Homework 5

By signing my name, I agree to abide by the Stern Code of Conduct ______________________

Question 1

This question is about the sampling distribution of the mean, and using $\bar{X}$ as an estimator for the parameter μ. Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) with the same distribution as X.

a. Recall that $E[\bar{X}] = E[X]$ and $\mathbf{Var}(\bar{X}) = \frac{\mathbf{Var}(X)}{n}$. Do we need to assume anything more about X for $\bar{X}$ to be an unbiased estimate of μ, and if so, what?

Solution: Since the bias is $E[\bar{X} - \mu] = E[\bar{X}] - \mu = E[X] - \mu$, we need to assume $E[X] = \mu$.

b. One way to write Chebyshev's inequality is $P(|Y - E[Y]| \geq k\sqrt{\mathbf{Var}(Y)}) \leq 1/k^2$. Let $Y = \bar{X}$ and write this inequality out in terms of $\bar{X}, E[X], k, \mathbf{Var}(X)$, and n.

Solution:

$$P\left(|\bar{X} - E[X]| \geq k\sqrt{\frac{\mathrm{Var}(X)}{n}}\right) \leq \frac{1}{k^2}$$

c. If we let $k = \sqrt{n}$, the answer from the previous part should simplify to $P(|\bar{X} - E[X]| \geq \sqrt{\mathbf{Var}(X)}) \leq 1/n$. If we want 99% probability that the sample mean is *within one standard deviation* of X away from $E[X]$, how large of a sample do we need?

Solution: If $1/n = 1/100$ then there is a $1 - 1/100 = 99\%$ probability of $\bar{X}$ being within σ from $E[X]$. So we need $n = 100$.

d. Suppose $\bar{X}$ is unbiased: $E[\bar{X}] = \mu$. Compute the mean squared error of this estimate, $\mathbf{MSE}(\bar{X})$.

Solution: Using the bias-variance decomposition, since we know the $\mathrm{Bias}(\bar{X})^2$ term is 0 (unbiased), we have $\mathrm{MSE}(\bar{X}) = \mathrm{Var}(\bar{X}) = \sigma^2/n$.

e. Assume now that $X \sim \mathbf{Ber}(p)$. For some constant c, $c\bar{X}$ has a distribution that you know. What is the constant, and what is the distribution?

Solution: If $c = n$ then $n\bar{X} = \sum_{i=1}^{n} X_i$ is a sum of n independent Bernoulli random variables with success probability p, hence $n\bar{X} \sim \mathrm{Bin}(n, p)$.

Question 2

This question will guide you through a *simulation study* in R to understand the bias of a certain estimator. Suppose $U_1, \ldots, U_n$ are independent and identically distributed as $U[0, \theta]$, with $\theta = 1$ and let $\hat{\theta} = \max\{U_1, \ldots, U_n\}$. Since we are generating this data ourselves, we know the true value of $\theta = 1$, so we can compute the bias of $\hat{\theta}$. Our goal will be to study how this bias decreases as the sample size n increases.

The following code creates a *function* that you can use to generate observations of $\hat{\theta}$.

```
theta_hat <- function(n) return(max(runif(n)))
```

You need to run this code once so that R learns the definition of `theta_hat`. After that, you can "call" the function by running, for example, `theta_hat(10)` to generate one observation of a maximum of sample size $n = 10$.

```
theta_hat(10)
```

```
## [1] 0.9207211
```

To generate a sample of many i.i.d. copies of $\hat{\theta}$, we use the `replicate` function:

```
replicate(10, theta_hat(10))
```

```
##  [1] 0.9004413 0.9898925 0.9436040 0.9708578 0.8303971 0.8675335 0.7628322
##  [8] 0.9685146 0.8851466 0.9208831
```

Finally, we estimate the bias by generating many samples of $\hat{\theta}$, taking their average, and subtracting θ:

```
mean(replicate(10000, theta_hat(10))) - 1
```

```
## [1] -0.09043724
```

a. Run this code again to estimate the bias when $\hat{\theta}$ is based on a sample of size $n = 100$, and again for a sample of size $n = 1000$.

Solution:

```
c(mean(replicate(10000, theta_hat(100))) - 1,
mean(replicate(10000, theta_hat(1000))) - 1)
```

```
## [1] -0.009974618 -0.001023195
```

b. Compute these answers to the answer I gave in class: compute $(n-1)/n - 1 = -1/n$ for the same values of n. Are the simulation estimates of bias reasonably close to the exact mathematical answer?

Solution: $-1/100 = -0.01, -1/1000 = -0.001$. Yes, the values from the simulation were fairly close to these.

c. Use the `sd` function to estimate the standard deviation of $\hat{\theta}$ based on $n = 10$ and on $n = 100$.

```
c(sd(replicate(10000, theta_hat(10))),
sd(replicate(10000, theta_hat(100))))
```

```
## [1] 0.083129768 0.009785584
```