This calculation is, of course, an approximation as the name suggests. Note that it does not usually give nice integer degrees of freedom, expect some decimal places. This is not an issue for computer programs that can get P-values for any d.f. It does complicate using our tables a little.

There is one additional "simplification". We know that the d.f. are at least the smaller of n_1-1 and n_2-1. But what if $n_1 = n_2 = n$? In this case the d.f. will be at least $n-1$. However, Satterthwaite's approximation will still, usually, yield a larger d.f.

Summary

There are two cases in two-sample t-tests. The case where $\sigma_1^2 = \sigma_2^2$ and the case where $\sigma_1^2 \neq \sigma_2^2$.

There are also some considerations for the cases where $n_1 = n_2$ and where $n_1 \neq n_2$.

Each of these cases alters the calculation of the standard error of the difference being tested and the degrees of freedom.

Variance	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$
$n_1 \neq n_2$	$S_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)$	$\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}$
$n_1 = n_2 = n$	$2S_p^2 \Big/ n$	$\dfrac{S_1^2 + S_2^2}{n}$

d.f.	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$
$n_1 \neq n_2$	$(n_1 - 1) + (n_2 - 1)$	$\geq \min[(n_1 - 1), (n_2 - 1)]$
$n_1 = n_2 = n$	$2n - 2$	$\geq n - 1$

For our purposes, we will generally use SAS to conduct two-sample t-tests, and will let SAS determine Satterthwaite's approximation when the variances are not equal?

How does SAS know if the variances are equal? How does it know what value of α you want to use? Good questions. Actually, SAS does not know or assume anything. We'll find out what it does later.

One last thought on testing for differences between two populations. The test we have been primarily discussing is the *t* test, a test of equality of means. However, if we find in the process of checking variance that the variances differ, then there are already some differences between the two populations that may be of interest.

Numerical example

Compare the ovarian weight of 14 fish, 7 randomly assigned to receive injections of gonadotropin (treatment group) and 7 assigned to receive a saline solution injection (control group). Both groups are treated identically except for the gonadotropin treatment. Ovarian weights are to be compared for equality one week after treatment. During the experiment two fish were lost due to causes not related to the treatment, so the experiment became unbalanced.

Raw data

Obs	Treatment	Control
1	134	70
2	146	85
3	104	94
4	119	83
5	124	97
6	*	77
7	*	80

Summary statistics

Statistic	Treatment	Control
n	5	7
ΣY_i	627	586
ΣY_i^2	79,625	49,588
$\bar{Y}$	125.4	83.7
SS	999	531
γ	4	6
S^2	249.8	88.6

Research question: Does the gonadotropin treatment affect the ovarian weight? (Note: this implies a non-directional alternative). First, which of the various situations for two-sample t-tests do we have? Obviously, $n_1 \neq n_2$. Now check the variances.

1) $H_0: \sigma_1^2 = \sigma_2^2$

2) $H_1: \sigma_1^2 \neq \sigma_2^2$

3) Assume $Y_i \sim NIDrv$, representing the usual assumptions of normality and independence.

4) $\alpha = 0.05$ and the critical value for 4, 6 d.f. is $F_{\alpha/2,4,6} = 6.23$.

5) We have the samples, and know that the variances are 249.8 and 88.6, and the d.f. are 4 and 6 respectively. The calculated value is (given that we have a nondirectional alternative and arbitrarily placing the largest variance in the numerator), F = 249.8/88.6 = 2.82 with 4, 6 d.f.

6) The critical value is larger than the calculated value. We therefore fail to reject the null hypothesis.

7) We can conclude that the two samples have sufficiently similar variances for pooling.

Pooling the variances.

Recall, $S_p^2 = \dfrac{\gamma_1 S_1^2 + \gamma_2 S_2^2}{\gamma_1 + \gamma_2} = \dfrac{SS_1 + SS_2}{\gamma_1 + \gamma_2}$

$$S_p^2 = \frac{4(249.8) + 6(88.6)}{4+6} = \frac{999+531}{4+6} = \frac{1530}{10} = 153 \text{ with 10 d.f.}$$

Now calculate the standard error for the test, $S_{\bar{d}}$, using the pooled variance.

For this case

$$S_{\bar{d}} = S_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{153\left(\frac{1}{5} + \frac{1}{7}\right)} = \sqrt{153(0.343)} = \sqrt{52.457} = 7.24\,, \text{ with 10 df}$$

Completing the two-sample t-test.

1) $H_0: \mu_1 - \mu_2 = \delta$. In this case we could state the null as $H_0: \mu_1 = \mu_2$ since $\delta = 0$.

