Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Exact distributions of the multinomial order statistics

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

MASTER OF SCIENCE
in
APPLIED MATHEMATICS
by

Anton Ogay
Delft, the Netherlands
August 2016

TU Delft Delft University of Technology

MSc THESIS APPLIED MATHEMATICS

Exact distributions of the multinomial order statistics

Algorithms for the exact distributions
and application in hypothesis testing

Anton OGAY

Delft University of Technology

Student number:	4408209	
Supervisor:	Dr. P. Cirillo,	TU Delft
Thesis committee:	Prof. dr. Cornelis W. Oosterlee,	TU Delft
	Dr. D. Kurowicka,	TU Delft
	Dr. J. -J. Cai,	TU Delft

An electronic version of this thesis is available at
http://repository.tudelft.nl/.

CONTENTS

1

INTRODUCTION

Multinomial distribution arises in various application areas: queuing theory, software reliability models, clinical trials and many others. Actually, data coming from any distribution with known cumulative distribution function for continuous case (or probability mass function for discrete data) via some transformation can be represented as the result of the multinomial experiment, which intuitively can be seen as the result of random throwing of N balls across m urns.
But in any context, the next question is common to appear - is the data really drawn from the multinomial distribution? Usually, it is interesting to test the null hypothesis that the underlying distribution is equiprobable multinomial.
The general framework that will be considered in this thesis is the following. Given a set of discrete i.i.d observations $N_1, N_2, \ldots, N_m$ with probability mass function P, we want to test the null hypothesis

$$H_0 : P = P_0,$$

vs

$$H_1 : P \neq P_0$$

where P_0 is the probability mass function of an equiprobable multinomial, which is given by

$$\begin{gathered} P_0(N_1 = n_1, N_2 = n_2, \ldots, N_m = n_m; p, m) = \frac{N!}{n_1! n_2! \ldots n_m!} p^N \\ \sum_{k=1}^{m} n_k = N, p = \frac{1}{m} \end{gathered} \tag{1.1}$$

1

Many procedures were proposed to answer this question. The classical way is to use the χ^2 goodness-of-fit test, which was first introduced in the paper by Pearson (1900) and based on the test statistic:

$$X^2 = \sum_{i=1}^{m} \frac{(N_i - Np)^2}{Np}.$$

Another popular procedure is to replace X^2 with the log likelihood ratio statistic:

$$G^2 = 2\sum_{i=1}^{m} N_i \ln(\frac{N_i}{Np})$$

Also well known statistics to use are: the Freeman-Turkey statistic, the Neyman modified X^2 statistic or the modified log likelihood ratio statistic.[8]

Young(1962)[11] revisited this problem and proposed two alternatives to the existing tests, with the test statistics based on the scaled range of the sample

$$W_m = (\max_{1\le k\le m} N_k - \min_{1\le k\le m} N_k)(\frac{m}{N})^{\frac{1}{2}},$$

or on the scaled mean

$$M_m = (mN)^{-\frac{1}{2}} \sum_{i=1}^{m} \left| N_i - \frac{N}{m} \right|.$$

Young also showed, that the range based statistics reveal power advantage under some alternatives.

But, all tests ,with the mentioned statistics, rely on the approximate distributions, which require original data to satisfy some conditions. For example, as a rule of thumb, use of Chi-squared approximation for the Pearson statistic is warranted only for values N and m, such that $\frac{N}{m} \ge 5$. But in some applications, these conditions are rarely satisfied, which leads to the inability to use these tests, or to the inaccuracy in the results. This problem can be solved, if the exact distribution of the used statistic is known.

So, the goal of this thesis is to develop the algorithm to compute the exact distribution of the range of multinomial sample, and use it to built a test based on the exact distribution of the statistic.

The outline of this thesis will go as following : Chapter 2 provides an overview of existing approximations for the range and the ordered values statistics distributions, and some existing procedures to compute the exact distributions. Chapter 3 is dedicated to the new version of an algorithm for the exact distribution of the multinomial range. Accuracy of the approximations is discussed in Chapter 4. In Chapter 5, the goodness-of-fit test based on multinomial range is discussed, along with its applications for the homogeneous Poisson process and the case studies in biometry.

2

EXACT AND APPROXIMATE DISTRIBUTIONS OF THE ORDERED VALUE STATISTICS AND THE RANGE

In order to perform hypothesis testing, the distribution of the test statistics has to be determined. This chapter introduces approximations and different computational methods for calculating the exact distribution functions of maximum, minimum and range for the multinomial distribution.
We first briefly discuss ways to approximate distribution of the range and the maximum, using results of Johnson & Young[7] and DasGupta[4]. We also give an overview of previously developed algorithms for the exact values, presented in papers of Corrado[3] and Rapperport[9].

2.1. APPROXIMATING DISTRIBUTIONS

First approximation to the multinomial distribution was introduced by Johnson & Young[7], which was based on multinormal limit for the multinomial sample. Later, Young[11] used this approximation to derive limiting distribution for the range of the sample.
The distribution of maximum, using Gumbel approximation, was initially introduced by Kolchin(1978)[5], but contained some typos and errors, which were corrected by DasGupta[4].
Since the null hypothesis we will use later, is that the underlying distribution is equiprob-

able multinomial, all approximations are derived under this condition. Next two sections follow original papers to provide details about approximations.

2

2.1.1. APPROXIMATION OF THE RANGE DISTRIBUTION

In order to define the approximating range distribution, we first look at the joint limiting distribution for a multinomial sample.

Consider, a set of m discrete random variables $n_1,\dots,n_m$ with probability mass function 1.1. For equiprobable case, expectation and variance of n_i are given by

$$E(n_i) = Np = \frac{N}{m}\,, Var(N_i) = Np(1-p).$$

Then straightforwardly from multidimensional central limit theorem, the joint distribution of standardized multinomial values ω_i

$$\omega_i = \frac{n_i - Np}{\sqrt{Np(1-p)}} = \frac{mn_i - N}{\sqrt{N(m-1)}}$$

with $N \to \infty$ converges to multinormal distribution with zero means, unit variances and covariance between w_i and w_j equal to $\frac{1}{1-k}$ for $i \neq j$.
Consider now a set of m independent unit normal variables $x_1,\dots,x_m$. Using the properties of normal distribution, it can be shown that the standardized deviates from the sample mean

$$t_1 = \left(\frac{m}{m-1}\right)(x_1 - \overline{x}),\dots,t_m = \left(\frac{m}{m-1}\right)(x_m - \overline{x})$$

are jointly distributed multinormally with zero means, unit variances and equal covariances $\frac{1}{1-k}$.
Hence, the distribution of standardized multinomial variables ω_i could be approximated by the joint distribution of the standardized mean deviates of $x_1,\dots,x_m$.
From this it follows, that the range distribution of ω_i's can be approximated by the distribution of the range of t_i's.

$$P(\max_{1\le i\le m}\omega_i - \min_{1\le i\le m}\omega_i \le r) \xrightarrow{D} P(\max_{1\le i\le m} t_i - \min_{1\le i\le m} t_i \le r)$$

Now, note that

$$\text{Range}(\omega_i) = \max_{1\le i\le m}\omega_i - \min_{1\le i\le m}\omega_i = \frac{m}{\sqrt{N(m-1)}}(\max_{1\le i\le m} n_i - \min_{1\le i\le m} n_i) = \frac{m}{\sqrt{N(m-1)}}\text{Range}(n_i)$$

Also

$$\text{Range}(t_i) = \left(\frac{m}{m-1}\right)(\max_{1\le i\le m} x_i - \min_{1\le i\le m} x_i) = \left(\frac{m}{m-1}\right)\text{Range}(x_i)$$

Finally the range distribution of multinomial sample, $n_1, \ldots, n_m$, could be approximated by the range distribution of m i.i.d standard normal variables

$$P\left(\max_{1\le i\le m} n_i - \min_{1\le i\le m} n_i \le r\right) \xrightarrow{D} P\left(\max_{1\le i\le m} x_i - \min_{1\le i\le m} x_i \le r\sqrt{\frac{m}{N}}\right) \tag{2.1}$$

The distribution of the range of m i.i.d standard normal variables is a known quantity, and can be easily computed:

$$P\left(\max_{1\le i\le m} x_i - \min_{1\le i\le m} x_i \le r\right) = m\int_{-\infty}^{\infty} f(x)\left(\int_x^{x+r} f(u)(d)u\right)^{m-1} \mathrm{d}x$$

where $f(x)$ is the probability density function of the standard normal variable.
Young[11] also noted, that the distribution of the range of a multinomial sample is necessarily discrete, with interval 1 for possible values, when $m > 2$, and interval 2 for the binomial case. So he introduced the continuity factor δ_k

$$P\left(\max_{1\le i\le m} n_i - \min_{1\le i\le m} n_i \le r\right) \xrightarrow{D} P\left(\max_{1\le i\le m} x_i - \min_{1\le i\le m} x_i \le (r+\delta_k)\sqrt{\frac{m}{N}}\right)$$

where

$$\delta_k = 1 \text{ for } k = 2, \delta_k = 0.5 \text{ for } k > 2$$

2.1.2. APPROXIMATION FOR THE MAXIMUM

Since, the initial formulation of the approximating distribution for the maximum of multinomial sample, proposed by Kolchin[5], contained some typos, here we provide corrected version of the theorem by DasGupta[4].
DasGupta stated the following:

Theorem. *If* $(n_1, n_2, \ldots, n_m) \sim Mult\left(N, \frac{1}{m}, \ldots, \frac{1}{m}\right)$.
Let

$$\mu = \frac{N}{m}, \omega = \frac{\log - 0.5\log\log m}{\mu}$$

And ϵ *us the unique positive root of the equation*

$$(1+\epsilon)\log(1+\epsilon) - \epsilon = \omega$$

Then the distribution of the maximum of multinomial sample converges in distribution to

$$P\left(\frac{\max n_i - \mu(1+\epsilon)}{\sqrt{\frac{N}{2m\log m}}} + 0.5\log(4\pi) \le z\right) \xrightarrow{D} F_{Gumbel}(z,0,0) = e^{-e^{-z}}$$

for all real z.

In this case, no continuity correction was introduced.

2.2. ALGORITHMS FOR THE EXACT DISTRIBUTIONS

2

Obtaining the exact distribution of the extremal order statistics (minimum, maximum and range) has been a problem of recurring interest in statistics. Few papers dealt with this issue. One of the first algorithms to obtain the exact distributions of the maximum and the minimum was proposed by Rapperport[9] in his unpublished thesis. Since the paper remained unprinted, Rapperport's idea has been neglected for a while. In 2010, Corrado[3] looked at this problem from a different angle, and suggested an alternative procedure.

2.2.1. STOCHASTIC REPRESENTATION ALGORITHM

One of the approaches to calculate the desired distributions, suggested by Corrado[3], is based on the stochastic matrix representation, which determines transition probabilities for the number of balls in urns. With n_k being the number of balls in urn k, the sequence $s_k = s_{k-1} + n_k$ describes the cumulative ball count from $s_0 = 0$ to $s_m = n$, where m is the total number of urns. Transition probability from s_{k-1} to s_k is defined as follows:

$$P(s_k|s_{k-1}, p_k^*) = \begin{cases} \binom{n-s_k}{s_k-s_{k-1}} (p_k^*)^{s_k-s_{k-1}} (1-p_k^*)^{n-s_k} & \text{if } s_k \geq s_k - 1 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

$$\text{where } p_k^* = p_k / \sum_{j=k}^{m} p_j$$

Using the probabilities 2.2, stochastic matrices are formed in the following way:

$$Q_k = \begin{bmatrix} P(0|0,p_k^*) & P(1|0,p_k^*) & \dots & P(n|0,p_k^*) \\ 0 & P(1|1,p_k^*) & \dots & P(n|1,p_k^*) \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} \tag{2.3}$$

$$Q_1 = [P(0|0,p_1)P(1|0,p_1)\dots P(n|0,p_1)]$$

$$Q_m^T = [111\dots1]$$

This representation turns out very handy, since it provides straightforward way to calculate desired distributions. For example, in order to calculate the exact probability of the maximum amount of balls in the urn being not more than r, all transition probabilities $P(s_k|s_{k-1}, p_k^*)$ for which $s_k - s_{k-1} > r$ should be set to 0. Product of modified stochastic matrices results in the exact probability $P(\max n_k \leq r)$. Distribution of the smallest order statistics can be obtained in the same way, simply changing the inequality sign.

Distribution of the range of multinomial sample also can be computed using the matrix representation. To simplify notation, we denote $Q_k(a_k, b_k)$ stochastic matrix for urn k, where $P(s_k|s_{k-1}, p_k^*) = 0$ for all $n_k > a_k$ or $n_k < b_k$. Introducing the set of all possible allocations of n balls among m urns as $\cap_{k=1}^{m} a_k \geq n_k \geq b_k$, we can express joint probability of maximum and minimum ball count as:

$$P(\cap_{k=1}^{m} a_k \geq n_k \geq b_k) = Q_1^1 \times \prod_{k=2}^{m-1} Q_k(a_k, b_k) \times Q_m \tag{2.4}$$

Note that the set of allocations, described above, have intersecting intervals. In order to calculate the probability of multinomial range exactly, intersection probabilities should be subtracted:

$$\begin{aligned} P(\max_{1\leq k\leq m} n_k - \min_{1\leq k\leq m} n_k < r) = \sum_{h=0}^{n-r+1} Q_1^1 \times \prod_{k=2}^{m} Q_k(h+r-1, h) \times Q_m - \\ \sum_{h=0}^{n-r} Q_1^1 \times \prod_{k=1}^{m} Q_k(h+r-1, h+1) \times Q_m \end{aligned} \tag{2.5}$$

The main advantage of the stochastic matrix representation approach is that it does not require equal urn probabilities, which can be very useful further for not equiprobable cases. On other hand, the stochastic matrix representation is an ad-hoc solution given that for every new composition, matrices should be redesigned and recalculated.

2.2.2. TREE-BASED ALGORITHM

In 1968, Rappeport[9] proposed an iterative algorithm for obtaining distributions of the order statistics for the multinomial sample. This algorithm is based on the representation of all possible outcomes of multinomial trial in a form of a tree. For example, random scatter of 6 balls across 3 urns can result in one of the paths through the tree, shown on the Figure2.1. Different levels of the tree represent the number of balls in the urn. The nodes - the number of urns with the specified amount of balls.

