Statistical and Mathematical Methods

Statistical and Mathematical Methods for Data Science
DS5003

Dr. Nasir Touheed

Random Variables

Random Variables

- A random variable is a function of an outcome,

$$X = f(\omega)$$

- Consider the experiment of tossing two coins. We can define X to be a random variable that measures the number of heads observed in the experiment. For the experiment, the sample space is shown below:

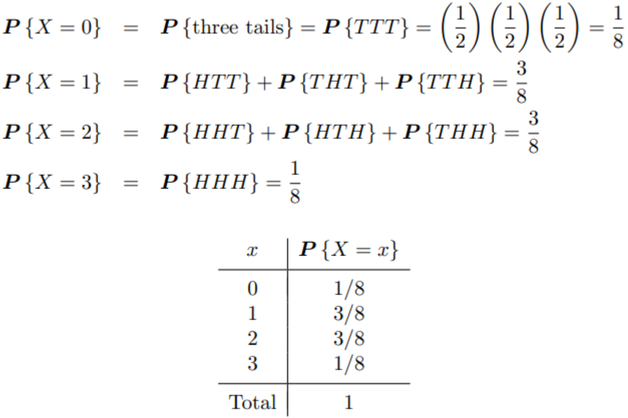
$$S = \{HH, HT, TH, TT\}$$

- There are 4 possible outcomes for the experiment, this is the domain of X.
- For each outcome, the associated value is shown as:

$$X(H,H) = 2$$
$$X(H,T) = 1$$
$$X(T,H) = 1$$
$$X(T,T) = 0$$

Example

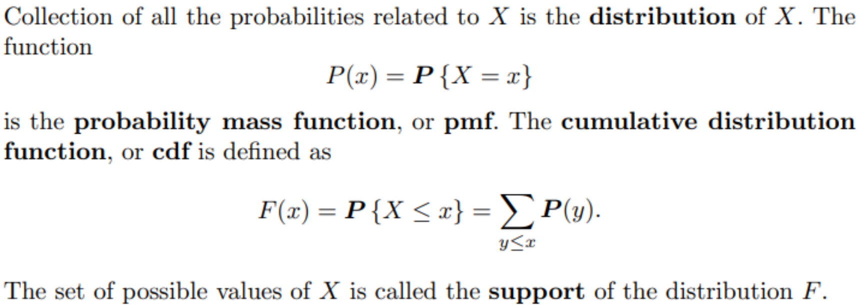
Consider an experiment of tossing 3 fair coins and counting the number of heads. Certainly, the same model suits the number of girls in a family with 3 children, the number of 1's in a random binary string of 3 characters, etc.



$$\begin{aligned}
\boldsymbol{P}\{X=0\} &= \boldsymbol{P}\{\text{three tails}\} = \boldsymbol{P}\{TTT\} = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = \frac{1}{8} \\
\boldsymbol{P}\{X=1\} &= \boldsymbol{P}\{HTT\} + \boldsymbol{P}\{THT\} + \boldsymbol{P}\{TTH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=2\} &= \boldsymbol{P}\{HHT\} + \boldsymbol{P}\{HTH\} + \boldsymbol{P}\{THH\} = \frac{3}{8} \\
\boldsymbol{P}\{X=3\} &= \boldsymbol{P}\{HHH\} = \frac{1}{8}
\end{aligned}$$

x	$\boldsymbol{P}\{X=x\}$
0	1/8
1	3/8
2	3/8
3	1/8
Total	1

Distribution of X



Collection of all the probabilities related to X is the **distribution** of X. The function

$$P(x) = \boldsymbol{P}\{X = x\}$$

is the **probability mass function**, or **pmf**. The **cumulative distribution function**, or **cdf** is defined as

$$F(x) = \boldsymbol{P}\{X \le x\} = \sum_{y \le x} \boldsymbol{P}(y).$$

The set of possible values of X is called the **support** of the distribution F.

