Managing_Describing_Analyzing_Data_Week2 Run Charts

Me

2023-09-18

Common Methods of Graphically Describing Sample Data

- Run charts
- Frequency Distributions *examples*
- ungrouped
- grouped
- relative

other examples - histograms - frequency polygons - box and whisker plots

note A run chart is best used to graphically represent time ordered data.

Step 1

Create a data file

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

Step 1.5 Make df

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

Step 2 Create run chart

require(lolcat)

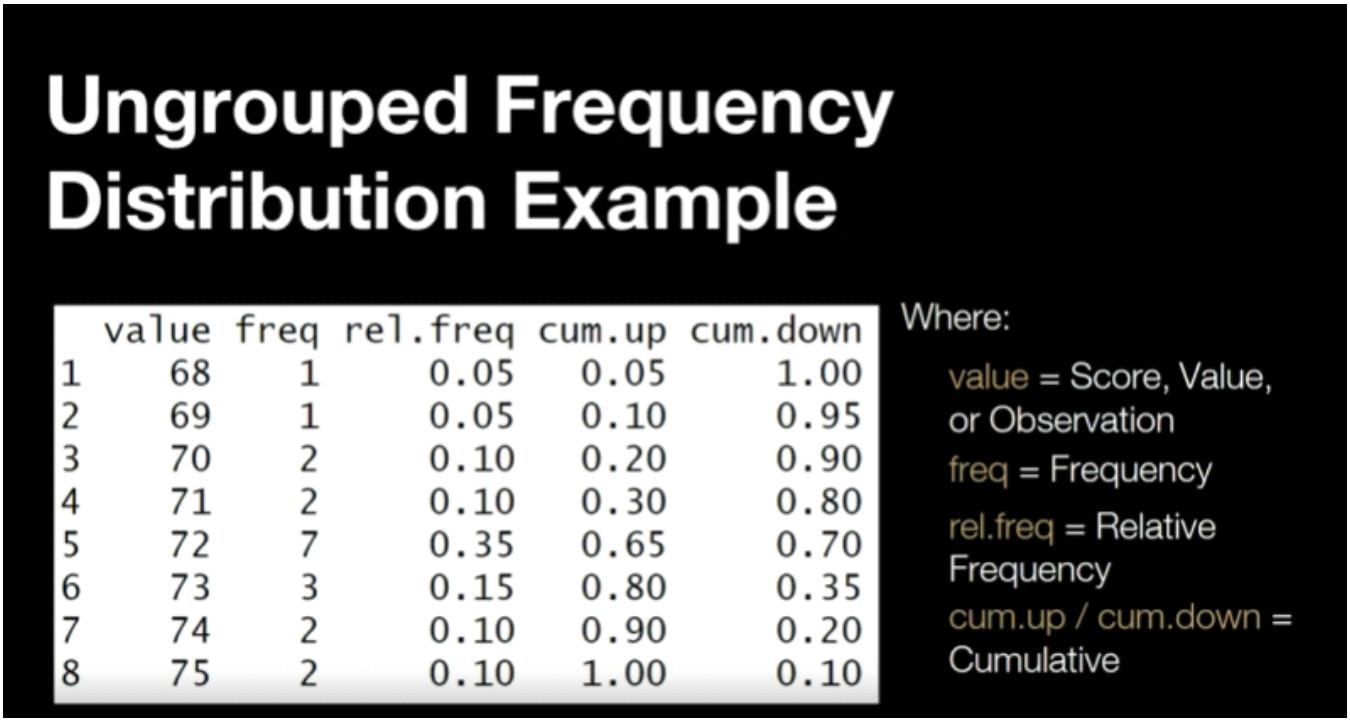
spc.run.chart(x=fanscfm)

Chaper 2 Frequency distributions

Ungrouped Use ungrouped when there are fewer than 20 data values

- Considered ungrouped when each row or class interval consists of only one score value or observation
- If the range of the data set is really large, ungrouped frequencies are too hard to use

Grouped When there are more than 20 data values. What class interval size is best? Rule of thumb is about 10 without going under. Or use 1,2,3,5 and then increase by intervals of 10.



Ungrouped frequency distribution

frequency.dist.ungrouped(fans$cfm)

Grouped frequency distribution Example

import data

run grouped frequency dist

frequency.dist.grouped(castings$weight)

Frequency polygon and Histogram *some examples of uses* - Evaluating manufacturing or business process. - Determining machine and process capabilities - Comparing vendor, material, process, operator, product characteristics

Both have identical rules Use grouped when there are more than 20 data values Use ungrouped when there are fewer than 20

a note about histograms When data is discrete values, such as counts, a histogram must be used. If data is continuous a frequency polygon or histogram may be used.

Frequency Polygon -A graph or chart that displays frequency of observations at each class interval (grouped) or value/score(ungrouped)

Similar to the frequency column of the frequency distribution. Often present a more representative when data are continuous

Becomes more smooth as sample size increases

Ungrouped Frequency polygon and histogram

frequency.polygon.ungrouped(fans$cfm)

hist.ungrouped(fans$cfm)

Grouped frequency polygon

frequency.polygon.grouped(castings$weight)

hist.grouped(castings$weight)

Histogram Patterns and Density Plots Center spread and shape can give clues to what the data is trying to convey.

Pattern 1 is very symmetric and bell shaped.Most likely number is going to be the mean. Or average.

Pattern 2 could be finance data

Pattern 3 could be the result of a natural limit, or sorting after collection.

