Probability and Mathematical Statistics in Data Science

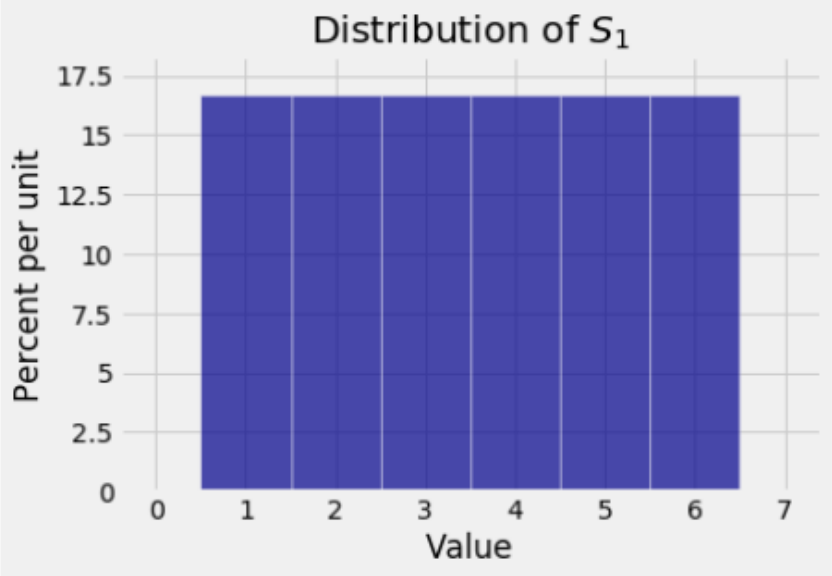
Lecture 20: Section 8.1: Distribution of the Sample Sum (and the Sample Mean)

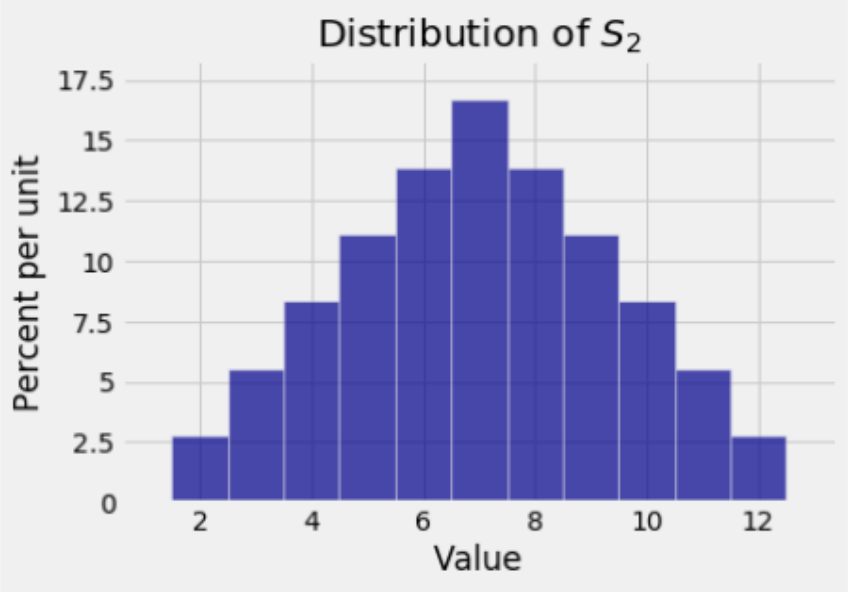
Normal Approximation to the Binomial

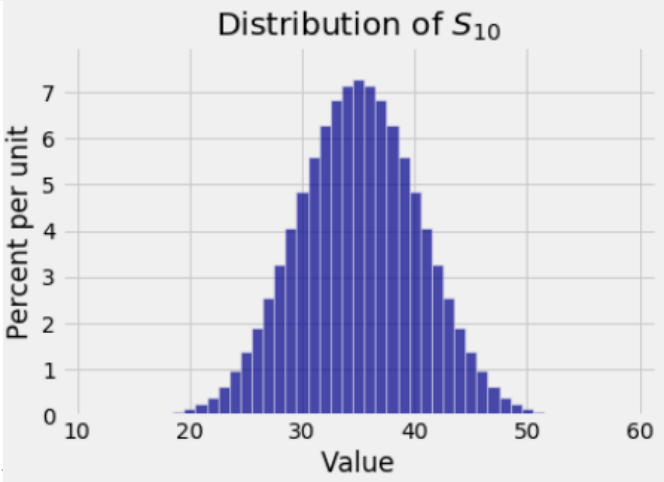
- If X1,X2,…,Xn are i.i.d. Bernoulli (p)random variables then Sn has the binomial (n,p) distribution.

- As n increases, the Binomial distribution approaches the Normal distribution

Uniform Distribution







Sums of random variables

- Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with mean μ and variance σ^2. Define S_n to be their sum:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

- We already know that $E(S_n) = \sum E(X_k) = n\mu$.

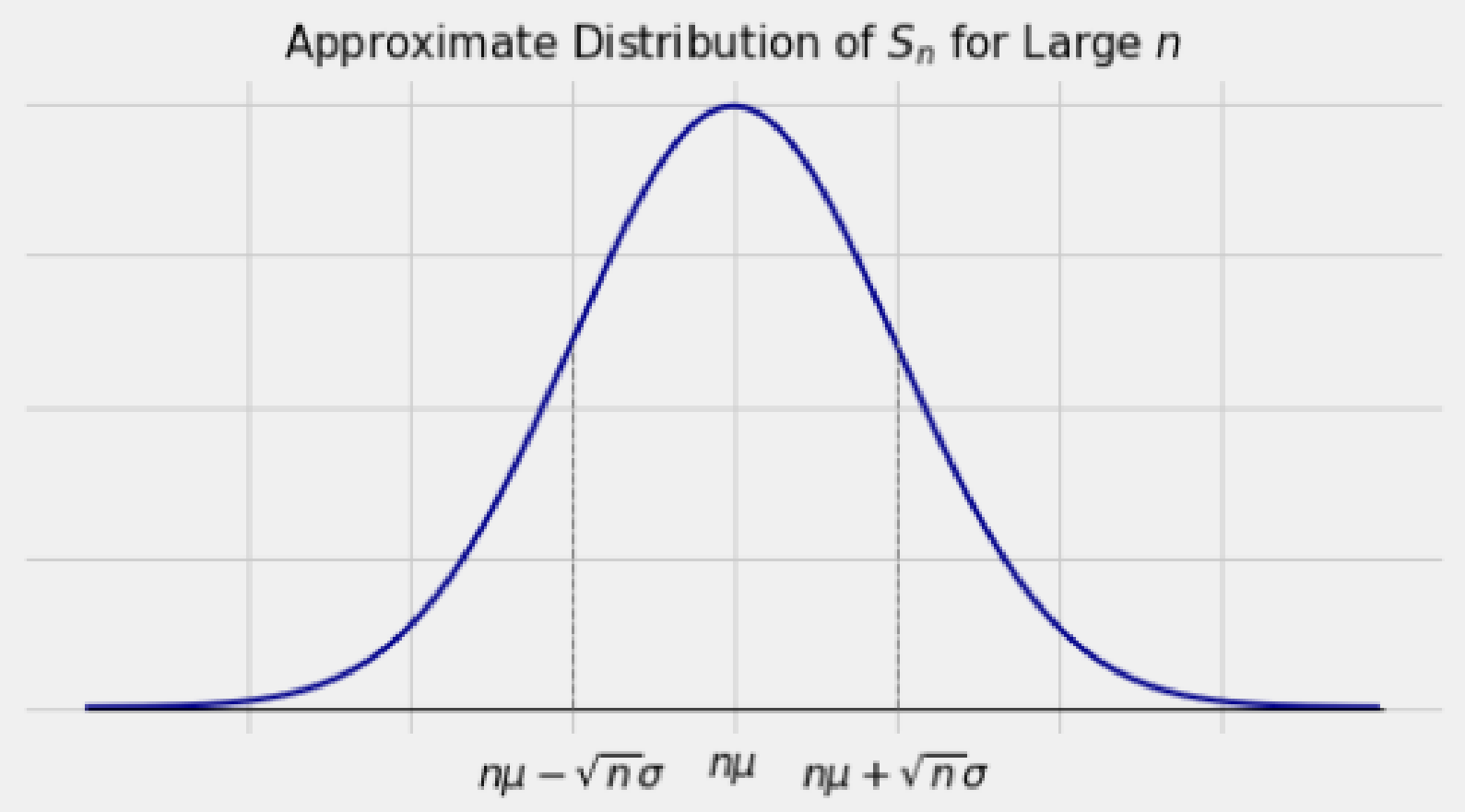
- Now we can further say that:

$$Var(S_n) = Var(X_1 + X_2 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2$$

