

6: Inference About a Mean: Hypothesis Testing

...the pitfall is in adopting procedures as things in their own right rather than by having regard to the central objectives the procedures are intended to achieve.—R. A. Cox (Armitage, 1983, p. 332)

[Testing Procedure](#) | [Type I and Type II Errors](#) | [One-Sample \$z\$ Test](#) | [One Sample \$t\$ Test](#)

Testing Procedure

We now enter an area of statistics that is frequently misunderstood—that of *statistical hypothesis testing* (also called *null hypothesis testing* and *significance testing*). When confronted with an observed difference, this form of statistical inference seeks to help us answer the initial question “ought I take any notice of that?” (Fisher, 1951; Lehmann, 1993).

Comment: Seldom is a single study sufficient for reaching a firm conclusion. Rather, each study must be seen as part of a larger picture to piece together existing theory.

There are two related methods of statistical hypothesis testing. One method is called *significance testing* (Fisher’s method), and the other method is called *fixed-level hypothesis testing* (Neymann–Pearson’s method). However, since both methods lead to substantially the same results, it is convenient to ignore subtle distinctions between the two for now, focusing instead on their similarities.

Let us break the statistical hypothesis testing procedure into the following steps:

- (A) The research question is stated in null and alternative forms
- (B) An error threshold for the decision is set (*used in fixed-level testing only*)
- (C) A test statistic is calculated and compared to a probability distribution for the sake of deriving a probability statement
- (D) A conclusion is reached

Step A: Null and alternative hypotheses

The first step of statistical testing is to convert the research question into null and alternative forms. We use the notation H_0 to represent the null hypothesis and H_1 (or H_a) to denote the alternative hypothesis. H_0 is a statement of “no difference.” This is the hypothesis that the researcher hopes to reject. H_1 opposes H_0 . *We retain the premise of the null hypothesis until proved otherwise.* This has a basis in [quasi-]deduction and is analogous to the presumption of innocence in a criminal trial.

Step B: Error threshold (α)

If we wish to reach a “yes-or-no decision,” fixed level testing must be pursued. (This is not always necessary, and is sometimes unwise.) To pursue fixed-level testing, we set an error threshold for the decision. The error threshold, called **alpha** (α), is the probability the researcher is willing to take of incorrectly rejecting a true H_0 . For example, the researcher may be willing to take a 1% chance of incorrectly rejecting a true H_0 . In such instances, $\alpha = .01$.

Step C: Test Statistic

A test statistic is calculated. There are different test statistics depending on the data being tested and question being asked. In this chapter, we introduce tests of single means. For single means tests, the null hypothesis is $H_0: \mu = \text{“some value”}$ and the test statistic is either a z_{stat} or t_{stat} . These statistics are introduced below.

Step D: Conclusion

We **convert the test statistic to a p value** by placing the test statistic on its appropriate probability distribution and determine the area under the curve beyond the test statistic.

With **fixed-level testing**, the p value is compared to the α level and this simple decision rule is applied:

When $p \leq \alpha$, H_0 is rejected.
When $p > \alpha$, H_0 is retained.

With **flexible significance testing**, the p value answers the question:

If the null hypothesis were true, what is the probability of observing the current test statistic or a test statistic that is more extreme than the current test statistic?

Thus, the smaller p value, the better the evidence against H_0 . As an *initial* rule-of-thumb we might say that we ought to take note of any p value approaching .05 (or less). In the parlance of statistics, such findings denote “statistical significance.”

Fallacies of Statistical Testing

As stated at the onset of this chapter, statistical testing is frequently misunderstood even by experienced scientists. Therefore, it may be useful to list some fallacies of hypothesis testing from the onset. Fallacies include:

1. Failure to reject the null hypothesis leads to its acceptance. (WRONG! Failure to reject the null hypothesis implies insufficient evidence for its rejection.)
2. The p value is the probability that the null hypothesis is incorrect. (WRONG! The p value is the probability of the current data or data that is more extreme assuming H_0 is true.)
3. $\alpha = .05$ is a standard with an objective basis. (WRONG! $\alpha = .05$ is merely a *convention* that has taken on unwise mechanical use.)
4. Small p values indicate large effects. (WRONG! p values tell you next to nothing about the size of a difference.)
5. Data show a theory to be true or false. (WRONG! Data can at best serve to bolster or refute a theory or claim.)
6. Statistical significance implies importance. (WRONG! WRONG! WRONG! Statistical significance says very little about the importance of a relation.)

