Variance and Standard Deviation Bias

davegoblue

Updated March 6, 2016

Background and Overview

Variance and standard deviation can be calculated on a population (divide by n) or sample (divide by n-1) basis. This code examines the methodology impact for 2,000 draws each of size-n from a normal distribution and a constant distribution. Each population has a known variance and standard deviation.

Experiment results confirm findings expected from the lierature, namely that using the “population variance” calculation on a sample produces a biased variance estimate, while using the “sample variance” calculation on a sample produces an inbiased variance estimate.

Due to concavity of SQRT(x), the standard deviation estimate is biased (too low) under both the “population sd” and “sample sd” methods. This is confirmed as a known artifact by the literature, and there are technical adjustments that can be (but usually are not) applied if an unbiased standard deviation estimate is needed.

Analysis and Results

First, we set up some counter variables and pre-set the random seed for reproducibility.

```
set.seed(0306160823) ## Set seed for reproducibility

myMax <- 50  ## Draws will be of sample size 1:myMaz
myDraws <- 2000  ## myDraws samples taken for each sample size

## These variables will hold the results from each draw
sdSam <- rep(0,myDraws)
sdPop <- rep(0,myDraws)
varSam <- rep(0,myDraws)
varPop <- rep(0,myDraws)

## These variables will hold the averages across draw for each sample size
mySDSam <- rep(0,myMax)
mySDPop <- rep(0,myMax)
myVarSam <- rep(0,myMax)
myVarPop <- rep(0,myMax)

## Hold the bias by metric, methodology, population, and sample size
## Bias will be calculated as (Estimate - True) / True
estBiasSD <- data.frame(knownPop=c(rep("Normal", 2*myMax),
                                   rep("Uniform", 2*myMax)
                                   ),
                        calcMethod=rep(c(rep("Population",myMax),
                                         rep("Sample",myMax)
                                         )
                                       ,2),
                        sampSize=rep(1:myMax,4),
                        biasEstimate=rep(0,4*myMax),
                        stringsAsFactors = FALSE
                        )

estBiasVar <- data.frame(knownPop=c(rep("Normal", 2*myMax),
                                    rep("Uniform", 2*myMax)
                                    ),
                         calcMethod=rep(c(rep("Population",myMax),
                                          rep("Sample",myMax)
                                          )
                                        ,2),
                         sampSize=rep(1:myMax,4),
                         biasEstimate=rep(0,4*myMax),
                         stringsAsFactors = FALSE
                         )
```

