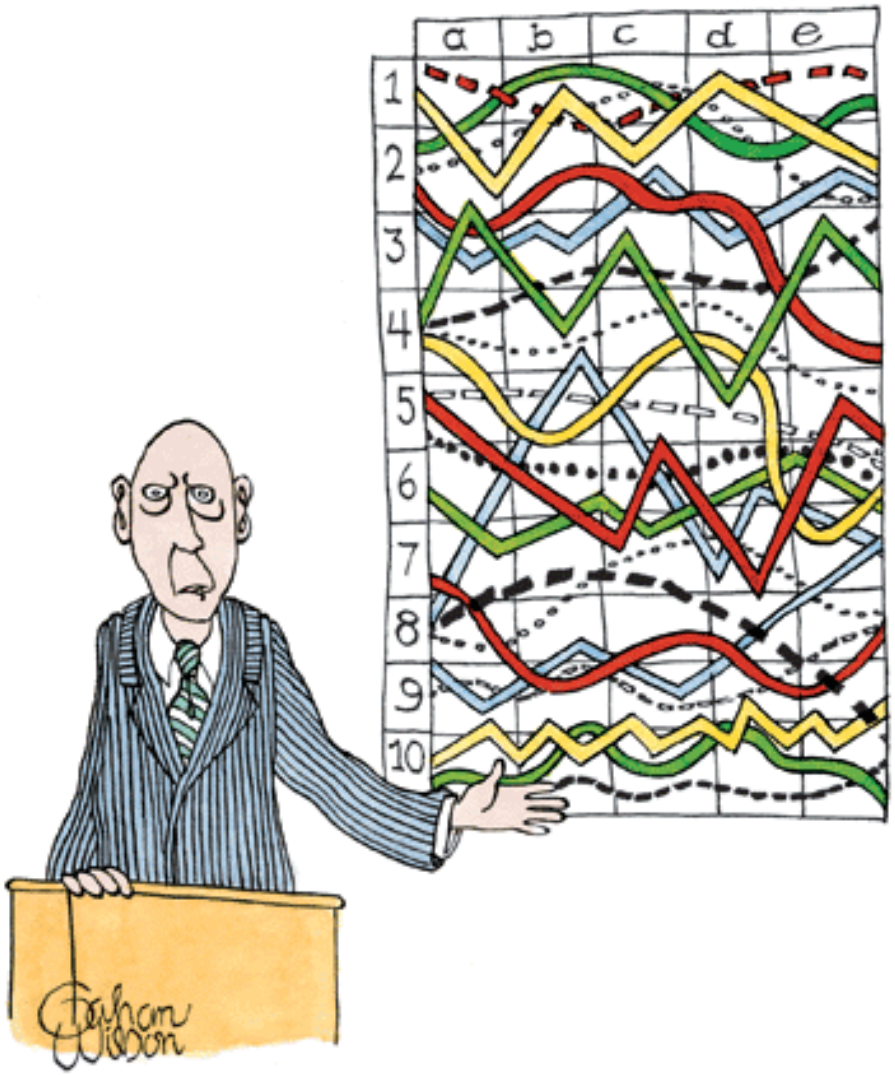
Applied Data Science

Bharathi H N

Assistant Professor
Department of Computer Engineering
K. J. Somaiya College of Engineering
Somaiya Vidyavihar University



"I'll pause for a moment so you can let this information sink in."

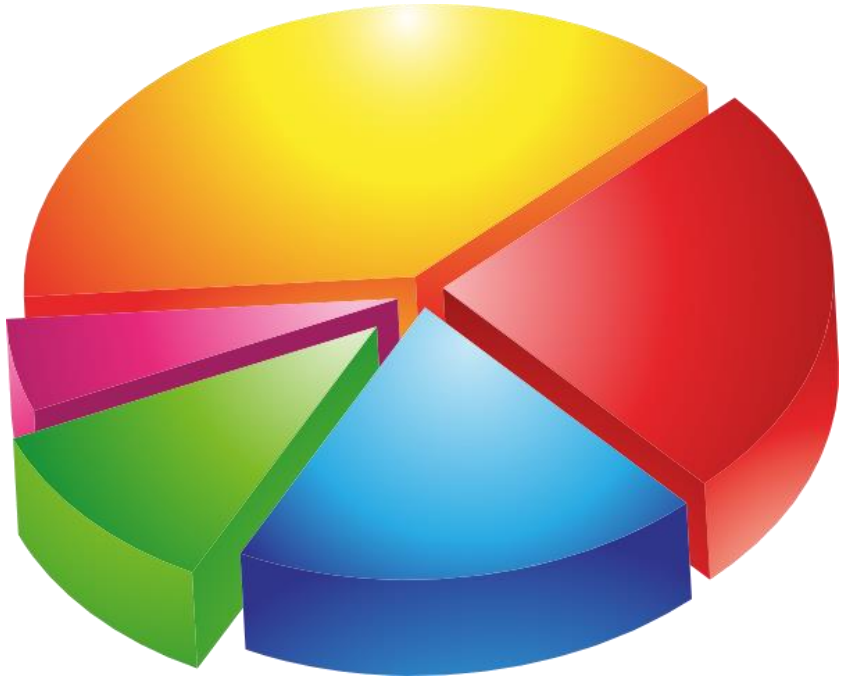
Everything You Wanted to Know about Statistics but Were Afraid to Ask

Andrew L. Luna, Ph.D

outline

- Section 1: Scientific Method, descriptive/inferential statistics, sampling, validity, and types of data.
- Section 2: Descriptive statistics, normal distribution, Central Limit Theorem, measures of central tendency, z scores, hypothesis testing.
- Section 3: Type I and Type II Error, Pearson R, Degrees of Freedom, Chi Square, t-test.
- Section 4: ANOVA and Regression

Connection?




UKRAINE
Kalan-
chak
Oblast
Armyansk
Armyansk
Skadovs'k
Isthmus
of Perekop
Krasnoperekopsk
Krasnoperekopsk
Krasnoperekopsk
Novooleksiiyka
Heniches'k
Kherson
Sivash
Lagoon System
Karkinit Bay
Rozdolne
Rozdolne
Dzhankoy
Dzhankoj
Pervomayske
Pervomayske
Dzhankoy
Azovske
Arabat Spit
Sea of Azov
Chorno-
morske
Chornomorske
Tarkhan Highlands
Crimean Prairie Lands
Krasnoh-
vardiyske
Nizhnyohirskyi
Cape
Kazantyp
Scholkine
Kerch
Novoozerne
Novoseliv'ske
Nizhnyohirskyi
Sovetskyi
Bay of Arabat
Baherove
Myrnyj
Saki
Autonomous Republic Crimea
Oktiabrs'ke
Sovetskyi
Kirovske
Lenine
Kerch
Strait of
RUSSIA
Yevpatoria
Saki
Yevpatoria
Lenine
Kerch Peninsula
Taman
Peninsula
Prymors'kyj
Kirovske
Saki
Simferopol
Zuia
Bilohirsk
Staryi Krym
Caffa Bay
Novofedorivka
Hresivs'kyi
Bilohirsk
Feodosiya
Kalamita Bay
Mykolaivka
Simferopol
Schebetovka
Ordszhonikidze
Feodosiya
Koktebel'
Sudak
Kurortne
Simferopol
Mountains
Sudak
Nowyj Svit
Kacha
Bakhchysarai
Alushta
Krimskiy
Trolejbus
Bakhchysarai
Roman Kosh
(1545 m)
Alushta
Inkerman
Sevastopol'
Partenit
Masandra
Hurzuf
Crimean
Yalta
City of Sewastopol'
Haspra
Yalta
Balaklava
Simeiz
Koreiz
Foros
Parkove
Alupka
Black Sea
RUSSIA
UKRAINE
Crimea
50 km

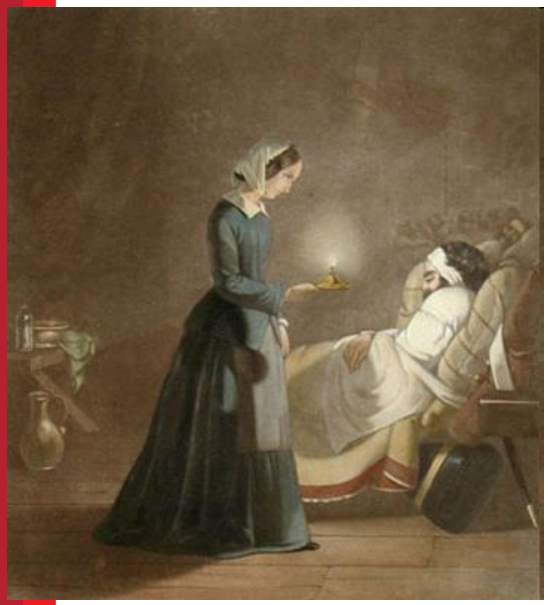
The Crimean War

The Crimean War (1853-1856) was a bloody battle between the Russians and the British Alliance (Great Brittan, France, Ottoman Empire, Kingdom of Sardinia) that saw great casualties on both sides.

"Half a league, half a league,
Half a league onward,
All in the valley of Death
Rode the six hundred.
"Forward, the Light Brigade!
"Charge for the guns!" he said:
Into the valley of Death
Rode the six hundred…"

Alfred, Lord Tennyson, "The Charge of the Light Brigade." Written to memorialize events in the Balaclava, Oct. 25, 1854.

Florence Nightingale “Lady with the Lamp”



Florence Nightingale (1820-1910)

Lo! in that hour of misery
A lady with a lamp I see
Pass through the glimmering gloom,
And flit from room to room.

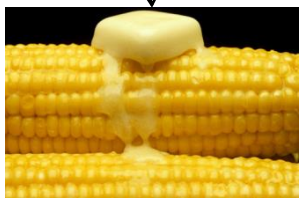
Henry Wadsworth Longfellow’s 1857 poem “Santa Filomena”

Florence Nightingale observed the horrific conditions of the wounded and was instrumental in convincing the British government to make sweeping changes in the sanitary conditions of the make-shift “hospitals.” Her work to make conditions more sanitary caused the mortality rate to decline from 44 percent to 2 percent within 6 months.

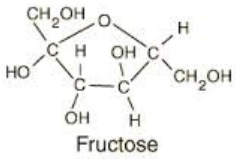
Nightingale wanted to create a visual representation of her argument on sanitary conditions in her reports to the British government. She saw that creating a circle denoting 100 percent of an event, and dividing that circle into segments, she could produce a simple graph that contained a lot of information…thus, Florence Nightingale created the PIE CHART!

Knowledge, Data, Information, and Decisions…

Data



Knowledge



Fructose

Information

New Knowledge/Decisions

The Scientific Method

- Scientific Method
  - The way researchers go about using knowledge and evidence to reach objective conclusions about the real world.
  - The analysis and interpretation of empirical evidence (facts from observation or experimentation) to confirm or disprove prior conceptions

Characteristics of the scientific method

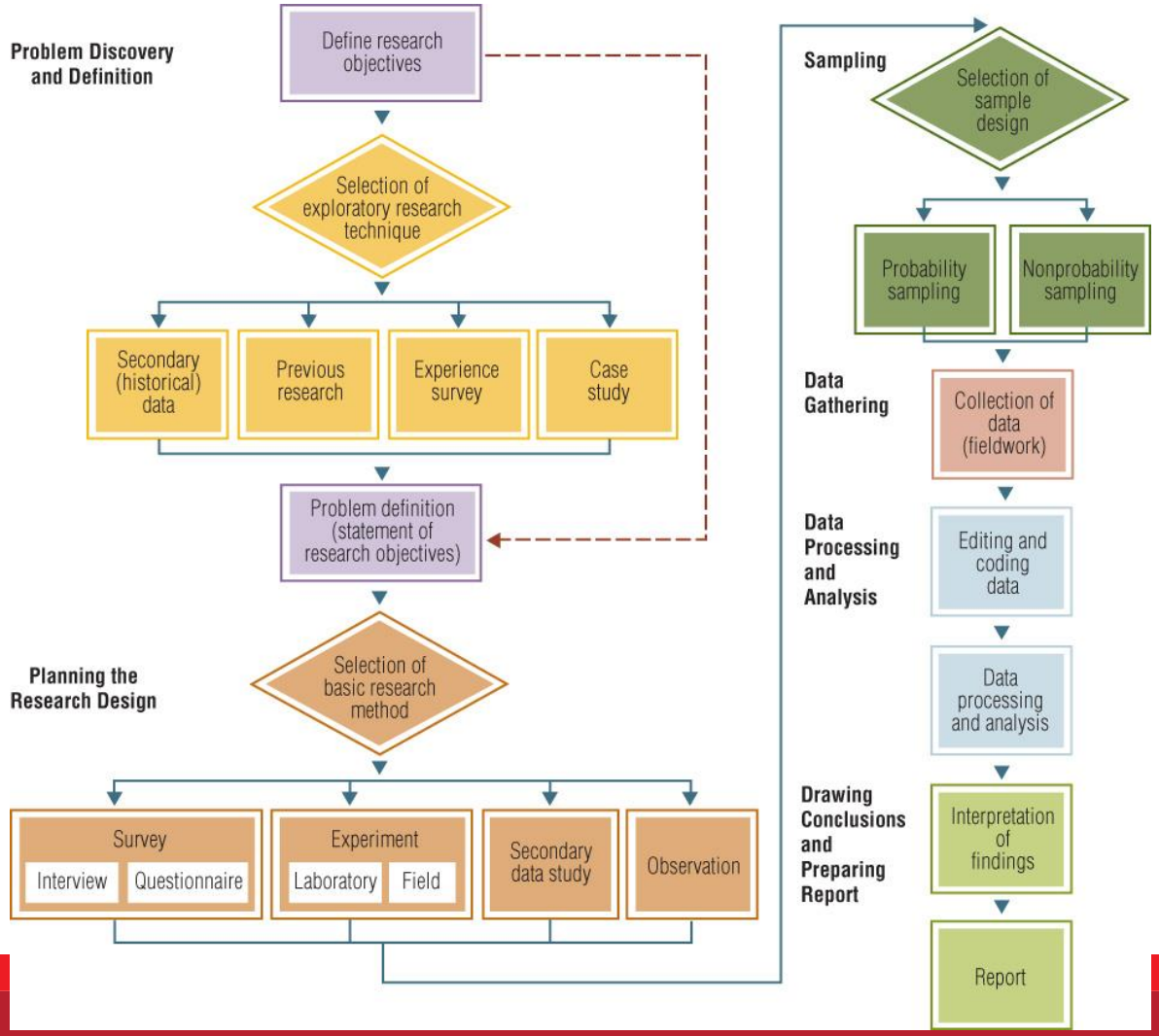
- **Scientific Research is Public** – Advances in science require freely available information (replication/peer scrutiny)
- **Science is Objective** – Science tries to rule out eccentricities of judgment by researchers and institutions. Wilhelm von Humboldt (1767-1835), founder University of Berlin (teaching, learning, research) "Lehrfreiheit," "Lernfreiheit," and "Freiheit der Wissenschaft"
- **Science is Empirical** – Researchers are concerned with a world that is knowable and potentially measurable. Researchers must be able to perceive and classify what they study and reject metaphysical and nonsensical explanations of events.

Characteristics of the scientific method, cont.

- **Science is Systematic and Cumulative** – No single research study stands alone, nor does it rise or fall by itself. Research also follows a specific method.
- **Theory** – A set of related propositions that presents a systematic view of phenomena by specifying relationsships among concepts
- **Law** – is a statement of fact meant to explain, in concise terms, an action or set of actions that is generally accepted to be true and universal
- **Science is Predictive** – Science is concerned with relating the present to the future (making predictions)
- **Science is Self-Correcting** – Changes in thoughts, theories, or laws are appropriate wen errors in previous research are uncovered

Flow chart of the scientific method

Note: Diamond-shaped boxes indicate stages in the research process in which a choice of one or more techniques must be made. The dotted line indicates an alternative path that skips exploratory research.



Two basic types of research

- **Qualitative Research (words) -** is by definition exploratory, and it is used when we don't know what to expect, to define the problem or develop an approach to the problem. It's also used to go deeper into issues of interest and explore nuances related to the problem at hand. Common data collection methods used in qualitative research are focus groups, in-depth interviews, uninterrupted observation, bulletin boards, and ethnographic participation/observation.
- **Quantitative Research (numbers) -** is conclusive in its purpose as it tries to quantify the problem and understand how prevalent it is by looking for projectable results to a larger population. Here we collect data through surveys (online, phone, paper), audits, points of purchase (purchase transactions), and other trend data.

