Data collection, Analysis and Inference

Subject Code: CPE-RPE,

May 2021,
SRM Univeristy-AP, Andhrapradesh

Lecture- 2:
Basic Statistical Distributions and their applications: Binomial, Poisson Distribution.

- Aim: To be able to recognize systems of discrete variable nature and their probability distributions.

Look at these examples

(1) In the experiment of rolling two dice, consider the event E that the sum of dice is 7

(2) In the experiment of tossing 4 coins, consider the event F that the outcome has at least 2 heads

(3) Four balls are randomly selected, without replacement, from an urn containing 20 balls numbered 1 through 20. Consider the event G that out of the 4 selected balls, the largest number is 10

- (1) In the experiment of rolling two dice, consider the event E that the sum of dice is 7

  Here, we are interested in the "sum" of the dice and not in the actual outcomes of individual dice

  Once the two dice are rolled, let the variable X denote the sum of dice.

  Then, **P(E) = P {X = 7}**

- (2) In the experiment of tossing 4 coins, consider the event F that the outcome has at least 2 heads

  Here, we are interested in the "number of total heads" and not in the actual outcome of individual toss. If we denote by the variable Y , the number of heads in the 4 tosses, then **P(F) = P {Y ≥ 2}**

- (3) Four balls are randomly selected, without replacement, from an urn containing 20 balls numbered 1 through 20.

  Consider the event G that out of the 4 selected balls, the largest number is 10

  Here, we are interested in the "largest numbered ball" and not in the actual sample of 4 balls

- Define the variable Z to be the largest number among the four selected balls

  Then, $P(G) = P\{Z = 10\}$

- In all the examples, we were interested in a "variable" whose value is dependent on the outcome of the experiment and we did not care about what the actual outcomes were!
- These quantities of interest are known as random variables.
- **Given an experiment whose sample space is S, a random variable X is real-valued function defined on the sample space S.**

$$\textbf{i.e., } X : S \to \mathbb{R}$$

- Example: Suppose that our experiment consists of tossing 3 fair coins. If we let Y denote the number of heads that appear in the three tosses.
- What are the values Y can take?

  $Y = 0, 1, 2, 3$
- $P\{Y = 0\} = P\{(T, T, T)\} = 1/8$
- $P\{Y = 1\} = P\{(T, T, H), (T, H, T), (H, T, T)\} = 3/8$
- $P\{Y = 2\} = P\{(T, H, H), (H, T, H), (H, H, T)\} = 3/8$
- $P\{Y = 3\} = P\{(H, H, H)\} = 1/8$

Cumulative distribution function

For a random variable X, the function F defined by

$$F(x) = P\{X \leq x\}, \quad -\infty < x < \infty$$

is called the **cumulative distribution function** or, simply, the **distribution function** of X.

Suppose $a \leq b$

Then, $\{X \leq a\} \subset \{X \leq b\}$

$$\Longrightarrow P\{X \leq a\} \leq P\{X \leq b\}$$

$$\Longrightarrow F(a) \leq F(b)$$

Note: **F (x) is a non-decreasing function of x**

- Now consider two experiments
- Experiment 1: Flipping a coin infinite no of times
- Experiment 2: Measuring lifetime of an electronic device
- Let X be the number of heads in experiment 1 and Y be the lifetime in hours in experiment 2
- X and Y are random variables. What are the values that X and Y can take?
- $X = 0, 1, 2, \ldots$ and $0 \leq Y \leq \infty$

Discrete

- What is the difference between X and Y ?
- X is taking only countably many values and Y is not so.
- Such random variables which can take at most a countable number of values is said to be discrete

- A random variable that takes only countable number of values is said to be a **discrete random variable**.
- For a discrete random variable X, we define the **probability mass function (p.m.f)**, $p(\cdot)$, of X by

$$p(a) = P\{X = a\} \text{ for every real number } a$$

Some properties of p.m.f

- Note that, as a function, $p : \mathbb{R} \rightarrow [0, 1]$
- For a given $b \in \mathbb{R}$, if the random variable does not take the value b, then $p(b) = P\{X = b\} = 0$
- Thus, if X can takes only the values $x_1, x_2, \ldots,$ then $p(x_i) \geq 0$ for $i = 1, 2, \ldots$

