26

Tests of Significance

Who would not say that the glosses [commentaries on the law] increase doubt and ignorance? It is more of a business to interpret the interpretations than to interpret the things.

—MICHEL DE MONTAIGNE (FRANCE, 1533–1592)[1]

1. INTRODUCTION

Was it due to chance, or something else? Statisticians have invented *tests of significance* to deal with this sort of question. Nowadays, it is almost impossible to read a research article without running across tests and significance levels. Therefore, it is a good idea to find out what they mean. The object in chapters 26 through 28 is to explain the ideas behind tests of significance, and the language. Some of the limitations will be pointed out in chapter 29. This section presents an example.

Suppose that a senator introduces a bill to simplify the tax code. The senator claims that his bill is revenue-neutral: on balance, tax revenues will stay the same. *Microsimulation models* are used to evaluate the impact of such bills, and tests of significance come in. Although the details are complicated, the idea is simple.

To evaluate the senator's claim, the Treasury Department will use a computer file of 100,000 representative tax returns. Each return shows the total tax payable under the old rules. From the detailed information on the form, the Treasury can compute the tax under the new rules and then look at the change:

$$\text{change} = \text{tax under new rules} - \text{tax under old rules.}$$

Signs matter. A plus-sign means that with the new rules, the Treasury would collect more from the taxpayer; a minus-sign, they would collect less. The senator thinks that on average, the pluses and minuses will balance.

In our (partly hypothetical) example, the work has been done on a pilot sample of 100 forms chosen at random from the file.[2] The sample average came out to −$219, so it looks like the new rules would cost the Treasury some money. But the standard deviation was quite large, at $725. A congressional aide is discussing these results with a Treasury official.

Aide. First off, I don't believe the SD. How can it be that much bigger than the average?

Treasury. Well, tax returns are all over the place. Some people pay nothing. That's about 20% of the returns, right there. Some pay a few thousand dollars. And some pay a few hundred thousand. The numbers have a really long tail.[3] The new rules make big changes for some taxpayers and have no impact on others. Believe me, we went over the program with a fine-tooth comb. It's right.

Aide. Now I suppose you're going to tell me that our proposal isn't revenue-neutral after all.

Treasury. If the bill passes, we lose around $200 per return. You may not think that's serious, but there are maybe 100 million returns. So we're talking about $20 billion. Or more.

Aide. Wait a minute. You only did this with a sample of 100 forms, right?

Treasury. Right.

Aide. And you keep telling me that the SD was $725. So this $219 is only a fraction of an SD. That's chance variation if I ever saw it.

Treasury. No, no. We need the SE, not the SD. To compute the SE, we should have a box model. The box has 100,000 tickets, one for each return in the file: the number on the ticket shows the change for that taxpayer. We drew 100 tickets at random for our sample. The data are like the 100 draws.

Aide. OK. So what does that tell us?

Treasury. What we're arguing about is the average of the 100,000 tickets in the box. You say that's $0. We say it's negative.

Aide. And we say that by the luck of the draw, you got too many big negative numbers in your sample. That's why the average of your sample is so negative.

Treasury. Well, that's where the SE comes in. To compute the SE exactly, we would need the SD of the box.

Aide. But you don't know that.

ıe Treasury would
l collect less. The
alance.
ɜn done on a pilot
mple average came
the Treasury some
. congressional aide

ıt much bigger than

ɪeople pay nothing.
pay a few thousand
The numbers have a
for some taxpayers
nt over the program

ɔposal isn't revenue-

. You may not think
rns. So we're talking

ıf 100 forms, right?

So this $219 is only
ɜver saw it.
te the SE, we should
ɔne for each return in
ıge for that taxpayer
he data are like these

100,000 tickets in the

too many big negative
ɜ of your sample is so

te the SE exactly, we

Treasury. Right. So we use the SD of the data to estimate the SD of the box.

Aide. That seems reasonable. Where do we go from there?

Treasury. The SE for the sum of 100 draws from the box is $\sqrt{100} \times \$725 = \$7{,}250$. That's the likely size of the chance error in the sum across all 100 sample forms. We're looking at the average, so we divide through by 100, and we get $\$7{,}250/100 \approx \$72$.

