Introduction to Data Analytics
DSE-2151

B.tech Data Science & Engineering
3rd Semester
Exploratory Analysis

Rohini R. Rao & Manjunath Hegde
Dept of Computer Applications
Sept 2021
(Slide set 2 out of 5)

Contents

Population vs. Sample

- **Population:**
  - A precise definition of all possible outcomes, measurements or values for which inferences will be made about.
  - Parameters are numbers that characterize a population
- **Sample:**
  - A portion of the population that is representative of the entire population.
  - Statistics are numbers that summarize the data collected from a sample of the population
  - Sample must be an unbiased, random sample from the entire population.

Data sources include:

- Surveys or polls – sampling & bias
- Experiments - For example, when studying the effects of a new drug, a double blind study is usually used
- Observational and other studies – when unethical to get experimental data
- Operational database
- Data warehouse or Historical data base
- Purchased data base

Data sets :

- contain data objects which represents an entity
- Data object is described by data fields called attributes
- Attributes are also called as dimension, feature , variable
- Values for a given attribute is observation
- A set of attribute values are also called as attribute vector

VIN	Manufacturer	Weight	Number of cylinders	Fuel efficiency
IM8GD9A_KP042788	Ford	2984	6	20
IC4GE9A_DQ1572481	Toyota	1795	4	34

Data types

- **Constant:** A variable where every data value is the same.
  - Example - PI
- **Dichotomous or Binary:** A variable where there are only two values
  - Example- Gender whose values can be male or female.
- **Discrete, Categoric or Nominal:** A variable that can only take a certain number of values (either text or numbers).
  - Example- the variable Color where values could be black, blue, red etc.
  - Ordinal Scale – low, medium, high, grades – A+,A,B,C,D,E,F
- **Continuous or Numeric:** A variable where an infinite number of numeric values are possible within a specific range.
  - Example- temperature between the minimum and maximum temperature, the variable could take any value

Table 3.9. Table of patient records

Name	Age	Gender	Blood group	Weight (kg)	Height (m)	Systolic blood pressure	Diastolic blood pressure	Temperature (°F)	Diabetes
P. Lee	35	Female	A Rh^+	50	1.52	68	112	98.7	0
R. Jones	52	Male	O Rh^-	115	1.77	110	154	98.5	1
J. Smith	45	Male	O Rh^+	96	1.83	88	136	98.8	0
A. Patel	70	Female	O Rh^-	41	1.55	76	125	98.6	0
M. Owen	24	Male	A Rh^-	79	1.82	65	105	98.7	0
S. Green	43	Male	O Rh^-	109	1.89	114	159	98.9	1
N. Cook	68	Male	A Rh^+	73	1.76	108	136	99.0	0
W. Hands	77	Female	O Rh^-	104	1.71	107	145	98.3	1
P. Rice	45	Female	O Rh^+	64	1.74	101	132	98.6	0
F. Marsh	28	Male	O Rh^+	136	1.78	121	165	98.7	1

Descriptive Statistics

Measuring Central Tendency

- **Arithmetic Mean**

  Let $x_1, x_2, x_3 \ldots x_n$ be a set of N values or observations. The mean of this set of values is

  $(x_1 + x_2 + x_3 \ldots x_n) / n$

  avg() function in SQL

- **Weighted Arithmetic Mean**

  $(w_1 x_1 + w_2 x_2 + w_3 x_3 \ldots w_n x_n) / n$

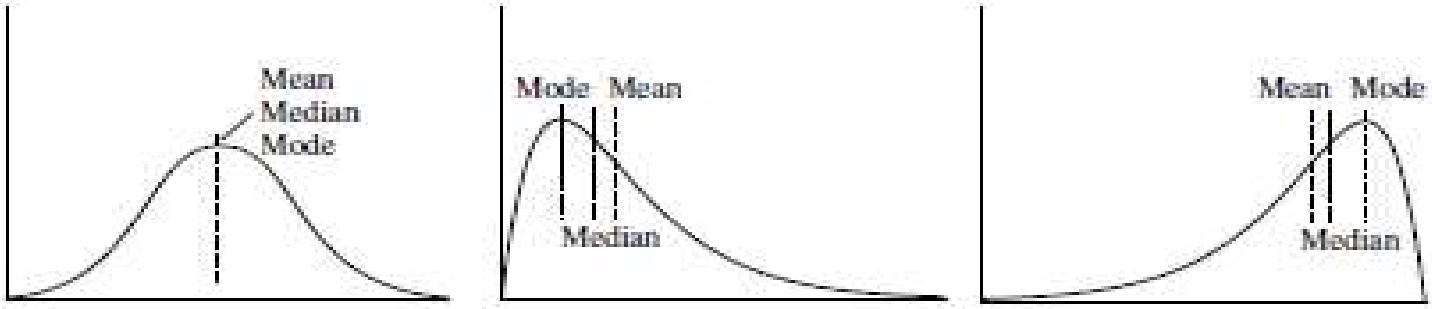
  - The mean of grouped data is obtained from the same formula by replacing the weight w_i with frequency f_i
  - Weight represents importance, occurrence frequency or significance

- **Trimmed mean**
  - remove top and bottom 2 %
- All are algebraic measures
- Advantages -it is easy to compute
- Disadvantage- it is sensitive to outlier or extreme values.

Descriptive Statistics
Measuring Central Tendency

- **Median**
  - Sort the N numbers .
  - If N is odd the middle number is median ie. median is the element at (n+1)/2 th position
  - If N is even the average of middle two values is median ie. Median is the arithmetic mean or average of the element at n/2 th position and n/2+1 position
  - Median is a holistic measure
  - **Advantages of Median**
    - divides the sample such that 50 % of the items are below it
    - is not influenced by extreme observations
    - is better than mean for skewed data
  - **Disadvantage**
    - is computational complexity and nature of distribution must be known
- **Mode**
  - for a set of data is the value that occurs most frequently in the set.
  - It is possible for the greatest frequency to correspond to several values which result in more than one mode – Multimodal
  - For unimodal frequency curves that are moderately skewed we have the following empirical relation
  - mean – mode = 3 * (mean – median)
  - Advantages
    - is not influenced by outliers and it coincides with sample observation
  - Disadvantage
    - need not be unique

Descriptive Statistics
Measuring Central Tendency

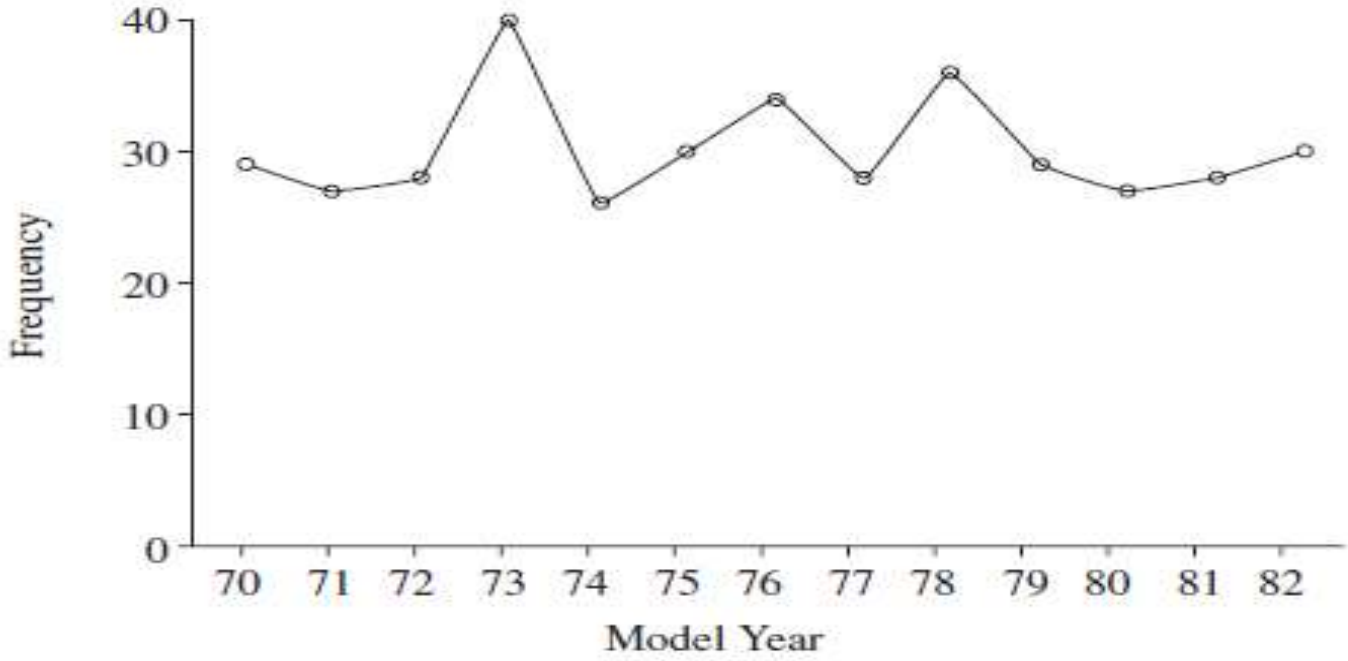


Mean, median, and mode of symmetric versus positively and negatively skewed data.

- **Midrange**
  - is the average of the largest and smallest values in the set.
  - In algebraic terms can be calculated using max() or min()

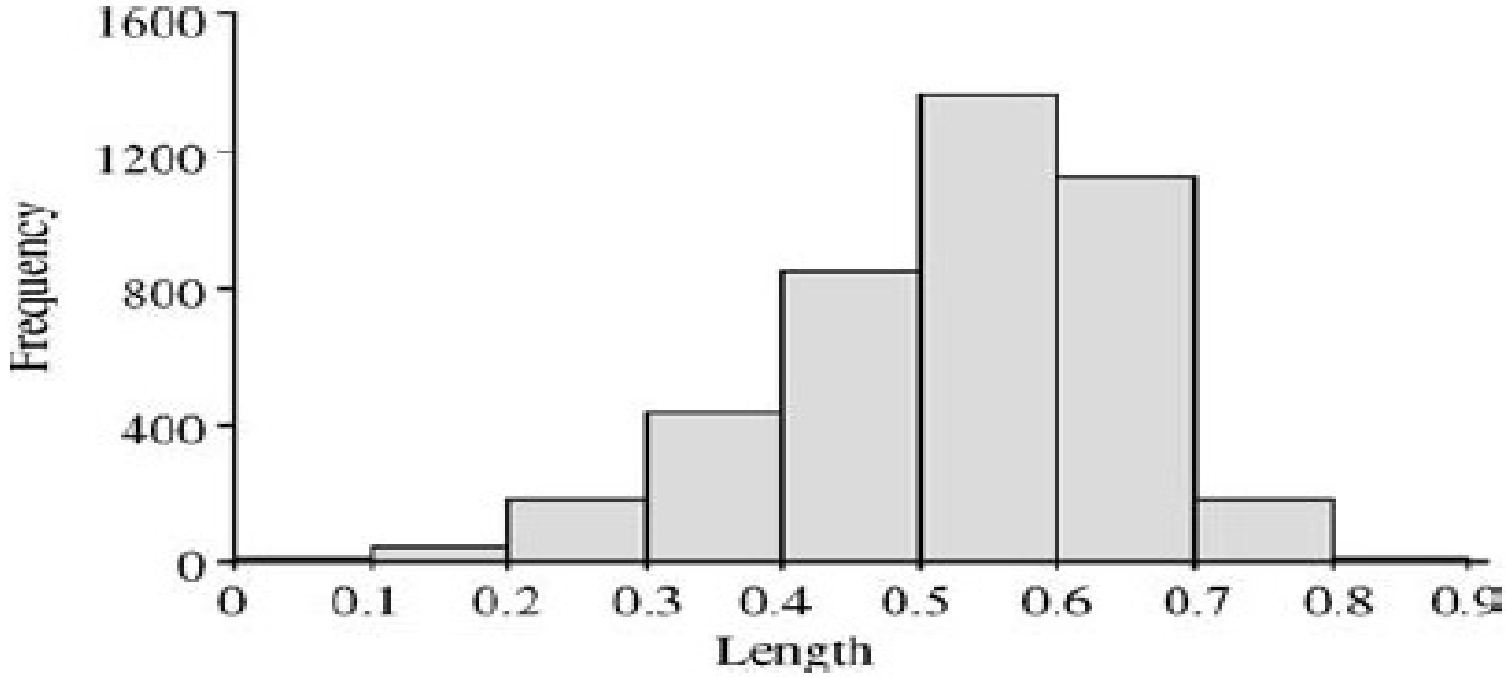
Descriptive Statistics
Visualizing Central Tendency

- Graphs enable us to visually identify trends, ranges, frequency distributions, relationships, outliers and make comparisons
- **Frequency Polygons**
  - plot information according to the number of observations for each value (or ranges of values) for a particular variable



Descriptive Statistics
Visualizing Central Tendency

- Graphs enable us to visually identify trends, ranges, frequency distributions, relationships, outliers and make comparisons
- **Histogram**
  - Useful for displaying the frequency distribution



Descriptive Statistics
Visualizing Central Tendency

- Graphs enable us to visually identify trends, ranges, frequency distributions, relationships, outliers and make comparisons
- **Histogram**
  - Useful for identifying outliers

