

Letters

A bootstrap approach is a superior statistical method for the comparison of non-normal data with differing variances

Phytologists have a primary interest in understanding plant growth, development and environmental responses. Experimentally, we rely on probing these by perturbing a system and monitoring changes, whether it be growth rate with temperature or gene expression in response to stress. To assess the significance of data from such experiments, frequentist statistics are used to ascertain the probability that a difference in a test statistic between conditions was due to chance (a P -value). When data are not normal, the adage is to use a nonparametric test for this analysis: the most common being the Mann–Whitney–Wilcoxon (MWW) test. Here, we explore conditions in which the MWW test is unsuitable and propose the use of a bootstrap approach instead.

Most experiments aim to assess whether some metric is greater or less under different conditions, by an analysis of the change in the median or mean of that metric. The MWW test challenges the null hypothesis that two data distributions are the same (Mann & Whitney, 1947), not whether the two distributions have the same median. Therefore, it is possible to find a significant difference in an MWW test with distributions that have the same median, but different variances (i.e. heteroskedastic data) (Hart, 2001).

These conditions are not uncommon in several experimental contexts, one being the measurement of symplastic cell-to-cell spread of green fluorescent protein (GFP) (Oparka *et al.*, 1999; Burch-Smith & Zambryski, 2010). These assays allow the experimenter to count the number of cells ('cell count'), or the number of concentric rings of cells ('cell layers'), to which GFP has spread from a single cell (Fig. 1a) as a measure of the status of connectivity. Neither cell nor layer counts are normally distributed (Fig. 1b–d, upper), so most studies use the nonparametric MWW test to compare conditions to identify factors that regulate the connection and communication between cells.

However, it is also clear that the shape of the distributions differs between experimentally compared conditions or genotypes (Fig. 1b,c) (Guseman *et al.*, 2010; Diao *et al.*, 2018; Cheval *et al.*, 2020). Thus, if an MWW test is used on cell count data, the difference in distribution shapes between conditions may lead to the erroneous conclusion that there is a significant difference in the amount of spread of GFP. Therefore, a different statistical method is required to properly interpret differences in nonparametric heteroskedastic data. For this, we propose a bootstrap method (Efron, 1979).

A P -value is defined as the probability of observing a value at least as extreme as the observed test static. This is done by a comparison with the null distribution, which describes the probability distribution of the test static when the null hypothesis is true. In the case of cell-to-cell connectivity, the test statistic is the difference in medians ($\hat{\theta}$). The null distribution describes the probability of observing a difference in medians when there is no true difference in the underlying data. Usually, a known distribution is used (e.g. t -distribution or F -distribution) but in this case it is unknown because the data do not follow parametric distributions (Fig. 1b,c).

Bootstrapping techniques can be used to generate a null distribution *de novo* from the observed data already collected if the samples are independent. This removes the requirement of using a known distribution. To do this, the observed data are sampled with replacement to generate a resample. This mimics what the experimenter has done originally when observing the true population. The relationship between multiple resamples and the observed data can be used to reveal how the observed data relate to the true population, and so estimate a P -value for the observation.

An example R function is provided to perform this analysis (Supporting Information Notes S1, *medianBootstrap.R*; <https://github.com/faulknerfalcons/Johnston-2020-Bootstrap>), which requires two arguments, that is two vectors of numbers: control and treatment. The function generates a null distribution to compare against by resampling each vector N times (by default 5000) and, for each resample, generates a resampled test statistic ($\hat{\theta}^*$). These N resampled test statistics are made into a null distribution by $|\hat{\theta}_n^* - \hat{\theta}|$ (Fig. 1b,c, lower) as suggested by Hall & Wilson (1991).

As this is a random sampling technique, an exact P -value cannot be calculated but is estimated by a Monte Carlo \hat{P} -value (Eqn 1). Thus, $\hat{\theta}$ is compared with the null distribution to find the chance of observing a value at least as extreme (line on Fig. 1b,c, lower). A + 1 is added to the numerator and denominator in Eqn 1 as suggested by Davison & Hinkley (1997): conceptually, this can be considered as including the observed sample among the bootstrap resamples.

$$\hat{p} = \frac{\sum_{n=1}^N I\left(\left|\hat{\theta}_n^* - \hat{\theta}\right| \geq \hat{\theta}\right) + 1}{N + 1} \quad \text{Eqn 1}$$

where $I(\cdot)$ is the indicator function and $\hat{\theta}_n^*$ is the n^{th} resampled test statistic.

