QCB 508 – Week 8

John D. Storey

Spring 2017

Contents

Nonparametric Statistics

Parametric Inference

Parametric inference is based on a family of known probability distributions governed by a defined parameter space.

The goal is to perform inference (or more generally statistics) on the values of the parameters.

Nonparametric Inference

Nonparametric inference or modeling can be described in two ways (not mutually exclusive):

1. An inference procedure or model that does not depend on or utilize the parametrized probability distribution from which the data are generated.
2. An inference procedure or model that may have a specific structure or based on a specific formula, but the complexity is adaptive and can grow to arbitrary levels of complexity as the sample size grows.

In *All of Nonparametric Statistics*, Larry Wasserman says:

> ... it is difficult to give a precise definition of nonparametric inference.... For the purposes of this book, we will use the phrase nonparametric inference to refer to a set of modern statistical methods that aim to keep the number of underlying assumptions as weak as possible.

He then lists five estimation examples (see Section 1.1): distributions, functionals, densities, regression curves, and Normal means.

Nonparametric Descriptive Statistics

Almost all of the exploratory data analysis methods we covered in the beginning of the course are nonparametric.

Sometimes the exploratory methods are calibrated by known probability distributions, but they are usually informative regardless of the underlying probability distribution (or lack thereof) of the data.

Semiparametric Inference

Semiparametric inference or modeling methods contain both parametric and nonparametric components.

An example is $X_i | \mu_i \sim \text{Normal}(\mu_i, 1)$ and $\mu_i \overset{\text{iid}}{\sim} F$ for some arbitrary distribution F.

Topics This Week

- Empirical distribution functions
- Bootstrap
- Permutation methods
- Goodness of fit
- Method of moments
- Semiparametric empirical Bayes

Empirical Distribution Functions

Definition

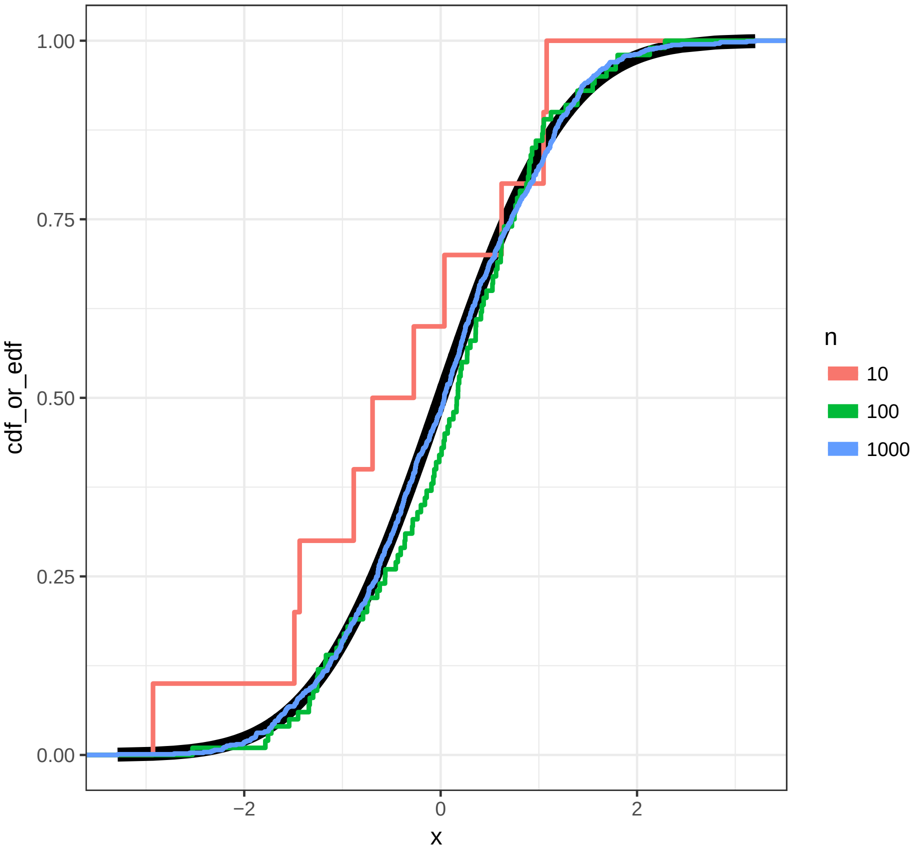
Suppose $X_1, X_2, \ldots, X_n \sim F$. The **empirical distribution function** (edf) – or **empirical cumulative distribution function** (ecdf) – is the distribution that puts probability $1/n$ on each observed value X_i.

