A graphical approach for goodness-of-fit of Poisson model

Davin P Padilla

University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Repository Citation
https://digitalscholarship.unlv.edu/rtds/1607

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
A GRAPHICAL APPROACH FOR GOODNESS-OF-FIT
OF POISSON MODEL

by

Davin P. Padilla
Bachelor of Arts in Mathematics
University of Hawaii at Hilo
2000

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science Degree in Mathematical Sciences
Department of Mathematical Sciences
College of Sciences

Graduate College
University of Nevada, Las Vegas
December 2003
Thesis Approval
The Graduate College
University of Nevada, Las Vegas

November 12, 2003

The Thesis prepared by

Davin P. Padilla

Entitled

A Graphical Approach for Goodness-of-Fit of Poisson Model

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Mathematical Sciences

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

PR/1017-53/1-00
ABSTRACT

A Graphical Approach for Goodness-of-Fit of Poisson Model

by

Davin P. Padilla

Dr. Ashok K. Singh, Examination Committee Chair
Professor of Statistics
University of Nevada, Las Vegas

Extensive work has been done on goodness-of-fit (GOF) tests for data assumed to have come from univariate continuous distributions; however, literature on GOF procedures for univariate discrete distributions is rather sparse in comparison. The Poisson distribution in particular has received much attention in the study of GOF tests due to its numerous applications as a model for observable phenomena. Hence, we survey existing GOF tests for Poissonity and present a useful guide to the most commonly used distribution-free GOF tests in practice. We then propose and investigate a graphical test of fit for the Poisson model that is based on a Poisson Q-Q plot, a squared correlation coefficient $R^2$ test statistic, and a sampling distribution of the $R^2$ test statistic simulated by parametric bootstrap. Similar methods exist for continuous distributions like the univariate normal and extreme-value distributions under regression tests of fit. Simulated examples as well as historically well-known Poisson data sets are then used to illustrate the proposed goodness-of-fit test for Poissonity.
# TABLE OF CONTENTS

ABSTRACT ................................................................................................................................. iii

LIST OF FIGURES ..................................................................................................................... vi

LIST OF TABLES ....................................................................................................................... vii

ACKNOWLEDGEMENTS ................................................................................................................. viii

CHAPTER 1  INTRODUCTION ................................................................................................. 1
  1.1 Graphical Goodness-of-Fit Procedures ................................................................. 2
  1.2 The Chi-Square Goodness-of-Fit Test .................................................................. 16
  1.3 Tests of Fit Based on EDF Statistics .................................................................... 20
  1.4 Purpose and Significance of Present Work........................................................... 23

CHAPTER 2  THE POISSON DISTRIBUTION ....................................................................... 27
  2.1 Applications of The Poisson Model ...................................................................... 32
  2.2 Existing Tests of Fit For Poissonity...................................................................... 41

CHAPTER 3  PROPOSED TEST FOR POISSONITY ................................................................. 50
  3.1 Construction of Poisson Q-Q Plot........................................................................ 51
  3.2 The $R^2$ Test Statistic ....................................................................................... 52
  3.3 The Simulated Sampling Distribution of $R^2$ .................................................... 55
  3.4 Power of $R^2$ Test for Poissonity ................................................................. 62
  3.5 Summary of Proposed Test For Poissonity........................................................ 62

CHAPTER 4  EXAMPLES ........................................................................................................... 64
  4.1 Death By Horse Kick ........................................................................................ 66
  4.2 Flying Bomb Hits ............................................................................................... 70
  4.3 Rainstorms In The Mid-West ............................................................................ 73
  4.4 Noxious Weed Seeds ......................................................................................... 76
  4.5 Alpha Emission From Radioactive Source ....................................................... 80

CHAPTER 5  DISCUSSION AND RECOMMENDATIONS ....................................................... 84

APPENDICES  NOTES ON THE POISSON DISTRIBUTION .................................................. 86
  A.1 Derivation of The Poisson Distribution .............................................................. 86
  A.2 Poisson Distribution and PDF Properties .......................................................... 88
  A.3 MGF of a Poisson Distribution ......................................................................... 89
  A.4 Mean and Variance of The Poisson Distribution ............................................. 90
REFERENCES ..................................................................................................................... 93
VITA ..................................................................................................................................... 99

v

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
LIST OF FIGURES

Figure 1.1 ECDF Plot of Sample From $N(100,25)$ ............................................................ 6
Figure 1.2 ECDF Plot of Sample From $\text{Poisson}(\lambda = 5)$ ............................................ 7
Figure 1.3 Discrepant Empirical and Hypothesized CDFs .............................................. 8
Figure 1.4 Standard Normal Theoretical Q-Q Plots .......................................................... 11
Figure 1.5 $\text{Poisson}(\lambda = 5)$ Theoretical Q-Q Plots ..................................................... 12
Figure 1.6 Q-Q Plots: Normals With Differing Locations and/or Spreads ......................... 14
Figure 1.7 Q-Q Plots With Outliers and Mixed Data ...................................................... 15
Figure 1.8 Comparison of Kolmogorov-Smirnov Statistics $D$ and $D^*$ ......................... 20
Figure 2.1 Comparison of Poisson PDFs With Differing $\lambda$ Values ............................. 29
Figure 2.2 Poisson CDF Plots With $\lambda$ Corresponding to Those in Figure 2.1 ............... 30
Figure 3.1 Q-Q Plot of Data Simulated from $\text{Poisson}(\lambda = 2)$ ............................... 52
Figure 3.2 S-PLUS Regression Output for Data Simulated from $\text{Poisson}(\lambda = 2)$ ...... 54
Figure 3.3 Program To Simulate Distribution of $R^2$ Under Poissonity .......................... 55
Figure 3.4 Empirical Distribution of $R^2_{25,2,16}$ .......................................................... 56
Figure 3.5 Specified Critical Value $R^2_{\alpha,5,0.05}$ ...................................................... 59
Figure 3.6 Output of Critical Values For $R^2_{25,2,16}$ Distribution ............................... 59
Figure 4.1 S-PLUS Program To Output components of $R^2$ Test ...................... 66
Figure 4.2 Q-Q Plot of Horse Kick Data with $R^2$ Statistic ........................................... 68
Figure 4.3 Sampling Distribution for $R^2_{350,0.64}$ with Critical Values ....................... 69
Figure 4.4 Q-Q Plot of Flying Bomb Hits Data with $R^2$ Statistic ............................... 71
Figure 4.5 Sampling Distribution for $R^2_{256,0.93}$ with Critical Values ..................... 72
Figure 4.6 Q-Q Plot of Rainstorm Data with $R^2$ Statistic ........................................... 75
Figure 4.7 Sampling Distribution for $R^2_{330,1.20}$ with Critical Values ....................... 76
Figure 4.8 Q-Q Plot of Weed Seed Data with $R^2$ Statistic ........................................... 78
Figure 4.9 Sampling Distribution for $R^2_{98,3.02}$ with Critical Values ....................... 79
Figure 4.10 Q-Q Plot of Alpha Emission Data with $R^2$ Statistic ............................... 81
Figure 4.11 Sampling Distribution for $R^2_{2608,3.87}$ with Critical Values ................... 82

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 1.1</td>
<td>Empirical CDF of Sample From $N(100,25)$</td>
<td>5</td>
</tr>
<tr>
<td>Table 1.2</td>
<td>Empirical CDF of Sample From $Poisson (\lambda = 5)$</td>
<td>6</td>
</tr>
<tr>
<td>Table 2.1</td>
<td>Death by Horse Kick of Members of The Prussian Army</td>
<td>34</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Connections To A Wrong Number</td>
<td>35</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Traffic Arrivals in 30-Second Intervals</td>
<td>36</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Flying-Bomb Hits on London During World War II</td>
<td>37</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>Noxious Weed Seeds in Subsamples of Meadow Grass</td>
<td>38</td>
</tr>
<tr>
<td>Table 2.6</td>
<td>10-Minute Excessive Rainstorms in Mid-West Region</td>
<td>39</td>
</tr>
<tr>
<td>Table 2.7</td>
<td>Alpha Emission From Radioactive Source</td>
<td>40</td>
</tr>
<tr>
<td>Table 2.8</td>
<td>Annual Number of Serious Earthquakes</td>
<td>41</td>
</tr>
<tr>
<td>Table 2.9</td>
<td>Data Simulated from $Poisson (\lambda = 2)$</td>
<td>50</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Selected Portions of The Empirical Distribution of $R_{25,2.16}^2$</td>
<td>57</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Lower-tail Critical Values $R_{x,\lambda;0.05}^2$</td>
<td>61</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Empirical and Binned Death By Horse Kick Data</td>
<td>67</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Selected Portion of Empirical Distribution of $R_{200,0.61}^2$</td>
<td>70</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Empirical and Binned Flying Bomb Hits Data</td>
<td>71</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Selected Portion of Empirical Distribution of $R_{576,0.93}^2$</td>
<td>73</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Empirical and Binned Rainstorm Data</td>
<td>74</td>
</tr>
<tr>
<td>Table 4.6</td>
<td>Selected Portion of Empirical Distribution of $R_{330,1.20}^2$</td>
<td>76</td>
</tr>
<tr>
<td>Table 4.7</td>
<td>Empirical and Binned Noxious Weed Seed Data</td>
<td>77</td>
</tr>
<tr>
<td>Table 4.8</td>
<td>Selected Portion of Empirical Distribution of $R_{98,3.02}^2$</td>
<td>79</td>
</tr>
<tr>
<td>Table 4.9</td>
<td>Empirical and Binned Alpha Emission Data</td>
<td>80</td>
</tr>
<tr>
<td>Table 4.10</td>
<td>Selected Portion of Empirical Distribution of $R_{2568,3.37}^2$</td>
<td>83</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

Accomplishments are rarely just a reflection of an individual’s own ambition; rather, it is a culmination of inner perseverance and the collective efforts of mentors, family, and friends in supporting, inspiring, and encouraging the individual to achieve their desired goals. My case is no exception. A very special thank you goes out to the following people:

☐ To my advisor, thesis committee chair, and mentor, Dr. Ashok K. Singh: your knowledge, compassion for students, and willingness to share your intelligent ideas with your students, serves as a model for all professors at the graduate level. You are truly a rare find. Thank you so much for investing your time in me, and for all of your help and guidance while I have been a graduate student here at UNLV.

☐ To Dr. Malwane Ananda, Dr. Rohan Dalpatadu, and Dr. Laxmi Gewali: thank you for taking the time out of your busy schedules to serve on my thesis examination committee.

☐ To my colleagues and friends, namely, Monica Solis, Tom Stanek, Maria Teruel, and Alex Berejnoi: thank you for helping me to balance the rigors of studying and research with extreme, but clean fun in this city of Las Vegas. My experience would not have been as enjoyable had our paths not crossed. I am taking home with me lots of fond (and crazy) memories, and I will surely miss all of you. I wish all of you much success wherever your futures may lead.
To Dr. Carole K. Miura: you are absolutely among the best mathematics and statistics professors out there. You helped me to understand the usefulness and power of statistics early on in my undergraduate career. Your inspiring motivation has continually been brewing in me ever since.

Finally, and most especially:

To my beloved family back home in Hilo, Hawaii: your love and endless encouragement and support has kept me emotionally strong and focused while I have been away. I love you guys. A special thanks goes out to Mom in particular: you always knew what to say and when to say the right things to help me believe that this moment was undoubtedly attainable. Thank you!
CHAPTER 1

INTRODUCTION

At the core of all statistical analyses, there exists a model that attempts to describe the underlying structure or relationship of some phenomena on which measurements are taken. Statistical tests, estimation procedures, and inference are based on these sampled measurements (or data) and a hypothesized model. Procedures used to verify and validate these model or distributional assumptions are known as goodness-of-fit (GOF) tests. Goodness-of-fit procedures, given a random sample $X_1, X_2, \ldots, X_n$, are used to test the hypothesis

$$H_0 : \text{Sample is from a population with distribution function } F(x)$$

$$H_1 : \text{Sample is not from a population with distribution function } F(x).$$

The null hypothesis $H_0$ is either simple, specifying the proposed theoretical distribution $F(x)$ completely with given parameter values; or composite, where $F(x)$ is stated without specifying values for its parameters. In most applications of goodness-of-fit, the alternative hypothesis $H_1$ is composite simply stating that $H_0$ is false.

Over the years, many have continued to recognize the fundamental importance of goodness-of-fit and thus, a vast number of test procedures and techniques have come about (see Cochran, W. G., 1954; Stephens, M.A., 1974; Wilk and Gnanadesikan, 1968; D’Agostino and Stephens, 1986; Rayner and Best, 1989; Gürtler and Henze, 2000; Aslan
and Zech, 2002; and Huber-Carol et al., 2002). These goodness-of-fit techniques are based on test statistics that measure in some way the consistency – or equivalently, the discrepancy – of a sample of data with the hypothesized distribution. These test procedures are either distribution dependent, where they are applicable to a specific distribution, or distribution-free, where they can be applied to an arbitrary distribution (Aslan and Zech, 2002). There exist both graphical techniques and formal numerical methods; some goodness-of-fit tests are a combination of both. In any case, involved and complicated procedures detract from usefulness and thus, practicality takes precedence.

We now present the reader with a resourceful synopsis of widely accepted and commonly used goodness-of-fit techniques in practice today. We reserve our discussion in this section to distribution-free goodness-of-fit tests that have a clear motivation, that are easily understood by the practical statistician, and those that have been well documented in the vast literature on the subject. Furthermore, we acknowledge within each respective goodness-of-fit test both advantages and setbacks, its potential adaptability to distributions of either the continuous or discrete type, and when the convenience of statistical software and computer applications exist. Although the techniques to be discussed are practical, we refer the reader to D’Agostino and Stephens (1986) and Huber-Carol et al. (2002) as the methods presented in this section may not necessarily be the most powerful that exists for the distribution in question.

1.1 Graphical Goodness-of-Fit Procedures

In goodness-of-fit problems, graphical techniques provide us with simple and effective means of evaluating the fit of a proposed probability model through visual
assessment (Gan, Koehler, and Thompson, 1991). These graphical techniques are valuable exploratory tools in helping the statistician to understand numerous relationships and characteristics present within the data that are not readily revealed by their numerically involved counterparts. Although graphical analysis is considered less formal than the numerical techniques to be discussed in later sections of this chapter, graphical analysis may supplement numerical methods. In general, it is recommended that formal numerical tests of fit procedures be preceded by graphical analyses (D’Agostino, 1986).

Here, we present two commonly used graphical goodness-of-fit techniques that may be applied to a distribution of either the continuous or discrete type. In particular, we discuss the *empirical cumulative distribution function* (ECDF) plot as well as the related *probability* plot, commonly referred to as the *theoretical quantile-quantile* (Q-Q) plot. By related, we mean that one is approximately equivalent to the other by a simple transformation on vertical axes.

**The Empirical CDF Plot**

Let $X_{(i)}$, $i = 1, 2, \ldots, n$, denote the $i$th order statistic of a sample of size $n$, so that $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ form the *order statistics* of a random sample $X_1, X_2, \ldots, X_n$. If

$$\#(X_i \leq x) = \# \text{ of observations} \leq x,$$  \hspace{1cm} (1.2)

we can define the *empirical distribution function* (EDF), also known as the *empirical cumulative distribution function* (ECDF), to be

$$F_n(x) = \begin{cases} 
0 & \text{if } x < x_{(i)} \\
\frac{\#(X_i \leq x)}{n} & x_{(i)} \leq x < x_{(n)} \\
1 & x_{(n)} \leq x 
\end{cases} \quad (1.3)$$
The empirical distribution function $F_n(x)$ is essentially a step function, which gives the relative frequency of the event that $X_{(i)} \leq x$. When plotted, $F_n(x)$ provides an exhaustive representation of the data that can be visually compared for consistency with the distribution $F(x)$ corresponding to $H_0$ in (1.1) (Wilk and Gnanadesikan, 1968; D’Agostino, 1986; Aslan and Zech, 2002). Furthermore, for large sample size $n$, $F_n(x)$ strongly converges to $F(x)$ for all $x$ (see Rényi, 1970; D’Agostino, 1986).

The use of the empirical CDF is independent of any specification of a parametric distribution and may usefully describe data even when random sampling has not been employed (Wilk and Gnanadesikan, 1968). Additional advantages of using the empirical CDF in data analysis as taken from Wilk and Gnanadesikan (1968) include: (i) it lends itself to graphical representation and immediately supplies direct information regarding the shape of the underlying distribution; (ii) the complexity of the graph or plot is independent of the number of observations; (iii) it is invariant under monotone transformation in the sense of quantiles; however, not in appearance; (iv) it is a robust supplier of information on location and scale or spread; (v) it is an effective indicator of peculiarities such as outliers; and (vi) it does not involve arbitrary binning or grouping difficulties that arise with the use of histograms.

To construct an empirical CDF plot:

1. Sort the $n$ observations, $x_i$, $i = 1, 2, \ldots, n$, to obtain the ordered statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

2. Calculate the empirical probability, $f_n(x) = \Pr(X_{(i)} = x) = \frac{\# \text{ of observations} = x}{n}$, for
\[ x_{(i)} \leq x \leq x_{(n)} \] Here, the empirical probability \( f_n(x) \) is the height of the step, or jump, at \( x \) in the space \( X_{(i)} \).

3. Plot the \( i \)th ordered observed value of the sample, \( x_{(i)}, i = 1, 2, \ldots, n \), as abscissa (horizontal axis) against its respective empirical cumulative probability, 

\[ F_n(x_{(i)}) = \Pr(X_{(i)} \leq x_{(i)}) = \sum_{j=1}^{i} f_n(x_{(j)}), \] as ordinate (vertical axis), for \( i = 1, 2, \ldots, n \).

The empirical CDF plot is a standard task function in most statistical software packages like the Statistical Analysis System (SAS), the Statistical Package for the Social Sciences (SPSS), and S-PLUS by Insightful Corp. Users of MINITAB or any statistics software program with built-in plotting utilities may also create an ECDF plot using the procedure above and the plot function from the drop-down graph menus; however, the actual step lines need to be drawn manually from the graph editing options.

