Violating Equidispersion: A Comparative Analysis of Poisson and Negative Binomial Regression

By: Hunter Evans
Advisor: Dr. Seals

Outline

1. Introduction
2. Methods
3. Results
4. Conclusion
5. Further Research

Statement of the Problem

- Statistical tests have related assumptions.
- Assumptions are also sometimes hard to test.
- What if someone used a model without testing or fulfilling certain assumptions?
- In particular, the equidispersion assumption of Poisson.

Relevance

- The real world is messy. Data is sloppy.
- What would happen if we did use Poisson on this real world over dispersed data?

Count Data: Why Can't we use the normal distribution?

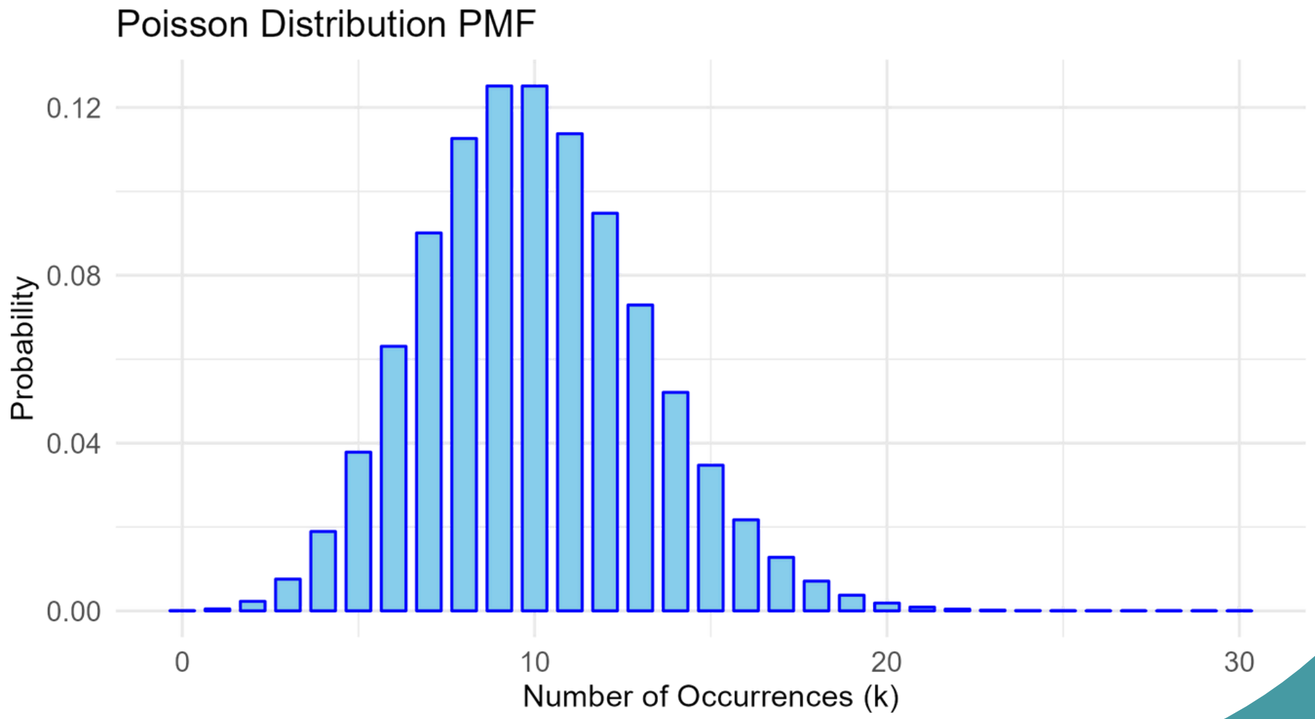
- Count data is nonnegative integers, ex. 0,1,2,3
- The normal distribution is a continuous probability function
- Count data is discrete
- Results would not make sense in the normal distribution

Poisson Regression

PMF: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$

MGF: $M_Y(t) = E[e^{tY}] = e^{\lambda(e^t - 1)}$

- Count data
- $\mu = \sigma^2$
- Uses 1 parameter: λ



Negative Binomial Regression

PMF: $$P(Y = y) = \frac{\Gamma(y + r)}{\Gamma(r)y!}\left(\frac{\mu}{\mu + r}\right)^{y}\left(\frac{r}{\mu + r}\right)^{r}$$

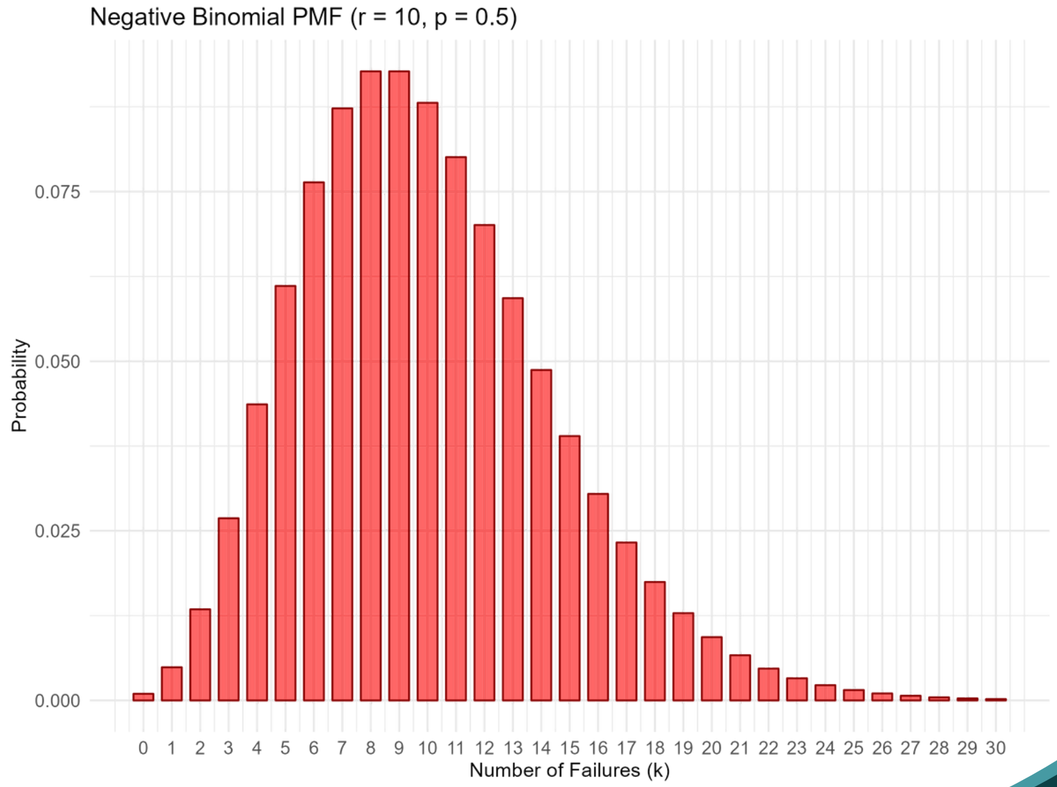
MGF: $$M_{NB}(t) = \left(\frac{p}{1 - (1 - p)e^{t}}\right)^{r}$$

- Count data
- Generalized model from

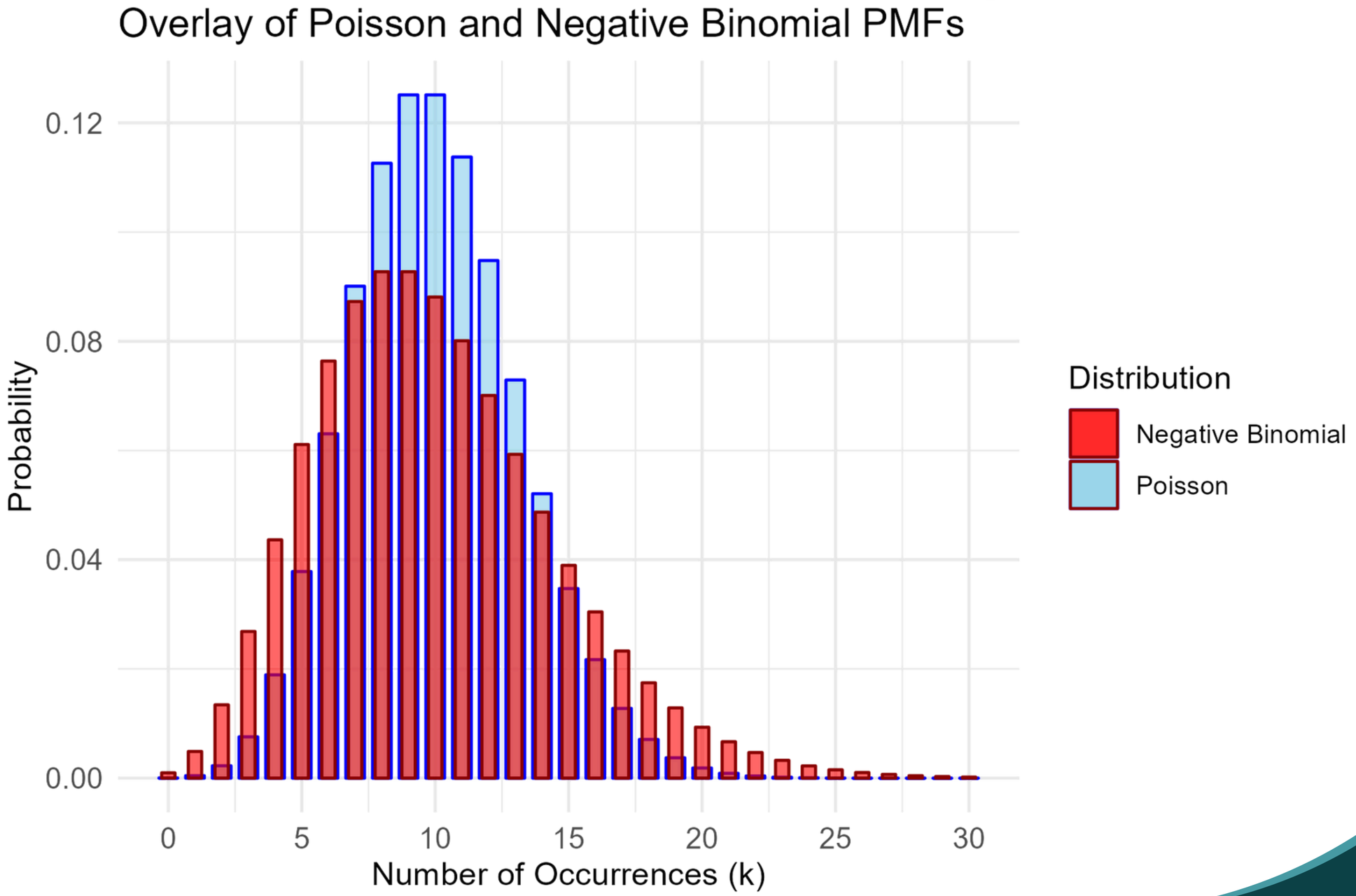
  Poisson
- 2 parameters:

