Math 3070 § 1. **Panzer Eg.: Simulation to Estimate** Name: Example
Treibergs **Variance of Estimator** July 5, 2011

In this example we consider unbiased estimators of the maximum of a discrete uniform distibution. To compare them, we simulate them and compare their variances. The better estimator is the one with lesser variance.

Larsen & Marx, *An Introduction to Mathematical Statistics and its Applications, 3th ed.,* Pearson, 2006 discuss this example studied by Leo Goodman in "Serial Number Analysis," *Journal of the American Statistical Association,* 1952. The Allies were interested in estimating the extent of Third Reich production of equipment, such as Mark I tanks. If there were N Mark I Tanks produced by a certain date then each of the existing tanks would bear one of the integers 1 to N. As the war progressed, some of the serial numbers would become known to the Allies, either by direct capture or from records captured when a command post was overrun. The war departments statisticians then had a sample of size M without replacement of "captured" serial numbers $1 \le X_1 < X_2 < \cdots < X_M \le N$. Their job was to estimate the production size N. From postwar records, it turned out that the statisticians did a much better job estimating than the other intelligence analysts.

Let us consider two estimators for N.

$$\hat\theta_1 = X_{\max} + \frac{1}{M-1}\sum_{i=1}^{M-1} (X_{i+1} - X_i - 1)\,,$$
$$\hat\theta_2 = \frac{M+1}{M} X_{\max} - 1.$$

The first is the maximum plus the average gap between observations. The second is an enlarged maximum, like the estimator to estimate the maximum for continuous uniform distribution.

We show that both of these estimators are unbiased. Notice that each combination of N serial numbers chosen M at a time is equally likely. Therefor the probability of any single combination

$$\mathrm{P}(X_1 = x_1 < X_2 = x_2 < \cdots < X_M = x_M) = c = \frac{1}{\binom{N}{M}}.$$

Let us figure out the pmf for the maximum of the sample, $X_{\max} = X_M$. The maximum of the sample ranges between $M \le X_{\max} \le N$. The cdf is

$$F(x) = \mathrm{P}(X_{\max} \le x) = c\binom{x}{M}$$

because if $X_{\max} \le x$ then the sample occurs in $1, \ldots, x$ and the binomial coefficient give the number of such combinations taken M at a time. Therefore, for $M \le x \le N$, the pmf is

$$p_{\max}(x) = \mathrm{P}(X_{\max} = x) = c\left[\binom{x}{M} - \binom{x-1}{M}\right] = c\binom{x-1}{M-1},$$

by the Pascal's triangle equation. We have $\sum_{i=M}^{N} p_{\max}(i) = 1$ since the sum telescopes,

$$\sum_{i=M}^{N} \binom{i-1}{M-1} = \binom{N}{M}. \tag{1}$$

Using (1), the expected value is thus

$$\begin{aligned}
\mathrm{E}(X_{\max}) &= \sum_{i=M}^{N} i\, p_{\max}(i) \\
&= c \sum_{i=M}^{N} i \binom{i-1}{M-1} \\
&= c \sum_{i=M}^{N} \frac{i(i-1)!}{(M-1)!(i-M)!} \\
&= cM \sum_{i=M}^{N} \frac{i!}{M!(i-M)!} \\
&= cM \sum_{i=M}^{N} \binom{i}{M} \\
&= cM \binom{N+1}{M+1} \\
&= \frac{M!(N-M)!}{N!} \cdot \frac{M(N+1)!}{(M+1)!(N-M)!} \\
&= \frac{M(N+1)}{M+1}.
\end{aligned}$$

It follows that the second estimator is unbiased

$$\mathrm{E}(\hat{\theta}_2) = \mathrm{E}\left(\frac{M+1}{M} X_{\max} - 1\right) = N.$$

The first estimator may be simplified because the sum telescopes,

$$\begin{aligned}
\hat{\theta}_1 &= X_{\max} + \frac{1}{M-1} \sum_{i=1}^{M-1} (X_{i+1} - X_i - 1) \\
&= X_{\max} + \frac{1}{M-1} (X_N - X_1) - 1 \\
&= \frac{M}{M-1} X_{\max} - \frac{1}{M-1} X_{\max} - 1
\end{aligned}$$

