Describing Through-Time Data: The Run Chart

Data Science for Quality Management: Describing Data Graphically

with **Wendy Martin**

Learning objective:

Construct a run chart using RStudio

Statistical Analysis

Statistical analysis has two parts:

- Graphics: pictures that provide a visual representation of what the numbers describe or identify

Statistical Analysis

- Numerics: numbers and statistical calculations which summarize and describe our data

Statistical Analysis

We always use both pictures and numbers (‘never present a picture without stats; never present stats without a picture’!)

Arranging and Presenting Data

The first step in the analysis and interpretation of data from a random sample is the arrangement and presentation of the data.

This should be done by first graphically describing the data.

Common Methods of Graphically Describing Sample Data

- Run Charts
- Frequency Distributions
  - ✓ Ungrouped
  - ✓ Grouped
  - ✓ Relative

Common Methods of Graphically Describing Sample Data

- Histograms
- Frequency Polygons
- Box and Whisker Plots

Presenting Data As Observed Through Time: Run Charts

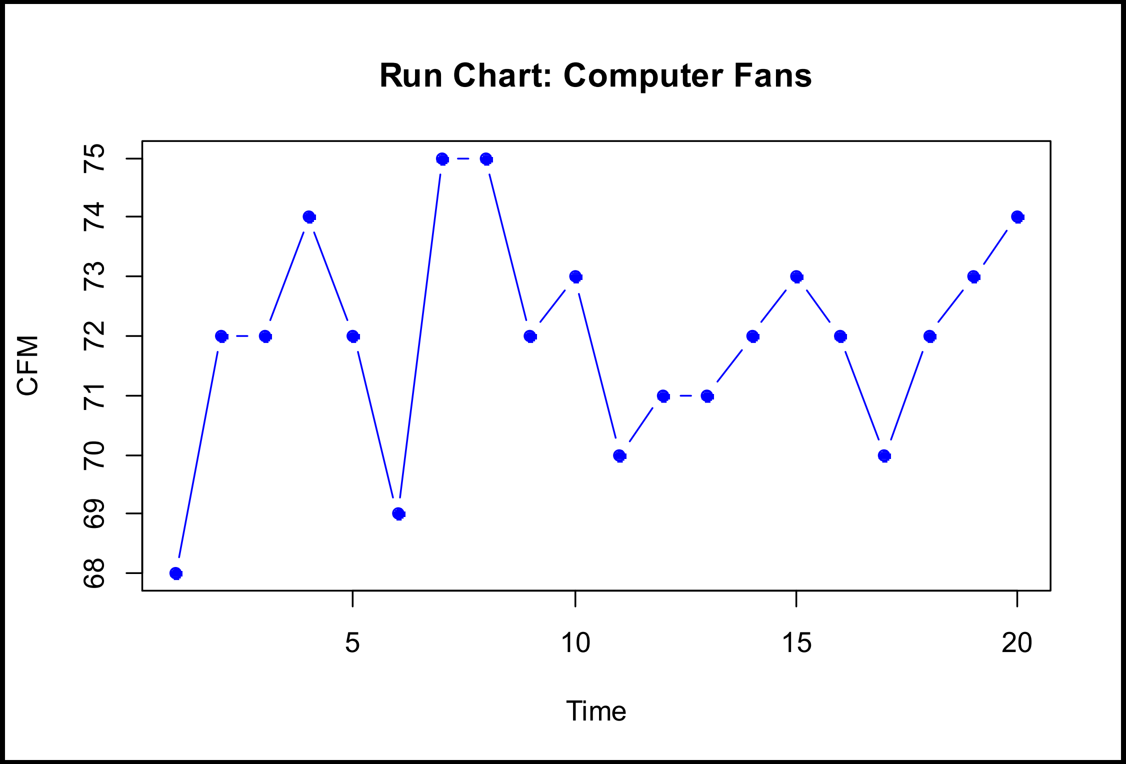
An engineer gathered 20 consecutive computer fans from a production line, keeping track of the order in which the fans were produced.

Presenting Data As Observed Through Time: Run Charts

Then these fans were tested for air flow in CFM. This testing produced the following data for the 20 fans, presented in time order.

Fans 1-10:	68	72	72	74	72	69	75	75	72	73
Fans 10-20:	70	71	71	72	73	72	70	72	73	74

Run Chart Example



Step 1: Create the Data File

Create a Vector

```
cfm <- c(68,72,72,74,72,69,75,75,72,73,70,71,71,72,73,72,70,72,73,74)
```

Store the Variable in a data frame

```
fans <- data.frame(cfm)
View(fans)
```

Step 2: Create the Run Chart

```
> require(lolcat)
> spc.run.chart(fans$cfm, main = "Run Chart: Computer Fans", ylab =
"CFM")
```

Step 3: Add a horizontal line

```
> abline(h=72)
```

Other Options for Customization

Point symbol: pch = (1-25)

Point size: cex =

Color: col = “red” (color name or hexadecimal code)

Line type: lty = (0-6)

Line width: lwd =

Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Frequency Distributions

Data Science for Quality Management: Describing Data Graphically

with **Wendy Martin**

Learning objectives:

Construct an ungrouped frequency distribution using RStudio

Construct a grouped frequency distribution using RStudio

Frequency Distributions

Frequency distributions provide us with a method for arranging and viewing data sets. This allows for easier interpretation and analysis of the data.

Ungrouped vs Grouped Frequency Distributions

Use ungrouped when there are fewer than 20 unique data values in the data set

Use grouped when there are more than 20 unique data values in the data set

Ungrouped Frequency Distribution

Using the same fan data as we employed for the run chart:

Fans 1-10:	68	72	72	74	72	69	75	75	72	73
Fans 10-20:	70	71	71	72	73	72	70	72	73	74

Ungrouped Frequency Distribution Example

	value	freq	rel.freq	cum.up	cum.down
1	68	1	0.05	0.05	1.00
2	69	1	0.05	0.10	0.95
3	70	2	0.10	0.20	0.90
4	71	2	0.10	0.30	0.80
5	72	7	0.35	0.65	0.70
6	73	3	0.15	0.80	0.35
7	74	2	0.10	0.90	0.20
8	75	2	0.10	1.00	0.10

Where:

- value = Score, Value, or Observation
- freq = Frequency
- rel.freq = Relative Frequency
- cum.up / cum.down = Cumulative

Ungrouped Frequency Distribution Example

	value	freq	rel.freq	cum.up	cum.down
1	68	1	0.05	0.05	1.00
2	69	1	0.05	0.10	0.95
3	70	2	0.10	0.20	0.90
4	71	2	0.10	0.30	0.80
5	72	7	0.35	0.65	0.70
6	73	3	0.15	0.80	0.35
7	74	2	0.10	0.90	0.20
8	75	2	0.10	1.00	0.10

Frequency distributions are considered ‘ungrouped’ when each row, or ‘class interval’, consists of only one score, value, or observation.

Ungrouped Frequency Distribution in R

```
> frequency.dist.ungrouped(fans$cfm)
```

Grouped Frequency Distribution

Ungrouped frequency distributions have one value for each class interval. Where the Range ($X_H - X_L$) of the data set is large, however, constructing a functional ungrouped frequency distribution becomes untenable.

Grouped Frequency Distribution

In these cases, we use a Grouped Frequency Distribution.

Grouped frequency distributions have a range of values associated with each interval.

- Example interval: 5 – 9
- Example interval: 1.230 - 1.234

Grouped Frequency Distribution Example

Forty (40) castings for use in a machining process have been randomly selected from an incoming lot from a supplier.

Grouped Frequency Distribution Example

Descriptive Statistics

Variable	Sample Size (n)	Mean	Std. Dev.	Low	High	Range
Weight	40	134.75	14.75	109	170	61

The data are initially arranged in an ungrouped frequency distribution:

Ungrouped Frequency Distribution

Too Many Intervals

	value	freq	rel.freq	cum.up	cum.down
1	109	1	0.025	0.025	1.000
2	111	1	0.025	0.050	0.975
3	117	1	0.025	0.075	0.950
4	118	1	0.025	0.100	0.925
5	120	1	0.025	0.125	0.900
6	121	1	0.025	0.150	0.875
7	122	2	0.050	0.200	0.850
8	124	2	0.050	0.250	0.800
9	125	1	0.025	0.275	0.750
10	126	2	0.050	0.325	0.725
11	128	2	0.050	0.375	0.675
12	129	3	0.075	0.450	0.625
13	130	1	0.025	0.475	0.550
14	131	2	0.050	0.525	0.525
15	132	1	0.025	0.550	0.475
16	133	1	0.025	0.575	0.450
17	134	1	0.025	0.600	0.425
18	135	2	0.050	0.650	0.400
19	137	1	0.025	0.675	0.350
20	139	1	0.025	0.700	0.325
21	143	2	0.050	0.750	0.300
22	146	1	0.025	0.775	0.250
23	148	2	0.050	0.825	0.225
24	152	1	0.025	0.850	0.175
25	155	2	0.050	0.900	0.150
26	158	1	0.025	0.925	0.100
27	162	1	0.025	0.950	0.075
28	165	1	0.025	0.975	0.050
29	170	1	0.025	1.000	0.025

Grouped Frequency Distribution

The data are then reorganized in a Grouped Frequency distribution

	l	min	midpoint	max	u	freq	rel.freq	cum.up	cum.down
1	[	105	107.5	110	)	1	0.025	0.025	1.000
2	[	110	112.5	115	)	1	0.025	0.050	0.975
3	[	115	117.5	120	)	2	0.050	0.100	0.950
4	[	120	122.5	125	)	6	0.150	0.250	0.900
5	[	125	127.5	130	)	8	0.200	0.450	0.750
6	[	130	132.5	135	)	6	0.150	0.600	0.550
7	[	135	137.5	140	)	4	0.100	0.700	0.400
8	[	140	142.5	145	)	2	0.050	0.750	0.300
9	[	145	147.5	150	)	3	0.075	0.825	0.250
10	[	150	152.5	155	)	1	0.025	0.850	0.175
11	[	155	157.5	160	)	3	0.075	0.925	0.150
12	[	160	162.5	165	)	1	0.025	0.950	0.075
13	[	165	167.5	170	)	1	0.025	0.975	0.050
14	[	170	172.5	175	)	1	0.025	1.000	0.025

Grouped Frequency Distribution in R

```
> frequency.dist.grouped(castings$weight)
```

Grouped Frequency Distribution

Important questions to answer:

- How many class intervals, optimally, should the frequency distribution have? How many is too few? Too many?

Grouped Frequency Distribution

- What class interval size is best for the data set we are attempting to portray in a frequency distribution?
- At what class interval should we start the grouped frequency distribution?

Constructing a Grouped Frequency Distribution

- Generate a frequency distribution with as close as you can get to 10 class intervals, without going under (divide the Range by 10 for an estimate of the class interval size you'll need);

Constructing a Grouped Frequency Distribution

- Use one of the following class interval sizes: 1, 2, 3, or 5; increasing the sizes in multiples of 10 where required (e.g. 10, 20, 30, 50, 100…)

Constructing a Grouped Frequency Distribution

- Start the first class interval with a number that is a multiple of the class interval size
- The first class interval must contain the lowest score in the data set (X_L)

Constructing a Grouped Frequency Distribution

- lolcat::freq.dist.grouped considers all of these rules to give an optimal result

Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Frequency Polygons and Histograms

Data Science for Quality Management: Describing Data Graphically

with **Wendy Martin**

Learning objectives:

Create a Frequency Polygon using RStudio
Create a histogram using RStudio

Frequency Polygons and Histograms

Useful for:

- Evaluating a manufacturing or business process
- Determining machine and process capabilities

Frequency Polygons and Histograms

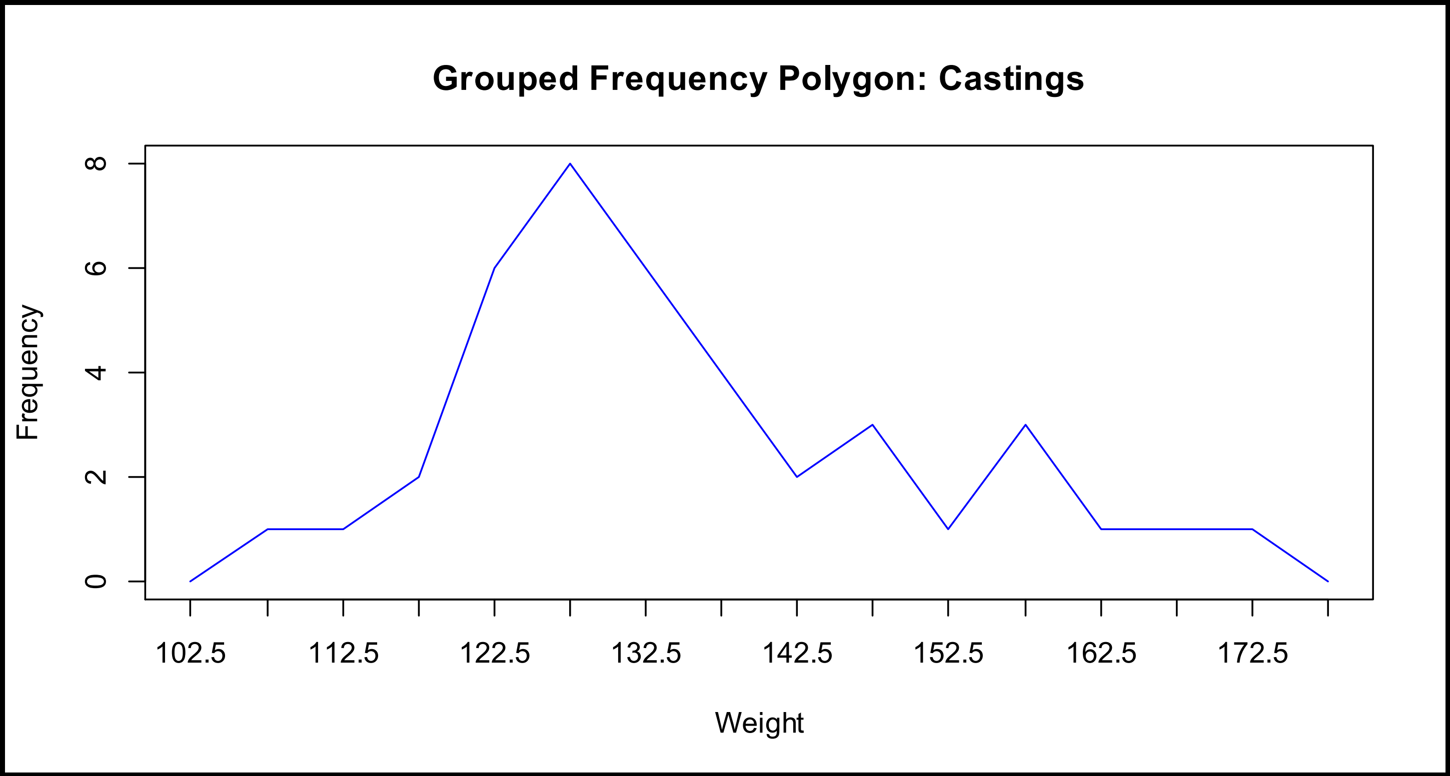
- Comparing material, vendor, operator, process and product characteristics

Ungrouped vs Grouped Frequency Histograms/Polygons

Use ungrouped when there are fewer than 20 unique data values in the data set

Use grouped when there are more than 20 unique data values in the data set

Frequency Polygons



Frequency Polygons

A graph or chart which represents the frequency of observations at each class interval (grouped) or value/score (ungrouped).

Similar to the frequency column of the frequency distribution.

Frequency Polygon: Advantages

Frequency polygons often present a more representative illustration of the data pattern when data are measured along a continuous scale.

Frequency Polygon: Advantages

The polygon becomes increasingly smooth and curve-like as the number of class intervals and sample size (n) increases, more closely representing the sampled population.

