Kahane et al. example: correlation, power-based bounds

Equivalence bounds based on power

In a paper published in 2015, Kahane and colleagues investigate moral dilemma vignettes in which participants have to decide whether or not they would sacrifice one person's life to save several other lives, or judge how morally admissible such actions are. Traditionally, greater endorsement for sacrificing a life to save others has been interpreted as a more "utilitarian" moral orientation, i.e. a stronger concern for the greater good (in total, fewer people lose their lives). In a number of studies, Kahane et al. contest this interpretation, for example by showing that greater endorsement for such "utilitarian" choices correlates with sub-clinical psychopathy, and, crucially, by comparing the traditionally used vignettes with a set of new ones that pit partial motivations against an impartial concern for the greater good (e.g. buying a new mobile phone vs. donating the money to save lives in a distant country, study 4).

Kahane et al. find no significant correlation between the perceived wrongness of a utilitarian choice in the classical sacrificial dilemmas and the new "greater good" dilemmas, $r(229) = -0.04$, $p = 0.525$ ($N = 231$[1]). They conclude that the classical vignettes fail to capture "true"" utilitarianism, but also grant that this conclusion is dependent on the power of their study and that better-powered future studies might overturn their verdict: "Thus, while we cannot rule out the possibility that such an association could emerge in future studies using an even larger number of subjects or different measures, we submit that, in light of the present results, a robust association between 'utilitarian' judgment and genuine concern for the greater good seems extremely unlikely." (p. 206). This inference — that the result would be surprising if there was a true effect of a size the study could have detected — can be formalized with an equivalence test for correlations and bounds set to an effect size the study had reasonable power to detect: With 231 participants, the study had 80% power to detect effects of $r = 0.18$. Given bounds of $\Delta_L = -0.18$ and $\Delta_U = 0.18$, we find that the result is indeed statistically equivalent, $r(229) = -0.04$, $p = 0.015$. This means that values smaller than $r = -0.18$ and larger than $r = 0.18$ can be rejected at an alpha level of 5% and used as an initial benchmark which future studies could challenge using larger samples and more narrow equivalence bounds.

[1]Study 4 in Kahane et al. (2015) had a final sample size of $N = 232$, but due to missing data in 1 case, the correlation reported here is based on a sample of only $N = 231$.