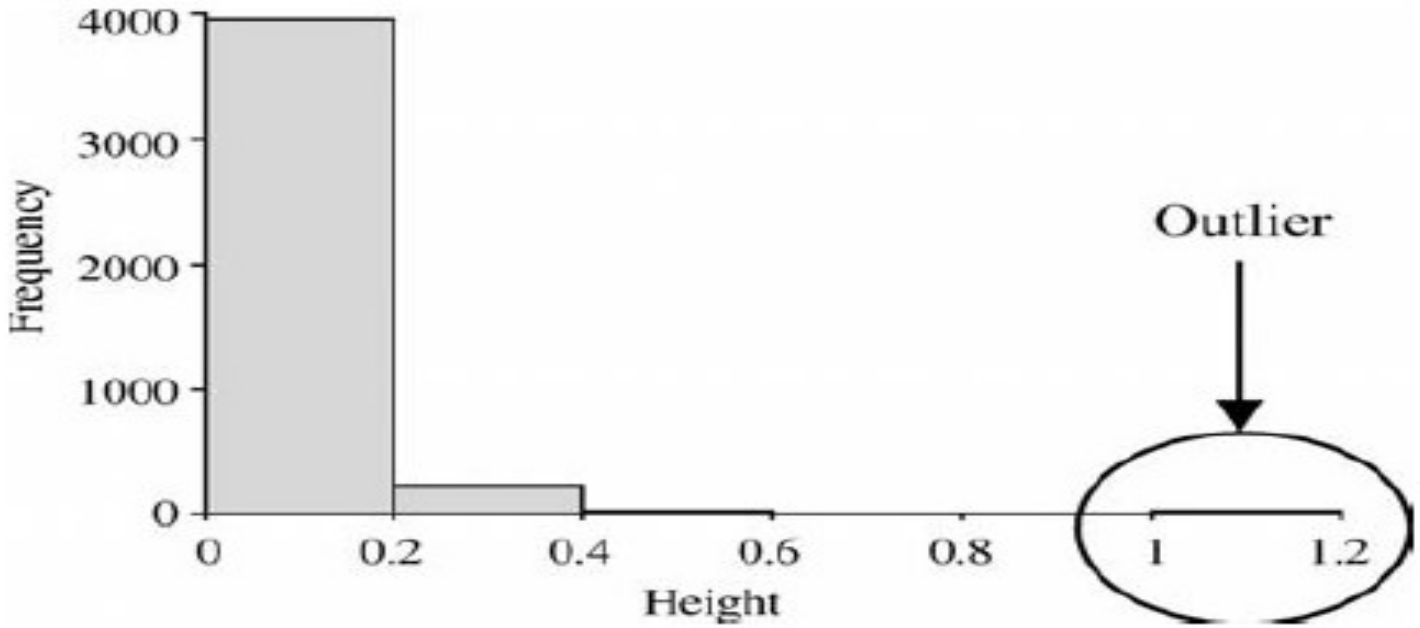


Table 3.9. Table of patient records

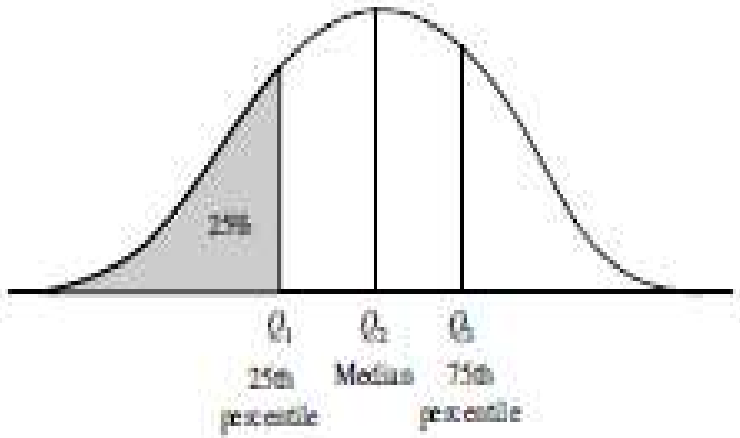
Name	Age	Gender	Blood group	Weight (kg)	Height (m)	Systolic blood pressure	Diastolic blood pressure	Temperature (°F)	Diabetes
P. Lee	35	Female	A Rh^+	50	1.52	68	112	98.7	0
R. Jones	52	Male	O Rh^-	115	1.77	110	154	98.5	1
J. Smith	45	Male	O Rh^+	96	1.83	88	136	98.8	0
A. Patel	70	Female	O Rh^-	41	1.55	76	125	98.6	0
M. Owen	24	Male	A Rh^-	79	1.82	65	105	98.7	0
S. Green	43	Male	O Rh^-	109	1.89	114	159	98.9	1
N. Cook	68	Male	A Rh^+	73	1.76	108	136	99.0	0
W. Hands	77	Female	O Rh^-	104	1.71	107	145	98.3	1
P. Rice	45	Female	O Rh^+	64	1.74	101	132	98.6	0
F. Marsh	28	Male	O Rh^+	136	1.78	121	165	98.7	1

Exploratory Analysis

- How many persons have diabetes?
- What is the average weight, average height?
- Tabulate the gender attribute
- What is the average , median, mode weight in males?
- What is the average , median, mode weight in females?
- Tabulate the BMI for Male vs Female.

Descriptive Statistics
Measuring Dispersion of Data

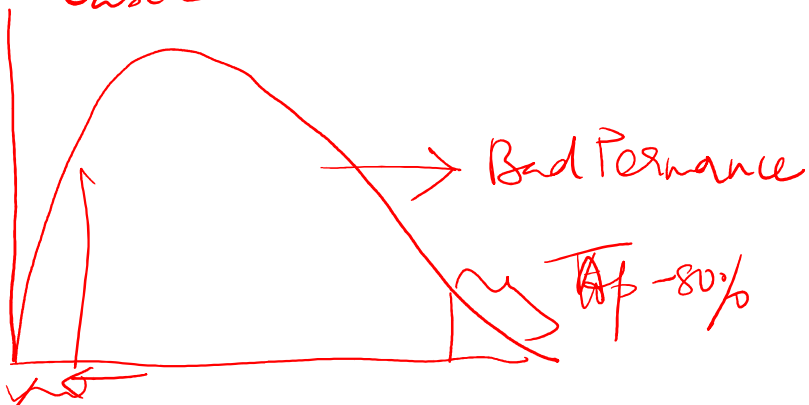
- Degree to which numerical data tend to spread is called as dispersion or variance of data.
- Common measures include
  - Range
  - five-number summary based on quartiles,
  - inter quartile range
  - standard deviation.



⇒ Relative Grading

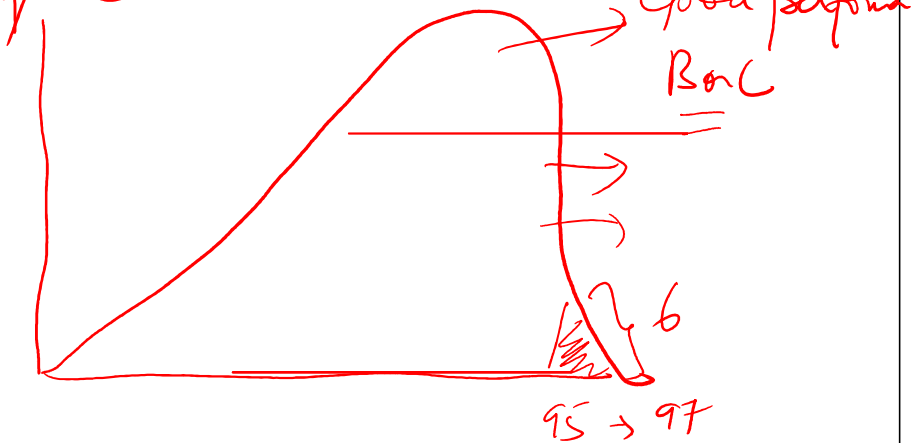
Case 1

Case 2



n = 60

Case 3



- 10% ✓ → 90th percentile
- 25% ✓ → Q_3

kth percentile

A+ - 90th percentile
A - 80th percentile
B - 70th
C - 60th
D - 50th
E - 40th

10th for Fail.

Descriptive Statistics
Measuring Dispersion of Data

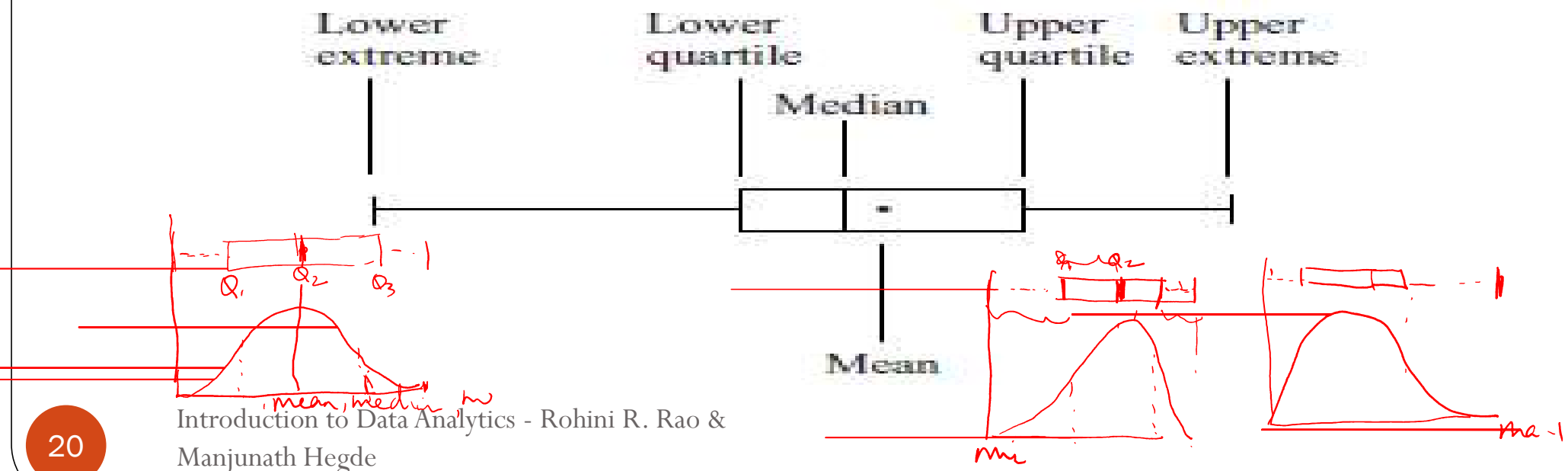
- **Range**
  - Let $x_1, x_2, \ldots x_n$ be a set of observations for some attribute. The range of the set is the difference between the max() and the min() values.
  - Easy to compute but sensitive to outliers
  - Coefficient of Range = (Max – Min)/(Max + Min)
- **The kth percentile** of a set of data in numerical order is the value xi having the property that k percent of the data lie at or below xi.
  - Median is 50^{th} percentile.
  - Quartile (Q_1) is 25^{th} percentile. First Quartile is the value below which one-forth of the observations will fall
  - **Quartile (Q_3) is 75^{th} percentile**. Third quartile is the value below which three-forth of the observations will fall
  - **Inter quartile range** - The distance between the first and third quartile.
  - IQR = $Q_3 - Q_1$
  - **Quartile Deviation** is one half of IQR
  - IQR is unaffected by outliers and provides supplementary information of the spread of observations around the centre of the sample.
- A common rule of thumb to identify suspected **outliers** is to single out values falling at least 1.5 * IQR above the third quartile or below the first quartile.
- The five-number summary of a distribution consists of the
  - Minimum, Q_1, Median, Q_3 , Maximum.
- **Box plots** are used to visualize them.

→ Lower Bound = $1.5 * IQR - Q_1$

→ Upper Bound = $1.5 * IQR + Q_3$

Descriptive Statistics
Visualizing Dispersion of Data

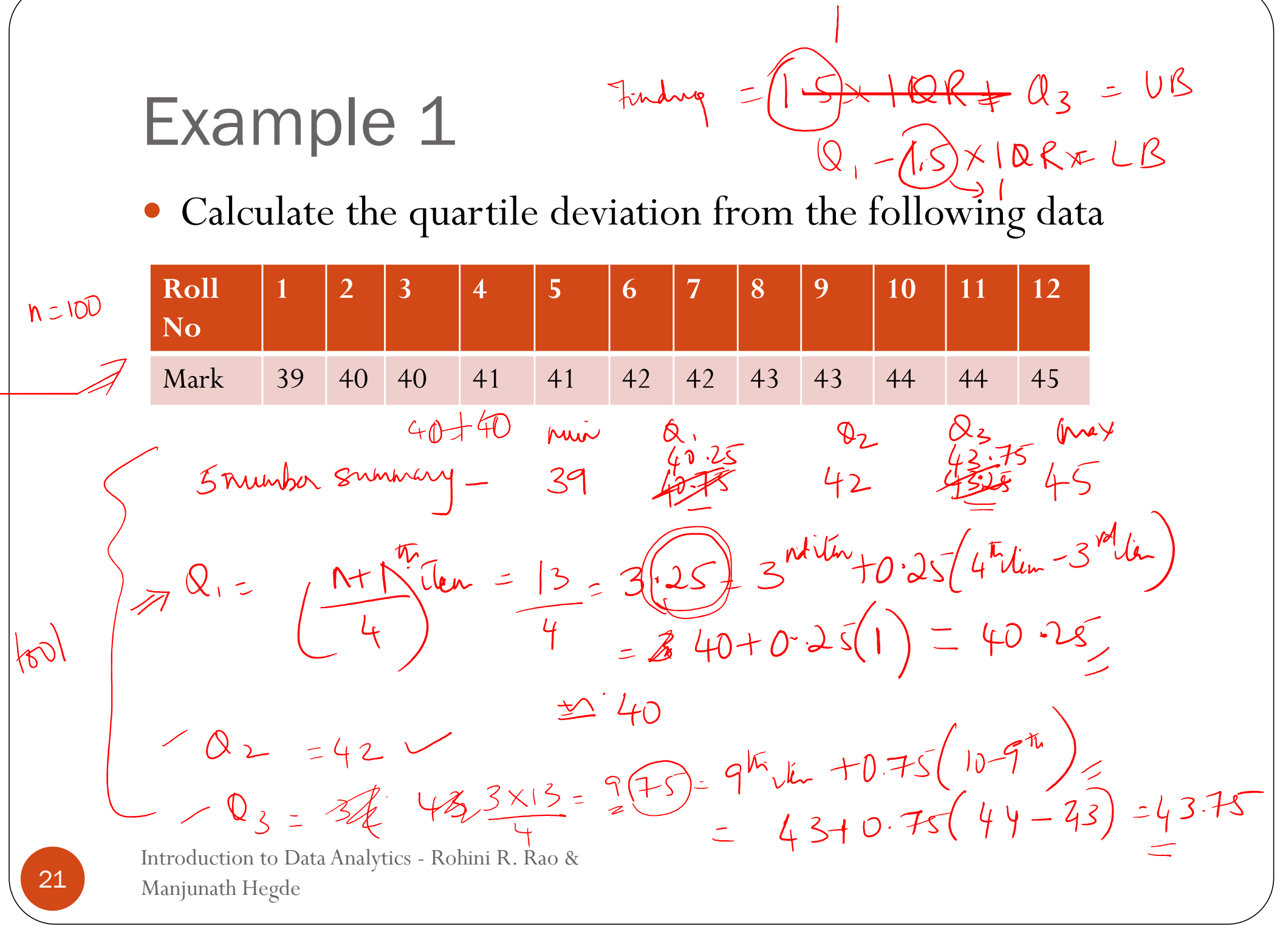
- **Box Plot**
  - provide a succinct summary of the overall distribution for a variable.
  - Five points are displayed:
    - the lower extreme value, the lower quartile, the median, the upper quartile, the upper extreme and the mean



Example 1

- Calculate the quartile deviation from the following data

Roll No	1	2	3	4	5	6	7	8	9	10	11	12
Mark	39	40	40	41	41	42	42	43	43	44	44	45



UB = 1.5*IQR + Q3 = 177 ≈ 100

Example 2

- Find 5 number summary
- Compare the performance of Male vs Female

(6M, 6F)

(10)

Sort Item wise ⇒

Roll No	1	2	3	4	5	6	7	8	9	10	11	12
Mark	23	25	73	65	43	65	89	35	58	100	77	87
Gender	M	F	M	M	F	F	F	M	F	M	F	M

