Bootstrap tests

Patrick Breheny

October 11

Introduction

- Conditioning on the observed data to obtain permutation tests is certainly an important idea and a useful tool for hypothesis testing
- Once we condition on the observed data, we can test hypotheses of whether or not the data is identically and independently drawn from the same distribution by permuting the observed values – or equivalently, drawing without replacement from the observed values
- One could also think about drawing *with replacement* from the observed values – *i.e.*, carrying out a *bootstrap test*

Advantages of permutation testing

- What are the ramifications of this difference?
- On the one hand, the p-values of a permutation test are exact conditional probabilities (up to computational limits) for all sample sizes
- Permutation tests do not make any effort to estimate the common distribution F; it is treated as a nuisance parameter
- In contrast, a bootstrap test estimates F using the empirical cdf and draws from $\hat{F}$ to estimate p-values
- Bootstrap tests therefore have no interpretation as exact probabilities, for any sample size

Advantages of bootstrap testing

- However, their probabilities are guaranteed to be accurate as the sample size becomes large
- Furthermore, bootstrap tests are more general, and can be applied to a wider range of hypotheses
- In contrast, permutation tests are typically limited to relatively simple hypotheses like H_0, H_1, and H_2

Testing for equality of means

- For example, consider again the two-sample problem where $X \sim F$ and $Y \sim G$, but instead of testing $H_0 : F = G$, we wish to test $H_0 : \mathbb{E}(X) = \mathbb{E}(Y)$
- In other words, F and G may differ in other ways — for example, $\mathbb{V}(X) \neq \mathbb{V}(Y)$ — but as long as $\mathbb{E}(X) = \mathbb{E}(Y)$, the null hypothesis is still satisfied
- This hypothesis cannot be tested with a permutation test, but can be tested with a bootstrap approach

Bootstrap approach

- The bootstrap test proceeds based on estimating F and G subject to the restriction that they must both have the same mean
- Specifically, let $\hat{F}_0$ put equal probability on the points $\tilde{x}_i = x_i - \bar{x} + \bar{z}$ and $\hat{G}_0$ put equal probability on the points $\tilde{y}_i = y_i - \bar{y} + \bar{z}$, where $\bar{z}$ is the mean of the combined sample
- This will leave everything about the distribution of X and Y the same, with the exception of their location, which are constrained to have the same center, $\bar{z}$

Bootstrap p-value: Procedure

(1) Form B bootstrap data sets $(\mathbf{x}^*, \mathbf{y}^*)$ by sampling $\mathbf{x}^*$ with replacement from $\tilde{\mathbf{x}}$ and $\mathbf{y}^*$ with replacement from $\tilde{\mathbf{y}}$

(2) Evaluate t_b^* for each data set $1, \ldots, B$:

$$t_b^* = \frac{\bar{x}_b^* - \bar{y}_b^*}{\sqrt{\hat{\sigma}_{xb}^{*2}/n + \hat{\sigma}_{yb}^{*2}/m}},$$

