Math 152 - Statistical Theory - Homework 8

Ethan Ashby

Due: Friday, October 16, 2020, midnight PDT

1: PodQ Describe one thing you learned from someone in your pod this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Rafa engaged us in a little conversatuion about the measure-theoretic interpretations of some of the stats concepts we've covered! Very cool!

4: R - blood pressure Consider a large study of the association between blood pressure and cardiovascular disease that found:

- 55 out of 3338 men with high bp died of cardiovascular disease ($\hat{p}_1 = 0.0165$)
- 21 out of 2676 men with low bp died of cardiovascular disease ($\hat{p}_2 = 0.0078$)

The parameter of interest in this study is called the relative risk:

$$RR = \theta = \frac{P(\text{death from CD} \,|\, \text{high bp})}{P(\text{death from CD} \,|\, \text{low bp})}$$

$$\hat{\theta} = \frac{\hat{p}_1}{\hat{p}_2} = \frac{0.0165}{0.0078} = 2.12$$

a. Why is bootstrapping a single time from the high bp group the same as taking 3338 random values from a Bernoulli distribution with probability 0.0165? (Or said differently, 1 binomial sample with `n=3338` and `p=0.0165`.) A single bootstrap form the high BP group is a sample with relacement of size 3338. The probability of getting a dead person per individual in your sample is $\hat{p}_1 = 0.0165$. A boostrap sample is tantamount to drawing 3338 individuals indepedently with replacement from this sample, where each person is labelled either a 0 (alive) or 1 (dead), and the probability of getting a 1 (dead) for each person is $\hat{p}_1 = 0.0165$. This matches the definition of drawing 3338 random values from a bernoulli distribution, which models a binary outcome with some fixed success probability.

b. Why is bootstrapping repeatedly (e.g., B = 500) from the high bp group the same as taking 500 binomial random values with `n=3338` and `p=0.0165`? If one BS sample is akin to one binomial sample with n=3338 and $\hat{p}_1 = 0.0165$, then 500 BS samples is like 500 binomial samples with n=3338 and $\hat{p}_1 = 0.0165$, because each BS sample is independent!

c. Use the code below to fill in the necessary values (also, change to `eval = TRUE` so that the code will run).

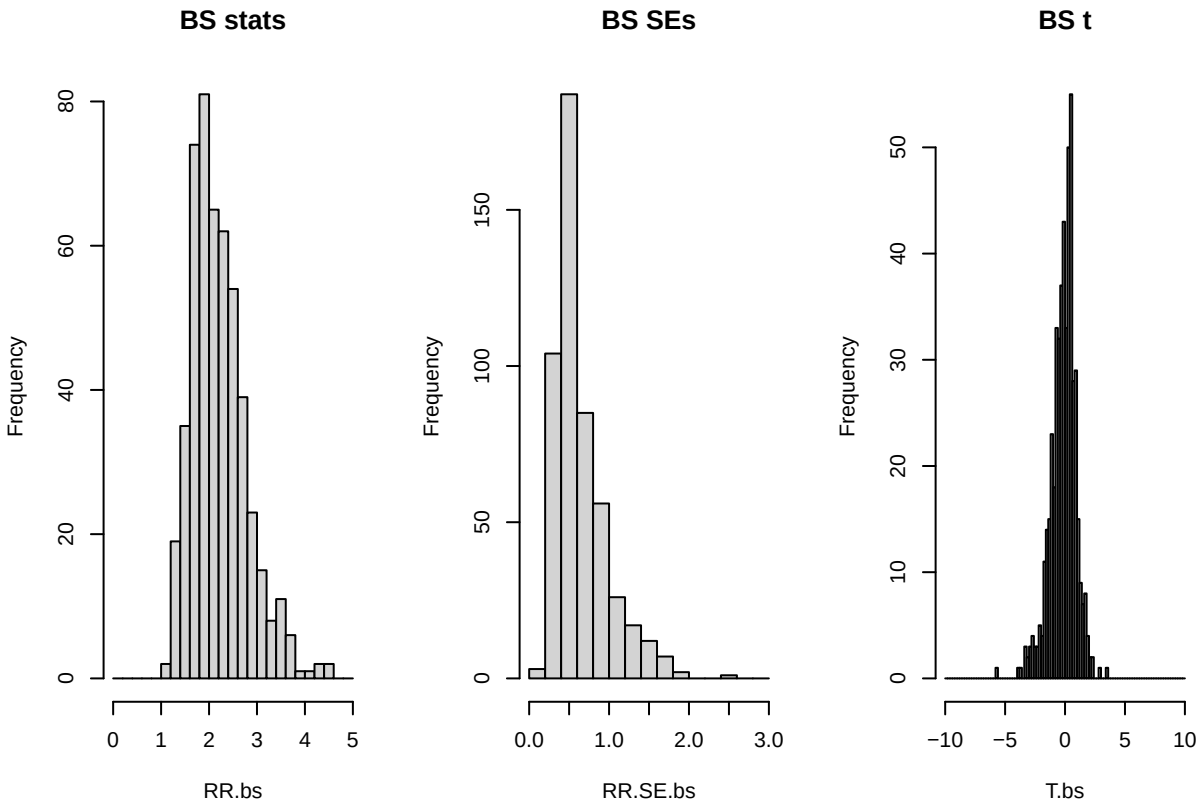
d. Create three histograms (use `hist()`) describing (1) bootstrapped statistic, (2) bootstrapped SEs, (3) bootstrapped T values.

