# 22S:166
# More on the Bootstrap

Lecture 9
September 24, 2008

Kate Cowles
374 SH, 335-0727
kcowles@stat.uiowa.edu



## Choosing the number of bootstrap datasets

- approximately 1000 to 2000 is minimum for reasonable performance in most cases
- choosing R = 999 or 1999 facilitates calculation of percentile confidence intervals (see below)



## Another version of the function for calculating the statistic for the city data

```
> meanratio
function(df, indices)
{
        # df must be a data frame with two colums, "x" and "u"
        mean( df[ indices, "x" ]) / mean( df[ indices, "u" ] )
}
```



## Running the bootstrap with different settings of R

```
> library(boot)
> data(city)
>
> boot.out <- boot( city, meanratio, R=999)

> boot.out

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = city, statistic = meanratio, R = 999)


Bootstrap Statistics :
    original     bias    std. error
t1* 1.520312 0.04122696   0.2168435
```

```
> boot.out <- boot( city, meanratio, R=999)
> boot.out

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = city, statistic = meanratio, R = 999)


Bootstrap Statistics :
    original     bias    std. error

t1* 1.520312 0.04515005   0.2256023

> boot.out <- boot( city, meanratio, R=999)
> boot.out

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = city, statistic = meanratio, R = 999)


Bootstrap Statistics :
    original     bias    std. error
t1* 1.520312 0.03724419   0.2098392

> boot.out <- boot( city, meanratio, R=1999)
> boot.out

ORDINARY NONPARAMETRIC BOOTSTRAP
```



```
Call:
boot(data = city, statistic = meanratio, R = 1999)


Bootstrap Statistics :
    original     bias    std. error
t1* 1.520312 0.04460204   0.2267071
> boot.out <- boot( city, meanratio, R=1999)
> boot.out

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = city, statistic = meanratio, R = 1999)


Bootstrap Statistics :
    original     bias    std. error
t1* 1.520312 0.02751536   0.2116137
>
```



# Interpreting the boot object

```
> names(boot.out)
 [1] "t0"        "t"         "R"         "data"      "seed"      "statistic'
 [7] "sim"       "call"      "stype"     "strata"    "weights"

> boot.out$t0     # thetahat from original data
[1] 1.520312

> hist(boot.out$t)    # histrogram of thetastars from bootstrap samples
> abline(v=boot.out$t0,lty=3)    # add vertical line at thetahat

> mean(boot.out$t)
[1] 1.547828

> mean(boot.out$t) - boot.out$t0     # bootstrap estimate of bias
[1] 0.02751536

> bbias <- mean(boot.out$t) - boot.out$t0
> boot.out$t0 - bbias             # bootstrap "unbiased" estimate

[1] 1.492797

> sd(boot.out$t)                  # bootstrap standard error
[1] 0.2116137
```



# More on bootstrap confidence intervals

```
> boot.ci.out <- boot.ci(boot.out)
Warning message:
In boot.ci(boot.out) : bootstrap variances needed for studentized intervals
> boot.ci.out
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1999 bootstrap replicates

CALL :
boot.ci(boot.out = boot.out)

Intervals :
Level      Normal              Basic
95%   ( 1.078,  1.908 )   ( 0.973,  1.796 )

Level     Percentile            BCa
95%   ( 1.245,  2.068 )   ( 1.258,  2.121 )
Calculations and Intervals on Original Scale
```

## Bootstrap confidence intervals continued

- Basic interval
  - if there was a function of the population quantity we're interested in $\theta$ and the estimator $\hat{\Theta}$ whose distribution was known, we could use the quantiles of this distribution to construct c.i. for $\theta$
  - since we don't have this, arbitrarily consider $W = (\hat{\Theta} - \theta)$
  - if we knew distribution of $W$, then two-sided level $100 \times (1 - \alpha)$ would be
    $$(\theta - w_{1-\frac{\alpha}{2}}, \theta - w_{\frac{\alpha}{2}})$$
  - bootstrap idea: use distribution of $W^* = (\hat{\Theta}^* - \hat{\theta})$ as approximation to distribution of $W$
  - 
```
> w <- sort(boot.out$t) - boot.out$t0
> hist(w)
```



```
> boot.out$t0 - w[c(1950,50)]
[1] 0.9727399 1.7959120
> boot.out$t0 - quantile(w,[c(.975,.025)
```

  - pros: may work well for medians
  - cons: bootstrap error (distribution of $W^*$ being a poor approximation to distribution of $W$) often is large



- Percentile interval of level (1-$\alpha$)
  - lower endpoint is $\frac{\alpha}{2}(R+1)$ entry in ordered bootstrap statistics
    item upper endpoint is $\left(1 - \frac{\alpha}{2}\right)(R+1)$ entry
  - 
```
> sort(boot.out$t)[c(50,1950)]
[1] 1.244713 2.067885
```

  - pros: simplicity
  - cons: may be very inaccurate if distribution of $\hat{\theta}$ is not close to symmetric



## Bias-correcting bootstrap percentile confidence intervals

- recall:
$$CDF(q) = Pr_*(\hat{\theta}^* \leq q) = \frac{\#\{\hat{\theta}^b \leq q\}}{B}$$
- if $CDF(\hat{\theta}) \neq .5$, then bias correction to percentile method c.i. may be in order
- let
$$z_0 = \Phi^{-1}(CDF(\hat{\theta}))$$
  - what Splus/R function evalutes $\Phi^{-1}$
- then bias-corrected $1 - \alpha$ c.i. is
$$\left[C\bar{D}F^{-1}(\Phi(2z_0 - z_{\alpha/2})), C\bar{D}F^{-1}(\Phi(2z_0 + z_{\alpha/2}))\right]$$
  - here $z_{\alpha/2}$ is upper $\alpha/2$ point of standard normal
    $\Phi(z_{\alpha/2}) = 1 - \alpha/2$