Stat 111 Homework 2

Kojin Oshiba

1. Sampling from major league baseball player data

(a) population mean, std dev

```
df          <- read.csv(file="baseball.csv", header=TRUE, sep=",")
K           <- nrow(df)
pop.mean    <- sum(df$Salary) / K
pop.stddev <- sqrt(sum((df$Salary-pop.mean)**2) / K)
cat("population mean:", pop.mean, '\n')
```

```
## population mean: 1183417
```

```
cat("population standard deviation:", pop.stddev)
```

```
## population standard deviation: 1389991
```

(b) random sample

The population $(y_1, ..., y_K)$ is fixed. The random sample $Y^* = (Y_1, ..., Y_5)'$ is random.

(c) std dev across sample averages

```
R = 100
calc_sample_avgs <- function(size) {
    replicate(R, {
                     sample.salaries <- sample(df$Salary,size=size,replace=T)
                     mean(sample.salaries)
                   })
}

cat("std dev across sample averages for n = 5:",sd(calc_sample_avgs(5)),'\n')
```

```
## std dev across sample averages for n = 5: 645339.4
```

```
cat("std dev across sample averages for n = 20:",sd(calc_sample_avgs(20)))
```

```
## std dev across sample averages for n = 20: 333336.3
```

The sample averages with $n = 20$ are better than those based on $n = 5$. This is because, by the law of large numbers, the standard deviation of a sample average as well as the stndard deviation across the 100 sample averages decrease. Hence, the sample averages with $n = 20$ are more likely to be a better proxy for μ.

(d)

i.

```
calc_sample_stddev_of_sample_avg <- function(size) {
  sample_avgs <- calc_sample_avgs(size)
  sqrt(sum((sample_avgs-mean(sample_avgs)) ** 2) / (R-1) )
}
cat("a sample standard deviation of a sample average using the 100 sample averages with sample size..."
```

```
## a sample standard deviation of a sample average using the 100 sample averages with sample size...
```

```
cat("n=5: ", calc_sample_stddev_of_sample_avg(5),'\n')
```

```
## n=5:  569049.8
```

```
cat("n=20:", calc_sample_stddev_of_sample_avg(20),'\n')
```

```
## n=20: 296502.3
```

```
cat("n=80:", calc_sample_stddev_of_sample_avg(80))
```

```
## n=80: 149495.6
```

ii.

salaries

iii.

sample mean salary

iv.

The standard deviation of $\bar{Y}^*$ is an unbiased estimate of σ.

2. Bootstraping from a major league baseball player sample

(a)

```
n = 20
Y.star = sample(df$Salary,size=n,replace=T)
sample.sigma.sq <- sum((Y.star-mean(Y.star)) ** 2) / (n - 1)
cat("std dev of the sample mean (formulaic):",sqrt(sample.sigma.sq / n))
```

```
## std dev of the sample mean (formulaic): 403685.3
```

(b)

```r
R = 5000
calc_bootsrap_stddev <- function (sample) {
  bootstrap <- sample(Y.star,size=n,replace=T)
  bootstrap.stddev <- sqrt(sum((bootstrap-mean(bootstrap)) ** 2) / (n - 1))
  bootstrap.stddev
}

bootsrap.stddevs <- replicate(R, calc_bootsrap_stddev(20))
cat("variance of the sample mean (bootstrap):",mean(bootsrap.stddevs))
```

```
## variance of the sample mean (bootstrap): 1742762
```

(c)

```r
bootsrap.stddevs <- replicate(R, calc_bootsrap_stddev(80))
cat("variance of the sample mean (bootstrap):",mean(bootsrap.stddevs))
```

```
## variance of the sample mean (bootstrap): 1747748
```

3. Binomial sampling

(a)

$$F_Y(y) = P(Y \leq y) = \sum_{i=0}^{\lfloor y \rfloor} \binom{7}{y} 0.1^i 0.9^{7-i}$$

(b)

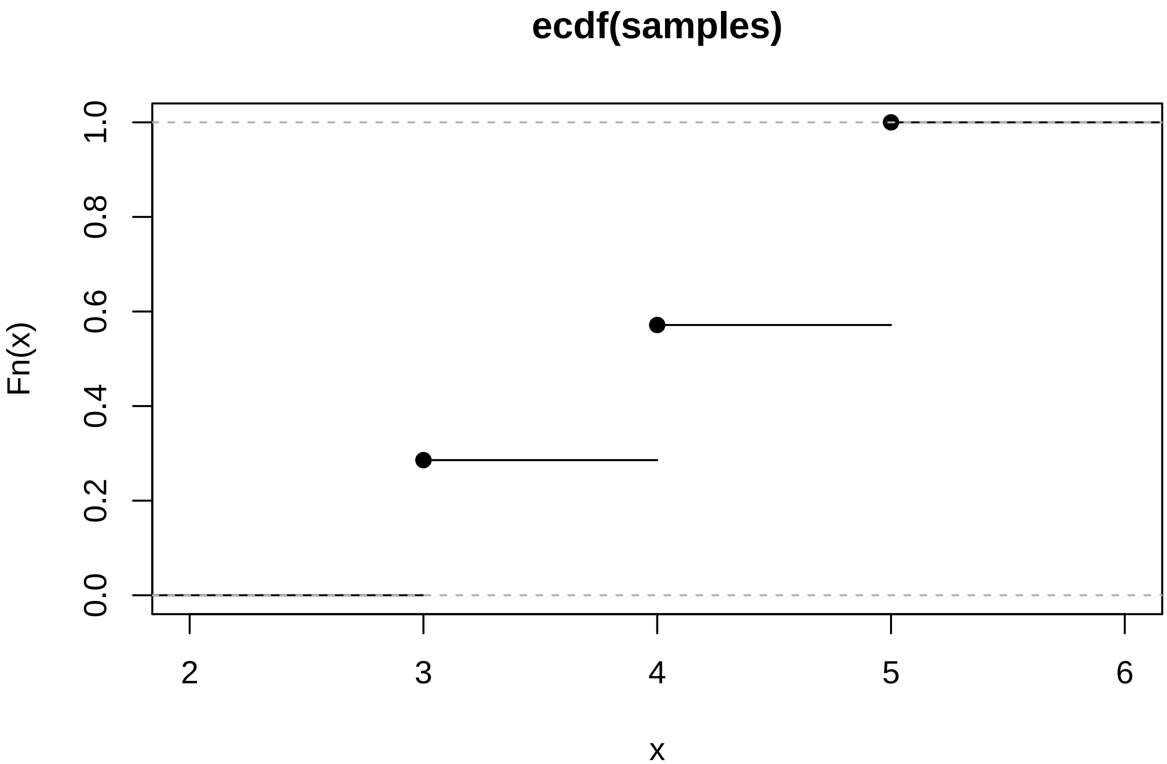
$$\begin{aligned} Var(\hat{F}_n(y)) &= Var(\frac{1}{n}\sum_{i=1}^{n} Var(\mathbf{1}_{Y_i \leq y})) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(\mathbf{1}_{Y_i \leq y}) \text{ } (\because Y_i\text{'s are sampled iid}) \end{aligned}$$

(c)

```r
n    <- 7
size <- 40
prob <- 0.1
samples <- rbinom(n, size, prob)
```

i.

```r
plot(ecdf(samples))
```



###

ii. It's pretty similar.

(d)

i.

```
ecdf.2s <- replicate(100, {
  samples <- rbinom(n, 40, prob)
  ecdf(samples)(2)
})
sd(ecdf.2s)
```

```
## [1] 0.1561229
```

ii.