### Table 1.1. Empirical CDF of Random Sample (\( n = 30 \)) from \( N(100, 25) \)

<table>
<thead>
<tr>
<th>( i )</th>
<th>( x_{(i)} )</th>
<th>( F_n(x_{(i)}) )</th>
<th>( i )</th>
<th>( x_{(i)} )</th>
<th>( F_n(x_{(i)}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.260</td>
<td>0.033</td>
<td>16</td>
<td>101.195</td>
<td>0.533</td>
</tr>
<tr>
<td>2</td>
<td>90.402</td>
<td>0.067</td>
<td>17</td>
<td>101.314</td>
<td>0.567</td>
</tr>
<tr>
<td>3</td>
<td>92.047</td>
<td>0.100</td>
<td>18</td>
<td>101.461</td>
<td>0.600</td>
</tr>
<tr>
<td>4</td>
<td>93.450</td>
<td>0.133</td>
<td>19</td>
<td>101.534</td>
<td>0.633</td>
</tr>
<tr>
<td>5</td>
<td>93.555</td>
<td>0.167</td>
<td>20</td>
<td>101.553</td>
<td>0.667</td>
</tr>
<tr>
<td>6</td>
<td>95.208</td>
<td>0.200</td>
<td>21</td>
<td>101.882</td>
<td>0.700</td>
</tr>
<tr>
<td>7</td>
<td>95.732</td>
<td>0.233</td>
<td>22</td>
<td>103.206</td>
<td>0.733</td>
</tr>
<tr>
<td>8</td>
<td>96.573</td>
<td>0.267</td>
<td>23</td>
<td>103.448</td>
<td>0.767</td>
</tr>
<tr>
<td>9</td>
<td>96.888</td>
<td>0.300</td>
<td>24</td>
<td>103.460</td>
<td>0.800</td>
</tr>
<tr>
<td>10</td>
<td>97.644</td>
<td>0.333</td>
<td>25</td>
<td>103.517</td>
<td>0.833</td>
</tr>
<tr>
<td>11</td>
<td>98.520</td>
<td>0.367</td>
<td>26</td>
<td>104.691</td>
<td>0.867</td>
</tr>
<tr>
<td>12</td>
<td>99.130</td>
<td>0.400</td>
<td>27</td>
<td>104.787</td>
<td>0.900</td>
</tr>
<tr>
<td>13</td>
<td>99.572</td>
<td>0.433</td>
<td>28</td>
<td>104.834</td>
<td>0.933</td>
</tr>
<tr>
<td>14</td>
<td>99.630</td>
<td>0.467</td>
<td>29</td>
<td>105.082</td>
<td>0.967</td>
</tr>
<tr>
<td>15</td>
<td>100.225</td>
<td>0.500</td>
<td>30</td>
<td>109.118</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 1.1. ECDF plot of a random sample \((n = 30)\) from \(N(100, 25)\) compared with the CDF plot of the theoretical \(N(100, 25)\).

Table 1.1 displays the empirical CDF \(F_n(x)\) of a sample of size \(n = 30\) from the continuous normal distribution with mean \(\mu = 100\) and standard deviation \(\sigma = 5\) (i.e.

<table>
<thead>
<tr>
<th>(i)</th>
<th>(x_{ij})</th>
<th>(freq(x_{ij}))</th>
<th>(F_n(x_{ij}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 2</td>
<td>1</td>
<td>2</td>
<td>0.067</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.100</td>
</tr>
<tr>
<td>4 - 8</td>
<td>3</td>
<td>5</td>
<td>0.267</td>
</tr>
<tr>
<td>9 - 13</td>
<td>4</td>
<td>5</td>
<td>0.433</td>
</tr>
<tr>
<td>14 - 18</td>
<td>5</td>
<td>5</td>
<td>0.600</td>
</tr>
<tr>
<td>19 - 22</td>
<td>6</td>
<td>4</td>
<td>0.733</td>
</tr>
<tr>
<td>23 - 25</td>
<td>7</td>
<td>3</td>
<td>0.833</td>
</tr>
<tr>
<td>26</td>
<td>8</td>
<td>1</td>
<td>0.867</td>
</tr>
<tr>
<td>27</td>
<td>9</td>
<td>1</td>
<td>0.900</td>
</tr>
<tr>
<td>28 - 29</td>
<td>10</td>
<td>2</td>
<td>0.967</td>
</tr>
<tr>
<td>30</td>
<td>11</td>
<td>1</td>
<td>1.000</td>
</tr>
</tbody>
</table>

Table 1.2. Empirical CDF of Random Sample \((n = 30)\) from Poisson \((\lambda = 5)\)
Figure 1.2. ECDF plot of a random sample \((n = 30)\) from \(\text{Poisson} (\lambda = 5)\) compared with the CDF plot of the theoretical \(\text{Poisson} (\lambda = 5)\).

\(\mathcal{N}(100, 25)\) and the corresponding ECDF plot is pictured in Figure 1.1. Similarly, Table 1.2 lists the empirical CDF of a sample of size \(n = 30\) from the discrete \(\text{Poisson}\) model with parameter \(\lambda = 5\), and is accompanied by the related ECDF plot in Figure 1.2. Through visual inspection, both \(F_n(x)\) in Figure 1.1 and Figure 1.2 appear to be in close conformity to the underlying distribution \(F(x)\). Contrary to this, Figure 1.3 clearly displays marked differences between the empirical and hypothesized CDFs being compared. In many situations, however, the consistency of the plot of \(F_n(x)\) with that of \(F(x)\) is difficult to perceive and evaluate with the human eye. In any case, if one is attempting to judge the consistency or discrepancy between the two plots through visual
assessment, it is probably easiest to base a decision on whether or not a set of points deviates from a straight line.

![Discrepant Empirical and Hypothesized CDFs](image)

Figure 1.3. Plot of discrepant empirical and hypothesized CDFs.

The Probability Plot or Theoretical Q-Q Plot

Let $Q(p)$, where $0 < p < 1$, denote the $p$th quantile. By quantile, or similarly a percentile, we mean the fraction (or percent) of points below a given value. For instance, the 0.85 quantile, or $Q(.85)$, is the point or value at which 85% of the data fall below and 15% lie above that given value. Hence, a probability plot or theoretical quantile-quantile (Q-Q) plot is a plot of empirical quantiles, or equivalently, the quantiles of a set of observed data, against the corresponding quantiles of the theoretical distribution, $F(x)$. We define and calculate the theoretical quantiles using
\[ Q_F(p_i) = F^{-1}(p) \quad (1.4) \]

for \( i = 1, 2, \ldots, n \). Here, \( F^{-1}(\cdot) \) denotes the inverse transformation of \( F(x) \), namely, the inverse CDF of \( X \) from \( F(x) \), while \( p_i \) is a plotting position which we will define as

\[ p_i = \frac{i - 0.5}{n} \quad (1.5a) \]

for \( i = 1, 2, \ldots, n \). Various plotting positions have been proposed with many of them being of the form

\[ p_i = \frac{i - c}{n - 2c + 1} \quad (1.5b) \]

for \( 0 \leq c \leq 1 \); however, the plotting position \( p_i \) defined in (1.5a) is typically used in practice (Looney and Gulledge, 1985; D’Agostino, 1986; Gan et. al, 1991). We refer the reader to Kimball (1960), Barnett (1975), Cunnane (1978), Chambers et. al (1983), Harter (1984), Looney and Gulledge (1984) and (1985), Harter and Weigund (1985), and D’Agostino (1986) for further discussion on selecting a plotting position \( p_i \), as the authors justify alternative plotting positions for specific distributions and for different applications of probability plotting. Finally, we define the empirical quantiles to be

\[ Q_{F_x}(p_i) = x_{(i)} \quad (1.6) \]

for \( i = 1, 2, \ldots, n \). Thus, the empirical quantiles are just the ordered observed data values themselves, \( x_{(i)} \).

To construct a theoretical Q-Q plot for a sample of size \( n \) from a distribution with location and scale parameters \( \alpha \) and \( \beta \), respectively:

1. Sort the \( n \) observations, \( x_i \), \( i = 1, 2, \ldots, n \), to obtain the ordered statistics.
\( x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)} \). Take \( x_{(i)} , \ i = 1, 2, \ldots, n \), to be the quantiles of the observed data, \( Q_{F_x}(p_i) \).

2. Calculate the quantiles for the hypothesized theoretical distribution \( F(x) \) using \( Q_{F_{x}}(p_i) \) defined in (1.4) for \( i = 1, 2, \ldots, n \). If \( \alpha \) and \( \beta \) are unknown, use \( \hat{\alpha} \) and \( \hat{\beta} \), which are respectively the maximum likelihood estimates (MLE) of \( \alpha \) and \( \beta \) calculated from the observed data, \( x_i , \ i = 1, 2, \ldots, n \), in the calculation of \( Q_{F_{x}}(p_i) \) (Gan et al, 1991).

3. Plot the empirical quantiles, \( x_{(i)} \), as abscissa (horizontal axis), against the calculated theoretical quantiles \( Q_{F_{x}}(p_i) \) as ordinate (vertical axis).

The probability plot or Q-Q plot is a routine task function in most statistical software packages like the Statistical Analysis System (SAS), the Statistical Package for the Social Sciences (SPSS), S-PLUS, and MINITAB. Some of these statistical packages even offer specific distributional Q-Q and probability plots; however, most are for commonly utilized continuous-type distributions like the standard normal distribution, \( N(0,1) \). In either the discrete or continuous case, a statistics software program with the power to sort data, calculate distributional quantiles, and one that has a basic plot function will suffice.

Standard normal theoretical Q-Q plots constructed using \( p_i \) defined in (1.5a) and random samples of size \( n = 10, 25, 50, \) and \( 100 \) from \( N(0,1) \) are shown in Figure 1.4.

Figure 1.5 displays discrete Poisson theoretical Q-Q plots obtained using the plotting position defined in (1.5a) and random samples of size \( n = 10, 25, 50, \) and \( 100 \) from a Poisson distribution with parameter \( \lambda = 10 \). Through visual inspection of Figures 1.4
and 1.5, it can easily be seen that there exists a positive linear relationship between the empirical quantiles, \( x_{(i)} \), and the hypothesized theoretical quantiles, \( Q_p(p_i) \).

![Continuous standard normal theoretical Q-Q plots with various sample sizes.](image)

Figures 1.4 (a) – (d). Continuous standard normal theoretical Q-Q plots with various sample sizes.

It is clearly evident from Figures 1.4 and 1.5 that the Q-Q plot is a graphical tool that lends itself nicely to a linear configuration, given that the theoretical distribution \( F(x) \) is a close approximation to the empirical distribution \( F_n(x) \). To see this, suppose that \( X \) and \( Y \) are identically distributed variables. Then the plot of \( Q_x(p_i) \) versus \( Q_y(p_i) \) will
intuitively be a straight line, namely \( y = x \), or in terms of quantiles \( Q_Y(p_i) = Q_X(p_i) \), through the origin \((0,0)\) with slope equal to 1. Hence, the line \( y = x \) is used as a reference measure of how well the hypothesized distribution \( F(x) \) fits the observed data; however, departures from the reference line \( y = x \) do not necessarily imply lack of fit.

In some cases, the Q-Q line will be a straight line that neither passes through the origin \((0,0)\) nor has slope equal to 1 and thus, departs from the reference line \( y = x \). For instance, the Q-Q line might differ from the line \( y = x \) by both an additive and
multiplicative constant. Here, we consider the variables $X$ and $Y$, where $Y$ is now a linear function of $X$, namely

$$Y = k \cdot X + c,$$  \hspace{1cm} (1.7a)

for constants $k$ and $c$. In terms of $X$-quantiles and $Y$-quantiles, this line will equivalently be

$$Q_y(p) = k \cdot Q_x(p) + c.$$  \hspace{1cm} (1.7b)

Thus, the corresponding Q-Q plot of $X$ and $Y$ will remain linear but with possible change in location (or intercept) and slope (or spread) from the reference line $y = x$. The effect on location and spread, as noted in Chambers et al. (1983), would approximately be

$$\text{location}(Y) = k \cdot \text{location}(X) + c$$  \hspace{1cm} (1.8)

and

$$\text{spread}(Y) = k \cdot \text{spread}(X).$$  \hspace{1cm} (1.9)

This valuable property of Q-Q plots is known as linear invariance (Wilk and Gnanadesikan, 1968). Linear invariance establishes that a single theoretical Q-Q plot not only compares a set of data to one theoretical distribution with specified parameters, but simultaneously to a whole family of that distribution with differing location and scale (or spread) parameters (Chambers et al., 1983). For instance, see Figures 1.6 (a) – (c), where a standard normal probability plot can be used to sufficiently test the fit of observed data arising from any arbitrary normal distribution. Hence, straightness of the theoretical Q-Q plot with shifts or tilts away from the reference line $y = x$ indicates that the empirical and hypothesized theoretical distributions are of the same family but differ in location and scale (or spread) parameters. Large and systematic departures from a straight line
Figures 1.6 (a) – (c). Q-Q plots of normal distributions with differing location and scale parameters.

then, are to be judged as a lack of fit between the reference distribution and the observed data. It is this straight-line test criterion that valuably makes the probability plot or theoretical Q-Q plot so practical and appealing.

The general usefulness of the Q-Q plot may be extended to obtaining informal estimates of unknown location and spread parameters of the observed data by estimating the respective intercept and slope of the Q-Q line from the probability plot (Chambers et al., 1983). This, of course, follows from the inversion of the linear invariance concept explained above. The Q-Q plot is also an effective indicator of possible outliers as well
Figure 1.7. (a) Q-Q plot of data with existing outliers. (b) Q-Q plot of data arising from a mixture of normal distributions.

as of contaminated data sets arising from a mixture of distributions as seen in Figures 1.7(a) and 1.7(b), respectively. The Q-Q plot can also be used to explore symmetry, skewness, and tail thickness in reference to a mismatch between the observed data and the hypothesized distribution. These distributional aspects are characterized by concave or convex curvature of the Q-Q line and curvature at both ends of the Q-Q line, respectively. See Wilk and Gnanadesikan (1968), Gerson (1975), Chambers et al. (1983), Harter (1984), and D’Agostino (1986) for further details on these applications of the Q-Q or probability plot.
It is obvious that both the ECDF plot and the Q-Q plot are useful tools for exploring distributional assumptions about a set of observed data. Furthermore, the simplicity of their construction and the opportunities for interpretation add to their merit. The power and interpretation of these plots, however, should be critically judged and evaluated in reference to the problem at hand. These plots can be extremely sensitive to random occurrences and variability in the data, especially in the case where the sample sizes are small, and may lead to hasty and incorrect conclusions about the underlying distribution (D’Agostino, 1986). Another limitation of these plots is that they compare the empirical distribution of one variable with that of a hypothesized distribution, thus, ignoring the relationship of this variable to other, possibly closer fitting, distributions (Chambers et. al, 1983). Hence, graphical goodness-of-fit procedures should be supplemented and used in conjunction with formal numerical goodness-of-fit techniques.

1.2 The Chi-Square ($\chi^2$) Goodness-of-Fit Test

Presented in 1900, the Karl Pearson chi-squared test remains among the oldest and most widely used formal statistical procedures in practice today (Moore, 1986). The chi-square goodness-of-fit test is a one-sample quantitative test that examines the discrepancies between the observed ($N_j$) and the expected ($np_j$) frequencies of $n$ observations grouped into $C$ classes with probability of occurrence $p_j$ for $j = 1, 2, \ldots, C$. The differences $N_j - np_j$ between observed cell frequencies and expected cell frequencies under the hypothesized distribution $F(x)$, explain a deviation
of the data $x_i$ from $F(x)$. Thus, the test statistic, due to Pearson, that is used as a measure of fit is

$$
\chi^2 = \sum_{j=1}^{c} \frac{(N_j - np_j)^2}{np_j},
$$

(1.10a)

where $\sum_{j=1}^{c} N_j = n$. Here, $\chi^2$ tends to be small when $H_0$ in (1.1) holds, and large when $H_0$ is false.

Under a simple $H_0$ and assuming the quantities $N_j - np_j$ have a limiting normal distribution, Pearson showed (see Moore, 1986) that $\chi^2$ defined in (1.10a) is approximately $\chi^2_{C-1}$ for large sample size $n$; however, for the case where we wish to test a composite $H_0$, R.A. Fisher (1924) showed that estimation of a parameter $\theta$ in $F(x)$ and the method used to estimate $\theta$ alters the large sample distribution of $\chi^2$ in (1.10a). Fisher argued that the appropriate method of fitting $\theta$ is through maximum likelihood estimation based on observed cell frequencies $N_j$ (Fisher, R.A., 1924). Hence, this led to the Pearson-Fisher chi-square test statistic

$$
\chi^2(\hat{\theta}_n) = \sum_{j=1}^{c} \frac{(N_j - np_j(\hat{\theta}_n))^2}{np_j(\hat{\theta}_n)},
$$

(1.10b)

where $\hat{\theta}_n$ is the MLE of the parameter $\theta$ (Moore, 1986; Rayner, G.D., 2002). Fisher then showed that under the null hypothesis, $\chi^2(\hat{\theta}_n)$ defined in (1.10b) has the $\chi^2_{C-t-1}$ distribution where $C$ is as defined before and $t$ being the number of parameters estimated. Hence, the limiting distribution of $\chi^2$ loses one additional degree of freedom for each

In applying the chi-square goodness-of-fit test to the alternatives in (1.1):

1. Calculate the expected cell frequencies, \( n\hat{p}_j \), for \( j = 1, 2, \ldots, C \), where \( \hat{p}_j \) depends on either the parameter values specified in \( H_0 \) or on the MLE, \( \hat{\theta}_n \). To ensure that the asymptotic properties of \( \chi^2 \) hold, it is often recommended that \( n\hat{p}_j \geq 5 \) for each class \( j \), and that the neighboring classes be combined if this requirement is not met (Cochran, 1954). Snedecor and Cochran (1954) suggest a less conservative rule that all \( n\hat{p}_j \) should be at least 1, with at least 80 percent being at least 5.

2. Sort and bin the frequency data into \( C \) non-overlapping classes or intervals according to the partitioning or grouping of the expected frequencies, \( n\hat{p}_j \), calculated above.

3. Calculate for each class, the quantity \( \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j} \), and obtain the test statistic by summing these quantities over the \( j \) classes or cells.