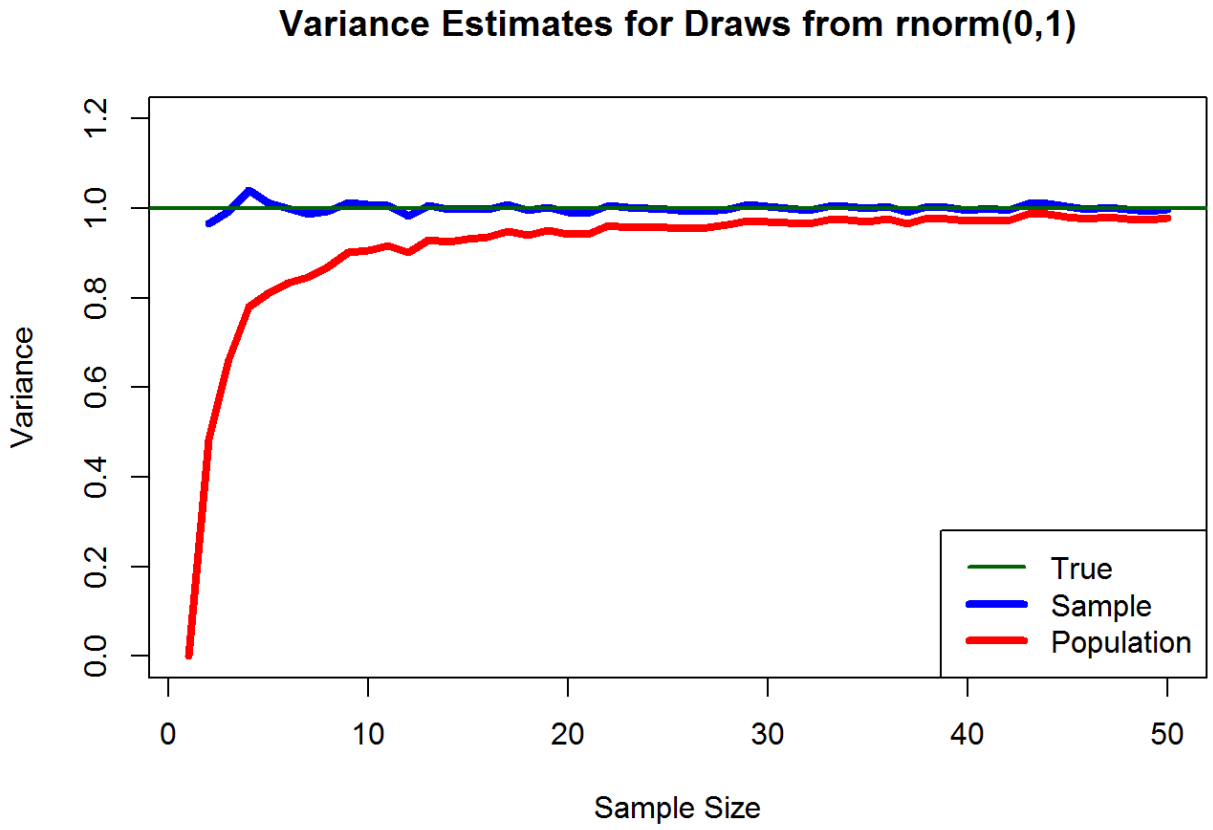
Results from the "standard normal" rnorm(0,1) distribution

We start with the rnorm(0,1) distribution and pull samples of size 1:myMax, each taken myDraws number of times. The standard deviations and variances are calculated on a population and sample basis and then averaged by sample size for later reporting.

```
for (intCtr in 1:myMax) {

    for (intCtr2 in 1:myDraws) {
        a <- rnorm(intCtr,mean=0,sd=1)
        sdSam[intCtr2] <- sd(a)
        sdPop[intCtr2] <- sd(c(a,mean(a))) ## numerator the same, denom + 1
        varSam[intCtr2] <- var(a)
        varPop[intCtr2] <- var(c(a,mean(a)))  ## same as above
    }

    mySDSam[intCtr] <- mean(sdSam)
    mySDPop[intCtr] <- mean(sdPop)
    myVarSam[intCtr] <- mean(varSam)
    myVarPop[intCtr] <- mean(varPop)

    estBiasSD[estBiasSD$knownPop=="Normal" &
              estBiasSD$calcMethod=="Sample" &
              estBiasSD$sampSize==intCtr,]$biasEstimate <- (mean(sdSam) - 1) /
1

    estBiasSD[estBiasSD$knownPop=="Normal" &
              estBiasSD$calcMethod=="Population" &
              estBiasSD$sampSize==intCtr,]$biasEstimate <- (mean(sdPop) - 1) /
1

    estBiasVar[estBiasVar$knownPop=="Normal" &
               estBiasVar$calcMethod=="Sample" &
               estBiasVar$sampSize==intCtr,]$biasEstimate <- (mean(varSam) -
1) / 1

    estBiasVar[estBiasVar$knownPop=="Normal" &
               estBiasVar$calcMethod=="Population" &
               estBiasVar$sampSize==intCtr,]$biasEstimate <- (mean(varPop) -
1) / 1
}
```