Stating a hypothesis or research question

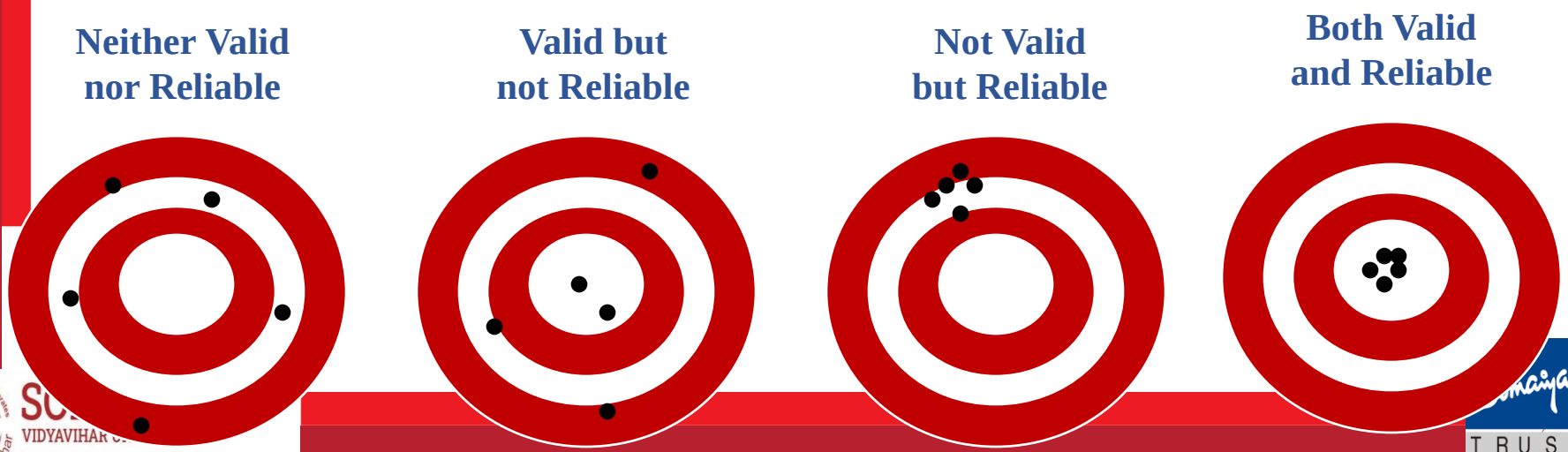
- **Research Question** – A formally stated question intended to provide indications about some; it is not limited to investigating relationships between variables. Used when the researcher is unsure about the nature of the problem under investigation.
- **Hypothesis** – a formal statement regarding the relationship between variables and is tested directly. The predicted relationship between the variables is either true or false.
  - Independent Variable (X_i)– the variable that is systematically varied by the researcher
  - Dependent Variable (Y_i) – the variable that is observed and whose value is presumed to depend on independent variables

Hypothesis vs. research question

- Research Question: “Does television content enrich a child’s imaginative capacities by offering materials and ideas for make-believe play?
- Hypothesis: The amount of time a child spends in make-believe play is directly related to the amount of time spent viewing make-believe play on television.
- Null Hypothesis: the denial or negation of a research hypothesis; the hypothesis of no difference
  - H_O: “There is no significant difference between the amount of time children engage in make-believe play and the amount of time children watch make-believe play on television.”

Data analysis and interpretation

- Every research study must be carefully planed and performed according to specific guidelines.
- When the analysis is completed, the researcher must step back and consider what has been discovered.
- The researcher must ask two questions:
  - Are the results internally and externally valid?
  - Are the results valid



Internal validity

- If y = f(*x*), control over the research conditions is necessary to eliminate the possibility of finding that y = f(*b*), where b is an extraneous variable.
- **Artifact** – Any variable that creates a possible but incorrect explanation of results. Also referred to as a confounding variable.
- The presence of an artifact indicates issues of internal validity; that is, the study has failed to investigate its hypothesis

What affects Internal validity

- **History** – various events that occur during a study may affect the subject's attitudes, opinions, and behavior.
- **Maturation** – Subjects' biological and psychological characteristics change during the course of a study (mainly longitudinal).
- **Testing** – The act of testing may cause artifacts depending on the environment, giving similar pre-tests/post-tests, and/or timing.
- **Instrumentation** – A situation where equipment malfunctions, observers become tired/casual, and/or interviewers may make mistakes.
- **Statistical regression** – Subjects who achieve either very high or very low scores on a test tend to regress to (move toward) the sample or population mean.

What affects internal validity, cont.

- **Experimental Mortality** – All research studies face the possibility that subjects will drop out for one reason or another.
- **Sample Selection** – When groups are not selected randomly or when they are not homogeneous
- **Demand Characteristics** – Subjects' reactions to experimental situations. Subjects who recognize the purpose of a study may produce only "good" data for researchers (Hawthorne Effect).
- **Experimenter Bias** – Researcher becomes swayed by a client's (or personal) wishes for a project's results (Blind vs. Double Blind).
- **Evaluation Apprehension** – Subjects are afraid of being measured or tested.
- **Causal Time Order** – An experiment's results are due not to the stimulus (independent) variable but rather to the effect of the dependent variable.

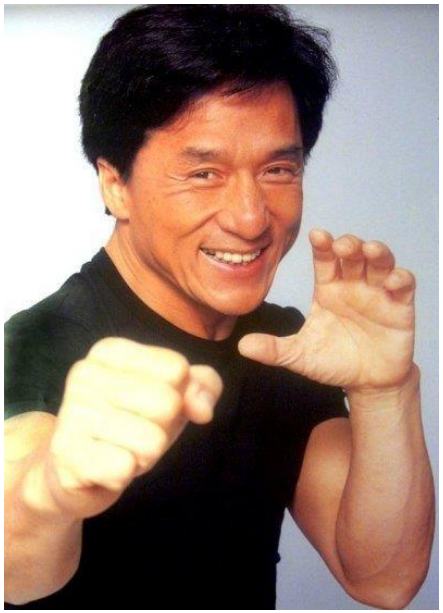
What affects internal validity, cont.

- **Diffusion or Imitation of Treatments** – Where respondents may have the opportunity to discuss the experiment/study with another respondent who hasn't yet participated.
- **Compensation** – The researcher treats the control group differently because of the belief that the group has been "deprived."
- **Compensatory Rivalry** – Subjects who know they are in the control group may work harder to perform differently or outperform the experimental group.
- **Demoralization** – Control group may feel demoralized or angry that they are not in the experiential group.

External validity

- How well the results or a study can be generalized across the population.
- Use random samples.
- Us heterogeneous (diverse) samples and replicate the study several times.
- Select a sample that is representative of the group to which the results will be generalized.

Sample Population



Probability versus Nonprobability Sampling

- Probability Sampling
  - A sampling technique in which every member of the population has a known, nonzero probability of selection.
- Nonprobability Sampling
  - A sampling technique in which units of the sample are selected on the basis of personal judgment or convenience.
  - The probability of any particular member of the population being chosen is unknown.

Replication

- **Replication -** the independent verification of a study and is designed to eliminate:
  - Design-specific results
  - Sample-specific results
  - Method-specific results
- **Literal Replication** – Involves the exact duplication of a previous study
- **Operational Replication** – attempts to duplicate only the sampling and experimental procedures of a previous study
- **Instrumental replications** – Attempts to duplicate the dependent measures used in a previous study.
- **Constructive Replication** – Attempts to test the validity of a previous study by not imitating the previous study.

concepts

- Building Blocks of Theory
- Abstract
- Represents broad, general ideas
- Not directly observable
- Examples:
  - Reality
  - Ideology
  - Commercialism
  - Value
  - Aesthetics

Theory

- Systematic; abstract explanation of some aspect of reality
- Primary goal is to provide a framework that links research and practice and contributes to making findings meaningful and generalizable
- Structure for interpretation of findings
- Means for summarizing and explaining observations for an isolated study
- Source to generate hypothesis
- Framework for guiding research
- Guide for selecting appropriate method
- Basis to describe, explain or predict factors influencing outcomes

constructs

- Concepts that are specified in such away they are observable in the real world
- Invented
- Examples
  - (Reality) Opinion, Choice
  - (Ideology) Conservatism, Liberalism, Libertarianism, Socialism
  - (Commercialism) Profit, Ratings
  - (Value) amount of information , newsworthiness, time spent
  - (Aesthetics) Color, Layout, Sound, Composition

Variables

- Concepts that are observable and measurable
- Have a dimension that can vary
- Narrow in meaning
- Examples:
  - Color classification
  - Loudness
  - Level of satisfaction/agreement
  - Amount of time spent
  - Media choice

Types and forms of variables

- Variable Types:
  - **Independent** – those that are systematically varied by the researcher
  - **Dependent** – those that are observed. Their values are resumed to depend on the effects of the independent variables

- Variable Forms:
  - **Discrete** – only includes a finite set of values (yes/no; republican/democrat; satisfied….not satisfied, etc.)
  - **Continuous** – takes on any value on a continuous scale (height, weight, length, time, etc.)

Scales: Concept

- A generalized idea about a class of objects, attributes, occurrences, or processes

Example: Satisfaction

Scales: Operational Definition

- Specifies what the researcher must do to measure the concept under investigation

Example: A 1-7 scale measuring the level of satisfaction; A measure of number of hours watching TV.

Media skepticism: conceptual definition

- Media skepticism - the degree to which individuals are skeptical toward the reality presented in the mass media. Media skepticism varies across individuals, from those who are mildly skeptical and accept most of what they see and hear in the media to those who completely discount and disbelieve the facts, values, and portrayal of reality in the media.

Media skepticism: operational definition

Please tell me how true each statement is about the news story. Is it very true, not very true, or not at all true?

1. The program was not accurate in its portrayal of the problem.
2. Most of the story was staged for entertainment purposes.
3. The presentation was slanted and unfair.

I believe national network news is fair in its portrayal of national news stories:

Strongly Disagree **Disagree** **Neutral** **Agree** **Strongly Agree**

Numbers, numbers everywhere

555-867-5309

9001

9

.05

3.5

97.5

502

4,832

834,722

77

999

65.87

36^2

.998

4001

51

.56732

1,248,965

9

21

2,387

672

145

999-99-9999

324

409

35.5

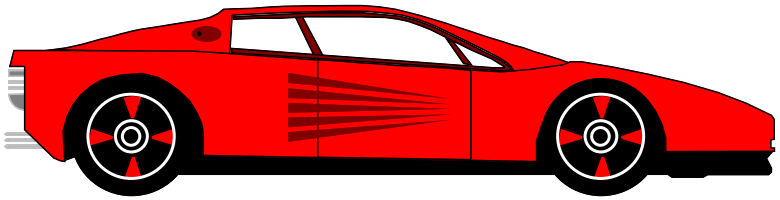
Scales

- Represents a composite measure of a variable
- Series of items arranged according to value for the purpose of quantification
  - Provides a range of values that correspond to different characteristics or amounts of a characteristic exhibited in observing a concept.
  - Scales come in four different levels: Nominal, Ordinal, Interval, and Ratio

Nominal Scale

- Indicates a difference

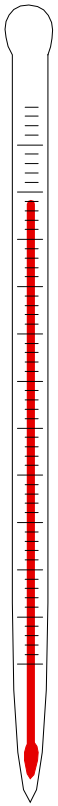
Ordinal Scale



- Indicates a difference
- Indicates the direction of the distance (e.g. more than or less than)

Interval Scale

32 f



0 c

- Indicates a difference
- Indicates the direction of the distance (e.g. more than or less than)
- Indicates the amount of the difference (in equal intervals)

Ratio Scale

- Indicates a difference
- Indicates the direction of the distance (e.g. more than or less than)
- Indicates the amount of the difference (in equal intervals)
- Indicates an absolute zero

Discussion/Test: Identify the Scale

- Sammy Sosa # 21
- Prices on the Stock Market
- Gender: Male = 1 or Female = 2
- Professorial rank: Asst. = 1, Assoc. = 2, Full = 3
- Number of Newspapers sold each day
- Amount of time a subject watches a television program
- Arbitron Rating
- Salary
- Satisfaction on a 1-7 Likert Scale
- How many times respondents return to a website
- Decibel level of a speaker
- Weight of paper

Things are not always what they seem to be…

- Radio Stations

- Does it show a difference?
- Does it show the direction of difference?
- Is the difference measured in equal intervals?
- Does the measure have an absolute zero?

Operational definitions: classroom project

- Provide operational definitions for the following:
  - Artistic quality
  - Objectionable song lyrics
  - Writing quality
  - Sexual content
  - Critical Thinking

Two sets of scores…

Group 1	Group 2
100, 100	91, 85
99, 98	81, 79
88, 77	78, 77
72, 68	73, 75
67, 52	72, 70
43, 42	65, 60

How can we analyze these numbers?

Choosing one of the groups… Descriptive statistics

Distribution of Responses

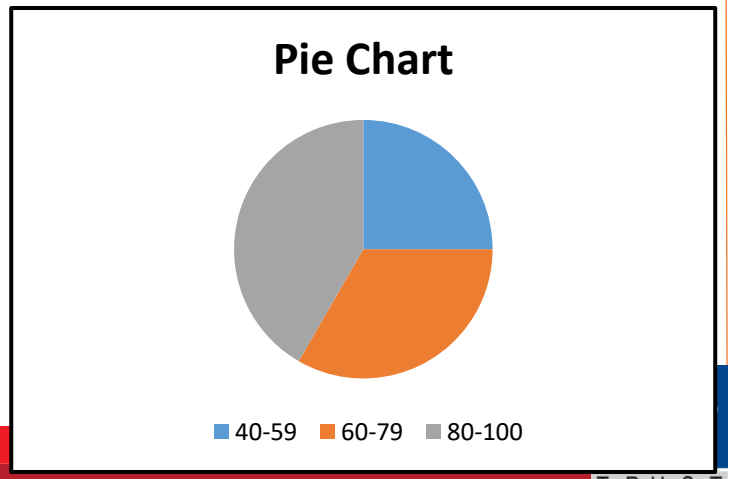
100, 100
99, 98
88, 77
72, 68
67, 52
43, 42

Frequency Distribution

Scores	Frequency (N = 12)
100	2
99	1
98	1
88	1
77	1
72	1
68	1
67	1
52	1
43	1
42	1

Frequency Distribution Grouped in Intervals

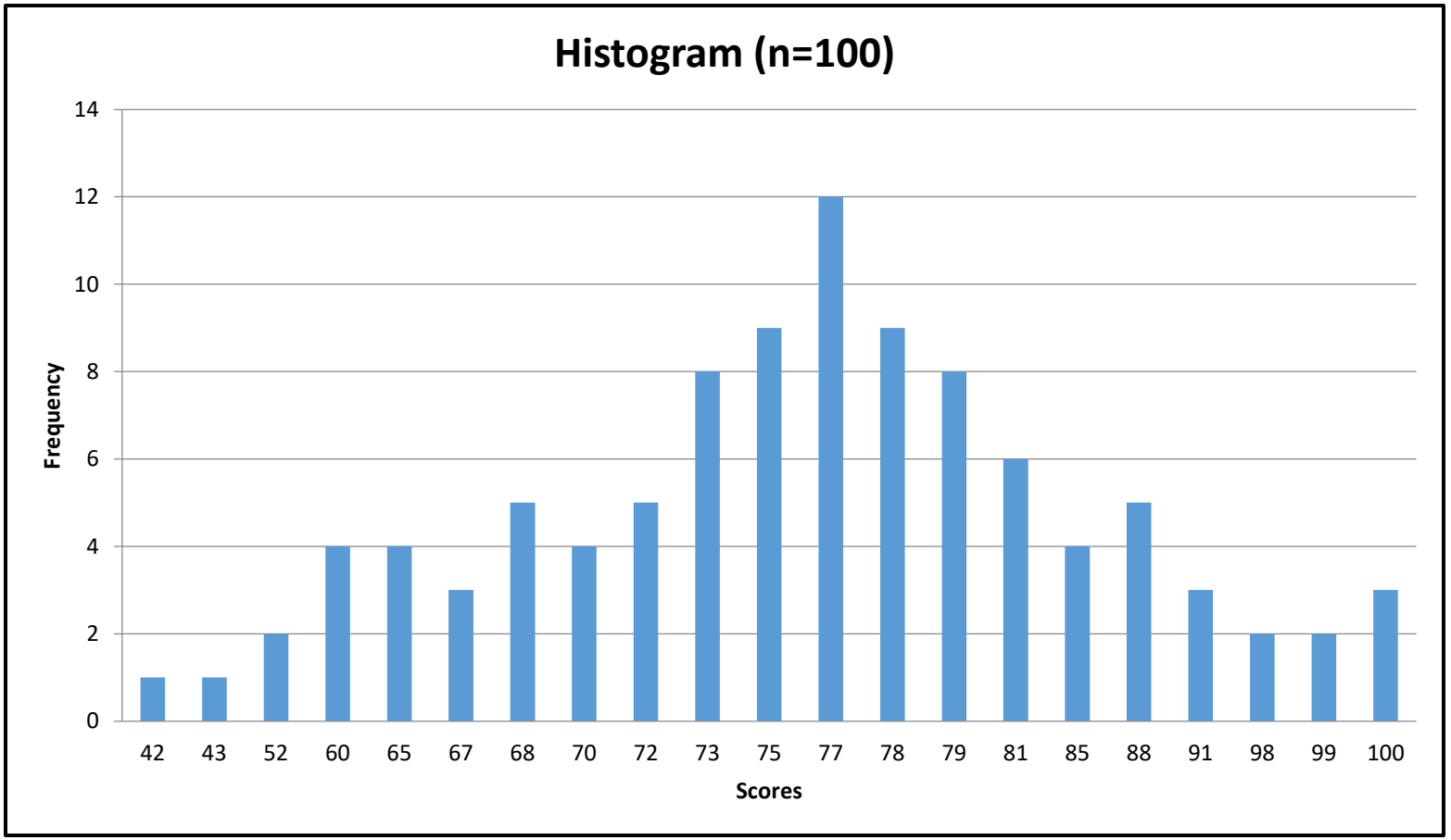
Scores	Frequency (N = 12)
40 - 59	3
60 - 79	4
80 - 100	5



