

The same basic logic underlies all statistical hypothesis testing. This toolkit illustrates the basic concepts using the most common tests, t-tests for differences between means.

Tests concerning means are common for at least three reasons. First, many questions of interest concern the average, or typical, values of variables (Is the average greater or less than some standard? Is the average greater or smaller in one group than another?). Second, the mean is the most stable of all the classical measures of a distribution. Third, the distributional properties of means are well understood, and convenient: the central limit theorem shows that means tend toward the normal distribution, which is simple (in the sense that it depends only on the mean and standard error).

Assumptions underlying t-tests:

-The means are normally distributed. That means that either the individual data are themselves normally distributed, or there are enough of them combined together in the means to make the means approximately normal anyhow (20-30 measurements are usually sufficient, depending on how non-normal their distributions are). Deviations from normality become increasingly important as one tries to make inferences about more distant tails of the distribution.

-The only difference between the hypothesis and the null hypothesis is that the true mean of the data would be in a different location; the distributions would have the same shape and width under both hypotheses.

## Outline of Hypothesis Testing

### 1. Formulate the hypothesis ( $H_1$ ) and the null hypothesis ( $H_0$ ).

These hypotheses must be *mutually exclusive* (they cannot be true simultaneously) and *jointly exhaustive* (there are no other possibilities). Both of these conditions must be met before you can legitimately conclude that if  $H_0$  is false, then  $H_1$  must be true.

Example:  $H_1$ : average levels of chromium in drinking water are above the federal 100 ppb standard  
 $H_0$ : average Cr levels in drinking water are at or below the federal standard.

Because the federal standard is fixed, rather than determined by measurement, this is a "single sample test".

Example:  $H_1$ : average levels of chromium in drinking water in Ourtown are higher than in Yourtown.  
 $H_0$ : levels in Ourtown are the same as, or lower than, in Yourtown.

Because you are comparing two sets of measurements (rather than one set against a benchmark), this is a "two sample test". The only difference is that one must account for the uncertainties in both samples (the benchmark in a one-sample test obviously has no uncertainty). This example can be rephrased as if it were a one-sample test: " $H_1$ : The average Cr level in Ourtown, minus the average in Yourtown, is greater than zero" and " $H_0$ : The difference between these averages is negative or zero". Again, the only difference is that, for a two-sample test, you need to pool the variance from your two samples.

### 2. Decide how much risk of "false positives" (Type I error rate= $\alpha$ ) and "false negatives" (Type II error rate= $\beta$ ) you are willing to tolerate.

$\alpha$  is the chance that, if the null hypothesis is actually true, you will mistakenly reject it.

$\beta$  is the chance that, if the null hypothesis is actually false, and the true distribution differs from it by an amount  $\delta$ , you will mistakenly accept it.

Example: If the average Cr level is really below the standard, I want to be 90% sure I don't "cry wolf"; therefore  $\alpha=0.1$ . If, on the other hand, average Cr is above the standard by more than 20 ppb, I want to be 95% sure I've detected the problem; therefore,  $\beta=0.05$  for  $\delta=20$ .

note: The choice of  $\alpha$  and  $\beta$  should be guided by the relative consequences of making either mistake.  $\alpha=0.05$  is conventional, but may not be appropriate for your circumstances.

### 3. Determine how the property of interest would be distributed, if the null hypothesis were true.

Example: Means are often roughly normally distributed (except means of small samples of very non-normal data). Normally distributed means whose standard errors are estimated by sampling (and thus also uncertain) follow Student's t distribution.

### 4. Determine the acceptance and rejection regions for the test statistic, and frame the decision rule.

The rejection region is the tail (or tails) of the distribution encompassing a total probability of  $\alpha$  that are least compatible with the null hypothesis. Thus, if the null hypothesis is true, the chance is  $\alpha$  that the property will fall within the rejection region. The acceptance region is the whole distribution except for the rejection

region: the range that encloses a probability of  $1-\alpha$  that is most compatible with the null hypothesis. The decision rule simply states the conditions under which you'll accept or reject the null hypothesis. Usually, the rule is simply "reject the null hypothesis if the property of interest falls within the rejection region, and accept the null otherwise."

