21DS636-Statistical Modelling

Probability Theory

Special Discrete Distribution

Dr. Deepa Gupta, Professor
Human Language Technology Lab
Department of Computer Science & Engineering
ASE-Bangalore

Special Probability distribution of Random Variable

- Discrete Random Variable:
  - Bernoulli distribution
  - Binomial Distribution
  - Multinominal Distribution
  - Poisson Distribution
  - …..etc.

- Continuous Random Variable:
  - Uniform Distribution
  - Exponential Distribution
  - Normal Distribution/Gaussian Distribution
  - ……..etc.

Bernoulli distribution

X: flip a coin, success is head

	T	H
X:	0	1
P(X=x)	1/2	1/2

- The Bernoulli distribution is the "coin flip" distribution.

- X is Bernoulli if its probability function is:

with prob

$$X = \begin{cases} 1 & w.p. \quad p \\ 0 & w.p. \quad 1-p \end{cases}$$

Abbreviation: Ber(*p*).

X	0	1
P(X=x)	1-p	p

p = ?

Ber(p)

X=**1 is usually interpreted as a "success."** E.g.:

X=1 for heads in coin toss
X=1 for male in survey
X=1 for defective in a test of product
X=1 for "made the sale" tracking performance

Bernoulli distribution

- Expectation:

$$E(X) = p(1) + (1-p)(0) = p$$

- Variance:

$$V(X) = E(X^2) - (E(X))^2$$
$$= p(1)^2 + (1-p)(0)^2 - (p)^2$$
$$= p - p^2 = p(1-p) = pq$$

$\downarrow q$

Binomial Setting/Binomial Experiment

i.e

Z - Binomial RV

The binomial distribution is just *n* independent Bernoulli i.e. added up.

- A binomial experiment has the following properties:
  1. experiment consists of n identical and independent trials i.e.
     - Fixed number *n* of observations
     - The *n* observations are independent

$Z = X_1 + X_2 + \dots + X_n$

Ber(p_1) Ber(p_2) Ber(p_n)

  2. Each observation falls into one of just two categories i.e. may be labeled "success" and "failure"
     - P(success) = p
     - P(failure) = q = 1 - p for all trials
     - The probability of success, *p*, is the same for each observation/trial

q

- The random variable of interest, X= The number of successes in the n trials.
- **X has a binomial distribution with parameters n and p i.e. Bin(*n*,*p*)**

prob of success

no. of trials fixed

Binomial Setting : Success(Example)

- In a shipment of 100 televisions, how many are defective?
  - counting the number of "successes" (defective televisions) out of 100

X : no. of defective televisions X: 0 1 2 3 4 5 ... 100

- A new procedure for treating breast cancer is tried on 25 patients; how many patients are cured?
  - counting the number of "successes" (cured patients) out of 25

Exmples

- A coin is flipped 10 times.
  - Success = head.
  - X = no. of heads in 10 trials n = 10 p = P(success) = $\frac{1}{2}$

    X = 0 1 2 3 - - - 10

- Twelve pregnant women selected at random, take a home pregnancy test.
  - Success = test says pregnant.
  - X = no. of women, taken home pregnancy test came +ve n = 12 p = 1/2

    = 0 1 2 - - - - 12

- Random guessing on a multiple choice exam. 25 questions. 4 answers per question.
  - Success = right answer.
  - X = no. of right answers in 25 questions n = 25 p = 1/4 q = 1 - p = 1 - 1/4 = 3/4

    X = 0 1 2 3 4 5 - 25

Binomial Distribution

- Let X = the count of successes in a binomial setting.
- The distribution of X is the ***binomial distribution*** **(pmf)** with parameters n and p.
  - ***n*** **is the number of observations/~~trails~~** trials
  - ***p*** **is the probability of a success on any one observation/~~trail~~,** trials **and** it is **fixed.**

0 1 2 - - n

  - ***X*** **takes on whole values between 0 and** ***n***

Binomial Distribution

→ 3 times tossing a coin

- Let's figure out a binomial r.v.'s probability function. Suppose we are looking at a binomial with n=3.
- We want P(X=0):
  - Can happen one way: 000 TTT
  - (1-p)(1-p)(1-p) = $(1-p)^3$

$P(TTT) = P(T)\,P(T)\,P(T) = (1-p)(1-p)(1-p)$

success — Head

X = 0 1 2 3

0 → TTT 000

- We want P(X=1):
  - Can happen three ways: 100, 010, 001 HTT THT TTH
  - p(1-p)(1-p)+(1-p)p(1-p)+(1-p)(1-p)p = $3p(1-p)^2$

$P(1) = p$

$P(0) = 1-p$

- We want P(X=2):
  - Can happen three ways: 110, 011, 101 HHT THH HTH
  - pp(1-p)+(1-p)pp+p(1-p)p = $3p^2(1-p)$
- We want P(X=3):
  - Can happen one way: 111 HHH
  - ppp = p^3

X	0	1	2	3
P(X=x)	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

Binomial distribution

- So, binomial r.v.'s probability function

with Prob.

$$X = \begin{cases} 0 & w.p. \quad (1-p)^3 \\ 1 & w.p. \quad 3p(1-p)^2 \\ 2 & w.p. \quad 3p^2(1-p) \\ 3 & w.p. \quad p^3 \end{cases}$$

$^3C_1 \, p^1 \, (1-p)^{3-1}$

$= 3$

$$P_X(x) = (\#\text{ of ways})\, p^x (1-p)^{n-x}$$
$$= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

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

$p(X=0) = {}^3C_0 \, p^0 \, (1-p)^{3-0}$

$= \frac{3!}{0!\,3!} = 1 \, (1-p)^3$

$1 = 0!$

What is P(x) for Binomial?