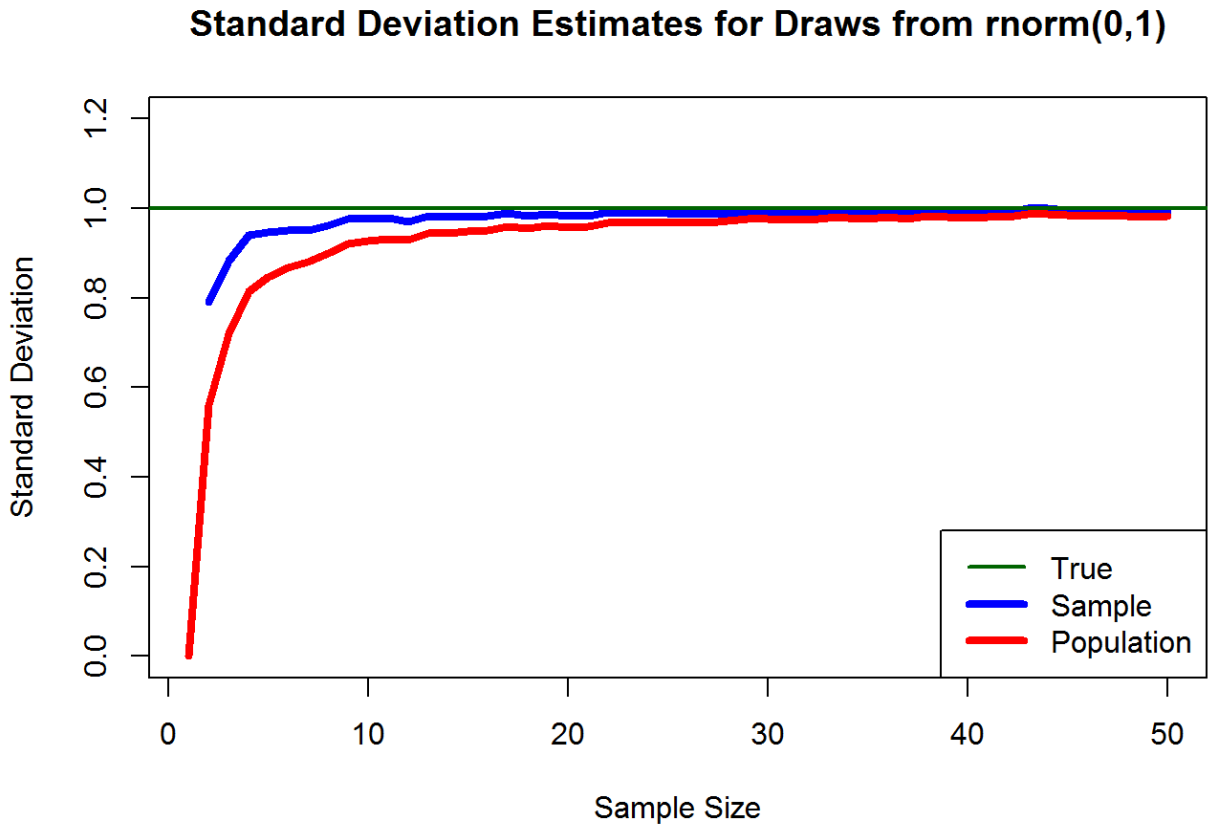
We graph the variance estimates (population, sample, true) of the samples drawn from the rnorm(0,1) distribution:

```
plot(x=1:intCtr,y=myVarSam,type="l",col="blue",ylim=c(0,1.2),lwd=4,
     xlab="Sample Size",ylab="Variance",
     main="Variance Estimates for Draws from rnorm(0,1)"
     )

lines(x=1:intCtr,y=myVarPop,col="red",lwd=4)

abline(h=1,col="dark green",lwd=2)

legend("bottomright",legend=c("True","Sample","Population"),
       col=c("dark green","blue","red"),lwd=c(2,4,4))
```



Further, we graph the standard deviation estimates (population, sample, true) of the samples drawn from the rnorm(0,1) distribution:

```
plot(x=1:intCtr,y=mySDSam,type="l",col="blue",ylim=c(0,1.2),lwd=4,
     xlab="Sample Size",ylab="Standard Deviation",
     main="Standard Deviation Estimates for Draws from rnorm(0,1)"
     )

lines(x=1:intCtr,y=mySDPop,col="red",lwd=4)

abline(h=1,col="dark green",lwd=2)

legend("bottomright",legend=c("True","Sample","Population"),
       col=c("dark green","blue","red"),lwd=c(2,4,4))
```



