Statistical and Mathematical Methods for Data Analysis

Dr. Faisal Bukhari

Associate Professor

Department of Data Science

Faculty of Computing and Information Technology

University of the Punjab

Textbooks

- ❑**Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer

- ❑**Elementary Statistics: Picturing the World, 6th Edition, Ron Larson and Betsy Farber**

- ❑**Elementary Statistics, 13th Edition, Mario F. Triola**

Reference books

- **Probability and Statistical Inference, Ninth Edition,** Robert V. Hogg, Elliot A. Tanis, Dale L. Zimmerman
- **Probability Demystified**, Allan G. Bluman
- **Schaum's Outline of Probability,** Second Edition, Seymour Lipschutz, Marc Lipson
- **Python for Probability, Statistics, and Machine Learning,** José Unpingco
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- **Think Stats: Probability and Statistics for Programmers,** Allen Downey

References

Readings for these lecture notes:

- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- **Probability Demystified**, Allan G. Bluman
- http://www.thefreedictionary.com/statistics
- **Discrete Mathematics and Its Application**, 7th Edition by Kenneth H. Rosen

These notes contain material from the above resources.

Distribution of points

Midterm = 30 points

Final term = 40 points

Sessional points = 30 points

I. Assignments = $2 \times 4 = 8$ points
II. Hands-on Python in class = $0.5 \times 6 = 3$ points
III. Quizzes = $2 \times 6 = 12$ points
IV. Journal/conference paper presentation = 5 points
V. Mini project (its report should be in an IEEE journal paper format) = 2 points

Or

The weightage of the project will be increased up to 10 points

What is Data Science?

Data Science is a fusion of multiples disciplines, including **statistics**, **computer science**, **information technology**, and **domain-specific fields**.

OR

Data Science is an umbrella that contain many other fields like **machine learning**, **data mining**, **big data**, **statistics**, **data visualization**, **data analytics**,...