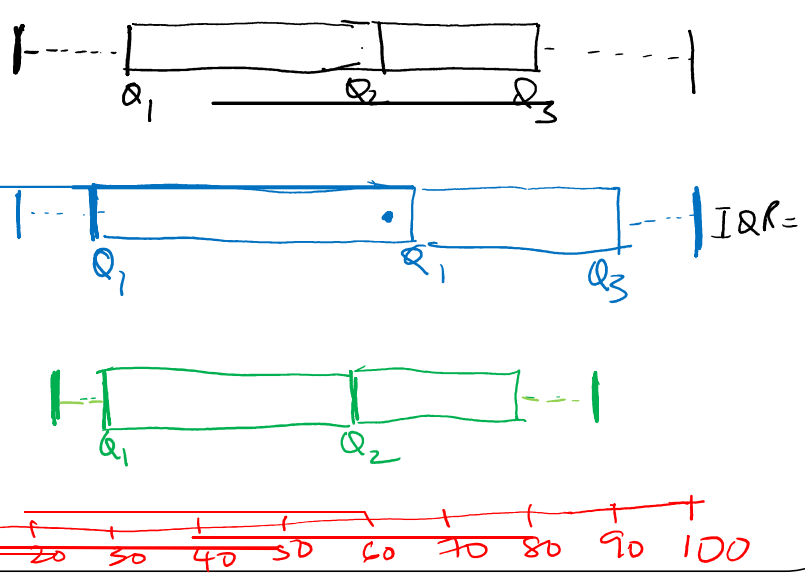
	Min	Q_1	Q_2	Q_3	Max	AVG
Overall	23	37	65	84.5	100	61.67
Male	23	32	69	90.25	100	63.83
Female	25	29.5	61.5	80	89	59.5

OVERALL

MALE

FEMALE

Upper Bound = 1.5*IQR + Q3 = 177 ≈ 100



Example 2

OVERALL - mean ± SD = 61.67 ±

MALE - mean ± SD = 63.83 ±

FEMALE - ~~mean ± SD~~ = 59.5 ±

Descriptive Statistics
Measuring Dispersion of Data

- **Variance**
  - measure of the deviation of a variable from the mean.
  - The sample variance formula is

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

  - The population variance is defined as

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

Descriptive Statistics
Measuring Dispersion of Data

- **Standard Deviation**
  - Also referred to as root mean square

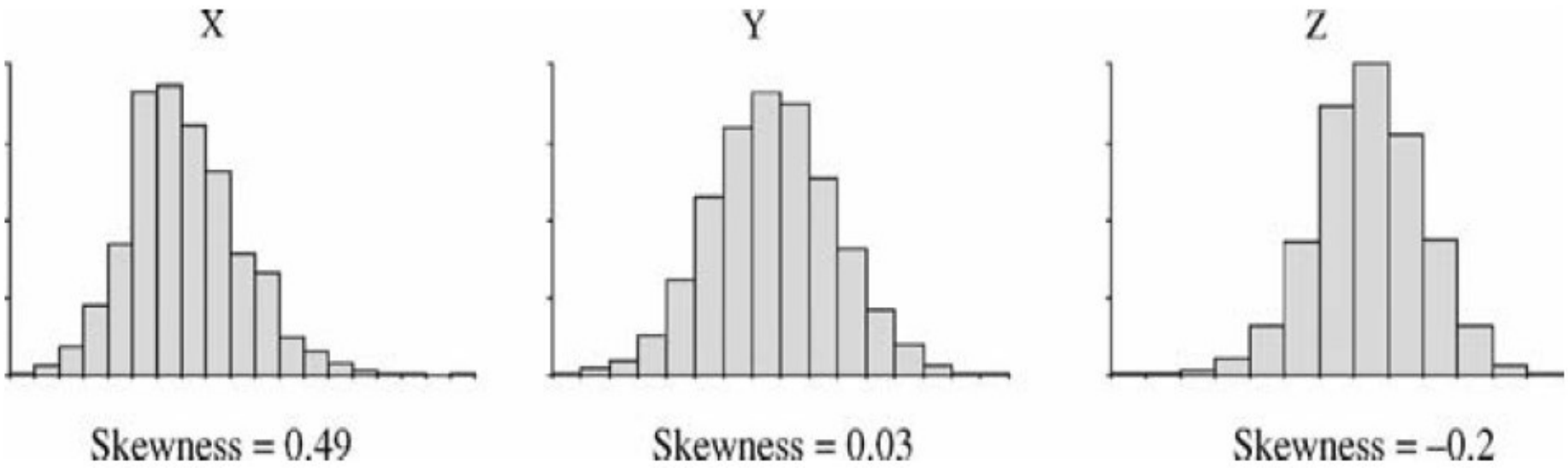
$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

  - If the frequency distribution is approximately normal, about 68% of all observations will fall within one standard deviation of the mean (34% less than and 34% greater than).
  - Approximately 95% of all observations fall within two standard deviations of the mean

Descriptive Statistics
Measuring Dispersion of Data

- **Skewness**
  - Quantifies the lack of symmetry or skewness in the distribution

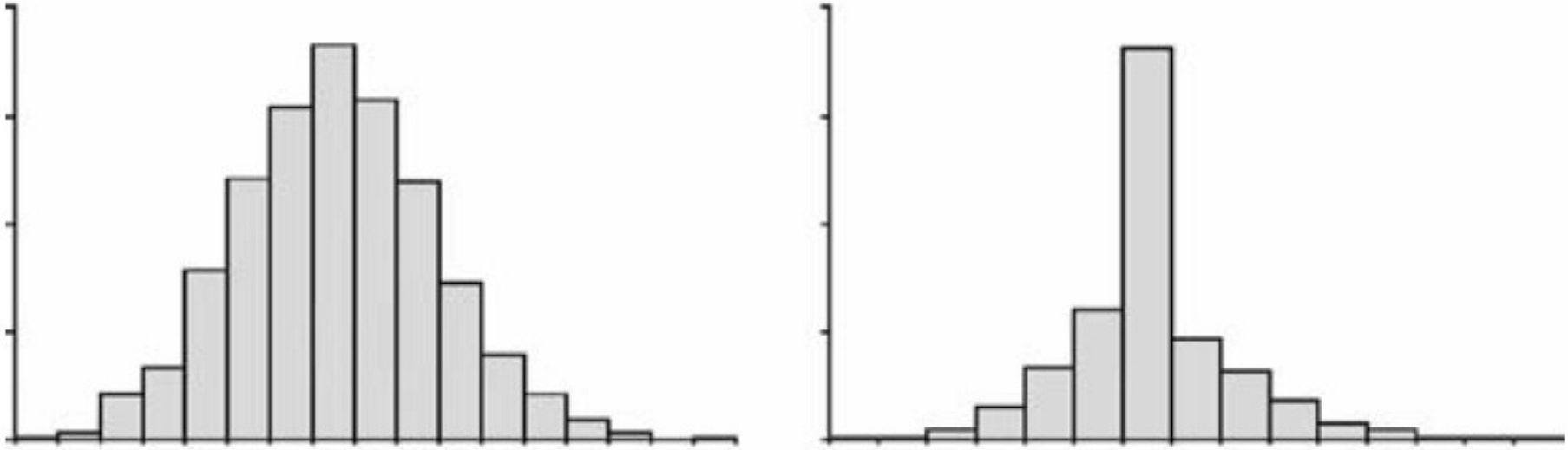
$$\text{skewness} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{(n-1)s^3}$$



Descriptive Statistics
Measuring Dispersion of Data

- **Kurtosis**
  - Measures the peak of the distribution

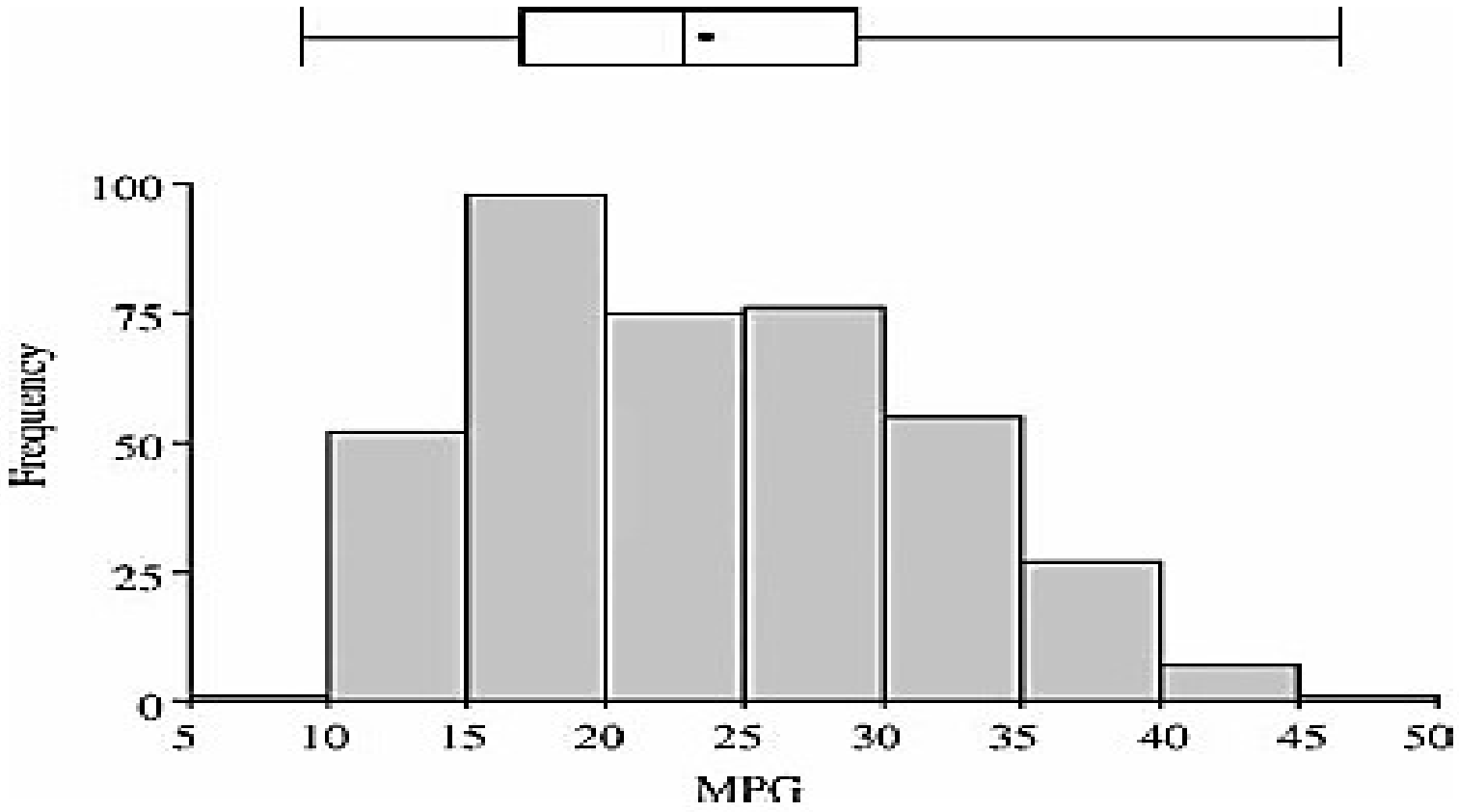
$$\text{kurtosis} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{(n-1)s^4}$$



Kurtosis = 0.08

Kurtosis = 1.97

Descriptive Statistics
Multiple Graphs



Example 3

- The following are scores made by 2 cricketers in 10 innings. Compare the 2 players

Innings	1	2	3	4	5	6	7	8	9	10
Cricketer A	31	48	13	51	38	43	50	36	47	82
Cricketer B	51	5	12	83	37	112	42	18	79	0

$\overline{\text{Cricketer A}}$ (Avg) = 43.9 ± SD =

$\overline{\text{Cricketer B}}$ (Avg) = 43.9 ± SD =

Coeff of Variance =

Case study 2: Breakfast Cereals Data Set

- **Abstract:**
  - Adults should consume between 20 and 35 grams of dietary fiber per day.
  - The recommended daily intake for calories is 2200 for women and 2900 for men.
  - Calories come in three food components. There are 9 calories per gram of fat, and 4 calories per gram of carbohydrate and protein.
  - Overall, in your diet, no more than 10% of your calories should be consumed from simple carbohydrates (sugars), and no more than 30% should come from fat. The RDI of protein is 50 grams for women and 63 grams for men. The balance of calories should be consumed in the form of complex carbohydrates (starches).
  - The average adult with no defined risk factors or other dietary restrictions should consume between 1800 and 2400 mg of sodium per day.
  - The type and amount of milk added to cereal can make a significant difference in the fat and protein content of your breakfast.

Types of data analysis

- What are the best cereals -gender wise?
- Is there a link between calories & shelves?
- Are the best rated cereals really healthy?
- What ingredients best predict the amount of calories a cereal contains?
- Does serving size play a role in the amount of calories?

An experiment or project idea

- *is to develop one's own rating system and find which cereal is the most healthy for an individual*
- *To make cereal recommendations for individuals based on requirements*

Descriptive Statistics
Contingency tables

- Also referred to as two-way cross tables
- Provides insight into the relationship between two variables.
- Variables must be categorical (dichotomous or discrete), or transformed to a categorical variable.

	Sex = Male	Sex = Female	Totals
Age (10.0 to 20.0)	847	810	1,657
Age (20.0 to 30.0)	4,878	3,176	8,054
Age (30.0 to 40.0)	6,037	2,576	8,613
Age (40.0 to 50.0)	5,014	2,161	7,175
Age (50.0 to 60.0)	3,191	1,227	4,418
Age (60.0 to 70.0)	1,403	612	2,015
Age (70.0 to 80.0)	337	171	508
Age (80.0 to 90.0)	54	24	78
Age (90.0 to 100.0)	29	14	43
Totals	21,790	10,771	32,561

Descriptive Statistics

Contingency table - (Miles per gallon vs. No. of Cylinders)

	Cylinders = 3	Cylinders = 4	Cylinders = 5	Cylinders = 6	Cylinders = 8	Totals
MPG (5.0 to 10.0)	0	0	0	0	1	1
MPG (10.0 to 15.0)	0	0	0	0	52	52
MPG (15.0 to 20.0)	2	4	0	47	45	98
MPG (20.0 to 25.0)	2	39	1	29	4	75
MPG (25.0 to 30.0)	0	70	1	4	1	76
MPG (30.0 to 35.0)	0	53	0	2	0	55
MPG (35.0 to 40.0)	0	25	1	1	0	27
MPG (40.0 to 45.0)	0	7	0	0	0	7
MPG (45.0 to 50.0)	0	1	0	0	0	1
Totals	4	199	3	83	103	392

Descriptive Statistics
Summary tables

- A single categorical variable is used to group the observations.
- Each row of the table represents a single group.
- Descriptive statistics that summarize a set of observations can be used
  - statistics are commonly used:
    - Mean: The average value.
    - Median: The value at the mid-point.
    - Sum: The sum over all observations in the group.
    - Minimum: The minimum value.
    - Maximum: The maximum value.
    - Standard deviation: A standardized measure of the deviation of a variable from the mean.
- Summary tables often show a count, number of observations (or percentage) that have that particular value (or range).

Example 1.