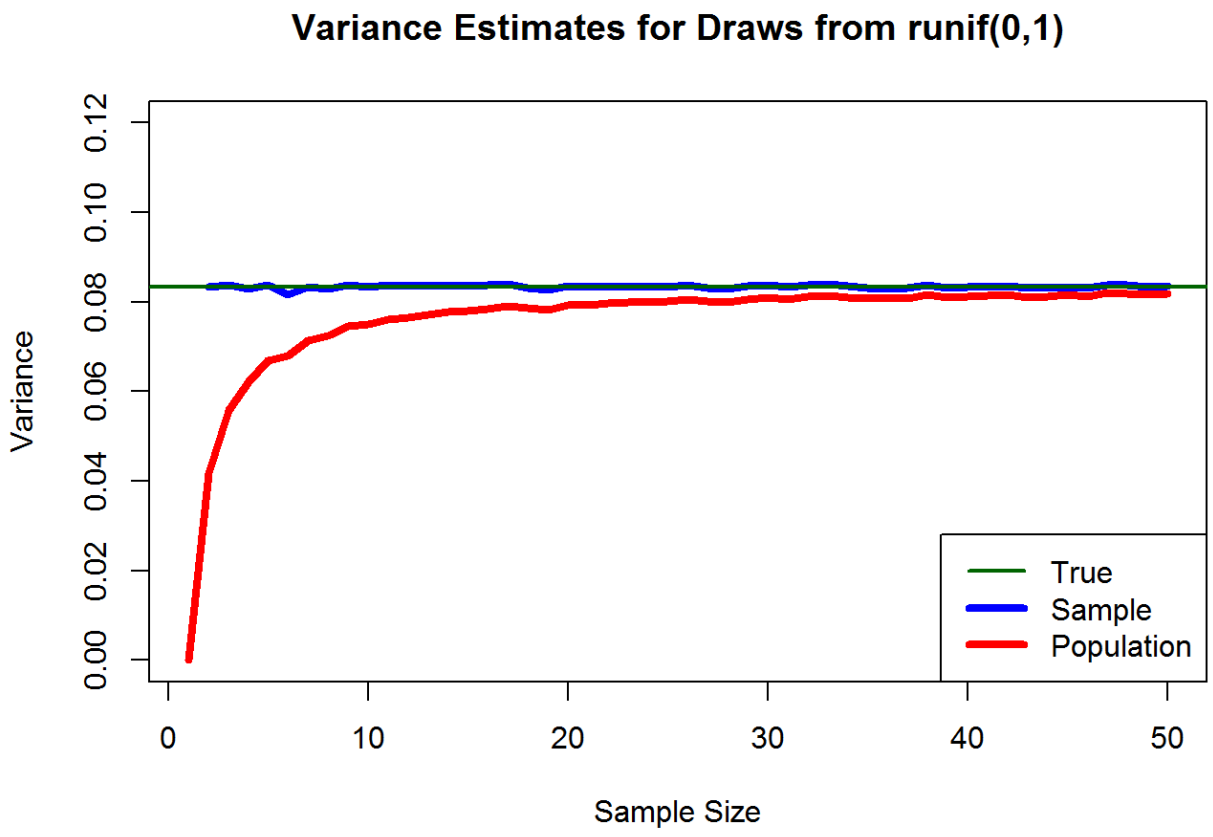
Results from the "uniform" runif(0,1) distribution