Pattern 4 Result of a process trying to conform to specification

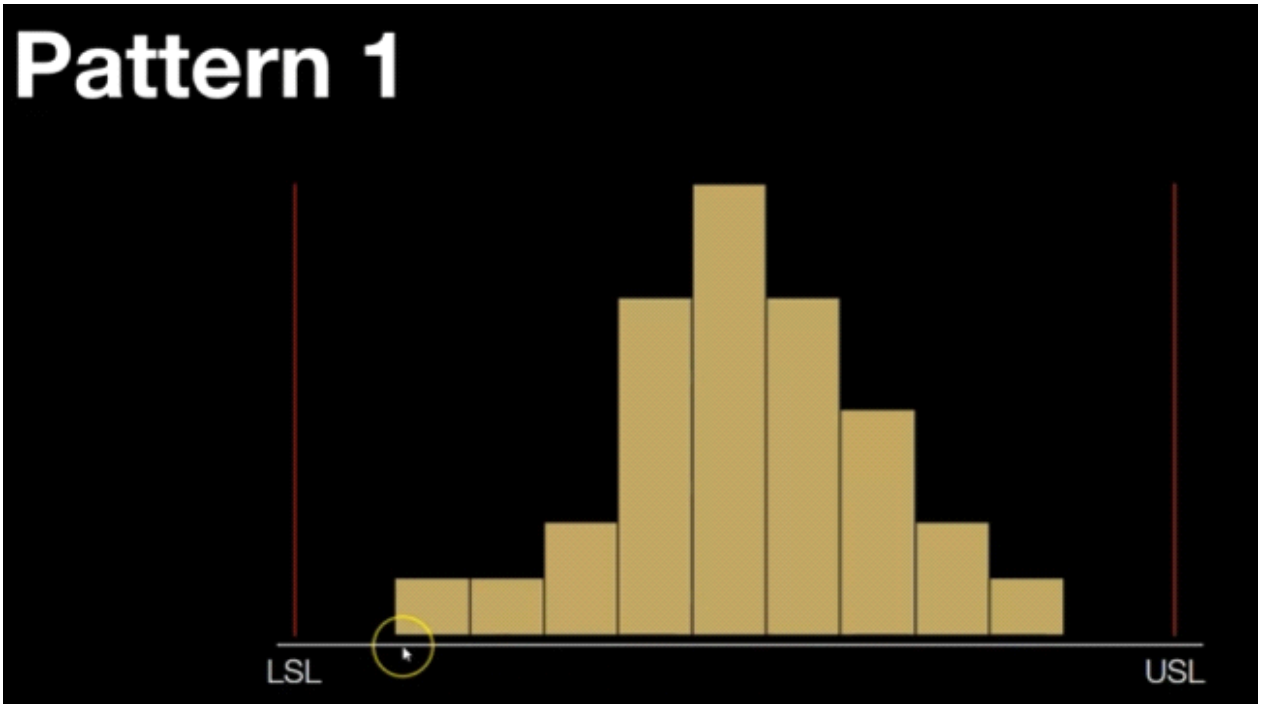
Pattern 5 might be the result of not wanting to have a process go out of specification? (I kind of don't get that one that much)

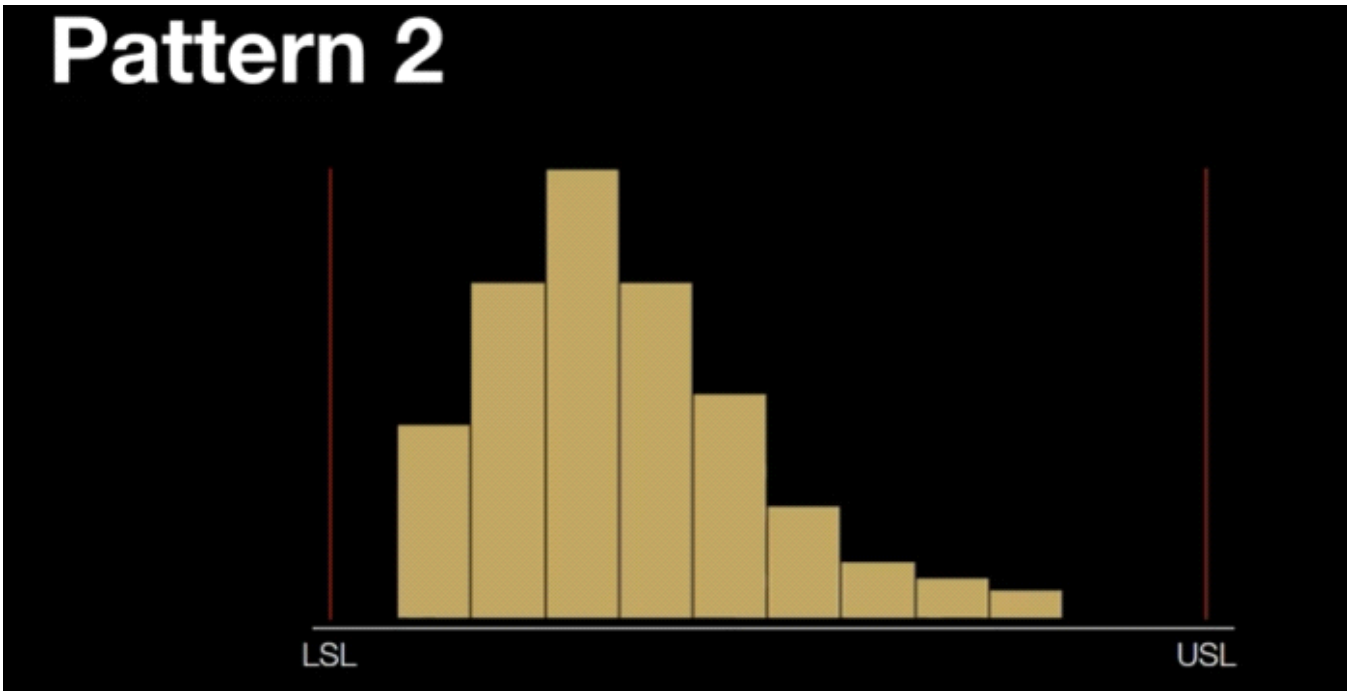
Pattern 6 could the the result of interpolation or lack of resolution with an instrument.