Table 4.8. Retail transaction data set

Customer	Store	Product category	Product description	Sale price ($)	Profit ($)
B.March	New York, NY	Laptop	DR2984	950	190
B.March	New York, NY	Printer	FW288	350	105
B.March	New York, NY	Scanner	BW9338	400	100
J.Bain	New York, NY	Scanner	BW9443	500	125
T.Goss	Washington, DC	Printer	FW199	200	60
T.Goss	Washington, DC	Scanner	BW39339	550	140
L.Nye	New York, NY	Desktop	LR21	600	60
L.Nye	New York, NY	Printer	FW299	300	90
S.Cann	Washington, DC	Desktop	LR21	600	60
E.Sims	Washington, DC	Laptop	DR2983	700	140
P.Judd	New York, NY	Desktop	LR22	700	70
P.Judd	New York, NY	Scanner	FJ3999	200	50
G.Hinton	Washington, DC	Laptop	DR2983	700	140
G.Hinton	Washington, DC	Desktop	LR21	600	60
G.Hinton	Washington, DC	Printer	FW288	350	105
G.Hinton	Washington, DC	Scanner	BW9443	500	125
H.Fu	New York, NY	Desktop	ZX88	450	45
H.Taylor	New York, NY	Scanner	BW9338	400	100

Example 1

1. Generate a contingency table summarizing the variables Store and product category
2. Generate the following summary tables:

a. Grouping by Customer and the showing a count of the number of observations and the sum of Sale price ($) for each row.

b. Grouping by Store and showing a count of the number of observations and the mean Sale price ($) for each row.

c. Grouping by Product category and showing a count of the number of observations and the sum of the Profit ($) for each row.

3. Create a histogram of Sales Price ($) using the following intervals: 0 to less than 250, 250 to less than 500, 500 to less than 750, 750 to less than 1000.
4. Create a scatterplot showing Sales price ($) against Profit ($).

① Store / Product Category

	NY	DC	Row Total
Laptop	1	2	3
Printer	2	2	4
Scanner	4	3	7
Desktop	3	1	4
Column Total	10	8	18

③

Store	Count	Mean(Sales Price)
NY	10	485
DC	8	528

Customer Wise Summary

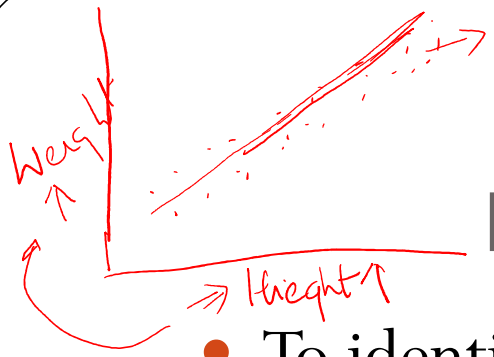
March	3	1700
Bain	1	500
Goss	2	750
Nye	2	900
Conn	1	600
Symms	1	700
Todd	2	900
Plinton	4	2150
Fry	1	450
Taylor	1	400

Product	Count	Sum (Profit)	Mean (Profit
Laptop	3	470	
Printer	4	360	
Scanner	7	640	
Desktop	4	295	

Example 2- Data Visualization

- MS Excel – Northwind Dataset

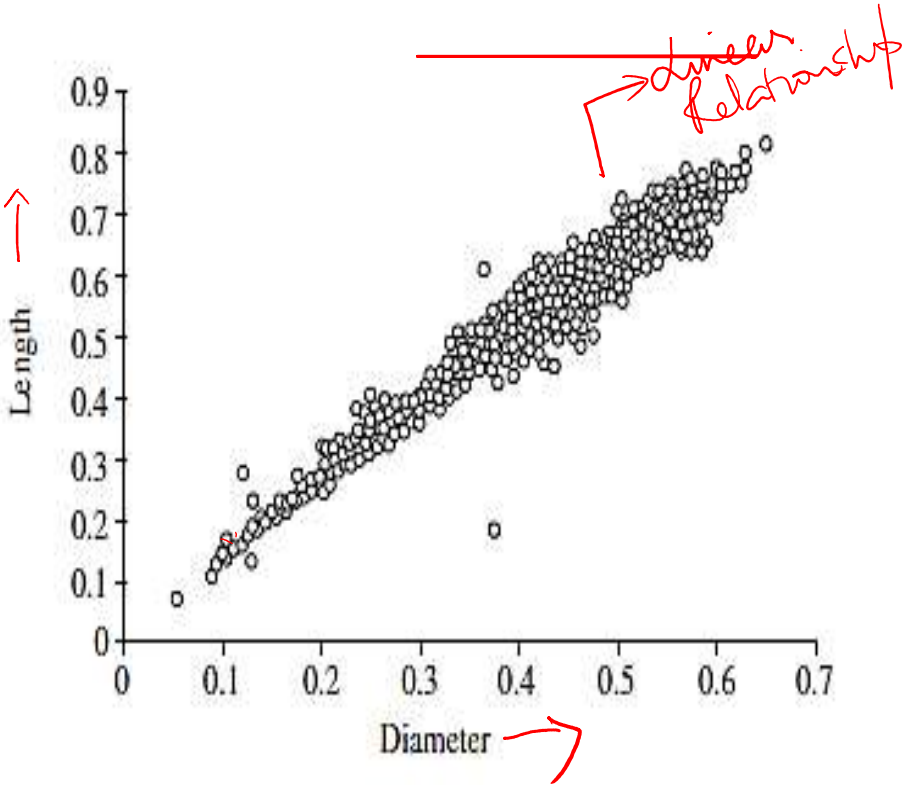
1. What is the total sales/sales person in categories of products?
2. What is the total sales for each company in categories of products?
3. What are the top 10 selling items
4. Visualize the Customer's region wise sales of items in each category
5. What is the sale of Boston Crab meat, customer wise?
6. Summarise % sales for each category of products yearwise
7. Report and analyse shipping country and category of products



Descriptive Statistics

Bivariate Analysis - Scatter Plots

- To identify relationships between two continuous variables
- Can be used to understand the type of relationships



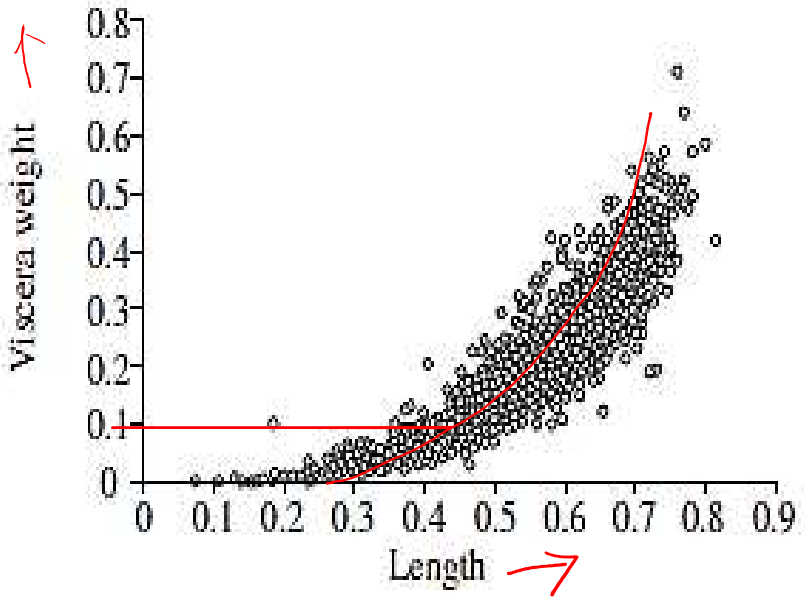
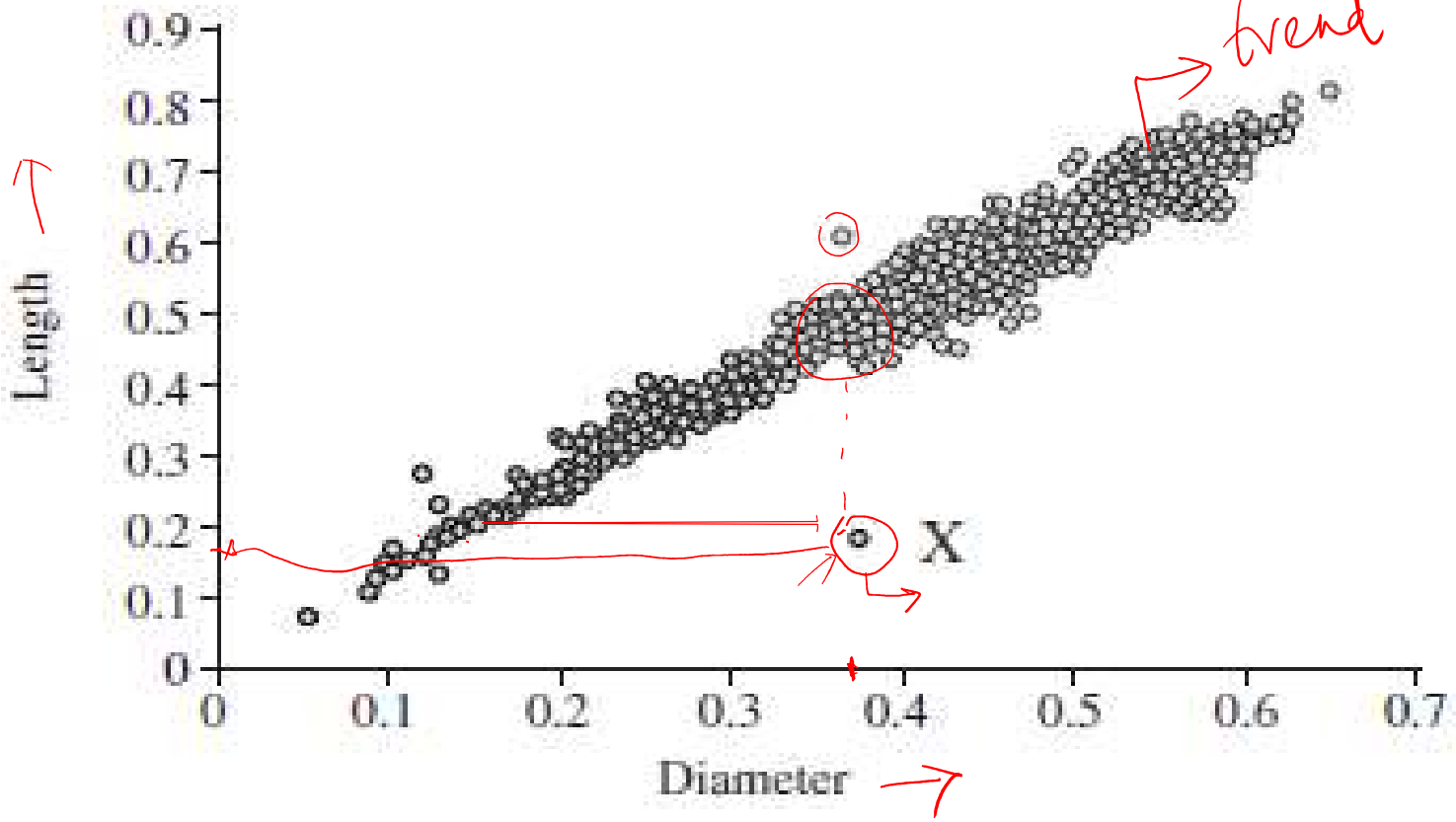


Figure 4.8. Scatterplot showing a nonlinear relationship

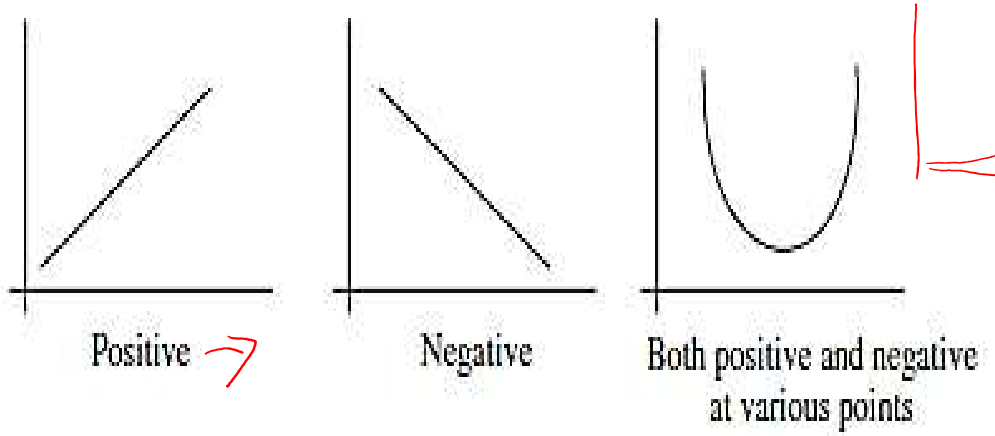
Descriptive Statistics
Bivariate Analysis - Scatter Plots

- To identify relationships between two continuous variables
- Can be used to find outliers



Comparative Analysis Correlation

- Correlation analysis
  - looks at associations between variables.
  - existence of an association between variables does not imply that one variable causes another
- Characteristics of the relationship can be measured:
  - Direction
    - Positive relationship
    - Negative relationship
  - Shape
    - Could be liner or nonlinear



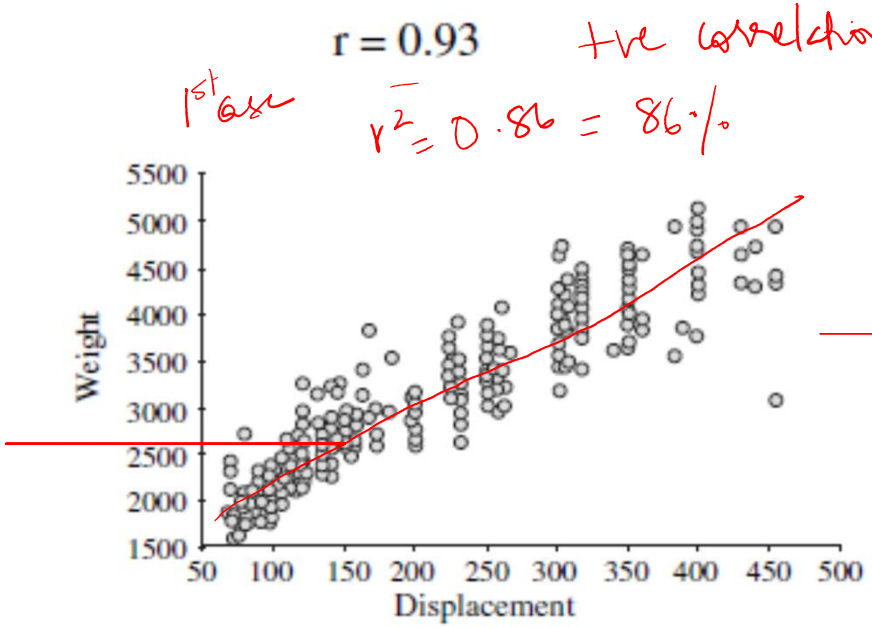
Pearson's

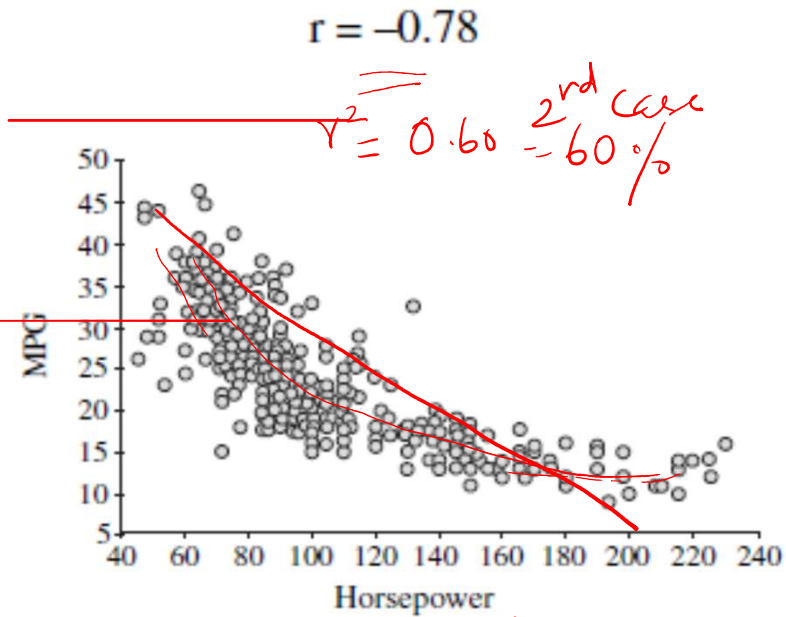
Comparative Analysis

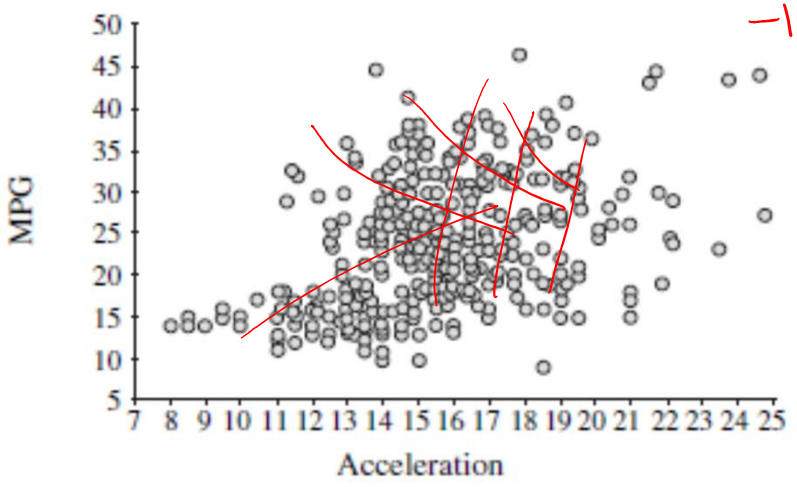
Correlation Coefficient for numeric variables

