

Spring 2018

# Score Test and Likelihood Ratio Test for Zero-Inflated Binomial Distribution and Geometric Distribution

Xiaogang Dai

Western Kentucky University, xiaogang.dai692@topper.wku.edu

Follow this and additional works at: <https://digitalcommons.wku.edu/theses>



Part of the [Applied Statistics Commons](#), [Other Applied Mathematics Commons](#), and the [Probability Commons](#)

---

## Recommended Citation

Dai, Xiaogang, "Score Test and Likelihood Ratio Test for Zero-Inflated Binomial Distribution and Geometric Distribution" (2018). *Masters Theses & Specialist Projects*. Paper 2447.  
<https://digitalcommons.wku.edu/theses/2447>

This Thesis is brought to you for free and open access by TopSCHOLAR®. It has been accepted for inclusion in Masters Theses & Specialist Projects by an authorized administrator of TopSCHOLAR®. For more information, please contact [topscholar@wku.edu](mailto:topscholar@wku.edu).

SCORE TEST AND LIKELIHOOD RATIO TEST FOR ZERO-INFLATED  
BINOMIAL DISTRIBUTION AND GEOMETRIC DISTRIBUTION

A Thesis  
Presented to  
The Faculty of the Department of Mathematics  
Western Kentucky University  
Bowling Green, Kentucky

In Partial Fulfillment  
Of the Requirements for the Degree  
Master of Science

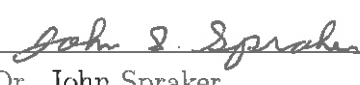
By  
Xiaogang Dai  
May 2018

SCORE TEST AND LIKELIHOOD RATIO TEST FOR ZERO-INFLATED  
BINOMIAL DISTRIBUTION AND GEOMETRIC DISTRIBUTION

Date Recommended 04/12/2018

  
\_\_\_\_\_  
Dr. Ngoc Nguyen, Director of Thesis

  
\_\_\_\_\_  
Dr. Dominic Lanphier

  
\_\_\_\_\_  
Dr. John Spraker

 4/20/18  
\_\_\_\_\_  
Dean, Graduate Studies and Research Date

## DEDICATION

I dedicate this thesis to my family, my friends, and the Western Kentucky University  
Mathematics Department.

## ACKNOWLEDGMENTS

I would like to thank my director of thesis, Dr. Nguyen for her guidance and support throughout my thesis project. Without her assistant and guidance, I would not have been able to successfully complete this thesis.

I would also like to thank Dr. Lanphier and Dr. Spraker, who have friendly agreed to be on my committee. I would like to thank the mathematics department of Western Kentucky University.

I would like to thank my parents for their understanding and encouragement and my friends, who gave the sincere support during the completion of my project.

## CONTENTS

List of Figures	vi
List of Tables	vii
ABSTRACT	viii
Chapter 1. Introduction	1
Chapter 2. Geometric distribution	3
2.1. Likelihood ratio test	3
2.2. Score test	5
Chapter 3. Zero-inflated binomial distribution	7
3.1. Likelihood ratio test	7
3.2. Score test	9
Chapter 4. Score test statistics for the zero-inflated binomial distribution	12
Chapter 5. Type I error and type II error	15
5.1. Type I error for the geometric distribution	15
5.2. Type II error for the geometric distribution	17
5.3. Type I error for the zero-inflated binomial distribution	20
5.4. Type II error for the zero-inflated binomial distribution	26
Chapter 6. Conclusion	30
Appendix	32
BIBLIOGRAPHY	59

## LIST OF FIGURES

1. The graph of the ratio of the two score test statistics for the zero-inflated binomial distribution(1000, 10, 0.5, 0.2), $z_I/z_O$ , plotted versus $n_0$ .	12
2. The graph of the ratio of the two score test statistics for the zero-inflated binomial distribution(1000, 10, 0.1, 0.2), $z_I/z_O$ , plotted versus $n_0$ .	14
3. Type I error for the geometric distribution.	16
4. Type II error with $n = 100, p_0 = 0.1$ for the geometric distribution.	17
5. Type II error with $n = 100, p_0 = 0.3$ for the geometric distribution.	18
6. Type II error with $p_0 = 0.05$ for the geometric distribution.	19
7. Type II error with $p_0 = 0.15$ for the geometric distribution.	19
8. Type II error with $p_0 = 0.2$ for the geometric distribution.	20
9. Type I error with $m = 10$ and $p = 0.1$ for zero-inflated binomial distribution.	21
10. Type I error with $m = 10$ and $p = 0.3$ for zero-inflated binomial distribution.	22
11. Type I error with $m = 20$ and $p = 0.1$ for zero-inflated binomial distribution.	22
12. Type I error with $m = 20$ and $p = 0.3$ for zero-inflated binomial distribution.	23
13. Type I error with $m = 20$ and $p = 0.3$ with larger sample size.	24
14. Type I error with $n = 1000$ and $m = 10$ .	25
15. Type II error with $m = 10, p = 0.1$ and $w = 0.2$ .	26
16. Type II error with $n = 1000, m = 10$ and $w = 0.2$ .	27
17. Type II error with $n = 1000, m = 10$ and $p = 0.1$ .	28

## LIST OF TABLES

1. Data from a zero-inflated binomial distribution with $m = 10, p = 0.5, w = 0.2$ .	12
2. Data from a zero-inflated binomial distribution with $m = 10, p = 0.1, w = 0.2$ .	13
3. Types of error.	15
4. The percentage of time that score test cannot be applied.	29



# SCORE TEST AND LIKELIHOOD RATIO TEST FOR ZERO-INFLATED BINOMIAL DISTRIBUTION AND GEOMETRIC DISTRIBUTION

Xiaogang Dai

59 Pages

Directed by: Dr. Ngoc Nguyen, Dr. Dominic Lanphier, Dr. John Spraker

Department of Mathematics

Western Kentucky University

The main purpose of this thesis is to compare the performance of the score test and the likelihood ratio test by computing type I errors and type II errors when the tests are applied to the geometric distribution and inflated binomial distribution. We first derive test statistics of the score test and the likelihood ratio test for both distributions. We then use the software package R to perform a simulation to study the behavior of the two tests. We derive the R codes to calculate the two types of error for each distribution. We create lots of samples to approximate the likelihood of type I error and type II error by changing the values of parameters.

In the first chapter, we discuss the motivation behind the work presented in this thesis. Also, we introduce the definitions used throughout the paper. In the second chapter, we derive test statistics for the likelihood ratio test and the score test for the geometric distribution. For the score test, we consider the score test using both the observed information matrix and the expected information matrix, and obtain the score test statistic  $z_O$  and  $z_I$ .

Chapter 3 discusses the likelihood ratio test and the score test for the inflated binomial distribution. The main parameter of interest is  $w$ , so  $p$  is a nuisance parameter in this case. We derive the likelihood ratio test statistics and the score test statistics to test  $w$ . In both tests, the nuisance parameter  $p$  is estimated using maximum likelihood estimator  $\hat{p}$ . We also consider the score test using both the observed and the expected information matrices.

Chapter 4 focuses on the score test in the inflated binomial distribution. We generate data to follow the zero inflated binomial distribution by using the package R. We plot the graph of the ratio of the two score test statistics for the sample data,  $z_I/z_O$ , in terms of different values of  $n_0$ , the number of zero values in the sample.

In chapter 5, we discuss and compare the use of the score test using two types of information matrices. We perform a simulation study to estimate the two types of errors when applying the test to the geometric distribution and the inflated binomial distribution. We plot the percentage of the two errors by fixing different parameters, such as the probability  $p$  and the number of trials  $m$ .

Finally, we conclude by briefly summarizing the results in chapter 6.

# CHAPTER 1

## INTRODUCTION

In statistics, a hypothesis is a statement about the numerical value of a population parameter. There are two types of hypotheses. The null hypothesis, denoted  $H_0$ , shows the hypothesis that will not be rejected unless the data provide convincing evidence that it is false. The alternative hypothesis, denoted  $H_1$  or  $H_a$ , shows the hypothesis that will be accepted only if the data provide convincing evidence of its truth. A test statistic is a sample statistic used in statistical hypothesis testing. A test statistic is computed based on a random sample drawn from a population and it measures the difference between the null hypothesis and what is observed in the sample. Generally, an extreme value of a test statistic indicates strong evidence against the null hypothesis. In hypothesis testing, a critical value is a value on the test distribution under the assumption that  $H_0$  is true, that is compared to the test statistic to determine whether to reject the null hypothesis. A type I error is an incorrect rejection of a true null hypothesis, while a type II error is incorrect acceptance of a false null hypothesis.

The score test, also known as Rao's score test, is a statistical test of a simple null hypothesis that a parameter of interest  $\theta$  is equal to some particular value  $\theta_0$ . We have the null hypothesis  $H_0 : \theta = \theta_0$ , and the alternative hypothesis  $H_1 : \theta \neq \theta_0$ . Let  $X = (X_1, \dots, X_n)$  be an independent and identically distributed sample from a probability density function  $p(x, \theta)$ , where parameter  $\theta = (\theta_1, \dots, \theta_r)^T$ . Let

$$P(X; \theta) = p(x_1; \theta) \dots p(x_n; \theta).$$

The score vector of Fisher is

$$S(\theta) = [s_1(\theta), \dots, s_r(\theta)]^T \quad \text{where } s_j(\theta) = \frac{1}{P} \frac{\partial P}{\partial \theta_j}; \quad j = 1, \dots, r.$$

The Fisher information matrix of order  $r \times r$  is defined by

$$I(\theta) = (i_{jk}(\theta)), \quad \text{where } i_{jk}(\theta) = E(s_j(\theta)s_k(\theta)).$$

Then the score test under the null hypothesis  $H_0 : \theta = \theta_0$  gives

$$RSS = S(\theta_0)^T [I(\theta_0)]^{-1} S(\theta_0).$$

The distribution of the score test statistic is a Chi-square distribution with 1 degree of freedom under the assumption that  $H_0$  is true.

The likelihood function is a function of the unknown parameter  $\theta$  given the data  $X$ .

$$P(X; \theta) = p(x_1; \theta) \dots p(x_n; \theta) = L(\theta; X).$$

We take the natural logarithm of the likelihood function, which is log-likelihood function:

$$\log L(\theta; X) = \sum_{i=1}^n \log P_i(X_i; \theta).$$

A likelihood ratio test is a statistical test used for comparing the goodness of fit of two statistical models, the null hypothesis  $H_0 : \theta = \theta_0$  and alternative hypothesis  $H_1 : \theta \neq \theta_0$ . The likelihood ratio test statistic, denoted by  $\lambda$ , is given by

$$\lambda = -2(L(\theta) - \sup L(\theta)),$$

where  $L(\theta)$  is the log-likelihood function. The distribution of the likelihood ratio test statistic is a chi-square distribution with 1 degree of freedom under the assumption that  $H_0$  is true.

## CHAPTER 2

### GEOMETRIC DISTRIBUTION

In probability theory, the geometric distribution is a discrete probability distribution that is used to model the number of failures until the first success when performing a sequence of Bernoulli trials. We define the probability mass function as follows:

$$Pr(X = i) = (1 - p)^i p, \quad i = 0, 1, 2, 3, \dots \quad (2.1)$$

where the random variable  $X$  denotes the number of failures until the first success.

For example, suppose an ordinary coin is tossed repeatedly until the first time a “Head” appears. The probability distribution of the number of Tails until the first Head is supported on the infinite set  $\{0, 1, 2, 3, \dots\}$  and follows the geometric distribution with  $p = \frac{1}{2}$ .

#### 2.1. Likelihood ratio test

We write  $L(p)$  to denote the log-likelihood function for the geometric distribution. Let  $n_i, i = 0, 1, 2, 3, \dots$ , denote the number of times in the sample that  $X_i = i$ . Given a random sample  $X_1, X_2, \dots, X_n$  from a geometric distribution with probability  $p$ , we can write

$$\begin{aligned} L(p) &= \sum_{i=0}^{\infty} n_i \log\{Pr(X = i)\} \\ &= n_0 \log(p) + \sum_{i=1}^{\infty} n_i \log(1 - p)^i + \log(p) \sum_{i=1}^{\infty} n_i \\ &= n \log(p) + \sum_{i=1}^{\infty} i n_i \log(1 - p), \end{aligned} \quad (2.2)$$

where  $n = \sum_{i=0}^{\infty} n_i$ . After deriving the log-likelihood function of geometric distribution, we perform the likelihood-ratio test. A simple hypothesis test has the models under both

the null and alternative hypotheses, which are expressed as:

$$\begin{aligned} H_0 : p &= p_0 \\ H_1 : p &\neq p_0. \end{aligned} \tag{2.3}$$

Under the null hypothesis, we use  $\lambda$  to denote the natural log of the likelihood ratio.  $\lambda$  is defined as follows:

$$\lambda = -2(L(p_0) - \sup L(p)),$$

where  $L(p)$  is the log-likelihood function, and  $\sup$  is the supremum function. To find  $\sup L(p)$ , we maximize  $L(p)$  in terms of  $p$ , this gives  $\sup L(p) = L(\hat{p})$ , where  $\hat{p} = \frac{n}{n+d}$ ,  $\sum_{i=1}^{\infty} in_i = d$ . The second-order derivative of the log-likelihood function at  $\hat{p}$  is less than 0, which makes sure that we get the maximum value at  $\hat{p}$ . The rules to reject or not reject the null hypothesis are as follows:

If  $\lambda \geq c$ , reject  $H_0$ ;

If  $\lambda < c$ , do not reject  $H_0$ .

