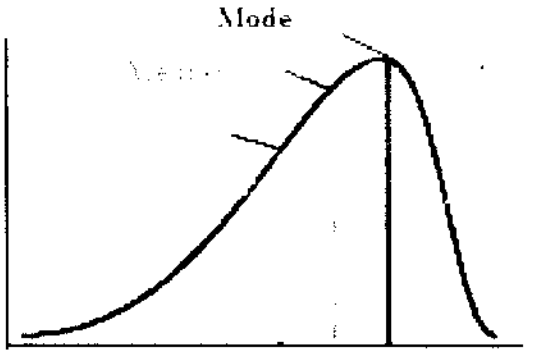
Instructor Notes Statistics in Agriculture

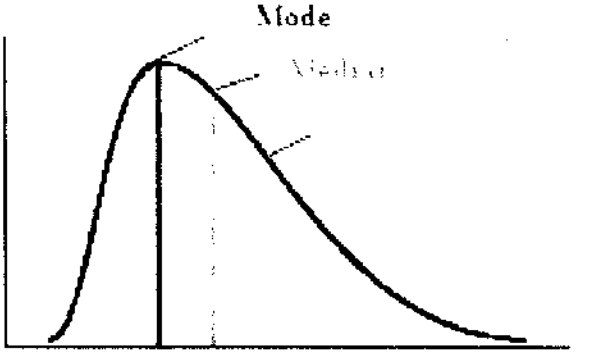
Lesson Objective: Students will be able to use statistical reasoning and techniques to gather, describe, and interpret agricultural data and to make decisions based upon statistical results. Students will describe data in terms of shape (symmetric, skewed, bell-shaped), center (mean, median, mode), and spread (standard deviation, range, interquartile range). Students will be able to make inferences and justify conclusions from experiments and observational studies.

Shape of data distribution: data can be organized in multiple ways—tables, scatterplots, box plots, and histograms. Different methods can reveal or create different impressions of the data. Scale of graphs is also an important part of interpretations.

Symmetric (bell-shaped) – when the mean and the median of a data set are the same.
Skewed – A set of data is skewed left when the mean is less than the median. A set of data is skewed right when the mean is greater than the median.

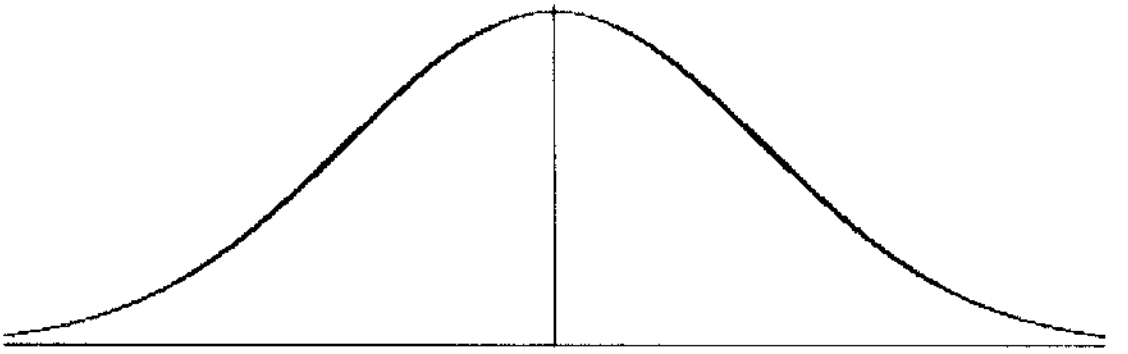


Left-Skewed (Negative Skewness)



Right-Skewed (Positive Skewness)

Symmetric, bell-shaped



Instructor Notes Statistics in Agriculture

Measures of Center: describes the center of the data distribution. Mean, median, and mode all describe the measure of center in a set of data; but each measure describes the center differently for various applications. It is important to realize which measure of center is the best choice for each application.

