MANUSCRIPT ACCEPTED FOR PUBLICATION IN PSYCHOPHYSIOLOGY
[APPEARED MARCH 2015].

A published version of the article can be found here:
http://onlinelibrary.wiley.com/doi/10.1111/psyp.12363/abstract

Resampling the Peak, some Do's and Don'ts

Alexia Zoumpoulaki, Abdulmajeed Alsufyani, Howard Bowman

University of Kent

Abstract

Resampling techniques are used widely within the ERP community to assess statistical significance and especially in the deception detection literature. Here we argue that because of statistical bias, bootstrap should not be used in combination with methods like peak – to –peak. Instead permutation tests provide a more appropriate alternative.

Keywords: bootstrap, permutation, significance testing, ERP, deception detection

Resampling the Peak, some Do's and Don'ts

As researchers, we are typically interested to demonstrate a difference between experimental conditions. This is usually done by rejecting the null hypothesis, which asserts that any difference in the sample datasets is the result of hypothesis-irrelevant (background) variation in the data. The process involves stating the experimental hypothesis, identifying the alternative (null) hypothesis, choosing and computing an appropriate statistic, determining the frequency distribution of the statistic under the null hypothesis and finally making a decision based on this distribution (Good, 2005) e.g. establishing a difference between two condition means using a t-test. But t-tests assume that the sampling distribution is normal/Gaussian. As this is not always the case, resampling techniques that avoid assumptions about the underlying distribution are often employed. Two of the most popular resampling methods are bootstrapping and permutation tests (Manly, 2006).

In the case of comparing two observed samples of size m and n, bootstrapping involves randomly resampling m data points with replacement from the first observed sample and n from the second, and calculating the statistic for the new samples. Repeating this procedure many times (>1000) allows one to approximate the statistic's distribution. Then, inferences can be made from this distribution based on its shape, center and spread. Bootstrap distributions are mainly used to calculate confidence intervals for a statistic (Hesterberg, Moore, Monaghan, Clipson, Epstein, 2005). Although some also have used them to reject a null hypothesis at an α level by showing the interval with probability $1-\alpha$ does not contain the hypothesized null value of the statistic.

Permutation tests allow one to generate an approximate null hypothesis distribution of the statistic. This involves randomly exchanging labels between the two data sets of observed values (one for each condition) and calculating the statistic for each new sample.

Again, repeating the process many times allows one to approximate the distribution of interest. The null hypothesis can be rejected at an α level if the true observed value is not contained within the interval with probability 1-α (Blair & Karniski, 1993).

Importantly, both methods are used in the ERP literature to reject null hypotheses. For example Blair and Karniski (1993), advocate permutation tests, while in the ERP lie detection literature, bootstrap tests are routinely used (Rosenfield, Miller, Rao, Soskins, 2001) (although (Bowman et al., 2013; Bowman et al., 2014) are exceptions). This letter explores the consequences of this choice. In particular, although both resampling techniques are based on random theory there is a clear difference between them. The bootstrap distribution is an approximation of the statistic distribution, while the permutation distribution is an approximation of the null hypothesis distribution (Hesterberg et al., 2005). This focus on the difference of the generated distributions would largely be of esoteric interest if the derived p-values were effectively the same. However, this is not the case for certain measures, and in particular, for the maximum, which is our point of focus. The problem results from the fact that bootstrap underestimates the maximum. One can find mathematical proofs of why bootstrap fails to approximate the maximum as well as other statistics that are on the boundary of the parameter space (Bickel & Freedman, 1989; Andrews, 2000; Abrevaya & Huang, 2005; Lehmann, Romano, 2006), but a simple example could help demonstrate this. Suppose we took a sample of size 10 from a Gaussian distribution. The sample set (rounded to the 3rd decimal) is:
A = {0.234, 3.488, -0.267, -0.244, -2.177, -0.405, -1.208, 0.229, -0.738, 0.288}, with a sample max of 3.488. If we were to create bootstrap samples from this observed sample there is a probability of 0.35 of not selecting the maximum value, $P(\text{not selecting max}) = \left(1 - \frac{1}{n}\right)^n = (1 - \frac{1}{10})^{10} = 0.35$. As the sample has a large variance (from the max) the rest of the values are

far from the max and the bias (distance of the mean of the bootstrap distribution from the original statistic) is going to be large. More specifically, the mean of the bootstrap distribution is 2.585 and the bias is -0.903 (Bias = 2.585-3.488). It is clear that in this case, the bootstrap distribution underestimates the maximum value of the sample and that this underestimation is large with respect to the dispersion of the sampling distribution.