So, the path, indicated blue on the Figure2.1, shows the case, when 6 balls are drawn into one urn and the other two urns are empty. Green branch of the tree represents the result of the experiment, when one urn contains 3 balls, another urn is filled with 2 balls, and the last one has 1 ball.

In the following sections, algorithms are also derived under assumption of the equiprobable multinomial. We begin with the algorithm for the maximum, since it is the basis for all succeeding algorithms.

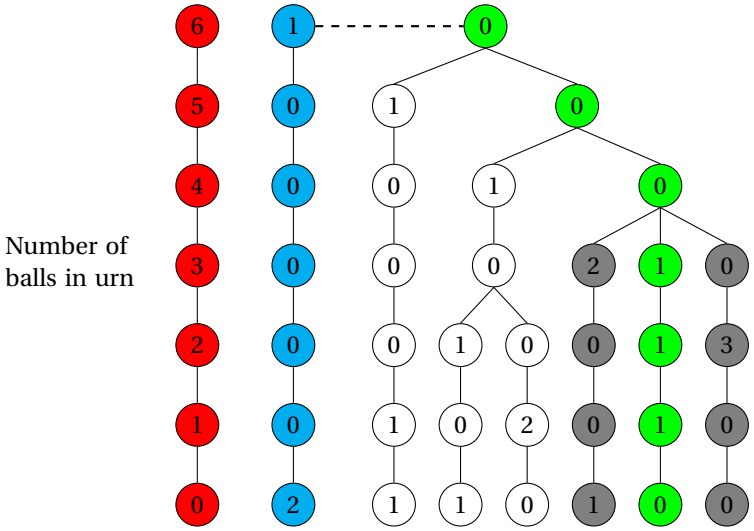

Figure 2.1: Tree representation of the throw of 6 balls into 3 urns

2.2.3. DISTRIBUTION OF THE LARGEST ORDERED VALUE

To compute the distribution of the highest order statistic

$$P(n_{<1>} \le r : N, m), \text{ with } n_{<1>} = \max_{1 \le k \le m} n_k$$

we need to sum up probabilities of all the paths, that have zero nodes for all levels from $r+1$ to N. For example, if want to compute the probability that no urn contain more than $r = 3$ balls, when $N = 6$ and $m = 3$, we need to sum probabilities of the paths, indicated gray and green on the Figure 2.1.

As a solution to this, Rappeport[9] developed an iterative procedure. The general idea is to represent $P(n_{<1>} \le r : N, m)$ in terms of $P(n_{<1>} \le r-1 : N, m)$ and find a way to compute this probability explicitly for some specific r.

We start from computing the probability, that the maximum amount of the balls in the urns equals exactly to r, and assuming, that such an urn is unique, i.e. all others urns have at least 1 ball less. Then using 1.1 and introducing operator W_2, which is nothing more than a sum over all possible values $n_{<2>} \dots n_{<m>}$, such that $n_{<1>} > n_{<2>}$, we can compute

$$P\left(n_{<1>} = r : N, m, n_{<1>} > n_{<2>}\right) = \frac{1}{r!} W_2 \left(\frac{N!}{m^N \prod_{i=2}^{m} n_{<i>}!} \frac{m!}{\prod_{k=0}^{r-1} (\#n_i = k)!} \right)$$

where $\#n_i = k$ denotes the number of n_i's equal to k.
The fraction

$$\frac{m!}{\prod_{k=0}^{r-1}(\#n_i = k)!}$$

arises from the simple combinatorial argument. Since all the bins are the same, the probability of given composition of $n_1,..n_m$ should be multiplied by total number of its unique permutations.
If we relax condition of the uniqueness of maximum and use $q = 0$ to mean that $n_{<1>} < r$, then

$$P\left(n_{<1>} =\leq r : N, m,\right) = \sum_q \frac{1}{r!q!}\frac{N!m!}{m^N} W_{q+1}\left(\frac{1}{\prod_{i=2}^{m} n_{<i>}!\prod_{k=0}^{r-1}(\#n_i = k)!}\right)$$

Clearly, the amount of bins, containing exactly r balls, can't be bigger than $\lfloor\frac{N}{r}\rfloor$, which defines the upper limit of the summation. Also, since all other $N - rq$ balls should be placed in $m - q$ urns, with maximum not exceeding $r - 1$, following inequality should be satisfied

$$(m - q)(r - 1) \geq (N - rq),$$

from which the lower bound for q is $\max(0, N - rm + m)$. So the range of the summation is defined as:

$$\max(0, N - rm + m) \leq q \leq \lfloor\frac{N}{r}\rfloor \tag{2.6}$$

Finally, noticing that

$$W_{q+1}\left(\frac{1}{\prod_{i=2}^{m} n_{<i>}!\prod_{k=0}^{r-1}(\#n_i = k)!}\frac{(m-q)!(N-rq)!}{(m-q)^{N-rq}}\right) = P\left(n_{<1>} \leq r-1 : N-rq, m-q\right),$$

we can write the iterative formula for the probability of maximum

$$P(n_{<1>} \leq r : N, m) = \sum_q A_q P(n_{<1>} \leq r-1 : N-rq, m-q) \tag{2.7}$$

$$A_q = \frac{N!m!}{m^N}\frac{1}{r!^q q!}\frac{(m-q)^{N-rq}}{(m-q)!(N-rq)!}$$

As was mentioned, to carry out iteration procedure it should be feasible to evaluate it for some particular value of r, which is clearly possible for $r = 1$.

$$P(n_{<1>} \leq 1 : N, m) = \begin{cases} \frac{m!}{m^N(m-N)!} & \text{if } m \geq N \\ 0 & m < N \end{cases} \tag{2.8}$$

DISTRIBUTION OF THE SUM OF THE HIGHEST ORDER STATISTIC

The algorithm for the maximum can also be used for the calculation of probability function for the sum of the highest order statistic.

As an example, we consider the case of the sum of the first three highest order statistics, $P(n_{<1>} + n_{<2>} + n_{<3>} \leq r : N, m)$. This example was also provided in the original paper, but contained some typos, which we corrected.

2

The framework of the method is to divide the probability function into the different terms, corresponding to the different ranges of $n_{<1>}$ and $n_{<2>}$. We can define three disjoint cases:

- $n_{<1>} \leq \frac{r}{3}$ - in this case, clearly

$$P(n_{<1>} + n_{<2>} + n_{<3>} \leq r : N, m) = P(n_{<1>} \leq \frac{r}{3} : N, m)$$

- $n_{<1>} > \frac{r}{3}$ and $n_{<2>} \leq \frac{r - n_{<1>}}{2}$ - here we can fix value of $n_{<1>} = t_1$, therefore reserve one urn to have exactly t_1 balls. If the maximum of the smaller sample, with $N = N - t_1$ and $m = m - 1$, will be smaller or equal than $\frac{r-t_1}{2}$, original inequality for the sum of the three highest order statistics will automatically hold. Total probability in this case is equal to the sum over all possible values of $n_{<1>}$.

$$P(n_{<1>} + n_{<2>} + n_{<3>} \leq r : N, m) = \sum_{t_1 = \lfloor \frac{r}{3} + 1 \rfloor}^{r} A_{t_1} P(n_{<1>} \leq \frac{r - t_1}{2} : N - t, m - 1)$$

$$A_{t_1} = \frac{N! m!}{m^N} \frac{1}{t_1!} \frac{(m-1)^{N-t_1}}{(m-1)!(N-t_1)!}$$

- $n_{<1>} > \frac{r}{3}$ and $n_{<2>} > \frac{r - n_{<1>}}{2}$ - analogously to the previous case, but we fix both values of $n_{<1>}$ and $n_{<2>}$.

$$P(\sum_{i=1}^{3} n_{<i>} \leq r : N, m) = \sum_{t_1 = \lfloor \frac{r}{3} + 1 \rfloor}^{r-1} \sum_{t_2 = \lfloor \frac{r-t_1}{2} + 1 \rfloor}^{\min(r, t_1 - r)} A_{t_1, t_2} B_{t_1, t_2} P(n_{<1>} \leq r - t_1 - t_2 : N - t_1 - t_2, m - 2)$$

$$A_{t_1, t_2} = \frac{N! m!}{m^N} \frac{(m-2)^{N-t_1-t_2}}{(m-2)!(N-t_1-t_2)!} \frac{1}{t_1! t_2!}$$

$$B_{t_1, t_2} = \begin{cases} \frac{1}{2!} & \text{if} t_1 = t_2 \\ 1 & \text{otherwise} \end{cases}$$

Coefficient B_{t_1, t_2} appears, because in this case values of t_1 and t_2 might be equal, so we need to account on their permutation.

Distribution of a sum of the J highest order statistics can be computed in the same way, splitting probability function into J terms. Each term corresponds to a particular range of values of $n_{<i>}$. On every interval we fix some of the $n_{<i>}$ and calculate the probability of the maximum on the smaller sample, restricting it in such a way, that original inequality holds. Summation over all possible values of fixed $n_{<i>}$ results in the total probability

for this range. The general formula for the distribution of a sum of the J highest order statistics is following

$$\begin{aligned} &P(\sum_{i=1}^{J} n_{<i>} \le r : N, m) = \\ &= P(n_{<1>} \le \frac{r}{J} : N, m) + \sum_{t_1=\lfloor \frac{r}{J}+1 \rfloor}^{r} A_{t_1} P(n_{<1>} \le \frac{r-t_1}{J-1} : N-t_1, m-1) + \\ &+ \cdots + \sum_{t_1=\lfloor \frac{r}{J}+1 \rfloor}^{r-J+2} \cdots \sum_{t_{J-1}} A_{t_1,\ldots,t_{J-1}} B_{t_1,\ldots,t_{J-1}} P(n_{<1>} \le r - \sum_{i=1}^{J-1} t_i : N - \sum_{i=1}^{J-1} t_i, m-J+1) \end{aligned} \tag{2.9}$$

If we denote I as the total number of summations for the particular range, then summation limits are defined as:

$$\begin{cases} \lfloor \frac{r}{J}+1 \rfloor \le t_1 \le r-I+1 \text{ - for } i=1 \\ \lfloor \frac{r-\sum_{i=1}^{I-1} t_i}{J-I+1}+1 \rfloor \le t_i \le \min(t_{I-1}, r-\sum_{i=1}^{I-1} t_i) \text{ - for } 2 \le i \le I \end{cases}$$

And coefficients A and B are calculated according to the following formula:

$$\begin{aligned} A_{t_1,\ldots,t_I} &= \frac{N!m!}{m^N} \frac{1}{\prod_{i=1}^{I}(n_{<i>})!} \frac{(m-I)^{(N-\sum_{i=1}^{I} t_i)}}{(m-I)!(N-\sum_{i=1}^{I} t_i)!} \\ B_{t_1,\ldots,t_I} &= \frac{1}{\prod_{k=t_1}^{t_I}(\#t_i=k)!} \end{aligned}$$

where $\#t_i = k$ denotes the number of t_i's equal to k.

2.2.4. DISTRIBUTION OF THE SMALLEST ORDER STATISTICS

The distribution of the smallest order statistics can be easily derived, simply using calculation of the sum of $m-1$ highest order statistics, since

$$P(\min_{1\le k\le m} n_k \ge r) = P(\sum_{i=1}^{m-1} n_{<i>} \le N-r : N, m) \tag{2.10}$$

However, this approach turns out very computationally inefficient already for quite small N and m.

Rapperport, in his thesis, suggested an idea how the algorithm 2.7 could be used to compute the distribution of the minimum, but no clear statement of the algorithm was made. We provide a detailed explanation on how to exploit this method in order to compute the minimum distribution.

Carrying out the iteration procedure, we can assign 0 probability to the branches that have urns with less than r balls - we define $P(n_{<1>} \le r-1 : N, m) = 0$ for all cases when

$N \neq 0$ or $m \neq 0$. Iterating through all possible values of the maximum - from r to N - we will compute the desired probability for the minimum.

$$\begin{aligned} &P(\min_{1\leq k\leq m} n_k \geq r) = \sum_{t=r}^{N} P(n_{<1>} \leq t : N, m), \text{ with} \\ &P(n_{<1>} \leq r-1 : N, m) = 1 \text{ if } N = 0, m = 0 \\ &P(n_{<1>} \leq r-1 : N, m) = 0 \text{ if } N \neq 0 \vee m \neq 0 \end{aligned} \tag{2.11}$$

As an example, we again consider the multinomial experiment, with $N = 6$ and $m = 3$, and compute the probability, that no bin has less than 2 balls, i.e. $P(\min n_i \geq 2 : N = 6, m = 3)$. Algorithm 2.11,on the first iteration will compute the probability of the green path on Figure 2.2. Following iterations will assign 0 probability to the gray branches, since they all have bins with 1 ball or empty urns.

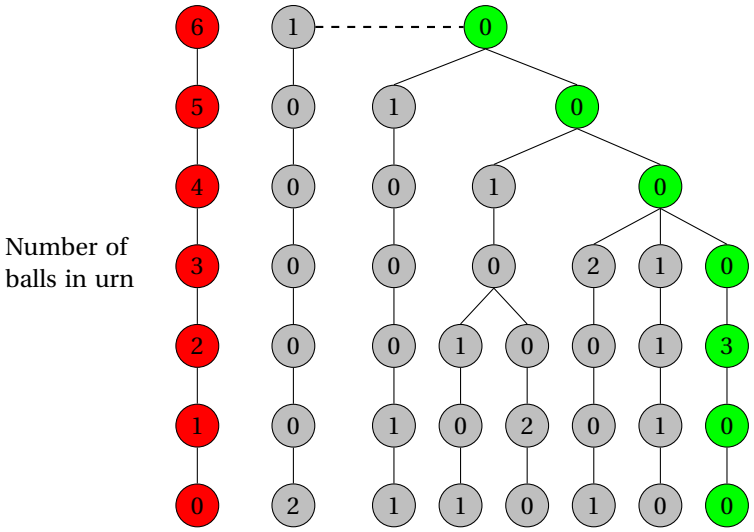

Figure 2.2: Outcome tree for the throw of 6 balls into 3 urns with visualization for the minimum algorithm

Now, the initial algorithm can be improved in a more computationally efficient way. We can notice that algorithm for the maximum traverse through all the branches of the outcome tree, while computing $P(P(n_{<1>} \leq N : N, m)$. So, we can find distribution of the minimum simply calculating $P(n_{<1>} \leq N : N, m)$ with the same "rules" for assigning 0 probability for branches as in 2.11. This addition reduces amount of iterations, since we don't calculate probabilities for the same branch multiple times.

$$\begin{aligned} &P(\min_{1\leq k\leq m} n_k \geq r) = P(n_{<1>} \leq N : N, m), \text{ where} \\ &P(n_{<1>} \leq r-1 : N, m) = 1 \text{ if } N = 0, m = 0 \\ &P(n_{<1>} \leq r-1 : N, m) = 0 \text{ if } N \neq 0 \vee m \neq 0 \end{aligned} \tag{2.12}$$

Both versions of the algorithm have been implemented and made available in this thesis. We also provide tables with the critical values for the described distributions in the Appendix B.

