-tailed tests. The point is easiest to see in a hypothetical example. Someone wants to test whether a coin is fair: does it land heads with probability 50%? The coin is tossed 100 times, and it lands heads on 61 of the tosses. If the coin is fair, the expected number of heads is 50, so the difference between 61 and 50 just represents chance variation. To test this null hypothesis, a box model is needed. The model consists of 100 draws from the box

$$\boxed{?? \ 0 \text{'s} \ ?? \ 1 \text{'s}} \quad 0 = \text{tails}, \ 1 = \text{heads.}$$

The fraction of 1's in this box is an unknown parameter, representing the probability of heads. The null hypothesis says that the fraction of 1's in the box is $1/2$. The test statistic is

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}} = \frac{61 - 50}{5} = 2.2$$

One investigator might formulate the alternative hypothesis that the coin is biased toward heads: in other words, that the fraction of 1's in the box is bigger than $1/2$. On this basis, large positive values of z favor the alternative hypothesis, but negative values of z do not. Therefore, values of z bigger than 2.2 favor the alternative hypothesis even more than the observed value does.

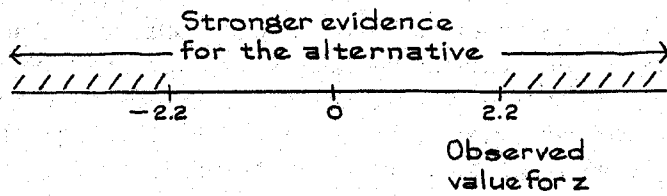


So P is figured as the area to the right of 2.2 under the normal curve:



Another investigator might formulate a different alternative hypothesis: that the probability of heads differs from 50%, in either direction. In other words, the fraction of 1's in the box differs from $1/2$, and may be bigger or smaller. On this basis, large positive values of z favor the alternative, but so do large negative values. If the number of heads is 2.2 SEs above the expected value of 50, that is bad for the null hypothesis. And if the number of heads is 2.2 SEs below the expected value, that is just as bad. The z -values more extreme than the observed 2.2 are:

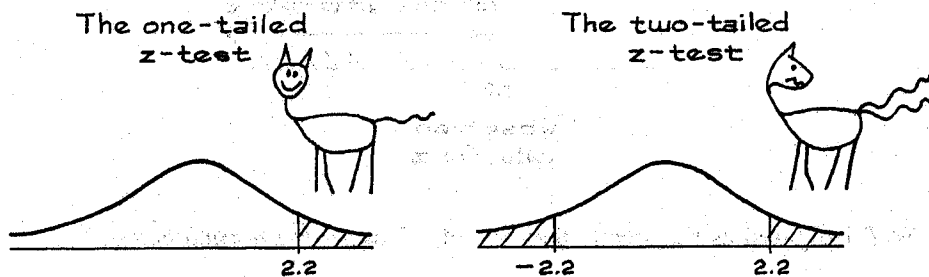
- 2.2 or more
- or
- -2.2 or less.



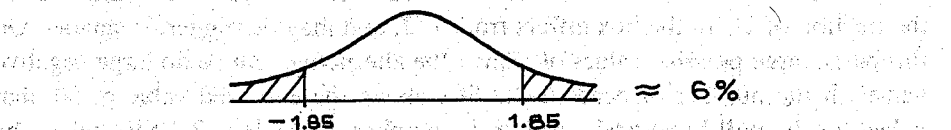
Now P is figured differently:



The first way of figuring P is the *one-tailed z-test*; the second is *two-tailed*. Which should be used? That depends on the precise form of the alternative hypothesis. It is a matter of seeing which z -values argue more strongly for the alternative hypothesis than the one computed from the data. The one-tailed test is appropriate when the alternative hypothesis says that the average of the box is bigger than a given value. The two-tailed test is appropriate when the alternative hypothesis says that the average of the box differs from the given value—bigger or smaller.



