Sample Size Estimation in the Multinomial Model

A Thesis

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science

with a

Major in Statistical Science

in the

College of Graduate Studies

University of Idaho

by

Martyna Lukaszewicz

Major Professor: Brian Dennis, Ph.D.

Committee Members: Erkan Buzbas, Ph.D.; Christopher Remien, Ph.D.

Department Administrator: Christopher Williams, Ph.D.

August 2018

Authorization to Submit Thesis

This thesis of Martyna Lukaszewicz, submitted for the degree of Master of Science with a Major in Statistical Science and titled “Sample Size Estimation in the Multinomial Model,” has been reviewed in the final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor: ________________________________ Date: ____________________

Brian Dennis, Ph.D.

Committee Members: ________________________________ Date: ____________________

Erkan Buzbas, Ph.D.

________________________________ Date: ____________________

Christopher Remien, Ph.D.

Department

Administrator: ________________________________ Date: ____________________

Christopher Williams, Ph.D.

Abstract

Predicting the timing of phenological events is important in agriculture, especially high revenue products. A project sponsored by USDA-ARS had the objective of adapting a previously developed model for estimating proportions of insects in different development stages as a function of temperature (degree) and time (days) for predicting bloom in almond orchards. Data for the model normally form a two-way table of counts, with rows corresponding to sample proportions of different development stages, and columns to sampling times. The data from the almond growers, however, proved problematic in that the percentages of trees in development stages were recorded but not the counts. Maximum likelihood estimation of model parameters was possible, but the variances of the estimates depend on sample size. This thesis reports a technique developed to estimate sample sizes of multinomial and product multinomial models with known proportions when empirical proportions are available but not the counts (sample size).

Acknowledgements

I would like to thank my advisor, Dr. Brian Dennis, for believing in me, encouraging me, and for his patience. I would also like to thank my committee members: Dr. Christopher Remien and Dr. Erkan Buzbas for dedicating their time to review this thesis. Furthermore, I would like to express my appreciation to Dr. Christopher Williams for his advice and insight throughout my master’s degree.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Prediction of the timing of developmental stages of plants and insects is important in agricultural management. "Phenology," or the timing of development stages, is a complex process depending on many factors such as weather and time [1]. According to the United States Department of Agriculture, the 2018-2019 U.S. wheat crop is projected at 1,821 million bushels, a 5 percent increase from previous year. The average projected farm price is between $4.50 and $5.50 per bushel [2]. Actual yields for an individual farmer will depend on the collective management actions taken by the farmer for pest control, pollination, and soil fertility. Such actions can vary greatly in effectiveness depending on the development stage of the crop and/or pest. It is, therefore important to improve methods of forecasting the phenology of plants as well as insect pests.

Dennis et al. [1] developed a model to predict proportions of insects in different development stages as a function of accumulated degree-days (DD). The data for the model is a two-way table of counts, with each row giving the counts of different development stages recorded in a sample of insects taken at a particular time. The model, known in the literature as the Dennis-Kemp model, specifies logistic functions for how the stage proportions change through time [3]. The functions contain unknown parameters requiring statistical estimation. The likelihood function is product-multinomial, each multinomial corresponding to one row of the data table. Various statistical inferences for the model have been presented (Dennis et al. 1986, Dennis and Kemp 1988) based on standard asymptotic theory for multinomial models (e.g. Bishop et al,1975) [4].

This project was motivated by a non-standard phenology data set that was collected by the California almond industry. Of critical importance in almond growing is the scheduling of placement of honeybees for pollination. The almond trees go through different phenological stages during a growing season, and the bees must be brought in at the onset of flowering for optimal production. There was a USDA-ARS project implemented to adapt the Dennis-Kemp phenology model for use by almond growers. Phenology data on almond trees had been collected by the almond growers over many years. However, the data proved to have a serious shortcoming: the two-way tables contained percentages rather than counts (each row adding to 100), and, to make things worse, the sample sizes corresponding to the row percents were not recorded. The question arose: can the sample sizes be estimated? Theoretically there is information about sample sizes in percent-only data, in that in a multinomial model the magnitude of the departures of empirical proportions from the modeled proportions, that is, the variability in the data, depends on sample size. It became apparent from the literature that the estimation of sample size in multinomial models with data on proportions but not counts had not been studied. Development of a model for sample size estimation when only the proportions are known will allow for expansion of estimation of variability of forecasting plant and insect phenology, not only to the data with known counts.

1.2 Previous Work

"Estimation of sample size in multinomial models" has many different meanings and contexts. Outlined here are some of the questions that have been previously addressed under that broad banner. Some of the questions involve survey design, that is, determination of how large a sample is needed to achieve some inferential goal. Other questions involve the sample size being unknown due to one or more missing counts, as in mark-recapture models (in which the count of animals not trapped is missing).

Eichenberger et al. [5] developed a model for sample size determination in survey design for groups that might be not detected by the sample. They formulated a technique for determining the smallest sample size necessary to ensure that a group is represented in the sample with probability of at least $1-\alpha$. In the multinomial phenology models, the group probabilities change through time. However, the Eichenberger et al. method nonetheless could prove useful for designing phenology field studies when the focus is on one particular row (time) of counts.

Thompson [6] proposed a method of selecting the smallest sample size n for a random sample from population with known multinomial proportions p_j , $j=1,2,...,r$, such that the probability will be at least $1-\alpha$ that all estimated proportions $\tilde{p}_1, \tilde{p}_2, ..., \tilde{p}_r$ will simultaneously be within specified distances of true population proportions. Chosen distances require some prior knowledge about the population of interest. Although Thompson's model does not apply to a product multinomial, it could be adapted for a particular sampling time of interest in a phenology study.

Otis et al. [7] summarized and improved earlier work from the 1950s of population size estimation for mark-recapture in a closed population model. In a population of size N , for a total of q sampling times, $i=1,2,...,q$, on each sampling occasion there are 2 possible outcomes; an individual is either captured or not captured. There are 2^q possible capture histories, $j=1,2,...,2^q$. For the number of individuals captured at ith sampling time y_j , the number of individuals not captured in the experiment is:

$$n_t = n - \sum_{i=1}^{t-1} y_j \,. \tag{1}$$

The last count is missing from the data, posing an estimation problem that is equivalent to having a missing sample size n. Mark-recapture differs from the problem investigated here in that actual counts are available in mark-recapture data (just not all of them).

1.3 Thesis Objective

This project proposes and evaluates a method to estimate sample size n for a multinomial model when the empirical proportions are known but not the counts. Sample size estimation is obtained by the method of moments approach, using the relationship of multinomial counts with the chi-squared distribution. Confidence intervals for n are developed as well. In chapter 2, a simplified version of the problem is studied: estimation of the sample size n in a single multinomial model, when the empirical proportions, but not the counts, are available. An estimate is developed and circumstances are described for when the estimate will work well. In chapter 3 the full problem raised by the nonavailability of counts in phenology data is tackled. Specifically, the problem of inference for the Dennis et al. [1] phenology model when the empirical proportions, but not the counts, is investigated. . Chapter 3 proposes rules for pooling sparse cells in datasets when the method described method in chapter 2 fails. The data from Blue Diamond® Growers Nonpareil almonds [8] serve to illustrate the concepts.

Chapter 2

Estimation of n with Known Proportions

2.1 Purpose

This section describes a proposed method for estimating the unknown sample count n for a multinomial model with r possible outcomes, $j = 1, 2, ..., r$, with known probabilities $p_1, p_2, ..., p_r$. The count data $y_1, y_2, \ldots, y_r$ along with the sample size n are assumed missing, but data in the form of empirical sample proportions $\tilde{p}_1, \tilde{p}_2, ..., \tilde{p}_r$, where $\tilde{p}_j = y_j / n$, are available.

2.2 The Method

Here is proposed a method to estimate sample size (n) using a moment estimate based on a chi-squared goodness-of-fit statistic. The purpose of developing the method was to study such estimation in a simple setting before adapting it to a more complex problem described in the next chapter [1]. As well, this chapter explores and maps out the scenarios for which the chi-squared estimate of sample size performs well.

Two multinomial goodness-of-fit tests, one based on Pearson's X^2 and one based on the log-likelihood ratio G^2, are described. The statistical properties of the two tests are compared in terms of performance when the sample counts are low. Next a moment estimate of sample size n is derived based on Pearson's X^2 statistic for use when the data consist only of sample proportions. Then improvements in the sample size estimate are discussed; in particular a bias correction is derived. The bias correction improves performance of a $100(1-\alpha)\%$ confidence interval for n .

2.3 Chi-squared

The Pearson goodness-of-fit statistic for a single multinomial with known group proportions p_1, p_2, …, p_r is given by

$$\mathrm{X}^2 = \sum_{j=1}^{r} \left[\frac{\left(y_j - np_j\right)^2}{np_j} \right]. \tag{2}$$

The Pearson statistic can be rewritten by factoring n out, thereby expressing the statistic in terms of known sample proportions and empirical proportions:

$$\mathrm{X}^2 = n\sum_{j=1}^{r} \left[\frac{\left(\tilde{p}_j - p_j\right)^2}{p_j} \right] = nD^2, \tag{3}$$

where the multiplier of n is a statistic D^2 giving the deviance of the empirical proportions from the model proportions. As n becomes large, the sampling distribution of the Pearson statistic asymptotically approaches a chi-squared distribution with $r-1$ degrees of freedom.

2.4 Log-Likelihood Ratio

The log-likelihood ratio statistic G^2, where

$$G^2 = -2n\sum_{j=1}^{r} \tilde{p}_j \log\left(\frac{p_j}{\tilde{p}_j}\right) = 2n\sum_{j=1}^{r} \tilde{p}_j \log\left(\frac{\tilde{p}_j}{p_j}\right) \tag{4}$$

tests the goodness-of-fit of a model. The statistic is also known as $-2\log\Lambda$ [9] where Λ is the likelihood ratio:

$$\Lambda = \frac{\left[\dfrac{n!}{y_1!y_2!...y_r!}\right] p_1^{y_1} p_2^{y_2} ... p_r^{y_r}}{\left[\dfrac{n!}{y_1!y_2!...y_r!}\right] \tilde{p}_1^{y_1} \tilde{p}_2^{y_2} ... \tilde{p}_r^{y_r}}. \tag{5}$$

When $\tilde{p}_j = 0$, $\tilde{p}_j \log\left(\frac{\tilde{p}_j}{p_j}\right) = 0$ by l'Hôpital's rule. The cell probabilities $p_1, p_2, ..., p_r$ are described by the model H_0. Under model H_1 cell probabilities are unconstrained except for $\sum_{j=1}^{r} p_j = 1$. When H_0 is true and n is large, $-2\log\Lambda$ and Pearson's statistic are asymptotically equivalent. The Taylor series expansion $\sum_{k=0}^{\infty}\left[\frac{f^{(i)}(x_0)}{k!}\right](x-x_0)^k$, of the function

$f(y) = y\log\left(\frac{y}{y_0}\right)$ about y_0 is:

$$f(y) = (y - y_0) - \frac{1}{2}(y - y_0)^2 \frac{1}{y_0} + \dots \tag{6}$$

The Taylor series approximation p_j applied to G^2 then yields

$$G^2 \approx 2n\sum_{j=1}^{r}\left(\tilde{p}_j - p_j\right) + n\sum_{j=1}^{r}\left[\frac{\left(\tilde{p}_j - p_j\right)^2}{p_j}\right], \tag{7}$$

and by the cancellation of the differences between $\tilde{p}_j$ and p_j [10], we have

$$G^2 \approx n\sum_{j=1}^{r}\left[\frac{\left(\tilde{p}_j - p_j\right)^2}{p_j}\right]. \tag{8}$$

2.5 Method of Moments

A method of moments estimate of the unknown parameter n is constructed by setting X^2 equal to its expected value, the degrees of freedom $r-1$. The moment estimate of n follows from algebraic solution and is $\mathrm{E}(\mathrm{X}^2)$ divided by the deviance statistic:

$$\tilde{N} = \frac{r-1}{D^2}. \tag{9}$$

The estimate increases as the deviance from chi-squared distribution decreases. Thus, the variability of the empirical proportions around the model proportions contains information for estimating n. The Pearson statistic was chosen for the basis of estimating n rather than the likelihood ratio statistic because the Pearson statistic is known to have superior asymptotic properties, such as for sparse tables [11].

