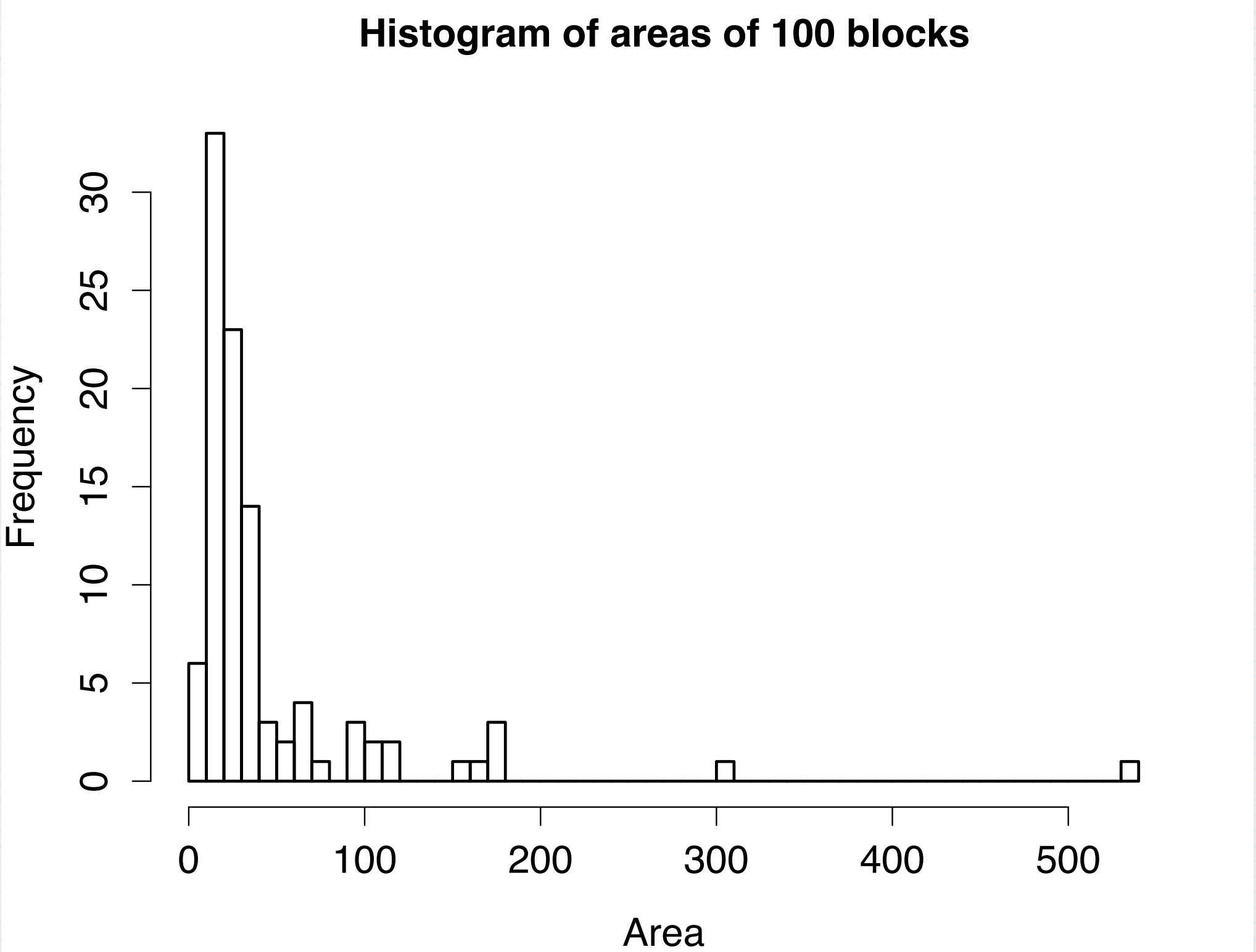
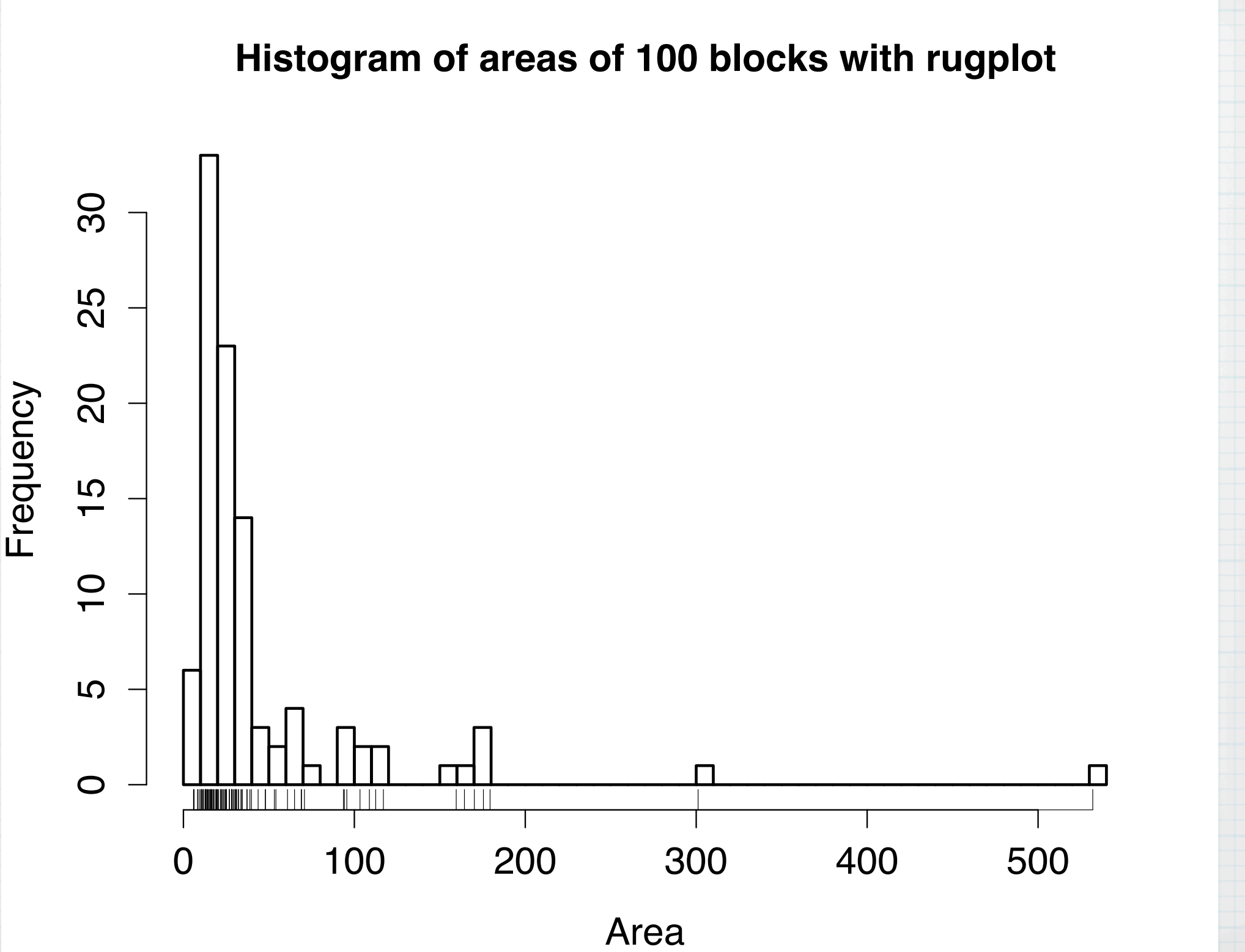
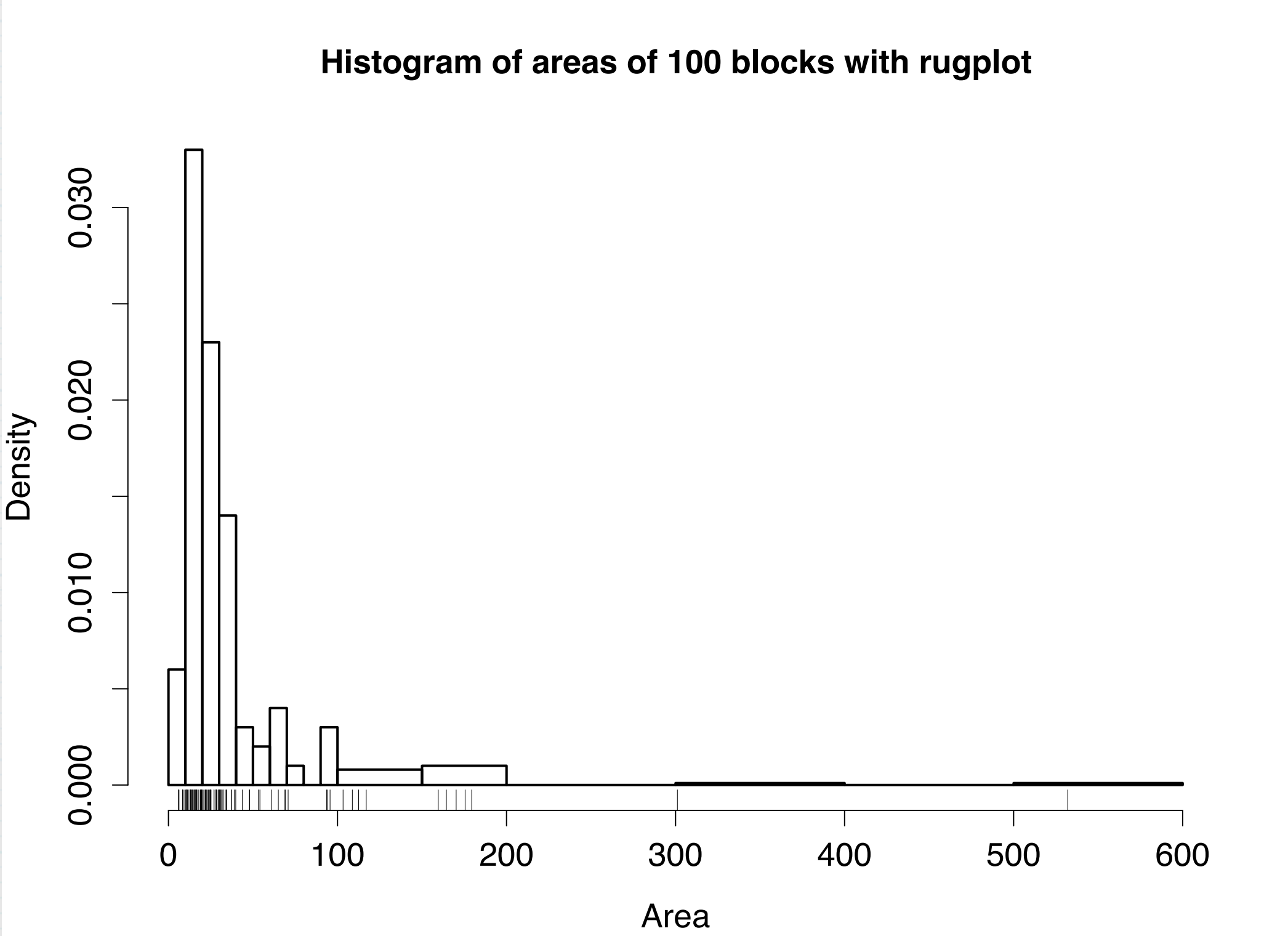
Sampling