- μ is the mean

- r deals with dispersion



Overlay of PMF: Poisson & Negative Binomial



MGF Negative Binomial Converges to MGF Poisson

$$\lim_{r\to\infty} M_{NB}(t) = \lim_{r\to\infty} \left(\frac{p}{1-(1-p)e^t}\right)^r$$

$$= \lim_{r\to\infty} \left(\frac{1-(1-p)}{1-(1-p)e^t}\right)^r$$

$$\text{Using } r(1-p) = \lambda \Rightarrow 1-p = \frac{\lambda}{r}$$

$$= \lim_{r\to\infty} \frac{\left(1+\frac{1}{r}(-\lambda)\right)^r}{\left(1+\frac{1}{r}e^t(-\lambda)\right)^r}$$

$$= \lim_{r\to\infty} \frac{e^{-\lambda}}{e^{-\lambda e^t}}$$

$$= e^{\lambda(e^t-1)} = M_Y(t)$$

Note: Lemma 2.3.14 in the Statistical Inference by Casella and Berger book states a useful limit where when we have $\lim_{n\to\infty} a_n = a$

$\lim_{n\to\infty}(1+\frac{a_n}{n})^n = e^a$

Goals of the study

- Compare Poisson and Negative Binomial at different level of overdispersion.

- Compare these by simulating data.

  - Bias, MSE, Standard Errors, Type 1 error, AIC

- See if results are comparable to theory

Simulation Study

- Simulation is the process of creating synthetic data
- We wanted to simulate data using R studio
- Allows to quantify bias by picking parameters
- Wrote a function to iterate data

The Variables

- n is sample size
- β_0 is the y - intercept
- β_1 is the slope
- $\mu = \exp(\beta_0 + \beta_1 * x)$
- θ is the dispersion parameter
- iterations are the number of times the function repeats

Cases

There are 5 cases:

- $\mu = \sigma^2$ ($\theta = 2000$)
- $\mu \cong \sigma^2$ ($\theta = 500$)
- $\mu < \sigma^2$ ($\theta = 50$)
- $\mu << \sigma^2$ ($\theta = 10$)
- $\mu <<< \sigma^2$ ($\theta = 1$)

All of these cases were also run at the following sample sizes:

- n = 25
- n = 100
- n = 500

Picking Parameters

- We let our parameters be $\beta_0 = 1.5$ and $\beta_1 = 0.25$
- Our original goal was to use an Earth & Environmental science based model
- Ran into convergence issues
- Reverted to arbitrary numbers

Expected Results

Based on our literature review:

- **We expect Bias and MSE to not differ much**
- **Poisson underestimating standard errors at high overdispersion**
- **This leads to Type 1 errors**
- **We also expect to see the Negative Binomial Model fit better at higher overdispersion**

What is Bias?

$$\text{Bias}(\hat{\beta}) = \mathbb{E}[\hat{\beta}] - \beta$$

$\mathbb{E}[\hat{\beta}]$ is the expected value of $\hat{\beta}$

β is the true value

- Estimates deviation from the true value

What is MSE?

$$\mathbf{MSE}(\hat{\beta}) = \mathbf{Var}(\hat{\beta}) + \Big[\mathbf{Bias}(\hat{\beta})\Big]^2$$

- Mean Squared Error
- Measures the average squared difference between actual and predicted values

Underestimating Standard Errors

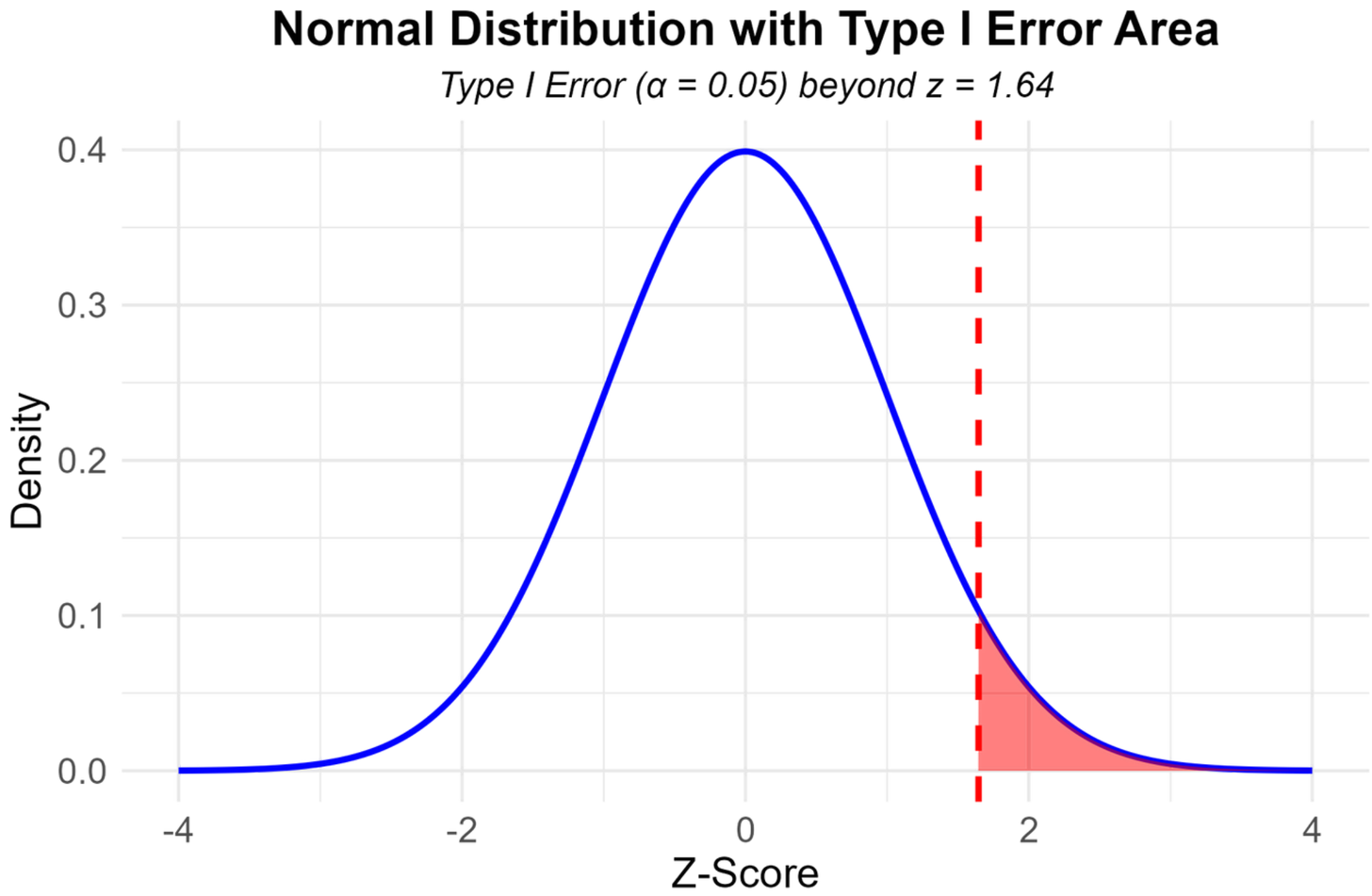
- Why does underestimating standard errors matter?
- Used the Wald Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- z (or t) = $\dfrac{\hat{\beta} - \beta_0}{\mathrm{SE}}$
- If standard errors are underestimated, z goes up
- This can result in type 1 error, and confidence intervals being messed up

Type 1 Error

- Often referred to as α
- False Positive
- Rejecting the null hypothesis when you are not supposed to