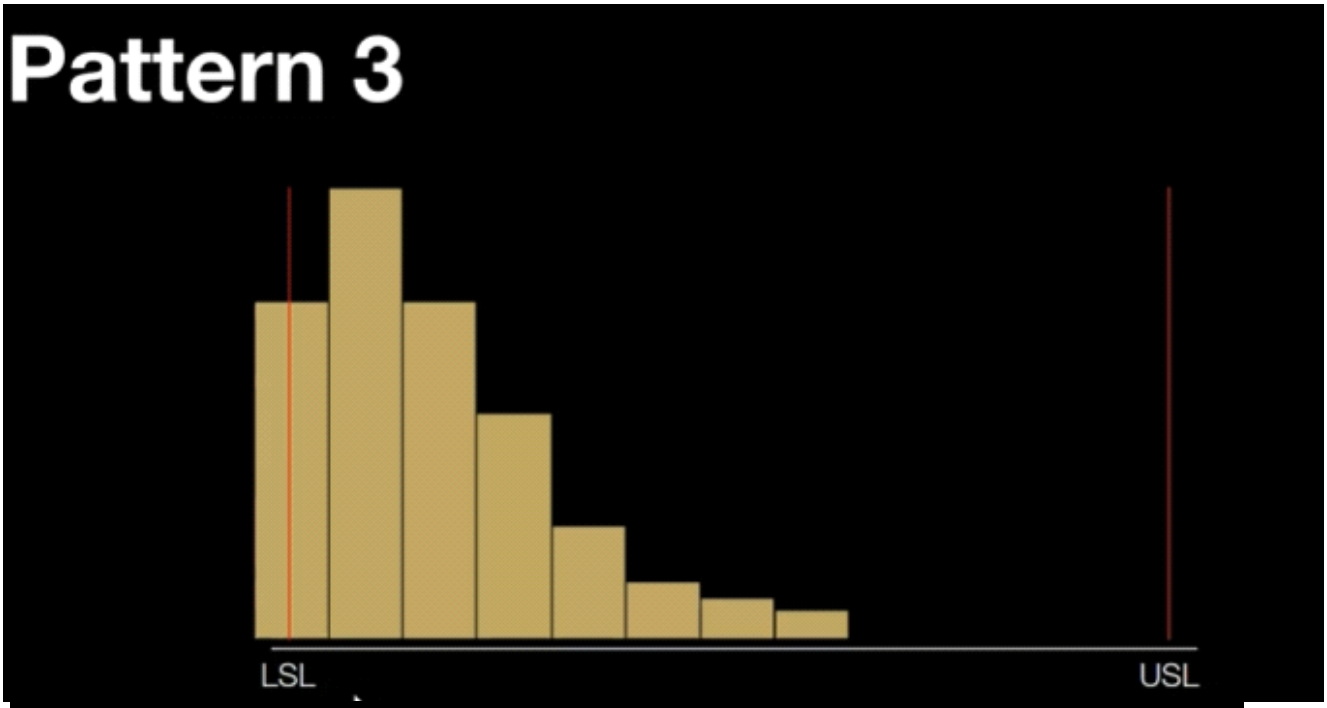
Pattern 7 could be the result of either there being no measurements in the center or some manufacturer sorting out center values.

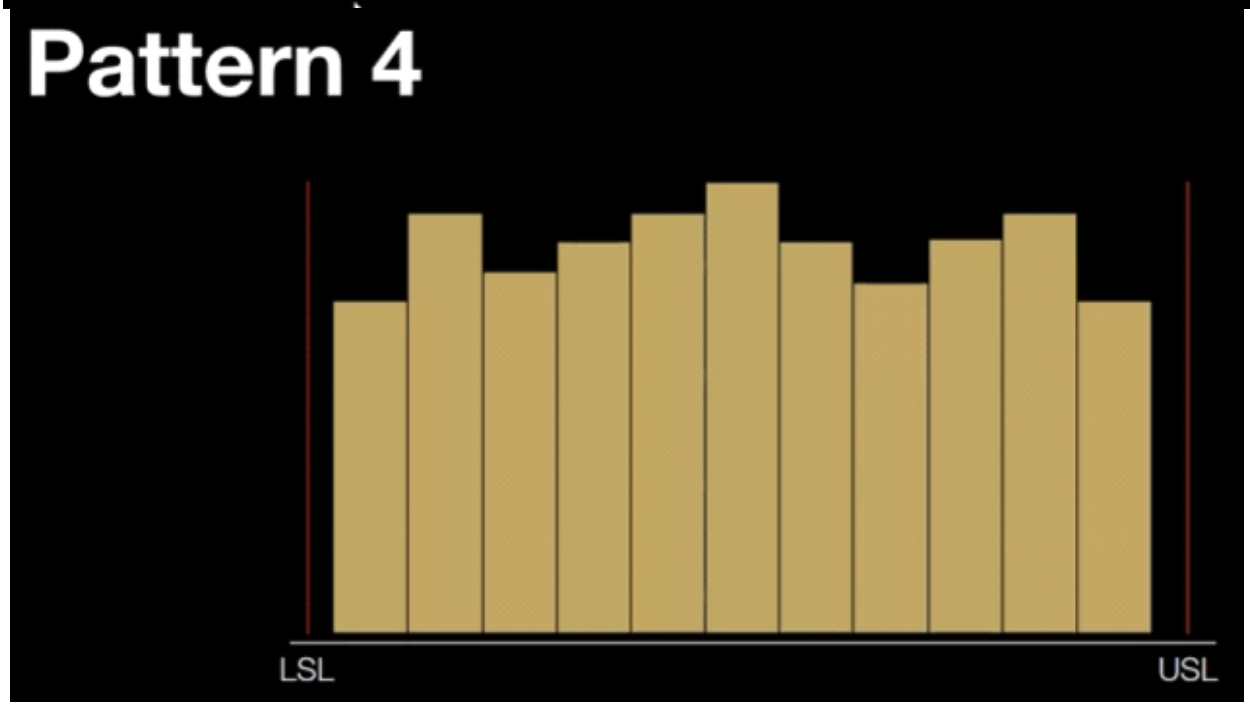
Pattern 8 Is the result of two separate process streams at work.