Aide. So?

Treasury. Well, suppose for a moment that you're right and the average of the box is really $0. Then you have to expect the average of the sample to be around $0. But we got −$219. That's 3 SEs below your expected value:

$$\frac{-\$219 - \$0}{\$72} \approx -3.$$

Aide. Hmmm.

Treasury. We have enough draws here so that we can use the normal approximation. The area to the left of −3 under the normal curve is about 0.1 of 1%. You're talking about 1 chance in 1,000.

Aide. Maybe, but where does the normal curve come in? The histogram for the data sure doesn't look normal.

Treasury. Right, but we're using the normal curve on the probability histogram for the average of the draws.

Aide. OK, I see what you're doing now.

Treasury. You can insist that the average of the box is $0, like the senator wants it to be. Or you can agree with us that it's negative. But if you stick with $0, you need a small miracle to explain the data—the sample average only has 1 chance in 1,000 to be that far below $0.

Aide. Maybe I give up for now. What do you think the impact of the new rules would be?

Treasury. We think the new rules would make us lose about $200 per return. That may not be a huge difference, but it's real. I mean, you can't just dismiss the sample average as chance variation.

Our first pass at testing is now complete. The issue in the dialog comes up over and over again: one side thinks a difference is real but a skeptic might say it's all chance variation. The skeptic can be fended off by a chance calculation, as in the dialog. This calculation is called a *test of significance*. The key idea: if an observed value is too many SEs away from its expected value, that is hard to explain by chance. Statisticians use rather technical language when making this sort of argument, and the next couple of sections will introduce the main terms: *null hypothesis, alternative hypothesis, test statistic*, and *P-value*.[4]

Exercise Set A

1. Fill in the blanks. In the example of this section:
   (a) The Treasury model had ______ tickets in the box, and ______ draws were made. Options:
   100   1,000   10,000   100,000
   (b) The SD of the box was ______ $725. Options:
   known to be   estimated from the data as
   (c) The −$219 is an ______ value. Options:
   observed   expected

2. In the example of this section:
   (a) Mr. Jones would pay $3,292 under the new rules and $3,117 under the old rules. His ticket would be marked ______. Fill in the blank, and explain briefly.
   (b) Ms. Smith's ticket is marked −$753. Is she better off under the new rules or the old? Explain briefly.

3. In the dialog, suppose tax collections for the 100 sample forms average $5,182 under the new rules and $5,217 under the old rules, but the SD of the 100 differences is still $725. Who wins now, the Treasury official or the congressional aide?

4. A die is rolled 100 times. The total number of spots is 368 instead of the expected 350. Can this be explained as a chance variation, or is the die loaded?

5. A die is rolled 1,000 times. The total number of spots is 3,680 instead of the expected 3,500. Can this be explained as a chance variation, or is the die loaded?

The answers to these exercises are on p. A91.

2. THE NULL AND THE ALTERNATIVE

In the example of the previous section, there was sample data for 100 taxpayers. Both sides saw the sample average of −$219. In statistical shorthand, the −$219 was "observed." The argument was about the interpretation: what do the sample forms tell us about the 100,000 forms in the whole file? The Treasury official claimed that the observed difference was "real." That may sound odd; of course −$219 is different from $0. But the question was whether the difference just reflected chance variation (as the aide said) or whether the new tax rules made a real difference—for all 100,000 forms in the file.

In order to convince the aide, the Treasury official set up a box model for the problem. The null hypothesis and the alternative hypothesis are statements about the box. Each hypothesis represents one side of the argument.

- Null hypothesis—the average of the box equals $0.
- Alternative hypothesis—the average of the box is less than $0.

In the dialog, the congressional aide is defending the null hypothesis. According to him, the sample average has an expected value of $0; the observed value turned out to be −$219 just by the luck of the draw. The Treasury official was arguing for the alternative hypothesis; she thinks the average of the box is negative. Her argument in a nutshell: the sample average is so far below $0 that the Congressional aide almost has to be wrong. Both sides agreed that the data were like 100 draws from a box. They disagreed about the average of the box.