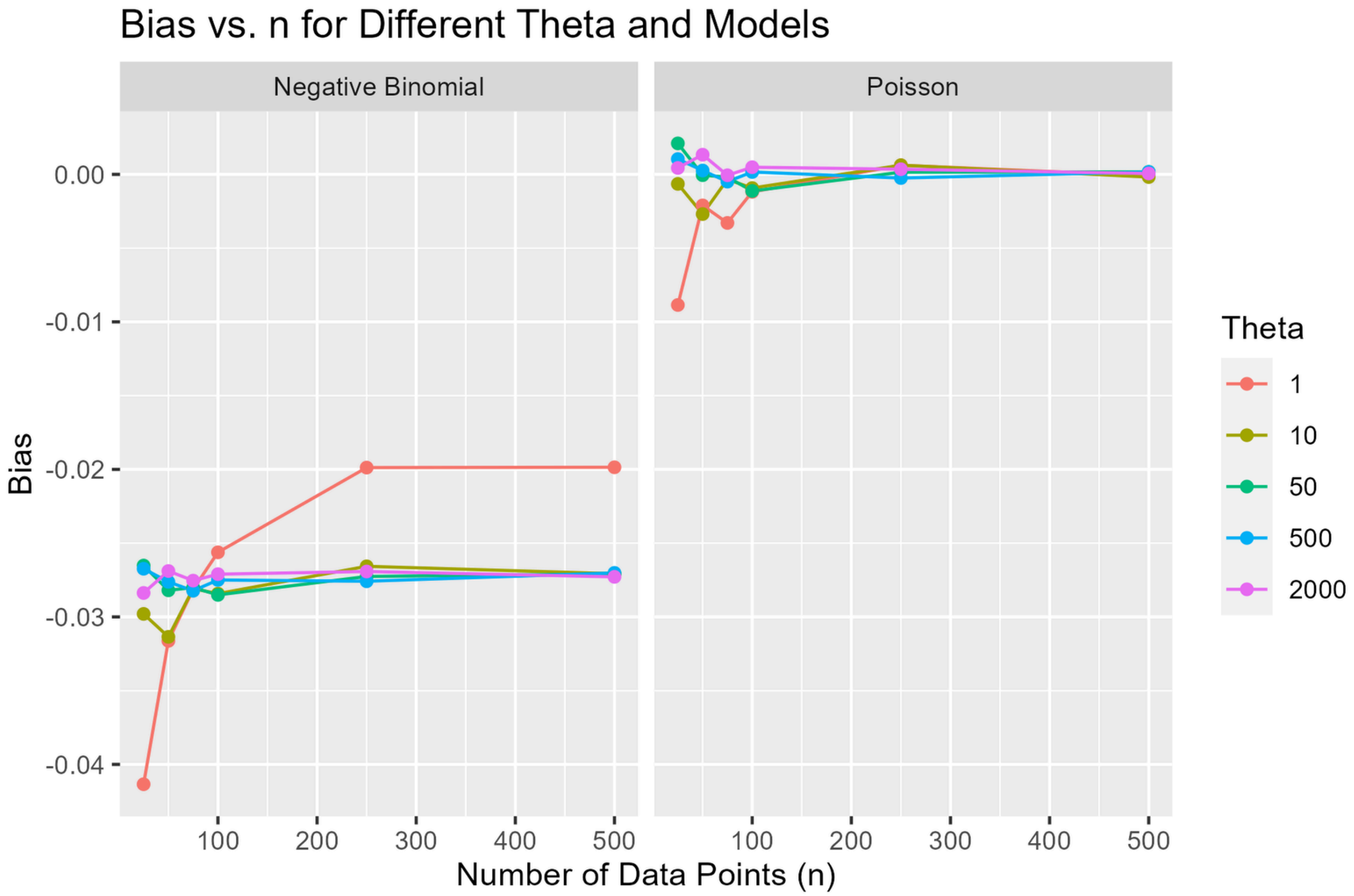


What is AIC?

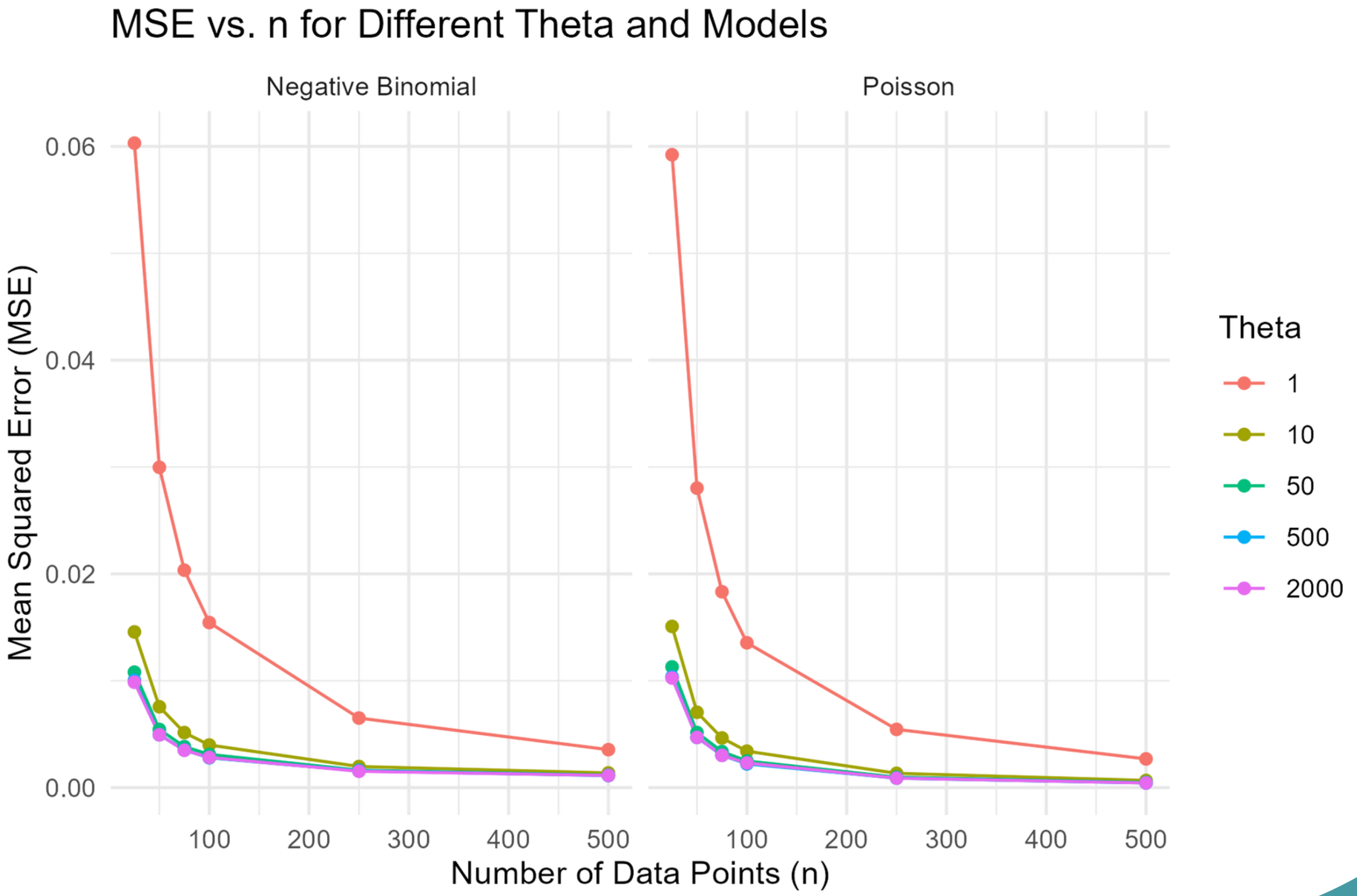
$$\mathrm{AIC} = 2k - 2\ln(L)$$

- Where k is the number of parameters and L is the likelihood of the model
- Measures quality of models
- Uses goodness of fit and parsimony (model simplicity)
- Helps decide on models while taking into account overcomplication
- Lower AIC means better model