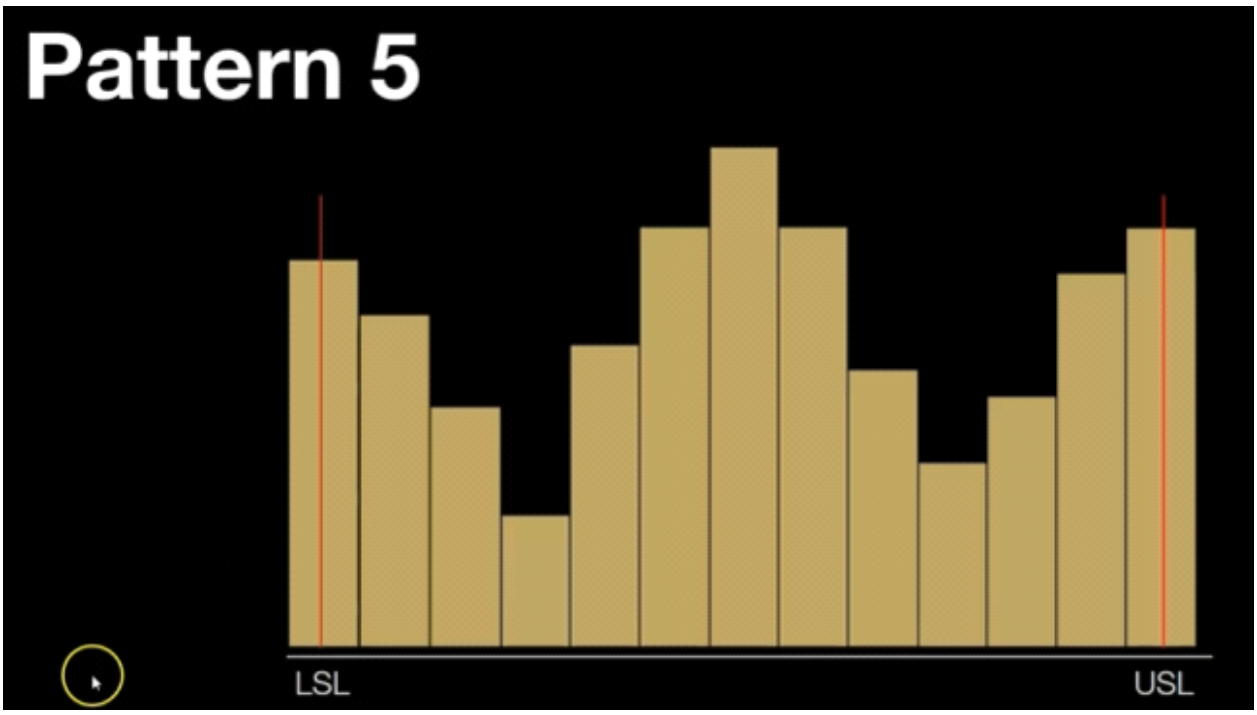
Pattern 9 Shows a possible out of control situation.

