MIT-WPU

॥ विश्वशान्तिर्ध्रुवं ध्रुवा ॥

Dr. Vishwanath Karad

MIT WORLD PEACE UNIVERSITY | PUNE

TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

UNIT - II

Statistics for Data Science

Data Science

T. Y. BTECH

SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

Contents

- Steps to summarizing Data
- Data Classification
- Types of Data
- What are Descriptive statistics
- Frequency tables and graphs, Histograms
- Central Tendency
- Mean, Median, Mode
- Dispersion
- Range, variance, standard deviation
- Quartiles, Percentiles
- Box Plots
- Bivariate Descriptive Statistics
  - Contingency Tables
  - Correlation
  - Regression

Three Steps to Summarize Data

1. Classify sample into different type

2. Depending on the type, use appropriate numerical summaries

3. Depending on the type, use appropriate visual summaries

Data Classification

- Data/Sample: $(X_1, \ldots, X_n)$

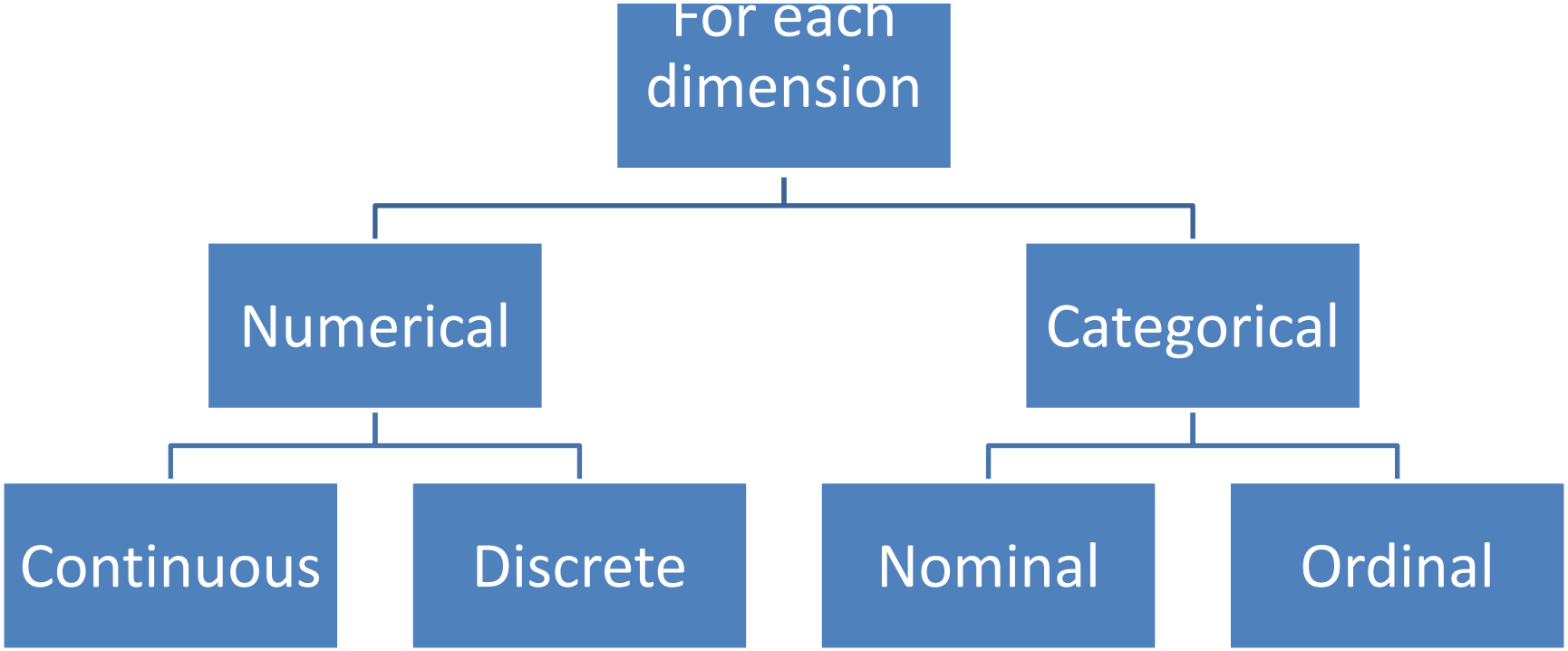
- Dimension of X_i (i.e. the number of measurements per unit i)
  - Univariate: one measurement for unit i (height)
  - Multivariate: multiple measurements for unit i (height, weight, sex)

- For each dimension, X_i can be numerical or categorical

- Numerical variables
  - Discrete: human population, natural numbers, (0,5,10,15,20,25,etc..)
  - Continuous: height, weight

- Categorical variables
  - Nominal: categories have no ordering (sex: male/female)
  - Ordinal: categories are ordered (grade: A/B/C/D/F, rating: high/low)

Data Types


For each dimension
Numerical
Categorical
Continuous
Discrete
Nominal
Ordinal

Primar y & Secondar y Data

- **Raw or Primary data:** when data collected having lot of unnecessar y, irrelevant & un wanted information

- **Treated or Secondary data:** when we treat & remove this unnecessar y, irrelevant & un wanted information

- **Cooked data:** when data collected not genuinely and is false and fictitious

Ungrouped & Grouped Data

Ungrouped data: when data presented or observed individually. For example if we observed no. of children in 6 families

2, 4, 6, 4, 6, 4

Grouped data: when we grouped the identical data by frequency. For example above data of children in 6 families can be grouped as:

No. of children	Families
2	1
4	3
6	2

or alternatively we can make classes:

No. of children	Frequency
2 - 4	4
5 - 7	2

Variable

A variable is something that can be changed, such as a characteristic or value. For example age, height, weight, blood pressure etc

Types of Variable

Independent variable: is typically the variable representing the value being manipulated or changed. For example smoking

Dependent variable: is the observed result of the independent variable being manipulated. For example ca of lung

Confounding variable: is associated with both exposure and disease. For example age is factor for many events

Categories of DATA

Quantitative or Numerical data

This data is used to describe a type of information that can be counted or expressed **numerically (numbers)**

2, 4 , 6, 8.5, 10.5

Quantitative or Numerical data (cont.)

This data is of **two** types

1. **Discrete Data:** it is in **whole numbers** or values and has **no fraction**. For example

Number of children in a family = 4

Number of patients in hospital = 320

2. **Continuous Data (Infinite Number):** measured on a **continuous scale**. It can be **in fraction**. For example

Height of a person = 5 feet 6 inches 5".6'

Temperature = 92.3 °F

Qualitative or Categorical data

This is **non numerical** data as

Male/Female, Short/Tall

This is of **two** types

1. **Nominal Data:** it has series of **unordered categories**

(one can not √ more than one at a time) For example

Sex = Male/Female Blood group = O/A/B/AB

2. **Ordinal or Ranked Data:** that has distinct **ordered/ranked categories.**
For example

Measurement of height can be = Short / Medium / Tall

Degree of pain can be = None / Mild /Moderate / Severe

Why Descriptive statistics?

- Who is a better ODI batsmen - Sachin or Muralidharan?
  - Batting average?
- Who is the reliable- Dhoni or Afridi?
  - Score variance
- A triangular series among Aus, Eng & Newziland ; Who will win?
  - Most number of wins - Mode
- I am going to buy shoes. Which brand has verity- Power or Adidas?
  - Price range - Range
- We used Average, Variance, Mode, Range to make some inferences. These are nothing but descriptive statistics
- Descriptive statistics tell us what happened in the past.
- Descriptive statistics avoid inferences but, they help us to get a feel of the data.
- Some times they are good enough to make an inference.

Descriptive Statistics

- A statistic or a measure that describes the data
  - Average salary of employees
- Describing data with tables and graphs (quantitative or categorical variables)
- Numerical descriptions
  - Center – Give some example measures of center of the data
  - Variability– Give some example measures of variability of the data
- Bivariate descriptions (In practice, most studies have several variables)
  - Dependency measures(Correlation)

Simple Descriptive Statistics

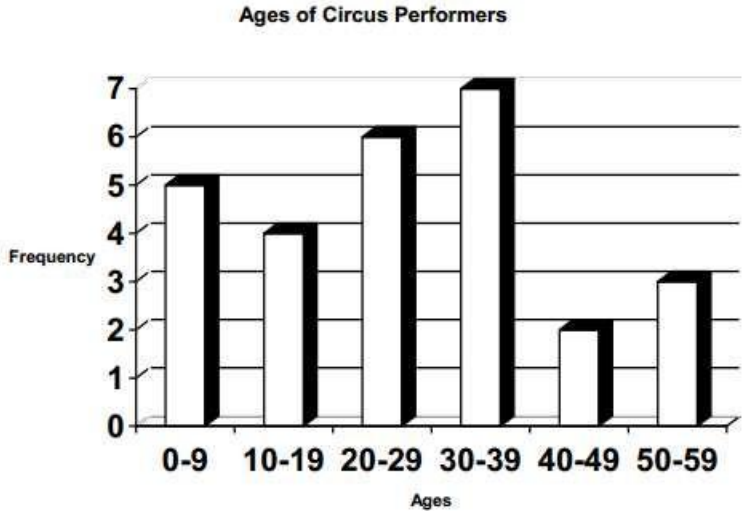
- N
- Sum
- Min
- Max
- Average
- Frequency of each level
- Variance
- Standard deviation

These simple descriptive statistics will be use in inferential statistics later.

Frequency tables & Histograms

- Frequency distribution: Lists possible values of variable and number of times each occurs

Ages of Circus Performers		
Intervals	Tally Marks	Frequency
0 – 9	𝍸	5
10 – 19	||||	4
20 – 29	𝍸 |	6
30 – 39	𝍸 ||	7
40 – 49	||	2
50 – 59	|||	3



Shapes of histograms

- Bell-shaped (IQ, SAT, political ideology in all U.S.)
- Skewed right
  - Example Annual income
  - No. times arrested
- Skewed left
  - Score on easy exam
  - Daily level if excitement in office
- Bimodal
  - Hardworking days in a year (Peaks near Mid year & year end
  - Appraisal)

Lab : Histogram

- Create a histogram on variable ‘actual’ in prdsale data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?
- Create a histogram on variable ‘msrp’ in cars data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?
- Create a histogram on variable ‘weight’ in cars data
  - How many modes?
  - What is the skewness?
  - What is its kurtosis?

Compare the above three histograms.