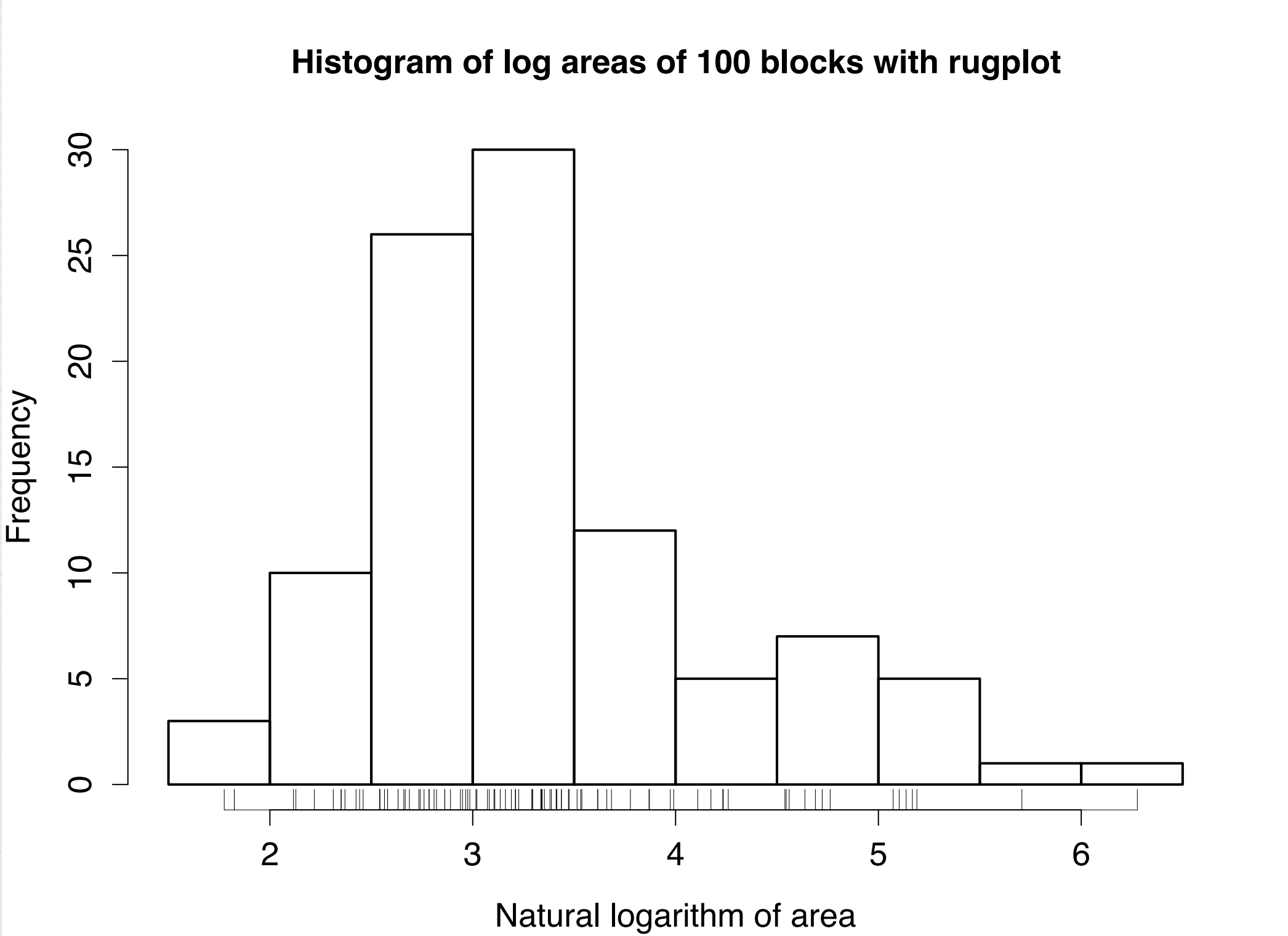
1/19/07

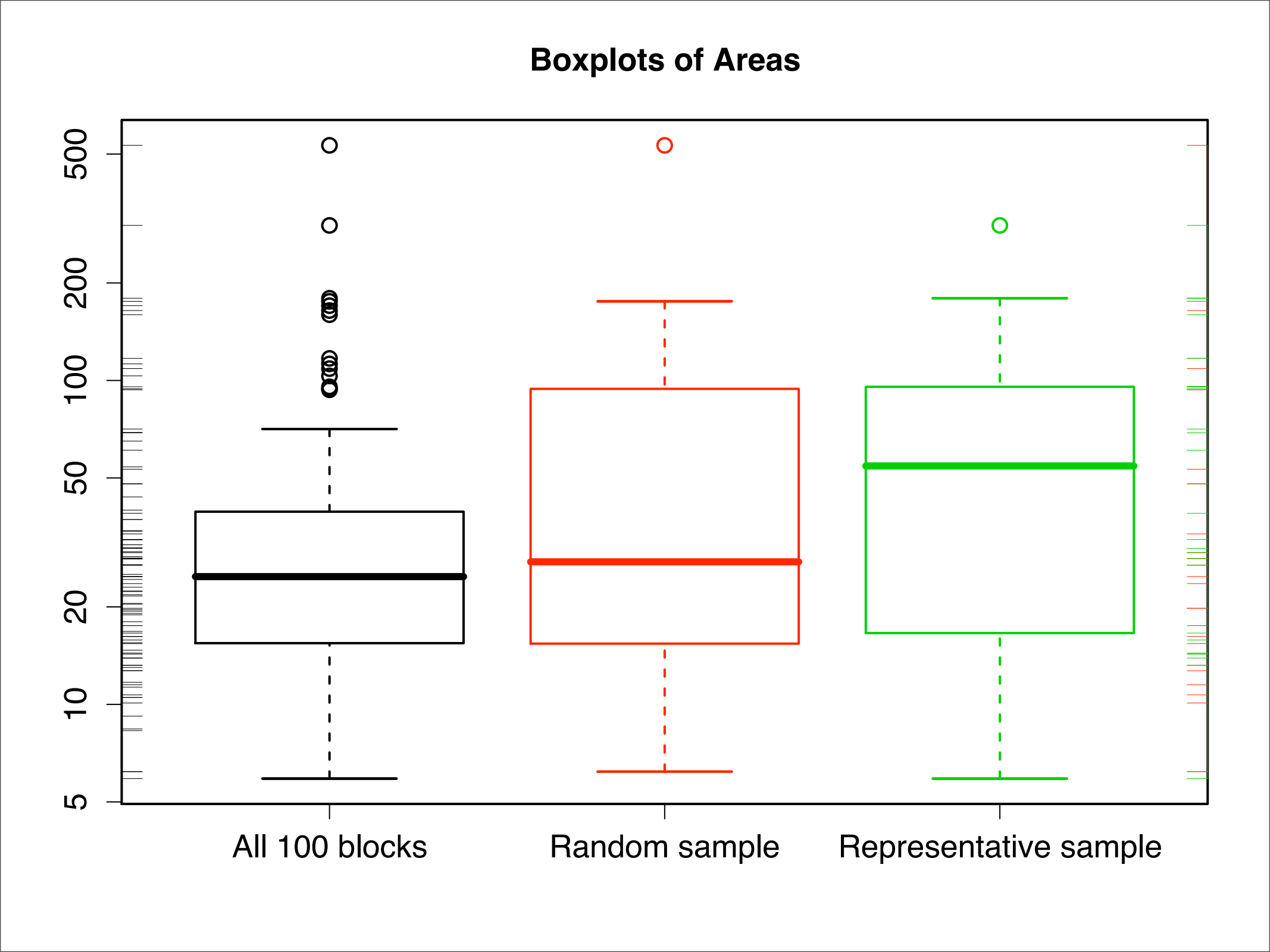
Summary of our sampling experiment

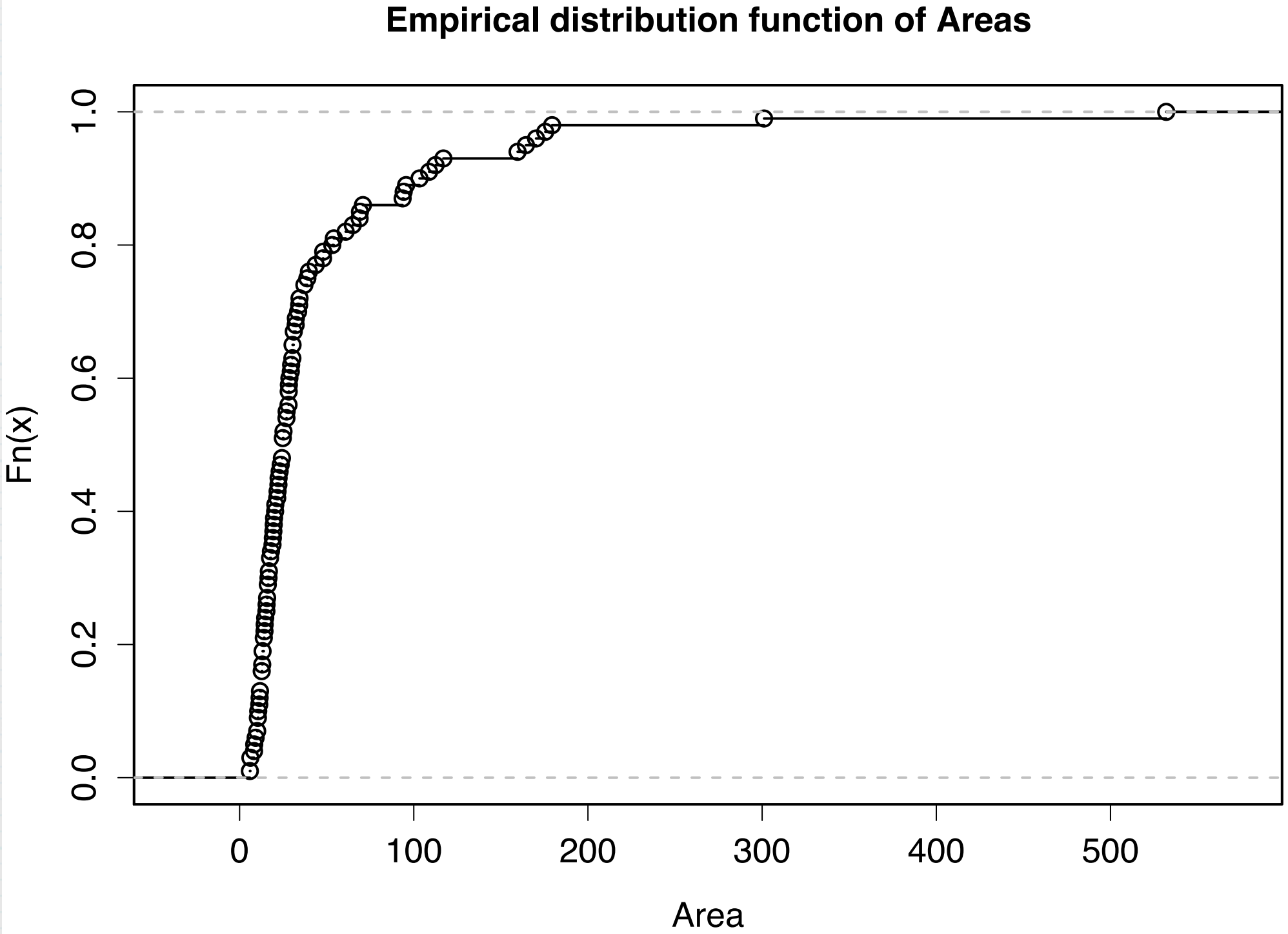

Histogram of areas of 100 blocks
Frequency
0
5
10
15
20
25
30
0
100
200
300
400
500
Area

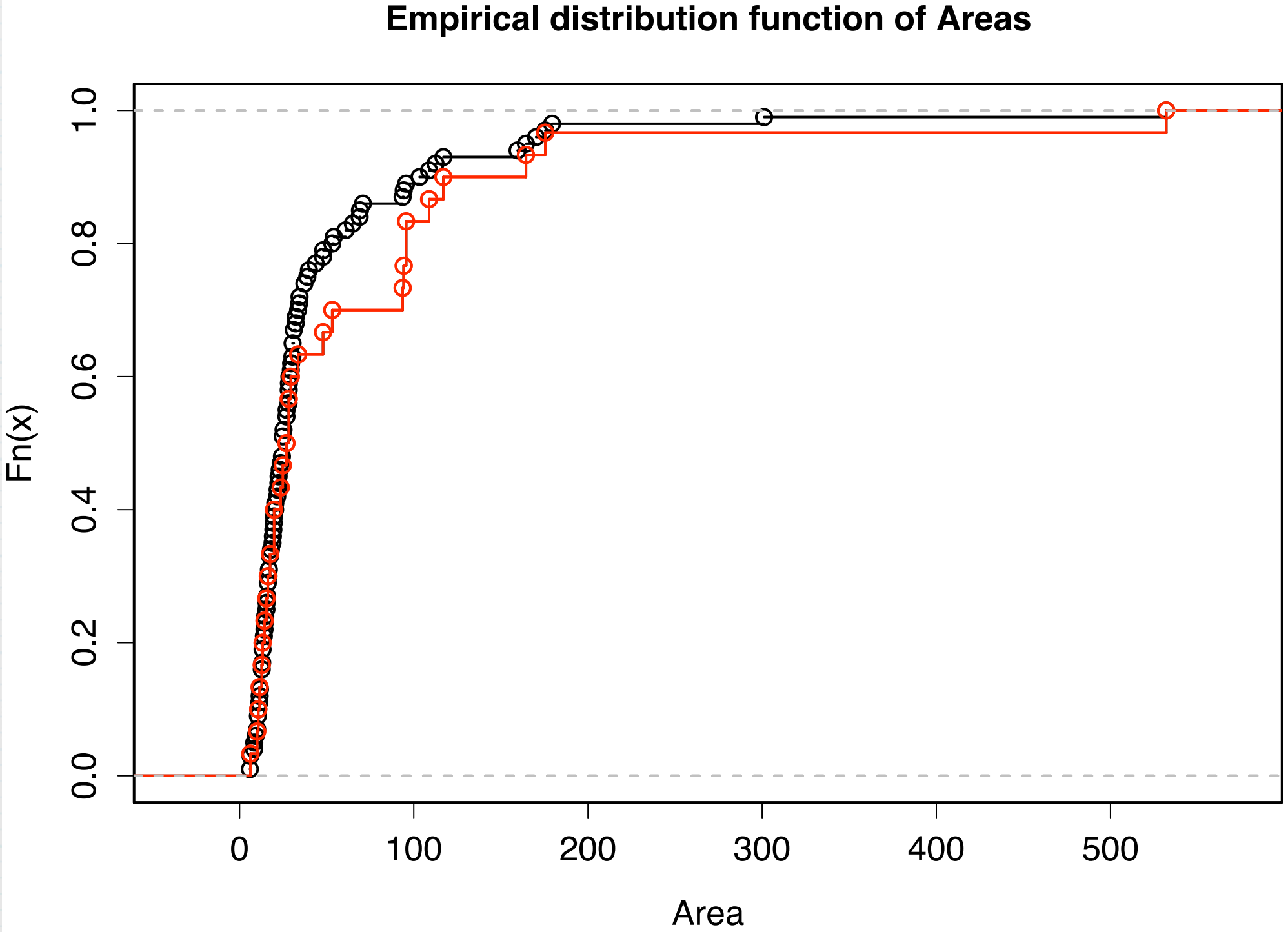

Histogram of areas of 100 blocks with rugplot
Frequency
0
5
10
15
20
25
30
0
100
200
300
400
500
Area