The value of  $c$  is calculated using a specified significance level  $\alpha$ .

Substituting  $\hat{p}$  to  $\lambda$  gives

$$\begin{aligned} \lambda &= -2(L(p_0) - L(\hat{p})) \\ &= -2(n \log(p_0) + \sum_{i=1}^{\infty} in_i \log(1 - p_0) - (n \log(\hat{p}) + \sum_{i=1}^{\infty} in_i \log(1 - \hat{p}))). \end{aligned} \tag{2.4}$$

Under the assumption that the null hypothesis is true and that the sample size  $n$  is sufficiently large, the test statistic  $\lambda$  approximately follows a chi-square distribution with one degree of freedom. Comparing  $\lambda$  with the critical value  $\chi_{\alpha}^2(1)$  at a significance level  $\alpha$ , we reject or do not reject  $H_0$ . Later, we discuss type I errors and type II errors in Chapter 5 using Equation (2.4).

## 2.2. Score test

The score test of the null hypothesis in (2.3) has test statistics of the form

$$z = U' J^{-1} U \quad (2.5)$$

where  $\mathbf{U}$  is the scores vector including the derivative of  $L$  with respect to the parameter  $p$ . The matrix  $\mathbf{J}$  is either the observed or the expected information matrix, denoted by  $\mathbf{O}$  and  $\mathbf{I}$ , given below,

$$\mathbf{O} = -\left(\frac{d^2 L}{dp^2}\right)$$

$$\mathbf{I} = \mathbb{E}[\mathbf{O}],$$

and both are evaluated at  $p = p_0$ .

We denote  $\sum_{i=1}^{\infty} n_i = n_+$  and  $\sum_{i=1}^{\infty} in_i = d$ . The first-order and second-order derivatives are given by

$$\frac{dL}{dp} = \frac{n}{p} - \frac{d}{1-p}$$

$$\frac{d^2 L}{dp^2} = -\frac{n}{p^2} - \frac{d}{(1-p)^2}.$$

Hence, the observed information matrix is given by

$$\mathbf{O}(p_0) = \left[ \frac{n}{p_0^2} + \frac{d}{(1-p_0^2)^2} \right].$$

In order to derive the expected information matrix, we can apply the facts that  $\mathbb{E}[n_0] = np$ ,  $\mathbb{E}[d] = \sum_{i=1}^{\infty} in(1-p)^i p = \frac{n(1-p)}{p}$ , and  $\mathbb{E}[n_+] = n - np$ . The expected information matrix is

$$\mathbb{E}[\mathbf{O}] = \left[ \frac{n}{p^2} + \frac{n}{p(1-p)} \right].$$

Therefore, from Equation (2.5), the two score test statistics are given by

$$z_O = \left(\frac{n}{p_0} - \frac{d}{1-p_0}\right) \left(-\frac{n}{p_0^2} - \frac{d}{(1-p_0)^2}\right)^{-1} \left(\frac{n}{p_0} - \frac{d}{1-p_0}\right)$$

$$z_I = \left(\frac{n}{p_0} - \frac{d}{1-p_0}\right) \left(\frac{n}{p_0^2} + \frac{n}{p_0(1-p_0)}\right)^{-1} \left(\frac{n}{p_0} - \frac{d}{1-p_0}\right).$$

We simplify the two score test statistics as:

$$z_O = \frac{(nq_0 - dp_0)^2}{nq_0^2 + dp_0^2}$$

$$z_I = \frac{(nq_0 - dp_0)^2}{nq_0},$$
(2.6)

where  $p_0$  is the hypothesized value of  $p$  under  $H_0$  and  $q_0 = 1 - p_0$ .



## CHAPTER 3

### ZERO-INFLATED BINOMIAL DISTRIBUTION

In probability, a zero-inflated binomial distribution is a distribution that allows for frequent zero-value observations and follows the binomial distribution as well. We define the probability mass function of the zero-inflated binomial distribution (ZIB) as follows,

$$\begin{aligned} Pr(X = 0) &= w + (1 - w)(1 - p)^m \\ Pr(X = i) &= (1 - w) \binom{m}{i} p^i (1 - p)^{m-i} \quad i = 1, 2, 3, \dots, m, \end{aligned} \quad (3.1)$$

where the random variable  $X$  denotes the number of successes in a sequence of  $m$  independent Bernoulli trials,  $w (> 0)$  is the zero-inflation probability.

#### 3.1. Likelihood ratio test

We use  $L(w, p)$  to denote the log-likelihood for a zero-inflated binomial distribution, and  $n_i, i = 0, 1, \dots, m$  denotes the number of times that  $X = i$  in  $m$  independent Bernoulli trials. The log-likelihood function of a zero-inflated binomial distribution given a random sample  $X_1, X_2, \dots, X_n$  from  $ZIB(w, m, p)$  is:

$$\begin{aligned} L(w, p) &= \sum_{i=0}^m n_i \log\{Pr(X = i)\} \\ &= n_0 \log(w + (1 - w)(1 - p)^m) + \sum_{i=1}^m n_i \log(1 - w) + \sum_{i=1}^m n_i \log \binom{m}{i} \\ &\quad + \sum_{i=1}^m i n_i \log(p) + \sum_{i=1}^m m n_i \log(1 - p) - \sum_{i=1}^m i n_i \log(1 - p) \end{aligned} \quad (3.2)$$

Now, we perform the likelihood ratio test for a zero-inflated binomial distribution to test the hypothesis:  $H_0 : w = 0$  and  $H_1 : w \neq 0$ .

In statistics, a nuisance parameter is any parameter which inference is not desired, but which needs to be accounted for in the analysis of those parameters which are of interest. In this situation,  $p$  is a nuisance parameter that is estimated by the maximum

likelihood estimator  $\hat{p}$  and we also derive the maximum likelihood estimator for  $w$  to be used in the likelihood ratio test. The first-order derivatives of  $L(w, p)$  are given by

$$\begin{aligned}\frac{\partial L}{\partial w} &= \frac{n_0(1-q^m)}{w+(1-w)q^m} - \frac{n-n_0}{1-w} \\ \frac{\partial L}{\partial p} &= -\frac{mn_0(1-w)q^{m-1}}{w+(1-w)q^m} + \frac{d}{p} - \frac{mn_+}{q} + \frac{d}{q},\end{aligned}\tag{3.3}$$

where  $q = 1-p$ ,  $n_+ = \sum_{i=1}^m n_i$ ,  $n = n_0 + n_+$  and  $d = \sum_{i=1}^m in_i$ . Under  $H_0$ , we plug  $w = 0$  in Equation (3.3) and let the first-order derivatives equal 0 to obtain the maximum estimator  $\hat{p}_1$

$$\left. \frac{\partial L}{\partial p} \right|_{w=0} = \frac{-mn + d}{1-p} + \frac{d}{p} = 0.$$

Solving this equation, we get  $\hat{p}_1 = \frac{d}{mn}$ . The second-order derivative of the log-likelihood function at  $\hat{p}$  when  $w = 0$  is less than 0, which gives that the log-likelihood function at  $\hat{p}$  achieves the maximum value.

Under  $H_1$ , we need to maximize  $L(w, p)$  with respect to both  $w$  and  $p$ . The maximum likelihood estimators under  $H_1$  are denoted by  $\hat{w}$  and  $\hat{p}_2$ . To obtain  $\hat{w}$  and  $\hat{p}_2$ , we set both  $\frac{\partial L}{\partial p} = 0$  and  $\frac{\partial L}{\partial w} = 0$ . After some steps of calculation, the maximum estimator  $\hat{w}$  and  $\hat{p}_2$  satisfy the system of equations below

$$\begin{aligned}\hat{w} &= \frac{n_0 - n\hat{q}_2^m}{n - n\hat{q}_2^m} \\ \hat{p}_2 &= \frac{d\hat{w} + (1-\hat{w})d\hat{q}_2^m}{m\hat{w}(n-n_0) - mn(1-\hat{w})\hat{q}_2^m},\end{aligned}\tag{3.4}$$

where  $\hat{q}_2 = 1 - \hat{p}_2$ . We use  $\lambda$  to denote the natural log of the likelihood ratio, which is defined by

$$\lambda = -2(L(0, \hat{p}_1) - L(\hat{w}, \hat{p}_2)).$$

Substituting  $w = 0$ ,  $\hat{p}_1$ ,  $\hat{w}$ , and  $\hat{p}_2$ , we obtain

$$\begin{aligned}\lambda &= -2(n_0 \log(1 - \hat{p}_1)^m + d \log(\hat{p}_1) + mn_+ \log(1 - \hat{p}_1) - d \log(1 - \hat{p}_1) \\ &\quad - n_0(\log(\hat{w} + (1 - \hat{w})\hat{q}_2)) - n_+ \log(1 - \hat{w}) - d \log(1 - \hat{q}_2) - mn_+ \log(\hat{q}_2) + d \log \hat{q}_2).\end{aligned}\tag{3.5}$$

We can find a critical value  $\chi_\alpha^2(1)$  at a significance level  $\alpha$  based on the chi-square distribution with 1 degree of freedom to compare with  $\lambda$ . If  $\lambda \geq c$ , reject  $H_0$ ; if  $\lambda < c$ , do not reject  $H_0$ . We will use Equation (3.5) to discuss type I errors and type II errors with changing values of parameters in chapter 5.

### 3.2. Score test

The two score tests of the null hypothesis at  $w = 0$  that we discuss have the same test statistic form as Equation (2.5).  $\mathbf{U}$  is a score vector containing the partial derivatives of  $L$  with respect to  $w$  and  $p$ , calculated when  $w = 0$  and  $p = \hat{p}_1$ , where  $\hat{p}_1$  is the maximum likelihood estimator of  $p$  when  $w = 0$ . The observed and expected information matrix, denoted by  $\mathbf{O}$  and  $\mathbf{I}$ , are given below

$$\mathbf{O}(w, p) = - \begin{pmatrix} \frac{\partial^2 L}{\partial w^2} & \frac{\partial^2 L}{\partial w \partial p} \\ \frac{\partial^2 L}{\partial w \partial p} & \frac{\partial^2 L}{\partial p^2} \end{pmatrix} \quad (3.6)$$

$$\mathbf{I}(w, p) = \mathbb{E}[\mathbf{O}],$$

both are evaluated at  $w = 0$  and  $p = \hat{p}_1$ .

