

Statistical and Mathematical Methods for Data Analysis

Dr. Faisal Bukhari

**Punjab University College of Information Technology
(PUCIT)**

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 Demystified**, Allan G. Bluman
- ❑ **Schaum's Outline of Probability and Statistics**
- ❑ **MATLAB Primer**, Seventh Edition
- ❑ **MATLAB Demystified** by McMahan, David

Reference books

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

References

Readings for these lecture notes:

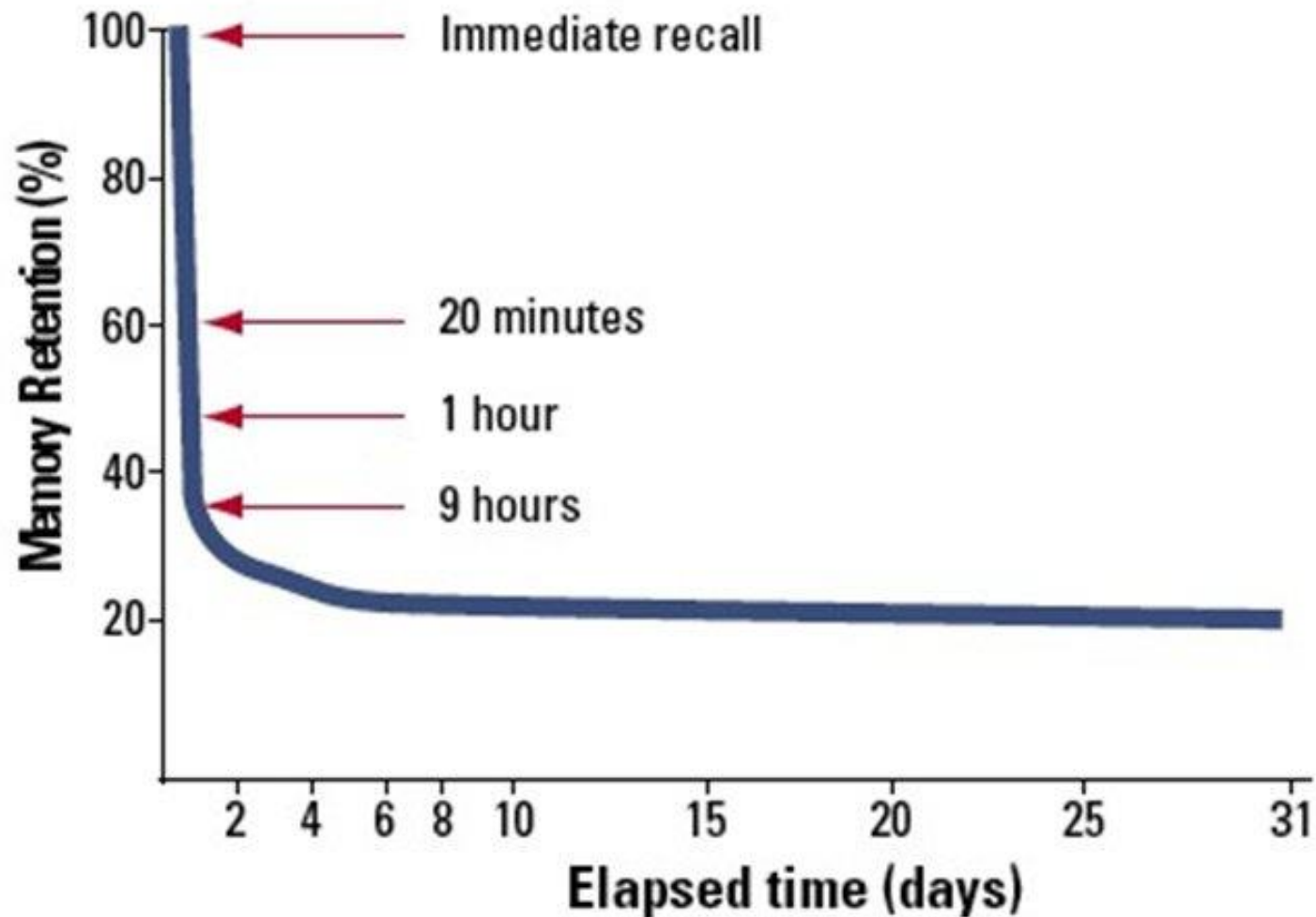
- ❑ Probability & Statistics for Engineers & Scientists, Ninth edition, Ronald E. Walpole, Raymond H. Myer
- ❑ <http://www.statisticshowto.com/geometric-distribution/>
- ❑ <https://peakmemory.me/category/forgetting-curve/>

These notes contain material from the above resources.

“If you want to know what a man's like, take a good look at how he treats his inferiors, not his equals.”