We continue with the runif(0,1) distribution and pull samples of size 1:myMax, each taken myDraws number of times. The standard deviations and variances are calculated on a population and sample basis and then averaged by sample size for later reporting.

```
for (intCtr in 1:myMax) {

    for (intCtr2 in 1:myDraws) {
        a <- runif(intCtr,min=0,max=1)
        sdSam[intCtr2] <- sd(a)
        sdPop[intCtr2] <- sd(c(a,mean(a))) ## numerator the same, denom + 1
        varSam[intCtr2] <- var(a)
        varPop[intCtr2] <- var(c(a,mean(a))) ## same as above
    }

    mySDSam[intCtr] <- mean(sdSam)
    mySDPop[intCtr] <- mean(sdPop)
    myVarSam[intCtr] <- mean(varSam)
    myVarPop[intCtr] <- mean(varPop)

    estBiasSD[estBiasSD$knownPop=="Uniform" &
              estBiasSD$calcMethod=="Sample" &
              estBiasSD$sampSize==intCtr,]$biasEstimate <- (mean(sdSam) - 0.288
675) / 0.288675

    estBiasSD[estBiasSD$knownPop=="Uniform" &
              estBiasSD$calcMethod=="Population" &
              estBiasSD$sampSize==intCtr,]$biasEstimate <- (mean(sdPop) - 0.288
675) / 0.288675

    estBiasVar[estBiasVar$knownPop=="Uniform" &
               estBiasVar$calcMethod=="Sample" &
               estBiasVar$sampSize==intCtr,]$biasEstimate <- (mean(varSam) - 0.
083333) / 0.083333

    estBiasVar[estBiasVar$knownPop=="Uniform" &
               estBiasVar$calcMethod=="Population" &
               estBiasVar$sampSize==intCtr,]$biasEstimate <- (mean(varPop) - 0.
083333) / 0.083333
}
```

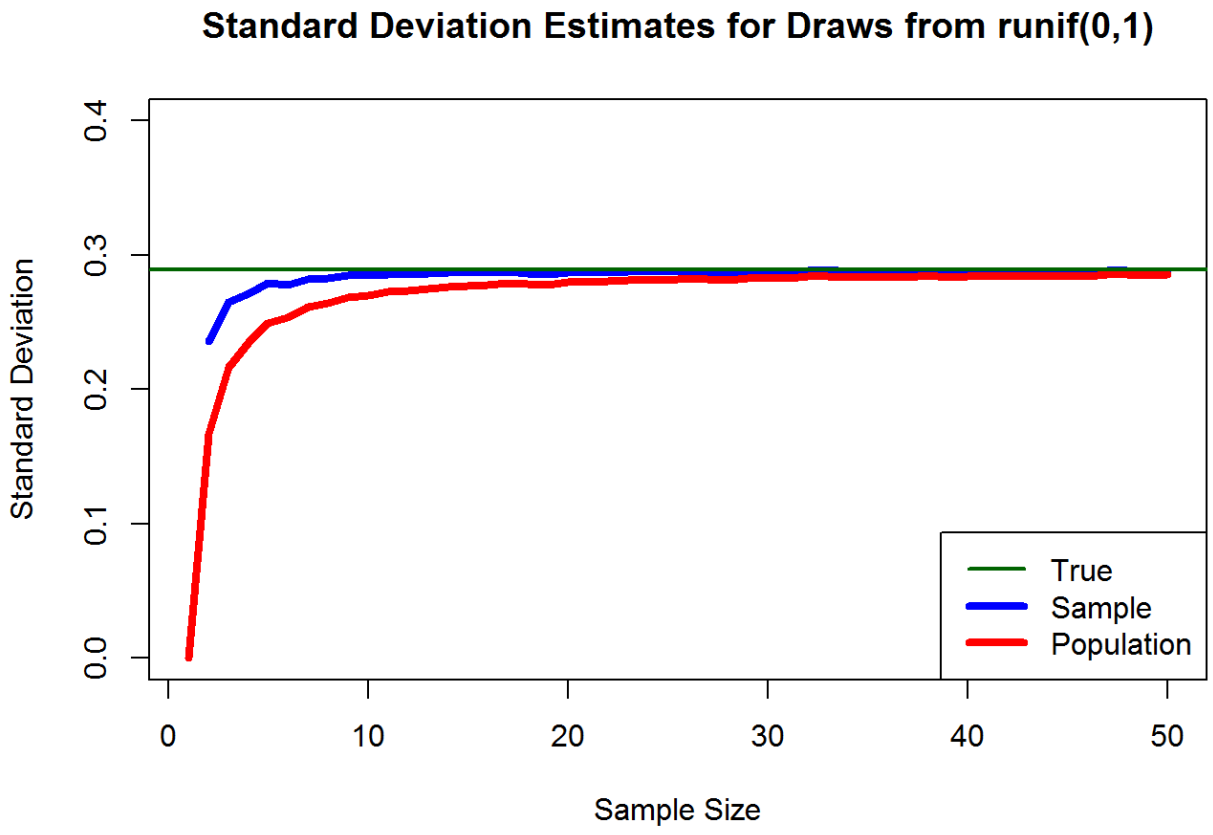
Next, we graph the variance estimates (population, sample, true) of the samples drawn from runif(0,1):

```
plot(x=1:intCtr,y=myVarSam,type="l",col="blue",ylim=c(0,0.12),lwd=4,
     xlab="Sample Size",ylab="Variance",
     main="Variance Estimates for Draws from runif(0,1)"
     )

lines(x=1:intCtr,y=myVarPop,col="red",lwd=4)

abline(h=.0833,col="dark green",lwd=2)

legend("bottomright",legend=c("True","Sample","Population"),
       col=c("dark green","blue","red"),lwd=c(2,4,4))
```



