Statistical and Mathematical Methods for Data Analysis

Dr. Syed 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:

- **Elementary Statistics: Picturing the World,** 6th Edition, Ron Larson and Betsy Farber
- **Probability & Statistics for Engineers & Scientists**, Ninth Edition, Ronald E. Walpole, Raymond H. Myer
- **Probability Demystified**, Allan G. Bluman
- **Practical Statistics for Data Scientists: 50 Essential Concepts,** Peter Bruce and Andrew Bruce
- https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/
- http://www.thefreedictionary.com/statistics

These notes contain material from the above three resources.

Distribution of points

Midterm = 30 points

Final term = 40 points

Sessional points = 30 points

I. Assignments = 2×4 = 10 points

II. Hands-on Python in class = 0.5×5 = 2.5 points

III. Quizzes = 0.5×5 = 2.5 points

IV. Journal/conference paper presentation = 5

V. Mini project (its report should be in an IEEE journal paper format) = 10 points

Target Journals

Some of the journals that are relevant to health care and the medical field, based on computer science.

1. **Medical Decision Making**, JCR Impact Factor (2017-18) = 2.793
2. **Health Informatics Journal**, JCR Impact Factor (2017-18) = 2.297
3. **Informatics for Health and Social Care**, JCR Impact Factor (2017-18) = 1.137
4. **Health Care Analysis**, , JCR Impact Factor (2017-18) = 1.043
5. **International Journal of Health Care Quality Assurance**, JCR Impact Factor (2017-18) = 1.218

Target Journals

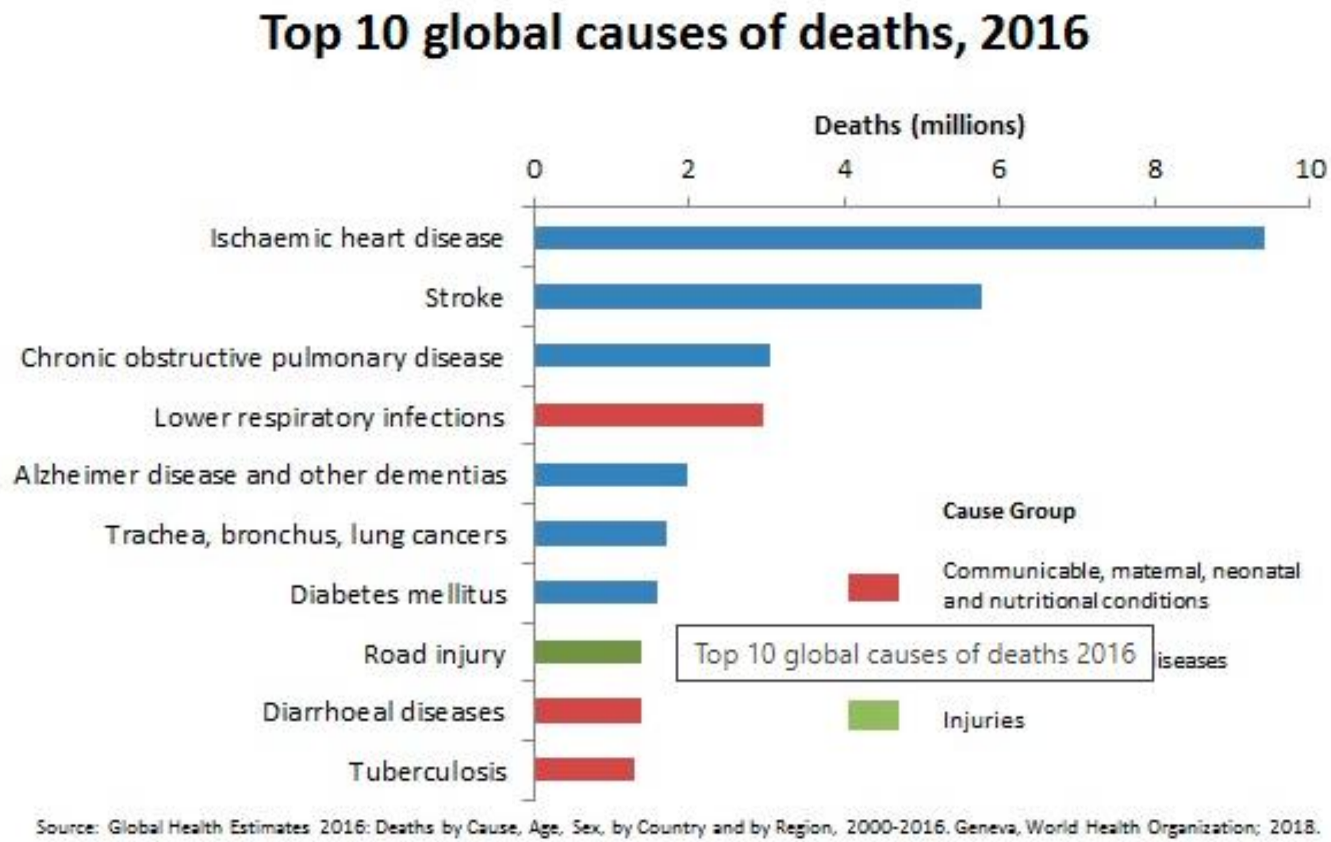
Some of the journals that are relevant to education, based on computer science.

