Sampling Distributions

Imagine homework assignment:

Each student in class picks 5 students at random from group who submitted surveys.

Each student asked to compute for the 5 students they pick:

Average height, average weight, SD height, SD weight, r.

Number forms 1 to 69 (number turned in).

Number students doing homework from 1 to 98.

Next pages refer to group of 331 students doing homework; 151 forms completed.

Beginning of results:

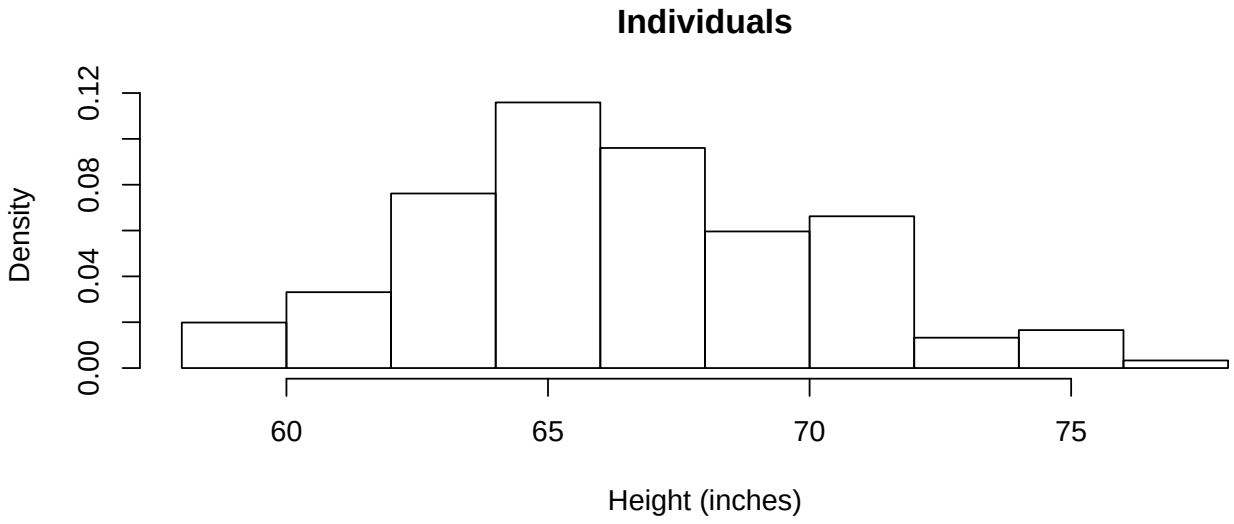
	Form #				
Student	1	2	3	4	5
1	24	27	87	71	85
2	74	16	88	150	138
3	63	19	8	53	45
⋮			⋮		

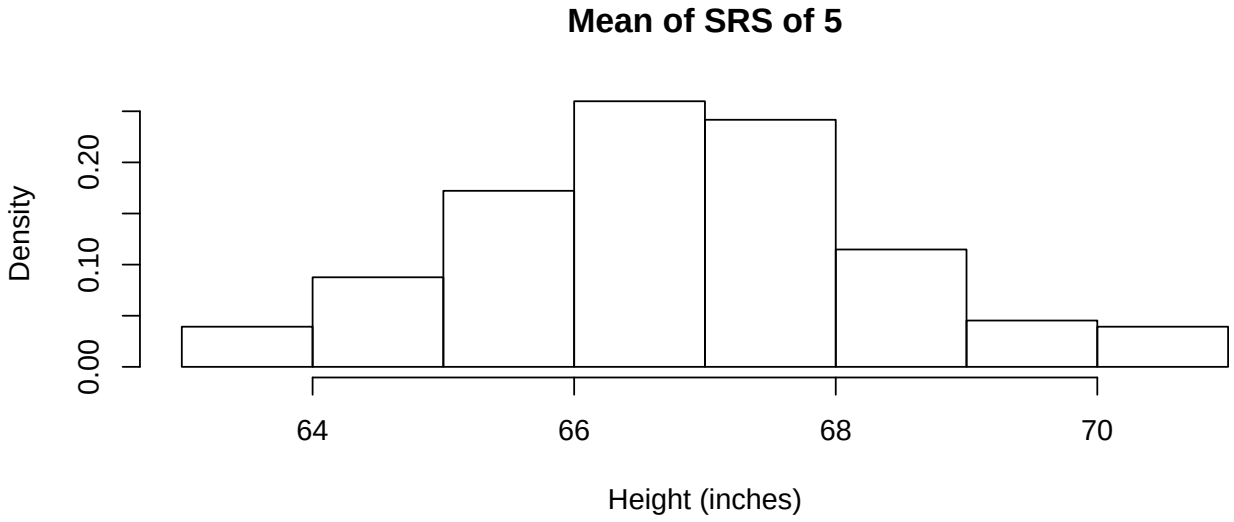
Now for each student get the forms

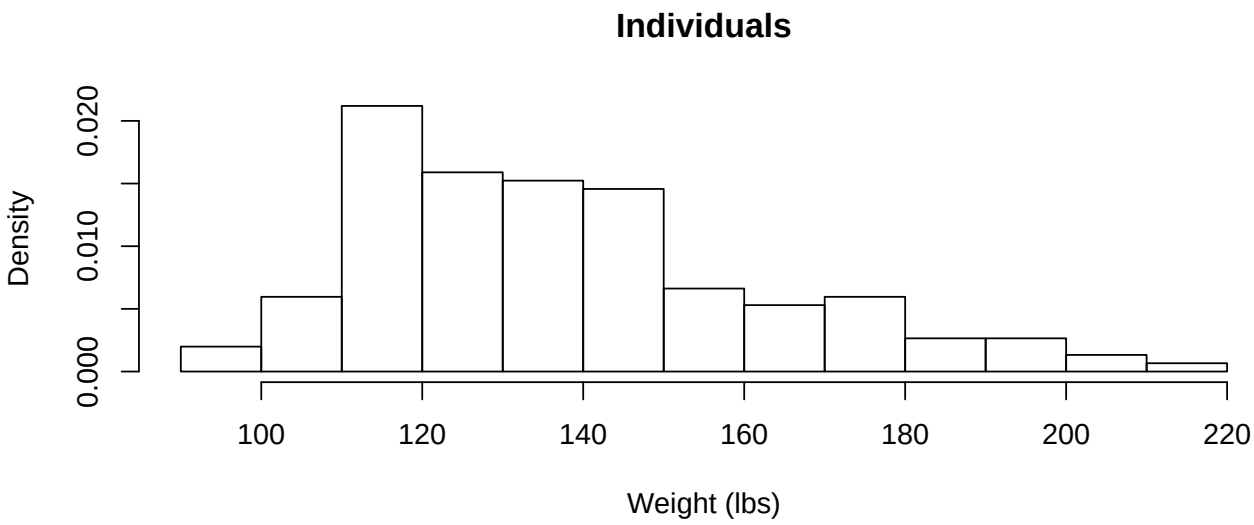
Get corresponding heights, weights and work out summary statistics:

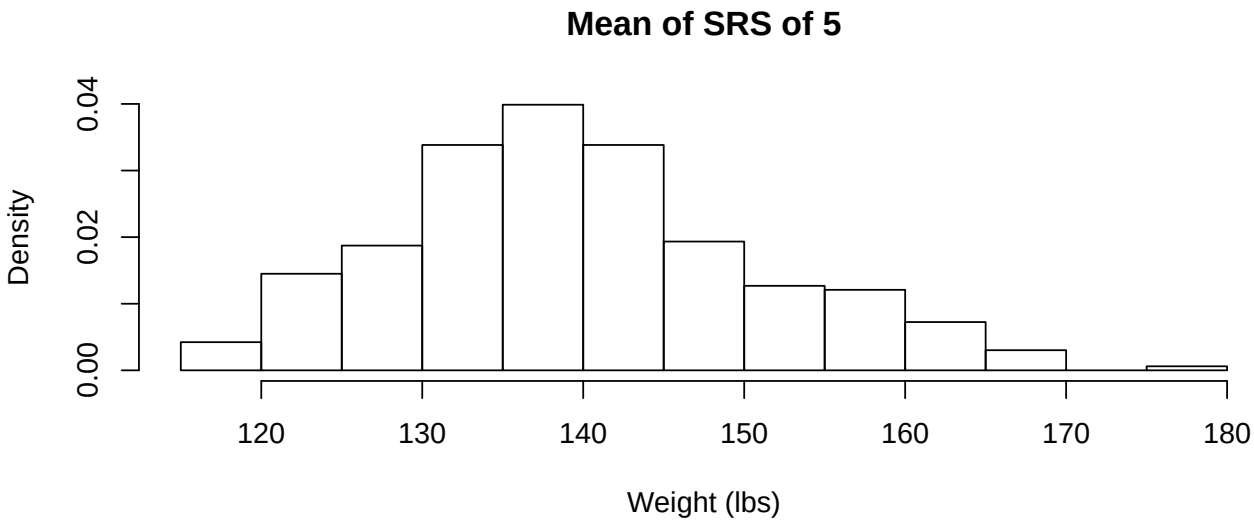
Stdnt	$\bar{H}$	s_H	$\bar{W}$	s_W	r
1	65.11	3.84	135.95	21.44	-0.06
2	67.17	3.67	131.58	16.56	0.84
3	64.40	2.61	118.00	9.75	0.53
4	69.05	4.64	153.19	27.47	0.91
5	64.78	3.87	143.96	27.05	0.32
⋮			⋮		

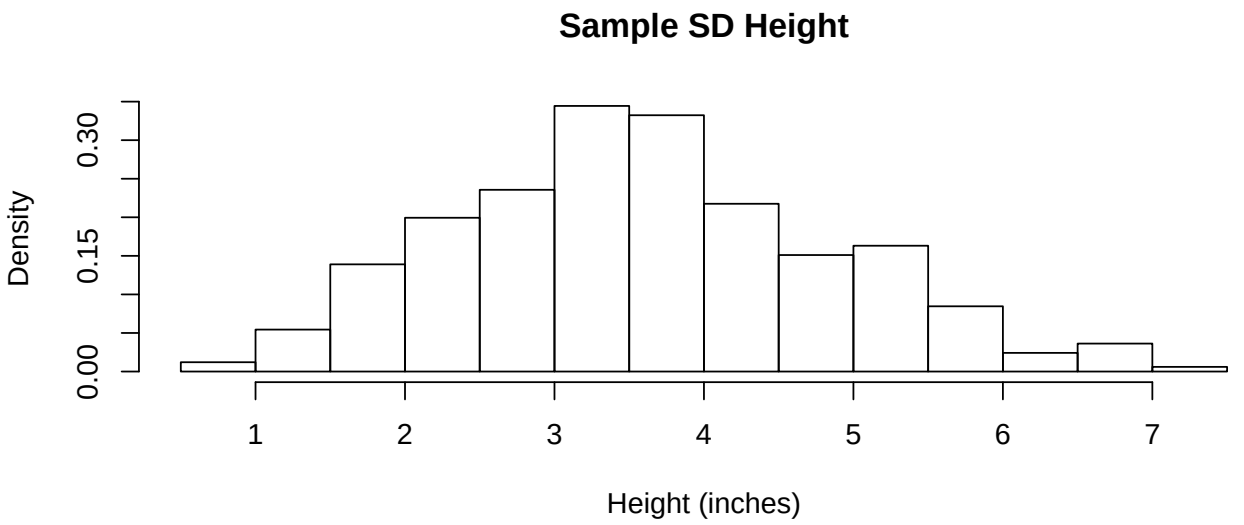
Look at histograms of the different sample means, etc.