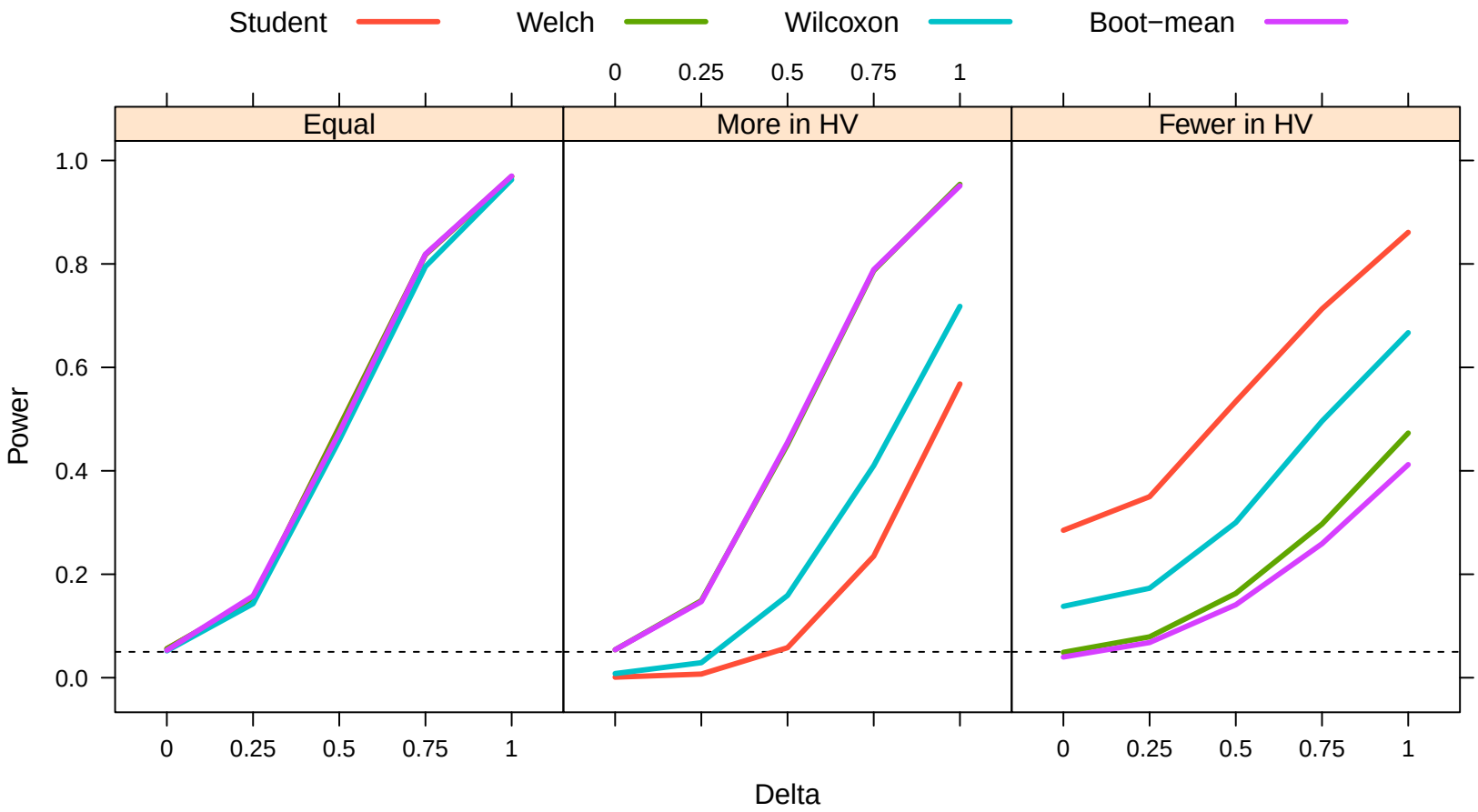
where $\bar{x}_b^*$ is the mean and $\hat{\sigma}_{xb}^{*2}$ is the variance of the bth bootstrap sample $\mathbf{x}_b^*$, and so on

(3) Approximate the achieved significance level by:

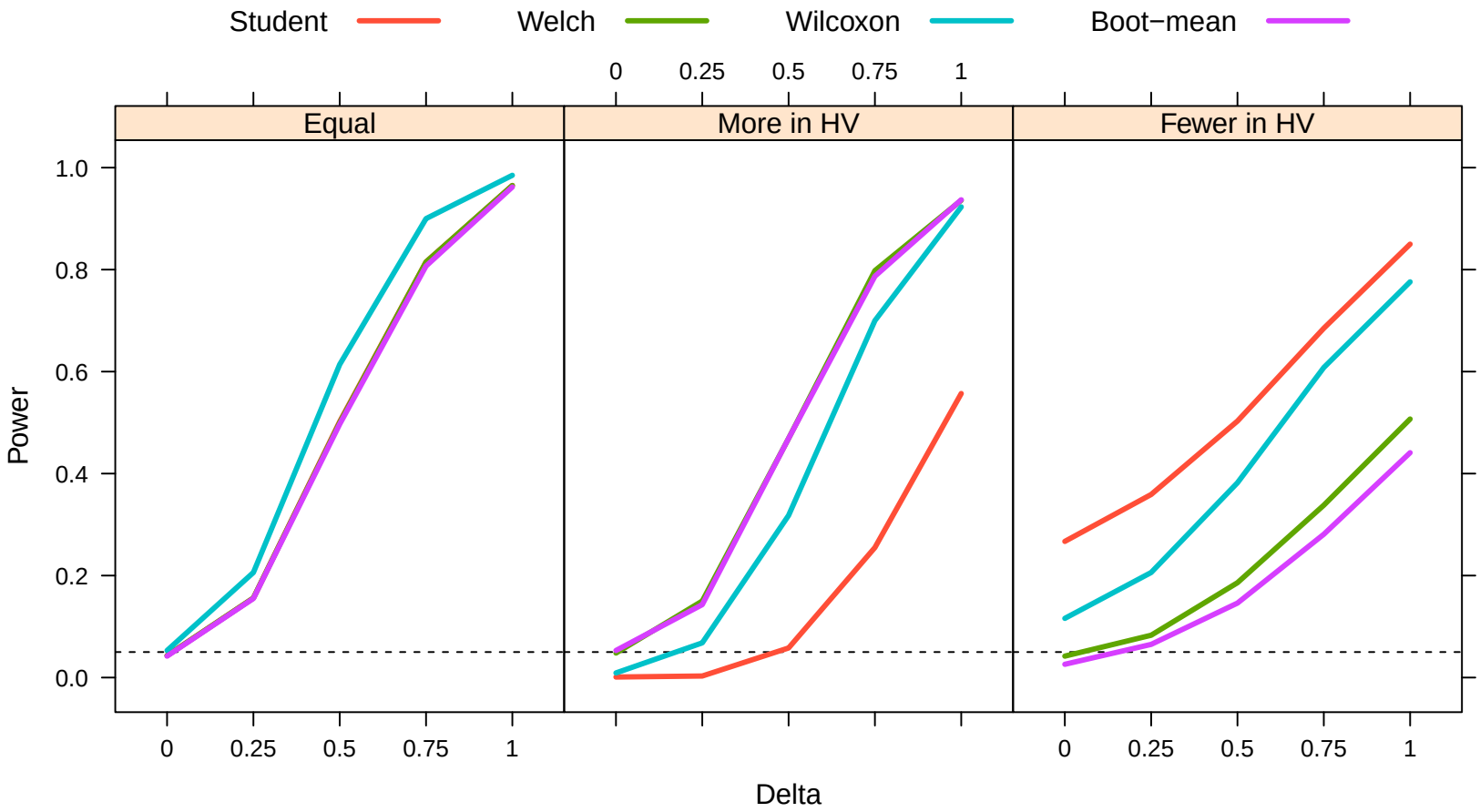
$$\widehat{ASL}_{boot} = B^{-1} \sum_b \mathbf{1}(|t_b^*| \geq |t_{obs}|)$$

