NUMERICAL COMPUTING
CST8233

By

Hala Own, Ph.D.

Outline

- Course Outline And Section Information(brightspace)
- Textbook And Software
- Evaluation Breakdown
- Contacting Your Professor
- Introduction To Numerical Computing
- Course Content
- Chapter #1 Number System
- Lab2

About This course

- The course will be delivered in virtual classroom format via synchronous and asynchronous video lectures at the regular class times.
- **You should be familiar with C or C++**
- You should be familiar with basic calculus and linear algebra.

Textbook and Software

- **Basic Technical Mathematics with Calculus** 10th Edition. Allyn J. Washington and Michelle Boue. ISBN 978-0-13-276283-0. Pearson. Additional course text material will be supplied by the instructor as required.

- Software Required: **Microsoft Visual Studio 2019**

- LMS : **Brightspace**

Evaluation Breakdown

Assessment	Frequency	weight	Comments
Assignments	3	20%	Assignment 1 6% Assignment 2 8% Assignment 3 6%
Class Activity	12	10%	Best of 10
Lab Work	12	15%	Best of 10
Quizzes	7	0%	For practice
Midterm Exam	1	25%	Week 7
Final Exam	1	30%	Date will be decided by the Registrar office

Labs

- There will be a zoom meeting for every lab. You can share your screen to demonstrate your work.
- Labs are assessed during the lab session.
- Attendance is mandatory (according to course outline policy).
- You must demo your lab(even if it does not work properly)
- The labs roughly follow the material presented in lecture.
- Lab grades count as 15% of your course grade.
- You are encouraged to work together with other students, but you must demo your own work.

Labs Grading

- Each lab will be given a grade of A,B,C or Zero. These grades correspond with the weights 1.5%, 1%, 0.5%,and 0.
- At the end of the semester, **best of 10 labs** are averaged and then integrated with your final grade .
- The grading criteria are:

Grade	Weight	Description
A	1.5%	All exercises were completed correctly.
B	1%	All exercises were attempted and are substantially correct.
C	0.5%	Little or nothing correct in the submission.
Zero	0	Lab was not ATTENDED

Assignments

- Three Assignments :
  - Assignment#1(6%)
  - Assignment #2(8%)
  - Assignment #3(6%)

Class Activities

- A simple quiz at the beginning of each lecture based on the topic which has been covered in the previous week 10% (**best of 10**)

Quizzes , Midterm And Final

- Quizzes : given at the end of each topics, you will not earn any mark for the quizzes, they are just for practice.
- Midterm and Final are both conducted Remotely.

- Midterm (**Week 7**)would take the entire Lecture regular length, **50 min**. (25%).

- Midterm covers the first **5 weeks** (introduction to numerical computing, number systems, floating point representation, Interpolation, Statistics, Liner Regression)

- **Final will cover all materials given during the course(30%)**

Contacting your Instructors

- Office hours: by appointment via Zoom
- Via email

ownh@algonquincollege.com

- If you need extra help send me an email, and we'll try to set up a mutually-convenient time for private Zoom meeting

Slido: Audience Interaction Platform

Course Contents

- Introduction To Numerical Computing
- Number System
- Floating Point Representation
- Polynomial Interpolation
- Descriptive Statistics
- Linear & Nonlinear Regression
- Taylor & Maclurin Series
- Numerical Integration/Differentiation
- Ordinary Differential Equations
- Solving Set of Linear Equation

What is Numerical Computing

Numerical computing is an interconnected combination of computer science and mathematics in which we develop and analyze algorithms for solving mathematical model approximately (numerically).

Why we need Numerical Computing

The numerical computational techniques are the technique by which mathematical problems are formulated and they can be solved with arithmetic operations. Those techniques are basically numerical methods.

Numerical method supports the solution of almost every type of problem.

The Goals of Numerical Computing

- Efficiency

  how much work is necessary (computer and human) to reach a solution

- Accuracy

  how close we are to the true value

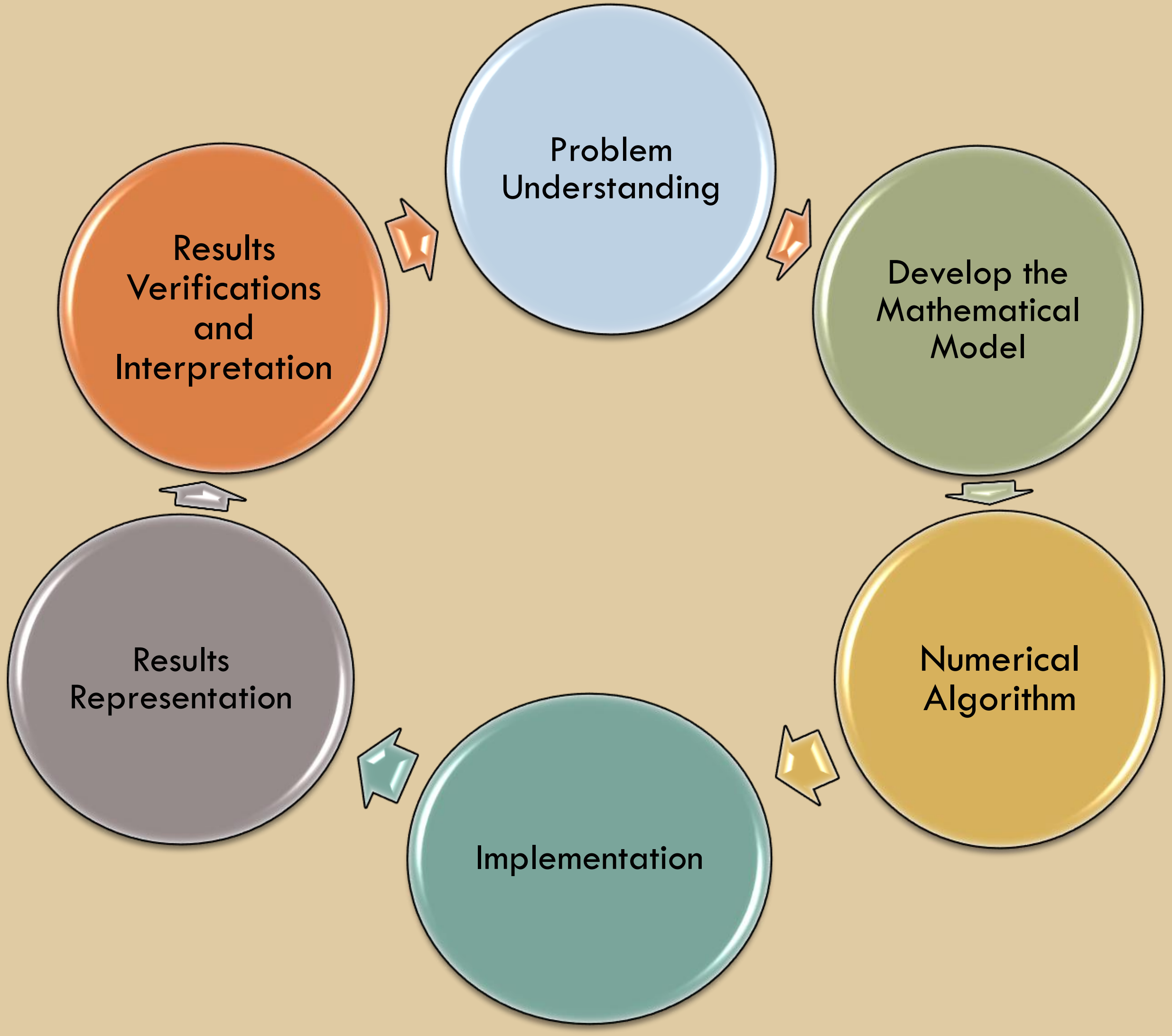
- Precision

  the level of details in our solution

Applications of Numerical Computing

- Biology: Study the genetic evolution(mutations etc) , enzyme production/kinetics, diffusion of virus epidemics etc.
- Finance: Model and predict the highly fluctuating market trends. Optimally pricing the options.
- Climate: Predict wind direction, humidity, temperature, pressure etc at a place in future times from the current data.
- Space missions: Predict the trajectory of the spacecrafts.
- Industry: Car companies can improve the crash safety of their vehicles by using computer simulations of car crashes. Such simulations essentially consist of solving partial differential equations numerically.

Numerical Computing Steps To Solve Problem



Course Overview

Number System And Floating-Point Representation Overview

- Number System Conversion
- Floating Point Representation

Computers cannot represent all real numbers exactly, so we face new challenges when designing computer algorithms for real numbers. Now, in addition to analyzing time and space complexity, we must be concerned with the accuracy of the resulting solutions.

What Is The Output Of The Following Program

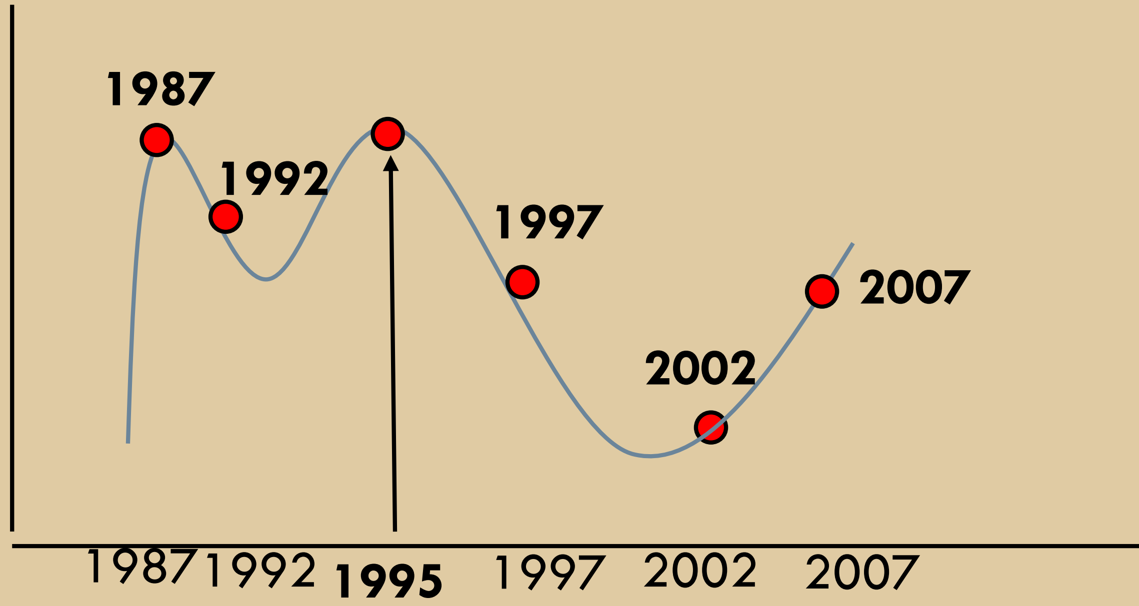

```
int main()
{
        float x;
        x=0.1+0.1+0.1 ;
        if (x==0.3)
            printf("True");
        else
            printf("False");

        return 0;

}
```

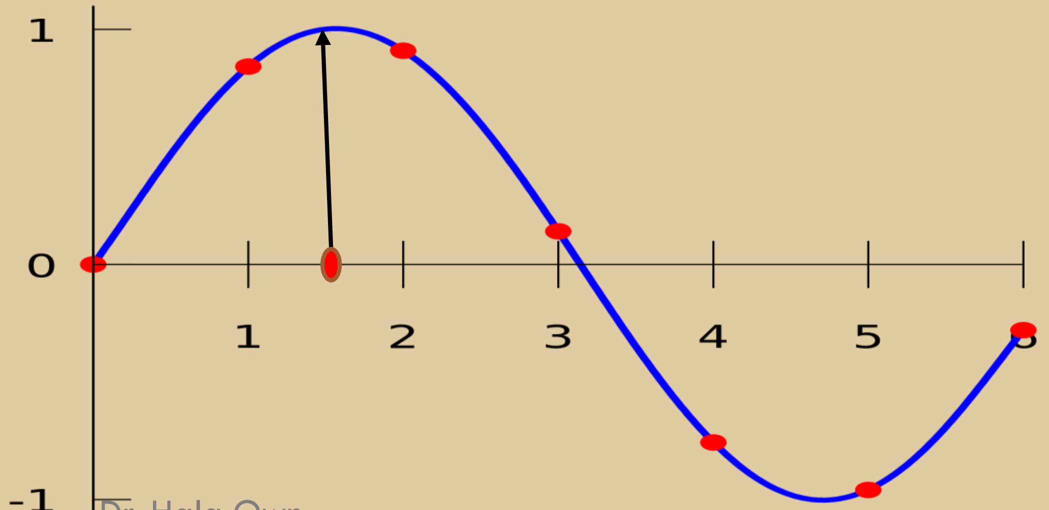
Interpolation Motivation

If we are finding out the number of sunspots in **1995** when we know the number of sunspots in the year 1987, 1992, 1997, 2002,2007 and so on. the process of finding the number of sunspot of **1995** is known as **interpolation**



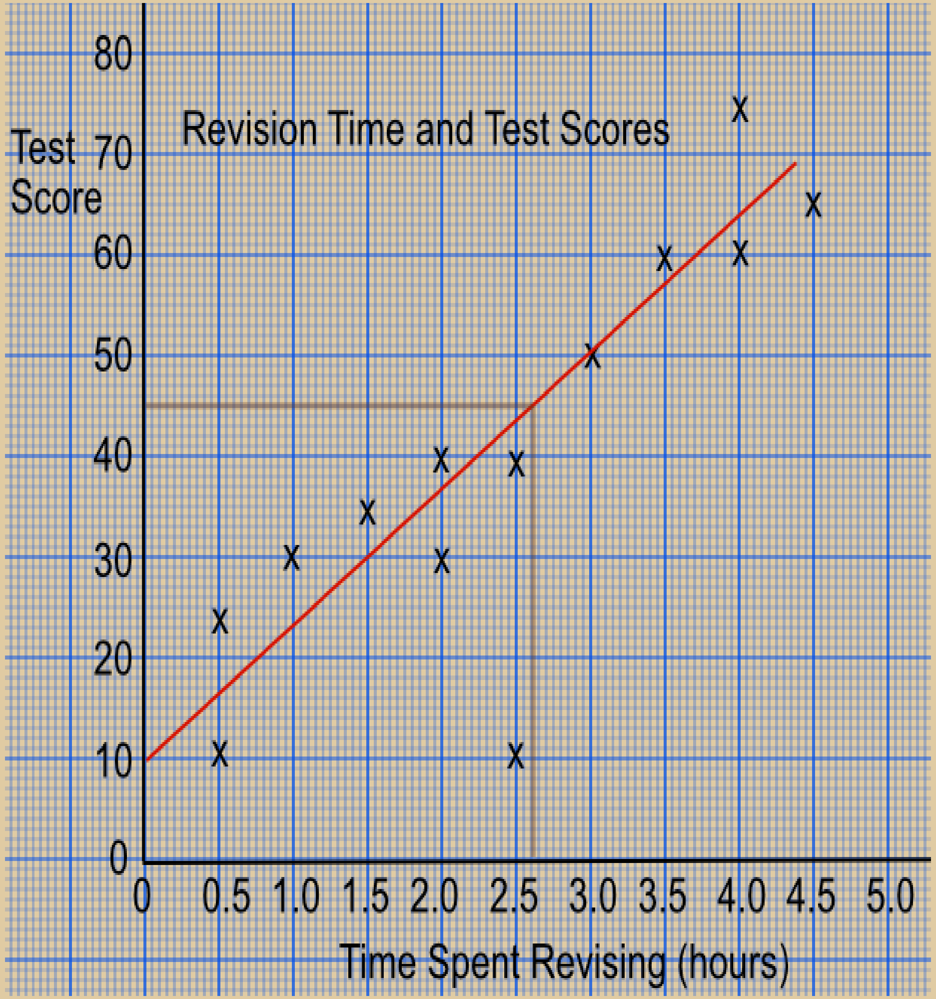
Polynomial Interpolation

- **Interpolation** is the process of deriving a simple function from a set of data points so that the function passes through all the points.
- **The goal** is to use this function to estimate a new data points within the range of a discrete set of known data points.



What Is The Linear Regression

- Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line.



Application of Regression Analysis :

- To predict the value of a difficult to measure variable, Y, based on an easy to measure variable, X.

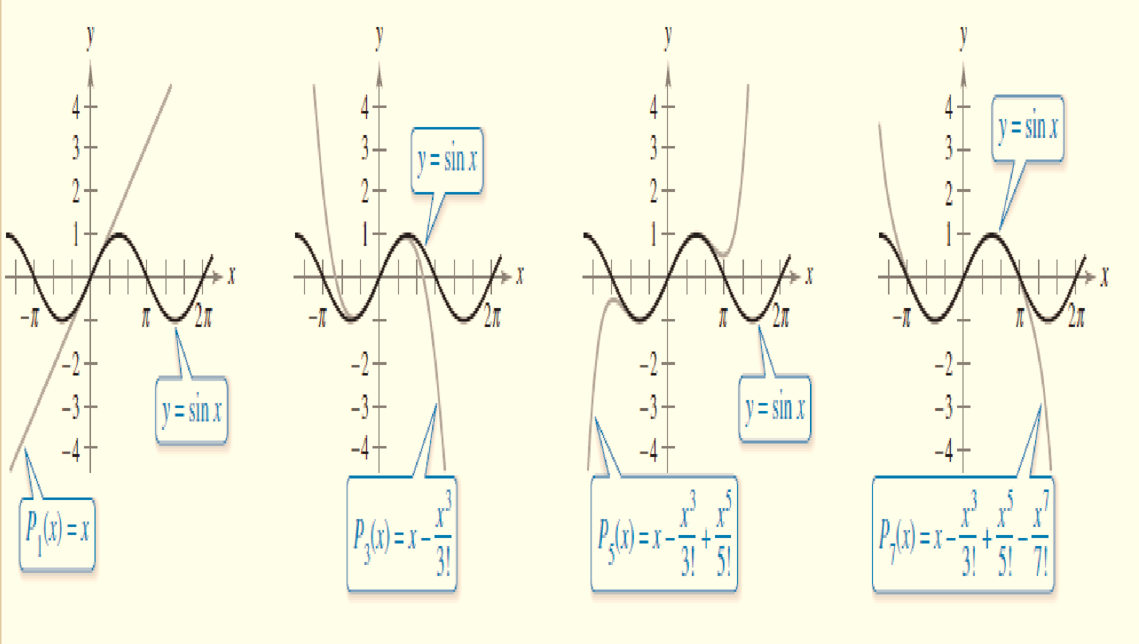
- Explain the impact of changes in an independent variable on the dependent variable.

Dependent variable: the variable you wish to explain(y)

Independent variable: the variable used to explain (x)

Taylor & Maclurin Series Overview

- Taylor & Maclurin Series
  - Evaluating functions like sin x, e^x etc
  - Absolute & Relative Error



As n increases, the graph of P_n more closely resembles the sine function.

Five Basic Maclaurin Expansions

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 \cdots = \sum_{n=0}^{\infty} x^n$$

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} \cdots = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n+1}}{(2n+1)!}$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{(2n)!}$$

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots = \sum_{n=0}^{\infty} \frac{(-1)^n x^{n+1}}{n+1}$$

Calcworksho

Numerical Integration/Differentiation

Evaluate the integral, $I = \int_a^b f(x)dx$ without doing the calculation analytically.

Necessary when either:

Integrand is too complicated to integrate analytically

$$\int_0^2 \frac{2+\cos(1+\sqrt{x})}{\sqrt{1+0.5x}} e^{0.5x} dx$$

Integrand is not precisely defined by an equation, i.e., we are given a set of data (x_i, y_i), $i = 1, 2, 3, \ldots, n$

Ordinary Differential Equation(ODE): Basics

- **Definition:** a differential equation is an equation that contains one or more ordinary derivatives of an unknown function.

Dependent variable

Initial Condition

ODE Example : $\frac{dy}{dx} + xe^{x} = \sin(x^2)$, y(5)=7.

Independent variable

Algebraic Equation

2+xy+5y2=k

- The solution of the differential equation is the function that satisfied the differential equation.
- The Initial Condition (IC), represents the solution of ODE at specific point(i.e the derivative of unknown function at specific point)

Ordinary Differential Equation(ODE):Motivation

- It is generally easier to relate **rates of change** of quantities than it is to develop algebraic formulas for specific values of those quantities under specific conditions.
- The function that results form the solution of a differential equation usually serves as the algebraic relation that someone can use to make predictions and get specific numerical values.

The two Methods are:

Euler's Method

Runge-Kutta Method

Solving Small Numbers of Equations

There are many ways to solve a system of linear equations:

- Graphical method
- Cramer's rule
- Method of elimination

For $n \leq 3$

- Numerical methods for solving larger number of linear equations:
  - **Naïve Gauss Elimination**
  - LU decompositions and matrix inversion

29

Solving Systems of Linear Equations:

It basically studies methods that can solve systems of linear equations. These methods include the **Gaussian Elimination**, **LU Decomposition**, and many other methods and techniques.

