A Review of the Behrens-Fisher Problem and Some of Its Analogs: Does the Same Size Fit All? *

Authors: Sudhir Paul
– Department of Mathematics and Statistics, University of Windsor, Ontario, Canada (smjp@uwindsor.ca)

You-Gan Wang
– School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD, Australia (you-gan.wang@qut.edu.au)

Insha Ullah
– School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD, Australia (insha.ullah@qut.edu.au)

Abstract:

- The traditional Behrens-Fisher (B-F) problem is to test the equality of the means μ_1 and μ_2 of two normal populations using two independent samples, when the quotient of the population variances is unknown. Welch [42] developed a frequentist approximate solution using a fractional number of degrees of freedom t-distribution. We make a a comprehensive review of the existing procedures, propose new procedures, evaluate these for size and power, and make recommendation for the B-F and its analogous problems for non-normal populations. On the other hand, we investigate and answer a question: does the same size fit all all, i.e. is the t-test with Welch's degree of freedom correction robust enough for the B-F problem analogs, and what sample size is appropriate to use a normal approximation to the Welch statistic.

Key-Words:

- *The Behrens-Fisher Problem; the Beta-binomial model; the Negative binomial model; the Weibull model.*

AMS Subject Classification:

- 49A05, 78B26.

*The opinions expressed in this text are those of the authors and do not necessarily reflect the views of any organization.

1. INTRODUCTION

The traditional Behrens-Fisher (B-F) [5, 20] problem is to test the equality of the means μ_1 and μ_2 of two independent normal populations where the variances σ_1^2 and σ_2^2 are unknown and unspecified. The problem arises when the ratio of the population variances is unknown as well. In the case of known Importance of this problem is well understood and its application is widespread [1, 12, 14, 15, 16].

Ever since the solution of this problem by [42], many papers have been written. See, for example, [7], [19], and [29]. These and similar other papers [9, 38] have attempted improvement, in terms of level and power, over the Welch procedure. More recently, non-parametric [14, 16, 21] and Bayesian [24, 45] procedures have also been developed.

However, independent samples from two two-parameter populations (other than the normal) arise in many situations. The problem then is to test the equality of two location (or some analogous) parameters when the dispersion (or some analogous) parameters are unknown and possibly different. These problems are analogous to the traditional Behrens-Fisher problem. Prior to 2014 not much have been written on the solution of the Behrens-Fisher analogous problems. Some (to our knowledge) problems analogous to the B-F problem that have been dealt with recently are (i) testing equality of two negative binomial means in presence of unequal dispersion parameters [30]; (ii) testing equality of scale parameters of two Weibull distributions in the presence of unequal shape parameters [2], and (iii) testing equality of two beta binomial proportions in the presence of unequal dispersion parameters [3].

When the sample sizes are small the two sample t-test (T_1) with Welch's [42] degree of freedom and for large sample sizes ($N = n_1 + n_2 > 30$) the standard normal statistic (T_N) (see, Section 2) are recommended by standard text books [23]. Many evidences have been shown in favour of the preference of the Welch T_1 over other procedures. See, for example, [7, 12, 29] for the standard BF problem. More recently [38] developed a jackknife based procedure and [9] developed a computationally intensive procedure for the BF problem. However, no systematic study has been conducted so far to determine the overall sample size required under which the normal approximation of the statistic T_N works.

The primary purpose of this paper is to make a comprehensive review of the existing procedures, evaluate these for size and power, and make recommendations for the standard BF and its analogous problems in some sense. For the standard BF and some of its analogous problems we also investigate performance of a new Monte-Carlo approach, the bootstrap and the rank counterparts. A recent study [30] suggests that the Welch T_1 does well in some non-normal situations, such as for samples from two negative binomial populations. Along with some other procedures performances of the Welch T_1 and the new Monte-Carlo

approach are investigated for samples from normal, two discrete models (count data and data in the form of proportions) and a survival model for a wide range of parameter spaces to reflect comparison of the means for variances which are same to very different.

The secondary purpose is to investigate and answer a question: does the same size fit all or in other words is the t-test with Welch's [42] degree of freedom correction robust enough for the BF problem analogs and what sample sizes are appropriate for the normal approximation of the statistic T_N.

Review, possible new procedures, simulations, and recommendations for the standard BF problem are given in Section 2. The BF analogues corresponding to the negative binomial, the beta binomial, and the Weibull are dealt with in Sections 3, 4 and 5 respectively. The concluding section (Section 6) provides some guide lines as to which procedure(s) to be used in each case. Some recommendations for possible future study are also provided in this section.

2. The Behrens-Fisher Problem: Two Normal Populations

2.1. Welch's t-Statistic

The well-known Behrens-Fisher (B-H) problem is to test the equality of the means μ_1 and μ_2 of two independent normal populations where the variances σ_1^2 and σ_2^2 are unknown and possibly unequal.

Let $Y_{i1}, \ldots, Y_{in_i}$ be a random sample from a population, $i = 1, 2$. Now, let $y_{i1}, \ldots, y_{in_i}$ be a corresponding sample realization with mean $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and variance $s_i^2 = \sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/(n_i - 1)$. If the samples come from normal populations with means μ_1 and μ_2 and unknown and possibly unequal variances σ_1^2 and σ_2^2, then

$$T_N = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

is asymptotically normally distributed with mean 0 and variance 1 when both n_1 and n_2 are sufficiently large. This is stated in many undergraduate text books in Mathematical Statistics [23].

However, when the sample sizes n_1 and n_2 are smaller the distribution of T_N, henceforth denoted by T_1, is approximately distributed as Student's t with degrees of freedom

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}\right)}$$

[42]. It is shown by [19] and [42] using simulations that the statistic

$$Z = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{(n_1-1)s_1^2}{(n_1^2-3n_1)} + \frac{(n_2-1)s_2^2}{(n_2^2-3n_2)}}}$$

might be preferable to the statistic T_1 because the former would maintain nominal level better than the later. However, [19] does not provide a degree of freedom for the above Z to be used as an approximation to the t-distribution. To this end [7] derive degrees of freedom and compare performance of T_1 with a few other statistics, such as the Wald, likelihood ratio and score statistics and the statistic Z, in terms of level and power and find that T_1 is still the best. However, there is an error in the degrees of freedom formula which later was corrected by [29]. After carrying out further simulations [29] finds that in addition to all the reasons given by [7] to prefer T_1 over Z, the former shows better power performance than the latter. See, [29] for further details.

To the best of our knowledge, to-date, the statistic T_1 is the best and is referred as the statistic to use in recent text books [23]. In this paper we attempt to do a comprehensive review of all available methods and develop a new Monte Carlo procedure.

2.2. The Likelihood, Score and Wald Tests [7]

The likelihood ratio statistic (LR), score statistic and Wald statistic, denoted by L, S and W, derived by Best and Rayner (1987) are

$$L = n_1 \log[(n_1-1)s_{10}^2/((n_1-1)s_1^2)] + n_2 \log[(n_2-1)s_{20}^2/((n_2-1)s_2^2)],$$

$$S = (\bar{y}_1 - \bar{y}_2)^2/((n_1-1)s_{10}^2/n_1^2 + (n_2-1)s_{20}^2/n_2^2),$$

and

$$W = (\bar{y}_1 - \bar{y}_2)^2/((n_1-1)s_1^2/n_1^2 + (n_2-1)s_2^2/n_2^2),$$

where $s_{i0}^2 = \sum_{j=1}^{n_i}(y_{ij}-\mu_0)^2/(n_i-1)$ and μ_0 is the solution to the cubic equation

$$\begin{aligned}
&-(n_1+n_2)\mu_0^3 + [(n_1+2n_2)\bar{y_1} + (n_2+2n_1)\bar{y_2}]\mu_0^2 \\
&-[n_1(n_2-1)s_2^2/n_2 + n_2(n_1-1)s_1^2/n_1 + 2(n_1+n_2)\bar{y_1}\bar{y_2} + n_2\bar{y_1}^2 + n_1\bar{y_2}^2]\mu_0 \\
&+[n_1\bar{y_1}\{(n_2-1)s_2^2/n_2 + \bar{y_2}^2\} + n_2\bar{y_2}\{(n_1-1)s_1^2/n_1 + \bar{y_1}^2\}] = 0
\end{aligned}$$

[31] give a brief description on the construction mechanism as well as the advantages of the $C(\alpha)$ or score tests over the LR and the Wald tests (see, [29] for details).

2.3. A Monte Carlo Procedure developed Using T_1

By examining the T_1-statistic, it is clear that the denominator is a convex combination of $\chi^2_{(n_1-1)}/(n_1-1)$ and $\chi^2_{(n_2-1)}/(n_2-1)$, and the combination proportion depends on the ratio of the two underlying population variances and the sample sizes. The t-distribution approximation becomes exact when $\tau = \sigma_2^2 n_1/\sigma_1^2 n_2 = 1$, and we expect the Monte Carlo method works better when τ is very different from 1. Theoretically, the p-value cannot be calculated under the null unless τ is specified. Under the null, the T_1 statistic follows an exact t distribution with degree of freedom being n_1-1, n_2-1 and (n_1+n_2-2) when τ takes 0, ∞ and 1. The new statistic, henceforth denoted by T, is

$$T = \frac{\frac{\bar{y}_1-\bar{y}_2}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}}{\sqrt{\frac{s_1^2/n_1+s_2^2/n_2}{\sigma_1^2/n_1+\sigma_2^2/n_2}}} = \frac{\mathcal{N}}{\sqrt{\mathcal{K}}}.$$

Here $\mathcal{N} \sim N(0,1)$. We now study the distribution of $\mathcal{K}$.

$$\begin{aligned}
\mathcal{K} &= \frac{s_1^2/n_1 + s_2^2/n_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2} \\
&\sim \frac{\frac{\chi^2_{n_1-1}}{n_1-1}\frac{\sigma_1^2}{n_1} + \frac{\chi^2_{(n_2-1)}}{n_2-1}\frac{\sigma_2^2}{n_2}}{\sigma_1^2/n_1 + \sigma_2^2/n_2} \\
&\sim \lambda\kappa_1 + (1-\lambda)\kappa_2,
\end{aligned}$$

where λ is a proportion parameter,$(\sigma_1^2/n_1)/(\sigma_1^2/n_1+\sigma_2^2/n_2)$, $\kappa_1 \sim \chi^2_{n_1-1}/(n_1-1)$, and $\kappa_2 \sim \chi^2_{n_2-1}/(n_2-1)$.