As \hat{P} is an estimate of P a 95% confidence interval should be constructed, where P will fall within this range 95% of the time (Wilson, 1927).

This method is not confounded by differences in variance or shape as with the MWW test. To illustrate this, we compared the

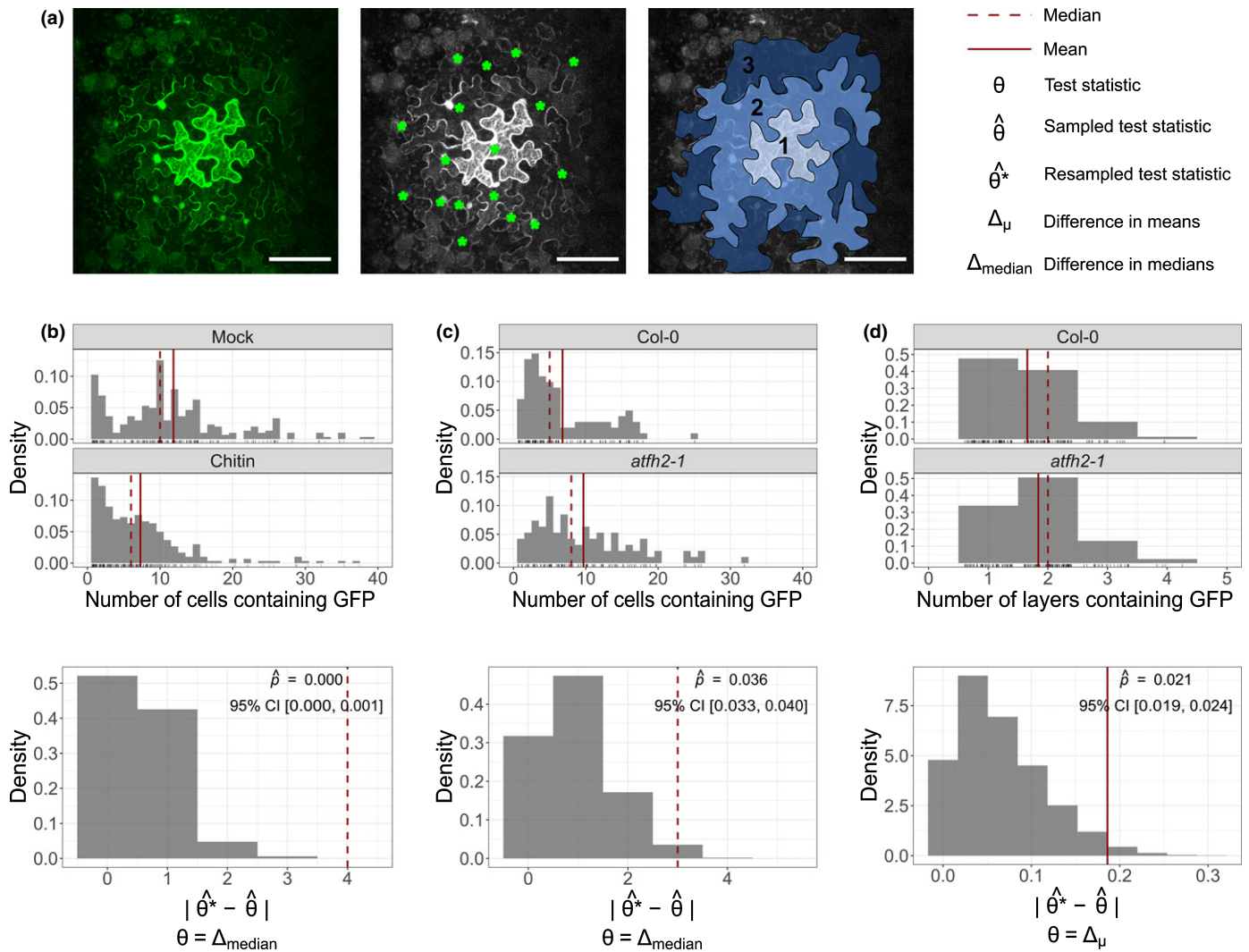


Fig. 1 Bootstrap statistics on GFP movement data. (a) An example image of GFP moving from a single transformation site. The degree of movement can either be counted as the number of fluorescent cells (denoted with stars, 17 cells) or the number of cell layers with GFP (blue overlays, three layers). Bars, 100 μm . (b–d) Top: Histograms of cell counts or layers, with the median and mean marked. Bottom: Bootstrap null distributions ($|\hat{\theta}^* - \hat{\theta}|$) for the differences in (b, c) median or (d) mean, with estimated \hat{P} -values and 95% confidence intervals (CI). The observed difference ($\hat{\theta}$) is marked by a red line. Data for (b) from Cheval *et al.* (2020) and data for (c, d) from Diao *et al.* (2018) are both under the CC-BY 4.0 licence.

