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

February 28, 2020

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. The Brambor, Clark and Golder (2006) procedure is overconfident when separately testing multiple hypotheses, as correctly stated in our paper. However, when *conjointly* testing multiple hypotheses, the Brambor, Clark and Golder (2006) procedure is appropriate. Our analysis and some alternative procedures we suggested for this case assumed a joint point null hypothesis which is inappropriate for a simultaneous test of multiple hypotheses. For this reason, the sections 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. The related "Prediction-corrected 90% Confidence Interval" in Figure 2 is based on this erroneous procedure and should be ignored. Finally, we believe that the procedure we suggest on pp. 1161-1163 to correct for overconfidence in the case of separate testing

*Thanks to William D. Berry, Carlisle Rainey, and the editors for commenting on an earlier version of this corrigendum.

of multiple hypothesis (based on Benjamini and Hochberg (1995)) is correct but subject to several limitations unstated in the paper; this corrigendum lays out those limitations and adds a more robust procedure to our `interactionTest` software. A version of our paper that strikes through errors is available at https://doi.org/10.7910/DVN/YMFJMM.

Details

The Brambor, Clark and Golder (2006) procedure is overconfident when separately testing multiple hypotheses, as correctly stated in our paper. However, the Brambor, Clark and Golder (2006) method produces confidence intervals associated with accurate significance tests when *conjointly* testing multiple hypotheses. This negates arguments in the section of the paper entitled "Underconfidence is Possible for Conjoint Tests of Theoretical Predictions." Contrary to our argument, in situations where a previously specified theory makes multiple predictions for $(\partial y/\partial x|z)$ at different values of z for a linear model:

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz \tag{1}$$

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 α 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 α will jointly reject both nulls at most α 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 root problem in our analysis 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:

$$\begin{aligned}
& \sup \Pr(\text{false positive}|ME_x^0 \leq 0 \vee ME_x^1 \geq 0) \\
= \; & \Pr\left[\left(\widehat{ME}_x^0 \text{ is stat. sig. and } > 0|ME_x^0 = 0\right) \wedge \left(\widehat{ME}_x^1 \text{ is stat. sig. and } < 0|ME_x^1 = 0\right)\right] \\
= \; & \Pr\left(\widehat{ME}_x^0 \text{ is stat. sig. and } > 0|ME_x^0 = 0\right) * \Pr\left(\widehat{ME}_x^1 \text{ is stat. sig. and } < 0|ME_x^1 = 0\right) \\
= \; & \alpha^2 = 0.05^2 = 0.0025
\end{aligned}$$

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, $\sup \Pr(\text{false positive}|ME_x^0 \leq 0 \vee ME_x^1 \geq 0)$ can be as high as the α 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.

For this reason, 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 `R` package (while leaving in the function to calculate a

t-statistic corresponding to a given false discovery rate).

Relatedly, 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.

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 do not adjudicate among these or other possible explanations. 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 (2005, 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, 2010). Confidence intervals created using the Benjamini and Yekutieli (2001) procedure possess this property under any relationship among the test statistics.

References

Benjamini, Y. and Y. Hochberg. 1995. "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 289–300. Theorem 1. p. 290-294.

Benjamini, Yoav. 2010. "Discovering the false discovery rate." *Journal of the Royal Statistical Society, Series B* 72(4):405–416.

Benjamini, Yoav and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics* 29(4):1165–1188.

Benjamini, Yoav and Daniel Yekutieli. 2005. "False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters." *Journal of the American Statistical Association* 100(469):71–81.

Berger, Roger L. 1982. "Multiparameter Hypothesis Testing and Acceptance Sampling." *Technometrics* 24(4):295–300.

Berger, Roger L. 1997. Likelihood ratio tests and intersection-union tests. In *Advances in statistical decision theory and applications*, ed. Subramanian Panchapakesan and Narayanaswamy Balakrishnan. Boston: Birkhäuser pp. 225–237.

Brambor, Thomas, William R. Clark and Matthew Golder. 2006. "Understanding interaction models: Improving empirical analyses." *Political Analysis* pp. 1–20. pp 75-76.

Casella, George and Roger L.. Berger. 2002. *Statistical Inference, Second Edition.* Belmont, CA: Brooks/Cole.

Clark, William R. and Matthew Golder. 2006. "Rehabilitating Duverger's theory." *Comparative Political Studies* 39(6):679–708.

Cohen, Arthur, Constantine Gatsonis and John I. Marden. 1983. "Hypothesis testing for marginal probabilities in a $2 \times 2 \times 2$ contingency table with conditional independence." *Journal of the American Statistical Association* 78(384):920–929.

Esarey, Justin and Jane Lawrence Sumner. 2018. "Marginal Effects in Interaction Models: Determining and Controlling the False Positive Rate." *Comparative Political Studies* 51(9):1144–1176. DOI: `https://doi.org/10.1177/0010414017730080`.

Lehmann, Erich L. 1952. "Testing multiparameter hypotheses." *The Annals of Mathematical Statistics* pp. 541–552.

Reiner-Benaim, Anat. 2007. "FDR Control by the BH Procedure for Two-Sided Correlated Tests with Implications to Gene Expression Data Analysis." *Biometrical Journal* 49(1):107–126.

Silvapulle, Mervyn J. and Pranab K. Sen. 2005. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions.* Hoboken, NJ: Wiley.