x = no. of success in n trials

x = 0, 1, 2, --- n

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

^nC_x

P(X = x)

p = prob. of success

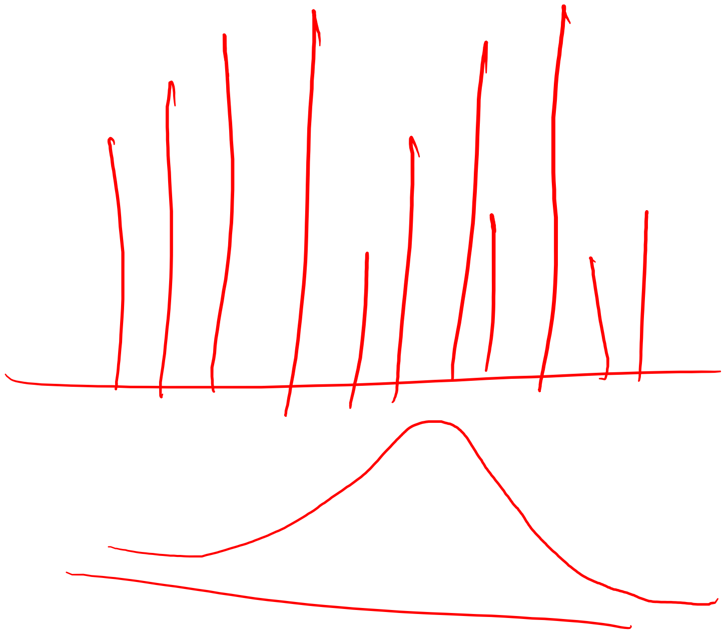
q = prob of failure

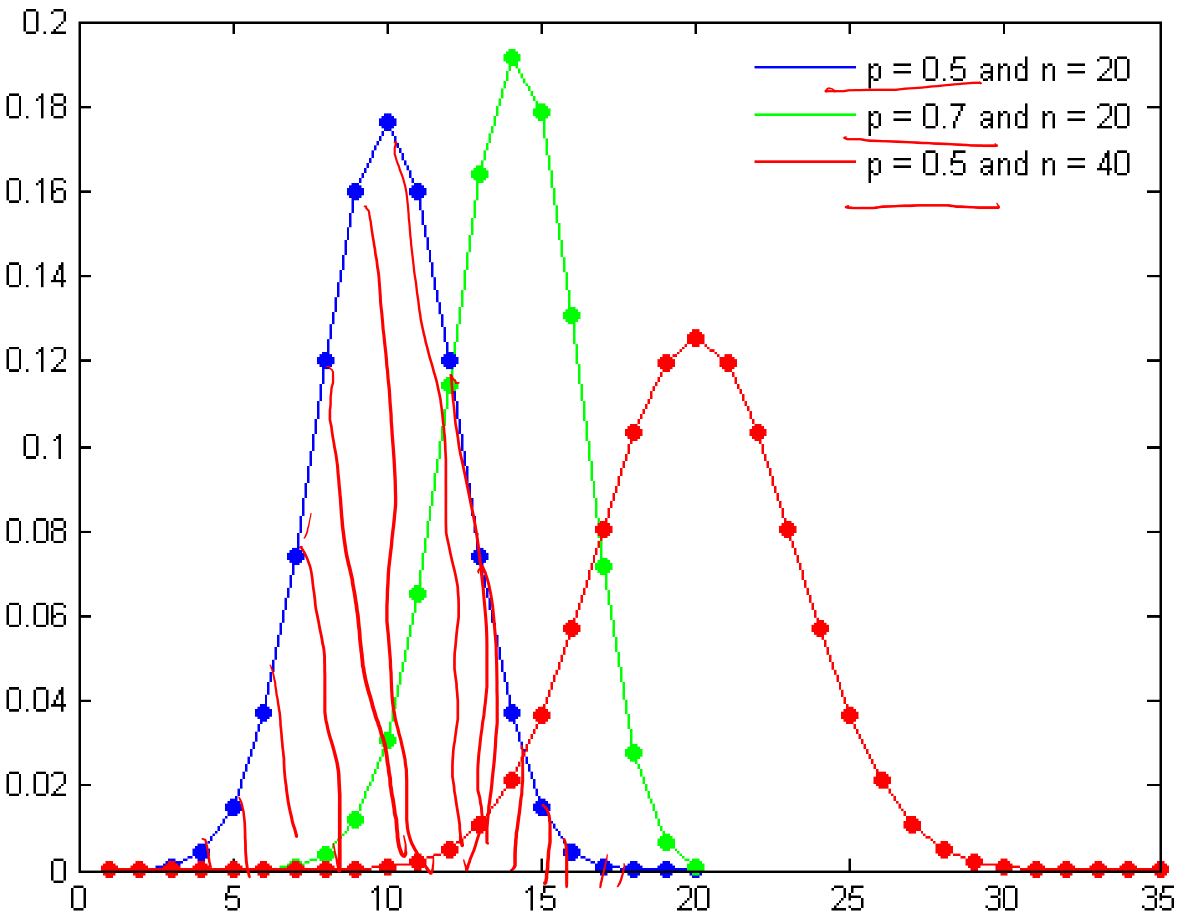
~~x = 0, 1, 2~~

x = 0, 1, 2, 3, --- n

Binomial distribution

- Typical shape of binomial:
  - Symmetric





AMRITA VISHWA VIDYAPEETHAM

$Z \sim$

Mean and Standard Deviation of Bin(*n*,*p*)

$Z = X_1 + X_2 + \cdots X_n$

- The mean (expected value) of a binomial random variable is

$X_i \sim Bern(p)$

$i = 1, 2 \ldots n$

$$\boxed{\mu = np}$$

$E[X_i] = p$

$V(X_i) = pq$

$i = 1, 2 \ldots n$

- The standard deviation of a binomial random variable is

$$\boxed{\sigma = \sqrt{npq}}$$

$V(Z) = V(X_1 + \cdots X_n)$

$= V(X_1) + V(X_2) + \cdots V(X_n)$

$= \underbrace{pq + pq + \cdots pq}_{n \text{ times}} = npq$

$E[Z] = E[X_1 + X_2 + \cdots X_n]$

$= E[X_1] + E[X_2] + \cdots E[X_n]$

$= \underbrace{p + p + \cdots p}_{n \text{ times}} = np$

Examples:

$P(X=x) = {}^nC_x\, p^x (1-p)^{n-x}$, $x = 0, 1, 2, \ldots$ $X \sim Bin(n, p)$

X = no. of success in n trials, p = prob. of success

1. The probability that you will win a certain game is 0.3. If you play the game 20 times, what is the probability that you will win at least 8 times?

