Hypothesis Testing with the Bootstrap

Noa Haas

Statistics M.Sc. Seminar, Spring 2017

Bootstrap and Resampling Methods

Bootstrap Hypothesis Testing

A bootstrap hypothesis test starts with a test statistic - $t(\boldsymbol{x})$ (not necessary an estimate of a parameter).

We seek an achieved significance level

$$ASL = Prob_{H_0}\{t(\boldsymbol{x}^*) \geq t(\boldsymbol{x})\}$$

Where the random variable $\boldsymbol{x}^*$ has a distribution specified by the null hypothesis H_0 - denote as F_0.

Bootstrap hypothesis testing uses a "plug-in" style to estimate F_0.

The Two-Sample Problem

We observe two independent random samples:

$$F \rightarrow \quad \boldsymbol{z} = (z_1, z_2, \ldots, z_n) \ \ independently\ of$$
$$G \rightarrow \quad \boldsymbol{y} = (y_1, y_2, \ldots, y_m)$$

And we wish to test the *null hypothesis* of no difference between F and G,

$$H_0 : F = G$$

Bootstrap Hypothesis Testing $F = G$

- Denote the combined sample by $\boldsymbol{x}$, and its empirical distribution by $\hat{F}_0$.
- Under H_0, $\hat{F}_0$ provides a non parametric estimate for the common population that gave rise to both $\boldsymbol{z}$ and $\boldsymbol{y}$.

1. Draw $\boldsymbol{B}$ samples of size $n + m$ **with replacement** from $\boldsymbol{x}$. Call the first n observations $\boldsymbol{z}^*$ and the remaining $m - \boldsymbol{y}^*$
2. Evaluate $t(\cdot)$ on each sample - $t(\boldsymbol{x}^{*\boldsymbol{b}})$
3. Approximate ASL_{boot} by

$$\widehat{ASL}_{boot} = \#\left\{t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right) \geq t(\boldsymbol{x})\right\}/B$$

* In the case that large values of $t(\boldsymbol{x}^{*\boldsymbol{b}})$ are evidence against H_0

Bootstrap Hypothesis Testing $F = G$ on the Mouse Data

A histogram of bootstrap replications of

$$t(\boldsymbol{x}) = \bar{z} - \bar{y}$$

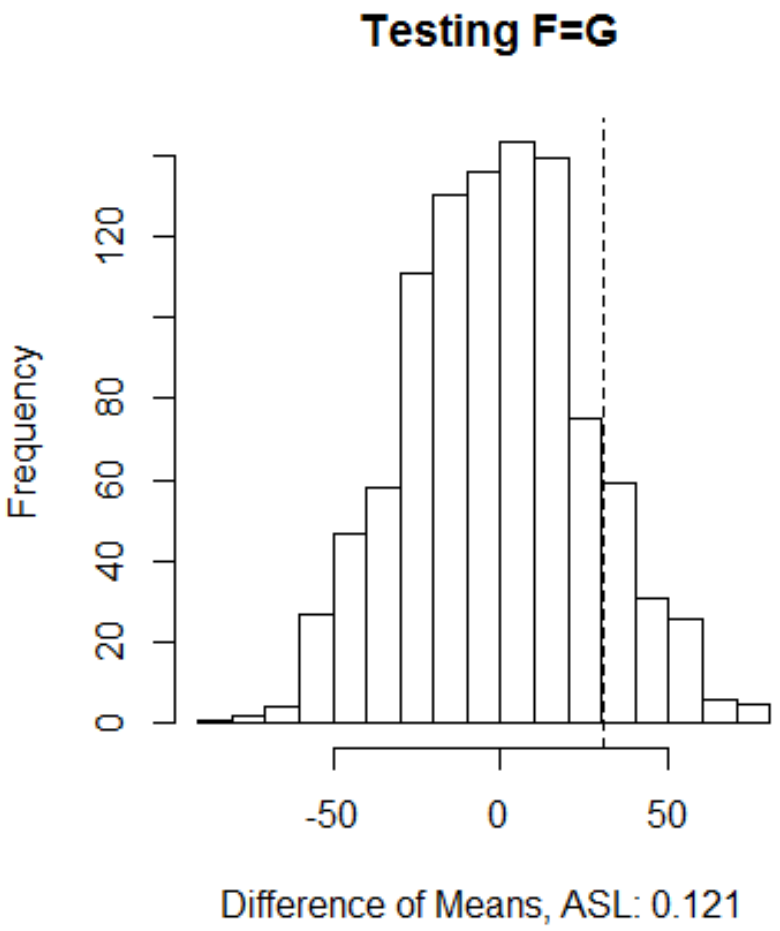
for testing $H_0: F = G$ on the mouse data. The proportion of values greater than 30.63 is .121.

