A special case – the paired t-test

One last case. In some circumstances the observations are not separate and distinct in the two samples. Sometimes they can be paired. This can be good, adding power to the design.

For example:

We want to test toothpaste. We may pair on the basis of twins, or siblings in assigning the toothpaste treatments.

We want to compare deodorants or hand lotions. We assign one arm or hand to one brand an the other to another brand.

In may drug and pharmaceutical studies done on rats or rabbits the treatments are paired on litter mates.

So, how does this pairing affect our analysis? The analysis is done by subtracting one category of the pair from the other category of the pair. In this way the pair values become difference values.

As a result, what originally appears to be a "two-sample t-test" becomes a one-sample t-test of the differences between the two members of each pair.

So, in many ways the paired t-test is easier.

Example: We already did an example of this type of analysis. Recall the Lucerne flowers whose seeds we compared for flowers at the top and bottom of the plant. This was paired and we took differences. The difference was "1" with a standard error of "0.5055".

SAS example 2c examined previously

```
            Tests for Location: Mu0=0
Test           -Statistic-    -----p Value------
Student's t    t  1.978141    Pr > |t|    0.0793
Sign           M         2    Pr >= |M|   0.3438
Signed Rank    S      19.5    Pr >= |S|   0.0469
```

So the paired t-test is an alternative analysis for certain data structures. It is better because it eliminates the "between pair" variation and compares the treatments "within pairs". This reduces variance.

However, note that the degrees of freedom are also cut in half. If the basis for pairing is not good, the variance is not reduced, but degrees of freedom are lost.

Summary

The SAS PROC TTEST provides all of the tests needed for two-sample t-tests. It provides the test of variance we need to start with, and it provides two alternative calculations, one for equal variance and one for unequal variance. We choose the appropriate case.

We also saw that several previous calculations, such as confidence intervals and sample size, are also feasible for the two-sample t-test case.

The paired t-test, where there is a good strong basis for pairing observations, can gain power by reducing between pair variation. However, if the basis for pairing is not good, we lose degrees of freedom and power.

Calculating a needed sample size

The Z-test and t-test use a similar formula. $Z = \dfrac{\overline{Y} - \mu_0}{\sqrt{\dfrac{\sigma^2}{n}}} = \dfrac{\overline{Y} - \mu_0}{\sigma / \sqrt{n}}$

Let's suppose we know everything in the formula except n. Do we really? Maybe not, but we can get some pretty good estimates.

Call the numerator ($\overline{Y} - \mu_0$) a difference, $\overline{d}$. It is some mean difference we want to be able to detect, so $\overline{d} = \overline{Y} - \mu_0$

The value σ^2 is a variance, the variance of the data that we will be sampling. We need this variance, or an estimate, S^2.

So we alter the formula to read. $Z = \dfrac{\overline{d}}{\sqrt{\dfrac{\sigma^2}{n}}} = \dfrac{\overline{d}}{\sigma / \sqrt{n}}$

What other values do we know? Do we know Z? No, but we know what Z we need to obtain significance. If we are doing a 2-tailed test, and we set $\alpha = 0.05$, then Z will be 1.96.

Any calculated value larger will be "more significant", any value smaller will not be significant.

So, if we want to detect significance at the 5% level, we can state that ...

We will get a significant difference if $Z = \dfrac{\overline{d}}{\sqrt{\dfrac{\sigma^2}{n}}} = \dfrac{\overline{d}}{\sigma / \sqrt{n}} \geq Z_{\alpha/2}$

We square both sides and solve for n. Then we will also SHOULD get a significant difference if $n \geq \dfrac{Z^2_{\alpha/2}\sigma^2}{\overline{d}^2}$. Then, if we know the values of $\overline{d}, \sigma^2$ and Z, we can solve the formula for n. If we are going to use a Z distribution we should have a known value of the variance (σ^2). If the variance is calculated from the sample, use the t distribution. This would give us the sample size needed to obtain "significance", in accordance with whatever Z value is chosen.