where t_{obs} is the observed value of the test statistic

Simulation results: Normal distribution



Simulation results: Double exponential distribution



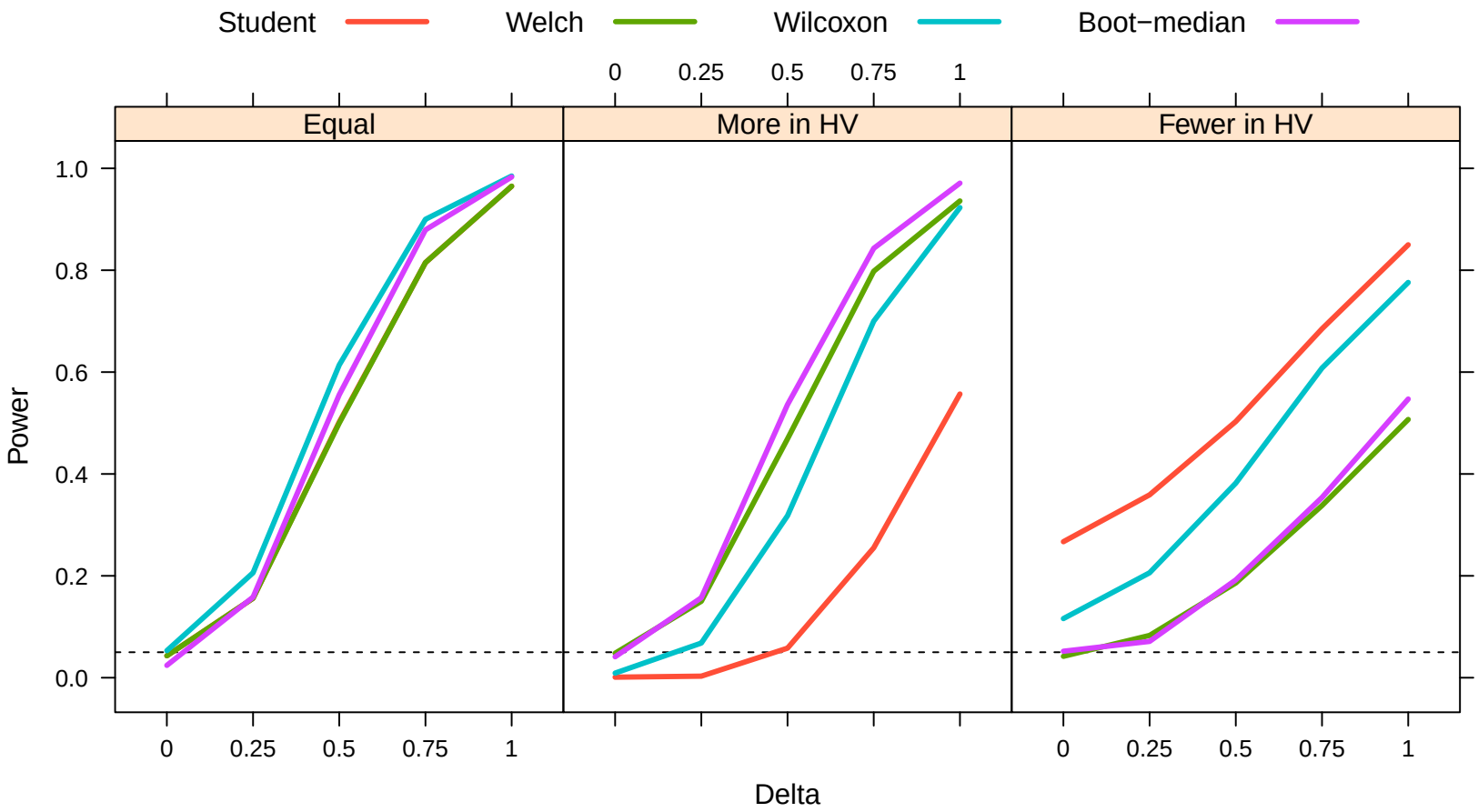
Bootstrap test vs. Welch's test

- It is perhaps not surprising that the results of this bootstrap test that we have developed are quite similar to that of Welch's test
- However, unlike Welch's test, we can easily extend this idea to other tests
- For example, perhaps we would have more power if we decided to test the hypothesis of equal medians

