MORE ON THE EQUAL-VARIANCE, TWO-SAMPLE T-TEST

I. Robustness: As pointed out in the quote from Box (in the notes for January 20), and as the discussion (when the example was introduced) of the example about comparing two computer packages suggests, we can't expect the assumptions of an inference procedure to apply exactly. A procedure is said to be *robust* to departures from a model assumption if the results are still reasonably accurate when the assumption is relaxed to some degree. Robustness is sometimes determined by theory, sometimes by computer simulations. For example, in the two-sample t-test above, if our samples are large enough, the Central Limit Theorem tells us that even if X and Y are not normally distributed, the distribution of $\overline{X}$ - $\overline{Y}$ is approximately normal if the sample sizes are large enough, so that the test statistic will still have a distribution that is approximately t with m + n -2 degrees of freedom. Of course, just how large is large enough will depend on the distributions of X and Y. Computer simulations have shown that moderate departures of X and Y from normality have little effect on the distribution of the t-statistic. Simulations have also shown that the equal-variance two-sample t-test is relatively robust to departures from the equal variance assumption, provided the two sample sizes are equal or nearly equal. However, lack of independence can cause serious problems -- the results of a t-test may be very misleading.

II. Another perspective on the two-sample, equal-variance t-test. Those of you who have had regression have seen that a certain t-test is equivalent to a certain F-test. The same is true here. The F-test perspective then allows us to generalize the method to situations where we are comparing more than two means and to some sampling methods other than simple random samples.

First we need more detail on t distributions: A *t-distribution with k degrees of freedom* is defined as the distribution of a random variable of the form $\dfrac{Z}{\sqrt{U/k}}$ where

- Z~N(0,1)
- U~ χ^2(k) (Chi-squared with k degrees of freedom.)
- Z and U are independent.

A *chi-squared distribution with k degrees of freedom* is defined as the distribution of a random variable that is a sum of squares of k independent, standard normal random variables.

The proof that our test statistic T for the equal-variance, two-sample t-test has a t-distribution follows from these facts:

- $$\mathrm{T} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{S^2}{m} + \dfrac{S^2}{n}}} = \frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma^2}{m} + \dfrac{\sigma^2}{n}}} \Bigg/ \sqrt{\frac{(m+n-2)S^2}{\sigma^2(m-n-2)}}$$ (algebra)

- Z = $\dfrac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\dfrac{\sigma^2}{m} + \dfrac{\sigma^2}{n}}}$ is standard normal (seen earlier)
- U = $\dfrac{(m+n-2)S^2}{\sigma^2}$ is chi-squared with m + n -2 degrees of freedom. (Can be proved using model assumptions)
- U and Z are independent (Can be proved using model assumptions.)

An *F-distribution F(ν_1, ν_2) with ν_1 degrees of freedom in the numerator and ν_2 degrees of freedom in the denominator* is the distribution of a random variable of the form $\dfrac{W/\nu_1}{U/\nu_2}$, where

- W ~ $\chi^2(\nu_1)$
- U ~ $\chi^2(\nu_2)$, and
- U and W are independent.

If we have a t random variable of the form T = $\dfrac{Z}{\sqrt{U/k}}$, where U and Z are as in the definition of t-distribution, then

$$T^2 = \frac{Z^2}{U/k}.$$

Now Z^2 is a chi-squared random variable with 1 degree of freedom, and U is chi-squared with k degrees of freedom, so T^2 is an F-distribution with 1 degree of freedom in the numerator and k degrees of freedom in the denominator. So we could do any t-test (with two-sided alternative) as an F-test, by using the square of the t-statistic.

III. (*Optional – I won't go over this in class.*) **Still Another Perspective**

Looking at the square of the t-statistic for the two-sample, equal-variance t-test in the case of equal sample sizes will give us some insight into generalizing the F-test to work for more than one sample and, eventually, to some other sampling designs as well.