3

ALGORITHM FOR THE RANGE DISTRIBUTION

In this chapter, we use same framework as Rapperport to develop a new method for calculating the range distribution for the multinomial sample. Unlike the solution, proposed by Corrado[3], our new algorithm doesn't require redesign for every new composition.

3.1. DISTRIBUTION OF THE RANGE

To make the idea of the algorithm clear, we first return to the example of the multinomial experiment with $N = 6$ balls and $m = 3$ urns and describe the steps to compute the probability of the range being smaller or equal to 3.
We can split this probability in two terms, corresponding to the different ranges of the maximum.

$$\begin{aligned} P(\max_{1\le k\le 3} n_k - \min_{1\le k\le 3} n_k \le 3) = P(\max_{1\le k\le 3} n_k - \min_{1\le k\le 3} n_k \le 3 : \max n_k \le 3) + \\ + P(\max_{1\le k\le 3} n_k - \min_{1\le k\le 3} n_k \le 3 : \max n_k > 3) \end{aligned} \tag{3.1}$$

First term can be simply computed using 2.7, since obviously

$$P(\max_{1\le k\le 3} n_k - \min_{1\le k\le 3} n_k \le 3 : \max n_k \le 3) = P(\max_{1\le k\le 3} n_k \le 3)$$

In order to compute the second term, we can use a similar idea as for the minimum. We iterate through all possible values for the maximum, assigning 0 probability to the branches that have urns with less than "current maximum - 3 " balls. For example, we compute $P(n_{<1>} \leq 4 : n_{<1>} > 3)$ with following conditions :

$$\begin{aligned} &P(n_{<1>} < 1 : N, m) = 1 \text{ if } N = 0, m = 0 \\ &P(n_{<1>} < 1 : N, m) = 0 \text{ if } N \neq 0 \vee m \neq 0 \end{aligned} \tag{3.2}$$

Using this, we will compute the probability of the green path on Figure3.1 and discard the grey one. Note, that we don't visit the three blue paths, since we consider the case in

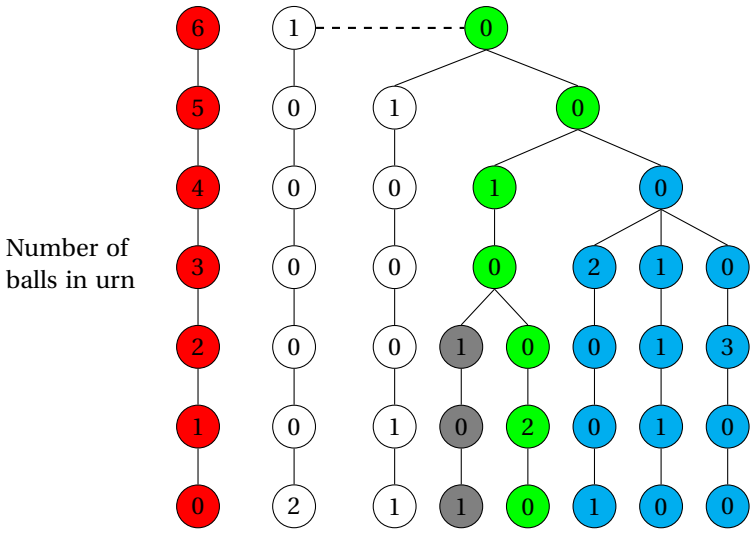

Figure 3.1: Selection of the tree branches in the range algorithm in case of N=6,m=3

which $\max n_k > 3$. Similarly, we compute $P(n_{<1>} \leq 5 : n_{<1>} > 4)$ and $P(n_{<1>} \leq 6 : n_{<1>} > 5)$ increasing minimum "allowed" amount of balls in urn to 2 and 3 correspondingly. Summing up all resulted probabilities, we obtain desired value for the range probability. Now, we will derive an equation for general case. Obviously, we can split the probability the same way as in the example:

$$\begin{aligned} P(\max n_i - \min n_i \leq r : N, m) = P(\max n_i \leq r : N, m) + \\ + P(\max n_i - \min n_i \leq r : N, m, \max n_i > r) \end{aligned} \tag{3.3}$$

The first term is easily computed using the algorithm for maximum 2.7, so now we will consider the second term.
Assume, for time being, that $\max n_i = n_{<1>} = n_{<q>} = r + 1 > n_{<q>}$ and denote $\min n_i =$

$n_{<m>}$. Then, we can rewrite the second term of 3.3:

$$P(n_{<1>} - n_{<m>} \le r : N, m, n_{<1>} = r+1) = \\ = \frac{1}{((r+1)!)^q q!} F_{q+1,1}\left(\frac{N!m!}{m^N} \frac{1}{\prod_{i=q+1}^{m} n_{<i>}! \prod_{m=t-r}^{r} (\#n_i = m)!}\right), \tag{3.4}$$

where $F_{q+1,t-r}$ is an operator that sums over all possible values of $n_{<q+1>}, \ldots, n_{<m>}$, such that $n_{<1>} = n_{<q>} > n_{<q+1>}$ and $n_{<m>} \ge 1$.

3

Summing over all possible values for q will result in the total probability for this case

$$P(n_{<1>} - n_{<m>} \le r : N, m, n_{<1>} = r+1) = \\ = \sum_q \frac{1}{((r+1)!)^q q!} \frac{N!m!}{m^N} F_{q+1,1}\left(\frac{1}{\prod_{i=q+1}^{m} n_{<i>}! \prod_{m=t-r}^{r} (\#n_i = m)!}\right) \tag{3.5}$$

Since, in this case $n_{<1>} > r$, range of q in summation is different from 2.6. To impose this condition, case of $q = 0$ should be excluded, so

$$\max(1, N - (r+1)m + m) \le q \le \lfloor \frac{N}{r+1} \rfloor$$

Now, if we multiply and divide 3.5 by $\frac{(m-q)^{(N-(r+1)q)}}{(m-q)!(N-(r+1)q)!}$, we can write it in the following form:

$$P(n_{<1>} - n_{<m>} \le r : N, m, n_{<1>} = r+1) = \sum_q \frac{1}{((r+1)!)^q q!} \frac{N!m!}{m^N} \\ \frac{(m-q)^{(N-(r+1)q)}}{(m-q)!(N-(r+1)q)!} F_{q+1,1}\left(\frac{1}{\prod_{i=q+1}^{m} n_{<i>}! \prod_{m=t-r}^{r} (\#n_i = m)!} \frac{(m-q)!(N-(r+1)q)!}{(m-q)^{(N-(r+1)q)}}\right) \tag{3.6}$$

We can notice, that

$$F_{q+1,1}\left(\frac{1}{\prod_{i=q+1}^{m} n_{<i>}! \prod_{m=t-r}^{r} (\#n_i = m)!} \frac{(m-q)!(N-(r+1)q)!}{(m-q)^{(N-(r+1)q)}}\right)$$

is simply another form of writing

$$P\left(\min n_i \ge 1 : N - (r+1)q, m - q, n_{<1>} \le r\right)$$

Let's recall the modified formula 2.12 for the distribution of minimum:

$$\begin{aligned}
&P(\min_{1\le k\le m} n_k \ge k) = P(n_{<1>} \le N : N, m), \text{ where}\\
&P(n_{<1>} \le k-1 : N, m) = 1 \text{ if } N = 0, m = 0\\
&P(n_{<1>} \le k-1 : N, m) = 0 \text{ if } N \ne 0 \vee m \ne 0
\end{aligned}$$

The condition $n_{<1>} \le r$ can be straightforwardly imposed to the algorithm by changing the threshold for the maximum. For the current case:

$$\begin{aligned}
&P\left(\min n_i \ge 1 : N-(r+1)q, m-q, n_{<1>} \le r\right) = P(n_{<1>} \le r : N, m), \text{ where}\\
&P(n_{<1>} \le 0 : N, m) = 1 \text{ if } N = 0, m = 0\\
&P(n_{<1>} \le 0 : N, m) = 0 \text{ if } N \ne 0 \vee m \ne 0
\end{aligned} \tag{3.7}$$

Plugging this into equation 3.6:

$$\begin{aligned}
&P(n_{<1>} - n_{<m>} \le r : N, m, n_{<1>} = r+1) = \sum_q \frac{1}{((r+1)!)^q q!} \frac{N!m!}{m^N}\\
&\frac{(m-q)^{(N-(r+1)q)}}{(m-q)!(N-(r+1)q)!} P(n_{<1>} \le r : N-(r+1)q, m-q),
\end{aligned} \tag{3.8}$$

with conditions

$$\begin{aligned}
&P(n_{<1>} \le 0 : N, m) = 1 \text{ if } N = 0, m = 0\\
&P(n_{<1>} \le 0 : N, m) = 0 \text{ if } N \ne 0 \vee m \ne 0
\end{aligned}$$

Using equation2.7 for the distribution of the maximum, we can write the desired probability as:

$$\begin{aligned}
&P(n_{<1>} - n_{<m>} \le r : N, m, n_{<1>} = r+1) = P(n_{<1>} \le r+1 : N, m), \text{ where}\\
&P(n_{<1>} \le 0 : N, m) = 1 \text{ if } N = 0, m = 0\\
&P(n_{<1>} \le 0 : N, m) = 0 \text{ if } N \ne 0 \vee m \ne 0
\end{aligned} \tag{3.9}$$

Finally, we can calculate the second term in 3.3 summing over all possible values of $n_{<1>}$

$$\begin{aligned}
&P(\max n_i - \min n_i \le r : N, m, \max n_i > r) = \sum_{t=r+1}^{N} P(n_{<1>} \le t : N, m, n_{<1>} > t-1), \text{ where}\\
&P(n_{<1>} \le t-r-1 : N, m) = 1 \text{ if } N = 0, m = 0\\
&P(n_{<1>} \le t-r-1 : N, m) = 0 \text{ if } N \ne 0 \vee m \ne 0
\end{aligned} \tag{3.10}$$

Condition $n_{<1>} > t-1$ is introduced at every summation step in order to avoid calculations for the same branches of the outcome tree. This results in different ranges of summations in the recursion, i.e.

$$\max(1, N-tm+m) \le q \le \lfloor \frac{N}{t} \rfloor \text{ if } n_{<1>} > r$$
$$\max(0, N-tm+m) \le q \le \lfloor \frac{N}{t} \rfloor \text{ if } n_{<1>} \le r$$

The range distribution can be evaluated via the following iteration procedure:

$$\begin{aligned} P(\max_{1\le k\le m} n_k - \min_{1\le k\le m} n_k \le r : N, m) &= P(n_{<1>} \le r : N, m) + \\ &+ \sum_{t=r+1}^{N} P(n_{<1>} \le t : N, m, n_{<1>} > t-1) \\ &\text{with conditions for the case } n_{<1>} > r \\ &P(n_{<1>} \le t-r-1 : N, m) = 1 \text{ if } N = 0, m = 0 \\ &P(n_{<1>} \le t-r-1 : N, m) = 0 \text{ if } N \ne 0 \vee m \ne 0 \end{aligned} \tag{3.11}$$

IMPLEMENTATION IN MATLAB

All the algorithms presented in this thesis were implemented in MATLAB. Even if the main application area of discussed algorithms are samples of small size, it is interesting to compare approximating and exact distributions for quite big values. It turns out that standard routine, implemented in MATLAB, allows calculation of the factorials up to 170. In order to avoid this problem, logarithm of gamma function was used, since

$$t! = \exp(\log(\Gamma(t-1))$$

Taking natural logarithm turned out very useful, since all formulas contain fractions with factorials in denominator and numerator, which were replaced by summation and subtraction.

In Appendix B, table of critical values for the range distribution is presented for the different combinations of the amount of observations and bins. For any other case, values could be obtained using the code from Appendix A.

4

ACCURACY OF THE APPROXIMATIONS

This chapter is dedicated to the analysis of accuracy of the approximating distributions for the maximum and the range of a multinomial sample. We compare approximate values with the exact ones, computed using the algorithms, discussed in Chapters 2-3. Along with exact values for different cases, we also provide some statistics about errors of the approximation.

4.1. ACCURACY OF THE MAXIMUM DISTRIBUTION APPROXIMATION

In this section, we examine calculation precision of the Gumbel approximation for the maximum distribution of multinomial sample. As was shown in 2.1.2, approximation is given by:

$$P\left(\frac{\max n_i - \mu(1+\epsilon)}{\sqrt{\frac{N}{2m\log m}}} + 0.5\log(4\pi) \le z\right) \xrightarrow{D} F_{Gumbel}(z,0,0) = e^{-e^{-z}}$$

We start with investigating accuracy of the approximation for the binomial case. Figure 4.2, 4.2, 4.3 show cumulative distribution functions when the number of observations is equal to 5,10 and 50.