2.6 Gamma-Chi-squared Relationship

For the purpose of evaluating the sampling distribution of $\tilde{N}$, we rewrite it as $\tilde{N} = n\left(r-1\right)/\mathrm{X}^2$, where $\mathrm{X}^2 \sim$ chi-squared($r-1$). A chi-squared random variable divided by a constant has a gamma distribution, and so $\tilde{n}$ is seen to be the reciprocal of a gamma random variable.

Suppose the random variable V has a gamma(α, β) distribution. The moment generating function of V is

$$m_V(s) = \mathrm{E}\left(e^{sV}\right) = \left(\frac{\beta}{\beta - s}\right)^{\alpha}, \tag{10}$$

and the probability density function is

$$f_V\left(v\right) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} v^{\alpha-1} e^{-\beta v}, \tag{11}$$

where α and β are positive parameters.

[12]. The chi-squared($r-1$) distribution of X^2 is a special case of a gamma distribution [12], with $\alpha = \dfrac{r-1}{2}$ and $\beta = 1/2$.

Let $Y = V/c$, where $c = n(r-1)$. The moment generating function of Y is then:

$$m_Y(s) = \mathrm{E}\left(e^{sV/c}\right) = m_V\left(\frac{s}{c}\right) = \left(\frac{c\beta}{c\beta - s}\right)^{\alpha}. \tag{12}$$

The random variable Y has a gamma($\alpha, c\beta$) distribution, with the same shape parameter as V but different rate parameter. The expected value of its reciprocal is:

$$\mathrm{E}\left(\frac{1}{Y}\right) = c\mathrm{E}\left(\frac{1}{V}\right). \tag{13}$$

2.7 Bias Correction and Confidence Intervals

We have established that the moment estimate of n can be written as $\tilde{N} = 1/Y$, where $Y \sim \text{gamma}\left(\frac{r-1}{2}, \frac{n(r-1)}{2}\right)$. A random variable from a gamma(α,β) distribution has its expected value of its inverse $\frac{1}{V}$ defined as $\int_0^\infty \left(\frac{1}{v}\right) f_V dv$ from the property of expected value [12]. In the case of gamma,

$$\mathrm{E}\left(\frac{1}{V}\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty v^{(\alpha-1)-1} e^{-\beta v} dv = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha - 1)}{\beta^{\alpha-1}} = \frac{\beta}{\alpha - 1}. \tag{14}$$

For a shape parameter $\frac{r-1}{2}$ and rate parameter $\frac{n(r-1)}{2}$ the expected value of the reciprocal of Y becomes

$$\mathrm{E}\left(\tilde{N}\right) = \mathrm{E}\left(\frac{1}{Y}\right) = \frac{n(r-1)/2}{\frac{r-1}{2} - 1} = \frac{n(r-1)}{(r-3)}. \tag{15}$$

Thus, the moment estimate $\tilde{N}$ is biased low by a proportion $\frac{(r-1)}{(r-3)}$. The bias is corrected by constructing a new moment estimate $\widehat{N}$:

$$\widehat{N} = \frac{(r-3)}{(r-1)}\tilde{N} = \frac{(r-3)}{D^2}. \tag{16}$$

Here D^2 is the deviance of proportions given in equation (3)

A $100(1-\alpha)$% confidence interval for n can be constructed from the relationship of the estimate $\widehat{N}$ with a chi-squared distribution. We have

$$\widehat{N} = \frac{n(r-3)}{\mathrm{X}^2}, \tag{17}$$

where X^2 has an asymptotic chi-squared($r-1$) distribution. Thus $n(r-3)/\widehat{N}$ is a pivotal quantity. Given that

$$P\left(\chi^2_{1-\alpha/2} < \mathrm{X}^2 < \chi^2_{\alpha/2}\right) \approx 1-\alpha, \tag{18}$$

an approximate $100(1-\alpha)$% confidence interval for n can be written as

$$\left(\frac{\chi^2_{1-\alpha/2}\widehat{N}}{r-3}, \frac{\chi^2_{\alpha/2}\widehat{N}}{r-3}\right), \tag{19}$$

which is equivalent to

$$\left(\frac{\chi^2_{1-\alpha/2}}{D^2}, \frac{\chi^2_{\alpha/2}}{D^2}\right). \tag{20}$$

The smaller the deviance between the observed and expected proportions, the larger the estimate of sample size becomes.

2.8 Convergence of Chi-squared

One can estimate the minimum size of n needed for the confidence interval coverage to still hold true. Here the confidence interval coverage is tested by performing a large number of simulations and counting the proportion of confidence intervals that include n, and comparing that proportion to $1-\alpha$. Because the point and interval estimates of n

were derived from the chi-squared goodness-of-fit statistic, we might expect the statistical properties of the estimates to depend heavily on the adequacy of the chi-squared approximation. The conventional rules for the chi-squared statistic to asymptotically converge are for the expected counts $e_j = np_j \geq 5$ for at least 80% of the cells and $e_j \geq 1$ for all j [13]. Another common, more conservative, approach is setting $e_j \geq 5$ for all j [12]. This leads to the conclusion that for higher probabilities smaller sample size is needed. Keeping variation in $\tilde{p}_j$ constant, multinomial distributions with lower r possible outcomes must have greater $\tilde{p}_j$ for each j for all probabilities to sum up to 1, and therefore smaller n to satisfy the requirements.

2.9 Results

This simulation used true cell proportions p_j , $j = 1, 2, ..., r$ for $r = 6$ with values of $p' = \left[1/12, 2/12, 3/12, 2/12, 2/12, 2/12\right]$, leaving n to be estimated. The values of n were set to 10, 50, 90 and 130. For each value of n, 10^4 simulated samples $Y_1, Y_2, ..., Y_r$ were generated from a single multinomial distribution with probability mass function

$$\mathrm{P}\left(Y_1 = y_1, Y_2 = y_2, ..., Y_r = y_r\right) = \left[\frac{n!}{y_1!y_2!...y_r!}\right] p_1^{y_1} p_2^{y_2} ... p_r^{y_r} . \tag{21}$$

Each simulated sample observed empirical proportions $\tilde{p}_j$, $j = 1, 2, ..., r$. The true and empirical proportions were then used to calculate point estimates (biased and unbiased) and 95% confidence intervals according to the chi-squared-based results from sections 2.5 through 2.7. The value of Pearson’s goodness-of-fit statistic was also recorded, in order to assess the chi-squared approximation (Table 2.1). The chi-squared plot in Figure 2.1 is useful for detecting possible outliers. The chi-squared statistics that deviated from the linearity suggest that the statistic differs from the expected value. As n increases, the simulated statistics tend to be closer to expected value of corresponding quantile.

Table 2.1: Estimates of parameters and actual coverage of sample size from 10^4 iterations.

n	10	50	90	130
mean($\tilde{N}$)	15.33306	80.61423	148.25858	213.11055
Var($\tilde{N}$)	181.2788323	9205.8829309	34260.964470	90523.84834
mean($\widehat{N}$)	9.199837	48.368536	88.955150	127.866327
Var($\widehat{N}$)	65.2603796	3314.1178551	12333.94721	32588.58540
mean(X^2)	5.003680	5.041748	5.000736	5.030245
Var(X^2)	9.537804238	9.73983669	9.65257563	10.084054308
$\widehat{N}$ actual coverage of 95% CI (%)	93.81	95.48	95.31	95.10

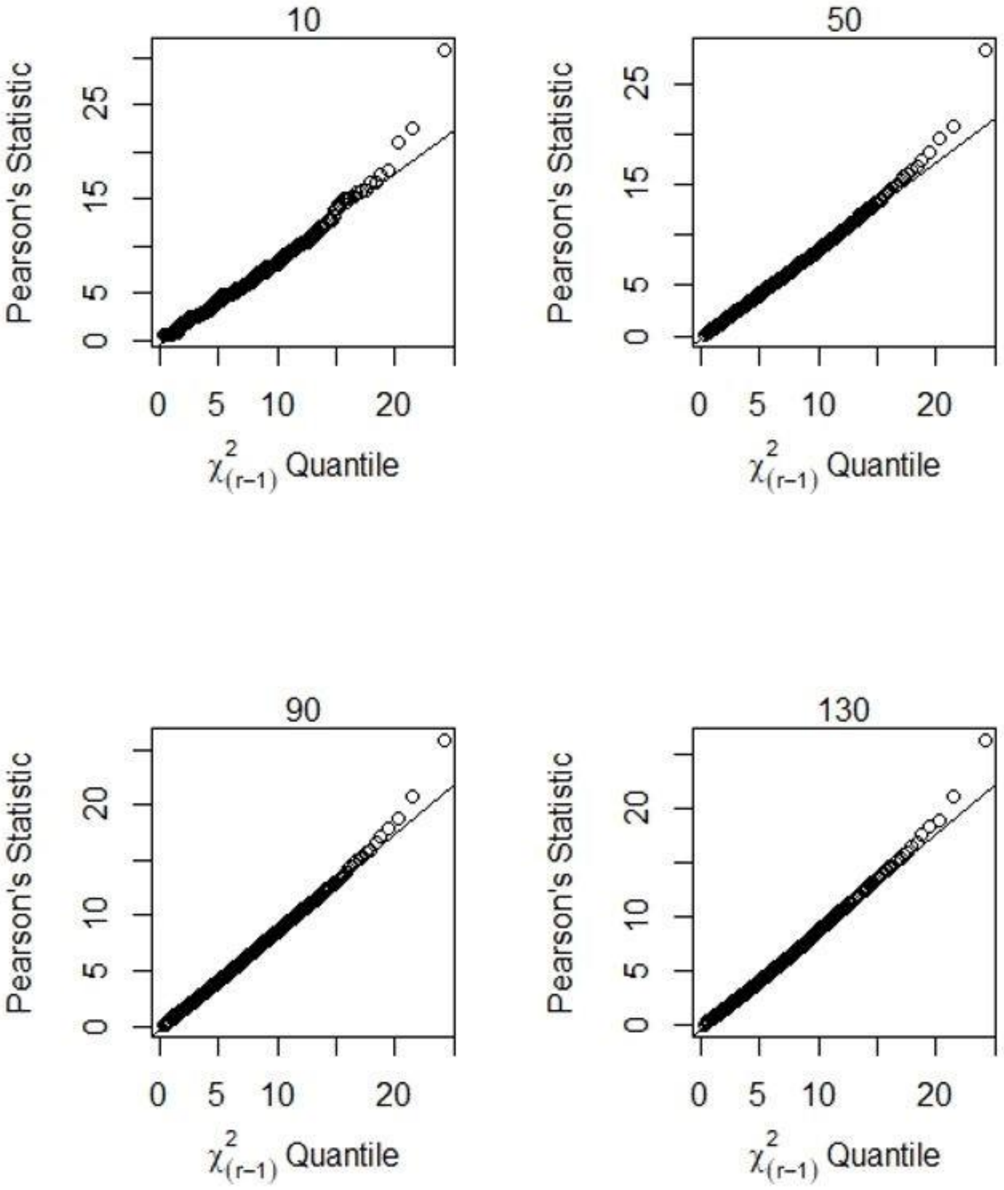

Figure 2.1: Single multinomial case. Sorted distances of Pearson’s chi-squared statistic for varying sample sizes over 10^4 iterations plotted against the quantiles of χ^2_{r-1} distribution.

