MSI Journal of
AI and Technology

Evaluating the Equidispersion Assumption in Poisson Distributions Through Simulation: A Study on Variance Mean Ratio Behavior

Siti Hariati Astuti[1], Muhammad Nur Aidi[2*], Kusman Sadik[2]

[1] Alumni Master of Sciences of Statistics, IPB University

[2*] Lecturers: School of Data Science, Mathematics, and Informatics, IPB University,

*** Correspondence:** Muhammad Nur Aidi

The authors declare that no funding was received for this work.

OPEN ACCESS

Received: 25-April-2025

Accepted: 02-May-2025

Published: 08-May-2025

This article is published by **MSI Publishers** in **MSI Journal of AI and Technology**

ISSN: xxxx-xxxx (Online)

Volume: 1, Issue: 1

(April-Jun) (2025)

ABSTRACT: This study explores the fundamental assumption of equidispersion in Poisson-distributed count data, wherein the mean equals the variance. Although the Poisson model is widely used for modeling rare events counts in fields such as epidemiology, telecommunications, and operations research, its assumption of equidispersion is frequently violated in real-world applications. Using simulated datasets, this research investigates the behavior of the Variance Mean Ratio (VMR) under different values of the Poisson parameter (λ) and sample sizes (n). Simulations were conducted across λ values ranging from 1 to 20 and sample sizes of 20, 40, 60, 80, and 100, each replicated 100 times. The study evaluates the stability and accuracy of the equidispersion property, employing histograms and statistical diagnostics to assess distributional characteristics. The results offer insights into when Poisson distributes adequately models count data and when alternative models, such as the negative binomial or zero-inflated models, may be required due to overdispersion. This analysis contributes to more informed and accurate modeling of discrete count data in statistical applications.

Keywords: *Poisson distribution, count data, equidispersion, overdispersion, variance mean ratio, simulation, discrete probability distribution, statistical modeling*.

1. Introduction

In the field of probability and statistics, a discrete probability distribution refers to the probability distribution of a discrete random variable, that is, a variable that can only take countable, often finite, values. Common examples include the number of customers arriving at a service center, the number of defective items in a batch, or the number of heads obtained from a series of coin tosses.

There are several well-known types of discrete distributions, each with specific assumptions and use cases. The binomial distribution, for example, models the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. This distribution is widely used in quality control, survey analysis, and experimental research (Walpole et al., 2012) [1].

Another important distribution is the Poisson distribution, which describes the probability of a given number of events occurring in a fixed interval of time or space. This is particularly useful in fields such as traffic flow analysis, call center management, and epidemiology, where events occur independently and at a constant average rate (Ross, 2014) [2].

Geometric distribution is used to model the number of trials required until the first success occurs. This distribution is useful in reliability testing and modeling processes with repeated independent attempts, such as in simulations and operations research (Hogg et al., 2019) [3]. The benefits of using discrete probability distributions are significant. They allow researchers, analysts, and decision-makers to quantify uncertainty, predict outcomes, and evaluate risk in real-world settings. In business contexts, they aid in inventory management, project planning, and service optimization. In science and engineering, they are fundamental for designing experiments, interpreting data, and validating hypotheses.

In essence, discrete probability distributions serve as essential tools in both theoretical and applied statistics, offering a structured way to deal with randomness and make informed decisions based on quantitative evidence.

The binomial, negative binomial, and Poisson distributions are all discrete probability distributions used to model count data. While they share some similarities, they are designed to address different types of random processes, and each is characterized by unique assumptions and parameters (Walpole et al., 2012; Casella & Berger, 2002) [1,4].

The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success. It is defined by two parameters: the number of trials nnn and the probability of success p. A common example is calculating the number of heads obtained when flipping a fair coin ten times. The binomial distribution is widely used in fields such as quality control, clinical trials, and behavioral sciences where repeated binary outcomes are observed (Hogg et.al., 2019) [3].