For every outcome ω, the variable X takes one and only one value x. This makes events $\{X = x\}$ disjoint and exhaustive

$$\sum_{x} P(x) = \sum_{x} P\{X = x\} = 1$$

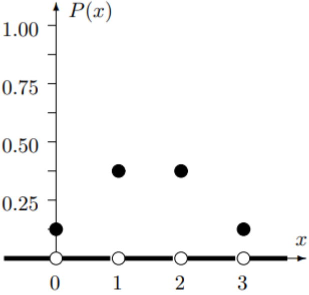
PMF and CMF Distribution of X

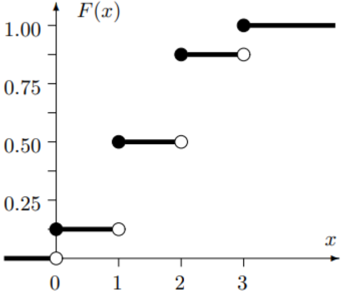
For any set

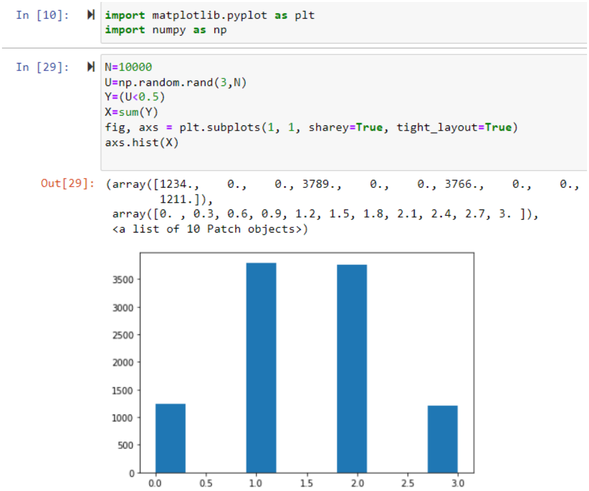
$$P\{X \in A\} = \sum_{x \in A} P(x)$$

When A is an interval, its probability can be computed directly from the cdf F(x),

$$P\{a < X \leq b\} = F(b) - F(a)$$







```
In [10]:  import matplotlib.pyplot as plt
          import numpy as np
```

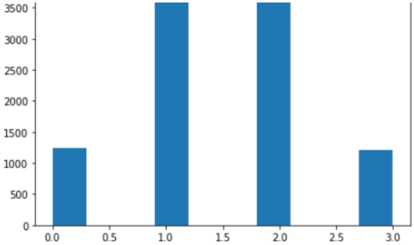
```
In [29]:  N=10000
          U=np.random.rand(3,N)
          Y=(U<0.5)
          X=sum(Y)
          fig, axs = plt.subplots(1, 1, sharey=True, tight_layout=True)
          axs.hist(X)
```

```
Out[29]: (array([1234.,    0.,    0., 3789.,    0.,    0., 3766.,    0.,    0.,
                 1211.]),
           array([0. , 0.3, 0.6, 0.9, 1.2, 1.5, 1.8, 2.1, 2.4, 2.7, 3. ]),
           <a list of 10 Patch objects>)
```

The following code simulate 3 coin tosses 10,000 times and produce a histogram of the obtained values of X

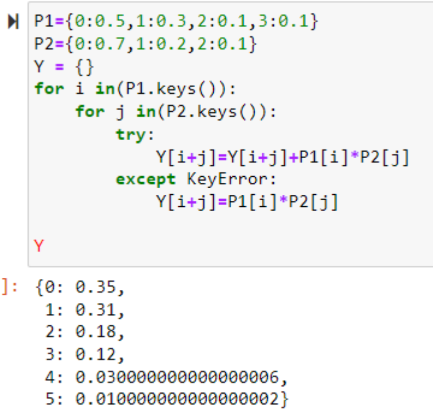
Histogram

- the two middle columns for X $=$ 1 and X $=$ 2 are about 3 times higher than the columns on each side.for X $=$ 0 and X $=$ 3.
- In a run of 10,000 simulations, values 1 and 2 are attained three times more often than 0 and 3.
- which is our pmf $P(0) = P(3) = 1/8, P(1) = P(2) = 3/8$