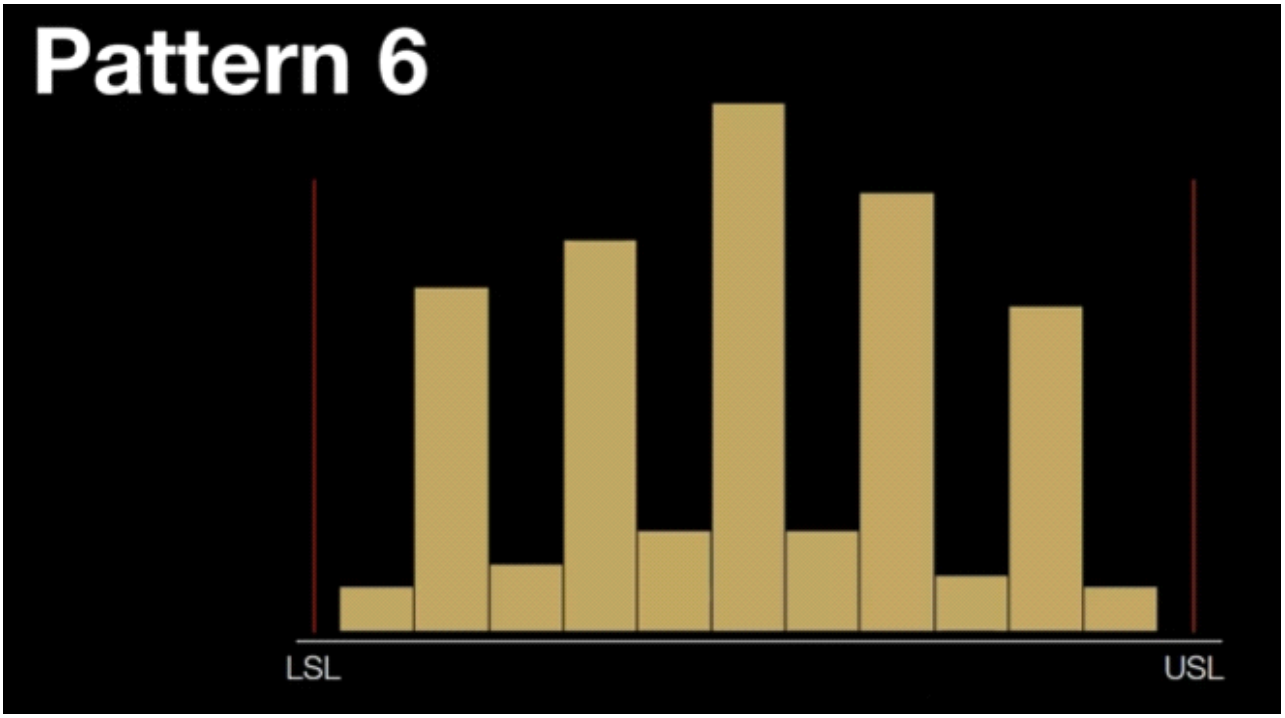

Pattern 1
LSL
USL

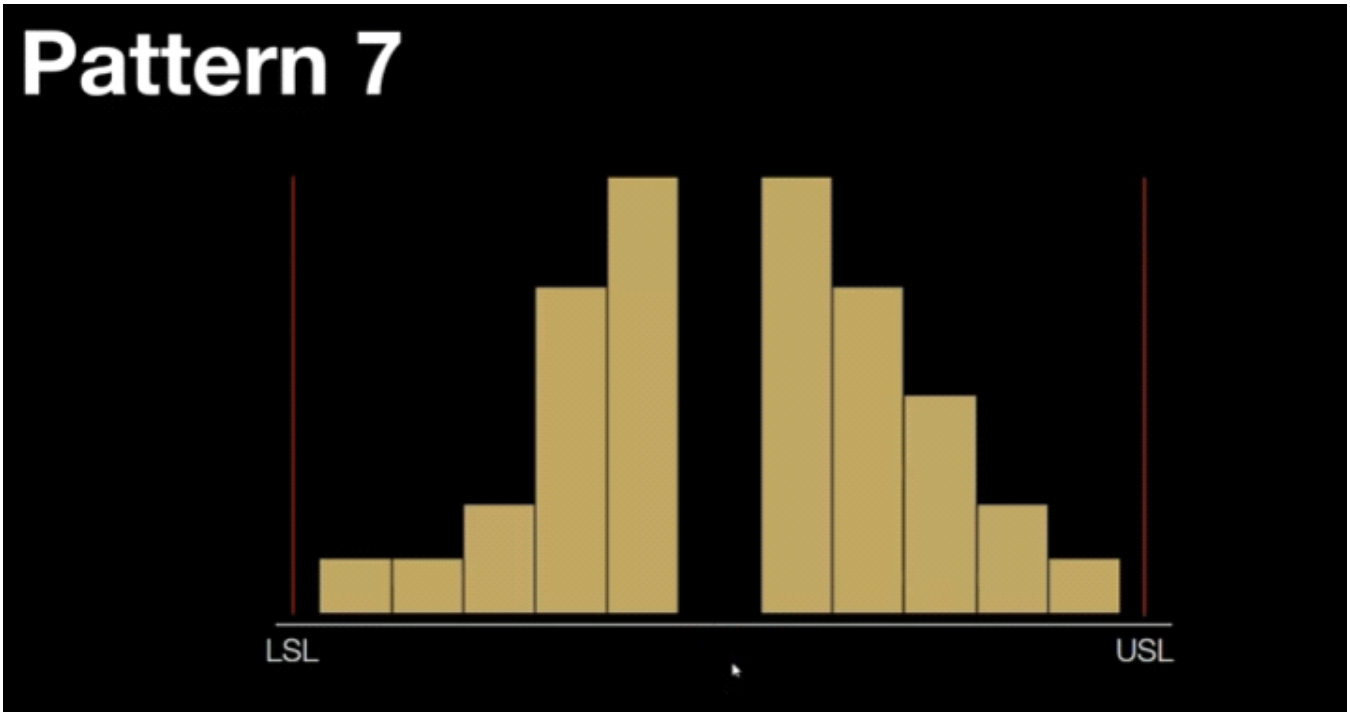

Pattern 2
LSL
USL

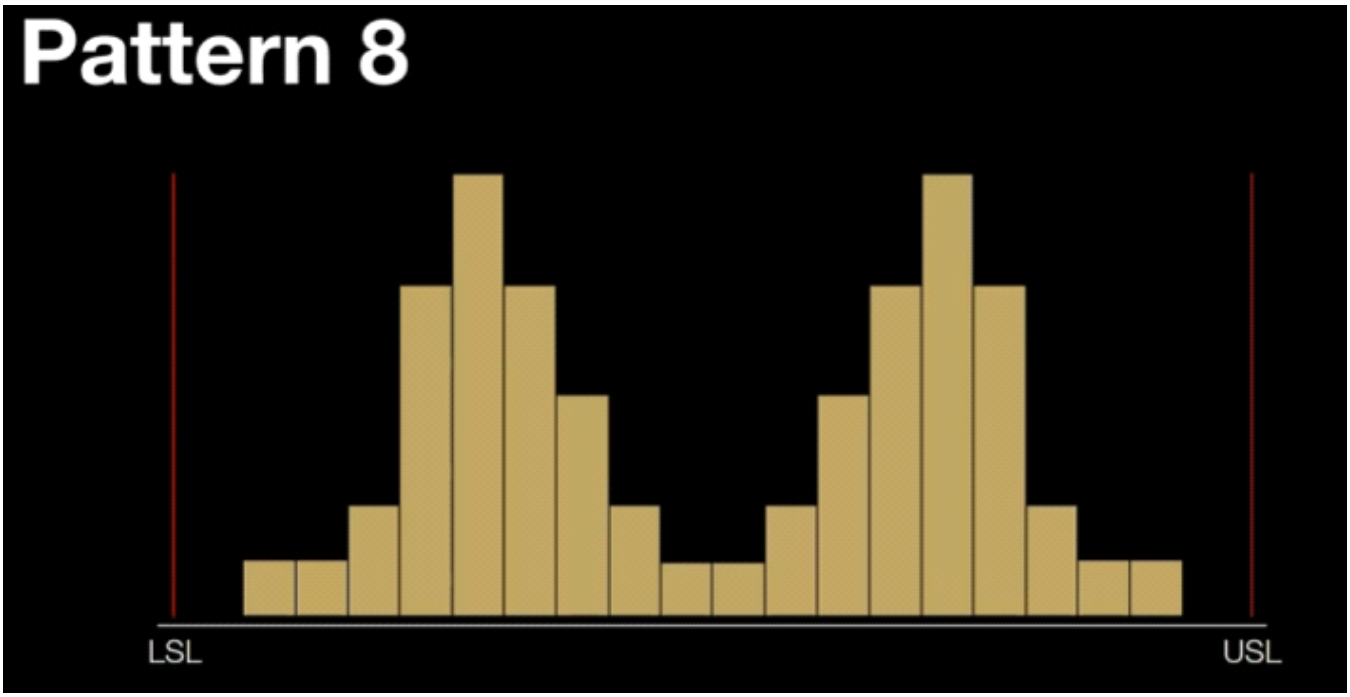

Pattern 3
LSL
USL

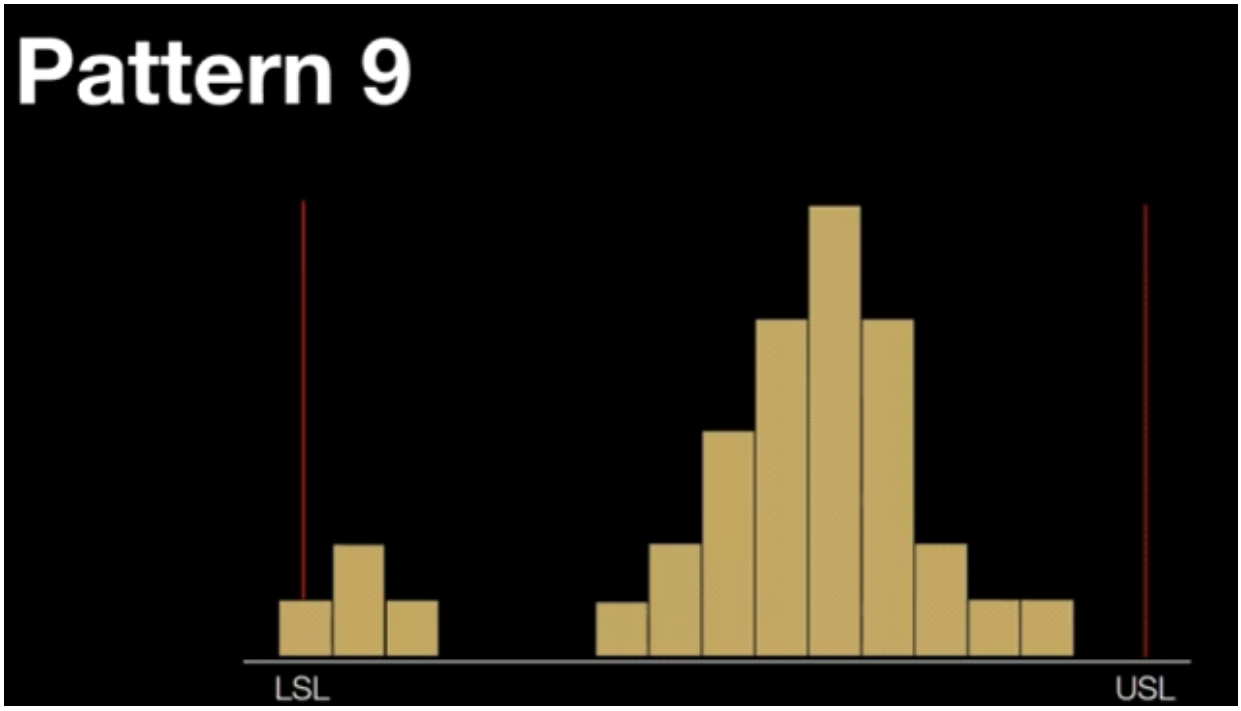

Pattern 4
LSL
USL


Pattern 5
LSL
USL


Pattern 6
LSL
USL


Pattern 7
LSL
USL


Pattern 8
LSL
USL



Grouped Histogram Now with a density plot Density plots are used when we have continuous data and need to visualize underlying probability distribution.

