Review of Fundamentals

In Statistics	In R
Y $\sim$ N(0,1)	`reps = 1e4` `Y = rnorm(reps)`
EY	`EY = mean(Y)`
var$Y$	`VY = var(Y)`
SD(Y)	`SY = sd(Y)`
IQR(Y)	`IY = IQR(Y)`
$f(Y) = \cos(Y)$	`fY = cos(Y)`
$E[\cos(Y)]$	`EfY = mean(fY)`
$P(Y > 2)$	`prb = mean(Y > 2)`
$P(\|Y\| > 1)$	`prb = mean(abs(Y) > 1)`

In Statistics	In R
$Y_1, \ldots, Y_n$ iid N(0,1)	`reps = 1e4` `n = 30` `Y = array(rnorm(n*reps), c(n,reps))`
$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$	`Ybar = apply(Y,2,mean)`
$E\bar{Y}$	`EYbar = mean(Ybar)`
$\text{var}\bar{Y}$	`VYbar = var(Ybar)`
$\text{IQR}(\bar{Y})$	`IYbar = IQR(Ybar)`
$\text{IQR}(Y_1, \ldots, Y_n)$	`IY = apply(Y,2,IQR)`
$E[\text{IQR}(Y_1, \ldots, Y_n)]$	`EIY = mean(IY)`
$\max(Y_1, \ldots, Y_n)$	`MY = apply(Y,2,max)`
$E[\max(Y_1, \ldots, Y_n)]$	`EMY = mean(MY)`
$\text{var}[\max(Y_1, \ldots, Y_n)]$	`VMY = var(MY)`
$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	`SampVar = apply(Y,2,var)`
$E\hat{\sigma}^2$	`ESampVar = mean(SampVar)`
$\text{var}\hat{\sigma}^2$	`VSampVar = var(SampVar)`

Estimation

We use sample statistics to estimate population parameters. This means that we have a sequence of random variables, $Y_1, \ldots, Y_n$ that represent our data, we then use these n random variables to estimate a population parameter. The sample statistic is just a function of these random variables, and hence the sample statistic is itself a random variable. Knowing the distribution of the random variable that is the sample statistic is important in order to attach meaning to any realization of it.

Let $Y_1, \ldots, Y_n$ be an independent and identically distributed sequence of random variables, let Y denote the random variable with the same distribution as any of the Y_i in our data sequence.

Sample Statistic $\hat{\theta}$	Population Parameter θ
$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$	EY
$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$	$\text{var}Y$
$\min(Y_1, \ldots, Y_n)$	$\min Y$
$\text{IQR}(Y_1, \ldots, Y_n)$	$\text{IQR}(Y)$
$\text{median}(Y_1, \ldots, Y_n)$	$\text{median}(Y)$

In General

$\hat{\theta} = h(Y_1, \ldots, Y_n)$	$g(\theta)$

Note that there could be many different sample statistics that estimate the same population parameter. We can assess the performance of these sample statistics (random variables) at estimating population parameters, by looking at parameters of the distribution of the sample statistic. In particular, we would like to know parameters like $E\hat{\theta}$ and $\text{var}\hat{\theta}$. Knowing these can help us evaluate the quality of the sample statistic. A natural hope is that $E\hat{\theta} = \theta$ and $\text{var}\hat{\theta}$ is very small.

Estimation Performance

Given a sample statistic $\hat{\theta}$, which is a random variable, we can assess its performance at estimating a population parameter θ by looking at its bias, variance, and mean-squared-error.

$$\text{Bias } \hat{\theta} = E\hat{\theta} - \theta$$

$$\text{MSE } \hat{\theta} = E(\hat{\theta} - \theta)^2$$

If one expands the square in the formula for the MSE, he/she would find that the MSE is equivalently expressed as:

$$\text{MSE } \hat{\theta} = \text{var } \hat{\theta} + \text{Bias}^2\ \theta$$

When estimating θ we want the MSE of our sample statistic $\hat{\theta}$ to be as small as possible, thus we want both the variance and the bias-squared of our sample statistic to be as small as possible. In some situations, choosing a biased estimator can yield better estimation performance than a competing unbiased estimator. How? Having the bias-squared increase from 0 to a small number can reduce the variance drastically, which can in turn drastically reduce the MSE.