e. Find three 95% intervals (note the `qnorm()` and `quantile()` functions in R):

   i. normal CI with BS SE
   ii. BS-t CI (don't forget to subtract the 0.975 quantile on the lower end)
   iii. BS percentile interval

f. Based on the distributions (histograms) seen above, interpret the intervals (using words like relative risk and blood pressure). Also, comment on whether or not each of the three intervals seems appropriate.

```
set.seed(10)
B = 500
M=10

#these are our first bs samples
highbp.bs <- rbinom(B, 3338, 0.0165)  # print this to see what it looks like!
lowbp.bs <- rbinom(B, 2676, 0.0078)

RR.bs <- c((highbp.bs/3338)/(lowbp.bs/2676))  # this should be a vector B long

RR.SE.bs <- c()  # will eventually be a vector that is B long

for(b in 1:B){
  highbp.bsbs <- rbinom(M, 3338, highbp.bs[b]/3338) #BS from each BS sample
  lowbp.bsbs <- rbinom(M, 2676, lowbp.bs[b]/2676)  #BS from each BS sample

  RR.bsbs <- (highbp.bsbs/3338)/(lowbp.bsbs/2676)

  RR.SE.bs <- c(RR.SE.bs, sd(RR.bsbs)) # keep the SE of the statistic from the double BS

}

T.bs <- (RR.bs-0.0165/0.0078)/RR.SE.bs

###### Histograms
par(mfrow=c(1,3))
hist(RR.bs, main="BS stats", breaks=seq(0,5,0.2))
hist(RR.SE.bs, main="BS SEs", breaks=seq(0,3,0.2))
hist(T.bs, main="BS t", breaks=seq(-10,10,0.2))
```




```
####### CIs
#normal CI
c("2.5%"=(0.0165/0.0078) #sample stat
- qnorm(0.975) * sd(RR.bs) #normal quantile and BS SE
  , "97.5%"=(0.0165/0.0078) + qnorm(0.975) * sd(RR.bs))
```

```
##      2.5%     97.5%
## 0.9744043 3.2563649
```

```
#t CI
c("2.5%"=(0.0165/0.0078) #sample stat
- quantile(T.bs, 0.975, names=FALSE) * sd(RR.bs) #t quantile and BS SE
  , "97.5%"=(0.0165/0.0078) - quantile(T.bs, 0.025, names=FALSE) * sd(RR.bs))
```

```
##     2.5%    97.5%
## 1.111216 3.681906
```

```
#percentile CI
c("2.5%"=quantile(RR.bs, 0.025, names=FALSE), "97.5%"=quantile(RR.bs, 0.975, names=FALSE))
```

```
##     2.5%    97.5%
## 1.340663 3.576716
```

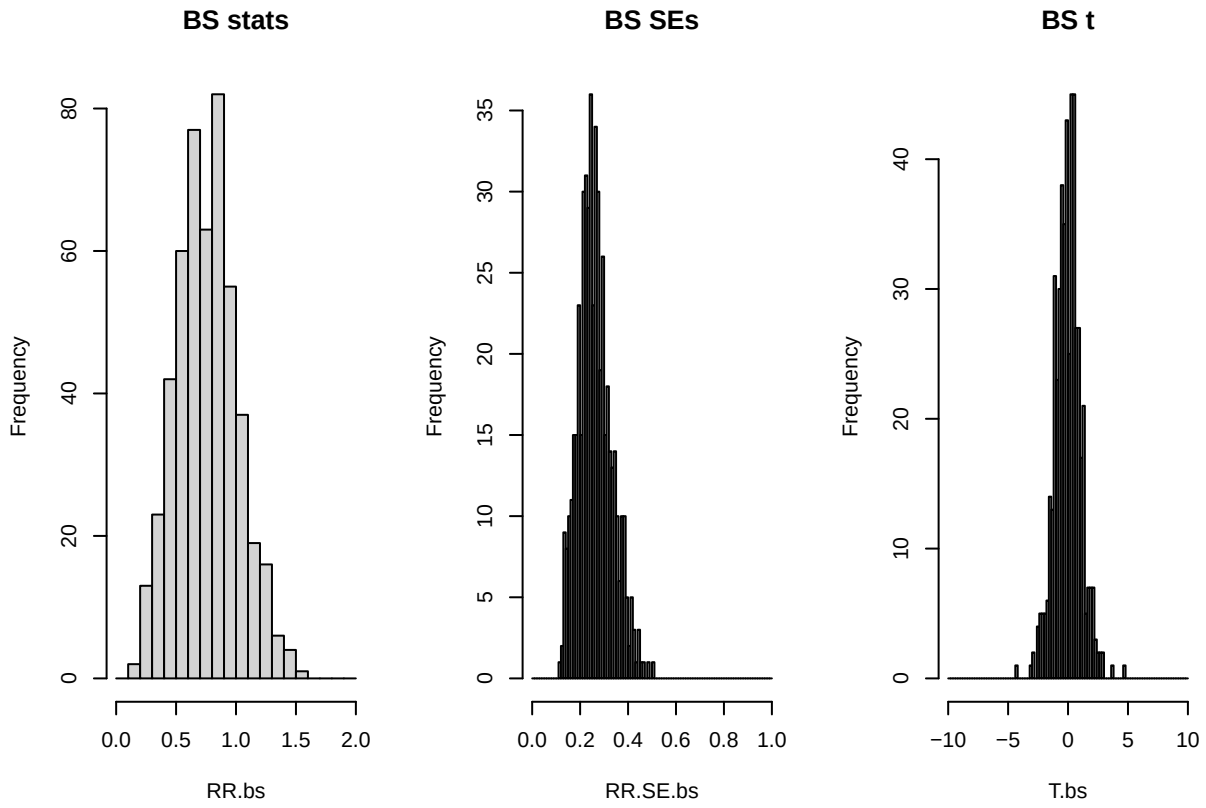
According to the Normal CI, we are 95% certain that the true RR of death in high BP compared to low BP lies within [0.97, 3.26]. According to the t-CI, we are 95% certain that the true RR of death in high BP compared to low BP lies within [1.11, 3.68]. According to the percentile-CI, we are 95% certain that the