> The null hypothesis expresses the idea that an observed difference is due to chance. To make a test of significance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box; it says that the difference is real.

The terminology may be unsettling; the "alternative hypothesis" is often what someone sets out to prove. The "null hypothesis" is then an alternative (and dull) explanation for the findings, in terms of chance variation. However, there is no help for it; the names are completely standard.

Every legitimate test of significance involves a box model. The test gets at the question of whether an observed difference is real, or just chance variation. A real difference is one that says something about the box, and isn't just a fluke of sampling. In the dialog, the argument was about the 100,000 numbers in the box, not the 100 numbers in the sample. A test of significance only makes sense in a debate about the box. This point will be discussed again in chapter 29, section 4.

Exercise Set B

1. In order to test a null hypothesis, you need
   (i) data
   (ii) a box model for the data
   (iii) both of the above
   (iv) none of the above

2. The ______ hypothesis says that the sample difference is just due to chance; the ______ hypothesis says that the sample difference points to a real difference. Fill in the blanks. Options: null, alternative.

3. In the dialog of section 1, the Treasury official needed to make a test of significance because:
   (i) she knew what was in the box but didn't know how the data were going to turn out, or
   (ii) she knew how the data had turned out but didn't know what was in the box.

   Choose one option, and explain briefly.

4. In the dialog, the null hypothesis says that the average of the ______ is $0. Options: sample, box.

5. One hundred draws are made at random with replacement from a box. The average of the draws is 102.7, and their SD is 10. Someone claims that the average of the box equals 100. Is this plausible? What if the average of the draws is 101.1?

The answers to these exercises are on p. A91.

3. TEST STATISTICS AND SIGNIFICANCE LEVELS

In the dialog of section 1, the Treasury official made a box model for the data. For the sake of argument, she temporarily assumed the null hypothesis to be right (the average of the box is $0). On this basis, she calculated how many SEs away the observed value of the sample average was from its expected value:

$$\frac{-\$219 - \$0}{\$72} \approx -3.$$

This is an example of a *test statistic.*

> A test statistic is used to measure the difference between the data and what is expected on the null hypothesis.

The Treasury official's test statistic is usually called z:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

Tests using the z-statistic are called *z-tests*. Keep the interpretation in mind.

> z says how many SEs away an observed value is from its expected value, where the expected value is calculated using the null hypothesis.

It is the null hypothesis which told the Treasury official to use $0 as the benchmark, and not some other number, in the numerator of z. That is the exact point where the null hypothesis comes into the procedure. Other null hypotheses will give different benchmarks in the numerator of z. The null hypothesis did not tell us the SD of the box: that had to be estimated from the data, in order to compute the SE in the denominator of z.

The z-statistic of -3 stopped the aide in his tracks. Why was it so intimidating? After all, 3 is not a very big number. The answer, of course, is that the area to the left of -3 under the normal curve is ridiculously small. The chance of getting a sample average 3 SEs or more below its expected value is about 1 in 1,000.

ɔm a box. The average
hat the average of the
e draws is 101.1?

a box model for the
the null hypothesis to
calculated how many
ɔm its expected value:

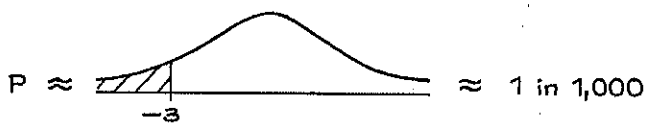
ɜtween the data

ɪrpretation in mind,

is from its ex-
ılated using the

to use $0 as the
:. That is the exact
her null hypothes
ll hypothesis did
data, in order to

s. Why was it
ver, of course,
lously small.
expected value



(From the table, the area is 0.135 of 1%; rounding off, we get 0.1 of 1%; this is 0.1 of 0.01 = 0.001 = 1/1,000.)

The chance of 1 in 1,000 overwhelmed the aide, and forced him to concede that the new rules would reduce the tax collections—not just for the sample but for all the forms in the file. This chance of 1 in 1,000 is called an *observed significance level*. The observed significance level is often denoted *P*, for probability, and referred to as a *P-value*. In the example, the *P*-value of the test was about 1 in 1,000.

