CollegeBoard
Advanced Placement Program

AP® Statistics

The Satterthwaite Formula for Degrees of Freedom in the Two-Sample t-Test

Michael Allwood
Brunswick School
Greenwich, Connecticut

connect to college success™
www.collegeboard.com

I. Introduction

What's the most complicated formula we encounter in AP Statistics? To me it's undoubtedly the formula for degrees of freedom in the two-sample *t*-test (the version of the test where we do not assume equal population variances):

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

Admittedly, we don't have to tell our students this formula. We can tell them to use the number of degrees of freedom given by the calculator (which is in fact the result of this formula), or we can tell them to resort to the "conservative" method of using the smaller of $n_1 - 1$ and $n_2 - 1$. Nonetheless, I've been intrigued over the years by this array of symbols and have been eager to know where it comes from.

The formula was developed by the statistician Franklin E. Satterthwaite and a derivation of the result is given in Satterthwaite's article in *Psychometrika* (vol. 6, no. 5, October 1941). My aim here is to translate Satterthwaite's work into terms that are easily understood by AP Statistics teachers. The mathematics involved might seem a little daunting at first, but apart perhaps from one or two steps in section V, no stage in the argument is beyond the concepts in AP Statistics. (Section V concerns two standard results connected with the chi-square distributions. These results can easily be accepted and their proofs omitted on the first reading.) It is also worth noting that section IV, concerning the test statistic in the **one**-sample *t*-test, is only included by way of an introduction to the work on Satterthwaite's formula. So this section, too, can be omitted by the reader who wants the quickest route to Satterthwaite's result.

II. A Definition of the Chi-Square Distributions

Let $Z_1, Z_2, \ldots, Z_n$ be independent random variables, each with distribution $N(0,1)$.

The χ^2 (chi-square) distribution with *n* degrees of freedom can be defined by

$$\chi_n^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2. \qquad (1)$$

III. A Definition of the *t*-Distributions

Let's suppose that *X* has distribution $N(\mu, \sigma)$ and that $X_1, \ldots, X_n$ is a random sample of values of *X*. As usual, we denote the mean and the standard deviation of the sample by $\bar{X}$ and s, respectively. In 1908, W. S. Gosset, a statistician working for Guinness in Dublin, Ireland, set about determining the distribution of

$$\frac{\bar{X} - \mu}{s/\sqrt{n}},$$

and it is this distribution that we refer to as the "*t*-distribution." Actually, we should refer to the "*t*-distributions" (plural), since the distribution of that statistic varies according to the value of *n*.

However, we **define** the *t*-distributions in the following way: Suppose that *Z* is a random variable whose distribution is $N(0,1)$, that *V* is a random variable whose distribution is χ^2 with *n* degrees of freedom, and that *Z* and *V* are independent. Then the *t*-distribution with *n* degrees of freedom is given by

$$t_n = \frac{Z}{\sqrt{V/n}}. \qquad (2)$$

Our task in the next section is to confirm that Gosset's *t*-statistic, $t = (\bar{X} - \mu)/(s/\sqrt{n})$, does, in fact, have a *t*-distribution.

IV. A Demonstration That $(\bar{X} - \mu)/(s/\sqrt{n})$ Has Distribution t_{n-1}

First,

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{s^2/\sigma^2}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}}.$$

Now we know that the distribution of

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \text{ is } N(0,1),$$

so according to the definition (2) of the *t*-distribution, we now need to show that $(n-1)s^2/\sigma^2$ is χ^2 distributed with $n-1$ degrees of freedom and that $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ and $(n-1)s^2/\sigma^2$ are independent. This second fact is equivalent to the independence of $\bar{X}$ and s when sampling from a normal distribution, and its proof is too complex for us to attempt here.[1] To show that $(n-1)s^2/\sigma^2$ is χ^2_{n-1}, we start by observing that

$$\frac{(n-1)s^2}{\sigma^2} = \frac{n-1}{\sigma^2} \cdot \frac{\sum (X_i - \bar{X})^2}{n-1} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2}.$$