X: no. of ~~games~~ win in 20 games

$X: 0, 1, 2, \ldots 20$

$p = 0.3 \quad q = 0.7 \quad X \sim Bin(20, 0.3)$

$P(X \geq 8) = 1 - P(X < 8) = 1 - \boxed{P(X=0) + P(X=1) + \ldots P(X=7)}$

$1 - \left[{}^{20}C_0 (0.3)^0 (1-0.3)^{20-0} + {}^{20}C_1 (0.3)^1\, 0.7^{20-1} + {}^{20}C_2 (0.3)^2\, 0.7^{18} + {}^{20}C_3 (0.3)^3 (0.7)^{17} + \ldots \right]$

2. The probability that you will win a certain game is 0.3. You play the game 20 times. What is the expected win?

$E[X] = np = 20 \times 0.3 = 6$

$Var(X) = npq = 20 \times 0.3 \times 0.7 = 6 \times 0.7 = 4.2$

$SD(X) = \sigma = \sqrt{npq} = \sqrt{4.2} = \square$

$\sqrt{Var X}$

30×0.4

3. A biased coin has a .6 chance of coming up heads. You flip it 50 times. What is the expected no. of heads and variance of this distribution?

X: no. of heads in 50 trials

$X: 0, 1, 2, \ldots 50$

$X \sim Bin(50, 0.6)$

$E[X] = np = 50 \times 0.6 = 30$

$V(X) = npq = 50 \times 0.6 \times 0.4 = 12$

Example:

- Suppose that 80% of adults with allergies report symptomatic relief with a specific medication. If the medication is given to 10 new patients with allergies, what is the probability that it is effective in exactly seven?

X: no. of adults get relief from allergies with a specific medication out of 10

X: 0 1 2 3 — — — 10

P/X

$p = 0.8 \quad q = 0.2$

$X \sim Bin(10, 0.8)$

$$P(X=7) = {}^{10}C_7 (0.8)^7 (1-0.8)^{10-7}$$

$$P(X=7) = {}^{10}C_7 (0.8)^7 (0.2)^3 = 0.203$$

$=$ $\boxed{}$

- Interpretation: There is a 20.13% probability that exactly 7 of 10 patients will report relief from symptoms when the probability that anyone reports relief is 80%

Multinomial Distribution

- The binomial distribution allows one to compute the probability of obtaining a given number of binary outcomes.
  - For example, it can be used to compute the probability of getting 6 heads out of 10coin flips.
  - The flip of a coin is a binary outcome because it has only two possible outcomes: heads and tails.
- The multinomial distribution can be used to compute the probabilities in situations in which there are more than two possible outcomes.
  - For example, suppose that two chess players had played numerous games and it was determined that the probability that Player A would win is 0.40, the probability that Player B would win is 0.35, and the probability that the game would end in a draw is 0.25.
  - The multinomial distribution can be used to answer questions such as:
    - "If these two chess players played 12 games, what is the probability that Player A would win 7 games, Player B would win 2 games, and the remaining 3 games would be drawn?"

Multinomial Experiment

- A multinomial experiment is a statistical experiment that has the following properties:
  - The experiment consists of *n* repeated trials.
  - Each trial has a discrete number of possible outcomes.
  - On any given trial, the probability that a particular outcome will occur is constant.
  - The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

Multinomial Distribution

- Statistical experiment with k outcomes
- Repeated independently *n* times
- Probability(Outcome *j*) = p_j, *j = 1, …, k*
- Number of times outcome *j* occurred is x_j, *j = 1, …, k*
- A **multivariate** distribution

$$p(x_1, \ldots, x_k) = \binom{n}{x_1 \; \cdots \; x_k} p_1^{x_1} \cdots p_k^{x_k}$$

Multinomial Distribution

$$p(x_1, \ldots, x_k) = \binom{n}{x_1 \; \cdots \; x_k} p_1^{x_1} \cdots p_k^{x_k}$$

$$x_j = 0, 1, \ldots, n$$

$$\sum_{j=1}^{k} x_j = n \qquad \sum_{j=1}^{k} p_j = 1 \qquad \binom{n}{n_1 \ldots n_k} = \frac{n!}{n_1! \ldots n_k!}$$

Denote by $\mathbf{x} \sim M(n, \mathbf{p})$

Example 1:

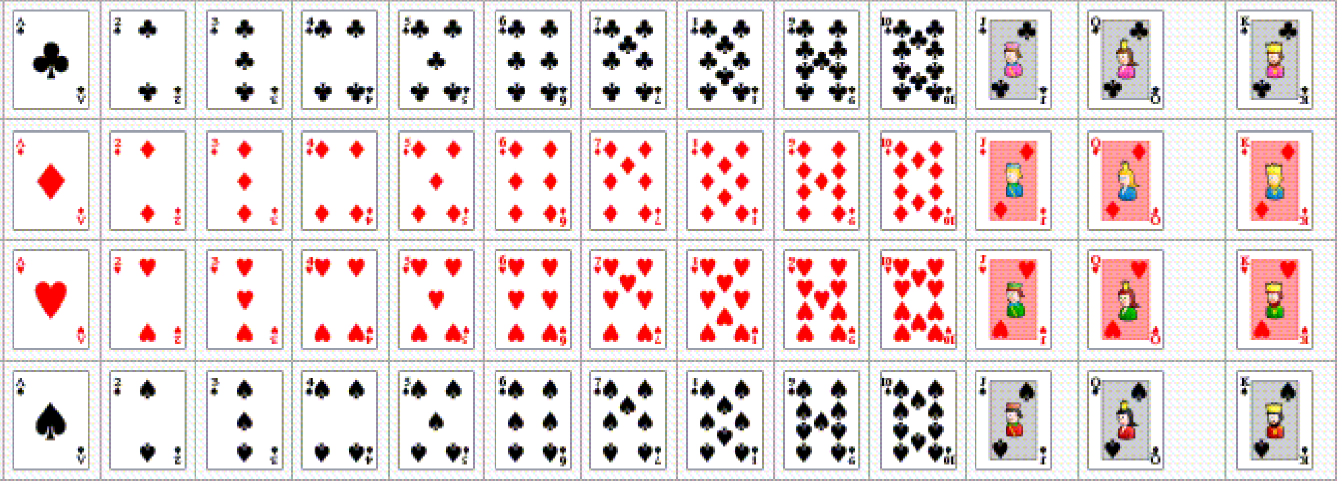
- Suppose we have a bowl with 10 marbles - 2 red marbles, 3 green marbles, and 5 blue marbles. We randomly select 4 marbles from the bowl, with replacement. What is the probability of selecting 2 green marbles and 2 blue marbles?