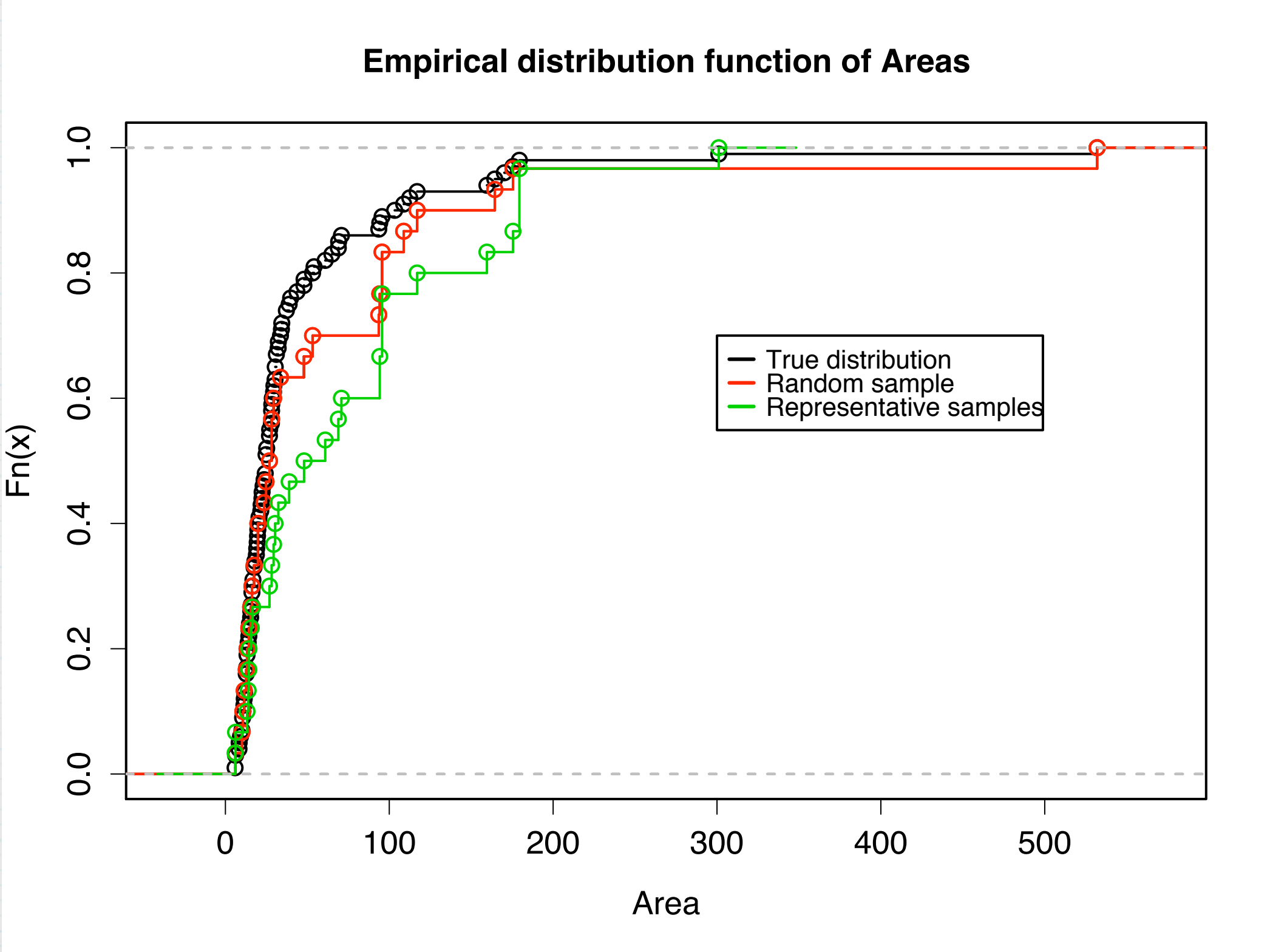

Histogram of areas of 100 blocks with rugplot
Density
0.000
0.010
0.020
0.030
0
100
200
300
400
500
600
Area

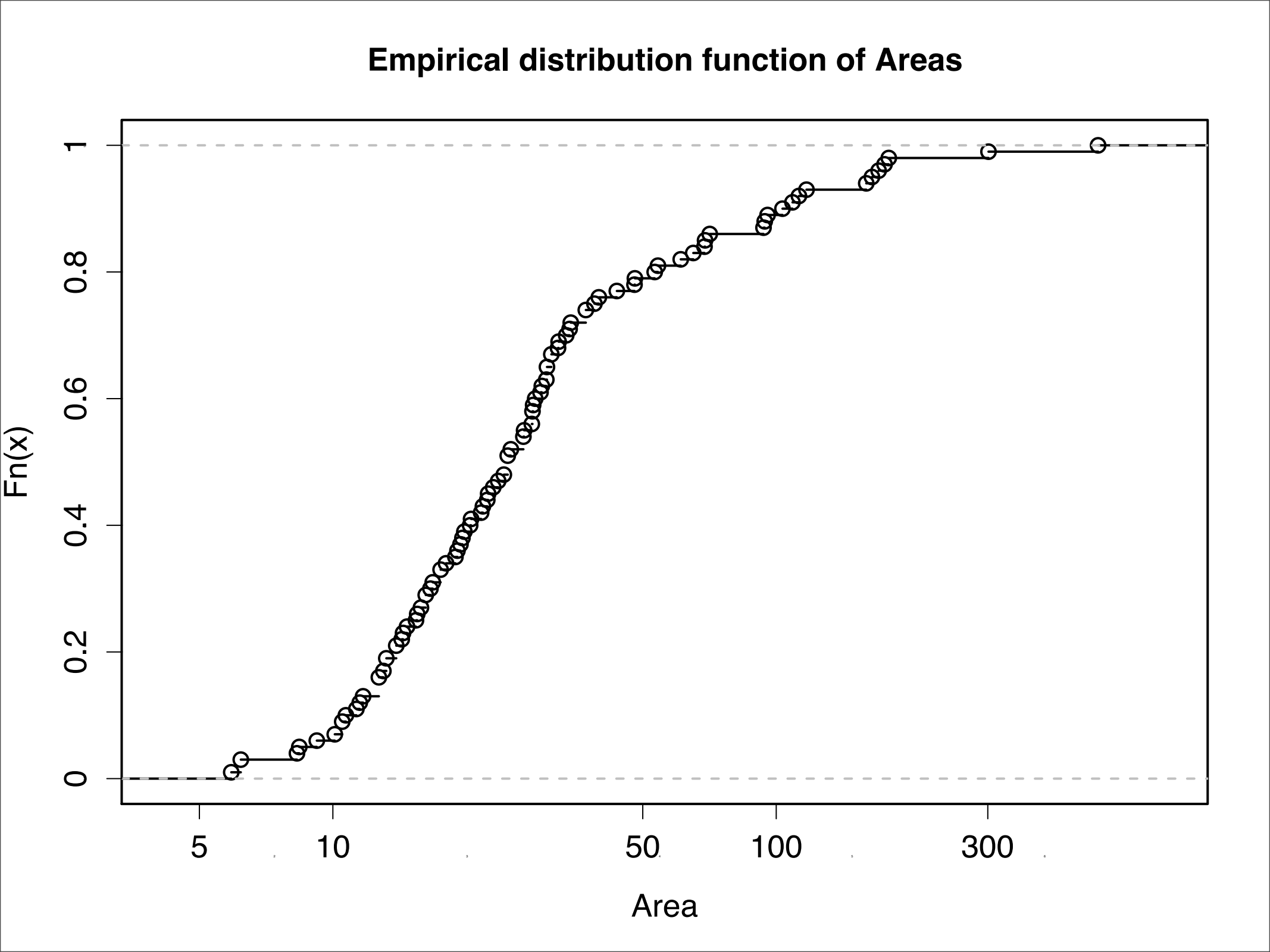

Histogram of log areas of 100 blocks with rugplot
Frequency
0
5
10
15
20
25
30
2
3
4
5
6
Natural logarithm of area

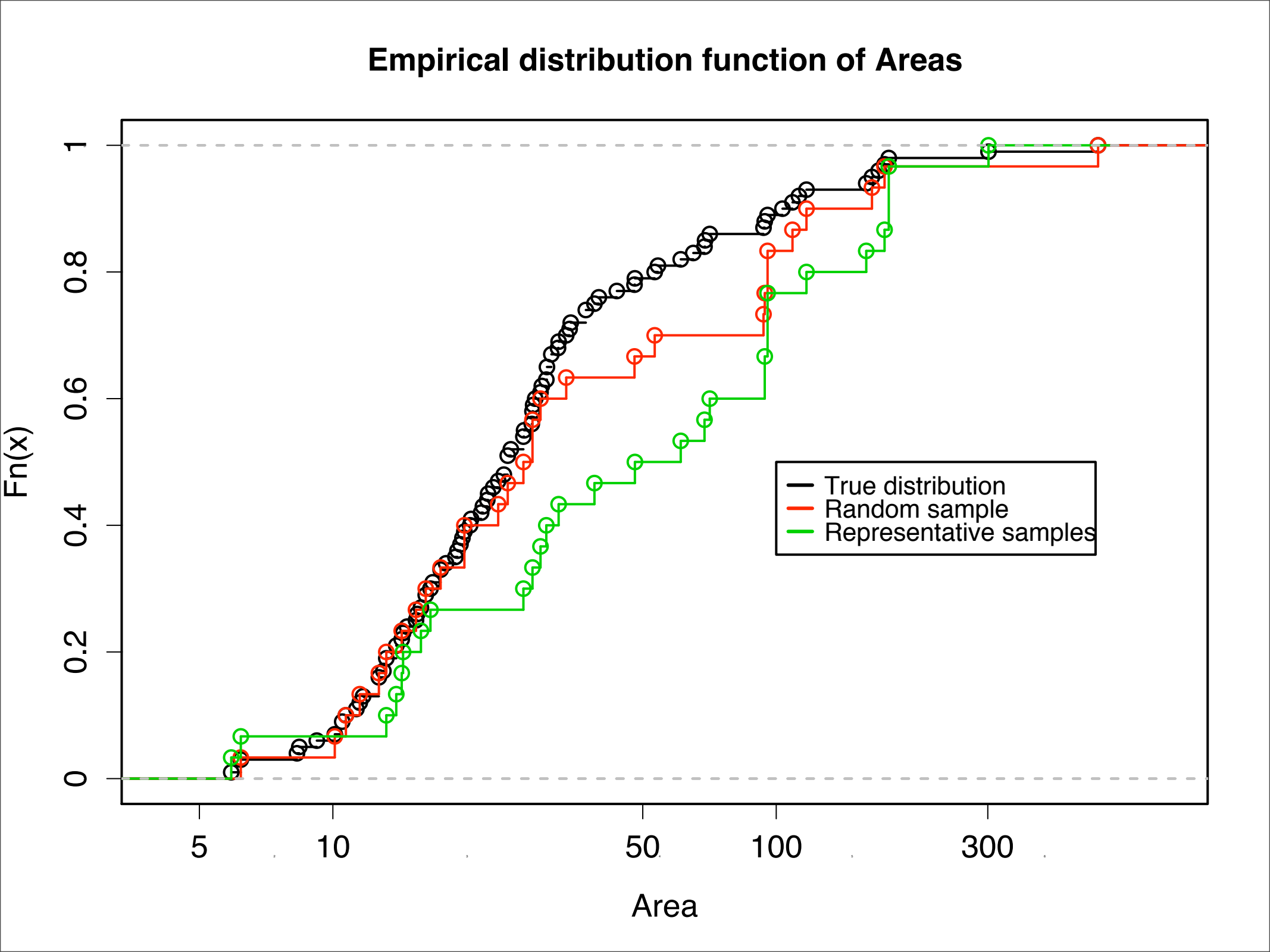

Boxplots of Areas
500
200
100
50
20
10
5
All 100 blocks
Random sample
Representative sample


Empirical distribution function of Areas
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
Area


Empirical distribution function of Areas
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
Area


Empirical distribution function of Areas
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
400
500
Area
True distribution
Random sample
Representative samples


Empirical distribution function of Areas
Fn(x)
0
0.2
0.4
0.6
0.8
1
5
10
50
100
300
Area


Empirical distribution function of Areas
Fn(x)
0
0.2
0.4
0.6
0.8
1
5
10
50
100
300
Area
True distribution
Random sample
Representative samples

Inferring (Estimating) Means

* Sampling from a population or from a probability distribution is like drawing from a box
* Sums or averages (or other functions) of the sample have a "sampling distribution"
* If it's a simple random sample (with or without replacement) the expectation of the sampling mean is the same as the mean of the numbers in the box.

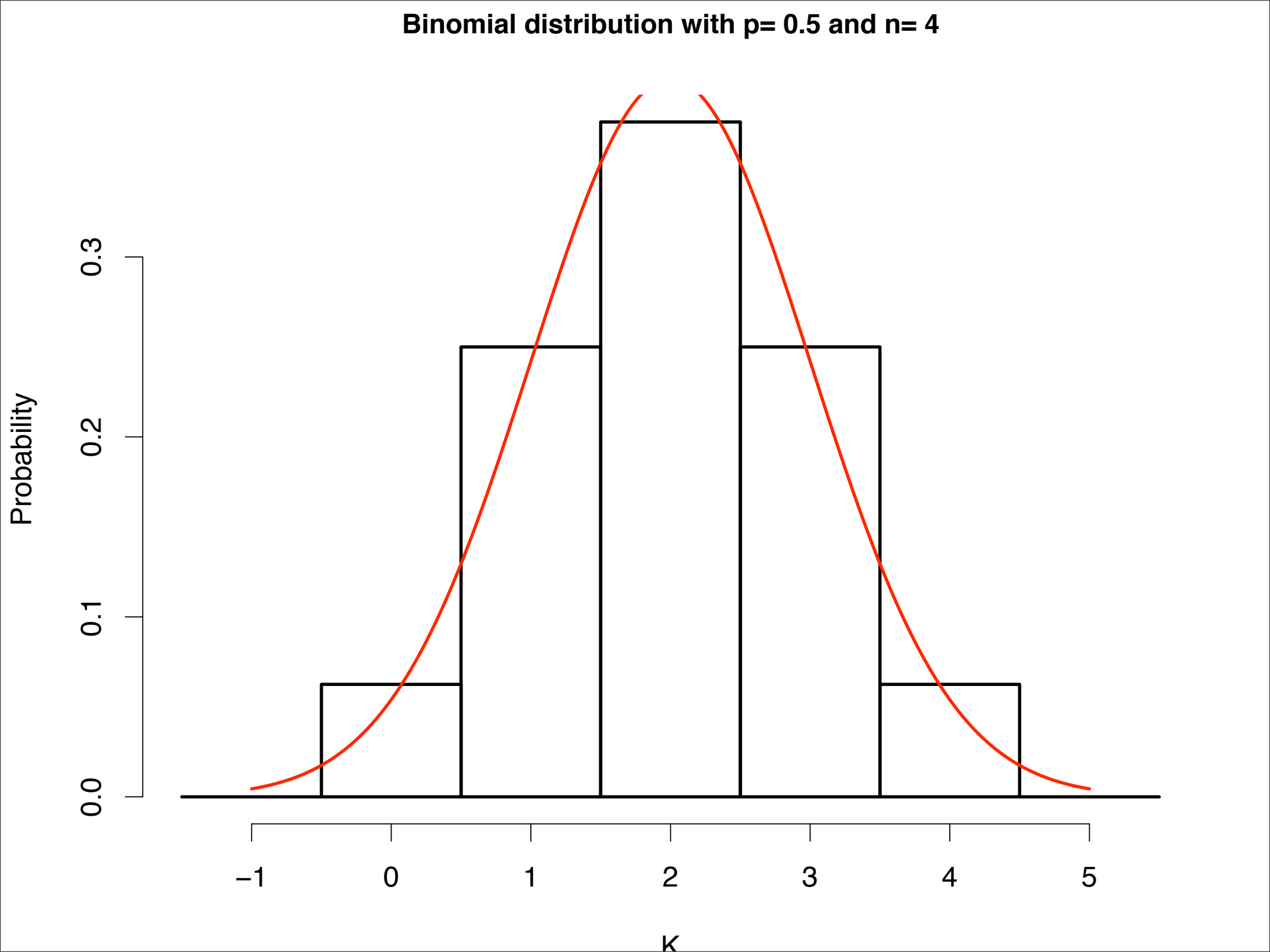
10 weighings of unknown weight average 10.5g. Known SD 2g, errors normally distributed

