Tales of the Terrible Twos

(Why can't I get a straight answer at this *t* party?)

compiled by

Chris Olsen
Cedar Rapids Community Schools
Cedar Rapids, IA 52403
COlsen@cr.k12.ia.us

Part 1: The Sampling Distributions

Preliminaries:

Up front, we need to make crystal clear that in all cases our sampling is simple and random. We may weave and dodge and mumble about assumptions in what follows, but there is no effective substitute for random sampling. While there are alternatives to simple random sampling in experiments, they all build on SRS. Now, for the REST of the story…

The two samples may be dependent or independent.

If the samples are dependent the quantities are subtracted to get differences, x_d, and X_d and s_d are calculated based on a sample of *n* pairs of differences…

A short note up front: although we usually (or at least sometimes) imagine that we are sampling from normally distributed populations, this is not a necessary requirement for any of the results for $\bar{X}_d$ that follow. We will note that for what follows, it must be so that

$$\frac{\bar{X}_d - \mu_d}{\sigma_d / \sqrt{n}} \text{ is normally distributed with mean 0 and variance 1.}$$

…If *n* is "large"…

… If the population of X_d is normal with mean μ_d, then the distribution of $\frac{\bar{X}_d - \mu_d}{s_d / \sqrt{n}}$ is approximately standard normal because as $n \to \infty$, $s_d^2 \to \sigma_d^2$. (This happy circumstance is known as "consistency" of the estimator.) For those who remember the definition of the limit from elementary calculus, an estimator, $\hat{\theta}$, is said to be **consistent** if for any $\varepsilon > 0, P\left(\left|\hat{\theta} - \theta\right| \geq \varepsilon\right) \to 0$ as $n \to \infty$. That is, $\hat{\theta}$ is consistent if, as the sample size gets larger, it is less and less likely that $\hat{\theta}$ will be further than ε from the true value of θ.

If the samples are dependent the quantities are subtracted to get differences, x_d, and $\overline{X}_d$ and s_d are calculated based on a sample of *n* pairs of differences…

…If *n* is "large"…

… If the population of X_d is not normal, then the distribution of $\dfrac{\overline{X}_d - \mu_d}{s_d/\sqrt{n}}$ is approximately standard normal, due to the Central Limit Theorem and the consistency mentioned above. Just how large *n* must be to make this happen depends on the nature of the non-normality of X_d; in any case, this is NOT our problem; it is a problem for the mathematical statisticians.

…If *n* is "small," s_d doesn't have much of a chance to be consistent, and…

… If the population of X_d is normal with mean μ_d, then the distribution of $\dfrac{\overline{X}_d - \mu_d}{s_d/\sqrt{n}}$ is a *t* distribution with n - 1 degrees of freedom.

… If the population of X_d is not normal,
then a transformation of the variable or a non-parametric technique may be used.

If the samples of size n_a from populations A, and size n_b from population B are independent the quantities $\overline{X}_a$, $\overline{X}_b$, s_a, and s_b are calculated.

…If n_a and n_b are "large"…

$\overline{X}_a - \overline{X}_b$ will be approximately normal by virtue of the Central Limit Theorem, with mean $\mu_a - \mu_b$, and estimated variance $\dfrac{s_a^{\,2}}{n_a} + \dfrac{s_b^{\,2}}{n_b}$ (again because of the consistency of the sample variance. Then $\dfrac{\left(\overline{X}_a - \overline{X}_b\right) - \left(\mu_a - \mu_b\right)}{\sqrt{\dfrac{s_a^{\,2}}{n_a} + \dfrac{s_b^{\,2}}{n_b}}}$

will be distributed as approximately standard normal. Just how large n_a and n_b must be to make this happen depends on the nature of the non-normality of populations A and B. In any case, this also is NOT our problem; it is a problem for the mathematical statisticians.

If the samples of size n_a from populations A, and size n_b from population B are <u>independent</u> the quantities $\overline{X}_a$, $\overline{X}_b$, s_a, and s_b are calculated.

…Even if n_a and n_b are <u>not both "large"</u>…

…If both populations are <u>normally distributed</u>…

…If the population variances, σ_a^2 and σ_b^2 are <u>equal</u>, then s_a^2 and s_b^2 are independent estimates of the common population variance, σ^2. The data from the two samples are combined, or "pooled" to form s_p^2, an estimate of σ^2. The formula for accomplishing this pooling is:

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

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

The distribution of $\overline{X}_a - \overline{X}_b$ has a mean of $\mu_a - \mu_b$ and an estimated variance of

$$s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)$$

The distribution of the statistic $\dfrac{(\overline{X}_a - \overline{X}_b) - (\mu_a - \mu_b)}{\sqrt{s_p^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}}$ is a *t* distribution with $n_a + n_b - 2$ degrees of freedom.