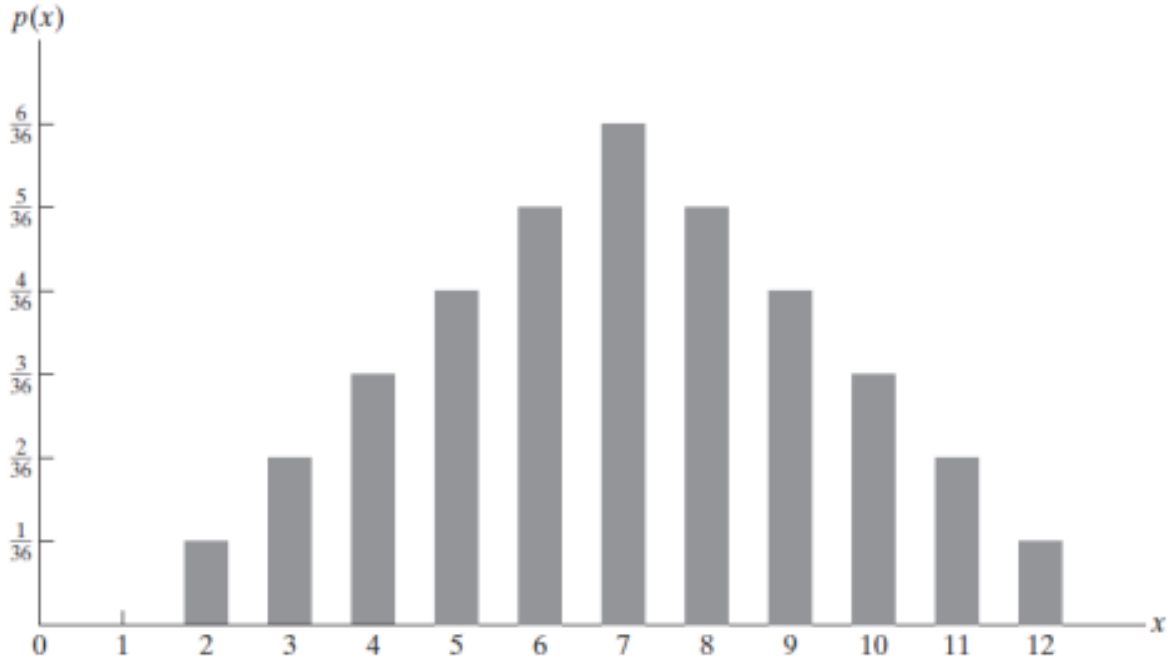
  $p(x) = 0$ for all other values of x.
- Since X must take one of the values x_i, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1$$

Example: Consider the experiment of rolling a pair of dice and the random variable X be the sum of the dice

We have the following

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$



x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

- **Example**: Five men and five women are ranked according to their scores on an examination. Assume that no two scores are alike and all 10! possible ranking are equally likely.

  Let X denote the highest ranking achieved by a woman.
- Find the probability mass function (p.m.f) of X.

- Solution: Step-1: Identify the values X can take

  First of all, note that the lowest possible position is 6, which means that all five women score worse than the five men.

  That is, X can take values 1, 2, 3, 4, 5, 6

- Step-2: Find $p(a) = P(X = a)$ for each value a that X can take

  For $X = 6$, we first need to find the number of different ways to arrange the 10 people such that the women all scored lower than the men.

- There are 5! ways to arrange the women and 5! ways to arrange the men, so $P\{X = 6\} = 5!5!/10!$

- Now consider the top woman scoring 5 th on the exam.
- There are 5 possible positions for the lower scoring women, and we have 4 women that must be assigned to these ranks.
- This can be accomplished in ($^{5}c_{4}$) different ways

Additionally, these 5 women can be arranged in 5! ways, and the men can be arranged in 5! ways. Thus,

$$P\{X = 5\} = \frac{\binom{5}{4} \cdot 5! \cdot 5!}{10!}$$

- Similarly for P{X = 4}, there are 6 positions for the 4 remaining women.
- The women can be arranged in 5! ways and the men can be arranged in 5! ways.