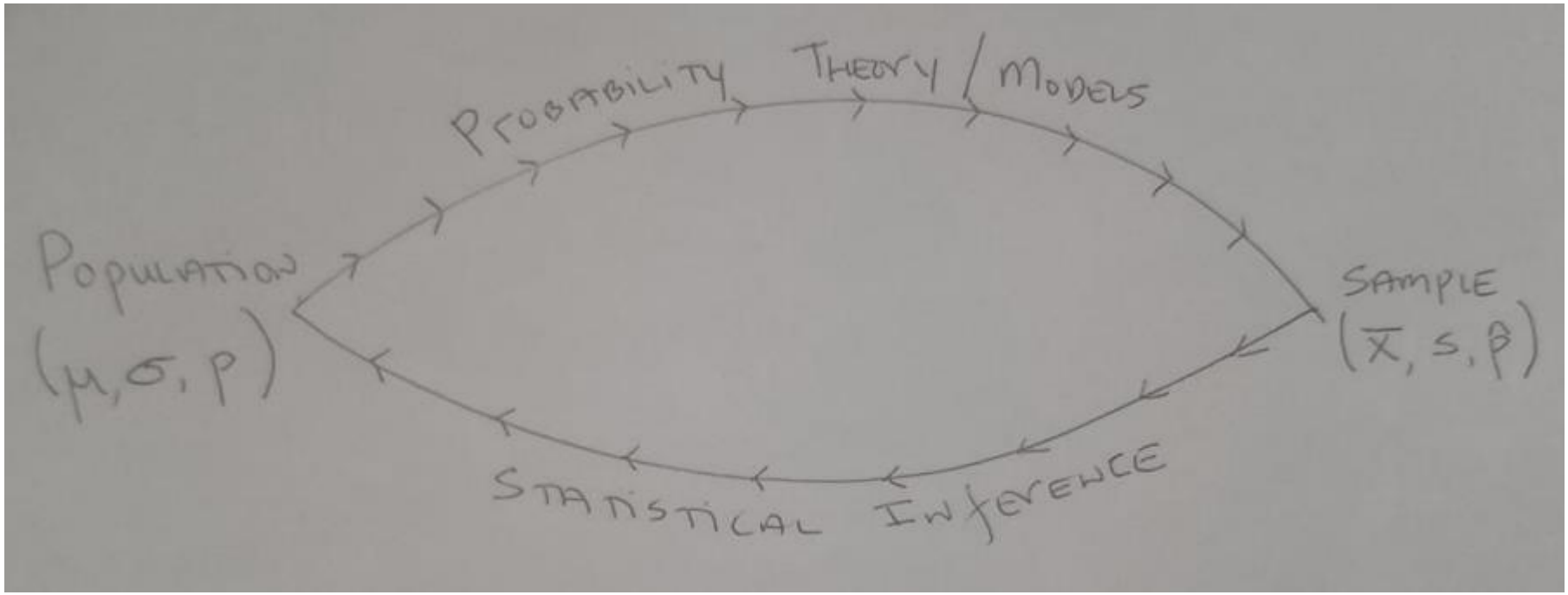
$$SD(S_n) = \sqrt{n}\,\sigma$$

- Notice that the expected value grows as n, but the sd grows as $\sqrt{n}$.

Central Limit Theorem



Population and Sample



The Distribution of the Sample Mean

- Let $X_1, X_2, \ldots, X_n$ be an IID sequence of random variables from a distribution with mean μ and variance σ^2. Sample Mean is $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.
- Then we can derive the expectation and variance of sample mean $\bar{X}$

$$E(\bar{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \mu$$

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Sample Mean of IID Normal

If $X_1, X_2, \ldots, X_n$ IID $\sim N(\mu, \sigma^2)$, then what is the distribution of $\bar{X}$?

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

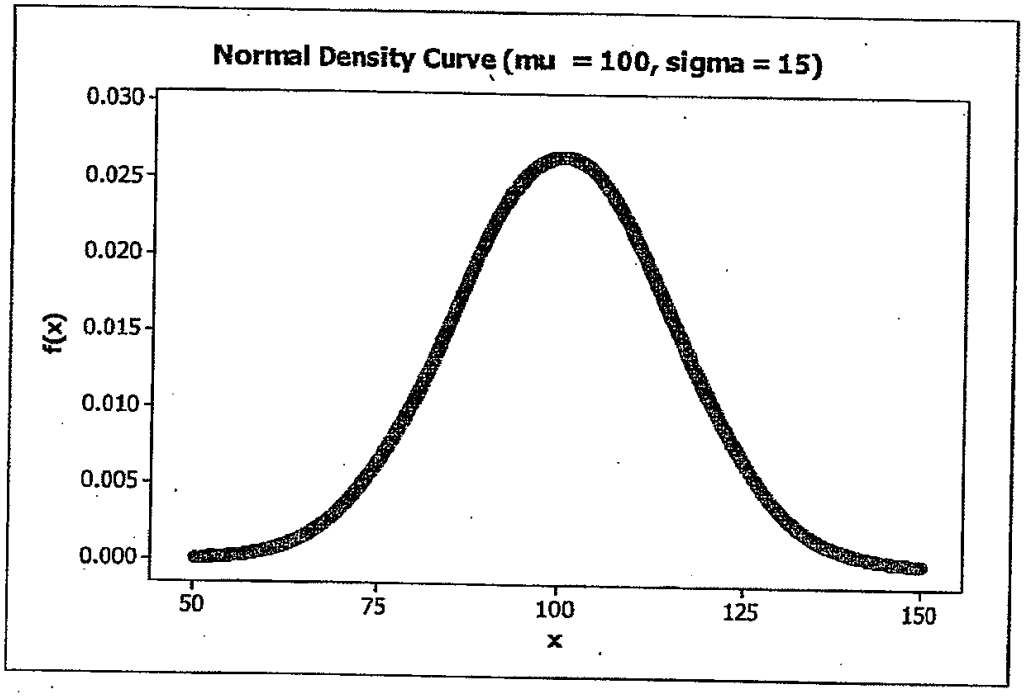
We can use the Central Limit Theorem (CLT) to derive this result.

We can also use the CLT to derive the distribution of Sample Mean when the sample is from a population that is not normal?

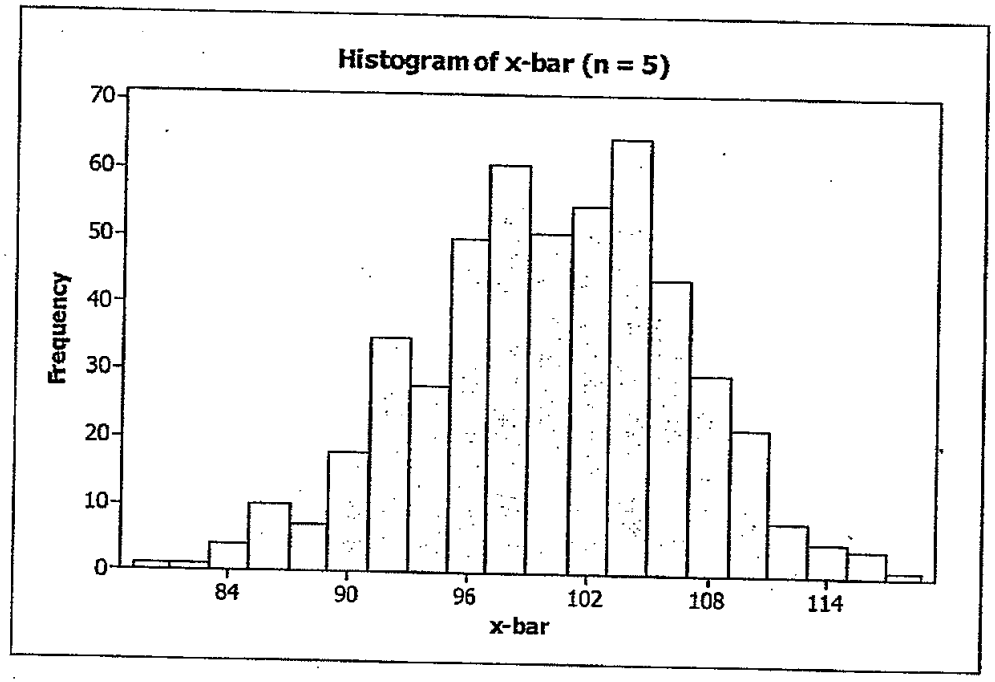
What to Expect of Sample Means : Population of measurements is bell-shaped, and a random sample is measured.

Examples of Simulation Experiments

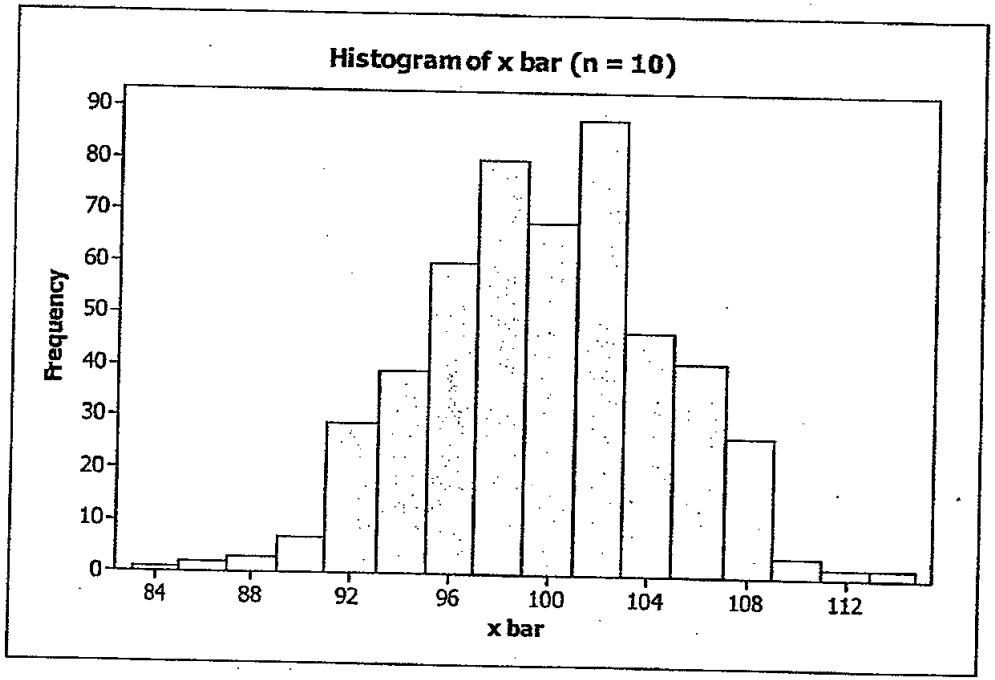
1. Consider as a population distribution the normal distribution with $\mu = 100$ and $\sigma = 15$ (the values of μ and σ are not important; the key feature of this example is a normal population distribution).