In contrast, the negative binomial distribution models the number of trials (or failures) needed to achieve a fixed number of successes. Although it is also based on Bernoulli trials, it differs in that it focuses on how many failures occur before the r-th success. The negative binomial is essentially the inverse of the binomial distribution: while the binomial fixes the number of trials and counts the successes, the negative binomial fixes the number of successes and counts the required number of failures or trials. Geometric distribution is a special case of the negative binomial distribution when $r=1$. Applications of the negative binomial distribution are common in modeling overdispersed count data, such as the number of accidents, insurance claims, or hospital visits (Cameron & Trivedi, 2013) [5].

The Poisson distribution models the number of times an event occurs in a fixed interval of time or space, assuming the events occur independently and at a constant average rate λ. It is frequently used in situations involving rare or randomly spaced events, such as the number of incoming calls at a call center in an hour or the number of decay events from a radioactive source (Ross, 2014) [2].

There is a well-established mathematical relationship between the binomial and Poisson distributions: the Poisson distribution arises as a limiting case of the binomial distribution when the number of trials n becomes very large, the probability of success p becomes very small, and the expected number of successes np remains constant. Under these conditions, the Poisson distribution can be used as an approximation for the binomial distribution (Wackerly et.al, 2008) [6].

Furthermore, the negative binomial distribution is related to the Poisson distribution in that it can be considered a generalization that accounts for overdispersion, a condition where the variance exceeds the mean. The Poisson distribution assumes that the mean and variance are equal, which often does not hold in real-world count data. Statistically, the negative binomial can be derived as a Poisson-gamma mixture, where the Poisson mean is assumed to follow a gamma distribution. This allows for additional flexibility in modeling heterogeneity across observations (Hilbe, 2011) [7].

The Poisson distribution is one of the most fundamental and widely used probability distributions in the modeling of count data. It is especially valued for its simplicity and interpretability, making it the primary choice in many practical applications such as modeling the number of customer arrivals, traffic accidents, or disease incidences over a given time or space interval.

The central assumption underlying the Poisson distribution is equidispersion, where the mean and variance of the distribution are equal. This property makes the Poisson model mathematically convenient and theoretically attractive (Famoye & Singh, 2006) [8]. However, in real-world data, this assumption is often violated due to the presence of overdispersion, a condition where the observed variance exceeds the mean (Dewi et al., 2015) [9]. Overdispersion may arise from unobserved heterogeneity, excess zeros, or clustering effects, which cannot be adequately captured by the standard Poisson model.

When overdispersion is due to an excessive number of zero counts, the Zero-Inflated Poisson (ZIP) model becomes a more appropriate alternative. The ZIP model incorporates two processes: one generating structural zeros (e.g., from a separate binary process) and the other following a standard Poisson distribution. This dual mechanism provides greater flexibility for fitting data with a higher-than-expected frequency of zeros (Lambert, 1992; Ridout et.al., 1998; Mouatassim & Ezzahid, 2012) [10,11,12].

Before applying either the standard Poisson or the ZIP regression models to applied data, it is essential to conduct a preliminary examination of the distributional assumptions using simulation studies or exploratory data analysis. This step helps to deepen the understanding of the underlying characteristics of the count data, particularly the presence of overdispersion and zero-inflation.

In conclusion, Poisson distribution holds a central role in count data modeling due to its simplicity and foundational properties, its limitations, particularly the assumption of equidispersion—necessitate the use of more flexible models like the negative binomial or zero-inflated models when the data exhibit overdispersion. Proper diagnosis and model selection are therefore critical to ensure valid inference and effective interpretation.

