Data Analysis

Idaho State University | Computer Science

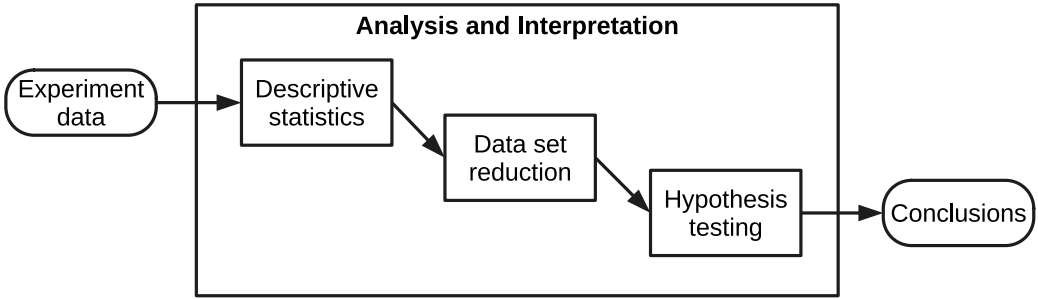
Isaac Griffith

CS 6620
Department of Informatics and Computer Science
Idaho State University

ROAR

Idaho State University
Computer Science

Analysis & Interpretation



ROAR

Descriptive Statistics

- Presentation and numerical processing of a data set
- Describe and graphically present interesting aspects of the data set
- **Goal**: get a feeling of how the data set is distributed
- May be used before hypothesis testing to understand nature of data
  - identify outliers
  - remove outliers
- Measurement scale restricts meaningful computations

Relevant Statistics

Scale Type	Measure of Central Tendency	Dispersion	Dependency
Nominal	Mode	Frequency	
Ordinal	Median, percentile	Interval of variation	Spearman corr. coeff., Kendall corr. coeff.
Interval	Mean, variance, range	Standard deviation	Pearson corr. coeff.
Ratio	Geometric mean	Coefficient of variation	

Descriptive Statistics

- Descriptive statistics tell us
  - **Measures of Central Tendency** - provide an estimate, or expectation, of the "center" of a distribution of values.
  - **Distribution** - provide a summary of the frequency of individual values or ranges of values for a variable
  - **Dispersion** - provide the spread of values around the central tendency
- These, along with some basic graphical analysis, allow us to reach some simple conclusions about the data.

Idaho State University
Computer Science

Measures of Central Tendency

Arithmetic Mean

- Calculated as: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}$
- **Ex**: (1, 1, 2, 4)
  - $\bar{x} = 2$

Median

- Calculation for $\tilde{x}$
  1. Sort the data
  2. If size of data is odd, select middle value
  3. Else, select middle two values and average
- **Ex**: (1, 4, 2, 1)
  - Sorted: (1, 1, 2, 4)
  - $\tilde{x} = 1.5$
- Also called the 50%-percentile, $x_{50\%}$

ROAR

Idaho State University
Computer Science

Measures of Central Tendency

Mode

- Represents the most commonly occurring sample
- Calculation:
  - count the number of samples for each unique value
  - select the value with the highest count
- Meaningful for nominal, ordinal, interval and ratio scales.

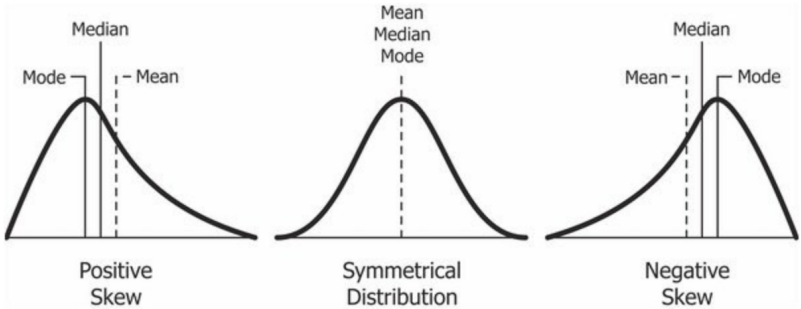
Geometric Mean

- Calculated: $\sqrt[n]{\prod_{i=1}^{n} x_i}$
- **Ex**: (1, 1, 2, 4), n = 4
  - Geometric Mean = 1.681792831
- Well defined if all samples are non-negative and meaningful for the ratio scale