Type I and Type II Errors

When we conduct a **fixed-level test**, we set the α level as a decision threshold. However, α level of the test addresses only false rejections of H_0 , and *not* false retentions. The two types of testing errors are labeled:

Type I errors = rejections of correct null hypotheses

Type II errors = retentions of incorrect null hypotheses

The consequences of any given test can thus be summarized:

		TRUTH	
		H_0 True	H_0 False
DECISION	Retain H_0	Correct Retention	Type II Error
	Reject H_0	Type I Error	Correct Rejection

If we compare a type I error to a *false positive* alarm (“an alarm without a fire”), a type II error is a false negative (“a fire without an alarm”). The probability of a type I error is alpha (α), and the probability of a type II error is beta (β):

$$\text{Pr}(\text{type I error}) = \alpha$$

$$\text{Pr}(\text{type II error}) = \beta$$

We also speak of the complements of α and β . The probability of *not* making a type I error ($1 - \alpha$) is called *confidence*. The probability of *not* making a type II error ($1 - \beta$) is called *power*. Thus:

$$\text{Pr}(\text{avoiding a type I error}) = 1 - \alpha = \textit{confidence}$$

$$\text{Pr}(\text{avoiding a type II error}) = 1 - \beta = \textit{power}$$

Conventional levels for *confidence* are .90, .95 and .99. Conventional levels for *power* are .80, .90, and .95.

One-Sample z Test

One-Sided Alternative

The one-sample z test compares a single mean to an “expectation” while using a known or assumed population standard deviation (σ). The test can be done in a one-sided or two-sided way. We start by considering one-sided alternatives.

Illustrative example (“IQ Data”). Suppose you hypothesize that IQs of children at a particular school are above average. It is known that Wechsler IQ scores are normally distributed with a mean of 100 and standard deviation of 15. A random sample of 9 children ($n = 9$) from the school shows a mean (\bar{x}) of 112.8.

(A) Null and Alternative Hypotheses. The **null hypothesis** is a statement of “no difference.” Let μ_0 represent the value of the mean under the null hypothesis. (This is as the null value for the parameter.) The null hypothesis may be denoted $H_0: \mu = \mu_0$. Notice that for one-sample tests of means (as we have here), the value of μ_0 will change depending on the research question. For illustrative example, $H_0: \mu = 100$ since an IQ value of 100 represents *no different* than typical.

Comment: For one-sided tests of this type, some texts identify the null as $H_0: \mu \leq \mu_0$ or $H_0: \mu \geq \mu_0$ depending on the direction of the alternative. This is pertinent for fixed-level decision but is unnecessary for flexible significance test. A statement flexible statement of the null hypothesis such as “ H_0 : no difference” is adequate for most purposes.

The **alternative hypothesis** can take *one* of three forms:

- A one-sided form to the right (looking for a mean that is greater than expected ($H_1: \mu > \mu_0$))
- A one-sided form to the left (looking for a mean that is less than expected ($H_1: \mu < \mu_0$))
- A two-sided form (looking for a mean that is different from expected ($H_1: \mu \neq \mu_0$))

We want to test whether the children in the sample have an average IQ that is significantly *higher* than average. Thus, a one-sided alternative to the right is used $H_1: \mu > 100$.

(B) Alpha. Declaring an alpha level is necessary when conducting a fixed level test. For the illustrative example, let $\alpha = .01$. (In flexible significance testing, a specific α level is unnecessary. Instead, the p value is used directly as a measure of evidence.)