We have the first-order derivatives in Equation (3.3), and the second-order derivatives are

$$\begin{aligned} \frac{\partial^2 L}{\partial w^2} &= -\frac{n_0(1 - (1 - p)^m)^2}{(w + (1 - w)(1 - p)^m)^2} - \frac{n_+}{(1 - w)^2} \\ \frac{\partial^2 L}{\partial w \partial p} &= \frac{n_0 m (1 - p)^{m-1} (w + (1 - w)(1 - p)^m) + n_0 m (1 - (1 - p)^m) (1 - w) (1 - p)^{m-1}}{(w + (1 - w)(1 - p)^m)^2} \\ \frac{\partial^2 L}{\partial p^2} &= \frac{n_0 w m (1 - w) (m - 1) (1 - p)^{m-2} - n_0 m (1 - w)^2 (1 - p)^{2m-2}}{(w + (1 - w)(1 - p)^m)^2}. \end{aligned}$$

Letting  $w = 0$  results in a simplification in the first-order derivatives:

$$\begin{aligned}\frac{\partial L}{\partial w} &= \frac{n_0}{(1-p)^m} - n \\ \frac{\partial L}{\partial p} &= \frac{d - mn}{1-p} + \frac{d}{p}.\end{aligned}\tag{3.7}$$

The corresponding second-order derivatives are as follows,

$$\begin{aligned}\frac{\partial^2 L}{\partial w^2} &= -\frac{n_0(1 - (1-p)^m)^2}{(1-p)^{2m}} - n_+ \\ \frac{\partial^2 L}{\partial w \partial p} &= \frac{n_0 m (1-p)^{m-1}}{(1-p)^{2m}} \\ \frac{\partial^2 L}{\partial p^2} &= \frac{d - mn}{(1-p)^2} - \frac{d}{p^2}.\end{aligned}$$

Hence, when  $w = 0$  and  $p = \hat{p}_1$ , the observed information matrix is

$$\mathbf{O}(0, \hat{p}_1) = \begin{pmatrix} \frac{n_0(1-(\hat{p}_1)^m)^2}{(\hat{p}_1)^{2m}} + n_+ & -\frac{n_0 m (1-\hat{p}_1)^{m-1}}{(\hat{p}_1)^{2m}} \\ -\frac{n_0 m (1-\hat{p}_1)^{m-1}}{(\hat{p}_1)^{2m}} & \frac{mn-d}{(\hat{p}_1)^2} + \frac{d}{\hat{p}_1^2} \end{pmatrix}.$$

Since  $\mathbb{E}[n_0] = n(1-p)^m$ ,  $\mathbb{E}[d] = mnp$ , and  $\mathbb{E}[n_+] = n - n(1-p)^m$ , we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2 L}{\partial w^2}\right] &= -\frac{n}{(1-p)^m} + n \\ \mathbb{E}\left[\frac{\partial^2 L}{\partial w \partial p}\right] &= \frac{nm}{1-p} \\ \mathbb{E}\left[\frac{\partial^2 L}{\partial p^2}\right] &= \frac{nm}{p(p-1)}.\end{aligned}$$

Hence when  $w = 0$  and  $p = \hat{p}_1$ , the expected information matrix is given by

$$\mathbf{I}(0, \hat{p}_1) = \begin{pmatrix} \frac{n}{(1-\hat{p}_1)^m} - n & -\frac{nm}{1-\hat{p}_1} \\ -\frac{nm}{1-\hat{p}_1} & \frac{mn}{\hat{p}_1(1-\hat{p}_1)} \end{pmatrix}.$$

When  $w = 0$  and  $p = \hat{p}_1$ ,  $\frac{\partial L}{\partial p} = 0$ , from Equation (3.7) and Equation (2.5), the two score test statistics have the simple form

$$z = \kappa(\hat{p}_1) \left( \frac{\partial L}{\partial p} \Big|_{w=0, p=\hat{p}_1} \right)^2 = \kappa(\hat{p}_1) \left( \frac{n_0}{(1-\hat{p}_1)^m} - n \right)^2$$

where  $\kappa(\hat{p}_1)$  is either the (1,1) the element of the inverted observed information matrix  $\mathbf{O}(0, \hat{p}_1)^{-1}$  or inverted expected information matrix  $\mathbf{I}(0, \hat{p}_1)^{-1}$ . After some steps of calculation and simplification, we find the score test statistic using the estimated observed information matrix:

$$z_O = \frac{(n_0 - n\hat{q}_1^m)^2 (nm\hat{p}_1^2 - d\hat{p}_1^2 + d\hat{q}_1^2)}{(n_0(1 - \hat{q}_1^m)^2 + n\hat{q}_1^{2m} - n_0\hat{q}_1^{2m})(nm\hat{p}_1^2 - d\hat{p}_1^2 + d\hat{q}_1^2) - n_0^2 m^2 \hat{p}_1^2} \quad (3.8)$$

where  $\hat{q}_1 = 1 - \hat{p}_1$ . The score test statistic using the estimated expected information matrix is

$$z_I = \frac{(n_0 - n\hat{q}_1^m)^2}{(n - n\hat{q}_1^m - mn\hat{p}_1\hat{q}_1^{m-1})\hat{q}_1^m}. \quad (3.9)$$

We apply these two score test statistics to find the ratio of  $z_I$  to  $z_O$  in Chapter 4, and to study type I errors and II errors of the score test in Chapter 5.

## CHAPTER 4

### SCORE TEST STATISTICS FOR THE ZERO-INFLATED BINOMIAL DISTRIBUTION

The data of Table 1 come from a simulation of a zero-inflated binomial distribution. We collect one sample from a zero-inflated binomial distribution with 10 trials, probability  $p = 0.5$  and the zero-inflated coefficient  $w = 0.2$ .

TABLE 1. Data from a zero-inflated binomial distribution with  $m = 10, p = 0.5, w = 0.2$ .

$i$	0	1	2	3	4	5	6	7	8	9	10
$n_i$	228	10	29	84	174	175	159	98	33	8	2

For the data in Table 1, we have the following values  $n_0 = 228, d = 3887$ , and  $\hat{p}_1 = 0.3887$ . We fix the value of  $n_i$  for  $i \in \{1, 2, 3, \dots, 10\}$  and regard  $n_0$  as a variable in Equation (3.8) and Equation (3.9). Then, we get the ratio  $z_I/z_O$  and the graph of the ratio of the two test statistics for the zero-inflated binomial distribution.

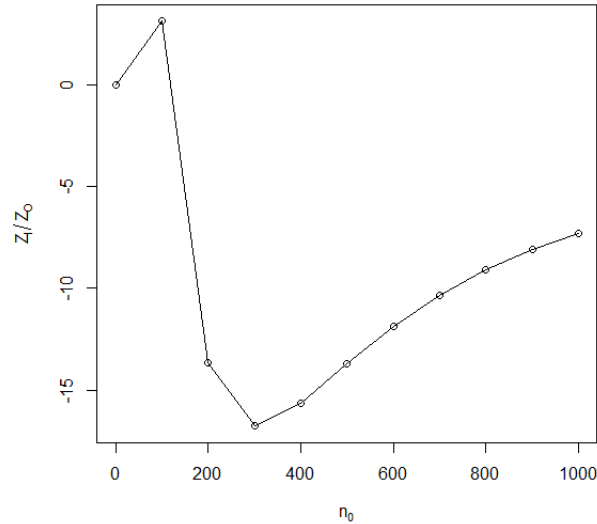


FIGURE 1. The graph of the ratio of the two score test statistics for the zero-inflated binomial distribution(1000, 10, 0.5, 0.2),  $z_I/z_O$ , plotted versus  $n_0$ .

We can see from Figure 1 that as  $n_0$  increases, the ratio increases as well, until the maximum value of the ratio is reached. Thereafter, the ratio decreases as  $n_0$  increases and produces a negative ratio until the minimum value of the ratio is achieved. Finally, the ratio increases as  $n_0$  increases but the ratio still keeps negative. The change in sign of the ratio occurs when the observed information matrix stops being positive-definite. Any value of  $n_0$  greater than 111 will produce a negative score test statistic by using the observed information matrix. The observed value of  $n_0 = 228$  is inside the area for  $n_0$  that results in a negative statistic. Thus the use of the observed information matrix in this case results in an invalid value of the test statistic, and thus the score test cannot be applied. As we can see, the negative values of score test statistic are obtained quite often.

The maximum likelihood estimator of the parameter  $p$  is  $\hat{p}_1 = 0.5030$  (when  $w = 0$ ). Substituting  $\hat{p}_1$  into  $z_I$  and  $z_O$ , we obtain  $z_I = 56657.78$  and  $z_O = -173.5659$ .

The following is another sample of a zero-inflated binomial distribution. The data set is collected from a zero-inflated binomial distribution with 10 trials, probability  $p = 0.1$  and zero-inflated coefficient  $w = 0.2$ . The data are as follows:

TABLE 2. Data from a zero-inflated binomial distribution with  $m = 10, p = 0.1, w = 0.2$ .

$i$	0	1	2	3	4	5	6	7	8	9	10
$n_i$	474	303	155	57	10	1	0	0	0	0	0

In this sample data, we have the following values  $n_0 = 474, d = 829$  and  $\hat{p}_1 = 0.0829$ . We do the same steps as in the former sample data. We change  $n_0$  from 0 to 1000 and fix the value of  $n_+$  to get the ratio  $z_I/z_O$  and plot the ratio of the two test statistics for the zero-inflated binomial distribution.

Figure 2 shows that as  $n_0$  increases, the ratio also increases, until the ratio reaches the maximum value. Thereafter, the ratio decreases as  $n_0$  increases and eventually goes

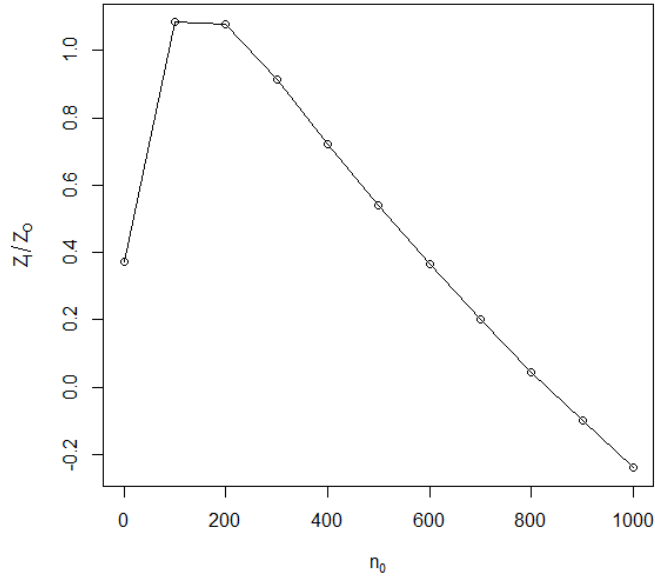


FIGURE 2. The graph of the ratio of the two score test statistics for the zero-inflated binomial distribution(1000, 10, 0.1, 0.2),  $z_I/z_O$ , plotted versus  $n_0$ .

to negative values. The change in sign of the ratio occurs at the value of  $n_0$  greater than 828, and the sign of the ratio is influenced when the observed information matrix stops being positive-definite. However, the observed value of  $n_0 = 474$  does not belong to the region for  $n_0$  that results in a negative statistic.

The maximum likelihood estimator of the parameter  $p$  is  $\hat{p}_1 = 0.1065$ . Plugging  $\hat{p}_1$  into  $z_I$  and  $z_O$ , we obtain  $z_I = 238.986$  and  $z_O = -367.305$ .



## CHAPTER 5

### TYPE I ERROR AND TYPE II ERROR

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis, while a type II error is incorrectly retaining a false null hypothesis. In this chapter, we focus on studying and comparing the percentage of type I errors and type II errors for the score test and the maximum likelihood ratio test when testing parameters for the geometric and the inflated binomial distributions. Table 3 is a summary of the two types error.

TABLE 3. Types of error.

	$H_0$ is true	$H_0$ is false
Reject $H_0$	Type I error, $P(\text{Type I}) = \alpha$	Correct decision
Do not reject $H_0$	Correct decision	Type II error, $P(\text{Type II}) = \beta$

We choose  $\alpha = 0.05$  as a threshold value to make the decision to reject or not reject the null hypothesis. Using the chi-square distribution with 1 degree of freedom, we find the critical value  $\lambda = 3.84146$  with the significance level of 0.05. Thereafter, we compare the critical value  $\lambda$  with the values of the likelihood ratio test and the score test to reach a decision about the null hypothesis  $H_0$ . We perform a simulation study by taking lots of sample data to estimate the percentage of type I errors and type II errors.

#### 5.1. Type I error for the geometric distribution

For the geometric distribution, we perform hypothesis testing regarding parameter  $p$ . We simulate 1000 sample data sets of the geometric distribution with  $p = 0.1$ . We have the following hypothesis,

$$H_0 : p = 0.1$$

$$H_1 : p \neq 0.1$$

We calculate type I errors under the likelihood ratio test and the score test using two score test statistics,  $z_I$  and  $z_O$ . Figure 3 shows type I errors for the geometric distribution with various values of the sample size  $n$ .

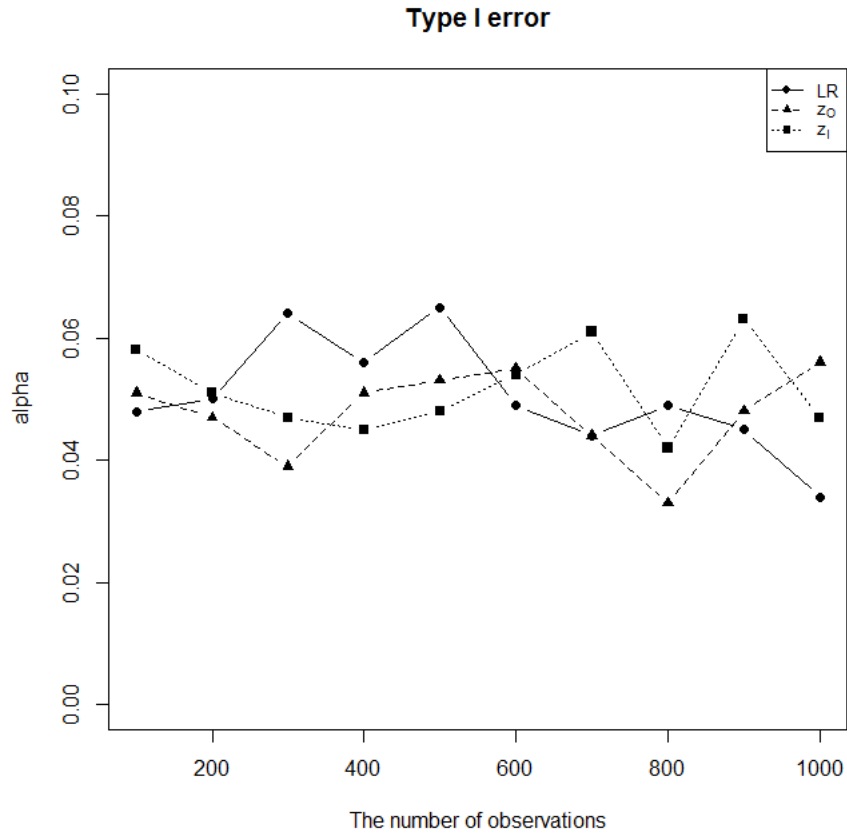


FIGURE 3. Type I error for the geometric distribution.