In order to simulate the Monte Carlo numbers from $\mathcal{K}$, we will need to provide a value for λ. Clearly, we can estimate λ by

$$\hat{\lambda} = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

We therefore obtained an approximate distribution for $\mathcal{K}$,

$$\tilde{\mathcal{K}} \sim \hat{\lambda}\kappa_1 + (1-\hat{\lambda})\kappa_2,$$

whose distribution can be easily obtained. The final distribution, using Monte Carlo procedure, can be approximated by $Z/\sqrt{\tilde{\mathcal{K}}}$ which is obtained by a random number from $N(0,1)$ and two independent random numbers from $\chi^2_{(n_1-1)}$ and $\chi^2_{(n_2-1)}$. Because κ_1 and κ_2 are independently simulated from $\hat{\lambda}$, we have $\mathrm{E}(\tilde{\mathcal{K}}) = 1$ and $\mathrm{var}(\tilde{\mathcal{K}}) = 2\hat{\lambda}^2/(n_1-1) + 2(1-\hat{\lambda})^2/(n_2-1)$.

If the variance ratio σ_2^2/σ_1^2 is known, the distribution of $\mathcal{K}$ above is known as a mixture of two χ^2 distributions and T (§2.3) becomes pivotal but it is generally

not an exact t distribution. However, if the variance ratio is given, one can use the pooled variance estimator and form a $t-$statistic with $n_1 + n_2 - 2$ degrees of freedom.

If $t-$distribution is used to approximate T, i.e., $\tilde{\mathcal{K}}$ is approximated by a chi-square distribution, the "best" degree of freedom by matching the variance $(\tilde{\mathcal{K}})$ to $\chi^2_{(d)}/d$ is

$$d = 2/\text{var}(\tilde{\mathcal{K}}) = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)\hat{\lambda}^2 + (n_1 - 1)(1 - \hat{\lambda})^2},$$

which is exactly the same as Welch's formula!

After developing this procedure we found that [18] also developed the same statistic. Similar idea has also been explored by [4] and [42]. However, they used an exact distribution which is complex to use and showed that the Welch approximation is remarkably accurate, even for small n_1 and n_2, provided that n_1 and n_2 are equal or nearly equal. Singh, Saxena, and Srivastava [38] developed a procedure similar to the one given above and [9] developed another Monte Carlo based procedure "Computational Approach Test" (CAT). Using a simulation study [9] find that the procedure developed by [38] is not as good as it has been claimed [9]. On the other hand the CAT procedure is quite computationally involved. For small sample sizes the CAT is quite conservative. In contrast our method, which is also Monte Carlo, is very easy to use and its performance is much better that that of CAT. This issue will be dealt with in a separate paper.

2.4. A Bootstrap Procedure [13]

A bootstrap test for the Behrens-Fisher problem is developed by [13]. Among the re-sampling methods, the two sample bootstrap test is the one that neither assumes equal variances nor does it require any distributional assumptions and offer a possible solution to the Behrens-Fisher problem [13]. All we need is a suitable test statistic and a null distribution under the hypothesis of equal population means. Manly (1997) recommends to use T_N as a test statistic, where,

$$T_N = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

is asymptotically normally distributed with mean 0 and variance 1 when both n_1 and n_2 are sufficiently large. The null distribution is approximated by the distribution of B values of T_N evaluated at each of the B bootstrap samples. The detailed algorithm proceeds as follows:

1. Calculate T_N using the observed two sample data.

2. Obtain a bootstrap samples of size n_i; say y^*_{ij}, from the adjusted y_{ij}, that is, from $y^{adj}_{ij} = y_{ij} - \bar{y}_i + \bar{y}$, where $\bar{y}$ is the overall mean.

3. Calculate
$$T^*_N = \frac{\bar{y}^*_1 - \bar{y}^*_2}{\sqrt{s^{2*}_1/n_1 + s^{2*}_2/n_2}}.$$

4. Repeat step 2 and 3 B times ($B = 999$); thereby obtaining 999 bootstrap values of T_N*.

5. For a two sided test, a difference between the means is significant if the observed value of $|T_N| > 100(1-\alpha/2)th$ values of T^*_N.

2.5. A Non-Parametric Procedure [21]

To address the Behrens-Fisher problem, the Mann-Whitney-Wilcoxon test [27, 43] is modified in [21]. Define P_{1i}, the number of y_2 observations less than y_{1i}, for $i = 1, \ldots, n_1$. Similarly, define P_{2j}, the number of y_1 observations less than y_{2j}, for $j = 1, \ldots, n_2$. The P_{1i} and P_{2j} are called the placements of y_1 and y_2, respectively [21]. Let $\bar{P}_1$ denotes the mean of y_1 placements and $\bar{P}_2$ the mean of y_2 placements. Also compute the quantities $V_1 = \sum_{i=1}^{n_1}(P_{1i} - \bar{P}_1)^2$ and $V_2 = \sum_{j=1}^{n_2}(P_{2j} - \bar{P}_2)^2$, then the Fligner-Policello statistic (modified Mann-Whitney-Wilcoxon statistic) is given by

$$\hat{U} = \frac{\sum_{i=1}^{n_1} P_{2i} - \sum_{j=1}^{n_2} P_{1j}}{2(V_1 + V_2 + \bar{P}_1\bar{P}_2)^{1/2}}.$$

For a two-sided test the null hypothesis of equal medians is rejected if $|\hat{U}| \geq u_{\alpha/2}$. The critical value $u_{\alpha/2}$ can be calculated exactly or estimated using Monte Carlo simulation for large n_1 and n_2. The procedure is also available in contributed R package *NSM3*.

2.6. Simulations

We have conducted a simulation study to compare the performance, in terms of level and power, of 10 statistics, namely, the statistic T_N, the Welch Statistic T_1, the new procedure T, the likelihood ratio statistic L, the Wald Test W, the score statistic S, the Fenstad statistic Z, the bootstap procedure BT, the Wilcoxon two sample non parametric procedure WC and the recent non-parametric procedure FP by [21]. To perform WC we used R function *wilcox.test()*.

To compare the statistics in terms of size, we considered $\mu_1 = \mu_2 = 1$, a range of values of $VR = \sigma^2_1/\sigma^2_2 = 1/25, 2/24, 3/23, \ldots, 24/2, 25/1$, and a nominal

level $\alpha = .05$. Note that this choice of variance ratios ensures comparison of the means for variances which are same to very different.

For sample sizes we considered equal and unequal n_1 and n_2. So, for example, n_1 was fixed at 5, 10, 15, 20, 25, 30. Then, for each fixed n_1, empirical levels were obtained for $n_2 = 5, 10, 15, 20, 25, 30$. These results are all given as graphs in Figures 1-6 in Appendix A1 in supplementary material. The graphs are in terms of size against $\rho = \log(\sigma_1^2/\sigma_2^2)$. All simulation results are based on 10,000 samples.

We now discuss the size results of the 10 statistics:

i. The statistics T_N and T_1: The statistic T_N is liberal, highly liberal for smaller n_1 and n_2. Even for $n_1 = n_2 = 30$, for which basic text books recommend its use, it is liberal, empirical level ranging, on average, from 0.0504 (when $VR \approx 1$) to 0.0618 (as VR is further and further away from 1). We then wanted to see what happens for larger n_1 and n_2. For this we extended the simulation study for (n_1, n_2)= (35,35), (40,40), (50,50), (60,60), (70,70), (80,80). Results are presented as graphs in Figure 7 in Appendix A1 in supplementary material. For $n_1 = n_2 = 35$, it holds level when $-1 < \rho < 1$. Otherwise, empirical level improves as the sample size increases. However, even at $n_1 = n_2 = 80$, this statistic is somewhat liberal, specially near $\rho = \pm 3$.

   For a close comparison between T_N and T_1 empirical level results for $n_1 = n_2 = 35, 40, 50, 60, 70, 80$ are given as graphs in Figure 1. It shows that even at $n_1 = n_2 = 80$ empirical levels of T_N are slightly larger than those of T_1; T_N is still slightly liberal.

ii. The statistics T_1 and T: For all situations studied, even for $n_1 = n_2 = 5$, these two statistics hold level very closely having almost identical empirical levels. For a more close comparison between these two statistics some graphs containing empirical levels are given in Figure 2. From these graphs we conclude that T performs better than T_1 only

   (a) for $n_1 = n_2$ and the variance ratio is moderate ($-.05 < \rho < .05$) and

   (b) for $n_1 \neq n_2$ and sample size of the sample with larger variance is larger.

   In all other situations, T_1, in general, performs better than or same as T.

iii. The non-parametric procedures WC and FP: The Wilcoxon test WC, in general, shows extreme behaviour. It is either conservative or liberal depending on the value of ρ or whether $n_1 < n_2$ or $n_1 > n_2$. The improved non-parametric procedure that is most recently introduced and is available in the R package, is substantially better than WC. The extreme behaviour moderates a lot compared to WC. However, in general, it also does not hold level. Only for $n_1 = n_2$ empirical level performance of this procedure is very close to that of T_1 and T (slightly better than that of T_1 and T when $n_1 = n_2 = 5$ and ρ is not too far from zero).

iv. The bootstrap procedure BT: Only in some instances, for example, for $n_1 = 5$, $n_2 = 25$ and $n_1 = 5$, $n_2 = 30$ and $\rho \leq 0$, level performance of this statistics is similar to those of T_1 and T. However, this is a computer intensive procedure.

v. The Fenstad Statistic Z: This statistic is conservative for smaller sample sizes and liberal for larger sample sizes. Its best performance is for $n_1 = n_2 = 20$, even then it is conservative.

vi The Statistics S, LR and W: The statistics LR and W are in general liberal and the statistic S is conservative. In a lot of situations, for example, for larger sample sizes the statistic S holds nominal level reasonably well (empirical size being very close to those of T_1 and T). Otherwise it is conservative.

For power comparison we considered all combinations of the sample sizes $n_1 = 5, 10, 15, 20, 25, 30$ and $n_2 = 5, 10, 15, 20, 25, 30$. The variance ratios considered were $VR = 1/16, 1/4, 1, 4, 16$. As in the study of performance in terms of size, the power study was done for the nominal level $\alpha = 0.05$. We use $\mu_1 = 1$ and $\mu_2 = \mu_1 + \tau$. The shift parameter τ is calculated as $\tau = \delta\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$ (see, [7]), where $\delta = 1, 2, 3$. Departure from equality of means for fixed but unequal variance is measured by τ. The power results are given in Tables 1 to 36 in Appendix A2 in supplementary material.