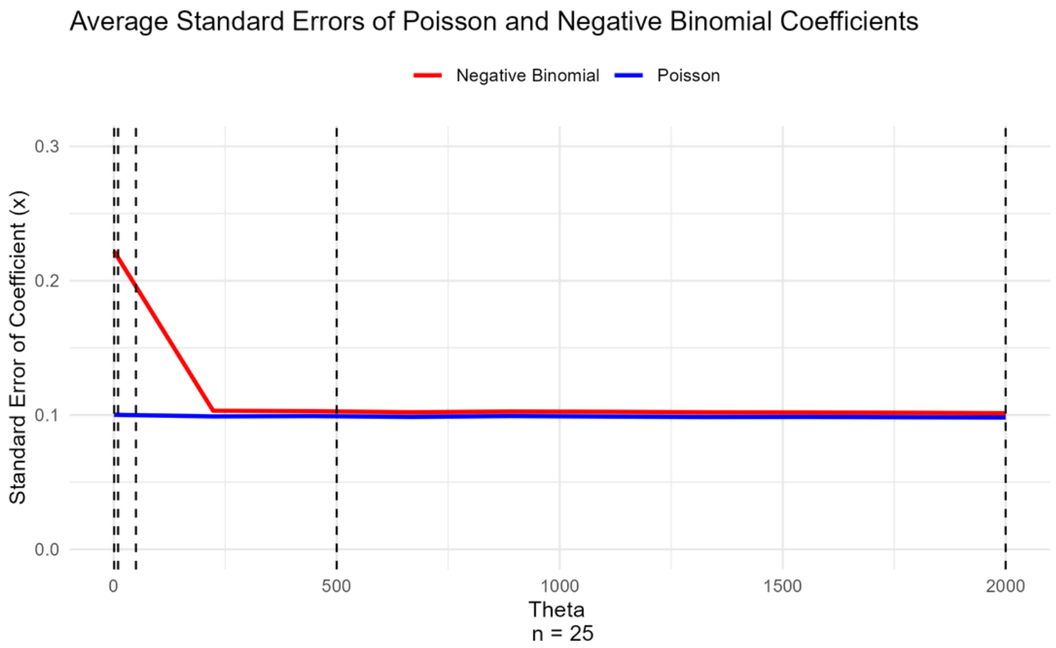
Bias Results

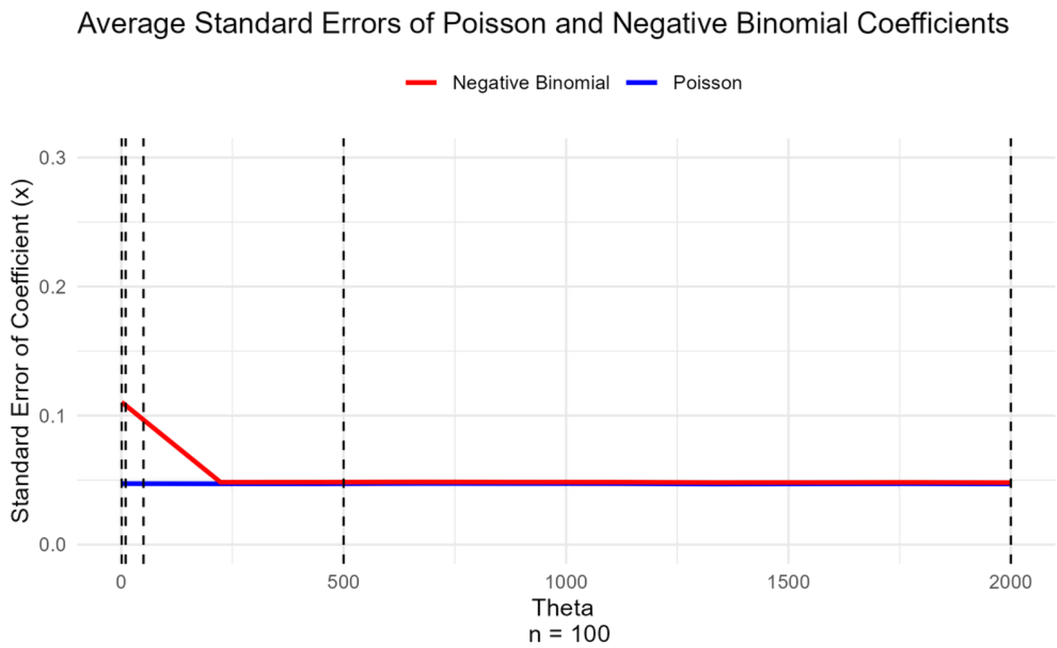


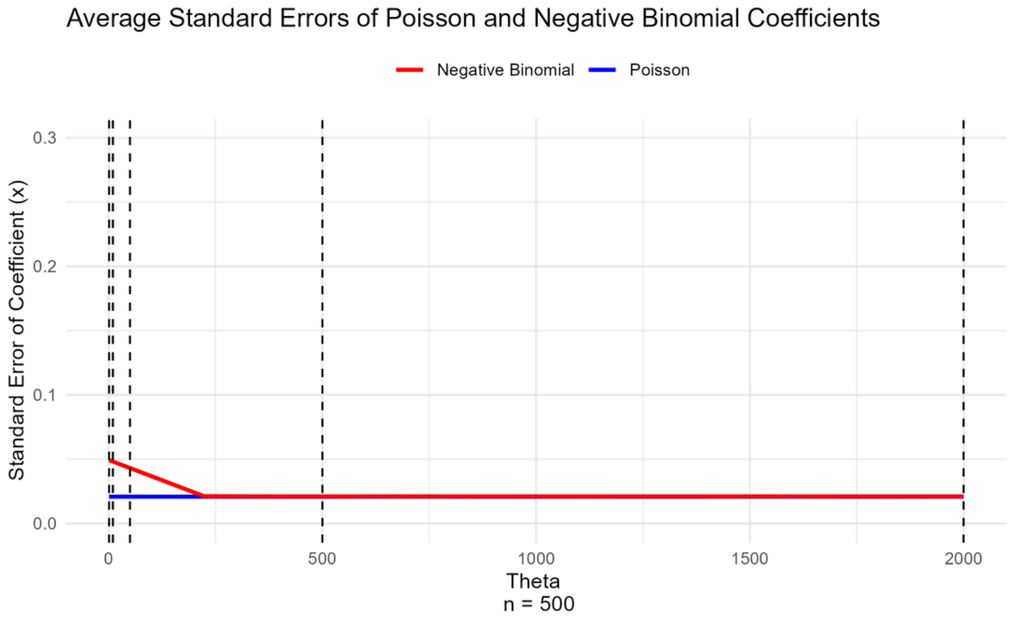
MSE Results


MSE vs. n for Different Theta and Models
Negative Binomial
Poisson
Mean Squared Error (MSE)
0.06
0.04
0.02
0.00
100
200
300
400
500
Number of Data Points (n)
Theta
1
10
50
500
2000