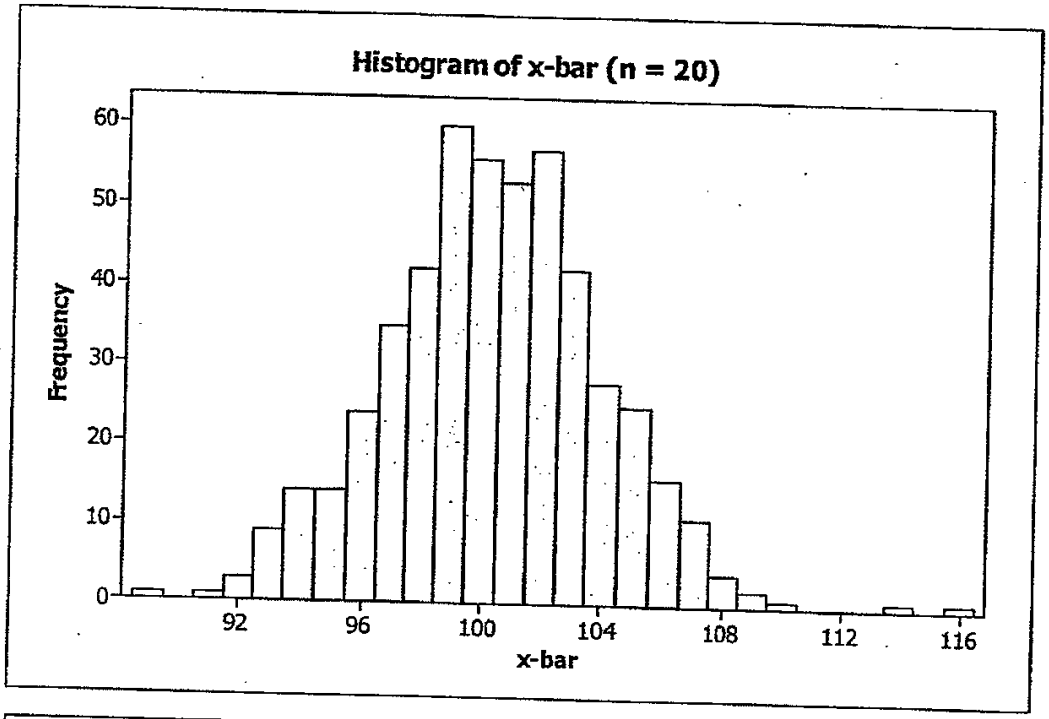


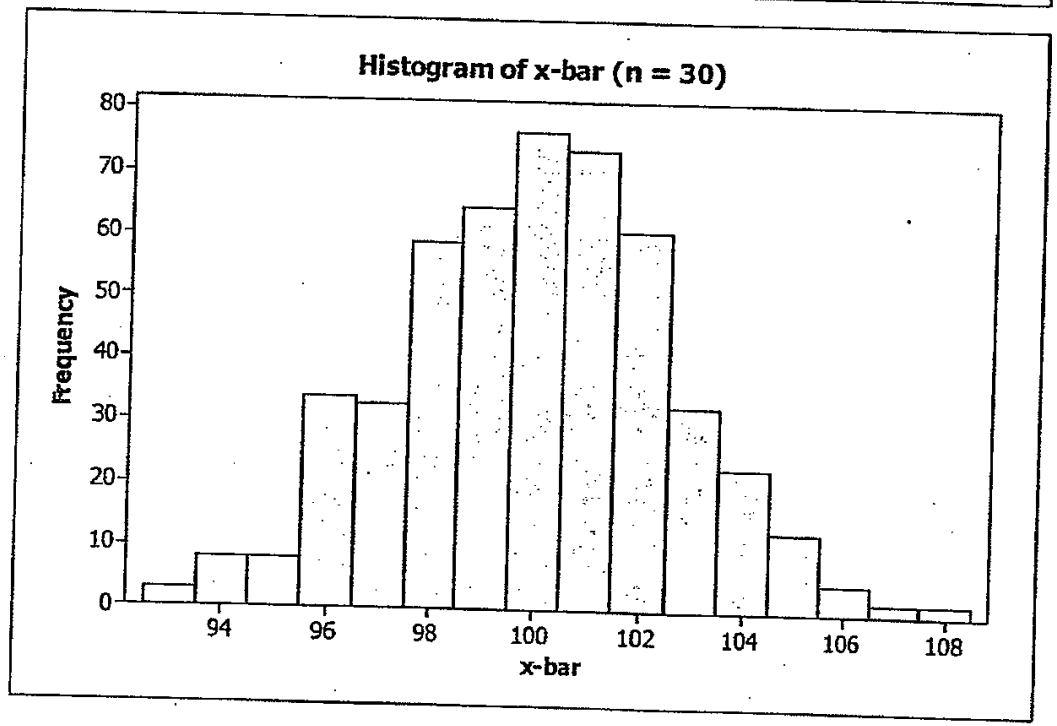
I first asked MINITAB to generate 500 samples, each of size n = 5, from this normal distribution (500 replications), and calculate the value of $\bar{x}$ for each sample. Here are the 500 resulting $\bar{x}$ values:

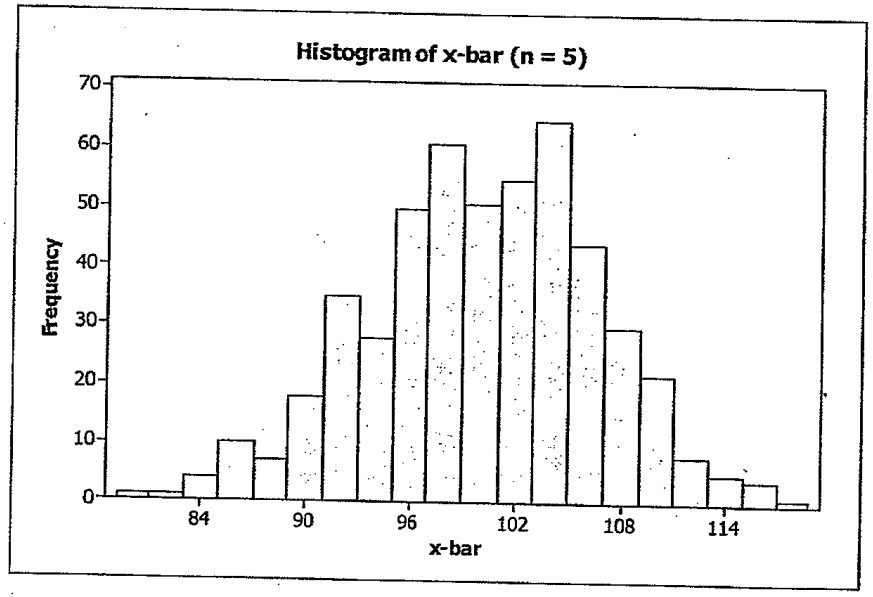

Histogram of x-bar (n = 5)
Frequency
0
10
20
30
40
50
60
70
84
90
96
102
108
114
x-bar

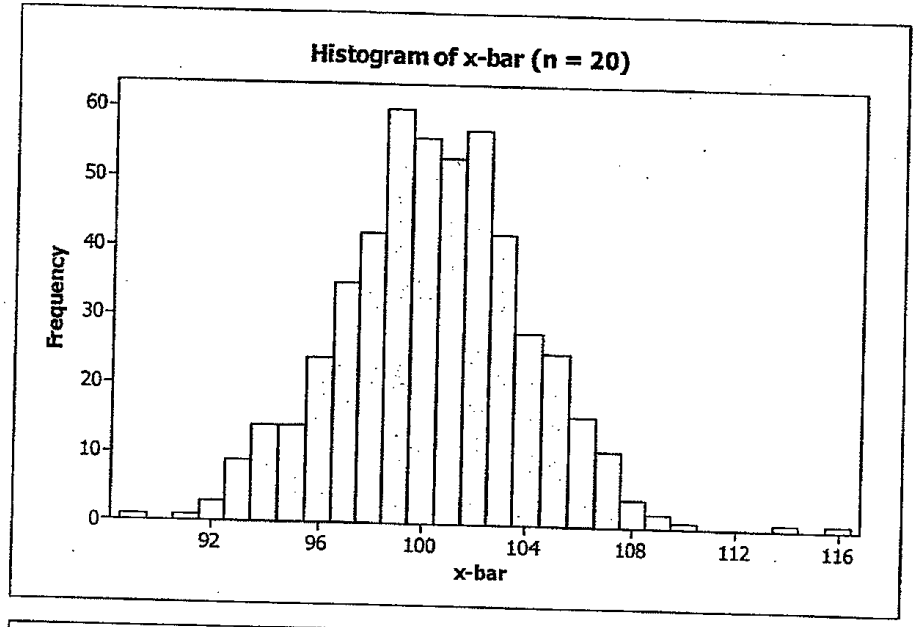
Now for n = 10:


Histogram of x bar (n = 10)
Frequency
0
10
20
30
40
50
60
70
80
90
84
88
92
96
100
104
108
112
x bar