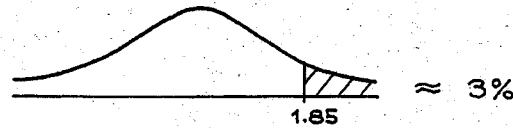
In principle, it doesn't matter very much whether investigators make one-tailed or two-tailed tests, as long as they say what they did. For instance, if they made a one-tailed test, and you think it should have been two-tailed, just double the P -value.⁶ To see why such a fuss is made over this issue, suppose a group of investigators makes a two-tailed z -test. They get $z = 1.85$, so $P \approx 6\%$.



Naturally, they want to publish. But as it stands, most journals won't touch the report—the result is not “statistically significant.”

What can they do? They could refine the experimental technique, gather more data, use sharper analytical methods. This is hard. The other possibility is

simpler: do a one-tailed test. It is the arbitrary lines at 5% and 1% which make the distinction between two-tailed and one-tailed tests loom so large.



Example 2. Cholesterol. In the 1970s, a randomized controlled double-blind experiment was performed to demonstrate the efficacy of a drug called "cholestyramine" in reducing blood cholesterol levels and preventing heart attacks. There were 3,806 subjects, who were all middle-aged men at high risk of heart attack; 1,906 were chosen at random for the treatment group and the remaining 1,900 were assigned to the control group. The subjects were followed for 7 years. The drug did reduce the cholesterol level in the treatment group (by about 8%). Furthermore, there were 155 heart attacks in the treatment group, and 187 in the control group: 8.1% versus 9.8%, $z \approx -1.8$, $P \approx 3.5\%$ (one-tailed). This was called "strong evidence" that cholesterol helps cause heart attacks.⁷

Discussion. With a two-tailed test, $P \approx 7\%$ and the difference is not significant. (The article was published in the *Journal of the American Medical Association*, whose editors are quite strict about the 5% line.) The investigators are overstating their results, and the emphasis on "statistical significance" encourages them to do so.

3. WAS THE RESULT IMPORTANT? (P. 554)

If a difference is statistically significant, then it is hard to explain away as a chance variation. But in this technical phrase, "significant" does not mean "important." Statistical significance and practical significance are two different

ideas.¹¹ The point is easiest to understand in the context of a hypothetical example (based on exercise 3, p. 509). Suppose that investigators want to compare WISC vocabulary scores for big-city and rural children, age 6 to 9. They take a simple random sample of 2,500 big-city children, and an independent simple random sample of 2,500 rural children. The big-city children average 26 on the test, and their SD is 10 points; the rural children only average 25, with the same SD of 10 points. What does this one-point difference mean? To find out, the investigators make a two-sample z -test. The SE for the difference can be estimated as 0.3, so

$$z \approx 1/0.3 \approx 3.3, \quad P \approx 5/10,000.$$

The difference between big-city children and rural children is highly significant, rural children are lagging behind in the development of language skills, and the investigators launch a crusade to pour money into rural schools.

The commonsense reaction must be: slow down. The z -test is only telling us that the one-point difference between the sample averages is almost impossible to explain as a chance variation. To focus the issue, suppose that the samples are a perfect image of the population, so that all the big-city children in the U.S. (not just the ones in the sample) would average 26 points on the WISC vocabulary scale, while the average for all the rural children in the U.S. would be 25 points. Then what? There is no more chance variation to worry about, so a test of significance cannot help. All the facts are in, and the problem is to find out what the difference means.

To do that, it is necessary to look at the WISC vocabulary scale itself. There are forty words which the child has to define. Two points are given for a correct definition, and one point for a partially correct definition. So the one-point difference between big-city and rural children only amounts to a partial understanding of one word out of forty. This is not a solid basis for a crusade. Quite the opposite: the investigators have proved there is almost no difference between big-city and rural children on the WISC vocabulary scale.¹²

Of course, the sample does not reflect the population perfectly, so a standard error should be attached to the estimate for the difference. Based on the two samples of 2,500 children, the difference in average scores between all the big-city and rural children in the U.S. would be estimated as 1 point, give or take 0.3 points or so. (The z -statistic is 3.3 because the difference has been estimated so accurately.)