- Generates values ranging from -1.0 to +1.0
- The value of r reflects how close to straight line the points lie.
- Positive numbers indicate a positive correlation
- Negative numbers indicate a negative correlation.
- If r is around 0 then there appears to be little or no relationship between the variables

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$


r = 0.93
Weight
Displacement

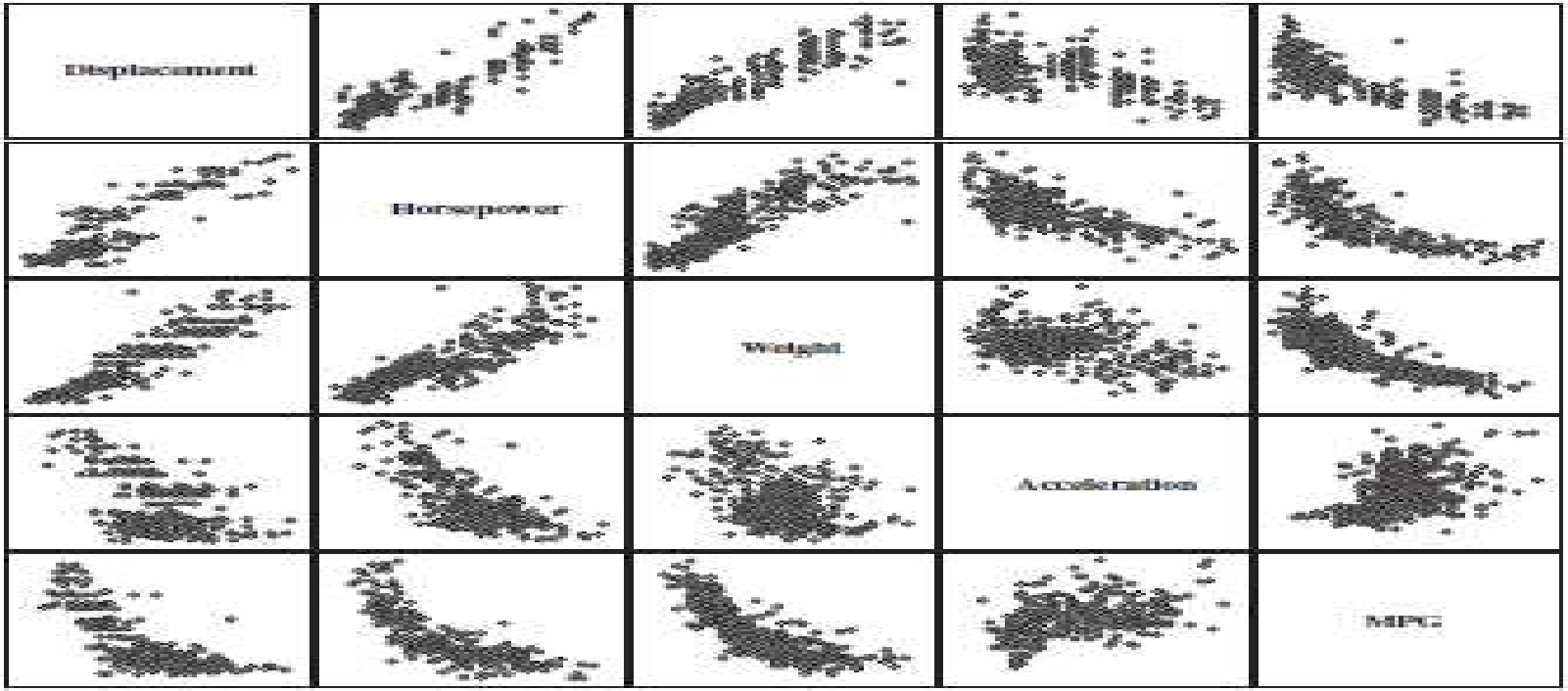

r = –0.78
MPG
Horsepower


MPG
Acceleration

Correlation Analysis
for more than two numeric variables

Table 5.13. Table displaying values for the correlation coefficient for five variables

	Displacement	Horsepower	Weight	Acceleration	MPG
Displacement	1	0.9	0.93	−0.54	−0.81
Horsepower	0.9	1	0.86	−0.69	−0.78
Weight	0.93	0.86	1	−0.42	−0.83
Acceleration	−0.54	−0.69	−0.42	1	0.42
MPG	−0.81	−0.78	−0.83	0.42	1



Correlation Analysis for more than two numeric variables

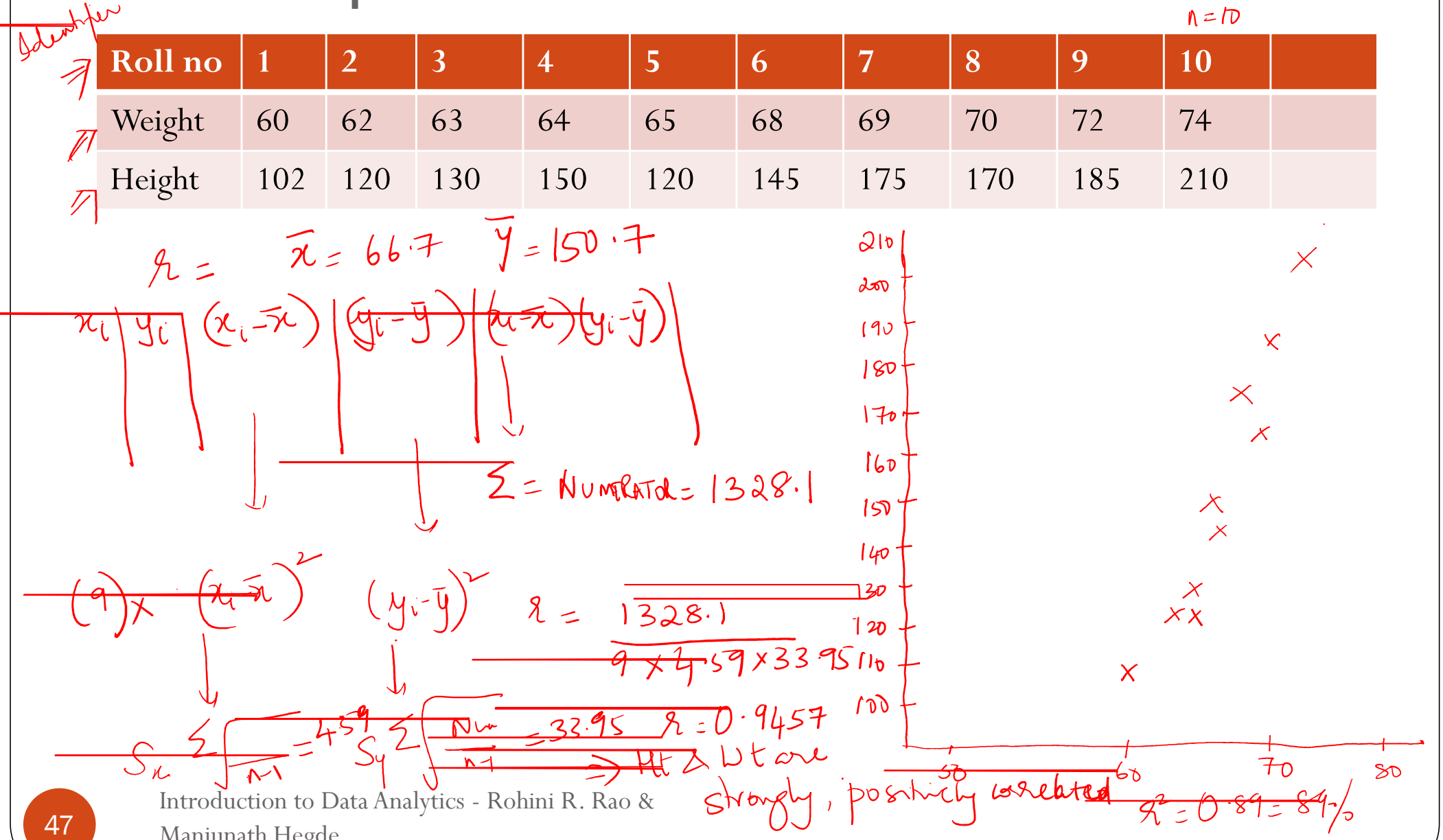
- Correlation Coefficient is squared to indicate percentage of variation that is explained by the regression line
- The coefficient of determination represents the percent of the data that is the closest to the line of best fit.

Table 5.14. Table displaying the value for r^2 for five variables

	Displacement	Horsepower	Weight	Acceleration	MPG
Displacement	1	0.81	0.87	0.29	0.66
Horsepower	0.81	1	0.74	0.48	0.61
Weight	0.87	0.74	1	0.18	0.69
Acceleration	0.29	0.48	0.18	1	0.18
MPG	0.66	0.61	0.69	0.18	1

Example 1 – Pearsons Correlation

Roll no	1	2	3	4	5	6	7	8	9	10	
Weight	60	62	63	64	65	68	69	70	72	74	
Height	102	120	130	150	120	145	175	170	185	210	



Example 2 – Pearsons Correlation

- Errors in test and study time

n	1	2	3	4	5	6	7	8	9	10
Study time (in mins)	90	100	130	150	180	200	220	300	350	400
No of errors	25	28	20	20	15	12	13	10	8	6

$n = 10$

$\bar{x} = 212$ mins

$\bar{y} = 15.7$ errors

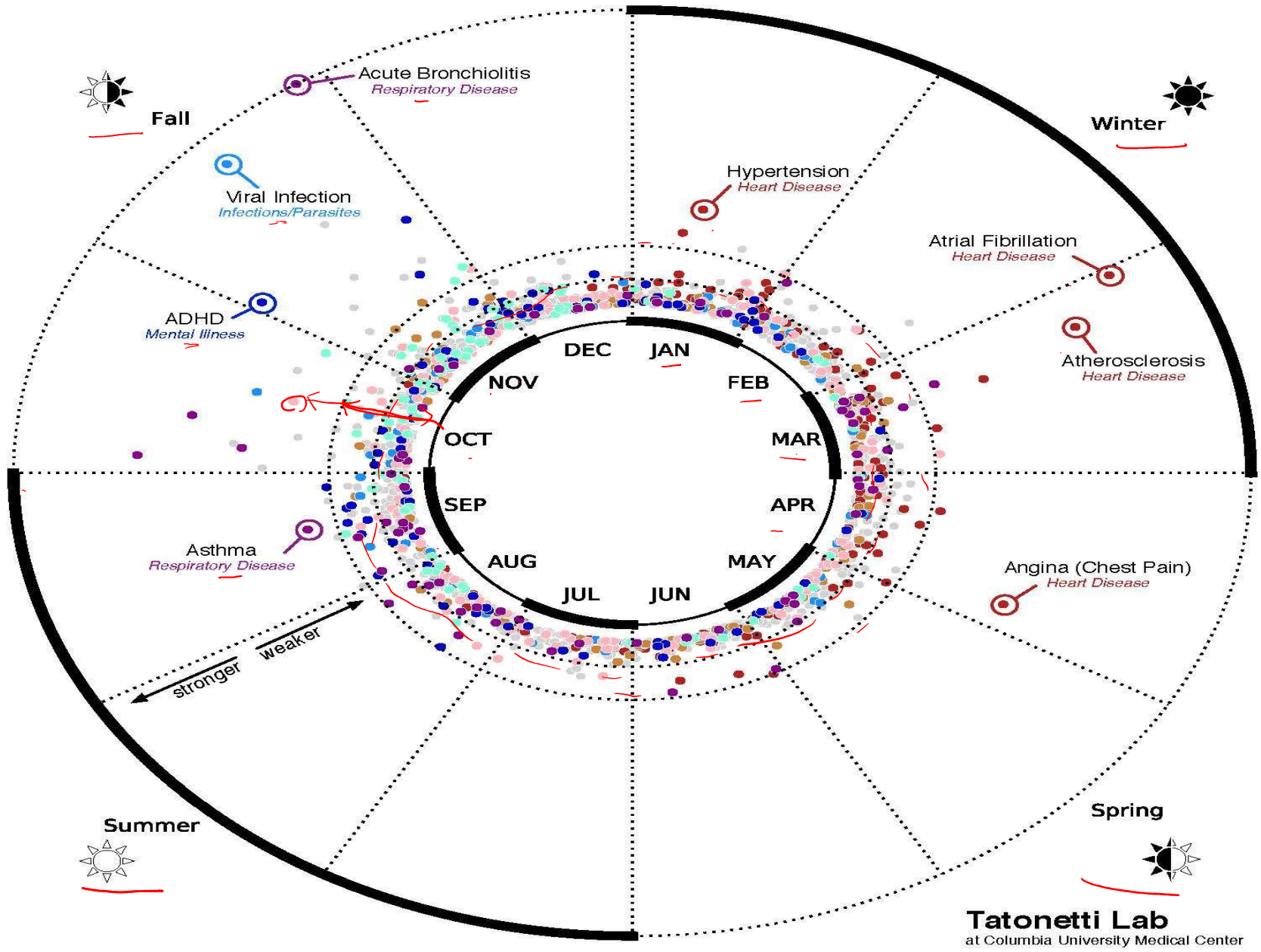
$r = \frac{\text{Numerator}}{9 \times S_x \times S_y} = \frac{-62.74}{9 \times 106.13 \times 7.32} = -0.926$

inference: strong negative correlation

$r^2 = 85\% = 0.85$

Case Study: Correlation between birth month & health

- Columbia University scientists found 55 diseases that correlated with the season of birth.
- compares 1,688 diseases against the birth dates and medical histories of 1.7 million patients treated at NewYork-Presbyterian Hospital/CUMC between 1985 and 2013.
- The study proved more than 1,600 associations and confirmed 39 links previously reported in the medical literature.
- The researchers performed statistical tests to check that the 55 diseases for which they found associations did not arise by chance.
- The risk related to birth month is relatively minor when compared to more influential variables like diet and exercise."
- For example
  - the study found that asthma risk is greatest for July and October babies. Danish study on the disease found that the peak risk was in the months (May and August).
  - For ADHD, around one in 675 occurrences could relate to being born in New York in November. This result matches a Swedish study showing peak rates of ADHD in November babies.
  - The researchers also found a relationship between birth month and nine types of heart disease, with people born in March facing the highest risk for atrial fibrillation, congestive heart failure, and mitral valve disorder. Austrian and Danish patient records found that those born in months with higher heart disease rates—March through June—had shorter life spans.
- Overall, the study indicated people born in May had the lowest disease risk, and those born in October the highest. This data could help scientists uncover new disease risk factors