type I error rate (false positives) between the MWW and *medianBootstrap* tests, when testing if there is a difference in medians between two populations for which there was no true difference in medians, that is $\theta = 0$ (Fig. 2). In this scenario, an error rate of 5% is expected at $\alpha = 0.05$. Equal samples ($n_A, n_B = 100$) for each population were drawn from normal distributions with the same median and three different shapes, simulated in R v.4.0.0 (R Core Team, 2020).

In the first instance, in which both distributions are equal ($A, B \sim N(0, 1)$), both the MWW and *medianBootstrap* methods gave a difference in medians about 5% of the time, as expected (4.5% (95% CI [3.4, 6.0]) and 4.9% (95% CI [3.7, 6.4]), respectively) (Fig. 2a). When variances differed between populations ($A \sim N(0, 1)$, $B \sim N(0, 5^2)$), the MWW test had a false-positive rate significantly higher than the set 5% of 7.5% (95% CI [6.0, 9.3]). Conversely, the false-positive rate of the

medianBootstrap method was correctly controlled at 4.7% (95% CI [3.6, 6.2]) (Fig. 2b). When two samples are drawn from populations with equal variance and median, but differing shape and mean ($A \sim N(1 - \frac{1}{\sqrt{2}}, \frac{3}{80})$, $B \sim \text{Beta}(1, 3)$), a *medianBootstrap* method finds a significant difference in 5.1% of the trials (95% CI [3.9, 6.6]), as expected, whereas an MWW test inflates the type I error rate to 17% (95% CI [15, 19]) (Fig. 2c).

The *medianBootstrap* method was robust to unequal sample sizes in all cases, provided the sample number was 10 or greater (Fig. 2). Thus, we recommend that both samples have at least 10 constituents for a robust result. Alongside this sample size requirement, samples must be independent and representative of the overall population. A bootstrapped difference in medians was used here as a comparison against the MWW test; it is worth noting that any test statistic, θ , can be computed in a bootstrapped