Ungrouped Frequency Polygon

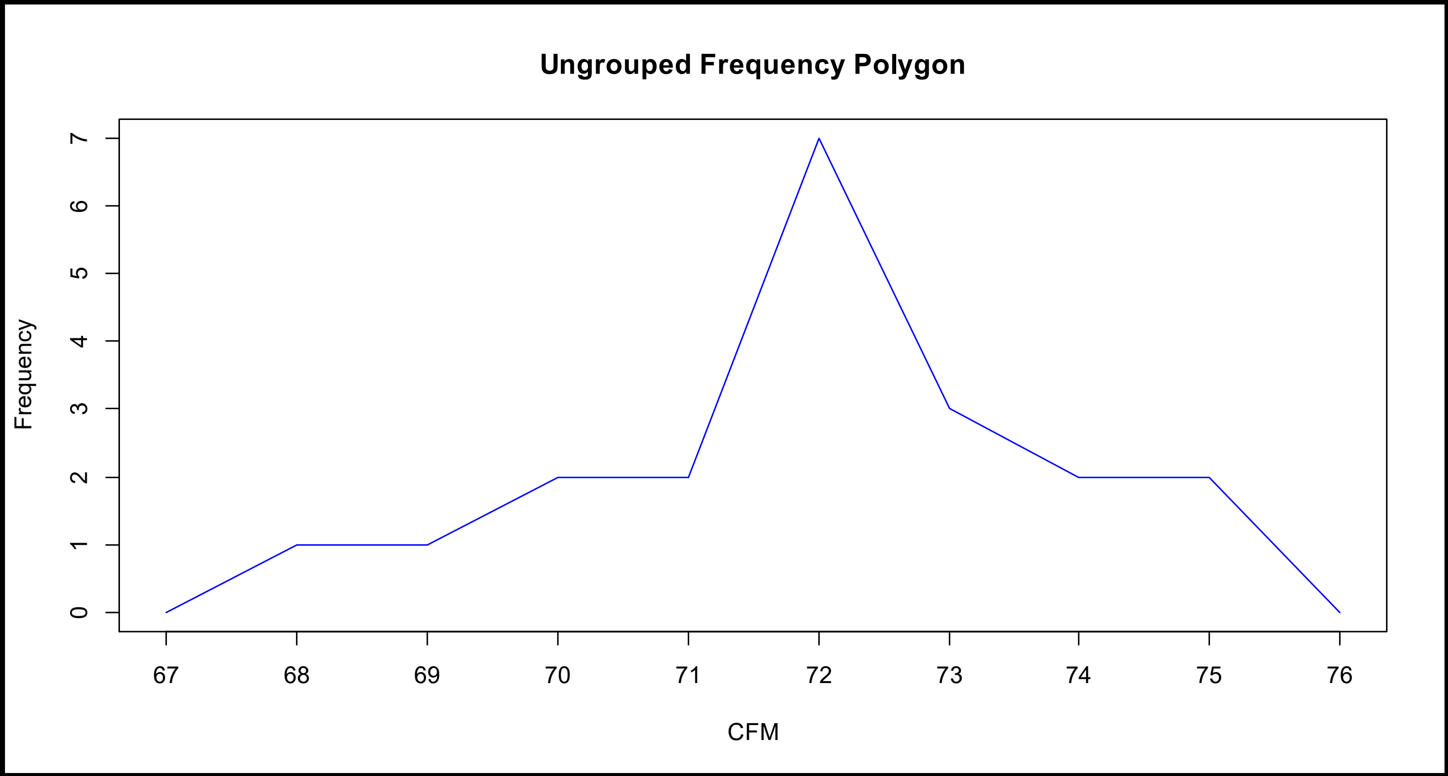
Using the same fan data as we employed for the ungrouped frequency distribution:

Fans 1-10:	68	72	72	74	72	69	75	75	72	73
Fans 10-20:	70	71	71	72	73	72	70	72	73	74

Ungrouped Frequency Polygon in R

```
> frequency.polygon.ungrouped(fans$cfm)
```

Ungrouped Frequency Polygon



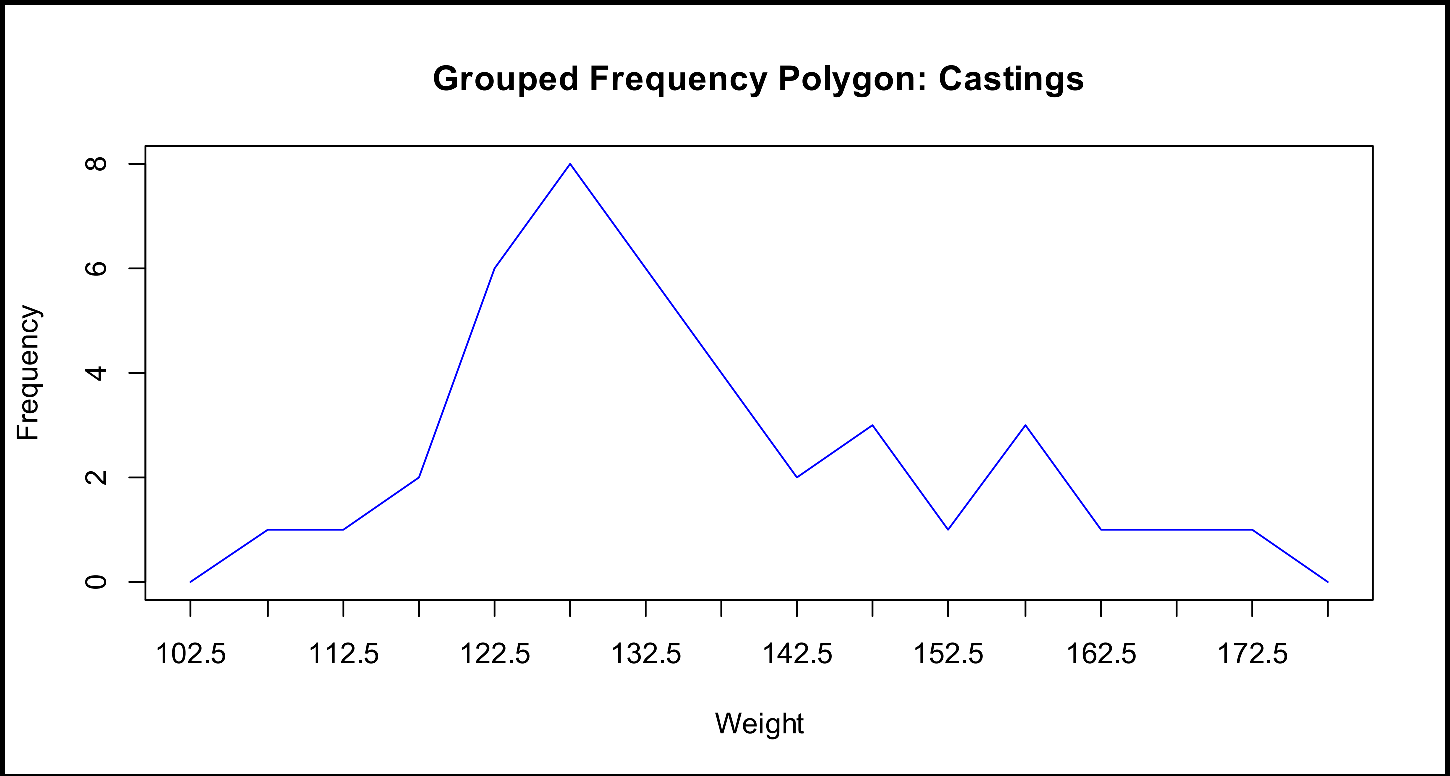
Grouped Frequency Polygon

Using the same castings data as we employed for the grouped frequency distribution:

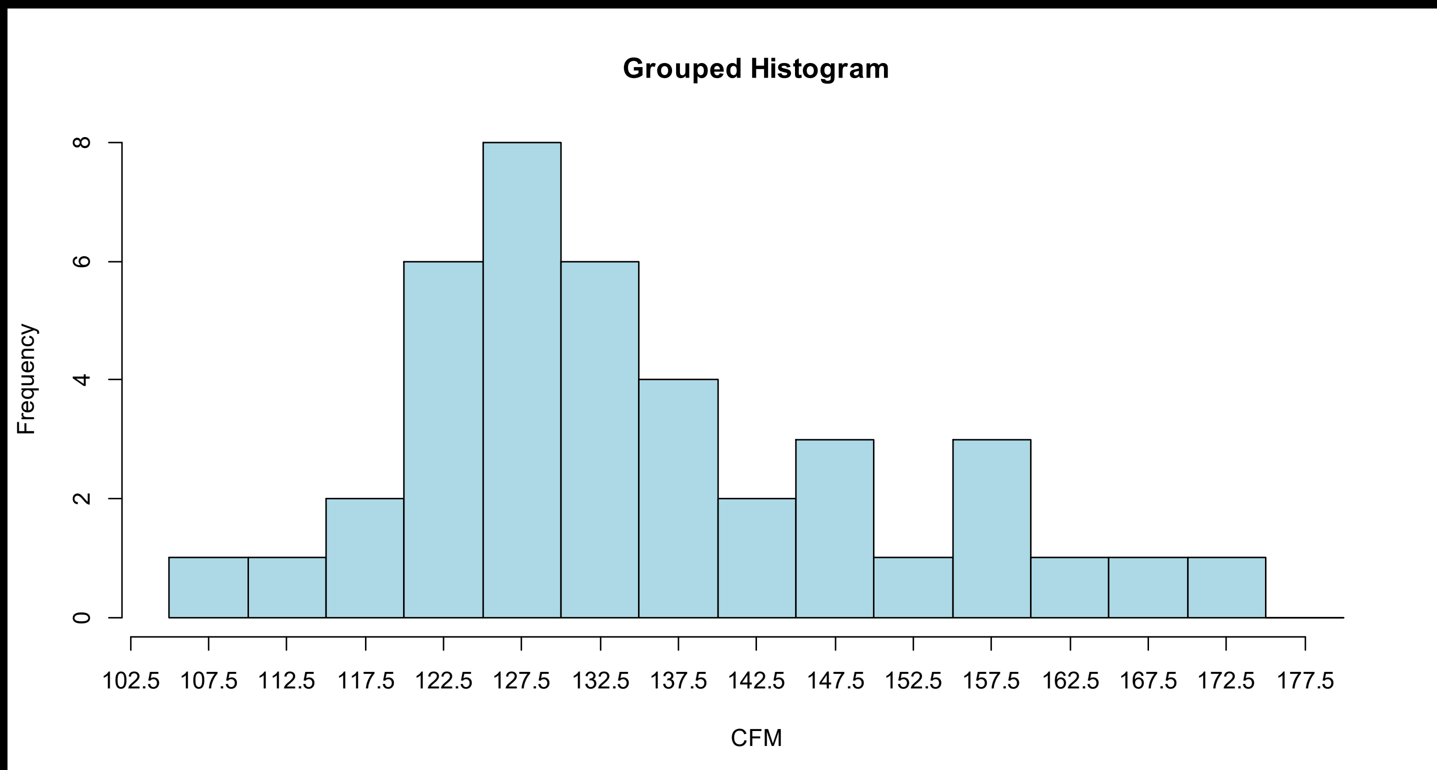
Grouped Frequency Polygon in R

```
> frequency.polygon.grouped(castings$weight)
```

Grouped Frequency Polygon



Histograms



Grouped Histogram
Frequency
0
2
4
6
8
102.5
107.5
112.5
117.5
122.5
127.5
132.5
137.5
142.5
147.5
152.5
157.5
162.5
167.5
172.5
177.5
CFM

Histograms

Similar to the frequency polygon, except that bars are used to represent the frequency of occurrence at each score or class interval.

Typically, each vertical bar in the histogram is centered above each class interval (or individual score).

Histogram: Advantages

Each bar or rectangular area clearly shows the relative magnitude of that class interval.

The area in each bar reflects the true proportion of the total number of observations occurring in the class interval.

A Note About Histograms

When the data represent discrete values, such as counts, histograms must be used.

When the data represent continuous values, a frequency polygon or histogram may be used.

Ungrouped Histogram

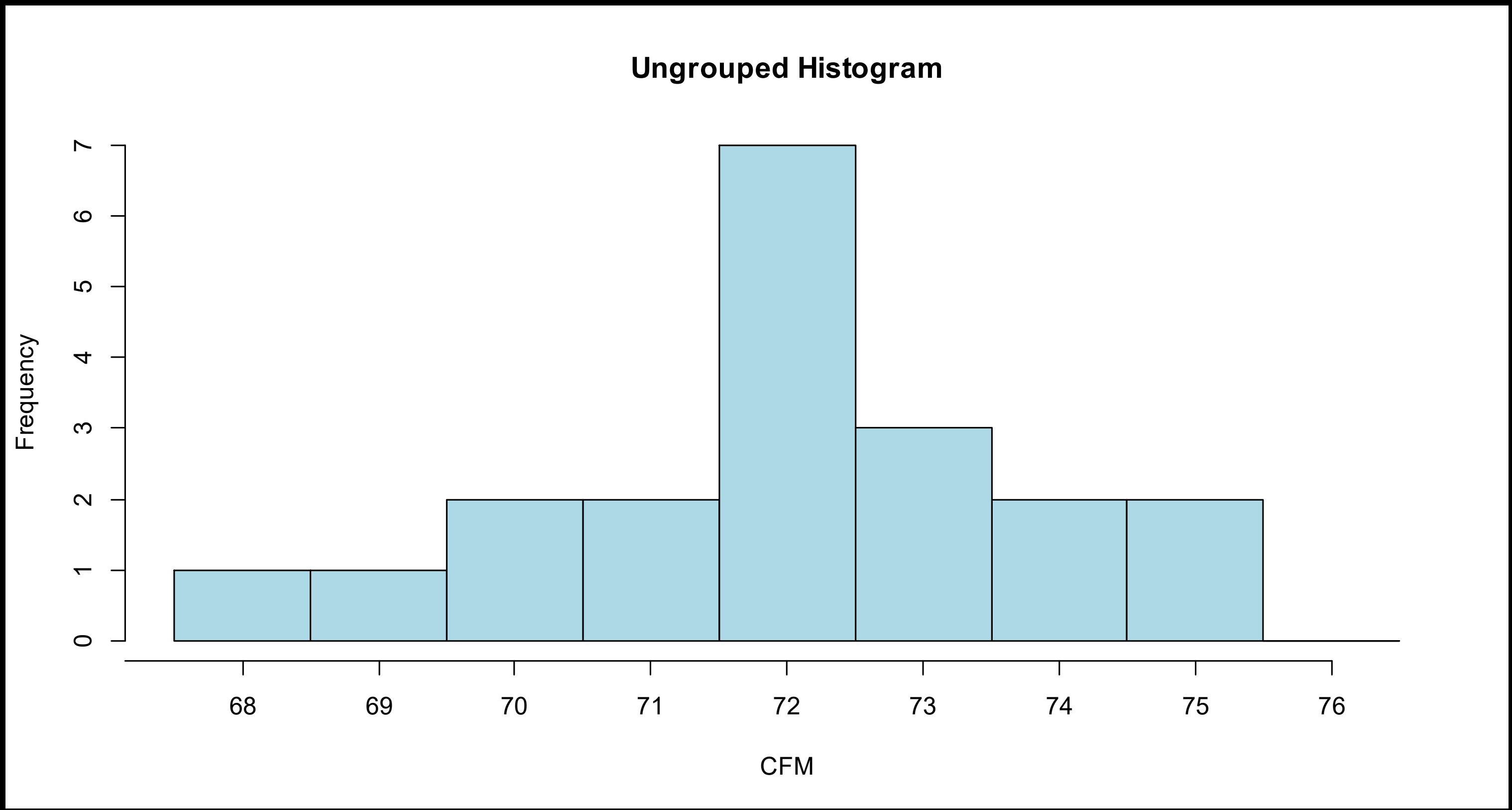
Using the same fan data as we employed for the ungrouped frequency polygon:

Fans 1-10:	68	72	72	74	72	69	75	75	72	73
Fans 10-20:	70	71	71	72	73	72	70	72	73	74