We first replace the sample mean $\bar{X}$ with the population mean μ and turn our attention to

$$\frac{\sum (X_i - \mu)^2}{\sigma^2} = \sum \left(\frac{X_i - \mu}{\sigma} \right)^2.$$

Since each X_i is independently $N(\mu,\sigma)$, each $(X_i-\mu)/\sigma$ is independently $N(0,1)$. So $\sum\left((X_i-\mu)/\sigma\right)^2$ is the sum of the squares of *n* independent $N(0,1)$ random variables, and therefore, according to the definition (1) of the χ^2 distributions, it is χ^2 distributed with *n* degrees of freedom.

Now,

$$\begin{aligned}\sum(X_i-\mu)^2 &= \sum\left[(X_i-\bar{X})+(\bar{X}-\mu)\right]^2\\ &= \sum\left[(X_i-\bar{X})^2+2(X_i-\bar{X})(\bar{X}-\mu)+(\bar{X}-\mu)^2\right]\\ &= \sum(X_i-\bar{X})^2+2(\bar{X}-\mu)\sum(X_i-\bar{X})+n(\bar{X}-\mu)^2.\end{aligned}$$

But $\sum(X_i-\bar{X})=\sum X_i-n\bar{X}=\sum X_i-n\frac{\sum X_i}{n}=0,$ so

$$\sum(X_i-\mu)^2=\sum(X_i-\bar{X})^2+n(\bar{X}-\mu)^2. \qquad (3)$$

Therefore, dividing by σ^2,

$$\frac{\sum(X_i-\mu)^2}{\sigma^2}=\frac{\sum(X_i-\bar{X})^2}{\sigma^2}+\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2. \qquad (4)$$

The fact that we have just established, (4), gives us the key to our argument: $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ is $N(0,1)$, and so $\left[(\bar{X}-\mu)/(\sigma/\sqrt{n})\right]^2$ is χ_1^2. Also, we established that $\sum(X_i-\mu)^2/\sigma^2$ is χ_n^2. Now we mentioned above that $(\bar{X}-\mu)/(\sigma/\sqrt{n})$ and $(n-1)s^2/\sigma^2$ (i.e., $\sum(X_i-\bar{X})^2/\sigma^2$) are independent when sampling from a normal distribution. So according to (4), $\sum(X_i-\bar{X})^2/\sigma^2$ has that distribution that must be independently added to χ_1^2 to give χ_n^2. Looking at the definition of the χ^2 distributions (1), we see that this distribution must be the sum of the squares of $n-1$ independent normally distributed random variables, that is, χ_{n-1}^2.

So we have shown that $\dfrac{\sum(X_i-\bar{X})^2}{\sigma^2}=\dfrac{(n-1)s^2}{\sigma^2}$ is χ_{n-1}^2.

Thus we have completed our demonstration that $\dfrac{\bar{X}-\mu}{s/\sqrt{n}}$ is *t* distributed with $n-1$ degrees of freedom.

V. The Mean and Variance of the Chi-Square Distribution with *n* Degrees of Freedom

In section II we defined the chi-square distribution with n degrees of freedom by $\chi_n^2 = Z_1{}^2 + Z_2{}^2 + \ldots + Z_n{}^2$, where $Z_1, Z_2, \ldots, Z_n$ are independent random variables, each with distribution $N(0,1)$.

Taking the expected value and the variance of both sides, we see that

$$E(\chi_n^2) = E(Z_1{}^2) + \ldots + E(Z_n{}^2),$$

and

$$\text{Var}(\chi_n^2) = \text{Var}(Z_1{}^2) + \ldots + \text{Var}(Z_n{}^2).$$

But all the instances of Z_i have identical distributions, so
$$E(\chi_n^2) = nE(Z^2),$$

and

$$\text{Var}(\chi_n^2) = n\text{Var}(Z^2),$$

where Z is the random variable with distribution $N(0,1)$.