Let $1(X_i \leq y) = 1$ if $X_i \leq y$ and $1(X_i \leq y) = 0$ if $X_i > y$.

$$\text{Random variable: } \hat{F}_{\boldsymbol{X}}(y) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq y)$$

$$\text{Observed variable: } \hat{F}_{\boldsymbol{x}}(y) = \frac{1}{n} \sum_{i=1}^{n} 1(x_i \leq y)$$

Example: Normal



Pointwise Convergence

Under our assumptions, by the strong law of large numbers for each $y \in \mathbb{R}$,

$$\hat{F}_{\boldsymbol{X}}(y) \xrightarrow{\text{a.s.}} F(y)$$

as $n \to \infty$.

Glivenko-Cantelli Theorem

Under our assumptions, we can get a much stronger convergence result:

$$\sup_{y \in \mathbb{R}} \left| \hat{F}_{\boldsymbol{X}}(y) - F(y) \right| \xrightarrow{\text{a.s.}} 0$$

as $n \to \infty$. Here, "sup" is short for *supremum*, which is a mathematical generalization of *maximum*.

This result says that even the worst difference between the edf and the true cdf converges with probability 1 to zero.

Dvoretzky-Kiefer-Wolfowitz (DKW) Inequality

This result gives us an upper bound on how far off the edf is from the true cdf, which allows us to construct confidence bands about the edf.

$$\Pr\left(\sup_{y\in\mathbb{R}}\left|\hat{F}_{\boldsymbol{X}}(y) - F(y)\right| > \epsilon\right) \leq 2\exp -2n\epsilon^2$$

As outlined in *All of Nonparametric Statistics*, setting

$$\epsilon_n = \sqrt{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)}$$

$$L(y) = \max\{\hat{F}_{\boldsymbol{X}}(y) - \epsilon_n, 0\}$$

$$U(y) = \min\{\hat{F}_{\boldsymbol{X}}(y) + \epsilon_n, 1\}$$

guarantees that $\Pr(L(y) \leq F(y) \leq U(y) \text{ for all } y) \geq 1-\alpha$.

Statistical Functionals

A **statistical functional** $T(F)$ is any function of F. Examples:

- $\mu(F) = \int x dF(x)$
- $\sigma^2(F) = \int (x-\mu(F))^2 dF(x)$
- $\text{median}(F) = F^{-1}(1/2)$

A **linear statistical functional** is such that $T(F) = \int a(x)dF(x)$.

Plug-In Estimator

A plug-in estimator of $T(F)$ based on the edf is $T(\hat{F}_{\boldsymbol{X}})$. Examples:

- $\hat{\mu} = \mu(\hat{F}_{\boldsymbol{X}}) = \int x\hat{F}_{\boldsymbol{X}}(x) = \frac{1}{n}\sum_{i=1}^n X_i$
- $\hat{\sigma}^2 = \sigma^2(\hat{F}_{\boldsymbol{X}}) = \int (x-\hat{\mu})^2\hat{F}_{\boldsymbol{X}}(x) = \frac{1}{n}\sum_{i=1}^n (X_i-\hat{\mu})^2$
- $\text{median}(\hat{F}_{\boldsymbol{X}}) = \hat{F}_{\boldsymbol{X}}^{-1}(1/2)$

EDF Standard Error

Suppose that $T(F) = \int a(x)dF(x)$ is a linear functional. Then:

$$\begin{aligned}
\mathrm{Var}(T(\hat{F}_{\boldsymbol{X}})) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(a(X_i)) = \frac{\mathrm{Var}_F(a(X))}{n} \\
\mathrm{se}(T(\hat{F}_{\boldsymbol{X}})) &= \sqrt{\frac{\mathrm{Var}_F(a(X))}{n}} \\
\hat{\mathrm{se}}(T(\hat{F}_{\boldsymbol{X}})) &= \sqrt{\frac{\mathrm{Var}_{\hat{F}_{\boldsymbol{X}}}(a(X))}{n}}
\end{aligned}$$

Note that

$$\mathrm{Var}_F(a(X)) = \int (a(x) - T(F))^2 dF(x)$$

because $T(F) = \int a(x)dF(x) = \mathrm{E}_F[a(X)]$. Likewise,

$$\mathrm{Var}_{\hat{F}_{\boldsymbol{X}}}(a(X)) = \sum_{i=1}^{n}(a(X_i) - T(\hat{F}_{\boldsymbol{X}}))^2$$

where $T(\hat{F}_{\boldsymbol{X}}) = \frac{1}{n}\sum_{i=1}^{n} a(X_i)$.

EDF CLT

Suppose that $\mathrm{Var}_F(a(X)) < \infty$. Then we have the following convergences as $n \rightarrow \infty$:

$$\frac{\mathrm{Var}_{\hat{F}_{\boldsymbol{X}}}(a(X))}{\mathrm{Var}_F(a(X))} \xrightarrow{P} 1 \ , \ \frac{\hat{\mathrm{se}}(T(\hat{F}_{\boldsymbol{X}}))}{\mathrm{se}(T(\hat{F}_{\boldsymbol{X}}))} \xrightarrow{P} 1$$

$$\frac{T(F) - T(\hat{F}_{\boldsymbol{X}})}{\hat{\mathrm{se}}(T(\hat{F}_{\boldsymbol{X}}))} \xrightarrow{D} \mathrm{Normal}(0, 1)$$

The estimators are very easy to calculate on real data, so this a powerful set of results.