Ungrouped Histogram in R

```
> hist.ungrouped(fans$cfm)
```

Ungrouped Histogram



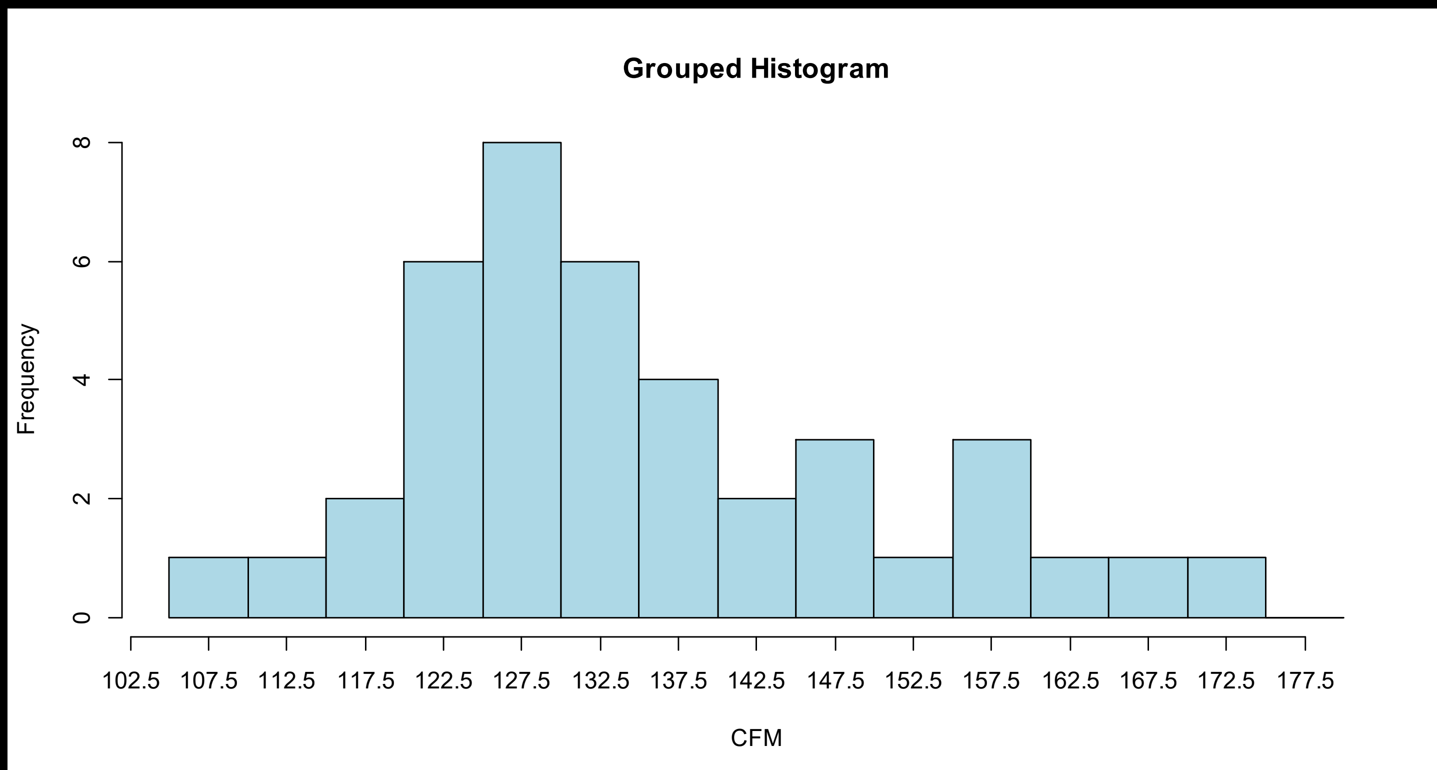
Grouped Histogram

Using the same castings data as we employed for the grouped frequency polygon:

Grouped Histogram in R

```
> hist.grouped(castings$weight)
```

Grouped Histogram



Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Histogram Patterns
Density Plots

Data Science for Quality Management: Describing Data Graphically

with **Wendy Martin**

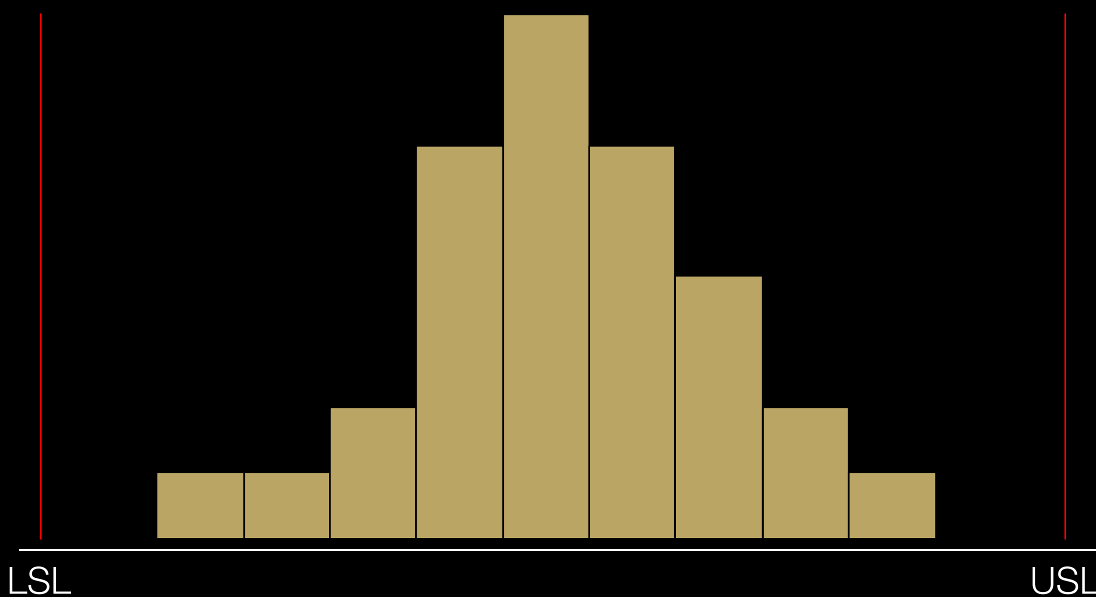
Learning objectives:

Interpret Histogram Patterns
Create a Density Plot using RStudio

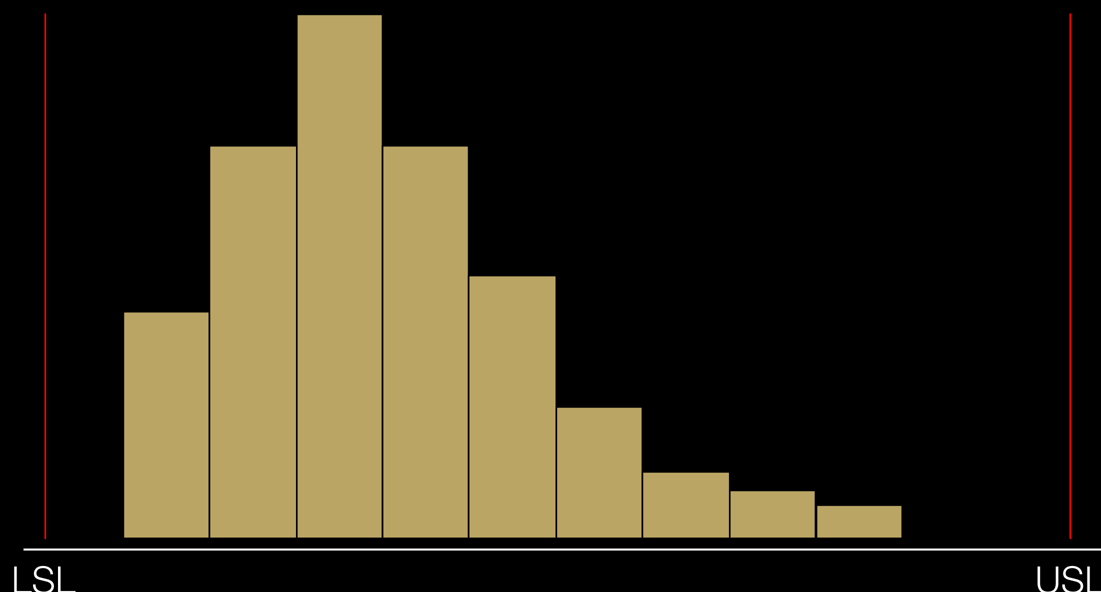
Histogram Patterns

The center, spread and shape of a histogram can give us clues as to what the data are telling us.

