Appendix: *t*-test with Welch correction*

This appendix presents additional simulation results in which the *t*-test with Welch (1936, 1938) correction was compared to parametric and permutational *t*-tests for two types of data distributions and for equal and unequal population variances. The result of a *t*-test is identical to that of an anova computed for two groups; the *t*-statistic is the square root of the *F*-statistic used in anova. The Welch correction, described in most textbooks of statistics (e.g., Scherrer, 1984, and Zar, 1999), for example, is a widely used solution to the Behrens-Fisher problem of testing for the difference in the means of two populations when the variances are unequal. These simulations, which led us to formulate recommendations with respect to the use of this correction, should prove useful to application domains where unequal variances are commonly encountered.

The Welch correction was designed to provide a valid *t*-test in the presence of unequal population variances. It consists of using a corrected number of degrees of freedom ν to assess the significance of the *t*-statistic computed as usual. ν is the next smaller integer of the value obtained from the following equation:

$$\nu = \frac{\left[(s_1^2/n_1) + (s_2^2/n_2) \right]^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}} \qquad (9)$$

where s_1^2 and s_2^2 are the sample variances of groups 1 and 2 respectively, whereas n_1 and n_2 are the number of observations in groups 1 and 2. When the variances are equal, equation 9 reduces to the usual formula $\nu = (n_1 + n_2 - 2)$ when the two groups have equal numbers of observations, but to a lower value when $n_1 \neq n_2$, making the test with Welch correction too conservative. We will see if the simulations can illustrate this bias, and what are its practical consequences, if any, for the users of the test.

1. Equal sample sizes

Simulations were carried out using two groups of data of equal sizes, with $n_1 = n_2 = \{10, 25, 50, 100\}$.

The first series of simulations used normal random deviates with mean 0 and standard deviations between 1 and 9, as specified in the graphs; the values were chosen in such a way that the standard deviations of the two reference populations added to 10. For the power study, the population mean of group 1 was 0 while that of group 2 was 5. 10000 simulations were run in each case, during which the following statistics were computed: a standard

---

* Excerpt from:

Legendre, P. and D. Borcard. Statistical comparison of univariate tests of homogeneity of variances. (Manuscript).

parametric *t*-test, a permutational *t*-test (using 499 random permutations), and a *t*-test with Welch correction. The *t*-test with Welch correction is expected to produce correct type I error when the variances are not homogeneous, whereas the permutational *t*-test is expected to do the same for skewed data.

• Figures 6a and 6e show that the *t*-test with Welch correction has correct type I error for any and all combinations of population variances, and for all sample sizes. The parametric and permutational *t*-tests are affected by inequality of the variances. The effect is strong when sample size is small (n_j = 10 in Fig. 6a), but disappears gradually as sample size increases. For example, with n_j = 50, the tests have slightly inflated type I error only in the most extreme case of inequality of the population variances (Fig. 6e). No inflation of type I error was found at n_j = 100 (results not illustrated). The power of the three tests is comparable when they are valid, i.e., when type I error is not larger than α (Figs. 6b and 6f).

In the second series of simulations, power base 1.4 data were used, as described in Section 2.2 of the main paper. Otherwise, the design of the simulations was the same as for normal data.

• When the variances are equal or nearly so ($\sigma_1 = \sigma_2$ in Figs. 6c and 6g), all three tests are valid for all sample sizes. The permutational *t*-test presents the advantage of having correct type I error whereas the other two forms of the test are too conservative; the permutational *t*-test also has the highest power (Figs. 6d and 6h).

• When the variances are unequal ($\sigma_1 \neq \sigma_2$, e.g., in Fig. 6g), type I error of all three tests becomes inflated to various degrees, so that the tests are invalid and should not be used. This was the case with all sample sizes used in the simulations, except with n_j = 10 (Fig. 6c); power of the tests is irrelevant when type I error is larger than α. The parametric and Welch-corrected *t*-test are valid when sample sizes are very small since type I error is not larger than α (Fig. 6c where n_j = 10), but power is so low that the tests are unusable (Fig. 6d).

2. Unequal sample sizes

Simulations were also conducted with different sample sizes chosen in such a way that $n_1 + n_2 = \{20, 50 \text{ or } 100\}$. Otherwise, the design of the simulations was the same as for equal sample sizes; the standard deviations were made to vary between 1 and 9, as in Fig. 6, the values being chosen in such a way that the standard deviations of the two reference populations added to 10. Simulations were carried out to measure type I error (with the population means equal) and power (with the population means unequal).