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \quad (1) \\
a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \quad (2) \\
&\dots \\
a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= b_n \quad (n)
\end{aligned}
$$

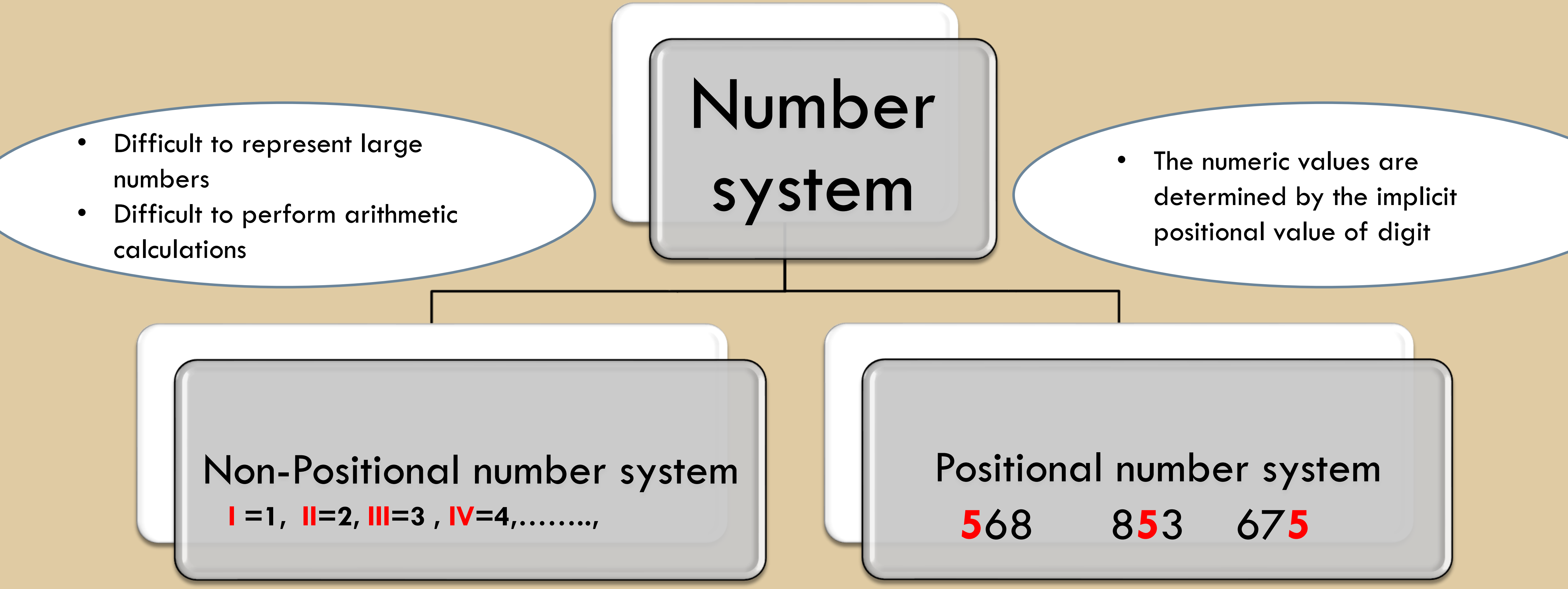
The goal is to solve: [A][X]=[C]

Chapter 1: Number System

Number System

- Basic Concept
- Decimal Number System
- Binary Number System
- Convert From Decimal To Binary
- Convert From Binary To Decimal
- Scientific Notation

Number System



Example Of Positional Number Systems

- Decimal
- Binary
- Octal
- Hexadecimal

Decimal Number System

- Base: "decem" (Latin) => 10 so the base is 10
- Digits (staring from 0 to base -1)0,1,2,3,4,5,6,7,8,9

- Example **3 4 2**

Each Number is multiplied by 10

position	- - -	4	3	2	1	0
Value	- - -	$10^4=10000$	$10^3=1000$	$10^2=100$	$10^1=10$	$10^0=1$

Binary Number System

- Base: “binarius” (Latin) => two so the base is 2
- Digits (staring from 0 to base -1)   0 , 1

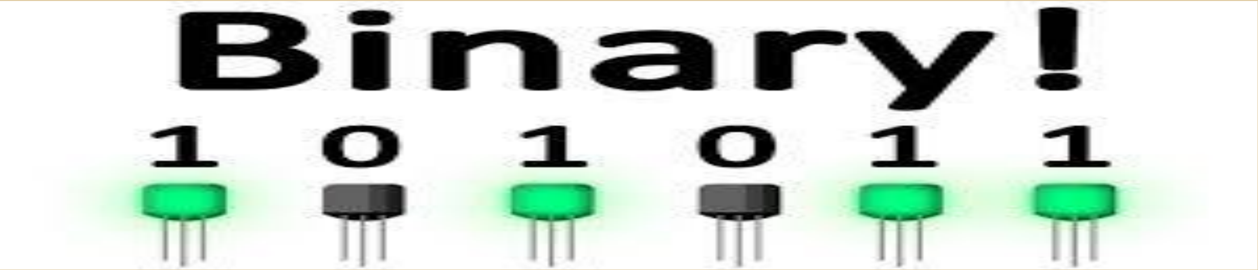
Each Number is multiplied by 2

Example $(11010)_2$

position	- - -	**5**	**4**	**3**	**2**	**1**	**0**
Position Value	- - -	$2^5=32$	$2^4=16$	$2^3=8$	$2^2=4$	$2^1=2$	$2^0=1$

How Information Stored Inside Computer

Used to model the series of electrical signals the computers used to represent information



Each digit of a binary number is called a **bit**.

A binary number with eight bits (i.e. digits) is called a **byte**.

[1]https://i.ytimg.com/vi/Xpk67YzOn5w/maxresdefault.jpg

[2]https://www.westfloridacomponents.com/mmgpj.stnenopmoC-cinortcelE/09F/scihparg/5

[3]https://encrypted-tbn9dNA:nbt=q?segami/moc.citatsg.0GcTLI4x4uFq2b-8f28U-yt7oaYQFjtDHM71qhzG4alXdRas7AB_6

Convert From Binary To Decimal

Position	8	7	6	5	4	3	2	1	0
Position Value	$2^8=256$	$2^7=128$	$2^6=64$	$2^5=32$	$2^4=16$	$2^3=8$	$2^2=4$	$2^1=2$	$2^0=1$

Each Number is doubled

$$value = \sum digit * Base^{position}$$

11011101

$(1*2^7)$ $(1*2^6)$ $(0*2^5)$ $(1*2^4)$ $(1*2^3)$ $(1*2^2)$ $(0*2^1)$ $(1*2^0)$

128 + 64 + 0 + 16 + 8 + 4 + 0 + 1

221

Your Turn

- Convert 1000001 to decimal

Solution:

65

Convert 10011001 to decimal

Solution:

153

Convert from Decimal to Binary

- First Method

1- Divide The Number In Half And Ignore Reminder

Divided by 2		Divided by 2		Divided by 2		Divided by 2	
1		2		5		10	21
1		0		1		0	1

2- Write 1 for odd numbers and 0 for even

$$(21)_{10}=(1\,0\,1\,0\,1)_2$$

Convert From Decimal To Binary(cont.)

- Second Method(division Method)

$21 \div 2 = 10$ R 1

$10 \div 2 = 5$ R 0

$5 \div 2 = 2$ R 1

$2 \div 2 = 1$ R 0

$1 \div 2 = 0$ R 1

$$(21)_{10} = (1\ 0\ 1\ 0\ 1)_2$$

Divide the number in half and Keep the reminder
STOP when quotient equals 0

Your Turn

- Convert to binary

165

Solution

10100101

- Convert to binary

364

Solution

101101100

Convert From Decimal Fraction To Binary

- Ex: 3.40625

0.40625 x 2 =0.8125 **0**
0.8125 x 2=1.625 1
0.625 x 2=1.25 1
0.25 x 2=0.5 0
0.5 x 2= 1.0 1

Multiply the fraction part by 2 and Keep the integral part
STOP when fraction equals 0

$$(3.140625)_{10}=(11.01101)$$

Convert From Decimal Fraction To Binary

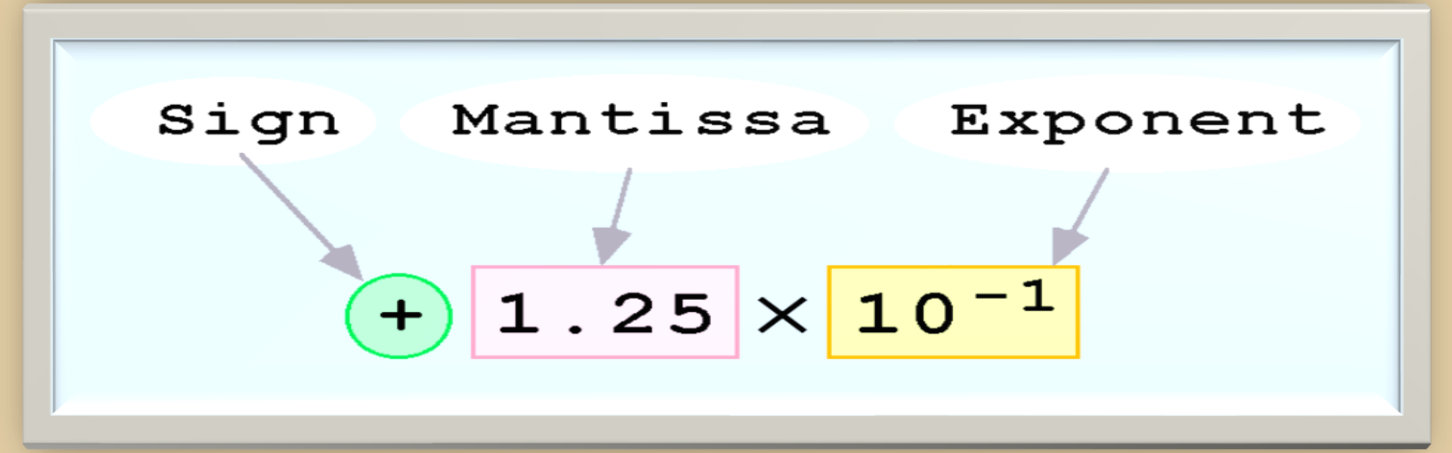
- Ex: 0.45

0.45 x 2 =0.9	0
0.9 x 2=1.8	1
0.8 x 2=1.6	1
0.6 x 2=1.2	1
0.2x 2= 0.4	0
0.4 x 2 = 0.8	0
0.8 x 2= 1.6	1
0.6x 2 =1.2	1

Multiply the fraction part by 2 and Keep the integral part
STOP when fraction equals 0

Scientific Notation

Numbers written in scientific notation have three components:

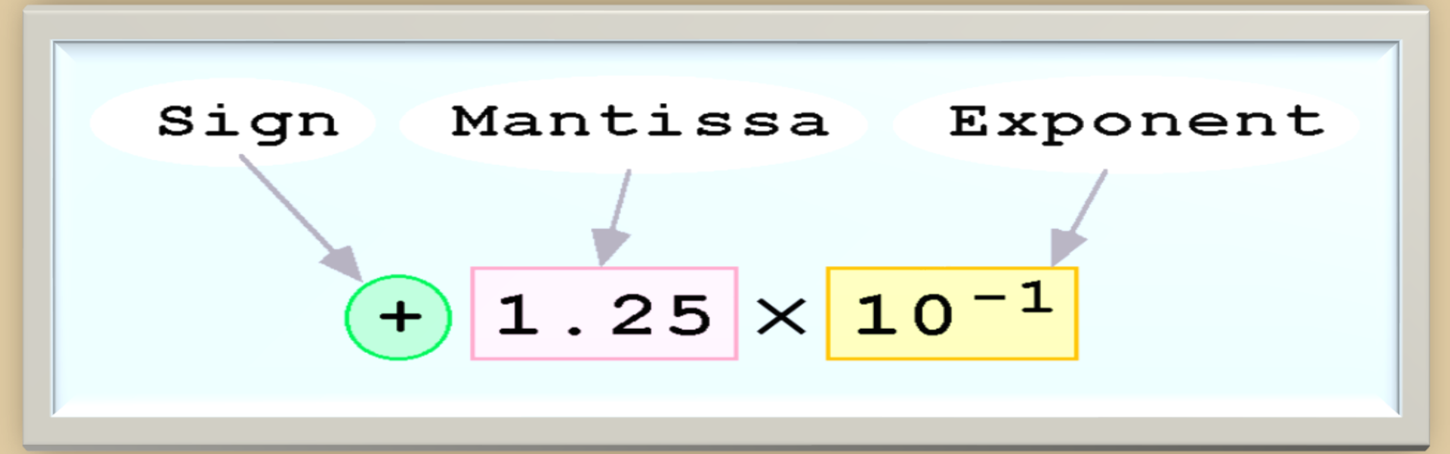


Mantissa must be between 1 and 10

Exponent must be integer

How To Covert A Number To Scientific Notation

1. Place the decimal point after non-zero digit and count the number of places the decimal point has moved.
2. If the decimal place has moved to the left, then multiply by a positive power of ten. If it has moved to the right, it will result in negative power of 10.



Why we Use Scientific Notation

- **Scientific Notation** makes the expression of very large or very small numbers much simpler:
- Example:

35000,000,000,000 → 3.5×10^{13}

0.00000000056547 → 5.6547×10^{-10}

Advantages :

- Standard representation of real number
- Easier in comparison
- Easier in arithmetic operation

Why we Use Scientific Notation

- Compare the following two numbers

152,000,000,008,500,000,00 940,000,004,000,000,000

Compare the following two numbers

3.3×10^{-16} and 3.2×10^{-14}

Scientific Notation In Binary Numbers

- the idea of scientific notation in binary numbers is similar, but uses powers of two

- Example:

- $6{,}584 = 1\ 1001\ 1011\ 1000_2 = 1\ .1001\ 1011\ 1000_2 \times 2^{12}$

Converting From Scientific Notation To Normal Representation

- If the power of 10 is positive, move the decimal point to the right; if negative move to the left.

- Examples

$$5.7 \times 10^{-7} = 0.00000057$$

$$3.01 \times 10^{3} = 3010$$

$$1.110101 \times 2^{5} = 111010.1$$

Examples

- Convert the following number to scientific notation
- 3040= 3.040×10^{3}
- -459.546= -4.59546×10^{2}
- 0.00252= 2.52×10^{-3}
- 1101.1010= 1.1011010×2^{3}
- 0.010101001= 1.01001×2^{-2}

Summary

Lab2: Number Systems

Due Week 2

- Discussion
- Q&A

Summary

- Course Section Information
- Course overview
- Number Systems
- Number System Conversions
- Scientific Notation

NUMERICAL COMPUTING
CST8233
Floating Point Representation

By

Hala Own, Ph.D.

Outline

- Floating Point Representation Background
- IEEE Standard Floating-Point Representation
- Tiny Machine Floating Point Representation Examples
- Machine Epsilon
- Absolute And Relative Error
- Floating Point disaster

Binary Real Number

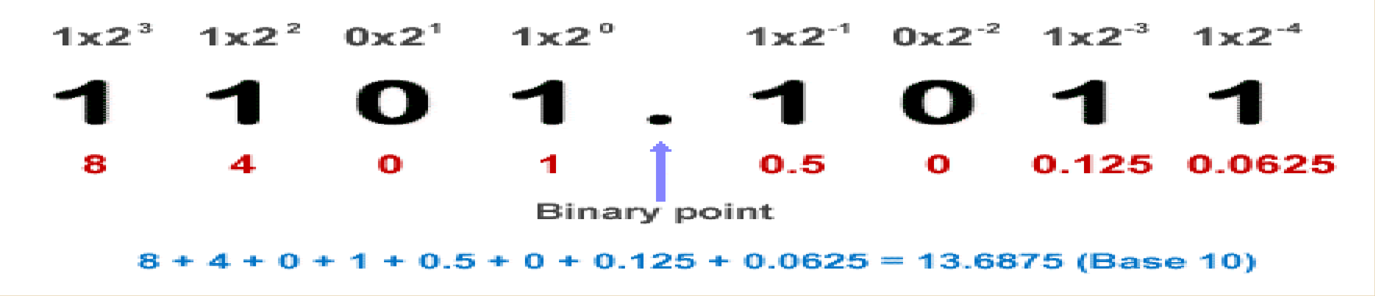
- The binary number $b_i\ b_{i-1}\ \ldots\ b_2\ b_1\ b_0\ .\ b_{-1} b_{-2}\ b_{-3}\ \ldots .\ b_{-j}$
- represents a particular (positive) sum. Each digit is multiplied by a power of two according to the following chart:

Bit	b_i	b_{i-1}	…..	b_2	b_1	b_0	.	b_{-1}	b_{-2}	b_{-3}
Position value	2^i	2^{i-1}		2^2	2^1	2^0	.	1/2	1/4	1/8

Representation:

Bits to the right of the binary point represent fractional powers of 2. This represents the rational number

Binary Real Number Example



Examples:

1- Convert 3.59375 to binary

$(3.59375)_{10} = (11.10011)_2$

	Result	Integer Part	Fractional part
0.59375*2	1.1875	1	0.1875
0.1875*2	0.375	0	0.375
0.375*2	0.75	0	0.75
0.75*2	1.5	1	0.5
0.5	1.0	1	0

2- Convert 0.1011 to decimal

$= (1 \times 2^{-1}) + (0 \times 2^{-2}) + (1 \times 2^{-3}) + (1 \times 2^{-4})$

$= 0.5 + 0 + 0.125 + 0.0625$

$= 0.6875$

Fixed-Point Representation

- To represent 3.625, we would use the first 24 bits to indicate 3, and we'd use the remaining 8 bits to represent 0.625. Thus, our 32-bit representation would be:

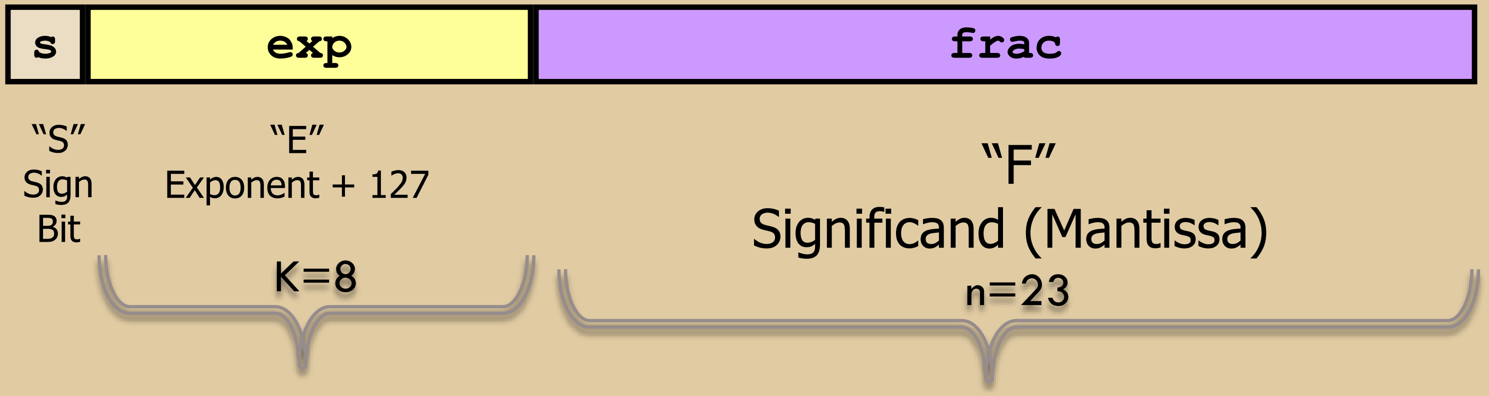
00000000 00000000 00000011 10100000.

Floating-Point Representation

- The **Floating-Point Representation** is how a computer stores a real number. It has features similar to that of scientific standard form. It is called floating-point because the decimal point “floats” to a normalized position.

Computer Representation of Real Numbers

IEEE Floating-point Format



Sizes

- **Single precision:** 8 `exp` bits(k), 23 `frac` bits 32 bits total(n)
- **Double precision:** 11 `exp` bits(k), 52 `frac` bits 64 bits total(n)

Floating-point Representation(cont.)

$6{,}584 = 1\ 1001\ 1011\ 1000_2 = 1\ .1001\ 1011\ 1000_2 \times 2^{12}$

- **The Hidden Bit**

The 1 before the decimal point is the "hidden bit" and is not stored in memory but the hardware remembers in any calculation that it is really there. It represents the number $2^0 = 1$

[1].01000000000000000000001

- **Bias (Excess representation)=**

Bias=$2^{k-1} - 1$

To allow for negative numbers in floating point we take our exponent and **add 127(bias)** to IEEE standard Format. **HOW??????**

Why a bias of 127

Consider the following three numbers

0.00000005 = 0 01100110 10101101011111110010101

1 = 0 01111111 00000000000000000000000

65536.5 = 0 10001111 00000000000000001000000

- The 3 numbers are listed in increasing order of magnitude,
- The exponents are in increasing order.
- We can do a simple comparison of the numbers
- **We don't have to decode** the floating-point numbers in order to compare them.