— J.K. Rowling, **Harry Potter and the Goblet of Fire**

Forgetting curve



Poisson Distribution [1]

Example: An automobile manufacturer is concerned about a fault in the braking mechanism of a particular model. The fault can, on rare occasions, cause a catastrophe at high speed. The distribution of the number of **cars per year** that will experience the fault is a Poisson random variable with $\lambda = 5$.

- (a) What is the probability that **at most 3** cars per year will experience a catastrophe?
- (b) What is the probability that **more than 1** car per year will experience a catastrophe?

Solution: Here $\lambda t = (5)(1) = 5$

$$P(x; \lambda t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, x = 0, 1, 2, \dots$$

$$\begin{aligned} \text{(a) } P(X \leq 3) &= \sum_{x=0}^3 p(x; 5) \\ &= 0.2650 \end{aligned}$$

$$\begin{aligned} \text{(b) } P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - \sum_{x=0}^1 p(x; 5) \\ &= 1 - 0.0404 \\ &= 0.9596 \end{aligned}$$

Poisson Distribution [2]

Example: Changes in airport procedures require considerable planning. Arrival rates of aircraft are important factors that must be taken into account. Suppose small aircraft arrive at a certain airport, according to a Poisson process, at the rate of **6 per hour**. Thus the Poisson parameter for arrivals for a period of hours is **$\lambda = 6$** .

(a) What is the probability that **exactly 4** small aircraft arrive during a **1-hour period**?

Poisson Distribution [3]

(b) What is the probability that **at least 4** arrive during a **1-hour period**?

(c) If we define a **working day** as **12 hours**, what is the probability that at least **75 small aircraft** arrive **during a day**?

Poisson Distribution [3]

Here $\lambda t = (6)(1) = 6$

$$P(x; \lambda t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, x = 0, 1, 2, \dots$$

$$(a) P(X = 4) = \frac{(6)^4 e^{-6}}{4!} = \mathbf{0.1339}$$

$$(b) P(X \geq 4) = 1 - P(x < 4) = 1 - \sum_{x=0}^3 p(x; 6) = 1 - 0.1512 = \mathbf{0.8488}$$

Here $\lambda t = (6)(12) = 72$

$$(c) P(X \geq 75) = 1 - P(x < 75) = 1 - \sum_{x=0}^{74} p(x; 72) = \mathbf{0.3773}.$$

Poisson Distribution using Python

a) What is the probability that **exactly 4** small aircraft arrive during a **1-hour period**?

```
from scipy.stats import poisson
mu = 6
x = 4
prob = round(poisson.pmf(x, mu), 4)
print('Probability that at least 4
arrive during a 1-hour period:', prob)
#0.1339
```

(b) What is the probability that **at least 4** arrive during a **1-hour period**?

```
x = [0, 1, 2, 3]
```

```
prob = 1 - round(sum(poisson.pmf(x,  
mu)), 4)
```

```
print('Probability that at least 4  
arrive during a 1-hour period:',  
prob)
```

#0.8488

(c) If we define a **working day** as **12 hours**, what is the probability that at least **75 small aircraft** arrive **during a day**?

```
mu = 12 * 6
```

```
x = range(0, 75)
```

```
#x = list(range(0, 75))
```

```
#print(x)
```

```
prob = 1 - round(sum(poisson.pmf(x,  
mu)), 4)
```

```
print('Probability that at least 75  
small aircraft arrive during a day:',  
prob)
```

```
#0.3773
```

Poisson approximation

The **Binomial distribution** converges towards the **Poisson distribution** as the number of trials goes to **infinity** while the product **np** remains fixed. Therefore the Poisson distribution with parameter **$\lambda = np$** can be used as an approximation to $b(n, p)$ of the binomial distribution if n is sufficiently large and p is sufficiently small.

According to two rules of thumb, this approximation is good if

$n \geq 20$ and $p \leq 0.05$, or if $n \geq 100$ and $np \leq 10$.

Poisson Distribution [4]

Formula:

$$f(x) = (e^{-\lambda} \lambda^x) / x! , x = 0, 1, 2, \dots$$

where, λ is an average rate of value, x is a Poisson random variable and e is the base of logarithm ($e = 2.718$).

Example:

Consider, in an office on **average 2 customers** arrived per day. Calculate the possibilities for exactly 3 customers to be arrived on today.

Step1: Find $e^{-\lambda}$.

where, $\lambda = 2$ and $e = 2.718$, $e^{-\lambda} = (2.718)^{-2} = 0.135$.

Step2: Find λ^x .

where, $\lambda = 2$ and $x = 3$, $\lambda^x = 2^3 = 8$.

Step3: Find $f(x)$.

$$f(x) = e^{-\lambda} \lambda^x / x!$$

$$f(3) = (0.135)(8) / 3! = 0.18.$$

Hence there are 18% possibilities for 3 customers to be arrived today

Geometric Distribution [1]

- Suppose we have a sequence of Bernoulli trials, each with a probability p of success and a probability $q = 1-p$ of failure. How many trials occur **before we obtain a success?**

Example

- A **search engine** goes through a list of sites looking for a **given key phrase**. Suppose the **search terminates** as soon as the **key phrase is found**. The number of sites visited is **Geometric**.

