

## Statistics 406 Problem Set 2

Due in lab, Tuesday October 3

1. Use simulation to estimate the bias, variance and MSE when using the sample median to estimate the population mean for standard normal and standard exponential populations. Consider sample sizes 10,20,40, and 80.

**Solution:** Here is the simulation code for Gaussian data. This is actually more than I asked for – here I get bias, variance and MSE for both the sample mean and the sample median.

```
bias_med <- NULL
Var_med <- NULL
MSE_med <- NULL

bias_mean <- NULL
Var_mean <- NULL
MSE_mean <- NULL

for (n in c(10,20,40,80))
{
  Z <- array(rnorm(n*1000), c(n,1000))
  Med <- apply(Z, 2, median)
  Mean <- apply(Z, 2, mean)

  bias_med <- c(bias_med, mean(Med) - 0)
  Var_med <- c(Var_med, var(Med))
  MSE_med <- c(MSE_med, mean((Med-0)^2))

  bias_mean <- c(bias_mean, mean(Mean) - 0)
  Var_mean <- c(Var_mean, var(Mean))
  MSE_mean <- c(MSE_mean, mean((Mean-0)^2))
}
```

My results for Gaussian data are below. The biases for both the mean and median are negligible. The sample median has greater variance and greater MSE than the sample mean for all sample sizes.

```
> bias_med
[1] -0.002481673  0.002527976  0.004563670  0.004829862
> Var_med
```

```

[1] 0.13283735 0.07122519 0.03805404 0.01963545
> MSE_med
[1] 0.14546650 0.07466551 0.03911633 0.01919594
> bias_mean
[1] -0.0099502337 0.0042899165 0.0002357599 0.0031641444
> Var_mean
[1] 0.09995863 0.04981747 0.02467007 0.01272365
> MSE_mean
[1] 0.10519139 0.05117167 0.02553403 0.01250219

```

Here is the simulation code for standard exponential data (again, I only asked for the results for the sample mean).

```

bias_med <- NULL
Var_med <- NULL
MSE_med <- NULL

bias_mean <- NULL
Var_mean <- NULL
MSE_mean <- NULL

for (n in c(10,20,40,80))
{
  Z <- array(rexp(n*1000), c(n,1000))
  Med <- apply(Z, 2, median)
  Mean <- apply(Z, 2, mean)

  bias_med <- c(bias_med, mean(Med)-1)
  Var_med <- c(Var_med, var(Med))
  MSE_med <- c(MSE_med, mean((Med-1)^2))

  bias_mean <- c(bias_mean, mean(Mean)-1)
  Var_mean <- c(Var_mean, var(Mean))
  MSE_mean <- c(MSE_mean, mean((Mean-1)^2))
}

```

The results for standard exponential data are below. Since the exponential distribution is right skewed, the population median is less than the population mean. Therefore the sample median is a negatively biased estimate of the population mean. The mean and median have similar variances in this case. The MSE for the median is much greater than the MSE for the mean.

```

> bias_med
[1] -0.2625701 -0.2827035 -0.3024182 -0.2962560
> Var_med
[1] 0.08960602 0.04555924 0.02467061 0.01326463
> MSE_med
[1] 0.1584595 0.1254349 0.1161027 0.1010190
> bias_mean
[1] -0.0019157912 -0.0011267070 -0.0050941639 -0.0006777711
> Var_mean
[1] 0.09066073 0.04823082 0.02528067 0.01309253
> MSE_mean
[1] 0.09057373 0.04818386 0.02528134 0.01307990

```

2. Use simulation to estimate the bias, variance and MSE when using the sample median to estimate the population median for standard normal and standard exponential populations. Consider sample sizes 10, 20, 40 and 80.

You will first need to calculate the population median values of the standard normal and exponential distributions. If  $p(x)$  is the density function, the median value  $Q$  satisfies the equation

$$\int_{-\infty}^Q p(x)dx = 1/2$$

for the normal population, and

$$\int_0^Q p(x)dx = 1/2$$

for the exponential population. You will need to calculate the exponential median directly. The normal median can be arrived at without explicit calculation by considering the symmetry of the normal density.

**Solution:** The exponential median is determined by

$$1/2 = \int_{-\infty}^Q \exp(-x)dx = \left|_0^Q -\exp(-x) \right| = 1 - \exp(-Q)$$

which gives  $Q = \log(2)$ .

```

Bias_normal <- NULL
Var_normal <- NULL
MSE_normal <- NULL

## Get the bias, variance and MSE for a normal population.
for (n in c(10,20,40,80))
{
  Z <- array(rnorm(n*1000), c(n,1000))
  Z <- apply(Z, 2, median)

  Bias_normal <- c(Bias_normal, mean(Z))
  Var_normal <- c(Var_normal, var(Z))
  MSE_normal <- c(MSE_normal, mean(Z^2))
}

Bias_exp <- NULL
Var_exp <- NULL
MSE_exp <- NULL

## Get the bias, variance and MSE for an exponential population.
for (n in c(10,20,40,80))
{
  Z <- array(rexp(n*1000), c(n,1000))
  Z <- apply(Z, 2, median)

  Bias_exp <- c(Bias_exp, mean(Z)-log(2))
  Var_exp <- c(Var_exp, var(Z))
  MSE_exp <- c(MSE_exp, mean((Z-log(2))^2))
}

```

My results are below. The bias is essentially zero for normal data, so in that case the variance and MSE are nearly the same. For exponential data, the bias is small for large sample sizes, but moderate for small sample sizes.