Floating-point Representation Example

- Example: 12.5

$\text{Bias}=2^{k-1}-1$

Step1 : convert the number to binary

$12.5=1100.1_2$

Step2: Normalize the number

$[1.]1001 \times 2^3$

Step3 : calculate the biased-exponents (Exponent + bias)

$127 + 3 = 130 = 1000\ 0010_2$

Step 4: represent the number in IEEE Single Precision Floating-point Format

sign	Biased exponent	Fraction
0	10000010	10010000000000000000000

Floating-point Representation Example

- Example: -1.0

Step1 : convert the number to binary

$1.0 = 1.0_2$

Step2: Normalize the number

$[1.]0 \times 2^0$

Step3 : calculate the exponents

$127+0 = 127 = = 0111\ 1111_2$

Step 4: represent the number in IEEE Single Precision Floating-point Format

1	01111111	00000000000000000000000

Floating-point Representation Example

- Example: 0.45

Step1 : convert the number to binary

$0.45 = 0.01110011001100110011001100110011\ldots\ldots\ldots\ldots_2$

Step2: Normalize the number

$[1.]\ 11001100110011001100110 \times 2^{-2}$

Step3 : calculate the exponents

$127+(-2) = 125 = = 01111101_2$

Step 4: represent the number in IEEE Single Precision Floating-point Format

0	01111101	11001100110011001100110

Approximated representation

Special Values

- Case1: exp = 0000…00, frac ≠ 000…0 ⟹ Denormalize number
- Case2: exp = 111…1, frac = 000…0 ⟹ ± ∞
- Case3: exp = 111…1, frac ≠ 000…0 ⟹ NAN
- Case4: exp = 00000….000, frac = 000…0 ⟹ ± 0

Floating-point Representation Decoding

$$V = (-1)^s \, x M x \, 2^E$$

Bias=$2^{k-1} - 1$

Normalized Values

Condition: exp ≠ 000…0 and exp ≠ 111…1

E= exponent - Bias M=1.f

Denormalized Values

Condition: exp = 000…0

M=f E=1-Bias

Floating-Point Representation Decoding: Example1

Normalized Values

Condition: exp ≠ 000…0 and exp ≠ 111…1

E= exponent - Bias

- Example: find the decimal number for the following floating point

  0 10001001 00000000000000000000000

Solution

First check this number represent which case (here normalized case)

S=0 → Positive number

step 1:**Calculate M**= 1.00000000000000000000000

Step 2: **Calculate E** = exp- bias =137-127=10

$$V = (-1)^s xMx2^E$$

Step 3: Apply the equation

$V= -1^0 \text{ x } 1\text{x}2^{10} = 1024_{10}$

Floating-Point Representation Decoding: Example

Normalized Values

Condition: exp ≠ 000…0 and exp ≠ 111…1

E= exponent - Bias

- Example: find the decimal number for the following floating point

  1 10000001 01000000000000000000000

- **E=**exp-Bias=129-127=2 (decimal)
- S=1 → negative number
- **M=1**. 01000000000000000000000

  = 1+ 0+1/4 = 1.25

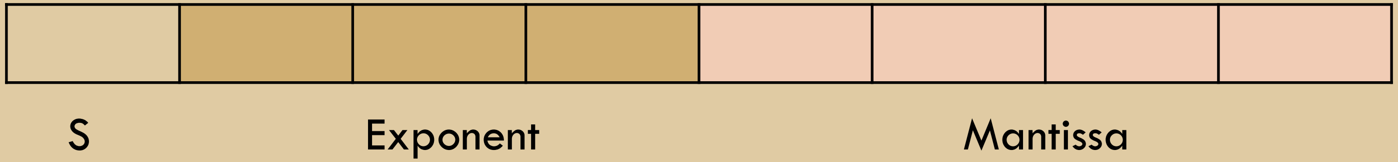
$$V = (-1)^s x M x 2^E$$

$V = -1^1 * 1.25 * 2^2 = (-5)_{10}$

Floating-point Representation hypothetical Tiny Machine

8-bit floating-point representation

one sign bit, three exponent bits (k =3), and four fraction bits (n = 4)



Bias=$2^{k-1}-1$

Floating-point Representation hypothetical Tiny Machine

- Consider an 8-bit floating-point representation based on the IEEE 754 floating-point standard with one sign bit, three exponent bits (k =3), and four fraction bits (n = 4). Complete the following table

Bias =3

Bias=$2^{k-1} - 1$

Step1: check the number Normalized.
Step2: calculate M=1.0100= 1+0+1/4+0+0= 1.25
Step 3: Calculate E = exp- bias = 101+bias=5-3=2
V= 1.25* 2^2
=1.25*4=5

Convert to binary= 14.5= 1110.1
Normalize= 1.1101
Exp= 3+bias=3+3=6= 011

	Bit pattern
5	01010100
-14.5	11101101
-2	11000000
0.65625	00100101
0.171875	0 000 1011

Floating-point Representation Decoding& Encoding Examples

Consider a 9-bit floating-point representation based on the IEEE 754 floating-point standard with one sign bit, four exponent bits (k =4), and four fraction bits (n = 4).

Bias =7

Bias=$2^{k-1}-1$

3/8= 0.375
In binary =0. 011

Number Description	Bit presentation
3/8	0 0101 1000
+∞	0 1111 0000
Largest number normalized	*011101111
Smallest positive number normalized	*000010000
Largest number Denormalized	*000001111
Smallest number Denormalized	*100000001
-13.9	110101011

-13.9=1101.11100
Normalize=1.10111100 *2^3

*try to decode these bit patterns

Important Characteristics of IEEE Floating Point Numbers

- **The smallest positive normalized**

0 0000 0001 000 0000 0000 0000 0000 0000

$=1.176*10^{-38}$

- **The Largest positive normalized**

0 111 11110 111 1111 1111 1111 1111 1111

The largest number $\sim 2 * 2^{254-127} = 2^{128} = 10^{38.53184} = 10^{38} * 10^{.53184}$

$= 3.403 * 10^{38}$

Overflow and Underflow

- $1.176*10^{-38} \leq X < 3.403 * 10^{38}$
- Floating-point overflow and underflow can cause programs to crash.
- Overflow occurs when there is no room to store the high-order bits resulting from a calculation.
- Underflow occurs when a value is too small to store, possibly resulting in division by zero.

Machine Epsilon

- The smallest change that can be made to any Floating-Point number is to add (subtract) 1 to the least significant bit of the mantissa.
- The value of the 23rd bit of the mantissa = 2^{-23} = $1.192*10^{-7}$

Always equal 2^{-n} → n is the number of bit reserved for mantissa

Is called **Machine precision**. It measures the amount of relative precision you have in the Floating-Point representation of a number.

1	**01111111**	**00000000000000000000000**
1	01111111	00000000000000000000001

The next number in the binary

Numerical Errors

- Round off& Chopping errors

Due to the way that digital computers store numbers

- Truncation Error

This is introduced by numerical methods used in solving mathematical problems(next chapter).

Absolute and Relative error

Absolute error= $|true\ value - numerical\ value|$

Relative error $= \left| \dfrac{absolute\ error}{true\ value} \right|$ *100%

Relative error < machine Epsilon

The Difference Between Absolute And Relative Error

- The exact distance between two cities = 100km and the measured distance is 99km

Absolute error= $|100-99|=1$km

Relative error $=|\frac{1}{100}|=0.01=1\%$

The exact distance between two cities = 2km and the measured distance is 1km

Absolute error= $|2-1|=1$km

Relative error $=|\frac{1}{2}|=0.5=50\%$

Rounding Error in 9 bit machine(k=4, n=4) (Example)

-13.9 → Encoding → 110101011

-13.9=1101.11100
Normalize=1.10111100 $*2^3$

110101011 → Decoding → -13.5

M=1+1/2+1/8+1/16=1.6875
E=10-7=3
V=- $1.6875*2^3$
= 13.5

Absolute error= |-13.9- (-13.5) | = 0.4

Relative error= $\left|\frac{0.4}{-13.9}\right|$= 0.0287 < 2^{-4}

< **0.0625 (Machine Epsilon)**

Truncation Error (example)

- $e^x = 1 + \mathrm{x} + \frac{x^2}{2!} + \frac{x^3}{2!} + \ldots\ldots\ldots\ldots\ldots$

$e^{0.7} = 2.0137$ analytical solution

Compute $e^{0.7}$ using the first five terms in the equation

$e^{0.7} = 2.0122$ numerical solution

Relative error $= \left|\frac{2.0137 - 2.0122}{2.0137}\right| = 0.00074$

Floating Point Disasters: The story of Ariane 5

June 4, 1996 an unmanned Ariane 5 rocket launched by the European Space Agency exploded just forty seconds after its lift-off from Kourou, French Guiana. The rocket was on its first voyage, after a decade of development costing $7 billion. The destroyed rocket and its cargo were valued at $500 million.

https://web.ma.utexas.edu/users/arbogast/misc/disasters.htm

Floating Point Disasters: The story of Ariane 5

The software converts 64- bit floating point numbers to a 16-bit integer representation of the horizontal bias. The software, written in Ada, was also used with Ariane 4 rocket subsystems but was not apparent then. Ariane 5's faster engines caused the 64-bit numbers to be larger than in the Ariane 4 (larger than a 16-bit integer can represent).

https://www.youtube.com/watch?v=W3YJeoYgozw

Patriot Missile Failure

Patriot Missile defense system misses scud –

- 28 people die
- System tracks time in tenths of second(
- Converted from integer to floating point number.
- Accumulated rounding error causes drift. 20% drift over 8 hours.
- Eventually (on 2/25/1991 system was on for 100 hours) causes range mis-estimation sufficiently large to miss incoming missiles.

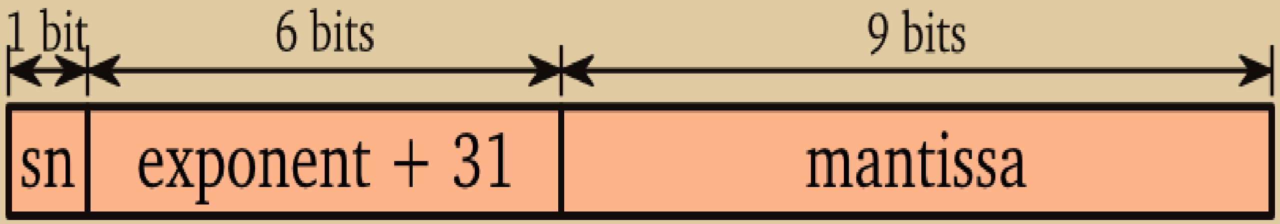
http://www.ima.umn.edu/~arnold/disasters/patriot.html.

The source of the Patriot missile bug.

- 0.1 is not exactly representable? Its nearest single precision

  0 011 1101 1 100 1100 1100 1100 1100 1101
- Now the 24 bit register in the Patriot stored instead 0.00011001100110011001100 introducing an error of 0.000000000000000000000000011001100... binary, or about 0.000000095 decimal.
- Multiplying by the number of tenths of a second in 100 hours gives $0.000000095 \times 100 \times 60 \times 60 \times 10 = 0.34$.).
- A Scud travels at about 1,676 meters per second, and so travels more than half a kilometer in this time. This was far enough that the incoming Scud was outside the "range gate" that the Patriot tracked

Tradeoff between Range and Precision



- Given a limited length for a floating-point representation, we have to compromise between **more mantissa bits (to get more precision)** and **more exponent bits (to get a wider range of numbers to represent).** For 16-bit floating-point numbers, the 6-and-9 split is a reasonable tradeoff of range versus precision.

Applications

- **Tensor Processing Units (TPUs)**

numerical computing libraries tend to stick to 32-bit floats by default. Half the size means the computations can be done faster, But lower precision comes with a cost.

Google's Tensor Processing Units instead use a modified 16-bit format for multiplication as part of their many optimizations for deep-learning tasks. The 8-bit exponent with 7-bit significand has just as many exponent bits as a 32-bit floating point number. And it turns out that in deep learning applications, this matters more than the significand bits. Also, when multiplying, the exponents can be added (easy) while the significand bits have to be multiplied (harder). Making the significand smaller makes the silicon that multiplies floats about 8 times smaller.

Lab3:

Due

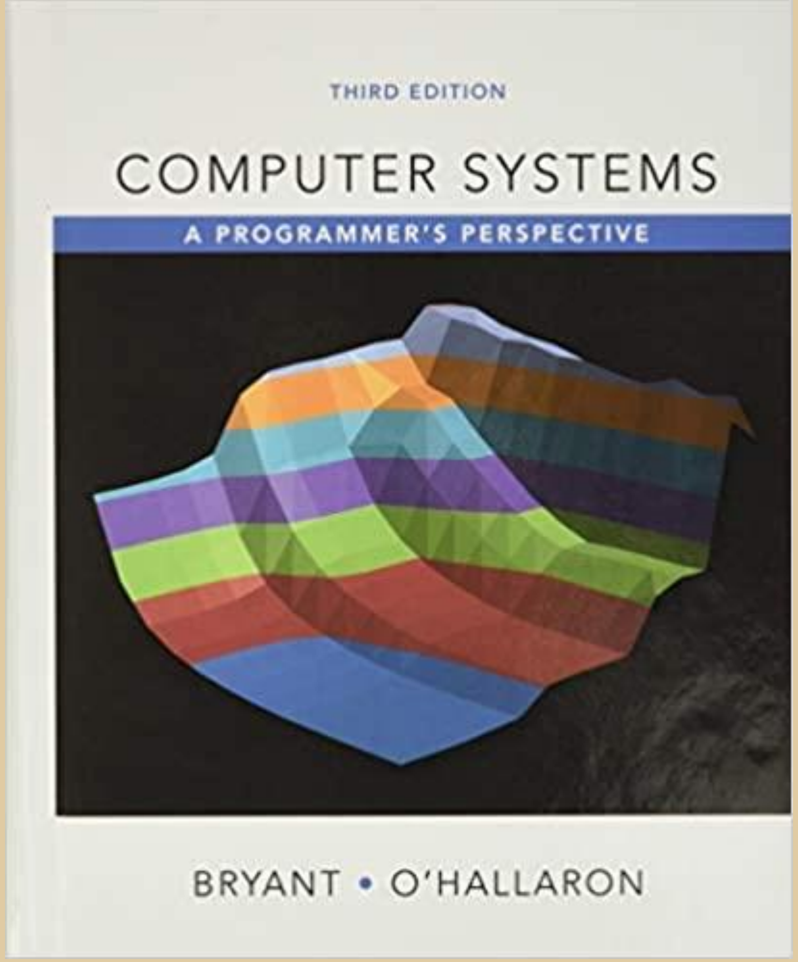
Week 3

Discussion

Q&A

Further Reading

- Read Chapter#2



Summary

- Scientific Notation
- Standard IEEE Floating Point Presentation
- Encoding and decoding Examples
- Machine epsilon
- Numerical Errors(round off and truncate error)

Floating Point in C\C++

- C guarantees two levels
- float: single precision
- double: double precision
- Conversions

Casting among int, float, and double changes numeric values

- Double or float to int:
  - truncates fractional part
  - like rounding toward zero
  - not defined when out of range: generally saturates to TMin or TMax
- int to double: exact conversion as long as int has ≤ 53-bit word size
- int to float: will round according to rounding mode.

ALGONQUIN COLLEGE

NUMERICAL COMPUTING
CST8233
Interpolation

By

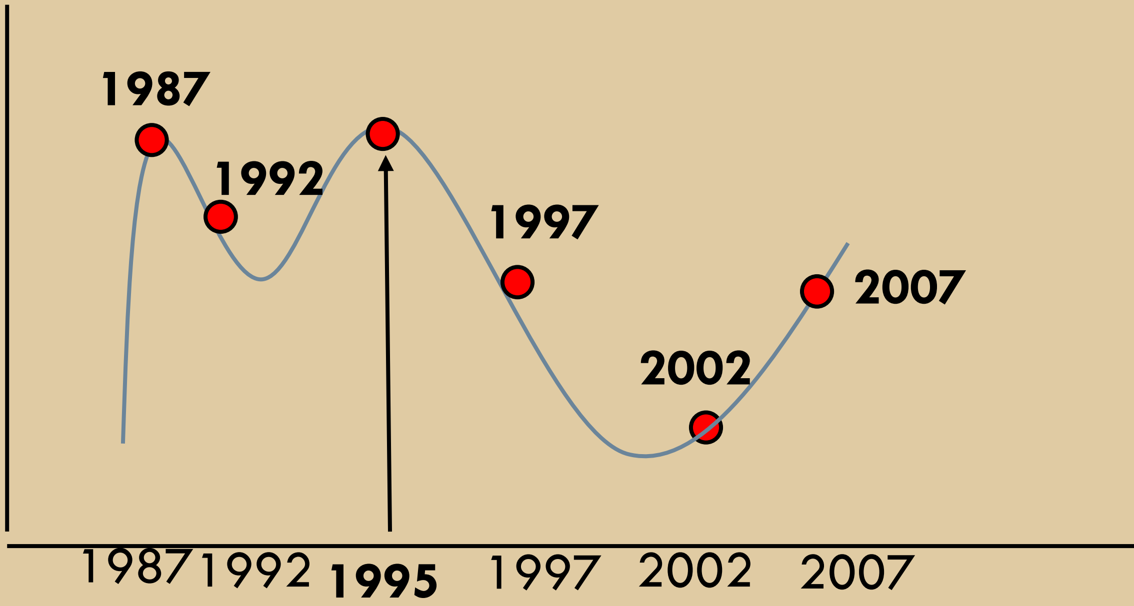
Hala Own, Ph.D.

Learning Objectives

- Interpolation Problem
- Lagrange Interpolation
- Computing Cardinal Function
- Truncation Error
- Applications of Interpolation
- Interpolation with Newton Method

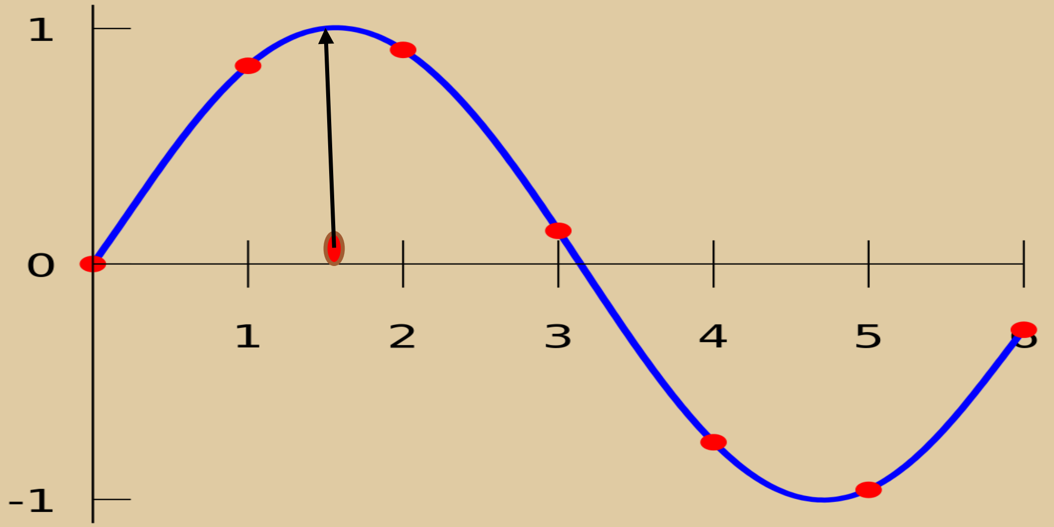
Interpolation Motivation

If we are finding out the number of sunspots in **1995** when we know the number of sunspots in the year 1987, 1992, 1997, 2002,2007 and so on. i.e. the figure of population are available for 1987, 1992, 1997, 2002,2007 etc., then the process of finding the population of **1995** is known as **interpolation**



What is Interpolation ?

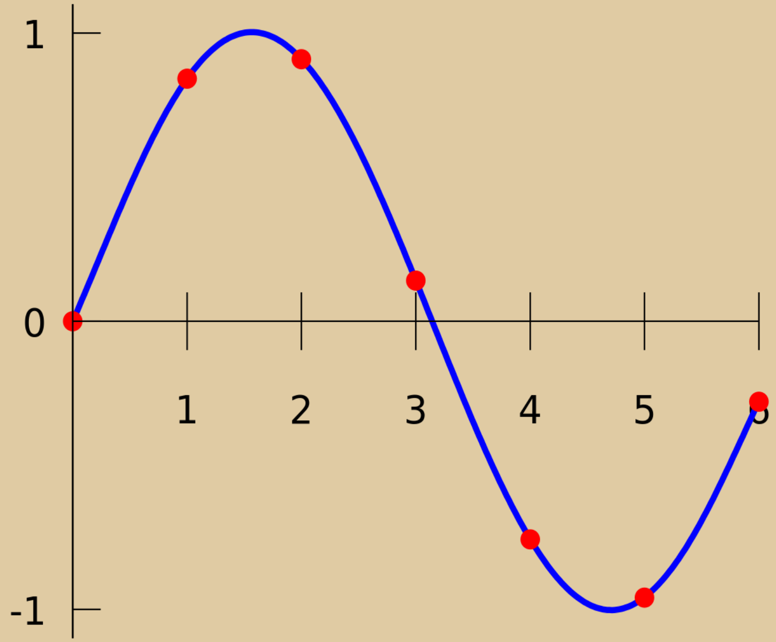
- **Interpolation** interpolation is a method of constructing new data points within the range of a discrete set of known data points.
- **Interpolation(another formal defination)** is the process of deriving a simple function from a set of data points so that the function passes through all the points.
- **The goal** is to use this function to estimate a new data points **within the range** of a discrete set of known data points.



Polynomial Interpolation

Polynomials are the most common choice of interpolants because they are easy to:

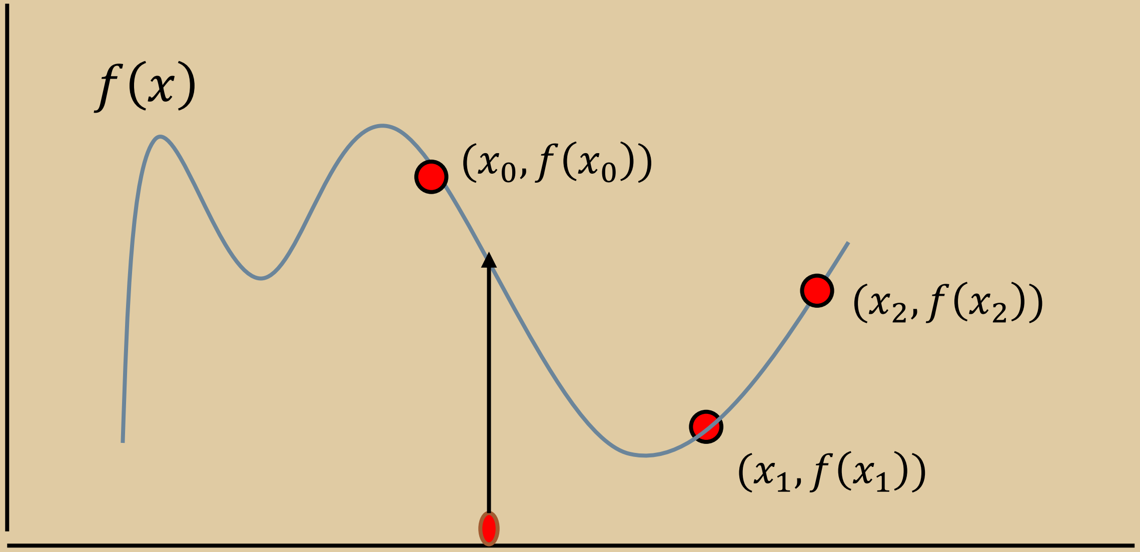
- Evaluate
- Differentiate, and
- Integrate.