(C) Test statistic: The test statistic for this problem is:

$$z_{\text{stat}} = \frac{\bar{x} - m_0}{SEM} \quad (6.1)$$

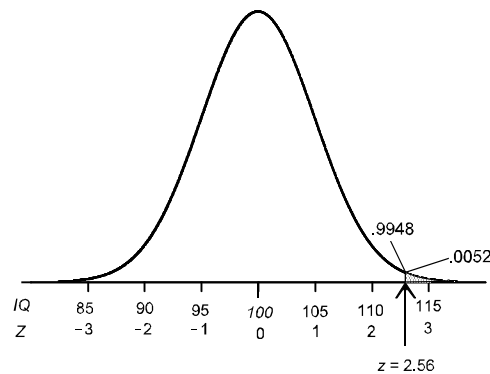
where:

- \bar{x} represents the sample mean,
- μ_0 represents the “null value” (the value of the mean under the null hypothesis), and
- SEM represents the standard error of the mean: $SEM = \sigma / \sqrt{n}$

For the illustrative data, $SEM = 15 / \sqrt{9} = 5$ and $z_{stat} = \frac{112.8 - 100}{2.56} = 2.56$. This z_{stat} indicates that the sample mean is 2.56 SEMs above the null value.

(D) Conclusion. The z_{stat} is converted to a p value by finding the *area under the curve* beyond it on a Z (standard normal) curve. Most computer programs print the p value directly. You can also use a Z table (Appendix 1) to determine the area.

Using the Z table (Table 1) we find that .9948 of the area is to the left of 2.56. We need the area to the right of 2.56. Using the law of complements we find this is equal to $1 - .9948 = .0052$. Thus, $p = .0052$ (Fig.)



With **fixed-level testing**, we would reject H_0 since the p value is less than the pre-specified α . We now say there is *significant evidence* to indicate the children have IQs that are above average. Like the children in Lake Woebegone, these children are above average.

With flexible **significance testing**, we are less rigid. The p value indicates that the probability of observing data this is more extreme than the current data, assuming H_0 is true, is only .0054. The evidence is significant in the sense of being unlikely if the null hypothesis were true.

Two-Sided Alternative

The previous illustration used a one-sided alternative to test the data. We will now use a two-sided test. Two-sided tests allow for unanticipated findings that are either “up” or “down” from what is expected.

Illustrative example. Let us once again use the “IQ” illustrative data. Recall that $\bar{x} = 112.8$ and $n = 9$. We assume $\sigma = 15$ and want to test whether the nine children in the sample are significantly *different* from average.

(A) Null and Alternative Hypotheses. The null hypothesis is $H_0: \mu = \mu_0$. However, since we are now open to difference that are either above or below expected, the alternative hypothesis is expressed in the form of an inequality $H_1: \mu \neq \mu_0$. For the illustrative data, $\mu_0 = 100$. Thus, the null and alternative hypotheses are $H_0: \mu = 100$ versus $H_1: \mu \neq 100$.

(B) Alpha Level. With fixed-level testing, we set alpha (e.g., let $\alpha = .01$). With flexible significance testing, we may skip this step.

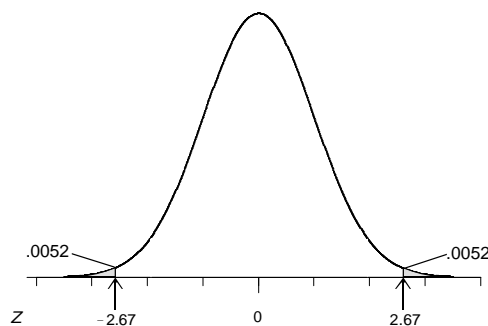
(C) Test Statistics. The test statistic is the same as before: the one-sample z statistic. Formula 6.1 is used to calculate the z_{stat} which is, once again, equal to 2.56.

(D) Conclusion. In light of the two-sided alternative, we could conceivably reject the null hypothesis in favor of the alternative if \bar{x} is either significantly greater than μ_0 or significantly less than μ_0 . Therefore, rejection regions lie in both tails of the “null” standard normal probability distribution. Although \bar{x} cannot fall into both tails of the probability distribution, we start by assuming it falls on the upper tail. If the alternative hypothesis was one-sided, the p value would be the area under the curve in the upper extent of the distribution, beyond the z_{stat} . Because this is a two-tailed test, we must account for both tails by doubling the probability in the tail.

If $\bar{x} > \mu_0$, the p value is *double* the area under the curve in the right-tail of Z distribution.

If $\bar{x} < \mu_0$, the p value is *double* the area under the curve in the left-tail of Z distribution.

Thus, for the illustrative example, $p = 2 \times .0052 = .0104$.



