

# BUS 232 Looking at Data -- Distributions

## Summary

## Terminology

### Distributions

- Values taken by a variable and how often it takes them

### Individuals

- described by variables

### Data Types

- Categorical - groupings. Can be numbers but usually not.
- Quantitative - we can do arithmetic. Numbers

### Pareto Chart

- Sorted bar chart

### center

- about half values above, half below

### spread

- look for values outside the overall pattern (outliers)

### Symmetric

- right and left sides are about mirrored

### skewed

- skewed to left -> left side longer from normal
- skewed to right -> right side longer from normal

### pth percentile

- p percent of distribution falls below it.
- 90 percentile of the class == top 10%

### quartiles

- first quartile = 25th percentile
- median = 50th percentile
- interquartile range: third quartile - first quartile

## Looking at Data -- Distributions

### To better understand a data set

1. Who?
2. What?
3. Why?

## Displaying categorical data

### Purpose

- summarize data to quickly get characteristics

### Process

- list categories

### Methods

- pie charts, bar graphs

## Steps of making a histogram

*In our example:* max = 8.9, min = 1.5

1. **Estimate the value range** Assume max is 9, assume min is 0.0. Estimated Value Range =  $9.0 - 0.0 = 9.0$   
2.0
2. **Dividing the range into N classes** Let  $W$  = the width of each class.  $W = R/N$ . eg) let  $N = 9$ .  $W = 9.0/9 = 1$ ; each class (9) has a width of 1.
3. Counting the frequencies. Make a chart with class, frequencies, and % frequencies as the columns.

## Stem plot

- split by a magnitude of 10
- stem is to the left of vertical line. if extend the graph, add more of the same stems
- leaf is to the right of vertical line
- have to indicate leaf unit (leaf unit = 1)
- can round the leaf (trimming the dataset)

## Describing distributions with numbers

### *central values*

- mean - the average of all values - more effected by outliers
- median - the midpoint of a distribution - not very effected by outliers
  - arrange from smallest to largest
  - if odd, it's the middle
  - if even it's the mean of the middle two numbers

### *measures of spread*

- range: largest-smallest
- interquartile range: third quartile - first quartile
  - third quartile get new median above median
  - first quartile get new median below median

- less useful for small datasets
- **boxplot** -- get the five number summary -- clearly show skewness  
*measures of shape*

Coming up next... variance and standard deviation!