The simulation results agree with the result derived earlier that $\tilde{N}$ is biased, and after the bias correction by multiplying $\tilde{N}$ with $\frac{r-3}{r-1}$, as in equation (16), the average of the 10^4 simulated values of the new estimate $\widehat{N}$ is much closer to n than that of $\tilde{N}$ (Table

2.1). The actual coverage of $100(1-\alpha)$% confidence interval at $\alpha = 0.05$ was overall closer to $100(1-\alpha)$% with $\widehat{N}$ than with $\tilde{N}$. The boxplots of scaled bias (the difference between the statistic and n, divided by n) for tested sample sizes over 10^4 iterations are shown in Figure 2.2 for $\tilde{N}$ and for $\widehat{N}$ in Figure 2.3. The dashed horizontal line corresponds to zero value on the y-axis of scaled bias. On the figures the bottom and the top of each of the three boxes are the 25th and 75th percentiles of simulated data, or Q1 and Q3 respectively, with the line segment near the middle of the box represents the 50th percentile- the median. The lower whisker is the Q1-1.5IQR, and the upper whisker is the Q3+1.5IQR, where IQR is the Q3-Q1 difference. In both cases the boxplot whiskers extending from 75th percentile mark were longer than the whiskers from the 25th percentile mark, low median, and many data points in the positive scaled bias direction, indicating distribution heavily skewed to the right. These results are consistent with the gamma distribution [14]. The Q3-Q1 range was smaller for $\widehat{N}$, indicating overall smaller scaled bias with $\widehat{N}$ than with $\tilde{N}$. The horizontal dashed line is centered closer to the center of boxes in Figure 2.2 than in Figure 2.3, indicating that $\widehat{N}$ is closer to true sample size than $\tilde{N}$ is.

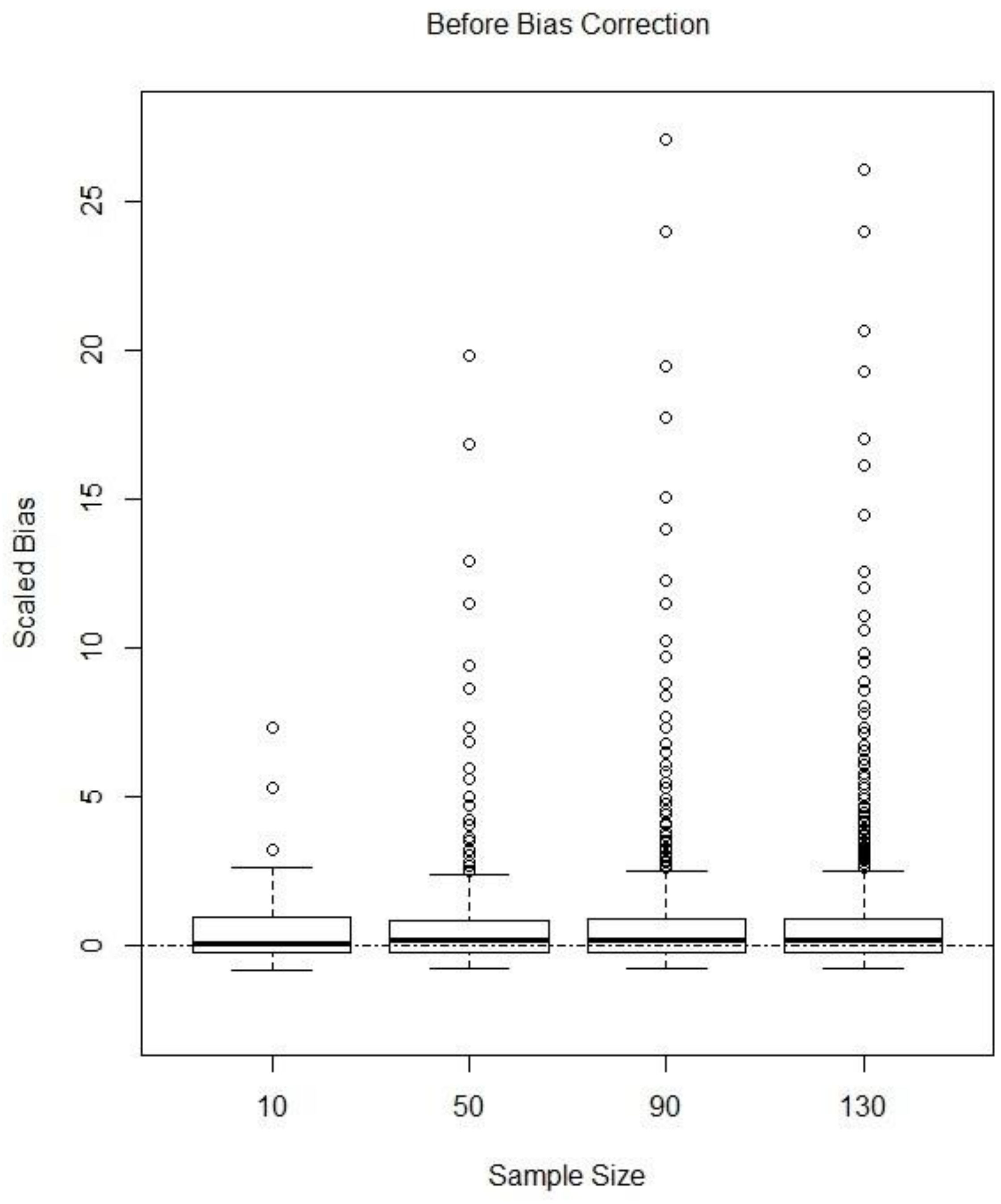

Figure 2.2: Single multinomial case. Boxplots of scaled difference $\frac{\tilde{N} - n}{n}$ for varying sample sizes over 10^4 iterations.

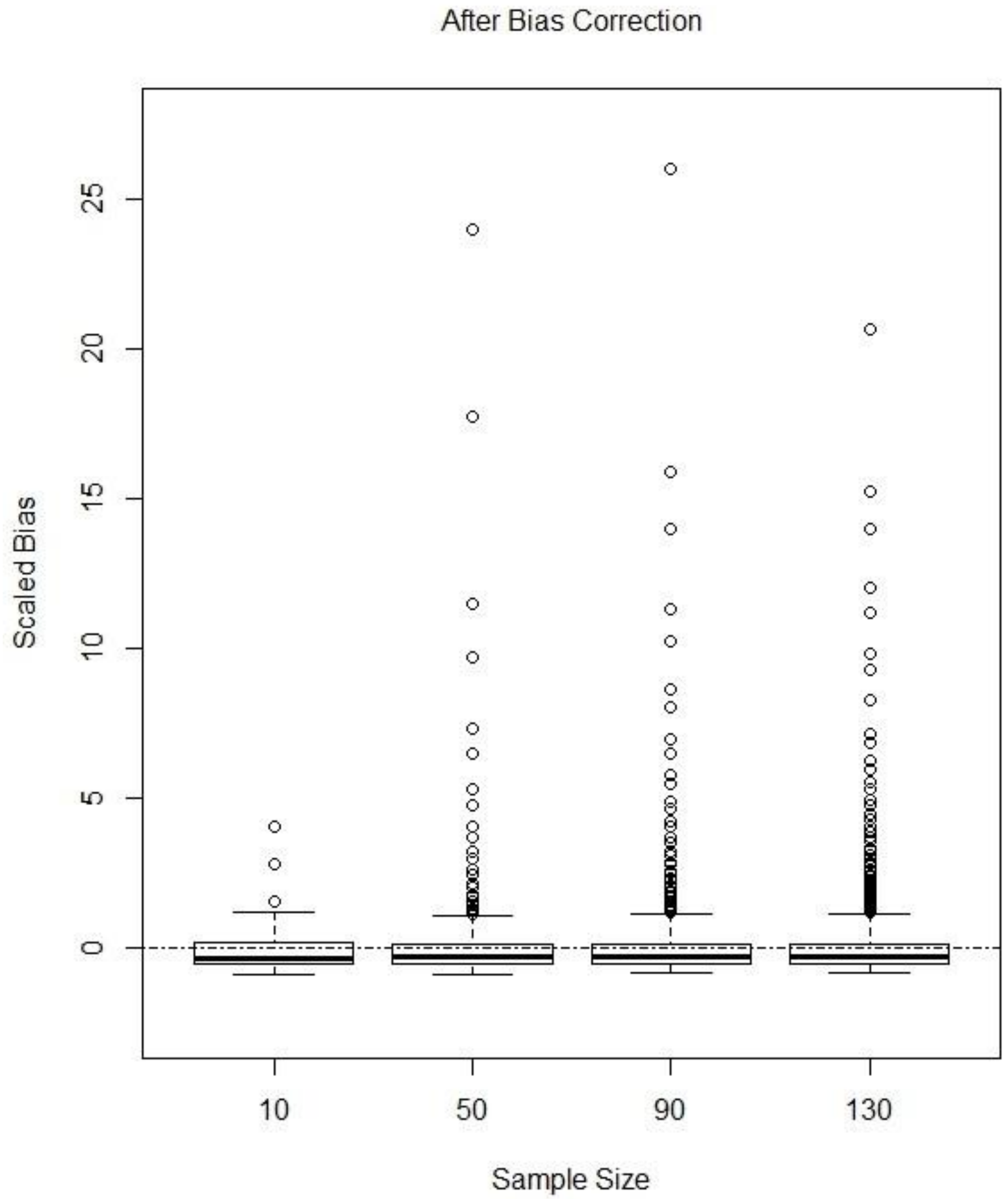

Figure 2.3: Single multinomial case. Boxplots of scaled difference $\frac{\widehat{N}-n}{n}$ for varying sample sizes over 10^4 iterations.

2.10 Conclusion and Discussion

Above was presented a method for estimating sample counts for a single multinomial model with r outcomes, when the proportions are known but the counts are missing. As shown in Section 2.9, the moment estimate corrected for bias $\widehat{N}$ converges quickly to the true value of n when known expected proportions $p_1, p_2, ..., p_r$ are non-sparse. When n is 50, approximately 57.14% of expected counts np_j are greater than 5, and the remaining expected counts being greater than 4. The results agree with the literature [13][12].

As shown in Table 2.1, an average of $\tilde{n}$ was 153.33%, 161.23%, 164.73%, and 163.93% for the n value for n of 10, 50, 90, and 130 respectively. The results are consistent with derived bias of n in equation (15). In the particular case of $r = 6$, the ratio is 1.67. For $k = r - 1$, the results of the mean value of X^2 reciprocal over 10^4 simulations are around $\frac{1}{k-2}$, and variance of around $\frac{2}{(k-2)^2(k-4)}$, the defined variance [15]. In same manner an expected variance of $\tilde{N}$, $n^2k^2\,\mathrm{Var}\left(\frac{1}{\mathrm{X}^2}\right)$, is $\frac{n^2k^2 2}{(k-2)^2(k-4)}$. As starting sample size n increases, the variance of $\tilde{N}$ over 10^4 iterations is closer to the expected variance of $\tilde{N}$.

The sample size estimation was obtained from using Pearson's X^2 statistic. The log-likelihood ratio statistic G^2 could be a possible direction of future work to explore a different approach for estimation of sample size n in the multinomial model [16].

Chapter 3

Dennis-Kemp Model with Same n for Each Time Interval

3.1 Purpose

This chapter implements a method for sample size n estimation described in previous chapter for use in forecasting plant development events in the agriculture of Blue Diamond® Growers Nonpareil almonds. The method is based on Dennis et al. [1] (known in the literature as the Dennis-Kemp model) which estimates the proportions of insects or plants in a given development stage as a function of time, measured in accumulated degree days. Typical stage development data are in the form of a two-way table counts, with the i^{th} row corresponding to a sample of size n_i at time t_i. The Dennis-Kemp model employs a product multinomial likelihood for count data, with the model proportions in each row (proportions of organisms in stage j at time t_i) being functions of time containing unknown parameters. In this chapter a method is proposed for estimation of sample size when only proportions are known for a product multinomial model with r development stages and q sampling times, as well as how to account for sparseness of contingency tables due to low or zero expected cell proportions.

3.2 Blue Diamond Almond Counts

The Blue Diamond® Growers keep track of development in almond orchards during each growing season for sampling times t_i, $i = 1,2,...,q$ in degree-days (DD) for r stages of tree development between dormancy and full bloom. A project was initiated under the USDA Agricultural Research Service to develop a phenology model for the almonds, in order to forecast the best time (10% bloom) to schedule honey bee placement for pollination. When the data were made available to USDA-ARS, the investigators became aware that the empirical proportions

$\tilde{p}_{ij} = y_{ij} / n_i$ for each sampling time t_i were recorded but not the counts nor the sample size.