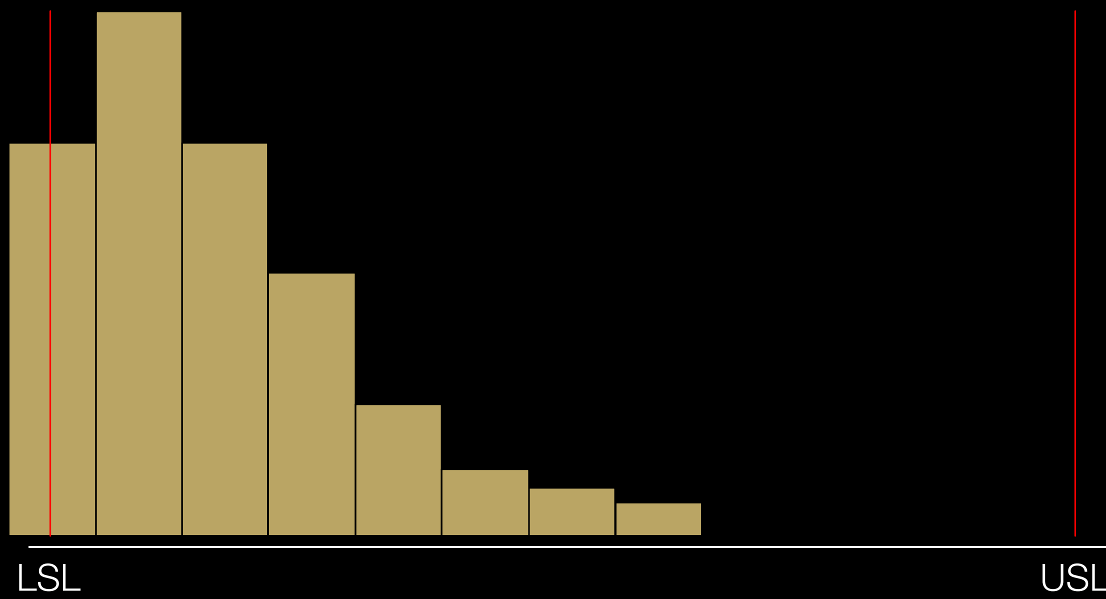
Pattern 1



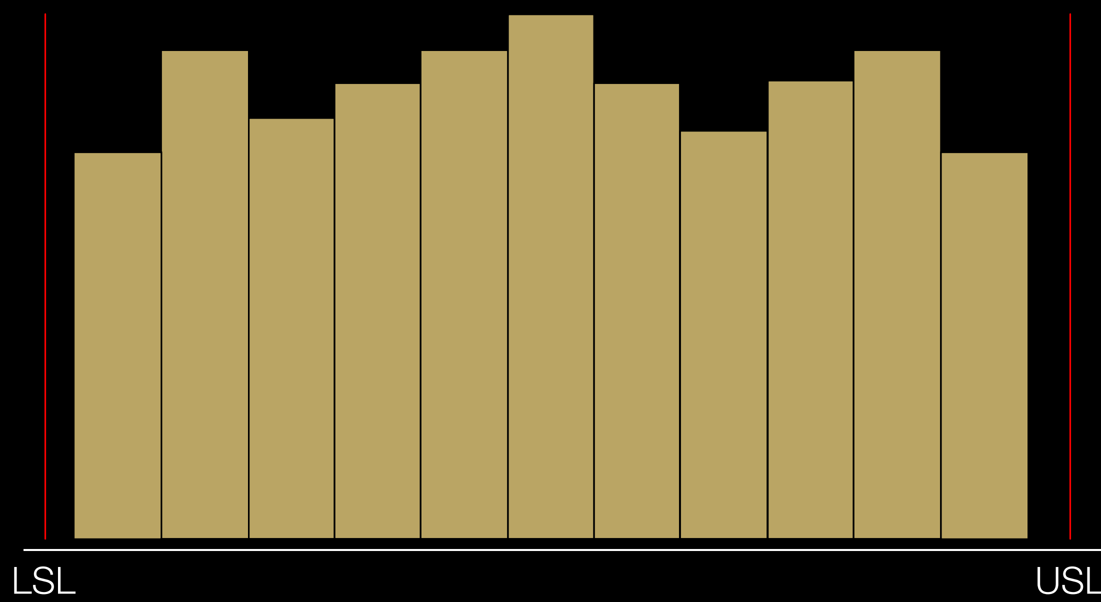
Pattern 2



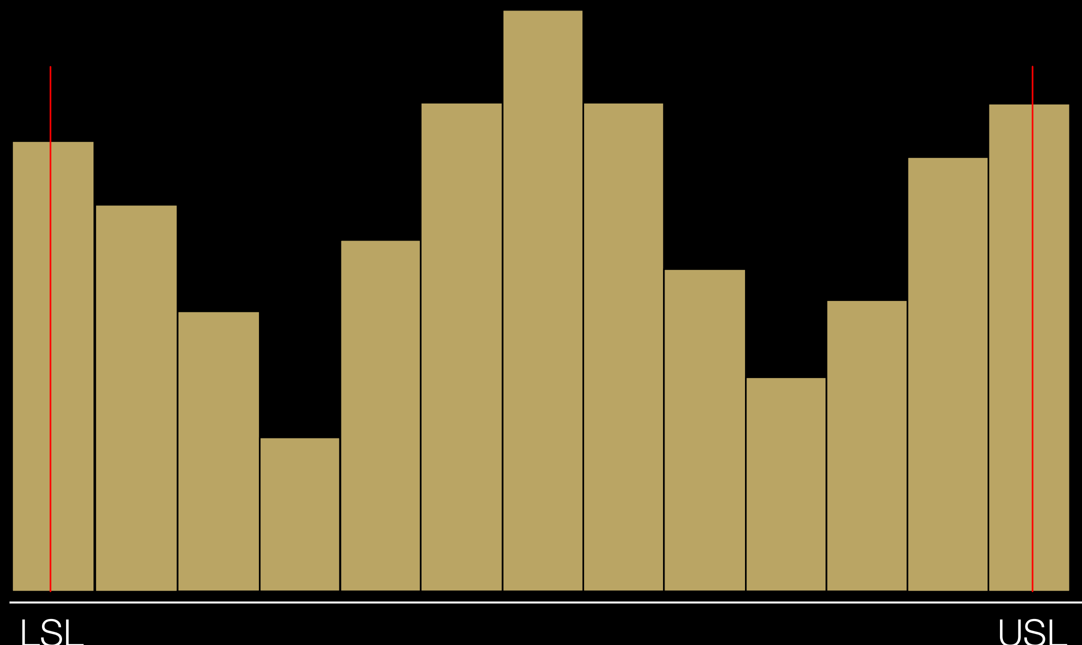
Pattern 3



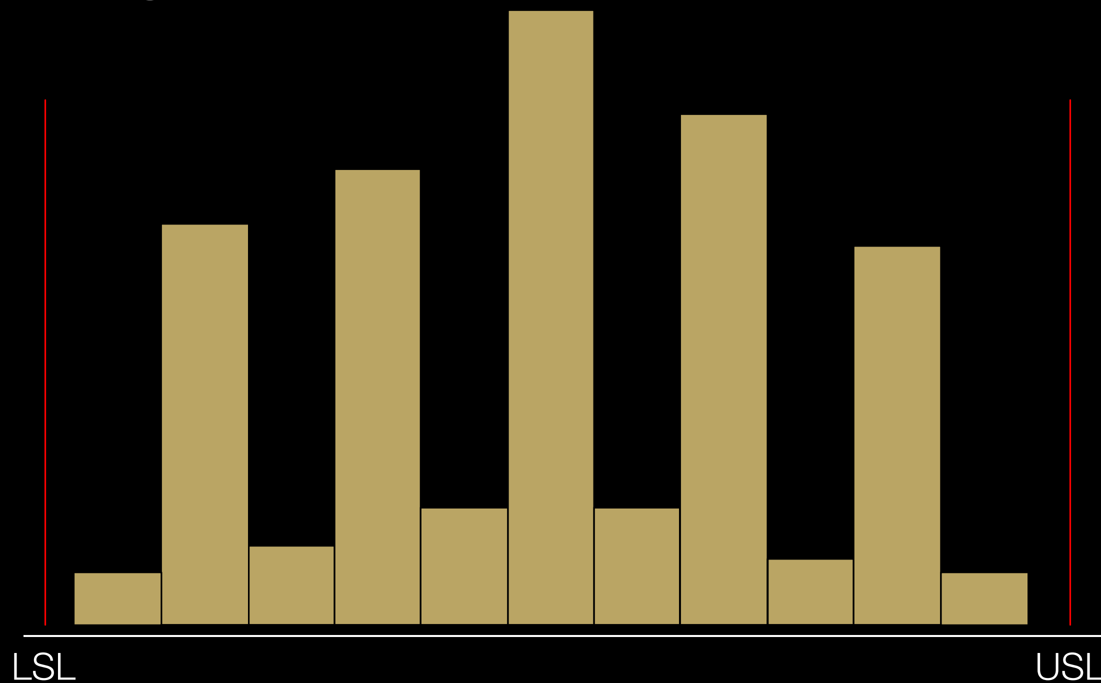
Pattern 4



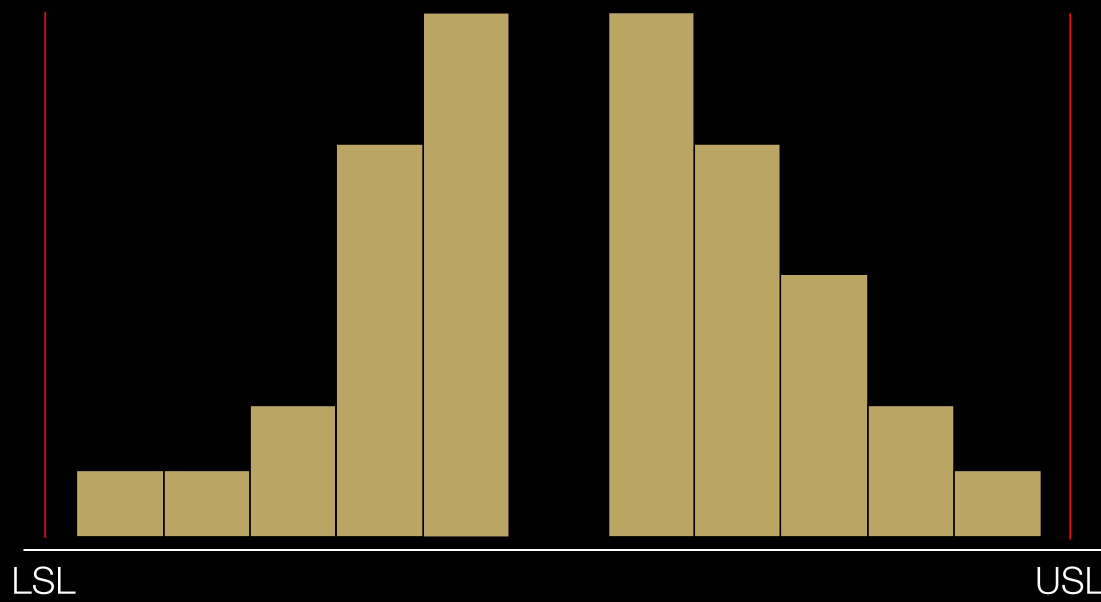
Pattern 5



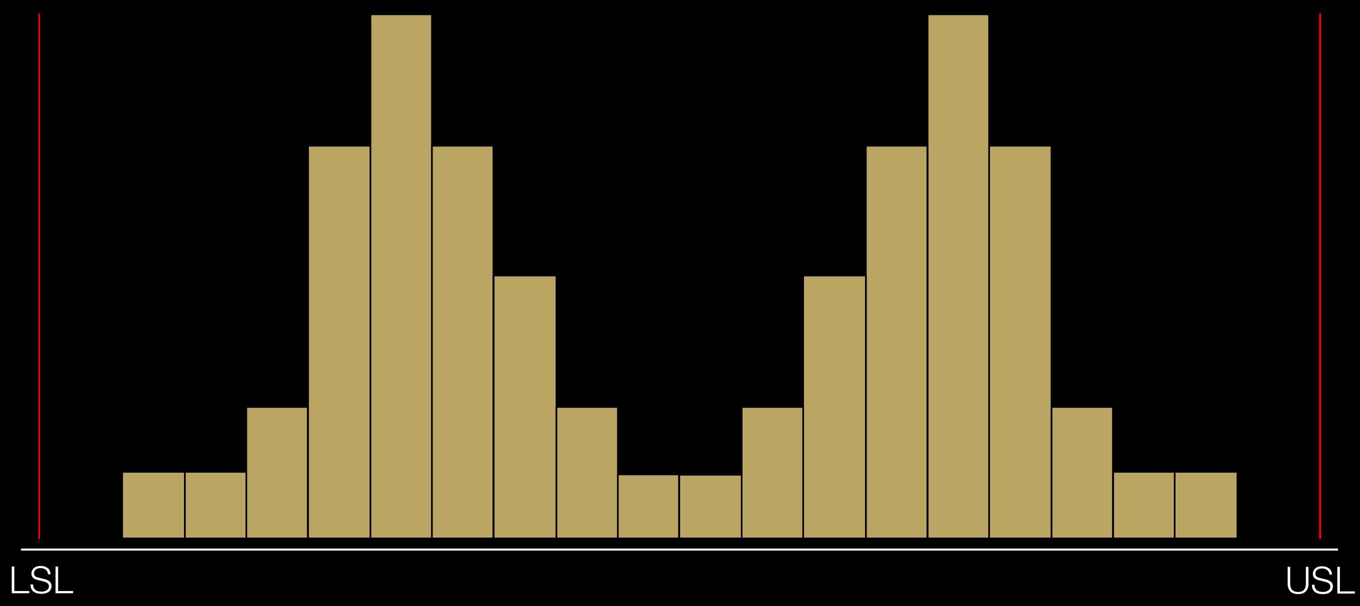
Pattern 6



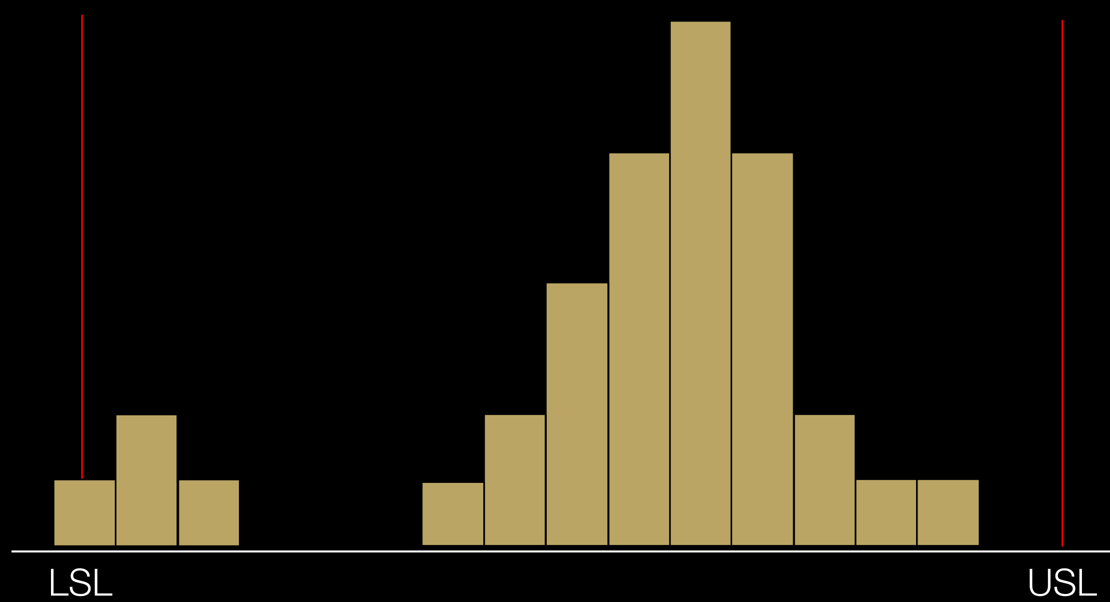
Pattern 7



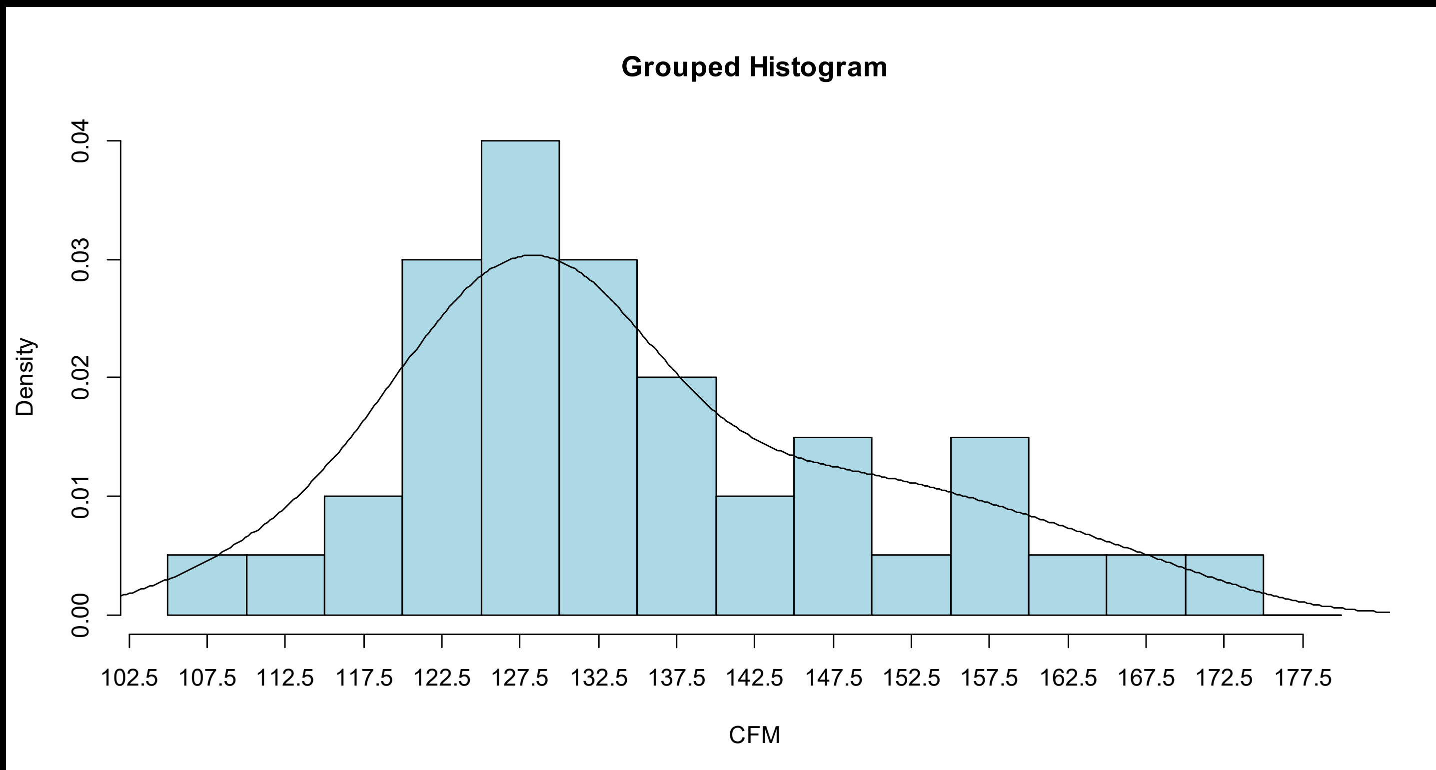
Pattern 8



Pattern 9



Density Plot



Density Plot

Similar to the frequency polygon, in that it is used with continuous data

Used to visualize an underlying probability distribution

Density Plot

When the data are continuous, we can use a density plot over a histogram.

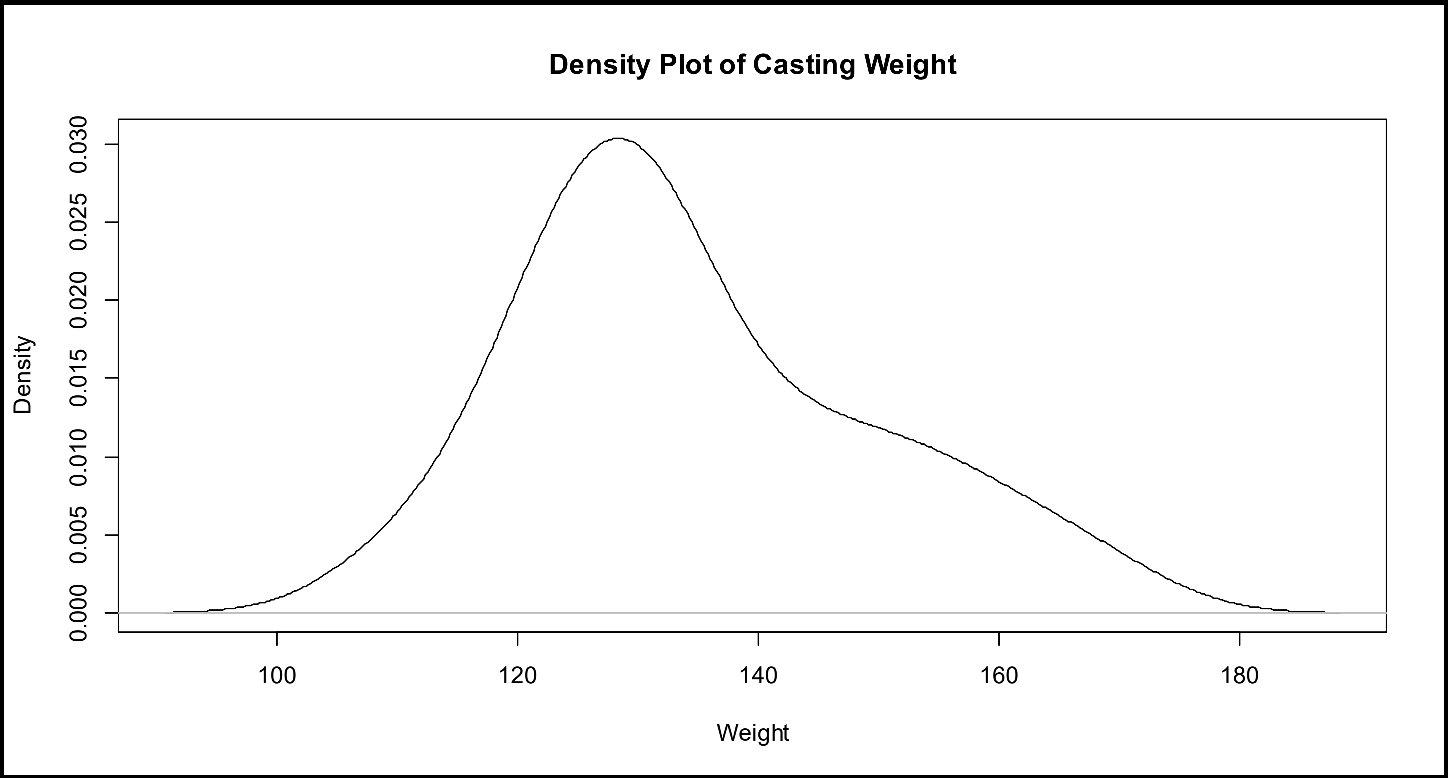
```
> hist.grouped(castings$weight, freq=F)
> lines(density(castings$weight))
```

Density Plot

The density plot can also be plotted without a histogram:

```
> plot(density(castings$weight)
```