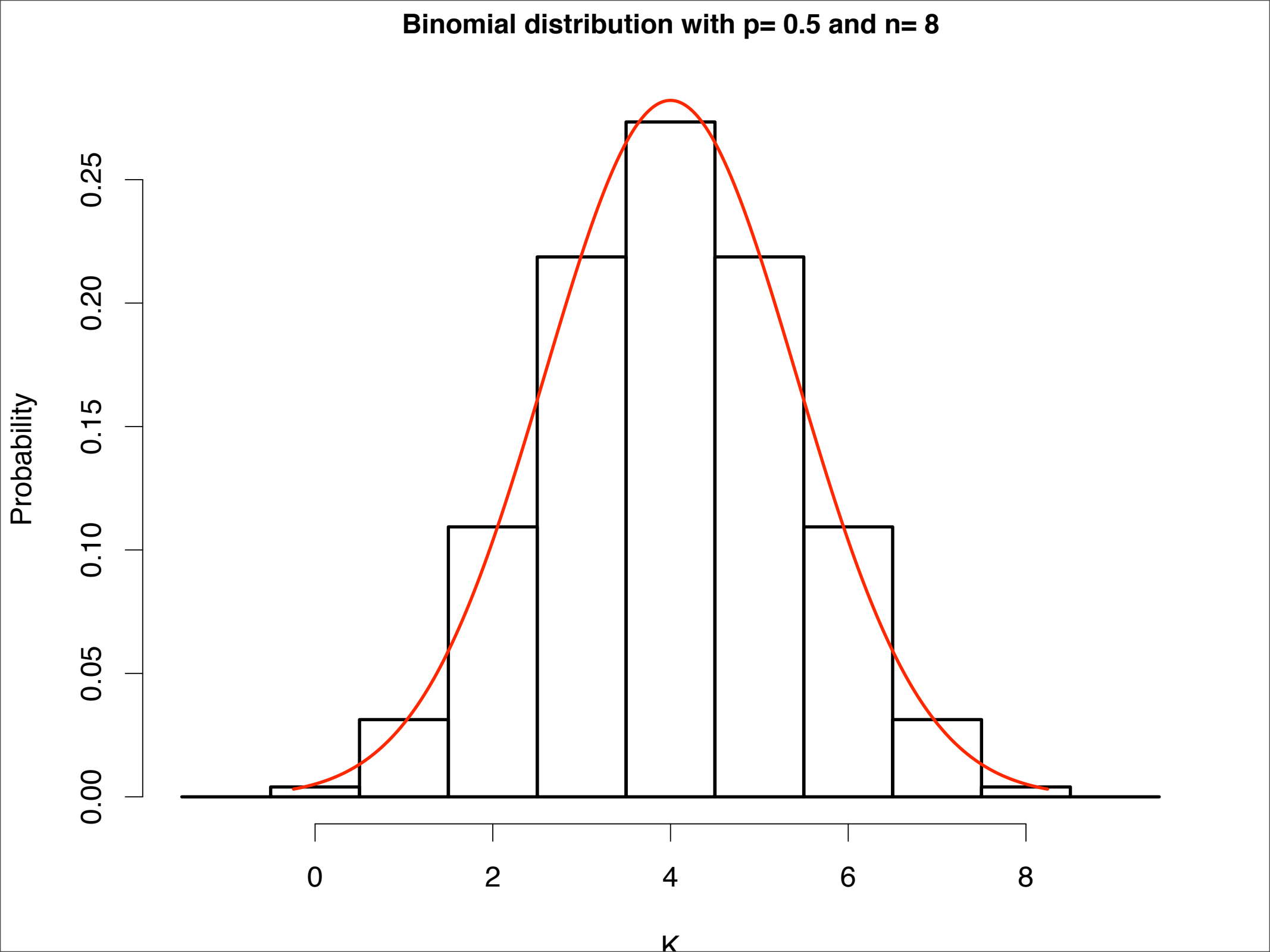
* What should you estimate for the true weight?
* How confident should you be about that answer?

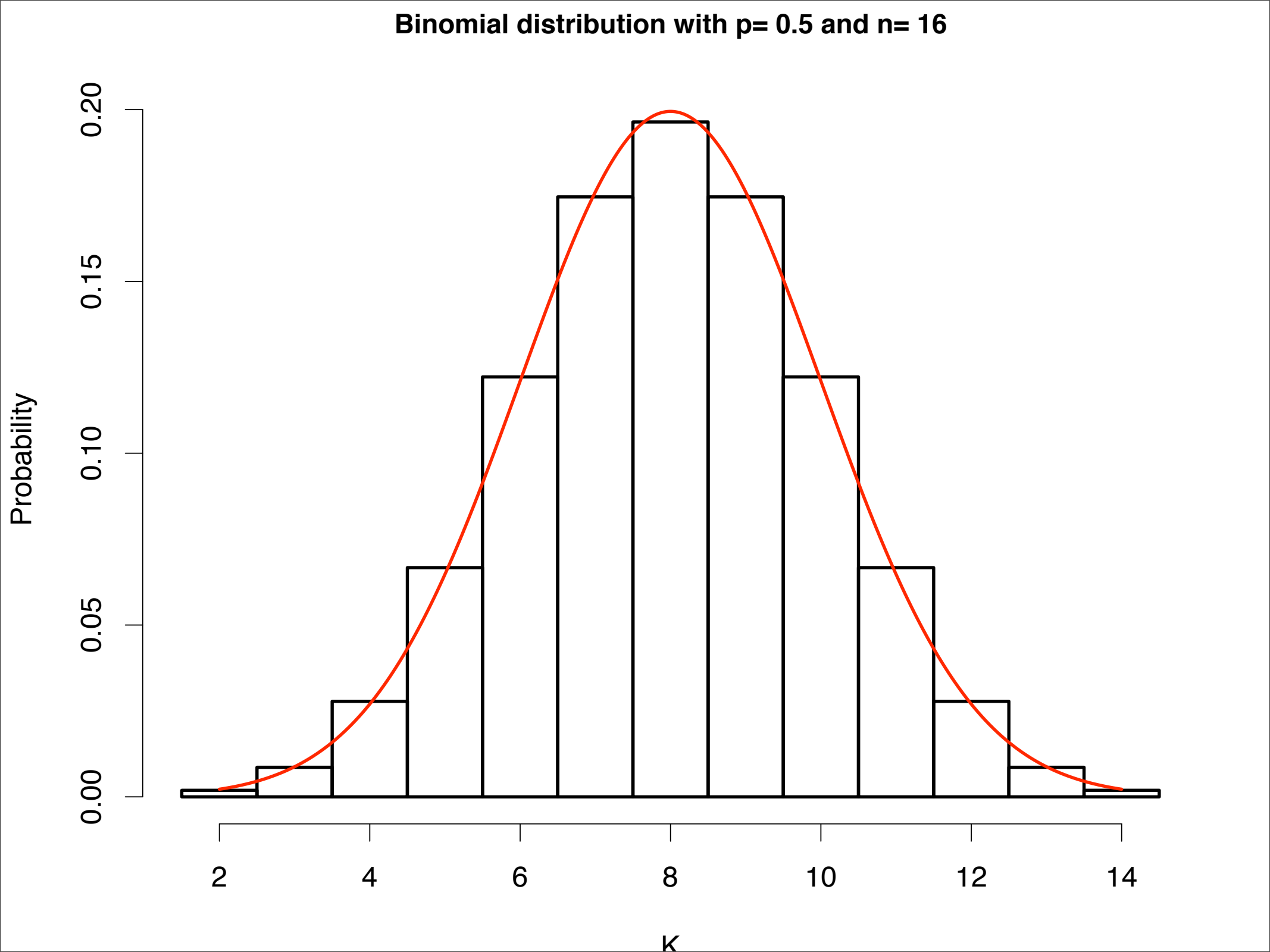
ESP experiment

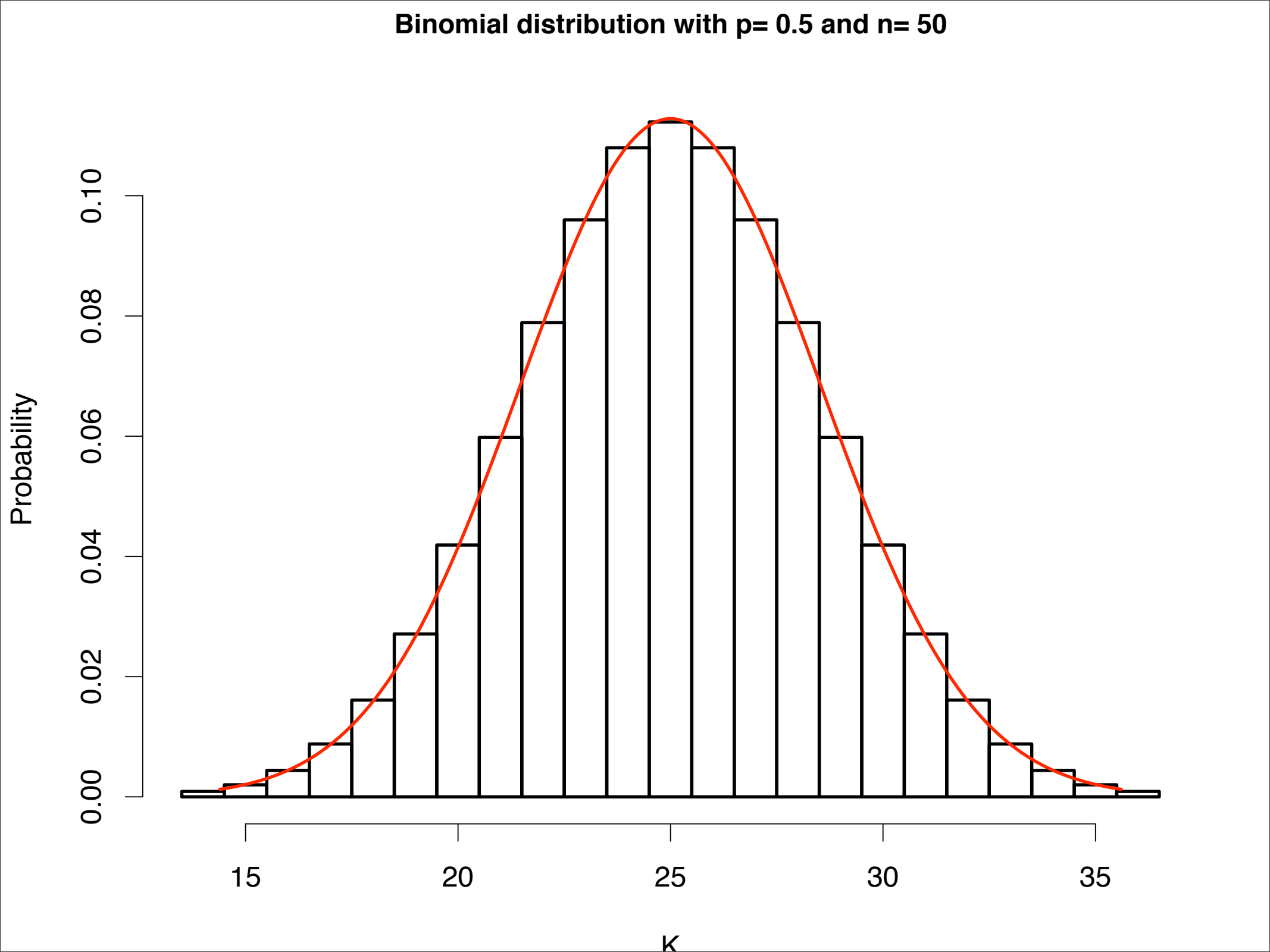
- Charles Tart (1970s): 7500 (500x15) attempts on "Aquarius machine"
- Subjects predict which of four lights will come on
- Signal tells them if they were right.
- 2,006 correct predictions in 7500 attempts

