STATISTICS FOR DATA SCIENCE

Poisson Distribution

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering

STATISTICS FOR DATA SCIENCE

Poisson Distribution

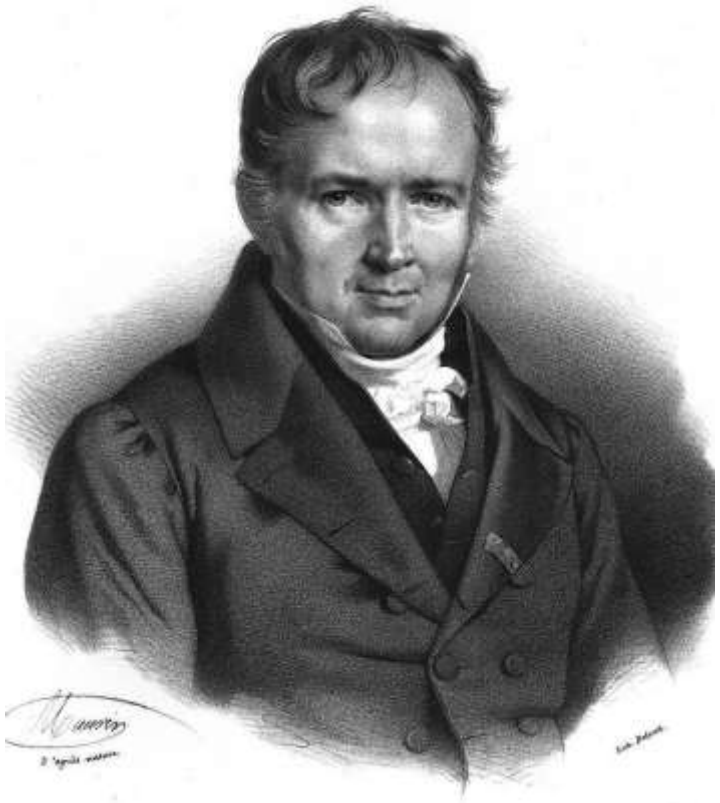
Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

STATISTICS FOR DATA SCIENCE

Topics to be covered...

- Poisson Distribution
- Probability Mass Function
- Students t-distribution
- Mean and Variance of Poisson Distribution
- Using the Poisson Distribution to Estimate a Rate
- Computing uncertainty of λ ^

Poisson Distribution



Siméon Denis Poisson
(1781–1840)

First derived Poisson distribution in 1837

STATISTICS FOR DATA SCIENCE

Poisson Distribution

A **Poisson distribution** is the probability distribution that results from a **Poisson experiment**.

Attributes of a Poisson Experiment:

1. The experiment results in **outcomes** that can be classified as **successes** or **failures**.
2. The average number of successes(**λ**) that occurs in a region(length, area, volume, period of time) is known.
3. The **probability** that a **success** will occur is **proportional** to the **size of the region**.
4. The probability that a success will occur in an **extremely small region** is virtually **zero**.

Source: www.stattrek.com

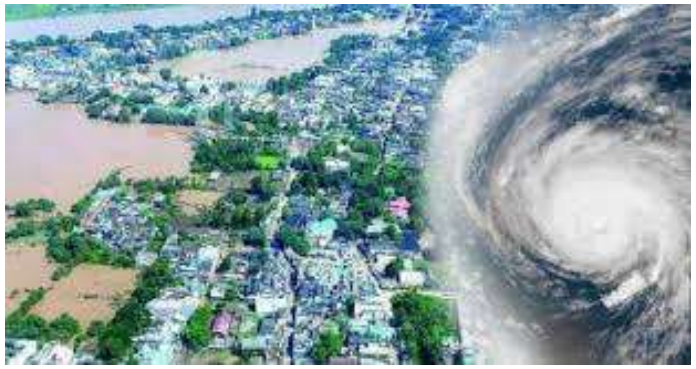
STATISTICS FOR DATA SCIENCE

Poisson Distribution

- Poisson distribution is used to describe number of occurrences of a (**rare)** event that occur randomly during a specified interval.
- The interval may be time, distance, area, or volume.
- It describes the frequency of “successes” in a test where a “success” is a rare event.
- Events with low frequency in a large population follow a Poisson distribution

Examples

- The number of deaths by horse kicking in the Prussian army (First application).
- The number of cyclones in a season.
- Arrival of Telephone calls, Customers, Traffic, Web requests.
- Estimating the number of mutations of DNA after exposure to radiation.
- Rare diseases (like Leukemia(cancer of the blood cells), but not AIDS because it is infectious and so not independent).



