Alternative effect size measurements corresponding to the two sample t-test

Andreas Kitsche

2017-11-21

If interest is in comparing the means of two (normally distributed) samples it is common practise to perform a two-sample t-test and report the corresponding p-value. Nevertheless, it has been widely criticized that the p-value does not provide a measure for the magnitude of the mean effect (e.g., Browne (2010)). This report provides an overview of existing alternatives recently published in the scientific literature that provide a more meaningful measurement of the effect size. Browne (2010) introduced closed form equations to translate a significant t-test p value and sample size into the probability of one treatment being more successful than another on a per individual basis $P(X^* > Y^*)$. This term was afterwards denoted as win probability by Hayter (2013) and he demonstrated the interpretation as "*what would happen if a single future observation were to be taken from either of the two treatments, with attention being directed towards which treatment would win by providing the better value.*" In addition Hayter (2013) introduced the corresponding confidence interval as well as the odds of X being greater than Y. He further introduced the transformation into Cohens effect size and the corresponding confidence intervals.

Example

In the depression trial, two groups of patients, one treatment (D) and one placebo (P) group, were compared Dmitrienko et al. (2005). The primary endpoint was the change from the baseline to the end of the 9-week acute treatment phase in the 17-item Hamilton depression rating scale total score (HAMD17 score). The scores range from -2 to 28, and therefore, we assume that this end-point is approximately normally distributed.

The goal is now to calculate different effect size measurements for the comparison of the treatment and the placebo group. Therefore we choose the function `WinPropRaw` from the R add on package `WinProp`.

```
#library(devtools)
#install_github("AKitsche/WinProp")
library(WinProp)
#install_github("AKitsche/poco")
library(poco)
library(dplyr)
data(Depression)
x <- Depression %>% filter(Group=="P") %>%  select(Score)
y <- Depression %>% filter(Group=="D") %>%  select(Score)
fm1 <- lm(Score ~  Group,data=Depression)
predict(fm1, interval=c("prediction"))
WinPropRaw(x=x$Score, y=y$Score, alpha=0.05, beta=0.95, var.equal=TRUE, alternative="two.sided")
#see also:
#t.test(x=x$Score, y=y$Score, var.equal=TRUE)
```

From the p-value of 8.2×10^{-8} of the two-sample t-test we can reject the null hypothesis of no treatment effect and conclude a significant treatment effect through the significance level α=0.05. A statistically significant difference between the sample means is also concluded from the confidence interval, since it does not contain 0.

As noted by Hayter (2013) there are arguments against decision making based on the evaluation of the confidence intervals: "*Criticism of this approach has centred on the fact that statistical significance does not necessarily equate to a meaningful difference in practice between the treatment means. In other words, even*

