

### Introduction to Hypothesis Testing Using the Single-Sample z-Test

**Inferential Statistics:** We want to know about the population (i.e., the general case), but we usually have information only from a sample. The procedures, rules, principles, and so forth involved in taking information from a sample and applying it to a population is called *statistical inference*. One form of inference is called *hypothesis testing*, in which we test hypotheses about the population (or general case) by looking at data from a sample (a specific case).

If you take a coin from your pocket and examine it carefully, can you tell if the coin is "fair?" A fair coin is one that would land heads side up 50% of the time and tails side up 50% of the time when tossed a very large number of times (billions, trillions, an infinite number). The obvious test is to toss the coin a few billion times and see what happens. Is anybody up for that? Umm, I didn't think so! (There is a famous case where a mathematician tossed a coin 10,000 times, but he had very little else to do, seeing as how he was in prison!)

Since populations are usually too large to be tested in their entirety, the only practical thing to do is to take a sample and look at that. So instead of tossing the coin a few billion times (which actually would still be just a sample of the infinite number of times the coin might be tossed), we'll settle for a sample of  $n = 80$  tosses. What should we expect to see?

**Hypotheses:** If the coin is fair, then we "should expect" 50% heads and 50% tails. If the coin is unfair, then we "should expect" something else. These statements, which predict the outcome of our little coin-tossing study, are called *hypotheses*.

The *null hypothesis* says, essentially, "nothing to see here." Everything is exactly as we expect. Nothing is out of the ordinary. The treatment won't work. No difference will occur between what we will see and what is "fair." The most important thing about the null hypothesis is that it makes a precise numerical prediction. This will form the basis of our statistical calculations. The null hypothesis is usually denoted by the symbol  $H_0$ .

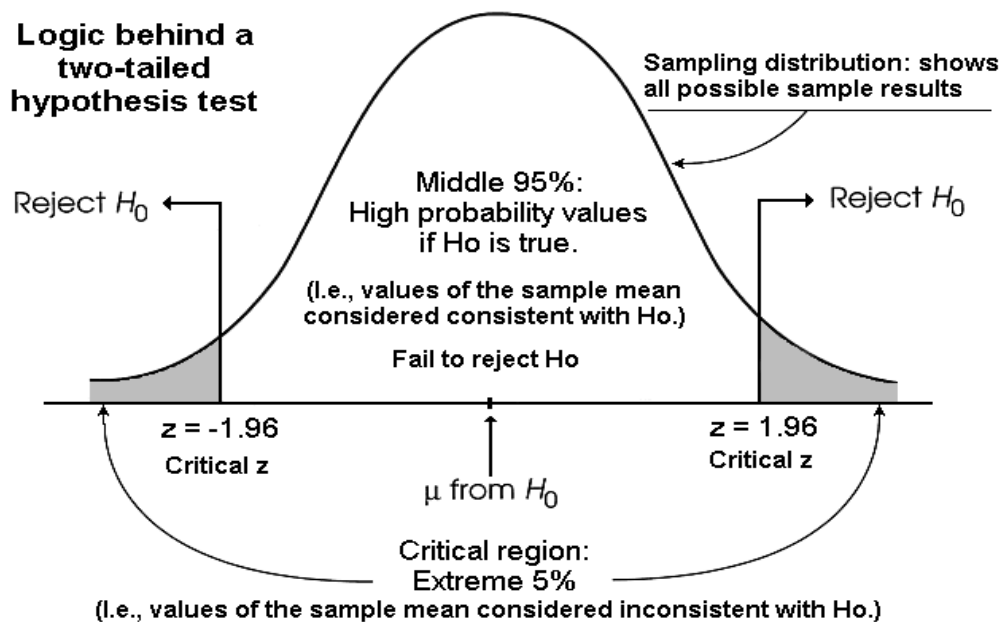
The *alternative hypothesis* says, essentially, "the null hypothesis is wrong." Something is going on here. The treatment will work. We are going to see a difference. The alternative hypothesis need not make a precise prediction (and usually doesn't). It is denoted by the symbol  $H_1$ .