Further, we graph the standard deviation estimates (population, sample, true) of the samples drawn from runif(0,1):

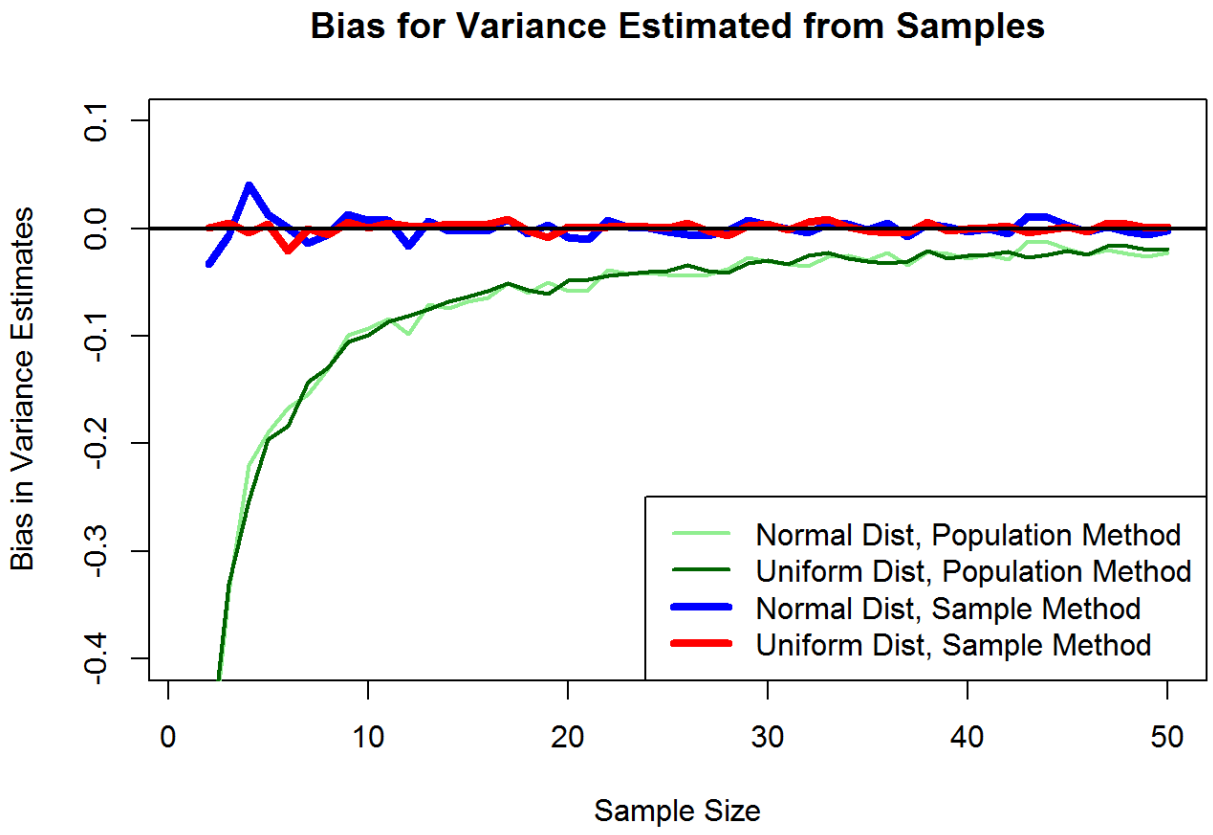
```
plot(x=1:intCtr,y=mySDSam,type="l",col="blue",ylim=c(0,0.4),lwd=4,
     xlab="Sample Size",ylab="Standard Deviation",
     main="Standard Deviation Estimates for Draws from runif(0,1)"
     )

lines(x=1:intCtr,y=mySDPop,col="red",lwd=4)

abline(h=.289,col="dark green",lwd=2)

legend("bottomright",legend=c("True","Sample","Population"),
       col=c("dark green","blue","red"),lwd=c(2,4,4))
```



