FOUNDATION YEAR 2025-2026

COURSE - DATA ANALYSIS

Lesson 7

Niccolò Maffezzoli, Ca' Foscari University of Venice

February 2, 2026

Remember: Poisson, Binomial, Gaussian formulas

- $P(k, p, n) = p^k(1-p)^{n-k}\frac{n!}{k!(n-k)!}$ *Binomial*
  - Fixed number of independent trials and each trial can only have one of two possible outcomes (usually success or failure). $\mu = np$, $\sigma = \sqrt{np(1-p)}$
- $P(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$ *Poisson*
  - The Poisson distribution is used to model the number of times an event occurs in a fixed interval of time or space, when the events occur independently and at a constant average rate λ; $\sigma = \sqrt{\lambda}$.
- $P(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ *Gaussian*

Exercise 1

Assume a Poisson distribution is involved and use the mean (i.e., λ) provided to find the indicated probability.

- λ=2; P(3)=? Answer: 0.1804
- λ=0.3; P(1)=? Answer: 0.2222
- λ=3/4; P(3)=? Answer: 0.0332
- λ=1/6; P(0)=? Answer: 0.8465

Exercise 2

Which gives the probability of seeing no threes's in six rolls of a standard die?

- $P = (1/6)^0 (5/6)^6 \frac{6!}{0!6!} = 0.335$
- $P = \frac{(1)^0 (e)^{-1}}{0!} = 0.368$

Answer: the correct is the first one, because the Binomial probability is defined for fixed number of trials.

Exercise 3

In the last 100 years, there have been 93 earthquakes measuring 6.0 or more on the Richter scale. What is the probability of having 3 earthquakes in the same year that all measure 6.0 or more? Answer: $\lambda = 93/100 = 0.93$. $P = \frac{(0.93)^3 (e)^{-0.93}}{3!} = 0.05289$
What is the probability of having less than 4 ? Answer: $P = P(0) + P(1) + P(2) + P(3) = 0.9850$

Exercise 4

Neuroblastoma, a rare form of malignant tumor, occurs in 11 out of a million children. In the 12,439 children of Oak Park, Illinois, 4 cases of neuroblastoma are reported. Assuming there was nothing special about Oak Park where the chance of neuroblastoma is higher than normal, find the probability of seeing more than a single case of neuroblastoma in a population this size. What can you likely conclude?

Answer: This is a Binomial (fixed number of trial n, constant p): $p = 11/1000000$, $n = 12439$. $P(k > 1) = 1 - (P(0) + P(1)) = 0.008548$.
Since p is small and n is large we can approximate using the Poisson with $\lambda = np = 11/100000 * 12439 = 0.1368$ is the average expected number of cases in the Oak Park.

$$P(k > 1) = 1 - (P(0) + P(1)) = 1 - \frac{(0.1368)^0 (e)^{-0.1368}}{0!} - \frac{(0.1368)^1 (e)^{-0.1368}}{1!} = 1 - e^{-0.1368} - 0.1368e^{-0.1368} = 0.008545.$$

We conclude that there is something special about Oak Park that makes the chance of neuroblastoma higher than 11 out of a million.

Exercise 5

Suppose the probability of suffering a fever from the flu vaccine is 0.005. If 1000 people are given the vaccine, use the Poisson distribution to approximate the probability that a) 1 person suffers a fever as a result; and b) more than 6 people suffer a fever as a result.

Answer:

$\lambda = np = 0.005 * 1000 = 5.$

a) $P(1) = \frac{(5)^1 (e)^{-5}}{1!} = 0.0337.$

b) $1 - [P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6)] = 0.2378.$

Calculate a) using a Binomial.

Answer:

p = 0.005, n = 1000.

a) $P(1) = (0.005)^1 (0.995)^{999} \frac{1000!}{1!999!} = 0.0334.$

Exercise 6

Telephone calls enter a college switchboard on the average of two every three minutes. What is the probability of 5 or more calls arriving in a 9-minute period?

