# Probability and Statstics

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,** $6^{th}$ Edition, Ron Larson and Betsy Farber
- **Elementary Statistics**, $13^{th}$ 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 McMahon, David

# 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/

- https://au.mathworks.com/help/stats/prob.normaldistribution.pdf.html#d117e648466

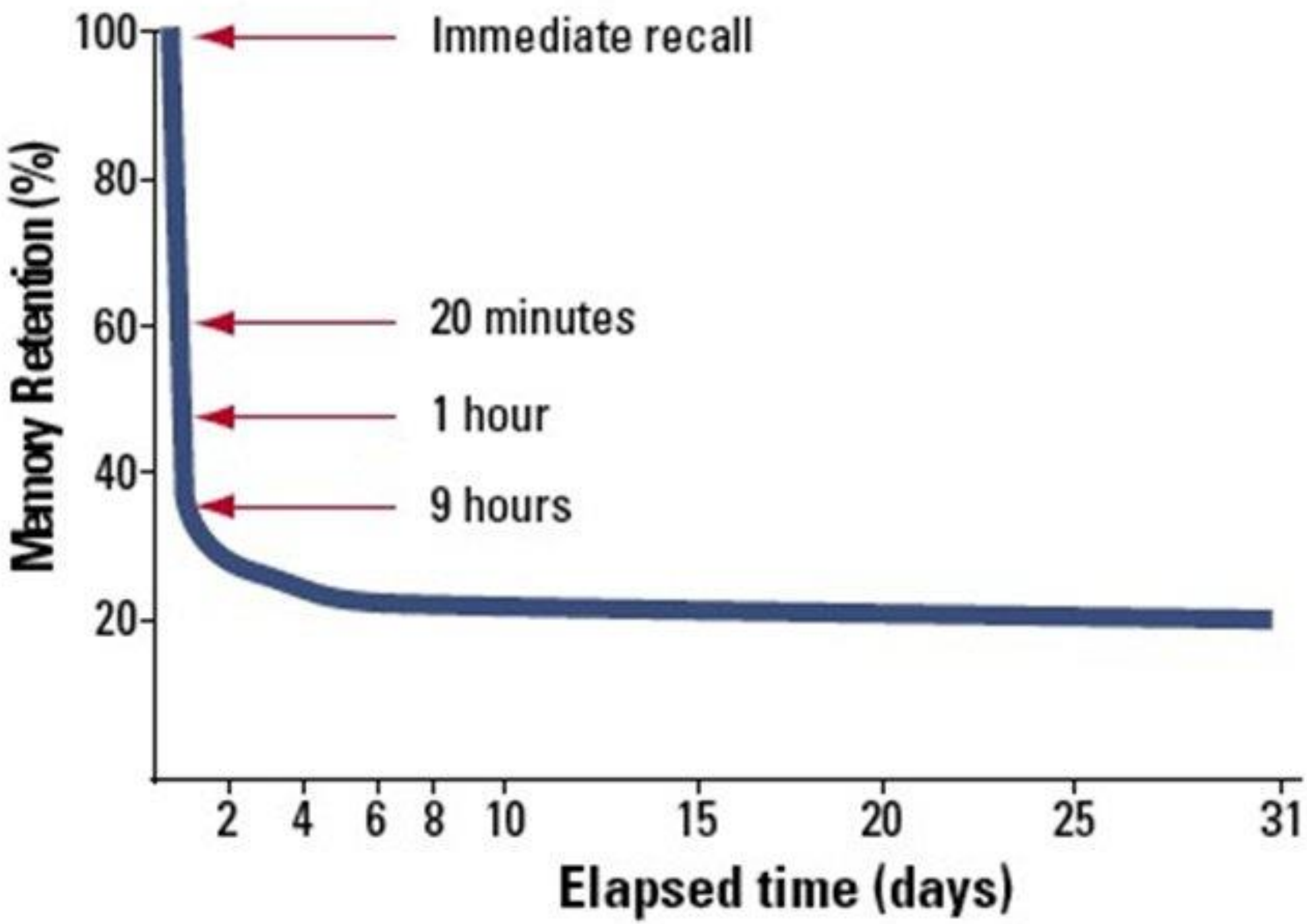
- **Elementary Statistics: Picturing the World,** $6^{th}$ Edition, Ron Larson and Betsy Farber


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 using MATLAB

**poisspdf** is Poisson probability density function in Matlab.

**Y = poisspdf(X, LAMBDA)** returns the Poisson probability density  function with parameter LAMBDA at the values in X

# 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:

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

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

**Step2:** Find $(\lambda t)^x$
where, $t = 1$, $\lambda t = 2$ and $x = 3$, $(\lambda t)^x = 2^3 = 8$.

**Step3:** Find $P(x; \lambda)$

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

$$P(3; 2) = \frac{(0.135)(8)}{3!} = 0.18.$$

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

# Geometric Distribution

- Many actions in life are **repeated until** a success occurs.