Because the sampling protocol appeared to be standard, it was assumed that the sample size at each time t_i stayed the same, i.e.:

$$n_i = n \,. \tag{22}$$

3.3 Model Description

The proportion of population in development stage j , $j = 1, 2, ..., r$ at t_i is described by the Dennis-Kemp (D-K) model [1] as:

$$\begin{aligned} p_{ij} &= 1/\left\{1+\exp\left[-\left(\frac{a_j - t_i}{\sqrt{\nu t_i}}\right)\right]\right\} \text{ for } j = 1, \\ p_{ij} &= 1/\left\{1+\exp\left[-\left(\frac{a_j - t_i}{\sqrt{\nu t_i}}\right)\right]\right\} - 1/\left\{1+\exp\left[-\left(\frac{a_{j-1} - t_i}{\sqrt{\nu t_i}}\right)\right]\right\} \text{ for } j = 2, ..., r-1, \\ p_{ij} &= 1/\left\{1+\exp\left[-\left(\frac{a_j - t_i}{\sqrt{\nu t_i}}\right)\right]\right\} - 1/\left\{1+\exp\left[-\left(\frac{a_{j-1} - t_i}{\sqrt{\nu t_i}}\right)\right]\right\} \text{ for } j = r \,. \end{aligned} \tag{23}$$

The a_j represents amount of development in DD needed to undergo stage j , and ν the variability of development rates within the population. The model assumes the underlying development level of an organism to be a continuous mean-increasing stochastic process, with the organism entering a discernable stage j after attaining development level a_{j-1}.

The likelihood for the D-K model can be expressed as a product multinomial

$$\mathrm{L}(a_1, a_2, ..., a_{r-1}, \nu) = \prod_{i=1}^{q} \binom{n}{y_{i1} y_{i2} ... y_{ir}} p_{i1}^{\;y_{i1}} p_{i2}^{\;y_{i2}} ... p_{ir}^{\;y_{ir}} \,, \tag{24}$$

where individual multinomials for each t_i are assumed to be independent [13]. Here the likelihood is written under the assumption of the same sample size for each sampling time. Numerical maximization is required to obtain maximum likelihood (ML) estimates of the unknown parameters:

$$\theta = \left[a_1, a_2, ..., a_{r-1}, \nu\right]' . \tag{25}$$

A computationally more stable technique is to maximize the log-likelihood

$$\log\mathrm{L}(\theta) = \sum_{i=1}^{q} \binom{n}{y_{i1} y_{i2} ... y_{ir}} + \sum_{i=1}^{q} \sum_{j=1}^{r} y_{ij} \log p_{ij} \,. \tag{26}$$

The sample sizes n are not needed for maximization of θ when rewriting the equation as:

$$\log\mathrm{L}(\theta) = \sum_{i=1}^{q} \binom{n}{y_{i1} y_{i2} ... y_{ir}} + n \sum_{i=1}^{q} \sum_{j=1}^{r} \tilde{p}_{ij} \log p_{ij} \tag{27}$$

and noting that the first term is a constant. The $\tilde{p}_{ij}$ corresponds to observed proportion of almonds that are at stage j for a given time point. The (log) likelihood is maximized when the double sum is maximized. Estimation of n was done on data of Blue Diamond Nonpareil type form year 2005 growing season. The data had recorded empirical proportions $\tilde{p}_{ij}$ but missing counts y_{ij}. The data with $q = 18$ sampling times originally consisted of $r = 7$ development stages; Dormant, Green Tip, Pink Bud, Popcorn, Bloom, Petal Fall, Jacket. For all of the recorded sampling times the observed proportion $\tilde{p}_{i1} = 0$ for development stage $j = 1$, corresponding to Dormant. The Dormant stage was excluded from the dataset, reducing the total number of stages of development to $r = 6$.

The estimates of expected proportions $\hat{p}_{i1}, \hat{p}_{i2}, ..., \hat{p}_{ir}$, $i = 1, 2, ..., q$ were derived from D-K model from equation (23), where the estimated proportions are a function of ML estimates $\left[\hat{a}_1, \hat{a}_2, ..., \hat{a}_{r-1}, \hat{v}\right]' = \left[695.593861, 769.754086, 816.238611, 919.064476, 952.973931, 1.086056\right]'$. The comparison of the fitted estimates of expected proportions as a function of Degree-Days from Dennis-Kemp model against the observed proportions are shown in Figure 3.1.

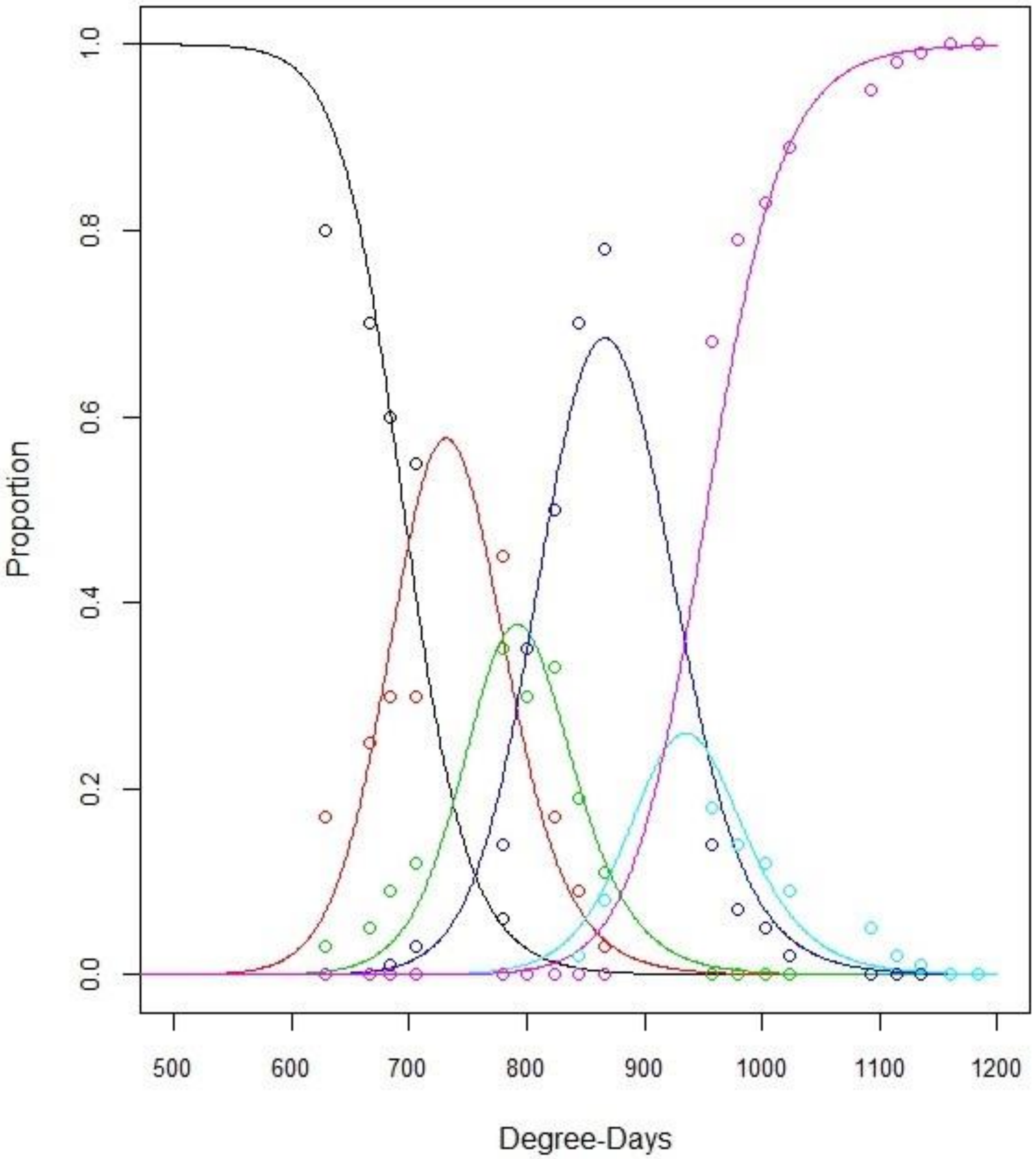

Figure 3.1: Line plots of expected proportions of almonds in stages 1-6 as a function of time (Degree-Days). Colored circles correspond to observed proportions from the data.

3.4 Variability of ML Estimates Depends on n

The maximization of the log-likelihood does not depend on sample size, however n does affect the variability of the ML estimates of parameters in θ. For sufficiently large samples ML estimates $\hat{\theta}$ follow a multivariate normal distribution with mean vector θ and variance-covariance matrix given by [13]

$$\mathrm{V}(\theta) = \left[\mathrm{I}(\theta)\right]^{-1}. \tag{28}$$

$\mathrm{V}(\theta)$ is inversely proportional to n. Here $\mathrm{I}(\theta)$ stands for the Fisher information matrix with k rows and l columns defined by

$$\mathrm{I}\left[\theta\right] = -\mathrm{E}\left[\frac{\partial^2 \log \mathrm{L}(\theta)}{\partial\theta_k \partial\theta_l}\right], \tag{29}$$

which for product multinomial models becomes

$$\mathrm{I}\left[\theta\right] = n\sum_{i=1}^{q}\sum_{j=1}^{r}\frac{1}{p_j(t_i)}\left[\frac{\partial p_j(t_i)}{\partial\theta_k}\right]\left[\frac{\partial p_j(t_i)}{\partial\theta_l}\right], \tag{30}$$

the second derivative corresponding to the curvature of log-likelihood. The variance-covariance matrix $\mathrm{V}(\theta)$ can be estimated with Hessian matrix:

$$\mathrm{J}(\hat{\theta}) = \frac{\partial^2 \log \mathrm{L}(\hat{\theta})}{\partial\hat{\theta}_k \partial\hat{\theta}_l} \tag{31}$$

[12] with $r \times r$ dimension for r parameters . The $100(1-\alpha)$ % Wald confidence interval for θ_j is:

$$\hat{\theta}_j \pm \mathrm{z}_{\alpha/2}\sqrt{\hat{\mathrm{v}}_{jj}}\ , \tag{32}$$

where $\hat{\mathrm{v}}_{jj}$ is the j^th^ element on the main diagonal of $\mathrm{V}(\hat{\theta})$.

3.5 Estimation of n from Pearson Goodness-of-Fit Statistic

The sample size needed for confidence intervals of D-K model ML is estimated with Pearson goodness-of-fit statistic. With r parameters and q sampling times, the model fits the data if:

$$\mathrm{X}^2 = n\left\{ \sum_{i=1}^{q} \sum_{j=1}^{r} \frac{\left[\tilde{p}_j(t_i) - \hat{p}_j(t_i)\right]^2}{\hat{p}_j(t_i)} \right\}, \tag{33}$$

where $\hat{p}_{ij}$ are estimates of p_{ij} from D-K model.

The observed proportions $\tilde{p}_{ij}$ are obtained from large number of simulations of q multinomial distributions with specified n and $\hat{p}_{ij}$. The expected value of X^2 for fitted model is the degrees of freedom of $q(r-1)$.

The reciprocal of a gamma-distributed random variable is derived in a similar manner as for a single multinomial case, but with different shape parameter. From X^2 ~chi-squared($q(r-1)$), we obtain the moment estimate

$$\tilde{N} = \frac{nq(r-1)}{\mathrm{X}^2} = \frac{q(r-1)}{D^2}. \tag{34}$$

3.6 Bias of Moment Estimate of n

As in Chapter 2 for a single multinomial, $\tilde{N}$ is biased. The expected value of $\tilde{N}$ is

$$\mathrm{E}\left(\tilde{N}\right) = \mathrm{E}\left(\frac{1}{Y}\right) = \frac{n\left[q(r-1)\right]}{q(r-1)-2}, \tag{35}$$

where $Y \sim \text{gamma}\,(\alpha, c\beta)$, $\alpha = \frac{q(r-1)}{2}$, $\beta = 1/2$, and $c = nq(r-1)$.