Example

- A program consists of two modules. The number of errors X_1 in the first module has the pmf $P_1(x)$, and the number of errors X_2 in the second module has the pmf $P_2(x)$, independently of X_1, where
- Find the pmf and cdf of $Y = X_1 + X_2$, the total number of errors



```
P1={0:0.5,1:0.3,2:0.1,3:0.1}
P2={0:0.7,1:0.2,2:0.1}
Y = {}
for i in(P1.keys()):
    for j in(P2.keys()):
        try:
            Y[i+j]=Y[i+j]+P1[i]*P2[j]
        except KeyError:
            Y[i+j]=P1[i]*P2[j]

Y
```

```
]: {0: 0.35,
    1: 0.31,
    2: 0.18,
    3: 0.12,
    4: 0.030000000000000006,
    5: 0.010000000000000002}
```

x	$P_1(x)$	$P_2(x)$
0	0.5	0.7
1	0.3	0.2
2	0.1	0.1
3	0.1	0

Types of Random Variables

- Discrete random variables: are random variables, whose range is a countable set. A countable set can be either a finite set or a countably infinite set. For instance, in the above example, X is a discrete variable as its range is a finite set $\{0, 1, 2\}$
- Continuous random variables, have a range in the forms of some interval, bounded or unbounded, of the real line. It can be e a union of several such intervals
- Mixed random variables are ones that are a mixture of both continuous and discrete variables. These variables are more complicated than the other two.

Examples of Random Variables

- A long jump is formally a continuous random variable because an athlete can jump any distance within some range. Results of a high jump, however, are discrete because the bar can only be placed on a finite number of heights.
- e. Examples of continuous variables include various times (software installation time, code execution time, connection time, waiting time, lifetime), also physical variables like weight, height, voltage.
- A job is sent to a printer.

Distribution of Random vectors

- Often we deal with several random variables simultaneously
- Computer Configuration
- Mobile Purchase
- Mobile Call Packages
- Car Selection
- Degree Selection

Joint and Marginal Distributions

- If X and Y are random variables, then the pair (X, Y) is a random vector.
- Its distribution is called the joint distribution of X and Y.
- Individual distributions of X and Y are then called the marginal distributions.
- Two vectors are equal,$(X, Y) = (x, y)$,, if X = x and Y = y.
- The joint probability mass function of X and Y is

$$P(x, y) = P\{(X, Y) = (x, y)\} = P\{X = x \cap Y = y\}$$

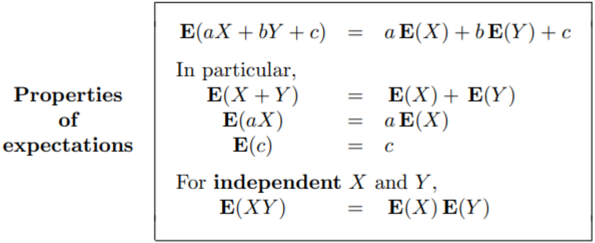
- $\sum_x \sum_y = 1$, because $\{(X, Y) = (x, y)\}$ are exhaustive and mutually exclusive events

A computer virus is trying to corrupt two files. The first file will be corrupted with probability 0.4. Independently of it, the second file will be corrupted with probability 0.3.

a. Compute the probability mass function (pmf) of X, the number of corrupted files.
b. Draw a graph of its cumulative distribution function (cdf).