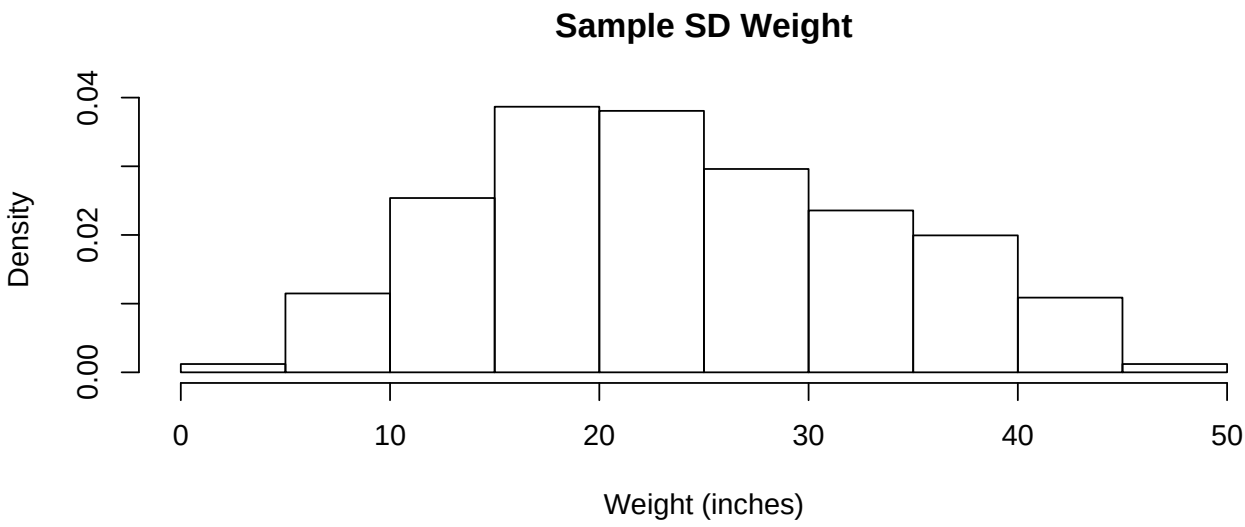


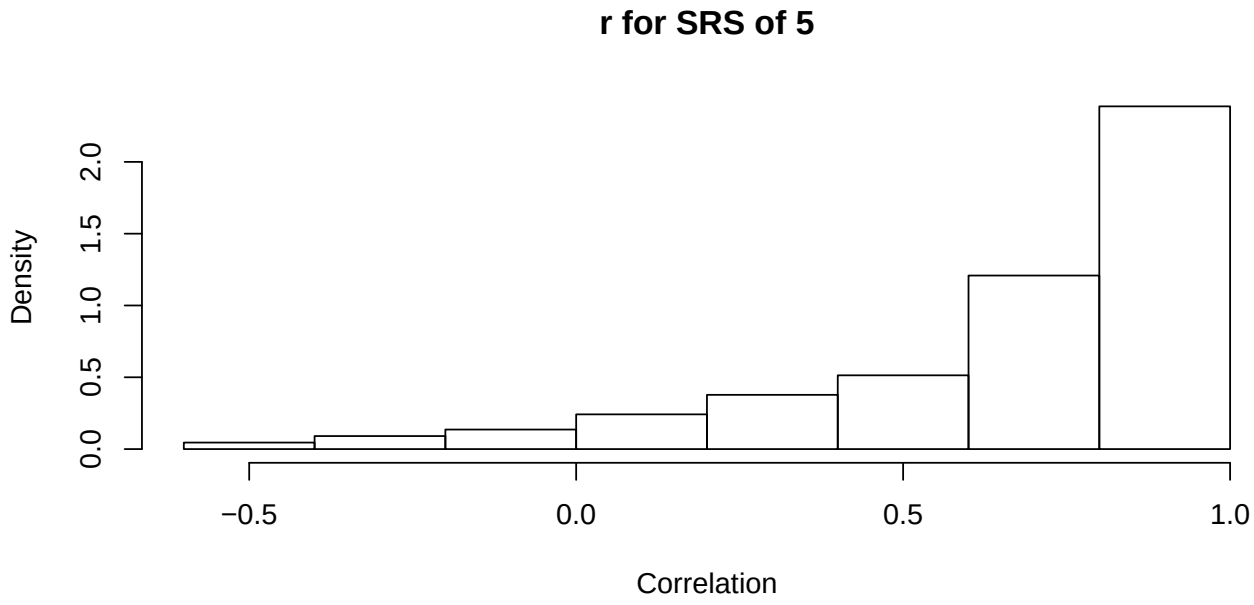



Individuals
Density
0.000
0.010
0.020
100
120
140
160
180
200
220
Weight (lbs)


Mean of SRS of 5
Density
0.00
0.02
0.04
120
130
140
150
160
170
180
Weight (lbs)


Sample SD Height
Density
0.00
0.15
0.30
1
2
3
4
5
6
7
Height (inches)


Sample SD Weight
Density
0.00
0.02
0.04
0
10
20
30
40
50
Weight (inches)



Commentary: if we had taken not 331 samples of size 5 but all possible samples of size 5 resulting histogram would have been:

The **Sampling Distribution** of the corresponding **Statistic**.