Data Science

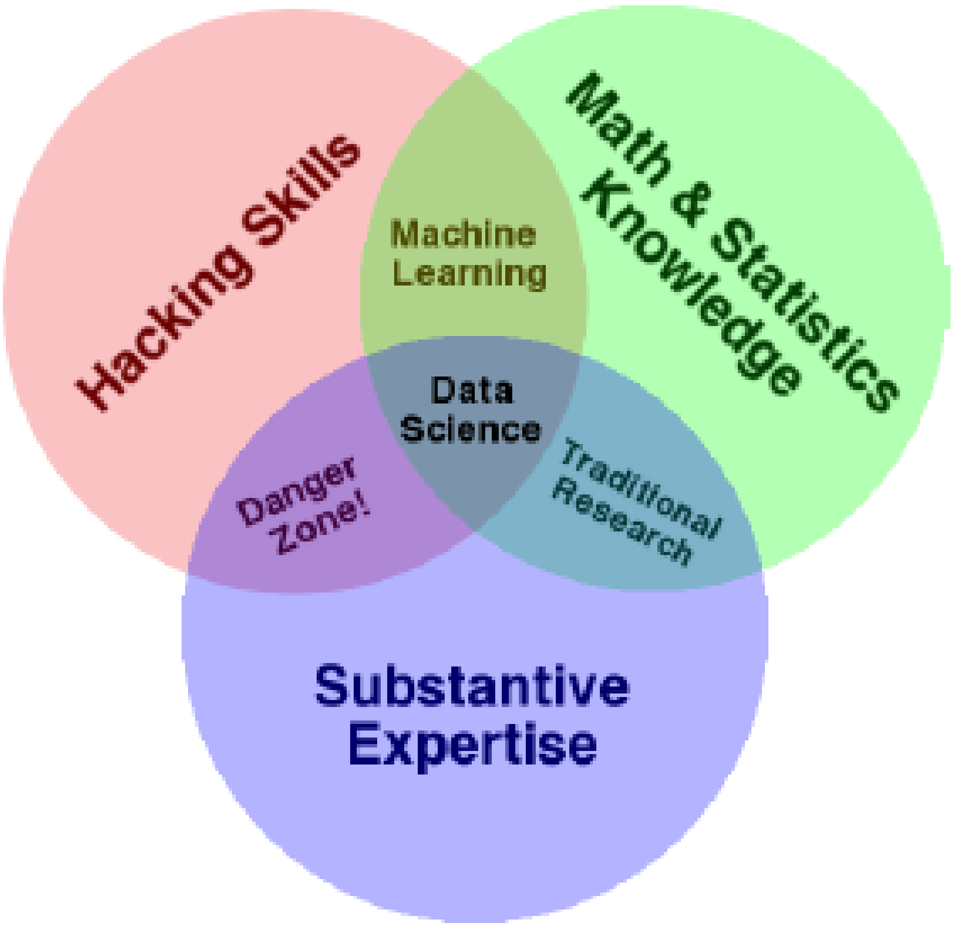


Figure 1-1. Drew Conway’s Venn diagram of data science

Set Theory

- **Set:** Any well **defined list** or **collection** of objects is called a ***set.***

  OR

  A set is an **unordered collection** of objects.

- **Element:** The objects comprising the **set** are called its ***elements*** or ***members.*** We write **p ∈ A** if p is an element in the set A

  OR

  The **objects** in a set are called the **elements**, or **members**, of the set. A set is said to contain elements.

- **Example** The set V of all vowels in the English alphabet can be written as V = {a , e, i, o, u} .

- **Example** The set O of odd positive integers less than 10 can be expressed by O = {1 , 3 , 5 , 7, 9}.

- **Example** {a , 2, Fred, New Jersey}

- **Note**: Although sets are usually used to group together **elements with common properties**, there is nothing that prevents a set from having **seemingly unrelated elements**.

Set builder notation

Another way to describe a set is to use set builder notation.

Example: The set 0 of all **odd positive integers** less than **10** can be written as

O = {x | x is an odd positive integer less than 10}

or

$O = \{x \in Z^+ \mid x \text{ is odd and } x < 10 \}$.

Note: The concept of a **datatype**, or type, in computer science is built upon the concept of a **set**.

Example: **boolean** is the name of the set {0, 1} together with operators on one or more elements of this set, such as **AND**, **OR**, and **NOT**

Set Theory

- **Subset:** If every element of ***A*** also belongs to a set ***B*****,** i.e. if $p \in A$ implies $p \in B$, then **A** is called a ***subset*** of ***B*** or is said to be ***contained*** in ***B*****;** this is denoted by $A \subset B$ or $B \supset A$

  OR

The set **A** is said to be a **subset** of **B if and only if every element of *A*** is also an **element of B**.

- **Note: Uppercase letters are usually used to denote sets**

❑Examples:

- ❑The set of all **odd positive integers less than 10** is a **subset** of the set of **all positive integers less than 10**.
- ❑The set of **rational numbers** is a **subset** of the set of **real numbers**.
- ❑The set of **all computer science majors** at your school is a **subset** of the **set of all students** at your school.
- ❑ The set of **all people in China** is a **subset** of the **set of all people in China** (that is, it is a subset of itself).

Theorem: For every set S,

(i) $\emptyset \subseteq S$ and

(ii) $S \subseteq S$

Proper subset

When we wish to emphasize that a **set A** is a **subset** of the **set B** but that **A ≠ B**, we write **A ⊂ B** and say that A is a **proper subset** of B. For **A ⊂ B** to be true, it must be the case that **A ⊆ B** and **there must exist** an **element x** of **B** that is **not an element** of **A**.

Note: Sets may have other sets as members.

A = {Ø, {a } , {b } , {a , b }}

Set Theory

- **Equal Set:** Two sets are ***equal*** if each is contained in the other; that is,

  A = B if and only if $\boldsymbol{A} \subset \boldsymbol{B}$ and $\boldsymbol{B} \subset \boldsymbol{A}$

- **Negation of Element, Subset and Equal Set:** The negations of $p \in \boldsymbol{A}$, $\boldsymbol{A} \subset \boldsymbol{B}$ and $\boldsymbol{A} = \boldsymbol{B}$ are written as

  $p \notin \boldsymbol{A}$, $\boldsymbol{A} \not\subset \boldsymbol{B}$ and $\boldsymbol{A} \neq \boldsymbol{B}$

Note: Lowercase letters are usually used to denote **elements** of sets.

Set Theory

We specify a particular set by either **listing its elements** or by **stating properties** which characterize the elements of the set. For example,

- **$A = \{1, 3, 5, 7, 9\}$**

means **A** is the set consisting of the numbers **1, 3, 5, 7** and **9;** and

- **$B = \{x : x \text{ is a prime number}, x < 15\}$**

means that ***B*** is the set of prime numbers less than **15**.

