6: Introduction to Hypothesis Testing

Significance testing is used to help make a judgment about a claim by addressing the question, Can the observed difference be attributed to chance? We break up significance testing into three (or four) steps:

Step A: Null and alternative hypotheses

The first step of hypothesis testing is to convert the research question into null and alterative hypotheses. We start with the **null hypothesis (H_0)**. The null hypothesis is a claim of "no difference." The opposing hypothesis is the **alternative hypothesis (H_1)**. The alternative hypothesis is a claim of "a difference in the population," and is the hypothesis the researcher often hopes to bolster. It is important to keep in mind that the null and alternative hypotheses reference population values, and *not* observed statistics.

Step B: Test statistic

We calculate a **test statistic** from the data. There are different types of test statistics. This chapter introduces the one-sample *z*-statistics. The z statistic will compare the observed sample mean to an expected population mean μ_0. Large test statistics indicate data are far from expected, providing evidence against the null hypothesis and in favor of the alternative hypothesis.

Step C: *p* Value and conclusion

The test statistic is converted to a conditional probability called a *P*-value. The *P*- value answers the question "If the null hypothesis were true, what is the probability of observing the current data or data that is more extreme?"

Small *p* values provide evidence against the null hypothesis because they say the observed data are unlikely when the null hypothesis is true. We apply the following **conventions**:

- When p value > .10 → the observed difference is "not significant"
- When p value $\le$ .10 → the observed difference is "marginally significant"
- When p value $\le$ .05 → the observed difference is "significant"
- When p value $\le$ .01 → the observed difference is "highly significant"

Use of "significant" in this context means "the observed difference is not likely due to chance." It does *not* mean of "important" or "meaningful."

Step D: Decision (optional)

Alpha (α) is a probability threshold for a decision. If $P \le \alpha$, we will reject the null hypothesis. Otherwise it will be retained for want of evidence.

One-Sample *z* Test

The one-sample z test is used to compare a mean from a single sample to an expected "norm." The norm for the test comes from a hypothetical value or observations in prior studies, and does not come from the current data. In addition, this test is used only when the population standard deviation σ is known from a prior source. Finally, data represent a SRS, and measurements that comprise the data are assumed to be accurate and meaningful.

Example ("Lake Wobegon"). Garrison Keller claims the children of Lake Wobegon are above average. You take a simple random sample of 9 children from Lake Wobegon and measure their intelligence with a Wechsler test and find the following scores: {116, 128, 125, 119, 89, 99, 105, 116, and 118}. The mean of this sample ($\bar{x}$) is 112.8. We know Wechsler scores are scaled to be Normally distributed with a mean of 100 and standard deviation of 15. Is this sample mean sufficiently different from a population mean μ of 100 to reject the null hypothesis of "no difference?"

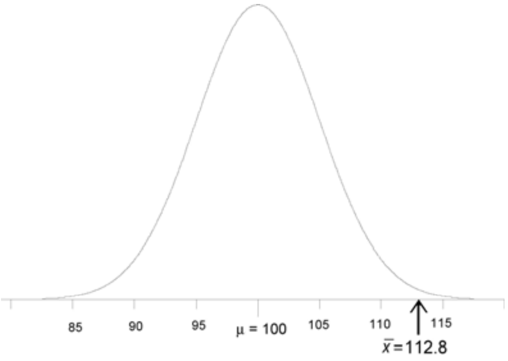
The null and alternative hypotheses

The claim being made in the illustrative example is that the population has higher than average intelligence. The null hypothesis is the population has average intelligence. Since an average intelligence score is 100, H_0: $\mu = 100$.

The alternative hypothesis claims the population has a higher than average intelligence. Therefore, H_1: $\mu > 100$. (The alternative hypothesis resembles the claim the investigator wishes to bolster.) It would be **incorrect** to state H_0: $\bar{x} = 100$. Inferential statements address the population, not the sample. The alternative hypothesis is **one-sided.** It is interested only in whether the population has a *higher* average core. It is not interested a *lower* than average score.

Test statistic

We use our knowledge of sampling distributions of the means (SDM) to help make judgments. Assuming the null hypothesis is true, the sampling distribution of $\bar{x}$ based on $n = 9$ would be Normal with a mean of 100 and standard error of $\sigma / \sqrt{n} = 15/\sqrt{9} = 5$. Therefore, under H_0, $\bar{x} \sim N(100, 5)$.



z-stat-SDM.ai

The observed $\bar{x}$ of 112.8 is out in the right-tail of the SDM.

