Statistical Methods I

EXST 7005 Course notes

James P Geaghan
Louisiana State University

The two-sample t-test

H_0: $\mu_1 - \mu_2 = \delta$

H_1: $\mu_1 - \mu_2 \neq \delta$,

a non-directional alternative (one tailed test) would specify a difference, either $>\delta$ or $<\delta$.

Commonly, δ is 0 (zero)

If H_0 is true, then

$$E(\bar{d}) = \mu_1 - \mu_2$$

$$\sigma_{\bar{d}}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If values of σ_1^2 and σ_2^2 were KNOWN, we could use a Z-test, $Z = \frac{\bar{d} - \delta}{\sigma_{\bar{d}}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

If values of σ_1^2 and σ_2^2 were NOT KNOWN, and had to be estimated from the samples, we would use a t-test, $t = \frac{\bar{d} - \delta}{S_{\bar{d}}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$.

Since the hypothesized difference is usually 0 (zero), the term $(\mu_1 - \mu_2)$ is usually zero, and the equation is often simplified to $t = \frac{\bar{d}}{S_{\bar{d}}}$,

Et voila, a two sample t-test!

This is a very common test, and it is the basis for many calculations used in regression and analysis of variance (ANOVA). It will crop up repeatedly as we progress in the course. It is very important!

$t = \frac{\bar{d} - \delta}{S_{\bar{d}}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$, often written just $t = \frac{\bar{d}}{S_{\bar{d}}}$ when δ or $\mu_1 - \mu_2$ is equal to zero.

The two-sample t-test

Unfortunately, this is not the end of the story. It turns out that there is some ambiguity about the degrees of freedom for the error variance, $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$. Is it n_1-1, or n_2-1, or somewhere in between, or maybe the sum?

Power considerations

POWER! We want the greatest possible power. It turns out that we get the greatest power (and our problems with degrees of freedom go away) if we can combine the two variance estimates into one, single, new and improved estimate! But we can only do this if the two variances are not different.

We can combine the two variance estimates into a single estimate if they are not too different. To determine if they are sufficiently similar we use an F test. Therefore, two-sample t-tests START WITH AN F TEST!

Pooling variances

If the two estimates of variance are sufficiently similar, as judged by the F test of variances (e.g. $H_0: \sigma_1^2 = \sigma_2^2$), then they can be combined. This is called "pooling variances", and is done as a weighted mean (or weighted average) of the variances. The weights are the degrees of freedom.

Weighted means or averages

The usual mean is calculated as $\bar{Y} = \sum_{i=1}^{n} Y_i \Big/ n$. The weighted mean is $\bar{Y} = \sum_{i=1}^{n} w_i Y_i \Big/ \sum_{i=1}^{n} w_i$, or the sum of the variable multiplied by the weights divided by the sum of the weights.

Pooled variances are calculated as Pooled $S^2 = S_p^2 = \sum_{j=1}^{k} \gamma_i S_i^2 \Big/ \sum_{j=1}^{k} \gamma_i$ where j will be j = 1 and 2 for groups 1 and 2. There could be more than 2 variances averaged in other situations.

Recall that $\gamma_j S_j^2 = SS_j$, so we can also calculate the sum of the corrected SS for each variable divided by the sum of the d.f. for each variable $S_p^2 = \sum \gamma_j S_j^2 \Big/ \sum \gamma_j = \sum SS_j \Big/ \sum \gamma_j$

Pooled variance calculation $S_p^2 = \dfrac{\gamma_1 S_1^2 + \gamma_2 S_2^2}{\gamma_1 + \gamma_2} = \dfrac{SS_1 + SS_2}{\gamma_1 + \gamma_2} = \dfrac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)}$

Two sample t-test variance estimates

From linear combinations we know that the variance of the sum is the sum of the variances. This is the GENERAL CASE. $\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}$. But, if we test $H_0: \sigma_1^2 = \sigma_2^2$ and fail to reject, we can pool the variances. The error variance is then $S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)$. One additional minor simplification is possible. If $n_1 = n_2 = n$, then we can place the pooled variance over a single n, $2S_p^2 \big/ n$.

So we now have a single, more powerful, pooled variance! What are it's degrees of freedom?

The first variance had a d.f.= $n_1 - 1 = \gamma_1$

The second variance had d.f.= $n_2 - 1 = \gamma_2$