Note that the transformation $x \mapsto N + 1 - x$ reverses the order of $1, 2, \ldots, N$ and swaps the maximum and minimum of a combination. Hence, the minimum of the sample ranges from $1 \le X_{\min} \le N - M + 1$ and the pmf becomes

$$p_{\min}(x) = p_{\max}(N + 1 - x) = c \binom{N-x}{M-1}.$$

The expectation also reflects

$$\mathrm{E}(X_{\min}) = N + 1 - \frac{M(N+1)}{M+1} = \frac{N+1}{M+1}.$$

It follows that the first estimator is unbiased too

$$\mathrm{E}(\hat{\theta}_1) = \mathrm{E}\left(\frac{M}{M-1} X_{\max} - \frac{1}{M-1} X_{\max} - 1\right) = \frac{M}{M-1} \cdot \frac{M(N+1)}{M+1} - \frac{1}{M-1} \cdot \frac{N+1}{M+1} - 1 = N.$$

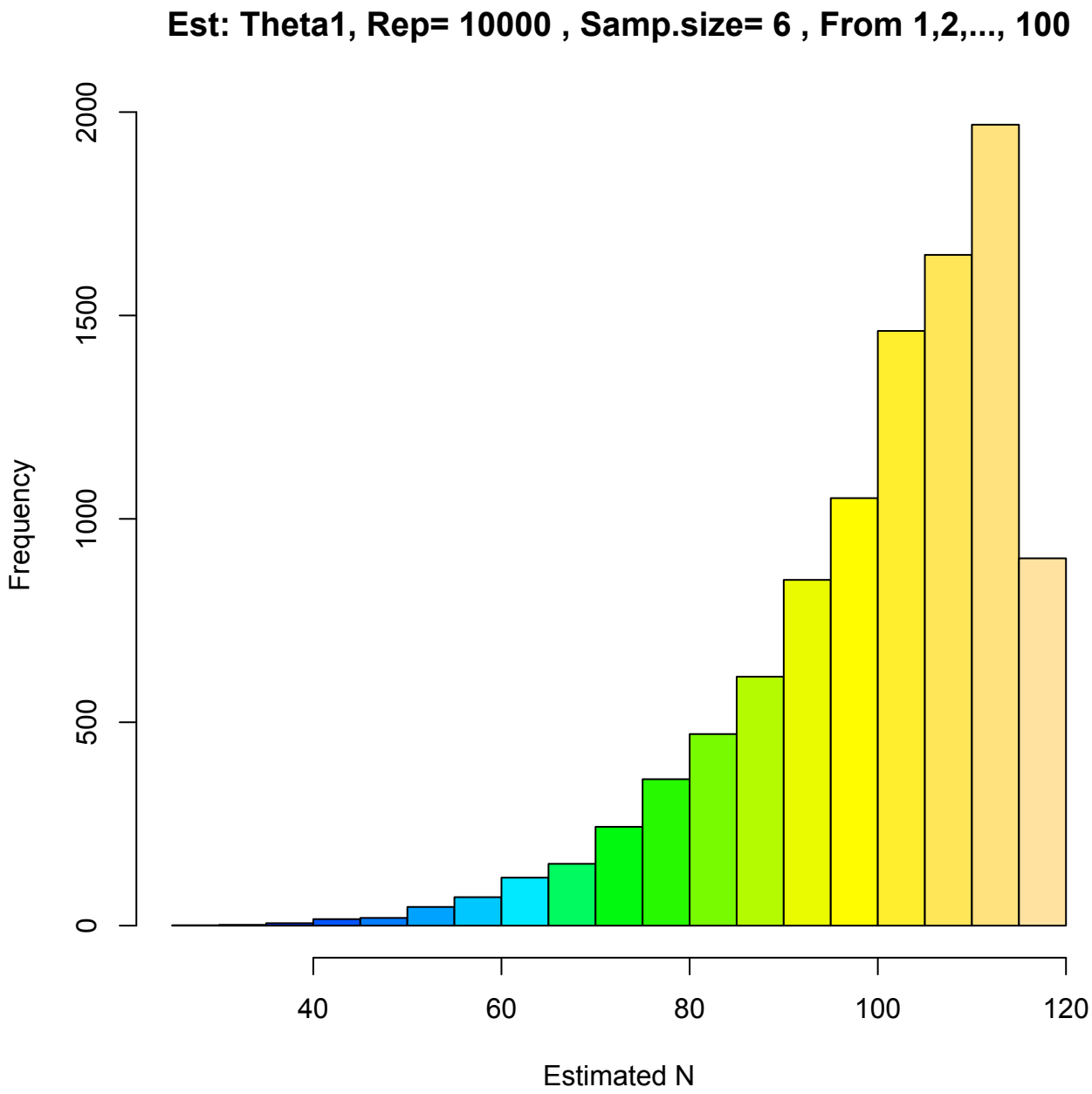
The standard error can be computed along similar lines with a longer computation. We simulate both statistics to see the difference in variance. Our simulation shows that $\hat{\theta}_2$ does slightly better. One can also prove that the second is slightly better since