1. **Computers & Education**, JCR Impact Factor (2017-18) = 5.627
2. **Computer Applications in Engineering Education**, JCR Impact Factor (2017-18) = **1.435**
3. **Journal of Computing in Higher Education**, JCR Impact Factor (2017-18) = **1.870**
4. **Acm Transactions on Computing Education**, , JCR Impact Factor (2017-18) = **1.356**
5. **Assessment & Evaluation In Higher Education**, JCR Impact Factor (2017-18) = **2.473**
6. **Educational Assessment Evaluation and Accountability**, JCR Impact Factor (2017-18) = **1.772**
7. **Computer Applications in Engineering Education = Impact Factor: 1.435**

- PAKISTAN: ROAD TRAFFIC ACCIDENTS
- Deaths = 30,046
- % = 2.42 (of total death in Pakistan)
- Rate = 17.12
- World Rank = 95
- According to the latest **WHO** data published in 2018 Road Traffic Accidents Deaths in Pakistan reached **30,046** or **2.42%** of total deaths. The age adjusted Death Rate is **17.12 per 100,000** of population ranks **Pakistan #95** in the world. Review other causes of death by clicking the links below or choose the full health profile.

Reference:https://www.worldlifeexpectancy.com/pakistan-road-traffic-accidents

- Road injuries **killed 1.4 million** people in 2016, about three-quarters (74%) of whom were men and boys.



Reference: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Basic concepts [1]

- **Probability** can be defined as the mathematics of chance.

- Statisticians use the word **experiment** to describe any process that **generates a set of data**.

OR

- A **probability experiment** is a chance process that leads to well defined outcomes or results. **For example**, tossing a coin can be considered a probability experiment since there are two well-defined outcomes—heads and tails.

Basic concepts [2]

- In probability theory, an **experiment or trial** is any procedure that can be **infinitely repeated** and has a **well-defined** set of possible **outcomes**, known as the **sample space**.

- An **outcome** of a probability experiment is the result of a single trial of a probability experiment.

Basic concepts [3]

- The set of all possible outcomes of a statistical experiment is called the **sample space** and is represented by the symbol ***S***.

OR

- The set of all outcomes of a probability experiment is called a **sample space**. Some sample spaces for various probability experiments are shown here.

Experiment	Sample space
Toss one coin	H, T
Roll a die	1, 2, 3, 4, 5, 6
Toss two coins	HH, HT, TH, TT

Basic concepts [4]

- Each outcome in a sample space is called an **element** or a **member** of the sample space, or simply a **sample point.**

- Each outcome of a probability experiment occurs at **random**.

- Each outcome of the experiment is **equally likely unless otherwise stated**.

Basic concepts [5]

- An **event** then usually consists of one or more outcomes of the sample space.

OR

- An **event** is a subset of a sample space.

- An event with one outcome is called a **simple event**.

- An event consists of two or more outcomes, it is called a **compound event**.

Example

A single die is rolled. List the outcomes in each event:

a. Getting an odd number

b. Getting a number greater than four

c. Getting less than one

Example cont.

Solution:

S = {1, 2, 3, 4, 5, 6}

a. Let **A** be the event contains the outcomes 1, 3, and 5.

A = {1, 3, 5}, n(A) = 3

b. Let **B** be the event contains the outcomes 5, and 6.

B = {5, 6}, n(B) = 2

c. Let **C** be the event that contains a number less than one

C = {}

Basic concepts [7]

Classical Probability:

The formula for determining the probability of an event **E** is

$$P(E) = \frac{n(E)}{n(S)}$$

OR

$$P(E) = \frac{\textbf{Number of outcomes contained in the event E}}{\textbf{Total number of outcomes in the sample space}}$$

Example:

Two coins are tossed; find the probability that both coins land heads up.

Solution:

S = {HH, HT, TH, and TT}

n(S) = 4

Let **A** be the event of getting a both heads

A = {HH}

n(A) = 1

P (A) = ¼ = **0.25 (or 25 %)**

Example:

A die is tossed; find the probability of each event:

a. Getting a two

b. Getting an even number

c. Getting a number less than 5

Example cont.