Example: If the true mean ( $\mu$ ) is 100 ppb, then the chance is  $\alpha$  of getting values of  $t = (\bar{x} - 100)/SE(\bar{x})$  that are greater than  $t_{\alpha, \nu}$  (where  $\nu=n-1$  is the number of degrees of freedom). The value  $t_{\alpha, \nu}$  is sometimes referred to as the "critical value" for  $t$ , since this is the value that separates values of  $t$  that are significant at  $\alpha$  from those that are not. Therefore, if  $t > t_{\alpha, \nu}$  reject  $H_0$ ; otherwise accept  $H_0$ .

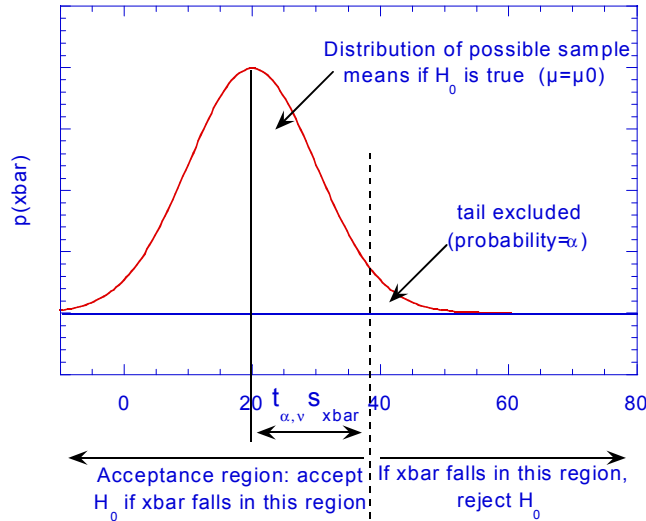


Figure 1: Acceptance and rejection regions in t-test for statistical significance of means.

**5. Now--and not before--look at the data, evaluate the test statistic, and follow the decision rule.**

Example: Five measurements give a mean of 106 ppb, with a standard deviation of 10 ppb. Thus the standard error is 4.47 ppb, and  $t$  (the number of standard errors by which the mean deviates from the 100 ppb benchmark) is thus  $t = (106 \text{ ppb} - 100 \text{ ppb}) / 4.47 \text{ ppb} = 1.34$ . This is less than the critical value of  $t$ ,  $t_{\alpha, \nu} = t_{0.1, 4} = 1.53$ , so the null hypothesis cannot be rejected (without exceeding our tolerance for risk of false positives). However,  $t$  falls between  $t_{0.1, 4}$  and  $t_{0.25, 4}$  so the statistical significance of the data lies between  $p=0.1$  and  $p=0.25$ .

**6. If the null hypothesis was not rejected, examine the power of the test.**

Example: In our example, could the test reliably detect violations of the standard by more than 20 ppb (see *power*, below)?

$$t_{\beta, \nu} = \frac{\delta}{s_{\bar{x}}} - t_{\alpha, \nu} = \frac{20 \text{ ppb}}{4.47 \text{ ppb}} - 1.53 = 2.94$$

The calculated value of  $t_{\beta, \nu}$  is slightly greater than  $t_{0.025, 4} = 2.78$ , so the probability of a "false negative" if the true mean violated the standard by 20 ppb would be just under 2.5%. Equivalently, the power of the test is about  $1-\beta=0.975$ : the test should be roughly 97.5% reliable in detecting deviations of 20 ppb above the standard.

**7. Report the results. Always include the measured magnitude of the property of interest (with standard error or confidence interval), not just the significance level.**

Example: "The average of the concentration measurements was  $106 \pm 4.5$  ppb (mean  $\pm$  SE). The standard is exceeded, but the exceedance is not statistically significant at  $\alpha=0.1$ ; we cannot conclude with 90% confidence that the true mean is greater than 100 ppb. The 90% confidence interval for the mean is 96 to 116 ppb ( $106 \text{ ppb} \pm t_{0.05, 4} * 4.47 \text{ ppb}$ ). The test meets our requirements for power; the test should be about 97% reliable in detecting levels more than 20 ppb above the standard."

---

**Common illegitimate inferences**


---

1. "The data are consistent with the hypothesis, therefore the hypothesis is valid."