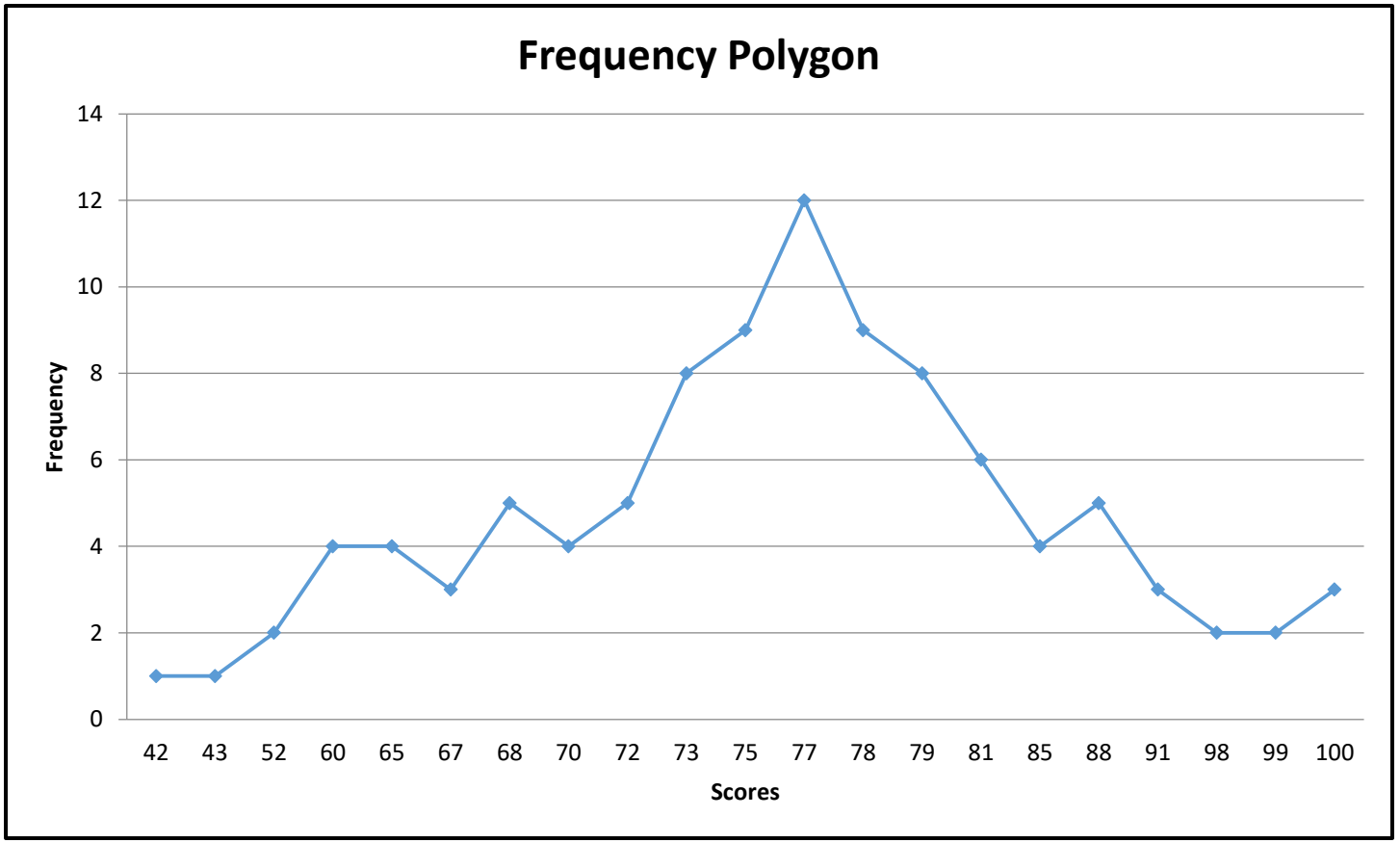
Frequency Distribution with Columns for Percentage, Cumulative Frequency, and Cumulative Percentage

Scores	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
100	2	8.33%	2	8.33%
99	1	4.17%	3	12.50%
98	1	4.17%	4	16.67%
91	1	4.17%	5	20.83%
88	1	4.17%	6	25.00%
85	1	4.17%	7	29.17%
81	1	4.17%	8	33.33%
79	1	4.17%	9	37.50%
78	1	4.17%	10	41.67%
77	2	8.33%	12	50.00%
75	1	4.17%	13	54.17%
73	1	4.17%	14	58.33%
72	2	8.33%	16	66.67%
70	1	4.17%	17	70.83%
68	1	4.17%	18	75.00%
67	1	4.17%	19	79.17%
65	1	4.17%	20	83.33%
60	1	4.17%	21	87.50%
52	1	4.17%	22	91.67%
43	1	4.17%	21	87.50%
42	1	4.17%	24	100.00%
N =	**24**	**100.00%**		

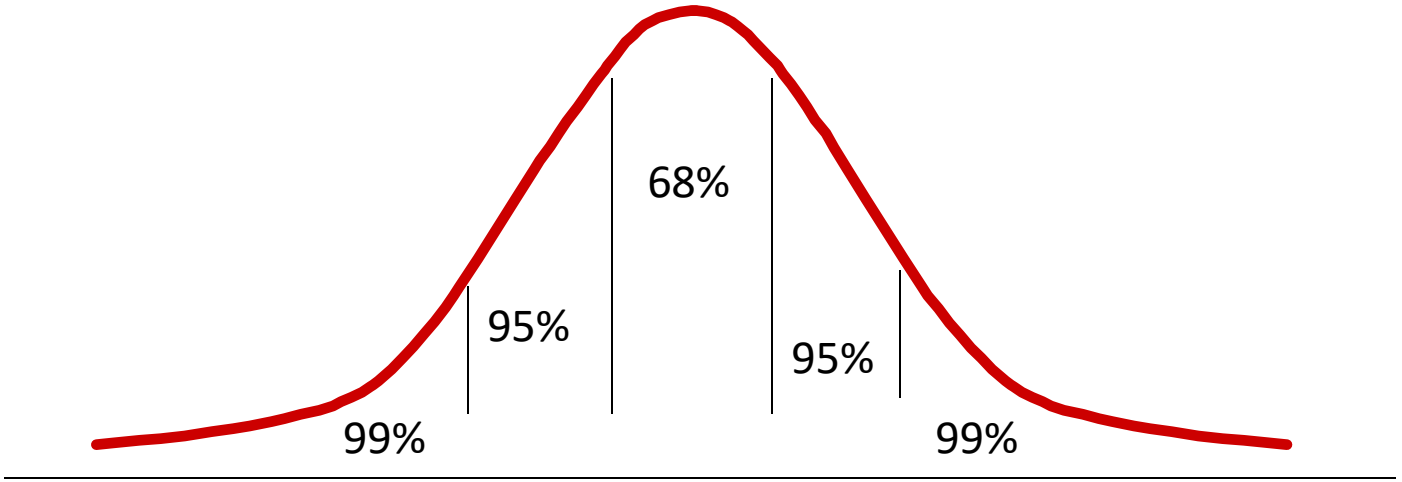
Creating a histogram (bar chart)



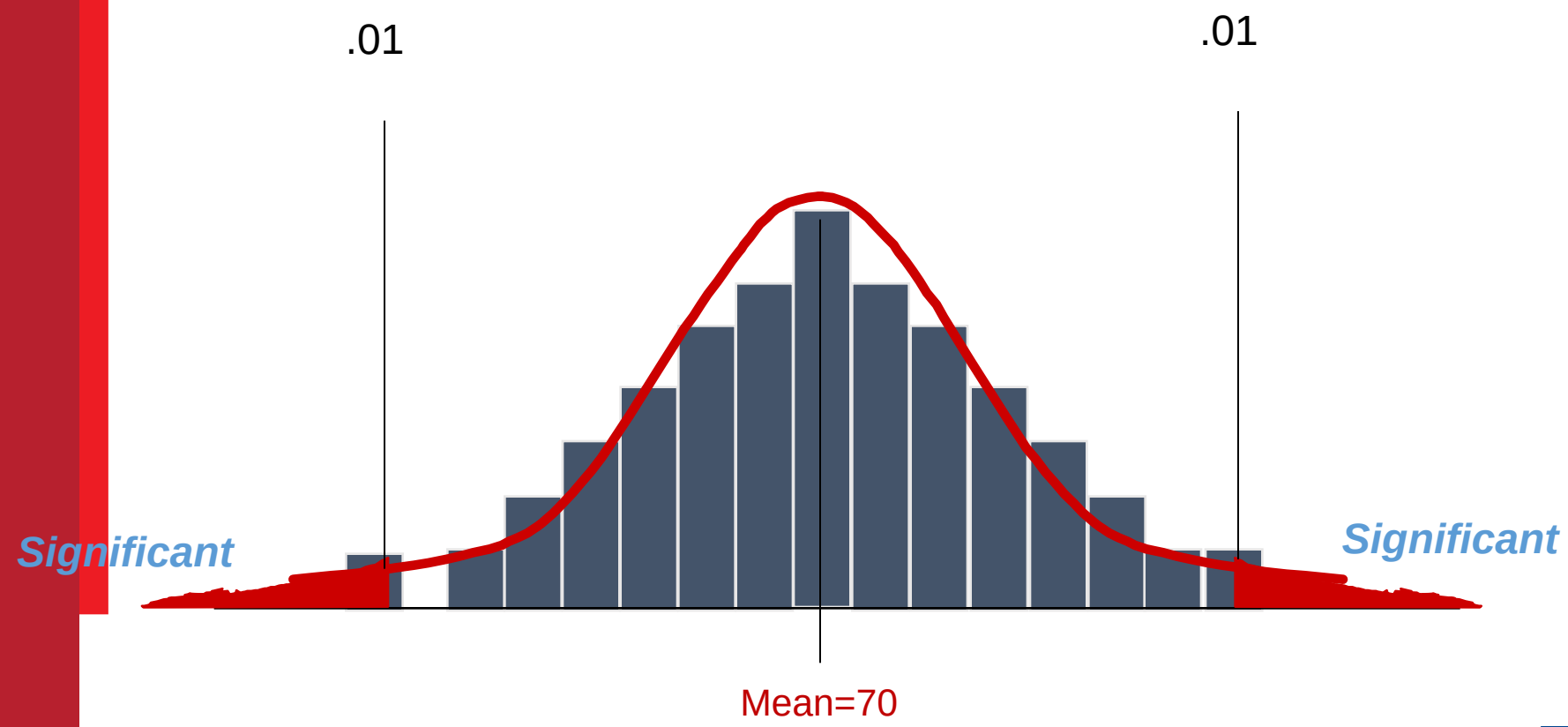
Creating a Frequency polygon



Normal Distribution



The Bell Curve



Central limit theorem

- In **probability theory**, the central limit theorem says that, under certain conditions, the sum of many independent identically-distributed **random variables**, when scaled appropriately, converges in distribution to a standard **normal distribution**.

Central Tendency

- These statistics answer the question: What is a typical score?
- The statistics provide information about the grouping of the numbers in a distribution by giving a single number that characterizes the entire distribution.
- Exactly what constitutes a “typical” score depends on the level of measurement and how the data will be used.
- For every distribution, three characteristic numbers can be identified:
  - Mode
  - Median
  - Mean

Measures of Central Tendency

- **Mean** - arithmetic average
  - –μ, Population; $\bar{x}$, sample
- **Median** - midpoint of the distribution
- **Mode** - the value that occurs most often

Mode Example

Find the score that occurs most frequently

98
88
81
74
72 ← Mode = 72
72
70
69
65
52

Median Example

Arrange in descending order and find the midpoint

Odd Number (N = 9)

98
88
81
74
72 ← Midpoint = 72
70
69
65
52

Even Number (N = 10)

98
88
81
74
72 ←
71 ← Midpoint = (72+71)/2 = 71.5
70
69
65
52

Different means

- Arithmetic Mean - the sum of all of the list divided by the number of items in the list