The estimate of n is corrected for bias:

$$\widehat{N} = \frac{q\left(r-1\right)-2}{q(r-1)}\tilde{N} = \frac{q\left(r-1\right)-2}{D^2}. \tag{36}$$

For the model with same r number of parameters, as the number of sampling times- q - increases, the bias of moment estimate decreases. For example, with a single multinomial $(q=1)$ from previous chapter , with $r=6$ the expected value is 66.67% too high, where with twenty sampling times the sample size is overestimated only by 2.27%.

3.7 Confidence Interval

The $100(1-\alpha)$ % confidence interval for sample size unbiased estimate for the product multinomial sampling model of q sampling times and r proportions per single multinomial is found similarly to as in Chapter 2, by finding a pivotal quantity based on the Pearson chi-squared statistic. Noting that

$$\mathrm{P}\left(\chi^2_{1-\alpha/2} < \mathrm{X}^2 < \chi^2_{\alpha/2}\right) \approx 1-\alpha \tag{37}$$

and that

$$\mathrm{X}^2 = \frac{n\left[q(r-1)-2\right]}{\widehat{N}} \tag{38}$$

is a pivotal quantity, an approximate $100\left(1-\alpha\right)$% confidence interval for n becomes

$$\left(\frac{\chi^2_{1-\alpha/2}\widehat{N}}{q(r-1)-2}, \frac{\chi^2_{\alpha/2}\widehat{N}}{q(r-1)-2}\right). \tag{39}$$

The equation (39) is equivalent to equation (20). The confidence interval is expected to capture n approximately $100(1-\alpha)$ % of occasions under hypothetical repeated sampling.

3.8 Convergence of n for q Sampling Times

Convergence rule for a chi-squared statistic for product multinomial is same as for a single multinomial case; $np_{ij} \geq 1$ for all j and $np_{ij} \geq 5$ for at least 80% of the cells or more rigorous $np_{ij} \geq 5$ for all j. The rule applies to each sampling time t_i separately under the assumption that multinomial distributions for different times are assumed to be independent.

3.9 Low Expected Counts

In the Blue Diamond® Nonpareil almond case, the data contains zeroes for some of the development stages per t_i. The maximization of $\log\mathrm{L}(\theta)$ equation for estimation of In D-K model [1] where the estimated multinomial probabilities $\hat{p}_{i1}, \hat{p}_{i2}, ..., \hat{p}_{ir}$ per t_i, $i = 1, 2, ..., q$ found by maximization of vector θ have their lower bound set to equal to no less than a chosen constant (10^{-6}). This allows for simulation of potentially non-zero observed proportions $\tilde{p}_{i1}, \tilde{p}_{i2}, ..., \tilde{p}_{ir}$ for $t_1, t_2, ..., t_q$. The simulated $\tilde{p}_{ij}$ for which a probability of random draw from multinomial distribution is set to minimum value will depend on that arbitrarily chosen lowest acceptable probability, leading to extremely low value of that $\tilde{p}_{ij}$. A few low or zero counts can strongly bias sample size estimation. Pooling low expected probabilities $\hat{p}_{i1}, \hat{p}_{i2}, ..., \hat{p}_{ir}$ together as a single stage [13] can be an alternative.

In the pooling case, instead of X^2 holding $q(r-1)$ degrees of freedom, each row of counts per t_i will contribute $r_i - 1$ degrees of freedom, with r_i being not necessarily the same for $i = 1, 2, ..., q$. Summation of $r_i - 1$ over q sampling times yields new degrees of freedom: $\sum_{i=1}^{q} r_i - q$, distributed with X_k^2,

$$k = \sum_{i=1}^{q} r_i - q. \tag{40}$$

The X^2 statistic is derived in similar manner as for a single multinomial and product multinomial case, from large number of simulations of $\mathrm{X}^2 = nD^2$ from equation (3), but the deviance statistic is equal to:

$$D^2 = \sum_{j=1}^{r_1} \frac{\left[\tilde{p}_{ij} - \hat{p}_{ij}\right]^2}{\hat{p}_{ij}} + \sum_{j=1}^{r_2} \frac{\left[\tilde{p}_{ij} - \hat{p}_{ij}\right]^2}{\hat{p}_{ij}} + \ldots + \sum_{j=1}^{r_q} \frac{\left[\tilde{p}_{ij} - \hat{p}_{ij}\right]^2}{\hat{p}_{ij}} \tag{41}$$

due to potentially different degrees of freedom per row.

The empirical sample size is derived by setting $\mathrm{E}(\mathrm{X}^2)$ to its expected degrees of freedom divided by the deviance statistic:

$$\tilde{N} = \left(\sum_{i=1}^{q} r_i - q \right) / D^2 . \tag{42}$$

The expected value of $\tilde{N}$ is:

$$\mathrm{E}\left(\tilde{N}\right) = \mathrm{E}\left(\frac{1}{Y}\right) = \frac{n\left(\sum_{i=1}^{q} r_i - q\right)}{\sum_{i=1}^{q} r_i - q - 2} \tag{43}$$

with $c = n\left(\sum_{i=1}^{q} r_i - q\right)$, $\alpha = \left(\sum_{i=1}^{q} r_i - q\right)/2$ and $\beta = 1/2$ for $Y \sim \text{gamma}(\alpha, c\beta)$. The moment estimate of sample size n is corrected for bias by setting to a new estimate:

$$\widehat{N} = \frac{\sum_{i=1}^{q} r_i - q - 2}{\sum_{i=1}^{q} r_i - q} \tilde{N} = \left(\sum_{i=1}^{q} r_i - q - 2 \right) / D^2 . \tag{44}$$

The approximate $100(1-\alpha)\%$ confidence interval for n for the pooling case is:

$$\left(\frac{\chi^2_{1-\alpha/2}\widehat{N}}{\sum_{i=1}^{q} r_i - q - 2}, \ \frac{\chi^2_{\alpha/2}\widehat{N}}{\sum_{i=1}^{q} r_i - q - 2} \right) \tag{45}$$

with pivotal quantity of $n\left(\sum_{i=1}^{q} r_i - q - 2 \right) / \widehat{N}$.

3.10 Results

Following are presented the results of sample size estimation method with and without pooling the cells with low expected probabilities from the D-K model. The model assumed $\hat{p}_{ij}$ estimated from D-K model to be the true expected proportions. The Nonpareil almond dataset from year 2005 consisted of eighteen sampling times t_i , $i = 1, 2, ..., q$, i.e. $q = 18$. The starting sample sizes were set to 50, 150, 250 and 500 before pooling, and 50, 250, 750, 1000 with pooling. The parameters $\tilde{N}$, $\widehat{N}$ and Pearson X^2 as well as the coverages of $\tilde{N}$ and $\widehat{N}$ were estimated from 10^4 iterations.

For the first case, the sample size estimation technique was evaluated for the unpooled model where the cell count per time point was kept constant at six for $r = 6$ stages of development of almond. The observed empirical proportions were simulated from a product multinomial:

$$\prod_{i=1}^{q} \left[\left(\frac{n_i\,!}{y_{i1}\,!\,y_{i2}\,!...y_{ir}\,!} \right) \hat{p}_{i1}{}^{y_{i1}}\, \hat{p}_{i2}{}^{y_{i2}} ...\hat{p}_{ir}{}^{y_{ir}} \right] \tag{46}$$

The Pearson X^2 from simulations was set to $q(r-1)$ the expected degrees of freedom. With 18 sampling times and 6 stages of development, E(X^2) being equal to 90. The first moment of n from equation (34) and set to new, unbiased moment estimate based on equation (36).

The second case involved pooling cells in the table with expected proportions $\hat{p}_{ij}$, so that each row that corresponded to specified t_i could potentially have different number of cells r_i. For the particular dataset cells with $\hat{p}_{ij} < 0.0035$ were combined with adjacent $\hat{p}_{ij}$ cells, except for last three sampling times, $i = 16$, $i = 17$ and $i = 18$, where cells were pooled with adjacent cell $\hat{p}_{i5} = 0.003488668$ for i=16, $\hat{p}_{i5} = 0.0017957$ for i=17, and $\hat{p}_{i5} = 0.0009966204$ for i=18 respectively, so that the table of pooled expected proportions had at least two cells per sampling time t_i.

Sorted distances of Pearson's chi-squared statistic from10^4 iterations were plotted against the quantiles of chi-squared distribution, $\chi^2_{q(r-1)}$ for first case (Figure 3.2), and after pooling χ^2_k , $k = \sum_{i=1}^{q} r_i - q$, $k = 51$ for second case. After pooling the chi-squared statistic approached the expected value corresponding of corresponding quantile (Figure 3.5)

Before pooling the variability in chi-squared statistic was large. For the first case with $q(r-1)$ equal to 90, the average of Pearson's X^2 statistic was close to expected value, (Table 3.1). With low coverages of sample size estimates- both $\tilde{N}$ and $\hat{N}$ - that were based only on Pearson's X^2 - the coverage failed to capture 95% of true n due to high variance of Pearson's X^2 statistic. High variance indicates that the simulated data do not fit the chi-squared distribution.

Table 3.1: Estimates of parameters and actual coverage of sample size from 10^4 iterations for Nonpareil almonds from year 2005 before pooling.

n	50	150	250	500
mean($\tilde{N}$)	73.02937	197.06024	312.37299	595.23978
Var($\tilde{N}$)	758.75875	4832.83463	11036.74841	35686.8419
mean($\hat{N}$)	71.4065	192.6811	305.4314	582.0122
Var($\hat{N}$)	725.4108324	4620.42856	10551.6765037	34118.383215
mean(X^2)	90.14077	89.60789	90.89638	89.46268
Var(X^2)	71440.731571	17025.323336	12380.91808	4011.53394
$\hat{N}$ actual coverage of 95% CI (%)	33.13	41.18	48.50	56.78

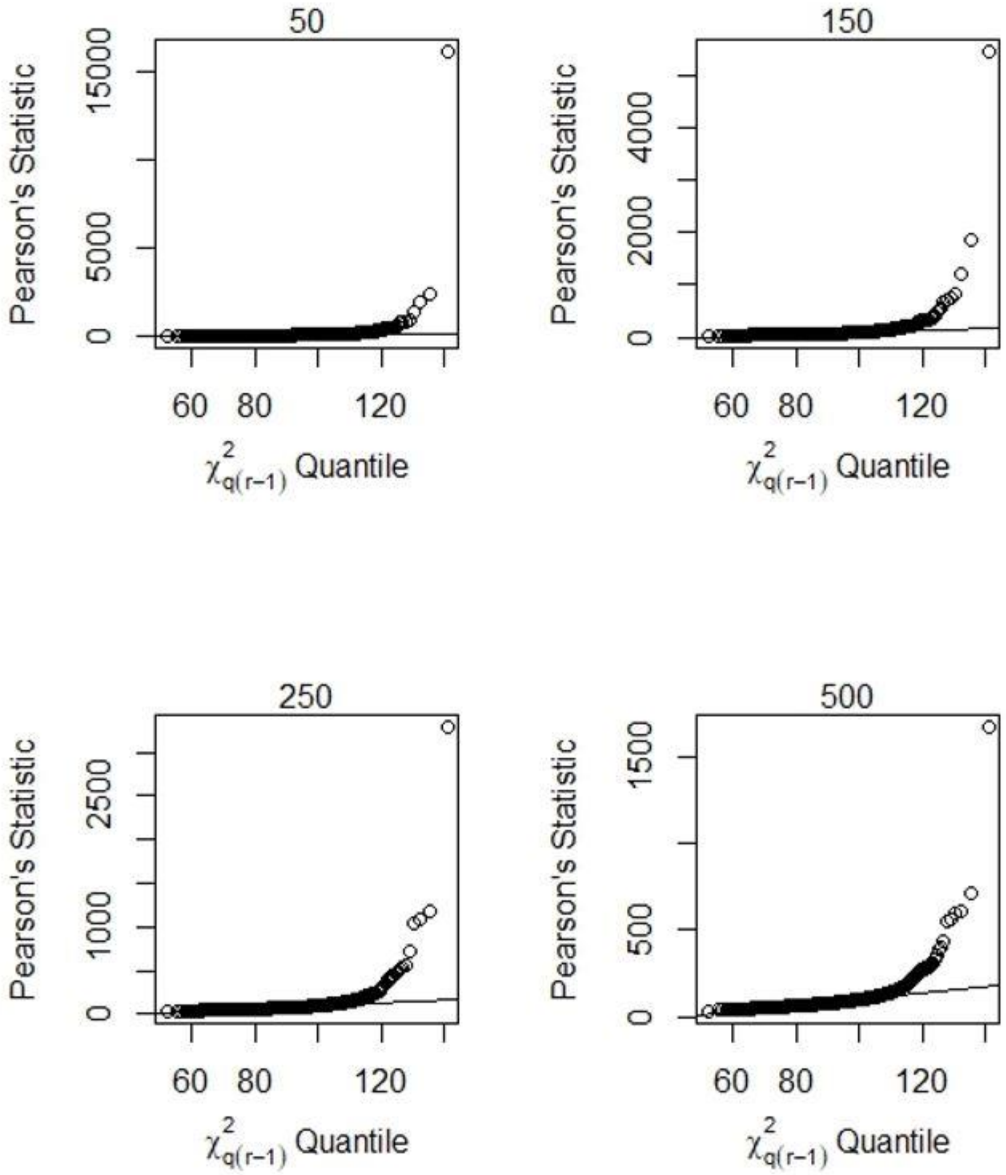

Figure 3.2: Product multinomial before pooling case. Sorted distances of Pearson's chi-squared statistic for varying sample sizes over 10^4 iterations plotted against the quantiles of $\chi^2_{q(r-1)}$ distribution.