Standard Error Results



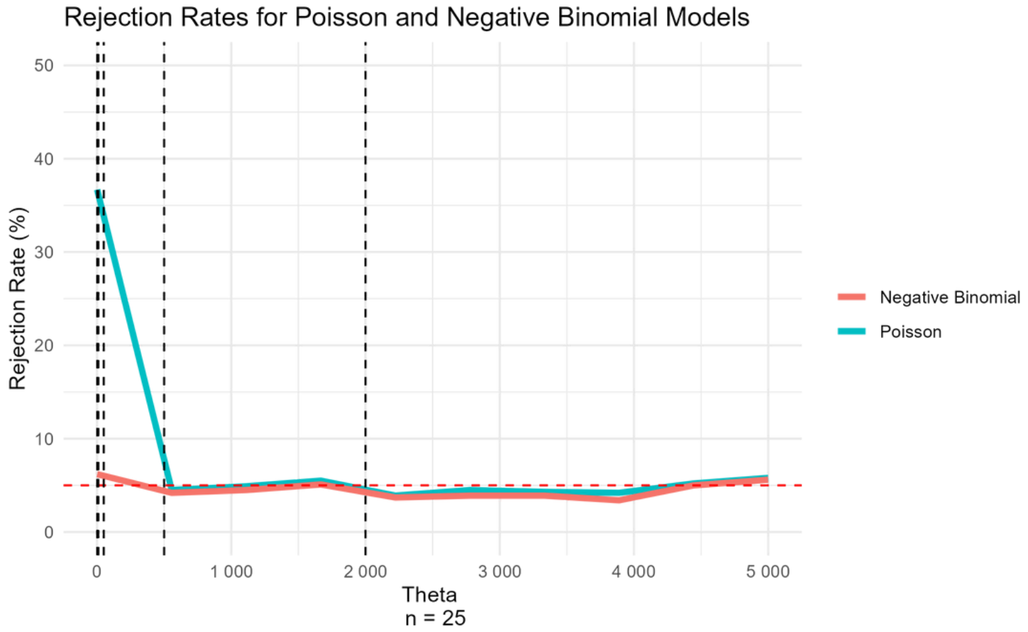


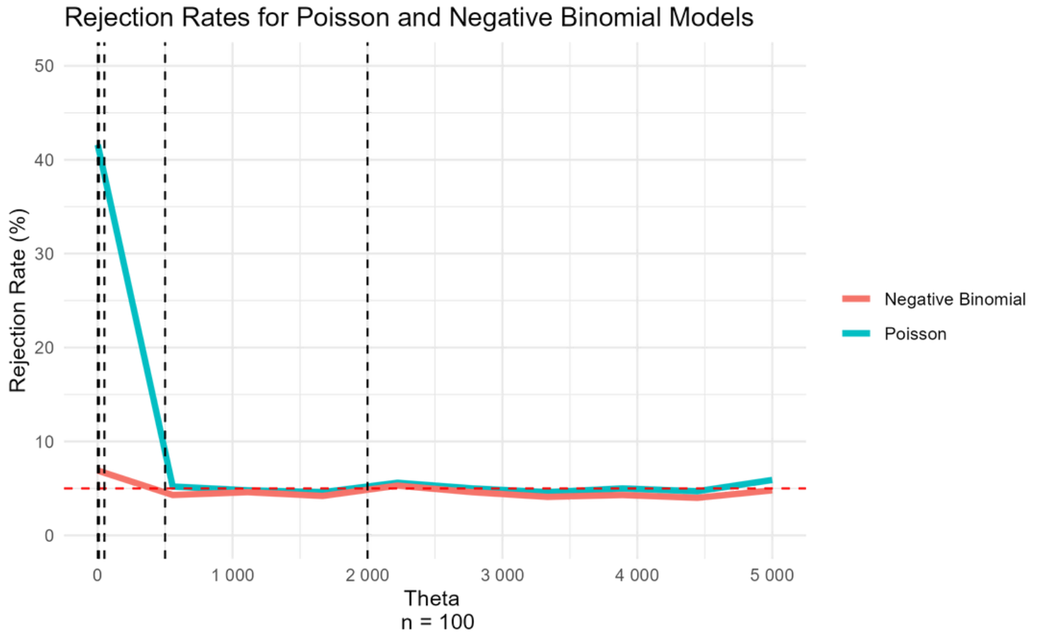


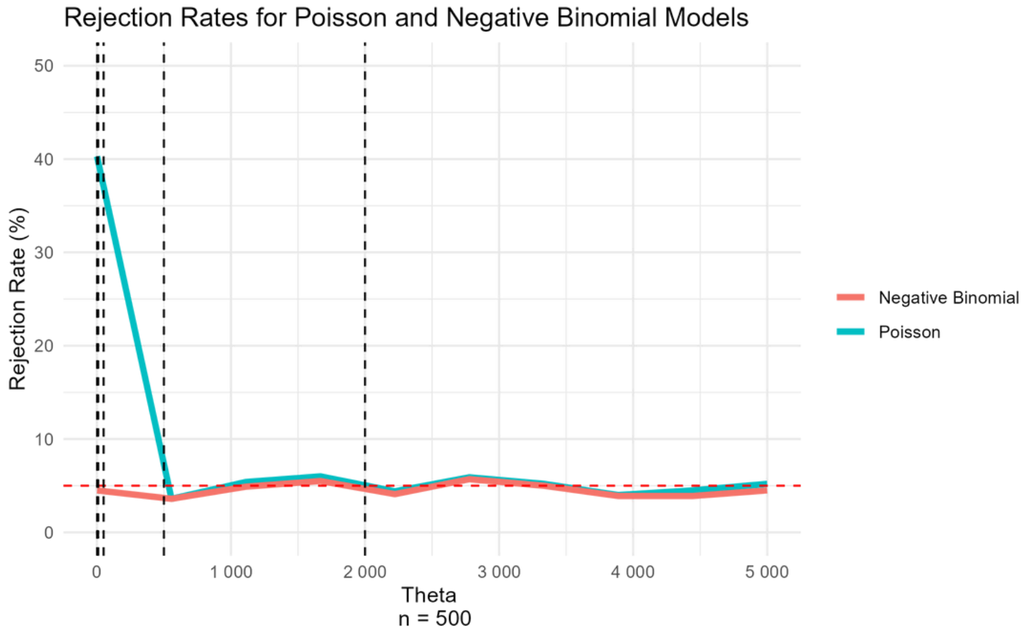
Overdispersed — Not Overdispersed

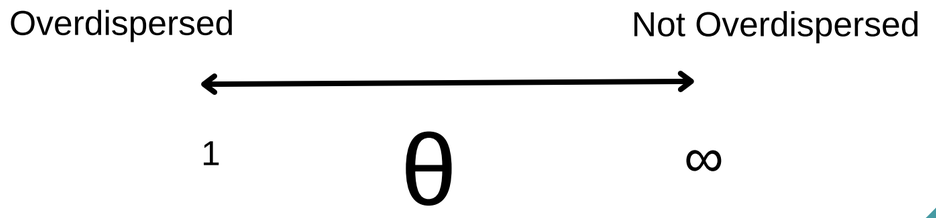
1 — θ — ∞

Rejection Rate Results

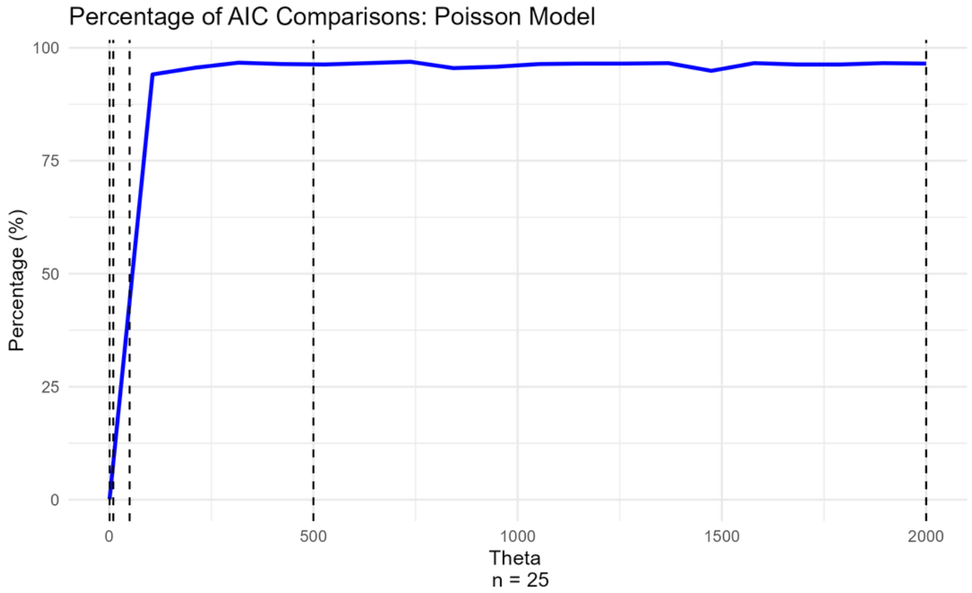


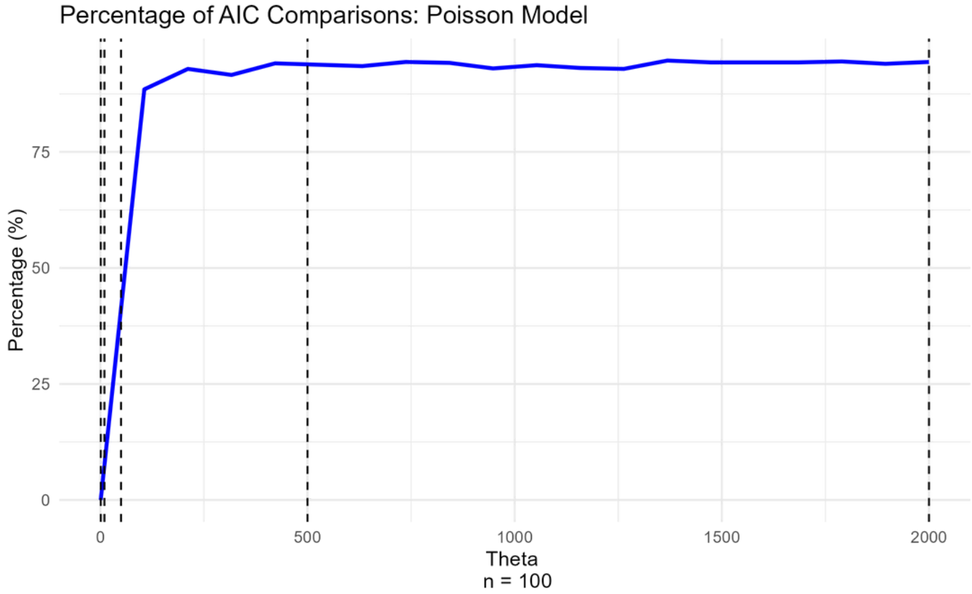


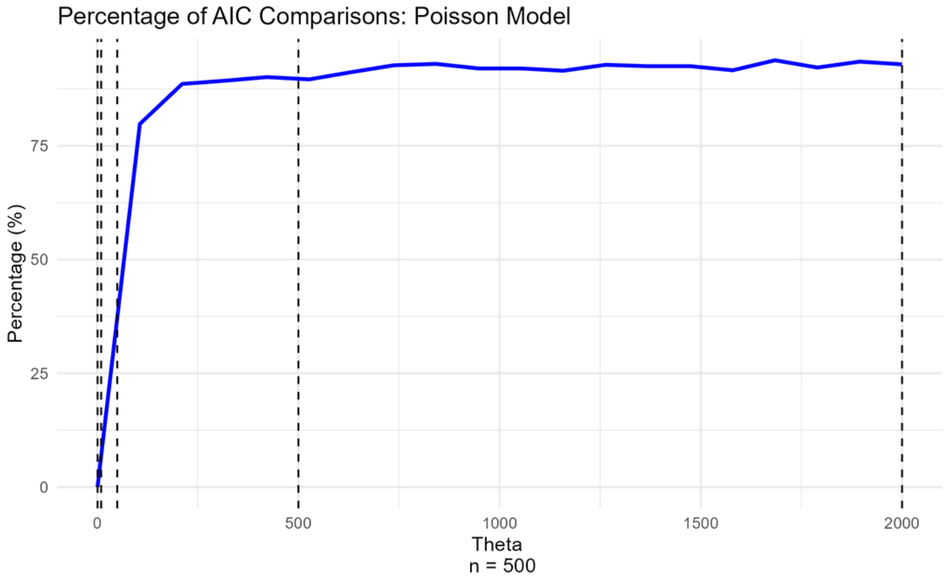




AIC Results







Overdispersed — Not Overdispersed

$1 \quad \theta \quad \infty$

Conclusion

Our results did agree with the theory.

At high levels of overdispersion:

- Standard errors of Poisson were underestimated
- Resulted in Type 1 error
- Which results in confidence interval being messed up

At the case of $\mu \cong \sigma^2$ (θ= 500)

- We start to see convergence
- Which would make Poisson viable
- In this case it might be better to use Poisson since it is easier to handle