Generic Example

Try an example where

$\overline{d} = 2$

$\sigma = 5,\ \sigma^2 = 25$

$Z = 1.96$

So what value of n would detect this difference with this variance and produce a value of Z equal to 1.96 (or greater)?

$n \geq \dfrac{Z^2_{\alpha/2}\sigma^2}{\overline{d}^2} = (1.96^2 * 5^2)/2^2 = 3.8416(25)/4 = 24.01$

since n $\geq$ 24.01, round up to 25.

Answer, n ≥ 25 would produce significant results. Guaranteed? Wouldn't this always produce significant results? Theoretically, within the limits of statistical probability of error, yes, but only if the difference was really 2. If the null hypothesis (no difference, $\mu = \mu_0$) was really true and we took larger samples, then we would get a better estimate of 0, and may never show significance.

Considering Type II Error

The formula we have seen contains only $Z_{\alpha/2}$ or $t_{\alpha/2}$, depending on whether we have σ^2 or S^2. However, a fuller version can contain consideration of the probability of Type II error (b).

We can often use Z when working with very large samples.

Remember that to work with TYPE II or β error we need to know the mean of the real distribution. However, in calculating sample size we have a difference, $\overline{d} = \overline{Y} - \mu_0$. So we can include consideration of type II error and power in calculating the sample size. The consideration of β error would be done by adding another Z or t for the error rate. Notice that below I switch to t distributions and use $n \geq \frac{(t_{\alpha/2} + t_{\beta})^2 S^2}{\overline{d}^2}$.

Other examples

We have done a number of tests, some yielding significant results and others not. If a test yields significant results (showing a significant difference between the observed and hypothesized values), then we don't need to examine sample size because the sample was big enough. However, some utility may be made of this information if we FAIL to reject the null hypothesis.

Note: Some textbooks give only the formula I originally gave for Z, without the β error consideration. What is the power if you use the formula omitting t_{β} from $n \geq \frac{(t_{\alpha/2} + t_{\beta})^2 S^2}{\overline{d}^2}$?

If you set t_{β} equal to zero the power is 0.50 and there is a 50% chance of making a Type II error.

An example with t values and β error included

Recall the Rhesus monkey experiment. We hypothesized no effect of a drug, and with a sample size of 10 were unable to reject the null hypothesis. However, we did observe a difference of +0.8 change in blood pressure after administering the drug. What if this change was real? What if we made a Type II error? How large a sample would we need to test for a difference of 0.8 if we also wanted 90% power?

So we want to know how large a sample we would need to get significance at the α=0.05 level if power was 0.90. In this case β = 0.10. To do this calculation we need a two tailed α and a one tailed β (we know that the observed change is +0.8). We will estimate the variance from the sample so we will use the t distribution. However, since we don't know the sample size, we don't know the degrees of freedom! Since we do not know the d.f. we will start off with some "reasonable" values for t_{α} and t_{β}. Then after we solve the equation we will have an estimate of the d.f. We can solve again with better values of t_{α} and t_{β}, and refine our estimate. After our second calculation we have even better estimates of d.f., so

we get new values for t_α and t_β and redo the calculations, etc, etc, until the estimate stabilizes.

So we will approximate to start with. Given the information,

$\alpha = 0.05$, so the value of t will be approximately 2

$\beta = 0.10$, so the value of t will be roughly 1.3

$\overline{d} = \overline{Y} - \mu_0 = 0.8$ from our previous results,

$S^2 = 9.0667$ from our previous results.

$$n \geq \frac{(t_{\alpha/2} + t_\beta)^2 S^2}{\overline{d}^2}$$

We do the calculations. $n \geq \frac{(2+1.3)^2 9.0667}{(0.8)^2} = \frac{(3.3)^2 9.0667}{0.64} = 154.27$

And now we have an estimate of n and the degrees of freedom, n = 155 and d.f. =154. We can refine our values for $t_{\alpha/2}$ and t_b.

for d.f. = 154, $t_{\alpha/2} = 1.97$ approx.

for d.f. = 154, $t_b = 1.287$ approx.