Density Plot

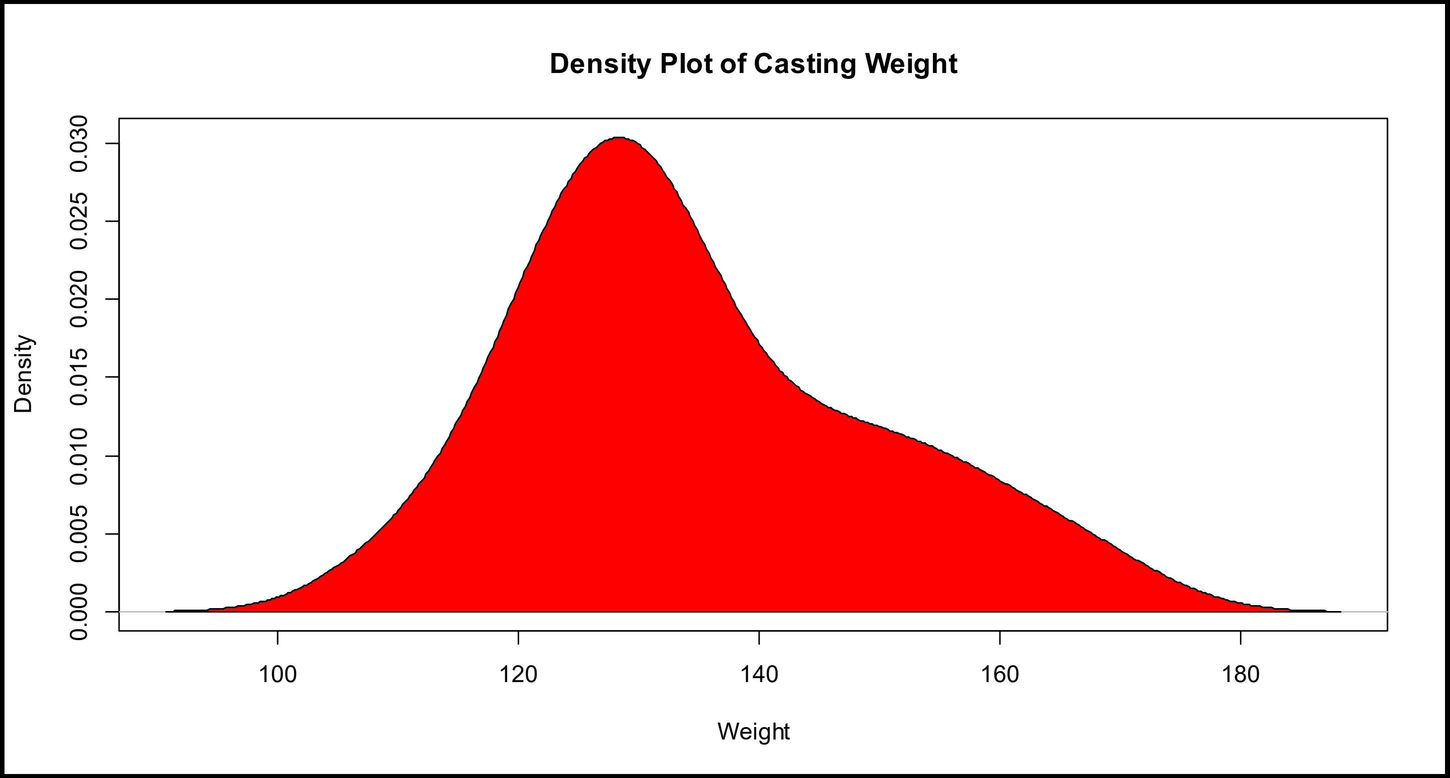


Density Plot

To fill a density plot with color:

```
> dp<-density(castings$weight)
> plot(dp, main="Density Plot of Casting
Weight", xlab="Weight")
> polygon(dp, col="red", border="black")
```

Density Plot



Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Box and Whisker Plots

Data Science for Quality Management: Describing Data Graphically

with **Wendy Martin**

Learning objective:

Create a Box and Whisker Plot using RStudio

Box and Whisker Plot

Box & Whisker Plots are used to display data corresponding to Percentiles, and typically from two or more sources or process streams, simultaneously

Box and Whisker Plot

One distinct advantages of this display is that the two sample data sets do not have to possess the same shape, but are directly comparable nonetheless.

Box and Whisker Plot

A second major advantage is that the Box & Whisker plot can display outliers; which we will see later can represent Special Causes of Variation.

5 Number Summary

Maximum

Q3 (3rd Quartile)

Median (Q2) (2nd Quartile)

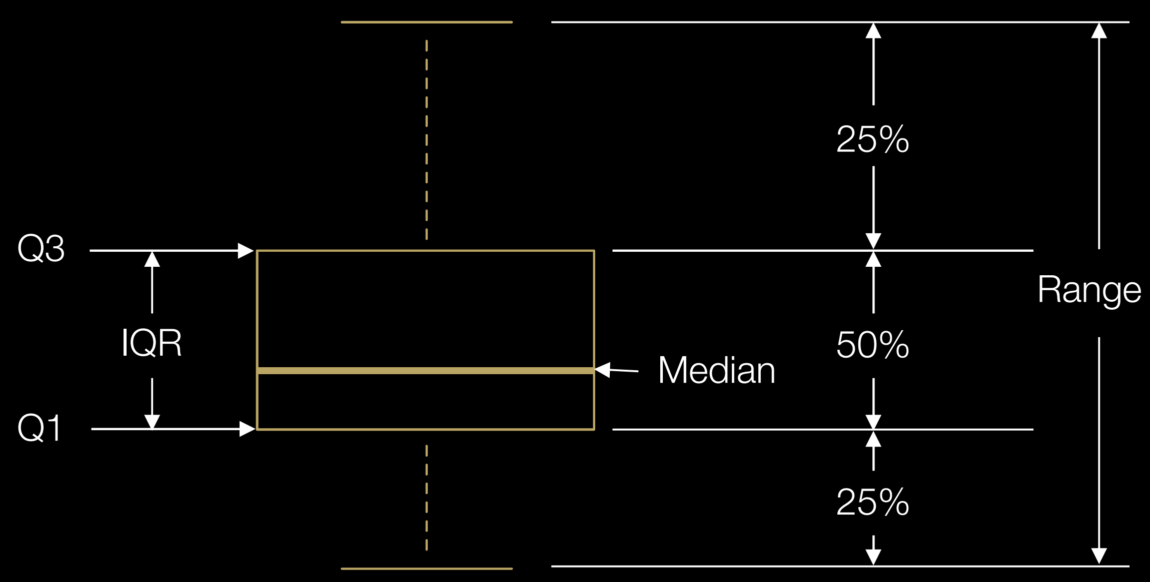
Q1 (1st Quartile)

Minimum

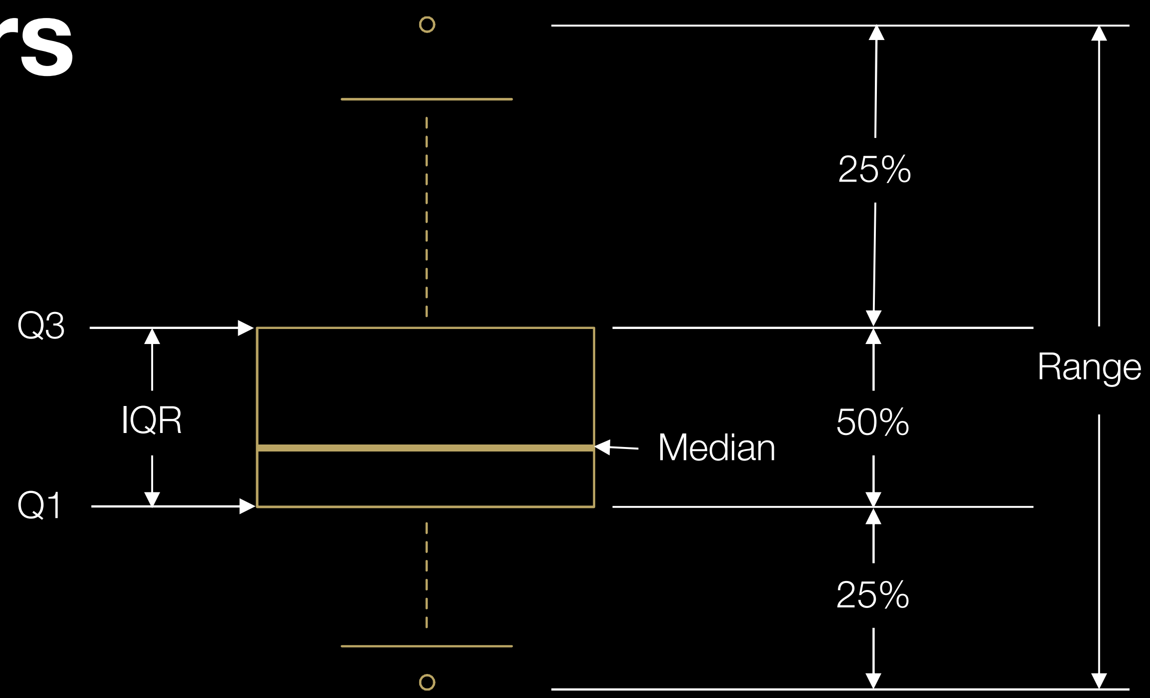
5 Number Summary

```
> summary(castings$weight)
```

Box and Whisker Plot



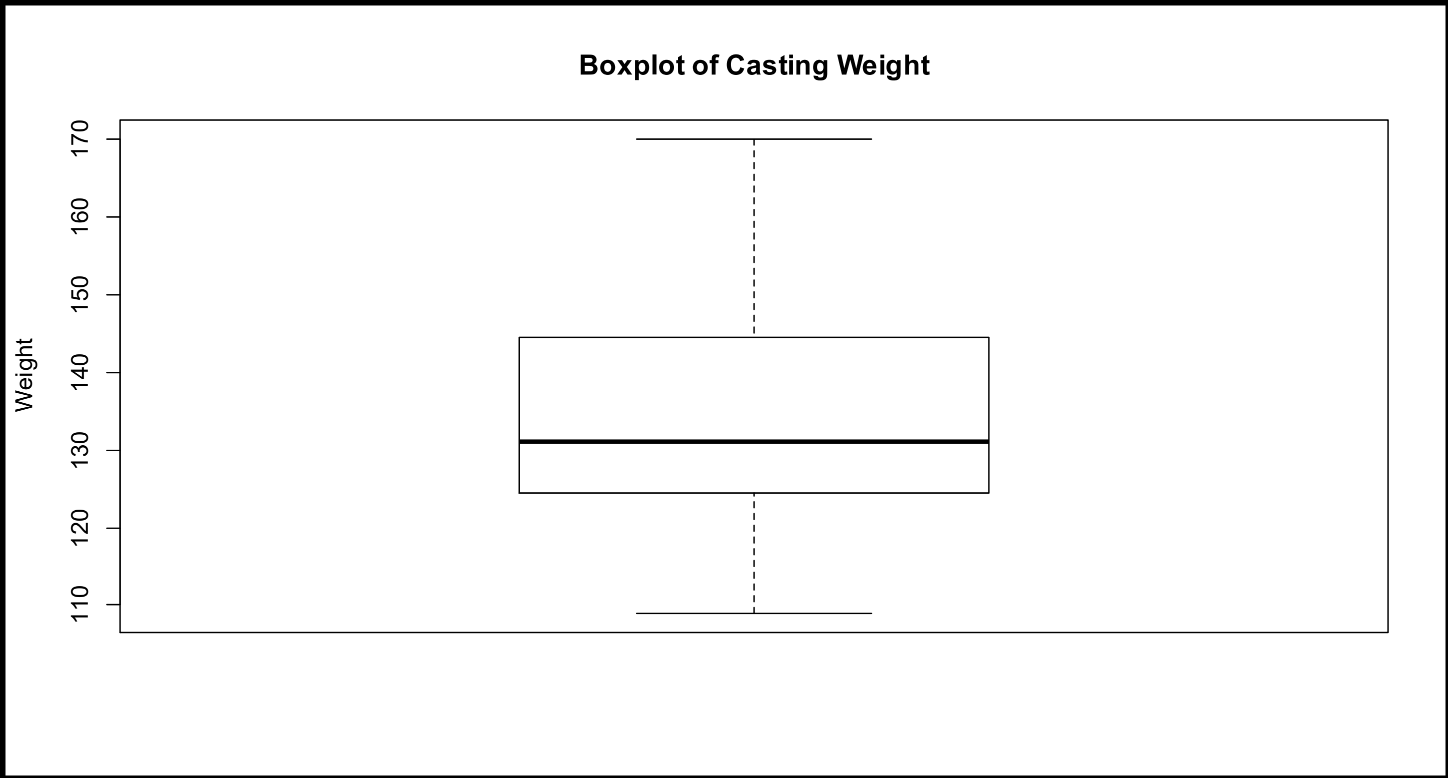
Box and Whisker Plot with Outliers



Box and Whisker Plot in R

```
> boxplot(castings$weight)
```

Box and Whisker Plot Example

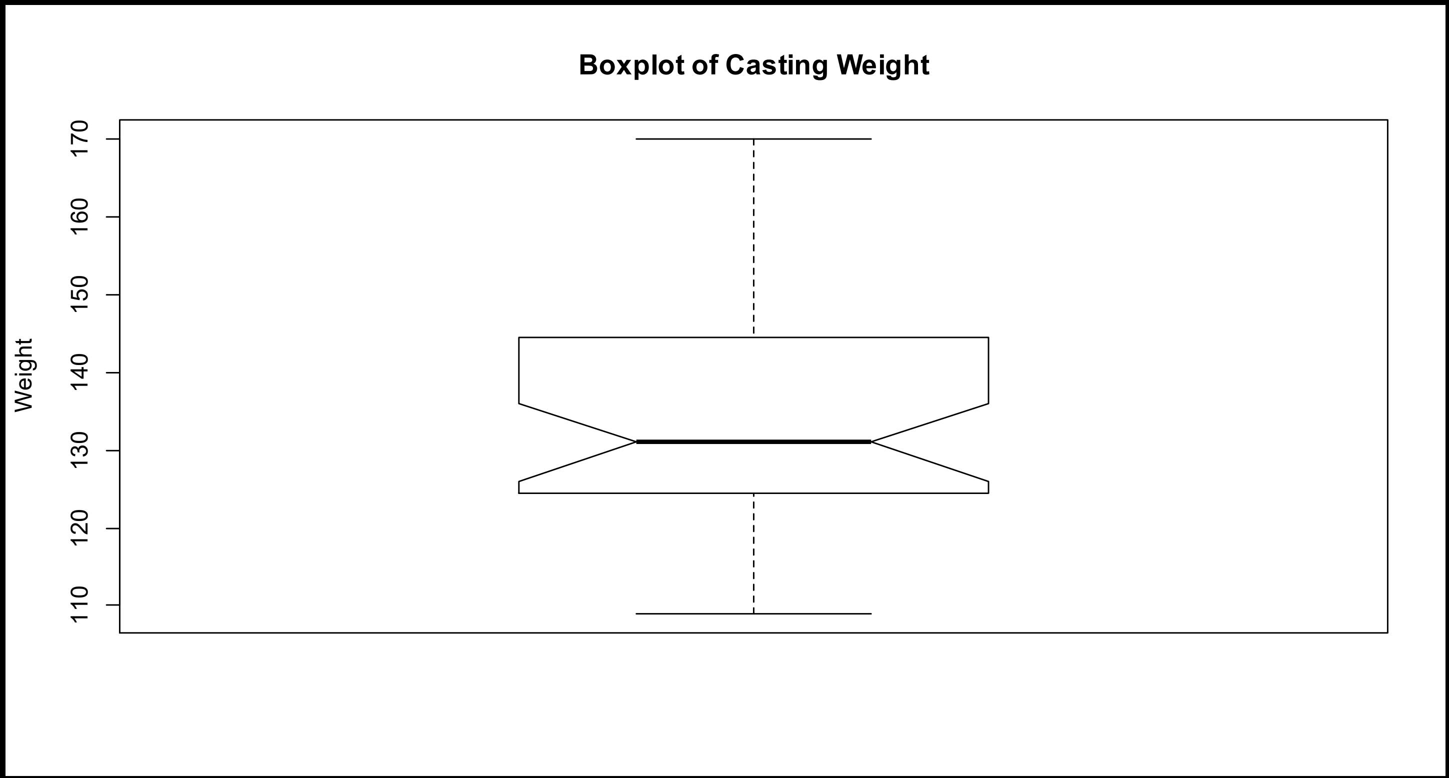


Notched Box and Whisker Plot

A notched Box and Whisker plot shows the 95% confidence interval of the median.

```
> boxplot(castings$weight, notch=T)
```