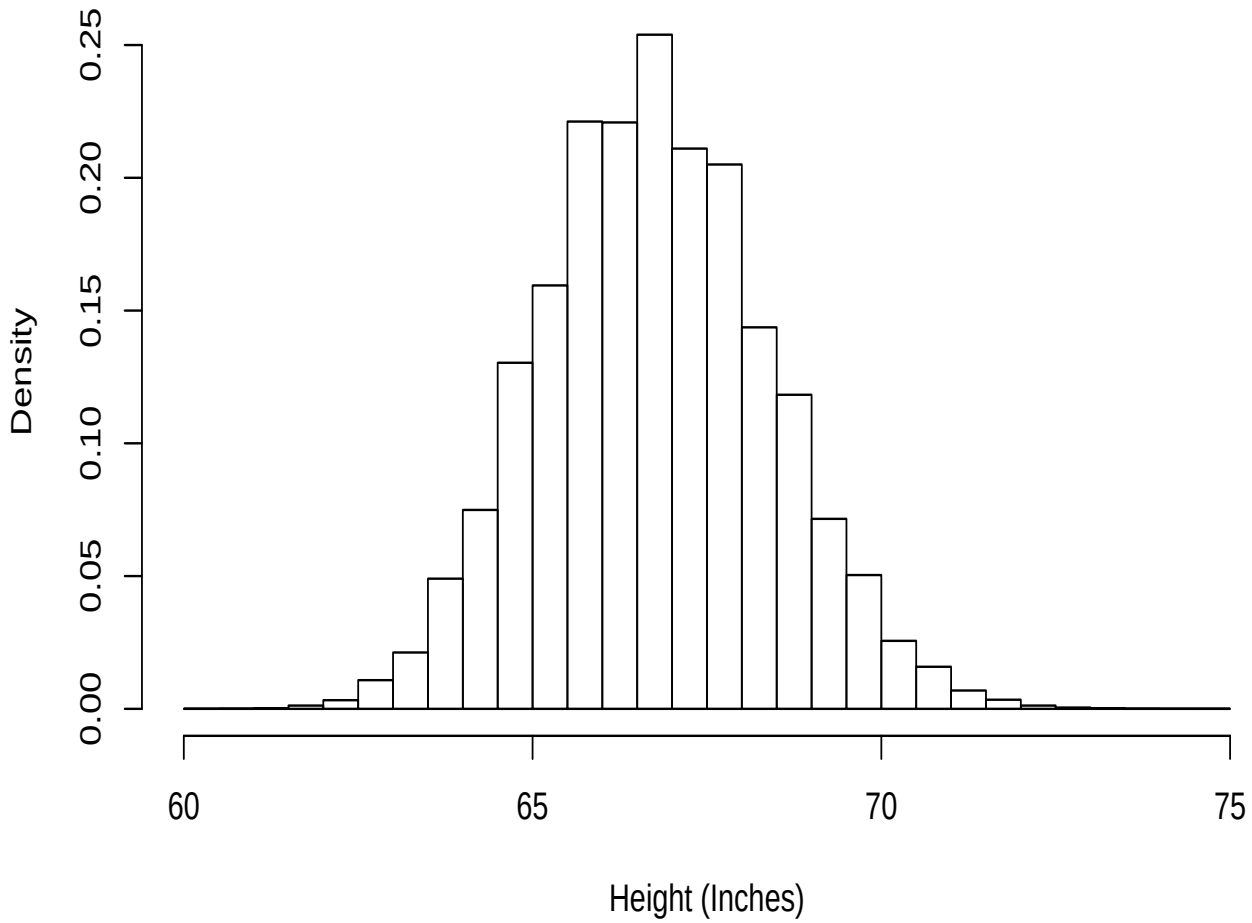
Jargon: a **Statistic** is a numerical summary of the sample or the data.

A **parameter** is a numerical summary of the population.

Too hard to take all samples of size 5.

Take 1,000,000 samples of size 5.

Histogram of sample mean for Simple Random Sample of 5 heights from 151.



Statistical Distribution Theory

Two kinds: **exact** and **approximate**.

Examples of exact theory.

1) If the population distribution is normal then the sampling distribution of the sample mean is normal.

2) If the population distribution has mean μ and standard deviation σ then the sampling distribution of $\bar{X}$ has mean μ and standard deviation $\sigma/\sqrt{n}$.

Jargon: we call the SD of the sample mean the **Standard Error** (SE) of the sample mean.

Very important point: To cut SE in half: need 4 times more data!

3) If the population distribution is normal then the sampling distribution of

$$t \equiv \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

(called a t-statistic) is called "Student's t on $n - 1$ degrees of freedom.

Approximate Distribution Theory

1) If the sample size n is large enough then the sample mean, $\bar{X}$ has a sampling distribution which is approximately normal.

This is the **Central Limit Theorem**.

Needed assumptions:

1) Infinite population.