Central tendency

- What is the flight fare from Bangalore to Delhi? 3500–Exact or average?
- What is central tendency? - Average
- Three types of Averages
  - Mean
  - Median
  - Mode

Mean

- Center of gravity
- Evenly partitions the sum of all measurement among all cases; average of all measures

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- Crucial for inferential statistics
- Mean is not very resistant to outliers –See in Median

Median

- What is the mean of [0.1 0.8 0.4 0.3 0.1 0.4 9.0 0.1 0.9 0.1] 0.3 1.0 0.3

- Guess without calculation – Around **0.5**?
- Now calculate the mean

- Median is exactly in the middle. Isn't mean exactly in the middle
- Order the observations in ascending or descending order and pick the middle observation
- less useful for inferential purposes
- More resistant to effects of outliers…

Calculation of Median

rim diameter (cm)

unit 1	unit 2	
9.7	9.0	
11.5	11.2	
11.6	11.3	
12.1	11.7	
12.4	12.2	
12.6	12.5	
12.9 <--	13.2	13.2
13.1	13.8	
13.5	14.0	
13.6	15.5	
14.8	15.6	
16.3	16.2	
26.9	16.4	

Mode

- How do you express average size of the shoes ?
  - 6.567 or 6?

- Mode is the most numerous category
- Can be more or less created by the grouping procedure
- For theoretical distributions—simply the location of the peak on the frequency distribution

Lab

- Run Proc means data product data
- What is the mean of ‘msrp’ in cars data?
- Is it reflecting the average value of price?
- What is median of ‘msrp’ in cars data?
- Is it reflecting the average value of price?
- Run Proc Univariate on weight varaibale in cars data. Find mean, Median & Mode.

Dispersion

Person1: What is the average depth of this river? 5 feet

Person2: I am 5.5 I can easily cross it(and starts crossing it)

Person 2: Help….help.

Person 1: Some times just knowing the central tendency is not sufficient

- Measures of dispersion summarize the degree of clustering/spread of cases, esp. with respect to central tendency…
  - range
  - variance
  - standard deviation

Range

- Max –Min

R: range(x)

unit 1	unit 2
9.7	9.0
11.5	11.2
11.6	11.3
12.1	11.7
12.4	12.2
12.6	12.5
13.1	13.2
13.5	13.8
13.6	14.0
14.8	15.5
16.3	15.6
26.9	16.2
	16.4

Variance

- Take deviation from Mean- It can be zero some times
- Hence take square of deviation from mean Take average of that
- Average mean squared distance is **variance**

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

- Units of variance are squared… this makes variance hard to interpret
- Eg : Mean length = 22.6 mm variance = 38 mm^2
- What does this mean??? –I don't Know

Standard Deviation

- Square root of variance

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

- Units are in same units as base measurements
- Mean = 22.6 mm standard deviation = 6.2 mm
- Mean +/- sd (16.4—28.8 mm)
  - should give at least some intuitive sense of where most of the cases lie, barring major effects of outliers

Quartiles & Percentiles

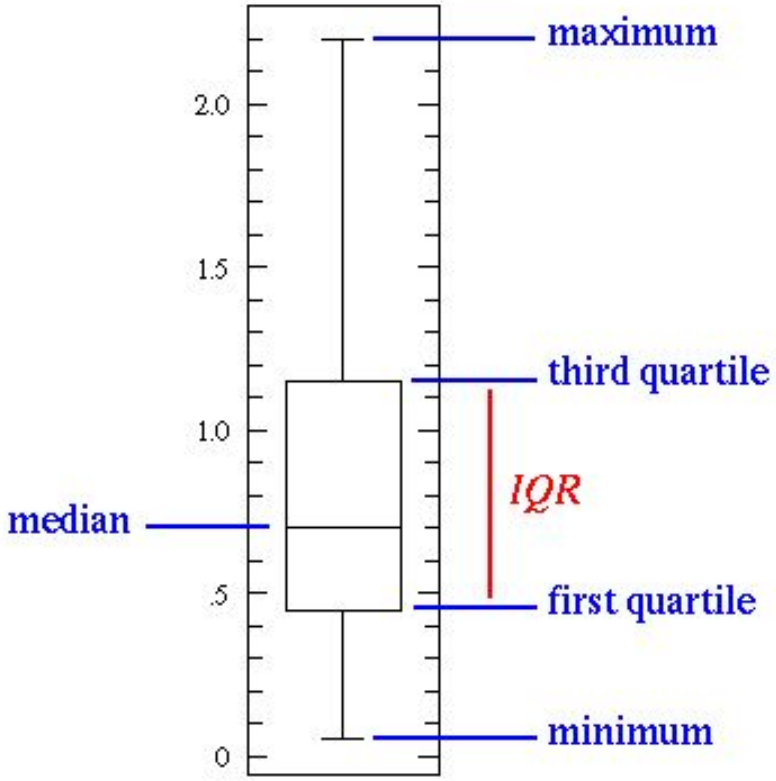
- pth percentile: p percent of observations below it, (100 - p)% above it.
- Like 95% of CAT percentile means 5% are above & 95% are below
- 1,2,3,4,5,6,7,8,9,10 - What is 25^{th} percentile?
- 1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20 - What is 25^{th} percentile? What is 80^{th} percentile?

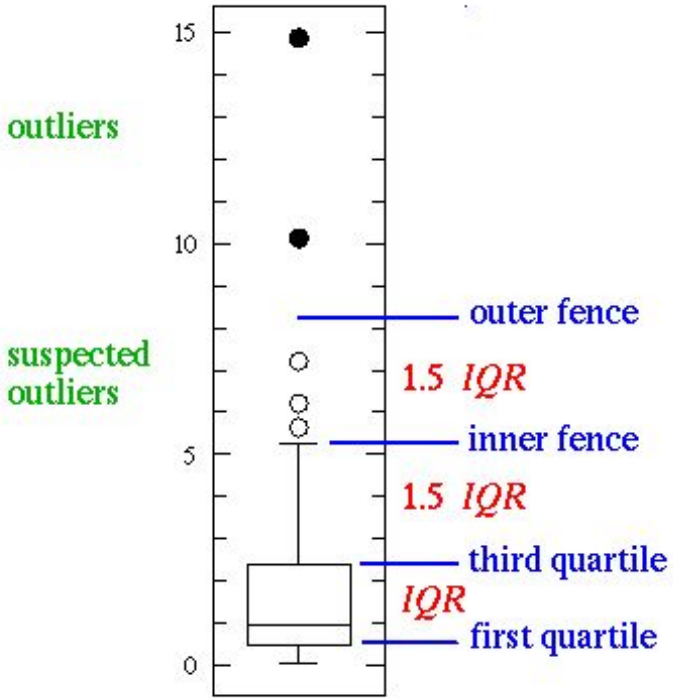
  - p = 50: median
  - p = 25: lower quartile (LQ)
  - p = 75: upper quartile (UQ)

- Interquartile range IQR = UQ - LQ

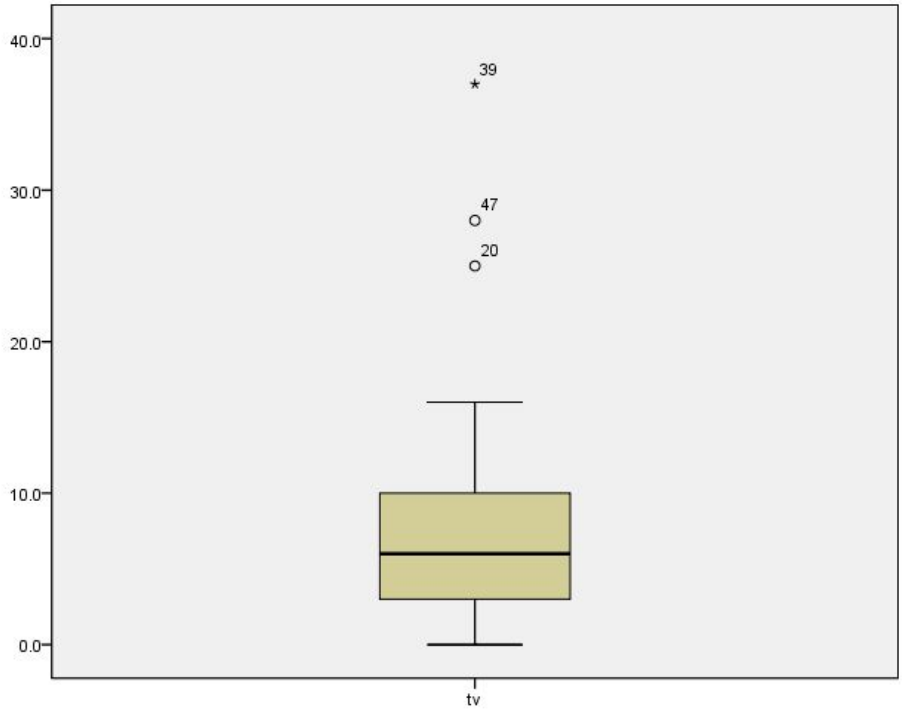
Box Plots

- Quartiles portrayed graphically by box plots





Box Plots



Example: weekly TV watching for *n*=60, 3 outliers

Box Plots Interpretation

- Box plots have box from LQ to UQ, with median marked. They portray a five-number summary of the data: Minimum, LQ, Median, UQ, Maximum
- Except for outliers identified separately
- **Outlier** = observation falling

  below LQ – 1.5(IQR) or above UQ + 1.5(IQR)
- Ex. If LQ = 2, UQ = 10, then IQR = 8 and outliers above 10 + 1.5(8) = 22

Lab

- Run proc univariate on a variable from sample data in sas default library(prd sale / cars)
- Run proc means on actual & predicted variables from product sales data
- What are the values of Range, Variance, SD
- What are 1,2,3 & 4 quartile values
- What is 95^{th} percentile?
- Use “all” option to display the box plots

Contingency Tables

- Cross classifications of categorical variables in which rows (typically) represent categories of explanatory variable and columns represent categories of response variable.
- Counts in "cells" of the table give the numbers of individuals at the corresponding combination of levels of the two variables

Example: Happiness and Family Income of 1993 families (GSS 2008 data: "happy," "finrela")

	Happiness			
Income	Very	Pretty	Nottoo	Total
Above Aver.	**164**	**233**	**26**	**423**
Average	293	473	117	883
Below Aver.	132	383	172	687
Total	589	1089	315	1993

Contingency tables

- Example: Percentage “very happy” is
  - 39% for above average income (164/423 = 0.39)
  - 33% for average income (293/883 = 0.33)
  - What percent for below average income?

	Happiness			
Income	Very	Pretty	Not oo	Total
Above	**164 (39%)**	**233 (55%)**	**26 (6%)**	**423**
Average	293 (33%)	473 (54%)	117 (13%)	883
Below	132 (19%)	383 (56%)	172 (25%)	687

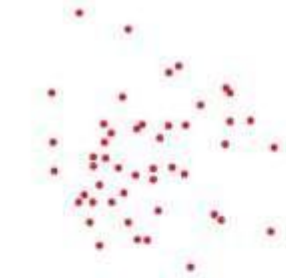
- What can we conclude? Is happiness depending on Income? Or Happiness is independent of Income?
- Inference questions for later chapters?