Solution:

S = {1, 2, 3, 4, 5, 6}

n(S) = 6

$$P(E) = \frac{\text{Number of outcomes contained in the event E}}{\text{Total number of outcomes in the sample space}}$$

a. Let A be the event of getting a “two”

A = {2}

n(A) = 1

P (A) = $^{1}/_{6}$ = 0.1667 (or 16.67%)

Example cont.

b. a. Let **B** be the event of getting a **“even number”**

A = {2, 4, 6}

n(A) = 3

P (B) = $^3/_6 = {}^1/_2$ **= 0.5 (or 50%)**

c. a. Let **C** be the event of getting a **“less than 5”**

C = {1, 2, 3, 4}

n(C) = 4

P (C) = $^4/_6 = {}^2/_3$ **= 0.6666 (or 66.67%)**

Basic concepts [8]

Rule 1: The probability of any event will always be a number from **zero to one**. Probabilities cannot be **negative** nor can they **be greater than one**.

Rule 2: When an event cannot occur, the probability will be **zero.**

Example: A die is rolled; find the probability of getting a 7.

Basic concepts [9]

Rule 3: When an event is certain to occur, the probability is **1**.

Example: A die is rolled; find the probability of getting a number less than 7.

Rule 4: The sum of the probabilities of all of the outcomes in the **sample space** is 1.

Example: $P(H) = {}^{1}/_{2}$, $P(T) = {}^{1}/_{2}$, $P(H) + P(T) = 1$.

Basic concepts [10]

Complement : The **complement** of an event A with respect to S is the subset of all elements of S that are not in A. We denote the complement of A by the symbol ***A' or $\bar{A}$ or A^c***

Rule 5: The probability that an event will not occur is equal to 1 minus the probability that the event will occur.

Example: P(H) = $^1/_2$, P(T) = 1- P(H) = $^1/_2$

Basic concepts

The **probability** of an event A is the sum of the weights of all **sample points** in A.

Therefore,

I. $0 \leq P(A) \leq 1$

II. $P(\varphi) = 0$

III. $P(S) = 1.$

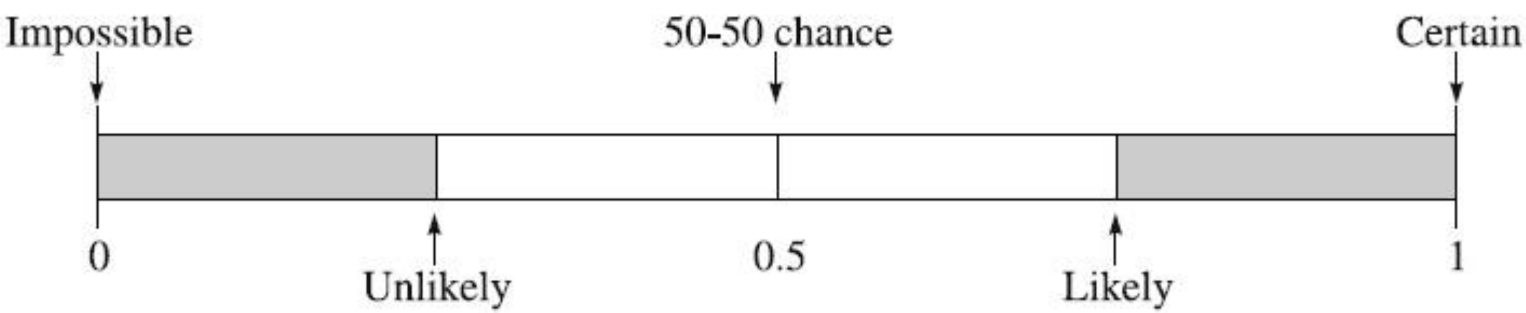
Basic concepts

- When the probability of an event is close to **zero**, the occurrence of the event is relatively **unlikely**. For example, if the chances that you will win a certain lottery are **0.00l** or one in one thousand, you probably won’t win, unless of course, you are very **“lucky.”**

- When the probability of an event is **0.5** or $^{1}/_{2}$, there is a **50–50 chance** that the event will happen—the same.

Basic concepts

When the probability of an event is close to one, the event is almost sure to occur. For example, if the chance of it snowing tomorrow is **90%,** more than **likely**, you'll see some snow.



Empirical Probability [1]

Probabilities can be computed for situations that do not use sample spaces. In such cases, frequency distributions are used and the probability is called **empirical probability**.

Rank	Frequency
Freshmen	4
Sophomores	6
Juniors	8
Seniors	7
TOTAL	25

Empirical Probability [2]

$$P(E) = \frac{\textbf{Frequency of E}}{\textbf{Sum of the frequencies}}$$

$P(E) = {}^{1}/_{4}$

Empirical probability is sometimes called **relative frequency probability**.

