HYPOTHESIS TESTING

Documents prepared for use in course B01.1305,
New York University, Stern School of Business

Cover photo: Yasgur farm, Woodstock, New York

revision date 15 JULY 2009

COMPARISONS BETWEEN HYPOTHESIS TESTS AND JURY DECISION-MAKING

General	Specific Example	Criminal Trial
Null Hypothesis	H_0: $\mu = 28$ (where μ is the unknown mean of some population)	Defendant is innocent
Alternative Hypothesis	H_1: $\mu \neq 28$	Defendant is guilty
Data	Sample $x_1, x_2, \ldots, x_n$	Testimony
Decision mechanism	t test	Jury deliberation
Accept null hypothesis H_0	Decide $\mu = 28$	Acquittal (decide innocent *or* insufficient evidence to convict)
Reject null hypothesis H_0	Decide $\mu \neq 28$	Convict (decide that defendant is guilty)
Type I error	Decide $\mu \neq 28$ when H_0 holds	Decide guilty when defendant is innocent
Type II error	Decide $\mu = 28$ when H_0 is wrong	Decide innocent when defendant is guilty

EXAMPLE 1: A health-care actuary has been investigating the cost of maintaining the cancer patients within its plan. These people have typically been running up costs at the rate of \$1,240 per month. A sample of 15 cases for November (the first 15 for which complete records were available) and an average cost of \$1,080, with a standard deviation of \$180. Is there any evidence of a significant change?

SOLUTION: Let's examine the steps to a standard solution.

Step 1: The hypothesis statement is H_0: μ = \$1,240 versus H_1: $\mu \neq$ \$1,240.

Observe that μ represents the true-but-unknown mean for November. The comparison value \$1,240 is the known traditional value to which you want to compare μ.

Do not be tempted into using H_1: $\mu <$ \$1,240. The value in the data should not prejudicially influence your choice of H_1. Also, you should not attempt to second-guess the researcher's motives; that is, you shouldn't try to create a story that suggests that the researcher was looking for smaller costs. In general, you'd prefer to stay away from one-sided alternative hypotheses.

Step 2: Level of significance $\alpha = 0.05$.

The story gives no suggestion as to the value of α. The choice 0.05 is the standard default.

Step 3: The test statistic will be $t = \sqrt{n}\,\frac{\overline{x} - \mu_0}{s}$. The null hypothesis will be rejected if $|\, t \,| \geq t_{\alpha/2;n-1}$. If $|\, t \,| < t_{\alpha/2;n-1}$ then H_0 will be accepted or judgment will be reserved.

At this point it would be helpful to recognize that the sample size is small; we should state the assumption that the data are sampled from a normal population.

In using this formula, we'll have $n = 15$, μ_0 = \$1,240 (the comparison value), and $\overline{x}$ =\$1,080 and s = \$180 will come from the sample. The value $t_{\alpha/2;n-1}$ is $t_{0.025;14}$ = 2.145.

The "judgment will be reserved" phrase allows for the possibility that you might end up accepting H_0 without really believing H_0. This happens frequently when the sample size is small.

Step 4: Compute $t = \sqrt{15}\,\frac{\$1,080 - \$1,240}{\$180} \approx -3.443$.

Step 5: Since | -3.443 | = 3.443 > 2.145, the null hypothesis is rejected. The November cases are significantly different.

> Plugging in the numbers and reaching the "reject" decision are routine. Observe that we declare a *significant* difference. The word *significant* has jargon status; specifically, it means that a null hypothesis has been rejected.

This discussion did not request a *p*-value. However, we can use the value 3.443 in the *t* table to make a statement. Using the line for 14 degrees of freedom, we find that

$$t_{0.005;14} = 2.977 < 3.443 < 3.787 = t_{0.001;14}$$

we see that H_0 would have been rejected with $\alpha = 0.01$ (for which $\alpha/2 = 0.005$) and would have been accepted with $\alpha = 0.002$ (for which $\alpha/2 = 0.001$). Thus we can make the statement $0.002 < p < 0.01$. Some users might simply write $p < 0.01$ ** , using the ** to denote significance at the 0.01 level.

> You can use Minitab to get more precise *p*-values. Use **Calc** ⇒ **Probability Distributions** ⇒ **t** and then fill in the details
>
> ⊙ **Cumulative probability**
>
> **Degrees of freedom:** 14
>
> **Input constant:** 3.443
>
> Minitab will respond with this:
>
> ```
>            x     P( X <= x )
>       3.4430          0.9980
> ```
>
> The excluded probability to the right is 1 - 0.9980 = 0.0020. The same probability appears below -3.443, so the *p*-value should be given as 0.0040.

Some people simply prefer confidence intervals as a method of summarizing. Here the 95% interval for μ is $\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$, which is $\$1,080 \pm 2.145 \frac{\$180}{\sqrt{15}}$. Numerically this is \$1,080 ± \$100 or (\$980 to \$1,180). It might be noted that the comparison value \$1,240 is outside this interval, consistent with the fact that H_0 was rejected at the 5% level.

EXAMPLE 2: The hourly French fried potato output by the Krisp-o-Matic fry machine is advertised to be 150 pounds. For the new machine purchased by the Burger Heaven drive-in, tests were run for 22 different one-hour periods, producing an average production of 143 pounds, with a standard deviation of 17 pounds. At the 5% level of significance, does the Burger Heaven management have grounds for complaints?

SOLUTION: Here are the steps for this problem.

Step 1: The hypothesis statement is H_0: $\mu = 150$ versus H_1: $\mu \neq 150$.

Observe that μ represents the true-but-unknown mean for the new Krisp-o-Matic machine. The comparison value 150 is the numerical claim, and we want to compare μ to 150.

It might seem that the whole problem was set up with H_1: $\mu < 150$ in mind. After all, the test could not possibly be designed to detect a machine that was performing better than advertised. However, in the absence of a blatant statement that the experiment was designed with a one-sided motive, we should use the two-sided alternative. As before, we should not let the value in the data influence the choice of H_1. Also as before, you should not attempt to second-guess the researcher's motives. In general, we really like to stay away from one-sided alternative hypotheses.

Step 2: Level of significance $\alpha = 0.05$.

The value 0.05 is requested. If the α value were left vague or unspecified, most users would take 0.05 as the default.

Step 3: The test statistic will be $t = \sqrt{n}\,\dfrac{\overline{x} - \mu_0}{s}$. The null hypothesis will be rejected if $|\, t\,| \geq t_{\alpha/2;n-1}$. If $|\, t\,| < t_{\alpha/2;n-1}$ then H_0 will be accepted or judgment will be reserved.

At this point it would be helpful to recognize that the sample size is small; we should state the assumption that the data are sampled from a normal population.

In using this formula, we'll have $n = 22$, $\mu_0 = 150$ (the comparison value). The numbers $\overline{x}$ =143 and $s = 17$ will come from the sample. The value $t_{\alpha/2;n-1}$ is $t_{0.025;21} = 2.080$.

The "judgment will be reserved" phrase allows for the possibility that you might end up accepting H_0 without really believing H_0. This happens frequently when the sample size is small.

Step 4: Compute $t = \sqrt{22}\,\frac{143-150}{17} \approx -1.931$

Step 5: Since | -1.931 | = 1.931 < 2.080, the null hypothesis is accepted. The results are not significant. The Krisp-o-Matic would be declared not significantly different from the claim.

> The phrase *not significant* means that the null hypothesis has been accepted. This does not mean that we really believe H_0 ; we might simply reserve judgment until we get more data.

The p-value would be reported as $p > 0.05$ (NS). The NS stands for not significant.

The 95% confidence interval for μ is $\bar{x} \pm t_{\alpha/2;n-1}\,\frac{s}{\sqrt{n}}$ which is $143 \pm 2.080\,\frac{17}{\sqrt{22}}$.
Numerically this is 143 ± 7.5 or (135.5, 150.5). It might be noted that the comparison value 150 is inside this interval, consistent with the fact that H_0 was accepted at the 5% level.

Finally, let's note what would have happened if someone had insisted that we use H_1: μ < 150. The rejection rule in step 3 would have been to reject H_0 if $t \leq -t_{\alpha;n-1}$ where we use $t_{0.05;21} = 1.721$. Then the action in step 5 would have noted that $t = -1.931$ and, since -1.931 < -1.721, we would have rejected H_0!

Notice that the acceptance or rejection of H_0 can depend on our psychological interpretation of the experimenter's mindset as to whether a one-sided or a two-sided test is intended. This odd situation is certainly an incentive to avoid one-sided tests.