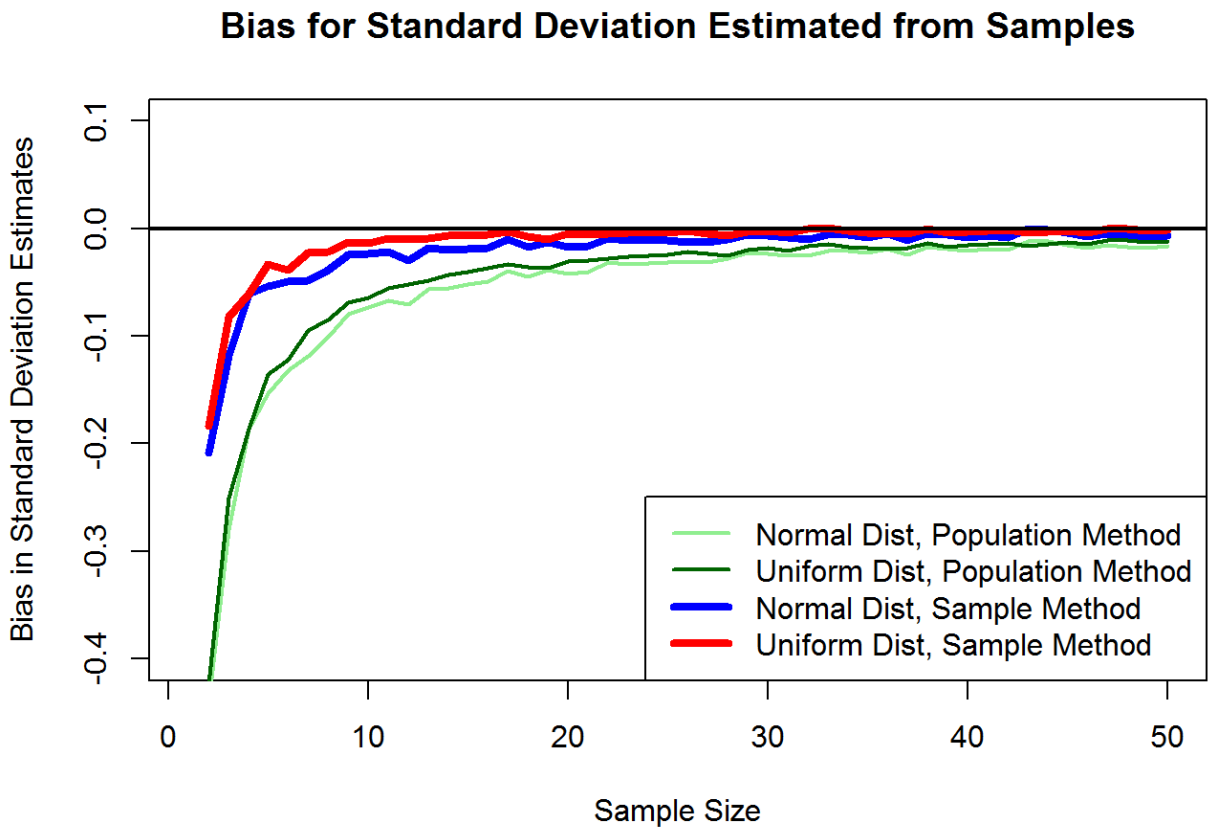
Overall bias by methodology and sample size