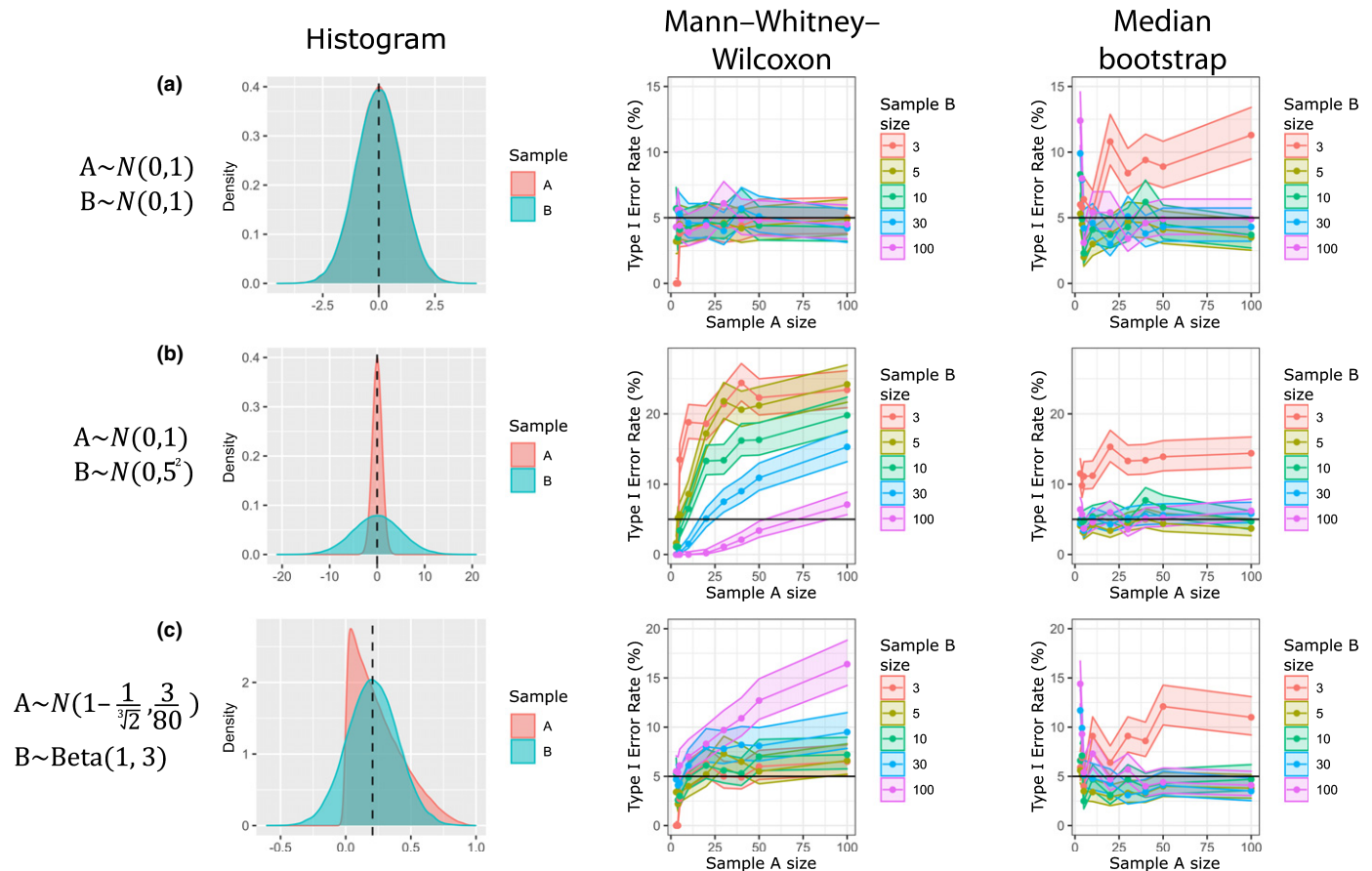


Fig. 2 Type I error rates are correctly controlled using a median bootstrap method. Three experiments were simulated in which data were compared from two samples drawn from different distributions with equal medians. The false-positive (type I error) rate was estimated for the null hypothesis that population medians are the same using the Mann–Whitney–Wilcoxon and *medianBootstrap* test in each experiment with 1000 simulations. The 95% confidence intervals for the type I error rate are indicated by shading. Each experiment was repeated with differing sample sizes for both A and B. (a) Two samples were drawn from identical distributions. (b) Two samples have the same mean, but differing variances. (c) Two samples have the same variance, but differ in shape and mean. The medians of each distribution are marked with a dashed line.

manner, provided the test is invariant to scaling. This is required, as a bootstrap test generates the null distribution by rescaling the distributions around 0. Thus, the test statistic must be the same at X and $X + n$, for example the median and arithmetic mean are suitable, but the geometric mean is not.

Therefore, the *medianBootstrap* method is a more appropriate analysis to identify differences in the spread of GFP, as cell count data exhibit unequal variances and differing distribution shapes between conditions and/or genotypes. Moreover, bootstrap testing can be extended to cell layer data, where means should be compared, as there is no difference in medians (Fig. 1d). An example of this extension is given in Notes S2, *medianBootstrap.html*.

We acknowledge alternative advanced statistical techniques, such as linear mixed effects models, for analysis of these data. However, they require more assumptions and are less user friendly, often leading to mistakes (Knief & Forstmeier, 2020). In addition, we note that data are more complex than a single metric and support the movement to present figures with the data points, as well as summary statistics (Weissgerber *et al.*, 2019). Moreover, while a summary statistic such as the median may remain unchanged, the distributions may still differ in a biologically relevant way, such as in a difference in variance. We consider this

bootstrap method a good, easy-to-use, superior alternative to MWW analysis of cell-to-cell movement data.



Acknowledgements

We thank Joanna Jennings (Department of Crop Genetics, John Innes Centre, UK) for providing the confocal micrograph in Fig. 1 (a) and Joshua Hodgson (Department of Medicine, University of Cambridge, UK) and Matthew Castle (Department of Genetics, University of Cambridge, UK) for constructive comments on the manuscript. Data presented in Fig. 1 come from (b) Fig. S2 of Cheval *et al.* (2020), (c) Fig. 2(d) Diao *et al.* (2018) and (d) Fig. 2(c) Diao *et al.* (2018) under use of the CC-BY 4.0 licence. MGJ is funded by a John Innes Foundation Studentship. Research in the Faulkner laboratory is supported by the Biotechnology and Biological Research Council (BB/L000466/1, BBS/E/J/000PR9796) and the European Research Council (725459, ‘INTERCELLAR’).