THE *t* TEST THROUGH MINITAB

In comparing two sets of data for the purpose of testing $H_0 : \mu_1 = \mu_2$, the arithmetic can be annoying, and it's useful to be able to use Minitab to perform the labor. We'll examine the data set used in Exercise 9.1 of Hildebrand-Ott-Gray, page 366. This concerns the impurities found in recycled aluminum from two different sources, identified as Source I and Source II. The units of measurement are "kilograms of impurities per hundred kilograms of product"; it seems that it would be simpler just to call this "percent impurities."

The data can be found in file EX0901.MTW, a Hildebrand-Ott-Gray exercise file.

The file looks like this in Minitab (rearranged to save page space):

Impurity	Source	Impurity	Source
3.8	I	1.8	II
3.5	I	2.2	II
4.1	I	1.3	II
2.5	I	5.1	II
3.6	I	4.0	II
4.3	I	4.7	II
2.1	I	3.3	II
2.9	I	4.3	II
3.2	I	4.2	II
3.7	I	2.5	II
2.8	I	5.4	II
2.7	I	4.6	II

There are two ways to lay out two samples of data.

One method would simply put sample 1 in one column (say C1) and put sample 2 in another column (say C2).

Another method puts all the values in a single column, using another column for identify purposes.

The second of these is used for this particular data file. In general, you'll find this to be much more convenient.

It's always helpful to begin work with some simple summaries. Use this:

Stat ⇒ **Basic Statistics** ⇒ **Descriptive Statistics** ⇒

Variables: Impurity

By variable: Source

OK

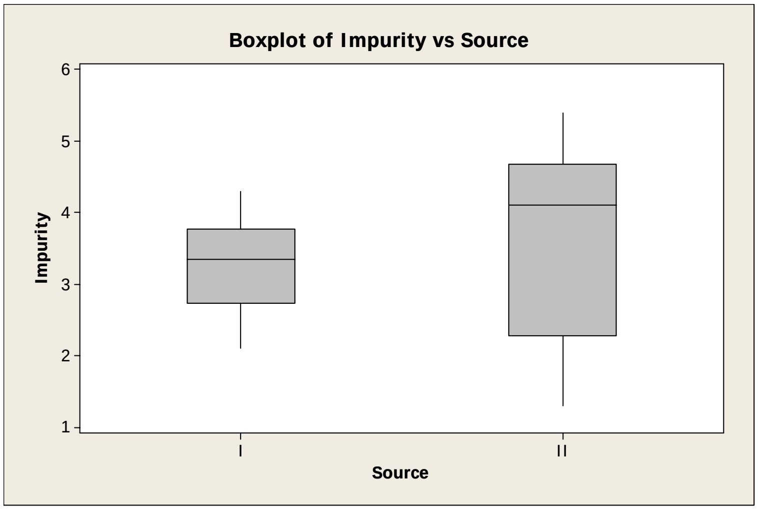
This will produce the following:

Descriptive Statistics: Impurity

```
Variable  Source     N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median
Impurity  I         12   0  3.267    0.195  0.676    2.100  2.725   3.350
          II        12   0  3.617    0.394  1.365    1.300  2.275   4.100

Variable  Source        Q3  Maximum
Impurity  I          3.775    4.300
          II         4.675    5.400
```

We see that the means are very slightly different. The standard deviations, however, are rather far apart. It's easy enough to get side-by-side boxplots by requesting **Graphs** from the **Display Descriptive Statistics** panel.



This picture certainly confirms our previous remark. The means are reasonably close, but the standard deviation is larger from Source II.

In comparing the two groups our major decision is whether we should assume that the standard deviations σ_1 and σ_2 are equal or not. (This is the same as asking whether the variances σ_1^2 and σ_2^2 are equal or not.) We'll do

Stat ⇒ **Basic statistics** ⇒ **2-sample t** ⇒

⊙ **Samples in one column**

Samples: Impurity

Subscripts: Source

⊙ **Assume equal variances**

OK ⇒

The resulting output is this:

Two-Sample T-Test and CI: Impurity, Source

```
Two-sample T for Impurity

Source          N   Mean  StDev  SE Mean
I              12  3.267  0.676     0.20
II             12   3.62   1.37     0.39

Difference = mu (I            ) - mu (II           )
Estimate for difference:  -0.350000
95% CI for difference:  (-1.261935, 0.561935)
T-Test of difference = 0 (vs not =): T-Value = -0.80
                                 P-Value = 0.435  DF = 22
Both use Pooled StDev = 1.0771
```

This repeats some of the previous arithmetic. It does reveal for us, however, that s_p = 1.08 (which is between s_1 = 0.676 and s_2 = 1.37) and also that t = -0.80. This is computed as $t = \dfrac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\dfrac{n_1+n_2}{n_1 n_2}}} = \sqrt{\dfrac{n_1 n_2}{n_1+n_2}}\,\dfrac{\overline{X}_1 - \overline{X}_2}{s_p}$. The statistic has 22 degrees of freedom. We are actually given the p-value as 0.435, so we certainly conclude that H_0 must be accepted.

Suppose that you decide *not* to make the assumption that σ_1 and σ_2 are equal. Just repeat the previous request but do not check the box **Assume equal variances**. This will get

Two-Sample T-Test and CI: Impurity, Source

```
Two-sample T for Impurity

Source          N   Mean  StDev  SE Mean
I              12  3.267  0.676     0.20
II             12   3.62   1.37     0.39


Difference = mu (I            ) - mu (II           )
Estimate for difference:  -0.350000
95% CI for difference:  (-1.282177, 0.582177)
T-Test of difference = 0 (vs not =): T-Value = -0.80
                                 P-Value = 0.438  DF = 16
```

The test statistic is computed as $t = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$, with the degrees of freedom computed as

$$\frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)\left(\dfrac{\dfrac{s_1^2}{n_1}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}\right)^2 + (n_1 - 1)\left(\dfrac{\dfrac{s_2^2}{n_2}}{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}\right)^2}$$, which Minitab truncates to the previous integer, here 16. While this procedure is elaborate, and while there is rather persuasive evidence that $\sigma_1 \neq \sigma_2$, the values produced by *t* are nearly identical (rounded to -0.80 for both) and the *p*-values are nearly identical (0.435 and 0.438).

You might wonder about a test for $H_0 : \sigma_1 = \sigma_2$ versus $H_1 : \sigma_1 \neq \sigma_2$. (We're now using the symbols H_0 and H_1 to refer to the hypotheses about the standard deviations.) Such a test is actually available through the commands

Stat ⇒ ANOVA ⇒ Test for Equal Variances ⇒

Response: Impurity

Factors: Source

OK

This will produce some elaborate output, including a picture:

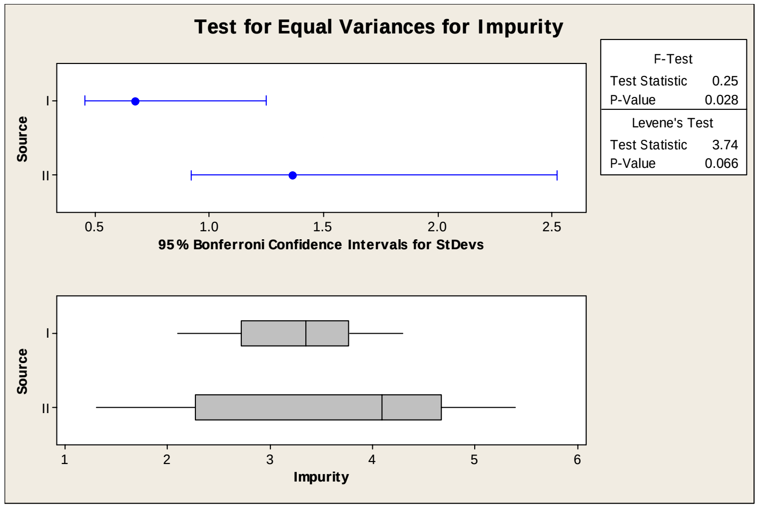
Test for Equal Variances: Impurity versus Source

```
95% Bonferroni confidence intervals for standard deviations

       Source   N     Lower    StDev    Upper
I              12  0.457120  0.67600  1.24958
II             12  0.923064  1.36504  2.52329


F-Test (normal distribution)
Test statistic = 0.25, p-value = 0.028


Levene's Test (any continuous distribution)
Test statistic = 3.74, p-value = 0.066
```



There are many parts of this that can be explored. For simplicity, we can just use the normal distribution test, which produces a *p*-value of 0.028. Formally this means that H_0 would be rejected as the 0.05 level of significance. Thus, it seems that we were wise to run the test to compare μ_1 and μ_2 without assuming equal standard deviations. We should point out however, that this test is highly dependent on the assumption of normal populations; if there is any reason to believe that the populations might be non-normal, then this test is highly unreliable.