The boxplots of scaled bias over 10^4 iterations for the unpooled case before bias correction (Figure 3.3) and after bias correction (Figure 3.4) look similar. The dashed horizontal line is a value of zero on the y-axis corresponding to *Scaled Bias*. As described in section 3.6, the bias of moment estimate $\tilde{N}$ is expected to be much smaller for product multinomial than for a single multinomial. For example, for $n = 500$ with $q(r-1)$ of 90, $\tilde{N}$ is larger from $\hat{N}$ only by approximately 2.27% (Table 3.1). The observed mean of $\tilde{N}$ and of $\hat{N}$ were greater than expected and the coverage was low before and after bias correction. The mean of the X^2 statistic was close to its expected value of k, where k stands for degrees of freedom. However, the variance of X^2 was larger than $2k$ [15].

The boxplot whiskers extending from 75th percentile mark were not much longer than the whiskers from the 25th percentile mark, which is inconsistent with the skewness of gamma distribution [14].

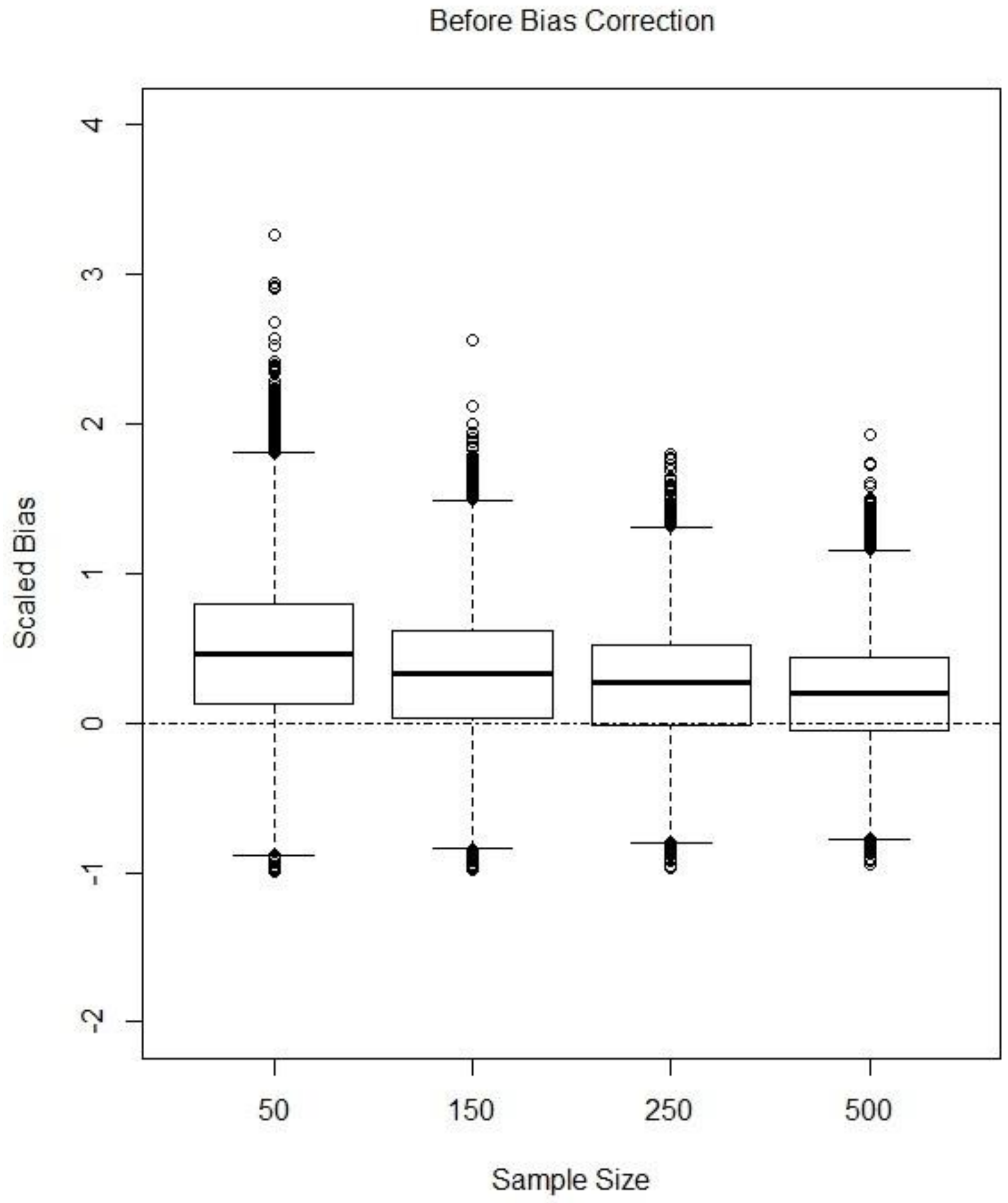

Figure 3.3: Boxplots of scaled difference $\frac{\tilde{N}-n}{n}$ for varying sample sizes over 10^4 iterations for unpooled case.

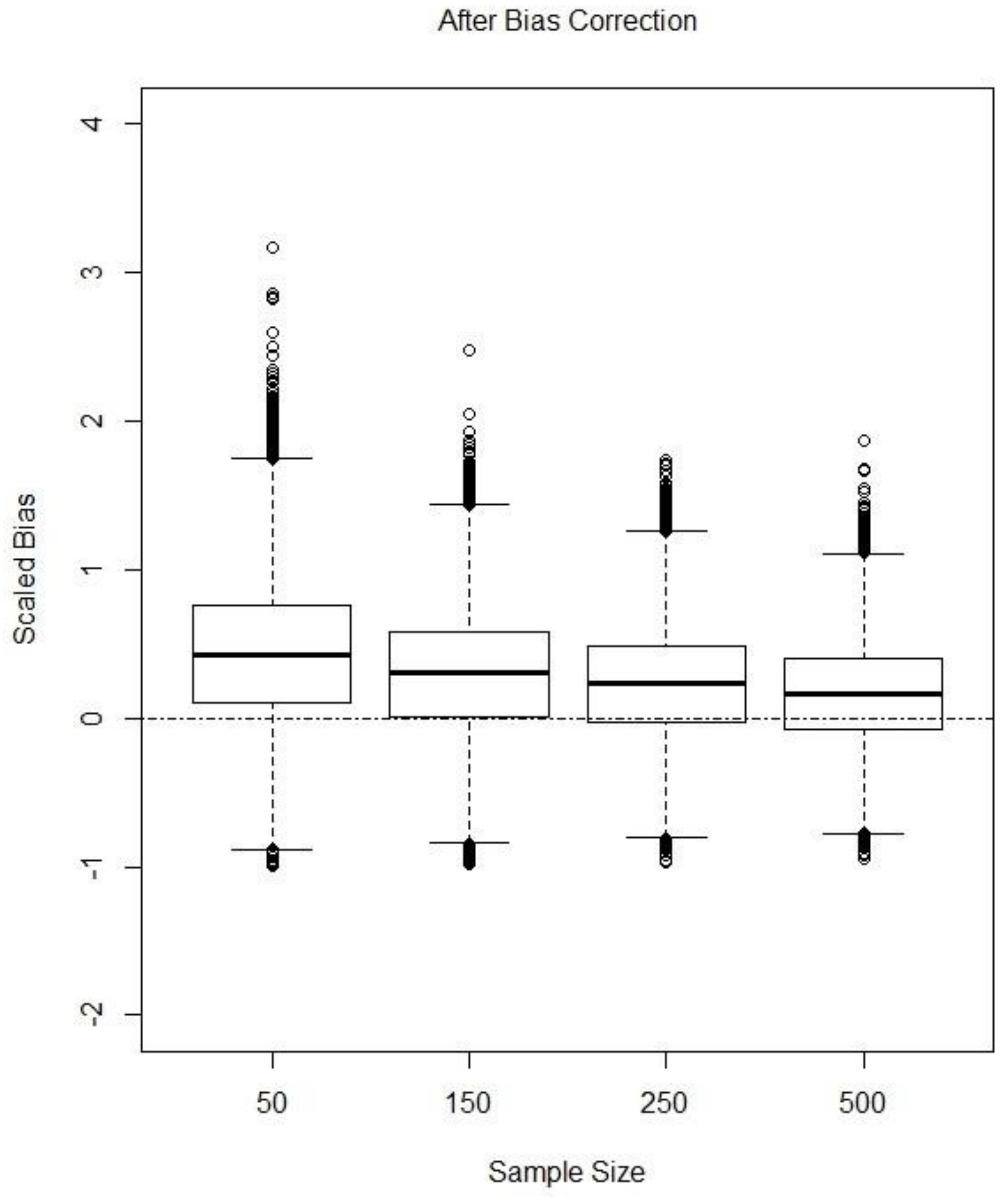

Figure 3.4: Boxplots of scaled difference $\frac{\widehat{N}-n}{n}$ for varying sample sizes over 10^4 iterations for unpooled case.

For comparison of how the summary statistics would look like with sample size sufficiently large, the unpooled case was re-ran separately with starting sample size n of 5 $\times 10^6$. The results are shown in Table 3.2.

Table 3.2: Estimates of parameters and actual coverage of sample size from 10^4 iterations for starting sample size of 5×10^6 for Nonpareil almonds data from year 2005 before pooling.

n	5×10^6
mean($\tilde{N}$)	$5{,}113{,}370\times 10^6$
Var($\tilde{N}$)	6.03916×10^{11}
mean($\hat{N}$)	4.999740×10^6
Var($\hat{N}$)	5.77374×10^{11}
mean(X^2)	90.00389
Var(X^2)	180.6244
$\hat{N}$ actual coverage of 95% CI (%)	95.11

According to Bernardo and Smith [15], an expected variance of the inverse chi-squared distribution is $\frac{2}{(k-2)^2(k-4)}$, or equivalently 3.003075×10^{-6} for k=90. From the relationship of expected value of $\tilde{N}$ defined in equation (43), or rewritten as:

$$\mathrm{E}(\tilde{N}) = kn\mathrm{E}\left(\frac{1}{\mathrm{X}^2}\right), \tag{47}$$

The variance of $\tilde{N}$ is defined as:

$$\mathrm{Var}(\tilde{N}) = \frac{n^2k^2 2}{(k-2)^2(k-4)} \tag{48}$$

(6.08123×10^{11} for k=90). The results in Table 3.2 are consistent with the literature. Both the estimates and the variances of estimates match with corresponding parameters and its variances of the chi-squared distribution.