Test Statistic: \[ \chi^2 = \sum_{j=1}^{C} \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j} \]

4. For any significance level \( \alpha \), where \( \alpha = \Pr \left( \chi^2 > \chi^2_{C-r-1;l-\alpha} \right) \), the decision rule is:

- Reject \( H_0 \) if \( \chi^2 > \chi^2_{C-r-1;l-\alpha} \) in favor of \( H_1 \) that there is lack of fit.
- Do not reject \( H_0 \) if \( \chi^2 \leq \chi^2_{C-r-1;l-\alpha} \) and say there is reasonable fit.

The chi-squared test is one of the most practical tests of fit due to its ease of use and flexibility in a variety of situations (see Cochran, 1954; Moore, 1986). The computation of the test statistic \( \chi^2 \) and of critical values for this test statistic is relatively simple with
the use of statistical software. In fact, the chi-square goodness-of-fit test is a built in routine in most statistical software packages like the Statistical Analysis System (SAS), the Statistical Package for the Social Sciences (SPSS), and S-PLUS. In any case, MINITAB or a comparable statistics software program that has the power to calculate critical values from a chi-square distribution, sort data, calculate expected distributional frequencies, and do basic computations from lists will be sufficient in employing the test.

In terms of flexibility, the chi-square test of fit applies to both continuous and discrete univariate distributions and can easily be adapted to the case when parameters of a distribution are estimated. It may also be extended to multivariate cases or when censored data is involved. Hence, the chi-square test is the most generally applicable test of fit (Moore, 1986). This flexibility, however, is one of several factors at the root of weakness and inherent problems of using the chi-square test of fit.

In addition to being based on large sample theory, the chi-square goodness-of-fit test also suffers from relative lack of power in the sense that it is often insensitive, and does not indicate significant results when the null hypothesis is actually false (Cochran, 1954; Moore, 1986). When the test does indicate significance, it does not shed light on the way in which the hypothesized distribution in (1.1) disagrees with the observed frequency distribution. The relative lack of power of the chi-square test of fit is in part due to the necessity to group data. The choice of binning here is arbitrary, especially for continuous distributions. Furthermore, following a rule that $n \hat{p}_j \geq 5$ will likely require grouping of classes or cells at the tails or extremes of the distribution – where differences between the two distributions are usually more pronounced – thus, disguising valuable distributional information. Hence, we introduce two alternative binning-free formal tests of fit based
on the empirical distribution function (EDF), which are superior in power, at least for the continuous case, to the chi-square goodness-of-fit test (Moore, 1986).

1.3 Tests Of Fit Based On EDF Statistics

Goodness-of-fit tests based on EDF statistics measure the discrepancy between the distribution $F(x)$ hypothesized in (1.1) and the empirical distribution function $F_n(x)$ defined in (1.3). There exist two major classes of EDF statistics, namely supremum statistics and quadratic deviation statistics. In this section, we direct attention to three of the leading statistics in the class of EDF tests of fit (see Stephens, 1974). For supremum statistics, we concentrate on the well-known Kolmogorov-Smirnov statistic $D$. For quadratic statistics, we discuss the Cramér-von Mises statistic $W^2$ and the Anderson-Darling statistic $A^2$.

![Comparison of Empirical and Theoretical CDFs](image)

Figure 1.8. Plot displaying the Kolmogorov-Smirnov statistics $D^+$ and $D^-$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Let $D^+$ denote the maximum positive deviation of $F_n(x)$ from $F(x)$. Likewise, let $D^-$ denote the maximum negative deviation of $F_n(x)$ from $F(x)$ (see Figure 1.8). The Kolmogorov-Smirnov test statistic $D$, as proposed by Kolmogorov, is then the maximum absolute deviation of the empirical distribution $F_n(x)$ from the hypothesized distribution $F(x)$. Hence, formal definitions for the statistics $D^+$, $D^-$, and $D$ are given as

$$D^+ = \sup_x \{F_n(x) - F(x)\}$$
(1.11a)

$$D^- = \sup_x \{F(x) - F_n(x)\}$$
(1.11b)

$$D = \sup_x \{|F_n(x) - F(x)|\}.$$  
(1.11c)

The Cramér-von Mises statistic $W^2$ and the Anderson-Darling statistic $A^2$ originate from a family of tests that measure the integrated quadratic deviation,

$$QD = n \int_{-\infty}^{\infty} \left[ F(x) - F_n(x) \right]^2 \psi(x) \, dF(x),$$
(1.12)

of $F_n(x)$ from $F(x)$, subject to a suitable weighting function $\psi(x)$ (Aslan and Zech, 2002). The weighting function determined for the Cramér-von Mises statistic $W^2$ is $\psi_{CM}(x) = 1$ and we obtain

$$W^2 = n \int_{-\infty}^{\infty} \left[ F(x) - F_n(x) \right]^2 dF(x).$$
(1.13)

If $\psi_{AD}(x) = \left( F(x) [1 - F(x)] \right)^{-1}$ is chosen as the appropriate weighting function, we arrive at the Anderson-Darling test statistic

$$A^2 = n \int_{-\infty}^{\infty} \frac{F(x) - F_n(x)}{F(x) [1 - F(x)]}^2 \, dF(x),$$
(1.14)
which heavily weights deviations near the extremes or tail ends of $F(x)$. See Stephens (1986) and Aslan and Zech (2002) for justification of weighting functions.

Although EDF statistics provide more powerful tests of fit than $\chi^2$ statistics, they are neither well suited for discrete distributions, nor adaptable in cases when the parameters of $F(x)$ must be estimated from the observed values. EDF statistics are also considered more difficult to compute than $\chi^2$ in (1.10a) and (1.10b). These limitations and difficulties of using EDF statistics have long prevented their wider use and application in practice (see Stephens, 1974 and 1986); however, contributions made over the years to the study of EDF statistics has made their use more practical for the case when the hypothesized distribution $F(x)$ is continuous and completely specified as well as for the distribution dependent cases when $F(x)$ being tested is normal or exponential, with parameters to be estimated (Stephens, 1974).

Having sorted the $n$ observations to obtain the ordered observed values $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, the Kolmogorov-Smirnov statistic $D$, the Cramér-von Mises statistic $W^2$, and the Anderson-Darling statistic $A^2$ can be calculated from the following computing formulas given in Stephens (1974):

1. **Kolmogorov-Smirnov Statistics** $D^+$, $D^-$, and $D$:

   $$D^+ = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - z_i \right\}$$  \hspace{1cm} (1.15a)

   $$D^- = \max_{1 \leq i \leq n} \left\{ z_i - \frac{(i-1)}{n} \right\}$$  \hspace{1cm} (1.15b)

   $$D = \max \left( D^+, D^- \right)$$  \hspace{1cm} (1.15c)

2. **Cramér-von Mises Statistic** $W^2$:
i = 1, 2

\[ W^2 = \sum_{i=1}^{n} \left( z_i - \frac{(2i-1)}{2n} \right)^2 + \frac{1}{12n} \]  

(1.16)

3. Anderson-Darling Statistic \( A^2 \):

\[
A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1) \left[ \ln z_i + \ln \left( 1 - z_{n+1-i} \right) \right] 
\]

(1.17)

In the computing formulas above, \( z_i = F(x_{(i)}) \) for \( i = 1, 2, \ldots, n \), is the cumulative probability of a continuous distribution at the value \( x_{(i)} \) that can be found in either standard tables or with the use of statistical software. We refer the reader to Stephens (1974) for specific conditions, as in when transformations to a uniform distribution or standardized distribution are employed, under which \( z_i \) is calculated. The calculated EDF statistic \( D \) (or \( W^2 \), or \( A^2 \)) is then compared to a corresponding tabled critical value given sample size \( n \) and percentage points or significance level. Published tables for these EDF statistics, either produced from simulation studies or modified from a previous source, can be found and referenced in: Marshall (1958); Lewis (1961); Van Soest (1967); Lilliefors (1967) and (1969); Stephens (1969), (1970a), (1970b), and (1974); Pearson and Hartley (1972); and Chen (2002). The decision rule then is to reject \( H_0 \) if the calculated EDF statistic exceeds its corresponding significant tabled value; otherwise, do not reject the null hypothesis.

1.4 Purpose and Significance of Present Work

Having surveyed the vast collection of literature on goodness-of-fit, it is clearly evident that test procedures for families of discrete distributions have not been researched as extensively as those for continuous distributions (see, e.g. D’Agostino and Stephens,
Only a few procedures for testing the fit of discrete distributions have been developed. Among these, the chi-square test of fit is undoubtedly the most popular and frequently used (see Johnson et al., 1992). Specific tests of fit for individual discrete models have been proposed (see, e.g. Pettit and Stephens, 1977; Lloyd, 1984; and Gürtler and Henze, 2000); however, in terms of an approach to goodness-of-fit, general applicability is more desirable than specific procedures.

The use of the probability generating function (PGF) of $F(x; \theta)$,

$$G(t; \theta) = E \{ t^X \} \quad \text{for } |t| \leq 1,$$

and its empirical counterpart,

$$G_n(t) = \frac{1}{n} \sum_{i=1}^{n} t^{X_i} \quad \text{for } |t| \leq 1,$$

in testing the fit of discrete distributions was proposed by Kocherlakota and Kocherlakota (1986) and further studied by Márques and Pérez-Abreu (1989). A quadratic-type test statistic, similar to that of the Cramér-von Mises family of tests, was proposed and defined as

$$d = \int_{0}^{1} \xi^2(t; \theta) \, dt$$

where,

$$\xi(t; \theta) = \sqrt{n} \left[ G_n(t) - G(t; \theta) \right].$$

Rayner and Best (1989) discussed tests of fit in general and have provided details on their own smooth goodness-of-fit tests. Klar (1999) presented a widely applicable goodness-of-fit test for discrete distributions based on the difference of the integrated distribution function (IDF),
\[
\Psi(t) = \int_t^\infty [1 - F(x)] \, dx ,
\]  
(1.22)

and its empirical counterpart,

\[
\Psi_n(t) = \int_t^\infty [1 - F_n(x)] \, dx .
\]  
(1.23)

To date, however, there exists no formal approach to testing fit for discrete distributions based on a visual or graphical component. Ord (1967) provides a general approach for graphical analysis for a class of discrete models; however, these methods do not involve the calculation of a formal test statistic.

In this paper, we present a graphical approach to goodness-of-fit of discrete distributions based on a theoretical Q-Q plot and the squared correlation coefficient, \(R^2\), obtained from unweighted least squares. Similar tests of fit have been developed for continuous distributions such as the univariate normal and extreme-value (or Gumbel) distributions and are referred to as regression tests of fit (see, e.g. Filliben, 1974; Looney and Gulledge, 1985; and Kinnison, 1989). These types of goodness-of-fit tests are based on simulation studies and calculations of tabled critical values and thus, are computer-intensive.

We illustrate our general approach with the well-known univariate discrete Poisson distribution. Testing the fit of random count data to the Poisson model is and continues to be of high interest (see, e.g. Hoaglin, 1980; Rueda et al., 1991; Baringhaus and Henze, 1992; Nakamura and Pérez-Abreu, 1993; Epps, 1995; Henze and Klar, 1995; Henze, 1996; Spinelli and Stephens, 1997; Kyriakoussis et al., 1998; Lee, 1998; Rueda and O'Reilly, 1999; and Gürtler and Henze, 2000). In Chapter 2, we present some recent and classical goodness-of-fit tests for Poissonity with particular emphasis on those that are
most commonly used in practice. We then propose and investigate our graphical based approach for testing the fit of a Poisson model in Chapter 3. The proposed test is then demonstrated with a simulated sample of Poisson data. In the remaining chapters, we apply the proposed test for Poissonity to data sets historically known to be Poisson, we compare these results to that of the frequently used $\chi^2$ goodness-of-fit test, and investigate appealing features of our new approach to testing Poissonity.
CHAPTER 2

THE POISSON DISTRIBUTION

The Poisson distribution bears the name of Simeon Denis Poisson (1781–1840), an accomplished French mathematician, who published his derivation of the Poisson probability model as a limiting case of a binomial distribution (see Poisson, 1837). In a binomial distribution, the probability that precisely \( x \) successes occur, \( \Pr[X = x] \), out of \( n \) trials of an event where \( \theta \), the probability of success, remains constant from event to event, is given by

\[
b(x; n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} \quad x = 0, 1, \ldots, n \quad \text{elsewhere.} \tag{2.1}
\]

Suppose now that the number of opportunities for an event to occur is very large \( (n \to \infty) \) while the product \( n\theta = \lambda \) remains a finite constant. This implies that the probability of occurrence becomes very small; namely, \( \theta \to 0 \). From direct analysis (see Appendix A.1), it can be established that a limiting distribution of \( b(x; n, \theta) \) in (2.1) is,

\[
\lim_{n\to\infty} \frac{\lambda^x e^{-\lambda}}{x!}.
\]

It can then be shown (see Appendix A.2) that the result in (2.2), defined for \( x = 0, 1, 2, \ldots \) and parameter \( \lambda > 0 \), satisfies the conditions of a probability density
function (PDF) of a discrete type random variable. Therefore, a random variable $X$ is said to have an underlying Poisson distribution with parameter $\lambda$, that is, $X \sim \text{Pois}(\lambda)$, if and only if its probability distribution is given by the function

$$p(x; \lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, 2, \ldots; \lambda > 0 \\ 0 & \text{elsewhere} \end{cases}$$

(2.3)

Thus, the Poisson distribution is a mathematical function that assigns probabilities to the number of occurrences of random events. Furthermore, the Poisson model is a power series distribution with infinite nonnegative integer support and belongs to the exponential family of distributions (Johnson et al., 1992).

The moment generating function (MGF) of a specified distribution, defined by the expectation

$$M(t) = E[e^{tx}]$$

for $t \in \mathbb{R}$,

(2.4)

is unique and completely determines the distribution of a random variable $X$ (see, e.g. Hogg and Craig, 1995). Thus, if $X \sim \text{Pois}(\lambda)$, its MGF is of the form

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

for $t \in \mathbb{R}$.

(2.5)

The derivation of $M_X(t; \lambda)$ in (2.5) is shown in Appendix A.3.

Likewise, if $X \sim \text{Pois}(\lambda)$, the mean of $X$ and the variance of $X$ are

$$E[X] = \mu = \lambda,$$

(2.6)

and

$$\text{Var}[X] = \sigma^2 = \lambda,$$

(2.7)
respectively. Thus, \( \mu = \sigma^2 = \lambda \). In this case, the parameter \( \lambda > 0 \), the mean or average number of occurrences of a particular event, is all that is needed to specify a Poisson distribution. The derivation of \( E[X] \) and \( Var[X] \) may be found in Appendix A.4.

Figure 2.1 displays Poisson density plots with various values of the single parameter \( \lambda \). Since \( \mu = \sigma^2 = \lambda \), note how the location and spread of each respective Poisson density changes with increase in \( \lambda \), which need not be an integer. Poisson densities with smaller values of \( \lambda \) tend to be more peaked and right-skewed, due to the conditions that
$\lambda > 0$ and $x = 0, 1, 2, \ldots$, whereas Poisson densities with larger $\lambda$ values tend to be less peaked and more symmetric. Corresponding Poisson CDFs defined by

$$
\Pr[X \leq x] = F(x; \lambda) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!},
$$

(2.8)

are shown in Figure 2.2.

![Poisson CDF plots](image)

**Figure 2.2 (a) – (d).** Poisson CDF plots with $\lambda$ corresponding to those in Figure 2.1.

It can be shown that the maximum likelihood estimate (MLE) of the parameter $\lambda$ is the sample or empirical mean, $\bar{X}$. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Poisson distribution with parameter $\lambda > 0$. We define the likelihood function as
\[ L(\lambda; x_1, x_2, \ldots, x_n) = \frac{\lambda^{\sum x_i} e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}, \quad (2.9) \]

which is \( \prod_{i=1}^{n} p(x_i; \lambda) \), the joint PDF of \( X_1, X_2, \ldots, X_n \). Let us consider

\[ \ln L(\lambda; x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i \ln(\lambda) - n\lambda - \ln\left(\prod_{i=1}^{n} x_i!\right). \quad (2.10) \]

Differentiating and maximizing \( \ln L(\lambda; x_1, x_2, \ldots, x_n) \) with respect to \( \lambda \), the parameter being estimated, we obtain

\[ \frac{d}{d\lambda} \left[ \ln L(\lambda; x_1, x_2, \ldots, x_n) \right] = \sum_{i=1}^{n} \frac{x_i}{\lambda} - n = 0 \quad (2.11) \]

\[ \Rightarrow n\lambda = \sum_{i=1}^{n} x_i \]

\[ \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}, \quad (2.12) \]

which is the empirical mean.

It can also be shown with the same amount of case, that \( \bar{x} \) is an unbiased estimate of \( \lambda \). By definition, any statistic whose mathematical expectation is equal to a parameter \( \theta \) is called an unbiased estimator of the parameter \( \theta \). Thus, suppose \( X_i \sim \text{Pois}(\lambda) \), \( i = 1, 2, \ldots, n \). Then

\[ E[\bar{x}] = \frac{1}{n} E\left[ \sum_{i=1}^{n} X_i \right] \quad (2.13) \]
Therefore, $\bar{X}$ is an unbiased estimate of $\lambda$.