A big sample is good because it enables the investigators to measure differences quite accurately—with small SEs. But the z -test compares a difference to its SE. Therefore, with a large sample even a small difference can lead to an impressive value for z . The z -test can be too sensitive for its own good.

The P -value of a test depends on the sample size. With a large sample, even a small difference can be “statistically significant,” that is, hard to explain by the luck of the draw. This doesn’t necessarily make it important. Conversely, an important difference may not be statistically significant if the sample is too small.

Example 3. As reported in section 2 of chapter 27, mathematics test scores increased from 300.4 in 1978 to 306.7 in 1992. These were averages based on nationwide samples; $z \approx 4.2$ and $P \approx 1/100,000$ (one-tailed). Does the increase matter?

Solution. The P -value says that the increase is hard to explain away as chance error. The P -value does not say whether the increase matters. More detailed analysis of the data suggests that each extra year of schooling is associated with about a 6-point increase in average test scores.¹³ On this basis, a 6-point increase is quite impressive. So is the reversal of half a century of declining scores.

4. THE ROLE OF THE MODEL

(p. 557)

To review briefly, a test of significance answers the question, "Is the difference due to chance?" But the test can't do its job until the word "chance" has been given a precise definition. That is where the box model comes in.¹⁸

To make sense out of a test of significance, a box model is needed.

This idea may be a little surprising, because the arithmetic of the test does not use the box model. Instead, the test seems to generate the chances directly from the data. However, that is an illusion. It is the box model which defines the chances. The formulas for the expected values and standard errors make a

tacit assumption: that the data are like draws from a box. So do the statistical tables—normal, t , and χ^2 . If the box model is wrong, the formulas and the tables do not apply, and may give silly results. This section discusses some examples.

Example 4. Census data show that in 1970, there were 203 million people in the U.S., of whom 9.8% were 65 or older. In 1990, there were 249 million people, of whom 12.5% were 65 or older.¹⁹ Is the difference in the percentages statistically significant?

Discussion. The arithmetic of a two-sample z -test is easy enough to do (especially with a calculator), but the result is close to meaningless. We have census data on the whole population. There is no sampling variability to worry about. Census data are subject to many small errors, but these are not like draws from a box. The aging of the population is real, and important. However, the concept of statistical significance does not apply.

If a test of significance is based on data for the whole population, watch out.

Example 5. Graduate Division records at the University of California, Berkeley, can be used to compare admission rates for men and women. For one year and one graduate major, this came out as follows: 825 men applied, and 61.7% were admitted; 108 women applied, and 82.4% were admitted.²⁰ Is the difference between admission rates for men and women statistically significant?

Discussion. Again, there is nothing to stop you from doing a two-sample z -test. However, to make sense out of the results, a box model would be needed, and there doesn't seem to be one in the neighborhood. Chance may enter in recruiting the applicants, or in deciding which applicants to admit. However, it is almost impossible to identify the pool of potential applicants; and even if you could, the actual applicants were not drawn from this pool by any probability method. Nor do departments admit candidates by drawing names from a hat (although that might not be such a bad idea). The concept of statistical significance does not apply.

Statisticians distinguish between samples drawn by probability methods and samples of convenience (section 4 of chapter 23). A sample of convenience consists of whoever is handy—students in a freshman psychology class, the first hundred people you bump into, or all the applicants to a given department in a given year. With a sample of convenience, the concept of chance becomes quite slippery, the phrase “the difference is due to chance” is hard to interpret, and so are P -values. Example 5 was based on a sample of convenience.²¹

If a test of significance is based on a sample of convenience, watch out.

Example 6. Academic gains were made by minority children in the Head-start preschool program, but tended to evaporate when the children went on to regular schools. As a result, Congress established Project Follow Through to provide continued support for minority children in regular schools. Seven sponsors were given contracts to run project classrooms according to different educational philosophies, and certain other classrooms were used as controls. The Stanford Research Institute (SRI) was hired to evaluate the project, for the Department of Health, Education, and Welfare.²² One important question was whether the project classrooms really were different from the control classrooms. (Of course, SRI also studied the effect of the programs on the children.)