Correlation

- **Correlation** describes strength of association between two variables
- Falls between -1 and +1, with sign indicating direction of association (formula & other details later)
- The larger the correlation in absolute value, the stronger the association (in terms of a straight line trend)
- **Examples**: (positive or negative, how strong?)
  - Mental impairment and life events, correlation =
  - GDP and fertility, correlation =
  - GDP and percent using Internet, correlation =

Strength of Association

- Correlation 0 No linear association
- Correlation 0 to 0.25 Negligible positive association
- Correlation 0.25-0.5 Weak positive association
- Correlation 0.5-0.75 Moderate positive association
- Correlation >0.75 Very Strong positive association
- What are the limits for negative correlation

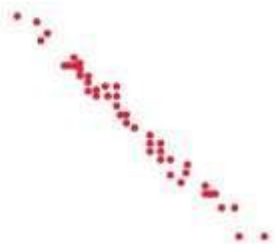

Correlation $r = 0$


Correlation $r = -0.3$


Correlation $r = 0.5$


Correlation $r = -0.7$


Correlation $r = 0.9$


Correlation $r = -0.99$

Regression

- **Regression analysis** gives line predicting *y* using *x*(algorithm & other details later)

- *y* = college GPA, *x* = high school GPA
- Predicted y = 0.234 + 1.002(x)

Lab

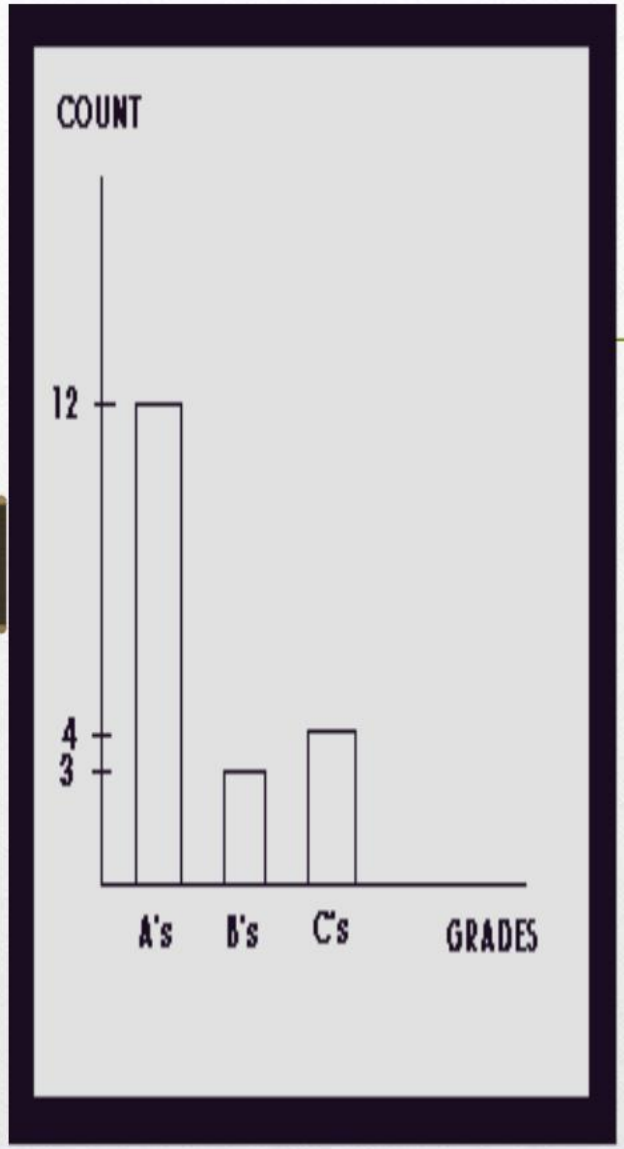
- Create a contingency table for product sales data
- Find contingency tables for
  - Region by product type
  - Division by Product type
- Find the correlation between actual sales and predicted sales.
- Find the correlation between weight & msrp in cars data

Graphs and their use

- Histogram & Box plots are used for continuous or scale variables like temperature, Bone density etc.
- Bar chart & Pie Charts are used to categorical or nominal variables like gender, name etc.
- Scatterplots . Used to measure to continuous variables.

BAR GRAPHS.

- **Bar graphs** are frequently used with the categorical data to compare the sizes of categories


COUNT
12
4
3
A's
B's
C's
GRADES

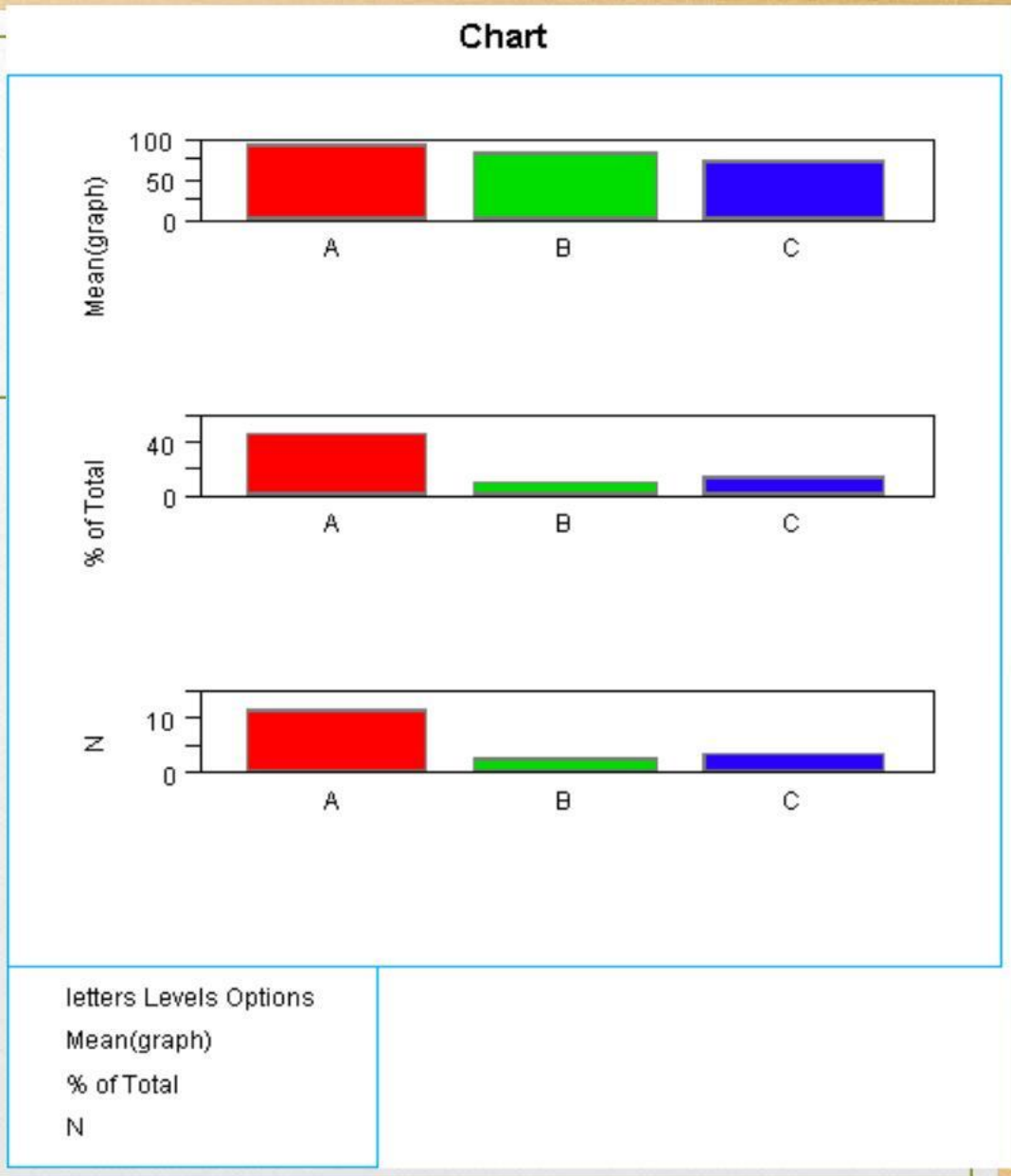
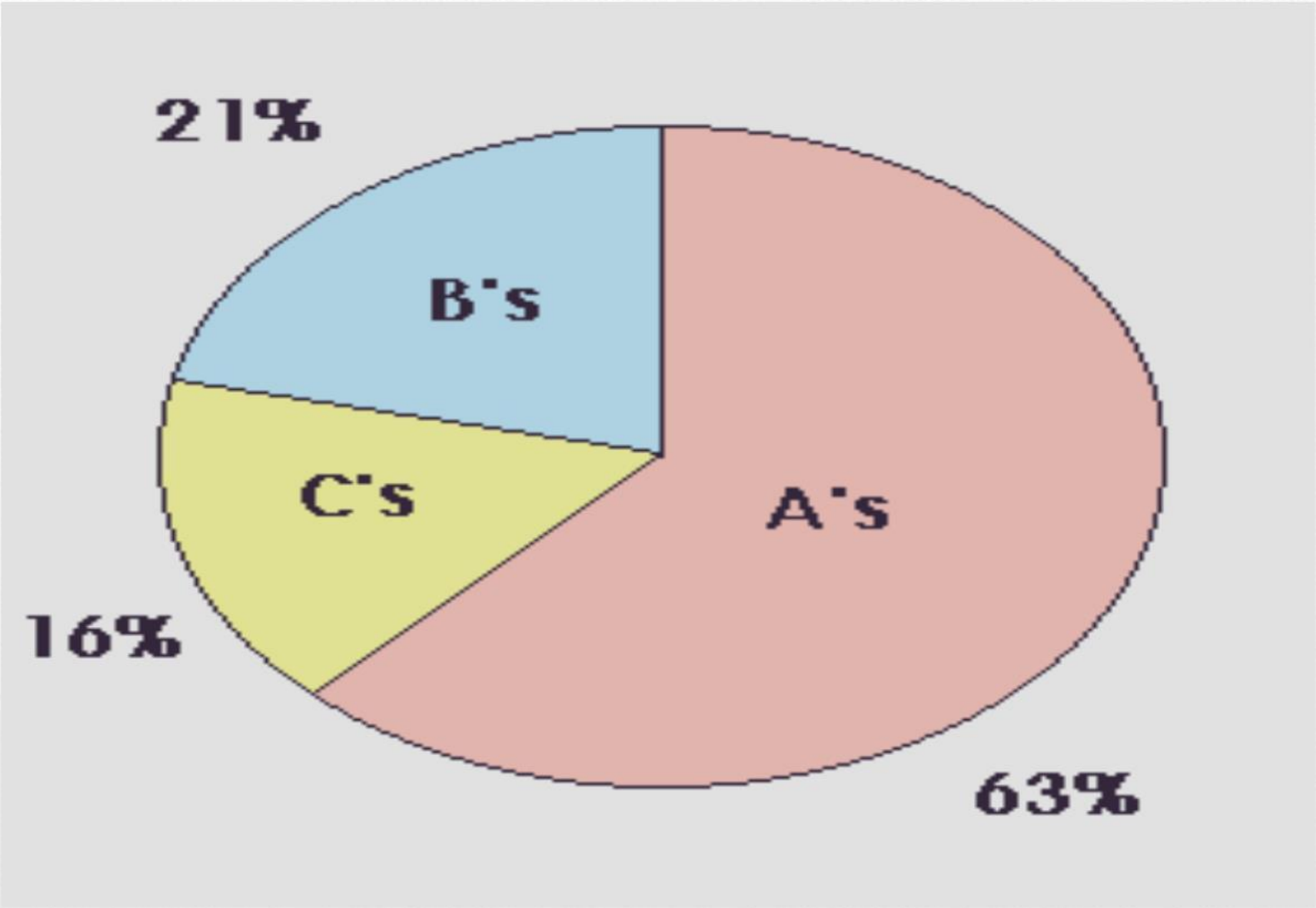


Chart
Mean(graph)
100
50
0
A
B
C
% of Total
40
0
A
B
C
N
10
0
A
B
C
letters Levels Options
Mean(graph)
% of Total
N

PIE CHARTS

- Like bar graphs, **pie charts** are best used with categorical data to help us see what percentage of the whole each category constitutes. Pie charts require all categories to be included in a graph. Each graph always represents the whole.
- One of the reasons why bar graphs are more flexible than pie charts is the fact that bar graphs compare selected categories, whereas pie charts must either compare all categories or none.