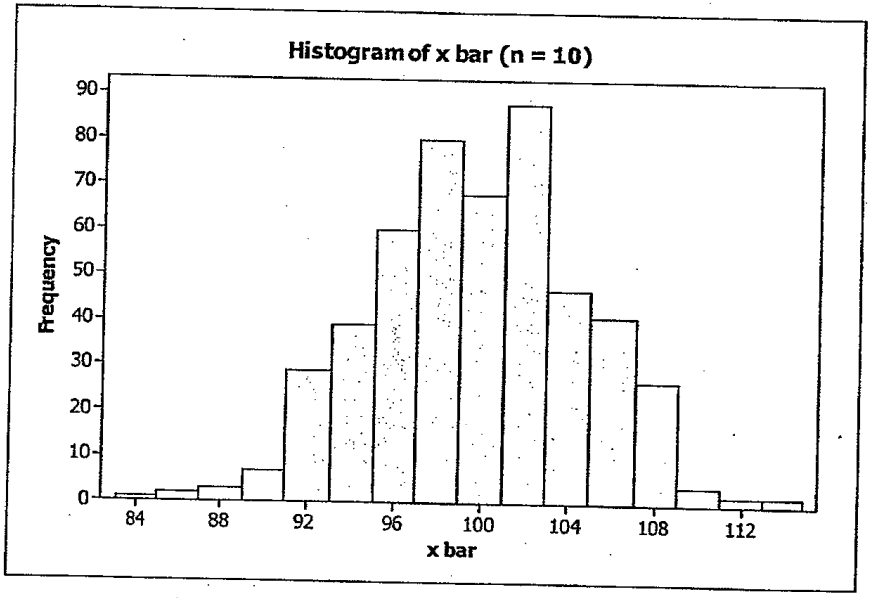

Histogram of x-bar (n = 20)
Frequency
0
10
20
30
40
50
60
92
96
100
104
108
112
116
x-bar

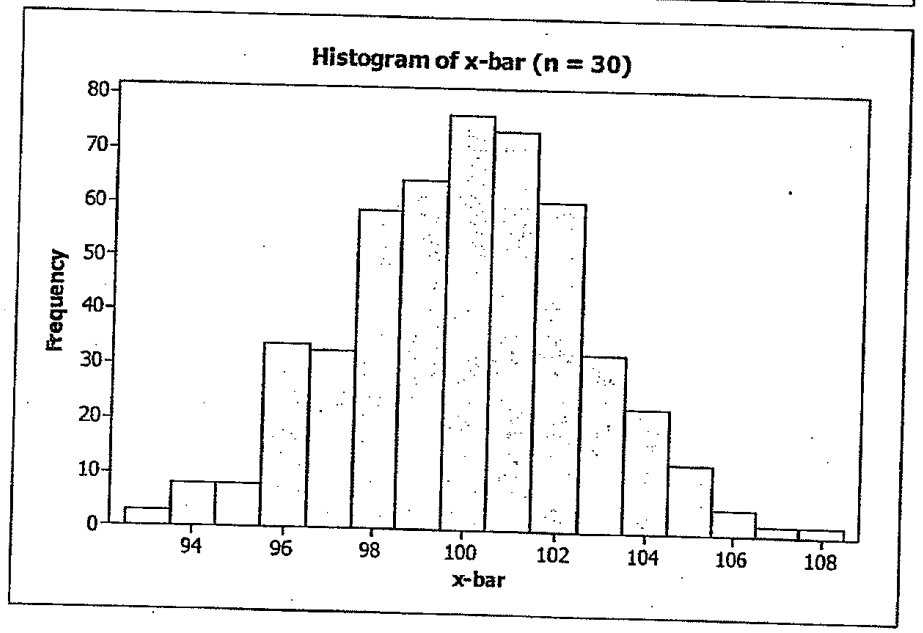

Histogram of x-bar (n = 30)
Frequency
0
10
20
30
40
50
60
70
80
94
96
98
100
102
104
106
108
x-bar


Histogram of x-bar (n = 5)
Frequency
x-bar
0
10
20
30
40
50
60
70
84
90
96
102
108
114


Histogram of x-bar (n = 20)
Frequency
x-bar
0
10
20
30
40
50
60
92
96
100
104
108
112
116

Now for n = 10:

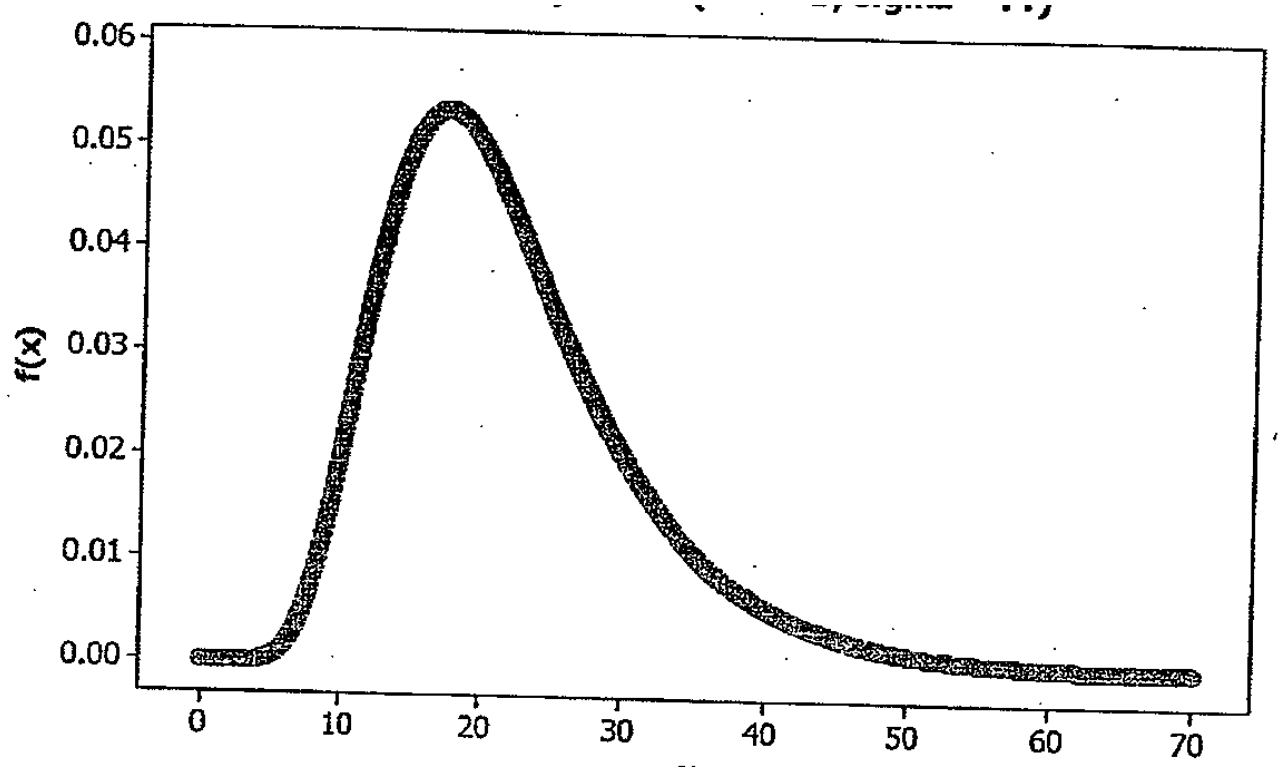

Histogram of x bar (n = 10)
Frequency
x bar
0
10
20
30
40
50
60
70
80
90
84
88
92
96
100
104
108
112

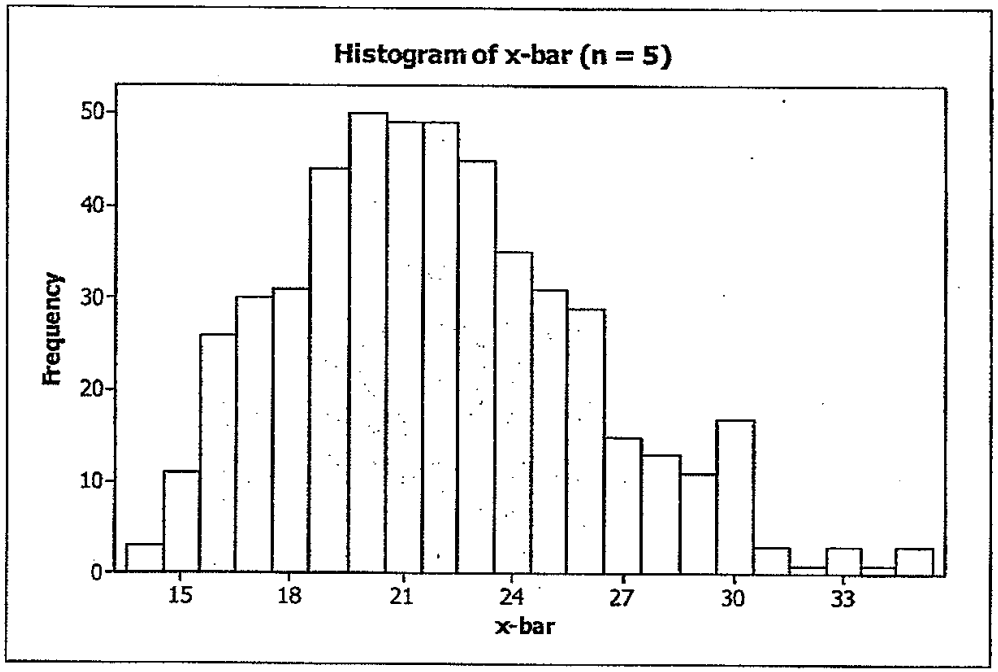

Histogram of x-bar (n = 30)
Frequency
x-bar
0
10
20
30
40
50
60
70
80
94
96
98
100
102
104
106
108