ROAR

Data Distribution

- If the data is symmetrically distributed
  - Arithmetic Mean and median are equal
- If the data is symmetrically distributed and has one unique maximum
  - Arithmetic Mean, Media, and Geometric Mean are equal
- If the data is asymmetric
  - The values will vary

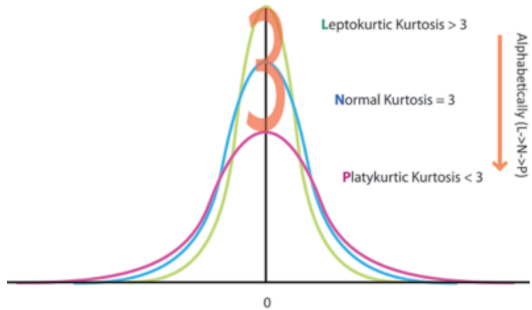


Idaho State University
Computer Science

Skewness and Kurtosis

- **Skewness** - measure of distortion from a symmetrical bell curve
  - **Positive** - tail on the right side is longer or fatter (median is greater than mode)
  - **Negative** - tail on the left side is longer or fatter (media is less than mode)

- **Kurtosis** - describes the extreme values in one versus the other tail (measure of outliers present)
  - **High Kurtosis** - heavy tails or outliers
  - **Low Kurtosis** - light tails or lack of outliers



ROAR

Idaho State University
Computer Science

Measures of Dispersion

Variance

- The average of the **squared** differences from the Mean.
  1. Finding the mean
  2. For each number in the data set: subtract the Mean and square the result
  3. Find the average of those squared differences
- Calculation: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **Ex**: (1, 1, 2, 4)
  - Mean: 2
  - Squared Diffs: (1, 1, 0, 4)
  - $s^2 = 1.5$

Standard Deviation

- Preferred over the variance as it retains the same dimension as the data itself
- Calculation: $s = \sqrt{s^2}$
- **Ex**: (1, 1, 2, 4)
  - $s = 1.224744871$

ROAR

Idaho State University
Computer Science

Measures of Dispersion

Range

- The difference between the max and min of a set of data.
- Calculation: $range = x_{max} - x_{min}$
- **Ex**: (1, 1, 2, 4).
  - The range is 3

Variation Interval

- Represented as the pair of the min and max of a set of data
- $(x_{min},\ x_{max})$
- **Ex**: (1, 1, 2, 4)
  - $(1,\ 4)$

Coefficient of Variation

- dispersion expressed in percentage of the mean
- Calculation: $100\frac{s}{x}$
- Has no dimension and is of the ratio scale

ROAR

Frequency Table

- **Frequency** of each data value provides a general view of dispersion
- A frequency table is constructed by
  - tabulating each unique value and the count of occurrence for each value
- **Relative Frequency** is calculated by dividing each frequency by the total number of samples
- **Ex**: (1, 1, 1, 2, 2, 3, 4, 4, 4, 5, 6, 6)

Value	Frequency	Relative Frequency
1	3	23%
2	2	15%
3	1	8%
4	3	23%
5	1	8%
6	2	15%
7	1	8%

Measures of Dependency

- When data consists of related pairs (x_i, y_i) from stochastic variables X and Y
- X and Y are relatable through some function $y = f(x)$
  - If we can state the function as, $y = \alpha + \beta x$
  - Then we can estimate using linear regression
- Regression fits the data to a curve

Sums of Squares

- The following sums are useful in computing the regression line: $y = \bar{y} + \beta(x - \bar{x})$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_{i=1}^{n} x_i y_i \right) - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right) \left(\sum_{i=1}^{n} y_i \right)$$

Regression Coefficients

- The slope of the line, β, is computed as:

$$\beta = \frac{S_{xy}}{S_{xx}}$$

- While the y-intercept, α, is computed as:

$$\alpha = \bar{y} - \beta\bar{x}$$

Idaho State University
Computer Science

Measures of Dependency

Covariance

- Provides a single number to quantify how much two data sets vary together
- Calculation: $c_{xy} = \frac{S_{xy}}{n-1}$
- meaningful for interval and ratio scales
- can be normalized using standard deviation of x_i and y_i
  - leads to Pearson's correlation

