Section 6.5 Bootstrap

20) The bootstrap generates pseudo data $T_1^*, \ldots, T_B^*$ for a statistic T_n that estimates θ) such that, under regularity conditions, applying certain large sample $100(1-\delta)\%$ PIs to the bootstrap sample results in a large sample $100(1-\delta)\%$ confidence intervals (CIs). The "$*$" means from the bootstrap resampling.

21) [p19] Applying the shorth PI to $T_1^*, \ldots, T_B^*$ gives the shorth CI

22) [p19] The "large sample" $100(1-\delta)\%$ percentile CI covers $\approx \lceil B(1-\delta) \rceil$ of the T_i^*. Let $k_1 = \lceil B\delta/2 \rceil$ and $k_2 = \lceil B(1-\delta/2) \rceil$. A common choice is $\left[\overline{T}^*_{(k_1)}, \overline{T}^*_{(k_2)} \right]$, which is the nonparametric PI of [19]' applied to the bootstrap sample. The shorth CI is the "shortest" percentile CI, that covers enough cases.

23) Nonparametric bootstrap: Let the data $Y_1, \ldots, Y_n$ be iid and let statistic $T_n = t(Y_1, \ldots, Y_n)$. Draw a sample of size n $Y_1^*, \ldots, Y_n^*$ with replacement from $Y_1, \ldots, Y_n$ and compute $T_1^* = t(Y_1^*, \ldots, Y_n^*)$. Repeat B times

to generate bootstrap sample
$T_1^*, \ldots, T_B^*$.

24) Know for E1) For any B, given B bootstrap samples, compute

$T_1^*, \ldots, T_B^*$ and the bagging estimator

$\bar{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ = sample mean of the bootstrap sample

Often $T_n = \bar{X}_n$, $MED(n)$, range $= Y_{(n)} - Y_{(1)}$

ex) $Y_1, Y_2 = 0, 4$ T_n = range $= x_{(n)} - x_{(1)} = 4 - 0 = 4$

bootstrap sample Y_i^*	T_i^*
0,0	0
0,4	4
4,0	4
4,4	0

$$\bar{T}^* = \frac{8}{4} = 2$$

See Example 2.10.

ex) data 1, 2, 5, 10, 50 $MED(n) = 5 = T_n$

bootstrap sample	ordered	T_i^*
2,10,1,2,2	1,2,2,2,10	$T_1^* = 2$
50,10,50,2,2	2,2,10,50,50	$T_2^* = 10$
10,50,2,1,1	1,1,2,10,50	$T_3^* = 2$

$$\bar{T}^* = \frac{\sum T_i^*}{B} = \frac{14}{3} = 4.6667$$

is $\left[\overline{T}^* - a_{(UB)}, \overline{T}^* + a_{(UB)}\right]$ is the closed interval centered at $\overline{T}^*$ just large enough to cover $U_B = \lceil B(1-\delta) \rceil$ of the T_i^* where $a_i = |T_i^* - \overline{T}^*|$.

Replace $\overline{T}^*$ by T_n for the Bretagnolle-Huber-Carol CI.

26) The percentile CI, CI in 25) and possibly the shorth CI are useful for robust statistics with good large sample theory and good bootstrap theory. $T_n = MED(n)$ is such a statistic.

27) often apply the bootstrap CIs for T_n and hope they simulate well. Then the CI and test are "ad hoc."

28) PB Bootstrap test: T_n estimates θ

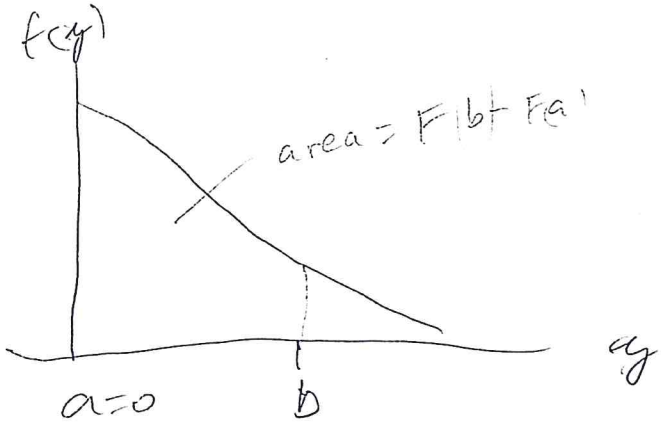
$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma_T^2)$.

$H_0: \theta = \theta_0$ vs $H_A: \theta \neq \theta_0$

reject H_0 if θ_0 is not in the

$100(1-\delta)\%$ CI for θ.

distributions will be useful see pp 111, Section 9.5



pdf of Y_T is $f(y)$ but needs to integrate to 1.

Let $F(y-) = P[Y < y]$. Assume $F(a) = F(a-)$

The truncated RV $Y_T = Y_T(a,b)$ with truncation points a and b has cdf

$$F_{Y_T}(y \mid a,b) = G(y) = \begin{cases} 0 & y < a \\ \dfrac{F(y) - F(a)}{F(b) - F(a)} & a \le y \le b \\ 1 & y > b \end{cases}$$

with mean $\mu_T = \mu_T(a,b) = \dfrac{\int_a^b y f(y)\,dy}{F(b) - F(a)}$

and variance $\sigma_T^2 = \sigma_T^2(a,b) = \dfrac{\int_a^b y^2 f(y)\,dy}{F(b) - F(a)} - \mu_T^2$.