In this case, we start the number of observations at 5 since we want to avoid the situation when  $\hat{p}$  equals 0 which results in an undefined  $\log(\hat{p})$ . Figure 3 shows that the type I errors of the geometric distribution when we choose different  $n$  are between 0.02 and 0.08. Comparing the performance of the likelihood ratio test and the score test, figure 3 shows that the two tests yield similar performances. Note that the percentage of type I error fluctuates around the chosen significance level of 0.05.

## 5.2. Type II error for the geometric distribution

For type II errors, we consider two situations, one by fixing  $n$ , the number of the observations, and another by fixing  $p_0$ , the hypothesized values. First, we fix  $n$  and estimate the likelihood of type II errors by the likelihood ratio test. We collect 1000 samples of 100 observations from the geometric distribution with probability  $p$  that takes different values. Then we let  $p$  vary around  $p_0$ , such as  $p_0 = 0.1$  and  $p_0 = 0.3$ , to observe the percentage of type II errors. Figure 4 shows the relationship between  $p_0$  and  $\beta$ . The hypothesis is

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

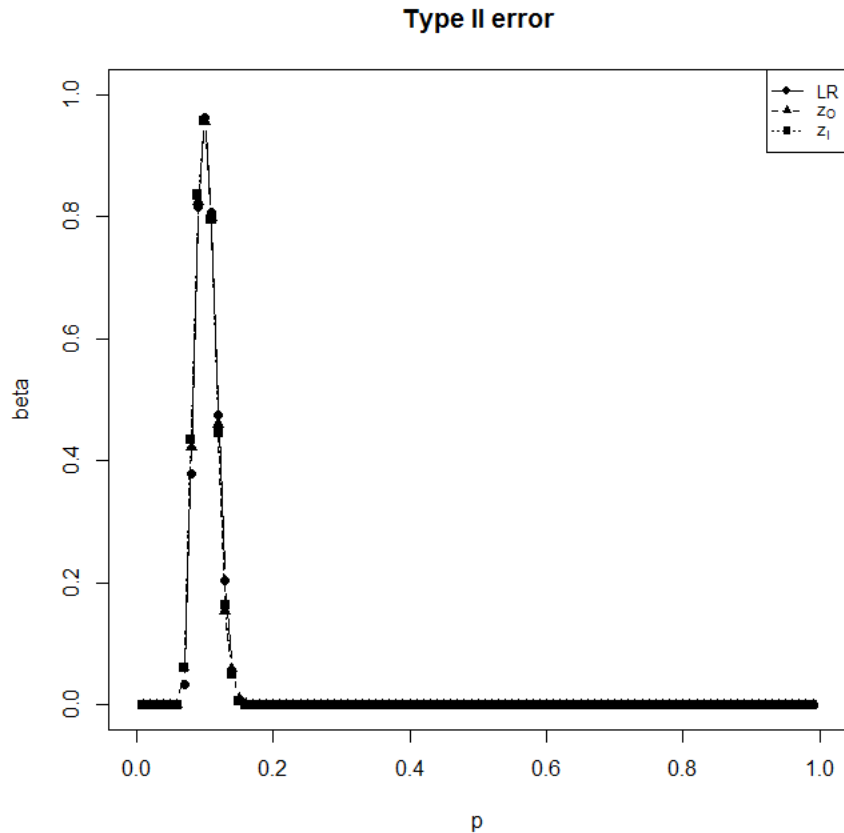


FIGURE 4. Type II error with  $n = 100, p_0 = 0.1$  for the geometric distribution.

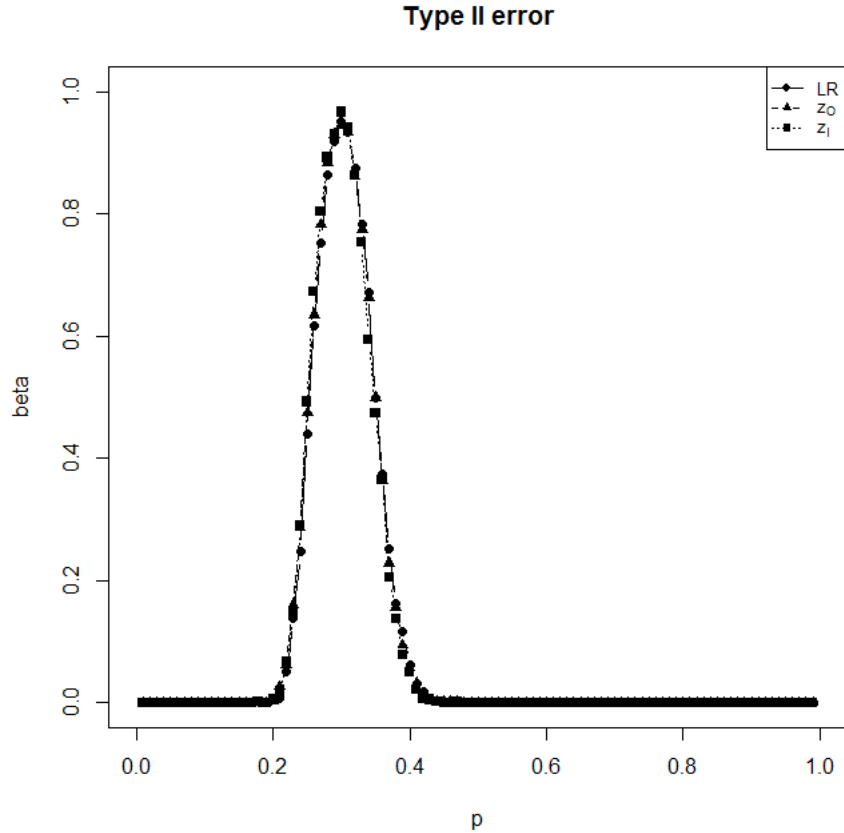


FIGURE 5. Type II error with  $n = 100, p_0 = 0.3$  for the geometric distribution.

Figure 4 and 5 show that the closer  $p$  is to  $p_0$ , the bigger the probability of a type II error. There is no big difference between the likelihood ratio test and the score test in terms of percentage of type II error.

Second, we want to study likelihood of type II error when  $n$  changes for some fixed values  $p_0$ . We choose  $p = 0.1$  to create the sample data and begin with  $n = 5$ . Figure 6, 7 and 8 show the percentage of type II error decreases as the number of observation increases for  $p_0 = 0.05, p_0 = 0.15$  and  $p_0 = 0.2$ . When  $p_0 = 0.05$  and  $n < 35$ , the likelihood ratio test has the smaller percentage of type II error than the score test. When  $p_0 = 0.15$  or  $0.2$  and  $n$  are small values, the score tests perform better than the likelihood ratio test; also, the closer the value  $p_0$  to the true value  $p$ , the higher the percentage of type II error in both tests.

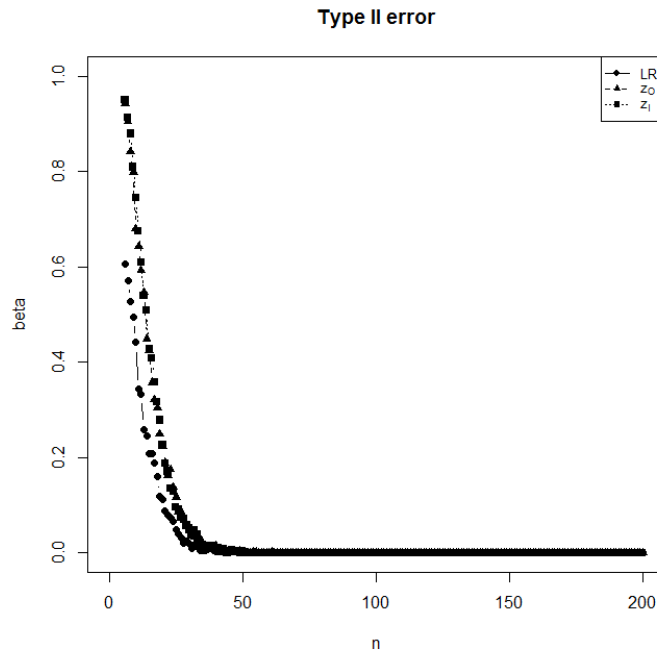


FIGURE 6. Type II error with  $p_0 = 0.05$  for the geometric distribution.

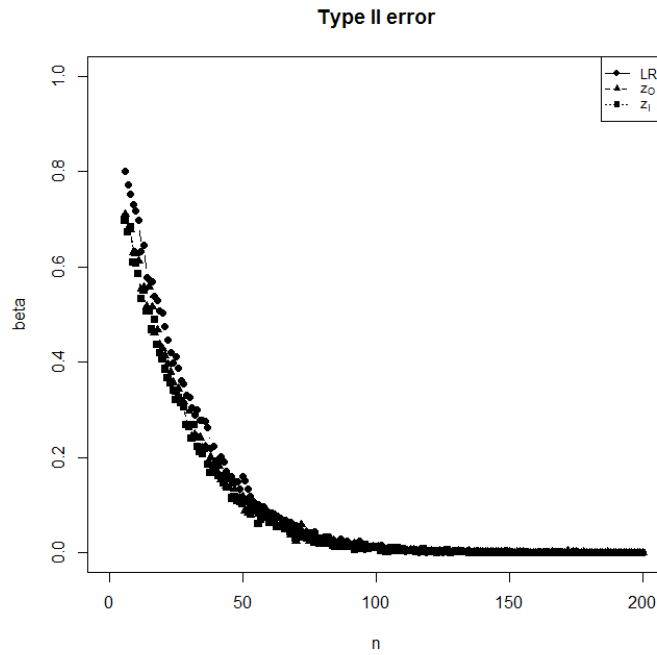


FIGURE 7. Type II error with  $p_0 = 0.15$  for the geometric distribution.

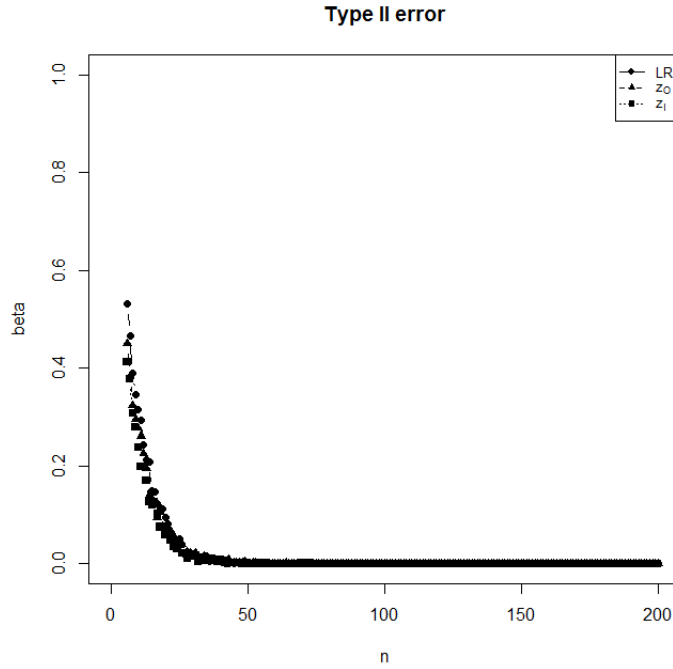


FIGURE 8. Type II error with  $p_0 = 0.2$  for the geometric distribution.

### 5.3. Type I error for the zero-inflated binomial distribution

For the zero-inflated binomial distribution, we compute the percentage of type I errors under the likelihood ratio test and the score test using  $z_O$  and  $z_I$ , for which we already have the formulas Equation (3.2), Equation (3.8) and Equation (3.9) in Chapter 3. We consider the hypothesis

$$H_0 : w = 0$$

$$H_1 : w \neq 0$$

We estimate the percentage of type I error with different values of sample size  $n$  under the likelihood ratio test and the score test. For each value of  $n$ , we discuss two different situations, one fixing the number of trials, and another fixing the probability of success in one trial. We let  $m$  equal 10 or 20 and  $p$  equal 0.1 or 0.3. Thus, we get

four different combinations. Under each combination, we want  $\alpha$  to hover around 0.05. We run 100 times to get the percentage of type I error in each case.