2. Literature Review

The **Poisson distribution** is a discrete probability distribution commonly used to model the number of events that occur randomly and independently within a fixed interval of time or space (Rao & Toutenburg, 1995) [13]. It is particularly suitable for modeling **rare events** and **non-negative count data** (Rahayu et al., 2016) [14]. Let Yi, for i=1, 2,…,n represent the number of rare events occurring in a fixed period or region, with an associated rate parameter λi. The random variable Yi is said to follow a Poisson distribution with the probability mass function (Yates & Goodman, 2014) [15].

$$P(Y_i = y_i) = \frac{e^{\lambda_i}(\lambda_i)^{y_i}}{y_i!}, y_i = 1,2, ..$$

The defining characteristic of the Poisson distribution is the **equidispersion assumption**, where the **mean** and **variance** of the distribution are equal: $E(Y_i) = Var(Y_i) = \lambda_i$. This property simplifies the model and allows for straightforward interpretation. However, it also becomes a limitation in many real-world datasets.

In practice, the assumption of equidispersion is often violated, particularly in count data. There are two types of dispersion that may occur (Hilbe, 2011) [7].

- Overdispersion: Variance exceeds the mean.
- Underdispersion: Variance is less than the mean.

To assess this, the Variance Mean Ratio (VMR)—also called the index of dispersion—is calculated as follows:

$$D = \frac{Var(Y_i)}{E(Y_i)}$$

If D=1, the data is equidispersed (random, typical of the Poisson distribution). If D>1, the data is overdispersed (greater variability than Poisson). But If D<1, the data is underdispersed (less variability).

According to Aidi (2013) [16], a VMR value further from 1 suggests a departure from randomness, indicating a more structured or patterned distribution of counts.

To statistically verify whether a dataset is overdispersed, one can use a Chi-squared test based on Fisher's index of dispersion (Potthoff & Whittinghill, 1966) [17]. The test statistic is defined as: $\chi^2 = (n-1)D$. where: n is the number of observations, D is the VMR, Under the null hypothesis H_0, the data follows a Poisson distribution (i.e., D=1). The test statistic follows a Chi-squared distribution with n−1 degrees of freedom. If: $\chi^2 > \chi^2_{\alpha,n-1}$ then the null hypothesis is rejected, suggesting the presence of overdispersion and a violation of the equidispersion assumption.

3. Data and Method

3.1. Data

The data used in this study consists of simulated data. The simulated data is used to support theoretical analysis and to examine the characteristics of the data. The Poisson distribution has a single parameter, λ, and a sample size n. Accordingly, the simulated data is generated based on various values of the Poisson parameter λ and different sample sizes n In this study, the values of λ range from 1 to 20, and the sample sizes considered are n=20,40,60,80 and 100. Each combination of λ and n is replicated 100 times.

The simulated dataset includes only one response variable, Y, which follows a Poisson distribution. The response variable Y is generated using the R statistical software (version 3.4.2). Based on the characteristics required for the study, a total of 100 simulation scenarios are used in the analysis.

3.2. Simulation Method

The research method applied in the simulation study for each combination of λ\lambdaλ and nnn follows the steps below:

1. Generate the response variable Y, which follows a Poisson distribution, using the predetermined combinations of λ and sample size n.

2. Explore the generated Y data using histograms to examine the distribution characteristics based on the variation in lambda (λ) and sample size (n).