Lastly, we graph bias by metric, methodology, population distribution, and sample size. Bias was previously calculated as (Metric - True) / True in all cases.

```
## Variance Bias Estimates
plot(x=1:intCtr,
     y=estBiasVar[estBiasVar$knownPop=="Normal" &
                  estBiasVar$calcMethod=="Sample",]$biasEstimate,
     type="l",col="blue",ylim=c(-0.4,0.1),lwd=4,
     xlab="Sample Size",ylab="Bias in Variance Estimates",
     main="Bias for Variance Estimated from Samples"
     )

lines(x=1:intCtr,
      y=estBiasVar[estBiasVar$knownPop=="Uniform" &
                   estBiasVar$calcMethod=="Sample",]$biasEstimate,
      col="red",lwd=4
     )

lines(x=1:intCtr,
      y=estBiasVar[estBiasVar$knownPop=="Normal" &
                   estBiasVar$calcMethod=="Population",]$biasEstimate,
      col="light green",lwd=2
     )

lines(x=1:intCtr,
      y=estBiasVar[estBiasVar$knownPop=="Uniform" &
                   estBiasVar$calcMethod=="Population",]$biasEstimate,
      col="dark green",lwd=2
     )

abline(h=0,col="black",lwd=2)

legend("bottomright",
       legend=c("Normal Dist, Population Method",
                "Uniform Dist, Population Method",
                "Normal Dist, Sample Method",
                "Uniform Dist, Sample Method"
                ),
       col=c("light green","dark green","blue","red"),
       lwd=c(2,2,4,4)
       )
```


Bias for Variance Estimated from Samples
Bias in Variance Estimates
0.1
0.0
-0.1
-0.2
-0.3
-0.4
0
10
20
30
40
50
Sample Size
Normal Dist, Population Method
Uniform Dist, Population Method
Normal Dist, Sample Method
Uniform Dist, Sample Method

```
## Standard Deviation Bias Estimates
plot(x=1:intCtr,
     y=estBiasSD[estBiasSD$knownPop=="Normal" &
                 estBiasSD$calcMethod=="Sample",]$biasEstimate,
     type="l",col="blue",ylim=c(-0.4,0.1),lwd=4,
     xlab="Sample Size",ylab="Bias in Standard Deviation Estimates",
     main="Bias for Standard Deviation Estimated from Samples"
     )

lines(x=1:intCtr,
      y=estBiasSD[estBiasSD$knownPop=="Uniform" &
                  estBiasSD$calcMethod=="Sample",]$biasEstimate,
      col="red",lwd=4
     )

lines(x=1:intCtr,
      y=estBiasSD[estBiasSD$knownPop=="Normal" &
                  estBiasSD$calcMethod=="Population",]$biasEstimate,
      col="light green",lwd=2
     )

lines(x=1:intCtr,
      y=estBiasSD[estBiasSD$knownPop=="Uniform" &
                  estBiasSD$calcMethod=="Population",]$biasEstimate,
      col="dark green",lwd=2
     )

abline(h=0,col="black",lwd=2)

legend("bottomright",
       legend=c("Normal Dist, Population Method",
                "Uniform Dist, Population Method",
                "Normal Dist, Sample Method",
                "Uniform Dist, Sample Method"
                ),
       col=c("light green","dark green","blue","red"),
       lwd=c(2,2,4,4)
       )
```



Conclusions

As described in the literature, the sample variance calculation applied to a sample creates an unbiased estimate of the population variance, while the population variance calculation applied to a sample creates a biased (too low) estimate of the population variance. The bias is worst with small-n.

Both the sample standard deviation calculation and the population standard deviation calculation applied to a sample create a biased (too low) estimate of the population standard deviation. This is a known consequence of the concavity of sqrt(x) and has the greatest impact for small-n. While there are technical corrections that can be applied, typically the sample standard deviation is considered to be a sufficiently good estimate for the population standard deviation.