Uniqness of the Polynomials

For The given interpolation points , there is a **unique n^{th} order** polynomial that passes through a given set of points.

Therefore, our methodology to estimating a new values between the given a set of data points, We first fit a function that exactly passes through these data points and then evaluate the new values using this function.



Interpolation, Formal Problem description

Given a set of **n+1** points, (x_0,y_0), (x_1,y_1),……………..(x_{n+1},y_{n+1})

Find an n[th] order polynomial

$$P_n = a_0 + a_1 x + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots + a_n x^n.$$

that passes through all points, such that:

$$P_n(x_i) = (Y_i) \; for \quad i = 0,1,2,\ldots,n+1$$

the goal is to find the coefficients a_0 , a_1 , . . ., a_n

Problem Constrains

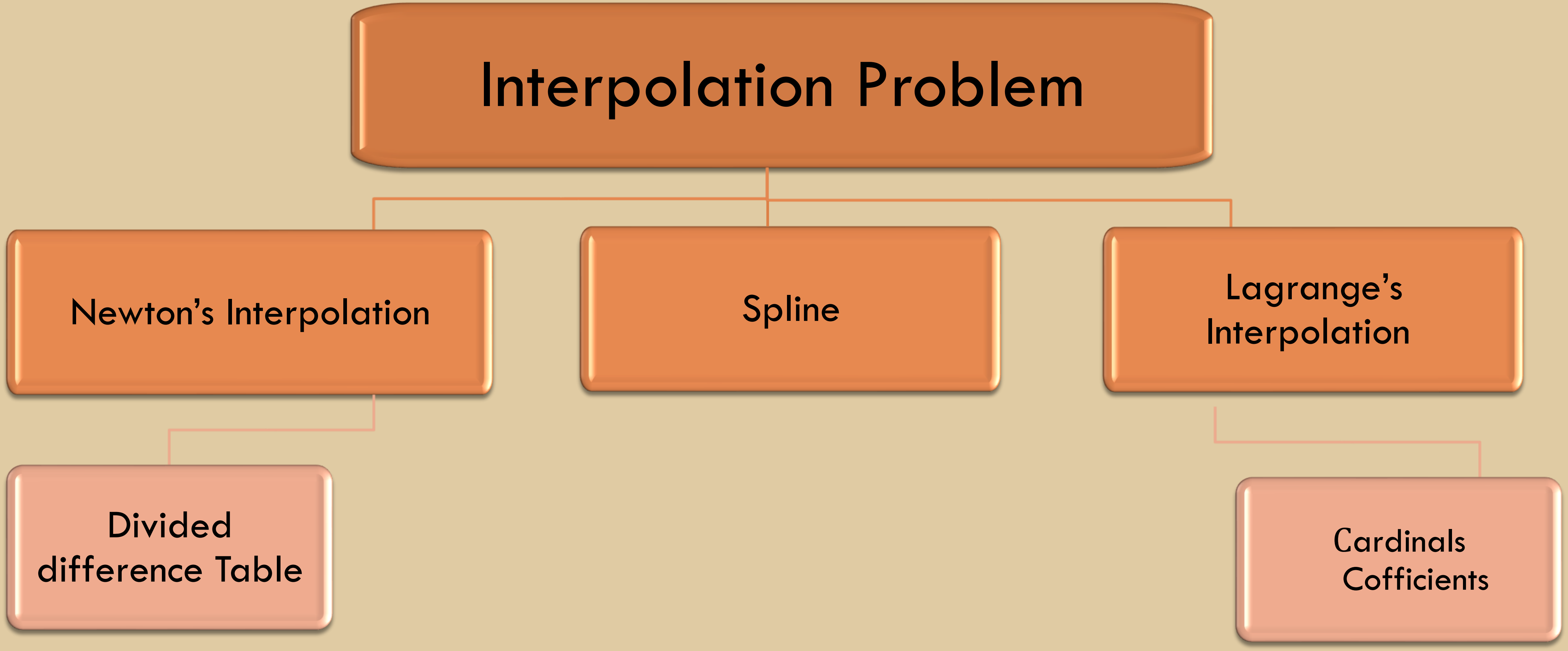
- The number of data points must exceed the degree of polynomial by 1.
- The unknown data point must be in between your given data points
- The generated polynomial must exactly pass through your given data points

Polynomial (review)

- Some polynomials have special names based on their degree and the number of terms they have.

Degree	Name
0	Constant
1	Linear
2	Quadratic
3	Cubic

Numerical Interpolation



Lagrange Interpolating Polynomials

Lagrange Interpolation Formula

Problem: Given the data set (x_0,y_0), (x_1,y_1),………….. (xn,yn),

$$f_n(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x)$$

$$\ell_i(x) = \prod_{j=0, j\neq i}^{n} \frac{(x - x_j)}{(x_i - x_j)}$$

$\ell_i(x)$ are called the cardinals.

The cardinals are n^{th} order polynomials:

$$\ell_i(x_j) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$$

$$f_n(x) = l_0(x_0) * Y_0 + l_1(x_1) * Y_1 + l_2(x_2) * Y_2 + \cdots l_n(x_n) * Y_n$$

Lagrange Linear Interpolation Example

- Given two points (2, 3) and (5,7), find the linear polynomial passing through the two points.

  Estimate the value when x= 3

x	F(x)
2	3
5	7

Solution: $P_1(x) = f(x_0)\ell_0(x) + f(x_1)\ell_1(x)$

$$\ell_0(x) = \frac{(x - x_1)}{(x_0 - x_1)}, \qquad \ell_1(x) = \frac{(x - x_0)}{(x_1 - x_0)}$$

$$P_1(x) = \frac{(x-5)}{(2-5)} \times 3 + \frac{(x-2)}{(5-2)} \times 7$$

$$P_1(x) = (5 - x) + \frac{7}{3}(x - 2)$$

$$P_1(3) = (5 - 3) + \frac{7}{3}(3 - 2) = 4.3333$$

Lagrange Interpolation Quadratic Example

Given the following data set find the polynomial to interpolate these data using Lagrange polynomial. Estimate the value when x= **0.5**

x	0	2	1
y	2	-1	7

$$P_2(x) = f(x_0)\ell_0(x) + f(x_1)\ell_1(x) + f(x_2)\ell_2(x)$$

$$\ell_0(x) = \frac{(x - x_1)}{(x_0 - x_1)}\frac{(x - x_2)}{(x_0 - x_2)} \qquad = \frac{(x-2)}{(0-2)}\frac{(x-1)}{(0-1)}$$

$$\ell_1(x) = \frac{(x - x_0)}{(x_1 - x_0)}\frac{(x - x_2)}{(x_1 - x_2)} \qquad = \frac{(x-0)}{(2-0)}\frac{(x-1)}{(2-1)}$$

$$\ell_2(x) = \frac{(x - x_0)}{(x_2 - x_0)}\frac{(x - x_1)}{(x_2 - x_1)} \qquad = \frac{(x-0)}{(1-0)}\frac{(x-2)}{(1-2)}$$

$$\frac{(x-2)}{-2}\frac{(x-1)}{-1} * 2 + \frac{x}{2}\frac{(x-1)}{1} * -1 + \frac{x}{1}\frac{(x-2)}{-1} * 7$$

$$P_2(x) = (x-2)(x-1) - \frac{1}{2}x(x-1) - 7x(x-2)$$

$$P_2(0.5) = (0.5-2)(0.5-1) - 0.5(0.5-1) - 7 * 0.5(0.5-2) = 6.25$$

Dr. Hala Own

Lagrange Interpolation Cubic Example

Given the following data set find the cardinal coefficient of these data using Lagrange polynomial. **What is he value of $\ell_0(1.5)$, $\ell_1(1.5)$, $\ell_2(1.5)$,**

$$P_4(x) = \sum_{i=0}^{4} f(x_i)\ell_i = \ell_0 + 3\ell_1 + 2\ell_2 + 5\ell_3 + 4\ell_4$$

$$\ell_0 = \frac{(x-1)}{(0-1)}\frac{(x-2)}{(0-2)}\frac{(x-3)}{(0-3)}\frac{(x-4)}{(0-4)} = \frac{(x-1)(x-2)(x-3)(x-4)}{24}$$

$$\ell_1 = \frac{(x-0)}{(1-0)}\frac{(x-2)}{(1-2)}\frac{(x-3)}{(1-3)}\frac{(x-4)}{(1-4)} = \frac{x(x-2)(x-3)(x-4)}{-6}$$

$$\ell_2 = \frac{(x-0)}{(2-0)}\frac{(x-1)}{(2-1)}\frac{(x-3)}{(2-3)}\frac{(x-4)}{(2-4)} = \frac{x(x-1)(x-3)(x-4)}{4}$$

$$\ell_3 = \frac{(x-0)}{(3-0)}\frac{(x-1)}{(3-1)}\frac{(x-2)}{(3-2)}\frac{(x-4)}{(3-4)} = \frac{x(x-1)(x-2)(x-4)}{-6}$$

$$\ell_4 = \frac{(x-0)}{(4-0)}\frac{(x-1)}{(4-1)}\frac{(x-2)}{(4-2)}\frac{(x-3)}{(4-3)} = \frac{x(x-1)(x-2)(x-3)}{24}$$

$$\ell_0(1.5) = \frac{(1.5-1)(1.5-2)(1.5-3)(1.5-4)}{24} \approx -\frac{0.9375}{24}$$

$$\approx -0.03906$$

x	y
0	1
1	3
2	2
3	5
4	4

Your Turn

Given the following data set find the polynomial to interpolate these data using Lagrange polynomial. Estimate the value when x= 8

$$P_3(x) = f(x_0)\ell_0(x) + f(x_1)\ell_1(x) + f(x_2)\ell_2(x) + f(x_3)\ell_3(x)$$

x	*f(x)*
0	3
2	60
4	90
10	120

$$f(x) = \frac{(x-2)(x-4)(x-10)}{(0-2)(0-4)(0-10)} \times 3 + \frac{(x-0)(x-4)(x-10)}{(2-0)(2-4)(2-10)} \times 60$$

$$+ \frac{(x-0)(x-2)(x-10)}{(4-0)(4-2)(4-10)} \times 90 + \frac{(x-0)(x-2)(x-4)}{(10-0)(10-2)(10-4)} \times 120$$

$$x = 8 \rightarrow f(8) \approx 109.8$$

Applying Lagrange for complicated function

- Following is a table of values for f(x) = tan x for a few values of x

Use linear interpolation to estimate tan(1.15).

x	Tan(x)
1	1.5574
1.1	1.9648
1.2	2.5722
1.3	6.6021

Soln.

Choose x= 1.1 and x= 1.2 (why?????)

$$P_1(x) = \frac{(1.15-1.2)}{(1.1-1.2)} * 1.9648 + \frac{(1.15-1.1)}{(1.2-1.1)} * 2.57 \approx 0.2685$$

The true value is tan (1.15) = 2.2345

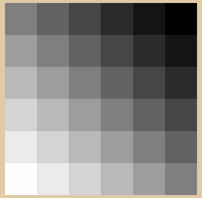
$$absolute\ error = |true\ value\ - Approximate\ value|$$

$$absolute\ error = ????$$

Check the results using higher order polynomial

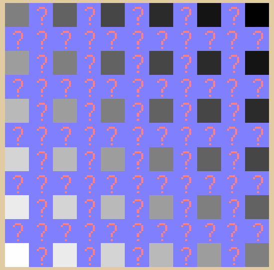
Interpolation Application: Image Resize Example

Image interpolation works in two directions and tries to achieve a best approximation of a pixel's color and intensity based on the values at surrounding pixels. The following example illustrates how resizing / enlargement works:

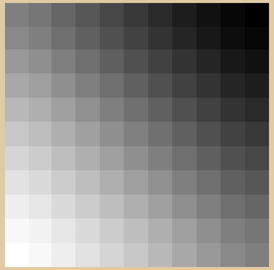


Original

Enlarge 183%



Before Interpolation



After Interpolation

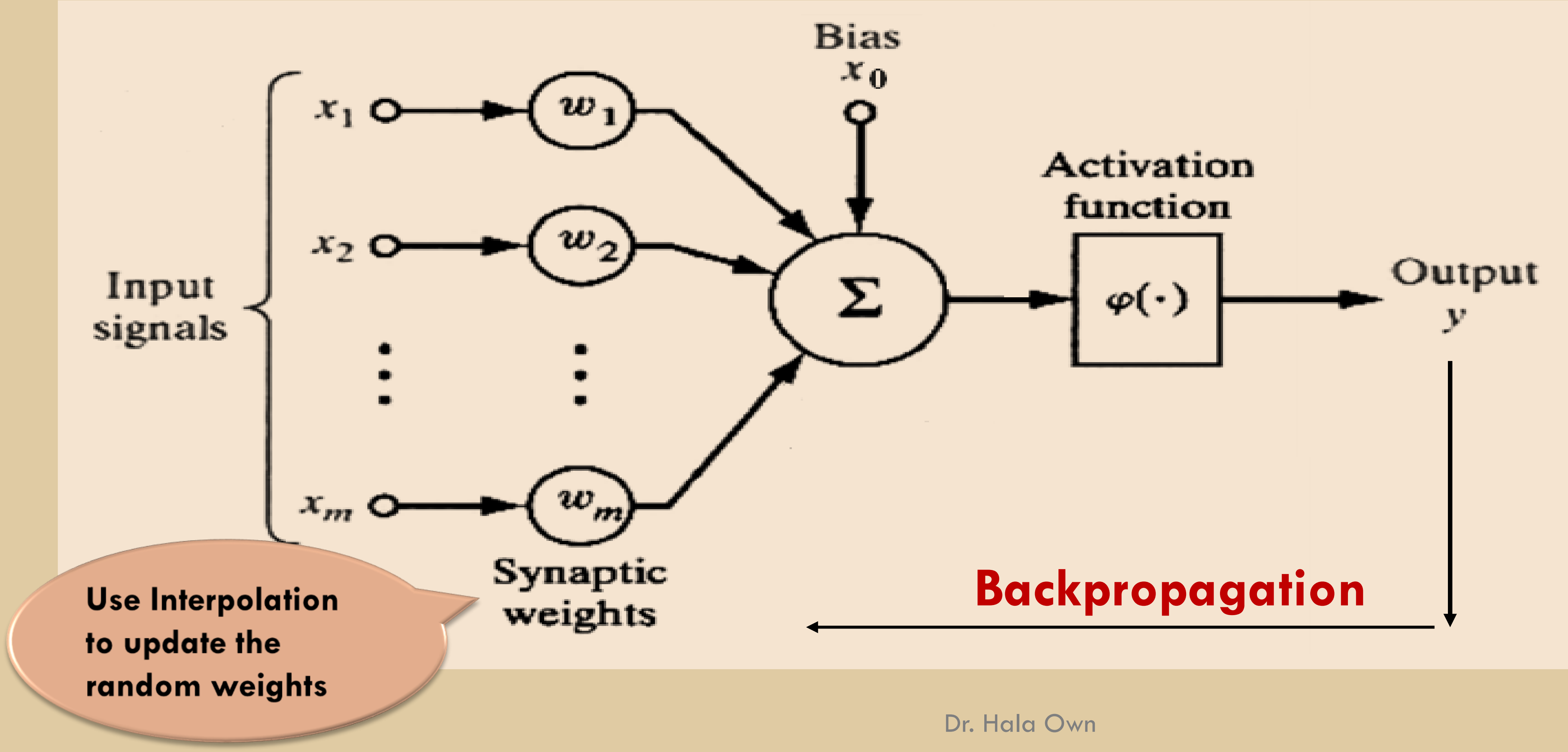
The more you know about the surrounding pixels, the better the interpolation will become, and interpolation can never add detail to your image which is not already present.

Interpolation Application: time Series Interpolation

Time series analysis has been critical in the development of modern economics and finance. Typically, a time series uses a regular polling interval, such as **once a week, once a month, or once a quarter**. However, some time series, such as the trading data, may be irregularly spaced. Working with time series sometimes forces the analyst to interpolate a reasonable value to fit into an analysis.

Year	Y
1996	210
1997	380
1998	600
1999	860
2000	1024
2001	1300
?	
2003	1500
2004	1850
2005	2031
2006	2100
2007	2424

The Lagrange Interpolation Polynomial for Neural Network Learning



Summary

- Applying Lagrange Interpolation *(Liner, Quadratic, and Cubic)*
- Applying Lagrange for complicated function *(Liner, Quadratic, and Cubic)*
- Truncate Error
- Interpolation Applications(Image resize, Time series, Neural Network)

Lab4: Interpolation

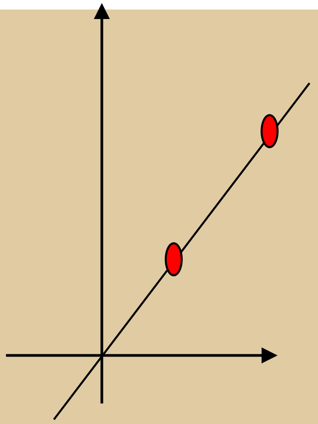
Due Week 4

- Discussion
- Q&A

Hybrid
Newton's Divided-difference Interpolating Polynomials

Linear Interpolation

Given any two points, $(x_0, f(x_0)), (x_1, f(x_1))$



The line that interpolates the two points is:

$$P_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0)$$

Example :

Find a polynomial that interpolates (1,2) and (2,4).

$$P_1(x) = 2 + \frac{4-2}{2-1}(x-1) \approx 2x$$

x	F(x)
1	2
2	4

Quadratic Interpolation

- Given any three points: $(x_0, f(x_0)), (x_1, f(x_1)), and\ (x_2, f(x_2))$
- The polynomial that interpolates the three points is:

x_0	$f(x)$
x_0	$f(x_0)$
x_1	$f(x_1)$
x_2	$f(x_2)$

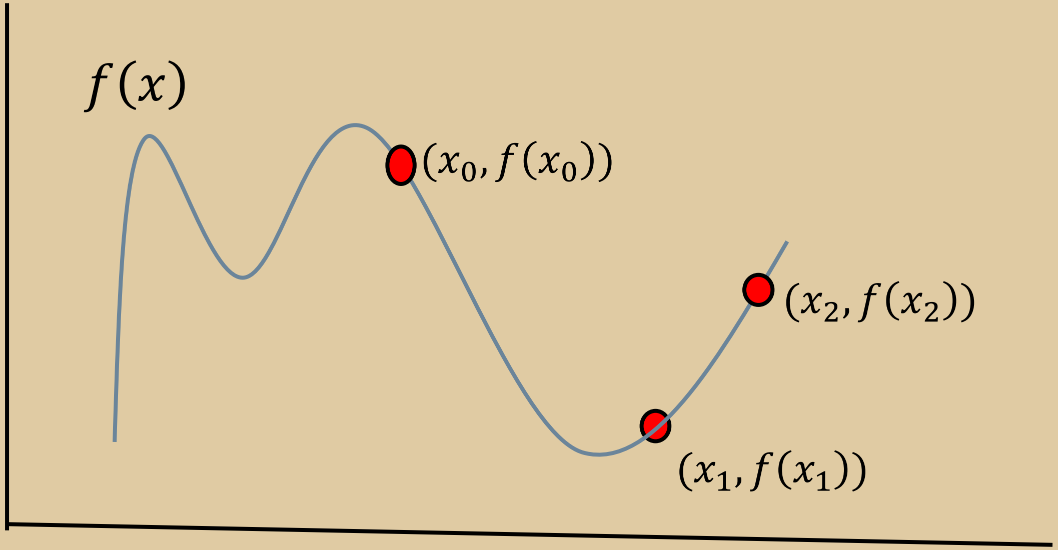
$$P_2(x) = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1)$$

where:

$$b_0 = f(x_0)$$

$$b_1 = f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

$$b_2 = f[x_0, x_1, x_2] = \frac{\frac{f(x_2) - f(x_1)}{x_2 - x_1} - \frac{f(x_1) - f(x_0)}{x_1 - x_0}}{x_2 - x_0}$$



Divided **Difference** Table

Entries of the divided difference table are obtained from the data table using simple operations.

x	F[ ]	F[,]	F[, ,]
0	-5	2	-4
1	-3	6	
-1	-15		

x_i	$f(x_i)$
0	-5
1	-3
-1	-15

The first two column of the table are the data columns.

Third column: First order differences.

Fourth column: Second order differences.