Bootstrap

Rationale

Suppose $X_1, X_2, \ldots, X_n \sim F$. If the edf $\hat{F}_{\boldsymbol{X}}$ is an accurate approximation for the true cdf F, then we can utilize $\hat{F}_{\boldsymbol{X}}$ in place of F to nonparametrically characterize the sampling distribution of a statistic $T(\boldsymbol{X})$.

This allows for the sampling distribution of more general statistics to be considered, such as the median or a percentile, as well as more traditional statistics, such as the mean, when the underlying distribution is unknown.

When we encounter modeling fitting, the bootstrap may be very useful for characterizing the sampling distribution of complex statistics we calculate from fitted models.

Big Picture

We calculate $T(\boldsymbol{x})$ on the observed data, and we also form the edf, $\hat{F}_{\boldsymbol{x}}$.

To approximate the sampling distribution of $T(\boldsymbol{X})$ we generate B random samples of n iid data points from $\hat{F}_{\boldsymbol{x}}$ and calculate $T(\boldsymbol{x}^{*(b)})$ for each bootstrap sample $b = 1, 2, \ldots, B$ where $\boldsymbol{x}^{*(b)} = (x_1^{*(b)}, x_2^{*(b)}, \ldots, x_n^{*(b)})^T$.

Sampling $X_1^*, \ldots, X_n^* \overset{\text{iid}}{\sim} \hat{F}_{\boldsymbol{x}}$ is accomplished by sampling n times *with replacement* from the observed data $x_1, x_2, \ldots, x_n$.

This means $\Pr\left(X^* = x_j\right) = \frac{1}{n}$ for all j.

Bootstrap Variance

For each bootstrap sample $\boldsymbol{x}^{*(b)} = (x_1^{*(b)}, x_2^{*(b)}, \ldots, x_n^{*(b)})^T$, calculate bootstrap statistic $T(\boldsymbol{x}^{*(b)})$.

Repeat this for $b = 1, 2, \ldots, B$.

Estimate the sampling variance of $T(\boldsymbol{x})$ by

$$\hat{\text{Var}}(T(\boldsymbol{x})) = \frac{1}{B} \sum_{b=1}^{B} \left(T\left(\boldsymbol{x}^{*(b)}\right) - \frac{1}{B} \sum_{k=1}^{B} T\left(\boldsymbol{x}^{*(k)}\right) \right)^2$$

Caveat

Why haven't we just been doing this the entire time?!

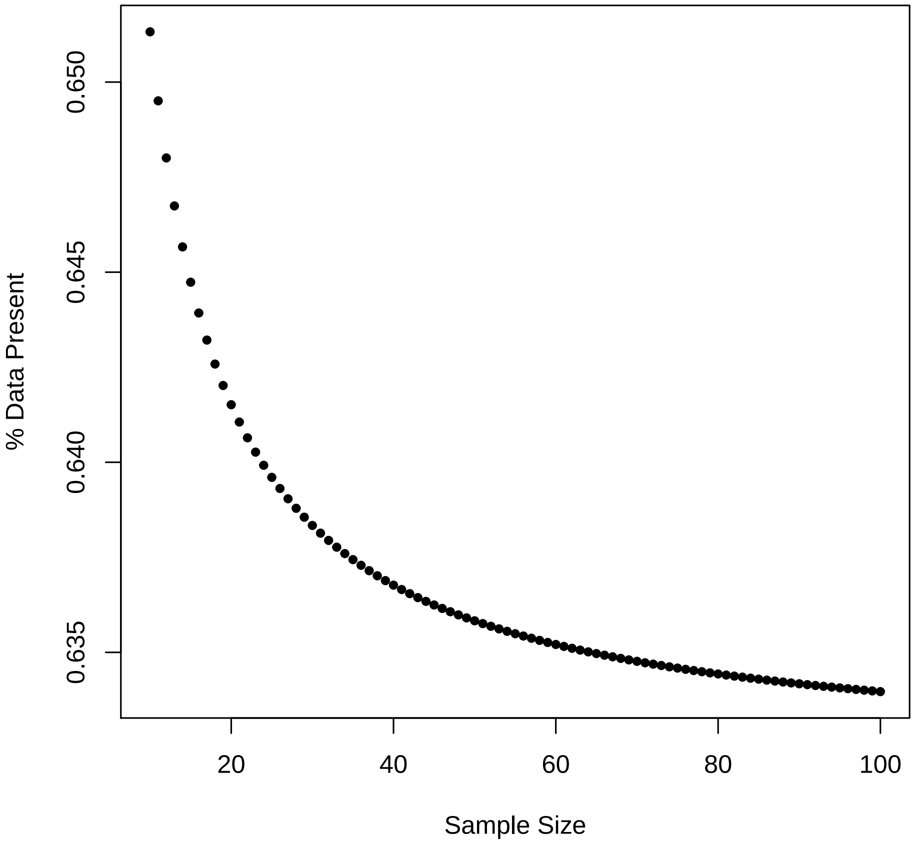
In *All of Nonparametric Statistics*, Larry Wasserman states:

> There is a tendency to treat the bootstrap as a panacea for all problems. But the bootstrap requires regularity conditions to yield valid answers. It should not be applied blindly.

The bootstrap is easy to motivate, but it is quite tricky to implement outside of the very standard problems. It sometimes requires deeper knowledge of statistical theory than likelihood-based inference.

Bootstrap Sample

For a sample of size n, what percentage of the data is present in any given bootstrap sample?



Bootstrap CIs

Suppose that $\theta = T(F)$ and $\hat{\theta} = T(\hat{F}_{\boldsymbol{x}})$.

We can use the bootstrap to generate data from $T(\hat{F}_{\boldsymbol{x}})$.

For $b = 1, 2, \ldots, B$, we draw $x_1^{*(b)}, x_2^{*(b)}, \ldots, x_n^{*(b)}$ as iid realiztions from $T(\hat{F}_{\boldsymbol{x}})$, and calculate $\hat{\theta}^{*(b)} = T(\hat{F}_{\boldsymbol{x}^{*(b)}})$.

Let p_α^* be the α percentile of $\left\{\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, \ldots, \hat{\theta}^{*(B)}\right\}$.

Let's discuss several ways of calculating confidence intervals for $\theta = T(F)$.

Invoking the CLT

If we have evidence that the central limit theorem can be applied, we can form the $(1-\alpha)$ CI as:

$$(T(\hat{F}_{\boldsymbol{x}}) - |z_{\alpha/2}| \, \mathrm{se}^*, T(\hat{F}_{\boldsymbol{x}}) + |z_{\alpha/2}| \, \mathrm{se}^*)$$

where se* is the bootstrap standard error calculated as

$$\mathrm{se}^* = \sqrt{\frac{1}{B} \sum_{b=1}^{B} \left(T\left(\hat{F}_{\boldsymbol{x}^{*(b)}}\right) - \frac{1}{B} \sum_{k=1}^{B} T\left(\hat{F}_{\boldsymbol{x}^{*(k)}}\right) \right)^2}.$$

Note that to get this confidence interval we need to justify that the following pivotal statistics are approximately Normal(0,1):

$$\frac{T(\hat{F}_{\boldsymbol{x}}) - T(F)}{\mathrm{se}(T(\hat{F}_{\boldsymbol{x}}))} \approx \frac{T(\hat{F}_{\boldsymbol{x}}) - T(F)}{\mathrm{se}^*}$$

Percentile Interval

If a *monotone* function $m(\cdot)$ exists so that $m\left(\hat{\theta}\right) \sim \mathrm{Normal}(m(\theta), b^2)$, then we can form the $(1-\alpha)$ CI as:

$$\left(p_{\alpha/2}^*, p_{1-\alpha/2}^*\right)$$

where recall that in general p_α^* is the α percentile of bootstrap estimates $\left\{\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, \ldots, \hat{\theta}^{*(B)}\right\}$

Pivotal Interval

Suppose we can calculate percentiles of $\hat{\theta} - \theta$, say q_α. Note that the α percentile of $\hat{\theta}$ is $q_\alpha + \theta$. The $1 - \alpha$ CI is

$$(\hat{\theta} - q_{1-\alpha/2}, \hat{\theta} - q_{\alpha/2})$$

which comes from:

$$\begin{aligned}
1 - \alpha &= \Pr(q_{\alpha/2} \leq \hat{\theta} - \theta \leq q_{1-\alpha/2}) \\
&= \Pr(-q_{1-\alpha/2} \leq \theta - \hat{\theta} \leq -q_{\alpha/2}) \\
&= \Pr(\hat{\theta} - q_{1-\alpha/2} \leq \theta \leq \hat{\theta} - q_{\alpha/2})
\end{aligned}$$

Suppose the sampling distribution of $\hat{\theta}^* - \hat{\theta}$ is an approximation for that of $\hat{\theta} - \theta$.

If p^*_α is the α percentile of $\hat{\theta}^*$ then, $p^*_\alpha - \hat{\theta}$ is the α percentile of $\hat{\theta}^* - \hat{\theta}$.