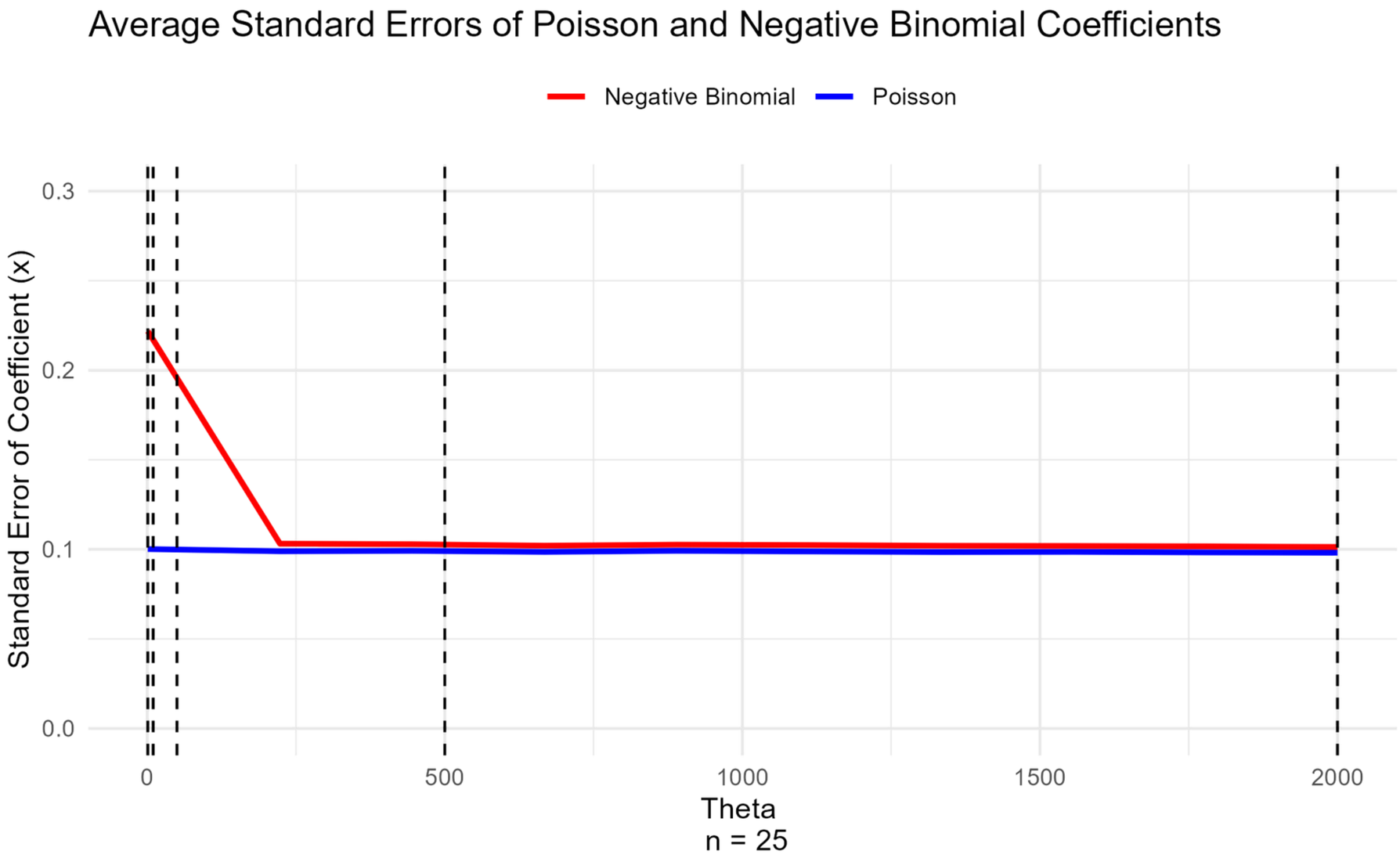
Further Research

- Using real world parameters for simulation study
- Comparing Negative Binomial vs. Quasi-Poisson Model
- Look into the impact of zero inflation

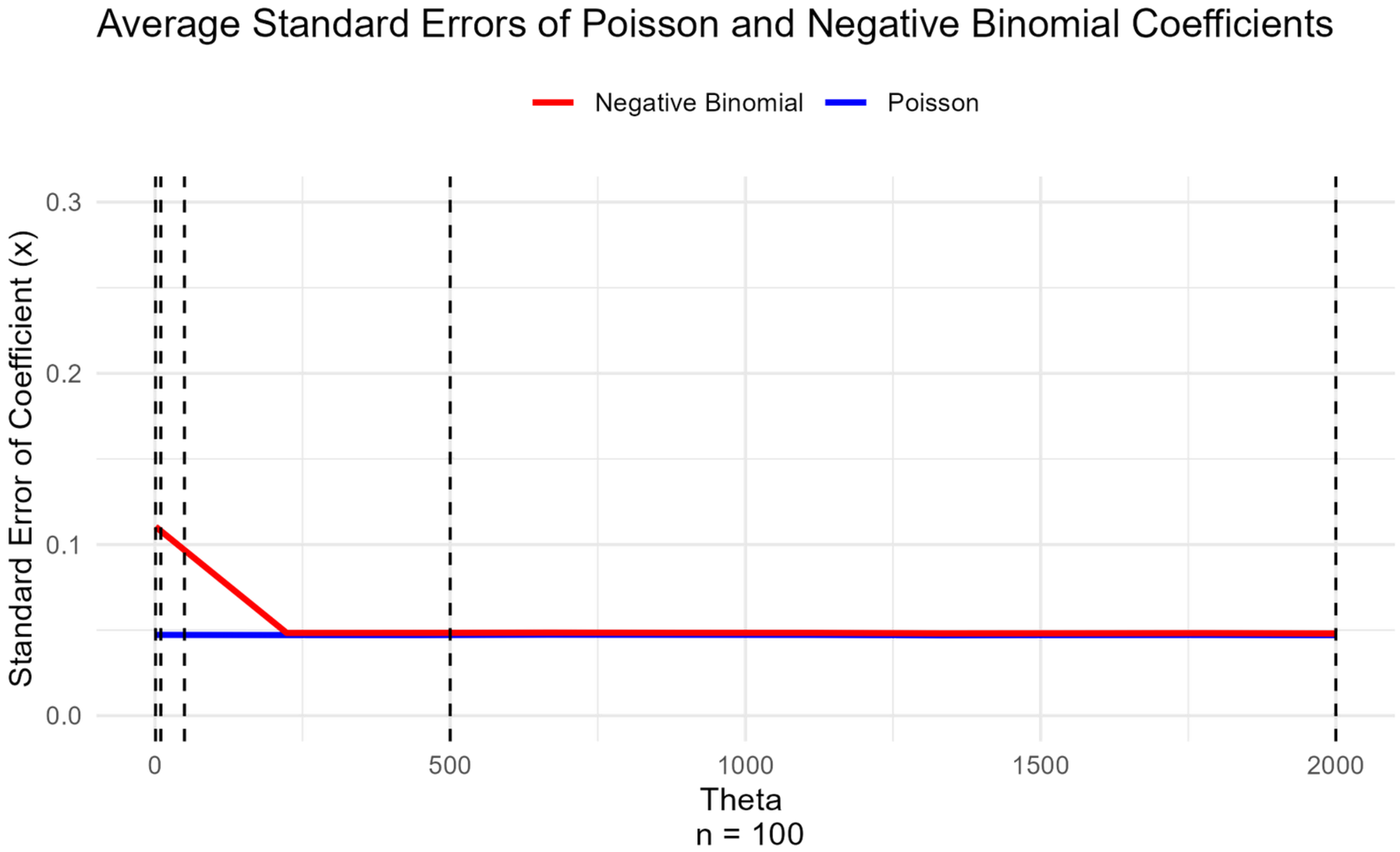
Thank you!

Questions?