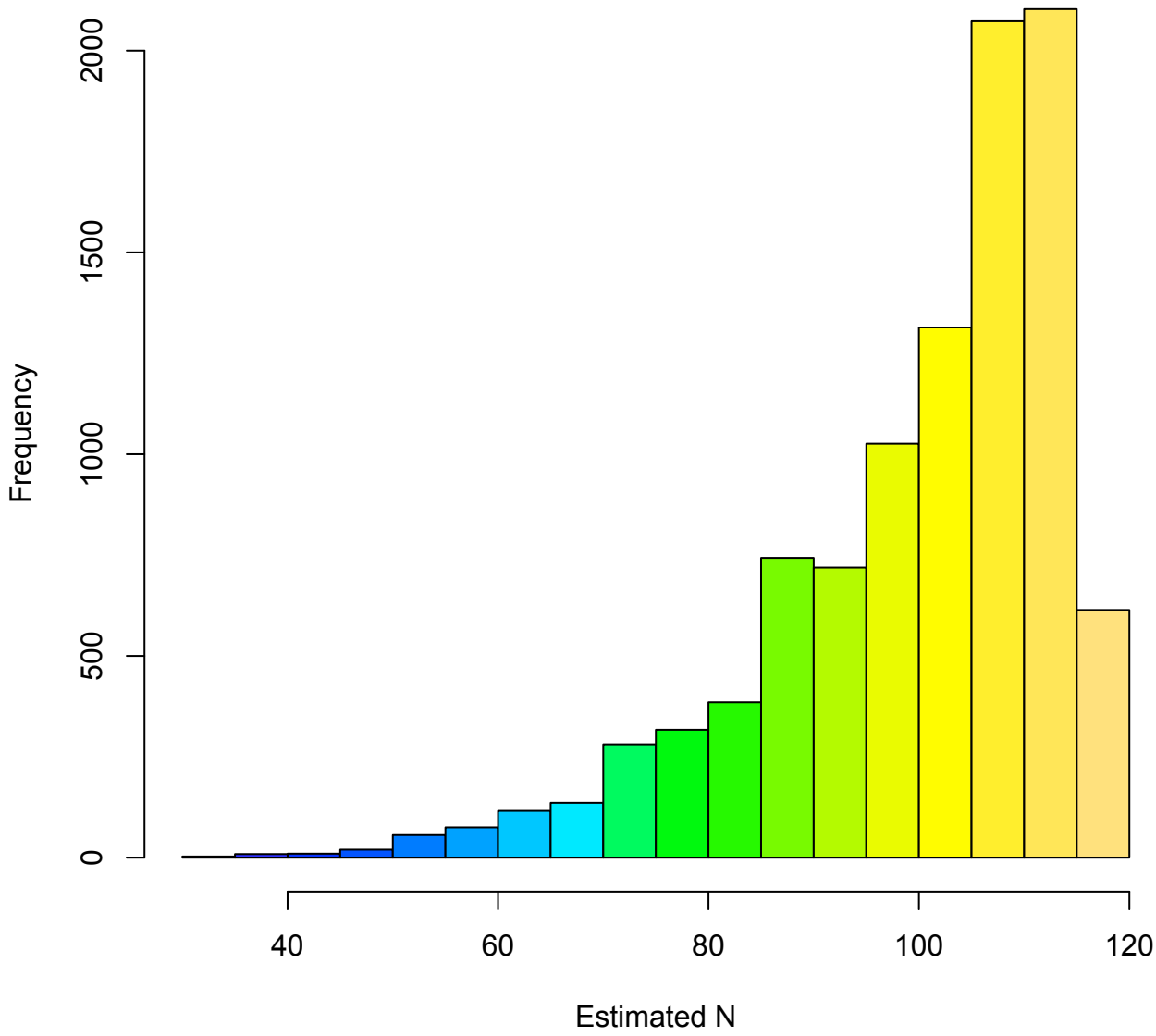
$$\frac{\mathrm{V}(\hat{\theta}_2)}{\mathrm{V}(\hat{\theta}_1)} = 1 - \frac{1}{n^2}.$$