Example 2:

- Suppose a card is drawn randomly from an ordinary deck of playing cards, and then put back in the deck. This exercise is repeated five times. What is the probability of drawing 1 spade, 1 heart, 1 diamond, and 2 clubs?

Example set of 52 poker playing cards

Suit	Ace	2	3	4	5	6	7	8	9	10	Jack	Queen	King
Clubs													
Diamonds													
Hearts													
Spades													



Example3:

Suppose that the racial/ethnic distribution in a large city is given by the table that follows. Consider these three options as the parameters of a multinomial distribution.

Black	Hispanic	Other
20%	15%	65%

Suppose that a jury of twelve members is chosen from this city in such a way that each resident has an equal probability of being selected independently of every other resident. There are a number of questions that we can ask of this type of distribution.

Let's find probability that the jury contains:

- three Black, two Hispanic, and seven Other members;
- four Black and eight Other members;
- at most one Black member.

To solve this problem, let $X = (X_1, X_2, X_3)$ where X_1 = number of Black members, X_2 = number of Hispanic members, and X_3 = number of Other members. Then X has a multinomial distribution with parameters $n = 12$ and $\pi = (.20, .15, .65)$. The answer to the first part is:

$$\begin{aligned} P(X_1 = 3, X_2 = 2, X_3 = 7) &= \frac{n!}{x_1!x_2!x_3!}\pi_1^{x_1}\pi_2^{x_2}\pi_3^{x_3} \\ &= \frac{12!}{3!2!7!}(0.20)^3(0.15)^2(0.65)^7 \\ &= 0.0699 \end{aligned}$$

The answer to the second part is:

$$\begin{aligned} P(X_1 = 4, X_2 = 0, X_3 = 8) &= \frac{12!}{4!0!8!}(0.20)^4(0.15)^0(0.65)^8 \\ &= 0.0252 \end{aligned}$$

For the last part, note that "at most one Black member" means $X_1 = 0$ or $X_1 = 1$. X_1 is a binomial random variable with $n = 12$ and $p = \pi_1 = .2$. Using the binomial probability distribution,

$$\begin{aligned} P(X_1 = 0) &= \frac{12!}{0!12!}(0.20)^0(0.8)^{12} \\ &= 0.0687 \end{aligned}$$

and

$$\begin{aligned} P(X_1 = 1) &= \frac{12!}{1!11!}(0.20)^1(0.8)^{11} \\ &= 0.2061 \end{aligned}$$

Therefore, the answer is:

$$P(X_1 = 0) + P(X_1 = 1) = 0.0687 + 0.2061 = 0.2748.$$

Poisson Distribution

- Many experimental situation occur in which we observe the counts of random events/rare events within a set unit of time, area, volume, length etc. For example,

    1. The number of births per hour during a given day.
    2. how many times a person becomes unemployed in a given year.
    3. The number of trades that a typical investor will make in a given day, which can be 0 (often), or 1, or 2, etc.

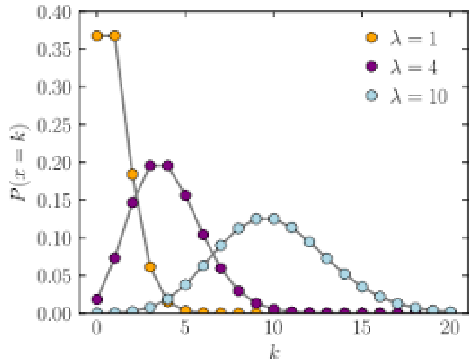
- In such situations, we are often interested in whether the events occur randomly in time or space, or not.

Poisson Distribution

- The Poisson distribution is used to model the number of random events occurring within a given **time interval**. i.e.
- A Poisson distribution can be used to estimate how likely it is that something will happen "X" number of times.
- The formula for the Poisson probability density (mass) function is
- $X \sim Pos(\lambda)$,

$$p(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

x=0,1,2,3,4,……..



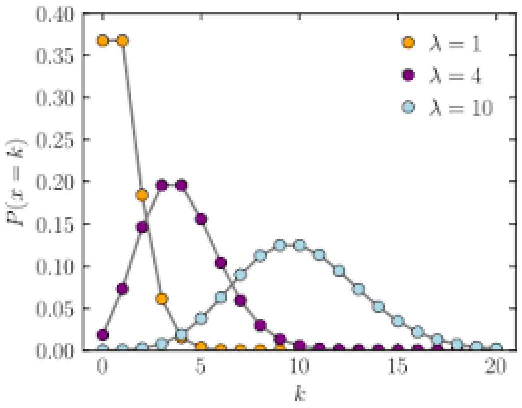
- λ is the shape parameter which indicates the average number of events in the given time interval.