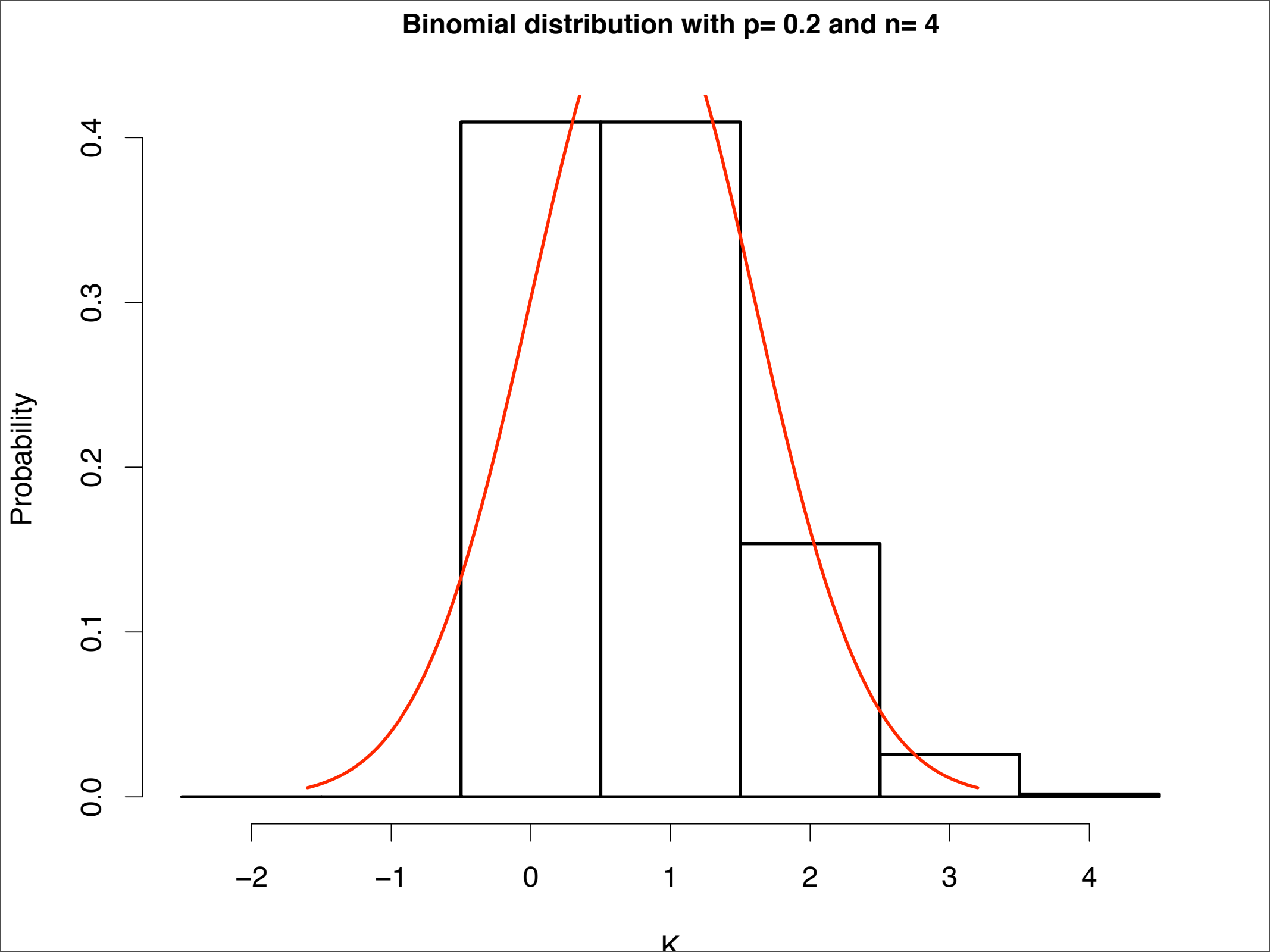
Normal approximation

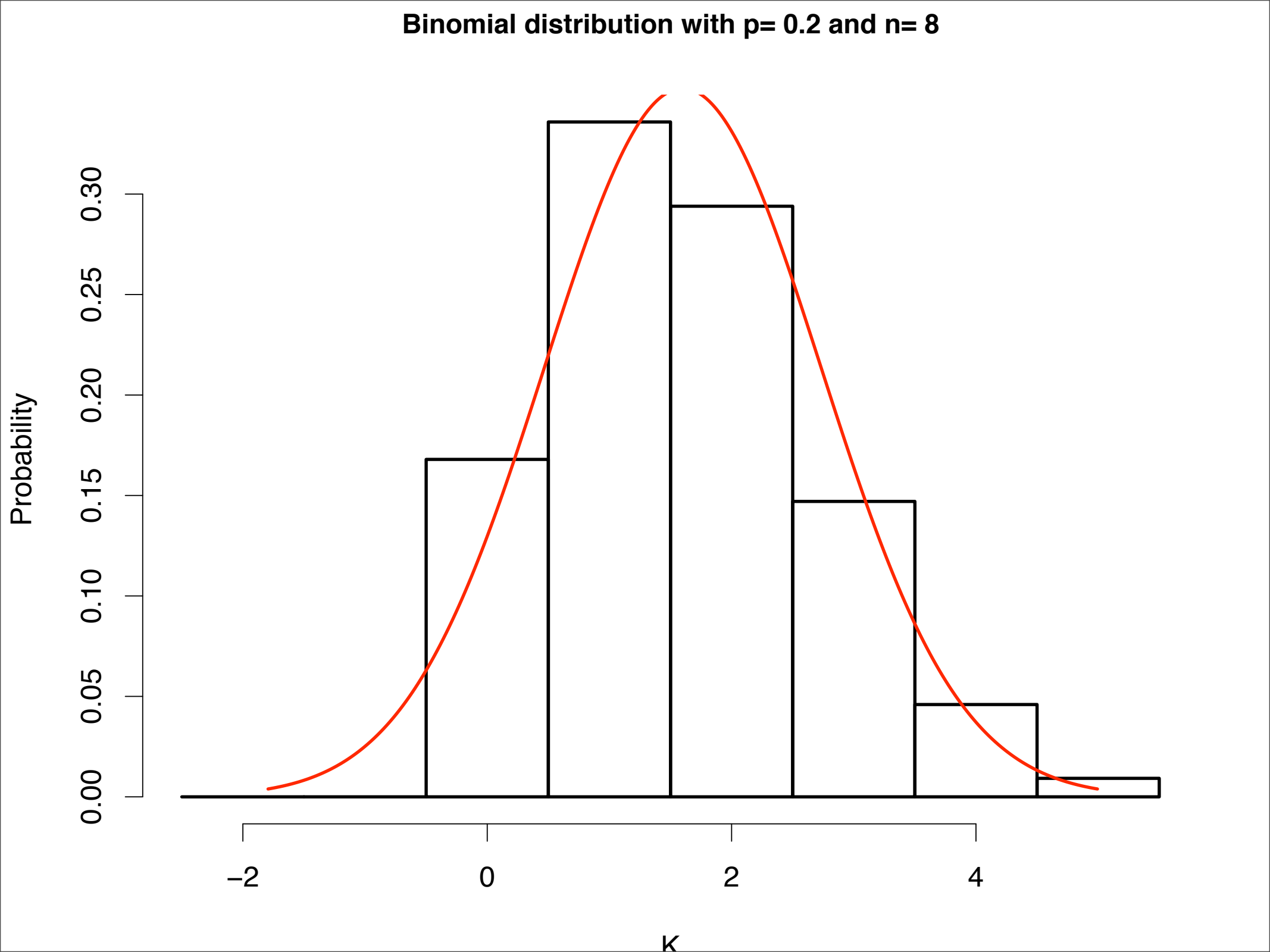

Binomial distribution with p= 0.5 and n= 4
Probability
0.0
0.1
0.2
0.3
−1
0
1
2
3
4
5
K

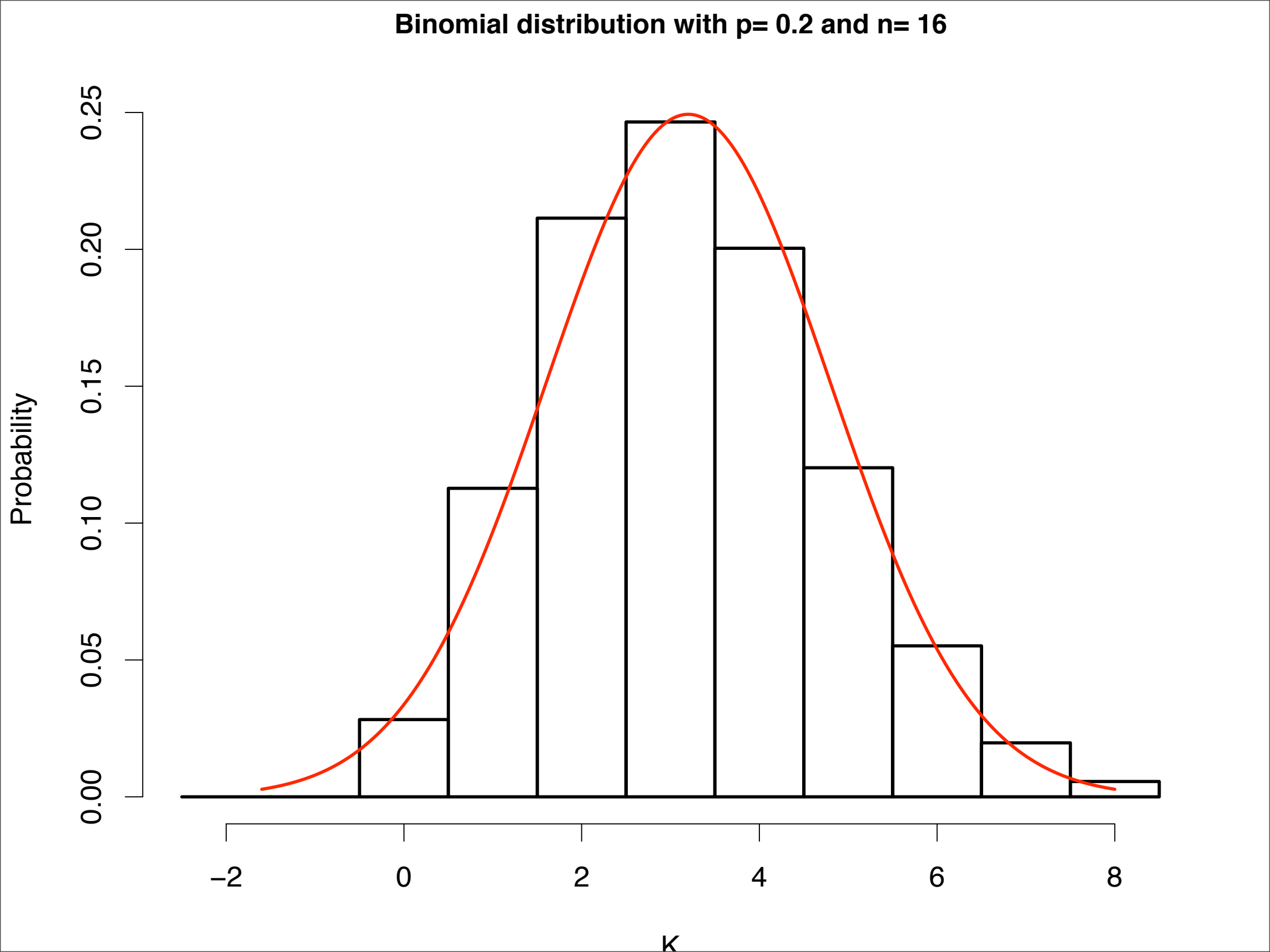

Binomial distribution with p= 0.5 and n= 8
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
K

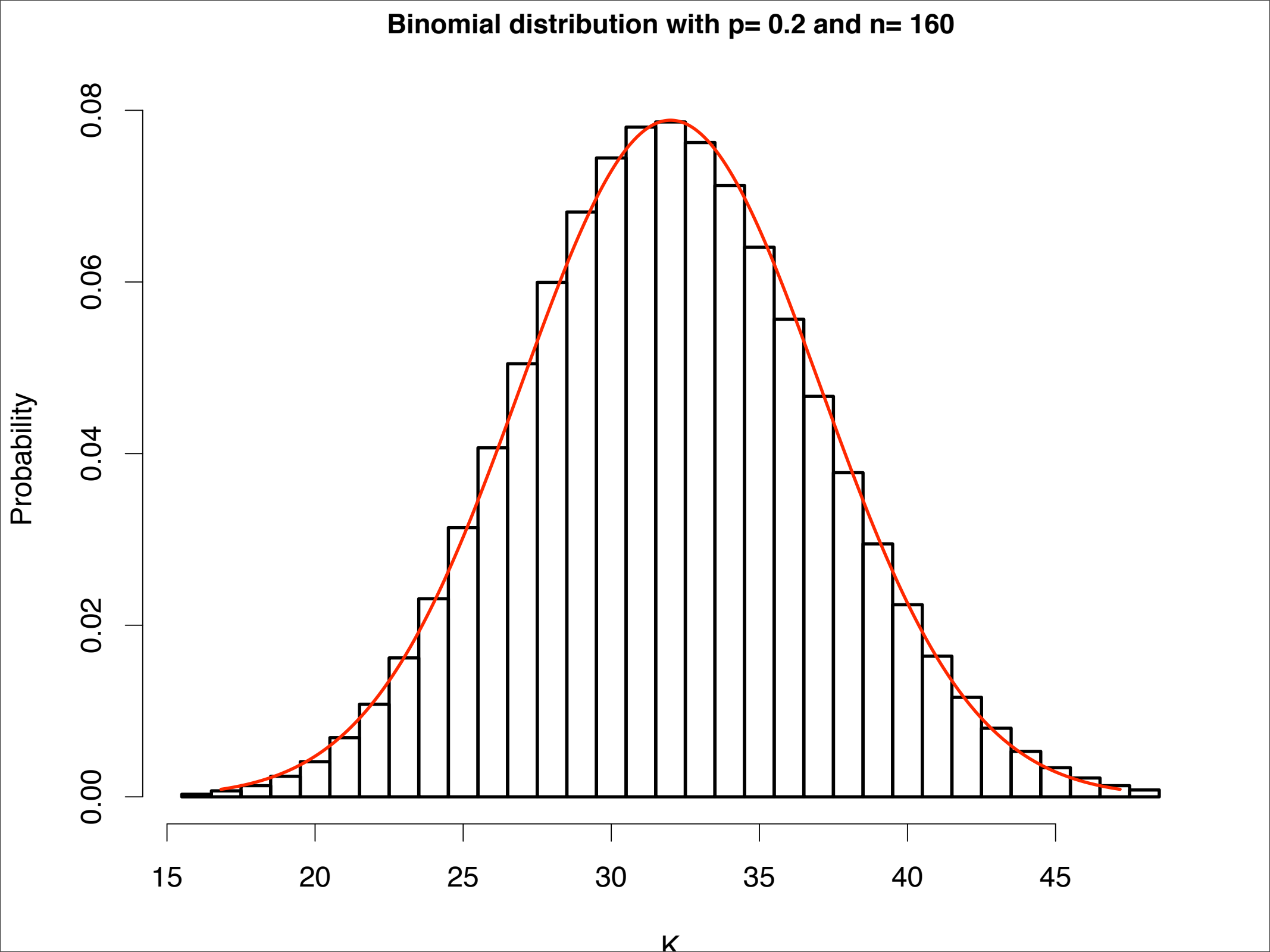

Binomial distribution with p= 0.5 and n= 16
Probability
0.00
0.05
0.10
0.15
0.20
2
4
6
8
10
12
14
K

