# Homework #6

**1. (Jackknife and Bootstrap, continuing from the previous h/w)** Using your knowledge of the definition expected value complete the following: One needs to estimate $\theta$, the frequency of days with 0 traffic accidents on a certain highway. The data are collected. During 40 days, there are 26 days with 0 accidents, 10 days with 1 accident, and 4 days with 2 accidents.
Statistician A estimates $\theta$ with a sample proportion $\hat{p} = 26/40 = 0.65$.
Statistician B argues that this method does not distinguish between the days with 1 accident and the days with 2 accidents, losing some valuable information. She suggests to model the number of accidents X by a Poisson distribution with parameter $\lambda$. Then we have $\theta = P\{X = 0\} = exp(-\lambda)$. She estimates $\lambda$ with $\hat{\lambda} = \bar{X}$. Then $\hat{\theta} = exp(-\hat{\lambda})$. However, this estimator is biased.

(a) Now we have three competing estimators - $\hat{p}$, $\hat{\theta}$, and $\hat{\theta_{JK}}$. Use bootstrap to estimate their standard deviations.

```r
set.seed(666)
accident <- sample( c(rep(0,26),rep(1,10),rep(2,4)) )
B <- 10000
n <- length(accident)
# estimations
p <- function(x){return( mean(x == 0) )} # p.hat
theta <- function(x){return( exp( - mean(x) ) )} # theta.hat
theta.jk <- function(x){ # theta.hat.jk
  jk <- theta(x) - jackknife( x, theta )$jack.bias
  return(jk)}
# container
p.boot <- theta.boot <- theta.jk.boot <- rep(NA,n)
# std. p.hat
for (i in 1:B){
  clone <- sample(accident, n, replace = T)
  p.boot[i] <- p(clone)
  theta.boot[i] <- theta(clone)
  theta.jk.boot[i] <- theta.jk(clone)
}
kable(cbind('$\\hat{Std}(\\hat{p})$' = sd(p.boot),
            '$\\hat{Std}(\\hat{\\theta})$' = sd(theta.boot),
            '$\\hat{Std}(\\hat{\\theta_{JK}})$' = sd(theta.jk.boot)),
      escape = F)
```

| $\hat{Std}(\hat{p})$ | $\hat{Std}(\hat{\theta})$ | $\hat{Std}(\hat{\theta_{JK}})$ |
|---|---|---|
| 0.0760843 | 0.0680744 | 0.0683877 |

(b) Compare our three estimators of $\theta$ according to their bias and standard error.
$\hat{p}$ is an unbias estimator, but since it lose some valuable information, it has the highest

standard error among three estimators. $\hat{\theta_{JK}}$ slightly reduces the bias of the $\hat{\theta}$, but also increases the standard error at the meanwhile.

**2.** We will now consider the `Boston` housing data set, from the `MASS` library.

(a) Based on this data set, provide an estimate for the population mean $\mu$ of `medv`, which is the median value of owner-occupied homes in \$1000s. Call this estimate $\hat{\mu}$.

```
mu.hat <- mean(Boston$medv)
mu.hat

## [1] 22.53281
```

The estimation: $\hat{\mu} = \sum_{i=1}^{n} medv_i = 22.53281$

(b) Provide an estimate of the standard error of $\hat{\mu}$ (as we know, $Std\bar{X} = \sigma/\sqrt{n}$).

```
n <- nrow(Boston)
s <- sd(Boston$medv)/sqrt(n)
s

## [1] 0.4088611
```

An estimate of the standard error of $\hat{\mu}$: $Std(\bar{X}) = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = 0.4088611$

(c) Now estimate the standard error of $\hat{\mu}$ using the **bootstrap**. How does this compare to your answer from (b)?

```
set.seed(666)
mu <- function(x,sample) {return( mean(x[sample]) )}
boot(Boston$medv, mu, R = 10000)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Boston$medv, statistic = mu, R = 10000)
##
##
## Bootstrap Statistics :
##     original       bias    std. error
## t1* 22.53281 0.001258636   0.4069227
```

The estimated standard error of $\hat{\mu}$ under bootstrap: $\hat{Std}(\bar{X}_{boot}) = 0.4058625$. It is very closed to the estimation in part (b).

(d) Based on your bootstrap estimate from (c), provide a 95 % confidence interval for $\mu$. A popular approximation is $\hat{\mu} \pm 2\hat{Std}(\hat{\mu})$. Compare it to the results obtained using `R` command `t.test(Boston$medv)`.

```
# bootstrap result
mu.boot <- boot(Boston$medv, mu, R = 10000)$t
cbind('lower bound' = mean(mu.boot)-2*sd(mu.boot),
      'upper bound' = mean(mu.boot)+2*sd(mu.boot))

##      lower bound upper bound
## [1,]    21.70625    23.35347

# t.test
t.test(Boston$medv)

##
##  One Sample t-test
##
## data:  Boston$medv
## t = 55.111, df = 505, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  21.72953 23.33608
## sample estimates:
## mean of x
##  22.53281
```

The 95% confidence interval for $\mu$ under bootstrap method is (21.71768, 23.35014), and that under t-test is (21.72953, 23.33608). Two intervals are very similar.

(e) Now, estimate M, the population median of `medv` with the sample median $\hat{M}$.

```
m.hat <- median(Boston$medv)
m.hat

## [1] 21.2
```

The sample median: $\hat{M} = 21.2$

(f) We now would like to estimate the standard error of $\hat{M}$, but unfortunately, there is no simple formula for computing the standard error of a sample median. Instead, estimate this standard error using the bootstrap.

```
m <- function(x,sample){return( median(x[sample]) )}
boot(Boston$medv, m, R = 10000)

##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = Boston$medv, statistic = m, R = 10000)
##
##
## Bootstrap Statistics :
##     original    bias    std. error
## t1*     21.2 -0.01053   0.3747355
```

The estimated standard error $\hat{Std}(\hat{M}_{boot}) = 0.3761129$