21%
B's
C's
A's
16%
63%

QUANTITATIVE VARIABLES

- **STEM PLOTS.**
- **Stemplots** (sometimes called stem-and-leaf plots) are used with quantitative data to display shapes of distributions, to organize numbers and make them more comprehensible.
- It is a descriptive technique which gives a good overall impression of the data. Stemplots include the actual numerical values of the observations, where each value is separated into two parts, a stem and a leaf.
- A stem is usually the first digit, or the leftmost digit(s), and a leaf is the final rightmost digit. We write the stems in a vertical column with the smallest at the top, and draw a vertical line to the right of the column. Finally, we write the leaves in the row to the right of the corresponding stem, starting with the smallest one.

STEM PLOTS.

- Grades. The average test grades of 19 students are as follows (on a scale from 0 to 100, with 100 being the highest score): **92 95 96 81 95 75 91 79 92 100 89 94 92 86 93 73 74 94 91**
- **Colour coordinated, in increasing order:**
- **73, 74, 75, 79, 81, 86, 89, 91, 91, 92, 92, 92, 93, 94, 94, 95, 95, 96, 100**

STEMPLOT#1:

stem	leaf
7	3 4
7	5 9
8	1
8	6 9
9	1 1 2 2 2 3 4 4
9	5 5 6
10	0
10	

STEMPLOT#2:

stem	leaf
7	3 4 5 9
8	1 6 9
9	1 1 2 2 2 3 4 4 5 5 6
10	0

Depending on the number of stems, different conclusions can be drawn about a given data set. In this example, even though both stemplots show a slight left-skew, stemplot #1 reflects that more evidently than stemplot #2.

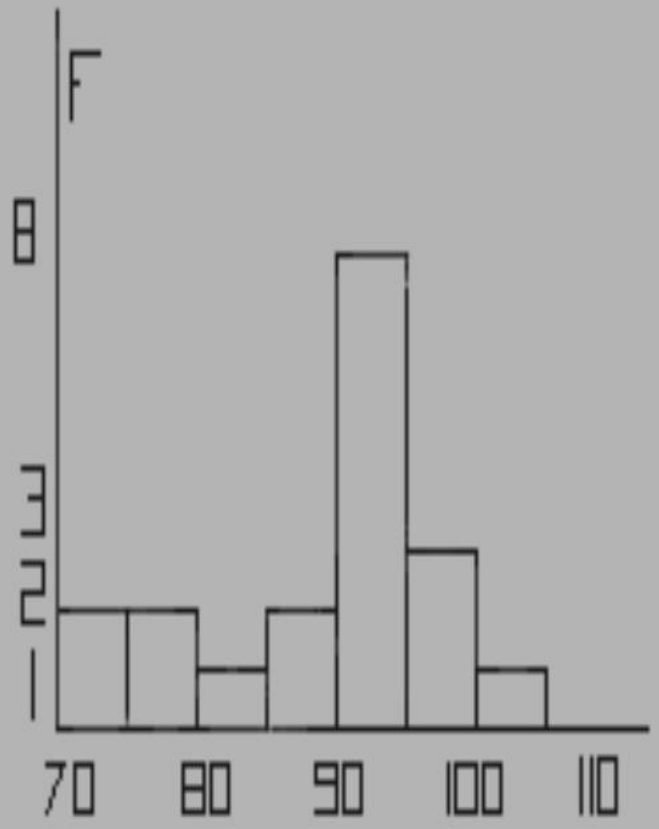
Stem and Leaf Plots

- .Simple way to order and display a data set.
- Abbreviate the observed data into two significant digits.

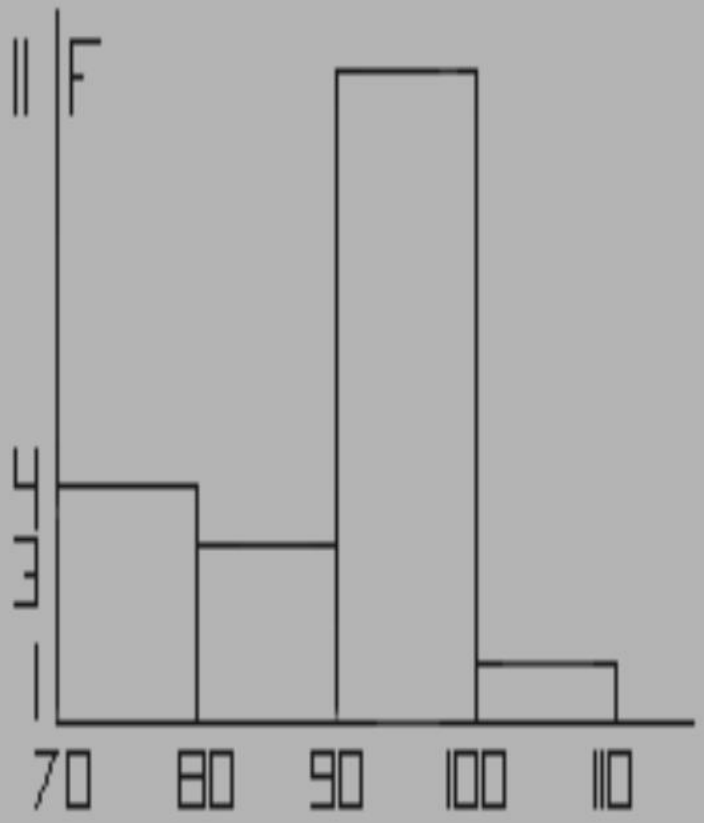
Stem	Leaf
0	6 1 4
1	1 3 5
2	6 2 0
3	2

HISTOGRAMS

- **Histograms** are yet another graphic way of presenting data to show the distribution of the observations. It is one of the most common forms of graphical presentation of a frequency distribution



1



2

BOXPLOTS

- **Boxplots** reveal the main features of a batch of data, i.e. how the data are spread out.
  - Any boxplot is a graph of the five-number summary: the minimum score, first quartile (Q1-the median of the lower half of all scores), the median, third quartile (Q3-the median of the upper half of all scores), and the maximum score, with suspected outliers plotted individually.

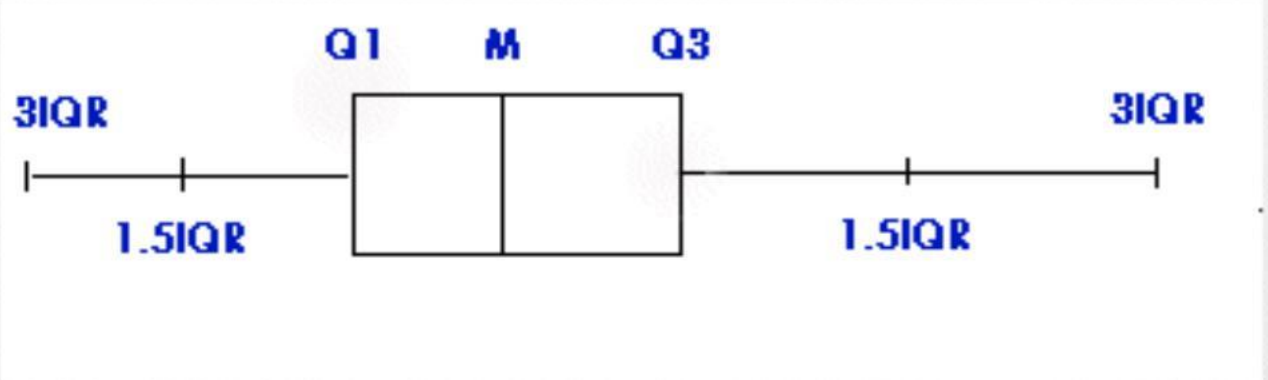
Continued (Explainable from Graph)

- The boxplot consists of a rectangular box, which represents the middle half of all scores (between Q1 and Q3). Approximately one-fourth of the values should fall between the minimum and Q1, and approximately one-fourth should fall between Q3 and the maximum. A line in the box marks the median. Lines called whiskers extend from the box out to the minimum and maximum scores that are not possible outliers. If an observation falls more than 1.5x IQR outside of the box, it is plotted individually as an outlier.