Mean of a Discrete Random Variable

- The mean of a discrete random variable, denoted by μ, is actually the mean of its probability Distribution.$\mu = \sum xP(x)$
- The mean of a discrete random variable x is also called its expected value and is denoted by E(x). $E(x) = \sum xP(x)$



Properties of expectations

$$\mathbf{E}(aX + bY + c) = a\,\mathbf{E}(X) + b\,\mathbf{E}(Y) + c$$

In particular,

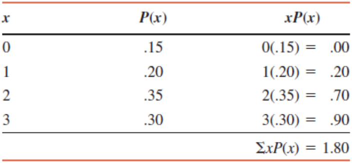
$$\begin{aligned}\mathbf{E}(X+Y) &= \mathbf{E}(X) + \mathbf{E}(Y)\\ \mathbf{E}(aX) &= a\,\mathbf{E}(X)\\ \mathbf{E}(c) &= c\end{aligned}$$

For **independent** X and Y,

$$\mathbf{E}(XY) = \mathbf{E}(X)\,\mathbf{E}(Y)$$

Examples

- Suppose that P(0) $= 0.75$ and P(1) $= 0.25$. Then, in a long run, X is equal 1 only 1/4 of times, otherwise it equals 0. Suppose we earn \$1 every time we see X $= 1$. On the average, we earn \$1 every four times, or \$0.25 per each observation
- Consider a variable that takes values 0 and 1 with probabilities P(0) $=$ P(1) $= 0.5$.
- Consider two users.One receives either 48 or 52 e-mail messages per day, with a 50-50% chance of each. The other receives either 0 or 100 e-mails, also with a 50-50% chance. Calculate $E(x)$ for both users.



x	$P(x)$	$xP(x)$
0	.15	0(.15) = .00
1	.20	1(.20) = .20
2	.35	2(.35) = .70
3	.30	3(.30) = .90
		$\Sigma xP(x) = 1.80$

Variance and Standard Deviation

- Expectation shows where the average value of a random variable is located, or where the variable is expected to be, plus or minus some error.
- How large could this "error" be, and how much can a variable vary around its expectation
- In Previous slide ,consider the first case, the actual number of e-mails is always close to 50, whereas it always differs from it by 50 in the second case.
- The first random variable, X, is more stable; it has low variability. The second variable, Y , has high variability.
- variability of a random variable is measured by its distance from the mean $\mu = E(X)$

Variance and Standard Deviation

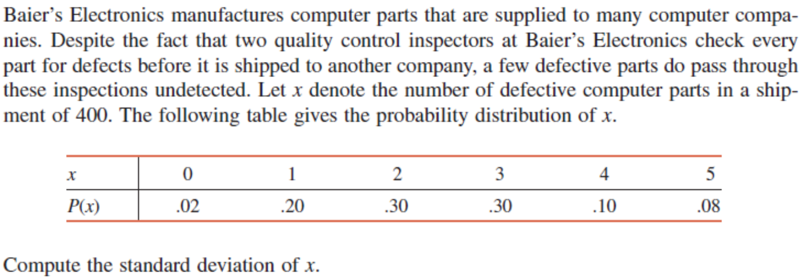
- Variance of a random variable is defined as the expected squared deviation from the mean. For discrete random variables, variance is

$$\sigma^2 = Var(x) = \sum_x (x - \mu)^2 P(x)$$

- Standard deviation is a square root of variance

$$\sigma = Std(X) = \sqrt{Var(X)}$$

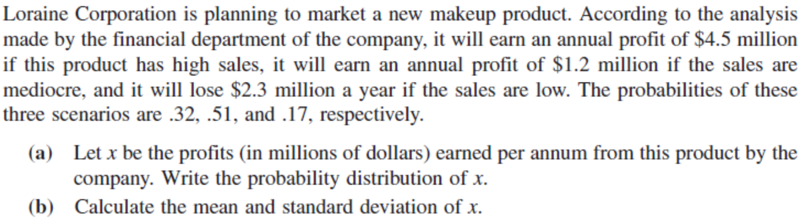
Example



Baier's Electronics manufactures computer parts that are supplied to many computer companies. Despite the fact that two quality control inspectors at Baier's Electronics check every part for defects before it is shipped to another company, a few defective parts do pass through these inspections undetected. Let x denote the number of defective computer parts in a shipment of 400. The following table gives the probability distribution of x.

x	0	1	2	3	4	5
$P(x)$	.02	.20	.30	.30	.10	.08

Compute the standard deviation of x.