The **test statistic** compares $\bar{x}$ to the hypothesized value (μ_0) as follows:

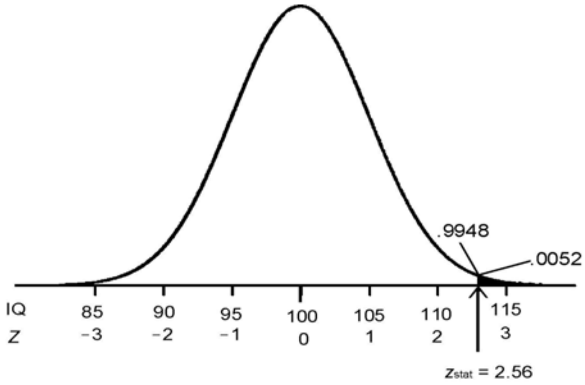
$$z_{stat} = \frac{\bar{x} - \mu_0}{SEM}$$

where *SEM* represents the standard error of the mean and is equal to $\sigma / \sqrt{n}$. The population standard deviation σ must be known in order to use this test statistic. (The z_{stat} quantifies how far $\bar{x}$ is from μ_0 in standard deviation units.)

For the illustrative example, $z_{stat} = \frac{112.8 - 100}{5} = 2.56$.

P-value and conclusion

To convert a z_{stat} to a *P*-value, find the area under the curve beyond the z_{stat} on a Standard Normal distribution. Use the Z table or a statistical package (e.g., *StaTable*) for this purpose. For the current problem we have:



z-stat-one-sided.ai

Therefore, $P = 0.0052$. This provides good evidence against H_0. The jargon is to say the difference is "significant." You can reject H_0.

Two-Sided Alternative

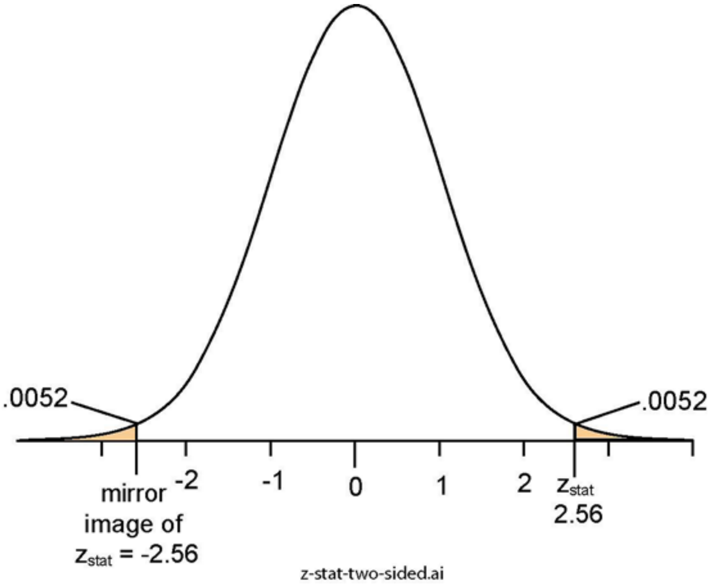
The two-sided *z* test makes no presupposition about the direction of the difference. The null hypothesis is the same as in the one-sided test: H_0: $\mu = \mu_0$. The alternative hypothesis is H_1: $\mu \neq \mu_0$. The test statistic is the same as the one-sample *z* test. Because we must consider sample means that might be above or below μ_0, the *p* value is "two-tailed."

Illustrative example (Two-sided alternative). We return to the Lake Wobegon illustrative example, where children are assumed to be "above average." The two-sided alternative allows for unanticipated findings that are either "up" or "down" from expected.

Let μ_0 represent the expected value of the population mean *under* the null hypothesis. In the Lake Wobegon example, $\mu_0 = 100$. Therefore, we test H_0: $\mu = 100$ versus H_1: $\mu \neq 100$.

The test statistic is $z_{stat} = \frac{\bar{x} - \mu_0}{SEM} = \frac{112.8 - 100}{5} = 2.56$ (same as before).

With two-sided alternatives, we reject the null hypothesis in favor of the alternative if the sample mean is either significantly greater or less than μ_0. Rejection regions for the test statistic lie in both tails of SDM. The effect is to double the size of the one-sided *p* value. The one-sided *p* value for the Lake Wobegon illustration was .0052. Therefore, the two-sided *p* value = 2 × .0052 = .0104.