Examples

1. Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. Conduct a simulation study to assess the performance of the sample mean at estimating the population mean.

   Solution:

   The sample statistic is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. The population parameter is $EX = \mu$. In this scenario we know from lecture that the distribution of $\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$. Thus $E\bar{X} = \mu$, and $\text{var}\bar{X} = \frac{\sigma^2}{n}$. Lets use R to simulate this:

```
## Population Parameters:
mu=4
sigma2=3

reps=1e4
n=30
X = array(rnorm(n*reps, mean=mu, sd=sqrt(sigma2)),c(n,reps))
Xbar = apply(X,2,mean)
EXbar = mean(Xbar)
VXbar = var(Xbar)

## Simualted
> c(EXbar, VXbar)
[1] 3.99777586 0.09941788

## Actual
> c(mu, sigma2/n)
[1] 4.0 0.1
```

   We can immediately find the bias and MSE of the sample mean:

$$\text{Bias } \bar{X} = E\bar{X} - \mu = \mu - \mu = 0$$

$$\text{MSE } \bar{X} = \text{var } \bar{X} + \text{Bias}^2 \ \bar{X} = \frac{\sigma^2}{n} + 0^2 = \frac{\sigma^2}{n}$$

   Lets use simulation to make sure that our MSE formula is correct:

```
MSE.Xbar = mean((Xbar - mu)^2)
> c(MSE.Xbar, sigma2/n)
[1] 0.09941288 0.10000000
```

Note that one could calculate the MSE of the sample mean using the formula:

```
## Using var() normalizes by 1/(reps-1)
MSE.Xbar = var(Xbar) + (mean(Xbar) - mu)^2
> c(MSE.Xbar, sigma2/n)
[1] 0.09942282 0.10000000

## Change this to normalize by 1/reps:
MSE.Xbar = var(Xbar)*(reps-1)/reps + (mean(Xbar) - mu)^2
>  c(MSE.Xbar, sigma2/n)
[1] 0.09941288 0.10000000
## same answer as the direct calculation.
```

2. Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. Conduct a simulation study to compare the performance of the $\frac{1}{n-1}$-scaled sample variance to the $\frac{1}{n}$-scaled sample variance for estimating the population variance. Specifically compare estimators:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

```
## Population Parameters
mu=4
sigma2=3
reps=1e4

MSE=NULL
BIAS2=NULL
VAR=NULL

for ( n in c(5,10,30,100,200,500) )
{
  ## Generate data
  X = array(rnorm(n*reps, mean=mu, sd=sqrt(sigma2)),c(n,reps))

  ## Compute the two estimators
  sigmaHat2 = apply(X,2,var)
  sigmaTilde2 = sigmaHat2*(n-1)/n

  ## Compute MSE, Bias, and Variance
  MSE.sH2 = mean((sigmaHat2-sigma2)^2)
  BIAS.sH2 = mean(sigmaHat2) - sigma2
  VAR.sH2 = MSE.sH2 - BIAS.sH2^2

  MSE.sT2 = mean((sigmaTilde2 - sigma2)^2)
  BIAS.sT2 = mean(sigmaTilde2) - sigma2
  VAR.sT2 = MSE.sT2 - BIAS.sT2^2

  ## store the results:
  MSE = rbind(MSE , c(n, MSE.sH2, MSE.sT2))
  BIAS2 = rbind(BIAS2, c(n, BIAS.sH2^2,  BIAS.sT2^2))
  VAR = rbind(VAR, c(n, VAR.sH2, VAR.sT2))
}
```

```
> MSE
     [,1]       [,2]       [,3]
[1,]    5 4.45883709 3.22123065
[2,]   10 2.05951595 1.74764565
[3,]   30 0.62477765 0.59611591
[4,]  100 0.18445330 0.18199031
[5,]  200 0.09041425 0.08978640
[6,]  500 0.03525574 0.03517871
```