In some cases, only one side of μ_0 is interesting.

For instance, results suggesting that really $\mu > \mu_0$ might be valuable while results suggesting that $\mu < \mu_0$ might be worthless. The formulation of the hypotheses is H_0: $\mu = \mu_0$ versus H_1: $\mu > \mu_0$.

This problem is written equivalently as H_0: $\mu \leq \mu_0$ versus H_1: $\mu > \mu_0$.

The procedure now is to reject if $t \geq t_{\alpha;n-1}$.

For instance, if $n = 20$ and $\alpha = 0.05$, then you reject H_0 when $t \geq 1.729$. Observe that no rejection of H_0 occurs for negative t; even $t = -500$ would not cause rejection of H_0.

Below is a parallel version of these statements for problems organized so that values below μ_0 are interesting.

≈≈

For instance, results suggesting that really $\mu < \mu_0$ might be valuable while results suggesting that $\mu > \mu_0$ might be worthless. The formulation of the hypotheses is H_0: $\mu = \mu_0$ versus H_1: $\mu < \mu_0$.

This problem is written equivalently as H_0: $\mu \geq \mu_0$ versus H_1: $\mu < \mu_0$.

The procedure now is to reject if $t \leq -t_{\alpha;n-1}$.

For instance, if $n = 20$ and $\alpha = 0.05$, then you reject H_0 when $t \leq -1.729$. Observe that no rejection of H_0 occurs for positive t; even $t = 500$ would not cause rejection of H_0.

≈≈

It is valuable to note that ***one-tail tests make it easier to reject H_0***, at least when $\bar{x}$ is on the interesting side of μ_0. For instance if $n = 20$, $\alpha = 0.05$, $\mu_0 = 1{,}200$, $\bar{x} = 1{,}260$, $s = 142$, then $t = \sqrt{20}\,\frac{1{,}260 - 1{,}200}{142} \approx 1.89$.

If the problem is H_0: $\mu = 1{,}200$ versus H_1: $\mu \neq 1{,}200$, then the procedure is to reject if $|t| \geq 2.093$. This would cause H_0 to be accepted.

If the problem is H_0: $\mu = 1{,}200$ versus H_1: $\mu > 1{,}200$, then the procedure is to reject if $t \geq 1.729$. This would cause H_0 to be rejected.

Your motives will be questioned whenever you do a one-tail test.

GOOD ADVICE: Always use two-sided tests, except for those cases which are clearly and blatantly one-sided from their description.

If you find yourself looking at the data to decide the form of H_1, then you should be doing a two-tail test.

The following situations are examples in which one-tail tests are definitely appropriate:

(a) Investigating legal limits in situations for which only results on one side of a stated limit mean trouble. These include problems on pollution allowances.

(b) Performing an audit in which only one side (undervaluing, say) is relevant.

If can sometimes be argued that one-sided tests should be used for these cases:

(c) Comparing a new product to a standard product.

(d) Investigating advertising claims or sales pitches about the merits of a product.

Virtually all other tests should be done two-sided. Particular cases to watch out for:

(e) Comparing two products, both of which are already on the market.

(f) Comparing two medical procedures.

(g) *t* tests on regression coefficients.

(h) Investigating scientific claims.

Special arguments can be invoked in almost every instance. If you have any doubts about whether you should be doing a one-tail test or a two-tail test.....then do a two-tail test.

This document takes the broad point of view that *every* test should be done two-sided, with exceptions only for those situations like (a) or (b) or for those situations in which there is a tradition of one-sided tests. Please see the last paragraph for an interesting defense of one-sided tests.

In dealing with hypothesis tests, there are rules about proper procedure. The prime concern is that a glance at the data influences the manner in which the test is conceived and conducted. Specifically....you must not inspect the data before you formulate H_0 and H_1. The concern is that you will decide whether this is a one-sided test or two-sided test after examining the data.

Consider this interesting situation. The Chow City Supermarket chain has just installed Magiceye optical scanning equipment at the checkout counters of its Mayville store. After one month of experimentation, management notes that the Magiceye system gives a checkout rate of 6.45 items per minute. This is computed as the number of line items processed per minute of time that a checkout counter is open, and necessarily involves time for bagging groceries and processing coupons. The manager of the Mayville store asked to try System *M*, an alternate system, and a couple of checkout counters were equiped with System *M*. Over the next few days, using different clerks, System *M* was evaluated for 50 separate one-hour periods. These 50 periods showed an average checkout rate of 6.61 items per minute, with a standard deviation of 0.62. What conclusions should be reached?

At company headquarters, the problem was formulated as $H_0 : \mu = 6.45$ versus $H_1 : \mu \neq 6.45$. (Here μ is the unknown mean for System *M*.) This was done as a two-sided problem because management expressed an interest in either side of 6.45. If System *M* does significantly worse, then the large purchase of Magiceye has been justified. If System *M* does significantly better, then additional installations might use System *M*. The test statistic was computed as $t = \sqrt{50}\,\frac{6.61 - 6.45}{0.62} \approx 1.825$.

The comparison point, using the 5% level of significance, is 2.011.

> This is the value $t_{0.025;49}$. This was obtained by quick interpolation between $t_{0.025;40} = 2.021$ and $t_{0.025;60} = 2.000$. If your t table lacks entries beyond 30 degrees of freedom, you are stuck using $t_{0.025;30} = 2.042$.

Since the t statistic is between -2.011 and +2.011, the null hypothesis is accepted. The conclusion is that there is no significant difference between Magiceye and System *M*.

The results were transmitted back to the manager of the Mayville store. He insisted that he wanted to test System *M* only because it would be an improvement. After all, why would he be interested in a system no better than what he already has? He redid the analysis.

The problem was formulated as $H_0 : \mu = 6.45$ versus alternative $H_1 : \mu > 6.45$. This was done as a one-sided problem because the Mayville manager expected to show that System *M* was superior. The test statistic was computed as $t = \sqrt{50}\,\frac{6.61 - 6.45}{0.62} \approx 1.825$.

The comparison point, using the 5% level of significance, is 1.678.

> This is the value $t_{0.05;49}$, obtained by quick interpolation between $t_{0.05;40} = 1.684$ and $t_{0.05;60} = 1.671$. If your t table lacks entries beyond 30 df, then you must use $t_{0.05;\infty} = 1.645$.

Since the value of t exceeds 1.678, the null hypothesis is rejected. The conclusion is that System *M* is significantly faster than Magiceye.

This example illustrates why many statisticians, especially those at regulatory agencies, are very wary of one-sided tests.

Now consider this example. The Ultra! soft drink company has retooled the chemical filtration step through which its cola drink is passed, in the expectation that the caffeine concentration will be reduced. The target concentration is 125 mg per twelve-ounce can, the current content of its UltraCola drink. The new method is used to produce 96 cans, and the caffeine concentration of each can is measured. The resulting 96 values have a mean concentration of 125.81 mg and a standard deviation of 4.2 mg.

The problem was formulated as a test of $H_0 : \mu = 125$ versus $H_1 : \mu \neq 125$, using μ to represent the true-but-unknown concentration using the retooled filtration step. The t statistic was computed as $t = \sqrt{96}\,\frac{125.81-125}{4.2} \approx 1.890$.

The comparison point, using the 5% level of significance, is $t_{0.025;95} = 1.992$.

> This was obtained by crude interpolation between $t_{0.025;60} = 2.000$ and $t_{0.025;120} = 1.980$. If your t table does not have lines past 30 df, use $t_{0.025;\infty} = 1.96$.

Since the computed value of t, namely 1.890, is between -1.992 and +1.992, the null hypothesis must be accepted. One would conclude that this change has had no significant impact on the caffeine concentration.

The analysis above was shown to a marketing specialist who refused to pass up a potentially interesting finding. He reinterpreted the situation as a desire to make a *higher* caffeine drink in order to tap into the "high-wired" market segment. Accordingly, using the same information, he rewrote the analysis as follows:

The problem was formulated as a test of $H_0 : \mu = 125$ versus $H_1 : \mu > 125$, using μ to represent the true-but-unknown concentration using the retooled filtration step. The t statistic was computed as $t = \sqrt{96}\,\frac{125.81-125}{4.2} \approx 1.890$.