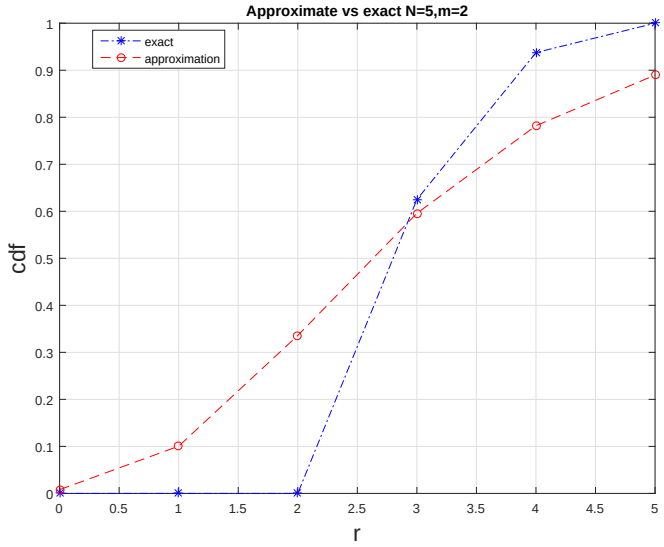

Figure 4.1: Exact maximum distribution and Gumbel approximation for N=5,m=2

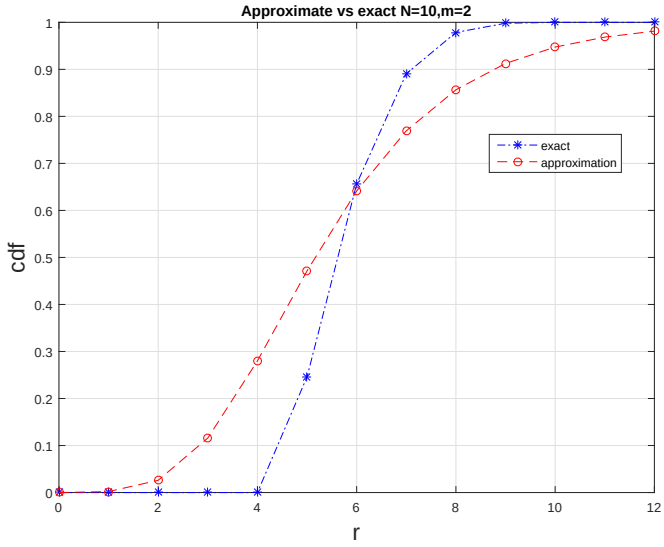

Figure 4.2: Exact maximum distribution and Gumbel approximation for N=10,m=2

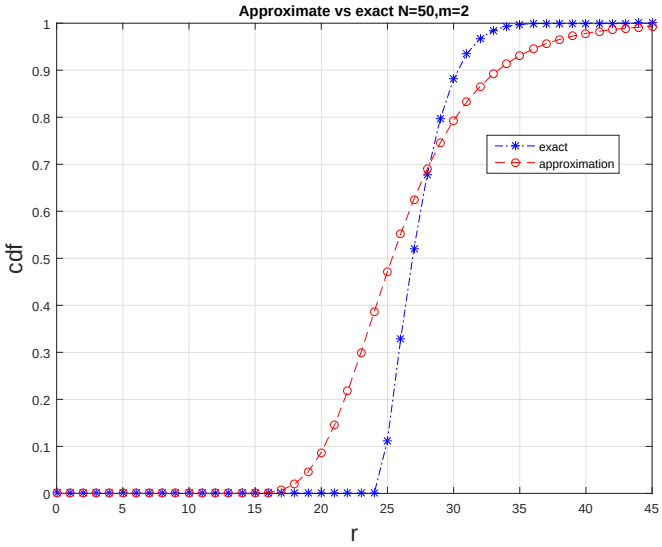

Figure 4.3: Exact maximum distribution and Gumbel approximation for N=50,m=2

It is clear that in the binomial case, even with quite big number of observations, approximating distributions differs quite significantly.
If we define the absolute and relative errors, as

$$abs = |p_{approx} - p_{ex}| \text{ and } rel = \frac{|p_{approx} - p_{ex}|}{p_{ex}},$$

the mean of the relative errors in the binomial case, even with 100 observations, is 0.16.
Table 4.1 reports the exact probabilities for the multinomial maximum and compares these with the approximate values for different combinations of number of observations and bins.

	N=25,m=5				N=100,m=25			
t	F(t) approx	F(t) exact	abs	rel	F(t) approx	F(t) exact	abs	rel
5	0,064	0,002	0,062	31,000	0,001	0,000	0,001	
6	0,292	0,147	0,145	0,986	0,121	0,022	0,099	4,500
7	0,576	0,490	0,086	0,176	0,552	0,240	0,312	1,300
8	0,781	0,769	0,012	0,016	0,846	0,596	0,25	0,419
9	0,895	0,913	0,018	0,020	0,954	0,837	0,117	0,140
10	0,951	0,972	0,021	0,022	0,987	0,945	0,042	0,044
	N=20,m=20				N=50,m=50			
2	0,185	0,121	0,064	0,529	0,013	0,003	0,01	3,333
3	0,864	0,705	0,159	0,226	0,767	0,370	1,071	1,073
4	0,987	0,948	0,039	0,041	0,984	0,848	0,161	0,160
5	0,998	0,993	0,005	0,005	0,999	0,976	0,023	0,024
6	1,000	1,000	0,000	0,000	1,000	0,997	0,003	0,003

Table 4.1: Approximation precision

Table 4.1 reveals quite significant difference in the probability values even for tail observations. For example, in the case of $N = 100, m = 25$, the $95th$ approximate percentile is 9, while the exact one is 11. This can influence the results of tests, based on maximum test statistics, quite strongly.
Figure 4.4 shows average absolute error for the different bin compositions.

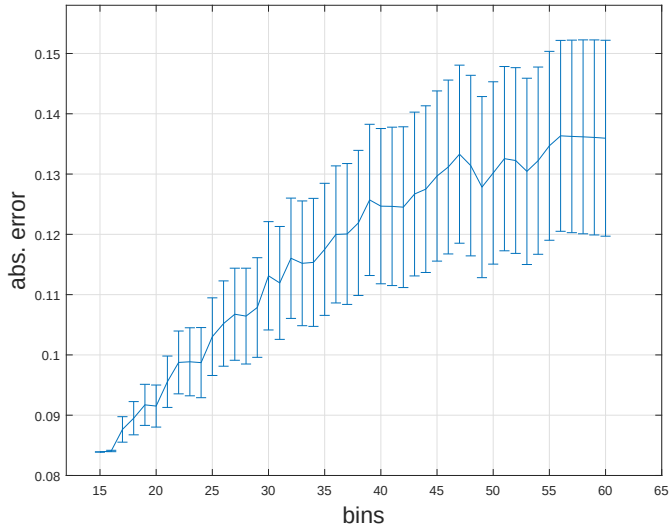

Figure 4.4: Average absolute error for the approximation of maximum distribution

The average absolute error remains around 0.11, even with growing amount of observations and bins, so we can conclude that the Gumbel approximation for the maximum distribution is not precise.

4.2. ACCURACY OF THE RANGE DISTRIBUTION APPROXIMATION

As was discussed in 2.1.1, the multinomial range distribution can be approximated by the range distribution of m i.i.d. standard normal variables.

$$P\left(\max_{1\le i\le m} n_i - \min_{1\le i\le m} n_i \le r\right) \xrightarrow{D} P\left(\max_{1\le i\le m} x_i - \min_{1\le i\le m} x_i \le (r+\delta_k)\sqrt{\frac{m}{N}}\right)$$

where

$$\delta_k = 1 \text{ for } k = 2, \delta_k = 0.5 \text{ for } k > 2$$

Unlike approximation to the maximum, approximate range distribution tend to perform better with growing number of observations, as it can be seen from Figures 4.5, 4.6.

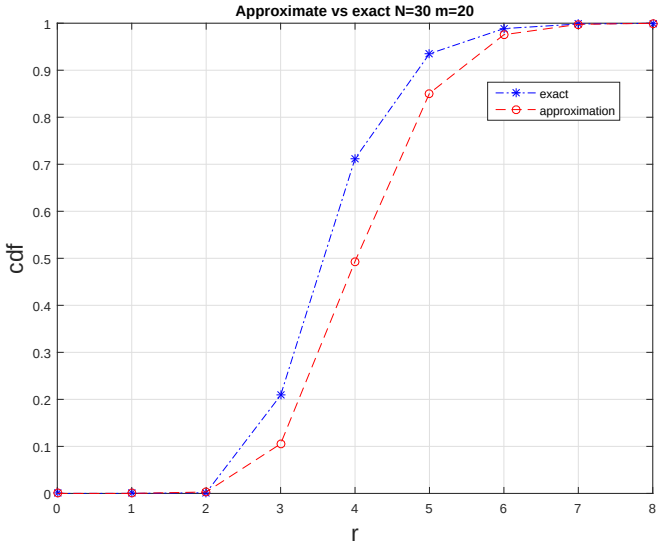

Figure 4.5: N=30,m=20

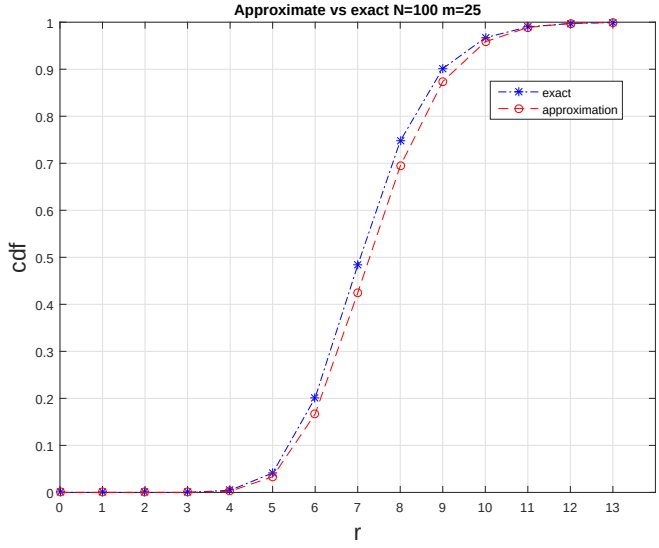

Figure 4.6: Exact range distribution and normal approximation for N=100,m=25

Normal approximation provides the reasonable degree of accuracy for not sparse cases. But in the sparse cases, the accuracy of the approximation tends to decrease, as shown in the Table 4.2.

	N=20,m=20				N=60,m=100			
t	F(t) approx	F(t) exact	abs	rel	F(t) approx	F(t) exact	abs	rel
1	0,000	0,000			0,000	0,000		
2	0,031	0,121	0,090	0,744	0,000	0,072	0,072	1,000
3	0,396	0,706	0,310	0,439	0,210	0,724	0,514	0,710
4	0,853	0,949	0,096	0,101	0,899	0,966	0,067	0,069
5	0,985	0,999	0,014	0,014	0,998	0,997	0,001	0,001
	N=60,m=60				N=50,m=60			
3	0,022	0,299	0,277	0,926	0,092	0,536	0,444	0,828
4	0,442	0,816	0,374	0,458	0,698	0,915	0,217	0,237
5	0,905	0,970	0,065	0,067	0,975	0,989	0,014	0,014
6	0,994	0,996	0,002	0,002	0,999	0,999	0,000	0,000
7	1,000	1,000	0,000	0,000	1,000	1,000	0,000	0,000

Table 4.2: Precision of the range approximation

Figure 4.7 shows that accuracy of the approximation fully depends on the ratio between the number of urns and the amount of the observations. Approximate distribution converges to the exact one for cases when $\frac{N}{m} \geq 3$.

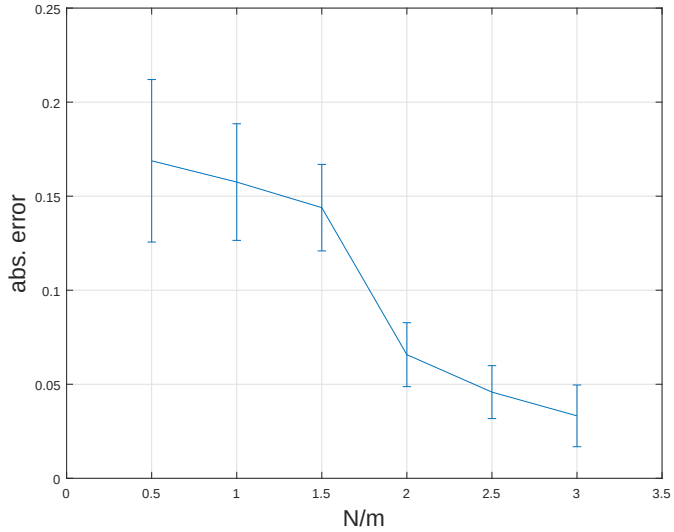

Figure 4.7: Average absolute error for the approximation of range distribution

5

STATISTICAL TESTS BASED ON THE MULTINOMIAL RANGE

Young(1962)[11] proposed a statistical test based on the range of multinomial sample and provided evidence of its power advantage against classical procedures, such χ^2 goodness-of-fits and the others. Young used approximating distribution for the test statistics, whose accuracy is strongly dependent both on the number of observations and the amount of bins, as was discussed in the previous chapter.
The goal of this chapter is to compare performance of the tests based on exact and approximate distributions for the test statistics under different specified alternatives. But first we start by discussing the fact, that the range based statistic has a very desirable property for a test statistics, such as unbiasedness.

5.1. UNBIASEDNESS OF THE RANGE BASED STATISTICS

According to Lehmann[6], a test procedure ϕ is called unbiased if the following conditions hold:

Definition 1. *Assume, a sample of observations $X = [X_1, \dots, X_n]$ from the known distribution $F(\cdot, \theta)$ is given. Consider the parametric hypothesis*

$$H_0 : \theta \in \Theta_0$$

vs

$$H_1 : \theta \in \Theta_1$$

where $\Theta_0 \cap \Theta_1 = \emptyset$ *and* $\Theta_0 \cup \Theta_1 = \Theta$.
Then, for a defined significance level α*, test procedure* ϕ_α *is unbiased if*

$$\begin{cases} \inf_{\theta^* \in \Theta_1} P\left(\phi_\alpha \text{ rejects } H_0 | \theta = \theta^*\right) \geq \alpha \\ \sup_{\theta^* \in \Theta_0} P\left(\phi_\alpha \text{ rejects } H_0 | \theta = \theta^*\right) \leq \alpha \end{cases} \tag{5.1}$$

Compiani[2], in his master thesis, proved the following theorem, regarding unbiasedness of the multinomial range statistics.

Theorem 1. *Suppose we are given a sample* $[N_1, \ldots, N_m]$ *from the multinomial distribution with parameters* N *and* $p = [p_1, \ldots, p_m]$. *Suppose we want to test the following parametric null hypothesis:*

$$H_0 : p = \left[\frac{1}{m}, \ldots, \frac{1}{m}\right]$$

against

$$H_1 : p \neq \left[\frac{1}{m}, \ldots, \frac{1}{m}\right]$$

Let

$$T = \max_{1 \leq i \leq m} n_i - \min_{1 \leq i \leq m} n_i$$

Then, for a fixed significance level α *and critical value* c_α*, test* ϕ_α *of a form:*

$$\phi_\alpha = \begin{cases} \text{Do not reject } H_0 \text{ if } T < c_\alpha \\ \text{Reject } H_0 \text{ if } T > c_\alpha \end{cases}$$

is unbiased.

