# 5601 Notes: The Subsampling Bootstrap

Charles J. Geyer

April 13, 2006

## 1 Web Pages

This handout accompanies the web pages

`http://www.stat.umn.edu/geyer/5601/examp/subboot.html`
`http://www.stat.umn.edu/geyer/5601/examp/subtoot.html`

## 2 History

The term "bootstrap" was coined by Efron (1979). He described both the nonparametric and parametric bootstrap. In particular, his nonparametric bootstrap is the procedure of resampling *with replacement* from the original sample *at the same sample size*, which is by far the most commonly used bootstrap procedure.

It wasn't long before people experimented with resampling at different sample sizes. But the key discovery in that area came later. Politis and Romano (1994) described resampling *without replacement* from the original sample *at smaller than the original sample size*.

This is different enough from Efron's idea that in their book (Politis, Romano, and Wolf, 1999) they don't call it "bootstrap" but just plain "subsampling".

Whatever you call it, here's why it is such an important innovation.

- Politis and Romano's subsampling bootstrap takes samples *without replacement* of size $b$ from the original sample of size $n$, generally with $b \ll n$ (read "$b$ much less than $n$"). Such samples are themselves samples of size $b$ from the true unknown distribution $F$ of the original sample.

- Efron's nonparametric bootstrap takes samples *with replacement* of size $n$ from the original sample of size $n$ (both sample sizes the same). Such samples are samples of size $n$ from the empirical distribution $\widehat{F}_n$ associated with the original sample.

Each of these procedures does the Wrong Thing.

- The Right Thing is samples of the right size $n$ from the right distribution $F$.
- The Politis and Romano thing is samples of the wrong size $b \ll n$ from the right distribution $F$.
- The Efron thing is samples of the right size $n$ from the wrong distribution $\widehat{F}_n$.

Both Wrong Things are wrong. We would like to do the Right Thing but we can't. (More precisely, we have exactly one such sample, the original data, and can't get more. Scientists may get more data, but that's of no interest to us statisticians.)

So which Wrong Thing do we want to do? Both have pluses and minuses. The Efron procedure is older, more widely used, and familiar to more people. It is also easier to use, at least in simple situations. But the Politis and Romano procedure has the great virtue of working in situations where the Efron bootstrap fails. The two main classes of such situations are presented in the following sections.

## 3 Stationary Time Series

A *time series* is a sequence $X_1$, $X_2$, ..., $X_n$ of possibly *dependent* random variables.

A times series is *stationary* if every consecutive block

$$X_{i+1}, X_{i+2}, \ldots, X_{i+b} \tag{1}$$

of length $b$ has the same (marginal) distribution. Roughly speaking, what actually happens changes over time, but the *probability distribution of what happens* does not change over time.

The ordinary (Efron) nonparametric bootstrap doesn't work for time series or any other form of dependent data. If the data are not IID then it makes no sense whatsoever to obtain IID bootstrap samples from $\widehat{F}_n$ (or any other distribution for that matter).

The (Politis and Romano) subsampling bootstrap does work for stationary time series. Under the stationarity assumption the $n - b - 1$ consecutive blocks (1) of length $b$ are identically distributed. Hence the estimators corresponding to such blocks

$$\theta_{bi}^* = s(X_{i+1}, X_{i+2}, \ldots X_{i+b})$$

are identically distributed (not IID just ID, since they are dependent) and analogous (in the "bootstrap world") to

$$\hat{\theta}_n = s(X_1, X_2, \ldots X_n)$$

(in the "real world"). The only problem we have to deal with is that

$$\operatorname{se}(\hat{\theta}_n) \approx \frac{c}{\sqrt{n}}$$

for some positive constant $c$ (assuming our estimator obeys the "square root law") whereas

$$\operatorname{se}(\theta_b^*) \approx \frac{c}{\sqrt{b}}$$

for the same positive constant $c$ (but different denominator). Thus we need to scale

$$\operatorname{se}(\hat{\theta}_n) \approx \operatorname{se}(\theta_b^*) \cdot \sqrt{\frac{b}{n}}$$

to get from $\operatorname{se}(\theta_b^*)$, which we can estimate by subsampling (as the standard deviation of the $n - b + 1$ quantities $\theta_{bi}^*$), to $\operatorname{se}(\hat{\theta}_n)$, which is the quantity we need to make a confidence interval for $\hat{\theta}_n$ and which we otherwise have no way to estimate.

