

Exploration of the Behrens distribution and the Welsh approximation with regard to the two sample T test

Introduction

Calculating a confidence interval or credible interval for the differences in means of two normally distributed samples is one of the more fundamental questions asked of statistics. However in text books it is nearly always assumed that these two sample problems have equal variance, so that C.I. can be calculated with ubiquitous Student's T distribution. This is rarely the state of affairs in the real world. The common approach taught in numerous textbooks, is the use of the Welsh approximation, which tries to approximate the effects of unequal variance by reducing the degrees of freedom to some usually non integer value.

The Behrens distribution, described in 1929, attempts to define the necessary two sample unequal variance distribution exactly, but even now only approximations exist for calculating densities, percentiles and quantiles.

Another possible problem is that reference priors do not exist for the Behrens-Fisher problem that would lead to solutions that would agree exactly with a Classical statistician's re-sampling perspective.

Note About How Behrens Distribution Is Calculated.

I have found the R code presented in Lee 2003 to be very slow and can give very unstable results. Instead an R package called "tdist" was used, which could efficiently calculate linear combinations of t random variables.

The B-F distribution is distributed as...

$$BF(df_1, df_2, \theta) \sim T_2 \cos(\theta) - T_1 \sin(\theta) \quad (1) \text{ [page 145 of Lee]}$$

where

$$\theta = \arctan((s_1/\sqrt{n_1}) * (s_2/\sqrt{n_2}))$$

$$T_1 \sim T(df = n_1 - 1)$$

$$T_2 \sim T(df = n_2 - 1)$$

θ can loosely be interpreted as a relative coefficient of variability between the two samples. $\theta = 45$ degrees and $n_1 = n_2$ implies that that two samples have equal variance and equal sample sizes, and thus the T distribution is a special case of the B-F distribution.

Thus the linear coefficients for the tdist package would be $[\cos(\theta), -\sin(\theta)]$.

Using The Behrens Distribution With Substantial Prior Information

Prior information for each sample can be specified in terms of v_0 , n_0 , θ_0 , and S_0 . Formulas from the top of Lee page 69 should be helpful.

$$\begin{aligned}v_1 &= v_0 + n \\n_1 &= n_0 + n \\ \theta_1 &= (n_0\theta_0 + n\bar{x}) / n_1 \\ S_1 &= S_0 + S + (n_0^{-1} + n^{-1})^{-1}(\theta_0 - \bar{x})^2 \\ s/\sqrt{n_1} &= \sqrt{S_1/v_1 n_1}\end{aligned}$$

$$\begin{aligned}\frac{\theta - \theta_1}{s/\sqrt{n_1}} &\sim t_{v_1} \\ \phi &\sim S_1 \chi_{v_1}^{-2}\end{aligned}$$

Note that the reference prior implies $v_0 = -1$, $n_0 = 0$, $S_0 = 0$ (θ_0 need not be defined). Also be aware that θ represents two different things, it was used in section 2.13 for specifying prior information, and in section 5.3 as a parameter for the Behrens-Fisher distribution.

To see how this prior information could be used in practice, observe that with a reference prior the statistic would be

$$BF \left(df_1 = n_x, df_2 = n_y, \theta = \arctan \left(\frac{s_x / \sqrt{n_x}}{s_y / \sqrt{n_y}} \right) \right)$$

While when incorporating prior information (using the formulas above) the statistic would be

$$BF \left(df_1 = v_{x0} + n_x, df_2 = v_{y0} + n_y, \theta = \arctan \left(\frac{\sqrt{\frac{S_{x0} + S_x + (n_{x0}^{-1} + n_x^{-1})^{-1}(\theta_{x0} - \bar{x})^2}{(v_{x0} + n_x)(n_{x0} + n_x)}}}{\sqrt{\frac{S_{y0} + S_y + (n_{y0}^{-1} + n_y^{-1})^{-1}(\theta_{y0} - \bar{y})^2}{(v_{y0} + n_y)(n_{y0} + n_y)}}} \right) \right)$$

WinBugs Example

I tried to extend the one sample test with mean and variance unknown for the two sample test with mean and variance unknown but was unsuccessful. Also it does not appear

necessary for this problem. It is easy enough to generate random variables from the BF distribution in R using the fact that $BF(df_1, df_2, \theta) \sim T_2 \cos(\theta) - T_1 \sin(\theta)$.