Pearson's Correlation, r

- Calculation: $r = \frac{c_{xy}}{s_x s_y} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$
- Value is between -1 and 1
  - $r = 0$ is a strong indicator of no relationship
  - $r \geq 0.7$ is a strong indicator of a positive relationship
  - $r \leq -0.7$ is a strong indicator of a negative relationship
- Assumptions:
  - Relationship is linear
  - Data is interval or ratio scale
  - Data is normally distributed

ROAR

Measures of Dependency

Spearman's ρ

- Non-parametric version of Pearson's Correlation
  - Used when Pearson assumptions are markedly violated
- Determines strength and direction of relationship
- Value between -1 and 1
- Calculated the same as Pearson's
  - Data is actually ranks of the original data
  - Data is sorted and numerically ranked
- Assumptions
  - Data is at least ordinal scale
  - Relationship is monotonic

Measures of Dependency

Kendall's Rank-Order Correlation, τ

- Another non-parametric correlation
- Counts agreements and disagreements in ranks of data
- Value between -1 and 1 indicating strength of disagreement/agreement
- Assumptions:
  - Data is at least ordinal scale

Idaho State University
Computer Science

Beyond Two Variables

- Multivariate Analysis
  - **MATH 4459/5559 Applied Multivariate Analysis**
    - PREREQ: MATH 2240 and MATH 3350
  - Multiple regression
  - PCA
  - Cluster Analysis
  - Discriminant Analysis

ROAR

Graphical Visualization

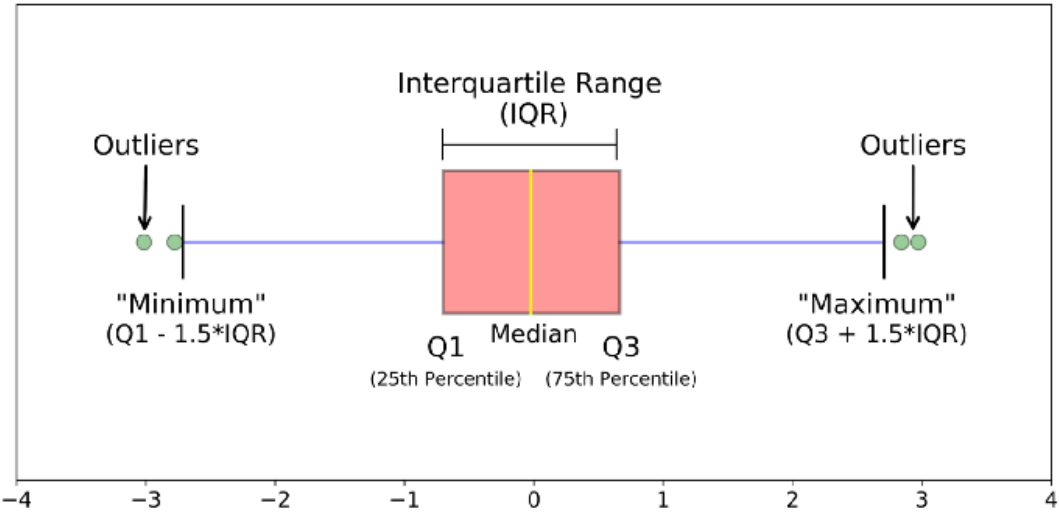
- Goal is to pair quantitative measures with visualizations to better understand the data.
- Basic Visualization Techniques
  - Scatterplots - shows dependencies between variables
  - Boxplots - shows dispersion and skewdness of samples
  - Histograms - shows an overview of the distribution density