# 4 IID Situations

## 4.1 Extreme Values

Suppose $X_1$, $X_2$, …, $X_n$ are IID Uniform$(0, \theta)$ random variables. Since the larger the sample the more the largest values crowd up against $\theta$, the natural estimator of $\theta$ is the maximum data value

$$\hat{\theta}_n = X_{(n)} = \max(X_1, X_2, \ldots, X_n).$$

This is in fact the maximum likelihood estimate.

The main statistical interest in this estimator is that it is a counter example to both the "square root law" and the "usual asymptotics" of maximum likelihood.

- The "rate" is $n$ rather than $\sqrt{n}$.
- The asymptotic distribution is not normal.

More precisely,

$$n(\theta - \hat{\theta}_n) \xrightarrow{\mathcal{D}} \text{Exponential}(1/\theta) \tag{2}$$

But to use the subsampling bootstrap, we need only know that the actual rate is $n$. We do not need to know the actual asymptotic distribution.

Actually, we do not even need to know the rate. By looking at the distribution of $\theta_b^*$ for different subsample sizes $b$ we can get an estimate of the rate (described in Section 6 below). But for now we'll assume we know the rate.

### 4.2 Other IID Situations

More generally, suppose $X_1$, $X_2$, ..., $X_n$ are IID from some distribution and we are interested in a parameter estimate

$$\hat{\theta}_n = s(X_1, X_2, \ldots, X_n).$$

And we assume that there is an asymptotic distribution at some "rate"

$$\tau_n(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \text{Something}, \tag{3}$$

where "Something" denotes any distribution whatsoever. Here $\tau_n$ can be any known sequence. Usually we have $\tau_n = \sqrt{n}$, in which case we say the estimator obeys the "square root law." But in the preceding section we saw an estimator for which we needed $\tau_n = n$. And on the homework we will see an estimator for which we will need $\tau_n = n^{1/3}$.

Under the IID assumption, the $\binom{n}{b}$ ways to choose a subsample of size $b$ without replacement from $X_1$, $X_2$, ..., $X_n$ are all identically distributed and the estimators

$$\theta_b^* = s(X_1^*, X_2^*, \ldots, X_b^*)$$

are analogous (in the "bootstrap world") to

$$\hat{\theta}_n = s(X_1, X_2, \ldots, X_n)$$

(in the "real world"). The only problem is that the distribution of $\theta_b^* - \theta$ differs from that of $\hat{\theta}_n - \theta$ by a factor of $\tau_b/\tau_n$. Hence, as we saw above, we need to rescale our standard errors by this factor. Or, as we will see below, we need to rescale any procedure we do.

## 5 Confidence Intervals

The fundamental idea of the subsampling bootstrap that some asymptotics (3) hold. It does not matter what the "Something" is (the asymptotic distribution of estimator) and it does not matter what the $\tau_n$ is (the rate). As long as we have asymptotics at all, the subsampling bootstrap works.

Then, trivially,

$$\tau_b(\hat{\theta}_b - \theta) \xrightarrow{\mathcal{D}} \text{Something}, \tag{4}$$

converges to the same "Something," because whether we index by $n$ or $b$ is merely a matter of notation. Usually, we write (4) as

$$\tau_b(\theta_b^* - \theta) \xrightarrow{\mathcal{D}} \text{Something} \tag{5}$$

to distinguish the estimator $\hat{\theta}_n$ for the full data and the estimator $\theta_b^*$ for a subsample, but it is the key feature of the subsampling bootstrap that $\hat{\theta}_b$ and $\theta_b^*$ are equal in distribution (because the subsampling is done without replacement as discussed in Sections 2 and 4).

