Understanding the t-test as a variance ratio test, and why $t^2 = F$.

ABSTRACT
This unpublished paper provides a rationale for regarding the t-test statistically as a variance ratio test, on the same basis as Fisher's F-test used in the analysis of variance. It is argued that regarding the t-test in this way both provides a heuristically convincing reason why in the special case of two groups the F statistic reduces to the square of the t statistic, and also gives a unified approach to the modifications of the formula for t in the cases where the variance in the groups has to be calculated separately for the two groups (the Behrens-Fisher problem).

The t-test is normally thought of as an extension of the z-test, suitable for situations when the variance of the population is not known and has to be estimated from the sample variance. The value of "t" is then calculated as the difference between the two sample means divided by the estimated pooled sample standard deviation (in the case of two independent samples, drawn from populations of equal variance). This is exactly analogous to the z-test, which uses the difference in the sample means divided by the known standard deviation.

However, there is another way of looking at the t-statistic, which has the advantage of linking the t-test to the next most complex statistical test, that of the one-way analysis of variance. It is often pointed out that when ANOVA is applied to just two groups, and when therefore one can calculate both a t-statistic and an F-statistic from the same data, it happens that the two are related by the simple formula: $t^2 = F$. This remark is seldom proved in statistics courses, nor even are any heuristic reasons are given why it should be so. This can leave some confusion in the minds of students. Clearly there is some connection between the t-test and two-group, one-way ANOVA, but what is it? The t-statistic is something resembling a z-value, and the F-statistic is essentially the ratio of two estimates of variance: why should there be any relationship between the two?

One way to look at this is to consider the t-statistic as the ratio of two estimates of standard deviation, which yields the square of t as being the ratio of two estimates of variance. Actually however it is more natural to come at it the other way round, and to write down an expression for the ratio of two variance estimates for a population, and derive an expression from it which then is seen to be identical to the usual formula for t^2. This approach even works for the tough case of different population standard deviations, ie the Behrens-Fisher situation, so it has some power as a heuristic approach.

The one-sample case

Why is the square of the t-statistic an estimate of the ratio of two variances? Starting with the simplest case, suppose we are dealing with the situation where a single sample t-test is applicable, where a sample $\{x_i\}$ of size N is being tested against the null hypothesis that it has been drawn from a population with known mean μ but unknown standard deviation σ. Suppose that m is the sample mean, so that

$m = \Sigma \{x_i\}/N$.

Suppose we assume the null hypothesis to be true, and that the sample really *is* drawn from a population of mean μ. Then we can treat the observed value m of the mean as a single sample, taken from the distribution space of all possible means of samples of size N drawn from a population of mean μ.

In this space of the possible values of m, the distribution of means does of course itself have a mean, and a standard deviation. The mean is easily seen to be μ, and the standard deviation is $\sigma/\sqrt{N}$ according to the usual formula, or in other words the variance is σ^2/N. But – and this is the crafty bit – we can gain an unbiassed estimate for the value of the variance for the means also *from our single value of m*. Namely, this estimate is $(m - \mu)^2$.

This last step may seem surprising. This is because we are doing two unusual things here. One is to estimate a parameter from a single value, and the other is to estimate a variance from the average square of the deviations from the known population mean (remember we are assuming the null hypothesis at this point).

Taking the second point first, we are so used to using the formula $s^2 = \Sigma(x_i - m)^2/(N - 1)$ for estimating a variance, that we may forget that it is equally legitimate (at least, the value we get also has the property of being an unbiassed estimate) to estimate the variance from the formula $\Sigma(x_i - \mu)^2/N$ where we have a sample of size N. Indeed, going back to first principles, this is just the *definition* of variance! When, as here, we have a sample of size N = 1 (we only have a single *mean* , however large the underlying sample is), then the formula $(m - \mu)^2$ for an estimate of the variance of the mean, hopefully begins to a bit look more convincing.

I will provide further evidence for this below in the Appendix. In the meantime, assuming this point has been accepted, what can we do with this estimate? We can of course equate it with the formula given above in terms of the unknown *population* standard deviation, and obtain:

$\sigma^2/N = (m - \mu)^2$ and so $\sigma^2 = N*(m - \mu)^2$ where we can now treat the right hand side as an estimate of the *population* variance.

Most readers will rightly have qualms about the sense of this procedure, if what we really want is an estimate of the variance σ^2. But of course we don't. We are really looking for evidence to reject the null. We are trying, like a tricky barrister cross-examining a hostile witness, to get the witness to entangle itself in a contradiction (we pretend that the sample is a witness for the null hypothesis), and we look to trip up the witness, enabling us to reject the null and obtain a significant result.