The correct inference is: *if* the hypothesis is valid, *then* you should observe data consistent with it. You cannot turn this inference around, because *other* hypotheses may *also* be consistent with the observed data.

The proper chain of inference is this: construct the hypothesis ( $H_A$ ) and the null hypothesis ( $H_0$ ) such that either one or the other must be true. Therefore if  $H_0$  is false, then  $H_A$  must be true. Next, determine what observable data are logically implied by  $H_0$ . If the data you observe are *inconsistent* with  $H_0$ , then  $H_0$  must be false, so  $H_A$  must be true.

In other words: the only way to prove  $H_A$  is true is by showing that the data are *inconsistent* with any other hypothesis.

2. "The null hypothesis was not rejected, therefore the null hypothesis is true."

*Failing to show* the null hypothesis is *false* is not the same thing as *showing* the null hypothesis is *true*.

3. "The effect is not statistically significant, therefore the effect is zero."

*Failure* to show that the effect *is not* zero (i.e., the effect is not statistically significant) is different from *success* in showing that the effect *is* zero.

The effect is *whatever it is*, plus or minus its associated uncertainty. Effects are statistically insignificant whenever their confidence intervals include the null hypothesis; that is, whenever they are *small compared to their uncertainties*. An effect can be statistically insignificant *either* because the effect is small, *or* because the uncertainty is large. Particularly when sample sizes are small, or the amount of random noise in the system is large, many important effects will be statistically insignificant.

4. "The effect is statistically significant, therefore the effect is important."

Again, the effect is *whatever it is*, plus or minus its associated uncertainty. Effects are statistically significant whenever their confidence intervals exclude the null hypothesis; that is, whenever they are *large compared to their uncertainties*. An effect can be statistically significant *either* because the effect is large, *or* because its uncertainty is small. Many trivial effects are statistically significant, usually because the sample size is so large that the uncertainties are tiny.

5. "This effect is more statistically significant than that effect, therefore it is more interesting/important."

At the risk of redundancy: the effects are *whatever they are*, plus or minus their associated uncertainties. That is why you need to *always* report the magnitude of the effects, with quantified uncertainties. If you only report statistical significance, readers cannot judge whether the effects are large enough to be important to them.

Whether an effect is important depends on *whether you care about it*. Whether an effect is statistically significant depends on *whether the observed data would be improbable* if the effect were absent (if the null hypothesis were true). Statistical significance and practical importance are not the same thing; indeed there is no logical connection between them.

6. "The null hypothesis was rejected at  $\alpha=0.05$ . Therefore, the probability is <5% that the null hypothesis is true, and therefore, the probability is >95% that the alternative hypothesis ( $H_A$ ) is true."

Statistical significance is a statement about your *data*, not about your *hypothesis*. Statistical significance of  $\alpha=0.05$  means that *if* the null hypothesis were true, *then* there is less than 5% chance that your data would deviate so far from it. You cannot turn that inference around; you cannot logically conclude that *if* the data deviate so far from the null hypothesis, *then* there is a chance of thus-and-such that the null hypothesis is true or false.

You can validly evaluate the probability that your hypothesis is true or false, but you can only do so by using Bayes' theorem (Bayesian inference). You cannot do it by classical hypothesis testing.

---

**Six Questions that Simple Statistics can Answer**


---

**1. Statistical significance**

Answers the question: *If  $H_0$  were actually true, what is the chance that  $\bar{x}$  would deviate from  $\mu_0$  (the true mean predicted by the null hypothesis) by at least the observed amount?*

-Determine by how many standard errors  $\bar{x}$  deviates from  $\mu_0$ :

$$t = \frac{\bar{x} - \mu_0}{s_{\bar{x}}} = \frac{\bar{x} - \mu_0}{\left( \frac{s_x}{\sqrt{n}} \right)}$$

-Given the number of degrees of freedom ( $v=n-1$ ), find (via a t-table), the two values of  $\alpha$  corresponding to values of  $t_{\alpha,v}$  that bracket the calculated value of  $t$  so that  $t_{\alpha',v} < t < t_{\alpha'',v}$ .