Why look at the area to the left of −3? The first point to notice: the data could have turned out differently, and then z would have been different too. For instance, if the sample average is −$239 and the SD is $590,

$$z = \frac{-\$239 - \$0}{\$59} \approx -4.1$$

This is stronger evidence against the null hypothesis: 4.1 SEs below $0 is even worse for the aide than 3 SEs. On the other hand, if the sample average is −$162 and the SD is $630,

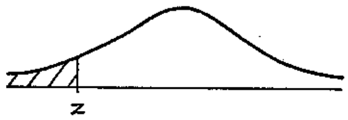
$$z = \frac{-\$162 - \$0}{\$63} \approx -2.6$$

This is weaker evidence. The area to the left of −3 represents the samples which give even more extreme z-values than the observed one, and stronger evidence against the null hypothesis.

> The observed significance level is the chance of getting a test statistic as extreme as, or more extreme than, the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller this chance is, the stronger the evidence against the null.

The z-test can be summarized as follows:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}, \quad P \approx$$



Since the test statistic z depends on the data, so does P. That is why P is called an "observed" significance level.

At this point, the logic of the z-test can be seen more clearly. It is an argument by contradiction, designed to show that the null hypothesis will lead to an absurd conclusion and must therefore be rejected. You look at the data, compute the test statistic, and get the observed significance level. Take, for instance, a P of 1 in 1,000. To interpret this number, you start by assuming that the null hypothesis is right. Next, you imagine many other investigators repeating the experiment. What the 1 in 1,000 says is that your test statistic is really far out: only one investigator in a thousand would get a test statistic as extreme as, or more extreme than, the one you got. The null hypothesis is creating absurdities, and should be rejected. In general, the smaller the observed significance level, the more you want to reject the null. The phrase "reject the null" emphasizes the point that with a test of significance, the argument is by contradiction.

Our interpretation of P may seem convoluted. It is convoluted. Unfortunately, simpler interpretations turn out to be wrong. If there were any justice in the world, P would be the probability of the null hypothesis given the data. However, that is wrong: indeed, P is computed using the null. Even worse, according to the frequency theory, there is no way to define the probability of the null hypothesis being right. The null is a statement about the box. No matter how often you do the draws, the null hypothesis is either always right or always wrong, because the box does not change.[5] (A similar point for confidence intervals is discussed in section 3 of chapter 21.) What the observed significance level gives is the chance of getting evidence against the null as strong as the evidence at hand—or stronger—if the null is true.

> The P-value of a test is the chance of getting a big test statistic—assuming the null hypothesis to be right. P is not the chance of the null hypothesis being right.

The z-test is used for reasonably large samples, when the normal approximation can be used on the probability histogram for the average of the draws. (The average has already been converted to standard units, by z.) With small samples, other techniques must be used, as discussed in section 6 below,

Exercise Set C

1. (a) Other things being equal, which of the following P-values is best for the null hypothesis? Explain briefly.

   0.1 of 1%   3%   17%   32%

   (b) Repeat, for the alternative hypothesis.

2. According to one investigator's model, the data are like 50 draws made at random from a large box. The null hypothesis says that the average of the box equals 100; the alternative says that the average of the box is more than 100. The average of the draws is 107.3, the SD is 22.1, and the SE for the sample average is 3.1. So

$$z = (107.3 - 100)/3.1 = 2.35 \text{ and } P = 1\%.$$

True or false, and explain:

(a) If the null hypothesis is right, there is only a 1% chance of getting a z bigger than 2.35.

(b) The probability of the null hypothesis given the data is 1%.

3. True or false, and explain:

(a) The observed significance level depends on the data.

(b) If the observed significance level is 5%, there are 95 chances in 100 for the alternative hypothesis to be right.

4. According to one investigator's model, the data are like 400 draws made at random from a large box. The null hypothesis says that the average of the box equals 50; the alternative says that the average of the box is more than 50. In fact, the data averaged out to 52.7, and the SD was 25. Compute z and P. What do you conclude?

5. In the previous exercise, the null hypothesis says that the average of the ______ is 50. Fill in the blank, using one of the options below, and explain briefly.

box   sample

6. In the dialog of section 1, suppose the Treasury official has only taken a sample of 10 forms. Should she use the normal curve to compute P? Answer yes or no, and explain briefly.