$$\bar{a} = \frac{a_1 + a_2 + a_3 + a_4 + \ldots + a_n}{n}$$

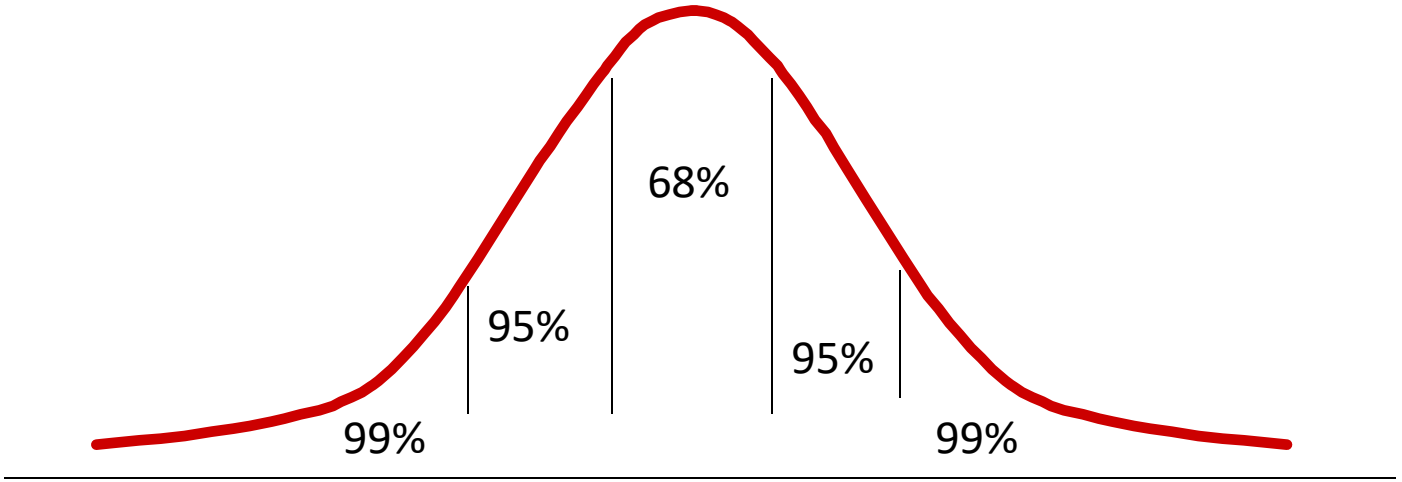
Arithmetic Mean Example

98
88
81
74
72
72
70
69
65
52

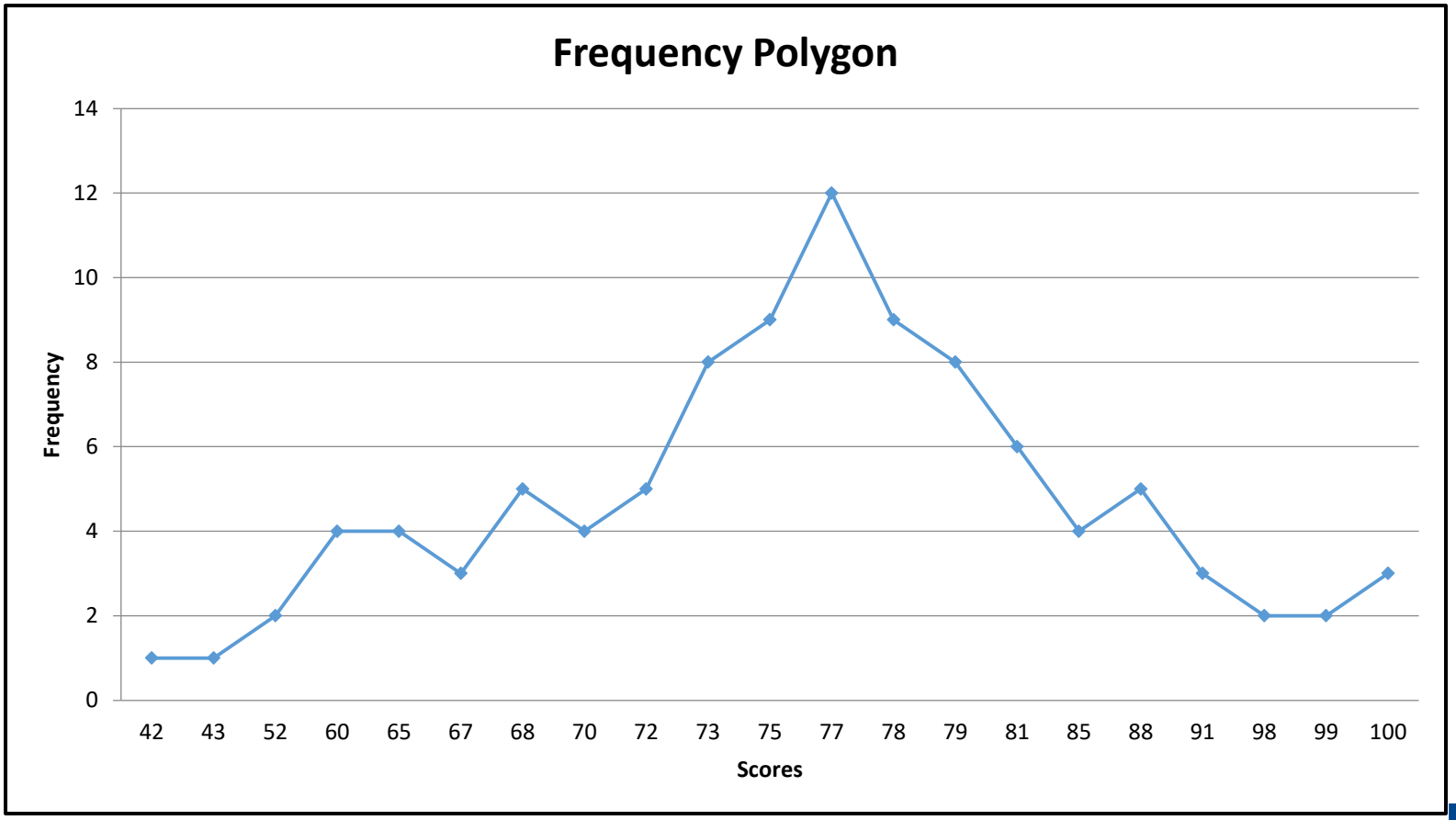
741

741\10 = 74.1

Normal Distribution

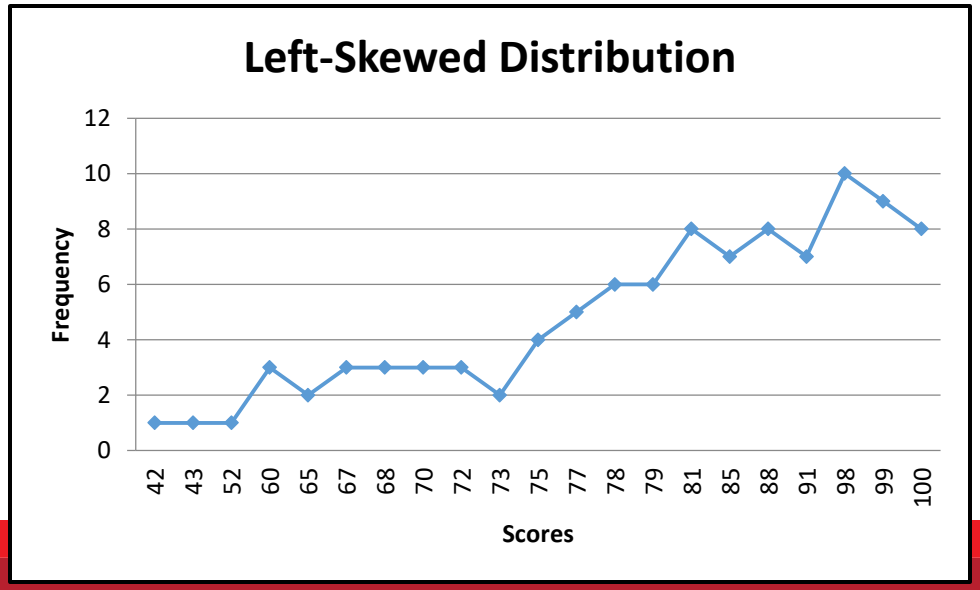


Frequency polygon of test score data



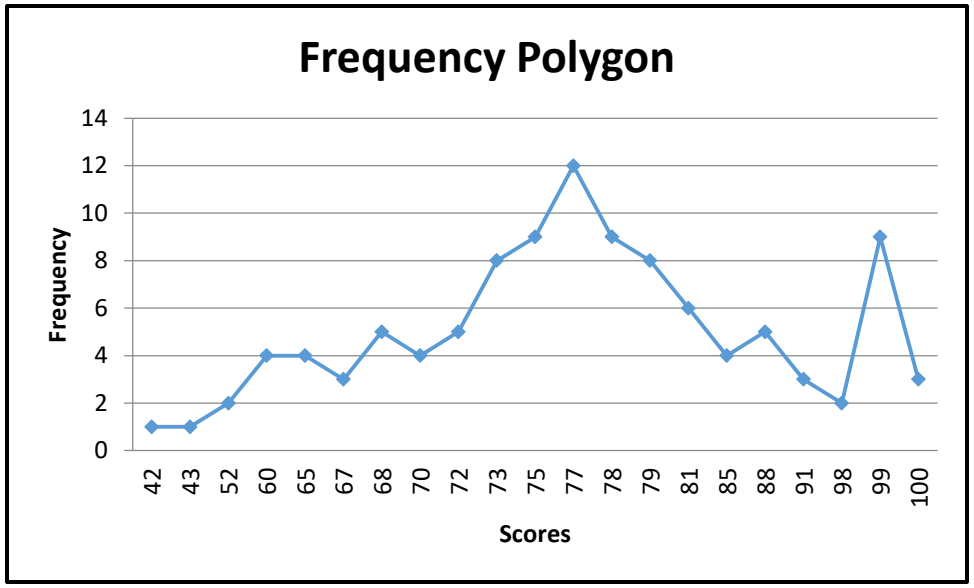
Skewness

- Refers to the concentration of scores around a particular point on the x-axis.
- If this concentration lies toward the low end of the scale, with the tail of the curve trailing off to the right, the curve is called a right skew.
- If the tail of the curve trails off to the left, it is a left skew.

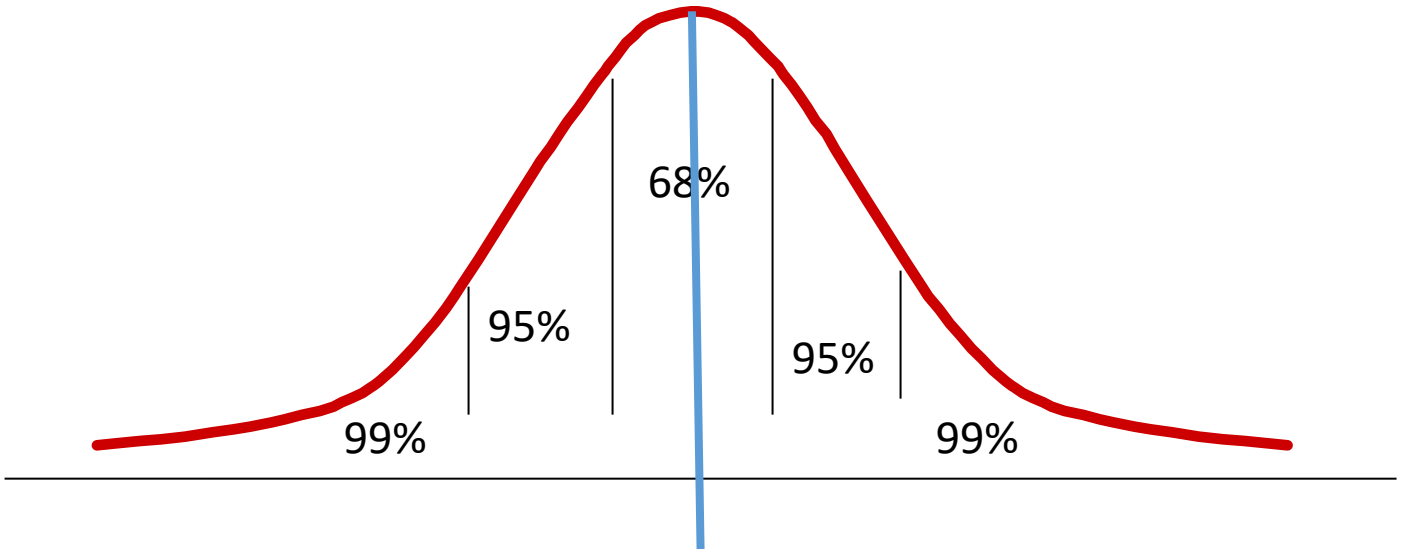


Skewness

- Skewness can occur when the frequency of just one score is clustered away from the mean.

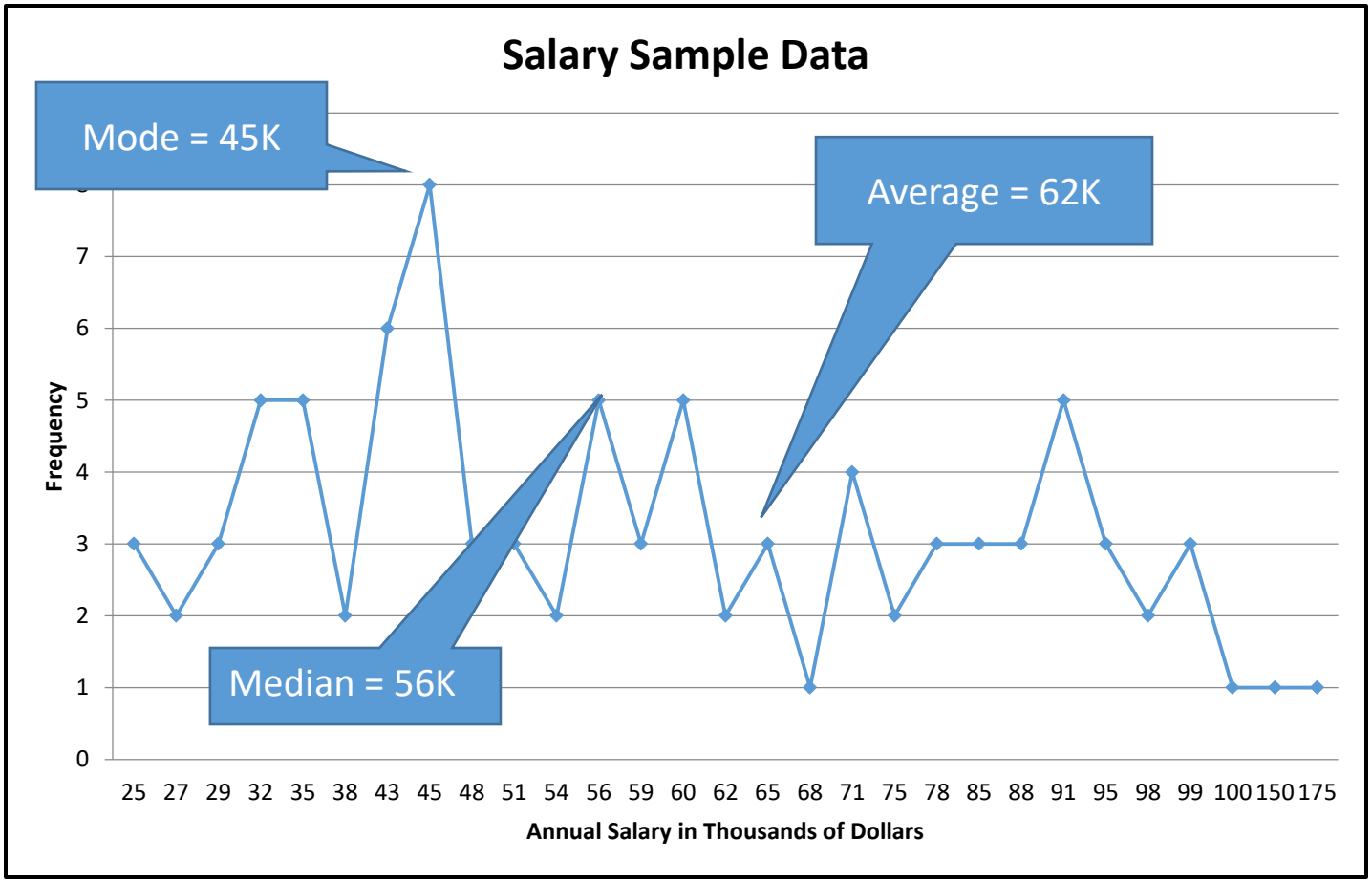


Normal Distribution



Mode = Median = Mean

When the Distribution may not be normal



Measures of Dispersion or Spread

- Range
- Variance
- Standard deviation

The Range as a Measure of Spread

- The range is the **distance** between the smallest and the largest value in the set.
- Range = largest value – smallest value

Group 1	Group 2
100, 100	91, 85
99, 98	81, 79
88, 77	78, 77
72, 68	73, 75
67, 52	72, 70
43, 42	65, 60

Range G1: 100 – 42 = 58

Range G2: 91 – 60 = 31

population Variance

$$S^2 = \frac{\Sigma(X_i - \overline{X})^2}{N}$$

Sample Variance

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n-1}$$

Variance

- A method of describing variation in a set of scores
- The higher the variance, the greater the variability and/or spread of scores

Variance Example

X		$\bar{X}$		$X - \bar{X}$		$X - \bar{X}^2$
98	- 74.1 =			23.90 =		571.21
88	- 74.1 =			13.90 =		193.21
81	- 74.1 =			6.90 =		47.61
74	- 74.1 =			-0.10 =		0.01
72	- 74.1 =			-2.10 =		4.41
72	- 74.1 =			-2.10 =		4.41
70	- 74.1 =			-4.10 =		16.81
69	- 74.1 =			-5.10 =		26.01
65	- 74.1 =			-9.10 =		82.81
52	- 74.1 =			-22.10 =		488.41
						1,434.90

Mean = 74.1

Population Variance (N)

1,434.90 \ 10 = 143.49

Sample Variance (n-1)

1,434.90 \ 9 = 159.43

Uses of the variance

- The variance is used in many higher-order calculations including:
  - T-test
  - Analysis of Variance (ANOVA)
  - Regression
- A variance value of zero indicates that all values within a set of numbers are identical
- All variances that are non-zero will be positive numbers. A large variance indicates that numbers in the set are far from the mean and each other, while a small variance indicates the opposite.

Standard Deviation

- Another method of describing variation in a set of scores
- The higher the standard deviation, the greater the variability and/or spread of scores

Sample Standard Deviation

$$S = \sqrt{\frac{\Sigma \left(X_i - \overline{X} \right)^2}{n-1}}$$

Standard Deviation Example

X		$\bar{X}$		X - $\bar{X}$		X – $\bar{X}^2$
98	-	74.1	=	23.90	=	571.21
88	-	74.1	=	13.90	=	193.21
81	-	74.1	=	6.90	=	47.61
74	-	74.1	=	-0.10	=	0.01
72	-	74.1	=	-2.10	=	4.41
72	-	74.1	=	-2.10	=	4.41
70	-	74.1	=	-4.10	=	16.81
69	-	74.1	=	-5.10	=	26.01
65	-	74.1	=	-9.10	=	82.81
52	-	74.1	=	-22.10	=	488.41
						1,434.90

Mean = 74.1

Population STD

1,434.90 \ 10 = 143.49

(SQRT) 143.49 = 11.98

Sample STD

1,434.90 \ 9 = 159.43

(SQRT) 159.43 = 12.63

Class assignment

- A survey was given to UNA students to find out how many hours per week they would listen to a student-run radio station. The sample responses were separated by gender. Determine the mean, range, variance, and standard deviation of each group.

Group A (Female)	Group B (Male)
15	30
25	15
12	21
7	12
3	26
32	20
17	5
16	24
9	18
24	10

Group one (females)

Range = 29

X	Mean	X-Mean	X-Mean^2			
15	16	-1	1			
25	16	9	81			
12	16	-4	16			
7	16	-9	81			
3	16	-13	169		718/9	79.78
32	16	16	256			
17	16	1	1		SQRT	8.93
16	16	0	0			
9	16	-7	49			
24	16	8	64			
16			**718**			

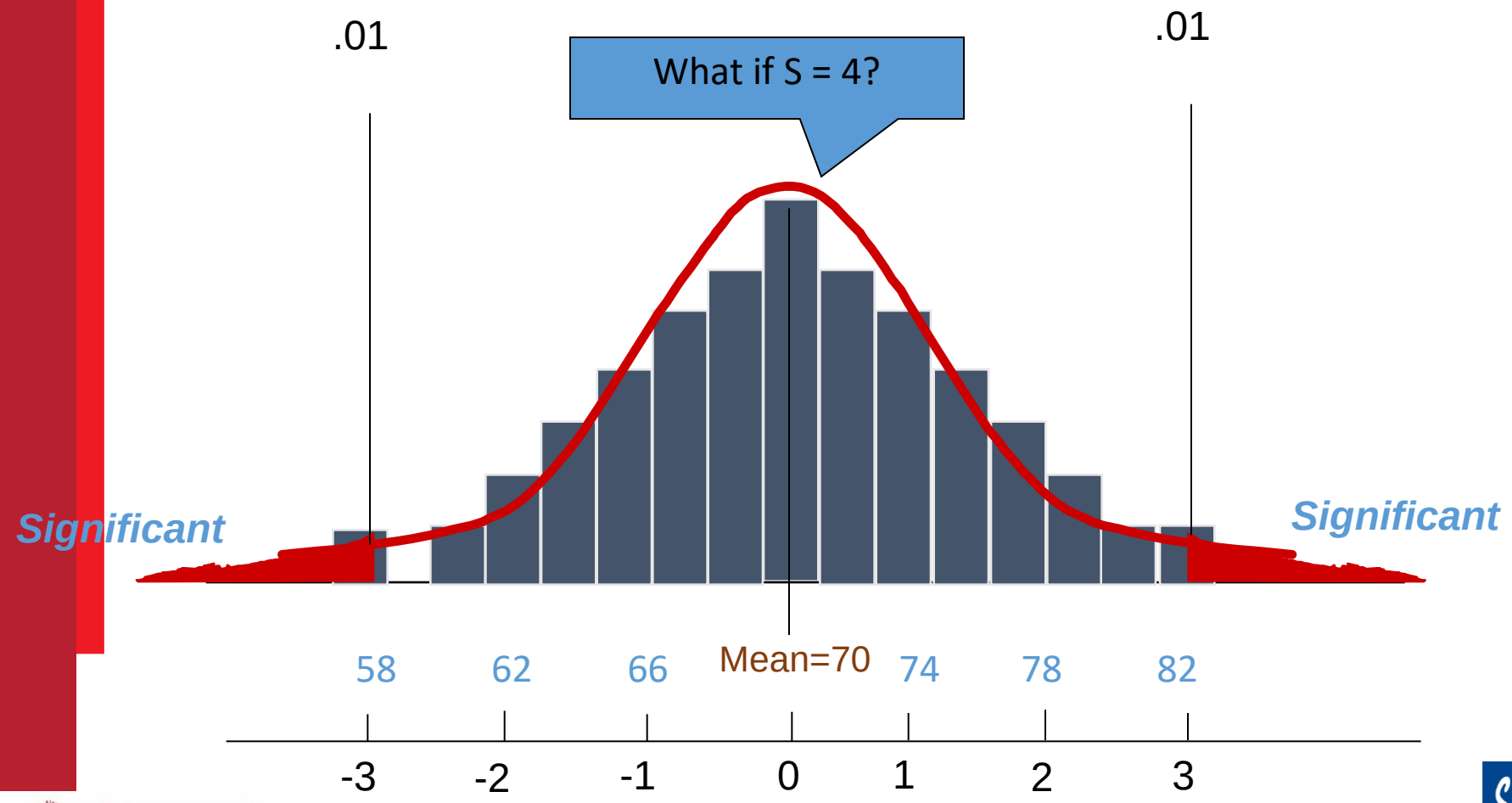
Group Two (Males)