Loraine Corporation is planning to market a new makeup product. According to the analysis made by the financial department of the company, it will earn an annual profit of $4.5 million if this product has high sales, it will earn an annual profit of $1.2 million if the sales are mediocre, and it will lose $2.3 million a year if the sales are low. The probabilities of these three scenarios are .32, .51, and .17, respectively.

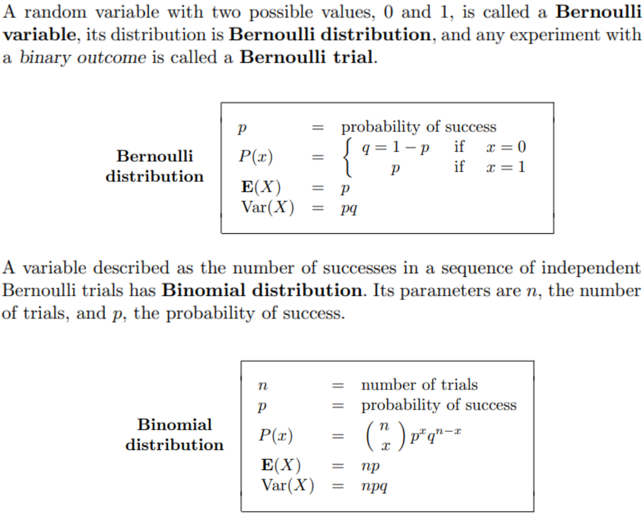
(a) Let x be the profits (in millions of dollars) earned per annum from this product by the company. Write the probability distribution of x.

(b) Calculate the mean and standard deviation of x.

Interpretation of the Standard Deviation

- According to Chebyshev's theorem, at least $(1 - 1/k^2) \times 100\%$ of the total area under a curve lies within k standard deviations of the mean, where k is any number greater than 1.
- if k $=$ 2,then at least 75% of the area under a curve lies between $\mu - 2\sigma$ and $mu + 2\sigma$.
- Chebyshev's inequality shows that in general, higher variance implies higher probabilities of large deviations, and this increases the risk for a random variable to take values far from its expectation.

Bernoulli and Binomial distribution



A random variable with two possible values, 0 and 1, is called a **Bernoulli variable**, its distribution is **Bernoulli distribution**, and any experiment with a *binary outcome* is called a **Bernoulli trial**.

Bernoulli distribution

$$\begin{array}{lcl} p & = & \text{probability of success} \\ P(x) & = & \begin{cases} q = 1-p & \text{if} \quad x=0 \\ p & \text{if} \quad x=1 \end{cases} \\ \mathbf{E}(X) & = & p \\ \mathrm{Var}(X) & = & pq \end{array}$$

A variable described as the number of successes in a sequence of independent Bernoulli trials has **Binomial distribution**. Its parameters are n, the number of trials, and p, the probability of success.

Binomial distribution

$$\begin{array}{lcl} n & = & \text{number of trials} \\ p & = & \text{probability of success} \\ P(x) & = & \binom{n}{x} p^x q^{n-x} \\ \mathbf{E}(X) & = & np \\ \mathrm{Var}(X) & = & npq \end{array}$$

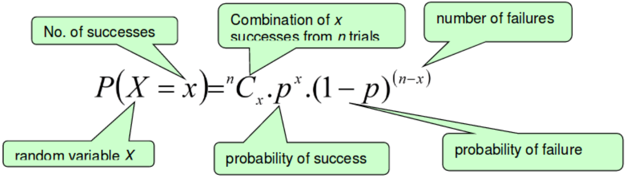
The Binomial Experiment:

An experiment that satisfies the following four conditions is called a binomial experiment.

a. There are n identical trials. In other words, the given experiment is repeated n times, where n is a positive integer. All of these repetitions are performed under identical conditions.
b. Each trial has two and only two outcomes. These outcomes are usually called a success and a failure, respectively. In case there are more than two outcomes for an experiment, we can combine outcomes into two events and then apply binomial probability distribution.
c. The probability of success is denoted by p and that of failure by q, and p + q= 1. The probabilities p and q remain constant for each trial.
d. The trials are independent. In other words, the outcome of one trial does not affect the outcome of another trial.

