

Advanced Statistics

DS2003 (BDS-4A)

Lecture 01

Instructor: Dr. Syed Mohammad Irteza

Assistant Professor, Department of Computer Science, FAST

15 February, 2022

Administrative Information

- Office: To be finalized soon, iA
- Email: smirteza@gmail.com
- Office Hours: (most probably Tuesday/Thursday)
- Website: To be announced soon, iA
- Class Schedule:
 - Tuesdays & Thursdays 1300 – 1430

My Academic & Research Background

- BS, MS and PhD in Computer Science from LUMS
 - 2002, 2005, 2018
- PhD Thesis: *Resilient Network Load Balancing for Datacenters*
 - Advisor – Dr. Ihsan Ayyub Qazi
- Google Scholar Page:
 - <https://scholar.google.com/citations?hl=en&user=wHazKsgAAAAJ>
- Main Research Interests:
 - Networking for Datacenters: network layer and transport layer protocols
 - Software Defined Networking

Teaching Experience

- **UMT** – October 2019 to February 2021
 - Grad-level courses → Advanced Networks, Advanced Computer Architecture
 - Programming Fundamentals, Artificial Intelligence, Computer Networks
- **GIFT University**, Gujranwala – April 2019 to August 2019
 - Software Requirement Engineering
- **LUMS** – September 2014 to May 2017
 - Co-instructor/grader for Intro to Programming (C++) and Machine Learning
- **Air University**, Islamabad – January 2010 to August 2012
 - Database Systems, Advanced Database Systems, Digital Logic Design, Computer Architecture, Distributed Systems, Number Theory
- **FAST**, Islamabad – Summer 2008
 - Programming for Engineers-II (C++)

Software Industry Experience

- ***Prosol Technologies***, Islamabad (partially now [Ciklum](#))
 - Software Consultant (Sep 2008 – Nov 2008)
 - Senior Software Engineer (Feb 2006 – Feb 2008)
 - Software Engineer (Apr 2004 – June 2004)
- ***Diyatech Pakistan*** ([Alachisoft](#)), Islamabad
 - Software Developer (Apr 2003 – Apr 2004)
- [InvestCorp](#), Bahrain
 - Junior Software Developer (June – July 2000)

Interest in Probability and Statistics?

- A levels → Pure Math and Statistics
- Undergraduate
 - Probability (Dr. Arif Zaman)
 - Statistics ([Dr. Asad Zaman](#))
 - Econometrics (Ms. Sadia Shaikh)
 - Intro to Mathematical Finance ([Dr. Kazim Khan](#))
 - Intro to Game Theory (Dr. Faisal Bari)
- MS → Stochastic Systems (Dr. Shahab Baqai)
- PhD → Applied Probability (Dr. Ihsan Ayyub Qazi)

Classroom Etiquette

- Please come on time
- Talking among each other is not acceptable, *while I am teaching*
- Leaving the class to attend a phone call *is not appreciated*
- Quizzes will in general be *unannounced*
 - They can be held at the start or end of class
- Cases of *plagiarism* (copying of other people's work) will lead to marks and/or grade *reductions*

Grading Policy – Tentative (*may be changed*)

- Quizzes & Assignments → 20% + 15%
 - If we have 7 or more quizzes, we will choose the best 5 or 6
 - All assignments will count to your grade
- Class Participation → 5%
- Midterm I and Midterm II → 20%
- Final Exam → 40%
 - Comprehensive exam (all course contents included)

Textbook

- To be finalized!

Course Objectives

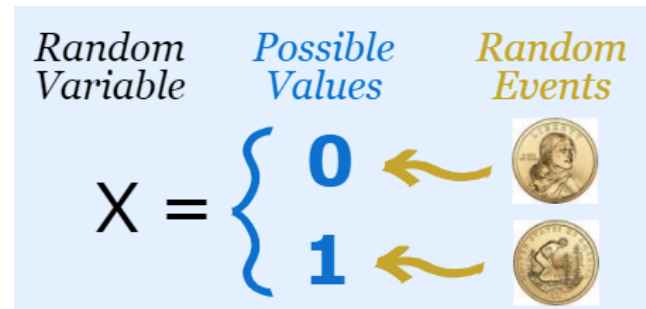
- This course aims at providing a deeper study of *advanced statistical concepts relevant to data science*.