So we redo the calculations with improved estimates.

$$n \geq \frac{(1.97+1.287)^2 9.0667}{(0.8)^2} = \frac{(3.257)^2 9.0667}{0.64} = 150.28$$

A little improvement! If we saw much change in the estimate of *n*, we could recalculate as often as necessary. Usually 3 or 4 recalculations are enough.

Summary

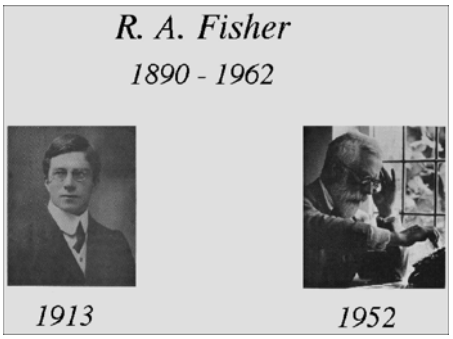
We developed a formula for calculating sample sizen that can be adapted for either t or Z distributions, $n \geq \frac{(t_{\alpha/2} + t_\beta)^2 S^2}{\overline{d}^2}$

We learned that we need input values of α, β, S^2 (or σ^2 for Z tests) and a value for the size of the difference to be detected ($\overline{d}$).

For the t-test, the first calculation was only approximate since we didn't know the degrees of freedom. However, after the initial calculation the estimate could be improved by the iterative recalculation of the estimate of *n* until it was stable.

Analysis of Variance (ANOVA)

R. A. Fisher – resolved a problem that had existed for some time. The hypothesis to be tested is H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ versus the alternative H_1: some μ_i is different. Conceptually, we have separate (and independent) samples, each giving a mean, and we want to know if they could have all come from the same population, or if is more likely that at least one came from a different population.



One way to do this is a series of t-tests.

If we want to test among 3 means we do 3 tests: 1 versus 2, 1 versus 3, 2 versus 3

For 4 means there are 6 tests. 1–2, 1–3, 1–4, 2–3, 2–4, and 3–4

For 5 means, 10 tests, etc.

This technique is unwieldy, and has other issues. When we do the first test, there is an α chance of error, and for each additional test another α chance of error. So if you do 3 or 6 or 10 tests, the chance of error on each and every test is α.

Overall, for the experiment, the chance of error for all tests together is much higher than α.

Bonferroni gave a formula that showed that the chance of error would be NO MORE than $\Sigma\alpha_i$. So if we do 3 tests, each with a 5% chance of error, the overall probability of error is no greater than 15%, 30 percent for 6 tests, 50% for 10 tests, etc.

Of course this is an upper bound. Other calculations are probably more realistic such as the calculation $\alpha' = 1-(1-\alpha)^{k-1}$ used by Duncan or $\alpha' = 1-(1-\alpha)^{k/2}$ from the Student-Newman-Keuls calculation (where k is the number of groups to be tested, α is the error rate for each test and α' is the error rate for the collection of tests). The table below gives some probabilities of error calculated by Bonferroni's, Duncan's and Student-Newman-Keuls' formulas for tests done at $\alpha = 0.05$.

Number of means	Pairwise tests	$(1-\alpha)$	Bonferroni (upper bound)	Duncan $[1-(1-\alpha)^{k-1}]$	Student-Newman-Keuls $[1-(1-\alpha)^{k/2}]$
2	1	0.95	0.05	0.0500	0.0500
3	3	0.86	0.15	0.0975	0.0741
4	6	0.74	0.30	0.1426	0.0975
5	10	0.6	0.50	0.1855	0.1204
6	15	0.46	0.75	0.2262	0.1426
7	21	0.34	1.05	0.2649	0.1643
10	45	0.1	2.25	0.3698	0.2262
50	1225	0	61.25	0.9190	0.7226

The bottom line: Splitting an experiment into a number of smaller tests is generally a poor idea. This applies at higher levels as well (i.e. splitting big ANOVAs into little ones). The solution: We need ONE test that will give us an accurate test with an α value of the desired level.