The results after pooling technique described earlier in this section are presented in Table 3.3. For sample size n of 1000, the coverage and Var(X^2) is close to 95% and to defined variance of:

$$\mathrm{Var}(X^2) = 2k, \tag{49}$$

respectively (Table 3.3).

Table 3.3: Estimates of parameters and actual coverage of sample size from 10^4 iterations for Nonpareil almonds from year 2005 after pooling.

n	50	250	750	1000
mean($\tilde{N}$)	53.0655	261.2525	783.9797	1040.1732
Var($\tilde{N}$)	172.781468	3114.690664	26980.26255	47067.287790
mean($\hat{N}$)	50.9845	251.0073	753.2354	999.3821
Var($\hat{N}$)	159.495696	2875.191189	24905.65566	43448.119179
mean(X^2)	51.06640	50.98479	50.84372	51.08287
Var(X^2)	168.5847550	114.0892897	104.7079488	105.6327902
$\tilde{N}$ actual coverage of 95% CI (%)	88.60	93.80	94.26	94.63
$\hat{N}$ actual coverage of 95% CI (%)	88.95	94.04	94.66	94.64

For the pooling case, cells with $\hat{p}_{ij} < 0.0035$ were combined with adjacent $\hat{p}_{ij}$ cells, except for last three sampling times i =16, i =17, and i =18, where cells were pooled with adjacent cell $\hat{p}_{i5} = 0.003488668$, $\hat{p}_{i5} = 0.0017957$, and $\hat{p}_{i5} = 0.0009966204$ respectively, so that the table of pooled expected proportions had at least two cells per sampling time, with exceptions to sampling times i =16, i =17, and i =18.

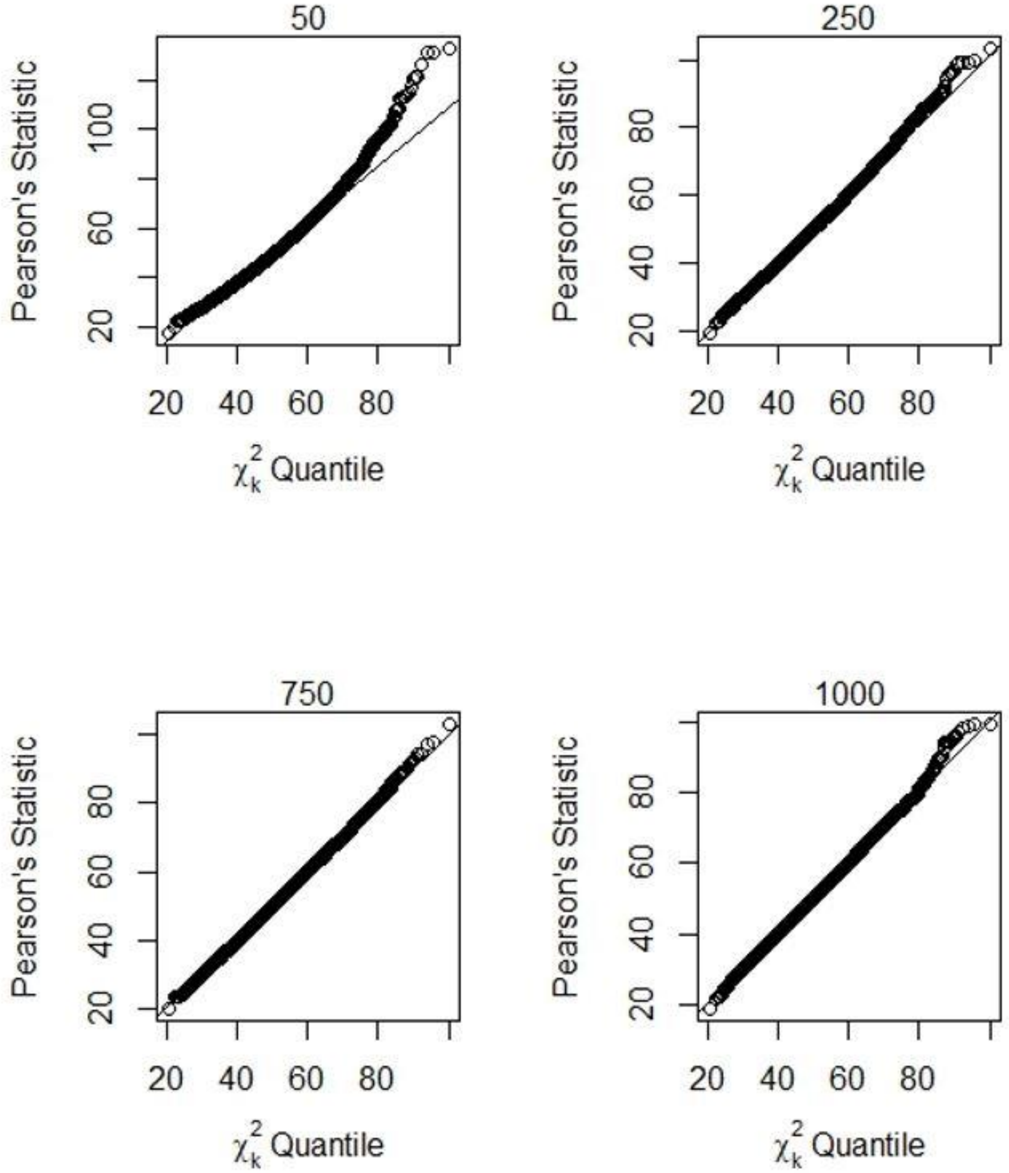

Figure 3.5: Product multinomial after pooling case. Sorted distances of Pearson's chi-squared statistic for varying sample sizes over 10^4 iterations plotted against the quantiles of χ^2_k distribution, $k = \sum_{i=1}^{q} r_i - q$.

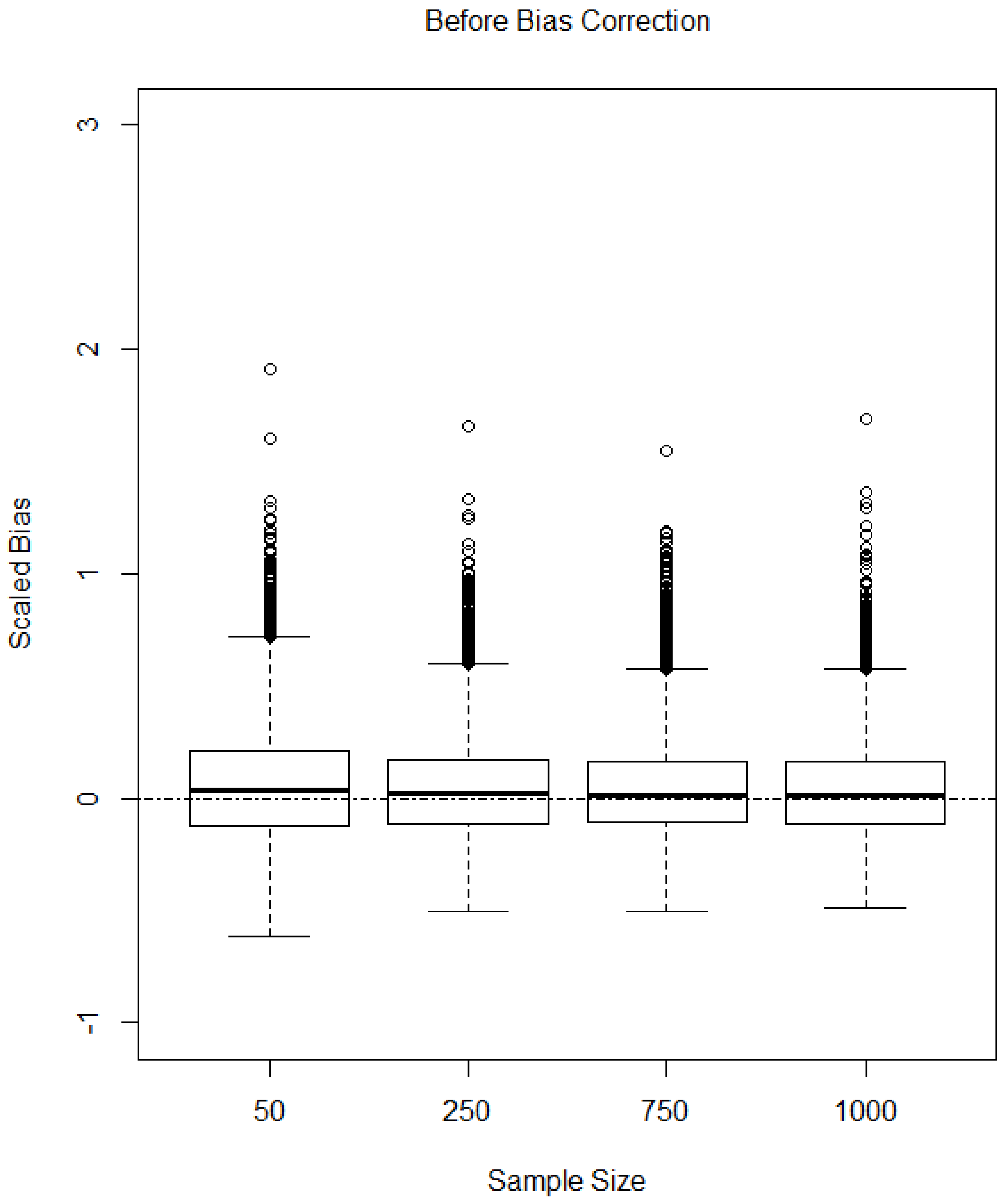

Figure 3.6: Boxplots of scaled difference $\frac{\tilde{N}-n}{n}$ for varying sample sizes over 10^4 iterations for pooled case.

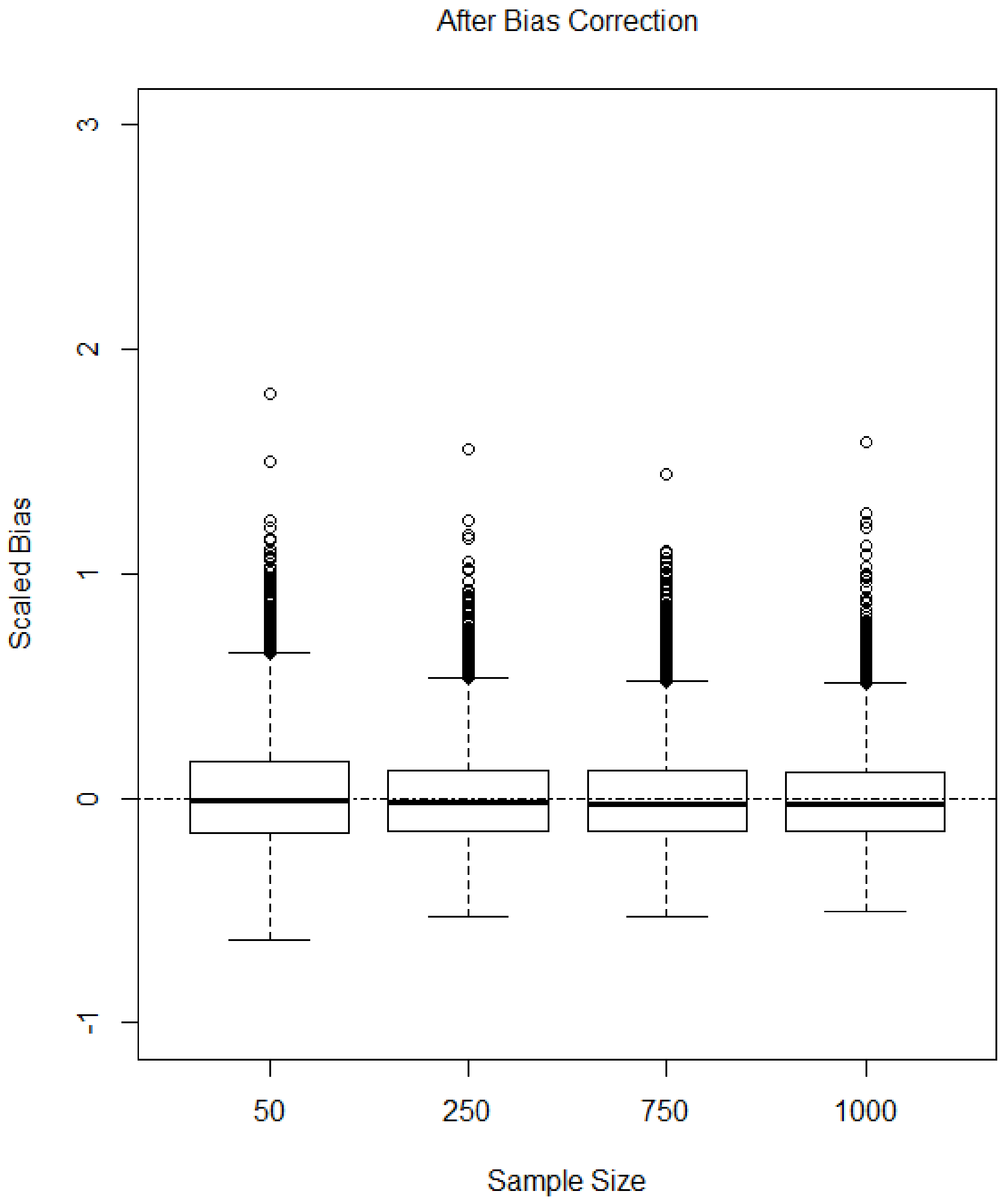

Figure 3.7: Boxplots of scaled differences $\frac{\widehat{N}-n}{n}$ for varying sample sizes over 10^4 iterations for pooled case.