Birth Month and Disease Incidence in 1.7 Million Patients
Fall
Winter
Spring
Summer
Acute Bronchiolitis
Respiratory Disease
Viral Infection
Infections/Parasites
ADHD
Mental Illness
Asthma
Respiratory Disease
Hypertension
Heart Disease
Atrial Fibrillation
Heart Disease
Atherosclerosis
Heart Disease
Angina (Chest Pain)
Heart Disease
DEC
JAN
FEB
MAR
APR
MAY
JUN
JUL
AUG
SEP
OCT
NOV
stronger
weaker
Tatonetti Lab
at Columbia University Medical Center

Comparative Analysis for discrete variables

- **Chi-square test**
  - is a hypothesis test for nominal variables
  - tests the hypothesis that *A* and *B* are *independent*, that is, there is no correlation between them
  - Hypothesis includes
    - Null hypothesis H_0: There is no relationship or they are independent
    - Alternate hypothesis H_a: There is a relationship or dependence
  - A contingency table or pivot table is constructed for variables
  - The chi-square test $$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$
  - Where expected frequency is $$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n},$$
  - The test is based on a significance level with (r-1) * (c-1) degrees of freedom

Example 1

- Are gender and preferred reading correlated?

	Male	Female	Total
Fiction	250 O_{11}	200 O_{12}	450
Non Fiction	50 O_{21}	1000 O_{22}	1050
Total	300	1200	

~~Step 1~~ H_0: Gender and ~~preferred reading~~ are independent / no association / no relationship

H_a: Gender and preferred reading are ~~dependent / associated~~ / CORRELATED

Example 1

- Are gender and preferred reading correlated?

	Male	Female	Total
Fiction	250 (e_{11} 90)	200 (e_{12} 360)	450
Non Fiction	50 (e_{21} 210)	1000 (e_{22} 840)	1050
Total	300	1200	(1500)

$n = 1500$

$e_{11} = \frac{\text{Count(male)} \times \text{Count(fiction)}}{n} = \frac{300 \times 450}{1500} =$

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

$\chi^2 = 93$

degrees of freedom (df) = (Rows − 1) × (cols − 1)

$= (2-1) \times (2-1) = 1$

$\chi^2 \geq \chi_c^2$ $\quad \chi_c^2 = 3.841$

Inference ⇒ Gender and preferred reading are associated/correlated

Example 2

	Male	Female	Totals
Supervisor	20 (11.67)	15 (23.34)	35
Shelf Stacker	20 (16.67)	30 (33.34)	50
Till Operator	10 (15)	35 (30)	45
Cleaner	10 (16.67)	40 (33.34)	50
	60	120	180

n = 18

H_0 : Type of jobs and Gender ~~are not associated~~

$\Rightarrow H_a$: Type of jobs and ~~Gender are associated or correlated~~

Age: 18 ← → 70

18-28 : young

48-38 : middle-old

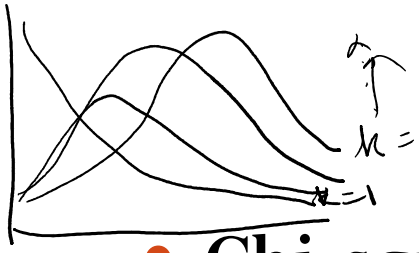
$\chi^2 = 16.422$ $\chi^2_c = 7.815$

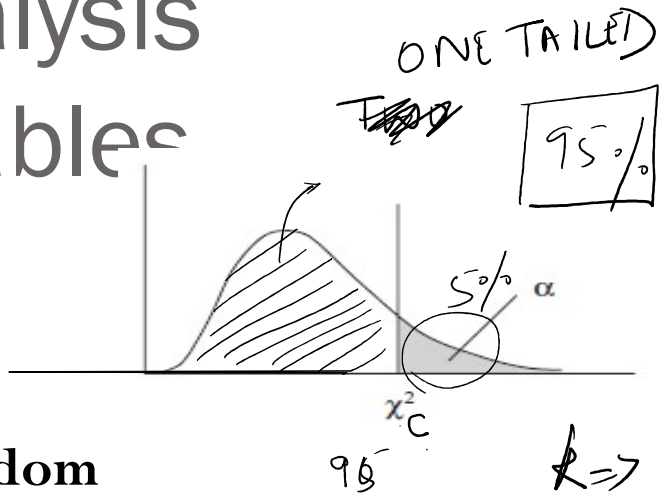
$df = (4-1) \times (2-1) = 3$ Inference => Type of job & Gender are correlated.

Example 2

	Male	Female	Totals
Supervisor	20	15	35
Shelf Stacker	20	30	50
Till Operator	10	35	45
Cleaner	10	40	50
	60	120	180

Comparative Analysis for discrete variables





- **Chi-square table –**
  - **Indicates distribution of sampling SD.**
  - **skewed distribution with degrees of freedom**

	Probability													
df	0.99	0.98	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.02	0.01	0.001
1	0.0^3157	0.0^3628	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.0201	0.0404	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210	13.815
3	0.115	0.185	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	0.297	0.429	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	0.554	0.752	1.145	1.610	2.343	3.000	4.351	6.064	7.289	9.236	11.070	13.388	15.086	20.515
6	0.872	1.134	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	1.239	1.564	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	1.646	2.032	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090	26.125
9	2.088	2.532	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	2.558	3.059	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	3.053	3.609	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	3.571	4.178	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	4.107	4.765	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	4.660	5.368	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	5.229	5.985	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697

Looking up critical chi-square value, for $df = 4$ and $\alpha = 0.05$

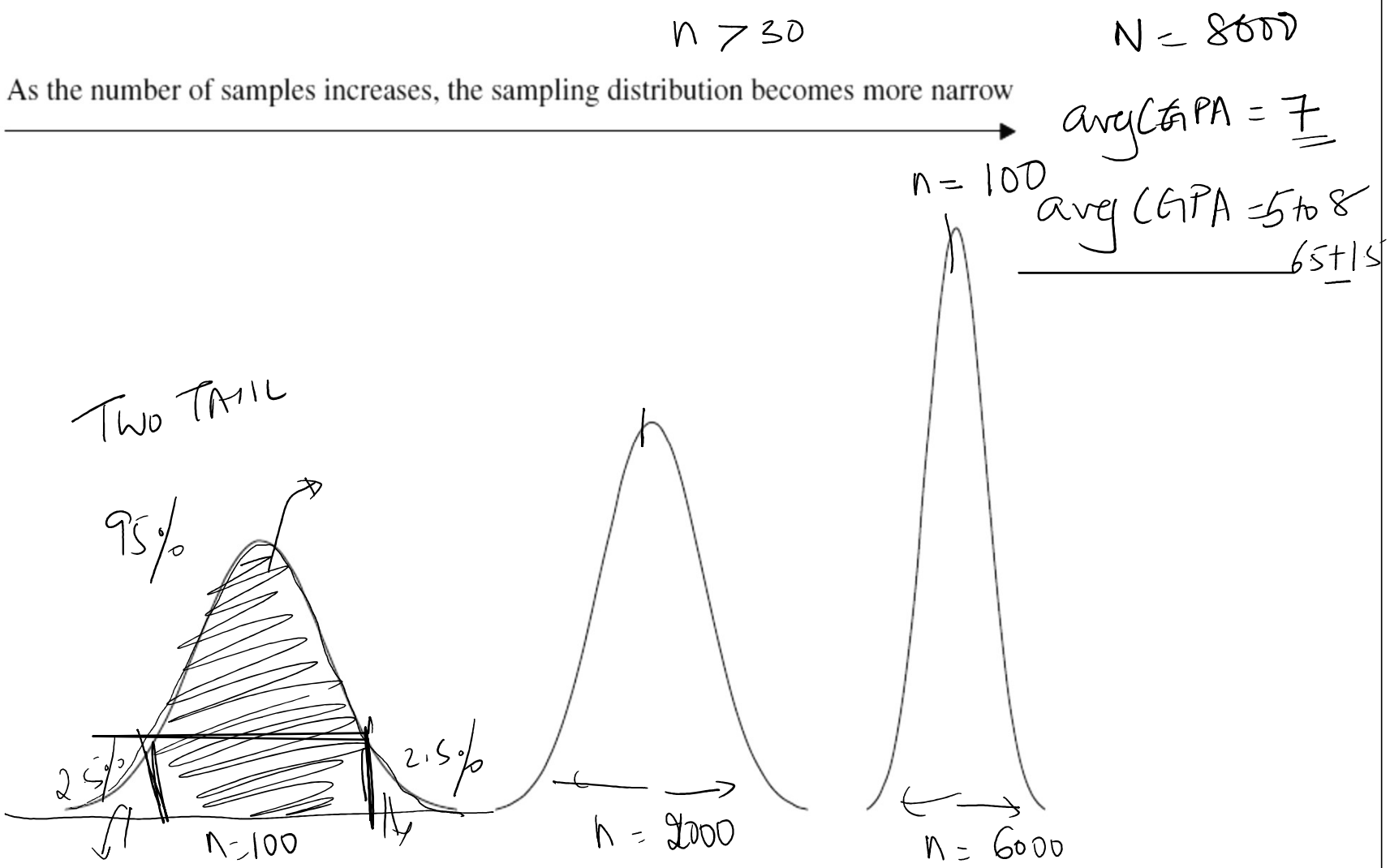
Inferential Analysis Overview

- **Sampling error**
  - If the process of selecting a random sample was repeated a number of times, the means from each sample would be different
- **Sampling distribution**
  - the distribution of the mean values follows a normal distribution for sample sizes greater than 30
  - The sampling distribution is normally distributed because of the central limit theorem
- **Standard error**
  - Relationship between the variance of the original variable and the number of observations in the sample to the sampling distribution is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

  - Where n is sample size and σ is std deviation of population *"s"*

Sampling Distribution of mean



Inferential Analysis Overview

- **Inferential statistics**
  - chance or probability that we will see a particular range of average values
  - Sampling distribution to make this assessment of probabilities.
  - From the central limit theorem , when n >=30 , for any population distribution, the sampling distribution of $\bar{X}$, is approximately normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$ where μ and σ are population values
  - The total area under the normal distribution curve is 1
  - The area between specific z-score ranges represents the probability that a value would lie within this range

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

Inferential Analysis Overview

- **Standard Error Calculation methods**
  - **Confidence intervals:**
    - A confidence interval allows us to make statements concerning the likely range that a population parameter (such as the mean) lies within.
  - **Hypothesis tests:**
    - A hypothesis test determines whether the data collected supports a specific claim.
    - A hypothesis test can refer to
      - a single group
      - refer to two groups
  - **Chi-square:**
    - The chi-square test is a statistical test procedure to understand whether a relationship exists between pairs of categorical variables.
  - **One-way analysis of variance:**
    - This test determines whether a relationship exists between three or more group means.

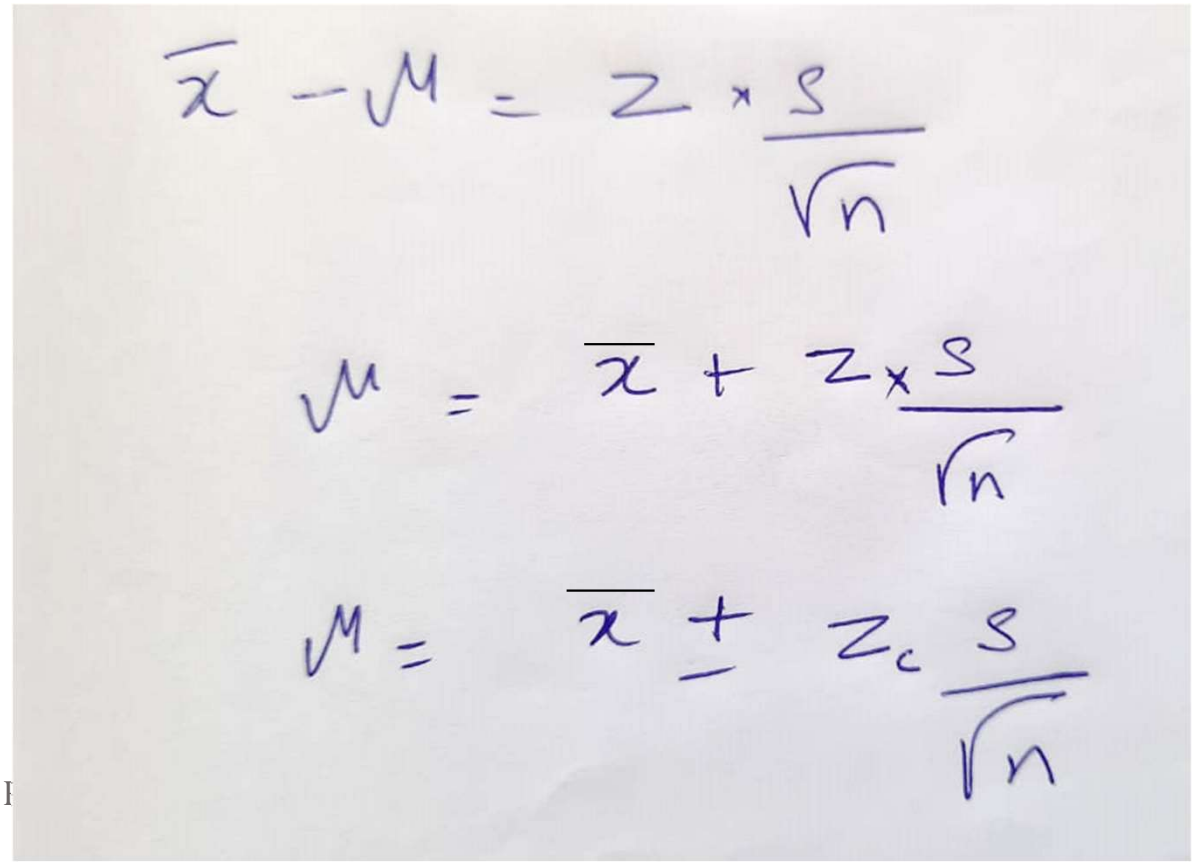
Table 5.3. Summary of inferential statistical tests

