STAT 421

Finite Population Sampling with Application to the Hypergeometric Distribution

Fritz Scholz

Suppose we have a population of $N = a + b$ balls, a white ones and b black ones. We randomly grab n of these N balls (without replacement) and denote by Y the number of white balls in the random grab. Y is then a hypergeometric random variable and its cumulative distribution function is given in **R** by

$$P(Y \le y) = \texttt{phyper}(\texttt{y}, \texttt{a}, \texttt{b}, \texttt{n}) \; .$$

We can view Y also as the sum of n numbers randomly drawn(without replacement) from a population consisting of the N numbers

$$\overbrace{1, \ldots, 1}^{a\ 1's}, \overbrace{0, \ldots, 0}^{b\ 0's} \; .$$

$\bar{Y} = Y/n$ would then be the proportion of $1's$ in the sample or it would be the average of n sampled values drawn without replacement from $(Z_1, \ldots, Z_N) = (1, \ldots, 1, 0, \ldots, 0)$. $\bar{Y}$ would also be the proportion of white balls in the random grab on n.

From slide 44 in `DoeFlux.pdf` we have mean and variance of $\bar{Y}$ given as

$$\begin{aligned} E(\bar{Y}) &= \bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i = \frac{1}{N}\left(\overbrace{1 + \ldots + 1}^{a \text{ terms}} + \overbrace{0 + \ldots + 0}^{b \text{ terms}}\right) = \frac{a}{N} \\ \text{and} \qquad \text{var}(\bar{Y}) &= \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{S^2}{n}\frac{N-n}{N} \end{aligned}$$

with

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \bar{Z})^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Z_i^2 - 2Z_i\bar{Z} + \bar{Z}^2\right) = \frac{1}{N-1}\left(\sum_{i=1}^{N} Z_i^2 - N\bar{Z}^2\right) \\ &= \frac{1}{N-1}\left(\sum_{i=1}^{N} Z_i - N\bar{Z}^2\right) = \frac{1}{N-1}\left(a - N(a/N)^2\right) = \frac{N}{N-1}\frac{a}{N}\left(1 - \frac{a}{N}\right) \end{aligned}$$

note $Z_i^2 = Z_i$ for $Z_i = 0, 1$. Thus

$$\text{var}(\bar{Y}) = \frac{1}{n}\frac{N}{N-1}\frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N} = \frac{1}{n}\frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N-1} \; .$$

Compare this with the variance of $\bar{Y}$ when we sample n numbers with replacement from $Z_1, \ldots, Z_N$, in which case $\bar{Y} = Y/n$ is a binomial proportion and Y is a binomial random variable with paramaters n and $p = a/N$. We then have

$$E(\bar{Y}) = p = \frac{a}{N} \quad \text{and} \quad \text{var}(\bar{Y}) = \frac{1}{n}p(1-p) = \frac{1}{n}\frac{a}{N}\left(1 - \frac{a}{N}\right) .$$

We see that in the hypergeometric sampling case the variance of $\bar{Y}$ has the additional finite population correction factor $(N-n)/(N-1)$, which degenerates to 0 when $n = N$. In that case we sample the full population and $\bar{Y} = \bar{Z}$ is no longer random, thus $\text{var}(\bar{Y}) = \text{var}(\bar{Z}) = 0$.

Returning to sampling without replacement (hypergeometric case) we get the mean and variance of the hypergeometric random variable $Y = n\bar{Y}$ as

$$\begin{aligned} E(Y) &= E(n\bar{Y}) = nE(\bar{Y}) = n\frac{a}{N} \\ \text{var}(Y) &= \text{var}(n\bar{Y}) = n^2\text{var}(\bar{Y}) = n^2\frac{1}{n}\frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N-1} = n\frac{a}{N}\left(1 - \frac{a}{N}\right)\frac{N-n}{N-1}. \end{aligned}$$

Again note the additional finite population correction factor $(N-n)/(N-1)$ multiplying the variance $np(1-p)$ for the binomial case.

Examining the CLT in Sampling from Finite Populations

In class I improvised looking at the distribution of a sum Y of n numbers, randomly drawn with replacement from $1, 2, \ldots, N$. On slide 44 we claimed that the CLT holds for averages $\bar{Y} = Y/n$ of such numbers and thus also for the sums Y. However, I also cautioned about the size of n, i.e., n should not be too small or be too close to N.