true RR of death in high BP compared to low BP lies within [1.34, 3.58]. I believe the Bootstrap t-CI is best in this situation. The Normal CI supposes that the sampling distribution of $\hat{RR}$ is distributed normally, but the histogram of the BS RR's show that the relative risk is right skewed. Bootstrap t-CI's are better at handling skewed sampling distributions than Normal CI and percentile CI's, so I would trust the t-CI the most.

g. Repeat c, d, e, f for a new statistic: $\ln(RR)$. n.b. the correct R function is `log()`.

$$\hat{\theta} = \ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) = \ln\left(\frac{0.0165}{0.0078}\right) = \ln(0.0165) - \ln(0.0078) = 0.7418$$

```
set.seed(10)
B = 500
M=10

#these are our first bs samples
highbp.bs <- rbinom(B, 3338, 0.0165)  # print this to see what it looks like!
lowbp.bs <- rbinom(B, 2676, 0.0078)

RR.bs <- c(log((highbp.bs/3338)/(lowbp.bs/2676)))  # this should be a vector B long

RR.SE.bs <- c()  # will eventually be a vector that is B long

for(b in 1:B){
  highbp.bsbs <- rbinom(M, 3338, highbp.bs[b]/3338) #BS from each BS sample
  lowbp.bsbs <- rbinom(M, 2676, lowbp.bs[b]/2676)  #BS from each BS sample

  RR.bsbs <- log(highbp.bsbs/3338/(lowbp.bsbs/2676))

  RR.SE.bs <- c(RR.SE.bs, sd(RR.bsbs)) # keep the SE of the statistic from the double BS

}

T.bs <- (RR.bs-log(0.0165/0.0078))/RR.SE.bs

###### Histograms
par(mfrow=c(1,3))
hist(RR.bs, main="BS stats", breaks=seq(0, 2, 0.1))
hist(RR.SE.bs, main="BS SEs", breaks=seq(0,1,0.01))
hist(T.bs, main="BS t", breaks=seq(-10,10,0.2))
```




```
####### CIs
#normal CI
c("2.5%"=log(0.0165/0.0078) #sample stat
- qnorm(0.975) * sd(RR.bs) #normal quantile and BS SE
  , "97.5%"=log(0.0165/0.0078) + qnorm(0.975) * sd(RR.bs))
```

```
##      2.5%     97.5%
## 0.2527119 1.2457614
```

```
#t CI
c("2.5%"=log(0.0165/0.0078) #sample stat
- quantile(T.bs, 0.975, names=FALSE) * sd(RR.bs) #t quantile and BS SE
  , "97.5%"=log(0.0165/0.0078) - quantile(T.bs, 0.025, names=FALSE) * sd(RR.bs))
```

```
##      2.5%     97.5%
## 0.2294725 1.3042847
```

```
#percentile CI
c("2.5%"=quantile(RR.bs, 0.025, names=FALSE), "97.5%"=quantile(RR.bs, 0.975, names=FALSE))
```

```
##      2.5%     97.5%
## 0.2931577 1.2744450
```

According to the normal confidence interval, we are 95% certain that the true RR of death in high BP group compared to low BP lies within [0.253, 1.246]. According to the t CI, we are 95% certain that the true RR of death in high BP group compared to low BP lies within [0.253, 1.246]. According to the percentile CI, we

are 95% certain that the true RR of death in high BP group compared to low BP group lies within [0.293, 1.274].

All these confidence intervals seem very reasonable, and this is because the log transformation of our statistic of interest made the BS distribution of the statistic normal. Thus, the log transformation "normalizes" the sampling distribution, making the bounds on our different confidence intervals more similar, and making normal CI work well. Normality is great!