So how would we interpret this result? With **fixed-level testing**, we would *retain* H_0 since the p value is greater than the pre-specified α . At the same time we would not the possibility of a type II error (false retention of H_0), which is especially relevant given the small sample size ($n = 9$). With flexible **significance testing**, we would note that the evidence of the null hypothesis is strong: the probability of observing a difference equal to or more extreme than the current data is only .0104, assuming the null hypothesis is

true.

Assumptions

All statistical inferences require distributional and validity assumptions. For now, let us focus on the the sampling and distributional assumption of the z test.

The main **sampling assumption** of the z test is that observations in the sample are **independent**. This means that data were derived by a simple random sample: each member of the population has an equal (and “independent”) probability of entering the sample. For example, if there were 900 children in the school from which the illustrative data were derived, each would have a $9 / 900$ (1%) chance of entering the sample. If, on the other hand, a particular class in the school was used to select the sample and all other classes were excluded, the sample would no longer be independent.

The main distributional assumption of the z test is that the sampling distribution of the *mean* (not the initial data) is **normal**. Because the distributional assumption is directed toward the hypothetical distribution of means, and distributions of means are influenced by the central limit theorem, modest departures from normality in the population are fully permissible. The property is often referred to as **robustness**, meaning that the test still derives meaningful results even in the face of distributional violations.

However, we must **guard against** severely skewed or symmetric distributions with long tails. Asymmetry and long tails can often be identified with EDA (e.g., via boxplots showing outside values, asymmetric boxes, or long whiskers on one the top or bottom).

Some statisticians offer rules of thumb for guarding against violations in the normality assumption. For example, some suggest:

- When the sample is small (say, $n < 10$), the population distribution should be close to normal.
- When the sample is moderate (say, $10 < n < 30$), the population should be fairly symmetric and mound-shaped with no outliers.
- When the sample is large (say, $n > 30$), the z test is permissible except in the case of extreme outliers and strong skewness.

Note, however, assessing the distribution of the population is difficult. You know only the shape of the sample distribution, and not the population distribution. Therefore, the distributional characteristics of the population are at best conjectural. In practice we rely on common sense, the robustness of the test, and a certain perceptiveness which views statistical models as non-realistic but useful tools of inference.

One-Sample t Test

In the prior section we introduced hypothesis testing a problem that involved the mean of a population. The z test was presented to test $H_0: \mu = \mu_0$ in either a one-sided to two-sided way. Recall that the z_{stat} used to conduct the test depended on knowing or assuming population standard deviation σ . In most instance, however, it is not reasonable to assume the population standard will be known. Thus, to test $H_0: \mu = \mu_0$ when the population standard deviation is not known we modify the z_{stat} by estimating σ with the standard deviation calculated in the sample, s . In such instances, a t tests statistic is used instead of a z statistic.

Illustrative Example (%ideal.sav). Let us consider data from a study in which body weights are expressed as a percentage of ideal. For example, a value of 100 will represent 100% of ideal body weight, a value of 120 will represent 20% above ideal body weight and so on. Data are:

107 119 99 114 120 104 88 114 124 116 101 121 152 100 125 114 95 117

The sample has 18 observations and a mean (\bar{x}) of 112.778. The population standard deviation is unknown, but the sample standard deviation (s) is calculated to be 14.424.

(A) Hypotheses: For this problem, let us ask whether the population mean is other than ideal by conducting a two-sided test. (One-sided tests are conducted in a similar way, except the p value would be one-tailed instead of two-tailed.) For the illustrative data, the two-sided hypotheses are $H_0: \mu = 100$ versus $H_1: \mu \neq 100$.

(B) Alpha (needed for fixed-level testing). Let us conduct a fixed level test with $\alpha = .05$.

(C) Test Statistic. Since the population standard deviation is unknown, we conduct a t test rather than a z test. The one-sample t statistic is:

$$t_{\text{stat}} = \frac{\bar{x} - \mathbf{m}_0}{sem} \quad (6.2)$$

where

\bar{x} represents the sample mean,

μ_0 represents the null value, and

sem represents the estimated standard error of the mean calculated from the sample as: $sem = s / \sqrt{n}$.