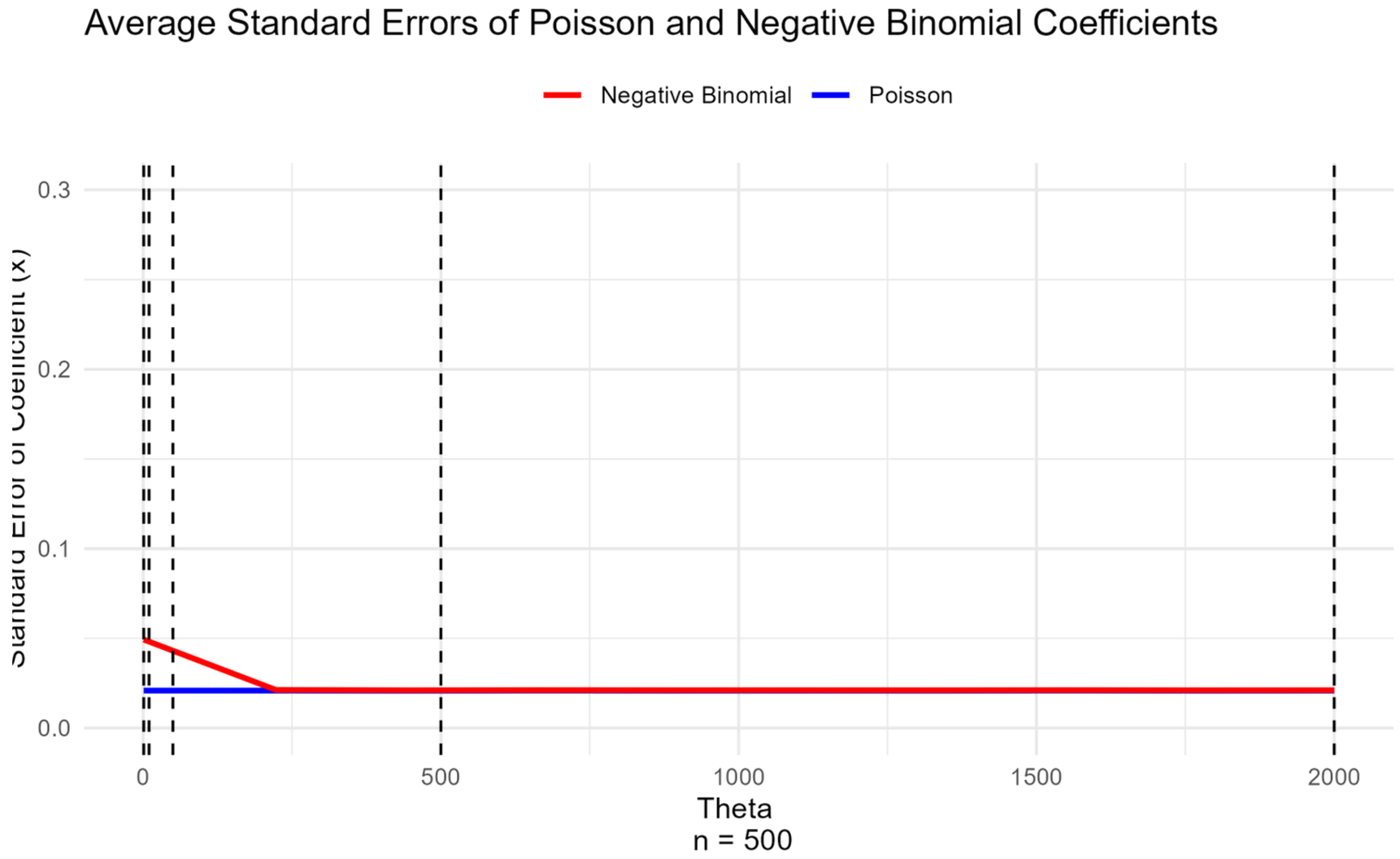
Standard Error: n = 25 Graph



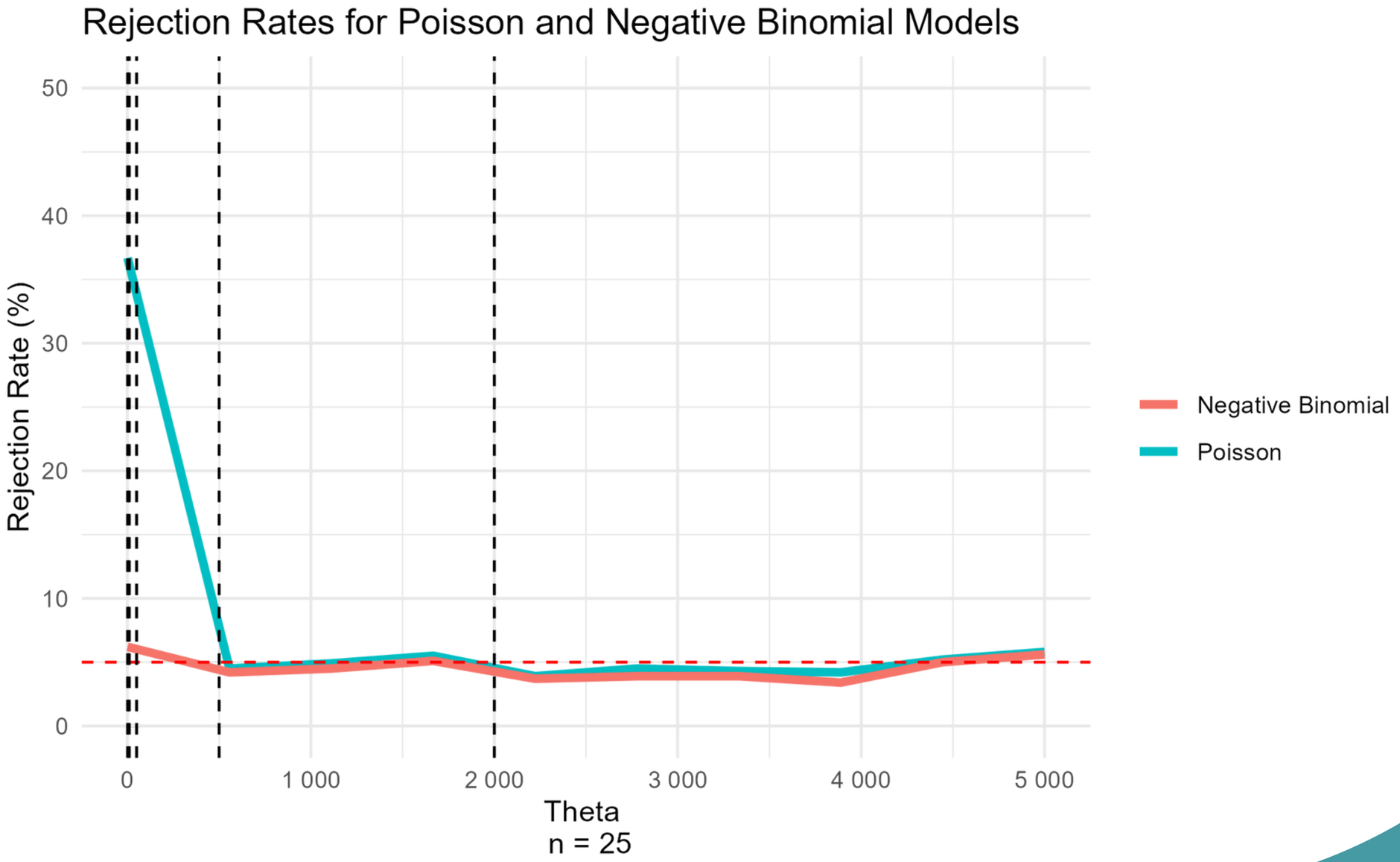
Standard Error: n = 100 Graph



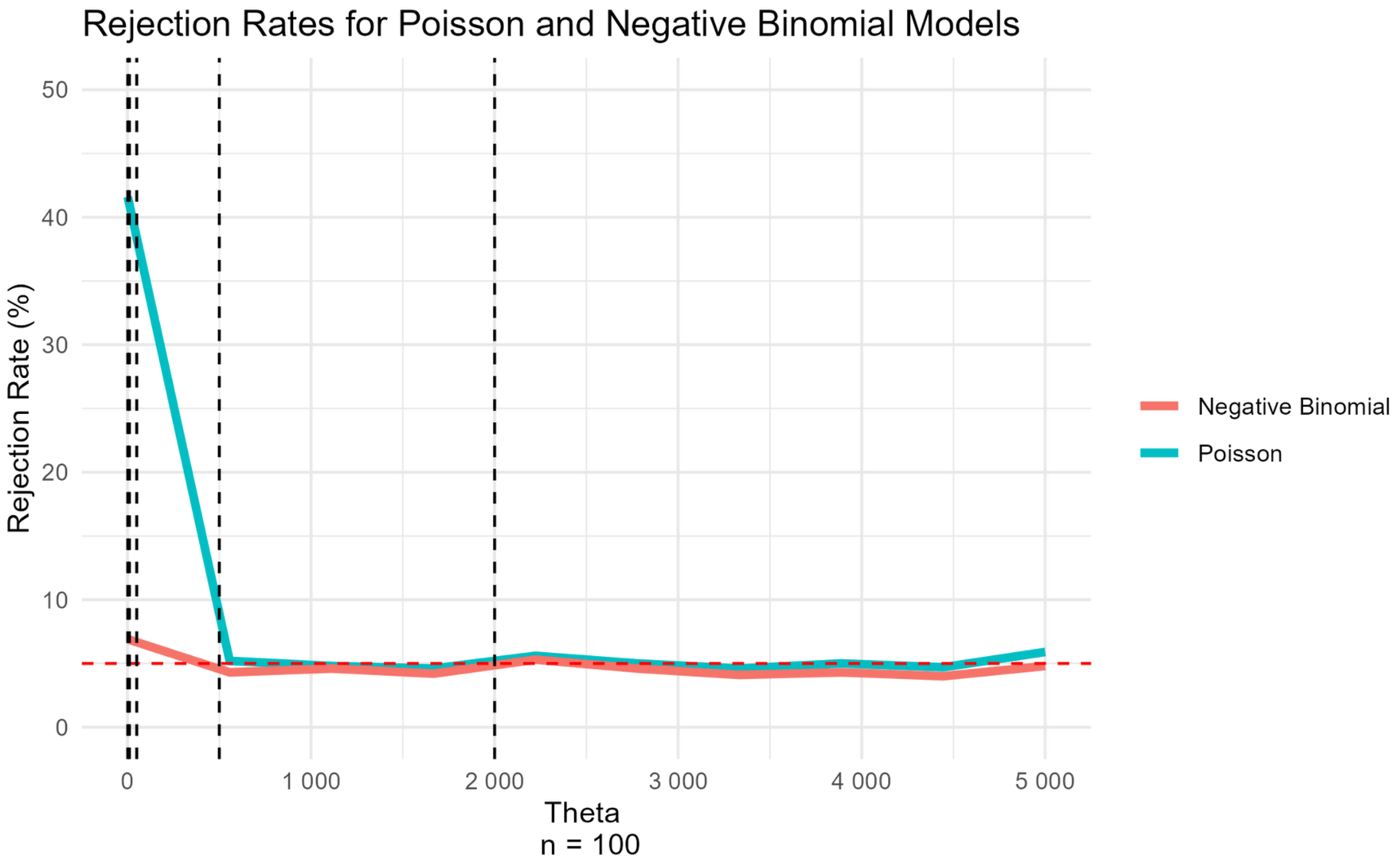
Standard Error: n = 500 Graph



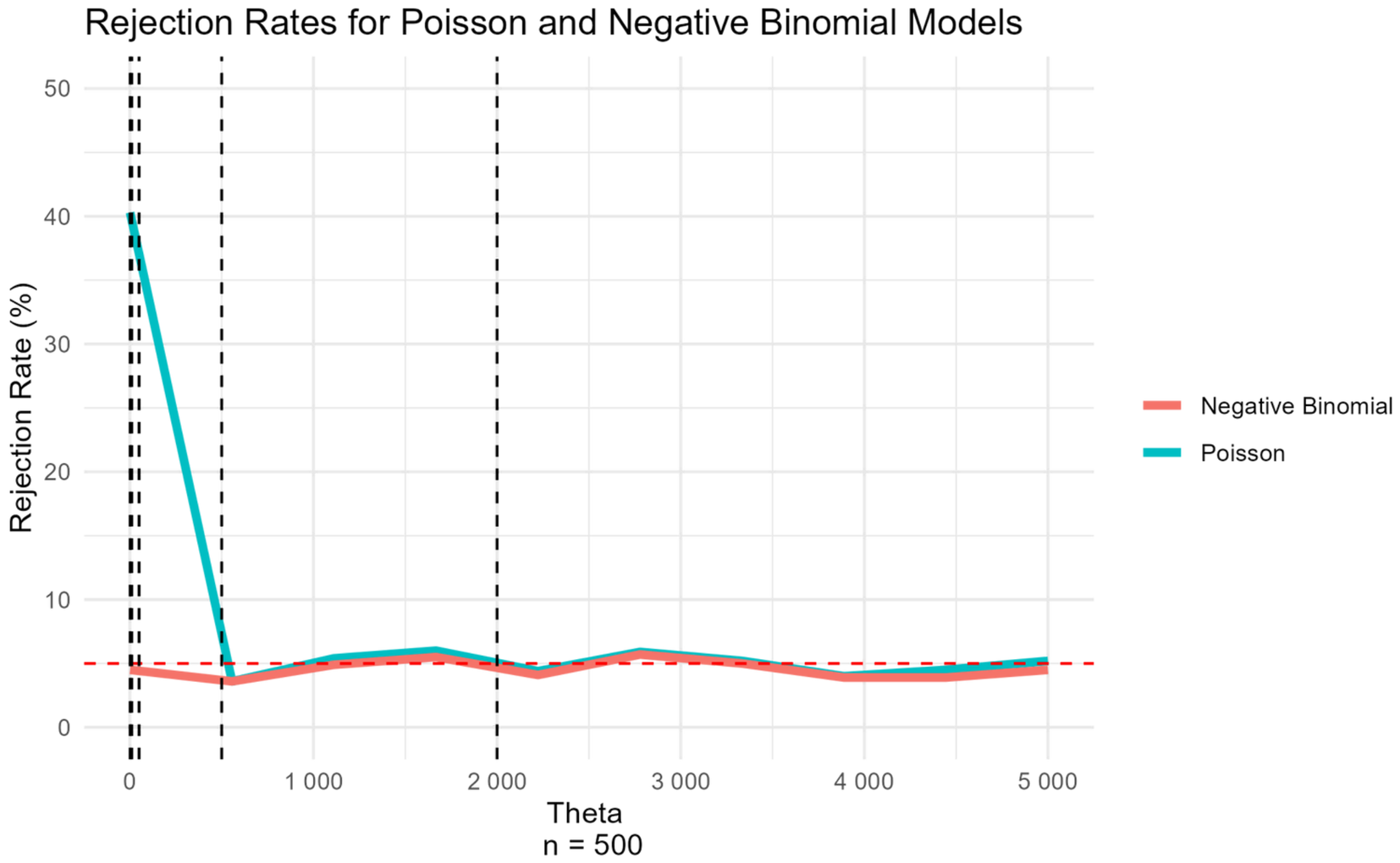