-The significance level of  $\bar{x}$  (denoted  $p$ ) then lies between  $\alpha'$  and  $\alpha''$  such that  $\alpha' > p > \alpha''$ .

note: If the test is two-tailed (that is, either values higher or lower than  $\mu_0$  would be incompatible with  $H_0$ ), and you read the  $\alpha$ 's from a table of one-tailed t values, then the p value must be doubled, to account for the equal probability on the other tail.

**2. Least Significant Difference (LSD)**

Answers the question: *What is the smallest deviation from  $\mu_0$  that would be statistically significant?*

-Look up the critical value of  $t$  for your desired significance level  $\alpha$  and your degrees of freedom  $v$ . Then, given your standard error  $s_{\bar{x}}$  or standard deviation  $s_x$ , solve:

$$LSD = t_{\alpha,v} s_{\bar{x}} = t_{\alpha,v} \frac{s_x}{\sqrt{n}}$$

note: If the test is two-tailed, then you need to use  $\alpha/2$  rather than  $\alpha$  in the expression above.

note: The LSD is simply the end point of the  $1-\alpha$  confidence interval for  $\bar{x}$  around  $\mu_0$ .

note: The LSD is the least significant difference, given your current sample size. You can make smaller differences significant by increasing the sample size.

Important note: The LSD is the smallest difference that, if it is detected, will be statistically significant. However, the *power* to detect differences as small as the LSD is very low (only 50%). That is, *if the true difference is equal to the LSD, you will fail to detect it half the time* (see *power*, below; also see Figure 3). You can only detect deviations from  $\mu_0$  reliably if they are much larger than the LSD.

**3. Least Significant Number (LSN)**

Answers the question: *What number of measurements would be required to make a given deviation from  $\mu_0$  statistically significant?*

-Look up the critical value of  $t$  for your desired significance level  $\alpha$  and your degrees of freedom  $v$ . Then, given your observed standard deviation  $s_x$ , solve:

$$n = \left( \frac{t_{\alpha,v} s_x}{\bar{x} - \mu_0} \right)^2$$

note: If the test is two-tailed, then you need to use  $\alpha/2$  rather than  $\alpha$  in the expression above.

note: The LSN is the sample size required to make the observed deviation from  $\mu_0$  statistically significant. Bigger deviations will be significant with smaller sample sizes than the LSN.

Important note: The LSN makes the observed deviation from  $\mu_0$  statistically significant, but the power to detect deviations that large is very low (only 50%) with sample sizes set at the LSN. That is, if there were another deviation just as large as the one you've observed, you would fail to detect it half the time. In fact, if you increase your sample size to the LSN, *there is only a 50-50 chance that the resulting measurements will be statistically significant after all* (see *power*, below; also see Figure 3). To *reliably* detect differences a large as the one you've observed, your sample size needs to be much larger than the LSN.

#### 4. Power

Answers the question: *If the true mean  $\mu$  actually differed from  $\mu_0$  by an amount  $\delta$ , what is the chance that the  $\bar{x}$  that you measure would let you reject  $H_0$  at a significance level of  $\alpha$ ?* In other words, if  $H_0$  is actually false by an amount  $\delta$ , what is the probability of correctly rejecting  $H_0$ ? Power measures how *reliably* you can detect an effect of a given size (see Figure 2).

-Look up  $t_{\alpha,v}$  for your desired significance level, and then calculate the corresponding  $t_{\beta,v}$ :

$$t_{\beta,v} = \frac{\delta}{s_{\bar{x}}} - t_{\alpha,v} = \left( \frac{s_x}{\sqrt{n}} \right) - t_{\alpha,v}$$

-Using a t-table, find the two t values that bracket  $t_{\beta,v}$  for  $v$  degrees of freedom. The "false negative" error rate,  $\beta$ , will be between the significance levels of the t values that bracket  $t_{\beta,v}$ . The corresponding power is  $1-\beta$ .