7. Many companies are experimenting with "flex-time," allowing employees to choose their schedules within broad limits set by management.[6] Among other things, flex-time is supposed to reduce absenteeism. One firm knows that in the past few years, employees have averaged 6.3 days off from work (apart from vacations). This year, the firm introduces flex-time. Management chooses a simple random sample of 100 employees to follow in detail, and at the end of the year, these employees average 5.5 days off from work, and the SD is 2.9 days. Does this mean that flex-time reduced absenteeism? Or is this a chance variation?

8. Repeat exercise 7 for a sample average of 5.9 days and an SD of 2.9 days.

The answers to these exercises are on pp. A91–92.

4. MAKING A TEST OF SIGNIFICANCE

Making a test of significance is a complicated job. You have to

- set up the null hypothesis, in terms of a box model for the data;
- pick a test statistic, to measure the difference between the data and what is expected on the null hypothesis;
- compute the observed significance level P.

The choice of test statistic depends on the model and the hypothesis being considered. The test discussed so far is the "one-sample z-test," which is based on the z-statistic. (Two-sample z-tests will be covered in chapter 27.) There are also "t-tests" based on the t-statistic (section 6 below), "χ^2-tests" based on the χ^2-statistic (chapter 28), and many others not even mentioned in this book. However, all tests follow the steps outlined above, and their P-values can be estimated in the same way.

It is natural to ask how small the observed significance level has to be before an investigator should reject the null hypothesis. Many statisticians draw the line at 5%.

- If P is less than 5%, the result is called *statistically significant.*

There is another line at 1%.

- If P is less than 1%, the result is called *highly significant.*

These somewhat arbitrary lines will be discussed again in section 1 of chapter 29.

Do not let the jargon distract you from the main idea. When the data are too far from the predictions of a theory, that means trouble. In statistics, the null hypothesis is rejected when the observed value is too many SEs away from the expected value.

Exercise Set D

1. True or false:
   (a) A "highly significant" result cannot possibly be due to chance.
   (b) If a difference is "highly significant," there is less than a 1% chance for the null hypothesis to be right.
   (c) If a difference is "highly significant," there is better than a 99% chance for the alternative hypothesis to be right.

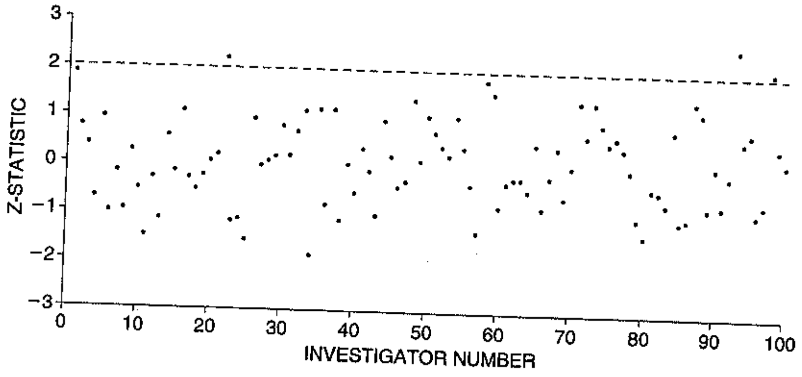
2. True or false:
   (a) If P is 43%, the null hypothesis looks plausible.
   (b) If P is 0.43 of 1%, the null hypothesis looks implausible.

3. True or false:
   (a) If the observed significance level is 4%, then the result is "statistically significant."
   (b) If the P-value of a test is 1.1%, the result is "highly significant."
   (c) If a difference is "highly significant," then P is less than 1%.
   (d) If the observed significance level is 3.6%, then $P = 3.6\%$.
   (e) If $z = 2.3$, then the observed value is 2.3 SEs above what is expected on the null hypothesis.

4. An investigator draws 250 tickets at random with replacement from a box. What is the chance that the average of the draws will be more than 2 SEs above the average of the box?