First, we let  $m = 10$  and  $p = 0.1$ . Figure 9 shows type I error at  $m = 10$  and  $p = 0.1$  with increasing sample sizes from 100 to 1000.

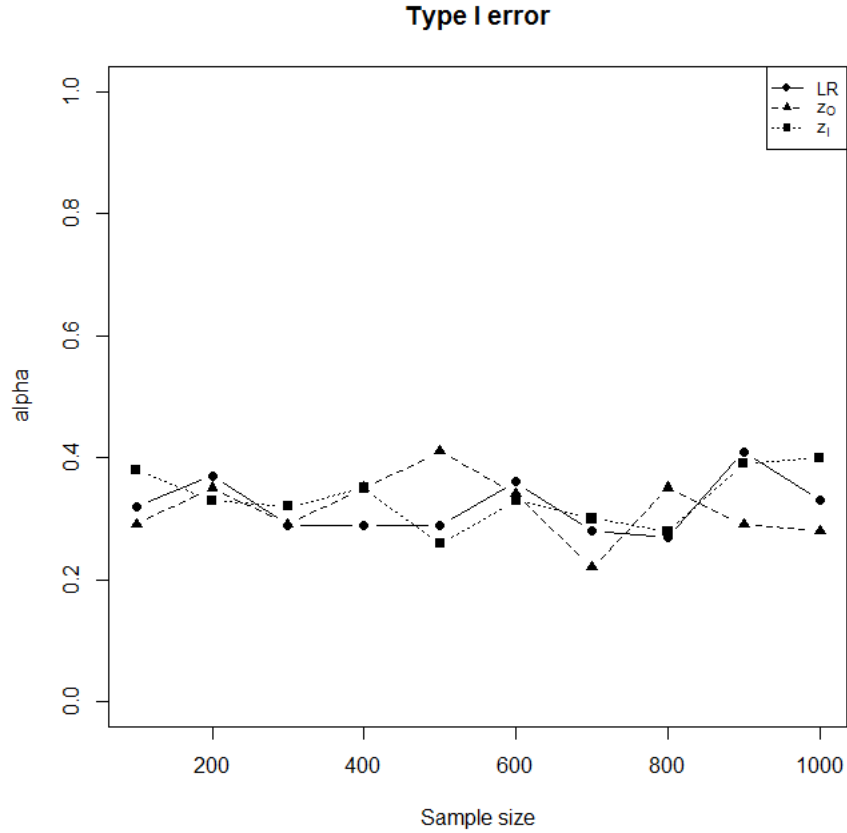


FIGURE 9. Type I error with  $m = 10$  and  $p = 0.1$  for zero-inflated binomial distribution.

Then we estimate type I error at  $m = 10$  and  $p = 0.3$  with increasing sample sizes from 100 to 1000. We show the result in Figure 10.

We also want to compute type I errors at  $m = 20$ ,  $p = 0.1$  and  $m = 20$ ,  $p = 0.3$  with increasing sample sizes. Figure 11 shows type I error with fixed  $m = 20$ ,  $p = 0.1$  and Figure 9 shows type I error with fixed  $m = 20$ ,  $p = 0.3$ .

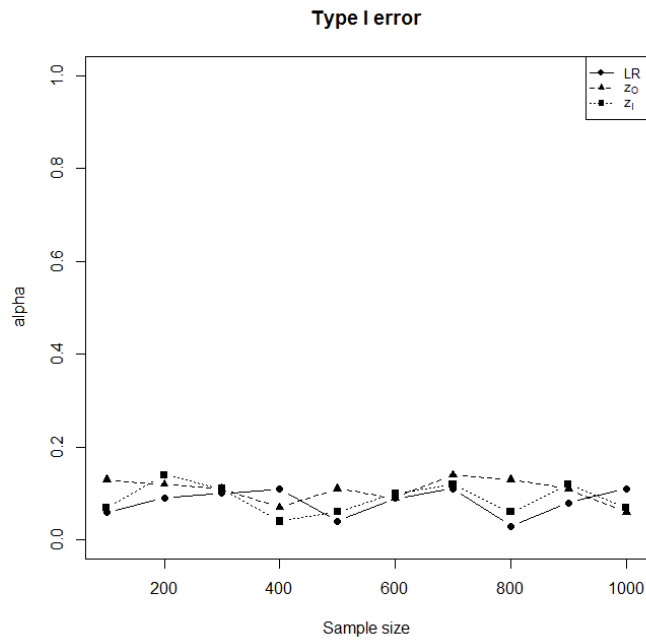


FIGURE 10. Type I error with  $m = 10$  and  $p = 0.3$  for zero-inflated binomial distribution.

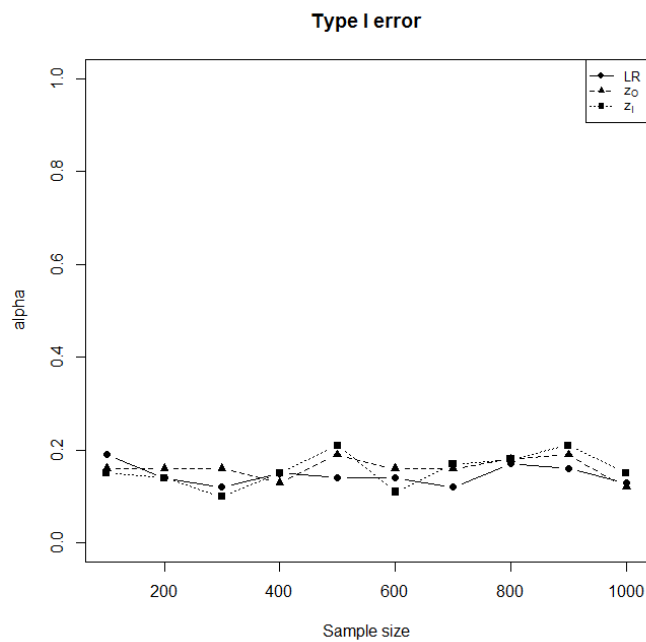


FIGURE 11. Type I error with  $m = 20$  and  $p = 0.1$  for zero-inflated binomial distribution.



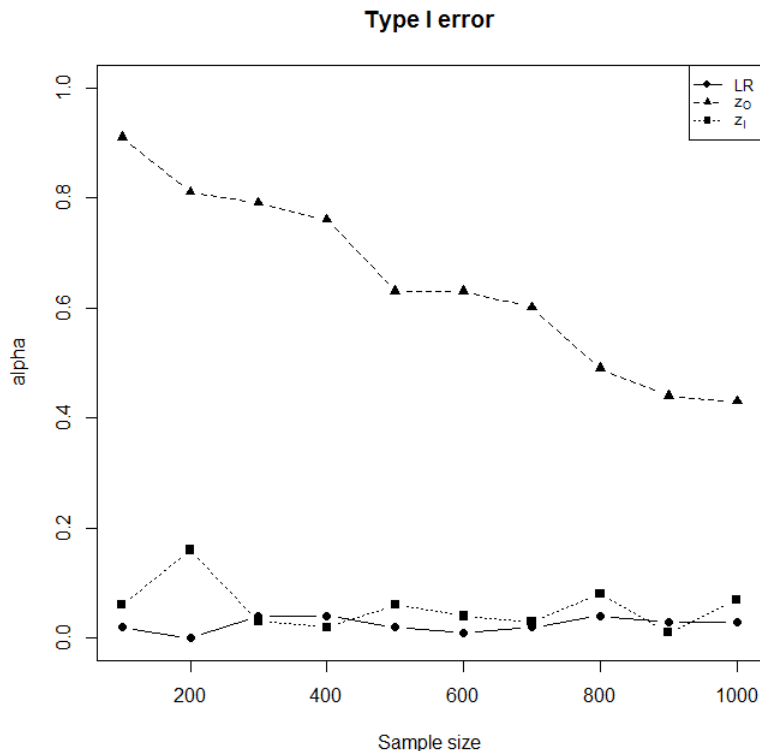


FIGURE 12. Type I error with  $m = 20$  and  $p = 0.3$  for zero-inflated binomial distribution.

Comparing Figure 9 with Figure 10, we find that when the number of trials is fixed at 10, the percentage of type I error under the likelihood ratio test and the score test is smaller with larger value of probability of success in one trial under the same sample size. Comparing Figure 9 with Figure 11, the graphs show that the smaller value of the number of trials, the larger the value of the type I error with fixed  $p = 0.1$ . However, comparing Figure 12 with Figure 10 or Figure 11, there is an abnormal type I error of  $z_0$ , which shows that type I error is significantly influenced by the sample size. We recall the formula to calculate  $z_0$ . If we choose  $m = 20, p = 0.3$ , there is no zero in a sample set with small sample size. The probability at  $X = 0$ ,  $P(X = 0) = 0.000798$ , is very small, which results that  $n_0 = 0$  in a small sample size. Then  $z_0 = n$ , where  $n$  is the sample size, is always greater than the critical value  $\chi_{0.05}^2 = 3.84146$  at 1 degree of freedom. In this case, we increase the sample size to 3,000 to see the change of type I

error. Figure 13 shows type I error at  $m = 20$ ,  $p = 0.3$  with maximum sample size 3,000. We can see that as the sample size increases, the percentage of type I error for score test

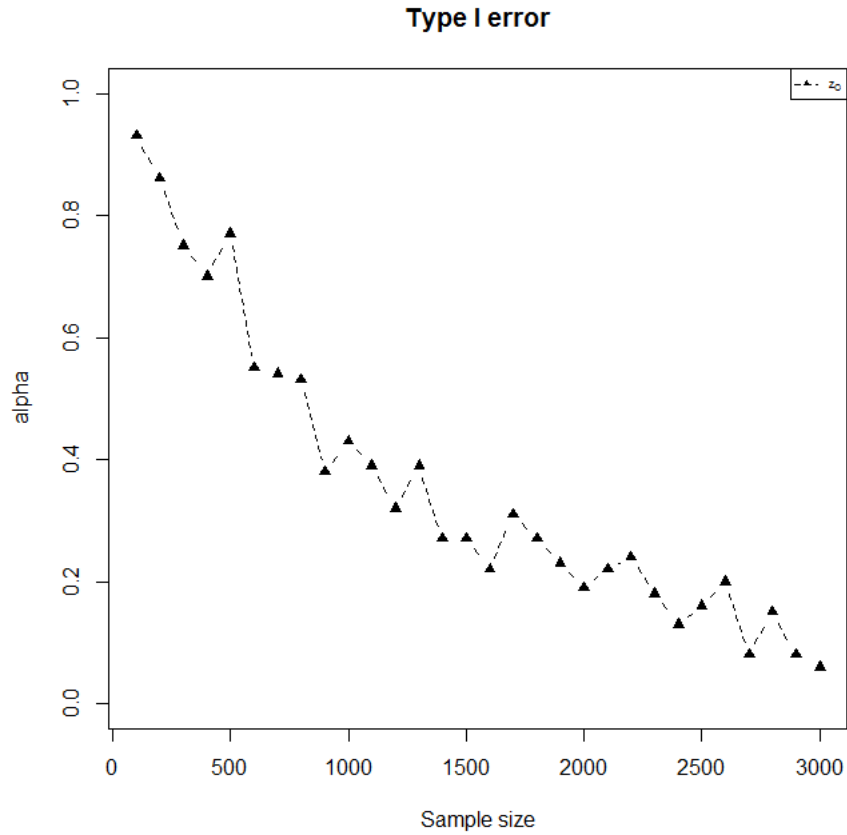


FIGURE 13. Type I error with  $m = 20$  and  $p = 0.3$  with larger sample size.

using test statistic  $Z_O$  significantly decreases. Figure 13 shows that if the sample size is large enough, type I error can reach the value below 0.05.

Comparing the values of percentages of type I error between the likelihood ratio test and the score test, figures 9 to 13 show that for the smaller number of trials  $m$  and probability  $p$  with the same sample size  $n$ , the difference between the two tests is insignificant. When the probability  $p$  is increasing, there is no zero in a sample which causes  $Z_O = n$  and  $Z_O$  is always greater than  $\chi^2_{0.05} = 3.84146$  at 1 degree of freedom. In this case, the likelihood ratio test performs better than the score test.

Next, we still discuss type I error under the likelihood ratio test and the score test when we fix the sample size and the number of trials with the probability of success in one trial being a variable. We run 100 times at  $n = 1000, m = 10$  to compute the percentage of type I error.