Divided Difference Table

x	F[ ]	F[,]	F[, ,]
0	-5	2	-4
1	-3	6	
-1	-15		

$$\frac{-3-(-5)}{1-0} = 2$$

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

x_i	y_i
0	-5
1	-3
-1	-15

Divided Difference Table

x	F[ ]	F[,]	F[, ,]
0	-5	2	-4
1	-3	6	
-1	-15		

$$\frac{-15-(-3)}{-1-1}=6$$

$$f[x_1,x_2]=\frac{f[x_2]-f[x_1]}{x_2-x_1}$$

x_i	y_i
0	-5
1	-3
-1	-15

Divided Difference Table

x	F[ ]	F[,]	F[, ,]
0	-5	2	-4
1	-3	6	
-1	-15		

$$\frac{6-(2)}{-1-(0)} = -4$$

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

x_i	y_i
0	-5
1	-3
-1	-15

Divided Difference Table

x	F[ ]	F[,]	F[, ,]
0	-5	2	-4
1	-3	6	
-1	-15		

x_i	y_i
0	-5
1	-3
-1	-15

$$P_2(x) = -5 + 2(x-0) - 4(x-0)(x-1)$$

$$P_2(x) = F[x_0] + F[x_0,x_1](x-x_0) + F[x_0,x_1,x_2](x-x_0)(x-x_1)$$

General n[th] Order Interpolation

Given any **n+1 points**: $\left(x_0, f(x_0)\right), \left(x_1, f(x_1)\right), \ldots, \left(x_n, f(x_n)\right)$

The **polynomial** that interpolates all points is:

$$P_n(x) \quad = b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + \ldots + b_n(x - x_0)\ldots(x - x_{n-1})$$

$$b_0 = f(x_0)$$

$$b_1 = f[x_0, x_1]$$

....

$$b_n = f[x_0, x_1, \ldots, x_n]$$

Divided Difference Table

x	F[X]	F[,]	F[, ,]	F[, , ,]
x_0	F[x0]	$F[x_0,x_1]$	$F[x_0,x_1,x_2]$	$F[x_0,x_1,x_2,x_3]$
x_1	$F[x_1]$	$F[x_1,x_2]$	$F[x_1,x_2,x_3]$	
x_2	$F[x_2]$	$F[x_2,x_3]$		
x_3	$F[x_3]$			

Newton Divided Difference General Equation

$$P_n(x) = \sum_{i=0}^{n} \left\{ F[x_0, x_1, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j) \right\}$$

Example 2:Higher Order Polynomial

- Construct P_4 polynomial to interpolate the data using divided differences Method.

x	f(x)
2	3
4	5
5	1
6	6
7	9

Example 2:Higher Order Polynomial,

Solution

x	f(x)	f[,]	f[, ,]	f[, , ,]	f[, , , ,]
2	**3**	**1**	**-1.6667**	**1.5417**	**-0.6750**
4	**5**	**-4**	**4.5**	**-1.8333**	
5	**1**	**5**	**-1**		
6	**6**	**3**			
7	**9**				

$$P_4(x) = b_0 + b_1(x-x_0) + b_2(x-x_0)(x-x_1) + b3(x-x_0)(x-x_1)(x-x_2) + b4(x-x_0)(x-x_1)(x-x_2)(x-x_3)$$

$$P_4 = 3 + 1(x-2) - 1.6667(x-2)(x-4) + 1.5417(x-2)(x-4)(x-5)$$
$$-0.6750(x-2)(x-4)(x-5)(x-6)$$

Divided Difference Table

i	x_i	$f(x_i)$	First	Second	Third
0	x_0	$f(x_0)$	$f[x_1, x_0]$	$f[x_2, x_1, x_0]$	$f[x_3, x_2, x_1, x_0]$
1	x_1	$f(x_1)$	$f[x_2, x_1]$	$f[x_3, x_2, x_1]$	
2	x_2	$f(x_2)$	$f[x_3, x_2]$		
3	x_3	$f(x_3)$			

The divided table difference table computes its entries starting from column 3 in a recursive. The higher-order differences are computed by taking differences of lower-order differences.

NUMERICAL COMPUTING

CST8233

Descriptive Statistics

By

Hala Own, Ph.D.

Learning Objectives

- Statistics general information
- Descriptive Statistics
  - Measures Of Tendency
  - Measure Of Variance
- Normal Distribution
- Standard Normal Distribution

Slido: Audience Interaction Platform

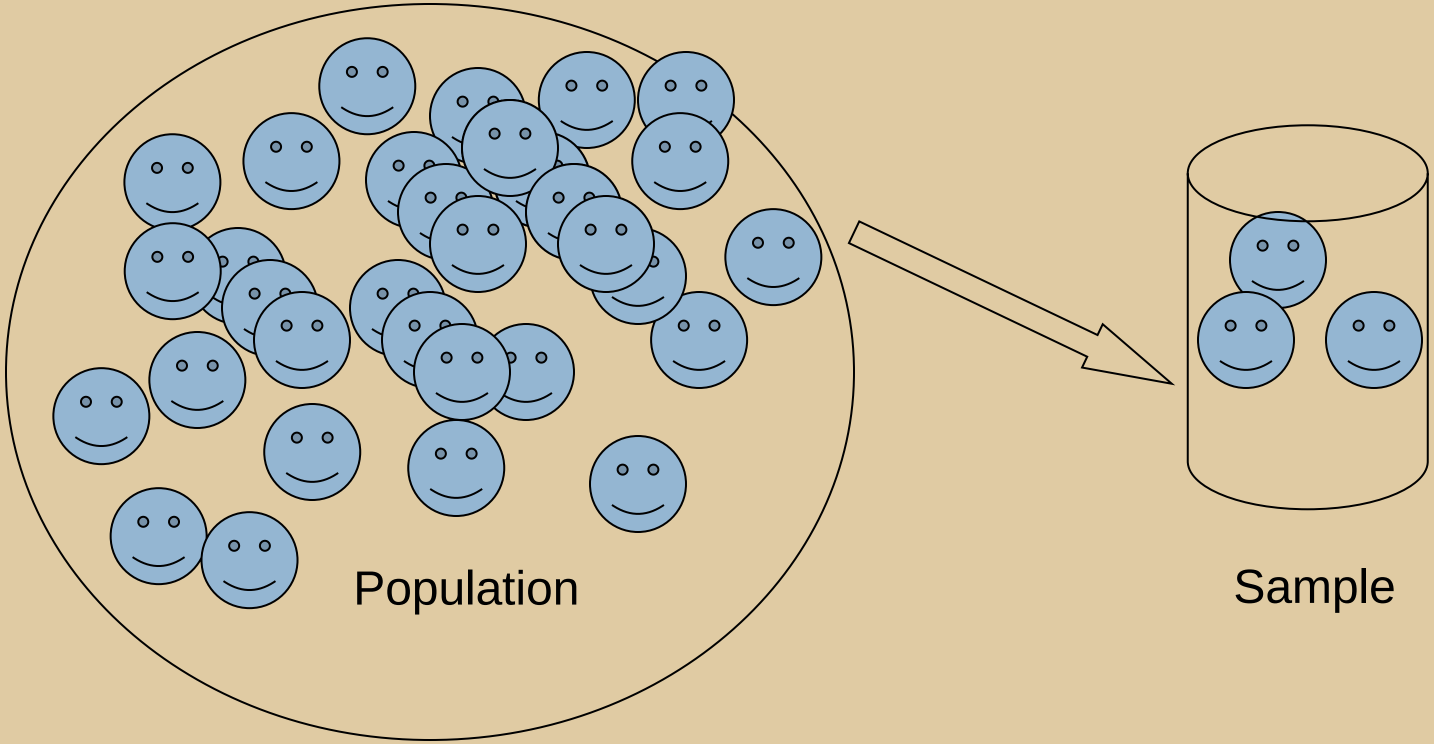
Join at

slido.com

#72284

Population and Sample

The entire group of individuals is called the **population.**



a **sample** is selected to represent the population in a research study. The goal is to use the results obtained from the sample to help answer questions about the population.

Statistics Basic Terminology

Data consists of information coming from observations, counts, measurements, or responses.

Quantitative data are measures of values or counts and are expressed as numbers.

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

Statistics is the science of collecting, organizing, analyzing, and interpreting data in order to make decisions.

Branches of Statistics

The study of statistics has two major branches: **descriptive statistics** and **inferential statistics**.

Statistics

Descriptive statistics

Involves the organization, summarization, and display of data.

Inferential statistics

Involves using a sample to draw conclusions about a population.

Descriptive Statistics

Descriptive statistics

Numerical

- Mean
- Variance
- Standard deviation
- Median
- Mode
- etc.

Graphical

- Histogram
- Boxplot
- Scatterplot
- etc.

Descriptive Statistics

Summarizing (Exploring) Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Variance
  - Standard Deviation

Slido: Audience Interaction Platform

Mean

important measure of "central tendency"

- The arithmetic average (add all of the scores together, then divide by the number of scores)

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Disadvantage

1. Means can be badly affected by outliers (data points with extreme values unlike the rest)
2. Outliers can make the mean a bad measure of central tendency or common experience

Example:
1+1+1+1+10=14/5=2.8

10+10+10+10+0=8

Median

1+1+1+1+10=14/5=2.8, median =1

10+10+10+10+0=8, median =10

- The middle value when a variable's values are ranked in order; the point that divides a distribution into two equal halves.

- When data are listed in order, the median is the point at which 50% of the cases are above and 50% below it.

Class A--IQs of 13 Students

89
93
97
98
102
106
109
110
115
119
12
131
140

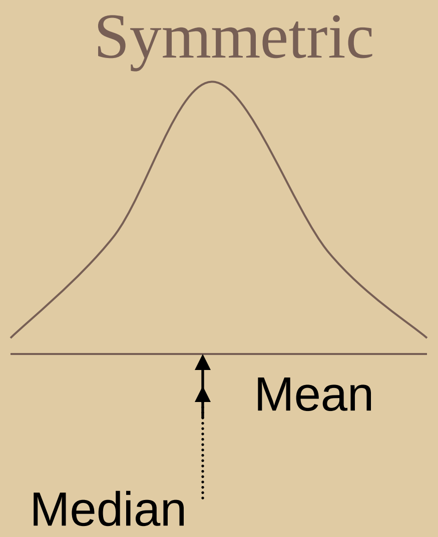
Median = 109

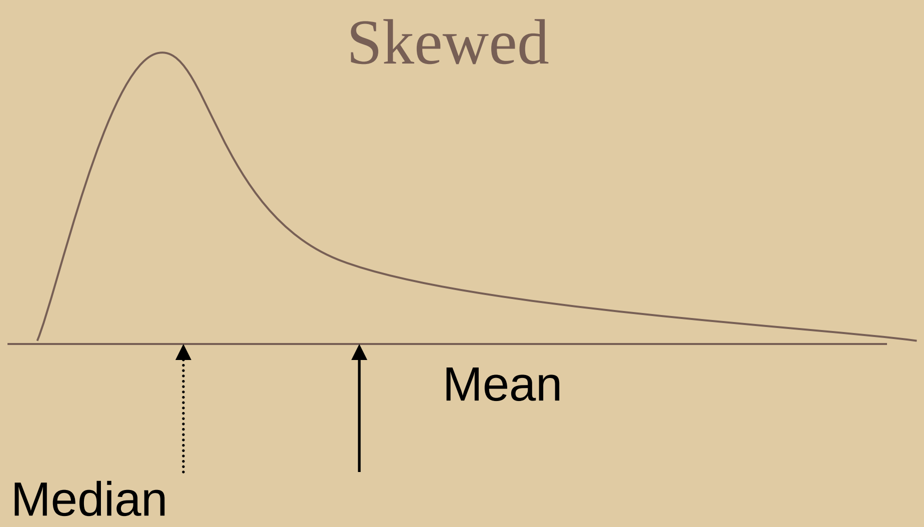
Advantage

- The median is unaffected by outliers, making it a better measure of central tendency, better describing the "typical person" than the mean when data are skewed.

Median vs Mean

If the recorded values for a variable form a **symmetric** distribution, the median and mean are identical. In **skewed data**, the mean lies further toward the skew than the median.





Mode

Mode: the most common value

1, 1, 2, 4, 6, 6, 6, 7, 7, 7, 7, 8, 8, 9, 9, 10, 12, 15

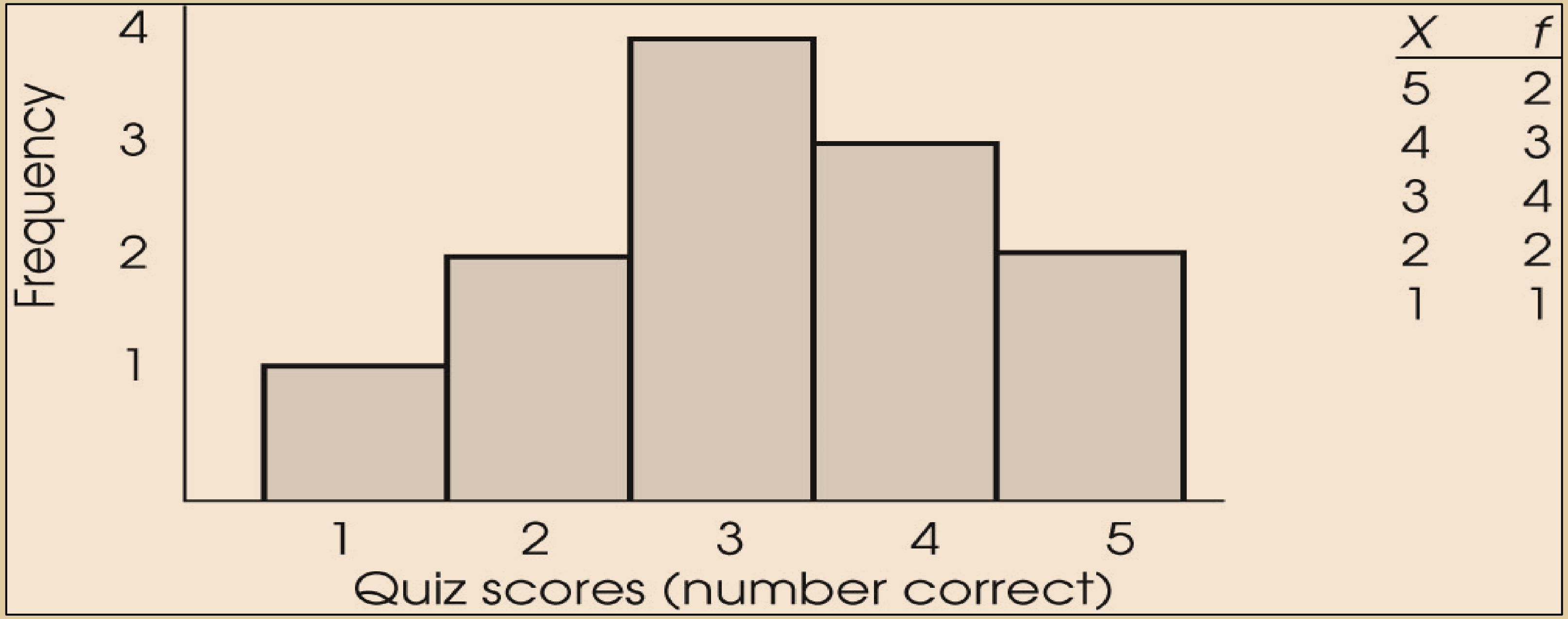
- It is possible to have more than one mode!
- It may mot be at the center of a distribution.

Frequency Distributions

- One method for simplifying and organizing data is to construct a **frequency distribution.**
- A **frequency distribution table** consists of at least two columns - one listing categories on the scale of measurement (X) and another for frequency (f).
- In the X column, values are listed from the highest to lowest, without skipping any.
- For the frequency column, tallies are determined for each value (how often each X value occurs in the data set). These tallies are the frequencies for each X value.

Histograms

In a **histogram**, a bar is centered above each score (or class interval) so that the height of the bar corresponds to the frequency and the width extends to the real limits, so that adjacent bars touch.



Descriptive Statistics

Summarizing Data:

- ✓ Central Tendency (or Groups' "Middle Values")
  - ✓ Mean
  - ✓ Median
  - ✓ Mode

- ❑ **Variation (or Summary of Differences Within Groups)**
  - ❑ **Range**
  - ❑ **Variance**
  - ❑ **Standard Deviation**

Measure of Variability

- Let's say we receive the final grades for the semester for the two classes:

ClassA	ClassB
75	75
80	100
70	50
77	85
73	65
75	98
90	52
60	

Measure of variability

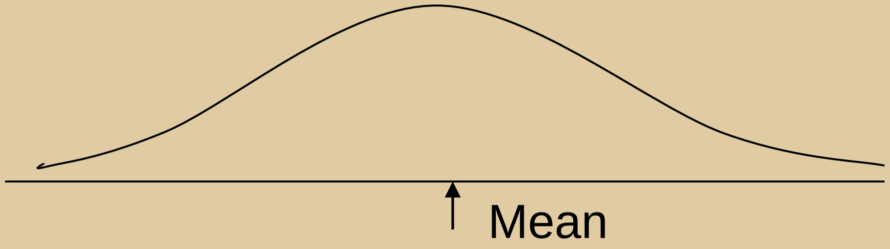
- If we measure the mean and median of both classes, we find that the mean and median are identical, both equal to **75**.

	ClassA	ClassB
	75	75
	80	100
	70	50
	77	85
	73	65
	75	98
	90	52
	60	
mean	**75**	**75**

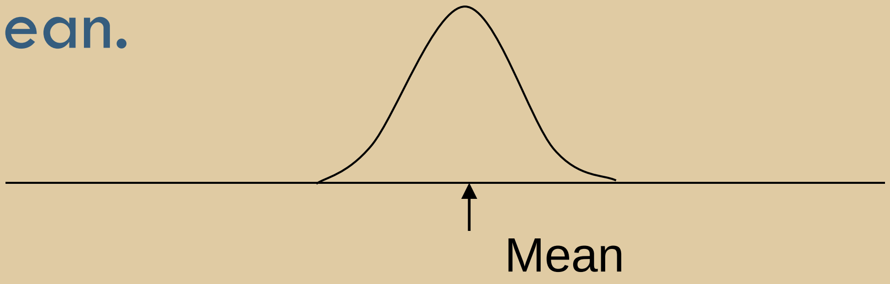
Variance

A measure of the spread of the data values . Variance help to determine the data spread size

The larger the variance, the further the individual cases are from the mean.



The smaller the variance, the closer the individual scores are to the mean.



Variance

- Most important measure of "dispersion"

Sum of squares called Deviation S_t

$$S_t = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2$$

Variance = $$\sigma^2 = \frac{S_t}{n-1} = \frac{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}{n-1}$$

Variance, Standard Deviation

Based on variance we can calculate standard deviation.

The square root of the variance reveals the **average deviation of the observations from the mean.**

Variance = $$\sigma^2 = \frac{S_t}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Standard Deviation = $$\sigma = \sqrt{\frac{S_t}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Variance, Standard Deviation

- Standard Deviation(SD) much easier to interpret

	ClassA	ClassB
	75	75
	80	100
	70	50
	77	85
	73	65
	75	98
	90	52
	60	
mean	75	75
SD	8.518887	20.44505

Calculate Variance: Example

Class A, sum of squares:

$(102 - 110.54)^2 + (115 - 110.54)^2 +$

$(126 - 110.54)^2 + (109 - 110.54)^2 +$

$(131 - 110.54)^2 + (89 - 110.54)^2 +$

$(98 - 110.54)^2 + (106 - 110.54)^2 +$

$(140 - 110.54)^2 + (119 - 110.54)^2 +$

$(93 - 110.54)^2 + (97 - 110.54)^2 +$

$(110 - 110.54) = 2825.39$

Variance = 2825.39/12= 235.45

Class A—IQs	of 13 Students
102	115
128	109
131	89
98	106
140	119
93	97
110	

$\bar{X} = 110.54$

Standard Deviation

For Class A, the Standard Deviation is:

$$\sqrt{235.45} = 15.34$$

Interpretation: The average of persons' deviation from the mean IQ of 110.54 is 15.34 IQ points.

Steps to calculate standard deviation

1. Calculate Deviation
2. Calculate Deviation squared
3. Calculate Sum of squares
4. Calculate Variance
5. Calculate Standard deviation $=\sqrt{variance}$

Standard Deviation Important Note

- Larger standard deviation means greater amounts of variation around the mean.

19 25 31 13 25 37

$\bar{Y} = 25$ $\bar{Y} = 25$

standard deviation = 3 standard deviation = 6

- Standard deviation = 0 only when all values are the same (only when you have a constant and not a “variable”)

Range

The spread, or the distance, between the lowest and highest values of a variable.

$$R = y_{\max} - y_{\min}$$

To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Class A Range = 140 - 89 = 51

Class B--IQs of 13 Students

127	162
131	103
96	111
80	109
93	87
120	105
109	

Class B Range = 162 - 80 = 82

Your Turn

Use the data in the following Table to calculate:

- Mean
- Median
- Mode
- Range of data,
- Variance
- Standard deviation

7	6	5	6	9

Solution

- Mean=6.6
- Median=6
- Mode=6
- Range=4
- Variance=2.3
- Standard deviation=1.516

Type of Data Distribution

- Normal Distribution
- Uniform distribution
- Skewed distribution

Normal Distribution Properties

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where μ is the mean and σ is the standard deviation

Properties of the normal Distribution:

- The mean is the center of this distribution and the highest point.
- The curve is symmetric about the mean. (The area to the left of the mean equals the area to the right of the mean.)
- The total area under the curve is equal to **one**.