The sample median is unbiased for symmetric distributions, and is approximately unbiased for any distribution when the sample size is large. For small sample sizes, the sample median is interpolated between two values. For example, when  $n = 10$  the sample median is the average of the 5<sup>th</sup> and 6<sup>th</sup> smallest elements. Due to the skew in the exponential distribution, this is slightly biased (we would be better off taking something closer to the 6<sup>th</sup> smallest element).

```
> Bias_normal
```

```

[1] 0.0053768355 0.0015853875 0.0009402255 -0.0017941483
> Var_normal
[1] 0.14674663 0.07106060 0.03996389 0.02120809
> MSE_normal
[1] 0.14662879 0.07099206 0.03992481 0.02119010
> Bias_exp
[1] 0.058244364 0.029960389 0.002982079 0.009224184
> Var_exp
[1] 0.09937712 0.05197702 0.02701005 0.01308496
> MSE_exp
[1] 0.10468953 0.05286045 0.02718827 0.01309556

```

3. Modify the simulations given in the notes for clustered data to calculate the bias, variance, and MSE for each cluster separately (i.e. do not average the MSE's across clusters, or select a single cluster as is done in the notes). You should use the same population cluster mean values for all simulation replications. Vary  $\lambda$  in increments of 0.05 to identify the best value of  $\lambda$  in terms of average MSE across the clusters.

**Solution:** Here is the code that is relevant to all parts of the question:

```

## Number of clusters.
nc <- 100

## Number of observations per cluster.
cs <- 5

## Shrinkage factor.
F <- seq(0, 1, 0.05)

## Save all estimates.
Estimate <- array(0, c(nc,1000))

## True cluster means.
CM <- rnorm(nc)

## Storage for bias, variance and MSE results.
MSE <- array(0, c(nc, length(F)))
Bias <- array(0, c(nc, length(F)))
Var <- array(0, c(nc, length(F)))

## Shrinkage factors.
for (k in 1:length(F))

```

```

{
  ## Simulation replications.
  for (r in 1:1000)
  {
    ## Estimated cluster means.
    Z <- array(0, nc)

    for (j in 1:nc)
    {
      z <- CM[j] + rnorm(cs)
      Z[j] <- mean(z)
    }

    ## The 'grand mean'
    GM <- mean(Z)

    ## Modify the cluster mean estimates by shrinking toward
    ## the grand mean.
    Estimate[,r] <- GM + F[k]*(Z-GM)
  }

  ## Bias, variance, and MSE results for each observation.
  for (j in 1:nc)
  {
    Bias[j,k] <- mean(Estimate[j,]) - CM[j]
    Var[j,k] <- var(Estimate[j,])
    MSE[j,k] <- mean((Estimate[j,] - CM[j])^2)
  }
}

## Use this to identify the best value of lambda. I get that the best
## value of lambda is 0.85, in column 17.
MSE_mean <- apply(MSE, 2, mean)

## Use these to answer part a.
Bias2_mean <- apply(Bias^2, 2, mean)
Var_mean <- apply(Var, 2, mean)
MSE_mean <- apply(MSE, 2, mean)

```

- (a) In the  $MSE = \text{bias}^2 + \text{variance}$  decomposition for the optimal value of  $\lambda$ , is the MSE mostly due to bias or mostly due to variance?

**Solution:** The optimal value of  $\lambda$  is around 0.85 (the smallest element of `MSE_mean`)

which occurs in column 17, see below). At  $\lambda = 0.85$ , the variance contributes about three times more to MSE than the squared bias. This can be seen from the following:

```
> MSE_mean[17]
[1] 0.1680416
> Var_mean[17]
[1] 0.1279959
> Bias2_mean[17]
[1] 0.04017362
```

- (b) Make scatterplots of the bias against the population cluster mean, the variance against the population cluster mean, and the MSE against the population cluster mean (three scatterplots in all).