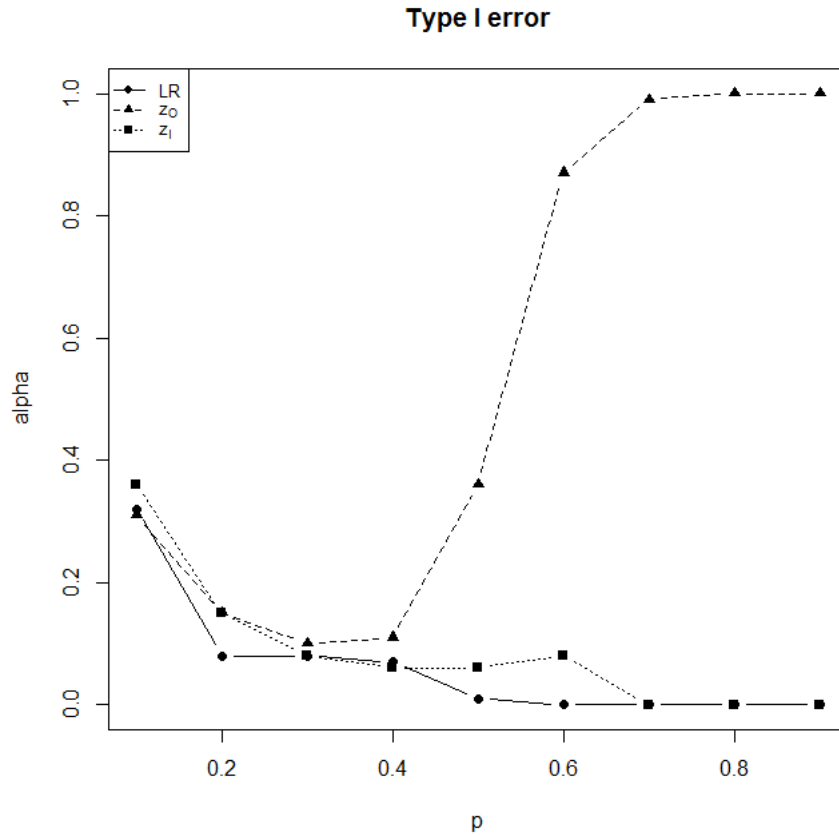


FIGURE 14. Type I error with  $n = 1000$  and  $m = 10$ .

Figure 14 shows type I error at different values of  $p$  with fixed  $n = 1000, m = 10$ . In this figure, the type I error under  $z_0$  is obviously increasing when  $p > 0.4$ , because when  $p$  is increasing, the probability of 0 in one sample data is decreasing. Once there is no 0 in the sample data,  $z_0$  will be  $n$ , which is always greater than the critical value 3.84146 and the score test leads to the wrong decision. Type I error under the likelihood ratio test and the score test using the expected information matrix decreases with increasing probability.

#### 5.4. Type II error for the zero-inflated binomial distribution

We discuss type II error in three situations by changing different parameters in the zero-inflated binomial distribution. In each case, we run 100 times to avoid some random mistake. Firstly, we fix  $m = 10, p = 0.1$  and  $w = 0.2$ , with the sample size changing from 100 to 2000.

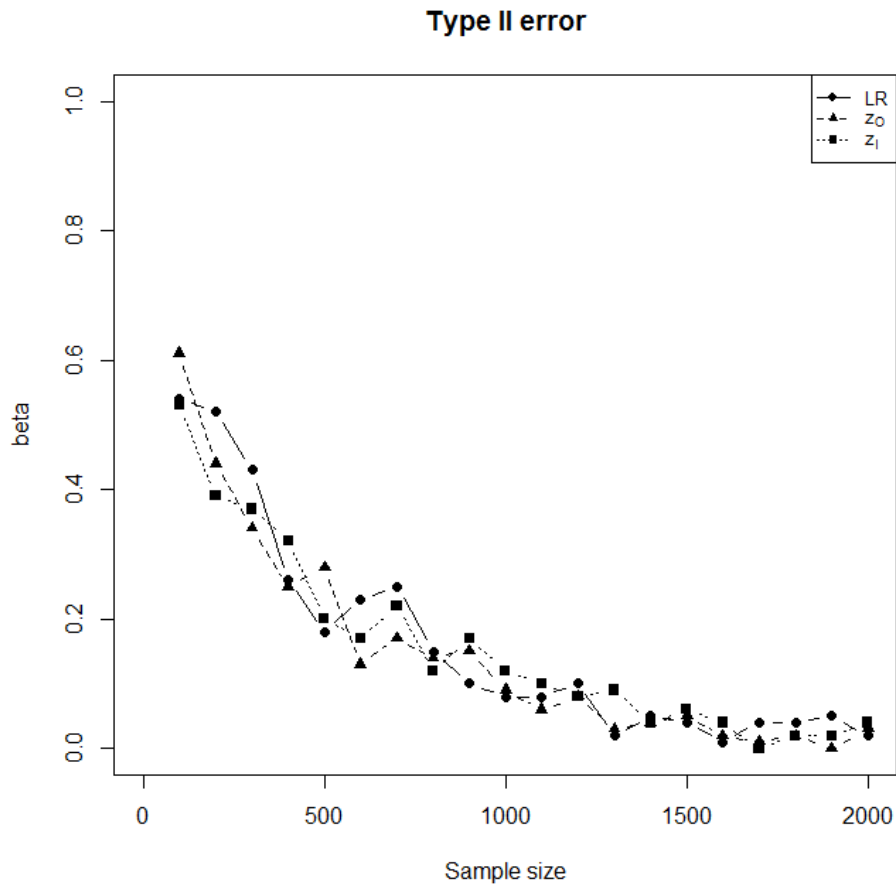


FIGURE 15. Type II error with  $m = 10, p = 0.1$  and  $w = 0.2$ .

Figure 15 shows the relationship between sample sizes with type II error under the likelihood ratio test and the score test fixing  $m, p$  and  $w$ . The type II error is decreasing when the sample size is increasing. If the sample size is big enough, the type II error

will reach 0. There is not a big difference in type II error between the two types of score tests.

Secondly, we fix  $n = 1000, m = 10$  and  $w = 0.2$ , choosing  $p$  as a variable.

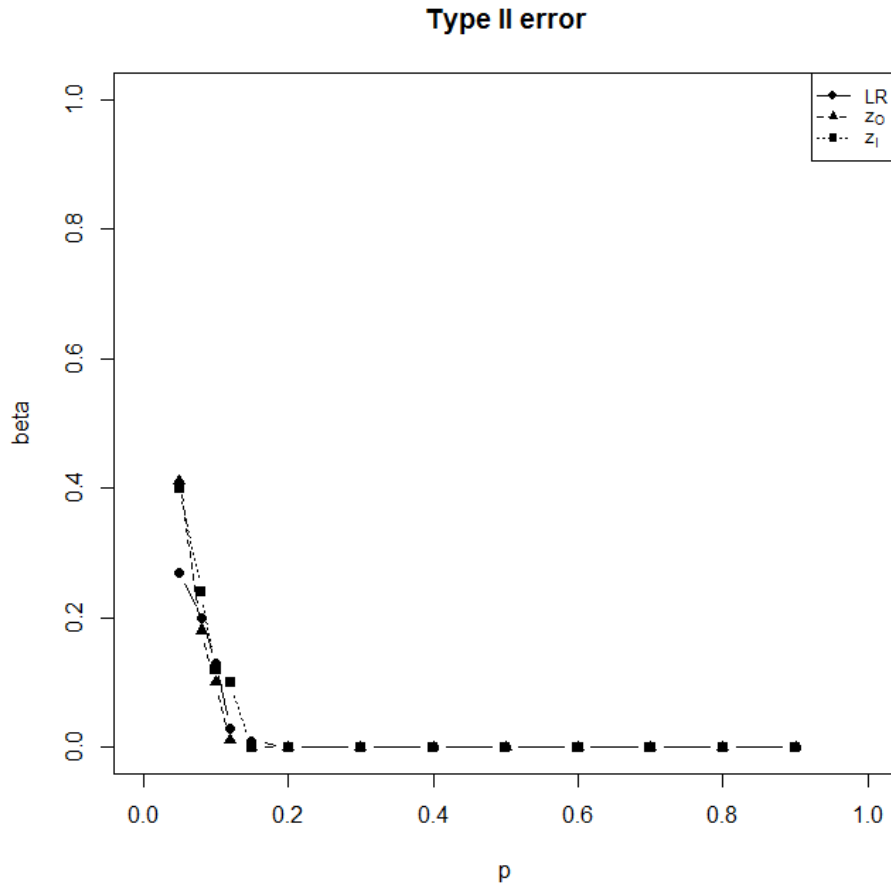


FIGURE 16. Type II error with  $n = 1000, m = 10$  and  $w = 0.2$ .

In Figure 16, we find that the percentage of type II error decreases with the probability increases and other parameters remain at  $n = 1000, m = 10$  and  $w = 0.2$ . At  $p = 0.05$ , the score test performs better than the likelihood ratio test. At other value of  $p$ , those type II errors under the likelihood ratio test and the score test are almost the same value.

Finally, we discuss the influence on type II errors by changing the zero-inflated parameter  $w$  and fixing  $n = 1000, m = 10$ , and  $p = 0.1$ .

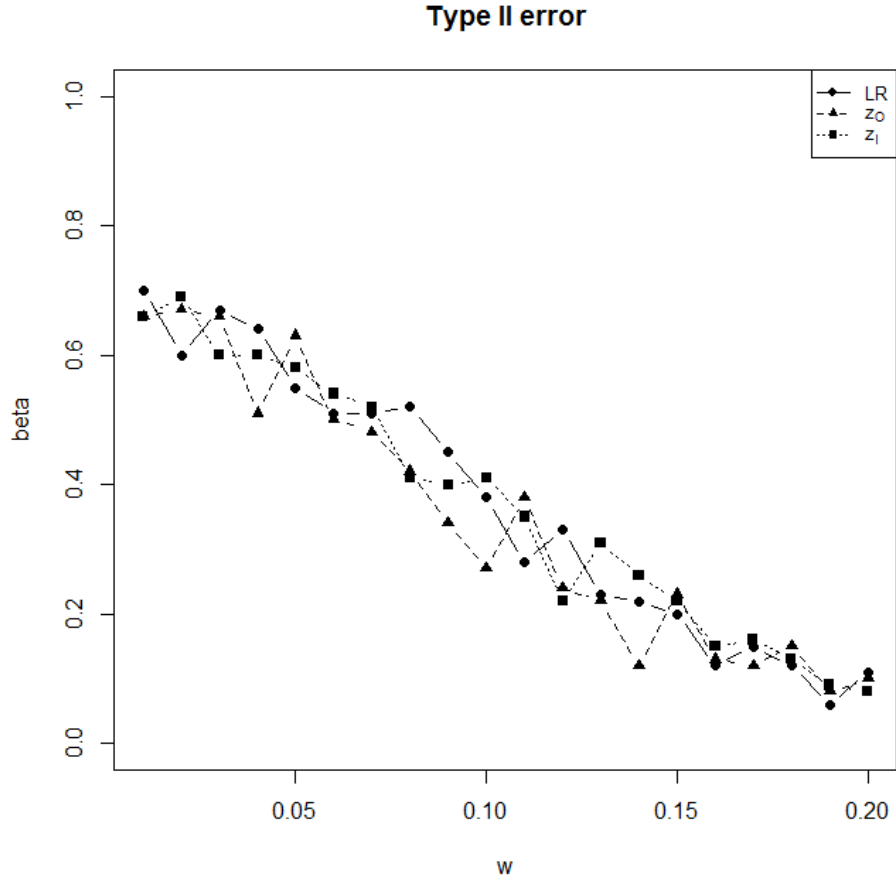


FIGURE 17. Type II error with  $n = 1000, m = 10$  and  $p = 0.1$ .

Figure 17 indicates that the percentage of type II errors under the likelihood ratio test and the score test decrease with the zero-inflated parameter being farther away from the hypothesised value zero. There is no substantial difference between the likelihood ratio test and the score test.

We only show percentages of type II errors when  $w$  is less than 0.2. If we increase the inflated parameter  $w$ , there will be many zeros in the sample data. In this case,  $Z_0$  might be less than 0, which is impossible under the chi-square distribution, and thus we cannot apply the score test using the observed information matrix. We want to show

the percentage of times that the score test cannot be applied. We fix the sample size  $n = 1000$  and choose  $m = 10, p = 0.1$ ;  $m = 10, p = 0.3$  and  $m = 20, p = 0.1$ . We run each case 1000 times.

TABLE 4. The percentage of time that score test cannot be applied.

w	m=10, p=0.1	m=10, p=0.3	m=20, p =0.1
0.1	0	0	0
0.2	0	0	0
0.3	0.029	0.57	0.004
0.4	0.332	1	0.559
0.5	0.77	1	0.999
0.6	0.947	1	1
0.7	0.987	1	1
0.8	0.986	1	1
0.9	0.977	1	1

Table 4 shows that the percentage of times that the score test cannot be applied is increasing with the inflated parameter increasing. When  $w$  is less than or equal 0.2, the percentage in each case is 0.