Mean $(\overline{X})$ – the sum of the data values divided by the number of data values (sample space, n). It is most commonly used as the "average," but can be skewed with extreme values (outliers).

$$\overline{x} = \frac{\sum n}{n}$$

Median (Med) – the middle value (if the sample space is odd) or the mean of the two middle values (if the sample space is even). Useful if there are outliers in the data set because it is less sensitive to extreme values.

Mode – the most frequently occurring value. Useful in marketing and inventory situations.

Measures of spread: describes the spread of the data distribution. Are the data points close together or spread out? Are the data points one-sided? Are there outliers?

Range – the difference between the maximum (highest data point) and the minimum (lowest data point). Range = maximum - minimum

Percentile - a number, x, from 0 to 100 that indicates the percent of the data that are less than or equal to x.

First quartile (Q_1) – 25th percentile or the point of the first quarter of the data.

Third quartile (Q_3) – 75th percentile or the point of the third quarter of the data.

Interquartile range (IQR) – the difference between the third quartile and the first quartile.

$$IQR = Q_3 - Q_1$$

Standard deviation (σ) – a measure of how much the values in the data set vary (deviate) from the mean.

$$\sigma = \sqrt{\frac{\sum\left(x - \overline{x}\right)^2}{n}}$$

Outliers – a data element with a substantially different value from the rest of the elements in the data set. Sometimes the outlier is an important part of the data; however, it can also represent a false reading. The rule for determining outliers or data lying more than 1.5 times the interquartile range away from the nearer quartile.

$Q_3 + (1.5 \times IQR)$ for outliers on the high end of the data set

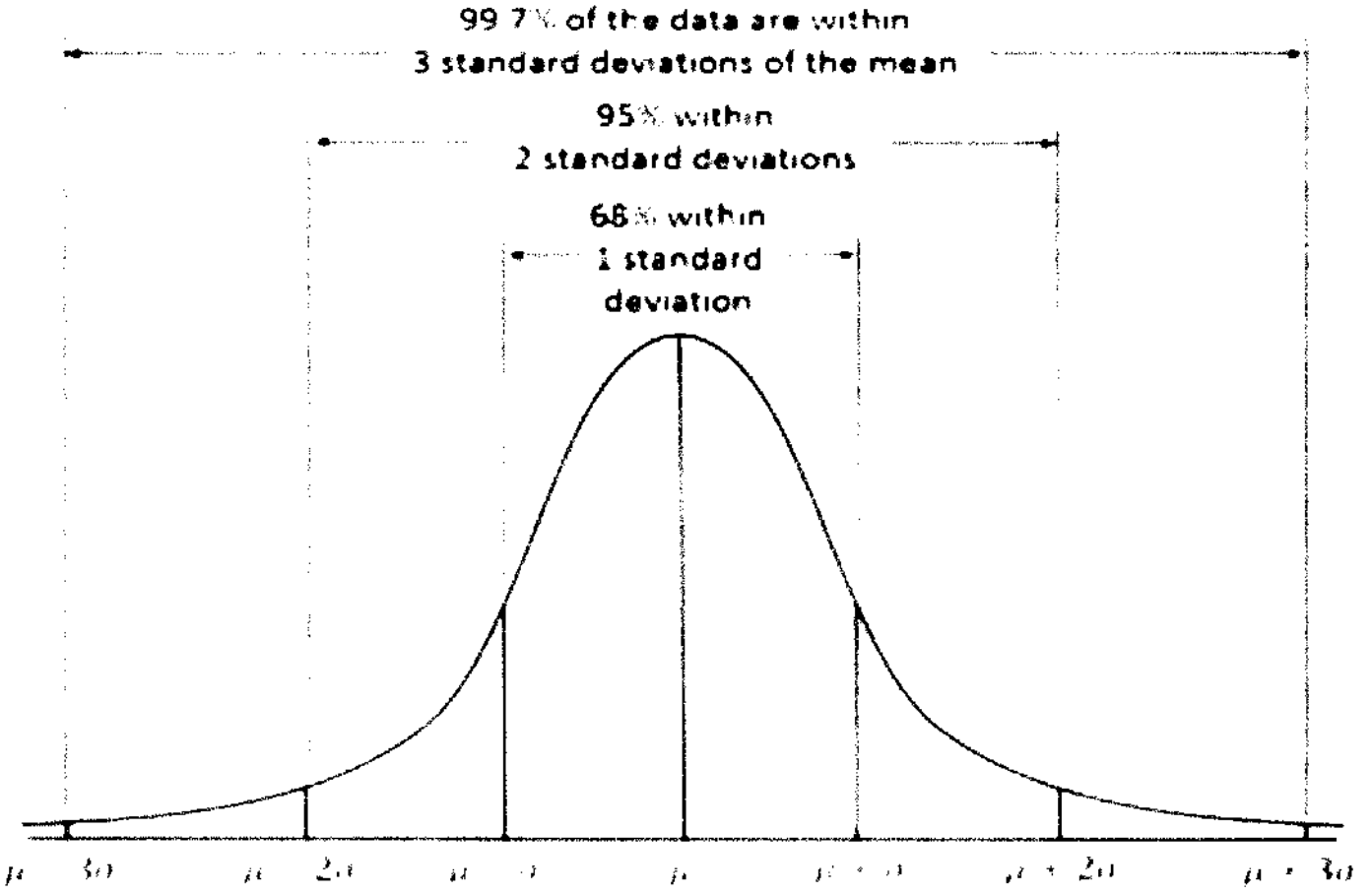
$Q_1 - (1.5 \times IQR)$ for outliers on the low end of the data set