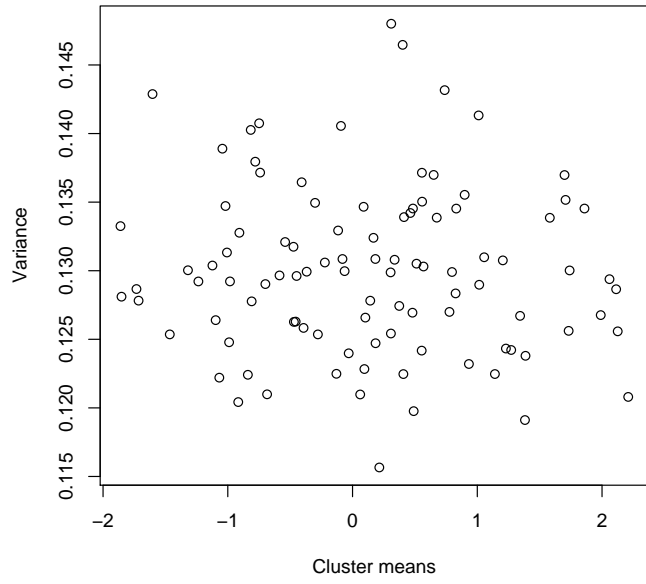
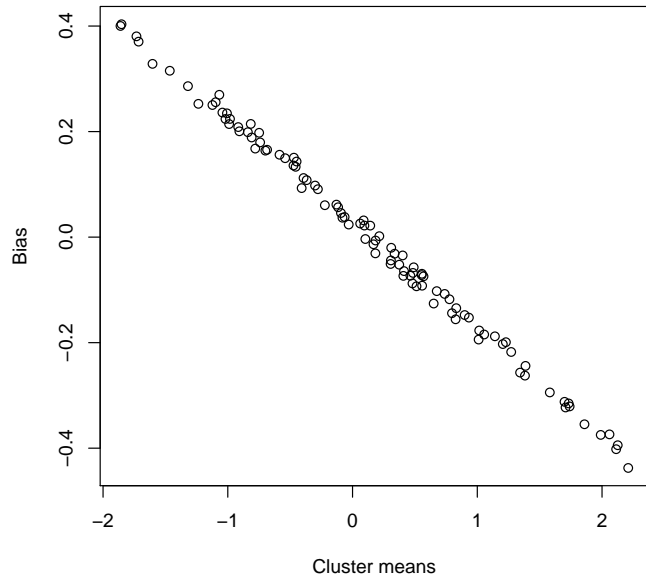
**Solution:** Use this code to generate the scatterplots:

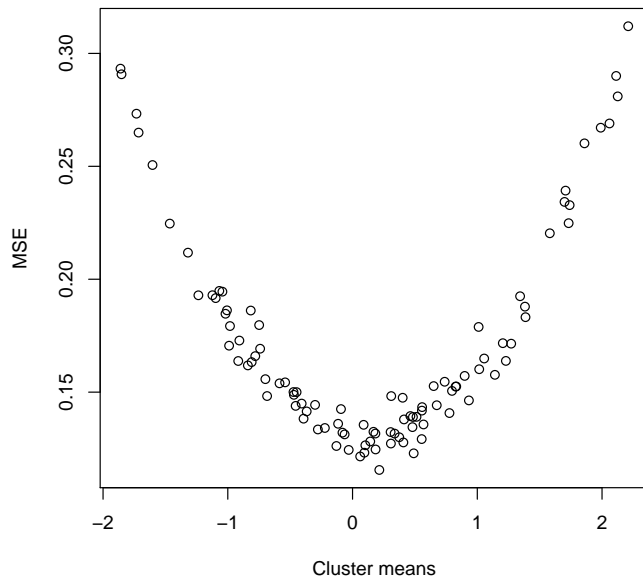
```
pdf('ps03-1.pdf')
plot(CM, Bias[,17], t='p', xlab='Cluster means', ylab='Bias')
dev.off()
```

```
pdf('ps03-2.pdf')
plot(CM, Var[,17], t='p', xlab='Cluster means', ylab='Variance')
dev.off()
```

```
pdf('ps03-3.pdf')
plot(CM, MSE[,17], t='p', xlab='Cluster means', ylab='MSE')
dev.off()
```

The scatterplots are below. They show that the MSE is smallest for clusters near the center of the distribution. The variances of the estimators vary from around 0.12 to 0.14 (substantially lower than  $1/5 = 0.2$  which would be the variance of the cluster averages). The bias is near zero for clusters at the center of the distribution. For clusters far from the center, the bias is substantial, and always points toward zero (e.g. cluster with low means are positively biased and clusters with high means are negatively biased).





- (c) Approximately how many individual clusters have greater MSE at the optimal value of  $\lambda$  compared to using  $\lambda = 0$ ?

**Solution:** To get this value, use

```
> sum(MSE[,17] > MSE[,1])
[1] 28
```

Therefore about 3/4's of the cluster means are accurated more accurately using this approach.

- (d) Is there a strong relationship between either bias or variance and the population cluster means?

**Solution:** The variance is not strongly related to the population cluster means. The bias is strongly related, in that clusters near the center of the range are approximately unbiased, while clusters in the tails of the range are biased toward the center.

- (e) Describe in terms of bias and variance why the overall MSE is less for certain nonzero values of  $\lambda$  compared to using  $\lambda = 0$ .

**Solution:**

When  $\lambda = 0$  the variance is  $1/5 = 0.2$  and the bias is zero for every cluster. When  $\lambda > 0$ , e.g.  $\lambda = 0.8$ , the variance is substantially reduced, to a value around 0.13, but at the cost of some bias. For the majority of clusters falling near the center of the

range, the bias is still relatively small. Therefore you come out ahead in terms of MSE. For around a fourth of the clusters, the squared bias is so large that it overwhelms the improvement in variance. However the average performance across the 100 clusters is still better.