The following function allows you to examine this CLT claim.

```
finite.population.CLT=function (N=20,n=7)
{
out=combn(1:N,n,FUN=sum) # random sum of n taken from 1:N
m=sum(1:n) #smallest possible sum
M=sum(N:(N-n+1)) #largest possible sum
mu=n*mean(1:N)
sig=sqrt(n*var(1:N)*(1-n/N))
```

```
hist.out=hist(out,breaks=seq(m-.5,M+.5,1),plot=F)
# this gives us the bar heights for proper plotting the
# second time, after we set ylim correctly.
z=seq(mu-4*sig,mu+4*sig,length.out=1000)
fz=dnorm(z,mu,sig)
yM=max(dnorm(mu,mu,sig),hist.out$density)
# this takes the maximum
# of density and box heights for proper hist plotting.
hist(out,breaks=seq(m-.5,M+.5,1),freq=F,ylim=c(0,yM),main="")

lines(z,fz,lwd=2,col="red")
}
```

The resulting plots from

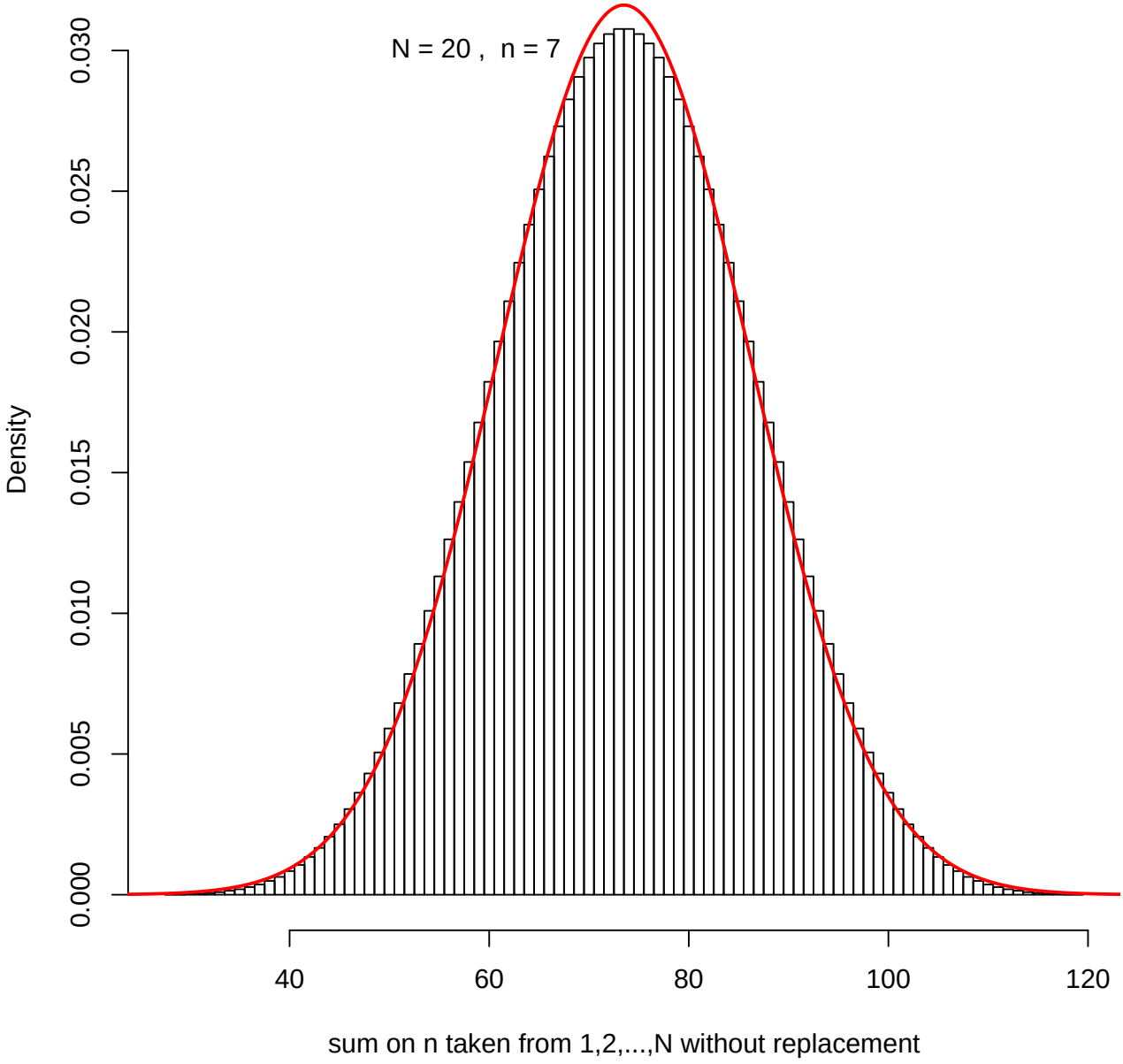
```
finite.population.CLT(20,7),
finite.population.CLT(20,2),
finite.population.CLT(20,1),
finite.population.CLT(20,19)
```