Answer:

$\lambda(3m) = 2 calls$. $\lambda(9m) = 6$ calls. $P(x \geq 5) = 1 - [P(0) + P(1) + P(2) + P(3) + P(4)] = 0.7149$

Exercise 7

After receiving a large shipment of computer chips, 800 chips are randomly selected. If 3 or fewer defective chips are found, the entire lot is accepted without inspecting the remaining chips in the lot. If 4 or more chips are defective, every chip in the entire lot is carefully inspected at the supplier's expense. Assume that the true proportion of nonconforming computer chips being supplied is 0.001. Use the Poisson distribution to approximate the probability the lot will be accepted.

Answer:

$p(defect) = 0.001$, $n = 800$. $\lambda = np = 0.8$ (average number of defective chips in the batch).

$P(accepted) = P(k \leq 3) = P(0) + P(1) + P(2) + P(3) = 0.9909$

Exercise 8

Wool fibre breaking strengths are normally distributed with mean $\mu = 23.56$ Newtons and standard deviation, $\sigma = 4.55$. What proportion of fibres would have a breaking strength of 14.45 or less? Draw the distribution and indicate the mean.

Answer:

Convert raw score x to z-score: $z = \frac{x-\mu}{\sigma} = \frac{14.45-23.56}{4.55} = -1.98$. z is negative because 14.45 lies on the left of the mean. We need to calculate $P(x < 14.45) = P(z < -1.98)$. From the table we read $P(-1.98 < z < 1.98) = 0.9523$.

$P(x < 14.45) = P(z < -1.98) = \frac{1-P(-1.98<z<1.98)}{2} = (1 - 0.9523)/2 = 0.023$. The proportion of fibres with a breaking strength of 14.46 or less is 2.3%.

Exercise 9

Carrots entering a processing factory have an average length of 15.3 cm, and standard deviation of 5.4 cm. If the lengths are approximately normally distributed, what is the maximum length x of the lowest 5% of the load? (I.e., what value x cuts off the lowest 5%?) Draw the distribution with the mean and the 5% cutoff.

Answer:

I need to calculate a score z so that $\frac{1-I(z)}{2} = 0.05 \rightarrow I(z) = 0.9$. From the table I can see $z = -1.645$. Important: I need a negative because z because it is below the mean. $z = \frac{x-\mu}{\sigma} \rightarrow x = z\sigma + \mu = -1.645 \cdot 5.4 + 15.3 = 6.417$ cm.

Exercise 10

The download time of a resource web page is normally distributed with a mean of 6.5 seconds and a standard deviation of 2.3 seconds.

- a) What proportion of page downloads take less than 5 seconds?
- b) What is the probability that the download time will be between 4 and 10 seconds?
- c) What is the fastest time for 35% of the downloads to be completed?

Answer:

- a). $z = \frac{x-\mu}{\sigma} = \frac{5-6.5}{2.3} = -0,652$. $I(x < 5s) = I(z < -0.652) = \frac{1-I(z)}{2} = \frac{1-0.4843}{2} = 0.258 = 25.\%$.
- b) $P(4 < x < 10)$?. Let's call $x_1 = 4$ and $x_2 = 10$. $z_1 = \frac{4-6.5}{2.3} = -1.09$. $z_2 = \frac{10-6.5}{2.3} = 1.52$. $P(4 < x < 10) = P(-1.09 < z < 1.52) = P(-1.09 < z < 0) + P(0 < z < 1.52) = \frac{0.7243}{2} + \frac{0.8715}{2} = 0.7979$. About 80% of downloads take between 4 and 10 seconds.
- c) I need to find z such that $P(z < -\bar{z}) = 0.35 \rightarrow P(-\bar{z} < z < \bar{z}) = 0.3$. From the tables I find that the corresponding z is $\bar{z} = -0.385$ (negative since $\bar{z} < 0$). $\bar{z} = \frac{\bar{x}-\mu}{\sigma} \rightarrow \bar{x} = \bar{z}\sigma + \mu = -0.385 \cdot 2.3 + 6.5 = 5.6\ s$. We conclude that 35% of downloads will be less than 5.6 seconds.