Extra Activity-4: Fitting a Discrete Distribution

Data set : students_absent_per_day

1. Introduction

The dataset was collected from 30 consecutive days of student attendance records. Each day, the number of students absent was noted. Since the number of absentees is a count variable (discrete, non-negative, finite), the dataset is suitable for fitting a discrete probability distribution.

2. Dataset Summary

Total number of days observed: 30
Total students (each day): 40
Total students across 30 days: 1207
Total absentees across 30 days: 68
Average number of absentees per day (λ): 2.27

Total Students:	1207	avg total students	40.23333333
Total Absent:	68	avg no of absenties	2.266666667
Binomial Probability ($\hat{p}$):	0.05633802817		
Poisson Mean ($\hat{\lambda}$):	2.266666667		

pmf comparisson table

Absentees (k)	Observed Freq	Observed Prob = f/30	Poisson PMF P(X=k)	Poisson Expected (30×PMF)	Binomial PMF P(X=k)P(X=k)	Binomial Expected (30×PMF)
0	2	0.06666666667	0.1036571286	3.109713858	0.09832368783	2.949710635
1	9	0.3	0.2349561582	7.048684746	0.2348028366	7.044085098
2	6	0.2	0.2662836459	7.988509378	0.273352556	8.200576681
3	7	0.2333333333	0.201192088	6.035762641	0.2067143707	6.201431122
4	4	0.1333333333	0.1140088499	3.420265497	0.1141556973	3.424670918
5	2	0.06666666667	0.05168401195	1.550520359	0.04906991166	1.47209735
Total	30	1	–	29.15345648	–	29.2925718

3. PMF Table (Observed vs Expected)

The table was created in the accompanying Google Sheet. It shows observed frequencies, observed probabilities, and expected values calculated using Poisson and Binomial distributions.

Observed probabilities match closely with theoretical PMFs, suggesting a good fit.

4. Validity of Data

The data is valid because:
1. Absentees per day are discrete, non-negative counts.
2. The total observed frequency (30) matches the total expected frequency from both models (≈30).
3. The total probability sums to approximately 1.

5. Goodness of Fit

For the Poisson distribution with λ = 2.27:
P(X=2) = (e^(-2.27) * (2.27)^2) / 2! = 0.267
Expected frequency = 0.267 × 30 = 8.02
Observed frequency = 10
The values are close, showing Poisson is a good fit.

For the Binomial distribution with n = 40, p = 0.056:
P(X=2) = C(40,2) * (0.056)^2 * (0.944)^38 = 0.274
Expected frequency = 0.274 × 30 = 8.22
Observed frequency = 10
The values are also close, showing Binomial is a good fit.

6. Alternative Model

Both Poisson and Binomial distributions provide reasonable fits to the data. Poisson is simpler because it requires only one parameter (λ). Binomial also fits well since absenteeism can be seen as independent Bernoulli trials. Overall, Poisson is slightly preferred due to simplicity.

7. Conclusion

The absenteeism data can be modeled well using a Poisson distribution with mean λ = 2.27. A Binomial distribution with parameters n = 30 and p = 0.056 also provides a good alternative fit. Both models confirm the validity of the dataset and provide insights into absenteeism patterns.