The *P*-value of 0.0104 provides good evidence against the null hypothesis.

Fallacies of Statistical Hypothesis Testing

The results of statistical tests are frequently misunderstood. Therefore, I'm going to list some of the fallacies of hypothesis testing here. It will be helpful to refer back to this list as you grapple with the interpretation of hypothesis tests results.

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. There is no sharp distinction between "significant" and "insignificant" results, only increasingly strong evidence as the *p* value gets smaller. Surely god loves $p = .06$ nearly as much as $p = .05$)
4. Small *p* values indicate large effects. (WRONG! *p* values tell you next to nothing about the size of an effect.)
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.)

One-Sample *t* Test

The prior section used a z_{stat} to test a sample mean against an expectation. The z_{stat} needed population standard deviation σ (without estimating it from the data) to determine the standard error of the mean. To conduct a one-sample test when the population standard deviation is not known, we use a variant of the z_{stat} called the t_{stat}. The advantage of the t_{stat} is that it can use sample standard deviation *s* instead of σ to formulate the estimated standard error of the mean.

Hypotheses: The null and alternative hypotheses are identical to those used by the *z* test. The null hypothesis is H_0: $\mu = \mu_0$. Alternatives are

H_1: $\mu \neq \mu_0$ (two-sided)
H_1: $\mu > \mu_0$ (one-sided to right)
H_1: $\mu < \mu_0$ (one-sided to left)

Test statistic: The one-sample *t* statistic is:

$$t_{stat} = \frac{\bar{x} - \mu_0}{sem}$$

where $\bar{x}$ represents the sample mean, μ_0 represents the expected value under the null hypothesis, and *sem* = s / √n. This statistic has $n - 1$ degrees of freedom.

***P*-value and conclusion:** The t_{stat} is converted to a *p* value with a computer program or *t* table. When using the *t* table, you will only be able to find boundaries for the *p* value. Small values of *P* provide evidence against H_0.

Illustrative Example (`%ideal.sav`). These data were introduced in the prior chapter. Briefly, each value represents body weight expressed as a percentage of ideal (e.g., 100 represents 100% of ideal). We want to test whether data provides sufficient evidence to support a non-ideal body weight in the population.

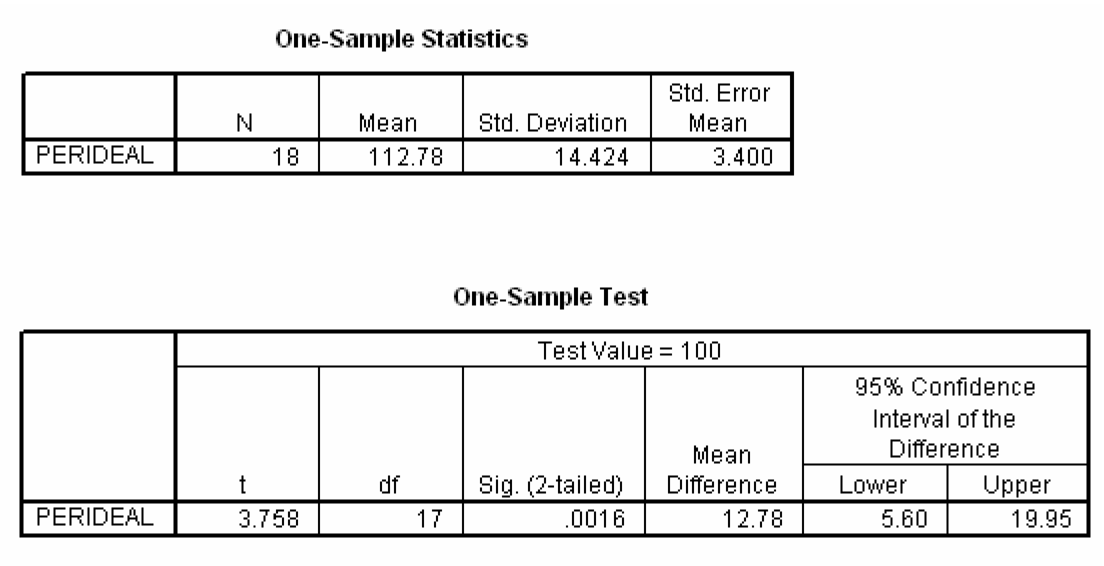
Hypotheses. If the mean body weight in the population was ideal, then μ would equal 100. Therefore, H_0: $\mu = 100$. The test can be one-sided or two-sided. Because we are prudent, we choose the two-sided alternative, which is H_0: $\mu \neq 100$.