Therefore, $p^*_\alpha - \hat{\theta}$ is the bootstrap estimate of q_α. Plugging this into $(\hat{\theta} - q_{1-\alpha/2}, \hat{\theta} - q_{\alpha/2})$, we get the following $(1 - \alpha)$ bootstrap CI:

$$\left(2\hat{\theta} - p^*_{1-\alpha/2}, 2\hat{\theta} - p^*_{\alpha/2}\right).$$

Studentized Pivotal Interval

In the previous scenario, we needed to assume that the sampling distribution of $\hat{\theta}^* - \hat{\theta}$ is an approximation for that of $\hat{\theta} - \theta$. Sometimes this will not be the case and instead we can studentize this pivotal quantity. That is, the distribution of

$$\frac{\hat{\theta} - \theta}{\hat{\text{se}}\left(\hat{\theta}\right)}$$

is well-approximated by that of

$$\frac{\hat{\theta}^* - \hat{\theta}}{\hat{\text{se}}\left(\hat{\theta}^*\right)}.$$

Let z^*_α be the α percentile of

$$\left\{ \frac{\hat{\theta}^{*(1)} - \hat{\theta}}{\hat{\text{se}}\left(\hat{\theta}^{*(1)}\right)}, \ldots, \frac{\hat{\theta}^{*(B)} - \hat{\theta}}{\hat{\text{se}}\left(\hat{\theta}^{*(B)}\right)} \right\}.$$

Then a $(1 - \alpha)$ bootstrap CI is

$$\left(\hat{\theta} - z^*_{1-\alpha/2}\hat{\text{se}}\left(\hat{\theta}\right), \hat{\theta} - z^*_{\alpha/2}\hat{\text{se}}\left(\hat{\theta}\right)\right).$$

Exercise: Why?

Bootstrap Hypothesis Testing

As we have seen, hypothesis testing and confidence intervals are very related. For a simple null hypothesis, a bootstrap hypothesis test p-value can be calculated by finding the minimum α for which the $(1-\alpha)$ CI does not contain the null hypothesis value. You showed this on your homework.

The general approach is to calculate a test statistic based on the observed data. Then the null distribution of this statistic is approximated by forming bootstrap test statistics under the scenario that the null hypothesis is true. This can often be accomplished because the $\hat{\theta}$ estimated from the observed data is the *population* parameter from the bootstrap distribution.

Example: *t*-test

Suppose $X_1, X_2, \ldots, X_m \sim F_X$ and $Y_1, Y_2, \ldots, Y_n \sim F_Y$. We wish to test $H_0 : \mu(F_X) = \mu(F_Y)$ vs $H_1 : \mu(F_X) \neq \mu(F_Y)$. Suppose that we know $\sigma^2(F_X) = \sigma^2(F_Y)$ (if not, it is straightforward to adjust the proecure below).

Our test statistic is

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) s^2}}$$

where s^2 is the pooled sample variance.

Note that the bootstrap distributions are such that $\mu(\hat{F}_{X^*}) = \overline{x}$ and $\mu(\hat{F}_{Y^*}) = \overline{y}$. Thus we want to center the bootstrap t-statistics about these known means.

Specifically, for a bootstrap data set $x^* = (x_1^*, x_2^*, \ldots, x_m^*)^T$ and $y^* = (y_1^*, y_2^*, \ldots, y_n^*)^T$, we form null t-statistic

$$t^* = \frac{\overline{x}^* - \overline{y}^* - (\overline{x} - \overline{y})}{\sqrt{\left(\frac{1}{m} + \frac{1}{n}\right) s^{2,*}}}$$

where again $s^{2,*}$ is the pooled sample variance.

In order to obtain a p-value, we calculate $t^{*(b)}$ for $b = 1, 2, \dots, B$ bootstrap data sets.

The p-value of t is then the proportion of bootstrap statistics as or more extreme than the observed statistic:

$$\text{p-value}(t) = \frac{1}{B} \sum_{b=1}^{B} 1 \left(|t^{*(b)}| \geq |t| \right).$$

Parametric Bootstrap

Suppose $X_1, X_2, \dots, X_n \overset{\text{iid}}{\sim} F_\theta$ for some parametric F_θ. We form estimate $\hat{\theta}$, but we don't have a known sampling distribution we can use to do inference with $\hat{\theta}$.

The parametric bootstrap generates bootstrap data sets from $F_{\hat{\theta}}$ rather than from the edf. It proceeds as we outlined above for these bootstrap data sets.