2) $H_0: \mu_1 - \mu_2 \neq \delta$ or $H_0: \mu_1 \neq \mu_2$

3) Assume $d_i \sim$ NIDr.v. $(\delta, \sigma_\delta^2)$. NOTE we have pooled the variances, so obviously we have assumed that all variance is homogeneous and equal to σ_δ^2.

4) $\alpha = 0.05$ and the critical value is 2.228 (a nondirectional alternative for α=0.05 and 10 df)

5) We have the samples and know that the means are 125.4 and 83.7. The calculated t value is:

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - 0}{S_{\bar{d}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{S_{\bar{d}}} = \frac{125.4 - 83.7}{7.24} = \frac{41.7}{7.24} = 5.76 \text{ with 10 d.f.}$$

6) The calculated value (5.76) clearly exceeds the critical value (2.228) value, so we would reject the null hypothesis.

7) Conclude that the gonadotropin treatment does affect the gonad weight of the fish. We can further state that the treatment increases the weight of gonads.

How about a confidence interval? Could we use a confidence interval here? You betcha!

Confidence interval for the difference between means

The general formula for a two-tailed confidence interval for normally distributed parameters is: "*Some parameter estimate* $\pm t_{\alpha/2}$ * *standard error*"

The difference between the means ($\delta = (\mu_1 - \mu_2)$) is another parameter for which we may wish to calculate a confidence interval. For the estimate of the difference between μ_1 and μ_2 we have already determined that for α=0.05 we have $t_{\alpha/2} = 2.228$ with 10 d.f.. We also found the estimate of the difference $(\bar{d} = (\bar{Y}_1 - \bar{Y}_2))$ is 41.7 and the std error of the difference, $(S_{\bar{d}} = S_{\bar{Y}_1 - \bar{Y}_2})$, is 7.24.

The confidence interval is then $\bar{d} \pm t_{\alpha/2} S_{\bar{Y}}$ or 41.7$\pm$ 2.228(7.24) and 41.7 $\pm$ 16.13. The probability statement is

$$P(\bar{d} - t_{\alpha/2} S_{\bar{d}} \leq \mu_1 - \mu_2 \leq \bar{d} + t_{\alpha/2} S_{\bar{d}}) = 1 - \alpha$$

$$P(25.57 \leq \mu_1 - \mu_2 \leq 57.83) = 0.95$$

Note that the interval does not contain zero. This observation is equivalent to doing a test of hypothesis against zero. Some statistical software calculates intervals instead of

doing hypothesis tests. This works for hypothesis tests against zero and is advantageous if the hypothesized value of δ is something other than zero. When software automatically tests for differences it almost always test for differences from zero.

Summary

Testing for differences between two means can be done with the two-sample t-test or two sample Z test if variances are known.

For two independently sampled populations the variance will be $\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}$, the variance of a linear combination of the means.

The problem is the d.f. for this expression are not known.

Degrees of freedom are known if the variances can be pooled, so we start our two-sample t-test with an F-test.

Variances are pooled, if not significantly different, by calculating a weighted mean.

$$S_p^2 = \frac{\gamma_1 S_1^2 + \gamma_2 S_2^2}{\gamma_1 + \gamma_2} = \frac{SS_1 + SS_2}{\gamma_1 + \gamma_2} = \frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)}$$

The error variance is given by $S_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)$

The standard error is $\sqrt{S_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2} \right)}$

If the variances cannot be pooled, the two-sample t-test can still be done, and degrees of freedom are approximated with Satterthwaite's approximation.

Once the standard error is calculated, the test proceeds as any other t-test.

Confidence intervals can also be calculated in lieu of doing the t-test.

SAS example 4 – PROC TTEST

We would normally do two-sample t-tests with the SAS procedure called PROC TTEST. This procedure has the structure

```
proc ttest data = dataset name;
class group variable;
var variable of interest;
```

The PROC statement functions like any other proc statement.

The VARIABLE or VAR statement works the same as in other procedures we have seen.

The CLASS statement is new. It specifies the variable that will allow SAS to distinguish between observations from the two groups to be tested.

PROC TTEST Example 4a

Example from Steele & Torrie (1980) Table 5.2.

Corn silage was fed to sheep and steers. The objective was to determine if the percent digestibility differed for the two types of animals.

Example 1: Raw data

Obs	Sheep	Steers
1	57.8	64.2
2	56.2	58.7
3	61.9	63.1
4	54.4	62.5
5	53.6	59.8
6	56.4	59.2
7	53.2	

Unfortunately this data is not structured properly for PROC TTEST. It has two variables (sheep and steers) giving the percent digestibility for sheep and steers separately.