Idaho State University
Computer Science

Scatterplots

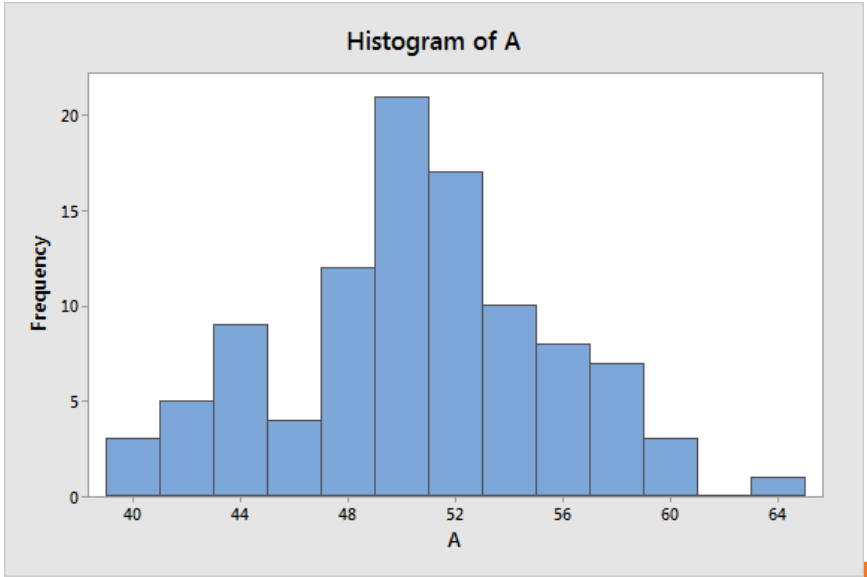
Idaho State University
Computer Science

Boxplots



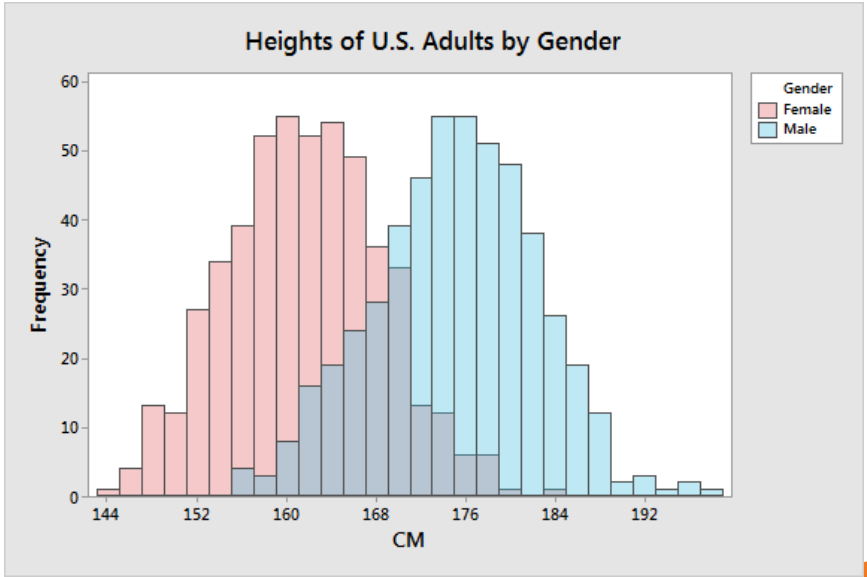
Idaho State University
Computer Science

Histograms

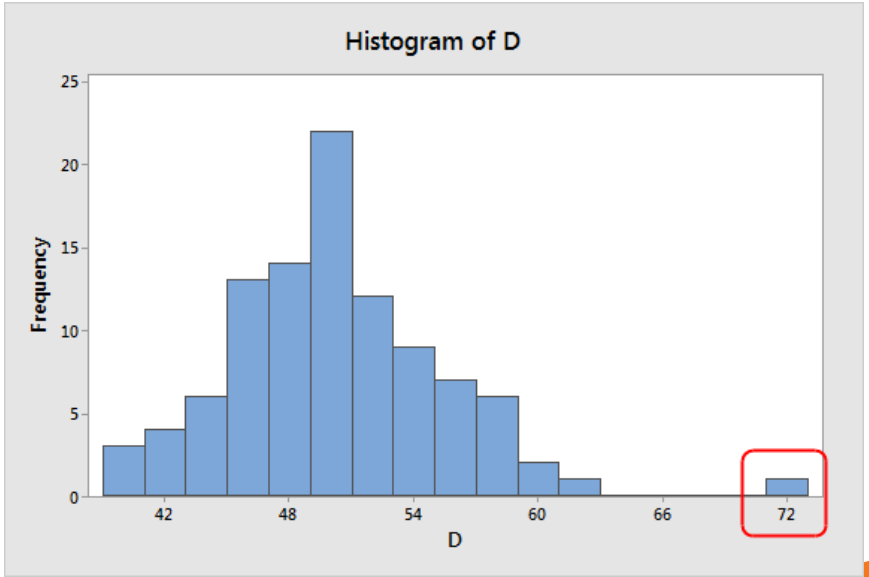


ROAR

Histograms

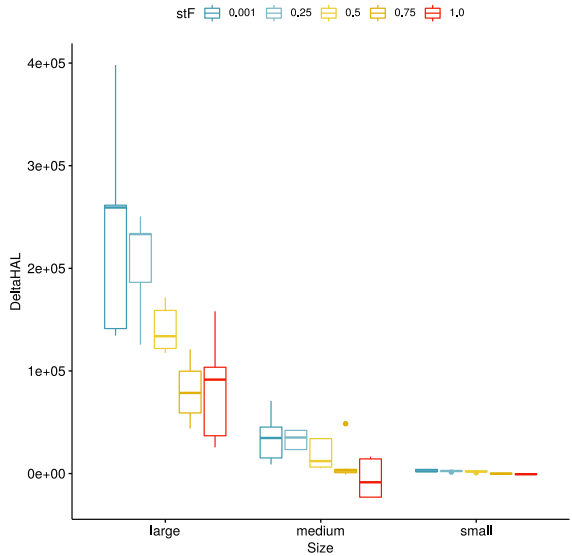


Histograms



Idaho State University
Computer Science

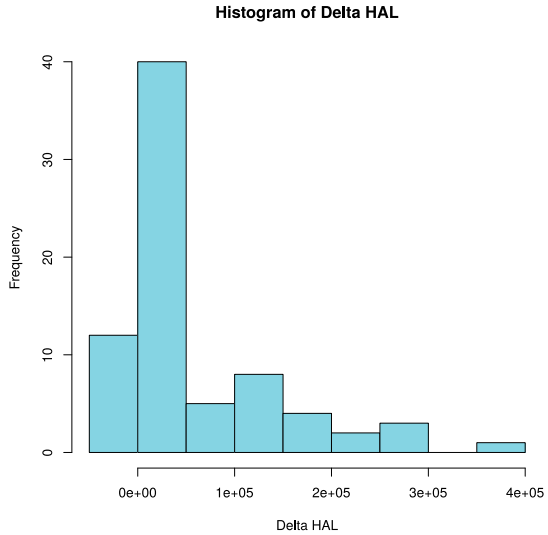
Examples



ROAR

Idaho State University