The Binomial Probability Distribution and Binomial Formula

- The random variable x that represents the number of successes in n trials for a binomial experiment is called a binomial random variable.
- The probability distribution of x in such experiments is called the binomial probability distribution.
- The binomial probability distribution is applied to find the probability of x successes in n trials for a binomial experiment.
- The number of successes x in such an experiment is a discrete random variable.



$$P(X = x) = {}^{n}C_{x} . p^{x} . (1 - p)^{(n-x)}$$

No. of successes

Combination of *x* successes from *n* trials

number of failures

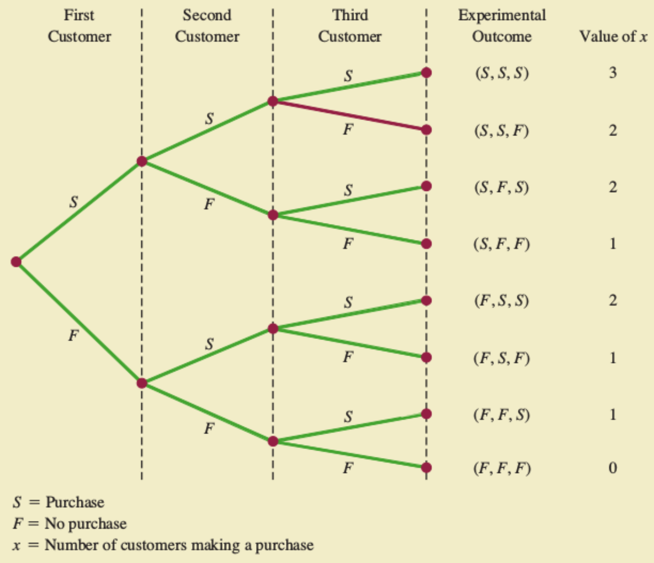
random variable *X*

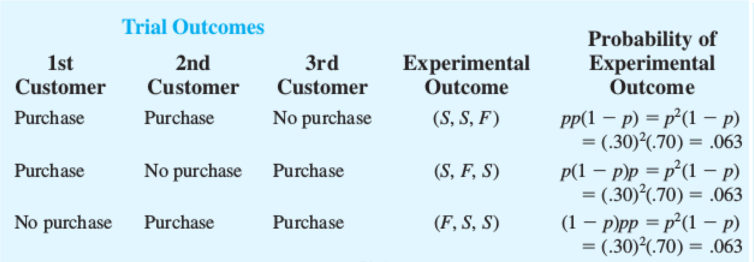
probability of success

probability of failure

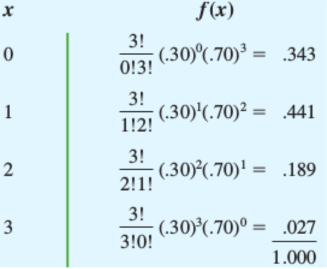
Example

- Question: Consider the purchase decisions of the next three customers who enter the Clothing Store. On the basis of past experience, the store manager estimates the probability that any one customer will make a purchase is .30. What is the probability that two of the next three customers will make a purchase?
- Let x be the number of Success in a sample of three.
- x can assume any of the values 0, 1, 2, and 3.
- it is a discrete random variable.