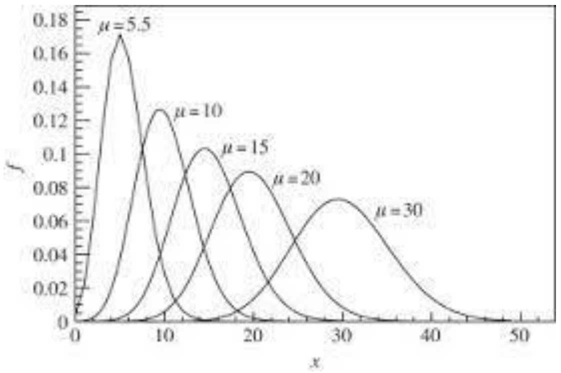
Poisson Distribution

- The Distribution arises when the events being counted, when events
    1. Occur independently such that the probability that two or more events occur simultaneously is zero, randomly in time;
    2. Uniformly (that is, the mean number of events in an interval is directly propositional to the length of the interval)

A random variable X has the Poisson distribution with parameter $\lambda(>0)$ if

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





Example 6: Poisson or not? : Which of the following is most likely to be well modelled by a Poisson distribution?

1. Number of trains arriving at Yeshwanthpur Railway station every hour
2. Number of lottery winners each year that live in Bangalore
3. Number of days between solar eclipses
4. Number of days until a component fails

Are they Poisson? Answers 6:

1. Number of trains arriving at Yeshwanthpur Railway every hour

   NO, (supposed to) arrive *regularly* on a timetable not at random

2. Number of lottery winners each year that live in Bangalore

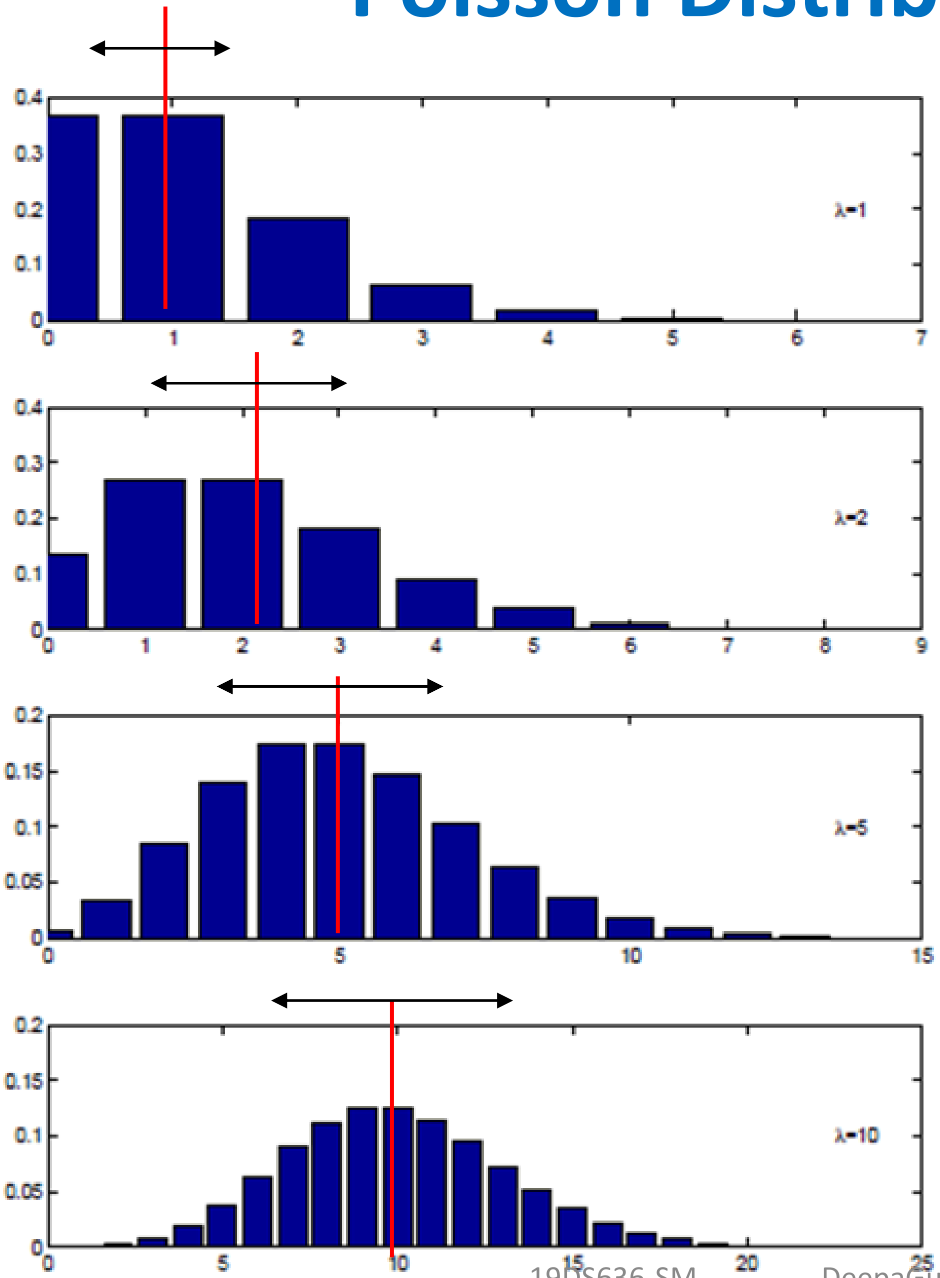
   Yes, is number of random events in fixed interval

3. Number of days between solar eclipses

   NO, solar eclipses are not random events and this is a time *between* random events, not the number in some fixed interval

4. Number of days until a component fails
   NO, random events, but this is time until a random event, not the *number* of random events