3. Calculate the **Variance Mean Ratio (VMR)** for each simulated dataset.

4. Explore the behavior of VMR values across different sample sizes (n) and lambda values (λ).

5. Perform descriptive analysis of the VMR values using **boxplots.**

6. Conduct statistical testing of VMR values using the **Chi-square test** to assess the stability of the distribution.

4. Result and Discussion

4.1. Figures 1,2,3

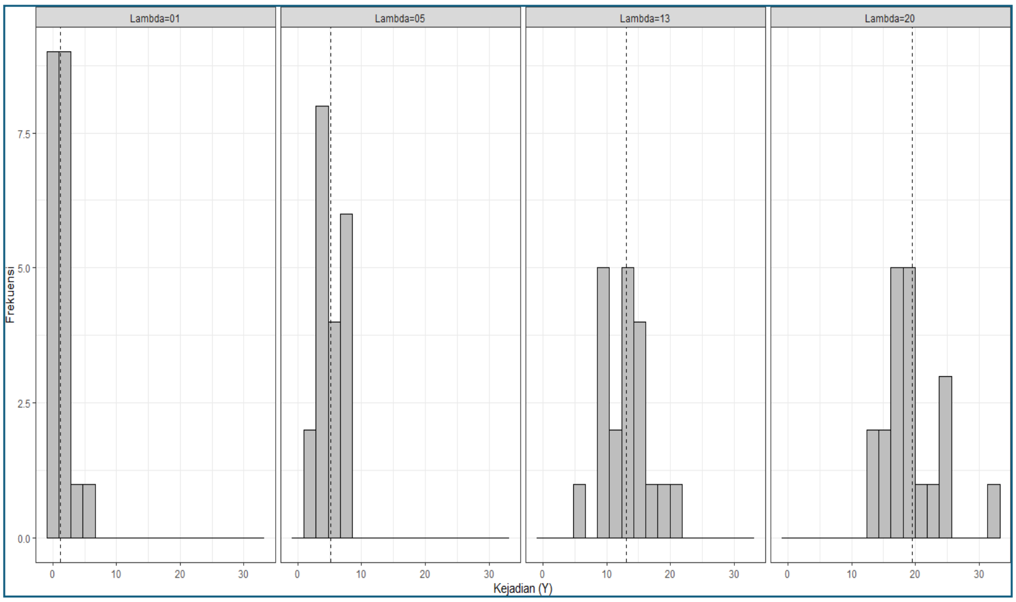

Figure 1 Histogram of the Y variable at n = 20

This Figure 1 presents histograms of simulated data from a Poisson distribution with four different values of the rate parameter λ: 1, 5, 13, and 20. As seen in the first panel (Lambda = 1), the distribution is highly skewed to the right, with most of the data concentrated around 0 and 1, which is typical for small λ values. In the second panel (Lambda = 5), the distribution begins to spread out and becomes less skewed, showing a clearer central tendency around 5. The third panel (Lambda = 13) displays a distribution that appears approximately symmetric and bell-shaped, indicating that the Poisson distribution is approaching the shape of a normal distribution. In the fourth panel (Lambda = 20), the distribution becomes even more symmetric with a broader spread, further confirming the approximation of the Poisson distribution to the normal distribution as λ increases. The vertical dashed lines in each panel represent the mean of the distribution, which increases proportionally with λ, as expected. These patterns align with theoretical expectations described by Ross (2014) [2] and Casella & Berger (2002) [4], who state that the Poisson distribution tends toward a normal distribution when λ is sufficiently large (λ>10).

