

Statistics 406 Problem Set 5

Due in lab, Tuesday October 24

1. Is the sample standard deviation biased? Do a simulation study to find out. Use sample sizes $n = 5, 10, 20$ in your study, and consider both standard normal and standard exponential distributions for the underlying data. Explain your findings in terms of the “Functions of random variables” section of the “Behavior of the sample mean” portion of the course notes.

Solution: We know that $\hat{\sigma}^2$ is an unbiased estimate of σ^2 . Since $f(x) = \sqrt{x}$ is concave, for any positive random variable, $E\sqrt{X} \leq \sqrt{EX}$. Therefore, letting $X = \hat{\sigma}^2$, we get that $E\hat{\sigma} \leq \sigma$. This is consistent with the results of the following simulation.

```
nrep <- 1e4
V <- NULL

## Loop over the sample sizes.
for (n in c(5,10,20))
{
  ## Generate nrep normal data sets of size n.
  X <- array(rnorm(n*nrep), c(n,nrep))

  ## Calculate the sample standard deviation for each simulated normal
  ## data set.
  SD <- apply(X, 2, sd)

  ## Get the average of the sample standard deviations. This should
  ## approximate E sigma_hat.
  sd1 <- mean(SD)

  ## Generate nrep exponential data sets of size n.
  X <- array(rexp(n*nrep), c(n,nrep))

  ## Calculate the sample standard deviation for each simulated normal
  ## data set.
  SD <- apply(X, 2, sd)

  ## Get the average of the sample standard deviations. This should
  ## approximate E sigma_hat.
  sd2 <- mean(SD)
```

```

## Save the results for the current sample size.
V <- rbind(V, c(sd1, sd2))
}

```

For the simulation, I get the following (column 1 is for the normal data, column 2 is for the exponential data; rows 1, 2, and 3 are for sample sizes 5, 10, and 20 respectively).

```

      [,1]      [,2]
[1,] 0.9400669 0.8710466
[2,] 0.9709772 0.9323119
[3,] 0.9887004 0.9614583

```

showing that: (i) the bias is negative, (ii) it is greater for the exponential distribution, and (iii) it diminishes as the sample size grows.

- The “bootstrap t-method” for confidence intervals for EX operates somewhat differently from the percentile method given in the course notes. To construct a t-method CI for EX , first calculate the mean \bar{X} and standard deviation $\hat{\sigma}$ for the actual data. Then generate $K = 1000$ non-parametric bootstrap data sets. For each bootstrap data set, calculate the sample mean \bar{X}_k and sample standard deviation $\hat{\sigma}_k$. Then let F denote the empirical 95th percentile of

$$\frac{|\bar{X}_k - \bar{X}|}{\hat{\sigma}_k}.$$

The 95% CI for EX is then taken to be

$$\bar{X} \pm F\hat{\sigma}.$$

- Compare coverage probabilities for the t-method CI and the percentile method CI.
- Compare widths for the bootstrap t-method CI and the classical parametric approach.

Solution:

```

## Sample size.
n <- 20

## Number of replications.
nrep <- 1000

```

```

## Space for the width data.
W <- array(0, c(nrep,3))

## Number of intervals that cover.
nc <- c(0,0,0)

## Simulation replications.
for (rep in 1:nrep)
{
  ## Actual data.
  X <- rnorm(n)

  ## Estimated mean and SD of the data. These will be the population
  ## mean and SD of the bootstrap data.
  mu1 <- mean(X)
  sd1 <- sd(X)

  ## Generate bootstrap samples.
  ii <- ceiling(n*runif(1000*n))
  B <- array(X[ii], c(n,1000))

  ## Use the bootstrap sets to define the CI.
  M <- apply(B, 2, mean)
  SD <- apply(B, 2, sd)
  F <- (M-mu1) / SD
  FA <- sort(abs(F))
  c <- FA[950]

  ## The CI.
  c1 <- mu1 - c*sd1
  c2 <- mu1 + c*sd1

  ## Check for coverage of the bootstrap CI.
  if ( (c1 < 0) * (c2 > 0) ) { nc[1] <- nc[1]+1 }

  ## Save the bootstrap width.
  W[rep,1] <- c2 - c1

  ## The bootstrap percentile method CI.
  MS <- sort(M)
  c1 <- MS[25]
}