5.2. RANDOMIZED STATISTICAL TESTS

One of the parameters of the statistical test, which should be defined beforehand, is the significance level α. That is to say, for some null hypothesis H_0 and alternative hypothesis H_1 we define the test function ϕ_α, using the test statistics $T = T(N_1, \ldots, N_m)$, as:

$$\phi_\alpha(N_1, \ldots, N_m) = \begin{cases} 0 \text{ if } T(N_1, \ldots, N_m) \notin \Delta_\alpha \\ 1 \text{ if } T(N_1, \ldots, N_m) \in \Delta_\alpha \end{cases} \tag{5.2}$$

where Δ_α is a subset of the support of test statistics T, such that

$$P\left(T(N_1, \ldots, N_m) \in \Delta_\alpha | H_0\right) = \alpha \tag{5.3}$$

In the following sections, we will use the multinomial range T_{range} as the test statistics. Range of the multinomial distribution is a discrete value, so there might be a case, that no such subset Δ_α exists, which satisfies 5.3. This leads to the fact that test of a form 5.2 will not have significance level α as desired.
In order to make significance level of the test exactly α we can modify the test function and introduce randomized statistical test. Consider a test function ϕ_α such that:

$$\phi_\alpha(N_1,\dots,N_m) = \begin{cases} 0 \text{ if } T(N_1,\dots,N_m) < \bar{c_\alpha} \\ \kappa \text{ if } T(N_1,\dots,N_m) = \bar{c_\alpha} \\ 1 \text{ if } T(N_1,\dots,N_m) > \bar{c_\alpha} \end{cases} \tag{5.4}$$

where $\kappa \in (0,1)$ satisfies following equation:

$$P\left(T(N_1,\dots,N_m) = \bar{c_\alpha} | H_0\right)\theta + P\left(T(N_1,\dots,N_m) > \bar{c_\alpha} | H_0\right) = \alpha \tag{5.5}$$

Equation 5.5 shows that the test defined in 5.4 has significance level exactly equal to α, as desired. In the randomized test 5.4, in the case of $T(N_1,\dots,N_m) = \bar{c_\alpha}$ we reject the null hypothesis with probability κ.
Critical value $\bar{c_\alpha}$ is the largest integer r, such that

$$P\left(T(N_1,\dots,N_m) \geq r | H_0\right) > \alpha,$$

therefore

$$\theta = \frac{\alpha - P\left(T(N_1,\dots,N_m) \geq \bar{c_\alpha} + 1 | H_0\right)}{P\left(T(N_1,\dots,N_m) = \bar{c_\alpha} | H_0\right)}.$$

5.3. GOODNESS-OF-FIT TESTS

Many goodness-of-fit tests can be reduced to the testing parametric hypothesis for a multinomial distribution. Suppose, we are given a set of i.i.d. observations $X_1,\dots,X_n$ from unknown distribution F, with null hypothesis $H_0 : F = F_0$. Support of the null distribution F_0 can be partitioned into m equiprobable non-overlapping sub-intervals $B_1,\dots,B_m$. We define random variables $N_1,\dots,N_m$ as counters of the events in the intervals $B_1,\dots,B_m$ correspondingly

$$N_i = \sum_{j=1}^{n} I(X_j \in B_i) \text{ for } 1 \leq i \leq m$$

Then, under H_0, auxiliary variables $N_1,\dots,N_m$ are distributed as follows:

$$N \sim \text{Multinomial}(n,p) \text{ with } p = \left[\frac{1}{m},\dots,\frac{1}{m}\right]$$

Therefore, the test becomes

$$H_0 : p = p_0$$

vs

$$H_1 : p \neq p_0$$

Remark. *If we define* $Y_i \equiv F_0(X_i)$*, by the probability integral transform, then* $Y_i \sim Uniform(0,1)$ *under* H_0 *for all* i*. This implies that, for any continuous theoretical distribution* F_0*, we can always transform the available data as shown above and obtain observations* Y_i *in the interval* $(0,1)$*. Then, sub-intervals* B_i*, defined as* $B_i = [\frac{i-1}{m}, \frac{i}{m}]$*, are equiprobable under distribution of* Y_i*. So, without loss of generality, we can always consider partition of a unit interval into m sub-intervals with length* $\frac{1}{m}$*.*

One of the important issues, in the construction of the goodness-of-fit test of this kind, is how to select the number of bins. Read and Cressie[10] provided the review of relevant literature on this topic. However, we will leave this problem outside of the scope of this work and try different compositions of bins and compare performance of the tests relying on approximate and exact distributions.

5

5.3.1. NORMAL VS LOGNORMAL

As an example, we will provide the power comparison of the exact and approximate tests under lognormal distribution as the alternative. The null distribution, that we consider, is the normal distribution with the mean $\mu = 1.3$ and the standard deviation $\sigma = 0.25$. Alternative distribution is the lognormal with zero mean and standard deviation $\sigma_{LN} = 0.25$. Densities of the distributions are presented in Figure5.1.

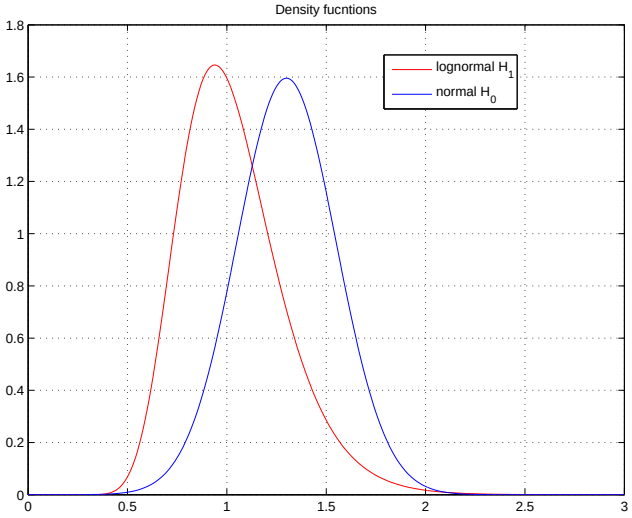

Figure 5.1: Normal(1.3,0.25) vs Lognormal(0,0.25)

Power of the exact and approximate test is computed by Monte-Carlo simulation. For each composition of the number of bins and the number of observations 3000 simulations were made.

The choice of the maximum sample size being not greater than 50 is motivated by the following fact. For all input values of N, the power of the two tests reaches the ceiling value of 1 for $N \leq 50$. Therefore comparing tests for the sample sizes bigger than 50 does not seem illustrative.

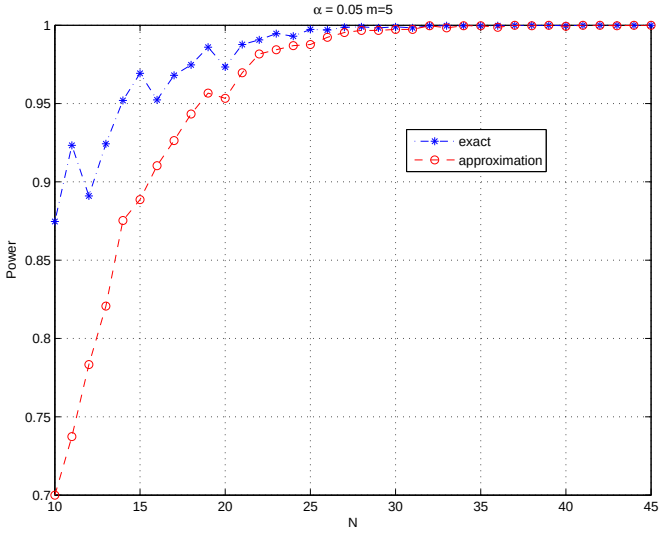

Figure 5.2: Power comparison for the case of 5 bins

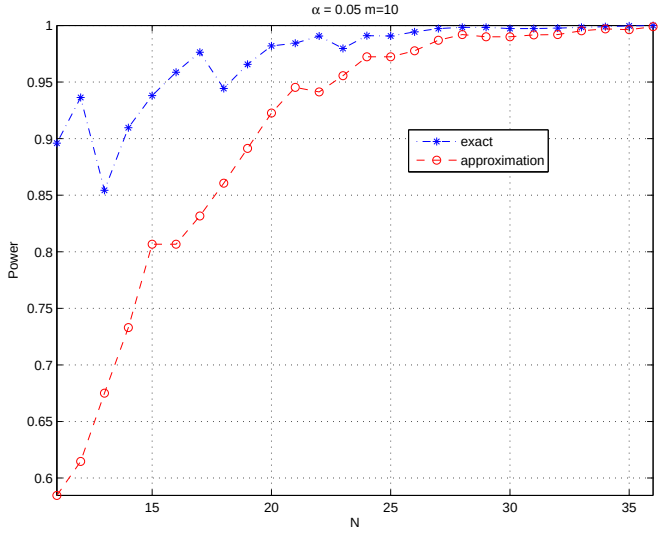

Figure 5.3: Power comparison for the case of 10 bins

Inspection of the Figures 5.2 - 5.3 reveals the power advantage of the exact distribution based test for all test sample size in case of the given alternative. In the case of the 5 bins, Figure 5.2, exact test outperforms the approximate test for sample sizes smaller than 25 and at least as powerful as approximate for bigger samples. Also, the Figures 5.2 - 5.3 suggest that the difference in the performance tend to grow with the growing amount of bins.

Figures 5.4 - 5.3 compare power of the tests for $m = 30$ and $m = 50$, respectively.

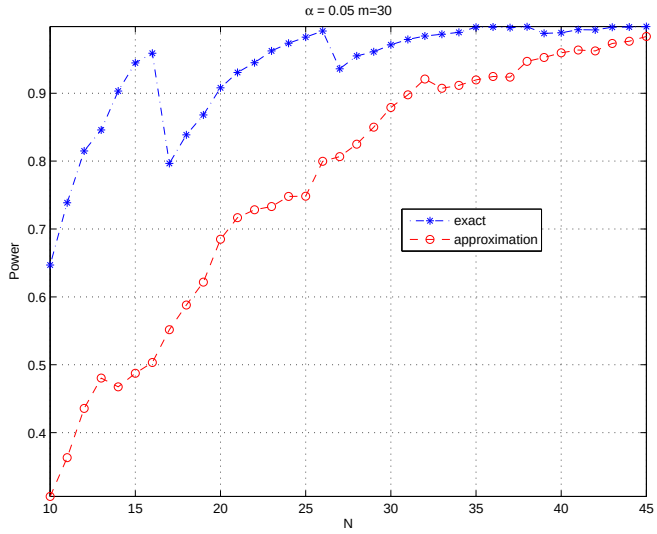

Figure 5.4: Power comparison for the case of 30 bins

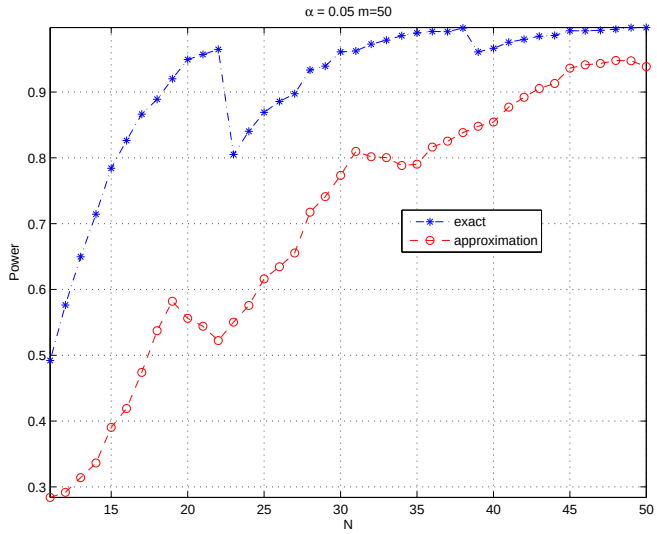

Figure 5.5: Power comparison for the case of 50 bins

Again, Figures 5.4 - 5.5 show that the exact distribution based test is more powerful than approximate based test for all values of N.
On the basis of Figures 5.2 - 5.5 we can conclude that the exact test performs better uniformly over all sample sizes, independently to the amount of bins. We also see, that the difference in power is increasing for the greater amount of bins.

5.3.2. POWER UNDER OTHER ALTERNATIVES

In order not to limit our attention to some particular distribution and provide more general comparison of the tests, in this section we study the performance of the test under the following class of the alternatives H_1:

$$H_1 : p_i = \frac{1}{m} + \frac{1}{\sqrt{N}} c_i \quad \text{,where } i = 1, \dots, m$$

With the c's being fixed set of constants, such that $\sum c_i = 0$, set of alternatives H_1 converges to the H_0 with $O\left(n^{-\frac{1}{2}}\right)$. This family of the alternatives was proposed by Cochran[1] as a method of strengthening chi-squared test. We again use Monte-Carlo simulation for different combinations of N and m to provide evidence of the power advantage of the exact test. For every pair of N and m we look at the behaviour of the test for different values of $\sum c^2$.
We start by investigating the behaviour on the small samples. In the following Figures, horizontal axis represent the value of $\sum c^2$.