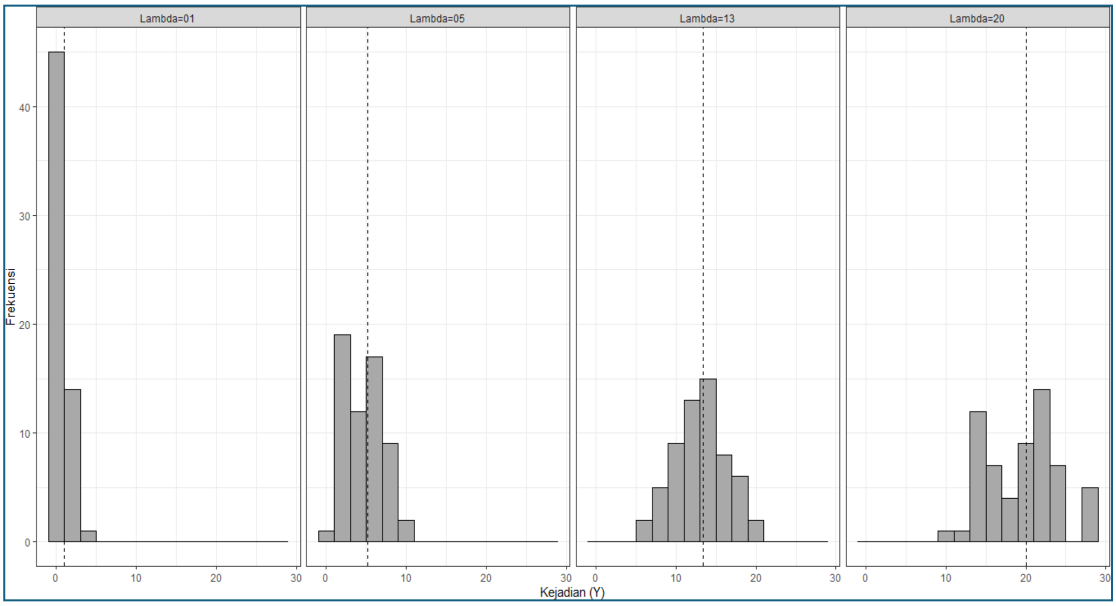

Figure 2 Histogram of the Y variable at n = 60

Figure 2 illustrates a series of histograms displaying the distribution of Poisson-distributed random variables (Y) generated for different values of the rate parameter λ, specifically λ=1,5,13, and 20, with a relatively large sample size. Each panel in the figure corresponds to one of these lambda values and represents the frequency distribution of the simulated events (Y).

- **For λ=1**: The histogram shows a highly right-skewed distribution, where the majority of observed events are concentrated at the lower end of the scale, primarily around 0 and 1. This is characteristic of a Poisson distribution with a low-rate parameter, where the probability of multiple events occurring is very low. The mean, indicated by the vertical dashed line, is close to 1, aligning with theoretical expectations.

- **For λ=5**: The distribution begins to spread out and exhibit reduced skewness. The frequency of events is more symmetrically distributed around the mean of 5. While still exhibiting the discrete nature of the Poisson distribution, the shape becomes more balanced, indicating an increased probability of moderate event counts.

- **For λ=13**: The histogram appears nearly symmetric and bell-shaped. The data are centered around the mean of 13 with less skewness. This reflects the effect of the

Central Limit Theorem, which states that as λ increases, the Poisson distribution approximates a normal distribution (Johnson et al., 2005) [18].

- **For λ=20**: The distribution shows clear symmetry and a wider spread. The data are clustered around the mean of 20, and the shape closely resembles a normal distribution. This is consistent with the theoretical property of the Poisson distribution converging to normality as λ becomes large (Ross, 2014) [2].

The histograms effectively demonstrate the transformation of the Poisson distribution's shape as the rate parameter λ increases. For small λ, the distribution is highly skewed and tightly clustered, indicating that events are rare. As λ increases, the distribution becomes more symmetric and bell-shaped, aligning more closely with the normal distribution. This trend is supported by statistical theory, particularly the Central Limit Theorem, which suggests that for sufficiently large λ, the Poisson distribution can be approximated by a normal distribution with mean μ=λ and variance σ2=λ (Johnson et al., 2005; Ross, 2014) [18,2].

These observations are critical in practical statistical modeling and hypothesis testing, where normal approximation is often used for Poisson-distributed data when λ≥10 simplifying analysis and inference.