We have just now got the witness to make a claim about the size of the population variance, of the form "$\sigma^2 = N*(m - \mu)^2$" and in a moment we are going to look at another estimate of the variance which we *know* to be reliable. If the two estimates are discrepant,

we can turn round to the first witness and accuse him of lying. Since the first witness is based on the null hypothesis, and no other questionable assumptions, this means the null must be false! Remember that the argument above about $N*(m - \mu)^2$ being an unbiassed estimate of variance holds *only on the assumption of the null hypothesis.*

So what is our gold standard estimate of the population variance, the equivalent of the Miss Marples character who notices everything and is never systematically wrong, and of course is always a model witness? It is our old friend $s^2 = \Sigma(x_i - m)^2/(N - 1)$, which is an unbiassed estimator of population variance *whether the null hypothesis is true or not.* So if there is a discrepancy between the two, or rather if the discrepancy is large enough to be significant, it can only be because the null hypothesis is false. If the null is true, the only large discrepancies possible can be those due to type I errors.

Therefore we can ask the question, do the two estimates of variance give results which are sufficiently close, or not? Or equivalently, if we take the ratio of the two, are they significantly different from one?

Since the first estimate (assuming the null hypothesis) is $N*(m - \mu)^2$
and the second is s^2, their ratio is

$$N*(m - \mu)^2/s^2$$

or $$(m - \mu)^2/\{s^2/N\}$$

If we take the square root of this we get just $(m - \mu)/\sqrt{\{s^2/N\}}$, which is exactly the usual formula for the t-statistic for the one-sample case. We have shown that in asking the question "is the t-statistic big enough for us to reject the null hypothesis?" we are equivalently asking the question "is the ratio between two estimates of population standard deviation, of which the numerator is only an unbiassed estimator if the null hypothesis is true, large enough to enable us to declare that the numerator is in fact biassed, and the null hypothesis is false?"

One question which might be asked at this point (by anyone who has not by now lost the will to live) is, "why look for an excessively *large* estimate of population variance on the basis of the null hypothesis? Wouldn't it be just as damning to the null if the estimate based on the null being true, were excessively *small*? The answer is that if the null hypothesis is violated, it will be because the actual population from which the sample is taken has a different mean from μ, say μ_1, and in that case it can be proved that the expected value of $(m - \mu)^2$ will be $(\mu - \mu_1)^2 + \sigma^2$, where the second term is the actual variance in both populations. This will always be larger than σ^2, so there is no point in looking for any kind of significance from a *small* ratio of variances: it contradicts the alternative hypothesis just as much as it does the null hypothesis!

The case of two independent groups, equal variances

The argument can be simply extended to the case of independent groups drawn from populations of equal variance, calling this variance as before σ^2. Suppose the samples are of size N_1 and N_2, with means m_1 and m_2, and suppose the within-sample standard deviation estimates are s_1 and s_2, using the usual notation.

Then if, as before, we look at m_1 and m_2 not as actual data but in terms of their possible distributions when the samples are taken many times, both means have distributions, and in particular they have variances. In fact the variance of mean m_1 will be σ^2/N_1 and that of mean m_2 will be σ^2/N_2. Therefore the variance of the difference $m_1 - m_2$ will be the sum of these variances, namely $\sigma^2(1/N_1 + 1/N_2)$.

Note that this deduction holds *whether or not* the populations have the same mean. But for the next step, we need the assumption of the null hypothesis, that the samples are drawn from the *same identical population*, for in this case, using a similar argument to the one-sample case, an unbiassed estimator of the variance of the difference between the two means is $(m_1 - m_2)^2$. This is because the expected value of $(m_1 - m_2)$ is zero. So equating the previous expression of the variance to this estimate gives

$$(m_1 - m_2)^2 = \sigma^2(1/N_1 + 1/N_2).$$

From which we can estimate σ^2 itself as

$$(m_1 - m_2)^2/(1/N_1 + 1/N_2).$$

This is the estimate from the actual sample mean difference, analogous to $N*(m - \mu)^2$ above in the single sample case.

The analogue of the estimate from within the sample is more complicated, because we have two samples and not just one. We can estimate the variance within sample 1 as s_1^2 and that from sample 2 as s_2^2 , and it seems reasonable to combine these by weighting them by the size of the corresponding degrees of freedom: larger samples should give a more accurate estimate of σ^2 than smaller ones.

