SEB QBIO310 HW5

Run Cmd+Shift+Enter Insert Chunk Cmd+Option+I Preview Cmd+Shift+K

1: Measuring the duration (s) of a courtship display in Anolis lizards, using green_sample = 100. Independent observations from a Normal(theta,1) distribution...theta is unknown...estimate it

```
# set parameters
num_samples <- 10000 # num of simulated samples
n <- 100 # sample size per simulation
theta <- 10 # true parameter value
stdev <- 1 # stdev of the normal distribution
set.seed(123) # accounts for reproducibility
```

a: What are the expectation and variance of the sample mean?

```
# simulating samples...each row is one sample of size n
samps <- matrix(rnorm(n * num_samples, mean = theta, sd = stdev), nrow =
num_samples)

# computing estimators
sample_mean <- apply(samps, 1, mean)
sample_median <- apply(samps, 1, median)
sample_midrange <- apply(samps, 1, function(x) (min(x) + max(x)) / 2)
```

Sample median is now proposed as an estimator with expectation theta and variance ~ 0.0157...for larger n the variance is ~ 1.57 * lilsigma^2/n Another proposal is the mid-range = (sigma_min(x) + sigma_max(x))/2 or just the average of the largest & smallest observation...mid-range has an expectation of theta and a variance of pi^2/(24*ln(n)) where ln(n) is the natural log of n b: Which of the proposed estimators (the mean, median, and mid-range), if any, are unbiased?

```
bias_mean <- mean(sample_mean) - theta
bias_median <- mean(sample_median) - theta
bias_midrange <- mean(sample_midrange) - theta
```

c: What is the MSE of each of the three estimators for n=100?

```
# calculating variances (MSEs)
var_mean <- var(sample_mean)
var_median <- var(sample_median)
var_midrange <- var(sample_midrange)
```

d: Which of the three estimators are consistent? All 3 estimators are consistent because as n increases their variances go towards zero, meaning the estimators converge in probability to the true parameter theta. Mid-range variance decreases at 1/ln(n) while mean & median decrease in variance at 1/n

e: For n=100 what are the relative efficiences of the mean relative to the mid-range and to the median?

```
# print results
cat("Normal Model (θ =", theta, "):\n")

## Normal Model (θ = 10 ):

cat("Sample Mean: Bias =", bias_mean, ", Variance =", var_mean, "\n")

## Sample Mean: Bias = -0.000521437 , Variance = 0.0100916

cat("Sample Median: Bias =", bias_median, ", Variance =", var_median, "\n")

## Sample Median: Bias = -0.0001908136 , Variance = 0.01561694

cat("Sample Mid-range: Bias =", bias_midrange, ", Variance =", var_midrange,
"\n")

## Sample Mid-range: Bias = -0.001831898 , Variance = 0.09408903
```

f: Out of the 3 proposals, which estimator do you prefer in this scenario? Why? Under the normal model(theta, 1) the sample mean is preferred because its unbiased and has the smallest variance (smallest MSE) among the 3 estimators, as well as being the max likelihood estimator.

2: Approximately 2% of the lizards have an atypical courtship display…Binomial(n, 0.02)…described by exponential(2) distribution…estimating theta=10

```
est.dist.expoultier <- function(samp.size, est.fun, gamma = 0.02, n.samps =
10000, theta = 10, lambda = 2, sd.maj = 1){
  n.target <- rbinom(1,samp.size*n.samps, 1-gamma)
  n.nontarget <- samp.size*n.samps - n.target
  samp.target <- rnorm(n.target, theta, sd.maj)
  samp.nontarget <- rexp(n.nontarget, lambda)
  allsamps <- sample(c(samp.target, samp.nontarget))
  allsamps.mat <- matrix(allsamps, nrow=n.samps)
  return(apply(allsamps.mat, 1, est.fun))
}
set.seed(123) # accounts for reproducibility
```

a: What is the approximate bias of each of the estimators?

```
# calculating estimations first
est_mean <- est.dist.expoultier(samp.size = n, est.fun = mean)
est_median <- est.dist.expoultier(samp.size = n, est.fun = median)
est_midrange <- est.dist.expoultier(samp.size = n, est.fun = function(x)
(min(x) + max(x)) / 2)

bias_mean <- mean(est_mean) - theta
bias_median <- mean(est_median) - theta
bias_midrange <- mean(est_midrange) - theta
```

b: What is the approximate MSE of each of the three estimators at n=100?

```
mse_mean <- var(est_mean) + bias_mean^2
mse_median <- var(est_median) + bias_median^2
mse_midrange <- var(est_midrange) + bias_midrange^2
```