Some Basics of Probability & Statistics

Random Variables

- A Random Variable is a set of *possible values* from a random experiment.
 - Example → Tossing a coin: we could get Heads or Tails.
 - Let's give them the values **Heads=0** and **Tails=1** and we have a Random Variable "X":

- Basically, $X = \{0, 1\}$



Note: We could choose Heads=100 and Tails=150 or other values if we want! It is our choice.

Random Variables

- So far:
 - We have an *experiment* (such as tossing a coin)
 - We give *values* to each event
 - The *set of values* is a *Random Variable*
- How is a Random Variable different to an algebra variable such as:
 - $x + 2 = 6$
- A Random Variable has a whole *set of values* and it could take any of those *values*, randomly
 - For example, $X = \{0, 1, 2, 3\}$
X could be 0, 1, 2, or 3, randomly, and they might each have a *different probability*

Random Variables

- We often use capital letters like ***X*** or ***Y*** for ***Random Variables***, to differentiate from the ***Algebra type of variables***
- Sample Space:
 - A Random Variable's set of values is the ***Sample Space***
- For example, if we throw a die once
 - $X = \text{"the score shown on the top face"}$
 - Thus the ***sample space*** is $\rightarrow \{1, 2, 3, 4, 5, 6\}$
 - In this case they are all equally likely, so the probability of any one is $1/6$

RV Example: How many heads in 3 coin tosses?

- X = "The number of Heads" is the Random Variable
- What is the sample space?



RV Example: How many heads in 3 coin tosses?

- X = "The number of Heads" is the Random Variable
- What is the sample space?
 - $\{0, 1, 2, 3\}$
- Are each of these *outcomes* equally *likely*?



RV Example: How many heads in 3 coin tosses?

- The *three coins* can land in *eight possible ways*:
 - Looking at the table we see just 1 case of Three Heads, but 3 cases of Two Heads, 3 cases of One Head, and 1 case of Zero Heads. So:
- $P(X = 3) = 1/8$
 - $P(X = 2) = 3/8$
 - $P(X = 1) = 3/8$
 - $P(X = 0) = 1/8$

		X = "Number of Heads"
HHH		3
HHT		2
HTH		2
HTT		1
THH		2
THT		1
TTH		1
TTT		0

RV Example: Sum of the scores of two dice

- The Random Variable is X = "The sum of the scores on the two dice"
- Can we make a table of all possible values?

	1 st Die						
2 nd Die		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

RV Example: Sum of the scores of two dice

- What is the probability of:

- $P(X = 7)$?
 - $6/36$
- $P(X = 11)$?
 - $2/36$

	1 st Die						
2 nd Die		1	2	3	4	5	6
	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

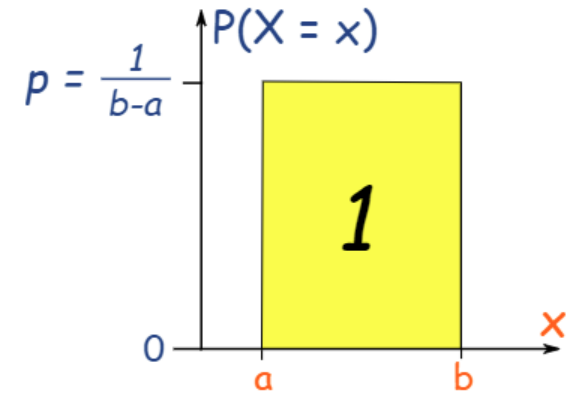
- What is the probability that the sum of the scores is 5, 6, 7 or 8?
 - $P(5 \leq X \leq 8) = P(X=5) + P(X=6) + P(X=7) + P(X=8)$
 $= (4+5+6+5)/36$
 $= 20/36 = 5/9$

Discrete and Continuous RVs

- Random Variables can be either *discrete* or *continuous*
 - The examples above are examples of *discrete data* (such as 1, 2, 3, 4, 5, and 6)
 - *Continuous data* can take any value within a range (such as a person's height)

The Uniform Distribution (Continuous RV)

- The Uniform Distribution (also called the Rectangular Distribution) is the simplest distribution.
- It has equal probability for all values of the Random variable between **a** and **b**:



The probability of any value between **a** and **b** is **p**

- We also know that **$p = 1/(b-a)$** , because the total of all probabilities must be 1, so:
 - The area of the rectangle is 1
 - $p \times (b-a) = 1$
 - $p = 1/(b-a)$

The Uniform Distribution (Continuous RV)

- We can write:

- $P(X = x) = 1/(b-a)$ for $a \leq x \leq b$
- $P(X = x) = 0$ otherwise

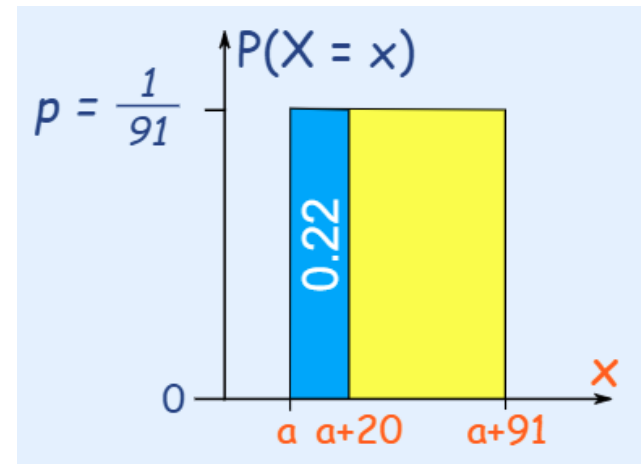


- Example: Old Faithful erupts every 91 minutes. You arrive there at random and wait for 20 minutes ... what is the probability you will see it erupt?
 - This is actually easy to calculate, 20 minutes out of 91 minutes is:
 - **$p = 20/91 = 0.22$** (to 2 decimals)
- But, let's use the Uniform Distribution for practice....

The Uniform Distribution (Continuous RV)

- To find the probability between a and $a+20$, find the blue area:

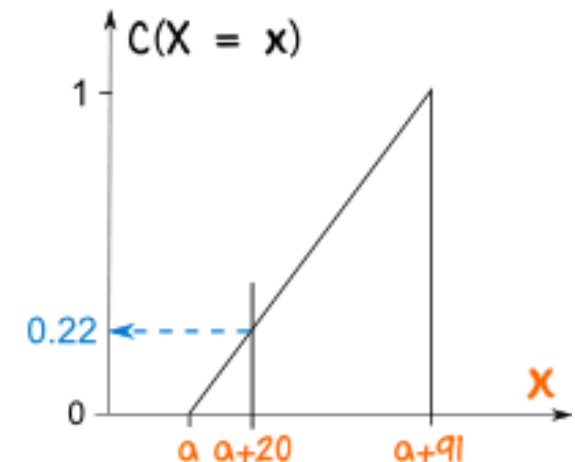
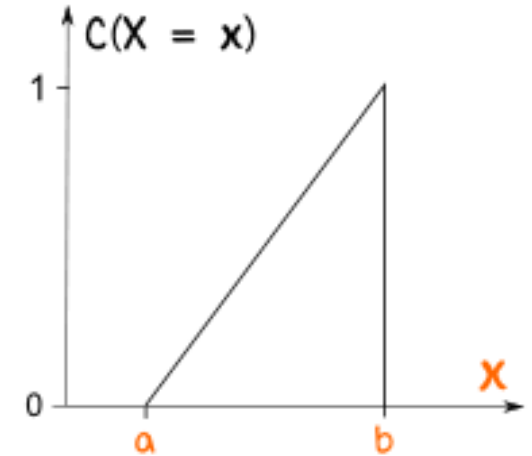
- Area = $(1/91) \times (a+20 - a)$
= $(1/91) \times 20$
= $20/91$
= **0.22** (to 2 decimals)



So there is a **0.22** probability you will see *Old Faithful erupt*.