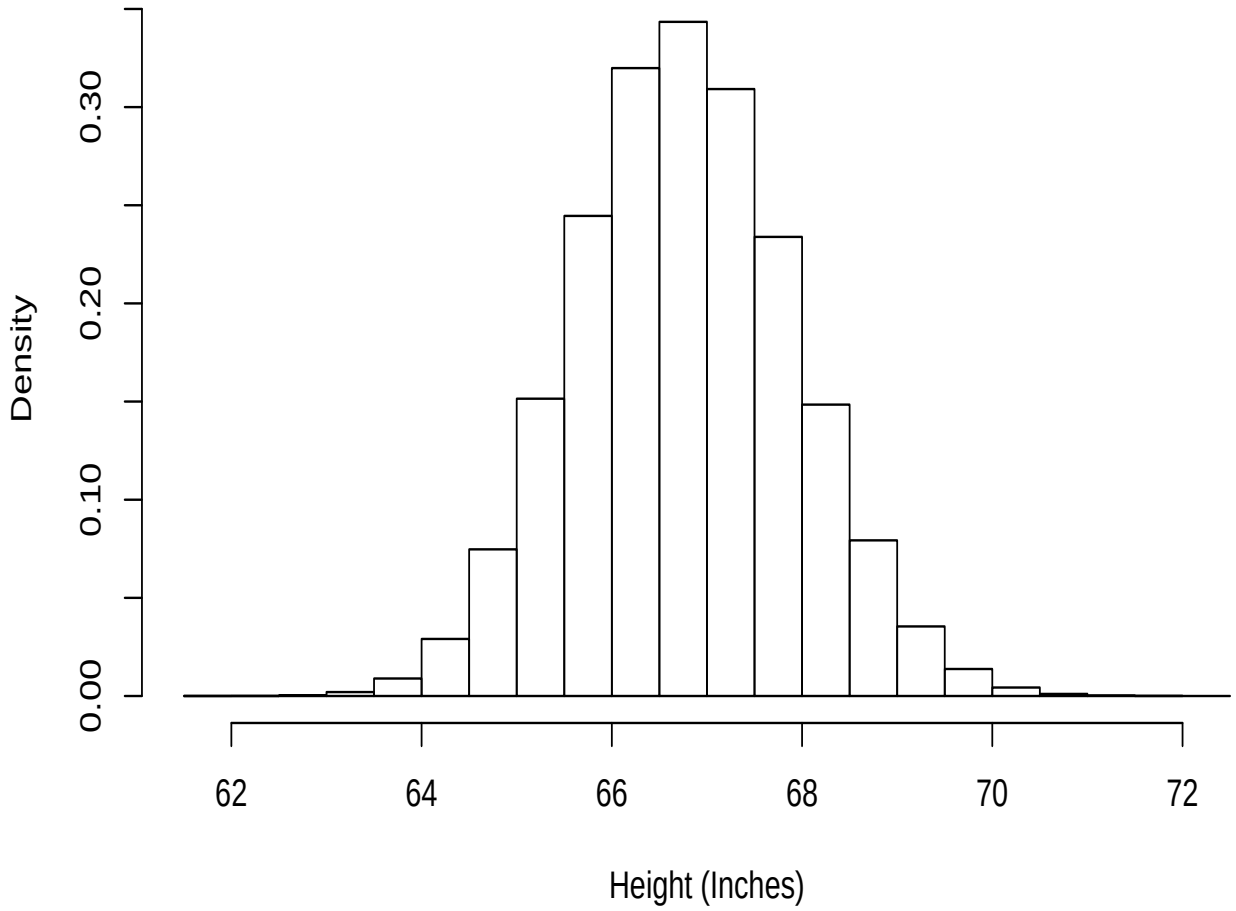
OR

2) Simple random sampling with replacement.

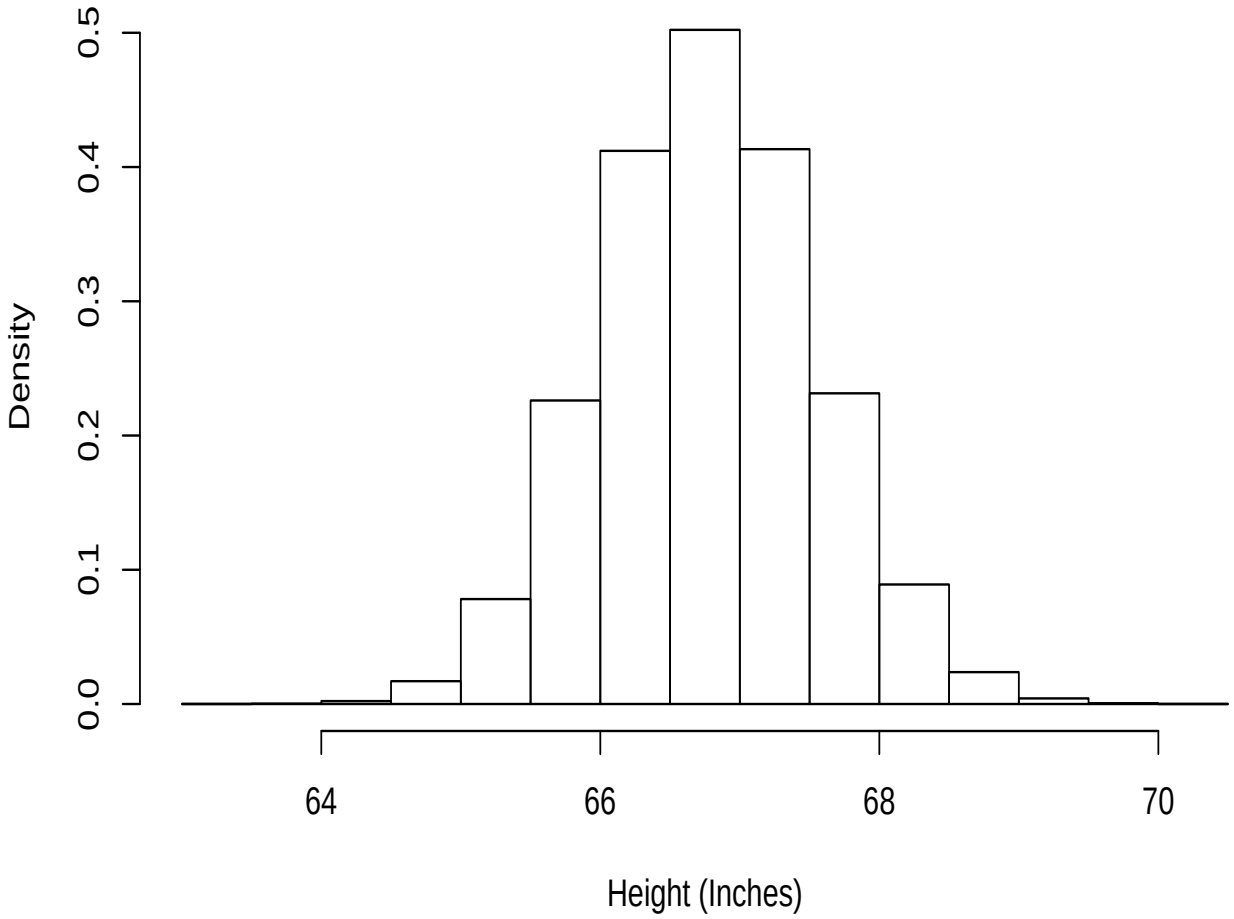
OR

3) SRS without replacement from pop of N individuals and n is tiny compared to N.