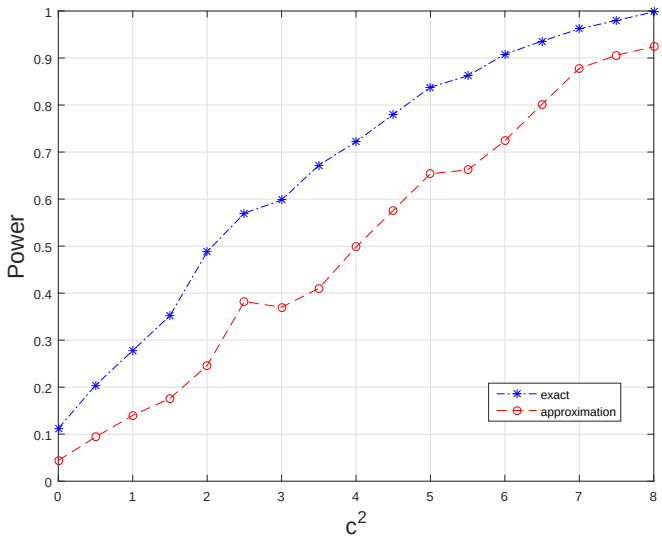

Figure 5.6: Power comparison for the case of 2 bins and N=10

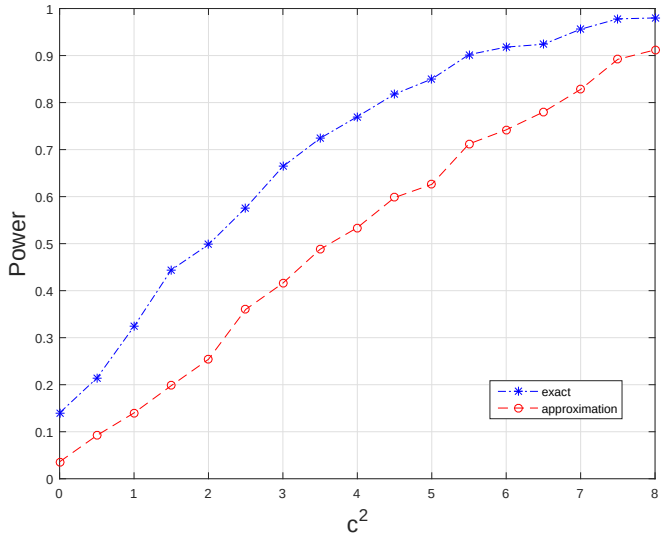

Figure 5.7: Power comparison for the case of 2 bins and N=12

As Figures 5.6 - 5.7 show, the exact test is more powerful than approximate one uniformly over all values of c^2. This is consistent with the remarks about behaviour of the tests that were made in previous section.
To support the advantage of the exact test and show independence from sample size and urn composition, we provide Figures 5.8 - 5.10

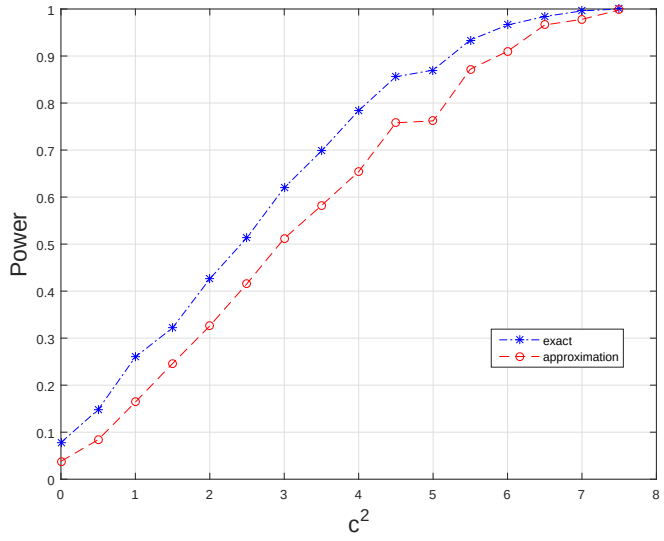

Figure 5.8: Power comparison for the case of 4 bins and N=30

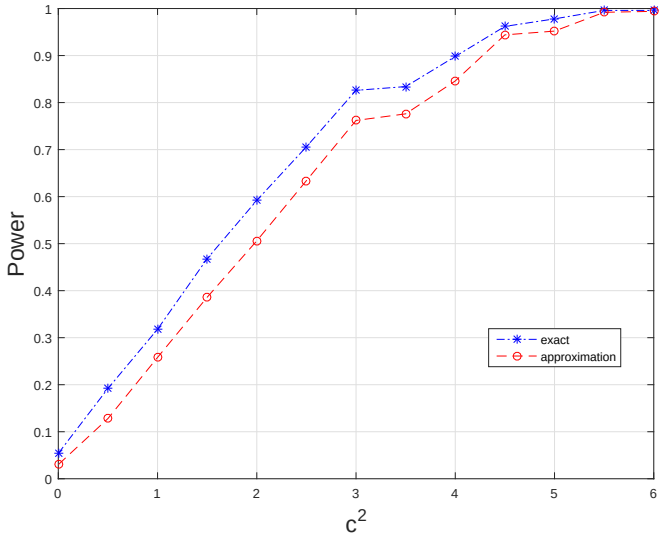

Figure 5.9: Power comparison for the case of 8 bins and N=160

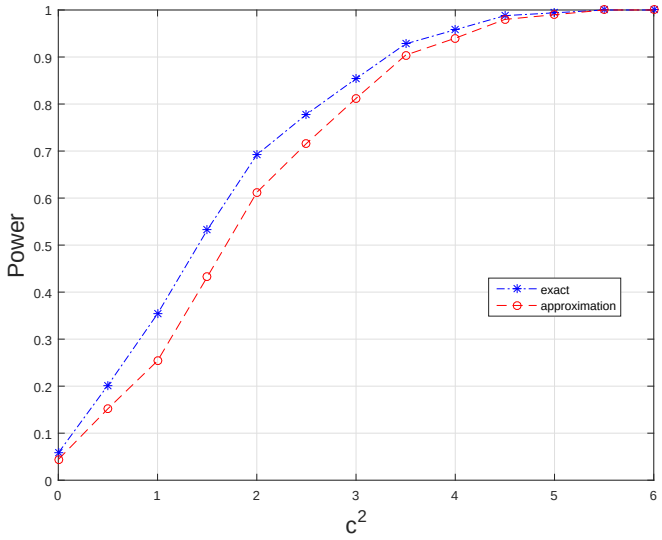

Figure 5.10: Power comparison for the case of 10 bins and N=160

Figures 5.8 - 5.10 show as well, that the gain of power for bigger samples becomes smaller, which can be explained by the growing ratio between amount of observations and number of bins due to the convergence of the approximate distribution to the exact one, as was explained in the Chapter 4.

5.4. TEST FOR THE HOMOGENEOUS POISSON PROCESS

The same kind of test could be applied for testing homogeneous Poisson process. Suppose, $M(t)$ - is homogeneous Poisson process. If we again partition time domain into m non-overlapping equal-length sub-intervals $B_1, \ldots, B_m$ and define number N_i as

$$N_i = \sum_{j=1}^{n} I(X_j \in B_i) \text{ for } 1 \le i \le m$$

Then distribution of random variables N_i is again equiprobable multinomial.
In order to study performance of the test non-homogeneous Poisson process should be simulated. We will use time-scale transformation of homogeneous Poisson process to generate NHPP. This method is based on simulation of the homogeneous Poisson process of the rate one.
Denote $N_1(t)$ as rate one HPP. Then inter arrival times T are distributed exponentially with intensity 1

$$P(T \ge t) = \exp(-t)$$

which can be rewritten

$$P(\Lambda^{-1}(T) \ge t) = \exp(-\Lambda(t)) \tag{5.6}$$

Let's also denote $\Lambda(x)$ - integrated rate function, which is nothing more than parameter of Poisson distribution of the number of points in any finite interval $(0, x]$, assuming that $\Lambda(0) = 0$. Then inter arrivals times T' for HNPP are distributed:

$$P(T' \ge t) = \exp(-\Lambda(t)) \tag{5.7}$$

From equations 5.6 - 5.7 we can conclude, that $T_1^{'}, T_2^{'}, \ldots$ are points of the NHPP with integrated rate function $\Lambda(t)$ if $T_1 = \Lambda(T_1^{'}), T_2 = \Lambda(T_2^{'})$ are points of a HPP with intensity 1. Hence, we can simulate NHPP simply generation exponential variables and taking inverse Λ^{-1} of the generated time instants.
Power comparison of the tests is presented in the Table 5.1, which reveals advantage of the exact test.

$\lambda = 2 + 0.01 * t$,m=30, T=20		$\lambda = 0.3 * t$, m = 20, T = 200	
Exact	Approximate	Exact	Approximate
0.184	0.037	0.427	0.371
$\lambda = 0.05 * t$, m= 15		$\lambda = 2 + sin(2\pi t)$,m=30	
Exact	Approximate	Exact	Approximate
0.133	0.059	0.213	0.084

Table 5.1: Power of the exact and approximate test in case of Poisson process

In the case of testing HPP, initial partition of the time domain is very important. For

example, we consider a case, in which the real intensity follows the harmonic function:

$$\lambda = 5\sin(\pi * t)$$

Figure 5.11 shows power of the tests with time domain split into 10 or 20 disjoint intervals. Blue and red lines represent case of 10 bins, black and green - 20.

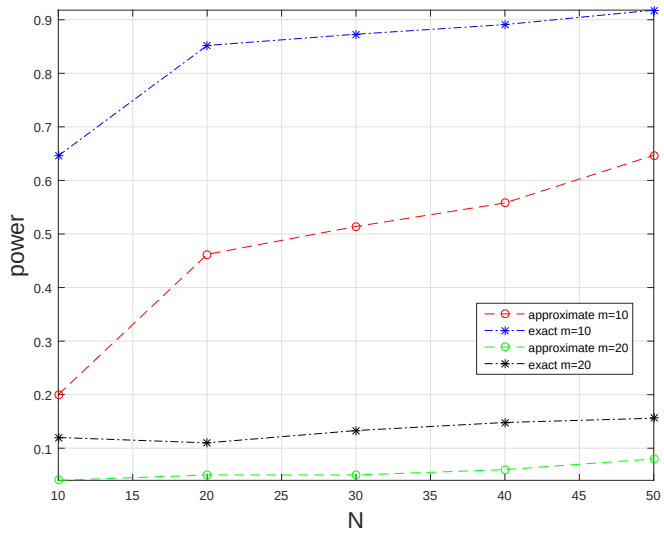

Figure 5.11: Power of the tests in case of harmonic intensity

From Figure 5.11, we see again the power advantage of the exact test. Huge power gain between cases with 10 and 20 bins can be explained by the following fact. If we split time domain into 20 non-overlapping equal-length intervals, each of them covers the period of intensity function, which averages out the effect of the harmonic function.
So,the exact range based test shows better performance than the approximate test in Poisson process application, but the performance itself is highly dependent on the initial urn composition.

5.5. A SIMPLE APPLICATION TO DISEASE CLUSTERING

We conclude by showing a simple application of the test to the issue of disease clustering. The problem of disease clustering is frequently of interest to epidemiologists and biomedical statisticians. From a statistical point of view, disease clustering usually have been approached as hypothesis testing problems. The main interest is to test a null hypothesis of no clustering, i.e., a common rate of disease across the study region, against an alternative hypothesis of clusters presence.
Numerous ways to construct such tests were proposed over the last years. One of the

approaches is based on the multinomial distribution. In this case, study region must be separated into equal clusters, in terms of the population. Then, obviously, in case of common rate of the disease across the region, the number of registered events should follow equiprobable multinomial distribution.

We will demonstrate this approach using well-known epidemiological dataset of diagnosed leukaemia cases over 8 counties in the upstate of New York. This data originated from the New York Sate Cancer Registry and was gathered during the 5-year period 1978-1982, with totally 584 individuals diagnosed with leukaemia over population of approximately 1 million people. Original dataset contains spatial information about registered events already split into 790 initial clusters with different population. Their distribution is shown on the Figure 5.12

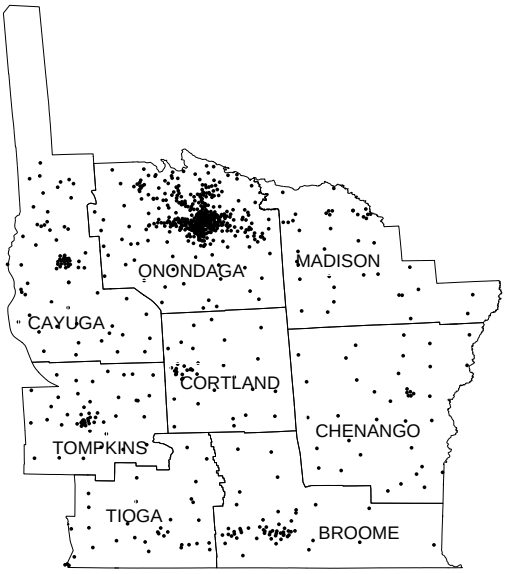

Figure 5.12: Distribution of leukaemia cases over 8 counties in the New York State

In order to perform such a test, we have to cluster datapoints into groups of approximately equal population. In a perfect case, the population of new clusters should be equal, but for the leukaemia dataset it is impossible due to the original grouping. To our knowledge, there is no existing unique algorithm for the equal size spatial clustering. For this case, we have followed next procedure:

1. Define the number of clusters
2. Select range for the cluster population
3. Create clusters, satisfying population range, around points with anomaly high initial population

4. Use k-means algorithm to create clusters for the rest of observations, initializing clusters with points of the highest population

5. Trade observations between clusters based on the population and distance, until population requirements are satisfied

Using this heuristics, we were able to create clusters with approximately equal population (some of the points were reassigned to the different cluster on the post-processing stage). In Figure 5.13, 32 cluster centroids are presented. In this case, population of the groups varies from 36036 to 39528, therefore average probability in cluster is 3.48 % with standard deviation of 0.08%, and we can assume that clusters are equiprobable.

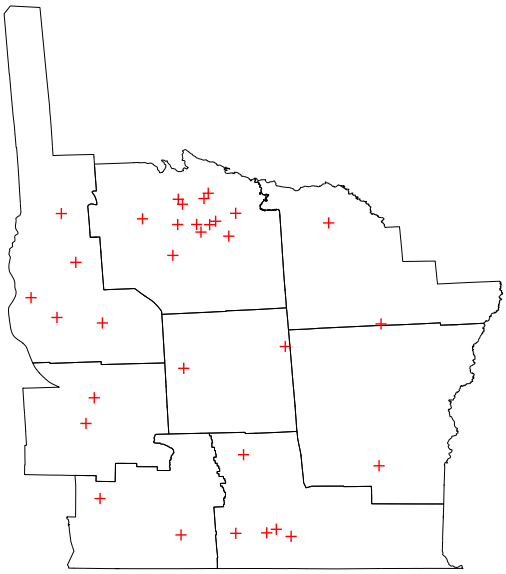

Figure 5.13: Centroids of the 32 clusters

For this grouping, the maximum amount of the individuals diagnosed with the leukaemia within one cluster is 39 versus the minimum of 3 cases registered. Total number of cases registered in each cluster is presented in Table5.2. Using multinomial test based on the range of the sample, null hypothesis is rejected for the 99% confidence level, since 99% quantile of the range distribution in the case of 584 events and 32 bins is 25. The null hypothesis is also rejected for the case of 25 clusters, which supports previous research about leukaemia clusters for this dataset, which are believed to be caused by hazardous waste sites in Broome, Onondaga, Cayuga and Cortland counties.