-If  $\delta$  itself is not statistically significant (that is,  $\delta/s_{\bar{x}} < t_{\alpha,v}$ ), then  $\beta$  is greater than 0.5 (see Figure 4). Estimate  $t_{\beta,v}$  as above (which will yield a negative value), then calculate  $t_{1-\beta,v} = -t_{\beta,v}$ . Then, as above, look up the t-values that bracket  $t_{1-\beta,v}$  (since  $t_{\beta,v}$  is negative,  $t_{1-\beta,v}$  is positive). Now the *power* itself will be between the significance level of the t values bracketing  $t_{1-\beta,v}$ . The "false negative" error rate is  $\beta=1$ -power.

note: If the test is two-tailed, then you need to use  $\alpha/2$  rather than  $\alpha$  in the expression above.

note: For sample sizes above 10 or so, one can use a Z-table instead; Z closely approximates  $t_{\beta,v}$ .

note: t values for estimating power are always one-tailed, whether the significance test itself is one-tailed or two tailed.

Important note: If the consequences of a false negative are dire, and  $H_0$  is not rejected, the *power* of the test may be much more important than its *statistical significance*. Low power means that given the uncertainty in the data, the test is unlikely to detect deviations from the null hypothesis of magnitude  $\delta$ .

To increase the power of a test, you can do two things (see the equation above): you can raise  $\alpha$  (reduce the significance level), or you can increase the sample size ( $n$ ). Also, if the variability in  $x$  ( $s_x$ ) is partly within your control (e.g., from analytical errors), reducing this variability will increase the power of the test. But in most environmental measurements, most of the variability in  $x$  is caused by real-world variation (from time to time, place to place, individual to individual, etc.). There is nothing you can do about this, unless you can find the source of that variability and correct for it in the data (such as with paired-sample testing).

#### 5. Minimum detectable difference

Answers the question: *By how much must the true mean  $\mu$  actually differ from  $\mu_0$ , so that the chance is  $1-\beta$  or better that the  $\bar{x}$  that you measure will let you reject  $H_0$ , at a significance level of  $\alpha$ ?* In other words, by how much must  $H_0$  actually be false, to make the probability  $1-\beta$  that you will correctly reject  $H_0$ ?

-Look up  $t_{\alpha,v}$  and  $t_{\beta,v}$  corresponding to your desired significance level and power. Then calculate (by rearranging the equation above):

$$\delta = (t_{\alpha,v} + t_{\beta,v})s_{\bar{x}} = (t_{\alpha,v} + t_{\beta,v})\frac{s_x}{\sqrt{n}}$$

-The basis for this expression can be seen directly in the geometry shown in Figure 2. The relationship between power and  $\delta$  can also be summarized graphically in a "power curve" (Figure 5).

-If the minimum detectable difference is uncomfortably large, then you can decide to tolerate larger error rates  $\alpha$  and  $\beta$ , or you can increase the sample size ( $n$ ).

#### 6. Sample size (for a specified significance and power)

Answers the question: *How large a sample do you need, so that if the true mean  $\mu$  actually differs from  $\mu_0$  by  $\delta$ , your chance of rejecting  $H_0$  (at significance level  $\alpha$ ) will be  $1-\beta$  or better?*

-Again, by rearranging the equations above to solve for  $n$ , we obtain directly,

$$n = (t_{\alpha,v} + t_{\beta,v})^2 \frac{s_x^2}{\delta^2}$$

-If the required sample size is impractical, you must decide to tolerate larger error rates ( $\alpha$  and  $\beta$ ), or decide that a bigger deviation  $\delta$  from  $H_0$  can go undetected.

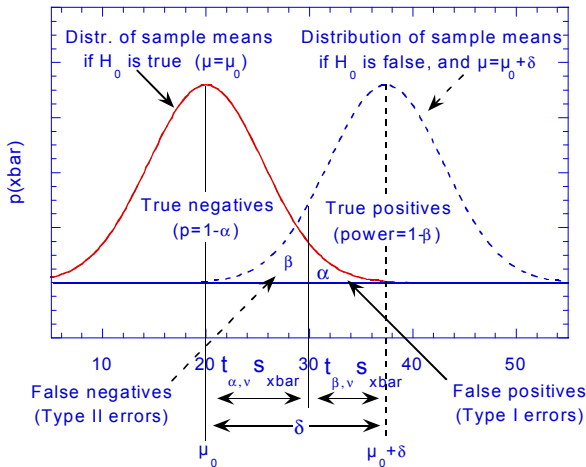


Figure 2. Tradeoff between power ( $1-\beta$ ), significance level ( $\alpha$ ), and minimum detectable difference ( $\delta$ ). Moving the significance threshold right or left shrinks  $\beta$  but increases  $\alpha$ , and vice versa. The only way to decrease  $\alpha$  and  $\beta$  simultaneously is by increasing  $\delta$  (moving the two distributions farther apart) or by increasing  $n$ , thus decreasing  $s_{\bar{x}}$  (making the two distributions narrower)