Now,

$$E(Z^2) = E\left[(Z-0)^2\right] = E\left[(Z-\mu_Z)^2\right] = \text{Var}(Z) = 1,$$

telling us that

$$E(\chi_n^2) = n \cdot 1 = n.$$

So we are left now with the task of finding $\text{Var}(Z^2)$.

Now,

$$\begin{aligned}\text{Var}(Z^2) &= E\left[(Z^2-\mu_{Z^2})^2\right] = E\left[(Z^2-1)^2\right] = E(Z^4 - 2Z^2 + 1) \\ &= E(Z^4) - 2E(Z^2) + 1 = E(Z^4) - 2 \cdot 1 + 1,\end{aligned}$$

so

$$\text{Var}(Z^2) = E(Z^4) - 1. \qquad (5)$$

To find $E(Z^4)$, we'll use the fact that for any continuous random variable X with probability density function f, and any exponent k,

$$E(X^k) = \int_{-\infty}^{\infty} x^k f(x)\,\mathrm{d}x,$$

and that the probability density function *f* of the $N(0,1)$ random variable is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2}.$$

Hence,

$$E(Z^4) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^4 \mathrm{e}^{-z^2/2} \mathrm{d}z.$$

From this, using integration by parts, we see that

$$\begin{aligned}
E(Z^4) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^3 \cdot z\mathrm{e}^{-z^2/2} \mathrm{d}z \\
&= \frac{1}{\sqrt{2\pi}} \left\{ \left[z^3 \cdot \left(-\mathrm{e}^{-z^2/2}\right) \right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} 3z^2 \left(-\mathrm{e}^{-z^2/2}\right) \mathrm{d}z \right\} \\
&= \frac{1}{\sqrt{2\pi}} \left\{ 0 + \int_{-\infty}^{\infty} 3z^2 \mathrm{e}^{-z^2/2} \mathrm{d}z \right\} \\
&= 3 \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 \mathrm{e}^{-z^2/2} \mathrm{d}z \\
&= 3E(Z^2) = 3 \cdot 1 = 3.
\end{aligned}$$

Hence, returning to (5), $\mathrm{Var}(Z^2) = 3 - 1 = 2,$ telling us that $\mathrm{Var}(\chi_n^2) = n \cdot 2 = 2n.$

So we have proved that $E(\chi_n^2) = n$ and $\mathrm{Var}(\chi_n^2) = 2n$. (6)

VI. Satterthwaite's Formula

In section IV we looked at the test statistic for the one-sample *t*-test, $(\overline{X} - \mu)\big/(s/\sqrt{n})$. We established that when sampling from a normal distribution and using the sample variance s^2 as an estimator for the population variance σ^2, the distribution of $(\overline{X} - \mu)\big/(s/\sqrt{n})$ is *t*, with $n-1$ degrees of freedom. This was a consequence of the fact that the distribution of $\dfrac{(n-1)s^2}{\sigma^2}$ is χ_{n-1}^2.

Note that n and σ are constants, and so the relevant fact here is that this particular multiple of s^2 is chi-square distributed.

Now we turn our attention to the two-sample *t*-test, and we're concerning ourselves with the version of the test where we don't assume that the two populations have equal variances. Here we're taking a random sample $X_1, \ldots, X_{n_1}$ from a random variable *X* with distribution $N(\mu_1, \sigma_1)$ and a random sample $Y_1, \ldots, Y_{n_2}$ from a random variable *Y* with distribution $N(\mu_2, \sigma_2)$. We say

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^{\,2}}{n_1} + \dfrac{s_2^{\,2}}{n_2}}}, \qquad (7)$$

and we would like to be able to say that this statistic has a *t*-distribution. But strictly speaking, it does not.

