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:

- **Probability and Statistics for Engineers and Scientists**, Ninth edition, Ronald E. Walpole, Raymond H. Myer
- **Probability Demystified**, Allan G. Bluman
- **Elementary Statistics**, 10th Edition, Mario F. Triola
- **A First Course in Probability,** Eighth Edition, Sheldon Ross

These notes contain material from the above books.

“Yesterday is history, tomorrow is a mystery, but today is a gift. That’s why we call it the present.”

—Attributed to A. A. Milne, Bill Keane, and Oogway, the wise turtle in Kung Fu Panda

Concept of a Random Variable [1]

- **Statistics** is concerned with making **inferences** (conclusions) about **populations** and **population characteristics**. Experiments are conducted with results that are subject to chance.

- The testing of a number of electronic components is an example of a **statistical experiment**, a term that is used to describe any process by which several chance **observations are generated**.

- It is often important to allocate a **numerical description** to the outcome.

Concept of a Random Variable [2]

- **For example**, the sample space giving a detailed description of each possible outcome when **three electronic components** are tested may be written

S = {NNN, NND, NDN, DNN, NDD, DND, DDN, DDD}

where **N** denotes **nondefective** and **D** denotes **defective**.

Concept of a Random Variable [3]

- One is naturally concerned with the number of defectives that occur. Thus, each point in the sample space will be assigned a numerical value of **0, 1, 2, or 3**.

These values are, of course, **random quantities** determined by the outcome of the experiment

Random variables [1]

Variables whose values are **due to chance** are called random variables.

OR

When an experiment is performed and it **produced different results under the same condition** is called random variable (r.v). It is usually denoted by capital letter **X**.

OR

A **random variable** is a function that associates **a real number** with **each element in the sample space**.

Random variables [2]

OR

A **random variable** is a variable (typically represented by *x*) that has **a single numerical value**, determined by chance, for each outcome of a procedure.

Random variables [3]

- We shall use a capital letter, say **X**, to denote a random variable and its **corresponding small letter**, **x** in this case, for one of its values.
- In the electronic component testing illustration above, we notice that the random variable **X** assumes the value **2** for all elements in the subset

E = {DDN, DND, NDD} of the **sample space S**.

- That is, each possible value of **X represents an event** that is a **subset of the sample space** for the given experiment

Random variables [1]

Example: Two balls are drawn in succession **without replacement** from an urn containing **4 red balls** and **3 black balls**. The possible outcomes and the **values y** of the random variable *Y*, where ***Y* is the number of red balls**, are

Sample Space	y
RR	2
RB	1
BR	1
BB	0

Discrete Sample Space vs. Continuous Sample Space.

- If a sample space contains a **finite number of possibilities or an unending sequence** with as many elements as there are whole numbers, it is called a **discrete sample space**.

- If a sample space contains an **infinite number of possibilities** equal to the number of points on a line segment, it is called a **continuous sample space**.

Discrete Random Variable vs. Continuous Random Variable.

- A random variable defined over **the discrete sample space** is called **discrete random variable.**

OR

- A random variable is called a **discrete random variable** if its set of possible outcomes is **countable**.

OR

- **Discrete random variable** has either **a finite number of values** or **a countable number of values**, where **"countable"** refers to the fact that there might be **infinitely many values**, but they can be associated with a **counting process**.

Discrete Random Variable vs. Continuous Random Variable.

- But a random variable whose set of possible values is an **entire interval of numbers is not discrete**. When a random variable can take on values on a continuous scale, it is called a **continuous random variable.**

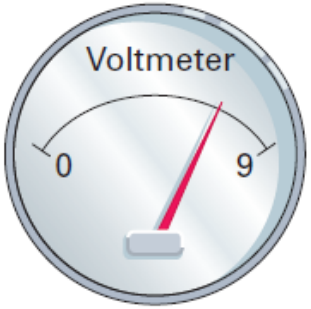
OR

- A **continuous random variable** has infinitely many values, and those values can be associated with measurements on a **continuous scale without gaps or interruptions**.

Devices Used to Count and Measure Discrete and Continuous Random Variables

(a) Discrete Random Variable: Count of the number of movie patrons.

(b) Continuous Random Variable: The measured voltage of a smoke detector battery.



Examples: Discrete vs. Continuous Random Variables.

- **Examples:**
- Let *x* the **number of eggs** that a hen lays in a day. This is a *discrete* random variable
- The **count of the number of statistics** students present in class on a given day is a whole number and is therefore a discrete random variable
- Let *x* the **amount of milk a cow produces in one day**. This is a ***continuous* random variable** because it can have any value over a continuous span.

Probability Distribution [1]

A **probability distribution** consists of the values of a random variable and their corresponding probabilities.

There are two kinds of **probability distributions**. They are **discrete** and **continuous**.

A **discrete variable** has a countable number of values (countable means values of zero, one, two, three, etc.). For example, **when four coins are tossed**, the outcomes for the number of heads obtained are zero, one, two, three, and four. **When a single die is rolled**, the outcomes are one, two, three, four, five, and six. These are examples of discrete variables.