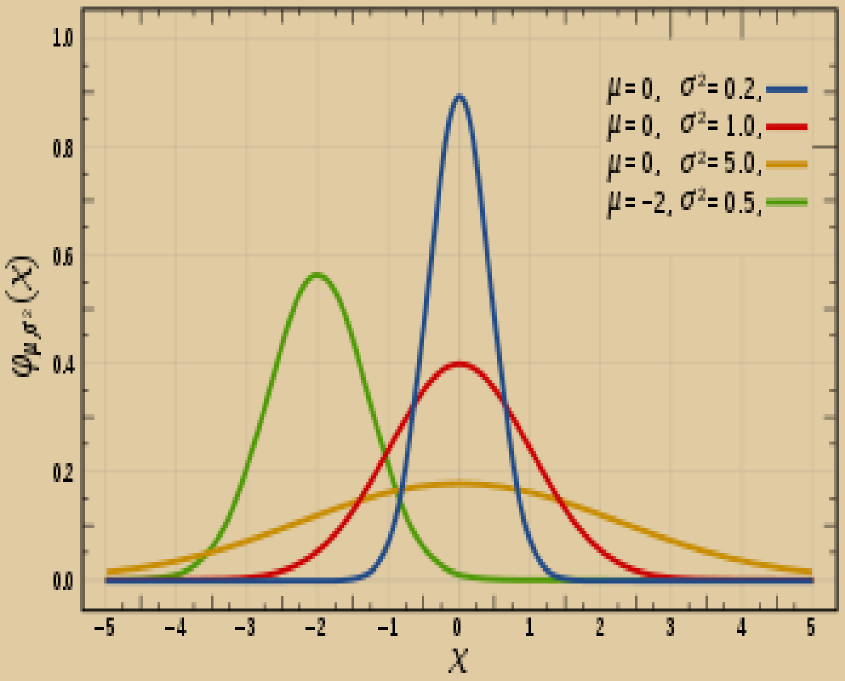
Standard Normal (Z) Distribution

https://demonstrations.wolfram.com/TheNormalDistribution/

- Problem: Unlimited number of possible normal distributions ($-\infty < \mu < \infty$, $\sigma > 0$)

Solution: Standardize the random variable to have mean 0 and standard deviation 1(z-score)

$$Y \sim N(\mu, \sigma) \quad \Rightarrow \quad Z = \frac{Y - \mu}{\sigma} \sim N(0,1)$$

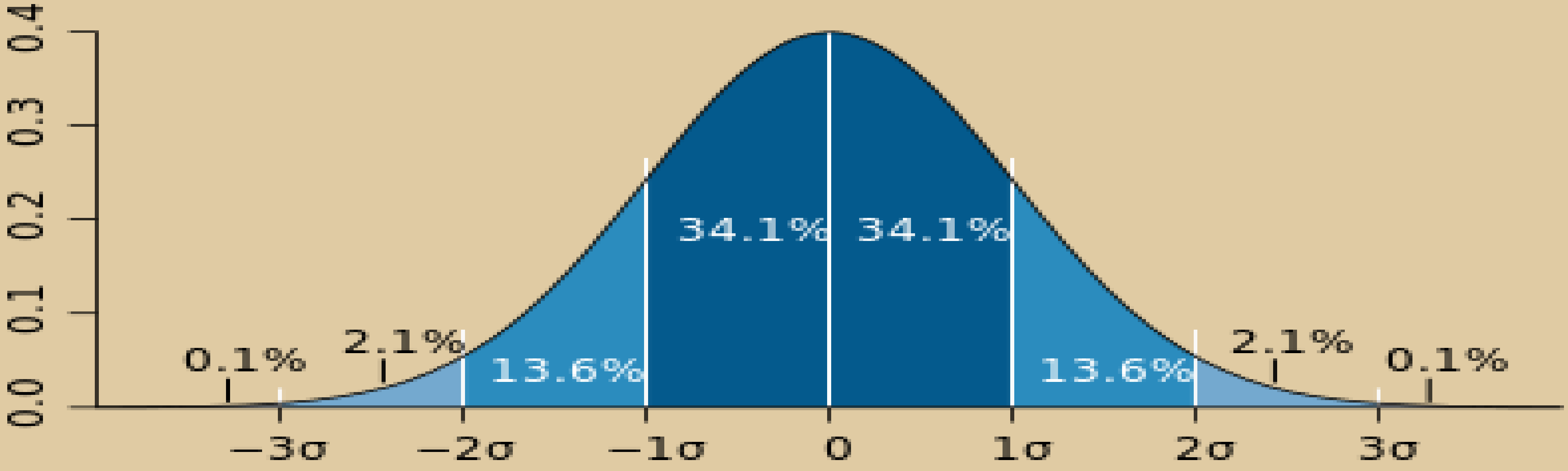


Somebody calculated all the integrals for the standard normal and put them in a table! So we never have to integrate!

Remember that *z*-scores tell us how far a value is from the mean.

Standard Normal Distribution

The process of *converting a value from a normal distribution to a value for the standard normal distribution* is called "**STANDARDIZING**" and requires the use of *z*-scores.



$P(Y \geq \mu) = 0.50 \quad P(\mu - \sigma \leq Y \leq \mu + \sigma) \approx 0.68 \quad P(\mu - 2\sigma \leq Y \leq \mu + 2\sigma) \approx 0.95$

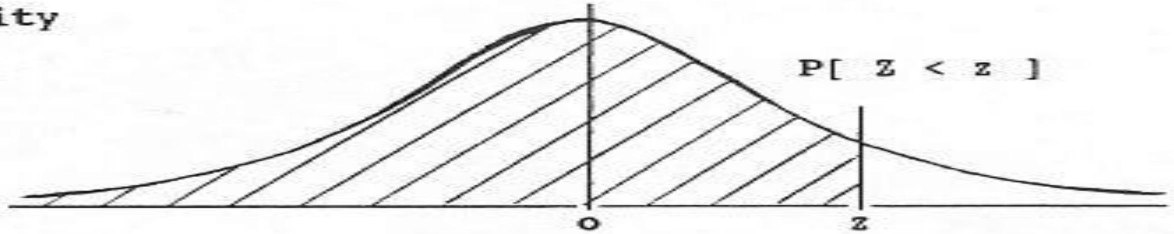
https://en.wikipedia.org/wiki/File:Standard_deviation_diagram.svg

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z
i.e.

$$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\ dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

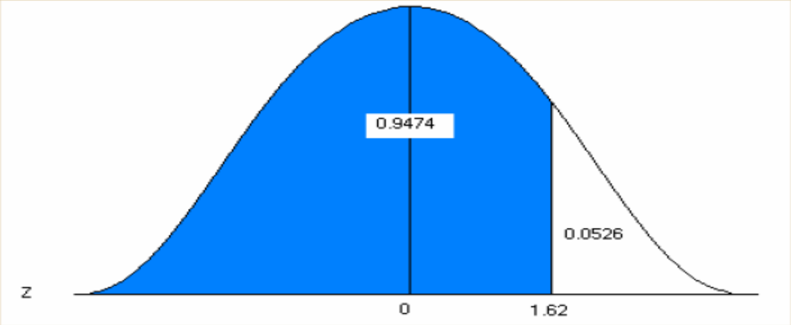
Dr. HalaOwn

Reading the Standard Normal Table

- What is the area associated with the Z-score **1.62**

- Read down the Z-column to get the first part of the Z-score (1.6).
- Read across the top row to get the second decimal place in the Z-score (0.02).
- The intersection of this row and column gives the area under the curve to the left of the Z-score.

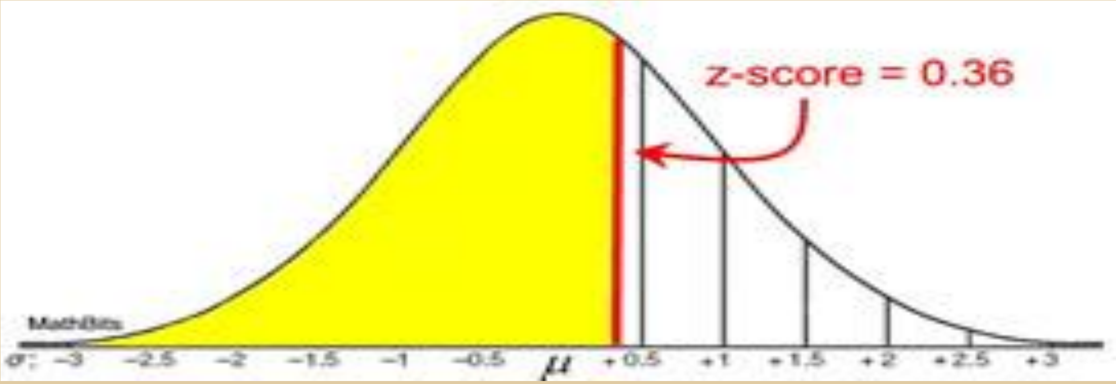
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9595	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633



Reading the Standard Normal Table

- **Example :** Find the probability that a variable has a *z*-score of less than **0.36.**

Solution:



Find the *z*-score in the table. The intersection shows 0.6406. The probability is 64.06%

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.00	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188
0.10	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142
0.20	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026
0.30	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803
0.40	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439
0.50	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904
0.60	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175

Normal distribution and probability

- What's the probability of getting a math SAT score of 575 or less, where the mean and SD of SAT exam is μ=500 and σ=50?

$$Y \sim N(\mu, \sigma) \quad \Rightarrow \quad Z = \frac{Y-\mu}{\sigma} \sim N(0,1)$$

- Solution : $Z = \frac{575-500}{50} = 1.5$

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.00	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769
1.10	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900
1.20	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796
1.30	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91308	0.91466
1.40	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922
1.50	**0.93319**	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179
1.60	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254

- i.e., A score of 575 is 1.5 standard deviations above the mean

look up Z= 1.5 in standard normal chart , the probability of getting a math SAT score of 575 or less, =0.9332

Normal distribution and probability Example

1- Find $P(Z < a)$ for $a = 1.43$

Solution: $P(Z \leq 1.43) = P(1.43) = 0.92364$

2- Find $P(Z > a)$ for $a = 1.43$

Solution: $P(Z > 1.43) =1- P(1.43)= 1- 0.92364 = 0.0764$

Normal distribution and probability Example

Find $P(a \le Z \le b)$ for $a = -1$ and $b = 1.5$

To solve: determine P(b) and P(a) and subtract P(b)-P(a).

$P(-1 \le Z \le 1.5)$

$P(Z > -1) = 0.15866$

$P(Z<1.5)= 0.93319$

0.93319- 0.15866= 0.7745

Your Turn

- Find $P(Z \leq a)$ for $a = 1.65, -1.65, 1.0, -1.0$

$P(Z \leq 1.65) = F(1.65) = .95$

Find $P(Z \geq a)$ for $a = 1.5, -1.5$

$P(Z \geq 1.5) = 1 - F(1.5) = 1 - .9332 = .0668$
$P(Z \geq -1.5) = F(1.5) = .9332$

Application of Z- score Example 2

If birth weights in a population are normally distributed with a mean of 109 g and a standard deviation of 13 g,

a. What is the chance of obtaining a birth weight of 141 g *or heavier* when sampling birth records at random?
b. What is the chance of obtaining a birth weight of 120 g *or lighter*?

Example 2 cont.

a. What is the chance of obtaining a birth weight of 141 g *or heavier* when sampling birth records at random?

$$Z = \frac{141 - 109}{13} = 2.46$$

Z of 2.46 corresponds to the area of: P(Z≥2.46) = 1-(.99305)= .0069 or .69 %

b. What is the chance of obtaining a birth weight of 120 *or lighter*?

$$Z = \frac{120 - 109}{13} = 0.85$$

From the chart → Z of 0.85 corresponds to the area of: P(Z≤0.85) = 0.8023= 80.23%

Your turn

a. Assume a normal distribution IQ test with $\mu = 100$ and $\sigma = 15$, What is the probability of randomly selecting an individual with an IQ score less than 130.

Soln.

Z= 100- 130/15= 2.0

$P(x < 130) = P(Z<2.0) = 0.9772$

Exercise

- Adult deer population weights are normally distributed with μ = 110 lb. and $\boldsymbol{\sigma}$ = 29.7 lb. The Biologist determine that a weight less than 82 lb. is unhealthy and you want to know what proportion of your population is unhealthy.

Soln.

$$z = \frac{82 - 110}{29.7} = -0.94$$

The *Z* value of 82 is 0.94 standard deviations below the mean.

$P(x<82) = P(Z<-0.94) = 0.1736$

- Approximately 17.36% of the population of adult deer is underweight

Application of Z- score : Standardize Collected Data

Mai score is 680 on the SAT mathematics test. The distribution of the SAT scores is normally distributed with a mean of 500 and a standard deviation of 100.

James takes the ACT mathematics test with score 27. the ACT scores is normally distributed with a mean of 18 and a standard deviation of 7.

Who has the higher score?

Why we Standardize our data, cont.

$$Y \sim N(\mu, \sigma) \quad \Rightarrow \quad Z = \frac{Y - \mu}{\sigma} \sim N(0,1)$$

Mai	James
μ=500 and σ=100 $Z = \dfrac{680 - 500}{100} = 1.8$	**μ=18 and σ=6** $Z = \dfrac{27 - 18}{6} = 1.5$

Mai score is higher than James's score

Lab5: Descriptive Statistics

Due

Week 5

Discussion

Q&A

Summary

- We use descriptive statistics to explore collected data
- Normal distribution
- Standard normal distribution
- Z- score
- Applications of Z- score

NUMERICAL COMPUTING

CST8233

Linear Regression

By

Hala Own, Ph.D.

Learning Objectives

- Introduction to Linear regression
- Define linear regression
- Least square Linear regression
- Identify errors of prediction in a scatter plot with a regression line
- Use regression analysis to predict the value of a dependent variable based on an independent variable
- The meaning of the regression coefficients

Make up Lecture Next Week

Join at

slido.com

#75289

Regression Analysis

- **Regression Analysis** is a statistical methodology to estimate the relationship of a **response or dependent** variable to a set of predictor or **independent variables.**

- when there is just one predictor variable, we will use **simple linear regression**. When there are two or more predictor variables, we use multiple linear regression.

What Is The Linear Regression

- Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line.

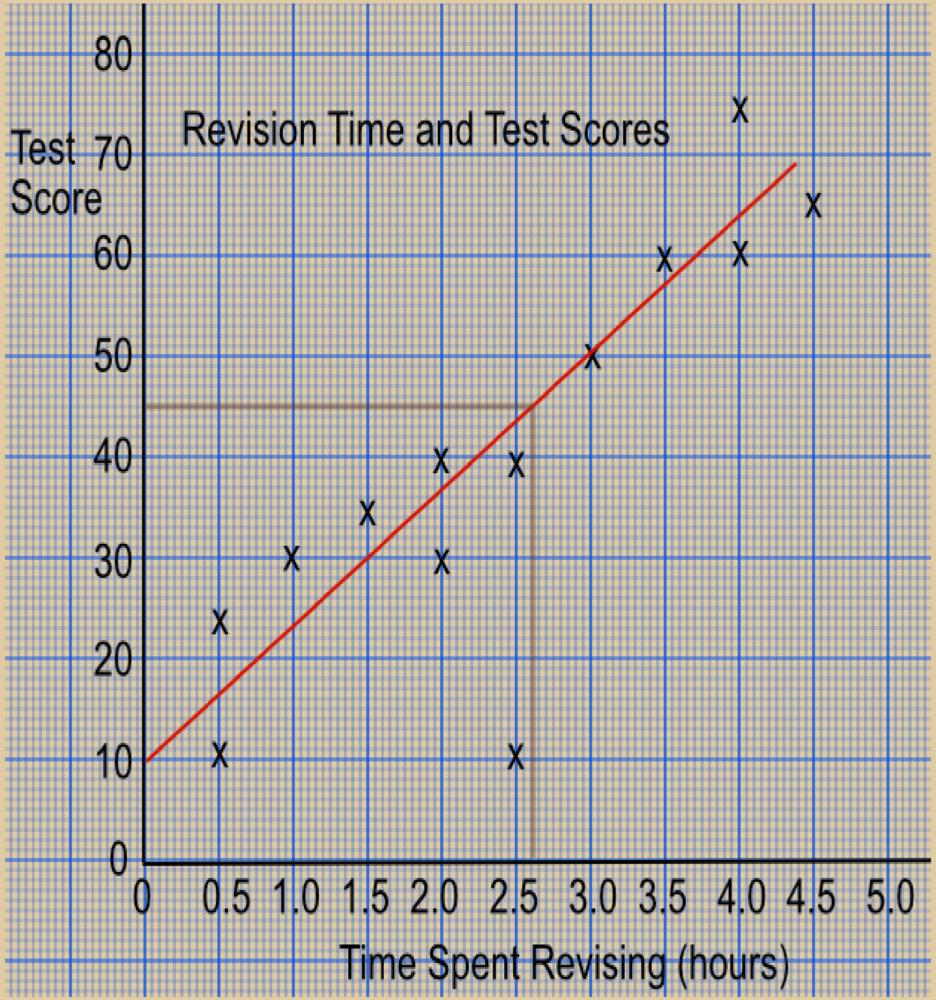
Motivation of Regression Analysis :

- To **predict** the value of a difficult to measure variable, Y, based on an easy to measure variable, X.

- **Explain** the impact of changes in an independent variable on the dependent variable.

Dependent variable: the variable you wish to explain(y)

Independent variable: the variable used to explain (x)



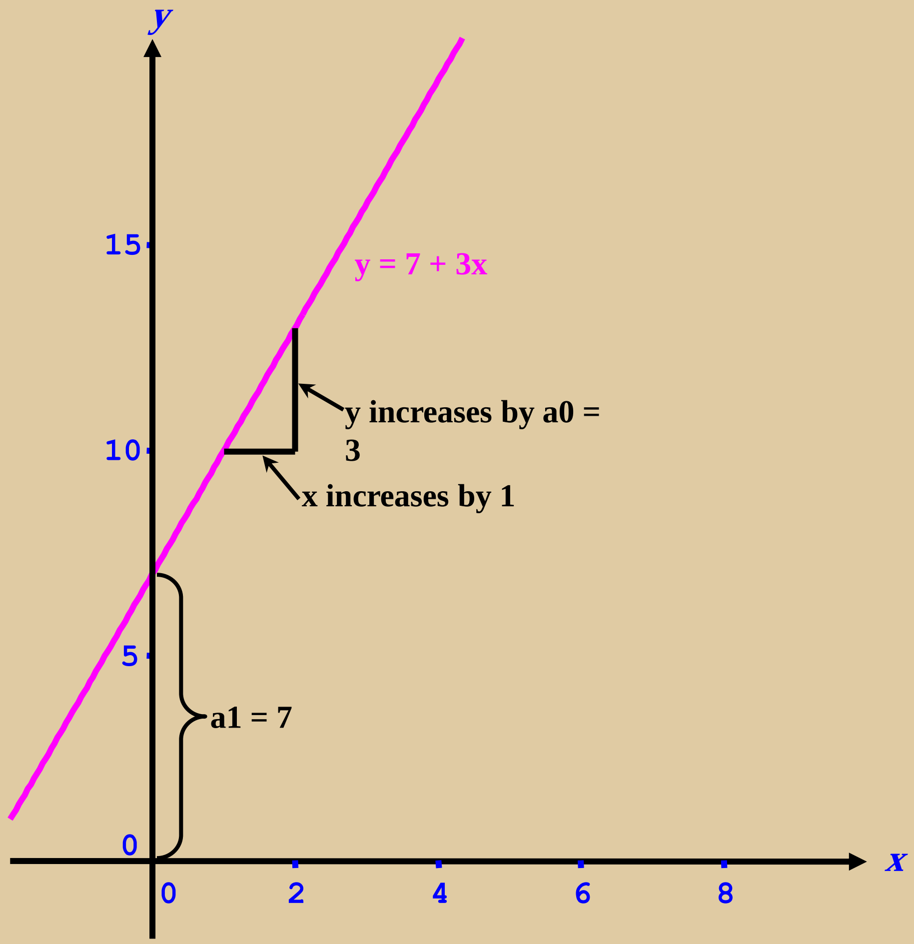
Linear Relation: Review

The general equation of a straight line is usually written as:

$$\boxed{y = a_0 x + a_1}$$

slope intercept

- Consider the equation: y=7+3x
  - The slope is 3.
    - For every increase of 1 in the x-variable, there will be an increase of 3 in the y-variable.
  - The intercept is 7.
    - When the x-variable is 0 i.e when the line cuts the y-axis at y= 7



The Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple: Only **one** independent variable, X

Linear :Relationship between X and Y is described by a linear function

Regression : Changes in Y are related to changes in X

Simple Linear Regression Equation

The simple linear regression equation estimates the regression line

Estimated (or predicted) Y

Estimate of the regression intercept

Estimate of the regression slope

Value of X for

Error

$$\widehat{Y} = a_0 + a_1 x + \varepsilon$$

where a_1 is the slope, the amount by which y changes when x increases by 1 unit and

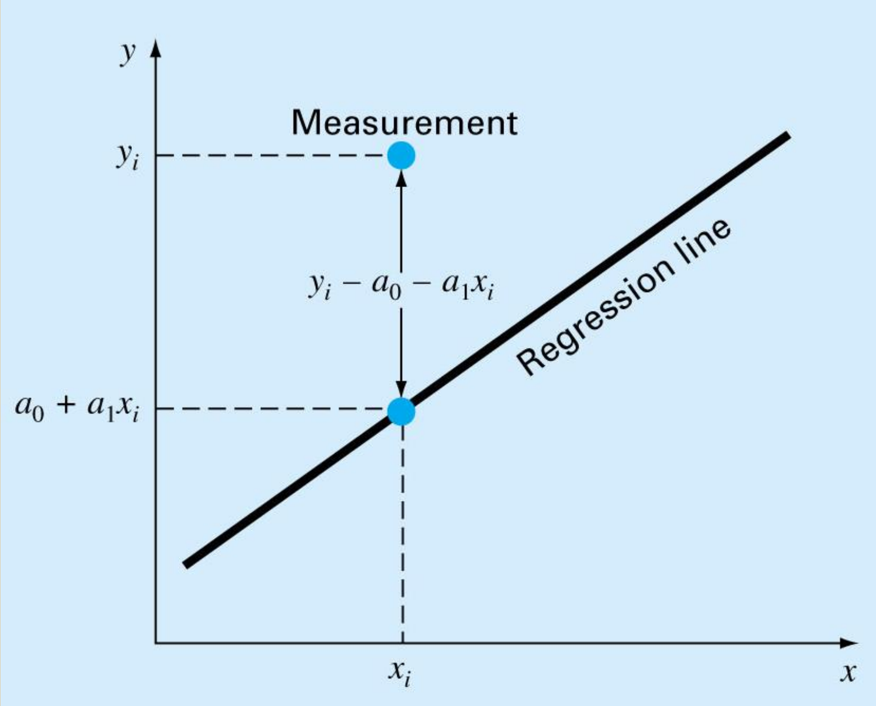
where a_0 is the intercept, the value of y when x = 0

Criteria For a “Best Fit”:(Quantification of Error of Linear Regression)

Best strategy is to *minimize the sum of the squares* of the residuals between the *measured*-y and the y *calculated* with the linear model:

$$SSE = \sum_{I=1}^{N} \left(y_{i,measured} - y_{i,estimated}\right)^2$$

$$SSE = \sum_{I=1}^{N} (y_i - a_0 - a_1 x_i)^2$$



Yields a unique line for a given set of data

Need to compute a_0 and a_1 such that SSE_r is minimized!

Linear Regression: Problem Description

Given n data points $(x_1, y_1),(x_2, y_2),......,(x_n, y_n)$

Find The line $y = a_0 + a_1 x$ that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

Problem Solution :

Minimizing the sum of squared error $SSE = \sum_{I=1}^{N} (y_i - a_0 - a_1 x_i)^2$

Solving for a_0 And a_1

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

Simple Linear Regression: Example1

- given the data that represent the hours of study vs the scores in math, Use the Linear regression model to predict what will be the math score if the student study for 8 hours.