$$\Longrightarrow P\{X=4\} = \frac{\binom{6}{4} \cdot 5! \cdot 5!}{10!}$$

- By exactly same argument, we get

$$P\{X=3\} = \frac{\binom{7}{4} \cdot 5! \cdot 5!}{10!}$$

$$P\{X=2\} = \frac{\binom{8}{4} \cdot 5! \cdot 5!}{10!}$$

$$P\{X=1\} = \frac{\binom{9}{4} \cdot 5! \cdot 5!}{10!}$$

- The cumulative distribution function of X is given by $F(a) = P\{X \leq a\}$ for every real number a
- If X is discrete, F is discontinuous precisely at the values which X takes
- The jump at each discontinuity $X = a$ is given by the p.m.f $p(a)$
- $P(a < X \leq b) = \sum_{x:a<x\leq b} p(x) = F(b) - F(a)$
- $P(a \leq X < b) = \sum_{x:a\leq x<b} p(x) = F(b) - F(a) - p(b) + p(a)$
- $P(a \leq X \leq b) = \sum_{x:a\leq x\leq b} p(x) = F(b) - F(a) + p(a)$

Expectations

If X is a discrete random variable having a p.m.f $p(x)$, then the **expectation**, or the **expected value**, of X, denoted by $E[X]$, is defined by

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

It is the average value that a random variable will take if we only repeat our experiment often enough!

Second order moment

- In fact, we may never observe the expected value!
- Let X be a discrete random variable with p.m.f p_X and $Y = X^2$
- Then Y is also a discrete random variable.
- What will be E [Y]?
-

Fact: $$E[Y] = \sum_{x:p_X(x)>0} x^2 p_X(x)$$

- Given a random variable X, its expected value, $E[X]$, is also referred to as the **mean** or the **first moment of** X.
- The quantity $E[X^n], n \geq 1$, is called the n^{th} **moment of** X

Fact 9. For random variables $X_1, X_2, \ldots, X_n$,

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right]$$

Variance as an expectation

If X is a random variable with mean $\mu = E[X]$, then the **variance** of X, denoted by **Var**(X), is defined by

$$\text{Var}(X) = E\left[(X - \mu)^2\right]$$

Result: $\text{Var}(X) = E[X^2] - (E[X])^2$

The square root of the $\text{Var}(X)$ is called the **Standard deviation** of X, and we denote it by $SD(X)$. That is,

$$SD(X) = \sqrt{\text{Var}(X)}$$

Bernoulli random variable

A random variable X is said to be a **Bernoulli random variable** if it takes only two values $0, 1$ and its probability mass function is given by

$$p(a) = \begin{cases} 1-p, & \text{if } a = 0, \\ p, & \text{if } a = 1, \\ 0, & \text{else,} \end{cases}$$

for some $p \in (0, 1)$.