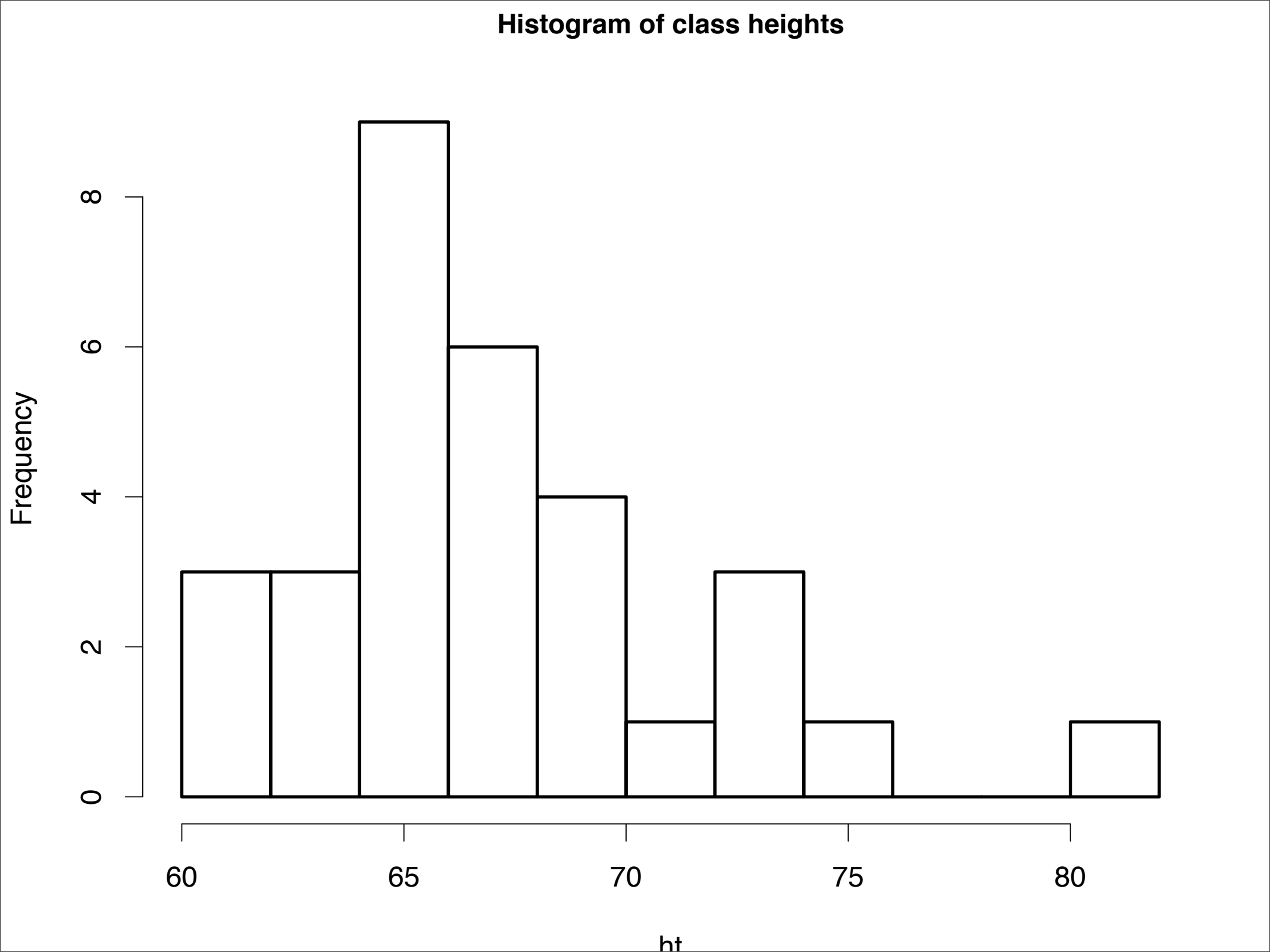

Binomial distribution with p= 0.5 and n= 50
Probability
0.00
0.02
0.04
0.06
0.08
0.10
15
20
25
30
35
K

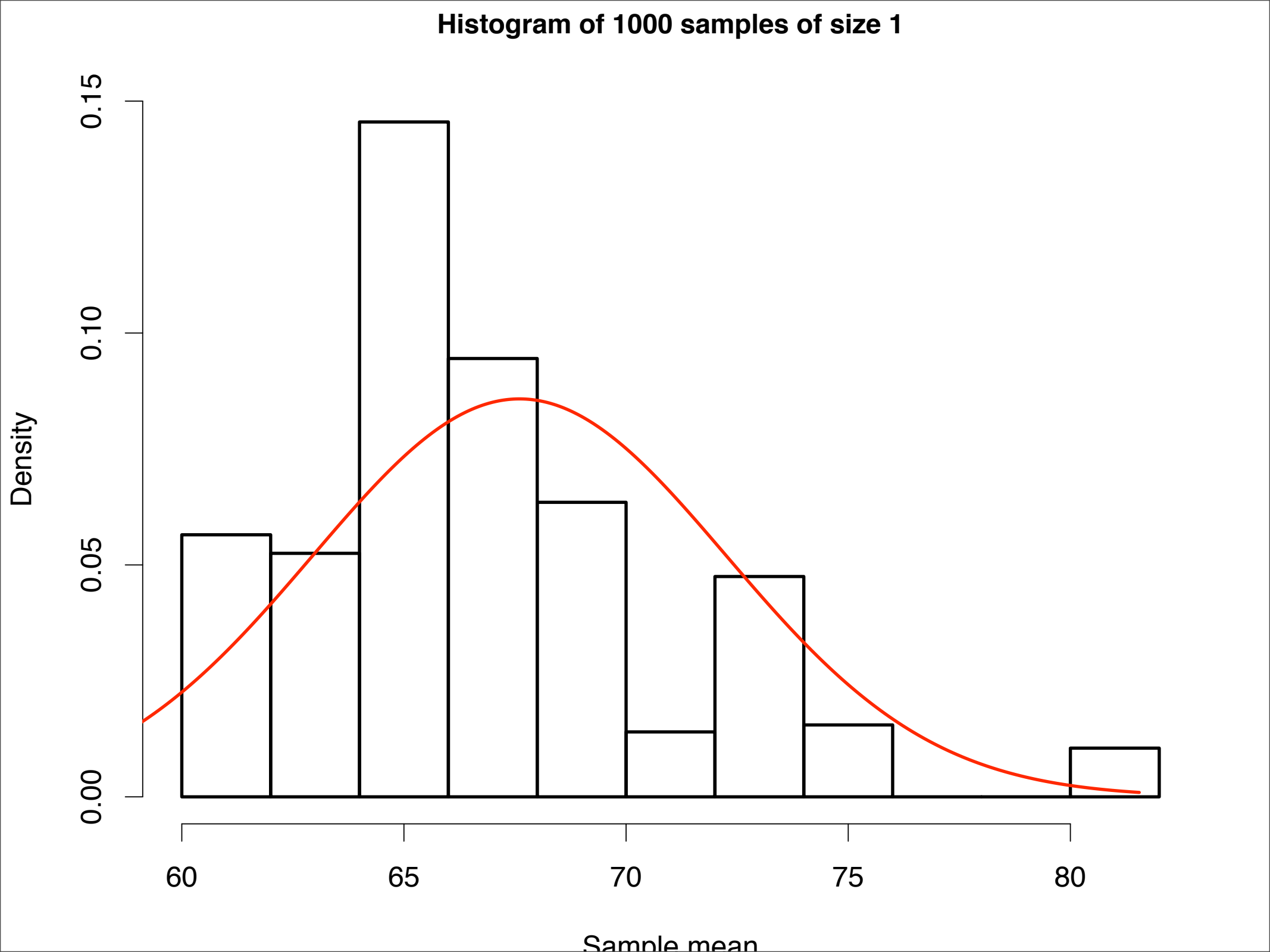

Binomial distribution with p= 0.2 and n= 4
Probability
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
3
4
K

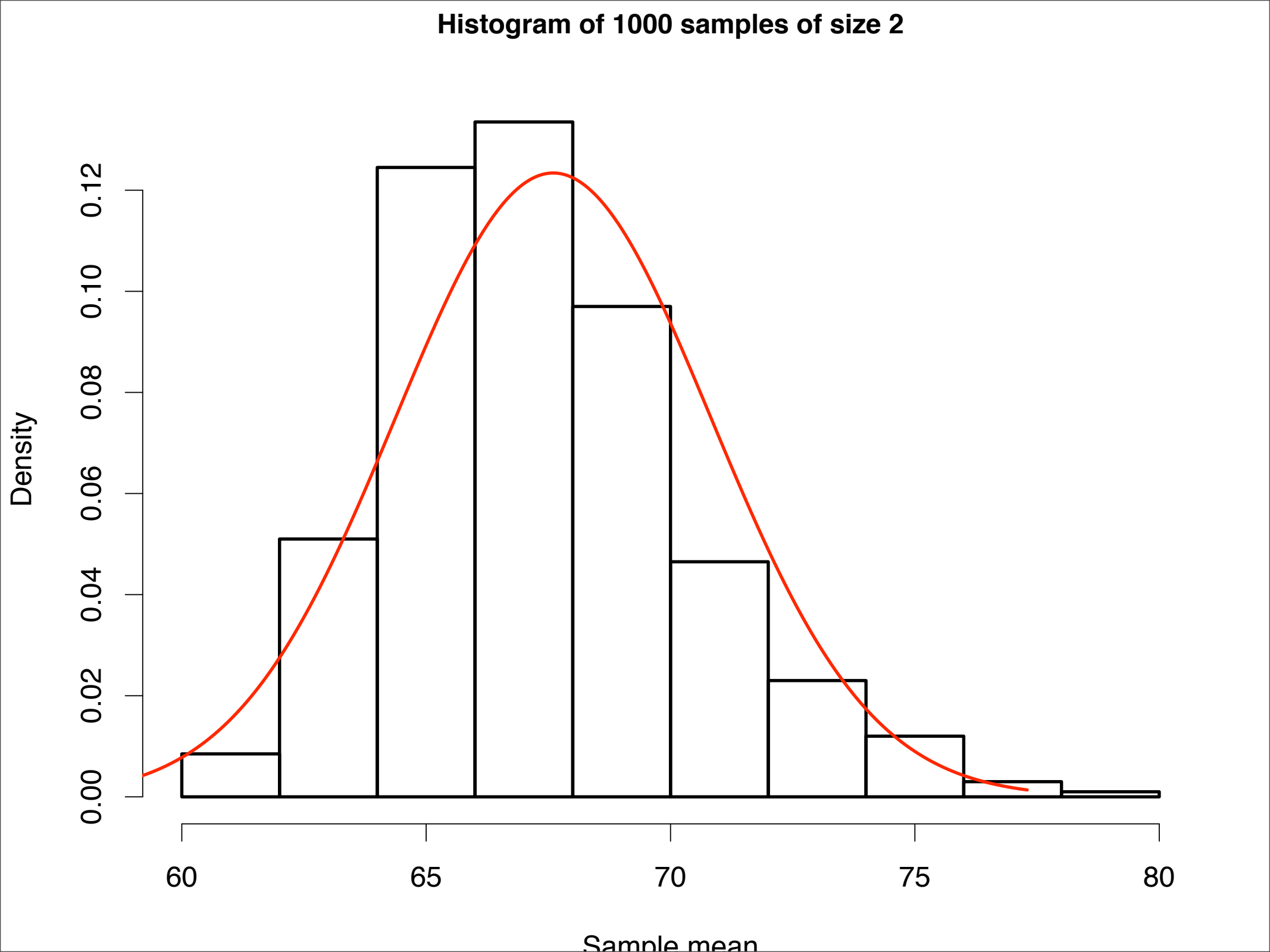

Binomial distribution with p= 0.2 and n= 8
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0.30
−2
0
2
4
K

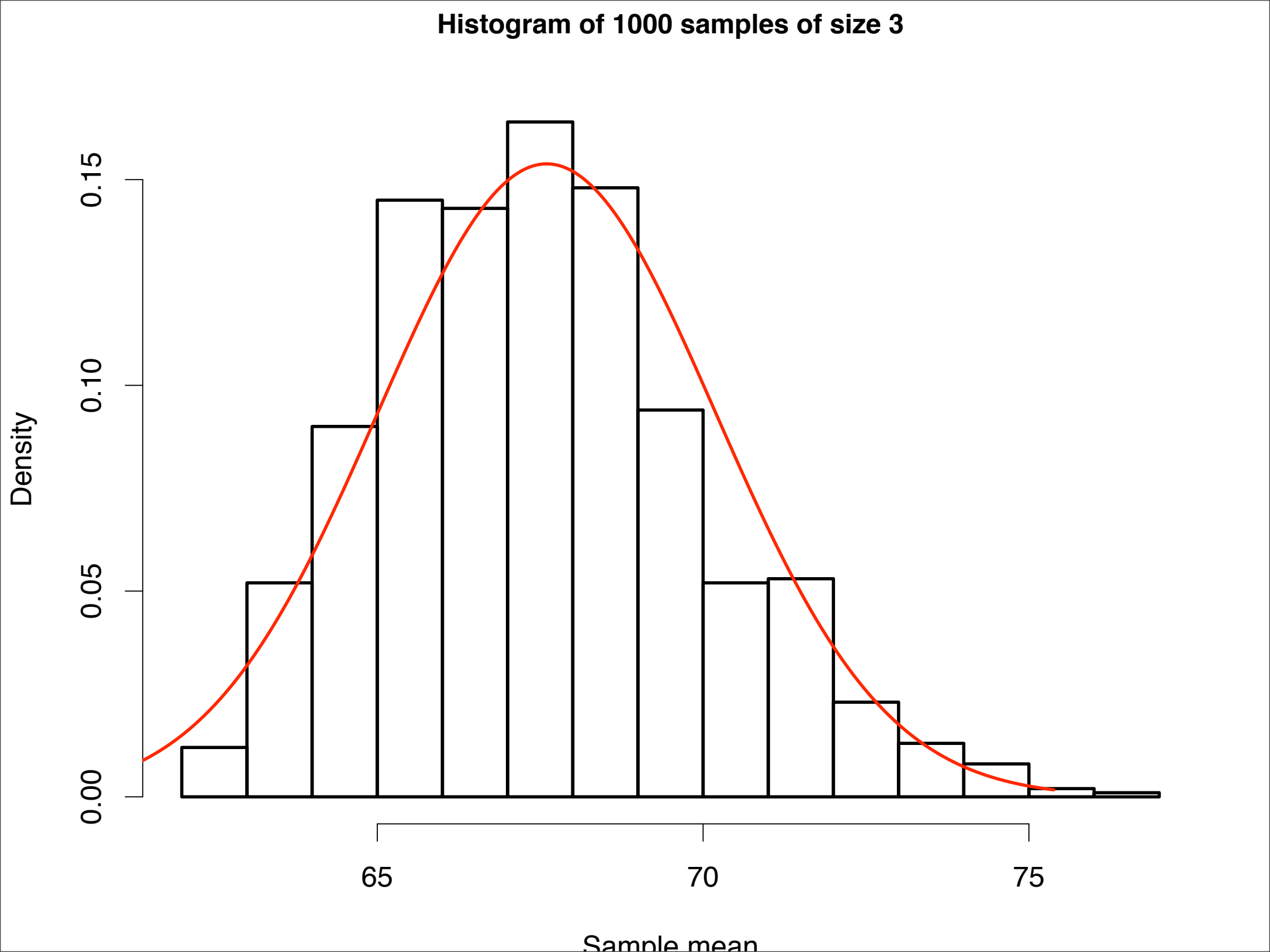

Binomial distribution with p= 0.2 and n= 16
Probability
0.00
0.05
0.10
0.15
0.20
0.25
−2
0
2
4
6
8
K