Sample Data

Data from Lee 5.7 problem 8, page 155, lengths of cuckoo bird's eggs.

```
>X = c(22.0, 23.9, 20.9, 23.8, 25.0, 24.0, 21.7, 23.8, 22.8, 23.1)
>Y = c(23.2, 22.0, 22.2, 21.2, 21.6, 21.9, 22.0, 22.9, 22.8)
```

Comparison With Welsh Approximation

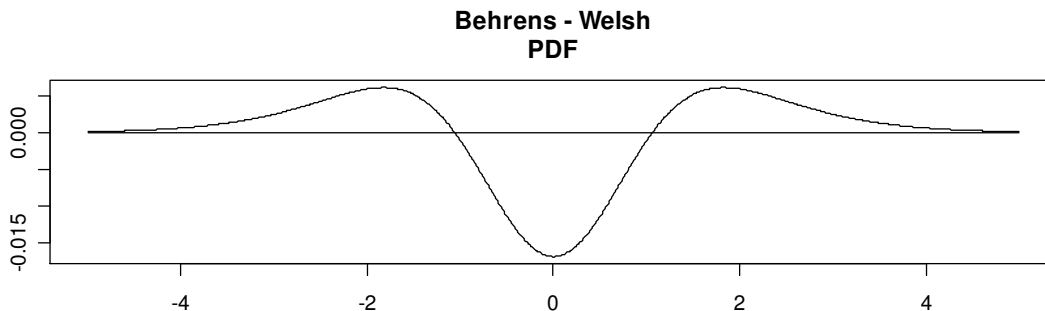
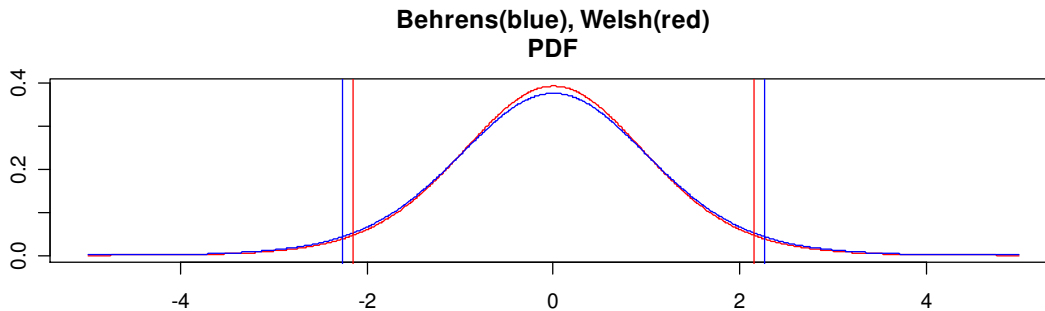
```
>t.test(X,Y)
>bf.test(X,Y)
>>null_diff( mu.x=mean(X), mu.y=mean(Y), s.x=sd(X), s.y=sd(Y),
            n.x=length(X), n.y=length(Y) )
```