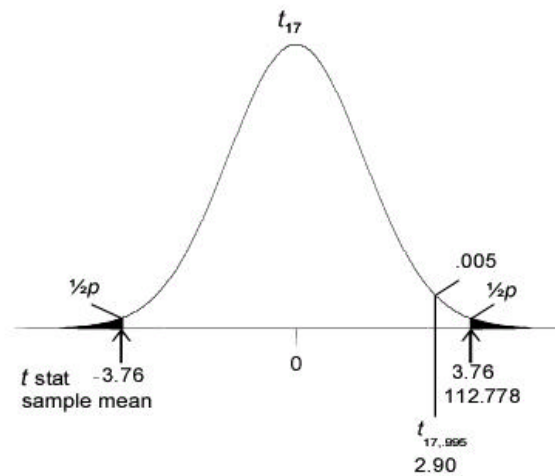
This statistic has $n - 1$ degrees of freedom.

For the illustrative example, $sem = s / \sqrt{n} = 14.424 / \sqrt{18} = 3.400$ and the $t_{\text{stat}} = (112.778 - 100) / (3.400) = 3.76$ with $df = 18 - 1 = 17$.

(D) Conclusion: Most computer programs calculate precise p values for t tests. However, when a computer program is unavailable, you must look up the approximate p value using a t table (e.g., Appendix 2). Be forewarned that the precise p value cannot be found in the t table. With the t table, you will be able to find bounds for the p value, as described below:

Recall that the two-sided p value is twice the area under the curve beyond the t_{stat} . Bounds for the p value can be derived as follows:

- Draw a t curve (which looks for current intents and purposes like a z curve). Recall that the curve is centered on 0 and the points of inflection of the curve are at approximately ± 1 standard errors from its.
- The current t_{stat} is 3.76 “standard deviations” (actually, standard errors) to the right of center. This places it in the far-right tail. Mark the test statistic in this approximate location and shade the area to its right. Also, mark -3.76 on the curve, and shade this region, since the current alternative is “two-tailed.” These shaded regions represent the p value.
- In the t table (Table xx), find the nearest known percentile just to the left of the t_{stat} . Thus, we find that the 99.5th percentile on the t distribution with 17 degrees of freedom ($t_{17,.995}$) is equal to 2.90. This landmark is associated with a right-tail region of .005.
- Since the shaded area (representing the half the p value) is beyond this landmark, we know it is less than .005. Double this is thus less than .01. Therefore, $p < .01$.



Various computer programs (e.g., SPSS, StaTable, and WinPepi > WHATIS.EXE) can be used to convert the t_{stat} to a more precise p value statement. In this instance, a t_{stat} of 3.76 with 17 degrees of freedom corresponds two-tailed $p = .0016$.

Interpretation of the p value. In pursuing fixed-level testing, the null hypothesis is rejected (since $p < \alpha$). With more flexible significance testing we note that “if the null hypothesis were true, the probability of the current observation, or observations more extreme, would be .0016. Either way, most investigators would agree that we should take note of these results.

SPSS: Use Analyze > Compare Means > One-Sample T Test to conduct this test. Select the variable you want to test and enter the null value in the field labeled “test value.”

Vocabulary

Null hypothesis (H_0) - A statement that declares that the observed difference is due to unexplained “chance.” It is the hypothesis the researcher hopes to reject.

Alternative hypothesis (H_1) - The opposite of the null hypothesis, declaring a non-chance difference.

Alpha (α) - The probability the researcher is willing to take of falsely rejecting an incorrect null hypothesis. In fixed-level testing, this serves as the cutoff point for making decisions about H_0 .

Test statistic - A statistic used to test the null hypothesis.

p value - A probability statement that answers the question “*If the null hypothesis were true*, what is the probability of observing the current data or data that is more extreme than the current data?.” It is the probability of the data conditional on the truth of H_0 . It is NOT the probability that the null hypothesis is true.

Type I error - a rejection of a true null hypothesis; a “false alarm.”

Type II error - a retention of an incorrect null hypothesis; “failure to sound the alarm.”

Confidence ($1 - \alpha$) - the complement of alpha; the probability of correctly retaining a true null hypothesis.

Beta (β) - the probability of a type II error; probability of a retaining a false null hypothesis.

Power ($1 - \beta$) - the complement of β ; the probability of avoiding a type II error; the probability of rejecting a false null hypothesis.