***For the 2-sample location problem, t-test for small n,***

***normal scores or Wilcoxon rank sum test for large n***

***Executive Summary Answer***

This is probably the opposite of what you were taught, but that thinking is wrong. Asymptotic theory has proven that the normal scores test is as powerful as the t-test if the raw data are normal (Gaussian) and as high or higher for every other distribution with finite variance. Hence, it should be routinely favored for large n. As a practical matter, I believe, with some theory and simulation results, that this switch should occur when n is 6 or more in both groups. For n=3 or less in both groups, the only choice that might reject the null hypothesis when α=.05 or less is the t-test or unequal variance t-test. For other small n sample sizes, discreteness lowers the power of the Wilcoxon rank sum test, and, to a much lesser extent, the normal scores test. When using the t-test for small n, the distribution should ideally be close to normal; an Anderson-Darling sum score can assist with a power transformation prior to the t-test. If it isn’t, the test might still be valid, but not that powerful. Also, the Welch t-test is generally preferable to the standard t-test.

My recommendation: For when either n1 or n2 is < 6, use the unequal variance t-test (Welch test), When n1 and n2 are both 6 or more, use the normal scores test (the Wilcoxon rank sum test is acceptable here, but not as good).

***Detailed Explanation***

**Use the t-test for small n**

If the two samples are both of size 3, you CANNOT get statistical significance with the 2-sided Wilcoxon rank sum test. Why not? If the three highest values are from group two, the probability is 1/20 = .05. But for a 2-sided test, the probability of group 2 having the three lowest values is also .05, so the p-value is 0.10. Thus, your only chance is the t-test. Don’t you need to believe that the raw data are (close to) normal? Yes, but that’s your only chance, and you can perform a power transformation to improve that chance (but that will be addressed separately).

**Use the normal scores test for large n**

The large n results between the normal scores test and the t-test are VERY clear: the normal scores test is better, as it has the same asymptotic relative efficiency (ARE) as the t-test when the raw data are normal (which the t-test was designed for) and has equal or greater power for every other distribution of finite variance (Chernoff, Savage, 1958). Wilcoxon rank sum test is nearly as good. Compared to the t-test, the minimum ARE for Wilcoxon is .864 and the maximum is infinity (it is .95 for normal data; see Hodges, Lehmann, 1956). The worst-case scenario implies that, asymptotically, it might take a 16% higher (=1/.864) sample size for the Wilcoxon to have the same power as the t-test, but it might take an infinitely larger sample for the t-test to have the same power as the Wilcoxon for some other distribution. Neither the Wilcoxon nor the normal scores test are always more asymptotically powerful than the other.

The only practical question, then, is where does “asymptotics” begin. Given the discreteness table below, I think when both n1 and n2 are 6 or more, we should shift to the normal scores test.

The table below gives the possible p-values for the Wilcoxon rank sum and normal scores test for various sample sizes. Note that when n1 = n2 = 3, there are only 5 possible p-values for a 2-sided Wilcoxon rank sum test, and 7 for the normal scores test. When n1 = n2 = 6, this increases to 37 possible p-values for a 2-sided Wilcoxon rank sum test, and 370 for the normal scores test. Thus, especially for the normal scores test, discreteness issues, which can prevent us from using the full desired α level, is no longer a serious concern.

**Discreteness of the nonparametric tests**

**Wilcoxon rank sum Normal scores Combinatoric possibilities**

**1-sided test 2-sided test 1-sided 2-sided 1-sided 2-sided**

**N/group # of Rsums Rsums/2**

3 10 5 14 7 20 10

3:4 13 7 27 14 35 18

4 17 8 42 21 70 35

4:5 20 10 81 41 126 73

5 26 13 122 61 252 126

6 37 18 370 185 924 462

7 50 25 3432 1716

8 65 32

9 82 41

10 101 50

Rsums = Rank sums

Combinatoric possibilities: (Obtained from N = n1 + n2, choose n1).

**Why is the recommendation qualitatively reversed from what the literature advises?**

The literature is just wrong, primarily because those who recommended it worried about the wrong things. One common recommendation is to shift from the Wilcoxon to the t-test when the sample sizes are 30 or more. This is because the t-test assumes a normal distribution, and it is enough that the sample means be normal, and 30 was thought to be a sufficient sample size to ensure that (due to the central limit theorem) the mean eventually converges to normality. In principle, a larger sample size, ensures that the t-test is a valid t-test, but it says nothing about the test’s power. Large outliers can kill the power of the t-test compared to the Wilcoxon or normal scores test, rendering it possibly infinitely worse. The ARE results presented above refer to power comparisons.