BOXPLOTS

- **FIVE-NUMBER SUMMARY:**
- MINIMUM
- 1ST QUARTILE
- MEDIAN
- 3RD QUARTILE
- MAXIMUM

IQR, or the interquartile range, is the distance between the first and third quartiles. IQR = Q3 - Q1



Summaries for numerical data

- Center/location: measures the "center" of the data
  - Examples: sample mean and sample median
- Spread/Dispersion: measures the "spread" or "fatness" of the data
  - Examples: sample variance, interquartile range
- Order/Rank: measures the ordering/ranking of the data
  - Examples: order statistics and sample quantiles

Summary	Type of Sample	Formula	Notes
	Continuous		• Summarizes the "center" of the data • Sensitive to outliers
	Continuous		• Summarizes the "spread" of the data • Outliers may inflate this value
	Continuous	i^{th} largest value of the sample	• Summarizes the order/rank of the data
	Continuous		• Summarizes the "center" of the data • Robust to outliers
	Continuous		• Summarizes the order/rank of the data • Robust to outliers
Sample Interquartile Range (Sample IQR)	Continuous		• Summarizes the "spread" of the data • Robust to outliers

Multivariate numerical data

- Each dimension in multivariate data is univariate and hence, we can use the numerical summaries from univariate data (e.g. sample mean, sample variance)

- However, to study two measurements and their relationship, there are numerical summaries to analyze it

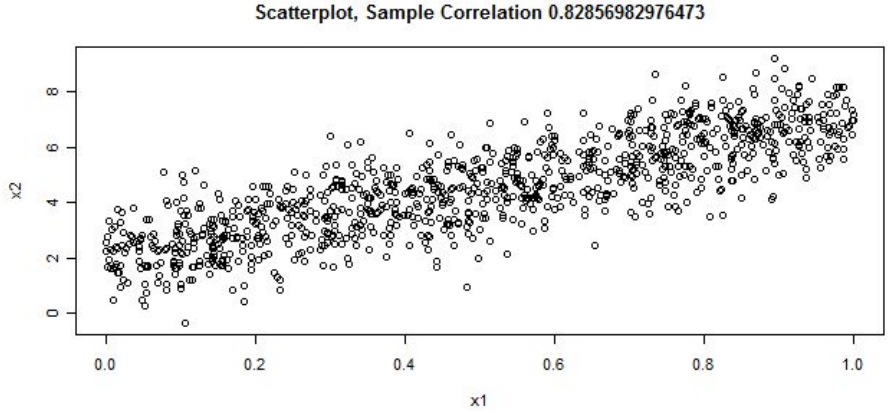
- Sample Correlation and Sample Covariance

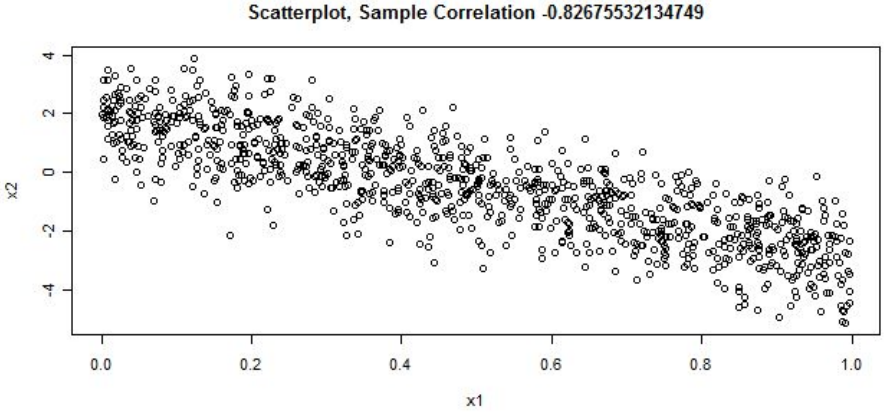
Sample Correlation and Covariance

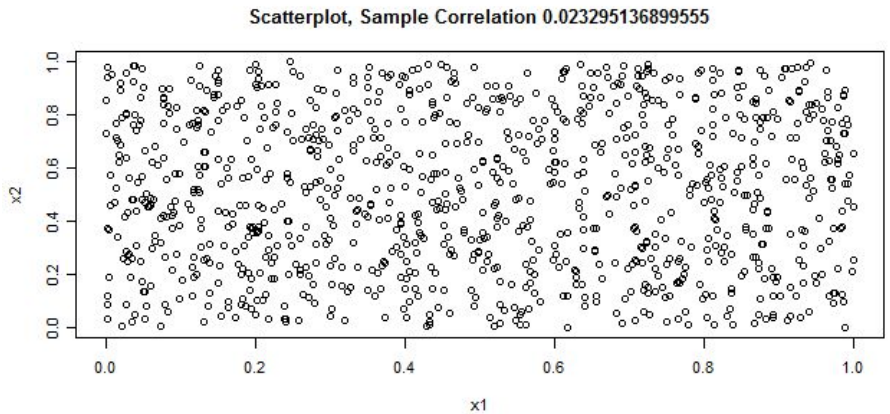
- Measures linear relationship between two measurements, X_{i1} and X_{i2}, where $X_i = (X_{i1}, X_{i2})$

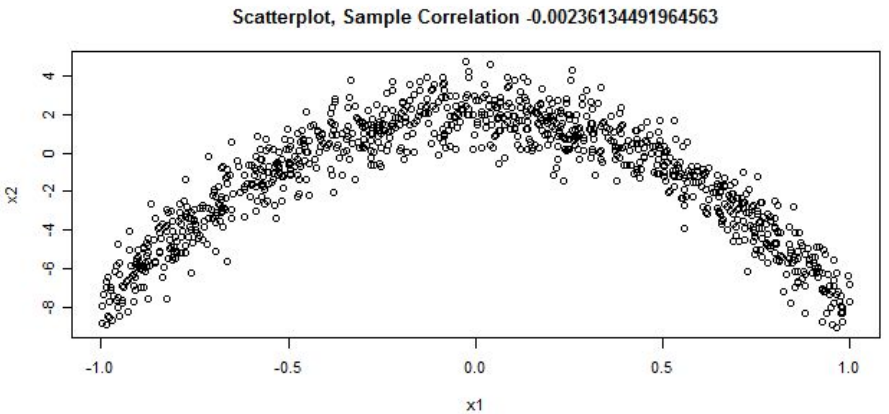
- $\hat{\rho} = \frac{\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(X_{i2}-\bar{X}_2)}{(n-1)\hat{\sigma}_{X_1}\hat{\sigma}_{X_2}}$
  - $-1 \leq \hat{\rho} \leq 1$
  - Sign indicates proportional (positive) or inversely proportional (negative) relationship
  - If X_{i1} and X_{i2} have a perfect linear relationship, $\hat{\rho} = 1$ or -1

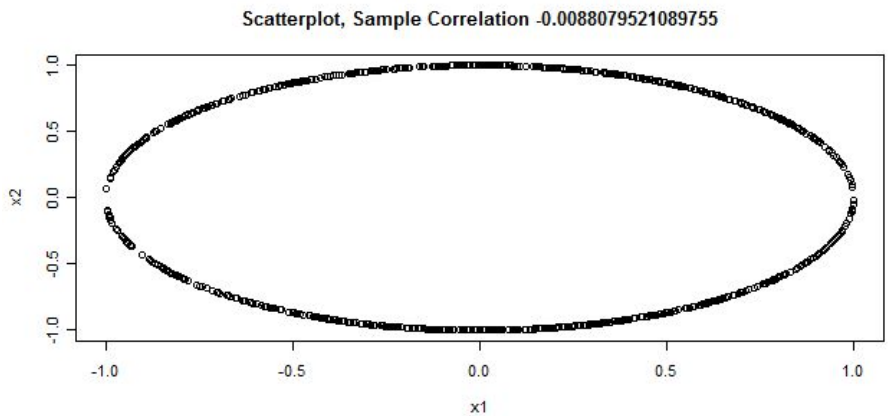
- Sample covariance $= \hat{\rho}\hat{\sigma}_{X_1}\hat{\sigma}_{X_2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$

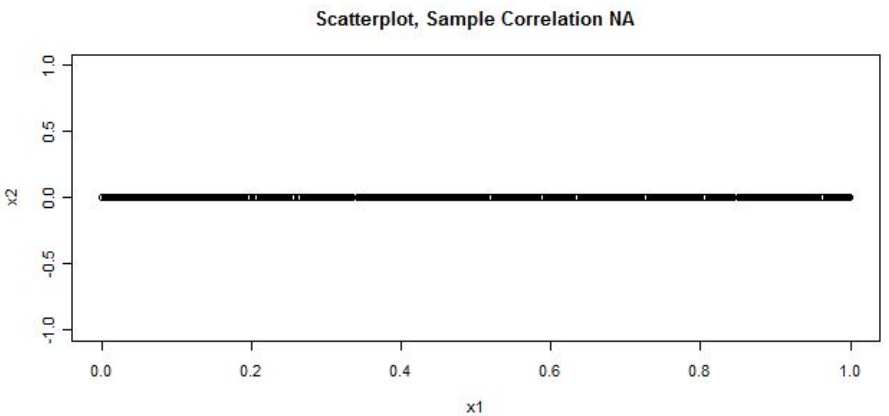

Scatterplot, Sample Correlation 0.82856982976473
x2
x1
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8


Scatterplot, Sample Correlation -0.82675532134749
x2
x1
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4


Scatterplot, Sample Correlation 0.023295136899555
x2
x1
0.0
0.2
0.4
0.6
0.8
1.0


Scatterplot, Sample Correlation -0.00236134491964563
x2
x1
-1.0
-0.5
0.0
0.5
1.0
-8
-6
-4
-2
0
2
4


Scatterplot, Sample Correlation -0.0088079521089755
x2
x1
-1.0
-0.5
0.0
0.5
1.0


Scatterplot, Sample Correlation NA
x2
x1
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0

How about categorical data?

Summaries for categorical data

- Frequency/Counts: how frequent is one category
- Generally use tables to count the frequency or proportions from the total
- Example: Stat 431 class composition

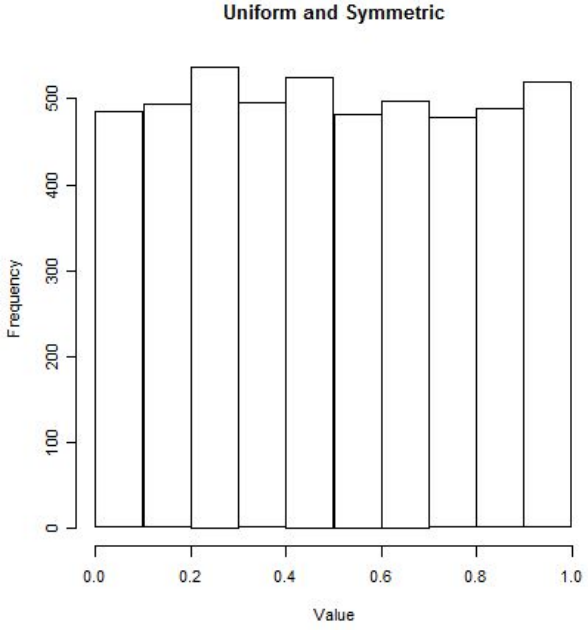
	Undergrad	Graduate	Staff
Counts	17	1	2
Proportions	0.85	0.05	0.1

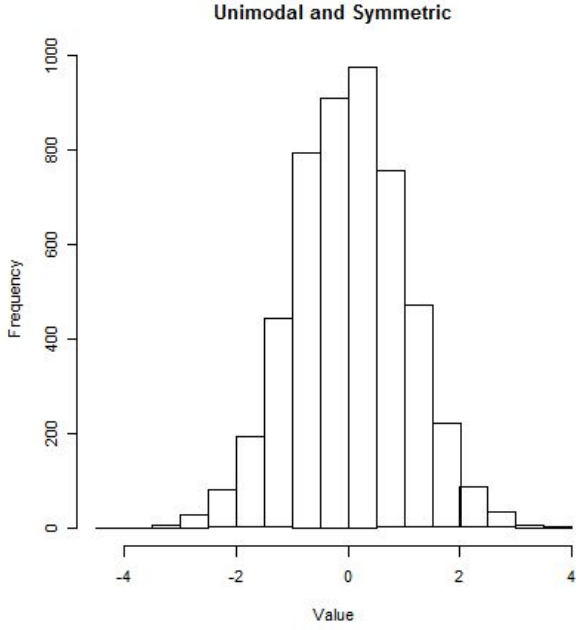
Are there visual summaries of the data?