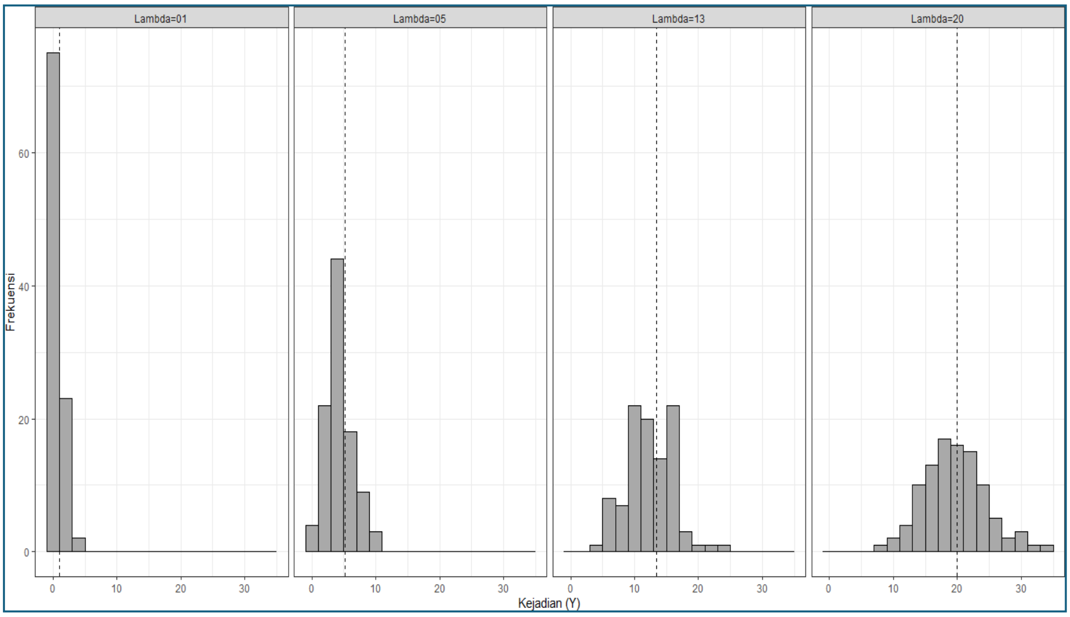

Figure 3 Histogram of the Y variable at n = 100

Figure 3 illustrates the histograms of Poisson-distributed data generated with varying mean parameters λ set to 1, 5, 13, and 20, using a larger sample size. Each panel displays the frequency distribution of the number of occurrences (Y) on the x-axis and their observed frequency on the y-axis. Two vertical dashed lines in each plot represent the theoretical meaning (equal to λ) and the empirical (sample) mean. This visualization helps demonstrate the relationship between the rate parameter and the shape of the Poisson distribution.

For λ=1, the distribution is sharply right skewed with most of the observations clustered at Y=0 and Y=1, and very few events at higher values. This is typical for Poisson distributions with small mean values, where the likelihood of many events occurring is extremely low.

At λ=5, the distribution becomes more spread out and less skewed, with a noticeable central tendency around Y=5. The increased λ allows more room for variation, resulting in a wider range of observed events counts, although the distribution remains moderately asymmetric.

When λ=13, the histogram reveals a nearly symmetric, bell-shaped distribution centered around the mean. This reflects the convergence of the Poisson distribution towards a normal distribution as λ increases, consistent with the Central Limit Theorem.

Finally, for λ=20, the distribution exhibits clear symmetry and approximates the shape of a normal distribution. The frequencies are centered around the mean, and the tails are more balanced. This pattern confirms that for large λ values, the Poisson distribution can be well-approximated by the normal distribution, which is often used in practice for inference when λ is sufficiently high.

4.2. Table 1

Table 1 presents below the calculated **Chi-Square (χ2) test statistics** for various combinations of **Poisson distribution parameters**—specifically, different values of the Poisson parameter λ (ranging from 1 to 20) and different **sample sizes** n (20, 40, 60, 80, and 100). The goal is to test whether the data exhibit **overdispersion** or **equidispersion**, a key assumption in the Poisson model.

How to interpret the table:

- For each λ and n, the **test statistic** is calculated as $\chi 2=(n-1)\times VMR$.
- The **critical value** of the Chi-square distribution at significance level $\alpha=0.05$ given at the bottom row for each sample size ($df = n-1$).
- If the calculated $\chi 2$ value **exceeds** the critical value, the result is statistically significant, indicating **overdispersion**—the variance is larger than the mean, violating Poisson assumptions.