For small n with at most 3 per group, mathematically t-tests can reject the null while Wilcoxon and normal scores tests can’t. Here’s an analogical explanation. The normality assumption from which the t-test is constructed is like a Bayesian prior distribution. Prior distributions are most beneficial when the sample size is small; for a very large sample size the data swamps out the prior distribution and thus dominates the posterior distribution conclusion. In other words, the normality assumption is only helpful (and necessary) when the sample size is small. What if the data aren’t normal? Practically, I believe that this still yields a valid test (the calculated p-value is at or below the true probability for a result as extreme as that seen), it will simply be less powerful, and thus you might fail to reject.

**What about equality of variances?**

The t-test also assumes that the variances are the same and the test statistic pools the variance. This assumption should be worried about, especially if the sample sizes of the two groups are not close to being equal in size. The solution is to use the unequal variance t-test (also called the Welch test or Satterthwaite test). The Welch test does still work when the variances are equal, it just has somewhat lower power. The Satterthwaite formulation essentially reduces the effective degrees of freedom from which the statistical significance is ascertained. When the observed sample variances are identical, the degrees of freedom are the same as for the classic t-test. Thus, the df is only reduced due to unequal sample variances, and the penalization is arguably appropriate.

Didn’t Boneau (1960) show that the t-test was robust to unequal variances and non-normality even for very small sample sizes? Partly yes, but his simulation study on distributions was flawed, and his work overall was only based on validity, not on power and not compared with other options like the Wilcoxon rank sum test. He looked at three distributions: normal, uniform, and exponential. T-test is designed for the normal distribution, the uniform distribution quickly converges to normality. The only real issue was behavior under the exponential distribution, which he didn’t program correctly. Moreover, the simulation study was done in 1960. While it was an advance in 1960, many much-improved studies were conducted later, yet this early work is the only one in the highly-cited papers database, and is still the primary reference on t-test assumptions in many of the introductory statistics books. It also has an extremely high diffusion number, being the 90th most diffuse paper among the 4492 in the highly-cited papers database. This idea has provided much of the license for overuse of the t-test.

His work on equality of variances was better, but still quite minimal in sample sizes considered. His abstract does note that under unequal variances, “the probability values will be quite different from the normal values”.

**What about preliminary tests?**

Preliminary tests have sometimes been recommended for testing for equality of variances, or normality of the data. I am not a big fan of them. One problem is that we would use exactly the opposite way that we typically use hypothesis tests. The usual goal is to reject the null. Here we are rooting for the null. Given that, some people recommend using α = .10 rather than α = .05. That is a better choice, but then why not, say, α = .20? Also, if the test power is low, we will favor what we want to favor – the null hypothesis. Regarding equality of variances, the Welch test penalty for unequal variances is reasonable, and correct under the assumptions of the test. For small n situation, I do recommend use of the Anderson-Darling (AD) test, but not as a test. I would calculate the AD test statistic for each sample separately for both groups, then add them together. I would then square-root the data and get a second AD sum. Then, I would log the data and obtain a third AD sum. (I also will sometimes defend trying a quarter-root transformation, and obtain a fourth AD sum). Then, I choose the power transformation (part of the Box-Cox transformation family) with the lowest AD sum. This exercise helps to find the distribution among the options that is closest to normality. Then, using that data transformation, I would conduct the Welch t-test.

**Conclusion**

Among parametric tests, I prefer the Welch unequal variance t-test to the t-test, as it eliminates the assumption of equal variances. Among nonparametric tests, I favor the normal scores test over the Wilcoxon rank sum test as its worst-case scenario compared to the t-test is not as low. The recommended switching place for where “asymptotics” begins is when both sample sizes are at least 6. An Anderson-Darling sum calculation can be useful for identifying a good power transformation.

**References**

Anderson, Darling, 1954. (#650) A test of goodness-of-fit.

Boneau, 1960. (#2816) The effects of violations of assumptions underlying the t test.

Box, Cox, 1964. (#95) An analysis of transformations.

Chernoff, Savage, 1958. Asymptotic normality and efficiency of certain nonparametric test statistics.

Hodges, Lehmann, 1956. The efficiency of some nonparametric competitors of the t-test.

Numbers denote their rank in the highly-cited papers database.