2.1 Applications of The Poisson Model

The mathematical conditions of an infinite number of trials and infinitesimal probability of occurrence are rarely achieved in practice. Thus, for practical purposes, we restate the Poisson distribution in (2.3) as follows: (i) if in a given experiment, the number of opportunities for an independent event to occur is large (e.g. $n \geq 30$), and; (ii) if the probability of this particular event occurring is small (e.g. $\theta \leq 0.05$), and; (iii) if the average number of occurrences of this event is a finite value, say $\lambda = n\theta$, then the probability that exactly $x$ of these events occurs is given by the Poisson model,

$$\Pr[X = x] = p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{where } x = 0, 1, 2, \ldots ; \lambda > 0.$$ 

The Poisson Process

The Poisson distribution is the counting distribution for a Poisson process. A Poisson process is a stochastic point process, which describes the situation of points (or events) in time or space according to probabilistic laws, and that satisfies certain special axioms (Haight, 1967). These axioms are presented as postulate systems by Doob (1953), Feller (1957), and Parzen (1962), which we conveniently express in elementary and summarized form.
Let $X(T)$ denote the number of events in a finite set, interval, or region $T \subseteq S$. We use $|T|$ to denote either the length or the size of $T$ depending on whether $T$ is a subset of the axis of time, or a subset of a two-dimensional (e.g. area) or three-dimensional (e.g. volume) space. Suppose that the following assumptions hold: (i) $\Pr[X(T) = x]$, the probability that exactly $x$ events occur in $T$, depends on $|T|$ and $x$ only; (ii) the number of events in non-overlapping intervals are independent, meaning, $X(T_1)$ and $X(T_2)$ are independent random variables if $T_1$ and $T_2$ are disjoint; and (iii) the probability that more than one event occurs in $T$ is small if $|T|$ is small. That is,

$$\frac{\Pr[X(T) \geq 2]}{|T|} \rightarrow 0 \quad \text{for} \quad |T| \rightarrow 0.$$  

(2.15)

It has been shown (see, e.g. Parzen, 1962) that there exists a number $\lambda > 0$ such that for all $T \subseteq S$, $X(T)$, the number of events in $T$, has a Poisson distribution with parameter $\lambda |T|$, or

$$X(T) \sim \text{Pois}(\lambda |T|).$$  

(2.16)

Thus, in this role, the Poisson distribution frequently appears in the enumeration of a wide assortment of random and often rare phenomena.

**Phenomena Fitting A Poisson Model**

The first records of the use of the Poisson distribution are attributed to Ladilaus von Bortkiewicz (1868 – 1931). Among several phenomena that Bortkiewicz successfully fit with the Poisson model, the best-known example is that of his study (see Bortkiewicz, 1898) of the number of men killed by the kick of a horse in ten cavalry corps of the
Prussian army during a period of 20 years (1875 – 1894). A summary of Bortkiewicz’s study with fitted Poisson distribution is shown in Table 2.1.

Table 2.1. Death by Horse Kick of Members of The Prussian Army (1875 – 1894)

<table>
<thead>
<tr>
<th>$x$ = number of deaths per corps-year</th>
<th>Observed number of corps-years during which $x$ fatalities occurred.</th>
<th>Theoretical number of corps-years during which $x$ deaths would be expected.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>109</td>
<td>108.67</td>
</tr>
<tr>
<td>1</td>
<td>65</td>
<td>66.29</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>20.22</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4.11</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.63</td>
</tr>
<tr>
<td>≥5</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>200.00</td>
</tr>
</tbody>
</table>

Not long thereafter, the Poisson distribution became firmly embedded in the repertoire of mathematicians, statisticians, and engineers and scientists of all types (Larsen, 1985). Motivated by Bortkiewicz’s work, the Poisson distribution has since been shown to be an excellent model in a variety of problems dealing with the counts of random events. For an extensive list of references and bibliographies of the applications mentioned in this and later sections, we refer the reader to Thorndike (1926), Feller (1957), Haight (1967), Barnes and Schuhl (1971), and Larsen (1985).

As a limiting form of a binomial distribution, the Poisson model naturally fits the frequency of occurrence of accidents. Studies in accident theory and proneness, as well as the number of deaths in automobile and train accidents along a certain stretch of road or rail track, and even coal mining disasters have all been associated with the Poisson
distribution. The Poisson model is also utilized in insurance problems, especially with collective risk theory and insurance claims.

Table 2.2. Connections To A Wrong Number

<table>
<thead>
<tr>
<th>$x = \text{number of wrong connections per period}$</th>
<th>Observed number of periods during which $x$ wrong connections occurred.</th>
<th>Theoretical number of periods during which $x$ wrong connections would be expected.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0.04</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.37</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1.63</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4.76</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>10.39</td>
</tr>
<tr>
<td>5</td>
<td>14</td>
<td>18.16</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>26.46</td>
</tr>
<tr>
<td>7</td>
<td>43</td>
<td>33.03</td>
</tr>
<tr>
<td>8</td>
<td>31</td>
<td>36.09</td>
</tr>
<tr>
<td>9</td>
<td>40</td>
<td>35.04</td>
</tr>
<tr>
<td>10</td>
<td>35</td>
<td>30.63</td>
</tr>
<tr>
<td>11</td>
<td>20</td>
<td>24.34</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>17.72</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>11.92</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td>7.44</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>4.33</td>
</tr>
<tr>
<td>$\geq 16$</td>
<td>2</td>
<td>4.65</td>
</tr>
</tbody>
</table>

| Total                                           | 267                                                                      | 267.00                                                                           |

Early engineering applications concerned with telephone traffic and switchboard problems have involved the Poisson model. For instance, the number of calls placed, the frequency of wrong numbers dialed, the number of accidental cut-offs, and the number of lines available in a given time interval has been shown to be Poisson. Table 2.2 is an example found in Thorndike (1926) of the number of connections to a wrong number.
Table 2.3. Traffic Arrivals In 30-Second Intervals (Durfee Avenue, Northbound)

<table>
<thead>
<tr>
<th>$x$ = number of vehicles arriving per 30-second interval</th>
<th>Observed frequency of intervals during which $x$ vehicles arrived.</th>
<th>Theoretical frequency of intervals during which $x$ vehicles would be expected to arrive.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>5.57</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>17.10</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>26.25</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>26.86</td>
</tr>
<tr>
<td>4</td>
<td>19</td>
<td>20.62</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>12.66</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6.48</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>2.84</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>1.09</td>
</tr>
<tr>
<td>$\geq 9$</td>
<td>1</td>
<td>0.52</td>
</tr>
<tr>
<td>120</td>
<td>120.00</td>
<td></td>
</tr>
</tbody>
</table>

Queuing theory, which deals with efficient scheduling of a sequence of tasks (or queues), also lends itself readily to the Poisson distribution. Besides telephone traffic, the theory of queues also extends to applications in road, rail, water, and air traffic. The scheduling of ships and barges, the number of cars passing a certain point, and the number of vacant spaces in a parking lot during a fixed time interval are instances where the Poisson model applies. An example of traffic arrival data analyzed in 30-second intervals found in Barnes and Schuhl (1971) and provided by the Los Angeles County Road Department is shown in Table 2.3. Consumer traffic may also be considered here. The arrival of customers in a given time period and the service termination points of customers are known Poisson processes.

The Poisson distribution also plays an important role in the field of commerce. Inventory theory, the number of transactions per day for a given stock, and even the
number of bids for a contract are instances where the Poisson law fits. The Poisson model is also used in industry for issues of reliability and quality control. Instances of machine breakdown often form a Poisson process while the number of defective items found in a batch of products or even the frequency of pages of a book containing exactly \( x \) misprints will frequently be Poisson.

<table>
<thead>
<tr>
<th>( x ) = number of hits per ( \frac{1}{3} \text{ km}^2 ) subregions</th>
<th>Observed number of subregions in which ( x ) hits occurred.</th>
<th>Theoretical number of subregions in which ( x ) hits would be expected to occur.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>229</td>
<td>227.26</td>
</tr>
<tr>
<td>1</td>
<td>211</td>
<td>211.35</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
<td>98.28</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>30.47</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>7.08</td>
</tr>
<tr>
<td>( \geq 5 )</td>
<td>1</td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td>576</td>
<td>576.00</td>
</tr>
</tbody>
</table>

Military applications exist as well. These include, but are not limited to, the number of times military land or aircraft require spare parts and the number of bullets hitting a target. Table 2.4 provides figures taken from Clark (1946) of the spatial distribution of flying bomb hits in south London during World War II. Sociological and demographical data may also be analyzed or described using the Poisson model. For example, the number of children per household, the number of isolates in a social group, and the number of daily deaths of senior citizens in a large city over a period of years has been treated with the Poisson distribution.
There are numerous applications of the Poisson model in various branches of science. In biology and medicine, the Poisson distribution has been used in: analyzing cell and virus counts of a given blood or solution sample; describing the spatial distribution of bacteria colonies in a plate of agar; studying of neural impulses; counting the number of defective teeth per individual; enumerating the number of chromosome interchanges induced by X-ray irradiation; and in generalizations of epidemic models to count number of victims of a specific disease (e.g. cancer, cholera, anthrax, malaria, etc.) per year.

Table 2.5. Number of Noxious Weed Seeds Found in Subsamples of Meadow Grass

<table>
<thead>
<tr>
<th>$x$ = number of noxious seeds per subsample</th>
<th>Observed number of subsamples with exactly $x$ noxious seeds.</th>
<th>Theoretical number of subsamples with exactly $x$ noxious seeds expected.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>4.78</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>14.44</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>21.81</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>21.95</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>16.58</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>10.01</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5.04</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>2.17</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.82</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.28</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td>$\geq 11$</td>
<td>0</td>
<td>0.03</td>
</tr>
<tr>
<td></td>
<td>98</td>
<td>98.00</td>
</tr>
</tbody>
</table>

In terms of agriculture and ecology, the distribution of plants and animals in either space or time is frequently Poisson. For instance, the spatial distribution of a certain plant over a fixed region and the number of fishes caught per day and the location at
which they were caught have been investigated using the Poisson model. Table 2.5 is an example found in Leggatt (1935) of the number of noxious weed seeds found in 1/4 oz subsamples of *Phleum praetense* (meadow grass). In archaeology, the Poisson distribution has been used to study the positions at which bones or artifacts were found. Likewise, meteorological applications involving the amount of rain or snow in a given month and the amount of excessive flooding in a specified region have made reference to the Poisson distribution. Table 2.6 displays data from Grant (1938) of frequencies of intense rainstorms of short duration per year measured at ten stations in the Mid-West geographic region of the continental United States.

### Table 2.6. 10-Minute Excessive Rainstorms In Mid-West Region

<table>
<thead>
<tr>
<th>$x$ = number of excessive rainstorms per station-year</th>
<th>Observed number of station-years with exactly $x$ rainstorms.</th>
<th>Theoretical number of station-years with exactly $x$ rainstorms expected.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>102</td>
<td>99.39</td>
</tr>
<tr>
<td>1</td>
<td>114</td>
<td>119.27</td>
</tr>
<tr>
<td>2</td>
<td>74</td>
<td>71.56</td>
</tr>
<tr>
<td>3</td>
<td>28</td>
<td>28.63</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>8.59</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2.06</td>
</tr>
<tr>
<td>$\geq 6$</td>
<td>0</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>330</td>
<td>330.00</td>
</tr>
</tbody>
</table>

Many references to particle physics and radioactivity have involved a Poisson process. A classic example is that of Rutherford and Geiger (1910) who studied alpha emission caused by radioactive decay of a quantity of polonium. The recorded number of $\alpha$-particles reaching a counter in this experiment is shown in Table 2.7.
Table 2.7. Alpha Emission From Radioactive Source

<table>
<thead>
<tr>
<th>$x$ = number of $\alpha$-particles recorded per $\frac{1}{4}$-minute intervals</th>
<th>Observed frequency of $\frac{1}{4}$-minute intervals during which $x$ $\alpha$-particles were recorded.</th>
<th>Theoretical frequency of $\frac{1}{4}$-minute intervals during which $x$ $\alpha$-particles would be expected.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>57</td>
<td>54.40</td>
</tr>
<tr>
<td>1</td>
<td>203</td>
<td>210.52</td>
</tr>
<tr>
<td>2</td>
<td>383</td>
<td>407.36</td>
</tr>
<tr>
<td>3</td>
<td>525</td>
<td>525.50</td>
</tr>
<tr>
<td>4</td>
<td>532</td>
<td>508.42</td>
</tr>
<tr>
<td>5</td>
<td>408</td>
<td>393.52</td>
</tr>
<tr>
<td>6</td>
<td>273</td>
<td>253.82</td>
</tr>
<tr>
<td>7</td>
<td>139</td>
<td>140.32</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>67.88</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>29.19</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>11.30</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>3.97</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1.28</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>0.38</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>0.11</td>
</tr>
<tr>
<td>$\geq 15$</td>
<td>0</td>
<td>0.04</td>
</tr>
</tbody>
</table>

| 2608 | 2608.00 |

Other interesting applications of the Poisson distribution are found in the fields of astronomy and seismology. In particular, the points of impact of meteorites and the number of comets observed over a period of time have been fitted to the Poisson distribution. Likewise, the frequencies of earthquakes in a given time period in seismically active regions have also been fitted to the Poisson distribution. Table 2.8 is an example of the annual number of serious earthquakes over a period of 75 years (1903 – 1977) found in Blæsild and Granfeldt (2003) and derived from The Open University (1981). In this example, an earthquake is considered serious if its magnitude is at least 7.5 on the Richter scale or if more than 100 people were killed.
Thus, the Poisson distribution and process leave much room for practical applications involving counts of random events. This intuitively supports the fact that the Poisson distribution is of highly sustained interest in the area of goodness-of-fit (see, e.g. Gürtler and Henze, 2000). In the remainder of this chapter, we discuss and survey existing tests for Poissonity with particular emphasis on those most frequently used in practice.

2.2 Existing Tests of Fit For Poissonity

Suppose $X_1, X_2, \ldots, X_n$ are independent observations on a discrete random variable $X$ defined over $\{0, 1, 2, \ldots\}$. Thus, in testing the null hypothesis

$$H_0 : X \sim Pois(\lambda)$$

against general alternatives, where $\lambda$ is either known or unknown, we can consider several existing goodness-of-fit procedures categorized as follows: graphical analysis; tests of chi-squared type; and tests based on the empirical distribution function, the integrated distribution function, and the probability generating function.
Graphical Analysis

There are several graphical tests for Poissonity in addition to the theoretical Q-Q plot discussed in Chapter 1. Hoaglin (1980) developed the *Poissonness plot*, which is similar in motivation to the probability or theoretical Q-Q plot. The Poissonness plot is based on the assumption that for some fixed value of $\lambda$, each observed frequency, $f_x$, where $\sum_x f_x = n$, equals the expected frequency,

$$m_x = n \cdot p(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \ x = 0, 1, 2, \ldots$$  \hspace{1cm} (2.18)

Taking natural logarithms on both sides of (2.18) and along with some algebraic manipulation yields

$$\ln \left( m_x \right) + \ln \left( x! \right) = x \cdot \ln \left( \lambda \right) + \left( \ln \left( n \right) - \lambda \right).$$  \hspace{1cm} (2.19)

It can easily be seen that the plot of $\ln \left( m_x \right) + \ln \left( x! \right)$ against $x$ results in a straight line with slope equal to $\ln \left( \lambda \right)$ and an intercept of $\ln \left( n \right) - \lambda$.

As with probability or theoretical Q-Q plots, the straightness of the Poissonness line is used to judge the fit of the data to a Poisson distribution. If the Poissonness line for a set of data is satisfactorily straight, either the MLE of $\lambda$, $\hat{\lambda}$ in (2.12), used in calculating $m_x$ or an estimate from the slope of the Poissonness line may be used to specify the Poisson distribution from which the data is assumed to have come. For further details, we refer the reader to Hoaglin (1980) and Hoaglin, Mosteller, and Tukey (1985).

Plots based on a ratio of successive observed frequencies have also been proposed. Dubey (1966) suggests plotting $\frac{f_x}{f_{x+1}}$ against $x$, for $x = 0, 1, 2, \ldots$, and showed that...
for a Poisson population, this plot should be a straight line with both intercept and slope equal to $\frac{1}{\lambda}$. Ord (1967) later found that
\[
  u_x = \frac{x f_x}{f_{x-1}}
\]
(2.20)
is a better diagnostic, and showed that plotting $u_x$ against $x$ should yield a straight line of the form
\[
  u_x = c_0 + c_1 x,
\]
(2.21)
for a number of discrete distributions. For the Poisson case, this relationship is
\[
  u_x = \lambda + 0 x;
\]
(2.22)
however, even though sample ratios cannot be expected to satisfy distributional relationships exactly, Ord (1967) suggests that sample plots of this sort give a fair indication of an appropriate type of distribution.

The construction of Poisson probability paper has also been suggested in the literature along with a number of rapid graphical tests for Poissonity for small sample cases. For references and details on these graphical techniques, we refer the reader to Johnson et al. (1992).

**Tests of Chi-Squared Type**

When the sample size $n$ is sufficiently large, Pearson's chi-square goodness-of-fit test based on the test statistic defined in (1.10a) may be used to test the null hypothesis in (2.17). In practice, however, the parameter $\lambda$ is often unknown and $\hat{\lambda}$ in (2.12) is used in calculating the expected cell frequencies. In this case, the Pearson-Fisher test statistic defined in (1.10b) is considered. Nevertheless, the classical $\chi^2$ goodness-of-fit test
discussed in Chapter 1, is frequently, if not preferably, used to test the null hypothesis in (2.17) (see, e.g. Johnson et al., 1992).

A well-known alternative to the classical $\chi^2$ goodness-of-fit test is the *index of dispersion test*, which is based on the result that the index (or coefficient) of dispersion, $ID$, of the Poisson distribution is equal to 1. Recall that if $X \sim \text{Pois}(\lambda)$, the mean and variance of $X$ is $E[X] = \lambda$ and $Var[X] = \lambda$, respectively. It then follows that

$$ID = \frac{Var[X]}{E[X]} = 1.$$  \hspace{1cm} (2.23)

Thus, it has to be expected that the ratio between the empirical variance,

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1},$$  \hspace{1cm} (2.24)

and the empirical mean,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$  \hspace{1cm} (2.25)

namely,

$$\widehat{ID} = \frac{s^2}{\bar{x}},$$  \hspace{1cm} (2.26)

would be close to 1 if in fact $X \sim \text{Pois}(\lambda)$. The test statistic $\widehat{ID}$ in (2.26) is attributed to Fisher et al. (1922) and is referred to as *Fisher's Index of Dispersion*.

For large values of $\lambda$, likewise, large sample size $n$, the distribution of $\widehat{ID}$ may be approximated by $\chi^2_{n-1}/(n-1)$. That is, as stated in Blæsild and Granfeldt (2003),

$$\widehat{ID} \sim \approx \chi^2_{n-1}/(n-1),$$  \hspace{1cm} (2.27)
which holds for \( n \geq 15 \) or for \( \lambda \geq 5 \). To test the null hypothesis in (2.17) with significance level \( \alpha \), the following decision rule is employed:

\[
\begin{cases}
\hat{ID} < \chi^2_{n-1;\alpha}/(n-1) \\
\text{Reject } H_0 \text{ if or} \\
\hat{ID} > \chi^2_{n-1;1-\alpha}/(n-1)
\end{cases}
\]

Thus, \( H_0 \) is rejected for large or small values of \( \hat{ID} \).