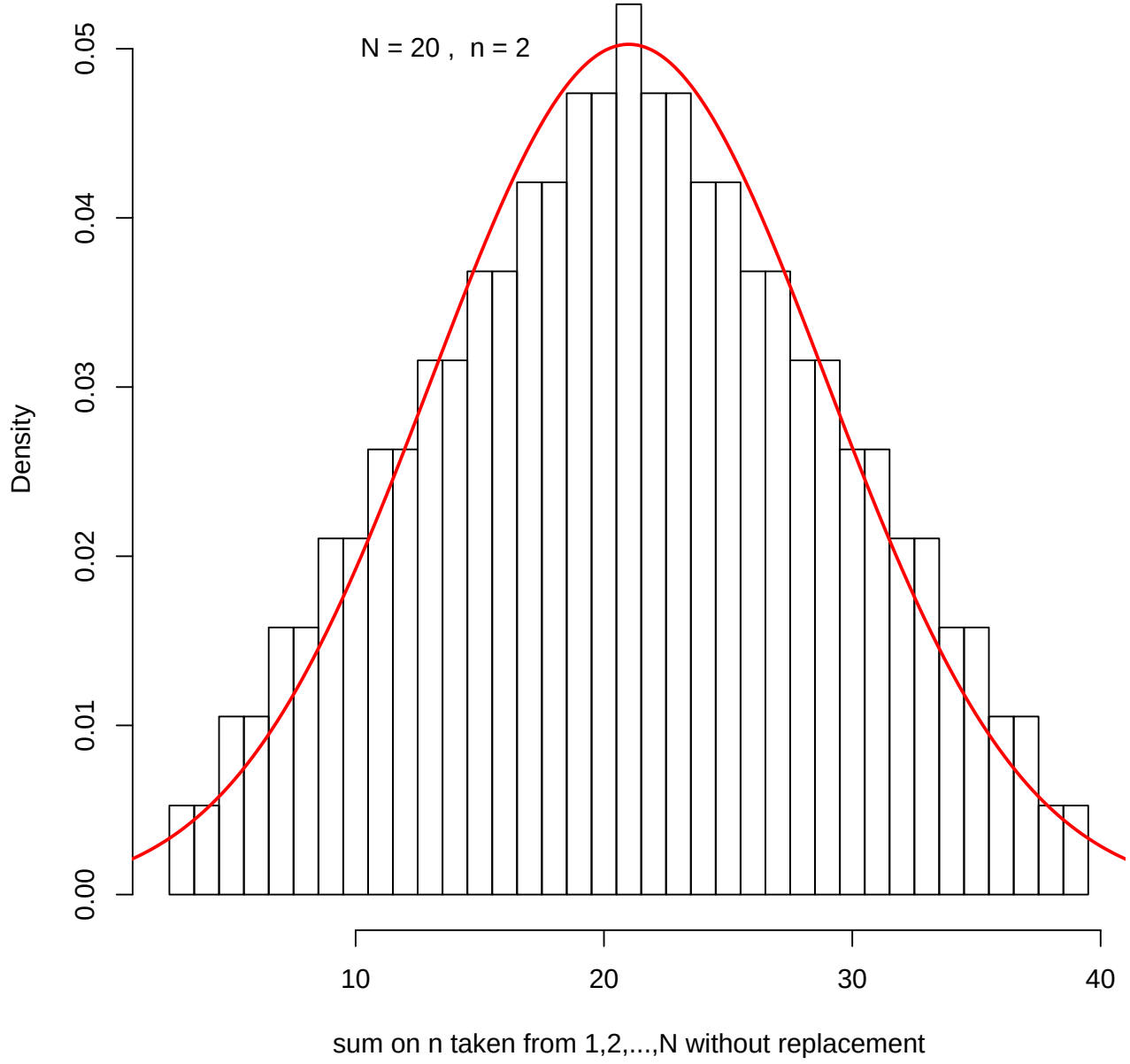
are shown below. Note that the last two plots are basically the same in character (except for location), because drawing 19 out of 20 is like drawing one from 20 to be left behind. The uniform nature of that histogram derives from the fact that each of the numbers $1, 2, \ldots, N$ has the same (uniform) chance $1/N$ of being selected in a sample of size 1.