	Continuous	Categorical	Number of groups	Number of variables
Confidence intervals	Yes	Yes	1	1
Hypothesis test	Yes	Yes	1 or 2	1
Chi-square	No	Yes	2+	2
One-way analysis of variance	Yes	No	3+	1

Confidence Intervals Formula

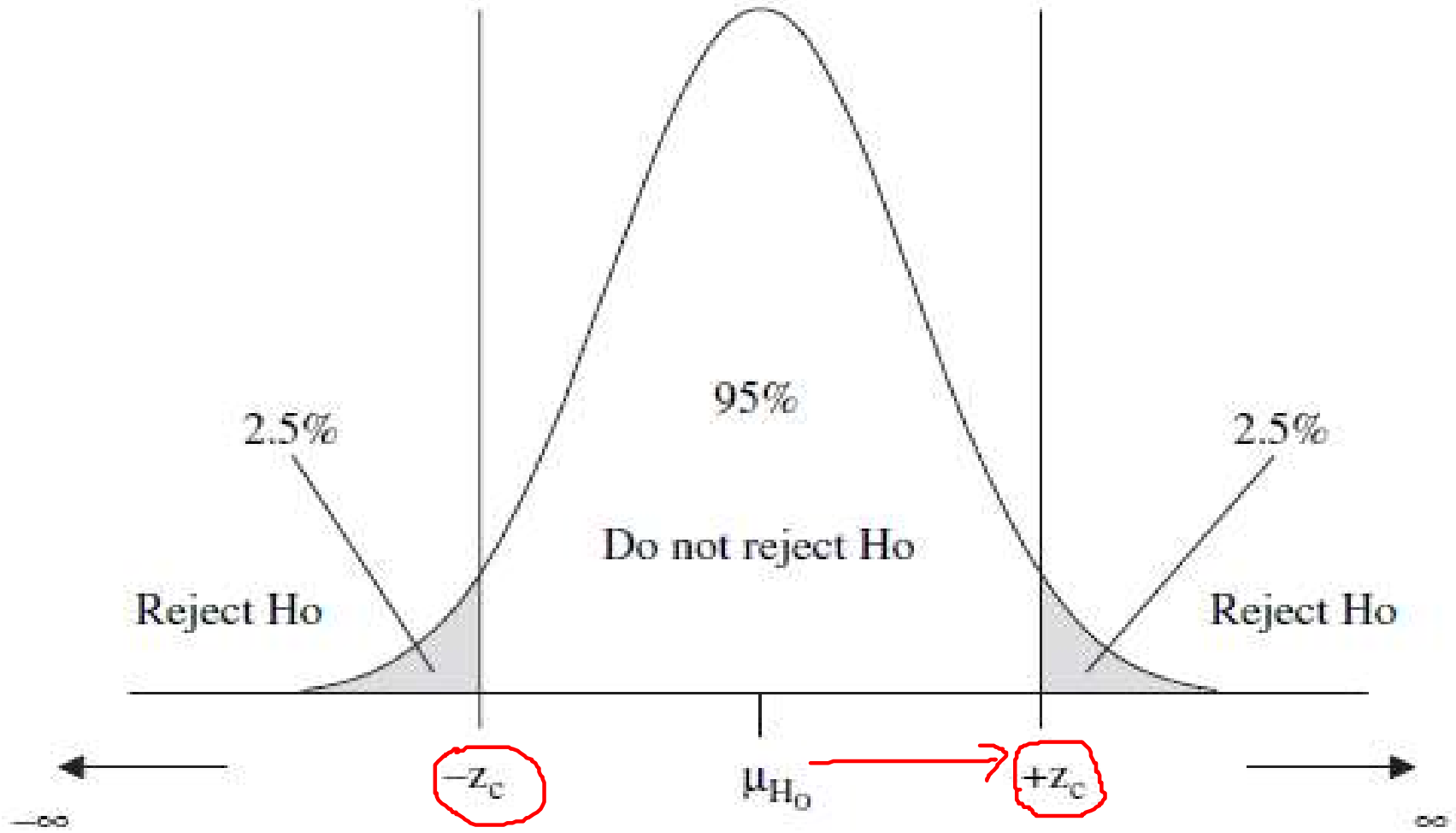
$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Confidence Levels -

95 %.



Determining critical z-score from normal distribution table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4841	0.4801	0.4761	0.4721	0.4641	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4091	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2644	0.2611	0.2579	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1563	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1094	0.1075	0.1057	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064

An area of 0.0250 has a *z-score* of 1.96

Common z-score values

Confidence Level	Area between 0 and z-score	Area in one tail (alpha/2)	z-score
50%	0.2500	0.2500	0.674
80%	0.4000	0.1000	1.282
90%	0.4500	0.0500	1.645
95%	0.4750	0.0250	1.960
98%	0.4900	0.0100	2.326
99%	0.4950	0.0050	2.576

Inferential Analysis
Standard Error Calculation methods

- **Confidence intervals:**
  - Range of values could be used as an estimate for a population
  - Commonly used confidence levels are 90%, 95% & 99%
- **Confidence Ranges for Continuous Variables**
  - Confidence levels
    - for large samples where z-score is number of standard deviations for a given confidence level , n >= 30

$$\bar{x} \pm z_C \frac{s}{\sqrt{n}}$$

    - For small samples

$$\bar{x} \pm t_C \frac{s}{\sqrt{n}}$$

- **Confidence Range for Categorical Variables**

$$p \pm z_C \sqrt{\frac{p(1-p)}{n}}$$

Problem 1 : Confidence Intervals

- Calculate a confidence interval for a set of 54 observations with a mean value of 33.25 and a standard deviation of 12.26. Let the confidence level be 95%

$$\bar{x} \pm z_C \frac{s}{\sqrt{n}}$$

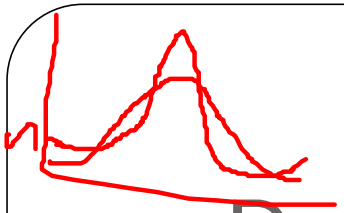
$$33.25 \pm 1.96 \frac{12.26}{\sqrt{54}}$$

$$33.25 \pm 3.27 = 29.98 \text{ to } 36.52$$

For n < 30 use t distribution

- Has fatter tails than the normal distribution
- Distribution will result in larger confidence intervals for smaller samples
- T- distribution based on degrees of freedom , which is 1 minus the sample size

$$\bar{x} \pm t_C \frac{s}{\sqrt{n}}$$



Determining critical t value based on df

	Upper tail area				
df	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947

Problem 2:
Confidence Intervals –small samples

- a set of 11 (n) observations was recorded and the mean value was calculated at 23.22 (x), with a standard deviation of 11.98 (s).

$$\bar{x} \pm t_C \frac{s}{\sqrt{n}}$$

$$23.22 \pm 2.228 \frac{11.98}{\sqrt{11}}$$

$$23.22 \pm 8.05 \quad \Rightarrow \quad 15.17 \text{ to } 31.27$$

Problem 3:

Confidence Intervals – Categorical Variables

- A factory is interested in the proportion of units produced with errors. To make this assessment, a sample of 300 units are tested for errors, and it is determined that 45 contain a problem. Compute the Confidence Interval for the population.

$p = \frac{45}{300} = 0.15$

$$p \pm z_c \sqrt{\frac{p(1-p)}{n}} \quad \text{or} \quad 0.15 \pm 1.96 \sqrt{\frac{0.15 \times (1-0.15)}{300}}$$

$$0.15 \pm 0.04 \Rightarrow 0.19 \text{ to } 0.11$$

Inferential Analysis
Standard Error Calculation methods

- **One-way analysis of variance (ANOVA)**
  - compares the means from three or more different groups
  - Can be applied to
    - cases where the groups are independent and random
    - the distributions are normal
    - the populations have similar variances.
  - hypothesis statement:
    - H_0: The sample means are equal
    - H_a: The sample means are not equal
  - The test has the following steps:

    1. Calculate group means and standard deviations

    2. Determine the within group variation (Mean Square Within)

$$MSW = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{N - k}$$

Inferential Analysis
Standard Error Calculation methods

- **One-way analysis of variance**
- The test has the following steps:

3. Determine the between group variation (Mean Square Between)

$$MSB = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

4. Determine the F-statistic, using the within and between group variation

$$F = \frac{MSB}{MSW}$$

5. Test the significance of the F-statistic

$$df_{within} = N - k = 25$$

$$df_{between} = k - 1 = 3$$

Problem 4

- For example, an on-line computer retail company has call centers in four different locations. These call centers are approximately the same size and handle a certain number of calls each day. An analysis of the different call centers based on the average number of calls processed each day is required.
- H_0: The sample means are equal
- H_a: The sample means are not equal

Problem 4:

Table 5.8. Calls processed by different call centers

Call center A	Call center B	Call center C	Call center D
136	124	142	149
145	131	145	157
139	128	139	154
132	130	145	155
141	129	143	151
143	135	141	156
138	132	138	
139		146	
n = 8	7	8	6

Problem 4

	Call Centre				
	A	B	C	D	
Count (n)	8	7	8	6	Total count N = 29
Mean ($\bar{x}_i$)	139.1	129.9	142.4	153.7	Average of means $\bar{\bar{x}}$ = 141.3
Variance (s_i^2)	16.4	11.8	8.6	9.5	k = 4

Problem 4

$$MSW = \frac{(8-1)\times 16.4 + (7-1)\times 11.8 + (8-1)\times 8.6 + (6-1)\times 9.5}{(29-4)}$$

$$MSW = 11.73$$

$$MSB = \frac{(8\times(139.1-141.3)^2) + (7\times(129.9-141.3)^2) + (8\times(142.4-141.3)^2) + (6\times(153.7-141.3)^2)}{4-1}$$

$$MSB = 626.89$$

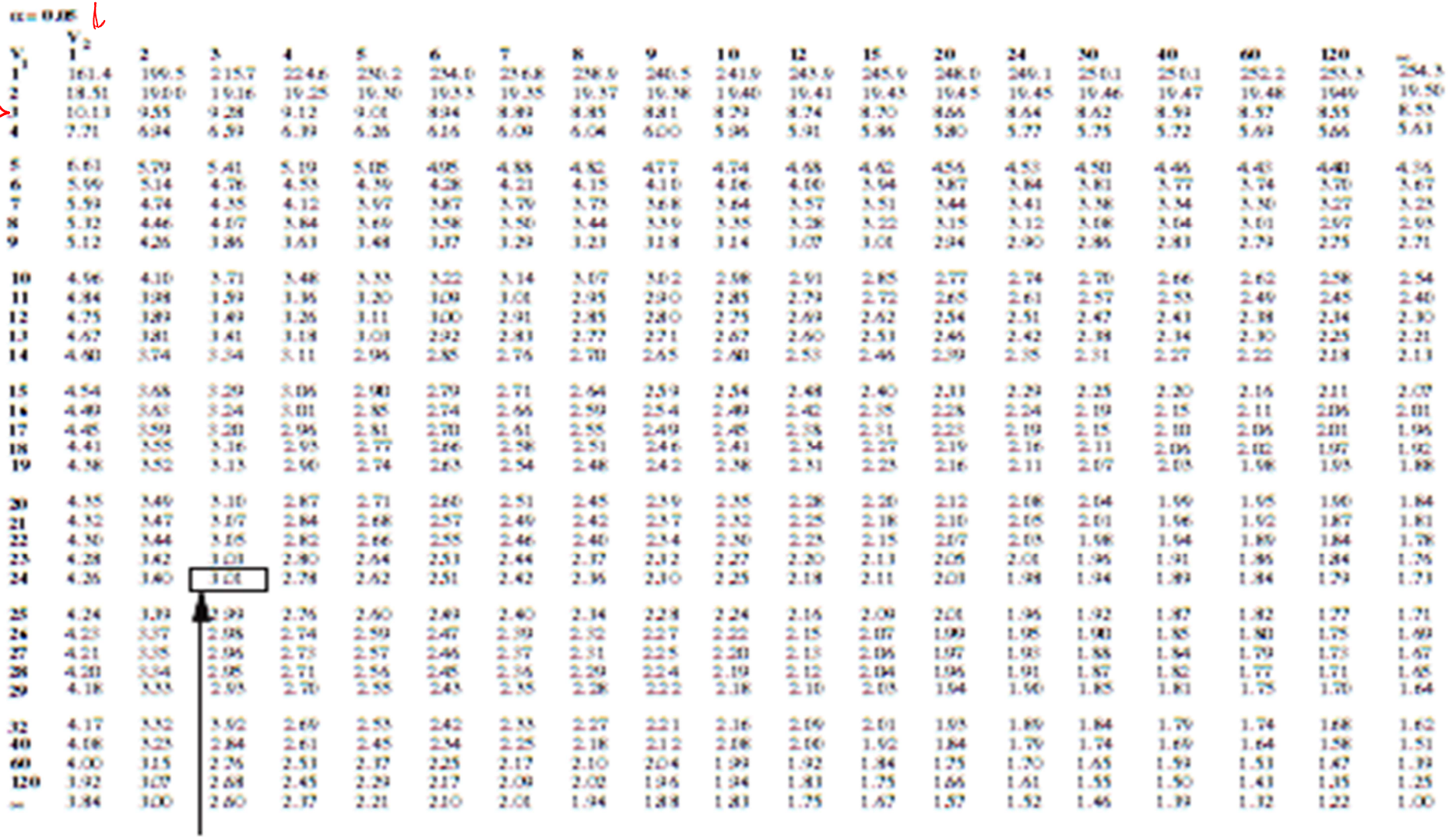
$$F = 53.44$$

$$df_{between} = 4 - 1 = 3$$
$$df_{within} = 29 - 4 = 25$$

Critical Value for F-Statistic is 2.99

the calculated F-statistic is greater than the critical value, we reject the null hypothesis. The means for the different call centers are not equal.