The comparison point, using the 5% level of significance, is $t_{0.05;95} = 1.663$.

> This was obtained by crude interpolation between $t_{0.05;60} = 1.671$ and $t_{0.05;120} = 1.658$. If your t table does not have lines past 30 df, use $t_{0.05;\infty} = 1.645$.

Since the computed value of t, namely 1.890, exceeds 1.663 the null hypothesis must be rejected. One would conclude that this change has had a significant increasing impact on the caffeine concentration.

The marketing specialist recommended that the product be developed under the name UltraVolt Cola. He may have ignored the fact that the typical soda drinker may not be able to distinguish 125 mg from (estimated) 125.81 mg.

Each of these examples illustrates that

> A nonsignificant result can sometimes be made significant by revising the procedure to be one-sided.
>
> A peek at the data can be a powerful influence to the person formulating the hypotheses.
>
> We must specify H_0 and H_1 without a look at the data.

As a practical question, we have to ask what to do if we see the data before we have a chance to specify the hypotheses. After all, there are cases in which the data are presented to you before you have any chance to react! In such a situation, your only reasonable strategy is to try to imagine what you would have done if you understood the experimental layout but had not seen the data. This is tough. It is probably reasonable to recommend a two-sided test at the 5% level. You are almost certainly unable to honorably recommend a one-sided test.

There is an interesting defense of the one-sided methodology. Suppose that you want to compare a new medical procedure to an existing procedure. The experiment needs to be done on human subjects. Experimental protocols must be developed in terms of statistical power, and there must be calculations supporting the notion that the sample size n will be large enough to achieve desired power. The needed value of n will be smaller for a one-sided test, meaning that *fewer subjects will be needed in the experiment*. This means that the knowledge will be obtaining while putting fewer subjects through the inferior medical procedure! This is discussed in “The Ethics of Sample Size: Two-Sided Testing and One-Sided Thinking,” by J. André Knottnerus and Lex M. Bouter, *Journal of Clinical Epidemiology*, vol 54, #2, February 2001, pp 109-110.

⚲ ⚲ ⚲ ⚲ ⚲ EXAMPLE OF A ONE-SIDED TESTING ENVIRONMENT ⚲ ⚲ ⚲ ⚲ ⚲

Hypothesis test example.

Regulations require that the impurity rate in ground flaxseed be 0.6%. That is, 1 kg of ground flaxseed may have no more than 6 g of impurities. You have a new shipment, and you check 15 samples, each of 1 kg. The impurity amounts were determined to have mean 7.2 g, with a standard deviation of 1.0 g. At the 5% level of significance, does this tell us whether this shipment might be impure?

SOLUTION: With $n = 15$, we will make the assumption that the values come from a normal population.

1. H_0: $\mu \leq 6$ g $\qquad$ H_1: $\mu > 6$ g

   Here μ is the true-but-unknown mean of the impurity amounts for this population (the new shipment).

2. $\alpha = 0.05$

3. Test statistic is $t = \sqrt{n}\,\dfrac{\overline{X} - 6}{s}$. We will reject H_0 if $t \geq t_{0.05;14} = 1.761$.

4. Find $t = \sqrt{15}\,\dfrac{7.2 - 6}{1.0} \approx 4.6476$.

5. Reject H_0. We do believe that $\mu > 6$.

In comparing two samples of measured data, there are several possibilities for structuring a test.

	Population standard deviations assumed equal; $\sigma_A = \sigma_B$	Population standard deviations σ_A and σ_B allowed to be different
Sample sizes n_A and n_B both large (30 or more)	Use t test with s_p ; degrees of freedom is n_A+n_B-2 ; as degrees of freedom is large this is approximately normal (whether the populations are assumed normal or not).	Use statistic $\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}}$ which will be approximately normal.
At least one of the sample sizes is small [must assume normal populations]	Use t test with s_p ; degrees of freedom is n_A+n_B-2.	Use statistic $\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}}$ which will approximately be t with approximate degrees of freedom noted at end of this document.

It is **strongly recommended** that you invoke the assumption $\sigma_A = \sigma_B$ unless the data are presenting very convincing evidence that these are unequal. As a quick approximation, you can reasonably believe that $\sigma_A = \sigma_B$ whenever $\frac{1}{2} \le \frac{s_A}{s_B} \le 2$. If you think that $\sigma_A \neq \sigma_B$, then you should seriously consider whether you really want to ask whether $\mu_A = \mu_B$.

Here are a number of illustrations of these calculations.

Two brands of commercial frying fat are to be compared in terms of saturated fat content, and the standard of comparison is expressed in terms of grams of saturated fat per tablespoon. A whole tablespoon contains 20 grams of fat, but only some of that amount is saturated.

Samples are obtained for brands A and B, resulting in the following:

Fat	Number of Samples	Average	Standard Deviation
A	40	6.02	0.86
B	50	6.41	0.90

Test whether or not the two fats have equal amounts of saturated fat. State your conclusion in terms of the p-value.

SOLUTION: Let μ_A and μ_B be the unknown population means. Also let σ_A and σ_B be the standard deviations. We will use $\overline{X}_A - \overline{X}_B$ to estimate and test μ_A - μ_B. It can be shown that the distribution of $\overline{X}_A - \overline{X}_B$ has mean μ_A - μ_B and standard deviation $\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$.
There are now several possible approaches.

Unequal standard deviations with large sample sizes: Use the Central Limit theorem to assert that $\overline{X}_A$ and $\overline{X}_B$ are approximately normally distributed. This will also allow us to conclude that the difference $\overline{X}_A - \overline{X}_B$ is also approximately distributed. We can then estimate the standard deviation of $\overline{X}_A - \overline{X}_B$ with $\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$. Alas, the ratio $\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$ will, for technical reasons, *not* follow a *t* distribution. However, based on the large sample sizes, this ratio will still be approximately normal. Thus, we base the test on the normal distribution. Formally, here are the steps for a test at the 0.05 level.

Step 1: $H_0 : \mu_A = \mu_B$ versus $H_1 : \mu_A \neq \mu_B$

Step 2: $\alpha = 0.05$

Step 3: The test statistic is $Z = \dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$. The null hypothesis will be rejected if $| Z | \geq z_{0.025} = 1.96$.

Step 4: The actual value is $Z = \dfrac{6.02 - 6.41}{\sqrt{\frac{0.86^2}{40} + \frac{0.90^2}{50}}} = \dfrac{-0.39}{0.1863} \approx -2.093$

Step 5: Since | -2.093 | > 1.96, the null hypothesis must be rejected.

The results would be deemed significant, and we can declare that the fats are significantly different.

Equal standard deviations with large sample sizes: We make one additional assumption here. We now assume that the population standard deviations are equal, meaning $\sigma_A = \sigma_B$. Use the symbol σ for the common value. In this case, the standard deviation of $\overline{X}_A - \overline{X}_B$ is now $\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}} = \sigma\sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = \sigma\sqrt{\frac{n_A + n_B}{n_A\, n_B}}$. The recommended estimate for σ is the *pooled standard deviation* s_p, defined through $s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}$. The test will be based on the statistic $\frac{\overline{X}_A - \overline{X}_B}{s_p\sqrt{\frac{n_A + n_B}{n_A n_B}}} = \sqrt{\frac{n_A n_B}{n_A + n_B}}\,\frac{\overline{X}_A - \overline{X}_B}{s_p}$. With large sample sizes, we need not make further detailed assumptions. However,

> if we make no further assumptions about the distributions of the *A* and *B* populations, the statistic will follow, approximately, a normal distribution.
>
> if we assume that the *A* and *B* populations themselves follow normal distributions, the statistic will follow *exactly* a *t* distribution with $n_A + n_B - 2$ degrees of freedom.

The *t* distribution with a large number of degrees of freedom is so close to the normal that it is not necessary to agonize about these final assumptions.

In executing the test, we change the work in steps 3, 4, and 5. This is now the following:

Step 3: The test statistic is $\frac{\overline{X}_A - \overline{X}_B}{s_p\sqrt{\frac{n_A + n_B}{n_A n_B}}} = \sqrt{\frac{n_A n_B}{n_A + n_B}}\,\frac{\overline{X}_A - \overline{X}_B}{s_p}$.

If we do not assume normal distributions for the *A* and *B* populations, then we call this statistic *Z* and we reject H_0 if $|Z| \geq 1.96$.

If we assume normal distributions for the *A* and *B* populations, then we call this statistic *t* with $n_A + n_B - 2 = 88$ degrees of freedom and we reject H_0 if $|t| \geq t_{0.025;88} = 1.9873$.