Probability Distribution [2]

A **continuous variable** has an infinite number of values between any two values.

Continuous variables are measured. For example, **temperature** is a **continuous variable** since the variable can assume any value between **108** and 208 or any other two temperatures or values for that matter.

Height and weight are **continuous variables**. Of course, we are limited by our measuring devices and values of continuous variables are usually ‘‘rounded off.’’

.

Discrete Probability Distributions [1]

Example: Construct a discrete probability distribution for the number of heads when three coins are tossed.

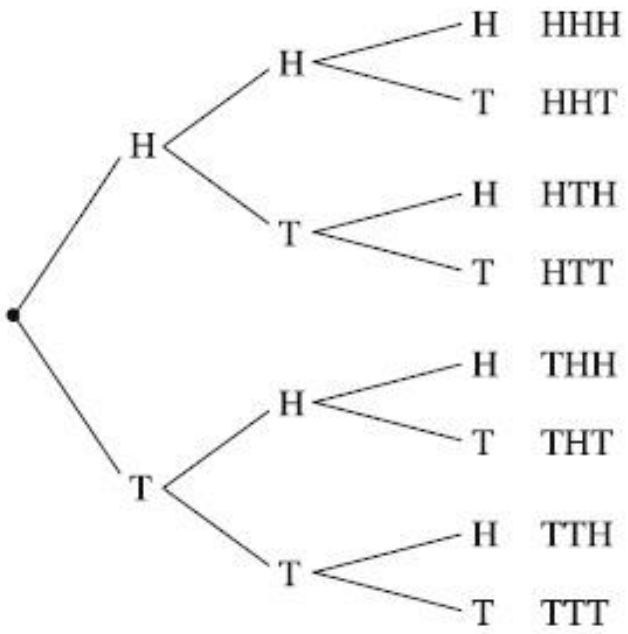
Solution: Recall that the sample space for tossing three coins is TTT, TTH, THT, HTT, HHT, HTH, THH, and HHH. The outcomes can be arranged according to the number of heads, as shown.

0 heads TTT

1 heads TTH

2 heads THH

3 heads HHH



Probability Distribution	
Value, x	Probability, P(x)
0	$\frac{1}{8} = 0.1250$
1	$\frac{3}{8} = 0.3750$
2	$\frac{3}{8} = 0.3750$
3	$\frac{1}{8} = 0.1250$
	$\sum P(x) = \frac{8}{8} = 1$

Discrete Probability Distributions [2]

- The **probability distribution** of a **discrete random variable X** is a list or table of the distinct numerical values of X and the probabilities associated with those values.
- The probability distribution is usually given in tabular form or in the form of an equation. For example, the discrete probability distribution for the previous problem is

x	1	2	3	4	5	6	
P(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\sum P(x) = \frac{6}{6} = 1$

Properties Of Discrete Probability Distribution

- Every probability distribution must satisfy the following two properties:

- Probability ranges from 0 to 1 i.e.,

  $0 \leq P(x) \leq 1$

- Sum of probabilities is one i.e.,

  $\sum P(x) = 1$

Discrete Probability Distributions

Some of the important type of discrete probability distributions are:

1. **Binomial Probability Distribution**
2. **Hypergeometric Distribution**
3. **Poisson Distribution**
4. **Geometric Distribution**
5. **Negative Binomial Distribution**
6. **Discrete Uniform Distribution**

Binomial Distribution [1]

A **binomial distribution** is obtained from a probability experiment called a binomial experiment. The experiment must satisfy these conditions:

1. Each trial can have only **two outcomes** or outcomes that can be **reduced to two outcomes**. The outcomes are usually considered as a success or a failure.

   have two classes or categories

2. There is a **fixed number** of trials.
3. The outcomes of each trial are **independent** of each other.
4. The probability of a **success** must remain the **same** for each trial.

- **Binomial probability distributions** allow us to deal with **circumstances** in which the outcomes belong to ***two*** **relevant categories**, such as **acceptable defective** or **survived died.** Other requirements are given in the following definition.

A Binomial Probability Distribution

- A **binomial probability distribution** results from a procedure that meets all the following requirements:

1. The procedure has a ***fixed number of trials***.

2. The trials must be ***independent***. (The outcome of any individual trial doesn't affect the probabilities in the other trials.)

3. Each trial must have all outcomes **classified into *two categories*** (commonly referred to as *success* and *failure*).

4. The probability of a **success remains the same** in all trials.

Notation for Binomial Probability Distributions

- **S and F (success and failure)** denote the two possible categories of all outcomes; ***p*** and ***q*** will denote the **probabilities of S and F**, respectively.

- **Note:** Be sure that *x* and *p* both refer to the *same* category being called a success.

Notation for Binomial Probability Distributions

P(S) = p	(p probability of a success)
P(F) = 1 – p = q	(q probability of a failure)
n	denotes the fixed number of trials.
x	denotes a specific number of successes in *n* trials, so *x* can be any whole number between 0 and *n*, inclusive.
p	denotes the probability of *success* in *one* of the *n* trials.
q	denotes the probability of *failure* in *one* of the *n* trials.
P(x)	denotes the probability of getting exactly *x* successes among the *n* trials.