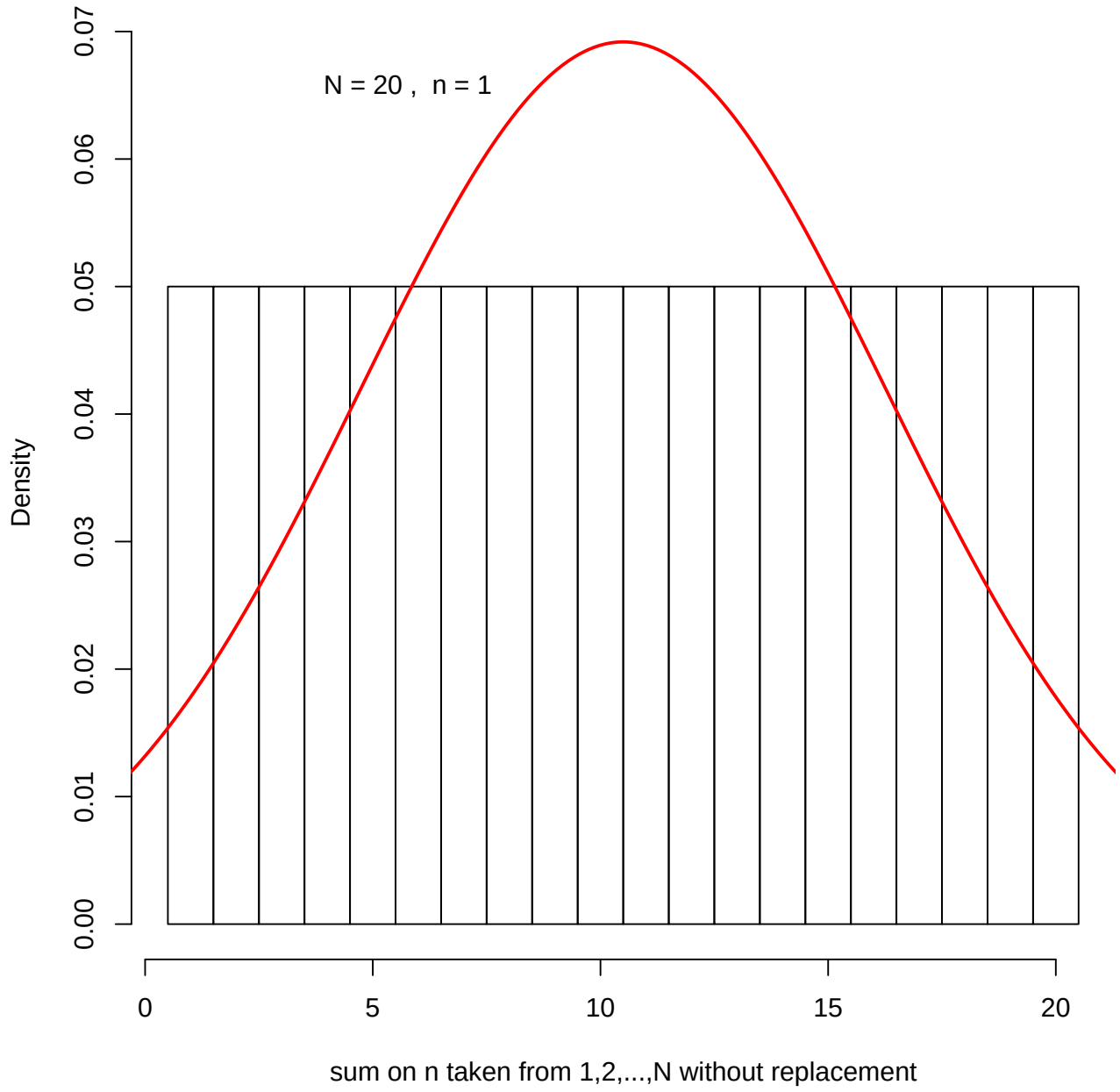
Note the difference in the two finite populations discussed. In the first case we sampled from