303 p21 The α trimmed mean

$$T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y(i) \quad \text{where}$$

$L_n = \lfloor n\alpha \rfloor$ and $U_n = n - L_n$. If $\alpha = .25$, say, the α trimmed mean is called the 25% trimmed mean.

A 95% CI for θ is $\hat{\theta} \pm t_{p,0.975} SE(\hat{\theta})$

where p = degrees of freedom is an integer ≥ 1,

$a.p.\left(t_p \leq t_{p,0.975}\right) = 0.975$ and $t_{p,0.975} \approx 1.96$ for $p > 30$.

2) I know for EI — find $\hat{\theta}$, p, $SE(\hat{\theta})$ for 3 estimators

	$\hat{\theta}$	p	$SE(\hat{\theta})$
I)	$\bar{Y}$	$n-1$	$S/\sqrt{n}$
II) p219	$\lfloor n/4 \rfloor$ trimmed mean	$U_n - L_n - 1$	$0.5\left[Y_{(U_n)} - Y_{(L_n+1)}\right]$

where $U_n = n - L_n$ and $L_n = \left\lfloor \frac{n}{2} \right\rfloor - \left\lceil \sqrt{\frac{n}{4}} \right\rceil$

Here $\lfloor x \rfloor$ = greatest integer function so $\lfloor 7.7 \rfloor = 7$

and $\lceil x \rceil$ = smallest integer $\geq x$ so $\lceil 7.7 \rceil = 8$

II) p219 T_n $\quad U_n - L_n - 1 \quad SE(T_n)$

3) p219 If T_n is the 25% trimmed mean,

$L_n = \left\lfloor \frac{n}{4} \right\rfloor$, $U_n = n - L_n$

$T_n = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)}$ estimates $\theta = \mu_T = \int_{y_{0.25}}^{y_{0.75}} 2y\, f(y)\, dy$

Order data, delete smallest and largest 25%

take sample mean of remaining half set of data

$$\begin{cases} \\ Y_{(i)} & L_n+1 \le i \le U_n \\ Y_{(U_n)} & i \ge U_n+1 \end{cases}$$

Let $S_n^2(d_1,\dots,d_n)$ = sample variance of the d_i

$$= \frac{\sum (d_i-\bar{d})^2}{n-1} = \frac{\sum d_i^2 - n(\bar{d})^2}{n-1}$$

Let $V_{SW}(L_n, U_n) = \frac{S_n^2(d_1,\dots,d_n)}{\left(\frac{U_n-L_n}{n}\right)^2}$

Then $SE(T_n) = \sqrt{\frac{V_{SW}(L_n,U_n)}{n}}$.

ex) data 66, 3, 8, 5, 2

sort data 2, 3, 5, 8, 6.6

earlier showed $\bar{Y} = 16.8$ $S = 27.5989$ $\rho = n-1 = 4$

So $SE(\bar{Y}) = \frac{S}{\sqrt{n}} = \frac{27.5989}{\sqrt{5}} = 12.343$

) $MED(n) = 5$

$L_n = \lfloor \frac{n}{2} \rfloor - \lceil \sqrt{\frac{n}{4}} \rceil = \lfloor \frac{5}{2} \rfloor - \lceil \sqrt{\frac{5}{4}} \rceil = \lfloor 2.5 \rfloor - \lceil 1.118 \rceil$

$= 2-2 = 0$ so $U_n = n - L_n = n = 5$

$SE(MED(n)) = 0.5\left(Y_{(U_n)} - Y_{(L_n+1)}\right) = \frac{Y_{(5)} - Y_{(1)}}{2}$

$= \frac{66-2}{2} = 32$

$\rho = U_n - L_n - 1 = 5-0-1 = 4$

CI's based on MEDIAN tend to be too long for small n; works better for $n > 100$

$L_n = \lfloor n \cdot .25 \rfloor = \lfloor \frac{5}{4} \rfloor = 1$, $U_n = n - L_n = 5 - 1 = 4$

2, | 3, 5, 8, | 66

$T_n = \frac{3+5+8}{3} = \frac{16}{3} \simeq \boxed{5.333}$

d_i's 3, 3, 5, 8, 8

$\bar{d} = \frac{3+3+5+8+8}{5} = \frac{27}{5} = 5.4$

$S_n^2(d_1, \ldots, d_n) = \frac{\sum d_i^2 - n(\bar{d})^2}{n-1} = \frac{171 - 5(5.4)^2}{4}$

$= \frac{25.2}{4} = 6.3$

$V_{SW} = \frac{6.3}{\left(\frac{4-1}{5}\right)^2} = 17.5$

$SE(T_n) = \sqrt{\frac{V_{SW}}{n}} = \sqrt{\frac{17.5}{5}} = \sqrt{3.5} = \boxed{1.871}$