Someone might argue that bootstrapping of very small-observed samples (such as in this example) is not advisable. In order to counter this concern we generated 1000 observed samples of size 1000 from a Gaussian distribution and for each one we generated the bootstrap distribution of the maximum. We found that the average bias was, $\overline{\text{Bias}} = -0.13$, the average variance from the max: $\overline{\text{Var}_{\text{max}}} = 11.7$ and that there was a correlation between the variance from the maximum in the observed sample and the bias.

Cor(var_max, bias) = -0.814

That is, the greater the variance of the original samples from the max, the more the bootstrap mean is below the sample max, and thus the bigger the bootstrap bias.

Figure 1 shows the statistic distribution as calculated for 1000 generated samples from a normal distribution for mean, max, and the correlation coefficient and difference in max between two such samples. Then the mean of the bootstrap distributions[1] from each one is over imposed. Bootstrap accurately approximates the distribution of means ($\overline{\text{Bias}} = 0.001$) and correlation coefficients ($\overline{\text{Bias}} = 0.0008$). Importantly, although the bias for the difference of two maxima is very small ($\overline{\text{Bias}} = 0.003$), the distribution of the difference between the bootstrapped maxima is narrower than the sampling distribution. This should be expected, as bootstrap underestimates larger maxima more than smaller ones, since the variance from the maximum in a sample tends

[1] All bootstrap distributions discussed in this letter are generated from 1000 resampling's

to increase with the maximum. Accordingly, more extreme differences of maxima (whether positive or negative) are pushed towards zero, since they respectively include one large maximum (which is underestimated a lot) and one small maximum (which is underestimated much less). Summary of the generated distributions is given in Table 1.

Efron proposed bootstrap in 1979 (Efron, 1979) and since then its weaknesses and strengths are widely known. But some of these problems have not been recognized in the ERP setting. For example in 1989, Wasserman and Bockenholt described how bootstraping could be used to make inferences about guilt from data collected in an ERP deception detection experiment (Farwell & Dochin, 1986). They used the difference of the correlation coefficient between Guilty knowledge and Task ERPs and Guilty knowledge and Irrelevant ERPs. They proposed statistical inference based upon whether the null hypothesis (difference of 0) was included in the 95% confidence interval of the bootstraped difference of correlation coefficients (Wasserman & Bockenhold, 1989). This method gained popularity and was used in the deception detection paradigm to compare the P300 component between conditions (Farwell & Donchin, 1991). But many considered that the correlation coefficient was not the most appropriate measure, since the P300 for the guilty knowledge may not resemble very closely that for the experimental task (Allen & Iacono, 1997). Instead, the difference of P300 amplitude between conditions was proposed (Rosenfeld et al., 2001). Amplitude was measured either with the peak-to-peak (p2p) or with the peak-to-baseline (p2b) method (Meixner & Rosenfeld, 2010; Hu & Rosenfeld, 2012). But as both of these measurements are in the boundary of the parameter space (p2p = difference between maximum and minimum and p2b = maximum from a baseline), bootstrap is, in fact, inappropriate. Since as just documented, bootstrap's underestimation increases with the extremity of the max/min, the distribution of the differences will be closer to

zero than in the null distribution, resulting in a loss of statistical power. To illustrate this, we generated two EEG datasets, each consisting of noise in the spectrum of human EEG[2] and then extracted a p-value by bootstrapping the p2p measure (Meixner & Rosenfeld, 2010). We repeated this process 10000 times, with the results in Figure 2. The distribution of p-values is not uniform, as it should be, since arbitrary noise data sets are equally likely to fall in any percentile of the null hypothesis distribution. Instead there is a bias away from extreme p-values and towards intermediate ones. At the same time, the p-values obtained from permutation tests on the same null samples are uniform. Additionally we performed the same analysis on white noise data, with the same results.