Set Theory

Example: The sets ***A*** and ***B*** in the previous slide can also be written as

A = {x : x is an odd number, z < 10}

and

B = {2, 3, 6, 7, 11, 13}

Example: We use the following special symbols:

N = the set of positive integers: 1, 2, 3, …

Z = the set of integers: … -3, -2, -1, 0, 1, 2, 3, …

R = the set of real numbers

Thus we have $\mathbf{N} \subset \boldsymbol{Z} \subset \mathbf{R}$

Example: ***Intervals*** on the real line, defined below, appear very often in mathematics. Here a and b are real numbers *with* $a < b$.

Open interval from a to b = $(a,b) = \{x : a < x < b\}$

Closed interval from a to b = $[a,b] = \{x : a \leq x \leq b\}$

Open-closed interval from a to b = $(a,b] = \{x : a < x \leq b\}$

Closed-open interval from a to b = $[a,b) = \{x : a \leq x < b\}$

The **open-closed** and **closed-open** intervals are also called ***half*-open**

Set Operations

- **Union:** Let **A** and **B** be arbitrary sets. The **union** of **A** and **B**, denoted by **A U B**, is the set of elements which belong to **A** or to **B**.

A U B = {x : x ∈ A or x ∈ B}. Here **"or"** is used in the sense of and/or.

OR

Let A and B be sets. The union of the sets A and B , denoted by **A U B**, is the set that contains those elements that are **either in A** or in **B** , or **in both**. An element x belongs to the union of the sets A and B if and only if x belongs to A or x belongs to B.

A U B = {x l x ∈ A v x ∈ B }.

Set Operations

EXAMPLE The union of the sets {1 , 3 , 5 } and {1 , 2, 3} is the set {1 , 2 , 3 , 5}

{1, 3, 5} U {1, 2, 3} = **{1, 2, 3, 5}.**

Set Operations

- **Intersection:** The intersection of A and B, denoted by **$A \cap B$**, is the set of elements belong which belong to both A and B

$A \cap B = \{x : x \in A$ and $x \in B\}$

OR

Let A and B be sets. The intersection of the sets A and B , denoted by A∩B, is the set containing those elements in both A and B.

$A \cap B = \{x \mid x \in A \land x \in B\}$.

Set Operations

Example: The intersection of the sets {1, 3 , 5} and {1, 2 , 3} is the set {1, 3}

{1, 3 , 5 } ∩ {1, 2, 3 } = {1, 3}.

Set Operations

- ***Disjoint:*** **If** $A \cap B = \emptyset$, that is, if **A** and ***B*** do not have any elements in common, then ***A*** and **B** are said to be ***disjoint*.**

  OR

  Two sets are called disjoint if their **intersection** is the **empty set**.

Set Operations

EXAMPLE Let A = {1, 3 , 5 , 7, 9} and B = {2, 4, 6, 8 , 10}.

$A \cap B = \emptyset$, A and B are disjoint.

Set Operations

- **Difference**: The ***difference*** of ***A*** and ***B*** or the ***relative complement*** of ***B*** with respect to ***A*,** denoted by ***A* \ *B***, is the set of elements which belong to ***A*** but not to *B***.**

  $A \setminus B = \{x : x \in A, x \notin B)$

  OR

  Let A and B be sets. The d**ifference** of A and B , denoted by **A - B**, is the set **containing those elements** that are in **A** but **not in B** . The difference of A and B is also called the complement of B with respect to A.

  $A - B = \{x \mid x \in A \wedge x \notin B\}$

Set Operations

EXAMPLE The difference of {1, 3 , 5} and {1, 2 , 3} is the set {5}

{1, 3 , 5} - {1, 2 , 3} = {5} .

{1, 2 , 3} - {1, 3 , 5} = {2}.