5. One hundred investigators set out to test the null hypothesis that the average of the numbers in a certain box equals 50. Each investigator takes 250 tickets at random with replacement, computes the average of the draws, and does a z-test. The results are plotted in the diagram on the next page: investigator #1 got a z-statistic of 1.9, this is plotted as the point (1, 1.9); investigator #2 got a z-statistic of 0.8, this is plotted as (2, 0.8); and so forth. Unknown to the investigators, the null hypothesis happens to be true.
   (a) True or false, and explain: The z-statistic is positive when the average of the draws is more than 50.

(b) How many investigators should get a positive z-statistic?
(c) How many of them should get a z-statistic bigger than 2? How many of them actually do?
(d) If $z = 2$, what is P?



The answers to these exercises are on p. A92.

5. ZERO-ONE BOXES

The z-test can also be used when the situation involves classifying and counting. It is a matter of putting 0's and 1's in the box (section 5 of chapter 17). This section will give an example. Charles Tart ran an experiment at the University of California, Davis, to demonstrate ESP.[7] Tart used a machine called the "Aquarius." The Aquarius has an electronic random number generator, and 4 "targets." Using its random number generator, the machine picks one of the 4 targets at random; it does not indicate which. Then, the subject guesses which target was chosen, by pushing a button. Finally, the machine lights up the target it picked, ringing a bell if the subject guessed right. The machine keeps track of the number of trials and the number of correct guesses.

Tart selected 15 subjects who were thought to be clairvoyant. Each of the subjects made 500 guesses on the Aquarius, for a total of 15 × 500 = 7,500 guesses. Out of this total, 2,006 were right. Of course, even if the subjects had no clairvoyant abilities whatsoever, they would still be right about 1/4 of the time. In other words, about 1/4 × 7,500 = 1,875 correct guesses are expected, just by chance. True, there is a surplus of 2,006 − 1,875 = 131 correct guesses, but can't this be explained as a chance variation?

Tart could—and did—fend off that explanation by making a test of significance. To set up a box model for the test, he assumed that the Aquarius generates numbers at random, so each of the 4 targets has 1 chance in 4 to be chosen. And he assumed (temporarily) that there is no ESP: a guess has 1 chance in 4 to be right.

The data consist of a record of the 7,500 guesses, showing whether each one is right or wrong. The null hypothesis says that the data are like 7,500 draws from the box

$$\boxed{\boxed{1}\ \boxed{0}\ \boxed{0}\ \boxed{0}} \qquad 1 = \text{right}, \quad 0 = \text{wrong}$$

The number of correct guesses is like the sum of 7,500 draws from the box. This completes the box model for the null hypothesis.

The machine is classifying each guess as right or wrong, and counting the number of correct guesses. That is why a zero-one box is needed. Once the null hypothesis has been translated into a box model, the z-test can be used:

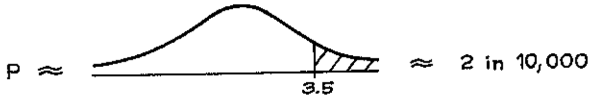
$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

The "observed" is 2,006, the number of correct guesses. The expected number of correct guesses comes from the null hypothesis, and is 1,875. The numerator of the z-statistic is 2,006 − 1,875 = 131, the surplus number of correct guesses.

Now for the denominator. You need the SE for the number of correct guesses. Look at the box model. In this example, the null hypothesis tells you exactly what is in the box: a 1 and three 0's. The SD of the box is $\sqrt{0.25 \times 0.75} \approx 0.43$. The SE is $\sqrt{7{,}500} \times 0.43 \approx 37$. So

$$z = 131/37 \approx 3.5$$

The observed value of 2,006 is 3.5 SEs above the expected value. And P is rather small:



The surplus of correct guesses is hard to explain away as a chance variation. Of course, this doesn't prove that ESP exists. For example, the Aquarius random number generator may not be very good (section 5 of chapter 29). Or the machine may be giving the subject some subtle clues as to which target it picked. There may be many reasonable explanations for the results, besides ESP. But chance variation isn't one of them. That is what the test of significance shows, finishing the ESP example.

The same z-statistic is used for ESP as for the tax example:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

Although the formula is the same, there are some differences between the z-test in this section and the z-test in section 1.