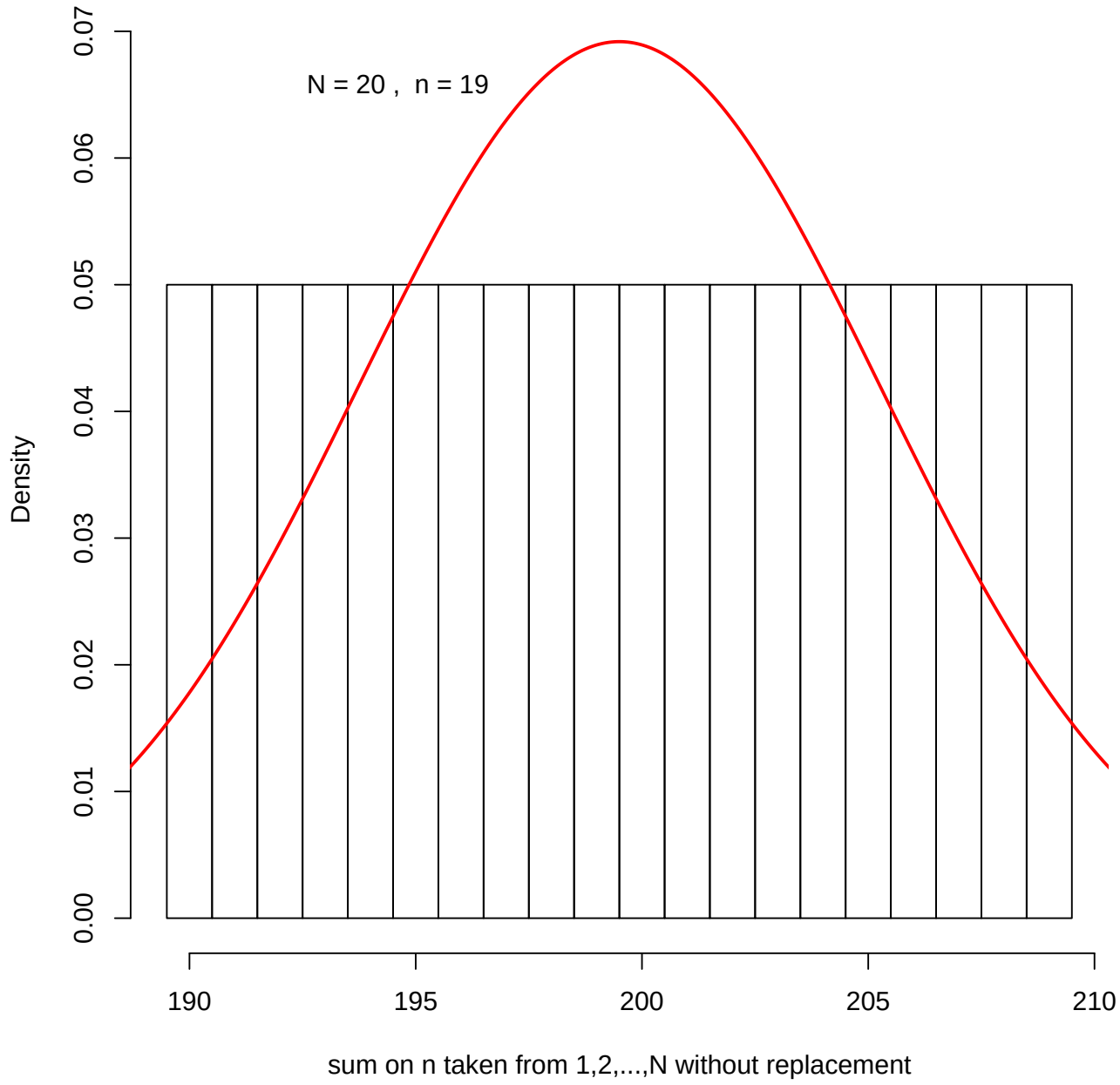
$$\overbrace{1, \ldots, 1}^{a\ 1's}, \overbrace{0, \ldots, 0}^{b\ 0's}$$

while in the second case we sampled from $1, 2, 3, \ldots, N$.


N = 20 , n = 7
Density
0.000
0.005
0.010
0.015
0.020
0.025
0.030
40
60
80
100
120
sum on n taken from 1,2,...,N without replacement


N = 20 , n = 2
Density
0.00
0.01
0.02
0.03
0.04
0.05
10
20
30
40
sum on n taken from 1,2,...,N without replacement


N = 20 , n = 1
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0
5
10
15
20
sum on n taken from 1,2,...,N without replacement


N = 20 , n = 19
Density
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
190
195
200
205
210
sum on n taken from 1,2,...,N without replacement