Poisson Distribution



Mean and variance

If $X \sim$ Poisson with mean λ, then i.e. *Pos*(λ)

$$p(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

X=0,1,2,3,4,

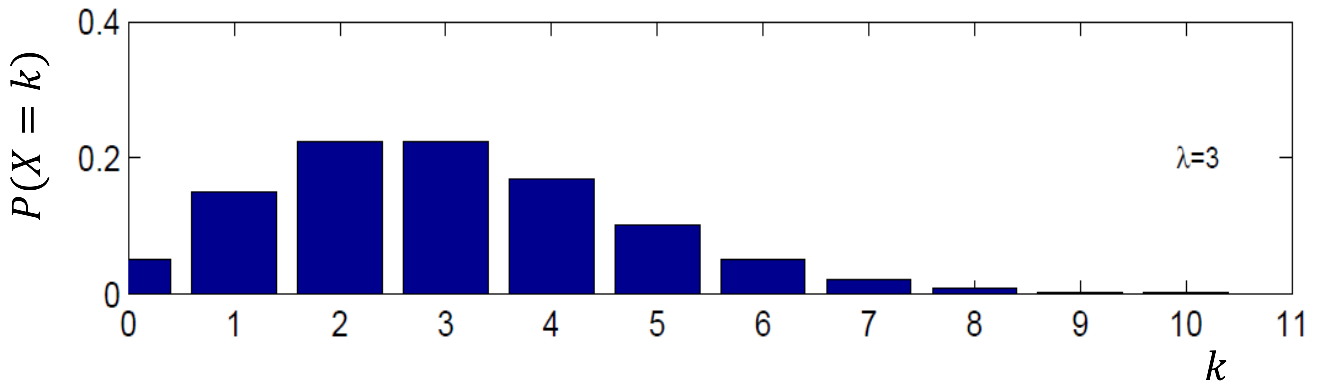
$$\mu = E(X) = \lambda$$
$$\sigma^2 = \mathrm{var}(X) = \lambda$$

We observe that the Poisson distributions

1. are unimodal;
2. exhibit positive skew (that decreases as λ increases);
3. are centered roughly on λ;
4. have variance (spread) that increases as λ increases

Example 1: On average lightning kills three people each year in the UK, $\lambda = 3$. What is the probability that only one person is killed this year?

Answer:
Assuming these are independent random events, the number of people killed in a given year therefore has a Poisson distribution:



Let the random variable X be the number of people killed in a year.

Poisson distribution $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ with $\lambda = 3$

$$\Rightarrow P(X = 1) = \frac{e^{-3}3^1}{1!} \approx 0.15$$

Changing the size of the interval

- Rule is as follows:
  - If X~Pos(λ) on 1 unit interval
  - then Y~Pos($k\lambda$) on k unit intervals.

Sum of two Poisson variables

Now suppose we know that

- in hospital A births occur randomly at an average rate of 2.3 births per hour
- in hospital B births occur randomly at an average rate of 3.1 births per hour

What is the probability that we observe 7 births in total from the two hospitals in a given 1 hour period?

To answer this question we can use the following rule

$$\boxed{\begin{array}{ll} \text{If} & X \sim \text{Po}(\lambda_1) \text{ on 1 unit interval,} \\ \text{and} & Y \sim \text{Po}(\lambda_2) \text{ on 1 unit interval,} \\ \text{then} & X+Y \sim \text{Po}(\lambda_1+\lambda_2) \text{ on 1 unit interval.} \end{array}}$$

Example 2: Suppose that trucks arrive at a receiving dock with an average arrival rate of 3 per hour. What is the probability exactly 5 trucks will arrive in a two-hour period?

Sol//

X ~ no. of trucks arrive at ... in per hour

$X \sim poi(3)$ $\lambda = 3$

$X \sim Poi(2\times3=6)$ in 2 hrs.

$P(X=5)$ in 2-hrs $= \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-6}(6)^5}{5!} =$

Example 3: The average number of homes sold by the Acme Realty company is 2 homes per day. What is the probability that exactly 3 homes will be sold tomorrow?

Example 4: Suppose the average number of lions seen on a 1-day safari is 5. What is the probability that tourists will see fewer than four lions on the next 1-day safari?

$\frac{e^{-\lambda}\lambda^{x}}{x!}$

X ~ no. of lions seen on a 1 day safari

$X \sim Po(5)$

$P(X<4) = P(X=0)+P(X=1)+P(X=2)+P(X=3)$

$=$

Example5: Telecommunications

Messages arrive at a switching centre at random and at an average rate of 1.2 per second.

a) Find the probability of 5 messages arriving in a 2-sec interval.

b) For how long can the operation of the centre be interrupted, if the probability of losing one or more messages is to be no more than 0.05?

$X \sim Po(1.2)$ per second

Sol //

$X \sim Po(1.2 \times 2 = 2.4)$ 2-sec interval

$P(X=5) = \frac{e^{-2.4}(2.4)^5}{5!}$

Sol //

$\boxed{P(X \geq 1) \leq 0.05}$

X : no. of messages in t seconds

Let t be the required time

$[1 - P(X<1)] \leq 0.05$

$[1 - P(X=0)] \leq 0.05$

$P(X=0) \geq 0.95$

$e^{-1.2t}(1.2t)^0 \geq 0.95$

$\lambda = 1.2t$, $e^{-1.2t} \geq 0.95$

$t \leq 0.043$ sec

- **Example 7:** *5. If $X \sim Po(2)$, $Y \sim Po(3)$ and $Z \sim Po(5)$, find:

  (a) $P(X+Y=0)$ (b) $P(X+Y=1)$ (c) $P(Z=0)$

  (d) $P(Z=1)$ (e) $P(X+Y \le 2)$ (f) $P(Z \le 2)$

A random variable X has the Poisson distribution with parameter $\lambda(>0)$ if

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

$X \sim Po(2)$

$Y \sim Po(3)$

$X+Y \sim Po(2+3=5)$

$P(X+Y=0) = \frac{e^{-5}(5)^0}{0!} = e^{-5}$

$P(X+Y \le 2) = P(X+Y=0) + P(X+Y=1) + P(X+Y=2)$

Approximation to the Binomial distribution

The Poisson distribution is an approximation to B(*n, p*), **when *n* is large and *p* is small** (e.g. if $np < 7$, say).