c: Rank these estimators in terms of their robustness to contaminating data from the exponential distribution. (not taught the expression yet so provide reason)

```
cat("Contaminated Data (θ =", theta, "):\n")

## Contaminated Data (θ = 10 ):

cat("Sample Mean: Bias =", bias_mean, ", MSE =", mse_mean, "\n")

## Sample Mean: Bias = -0.1894965 , MSE = 0.06398765

cat("Sample Median: Bias =", bias_median, ", MSE =", mse_median, "\n")

## Sample Median: Bias = -0.02570801 , MSE = 0.0171055

cat("Sample Mid-range: Bias =", bias_midrange, ", MSE =", mse_midrange, "\n")

## Sample Mid-range: Bias = -3.129753 , MSE = 11.36665
```

The most robust is the sample median has a very low bias (-0.0257) and the smallest MSE (0.0171) meaning its not heavily affected by exponential outliers. Also, because the median depends on the middle order stat rather the extreme values, it resists contamination.

d: Which estimator do you most prefer in this scenario? Why? The sample median is the preferred estimator because it has minimal bias and the lowest MSE indicate its stable and provides reliable estimates. It's resistance to the influence of extreme values makes it the best choice.

3: Continuous uniform(a,b) distribution (denying lizards with atypical displays). Use the function below to compare the performance of the 3 estimators from (1) as estimators of (a+b)/2, the expectation of the Uniform(a, b) distribution. Assume a = 7 and b = 13

```
est.dist.unif <- function(samp.size, est.fun, a, b, n.samps = 10000){
  samps <- runif(samp.size*n.samps, min = a, max = b)
  samp.mat <- matrix(samps, nrow = n.samps)
  return(apply(samp.mat, 1, est.fun))
}

a <- 7
b <- 13
set.seed(123) # for reproducibility
```

a: What is the approximate bias of each of the estimators?

```
unif_mean <- est.dist.unif(samp.size = n, est.fun = mean, a = a, b = b)
unif_median <- est.dist.unif(samp.size = n, est.fun = median, a = a, b = b)
```

```r
unif_midrange <- est.dist.unif(samp.size = n, est.fun = function(x) (min(x) +
max(x)) / 2, a = a, b = b)

true_mean <- (a + b) / 2

bias_mean <- mean(unif_mean) - true_mean
bias_median <- mean(unif_median) - true_mean
bias_midrange <- mean(unif_midrange) - true_mean
```

b: What is the approximate MSE of each of the three estimators?

```r
mse_mean <- var(unif_mean) + bias_mean^2
mse_median <- var(unif_median) + bias_median^2
mse_midrange <- var(unif_midrange) + bias_midrange^2
```

c: Which of the three estimators are consistent? (Don't need to attempt to prove it, make a conjecture and justify it on the basis on simulated samples up to size 10,000)

```r
cat("Uniform Distribution (a =", a, "b =", b, "):\n")
```

```
## Uniform Distribution (a = 7 b = 13 ):
```

```r
cat("Sample Mean: Bias =", bias_mean, ", MSE =", mse_mean, "\n")
```

```
## Sample Mean: Bias = -0.002826009 , MSE = 0.03029357
```

```r
cat("Sample Median: Bias =", bias_median, ", MSE =", mse_median, "\n")
```

```
## Sample Median: Bias = 0.004883452 , MSE = 0.08562295
```

```r
cat("Sample Mid-range: Bias =", bias_midrange, ", MSE =", mse_midrange, "\n")
```

```
## Sample Mid-range: Bias = 0.0001059179 , MSE = 0.001741446
```

All 3 estimators are consistent because the biases are all nearly zero and we expect these biases and their corresponding MSes to decrease further as n increases. The rates at which their variances decrease might differ, all 3 converge in probability to the true mean (10) for the uniform(7, 13) distribution

d: For n=100, what is the relative efficiency of the midrange relative to the mean and to the median? The relative efficiency of the mid-range relative to the mean is the (sample mean MSE) / (sample midrange MSE) ~ 17.4 The relative efficiency of the midrange relative to the median is thhe (sample median MSE) / (sample midrange MSE) ~ 49.2 This means the midrange is about 49x more efficient than the sample mean.

e: Which estimator do you most prefer in this scenario? Why? The sample midrange is the most preferred in this scenario because its MSE is dramatically lower, has very little bias, and its very suitable for the (7, 13) data distribution.