The concept

We are familiar with variance. $S^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} = \frac{SS}{d.f.}$

We are familiar with the pooled variance $S_p^2 = \frac{\gamma_1 S_1^2 + \gamma_2 S_2^2}{\gamma_1 + \gamma_2} = \frac{SS_1 + SS_2}{(n_1 - 1) + (n_2 - 1)}$

We are familiar with the variance of the means. But we never get "multiple" estimates of the mean and calculate a variance from those. The calculation we use to get the variance of the means comes from statistical theory, $S_{\bar{Y}}^2 = \frac{S^2}{n}$. Could we actually get multiple estimates of the means and calculate a sum of squared deviations of the various means from an overall mean and get variance of the means from that?

Yes, we could, and using the formula $S_{\bar{Y}}^2 = \frac{S^2}{n} = \frac{\sum_{i=1}^{k} (\bar{Y}_{i.} - \bar{\bar{Y}}_{..})^2}{k-1}$ should give the same value.

Suppose we have some values from a number of different samples, perhaps taken at different sites. The values would be Y_{ij}, where the sites are i=1, 2, ..., k, and the observations from within the sites are j = 1, 2, 3, ..., n_i. For each site we calculate a value of the mean. We then take the various means (k different means) and calculate a variance among those. This would also give the "variance of the means".

The LOGIC

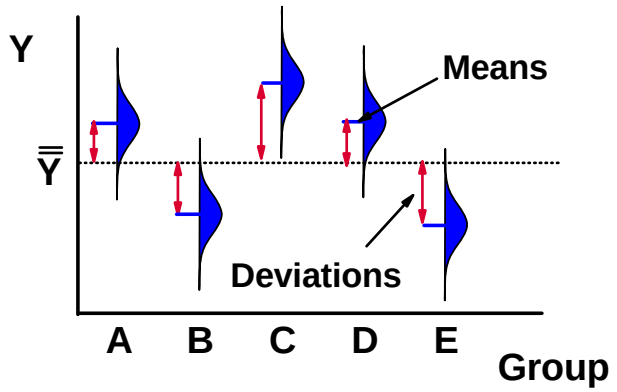
Remember, we want to test

H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$

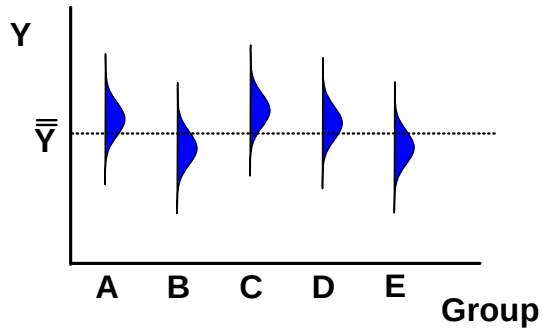
We have a bunch of means and we want to know if they were drawn from the same population or different populations. We also have a bunch of samples each with its own variance (S^2). If we can assume homogeneous variance (all variances equal) then we could POOL the multiple estimates of variance. So, to start with we will take the variances from each of the groups and pool them into one new & improved estimate of variance. This will be the very best estimate of variance that we will get (if the assumption is met).

$$S_p^2 = \frac{SS_1 + SS_2 + SS_3 + SS_4 + SS_5}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1) + (n_5 - 1)}$$

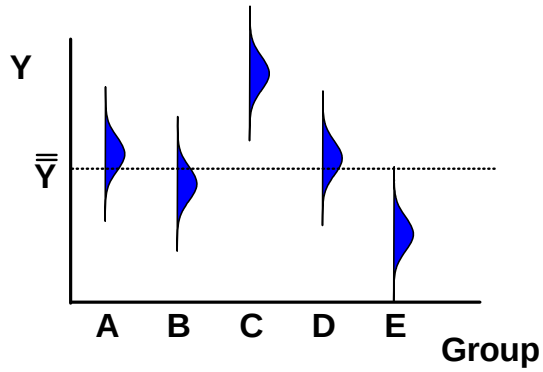
Now, think about the means. If the NULL HYPOTHESIS IS TRUE, then we could calculate the variance of the means from the multiple means. This would estimate $S_{\bar{Y}}^2$, the variance of the means. We would take the deviations of each $\bar{Y}_{i.}$ from the overall mean, $\bar{\bar{Y}}_{..}$, and get a variance from that.