1) In section 1, the SE was for an average; here, the SE is for the number of correct guesses. To work out z, first decide what is "observed" in the numerator. Are you dealing with a sum, an average, a number, or a percent? That will tell you which SE to use in the denominator. In the ESP example, the number of correct guesses was observed; that is why the SE for the number goes into the denominator, as indicated by the sketch at the top of the next page.

whether each
e 7,500 draws

n the box. This

d counting the
. Once the null
used:

pected number
The numerator
orrect guesses.
ber of correct
ypothesis tells
of the box is

And *P* is rather

00

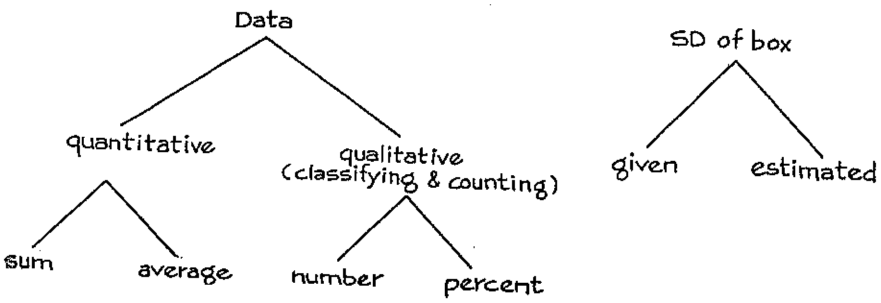
hance variation
quarius random
Or the machine
it picked. There
SP. But chance
shows, finishing

tween the

or the number
n the numerator
t? That will
, the number of
er goes into the
ge.

$$z = \frac{\overset{\text{number}}{\text{observed}} - \overset{\text{number}}{\text{expected}}}{\underset{\text{for number}}{\text{SE}}}$$

2) In section 1, the SD of the box was unknown; the Treasury official had to estimate it from the data. With the ESP example, the SD of the box is given by the null hypothesis: you do not have to estimate it. The diagram below summarizes points 1) and 2).



3) With the Treasury example, there was an alternative hypothesis about the box; its average was negative. With the ESP example, there is no sensible way to set up the alternative hypothesis as a box model. The reason: if the subjects do have ESP, the chances for each guess to be right may well depend on the outcomes of previous trials, and may change from trial to trial. Then the data will not be like draws from a box.[8]

4) In section 1, the data were like draws from a box, because the Treasury official took a simple random sample of forms; the argument was only about the average of the box. Here, part of the question is *whether* the data are like draws from a box—any box.

Chapters 19–24 were about *estimation*—how big is the difference? how reliable is the estimate? *Testing* puts another spin on the question—is the difference real, or due to chance?

Exercise Set E

This exercise set also covers material from previous sections.

1. In Tart's experiment, the null hypothesis says that ______. Fill in the blank, using one of the options below.
   (i) The data are like 7,500 draws from the box $\boxed{\boxed{0}\,\boxed{0}\,\boxed{0}\,\boxed{1}}$.
   (ii) The data are like 7,500 draws from the box $\boxed{\boxed{0}\,\boxed{0}\,\boxed{1}}$.
   (iii) The fraction of 1's in the box is 2,006/7,500.
   (iv) The fraction of 1's among the draws is 2,006/7,500.
   (v) ESP is real.

2. As part of a statistics project, Mr. Frank Alpert approached the first 100 students he saw one day on Sproul Plaza at the University of California, Berkeley, and found out the school or college in which they enrolled. His sample included 53 men and 47 women. From Registrar's data, 25,000 students were registered at Berkeley that term, and 67% were male. Was his sampling procedure like taking a simple random sample?

   Fill in the blanks. That will lead you step by step to the box model for the null hypothesis. (There is no alternative hypothesis about the box.)

   (a) There is one ticket in the box for each ______.

   person in the sample   student registered at Berkeley that term

   (b) The ticket is marked ______ for the men and ______ for the women.

   (c) The number of tickets in the box is ______ and the number of draws is ______. Options: 100, 25,000.

   (d) The null hypothesis says that the sample is like ______ ______ made at random from the box. (The first blank must be filled in with a number; the second, with a word.)