In summary, bootstrap tests with p2p or p2b, as commonly used in the deception detection literature, are biased. Most significantly, the method will tend to push small p-values, which might otherwise be significant, up towards 0.5. This will induce an unnecessary loss of statistical power, suggesting that existing studies may have underestimated the effectiveness of their deception detection methods. Although permutation tests might have their limitations, permuting p2p or p2b measurements suffers no such bias and should be the inferential method of choice in this context.

[2] The script used to generate noise can be found at http://www.cs.bris.ac.uk/~rafal/phasereset/.

References

Abrevaya, J., & Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica, 73*(4), 1175-1204.

Allen, J. J., & Iacono, W. G. (1997). A comparison of methods for the analysis of event- related potentials in deception detection. *Psychophysiology,* 34(2), 234-240.

Andrews, D. W. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica, 68*(2), 399-405.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics,* 1196-1217.

Blair, R. C. & Karniski, W. (1993). An alternative method for significance testing of waveform difference potentials. *Psychophysiology, 30*(5), 518-524.

Bowman, H., Filetti, M., Alsufyani, A., Janssen, D., & Su, L. (2014). Countering countermeasures: Detecting identity lies by detecting conscious breakthrough. *PloS One, 9*(3), e90595.

Bowman, H., Filetti, M., Janssen, D., Su, L., Alsufyani, A., & Wyble, B. (2013). Subliminal salience search illustrated: EEG identity and deception detection on the fringe of awareness. *PloS One, 8*(1)

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics,* 1-26.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* (Vol. 57). CRC press.

Farwell, L. A., & Donchin, E. (1991). The truth will out: Interrogative polygraphy ("Lie detection") with Event- Related brain potentials. *Psychophysiology, 28*(5), 531-547.

Farwell, L., & Donchin, E. (1986). The 'brain detector': P300 in the detection of deception [Abstract]. *Psychophysiology, 23*(4) 434.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses.* New York, NY: Springer Series in Statistics

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., & Epstein, R. (2005). Bootstrap methods and permutation tests. *Introduction to the Practice of Statistics, 5*, 1-70.

Hu, X., & Rosenfeld, J. P. (2012). Combining the P300- complex trial- based concealed information test and the reaction time- based autobiographical implicit association test in concealed memory detection. *Psychophysiology, 49*(8), 1090-1100.

Lehmann, E. L., & Romano, J. P. (2006). *Testing statistical hypotheses*. New York, NY: Springer.

Manly, B. F. (2006). Randomization, bootstrap and monte carlo methods in biology. London: Chapman & Hall.

Meixner, J. B., & Rosenfeld, J. P. (2010). Countermeasure mechanisms in a P300 based concealed information test. *Psychophysiology*, 47(1), 57-65.

Rosenfeld, J. P., Miller, A. R., Rao, A., & Soskins, M. (2001). Event-related potentials in detection of deception. *Handbook of Polygraphy.* New York , NY: Academic Press.

Wasserman, S., & Bockenholt, U. (1989). Bootstrapping: Applications to psychophysiology. *Psychophysiology*, 26(2), 208-221.

Table 1

Summary of distributions. Comparison of the distributions of four statistics (mean, max, correlation coefficient, difference between maxima) across 1000 samplings from a normal distribution and the mean of the bootstrap distribution generated for each sampling.