Calculating

$$t(\boldsymbol{x}) = \frac{\bar{z} - \bar{y}}{\bar{\sigma}\sqrt{1/n + 1/m}}$$

(approximate pivotal) for the same replications produced

$$\widehat{ASL}_{boot} = .128$$



Testing Equality of Means

- Instead of testing $H_0: F = G$, we wish to test $\mathrm{H}_0: \mu_z = \mu_y$, without assuming equal variances. We need estimates of F and G that use only the assumption of common mean

1. Define points $\tilde{z}_i = z_i - \bar{z} + \bar{x}, i = 1, \ldots, n$, and $\tilde{y}_i = y_i - \bar{y} + \bar{x}, i = 1, \ldots, m$. The empirical distributions of $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{y}}$ shares a common mean.
2. Draw $\boldsymbol{B}$ bootstrap samples with replacement $(\boldsymbol{z}^*, \boldsymbol{y}^*)$ from $\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_n$ and $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_m$ respectivly
3. Evaluate $t(\cdot)$ on each sample -

$$t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right) = \frac{\bar{z}^* - \bar{y}^*}{\sqrt{\bar{\sigma}_z^*\, 1/n + \bar{\sigma}_y^*\, 1/m}}$$

4. Approximate ASL_{boot} by

$$\widehat{ASL}_{boot} = \#\left\{t\left(\boldsymbol{x}^{*\boldsymbol{b}}\right) \geq t(\boldsymbol{x})\right\}/B$$

Permutation Test VS Bootstrap Hypothesis Testing

- Accuracy: In the two-sample problem, ASL_{perm} is the exact probability of obtaining a test statistic as extreme as the one observed. In contrast, the bootstrap explicitly samples from estimated probability mechanism. $\widehat{ASL}_{boot}$ has no interpretation as an exact probability.
- Flexibility: When special symmetry isn't required, the bootstrap testing can be applied much more generally than the permutation test. (Like in the two sample problem – permutation test is limited to $H_0: F = G$, or in the one-sample problem)

The One-Sample Problem

We observe a random sample:

$$F \rightarrow \quad \boldsymbol{z} = (z_1, z_2, \ldots, z_n)$$

And we wish to test whether the mean of the population equals to some predetermine value μ_0 −

$$H_0: \mu_z = \mu_0$$

Bootstrap Hypothesis Testing $\mu_z = \mu_0$

What is the appropriate way to estimate the null distribution?

The empirical distribution $\hat{F}$ is not an appropriate estimation, because it does not obey H_0.

As before, we can use the empirical distribution of the points:

$$\tilde{z}_i = z_i - \bar{z} + \mu_0, i = 1, \dots, n$$

Which has a mean of μ_0.