If the samples of size n_a from populations A, and size n_b from population B are <u>independent</u> the quantities $\overline{X}_a$, $\overline{X}_b$, s_a, and s_b are calculated.

…If n_a and n_b are <u>not both "large"</u> …

…If both populations are <u>normally distributed</u>…

…If the population variances are <u>not equal</u>,

Neither $\dfrac{\left(\overline{X}_a - \overline{X}_b\right) - \left(\mu_a - \mu_b\right)}{\sqrt{\dfrac{s_a^2}{n_1} + \dfrac{s_b^2}{n_2}}}$ nor $\dfrac{\left(\overline{X}_a - \overline{X}_b\right) - \left(\mu_a - \mu_b\right)}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ has a *t* distribution,

but the formula on the left is <u>close</u> if you use $df = \dfrac{\left(\dfrac{{s_1}^2}{n_1} + \dfrac{{s_2}^2}{n_2}\right)^2}{\dfrac{\left({s_1}^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left({s_2}^2/n_2\right)^2}{n_2 - 1}}$.

(Note: Just what the distribution actually is, is a celebrated <u>and unsolved</u> problem in statistics, known as the Behrens-Fisher problem.)

…Even if at least one of the populations are <u>not normally distributed</u>…

We could transform the variables or use a non-parametric technique.

Part 2: Comparing 2 means for fun and profit (Hey, Buddy -- Gotta match?)

Moving from making inferences about one population mean to making inferences about two population means is not just twice the trouble; with the added mean comes added headaches in determining the sampling distributions of the relevant statistics.

The two samples may be dependent or independent.

Orientation to the idea of independent samples

When a simple random sample of size *n* is drawn from a population, each combination of *n* elements from the population is equally probable. Suppose we were to draw two samples of size *n* one sample from population A, one from population B. These two samples would be independent if the probability of a particular combination of *n* elements from one sample is independent of the probability of a particular combination of *n* elements from the other sample.

While we would always wish to draw independent samples when sampling from a single population, it is not always true that we would wish to do so when sampling from two populations. In some cases we might wish to treat our data from two different samples in a -- to coin a phrase -- Noah-like manner: we might grab samples in pairs. Now, while it may be obvious (even in a family-oriented statistics class) why Noah would want animals in pairs, it may not be clear why this would be so in normal non-flood circumstances.

There are many ways in which such sampling might arise in statistical investigations.

1. Self-pairing occurs when two treatments are applied to the same individual. For example, we might want to compare two kinds of shaving devices by having men shave one side with Brand A and the other side with Brand B.

2. Natural-pairing may occur in some experimentation where twins or litter-mates are used for comparison.

3. Artificial-pairing may be used to "match" subjects on some characteristic related to the response variable.

4. Experimentally-imposed pairing may occur due to some aspect of the experiment. In that case the pairing is a special case of blocking.

Two different views of the data.
(Note: we will temporarily assume normality, etc.)

When we are dealing with independent sampling from two populations, we view an individual data element like this:

Observation from population = mean of population + random error.

We will need to descend into the depths of algebra to be clearer. We would more specifically say that the *i*th observation from population A is equal to the mean of population A plus the random error in the *i*th observation. That is,

$$x_{Ai} = \mu_A + \varepsilon_i$$

Suppose we are doing an experiment with two treatments. Then our original population is split, according to our random assignment into two "theoretical" populations, those who could have gotten treatment A and those who could have gotten treatment B. We would symbolize observations from the two treatments like this:

$$x_{Ai} = \mu_A + \varepsilon_i$$
$$x_{Bj} = \mu_B + \varepsilon_j$$

From our algebra of random variables we can create the random variable, "difference of sample means" by subtracting…

$$\overline{X}_A - \overline{X}_B = (\mu_A - \mu_B) + (\overline{\varepsilon}_A - \overline{\varepsilon}_B)$$

When we have "matched" pairs, on the other hand, we view our data as coming from two populations but appearing in pairs, like this:

$$x_{Ai} = \mu_A + \pi_i + \varepsilon_i$$
$$x_{Bi} = \mu_B + \pi_i + \varepsilon_i$$

Notice that the subscripts are now both *i*'s, indicating that these are two observations but they form the *i*th pair. We can again, using the algebra of random variables, create the random variable, "difference of sample means" by subtracting…

$$\overline{X}_A - \overline{X}_B = (\mu_A - \mu_B) + (\pi_i - \pi_i) + (\overline{\varepsilon}_A - \overline{\varepsilon}_B)$$

and simplifying,

$$\overline{X}_A - \overline{X}_B = (\mu_A - \mu_B) + (\overline{\varepsilon}_A - \overline{\varepsilon}_B).$$

Now, the crucial difference in these different ways of viewing an observation lies in the calculation of the variance. In a sense, there is a "commonality" to each pair, a certain amount of variation that is "accounted for" in the pairing rather than contributing to the variability of the sample statistic. The effect of the commonality in each pair is to make the observations correlated.