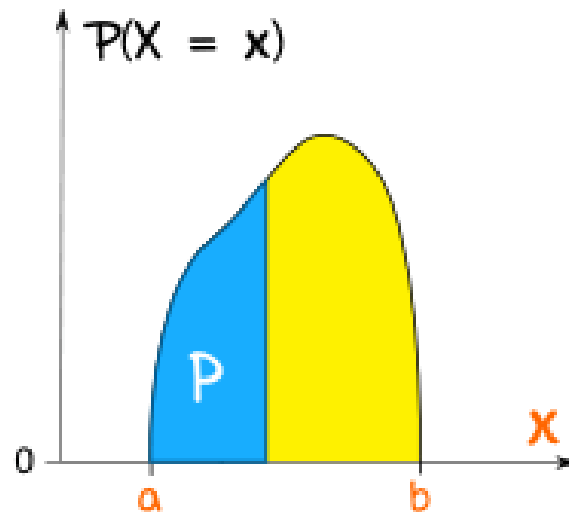
Cumulative Uniform Distribution

- We can have the Uniform Distribution as a **cumulative** (adding up as it goes along) distribution:
 - This type of thing is called a "**Cumulative distribution function**", often shortened to "CDF"
- Let's use the **CDF** of the previous Uniform Distribution to work out the probability:
 - At **$a+20$** the probability has accumulated to about **0.22**



Other Distributions

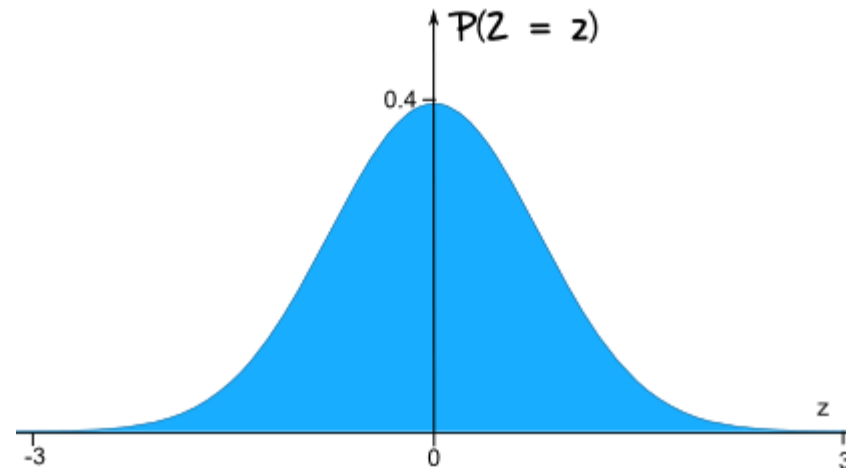
- Knowing how to use the *Uniform Distribution* helps when dealing with more *complicated distributions* like this one:



The general name for any of these is *probability density function* or *PDF*

The Normal Distribution

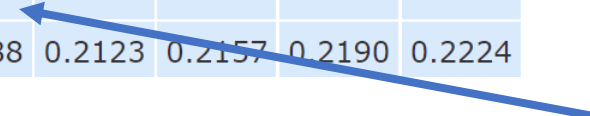
- The most important continuous distribution is the *standard normal distribution*
- It is so important the Random Variable has its own special letter **Z**.
- The graph for **Z** is a *symmetrical bell-shaped curve*:



The Normal Distribution

- Usually we want to find the probability of Z being between certain values.
 - Example: $P(0 < Z < 0.45)$
 - What is the probability that Z is between 0 and 0.45?
 - This is found by using the *Standard Normal Distribution Table*

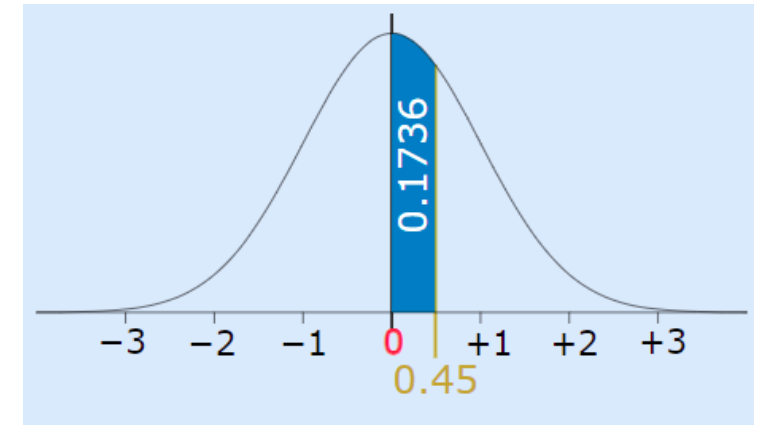
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224