The basic assumptions of the subsampling bootstrap are

$$\begin{aligned} b &\to \infty \\ \frac{b}{n} &\to 0 \\ \tau_b &\to \infty \\ \frac{\tau_b}{\tau_n} &\to 0 \end{aligned} \tag{6}$$

where $n$ is the sample size and $b$ the subsample size.

Under these assumptions

$$\tau_b(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} 0 \tag{7}$$

just because we would need to multiply by $\tau_n$ rather than $\tau_b$ to get a nonzero limit and $\tau_b/\tau_n$ goes to zero.

Subtracting (7) from (5) gives

$$\tau_b(\theta_b^* - \hat{\theta}_n) \xrightarrow{\mathcal{D}} \text{Something}, \tag{8}$$

where "Something" denotes the same distribution as in (3).

To summarize where we have gotten, the subsampling bootstrap is based on the assumptions (6) and the convergence in distribution (3). It then

follows from asymptotic theory that (8) describes the same asymptotics as (3).

It does not matter what the limiting distribution is because we approximate it using the subsampling bootstrap. Suppose the limiting distribution, the "Something" in (3) has distribution function $F$. We don't know the functional form of $F$ but we can approximate it by the empirical distribution function $F_b^*$ of the left hand side of (8) using bootstrap subsampling to simulate $\theta_b^*$.

We know that for large $n$

$$F^{-1}(\alpha/2) < \tau_n(\hat{\theta}_n - \theta) < F^{-1}(1 - \alpha/2) \tag{9}$$

occurs with probability approximately $1 - \alpha$. That's what the convergence in distribution statement (3) means when $F$ is the distribution function of the "Something" on the right hand side. $F^{-1}(\alpha/2)$ is the $\alpha/2$ quantile of this distribution and $F^{-1}(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile. Thus if $Y$ is a random variable having this distribution and the distribution is continuous, the probability that

$$F^{-1}(\alpha/2) < Y < F^{-1}(1 - \alpha/2) \tag{10}$$

is $1 - \alpha$. Since we are assuming $Y$ and $\tau_n(\hat{\theta}_n - \theta)$ have approximately the same distribution for large $n$, (9) has approximately the same probability as (10).

Of course, we don't know $F$, but $F_b^*$ converges to $F$, so for large $b$ and $n$, we have

$$F_b^{*-1}(\alpha/2) < \tau_n(\hat{\theta}_n - \theta) < F_b^{*-1}(1 - \alpha/2) \tag{11}$$

with probability close to $1 - \alpha$. Rearranging (11) to put $\theta$ in the middle by itself gives

$$\hat{\theta}_n - \tau_n^{-1} F_b^{*-1}(1 - \alpha/2) < \theta < \hat{\theta}_n - \tau_n^{-1} F_b^{*-1}(\alpha/2) \tag{12}$$

which is the way subsampling bootstrap confidence intervals are done.

In practice, we don't explicitly calculate empirical CDFs and invert them. We use the R `quantile` function to directly calculate quantiles. Assuming we already have calculated the estimator `theta.hat` for the original data having sample size `n` and a vector `theta.star` of estimators for bootstrap subsamples of size `b` and have previously defined a function `tau` that calculates the "rate" and a significance level `alpha`, the following three R statements calculate the confidence interval.

```
z.star <- tau(b) * (theta.star - theta.hat)
crit.val <- quantile(z.star, probs = c(1 - alpha / 2, alpha / 2))
theta.hat - crit.val / tau(n)
```

### 5.1 A Time Series Example

Here we redo the time series example from the web pages. First the statements

```
> library(bootstrap)
> y <- lutenhorm[, 4]
> n <- length(y)
> foo <- function(w) {
+     z <- w - mean(w)
+     m <- length(w)
+     out <- lm(z[-1] ~ z[-m] + 0)
+     as.numeric(coefficients(out))
+ }
> beta.hat <- foo(y)
> beta.hat

[1] 0.5857651
```

create a time series `y` and an estimator of the AR(1) coefficient calculated by the function `foo`.

