Chapter 3D

Random Behavior of Means

Some Estimates

Parameter	Measure	Statistic
μ	Mean of a single population	$\bar{X}$
σ^2	Variance of a single population	S^2
σ	Standard deviation of a single population	S
p	Proportion of a single population	$\hat{p}$
$\mu_1 - \mu_2$	Difference in means of two populations	$\bar{X}_1 - \bar{X}_2$
$p_1 - p_2$	Difference in proportions of two populations	$\hat{p}_1 - \hat{p}_2$

- To estimate the mean of a population, we could use the Sample mean ($\bar{X}$).
- Is the sample mean a good estimate?

Population Mean

- Parameter labeled, μ.
- Often too large to calculate or too difficult to access.
- If a probability distribution can represent this population, then the population mean is considered the mean of a random variable.
- Consider estimating it with the sample mean. Would this be a "Good" Estimate?

A "Good" Estimate

An estimation method should be both accurate and precise.

- Accurate – The method measures what it intended; correctly estimates the population parameter.

- Precise – If the method is repeated, the estimates are very consistent.

To be a good golfer, we need to be both *accurate* (tends to hit the ball near the cup) and *precise* (shot is repeatable, consistent).

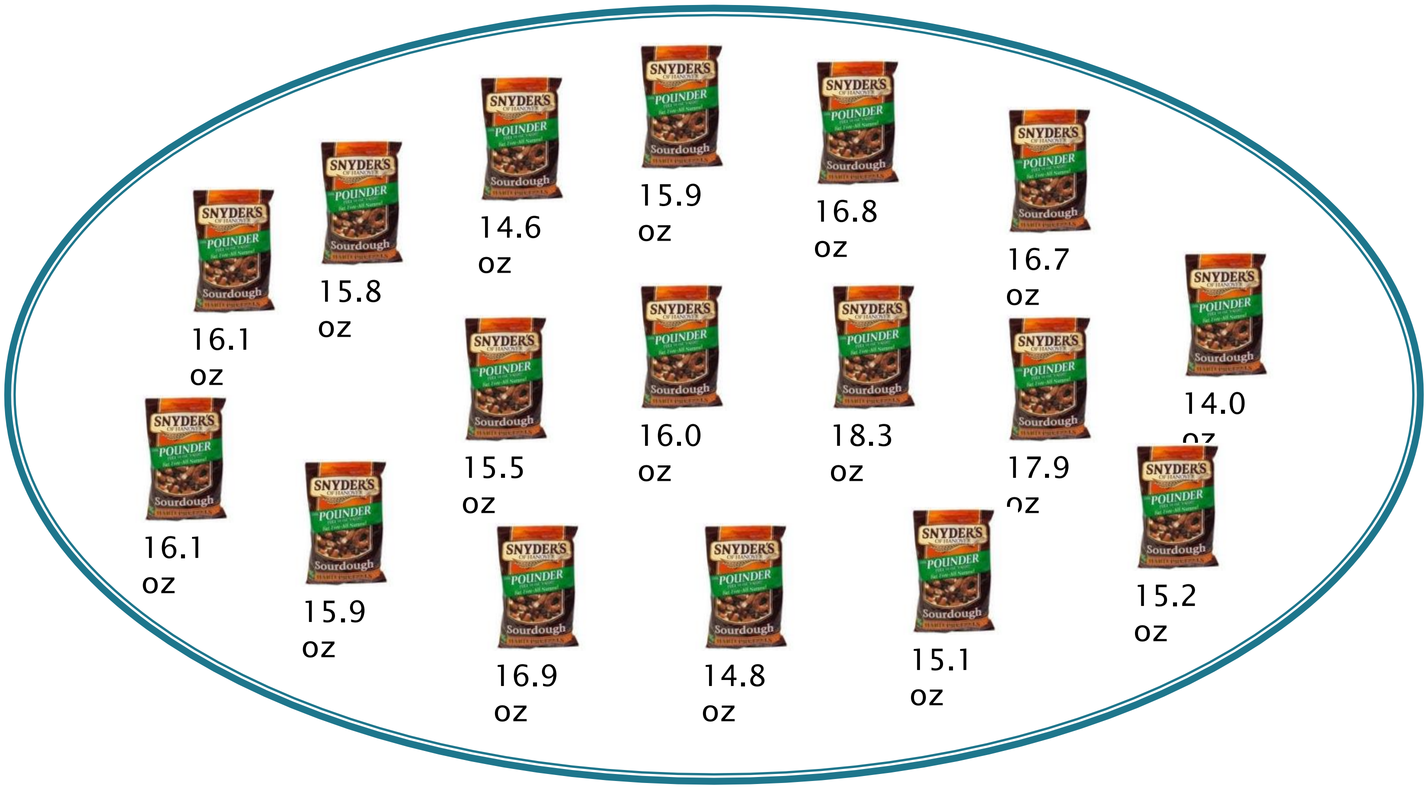
An accurate and precise estimate is called an *Unbiased* estimate

Estimating the mean (normal population)

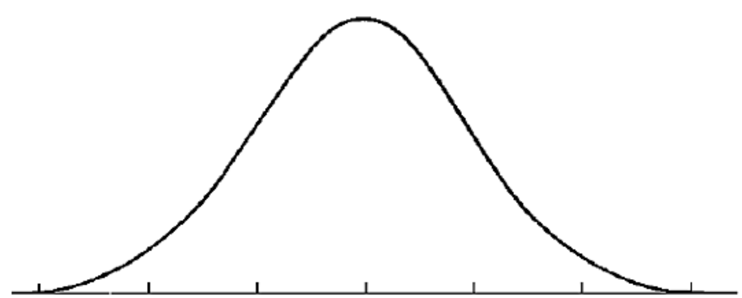
- Consider the distribution of weights of bags of pretzels. Assume the population distribution of weights is normal with $\mu = 16$ and $\sigma = 5$

- Imagine taking multiple samples from this population

Population Distribution (Normal)

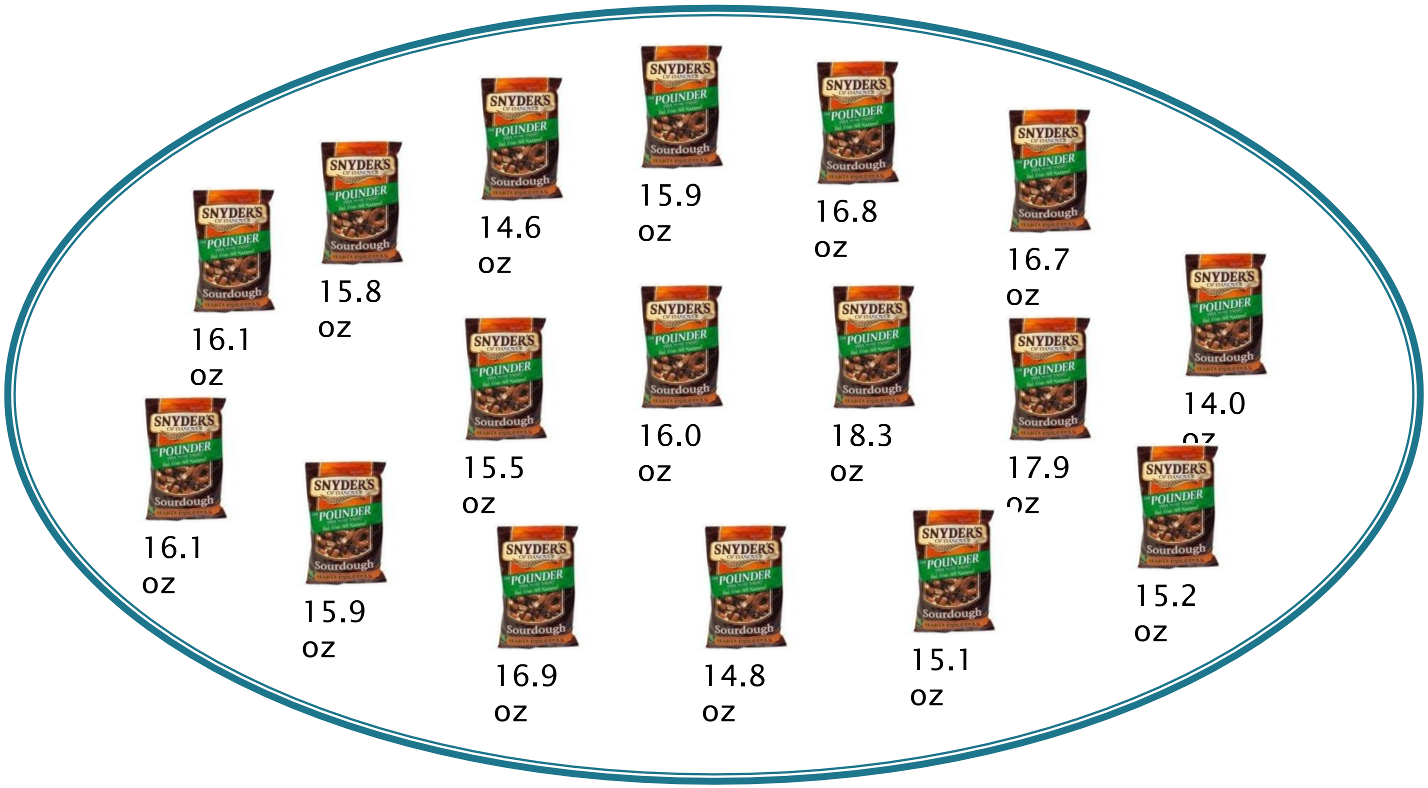


Shape of the population of weights:



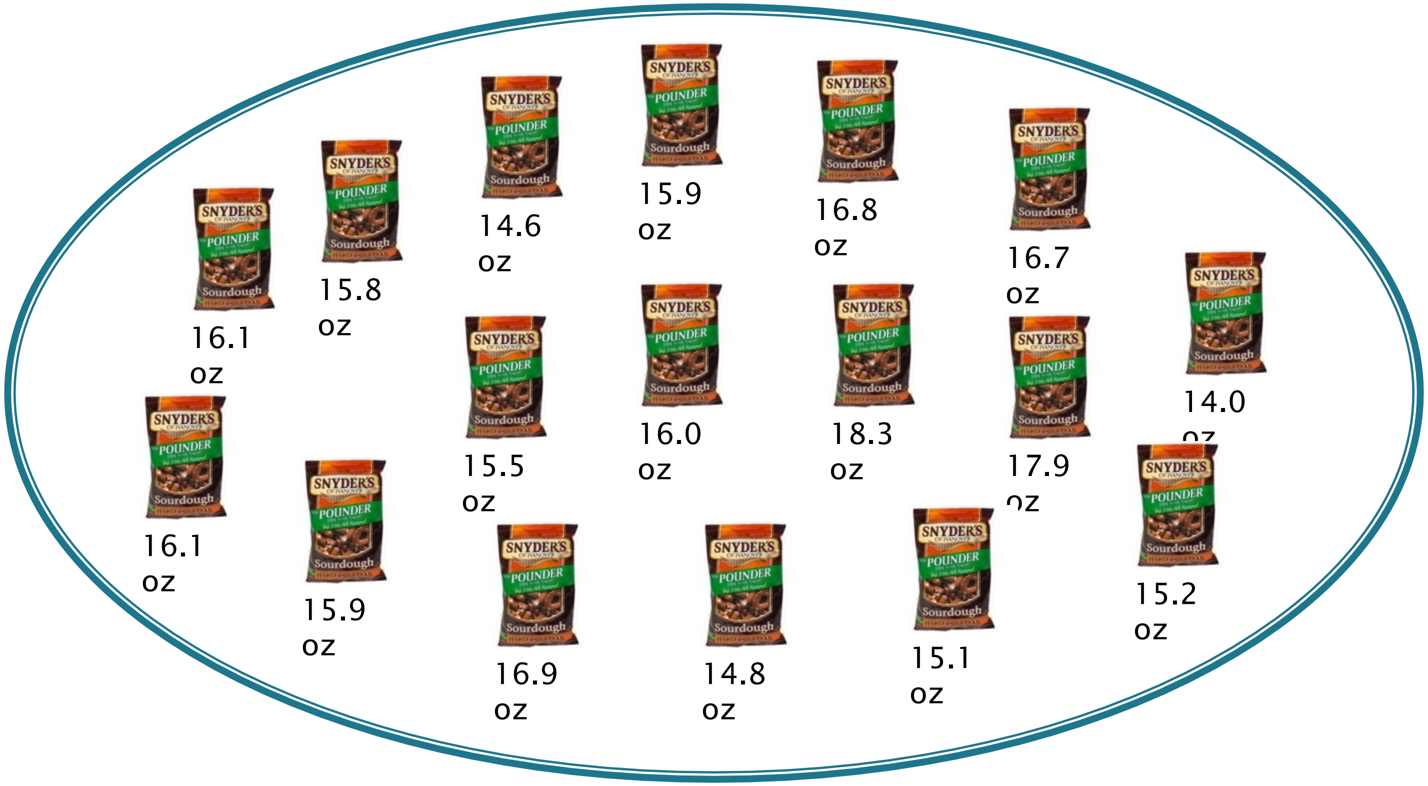
X, Weights

Population of Weights

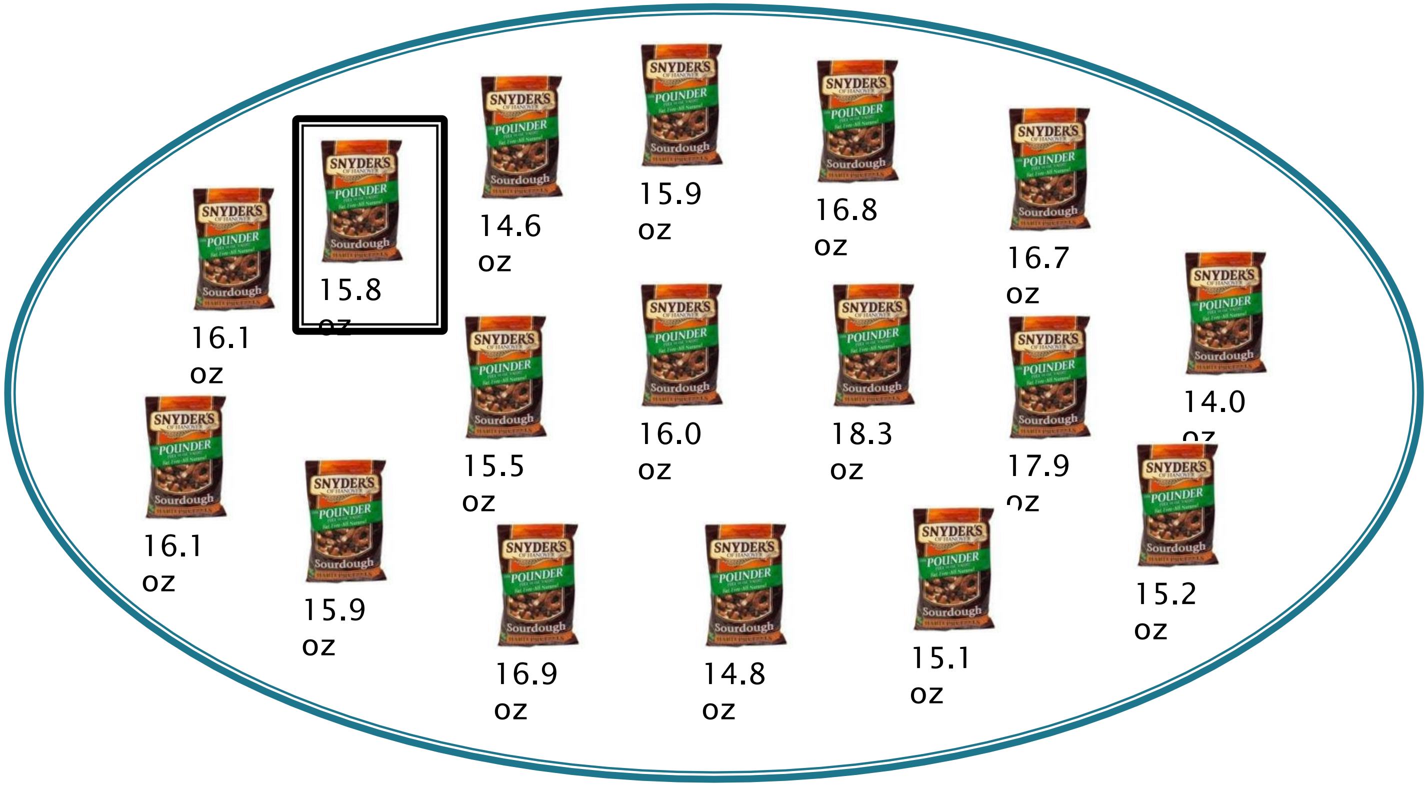


Impossible to obtain all the weights in the population

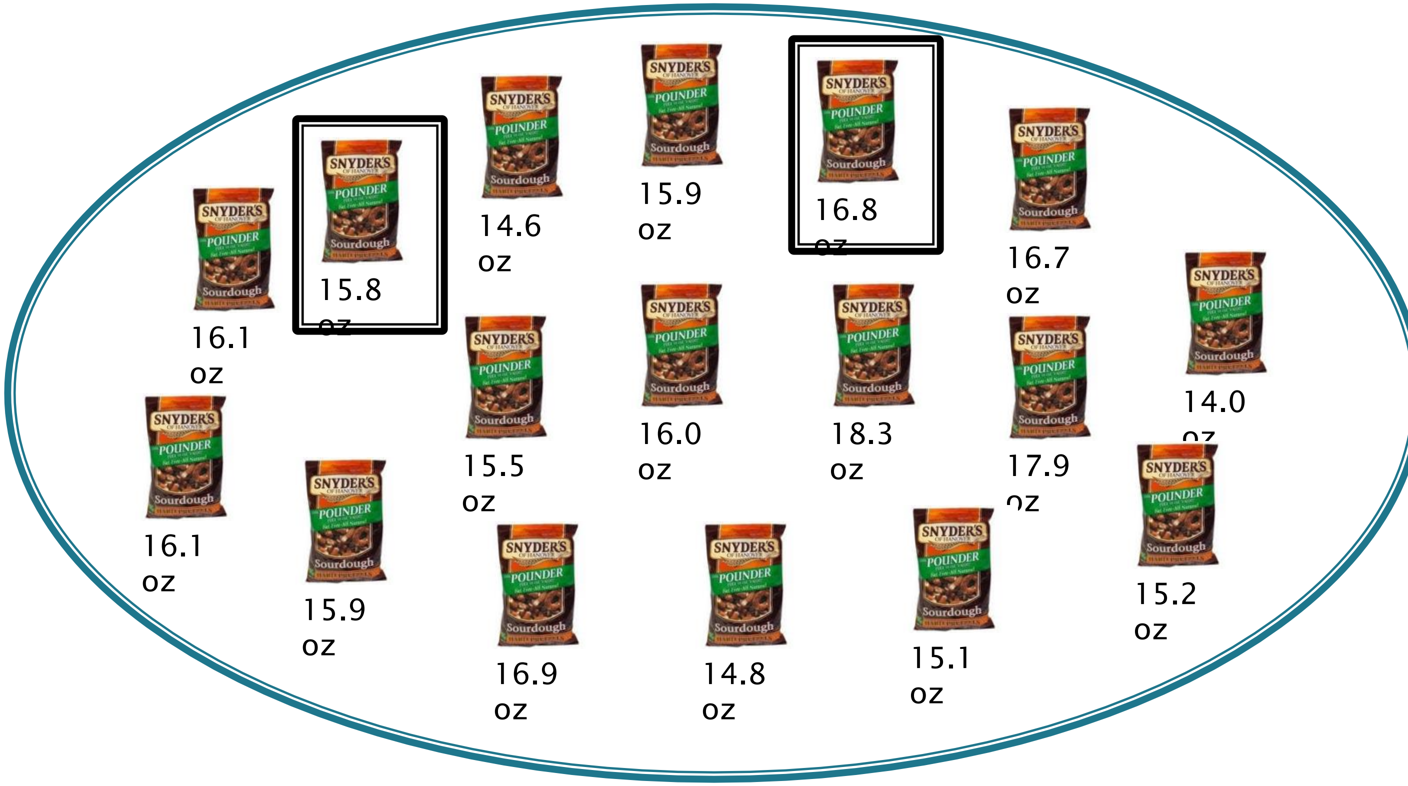
Sample from this population



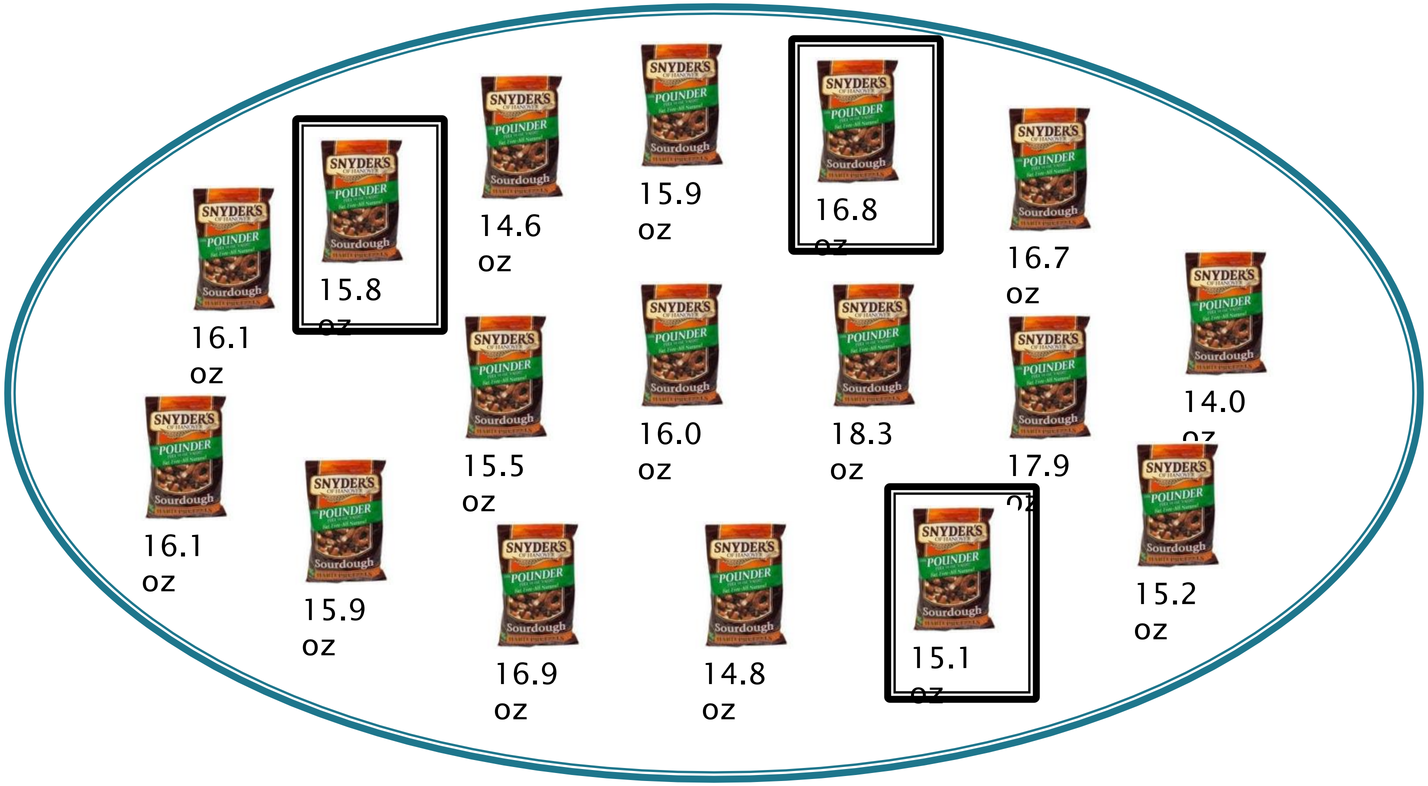
Sample from this population



Sample from this population

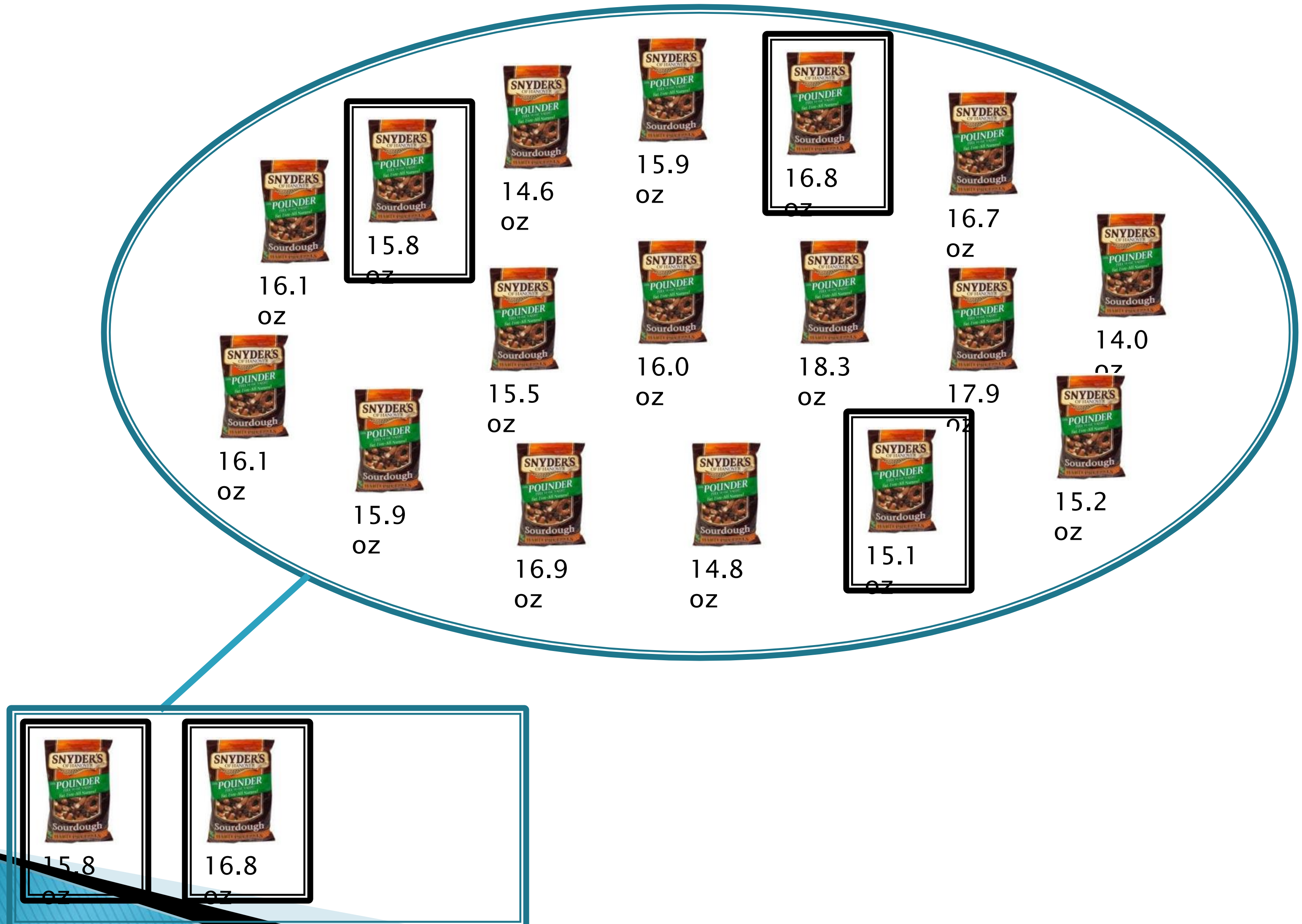


Sample from this population

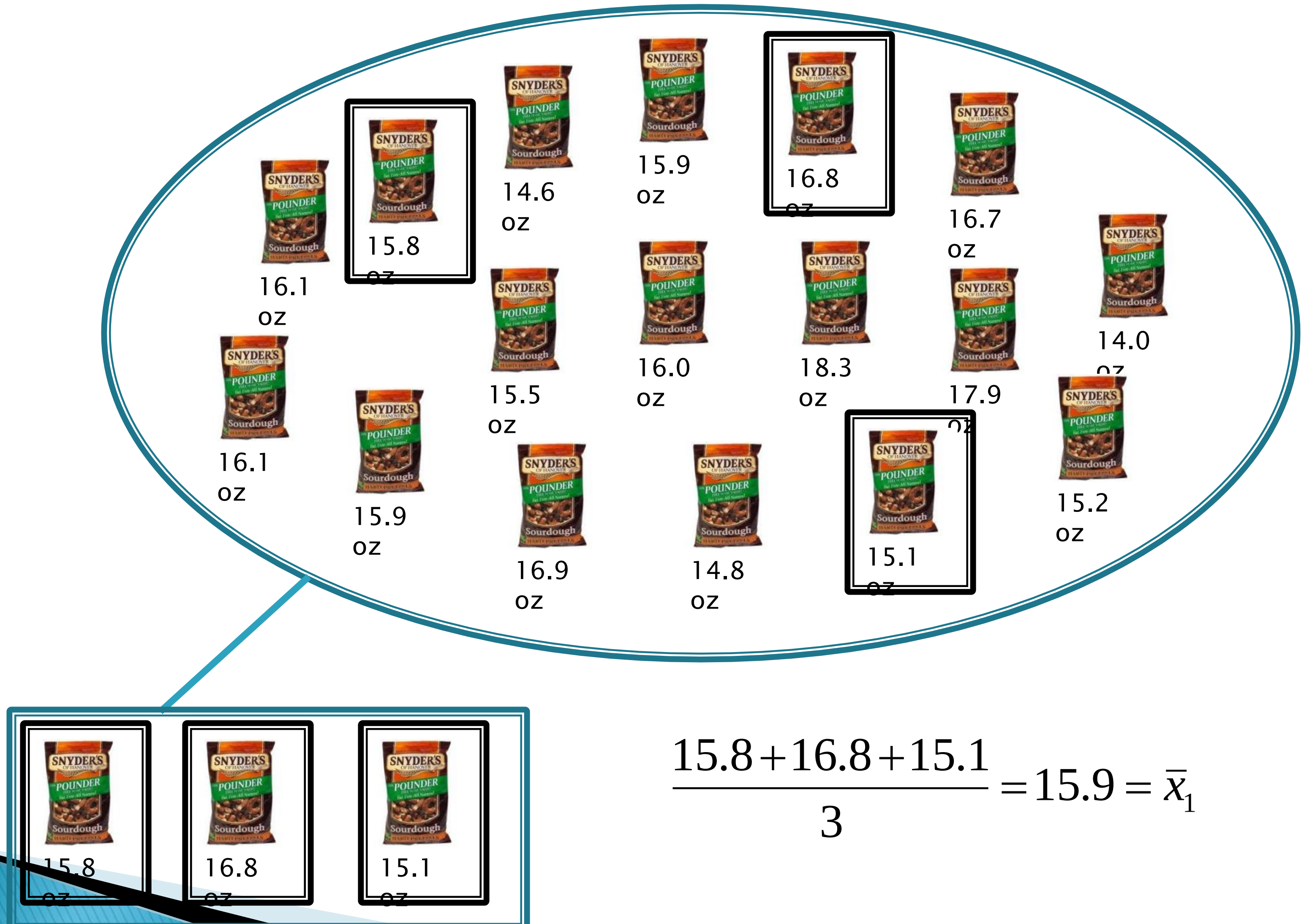


Sample from this population

Sample from this population



Sample from this population

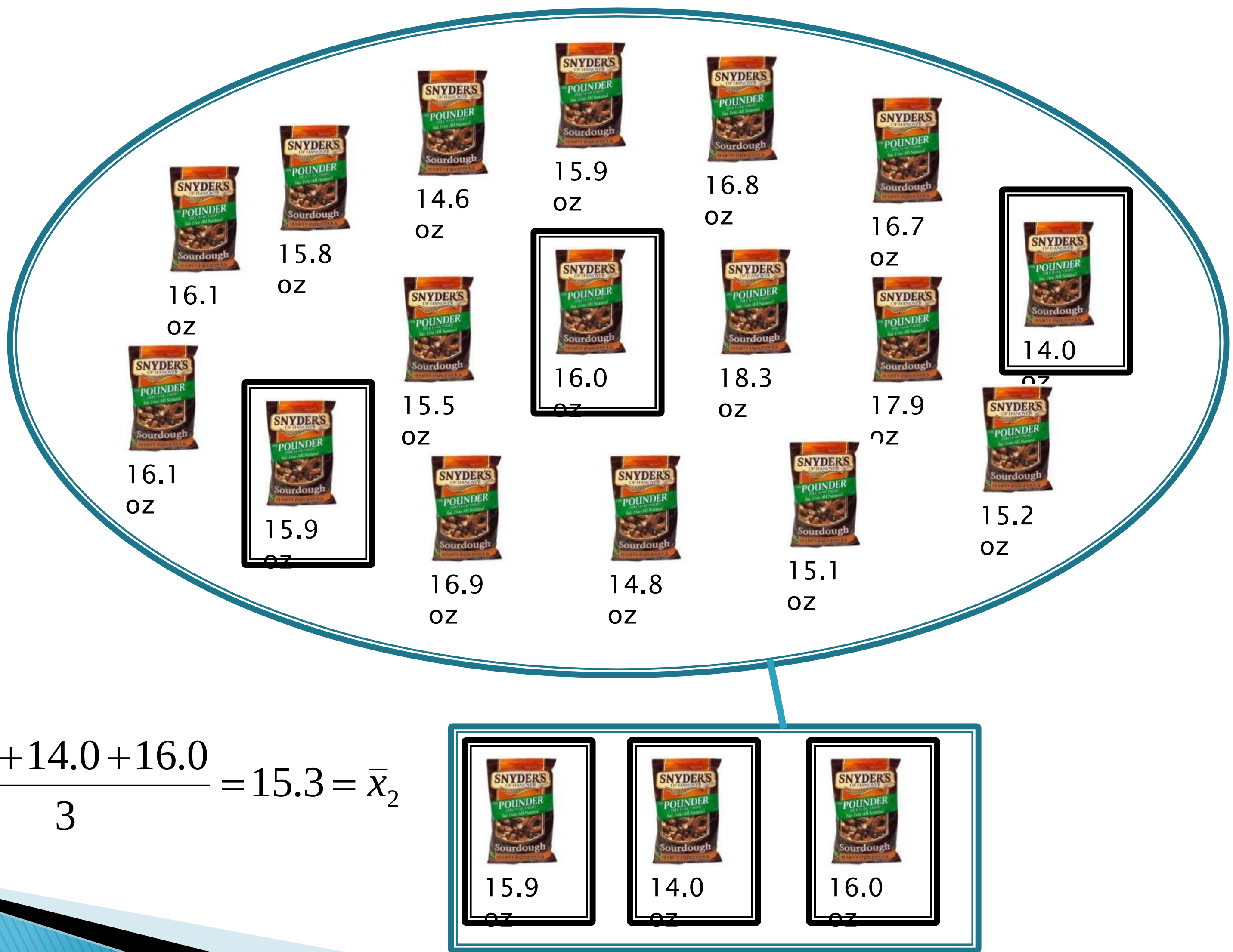


$$\frac{15.8+16.8+15.1}{3}=15.9=\bar{x}_1$$

Sample Mean is a Good Estimate

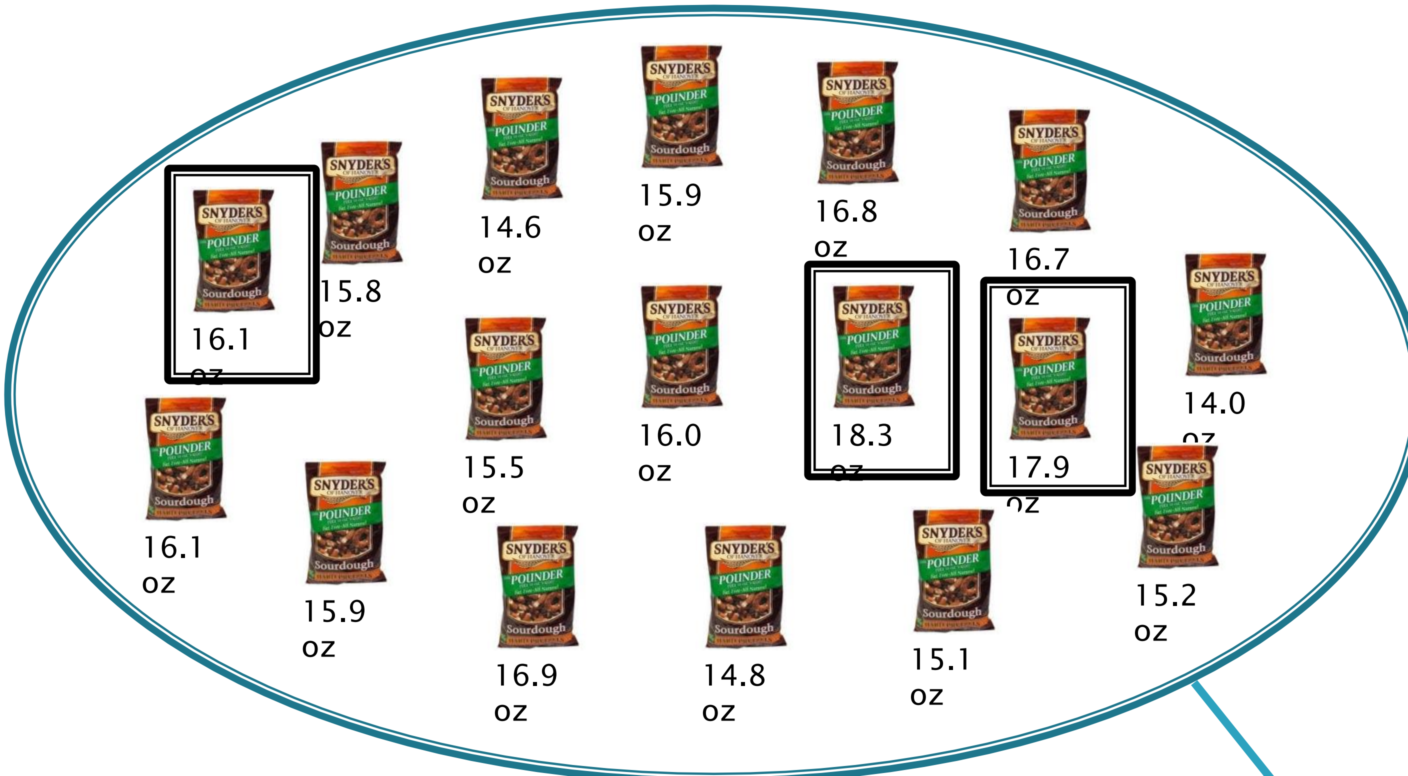
- But, how close is $\bar{x}$ to the unknown μ?
- This sample mean that we just found comes from a distribution of sample means.
- Do you think all samples will result in the same sample mean?

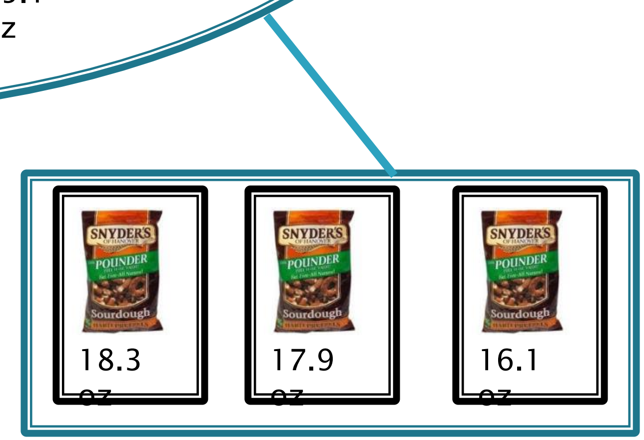
Sample from this population



$$\frac{15.9+14.0+16.0}{3} = 15.3 = \bar{x}_2$$

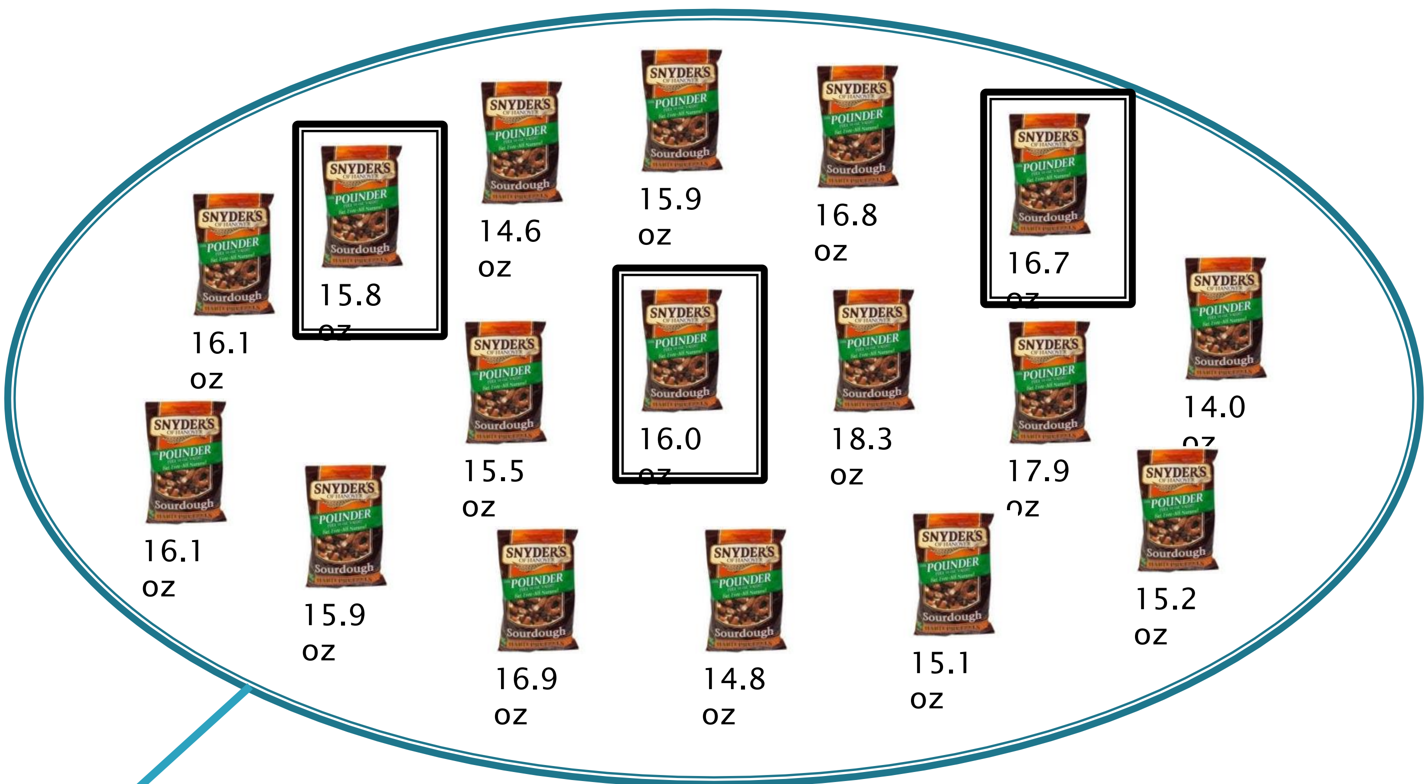
Sample from this population

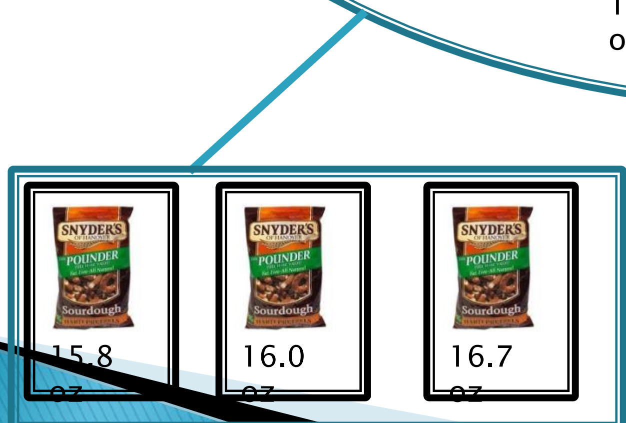




$$\frac{18.3+17.9+16.1}{3}=17.43=\bar{x}_3$$

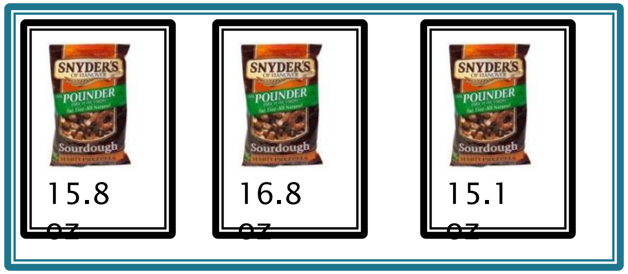
Sample from this population





$$\frac{15.8+16.0+15.7}{3}=15.83=\bar{x}_4$$

Repeated samples



$$\frac{15.8+16.8+15.1}{3}=15.9=\bar{x}_1$$

$$\frac{15.9+14.0+16.0}{3}=15.3=\bar{x}_2$$

$$\frac{18.3+17.9+16.1}{3}=17.43=\bar{x}_3$$

$$\frac{15.8+16.0+15.7}{3}=15.83=\bar{x}_4$$

Sampling Variability

- The value of a statistic varies in repeated random sampling

- Main idea: to see how trustworthy a procedure is, ask what would happen if we repeated it many times.

Sampling Distribution

- The **sampling distribution of a statistic** is the distribution of all possible values taken by the statistic when all possible samples of a fixed size n are taken from the population.

- It is a theoretical idea; in reality, we do not actually build it (though today we will simulate it).

- The sampling distribution of a statistic is the **probability distribution** of that statistic.

Accuracy and Pecision Connection

- We need an estimation method that aims in the right direction (accurate).
- Also, we need an estimation method that if we repeat the process we would arrive at nearly the same estimate (precise).
- We measure accuracy and precision using simulation.
  - We think about an estimate's accuracy by considering bias (which focuses on center).
  - We will measure an estimate's precision with a statistic called the standard error (which focuses on spread).

Sampling in Minitab

- Create a hypothetical population
  - Calc -> Random Data – > Normal (enter parameters and N)

- Using the pull down menus or commands in the session window will only allow you to take one sample at a time.
- If we want to take multiple samples at once, press “control and L” to open the command line editor
- Type (or copy and paste) the following into this window

```
sample 5 c1 c2
sample 5 c1 c3
sample 5 c1 c4
……..
```

- You can read the first command as “take a sample of 5 from c1 and store it in c2”

Sampling Distribution of the sample mean (normal population)

- We will take many random samples of a given size *n* from a population with mean *m* and standard deviation *s*.

- Some sample means will be above the population mean *m* and some will be below, making up the sampling distribution.

- We will begin with the normal “population” distribution (100,000 values) of weights with $\mu = 16$ and $\sigma = 5$.

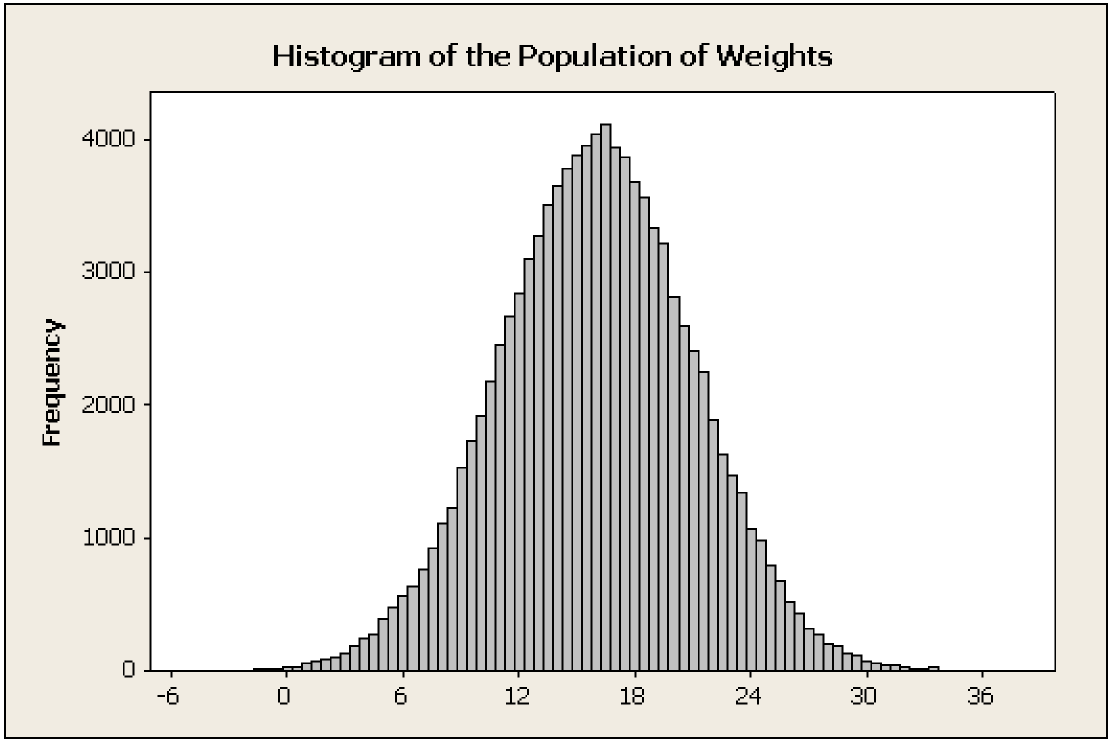
- Let’s simulate taking 1000 samples and graphing their means in Minitab (CLT Normal)

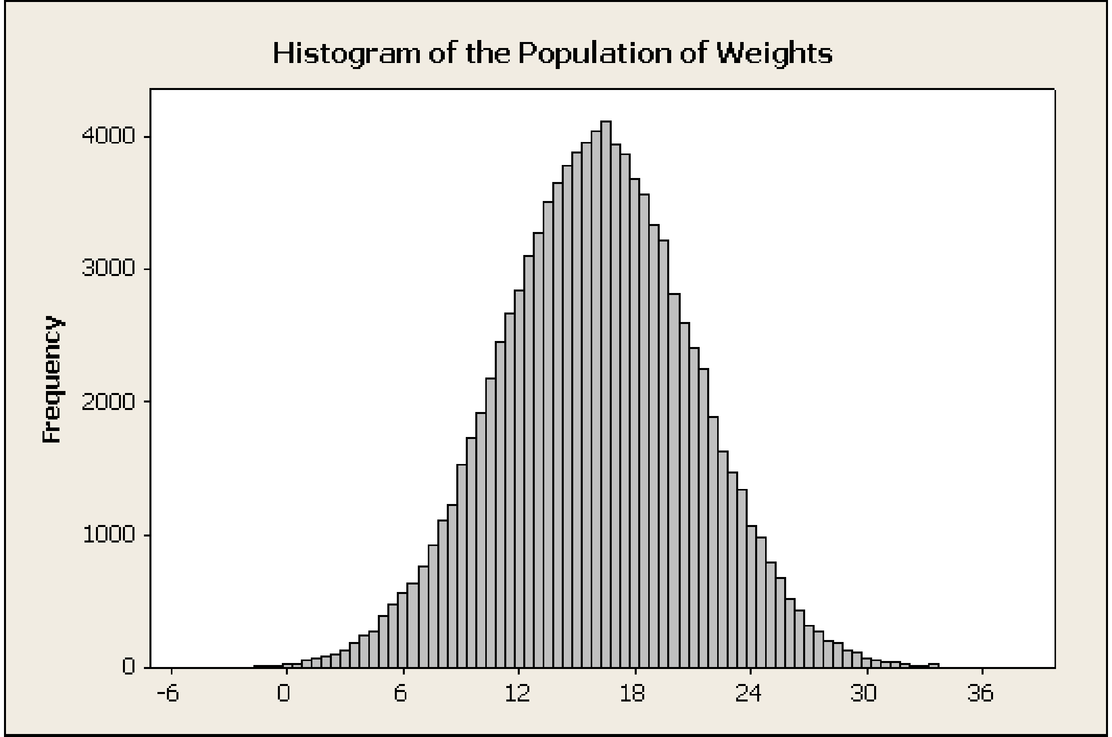
Questions to think about

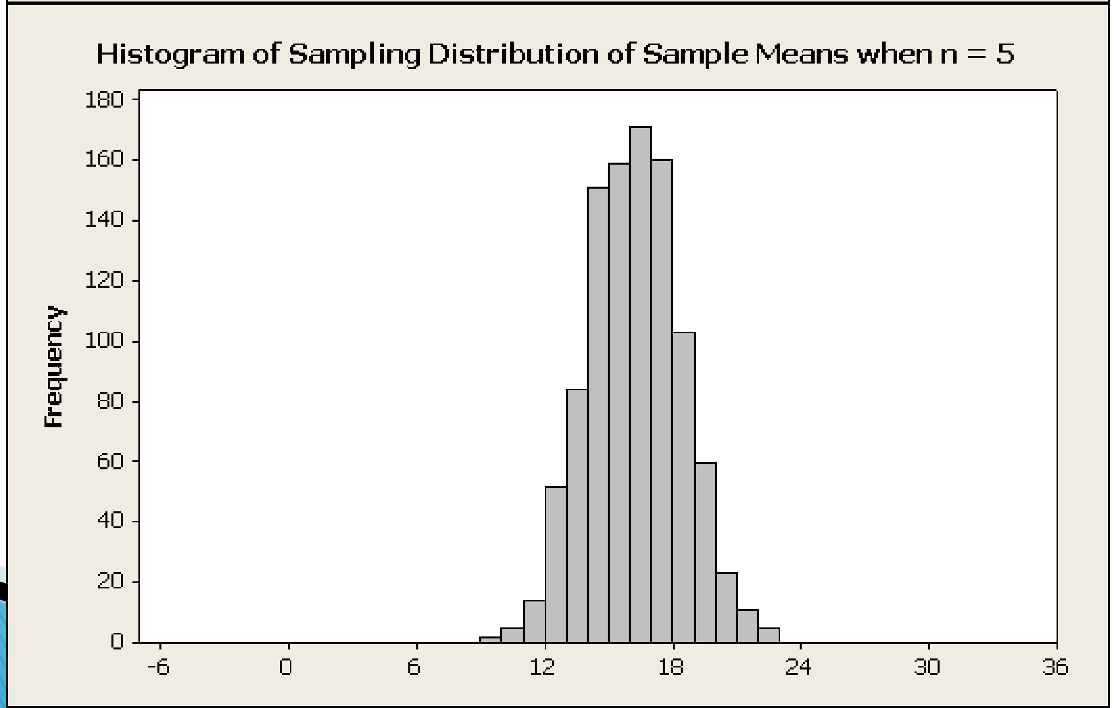
- What does the shape of the sampling distribution depend on?
- What statistical value will be found at the center of the sampling distribution?
- How will the spread of the sampling distribution compare to the spread of the population distribution?
- Does the spread depend on a certain quantity?