```

```

c2 <- MS[975]

## Check for coverage of the bootstrap percentile CI.
if ( (c1 < 0) * (c2 > 0) ) { nc[2] <- nc[2]+1 }

## Save the bootstrap percentile method width.
W[rep,2] <- c2 - c1

# The classical CI.
c1 <- mu1 - qt(0.975, n-1)*sd1/sqrt(n)
c2 <- mu1 + qt(0.975, n-1)*sd1/sqrt(n)

## Check for coverage of the classical CI.
if ( (c1 < 0) * (c2 > 0) ) { nc[3] <- nc[3]+1 }

## Save the classical method width.
W[rep,3] <- c2 - c1
}

## The estimated coverage probability.
cp <- nc/nrep

## The average width.
AW <- apply(W, 2, mean)

```

I get this:

```

> AW
[1] 0.9422692 0.8461670 0.9282856
> cp
[1] 0.958 0.936 0.957

```

The classical and bootstrap-t methods have good coverage probabilities, while the percentile t-method has somewhat low coverage. The percentile t-method has the narrowest intervals, but this is due to the lower coverage. The widths for the bootstrap-t and classical intervals are comparable, and the classical method comes out ahead. The classical approach performs best in this case since the data are normal.

3. How accurate is the parametric bootstrap estimate of $\text{var } \hat{Q}$, where \hat{Q} is a sample quantile? Suppose we are interested in the 50th percentile (the median) and the 90th percentile of

an unknown distribution. We are interested in data sets of size 100 from standard normal and t_8 (t distribution with 8 degrees of freedom) distributions.

Note that

```
Q <- apply(X, 2, quantile, p)
```

will calculate p^{th} quantiles for every column of X.

- (a) Use direct simulation (not bootstrapping) to estimate the variance of sample estimates of the 50th and 90th percentiles, for both standard normal and t_8 (t distribution with 8 degrees of freedom) populations. Which quantile is harder to estimate, and which distribution provides less informative data?

Solution:

```
n <- 100
V <- NULL

## Loop over the two distributions.
for (k in c(1,2))
{
  ## Generate the data.
  if (k==1) { X <- array(rnorm(n*1e4), c(n,1e4)) }
  else      { X <- array(rt(n*1e4, 8), c(n,1e4)) }

  # Calculate variance estimates for both quantiles.
  for (p in c(0.5,0.9))
  {
    Q <- apply(X, 2, quantile, p)
    V <- c(V, var(Q))
  }
}
```

I get this:

```
> V
[1] 0.01546619 0.02816679 0.01625445 0.04088814
```

For both distributions, it is harder to estimate the 90th percentile than the median. For both percentiles, it is harder to estimate with t_8 data compared to Gaussian data (but this holds to a much greater extent for the 90th percentile).

- (b) Use the parametric bootstrap to estimate $\text{var } \hat{Q}$, based on a normal working model (i.e. generate data from a member of the normal distribution family when performing the bootstrap). For both quantiles and both data populations, generate 200 replicated data sets. For each data set, use 200 bootstrap samples to estimate the variance of \hat{Q} . Then make a histogram of the variance estimates for the 200 replicated data sets. Your result should be four histograms. Compare your results to part (a).

Solution:

```
## Number of simulation replications.
nrep <- 200

## Sample size.
n <- 100

## The quantiles we are looking at.
P <- c(0.5, 0.9)

## Loop over the quantiles.
for (pi in c(1,2))
{
  ## Loop over the distributions.
  for (k in c(1,2))
  {
    ## Simulation replications.
    for (rep in 1:nrep)
    {
      ## Generate an 'actual' data set.
      if (k == 1) { X <- rnorm(n) }
      else      { X <- rt(n, 8) }

      ## Get quantile estimates from parameteric bootstrap data sets derived
      ## from X.
      mu <- mean(X)
      sd <- sd(X)
      Z <- mu + sd*rnorm(n*200)
      Z <- array(Z, c(n, 200))
      Q <- apply(Z, 2, quantile, P[pi])

      ## The variability of the quantile estimates.
      V[rep] <- var(Q)
    }
  }
}
```

```

print(mean(V))

## Generate a histogram of the quantile estimate variances
## (you don't need to know about sprintf for the course).
pdf(sprintf('ps05-hist-%d-%d.pdf', pi, k))
hist(V)
dev.off()
}
}