Step 4: Begin by finding $s_p^2 = \frac{(40-1)0.86^2 + (50-1)0.90^2}{40+50-2} \approx 0.7788$. This leads to $s_p \approx 0.8825$. The value of the test statistic is then $\sqrt{\frac{40 \cdot 50}{40+50}}\,\frac{-0.39}{0.8825} \approx -2.083$.

Step 5: Whether we made the assumptions leading to Z or to t, the null hypotheses H_0 would be rejected.

You might notice that the values of Z and t are very close. This is usually the case.

This example had quite a lot of data, 90 observations in all. Suppose, hypothetically, that we had acquired the same data with smaller samples:

Fat	Number of Samples	Average	Standard Deviation
A	12	6.02	0.86
B	15	6.41	0.90

In this case we would not be able to invoke the Central Limit theorem, so we *must* take as an assumption that the data are independent samples from two normal populations. We simply can't deal with small sample sizes unless we make this assumption.

There is a now a question as to whether we should use the assumption that the two normal populations have the same standard deviation. Let's first make this assumption. It certainly looks reasonable, given the numeric values for s_A and s_B.

Equal standard deviations with small sample sizes: Proceed as follows.

Step 1: $H_0 : \mu_A = \mu_B$ versus $H_1 : \mu_A \neq \mu_B$

Step 2: $\alpha = 0.05$

Step 3: The test statistic is $\dfrac{\overline{X}_A - \overline{X}_B}{s_p \sqrt{\dfrac{n_A + n_B}{n_A n_B}}} = \sqrt{\dfrac{n_A n_B}{n_A + n_B}}\, \dfrac{\overline{X}_A - \overline{X}_B}{s_p}$. This statistic is t with $n_A + n_B - 2 = 25$ degrees of freedom and we reject H_0 if $|\, t \,| \geq t_{0.025;25}$ = 2.0595.

Step 4: Begin by finding $s_p^2 = \dfrac{(12-1)0.86^2 + (15-1)0.90^2}{12+15-2} \approx 0.7754$. This leads to $s_p \approx 0.8806$. The value of the test statistic is then $\sqrt{\dfrac{12 \cdot 15}{12+15}}\, \dfrac{-0.39}{0.8806} \approx -1.144$.

Step 5: The null hypotheses H_0 must be accepted.

Unequal standard deviations with small sample sizes: Finally, we could repeat this without making the assumption that the population standard deviations are the same. This would lead to the statistic $\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}}$. In the case in which we are willing to assume normal distributions with *un*equal standard deviations, this statistic will have an approximate *t* distribution. There is even an approximate degrees of freedom calculation for this:

$$\frac{(n_A - 1)(n_B - 1)}{(n_B - 1)\left(\dfrac{\dfrac{s_A^2}{n_A}}{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}\right)^2 + (n_A - 1)\left(\dfrac{\dfrac{s_B^2}{n_B}}{\dfrac{s_A^2}{n_A} + \dfrac{s_B^2}{n_B}}\right)^2}$$

This expression appears in item **3** at the top of page 363 of Hildebrand, Ott, and Gray. This number is frequently truncated to the previous integer.

The calculation of the test statistic (for step 4) would be $\dfrac{6.02 - 6.41}{\sqrt{\dfrac{0.86^2}{12} + \dfrac{0.90^2}{15}}} = \dfrac{-0.39}{0.3400}$ ≈ -1.147. This would again lead to acceptance of H_0.

The conclusion is of course influenced heavily by the sample sizes. However, the conclusion is quite robust to changes in the assumptions. This is fortunate!

Suppose that you wish to compare μ_X and μ_Y , and that your data consists of the two independent samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$. Minitab will do all the messy arithmetic of the previous section. Data can be given to Minitab in three ways:

* The X-values and Y-values appear in a single column (of length m+n). Another column notes the group identities. Here is data in this form:

↓	C1	C2
	Wght	Group
1	13.6	1
2	13.8	1
3	17.6	2
4	15.0	1
5	16.8	2
6	17.1	1
7	17.4	2
8	14.2	1

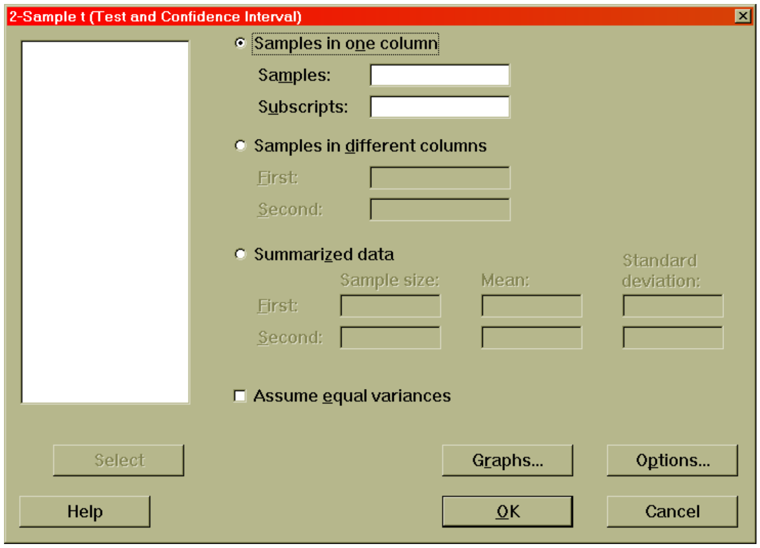
  The X and Y values may be interspersed, as they are here. The identifying variable can be numeric, and here we use 1 ⇔ X and 2 ⇔ Y. The identifying variable can also be alphabetic.

* The X-values appear in one column (of length m) and the Y-values appear in another column (of length n). Here are the same data in this form:

↓	C1	C2
	X	Y
1	14.2	17.4
2	15.0	16.8
3	13.8	17.6
4	17.1	
5	13.6	

* The information can be presented in Minitab 14 in summarized form, giving just the sample sizes (m and n), the means ($\overline{X}$ and $\overline{Y}$), and the sample standard deviations (s_X and s_Y). Textbook problems are often presented in this form, as were all the examples of the previous section.

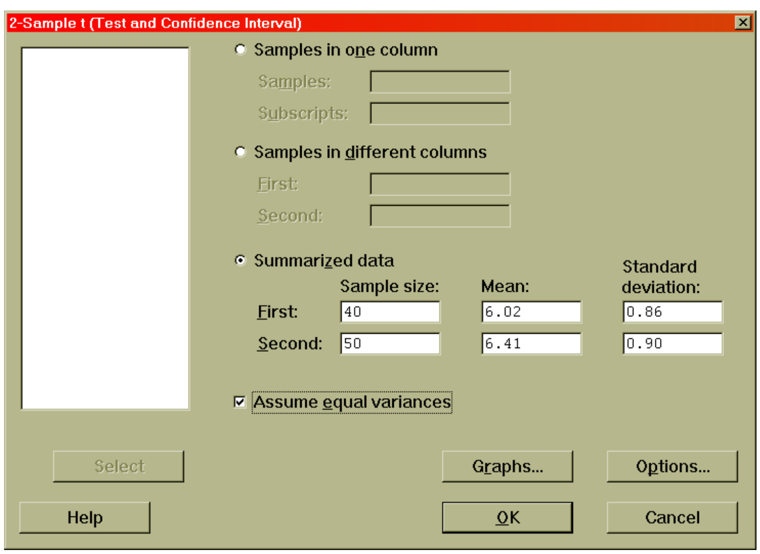
Use the sequence **Stat** ⇒ **Basic Statistics** ⇒ **2-Sample t**. You will be given the immediate choice as to the organization of your data:



Consider, for example, the donut fat example of the previous section. The data details were given in summarized form as

Fat	Number of Samples	Average	Standard Deviation
A	40	6.02	0.86
B	50	6.41	0.90

We present this to Minitab as follows:



Observe that we have checked the box **Assume equal variances**, as seems reasonable (but see the discussion below).

Minitab's results are these:

Two-Sample T-Test and CI

```
Sample   N   Mean  StDev  SE Mean
1       40  6.020  0.860     0.14
2       50  6.410  0.900     0.13

Difference = mu (1) - mu (2)
Estimate for difference:  -0.390000
95% CI for difference:  (-0.762032, -0.017968)
T-Test of difference = 0 (vs not =): T-Value = -2.08
       P-Value = 0.040  DF = 88
Both use Pooled StDev = 0.8825
```

If you use the equal variances choice, you are making these assumptions that $\sigma_X = \sigma_Y$ and also that the samples come from normal populations. The resulting distribution for the test statistic is t with $m + n - 2$ degrees of freedom.