X	Mean	X-Mean	X-Mean^2			
30	18	12	144			
15	18	-3	9			
21	18	3	9			
12	18	-6	36			
26	18	8	64			
20	18	2	4		535/9	59.44
5	18	-13	169			
24	18	6	36		SQRT	7.71
18	18	0	0			
10	18	-8	64			
18			535			

Results

Radio Listening Results				
Group	Average	Range	Variance	S
Females	16	29	79.78	8.93
Males	18	22	59.44	7.71

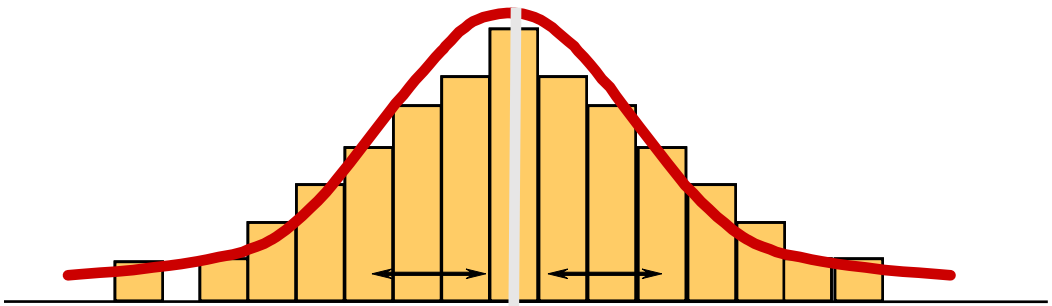
Standard Deviation on Bell Curve



How Variability and Standard Deviation Work…

Class A

100, 100
99, 98
88, 77
72, 68
67, 52
43, 42



Mean

Class B

91, 85
81, 79
78, 77
73, 75
72, 70
65, 60

Mean = 75.5

STD = 21.93

Mean = 75.5

STD = 8.42

How Do We Use This Stuff?

- The type of data determines what kind of measures you can use
- Higher order data can be used with higher order statistics

When scores don't compare

- A student takes the ACT test (11-36) and scores a 22…
- The same student takes the SAT (590-1,600) and scores a 750…
- The same student takes the TOFFEL (0-120) and scores a 92…
- How can we tell if the student did better/worse on one score in relation to the other scores?
- ANSWER: Standardize or Normalize the scores
- HOW: Z-Scores!

Z-Scores

- In statistics, the standard score is the (signed) number of standard deviations an observation or datum is above or below the mean.
- A positive standard score represents a datum above the mean, while a negative standard score represents a datum below the mean.
- It is a dimensionless quantity obtained by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This conversion process is called standardizing or normalizing.
- Standard scores are also called z-values, z-scores, normal scores, and standardized variables.

Z-score formula

$$z = \frac{X - \bar{X}}{S}$$

Z-Scores with positive numbers are above the mean while Z-Scores with negative numbers are below the mean.

Z-scores, cont.

- It is a little awkward in discussing a score or observation to have to say that it is “2 standard deviations above the mean” or “1.5 standard deviations below the mean.”

- To make it a little easier to pinpoint the location of a score in any distribution, the z-score was developed.

- The z-score is simply a way of telling how far a score is from the mean in standard deviation units.

Calculating the z-score

- If the observed value (individual score) = 9; the mean = 6; and the standard deviation = 2.68:

$$z = \frac{x - \bar{x}}{S} = \frac{9-6}{2.68} = \frac{3}{2.68} = 1.12$$

Z-Scores, cont.

- A z-score may also be used to find the location of a score that is a normally distributed variable.
- Using an example of a population of IQ test scores where the individual score = 80; population mean = 100; and the population standard deviation = 16…

$$z = \frac{X - \mu}{\partial} = \frac{80 - 100}{16} = \frac{-20}{16} = -1.25$$

Comparing z-scores

- Z-scores allow the researcher to make comparisons between different distributions.

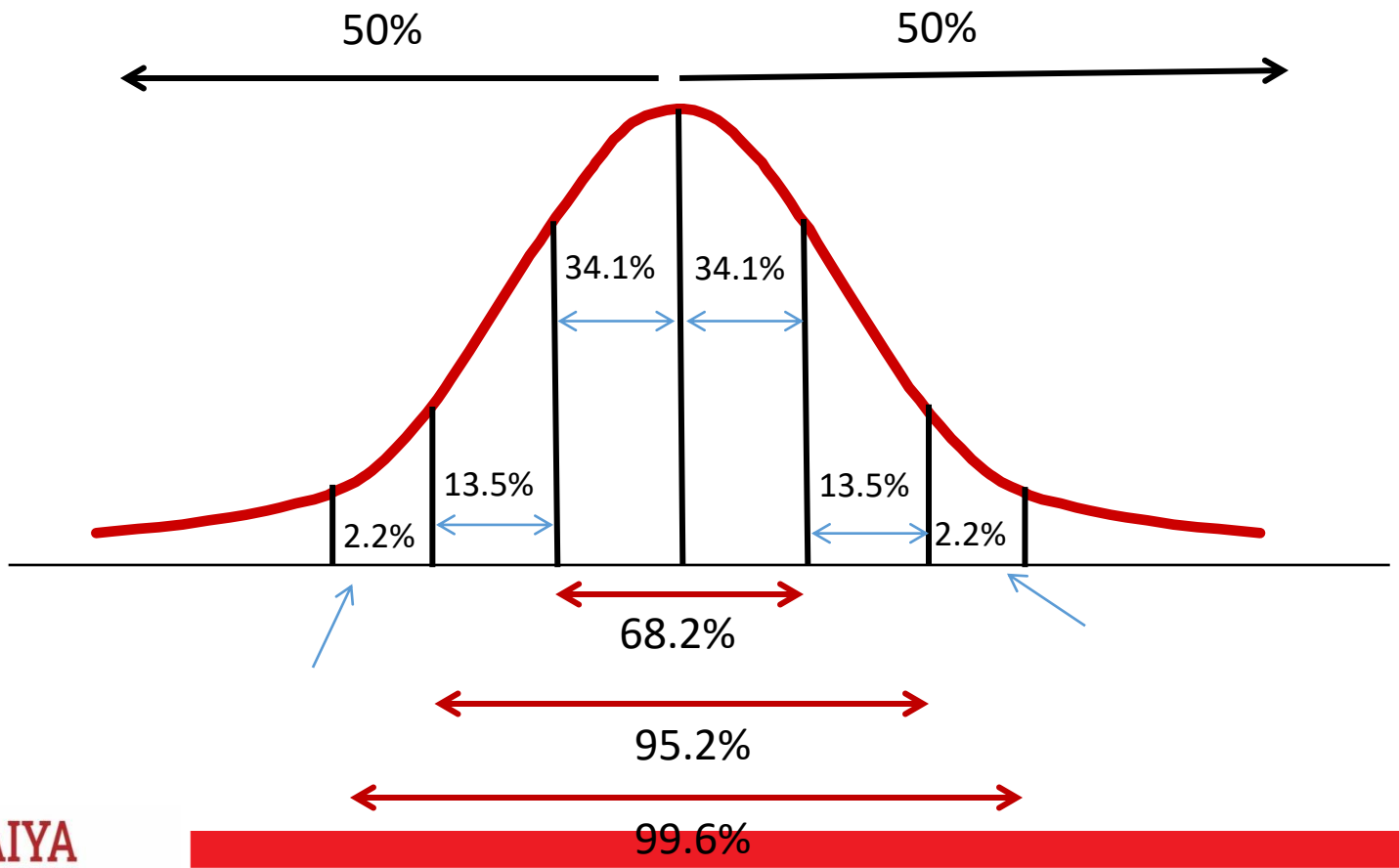
Mathematics	Natural Science	English
μ = 75	μ = 103	μ = 52
σ = 6	σ = 14	σ = 4
X = 78	X = 115	X = 57

Mathematics $$z = \frac{X - \mu}{\sigma} = \frac{78 - 75}{6} = \frac{3}{6} = 0.5$$

Natural Science $$z = \frac{115 - 103}{14} = \frac{12}{14} = 0.86$$

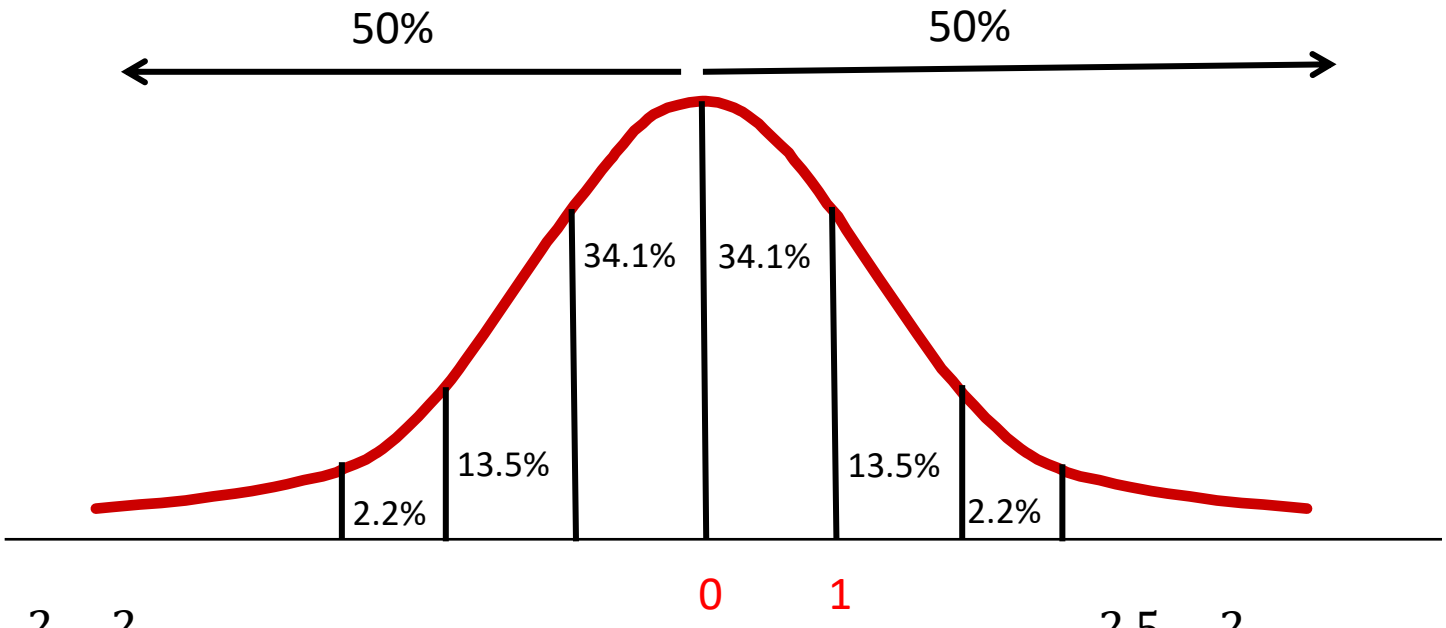
English $$z = \frac{57 - 52}{4} = \frac{5}{4} = 1.25$$

Area under the normal curve



Area under the normal curve

- TV viewing is normally distributed with a mean of 2 hours per day and standard deviation of 0.5. What proportion of the population watches between 2 and 2.5 hours of TV?

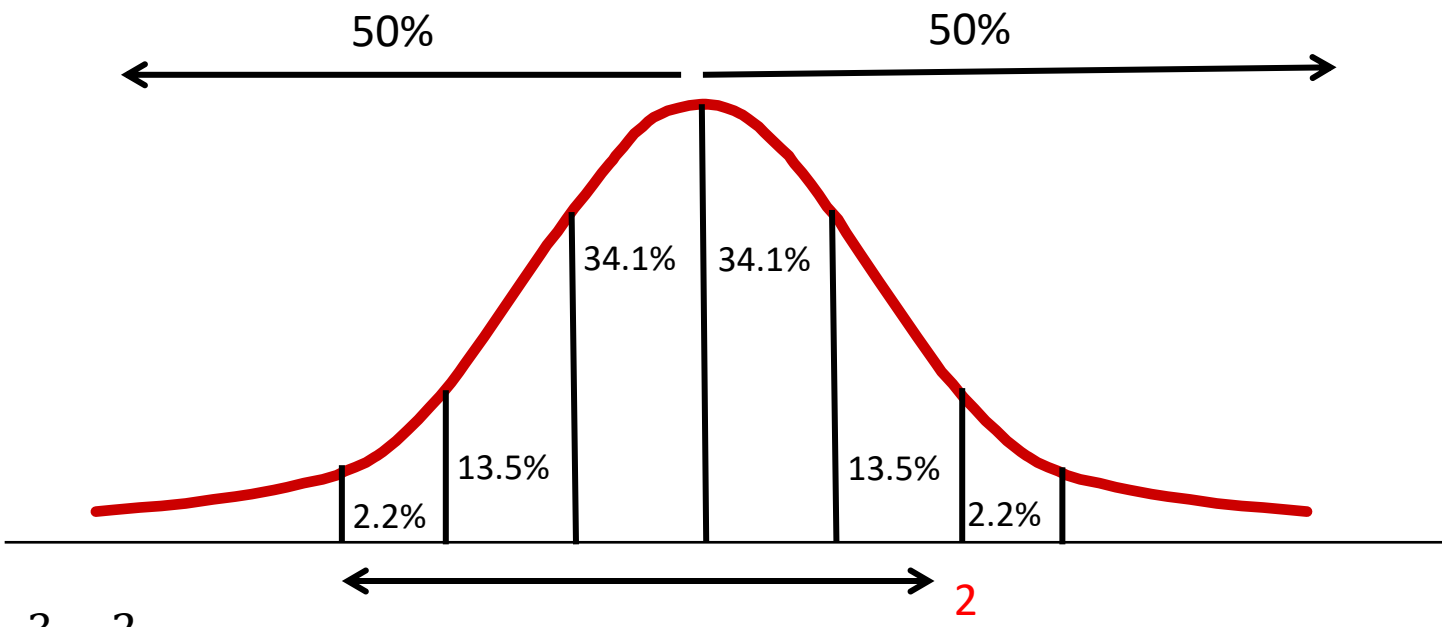


$$\frac{2-2}{.5}=0 \qquad \frac{2.5-2}{.5}=1$$

Answer = 34%

Area under the normal curve

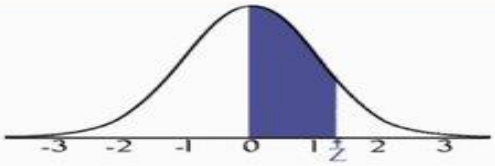
- How many watches more than 3 hours per day?



$$\frac{3-2}{.5} = 2$$

Answer = 2.2%

Area under the normal curve



STANDARD NORMAL TABLE (z)

Entries in the table give the area under the curve between the mean and z standard deviations above the mean. For example, for $z = 1.25$ the area under the curve between the mean (0) and z is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

Area under the normal curve

- Go to z-score table on-line
- Assume the z-score of a normally distributed variable is 2.67
- First find the row with 2.6 , then go to the column of .07 (second decimal place in z).
- At the intersection of the 2.6 row and the .07 column is the number .4962.
- Therefore, the area between the mean of the curve (midpoint) and a z-score of 2.67, is .4962 or approximately 46%

Final example

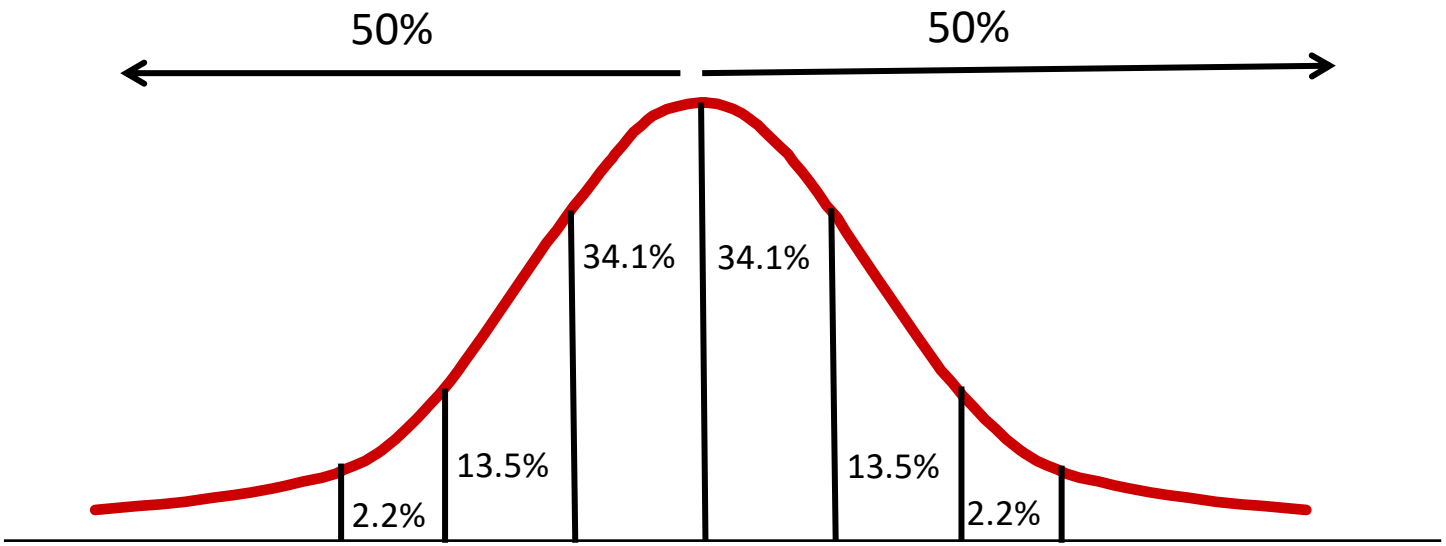
- What is the distance from the midpoint of a curve to the z-score of -1.32?