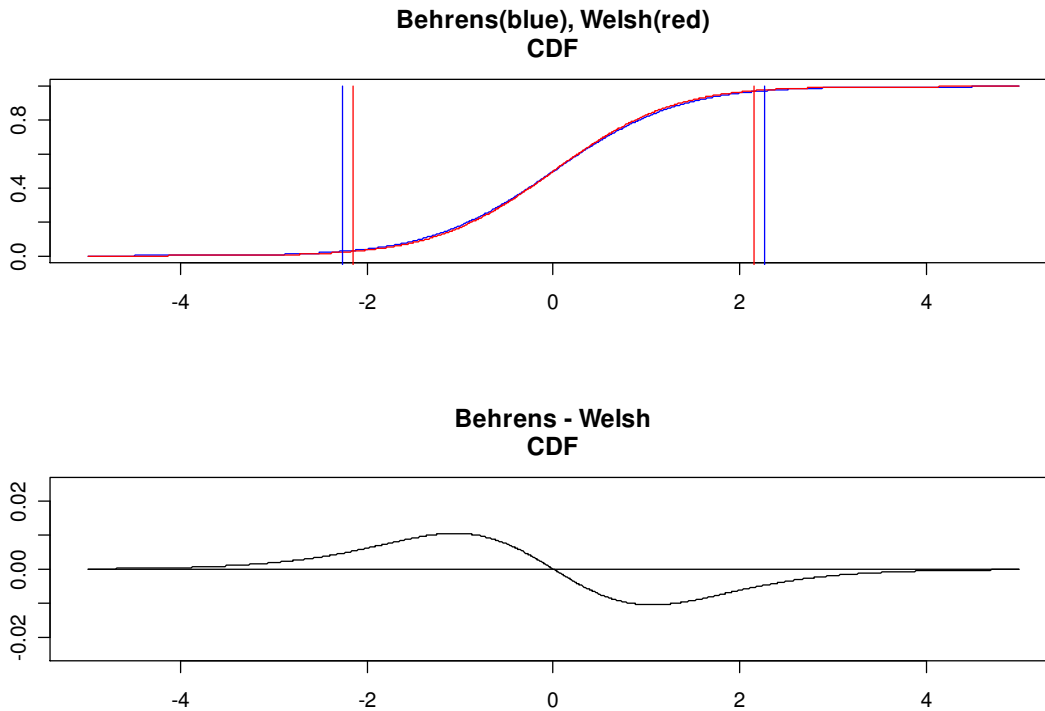
Welsh

```
t = 1.9924, df = 13.796, p-value = 0.0665
95 percent confidence interval: [ -0.07019846  1.87019846 ]
```

BF

```
df1 = 9, df2 = 8, theta = 61.33793 degrees, p-value = 0.07906503
95 percent confidence interval: [ -0.1212955  1.9212955 ]
```





The differences can be hard to see when looking at the superimposed PDFs, but the second figure shows the difference (ie the “Behrens pdf” – the “welsh t pdf”). The next two graphs show the same thing but compares CDFs instead. What is important to note, is that regardless of what significance level you choose, the Welsh T will give you larger C.I.s. This is easiest to see in the last figure because for quantiles less than zero, the Behrens distribution will give you a larger p value, while for quantiles larger than zero the opposite effect is observed.

Substantial Prior Information Example

Using the same data as in the previous example, suppose some prior information was available, and we will use the same information for both samples.

$v_0=20$
 $n_0=5$
 $S_0=40$
 $\theta_0=22$

Thus the sample means should be about 22mm and are worth perhaps 5 observations, and the variance is expected to be approximately 2.2 with standard deviation of 0.8.

$$BF \left(df_1 = v_{x_0} + n_x, df_2 = v_{y_0} + n_y, \theta = \arctan \left(\frac{\sqrt{\frac{S_{x_0} + S_x + (n_{x_0}^{-1} + n_x^{-1})^{-1} (\theta_{x_0} - \bar{x})^2}{(v_{x_0} + n_x)(n_{x_0} + n_x)}}}{\sqrt{\frac{S_{y_0} + S_y + (n_{y_0}^{-1} + n_y^{-1})^{-1} (\theta_{y_0} - \bar{y})^2}{(v_{y_0} + n_y)(n_{y_0} + n_y)}}}} \right) \right)$$

Df1=15, df2=14,

```
> atan( sqrt(40+sd(X)*9 + 1/(1/5+1/10)*(22-mean(X))^2)
+ /sqrt(40+sd(Y)*8 + 1/(1/5+1/9)*(22-mean(Y))^2)
+ )*180/pi
[1] 47.84711
```