Critical f-statistic



$\alpha = 0.05$	v_2																		
v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	1949	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.16	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
32	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Critical F-statistic where $\alpha = 0.05$, $df_{between} = 4$ (v_1) and $df_{within} = 24$ (v_2)

Inferential Analysis
Standard Error Calculation methods

- **Hypothesis Testing**
  - Null hypothesis (H_o):
    - Claim that a particular population parameter (e.g. mean) equals a specific value.
  - Alternative hypothesis (H_a):
    - Conclusion that we would be interested in reaching if the null hypothesis is rejected.
    - Also called research hypothesis
- **Hypothesis Assessment**
  - The statistic of interest from the sample is calculated
  - difference between the value claimed in the hypothesis statement and the calculated sample statistic.
  - **Approach 1 – α – significance level of test**
    - For large sample sets - identifying where the hypothesis test result is located on the normal distribution curve of the sampling distribution, will determine whether the null hypothesis is rejected.
  - Approach 2 – statistic looked up a Normal Distribution table to find p-value.
    - If hypothesis test is two-sided double p-value & compare with α
    - If p-value is lesser than α then null hypothesis is rejected

Inferential Analysis
Standard Error Calculation methods

- Hypothesis Test: Single Group, Continuous Data

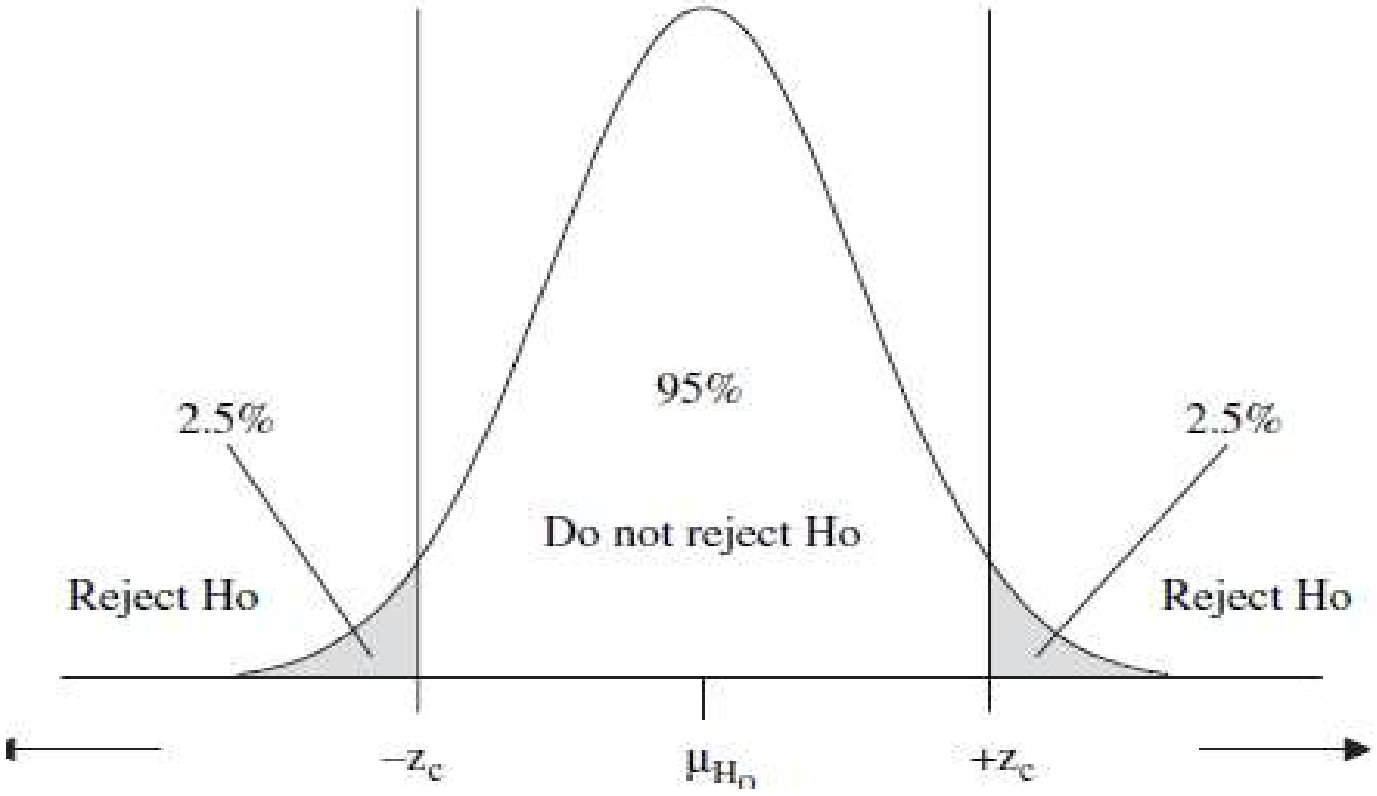
$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- Hypothesis Test: Single Group, Categorical Data

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Inferential Analysis
Standard Error Calculation methods

- **Hypothesis Testing – two tailed test**
  - **Ha: $\mu \neq 0$**



Problem 5

- To test the hypothesis that the number of days to process a passport is 12 (m), 45 passport applications were randomly selected and timed .The average time to process the passport application was 12.1 (x) and the standard deviation was 0.23 (s) and confidence level was set to 95%.

Claim: The average time to process a passport is 12 days

$$H_0 : \mu = 12$$
$$H_a : \mu \neq 12$$

$$z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = 2.9$$

$z > z_c \Rightarrow$ reject null

$\Rightarrow$ accept null

Z-score of 2.9 > 1.96
we reject the null hypothesis
the average number of days to process a passport is not 12 days.

Problem 6

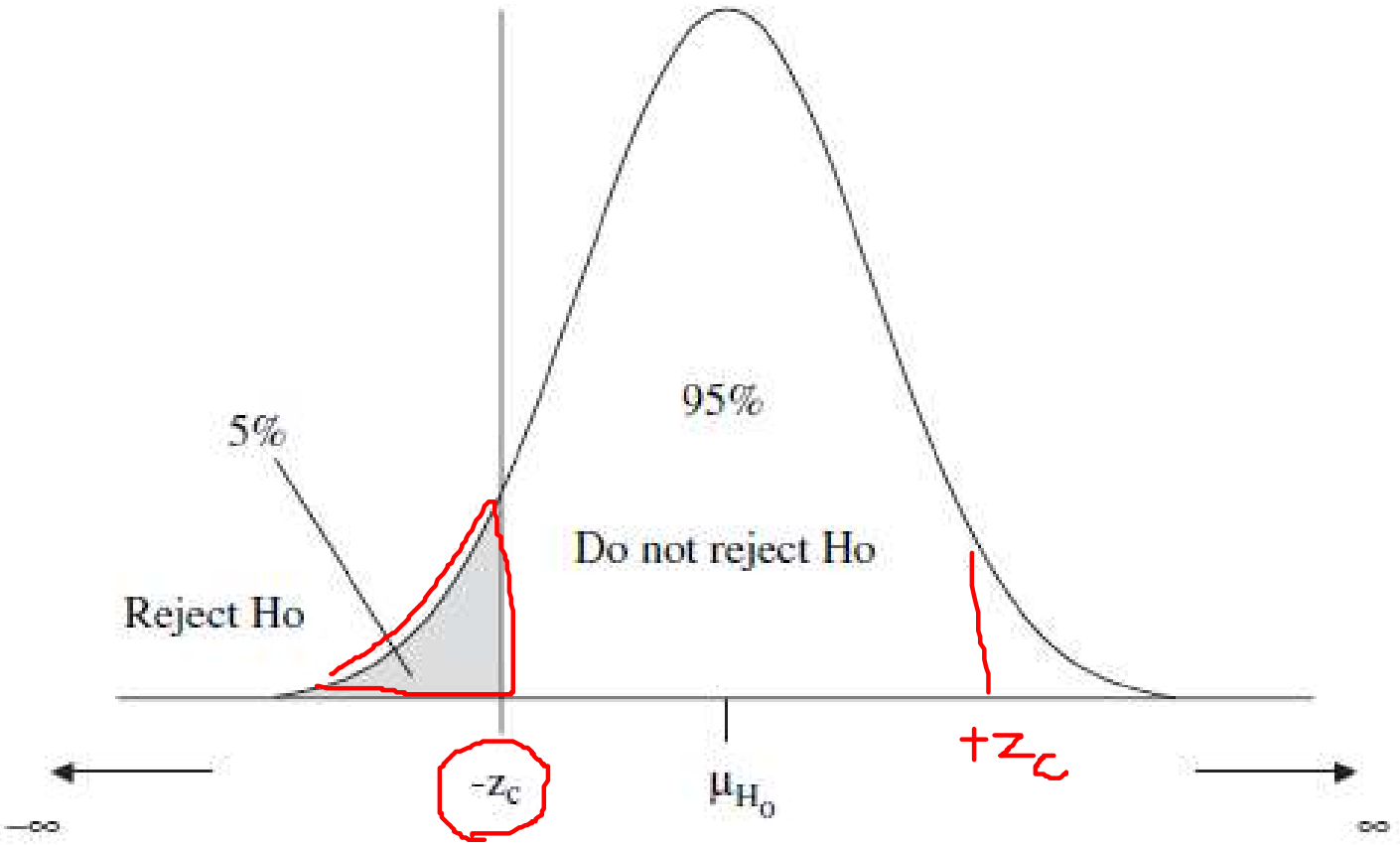
- To test the claim that more than eight out of ten dog owners prefer a certain brand (brand X) of dog food
- Claim : $H_0 : \pi = 0.8$, $H_a : \pi > 0.8$
- To test this hypothesis, 40 random dog owners (n = 40) were questioned and the proportion that responded that they preferred brand X was 33 out of 40 .

$$z = \frac{\rho - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

- Z = 0.395
- $Z_c = 1.65$
- we do not reject the null hypothesis.

Inferential Analysis
Standard Error Calculation methods

- **Hypothesis Testing – one tailed test**



Determining critical z-score from normal distribution table

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4841	0.4801	0.4761	0.4721	0.4641	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4091	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2644	0.2611	0.2579	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2297	0.2266	0.2236	0.2207	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1563	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1094	0.1075	0.1057	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064

An area of 0.0250 has a *z-score* of 1.96

Inferential Analysis
Standard Error Calculation methods

- Hypothesis Test: Two Groups, Continuous Data
  - When group size is less than 30

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

  - When group size is greater than 30

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Hypothesis Test: Two Groups, Categorical Data

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Problem 7

- Claim: A new drug reduces the number of strokes.
  - where p1 is the proportion of the population with strokes taking the new medicine
  - and p2 is the proportion of the population with strokes taking the placebo.

$$H_0: \pi_1 = \pi_2$$

$$H_a: \pi_1 < \pi_2$$

Problem 7

- There were 10,004 patients in the first group who took the medicine (n_1). and of these 213 had strokes (X_1). There were 10,013 patients in group 2 that did not take the medicine and took a placebo instead (n_2) and in this group 342 patients had a stroke (X_2).

	Takes medicine	Takes placebo	Total
Has strokes	213	342	555
No strokes	9,791	9,671	19,462
Totals	10,004	10,013	20,017

$$p = \frac{X_1 + X_2}{n_1 + n_2} \qquad p_1 = \frac{X_1}{n_1}$$

Problem 7

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$Z = -5.5$$

n > 30 Normal
95% tail = 1 − 0.95 = 0.05
Z = −1.65

- we reject the null hypothesis
- Conclude the number of strokes for the group taking the medicine is lower than the group that does not take the medicine.

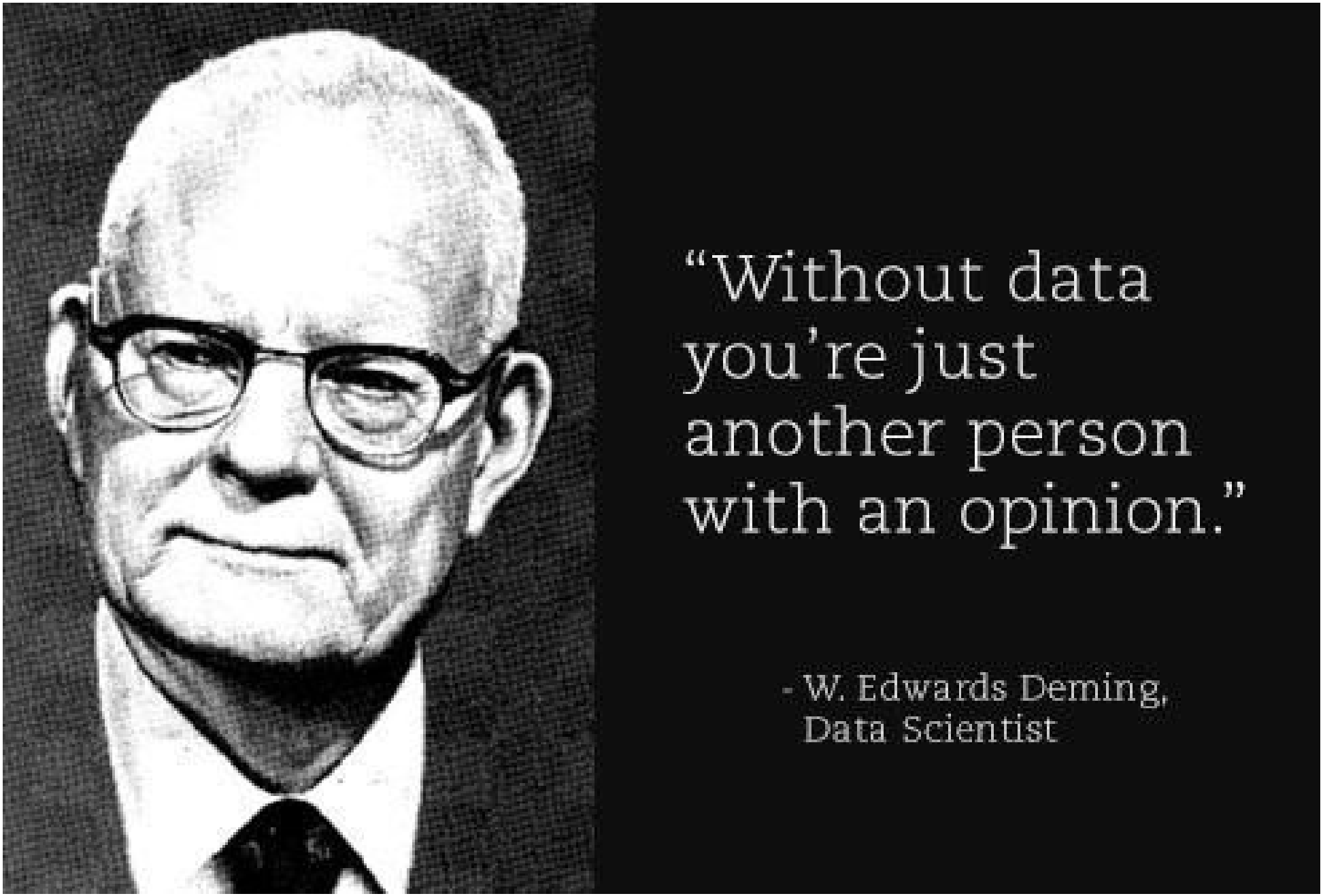
Inferential Analysis
Standard Error Calculation methods

- **Paired Test**
  - Claim: There is no difference in the wear of shoes made from material X compared to shoes made from material Y.
  - To test this claim, the null and alternative hypothesis are set up:
    - H_o: $\mu_D = 0$
    - H_a: $\mu_D \neq 0$
    - where μ is the difference between the wear of shoes made with material X and the wear of shoes made with material Y.
  - Formula used

$$t = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

Errors

- hypothesis test is based on a sample and samples vary
- potential errors are:
  - Type I Error: null hypothesis is rejected when it really should not be. These errors are minimized by setting the value of alpha low.
  - Type II Error: null hypothesis is not rejected when it should have been. These errors are minimized by increasing the number of observations in the sample.


"Without data
you're just
another person
with an opinion."
- W. Edwards Deming,
Data Scientist

References

- Glenn J. Myatt, Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining, John Wiley, November 2006.

- G. Shmueli, N. R. Patel, and P.C. Bruce, Data Mining for Business Intelligence, John Wiley and Sons, 2010