In that case, if $X \sim B(n,p)$ then $P(X = k) \approx \frac{e^{-\lambda}\lambda^k}{k!}$ Where $\boxed{\lambda = np}$

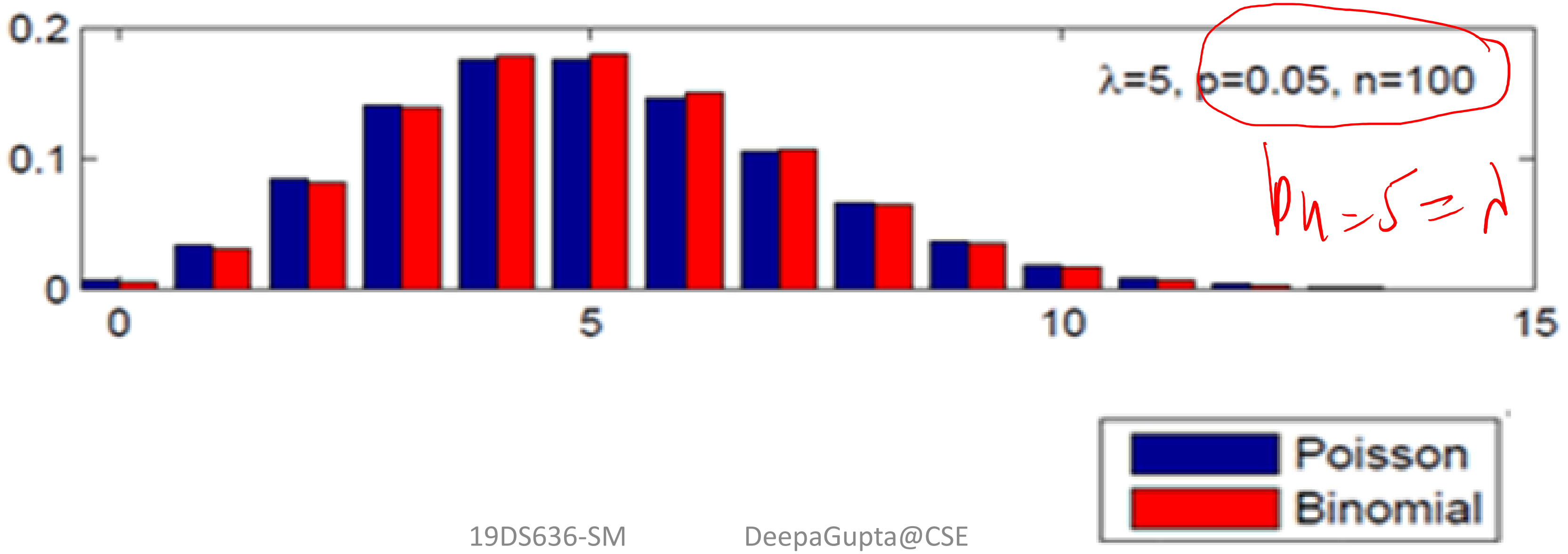
i.e. X is approximately Poisson, with mean $\lambda = np$.