We now discuss the power results.

i. The statistic T_N: It shows highest power which is not surprising as it is also highly liberal. It is interesting to note that, even though T_N is more liberal than T_1 and T for $n_1 = n_2 = 30$, it is only slightly more powerful. For large and equal sample sizes ($n_1 = n_2 = 80$) in which its empirical level is close to the nominal level power of this statistic is similar to that of T_1. A Power graph of T_N, T_1 and T for $n_1 = n_2 = 80$ and $\delta = 2$ against $VR = 1/16, 1/4, 1, 4, 16$ is given in Figure 3(a). The statistics T_1 and T show almost indistinguishable power, where as T_N shows slightly larger power. This is in line with the finding that T_N is slightly liberal.

ii. The statistics T_1 and T: Both these statistics show similar power. Power increases as δ increases. See, for instance, power graphs of both these statistics for $n_1 = n_2 = 15$, $\delta = 1$ and $\delta = 2$ against $VR = 1/16, 1/4, 1, 4, 16$ in Figures 3(b, c).

iii. As expected, power of all the other statistics L, W and Z or the procedures BT, WC and FP is more or less than that of T_1 and T depending on whether they are liberal or conservative.

We now examine a situation $n_1 = n_2 = 5$, $\rho = 1/1.69$ from $n_1 = n_2$ in which empirical level performance of the procedure FP is very close to that of T_1 and T. The power graph is given in Figure 3(d) (power against

$\delta = 0, 1, 2, 3$). It shows that power of all three procedures increase as δ increases (as expected). However, as δ increases, power of FP does not increase as fast as the power of T_1 and T. In general, for smaller and equal sample sizes, level performances of the statistics T_1, T, BT, and FP are similar and hold level reasonably close to the nominal. However, in these situations power of the procedure FP is similar or somewhat smaller in comparison to that of the other three statistics or procedures.

iv. The Statistic S: In all those situations in which (for larger sample sizes and for $\rho < 0$) this statistic holds nominal level reasonably well (empirical size being very close to those of T_1 and T) the power of this statistic is also close to those of T_1 and T. Otherwise it is less powerful as expected.

2.7. An Example

This is a set of data from [**?**,]p.83]Lehman1975Nonparametrics. The data which refer to driving times from a person's home to work, measured for two different routes, are 6.5, 6.8, 7.1, 7.3, 10.2 ($n_1 = 5$, $\bar{x}_1 = 7.58$, $s_1^2 = 2.237$) and 5.5, 5.8, 5.9, 6.0, 6.0, 6.0, 6.3, 6.3, 6.4, 6.5, 6.5 ($n_2 = 11$, $\bar{x}_2 = 6.136$, $s_2^2 = 0.073$). The means are different with very different variances. By examining the overall findings of the simulation results above, we see that the only statistic that is appropriate here is the statistic T_1 as $n_1 = 5$, $n_2 = 11$, $s_1^2 = 2.237$, $s_2^2 = 0.073$ are contrary to the situation in which the statistic T or the procedure FP is appropriate.

For these data the p-values of the statistics T_N, T_1, T, L, W, S, Z, BT, WC, FP are 0.0321, 0.0968, 0.0961, 0.0500, 0.0167, 0.1009, 0.0327, 0.3395, 0.0030, 0.0000 respectively.

Now, the value of $T_1 = 2.1426$ with p-value=.0968 indicates that means of the two groups are not different at 10% level of significance.

However, note that (from Figure 1(b) of the supplementary material) both T and T_1 hold level for $n_1 = 5$, $n_2 = 10$ and $\rho > 3$ and their p-values (.0968 and .0961) are also very similar. The same is more or less true for S whose empirical level is below 0.05 but not too much (again from Figure 1(b) of the supplementary material). The p-value of 0.10 for S is also not too different from those of T and T_1. The overall conclusion using the p-values coincide with the findings in Figure 1(b) of the supplementary material. But, since $n_1 = 5$, $n_2 = 11$ and $\hat{\rho} > 3$ for these data the conclusion is that the the hypothesis of equality of the means can be accepted at 10% level of significance. However, at 5% level of significance there is evidence that the two means are different.

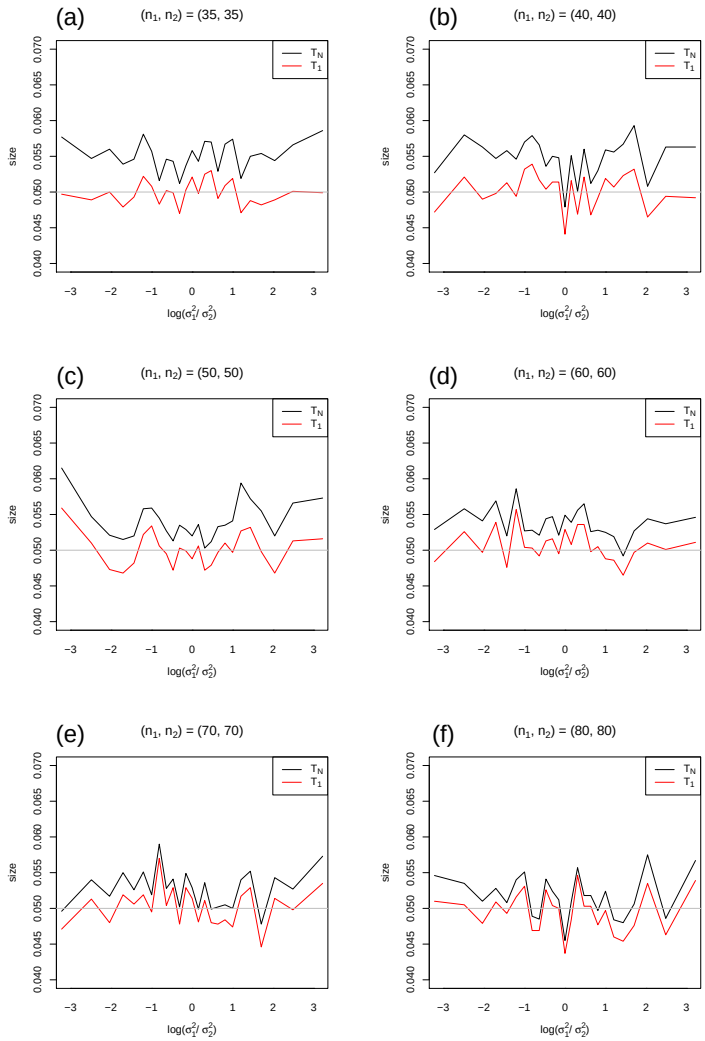

Figure 1: Plots of graphs showing empirical levels of the statistics T_N and T_1 for large sample sizes.

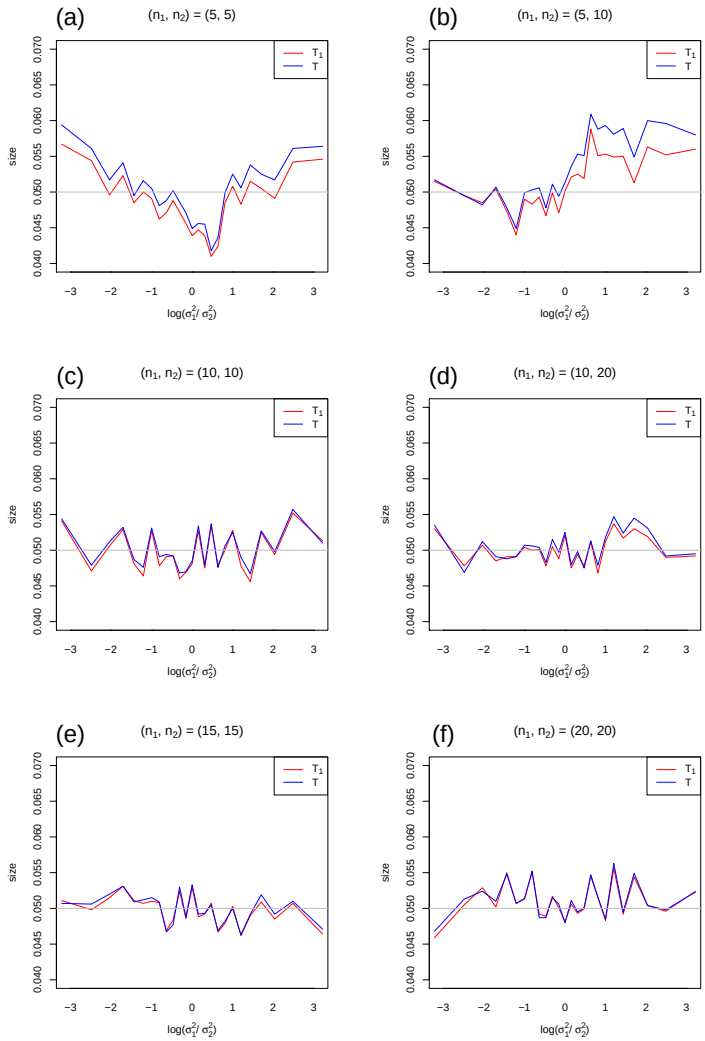

Figure 2: Plots of graphs showing empirical levels of the statistics T_1 and T under certain conditions explained in the text.

3. Two negative binomial populations

3.1. The Negative Binomial Formulation

The most convenient form of the negative binomial distribution, henceforth denoted by $NB(\mu, c)$ is

$$f(y|\mu,c) = Pr(Y = y|\mu,c) = \frac{\Gamma(y + c^{-1})}{y!\Gamma(c^{-1})} \left(\frac{c\mu}{1+c\mu}\right)^{y} \left(\frac{1}{1+c\mu}\right)^{c^{-1}}, \tag{3.1}$$