Author contributions

MGJ and CF designed, discussed and wrote up the research. MGJ performed the analysis.

ORCID

Christine Faulkner  <https://orcid.org/0000-0003-3905-8077>
 Matthew G. Johnston  <https://orcid.org/0000-0003-1141-6135>

Matthew G. Johnston*  and Christine Faulkner 

Department of Crop Genetics, John Innes Centre, Norwich,
 NR4 7UH, UK

(*Author for correspondence: email matthew.johnston@jic.ac.uk)

References

- Burch-Smith TM, Zambryski PC. 2010. Loss of increased size exclusion limit (ise1 or ise2) increases the formation of secondary plasmodesmata. *Current Biology* 20: 989–993.
- Cheval C, Samwald S, Johnston MG, de Keijzer J, Breakspear A, Liu X, Bellandi A, Kadota Y, Zipfel C, Faulkner C. 2020. Chitin perception in plasmodesmata characterizes submembrane immune-signaling specificity in plants. *Proceedings of the National Academy of Sciences, USA* 117: 9621–9629.
- Davison AC, Hinkley DV. 1997. *Bootstrap methods and their application*. Cambridge, UK: Cambridge University Press.
- Diao M, Ren S, Wang Q, Qian L, Shen J, Liu Y, Huang S. 2018. Arabidopsis formin 2 regulates cell-to-cell trafficking by capping and stabilizing actin filaments at plasmodesmata. *eLife* 7: e36316.
- Efron B. 1979. Bootstrap methods: another look at the Jackknife. *Annals of Statistics* 7: 1–26.
- Guseman JM, Lee JS, Bogenschutz NL, Peterson KM, Virata RE, Xie B, Kanaoka MM, Hong Z, Torii KU. 2010. Dysregulation of cell-to-cell connectivity and stomatal patterning by loss-of-function mutation in Arabidopsis chorus (GLUCAN SYNTHASE-LIKE 8). *Development* 137: 1731–1741.
- Hall P, Wilson SR. 1991. Two guidelines for bootstrap hypothesis testing. *Biometrics* 47: 757–762.
- Hart A. 2001. Mann–Whitney test is not just a test of medians: differences in spread can be important. *British Medical Journal* 323: 391–393.
- Knief U, Forstmeier W. 2020. Violating the normality assumption may be the lesser of two evils. *bioRxiv*: 498931.
- Mann HB, Whitney DR. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18: 50–60.
- Oparka KJ, Roberts AG, Boevink P, Cruz SS, Roberts I, Pradel KS, Imlau A, Kotlizky G, Sauer N, Epel B. 1999. Simple, but not branched, plasmodesmata allow the nonspecific trafficking of proteins in developing tobacco leaves. *Cell* 97: 743–754.
- R Core Team. 2020. *R: a language and environment for statistical computing, v.4.0.0*. Vienna, Austria: R Foundation for Statistical Computing.
- Weissgerber TL, Winham SJ, Heinzen EP, Milin-Lazovic JS, Garcia-Valencia O, Bukumiric Z, Savic MD, Garovic VD, Milic NM. 2019. Reveal, don't conceal: transforming data visualization to improve transparency. *Circulation* 140: 1506–1518.
- Wilson EB. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 22: 209–212.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Notes S1 A short R script that provides a bootstrap function (*medianBootstrap()*) for the comparison of medians between two datasets, which returns a \hat{P} -value and associated confidence intervals. It can be used to directly replace the Mann–Whitney–Wilcoxon *wilcox.test()* function from base R. *medianBootstrap(data1, data2, N = 5000, alpha = 0.05)*, requires two vectors of data (data1 and data2) and accepts to optional arguments: the number of bootstraps (N, default 5000) and the significance level for the constructing the confidence intervals (alpha, default 0.05).

Notes S2 A primer on how to use the *medianBootstrap.R* code, with an example use case of the *medianBootstrap()* function; the code is then extended to replicate the functionality of a one-way ANOVA, allowing multiple-to-one bootstrap comparisons, as well as providing functions for the comparison of means and plotting the null distribution.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Key words: bootstrap, cell-to-cell movement, heteroskedastic, Mann–Whitney–Wilcoxon, null hypothesis significance testing, plasmodesmata, statistics.

Received, 11 September 2020; accepted, 15 December 2020.