The actual formula for this weighted variance estimate is

$$\{(N_1 - 1)*s_1^2 + (N_2 - 1)*s_2^2\}/(N_1 + N_2 - 2)\}$$

If we call this "pooled variance estimate" s_p^2, we get, taking the ratio of the two variance estimates,

$$\{(m_1 - m_2)^2/(1/N_1 + 1/N_2)\}/s_p^2$$

$$= \{(m_1 - m_2)^2/\{s_p^2 *(1/N_1 + 1/N_2)\}.$$

Taking square roots of top and bottom, we have the usual formula for a t-statistic, now expressed, if you like to think of it that way, as the ratio of two standard deviations:

$t = (m_1 - m_2)/\sqrt{\{s_p^2 * (1/N_1 + 1/N_2)\}}$.

The Behrens-Fisher case: population variances not necessarily equal

Surprisingly, this approach even works with the difficult Behrens-Fisher problem, in which the samples are drawn from populations which may have *different* variances, but where the null hypothesis is still that they have the same mean.

In this case, $(m_1 - m_2)^2$ is still an unbiassed estimator of the variance of a certain random variable. This variable is a little complicated to define but we can do so as follows: on the assumption that the null hypothesis is true, so $\mu_1 = \mu_2$, say, the statistic is the *difference between the means of samples of size N_1 and N_2 drawn from the two populations respectively* (note that we must now talk about *two* populations even if the null hypothesis is true, because the null hypothesis ensures that the population *means* are equal, but not necessarily the *standard deviations*).

Even without assuming the null hypothesis, we can calculate the variance of the $m_1 - m_2$ statistic using the data internal to the two samples to estimate σ_1 and σ_2. The obvious estimates to take for these are s_1 and s_2 respectively. In that case, the variance of $m_1 - m_2$ will be the sum of the individual variances. The variance of m_1 is then estimated as s_1^2/N_1, and that of m_2 as s_2^2/N_2. So the variance of $m_1 - m_2$ on this calculation is $(s_1^2/N_1 + s_2^2/N_2)$.

Taking the ratio gives us

$(m_1 - m_2)^2/(s_1^2/N_1 + s_2^2/N_2)$

and taking the square root in the usual way gives the standard expression for the Behrens-Fisher statistic t':

$t' = (m_1 - m_2)/\sqrt{(s_1^2/N_1 + s_2^2/N_2)}$.

Which is what we set out to prove, as it shows that the variance-ratio approach produces the same result as the approach in the textbooks.

APPENDIX

The total sum of squares in a t-test can be dissected just as it is in an ANOVA, but the algebra is simpler, considerably so for the one-sample case which is the only one dealt with in detail here. We can write in the usual way:

$$\Sigma(x_i - \mu)^2 = \Sigma(x_i - m + m - \mu)^2 = \Sigma(x_i - m)^2 + \Sigma(m - \mu)^2 + 2\Sigma(x_i - m)(m - \mu).$$

The third term on the right hand side is equal to zero, as $\Sigma(x_i - m) = 0$ by definition of m as the sample mean, so finally

$\Sigma(x_i - \mu)^2 = \Sigma(x_i - m)^2 + N*(m - \mu)^2$ …… (†)

Hidden in this formula, like the animals in a puzzle picture, there are in fact three estimates of the variance of the underlying population which are unbiassed *provided that the null hypothesis is true.* In this case, where we have no doubts that the sample did come from the population we can estimate the unknown variance as $\Sigma(x_i - \mu)^2/N$, with N and not N – 1 in the denominator (the correction to N – 1 is only necessary when we are estimating the population mean from the sample itself and therefore lose a degree of freedom doing so: it is not needed when we *know* the population mean already). Call this estimate, est_1.

There is another way in which we can estimate the variance. The formula $\Sigma(x_i - m)^2/(N - 1)$ is known to provide an unbiassed estimate of variance.

If we write $\Sigma(x_i - m)^2/(N - 1)$ as est_2, the second estimate, the formula (*) above becomes

$N*est_1 = (N - 1)* est_2 + N.(m - \mu)^2$

Suppose we now take the expectations of both sides. Because both est_1 and est_2 are unbiassed, their expectations are both equal to σ^2.

Therefore we get $N*\sigma^2 = (N - 1)*\sigma^2 + E\{N.(m - \mu)^2\}$

And finally $E\{N.(m - \mu)^2\} = \sigma^2$, which for sceptics is the long-awaited proof that $N.(m - \mu)^2$ really does provide an unbiassed estimator for the population variance, on the assumption that the null hypothesis is true.