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

x	y	x^2	xy
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
Σx: 26	**Σy: 41**	**Σx²: 168**	**Σxy: 263**

a_1=((5 * 263) – (26 * 41))/ ((**5 * (168 – 26²))**

249 / **164 = 1.5183**

a_0= 8.2 –(1.5183*5.2)= **0.3049**

$$a_0 = \bar{y} - a_1 \bar{x}$$

$$\hat{y} = a_1 x + a_0$$

$$\hat{y} = 1.518\, x + 0.305$$

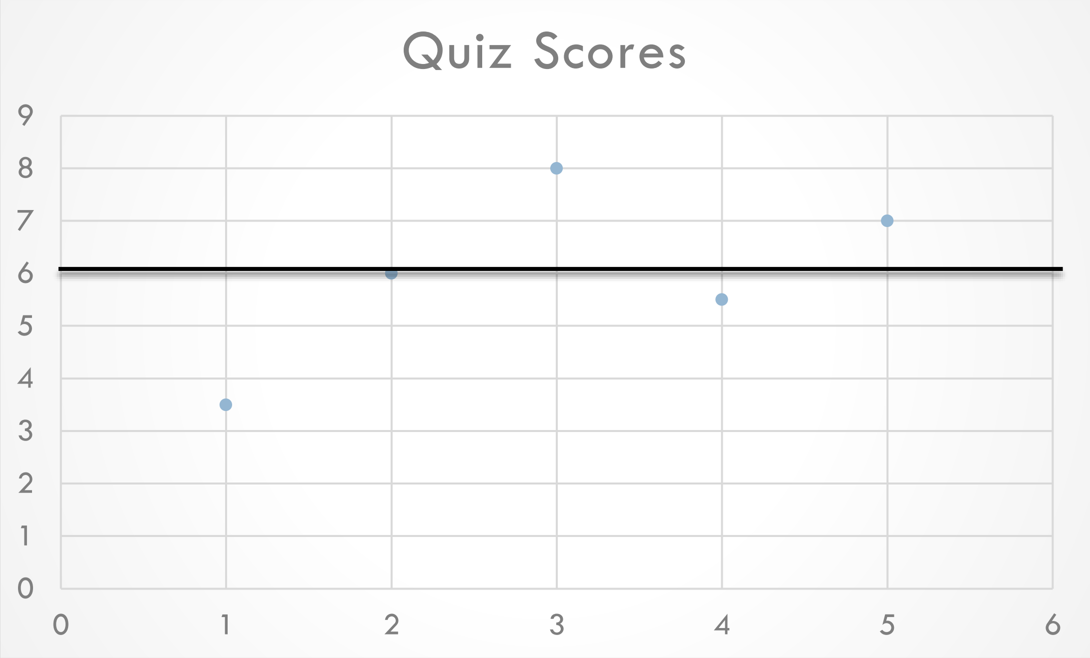
score if the student study for 8 hours **= 1.518 * 8 + 0.305 = 12.45**

"x" Hours of study	"y" grdes
2	4
3	5
5	7
7	10
9	15

Error Analysis

Student	Quiz Scores
1	3.5
2	6
3	8
4	5.5
5	7

Mean=6



Mean=6

The mean value now divides the data with values above the mean and values below it. This is called the **spread** of the data.

Goodness of Fit : SUM OF SQUARE RESIDUAL

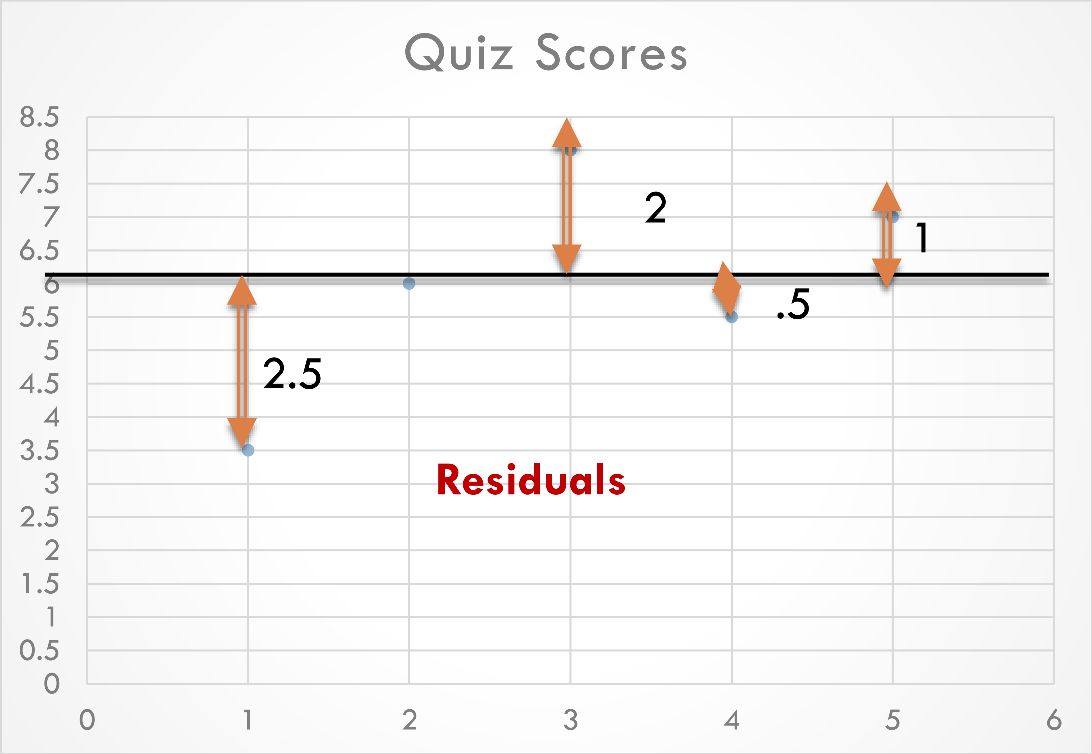
The aggregate distance of all points from the mean can be defined as Variance = $\sum_{i=0}^{n}(y_i - \bar{y})^2$

The average distance from the mean is: **standard deviation** $\sqrt{\frac{variance}{n-1}}$

student	Quiz Scores	RESIDUAL (x-mean)	RESIDUAL² $(x\text{-mean})^2$
1	3.5	-2.5	6.25
2	6	0	0
3	8	2	4
4	5.5	-0.5	0.25
5	7	-1	1

The larger the standard deviation, the farther the data are spread from the mean

Mean=6



SUM OF SQUARE RESIDUAL (SSR) =11.5

$\sigma(STANDARD\ ERROR) = 1.957$

Linear Regression: The Standard Error

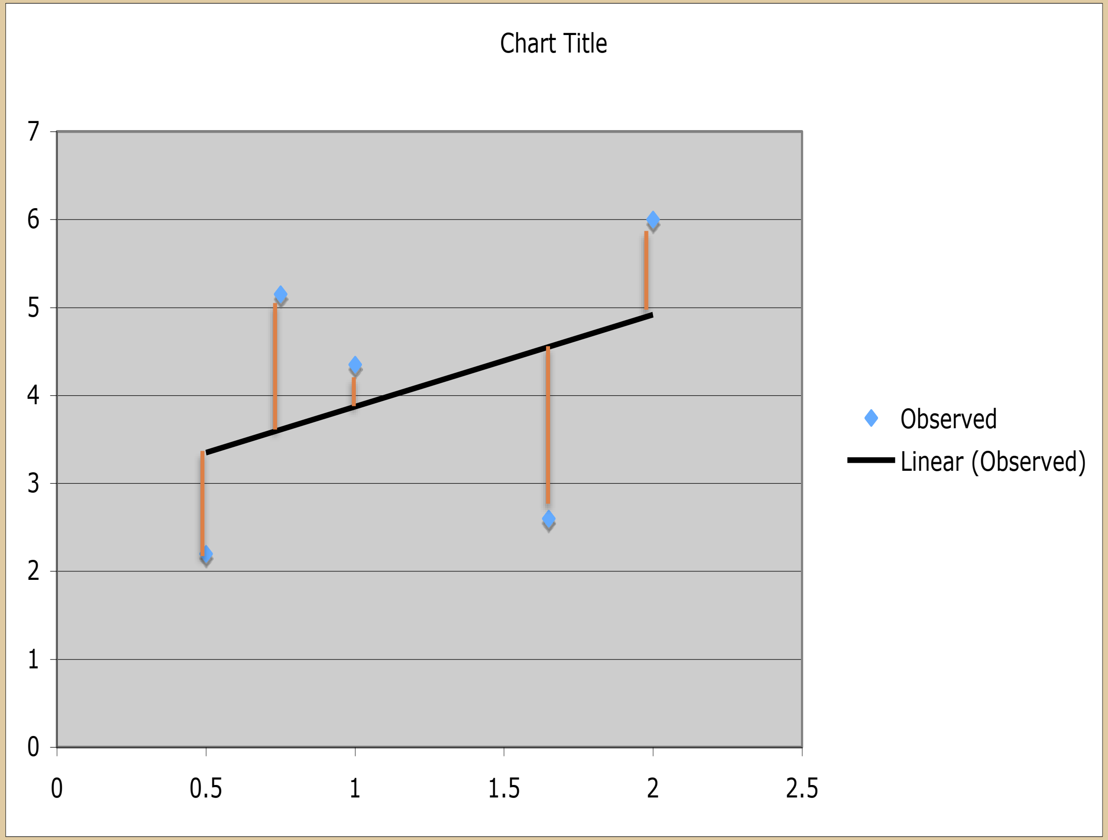
- The regression line divides the data with values above it and below it. This is called the **spread** of the data around the regression line. Each data point falls at a distance $\hat{y}_i - y_i$ from the regression line.
- The aggregate distance of all points from the mean can be defined as:

SSE = Sum of the squares of residuals around the regression line.

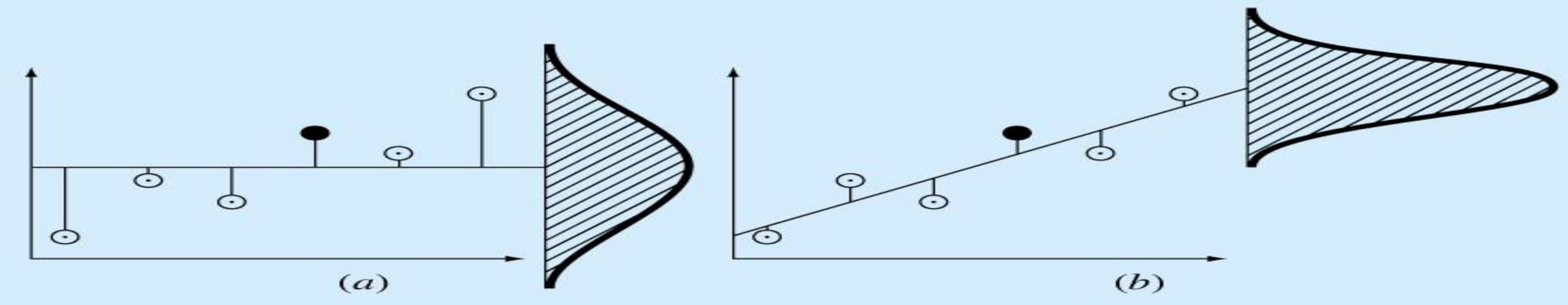
$$\sum_{i=0}^{n} (\hat{y}_i - y_i)^2$$

We can now compute an average distance from the regression line. "standard deviation" for the regression line

$$\sigma(STANDARD\ ERROR\ of\ estimate\ ) = \sqrt{\frac{SSE}{N-2}}$$



Simple Linear Regression: Calculate prediction error



(a) around the mean *(b)around the best-fit line*

Standard Deviation

$$\sigma = \sqrt{\frac{S_t}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

$$\sigma(STANDARD\ ERROR) = \sqrt{\frac{SSE}{N-2}}$$

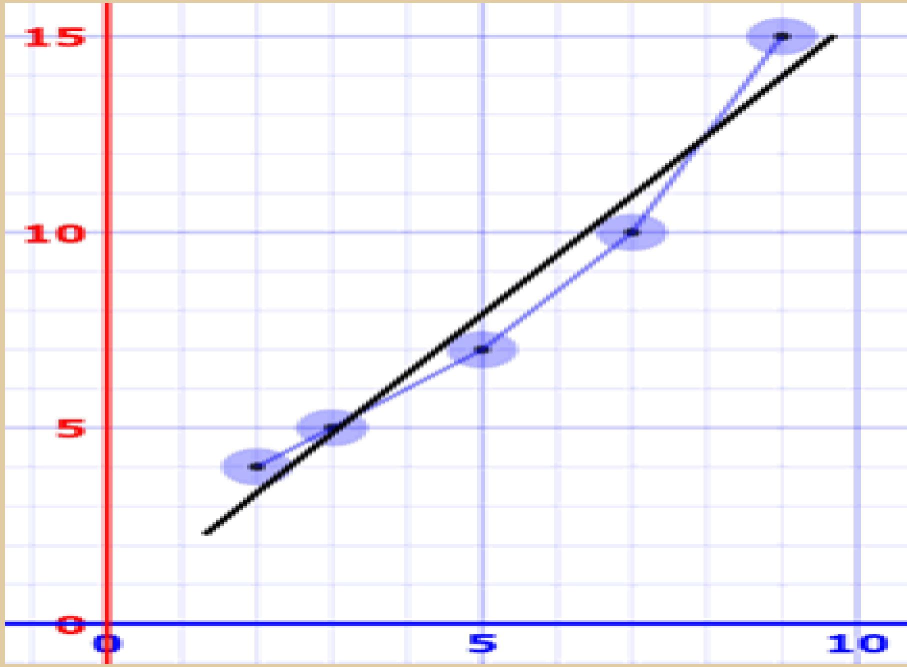
SSE = Sum of the squares of residuals around the regression line.

Std. Error - This determines the level of variability associated with the estimates.

Smaller the standard error of an estimate is, more accurate will be the predictions.

Simple Linear Regression: Example1 cont.
Calculate the Standard Error

x	y	y = 1.518x + 0.305	Error= (y - y-predicted)	square errors ***SSE***
2	4	3.34	-0.66	0.4356
3	5	4.86	-0.14	0.0196
5	7	7.89	0.89	0.7921
7	10	10.93	0.93	0.8649
9	15	13.97	-1.03	1.0609



Sum of square errors (SSE) around the regression line = 3.1731

$\sigma(STANDARD\ ERROR)$ = 1.028445

$$\sigma(STANDARD\ ERROR) = \sqrt{\frac{SSE}{N-2}}$$

Properties of the Regression Line

- **The line minimizes the sum of squared differences between observed values (the *y* values) and predicted values (the ŷ values computed from the regression equation).**
- The regression line passes through the mean of the *X* values and through the mean of the *Y* values .
- The regression constant (a_0) is equal to the **y intercept** of the regression line.
- The regression coefficient (a_1) is the average change in the dependent variable (*Y*) for a 1-unit change in the independent variable (*X*). It is the **slope** of the regression line.
- The least squares regression line is the only straight line that has all of these properties.

Simple Linear Regression: Example2

Nine volunteers are tested before and after a training programme. Find the line of best fit for the posterior (after training) scores as a function of the prior (before training) scores.

Volunteer:	1	2	3	4	5	6	7	8	9
After training:	75	66	69	45	54	85	58	91	62
Before training:	72	65	64	39	51	85	52	92	58

Let Y = score after training and X = score before training.

Simple Linear Regression: Example2 cont.

i	x_i	y_i	x_i^2	$x_i \cdot y_i$
1	72	75	5184	5400
2	65	66	4225	4290
3	64	69	4096	4416
4	39	45	1521	1755
5	51	54	2601	2754
6	85	85	7225	7225
7	52	58	2704	3016
8	92	91	8464	8372
9	58	62	3364	3596
Sum:	578	605	39384	40824

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \overline{y} - a_1 \overline{x}$$

a_1=17726/20372= 0.870116

a_0=11.34145

$$\hat{y} = a_1\, x + a_0$$

$$\hat{y} = 0.87\, x + 11.34$$

x_i	y_i	$\hat{y}_i$	E_i	E_i^2
72	75	73.98979	1.0102	1.0205
65	66	67.89898	−1.8990	3.6061
64	69	67.02886	1.9711	3.8854
39	45	45.27597	−0.2760	0.0762
51	54	55.71736	−1.7174	2.9493
85	85	85.30130	−0.3013	0.0908
52	58	56.58747	1.4125	1.9952
92	91	91.39211	−0.3921	0.1537
58	62	61.80817	0.1918	0.0368

Sum of square errors (SSE) = 13.8141

THE STANDARD ERROE =**1.4047**

Linear Regression Practice Example 3

Use least-squares regression to fit a straight line to

x	1	3	5	7	10	12	13	16	18	20
y	4	5	6	5	8	7	6	9	12	11

$$n = 10$$

$$\sum x_i = 105$$

$$\sum y_i = 73$$

$$\bar{x} = 10.5$$

$$\bar{y} = 7.3$$

$$\sum x_i^2 = 1477$$

$$\sum x_i y_i = 906$$

$$a_1 = \frac{n\sum(x_i y_i) - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{10*906 - 105*73}{10*1477 - 105^2} = 0.3725$$

$$a_0 = 7.3 - 0.3725*10.5 = 3.3888$$

YOUR TURN

Problem Statement

Last year, five randomly selected students took a math aptitude test before they began their statistics course. The Statistics Department has three questions.

•What linear regression equation best predicts statistics performance, based on math aptitude scores?

•If a student made an 80 on the aptitude test, what grade would we expect her to make in statistics?

•How well does the regression equation fit the data?

Student	x_i	y_i
1	95	85
2	85	95
3	80	70
4	70	65
5	60	70

$a_1 = 0.644$

$a_0 = 26.768$

$$\hat{y} = 26.768 + 0.644\,x\ .$$

$\hat{Y}(\mathbf{80}) = = 26.768 + 0.644 * \mathbf{80}$

$\hat{y} = 26.768 + 51.52 = 78.288$

SSE ??? ,

Multiple Regression Model

Estimated Multiple Regression Equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

A simple random sample is used to compute sample statistics b_0, b_1, b_2, . . . ,

Least Squares Method

The formulas for the regression coefficients b_0, b_1, b_2, . . . b_p involve the use of matrix algebra.

We will rely on computer software packages to perform the calculations.

Least Squares Criterion

$$\boxed{\min \sum (y_i - \hat{y}_i)^2}$$

Multiple Regression Model: Example

A software firm collected data for a sample of 20 computer programmers. A suggestion was made that regression analysis could be used to determine if the salary was related to the years of experience and the score on the firm's programmer aptitude test.

Exper.	Score	Salary	Exper.	Score	Salary
4	78	24.0	9	88	38.0
7	100	43.0	2	73	26.6
1	86	23.7	10	75	36.2
5	82	34.3	5	81	31.6
8	86	35.8	6	74	29.0
10	84	38.0	8	87	34.0
0	75	22.2	4	79	30.1
1	80	23.1	6	94	33.9
6	83	30.0	3	70	28.2
6	91	33.0	3	89	30.0

Estimated Regression Equation

- If we feed the data to SPSS package the output will be the following:

coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.143	6.147		.511	.616
	Exper	1.404	.198	.741	7.083	.000
	Score	.251	.077	.340	3.252	.005

a Dependent Variable: salary

SALARY = 3.143 + 1.404(EXPER) + 0.251(SCORE)

Interpreting the Coefficients

In multiple regression analysis, we interpret each regression coefficient as follows:

b_i represents an estimate of the change in y corresponding to a 1-unit increase in x_i when all other independent variables are **held constant**.

$\boxed{b_1 = 1.404}$

Salary is expected to increase by \$1,404 for each additional year of experience (when the variable *score on programmer attitude test* is held constant).

$\boxed{b_2 = 0.251}$

Salary is expected to increase by \$251 for each additional point scored on the programmer aptitude test (when the variable *years of experience* is held constant).

Hybrid

Logistic Regression

Logistic Regression

- Regression used to fit a curve to data in which the dependent variable is **binary**
- Typical application:
  - We might want to predict response to treatment, where we might code survivors as **1** and those who don't survive as 0, cancer or not cancer, approved or not approved, so on ….
  - Logistic regression can be used for classifying a new observation into one of the classes, based on the values of its predictor variables (called "classification").