Computer Science

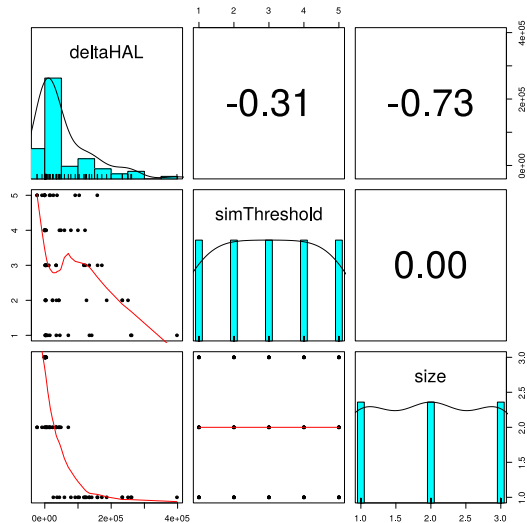
Examples



ROAR

Idaho State University
Computer Science

Examples



ROAR

Data Set Reduction

- Garbage in = Garbage out
  - Quality results require quality input data
- Errors
  - Systematic
  - Outliers
- Outlier Detection
  - Scatterplots, Histograms, and other graphical techniques
  - Computational techniques
- Once found you must decide to remove or not remove them
- Value of data
  - PCA

Are there any questions?