We need one variable with percent digestibility for both and a second variable specifying the type of animal.

This can be fixed in the data step.

In program note the following;

Change of the data structure from multivarite to univariate style.

The proc ttest statement

Note intermediate statistics, especially the confidence intervals for both means and standard deviations

The test the hypothesis for both means and variances are discussed below.

Interpreting the SAS Output

First examine the last lines

```
   Equality of Variances
   Variable    Method          Num DF      Den DF      F Value     Pr > F
   percent     Folded F            6           5         1.70      0.5764
```

SAS is testing the Equality of Variances (H_0: $\sigma_1^2 = \sigma_2^2$). Notice that SAS provides a "folded F".

Most SAS F tests are one-tailed, but this is one of the few places that SAS does a two-tailed F test (a "folded F"). SAS gives the d.f. and the probability of a greater F by random chance. We would usually set $\alpha = 0.05$, and would reject any P-value less than this and fail to reject any value greater than this. In this case we fail to reject.

Exactly what did SAS do with the "folded F". Recall the two-tailed F allows you to place the larger F in the numerator, but you must use $\alpha/2$ as a critical value. This is what SAS has done. When SAS gave the P value of 0.5764, it is a two tailed P value.

So we conclude that the variances do not differ. If doing the test by hand we would now pool the variances to calculate the standard error.

NOW, look at the PROC TTEST output, above the F test.

t-tests

Here SAS provides results for both types of test, one calculated using equal variances and another done with unequal variances and the user chooses which is appropriate for their case. Since we had equal variances according to the F test we just examined, we would use the first line.

```
   Variable   Method          Variances    DF   t Value   Pr > |t|
   percent    Pooled          Equal        11    -3.34     0.0065
   percent    Satterthwaite   Unequal      10.9  -3.42     0.0058
```

From the first line we see that the calculated t value was -3.3442 with 11 d.f.   The probability of getting a greater value by random chance (i. e. the H_0) is 0.0065, not very likely. We would conclude that there are statistically significant differences between the two animals in terms of silage digestibility.

What about the other line, for unequal variances?

```
   Variable   Method          Variances    DF   t Value   Pr > |t|
   percent    Pooled          Equal        11    -3.34     0.0065
   percent    Satterthwaite   Unequal      10.9  -3.42     0.0058
```

This line would be used if we rejected the F test of equal variances. In this particular case the conclusion would be the same since we would also reject H_0. Notice that the d.f. for the calculations for unequal variance are not integer. This is because Satterthwaite's approximation was used to estimate the variances. Since the variances were actually "equal", the estimate is close to $(n_1-1) + (n_2-1) = 11$.

From the SAS STATISTICS output we can conclude that the digestibility is higher for the steers, by about 5 percent.

Example 4b: from Steele & Torrie (1980) Table 5.6

Determine if there is a difference in the percent fine gravel found in surface soils. The data is from a study comparing characteristics of soil categorized as "good" or "poor".

The raw data

Good	Poor
5.9	7.6
3.8	0.4
6.5	1.1
18.3	3.2
18.2	6.5
16.1	4.1
7.6	4.7

Percent fine sand in good and poor soils

This data is also in the form of two separate variables and must be adjusted to accommodate the data structure needed by PROC TTEST.

In program note the following;

Change of the data structure from multivarite to univariate style.

The proc ttest statement

Note intermediate statistics, especially
the confidence intervals for both means and standard deviations

The test the hypothesis for both means and variances are discussed below.

In this case the variances are not quite different, though it is a close call and there is a pretty good chance of Type II error. Fortunately, the result is the same with either test.

If we go strictly by the "$\alpha = 0.05$" decision rule that we usually use, we would fail to reject the hypothesis of equal variances.

We would then examine the line for equal variances and conclude that there was indeed a difference between the good and poor quality soil in terms of the fine sand present.

The intermediate statistics show that the good soil had about 7 percent more fine sand.

```
Statistics                    Lower CL          Upper CL  Lower CL
Variable  soilqual    N      Mean     Mean      Mean    Std Dev  Std Dev
percent   good        7    5.0559   10.914    16.773    4.0819   6.3344
percent   poor        7    1.5048   3.9429    6.3809    1.6987   2.6362
```

Example 4c: Steele & Torrie (1980) Exercise 5.5.6

The weights in grams of 10 male and 10 female juvenile ring-necked pheasants trapped in January in Wisconsin are given. Test the H_0 that males were 350 grams heavier than females.