The first series of simulations used normal random deviates with mean 0 and standard deviations between 1 and 9 summing to 10 for the two groups, as specified in the graphs.

• When the population variances are equal, the parametric and permutational *t*-tests have correct type I error for any combination of sample sizes (Fig. 7a), whereas the test with

Welch correction becomes too conservative when sample sizes are strongly unequal. Power of all tests decreases as the sample sizes become more unequal (Fig. 7b); the *t*-test with Welch correction has lower power than the other forms in the most extreme cases of inequality of the sample sizes. Of course, power of all tests increases as $n_1 + n_2$ grows from 20 to 50 to 100 (not illustrated).

• When the population variances are unequal, all tests have correct type I error for equal group sizes (e.g., Fig. 7c, except for a slight inflation of type I error in the most extreme case of inequality of the population variances, already shown in Fig. 6e), but type I error becomes increasingly too conservative as the group sizes become more unequal. All tests remain valid with sample sizes $n_1 + n_2 = 50$ or 100, but for strongly unequal group sizes, power becomes too low for the tests to be useful (Fig. 7d). We already know from Fig. 6a that for very small sample sizes, such as $n_1 + n_2 = 20$, there is a small inflation of type I error of the parametric and permutational *t*-tests in the case of equal group sizes; this is quickly compensated by the conservativeness of these tests in the case of unequal group sizes.

The second series of simulations, based upon power base 1.4 random deviates (skewed distribution), gave the following results.

• When the population variances are equal, only the permutational *t*-test has correct type I error for any combination of sample sizes and for all $n_1 + n_2 = \{20, 50 \text{ or } 100\}$ subjected to simulations (Fig. 7e); it also has good power (Fig. 7f). The test with Welch correction is too conservative for all combinations of sample sizes; power is reduced compared to the power of the permutational *t*-test. The parametric *t*-test has conservative type I error in most cases, but when the samples become strongly unequal in size, it has inflated type I error. In the area where the parametric *t*-test is valid (sample sizes equal or moderately unequal), its power is less than that of the permutational *t*-test.

• With unequal variances, the behavior of the three tests becomes erratic (e.g., Fig. 7g). When the tests are valid, they have poor power so that they are useless (Fig. 7h).

3. Recommendations

The recommendations that can be derived from our simulation results are complex. They are presented in tabular form (Table II). To summarize, the test with Welch correction is useful when the data are normal, sample sizes are small, and the variances are heterogeneous. Otherwise, use the parametric *t*-test for normal data, or the permutational *t*-test for skewed data. For heteroscedastic data that cannot be normalized, a nonparametric test should be used.