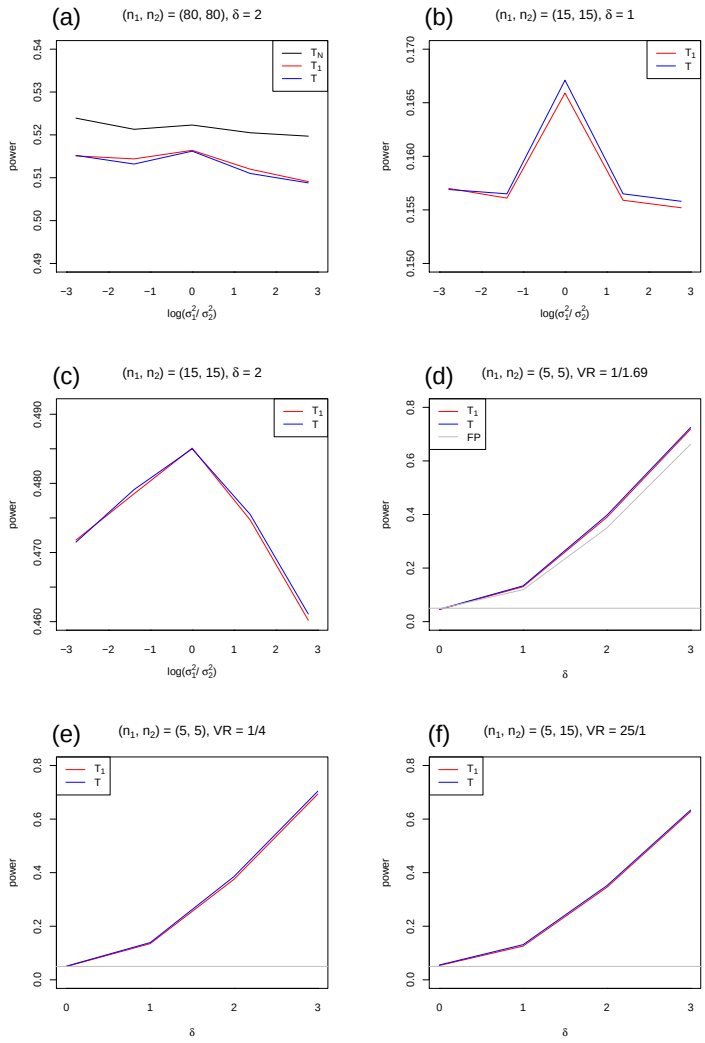

Figure 3: Plots of graphs showing empirical power (a) of the statistics T_N, T_1, and T for $(n_1, n_2) = (80, 80)$ and $\delta = 2$; (b) of the statistics T_1 and T for $(n_1, n_2) = (15, 15)$ and $\delta = 1$; (c) of the statistics T_1 and T for $(n_1, n_2) = (15, 15)$ and $\delta = 2$; (d) of the statistics T_1, T, and FP for $(n_1, n_2) = (5, 5)$ and $VR = 1/1.69$; (e) of the statistics T_1 and T for $(n_1, n_2) = (5, 5)$ and $VR = 1/4$; (f) of the statistics T_1 and T for $(n_1, n_2) = (5, 15)$ and $VR = 25/1$.

for $y = 0, 1, \ldots, \mu > 0$ [32, 33]. See, [30] for further details.

Now, let $y_{i1}, \ldots, y_{in_i}$ be a sample realization from $NB(\mu_i, c_i)$, $i = 1, 2$. Our problem is to test $H_0 : \mu_1 = \mu_2$, where c_1 and c_2 are unspecified. To test this hypothesis [30] develop a likelihood ratio test L, a likelihood ratio test based on the bias corrected maximum likelihood estimates of the nuisance parameters

$L(bc)$, a score test T^2_{NB} (henceforth denoted by S), a score test based on the bias corrected maximum likelihood estimates of the nuisance parameters $S(bc)$, a $C(\alpha)$ test based on the method of moments estimates of the nuisance parameters. [30] show that this later statistic, if Welch's [42] degree of freedom correction is applied, becomes identical to Welch's t-statistic T_1.

[30] investigated by simulations, for level and power, the statistics L, $L(bc)$, S, $S(bc)$, T_1, and the statistic T_N (pretending that negative binomial data can be treated as normal $N(\mu, \sigma^2)$ data). Their simulation study showed no advantage of the bias corrected statistics $L(bc)$ and $S(bc)$ over their uncorrected counterparts. So, here and in subsequent sections any statistic based on bias corrected estimates of the nuisance parameters will not be discussed. The remaining four statistics and the new statistic T developed in Section 2 for normal data are given below.

3.2. The likelihood Ratio Test

The likelihood ratio test is fully described and all necessary results are developed in [30]. So, to save space we omit this from presentation in this paper and refer the reader to that paper.

3.3. The Score Test

The score test statistic (for derivation see, [30]) is

$$S = \sum_{i=1}^{2} \frac{n_i(\bar{y}_i - \tilde{\mu}_0)^2}{\tilde{\mu}_0(1 + \tilde{\mu}_0\tilde{c}_{i0})},$$

which has an asymptotic $\chi^2(1)$ as $n \to \infty$, where $n = n_1 + n_2$.

3.4. The Other Three Statistics T_N, T_1 and T

These three statistics are given in Section 2.1 for data that come from normal distribution. Here the same statistics are used for negative binomial data as if these are normally distributed data.

Apart from the statistic T, which is newly introduced in Section 2.3, [30] show by simulations that for moderate to large sample sizes, in general, the statistic T_1 shows best overall performance in terms of size and power and it is easy to calculate. For large sample sizes, for example, for $n_1 = n_2 = 50$, all four statistics, L, S, T_1, T_N do well in terms of level and their power performances are also similar.

3.5. Simulations

We have conducted a simulation study to compare the 5 statistics T_N, T_1, T, L, and S, the bootstrap procedure BT and the two non-parametric procedures WC and FP. The three statistics T_N, T_1, and T and the three procedures BT, WC, and the FP are applied here exactly the same way as in the case of normally distributed data in Sections 2.4 and 2.5 respectively.

To compare the statistics in terms of size, we considered all combinations of the sample sizes $n_1 = 5, 10, 15, 20, 25, 30$ and $n_2 = 5, 10, 15, 20, 25, 30$, $\mu_1 = \mu_2 = 2$, $c_1 = .10, .25, .40, .55, .70, .85, 1$, $c_2 = .10, .25, .40, .55, .70, .85, 1$, and a nominal level $\alpha = .05$. These results are all given as graphs in Figures 1-6 in Appendix B1 in Supplementary Material. The graphs are in terms of size against $\rho = \log(c_1/c_2)$. All simulation results are based on 10,000 samples. A discussion of the size results is given in what follows.

i. For $n_1 = n_2 = 5, 10$, the L statistic holds level most effectively (though somewhat conservative for $n_1 = n_2 = 5$ and somewhat liberal for $n_1 = n_2 = 10$), This finding is in line with Paul and Alam (2014). In these situations another statistic that is competing with L having very similar level is T_N.

ii. For the smaller of n_1 and n_2 equal to 5 and the other equal to 10 to 30, the L statistic performs best, although consistently somewhat conservative. In these situations, for all other statistics no consistent pattern emerges. For example, T_N is mainly very highly liberal, only in a very few situations its empirical level is close to the nominal level. For the smaller of n_1 and n_2 equal to 10 and the other equal to 10 to 30, the L statistic performs best, although consistently somewhat liberal. In these situations the other statistics are either liberal or conservative. For unequal sample sizes, smaller of n_1 and n_2 less than 20 and the other up to 30 the L statistic seems to perform best.

iii. For the smaller of n_1 and n_2 equal to or greater than 20 and the other also equal to or greater than 20, overall, the best performing procedures are through the use of the statistic T_1 or T or the score test statistc S. At $n_1 = n_2 = 30$ empirical level of all these 3 procedures are very close to the nominal level.

For power comparison we consider all combinations of the sample sizes $n_1 = 5, 10, 15, 20, 25, 30$ and $n_2 = 5, 10, 15, 20, 25, 30$. We use $\mu_1 = 1$, $c_1 = .1$, $c_2 = .10, .25, .40, .55, .70, .85, 1$, and $\mu_2 = \mu_1 + \delta$, for $\delta = 1.0, 1.5, 2.0$. As in the study of performance in terms of size, the power study was done for the nominal

level $\alpha = 0.05$. All simulation results are based on 10,000 samples. A discussion of the power results is given in what follows.

We first concentrate on the L statistic which seems to be doing better in terms of size for the smaller of n_1 and n_2 less than 20 and the other up to 30. The power results are given in Tables 1 to 27 in Appendix B2 in supplementary material. In general, the L statistic shows highest power. Only in some situations the statistic T_N or T_1 or T show higher power, but in these situations these later statistics are also liberal.

Now we discuss power performance of the statistics T_1, T and S which perform best in terms of size starting at $n_1 = 20$ and $n_2 = 20$. Here we compare these only with the L statistic as it is, in general, liberal or conservative but not too much. The power results are given in Tables 28 to 36 in Appendix B2 in supplementary material. The L statistic, in general, is somewhat more powerful than the other three statistics, but it is also slightly liberal in comparison to the other three statistics. The other 3 statistic show similar power. For example, for $n_1 = n_2 = 20$ and $c_2 = .7$ empirical level of L is close to 0.06 and those of the other three are close 0.05 (see, graph for $n_1 = n_2 = 20$ in Figure 4). The powers for L, T_1, T and S, $\delta = 2$, are 0.694, 0.554, 0.555 and 0.572 respectively (see, Table 22).

In general, power decreases as the value of c_2 goes further away from $c_1 = .10$ and increases as the sample size increases.

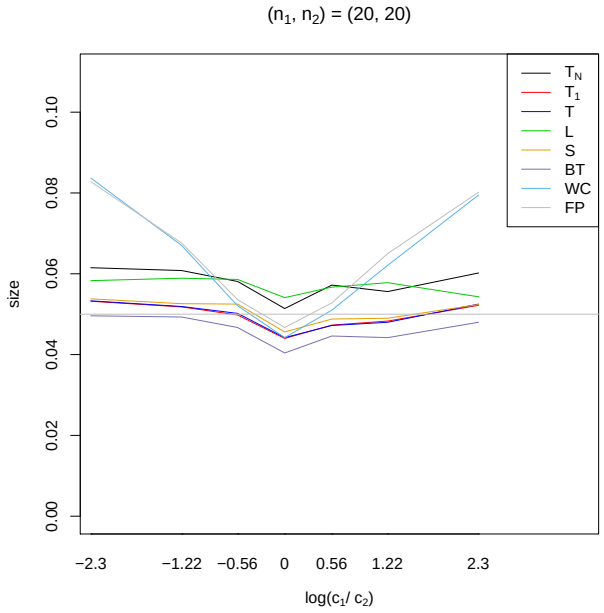

Figure 4: Plots of graphs showing empirical levels of all the statistics for $(n_1, n_2) = (20, 20)$.

Table 1: Frequency of patients by number of lesions on each patients angiogram [6].

Number of lesions (y_{ij})	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$\hat{\mu}$	$\hat{c}$
Cholestyramine	5	4	6	5	7	7	6	6	7	2	2	1	0	0	1	4.932	0.250
Placebo	2	4	6	4	6	9	7	5	2	4	4	2	0	2	0	5.509	0.185

3.6. An Example

[36] presents a set of data, originally given by [6], to see the effectiveness of a treatment (Cholestyramin), in comparison to a placebo, in reducing the number of vascular lesions. The data are given in Table 1, which refer to the observed number of vascular lesions on each patient's angiogram in the treatment group as well as in the control group (placebo).

The maximum likelihood estimates of μ and c based on a negative binomial model for the two groups are given in this table as well. The $\hat{\mu}$'s and the $\hat{c}$'s both differ. We now apply the statistics T_1, T and T_{NB}^2 to test the equality of the two means. The values of T_1, T and S with p-value in the parenthesis are -0.379(.705), -0.379(.704), and 0.146(.702) respectively. Based of the p-values which are very close the difference is not significant.