Binomial random variable

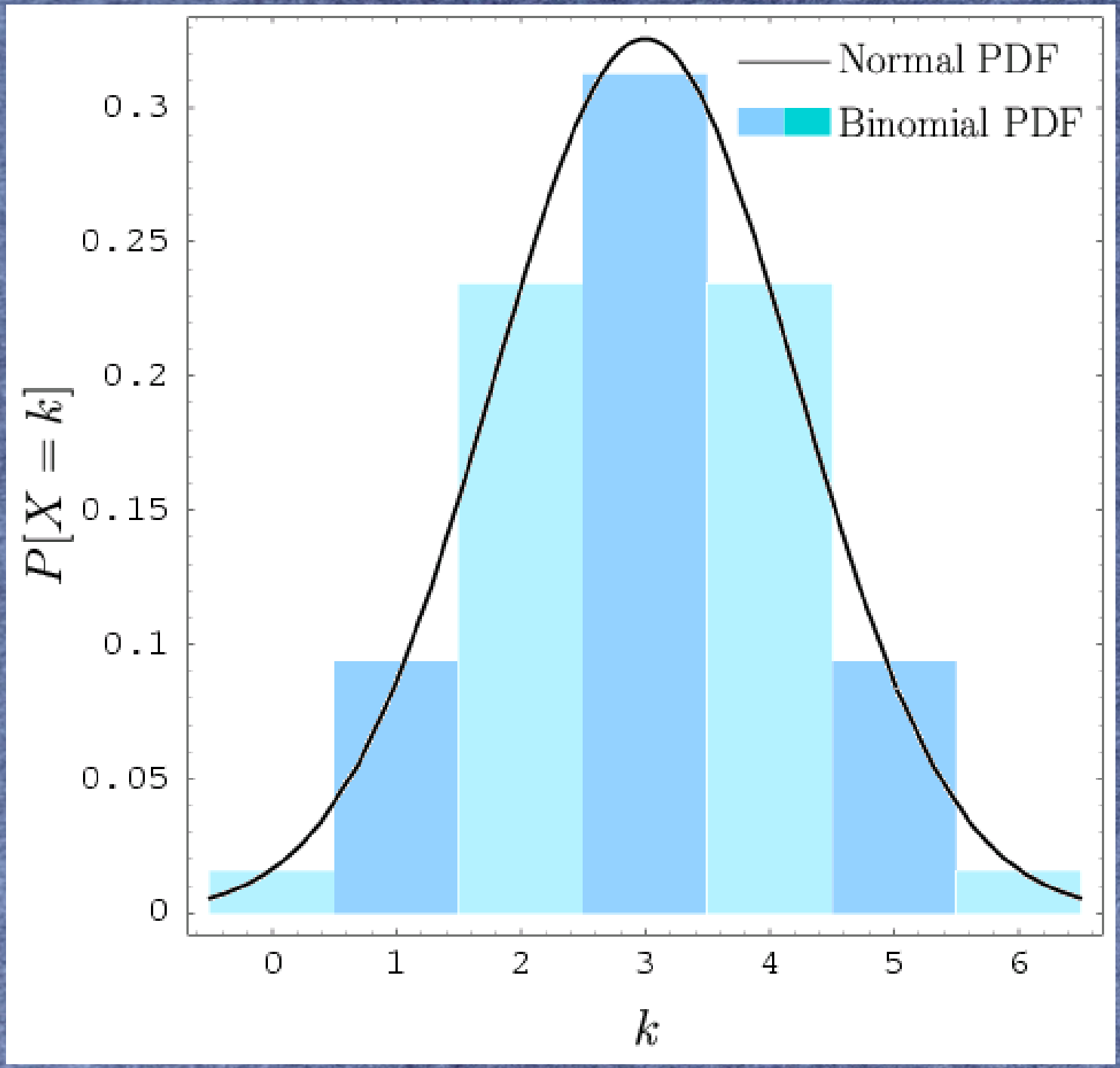
- If X denotes the number of successes that occur in n trials, then X is said to be a binomial random variable with parameters (n, p).

If X is a binomial random variable with parameters (n, p), then the p.m.f of X is given by

$$p(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & k = 0, 1, 2, \ldots, n, \\ 0, & \text{else.} \end{cases}$$

By binomial theorem,

$$\sum_{k=0}^{\infty} p(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1$$


Normal PDF
Binomial PDF
0.3
0.25
0.2
0.15
0.1
0.05
0
$P[X = k]$
0
1
2
3
4
5
6
k

- Example: Five fair coins are flipped. If the outcomes are assumed independent, find the probability mass function of the number of heads obtained.
- X be the number of heads (successes) that appear.
- X is a binomial random variable with parameters

$$(n = 5, p = 1/2)$$

- If you wish to calculate, say, $P\{X = 3\}$, just substitute 3 in place of k

Properties of Binomial Random Variables

For a binomial random variable X with parameters (n, p), we have

$$E[X] = np$$

$$\text{Var}(X) = np(1-p)$$

$$E[X^k] = npE[(Y+1)^{k-1}] \quad \text{where } Y \sim \text{Bin(n-1,p)}$$

IS IT BINOMIAL?
FOUR CONDITIONS TO CHECK.

(1) The trials are independent.

(2) The number of trials, n, is fixed.

(3) Each trial outcome can be classified as a success or failure.

(4) The probability of a success, p, is the same for each trial.

Arachnophobia. A Gallup Poll found that 7% of teenagers (ages 13 to 17) suffer from arachnophobia and are extremely afraid of spiders. At a summer camp there are 10 teenagers sleeping in each tent. Assume that these 10 teenagers are independent of each other.

