Corrigendum to “Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate”∗

Justin Esarey
Wake Forest University
Department of Politics and International Affairs
justin@justinesarey.com

Jane L. Sumner
University of Minnesota
Department of Political Science
jlsumner@umn.edu

June 17, 2019

After publication of Esarey and Sumner (2018), we recognized substantively significant errors in the paper. We correct those errors here. We discovered these errors as part of work on another, unrelated paper to which one co-author hoped to apply similar methods. To summarize, the section titled “Underconfidence is possible for conjoint tests of theoretical predictions” and the subsections “Underconfidence corrections for estimated marginal effects” and “Suggestion 3: specify theories with multiple predictions in advance and used bootstrapped critical t statistics to maximize empirical power,” including Tables 3 and 5, are incorrect. Our procedure assumes a point null hypothesis, but the appropriate null hypothesis for the joint test does not correspond to this point null. The related “Prediction-corrected 90% Confidence Interval” in Figure 2 is based on this erroneous procedure and should be ignored. Finally, the procedure we suggest on pp. 1161-1163 to correct for overconfidence based on Benjamini and Hochberg (1995) is subject to several limitations unstated in the paper; this corrigendum lays out those limitations and adds a more robust procedure to our interactiTest software.

∗Thanks to William D. Berry and Carlisle Rainey for commenting on an earlier version of this corrigendum.
Details

First, the section of the paper entitled “Underconfidence is Possible for Conjoint Tests of Theoretical Predictions” is incorrect. Contrary to our argument, the Brambor, Clark and Golder (2006) method produces confidence intervals associated with accurate significance tests when jointly testing multiple hypotheses. However, it is important to emphasize that the Brambor, Clark and Golder (2006) procedure is still overconfident when separately testing multiple hypotheses. But in situations where a previously specified theory makes multiple predictions for $\frac{\partial y}{\partial x}|z$ at different values of $z$ for a linear model:

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz$$

(1)

then testing these hypotheses using $(1 - \alpha)$ confidence intervals generated using the Brambor, Clark and Golder (2006) method will jointly reject all the null hypotheses at most $\alpha$ proportion of the time. For example, if $z \in \{0, 1\}$, and a researcher pre-specifies alternative hypotheses that $(\partial y/\partial x|z = 0) < 0$ and $(\partial y/\partial x|z = 1) > 0$, separate $t$-tests rejecting each null separately using $t$-tests with size $\alpha$ will jointly reject both nulls at most $\alpha$ proportion of the time. This is discussed and proved in Silvapulle and Sen (2005, Section 5.3), especially in proposition 5.3.1, who in turn cite (inter alia) Lehmann (1952); Berger (1982); Cohen; Gatsonis and Marden (1983); and Berger (1997). It is also discussed in Casella and Berger (2002, Section 8.2.3 and 8.3.3).

The problem lies in the appropriate null hypothesis corresponding to a joint hypothesis test. Let $(\partial y/\partial x|z = z_0)$ be abbreviated as $ME_{x}^{z_0}$. Esarey and Sumner (2018, p. 1157) states
that:

\[
\sup \Pr(\text{false positive} | ME_x^0 \leq 0 \lor ME_x^1 \geq 0) \\
= \Pr \left( \left( \hat{ME}_x^0 \text{ is stat. sig. and } > 0 | ME_x^0 = 0 \right) \land \left( \hat{ME}_x^1 \text{ is stat. sig. and } < 0 | ME_x^1 = 0 \right) \right) \\
= \Pr (\hat{ME}_x^0 \text{ is stat. sig. and } > 0 | ME_x^0 = 0) \cdot \Pr (\hat{ME}_x^1 \text{ is stat. sig. and } < 0 | ME_x^1 = 0) \\
= \alpha^2 = 0.05^2 = 0.0025
\]

This calculation would be correct if \( ME_x^0 = 0 \) and \( ME_x^1 = 0 \), a point null hypothesis. However, the null hypothesis space stated in the problem includes the possibility (for example) that \( ME_x^0 \) is large but \( ME_x^1 = 0 \). Consequently, \( \sup \Pr(\text{false positive} | ME_x^0 \leq 0 \lor ME_x^1 \geq 0) \) can be as high as the \( \alpha \) value of any of the individual tests (in this case, 0.05), as stated in Silvapulle and Sen (2005) and Casella and Berger (2002). All the quantities in Table 3 of Esarey and Sumner (2018) are based on similar miscalculations.

Second, the subsection of Esarey and Sumner (2018) titled “Underconfidence corrections for estimated marginal effects” and related material in the subsection titled “Specify theories with multiple predictions in advance and use bootstrapped critical-t statistics to maximize empirical power” is incorrect. In particular, the bootstrapped critical-t statistics reported in Table 5 of Esarey and Sumner (2018) should not be used. Those statistics correspond (for non-zero predictions) to \( t \)-values that occur 5% of the time or less in the situation where all marginal effects are equal to zero, the point null hypothesis discussed above. Table 5 in this paper confirms that the procedure works, but for the point null hypothesis for which all marginal effects are simultaneously equal to zero. This is not the appropriate null hypothesis space, as discussed above. We have removed the corresponding function for calculating these \( t \)-statistics from our \texttt{R} package (while leaving in the function to calculate a \( t \)-statistic corresponding to a given false discovery rate).

Third, in our reanalysis of Clark and Golder (2006), the appropriate confidence intervals
in Figure 2 for a joint test of the proposed hypotheses is the original 90% confidence interval reported by Clark and Golder (2006). The “prediction-corrected” 90% confidence interval should be ignored, as it uses the procedures above.

Fourth, the portions of Esarey and Sumner (2018) related to overconfidence of confidence intervals described in Brambor, Clark and Golder (2006) in situations where individual marginal effects are being tested and reported, and the related FDR and FWER-controlling corrections, are (to our knowledge) correct but subject to several qualifications unstated in the paper. Contrary to a statement on p. 1162 of Esarey and Sumner (2018), the Benjamini and Hochberg (1995) procedure is only formally proved to control the false discovery rate under independence of test statistics or positive regression dependency in the subset of true null hypotheses (PRDS) (Benjamini and Yekutieli, 2001). Our simulation evidence (see also Reiner-Benaim, 2007) appears to indicate that the procedure works adequately in the regression interaction context, possibly because $t$-statistics for a two-sided test are PRDS in the situations that our simulations cover or possibly because the procedure is robust to situations outside what has been formally proved. We have added an option to our interactionTest software package to allow use of an alternative procedure presented by Benjamini and Yekutieli (2001) that is robust to any correlation among $t$-statistics, but is more conservative than the Benjamini and Hochberg (1995) procedure.

Finally, the properties of confidence intervals constructed using the Benjamini and Hochberg (1995) procedure in our paper (on p. 1162) are subject to special properties unstated in our paper but discussed in Benjamini and Yekutieli (2003, p. 72). In particular, these confidence intervals are formally proved to control the false coverage rate (i.e., “the expected proportion of parameters not covered by their [confidence intervals] among the selected parameters [statistically significant effects under the Benjamini and Hochberg (1995) procedure”), not necessarily the overall false coverage probability for all parameters (i.e., those where the null hypothesis is not rejected), when the test statistics are independent (see also Benjamini,
Confidence intervals created using the Benjamini and Yekutieli (2001) procedure possess this property under any relationship among the test statistics.

References