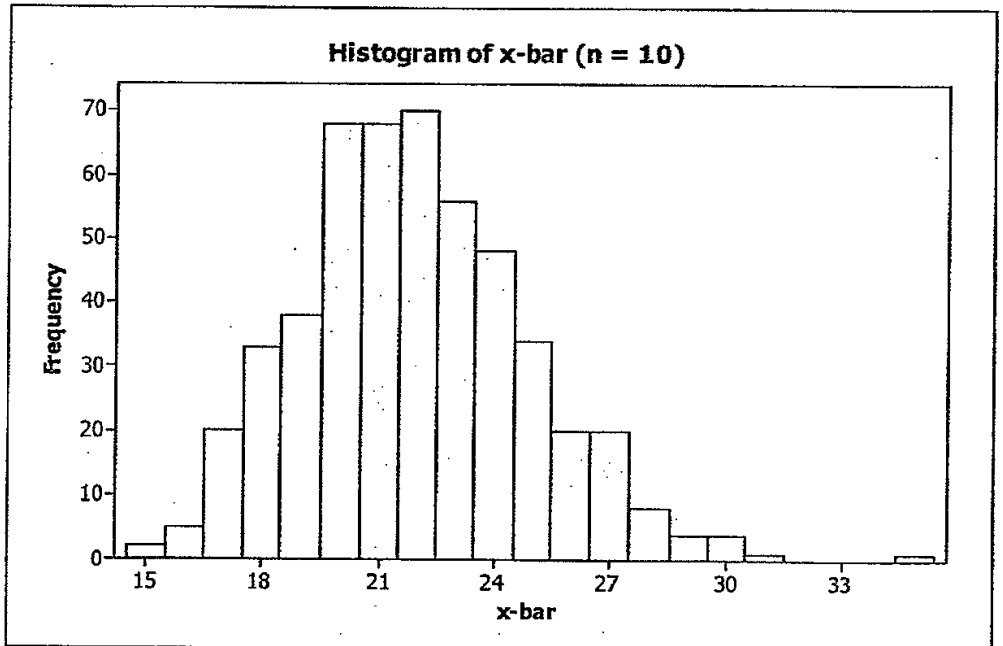
What to Expect of Sample Means: Population of measurements of interest is not bell-shaped, but a large random sample is measured.

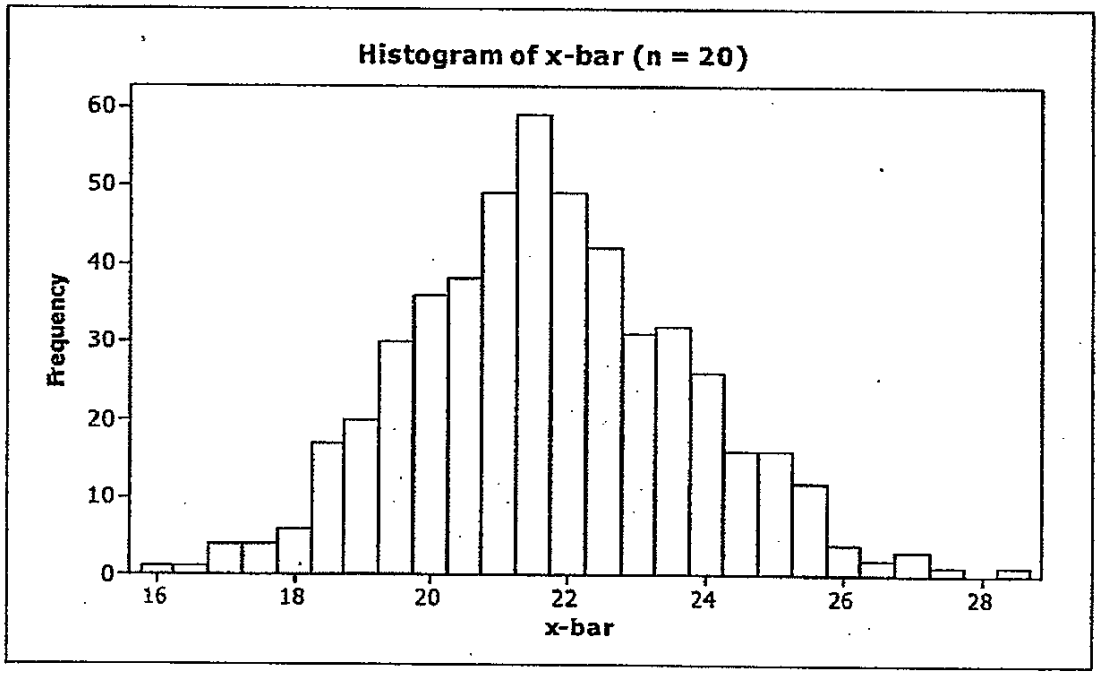
- Sample of size 30 is considered "large," but if there are extreme outliers, better to have a larger sample.