- Find the row 1.3
- Then find the column .02
- At the intersection of the row 1.3 and the column of .02 is .

- The distance from the midpoint of a curve to the z-score of -is 40.66%

- No matter if the z-score is negative or positive, the area is always positive.

The normal curve



Interpretation

- Interpretation
  - The process of drawing inferences from the analysis results.
  - Inferences drawn from interpretations lead to managerial implications and decisions.
  - From a management perspective, the qualitative meaning of the data and their managerial implications are an important aspect of the interpretation.

Inferential Statistics Provide Two Environments:

- Test for Difference – To test whether a significant difference exists between groups
- Tests for relationship – To test whether a significant relationship exist between a dependent (Y) and independent (X) variable/s
  - Relationship may also be predictive

Hypothesis Testing Using Basic Statistics

- Univariate Statistical Analysis
  - Tests of hypotheses involving only one variable.
- Bivariate Statistical Analysis
  - Tests of hypotheses involving two variables.
- Multivariate Statistical Analysis
  - Statistical analysis involving three or more variables or sets of variables.

Hypothesis Testing Procedure

- Process
  - The specifically stated hypothesis is derived from the research objectives.
  - A sample is obtained and the relevant variable is measured.
  - The measured sample value is compared to the value either stated explicitly or implied in the hypothesis.
    - If the value is consistent with the hypothesis, the hypothesis is supported.
    - If the value is not consistent with the hypothesis, the hypothesis is not supported.

Hypothesis Testing Procedure, Cont.

- H_0 – Null Hypothesis
  - "There is no significant difference/relationship between groups"
- H_a – Alternative Hypothesis
  - "There is a significant difference/relationship between groups"
- Always state your Hypothesis/es in the Null form
- The object of the research is to either reject or accept the Null Hypothesis/es

Significance Levels and p-values

- Significance Level
  - A critical probability associated with a statistical hypothesis test that indicates how likely an inference supporting a difference between an observed value and some statistical expectation is true.
  - The acceptable level of Type I error.
- p-value
  - Probability value, or the observed or computed significance level.
    - p-values are compared to significance levels to test hypotheses.

Experimental Research: What happens?

An hypothesis (educated guess) and then tested. Possible outcomes:

Something Not Will Happen / It Happens	Something Will Not Happen / It Does Not Happen
Something Will Happen / It Happens	Something Will Happen / It Does Not Happen

Type I and Type II Errors

- Type I Error
  - An error caused by rejecting the null hypothesis when it should be accepted (false positive).
  - Has a probability of alpha (α).
  - Practically, a Type I error occurs when the researcher concludes that a relationship or difference exists in the population when in reality it does not exist.
  - “There really are no monsters under the bed.”

Type I and Type II Errors (cont'd)

- Type II Error
  - An error caused by failing to reject the null hypothesis when the hypothesis should be rejected (false negative).
  - Has a probability of beta (β).
  - Practically, a Type II error occurs when a researcher concludes that no relationship or difference exists when in fact one does exist.
  - "There really are monsters under the bed."

Type I and II Errors and Fire Alarms?

	FIRE	NO FIRE
NO ALARM	TYPE I	NO ERROR
Alarm	NO ERROR	TYPE II

	H_0 is False	H_0 is True
ACCEPT H_0	TYPE I	NO ERROR
REJECT H_0	NO ERROR	TYPE II

Recapitulation of the Research Process

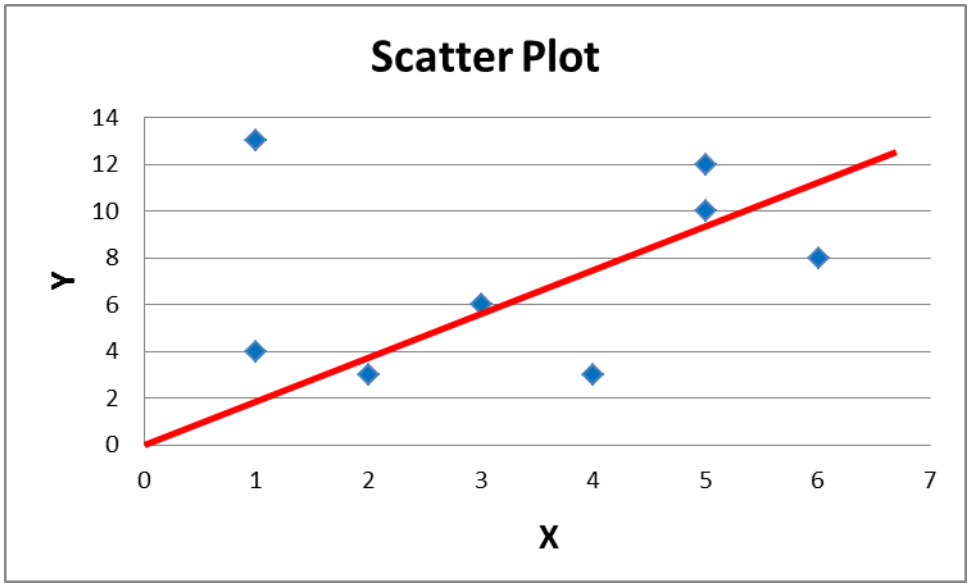
- Collect Data
- Run Descriptive Statistics
- Develop Null Hypothesis/es
- Determine the Type of Data
- Determine the Type of Test/s (based on type of data)
- If test produces a significant p-value, REJECT the Null Hypothesis. If the test does not produce a significant p-value, ACCEPT the Null Hypothesis.
- Remember that, due to error, statistical tests only support hypotheses and can NOT prove a phenomenon

Data Type v. Statistics Used

Data Type	Statistics Used
Nominal	Frequency, percentages, modes
Ordinal	Frequency, percentages, modes, median, range, percentile, ranking
Interval	Frequency, percentages, modes, median, range, percentile, ranking average, variance, SD, t-tests, ANOVAs, Pearson Rs, regression
Ratio	Frequency, percentages, modes, median, range, percentile, ranking average, variance, SD, t-tests, ratios, ANOVAs, Pearson Rs, regression

Pearson R Correlation Coefficient

X	Y
1	4
3	6
5	10
5	12
1	13
2	3
4	3
6	8



Pearson R Correlation Coefficient

A measure of how well a linear equation describes the relation between two variables *X* and *Y* measured on the same object

	X	Y	$x - \bar{x}$	$y - \bar{y}$	xy	x^2	y^2
	1	4	-3	-5	15	9	25
	3	6	-1	-3	3	1	9
	5	10	1	1	1	1	1
	5	12	1	3	3	1	9
	1	13	2	4	8	4	16
Total	20	45	0	0	30	16	60
Mean	4	9	0	0	6		

Calculation of Pearson R

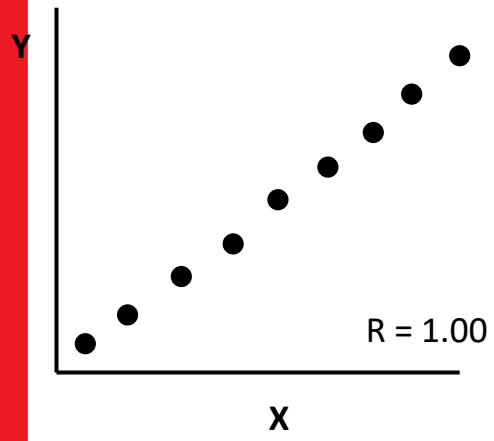
$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

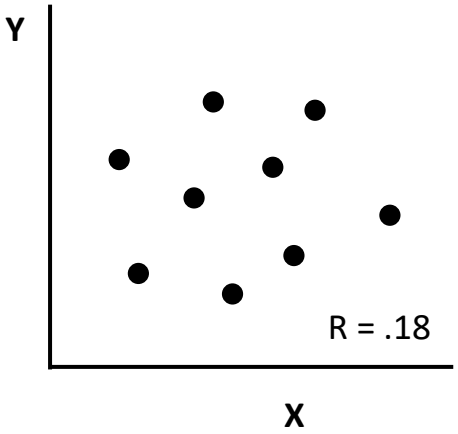
$$r = \frac{30}{\sqrt{(16)(60)}} = \frac{30}{\sqrt{960}} = \frac{30}{30.984} = 0.968$$

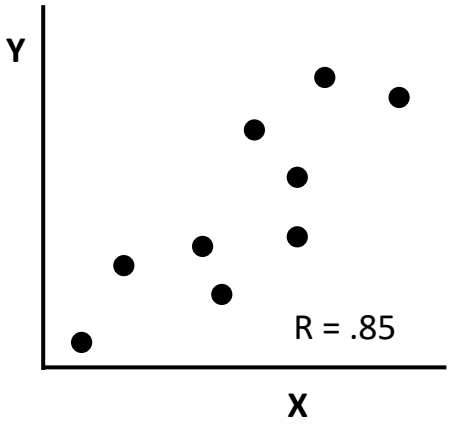
Alternative Formula

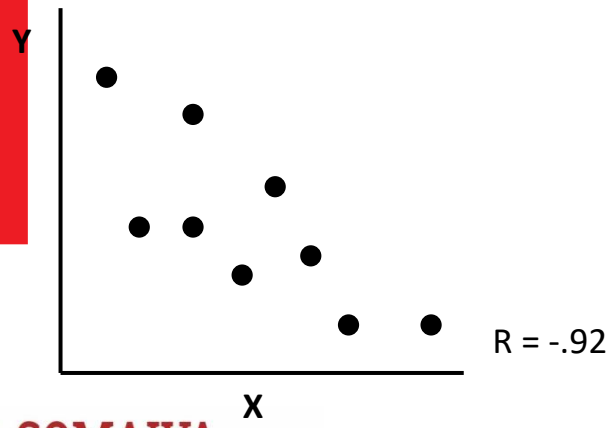
$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{N}} \sqrt{\sum Y^2 - \frac{(\sum Y)^2}{N}}}$$

How Can R's Be Used?









R's of 1.00 or -1.00 are perfect correlations

The closer R comes to 1, the more related the X and Y scores are to each other

R-Squared is an important statistic that indicates the variance of Y that is attributed to by the variance of X (.04, .73)

Concept of degrees of freedom

Choosing Classes for Academic Program

Class I
Class D
Class M
Class G
Class L
Class A
Class F
Class J
Class B
Class P
Class N
Class K
Class H
Class C
Class E
Class O

16 Classes to Graduate

Degrees of Freedom

- The number of values in a study that are free to vary.
- A data set contains a number of observations, say, *n*. They constitute *n* individual pieces of information. These pieces of information can be used either to estimate parameters or variability. In general, each item being estimated costs one degree of freedom. The remaining degrees of freedom are used to estimate variability. All we have to do is count properly.
- **A single sample:** There are *n* observations. There's one parameter (the mean) that needs to be estimated. That leaves *n-1* degrees of freedom for estimating variability.
- **Two samples:** There are n_1+n_2 observations. There are two means to be estimated. That leaves n_1+n_2-2 degrees of freedom for estimating variability.

Testing for Significant Difference

- Testing for significant difference is a type of inferential statistic
- One may test difference based on any type of data
- Determining what type of test to use is based on what type of data are to be tested.

Testing Difference

- Testing difference of gender to favorite form of media
- Gender: M or F
- Media: Newspaper, Radio, TV, Internet
- Data: Nominal
- Test: Chi Square

- Testing difference of gender to answers on a Likert scale
- Gender: M or F
- Likert Scale: 1, 2, 3, 4, 5
- Data: Interval
- Test: t-test

What is a Null Hypothesis?

- A type of hypothesis used in statistics that proposes that no statistical significance exists in a set of given observations.
- The null hypothesis attempts to show that no variation exists between variables, or that a single variable is no different than zero.
- It is presumed to be true until statistical evidence nullifies it for an alternative hypothesis.

Examples

- **Example 1:** Three unrelated groups of people choose what they believe to be the best color scheme for a given website.
- **The null hypothesis is:** There is no difference between color scheme choice and type of group
- **Example 2:** Males and Females rate their level of satisfaction to a magazine using a 1-5 scale
- **The null hypothesis is:** There is no difference between satisfaction level and gender

Chi Square

A chi square (X^2) statistic is used to investigate whether distributions of categorical (i.e. nominal/ordinal) variables differ from one another.

General Notation for a chi square 2x2 Contingency Table

Variable 1

Variable 2	Data Type 1	Data Type 2	Totals
Category 1	a	b	a+b
Category 2	c	d	c+d
Total	a+c	b+d	a+b+c+d

$$x^2 = \frac{(ad - bc)^2(a + b + c + d)}{(a + b)(c + d)(b + d)(a + c)}$$

Chi square Steps

- Collect observed frequency data
- Calculate expected frequency data
- Determine Degrees of Freedom
- Calculate the chi square
- If the chi square statistic exceeds the probability or table value (based upon a p-value of x and n degrees of freedom) the null hypothesis should be rejected.

Two questions from a questionnaire…

- Do you like the television program? (Yes or No)
- What is your gender? (Male or Female)

Gender and Choice Preference

H_0: There is no difference between gender and choice

Actual Data

	Male	Female	Total
Like	36	14	50
Dislike	30	25	55
Total	66	39	105

Column Total

Row Total

Grand Total

To find the expected frequencies, assume independence of the rows and columns. Multiply the row total to the column total and divide by grand total

$$ef = \frac{rt * ct}{gt} OR \frac{50 * 66}{105} = 31.43$$

Chi square

Expected Frequencies

	Male	Female	Total
Like	31.43	18.58	**50.01**
Dislike	34.58	20.43	**55.01**
Total	**66.01**	**39.01**	**105.02**

The number of degrees of freedom is calculated for an x-by-y table as (x-1) (y-1), so in this case (2-1) (2-1) = 1*1 = 1. The degrees of freedom is 1.

Chi square Calculations

O	E	O-E	$(O-E)^2/E$
36	31.43	4.57	.67
14	18.58	-4.58	1.13
30	34.58	-4.58	.61
25	20.43	4.57	1.03

Chi square observed statistic = 3.44

Chi square

Probability Level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.51

Chi Square (Observed statistic) = 3.44
Probability Level (df=1 and .05) = 3.841 (Table Value)
So, Chi Square statistic < Probability Level (Table Value)
Accept Null Hypothesis

Check Critical Value Table for Chi Square Distribution on Page 448 of text

Results of Chi square Test

There is no significant difference between product choice and gender.

Chi square Test for Independence

- Involves observations greater than 2x2
- Same process for the Chi square test
- Indicates independence or dependence of three or more variables…but that is all it tells

Two Questions…

- What is your favorite color scheme for the website? (Blue, Red, or Green)
- There are three groups (Rock music, Country music, jazz music)

Chi Square

H_0: Group is independent of color choice

Actual Data

	Blue	Red	Green	Total
Rock	11	6	4	**21**
Jazz	12	7	7	**26**
Country	7	7	14	**28**
Total	**30**	**20**	**25**	**75**

Row Total

Column Total

Grand Total

To find the expected frequencies, assume independence of the rows and columns. Multiply the row total to the column total and divide by grand total

$$ef = \frac{rt * ct}{gt} OR \frac{21*30}{75} = 8.4$$

Chi Square

Expected Frequencies

	Blue	Red	Green	Total
Rock	8.4	5.6	7.0	**21**
Jazz	10.4	6.9	8.7	**26**
Country	11.2	7.5	9.3	**28**
Total	**30**	**20**	**25**	**75**

The number of degrees of freedom is calculated for an x-by-y table as (x-1) (y-1), so in this case (3-1) (3-1) = 2*2 = 4. The degrees of freedom is 4.