```

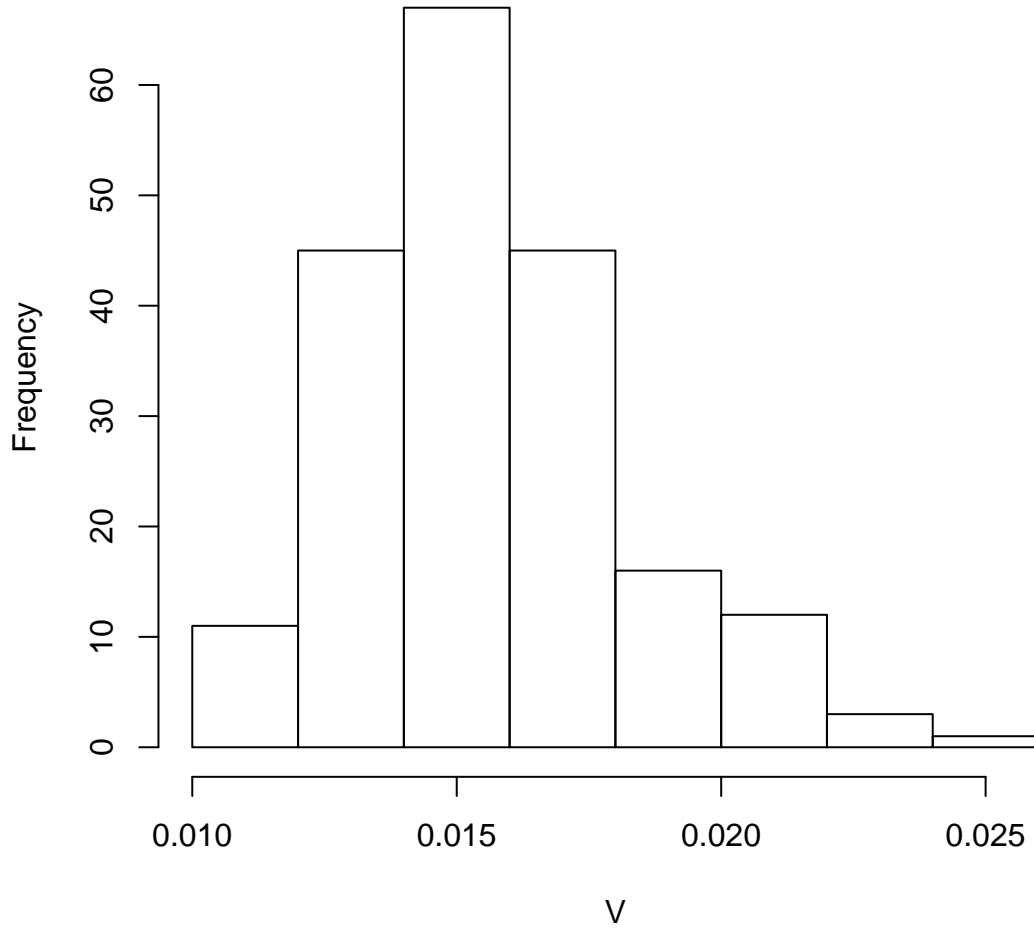
I get the following variances (the “actual variance” comes from part (a)).

		Bootstrap variance	Actual variance
Median	Normal	0.016	0.015
Median	t_8	0.021	0.016
90 th percentile	Normal	0.029	0.028
90 th percentile	t_8	0.038	0.041

Here are the histograms:

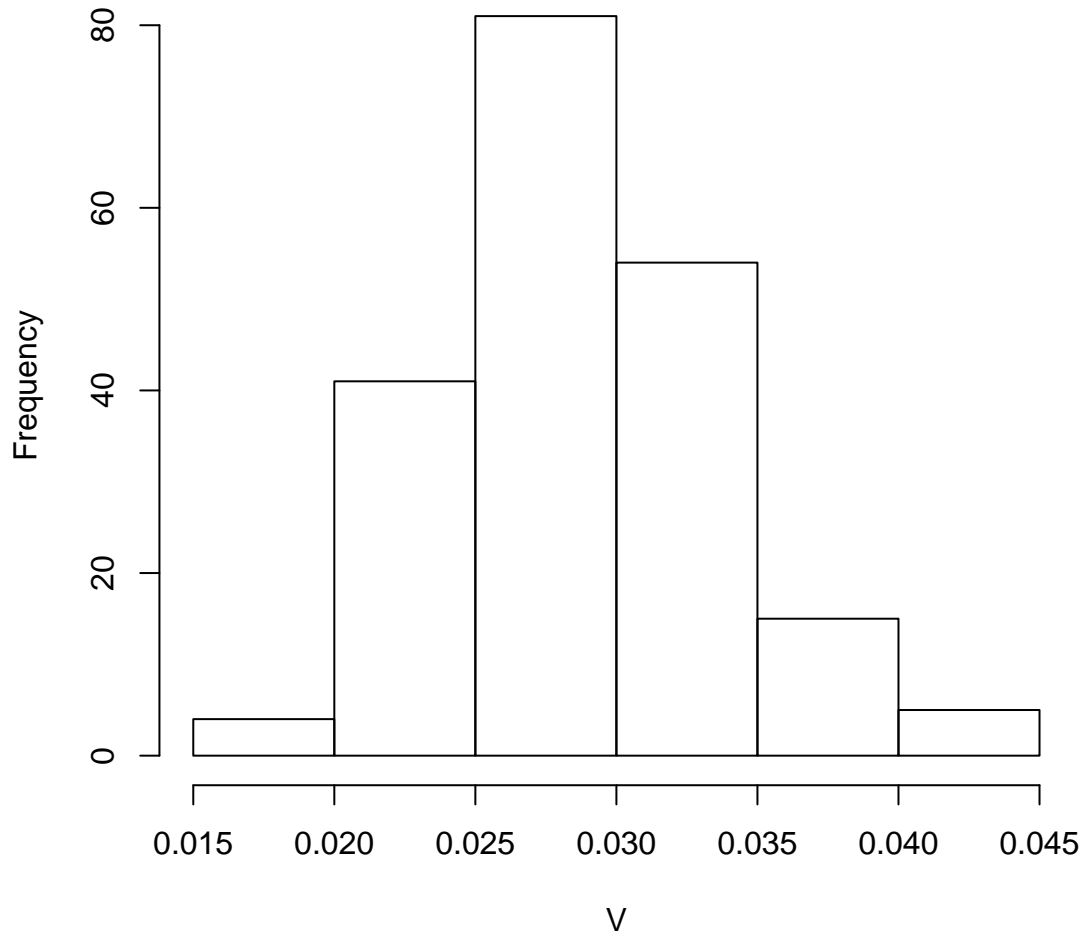
Median, normal data

Histogram of V



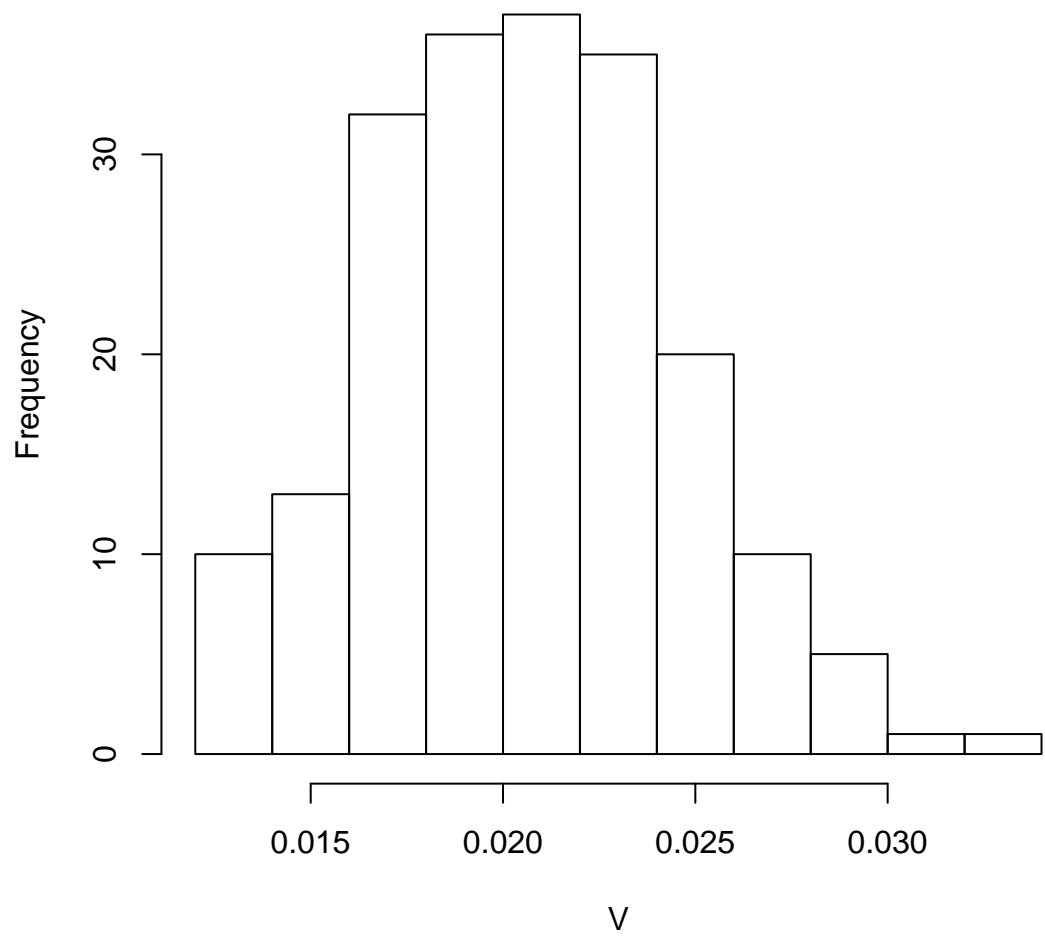
90th percentile, normal data

Histogram of V



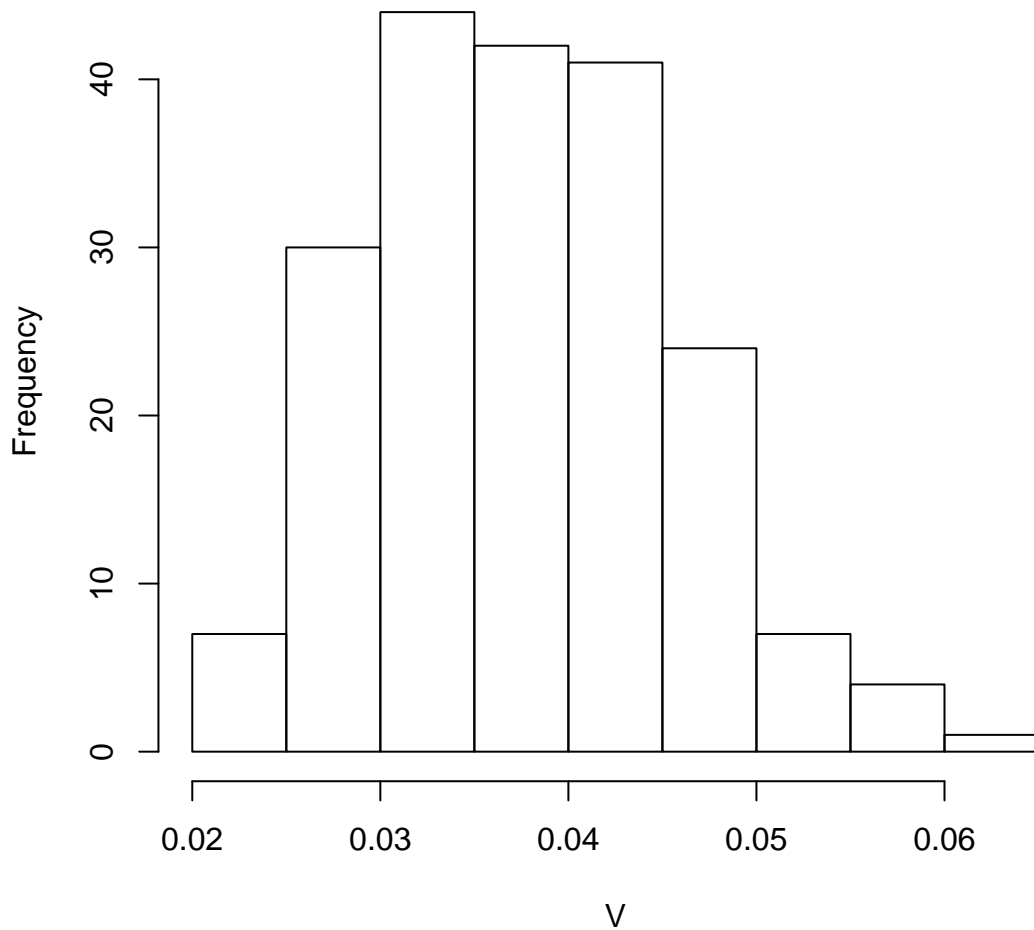
Median, t_8 data

Histogram of V



90th percentile, t_8 data

Histogram of V



4. Suppose we observe values that are uniformly distributed on the interval $(0, K)$, where K is an unknown parameter that needs to be estimated. Use the maximum observed value in a dataset, denoted \hat{K} , to estimate K . Use simulation to estimate $n\text{var}(\hat{K})$ for $n = 10, 100, 1000, 10000$. In what important way does \hat{K} behave differently from \bar{X} ?

Solution:

```
V <- NULL
```

```
for (n in c(10,100,1000,10000))  
{  
  X <- array(runif(n*1000), c(n,1000))  
  M <- apply(X, 2, max)  
  V <- c(V, n*var(M))  
}
```

I get:

```
6.758194e-02 1.040442e-02 9.692937e-04 9.687662e-05
```

Since these numbers decrease, we conclude that the variance of the sample maximum decreases faster than the variance of the sample mean as the sample size increases.