Bootstrap Hypothesis Testing $\mu_z = \mu_0$

The test will be based on the approximate distribution of the test statistic $t(\boldsymbol{z}) = \frac{\bar{z}-\mu_0}{\bar{\sigma}/\sqrt{n}}$

We sample $\boldsymbol{B}$ times $\tilde{z}_1^*, \ldots, \tilde{z}_n^*$ with replacement from $\tilde{z}_1, \ldots, \tilde{z}_n$, and for each sample compute

$$t(\tilde{\boldsymbol{z}}^*) = \frac{\bar{\tilde{z}} - \mu_0}{\bar{\tilde{\sigma}}/\sqrt{n}}$$

And the estimated ASL is given by

$$\widehat{ASL}_{boot} = \#\left\{t\left(\tilde{\boldsymbol{z}}^{*\boldsymbol{b}}\right) \geq t(\boldsymbol{z})\right\}/B$$

* In the case that **large** values of $t\left(\tilde{\boldsymbol{z}}^{*\boldsymbol{b}}\right)$ are evidence against H_0

Testing $\mu_z = \mu_0$ on the Mouse Data

Taking $\mu_0 = 129$, the observed value of the test statistic is

$$t(\boldsymbol{z}) = \frac{86.9 - 129}{66.8/\sqrt{7}} = -1.67$$

(When estimating σ with the unbiased estimator for standard deviation). For 94 of 1000 bootstrap samples, $t(\tilde{\boldsymbol{z}}^*)$ was smaller than -1.67, and therefor

$$\widehat{ASL}_{boot} = .094$$

For reference, the student's t-test result for the same null hypothesis on that data gives us

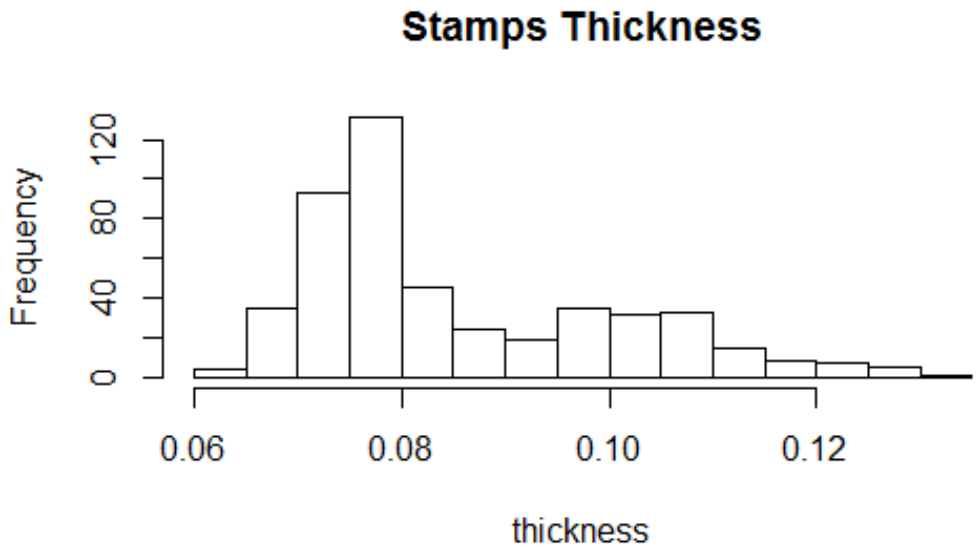
$$ASL = Prob\left\{t_6 < -\frac{42.1}{66.8/\sqrt{7}}\right\} = 0.07$$

Testing Multimodality of a Population

A mode is defined to be a local maximum or “bump” of the population density

The data: $x_1, \ldots, x_{485}$ Mexican stamps’ thickness from 1872.

The number of modes is suggestive of the number of distinct type of paper used in the printing.