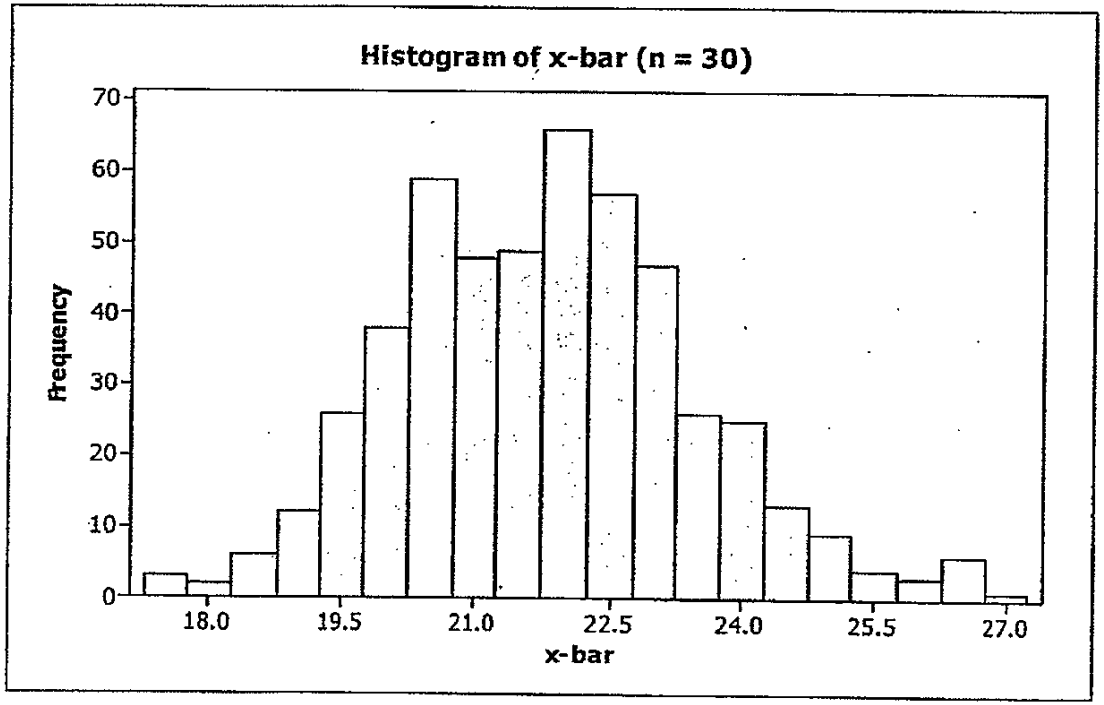
Population Mean is 21.76

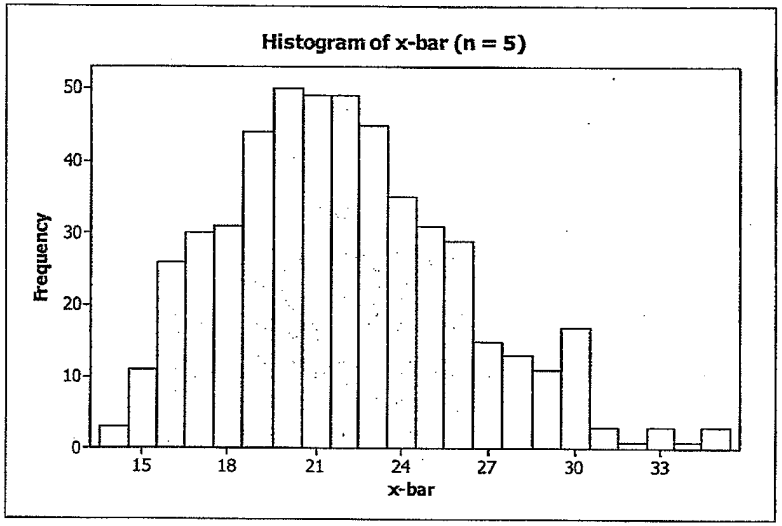


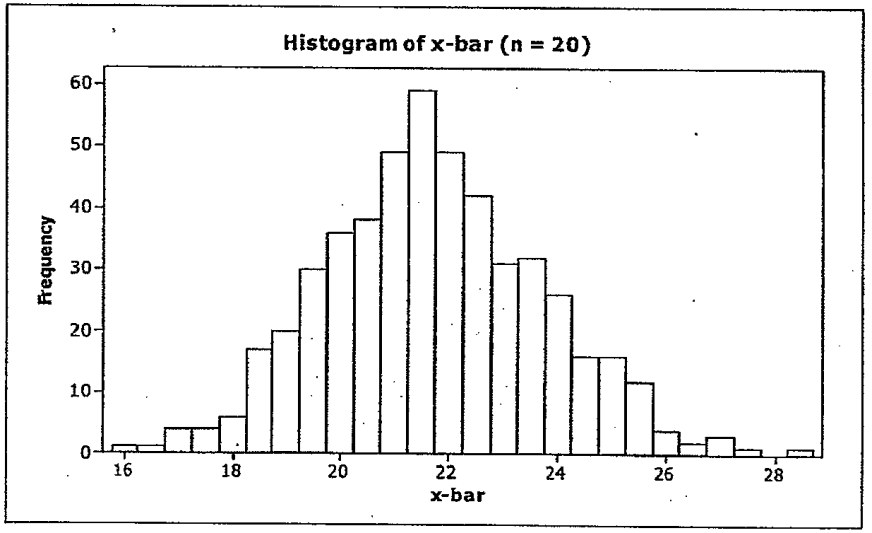

Histogram of x-bar (n = 5)
Frequency
0
10
20
30
40
50
15
18
21
24
27
30
33
x-bar

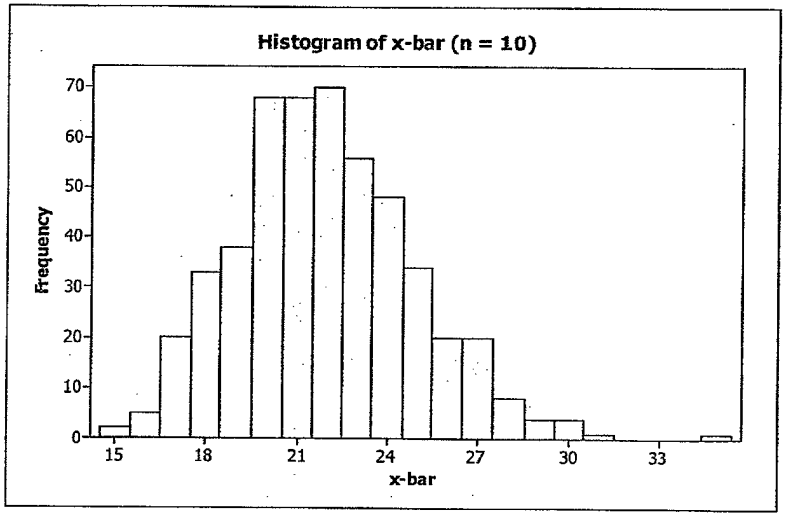

Histogram of x-bar (n = 10)
Frequency
0
10
20
30
40
50
60
70
15
18
21
24
27
30
33
x-bar

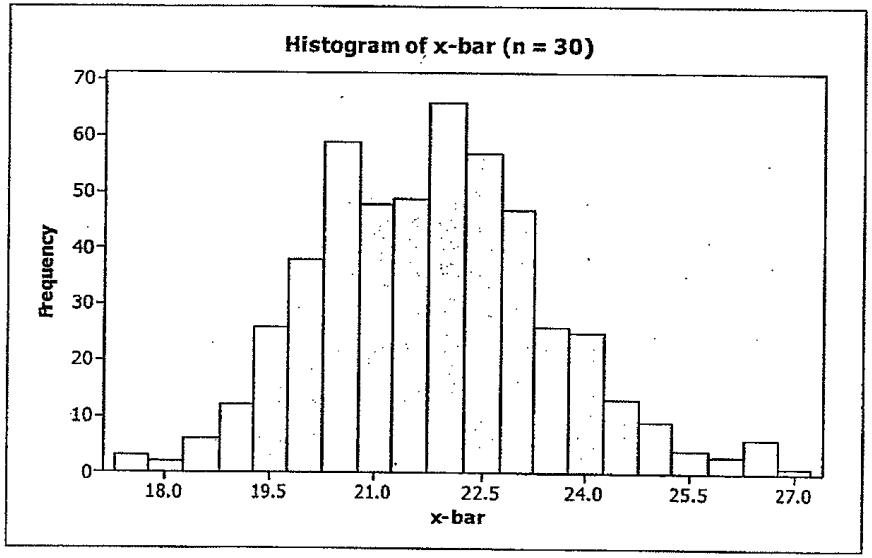

Histogram of x-bar (n = 20)
Frequency
0
10
20
30
40
50
60
16
18
20
22
24
26
28
x-bar


Histogram of x-bar (n = 30)
Frequency
0
10
20
30
40
50
60
70
18.0
19.5
21.0
22.5
24.0
25.5
27.0
x-bar

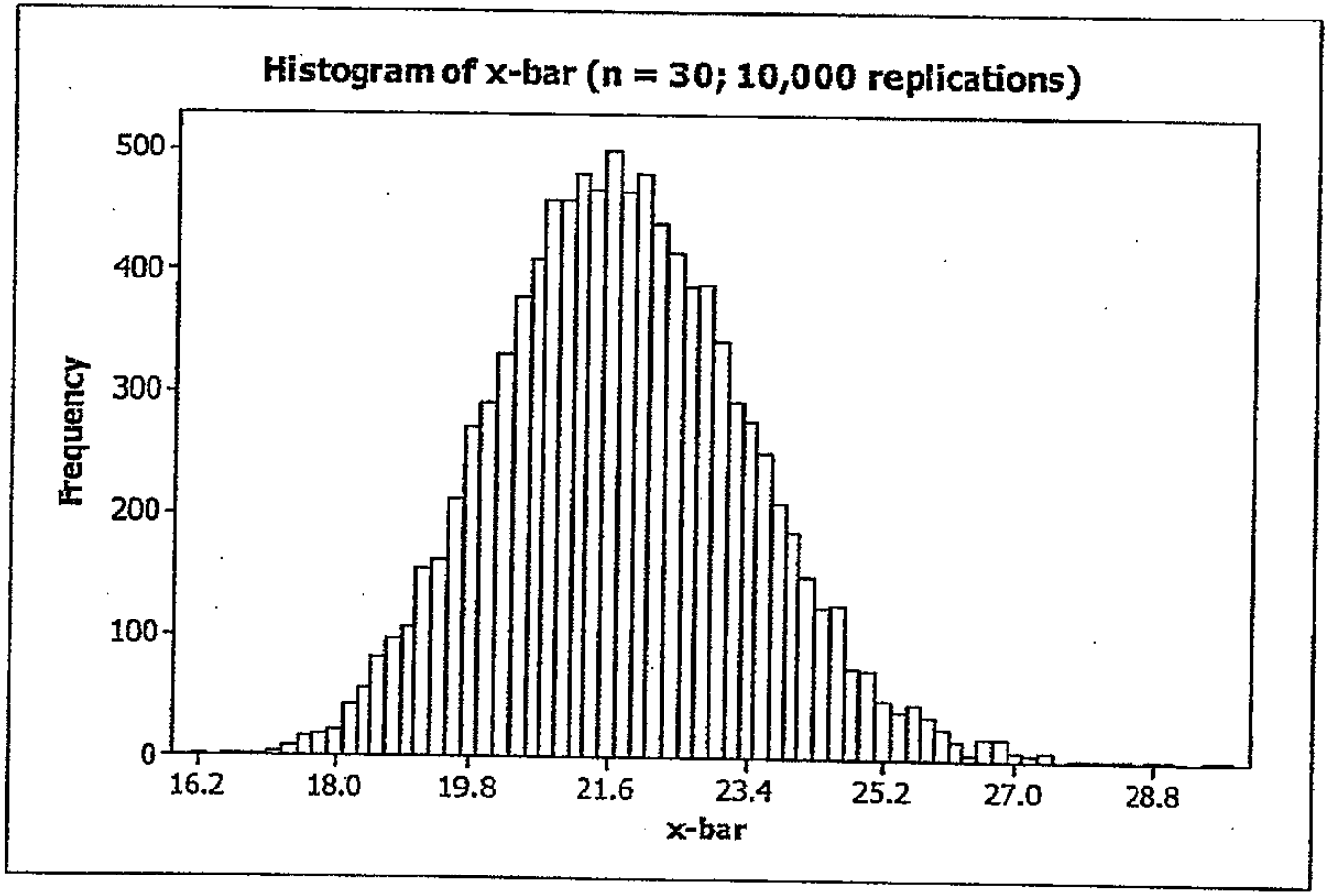

Histogram of x-bar (n = 5)
Frequency
0
10
20
30
40
50
15
18
21
24
27
30
33
x-bar


Histogram of x-bar (n = 20)
Frequency
0
10
20
30
40
50
60
16
18
20
22
24
26
28
x-bar


Histogram of x-bar (n = 10)
Frequency
0
10
20
30
40
50
60
70
15
18
21
24
27
30
33
x-bar


Histogram of x-bar (n = 30)
Frequency
0
10
20
30
40
50
60
70
18.0
19.5
21.0
22.5
24.0
25.5
27.0
x-bar

What to Expect of Sample Means: Population of measurements of interest is not bell-shaped, but a large random sample is measured.



Very close to normal-land!

Variation in Sample Statistics

source: https://mcu.edu/

Sample Means from 12 samples with a sample size equal to 16 men

68.3, 68.7, 69.2, 69.4, 69.6, 69.9, 70.1, 70.3, 70.5, 70.9, 71.1, 71.4.