A related alternative to Fisher’s index of dispersion test is the variance test, as referred to by Cochran (1954). The variance test is based on the test statistic

\[
\chi^2_{\text{var}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}}.
\] (2.28)

It can be seen that the relationship between \( \hat{ID} \) in (2.26) and \( \chi^2_{\text{var}} \) in (2.28) is

\[
\chi^2_{\text{var}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}} = \frac{(n-1)s^2}{\bar{x}} = (n-1)\cdot \hat{ID},
\] (2.29)

and that

\[
\chi^2_{\text{var}} \sim \chi^2_{n-1}.
\] (2.30)

The proof, attributed to Fisher, is given in various steps in Cochran (1936), Kathirgamatamby (1953), and Lancaster (1957).

It is noteworthy to point out that Fisher’s index of dispersion test and the variance test do not test the same hypotheses. Fisher’s index of dispersion is derived from a model where the observations \( x_i \) are independent and identically distributed according to an unspecified distribution, thus testing the hypothesis in (2.17), that the common distribution is the Poisson distribution (see Blæsild and Granfeldt, 2003). On the other
hand, $\chi^2_{var}$ is derived from a model where the $x_i$ are independent and Poisson distributed, but not necessarily identically distributed. Thus, the hypotheses being tested with $\chi^2_{var}$ is precisely

$$H'_0: x_i \text{ are identically Poisson distributed}$$

versus $H'_1: x_i \text{ follow independent Poisson distributions with different means, } \lambda_j$ (see Snedecor and Cochran, 1991; Brown and Zhao, 2002; and Blæsild and Granfeldt, 2003). Hence, this test is sometimes called the test for homogeneity of the Poisson distribution. Here, we use the decision rule

$$\text{Reject } H'_0 \text{ if } \chi^2_{var} > \chi^2_{n-1,1-\alpha}.$$  

This test has been shown to be more powerful than the general-purpose $\chi^2$ goodness-of-fit test (see Berkson, 1940), due to its sensitivity in detecting $H'_1$.

Other tests involving functionals of $\chi^2_{var}$ such as Neyman's smooth test of Poissonity and its various forms may also be considered. These are discussed in Rayner and Best (1989), Gürtler and Henze (2000), and Brown and Zhao (2002).

**Tests Based on The Empirical Distribution Function**

As in the continuous case, goodness-of-fit tests for Poissonity based on EDF statistics measure the discrepancy between $F(x; \lambda)$, the distribution function of the Poisson model given in (2.8), and the empirical distribution function $F_n(x)$. For instance, Henze (1996) has investigated tests for Poissonity based on functionals of the measure

$$Z = \sqrt{n} \left[ F_n(x) - F(x; \lambda) \right], \quad x \geq 0,$$

(2.32)
where $F(x; \hat{\lambda})$ is the estimated distribution function under Poissonity. Examples of test statistics considered include a Kolmogorov-Smirnov type statistic

$$K = \sup_{x \geq 0} \left\{ \sqrt{n} \left| F_n(x) - F(x; \hat{\lambda}) \right| \right\}$$

(2.33)

and a Cramér-von Mises type statistic

$$C = n \sum_{x=0}^{\infty} \left[ F_n(x) - F(x; \hat{\lambda}) \right]^2 f(x; \hat{\lambda}),$$

(2.34)

where $f(x; \hat{\lambda}) = p(x; \hat{\lambda})$ defined in (2.3) with $\hat{\lambda}$ replacing $\lambda$. We refer the reader to Henze (1996), and Gürtler and Henze (2000) for a deeper investigation of these test statistics and their modifications.

**Tests Based on The Integrated Distribution Function**

Klar (1999) introduced a widely applicable goodness-of-fit test for discrete distributions based on the integrated distribution function (IDF) $\Psi(t)$ defined in (1.22). In particular, the discrepancy between the empirical integrated distribution function $\Psi_n(t)$ defined in (1.23) and the estimated IDF under Poissonity

$$\hat{\Psi}(t) = \int_{t}^{\infty} \left[ 1 - F(x; \hat{\lambda}) \right] dx$$

(2.35)

is used to obtain a Kolmogorov-Smirnov type test statistic of the form

$$I = \sup_{t \geq 0} \left\{ \sqrt{n} \left| \Psi_n(t) - \hat{\Psi}(t) \right| \right\}.$$  

(2.36)

Here, the null hypothesis is rejected for large values of $I$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Tests Based on The Probability Generating Function

Since the distribution of $X$ is determined by its probability generating function $G(t; \theta)$ defined in (1.18), it has been proposed (see, e.g. Kocherlakota and Kocherlakota, 1986; Márques and Pérez-Abreu, 1989; and Rueda et al., 1991) that a goodness-of-fit test be based on the difference of the probability generating function $G(t; \theta)$ and its empirical counterpart $G_n(t)$ defined in (1.19). Tests of this sort are based on a functional of the difference measure $\xi(t; \theta)$ defined in (1.21). In the Poisson case,

$$\xi(t; \hat{\theta}) = \sqrt{n} \left[ G_n(t) - G(t; \hat{\theta}) \right]$$  \hspace{1cm} (2.37)

where

$$G(t; \hat{\lambda}) = e^{\hat{\lambda}t(1-1)} \text{ for } |t| \leq 1.$$  \hspace{1cm} (2.38)

Rueda et al. (1991) has suggested a quadratic-type test statistic, similar to that of the Cramér-von Mises family of tests, defined by

$$d = \int_0^1 \xi^2(t; \hat{\lambda}) \, dt.$$  \hspace{1cm} (2.39)

Alternatives to $d$, such as the statistics $T$ and $V$, have also been proposed to test the null hypothesis in (2.17). For definitions and details on these, we refer the reader to Baringhaus and Henze (1992), Nakamura and Pérez-Abreu (1993), and Gürtler and Henze (2000).

Having surveyed much of the literature on existing goodness-of-fit tests for the Poisson distribution, we find it noteworthy that to date, no formal graphical based goodness-of-fit test for Poissonity exists. Motivated by this fact, and the wide ranging applications of the Poisson model to real life phenomena, as well as the inherent
weaknesses of the frequently used classical $\chi^2$ test of fit, we are now ready to propose a very practical goodness-of-fit test based on both a graphical component and the calculation of a formal test statistic.
CHAPTER 3

PROPOSED TEST FOR POISSONITY

Our proposed graphical approach to testing the composite null hypothesis of Poissonity preserves much of the same motivations of earlier studies by Filliben (1975), who proposed the probability plot correlation coefficient test for normality, and by Kinnison (1989), who proposed the correlation goodness-of-fit test for the extreme-value (Gumbel) distribution. We base our proposed test for Poissonity on: the construction of a Poisson Q-Q plot; the use of a squared correlation coefficient $R^2$ as a formal test statistic; and the comparison of this $R^2$ test statistic against a specified critical value obtained from a simulated distribution of $R^2$ under Poissonity. Throughout this chapter, we use a simulated data set of size $n = 25$, generated from a Poisson distribution with $\lambda = 2$ (see Table 3.1) to illustrate each component of our proposed test procedure.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>freq($x_i$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

$\bar{x} = 2.16$

$n = 25$
3.1 Construction of Poisson Q-Q Plot

The starting point for our method of testing Poissonity is rooted in the information given in a Poisson Q-Q plot. A Poisson Q-Q plot, as here used, is defined as a plot of the \(i\)th ordered statistic \(X_{(i)}\), versus the expected quantiles under Poissonity. Here, we shall denote these quantiles by \(Q_i\) and define them as

\[
Q_i = F^{-1}\left(p_i; \hat{\lambda}\right),
\]

(3.1)

where \(F^{-1}(\cdot)\) is the inverse transform of the estimated Poisson distribution \(F(x; \hat{\lambda})\). We use \(p_i\) to denote a vector of plotting positions defined by

\[
p_i = i - 0.5 \frac{n}{n},
\]

(3.2)

for \(i = 1, 2, \ldots, n\), and \(\hat{\lambda}\) to denote the MLE of the parameter \(\lambda\), that is defined to be the empirical mean, \(\bar{X}\).

As noted in Section 1.1 of Chapter 1, if the sample is in fact generated from the hypothesized (\(\lambda\) unspecified) Poisson distribution, then the plot of \(X_{(i)}\) versus \(Q_i\) will be approximately linear. Hence, the straightness of the Q-Q plot can be used to informally and visually evaluate the fit of the sample to the hypothesized Poisson distribution. Large and systematic departures from a straight-line configuration suggest lack of fit. At this exploratory stage of the proposed test procedure, the Poisson Q-Q plot may also be used to identify any peculiarities such as outliers and possible contamination that exist within the data. We refer the reader to Section 1.1 of Chapter 1 for further details as the Q-Q plot leaves much room for interpretation.
The Poisson Q-Q plot of the simulated data in Table 3.1 is shown above in Figure 3.1. As expected, the Q-Q plot exhibits a definite linear association between the order statistics $X_{(i)}$ and the expected quantiles $Q_i$ under Poissonity. Furthermore, no peculiarities appear to exist in the data according to the Q-Q plot. Thus, we proceed to the calculation of a formal test statistic, namely $R^2$, to supplement and confirm the information provided in the Poisson Q-Q plot.

3.2 The $R^2$ Test Statistic

A conceptually simple and obvious choice for evaluating the strength or degree of linear association between two variables is the coefficient of determination $R^2$, otherwise known as the square of the correlation coefficient $R$. Thus, we define the proposed test
statistic $R^2$ as the square of the correlation coefficient between $X_{(i)}$, the ordered observations, and $Q$, the expected quantiles from a Poisson distribution with estimated parameter $\hat{\lambda} = X$. We note here, that in terms of least-squares regression and depending on the context of the problem at hand, we may interpret $R^2$ as the proportionate reduction in total variation with the use of a certain predictor variable; however, for the purposes of our proposed test procedure, we take $R^2$ to simply be a measure of linearity of the Poisson Q-Q plot, where this measure may take on possible values ranging from 0 to 1. That is,

$$0 \leq R^2 \leq 1.$$  \hspace{1cm} (3.3)

In practice, $R^2$ is not likely to be 0 or 1, but somewhere between these limits. The closer $R^2$ is to 1, the stronger we say is the degree of linear association between $X_{(i)}$ and $Q$. We also note that although $R^2$ and $R$ may both be used as measures of the degree of linearity between two variables, there is a tendency in most applied work to use the correlation coefficient $R$ rather than $R^2$ (see, e.g. Neter et al., 1996). This is perhaps due to the property that for any $R^2$ other than 0 or 1,

$$R^2 < |R|.$$  \hspace{1cm} (3.4)

Hence, when presented, the correlation coefficient $R$ may give an impression of greater linear association between two variables than does $R^2$.

$R^2$ can easily be calculated by squaring the result of the computational formula for the correlation coefficient $R$ given by

$$R = \frac{\sum(X_{(i)} - \bar{X})(Q_i - \bar{Q})}{\sqrt{\sum(X_{(i)} - \bar{X})^2 \sum(Q_i - \bar{Q})^2}}.$$  \hspace{1cm} (3.5)
Likewise, most statistical calculators and software packages automatically compute and provide the $R^2$ statistic, either in decimal or percent form, within the least squares regression output.

![S-PLUS regression output](image)

**Figure 3.2.** S-PLUS regression output for data simulated from Poisson ($\lambda = 2$)

For example, the calculated $R^2$ value between the simulated data in Table 3.1 and the respective theoretical Poisson quantiles $Q_i$ can be found in the S-PLUS output from least squares regression of $Q_i$ on $X_i$ shown above in Figure 3.2. For our simulated Poisson data set, the calculated $R^2$ statistic is 0.9666, and as suspected from the Poisson Q-Q plot in Figure 3.1, this $R^2$ value appears to confirm the relatively strong linear association between $X_i$ and $Q_i$, thus, suggesting that $X$ could possibly be Poisson distributed; however, before making a conclusive decision, we first investigate the sampling distribution of the test statistic $R^2$ under Poissonity.
3.3 The Sampling Distribution of $R^2$

We investigate the distribution of the test statistic $R^2$ under Poissonity with empirical sampling through a *parametric bootstrap* simulation (see, e.g., Efron and Tibshirani, 1993). The concept is rather simple: (1) simulate a Poisson data set of size $n$ using the parameter $\hat{\lambda} = \bar{X}$ estimated from the original sample; (2) generate $Q_i$, the expected quantiles under Poissonity, using $p_i$ in (3.2) and $\hat{\lambda} = \bar{X}$; (3) compute the goodness-of-fit statistic $R^2_{n,\hat{\lambda}}$ for this data; (4) repeat steps (1) – (3) many times, say $N = 10,000$ iterations; and (5) tabulate the resulting empirical distribution of the statistic $R^2_{n,\hat{\lambda}}$.

![Program used in S-PLUS to simulate the distribution of $R^2$ under Poissonity.](image)

Figure 3.3. Program used in S-PLUS to simulate the distribution of $R^2$ under Poissonity.
The simulation program shown above in Figure 3.3 accomplishes such a task. The script is written in the S language (Version 4) for S-PLUS 6.0+, which is an interpretive rather than compiled language (see, e.g. Insightful Corp., 2001). The code may easily be modified and rewritten for use in other consoles such as the Statistical Analysis System (SAS). Furthermore, the application may run more efficiently time-wise if coded in a compiled language such as Fortran and C. Nevertheless, the simulation program provided, when executed in S-PLUS, suffices.

In regards to the simulated Poisson data set in Table 3.1, we specify $n = 25$ and $\lambda = 2.16$ in the code of Figure 3.3 and execute the program in S-PLUS. The $R^2_{25,2.16}$ values for each of 10,000 iterations get stored and sorted in a vector “Rsq”. The output is summarized in Figure 3.4. Selected portions of the empirical distribution of $R^2_{25,2.16}$, read left to right by rows where $[j]$ is the $j$th ordered value, are shown in Table 3.2.
Table 3.2. Selected Portions of The Empirical Distribution of $R^2_{25,2.16}$.

<table>
<thead>
<tr>
<th>Ordered $R^2_{25,2.16}$ Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] 0.6689 0.6812 0.6853 0.7005 0.7013 0.7111 0.7146 0.7163 0.7191 0.7224</td>
</tr>
<tr>
<td>[11] 0.7243 0.7266 0.7273 0.7303 0.7305 0.7328 0.7331 0.7351 0.7385 0.7386</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[81] 0.7704 0.7710 0.7711 0.7714 0.7714 0.7715 0.7715 0.7717 0.7722 0.7725</td>
</tr>
<tr>
<td>[91] 0.7729 0.7730 0.7734 0.7737 0.7741 0.7749 0.7752 0.7753 0.7757 0.7759</td>
</tr>
<tr>
<td>[101] 0.7759 0.7761 0.7762 0.7770 0.7772 0.7772 0.7780 0.7780 0.7793 0.7795</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[481] 0.8221 0.8222 0.8222 0.8222 0.8223 0.8225 0.8225 0.8226 0.8227 0.8229</td>
</tr>
<tr>
<td>[491] 0.8229 0.8230 0.8230 0.8230 0.8231 0.8231 0.8233 0.8233 0.8233 0.8234</td>
</tr>
<tr>
<td>[501] 0.8235 0.8235 0.8235 0.8235 0.8237 0.8238 0.8239 0.8239 0.8239 0.8244</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[801] 0.8399 0.8399 0.8399 0.8399 0.8399 0.8400 0.8400 0.8400 0.8400 0.8400</td>
</tr>
<tr>
<td>[811] 0.8400 0.8400 0.8400 0.8400 0.8400 0.8400 0.8402 0.8402 0.8402 0.8402</td>
</tr>
<tr>
<td>[821] 0.8402 0.8403 0.8403 0.8403 0.8403 0.8403 0.8405 0.8405 0.8405 0.8407</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[981] 0.8447 0.8447 0.8447 0.8447 0.8447 0.8447 0.8447 0.8448 0.8448 0.8448</td>
</tr>
<tr>
<td>[991] 0.8448 0.8448 0.8448 0.8448 0.8448 0.8450 0.8450 0.8450 0.8452 0.8452</td>
</tr>
<tr>
<td>[1001] 0.8452 0.8453 0.8454 0.8454 0.8454 0.8454 0.8455 0.8456 0.8456 0.8457</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[2001] 0.8676 0.8676 0.8678 0.8679 0.8679 0.8681 0.8681 0.8681 0.8681 0.8682</td>
</tr>
<tr>
<td>[201] 0.8682 0.8682 0.8682 0.8682 0.8682 0.8682 0.8683 0.8683 0.8683 0.8683</td>
</tr>
<tr>
<td>[2021] 0.8683 0.8683 0.8683 0.8683 0.8683 0.8683 0.8683 0.8683 0.8683 0.8683</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[3001] 0.8852 0.8852 0.8852 0.8852 0.8852 0.8852 0.8852 0.8852 0.8852 0.8853</td>
</tr>
<tr>
<td>[301] 0.8853 0.8853 0.8853 0.8853 0.8853 0.8853 0.8853 0.8854 0.8854 0.8854</td>
</tr>
<tr>
<td>[3021] 0.8854 0.8854 0.8854 0.8854 0.8854 0.8854 0.8854 0.8854 0.8854 0.8854</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[4501] 0.8999 0.8999 0.8999 0.8999 0.8999 0.8999 0.8999 0.8999 0.9000 0.9000</td>
</tr>
<tr>
<td>[451] 0.8999 0.8999 0.8999 0.8999 0.8999 0.8999 0.8999 0.9000 0.9000 0.9000</td>
</tr>
<tr>
<td>[4521] 0.9001 0.9001 0.9001 0.9001 0.9002 0.9002 0.9002 0.9002 0.9002 0.9002</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[7501] 0.9317 0.9317 0.9317 0.9317 0.9317 0.9317 0.9318 0.9318 0.9318 0.9318</td>
</tr>
<tr>
<td>[751] 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320</td>
</tr>
<tr>
<td>[7521] 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320 0.9320 0.9321 0.9321</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[9651] 0.9660 0.9661 0.9662 0.9662 0.9662 0.9662 0.9662 0.9662 0.9662 0.9662</td>
</tr>
<tr>
<td>[9661] 0.9663 0.9663 0.9663 0.9663 0.9663 0.9663 0.9663 0.9663 0.9666 0.9666</td>
</tr>
<tr>
<td>[9671] 0.9666 0.9666 0.9666 0.9666 0.9666 0.9667 0.9667 0.9667 0.9668 0.9668</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[9901] 0.9822 0.9823 0.9823 0.9823 0.9824 0.9824 0.9824 0.9825 0.9825 0.9825</td>
</tr>
<tr>
<td>[9911] 0.9825 0.9827 0.9827 0.9828 0.9828 0.9828 0.9828 0.9829 0.9829 0.9829</td>
</tr>
<tr>
<td>[9921] 0.9829 0.9835 0.9836 0.9836 0.9839 0.9839 0.9840 0.9840 0.9843 0.9843</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>[9981] 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000</td>
</tr>
<tr>
<td>[9991] 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000 1.0000</td>
</tr>
</tbody>
</table>
We now recall that the rationale behind the $R^2$ statistic is that underlying Poissonity will tend to yield near-linear Poisson Q-Q plots, which in turn will be reflected by high $R^2$ values close to 1. $R^2$ values that are too small would indicate a lack of fit. Thus, we look within the left tail of the distribution of $R^2 \AST$ for critical values (i.e. a left-tail test) on which to base a decision rule on.