Set Operations

- **Complement:** The ***absolute complement*** or, simply, ***complement*** of A, denoted by A^c is the set of elements which do not belong to ***A:***

  $A^c = \{x : x \in U, x \notin A)$

  - That is, A^c is the difference of the universal set **U** and ***A.***

  OR

  Let **U** be the universal set. The **complement** of the set A, denoted by $\bar{A}$, is the complement of A with respect to U. In other words, the complement of the set A is U - A . $\bar{A}$**= {x | x ∉ A}**

Set Operations

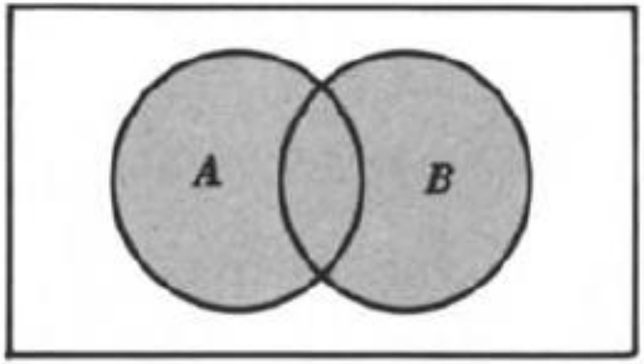
EXAMPLE Let A = {a , e, i , 0, u } (where the universal set is the set of letters of the English alphabet).

$\overline{\mathbf{A}}$ = {b, c, d, j, g, h , j, k, l , m , n , p , q , r, S , t , v , w , x , y , z} .

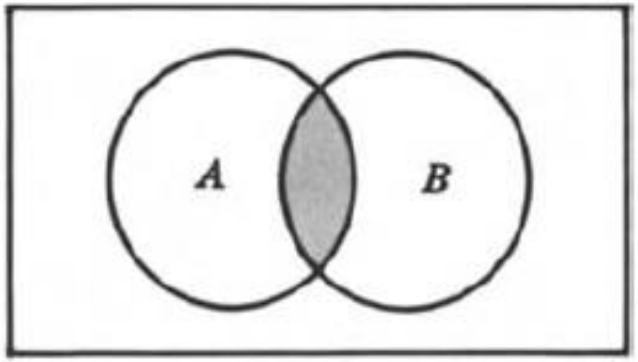
Set Operations

- **Example:** The diagrams on next slide, called **Venn diagrams**, illustrate the set operations discussed in the previous slides. Here sets are represented by simple plane areas and ***U*****,** the **universal set**, **by** the area in the entire rectangle.

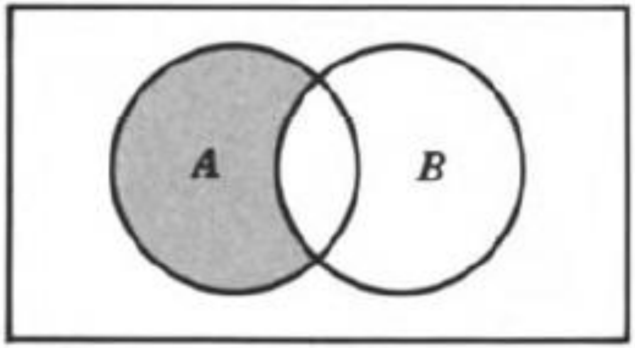
Example cont.



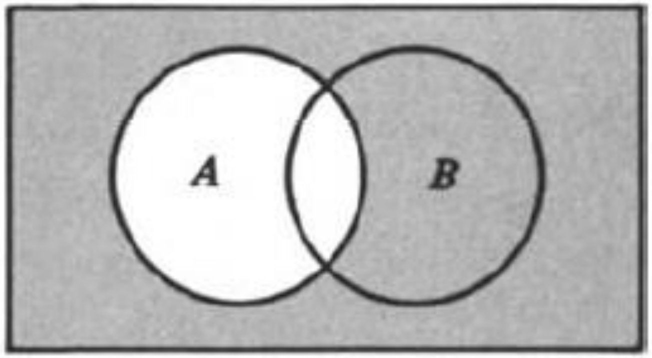
$A \cup B$ *is shaded*



$A \cap B$ *is shaded*



$A \setminus B$ *is shaded*



A^c *is shaded*

De Morgan's law

I. $(A \cup B)^C = A^c \cap B^c$

II. $(A \cap B)^C = A^c \cup B^c$

❑**Example:** Let U = {1, 3, 5, 7, 9, 2, 6, 4, 8, 10}, A = {3, 2, 7, 5 ,8, 9}, and B = {2, 5, 4, 8, 10}. Prove De Morgan's law of intersection.