In this case the data is in the form needed, one variable for weight and one for sex.

Raw data

Sex	Weight
Female	1061
Female	1065
Female	1092
Female	1017
Female	1021
Female	1138
Female	1143
Female	1094
Female	1270
Female	1028

Sex	Weight
Male	1293
Male	1380
Male	1614
Male	1497
Male	1340
Male	1643
Male	1466
Male	1627
Male	1383
Male	1711

There was, however, one little problem with this analysis. The hypothesis requested was not simply $H_0: \mu_{male} = \mu_{female}$, it was $H_0: \mu_{male} = \mu_{female} + 350$, or $H_0: \mu_{male} - \mu_{female} = 350$. SAS does not have provisions to specify an alternative other than zero, but if we subtract 350 from the males, we could then test for equality. We know from our discussion of transformations that the variances will be unaffected.

So we create a new variable called adjwt for "adjusted weight". See the calculations in the SAS program.

```
    ...
    8      Female      1094      1094
    9      Female      1270      1270
   10      Female      1028      1028
   11      Male        1293       943
   12      Male        1380      1030
   13      Male        1614      1264
    ...
```

See SAS OUTPUT Appendix 4c

Note intermediate statistics

Note test the hypothesis for both means and variances.

Note that in the PROC TTEST there is another calculation in the statistics. This is the "Diff" which also gets its calculated value and confidence interval. This difference is not a paired difference, it is the simple difference between the two group means.

Statistics Variable	sex	N	Lower CL Mean	Mean	Upper CL Mean
AdjWT	Female	10	1038.1	1092.9	1147.7
AdjWT	Male	10	1041	1145.4	1249.8
AdjWT	Diff (1-2)		-162	-52.5	56.989

Interpretation of the SAS output

First, we fail to reject $H_0: \sigma_1^2 = \sigma_2^2$ again (barely). But the weights do not differ either way (examining $Pr > |t|$). So we fail to reject $H_0: \mu_1 = \mu_2$, but remember we added 350 to the males. So actually we conclude that the males are greater by an amount not different from 350 grams.

A special case – the paired t-test

One last case. In some circumstances the observations are not separate and distinct in the two samples. Sometimes they can be paired. This can be good, adding power to the design.

For example:

We want to test toothpaste. We may pair on the basis of twins, or siblings in assigning the toothpaste treatments.

We want to compare deodorants or hand lotions. We assign one arm or hand to one brand an the other to another brand.

In may drug and pharmaceutical studies done on rats or rabbits the treatments are paired on litter mates.

So, how does this pairing affect our analysis? The analysis is done by subtracting one category of the pair from the other category of the pair. In this way the pair values become difference values.

As a result, what originally appears to be a "two-sample t-test" becomes a one-sample t-test of the differences between the two members of each pair.

So, in many ways the paired t-test is easier.

Example: We already did an example of this type of analysis. Recall the Lucerne flowers whose seeds we compared for flowers at the top and bottom of the plant. This was paired and we took differences.   The difference was "1" with a standard error of "0.5055".

SAS example 2c examined previously

```
          Tests for Location: Mu0=0
 Test              -Statistic-      -----p Value------
 Student's t    t  1.978141     Pr > |t|    0.0793
 Sign           M         2     Pr >= |M|   0.3438
 Signed Rank    S      19.5     Pr >= |S|   0.0469
```

So the paired t-test is an alternative analysis for certain data structures. It is better because it eliminates the "between pair" variation and compares the treatments "within pairs". This reduces variance.

However, note that the degrees of freedom are also cut in half. If the basis for pairing is not good, the variance is not reduced, but degrees of freedom are lost.

Summary

The SAS PROC TTEST provides all of the tests needed for two-sample t-tests. It provides the test of variance we need to start with, and it provides two alternative calculations, one for equal variance and one for unequal variance. We choose the appropriate case.

We also saw that several previous calculations, such as confidence intervals and sample size, are also feasible for the two-sample t-test case.

The paired t-test, where there is a good strong basis for pairing observations, can gain power by reducing between pair variation. However, if the basis for pairing is not good, we lose degrees of freedom and power.

Calculating a needed sample size

The Z-test and t-test use a similar formula. $Z = \dfrac{\overline{Y} - \mu_0}{\sqrt{\dfrac{\sigma^2}{n}}} = \dfrac{\overline{Y} - \mu_0}{\sigma / \sqrt{n}}$

Let's suppose we know everything in the formula except n. Do we really? Maybe not, but we can get some pretty good estimates.