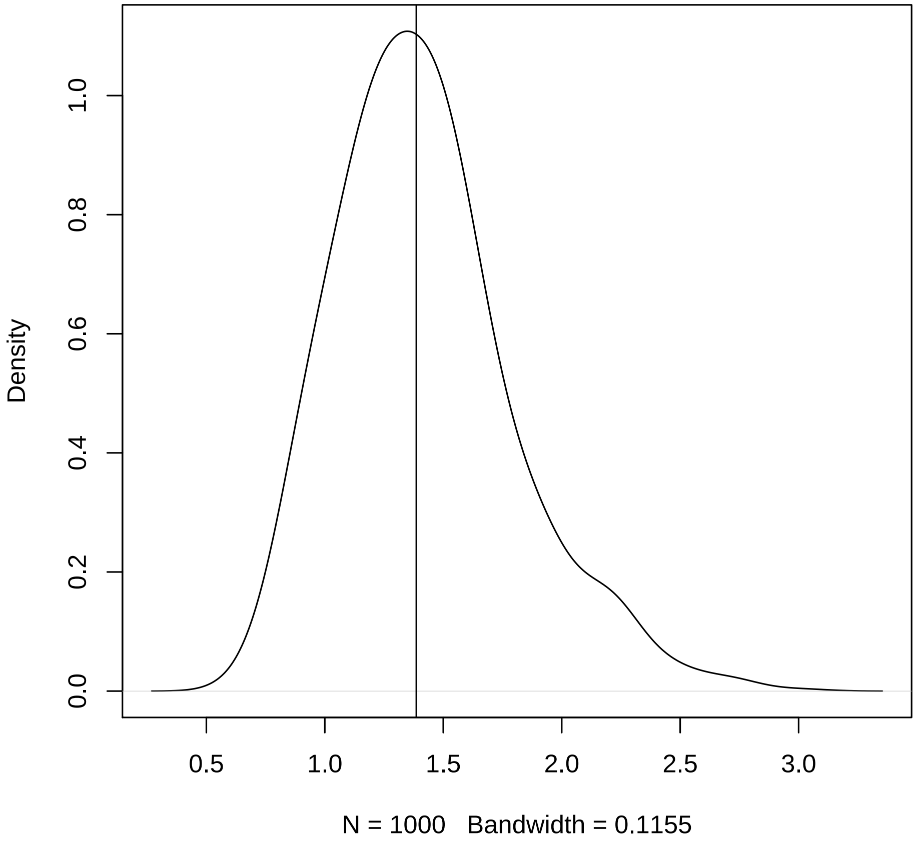
Example: Exponential Data

In the homework, you will be performing a bootstrap t-test of the mean and a bootstrap percentile CI of the median for the following Exponential(λ) data:

```
> set.seed(1111)
> pop.mean <- 2
> X <- matrix(rexp(1000*30, rate=1/pop.mean), nrow=1000, ncol=30)
```

Let's construct a pivotal bootstrap CI of the median here instead.

```
> # population median 2*log(2)
> pop_med <- qexp(0.5, rate=1/pop.mean); pop_med
[1] 1.386294
>
> obs_meds <- apply(X, 1, median)
> plot(density(obs_meds, adj=1.5), main=" "); abline(v=pop_med)
```