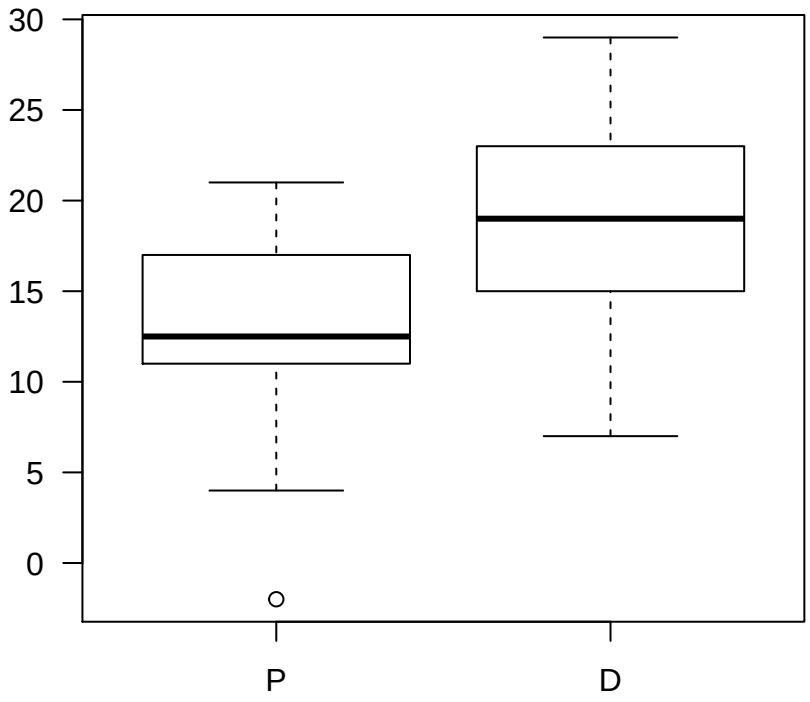
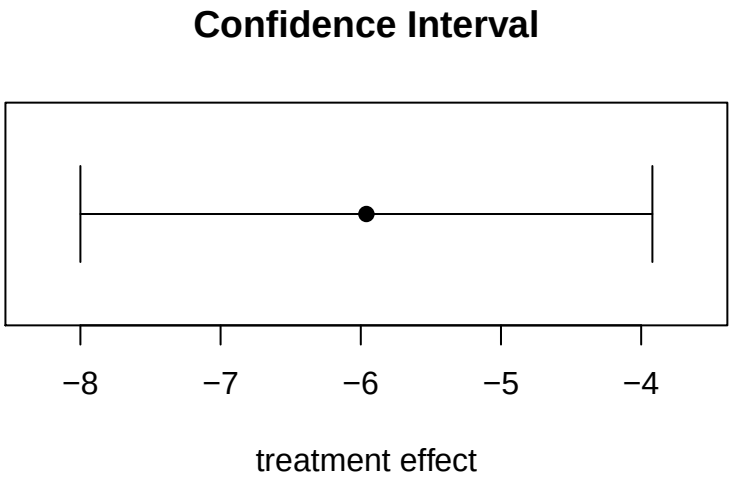


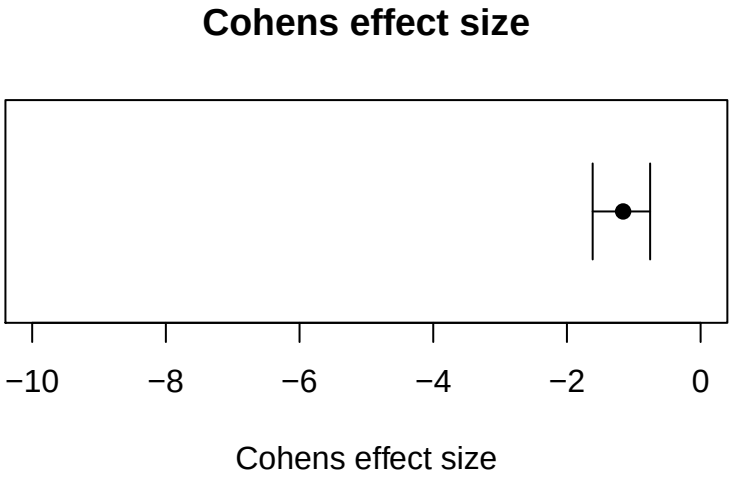
Figure 1: In the depression trial, two groups of patients, one treatment and one placebo group, were compared. The primary endpoint was the change from the baseline to the end of the 9-week acute treatment phase in the 17-item Hamilton depression rating scale total score (HAMD17 score).

though the difference between the two treatment means may be statistically significant, the actual magnitude of the difference may not be particularly important,. . . ."



Inferences based on the standardized mean difference $(\mu_1 - \mu_2)/\sigma$ (Cohens effect size) provide a