**For instance**, you might have to send an **e-mail several** times before it is **successfully sent**. A situation such as this can be represented by a **geometric distribution.**

# Geometric Distribution

A **geometric distribution** is a discrete probability distribution of a random variable *x* that satisfies these conditions.

**1.** A trial is **repeated** until a **success occurs**.

**2.** The repeated trials are **independent** of each other.

**3.** The probability of success *p* is the same for each trial.

**4.** The random variable *x* represents the number of the trial in which the **first success occurs**.

.

# Geometric Distribution

- The probability that the first success will occur on trial number $x$ is **$g(x; p) = p\, q^{x-1}, x = 1, 2, 3, \cdots$**

In other words, when the first success occurs on the third trial, the outcome is **FFS**, and the probability is

$P(3) = q \times q \times p$, or $P(3) = p \times q^2$.

# 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,…}, and for k ≥1,

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

**Alternative form**

**$g(x; p) = p\, q^{x}$, $x = 0, 1, 2, 3, \cdots$**

# 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.**

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

# 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** Basketball player LeBron James makes a **free throw** shot about **75%** of the time. Find the probability that the **first free** throw shot he makes occurs on the **third or fourth attempt**. (Source: National Basketball Association)

## Solution

Let x denotes number of attempts to get **first free** throw

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

p = 0.75 and q = 0.25

$g(3, 0.75) = (0.75)(0.25)^{3-1}$

$= 0.046875.$

$g(4, 0.75) = (0.75)(0.25)^{4-1}$

$= 0.011719.$

Since events are independent

$P(X = 3 \; or \; X = 4) = 0.046875 + 0.011719$

$= .059$

- Even though theoretically a **success may never occur**, the **geometric distribution** is a discrete probability distribution because the values of $x$ can be listed: 1, 2, 3, . . . .
- **Notice** that as ***x* becomes larger**, $P(x)$ gets **closer to zero**.

For instance, $P(15) = g(15, 0.75)$

$$= (0.75)(0.25)^{15-1}$$

$$= 0.0000000028.$$

**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:**

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

$= (0.97^4)(0.03)$

$= 0.0266$

**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

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

$g(5;0.01) = (0.01)(0.99)^{5-1}$

$= 0.0096$

# Geometric Distribution

## Syntax

**Y = geopdf(X,P)**

## Description

Y = geopdf(X,P) computes the geometric pdf at each of the values in X using the corresponding probabilities in P.

X and P can be vectors, matrices, or multidimensional arrays that all have the same size. A scalar input is expanded to a constant array with the same dimensions as the other input. The parameters in P must lie on the interval **[0 1].**

**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

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

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

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

$$= 0.041.$$

# Matlab code

```
p = 0.05
x = 4
prob = geopdf(x, p)

display(prob)

 % 0.0407
```

# 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 $x_1$, $x_1$, $x_2$, …, $x_k$ with equal probabilities, then the discrete uniform distribution is given by

$$\mathbf{P(x;\,k)} = \frac{\mathbf{1}}{\mathbf{k}}\,, \qquad \mathbf{x_1, x_2, x_3, \ldots, x_k}$$

# 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 sample1 space ***S* = {40, 60, 75, 100}** occurs with probability 1/4. Therefore, we have a uniform distribution, with probability

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

# Discrete Uniform Distribution using MATLAB [1]

## Syntax

## Y = unidpdf(X,N)

## Description

Y = unidpdf(X,N) computes the discrete uniform pdf at each of the values in X using the corresponding maximum observable value inN. X and N can be vectors, matrices, or multidimensional arrays that have the same size. A scalar input is expanded to a constant array with the same dimensions as the other inputs. The parameters in N must be positive integers.

# Discrete Uniform Distribution using MATLAB [2]

## Examples

For fixed n, the uniform discrete pdf is a constant.

```
>> y = unidpdf(1:10, 10)
y =   0.1000    0.1000    0.1000    0.1000    0.1000    0.1000
      0.1000    0.1000         0.1000    0.1000

>> y = unidpdf(1:6, 6)
y =   0.1667    0.1667    0.1667    0.1667    0.1667    0.1667
```