3.11 Conclusion and Discussion

In general, the proposed method of pooling is an alternative solution to sample size estimation of multinomial sampling model with known expected proportions.

The minimum proposed expected count $n\hat{p}_{ij}$ per row is 3.5, allowing for few exceptions to the rule. The particular dataset had exceptions to the rule 16.67% of the times. It is recommended that the rule for pooling is not violated more than the tested limit. The proposed method is an improvement over existing technique [17] that does not allow exceptions to the pooling. It is also more relevant than the method proposed by Otis et al. [7] for mark-recapture in closed populations.

The developed model is an extension from Dennis-Kemp model [1] in which the maximization of parameter estimates does not depend on the sample size, however their variability does. The expected proportions of almond data are a function of time, and the implementation of developed sample size estimation with previously developed model can be applied to future phenology data.

For Blue Diamond almond data, an assumption of constant sample size over sampling times is adequate but the model cannot be applied to animal datasets or other plant data with changing population size over time. A thorough knowledge of studied population is needed.

References

[1] B. Dennis, W. Kemp, and R. Beckwith, “Stochastic Model of Insect Phenology : Estimation and Testing,” vol. 546, pp. 540–546, 1986.

[2] “World Agricultural Supply and Demand Estimates,” 2018.

[3] W. P. Kemp, B. Dennis, and R. C. Beckwith, “Stochastic Phenology Model for the Western Spruce Budworm (Lepidoptera: Tortricidae),” *Environ. Entomol.*, vol. 15, no. 3, pp. 547–554, 1986.

[4] Y. Bishop, S. Fienberg, and P. Holland, *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: M.I.T. Press, 1975.

[5] P. Eichenberger, J. Potterat, and B. Hulliger, “Two Measures for Sample Size Determination,” vol. 5, no. 1, pp. 27–37, 2011.

[6] S. K. Thompson, “Sample Size for Estimating Multinomial Proportions,” vol. 41, no. 1, pp. 42–46, 1987.

[7] D. L. Otis, K. P. Burnham, G. C. White, and D. R. Anderson, “Statistical Inference from Capture Data on Closed Animal Populations,” *Wildl. Monogr.*, no. 62, pp. 3–135, 1978.

[8] “Bloom/Harvest Reports.” [Online]. Available: http://www.bdingredients.com/category/almond-bloom-harvest-reports/. [Accessed: 13-May-2018].

[9] S. S. Wilks, “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses,” *Ann. Math. Stat.*, vol. 9, no. 1, pp. 60–62, 1938.

[10] J. Rice, *Mathematical Statistics and Data Analysis*, 3rd ed. 2007.

[11] T. R. C. Read and N. A. C. Cressie, *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer New York, 1988.

[12] F. Samaniego, *Stochastic Modeling and Mathematical Statistics*. 2014.

[13] A. Tamhane and D. Dunlop, *Statistics and Data Analysis*. 2000.

[14] R. Hogg, J. McKean, and A. Craig, *Introduction to Mathematical Statistics*, 6th ed. 2005.

[15] J. M. Bernardo and A. F. M. Smith, *Bayesian theory*. Wiley, 2000.

[16] J. Neyman, "Contribution to the Theory of the Chi-Square Test," 1949.

[17] J. L. Bresnahan and M. M. Shapiro, "A general equation and technique for the exact partitioning of chi-square contingency tables," *Psychol. Bull.*, vol. 66, no. 4, pp. 252–262, 1966.

Appendix 1

This appendix contains derivation of unbiased estimate of sample size n in a multinomial model. The chi-squared distribution with k degrees of freedom is denoted by X_k^2.

$k = r - 1$ for a single multinomial with r cell probabilities with $\sum_{j=1}^{r} p_j = 1$ constraint,

$k = q(r-1)$ for a product multinomial (q independent rows) and $\sum_{j=1}^{r} p_{ij} = 1$,

$k = q\left(\sum_{i=1}^{q} r_i - 1\right)$ for a product multinomial of varying numbers of cell probabilities

$r_1, r_2, ..., r_q$, $\sum_{j=1}^{r_i} p_{ij} = 1$:

$$\mathrm{X}^2 = nD^2 \sim \chi_k^2$$

$$E(\mathrm{X}^2) = nE(D^2) = k$$

Set $nD^2 = k$, so that the moment estimate $\tilde{N} = \dfrac{k}{D^2}$.

$$E(\tilde{N}) = kE\left(\frac{1}{D^2}\right) = knE\left(\frac{1}{\mathrm{X}^2}\right) = \frac{kn\left(1/2\right)^{k/2}}{\Gamma\left(k/2\right)}\int_0^\infty y^{\left(\frac{k}{2}-1\right)-1} e^{-\left(\frac{1}{2}\right)y} dy$$

$$= \frac{kn\left(1/2\right)^{k/2}}{\Gamma\left(k/2\right)} \frac{\Gamma\left(\frac{k}{2}-1\right)}{\left(\frac{1}{2}\right)^{\left(\frac{k}{2}-1\right)}} = \frac{kn\left(1/2\right)\Gamma\left(\frac{k}{2}-1\right)}{\left(\frac{k}{2}-1\right)\Gamma\left(\frac{k}{2}-1\right)}$$

$$= \frac{kn\left(1/2\right)}{\left(k-2\right)\left(1/2\right)} = n\frac{k}{k-2}$$

$\widehat{N} = \tilde{N}\dfrac{k-2}{k}$ unbiased estimate.

Appendix 2

Sample code for R program.

```
# expected probabilities for 6 sampling times:
prob.t1=c(0.895, 0.1, 0.005)
prob.t2=c(0.22, 0.62, 0.15 ,0.01)
prob.t3=c( 0.01, 0.275 ,0.6, 0.1065,0.005,0.0035)
prob.t4=c(0.004, 0.146 ,0.85)
prob.t5=c(0.005,0.57,0.425)
prob.t6=c(0.001, 0.999)   # violation of minimum of 0.0035 for expected probability

q=6  # 6 sampling times

prob.adj=vector("list",q)

prob.adj[[1]]=prob.t1
prob.adj[[2]]=prob.t2
prob.adj[[3]]=prob.t3
prob.adj[[4]]=prob.t4
prob.adj[[5]]=prob.t5
prob.adj[[6]]=prob.t6

n.0=c(rep(50,q),rep(250,q),rep(750,q),rep(1000,q))   # choose starting n value, same for 
each t
l=length(n.0)
prob.0=rep(prob.adj,l/q)
nsim=1000 # choose number of simulations
N.emp=matrix(0,nsim,l/q)
p.log.p=vector("list",l)
p.emp=vector("list",l)  # empirical proportions of simulations
chi.sq.int=matrix(0,nsim,l)
chi.sq=matrix(0,nsim,l/q)
g.sq.int=matrix(0,nsim,l)
g.sq=matrix(0,nsim,l/q)
k=length(unlist(prob.adj))-q  # degrees of freedom of chi-squared



for(jj in 1:nsim){
  for (i in 1:l){
    p.emp[[i]]=t(rmultinom(nsim,size=n.0[i],prob=unlist(prob.0[[i]]))/n.0[i])
    chi.sq.int[jj,i]=n.0[i]*sum((p.emp[[i]][jj,]-prob.0[[i]])^2/prob.0[[i]])

    for (m in 1:(l/q) ){
      chi.sq[,m]=rowSums(chi.sq.int[,((m-1)*q+1):(m*q)])
      N.emp[jj,m]=n.0[m*q]*k/chi.sq[jj,m]
    }

p.log.p[[i]]=p.emp[[i]]*log(p.emp[[i]]/matrix(prob.0[[i]],nsim,length(prob.0[[i]]),byrow = 
TRUE))
  }
}

for(jjj in 1:nsim){
  for (iii in 1:l){
    for (s in 1:ncol(unlist(p.log.p[[iii]]))){
      if(is.nan(p.log.p[[iii]][jjj,s])==TRUE){
        p.log.p[[iii]][jjj,s]=0                # l'Hopital's Rule
      }
    }
    g.sq.int[,iii]=2*n.0[iii]*rowSums(unlist(p.log.p[[iii]]))
    for (m in 1:(l/q) ){
      g.sq[,m]=rowSums(g.sq.int[,((m-1)*q+1):(m*q)])
    }
  }
}


for (m in 1:(l/q)){
  qqplot(qchisq(ppoints(nsim),df=k), g.sq[,m],xlab = expression(paste(Chi^2, (k), 
~"Quantile")),ylab=expression(paste(-2*log*Lambda)),cex.lab=1)
  qqline(distribution = function(p) qchisq(p,df=k),(g.sq[,m]))
  mtext(unique(n.0)[m])
}


chi.sq.N.emp.low=matrix(0,nsim,l/q)
chi.sq.N.emp.up=matrix(0,nsim,l/q)
count.N.emp=matrix(0,nsim,l/q)
count.N.hat=matrix(0,nsim,l/q)
V.inverse=1/(k-2)


for (jj in 1:nsim){
  for (m in 1:(l/q)){
    chi.sq.N.emp.low[jj,m]=qchisq(0.05/2,df=k)*V.inverse*N.emp[jj,m]  # upper confidence 
interval for N.emp at alpha=0.05

    chi.sq.N.emp.up[jj,m]=qchisq(0.05/2,df=k,lower.tail=FALSE)*V.inverse*N.emp[jj,m] # 
lower confidence interval for N.emp at alpha=0.05

    if(n.0[m*q] > chi.sq.N.emp.low[jj,m]  & n.0[m*q] < chi.sq.N.emp.up[jj,m] )
    {
      count.N.emp[jj,m]=1
    }
    else{
      count.N.emp[jj,m]=0
    }
  }
}

for (jj in 1:nsim){
  for (m in 1:(l/q)){
    if(n.0[m*q] > (chi.sq.N.emp.low[jj,m]*(k-2)/k)  & n.0[m*q] < 
(chi.sq.N.emp.up[jj,m]*(k-2)/k))
    {
      count.N.hat[jj,m]=1
    }
    else{
      count.N.hat[jj,m]=0
    }
  }
}


# table with resuls of actual coverage at alpha=0.05 of N.emp and N.hat=N.emp*(k-2)/k
coverage.95=rbind(colSums(count.N.emp)/nsim,colSums(count.N.hat)/nsim)
row.names(coverage.95)=c("N.emp","N.emp(k-2)/k")
coverage.95
```