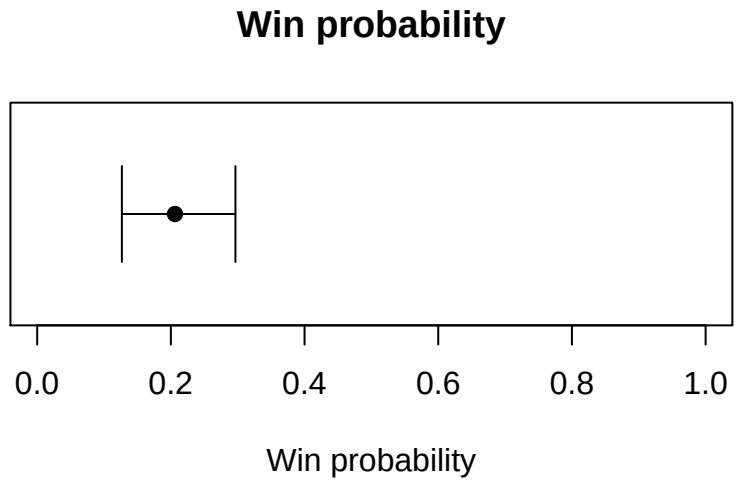
The value of Cohens effect size is -1.16 with an interval of -1.62--0.76



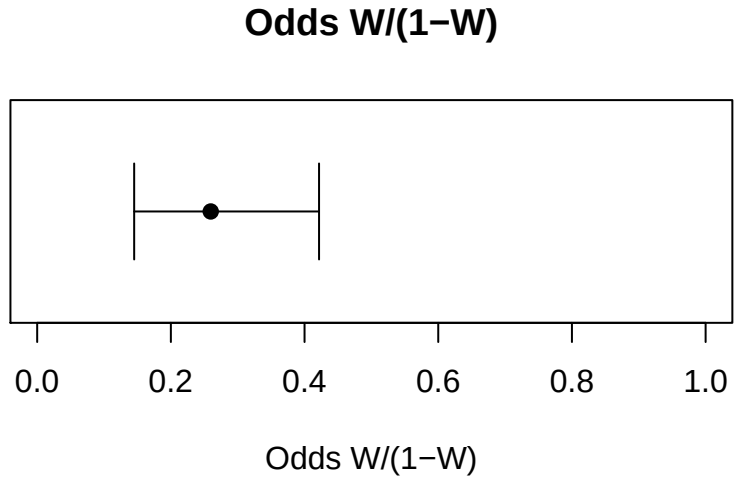
Browne (2010) proposed that inference about potential future observations from the treatments provides valuable assistance to the problem of choosing between the two treatments. Hayter (2013) introduced inferences about the difference of potential future observations X_D^* and X_P^*. He definded Win-probability $P(X_D^* > X_P^*)$ which allows a practitioner to consider what would happen if a single future observation were to be taken from either of the two treatments, with attention being directed towards which treatment would win by providing the better value.

The Win-probability introduced by Hayter (2013) directly adresses the question of whether or not an individual´s choice of treatment may have an effect on its hamilton depression score. The 95% confidence interval for the Win-probability is $P(X_D^* > X_P^*) \in (0.13,0.3)$, and since this probability my be less than 50%, it can be guaranteed that it is more likely that the dose treatment will be effective. Furthermore, this

confidence interval also provides the additional pertinent information that there may be as much as an 12.67% chance that the dose treatment will improve an individuals hamilton depression score.



In addition to the Win-probabilities, Browne (2010) introduced the odds of Win-probabilities $W/(1-W)$. The 95% lower and upper confidence limit for the odds are given by 0.15 and 0.42