	Summary of distribution statistic across 1000 samplings					
Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
mean	-0.08648	-0.02137	0.00126	0.001619	0.02321	0.109
max	2.47	3.002	3.204	3.245	3.43	5.285
diff_cor	-0.1114	-0.01988	0.001059	0.0008311	0.02287	0.1087
diff_max	-1.875	-0.3178	-0.0243	-0.01195	0.3141	2.257
	Summary of the histogram of means of bootstrap distributions across the same 1000 samples					
mean	-0.08591	-0.02101	0.001387	0.001652	0.02334	0.1087
max	2.43	2.915	3.08	3.109	3.267	4.523
diff_cor	-0.1118	-0.02002	0.0007477	0.00084	0.02296	0.1085
diff_max	-1.189	-0.2606	-0.008414	-0.009064	0.2455	1.518

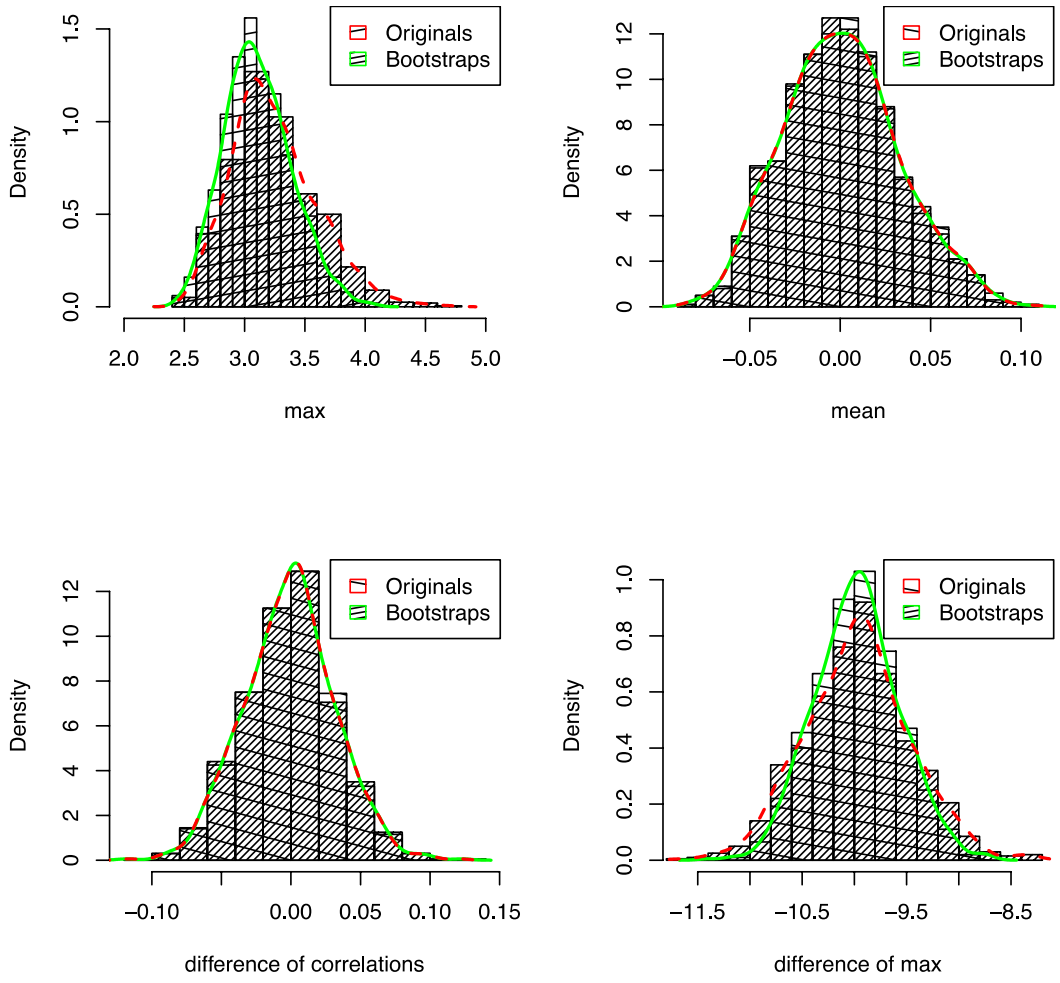

Figure 1. Comparison of sampling distributions vs the histogram of means of bootstrap distributions. Each of the four statistics was applied on 1000 values sampled from normal distributions in order to obtain the sampling distributions. The process was repeated 1000 times..