Chi Square Calculations

O	E	O-E	$(O-E)^2/E$
11	8.4	2.6	.805
6	5.6	.4	.029
4	7	3	1.286
12	10.4	1.6	.246
7	6.9	.1	.001
7	8.7	1.7	.332
7	11.2	4.2	1.575
7	7.5	.5	.033
14	9.3	4.7	2.375

Chi Square observed statistic = 6.682

Chi Square Calculations, cont.

Probability Level (alpha)

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.51

Chi Square (Observed statistic) = 6.682

Probability Level (df=4 and .05) = 9.488 (Table Value)

So, Chi Square observed statistic < Probability level (table value)

Accept Null Hypothesis

Check Critical Value Table for Chi Square Distribution on page 448 of text

Chi square Test Results

There is no significant difference between group and choice, therefore, group and choice are independent of each other.

What's the Connection?

$$\longleftrightarrow \quad t = \frac{\bar{x}_1 - \bar{x}_2}{S_{x_1 - x_2}}$$

Gosset, Beer, and Statistics…

William S. Gosset (1876-1937) was a famous statistician who worked for Guiness. He was a friend and colleague of Karl Pearson and the two wrote many statistical papers together. Statistics, during that time involved very large samples, and Gosset needed something to test difference between smaller samples.

Gosset discovered a new statistic and wanted to write about it. However, Guiness had a bad experience with publishing when another academic article caused the beer company to lose some trade secrets.

Because Gosset knew this statistic would be helpful to all, he published it under the pseudonym of "Student."

William Gosset

The t test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{x_1 - x_2}}$$

$\bar{x}_1 =$ Mean for group 1

$\bar{x}_2 =$ Mean for group 2

$S_{\bar{x}_1 - \bar{x}_2} =$ Pooled, or combined, standard error of difference between means

The pooled estimate of the standard error is a better estimate of the standard error than one based of independent samples.

Uses of the t test

- Assesses whether the mean of a group of scores is *statistically* different from the population (One sample t test)
- Assesses whether the means of two groups of scores are *statistically* different from each other (Two sample t test)
- Cannot be used with more than two samples (ANOVA)

Sample Data

Group 1	Group 2
$\bar{x}_1 = 16.5$	$\bar{x}_2 = 12.2$
$S_1 = 2.1$	$S_2 = 2.6$
$n_1 = 21$	$n_2 = 14$

Null Hypothesis

$$H_0 : \mu_1 = \mu_2$$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{x_1 - x_2}}$$

Step 1: Pooled Estimate of the Standard Error

$$S_{\bar{x}_1-\bar{x}_2} - \sqrt{\left(\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$S_1^2 =$ Variance of group 1

$S_2^2 =$ Variance of group 2

$n_1 =$ Sample size of group 1

$n_2 =$ Sample size of group 2

Group 1	Group 2
$\bar{x}_1 = 16.5$	$\bar{x}_2 = 12.2$
$S_1 = 2.1$	$S_2 = 2.6$
$n_1 = 21$	$n_2 = 14$

Step 1: Calculating the Pooled Estimate of the Standard Error

$$S_{\bar{x}_1-\bar{x}_2} - \sqrt{\left(\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

$$S_{\bar{x}_1-\bar{x}_2} - \sqrt{\left(\frac{(20)(2.1)^2+(13)(2.6)^2}{33}\right)\left(\frac{1}{21}+\frac{1}{14}\right)}$$

=0.797

Step 2: Calculate the t-statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S_{x_1 - x_2}}$$

$$t = \frac{16.5 - 12.2}{0.797} = \frac{4.3}{0.797} = 5.395$$

Step 3: Calculate Degrees of Freedom

- In a test of two means, the degrees of freedom are calculated: d.f. =n-k
- n = total for both groups 1 and 2 (35)
- k = number of groups
- Therefore, d.f. = 33 (21+14-2)
- Go to the tabled values of the t-distribution on website. See if the observed statistic of 5.395 surpasses the table value on the chart given 33 d.f. and a .05 significance level

Step 3: Compare Critical Value to Observed Value

Observed statistic= 5.39

Df	0.10	0.05	0.02	0.01
30	1.697	2.042	2.457	2.750
31	1.659	2.040	2.453	2.744
32	1.694	2.037	2.449	2.738
33	1.692	2.035	2.445	2.733
34	1.691	2.032	2.441	2.728

If Observed statistic exceeds Table Value:

Reject H_0

So What Does Rejecting the Null Tell Us?

Group 1	Group 2
$\bar{x}_1 = 16.5$	$\bar{x}_2 = 12.2$
$S_1 = 2.1$	$S_2 = 2.6$
$n_1 = 21$	$n_2 = 14$

Based on the .05 level of statistical significance, Group 1 scored significantly higher than Group 2

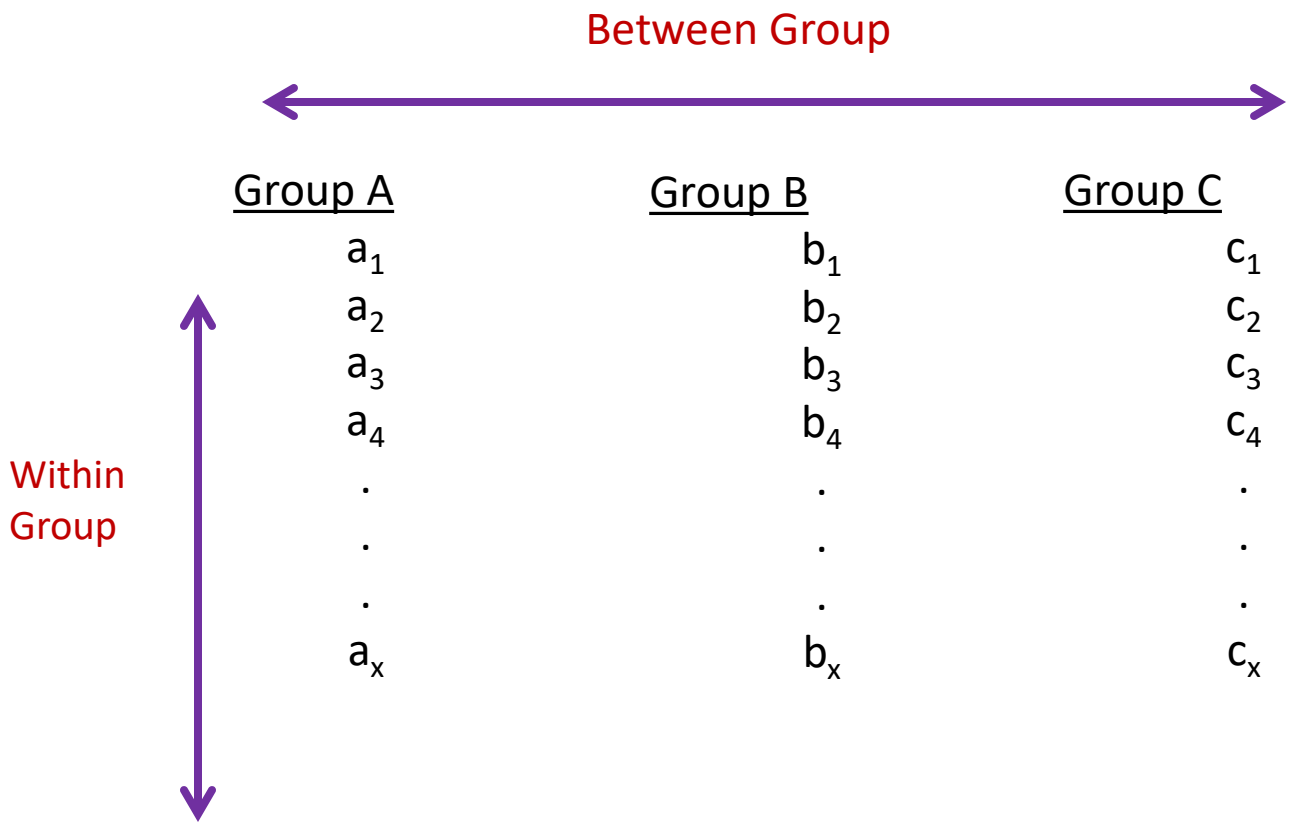
ANOVA Definition

- In statistics, **analysis of variance** (**ANOVA**) is a collection of statistical models, and their associated procedures, in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation.
- In its simplest form ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes *t*-test to more than two groups.
- Doing multiple two-sample t-tests would result in an increased chance of committing a type I error. For this reason, ANOVAs are useful in comparing two, three or more means.

Variability is the Key to ANOVA

- Between group variability and within group variability are both components of the total variability in the combined distributions
- When we compute between and within group variability we partition the total variability into the two components.
- Therefore: Between variability + Within variability = Total variability

Visual of Between and Within Group Variability



ANOVA Hypothesis Testing

- Tests hypotheses that involve comparisons of two or more populations
- The overall ANOVA test will indicate if a difference exists between any of the groups
- However, the test will not specify which groups are different
- Therefore, the research hypothesis will state that there are no significant difference between any of the groups

$$H_0: \mu_1 = \mu_2 = \mu_3$$

ANOVA Assumptions

- Random sampling of the source population (cannot test)
- Independent measures within each sample, yielding uncorrelated response residuals (cannot test)
- Homogeneous variance across all the sampled populations (can test)
  - Ratio of the largest to smallest variance (F-ratio)
  - Compare F-ratio to the F-Max table
  - If F-ratio exceeds table value, variance are not equal
- Response residuals do not deviate from a normal distribution (can test)
  - Run a normal test of data by group

ANOVA Computations Table

	SS	df	MF	F
Between (Model)	SS(B)	k-1	$\frac{SS(B)}{k-1}$	$\frac{MS(B)}{MS(W)}$
Within (Error)	SS(W)	N-k	$\frac{SS(W)}{N-k}$	
Total	SS(W)+SS(B)	N-1		

ANOVA Data

Group 1	Group 2	Group 3
5	3	1
2	3	0
5	0	1
4	2	2
2	2	1
$\Sigma x_1=18$	$\Sigma x_2=10$	$\Sigma x_3=5$
$\Sigma x^2_1=74$	$\Sigma x^2_2=26$	$\Sigma x^2_3=7$

Calculating Total Sum of Squares

$$SS_T = \sum x^2{}_{\mathrm{T}} - \left(\frac{\sum x_T}{N_T}\right)^2$$

$$SS_T = 107 - \frac{(33)^2}{15}$$

$$SS_T = 107 - \frac{1089}{15} = 107 - 72.6 = \mathbf{34.4}$$

Calculating Sum of Squares Within

$$SS_W = \left(\Sigma x^2{}_1 - \frac{(\sum x_1)^2}{n_1}\right) + \left(\Sigma x^2{}_2 - \frac{(\sum x_1)^2}{n_2}\right) + \left(\Sigma x^2{}_1 - \frac{(\sum x_1)^2}{n_1}\right)$$

$$SS_W = \left(74 - \frac{(18)^2}{5}\right) + \left(26 - \frac{(10)^2}{5}\right) + \left(7 - \frac{(5)^2}{5}\right)$$

$$SS_w = \left(74 - \frac{324}{5}\right) + \left(26 - \frac{100}{5}\right) + \left(7 - \frac{25}{5}\right)$$

$$SS_w = (74 - 64.8) + (26 - 20) + (7 - 5)$$

$$SS_W = 9.2 + 6 + 2 = \mathbf{17.2}$$

Calculating Sum of Squares Between

$$SS_B = \frac{(\sum x_1)^2}{n_1} + \frac{(\sum x_2)^2}{n_2} + \frac{(\sum x_3)^2}{n_3} - \frac{(\sum X_T)^2}{N_T}$$

$$SS_B = \frac{(18)^2}{5} + \frac{(10)^2}{5} + \frac{(5)^2}{5} - \frac{(33)^2}{15}$$

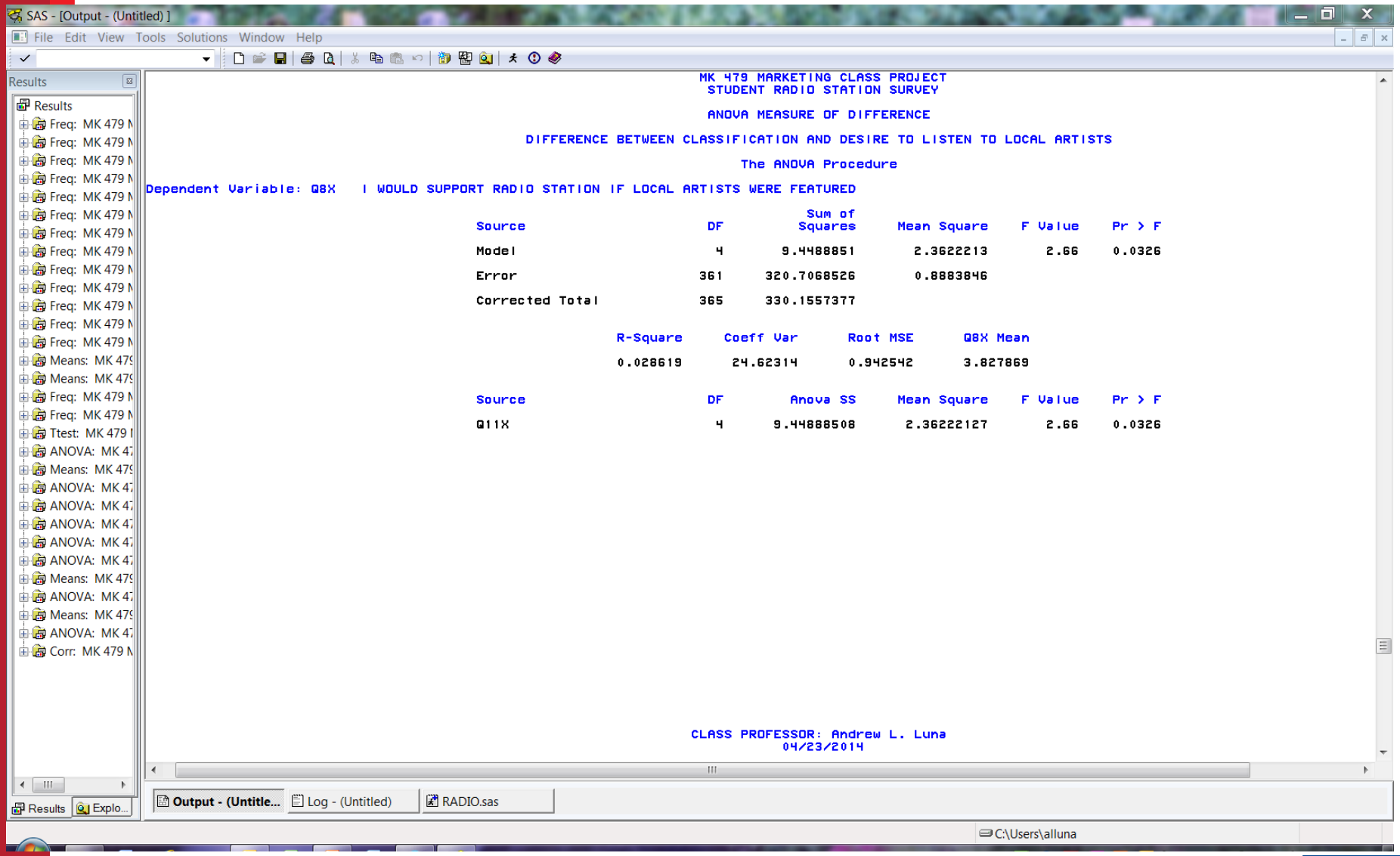
$$SS_B = \frac{324}{5} + \frac{100}{5} + \frac{25}{5} - \frac{1089}{15}$$

$$SS_B = 64.8 + 20 + 5 - 72.6 = \mathbf{17.2}$$

Complete the ANOVA Table

	SS	df	MF	F
Between (Model)	SS(B) **17.2**	k-1 **2**	$\frac{SS(B)}{k-1}$ **8.6**	$\frac{MS(B)}{MS(W)}$ **6**
Within (Error)	SS(W) **17.2**	N-k **12**	$\frac{SS(W)}{N-k}$ **1.43**	
Total	SS(W)+SS(B) **34.4**	N-1 **14**		

If the F statistic is higher than the F probability table, reject the null hypothesis


SAS - [Output - (Untitled)]
File Edit View Tools Solutions Window Help
Results
Results
Freq: MK 479 N
Means: MK 479
ANOVA: MK 47
Ttest: MK 479 I
Corr: MK 479 N
Results
Explo...
MK 479 MARKETING CLASS PROJECT
STUDENT RADIO STATION SURVEY
ANOVA MEASURE OF DIFFERENCE
DIFFERENCE BETWEEN CLASSIFICATION AND DESIRE TO LISTEN TO LOCAL ARTISTS
The ANOVA Procedure
Dependent Variable: Q8X I WOULD SUPPORT RADIO STATION IF LOCAL ARTISTS WERE FEATURED
Source DF Sum of Squares Mean Square F Value Pr > F
Model 4 9.4488851 2.3622213 2.66 0.0326
Error 361 320.7068526 0.8883846
Corrected Total 365 330.1557377
R-Square Coeff Var Root MSE Q8X Mean
0.028619 24.62314 0.942542 3.827869
Source DF Anova SS Mean Square F Value Pr > F
Q11X 4 9.44888508 2.36222127 2.66 0.0326
CLASS PROFESSOR: Andrew L. Luna
04/23/2014
Output - (Untitle...
Log - (Untitled)
RADIO.sas
C:\Users\alluna

You Are Not Done Yet!!!