Examples

- The number of calls coming per minute into a hotels reservation Center.
- The number of particles emitted by a radioactive source in a given time.
- The number of births per hour during a given day.
- The number of patients arriving in an emergency room between 11 -12 pm.
- The number of car accidents in a day.

In such situations we are often interested in whether the events occur randomly in time or space.

Law of small numbers by von Bortkiewicz(1898)

Showed how the Poisson distribution could be used to explain statistical regularities in the occurrence of rare events.

As examples von Bortkiewicz considered:

1) The number of suicides of children in different years
2) The number of suicides of women in different states and years
3) The number of accidental deaths in different years and
4) **The number of deaths from horse kicks in the Prussian Army in different years.**

Law of small numbers by Von Bortkiewicz(1898)

- The last example provided the most extensive data and has become a classical example of the Poisson distribution.

- He extracted from official records the number of deaths from horse kicks in 14 army corps over the 20 year period 1875-1894, obtaining 280 observations in all.

- He argued that the chance that a **particular soldier should be killed by a horse in a year was extremely small,** but the no of men in corps was very large, so that the no of deaths in a cavalry corps in a year should follow the Poisson distribution.

STATISTICS FOR DATA SCIENCE

Poisson Distribution

- One way to think of the Poisson distribution is as an approximation to the binomial distribution when n is large and p is small.

Example

- A mass contains 10,000 atoms of a radioactive substance. The probability that a given atom will decay in a one- minute time period is 0.0002. Let X represent the number of atoms that decay in one minute. Now each atom can be thought of as a Bernoulli trial, where success occurs if the atom decays.

STATISTICS FOR DATA SCIENCE

Poisson Distribution - Example

- Thus X is the number of successes in 10,000 independent Bernoulli trials, each with success probability 0.0002, so the distribution of
- X is Bin(10,000, 0.0002).
- The mean of X is $\mu X = (10,000)\ (0.0002) = 2$.

STATISTICS FOR DATA SCIENCE

Poisson Distribution - Example

- Another mass contains 5000 atoms, and each of these atoms has probability 0.0004 of decaying in a one-minute time interval.
- Let Y represent the number of atoms that decay in one minute from this mass. By the reasoning in the previous paragraph,
- $Y \sim Bin(5000, 0.0004)$ and
- $\mu Y = (5000)(0.0004) = 2$.

STATISTICS FOR DATA SCIENCE

Poisson Distribution - Example

- In each of these cases, the number of trials n and the success probability p are different, but the mean number of successes, which is equal to the product np, is the same.
- Compute the probability that exactly three atoms decay in one minute for each of these masses using the binomial probability mass function.

STATISTICS FOR DATA SCIENCE

Poisson Distribution - Example

$$P(X = 3) = \frac{10{,}000!}{3!\,9997!}(0.0002)^3(0.9998)^{9997} = 0.180465091$$

$$P(Y = 3) = \frac{5000!}{3!\,4997!}(0.0004)^3(0.9996)^{4997} = 0.180483143.$$

- When n is large and p is small the mass function depends on the mean np, and very little on the specific values of n and p.
- We can therefore approximate the binomial mass function with a quantity that depends on the product np only.
- Specifically, if n is large and p is small, and we let $\lambda = np$, it can be shown by advanced methods that for all x,

Text-Book-Chapter 4.3- Page.No:216

STATISTICS FOR DATA SCIENCE

Probability Mass Function

$$\frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \approx e^{-\lambda}\frac{\lambda^x}{x!} \tag{4.8}$$

We are led to define a new probability mass function, called the Poisson probability mass function. The Poisson probability mass function is defined by

$$p(x) = P(X = x) = \begin{cases} e^{-\lambda}\dfrac{\lambda^x}{x!} & \text{if } x \text{ is a non-negative integer} \\ 0 & \text{otherwise} \end{cases} \tag{4.9}$$

If X is a random variable whose probability mass function is given by Equation (4.9), then X is said to have the **Poisson distribution** with parameter λ. The notation is $X \sim \text{Poisson}(\lambda)$.