$X \sim Po(\lambda)$

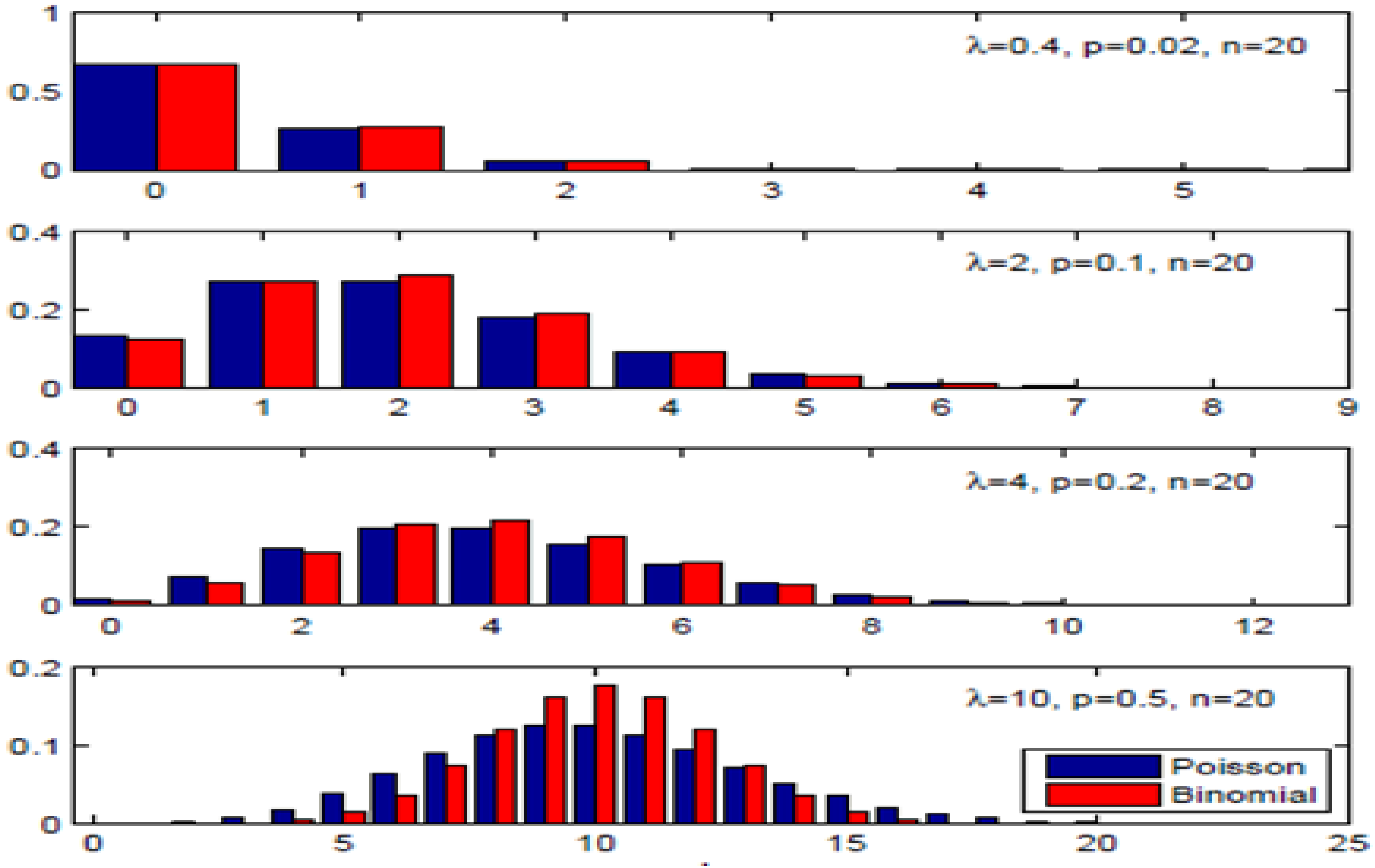
$E[X] = \lambda$

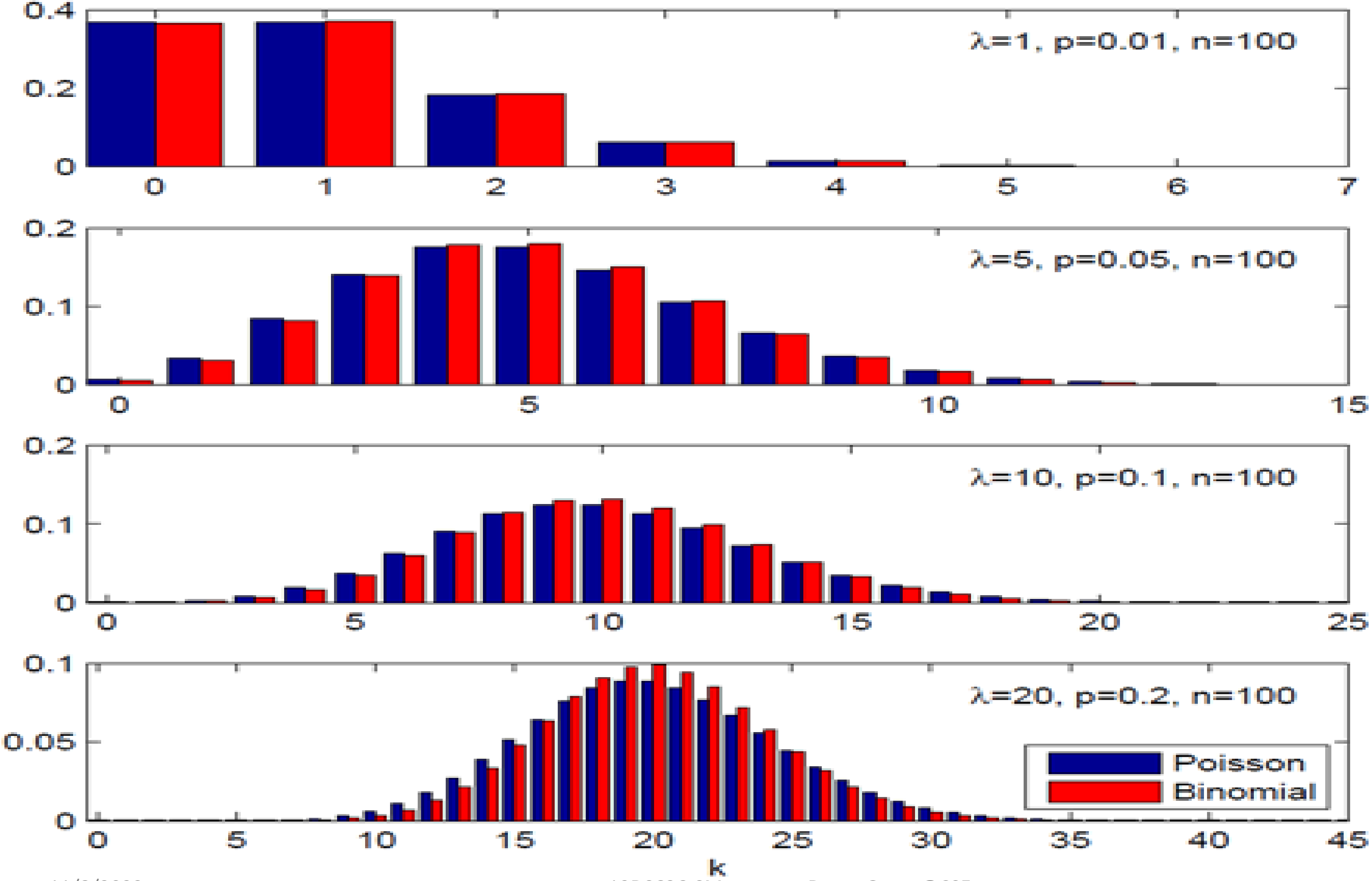
$X \sim Bin(n,p)$

$E[X] = np$



$pn = 5 = \lambda$


λ=0.4, p=0.02, n=20
λ=2, p=0.1, n=20
λ=4, p=0.2, n=20
λ=10, p=0.5, n=20
Poisson
Binomial


λ=1, p=0.01, n=100
λ=5, p=0.05, n=100
λ=10, p=0.1, n=100
λ=20, p=0.2, n=100
Poisson
Binomial
k

Example 7 : The probability of a certain part failing within ten years is 10^{-6}. Five million of the parts have been sold so far. What is the probability that three or more will fail within ten years?

Answer:

Let X = number failing in ten years, out of 5,000,000; $X \sim B(5000000, 10^{-6})$

Evaluating the Binomial probabilities is rather awkward; better to use the Poisson approximation.

X has approximately Poisson distribution with $\lambda = np = 5000000 \times 10^{-6} = 5$.

$$\text{P(Three or more fail)} = P(X \geq 3) = 1 - P(X=0) - P(X=1) - P(X=2)$$

$$= 1 - \frac{e^{-5}5^0}{0!} - \frac{e^{-5}5^1}{1!} - \frac{e^{-5}5^2}{2!}$$

$$= 1 - e^{-5}(1 + 5 + 12.5) = 0.875$$

For such small p and large n the Poisson approximation is very accurate (exact result is also 0.875 to three significant figures).