Based on these results for $\sigma^2 = 3$, the sample statistic $\tilde{\sigma}^2$ has a lower MSE than $\hat{\sigma}^2$ for estimating the population variance σ^2. This is particularly noticeable for small sample sizes. Why did this happen? Why don't we use $\tilde{\sigma}^2$ all of the time? Let's look at the Bias/Variance decomposition of the MSE's we calculated:

```
> BIAS2
     [,1]         [,2]         [,3]
[1,]    5 6.226055e-05 3.676148e-01
[2,]   10 3.825840e-04 7.974763e-02
[3,]   30 1.410243e-04 1.242768e-02
[4,]  100 2.682132e-05 1.233916e-03
[5,]  200 2.699204e-06 2.767136e-04
[6,]  500 5.407065e-06 6.923337e-05
> VAR
     [,1]       [,2]       [,3]
[1,]    5 4.45877483 2.85361589
[2,]   10 2.05913337 1.66789803
[3,]   30 0.62463662 0.58368822
[4,]  100 0.18442648 0.18075640
[5,]  200 0.09041155 0.08950969
[6,]  500 0.03525033 0.03510947
```

It is clear that $\hat{\sigma}^2$ is unbiased as we know from theory; however, one could be more inclined to use $\tilde{\sigma}^2$ because it has a lower MSE for $\sigma^2 = 3$. It turns out that $\tilde{\sigma}^2$ has a lower MSE than $\hat{\sigma}^2$ regardless of the value of σ^2. Let's derive this: From the class notes we have the fact that for the normal data considered here,

$$\begin{aligned}
E\hat{\sigma}^2 &= \sigma^2 \\
\text{Bias}^2\ \hat{\sigma}^2 &= (\sigma^2 - \sigma^2)^2 = 0^2 = 0 \\
\text{var}\ \hat{\sigma}^2 &= \sigma^4 \frac{2}{n-1} \\
\text{MSE}\ \hat{\sigma}^2 &= \text{Bias}^2\ \hat{\sigma}^2 + \text{var}\ \hat{\sigma}^2 \\
&= 0^2 + \sigma^4 \frac{2}{n-1} \\
&= \sigma^4 \frac{2}{n-1}
\end{aligned}$$

Now for $\tilde{\sigma}^2$ we know that $\tilde{\sigma}^2 = \frac{n-1}{n}\hat{\sigma}^2$, so we have:

$$\begin{aligned}
E\tilde{\sigma}^2 &= E\frac{n-1}{n}\hat{\sigma}^2 = \frac{n-1}{n}\sigma^2 \\
\text{Bias}^2\ \tilde{\sigma}^2 &= (\frac{n-1}{n}\sigma^2 - \sigma^2)^2 = \frac{1}{n^2}\sigma^4 \\
\text{var}\ \tilde{\sigma}^2 &= \left(\frac{n-1}{n}\right)^2 \text{var}\ \hat{\sigma}^2 = \sigma^4 \frac{2}{n-1}\left(\frac{n-1}{n}\right)^2 \\
\text{MSE}\ \tilde{\sigma}^2 &= \text{Bias}^2\ \tilde{\sigma}^2 + \text{var}\ \tilde{\sigma}^2 \\
&= \frac{1}{n^2}\sigma^4 + \sigma^4 \frac{2}{n-1}\left(\frac{n-1}{n}\right)^2 \\
&= \sigma^4\left(\frac{2}{n} - \frac{1}{n^2}\right)
\end{aligned}$$

Thus for $n \geq 2$:

$$\begin{aligned}
\text{MSE}\ \hat{\sigma}^2 &= \sigma^4 \frac{2}{n-1} \\
\text{MSE}\ \tilde{\sigma}^2 &= \sigma^4\left(\frac{2}{n} - \frac{1}{n^2}\right)
\end{aligned}$$

Hence MSE $\tilde{\sigma}^2 <$ MSE $\hat{\sigma}^2$. There's more to the story, one can verify that $\tilde{\sigma}^2$ is also the MLE of σ^2. So when I approach applied statistics problems for which I need to estimate the population variance of normal data, I prefer $\tilde{\sigma}^2$. For large sample sizes, the difference is very small; however, given the knowledge that $\tilde{\sigma}^2$ has a lower MSE than its competitor, one is responsible to act.

In addition, if we knew μ in advance, but σ^2 was unknown, then:

$$\frac{1}{n}\sum_i (X_i - \mu)^2$$

is an unbiased estimator of σ^2.