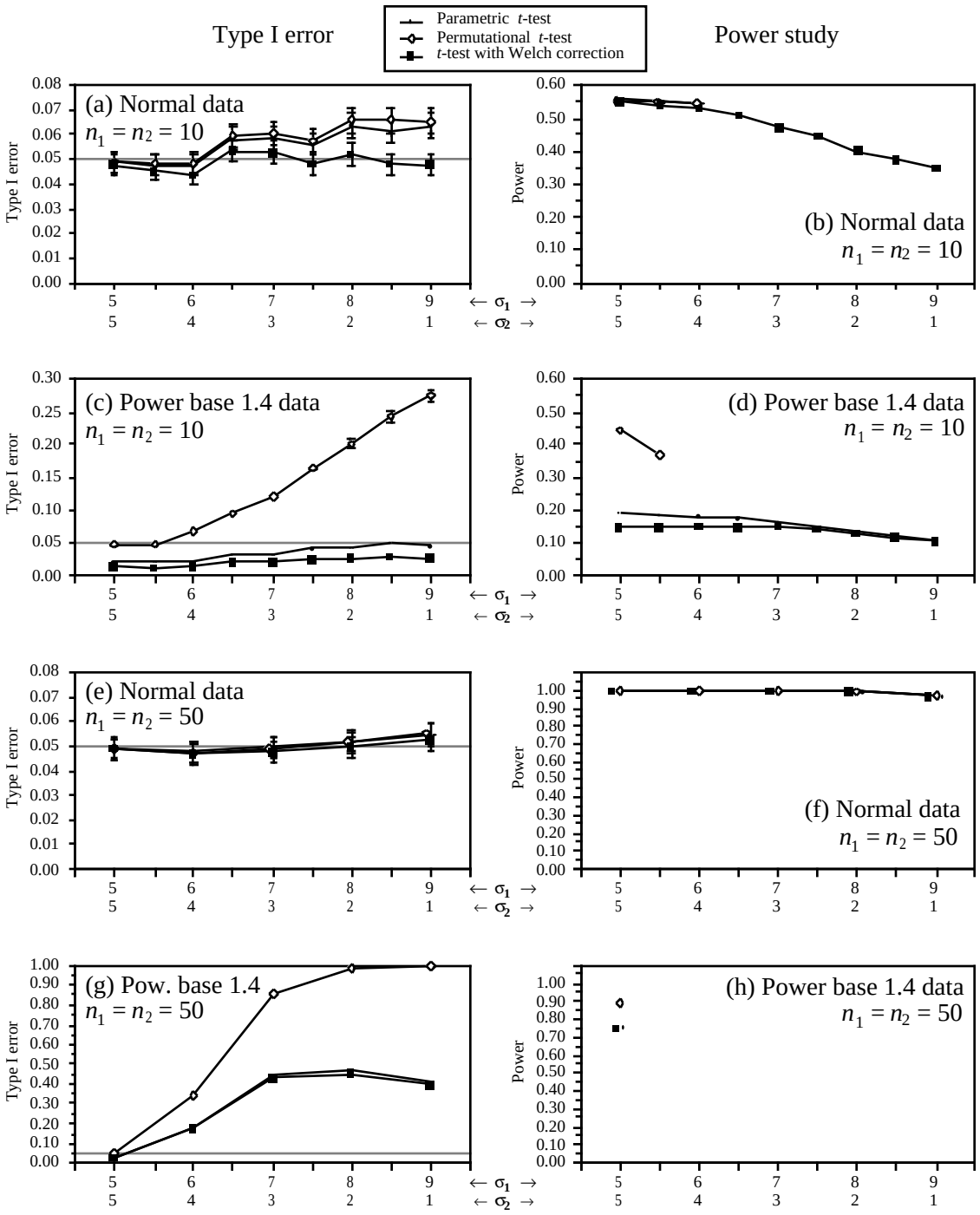

Figure 6 (a) Type I error and 95% confidence intervals (error bars) for *t*-tests of difference between two group means, at $\alpha = 0.05$, for two groups of normal data ($n_1 = n_2 = 10$) with equal means, as a function of the two population standard deviations (σ_k, abscissa). (b) Power simulation results for the same data. (c, d) Same as (a, b) for power base 1.4 data. (e, f, g, h) Same as (a, b, c, d) using $n_1 = n_2 = 50$. When they were closer than the size of the symbols, the 95% confidence error bars were omitted. Each point summarizes 10000 simulations.