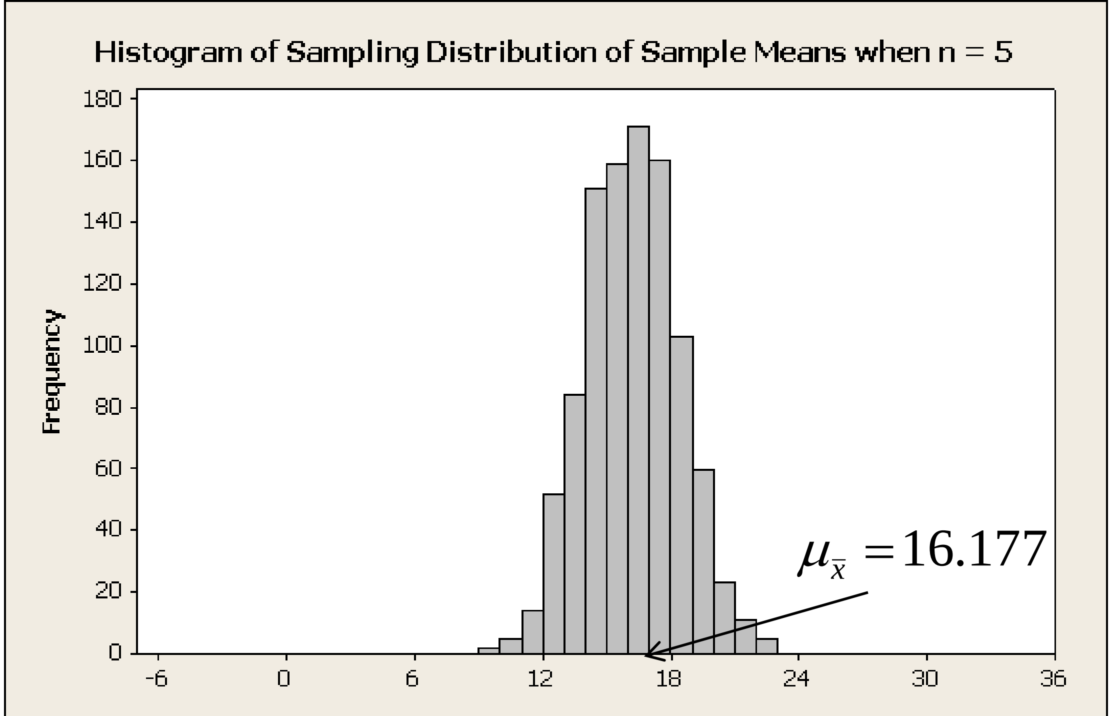
Sampling Distribution example

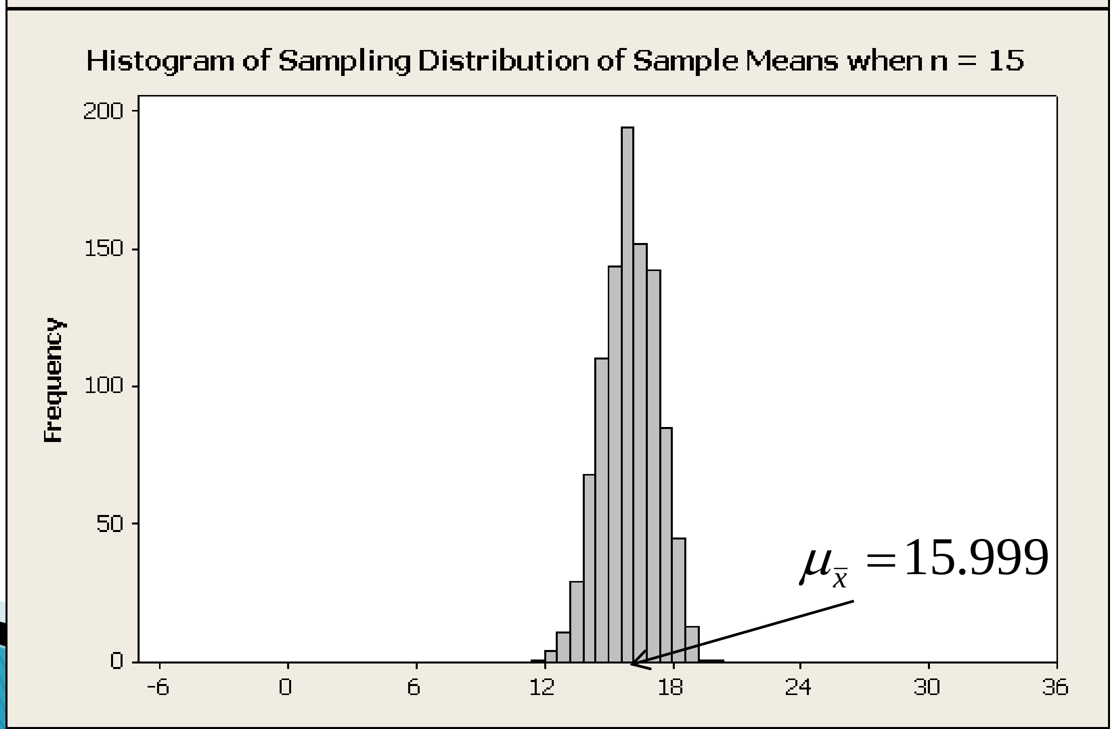
- Our population looks something like this:

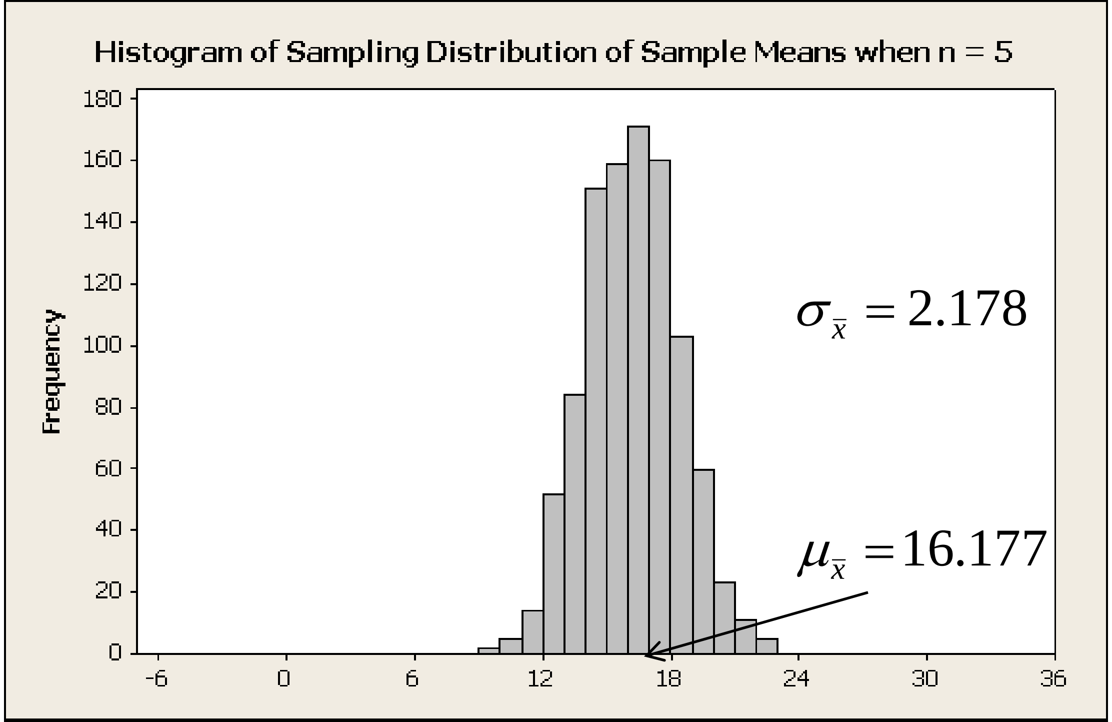


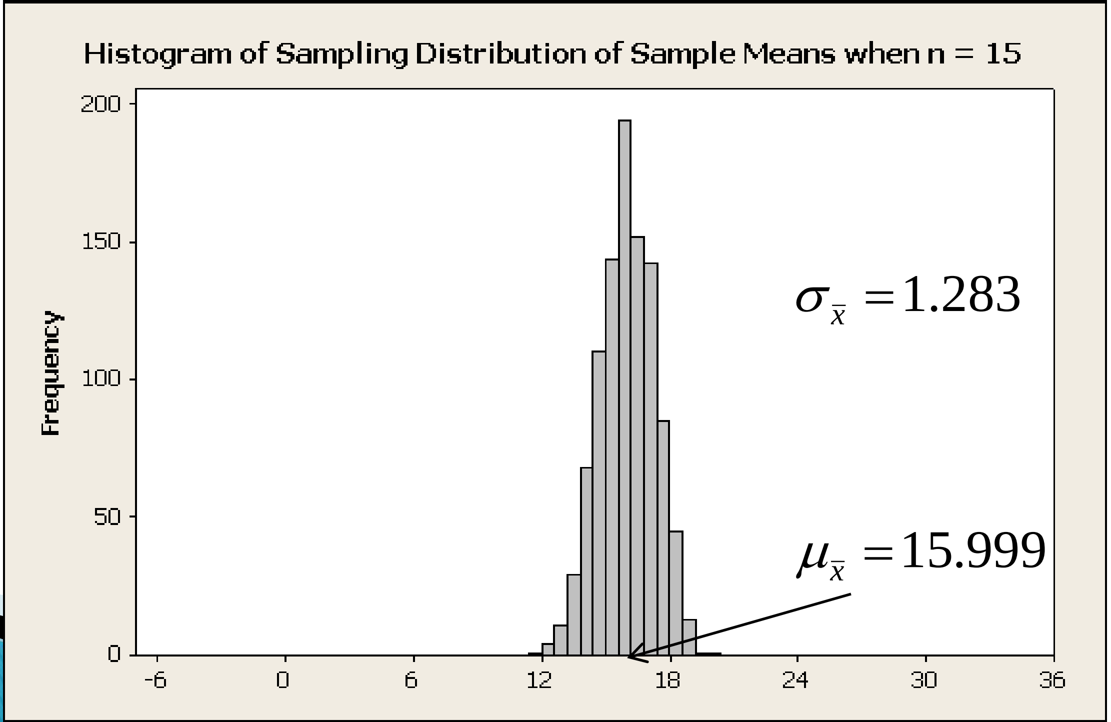

Histogram of the Population of Weights
Frequency
4000
3000
2000
1000
0
-6
0
6
12
18
24
30
36

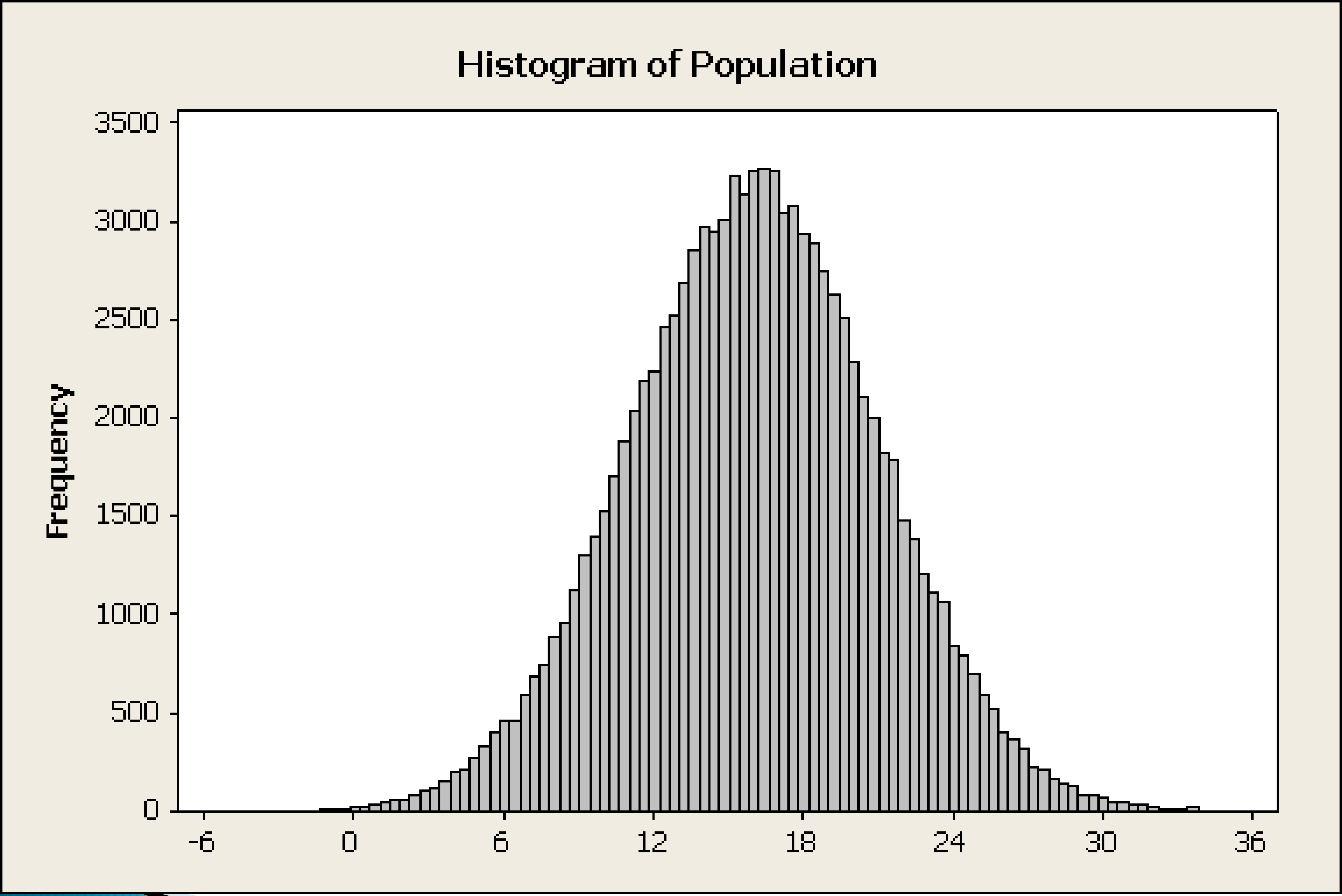

Histogram of Sampling Distribution of Sample Means when n = 5
Frequency
180
160
140
120
100
80
60
40
20
0
-6
0
6
12
18
24
30
36

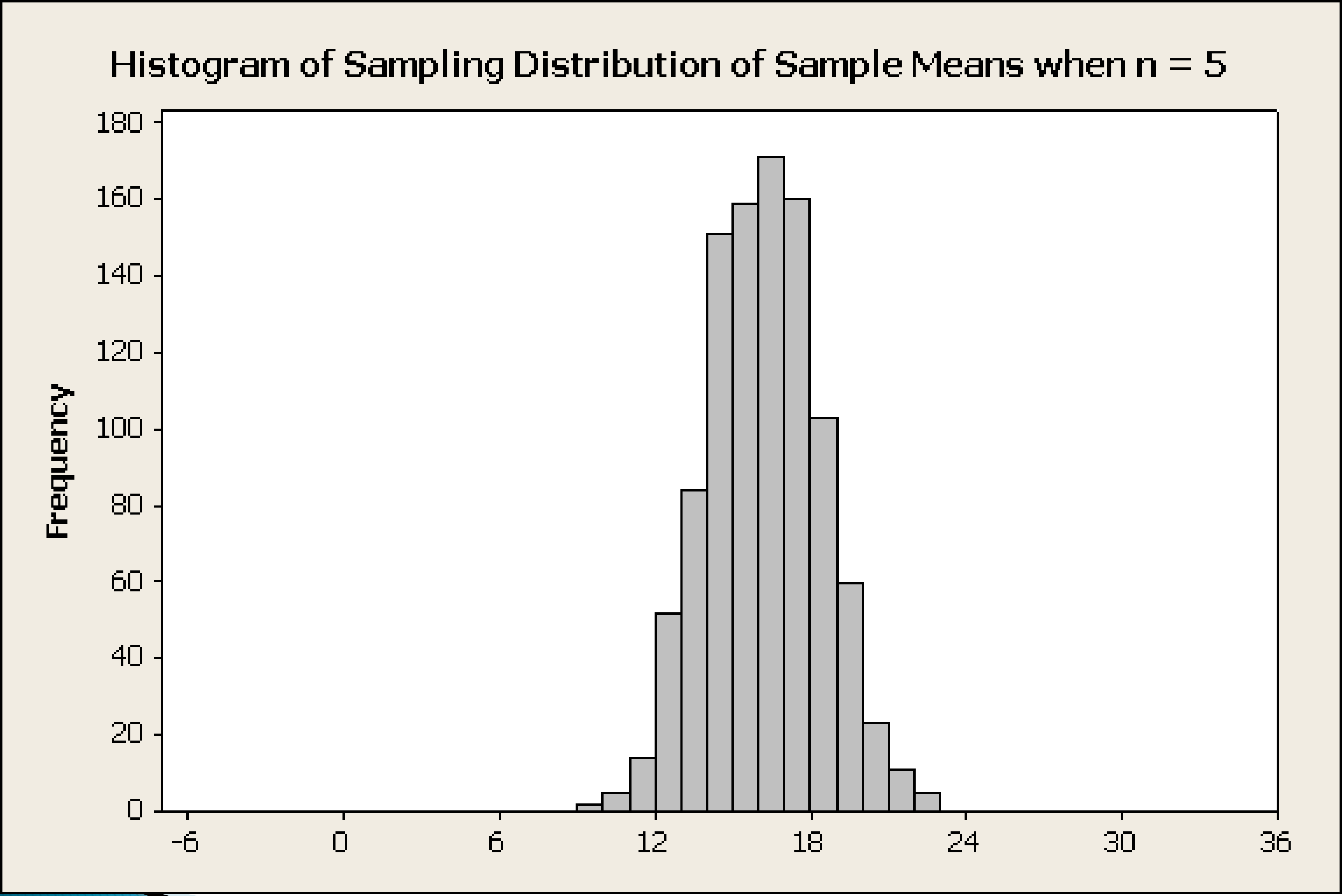

Histogram of Sampling Distribution of Sample Means when n = 5
Frequency
180
160
140
120
100
80
60
40
20
0
-6
0
6
12
18
24
30
36
$\mu_{\bar{x}} = 16.177$

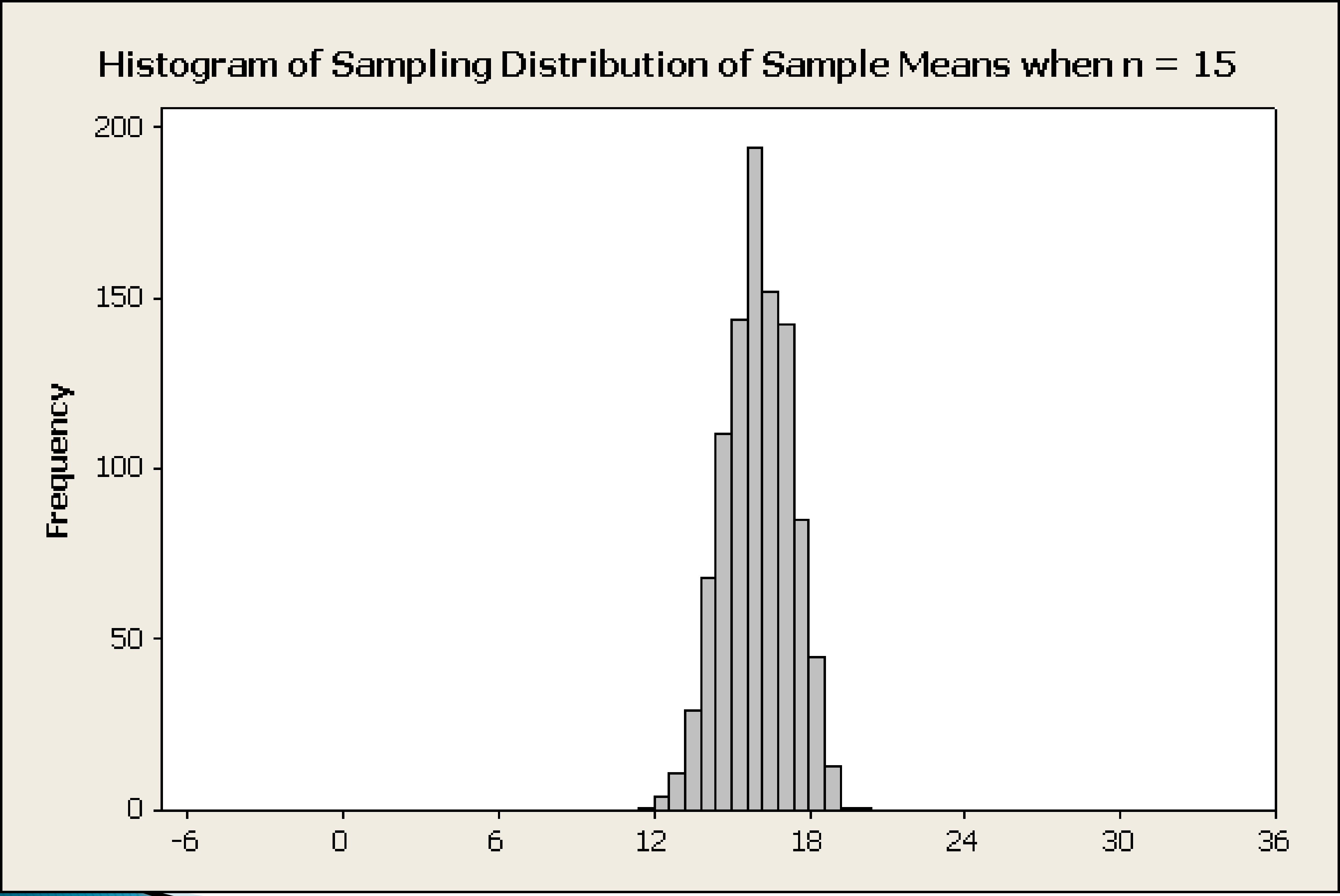

Histogram of Sampling Distribution of Sample Means when n = 15
Frequency
200
150
100
50
0
-6
0
6
12
18
24
30
36
$\mu_{\bar{x}} = 15.999$

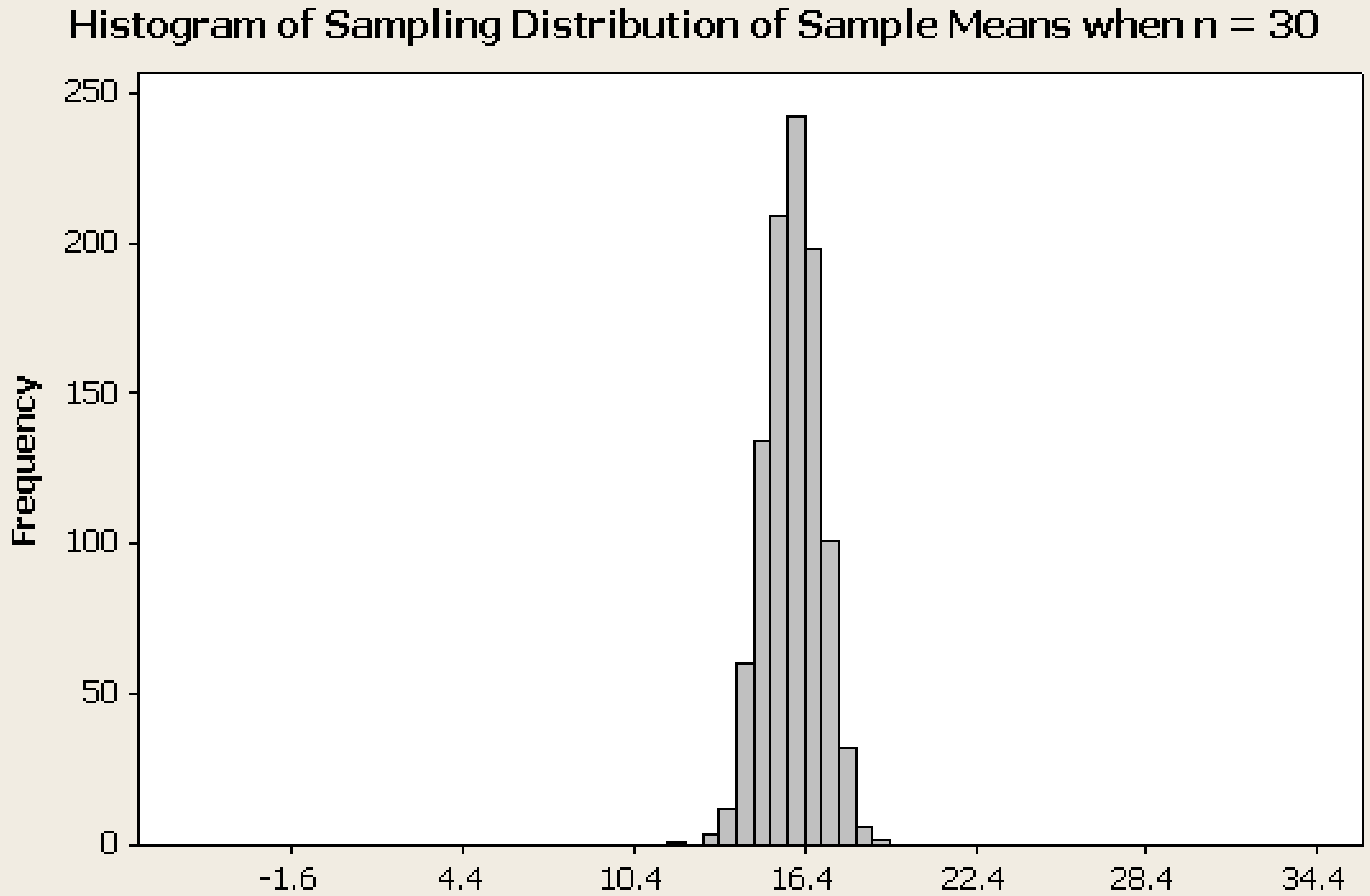

Histogram of Sampling Distribution of Sample Means when n = 5
Frequency
$\sigma_{\bar{x}} = 2.178$
$\mu_{\bar{x}} = 16.177$

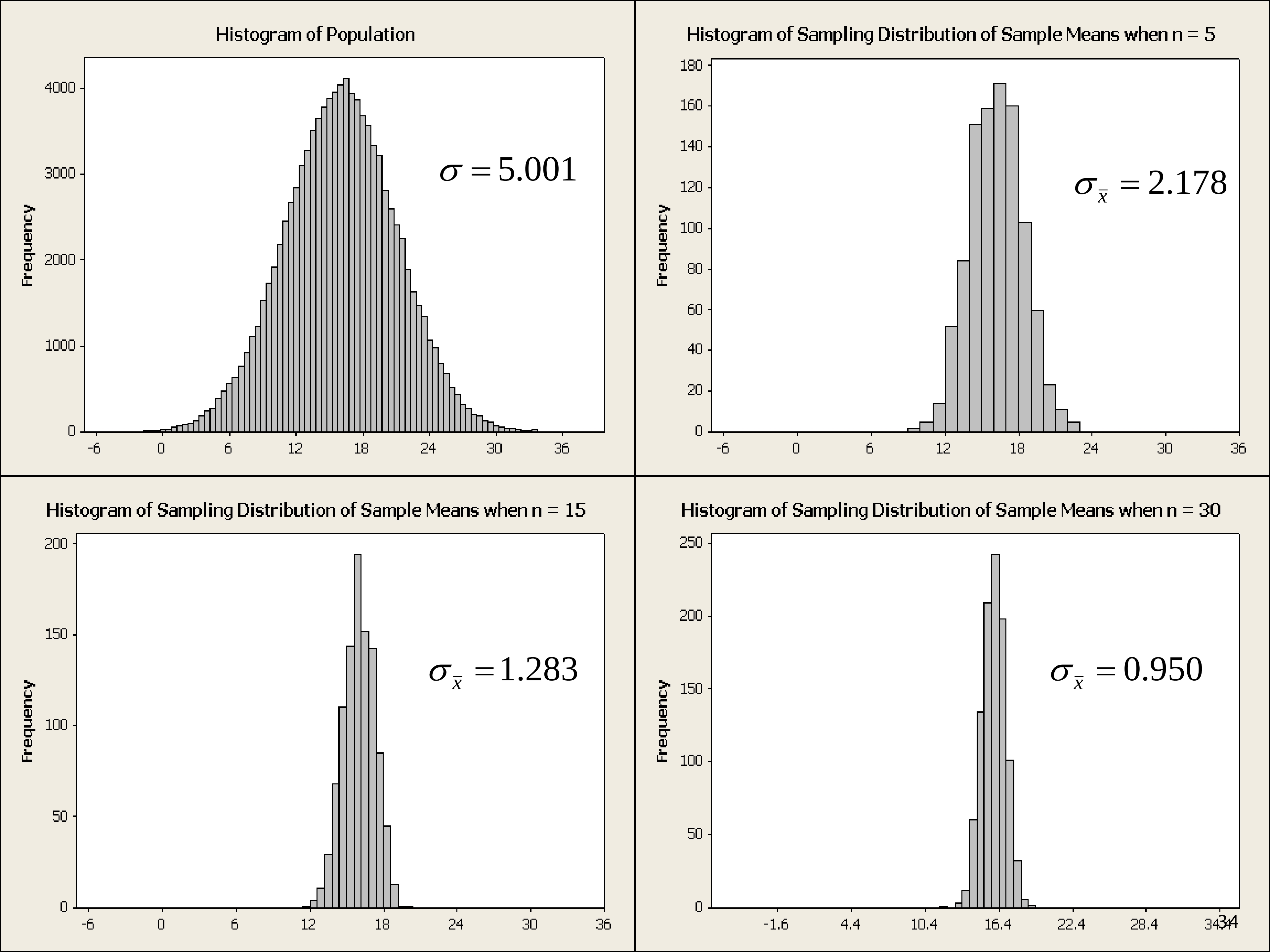

Histogram of Sampling Distribution of Sample Means when n = 15
Frequency
$\sigma_{\bar{x}} = 1.283$
$\mu_{\bar{x}} = 15.999$


Histogram of Population
Frequency
3500
3000
2500
2000
1500
1000
500
0
-6
0
6
12
18
24
30
36


Histogram of Sampling Distribution of Sample Means when n = 5
Frequency
180
160
140
120
100
80
60
40
20
0
-6
0
6
12
18
24
30
36


Histogram of Sampling Distribution of Sample Means when n = 15
Frequency
200
150
100
50
0
-6
0
6
12
18
24
30
36


Histogram of Sampling Distribution of Sample Means when n = 30
Frequency
250
200
150
100
50
0
-1.6
4.4
10.4
16.4
22.4
28.4
34.4


Histogram of Population
$\sigma = 5.001$
Frequency
4000
3000
2000
1000
0
-6
0
6
12
18
24
30
36
Histogram of Sampling Distribution of Sample Means when n = 5
$\sigma_{\bar{x}} = 2.178$
Frequency
180
160
140
120
100
80
60
40
20
0
-6
0
6
12
18
24
30
36
Histogram of Sampling Distribution of Sample Means when n = 15
$\sigma_{\bar{x}} = 1.283$
Frequency
200
150
100
50
0
-6
0
6
12
18
24
30
36
Histogram of Sampling Distribution of Sample Means when n = 30
$\sigma_{\bar{x}} = 0.950$
Frequency
250
200
150
100
50
0
-1.6
4.4
10.4
16.4
22.4
28.4
34.4

Questions and answers (normal population)

- What is the shape of the sampling distribution?
  - Our sampling distribution looks normal.
- What statistical value will be found at the center of the sampling distribution?
  - The mean of the sample means will be very close to the population mean.
- How will the spread of the sampling distribution compare to the spread of the population distribution?
  - The spread of our sampling distribution is smaller than that of the population
- Does the spread depend on a certain quantity?
  - The bigger our sample the smaller the spread of the sampling distribution

From my simulations…

- Population: $\mu = 15.995$, $\sigma = 5.001$
- Sampling distributions when
  - n = 5: $\mu_{\bar{x}} = 16.177, \sigma_{\bar{x}} = 2.178 \approx \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{5}} = 2.236$
  - n = 15: $\mu_{\bar{x}} = 15.999, \sigma_{\bar{x}} = 1.283 \approx \frac{5}{\sqrt{15}} = 1.291$
  - n = 30: $\mu_{\bar{x}} = 15.978, \sigma_{\bar{x}} = 0.950 \approx \frac{5}{\sqrt{30}} = 0.913$

The Standard Error

- The standard error is another name for the spread, or standard deviation, of a sampling distribution

- The Standard Error for a sample mean is found by:

$$\frac{\sigma}{\sqrt{n}}$$

Shape, Center, and Spread

- If our population is normal, the shape of our sampling distribution of the sample mean will be approximately normal regardless of sample size

- The **mean** of the sampling distribution is equal to the population mean μ.