Let's look into this a little more deeply. The variance of $\bar{X} - \bar{Y}$ is [2]

$$\sigma_B^{\,2} = \frac{\sigma_1^{\,2}}{n_1} + \frac{\sigma_2^{\,2}}{n_2},$$

and, as an estimator for $\sigma_B^{\,2}$, we're using

$$s_B^{\,2} = \frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}.$$

For *t* to be *t*-distributed, there would have to be some multiple of $s_B^{\,2}$ that is chi-squared distributed -- and this is not the case. (If we try to analyze $s_B^{\,2}$ in the same way we analyzed s^2 in section IV , it becomes clearer that no multiple of $s_B^{\,2}$ can be chi-square distributed.)

However, remember that in the one-sample case, $(n-1)s^2/\sigma^2$ had a chi-square distribution with $n-1$ degrees of freedom. By analogy, we would like here to be able to say that, for some value of *r*, $rs_B^{\,2}/\sigma_B^{\,2}$ has a chi-square distribution with *r* degrees of freedom. Satterthwaite found the true distribution of $s_B^{\,2}$ and showed that if *r* is chosen so that the variance of the chi-square distribution with *r* degrees of freedom is equal to the true variance of $rs_B^{\,2}/\sigma_B^{\,2}$, then, under certain conditions, this chi-square distribution with *r* degrees of freedom is a good approximation to the true distribution of $rs_B^{\,2}/\sigma_B^{\,2}$. (In practice, we summarize the conditions by requiring that both n_1 and n_2 be reasonably large -- for example, that n_1 and n_2 both be greater than 5.)[3] Our task here is to derive the formula for this value of *r*.

So from this point, we are assuming that $rs_B^{\,2}/\sigma_B^{\,2}$ has distribution χ_r^2. In which case, using (6),

$$\mathrm{Var}\left(\frac{r{s_B}^2}{{\sigma_B}^2}\right) = 2r. \qquad (8)$$

Now, using the elementary rule for variances of random variables, $\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, we can also say that

$$\mathrm{Var}\left(\frac{r{s_B}^2}{{\sigma_B}^2}\right) = \frac{r^2}{{\sigma_B}^4}\mathrm{Var}({s_B}^2). \qquad (9)$$

Hence, using (8) and (9),

$$2r = \frac{r^2}{{\sigma_B}^4}\mathrm{Var}({s_B}^2),$$

giving

$$\frac{2}{r} = \frac{1}{{\sigma_B}^4}\mathrm{Var}({s_B}^2). \qquad (10)$$

Now,

$${s_B}^2 = \frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2},$$

and s_1 and s_2 are independent, so

$$\mathrm{Var}({s_B}^2) = \frac{1}{{n_1}^2}\mathrm{Var}({s_1}^2) + \frac{1}{{n_2}^2}\mathrm{Var}({s_2}^2). \qquad (11)$$

We know that $(n_1 - 1){s_1}^2/{\sigma_1}^2$ has a chi-square distribution with $n_1 - 1$ degrees of freedom, and so, using (6) again,

$$\mathrm{Var}\left[\frac{(n_1 - 1){s_1}^2}{{\sigma_1}^2}\right] = 2(n_1 - 1).$$

Therefore,

$$\frac{(n_1 - 1)^2}{{\sigma_1}^4}\mathrm{Var}({s_1}^2) = 2(n_1 - 1),$$

and so

$$\mathrm{Var}(s_1^{\,2}) = \frac{2\sigma_1^{\,4}}{n_1 - 1}.$$

Similarly,

$$\mathrm{Var}(s_2^{\,2}) = \frac{2\sigma_2^{\,4}}{n_2 - 1}.$$

Hence, returning to (11),

$$\mathrm{Var}(s_B^{\,2}) = \frac{1}{n_1^{\,2}} \cdot \frac{2\sigma_1^{\,4}}{n_1 - 1} + \frac{1}{n_2^{\,2}} \cdot \frac{2\sigma_2^{\,4}}{n_2 - 1}.$$