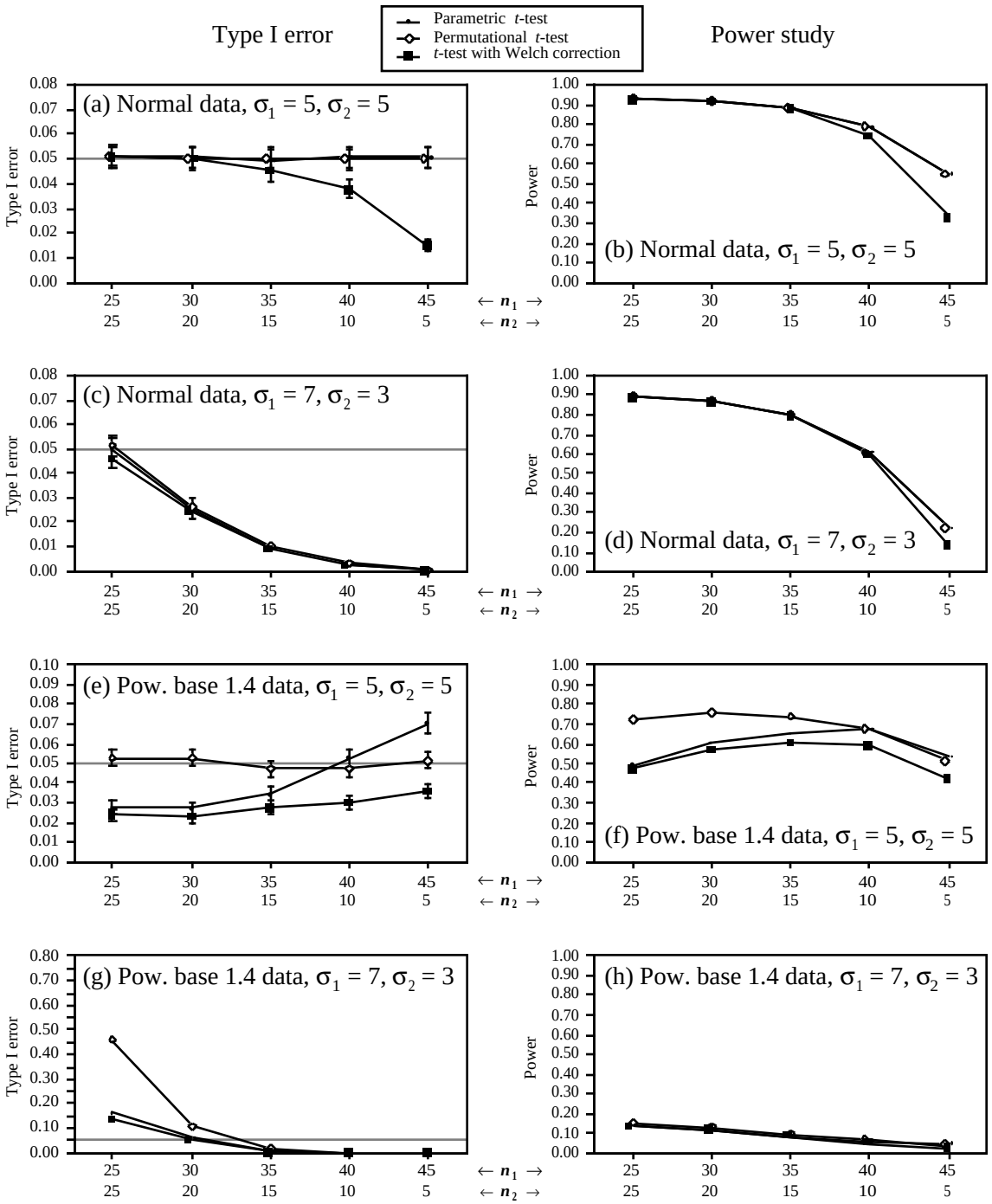

Figure 7 (a) Type I error and 95% confidence intervals (error bars) for *t*-tests of difference between two group means, at $\alpha = 0.05$, for normal data with equal population standard deviations (σ_j), as a function of the sample sizes (n_j, abscissa); $n_1 + n_2 = 50$. (b) Power simulation results for the same data. (c, d) Same as (a, b) for unequal population standard deviations (σ_j). (e, f, g, h) Same as (a, b, c, d) using power base 1.4 data. When they were closer than the size of the symbols, the 95% confidence error bars have been omitted.

Table II Recommendations for *t*-test of equality of two means.

1. Sample sizes equal?

Yes ⇒ go to 2

No ⇒ go to 6

2. Equal sample sizes. Distribution:

Normal ⇒ go to 3

Skewed ⇒ go to 5

3. Normal distributions. THV result:

Variances homogeneous ⇒ use any one of the 3 tests (most simple: parametric *t*-test).

Variances unequal ⇒ go to 4

4. Variances unequal. Sample size:

Small ⇒ use the *t*-test with Welch correction.

Large ⇒ use any one of the 3 tests (most simple: parametric *t*-test).

5. Skewed distributions. THV result:

Variances homogeneous ⇒ all 3 tests are valid, but the permutational *t*-test is preferable because it has correct type I error and the highest power.

Variances unequal ⇒ normalize the data or use a nonparametric test (Wilcoxon-Mann-Whitney test, median test, Kolmogorov-Smirnov two-sample test, etc.).

6. Unequal sample sizes. Distribution:

Normal ⇒ go to 7

Skewed ⇒ go to 8

7. Normal distributions. THV result:

Variances homogeneous ⇒ use the parametric or permutational *t*-tests (most simple: parametric *t*-test).

Variances unequal ⇒ use any one of the 3 tests (most simple: parametric *t*-test). Power is low when the sample sizes are strongly unequal; avoid the Welch-corrected *t*-test in the most extreme cases of sample size inequality (lower power).

8. Skewed distributions. THV result

Variances homogeneous ⇒ use the permutational *t*-test.

Variances unequal ⇒ normalize the data or use a nonparametric test (Wilcoxon-Mann-Whitney test, median test, Kolmogorov-Smirnov two-sample test, etc.).