To plot the density data over the histogram hist.grouped(castings$weight,freq=F)

lines(density(castings$weight))

To plot just the density plot

plot(density(castings$weight))

Box and whisker plots

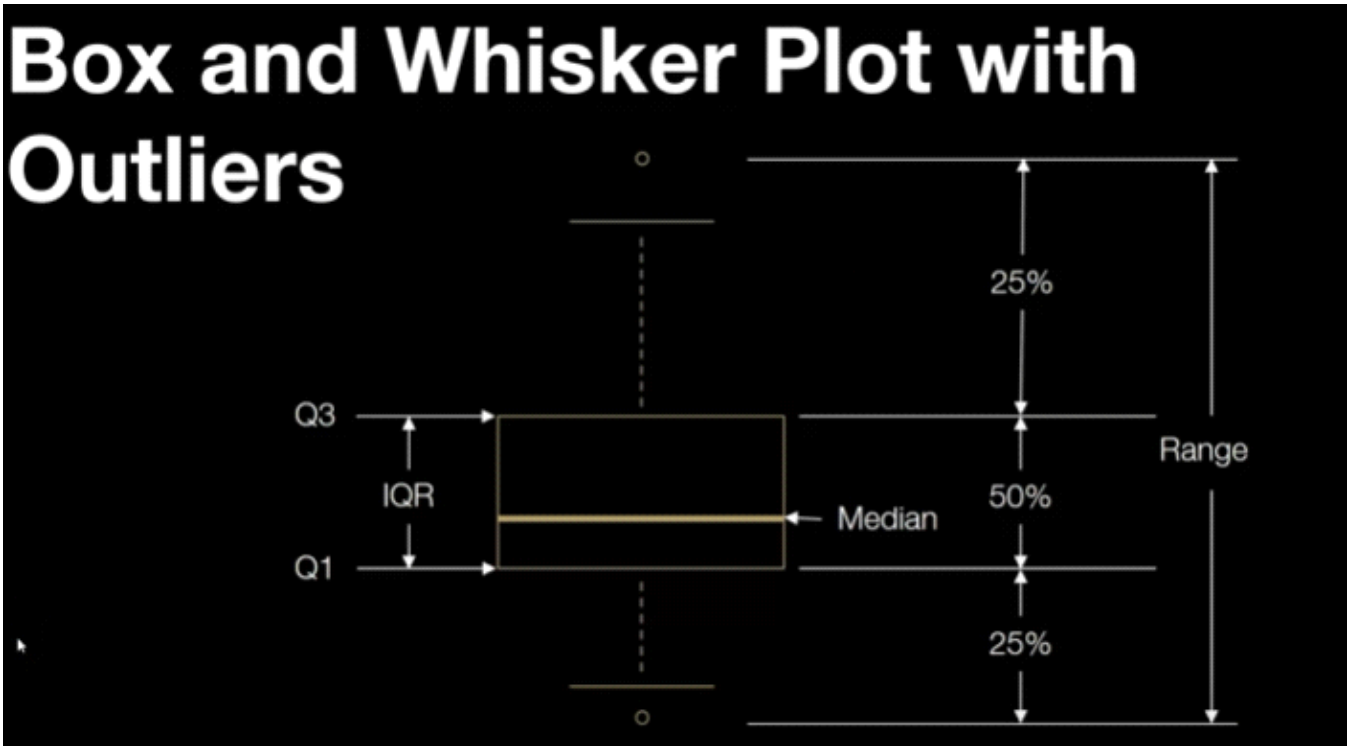
Use the summary command to get the 5 number summary Start with that

summary(castings$weight)

Consists of a 5 number summary

-Maximum - Q3 25% of data lie above Q3 - Median (Q2)- Known as interquartile range, where 50% of the data lie. Distance between Q3 and Q1 - Q1- Data value at which 25% lowest values fall - Minimum- Lowest value To determine if a value is an outlier- multiply the distance between Q3 and Q1 by 1.5, and we add it to Q3 to produce the 'inner fence'.

If you add 3X box length to Q3 value or subtract that same value from Q1 any value found outside that value is considered a wild outlier.



box and whisker plot summary

Box plot now

boxplot(castings$weight)

To show a 95% confint of the median add notch = T to the end

boxplot(castings$weight,notch=T)

boxplot to compare data

boxplot(y~x, data= dataframe)

Measures of central tendency and position Note! A grouped mean and weighted mean are different words for the same thing

Step 1) Create a vector

weight <-c(65,67,36,37,36,57,53,39,38,58) preform <-data.frame(weight)

Step 2) look at measures of central tendency

mean is the mean, sum of all deviations must equal 0.

Note the calculated value for the ungrouped mean uses all of the data points and can be affected by extreme values.