Since this is a time series, we calculate subsampling bootstrap estimates by

```
> b <- 8
> nboot <- n - b + 1
> beta.star <- double(nboot)
> for (i in 1:nboot) {
+     y.star <- y[seq(i, i + b - 1)]
+     beta.star[i] <- foo(y.star)
+ }
```

Figure 1 shows the histogram of `beta.star` with the position of `beta.hat` shown (as usual by a vertical dotted line). It is made by the code

```
> hist(beta.star)
> abline(v = beta.hat, lty = 2)
```

We can see from the figure that the distribution of $\beta^*$ is badly biased and hence (assuming a close analogy between the "bootstrap world" and the "real world") so is the distribution of $\hat{\beta}$.

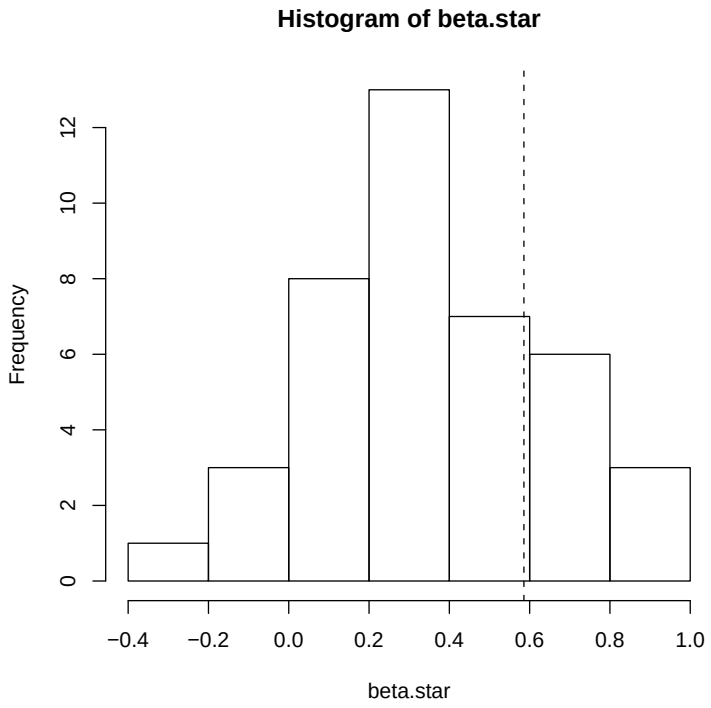


Figure 1: Histogram of $\beta^*$ with position of $\hat{\beta}$ shown by dashed line.

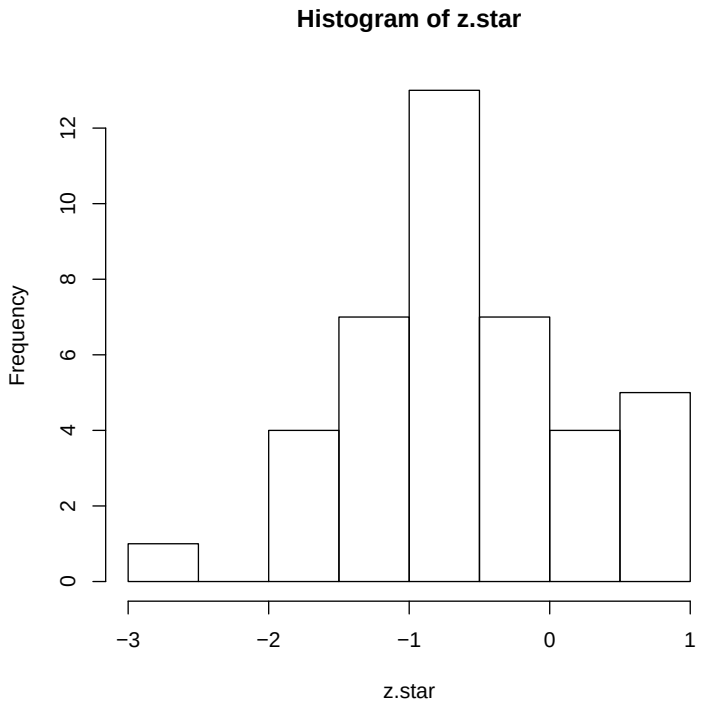


Figure 2: Histogram of $z^*$ given by (13)