Although the alternative hypothesis does not have to make a precise prediction, it does sometimes predict the direction the effect will go in. This is called a *directional alternative hypothesis*. (Sometimes this is called a "one-tailed" hypothesis, which is not formally correct.) Thus, "we will see more than 50% heads" would be a directional hypothesis. If the alternative hypothesis does not predict a direction, it's called a *nondirectional alternative hypothesis*. "We don't know whether there will be more heads or more tails, but it isn't going to be 50:50." This would be a nondirectional hypothesis.

**Sampling Error:** The rule is, *samples are always wrong. What we expect to happen never does happen.* Another way to say this is, *any possible result can happen by random chance alone.*

That coin might be perfectly fair, and we might toss it 80 times and get 80 heads just by dumb luck alone. It's not likely, but it's possible. *A single hypothesis test will not allow us to say definitely what's true and what's false.* It will allow us to say only what's *probably true* and what's *probably false*.

**The Sampling Distribution:** What we need to see is the distribution of possible results under the null hypothesis. What might our samples look like if the null hypothesis is correct. When we look at this *sampling distribution*, we will see that some results are likely to happen if the null is true, and some are unlikely to happen if the null is true. Based on these facts, we can establish a *decision criterion*. In other words, we need to "draw a line in the sand," so to speak. Results that are likely to happen under the null will be considered consistent with the null hypothesis. Results that are unlikely to happen under the null will be considered inconsistent with the null. The easiest cases to deal with are those where the sampling distribution is a normal distribution.



**Decision Errors:** In the end, we will make a statistical decision, which will be either, "Reject the null hypothesis," or, "Fail to reject the null hypothesis." If we toss the coin 80 times and get 50 heads, this might lead us to reject the null and call the coin unfair. We might be wrong, because 50 heads could occur just by chance even if the coin is fair (due to nothing more than *sampling error*). Or we might toss the coin 80 times, see 45 heads, and conclude this is consistent with the null that the coin is fair. Again, we could be wrong. Sampling error might have given us a "fair-looking" result when, in fact, the coin is biased. Thus, there are two decision errors we might make.

|                             | the truth (unknown to researcher) |                         |
|-----------------------------|-----------------------------------|-------------------------|
|                             | H <sub>0</sub> is true            | H <sub>0</sub> is false |
| H <sub>0</sub> rejected     | Type I error                      | correct decision        |
| H <sub>0</sub> not rejected | correct decision                  | Type II error           |

## Terms:

- 1) null hypothesis
- 2) alternative hypothesis
- 3) directional hypothesis
- 4) nondirectional hypothesis
- 5) one-tailed test
- 6) two-tailed test
- 7) decision criterion
- 8) significance level ("alpha level")
- 9) test statistic
- 10) critical value
- 11) statistical significance
- 12) Type I error
- 13) Type II error
- 14) effect size
- 15) power

**The Single-Sample z-Test (test of a single sample mean when  $\sigma$  is known or can be very accurately estimated):** Here are some examples.

Scores (verbal + quantitative) on the Scholastic Assessment Test (SAT) are normally distributed with  $\mu = 1000$  and  $\sigma = 200$  in the population. A high school counselor developed a special SAT preparation course. To determine if the course has a beneficial effect on SAT performance, the counselor will select a random sample of  $n = 25$  students. Test the researcher's hypothesis that the prep course has a beneficial effect.

**Step 1)** State the null and alternative hypotheses.

$H_0$ : The SAT prep course has no effect on mean SAT scores.  $\mu_0 = 1000$ .

$H_1$ : The SAT prep course changes the mean of SAT scores.

**Step 2)** Establish a decision criterion. In other words, set the alpha level.

$\alpha = .05$ . The alternative hypothesis is nondirectional, so we are doing a two-tailed test. Therefore, we will split the critical region (or "rejection region") between the two tails of the distribution (as in the diagram above). The critical value of the test statistic will cut off this part of the distribution. Therefore,  $z_{crit} = \pm 1.96$ .

**Step 3)** Go out and collect the data and calculate summary statistics. Then get the calculated value of the test statistic.

A sample of  $n = 25$  students was selected. Each student completed the prep course and then took the SAT. The mean score of the sample was found to be  $M = 1090$ .

The amount of sampling error we "should expect" is given by the s.e.m., which is now

calculated.

$$s.e.m. = \sigma / \sqrt{n} = 200 / \sqrt{25} = 200 / 5 = 40$$

Now we can calculate the test statistic:

$$z = \frac{M - \mu_0}{s.e.m.} = (1090 - 1000) / 40 = +2.25$$

**Step 4)** Make a statistical decision concerning the null hypothesis.

The null hypothesis is rejected. This is so since the obtained result was in the area of the sampling distribution judged to be too unlikely to be consistent with the null, i.e., the rejection region.