So, by (10),

$$\frac{2}{r} = \frac{1}{\sigma_B^{\,4}} \left(\frac{1}{n_1^{\,2}} \cdot \frac{2\sigma_1^{\,4}}{n_1 - 1} + \frac{1}{n_2^{\,2}} \cdot \frac{2\sigma_2^{\,4}}{n_2 - 1} \right),$$

which gives us

$$r = \frac{\sigma_B^{\,2}}{\frac{1}{n_1^{\,2}} \cdot \frac{\sigma_1^{\,4}}{n_1 - 1} + \frac{1}{n_2^{\,2}} \cdot \frac{\sigma_2^{\,4}}{n_2 - 1}}. \qquad (12)$$

In practice, the values of the population variances, σ_1^2 and $\sigma_2^{\,2}$, are unknown, and so we replace $\sigma_1^{\,2}$, $\sigma_2^{\,2}$, and $\sigma_B^{\,2}$ by their estimators $s_1^{\,2}$, $s_2^{\,2}$, and $s_B^{\,2}$. Also, $s_B^{\,2} = s_1^{\,2}/n_1 + s_2^{\,2}/n_2$.

So, from (12),

$$r = \frac{s_B^{\,4}}{\frac{1}{n_1^{\,2}} \cdot \frac{s_1^{\,4}}{n_1 - 1} + \frac{1}{n_2^{\,2}} \cdot \frac{s_2^{\,4}}{n_2 - 1}} = \frac{\left(\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2} \right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^{\,2}}{n_1} \right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^{\,2}}{n_2} \right)^2},$$

which is the result that we wanted to prove.

For the sake of completeness, we should verify that, given this approximate χ_r^2 distribution for $rs_B^{\,2}/\sigma_B^{\,2}$, the two-sample *t*-statistic does indeed have an approximate *t*-distribution.

Recall from section III that the *t*-distribution with *n* degrees of freedom is defined by

$$t_n = \frac{Z}{\sqrt{V/n}},$$

where Z is $N(0,1)$, V is χ_n^2, and Z and V are independent.

In the one-sample case, we had

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{s^2/\sigma^2}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{\frac{(n-1)s^2/\sigma^2}{n-1}}}.$$

The numerator has distribution $N(0,1)$, $(n-1)s^2/\sigma^2$ has distribution χ_{n-1}^2, and we had to accept the fact that these random variables were independent.

Now in the two-sample case, we have

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^{\ 2}}{n_1} + \frac{s_2^{\ 2}}{n_2}}} = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{s_B} = \frac{\left[(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)\right]/\sigma_B}{\sqrt{\frac{rs_B^{\ 2}/\sigma_B^{\ 2}}{r}}}.$$

The numerator has distribution $N(0,1)$, $rs_B^{\ 2}/\sigma_B^{\ 2}$ has approximate distribution χ_r^2, and so, assuming the independence of these random variables, we have obtained the fact that the two-sample *t*-statistic has an approximate *t*-distribution.

Endnotes

1. Proofs of this are given in many mathematical statistics textbooks, for example, Marx Larsen, *An Introduction to Mathematical Statistics and Its Applications*, 3rd ed., p. 455. Copyright 2001, 1986, 1981 by Prentice-Hall, Inc., Upper Saddle River, NJ 07458

2. We use the subscript *B* here since this is the subscript that Satterthwaite himself used.

3. Yates, Moore, and Starnes, *The Practice of Statistics*, 3rd ed., p. 792. Copyright 2008 by W.H. Freeman and Company, 41 Madison Avenue, New York, NY 10010

Michael Allwood was born in the UK. He attended Oxford University, where he gained an undergraduate degree in mathematics. From 1985 to 1997 he taught mathematics at Westminster School in London − one of the UK's best known private schools. He moved to the US in 1997 and took a position teaching mathematics at Brunswick School in Greenwich, CT, where he is now Chairman of the mathematics department.

Michael has been a reader for AP Statistics since 2001 and a Faculty Consultant for the College Board since 2002.