Two different formula:

With independent samples, the difference in sample means is a random variable, normally distributed (since the original random variables, A and B, are assumed normal). The mean and variance of $\bar{X}_A - \bar{X}_B$ are:

$$\mu_{\bar{X}_A - \bar{X}_B} = \mu_A - \mu_B$$

$$\sigma^2{}_{\bar{X}_A - \bar{X}_B} = \frac{\sigma^2{}_A}{n} + \frac{\sigma^2{}_B}{n}$$

With "matched" or "paired" data, the difference between the pairs is calculated,

$$d_i = x_{Ai} - x_{Bi}$$

and the differences are analyzed. The distribution of differences is normal. The mean and variance of the random variable are:

$$\mu_D = \mu_A - \mu_B$$

$$\sigma^2{}_D = \frac{\sigma^2{}_A}{n} + \frac{\sigma^2{}_B}{n} - 2\rho\sigma_A\sigma_B$$

In actual practice we use estimates of these quantities and perform the mathematically equivalent procedure of testing the hypothesis the mean difference is a constant, or constructing a confidence interval for the mean difference.

The differences:

Theoretically, your choice should be clear. If there is a natural reason for "pairing" the data, a known common contribution to the variability of the observations, you should do so and use the statistics for the "matched pairs *t*." If there is no reason for the pairing that can be seen, you should use the independent *t* procedures.

A natural reason for pairing the data would be a known, identifiable, measurable, and potentially powerful variable that could be affecting the response variable -- that is, a confounding variable that is non-trivially correlated with the response variable. The effect of the paired *t* procedures would be to eliminate the effect of the confounding variable.

As a practical matter, your choice may not be all that clear. First of all, with small data sets, you may have non-zero correlations "by chance." One should not choose to analyze one's data using a paired *t* simply because you have found a correlation in the data! You should only analyze the data using a paired *t* if you have a suspiciously large correlation AND you can explain the reason for it! If you are designing an experiment from scratch, and want to decide whether or not to pair, then you need to consider the tradeoffs before you gather the data.

The tradeoffs are:

1. Pairing in the face of a confounding variable frequently reduces variance! (Yay!)

2. Pairing in the face of a confounding variable reduces the number of degrees of freedom! (Boo!)

So your design of experiment decision must be based on guesses -- excuse me, estimates -- of the effects of pairing. If the combination of decrease in variance and lessening of degrees of freedom result in an overall decrease in the standard error of the statistic, then from a statistical standpoint you should use matched pairs. If, however, the added trouble to make the matches outweighs the gain, then the matching may not be justified.

Part 3: Robustness and the two-sample Student's *t*: Guidelines

Ref: Havlicek, Larry L., Peterson, Nancy L. Robustness of the *t* test: A guide for researchers on effect of violations of assumptions. Psychological Reports, 1974, 34, 1095-1114.

The following guidelines are offered to those who would use pooling with the independent *t* test statistic in the face of potential violations of assumptions. These results were obtained through simulation, and the phrase "obtained *t* distribution" refers to the approximate distribution of the test statistic as seen in their simulation.

The assumptions:

1. The observation must be independent.
2. The observations must be drawn from normally distributed populations.
3. The populations must have the same variance. ("Homogeneity of variance.")

The guidelines:

1. Ordinal and percentile transformations of scale values have little effect on the obtained *t* distributions.

2. When sampling from normal distributions, with equal sample sizes, or samples which differ very little in size, large differences in variance do not influence the obtained *t* distributions.

3. When sampling from normal distributions with unequal sample sizes differences in variances have little distorting effect on the obtained *t* distributions.

4. When sampling from two different shapes of distributions, normal and skewed, with equal variances and with either equal or unequal sample sizes, little distortion occurs in the obtained *t* distributions.

5. When sampling from two different shapes of distributions, normal and skewed, with unequal variances and with either equal or unequal sample sizes, serious distortion occurs in the obtained *t* distributions.

6. When sampling from two non-normal distributions of the same shape, i.e. both skewed in the same direction and with equal variances, there is little distortion in the obtained *t* distributions for both equal and unequal sample sizes.

7. When sampling from two non-normal distributions of the same shape, i.e. both skewed in the same direction and with unequal variances, there is considerable distortion in the obtained *t* distributions for both equal and unequal sample sizes.

8. When sampling from two non-normal distributions of the same shape but skewed in opposite directions, there are serious distortions in the obtained *t* distributions with equal or unequal variances and with equal or unequal sample sizes.

9. When sampling from two non-normal distributions with different shapes and with equal variances, there is little distortion in the obtained *t* distributions with equal and unequal variances and with equal and unequal sample sizes.

10. When sampling from two non-normal distributions with different shapes and with unequal variances, there is considerable distortion in the obtained *t* distributions for samples of unequal sample sizes.