First Customer
Second Customer
Third Customer
Experimental Outcome
Value of x
S
F
S
F
S
F
S
F
S
F
S
F
S
F
(S, S, S) 3
(S, S, F) 2
(S, F, S) 2
(S, F, F) 1
(F, S, S) 2
(F, S, F) 1
(F, F, S) 1
(F, F, F) 0
S = Purchase
F = No purchase
x = Number of customers making a purchase



Trial Outcomes				
1st Customer	2nd Customer	3rd Customer	Experimental Outcome	Probability of Experimental Outcome
Purchase	Purchase	No purchase	(S, S, F)	$pp(1-p) = p^2(1-p)$ $= (.30)^2(.70) = .063$
Purchase	No purchase	Purchase	(S, F, S)	$p(1-p)p = p^2(1-p)$ $= (.30)^2(.70) = .063$
No purchase	Purchase	Purchase	(F, S, S)	$(1-p)pp = p^2(1-p)$ $= (.30)^2(.70) = .063$



x	$f(x)$
0	$\frac{3!}{0!3!}(.30)^0(.70)^3 = .343$
1	$\frac{3!}{1!2!}(.30)^1(.70)^2 = .441$
2	$\frac{3!}{2!1!}(.30)^2(.70)^1 = .189$
3	$\frac{3!}{3!0!}(.30)^3(.70)^0 = .027$
	1.000

Question: What is the probability of making exactly four sales to 10 customers entering the store.

we have a binomial experiment with n = 10, x = 4, and p =.30

Example

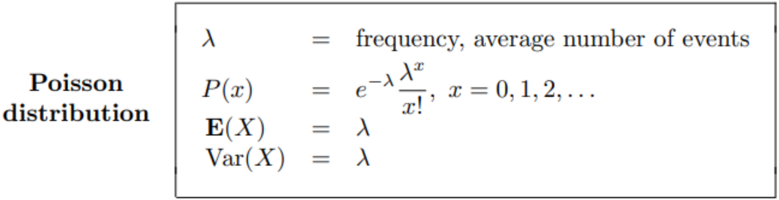
- Five percent of all DVD players manufactured by a large electronics company are defective. Three DVD players are randomly selected from the production line of this company. What is the probability that exactly one of these three DVD players is defective
- At the Express House Delivery Service, providing high-quality service to customers is the top priority of the management. The company guarantees a refund of all charges if a package it is delivering does not arrive at its destination by the specified time. It is known from past data that despite all efforts, 2% of the packages mailed through this company do not arrive at their destinations within the specified time. Suppose a corporation mails 10 packages through Express House Delivery Service on a certain day.

  Calculating the probability using the binomial formula.

  - Find the probability that exactly one of these 10 packages will not arrive at its destination within the specified time.
  - Find the probability that at most one of these 10 packages will not arrive at its destination within the specified time.
- An exciting computer game is released. Sixty percent of players complete all the levels. Thirty percent of them will then buy an advanced version of the game. Among 15 users, what is the expected number of people who will buy the advanced version? What is the probability that at least two people will buy it?

The Poisson Probability Distribution:

- The number of rare events occurring within a fixed period of time has Poisson distribution.
- The following examples also qualify for the application of the Poisson probability distribution.
- The number of accidents that occur on a given highway during a 1-week period
- The number of customers entering a grocery store during a 1-hour interval
- The number of television sets sold at a department store during a given week
- The following three conditions must be satisfied to apply the Poisson probability distribution.
  - x is a discrete random variable.
  - The occurrences are random.
  - The occurrences are independent.



Poisson distribution

$$\begin{aligned}
\lambda &= \text{frequency, average number of events}\\
P(x) &= e^{-\lambda}\frac{\lambda^x}{x!},\ x = 0, 1, 2, \ldots\\
\mathbf{E}(X) &= \lambda\\
\mathrm{Var}(X) &= \lambda
\end{aligned}$$

- PROPERTIES OF A POISSON EXPERIMENT
- The probability of an occurrence is the same for any two intervals of equal length.
- The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.

- On average, a household receives 9.5 telemarketing phone calls per week. Using the Poisson probability distribution formula, find the probability that a randomly selected household receives exactly 6 telemarketing phone calls during a given week.
- A washing machine in a laundromat breaks down an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month this machine will have
  - exactly two breakdowns
  - at most one breakdown
- The number of emails that I get in a weekday can be modeled by a Poisson distribution with an average of 0.2 emails per minute.
  - What is the probability that I get no emails in an interval of length 5 minutes?
  - What is the probability that I get more than 3 emails in an interval of length 10 minutes?