Actually, we aren't much interested in this particular histogram. We really want to look at the distribution of

$$z = \sqrt{n}(\hat{\beta} - \beta)$$

which is approximated by

$$z^* = \sqrt{b}(\beta^* - \hat{\beta}) \tag{13}$$

Figure 2 shows this histogram, which is made by the code

```
> z.star <- sqrt(b) * (beta.star - beta.hat)
> hist(z.star)
```

We use quantiles of the distribution shown by Figure 2 to calculate critical values. For a 95% confidence interval, this goes as follows

```
> conf.level <- 0.95
> alpha <- 1 - conf.level
> crit.val <- quantile(z.star, probs = c(1 - alpha/2,
+     alpha/2))
> beta.hat - crit.val/sqrt(n)

    97.5%      2.5%
0.4719918 0.8641141
```

# 6 Estimating the Rate

We aren't always lucky enough to know the rate of convergence $\tau_b$. But even if we don't, we can estimate the rate from looking at the distribution of $\theta_b^*$ for different subsample sizes $b$. The method described here is that of Chapter 8 of Politis, et al. (1999).

Suppose $\tau_n = n^\beta$ for some constant $\beta$. This is the usual case. It includes the "square root law" ($\beta = 1/2$) found in the usual asymptotics and most other examples of interest. Under this supposition, what we need to do is estimate the unknown constant $\beta$.

To do this we take subsamples at different sizes. For each sample size $b$ we look at the distribution of $\theta_b^* - \hat{\theta}_n$. For each such distribution we determine several quantiles. Let $G_b^{*-1}(t)$ denote the $t$-th quantile of the distribution of $\theta_b^* - \hat{\theta}_n$. Then

$$G_b^{*-1}(t) \approx b^{-\beta} F^{-1}(t) \tag{14}$$

where $F$ is the CDF of the asymptotic distribution of $n^\beta(\hat{\theta}_n - \theta)$. A proof of this is rather technical (Politis, et al., 1999, Chapter 8), but we will give a "proof by picture" in the example.

We want to use (14) for different $b$ and $t$ to estimate $\beta$. It would help if we take logs, but there is a problem that $G_b^{*-1}(t)$ need not be positive, it being the quantile of a distribution centered at zero. Thus we take differences. When $s < t$

$$G_b^{*-1}(t) - G_b^{*-1}(s) \approx b^{-\beta}\bigl[F^{-1}(t) - F^{-1}(s)\bigr] \tag{15}$$

will be positive, and we can take logs

$$\log\bigl[G_b^{*-1}(t) - G_b^{*-1}(s)\bigr] \approx -\beta \log(b) + \log\bigl[F^{-1}(t) - F^{-1}(s)\bigr] \tag{16}$$

and this suggests estimating $\beta$ by linear regression.

Choose several sample sizes $b_j$ and several pairs of quantiles $s_i < t_i$. Politis, et al. (1999, Chapter 8) suggest choosing $s_i < 0.5 < t_i$, but this is not essential. Then define

$$y_{ij} = \log\bigl[{G_{b_j}^*}^{-1}(t_i) - {G_{b_j}^*}^{-1}(s_i)\bigr],$$

which are calculated as differences of sample quantiles of the distribution of $\theta_{b_j}^* - \hat{\theta}_n$, and define

$$c_i = \log\bigl[F^{-1}(t_i) - F^{-1}(s_i)\bigr]$$

which cannot be calculated because we do not know $F$ but this will not matter. Now average over the index $i$ (over pairs of quantiles). Let $\bar{y}_j$ be the average of the $y_{ij}$, and let $\bar{c}$ be the average of the $c_i$. Then

$$\bar{y}_j \approx -\beta \log(b_j) + \bar{c}$$

and Politis, et al. (1999) recommend regressing $\bar{y}_j$ on $\log(b_j)$ to estimate $\beta$.