$\rho = U_n - L_n - 1 = 4 - 1 - 1 = \boxed{2}$

---

using rpack

$y \leftarrow c(66, 3, 8, 5, 2)$

a) cci(y) int = CI = (17.469, 51.069) ← way too long

mean = $\bar{Y} = 16.8$

SE = $SE(\bar{Y}) = 12.343$

b) tmci CI = [−83.846, 93.846] ← way too long

MED = MED(n) = 5

seb = SE(MED(n)) = 32

c) tmci(y) CI = [−2.716, 13.383] good

Check $T_n \pm t_{2,.975} SE(T_n) = \frac{16}{3} \pm 4.303(1.871)$

↑ t table, computer is more accurate

$\approx 5.3333 \pm 8.0509 = [-2.718, 13.384]$

p39–40

34) These are large sample CI's (n should be large, roughly want n > 20).

Assumptions i) $Y_1, ..., Y_n$ are iid with a pdf

ii) $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N(0, V_\theta)$ where $\lim_{n\to\infty} n \, var(\hat{\theta}) = V_\theta$

so $\hat{\theta} \approx N(\theta, [SE(\hat{\theta})]^2)$

often $\lim_{n\to\infty} n \, var(\hat{\theta}) = V_\theta$

eg $var \bar{Y} = \sigma^2/n$, $SE(\bar{Y}) = \frac{S}{\sqrt{n}}$

and $P\left(-z_{.975} < \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} < z_{.975}\right) \approx .95$ ← for large n

or $P(-z_{.975} SE(\hat{\theta}) < \hat{\theta} - \theta < z_{.975} SE(\hat{\theta})) \approx .95$

or $P(-\hat{\theta} - z_{.975} SE(\hat{\theta}) < -\theta < -\hat{\theta} + z_{.975} SE(\hat{\theta})) \approx .95$

or $P(\hat{\theta} - z_{.975} SE(\hat{\theta}) < \theta < \hat{\theta} + z_{.975} SE(\hat{\theta})) \approx .95$

So $P\left(\hat{\theta} - z_{.975}\, SE(\hat{\theta}) < \theta < \hat{\theta} + z_{.975}\, SE(\hat{\theta})\right) \simeq .95$

1) For smaller n replacing $z_{.975}$ by $t_{p,.975}$ works often so

$\hat{\theta} \pm t_{p,.975}\, SE(\hat{\theta})$ is a good CI.

ex) $\hat{\theta} = \bar{Y}$ then $\sqrt{Var(\bar{Y})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

and $SE(\bar{Y}) = \frac{s}{\sqrt{n}}$, $s^2 \simeq \sigma^2$ for large n

§2.14 35) Simulation relies on computer generated pseudo random variables that can, for many purposes, be regarded as iid random variables $Y_1, \ldots, Y_n$ from a specified distribution. The "runs" are independent.

ex) ($Y_1, Y_2, Y_3 \sim N(0,1)$)

				MED($\bar{Y}$)
rnorm(3) run 1:	.861	.194	1.699	.861
rnorm(3) run 2:	−1.418	.579	.409	.409

ex) length of pregnancy $\simeq N(\mu = 266, \sigma^2 = (16)^2)$

rnorm(3, mean=266, SD=16)

				max
run 1	266.34	281.46	251.898	281.46
run 2	261.81	232.64	266.07	266.07

start a year, what will be the longest pregnancy?

To simulate 100 such hospitals use the command

X ← matrix(rnorm(100000, mean=266, sd=16), nrow=100, ncol=1000)

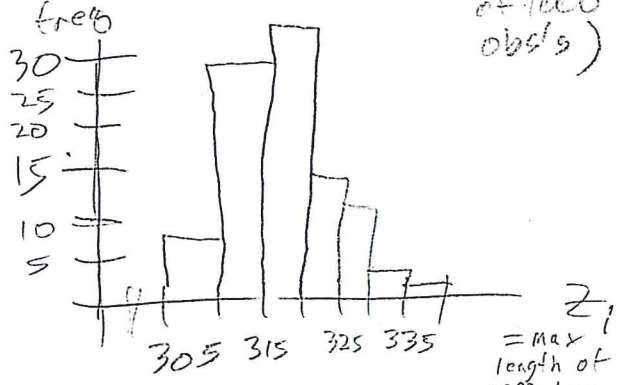
X is 100 × 1000 and each row represents a hospital with 1000 pregnancy lengths.

hist(X) to see all births at the 100 hospitals

Z ← apply(X, 1, max) ← find the max of each row (= sample of 1000 obs's)

min(Z)	307.76
max(Z)	337.83
mean(Z)	317.70
median(Z)	316.52
sqrt(var(Z))	6.39
mad(Z, constant=1)	4.15

hist(Z)



(Actually the right tail of pregnancy lengths may not be normal. Hospital might induce labor or do a csection for 30 or so days overdue.)

36) The point of simulation is that you can do large numbers of computer experiments often in seconds.

37) Applications: i) check computer programs: does simulation match theory

ii) demonstrate 10 or 20 graphs in rapid succession so user can see typical graph

ex HW 1.3d 4 histograms of normal data with sample size n=100. Most of the histograms do