Mean: Calculations

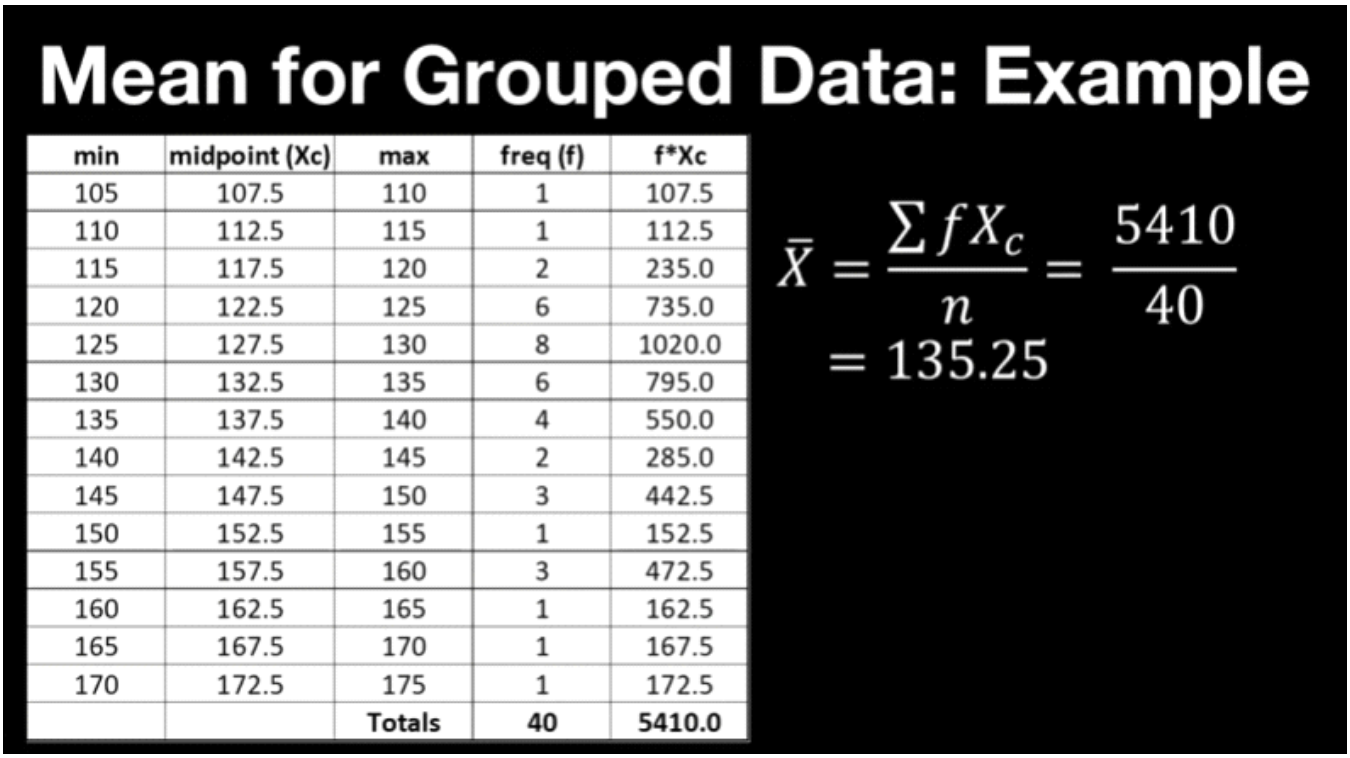
- Ungrouped Data: $\bar{X} = \frac{\sum X}{n}$
- Grouped Data: $\bar{X} = \frac{\sum f X_c}{n}$
- Weighted Mean: $\bar{X} = \frac{\sum w_j X}{w_j n_j}$

Mean 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



1) Assign to local variable

fdcast <- frequency.dist.grouped(castings$weight)

- Look over relevant data

str(fdcast)

- Create a vector of midpoints and frequencies midpts <-freqfreq
- use the weighted mean command

weighted.mean(x=midpts,w = freq)

For the median Median is exact middle of data. And a measure of position. *advantage* Easy to understand Not affected by extremes

disadvantages Does not take magnitude of values into account.

Another note: Make sure to order the data

For the mode

Most frequently occurring value For a population the mode is the peak of the distribution curve

For percentiles and quantiles

Pth percentile is the value that P% of values fall at or below, and 100-P% fall above it.

use the quantile function(x=value, probs=probability)

quantile(x=preform$weight, prob=0.30)

Quartile

Postions at specific percentiles, 25,50,75,100

example find first and 3rd quartile

quantile(x=preform$weight,probs=c(0.25,0.75))

Measures of Dispersion

Range difference between highest and lowest value in each data set- only depends on max and min, but very sensitive to outliers

range(preform$weight)

Find the IQR

IQR(preform$weight)

Measures of Shape

skewness describes departure from symmetry on the x axis, kurtosis describes the peakedness on the y-axis.

skeweness = gamma3

symmetric distributions have 0 skeweness Most important calculations use third and fourth moments about the mean, moments are the average of the deviations from the mean raised to some power

to calculate in r studio (skewness) skewness(castings$weight)

kurtosis = g4

degree of peakedness no kurtosis is mesokurtic negative is platykurtic

calculate kurtosis kurtosis(castings$weight)

you can also use summary function

summary.continuous(castings$weight, stat.sd=T)

Measures of relationship First are the data nominal, ordinal or continuous

coefficient of correlation- continuous data coefficient of association- discrete data

example relationship between type of metal used and whether the part is defective or not defective

correlations, 1 is perfect -1 is also perfect anything else is not. Correlation does not mean association.

- make sure data is in correct format (dependent or independent) castnew <- transform.independent.format.to.dependent.format(fx=weight~mold, data = castings3)
- Calculate pearson product moment correlation

cor(x=castnewcell.2, method='pearson')

- create a scatterplot and then abline an lm over it plot(castnewcell.2) abline(lm(castnewcell.2))