Distribution of Men's Heights

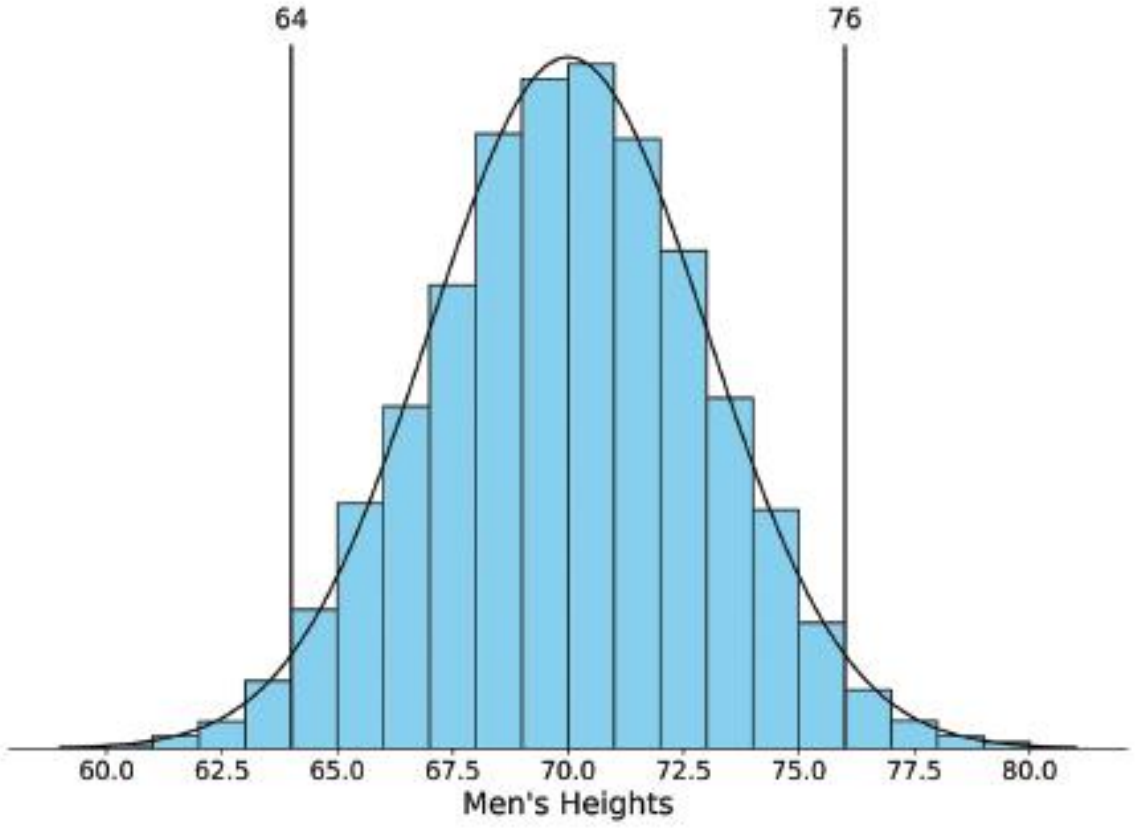
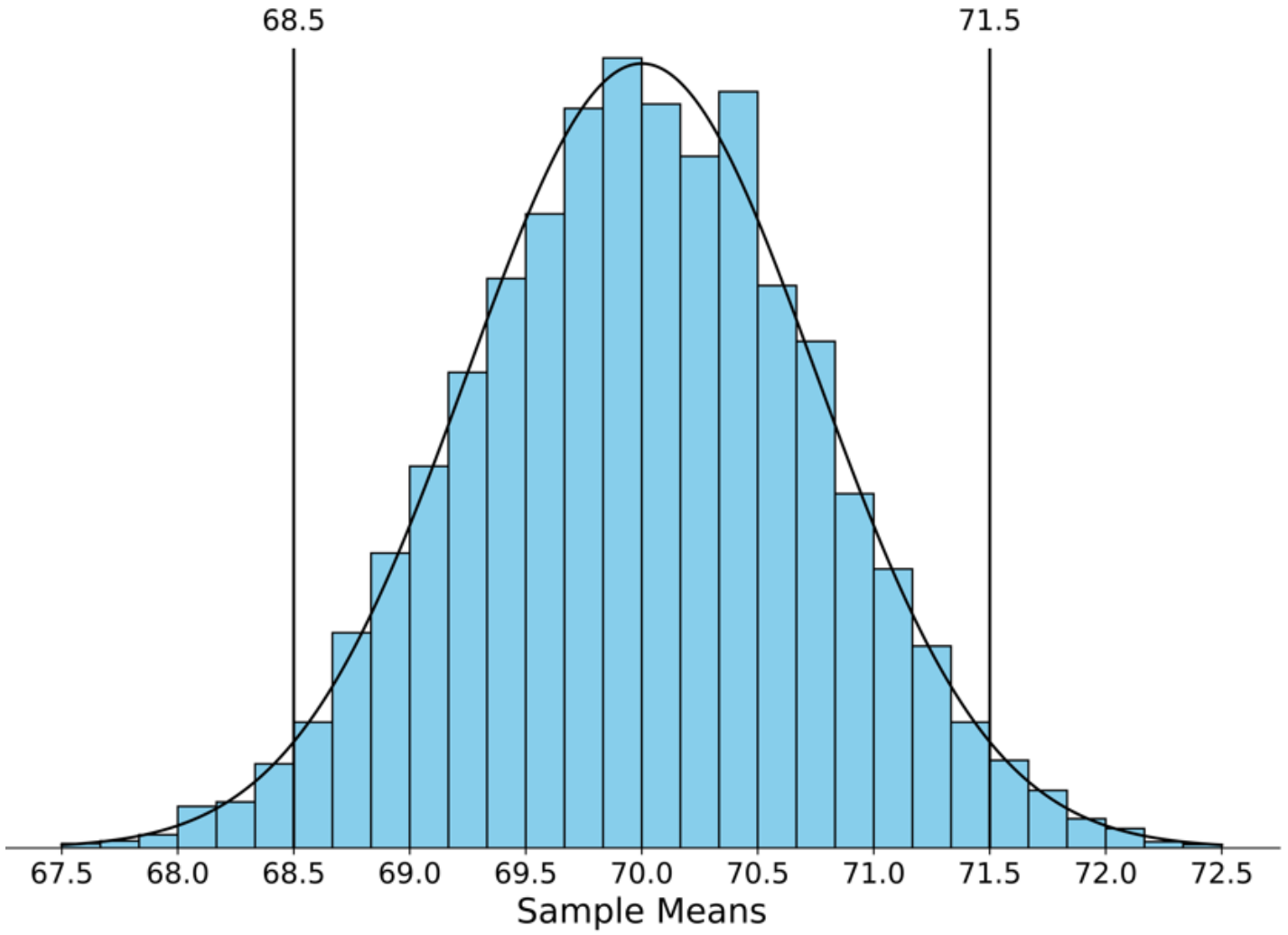
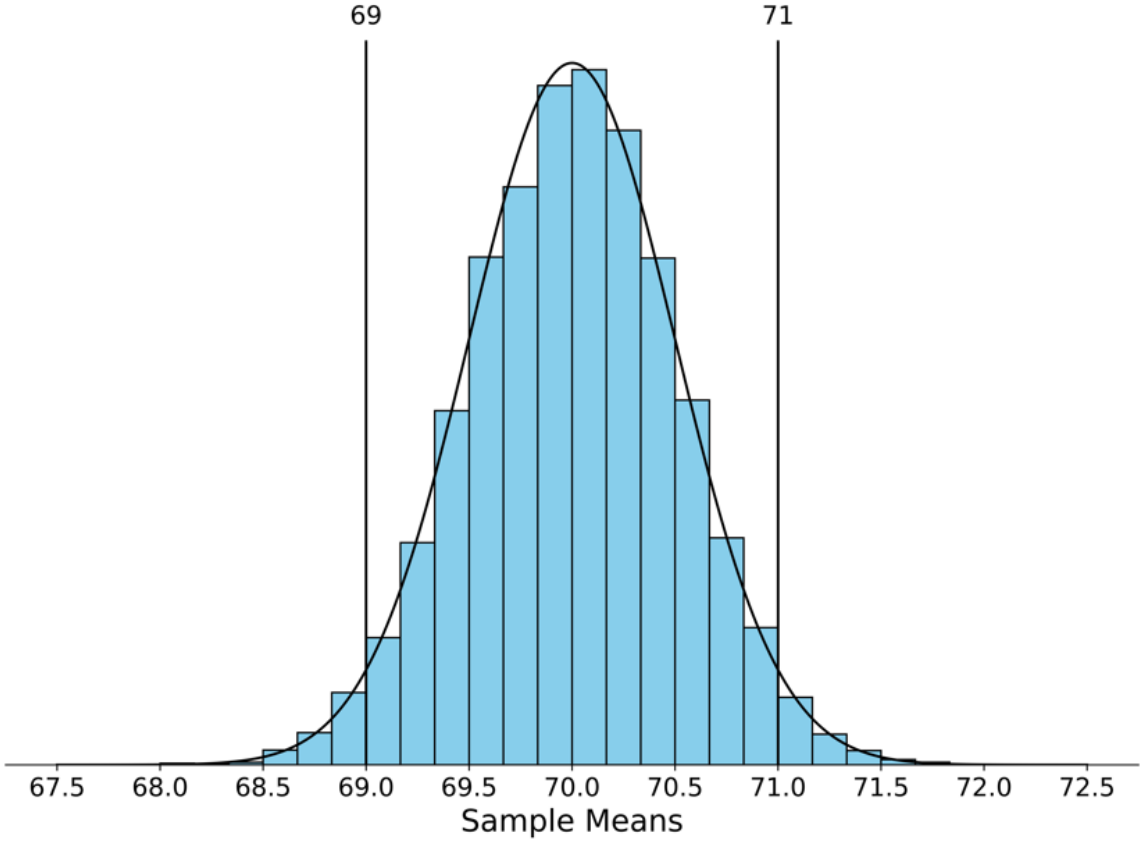