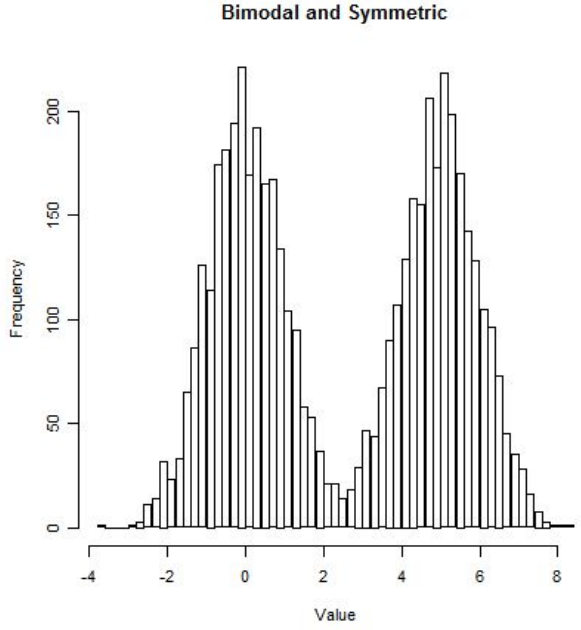
Histograms, boxplots, scatterplots, and QQ plots

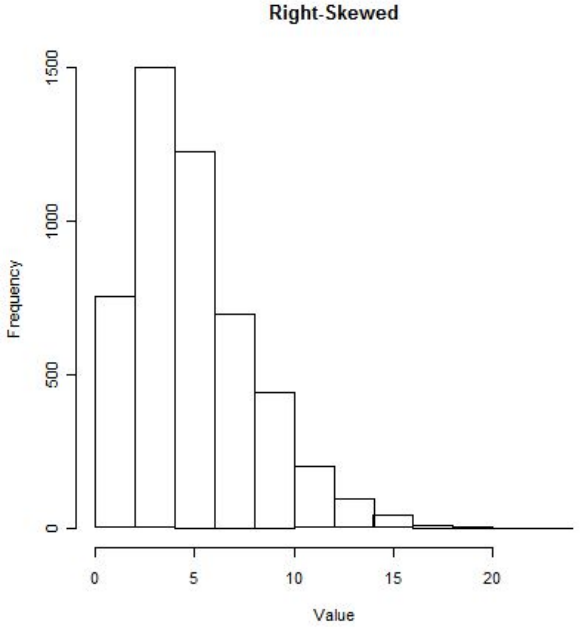
Histograms

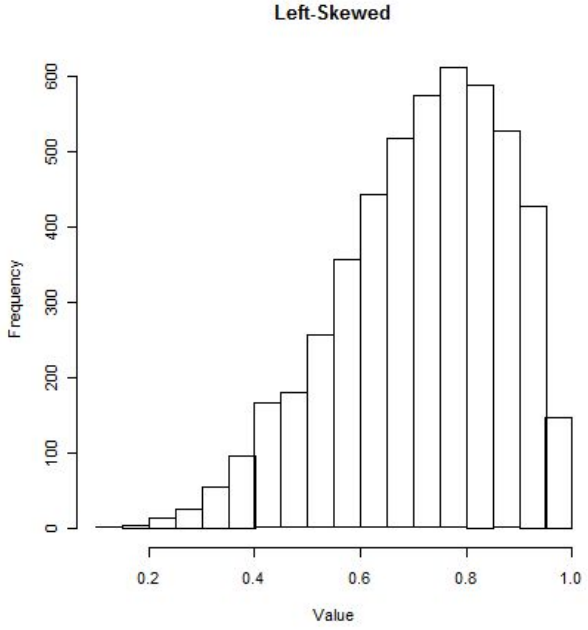
- For numerical data
- A method to show the "shape" of the data by tallying frequencies of the measurements in the sample
- Characteristics to look for:
  - Modality: Uniform, unimodal, bimodal, etc.
  - Skew: Symmetric (no skew), right/positive-skewed, left/negative-skewed distributions
  - Quantiles: Fat tails/skinny tails
  - Outliers

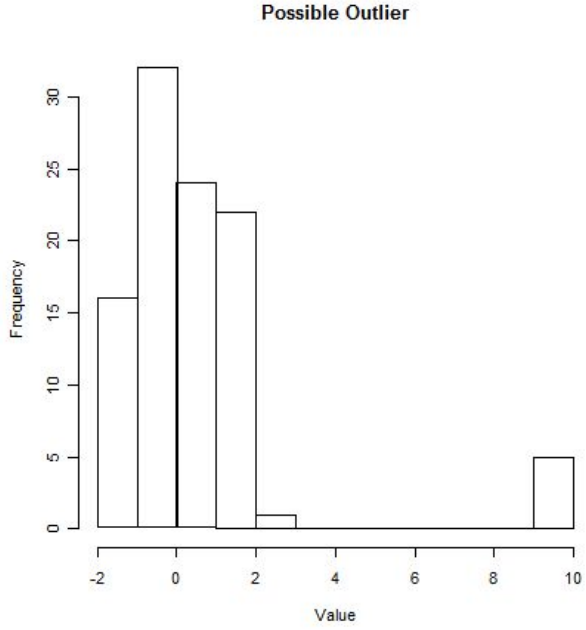

Uniform and Symmetric
Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
Value