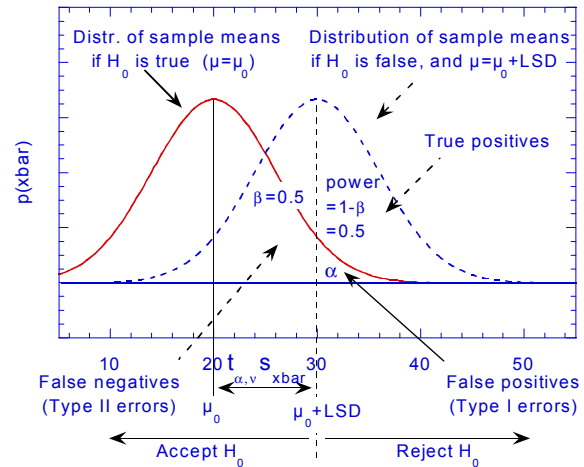


Figure 3. If the true mean  $\mu$  differs from null hypothesis mean  $\mu_0$  by an amount equal to the least significant difference (LSD), then half of all sample means  $\bar{x}$  will be statistically significant (to the right of  $\mu$ ) and the other half will not. Thus, setting sample size by LSN or LSD yields tests with only 50% power.

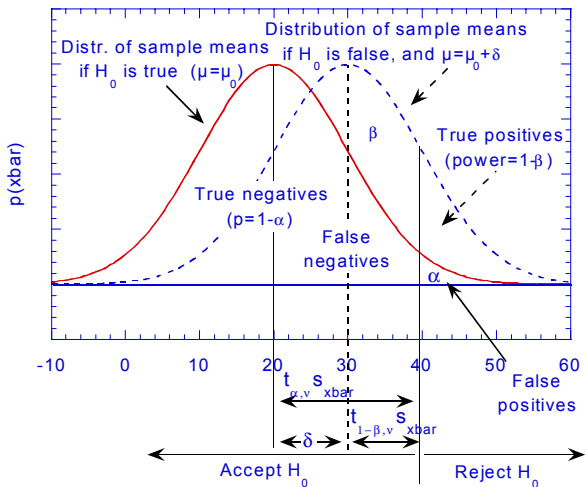


Figure 4. Test with very low power. If true mean  $\mu$  lies within the acceptance region, most of the sample means (distribution shown by dashed curve) will also fall within the acceptance region (false negatives). Few of the sample means will differ enough from  $\mu_0$  to be statistically significant (true positives). High false negative rate makes negative results inconclusive.

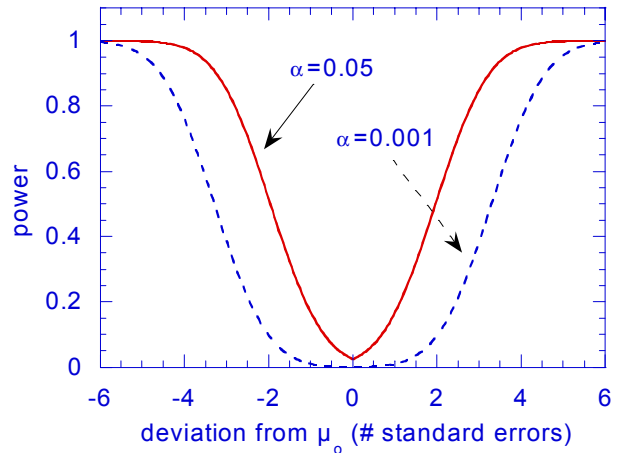


Figure 5. Power curves, showing the power to detect a given deviation ( $\delta$ ) from the null hypothesis, measured in standard errors (that is, the x-axis is  $\delta/s_{\bar{x}}$ ). Note that making significance tests more stringent widens the zone of low power. Curves are calculated under the large-sample approximation (that is, using normal distribution to approximate the t-distribution).