Probability Review

The **probability** that an event will occur is the fraction of times you expect to see that event in many trials. If the probability of an event occurring is Y, then the probability of the event not occurring is 1-Y

The **odds** are defined as the probability that the event will occur divided by the probability that the event will not occur.
Unlike **probability,** the odds are not constrained to lie between 0 and 1 but can take any value from zero to infinity.
If the probability of Success is P, then the odds of that event is: $\boxed{odds = \frac{P}{1-P}}$

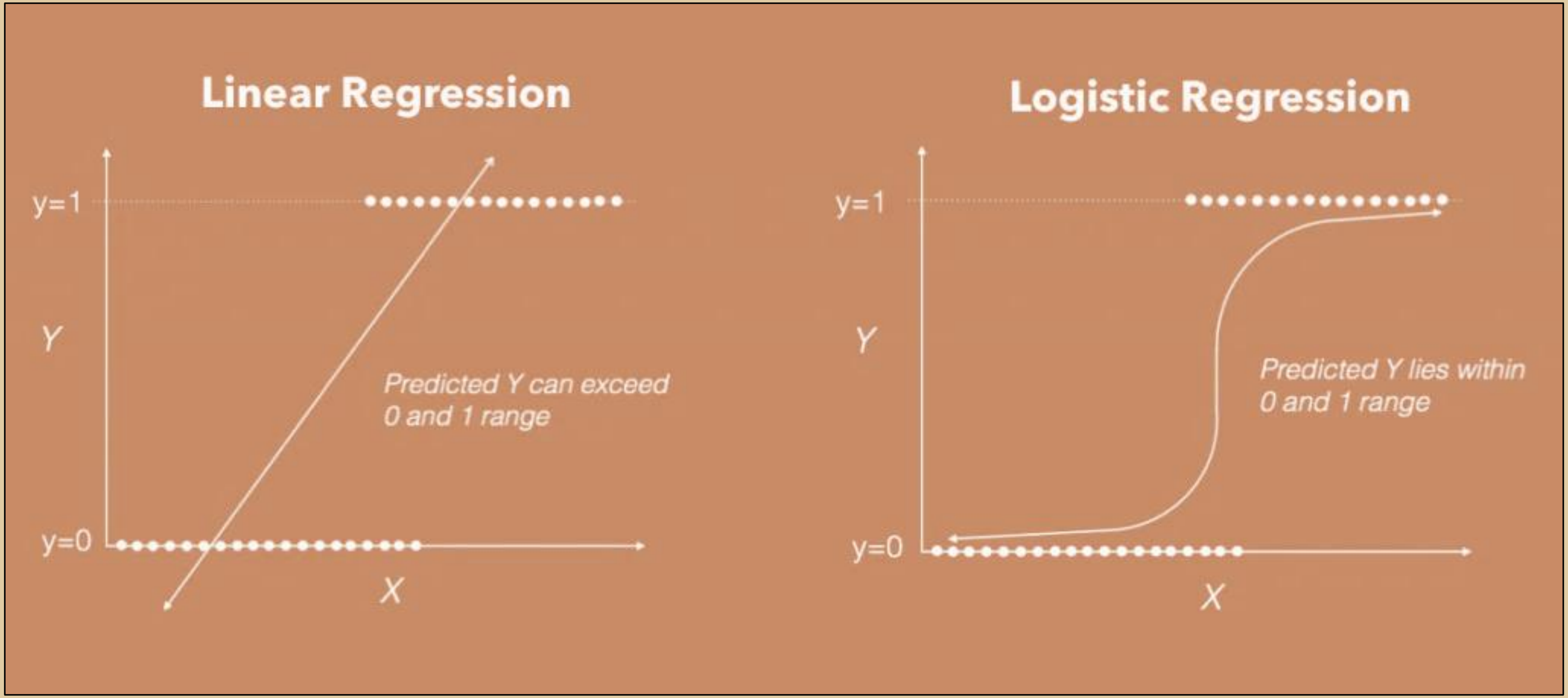
$$\text{Odds}(\text{event}) = \frac{\text{Probability}(\text{event})}{\text{1-Probability}(\text{event})}$$

$$\text{Probability}(\text{event}) = \frac{\text{Odds}(\text{event})}{1+\text{Odds}(\text{event})}$$

Logistic Regression

- In logistic regression we take two steps:
  - the first step yields estimates of the probabilities of belonging to each class.
    - In the binary case we get an estimate of $P(Y = 1)= P(Y = 0)$.
- In the next step we use
  - a cutoff value on these probabilities in order to classify each case to one of the classes.
    - In a binary case, a cutoff of 0.5 means that cases with an estimated probability of $P(Y = 1) > 0.5$ are classified as belonging to class 1, whereas cases with $P(Y = 1) < 0.5$ are classified as belonging to class 0.

A Sigmoid Function



The Logistic Regression Model

The "logit" model solves these problems:

$$\text{logit(P)} = \ln\left(\frac{P}{(1-P)}\right) = \alpha + \beta\text{X} + Error$$

- p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"
- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is:

$$P = \frac{1}{(1+e^{-\alpha - \beta \text{X}})}$$

Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the parameters of a model.

- MLE means choosing the parameter that maximize the probability of *Y* values in the sample with the given *X* values. the likelihood is the same as the probability. It calculates the ***joint probability*** of all the values of the dependent variable

- MLE involves finding the coefficients (α, β)

Logistic Regression :Example

- A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and the Rank the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

ADMIT	GRE	GPA	RANK
0	380	3.61	3
1	660	3.67	3
1	800	4	1
1	640	3.19	4
0	520	2.93	4
1	760	3	2
1	560	2.98	1
0	400	3.08	2
1	540	3.39	3
0	700	3.92	2
0	800	4	4
0	440	3.22	1
1	760	4	1
0	700	3.08	2
1	700	4	1
0	480	3.44	3
0	780	3.87	4
0	360	2.56	3
0	800	3.75	2

Confusion Matrix

Classification Table[a]

Observed			Predicted admit 0	Predicted admit 1	Percentage Correct
Step 1	admit	0	254	19	93.0
		1	97	30	23.6
	Overall Percentage				71.0

a. The cut value is .500

Application of Linear Regression

- Predictive Analytics
- Operation Efficiency: In a call center, we can analyze the relationship between wait times of callers and number of complaints
- Supporting Decisions: regression analysis leads the way to smarter and more accurate decisions.
- Correcting Errors: For example, a retail store manager may believe that extending shopping hours will greatly increase sales. RA, however, may indicate that the increase in revenue might not be sufficient to support the rise in operating expenses due to longer working hours (such as additional employee labor charges).
- https://towardsdatascience.com/linear-regression-in-real-life-4a78d7159f16

Lab6: Linear Regression

Due
Week 6

Discussion

Q&A

Summary

Simple Linear Regression model

Minimum sum of square error

Standard error

Multiple Linear regression

Practical examples

NUMERICAL COMPUTING
CST8233
non-Linear Regression

By

Hala Own, Ph.D.

Learning Objectives

- Use transformations involving powers and roots to achieve linearity for a relationship between two variables
- Make predictions from a least-squares regression line involving transformed data
- Use transformations involving logarithms to achieve linearity for a relationship between two variables
- Determine which of several transformations does a better job of producing a linear relationship

Nonlinear Regression

Some popular nonlinear regression models:

1. Exponential model: $(y = ae^{bx})$
2. Power model: $(y = ax^{b})$
3. Saturation growth model: $\left(y = \dfrac{ax}{b + x}\right)$
4. Polynomial model: $(y = a_0 + a_1 x + \ldots + a_m x^{m})$

The best transformation method (exponential model, quadratic model, reciprocal model, etc.) will depend on nature of the original data. The only way to determine which method is best is to try each and compare the result (i.e., SSE). The best method will yield the (MINIMUM SSE).

Non-linear Simple Regression

- If the relation between the dependent variable and the independent variable is nonlinear

  Option1: Use another technique

  Option 2: the data can be transformed so that linear regression can still be used.

- The latter technique is frequently used to fit the set of data.

Transforming to Achieve Linearity

Applying a function such as the logarithm or square root to a quantitative variable is called **transforming** the data.

General Procedure:

1. Linearize the model around the current parameter values. This results in a linearized objective-function surface.

2. Using the normal equations, obtain the corresponding prediction for the response variable *y*, we have to **"undo"** the transformation to return to the original units of measurement.

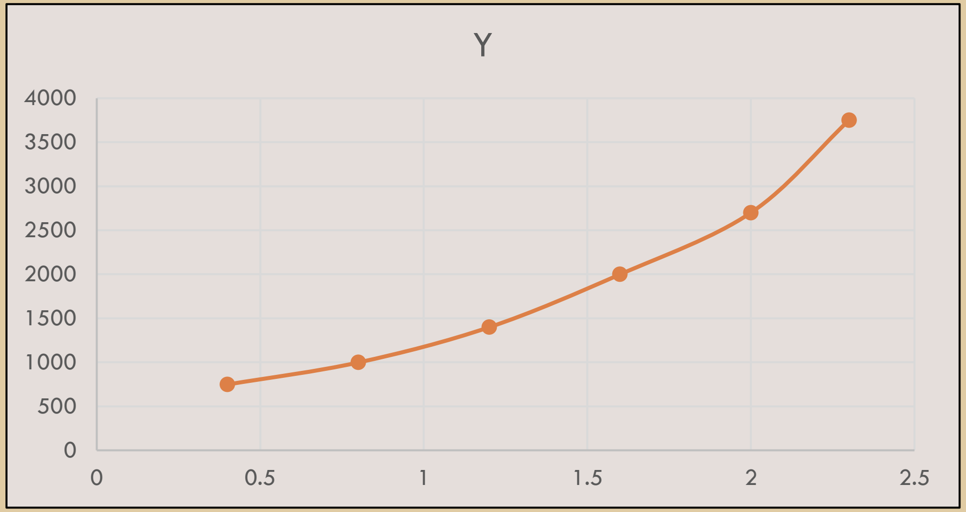
How to choose your nonlinear function

Some considerations when choosing a nonlinear function for curve fitting are as follows:

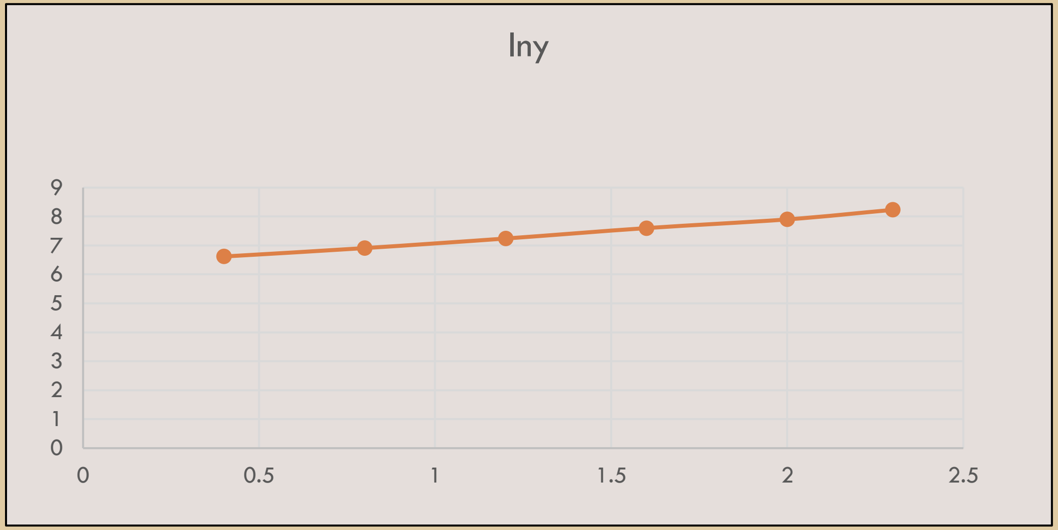
- Exponential functions cannot pass through the origin.
- Exponential functions can only fit data with all positive ys, or all negative ys.
- Logarithmic functions cannot include $x = 0$ or negative values of x.
- For power function $y = 0$ when $x = 0$.

Exponential Model

$$\left(y = ae^{bx}\right)$$



Linearize



Exponential Model:

Exponential model given as, $(y = ae^{bx})$

$$\underbrace{\ln y}_{K} = \underbrace{\ln a}_{a_0} + \underbrace{b}_{a_1}x$$

$$K = a_0 + a_1 x$$

This is a linear relationship between K and x

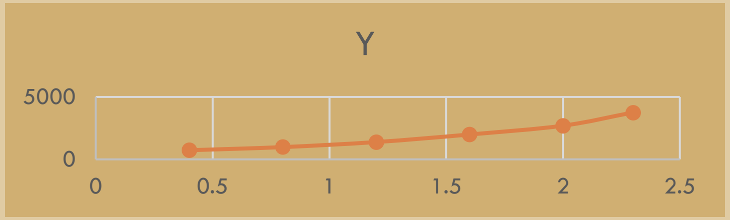
$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

Important step : the original constants of the model are found as $b = a_1$ but we need to add extra step here since $a_0 = \ln(a)$, then $\boldsymbol{a = e^{a_0}}$

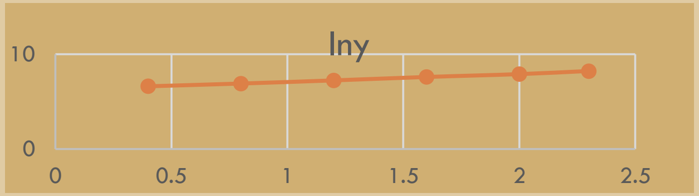
$$\sigma(STANDARD\ ERROR) = \sqrt{\frac{SSE}{N-2}}$$

Exponential Simple Regression Model example



$(y = ae^{bx})$

Linearize



$\ln(y) = \ln a_0 + a_1 x$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

	X	Y
	0.4	750
	0.8	1000
	1.2	1400
	1.6	2000
	2	2700
	2.3	3750
sum	8.3	11600
mean	1.383333	1933.333

Ln(y)	x^2	x*ln(y)	predicted $\hat{y} = 518e^{0.841406x}$	y-$\hat{y}$	(y-$\hat{y}$ $)^2$
6.62	0.16	2.648	725.1476159	24.85238411	617.6409957
6.91	0.64	5.528	1015.13333	15.1333298-	229.0176709
7.24	1.44	8.688	1421.084004	21.08400371-	444.5352123
7.6	2.56	12.16	1989.373894	10.62610565	112.9141212
7.9	4	15.8	2784.922271	84.92227145-	7211.792188
8.23	5.29	18.929	3584.145041	165.8549592	27507.86748
44.5	14.09	63.753			36123.76767
7.416667					

$a_1 = 0.841406$, $a_0 = 6.252722$

Switch back to the original equation since a_0=ln(a) **then a= e^{a0} = $e^{6.252722}$ =**518, b =a1, therefore the exponential equation is

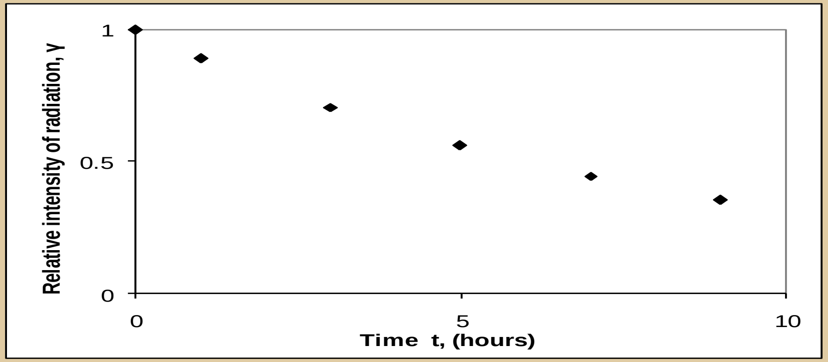
$$\hat{y} = 518e^{0.841406x}$$

Check the equation for **x = 1.2**, $\hat{y} = 518e^{0.841406*1.2}$=1421 and SSE= 36123.76767 , with standard error= 95.03127

Example 2:

Many patients get concerned when a test involves injection of a radioactive material. For example, for scanning a gallbladder. Below is given the relative intensity of radiation as a function of time.

t(hrs)	0	1	3	5	7	9
R	1.000	0.891	0.708	0.562	0.447	0.355



Find:

a) The value of the regression constants a and b

b) Radiation intensity after 24 hours

The relative intensity is related to time by the equation

R= ae^{bt}

Example 2: cont.

$$(y = ae^{bx})$$

$$a_0 = \bar{y} - a_1\bar{x}$$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

1- Calculating a_0, a_1

$$a_1 = \frac{6(-18.990) - (25)(-2.8778)}{6(165.00) - (25)^2} = -0.11505$$

$$a_0 = \frac{-2.8778}{6} - (-0.11505)\frac{25}{6} = -2.6150 \times 10^{-4}$$

i	t_i	R_i	$\ln R_i$	$t_i * \ln R_i$	t_i^2
1	0	1	0.00000	0.0000	0.0000
2	1	0.891	−0.11541	−0.11541	1.0000
3	3	0.708	−0.34531	−1.0359	9.0000
4	5	0.562	−0.57625	−2.8813	25.000
5	7	0.447	−0.80520	−5.6364	49.000
6	9	0.355	−1.0356	−9.3207	81.000
	25.000		−2.8778	−18.990	165.00

2- undo step: Since

$$a_0 = \ln(a)$$

$$a = e^{a_0} = e^{-2.6150\times10^{-4}} = 0.99974$$

b $= a_1 = -0.11505$ Resulting model is

$$\boxed{R = 0.99974 \times e^{-0.11505t}}$$

b) The relative intensity of radiation after 24 hours is then $\mathbf{R = 0.99974 e^{-0.11505(24)}} = 0.063200$

Power Model:

Exponential model given as, $(y = ax^b)$

$$\underbrace{\log(y)}_{K} = \underbrace{\log a}_{a_0} + \underbrace{b}_{a_1} \underbrace{\log(x)}_{z}$$

$$K = a_0 + a_1 z$$

$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{y} - a_1 \bar{x}$$

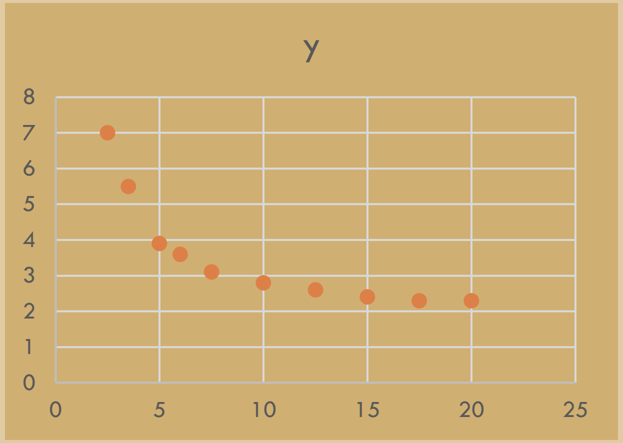
This is a linear relationship between K and z

Important step : the original constants of the model are found as since $a_0=\log(a)$ then **$a = 10^{a_0}$ and $b = a_1$**

$$a_0 = \bar{y} - a_1 \bar{x}$$

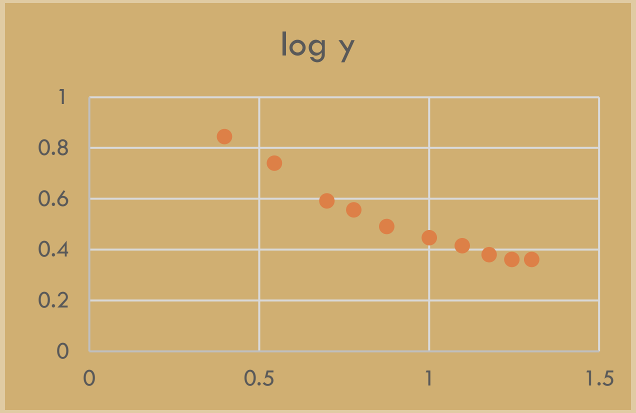
$$a_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Power Simple Regression Model example



$y = ax^b$

Linearize



log (y)= log a + b log(x)

	x	y
	2.5	7
	3.5	5.5
	5	3.9
	6	3.6
	7.5	3.1
	10	2.8
	12.5	2.6
	15	2.4
	17.5	2.3
	20	2.3
sum	99.5	35.5
mean	9.95	3.55

log x	log y	(Logx)²	(logx)*(logy)
0.39794	0.845098	0.158356	0.336298
0.544068	0.740363	0.29601	0.402808
0.69897	0.591065	0.488559	0.413136
0.7781513	0.556303	0.605519	0.432887
0.8750613	0.491362	0.765732	0.429972
1.00000000	0.447158	1	0.447158
1.09691	0.414973	1.203212	0.455188
1.1760913	0.380211	1.383191	0.447163
1.243038	0.361728	1.545144	0.449641
1.30103	0.361728	1.692679	0.470619
9.1112599	5.189988	9.138402	4.284871
0.911126	0.518999		

a_1=	-0.53037
a_0=	1.002229

Switch back to the original equation

a= 10^{a0} =10.05

, b =a_1= -0.53037 ,

the Power equation is

$$\hat{y} = 10.05x^{-0.53037}$$

Power Simple Regression Model example, cont.

Check the equation for **x = 5**, then $\hat{y} = 10.05 * 5^{-0.53037}$=4.28

Or x= 15, then $\hat{y} = 10.05 * 15^{-0.53037}$=2.39 .

Standard Error = 0.399498

$\sigma(STANDARD\ ERROR) = \sqrt{\frac{SSE}{N-2}}$

$\hat{y}=10.05x^{-0.53037}$	y-$\hat{y}$	(y-$\hat{y}$)^2
6.12743139	0.87256861	0.761375979
5.109393108	0.390606892	0.152573744
4.214269057	0.314269057-	0.09876504
3.819129515	0.219129515-	0.048017745
3.385579213	0.285579213-	0.081555487
2.898451661	0.098451661-	0.009692729
2.569417364	0.030582636	0.000935298
2.328502893	0.071497107	0.005111836
2.142523112	0.157476888	0.02479897
1.993470734	0.306529266	0.093960191
4.280094329	31.21990567	1.27678702

Power Simple Regression Model: Example 2

Determine the predicted equation for the following data, where the power law fits this data.

x	y
1	0.5
2	1.7
3	3.4
4	5.7
5	8.4

Transformed data

	x	y	log x	log y	$(\log x)^2$	Logx * logy
	1	0.5	0	0.3-	0	0
	2	1.7	0.301	0.226	0.090601	0.068026
	3	3.4	0.477	0.531	0.227529	0.253287
	4	5.7	0.602	0.756	0.362404	0.455112
	5	8.4	0.699	0.924	0.488601	0.645876
sum	15	19.7	2.079	2.136	1.169135	1.422301
mean	3	3.94	0.4158	0.4272		

$a_1 =$ 1.753 and $a_0 =$ -0.300 ,

Linear Regression yields the result: **log y = 1.75*log x – 0.300**

Switch back to the original equation a= 10^{a0} =0.511 and

b $= a_1 =$ 1.753

$\hat{y}$ **= 0.5 $x^{1.75}$**

Exercise

- Determine an equation to predict metabolism rate as a function of mass based on the following data. Use it to predict the metabolism rate of a 200-kg tiger

Mass (kg)	Metabolism (kCal/day)	log Mass	log Met
300	5600	2.477121	3.748188
70	1700	1.845098	3.230449
60	1100	1.778151	3.041393
2	100	0.30103	2
0.3	30	-0.52288	1.477121

$b = 0.7497$ and $a = 10^{1.818} = 65.768$, and the nonlinear regression equation is :

Metabolism = $65.768\text{Mass}^{0.7497}$

Lab7: Linear Regression

Due

Week 7

Discussion

Q&A