

Bootstrap estimate of Standard Error (SE) Example 2: Correlation Coefficient.

Crime Data: The data in file `crime.txt` are crime-related and demographic statistics for 47 US states in 1960. The data were collected from the FBI's Uniform Crime Report and other government agencies to determine how the variable crime depends on other variables measured in the study. 14 variables were measured in the study. We will consider two variables:

Column 1: R = crime rate = number of offenses reported to police per million population;
Column 5: $Ex0$ = 1960 per capita expenditure on police by state and local government.

```
> # File bootCORR.S: S-plus code for Bootstrap Example 2.
> # -----
> crime <- read.table("crime.txt", header=T)
> crime
      R Age S  Ed Ex0 Ex1  LF   M   N  NW  U1 U2  W  X
1  79.1 151 1  91  58  56 510  950  33 301 108 41 394 261
2 163.5 143 0 113 103  95 583 1012  13 102  96 36 557 194
3  57.8 142 1  89  45  44 533  969  18 219  94 33 318 250
4 196.9 136 0 121 149 141 577  994 157  80 102 39 673 167
...
> plot(Ex0, R, xlim=c(0,200), ylim=c(0,200),
      xlab="police expenditure", ylab="crime rate",
      main="Crime Data")
```

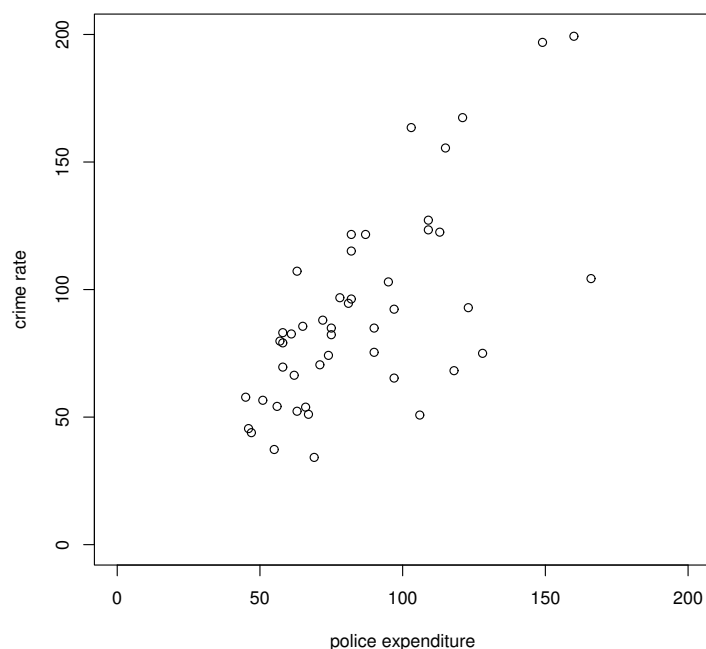


Figure 1: Crime Data.

Correlation Coefficient, θ :

```
> cor(R, Ex0)
[1] 0.6876044
```

Random Sample:

```
> my.sample <- sample(1:47, 15)
> my.data <- crime[my.sample, c(1,5)]
```

Sample Correlation Coefficient, $\hat{\theta}$:

```
> cor(R[my.sample], Ex0[my.sample])
[1] 0.75792
```

How accurate is the sample correlation coefficient, $\hat{\theta}$, as an estimate of the population correlation coefficient, θ ? Use bootstrap to estimate the standard error of $\hat{\theta}$.

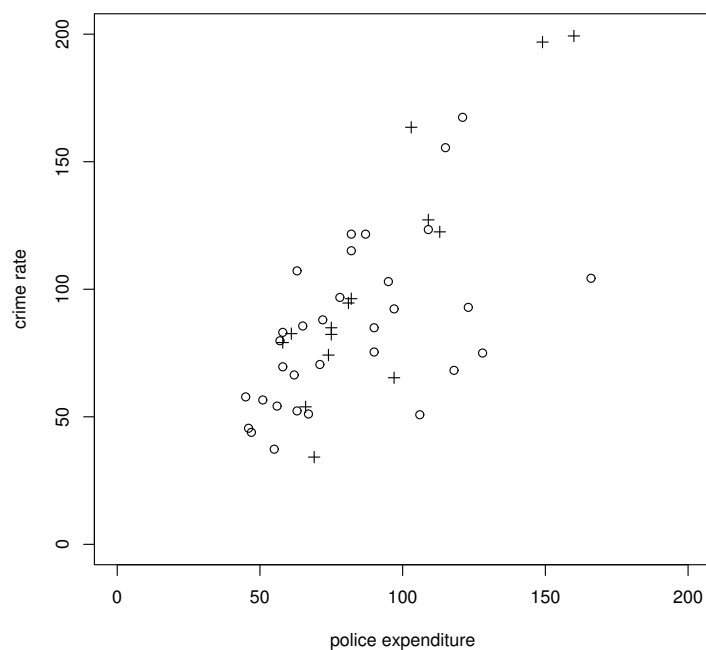


Figure 2: Sampled Data.

