

The Two-Sample t Test

The two-sample t test is probably the most used statistical test. Unfortunately, as alluded to in David Moore's essay *Teaching Statistics as a Respectable Subject* in Chapter 1 of Perspectives on Contemporary Statistics, MAA Notes Number 21, it may also be the least understood and most abused statistical test. Moore suggests that "the theoretical merit of a statistical procedure is not the same as its practical merit...theory is an imperfect guide to practice" (p. 18-19). We will see his point as we consider the two-sample t -procedures.

The Pooled T Test

Consider two normal populations, one with mean \mathbf{m}_1 and variance \mathbf{s}_1^2 , the other with mean \mathbf{m}_2 and variance \mathbf{s}_2^2 . Let \bar{Y}_1 and \bar{Y}_2 be means of two independent random samples of sizes n_1 and n_2 . Let S_1 and S_2 be the standard deviations of those samples.

What is the distribution of $\frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$?

If $\mathbf{s}_1^2 = \mathbf{s}_2^2 = \mathbf{s}^2$ we can get an unbiased estimator of this common variance, \mathbf{s}^2 , by pooling to get

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$

We can use this expression for S_p^2 to define W .

$$\begin{aligned} W &= \frac{(n_1 + n_2 - 2)S_p^2}{\mathbf{s}^2} \\ &= \frac{(n_1 + n_2 - 2)}{\mathbf{s}^2} \cdot \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)} \\ &= \frac{(n_1 - 1)S_1^2}{\mathbf{s}^2} + \frac{(n_2 - 1)S_2^2}{\mathbf{s}^2} \end{aligned}$$

We know that $\frac{(n_1 - 1)S_1^2}{\mathbf{s}_1^2} \sim \mathbf{c}_{n_1 - 1}^2$, and $\frac{(n_2 - 1)S_2^2}{\mathbf{s}_2^2} \sim \mathbf{c}_{n_2 - 1}^2$. We also know that the sum of

independent \mathbf{c}^2 random variables is a \mathbf{c}^2 random variable. So $W = \frac{(n_1 + n_2 - 2)S_p^2}{\mathbf{s}^2}$ has a Chi-square distribution with $n_1 + n_2 - 2$ degrees of freedom.

Now, consider the statistic, $\frac{\left(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\mathbf{s}^2/n_1 + \mathbf{s}^2/n_2}} \right)}{\sqrt{\frac{(n_1 + n_2 - 2) S_p^2 / \mathbf{s}^2}{(n_1 + n_2)}}}$. This statistic is of the form $\frac{N(0,1)}{\sqrt{c^2/df}}$, a normal

divided by an independent Chi-square divided by its degrees of freedom. We know that this is the definition of a t -distribution with $n_1 + n_2 - 2$ degrees of freedom. Simplifying the algebra, the statistic becomes:

$$\begin{aligned} & \frac{\left(\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\mathbf{s}^2/n_1 + \mathbf{s}^2/n_2}} \right)}{\sqrt{\frac{(n_1 + n_2 - 2) S_p^2 / \mathbf{s}^2}{(n_1 + n_2)}}} \\ &= \frac{\bar{Y}_1 - \bar{Y}_2}{\mathbf{s} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \cdot \frac{\mathbf{s}}{S_p} \\ &= \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \end{aligned}$$

Notice that the terms \mathbf{s}^2 and $n_1 + n_2 - 2$ cancel. This result means that if we want to test $H_0 : \mathbf{m}_1 - \mathbf{m}_2 = 0$ vs. $H_a : \mathbf{m}_1 - \mathbf{m}_2 \neq 0$, the statistic

$$T_p = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t -distribution with $n_1 + n_2 - 2$ degrees of freedom. If $\mathbf{s}_1 = \mathbf{s}_2$, independent normal samples and $\mathbf{m}_1 = \mathbf{m}_2$, this is not an approximation, but an exact t -distribution.

To review, if the samples are selected independently and randomly from normal populations with common standard deviations, and if the null hypothesis $\mathbf{m}_1 = \mathbf{m}_2$ is true, then T_p has a t distribution with $n_1 + n_2 - 2$ degrees of freedom. However, the test based on this statistic is not robust with respect to departures from the assumption of common variances, especially if n_1 and n_2 are unequal. And we never can be sure that $\mathbf{s}_1 = \mathbf{s}_2$. Many statistical tests for $\mathbf{s}_1 = \mathbf{s}_2$ are not very useful because they are very sensitive to violations of normality. As George Box says, "To make a preliminary test on variances is rather like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port." Given all of this, using a t test in this situation is a reasonable solution to the wrong problem. But as John

Tukey suggests, "an approximate solution to the right problem is better than an exact solution to the wrong problem." The unpooled t -test is that approximate solution.