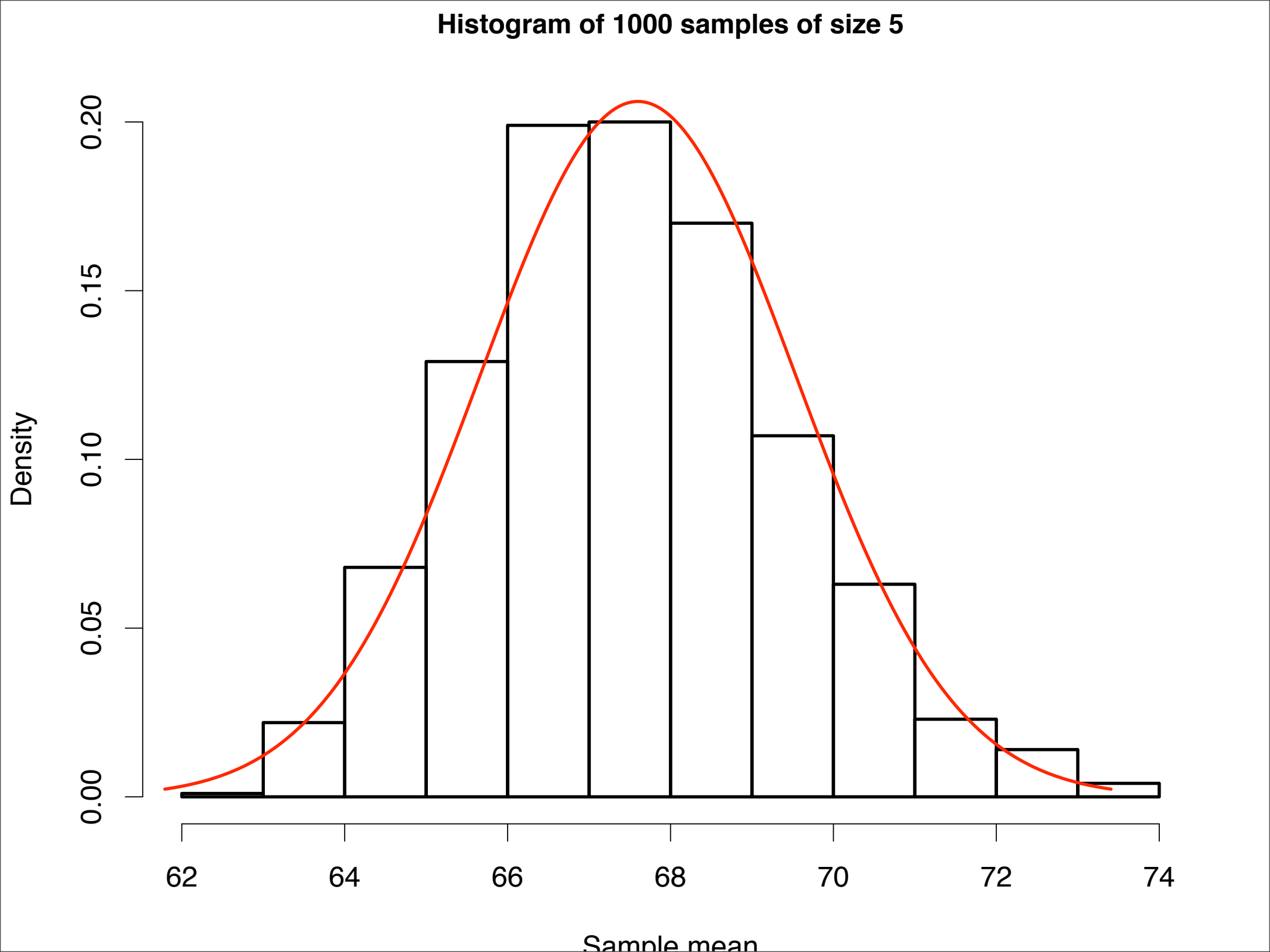

Binomial distribution with p= 0.2 and n= 160
Probability
0.00
0.02
0.04
0.06
0.08
15
20
25
30
35
40
45
K

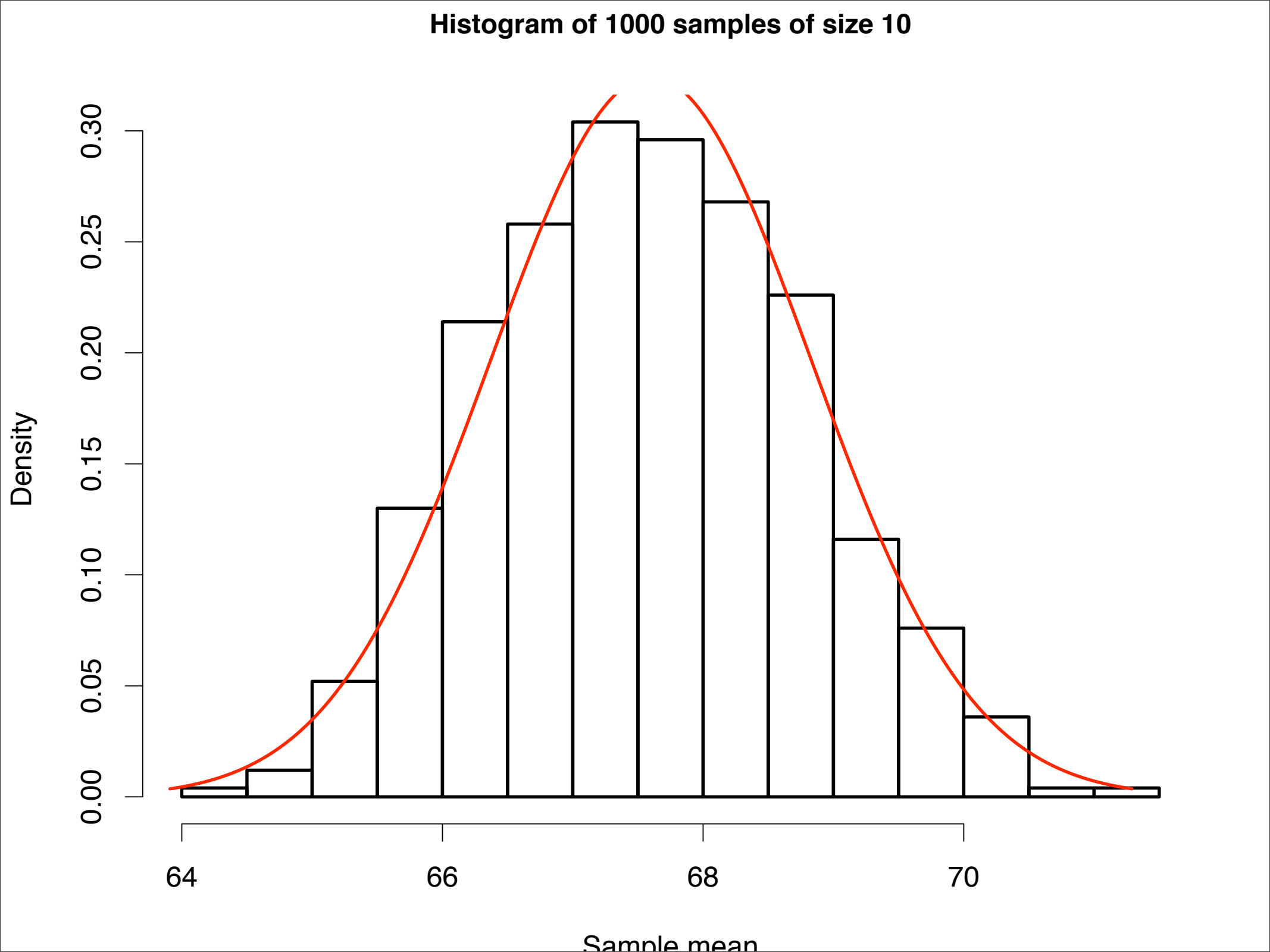

Histogram of class heights
Frequency
0
2
4
6
8
60
65
70
75
80
ht

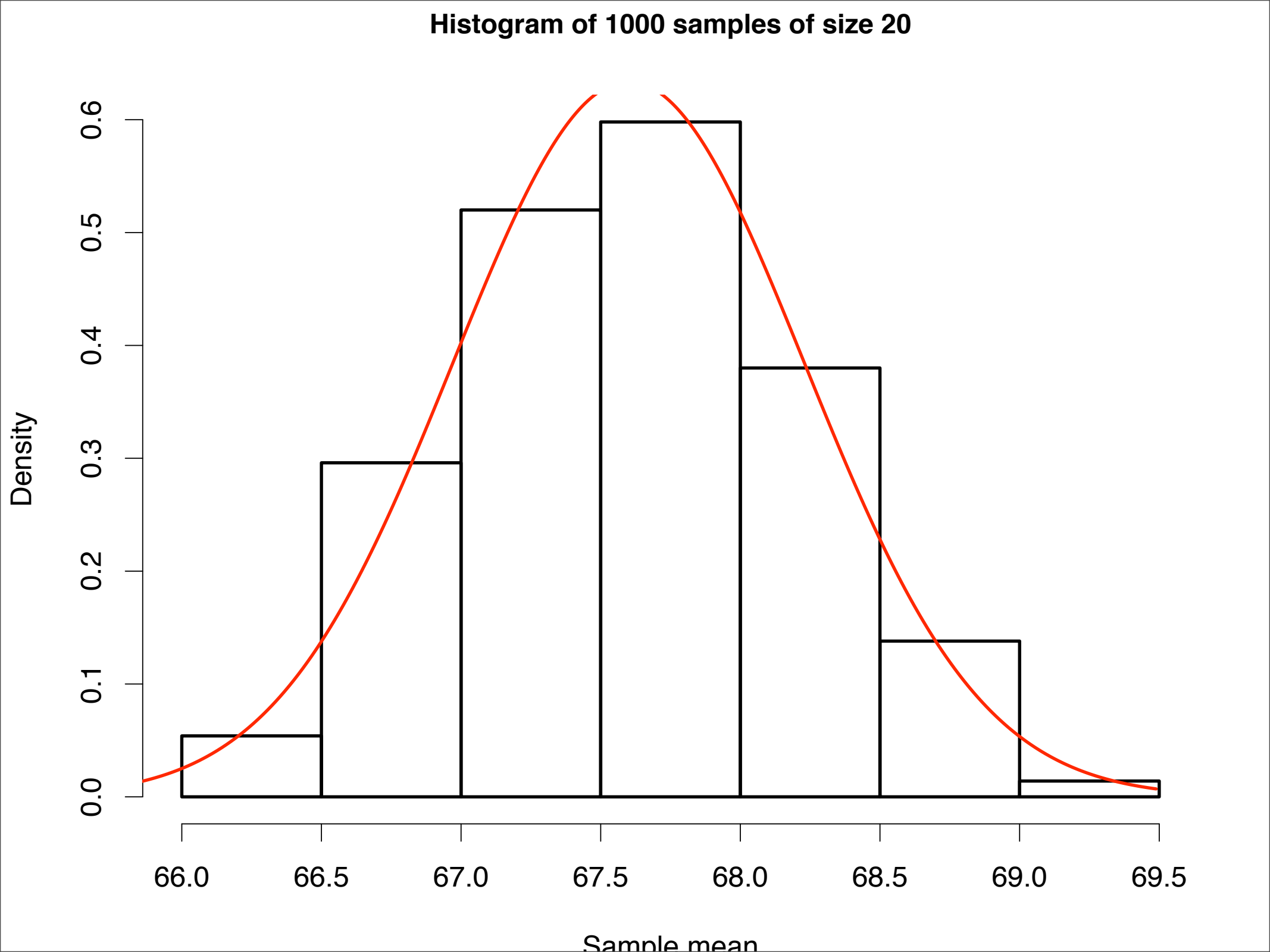

Histogram of 1000 samples of size 1
Density
0.00
0.05
0.10
0.15
60
65
70
75
80
Sample mean

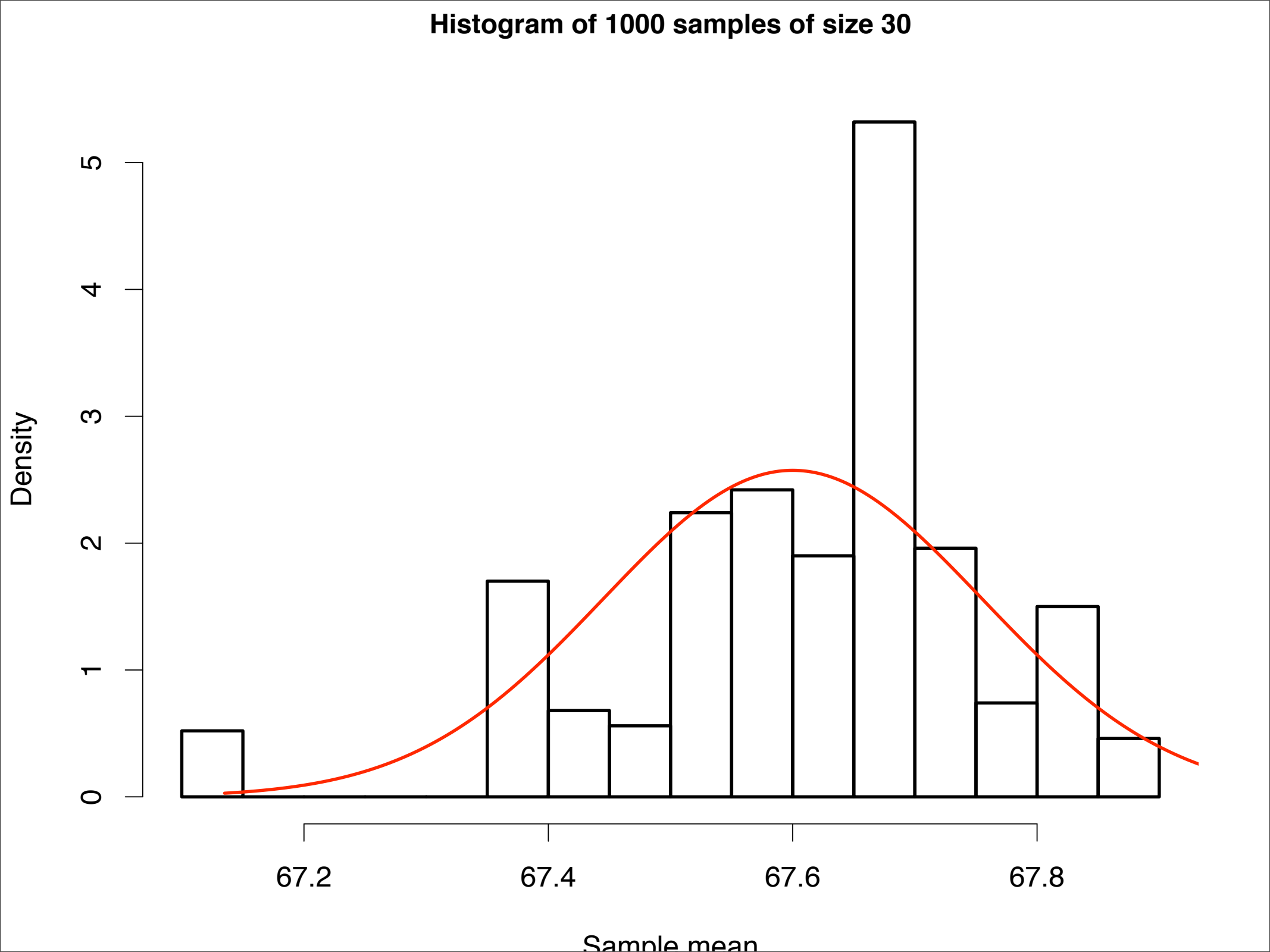

Histogram of 1000 samples of size 2
Density
0.00
0.02
0.04
0.06
0.08
0.10
0.12
60
65
70
75
80
Sample mean