Text-Book-Chapter 4.3- Page.No:216

STATISTICS FOR DATA SCIENCE

Probability Mass Function - Example

If $X \sim \text{Poisson}(3)$, compute $P(X = 2)$, $P(X = 10)$, $P(X = 0)$, $P(X = -1)$, and $P(X = 0.5)$.

Solution

Using the probability mass function (4.9), with $\lambda = 3$, we obtain

$$P(X = 2) = e^{-3}\frac{3^2}{2!} = 0.2240$$

$$P(X = 10) = e^{-3}\frac{3^{10}}{10!} = 0.0008$$

$$P(X = 0) = e^{-3}\frac{3^0}{0!} = 0.0498$$

$P(X = -1) = 0$ because -1 is not a non-negative integer

$P(X = 0.5) = 0$ because 0.5 is not a non-negative integer

Probability Mass Function - Example

What is the probability that 8 or more accidents happen?

$$P(x \geq 8) = \mathbf{1 - P(x < 8)}$$
$$= \mathbf{1 - P(x \leq 7)}$$
$$= \mathbf{1 - .999 = .001}$$

k	$\mu = 2$
0	.135
1	.406
2	.677
3	.857
4	.947
5	.983
6	.995
7	.999
8	1.000

STATISTICS FOR DATA SCIENCE

Probability Mass Function - Example

If $X \sim \text{Poisson}(4)$, compute $P(X \leq 2)$ and $P(X > 1)$.

Solution

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= e^{-4}\frac{4^0}{0!} + e^{-4}\frac{4^1}{1!} + e^{-4}\frac{4^2}{2!} \\ &= 0.0183 + 0.0733 + 0.1465 \\ &= 0.2381 \end{aligned}$$

To find $P(X > 1)$, we might try to start by writing

$$P(X > 1) = P(X = 2) + P(X = 3) + \cdots$$

This leads to an infinite sum that is difficult to compute. Instead, we write

$$\begin{aligned} P(X > 1) &= 1 - P(X \leq 1) \\ &= 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - \left(e^{-4}\frac{4^0}{0!} + e^{-4}\frac{4^1}{1!}\right) \\ &= 1 - (0.0183 + 0.0733) \\ &= 0.9084 \end{aligned}$$

STATISTICS FOR DATA SCIENCE

Students t-Distribution

William Sealy Gosset (pen name Student)
(1876 – 1937)
English statistician
Famous for Student's t-distribution

STATISTICS FOR DATA SCIENCE

Students t-Distribution

- The first biological application of the Poisson distribution was given by 'Student' (1907) in his paper on the **error of counting yeast cells in a haemocytometer**(instrument for counting the no of cells in a cell suspension.), although he was unaware of the work of Poisson and von Bortkiewicz and derived the distribution afresh.

- W.S. Gosset who was employed by Messrs Guinness in Dublin, Ireland.

- Guinness is one of the most successful beer brands worldwide.

STATISTICS FOR DATA SCIENCE

Students t-Distribution

- A normal distribution describes a full population, t-distributions describe samples drawn from a full population;
- Accordingly, the t-distribution for each sample size is different, and the larger the sample, the more the distribution resembles a normal distribution.
- The t-distribution plays a role in a number of widely used statistical analyses, including Student's t-test for assessing the statistical significance of
  - the difference between two sample means,
  - the construction of confidence intervals for the difference between two population means, and
  - in linear regression analysis.

STATISTICS FOR DATA SCIENCE

Mean and Variance of Poisson Distribution

Has a single parameter (mean of the distribution)
Theoretical range of the random variable is zero to infinity

Mean and Variance

$$\mu_X = \lambda \qquad \sigma_X^2 = \lambda$$

In Poisson distribution, mean = variance = λ. This is the **acid test** to be applied to any data which might appear to confirm to Poisson distribution

Examples

If electricity power failures occur according to a Poisson distribution with an average of 3 failures every twenty weeks, calculate the probability that there will not be more than one failure during a particular week.

Solution:

Average no of failures per 20 weeks = $\lambda = 3$

X denote the number of failures per week X ~Poisson(λt)

X ~ Poisson(3/ 20) => X ~ Poisson(0.15)

"Not more than one failure" means we need to include the probabilities for "0 failures" plus "1 failure".

$$P(x_0) + P(x_1) = \frac{e^{-0.15}0.15^0}{0!} + \frac{e^{-0.15}0.15^1}{1!} = 0.98981$$

Examples

A life insurance salesman sells on the average 3 life insurance policies per week. Use Poisson's law to calculate the probability that in a given week he will sell.

