The TOST As A Method Of Similarity Testing In Linguistics

computational and experimental approaches, methodology, equivalence testing, data simulation

INTRODUCTION Classical analyses typically test for differences and their null hypotheses state that the compared samples come from the same population. If negative, the outcome is *insufficient evidence* to assume a difference between the samples; which is not, though, sufficient to assume *equivalence* (Altman and Bland, 1995), or similarity for that matter. Linguistics heavily relies on classical tests (e.g. all 16 experimental talks at the LSA 2013 used classical tests). However, they are insufficient for many linguistic questions. Consider $RQ_{1\text{-}3}$ (p.2). Negative results for $RQ_{1\text{-}3}$ would probably go unreported. This disincentivises such research (Bakker, van Dijk, and Wikkerts, 2012) and the field might miss out. An similarity test would be more suitable.

THE TOST The TOST, attributed to Westlake (1976), is one of the most common similarity tests (Richter and Richter, 2002). It performs *t*wo *o*ne-*s*ided *t*-tests and the null hypotheses are ($H0_1$): the difference in means of the two samples is bigger than a pre-set boundary δ and ($H0_2$): the difference is smaller than −δ.

$$H0_1:\ \mu_1 - \mu_2 > \delta \qquad H0_2:\ \mu_1 - \mu_2 < -\delta$$

A positive outcome (rejecting both nulls) denotes *similarity within the range δ*. The researcher sets δ based on her knowledge of previous research. However, this leaves room for subjectiveness (Clark, 2009). Hence, our goal is to find an objective way to set δ.

DATA SIMULATION The "right" δ-value is the value that gives a positive test outcome (indicating similarity) with statistical power at $1-\alpha = 95\%$ and $1-\beta = 80\%$. To observe how the desired δ-values behave for different data, we simulate a "two-samples-one-population" setting for various datasets (24 in total; p.2) over various Ns (3 to 50000). In the simulations, we "TOSTed" random pairs of subsets from a dataset, over and over again. In total, we simulated $\sim 2.1\times10^{12}$ data points.

PREDICTING AND VALIDATING δ We found a relationship between observed δ (δ_{obs}; from our simulations) and the subsets' pooled standard deviation (s_p). This relationship is near-constant for N_p (pooled from each pair of subsamples) and we call its quotient τ (the *Tübingen Quotient*; τ comes from δ_{obs}, thus τ_{obs}; see f_1).

f_1: $\tau_{obs} = s_p \div \delta_{obs}$ (for constant Np) $\qquad$ f_2: $\tau_{pred} = (\sqrt{N_p}) \div 4.581$ $\qquad$ f_3: $\delta_{pred} = s_p \div \tau_{pred}$

Fig. 1 shows τ_{obs} over increasing Ns_p. Curve-fitting τ_{obs} led to f_2, which *predicts* τ (τ_{pred}). f_2 and the 4.581 are our critical findings, because: *by reversing f_1 to f_3, f_2 can be used to objectively set δ* (δ_{pred}). In a validation phase, we then compared τ_{obs} to τ_{pred}. For large parts, they match within ±0.1% (Fig. 2). Further simulations indicate that our results also apply to non-linguistic data.

CONCLUSION In our view, the TOST similarity test is a useful tool in a linguist's repertoire, allowing to investigate research questions that ask for similarity. So far, the lack of instructions to objectively set δ might have been a barrier to use this test. The present work outlined such guidelines and we hope that they will help boost similarity testing in linguistics.

word count (body): 489

Additional Materials

$RQ_{1\text{-}3}$

RQ_1: Can highly experienced L2 learners attain a native-like level of language production?
RQ_2: At which age do teenagers typically reach adult-like reading times?
RQ_3: Are resumptive pronouns perceived as equally bad across modalities?

THE DATASETS

Source: authors or colleagues (all 24 datasets). *Areas*: syntax (13), phonetics (8), psycho-linguistics (3). *Units*: Likert-Scale data (13), normalised Likert-Scale data (4), Hz (4), ms (3). *Aggregation*: aggregated (18), non-aggregated (6). *Size of Datasets*: 42 to 152, mean = 85.79.

GRAPHS

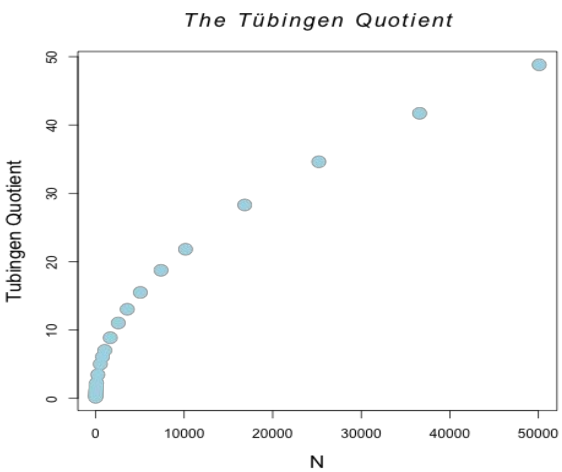

Fig. 1: τ_{obs} (y-axis) over increasing N_p (x-axis)

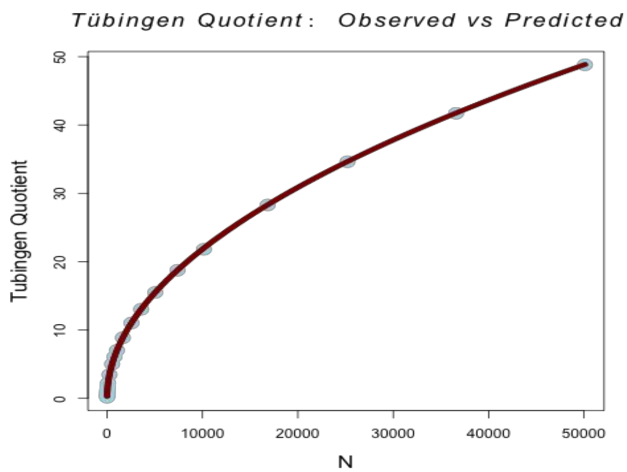

Fig. 2: τ_{obs} (y: blue) vs τ_{pred} (y: red) over increasing N_p (x)

REFERENCES

Altman, D. G., Bland, J. M. (1995). Absence Of Evidence Is Not Evidence Of Absence. British Medical Journal 311, 485.
Clark, M. (2009). Equivalence Testing [PowerPoint slides]. Retrieved 16 Dec 2013 from: www.unt.edu/rss/class/mike/5700/Equivalence%20testing.ppt.
Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. Perspectives on Psychological Science 7, 543-554.
Richter, S. J., Richter, C. (2002). A Method For Determining Equivalence In Industrial Applications. Quality Engineering 14 (3), 375-380.
Westlake, W.J. (1976). Symmetric Confidence Intervals for Bioequivalence Trials. Biometrics 32, 741-744.