Notched Box and Whisker Plot



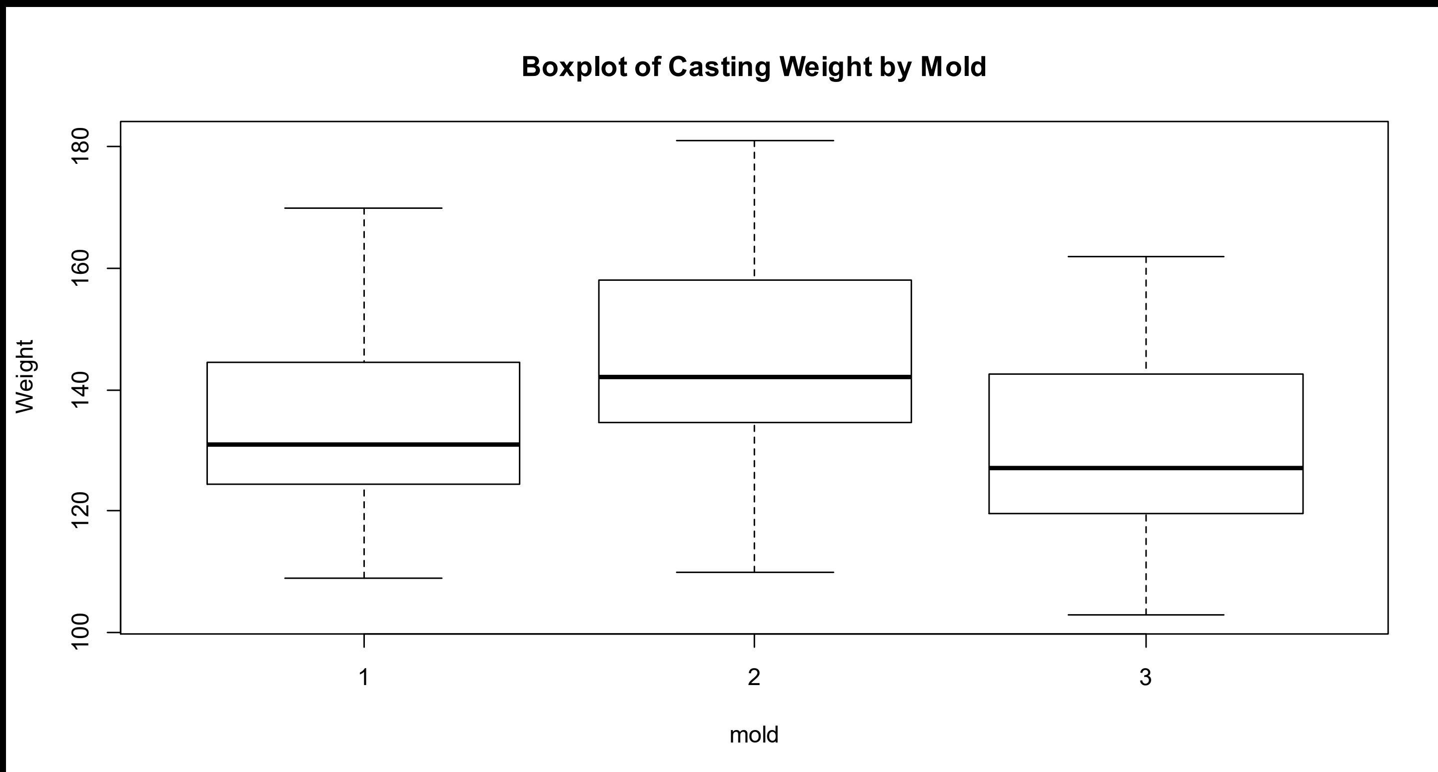
Boxplot to Compare Groups

```
> boxplot(y ~ x, data = data.frame)
```

Boxplot to Compare Groups

```
> boxplot(y ~ x, data = data.frame)
> boxplot(weight ~ mold, data = castings3)
```

Boxplot to Compare Groups



Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Measures of Central Tendency

Data Science for Quality Management: Describing Data Numerically

with **Wendy Martin**

Learning objectives:

Calculate the sample mean for ungrouped and grouped data and the weighted mean

Calculate the sample median for ungrouped data

Find the sample mode or modes

5 Aspects of Data

Location or Central Tendency

Spread or Dispersion (Variability)

Shape

Time Sequence

Relationship

Sample Data

- Preforms for a compression molding process were randomly sampled
- Sample size (n) is 10
- Each Preform was then weighed on a gram scale

Sample Data

- Suppose the resultant data appeared as:

  65 67 36 37 36 57 53 39 38 58

- We will use this sample data set to demonstrate the calculation of various statistics

Create Data File:

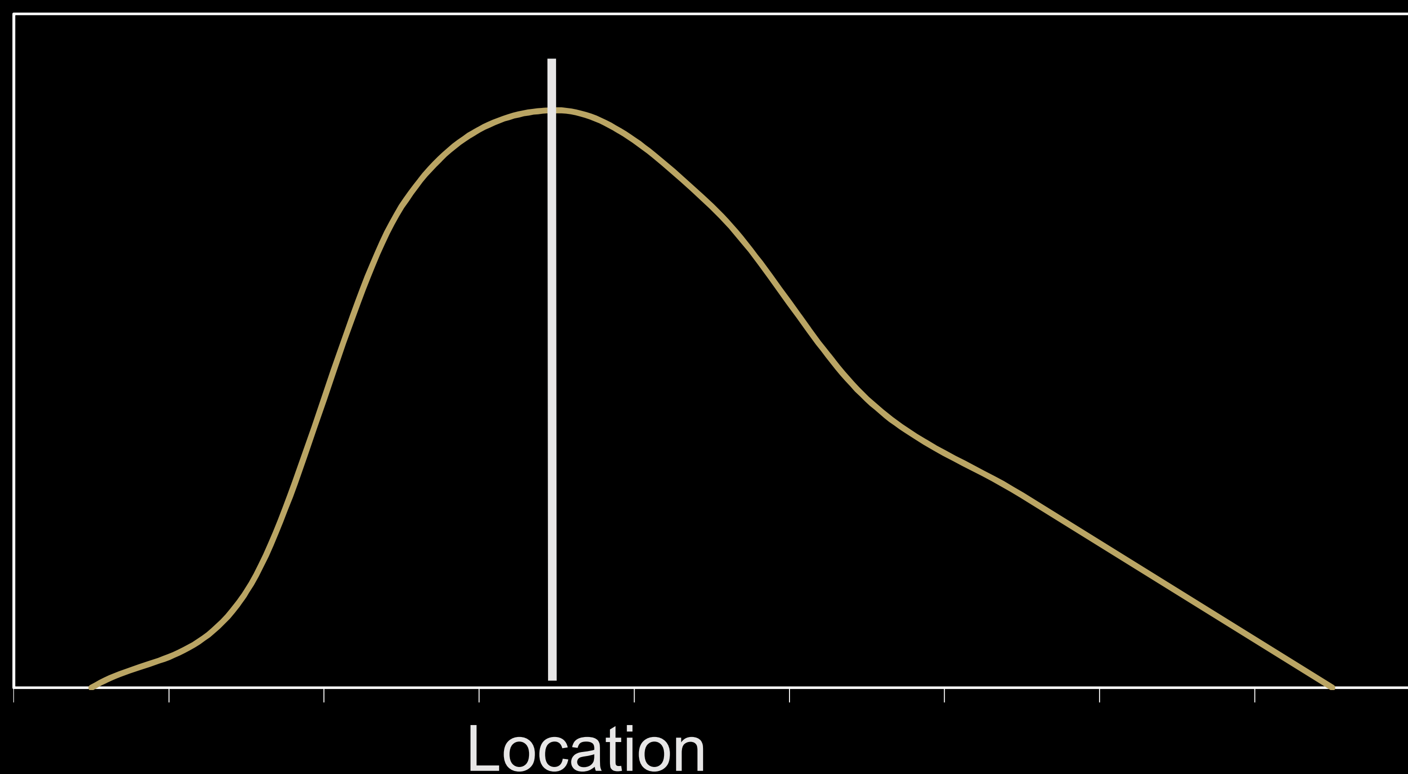
- Create a vector:

```
weight <-c(65,67,36,37,36,57,53,39,38,58)
```

- Store the variable in a data frame:

```
preform <- data.frame(weight)
View(preform)
```

Measures of Central Tendency



Measures of Central Tendency

Measures of location, sometimes called measures of central tendency, describe a middle or central point or tendency of a distribution.

- Mean, Median, Mode

The Mean

- Arithmetic average
- Can be thought of as the “center of gravity” of the frequency distribution
- The value in which the sum of all deviations from this value are zero
- Symbols: population (μ) and sample ($\bar{X}$)

Mean: Calculations

- Ungrouped Data: $\bar{X} = \frac{\sum X}{n}$

- Grouped Data: $\bar{X} = \frac{\sum f X_c}{n}$

- Weighted Mean: $\bar{X} = \frac{\sum w_j X}{w_j n_j}$

Mean: Advantages

- Easy to understand
- Simple to calculate
- Every data set possesses an arithmetic mean

Mean: Disadvantages

- Affected by extreme measures or values

Mean: Example

- For our ungrouped preform data set, the calculation for the mean is as follows:
- Ungrouped Data: $\bar{X} = \frac{\sum X}{n} = \frac{486}{10} = 48.6$

How to Calculate in RStudio

- In R Studio:

```
> mean(preform$weight)
```

Mean for Grouped Data

- Formula for Grouped Data: $\bar{X} = \frac{\sum f X_c}{n}$

where

- X_c = the midpoint of each class interval
- f = the frequency associated with each class interval

Mean for Grouped Data: Example

- Frequency Distribution for the Casting Weight data from Module 2

	l	min	midpoint	max	u	freq	rel.freq	cum.up	cum.down
1	[	105	107.5	110	)	1	0.025	0.025	1.000
2	[	110	112.5	115	)	1	0.025	0.050	0.975
3	[	115	117.5	120	)	2	0.050	0.100	0.950
4	[	120	122.5	125	)	6	0.150	0.250	0.900
5	[	125	127.5	130	)	8	0.200	0.450	0.750
6	[	130	132.5	135	)	6	0.150	0.600	0.550
7	[	135	137.5	140	)	4	0.100	0.700	0.400
8	[	140	142.5	145	)	2	0.050	0.750	0.300
9	[	145	147.5	150	)	3	0.075	0.825	0.250
10	[	150	152.5	155	)	1	0.025	0.850	0.175
11	[	155	157.5	160	)	3	0.075	0.925	0.150
12	[	160	162.5	165	)	1	0.025	0.950	0.075
13	[	165	167.5	170	)	1	0.025	0.975	0.050
14	[	170	172.5	175	)	1	0.025	1.000	0.025

Mean for Grouped Data: Example

min	midpoint (Xc)	max	freq (f)	f*Xc
105	107.5	110	1	107.5
110	112.5	115	1	112.5
115	117.5	120	2	235.0
120	122.5	125	6	735.0
125	127.5	130	8	1020.0
130	132.5	135	6	795.0
135	137.5	140	4	550.0
140	142.5	145	2	285.0
145	147.5	150	3	442.5
150	152.5	155	1	152.5
155	157.5	160	3	472.5
160	162.5	165	1	162.5
165	167.5	170	1	167.5
170	172.5	175	1	172.5
		Totals	**40**	**5410.0**

$$\bar{X} = \frac{\sum f X_c}{n} = \frac{5410}{40}$$

$$= 135.25$$

How to Calculate in RStudio

- In R Studio:

```
> fdcast<-
frequency.dist.grouped(castings$weight)
> (midpts<-fdcast$midpoint)
> (freq<-fdcast$freq)
> weighted.mean(x = midpts, w = freq)
```

Weighted Mean

- Formula for Weighted Mean: $\bar{X}_w = \frac{\sum wX}{\sum w}$

where

- X = a value
- w = the weight associated with a value

Weighted Mean: Example

In a statistics class, there are three exams, each totaling 100 points. A student scores 88, 85 and 92. The first exam was easier than the last two, so it was weighted less.

Weighted Mean: Example

- Exam 1: 20 % of the grade (0.2 in decimal form)
- Exam 2: 40 % of the grade (0.4 in decimal form)
- Exam 3: 40 % of the grade (0.4 in decimal form)

- What is the final weighted mean for the student in the class?

Weighted Mean: Example

$$\bar{X}_w = \frac{\sum wX}{\sum w}$$

$$= \frac{(0.2 * 88) + (0.4 * 85) + (0.4 * 92)}{(0.2 + 0.4 + 0.4)}$$

$$= \frac{17.6 + 34 + 36.8}{1} = 88.4$$

How to Calculate in RStudio

- In R Studio:

```
> wt<-c(0.2, 0.4, 0.4)
> x<-c(88, 85, 92)
> weighted.mean(x = x, w = wt)
```

The Median

- The median is the value at or below which 50% of the data fall, or at or above which 50% of the data fall
- The median is a measure of position and is the middle value in a sorted array of data
- Symbols: population (M) and sample ($\tilde{X}$)

Median: Example

Values	Median
2 4 6 12 14	6
2 4 6 55 99	6
1 4 5 5 5	5
1 2 5 6 12 15	5.5

Median: Example

For our ungrouped preform data set:

- First, the data set is sorted from low to high
- 36 36 37 38 39 53 57 58 65 67

Median: Example

- We note the median may be found in the (n + 1)/2th position, or (10 + 1)/2 = 5.5 position
- 36 36 37 38 39 53 57 58 65 67

How to Calculate in RStudio

- In R Studio:

```
> median(preform$weight)
```

Median: Advantages

- Easy to understand
- Not affected by extreme values

Median: Disadvantages

- The median does not take the relative magnitude of the values into account

The Mode

- The mode is the most frequently occurring value in a data set
- For a population, the mode is the peak of the population distribution curve
- Symbols: population (M_o) and sample (X_{mode})

Mode: Example

- For our preform data set (sorted)
- 36 36 37 38 39 53 57 58 65 67

- The mode is 36

Mode: Advantages

- Not affected by extreme values
- Can be used with categorical data

Mode: Disadvantages

- The data set may not have a modal value. For example, it is possible that no two values are alike
- The data set may contain too many modal values to be useful

How to Calculate in RStudio

- In R Studio:

```
> table(preform$weight) or
> sample.mode(preform$weight)
```

Example: Central Tendency

$170,000 $170,000 $170,000 $170,000

Mean = $170,000

Median = $170,000

Mode = $170,000

Example: Central Tendency

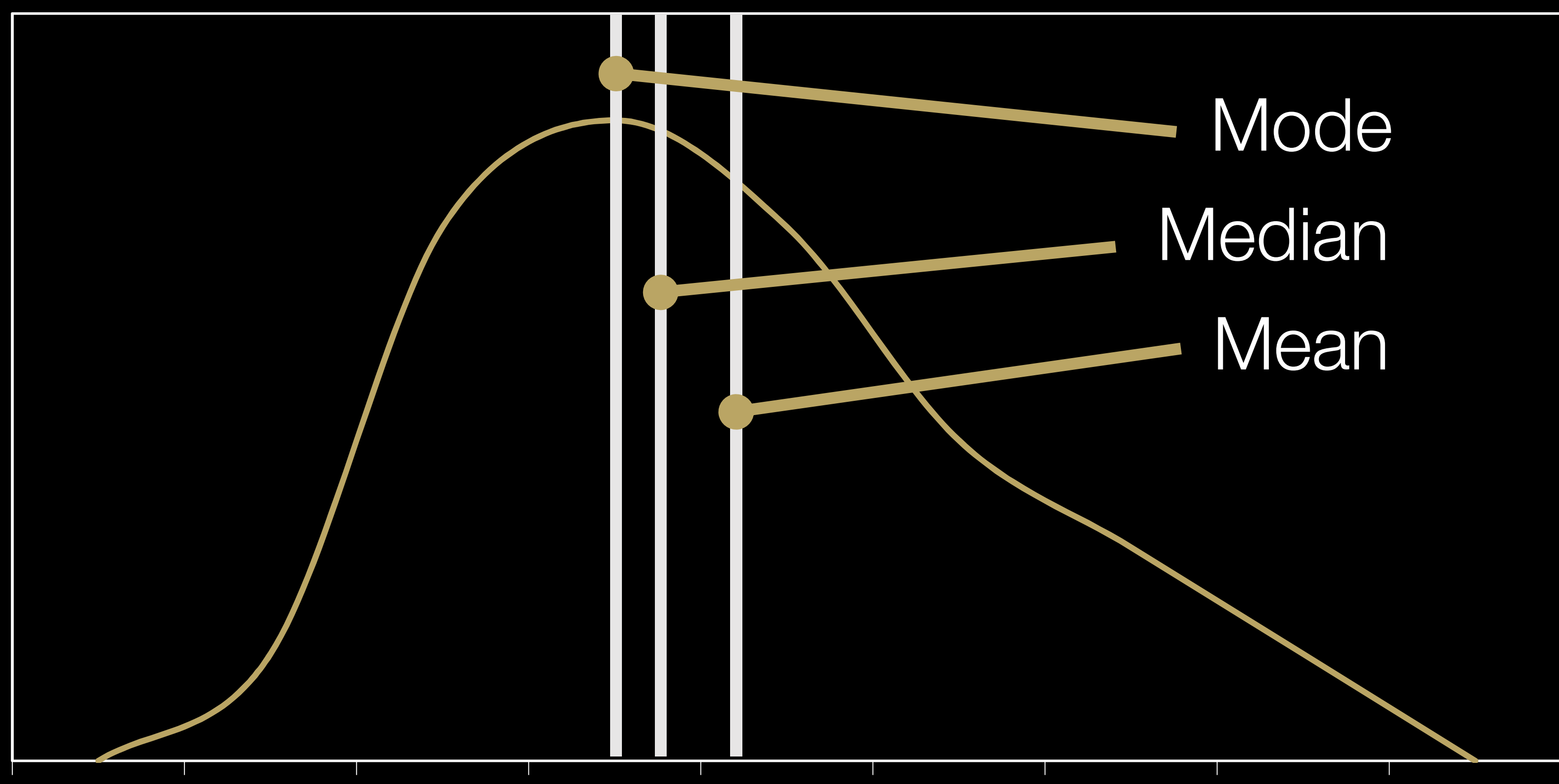
$170,000 $170,000 $170,000 $170,000
$17,000,000

Mean = $3.536 Million

Median = $170,000

Mode = $170,000

Measures of Location



Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Measures of Position

Data Science for Quality Management: Describing Data Numerically

with **Wendy Martin**

Learning objectives:

Calculate measures of position

Find the low and high values of a data set

Find sample quartiles and percentiles

Measures of Position

Measures of position, or relative standing, display values representing position or order in the data set or distribution

Measures of Position

They describe the relationship of a measure to the rest of the data

- Low and High
- Percentiles
- Quartiles

Low and High Scores

- Low and high are the lowest and highest values in the data set. These may not exist for populations, unless bounded.