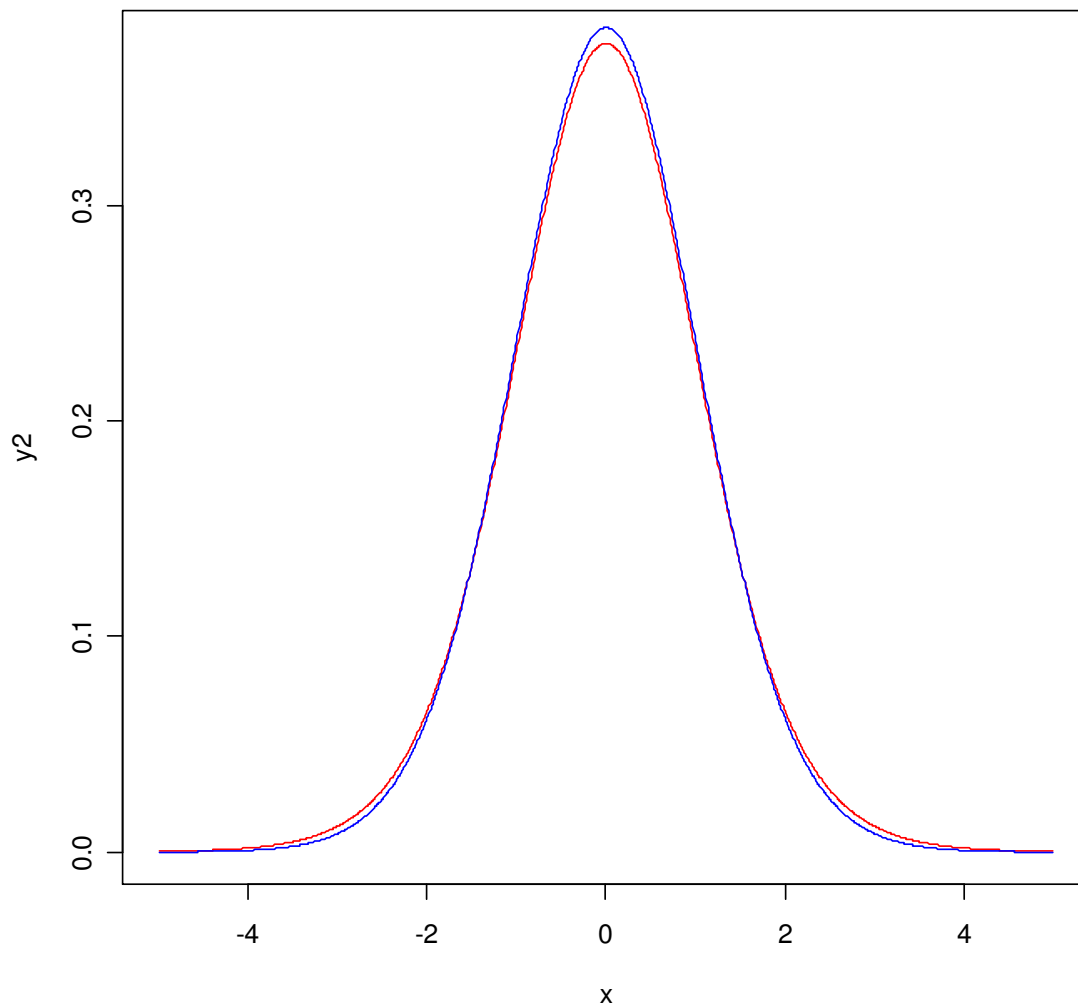
Thus the distribution would be BF(15, 14, 47.84711). This is slightly narrower distribution due the increase in the degrees of freedom for the two samples. Using identical prior information has also biased theta toward 45 degrees; the value at which the BF distribution behaves like a pooled two sample t test.

```
> pbf(.975, 9, 8, 61.33793) # reference prior
      1
0.8163274
> pbf(.975, 15, 14, 47.84711) # substantial prior
      1
0.8228224
```

However at the tails it turns out the increase in the degrees of freedom did not counter act the effect of driving theta toward 45 degrees, and in fact the substantial prior information has made the C.I. wider.

```
> y1=y2=x=seq(-5, 5, 0.01)
> for( i in 1:length(x) ){
+ y1[i]=dbf(x[i], 15, 14, 47.84711)
+ y2[i]=dbf(x[i], 9, 8, 61.33793) }
> plot(x, y2, main="POSTERIOR\nred = reference, blue = substantial
prior", type="l", col="red")
> lines(x, y1, type="l", col="blue")
```

POSTERIOR
red = reference, blue = substantial prior



Conclusion

The Welsh Approximation has always been a “fudge” that was needed when models had to be simplified such that tables of t distributions could be used instead of using a computer to calculate or precisely estimate these things. With the prevalence of computer packages like R for calculating pvalues instead of looking them up in a distribution table, the simplifications of the Welsh Approximation looks like a very dated technique. More stable answers are now obtainable with the use of the Behrens distribution.

Source for the functions I created available at <http://jbcovlin.fastmail.fm/behrens/>,
Namely project.R and bf.R.