1) Some policies

2) 2 or more policies but less than 5 policies.

3) Assuming that there are 5 working days per week, what is the probability that in a given day he will sell one policy?

Solution

$\lambda = 3$

1)Some policies means P(X > 0)

P(X > 0) = 1 – P(X = 0)

$$P(x_0) = \frac{e^{-3}3^0}{0!} = 4.9787 \times 10^{-2}$$

2) 2 or more policies but less than 5 policies.

P(2 <= X < 5) = P(X =2) + P(X =3) + P(X = 4)

$$= P(x_2) + P(x_3) + P(x_4)$$

$$= \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!} + \frac{e^{-3}3^4}{4!}$$

$$= 0.61611$$

3) Assuming that there are 5 working days per week, what is the probability that in a given day he will sell one policy?

Average number of policies sold per day: $\frac{3}{5} = 0.6$

on a given day, $P(X) = \frac{e^{-0.6}(0.6)^1}{1!} = 0.32929$

Using the Poisson Distribution to Estimate a Rate

Let λ denote the mean number of events that occur in one unit of time or space.

Let X denote the number of events that are observed to occur in t units of time or space.

Then,

$$X \sim Poisson(\lambda t)$$

Where λ is estimated with $\hat{\lambda} = X / t$

Example

A microbiologist wants to estimate the concentration of a certain type of bacterium in a wastewater sample.

She puts a 0.5 mL sample of the waste-water on a microscope slide and counts 39 bacteria.

Estimate the concentration of bacteria per mL, in this waste-water.

Example

Solution:

Let X represent the number of bacteria observed in 0.5 mL.

Let λ represent the true concentration in bacteria per mL.

Then $X \sim Poisson(0.5\lambda)$.

The observed value of X is 39.

The estimated concentration is $\lambda = 39/0.5 = 78$.

Computing bias of λ^

Bias – is intentional or unintentional favoring of one outcome over the other in the population.

In statistics, Bias of an estimator is the difference between estimator's expected value and true value of parameter being estimated.

$$\mu_{\hat{\lambda}} - \lambda \longrightarrow \text{Eq (1)}$$

Since $\hat{\lambda} = X/t$,

$$\mu_{\hat{\lambda}} = \mu_{X/t} = \frac{\mu_X}{t}$$

$$= \frac{\lambda t}{t} = \lambda$$

Substitute μ_λ value in Equation (1) it becomes 0 and $\hat{\lambda}$

becomes unbiased.

Computing uncertainty of λ^

Uncertainty – is the standard deviation of sample proportion.

$$\sigma_{\hat{\lambda}} = \sigma_{x} / t$$

$$= \text{sqrt}\,(\lambda\, t)/t$$

$$= \text{sqrt}(\lambda/t)$$

As λ is unknown when computing uncertainty , we approximate it with λ^

Example

A 5 mL sample of a suspension is withdrawn, and 47 particles are counted. Estimate the mean number of particles per mL, and find the uncertainty in the estimate.

Solution

The number of particles counted is $X = 47$. The volume withdrawn is $t = 5$ mL.
The estimated mean number of particles per mL is

$$\hat{\lambda} = \frac{47}{5} = 9.4$$

The uncertainty in the estimate is

$$\sigma_{\hat{\lambda}} = \sqrt{\frac{\lambda}{t}}$$

$$= \sqrt{\frac{9.4}{5}} \qquad \text{approximating } \lambda \text{ with } \hat{\lambda} = 9.4$$

$$= 1.4$$

Example

A certain mass of a radioactive substance emits alpha particles at a mean rate of λ particles per second. A physicist counts 1594 emissions in 100 seconds. Estimate λ, and find the uncertainty in the estimate.

Solution

The estimate of λ is $\hat{\lambda} = 1594/100 = 15.94$ emissions per second. The uncertainty is

$$\sigma_{\hat{\lambda}} = \sqrt{\frac{\lambda}{t}}$$

$$= \sqrt{\frac{15.94}{100}} \qquad \text{approximating } \lambda \text{ with } \hat{\lambda} = 15.94$$

$$= 0.40$$

THANK YOU

Prof. Uma D
Prof. Suganthi S
Prof. Silviya Nancy J

Department of Computer Science and Engineering