z-score - the number of standard deviations that a value is from the mean.

$$z = \frac{x - \bar{x}}{\sigma}$$

Normal distribution – shows data that vary randomly from the mean, forms a bell-shaped curve.

Standard normal curve – normal distribution centered on the y-axis with a mean of 1 and standard deviation of 1. When a data set is normally distributed, about 68% of the data fall within one standard deviation of the mean. About 95% of the data fall within two standard deviations of the mean. About 99.7% of the data fall within three standard deviations of the mean.



Example 1: Find the following for the year 2001 using the table below:

Farm Income in Midwestern States (in millions of dollars)

State	2001	2002
Iowa	10,653	10,834
Kansas	7,979	7,862
Minnesota	7,537	7,478
Missouri	4,723	4,402
Nebraska	9,221	9,589
North Dakota	2,938	3,223
South Dakota	3,897	3,779

Source: U.S. Department of Agriculture

Mean = 6707

Median = 7537

Range = 7715 Range = max – min = 10653-2938

1st Quartile = 3897 IQR – Q3 – Q1 = 9221 - 3897

3rd Quartile = 9221

IQR = 5324

Outliers = None Q3 + (1.5 x 5324) = 17207
(if applicable) Q1 – (1.5 x 5324) = -4089

Standard deviation = 2679

Which state's 2001 income has a z-score of about 1.5? Iowa

$$1.5 = \frac{x - 6707}{2679}$$
$$x = 10725$$

Try it: Find the following information for the year 2002 using the table above.

Mean = 6738

Median = 7478

Range = 7055 Range = 10834 - 3779

1st Quartile = 3779 IQR = Q3 Q1 = 9589 - 3779

3rd Quartile = 9589

IQR = 5810

Outliers = _____None_____ (if applicable)

Q3 + (1.5 x 5810) = 18304
Q1 – (1.5 x 5810) = -4936

Standard deviation = _____2759_____

Which state's 2002 income has a z-score of about 1.0? Nebraska

$$1.0 = \frac{x - 6738}{2759}$$
$$x = 9497$$

In which year did the farm income cluster more tightly around the mean? **The year 2001. The lower the standard deviation, the closer the values are to the mean.**