- The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where n is the sample size.

Sampling Distribution of the Sample mean

- A random sample of size *n* is taken from a ***normal population*** with mean μ and variance σ^2.

- A linear function ($\bar{x}$) of normal and independent random variables is itself normally distributed.

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \text{ has a normal distribution}$$

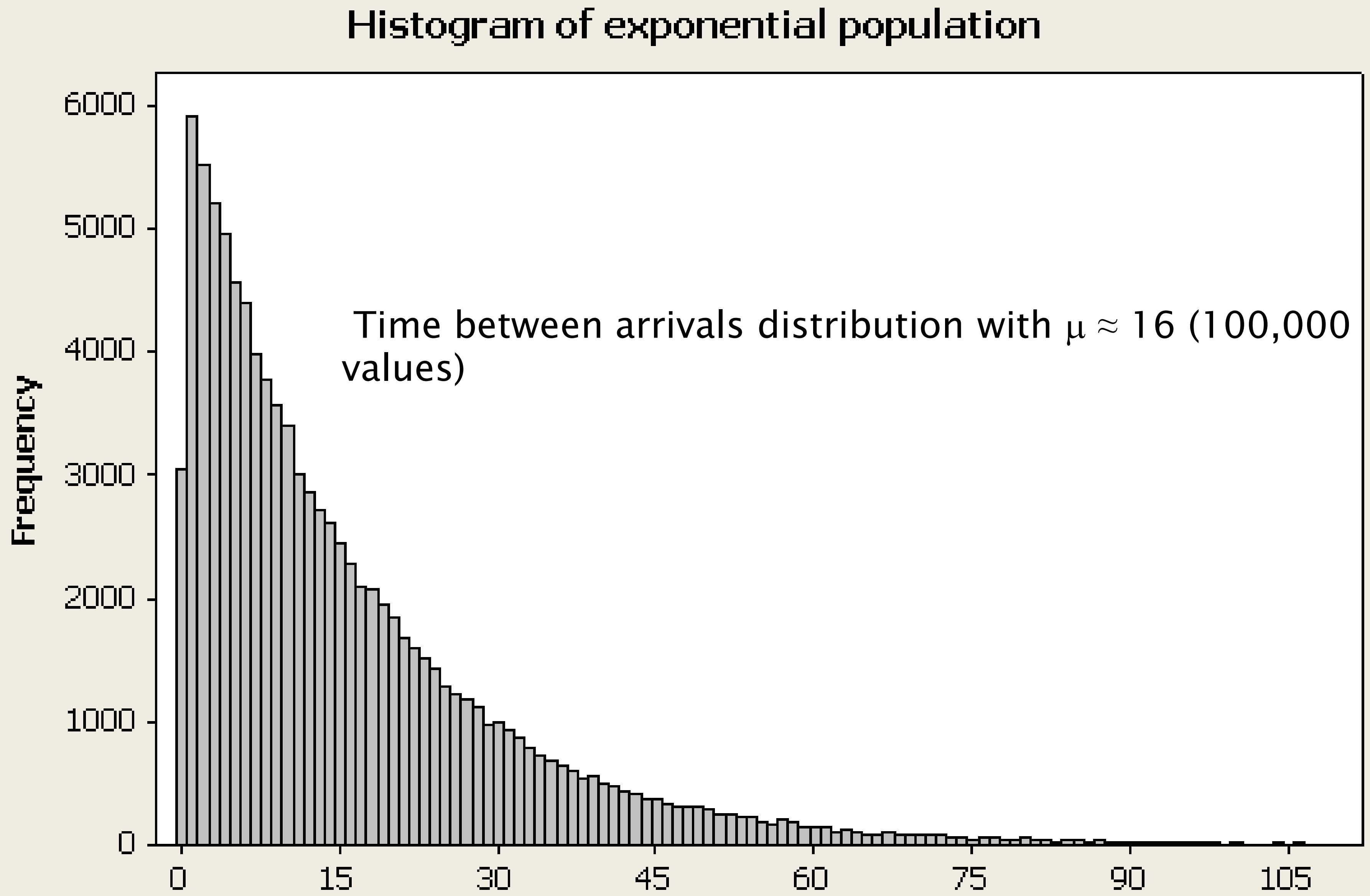
$$\text{with mean } \mu_{\bar{X}} = \frac{\mu + \mu + \ldots + \mu}{n} = \mu$$

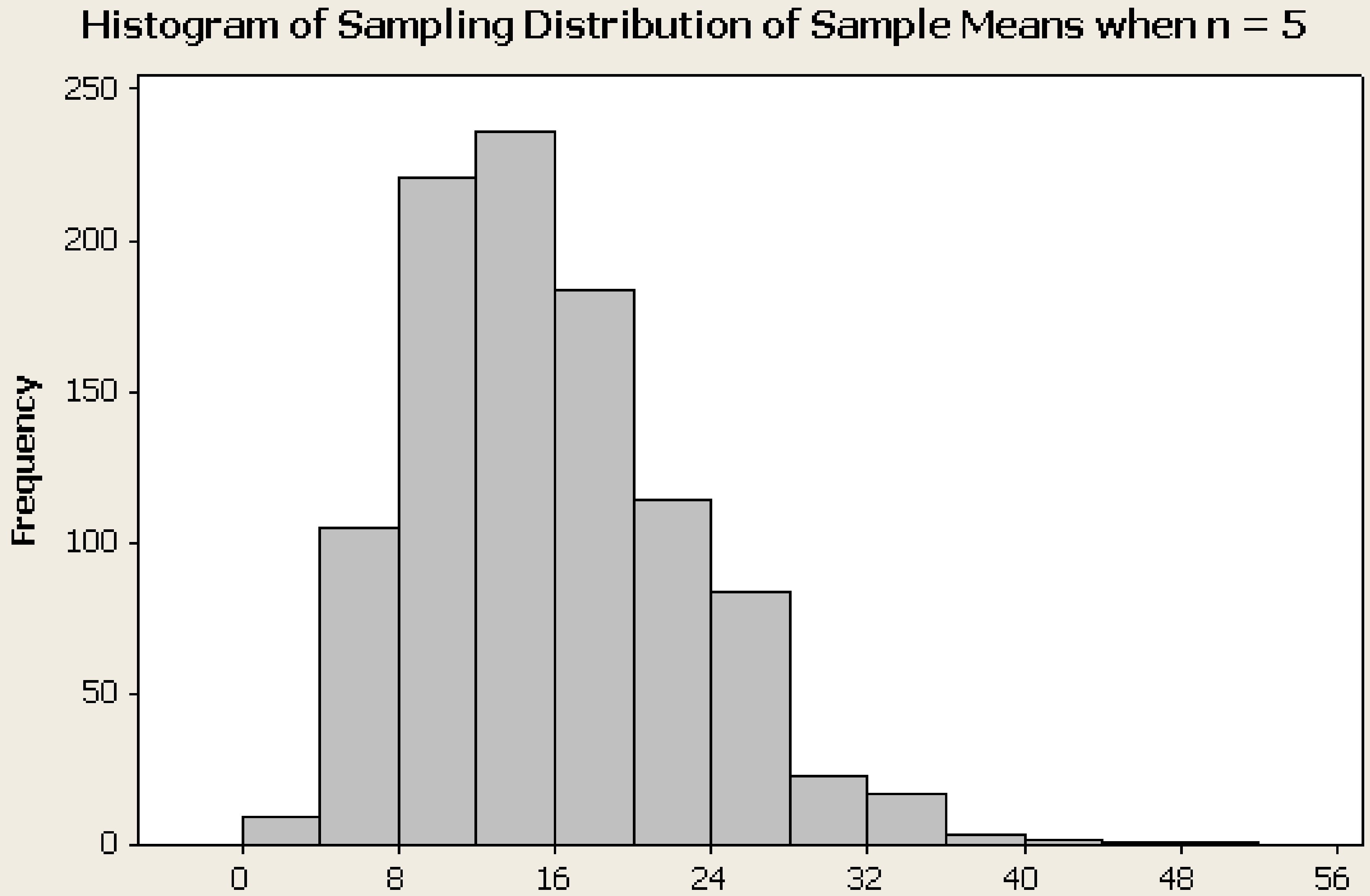
$$\text{and variance } \sigma_{\bar{X}}^2 = \frac{\sigma^2 + \sigma^2 + \ldots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

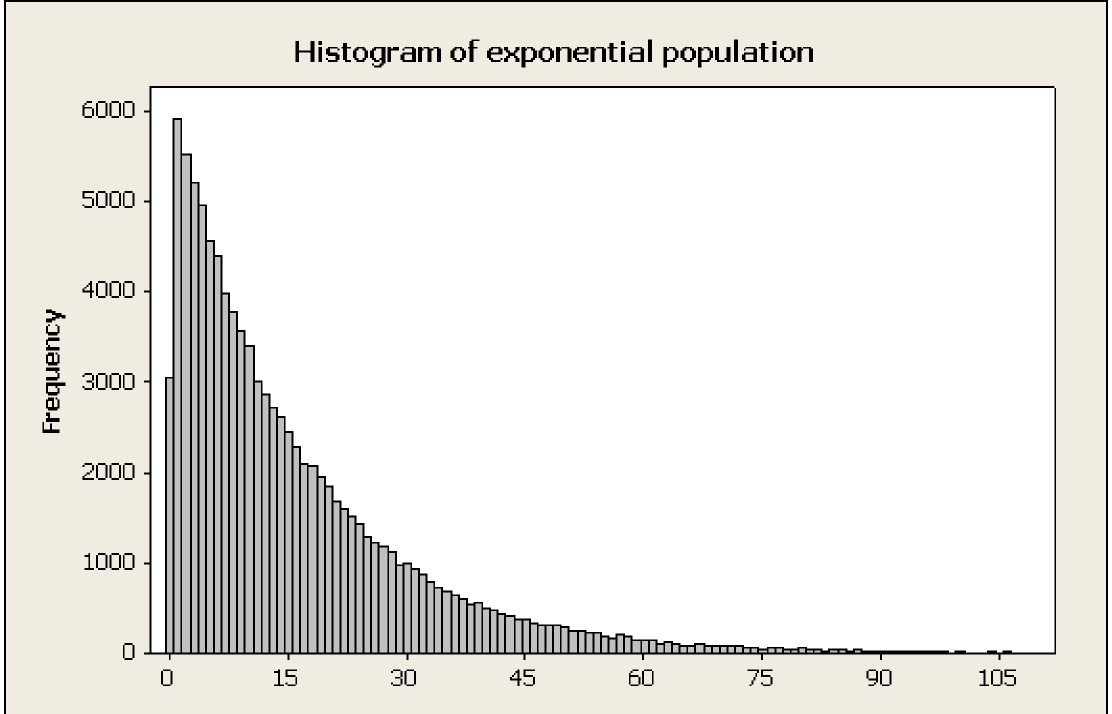
Sampling Distribution of the sample mean (non– normal population)

- Consider the time between arrivals of vehicles at a particular intersection. Assume an exponential distribution with $\mu = \sigma = 16$.

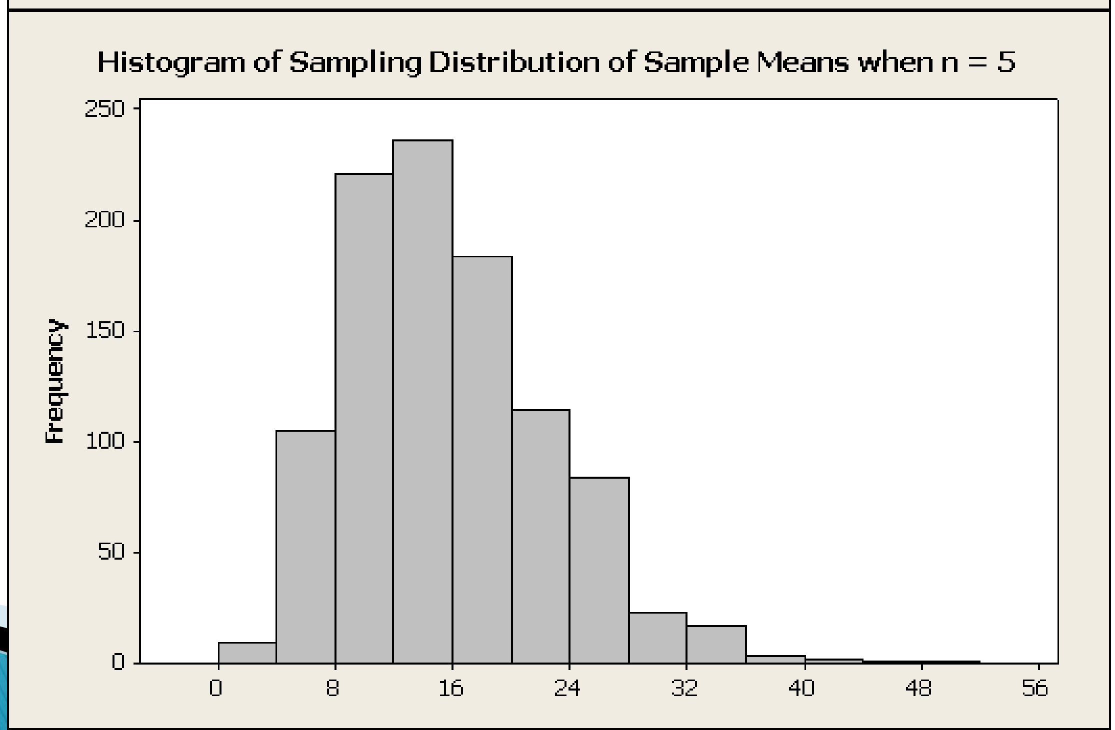
- Same procedure as earlier example (normal)
  - Took 1,000 samples of size 5 from the 100,000 exponential times in Minitab.
  - Calculated 1,000 means
  - Graphed those means in a histogram
  - Repeated this process using n = 15 and n = 30.


Histogram of exponential population
Frequency
6000
5000
4000
3000
2000
1000
0
0
15
30
45
60
75
90
105
Time between arrivals distribution with $\mu \approx 16$ (100,000 values)

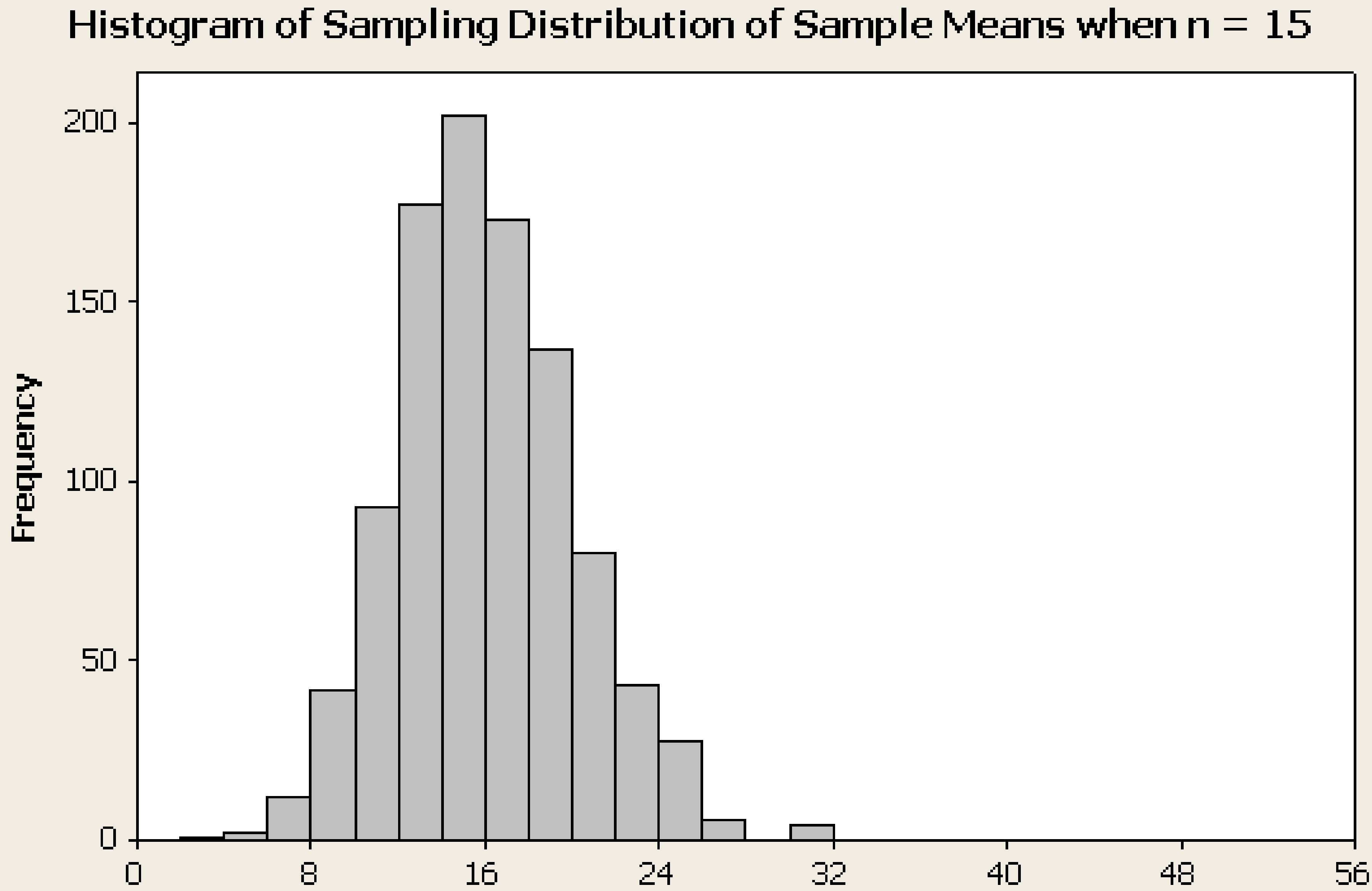

Histogram of Sampling Distribution of Sample Means when n = 5
Frequency
250
200
150
100
50
0
0
8
16
24
32
40
48
56

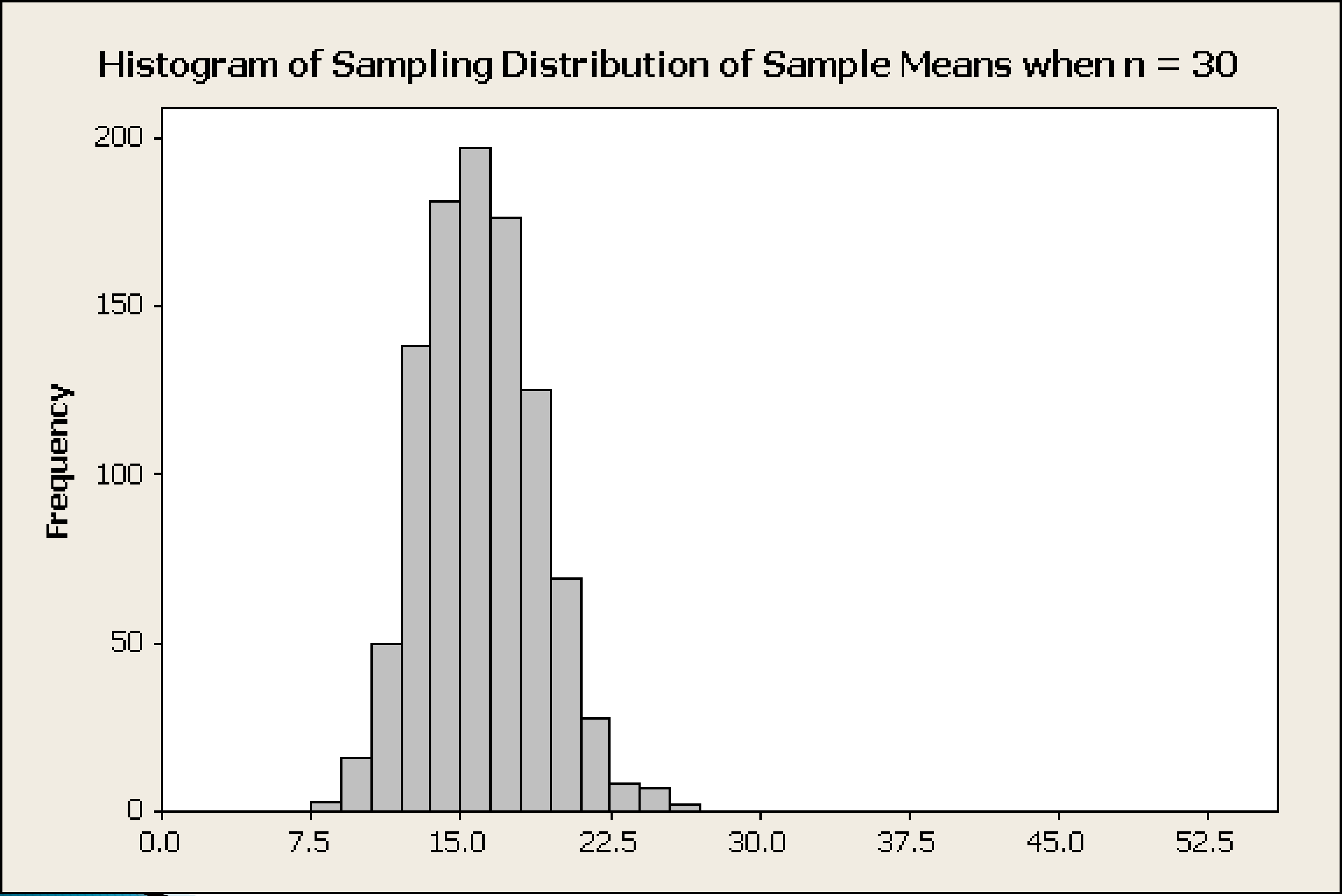


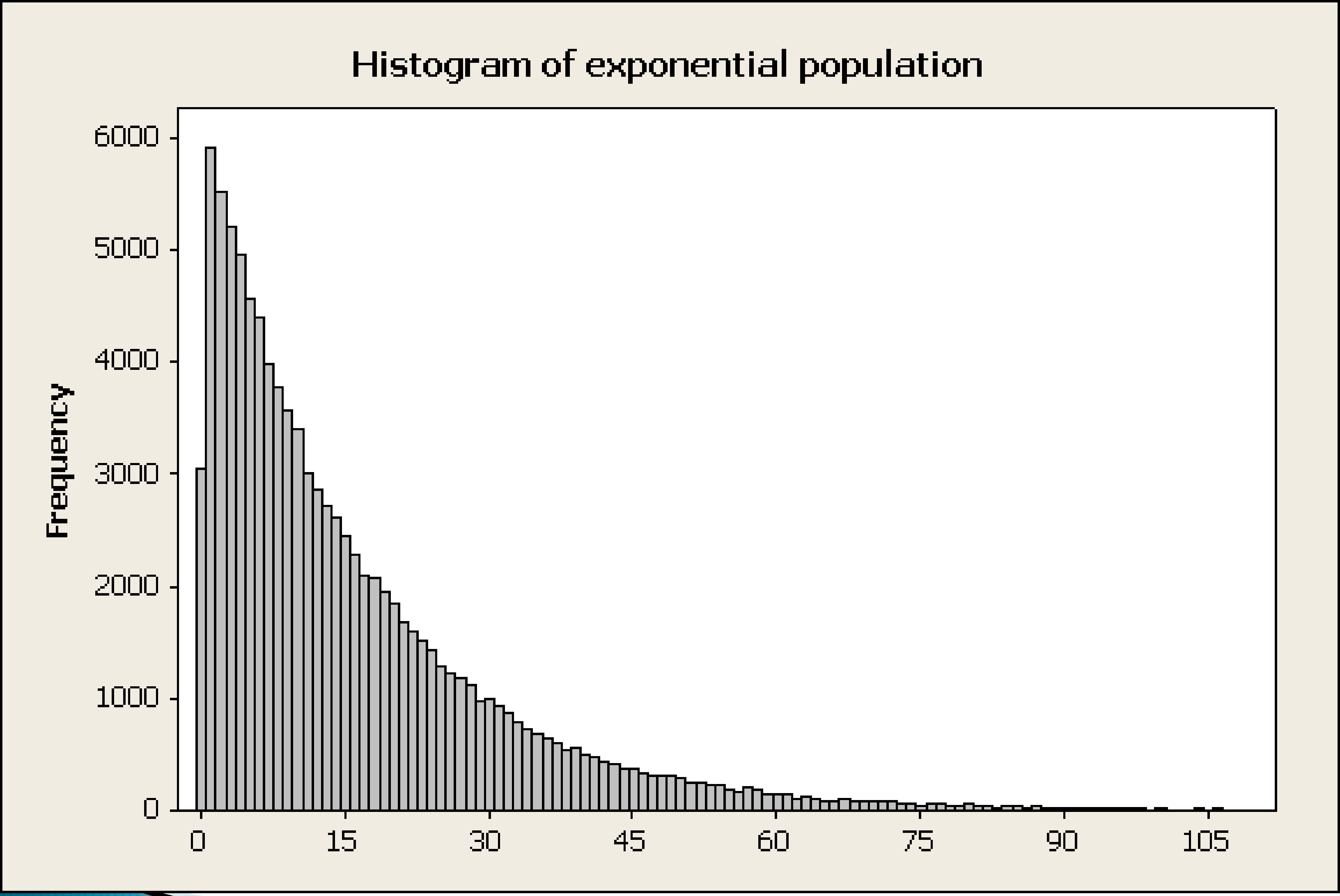
Histogram of exponential population
Frequency
6000
5000
4000
3000
2000
1000
0
0
15
30
45
60
75
90
105