We don't need a regression routine, because this is simple linear regression, so

$$\hat{\beta} = -\frac{\operatorname{cov}\{\bar{y}, \log(b)\}}{\operatorname{var}\{\log(b)\}}$$

does the job.

## 6.1 An IID Example

The data for the example on the web page can be loaded into R by

```
> X <- read.table(url("http://www.stat.umn.edu/geyer/5601/mydata/big-unif.txt"),
+     header = TRUE)
> names(X)

[1] "x"

> attach(X)
```

(this is essentially what Rweb does when loading a file from a dataset URL).

The data are IID Uniform$(0, \theta)$ and have "rate" $\tau_n = n$, but we pretend for the purposes of this example that we don't know this. The estimator of interest is calculated by the `max` function.

```
> theta.hat <- max(x)
> theta.hat
```

```
[1] 2.717583
```

The following code (quite tricky, explained below) calculated several bootstrap subsamples with different subsample sizes $b$

```
> nboot <- 20000 - 1
> b <- c(40, 60, 90, 135)
> b <- sort(b)
> theta.star <- matrix(NA, nboot, length(b))
> for (i in 1:nboot) {
+     x.star <- x
+     for (j in length(b):1) {
+         x.star <- sample(x.star, b[j], replace = FALSE)
+         theta.star[i, j] <- max(x.star)
+     }
+ }
```

Here `b` is a vector of subsample sizes (we need at least two) chosen with no very scientific theory. They should all be large compared to 1 and small compared to $n$, which in this case is

```
> length(x)
```

```
[1] 10000
```

Thus very roughly, all of the $b$'s should be about $\sqrt{10000} = 100$. Also, since we intend to regress $\log(\bar{y})$ on $\log(b)$, we choose the $\log(b)$ equally spaced, which means the $b$ themselves increase by a constant factor, here 1.5 (that is 60 is $1.5 \times 40$ and 90 is $1.5 \times 60$ and so forth). This type of spacing isn't necessary, but makes the plots below look nicer.

This code uses a minor innovation of our own (not from Politis and Romano) which is to use the *principle of common random numbers* from Monte Carlo theory. This says that when comparing random (simulated) objects (here the subsamples for different $b$) it always helps to use the same random numbers as much as possible. We do this by reusing the basic Politis and Romano idea: a subsample of size $b$ *without replacement* from a sample of size $n$ does not change the distribution. It is just like directly taking a sample of size $b$ from the population.

Thus we first take our largest subsample, here of size 135, and then (here's where the principle of common random numbers comes in) take our next largest subsample, here of size 90, from the subsample of size 135 (rather

from the original sample of size 1000). This makes the two subsamples of different sizes positively correlated and improves everything else we do. The smaller subsamples of size 60 and 40 are done analogously.

The estimates for each sample are stored in an `nboot` by `length(b)` matrix `theta.star`.

Figure 3 shows parallel boxplots of $\theta_b^* - \hat{\theta}_n$ for the different sample sizes. It is made by the code

```
> zlist <- list()
> for (i in 1:length(b)) {
+     zlist[[i]] <- theta.star[, i] - theta.hat
+ }
> names(zlist) <- b
> boxplot(zlist, xlab = "subsample size", ylab = expression(hat(theta)[b] -
+     hat(theta)[n]))
```

From Figure 3 we see several things.

- The spread of the $\theta_b^* - \hat{\theta}_n$ values decreases as $b$ increases.
- The $\theta_b^* - \hat{\theta}_n$ values are all negative.

But this figure doesn't tell us directly anything about the rate.