Test statistic. Eighteen (n = 18) subjects demonstrate the following data {107, 119, 99, 114, 120, 104, 88, 114, 124, 116, 101, 121, 152, 100, 125, 114, 95, 117}. The sample mean $\bar{x}$ = 112.778. The sample standard deviation *s* = 14.424. The *sem* = 14.424 / √18 = 3.400.

The $t_{stat} = \dfrac{112.778 - 100}{3.400}$ = 3.76 with df = 18 − 1 = 17. The t_{stat} of 3.76 shows that the sample mean of 112.778 is 3.76 standard errors greater than the null value.

***P*-Value and conclusion.** The two-sided *P*-value = 0.0016, indicating that values as far from 100 as $\bar{x}$ = 112.228 would occur 0.16% of the time *if* H_0 were true. This provides good evidence against H_0. We reject H_0: $\mu = 100$ and conclude the difference is significant.

Here's output from SPSS for the problem:



One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
PERIDEAL	18	112.78	14.424	3.400

One-Sample Test

	Test Value = 100					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
PERIDEAL	3.758	17	.0016	12.78	5.60	19.95

If software is unavailable, use a t table to convert the t_{stat} to the P-value. Follow this procedure:

- Draw a t curve. The t distribution curve is centered on 0 with inflection points at ±1. Label the X-axis with tick marks approximately one standard deviation from each other. By the time you get past ±3 standard deviations, the curve should almost be touching the X-axis as an asymptote.
- Place $|t_{stat}|$ on the horizontal axis of the curve in its approximate location. Shade the region in the right-tail. For two-sided alternatives, place $-|t_{stat}|$ on the negative side of the curve.
- Use the t table to find t landmarks just to the left and just to the right of the $|t_{stat}|$. The one-tailed P-value is between these two values. Double the one-side p value to get the two-sided p value.

Here's the visualization for the illustrative example:

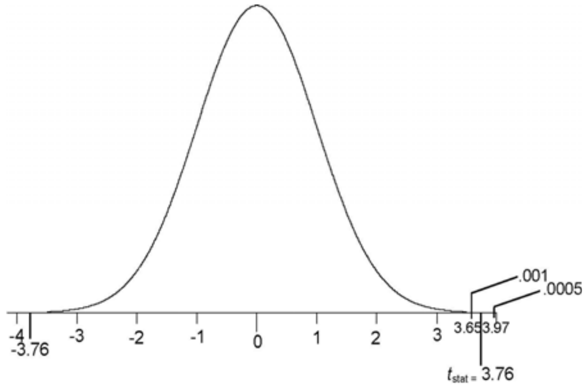


Fig:t_stat=3.76.ai

The t_{stat} (3.76) is between critical values of 3.65 (right-tail region of .001) and 3.97 (right-tail region of .0005). Therefore, the one-tail P-value is greater than 0.0005 and less than 0.001. The two-tail P-value is greater than 0.001 and less than 0.002.

Vocabulary

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

Alternative hypothesis (H_1) - The opposite of the null hypothesis. The hypothesis the researcher hopes to bolster.

Alpha (α) - The probability the researcher is willing to take in falsely rejecting a true null hypothesis.

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 H0. 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.

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.

Comment: Statistical hypothesis testing is not the same thing as scientific hypothesis testing. Joseph Goldberger (1923) studied the cause of pellagra in a three year study completed from 1914 - 1917. The study was carried out at four orphanages and in a state sanitarium at a time when pellagra was endemic. Goldberger modified diets at the institutions by reducing maize content and adding legumes and fresh animal protein (meat, milk, and eggs). More than four hundred pellagrins and non-pellagrins were studied. There was a single recurrence among the pellagrins and not a single case among the non-pellagrins following initiation of dietary changes. Following discontinuation of the study, diets returned to their former state, and so did the state of health of the inmates: the incidence of pellagra returned to approximately 40 percent. Resumption of the modified diet for 14 months was followed by complete disappearance of the disease. This type of challenge-dechallange-rechallenge study did warranted a statistical test; the conclusion that pellagra was prevented by a diet high in legumes and animal protein was clear.