$(A \cap B)^C = A^c \cup B^c$

Solution:

LHS = $(A \cap B)^C$

$A \cap B$ = {3, 2, 7, 5, 8, 9} $\cap$ {2, 5, 4, 8, 10}

= {2, 5, 8}

$(A \cap B)^C$ = {1, 3, 5, 7, 9, 2, 6, 4, 8, 10} - {2, 5, 8}

= {1, 3, 7, 9, 6, 4, 10}

LHS = {1, 3, 4, 6, 7, 9, 10}

RHS = $A^c \cup B^c$

A^c = U – A

= {1, 3, 5, 7, 9, 2, 6, 4, 8, 10} **-** {3, 2, 7, 5 ,8, 9}

= {1, 4, 6, 10}

B^c = U – B

= {1, 3, 5, 7, 9, 2, 6, 4, 8, 10} **-** {2, 5, 4, 8, 10}

= {1, 3, 6, 7, 9}

RHS = {1, 4, 6, 10} **U** {1, 3, 6, 7, 9}

= {1, 3, 4, 6, 7, 9, 10}

RHS = LHS

Cardinality of a set

- Let **S** be a **set**. If there are exactly **n distinct** elements in **S** where **n** is a **nonnegative integer**, we say that **S** is a **finite set** and that **n** is the **cardinality** of **S**. The cardinality of **S** is denoted by $|S|$.

Cardinality of a set

- **Example** Let **A** be the set of **odd positive integers** less than 10. Then $|A|$**= 5**.

- **Example** Let **S** be the **set of integers** in the English alphabet. Then $|A|$ **= 26**.

- **Example** Because the null set has no elements, it follows that $|\emptyset|$**= 0**.

Infinite, not finite, and power set

Definition A set is said to be **infinite** if it is **not finite**.

Example: The set of positive integers is **infinite.**

Definition Given a set **S**, the **power set** of **S** is the set of **all subsets** of the **set S**. The power set of **S** is denoted by **P(S).**

Example: What is the **power set** of the set to {0, 1 , 2} ?

Solution:

The power set **P({ 0, 1 , 2 })** is the set of all subsets of to {0, 1 , 2 } .

Hence,

P ({0, 1 , 2 }) = {{Ø}, {0}, {1}, {2 }, {0, 1}, {0, 2}, { 1 , 2}, {0, 1 , 2}} .

Example What is the power set of the **empty set**? What is the power set of the set **{Ø}?**

Solution: The **empty set** has exactly **one subset**, namely, itself.

$P(\emptyset) = \{\emptyset\}$.

The **set $\{\emptyset\}$** has exactly **two subsets**, namely, **$\emptyset$** and the set **$\{\emptyset\}$** itself. Therefore,

$P(\{\emptyset\}) = \{\emptyset, \{\emptyset\}\}$

Note:

No of elements in a power set: If a set has n elements, then its power set has **2^n elements**.

Cartesian Products [1]

The ordered n-tuple $(a_1 , a_2 , . . . , a_n)$ is **the ordered collection** that has **a_1 as its first element**, **a_2 as its second element**, . . . , and **a_n as its nth element**.

2-tuples are called **ordered pairs**. The ordered pairs **(a , b)** and **(c, d)** are equal if and only if **a = c** and **b = d**.

Note: (a , b) and (b, a) are not equal unless a = b.

Definition Let **A** and **B** be **sets**. The **Cartesian product** of **A** and **B**, denoted by **$A \times B$**, is the set of all ordered pairs **(a , b),** where **$a \in A$** and **$b \in B$** .

$$A \times B = \{(a , b) \mid a \in A \land b \in B \} .$$

Cartesian Products [2]

EXAMPLE: What is the Cartesian product of

$A = \{1, 2\}$ and $B = \{a, b, c\}$?

Solution:

The Cartesian product $A \times B$ is

$A \times B = \{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c)\}$.

Relation

- A **subset R** of the **Cartesian product A × B** is called a relation from the **set A** to the **set B**. The elements of **R** are **ordered pairs,** where the **first element** belongs to **A** and the **second** to **B**.

R = {(a , 0), (a , 1), (a , 3), (b, 1), (b, 2), (c, 0), (c, 3)} is a relation from the set {a , b , c} to the set to

{ 1 , 2, 3 }

- The **Cartesian products A × B** and **B × A** are not equal, unless **A = Ø** or **B = Ø** (so that A x B = Ø) or A = B