To see whether or not there were real differences, SRI devised an implementation score to compare project classrooms with control classrooms. This score involved observing the classrooms to determine, for instance, the amount of time children spent playing, working independently, asking questions of the teacher, and so on. The results for one sponsor, Far West Laboratory, are shown in table 2.

Table 2. SRI Implementation Scores for 20 Far West Laboratory classrooms. Scores are between 0 and 100.

Site	Classroom scores			
Berkeley	73	79	76	72
Duluth	76	84	81	80
Lebanon	82	76	84	81
Salt Lake City	81	86	76	80
Tacoma	78	72	78	71

The average of these 20 scores is about 78; their SD is about 4.2. The average score for the control classrooms was about 60, so the difference is 18 points. As far as the SRI implementation score is concerned, the Far West classrooms are very different from the control classrooms. So far, so good. However, SRI was not satisfied. They wished to make a z-test,

to test whether the average implementation score for Follow Through was significantly greater than the average for Non-Follow Through.

The computation is as follows.²³ The SE for the sum of the scores is estimated as $\sqrt{20} \times 4.2 \approx 19$. The SE for their average is $19/20 \approx 1$ and $z \approx (78 - 60)/1 = 18$. Now



The inference is:

the overall Far West classroom average is significantly different from the Non-Follow Through classroom average of 60.

* *Discussion.* The arithmetic is all in order, and the procedure may seem reasonable at first. But there is a real problem, because SRI did not have a chance model for the data. And it is hard to invent a plausible one. SRI might be thinking of the 20 treatment classrooms as a sample from the population of all classrooms. But they didn't choose their 20 classrooms by simple random sampling, or even by some more complicated probability method. In fact, no clear procedure for choosing the classrooms was described in the report. It was a sample of convenience, pure and simple.

SRI might be thinking of measurement error. Is there some "exact value" for Far West, which may or may not be different from the one for controls? If so, is this a single number? Or does it depend on the site? on the classroom? the teacher? the students? the year? Or are these part of the error box? If so, isn't the error box different from classroom to classroom, or site to site? Aren't the errors dependent?

The report covers 500 pages, and there isn't a single one which touches on these problems. It is taken as self-evident that a test of significance can be used to compare the average of any sample, no matter where it comes from, with an external standard. The whole argument to show that the project classrooms differ from the controls rests on these tests, and the tests rest on nothing. SRI does not have a simple random sample of size 20, or 20 repeated measurements on the same quantity. It has 20 numbers. These numbers have chance components, but almost nothing is understood about the mechanism which generated them. Under these conditions, a test of significance is an act of intellectual desperation.

We went down to SRI to discuss these issues with the investigators. They insisted that they had taken very good statistical advice when designing their study, and were only doing what everybody else did. We pressed our arguments. The discussion went on for several hours. Eventually, the senior investigator said:

Look. When we designed this study, one of our consultants explained that some day, someone would arrive out of the blue and say that none of our statistics made any sense. So you see, everything was very carefully considered.

5. DOES THE DIFFERENCE PROVE THE POINT?

Usually, an investigator collects data to prove a point. If the results can be explained by chance, the data may not prove anything. So the investigator makes a test of significance to show that the difference was real. However, the test has to be told what "chance" means. That is what the box model does, and if the investigator gets the box model wrong, the results of the test may be quite misleading. Section 4 made this point, and the discussion continues here.

For example, take an ESP experiment in which a die is rolled, and the subject tries to make it land showing six spots.³⁰ This is repeated 720 times, and the die lands six in 143 of these trials. If the die is fair, and the subject's efforts have no effect, the die has 1 chance in 6 to land six. So in 720 trials, the expected number of sixes is 120. There is a surplus of $143 - 120 = 23$ sixes.

Is this difference real, or a chance variation? This is where a test of significance comes in. The null hypothesis can be formulated in terms of a box model: the number of sixes is like the sum of 720 draws from the box $\boxed{0} \boxed{0} \boxed{0} \boxed{0} \boxed{0} \boxed{1}$. The SE for the sum of the draws is

$$\sqrt{720} \times \sqrt{1/6 \times 5/6} = 10.$$

So $z = (143 - 120)/10 = 2.3$, and $P \approx 1\%$. The difference looks real.