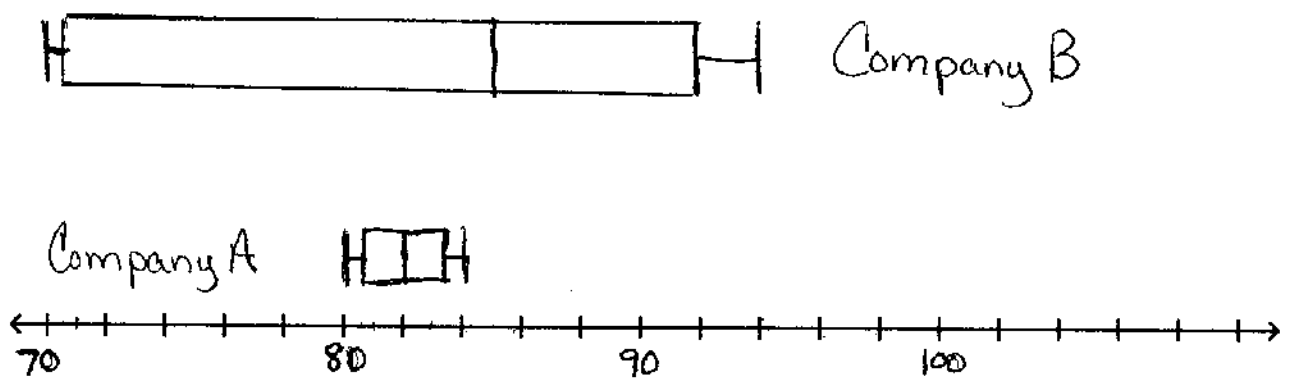
Try it: A parts store manager sells five different sizes of belts for hay balers. The computer system tracks the number of belts sold by the part number. Which measure of center would be the best choice for the store manager to use when reordering parts? Explain.

The mode would be the best measure of center for this scenario because for inventory you would want to order the parts bought the most.

Example 2: Fertilizers from 2 companies were tested on apple trees. The numbers of edible apples per tree are shown below. Each company reported an "average" of 82 edible apples per tree.
Fertilizer from company A: 81, 82, 83, 80, 84
Fertilizer from company B: 70, 90, 85, 71, 94

Company A	**Company B**
Mean = 82	Mean = 82
Median = 82	Median = 85
Min = 80	Min = 70
Max = 84	Max = 94
Q1 = 80.5	Q1 = 70.5
Q3 = 83.5	Q3 = 92
σ = 1.4	σ = 9.8
IQR = 3	IQR = 21.5
Outliers? None	Outliers? None

Graph the box plot for the data for both companies on the same graph. Use an appropriate scale.



Numbers of edible apples

State which measure of center was used by each company.

Company A could use the mean or median since both are 82. Company B used the mean.

Which company's product showed more consistent results? Explain.

Company A's product shows more consistent results because there is less spread in the data as seen by the box plots. Also, the standard deviation for Company A $\sigma = 1.4$ is much smaller than Company B's $\sigma = 9.8$ which indicates data values closer to the mean.

Find the number of edible apples for a z-score of 1.5 for both Company A and Company B.

Company A

$$1.5 = \frac{x - 82}{1.4}$$

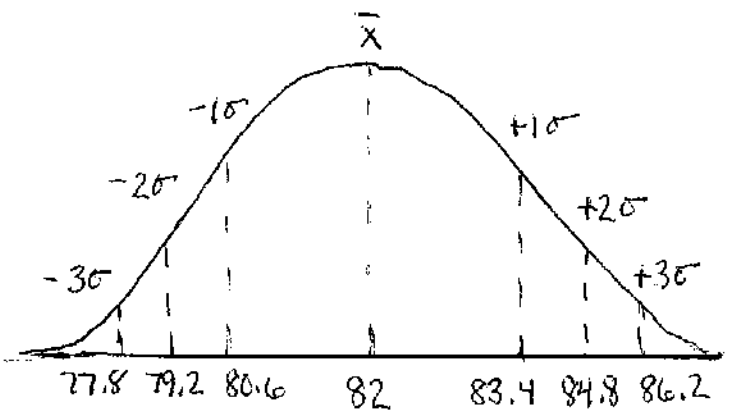
$$x = 84.1$$

Company B

$$1.5 = \frac{x - 82}{9.8}$$

$$x = 96.7$$

Draw a normal distribution for Company A



Draw a normal distribution for Company B