If the null hypothesis is true, the means should be pretty close to the overall mean. They won't be exactly equal to the overall mean because of random sampling variation in the individual observations.



However, if the null hypothesis is false, then some mean will be different! At least one, maybe several.



So we take the Sum of squared deviations, divide by the degrees of freedom and we get an estimate of the variance of the means, $S^2_{\bar{Y}} = \frac{\sum_{i=1}^{k}(\bar{Y}_{i.} - \bar{\bar{Y}}_{..})^2}{k-1}$. But this does not exactly estimate the variance, it estimates the variance of the means, that is the variance divided by the sample size! The sample size is the number of observations in each mean. $S^2_{\bar{Y}} = \frac{\sum_{i=1}^{k}(\bar{Y}_{i.} - \bar{\bar{Y}}_{..})^2}{k-1} = \frac{S^2}{n}$. In order to estimate the variance we must multiply this estimate by *n*, the sample size, $nS^2_{\bar{Y}} = \frac{nS^2}{n} = S^2$, giving a second estimate of the variance. This is obviously easier if each sample size is the same (i. e. the experiment is balanced). We will usually use the calculations for a balanced design, but the analysis can readily be done if the data is not balanced. It's just a little more complicated.

The Solution

So what have we got?

One variance estimate that is pooled across all of the samples because the variances are equal (an assumption, sometimes testable). This is the best estimate of random error.

And another variance that should be the same IF the null hypothesis is TRUE.

The second mean (from the variances) may not be the same if the null hypothesis is false, depending on how great the departure from the null hypothesis. Not only will the second

variance from the mean not be the same, IT WILL BE LARGER! Why? Because when we are testing means for equality we will not consider rejecting if the means are too similar, only if they are too different and large differences in means yield large deviations which produce an overly large variance. So this will be a one tailed test.

And how to we go about testing these two variances for equality? Testing for equality of variances requires an F-test, of course.

If H_0: $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ is true, then $S_p^2 = nS_{\bar{Y}}^2$

If H_1: some μ_i is different, then $S_p^2 < nS_{\bar{Y}}^2$

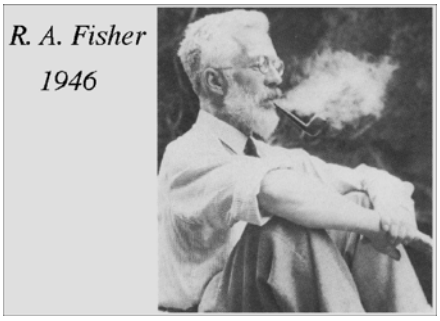
For a one tailed F test we put the ONE WE EXPECT TO BE LARGER IN THE NUMERATOR.

$$F = nS_{\bar{Y}}^2 \Big/ S_p^2$$

And that is Analysis of Variance.

We are actually testing means, but we are doing it by turning them into variances; one pooled variance from within the groups, called the "pooled within variance" and one variance from between groups or among groups called the "variance among groups" or "between group variance". If the variances are not significantly different as judged by the F test, then we cannot reject the null hypothesis. It is possible, as usual, that we make a Type II error with some unknown probability (β). If the variances are judged to not be the same, then the null hypothesis is probably not true. Of course we may have made a Type I error, with a known probability of α.