We shall denote these critical values as $R^2 \AST\alpha$, where $\alpha$ is a specified significance level, which for our proposed test, is defined as

$$\alpha = \Pr\left(R^2 < R^2 \AST\alpha \mid H_0 \text{ True}\right). \quad (3.6)$$

For some fixed value of $\alpha$, the critical value $R^2 \AST\alpha$ would correspond to the $(\alpha \cdot 10000)$th smallest ordered $R^2$ value, denoted by $R^2 \left[\alpha \cdot 10000\right]$, from the empirical distribution of $R^2 \AST$. For instance, if we fix $\alpha$ at .05,

$$R^2 \AST.05 = R^2 \left[500\right]. \quad (3.7)$$

See Figure 3.5 for a graphical interpretation of the specified critical value $R^2 \AST.05$. Likewise, for $\alpha = .01$ and $\alpha = .10$, we have

$$R^2 \AST.01 = R^2 \left[100\right] \quad (3.8)$$

and

$$R^2 \AST.10 = R^2 \left[1000\right], \quad (3.9)$$

respectively. Thus, the decision rule for our proposed test for Poissonity is

Reject $H_0$ in favor of $H_1 \text{ if } R^2 < R^2 \AST\alpha$.

Do not reject $H_0$ if $R^2 \geq R^2 \AST\alpha$. \quad (3.10)
Referring back to our simulated data set in Table 3.1 with \( n = 25 \) and \( \hat{\lambda} = 2.16 \), we see that the critical values \( R^2_{25,2.16,\alpha} \) for \( \alpha = .01, .05, \) and .10 are highlighted in Table 3.2 and found in the simulation output shown in Figure 3.6. If we wish to test \( H_0 \) at \( \alpha = .05 \)

```
data.frame(crv, row.names = c("(a=.01)", "(a=.05)", "(a=.10)"))

  (a=.01) 0.7759
  (a=.05) 0.8234
  (a=.10) 0.8452
```

Figure 3.6. Output of critical values from simulated sampling distribution of \( R^2_{25,2.16} \).
we consider the critical value $R^2_{25,2.16,.05} = 0.8234$. Thus, since $R^2 = 0.9666 > R^2_{25,2.16,.05}$, we do not reject $H_0$ and say that there is reasonable fit at the .05 significance level of the sample $X$ to a Poisson distribution.

We can also consider the $P$-value of the test, which is the observed “tail” probability of a statistic, say $R^2$, being at least as extreme as the particular observed value when $H_0$ is true. In our case, evidence in support of $H_0$ would be

$$P\text{-value} = \Pr\left(R^2 \leq 0.9666 | H_0 \text{ true}\right) = 0.9674.$$ 

Here, the $\Pr\left(R^2 \leq 0.9666 | H_0 \text{ true}\right) = 0.9674$ since 0.9666 corresponds to $R^2[9674]$, which in terms of the simulated sampling distribution of $R^2_{25,2.16}$, is the 0.9674 percentage point. Thus, since $P\text{-value} = 0.9674 > \alpha = .05$, we do not reject $H_0$. It can also be seen that $H_0$ is not rejected for $\alpha = .01$ and .10 as well. Hence, we conclude as before that the sample $X$ is distributed as Poisson.

We note here that, unlike the correlation coefficient test for composite normality proposed by Filliben (1975), we were unable to generate a concise table of critical values for the test statistic $R^2$ under Poissonity solely based on sample size $n$; rather, we base the sampling distribution of $R^2$ under Poissonity on both $n$ and $\hat{\lambda}$, and hence the notation, $R^2_{n,\hat{\lambda}}$. Filliben (1975) uses the standard normal transformation $Z$ on the observed data, where the distribution is known to be $N(0,1)$. Hence, the distribution of the normality test statistic $R$ depends only on the sample size $n$ and the quantiles of the standardized normal distribution, $N(0,1)$. It is also known that $R$ is location and scale.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
invariant and statistically independent of $\bar{X}$ and $S$ (see Filliben, 1975). Thus, for the normal case, $R$ does not depend on the unknown normal location and scale parameters $\mu$ and $\sigma$, respectively, and the simulation need only be performed for each $n$.

For the Poisson distribution, $\lambda$ is not a scale parameter and we observed that running the simulation program for a given sample size $n$ and varying $\lambda$ resulted in changes in the percentage points of the simulated sampling distribution of $R^2$. In particular, for a specified sample size $n$, increases in $\lambda$ tend to squeeze the location of the distribution of $R^2_{n,\lambda}$ closer and closer to 1. Thus, for our proposed test procedure, $R^2$ depends on both sample size $n$ and the quantiles of $\text{Poisson}(\lambda)$ meaning, that the simulation program in Figure 3.3 must be executed per $n$, per $\lambda$ in order to generate the critical values $R^2_{n,\lambda,\alpha}$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>0.5</th>
<th>1.0</th>
<th>1.5</th>
<th>2.0</th>
<th>3.0</th>
<th>4.0</th>
<th>5.0</th>
<th>6.0</th>
<th>7.0</th>
<th>8.0</th>
<th>9.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.5122</td>
<td>.5643</td>
<td>.6750</td>
<td>.7111</td>
<td>.7588</td>
<td>.7760</td>
<td>.7900</td>
<td>.7991</td>
<td>.8086</td>
<td>.8115</td>
<td>.8188</td>
<td>.8192</td>
</tr>
<tr>
<td>15</td>
<td>.5946</td>
<td>.6612</td>
<td>.7352</td>
<td>.7607</td>
<td>.8067</td>
<td>.8239</td>
<td>.8378</td>
<td>.8443</td>
<td>.8516</td>
<td>.8548</td>
<td>.8585</td>
<td>.8616</td>
</tr>
<tr>
<td>20</td>
<td>.6522</td>
<td>.7045</td>
<td>.7664</td>
<td>.7939</td>
<td>.8345</td>
<td>.8493</td>
<td>.8619</td>
<td>.8680</td>
<td>.8763</td>
<td>.8794</td>
<td>.8802</td>
<td>.8867</td>
</tr>
<tr>
<td>25</td>
<td>.6667</td>
<td>.7345</td>
<td>.7951</td>
<td>.8188</td>
<td>.8488</td>
<td>.8665</td>
<td>.8787</td>
<td>.8854</td>
<td>.8909</td>
<td>.8942</td>
<td>.8977</td>
<td>.8998</td>
</tr>
<tr>
<td>30</td>
<td>.6947</td>
<td>.7661</td>
<td>.8086</td>
<td>.8340</td>
<td>.8655</td>
<td>.8807</td>
<td>.8897</td>
<td>.8985</td>
<td>.9028</td>
<td>.9072</td>
<td>.9087</td>
<td>.9109</td>
</tr>
<tr>
<td>35</td>
<td>.7098</td>
<td>.7838</td>
<td>.8212</td>
<td>.8492</td>
<td>.8724</td>
<td>.8881</td>
<td>.8995</td>
<td>.9068</td>
<td>.9094</td>
<td>.9137</td>
<td>.9177</td>
<td>.9200</td>
</tr>
<tr>
<td>40</td>
<td>.7302</td>
<td>.8002</td>
<td>.8348</td>
<td>.8559</td>
<td>.8822</td>
<td>.8981</td>
<td>.9060</td>
<td>.9136</td>
<td>.9166</td>
<td>.9220</td>
<td>.9235</td>
<td>.9257</td>
</tr>
<tr>
<td>45</td>
<td>.7447</td>
<td>.8141</td>
<td>.8446</td>
<td>.8645</td>
<td>.8904</td>
<td>.9036</td>
<td>.9130</td>
<td>.9191</td>
<td>.9215</td>
<td>.9254</td>
<td>.9287</td>
<td>.9310</td>
</tr>
<tr>
<td>50</td>
<td>.7568</td>
<td>.8196</td>
<td>.8539</td>
<td>.8703</td>
<td>.8932</td>
<td>.9081</td>
<td>.9162</td>
<td>.9232</td>
<td>.9267</td>
<td>.9307</td>
<td>.9341</td>
<td>.9353</td>
</tr>
<tr>
<td>55</td>
<td>.7682</td>
<td>.8311</td>
<td>.8599</td>
<td>.8765</td>
<td>.8992</td>
<td>.9128</td>
<td>.9194</td>
<td>.9266</td>
<td>.9309</td>
<td>.9349</td>
<td>.9367</td>
<td>.9392</td>
</tr>
<tr>
<td>60</td>
<td>.7776</td>
<td>.8366</td>
<td>.8660</td>
<td>.8829</td>
<td>.9036</td>
<td>.9159</td>
<td>.9241</td>
<td>.9309</td>
<td>.9340</td>
<td>.9378</td>
<td>.9399</td>
<td>.9420</td>
</tr>
<tr>
<td>65</td>
<td>.7859</td>
<td>.8450</td>
<td>.8710</td>
<td>.8874</td>
<td>.9081</td>
<td>.9190</td>
<td>.9273</td>
<td>.9338</td>
<td>.9375</td>
<td>.9404</td>
<td>.9430</td>
<td>.9448</td>
</tr>
<tr>
<td>70</td>
<td>.7972</td>
<td>.8487</td>
<td>.8757</td>
<td>.8915</td>
<td>.9108</td>
<td>.9213</td>
<td>.9303</td>
<td>.9355</td>
<td>.9395</td>
<td>.9428</td>
<td>.9455</td>
<td>.9473</td>
</tr>
<tr>
<td>75</td>
<td>.8051</td>
<td>.8526</td>
<td>.8788</td>
<td>.8954</td>
<td>.9139</td>
<td>.9236</td>
<td>.9321</td>
<td>.9379</td>
<td>.9417</td>
<td>.9450</td>
<td>.9473</td>
<td>.9494</td>
</tr>
<tr>
<td>80</td>
<td>.8063</td>
<td>.8567</td>
<td>.8811</td>
<td>.8989</td>
<td>.9171</td>
<td>.9262</td>
<td>.9342</td>
<td>.9394</td>
<td>.9440</td>
<td>.9470</td>
<td>.9490</td>
<td>.9510</td>
</tr>
<tr>
<td>85</td>
<td>.8091</td>
<td>.8618</td>
<td>.8857</td>
<td>.9017</td>
<td>.9184</td>
<td>.9292</td>
<td>.9364</td>
<td>.9418</td>
<td>.9456</td>
<td>.9488</td>
<td>.9509</td>
<td>.9528</td>
</tr>
<tr>
<td>90</td>
<td>.8159</td>
<td>.8643</td>
<td>.8886</td>
<td>.9031</td>
<td>.9211</td>
<td>.9319</td>
<td>.9380</td>
<td>.9436</td>
<td>.9478</td>
<td>.9499</td>
<td>.9521</td>
<td>.9550</td>
</tr>
<tr>
<td>95</td>
<td>.8175</td>
<td>.8689</td>
<td>.8910</td>
<td>.9062</td>
<td>.9225</td>
<td>.9337</td>
<td>.9403</td>
<td>.9443</td>
<td>.9488</td>
<td>.9518</td>
<td>.9539</td>
<td>.9562</td>
</tr>
<tr>
<td>100</td>
<td>.8245</td>
<td>.8725</td>
<td>.8938</td>
<td>.9085</td>
<td>.9253</td>
<td>.9356</td>
<td>.9415</td>
<td>.9463</td>
<td>.9502</td>
<td>.9528</td>
<td>.9553</td>
<td>.9569</td>
</tr>
</tbody>
</table>

Table 3.3. Lower-Tail Critical Values $R^2_{n,\lambda,\alpha}$.
Publishing critical value tables for our proposed test for Poissonity would be extremely tedious and extensive. Nevertheless, we provide a table (see Table 3.3) for $R^2_{n,\hat{\lambda};.05}$ for sample sizes $n = 5(5)100$ and estimated parameter $\hat{\lambda} = 0.5(.5)2(1)10$.

### 3.4 Power of $R^2$ Test for Poissonity

The power of a test is defined to be the probability of correctly rejecting $H_0$. That is, $\Pr(\text{reject } H_0 \mid H_1 \text{ true})$. As a general approach, the power of the proposed test for Poissonity can be investigated by replacing the Poisson random number generator, `rpois(n,\lambda)`, in the simulation program of Figure 3.3 with generators for other statistical distributions. We can then run the simulation program to generate say 10,000 samples for each distribution and each sample size $n$ and calculate the goodness-of-fit statistic $R^2$ to the Poisson distribution for each sample. We would expect these $R^2$ values to be typically lower than those resulting from using the Poisson random number generator. We would then sort these resulting squared correlations, and determine the proportion of the resulting $R^2$ values that were less than some critical value $R^2_{n,\hat{\lambda};\alpha}$. We did not, however, investigate the power of the proposed $R^2$ test for Poissonity in this paper. Nevertheless, it is a subject of much interest for future and related works.

### 3.5 Summary of Proposed Test For Poissonity

In using the proposed squared correlation coefficient test for Poissonity to test the composite hypothesis
\( H_0 : \) Sample is Poisson distributed  
\( H_1 : \) Sample is not Poisson distributed

we suggest the following sequence:

1. *Construct a Poisson Q-Q plot.* Plot the \( i \)th ordered observation \( X_{(i)} \), \( i = 1, 2, \ldots, n \), against \( Q_i \), the expected quantiles under Poissonity defined in (3.1). Visually evaluate the linearity of the Poisson Q-Q plot. Straight-line configurations suggest a good fit of the sample to a Poisson distribution, while large and systematic departures from linearity are to be judged as lack of fit.

2. *Calculate the \( R^2 \) test statistic.* \( R^2 \) can be computed directly by squaring the correlation coefficient between \( X_{(i)} \) and \( Q_i \) given by the formula in (3.5) or found in the output window of most statistical software packages having regressed \( Q_i \) on \( X_{(i)} \).

3. *Simulate the sampling distribution of \( R^{2}_{n, \hat{\lambda}} \).* The code provided in Figure 3.3 is readily employable in S-PLUS (Version 6.0 or higher). Likewise, the simulation program can easily be rewritten and modified for use in programming consoles that the user may be more familiar with.

4. *Decision Rule.* Since the test is lower-tail and \( R^2 \) tends to near-unity when \( H_0 \) holds, the decision rule is as follows when the risk of a Type I error is to be controlled at \( \alpha \):

   \[
   \begin{align*}
   \text{If } R^2 < R^2_{n, \hat{\lambda}, \alpha}, & \text{ reject } H_0 \text{ in favor of } H_1, \\
   \text{If } R^2 \geq R^2_{n, \hat{\lambda}, \alpha}, & \text{ conclude } H_0,
   \end{align*}
   \]

   where \( R^2_{n, \hat{\lambda}, \alpha} \) is the \( \alpha \)-percentile, or equivalently, the \((\alpha \cdot 10000)th\) smallest ordered value of the simulated \( R^2_{n, \hat{\lambda}} \) distribution.
EXAMPLES

We now apply the proposed squared correlation coefficient test for Poissonity to some historical and well-known Poisson data sets. The data considered here either represents a distribution of random events or points along an interval of the time axis, or a distribution of random events or points in an area region or some volume space. In each of the foregoing examples, we evaluate the fit of the observed sample to a Poisson distribution, having a parameter $\hat{\lambda} = \bar{X}$ computed from the observed data, by means of both the proposed test for Poissonity and the commonly used $\chi^2$ goodness-of-fit test. Thus, to test the composite hypothesis

$$H_0: \text{Observed sample is Poisson distributed} \quad \text{vs.} \quad H_1: \text{Sample is not Poisson distributed}$$

at some fixed significance level $\alpha$, we recall the tests of fit to be used as follows:

- $\chi^2$ Goodness-of-Fit Test:

  Test Statistic: $\chi^2 = \sum_{j=1}^{C} \frac{(N_j - n\hat{p}_j)^2}{n\hat{p}_j}$, where

  $N_j = \text{observed cell frequencies}$
  $n\hat{p}_j = \text{expected cell frequencies under Poissonity with } \hat{\lambda} = \bar{X}.$

  Decision Rule: If $\chi^2 > \chi^2_{C - r - 1, \alpha}$, reject $H_0$ in favor of $H_1$.
Otherwise, conclude $H_0$, where

\[ C = \# \text{ of cells (or classes) based on } n\hat{p}_j \geq 5. \]
\[ t = \# \text{ of estimated parameters}. \]

**$R^2$ Test for Poissonity:**

- **Test Statistic:**  
  (i) *Informal:* linearity of Poisson Q-Q plot of ordered observations $X_{(i)}$, $i = 1, 2, \ldots, n$, and $Q_i$, the expected quantiles under Poissonity with $\hat{\lambda} = \bar{X}$.