> If both of m and n are large (meaning 30 or more), you may relax the assumption about normal populations. Technically, the distribution of the test statistic is then approximately normal, but t with a large number of degrees of freedom is very close to normal.

The confidence interval given here is $(\bar{X} - \bar{Y}) \pm t_{\alpha/2;m+n-2}\, s_p \sqrt{\frac{m+n}{mn}}$.

Suppose that, out of curiousity, you removed the equal variances assumption. The results would be these:

Two-Sample T-Test and CI

```
Sample   N   Mean  StDev  SE Mean
1       40  6.020  0.860     0.14
2       50  6.410  0.900     0.13


Difference = mu (1) - mu (2)
Estimate for difference:  -0.390000
95% CI for difference:  (-0.760320, -0.019680)
T-Test of difference = 0 (vs not =): T-Value = -2.09
       P-Value = 0.039  DF = 85
```

The inferential conclusions are nearly identical. The p-values are very close, and the confidence intervals are very similar.

If you do not use the equal variances choice, you are allowing that $\sigma_X \neq \sigma_Y$, but you are still assuming that the samples come from normal populations. The resulting distribution for the test statistic is approximately t , and Minitab reports the degrees of freedom as λ, the integer just below

$$\frac{(m-1)(n-1)}{(n-1)\left(\dfrac{\frac{s_X^2}{m}}{\frac{s_X^2}{m}+\frac{s_Y^2}{n}}\right)^2 + (m-1)\left(\dfrac{\frac{s_Y^2}{n}}{\frac{s_X^2}{m}+\frac{s_Y^2}{n}}\right)^2}$$

If both of m and n are large (meaning 30 or more), you may relax the assumption about normal populations. The distribution of the test statistic is then approximately normal, and Minitab should have labeled it as Z. Of course, t with a large number of degrees of freedom is very close to normal.

The confidence interval given when you allow $\sigma_X \neq \sigma_Y$ is $(\overline{X} - \overline{Y}) \pm t_{\alpha/2;\lambda} \sqrt{\frac{s_x^2}{m}+\frac{s_y^2}{n}}$, where λ is the degrees of freedom value described just above.

The substantive conclusions rarely depend on whether or not you check the box **Assume equal variances**. The only time this will matter is for data with

m and n very unequal (say one is three times the other)

and also

s_X and s_Y very unequal (say one is three times the other)

You can formally investigate the appropriateness of the assumption $\sigma_X = \sigma_Y$ through **Stat ⇒ Basic Statistics ⇒ 2 Variances**.

The two-group comparison discussed in the section "Comparing the means of two groups" gets confusing because there are two different forms for the test.

If we are willing to assume $\sigma_A = \sigma_B$, we use $\sqrt{\frac{mn}{m+n}}\ \frac{\overline{X}_A - \overline{X}_B}{s_p}$. The null distribution of this statistic is t_{m+n-2} [*] .

If we wish to allow $\sigma_A \neq \sigma_B$, we use $\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\dfrac{s_A^2}{m} + \dfrac{s_B^2}{n}}}$. Depending on nuanced assumptions, this distribution is either described as approximately normal [*] or as approximately *t* with degrees of freedom [*] given on page 27.

In any case, the three target distributions noted with [*] are very similar. If we are going to ask whether any of this matters at all, we need to ask when (if ever) the two versions of the computation lead to very different answers.

The two forms are

$$\sqrt{\frac{mn}{m+n}}\ \frac{\overline{X}_A - \overline{X}_B}{s_p} \quad \text{and} \quad \frac{\overline{X}_A - \overline{X}_B}{\sqrt{\dfrac{s_A^2}{m} + \dfrac{s_B^2}{n}}}$$

Let's look at the ratio:

$$\frac{\sqrt{\dfrac{mn}{m+n}}\ \dfrac{\overline{X}_A - \overline{X}_B}{s_p}}{\dfrac{\overline{X}_A - \overline{X}_B}{\sqrt{\dfrac{s_A^2}{m} + \dfrac{s_B^2}{n}}}} = \frac{\sqrt{\dfrac{mn}{m+n}}\ \dfrac{1}{s_p}}{\dfrac{1}{\sqrt{\dfrac{s_A^2}{m} + \dfrac{s_B^2}{n}}}} = = \sqrt{\frac{s_A^2}{m} + \frac{s_B^2}{n}} \times \sqrt{\frac{mn}{m+n}} \times \frac{1}{s_p}$$

$$= \sqrt{\frac{ns_A^2 + ms_B^2}{mn}} \times \sqrt{\frac{mn}{m+n}} \times \frac{1}{\sqrt{\dfrac{(m-1)s_A^2 + (n-1)s_B^2}{m+n-2}}}$$

$$= \sqrt{\frac{ns_A^2 + ms_B^2}{(m-1)s_A^2 + (n-1)s_B^2}} \times \sqrt{\frac{m+n-2}{m+n}}$$

The second square root is very close to 1. So the item here that *really* matters is under the first square root. This is

$$\frac{ns_A^2 \;+\; ms_B^2}{(m-1)s_A^2+(n-1)s_B^2} \;=\; \frac{n \;+\; m\,\dfrac{s_B^2}{s_A^2}}{(m-1) \;+\; (n-1)\,\dfrac{s_B^2}{s_A^2}} \;\approx\; \frac{n \;+\; m\,\dfrac{s_B^2}{s_A^2}}{m \;+\; n\,\dfrac{s_B^2}{s_A^2}} \;=\; \frac{1 \;+\; \dfrac{m}{n}\times\dfrac{s_B^2}{s_A^2}}{\dfrac{m}{n} \;+\; \dfrac{s_B^2}{s_A^2}}$$

This looks like $\dfrac{1 \;+\; ab}{a \;+\; b}$, where $a = \dfrac{m}{n}$ is the sample size ratio and $b = \dfrac{s_B^2}{s_A^2}$ is the variance ratio.

Observe now the following:

If a is near 1 (roughly equal sample sizes) then $\dfrac{1 \;+\; ab}{a \;+\; b}$ is near 1.

If b is near 1 (roughly equal standard deviations) then $\dfrac{1 \;+\; ab}{a \;+\; b}$ is near 1.

This leads to this very interesting conclusion:

The two forms of the test statistic, meaning $\sqrt{\dfrac{mn}{m+n}}\,\dfrac{\overline{X}_A-\overline{X}_B}{s_p}$ and $\dfrac{\overline{X}_A \;-\; \overline{X}_B}{\sqrt{\dfrac{s_A^2}{m}+\dfrac{s_B^2}{n}}}$,

can be very different *only* when *both*

the sample sizes are very unequal

and also

the sample standard deviations are very unequal.

It's hard to establish a threshold for concern, but certainly we should be wary if

the larger sample size is more than three times the smaller sample size

and

the larger sample variance is more than three times the smaller sample variance

Remember, of course, that the variance is the square of the standard deviation.

We can now give a clean summary of all the tests and confidence intervals for the common one-sample and two-sample problems.

Data	Assumptions	Hypotheses	Test Statistic (confidence interval)
$X_1, X_2, \ldots, X_n$	Sample from continuous population with mean μ; must assume normal if $n < 30$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$	$t_{n-1} = \sqrt{n}\,\dfrac{\overline{X} - \mu_0}{s}$ (confidence interval for μ is $\overline{X} \pm t_{\alpha/2;n-1} \dfrac{s}{\sqrt{n}}$)
$X_1, X_2, \ldots, X_m$ $Y_1, Y_2, \ldots, Y_n$	Independent samples from continuous populations with mean μ_x and μ_y and common standard deviation σ; must assume normal if either $m < 30$ or $n < 30$	$H_0: \mu_x = \mu_y$ $H_1: \mu_x \neq \mu_y$	$t_{m+n-2} = \sqrt{\dfrac{mn}{m+n}}\,\dfrac{\overline{X} - \overline{Y}}{s_p}$ (confidence interval for $\mu_x - \mu_y$ is $(\overline{X} - \overline{Y}) \pm t_{\alpha/2;m+n-2}\, s_p \sqrt{\dfrac{m+n}{mn}}$)
$X_1, X_2, \ldots, X_m$ $Y_1, Y_2, \ldots, Y_n$	Independent samples from continuous populations with mean μ_x and μ_y and possibly unequal standard deviations σ_x and σ_y; must assume normal if either $m < 30$ or $n < 30$	$H_0: \mu_x = \mu_y$ $H_1: \mu_x \neq \mu_y$	$Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\dfrac{s_x^2}{m} + \dfrac{s_y^2}{n}}}$; when normal distributions are assumed this can be described as an approximate t (confidence interval for $\mu_x - \mu_y$ is $(\overline{X} - \overline{Y}) \pm z_{\alpha/2} \sqrt{\dfrac{s_x^2}{m} + \dfrac{s_y^2}{n}}$)
$(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$	Paired data, with the differences $D_i = Y_i - X_i$ being a sample from a population with mean μ_D; must assume normal if $n < 30$	$H_0: \mu_D = 0$ $H_1: \mu_D \neq 0$	$t_{n-1} = \sqrt{n}\,\dfrac{\overline{D}}{s_D}$ (confidence interval for $\mu_x - \mu_y = \mu_D$ is $\overline{D} \pm t_{\alpha/2;n-1} \dfrac{s_D}{\sqrt{n}}$)