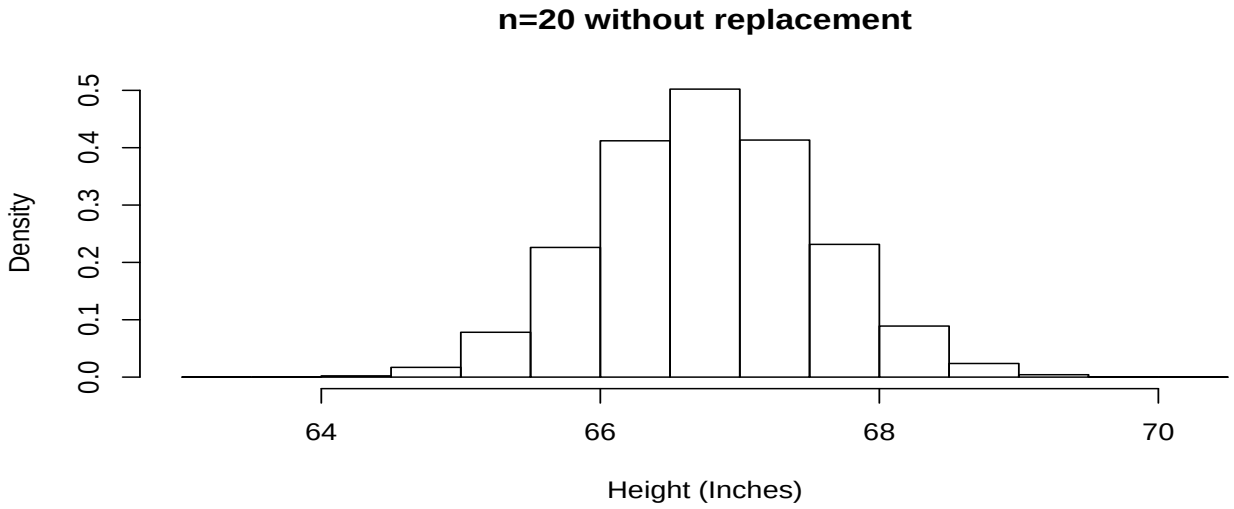
Histogram of sample mean for Simple Random Sample of 10 heights from 151.

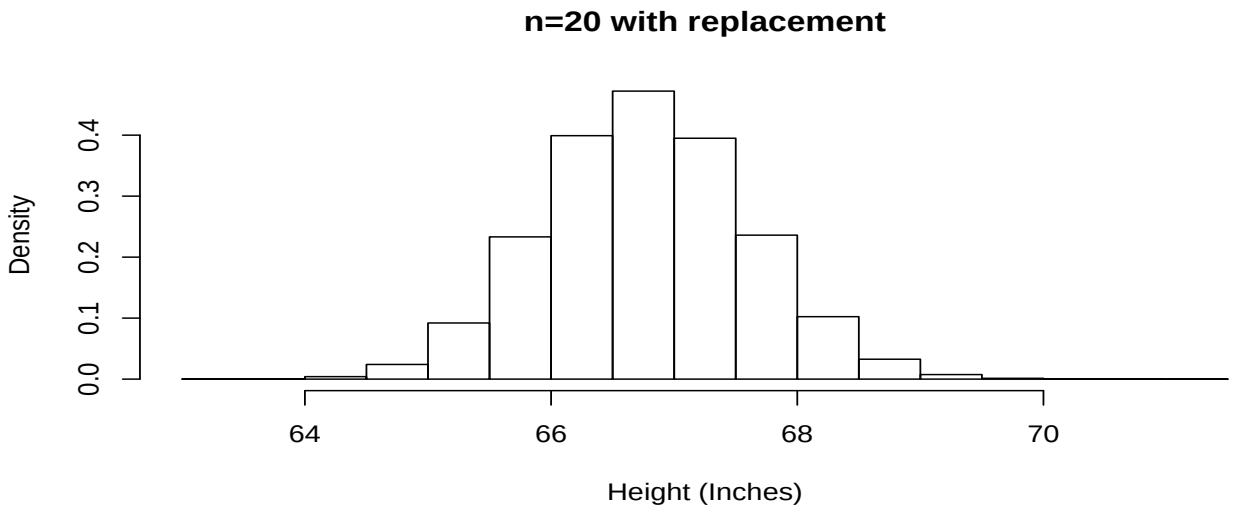


Histogram of sample mean for Simple Random Sample of 20 heights from 151.