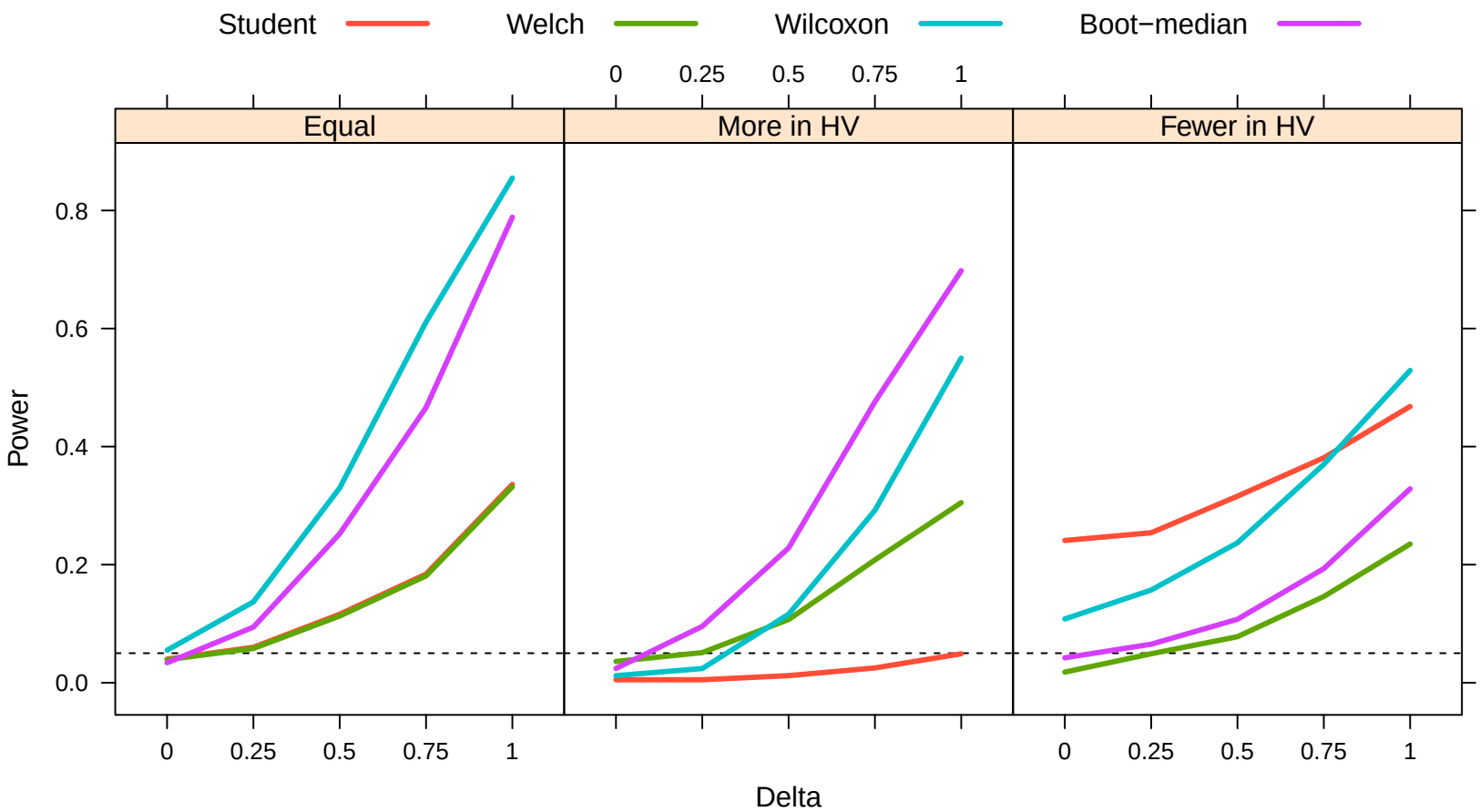
Testing equality of medians

- An alternative bootstrap test, then, is to estimate $\hat{F}_0$ by placing equal mass at all $x_i - \tilde{x} + \tilde{z}$, where $\tilde{x}$ is the median of the $\{x_i\}$'s and $\tilde{z}$ is the median of the combined sample, and so on for $\hat{G}_0$
- A possible test statistic would then be the difference of medians between the resampled values from $\hat{F}_0$ and $\hat{G}_0$

Simulation results: Double exponential distribution



Simulation results: Mixture of normals



Homework

Homework: At a conference once, I heard someone say that Wilcoxon rank-sum tests are not valid level α tests for the two-group comparison problem when the variances of the two groups are different. Is this correct? Why or why not?