Now for the main question: Does the difference prove that ESP exists? As it turned out, in another part of the experiment the subject tried to make the die land showing aces, and got too many sixes. In fact, whatever number the subject tried for, there were too many sixes. The test proves that the die was biased, not that ESP exists.

A test of significance can only tell you that a difference is there. It cannot tell you the cause of the difference. The difference could be there because the investigator got the box model wrong, or made some other mistake in designing the study.

A test of significance does not check the design of the study.

In the ESP experiment, why did the z -test lead us astray? It didn't. The test was asked whether there were too many sixes to explain by chance. And it answered, correctly, that there were. But the test was told what "chance" meant: rolling a fair die. This assumption was used to compute the expected value and SE in the formula for z . Tests of significance have to be told what chances to use. If the investigator gets the box model wrong, as in the ESP example, do not blame the test.

Example 7. Tart's experiment on ESP was discussed in section 5 of chapter 26. A machine called the "Aquarius" picked one of 4 targets at random, and subjects tried to guess which one. The subjects scored 2,006 correct guesses in 7,500 tries, compared to the chance level of $1/4 \times 7,500 = 1,875$. The difference was $2,006 - 1,875 = 131$, $z \approx 3.5$, and $P \approx 2/10,000$ (one-tailed). What does this prove?

Discussion. The difference is hard to explain as a chance variation. That is what the z -test shows. But was it ESP? To rule out other explanations, we have to look at the design of the study. Eventually, statisticians got around to checking Tart's random number generators. These generators had a flaw: they seldom picked the same target twice in a row. In the experiment, the Aquarius lit up the target after each guess. Subjects who noticed the pattern, or picked new targets each time for some other reason, may have improved their chances due to the flaw in the random number generator. Tart's box model—which defined the chances for the test—did not correspond to what the random number generator was really doing.

Tart began by denying that the non-randomness in the numbers made any difference. Eventually, he replicated the experiment. He used better random number generators, and tightened up the design in other ways too. In the replication, subjects guessed at about the chance level: there was no ESP. The subjects in both experiments were students at the University of California, Davis. Tart's main explanation for the failure to replicate—"a dramatic change" in student attitudes between experiments.³¹

In the last year or two, students have become more serious, competitive and achievement-oriented than they were at the time of the first experiment. Such "uptight" attitudes are less compatible with strong interest and motivation to explore and develop a "useless" talent such as ESP. Indeed, we noticed that quite a few of our participants in the present experiment did not seem to really "get into" the experiment and were anxious to "get it over with."

6. CONCLUSION

When a client is going to be cross-examined, lawyers often give the following advice:

Listen to the question, and answer the question. Don't answer the question they should have asked, or the one you wanted them to ask. Just answer the question they really asked.

Tests of significance follow a completely different strategy. Whatever you ask, they answer one and only one question:

How easy is it to explain the difference between the data and what is expected on the null hypothesis, on the basis of chance variation alone?

Chance variation is defined by a box model. This model is specified (explicitly or implicitly) by the investigator. The test will not check to see whether this model is relevant or plausible. The test will not measure the size of a difference, or its importance. And it will not identify the cause of the difference. So a test can answer only one very specific question. Often that's the wrong question to ask. Then the problem should not be resolved by testing, but by estimation. This involves making a chance model for the data, defining what you want to estimate in terms of the model, estimating it from the data, and attaching a standard error to the estimate.

Nowadays, tests of significance are extremely popular. One reason is that the tests are part of an impressive and well-developed mathematical theory. Another reason is that many investigators just cannot be bothered to set up chance models. The language of testing makes it easy to bypass the model, and talk about "statistically significant" results. This sounds so impressive, and there is so much mathematical machinery clanking around in the background, that

tests seem truly scientific—even when they are complete nonsense. St. Exupéry understood this kind of problem very well:

When a mystery is too overpowering, one dare not disobey.

—*The Little Prince*³²