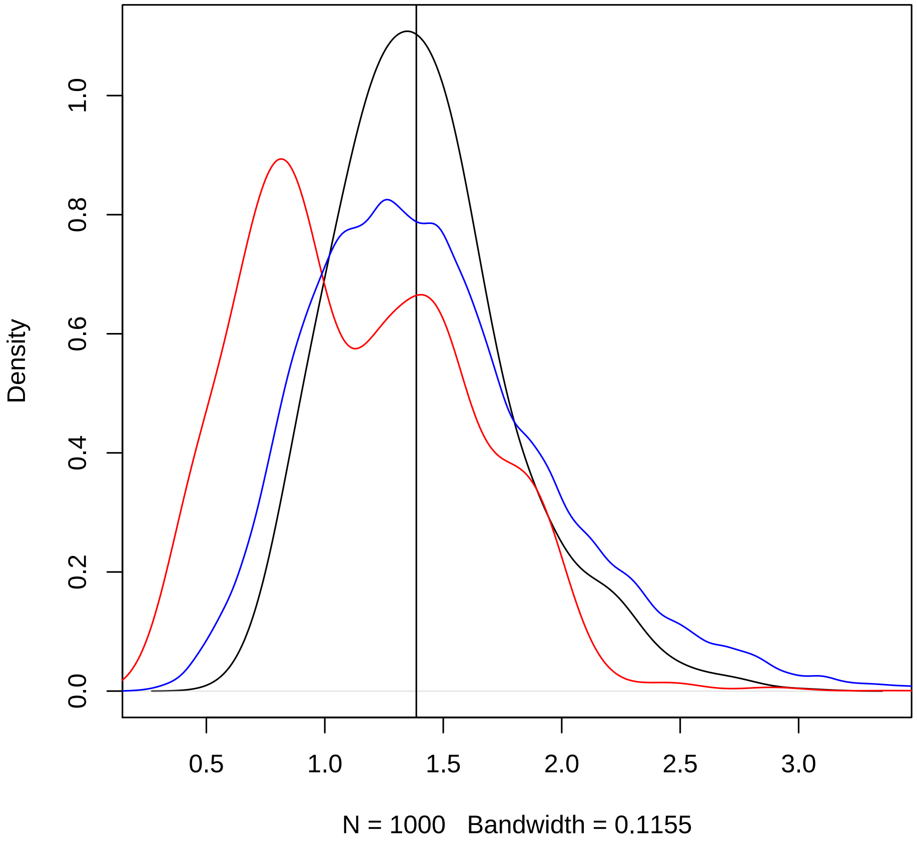


Some embarrassingly inefficient code to calculate bootstrap medians.

```
> B <- 1000
> boot_meds <- matrix(0, nrow=1000, ncol=B)
>
> for(b in 1:B) {
+   idx <- sample(1:30, replace=TRUE)
+   boot_meds[,b] <- apply(X[,idx], 1, median)
+ }
```

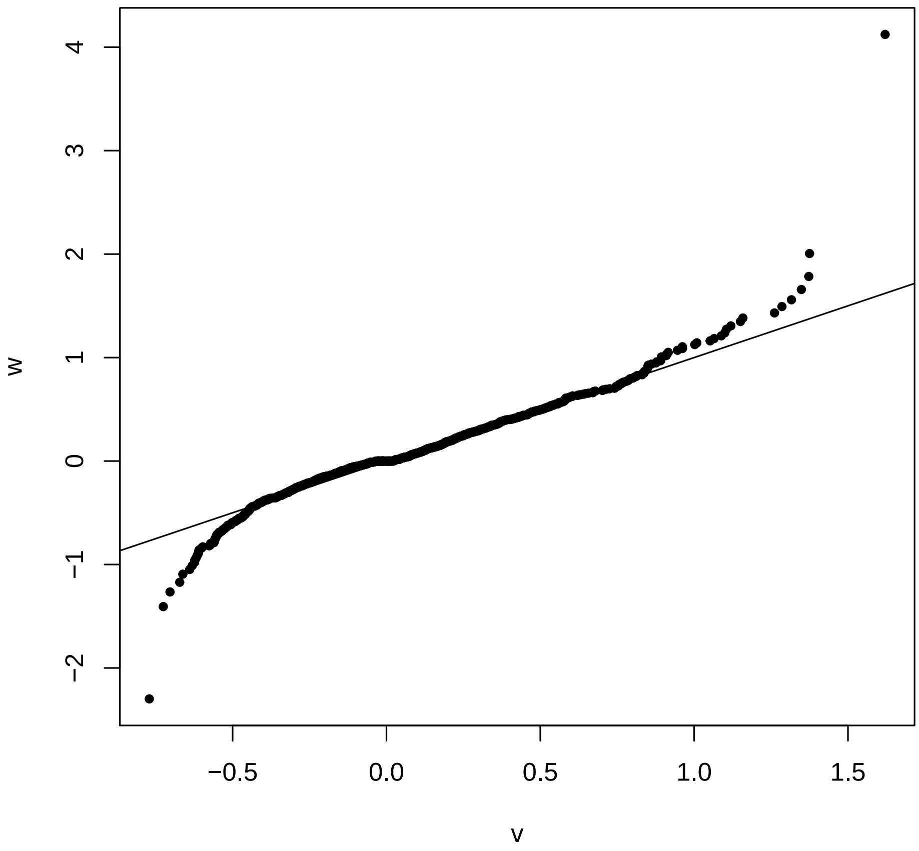
Plot the bootstrap medians.

```
> plot(density(obs_meds, adj=1.5), main=" "); abline(v=pop_med)
> lines(density(as.vector(boot_meds[1:4,]), adj=1.5), col="red")
> lines(density(as.vector(boot_meds), adj=1.5), col="blue")
```



Compare sampling distribution of $\hat{\theta} - \theta$ to $\hat{\theta}^* - \hat{\theta}$.

```
> v <- obs_meds - pop_med
> w <- as.vector(boot_meds - obs_meds)
> qqplot(v, w, pch=20); abline(0,1)
```