Under the null hypothesis $\mu_X = \mu_Y$, the t-statistic is

$$T = \frac{\bar{X} - \bar{Y}}{S\sqrt{\dfrac{1}{m} + \dfrac{1}{n}}},$$

Our additional restriction of equal sample sizes means m = n. So

$$S^2 = \frac{(n-1)S_x^2 + (n-1)S_Y^2}{(n-1)+(n-1)} = \frac{S_x^2 + S_Y^2}{2}$$

and

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\left(S_x^2 + S_Y^2\right)}{2}}\sqrt{\frac{2}{n}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\left(S_x^2 + S_Y^2\right)}{n}}}.$$

Then our F statistic is

$$T^2 = \frac{\left(\overline{X} - \overline{Y}\right)^2}{\left[\frac{\left(S_x^2 + S_Y^2\right)}{n}\right]},$$

which is equivalent to

$$\frac{\frac{n}{2}\left(\overline{X} - \overline{Y}\right)^2}{\frac{1}{2}\left(S_x^2 + S_Y^2\right)}.$$

With our assumption that m = n, the denominator in this re-expression is just our pooled estimator of σ^2, the common variance of the two populations.

If the null hypothesis is true, then the two distributions (of X and Y) are the same -- so we may consider our two samples to be two samples of size n from the same $N(\mu,\sigma^2)$ distribution. But we know that the sample means of samples of size n from an $N(\mu,\sigma^2)$ distribution have an $N(\mu,\sigma^2/n)$ distribution (the sampling distribution). Now the sample variance of a distribution is an unbiased estimator of the population variance of that distribution. Applying this to our $N(\mu,\sigma^2/n)$ sampling distribution, we conclude that the random variable

$$S_b^2 = \frac{\left(\overline{X} - \frac{\overline{X} + \overline{Y}}{2}\right)^2 + \left(\overline{Y} - \frac{\overline{X} + \overline{Y}}{2}\right)^2}{2-1}$$

(i.e., the sample variance for the sample $\left\{\overline{X},\overline{Y}\right\}$ from the distribution of sample means) is an unbiased estimator of σ^2/n. (The b stands for "between sample")

Using algebra,

$$S_b^2 = \left(\frac{\overline{X} - \overline{Y}}{2}\right)^2 + \left(\frac{\overline{X} - \overline{Y}}{2}\right)^2 = \frac{1}{2}\left(\overline{X} - \overline{Y}\right)^2.$$

Thus, if the null hypothesis is true, the numerator $\frac{n}{2}\left(\overline{X} - \overline{Y}\right)^2$ of T^2 is an unbiased estimator of σ^2, so we expect the quotient in T^2 to be close to 1. It can be proved that if the null hypothesis is false, then the ratio T^2 is *greater* than 1. So the F-test (equivalent to the t-test) can be interpreted as a test for the ratio of two estimates of σ^2.

This idea can be generalized to more than two samples: We form the sample variance for each sample, take the mean of these sample variances as one estimate of the common population variance σ^2, and compare with a "between sample" estimate of σ^2. With suitable modifications, this works, and is the idea behind the method of Analysis of Variance. However, we may, as above, multiply the numerator and denominator in the F-statistic by constants to make interpretations and/or formulas easier. In the notation used in the textbook, for the special case n = m considered here, we would express the F-statistic as

$$\frac{SST}{SSE/(2n-2)},$$

where SST (the *sum of squares for treatments* or *treatment sum of squares*) is

$$\text{SST} = \left(\bar{X} - \frac{\bar{X}+\bar{Y}}{2}\right)^2 + \left(\bar{Y} - \frac{\bar{X}+\bar{Y}}{2}\right)^2,$$

and SSE (the *sum of squares for error* or *error sum of squares*) is

$$\text{SSE} = \sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

($\frac{\bar{X}+\bar{Y}}{2}$ is sometimes called the *grand mean*, abbreviated GM.)

Exercises: *(These exercises can be helpful to get you familiar with working with the notation.)*

i. Go through the algebra to check that the two expressions for the F-statistic are equivalent.

ii. Express GM, SST and SSE using the following notation:

The sample from the first random variable is $Y_{11}, Y_{12}, \ldots, Y_{1n}$, and the sample from the second random variable is $Y_{21}, Y_{22}, \ldots, Y_{2n}$. (In other words, the random variable representing the t^{th} observation from the i^{th} population, for i = 1,2, is Y_{it}.) We will need to use double subscripts when we go to more than 2 populations; typically, populations in this class will be defined by treatments.