We now show how to apply the bootstrap critical value method using the likelihood ratio statistic L for small sample sizes. For this we take a sample of size $n_1 = 15$ with replacement from the treatment group and a sample of size $n_1 = 10$ with replacement from the control group which are given below
Treatment group: 8 8 10 5 2 0 0 7 3 1 1 3 8 6 0
Placebo group: 1 1 2 9 13 4 6 9 10 6
Suppose these are the observed data for the two groups. For these data the value of L is 1.26 and the bootstrap 95% critical value is 5.16 which indicates that the difference between the two means is not significant.

The bootstrap critical value is obtained as: from the sampled data of $n_1 = 15$ and $n_2 = 10$ above we take 10000 pairs of samples (one sample of size 15 from the treatment group and one sample of size 10 from the control group) with replacement. For each pair of samples we obtain the value of L. Then the bootstrap critical value is the 9500^{th} value of the ordered (from smallest) L values.

4. Two beta-binomial populations

4.1. The Beta-Binomial Formulation

For modelling data in the form of proportions with extra-dispersion the most popular model is the extended beta-binomial distribution of [34]. Let $y|p \sim$ binomial(m, p), where p is a beta random variable with mean π and variance $\pi(1-\pi)\phi$, where ϕ is an extra dispersion parameter. Then the unconditional distribution of y is the extended beta-binomial distribution of [34] for which the pmf is given in what follows.

$$Pr(y|\pi,\phi) = \binom{m}{y} \frac{\prod_{r=0}^{y-1}[\pi(1-\phi)+r\phi] \prod_{r=0}^{m-y-1}[(1-\pi)(1-\phi)+r\phi]}{\prod_{r=0}^{m-1}[(1-\phi)+r\phi]} \tag{4.1}$$

with mean $m\pi$ and variance $m\pi(1-\pi)(1+(m-1)\phi)$, where $0 \leq \pi \leq 1$, and $\phi \geq \max[-\pi/(m-1), -(1-\pi)/(m-1)]$.

Denote this probability mass function by BB(m, π, ϕ). Now, let y_{i1}/m_{i1}, . . ., y_{in_i}/m_{in_i} be a sample realization from BB(m_{ij}, π_i, ϕ_i), $i = 1, 2$, $j = 1, \ldots, m_{in_i}$. Our purpose is to test $H_0 : \pi_1 = \pi_2$ with ϕ_1 and ϕ_2 being unspecified. [3] develop eight tests, namely, a likelihood ratio test, a $C(\alpha)$ (score) test based on the maximum likelihood estimates of nuisance parameters, a $C(\alpha)$ test based on the [**?**] method of moments estimates of the nuisance parameters, a $C(\alpha)$ test based on the quasi-likelihood and the method of moments estimates of the nuisance parameters by [8], a $C(\alpha)$ test based on the quasi-likelihood and the method of moments estimates of the nuisance parameters by [39], a $C(\alpha)$ test based on extended quasi-likelihood estimates of the nuisance parameters, and two non-parametric tests by [35]. See, [3] for further details.

By doing an extensive simulation study [3] show that none of the statistics, except the $C(\alpha)$ statistic C_{BB}, does well in terms of level and power. The statistic C_{BB} holds nominal level most effectively (close to the nominal level) and it is at least as powerful as any other statistic which is not liberal. It has the simplest formula, is based on estimates of the nuisance parameters only under the null hypothesis and is easiest to calculate. Also, it is robust in the sense that no distributional assumption is required to develop this statistic.

In this paper we compare the performance C_{BB} with the statistics T_N, T_1 and T, the bootstrap procedure BT and the two non-parametric procedures WC and FP. These are described below for the application to data in the form of proportions.

4.2. The Statistic C_{BB}

The statistic C_{BB} is (detailed derivation is given in [3] $C_{BB} = C^2/(A - A^2/B)$, which is distributed as chi-squared, asymptotically, as $n \to \infty$ ($n = n_1 + n_2$), with 1 degree of freedom, where

$$C = \sum_{j=1}^{n_1} \left[\frac{1}{1 + (m_{1j} - 1)\phi_1} \left\{ \frac{y_{1j}}{\pi} - \frac{m_{1j} - y_{1j}}{1 - \pi} \right\} \right],$$

$$A = \sum_{j=1}^{n_1} \left[\frac{1}{1 + (m_{1j} - 1)\phi_1} \left\{ \frac{m_{1j}}{\pi(1 - \pi)} \right\} \right]$$

and

$$B = \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left[\frac{1}{1 + (m_{ij} - 1)\phi_i} \left\{ \frac{m_{ij}}{\pi(1 - \pi)} \right\} \right].$$

The parameters π, ϕ_1 and ϕ_2 in C, A and B are replaced by the maximum extended quasi-likelihood estimates $\hat{\pi}$, $\hat{\phi}_1$ and $\hat{\phi}_2$ obtained by solving

$$\sum_{i=1}^{2} \sum_{j=1}^{n_i} \left[\frac{1}{1 + (m_{ij} - 1)\phi_i} \left\{ \frac{y_{ij}}{\pi} - \frac{m_{ij} - y_{ij}}{1 - \pi} \right\} \right] = 0,$$

$$\sum_{j=1}^{n_1} \left[\frac{m_{1j} - 1}{\{1 + (m_{1j} - 1)\phi_1\}^2} \left\{ y_{1j} \log\left(\frac{z_{1j}}{\pi} \right) + (m_{1j} - y_{1j}) \log\left(\frac{1 - z_{1j}}{1 - \pi} \right) \right. \right.$$

$$\left. \left. - \frac{1 + (m_{1j} - 1)\phi_1}{2} \right\} \right] = 0$$

and

$$\sum_{j=1}^{n_2} \left[\frac{m_{2j} - 1}{\{1 + (m_{2j} - 1)\phi_2\}^2} \left\{ y_{2j} \log\left(\frac{z_{2j}}{\pi} \right) + (m_{2j} - y_{2j}) \log\left(\frac{1 - z_{2j}}{1 - \pi} \right) \right. \right.$$

$$\left. \left. - \frac{1 + (m_{2j} - 1)\phi_2}{2} \right\} \right] = 0$$

simultaneously.

4.3. The Bootstrap Procedure

The bootstrap procedure is developed here for data in the form of proportions (e.g. x/n) as follows:

1. Calculate the continuous data in the form of proportions for the two samples as $p_{ij} = y_{ij}/m_{ij}$, $i = 1, 2$, $j = 1, \ldots, m_{in_i}$. Let $\bar{p}_i = \sum_{j=1}^{n_i} p_{ij}/n_i$ and

$s^2_{iP} = \sum_{j=1}^{n_i}(p_{ij} - \bar{p}_i)^2/(n_i - 1)$. Then, define a statistic T_P, analogous to T_N, as

$$T_P = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\frac{s^2_{1P}}{n_1} + \frac{s^2_{2P}}{n_2}}}.$$

2. Obtain a bootstrap sample of size n_i; say p^*_{ij}, from the adjusted p_{ij}, that is, from $p^{adj}_{ij} = p_{ij} - \bar{p}_i + \bar{p}$, where $\bar{p}$ is the overall mean of p_{ij} .

3. Calculate

$$T^*_P = \frac{\bar{p}^*_1 - \bar{p}^*_2}{\sqrt{s^{2*}_{1P}/n_1 + s^{2*}_{2P}/n_2}}.$$

4. Repeat step 2 and 3 B times ($B = 999$); thereby obtaining 999 bootstrap values of T_P*.

5. For a two sided test, a difference between the means is significant if the observed value of $|T_P| >$ ($100(1 - \alpha/2)th$ values of T^*_P.

4.4. The Other Three Statistics T_N, T_1, and T and The three Procedures BT, WC, and FP

Calculation of the three statistics T_N, T_1, and T and the three procedures BT, WC, and FP proceed by considering the p_{ij}, as y_{ij} in Section 2.

4.5. Simulations

We have conducted a simulation study to compare, in terms of level and power, the statistics C_{BB}, T_N, T_1 and T, the bootstrap procedure BT and the two non-parametric procedures WC and FP.

To generate data y_{ij} from $BB(m_{ij}, \pi_i, \phi_i)$, we take random samples with replacement of $n_1 = 5, 10, 15, 20, 25, 30$ litters with the litter sizes m_{1j}, $j = 1, \ldots, 27$ of the control group (Group 1) and $n_2 = 5, 10, 15, 20, 25, 30$ litters with the litter sizes m_{2j}, $j = 1, \ldots, 21$ of the medium group (Group 2) of Paul (1982). The m_{1j}, $j = 1, \ldots, 27$ of group 1 were 12, 7, 6, 6, 7, 8, 10, 7, 8, 6, 11, 7, 8, 9, 2, 7, 9, 7, 11, 10, 4, 8, 10, 12, 8, 7, 8 and m_{2j} of group 2 were 4, 4, 9, 8, 9, 7, 8, 9, 6, 4, 6, 7, 3, 13, 6, 8, 11, 7, 6, 10, 6. Note that our simulation study is much more extensive in comparison to [3]. Where as [3] consider fixed sample sizes ($n_1 = 27$ and $n_1 = 21$), we consider random samples of different sizes given above. The different combinations of parameter values are also much more extensive in our study.

For empirical levels we considered $\pi_1 = \pi_2 = \pi = 0.05, 0.10, 0.20, 0.40, 0.50$ and (ϕ_1, ϕ_2) = (0.05, 0.50), (0.10, 0.40), (0.15, 0.30), (0.20, 0.20), (0.30, 0.15), (0.40, 0.10), (0.50, 0.05).

For power comparison the values of π_1 and π_2 considered were according to the formula $\pi_2 = \pi_1 + \delta$ with $\pi_1 = 0.05, 0.10, 0.20, 0.40$ and $\delta = 0.05, 0.10, 0.20$. That is, for each value of π_1 power has been simulated for three increments $0.05, 0.10, 0.20$. The same combination of values (ϕ_1, ϕ_2) were chosen as in the study of level performance.

All simulation results are based on 10,000 good samples. The definition of good samples here is "those samples for which the estimating equations converged within the permitted range $\cap_j (-1/(n_{ij} - 1)) < \phi_i < 1$, $i = 1, 2$. For more details see [3].

The empirical level results are summarized in Figures 1-36 in Appendix C1 and empirical power results are summarized in Tables 1-36 in Appendix C2 in Supplementary Material. The Level results are graphed against $\log(\phi_1/\phi_2)$ and power tables are in terms of VR = (ϕ_1/ϕ_2).