Bootstrap Standard Error:

```
B = 100 ⇒ seboot = 0.1459717
B = 1000 ⇒ seboot = 0.1216236
B = 3000 ⇒ seboot = 0.1352250
```

Details:

```

> theta <- function(x, xdata)
  {
    cor(xdata[x,1], xdata[x,2])
  }

> boot1 <- bootstrap(1:15, nboot=100, theta, my.data)
> sqrt(var(boot1$thetastar))
[1] 0.1459717

> boot2 <- bootstrap(1:15, nboot=100, theta, my.data)
> sqrt(var(boot2$thetastar))
[1] 0.1216236

> boot3 <- bootstrap(1:15, nboot=100, theta, my.data)
> sqrt(var(boot3$thetastar))
[1] 0.1352250

> hist(boot3$thetastar, xlab="correlation", main="3000 replications")

```

Left panel of Figure 3 shows the histogram of 3000 bootstrap replicates of correlation coefficient. This is bootstrap approximation of the distribution of sample correlation coefficient. However, since we have the whole population (i.e. 47 states), we can estimate the distribution of sample correlation coefficient by taking random samples of size 15 from this population:

```

> N <- 3000
> r <- rep(0, N)
> for (n in 1:N) {
  my.sample <- sample(1:47, 15, replace=F)
  my.data <- crime[my.sample, c(1,5)]
  r[n] <- cor(my.data[, 1], my.data[, 2])
}
> hist(r, xlab="correlation", main="3000 Random samples")

```

Comparison between left and right panels of Figure 3 shows that bootstrap approximates the true distribution well. See chapter 6 of Efron and Tibshirani.

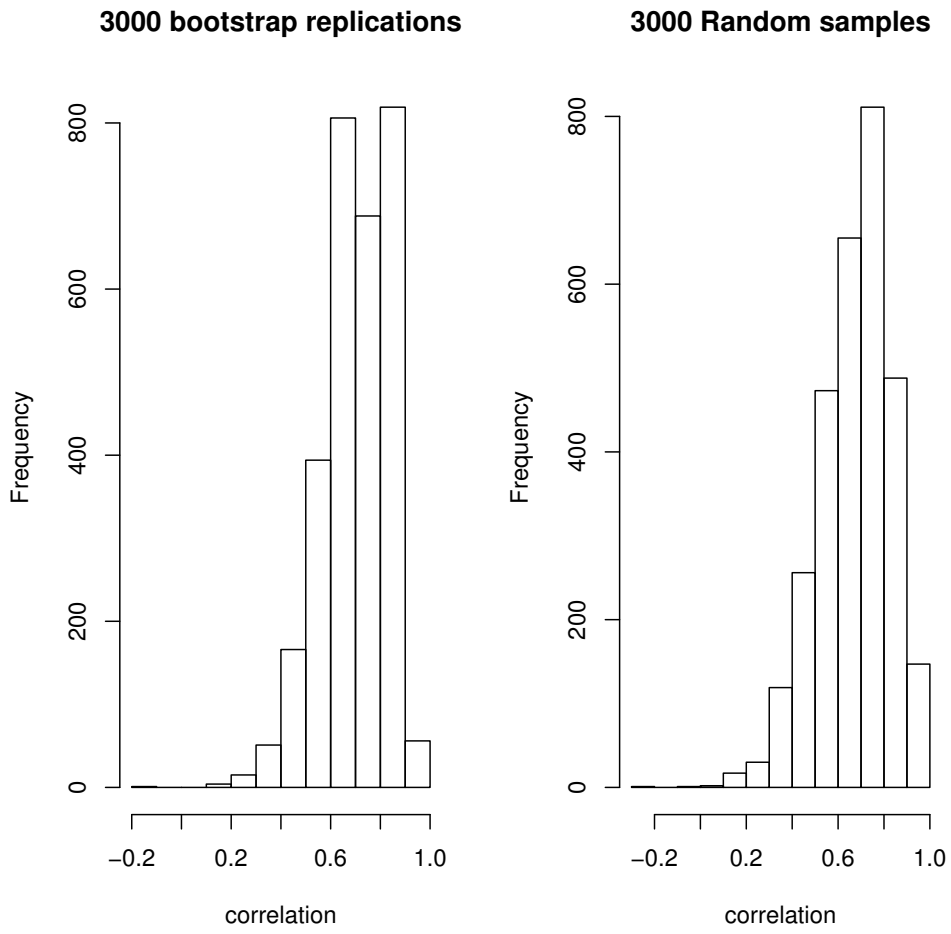


Figure 3:
 Left panel: histogram of 3000 bootstrap replications of $\hat{\theta}$.
 Right panel: histogram of $\hat{\theta}$ based on 3000 samples of size 15 from the population of 47 states.