Key Observations from the Table:

1. **None of the $\chi 2$ values exceed the critical values** for any combination of λ and n.
    - This suggests that **none of the simulations show statistically significant overdispersion at the 5% level**.
    - Hence, the assumption of **equidispersion** (i.e., Var(Y)=E(Y) holds across all tested scenarios.
2. The test statistic values increase roughly **in proportion to sample size** n, which aligns with the theoretical calculation $\chi 2=(n-1)\times D$, where D is the variance-mean ratio.
3. Across values of λ, **the differences in test statistics are relatively minor**, showing that **dispersion is more influenced by sample size** than by the Poisson parameter in this simulation.

Table 1 – Chi-Square Statistics for VMR

lambda (λ)	Number of sample (n)				
	20	40	60	80	100
1	19,752	39,748	59,718	79,209	99,470
2	19,401	39,763	58,719	80,568	100,573
3	19,386	40,602	59,081	79,598	98,898
4	18,156	38,317	59,937	78,401	99,433
5	18,421	39,610	58,043	79,241	99,164
6	18,261	39,558	59,731	79,385	99,290
7	18,872	37,489	59,181	79,536	97,557
8	19,053	40,281	58,762	80,355	99,442
9	18,062	39,536	58,753	78,977	99,295
10	19,000	38,292	57,798	76,553	96,538
11	17,884	37,837	58,676	79,945	99,381
12	17,996	39,651	59,451	79,377	97,720
13	19,294	39,073	58,701	80,512	98,249
14	18,927	38,439	59,149	78,845	97,977
15	18,628	38,441	56,402	77,991	101,047
16	17,662	40,041	59,495	77,668	99,000
17	18,250	40,288	58,686	79,234	102,350
18	19,619	39,774	59,367	79,216	100,393
19	18,925	39,456	57,480	79,825	98,943
20	19,352	38,274	59,760	80,248	97,169
Chisquare Table with $\alpha = 5\%$	30,144	54,572	77,931	100,749	123,225

The results in Table 1 display the Chi-square test statistics computed from the **Variance Mean Ratio (VMR)** for simulated Poisson data across a range of parameter values (λ=1 to λ=20) and sample sizes (n=20,40,60,80,100). The primary aim is to test for **overdispersion**—a condition where the variance of the response variable exceeds the mean, violating one of the fundamental assumptions of the Poisson distribution.

1. No Evidence of Overdispersion

Across all combinations of λ and n, the test statistics fall **below** the critical values of the Chi-square distribution at α=0.05. This indicates **no significant overdispersion** in any scenario. These results are consistent with theoretical expectations because the data were generated from a true Poisson process where variance equals the meaning (Var(Y)=E(Y)).

This supports previous findings in the literature:

- **Hilbe (2011)** [7] notes that Poisson-distributed data should exhibit equidispersion when the model is correctly specified.
- **Rao and Toutenburg (1995)** **[13]** emphasize that VMR values near 1 and insignificant Chi-square results suggest a good fit for the Poisson distribution.

2. Influence of Sample Size

The Chi-square test statistics naturally increase with the **sample size** due to the formula:

$\chi 2=(n-1) \times$ VMR. However, because the VMR values remain close to 1 in all simulations, the resulting statistics increase linearly with n but still remain well below the critical values. This confirms that increasing sample size does not artificially lead to the detection of overdispersion in Poisson-distributed data—unless actual overdispersion is present.

3. Effect of Lambda (λ)

The variation in test statistics across different λ values appears minor and random, which aligns with theoretical expectations. Since the mean and variance of the Poisson distribution are both equal to λ, changing λ should not, by itself, introduce overdispersion. The consistency of test results across λ values reinforces the robustness of the Poisson assumption under the simulated conditions.