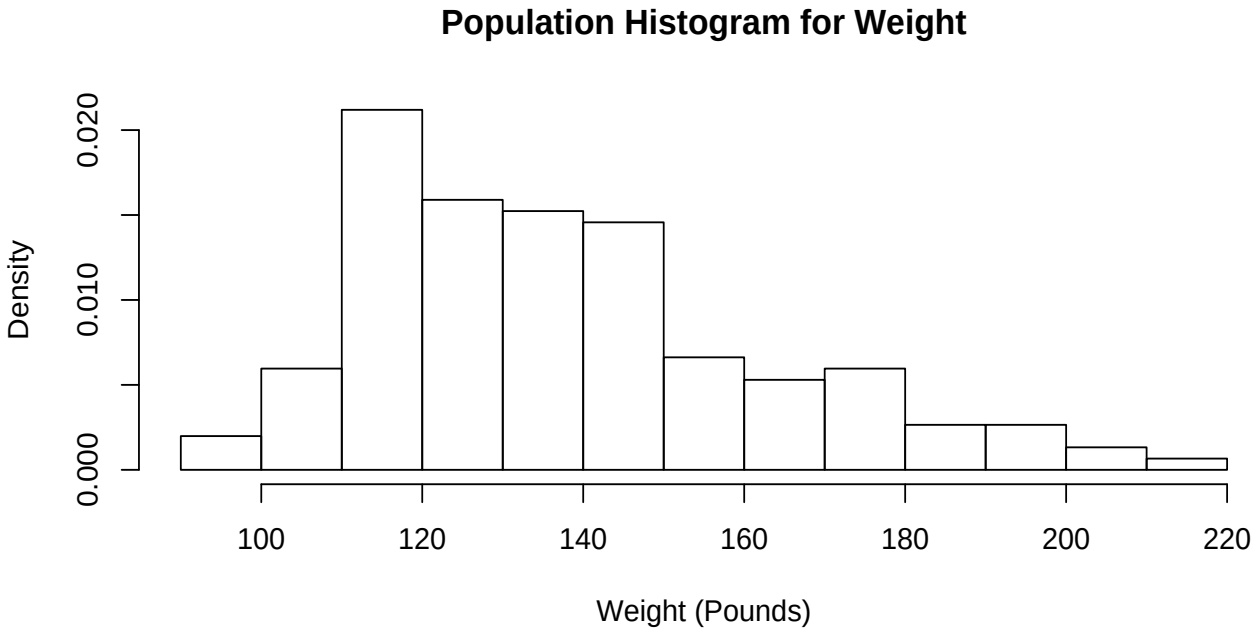


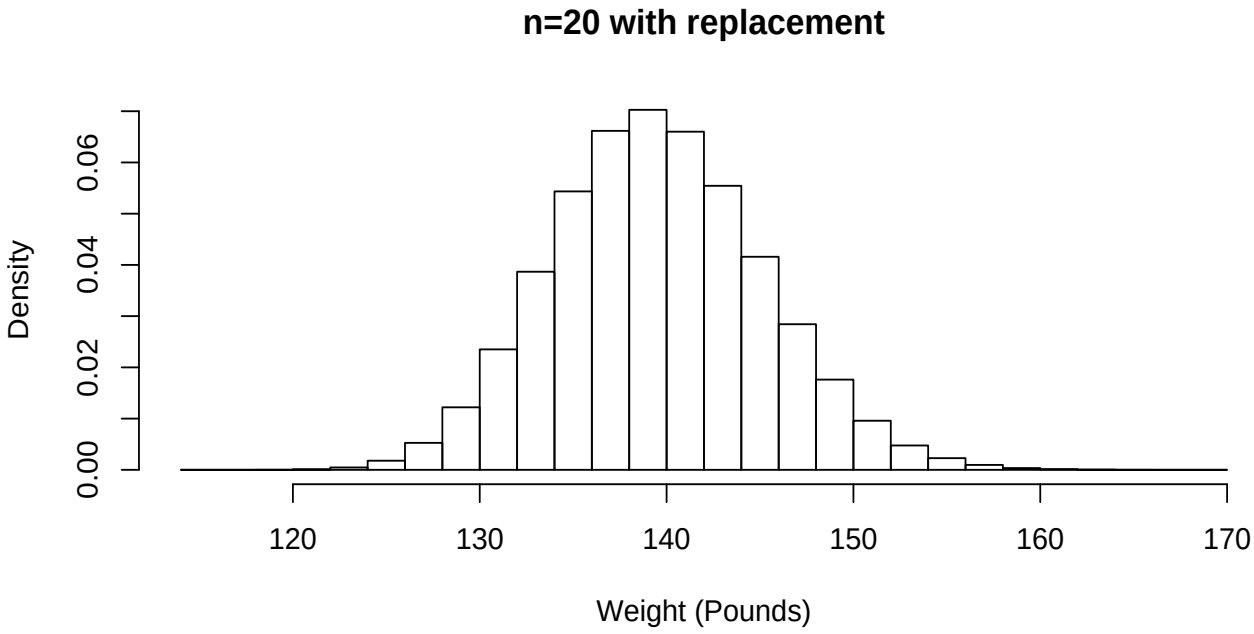
Histogram of sample mean for Simple Random Sample **with replacement** of 20 heights from 151.





Weights: $n = 20$, with replacement.





Using the Central Limit Theorem.

Draw sample of $n = 20$, with replacement, from 151 student heights.

Work out $\bar{H}$, the sample mean.

Compute chance that sample mean is in the range 65.1 to 69.4 inches.

Steps:

1) Compute mean and standard deviation of $\bar{H}$,

a) the mean of $\bar{H}$ is the same as the mean of the population heights, $\mu = 66.77$ inches.

b) the SD of $\bar{H}$ is

$$\frac{\sigma}{\sqrt{n}} = \frac{3.73}{\sqrt{20}} = 0.879.$$

2) Convert range to standard deviation units:

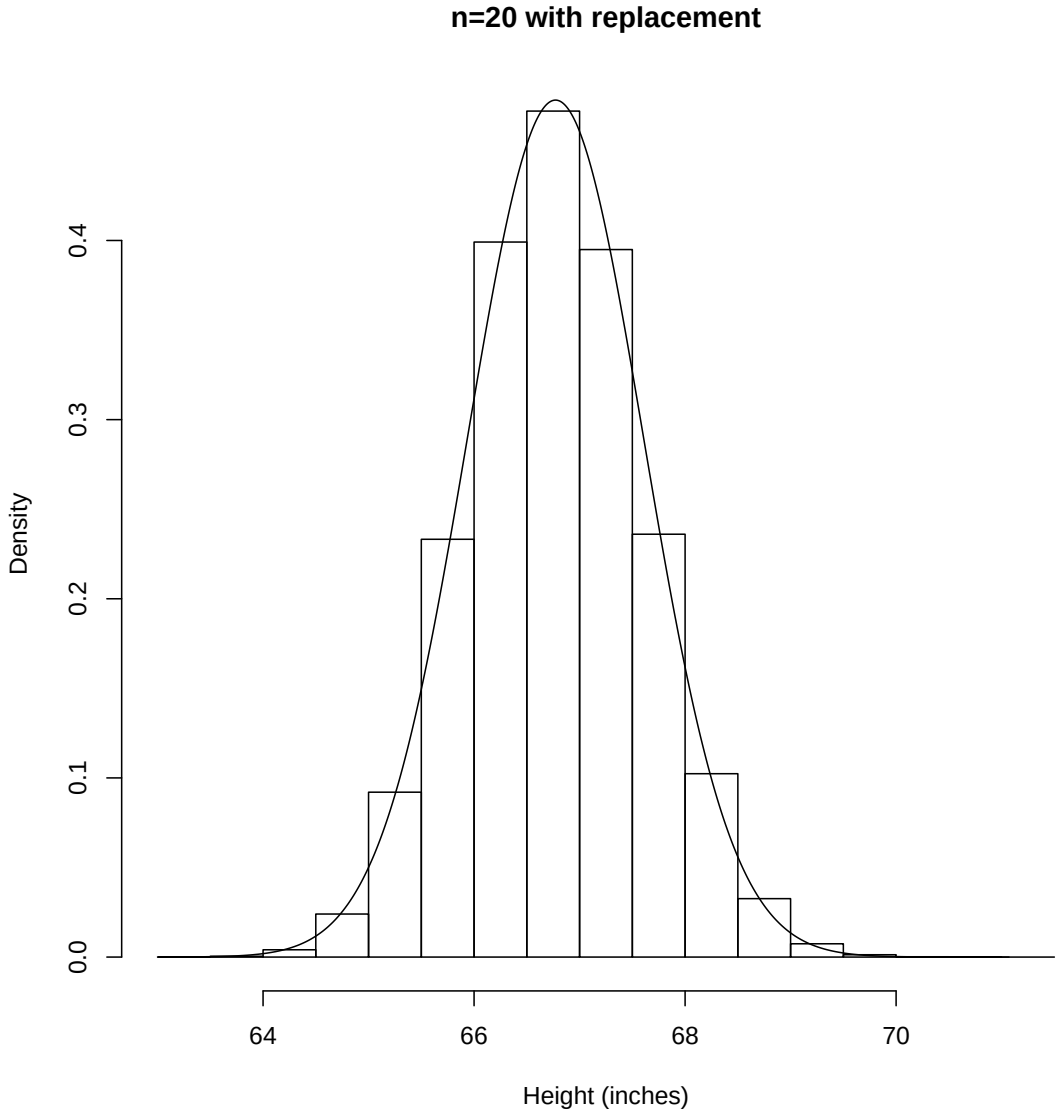
$$\frac{65.1 - 66.77}{0.879} \text{ to } \frac{69.4 - 66.77}{0.879}$$