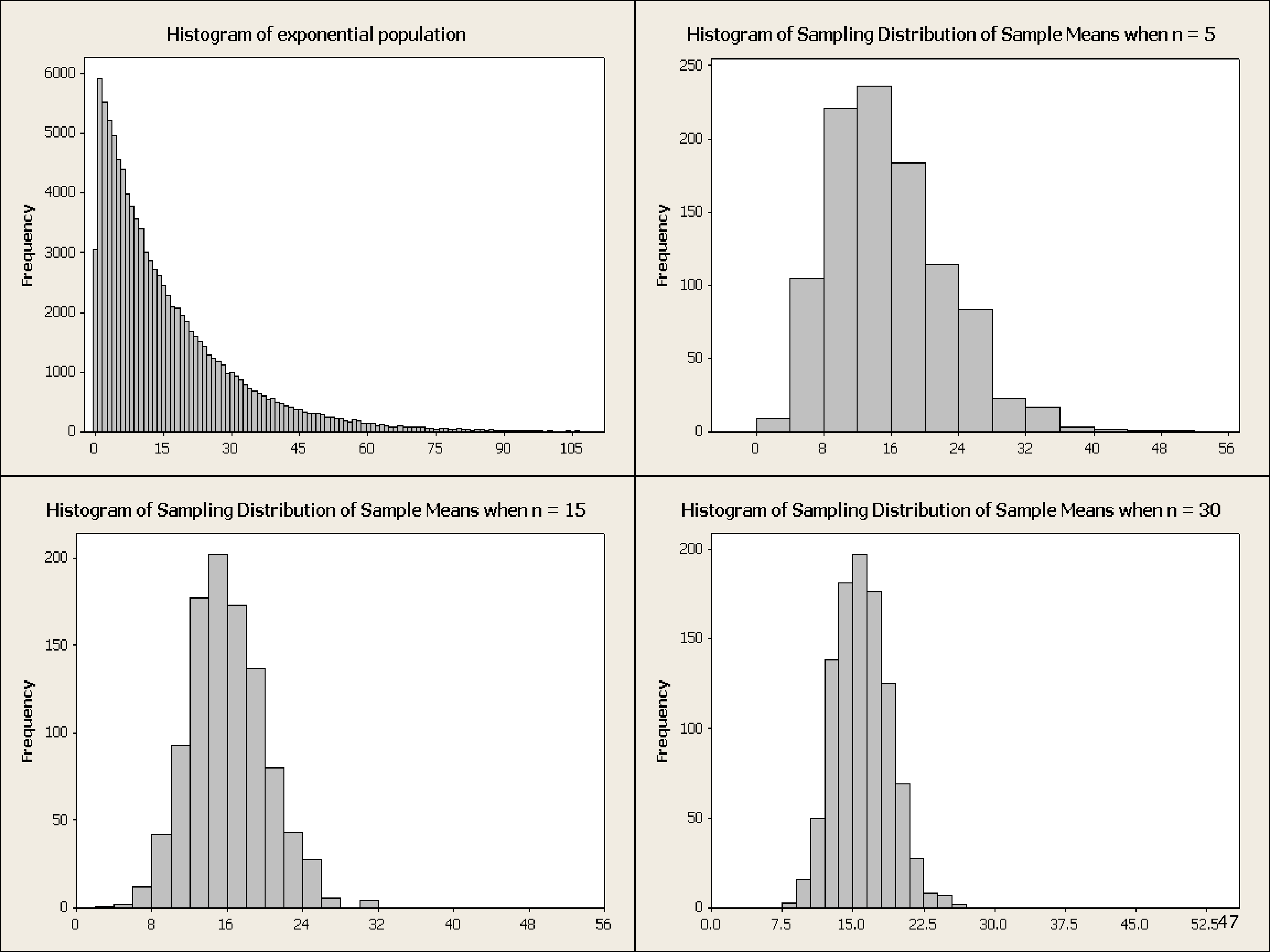


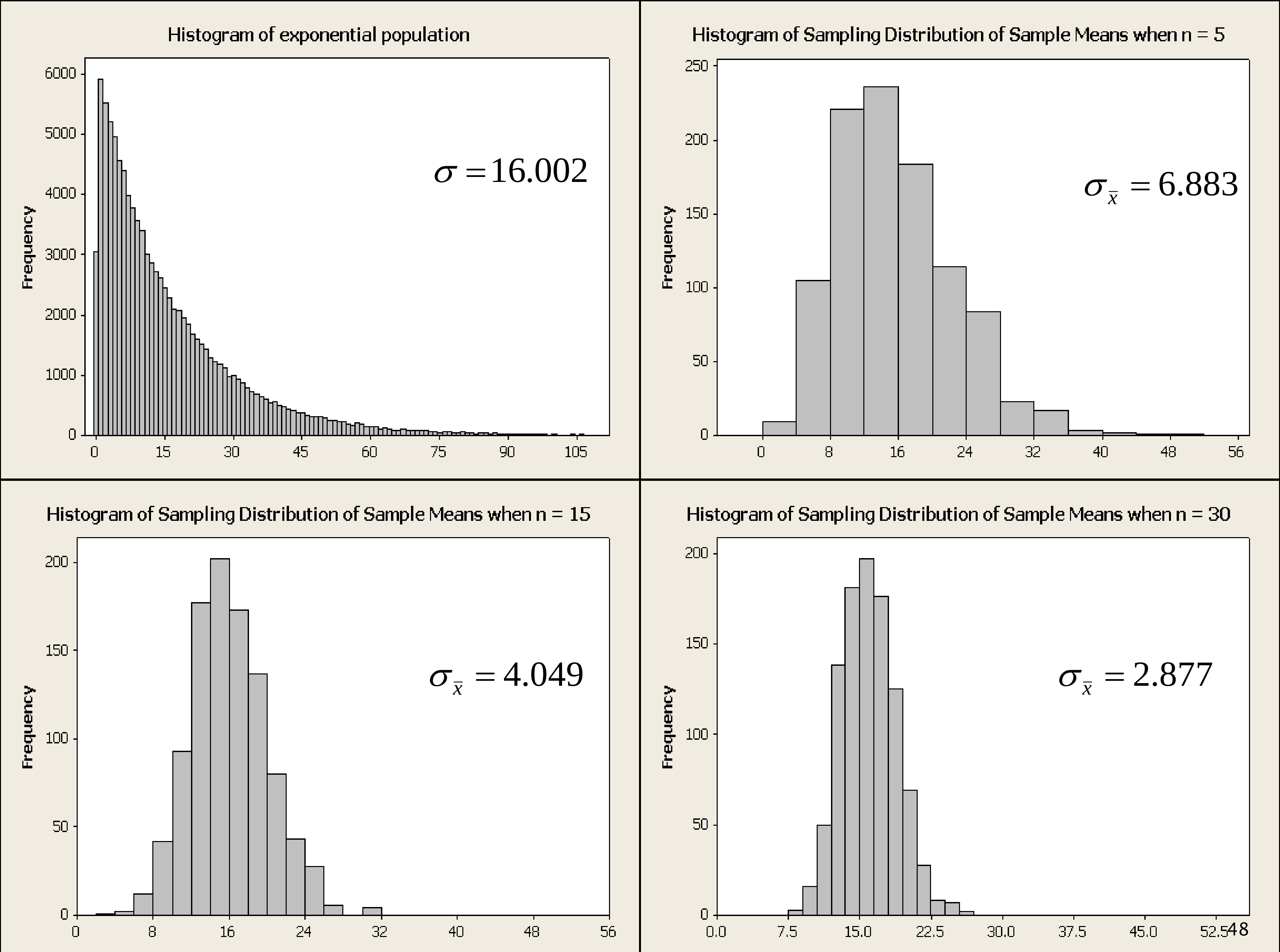
Histogram of Sampling Distribution of Sample Means when n = 5
Frequency
250
200
150
100
50
0
0
8
16
24
32
40
48
56


Histogram of Sampling Distribution of Sample Means when n = 15
Frequency
200
150
100
50
0
0
8
16
24
32
40
48
56


Histogram of Sampling Distribution of Sample Means when n = 30
Frequency
200
150
100
50
0
0.0
7.5
15.0
22.5
30.0
37.5
45.0
52.5


Histogram of exponential population
Frequency
6000
5000
4000
3000
2000
1000
0
0
15
30
45
60
75
90
105


Histogram of exponential population
Frequency
6000
5000
4000
3000
2000
1000
0
0
15
30
45
60
75
90
105
Histogram of Sampling Distribution of Sample Means when n = 5
Frequency
250
200
150
100
50
0
0
8
16
24
32
40
48
56
Histogram of Sampling Distribution of Sample Means when n = 15
Frequency
200
150
100
50
0
0
8
16
24
32
40
48
56
Histogram of Sampling Distribution of Sample Means when n = 30
Frequency
200
150
100
50
0
0.0
7.5
15.0
22.5
30.0
37.5
45.0
52.5


Histogram of exponential population
$\sigma = 16.002$
Frequency
0 1000 2000 3000 4000 5000 6000
0 15 30 45 60 75 90 105
Histogram of Sampling Distribution of Sample Means when n = 5
$\sigma_{\bar{x}} = 6.883$
Frequency
0 50 100 150 200 250
0 8 16 24 32 40 48 56
Histogram of Sampling Distribution of Sample Means when n = 15
$\sigma_{\bar{x}} = 4.049$
Frequency
0 50 100 150 200
0 8 16 24 32 40 48 56
Histogram of Sampling Distribution of Sample Means when n = 30
$\sigma_{\bar{x}} = 2.877$
Frequency
0 50 100 150 200
0.0 7.5 15.0 22.5 30.0 37.5 45.0 52.5

Questions and answers (non-normal population)

- What is the shape of the sampling distribution?
  - Our sampling distribution still looks normal, even more so as our sample size gets large (n=30 yields best results)
- What statistical value will be found at the center of the sampling distribution?
  - The mean of the sample means is still very close to the population mean.
- How will the spread of the sampling distribution compare to the spread of the population distribution?
  - The spread of our sampling distribution is smaller than that of the population.
- Does the spread depend on a certain quantity?
  - The bigger our sample the smaller the spread of the sampling distribution (and the more normal it begins to look)

From my simulations…

- Population: $\mu = 15.967$, $\sigma = 16.002$

- Sampling distributions when
  - n = 5: $\mu_{\bar{x}} = 15.754, \sigma_{\bar{x}} = 6.883 \approx \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{5}} = 7.155$
  - n = 15: $\mu_{\bar{x}} = 16.003, \sigma_{\bar{x}} = 4.049 \approx \frac{16}{\sqrt{15}} = 4.131$
  - n = 30: $\mu_{\bar{x}} = 15.964, \sigma_{\bar{x}} = 2.877 \approx \frac{16}{\sqrt{30}} = 2.921$

Shape, Center, and Spread

- If our population is non-normal, the shape of our sampling distribution of the sample mean will be approximately normal depending on the sample size of each sample (we'll use n $\geq$ 30)

- The **mean** of the sampling distribution is equal to the population mean μ.

- The **standard deviation** of the sampling distribution is $\sigma/\sqrt{n}$, where *n* is the sample size.

Central Limit Theorem

When randomly sampling from any population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ is approximately normal with mean = μ and s.d = $\frac{\sigma}{\sqrt{n}}$ when the sample size, n, is "sufficiently large".

- Note: In general, we can assume n ≥ 30 is "sufficiently large"

Implications of the CLT

1.) If the rv X is normal the distribution of the sample means is normal no matter what sample size is taken.

If our Population ~ N(μ,σ)

Then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ for **ANY** n.

2.) If the rv X is non-normal, the distribution of sample means is approximately normal for a "sufficiently large" sample size ($n \geq 30$)

If our Population ~ ?(μ,σ) aka non-normal

Then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ **IF** $n > 30$

Implications of the CLT cont…

- How does all of that help us?

- We can assume normality of the sampling distribution and standardize to find probabilities about the sample mean

$$\text{In words}: z = \frac{(\text{value} - \text{mean})}{\text{standard deviation}}$$

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Implications of the CLT

There are certain types of problems that we now can do assuming the CLT holds:

- Find probabilities associated with a **single individual** from a **Normal** Population (already know)
- Find probabilities associated with a **small** sample from a **Normal** Population
- Find probabilities associated with a **large** sample from a **Normal** Population
- Find probabilities associated with a **large** sample from a **Non-Normal** Population

Implications of the CLT

Can't Do (Yet):

- Find probabilities associated with a **single** individual from a **Non-Normal** Population
- Find probabilities associated with a **small** sample from a **Non-Normal** Population

CLT example 1

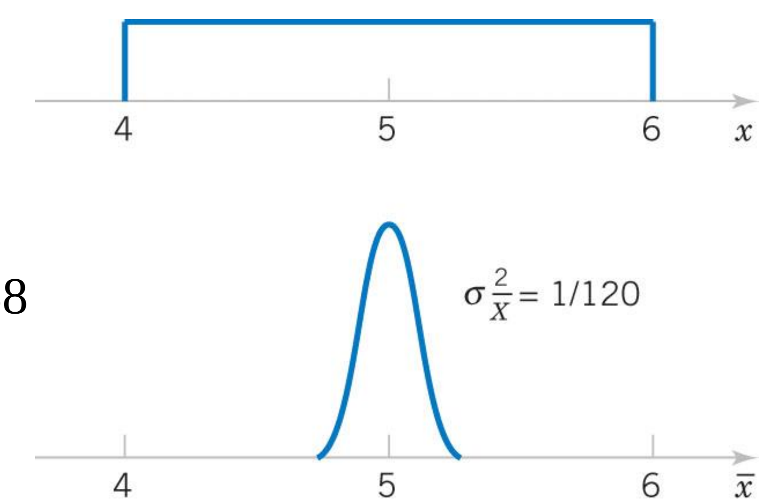
- Suppose that a random variable X has a continuous uniform distribution:
$$f(x)=\begin{cases} 1/2, & 4\le x\le 6 \\ 0, & \text{otherwise} \end{cases}$$
- Describe the distribution of the sample mean of a random sample of size $n = 40$

Distribution is normal by the CLT.

$$\mu=\frac{b+a}{2}=\frac{6+4}{2}=5.0$$

$$\sigma=\sqrt{\frac{(b-a)^2}{12}}=\sqrt{\frac{(6-4)^2}{12}}=\sqrt{\frac{1}{3}}=0.58$$

$$\sigma_{\overline{X}}=\sqrt{\frac{\sigma^2}{n}}=\sqrt{\frac{1/3}{40}}=\sqrt{\frac{1}{120}}=0.09$$



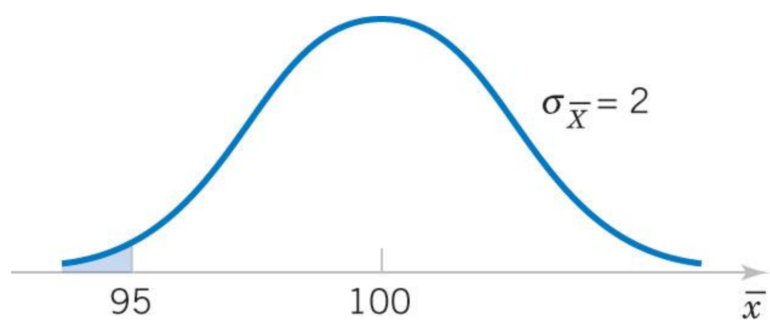
CLT Example 2

- An electronics company manufactures resistors having a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. What is the probability that a random sample of $n = 25$ resistors will have an average resistance of less than 95 ohms?

$$\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2.0$$

$$P\left(\bar{X} < 95\right) = \Phi\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}\right) = \Phi\left(\frac{95 - 100}{2}\right)$$

$$= \Phi\left(-2.5\right) = 0.0062$$



0.0062	= NORMSDIST(-2.5)
A rare event at less than 1%.	

Checking Normality Visually

- Many times, we need to know if a sample seems to come from a Normal Distribution.

- There are numerical ways to do this, but now we will focus on Visual Methods. You could use:
  - Histograms with overlaying Normal density curves
    - Could be issues w/ sample size, bin size, etc…
    - Outliers not always obvious
  - Normal Probability (Q-Q) plots
    - Best option

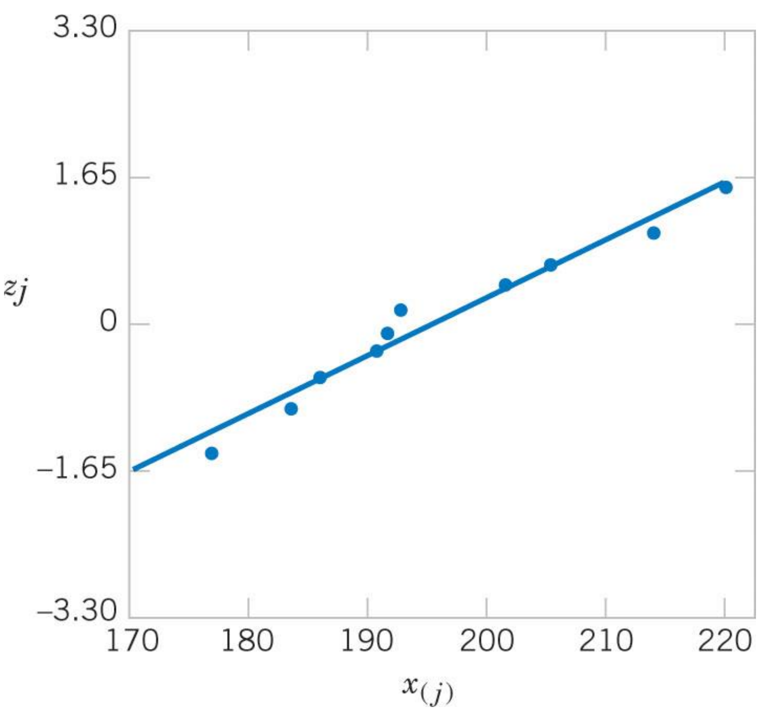
How To Build a Probability Plot

To construct a probability plot:

- Sort the data observations in ascending order: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$.
- Pair each observation with either it's quantile or it's Z score (for normal)
- If the paired numbers form a straight line, it is reasonable to assume that the data follows the proposed distribution.

Probability Plot on Ordinary Axes

A normal probability plot can be plotted on ordinary axes using z-values. The normal probability scale is not used.

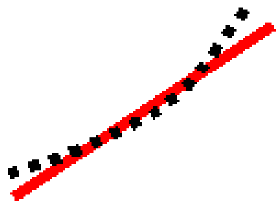


Calculations for Constructing a Normal Probability Plot			
j	$x_{(j)}$	$(j-0.5)/10$	z_j
1	176	0.05	-1.64
2	183	0.15	-1.04
3	185	0.25	-0.67
4	190	0.35	-0.39
5	191	0.45	-0.13
6	192	0.55	0.13
7	201	0.65	0.39
8	205	0.75	0.67
9	214	0.85	1.04
10	220	0.95	1.64

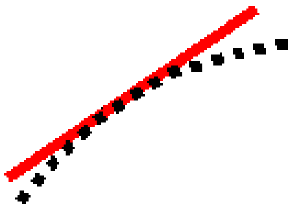
Use of the Probability Plot

- The probability plot can identify variations from a normal distribution shape.
  - Light (short) tails of the distribution – more peaked.
  - Heavy (Long) tails of the distribution – less peaked.
  - Skewed distributions can also be identified
- Larger samples increase the clarity of the conclusions reached.

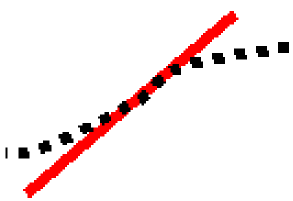
Probability Plot Variations



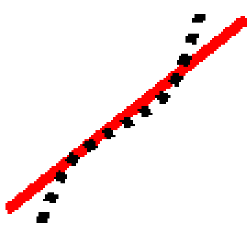
Right Skew - If the plotted points appear to bend up and to the left of the normal line that indicates a long tail to the right.



Left Skew - If the plotted points bend down and to the right of the normal line that indicates a long tail to the left.



Short Tails - An S shaped-curve indicates shorter than normal tails, i.e. less variance than expected.



Long Tails - A curve which starts below the normal line, bends to follow it, and ends above it indicates long tails. That is, you are seeing more variance than you would expect in a normal distribution.

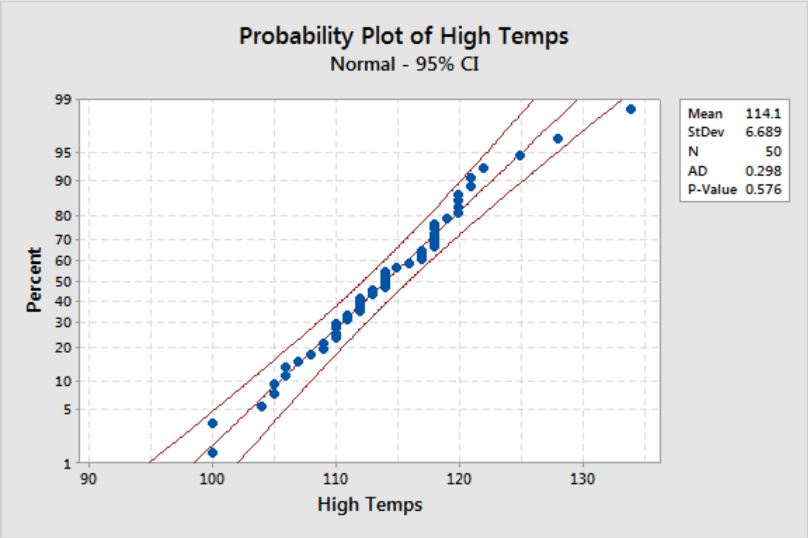
Probability Plot Minitab Example

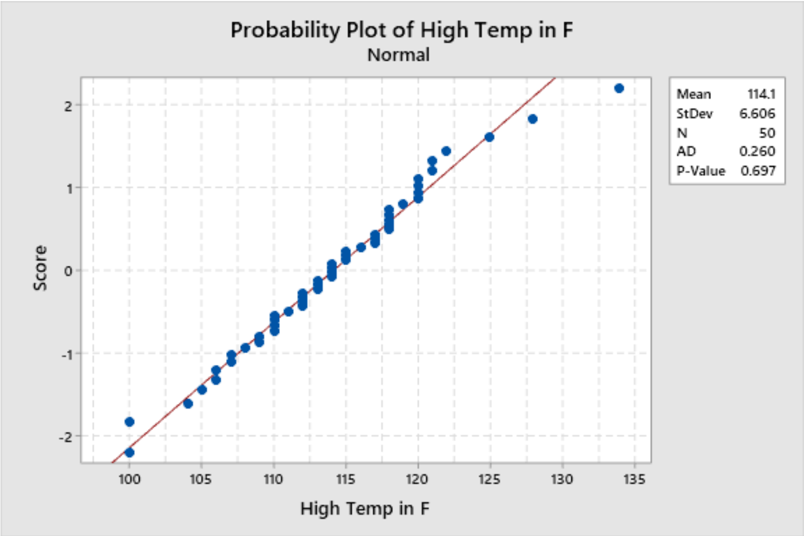
Default:

- Using High Temps Data
- Graph > Probability Plot > single > Choose Data
  - Distribution button > choose distribution (normal here)

Edits:

- Distribution Button-> Data Display Tab -> Uncheck Show C.I.
- Scale Button-> Y-Scale-> Score





Behavior of Means (σ unKnown)

- So we can use one sample mean to find probabilities. However, there is one problem…

- We most likely will not have knowledge of the population standard deviation σ, but we can estimate it.

Estimate σ

- Estimate the population standard deviation σ with the sample standard deviation, s.
- s is known to be a good estimate of σ.
- s is a statistic calculated from the sample data.

Formulas

- Population standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \mu\right)^2}{N}}$$

- Sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Information

- There exists only one value of σ for a population.

- Each sample one takes produces another different sample standard deviation, s.
  - S has it's own sampling distribution which we will discuss later

- s is an unbiased estimate of σ only when we divide by n – 1 in the formula.
  - We have n – 1 degrees of freedom

Conditions for Inference with t

- Data still need to be collected randomly and independently

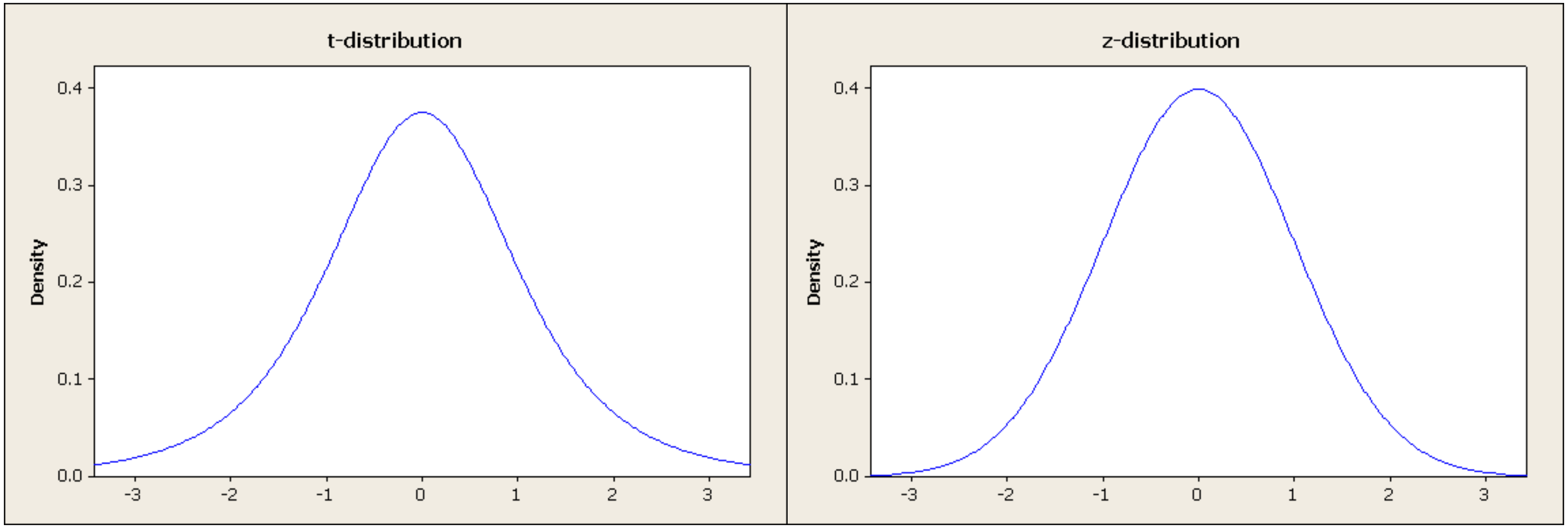
- In addition, either the population must be normally distributed or the sample size must be fairly large ($n \geq 30$).

- However, since we are estimating σ with s, we need to introduce a new distribution to use for inference.

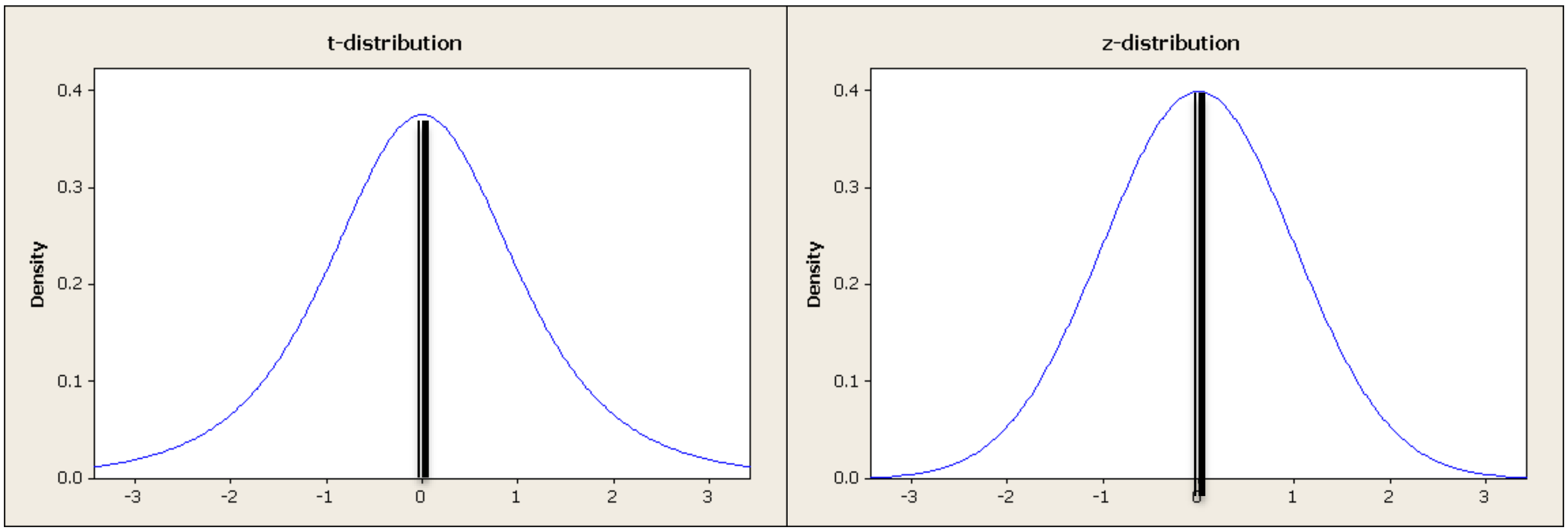
t-distribution

- Used when σ is unknown.
- Family of t-distributions that depend on degrees of freedom $(n - 1)$.
- There is a different t-distribution curve for each degree of freedom.