The rate estimation method of Politis and Romano requires us to calculate some differences of quantiles of these distributions. We estimate them by differences of order statistics.

```
> qlist <- list()
> k <- (nboot + 1) * seq(0.05, 0.45, 0.05)
> l <- (nboot + 1) * seq(0.55, 0.95, 0.05)
> k

[1] 1000 2000 3000 4000 5000 6000 7000 8000 9000

> l

[1] 11000 12000 13000 14000 15000 16000 17000 18000 19000

> for (i in 1:length(b)) {
+     z.star <- zlist[[i]]
+     sz.star <- sort(z.star, partial = c(k, l))
+     qlist[[i]] <- sz.star[l] - sz.star[k]
+ }
```

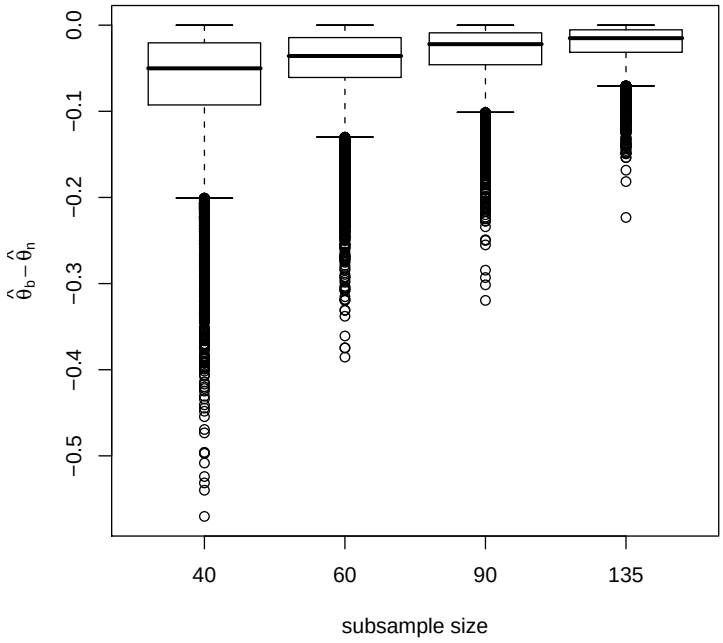


Figure 3: Parallel Boxplots of $\theta_b^* - \hat{\theta}_n$.

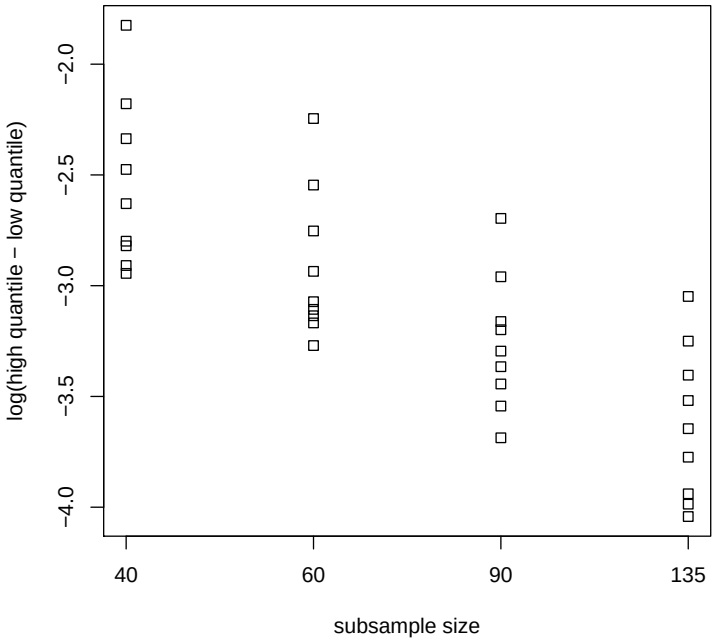


Figure 4: Stripchart of Subsample Quantiles.

