

# Validation of Some Functions in the exact2x2 R package

Michael P. Fay, Sally Hunsberger

December 8, 2021

## Summary

This document details some of the checks that were done to ensure that the `uncondExact2x2`, `boschloo` and `binomMeld.test` functions were giving correct and reasonable results.

## 1 Introduction

Because some of the options in the functions `uncondExact2x2` and `boschloo` produce tests and confidence intervals that are not available elsewhere, there is not necessarily an existing gold standard to check each option. Nevertheless, we do check some options for which there are functions available in other R packages (most notably, the `Exact` package) or other software (SAS and StatXact).

For the `binomMeld.test` function we create two different methods to calculate the p-values and confidence intervals and compare them. See Section 10 for those tests.

## 2 The Fast Algorithm

For some options, specifically, in the case when:

```
(parmtype=="difference") & (tsmethod == "central") &
(gamma == 0) &
(method == "simple" | method=="user-fixed" | method=="FisherAdj")
```

Then the confidence intervals can be calculated using a faster algorithm, since we have monotonicity in the one-sided p-values. We test this by checking that the confidence intervals and p-values have unified inferences. Specifically, we calculate the 95% confidence interval, then check the p-value for the test with `nullparm` equal to one limit of the interval and see if the p-value is equal to 0.05. The check for `method=="FisherAdj"` is in `\tests\testthat\test_FisherAdj.R`.

## 3 FisherAdj: The Slow Algorithm

For `parmtype="ratio"` or `parmtype="oddsratio"` with `method=="FisherAdj"` we must use the slower algorithm. We checked `method=="FisherAdj"` for whether the confidence intervals and p-values have unified inferences in `\tests\testthat\test_FisherAdj.R`.

## 4 Boschloo Test

The Boschloo test is run with the `boschloo` function. There is a separate function because unlike most unconditional exact tests, the traditional way of calculating the `alternative="less"` option uses a different ordering function than calculating the `alternative="greater"` option. The `boschloo` uses the appropriate one-sided p-value from the Fisher's exact test for ordering. We validate the p-values by testing them against the `Exact` R package and StatXact Version 11. They match for the data that we tested. The tests on the p-values are in `\tests\testthat\test_Exact.R` and `\inst\slowTests\test_StatXact.R`. Note that the default for `alternative="two.sided"` in our `boschloo` function is `tsmethod="central"`, but the traditional two-sided p-value for Boschloo's test uses `tsmethod="minlike"`. Thus, we only tested the latter. There is no software of which we are aware that calculates the confidence intervals that match Boschloo's test.

## 5 Confidence Intervals on Odds Ratios with method="score"

Agresti and Min (2002) discussed two ways of calculating the exact unconditional confidence intervals for the odds ratio using the score statistic for ordering. The two ways correspond to the two options in `tsmethod`. But the function `uncondExact2x2` with `parmtpe="oddsratio"` and `method="score"` does not match the results in the paper for either `tsmethod`, even though we used a more precise grid than the default (`control$nPgrid=1000`). This could be due to using different algorithms for the calculation. Although Agresti and Min (2002) state that "There is no assurance that an algorithm such as just described produces the correct interval", the same can be said of the algorithms for `uncondExact2x2`. So it is not clear which software is closer to the true values. The tests are in `\inst\slowTests\test_OddsRatioScore.R`.

## 6 The E+M Method

We check the E+M method of Lloyd (2008) by reproducing the some results on the example in Table 1 of that paper. We reproduce the p-values using the Wald-pooled difference example (T rows: top is one-sided, bottom is two-sided) and the E+M column (results are  $p=0.02518$  [one-sided] and  $p=0.03681$  [two-sided]). This test is in `\inst\slowTests\test_EplusM.R`.

## 7 Berger and Boos Adjustment

Berger and Boos (1994) developed a method to adjust for nuisance parameters that tends to have better power when applied to the 2x2 unconditional exact tests. We calculated their adjustment for the data example in that paper, and the results agree between our `uncondExact2x2` and the `exact.test` from the Exact R package. Both these functions differ by a small amount from the value given in the paper (in the paper,  $p=0.037$ , while the functions give  $p=0.03781$ ). The test is in `\tests\testthat\test_Exact.R`.

The `uncondExact2x2` results were fairly close for the examples we checked in StatXact 11 (difference in log values within 0.02). See `\inst\slowTests\test_StatXact.R`.

## 8 Comparison with StatXact

We ran a comparison with StatXact version 11. The tests are in `\inst\slowTests\test_StatXact.R`. We tested the `parmtpe="difference"` and `parmtpe="ratio"` with both `tsmethod` options with the `method="score"`. This is associated with either "invert two one-sided tests" (for `tsmethod="central"`) or "invert a two-sided test" (for `tsmethod="square"`). We found that in most cases with `tsmethod="central"` the agreement with "invert two one-sided tests" was close. But we found some big differences when `tsmethod="square"` and "invert a two-sided test".

We mention one case where we can show clearly that the StatXact 11 is wrong (in the other cases, it is not as clear which software is more accurate). The case is  $x_1/n_1 = 0/7$  and  $x_2/n_2 = 13/13$ . StatXact 11 using "invert a two-sided test" and `Gamma=0` gives a 95% confidence interval on that difference as (0.6088, 0.6101). This is wrong, because when  $\theta_1 = 0$  and  $\theta_2 = 1$  then the true differences is 1, and the probability of observing  $x_1/n_1 = 0/7$  and  $x_2/n_2 = 13/13$  is also 1. So the actual coverage under that assumption is 0, and the confidence interval is clearly not exact. For  $\theta_1 = \epsilon$  and  $\theta_2 = 1 - \epsilon$  with  $\epsilon$  a very small positive value, we would similarly get coverage close to 0. In contrast, the `uncondExact2x2` function with `parmtpe="difference"`, `tsmethod="square"`, and `gamma=0` has the 95% confidence interval as (0.6088, 1). In this case the coverage is 100% (assuming the CI includes 1) when  $\theta_1 = 0$  and  $\theta_2 = 1$ .

Comparison of Boschloo values has previously been mentioned in Section 4.

## 9 Comparison with SAS

We check some results and compared them with SAS version 9.4. Proc Freq has the Exact option, that gives p-values for the Barnard test (`uncondExact2x2` test of  $\theta_1 = \theta_2$  with `method="wald-pooled"` and `parm-`

type="difference"). The Exact option also gives confidence intervals for the ratio and difference (uncondExact2x2 with method="score").

The methods agree, and the tests are in `\inst\slowTests\test_SAS.R`. The default control value for `control$nPgrid` option of the algorithm in `uncondExact2x2` must be increased for a few of the values to match. This suggests that the default for SAS is more accurate than the default of `control$nPgrid=100`.

## 10 The `binomMeld.test` function

The option `nmc` in the `binomMeld.test` function allows calculation by Monte Carlo simulation. That provides a separate check for the numeric integration calculations. We compare the two calculation methods in `\inst\slowTests\test_binomMeldtest.R`.

## References

- Agresti, A., and Min, Y. (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics*, 3(3), 379-386.
- Berger, R. L., and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427), 1012-1016.
- Lloyd, C (2008). Exact p-values for discrete models obtained by estimation and maximization. *Australian & New Zealand Journal of Statistics*, 50(4): 329-345.