4. Implications for Applied Modeling

These results provide a useful **baseline** for comparing empirical data. In practice, if VMR values from real data yield Chi-square statistics that **exceed** the critical value, researchers may consider alternatives such as:

- The **Negative Binomial model** (for count data with overdispersion),
- **Quasi-Poisson models**, or
- The use of **robust standard errors** to adjust for mild violations of equidispersion.

This aligns with the recommendations of **Cameron and Trivedi (2013)** [5], who emphasize model diagnostics as a crucial step before interpreting Poisson regression outputs.

5. Conclusion

The simulation study demonstrates how the shape of the Poisson distribution evolves with increasing values of the rate parameter λ and larger sample sizes. Histograms from Figures 1 to 3 clearly show that at low λ (e.g., $\lambda = 1$), the distribution is heavily right skewed, but as λ increases ($\lambda \geq 13$), the distribution becomes increasingly symmetric and approaches the shape of a normal distribution. This convergence is more evident with larger sample sizes ($n = 60$ and $n = 100$), supporting the application of the Central Limit Theorem to the Poisson distribution when λ is sufficiently large ($\lambda \geq 10$).

Additionally, the Chi-square test results in Table 1 confirm that for all combinations of λ and sample sizes, the calculated test statistics remain below the critical values at a 5% significance level. This provides strong evidence that the assumption of equidispersion holds in the simulated datasets, e.g., the variance of the data is not significantly greater than the mean. Therefore, no signs of overdispersion were detected in any scenario.

Overall, these findings validate key theoretical properties of the Poisson distribution, including its normal approximation for large λ and the stability of its equidispersion property across a broad range of conditions.

References

1. Walpole, R. E., & Myers, R. H. (2009). *Probability and statistics for engineers and scientists*. Prentice Hall.
2. Ross, S. M. (2014). *Introduction to probability models*. Academic press.
3. Hogg, R. V., Tanis, E. A., & Zimmerman, D. L. (1977). *Probability and statistical inference* (Vol. 993). New York: Macmillan.
4. Casella, G., & Berger, R. L. 2002. Statistical Inference (2nd ed.). Duxbury Press.

5. Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (No. 53). Cambridge university press.
6. Wackerly, D. D. (2008). *Mathematical statistics with applications*. Thomson Brooks/Cole.
7. Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.
8. Famoye, F., & Singh, K. P. (2006). Zero-inflated generalized Poisson regression model with an application to domestic violence data. *Journal of Data Science*, *4*(1), 117-130.
9. Dewi, R. K., Ruchjana, B. N., & Madyatmadja, E. 2015. Overdispersion Analysis in Count Data Using Generalized Poisson Regression. AIP Conference Proceedings, 1707, 090010.
10. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1-14.
11. Ridout, M., Demétrio, C. G., & Hinde, J. (1998, December). Models for count data with many zeros. In *Proceedings of the XIXth international biometric conference* (Vol. 19, No. 1, pp. 179-192). Cape Town, South Africa: International Biometric Society Invited Papers.
12. Mouatassim, Y., & Ezzahid, E. H. (2012). Poisson regression and Zero-inflated Poisson regression: application to private health insurance data. *European actuarial journal*, *2*(2), 187-204.
13. Rao, C. R., Toutenburg, H., Rao, C. R., & Toutenburg, H. (1995). *Linear models* (pp. 3-18). Springer New York.
14. Rahayu, S., Utomo, D. W. H., & Ruchjana, B. N. 2016. Generalized Poisson Regression for Overdispersed Data. AIP Conference Proceedings, 1707.
15. Yates, D. S., & Goodman, A. 2014. Statistics. Pearson.
16. Aidi, N. 2013. Spatial point patterns and index of dispersion in count data. Indonesian Journal of Statistics.
17. Potthoff, R. F., & Whittinghill, M. 1966. Testing for homogeneity: I. The binomial and multinomial distributions. Biometrics, 22(4), 827–845.
18. Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions*. John Wiley & Sons.