.

Let the random variable X be the number of trials needed to obtain a success. Then X has values in the range $\{1, 2, \dots\}$, and for $k \geq 1$,

$$g(x; p) = p q^{x-1}, x = 1, 2, 3, \dots$$

Alternative form

$$g(x; p) = p q^x, x = 0, 1, 2, 3, \dots$$

Geometric Distribution [2]

Mean = $1/p$ and Variance = q/p^2

In the theory of **probability and statistics**, a **Bernoulli trial** is an experiment whose outcome is random and can be either of **two possible** outcomes, “**success**” and “**failure**”.

Geometric Distribution [3]

Conditions:

An experiment consists of repeating trials **until first success**.

Each trial has **two possible outcomes**.

A success with probability **p**.

A failure with probability **q** = 1 – p.

Repeated trials are **independent**.

x = number of trials to first success

x is a **GEOMETRIC RANDOM VARIABLE**.

$$g(x; p) = q^{x-1}p, x = 1, 2, 3, \dots$$

Assumptions for the Geometric Distribution

The three assumptions are:

- ❑ There are **two possible outcomes** for each trial (success or failure).
- ❑ The trials are **independent**.
- ❑ The **probability of success** is the same for each trial.

Example From past experience it is known that **3%** of accounts in a large accounting population are in **error**. What is the probability that **5 accounts** are audited **before** an account in **error** is found?

Solution:

$$g(x; p) = q^{x-1} p, x = 1, 2, 3, \dots$$

$$P(X = 5) = P(\text{1st 4 correctly stated}) P(\text{5th in error})$$

$$= g(x; p) = q^{x-1} p, x = 1, 2, 3, \dots$$

$$= (0.97)^{5-1} (0.03)$$

$$= 0.0266$$

Example: In a certain manufacturing process it is known that, on the average, **1** in every **100**, items is defective. What is the probability that the **fifth item** inspected is the **first defective** item found?

Solution: Using the geometric distribution with $x = 5$ and

$p = 1/100 = 0.01$, $q = 0.99$, we have

$$g(x; p) = p q^{x-1}, \quad x = 1, 2, 3, \dots$$

$$\begin{aligned} g(5; 0.01) &= (0.01)(0.99)^{5-1} \\ &= 0.0096 \end{aligned}$$

Python code

```
from scipy.stats import geom
p = 1/100
x = 5
prob = round(geom.pmf(x, p), 4)
print('The probability that the fifth
item inspected is the first defective
item found :', prob)
# 0.0096
```

Example: At “busy time” a telephone exchange is very near capacity, so callers have difficulty placing their calls. It may be of interest to know the number of attempts necessary in order to gain a connection. Suppose that we let $p = 0.05$ be the probability of a connection during busy time. We are interested in knowing the probability that **5 attempts** are necessary for a successful call.

Solution:

Using the geometric distribution with $x = 5$ and $p = 0.05$ yields

$$g(x; p) = p q^{x-1}, x = 1, 2, 3, \dots$$

$$P(X = x) = g(5; 0.05)$$

$$= (0.05) (0.95)^{5-1}$$

$$= 0.041.$$

Python code

```
p = 0.05
```

```
x = 5
```

```
prob = round(geom.pmf(x, p), 4)
```

```
print('The probability that 5  
attempts are necessary for a  
successful call :', prob)
```


Discrete Uniform Distribution [1]

If a random variable has any of n possible values that are **equally probable**, then it has a discrete uniform distribution. The probability of any outcome k_i is **$1/n$** .

A simple example of the discrete uniform distribution is throwing a fair die. The possible values of k are **1, 2, 3, 4, 5, 6**; and each time the die is thrown, the probability of a given score is **$1/6$** .

Discrete Uniform Distribution [2]

Generating random numbers are the prime application of uniform distribution. The basic random numbers are **0, 1, 2, 3, 4, 5, 6, 7, 8, 9**. Each with probability equal to **$1/10$** .

For **two digit random numbers** the probability of selecting a particular random variable will be **$1/100$** .

Discrete Uniform Distribution [3]

If the random variable X assumes the values $x_1, x_1, x_2, \dots, x_k$ with equal probabilities, then the discrete uniform distribution is given by

$$P(x; k) = \frac{1}{k}, \quad x_1, x_2, x_3, \dots, x_k$$

Discrete Uniform Distribution [4]

When a light bulb is selected at random from a box that contains a 40-watt bulb, a 60-watt bulb, a 75-watt bulb, and a 100-watt bulb, each element of the sample space $S = \{40, 60, 75, 100\}$ occurs with probability $1/4$. Therefore, we have a uniform distribution, with probability

$$P(x; k) = \frac{1}{4}, \quad x = 40, 60, 75, 100$$