Call the numerator ($\overline{Y} - \mu_0$) a difference, $\overline{d}$. It is some mean difference we want to be able to detect, so $\overline{d} = \overline{Y} - \mu_0$

The value σ^2 is a variance, the variance of the data that we will be sampling. We need this variance, or an estimate, S^2.

So we alter the formula to read. $Z = \dfrac{\overline{d}}{\sqrt{\dfrac{\sigma^2}{n}}} = \dfrac{\overline{d}}{\sigma / \sqrt{n}}$

What other values do we know? Do we know Z? No, but we know what Z we need to obtain significance. If we are doing a 2-tailed test, and we set $\alpha = 0.05$, then Z will be 1.96.

Any calculated value larger will be "more significant", any value smaller will not be significant.

So, if we want to detect significance at the 5% level, we can state that ...

We will get a significant difference if $Z = \dfrac{\overline{d}}{\sqrt{\dfrac{\sigma^2}{n}}} = \dfrac{\overline{d}}{\sigma / \sqrt{n}} \geq Z_{\alpha/2}$

We square both sides and solve for n. Then we will also SHOULD get a significant difference if $n \geq \dfrac{Z_{\alpha/2}^2 \sigma^2}{\overline{d}^2}$. Then, if we know the values of $\overline{d}, \sigma^2$ and Z, we can solve the formula for n. If we are going to use a Z distribution we should have a known value of the variance (σ^2). If the variance is calculated from the sample, use the t distribution. This would give us the sample size needed to obtain "significance", in accordance with whatever Z value is chosen.

Generic Example

Try an example where

$\overline{d} = 2$

$\sigma = 5$, $\sigma^2 = 25$

$Z = 1.96$

So what value of n would detect this difference with this variance and produce a value of Z equal to 1.96 (or greater)?

$n \geq \dfrac{Z_{\alpha/2}^2 \sigma^2}{\overline{d}^2} = (1.96^2 * 5^2)/2^2 = 3.8416(25)/4 = 24.01$

since n $\geq$ 24.01, round up to 25.

Answer, $n \geq 25$ would produce significant results. Guaranteed? Wouldn't this always produce significant results? Theoretically, within the limits of statistical probability of error, yes, but only if the difference was really 2. If the null hypothesis (no difference, $\mu = \mu_0$) was really true and we took larger samples, then we would get a better estimate of 0, and may never show significance.

Considering Type II Error

The formula we have seen contains only $Z_{\alpha/2}$ or $t_{\alpha/2}$, depending on whether we have σ^2 or S^2. However, a fuller version can contain consideration of the probability of Type II error (b).

We can often use Z when working with very large samples.

Remember that to work with TYPE II or β error we need to know the mean of the real distribution. However, in calculating sample size we have a difference, $\overline{d} = \overline{Y} - \mu_0$. So we can include consideration of type II error and power in calculating the sample size. The consideration of β error would be done by adding another Z or t for the error rate. Notice that below I switch to t distributions and use $n \geq \dfrac{(t_{\alpha/2} + t_\beta)^2 S^2}{\overline{d}^2}$.

Other examples

We have done a number of tests, some yielding significant results and others not. If a test yields significant results (showing a significant difference between the observed and hypothesized values), then we don't need to examine sample size because the sample was big enough. However, some utility may be made of this information if we FAIL to reject the null hypothesis.

Note: Some textbooks give only the formula I originally gave for Z, without the β error consideration.   What is the power if you use the formula omitting t_β from $n \geq \dfrac{(t_{\alpha/2} + t_\beta)^2 S^2}{\overline{d}^2}$?

If you set t_β equal to zero the power is 0.50 and there is a 50% chance of making a Type II error.

An example with t values and β error included

Recall the Rhesus monkey experiment. We hypothesized no effect of a drug, and with a sample size of 10 were unable to reject the null hypothesis. However, we did observe a difference of +0.8 change in blood pressure after administering the drug. What if this change was real? What if we made a Type II error? How large a sample would we need to test for a difference of 0.8 if we also wanted 90% power?

So we want to know how large a sample we would need to get significance at the α=0.05 level if power was 0.90. In this case $\beta = 0.10$. To do this calculation we need a two tailed α and a one tailed β (we know that the observed change is +0.8). We will estimate the variance from the sample so we will use the t distribution. However, since we don't know the sample size, we don't know the degrees of freedom! Since we do not know the d.f. we will start off with some "reasonable" values for t_α and t_β. Then after we solve the equation we will have an estimate of the d.f.   We can solve again with better values of t_α and t_β, and refine our estimate. After our second calculation we have even better estimates of d.f., so