We now discuss the size results of the 7 statistics:

(i) The statistics T_N: In general, the statistic T_N does not show any consistent behaviour, although shows mostly highly liberal behaviour.

(ii) The statistics T_1 and T: In general, level performance of these two statistics are similar. These two statistics hold level reasonably well when n_1, n_2 and π are all large, for example, for $n_1 \geq 20$ and $n_2 \geq 20$ and π ($\geq .2$). See Figures 22, 23, 24, 28, 29, 30, 34, 35 and 36 in Appendix C1 of the supplementary material. For some other situations, for example for $n_1 = n_2 = 10, 15$ and $\pi \geq .20$, performance of these two statistics are also the best and hold nominal level reasonably well. See Figures 8 and 15 in Appendix C1 of the supplementary material.

(iii) The statistic C_{BB}, recommended by Alam and Paul (2017): In some small sample size situations this statistic holds level reasonably well. See, for example, the situations in which one of the sample size is large and π is not too large ($\pi = .20$) (graphs (c) in figures 8-13, 15-18, 21-24, 28-36 in Appendic C of the supplementary material). For small π (in case of some of π=.05, .10, and .20) C_{BB} performs best (see Figures 8(a,b,c), 9(c), 10(c), 11(c), 14(c), 15(c), 16(c), 17(c), 21(c) in Appendix C1 of the supplementary material).

For large sample sizes ($n_1 \geq 20$ and $n_2 \geq 20$) level performance of C_{BB} is close to those of T_1 and T for $\pi = .2$. However, as π increases from .2 it shows conservative behaviour (see Figures 22, 23, 24, 28, 29, 30, 34, 35 and 36 in Appendix C1 of the supplementary material).

(iv) Performance of all other statistics are erratic at the best.

Next we discuss power performance.

(i) Since level performance of the statistic T_N, the bootstrap procedure BT and the two non-parametric procedures WC and FP are, in general, not satisfactory, we do not discuss their power performances, although power results are given in the supplementary material.

(ii) Power of T_1 and T are similar in all situations studied. Note from the level results that for large sample sizes ($n_1 \geq 20$ and $n_2 \geq 20$) level performance of C_{BB} is close to those of T_1 and T for $\pi = .2$ and as π increases from .2 it shows conservative behaviour. In all these situations power of C_{BB} is the best. That means, C_{BB} shows higher power even in situations where it is conservative but T_1 and T hold level. So, in these situations, unless C_{BB} can be adjusted to hold level we can not recommend its use. Power of C_{BB}, in most small sample sizes and small π ($< .2$) situations in which it holds level, in general, is larger or similar to those of T_1 and T.

4.6. Two Examples

Example 1. Here, for illustrative purposes, we use data from an experiment, given in [40], to identify in utero damage in laboratory rodents after exposer to boric acid. The study design involved four doses of boric acid. The compound was administered to pregnant female mice during the first 17 days of gestation, after which the dams were sacrificed and their litters examined. Table 2 lists the number of dead embryos and total number of implants at each of the four exposure doses: $d1 = 0$ (control), $d2 = 0.1$, $d3 = 0.2$, and $d4 = 0.4$ (as percent boric acid in feed).

The maximum likelihood estimates of the parameters (π, ϕ) for the four dose groups are also given in Table 2. It shows that the estimates of the $\hat{\pi}$'s are different and also the estimates of the $\hat{\phi}$'s are different. Now, suppose we want to compare π of the control group ($d_1 = 0$) with that of the dose group 4 ($d_4 = .4$). That is, we want to test $H_0 : \pi_1 = \pi_4$.

Now, the maximum likelihood estimate of π_1 is 0.069. If we assume that 0.069 is the true value of π_1 and $H_0 : \pi_1 = \pi_4$ is true, then, under the null hypothesis, the value of the common π is 0.069. Further, the sample sizes in the two groups are 27 and 26 which are between (25,25) and (30,30). Now, looking at Figures 29, 30, 35 and 36 in Appendix C1 of the supplementary material we see that none of the statistics hold nominal level for $\pi = 0.069$ and sample sizes $n_1 = 27$ and $n_2 = 26$. So, we apply a Monte Carlo Procedure (MCP) similar to the parametric bootstrap. For this we consider

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Table 2: Per-litter data from Teratological study of boric acid (Stalon, et al. (2000). (i) Number of dead embryos. (ii) Total number of implants. Doses $d_1 = 0$ (control), $d_2 = 0.1$, $d_3 = 0.2$, $d_4 = 0.4$.

Dose Group															$\hat{\pi}$	$\hat{\phi}$
$d_1 = 0$	(i)	0	0	1	1	1	2	0	0	1	2	0	0	3	.0692	.0219
	1	0	0	2	3	0	2	0	0	2	1	1	0	0		
	(ii)	15	3	9	12	13	13	16	11	11	8	14	13	14		
	13	8	13	14	14	11	12	15	15	14	11	16	12	14		
$d_2 = 0.1$	(i)	0	1	1	0	2	0	0	3	0	2	3	1	1	.0968	.0058
	0	0	0	1	0	2	2	2	3	1	0	1	1	1		
	(ii)	6	14	12	10	14	12	14	14	10	12	13	11	11		
	11	13	10	12	11	10	12	15	12	12	12	12	13	15		
$d_3 = 0.2$	(i)	1	0	0	0	0	0	4	0	0	1	2	0	1	.0521	.0245
	1	0	0	1	0	1	0	0	1	2	1	0	0	1		
	(ii)	12	12	11	13	12	14	15	14	12	6	13	10	14		
	12	10	9	12	13	14	13	14	13	12	14	13	12	7		
$d_4 = 0.4$	(i)	12	1	0	2	2	4	0	1	0	1	3	0	1	.2234	.2497
	0	3	2	3	3	1	1	8	0	2	8	4	2			
	(ii)	12	12	13	8	12	13	13	13	12	9	9	11	14		
	10	12	21	10	11	11	11	14	15	13	11	12	12			

Note that if we apply a t-test with Welch's degree of freedom, it becomes the procedure T_1. We now do the test by obtaining approximate critical values, for a two sided test, of the exact distribution of t which are calculated as what is given below.

Keep m_{ij} fixed as given in the two groups, $j = 1, \ldots, 27$ for $i = 1$ and $j = 1, \ldots, 26$ for $i = 2$. Now, generate random numbers from $BB(m_{1j}, 0.069, 0.0218)$ for $j = 1, \ldots, 27$ and random numbers from $BB(m_{2j}, 0.069, 0.2496)$ for $j = 1, \ldots, 26$. This gives one sample for which calculate the value of t. Repeat this procedure and generate 100,000 samples and thereby 100,000 values of t. Order these 100,000 values from the smallest to the largest. The 2500^{th} and the 97500^{th} values are the 2.5% and the 97.5% critical values.

Now, the value of t from the data in the dose groups $d_1 = 0$ and $d_4 = .4$ is -2.8182. If -2.8182 does not fall between the 2.5% and the 97.5% critical values reject the null hypothesis of equality of the two proportions at 5% level of significance.

Following the procedure described above, the 2.5% and the 97.5% critical values obtained are -1.673003 and 2.637581 respectively. Since $T_1 = -2.8182$ falls in the rejection region the null hypothesis $H_0 : \pi_1 = \pi_4$ is rejected.

To check whether this procedure works we did some further simulations. For empirical level we again obtained 100,000 values of t as above with $\pi = 0.069$. We then calculated the proportion of t values that fall outside (-1.673003, 2.637581).

Table 3: Power Table of T_1 and MCP, $\pi = 0.069 + \delta$, $\delta = 0, 02, .04, \ldots, .14$.

δ	0.00	0.02	0.04	0.06	0.08	0.10	0.12	0.14
T_1	0.068	0.054	0.119	0.248	0.424	0.604	0.756	0.866
MCP	0.052	0.100	0.227	0.406	0.599	0.757	0.868	0.938

Table 4: Data from an in vivo cytogenetic assay [44].

Dose Group	No. of aberrant cells in 50 cells per animal										$\hat{\pi}$	$\hat{\phi}$
Negative control	0	4	0	0	4	0	1	1	0	0	0.0199	0.0447
Low Dose	1	0	3	0	1	0	3	0	0	1	0.0180	0.0125
Medium Dose	6	5	0	3	7	1	1	0	0	0	0.0454	0.0690
High Dose	3	2	1	6	4	0	0	0	0	5	0.0417	0.0476

When this proportion is multiplied by 100 we obtain the empirical level. For power we do exactly the same as above but now take $\pi = 0.069 + \delta$, where $\delta = 0.02, .04, \ldots, .14$. The power results are given in Table 3.

To compare the performance of the above Monte Carlo method with that of T_1 we extended the simulation study by obtaining the proportion of the 100,000 samples for which $|t| >$ the critical value of T_1 with Welch's degree of freedom. Results are also given in Table 3, which show that the new Monte Carlo procedure holds level almost exactly, the Welch T_1-test is somewhat liberal and yet the new procedure shows higher power compared to T_1.

Example 2. A data set from [44] of an in vivo cytogenetic assay is given Table 4. In this example, the sample sizes $n_1 = n_2 = 10$ are small in which the extended quasi-likelihood based score test C_{BB} does well (see Figure 8(a,b,c) in Appendix C1 of the supplementary materia). For illustrative purpose, we test the equality of proportions in the first two groups. For this the value of $C_{BB} = 0.0171$ with p-value=0.8660 showing strong support for the null hypothesis of the two proportions.

5. Two Weibull populations

5.1. The Weibull Formulation

Data in the form of survival times arise in many fields of studies such as engineering, manufacturing, aeronautics and bio-medical sciences. See [28] for a recent review. The two parameter Weibull random variable Y with shape

parameter β and scale parameter α has the probability density function

$$f(y;\alpha,\beta)=\frac{\beta}{\alpha}\left(\frac{y}{\alpha}\right)^{(\beta-1)}\exp\left[-\left(\frac{y}{\alpha}\right)^{\beta}\right];\ y\geq 0;\ \beta,\alpha>0. \tag{5.1}$$

The mean and variance of Y are $\mu=\alpha\Gamma(1+1/\beta)$ and $\sigma^2=\alpha^2[\Gamma(1+2/\beta)-\{\Gamma(1+1/\beta)\}^2]$ respectively.

In some practical data analytic problems lifetimes or survival times data arise in the form of two samples following two independent Weibull populations with different shape and scale parameters. Let $y_{11},\ldots,y_{1n_1}$ and $y_{21},\ldots,y_{2n_2}$ be samples from two independent Weibull populations with parameters (α_1,β_1) and (α_2,β_2) respectively. In such a situation it may be of interest to test the equality of the scale parameters with the shape parameters being unspecified. That is to test the null hypothesis $H_0:\alpha_1=\alpha_2$, where β_1 and β_2 are unspecified.