   (e) The percentage of 1's in the box is ______. Options: 53%, 67%.

3. (This continues exercise 2.) Fill in the blanks. That will lead you step by step to z and P.

   (a) The observed number of men is ______.

   (b) The expected number of men is ______.

   (c) If the null hypothesis is right, the number of men in the sample is like the ______ of the draws from the box. Options: sum, average.

   (d) The SE for the number of men is ______.

   (e) $z =$ ______ and $P =$ ______.

4. (This continues exercises 2 and 3.) Was Alpert's sampling procedure like taking a simple random sample? Answer yes or no, and explain briefly.

5. This also continues exercises 2 and 3.

   (a) In 3(b), the expected number was ______.

   computed from the null hypothesis   estimated from the data

   (b) In 3(d), the SE was ______.

   computed from the null hypothesis   estimated from the data

6. Another ESP experiment used the "Ten Choice Trainer." This is like the Aquarius but with 10 targets instead of 4. Suppose that in 1,000 trials, a subject scores 173 correct guesses.

   (a) Set up the null hypothesis as a box model.

   (b) The SD of the box is ______. Fill in the blank, using one of the options below, and explain briefly.

   $$\sqrt{0.1 \times 0.9} \qquad \sqrt{0.173 \times 0.827}$$

   (c) Make the z-test.

   (d) What do you conclude?

7. A coin is tossed 10,000 times, and it lands heads 5,167 times. Is the chance of heads equal to 50%? Or are there too many heads for that?

   (a) Formulate the null and alternative hypotheses in terms of a box model.

100 students
erkeley, and
included 53
registered at
e like taking

1 for the null

that term
: women.
ber of draws

___ made at
a number; the

67%.

tep by step to

sample is like
age.

ure like taking

m the data

m the data

e the Aquarius
ject scores 1??

: of the options

s the chance of

a box model

(b) Compute z and P.
(c) What do you conclude?

8. Repeat exercise 7 if the coin lands heads 5,067 times, as it did for Kerrich (section 1 of chapter 16).

9. One hundred draws are made at random with replacement from a box of tickets; each ticket has a number written on it. The average of the draws is 29 and the SD of the draws is 40. You see a statistician make the following calculation:

$$z = \frac{29 - 20}{4} = 2.25, \quad P \approx 1\%$$

(a) She seems to be testing the null hypothesis that the average of the ______ is 20. Options: box, sample.
(b) True or false: there is about a 1% chance for the null hypothesis to be right.

Explain briefly.

10. A colony of laboratory mice consisted of several hundred animals. Their average weight was about 30 grams, and the SD was about 5 grams. As part of an experiment, graduate students were instructed to choose 25 animals haphazardly, without any definite method.[9] The average weight of these animals turned out to be around 33 grams, and the SD was about 7 grams. Is choosing animals haphazardly the same as drawing them at random? Or is 33 grams too far above average for that?

Discuss briefly; formulate the null hypothesis as a box model; compute z and P. (There is no need to formulate an alternative hypothesis about the box; you must decide whether the null hypothesis tells you the SD of the box: if not, you have to estimate the SD from the data.)

11. (Hard.) Discount stores often introduce new merchandise at a special low price in order to induce people to try it. However, a psychologist predicted that this practice would actually reduce sales. With the cooperation of a discount chain, an experiment was performed to test the prediction.[10] Twenty-five pairs of stores were selected, matched according to such characteristics as location and sales volume. These stores did not advertise, and displayed their merchandise in similar ways.

A new kind of cookie was introduced in all 50 stores. For each pair of stores, one was chosen at random to introduce the cookies at a special low price, the price increasing to its regular level after two weeks; the other store in the pair introduced the cookies at the regular price. Total sales of the cookies were computed for each store for six weeks from the time they were introduced.

In 18 of the 25 pairs, the store which introduced the cookies at the regular price turned out to have sold more of them than the other store. Can this result be explained as a chance variation? Or does it support the prediction that introducing merchandise at a low price reduces long-run sales? (Formulate the null hypothesis as a box model; there is no alternative hypothesis about the box.)

The answers to these exercises are on pp. A92–94.