(a) Calculate the probability that at least one of them suffers from arachnophobia.

(b) Calculate the probability that exactly 2 of them suffer from arachnophobia.

(c) Calculate the probability that at most 1 of them suffers from arachnophobia.

(d) If the camp counselor wants to make sure no more than 1 teenager in each tent is afraid of spiders, does it seem reasonable for him to randomly assign teenagers to tents?

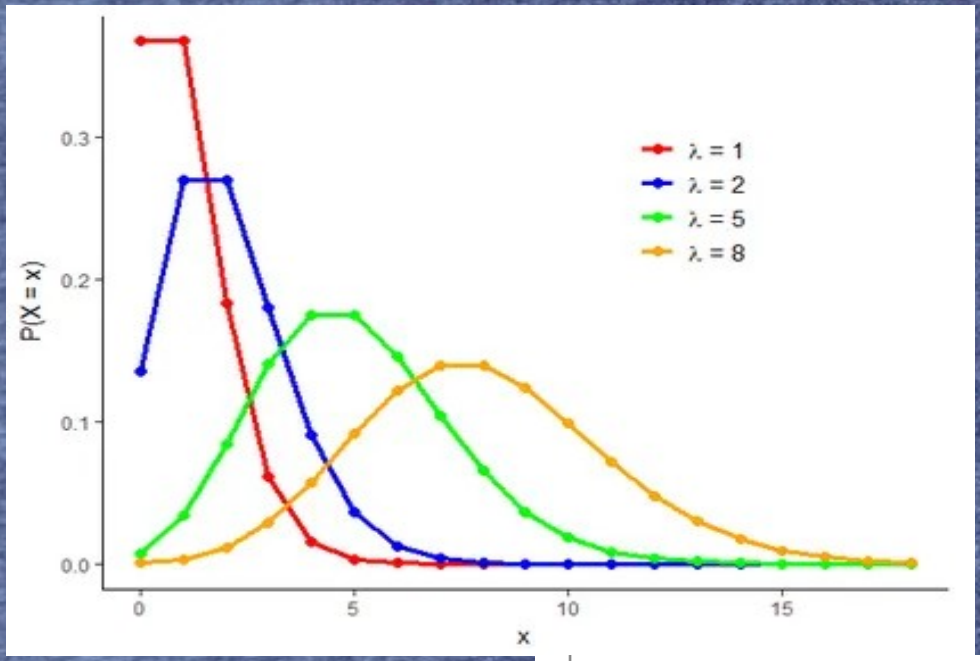
Poisson random variable

A random variable X that takes on one of the values $0, 1, 2, \ldots$ is said to be a **Poisson random variable** with parameter λ if, for some $\lambda > 0$,