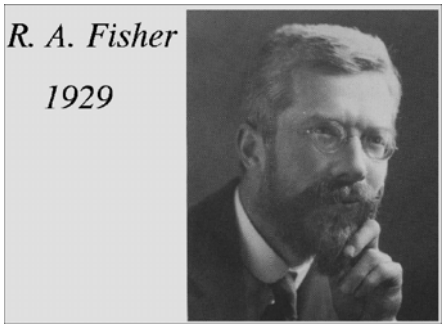
Some of the calculations later, but this is the basic idea.



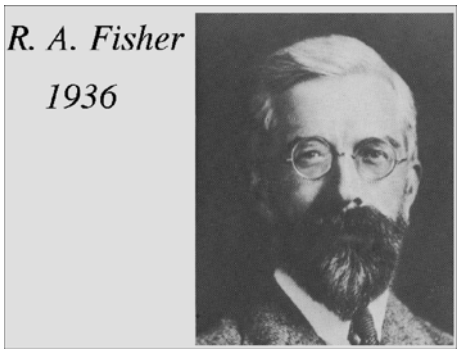
R. A. Fisher

Ronald Aylmer Fisher is sometimes called the father of modern statistics. Some of his major contributions include the development of the basics of design of experiments and Analysis of Variance.

Born in London 1890, he had very poor eyesight that prevented him from learning by electric light. He had to learn by having things read out to him. He developed the ability view problems geometrically and to figure mathematical equations in his head. In 1909 he won a scholarship to Cambridge.



He left an academic position teaching mathematics for a position at Rothamsted Agricultural Experiment Station. In this environment he developed many applied analyses for testing experimental hypotheses (Analysis of Variance, circa 1918), and provided much of the foundation for modern statistics.



We will see other analyses (in addition to ANOVA) developed by Fisher. Some other contributions by Fisher include the first use of the term "null hypothesis", development of the F distribution, of the Least Significant Difference, maximum likelihood estimation and contributed to the early use nonparametric statistics.

Terminology used in Analysis of Variance

Treatment – different experimental populations that are contained in an experiment and undergo some application or manipulation by the experimenter

Control or check – a "treatment" that receives no experimental manipulation

Experimental Unit – the unit to which a treatment is applied

Sampling Unit – the unit that is sampled or measured

The linear model is given by $Y_{ij} = \mu_i + \varepsilon_{ij}$ or $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

where $\tau_i = (\mu_i - \mu_.)$ is estimated by $\hat{\tau}_i = (\bar{Y}_{i.} - \bar{Y}_{..})$

The calculation of treatment Sum of Squares for treatments is a sum of the squared treatment effects $SS_{Treatments} = n\sum_{i=1}^{t}(\bar{Y}_{i.} - \bar{Y}_{..})^2$.

The calculation of treatment Mean Square is a sum of squared effects divided by the degrees of freedom. A variance? $MS_{Treatments} = \frac{n\sum_{i=1}^{t}(\bar{Y}_{i.} - \bar{Y}_{..})^2}{t-1} = \frac{n\sum_{i=1}^{t}\tau_i^2}{t-1}$

A random treatment effect estimates a variance component. In order for treatments to be random, they should be a random selection from a large (theoretically ∞) number of treatments. Inferences developed from random treatments are for all the possible treatment levels.

Examples of random effects

The term used for the error in an experiment are always random. They represent random variation. This variation comes from the experimental unit and sometimes the sampling unit.

Compare production rice varieties, where rice varieties represent a random sample from the world's rice varieties.

Estimate the alcohol content of beer, where the beers tested are randomly sampled from all the beers in the population of interest (world, national).

Oxygen levels in bayous, where randomly selected bayous represent all bayous in the state.

A treatment is FIXED if all possible levels, or all levels of interest, are included in the experiment. The treatment levels are selected by the investigator and are probably not chosen from a very large number of possible values.

A fixed treatment estimates the sum of squared fixed effects for the treatments being investigated. This is NOT a variance, but the calculation is the same, $\sum_{i=1}^{t}\tau_i^2 \Big/ t-1$.

Examples of fixed effects

- Experiment includes all of the 7 rice varieties commonly grown in Louisiana
- Beers are limited to the 5 micro-breweries in Anchorage, Alaska.