Figure 4 shows a "stripchart" of the logs of these quantities for the different sample sizes. It is made by the code

```
> names(qlist) <- b
> lqlist <- lapply(qlist, log)
> stripchart(lqlist, xlab = "subsample size", ylab = "log(high quantile - low quantil
+     vertical = TRUE)
```

From Figure 4 we see several things.

- The four "stripcharts" are about equally spread (on this log scale).
- If we think of these points as a scatter plot, a simple linear regression doesn't look too bad.

Of course, these points are not typical regression data (not a straight line plus IID errors). We are only using linear regression as tool to put a

straight line on this plot.

Our estimate of $\beta$ is minus the slope of this "regression line" and is calculated by

```
> y <- sapply(lqlist, mean)
> beta <- cov(-y, log(b))/var(log(b))
> beta

[1] 0.8822342
```

and, just out of curiosity, we also calculate the intercept

```
> inter <- mean(y) + beta * mean(log(b))
```

and add the "regression line" to the plot giving Figure 5. It is made by the code

```
> plot(log(rep(b, each = length(k))), unlist(lqlist),
+     xlab = "log(subsample size)", ylab = "log(high quantile - low quantile)")
> abline(inter, -beta)
```

From Figure 5 we can see how the method works.

It's certainly not the one and only way to get a rate estimate. And the rate estimate isn't all that good, 0.8822 when we know from theory that the correct rate is exactly 1.

Now, finally, we are ready for a confidence interval calculation, which proceeds exactly like the preceding example, except we use "rate" $\tau_n = n^\beta$ where $\beta$ is the estimate just obtained.

```
> conf.level <- 0.95
> alpha <- 1 - conf.level
> m <- 3
> b <- b[m]
> b

[1] 90

> theta.star <- theta.star[, m]
> crit.val <- sort(z.star)[(nboot + 1) * alpha]
> theta.hat - c(0, crit.val)/n^beta

[1] 2.717583 2.719533
```

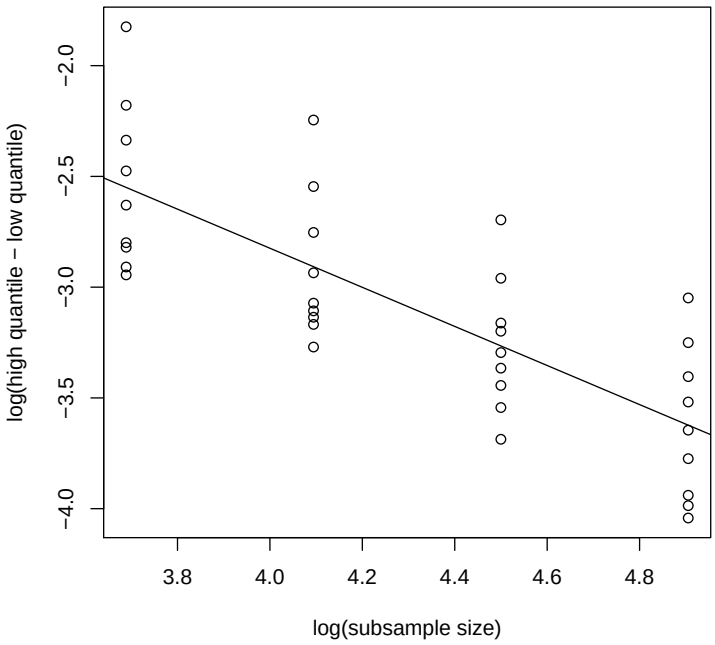


Figure 5: Stripchart of Subsample Quantiles with "Regression Line."

Note this is a one-sided confidence interval, because in this particular example we know that $\hat{\theta} < \theta$ with probability one. In general, we want a two-sided interval, which we leave to the web page.

Also a little tricky, we have not obtained a new bootstrap of $\theta^*$ but rather used one we already had, the one with $b = 90$, with is `theta.star[ , 3]` the way we stored them.

## References


Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1–26.

Politis, D. N.; Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Annals of Statistics*, 22, 2031–2050.

Politis, D. N.; Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer Verlag.