- Start at the row for 0.4, and read along until 0.45: there is the value 0.1736
- Source: [Standard Normal Distribution Table \(mathsisfun.com\)](https://mathsisfun.com/standard-normal-distribution-table.html)

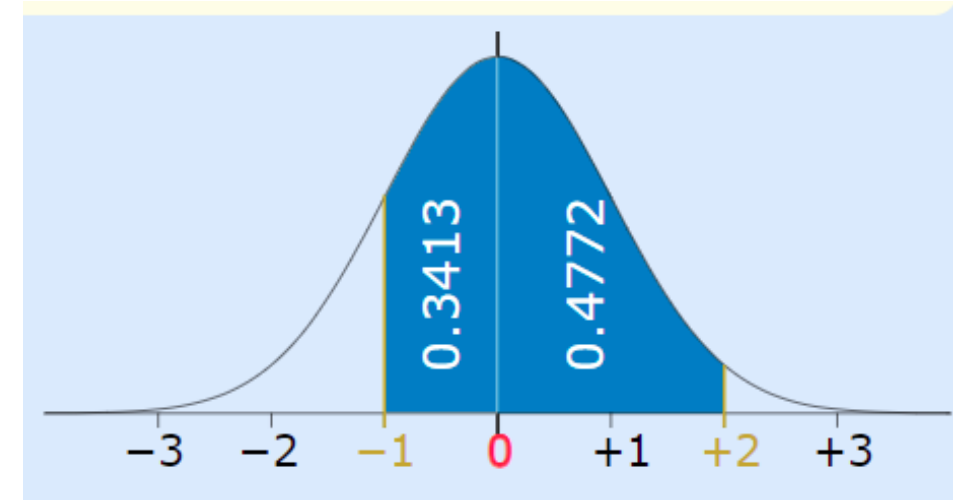
The Normal Distribution

- Example: Percent of Population Between **0** and **0.45**
 - And 0.1736 is **17.36%**
 - So 17.36% of the population are between 0 and 0.45 Standard Deviations from the Mean.
- Because the curve is *symmetrical*, the same table can be used for values going either direction, so a **negative 0.45** also has an area of **0.1736**



The Normal Distribution

- Example: Percent of Population Z Between -1 and 2
- From -1 to 0 is the same as from 0 to $+1$:
 - At the row for 1.0, first column 1.00, there is the value 0.3413
- From 0 to $+2$ is:
 - At the row for 2.0, first column 2.00, there is the value 0.4772
- Add the two to get the total between -1 and 2 :
- $0.3413 + 0.4772 = 0.8185$
 - And 0.8185 is 81.85%
 - So 81.85% of the population are between -1 and $+2$ Standard Deviations from the Mean.



IID

- In statistics, we commonly deal with *random samples*.
- A random sample can be thought of as a set of objects that are chosen randomly. Or, more formally, it's “a sequence of *independent, identically distributed (IID)* random variables”.
- In other words, the terms *random sample* and *IID* are basically one and the same.

IID

- In statistics, we usually say “random sample,” but in probability it’s more common to say “IID.”
- *Identically Distributed* means that there are no overall trends—the distribution doesn’t fluctuate and all items in the sample are taken from the same probability distribution.
- *Independent* means that the sample items are all independent events. In other words, they aren’t connected to each other in any way

Bernoulli Trials

- A single trial or experiment, for example, a *coin toss*
- Independent repeated trials of an experiment with exactly two possible outcomes are called Bernoulli trials.
- Call one of the outcomes *success* and the other outcome *failure*.
- Let p be the probability of *success* in a Bernoulli trial, and q be the probability of *failure*.
- Then the probability of *success* and the probability of *failure* sum to *one*, since these are complementary events:
 - "success" and "failure" are *mutually exclusive and exhaustive*.
- Thus one has the following relations:
 - $p + q = 1$

Sources

- <https://www.mathsisfun.com/data/random-variables.html>
- [https://en.wikipedia.org/wiki/Independent and identically distributed random variables](https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables)
- [https://en.wikipedia.org/wiki/Bernoulli trial](https://en.wikipedia.org/wiki/Bernoulli_trial)