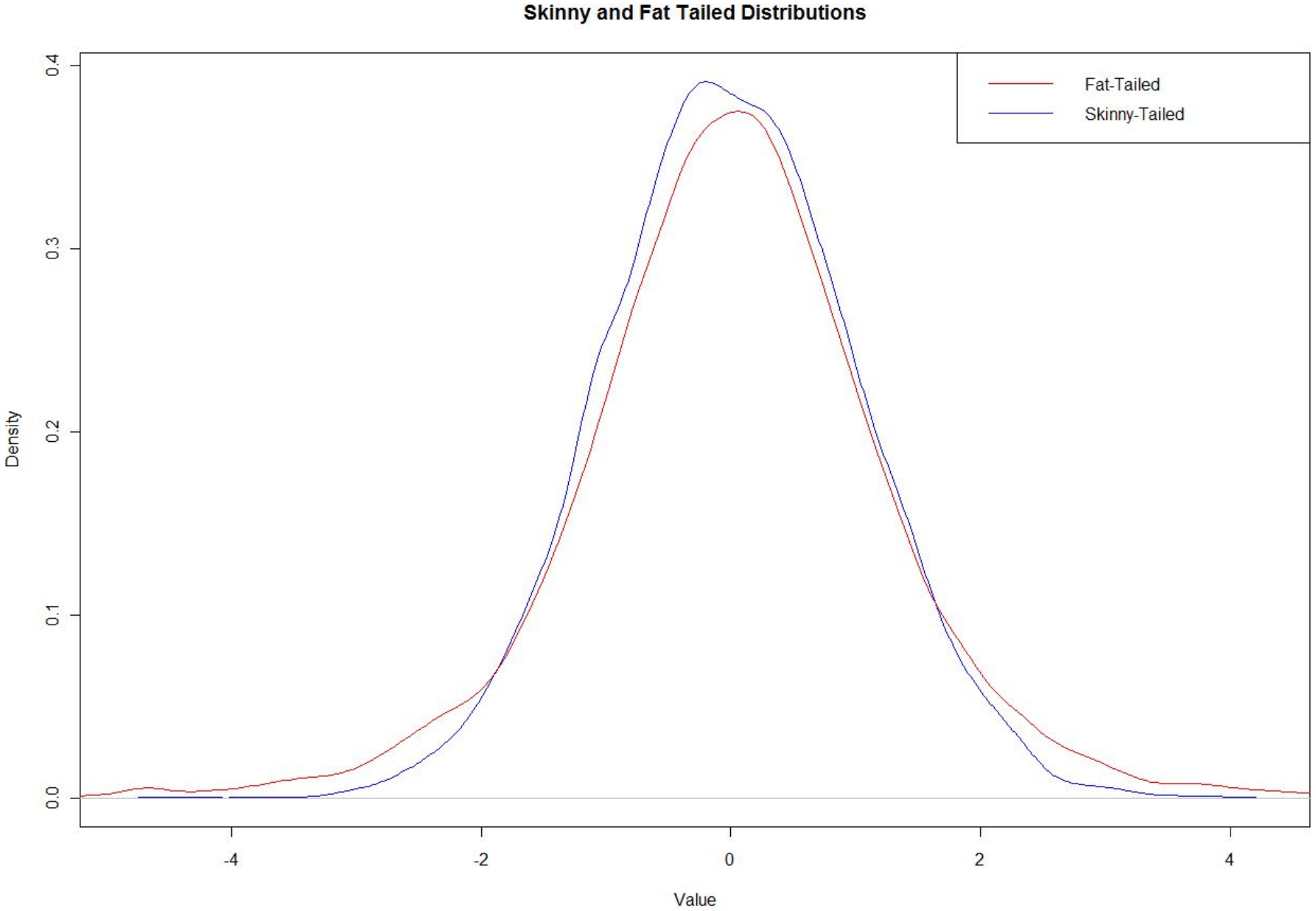

Unimodal and Symmetric
Frequency
0
200
400
600
800
1000
-4
-2
0
2
4
Value


Bimodal and Symmetric
Frequency
0
50
100
150
200
-4
-2
0
2
4
6
8
Value


Right-Skewed
Frequency
0
500
1000
1500
0
5
10
15
20
Value

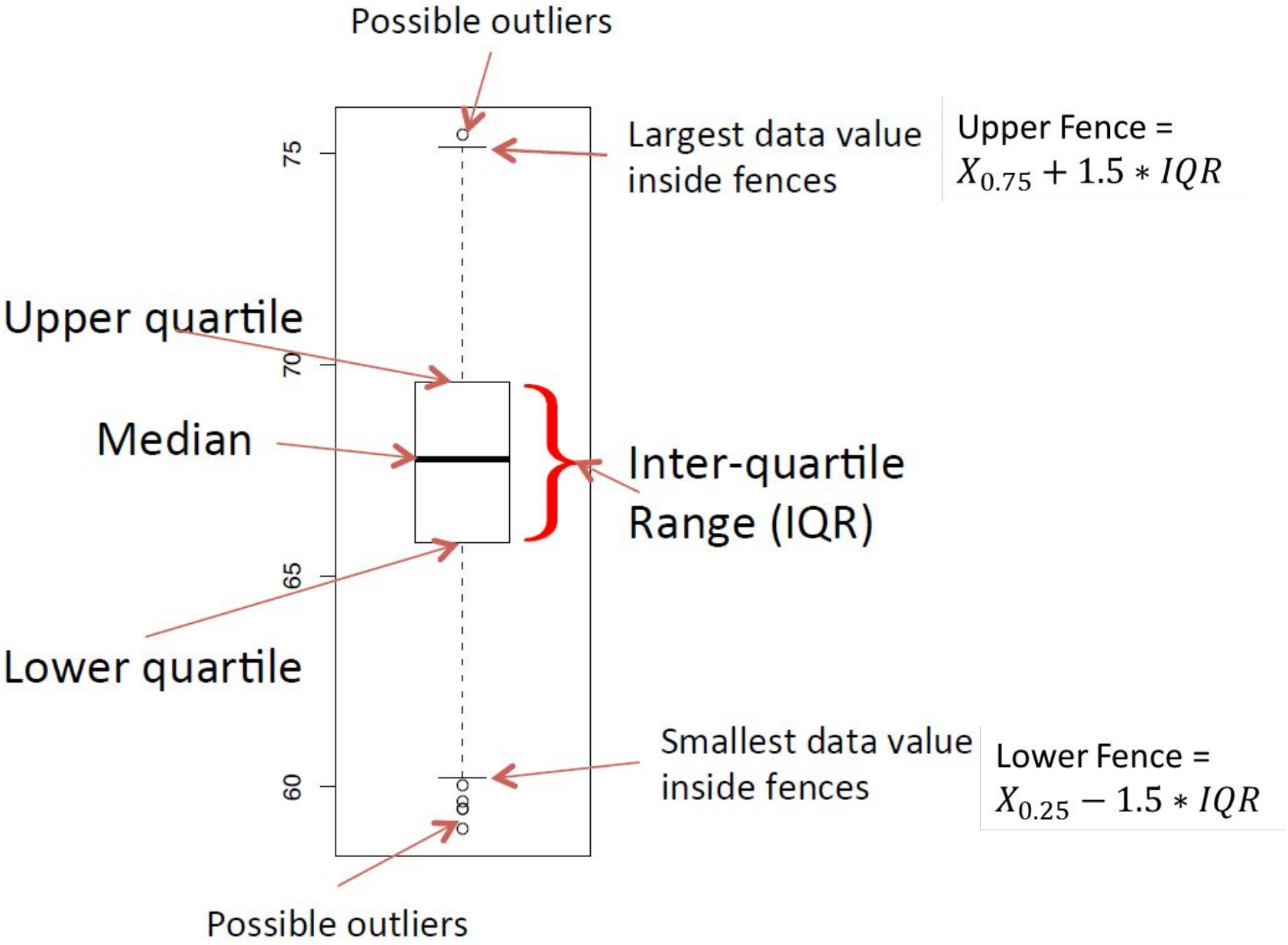

Left-Skewed
Frequency
0
100
200
300
400
500
600
0.2
0.4
0.6
0.8
1.0
Value

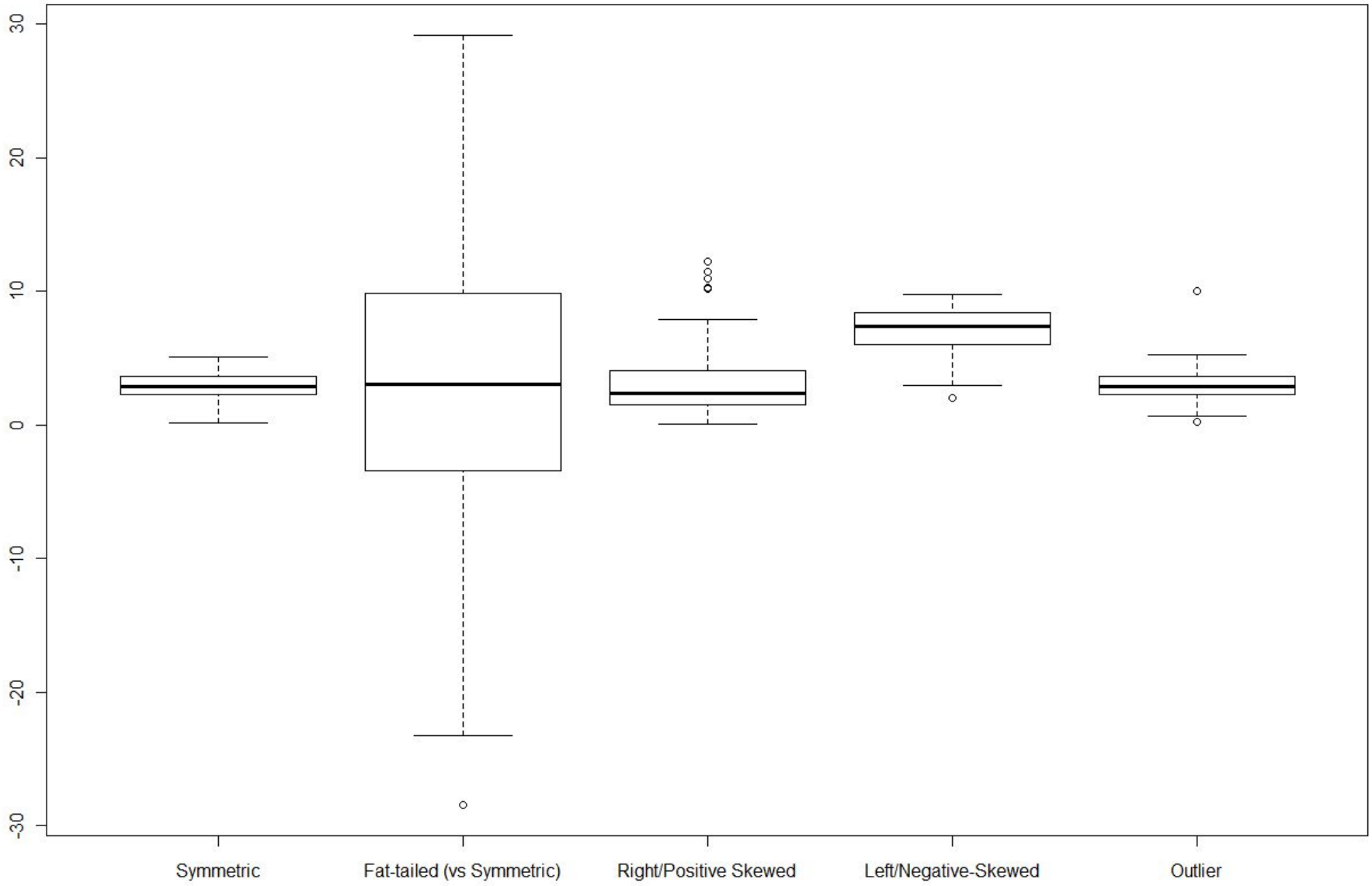

Possible Outlier
Frequency
0
5
10
15
20
25
30
-2
0
2
4
6
8
10
Value


Skinny and Fat Tailed Distributions
Fat-Tailed
Skinny-Tailed
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
Value

Boxplots

- For numerical data
- Another way to visualize the "shape" of the data. Can identify…
  - Symmetric, right/positive-skewed, and left/negative-skewed distributions
  - Fat tails/skinny tails
  - Outliers
- However, boxplots cannot identify modes (e.g. unimodal, bimodal, etc.)


Possible outliers
Largest data value inside fences
Upper Fence = $X_{0.75} + 1.5 * IQR$
Upper quartile
Median
Inter-quartile Range (IQR)
Lower quartile
Smallest data value inside fences
Lower Fence = $X_{0.25} - 1.5 * IQR$
Possible outliers
75
70
65
60

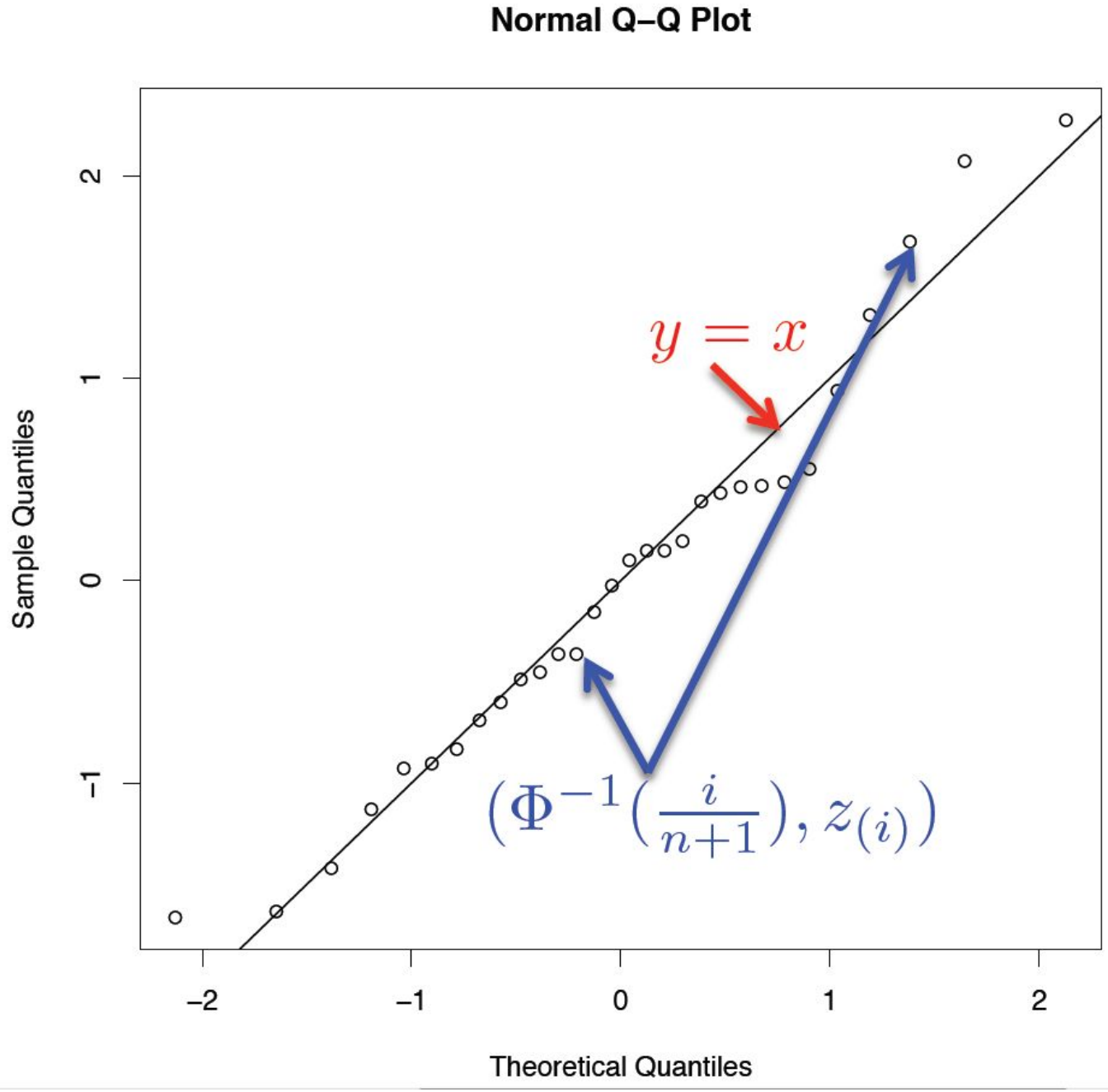

30
20
10
0
-10
-20
-30
Symmetric
Fat-tailed (vs Symmetric)
Right/Positive Skewed
Left/Negative-Skewed
Outlier

Quantile-Quantile Plots (QQ Plots)

- For numerical data: visually compare collected data with a known distribution

- Most common one is the Normal QQ plots
  - We check to see whether the sample follows a normal distribution
  - This is a common assumption in statistical inference that your sample comes from a normal distribution

- Summary: If your scatterplot "hugs" the line, there is good reason to believe that your data follows the said distribution.