R Session:

```
R version 2.10.1 (2009-12-14)
Copyright (C) 2009 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.31 (5538) powerpc-apple-darwin8.11.1]

[Workspace restored from /Users/andrejstreibergs/.RData]

> ################ SIMULATE THE FIRST ESTIMATOR OF  N  ##########################
> B <- 100
> n <- 6
> # rnd sample of size n from uniform discrete
> # 1,...,N
> # how to estimate N?
> # First N1 = max(samp)+ 1/(n-1) sum_{i>j}(yi-yj-1)
> # Second N2 = (n+1)/n max(samp)-1
> th1 <- function(x){(n*max(x) - min(x))/(n-1) - 1}
> th2 <- function(x){(n+1)*max(x)/n - 1}
>
> M <- 100
>
> v <- replicate(B,th1(sample(1:M,n)))
> hist(v)
> summary(v)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  26.40   92.60  103.60   99.94  111.00  118.80
> sd(v)
[1] 14.41988
>
```

```
> v <- replicate(B,th1(sample(1:M,n)))
> hist(v)
> summary(v)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   29.6    92.4   103.6   100.1   111.0   118.8
> sd(v)
[1] 14.22717

>
>
> hist(v, col = topo.colors(20), xlab = "Estimated N",
+ main = paste("Est: Theta1, Rep=", B, ", Samp.size=", n,
+ ", From 1,2,...,",M))
> # M3074MaxEst1.pdf
>
> ################ SIMULATE THE SECOND ESTIMATOR OF  N  ########################
> v <- replicate(B,th2(sample(1:M,n)))
> summary(v)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.33   92.33  104.00  100.00  111.00  115.70
> sd(v)
[1] 14.06135
>
> v <- replicate(B,th2(sample(1:M,n)))
> summary(v)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  24.67   93.50  104.00  100.10  111.00  115.70
> sd(v)
[1] 14.05302
>
> v <- replicate(B,th2(sample(1:M,n)))
> summary(v)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  31.67   92.33  104.00   99.96  111.00  115.70
> sd(v)
[1] 13.90838
>
> v <- replicate(B,th2(sample(1:M,n)))
> summary(v)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   30.5    93.5   104.0   100.1   111.0   115.7
> sd(v)
[1] 14.04389

> # The second estimator has slightly smaller variance, hence is better.
> hist(v, col = topo.colors(20), xlab="Estimated N",
+ main = paste("Est: Theta2, Rep=", B, ", Samp.size=", n,
+ ", From 1,2,...,", M))
> # M3074MaxEst2.pdf
```


Est: Theta1, Rep= 10000 , Samp.size= 6 , From 1,2,..., 100
Frequency
0
500
1000
1500
2000
40
60
80
100
120
Estimated N


Est: Theta2, Rep= 10000 , Samp.size= 6 , From 1,2,..., 100
Frequency
0
500
1000
1500
2000
40
60
80
100
120
Estimated N