**Step 5)** Write a conclusion that describes the results of the statistical analysis. (*The following example shows how this might be written in a manuscript being prepared for publication in a journal that follows APA publication style.*)

The SAT preparation course had a significant effect on the mean SAT score,  $z = 2.25$ ,  $p < .05$ , one-tailed. The mean score of the students who completed the SAT preparation course ( $M = 1090$ ) was significantly higher than the mean score of students in the general population ( $\mu = 1000$ ).

-----

In the general population of adults, scores on a standardized short-term memory test are normally distributed, with  $\mu = 50$  and  $\sigma = 6$ . A neuropsychologist wants to conduct a study to determine if chronic smokers of marijuana would perform worse than the general population (since it is generally believed that marijuana smoking impairs STM). A random sample of  $n = 16$  adult chronic marijuana smokers will be obtained and given the memory test to test his hypothesis.

**Step 1)** State the null and alternative hypotheses.

$H_0$ :

$H_1$ :

**Step 2)** Establish a decision criterion. In other words, set the alpha level.

$\alpha = .05$ . The alternative hypothesis is directional, so we are doing a one-tailed test. Therefore, we will put the entire critical region (or "rejection region") in one tail of the distribution. Since we are looking for a result in the lower tail to confirm  $H_1$ , that's where

our critical region ("rejection region") goes. The critical value of the test statistic will cut off this part of the distribution. Therefore,  $z_{\text{crit}} = -1.645$ .

**Step 3)** Go out and collect the data and calculate summary statistics. Then get the calculated value of the test statistic.

The sample of  $n = 16$  subjects gave the following scores on the memory test:

40 49 46 48 51 47 49 45 44 50 48 46 42 52 47 48

**Step 4)** Make a statistical decision concerning the null hypothesis.

**Step 5)** Write a conclusion that describes the results of the statistical analysis.

Question: How would the results of this analysis be different if alpha had been set at .01?

-----

On a vocational interest inventory that measures interests in several categories, a very large standardization group of adults has an average score on the literary scale of  $\mu = 22$  with  $\sigma = 4$ . A researcher would like to determine if scientists differ from the general population in terms of writing interests.

**Step 1)** State the null and alternative hypotheses.

$H_0$ :

$H_1$ :

**Step 2)** Establish a decision criterion. In other words, set the alpha level.

**Step 3)** Go out and collect the data and calculate summary statistics. Then get the calculated value of the test statistic.

The random sample of  $n = 8$  scientists gave the following scores on the memory test:

21 20 23 28 30 24 23 19

**Step 4)** Make a statistical decision concerning the null hypothesis.

**Step 5)** Write a conclusion that describes the results of the statistical analysis.

-----

**Assumptions of the Single-Sample z-Test:**

- 1) Random sampling from the population of interest.
- 2) A normally distributed sampling distribution for the mean.
- 3) Treatment does not change the variability of the scores.
- 4) All scores are independently measured.

**Factors Affecting the Power of the Single-Sample z-Test:**

- 1) The size of the effect,  $M - \mu$  (the bigger the better).
- 2) The size of  $\sigma$  (the smaller the better).
- 3) The alpha level ( $\alpha = .01$  less powerful than  $\alpha = .05$ ).
- 4) The directionality of the test (one-tailed more powerful).
- 5) The sample size,  $n$  (the bigger the better).

**A Note on Confidence Intervals:**

Some statisticians would now (or even earlier) calculate a confidence interval for the population mean using the sample data. These techniques are covered in Chapter 12 of the textbook, which you are not responsible for. But just for your information, the confidence interval in the first problem above would be calculated as  $M \pm Z_{\text{crit},2\text{-tailed}} \times \text{s.e.m.}$  So we can say with a confidence of  $100 \times (1 - \alpha)\%$ , or 95% in this case, that the true population mean of people who take the SAT prep course is in the interval  $1090 \pm 1.96 \times 40$ , or 1011.6 to 1168.4. Notice that this does not include the null hypothesized value of 1000, so this also allows the null to be rejected.