Statistics and Linguistic Applications
Hale
February 24, 2010

Two-sample test

The same t statistic can be used different ways:

one sample t test	can judge whether to reject the null hypothesis that μ is some particular constant value
two sample t test	can judge whether to reject the null hypothesis that $\mu_a = \mu_b$

Disparity between sample means across two samples could just be due to (random) sampling error, or it could be evidence that that two samples come from populations that indeed have different means. How can we make a probability statement about the wackiness of the two samples we actually got? Following Vasishth §3.16.1, lets rephrase the hypothesis by naming the supposed difference between population means δ.

$$\begin{aligned} \mathcal{H}_0: &\quad \mu_a - \mu_b = \delta = 0 \\ \mathcal{H}_0: &\quad \delta = 0 \end{aligned}$$

The question at hand concerns the difference, δ, between the population means from which two different samples, X_a and X_b, have been obtained. Using elementary properties of variance (see slide 4 on February 17th) we can deduce that $Var[\bar{X}_a] = \frac{\sigma_a^2}{n_a}$ and correspondingly $Var[\bar{X}_b] = \frac{\sigma_b^2}{n_b}$. As a simplifying assumption, let us presume that $\sigma_a^2 = \sigma_b^2 = \sigma^2$, in other words that the samples come from populations whose variance happens to be exactly the same. So the variance of the difference of sample means, $Var[\bar{X}_a - \bar{X}_b] = \frac{\sigma^2}{n_a} + \frac{\sigma^2}{n_b}$. In terms of standard deviations, which are after all just square roots of variances, we have

$$\sigma_{\bar{X}_a - \bar{X}_b} = \sqrt{\frac{\sigma^2}{n_a} + \frac{\sigma^2}{n_b}} = \sqrt{\sigma^2 \left(\frac{1}{n_a} + \frac{1}{n_b} \right)} = \sigma \sqrt{\frac{1}{n_a} + \frac{1}{n_b}} \tag{1}$$

Expression 1 is the standard deviation of the difference of the sample means. It gives us an appropriate denominator for a standardized variable reflecting the sampling distribution of δ. Under the null hypothesis, this quantity would be Normally distributed with mean zero and variance one. The subscript 0 is meant to invoke the dependence on $\mathcal{H}_0$.

$$Z_0 = \frac{\bar{X}_a - \bar{X}_b}{\sigma \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \tag{2}$$

In most scientific situations, we don't know the value of σ and have to estimate it. What if we have a large n_b but only small n_a? Shouldn't our estimate of σ rely more heavily on the bigger sample, X_b? This idea is the basis of the pooled variance, s_p^2. This value is just the weighted average of the two samples' respective variances.

$$s_p^2 = \frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{(n_a - 1) + (n_b - 1)}$$

Replacing σ with the pooled value $\sqrt{s_p^2} = s_p$ in expression 2 yields a new quantity that has a Student's t distribution with $n_a + n_b - 2$ degrees of freedom. This quantity builds-in the assumption that disparities in the sample variances are due to random sampling error.

$$t_0 = \frac{\bar{X}_a - \bar{X}_b}{s_p \sqrt{\frac{1}{n_a} + \frac{1}{n_b}}} \tag{3}$$

Under the assumption of $\delta = 0$, we would expect $100(1 - \alpha)$ percent of the values of t_0 to fall between $-t_{\alpha/2}$ and $t_{\alpha/2}$. A sample producing a t_0 value outside these limits would be unusual if the null hypothesis were true, and is evidence that $\mathcal{H}_0$ should be rejected.

Testing to see if $\sigma_a^2 = \sigma_b^2$

Johnson invokes a theorem that we don't have the wherewithal to prove (yet). Let two independent random samples of sizes n_a and n_b, respectively, be drawn from two Normal populations with variances σ_a^2 and σ_b^2. Then if the variances of the random samples are given by s_a^2 and s_b^2 we have

$$F = \frac{s_a^2/\sigma_a^2}{s_b^2/\sigma_b^2}$$

This ratio follows an F distribution with $n_a - 1$ and $n_b - 1$ degrees of freedom. Under the assumption that both samples come from equi-variable populations, the dependence on the population variance vanishes.

$$F = \frac{s_a^2}{s_b^2}$$

On this $\sigma_a = \sigma_b$ hypothesis, any deviation from a 1:1 ratio between sample variances has to be due to random sampling error. We thus expect F-ratios from equal-variance populations to cluster near 1. If this statistic strays far from 1, this is grounds for abandoning the assumption that X_a and X_b really come from populations with the same amount of variability.