Cluster	Registered	Cluster	Registered	Cluster	Registered	Cluster	Registered
1	34	9	18	17	25	25	14
2	28	10	14	18	20	26	12
3	13	11	17	19	9	27	24
4	23	12	39	20	21	28	17
5	23	13	20	21	24	29	5
6	20	14	17	22	12	30	4
7	18	15	14	23	31	31	3
8	27	16	13	24	24	32	3

Table 5.2: Number of patients diagnosed leukaemia in each of 32 clusters

We want to notice that in this particular case there is no difference in using test based on exact or approximate distribution, since for $N = 584$ approximate distribution converges to the exact one. For smaller datasets the use of the exact test is preferable.
This method may not allow to identify the exact disease clusters, but rejection of the null hypothesis is a strong indication of their presence and might be a powerful tool for the early stages of the research.

6

CONCLUSIONS

In this thesis, we have considered different class of tests used to tackle goodness-of-fit problem: multinomial range test based on the exact distribution of the statistics. Using results from the work of Rapperport, we developed a new algorithm for computing the exact distribution of the multinomial range. Our algorithm relies on the tree representation of the outcome of a multinomial experiment. Unlike the other algorithms, available in the literature, our version is uniform for any combination of the number of observations and bins. Presented algorithm require no modification for every new input. As far as we know, this is a new contribution to the algorithms of this family.

In fact, we also developed new algorithm for the distribution of multinomial minimum along with the correction of the previously developed algorithm for the distribution of the sum of multinomial highest order statistics. All the algorithms were developed under hypothesis of the equiprobable multinomial, so future research can be done to create modifications for not equiprobable cases.

Given this result, we studied the behaviour of the approximating distributions of the maximum and the range. We discovered that the accuracy of the approximations is strongly depended on the combination of the number of urns and amount of observations. This result appears to be interesting, since approximate distribution is widely used for the hypothesis testing in applications.

Finally, we focused on the hypothesis testing problem. Previous results from the literature provide evidence of power advantage of the range based test among well-known goodness-of-fit test statistics, such χ^2-test and the others. It was also proved before by Compiani, that the range based statistics is unbiased. We performed our simulations to

compare tests based on the exact and the approximate distributions. We found that exact test is more powerful than the approximate for different classes of alternatives. We first showed the power advantage in the case of the particular distributions, but then also discovered same behaviour under closer class of alternatives. Power advantage of the exact test holds independently from the sample size and the urn composition.
We also provided simulation results for the particular applications. In application to the homogeneous Poisson process, exact test showed again power advantage, but for some alternatives, the power value is highly dependent on the partition of the time domain.
For disease clustering problem, we rejected the hypothesis of a common rate of leukaemia over upstate of New York. This results is supported by previous research. In this case, results of the exact and the approximate tests were identical due to the large size of the sample.

A

MATLAB CODES

A.1. ALGORITHM FOR THE MAXIMUM

```
function [P] = max_order_statistic(t,N,I)
%% Script calculates probability of highest order statistic being <=
    t
% Under equiprobable multinomial
% Input : N – number of balls
%   I – number of urns(cells)
%  t – argument of cdf
P = 0;
if t == 0 && N~=0
    P = 0;
    return
end
if t==0 && N==0
    P=1;
    return
end
if t>=N
    P = 1;
    return
end
```

A

```
if N==0 && I~=0
    P=1;
    return
end
if N==0 && I==0
    P=1;
    return
end
common_term = gammaln(N+1)+gammaln(I+1)-N*log(I);
switch t
    case 1
        if I>=N
            P = exp(gammaln(I+1)-gammaln(I-N+1)-N*log(I)); %
                explicit calculation of P(n<1><=1)
              calc = [calc;t,N,I,P];
        else
            P = 0;
            calc = [calc;t,N,I,P];
        end
    otherwise
        % range of summation for q
        LowSum = max(0,N-t*I+I);
        UpSum = floor(N/t);
        for q = LowSum:UpSum
            summ_term = (-q*gammaln(t+1)-gammaln(q+1)-gammaln(I-q+1)
                -gammaln(N-t*q+1));
            if I==q
                summ_term_nominator = 0;
            else
                summ_term_nominator = (N-t*q)*log(I-q);
            end
            coef = exp(common_term+summ_term+summ_term_nominator);
                [temp]  = max_order_statistic(t-1,N-t*q,I-q);

            P = P + coef*temp;

        end
end
```

```
end
```


A.2. SUM OF J HIGHEST ORDER STATISTICS

```
function [ P ] = highest_order_statistics( t,N,I,J )
%% Probability of the sum of the first J highest order statistics
    being smaller than t
%   P(sum(1:J)n<i> <= t : N,I) under H1 – equiprobable multinomial
% Input: N– number of trials
% I – number of cells(urns)
% t – argument of cdf
% J – number of highest order statistics
if J>I
    error('J should be smaller or equal than I')
end
if J==I && (t<N )
    error(' Total sum is every time equal to N')
end

if t == 0 && N~=0
    P = 0;
    return
end
if t==0 && N==0
    P=1;
    return
end
if t>=N
    P = 1;
    return
end

%% first term if n<1> <= t/J
P = max_order_statistic(floor(t/J),N,I);
%% recursive summation over all possible options
for sum_depth = 1 : J–1
    rangeArg = [];
    cur_depth = 1;
```

A

```
    P = P + recursive_sum(t,N,I,J,sum_depth,cur_depth,rangeArg);
end
end
function S = recursive_sum(t,N,I,J,sum_depth,cur_depth,rangeArg)
%% auxiliary function for calculating sum of nested loops
S = 0;
if cur_depth <= sum_depth % either increment summation depth or
    calculate the term
    if cur_depth == 1
        cur_range(1) = floor(t/J+1);
        cur_range(2) = t-sum_depth+1;
    else
        cur_range(1) = floor((t - sum(rangeArg(1:cur_depth-1)))/(J-
            cur_depth+1)+1);
        cur_range(2) = min( rangeArg(cur_depth-1), t - sum(rangeArg
            (1:cur_depth-1)));
    end
    for r=cur_range(1):cur_range(2)
       rangeArg(cur_depth) = r;
        S = S + recursive_sum(t,N,I,J,sum_depth,cur_depth+1,rangeArg
            );
    end
else
    prob_arg = floor( (t-sum(rangeArg))/(J-sum_depth));
    temp_p = max_order_statistic(prob_arg, N - sum(rangeArg),I-
        sum_depth);
    common_term = gammaln(N+1) + gammaln(I+1) - N*log(I);
    coef = (N-sum(rangeArg))*log(I-sum_depth) - gammaln(I-sum_depth
        +1) - gammaln( N - sum(rangeArg)+1);
    for k=1:numel(rangeArg)
        coef = coef - gammaln(rangeArg(k)+1);
    end
    equal_statistics = unique(rangeArg);
    for k=1:numel(equal_statistics)
        temp = numel(find(rangeArg == equal_statistics(k)));
        coef = coef   - gammaln(temp+1);
    end
    S  = temp_p*exp(common_term+coef);
```

```
66
67 end
68 end
```

A.3. ALGORITHM FOR THE MINIMUM

```
1  function P = smallest_order_value( t,N,I )
2  %% Function to calculate the probability of smallest order statistic
       to be >= than t for equiprobable multinomial
3  % Input:
4  % t  - argument of "survival" function
5  % N - number of balls
6  % I  - number of cells
7
8  P = 0;
9  % add for exceptions for "naive" input
10 if t>floor(N/I)
11     P=0;
12     return
13 end
14 if t==0
15     P=1;
16     return
17 end
18     aux = max_for_min(N,N,I,calc,t);
19     P = P + aux;
20 end
21
22 function [aux,calc] = max_for_min(t_max,N,I,calc,t)
23 aux = 0;
24 if t_max<t
25     if N==0 && I==0
26         aux=1;
27         return
28     else
29         aux = 0;
30         return
31     end
```

A

```
else
    if N==0 && I == 0
        aux=1;
        return
    end
    if t_max==1
        if I==N
            aux = exp(gammaln(I+1)-gammaln(I-N+1)-N*log(I)); %
                explicit calculation of P(n<1><=1)
            return
        else
            aux= 0;
            return
        end
     end
    if N==0 && I~=0
        aux=0;
        return
    end
    common_term = gammaln(N+1)+gammaln(I+1)-N*log(I);
    LowSum = max(0,N-t_max*I+I);
    UpSum = floor(N/t_max);
    for q = LowSum:UpSum
        summ_term = (-q*gammaln(t_max+1)-gammaln(q+1)-gammaln(I-q+1)
            -gammaln(N-t_max*q+1));
        if I==q
            summ_term_nominator = 0;
        else
            summ_term_nominator = (N-t_max*q)*log(I-q);
        end
        coef = exp(common_term+summ_term+summ_term_nominator);
        [temp]  = max_for_min(t_max-1,N-t_max*q,I-q,t);
        aux = aux + coef*temp;
    end
end
end
```

A.4. ALGORITHM FOR THE RANGE

```
function [P] = range_probability_ver2( t,N,I )
%% Function to calculate the probability of the range to be < =than
    t
%for equiprobable multinomial
% Input:
% r  – argument of "cdf" function
% N – number of balls
% I  – number of cells
P = 0;
t= floor(t);
if t>N
    P=1;
    return
end

[P]=max_order_statistic(t,N,I);
prev=[t,N,I];
for t_max = t+1:N
    [aux] = max_for_range(t_max,N,I,prev,t);
    P = P + aux;
    prev = [t_max,N,I];
end
end

function [aux] = max_for_range(t_max,N,I,prev,t)

aux=0;
if [t_max,N,I]==prev
    aux = 0;
    return
end
if prev(1)+1-t_max>t
    if N==0 && I==0
        aux=1;
        return
    else
        aux = 0;
        return
```

A

```
    end
else
    if N==0 && I == 0
        aux=1;
        return
    end
    if t_max==1
        if I==N
            aux = exp(gammaln(I+1)-gammaln(I-N+1)-N*log(I)); %
                explicit calculation of P(n<1><=1)
            return
        else
            aux= 0;
                        return
        end
    end
    if N==0 && I~=0
        aux=0;
                return
    end

    common_term = gammaln(N+1)+gammaln(I+1)-N*log(I);
    LowSum = max(0,N-t_max*I+I);
    UpSum = floor(N/t_max);
    for q = LowSum:UpSum
        summ_term = (-q*gammaln(t_max+1)-gammaln(q+1)-gammaln(I-q+1)
            -gammaln(N-t_max*q+1));
        if I==q
            summ_term_nominator = 0;
        else
            summ_term_nominator = (N-t_max*q)*log(I-q);
        end
        coef = exp(common_term+summ_term+summ_term_nominator);
        [temp]  = max_for_range(t_max-1,N-t_max*q,I-q,prev,t);
        aux = aux + coef*temp;
    end
end
end
```

B

VALUES OF THE EXACT DISTRIBUTIONS

B.1. DISTRIBUTION OF THE MAXIMUM

The following table provides critical values for the maximum distribution. For example, in the case with $N = 10$ and $m = 5$

$$P(\max n_i \leq 3 : N = 10, m = 5) = 0.433$$

r		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
N	m															
5	5	0,710	0,966	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
10	5	0,012	0,433	0,836	0,968	0,996	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
15	5	0,000	0,006	0,284	0,700	0,910	0,979	0,996	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	5	0,000	0,000	0,003	0,200	0,584	0,840	0,950	0,987	0,997	0,999	1,000	1,000	1,000	1,000	1,000
25	5	0,000	0,000	0,000	0,002	0,147	0,490	0,769	0,913	0,972	0,992	0,998	1,000	1,000	1,000	1,000
30	5	0,000	0,000	0,000	0,000	0,001	0,113	0,415	0,701	0,872	0,953	0,984	0,995	0,999	1,000	1,000
10	10	0,396	0,873	0,984	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	10	0,000	0,127	0,603	0,889	0,976	0,996	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	10	0,000	0,000	0,049	0,394	0,753	0,923	0,980	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000
40	10	0,000	0,000	0,000	0,022	0,259	0,617	0,848	0,950	0,985	0,996	0,999	1,000	1,000	1,000	1,000
50	10	0,000	0,000	0,000	0,000	0,011	0,173	0,498	0,765	0,908	0,968	0,990	0,997	0,999	1,000	1,000
60	10	0,000	0,000	0,000	0,000	0,000	0,006	0,119	0,401	0,681	0,857	0,944	0,980	0,993	0,998	0,999
15	15	0,219	0,785	0,966	0,996	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	15	0,000	0,037	0,432	0,813	0,955	0,991	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
45	15	0,000	0,000	0,009	0,221	0,620	0,867	0,962	0,990	0,998	1,000	1,000	1,000	1,000	1,000	1,000
60	15	0,000	0,000	0,000	0,002	0,114	0,451	0,755	0,911	0,972	0,992	0,998	1,000	1,000	1,000	1,000
75	15	0,000	0,000	0,000	0,000	0,001	0,061	0,321	0,638	0,844	0,942	0,981	0,994	0,998	1,000	1,000
90	15	0,000	0,000	0,000	0,000	0,000	0,000	0,034	0,228	0,530	0,769	0,902	0,963	0,987	0,996	0,999
20	20	0,121	0,706	0,949	0,993	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	20	0,000	0,011	0,309	0,742	0,933	0,986	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	20	0,000	0,000	0,001	0,123	0,510	0,814	0,944	0,985	0,997	0,999	1,000	1,000	1,000	1,000	1,000
80	20	0,000	0,000	0,000	0,000	0,050	0,329	0,671	0,874	0,958	0,988	0,997	0,999	1,000	1,000	1,000
100	20	0,000	0,000	0,000	0,000	0,000	0,021	0,207	0,532	0,785	0,916	0,971	0,991	0,997	0,999	1,000
120	20	0,000	0,000	0,000	0,000	0,000	0,000	0,010	0,129	0,411	0,689	0,862	0,945	0,980	0,993	0,998
25	25	0,067	0,634	0,931	0,991	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
50	25	0,000	0,003	0,222	0,678	0,912	0,981	0,996	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
75	25	0,000	0,000	0,000	0,069	0,419	0,764	0,926	0,980	0,995	0,999	1,000	1,000	1,000	1,000	1,000
100	25	0,000	0,000	0,000	0,000	0,022	0,240	0,596	0,837	0,945	0,983	0,995	0,999	1,000	1,000	1,000
125	25	0,000	0,000	0,000	0,000	0,000	0,007	0,133	0,443	0,730	0,891	0,961	0,987	0,996	0,999	1,000
150	25	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,073	0,319	0,617	0,823	0,928	0,973	0,991	0,997
30	30	0,037	0,570	0,914	0,988	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	30	0,000	0,001	0,159	0,619	0,891	0,975	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
90	30	0,000	0,000	0,000	0,039	0,345	0,717	0,908	0,975	0,994	0,999	1,000	1,000	1,000	1,000	1,000
120	30	0,000	0,000	0,000	0,000	0,010	0,175	0,529	0,802	0,931	0,979	0,994	0,998	1,000	1,000	1,000
150	30	0,000	0,000	0,000	0,000	0,000	0,003	0,086	0,369	0,678	0,866	0,952	0,984	0,995	0,999	1,000
180	30	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,042	0,248	0,553	0,786	0,911	0,967	0,988	0,996
40	40	0,011	0,459	0,880	0,982	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
80	40	0,000	0,000	0,081	0,516	0,850	0,965	0,993	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
120	40	0,000	0,000	0,000	0,012	0,233	0,631	0,873	0,964	0,991	0,998	1,000	1,000	1,000	1,000	1,000
160	40	0,000	0,000	0,000	0,000	0,002	0,093	0,417	0,737	0,905	0,970	0,991	0,998	0,999	1,000	1,000
200	40	0,000	0,000	0,000	0,000	0,000	0,000	0,035	0,256	0,585	0,819	0,932	0,977	0,993	0,998	0,999
50	50	0,003	0,370	0,848	0,976	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
100	50	0,000	0,000	0,042	0,430	0,811	0,954	0,991	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000
150	50	0,000	0,000	0,000	0,004	0,157	0,556	0,840	0,953	0,988	0,997	0,999	1,000	1,000	1,000	1,000
200	50	0,000	0,000	0,000	0,000	0,000	0,049	0,329	0,677	0,879	0,961	0,989	0,997	0,999	1,000	1,000