## CHAPTER 6

### CONCLUSION

We derived formulas to compute the likelihood ratio test statistic and the score test statistic for the geometric distribution and the zero-inflated binomial distribution. The two score test statistics are calculated by the estimated observed information matrix and the expected information matrix. We focus on computing score test statistics for the zero-inflated binomial distribution to calculate the ratio  $z_I/z_O$  with fixed  $n_i, i = 1, \dots, m$ . There are many instances in which the score test statistics using the observed information matrix are negative, which leads to a negative ratio  $z_I/z_O$ .

We study type I errors and type II errors for the geometric distribution and the zero-inflated binomial distribution. For the geometric distribution, we consider the null hypothesis and the alternative hypothesis,  $H_0 : p = 0.1$  and  $H_0 : p \neq 0.1$ . We find that the percentages of type I errors with various values of sample size for the likelihood ratio test and the score test are between 0.02 and 0.08. For type II errors, we have  $H_0 : p = p_0$  and  $H_0 : p \neq p_0$ . The percentages of type II error for the two tests are almost the same with fixed sample size  $n$ . The closer  $p$  is to  $p_0$ , the bigger the value of type II error is. The percentages of type II errors in both tests decrease with increasing sample size.

For the zero-inflated binomial distribution,  $p$  is a nuisance parameter. We have the null hypothesis and the alternative hypothesis,  $H_0 : w = 0$  and  $H_0 : w \neq 0$ . The percentages of type I errors under the two tests are almost the same when  $m$  and  $p$  are small. When  $m$  and  $p$  are increasing, the percentage of type I errors of the score test statistics using the estimated observed information matrix is extremely high because of no zero is in samples with small size. For different values of  $p$ , the percentage of type I errors of the score test using the estimated expected information matrix and the likelihood ratio test decrease as the probability  $p$  increases, while that of the score test statistic using the estimated observed information matrix decreases at first and then



increases as  $p$  increases. For type II errors, there is no substantial difference between the likelihood ratio test and the score test. The percentage of type II errors is decreasing when one of the parameters,  $n, p, w$  is increasing while others are fixed.

## APPENDIX

```
#Code to create data for table 1 and 2
r.Zib<-function(n,m,p,w)
{
  U<-runif(n)
  X<-rep(0,n)
  for(i in 1:n)
  {
    if(U[i]<Zib.cdf(m,0,p,w))
    {
      X[i]<-0
    }
    else
    {
      B=FALSE
      I=0
      while(B==FALSE)
      {
        int<-c(Zib.cdf(m,I,p,w),Zib.cdf(m,I+1,p,w))
        if((U[i]>int[1])&(U[i]<int[2]))
        {
          X[i]<-I+1
          B=TRUE
        }
      }
    }
  }
}
```

```

I=I+1
}
}
}
}
return (X)
}

```

```

Zib.cdf<-function (m, j , p , w)
{
  return (w+(1-w)*(1-p)^m+(1-w)*sum(pbinom(j , size=m, prob=p))
-(1-w)*dbinom(0 , size=m, prob=p))
}

```

```

r.Zib(1000,10,0.5,0.2)

```

```

r.Zib(1000,10,0.1,0.2)

```

**#Figure 1 and 2: ratio**

```

O<-function (n0 , n , d , m , p)
{
  return (((n0-n*(1-p)^m)^2*(m*n*p^2-d*p^2+d*(1-p)^2)/((n0*(1
-(1-p)^m)^2+n*(1-p)^(2*m)-n0*(1-p)^(2*m))*(m*n*p^2-d*p^2
+d*(1-p)^2)-n0^2*m^2*p^2))
}
I<-function (n0 , n , m , p)
{
  return ((n0/(1-p)^m-n)^2*(1-p)^m/(n-n*(1-p)^m-n*m*p*(1-p)^(m-1)))
}

```

```

}
ratio<-function(n0,n,d,m,p)
{
  return(return(I(n0,n,m,p)/O(n0,n,d,m,p)))
}
v<-c(0,100,200,300,400,500,600,700,800,900,1000)
r.Zib(1000,10,0.5,0.2)
v1<-c(ratio(0,772,3887,10,0.5035),ratio(100,872,3887,10,0.4458),
ratio(200,972,3887,10,0.3999),ratio(300,1072,3887,10,0.3626),
ratio(400,1172,3887,10,0.3317),ratio(500,1272,3887,10,0.3056),
ratio(600,1372,3887,10,0.2833),ratio(700,1472,3887,10,0.2641),
ratio(800,1572,3887,10,0.2473),ratio(900,1672,3887,10,0.2325),
ratio(1000,1777,3887,10,0.2194))
r.Zib(1000,10,0.1,0.2)
v3<c(ratio(0,526,829,10,0.1576),ratio(100,626,829,10,0.1324),
ratio(200,726,829,10,0.1142),ratio(300,826,829,10,0.1004),
ratio(400,926,829,10,0.0896),ratio(500,1026,829,10,0.0808),
ratio(600,1126,829,10,0.0736),ratio(700,1226,829,10,0.0676),
ratio(800,1326,829,10,0.0625),ratio(900,1426,829,10,0.0581),
ratio(1000,1526,829,10,0.0543))
plot(v,v1,type="o",xlab=expression("n" [0]),ylab=expression("Z" [I]/
"Z" [O]))
plot(v,v3,type="o",xlab=expression("n" [0]),ylab=expression("Z" [I]/
"Z" [O]))
#Figure 3:type I error for geometric distribution
#likelihood ratio test

```

```

compare<-function (n , p)
{
x<-rgeom (n , p)
d<-sum (x)
phat<-n / (n+d)
lamda<-(-2 * (n * log (p) + d * log (1 - p) - (n * log (phat) + d * log (1 - phat))))
countML<-0
if (lamda > 3.84146)
countML=countML+1
return (countML)
}
diffn<- function (m, n , p)
{
Z<-rep (0 , n)
for (i in 1:n)
{
Y<-rep (0 , m)
for (j in 1:m)
{
Y[j] <- compare (100 * n , p)
}
Z[i] <- sum (Y) / m
}
return (Z)
}
#z_0

```

```

compareO<-function(n,p)
{
x<-rgeom(n,p)
d<-sum(x)
phat<-p
ZO<-((n*(1-phat)-d*phat)^2/((n*(1-phat)^2+d*phat^2)))
countO<-0
if(ZO>3.84146)
countO=countO+1
return(countO)
}
diffnO<-function(m,n,p)
{
Z<-rep(0,n)
for(i in 1:n)
{
Y<-rep(0,m)
for(j in 1:m)
{
Y[j]<-compareO(100*n,p)
}
Z[i]<-sum(Y)/m
}
return(Z)
}#z-I
compareI<-function(n,p)

```

```

{
x<-rgeom(n,p)
d<-sum(x)
phat<-p
ZI<-((n*(1-phat)-d*phat)^2/(n*(1-phat)))
countI<-0
if(ZI>3.84146)
countI=countI+1
return(countI)
}
diffnI<-function(m,n,p)
{
Z<-rep(0,n)
for(i in 1:n)
{
Y<-rep(0,m)
for(j in 1:m)
{
Y[j]<-compareI(100*n,p)
}
Z[i]<-sum(Y)/m
}
return(Z)
}
#the graph of type I error
Final<-function(m,n,p)

```

```

{
v<-seq(100,1000,by=100)
v1<-diffn(m,n,p)
v2<-diffnO(m,n,p)
v3<-diffnI(m,n,p)
plot(v,v1,type="b",pch=19,xlab="The number of observations",
ylab="alpha",main="Type I error",ylim=c(0,0.1),xlim=c(100,1000))
lines(v,v2,pch=17,type="b",lty=2)
lines(v,v3,pch=15,type="b",lty=3)
legend("topright",legend=expression("LR", "z" [O], "z" [I]),pch=
c(19,17,15),lty=1:3,cex=0.8)
}

```

**#Figure 4 and 5**

```

compare2<-function(n,p,p0)
{
x<-rgeom(n,p)
d<-sum(x)
phat<-n/(n+d)
beta<-(-2*(n*log(p0)+d*log(1-p0)-(n*log(phat)+d*log(1-phat))))
countML<-0
if(beta<3.84146)
countML=countML+1
return(countML)
}
per2<-function(m,n,p,p0)
{

```



```

Y<-rep(0,m)
for(j in 1:m)
{
Y[j]<-compare2(n,p,p0)
}
sum(Y)
return(sum(Y)/m)
}
diffnfixLRT<-function(m,n,p,p0)
{
Z<-rep(0,p0)
for(i in 1:p0)
{
Z[i]<-per2(m,n,p,i/100)
}
return(Z)
}
#z_O
compareO2<-function(n,p,p0)
{
x<-rgeom(n,p)
d<-sum(x)
ZO<-(((n*(1-p0)-d*p0)^2/((n*(1-p0)^2+d*p0^2)))
countO<-0
if(ZO<3.84146)
countO=countO+1

```

```

return (countO)
}
perO2<-function (m, n , p , p0)
{
Y<-rep (0 ,m)
for (j in 1:m)
{
Y[j]<-compareO2 (n , p , p0)
}
sum(Y)
return (sum(Y) /m)
}
diffnfixO2<-function (m, n , p , p0)
{
Z<-rep (0 ,p0)
for (i in 1:p0)
{
Z[ i ]<-perO2 (m, n , p , i / 100)
}
return (Z)
}
#z - I
compareI2<-function (n , p , p0)
{
x<-rgeom (n , p)
d<-sum (x)

```

```

ZI<-((n*(1-p0)-d*p0)^2/(n*(1-p0)))
countI<-0
if (ZI < 3.84146)
countI=countI+1
return (countI)
}
perI2<-function (m, n, p, p0)
{
Y<-rep (0, m)
for (j in 1:m)
{
Y[j]<-compareI2 (n, p, p0)
}
sum(Y)
return (sum(Y)/m)
}
diffnfixI2<-function (m, n, p, p0)
{
Z<-rep (0, p0)
for (i in 1:p0)
{
Z[i]<-perI2 (m, n, p, i/100)
}
return (Z)
}
Figure4<-function (m, n, p, p0)

```

```

{
  v<-seq(0.01,0.99,by=0.01)
v1<-diffnfixLRT(m,n,p,p0)
v2<-diffnfixO2(m,n,p,p0)
v3<-diffnfixI2(m,n,p,p0)
plot(v,v1,type="b",pch=19,xlab="p",ylab="beta",
main="Type II error",ylim=c(0,1),xlim=c(0,1))
lines(v,v2,pch=17,type="b",lty=2)
lines(v,v3,pch=15,type="b",lty=3)
legend("topright",legend=expression("LR", "z" [O], "z" [I]),
pch=c(19,17,15),lty=1:3,cex=0.8)
}

#figure 6,7,8
#likelihood ratio test
diffnLRT<-function(m,n,p,p0)
{
Z<-rep(0,n)
for(i in 5:n)
{
Z[i]<-per2(m,i,p,p0)
}
return(Z)
}

#z_O
diffnOgeo<-function(m,n,p,p0)
{

```

```

Z<-rep(0,n)
for(i in 5:n)
{
Z[i]<-perO2(m,i,p,p0)
}
return(Z)
}

#z-I
diffnIgeo<-function(m,n,p,p0)
{
Z<-rep(0,n)
for(i in 5:n)
{
Z[i]<-perI2(m,i,p,p0)
}
return(Z)
}

Figure6<-function(m,n,p,p0)
{
v<-tail(seq(1,n,by=1),n-5)
v1<-tail(diffnLRT(m,n,p,p0),n-5)
v2<-tail(diffnOgeo(m,n,p,p0),n-5)
v3<-tail(diffnIgeo(m,n,p,p0),n-5)
plot(v,v1,type="b",pch=19,xlab="n",ylab="beta",
main="Type II error",ylim=c(0,1),xlim=c(0,n))
lines(v,v2,pch=17,type="b",lty=2)

```

```

lines(v, v3, pch=15, type="b", lty=3)
legend("topright", legend=expression("LR", "z" [O], "z" [I]),
pch=c(19, 17, 15), lty=1:3, cex=0.8)
}
#figure9 to12
#likelihood ratio test
IBlamda<-function(n, m, p, w)
{
d<-sum(r.Zib(n, m, p, w))
numberofzero<-c(r.Zib(n, m, p, w))
n0<-length(which(numberofzero==0))
phat1<-d/(m*n)
qhat2<-uniroot(function(x)-d*n0*x^m+(m*n*n0-m*n0*n0)*x
+d*n0+m*n0*n0-m*n*n0, lower=0, upper=0.9999)$root
what2<-(n0-n*qhat2^m)/(n-n*qhat2^m)
ifelse(qhat2==0, 0, -2*(n0*log((1-phat1)^m)+d*log(phat1)
+m*(n-n0)*log(1-phat1)-d*log(1-phat1)-n0*(log(what2+
(1-what2)*qhat2^m))-(n-n0)*log(1-what2)-d*log(1-qhat2)-
m*(n-n0)*log(qhat2)+d*log(qhat2)))
}
IBlrt<-function(n, m, p, w)
{
lamda<-IBlamda(n, m, p, w)
count<-0
if(lamda>3.84146)
count=count+1