Binomial Distribution [2]

Example: Explain why the probability experiment of tossing three coins is a binomial experiment.

Solution:

1. There are only **two outcomes** for each trial, head and tail. Depending on the situation, either heads or tails can be defined as a success and the other as a failure.
2. There is a **fixed number** of trials. In this case, there are three trials since three coins are tossed or one coin is tossed three times.
3. The outcomes are **independent** since tossing one coin does not effect the outcome of the other two tosses.
4. The probability of a **success** (say heads) is $\frac{1}{2}$ and it **does not change**. Hence the experiment meets the conditions of a binomial experiment.

Binomial Distribution [3]

The binomial probability formula is used to compute **probabilities for binomial random** variables. The binomial probability formula is given as:

$$b(x; n, p) = c_x^n p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

where $c_x^n = \frac{n!}{x!(n-x)!}$

OR

$$b(x; n, p) = \binom{n}{x} p^x q^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

where **n** = the total number of trials

x = the number of successes (0, 1, 2, 3, . . . , n)

p = the probability of a success

q = the probability of a failure

p + q = 1

- **Example** Five fair coins are flipped. If the outcomes are assumed independent, find the probability mass function of the number of heads obtained.

Probability mass function <=> Probability distributions

Solution

Here n = 5 **(Total number of coins)**

$p = \frac{1}{2}$ **(Probability of head)**

$q = 1 - p = 1 - \frac{1}{2} = \frac{1}{2}$ **(Probability of tail)**

Let X denotes number of heads

$$\mathbf{b(x; n, p) = \binom{n}{x} p^x q^{n-x},\ x = 0, 1, 2, \ldots, n}$$

$$\therefore b\left(x; \mathbf{5}, \mathbf{\frac{1}{2}}\right) = \binom{5}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{5-x},\ x = 0, 1, 2, \ldots, 5$$

$$P(X = 0) = \binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$P(X = 1) = \binom{5}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^4 = \frac{5}{32}$$

$$P(X = 2) = \binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32}$$

$$P(X = 3) = \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32}$$

$$P(X = 4) = \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 = \frac{5}{32}$$

$$P(X = 5) = \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = \frac{1}{32}$$

Probability Distribution

X	P(X = x)
0	$= \binom{5}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^5 = \frac{1}{32}$
1	$= \binom{5}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^4 = \frac{5}{32}$
2	$= \binom{5}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^3 = \frac{10}{32}$
3	$= \binom{5}{3}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^2 = \frac{10}{32}$
4	$= \binom{5}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^1 = \frac{5}{32}$
5	$= \binom{5}{5}\left(\frac{1}{2}\right)^5\left(\frac{1}{2}\right)^0 = \frac{1}{32}$
	$\sum_{i=0}^{5} P_i = 1$

SciPy

SciPy (pronounced “Sigh Pie”) is a Python-based ecosystem of open-source software for mathematics, science, and engineering. In particular, these are some of the **core packages**:

NumPy
Base N-dimensional array package

SciPy library
Fundamental library for scientific computing

Matplotlib
Comprehensive 2-D plotting

IPython
Enhanced interactive console

SymPy
Symbolic mathematics

pandas
Data structures & analysis

Reference: https://www.scipy.org/

SciPy library

- The SciPy library is one of the core packages that make up the SciPy stack. It provides many user-friendly and efficient numerical routines, such as routines for numerical integration, interpolation, optimization, linear algebra, and statistics.

Reference: https://www.scipy.org

Python code

```
from scipy.stats import binom
n = 5                  # Total number of coins
p = 0.5                # Probability of head
# Let x denotes number of heads
x = [0, 1, 2, 3, 4, 5]
#Compute probabilities
prob = binom.pmf(x, n, p)
#print probabilities
print(prob)
#[0.03125 0.15625 0.3125  0.3125
#0.15625 0.03125]
```

```
sumOfProb = sum(prob)

print('Sum of probabilities is:',
sumOfProb)

# Sum of probabilities is: 1.0
```

Example It is known that screws produced by a certain company will be **defective** with **probability .01**, independently of each other. The company sells the **screws in packages of 10** and offers a money-back guarantee that **at most 1 of the 10 screws is defective.** What proportion of packages sold must the company replace?

Here n = 10 **(Total number of screws)**

p = 0.01 **(Probability of defective)**

q = 0.99 **(Probability of non defective)**

Let X denotes number of defective screws

$$\mathbf{b(x; n, p) = \binom{n}{x} p^x q^{n-x},\ x = 0, 1, 2, \ldots, n}$$

$P(X \leq 1) = P(X = 0) + P(X = 1)$

$$= \binom{10}{0}(0.01)^0(0.99)^{10} + \binom{10}{1}(0.01)^1(0.99)^9$$

$$= 0.9044 + 0.0914 = 0.9958$$

Therefore probability that a package will have to be replaced is = **$P(X > 1) = 1 - P(X \leq 1)$**

= 1 - 0.9958 **= 0.0042 or 0.4200%**