- Symbols: sample (X_L and X_H)

How to Calculate in RStudio

- In R Studio:

```
> min(preform$weight)
> max(preform$weight)
```

Or

```
> summary(preform$weight)
```

Percentiles

- The P^{th} percentile is the value that P% of the values fall at or below and (100 - P%) fall above it
- Symbols: no common symbols used, but generally written simply as ”P^{th} percentile"

Percentiles: Calculations

- First sort the data from low to high
- The Pth percentile is found in the 1+P(n-1)/100th position (P in a proportion)

Percentiles: Example

For our preform data set, 30th percentile:

- Data sorted from low to high:
- 36 36 37 38 39 53 57 58 65 67
- The 30th percentile is found in the $1 + 0.30(n-1)^{th}$ position, or $1 + 0.30(10-1) = 3.7$ (between the 3rd and 4th value)

Percentiles: Example

- 36 36 37 38 39 53 57 58 65 67
- Using the fraction of 3.7 (0.7), the percentile is 0.7 times the range between the 3rd and 4th value above the 3rd value or 37+0.7(38-37)
- The 30th percentile is 37.7

How to Calculate in RStudio

- In R Studio:

```
> quantile(x = preform$weight, probs =
0.30)
```

Read more at:
https://www.rdocumentation.org/packages/stats/versions/3.4.3/topics/quantile

Quartiles

- Quartiles are the 25th, 50th, 75th and 100th percentiles

- Symbols: Q_i

Quartiles: Calculations

- Q1: The 25th percentile, found in the $1+(n-1)/4^{th}$ position $\{1+((n-1)*.25^{th})\}$
- Q2: The median $\{1+(n-1)*.50^{th}\}$
- Q3: The 75th percentile, found in the $1+3(n-1)/4^{th}$ position $\{1+(n-1)*.75^{th}\}$
- Q4: The highest value, X_H

Quartiles: Example

For our preform data set, the 1st and 3rd quartiles are found as follows

- Q1 is found in the (1+(10-1)/4th position or 3.25 = 37.25
- Q3 is found in the (1+3(10-1)/4th position or 7.75 = 57.75

How to Calculate in RStudio

- In R Studio:

```
> quantile(x = preform$weight, probs =
0.25)

Or
> summary(preform$weight)
```

Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Measures of Dispersion

Data Science for Quality Management: Describing Data Numerically

with **Wendy Martin**

Learning objectives:

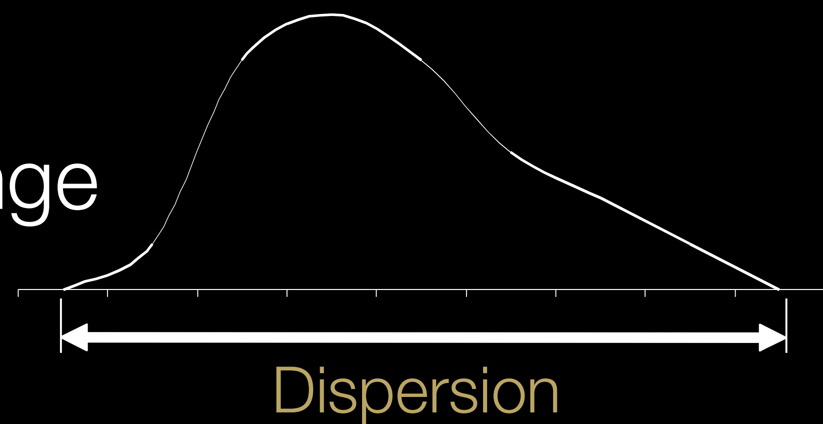
Calculate the sample range
Calculate the interquartile range
Calculate the sample standard deviation
Calculate the sample variance

Measures of Dispersion

Measures of dispersion reflect the variation or spread in a data set or distribution. Some of the common measures of dispersion are:

Measures of Dispersion

- Range
- Interquartile Range
- Semi-Interquartile Range
- Standard Deviation
- Variance



The Range

- The range is the difference between the highest and lowest value in a data set
- Symbols:
  Population (generally does not exist)
  Sample (R)

The Range

- Calculations: $R = X_H - X_L$
- Example:
- For our sample data set, the low is 36 and the high is 67
- The range is: $R = 67 - 36 = 31$

The Range

Advantages

- Depends on only two values - Maximum minus minimum
- Easy to understand

Disadvantages

- Extremely sensitive to “outliers”

How to Calculate in RStudio

- In R Studio:

```
> range(preform$weight)
> rng<-range(preform$weight)
> rng[2]-rng[1]
```

The Interquartile Range

- The Interquartile Range is the range of the middle 50% of the data or distribution
- Symbols:
  Population or sample, IQR or IQ range
- Calculations:
  IQR = Q3 - Q1

Interquartile Range: Example

- For our preform data set, Q1 is 37.25 and Q3 is 57.75, the interquartile range is:

- IQR = 57.75 – 37.25 = 20.5

How to Calculate in RStudio

- In R Studio:

```
> IQR(preform$weight)
```

The Standard Deviation

- The standard deviation is a measure of variation that includes all data values in its calculation
- The standard deviation is the square-root of the average squared distance values fall from the mean

Standard Deviation: Calculations

- For a sample

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n-1}}$$

Standard Deviation: Example

- For our sample data set, with a mean of 48.6

$$s = \sqrt{\frac{\sum(X - \bar{X})^2}{n - 1}} = \sqrt{\frac{\sum(X - 48.6)^2}{9}} = \sqrt{\frac{1442.40}{9}} = 12.66$$

Standard Deviation: Example 2

65 67 36 37 36 57 53 39 38 58

1. Calculate the mean: 48.6
2. Calculate deviations from the mean for each value

   16.4 18.4 -12.6 -11.6 -12.6 8.4 4.4 -9.6 -10.6 9.4

Standard Deviation: Example 2

3. Square each deviation

269.96	338.56	158.76	134.56
158.76	70.56	19.36	92.16
112.36	88.36		

4. Sum the squared deviations: 1442.40

Standard Deviation: Example 2

5. Divide the sum of the squared deviations by $(n - 1)$ and then take the square root of this value

   $s = 12.66$

How to Calculate in RStudio

- In R Studio:

```
> sd(preform$weight)
```

The Variance

- The variance is the square of the standard deviation
- The variance is the average squared distance values fall from the mean
- Symbols: Population (σ^2) and Sample (s^2)

Variance: Calculations

- For a sample

$$s^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

Variance: Calculations

- For our sample preform data set, in which the standard deviation is 12.6596 (using four decimal places), the variance is:
- $s^2 = (12.6596)^2 = 160.27$

How to Calculate in RStudio

- In R Studio:

```
> var(preform$weight)
```

Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Measures of Shape

Data Science for Quality Management: Describing Data Numerically

with **Wendy Martin**

Learning objectives:

Discriminate between skewness & kurtosis
Calculate the sample skewness & kurtosis

Measures of Shape

Measures of shape reflect the type of distribution sampled.

- Skewness is concerned with the symmetrical nature of the distribution, and
- Kurtosis is concerned with the peakedness of the distribution.

Skewness

- Skewness is the degree of departure from symmetry of a distribution

- Symbols
  Population (γ_3) and
  Sample (g_3)

Skewness

- Measures “lopsidedness”
- Symmetric distributions have zero skewness

Skewness: Calculations

- The most important group of measures of skewness and kurtosis use the third and fourth moments about the mean
- Moments about the mean are the average of the deviations from the mean raised to some power

Skewness: Calculations

- The r[th] moment about the mean is:

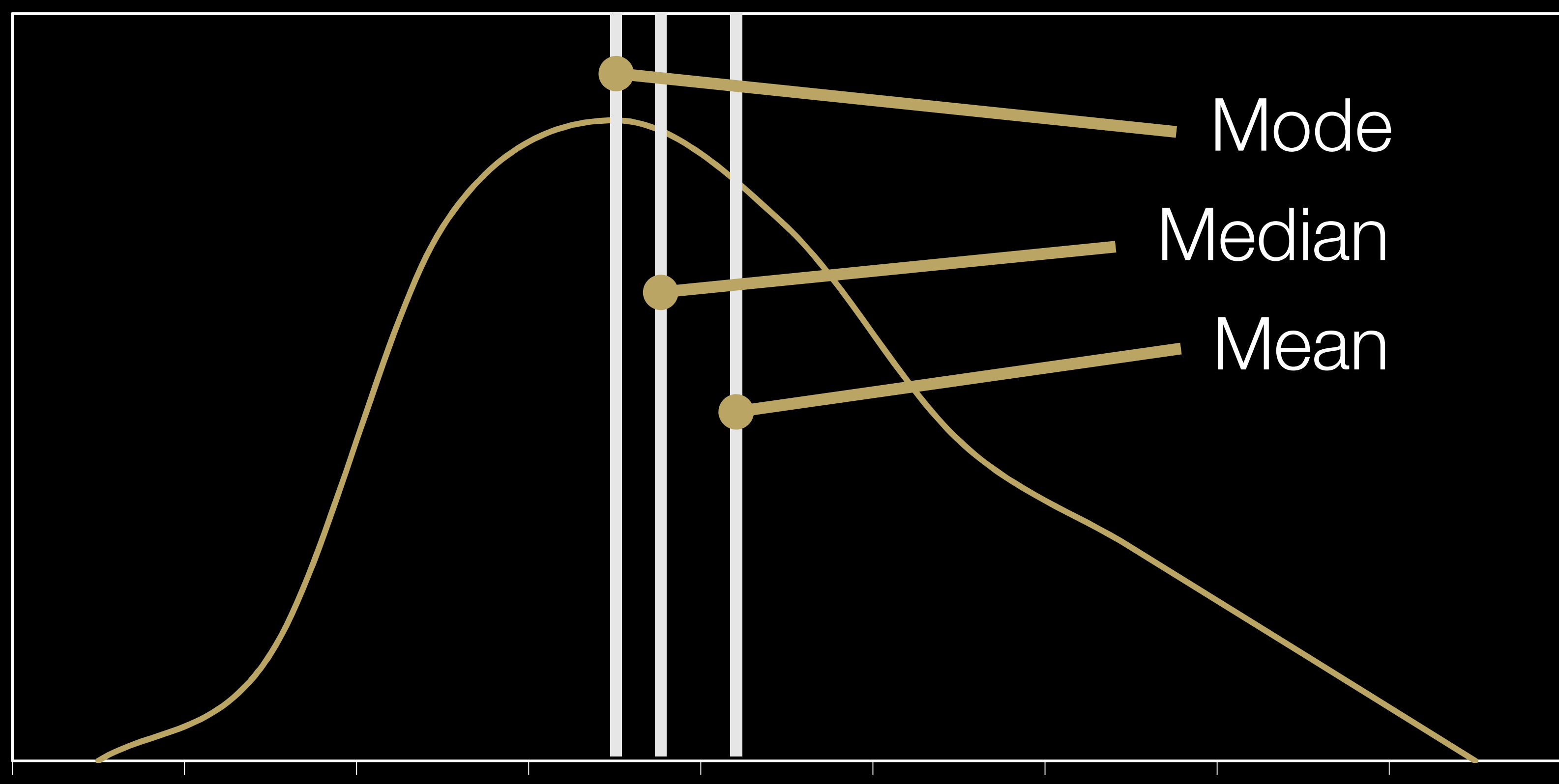
$$m_r = \frac{\sum(X - \bar{X})^r}{n}$$

Skewness: Calculations

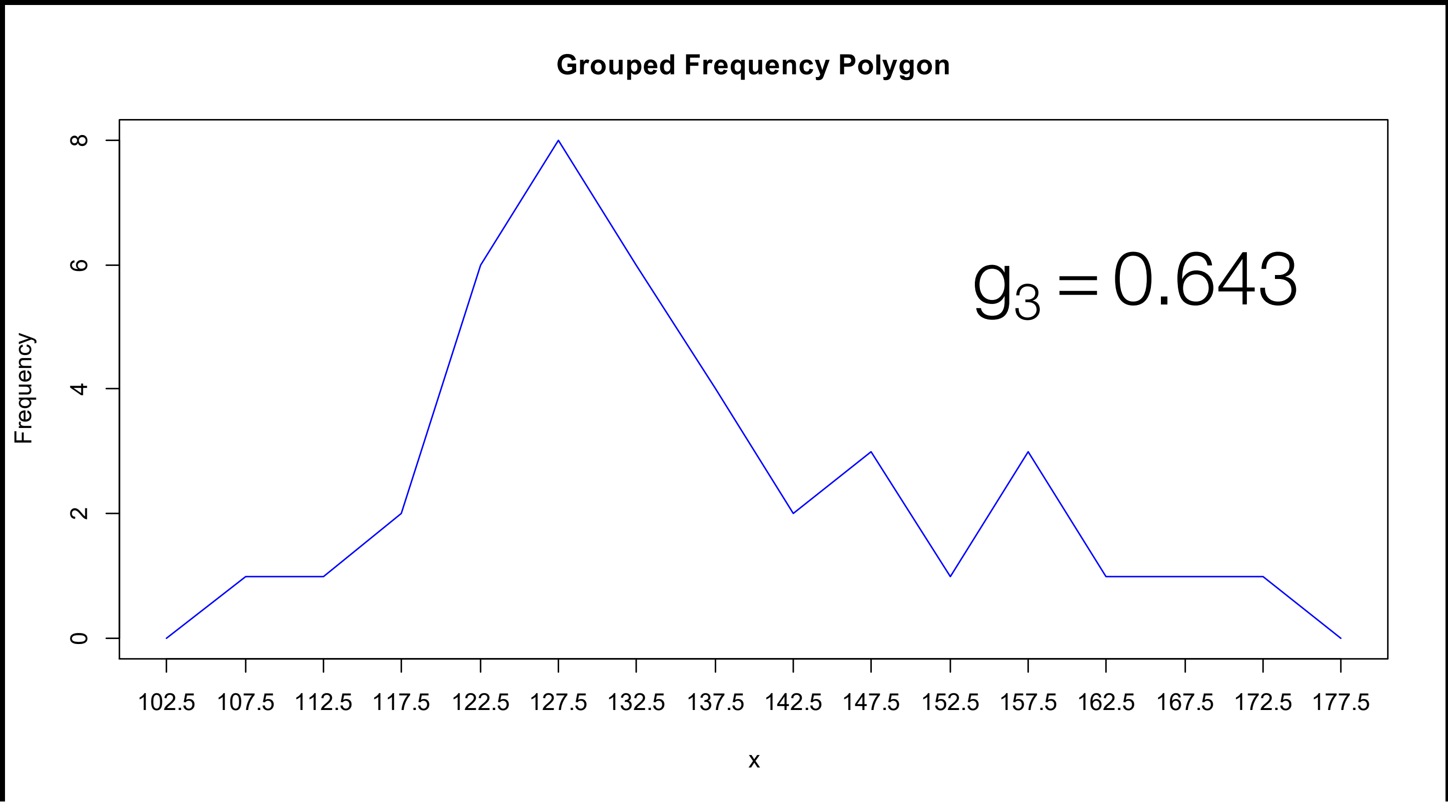
- A measure of skewness may then be calculated as follows
- The sign displays the direction of skewness

$$g_3 = \left[\frac{\sqrt{n(n-1)}}{n-2}\right] x \frac{m_3}{m_2^{3/2}}$$

Skewed Distributions



Skewed Distributions



How to Calculate in RStudio

- In R Studio:

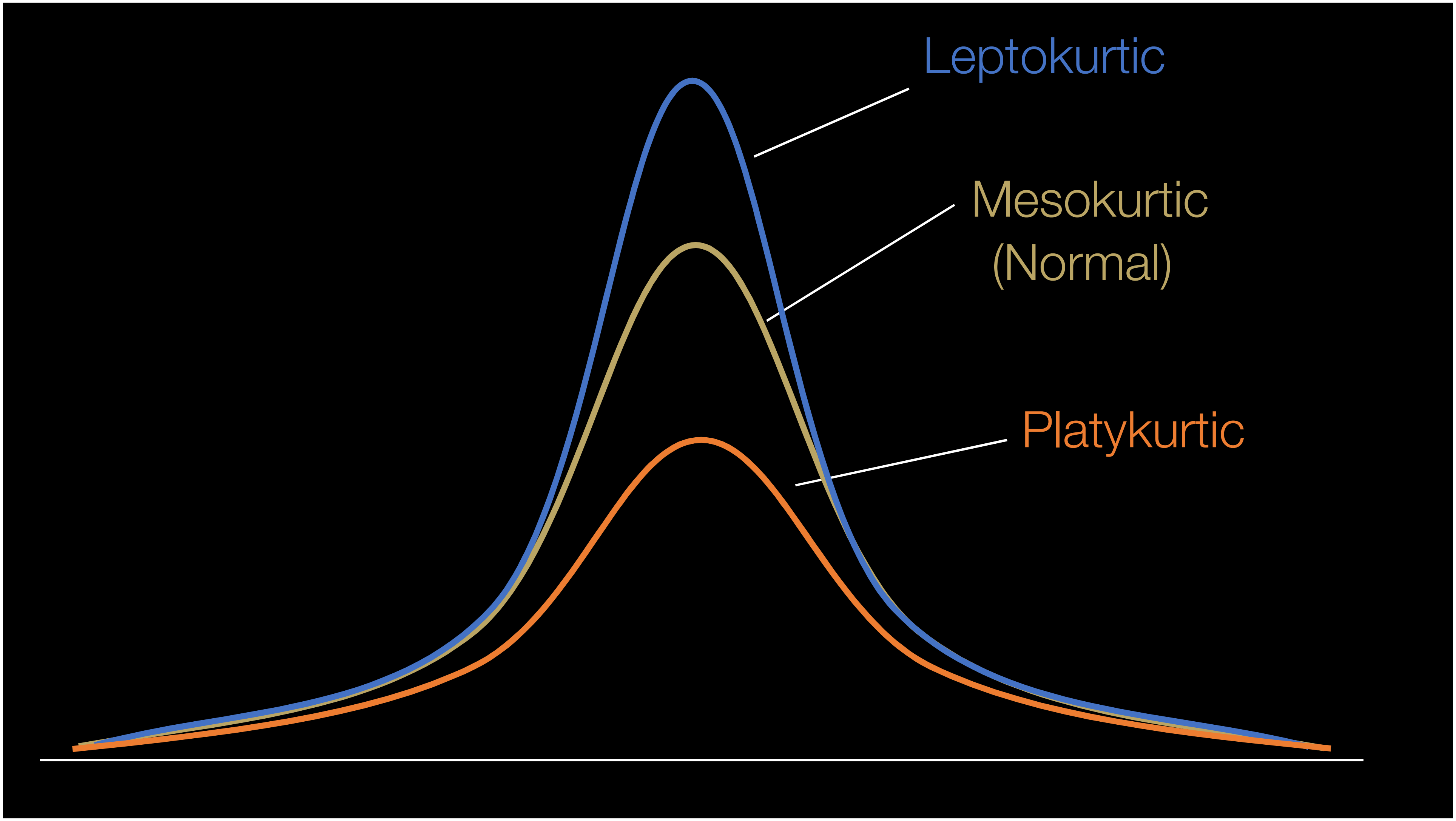
```
> skewness(castings$weight)
```

Kurtosis

- Kurtosis is the degree of peakedness of a distribution
- An intermediate distribution, with zero kurtosis, is known as a mesokurtic distribution

Kurtosis

- A symmetrical leptokurtic distribution has a higher peak and has heavier tails, and has positive kurtosis
- A symmetrical platykurtic distribution has a lower peak and lighter tails, and has negative kurtosis

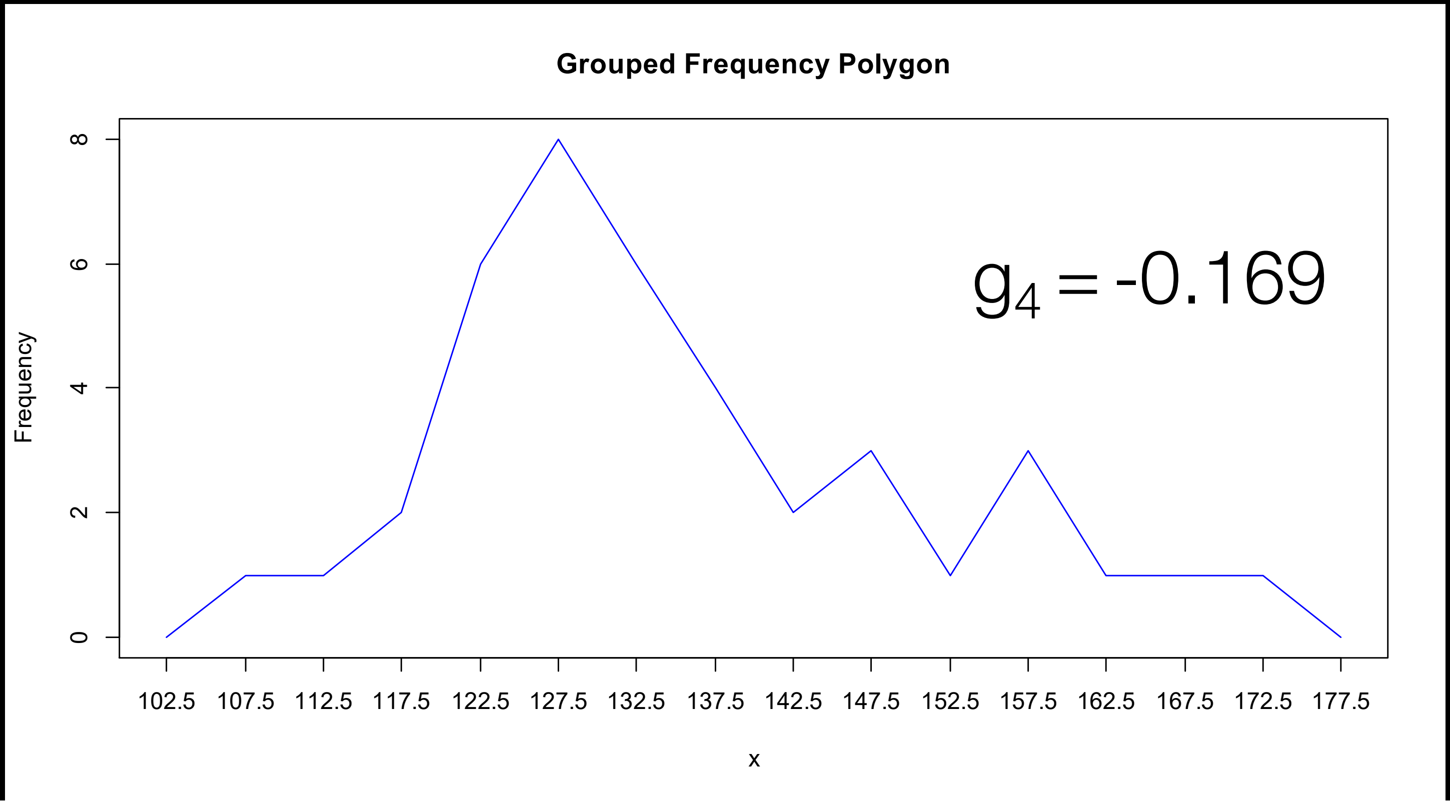

Leptokurtic
Mesokurtic
(Normal)
Platykurtic

Kurtosis: Calculations

- Symbols
  Population (γ_4) and
  Sample (g_4)

$$g_4 = \left[\frac{(n-1)(n+1)}{(n-2)(n-3)}\right] x \frac{m_4}{m_2^2} - 3\left[\frac{(n-1)^2}{(n-2)(n-3)}\right]$$

Skewed Distributions



Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005

Measures of Relationship

Data Science for Quality Management: Describing Data Numerically

with **Wendy Martin**

Learning objectives:

Discriminate between correlation & association
Calculate correlation for two variables

Measures of Relationship

Correlation and association are measures of the strength of a relationship between two variables.

Measures of Relationship

Before we calculate statistics related to relationship, we must first properly classify each variable.

- Nominal
- Ordinal
- Continuous

Correlation

- Where both variables are continuous, the statistic employed to measure the relationship may be referred to as a Coefficient of Correlation

Association

- Where both variables are nominal, the statistic employed to measure the relationship may be referred to as a Coefficient of Association

Correlation and Association

- Coefficients of Correlation and Association can vary given all possible combinations of nominal, ordinal, and continuous data that can occur

Coefficient of Correlation

- The most frequently used coefficient of correlation used is the Pearson Product-Moment Coefficient of Correlation.
- Symbols
  Population: ρ_{xy}
  Sample: r_{xy}

Coefficient of Correlation

- The most frequently used coefficients of correlation is the Pearson Product-Moment Coefficient of Correlation.
- Symbols
  Population: ρ_{xy}
  Sample: r_{xy}

Product Moment Coefficient

$$r_{xy} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$$

Product Moment Coefficient

Two components:

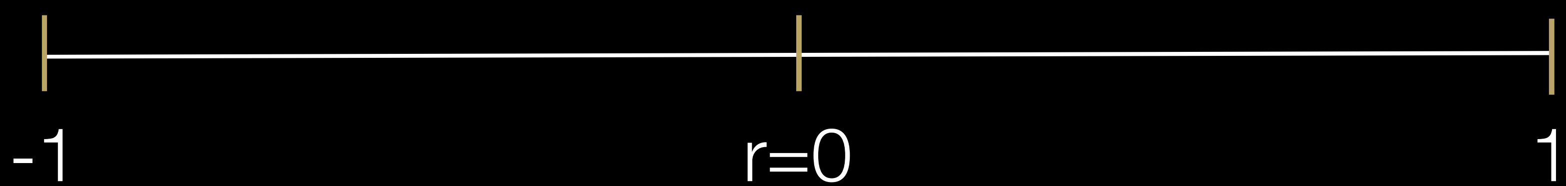
- Sign (+ or -)
- Numeric Value

Product Moment Coefficient

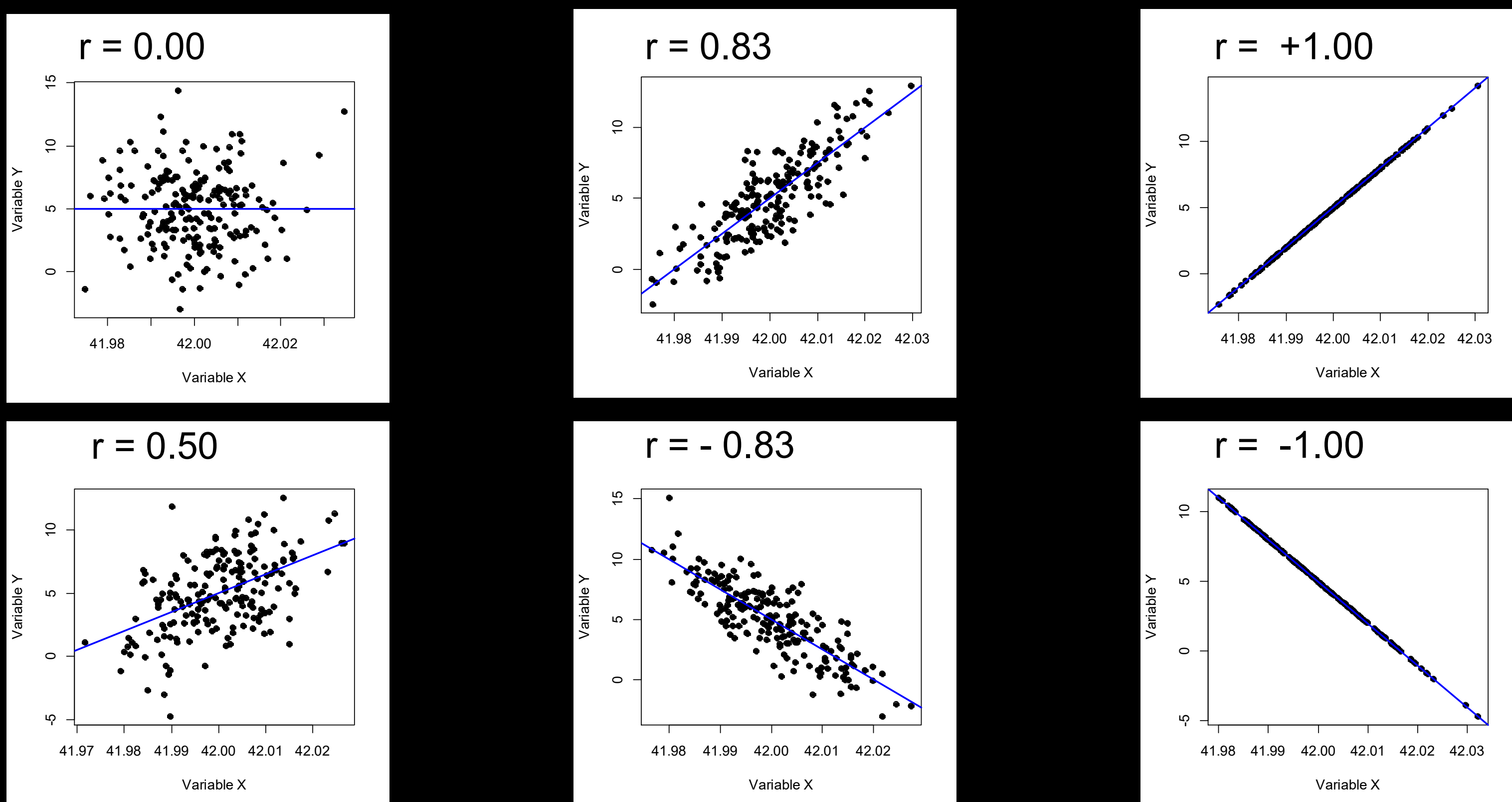
Sign (+ or -) gives the direction of the relationship

- Positive: As one variable increases in magnitude, the other variable increases
- Negative: As one variable increases in magnitude, the other variable decreases

Product Moment Coefficient



Scatterplot Examples



Sources

The material used in the PowerPoint presentations associated with this course was drawn from a number of sources. Specifically, much of the content included was adopted or adapted from the following previously-published material:

- Luftig, J. An Introduction to Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1982
- Luftig, J. Advanced Statistical Process Control & Capability. Luftig & Associates, Inc. Farmington Hills, MI, 1984.
- Luftig, J. A Quality Improvement Strategy for Critical Product and Process Characteristics. Luftig & Associates, Inc. Farmington Hills, MI, 1991
- Luftig, J. Guidelines for Reporting the Capability of Critical Product Characteristics. Anheuser-Busch Companies, St. Louis, MO. 1994
- Spooner-Jordan, V. Understanding Variation. Luftig & Warren International, Southfield, MI 1996
- Luftig, J. and Petrovich, M. Quality with Confidence in Manufacturing. SPSS, Inc. Chicago, IL 1997
- Littlejohn, R., Ouellette, S., & Petrovich, M. Black Belt Business Improvement Specialist Training, Luftig & Warren International, 2000
- Ouellette, S. Six Sigma Champion Training, ROI Alliance, LLC & Luftig & Warren, International, Southfield, MI 2005