The Unpooled T Test

A natural test statistic for comparing \mathbf{m}_1 and \mathbf{m}_2 is:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}.$$

This statistic does not have a t distribution even if $\mathbf{S}_1 = \mathbf{S}_2$ and $\mathbf{m}_1 = \mathbf{m}_2$. However, regardless of whether $\mathbf{S}_1 = \mathbf{S}_2$, the critical values of its null distribution (that is, when $\mathbf{m}_1 = \mathbf{m}_2$) can be approximated quite well by those of a t distribution whose degrees of freedom depend on the observed data. The degrees of freedom are given as:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

When $n_1 = n_2$ and $s_1 = s_2$ this expression is equivalent to $n_1 + n_2 - 2$ degrees of freedom. The value of df is always between $\min(n_1 - 1, n_2 - 1)$ and $n_1 + n_2 - 2$.

It seems advisable in an introductory statistics course just to teach and use the unpooled procedure for tests of $\mathbf{m}_1 = \mathbf{m}_2$ and for confidence intervals for $\mathbf{m}_1 - \mathbf{m}_2$. Teaching both methods in a first course would likely only confuse students. Minitab uses the *unpooled* procedure by default. This seems wise!

The unpooled procedure loses only a little power if in fact $\mathbf{S}_1 = \mathbf{S}_2$. If \mathbf{S}_1 and \mathbf{S}_2 are unequal, the unpooled procedure maintains the prescribed α -level and retains good power. If you use the pooled test statistic when it is inappropriate, you may be seriously misled. That is, the probabilities of various errors may be far from what you think they are. For example, you may do a test with an "advertised" or nominal $\alpha = 0.05$ and yet, in fact, α might be 0.20. Or you might calculate a supposed 95% confidence interval but, in fact, the confidence level could be only 80%. Using simulation we can demonstrate the truth of these comments.

The following calculator program can illustrate this (more slowly and less accurately than a computer, but it suffices). It selects 20 values from a normal distribution with mean 100 and standard deviation 40 and 30 values from a normal distribution with mean 100 and standard deviation 10. This violates the equal variance condition for pooling. Both the pooled value of t with 48 degrees of freedom (in this program called P) and the unpooled value of t with the conservative 19 degrees of freedom (in this program called U) are computed. The number of times the true null hypothesis is rejected is stored in lists 5 and 6. The program takes approximately 40 minutes to run.

PROGRAM: POOL

```

ClrList L1 , L2 , L5 , L6 , LA , LB , LC , LD , LV , LP , LU
For(X,1,500)
  randNorm(100,40,20) → L1
  randNorm(100,10,30) → L2
  mean(L1)→A
  variance(L1)→B
  mean(L2)→C
  variance(L2)→D
  (19B+29D)/48→V
  (A - C)/(√(V/20+V/30))→P
  (A - C)/(√(B/20+D/30))→U
  If abs(P) ≥ 2.011
  Then
  1→L5
  Else
  0→L5
  End
  If abs(U) ≥ 2.094
  Then
  1→L6
  Else
  0→L6
  End
End

```

Program Notes

Repeats process 500 times
Selects 20 from N(100, 40)
Selects 30 from (N100, 10)

Computes pooled variance
Finds t-score for pooled variance
Finds t-score for unpooled variance
Compares t-score with critical value 48 df

Assigns 1 for reject null

Assigns 0 for fail to reject

Compares t-score with critical value 19 df

Assigns 1 for reject null

Assigns 0 for fail to reject

The result of one run of the program is that the true null hypothesis was rejected 9.6% of the time with pooling (48 out of 500) and 4.4% of the time using the unpooled, approximate technique (22 out of 500). The prescribed value of α was 0.05.

How accurate are these results? This is an example of a Bernoulli trial. We have 500 repetitions of either reject (success) or fail to reject (failure). We can create confidence intervals for the true proportion and see if they contain 0.05.

For the pooled setting, we have $0.096 \pm 1.96 \sqrt{\frac{0.096(0.904)}{500}} = 0.096 \pm 0.0258$. This gives a 95% confidence interval for the true proportion of (0.070, 0.122). Notice 0.05 is not in this interval.

For the unpooled setting, we have $0.044 \pm 1.96 \sqrt{\frac{0.044(0.956)}{500}} = 0.044 \pm 0.018$. This gives a 95% confidence interval for the true proportion of (0.026, 0.062). Notice 0.05 is in this interval.