  (ii) *Formal:* $R^2$, the squared correlation coefficient between $X_{(i)}$ and $Q_i$.

- **Decision Rule:**  
  If $R^2 < R^2_{n,\lambda,\alpha}$, reject $H_0$ in favor of $H_1$;

  Otherwise, conclude $H_0$, where

\[
R^2_{n,\lambda,\alpha} = \alpha \text{th percentage point of the simulated sampling distribution of } R^2_{n,\lambda}.
\]

We provide the script shown below in Figure 4.1, that when executed in S-PLUS, conveniently and automatically constructs the Poisson Q-Q plot with fitted regression line (of $Q_i$ on $X_{(i)}$) and computes the required test statistic $R^2$. For observed sample data presented in frequency form, the user need only input the sample size $n$, and the number of successes $x$, and the observed frequency of successes $f_i$ in the `rep(x,f)` function in the provided code.
4.1 Death By Horse Kick

The first records of the use of the Poisson distribution to model the underlying distribution of an observed sample is that of Bortkiewicz's study of the frequency of death among members of the Prussian Army due to the kick of a horse. With the aid of Prussian officials during the late 19th century, Bortkiewicz gathered information on the hazards that horses posed to cavalry soldiers (Larsen, 1985). A total of 10 cavalry corps was monitored over a period of 20 years (1875 – 1894). Recorded for each year and each corps was $X$, the number of fatalities due to horse kick. Table 4.1(a) displays the empirical distribution of $X$ for the $n = 200$ corps-years, along with the fitted Poisson
distribution using \( \hat{\lambda} = 0.61 \). Binned data for use with the \( \chi^2 \) goodness-of-fit test is shown in Table 4.1(b). We note here, that in Table 4.1 and all subsequent tables to follow, we define the following notations:

\[
\begin{align*}
    f_x &= \text{frequency of } x \\
    \hat{\lambda} &= \frac{\sum x \cdot f_x}{n} = \bar{x} \\
    n\hat{\lambda}_x &= \text{expected frequency of } x \text{ under Poissonity} \\
    j &= j\text{th cell or class} \\
    N_j &= \text{observed frequency of cell } j \\
    np_j &= \text{expected frequency of cell } j \text{ under Poissonity}
\end{align*}
\]

Table 4.1. Empirical and Binned Death By Horse Kick Data

<table>
<thead>
<tr>
<th>(a) Empirical Data</th>
<th>(b) Binned Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>( f_x )</td>
</tr>
<tr>
<td>0</td>
<td>109</td>
</tr>
<tr>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>( \geq 5 )</td>
<td>0</td>
</tr>
</tbody>
</table>

\( n = 200 \) \\
\( \bar{x} = 0.61 \)

We first use the \( \chi^2 \) goodness-of-fit test to illustrate the fit of the horse kick data in Table 4.1 to a Poisson distribution. We calculate \( \hat{\lambda} \), the mean number of fatalities per corps-year due to horse kick, to be \( \bar{x} = 0.61 \). By inspection of Table 4.1(a), we see that there is excellent agreement between observed and expected frequencies of deaths per corps-year. In proceeding with the \( \chi^2 \) goodness-of-fit test, we combine the last three
classes in Table 4.1(a) to give a minimum approximate expectation \( n\hat{p}_j \) of 5 for each class \( j \). The binned data is found in Table 4.1(b). This makes the number of classes \( C = 4 \). Since one parameter, namely \( \lambda \), was estimated in fitting the distribution, \( \chi^2 \) has \( 4 - 1 - 1 = 2 \) df. Controlling \( \alpha \) at .05, we calculate the test statistic to be \( \chi^2 = 0.323 \) and compare this with the critical value \( \chi^2_{2.95} = 5.991 \). Thus, since \( \chi^2 = 0.323 < \chi^2_{2.95} \), we conclude \( H_0 \), that there is reasonable fit of the sample to a Poisson distribution. Likewise, the \( P \)-value is \( \text{Pr}(\chi^2 \geq 0.323) = .8509 \), so the discrepancies between the observed and expected frequencies under Poissonity are not unusually large.

![Q-Q Plot of Death By Horse Kick Data](image)

Figure 4.2. Q-Q plot of death by horse kick data with \( R^2 \) test statistic.

In applying the proposed \( R^2 \) test for Poissonity to the death by horse kick data, we consider the empirical data in Table 4.1(a). We first construct a Poisson Q-Q plot of the \( i \)th ordered statistic \( X_{(i)} \) against the expected Poisson quantiles \( Q_i \) using \( \hat{\lambda} = 0.61 \). The
Q-Q plot of the death by horse kick data shown in Figure 4.2 does exhibit a linear association between observed and expected quantiles under Poissonity. Hence, we calculate the test statistic in Figure 4.2 to be \( R^2 = 0.9838 \).

![Sampling Distribution of \( R^2_{200,0.61} \)](image)

Figure 4.3. Simulated sampling distribution for \( R^2_{200,0.61} \) with critical values.

With \( n = 200 \) and \( \hat{\lambda} = 0.61 \) entered in the simulation program given in Figure 3.3, we obtain the sampling distribution for \( R^2_{200,0.61} \) shown in Figure 4.3. Also shown are the critical values \( R^2_{200,0.61; \alpha} \) for \( \alpha = .01, .05, \) and .10. Like the \( \chi^2 \) goodness-of-fit, we fix \( \alpha \) at .05 and conclude \( H_0 \) since \( R^2 = 0.9838 \) is well above the critical value \( R^2_{200,0.61; .05} = 0.8842 \). In fact, the observed value for \( R^2 \) corresponds to the 9431st smallest sorted value from the simulated distribution of \( R^2_{200,0.61} \) (see Table 4.2), thus a \( P \)-value of
Pr\left( R^2 \leq 0.9838 \right) = .9431. On the basis of the $R^2$ test, there is no evidence to contradict the hypothesis of Poissonity.

Table 4.2. Selected Portion of Empirical Distribution of $R^2_{200,0.61}$.

<table>
<thead>
<tr>
<th>Ordered $R^2_{200,0.61}$ Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>:</td>
</tr>
<tr>
<td>[9421] 0.9838 0.9838 0.9838 0.9838 0.9838 0.9838 0.9838 0.9838 0.9838</td>
</tr>
<tr>
<td>[9431] 0.9838 0.9839 0.9839 0.9839 0.9839 0.9839 0.9839 0.9839</td>
</tr>
<tr>
<td>[9441] 0.9839 0.9839 0.9839 0.9839 0.9839 0.9839 0.9839 0.9839 0.9839</td>
</tr>
<tr>
<td>:</td>
</tr>
</tbody>
</table>

4.2 Flying Bomb Hits

As an example of spatial distribution of random points, we consider the number of falling bombs and their points of impact in south London during WWII. The region of the city considered here, was divided into $n = 576$ subregions, each with an area of $1/4$ km$^2$. The number of subregions with $X$ number of hits can be found in Table 4.3(a) accompanied by the expected frequencies under Poissonity with $\lambda = 0.93$.

We calculate $\hat{\lambda}$, the mean number of flying bomb hits per $1/4$ km$^2$ subregion, to be $\bar{x} = 0.93$. We see from Table 4.3(a), a close consistency between observed and expected frequencies of flying bomb hits per subregion. In conducting the $\chi^2$ goodness-of-fit test we combine the last two classes in Table 4.3(a) in accordance with the rule of thumb that each $np_j \geq 5$. The binned data is shown in Table 4.3(b). This makes the number of classes $C = 5$. Having estimated $\lambda$, $\chi^2$ then has $5 - 1 - 1 = 3$ df. With $\chi^2 = 1.018$, we
Table 4.3. Empirical and Binned Flying Bomb Hits Data

<table>
<thead>
<tr>
<th>(a) Empirical Data</th>
<th>(b) Binned Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$f_x$</td>
</tr>
<tr>
<td>0</td>
<td>229</td>
</tr>
<tr>
<td>1</td>
<td>211</td>
</tr>
<tr>
<td>2</td>
<td>93</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>$\geq 5$</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>$n = 576$</td>
</tr>
</tbody>
</table>

fix $\alpha$ at .05 and consider the critical value $\chi^2_{3.05} = 7.815$. Thus, we conclude $H_0$, that there is reasonable fit of this sample of flying bomb hits to a Poisson distribution. The $P$-value, $\Pr(\chi^2 \geq 1.018) = .7969$, also supports the conclusion of $H_0$.

Figure 4.4. Q-Q plot of flying bomb hits data with $R^2$ test statistic.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
For the proposed $R^2$ test for Poissonity, we first consider the Poisson Q-Q plot of the flying bomb hits data, shown in Figure 4.4. The Q-Q plot does indicate a linear relationship between observed and expected quantiles under Poissonity. Hence, we calculate the test statistic in Figure 4.4 to be $R^2 = 0.9852$.

![Sampling Distribution of $R^2_{576,0.93}$](image)

Figure 4.5. Simulated sampling distribution for $R^2_{576,0.93}$ with critical values.

Running the simulation program in Figure 3.3 with $n = 576$ and $\hat{\lambda} = 0.93$, we obtain the sampling distribution for $R^2_{576,0.93}$ shown in Figure 4.5. Also shown are the critical values $R^2_{576,0.93,\alpha}$ for $\alpha = .01, .05, \text{and} .10$. Controlling $\alpha$ at .05, we conclude $H_0$ since $R^2 = 0.9838 > R^2_{576,0.93; .05} = 0.9429$. Likewise, the observed value for $R^2$ corresponds to the 8989th smallest sorted value from the simulated distribution of $R^2_{576,0.93}$ (see Table

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
4.4), thus the \( P \)-value for the test is \( \Pr \left( R^2 \leq 0.9852 \right) = .8989 \). Hence, there is enough support for \( H_0 \).

Table 4.4. Selected Portion of Empirical Distribution of \( R^2_{576,0.93} \).

<table>
<thead>
<tr>
<th>Ordered ( R^2_{576,0.93} ) Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>:</td>
</tr>
<tr>
<td>[8971] 0.9852 0.9852 0.9852 0.9852 0.9852 0.9852 0.9852 0.9852 0.9852</td>
</tr>
<tr>
<td>[8981] 0.9852 0.9852 0.9852 0.9852 0.9852 0.9852 0.9852 0.9852 0.9853</td>
</tr>
<tr>
<td>[8991] 0.9853 0.9853 0.9853 0.9853 0.9853 0.9853 0.9853 0.9853 0.9853</td>
</tr>
<tr>
<td>:</td>
</tr>
</tbody>
</table>

4.3 Rainstorms In The Mid-West

The following example found in Grant (1938), is derived from an engineering problem involving the occurrence of excessive rainfall. The area region considered in this application is that bounded by a line passing through St. Paul, MN., Detroit, MI., Knoxville, TN., Memphis, TN., Dodge City, KS., and Yankton, SD., where it is found to be homogeneous with respect to the frequencies of the intense storms which were studied. The data presented in Table 4.5 is a 33-year record of 10-minute intense rainstorms obtained from ten stations within the above-specified region of the Mid-West United States. This amounts to \( n = 330 \) station-years, where it has been shown in Grant (1938), that the storms at each station were independent events subject to the same system of chance causes. Thus, \( X \) denotes the number of intense 10-minute rainstorms per station-year.
We estimate the parameter \( \lambda \) with \( \bar{x} = 1.20 \) and interpret this as the mean number of intense 10-minute rainstorms per station-year. From the last two columns of Table 4.5, we see that the observed and expected frequencies of rainstorms per station-year are in close agreement. To apply the \( \chi^2 \) goodness-of-fit test, we combine the last three classes of Table 4.5(a) to give a minimum expected frequency \( n\hat{p}_j \) of 5 for each class \( j \). The binned data is found in Table 4.5(b). This makes the number of classes \( C = 5 \). Hence, \( \chi^2 \) has \( 5 - 1 - 1 = 3 \) df. Controlling \( \alpha \) at .05, we calculate the test statistic to be \( \chi^2 = 0.464 \) and compare this statistic with the critical value \( \chi^2_{3,95} = 7.815 \). Since \( \chi^2 = 0.464 < \chi^2_{3,95} \), we conclude \( H_0 \), that there is reasonable fit of the sample to a Poisson distribution. Likewise, the \( P \)-value is \( Pr(\chi^2 \geq 0.464) = .9267 \), and further supports the conclusion that a Poisson distribution is suitable for modeling this sample of rainstorm data.
To apply the $R^2$ test for Poissonity, we first plot the empirical quantiles based on the data in Table 4.5(a) against the expected Poisson quantiles $Q_i$ using $\lambda = 1.20$ to obtain the Q-Q plot in Figure 4.6. A linear association between observed and expected quantiles under Poissonity is evident in the Q-Q plot and is reflected in the test statistic $R^2 = 0.9799$.

Specifying $n = 330$ and $\lambda = 1.20$, we run the simulation program given in Figure 3.3 and obtain the sampling distribution for $R^2_{330,1.20}$ shown in Figure 4.7. The critical values $R^2_{330,1.20,\alpha}$ for $\alpha = .01$, .05, and .10 are also given. At the .05 significance level, we conclude $H_0$ since $R^2 = 0.9799$ is well above the critical value $R^2_{330,1.20,.05} = 0.9338$. Furthermore, the observed value for $R^2$ corresponds to the 8906th smallest sorted value from the simulated distribution of $R^2_{330,1.20}$ (see Table 4.6), thus the $P$-value is
Pr\(R^2 \leq 0.9799\) = .8906. As before, we conclude \(H_0\), that the rainstorm data in Table 4.5 is Poisson distributed.

![Sampling Distribution of \(R^2\)](image)

Figure 4.7. Simulated sampling distribution for \(R^2\) with critical values.

<table>
<thead>
<tr>
<th>Ordered (R^2) Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
<tr>
<td>[8911] 0.9800</td>
</tr>
</tbody>
</table>

Table 4.6. Selected Portion of Empirical Distribution of \(R^2\).

4.4 Noxious Weed Seeds

The data found in Table 4.7 is taken from Legatt (1935). The sample represents the number of noxious weed seeds \(X\), found in \(n = 98\) subsamples of *Phleum praetense*
(meadow grass). Each of the 98 subsamples weighed 1/4 oz and obviously contained many seeds, of which only a small percentage were noxious.

Table 4.7. Empirical and Binned Noxious Weed Seed Data

<table>
<thead>
<tr>
<th>(a) Empirical Data</th>
<th>(b) Binned Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$f_x$</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>$\geq 11$</td>
<td>0</td>
</tr>
</tbody>
</table>

$C = 7$  $n = 98$  $\bar{x} = 3.02$  $\chi^2 = 3.802$  $df = 7 - 1 - 1 = 5$

We calculate $\hat{\lambda}$, the mean number of noxious weed seeds per 1/4 oz subsample, to be $\bar{x} = 3.02$. The agreement between observed and expected frequencies seems good with the exception of $x = 2$ and $x = 3$, where the expected frequencies for these classes are nearly equal while the observed frequencies are 26 and 16. For the $\chi^2$ goodness-of-fit test, we combine the last six classes in Table 4.7(a) in accordance with the rule $n\hat{p}_j \geq 5$ for all $j$, making the number of classes $C = 7$. The binned data is shown in Table 4.7(b). Having estimated the parameter $\hat{\lambda}$, $\chi^2$ then has $7 - 1 - 1 = 5$ df. Supposing we fix $\alpha$ at .05, the calculated test statistic $\chi^2 = 3.802$ is then compared with the critical value.
Thus, since $\chi^2 = 3.802 < \chi^2_{5,95}$, we conclude $H_0$, that there is goodness-of-fit of this sample of noxious weed seeds to a Poisson distribution. Likewise, the $P$-value is $\Pr(\chi^2 \geq 3.802) = .5783$, so $H_0$ is supported. The discrepancies between the observed and expected frequencies under Poissonity are not unusually large.

![Q-Q Plot of Weed Seed Data](image)

Figure 4.8. Q-Q plot of weed seed data with $R^2$ test statistic.

For the $R^2$ test for Poissonity, we start with the Poisson Q-Q plot of weed seed data in Figure 4.8, where a clear linear association between the empirical quantiles, from the data in Table 4.7(a), and the expected Poisson quantiles $Q_i$, with $\hat{\lambda} = 3.02$, is seen. The strength of the linearity displayed in the Q-Q plot is reflected by the calculated test statistic $R^2 = 0.9574$.

Running the simulation program in Figure 3.3 having specified $n = 98$ and $\hat{\lambda} = 3.02$ yields the sampling distribution for $R^2_{98,3.02}$ shown in Figure 4.9. With $\alpha = .05$, we
conclude $H_0$ noting that $R^2 = 0.9574 > R^2_{98,3,02; 0.05} = 0.9242$. With the observed value for $R^2$ corresponding to the 5814th smallest sorted value from the simulated distribution of $R^2_{98,3,02}$ (see Table 4.8), we calculate the $P$-value to be $\Pr(R^2 \leq 0.9574) = .5814$. Hence, the number of weed seeds found in the 98 subsamples of meadow grass seeds can be modeled with a Poisson distribution.

Table 4.8. Selected Portion of Empirical Distribution of $R^2_{98,3,02}$.

<table>
<thead>
<tr>
<th>Ordered $R^2_{98,3,02}$ Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
</tr>
<tr>
<td>[5791] 0.9573 0.9573 0.9573 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574</td>
</tr>
<tr>
<td>[5801] 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574</td>
</tr>
<tr>
<td>[5811] 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574 0.9574</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
4.5 Alpha Emission From Radioactive Source

Radioactive disintegration can be monitored by the emission of $\alpha$-particles by some radioactive substance. The number of $\alpha$-particles reaching a given portion of space during some relatively short period of time is a classic example of phenomena obeying a Poisson model. Among the early research efforts in radioactivity was a famous study of alpha emission done by Rutherford and Geiger (1910). Their experiment consisted of a source of the element polonium placed a short distance from a counting screen. The number of $\alpha$-particles $X$ impinging on the screen, were recorded for each of $n = 2608$ 1/8 minute intervals. The empirical data can be found in the first two columns of Table 4.9(a). Table 4.9(b) displays binned data for this final example.