Data	Assumptions	Hypotheses	Test Statistic (confidence interval)
$X \sim \text{Bin}(n, p)$	Binomial with $n \geq 30$	$H_0: p = p_0$ $H_1: p \neq p_0$	$Z = \sqrt{n}\, \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}}$ confidence interval for p is $\tilde{p} \pm z_{\alpha/2} \sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{n}}$
$X \sim \text{Bin}(m, p_x)$ $Y \sim \text{Bin}(n, p_y)$	Independent binomials with $m \geq 30$ and $n \geq 30$	$H_0: p_x = p_y$ $H_1: p_x \neq p_y$	χ^2 (confidence interval for parameter function $\dfrac{p_x/(1-p_x)}{p_y/(1-p_y)}$ requires special methods)

In the first four rows of this table, the confidence intervals and tests correspond exactly. This means that (for the first row)

H_0: $\mu = \mu_0$ will be accepted at level α

if and only if

value μ_0 is inside confidence interval $\overline{X} \pm t_{\alpha/2;n-1} \dfrac{s}{\sqrt{n}}$

This is routine to show.

Observe that H_0 is accepted if and only if

$$\sqrt{n}\, \frac{\left|\overline{X} - \mu_0\right|}{s} < t_{\alpha/2;n-1}$$

Equivalently, this can be stated as

$$\left|\overline{X} - \mu_0\right| < t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

We could also say this as

$$-\, t_{\alpha/2;n-1} \frac{s}{\sqrt{n}} < \overline{X} - \mu_0 < t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

This last inequality can be rearranged as

$$\overline{X} - t_{\alpha/2;n-1} \frac{s}{\sqrt{n}} < \mu_0 < \overline{X} + t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$$

This is precisely the condition that μ_0 is inside the confidence interval.

Now let's deal with the binomial random variable X with n and π. In general we don't know π, so we use the estimate $\hat{\pi} = \frac{X}{n}$. We also noted that SE($\hat{\pi}$) = $\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$. We had the 1-$\alpha$ confidence interval by the Agrest-Coull method.

Now let's consider a test of the null hypothesis H_0: $\pi = \pi_0$, where π_0 is some specified comparison value. The alternative will be H_1: $\pi \neq \pi_0$. If you like one-sided tests, then you can modify all this stuff in the obvious way. If H_0 is true, then the SD of $\hat{\pi}$ is $\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$. This is *not* a standard error. If H_0 holds, then we don't have to estimate SD($\hat{\pi}$). Thus it follows, if the sample size is reasonably large, that the distribution of $Z = \frac{\hat{\pi} - \pi_0}{\mathrm{SD}(\hat{\pi})}$ is approximately standard normal. This leads to the test based on Z.

To what uses can we put hypothesis tests? This is an interesting question, because we often have alternate ways of dealing with data.

We could use our data to test the null hypothesis H_0: $\theta = \theta_0$ against an alternative H_1.

> Here θ is the true-but-unknown parameter, and θ_0 is a specified comparison value. In most cases θ_0 is an obvious baseline value (zero for a regression coefficient, one for a risk ratio, zero for a product difference, and so on).
>
> The alternative could be H_1: $\theta \neq \theta_0$, which is called a two-sided (or two-tailed) alternative. In many cases H_1: $\theta > \theta_0$ because we are interested only in θ-values which are larger than θ_0 . This is called a one-sided (or one-tailed) alternative. There are also cases H_1: $\theta < \theta_0$ because we are interested only in θ-values which are smaller than θ_0 ; these cases are also called one-sided.
>
>> In most cases, the two-sided version of H_1 is preferred, unless there is obvious *a priori* interest in a one-sided statement. This formalization has to be part of the investigation protocol. It is considered unacceptable to specify H_1 after an examination of the data.

The most obvious competitor for a hypothesis test is a confidence interval. This is a statement of the form "I am 95% confident that the true-but-unknown value of θ is in the interval 38.5 ± 8.4." In any application, this interval is given numerically, but you will encounter algebraic forms such as $\bar{x} \pm t_{\alpha/2;n-1} \frac{s}{\sqrt{n}}$.

The hypothesis test seems to make a yes-or-no decision about H_0 , while the confidence interval makes a data-based suggestion as to the location of θ.

It is important to note that either of these methods could be in error.

>> The confidence interval might not include the true value of θ. If you routinely use 95% confidence intervals, then in the long run about 5% of your intervals will not contain the target value. This is understood, and it's implicit in the notion of 95% confidence.

> The hypothesis test might lead to a wrong decision.
>
> (1) If H_0 is correct and you end up rejecting H_0, then you have made a Type I error. From a statistical point of view, we try to control this error, and tests use the notion of *level of significance* as an upper bound on the probability of Type I error. This upper bound is usually called α, and its value is most often 0.05. We design the test so that
>
> $$P[\text{ reject } H_0 \mid H_0 \text{ true }] \leq \alpha$$

Statisticians are very much aware of this type of error, and some are reluctant to utter the phrase "I reject the null hypothesis." These people will use phrases like "the results are statistically significant."

In the legal comparison, a Type I error corresponds to finding the defendant guilty when in fact the defendant is innocent. The law certainly finds $\alpha = 0.05$ too high for use in a criminal trial, but the 0.05 standard can be used in relation to monetary awards

(2) If H_0 is incorrect and you end up accepting H_0 , then you have made a Type II error. As the hypothesis testing game has been set up, it is very hard to give numbers for Type II error. This happens because there are many ways for H_0 to be false.

Suppose that you are testing H_0: $\theta = 400$ versus H_1: $\theta \neq 400$ and you have a sample of $n = 40$ data points. You are very unlikely to make a Type II error if the true value of θ is 900, but you have a large probability of Type II error if the true value of θ is 402.

Most hypothesis tests operate so that the probability of Type II error drops as n grows. A sample of size $n = 50$ is better than a sample of size $n = 40$, whether θ is 900 or 402.

Statisticians are aware of this type of error as well, and some do not like to say "I accept the null hypothesis." Alternate phrases are "the results are not statistically significant," "I cannot reject the null hypothesis," or "I reserve judgment."

So how do hypothesis tests get used?

Some situations call for a clear accept-or-reject action. We might have to decide which make of photocopier to use when the current contract expires. We might have to decide whether an environmentally sensitive lake should be opened for sport fishing for the next season. These situations require a careful evaluation of the costs and an appreciation for the consequences of Type I and Type II errors. Legal decisions are of course in this context.

Some situations ask for an opinion about whether a relationship exists. For example, in a regression of Y on X, using the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, it's common to test the null hypothesis $H_0 : \beta_1 = 0$ against H_1: $\beta_1 \neq 0$. We do this to see if there is a relationship between X and Y. It's possible that no actions will be associated with the decisions, and we're doing all this just to satisfy our curiosity.

If X is a policy variable and if Y is a consequence, the result of the hypothesis test might decide our future actions. Even so, it might be much more useful to have a confidence interval for β_1, a quantitative assessment of the relationship.

As noted above, the probability of making a Type II error depends on the true value of θ and depends also on the sample size n. The following symbols and expressions are used in describing items related to Type II error:

β = P[Type II error]	Beta depends on the true value of θ and depends also on n.
$1 - \beta$	This is the *power* of the test. It depends on the true value of θ and also on n.
$\beta(\theta)$ = power curve	This assumes that the sample size n has been fixed, and it gives the probability of rejecting H_0 as a function of θ.

The function $1 - \beta(\theta)$ is called the operating characteristic curve, or the OC curve. (This terminology is not universal.)