For this problem [2] develop four test statistics, namely, a likelihood ratio statistic, a score statistic, and two $C(\alpha)$ statistics; one of which is based on the method of moments estimates of the nuisance parameters by [11] and the other is based on the method of moments estimates of the nuisance parameters by [41]. However, through a simulation study they show that the two statistics based on the method of moments estimates of the nuisance parameters perform best.

However, the actual analog of the Behrens-Fisher problem is to test $H_0:\mu_1=\mu_2$ with σ_1^2 and σ_2^2 being unspecified. To deal with this problem we develop a score test in Section 5.2. In Section 5.3 we conduct a simulation study to compare this statistic for level and power with the statistics T_N, T_1 and T, and the procedures BT, WC and the FP.

5.2. The Score Test

A score test statistic (derivation is given in Appendix E of the supplementary material) for testing $H_0:\mu_1=\mu_2$, where σ_1^2 and σ_2^2 are unknown and unspecified is given by $S_w=S^2/I$, where

$$S=\frac{1}{\Gamma(1+\beta_1^{-1})}\left[-\frac{n_1\beta_1}{\mu}+\frac{\beta_1\{\Gamma(1+\beta_1^{-1})\}^{\beta_1}}{\mu^{\beta_1+1}}\sum_{j=1}^{n_1}y_{1j}^{\beta_1}\right]$$
$$+\frac{1}{\Gamma(1+\beta_2^{-1})}\left[\frac{n_2\beta_2}{\mu}-\frac{\beta_2\{\Gamma(1+\beta_2^{-1})\}^{\beta_2}}{\mu^{\beta_2+1}}\sum_{j=1}^{n_2}y_{2j}^{\beta_2}\right]$$

and

$$I = \frac{1}{\Gamma(1+\beta_1^{-1})}\left\{\frac{n_1\beta_1}{\mu^2} - \frac{\beta_1(\beta_1+1)\{\Gamma(1+\beta_1^{-1})\}^{\beta_1}}{\mu^{\beta_1+2}}\sum_{j=1}^{n_1}\mathrm{E}(y_{1j}^{\beta_1})\right\} - \frac{1}{\Gamma(1+\beta_2^{-1})}\left\{\frac{n_2\beta_2}{\mu^2} - \frac{\beta_2(\beta_2+1)\{\Gamma(1+\beta_2^{-1})\}^{\beta_2}}{\mu^{\beta_2+2}}\sum_{j=1}^{n_2}\mathrm{E}(y_{2j}^{\beta_2})\right\}.$$

In S and I the quantity, such as $\mathrm{E}(y^{\beta_i})$ is calculated as $\mathrm{E}(y^{\beta_i}) = \int_0^\infty y_i^\beta f(y,\mu,\beta_i)dy$. Of course, the parameters μ, β_1 and β_2 in S and I are to be replaced by their maximum likelihood estimates $\hat{\mu}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ which are obtained by maximizing the log-likelihood function

$$l = \sum_{i=1}^{2}\left[\frac{1}{\Gamma(1+\beta_i^{-1})}\left\{n_i\log\left(\frac{\beta_i\Gamma(1+\beta_i^{-1})}{\mu}\right) + (\beta_i-1)\left\{\sum_{j=1}^{n_i}\log(y_{ij}) - n_i\log\left(\frac{\mu}{\Gamma(1+\beta_i^{-1})}\right)\right\}\right\}\right] - \sum_{i=1}^{2}\frac{\left\{\Gamma(1+\beta_i^{-1})\right\}^{\beta_i-1}}{\mu^{\beta_i}}\sum_{j=1}^{n_i}y_{ij}^{\beta_i}$$

with respect to the parameters μ, β_1 and β_2. The distribution of S_w is asymptotically distributed as chi-square with one degrees of freedom.

5.3. Simulations

We have conducted a simulation study to compare the statistic S_w, in terms of level and power, with the three statistics T_N, T_1 and T, and the three procedures BT, WC, and FP. These statistics are applied here exactly the same way as in the case of normally distributed data studied in Sections 2.4 and 2.5. As in the two previous sections we use the Weibull data as if the data come from two normal populations.

To compare the statistics in terms of size and power, we considered the sample sizes $n_1 = 5, 10, 15, 20, 25, 30$ and $n_2 = 5, 10, 15, 20, 25, 30$. We generate data from the Weibull (α_1, β_1) and Weibull (α_2, β_2) populations. For size comparison, in order to comply with equal means condition, we fix the values of α_1, β_1, and β_2; and evaluate the expression $\{\alpha_1\Gamma(1+1/\beta_1)\}/\{\alpha_2\Gamma(1+1/\beta_2)\} = 1$ to obtain the value of α_2. For power comparison, we again fix the values of α_1, β_1, and β_2; but evaluate the expression $\{\alpha_1\Gamma(1+1/\beta_1)\}/\{\alpha_2\Gamma(1+1/\beta_2)\} = 1/(1+\delta)$ with $\delta = .1, .2, .3$, to obtain the value of α_2. Both the size and power are calculated for all combinations of $\beta_1 = 1, 2, 3, 4, 5$ and $\beta_2 = 2, 3, 4$ while fixing $\alpha_1 = 1$ and determining α_2 from the expressions given above.

The size results are all given as graphs in Figures 1-36 and the power results are all given in Tables 1-36 in Appendix D2 in supplementary material.

The graphs are in terms of size against $\rho = \log(\sigma_1^2/\sigma_2^2)$. All simulation results are based on 10,000 samples.

We now discuss the size results, of the 7 statistics, given in Figures 1-36 in Appendix D1 in the supplementary material:

(i) The statistic T_N: The statistic T_N is liberal, highly liberal for smaller n_1 and n_2. Even for $n_1 = n_2 = 30$ it is liberal, empirical level ranging, on average, from 0.0525 (when $VR \approx 1$) to 0.0781 (as VR is further and further away from 1).

(ii) The statistics T_1 and T: Overall, these two statistics perform best, even for smaller sample sizes, holding empirical levels closer to the nominal. Only exceptions are when the sample size differences are large as well as when the differences between the variances are large, also when $n_2 > n_1$ as well as $\sigma_2^2 > \sigma_1^2$. In these situations both of these statistics can be quite liberal, although T_1 is slightly better than T. See, for example, Figures 4, 5, 11, 12, 25 of Appendix D1 of the the supplementary material.

(iii) Behaviour of the remaining four statistics or procedures are inconsistent, sometimes very liberal and sometimes very conservative. The exceptions are for

  (a) FP for $n_1 = n_2$ which does as well as T_1 and T in some cases (see, for example, Figure 1),

  (b) BT, irrespective of sample sizes, which does as well or better than T_1 and T (see, for example, Figure 5).

Next we discuss power performance using the power results given Tables 1-36 in Appendix D2 in the supplementary material.
Since the procedures T_N, WC, and S_w have highly inconsistent behaviour in terms of level, we omit these from power discussion. Power of T_1 and T are similar. However, T shows some edge over T_1. In general, these show higher power than FP and BT. Even in the situations in which FP and BT have slight advantage in terms level, T_1 and T maintain higher power.

5.4. An Example

[17] give data on survival times (in weeks) for two groups of patients who died of acute myelogenous leukemia. Patients were classified into the two groups according to the presence or absence of a morphologic characteristic of white cells. Patients termed AG positive were identified by the presence of Auer rods and/or significant granulature of the leukemic cells in the bone marrow at diagnosis. For the AG negative patients these factors were absent. The survival times for 17 patients in the AG positive group were: 65, 156, 100, 134, 16, 108, 121, 4, 39,

143, 56, 26, 22, 1, 1, 5, 65 and those for 16 patients in the AG negative group were: 56, 65, 17, 17, 16, 22, 3, 4, 2, 3, 8, 4, 3, 30, 4, 43.

We now test the equality of the mean survival times in the two groups. As the performance of the five statistics or procedures T_N, WC, BT, FP, and C_W are far less than satisfactory we do not consider them any further. The values of T_1 and T with corresponding p-values in the parenthesis are 3.1124 (0.0054) and 3.1124 (0.0047) respectively leading the conclusion that the two means are not the same.

6. Discussion

We do a comprehensive review of the standard Behrens Fisher (BF) problem and some of its analogs. Among the B-F analogous problems we deal with the two parameters negative binomial, the Beta-binomial, and the two parameter Weibull. In each case a number of procedures are either reviewed or developed and extensive simulation studies are conducted to study the properties of the procedures in terms of size and power. Some new results and findings are shown and examples of application are given in all cases.

If the variance ratio is known, the mixing parameter λ in $\mathcal{K}$ is then known, so the distribution of T (§2.3) becomes pivotal, which is not an exact t-distribution. In fact, if the variance ratio is given, one should use the pooled variance estimator which can lead to a t-statistic. For other distributions other than the normal cases, it is the same story but in an asymptotical sense. The tests based on t-distributions or chi-square distributions or any other derived from "normal distributions all become asymptotical approximations. Therefore, if there is some reason to specify the variance ratio σ_2/σ_1, the traditional two independent samples Student t-test or Welch test are usable but both are approximations.

A review paper can possibly be never complete given that a vast literature is available. Here also we do not make such a claim. For example, we do not consider the Bayesian methods [24, 45] to the solve Behrens Fisher problem.

For the standard Behrens Fisher problem we studied 10 procedures T_N, T_1, T, L, W, S, Z, BT, WC and FP including a new procedure T. Based on the finding through extensive simulation study we recommend that the statistic T_N be used only when the two means are visibly different or if the sample sizes are large, such as, $min(n_1, n_2) \geq 80$ (only at this sample size level the Central Limit Theorem reasonably takes hold); otherwise use T_1 except for

(i) (a) $n_1 = n_2$ and the variance ratio is not extreme (close to 1/25 or 25/1 limits).

(b) for $n_1 \neq n_2$ and sample size of the sample with larger variance is larger, in which case use T.

(ii) for smaller and equal sample sizes use the procedure FP.

For the negative binomial BF Problem we studied five statistics T_N, T_1, T, LR_{NB}, and T^2_{NB} and the bootstrap procedure BT and two non-parametric procedures WC and FP. Note that six of these T_N, T_1, T, BT, WC and FP are the same as those used for the standard BF problem. We recommend that for the smaller of n_1 and n_2 less than 20 and the other up to 30 the LR statistic, although somewhat liberal or conservative, be used. In these situations, in general, it is most powerful. However, for some extra effort, it would be advisable to use the bootstrap p-value based on this statistic. For the sample sizes stating at $n_1 = 20$ and $n_2 = 20$ (n_1 equal to or not equal to n_2) the statistics T_1, T and S all hold level reasonably well and at $n_1 = n_2 = 30$ empirical level of all these 3 procedures are very close to the nominal level. In these situations these statistics are also, in general, most powerful and therefore recommended. The practitioner can use any one of them.