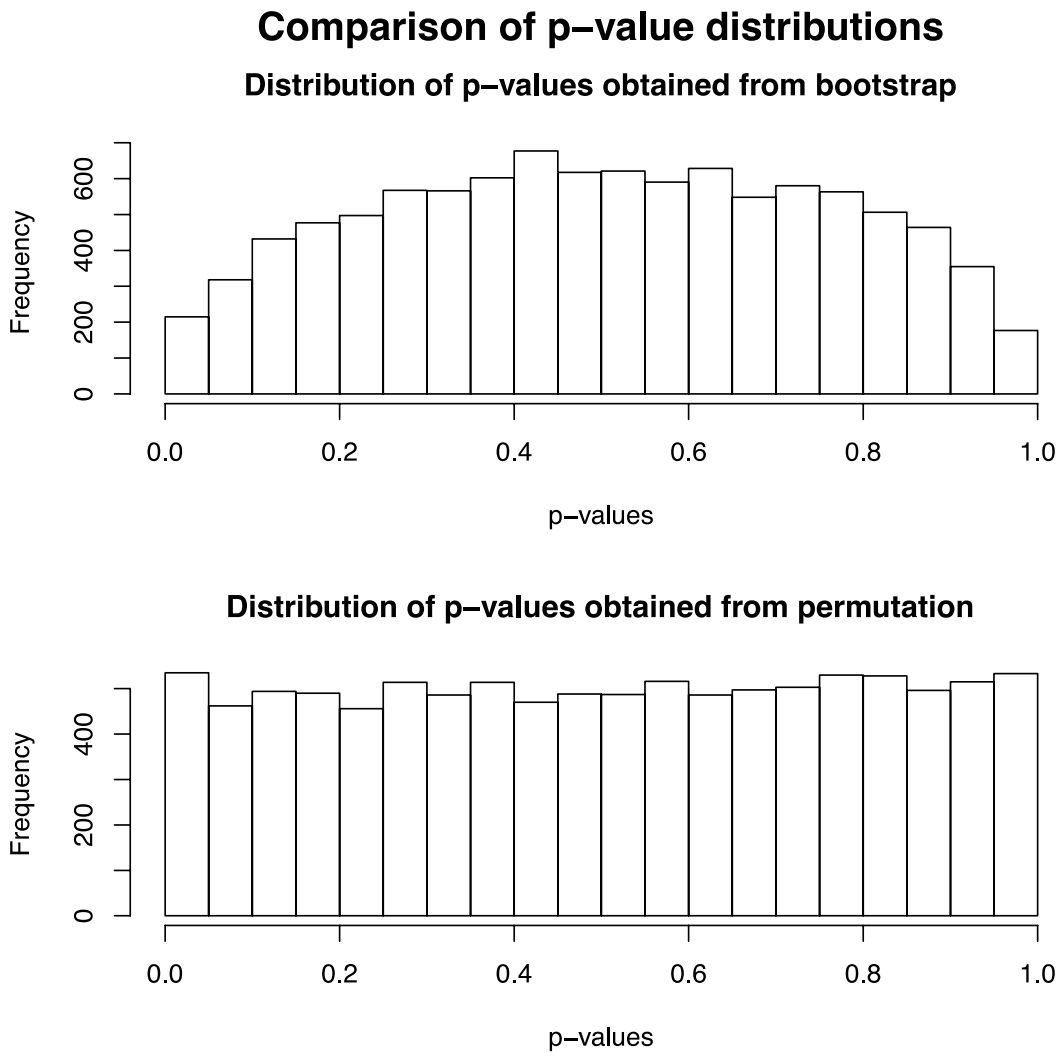

Figure 2. Comparison of p-values obtained from 10000 bootstrap tests vs 10000 permutation tests. P-values were obtained for p2p measurement on simulated noise EEG data. Using the Chi-squared test to check for uniformity, we can reject the hypothesis that the bootstrap distribution is uniform (p-value < 0.00000000000000022, while for permutation p-value = 0.3915).