<table>
<thead>
<tr>
<th>Table 4.9. Empirical and Binned Alpha Emission Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Empirical Data</td>
</tr>
<tr>
<td>$x$</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>$n = 2608$</td>
</tr>
<tr>
<td>$\bar{x} = 3.87$</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Here, the mean number of $\alpha$-particles emitted per 1/8-minute interval, as estimated from the empirical data, is $\hat{\lambda} = 3.87$. We see from Table 4.9(a) that the numbers for observed and expected frequencies are consistent with one another with some exceptions for $x = 2, 4, 5, 6, \text{ and } 8$; however, given the very large sample size, these differences are relatively small. In evaluating the fit of the sample to a Poisson distribution with the $\chi^2$ goodness-of-fit test, we combine the last five classes in Table 4.9(a) according to the rule of thumb $n\hat{\lambda}_j \geq 5$ to obtained the binned data in Table 4.9(b). Thus, the number of classes we consider is $C = 12$ and $\chi^2$ then has $12 - 1 - 1 = 10$ df. If we compare the calculated test statistic $\chi^2 = 12.974$ with the critical value $\chi^2_{10,0.95} = 15.987$, we conclude $H_0$, that there is reasonable fit of the sample to a Poisson distribution. The same conclusion would be made had we considered the $P$-value of the test given by $\Pr(\chi^2 \geq 12.974) = .2251$, which is still greater than the specified significance level $\alpha$.

![Q-Q Plot of Alpha Emission Data](image)

Figure 4.10. Q-Q plot of alpha emission data with $R^2$ test statistic.
To use the $R^2$ test for Poissonity on the alpha emission data, we begin by constructing the Poisson Q-Q plot of the ordered statistics $X_{(i)}$ and the expected quantiles $Q_i$ under Poissonity with $\hat{\lambda} = 3.87$ shown in Figure 4.10. The Q-Q plot clearly exhibits a linear association between $X_{(i)}$ and $Q_i$ and the strength of the observed linear association between the two variables is explained in the test statistic $R^2 = 0.9864$.

![Sampling Distribution of $R^2_{2608,3.87}$](image)

Figure 4.11. Simulated sampling distribution for $R^2_{2608,3.87}$ with critical values.

To compare this observed test statistic with a specified critical value, we first run the simulation program in Figure 3.3 with $n = 2608$ and $\hat{\lambda} = 3.87$ to obtain the sampling distribution for $R^2_{2608,3.87}$ shown in Figure 4.11. To test $H_0$ at $\alpha = .05$, we consider the critical value $R^2_{2608,3.87,.05} = 0.9862$. Since $R^2 = 0.9864 > R^2_{2608,3.87,.05}$, we conclude $H_0$ as we did when using the $\chi^2$ test. The $P$-value is $\Pr(R^2 \leq 0.9864) = .0577$ since the
observed value for the test statistic $R^2$ corresponds to the 577th smallest sorted value of the simulated sampling distribution of $R^2_{2608,3.87}$ (see Table 4.10). The $P$-value is still greater than $\alpha$, although not by much, and our decision to conclude $H_0$ at the .05 significance level remains the same.

Table 4.10. Selected Portion of Empirical Distribution of $R^2_{2608,3.87}$.

<table>
<thead>
<tr>
<th>Ordered $R^2_{2608,3.87}$ Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>:</td>
</tr>
<tr>
<td>[551] 0.9863 0.9863 0.9863 0.9863 0.9863 0.9863 0.9863 0.9864 0.9864 0.9864</td>
</tr>
<tr>
<td>[561] 0.9864 0.9864 0.9864 0.9864 0.9864 0.9864 0.9864 0.9864 0.9864 0.9864</td>
</tr>
<tr>
<td>[571] 0.9864 0.9864 0.9864 0.9864 0.9864 0.9864 0.9864 0.9865 0.9865 0.9865</td>
</tr>
<tr>
<td>:</td>
</tr>
</tbody>
</table>

We note here, that had we controlled $\alpha$ at say .10, we would have concluded otherwise in this example, that the sample is not Poisson distributed. With respect to the $\chi^2$ goodness-of-fit approach to this example, we would still conclude $H_0$ at the .10 significance level given that the $P$-value for the observed $\chi^2$ statistic was calculated to be .2251. Thus, as mentioned in Chapter 3, future studies on the power of the $R^2$ test for Poissonity for a fixed significance level $\alpha$ is needed to determine the sensitivity of the proposed test in detecting $H_i$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
CHAPTER 5

DISCUSSION AND RECOMMENDATIONS

We have taken a broad survey of the methodologies behind goodness-of-fit for univariate distributions. Having researched the large collection of literature on goodness-of-fit, it became evident that techniques for univariate continuous distributions outweigh those proposed for univariate discrete distributions in both number and practical appeal. In particular, we found no graphical based test of fit for the discrete case. Hence, in this thesis, we examined a formal graphical goodness-of-fit approach for discrete-type distributions and have illustrated this proposed test through the Poisson model for which many applications exist.

The proposed graphical approach for testing Poissonity is rooted in the information contained in a Poisson Q-Q plot, where good distributional fits result in near-linear plots. As a simple measure of linearity and as a means of summarizing the information in the Poisson Q-Q plot, we proposed the squared correlation coefficient $R^2$ between $X_{(i)}$ and $Q_i$ as a formal test statistic, which is conveniently provided in the output of most statistical software packages having regressed $Q_i$ on $X_{(i)}$. The distribution of the test statistic $R^2$ was then determined with empirical sampling through a parametric bootstrap simulation for which the code is provided. With the sampling distribution of $R^2$ under Poissonity known and with an underlying rationale that near-linear Q-Q plots will be
reflected by high values of \( R^2 \), we defined a lower-tail rejection region and decision rule to test the composite hypothesis of Poissonity. Thus, the proposed \( R^2 \) test for Poissonity: (i) lends itself to *graphical representation*; (ii) is *conceptually simple*; and (iii) is *computationally convenient* through computer application. Furthermore, the methodology behind the \( R^2 \) test for Poissonity can be extended to investigate hypotheses for goodness-of-fit of other discrete distributions.

In applying the proposed \( R^2 \) test for Poissonity to several historical and well-known Poisson data sets for a specified significance level \( \alpha \), the proposed GOF test gave the same conclusions as the classical and commonly used \( \chi^2 \) goodness-of-fit test. The power of the proposed \( R^2 \) test for Poissonity, however, was not investigated in this paper; rather, a general approach to the problem was given. Thus, as a recommendation for future and related work on the subject of testing Poissonity with the \( R^2 \) test, we suggest a power comparison study be done to determine the sensitivity of the \( R^2 \) test in detecting the alternative \( H_1 \), relative to other existing and commonly used goodness-of-fit tests for Poissonity.
NOTES ON THE POISSON DISTRIBUTION

A random variable $X$, which denotes discrete numbers of successes, is said to have an underlying Poisson distribution with parameter $\lambda$, that is, $X \sim \text{Pois}(\lambda)$, if and only if its probability distribution is given by the function

$$p(x; \lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} & x = 0, 1, 2, \ldots ; \lambda > 0 \\ 0 & \text{elsewhere} \end{cases}$$

(A.1)

A.1 Derivation of The Poisson Distribution

The Poisson distribution can be derived as a limiting case of the binomial distribution. By definition, a random variable $X$ has a binomial distribution, and it is referred to as a binomial random variable, if and only if its probability distribution is given by

$$b(x; n, \theta) = \begin{cases} \binom{n}{x} \theta^x (1 - \theta)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{elsewhere} \end{cases}$$

(A.2)

where $n$ is the number of trials, $x$ denotes the number of successes, and $\theta$ being the probability of success, which remains constant from trial to trial. Now consider the case where $n$ is very large ($n \to \infty$), the probability of success $\theta$ becomes small ($\theta \to 0$),
while the product \( n\theta \) remains constant, say \( n\theta = \lambda \). Hence, \( \theta = \frac{\lambda}{n} \) and we can write

\[
b(x; n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}
\]

\[
= \binom{n}{x} \left( \frac{\lambda}{n} \right)^x \left( 1 - \frac{\lambda}{n} \right)^{n-x}
\]

\[
= \frac{n!}{x!(n-x)!} \left( \frac{\lambda}{n} \right)^x \left( 1 - \frac{\lambda}{n} \right)^{n-x}.
\]

Expanding \( \frac{n!}{x!(n-x)!} \), we obtain

\[
\frac{n!}{x!(n-x)!} = \frac{n(n-1)(n-2) \cdots (n-x+1)}{x!},
\]

and it can be seen that there are \( x \) factors in \( n(n-1)(n-2) \cdots (n-x+1) \). By dividing each of these \( x \) factors by \( n \) from \( \left( \frac{\lambda}{n} \right)^x = \frac{\lambda^x}{n^x} \), we can rewrite \( b(x; n, \theta) \) in (A.3) as

\[
b(x; n, \theta) = \frac{1}{x!} \left( 1 - \frac{1}{n} \right) \left( 1 - \frac{2}{n} \right) \cdots \left( 1 - \frac{x-1}{n} \right) \left( 1 - \frac{\lambda}{n} \right)^{n-x}.
\]

Rewriting \( 1 - \frac{\lambda}{n} \) as

\[
\left( 1 - \frac{\lambda}{n} \right)^{n-x} = \left( \frac{1 - \lambda}{n} \right)^{-\frac{x}{n}} \left( \frac{1 - \lambda}{n} \right)^{-\frac{x}{n}} = \left( \frac{1}{n} \right)^{-\frac{x}{n}} \left( 1 - \frac{\lambda}{n} \right)^{-\frac{x}{n}},
\]

we obtain
\[
\frac{1}{x!} \prod_{k=0}^{x-1} \frac{1}{n} = \frac{1}{x!} \prod_{k=0}^{x-1} \left(1 - \frac{k}{n}\right) 
\]

\[(A.7)\]

For fixed \(x\), if we let \(n \to \infty\), while \(\lambda = n\theta\) remains constant, it can be seen that

\[
\lim_{n \to \infty, \lambda = n\theta} \frac{1}{x!} \prod_{k=0}^{x-1} \left(1 - \frac{k}{n}\right) = \frac{1}{x!},
\]

\[
\lim_{n \to \infty, \lambda = n\theta} \left(1 - \frac{\lambda}{n}\right) = e^{-\lambda},
\]

and

\[
\lim_{n \to \infty, \lambda = n\theta} \left(1 - \frac{\lambda}{n}\right)^{-x} = 1.
\]

Thus, for fixed \(x\) and as \(n \to \infty\) while \(\lambda = n\theta\) remains constant, the limiting distribution of the \(b(x;n,\theta)\) is

\[
\lim_{n \to \infty, \lambda = n\theta} b(x;n,\theta) = \frac{\lambda^x e^{-\lambda}}{x!},
\]

\[(A.8)\]

which if defined for \(x = 0, 1, 2, \ldots\) and \(\lambda > 0\), is the Poisson distribution in (A.1).

### A.2 Poisson Distribution and PDF Properties

Consider the function \(f(x;\lambda)\) defined by

\[
f(x;\lambda) = \begin{cases} 
\frac{\lambda^x e^{-\lambda}}{x!} & \text{for } x = 0, 1, 2, \ldots \\
0 & \text{elsewhere}
\end{cases}
\]

\[(A.9)\]

with parameter \(\lambda > 0\). Since \(\lambda > 0\), then \(f(x) \geq 0\) and

\[
\sum_{x=0}^{\infty} f(x;\lambda) = \sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!}.
\]

\[(A.10)\]
Since,
\[
\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = 1 + \frac{\lambda}{2!} + \frac{\lambda^3}{3!} + \ldots
\]
converges for all values of \(\lambda\), to \(e^\lambda\), we have
\[
\sum_x f(x; \lambda) = \ldots = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} \cdot e^\lambda = 1. \tag{A.11}
\]
Therefore, \(f(x; \lambda)\) in (A.9) satisfies the conditions of being a probability density function (PDF) of a discrete type random variable and any such \(f(x; \lambda)\) is known as a Poisson PDF.

A.3 MGF of a Poisson Distribution

The moment generating function (MGF) of a discrete random variable \(X\), denoted by \(M_X(t)\), is defined as
\[
M_X(t) = E[e^{tx}] = \sum_x e^{tx} \cdot f(x) \quad \text{for } t \in \mathbb{R} \tag{A.12}
\]
It follows then that the moment generating function of the Poisson distribution is given by
\[
M_X(t; \lambda) = \sum_{x=0}^{\infty} e^{tx} \cdot \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \tag{A.13}
\]
\[
= e^{-\lambda} \left[ 1 + \frac{\lambda e^t}{1!} + \frac{(\lambda e^t)^2}{2!} + \frac{(\lambda e^t)^3}{3!} + \ldots \right].
\]
Since \( 1 + \frac{(\lambda e^t)}{1!} + \frac{(\lambda e^t)^2}{2!} + \frac{(\lambda e^t)^3}{3!} + \cdots \) in (A.13) converges to \( e^{\lambda t} \), we can rewrite (A.13) as

\[
M_X(t; \lambda) = e^{-\lambda} \cdot e^{\lambda t} = e^{\lambda(t-1)}.
\]

Hence, if \( X \sim \text{Pois}(\lambda) \), it has a moment generating function of the form

\[
M_X(t; \lambda) = e^{\lambda(t-1)} \quad \text{for } t \in \mathbb{R}.
\]

### A.4 Mean and Variance of The Poisson Distribution

Using \( M_X(t; \lambda) \) in (A.15), we can derive the mean \( \mu \) and the variance \( \sigma^2 \) of the Poisson distribution as follows:

(i) The mean of a probability distribution is defined as \( \mu = M'(0) \). Thus, in the Poisson case,

\[
\mu = M'(0) = \frac{d}{dt} \left[ e^{\lambda(t-1)} \right]_{t=0} = e^{\lambda(t-1)} \cdot (\lambda e^t) \bigg|_{t=0} = \lambda.
\]

(ii) The variance of a probability distribution is defined as \( \sigma^2 = M''(0) - \mu^2 \). Here,

\[
M''(0) = \frac{d}{dt} \left[ M'(t; \lambda) \right]_{t=0} = e^{\lambda(t-1)}(\lambda e^t) + e^{\lambda(t-1)}(\lambda e^t) \cdot (\lambda e^t) \bigg|_{t=0} = \lambda + \lambda^2.
\]
It then follows that the variance $\sigma^2$ of the Poisson distribution is

$$\sigma^2 = \lambda + \lambda^2 - \lambda^2 = \lambda. \quad (A.18)$$

Thus, in the Poisson case,

$$\mu = \sigma^2 = \lambda \quad (A.19)$$

and the parameter $\lambda > 0$ is all that is needed to specify the Poisson distribution.

We may also derive the mean and variance of a distribution using the original definitions as follows:

(i) The mean $\mu$ of a random variable $X$ is defined as $E[X] = \sum_{x} x \cdot Pr[X = x]$. Thus,

in the Poisson case,

$$E[X] = \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}(x-1)!}{(x-1)!}$$

$$= \lambda e^{-\lambda} \left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots + \frac{\lambda^{n-1}}{(n-1)!} + \cdots\right). \quad (A.20)$$

It can be seen that $1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots + \frac{\lambda^{n-1}}{(n-1)!} + \cdots$ in (A.20) converges to $e^\lambda$ for all values of $\lambda$. It then follows that

$$E[X] = e^{-\lambda} \cdot \lambda \cdot e^\lambda = \lambda. \quad (A.21)$$

(ii) The variance $\sigma^2$ of a random variable $X$ is defined as $Var[X] = E[(X - \mu)^2]$

$$= E[X^2] - (E[X])^2. \quad \text{Thus, in the Poisson case, } Var[X] = E[X^2] - \lambda^2. \quad \text{Here,}$$
\[ E[X^2] = \sum_{x=0}^{\infty} x^2 \cdot \Pr[X = x] \]
\[ = \sum_{x=0}^{\infty} x^2 \cdot \frac{\lambda^x e^{-\lambda}}{x!} \], substituting \( x^2 = x(x-1) + x \)
\[ = \sum_{x=0}^{\infty} \left[ x(x-1) + x \right] \cdot \frac{\lambda^x e^{-\lambda}}{x!} \]
\[ = \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\lambda^x e^{-\lambda}}{x!} + \sum_{x=0}^{\infty} x \cdot \frac{\lambda^x e^{-\lambda}}{x!} \]
\[ = \left( e^{-\lambda} \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\lambda^x}{x!} \right) + \lambda. \]  

Since
\[ \sum_{x=0}^{\infty} x(x-1) \cdot \frac{\lambda^x}{x!} = \lambda^2 \left( 1 + \frac{\lambda}{2!} + \frac{\lambda^2}{3!} + \cdots \right), \]  

we can rewrite \( E[X^2] \) as
\[ E[X^2] = e^{-\lambda} \left( \lambda^2 \cdot e^{\lambda} \right) + \lambda \]
\[ = \lambda^2 + \lambda \]  

Finally,
\[ Var[X] = E[X^2] - \lambda^2 \]
\[ = \lambda^2 + \lambda - \lambda^2 \]
\[ = \lambda. \]
REFERENCES


93


VITA

Graduate College
University of Nevada, Las Vegas

Davin P. Padilla

Home Address:
690 Kilaha Place
Hilo, Hawaii 96720

Degrees:
Bachelor of Arts in Mathematics (2000)
University of Hawaii at Hilo

Certifications:
Teacher Education Program, University of Hawaii at Hilo

Special Honors and Awards:
Graduate Assistant (2001 – 2003)
Department of Mathematical Sciences, University of Nevada, Las Vegas

Outstanding Senior in Mathematics Award (2000)
Department of Mathematics, University of Hawaii at Hilo

University of Hawaii Presidential Scholar Award (1998 – 2000)
University of Hawaii at Hilo

Salutatorian Award (1996)
Class of 1996, Hilo High School, Hilo, Hawaii

Eagle Scout (1995)
Boy Scouts of America, Troop 78, Hilo, Hawaii

Thesis Title: A Graphical Approach for Goodness-of-Fit of Poisson Model

Thesis Examination Committee:
Chairperson, Dr. Ashok K. Singh, Ph. D.
Committee Member, Dr. Malwane Ananda, Ph. D.
Committee Member, Dr. Rohan Dalpatadu, Ph. D.
Graduate Faculty Representative, Dr. Laxmi Gewali, Ph. D.