Law of large numbers

- In probability theory, the **law of large numbers (LLN)** is a theorem that describes the **result** of performing the **same experiment a large number of times**.

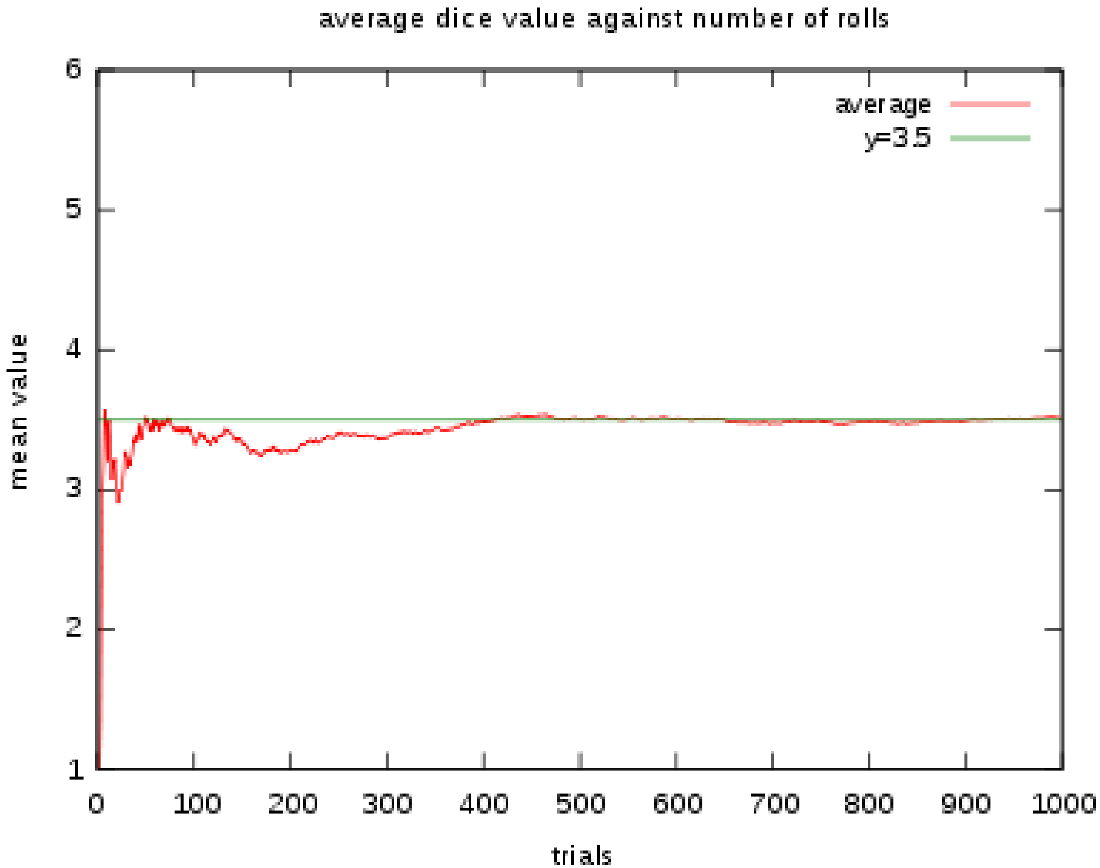
- According to the law, the **average** of the results obtained from a **large number of trials** should be close to the **expected value**, and will tend to become **closer** as more trials are performed.

Law of large numbers

- The LLN is important because it **"guarantees" stable long-term** results for the averages of some random events.

- **For example**, while a casino may lose money in a single spin of the **roulette** wheel, its earnings will tend towards a predictable percentage over a large number of spins.

Out come of a die
1
2
3
4
5
6
$\sum x = 21$
$\bar{x} = \frac{\sum x}{n} = \frac{21}{6} = 3.5$



An illustration of the law of large numbers using a particular run of rolls of a single die. As the number of rolls in this run increases, the **average** of the values of all the results approaches **3.5**.

Law of Large Numbers

Questions:

What happens if we toss the coin **100 times** ? Will we get **50** heads?

What will happen if we toss a coin **1000 times**? Will we get exactly **500** heads?

Law of Large Numbers

- **Solution:** Probably not.

- However, as the number of **tosses increases**, the ratio of the number of heads to the total number of tosses will get closer to $^{1}/_{2}$.

- This phenomenon is known as the **law of large numbers**.

Law of Large Numbers

- **Solution:** Probably not.

- However, as the number of **tosses increases**, the ratio of the number of heads to the total number of tosses will get closer to $\mathbf{1/2}$.

- This phenomenon is known as the **law of large numbers**.

Suggested Readings

2.1 Sample space

2.2 Events

2.3 Counting Sample Points