Compare t and z distributions



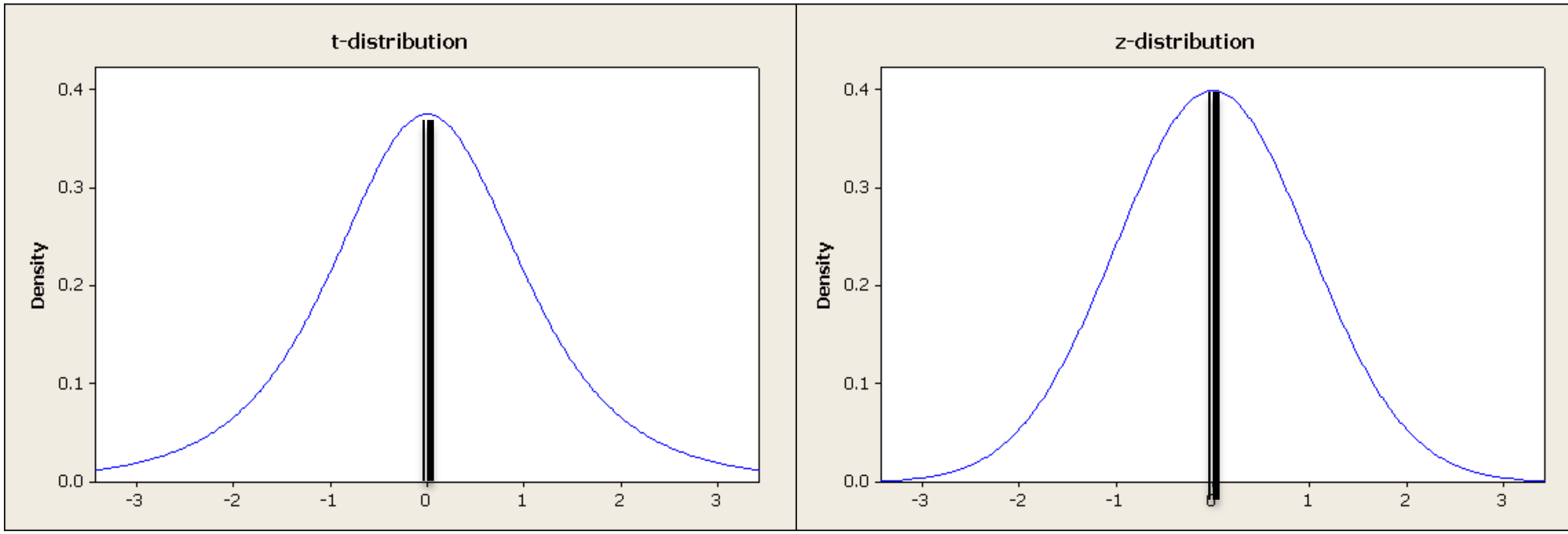
Compare t and z distributions



<u>Similarities</u>

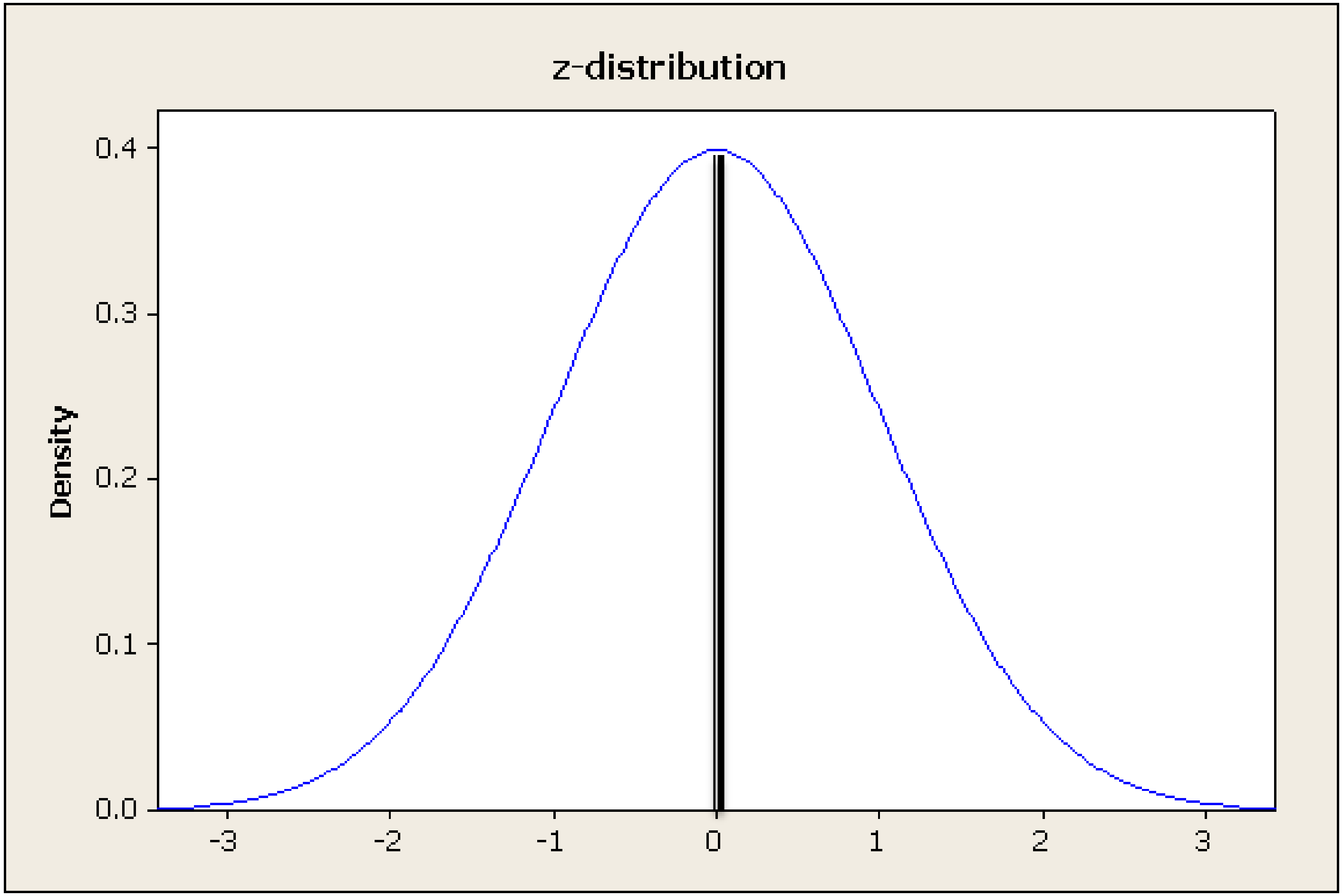
- Bell Shaped
- Symmetrical
- Centered at 0

Compare t and z distributions

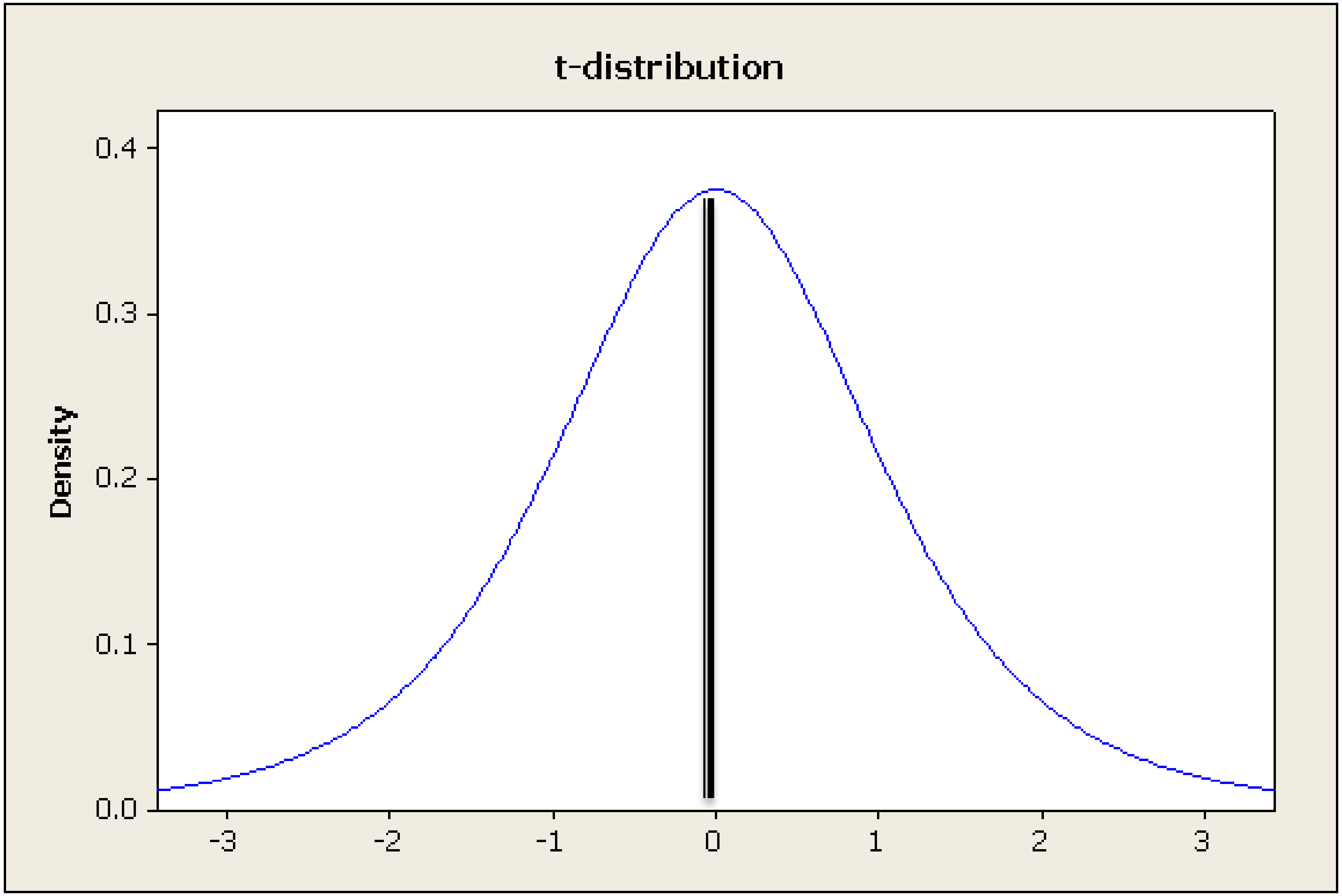


Differences

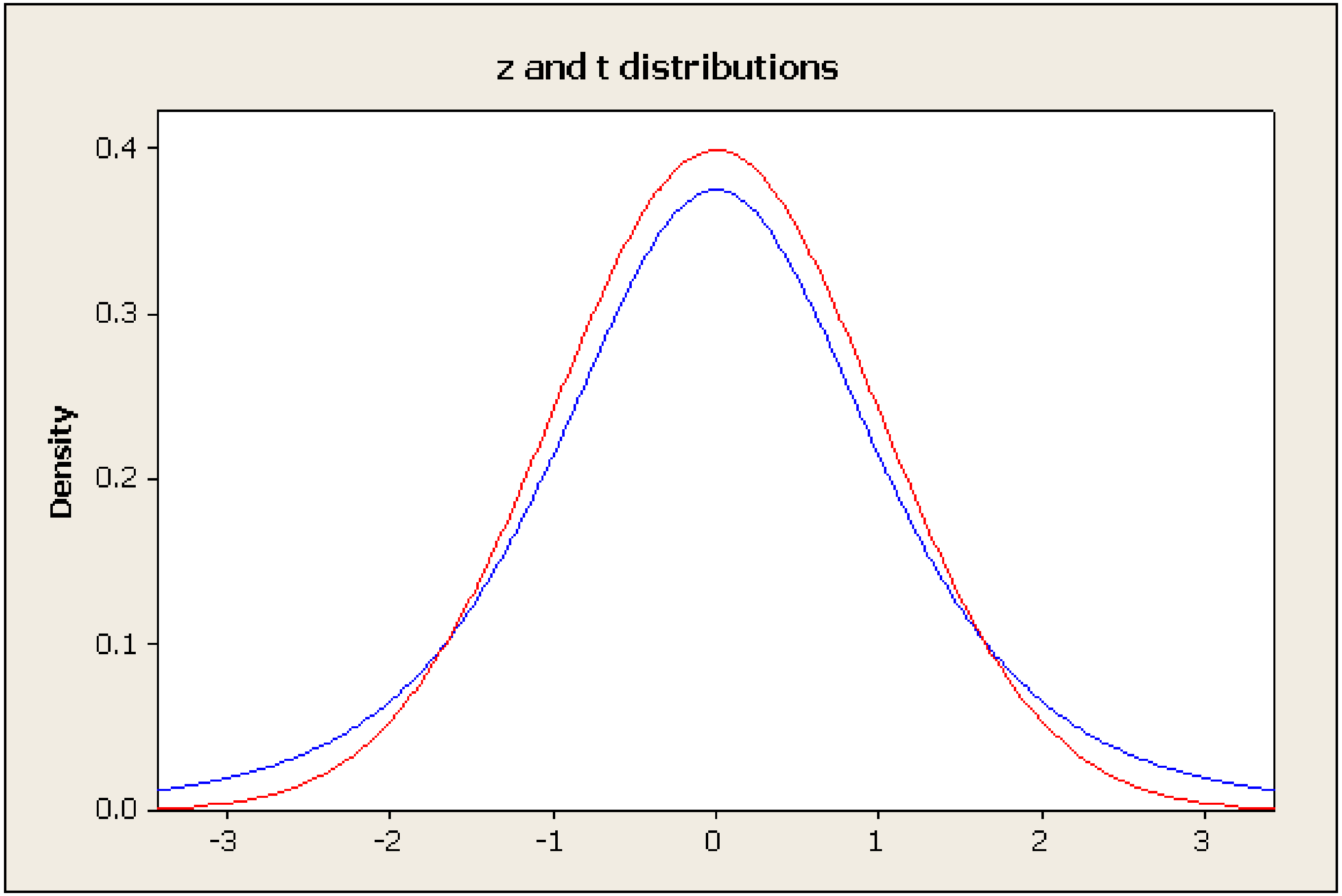
- t distribution is more spread out than z distribution
- Standard deviation of t distribution > 1
- More area in the tails of the t-distribution



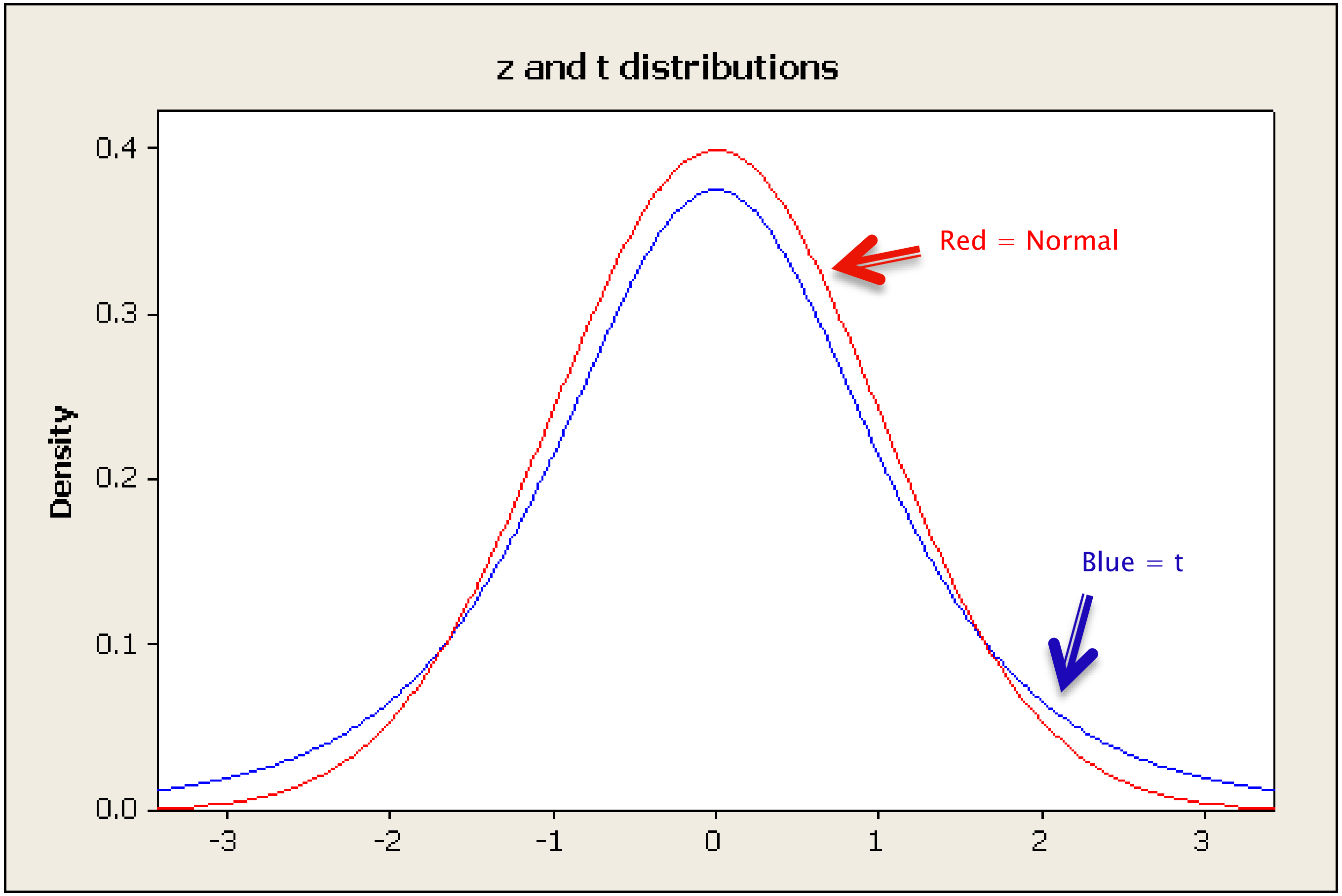
z-distribution
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3



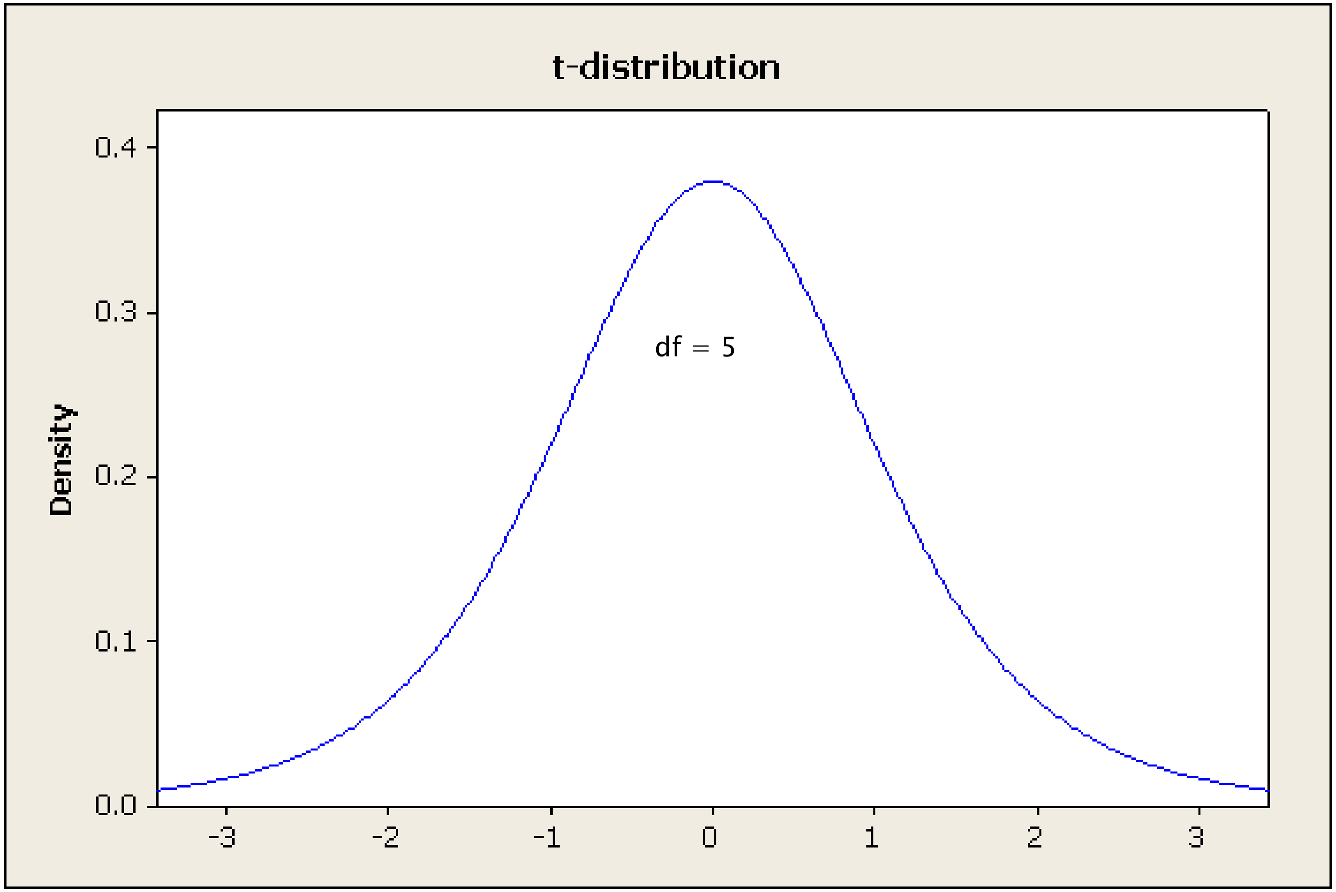
t-distribution
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3



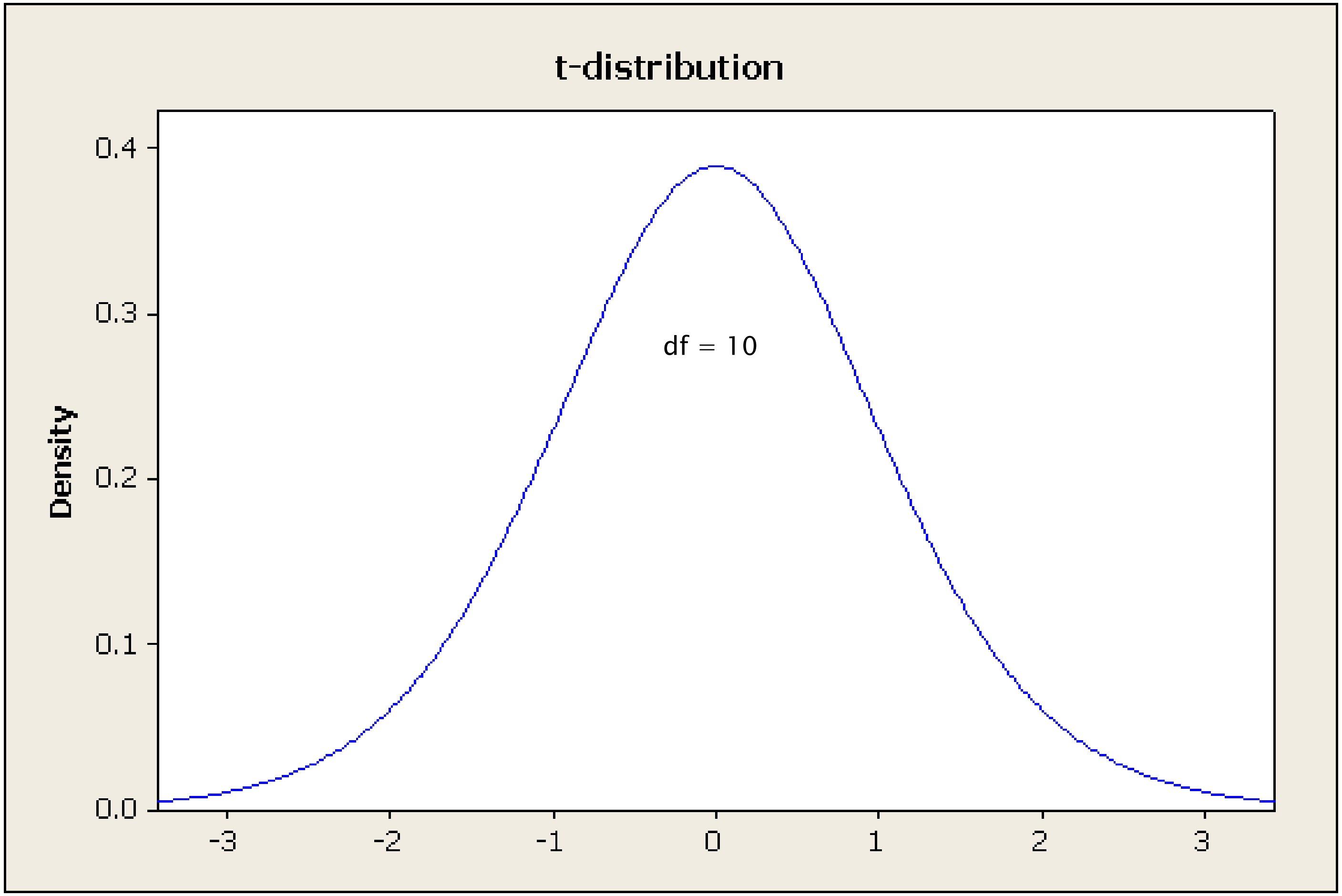
z and t distributions
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