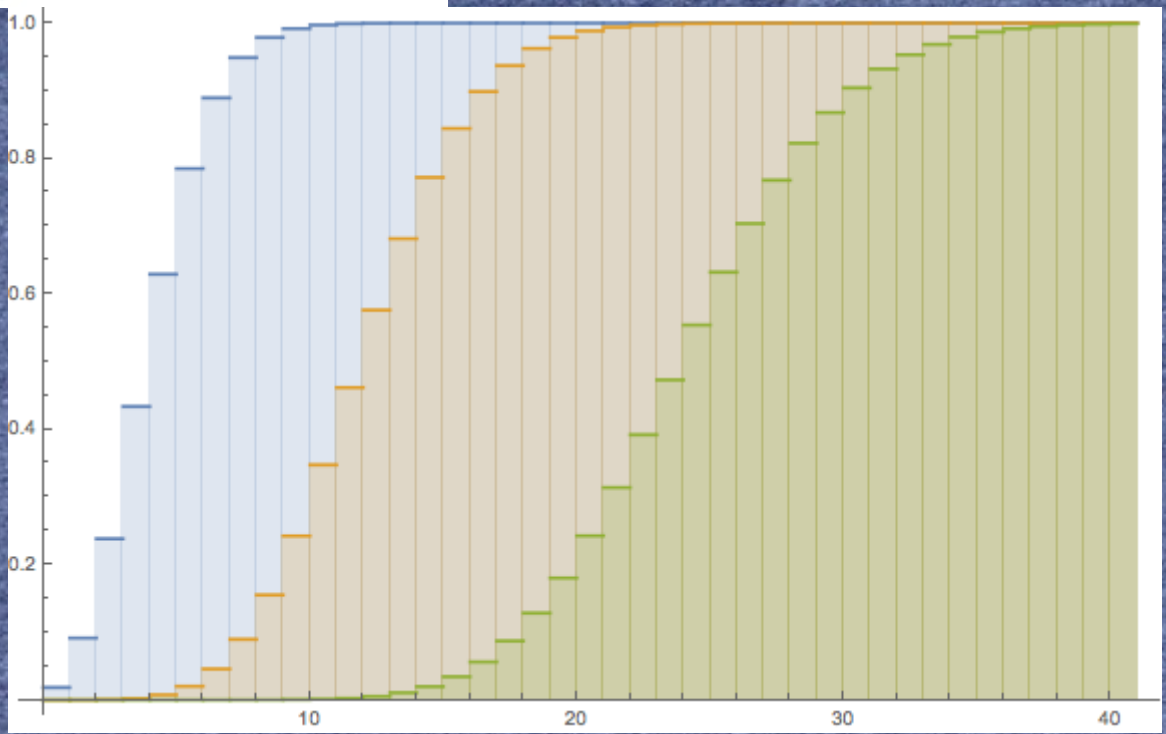
$$p(i) = P\{X = i\} = e^{-\lambda}\frac{\lambda^i}{i!}, \quad i = 0, 1, 2, \ldots$$

$$\sum_{i=0}^{\infty} p(i) = e^{-\lambda}\left(\sum_{i=0} \frac{\lambda^i}{i!}\right) = e^{-\lambda}e^{\lambda} = 1$$

p.m.f and c.d.f



$$F(x,\lambda) = \sum_{k=0}^{x} \frac{e^{-\lambda}\ \lambda^{x}}{k!}$$



Some examples of random variables which obey Poisson probability law

- The number of misprints on a page (or a group of pages) of a book
- The number of people in a community who survive to age 100.
- The number of packages of biscuits sold in a particular store each day
- The number of customers entering a post office on a given day .
- The number of vacancies occurring during a year in any of the government departments
- The number of α-particles discharged in a fixed period of time from some radioactive material

Observe that in each case the number of objects is very large and the probability is very small!

For a Poisson random variable X with parameter λ, we have

$$p(i) = \begin{cases} e^{-\lambda}\frac{\lambda^i}{i!}, & i = 0, 1, 2, \ldots \\ 0, \text{ else.} \end{cases}$$

$$E[X] = \lambda$$

$$\text{Var}(X) = \lambda$$

A Poisson experiment is a statistical experiment that has the following properties:

- The experiment results in outcomes that can be classified as successes or failures
- The average number of successes (λ) that occurs in a
- specified region is known.
- The probability that a success will occur is proportional to the size of the region
  - **The probability that a success will occur in an extremely small region is virtually zero**

- Example: Vehicles pass through a junction on a busy road at an average rate of 300 per hour.

  (a) Find the probability that none passes in a given minute.

  (b) What is the expected number passing in two minutes?

  (c) Find the probability that this expected number actually pass through in a given two-minute period.

  Solution :

  X denote the number of vehicles that pass through the junction per minute

  X follows Poisson distribution

  The average number of cars per minute is $\lambda = 300/60 = 5$

Thus, $p(i) = P\{X = i\} = e^{-5}\frac{5^i}{i!}, \quad i = 0, 1, 2, \ldots$

(a) $P\{X = 0\} = e^{-5}$

(b) Per minute average is 5

$\Rightarrow$ per 2-minute average is 300/30= 10

Thus, if Y denoted the number of vehicles passing through the junction per 2 minutes, then Y is also a Poisson random variable with parameter 10.

$\Rightarrow E[Y] = 10$

(c) We have $P\{Y = i\} = e^{-10}\frac{10^i}{i!}, \quad i = 0, 1, 2, \ldots$

- We need to find $P\{Y = 10\}$

$$\Longrightarrow P\{Y = 10\} = e^{-10}\frac{10^{10}}{10!} \approx 0.12511$$

- END