```

```

return (count)
}
perib<-function(l,n,m,p,w)
{
Y<-rep(0,l)
for(i in 1:l)
{
Y[i]<-IBlrt(n,m,p,w)
}
sum(Y)
return(sum(Y)/l)
}
diffnib<-function(l,n,m,p,w)
{
Z<-rep(0,n)
for(i in 1:n)
{
Z[i]<-perib(l,i*100,m,p,w)
}
return(Z)
}
IBlamdazo<-function(n,m,p,w)
{
d<-sum(r.Zib(n,m,p,w))
numberofzero<-c(r.Zib(n,m,p,w))
n0<-length(which(numberofzero==0))

```

```

phat1<-d/(m*n)
qhat1<-1-phat1
lamda<-(n0-n*qhat1^m)^2*((m*n-d)*phat1^2+d*qhat1^2)
/((n0*(1-qhat1^m)^2+(n-n0)*qhat1^(2*m))*((m*n-d)*phat1^2
+d*qhat1^2)-n0^2*m^2*phat1^2)
return(lamda)
}
IBzo<-function(n,m,p,w)
{
lamda<-IBlamdazo(n,m,p,w)
count<-0
if(lamda>3.84146)
count=count+1
return(count)
}
peribzo<-function(l,n,m,p,w)
{
Y<-rep(0,l)
for(i in 1:l)
{
Y[i]<-IBzo(n,m,p,w)
}
sum(Y)
return(sum(Y)/l)
}
diffnibzo<-function(l,n,m,p,w)

```



```

{
Z<-rep(0,n)
for(i in 1:n)
{
Z[i]<-peribzo(1,100*i,m,p,w)
}
return(Z)
}
#z-I
IBlamdazi<-function(n,m,p,w)
{
d<-sum(r.Zib(n,m,p,w))
numberofzero<-c(r.Zib(n,m,p,w))
n0<-length(which(numberofzero==0))
phat1<-d/(m*n)
qhat1<-1-phat1
lamda<-(n0-n*qhat1^m)^2/((n-n*qhat1^m-m*n*phat1*qhat1^(m-1))
*qhat1^m)
return(lamda)
}
IBzi<-function(n,m,p,w)
{
lamda<-IBlamdazi(n,m,p,w)
count<-0
if(lamda>3.84146)
count=count+1
}

```

```

    return(count)
  }
peribzi<-function(l,n,m,p,w)
{
  Y<-rep(0,l)
  for(i in 1:l)
  {
    Y[i]<-IBzi(n,m,p,w)
  }
  sum(Y)
  return(sum(Y)/l)
}
diffnibzi<-function(l,n,m,p,w)
{
  Z<-rep(0,n)
  for(i in 1:n)
  {
    Z[i]<-peribzi(l,100*i,m,p,w)
  }
  return(Z)
}
Figure9<-function(l,n,m,p,w)
{
  v<-seq(100,1000,by=100)
  v1<-diffnib(l,n,m,p,w)
  v2<-diffnibzo(l,n,m,p,w)

```

```

v3<-diffnibzi(l,n,m,p,w)
plot(v,v1,type="b",pch=19,xlab="Sample size",ylab="alpha",
main="Type I error",ylim=c(0,1),xlim=c(100,1000))
lines(v,v2,pch=17,type="b",lty=2)
lines(v,v3,pch=15,type="b",lty=3)
legend("topright",legend=expression("LR", "z" [O], "z" [I]),
pch=c(19,17,15),lty=1:3,cex=0.8)
}

```

**#figure13**

```

Figure13<-function(l,n,m,p,w)
{
v<-seq(100,3000,by=100)
v1<-diffnibzo(l,n,m,p,w)
plot(v,v1,pch=17,type="b",lty=2,xlab="Sample size",ylab="alpha",
main="Type I error",ylim=c(0,1),xlim=c(100,3000))
legend("topright",legend=expression("z" [O]),lty=2,pch=17,cex=0.6)
}

```

**#figure14**

**#LRT**

```

diffpib<-function(l,n,m,p,w)
{
Z<-rep(0,p)
for(i in 1:p)
{
Z[i]<-perib(l,n,m,0.1*i,w)
}
}

```

```

return (Z)
}
#z_O
diffpibzo<-function(l,n,m,p,w)
{
Z<-rep(0,p)
for(i in 1:p)
{
Z[i]<-peribzo(l,n,m,0.1*i,w)
}
return (Z)
}
#z_I
diffpibzi<-function(l,n,m,p,w)
{
Z<-rep(0,p)
for(i in 1:p)
{
Z[i]<-peribzi(l,n,m,0.1*i,w)
}
return (Z)
}
Figure14<-function(l,n,m,p,w)
{
v<-seq(0.1,0.9,by=0.1)
v1<-diffpib(l,n,m,p,w)

```

```

v2<-diffpibzo(1,n,m,p,w)
v3<-diffpibzi(1,n,m,p,w)
plot(v,v1,type="b",pch=19,xlab="p",ylab="alpha",main="Type I
  error",
ylim=c(0,1),xlim=c(0.1,0.9))
lines(v,v2,pch=17,type="b",lty=2)
lines(v,v3,pch=15,type="b",lty=3)
legend("topleft",legend=expression("LR","z"[O],"z"[I]),pch=c(19,
17,15),
lty=1:3,cex=0.8)
}
#figure15
#LRT
IBlrt2<-function(n,m,p,w)
{
beta<-IBlamda(n,m,p,w)
count<-0
if(beta<3.84146)
count=count+1
return(count)
}
perlrt2<-function(l,n,m,p,w)
{
Y<-rep(0,l)
for(i in 1:l)
{

```

```

Y[i] <- IBlrt2(n, m, p, w)
}
sum(Y)
return(sum(Y)/l)
}
diffnlrt <- function(l, n, m, p, w)
{
Z <- rep(0, n)
for(i in 1:n)
{
Z[i] <- perlrt2(l, 100*i, m, p, w)
}
return(Z)
}
#z_0
IBzo2 <- function(n, m, p, w)
{
beta <- IBlamdazo(n, m, p, w)
count <- 0
if(beta < 3.84146)
count = count + 1
return(count)
}
perzo2 <- function(l, n, m, p, w)
{
Y <- rep(0, l)

```

```

for(i in 1:l)
{
Y[i]<-IBzo2(n,m,p,w)
}
sum(Y)
return(sum(Y)/l)
}
diffnzo<-function(l,n,m,p,w)
{
Z<-rep(0,n)
for(i in 1:n)
{
Z[i]<-perzo2(l,100*i,m,p,w)
}
return(Z)
}
#z-I
IBzi2<-function(n,m,p,w)
{
beta<-IBlamdazi(n,m,p,w)
count<-0
if(beta<3.84146)
count=count+1
return(count)
}
perzi2<-function(l,n,m,p,w)

```

```

{
Y<-rep(0,l)
for(i in 1:l)
{
Y[i]<-IBzi2(n,m,p,w)
}
sum(Y)
return(sum(Y)/l)
}
diffnzi<-function(l,n,m,p,w)
{
Z<-rep(0,n)
for(i in 1:n)
{
Z[i]<-perzi2(l,100*i,m,p,w)
}
return(Z)
}
Figure15<-function(l,n,m,p,w)
{
v<-seq(100,2000,by=100)
v1<-diffnlrt(l,n,m,p,w)
v2<-diffnzo(l,n,m,p,w)
v3<-diffnzi(l,n,m,p,w)
plot(v,v1,type="b",pch=19,xlab="Sample size",ylab="beta",
main="Type II error",ylim=c(0,1),xlim=c(0,2000))

```



```

lines (v ,v2 ,pch=17,type="b" ,lty =2)
lines (v ,v3 ,pch=15,type="b" ,lty =3)
legend (" topright" ,legend=expression ("LR" ," z" [O] ," z" [I]) ,
pch=c (19 ,17 ,15) ,lty =1:3 ,cex=0.8)
}
#figure16
type2diffp<-function (l ,n,m,p,w)
{
v1<-perlrt2 (l ,n,m,p,w)
v2<-perzo2 (l ,n,m,p,w)
v3<-perzi2 (l ,n,m,p,w)
return ( list (v1 ,v2 ,v3))
}
v<-c (0.05 ,0.08 ,0.1 ,0.12 ,0.15 ,0.2 ,0.3 ,0.4 ,0.5 ,0.6 ,0.7 ,0.8 ,0.9)
v1<-c (0.27 ,0.2 ,0.13 ,0.03 ,0.01 ,0 ,0 ,0 ,0 ,0 ,0 ,0 ,0)
v2<-c (0.41 ,0.18 ,0.1 ,0.01 ,0 ,0 ,0 ,0 ,0 ,0 ,0 ,0 ,0)
v3<-c (0.4 ,0.24 ,0.12 ,0.1 ,0 ,0 ,0 ,0 ,0 ,0 ,0 ,0 ,0)
plot (v ,v1 ,type="b" ,pch=19,xlab="p" ,ylab="beta" ,
main="Type II error" ,ylim=c (0 ,1) ,xlim=c (0 ,1))
lines (v ,v2 ,pch=17,type="b" ,lty =2)
lines (v ,v3 ,pch=15,type="b" ,lty =3)
legend (" topright" ,legend=expression ("LR" ," z" [O] ," z" [I]) ,
pch=c (19 ,17 ,15) ,lty =1:3 ,cex=0.8)
#figure17
#LRT
diffwlrt<-function (l ,n,m,p,w)

```

```

{
Z<-rep(0,w)
for(i in 1:w)
{
Z[i]<-perlrt2(l,n,m,p,0.01*i)
}
return(Z)
}
#z-O
diffwzo<-function(l,n,m,p,w)
{
Z<-rep(0,w)
for(i in 1:w)
{
Z[i]<-perzo2(l,n,m,p,0.01*i)
}
return(Z)
}
#z-I
diffwzi<-function(l,n,m,p,w)
{
Z<-rep(0,w)
for(i in 1:w)
{
Z[i]<-perzi2(l,n,m,p,0.01*i)
}
}

```

```

return (Z)
}
Figure17<-function (l ,n,m,p,w)
{
v<-seq (0.01 ,0.2 ,by=0.01)
v1<-diffwlr t (l ,n,m,p,w)
v2<-diffwzo (l ,n,m,p,w)
v3<-diffwzi (l ,n,m,p,w)
plot (v ,v1 ,type="b" ,pch=19,xlab="w" ,ylab=" beta" ,
main=" Type II error" ,ylim=c (0 ,1) ,xlim=c (0.01 ,0.2))
lines (v ,v2 ,pch=17,type="b" ,lty=2)
lines (v ,v3 ,pch=15,type="b" ,lty=3)
legend (" topright" ,legend=expression ("LR" ," z" [O] ," z" [I] ) ,
pch=c (19 ,17 ,15) ,lty =1:3 ,cex=0.8)
}
###Table4
invalidtest<-function (n,m,p,w)
{
beta<-IBlamdazo (n,m,p,w)
count<-0
if (beta <0)
count=count+1
return (count)
}
perinvalid<-function (k,n,m,p,w)
{

```

```

Y<-rep(0,k)
for(i in 1:k)
{
Y[i]<-invalidtest(n,m,p,w)
}
sum(Y)
return(sum(Y)/k)
}
perinvaliddiff<-function(k,n,m,p,w)
{
Y<-rep(0,w)
for(i in 1:w)
{
Y[i]<-perinvalid(k,n,m,p,0.1*i)
}
return(Y)
}
perinvaliddiff(1000,1000,10,0.1,9)
[1] 0.000 0.000 0.029 0.332 0.770 0.947 0.987 0.986 0.977
perinvaliddiff(1000,1000,10,0.3,9)
[1] 0.00 0.00 0.57 1.00 1.00 1.00 1.00 1.00 1.00
perinvaliddiff(1000,1000,20,0.1,9)
[1] 0.000 0.000 0.004 0.559 0.999 1.000 1.000 1.000 1.000
perinvaliddiff(1000,1000,20,0.3,9)
[1] 0 1 1 1 1 1 1 1 1

```

## BIBLIOGRAPHY

- [1] Morgan, B. J. T. Palmer, K. J. and Ridout, M. S. (2007). “Negative Score Test Statistic”, *The American Statistical Association*, vol. 61. pp. 285-288.
- [2] George, C. and Roger, L. B. (2002). *Statistical Inference*, 2nd ed., Thomson Learning.
- [3] *Sampling from discrete distributions*. <http://dept.stat.lsa.umich.edu/~jasoneg/Stat406/lab5.pdf>.
- [4] Rao, C. R. (1973). *Linear Statistical Inference*, 2nd ed., New York: Wiley.
- [5] Ferguson, T. (1982). “An Inconsistent Maximum Likelihood Estimate,” *Journal of the American Statistical Association*, 77, 831-34.
- [6] R software. <https://www.r-project.org/foundation/donors.html>