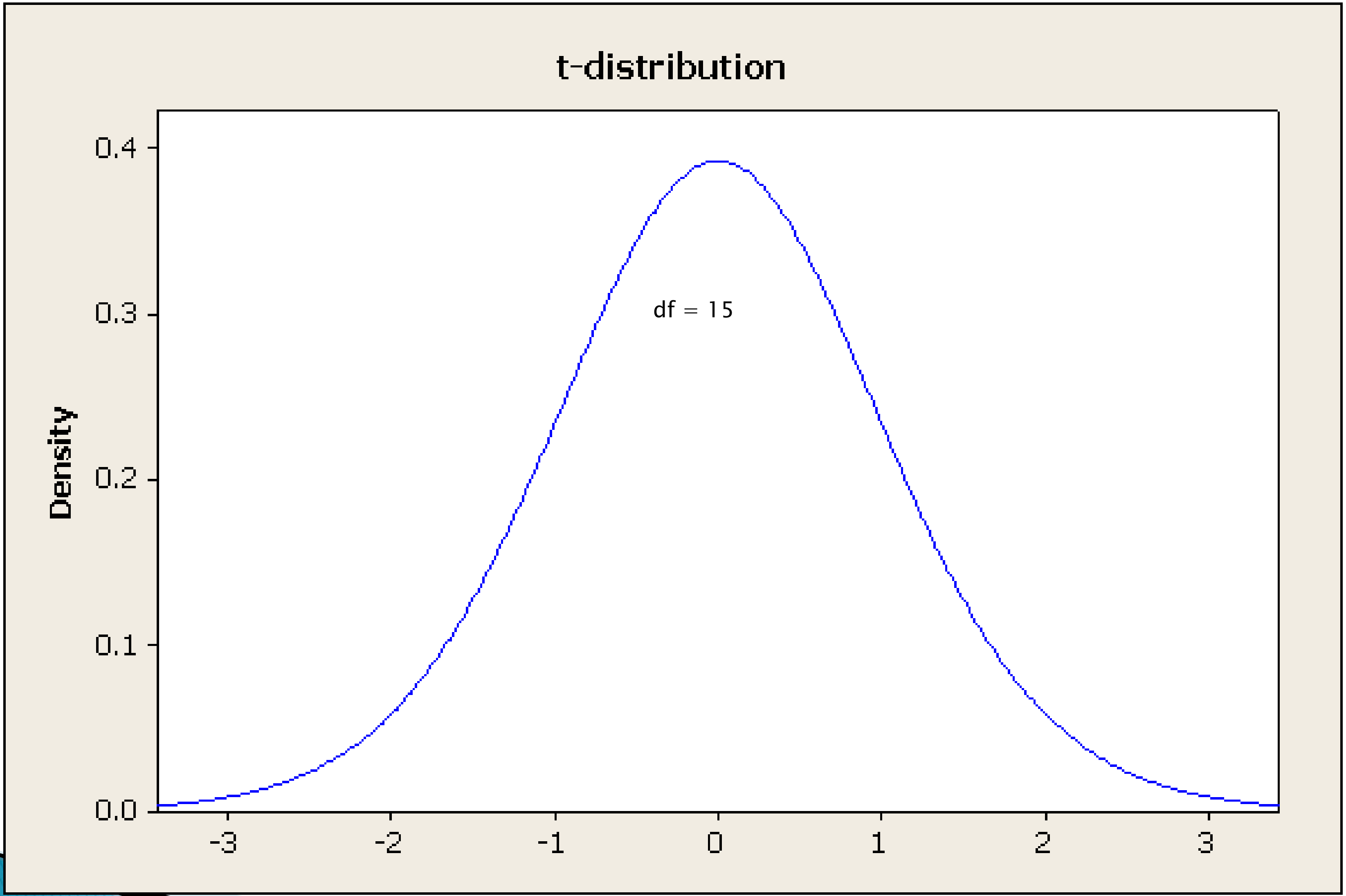


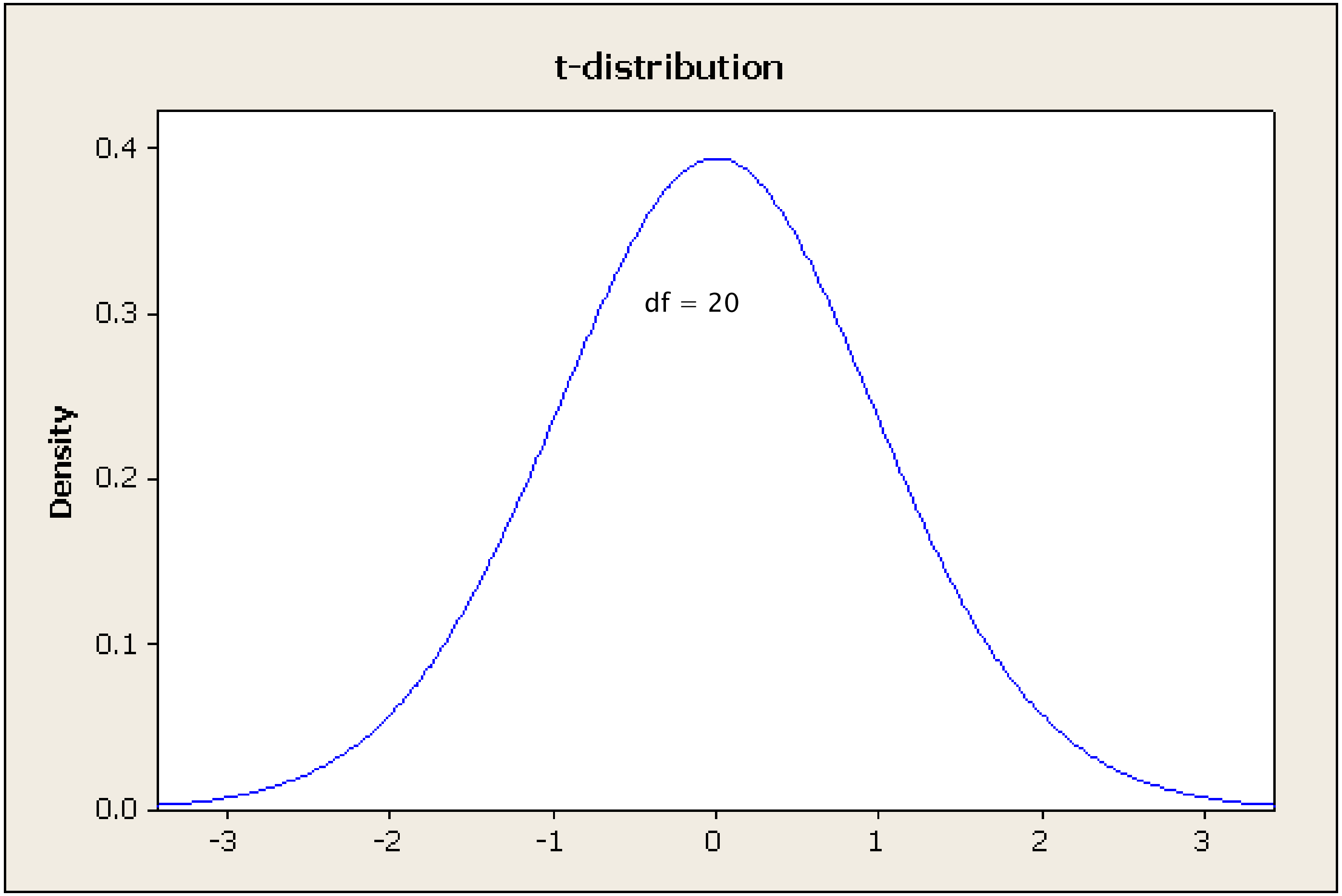
z and t distributions
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
Red = Normal
Blue = t

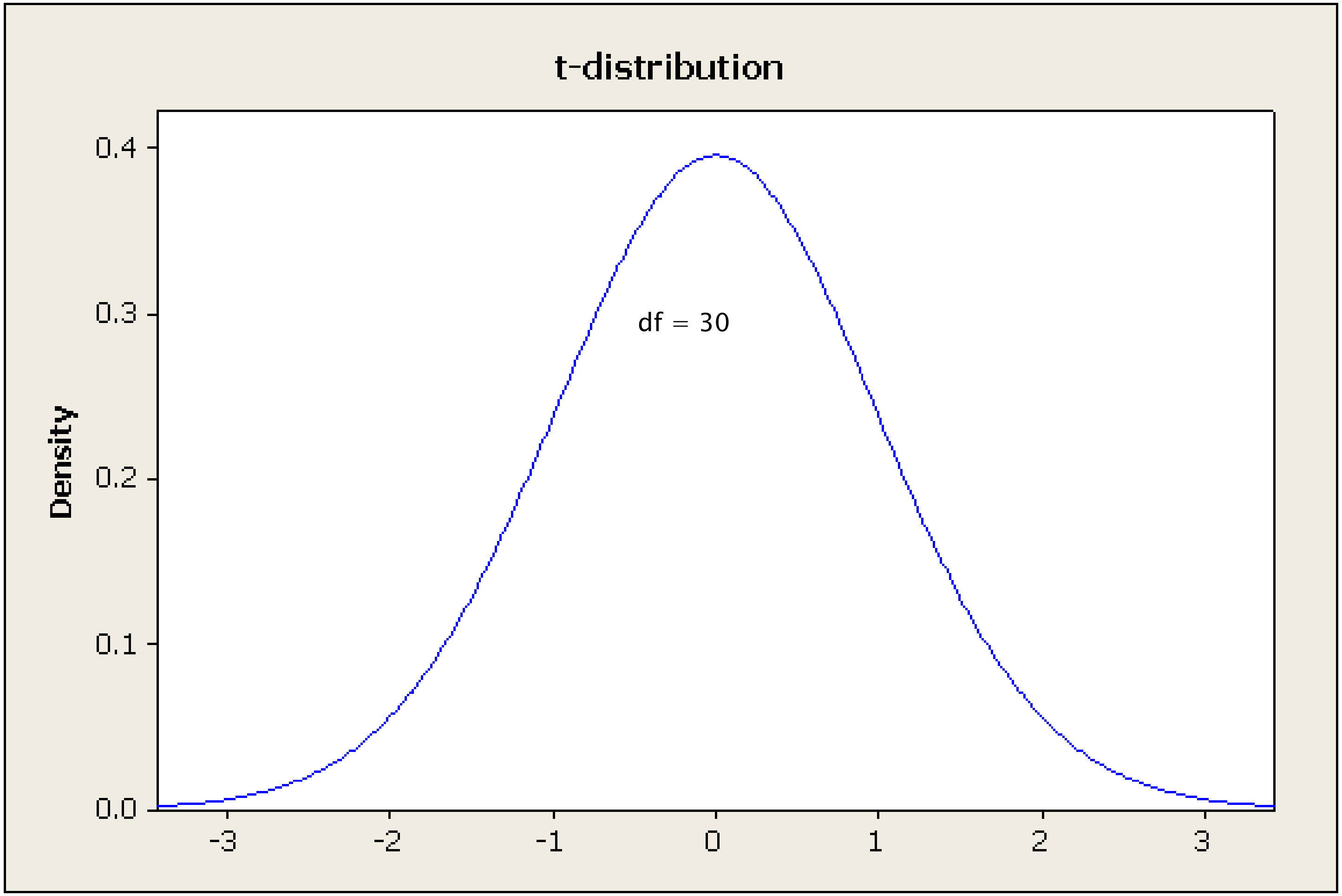


t-distribution
Density
df = 5
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

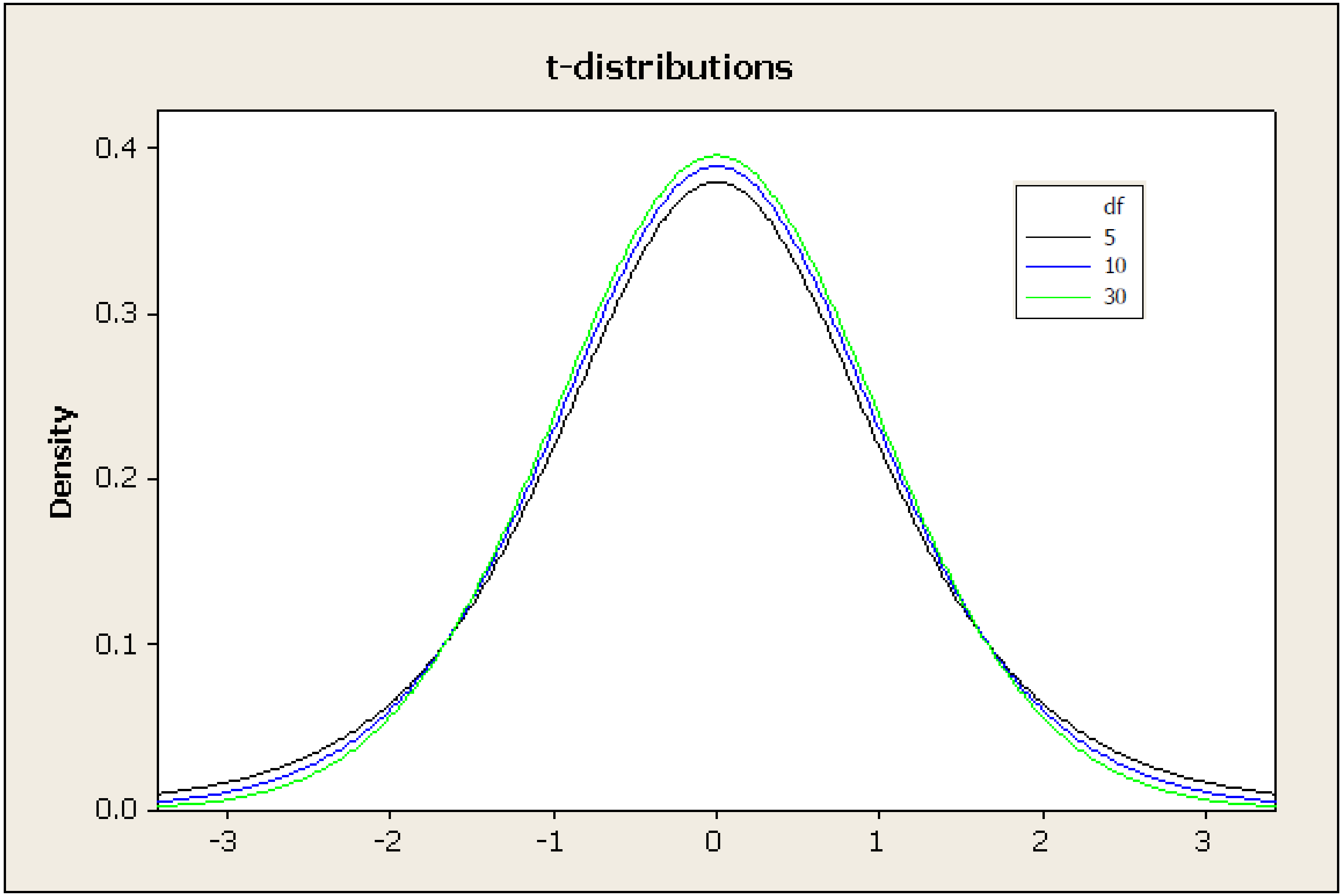

t-distribution
Density
df = 10
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3

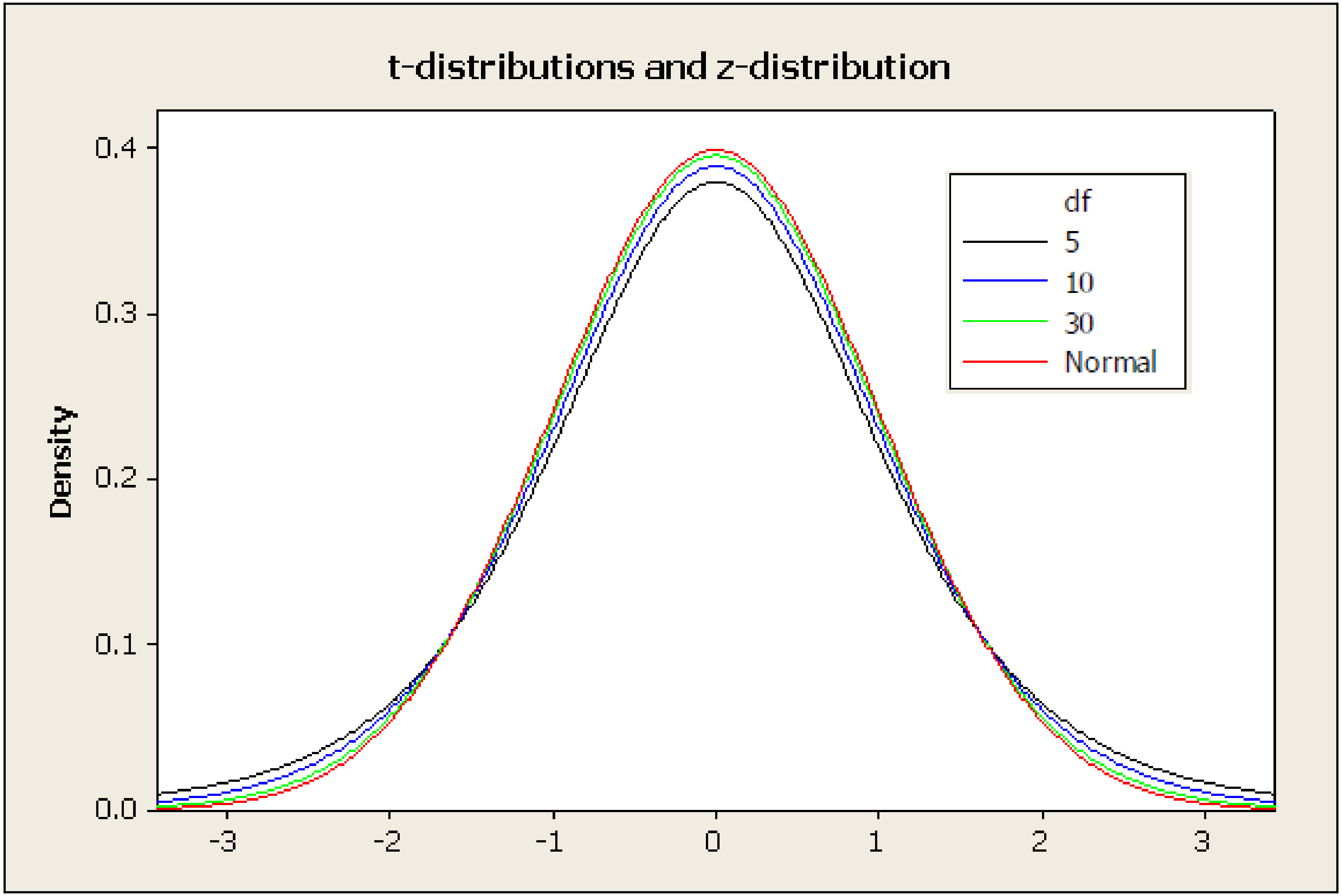

t-distribution
Density
0.4
0.3
0.2
0.1
0.0
df = 15
-3
-2
-1
0
1
2
3


t-distribution
Density
0.4
0.3
0.2
0.1
0.0
df = 20
-3
-2
-1
0
1
2
3



t-distribution
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
df = 30


t-distributions
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
df
5
10
30


t-distributions and z-distribution
Density
0.4
0.3
0.2
0.1
0.0
-3
-2
-1
0
1
2
3
df
5
10
30
Normal

Summary

- As the degrees of freedom increase, the t-distribution's shape gets closer and closer to a z-distribution.
- At around $n = 30\text{–}40$ the difference is nearly indistinguishable.

Use of t-distribution

- Used when σ is unknown typically for small samples
- Most appropriate when the population we are sampling from is *normal* but can be used when it:
  - Does not contain outliers
  - Is not extremely skewed
- Assumptions can be eased if one collects a larger sample.

T Distribution Example

- The CEO of light bulbs manufacturing company claims that an average light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days.

- If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?

- What distribution should we use? Check assumptions:

T Distribution Example cont…

- We have no information about the population except a claimed mean
  - $\mu = 300$
- We have more information about the sample:
  - $n = 15$
  - $\bar{x} = 290$
  - $s = 50$
- We do not know σ or have $n > 30$, CLT does not hold
  - Must use the T distribution

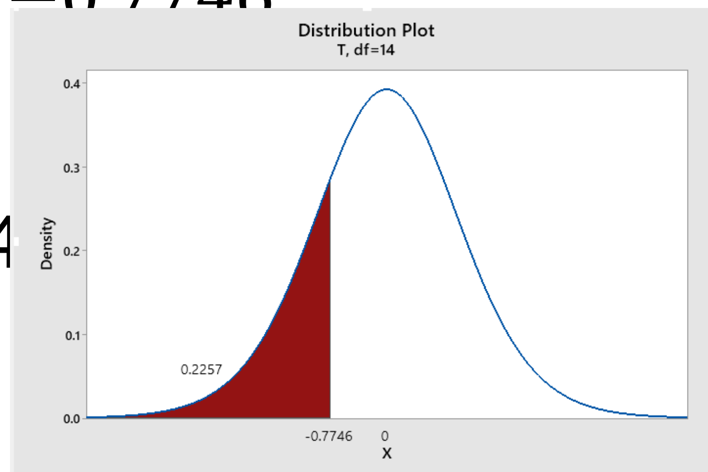
T Distribution Example cont…

- Find P(X ≤ 290) form the t distribution w/ df=15-1=14

$$\frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}} = \frac{290-300}{\frac{50}{\sqrt{15}}} = -0.7746$$

- From the T table w/ df=14
  - P(T ≤ -0.7746) = 0.2257

0.22573 =T.DIST(-0.7746,14,1)



Approximating Discrete Distributions

- Important Discrete Distributions:
  - Poisson – The number of events (x) likely to happen on a fixed interval with rate λ
  - Binomial – Probability of x successes in a fixed number of trials (n) with (p) probability of success

- We know how to solve these, but what if our numbers get really big?

Binomial example

- In a digital communication channel, assume that the number of bits received in error can be modeled by a binomial random variable.

- The probability that a bit is received in error is 0.00001 (10^{-5}). If 16 million bits are transmitted, what is the probability that 150 or fewer errors occur?

- Let X denote the number of errors. Can we solve this?

$$P\left(X \le 150\right) = \sum_{x=0}^{150} C_x^{16000000} \left(10^{-5}\right)^x \left(1-10^{-5}\right)^{16000000-x}$$

- Technically, yes, but too hard manually.

Approximating w/ the Normal

- So what if our numbers get really big?
  - We can approximate these distributions with the Normal
  - We will focus on this with the binomial, but it can also be done in a similar manner with the Poisson

- Let’s visualize this in Minitab

- If n is large, and p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.

Normal Approximation for the Binomial

- Recall the Binomial Mean and SD.

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

$$\sigma = \sqrt{np(1-p)}$$

- Then if we can say :

$$X \text{ is approx. } N\left(np, \sqrt{np(1-p)}\right)$$

- A common rule of thumb, we will use the approximation for values of n and p that satisfy both:

$np \geq 5$ and $n(1-p) \geq 5$

Graduation Example

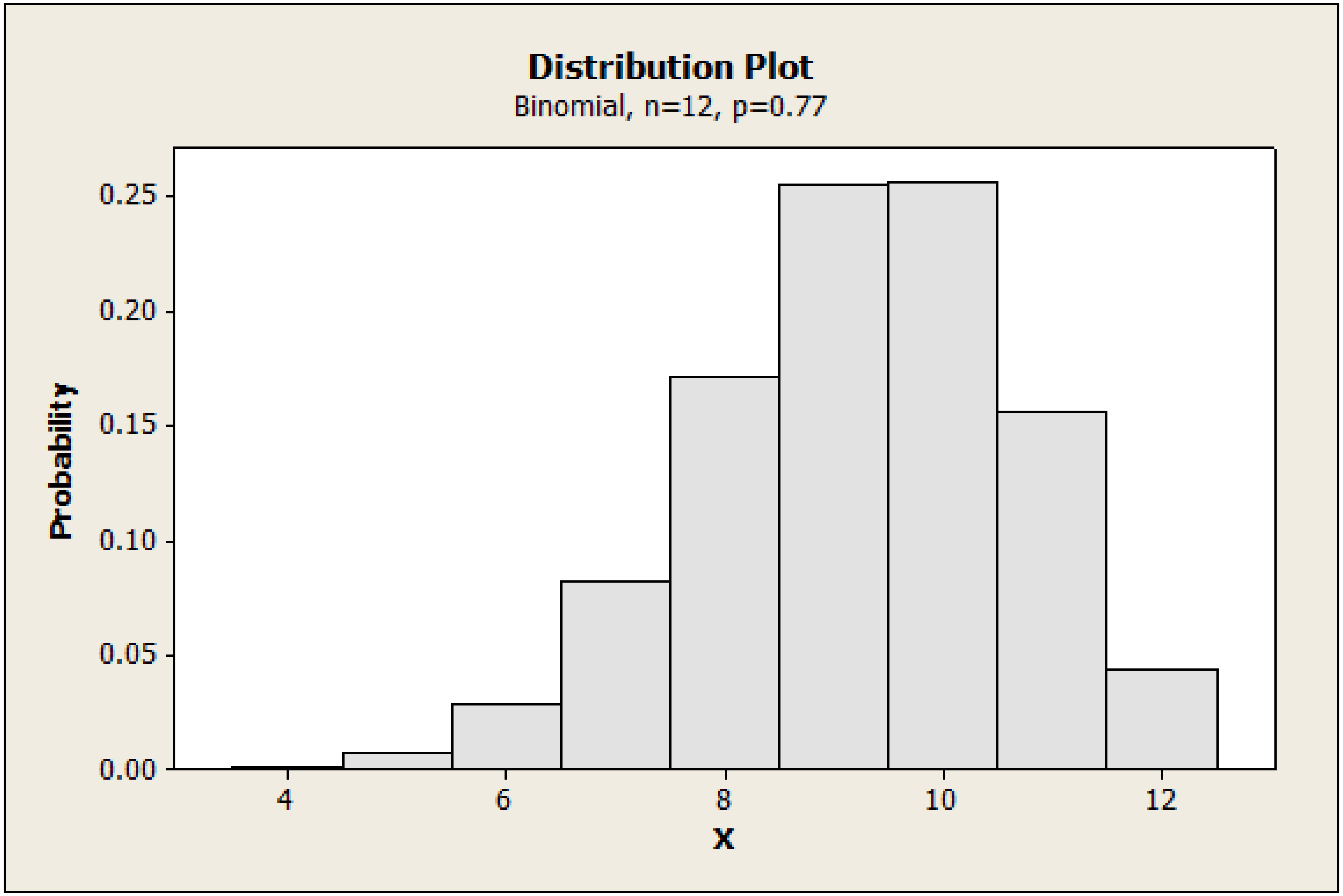
- Suppose the probability that on entering college, a student will graduate in 4 years is 0.77. An academic advisor is advising 12 freshmen.

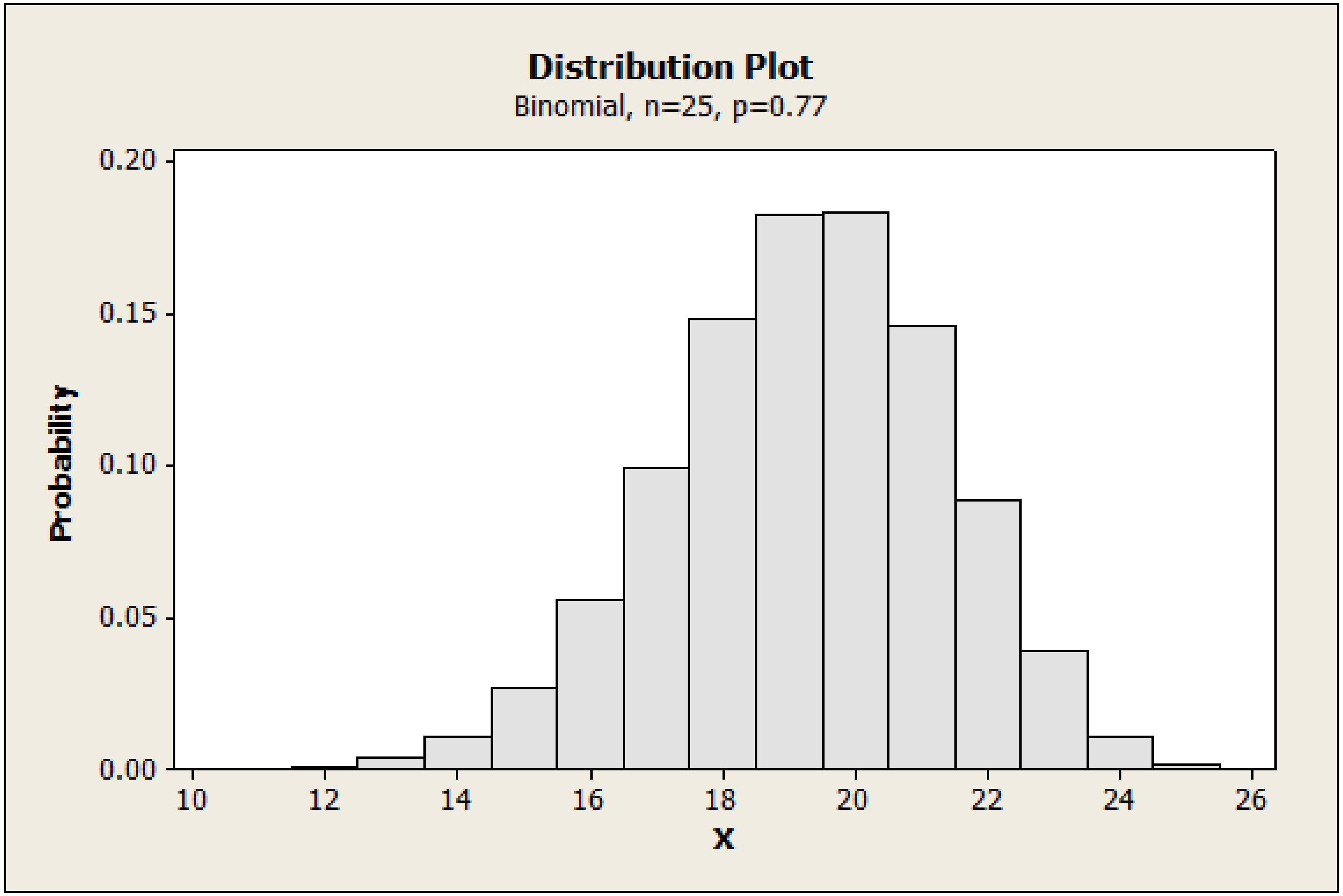
- Would the approximation work?