Pictures of the power curve are very interesting. Suppose that we have a single sample $x_1, x_2, \ldots, x_n$ and we wish to test the null hypothesis H_0: θ = population mean = 400 versus alternative H_1: $\theta \neq 400$. The conventional statistical procedure is the t test, based on the statistic

$$t = \sqrt{n}\,\frac{\overline{x} - 400}{s}$$

Here $\overline{x}$ represents the sample mean and s represents the sample standard deviation. This kind of test is usually done at the 0.05 significance level. This means, to an excellent approximation, that

we will be led to accept H_0 if $-2 \leq t \leq 2$

we will be led to reject H_0 if $| t | > 2$

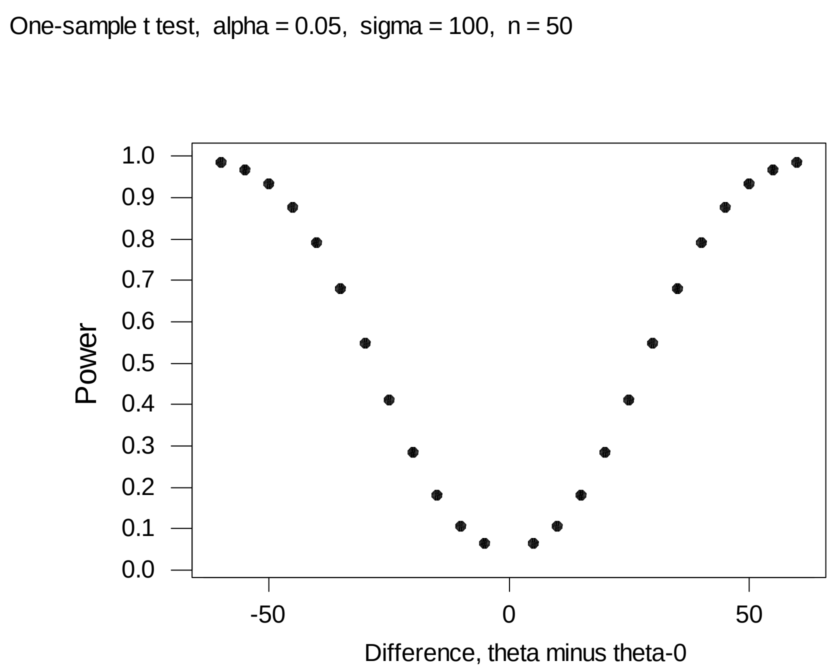
Suppose that the data produce $\overline{x}$ = 410 and s = 98 with a sample of n = 50. The value of the t statistic would be $\sqrt{50}\,\frac{410 - 400}{98} \approx 0.72$. This value would lead to accepting H_0 ; the data are not able to convince us that the population mean θ is not 400. The departure of the data value $\overline{x}$ from the target value 400 is simply not shocking.

Suppose though that we produced the same $\overline{x}$ = 410 and the same s = 98, but we did this with a sample of n = 500. This time the t statistic would be $\sqrt{500}\,\frac{410 - 400}{98} \approx 2.28$ and we would be led to rejecting H_0 .

Perhaps the population mean θ really is near 410. This is a short distance from 400, in light of the size of the standard deviation. The larger sample size gives us the ability to make a smaller difference significant.

It can be proved that, if H_0 is correct, then the probability of incorrectly rejecting H_0 (thus committing Type I error) will be exactly 0.05 *no matter what the sample size*.

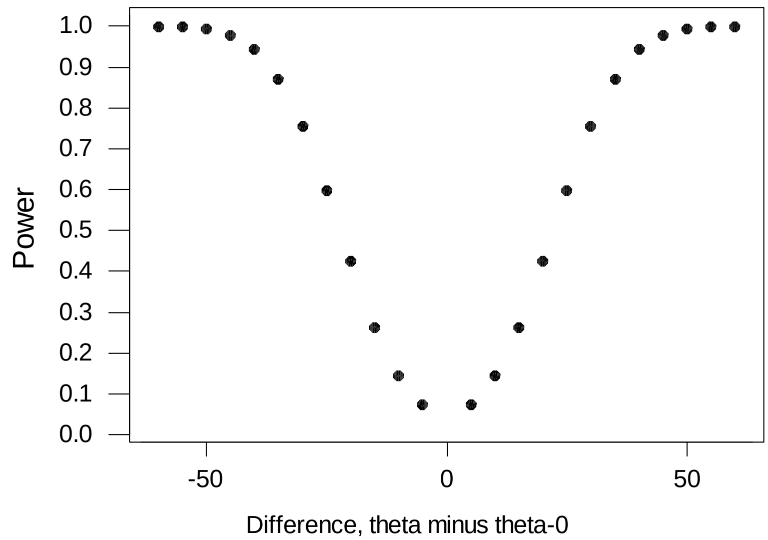
The behavior of this procedure when H_0 is false can be seen best by examining the power curve. Suppose that the population standard deviation is σ = 100. This neatly matches the sample standard deviation s = 98 in the example. The graph below gives the power, meaning the probability of rejecting H_0. The sample size has been fixed here at n = 50.



If the difference between the true θ and the comparison value 400 is about 10, and if the true standard deviation of the population is 100, there is only about a 15% probability of rejecting H_0 . Here's another way to say this: if H_0 is false with the actual θ being 410, then this procedure has only a 15% chance of doing the right thing. This test has low power when the alternative value is 410.

The next picture deals with exactly the same problem, except that a sample size of n = 80 is used.

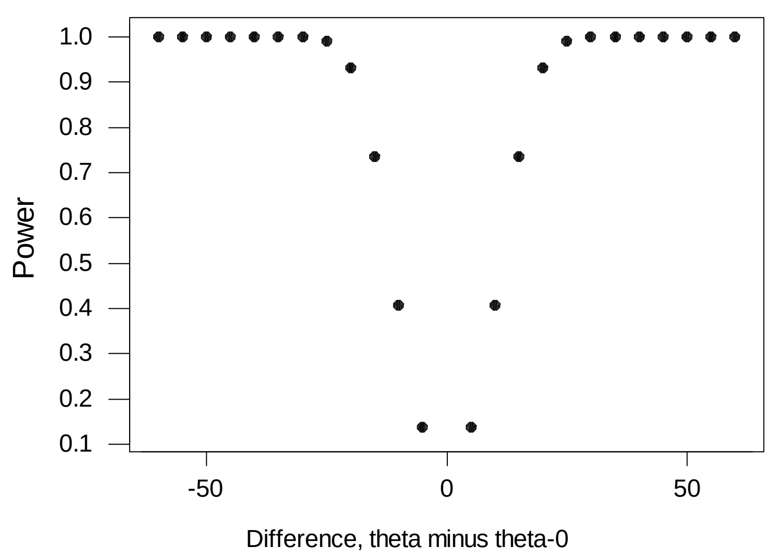
One-sample t test, alpha = 0.05, sigma = 100, n = 80



This graph has a narrower valley, meaning that the power is greater at every θ value.

With a sample size of n = 300, this is the power curve:

One-sample t test, alpha = 0.05, n = 300



All curves have their bottoms at the point (0, 0.05), corresponding to the 0.05 probability of rejecting H_0 when it is true.

So here are some useful points:

* The power is greater if θ is far from the comparison value specified by the null hypothesis.
* The power is greater if the sample size n is larger.
* If you want to accept H_0 you should try to use a small sample size.
* If you want to reject H_0 you should try to use a large sample size.

Suppose that the problem is to test the null hypothesis H_0: θ = 400 versus alternative H_1: θ ≠ 400. Suppose that, in advance of collecting the data, you believe that the true value of θ is near 410 and that the population standard deviation is near 100.

> If you would really like to see that H_0 is accepted, then you would look at the first picture above and note that n = 50 is very likely to lead to accepting H_0 . You will recommend a sample of size 50.
>
> If you would like to see that H_0 is rejected, then you would look at a series (varying over n) of pictures like those here until you get one with high power when θ is near 410 and the standard deviation is near 100. You will recommend a sample of size 500. Technically, there's a formula that can be used for this purpose, so you do not have to create and examine all these graphs.

There is an obvious tension here. This tension will be resolved in a legal sense by the requirement that the statistical procedure have enough power to detect reasonably interesting alternative hypotheses. This still leaves plenty of room to haggle. After all, we have to decide what "reasonably interesting" means. In the example given here, θ = 410 (relative to σ = 100) would probably not be deemed reasonably interesting, as it is only $\frac{410 - 400}{100} = 0.1$ of a standard deviation off the null hypothesis. Certainly θ = 450 would be considered reasonably interesting.