Figure 6.2: Population Distribution of Men's Heights in the United States

Distribution of Sample Mean Heights based on Sample Sizes of 16 Men

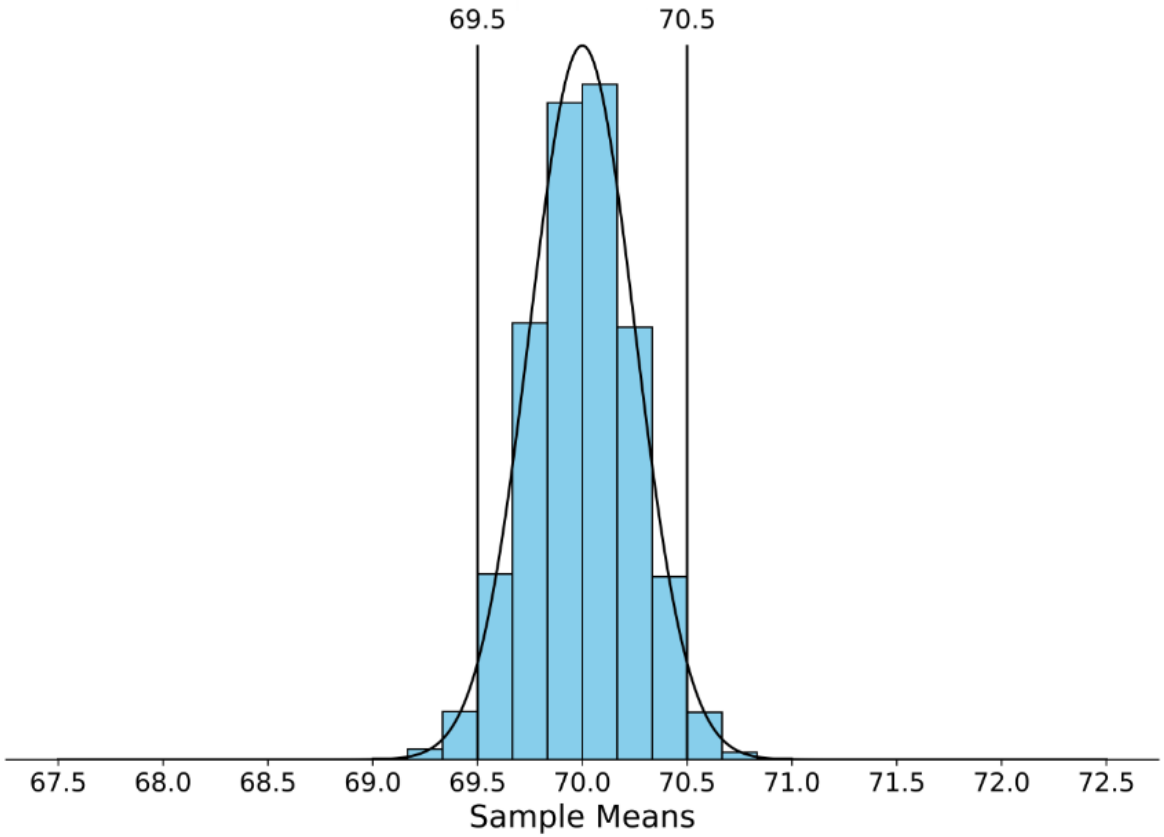


Distribution of Sample Mean Heights based on Sample Sizes of 36 Men



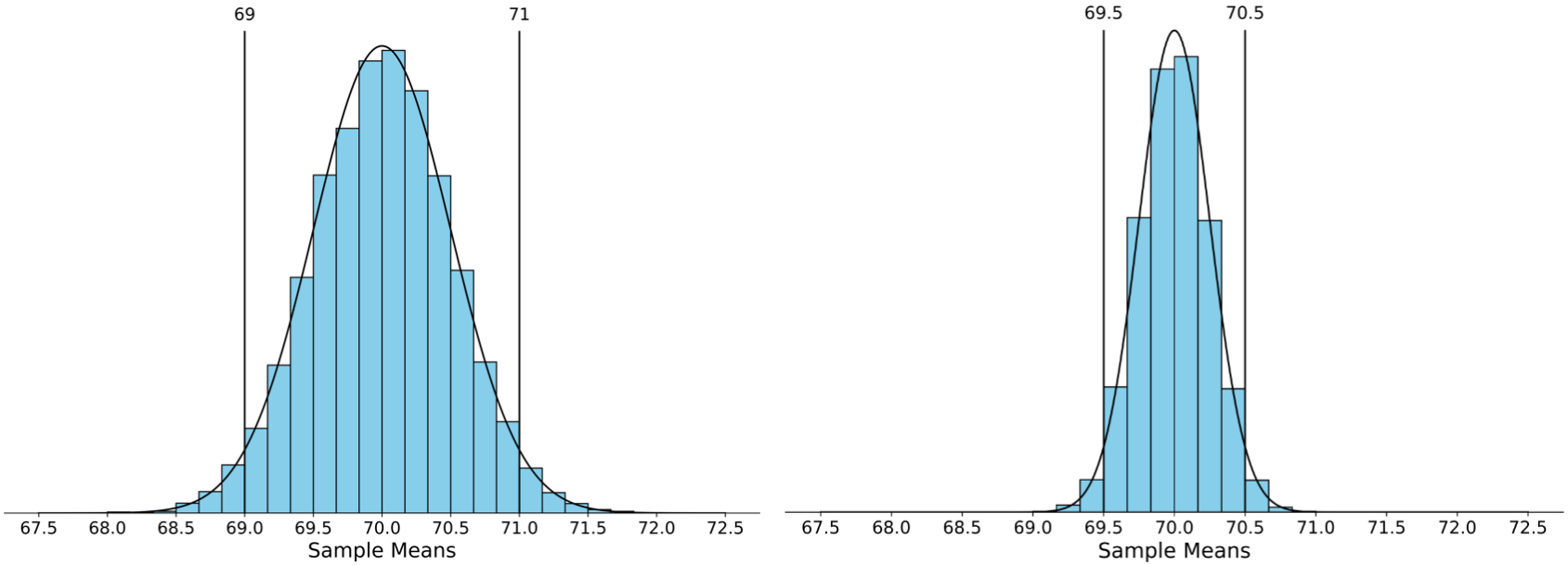
- For a sample of 36 men the standard error is 3/sqrt(36) = 0.5 inches
- 95% of possible sample mean heights between 69 inches and 71 inches

Distribution of Sample Mean Heights based on Sample Sizes of 144 Men



- For a sample of 144 men the standard error is 3/sqrt(144) = 0.25 inches
- 95% chance our sample mean will be between 69.5 inches and 71.5 inches

The Width of the Sampling Distribution



- The larger the sample size, the narrower the sampling distribution
- In reality, we don't know what the population mean height is
- Our sample mean height is one of a range of possible sample mean heights normally distributed around the unknown truth (or population mean)

Sampling Distribution of Men's Heights

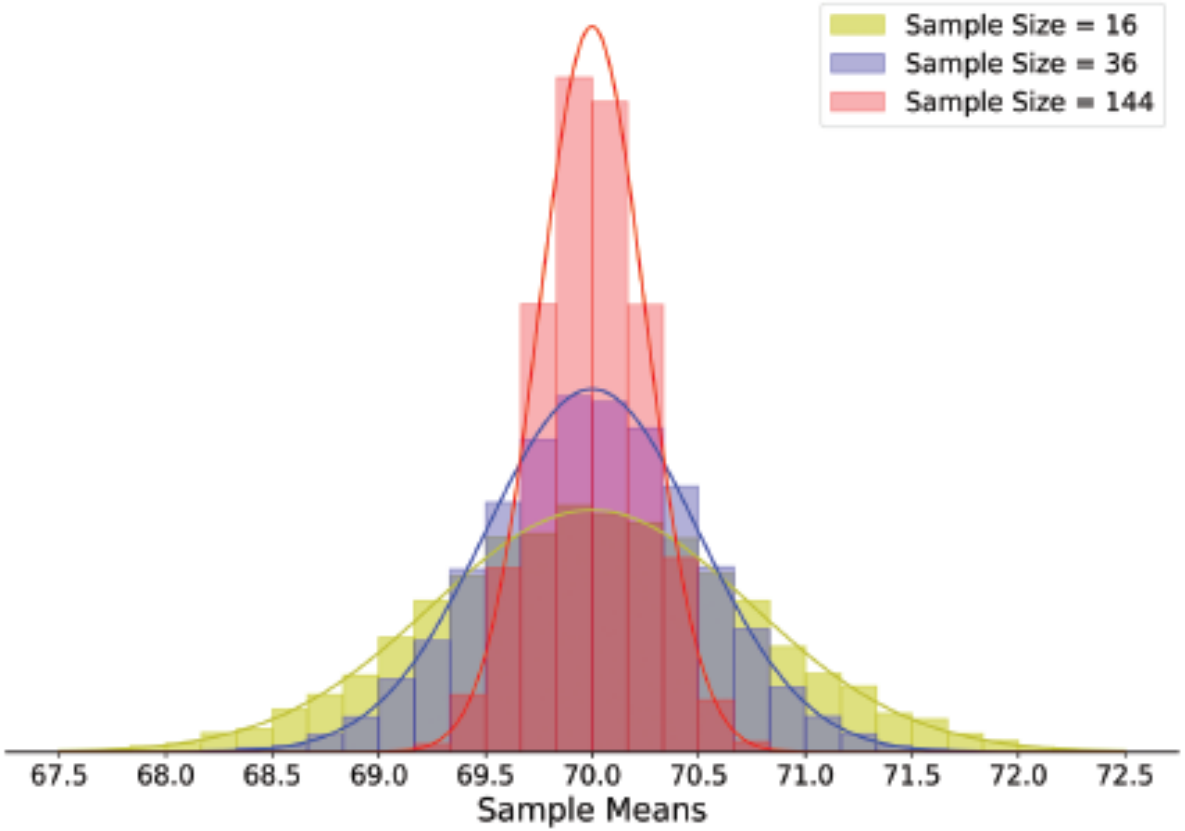


Figure 6.5: Simulation of 10,000 Sample Mean Heights—Sample Sizes Equal to 16, 36 and 144 Men