- If the ANOVA test determines a difference exists, it will not indicate where the difference is located
- You must run a follow-up test to determine where the differences may be

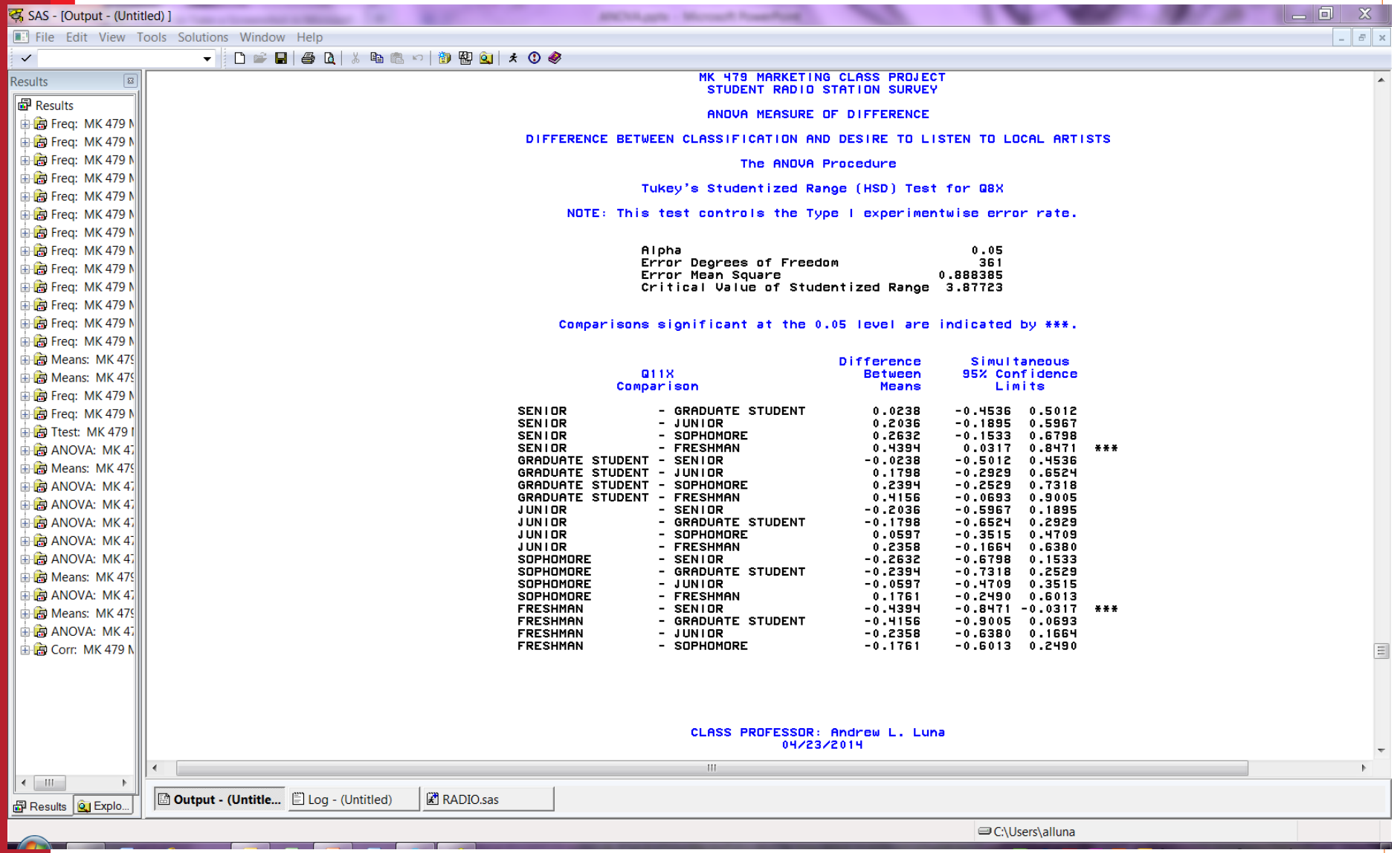
G1 compared to G2
G1 compared to G3
G2 compared to G3

Running the Tukey Test

- The "Honestly Significantly Different" (HSD) test proposed by the statistician John Tukey is based on what is called the "studentized range distribution.“
- To test all pairwise comparisons among means using the Tukey HSD, compute t for each pair of means using the formula:

$$t_s = \frac{M_i - M_j}{\sqrt{\frac{MSE}{n_h}}}$$

Where $M_i - M_j$ is the difference ith and jth means, MSE is the Mean Square Error, and n_h is the harmonic mean of the sample sizes of groups i and j.


SAS - [Output - (Untitled)]
File Edit View Tools Solutions Window Help
Results
Results
Freq: MK 479 N
Means: MK 479
ANOVA: MK 47
Ttest: MK 479 I
Corr: MK 479 N
Results
Explo...
MK 479 MARKETING CLASS PROJECT
STUDENT RADIO STATION SURVEY
ANOVA MEASURE OF DIFFERENCE
DIFFERENCE BETWEEN CLASSIFICATION AND DESIRE TO LISTEN TO LOCAL ARTISTS
The ANOVA Procedure
Tukey's Studentized Range (HSD) Test for Q8X
NOTE: This test controls the Type I experimentwise error rate.
Alpha 0.05
Error Degrees of Freedom 361
Error Mean Square 0.888385
Critical Value of Studentized Range 3.87723
Comparisons significant at the 0.05 level are indicated by ***.
Q11X Comparison
Difference Between Means
Simultaneous 95% Confidence Limits
SENIOR - GRADUATE STUDENT 0.0238 -0.4536 0.5012
SENIOR - JUNIOR 0.2036 -0.1895 0.5967
SENIOR - SOPHOMORE 0.2632 -0.1533 0.6798
SENIOR - FRESHMAN 0.4394 0.0317 0.8471 ***
GRADUATE STUDENT - SENIOR -0.0238 -0.5012 0.4536
GRADUATE STUDENT - JUNIOR 0.1798 -0.2929 0.6524
GRADUATE STUDENT - SOPHOMORE 0.2394 -0.2529 0.7318
GRADUATE STUDENT - FRESHMAN 0.4156 -0.0693 0.9005
JUNIOR - SENIOR -0.2036 -0.5967 0.1895
JUNIOR - GRADUATE STUDENT -0.1798 -0.6524 0.2929
JUNIOR - SOPHOMORE 0.0597 -0.3515 0.4709
JUNIOR - FRESHMAN 0.2358 -0.1664 0.6380
SOPHOMORE - SENIOR -0.2632 -0.6798 0.1533
SOPHOMORE - GRADUATE STUDENT -0.2394 -0.7318 0.2529
SOPHOMORE - JUNIOR -0.0597 -0.4709 0.3515
SOPHOMORE - FRESHMAN 0.1761 -0.2490 0.6013
FRESHMAN - SENIOR -0.4394 -0.8471 -0.0317 ***
FRESHMAN - GRADUATE STUDENT -0.4156 -0.9005 0.0693
FRESHMAN - JUNIOR -0.2358 -0.6380 0.1664
FRESHMAN - SOPHOMORE -0.1761 -0.6013 0.2490
CLASS PROFESSOR: Andrew L. Luna
04/23/2014
Output - (Untitle...
Log - (Untitled)
RADIO.sas
C:\Users\alluna

Results of the ANOVA and Follow-Up Tests

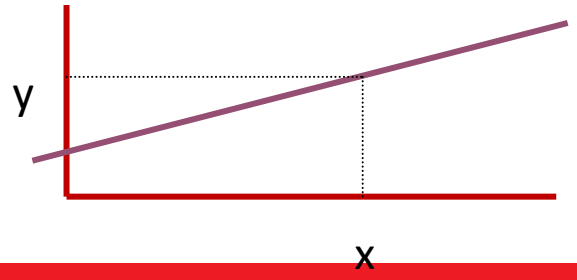
- If the F-statistic is significant, then the ANOVA indicates a significant difference
- The follow-up test will indicate where the differences are
- You may now state that you reject the null hypothesis and indicate which groups were significantly different from each other

Regression Analysis

- The description of the nature of the relationship between two or more variables
- It is concerned with the problem of describing or estimating the value of the dependent variable on the basis of one or more independent variables.

Regression Analysis

Around the turn of the century, geneticist Francis Galton discovered a phenomenon called Regression Toward The Mean. Seeking laws of inheritance, he found that sons' heights tended to regress toward the mean height of the population, compared to their fathers' heights. Tall fathers tended to have somewhat shorter sons, and vice versa.



Predictive Versus Explanatory Regression Analysis

- Prediction – to develop a model to predict future values of a response variable (Y) based on its relationships with predictor variables (X's)
- Explanatory Analysis – to develop an understanding of the relationships between response variable and predictor variables

Problem Statement

- A regression model will be used to try to explain the relationship between departmental budget allocations and those variables that could contribute to the variance in these allocations.

$$Bud.Alloc.\int[x_1, x_2, x_3 \ldots x_i]$$

Simple Regression Model

$$(y) = a + bx$$

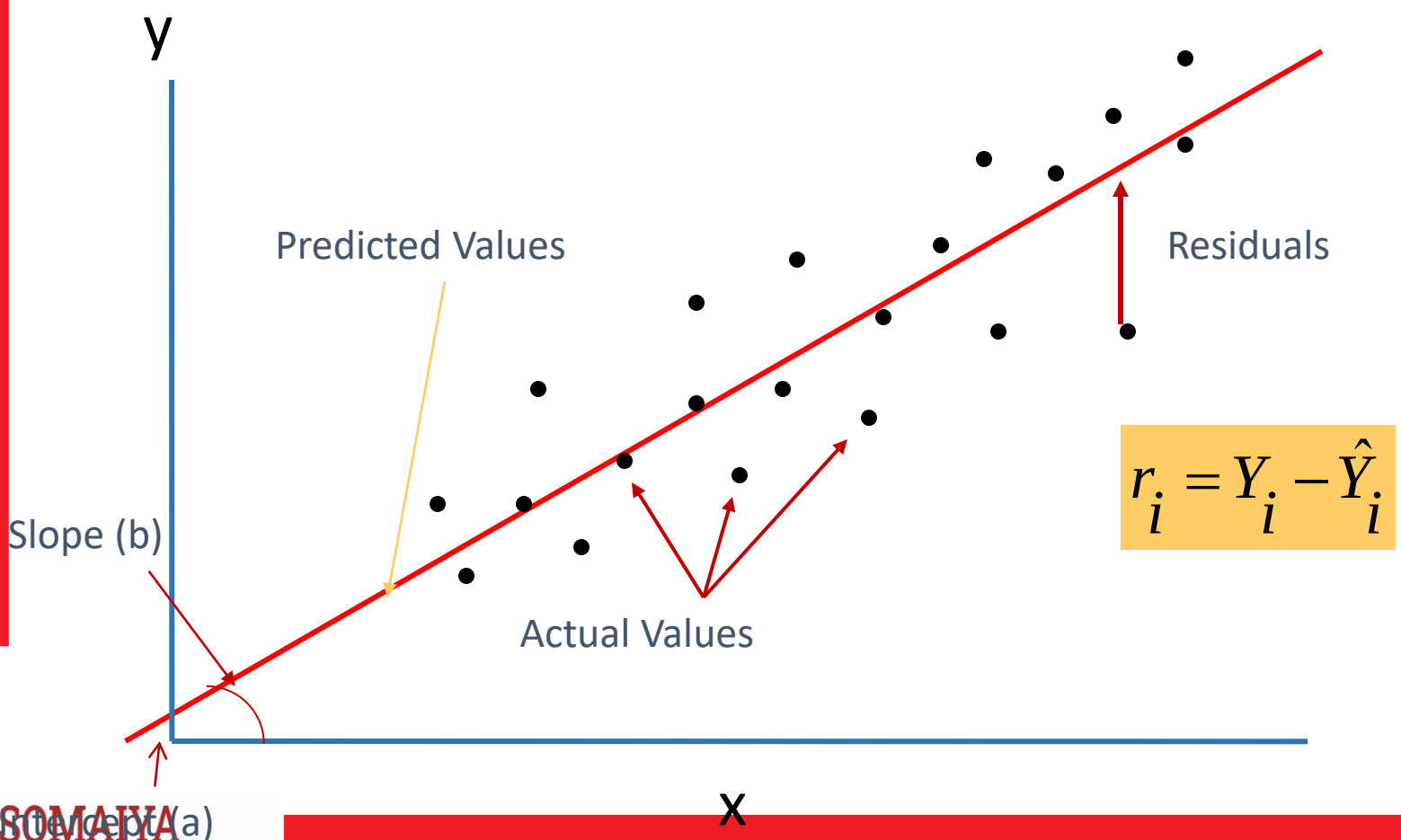
$$\textbf{\textit{Slope}}\ (\boldsymbol{b}) = (N\Sigma XY\ -(\Sigma X)(\Sigma Y)))/(N\Sigma X^2\ -(\Sigma X)^2)$$

$$\textbf{\textit{Intercept}}\ (\boldsymbol{a}) = (\Sigma Y - b(\Sigma X))/N$$

Where:

y = Dependent Variable
x = Independent Variable
b = Slope of Regression Line
a = Intercept point of line
N = Number of values
X = First Score
Y = Second Score
ΣXY = Sum of the product of 1st & 2nd scores
ΣX = Sum of First Scores
ΣY = Sum of Second Scores
ΣX^2 = Sum of squared First Scores

Simple regression model



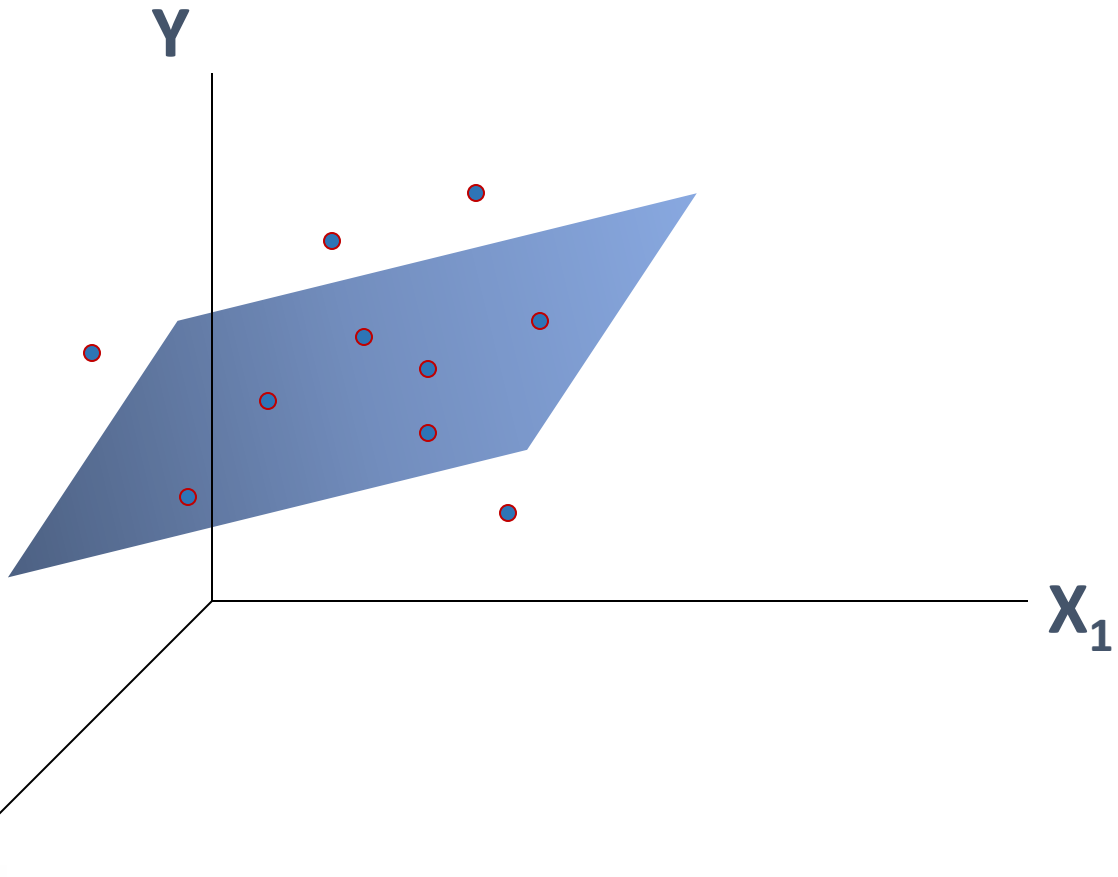
$$r_i = Y_i - \hat{Y}_i$$

Simple vs. Multiple Regression

Simple: $Y = a + bx$

Multiple: $Y = a + b_1X_1 + b_2 X_2 + b_3X_3 \ldots + b_iX_i$

Multiple regression model



Question ?

Reference

Andrew L. Luna
Director
Institutional Research, Planning, and Assessment
The University of North Alabama
alluna@una.edu
Phone: 256.765.4221