which works out to -2.00 to 2.99.

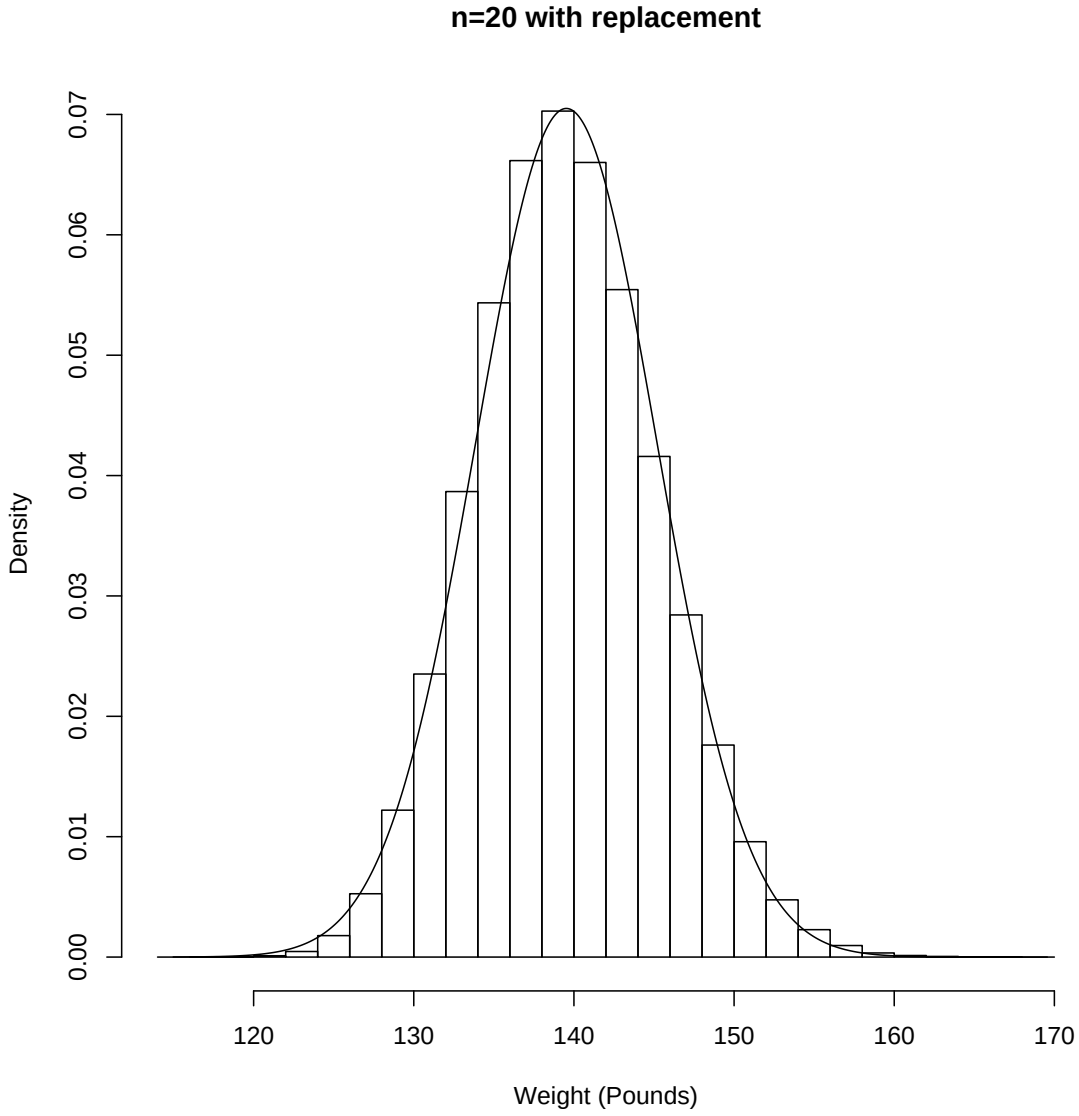
3) Find the corresponding area under the normal curve:

$$0.9986 - 0.0228 = 0.9768 \approx 97.7\%$$

Some graphical evidence. Heights, $n = 20$, with replacement.



Some graphical evidence. Weights, $n = 20$, with replacement.



Comments:

1) Histogram of individual heights was more normal than histogram of individual weights.

2) So normal approx better for heights than weights.

3) Actual fraction of $\bar{H}$ values in range 65.1 to 69.4 is 97.92%; pretty good approximation.