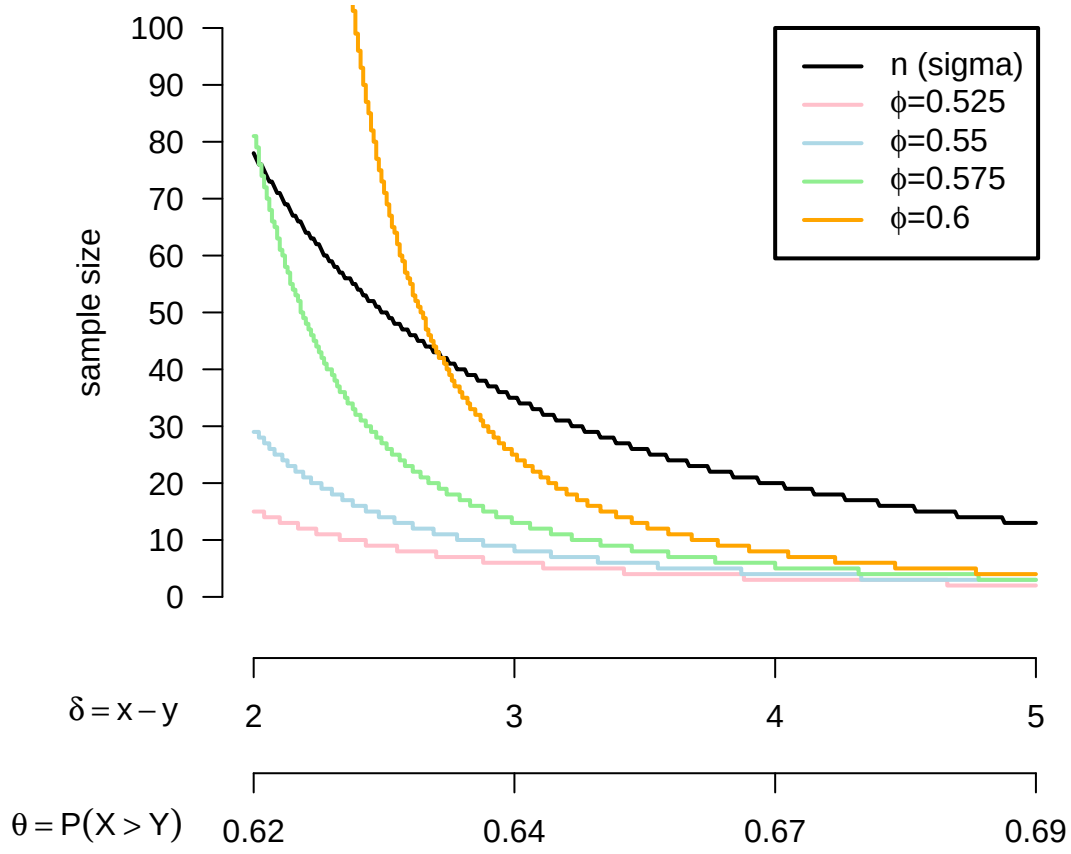


Sample size calculation to demonstrate clinical relevance

Kieser, Friede, and Gondan (2013) presented a new approach for the assessment of clinical relevance based on the so-called relative effect (or probabilistic index), which corresponds to the win probability. In addition, they provide a sample size formula based on the approximate sample size formula for the t-test, to observe a relative effect ϕ . As noted by Kieser, Friede, and Gondan (2013) $n_{sig+rel} = max(n_{sig}, n_{rel})$ gives the sample size per group required for the simultaneous proof of statistical significance and clinical relevance. The function `Nrel` provides the functionality to calculate those sample sizes.

In the following we demonstrate the relationship between the required sample size per group and the difference

in expectation. It has to be noted that the estimated relative effect θ is linked to the difference in expectation δ via $\theta = \Phi((\delta/\sigma)/\sqrt{2})$, where $\Phi()$ is the standard normal distribution function. We consider a standard deviation of $\sigma = 5$ and difference in expectation from 2 to 5 with increments of 0.01 for differing values of ϕ (0.525, 0.55, 0.575 and 0.6).



References

Browne, Richard H. 2010. "The t -Test p Value and Its Relationship to the Effect Size and P (X > Y)." *The American Statistician* 64 (1): 30–33. doi:10.1198/tast.2010.08261.

Dmitrienko, A, G Molenberghs, C Chuang-Stein, and W Offen. 2005. *Analysis of Clinical Trials using SAS: A Practical Guide*. Cary, NC, SAS Institute Inc.

Hayter, AJ. 2013. "Inferences on the difference between future observations for comparing two treatments." *Journal of Applied Statistics* 40 (4): 887–900. http://www.tandfonline.com/doi/abs/10.1080/02664763.2012.758245.

Kieser, Meinhard, Tim Friede, and Matthias Gondan. 2013. "Assessment of statistical significance and clinical relevance." *Statistics in Medicine* 32 (10): 1707–19. doi:10.1002/sim.5634.