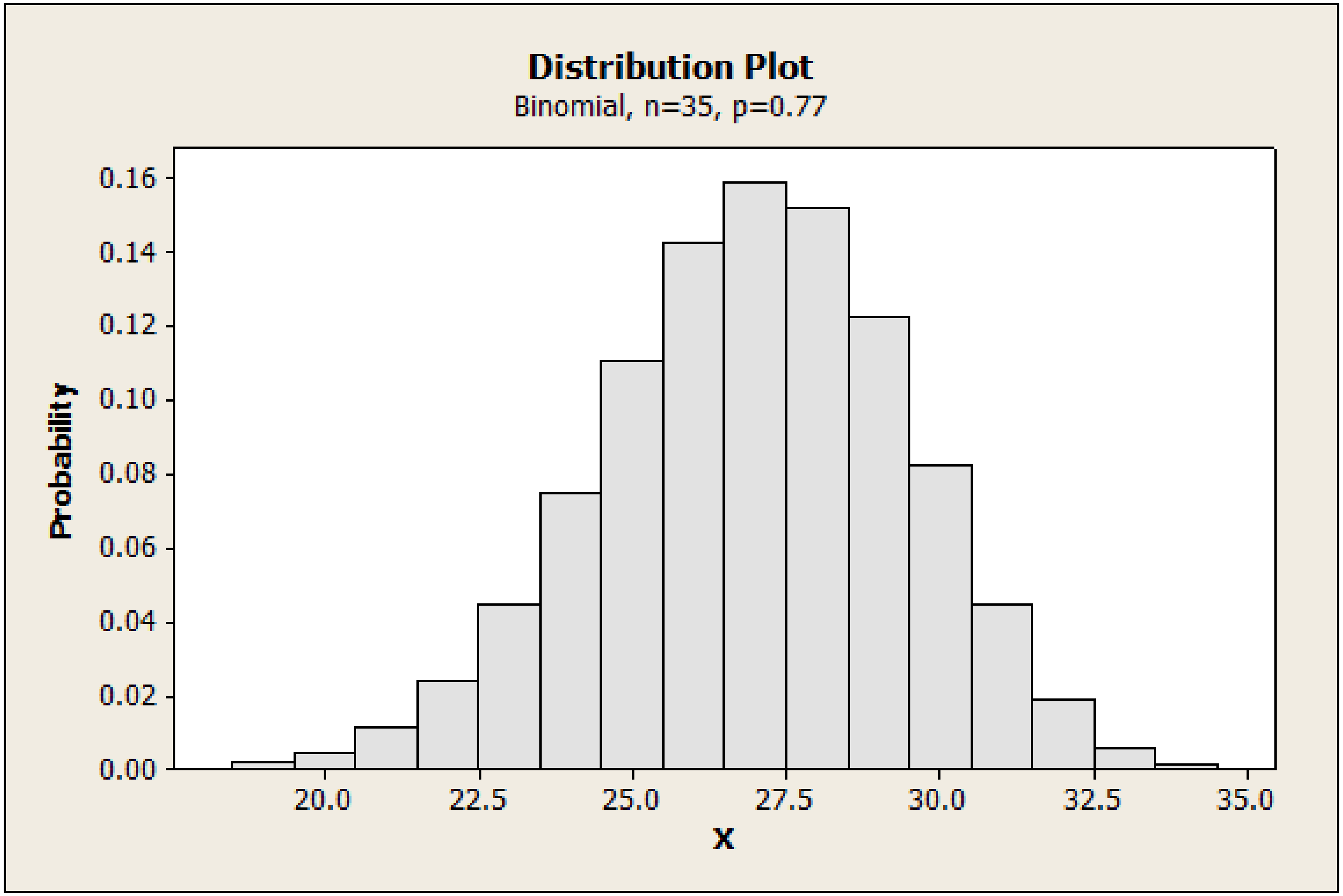
$$12(0.77) = 9.24 \text{ and } 12(1 - 0.77) = 2.76$$

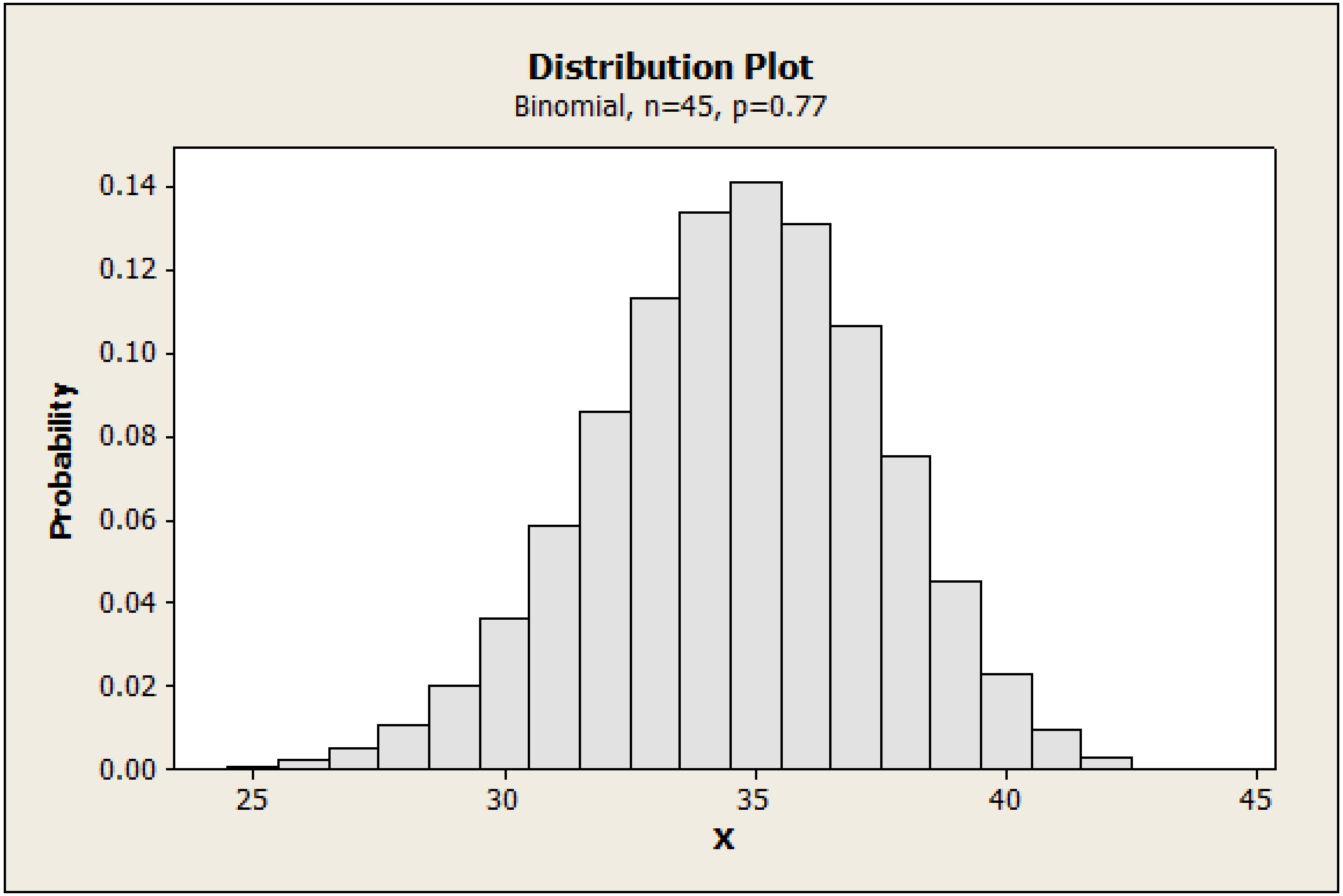
- We do not meet the criteria.

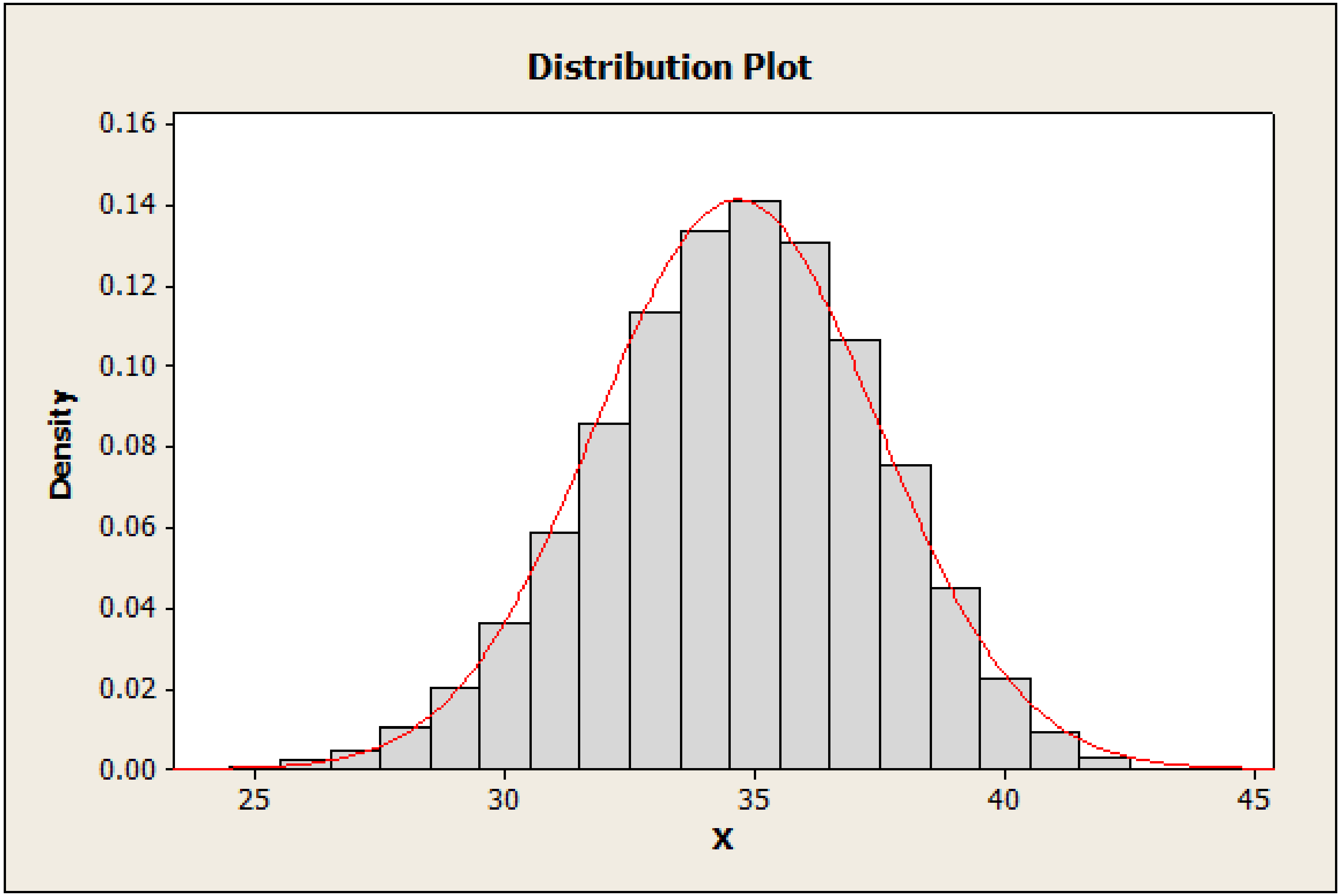
- Let's see what that distribution looks like…


Distribution Plot
Binomial, n=12, p=0.77
Probability
0.25
0.20
0.15
0.10
0.05
0.00
4
6
8
10
12
X


Distribution Plot
Binomial, n=25, p=0.77
Probability
0.20
0.15
0.10
0.05
0.00
10
12
14
16
18
20
22
24
26
X


Distribution Plot
Binomial, n=35, p=0.77
Probability
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
20.0
22.5
25.0
27.5
30.0
32.5
35.0
X


Distribution Plot
Binomial, n=45, p=0.77
Probability
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
25
30
35
40
45
X


Distribution Plot
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
25
30
35
40
45
X

Graduation Example

- Now would the approximation work?

$45(0.77) = 34.65$ and $45(1 - 0.77) = 10.35$

- Both calculations are greater than 10.
- Thus, this binomial distribution can be approximated with the normal distribution.

Graduation Example

- Start with X ~ B(45, 0.77)
- We meet our criteria
- Then calculate the mean and standard deviation of this binomial distribution

$$\mu = np = 45(0.77) = 34.65$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{45(0.77)(0.23)} = 2.823$$

- Thus our approximate distribution is:

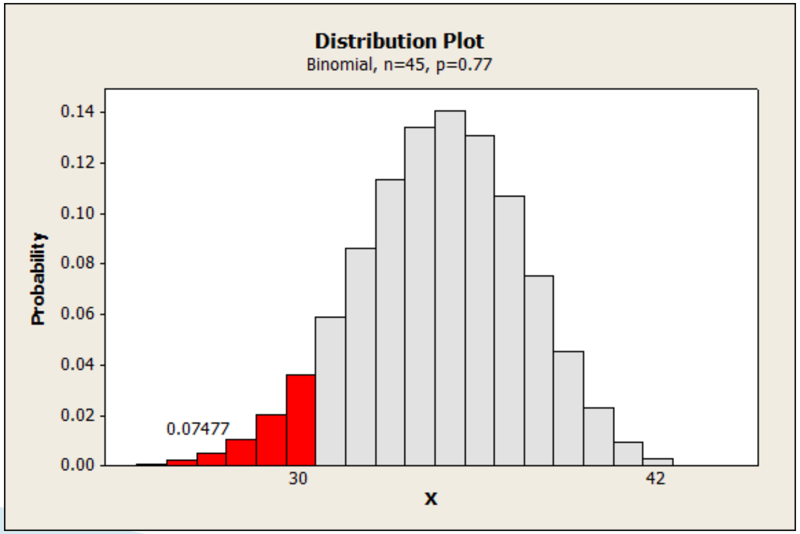
X is approx. N(34.65, 2.823)

Graduation Example

- From this SRS of 45, what is the probability that 30 or less graduate?

- Exact Binomial probability:

$$P(X \leq 30) = 0.075$$



Graduation Example

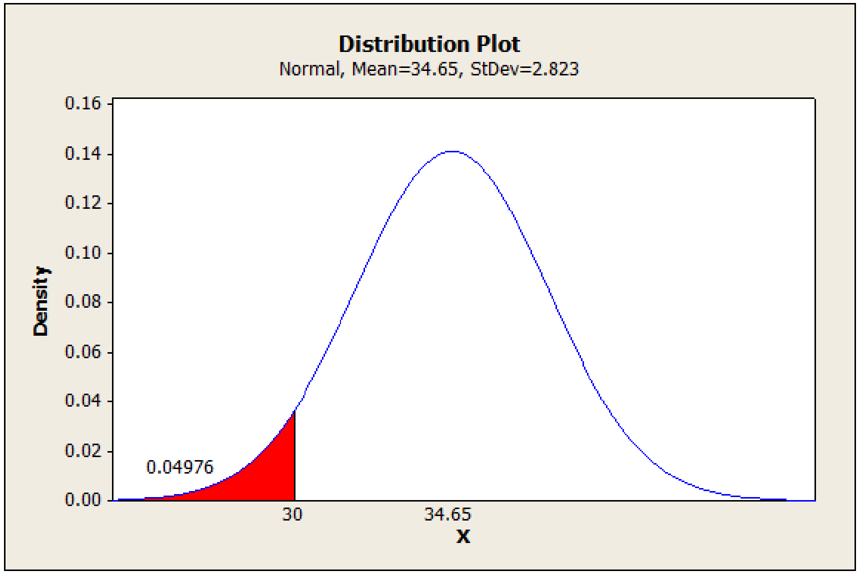
- Consider the normal approximation:

  N(34.65, 2.823)

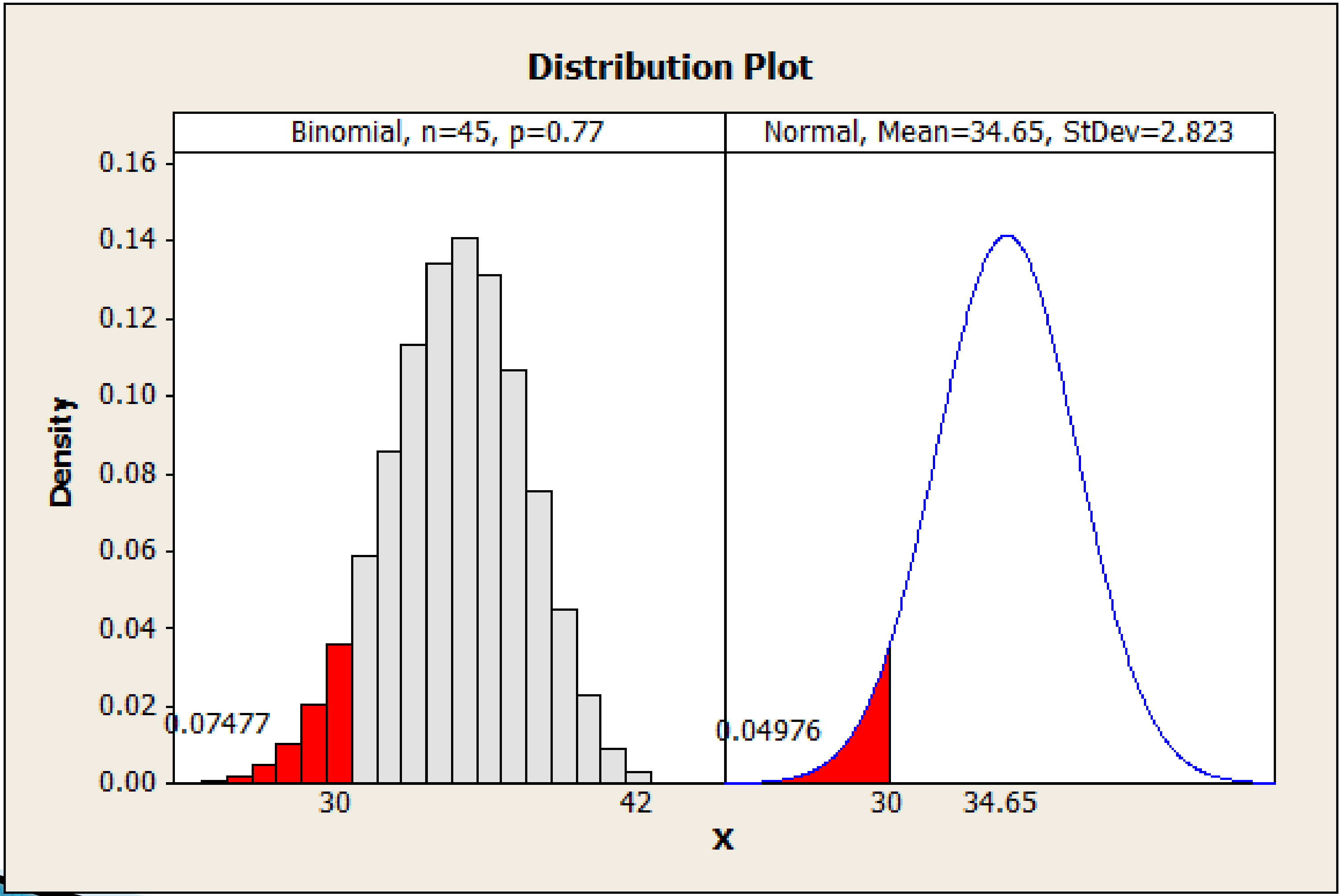
- P(X $\leq$ 30)

$$z = \frac{30 - 34.65}{2.823} = -1.65$$

- P(X $\leq$ 30) = 0.0495



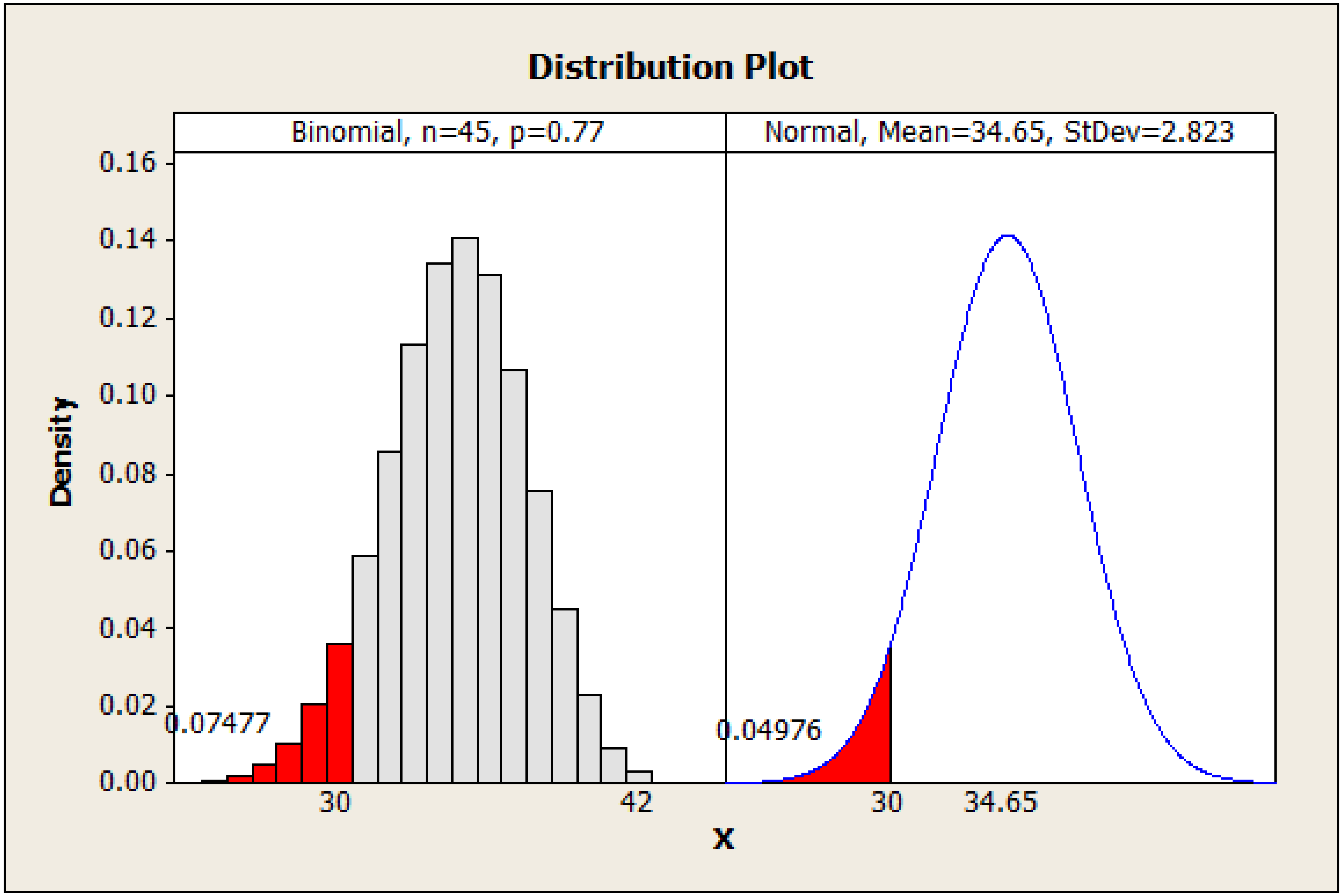
Comparison

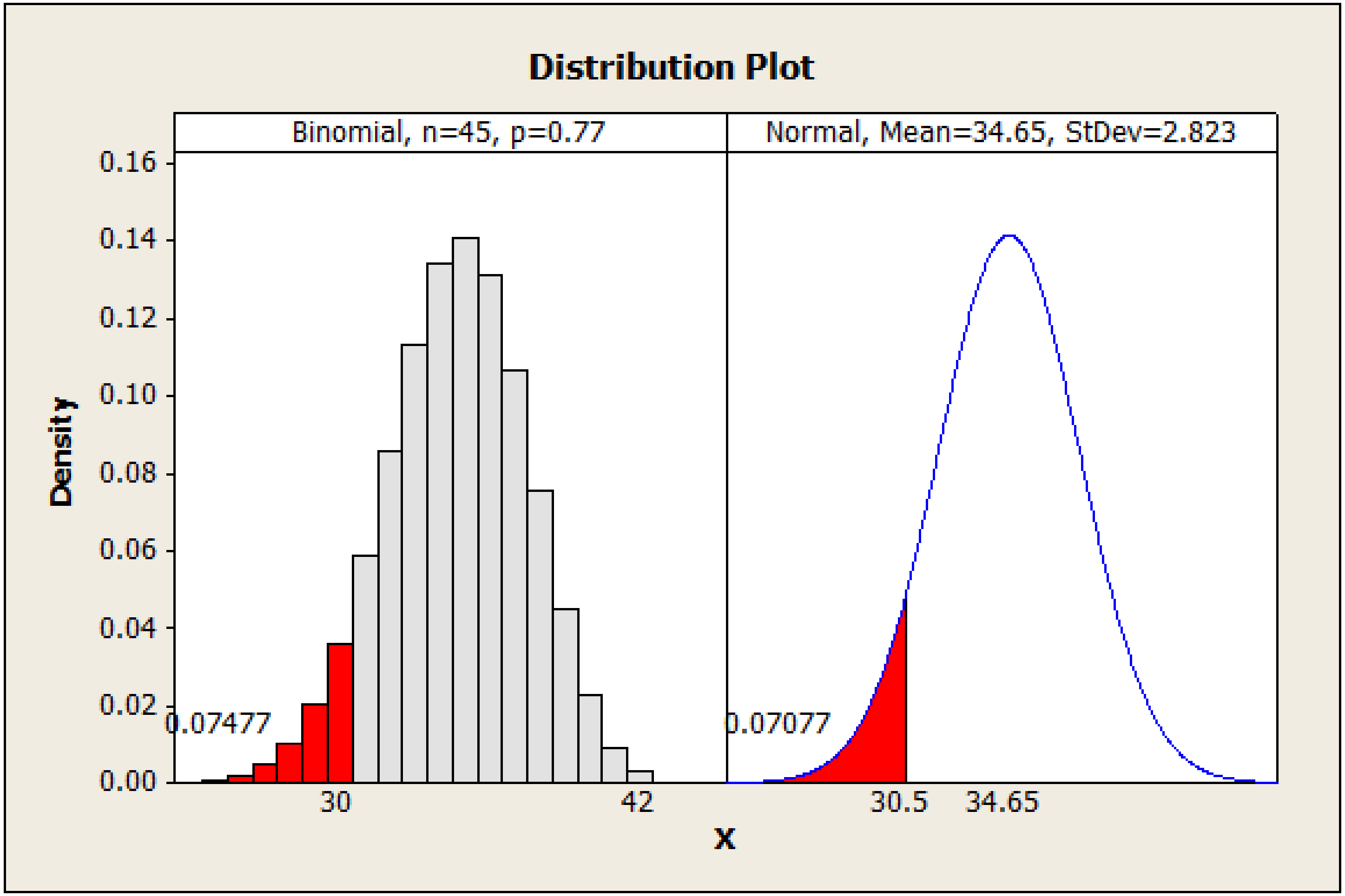


Distribution Plot
Binomial, n=45, p=0.77
Normal, Mean=34.65, StDev=2.823
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.07477
0.04976
30
42
30
34.65
X

Normal Approximation

- The normal approximation is not perfect.
- A continuity correction can be made to improve the approximation.
- Adding 0.5 to our x value utilizes what we call the continuity correction


Distribution Plot
Binomial, n=45, p=0.77
Normal, Mean=34.65, StDev=2.823
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.07477
0.04976
30
42
30
34.65
X


Distribution Plot
Binomial, n=45, p=0.77
Normal, Mean=34.65, StDev=2.823
Density
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0.07477
0.07077
30
42
30.5
34.65
X

Normal Approximation to the Poisson

- Let X be a Poisson RV w/ mean = λ = VAR

- Then we can apply similar ideas and use:

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

- Typically works when :

$$\lambda \geq 5$$

Normal Approximation to Poisson

- Assume that the number of asbestos particles in a square meter of dust on a surface follows a Poisson distribution with a mean of 1000. If a square meter of dust is analyzed, what is the probability that 950 or fewer particles are found?

$$P(X \le 950) = \sum_{x=0}^{950} \frac{e^{-1000} 1000^x}{x!} \qquad \text{... too hard manually!}$$

$$\approx P(X < 950.5) = P\left(Z < \frac{950.5 - 1000}{\sqrt{1000}}\right)$$

$$= P(Z < -1.57) = 0.058$$