Rejection Rate: n = 25 Graph



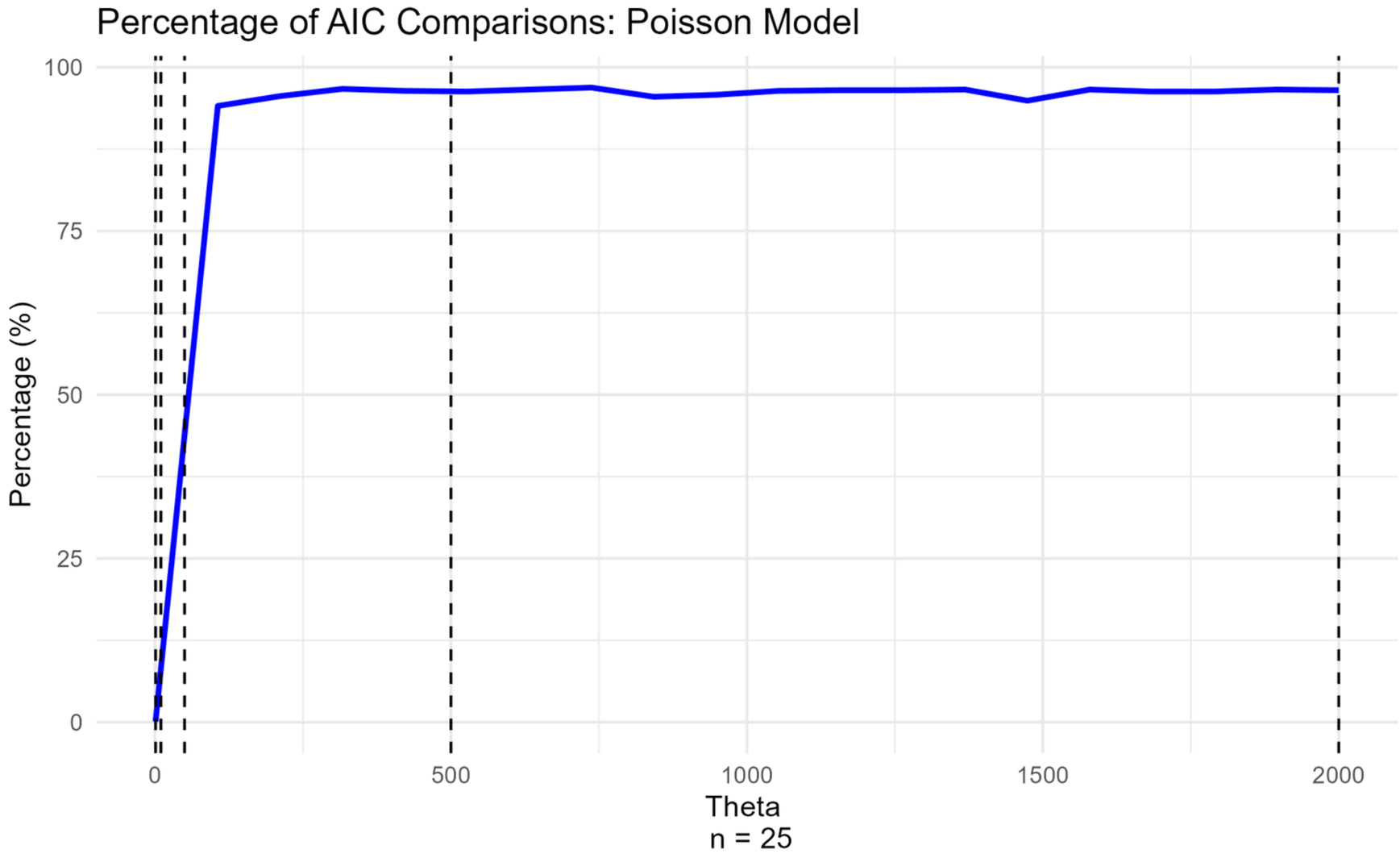
Rejection Rate: n = 100 Graph



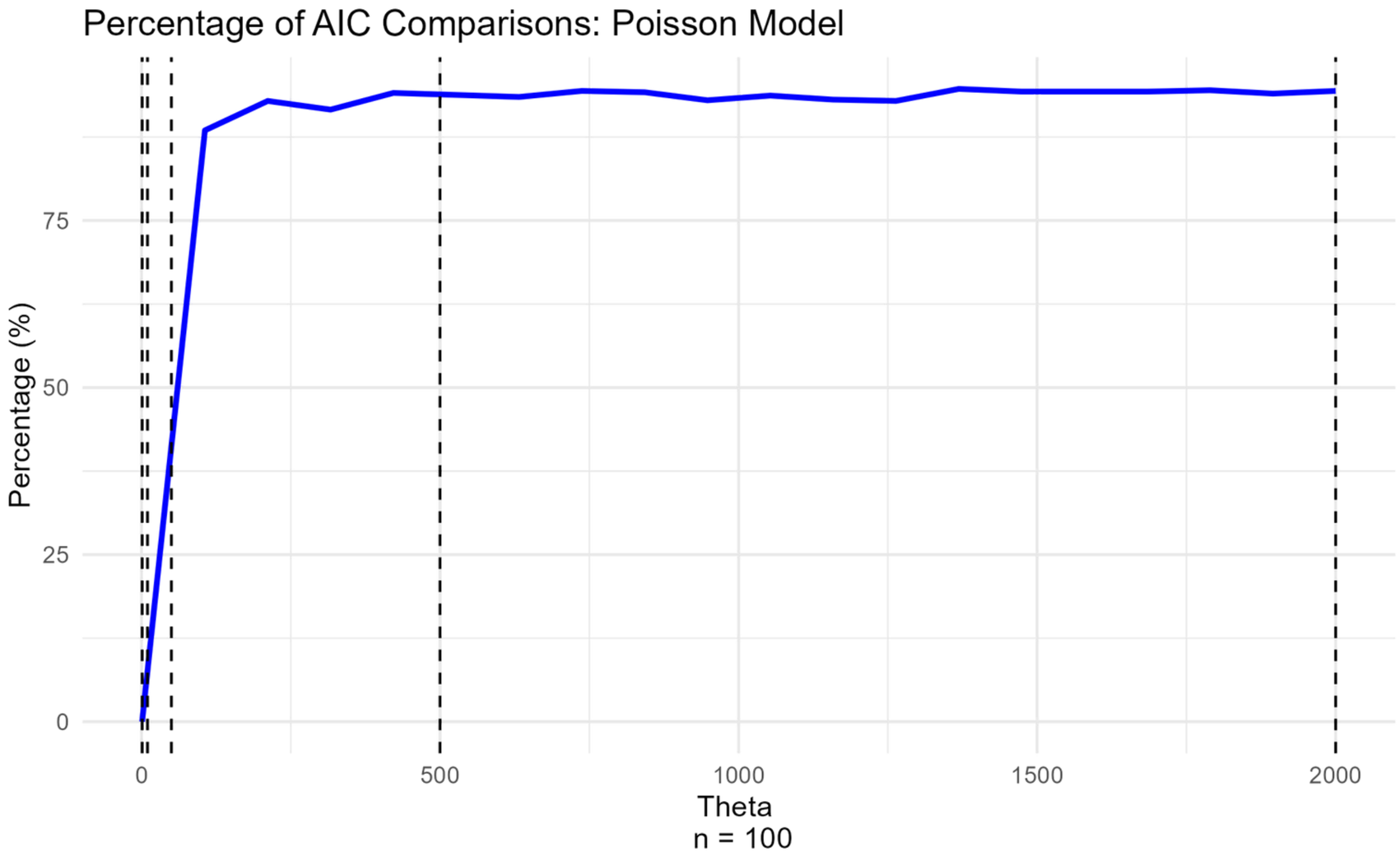
Rejection Rate: n = 500 Graph



AIC : n = 25 Graph



AIC : n = 100 Graph



AIC : n = 500 Graph