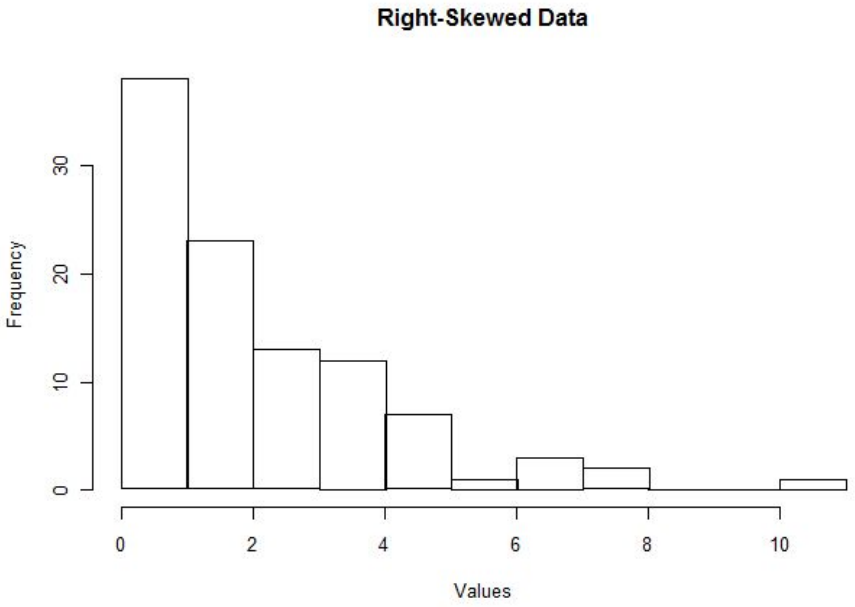

Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
2
1
0
−1
−2
−1
0
1
2
$y = x$
$(\Phi^{-1}(\frac{i}{n+1}), z_{(i)})$

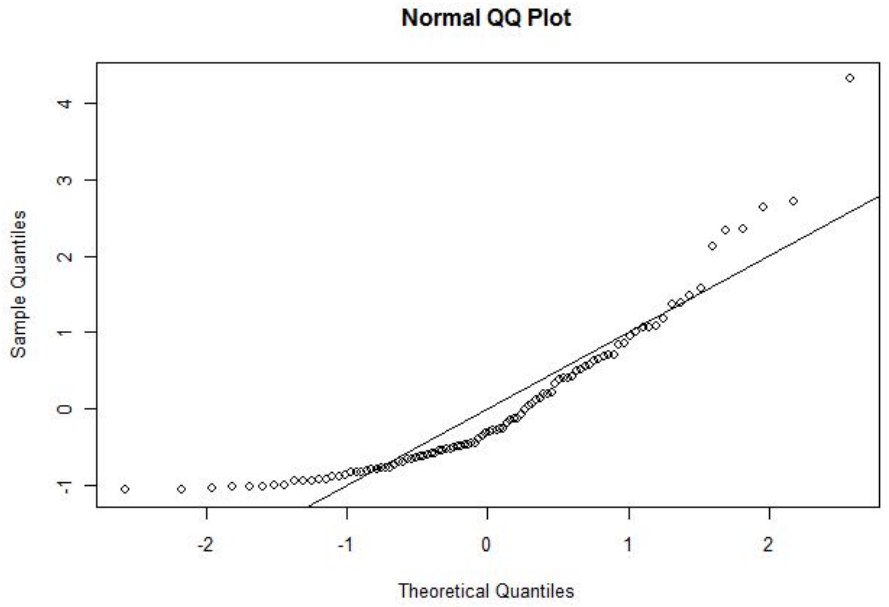
Making a Normal QQ plot

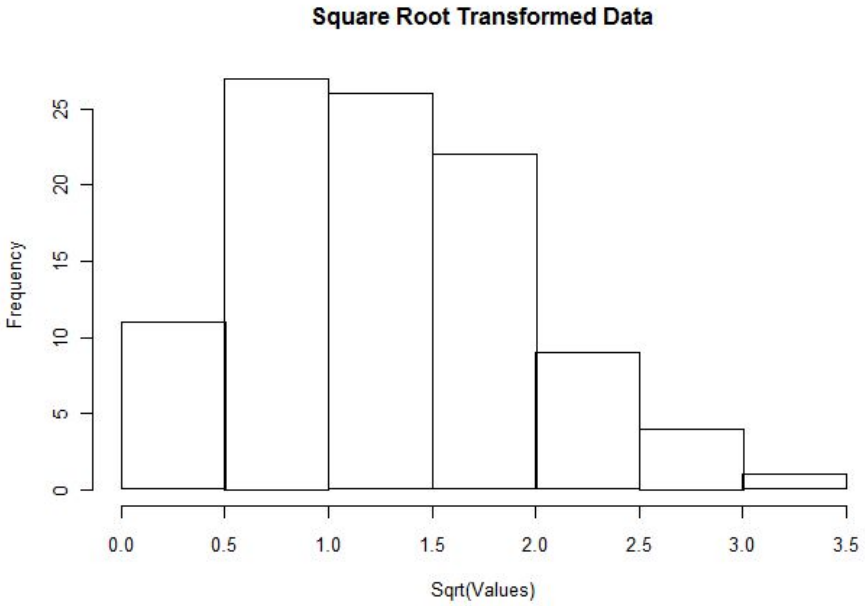
1. Compute z-scores: $\mathrm{Z_i} = \frac{X_i - \bar{\mathrm{X}}}{\hat{\sigma}}$
2. Plot $\frac{i}{n+1}$th theoretical normal quantile against ith ordered z-scores (i.e. $\left(\Phi\left(\frac{i}{n+1}\right)^{-1}, Z_{(i)}\right)$
   - Remember, $Z_{(i)}$ is the $\frac{i}{n+1}$ sample quantile (see numerical summary table)
3. Plot $Y = X$ line to compare the sample to the theoretical normal quantile

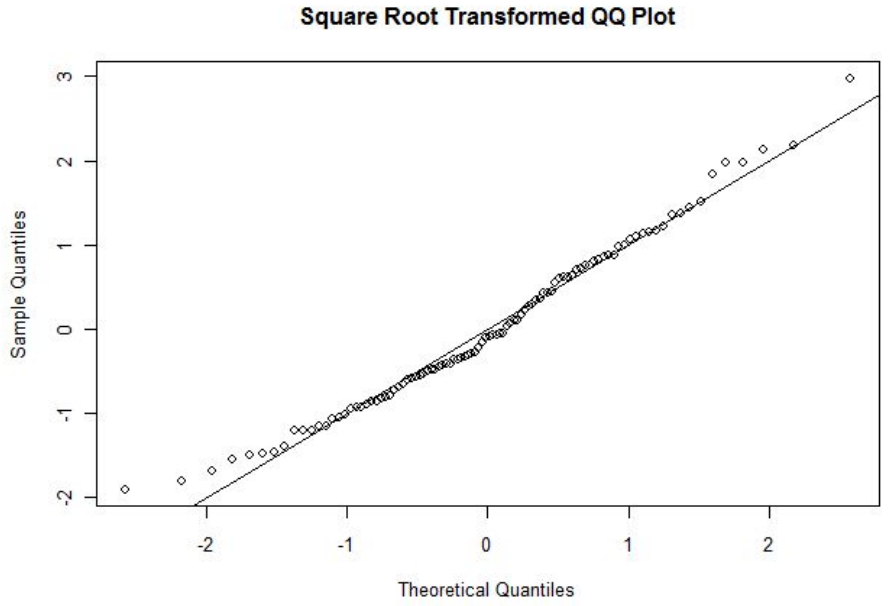
If your data is not normal…

- You can perform transformations to make it look normal
- For right/positively-skewed data: Log/square root
- For left/negatively-skewed data: exponential/square


Right-Skewed Data
Frequency
0
10
20
30
0
2
4
6
8
10
Values


Normal QQ Plot
Sample Quantiles
-1
0
1
2
3
4
-2
-1
0
1
2
Theoretical Quantiles


Square Root Transformed Data
Frequency
0
5
10
15
20
25
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Sqrt(Values)


Square Root Transformed QQ Plot
Sample Quantiles
-2
-1
0
1
2
3
-2
-1
0
1
2
Theoretical Quantiles

Comparing the three visual techniques

Histograms

- Advantages:
  - With properly-sized bins, histograms can summarize any shape of the data (modes, skew, quantiles, outliers)
- Disadvantages:
  - Difficult to compare side-by-side (takes up too much space in a plot)
  - Depending on the size of the bins, interpretation may be different

Boxplots

- Advantages:
  - Don't have to tweak with "graphical" parameters (i.e. bin size in histograms)
  - Summarize skew, quantiles, and outliers
  - Can compare several measurements side-by-side
- Disadvantages:
  - Cannot distinguish modes!

QQ Plots

- Advantages:
  - Can identify whether the data came from a certain distribution
  - Don't have to tweak with "graphical" parameters (i.e. bin size in histograms)
  - Summarize quantiles
- Disadvantages:
  - Difficult to compare side-by-side
  - Difficult to distinguish skews, modes, and outliers

Scatterplots

- For multidimensional, numerical data:
  $\mathrm{X_i} = (X_{i1}, X_{i2}, \ldots, X_{ip})$

- Plot points on a p dimensional axis

- Characteristics to look for:
  - Clusters
  - General patterns

- See previous slide on sample correlation for examples.
  See R code for cool 3D animation of the scatterplot

Lecture Summary

- Once we obtain a sample, we want to summarize it.
- There are numerical and visual summaries
  - Numerical summaries depend on the data type (numerical or categorical)
  - Graphical summaries discussed here are mostly designed for numerical data
- We can also look at multidimensional data and examine the relationship between two measurement
  - E.g. sample correlation and scatterplots