For the beta-binomial BF problem we have studied seven statistics or procedures C_{BB}, T_N, T_1, T, BT, WC, and FP. For larger sample sizes (n_1 or n_2 ≥ 20) and for large π ($\geq .2$) the statistics T_1 and T are the best and therefore recommended. For small sample sizes and small π ($< .2$) we recommend to use the statistic C_{BB}. In all other situations we recommend a procedure similar to the parametric bootstrap given in example 1 in Section 4.5.

The results of the statistics T_1 and T are interesting. Even though here we are not dealing with normal data, the level properties, for large sample sizes and large π (n_1, $n_2 \geq 20$, and $\geq .2$), show to be similar to those for normally distributed data. The reason, in our opinion, is that the transformation of the discrete (binomial) data y_{ij} to continuous (proportions) data $p_{ij} = n_{ij}/y_{ij}$ does the trick in this situation.

For the Weibull BF problem also we have studied seven statistics S_w, T_N, T_1, T, BT, WC and FP. Based on extensive simulation studies we recommend that the statistic T_1 or T be used for larger sample sizes (n_1 and n_2 both larger than 25), otherwise use the bootstrap p-value or the approximate critical value of the exact distribution of the statistic based on T_1 or T.

The interesting overall finding is that the statistic T_1 or T can be used for all the cases studied here for sample sizes larger than 25 except for the beta-binomial samples in which the additional requirement is that π be large ($\geq .2$) . For smaller sample sizes, specific recommendations given above, on a case by case basis, should be followed. The statistic T_N should never be used in the BF or BF analogous problems unless the two sample sizes are very large.

It will be interesting to find through further studies whether these recommendations are applicable in other BF analogous problems, such as, testing equality of means of two gamma, extreme value and log-normal or other similar survival populations having possibly different variances. In some large sample

size situations or in sparse (beta-binomial with $\pi \leq .1$) situations for data in the form of proportions we recommended using a parametric bootstrap type procedure. Further research in this area should focus on improvements in performance, specially in terms of levels, of some of the statistics, such as the statistic C_{BB}.

For testing the equality of the scale parameters with the shape parameters being unspecified of two Weibull populations [2] develop four test statistics of which they recommend the statistics based on two different method of moments estimates of the nuisance parameters. It will be of interest to develop these later two statistics for testing $H_0 : \mu_1 = \mu_2$ with σ_1^2 and σ_2^2 being unspecified and compare with the statistics recommended in this paper.

Acknowledgements

This research was partially supported by the Natural Sciences and Engineering Research Council of Canada and the Australian Research Council Discovery Project (ARC DP130100766 and DP160104292). The paper was completed while Sudhir Paul was visiting School of Mathematical Sciences, Queensland University of Technology during January- March, 2016 and 2017.

Supplementary Material

Supplementary material for "Empirical level and Power" that includes graphs of empirical levels and tables of empirical power referred to in Sections 2, 3, 4 and 5 and derivation of the score test referred to in Section 5.2 are available as Appendix A, Appendix B, Appendix C, Appendix D, and Appendix E in https://dataverse.scholarsportal.info/dataverse/sudhirpaul. Empirical level graphs and empirical power tables for the normal BF problem are in Appendix A1 and Appendix A2 respectively. The (level graphs, power tables) for the negative binomial, beta-binomial, and the Weibull BF analogous problems are in Appendices (B1, B2), (C1, C2) and, (D1, D2) respectively.

REFERENCES

[1] Algina, J. and Oshima, T. C. and Lin, W.-Y. (1994). Type I error rates for Welchs test and Jamess second-order test under nonnormality and inequality of variance when there are two groups, *Journal of Educational Statistics* **19**, 3, 275–291.

[2] Alam, K. and Paul, S. (2015). Testing equality of scale parameters of two Weibull distributions in the presence of unequal shape parameters. *Acta Comment. Univ.*

Tartu. Math. **19**, 1, 11–26.

[3] Alam, K. and Paul, S. (2017). Testing Equality of Two Beta Binomial Proportions in the Presence of Unequal Extra-dispersion Parameters. *Comm. Statist. Simulation Comput.* **45**, 4, 1–25.

[4] Barnard, G. A. (1984). Comparing the means of two independent samples. *Applied Statistics* **33**, 266–271.

[5] Behrens W. H. V. (1929). Ein Beitrang Zur Fehlerberchnung bei wenigen beobachtungen. *Landwirtsch Jahrbucher* **68**, 807–837.

[6] Brensike, J. F., Kelsey, S. F., Passamani, E. R., Fisher, M. R., Richardson, J. M., Loh, I. K., Stone, N. J., Aldrich, R. F., Battaglini, J. W., Moriarty, D. J., Myrianthopoulos, M. B., Detre, K. M., Epstein, S. E., Levy, R. I. (1982). National Heart, Lung, and Blood Institute Type II coronary intervention study: design, methods, and baseline characteristics. *Control Clin. Trials* **3**,2 91–111.

[7] Best, D. J. and Rayner J. C. W. (1987). Welch's Approximate Solution for the Behrens-Fisher Problem. *Technometrics* **29**, 2, 205–210.

[8] Breslow, N. E. (1990). Tests of hypotheses in over-dispersed Poisson regression and other quasi-likelihood models. *J. Amer. Statist. Assoc.* **85**, 565–571.

[9] Chang, C.-H. and Pal, N. (2008). A revisit to the Behrens–Fisher problem: comparison of five test methods. *Communications in Statistics–Simulation and Computation* **37**, 6, 1064–1085.

[10] Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **30**, 248–275.

[11] Cran, G. W. (1988). Moment estimators for the 3-parameter Weibull distribution. *IEEE Trans. Rel.* **37**, 4, 360–363.

[12] Derrick, B. and Toher, D. and White, P. (2016). Why Welchs test is Type I error robust. *The Quantitative Methods in Psychology* **12**, 1, 30–38.

[13] Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the bootstrap.* Chapman and Hall, London.

[14] Fagerland, M. W. and Sandvik, L. (2009). The wilcoxon–mann–whitney test under scrutiny. *Statistics in medicine* **28**, 10, 1487–1497.

[15] Fagerland, M. W. and Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary clinical trials* **30**, 5, 490–496.

[16] Fay, M. P. and Proschan, M. A. (2010). Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics surveys* **4**, 1–39.

[17] Feigl, P. and Zelen, M.(1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics* **21**, 826–838.

[18] Feiveson, A. H. and Delaney, F. C. (1968). The Distribution and Properties of a Weighted Sum of Chi Squares. *Manned Spacecraft Center, Houston, Texas* **1**, NASA TN D–4575.

[19] Fenstad, G. V. (1983). A Comparison between the U and V tests in the Behrens-Fisher Problem. *Biometrika* **76**, 1, 300–302.

[20] Fisher, R. A. (1936). The Statistical Utilization of Multiple Measurements. *A. Eug.* **11** 377–386.

[21] Fligner, M. H. and Policello, C. C. (1981). Robust rank procedures for the Behrens-Fisher problem. *J. Amer. Statist. Assoc.* **76**, 4, 162–168.

[22] Gail, M. H., Santner, T. J. and Brown, C. C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumour. *Biometrics* **36**, 225–231.

[23] Hogg, R. V. and Tanis, E. A. (2010). *Probability and Statistical Inference* Pearson, New York.

[24] Kim, S-H. and Cohen, A. H. (1998). On the Behrens-Fisher Problem: A Review. *J. Educ. Behav. Stat.* **23**, 4, 356–377.

[25] Lehman, E. L. and D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Halden-Day: U.S.A.

[26] Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd ed. London: Chapman and Hall.

[27] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Biometrics* , 50–60.

[28] Murthy, D. N. P., Xie, M. and Jiang, R. (2004). Weibull Models. John Wiley & Sons, Inc., New Jersey.

[29] Paul, S. R. (1992). Comment on Best and Rayner (1987). *Technometrics* **34**, 2, 249–250.

[30] Paul, S. and Alam, K. (2014). Testing equality of two negative binomial means in presence of unequal over-dispersion parameters: a BehrensFisher problem analog *J. Stat. Comput. Simul.* **45**, 15, 3140–3153.

[31] Paul, S. R. and Islam, A. (1995). Analysis of proportions in the presence of over-/under-dispersion. *Biometrics* **51**, 1400–1410.

[32] Paul, S. R. and Plackett, R. L. (1978). Inference sensitivity for Poisson mixtures. *Biometrika* **65**, 3, 591–602.

[33] Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics* **46**, 863–867.

[34] Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.* **81**, 394, 321–327.

[35] Rao, J. N. K. and Scott, A. J. (1992). A simple method for the analysis of clustered binary data. *Biometrics* **48**, 577–585.

[36] Saha, K. (2013). Interval estimation of the mean difference in the analysis of over-dispersed count data. *Biom. J.* **55**, 1, 114–133.

[37] Saha, K. and Paul, S. R. (2005). Bias corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* **61**, 1, 179–185.

[38] Singh, P. and Saxena, K. K. and Srivastava, O. P. (2002). Power comparisons of solutions to the Behrens-Fisher problem. *American Journal of Mathematical and Management Sciences* **22**, 3-4, 233–250.

[39] Srivastava, M. S. and Wu, Y. (1993). Local efficiency of moment estimators in beta-binomial model. *Comm. Statist. Theory Methods* **22**, 9, 257–261.

[40] Slaton, T. L., Piegorsch, W. W. and Durham, S. D. (2000). Estimation and Testing with Overdispersed Proportions Using the Beta-Logistic Regression Model of Heckman and Willis. *Biometrics* **56**, 1, 125–133.

[41] Teimouri, M. and Gupta, A. K. (2013). On the three-parameter Weibull distribution shape parameter estimation. *J. Data Sci.*. **11**, 3, 403–414.

[42] Welch, B.L. (1938). The significance of the difference between two means when the population variances are unequal, *Biometrika*, **29**, 3/4, 350–362.

[43] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics bulletin* **1**, 6, 80–83.

[44] Williams, D. A. (1988). Extra-binomial variation in toxicology. *Proceedings of the XIVth International Workshop on Statistical Modelling* 165–174.

[45] Yin, Y. and Li, B. (2014). Analysis of the Behrens-Fisher Problem based on Bayesian Evidence. *J. Appl. Math.* **2014**.