Table B.1: Distribution of the multinomial maximum

B.2. DISTRIBUTION OF THE RANGE

The following table provides critical values for the range distribution. For example, in the case with $N = 15$ and $m = 15$

$$P(\max n_i - \min n_i \leq 3 : N = 15, m = 15) = 0.51$$

t		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
N	m																
5	5	0,038	0,710	0,966	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
10	5	0,012	0,321	0,601	0,867	0,971	0,996	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
15	5	0,006	0,181	0,386	0,659	0,854	0,953	0,988	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	5	0,003	0,116	0,265	0,500	0,716	0,868	0,949	0,983	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000
25	5	0,002	0,081	0,193	0,388	0,595	0,770	0,887	0,952	0,982	0,994	0,998	1,000	1,000	1,000	1,000	1,000
30	5	0,001	0,059	0,146	0,308	0,497	0,676	0,815	0,906	0,958	0,983	0,994	0,998	0,999	1,000	1,000	1,000
10	10	0,000	0,396	0,873	0,984	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	10	0,000	0,060	0,237	0,640	0,896	0,977	0,996	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	10	0,000	0,016	0,079	0,313	0,612	0,841	0,949	0,986	0,997	0,999	1,000	1,000	1,000	1,000	1,000	1,000
40	10	0,000	0,006	0,032	0,158	0,380	0,637	0,831	0,935	0,978	0,994	0,998	1,000	1,000	1,000	1,000	1,000
50	10	0,000	0,002	0,015	0,086	0,236	0,458	0,679	0,840	0,932	0,974	0,991	0,997	0,999	1,000	1,000	1,000
60	10	0,000	0,001	0,008	0,050	0,151	0,325	0,535	0,724	0,858	0,935	0,973	0,990	0,997	0,999	1,000	1,000
15	15	0,010	0,061	0,510	0,875	0,978	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	15	0,000	0,016	0,079	0,313	0,612	0,841	0,949	0,986	0,997	0,999	1,000	1,000	1,000	1,000	1,000	1,000
45	15	0,000	0,001	0,028	0,111	0,299	0,542	0,756	0,892	0,959	0,986	0,996	0,999	1,000	1,000	1,000	1,000
60	15	0,000	0,001	0,008	0,050	0,151	0,325	0,535	0,724	0,858	0,935	0,973	0,990	0,997	0,999	1,000	1,000
75	15	0,000	0,000	0,005	0,023	0,082	0,198	0,365	0,553	0,719	0,843	0,920	0,963	0,984	0,994	0,998	0,999
90	15	0,000	0,000	0,002	0,013	0,047	0,124	0,250	0,412	0,581	0,729	0,841	0,914	0,957	0,980	0,991	0,996
20	20	0,000	0,121	0,706	0,949	0,993	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	20	0,000	0,002	0,029	0,321	0,745	0,934	0,986	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	20	0,000	0,000	0,002	0,065	0,274	0,607	0,850	0,954	0,988	0,997	0,999	1,000	1,000	1,000	1,000	1,000
80	20	0,000	0,000	0,000	0,015	0,089	0,301	0,586	0,812	0,929	0,977	0,993	0,998	1,000	1,000	1,000	1,000
100	20	0,000	0,000	0,000	0,004	0,030	0,138	0,349	0,603	0,802	0,917	0,969	0,990	0,997	0,999	1,000	1,000
25	25	0,000	0,067	0,634	0,931	0,991	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
50	25	0,000	0,000	0,010	0,227	0,680	0,912	0,981	0,996	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
75	25	0,000	0,000	0,000	0,030	0,178	0,505	0,799	0,936	0,983	0,996	0,999	1,000	1,000	1,000	1,000	1,000
100	25	0,000	0,000	0,000	0,005	0,041	0,201	0,483	0,747	0,901	0,967	0,990	0,997	0,999	1,000	1,000	1,000
30	30	0,000	0,037	0,570	0,914	0,988	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	30	0,000	0,000	0,003	0,161	0,620	0,891	0,975	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	40	0,000	0,000	0,201	0,720	0,941	0,990	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
80	40	0,000	0,000	0,001	0,027	0,195	0,587	0,857	0,960	0,990	0,998	1,000	1,000	1,000	1,000	1,000	1,000
50	50	0,000	0,003	0,370	0,848	0,976	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
100	50	0,000	0,000	0,000	0,042	0,430	0,811	0,954	0,991	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Table B.2: Distribution of the multinomial range

B.3. DISTRIBUTION OF THE FIRST TWO HIGHEST ORDER STATISTICS

The following table provides critical values for the distribution of the sum of the maximum and the second maximum. For example, in the case with $N = 10$ and $m = 10$

$$P(n_{<1>} + n_{<2>} \leq 5 : N = 10, m = 10) = 0.811$$

t		5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
N	m																		
10	5	0,166	0,588	0,887	0,984	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
15	5	0,000	0,006	0,088	0,392	0,722	0,913	0,981	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	5	0,000	0,000	0,000	0,003	0,054	0,277	0,580	0,817	0,939	0,984	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000
25	5	0,000	0,000	0,000	0,000	0,000	0,002	0,037	0,205	0,469	0,720	0,881	0,959	0,988	0,997	0,999	1,000	1,000	1,000
30	5	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,027	0,157	0,385	0,631	0,816	0,923	0,972	0,992	0,998	1,000
10	10	0,811	0,967	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	10	0,001	0,131	0,415	0,736	0,913	0,978	0,996	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	10	0,000	0,000	0,000	0,050	0,206	0,494	0,745	0,898	0,966	0,991	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	10	0,000	0,000	0,000	0,000	0,000	0,023	0,109	0,325	0,573	0,779	0,903	0,963	0,988	0,996	0,999	1,000	1,000	1,000
50	10	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,011	0,061	0,216	0,432	0,653	0,816	0,915	0,965	0,987	0,996	0,999
60	10	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,006	0,036	0,148	0,325	0,538	0,721	0,852	0,929	0,969
15	15	0,592	0,884	0,978	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	15	0,000	0,037	0,172	0,501	0,772	0,922	0,978	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
45	15	0,000	0,000	0,000	0,009	0,052	0,251	0,511	0,749	0,893	0,962	0,988	0,997	0,999	1,000	1,000	1,000	1,000	1,000
60	15	0,000	0,000	0,000	0,000	0,000	0,002	0,018	0,127	0,316	0,561	0,760	0,888	0,954	0,983	0,994	0,998	1,000	1,000
75	15	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,007	0,067	0,193	0,405	0,616	0,786	0,894	0,953	0,981	0,993
90	15	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,037	0,119	0,287	0,485	0,674	0,815	0,905
20	20	0,410	0,788	0,948	0,991	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	20	0,000	0,011	0,066	0,339	0,635	0,852	0,952	0,987	0,997	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	20	0,000	0,000	0,000	0,001	0,012	0,131	0,340	0,610	0,808	0,922	0,972	0,991	0,998	0,999	1,000	1,000	1,000	1,000
80	20	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,052	0,169	0,396	0,622	0,801	0,908	0,963	0,986	0,995	0,999	1,000
100	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,022	0,084	0,247	0,451	0,658	0,811	0,907	0,959	0,983
25	25	0,273	0,695	0,912	0,982	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
50	25	0,000	0,003	0,024	0,233	0,515	0,779	0,919	0,976	0,994	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	30	0,176	0,612	0,873	0,971	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	30	0,000	0,001	0,008	0,163	0,412	0,708	0,884	0,963	0,990	0,998	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	40	0,070	0,478	0,790	0,945	0,989	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
80	40	0,000	0,000	0,001	0,082	0,257	0,579	0,808	0,932	0,979	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
50	50	0,026	0,378	0,708	0,915	0,981	0,997	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
100	50	0,000	0,000	0,000	0,042	0,156	0,471	0,733	0,897	0,967	0,991	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Table B.3: Distribution of the $n_{<1>} + n_{<2>}$

B.4. DISTRIBUTION OF THE FIRST THREE HIGHEST ORDER STATISTICS

The following table provides critical values for the distribution of the sum of the first three highest order statistics. For example, in the case with $N = 15$ and $m = 15$

$$P(n_{<1>} + n_{<2>} + n_{<3>} \leq 8 : N = 15, m = 15) = 0.869$$

t		8	9	10	11	12	13	14	15	16	18	20	22	24	26	28	30	32	34
N	m																		
10	10	0,987	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	10	0,020	0,174	0,468	0,747	0,913	0,978	0,996	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	10	0,000	0,000	0,000	0,005	0,064	0,230	0,472	0,708	0,869	0,986	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	10	0,000	0,000	0,000	0,000	0,000	0,000	0,002	0,028	0,120	0,509	0,856	0,977	0,998	1,000	1,000	1,000	1,000	1,000
50	10	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,014	0,181	0,559	0,858	0,972	0,996	1,000	1,000	1,000
60	10	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,040	0,254	0,610	0,868	0,970	0,995	0,999
15	15	0,869	0,972	0,996	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	15	0,000	0,038	0,174	0,405	0,669	0,856	0,950	0,986	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
45	15	0,000	0,000	0,000	0,000	0,009	0,052	0,160	0,362	0,592	0,898	0,986	0,999	1,000	1,000	1,000	1,000	1,000	1,000
60	15	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,018	0,186	0,571	0,866	0,974	0,997	1,000	1,000	1,000	1,000
75	15	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,001	0,028	0,226	0,578	0,850	0,964	0,994	0,999	1,000
90	15	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,053	0,269	0,598	0,846	0,958	0,991
20	20	0,714	0,908	0,979	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	20	0,000	0,011	0,066	0,197	0,448	0,697	0,864	0,950	0,985	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	20	0,000	0,000	0,000	0,000	0,001	0,012	0,047	0,172	0,373	0,770	0,953	0,994	1,000	1,000	1,000	1,000	1,000	1,000
80	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,003	0,066	0,351	0,718	0,923	0,986	0,998	1,000	1,000	1,000
100	20	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,004	0,089	0,358	0,695	0,901	0,977	0,996	0,999
25	25	0,558	0,823	0,949	0,989	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
50	25	0,000	0,003	0,024	0,087	0,290	0,550	0,763	0,898	0,963	0,997	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
30	30	0,416	0,728	0,909	0,977	0,995	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
60	30	0,000	0,001	0,008	0,036	0,189	0,429	0,662	0,837	0,934	0,993	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
40	40	0,209	0,553	0,816	0,941	0,985	0,997	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
80	40	0,000	0,000	0,001	0,005	0,086	0,260	0,479	0,705	0,862	0,981	0,998	1,000	1,000	1,000	1,000	1,000	1,000	1,000
50	50	0,096	0,417	0,723	0,896	0,970	0,993	0,999	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
100	50	0,000	0,000	0,000	0,001	0,042	0,156	0,332	0,577	0,780	0,962	0,996	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Table B.4: Distribution of the $n_{<1>} + n_{<2>} + n_{<3>}$

REFERENCES

[1] William G. Cochran. Some methods for strengthening the common x2 tests. *Biometrics*, 10(4):417–451, 1954.

[2] Giovanni Compiani. Some considerations on the goodness-of-fit problem. Bocconi University, 2012.

[3] Charles J. Corrado. The exact distribution of the maximum, minimum and the range of Multinomial/Dirichlet and Multivariate Hypergeometric frequencies. *Statistics and Computing*, 21(3):349–359, 2011.

[4] Anirban Dasgupta. Exact tail probabilities and percentiles of the multinomial maximum.

[5] V. Kolchin, B. Sevast'yanov, and V. Chistykov. *Random allocations*. V.H Wnston & Sons, 1978.

[6] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. 2006.

[7] D. H. Young N. L. Johnson. Some applications of two approximations to the multinomial distribution. *Biometrika*, 47(3/4):463–469, 1960.

[8] Timothy R. C. Read Noel Cressie. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464, 1984.

[9] Michael A. Rappeport. Algorithms and computational procedures for the application of order statistics to queing problems. unpublished dissertation, 1968.

[10] Timothy R.C. Read and Noel A.C. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data*, volume 1. Springer-Verlag, 1988.

[11] D.H. Young. Two alternatives to the standard x2 -test of the hypothesis of equal cell frequencies. *Biometrika*, (49):107–116, 1962.