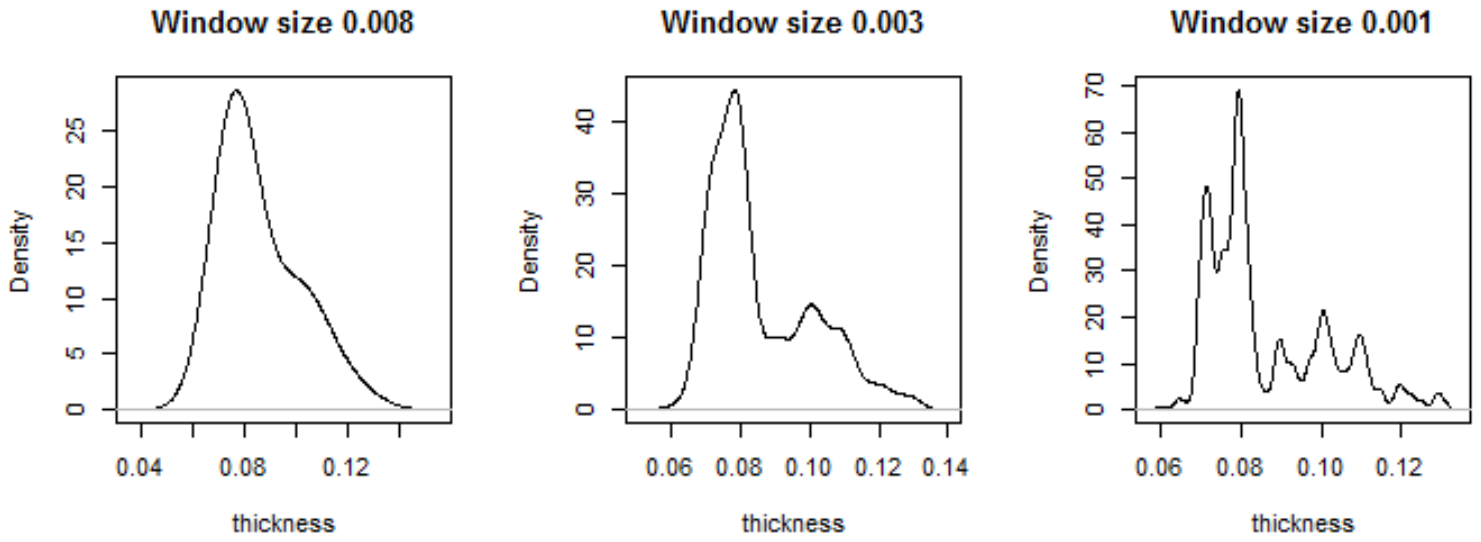
Testing Multimodality of a Population

Since the histogram is not smooth, it is difficult to tell from it whether there are more than one mode.

A *Gaussian kernel density* with window size h estimate can be used in order to obtain a smoother estimate:

$$\hat{f}(t;h) = \frac{1}{nh}\sum_{i=1}^{n}\phi\left(\frac{t-x_i}{h}\right)$$

As h increases, the number of modes in the density estimate is non-increasing

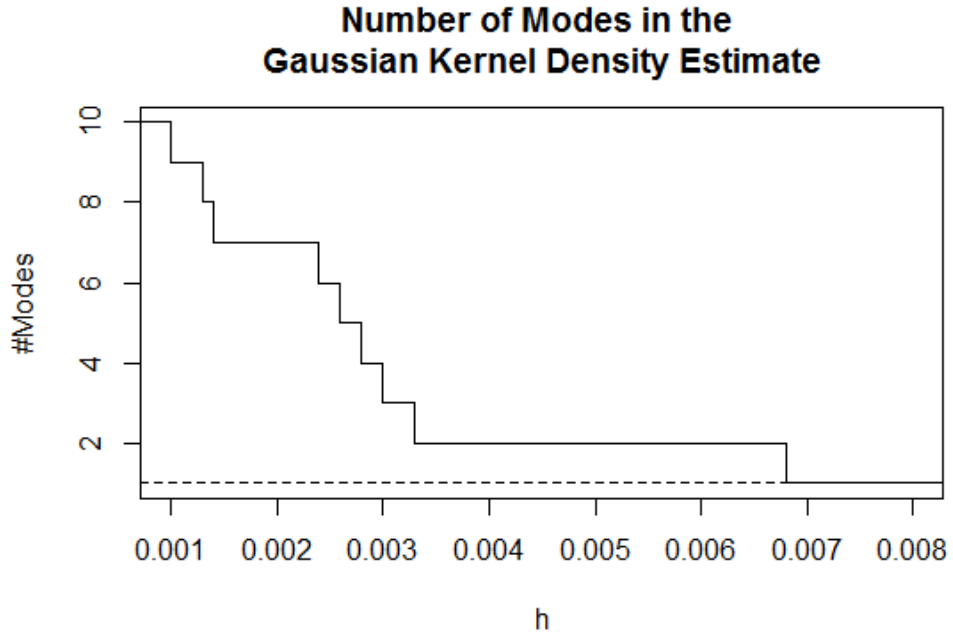


Testing Multimodality of a Population

The null hypothesis:

$$H_0: number\ of\ modes = 1$$

Versus $number\ of\ modes > 1$. Since the number of modes decreases as h increases, there is a smallest value of h such that $\hat{f}(t;h)$ has one mode. Call it $\hat{h}_1$. In our case, $\hat{h}_1 \approx .0068$.



Testing Multimodality of a Population

It seems reasonable to use $\hat{f}(t; \hat{h}_1)$ as the estimated null distribution for our test of H_0. It is the density estimate that uses least amount of smoothing among all estimated with one mode (conservative).

A small adjustment to $\hat{f}$ is needed because the formula artificially increases the variance of the estimate with $\hat{h}_1^2$. Let $\hat{g}(\cdot; \hat{h}_1)$ be the rescale estimate, that imposes variance equal to the sample variance.

A natural choice for a test statistic is $\hat{h}_1$ - a large value of $\hat{h}_1$ is evidence against H_0.

Putting all of this together, the achieved significance level is

$$ASL_{boot} = Prob_{\hat{g}(\cdot;\hat{h}_1)}\{\hat{h}_1^* > \hat{h}_1\}$$

Where each bootstrap sample $\boldsymbol{x}^*$ is drawn from $\hat{g}(\cdot; \hat{h}_1)$

Testing Multimodality of a Population

The sampling from $\hat{g}(\cdot; \hat{h}_1)$ is given by:

$$x_i^* = \bar{x} + \left(1 + \hat{h}_1^2/\hat{\sigma}^2\right)^{-\frac{1}{2}}\left(y_i^* - \bar{x} + \hat{h}_1\epsilon_i\right); i = 1, \dots, n$$

Where $y_1^*, \dots, y_n^*$ are sampled with replacement from $x_1, \dots, x_n$, and ϵ_i are standard normal random variables. (called *smoothed bootstrap*)

In the stamps data, out of 500 bootstrap samples, none had $\hat{h}_1^* > .0068$, so $\widehat{ASL}_{boot} = 0$.

The results can be interpreted in sequential manner, moving on to higher values of the least amount of modes. (Silverman 1981)

When testing the same for $H_0: number\ of\ modes = 2$, 146 samples out of 500 had $\hat{h}_2^* > .0033$, which translates to $\widehat{ASL}_{boot} = 0.292$.

In our case, the inference process will end here.

Summary

A bootstrap hypothesis test is carried out using the followings:

a) A test statistic $t(\boldsymbol{x})$

b) An approximate null distribution $\hat{F}_0$ for the data under H_0

Given these, we generate B bootstrap values of $t(\boldsymbol{x}^*)$ under $\hat{F}_0$ and estimate the achieved significance level by

$$\widehat{ASL}_{boot} = \#\{t(\boldsymbol{x}^{*\boldsymbol{b}}) \geq t(\boldsymbol{x})\}/B$$

The choice of test statistic $t(\boldsymbol{x})$ and the estimate of the null distribution will determine the power of the test. In the stamp example, if the actual population density is bimodal, but the Gaussian kernel density does not approximate it accurately, then the suggested test will not have high power.

Bootstrap tests are useful when the alternative hypothesis is not well specified. In cases where there is parametric alternative hypothesis, likelihood or Bayesian methods might be preferable.