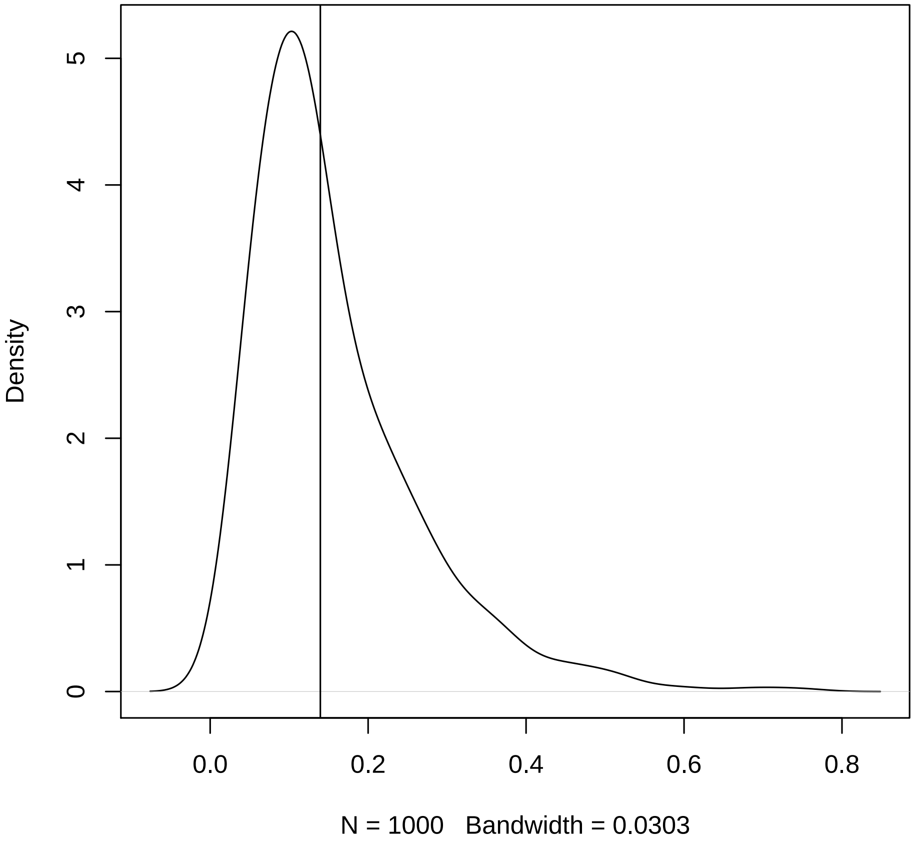


Does a 95% bootstrap pivotal interval provide coverage?

```
> ci_lower <- apply(boot_meds, 1, quantile, probs=0.975)
> ci_upper <- apply(boot_meds, 1, quantile, probs=0.025)
>
> ci_lower <- 2*obs_meds - ci_lower
> ci_upper <- 2*obs_meds - ci_upper
>
> ci_lower[1]; ci_upper[1]
[1] 0.8958224
[1] 2.113859
>
> cover <- (pop_med >= ci_lower) & (pop_med <= ci_upper)
> mean(cover)
[1] 0.809
>
> # :-(
```

Let’s check the bootstrap variances.

```
> sampling_var <- var(obs_meds)
> boot_var <- apply(boot_meds, 1, var)
> plot(density(boot_var, adj=1.5), main=" ")
> abline(v=sampling_var)
```



Bootstrap Subtleties

Comment

Normal Data

Verify Sampling Distribution

Bootstrap Columns

Bootstrap Rows

Studentize Rows

Permutation Methods

Rationale

Permutation Test

Wilcoxon Signed Rank-Sum Test

Wilcoxon Rank Sum Test

Permutation *t*-test

Goodness of Fit

Rationale

Chi-Square GoF Test

Example

Chi-Square Contingency Test

Example

Kolmogorov–Smirnov Test

Example

Method of Moments

Definition

Example: Normal

Goodness of Fit

Semiparametric Empirical Bayes

Source Code

Session Information

```
> sessionInfo()
R version 3.3.2 (2016-10-31)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: macOS Sierra 10.12.4

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
 [1] dplyr_0.5.0     purrr_0.2.2     readr_1.0.0
 [4] tidyr_0.6.1     tibble_1.2      ggplot2_2.2.1
 [7] tidyverse_1.1.1 knitr_1.15.1    magrittr_1.5
[10] devtools_1.12.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.9      plyr_1.8.4       forcats_0.2.0
 [4] tools_3.3.2      digest_0.6.12    lubridate_1.6.0
 [7] jsonlite_1.2     evaluate_0.10    memoise_1.0.0
[10] nlme_3.1-131     gtable_0.2.0     lattice_0.20-34
[13] psych_1.6.12     DBI_0.5-1        yaml_2.1.14
[16] parallel_3.3.2   haven_1.0.0      xml2_1.1.1
[19] withr_1.0.2      stringr_1.1.0    httr_1.2.1
[22] hms_0.3          rprojroot_1.2    grid_3.3.2
[25] R6_2.2.0         readxl_0.1.1     foreign_0.8-67
[28] rmarkdown_1.3    modelr_0.1.0     reshape2_1.4.2
[31] backports_1.0.5  scales_0.4.1     htmltools_0.3.5
[34] rvest_0.3.2      assertthat_0.1   mnormt_1.5-5
[37] colorspace_1.3-2 labeling_0.3     stringi_1.1.2
[40] lazyeval_0.2.0   munsell_0.4.3    broom_0.4.2
```