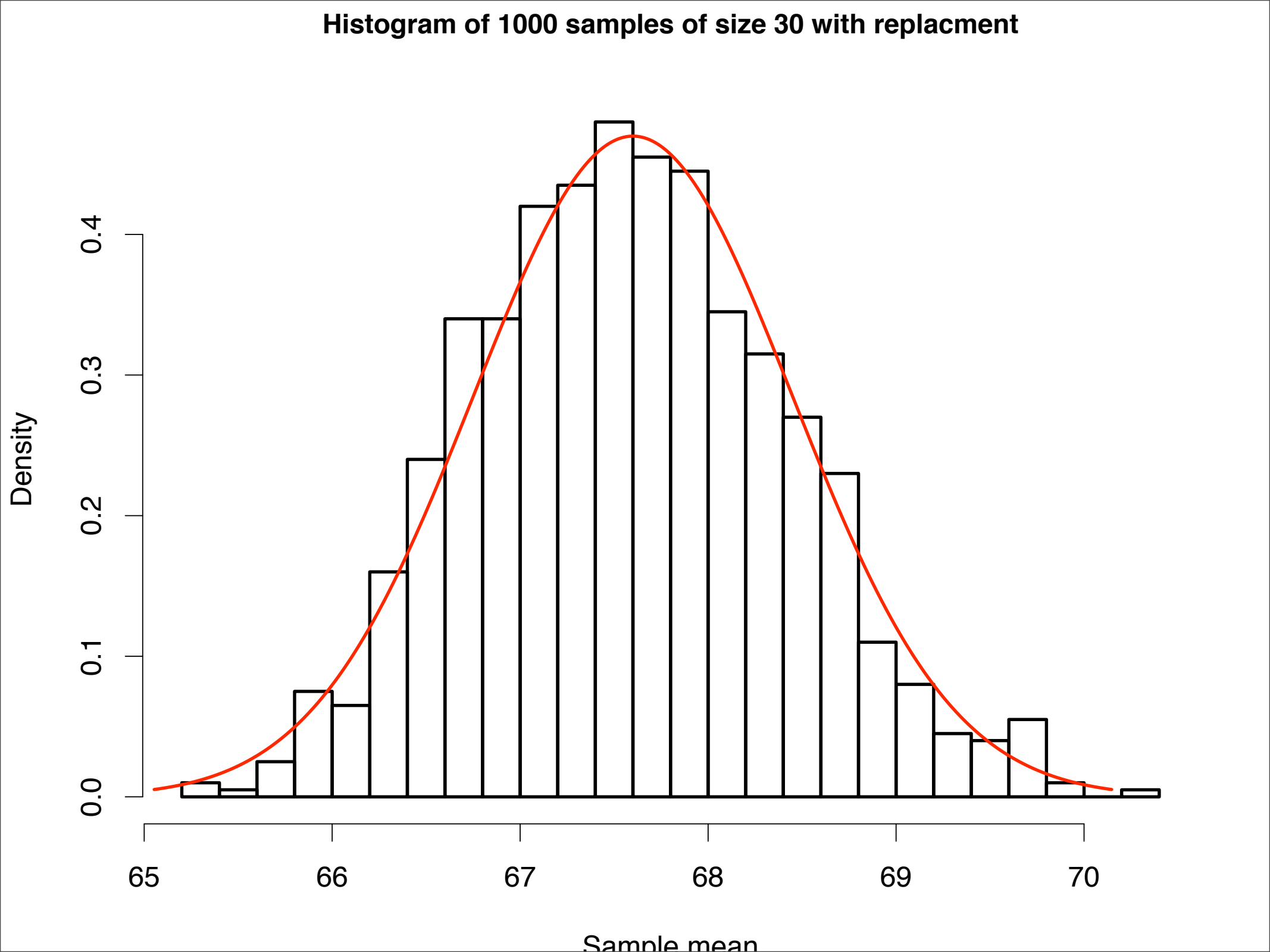

Histogram of 1000 samples of size 3
Density
0.00
0.05
0.10
0.15
65
70
75
Sample mean

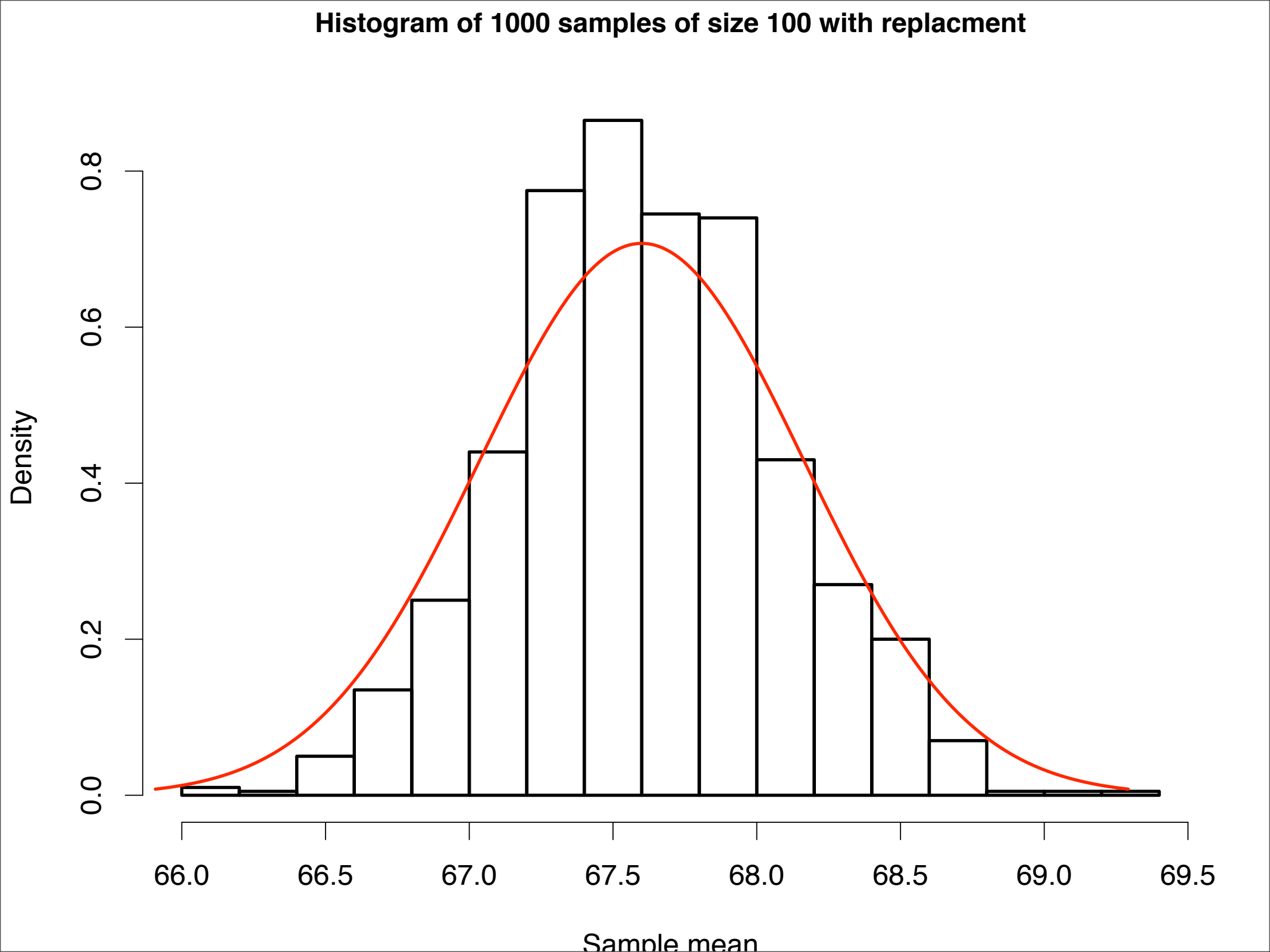

Histogram of 1000 samples of size 5
Density
0.00
0.05
0.10
0.15
0.20
62
64
66
68
70
72
74
Sample mean

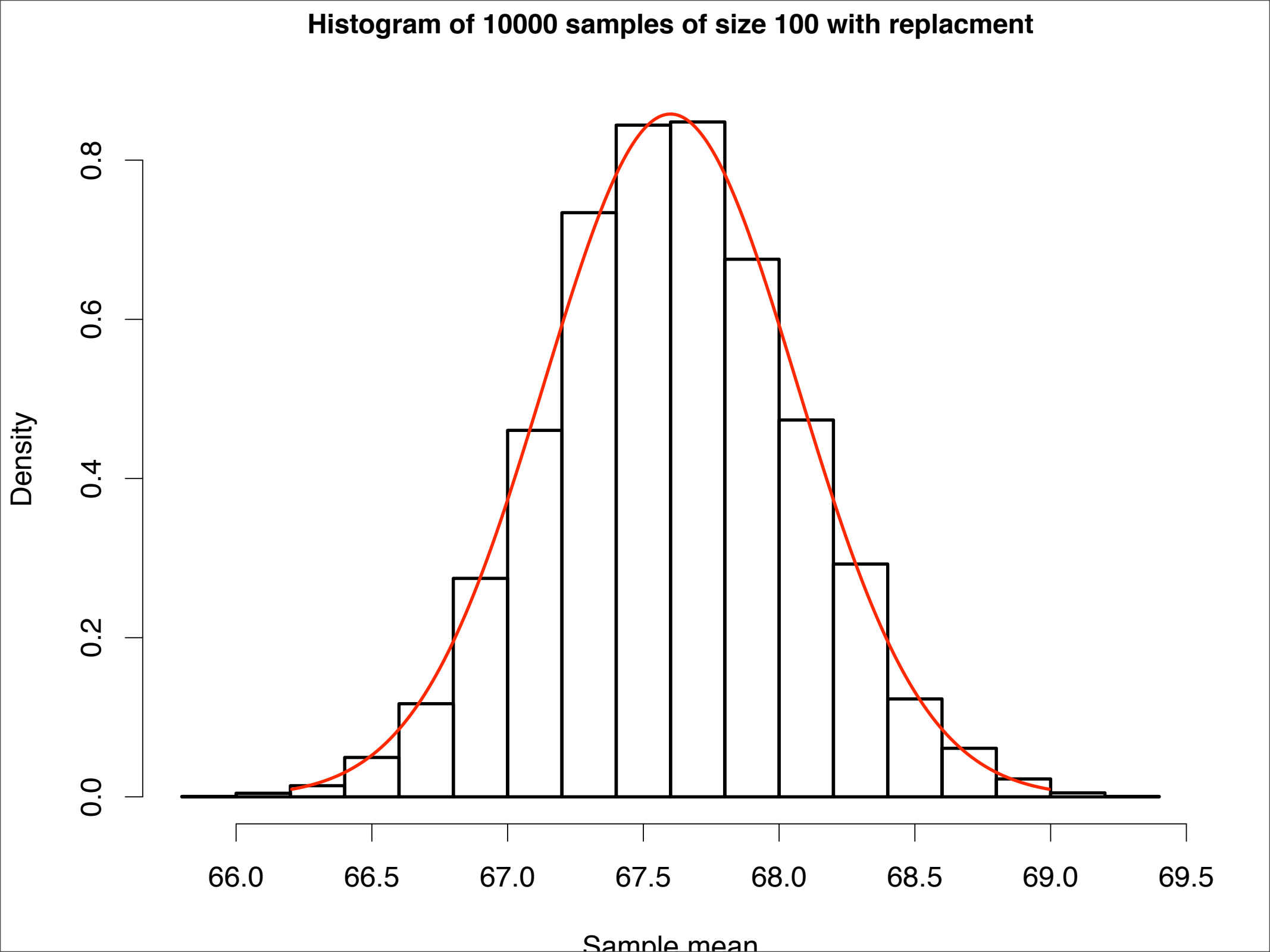

Histogram of 1000 samples of size 10
Density
0.00
0.05
0.10
0.15
0.20
0.25
0.30
64
66
68
70
Sample mean


Histogram of 1000 samples of size 20
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
66.0
66.5
67.0
67.5
68.0
68.5
69.0
69.5
Sample mean


Histogram of 1000 samples of size 30
Density
0
1
2
3
4
5
67.2
67.4
67.6
67.8
Sample mean


Histogram of 1000 samples of size 30 with replacment
Density
0.0
0.1
0.2
0.3
0.4
65
66
67
68
69
70
Sample mean


Histogram of 1000 samples of size 100 with replacment
Density
0.0
0.2
0.4
0.6
0.8
66.0
66.5
67.0
67.5
68.0
68.5
69.0
69.5
Sample mean


Histogram of 10000 samples of size 100 with replacment
Density
0.0
0.2
0.4
0.6
0.8
66.0
66.5
67.0
67.5
68.0
68.5
69.0
69.5
Sample mean

Goals

* Summarize the data
* Model the source of the data
* Use the data to infer details of the source

"How could the data have been different?"

* Were data sampled from a larger population?
* Was the population generated by a random process?

Summary

* Data are assumed to derive from random processes
* We use probability theory to make up models, compute implications
* We use statistics to infer backward from data to the particular model
* Results depend on the accuracy of the assumptions in the general model
* Averaging reduces variance