When $\sigma_a^2 \neq \sigma_b^2$

If we don't pool the variance, then standard error of the difference of the means $\sigma_{\bar{X}_a - \bar{X}_b}$ must acknowledge roles for both sources of variability.

$$t_0 = \frac{\bar{X}_a - \bar{X}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \tag{4}$$

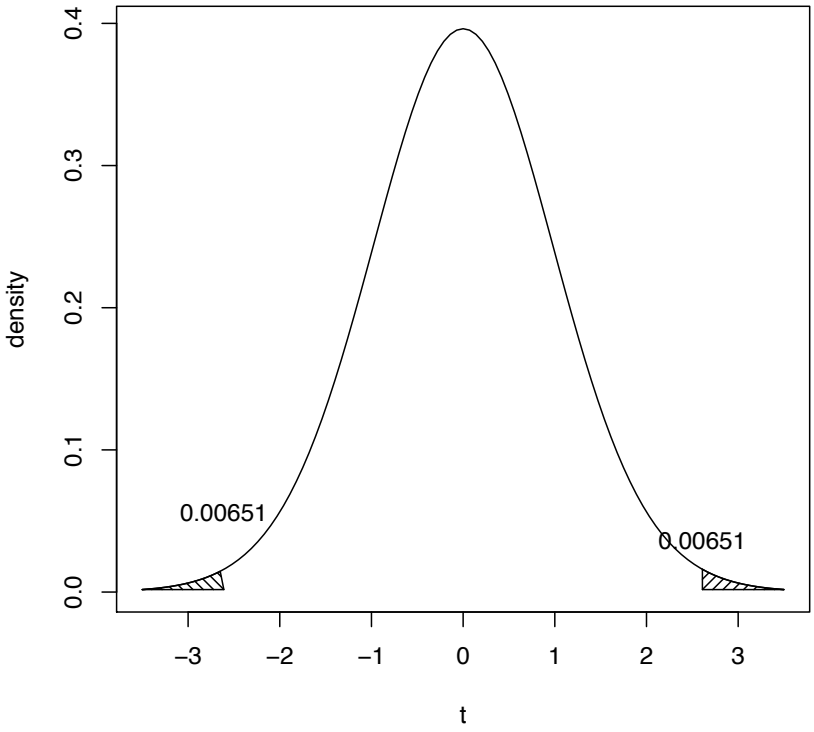
The statistic in 4 does not follow exactly the same t-distribution as the expression in 3. However, it is possible to approximate the "effective" degrees of freedom using the Welch correction, which Johnson writes out on page 78. It is this corrected t-statistic that R calculates when you invoke `t.test` without special parameters, hence the decimal df's.

```
> t.test(VOT[year == "1971"], VOT[year == "2001"])

        Welch Two Sample t-test

data:  VOT[year == "1971"] and VOT[year == "2001"]
t = 2.6137, df = 36.825, p-value = 0.01290
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.480602 51.211706
sample estimates:
mean of x mean of y
113.50000  84.65385
```

The scientific conclusion is that the passage of time, and perhaps the concomitant exposure to English, has reduced Durbin Feeling's voice onset time for consonants like [t] and [k].

Ruling out one direction of VOT change on conceptual grounds leads to a larger rejection region.

```
> shade.tails(2.61, df = 36.825, tail = "both")
```



Paired t test

In a *paired* design, rather than subtracting summaries of multi-measurement samples, one subtracts single measurements of the exact same subject. The experimenter suffers the effects of inopportune sampling to the same degree both times, and it thus cancels out.

In an example borrowed from Butler (1985), each of ten human subjects was asked to read a pair of sentences that contained the same vowel in two different distributional environments.

The dependent variable (tabulated in figure 1) was vowel length. The investigator predicts that Environment 2 is a lengthening environment, whereas Environment 1 is not.

Subject number	Environment 1	Environment 2
1	22	26
2	18	22
3	26	27
4	17	15
5	19	24
6	23	27
7	15	17
8	16	20
9	19	17
10	25	30

Figure 1: Vowel length in unspecified units

The alternative hypothesis is that Environment 2's mean vowel length is greater than Environment 1's. Any lack of difference E1 − E2 or a positive difference is consistent with the null.

```
> env1 <- c(22, 18, 26, 17, 19, 23, 15, 16, 19, 25)
> env2 <- c(26, 22, 27, 15, 24, 27, 17, 20, 17, 30)
> t.test(env1, env2, paired = T, alternative = "less")

        Paired t-test

data:  env1 and env2
t = -2.9531, df = 9, p-value = 0.00807
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.9481568
sample estimates:
mean of the differences
                  -2.5
```