A Recursive Formulation for the Rank-Sum Statistic used to detect Genomic Copy Number Variation.

Nam Tran Hoang Department of Mathematics and Computer Science
Ohio Wesleyan University, Delaware, Ohio 43015
nntranho@owu.edu

February 14, 2016

Abstract

Copy number variation (CNV) results from duplications and deletions of genomic DNA. Since CNVs were found to correlate with a number of genetic diseases, detecting and characterizing CNV is a major goal of genetic research. Recently, a rank-based method has been developed to analyze raw CNV.

This method involves a rank comparison of a sample DNA across multiple DNA sections, against multiple controls. The overall CNV of the sample is then determined by a statistical comparison of the sample's Rank-Sum against the discrete null distribution. As such, the accuracy of this method depends, to a large degree, on an accurate representation of the null distribution. So far, the exact null distribution has only been approximated using the continuous Irwin-Hall distribution.

This study includes the rigorous proof of several recursive formulations for the weights of the random Rank-Sum statistic. Unexpectedly, these recursive formulae give the generalized form of the binomial coefficients. The descriptive statistics of the exact null distribution are also derived.

The approximated Irwin-Hall distribution is compared to the exact null distribution, from which it is shown to underestimate the standard deviation and overestimate the kurtosis. Using data simulations, the approximated Irwin-Hall distribution also increases the likelihood of type I error (false positive) and gives an overstatement of the test power. Hence, the use of these recursive formulae improves the ability of this rank-based method to detect CNV.

1 Introduction

Genetic information is encoded within the DNA molecule. DNA, though replication of itself, transcription to RNA, and the following translation to protein, propagates genetic materials to biological expressions. This concept embodies the central dogma of modern biology, and in reverse, it traces an aberrant condition back to cellular/molecular defects. With the aid of recent sequencing techniques, copy number variation has been identified as a major component of genomic diversity [4].

Copy number variation (CNV) refers to variations in the number of repeats of a particular section/region of DNA. Although the repetition can arise in many forms (adjacent or non-adjacent segments, preserved or inverted sequences, and other patterns), so far only variable number tandem repeats (VNTR), in which the repetitive segments of preserved sequences are adjacent to one another, are associated with copy number variation. This simplification, in turn, attributes the origin of copy number variation to simple deletions or duplications in the genome: deletions cause low copy number variation, and similarly duplications give rise to high copy number variation. Variable number tandem repeats are further classified into two subgroups: minisatellite DNAs (length of 10-60 base pairs) and microsatellite DNAs (length of 2-5 base pairs). These satellite DNAs can be coding or non-coding/regulatory, linking their copy number variation back to variation in gene/protein expression in either case.

Variable number tandem repeats and copy number variations exist in both prokaryotes and eukaryotes. In humans, CNV has been found in all populations, and since the completion of The Human Genome Project, hundreds of new repetitive regions have been identified [4]. Most CNVs are inherited and stable. Nevertheless, post-fertilization CNVs have been observed both among homologous twins [8] and between cells from the same individual [7]. Evidence supporting the correlation between physical traits and CNVs in humans is ample: the long form dopamine receptor DRD4-7R and attention deficit/hyperactivity disorder [3]; the polymorphism of amylase and starch digestion [2] [5]; and the low copy number of FCGR3B (gene of CD16 cell surface immunoglobulin receptor) and susceptibility to inflammatory autoimmune disorders [6] [10].

CNVs can be detected through a number of techniques: fluorescent in situ hybridization, DNA microarrays [9], next-generation sequencing, etc. Since CNVs correlate to a number of genetic conditions, many statistical tests have been devised to assess the overall CNV of an individual [12]. Recently, LaFramboise et al. (2009) mentioned a particular rank-based method that is both simple and flexible in concept, and relevant to clinical testing [11]. In detail, this method ranks the raw CNV scores of an individual against the similar scores for N other controls, respective to k DNA sections/regions considered. The sum of all the ranked copy number variations of the individual is called the Rank-Sum statistic. The scheme of the procedure is illustrated in Figure 1. Previously, the null distribution for the Rank-Sum (i.e. assuming all variations are purely random) has only been approximated using a modified formulation of the Irwin-Hall distribution [11].

In Section 2, several recursive formulations describing the random Rank-Sum statistic are derived. Through exploring the properties of these formulations, a similarity with binomial coefficients is observed, which eventually leads to an unexpected, fascinating conclusion in Section 3: the Rank-Sum statistic is itself a generalized form of the binomial coefficients.

In Section 5, the probability mass function is obtained, and the exact null distribution is constructed. The cumulative distribution function and other descriptive statistics (mean μ, standard deviation σ, skewness γ, kurtosis κ) are all derived.

Then, data simulations of high Rank-Sum scenarios are performed for both the exact null distribution and the approximated Irwin-Hall distribution. From the simulations, conclusions about the test's type I error α (false positive), type II error β (false negative), and the test's power are deduced. Comparisons are made between the two distributions in Section 6 and Section 7.

Raw CNV		Sample	N controls				
			1	2	3	4	5
k sections	1	5.1	8.8	2.7	4.1	9.0	6.2
	2	0.7	9.3	5.2	6.1	3.7	0.5
	3	7.1	7.9	3.9	3.1	1.3	9.7
	4	0.3	2.4	5.5	6.9	3.2	4.2

Ranked CNV		Sample	N controls				
			1	2	3	4	5
k sections	1	2	4	0	1	5	3
	2	1	5	3	4	2	0
	3	3	4	2	1	0	5
	4	0	1	4	5	2	3
Rank Sum		6					

Figure 1: Example of the rank-base method, with $k = 4$, $N = 5$

Future works are explored in Section 8, which discuss the prospect of applying the recursive formula in analyzing the term x^R of the polinomial $(1 + x + x^2 + \cdots + x^N)^k$ in Number Theory, and in determining the number of ways of obtaining a system of k particles across N quantized energy levels, having a total energy of R in Statistical Mechanics.

2 Constructing the Exact Null Distribution

The process of deriving the Rank-Sum, as described in Section 1, is translated into the following mathematical question: given k integers $x_1, x_2, \ldots, x_k$ such that $0 \le x_i \le N$, what is the probability of obtaining R by summing all x_i? In this set up, the k integers x_i corresponds to the ranked copy number variations of k DNA section, and the restriction on each integer to be between 0 and N incorporates the ranking across N controls of each DNA section. This question is formalized in Definition 2.1 below.

Definition 2.1. The number of distinct solutions to the system

$$x_1 + x_2 + \cdots + x_k = R \qquad (0 \le x_i \le N)$$

is denoted by:

$$\binom{k}{R}_N$$

The number of distinct solutions is written as $\binom{k}{R}_N$ because of its similarity to the binomial coefficients $\binom{k}{R}$. This similarity will be investigated in detail in Section 3. Following the definition, a simple result for $\binom{1}{R}_N$ is derived in Lemma 2.2.

Lemma 2.2.

$$\binom{1}{R}_N = \begin{cases} 1 & \text{if } 0 \le R \le N \\ 0 & \text{if } R < 0 \text{ or } R > N \end{cases}$$

Proof. $\binom{1}{R}_N$ denotes the number of ways of obtaining R from 1 integer: $x_1 = R$, $0 \le x_i \le N$.

If $0 \le R \le N$, there is one way of obtaining R, by setting $x_1 = R$; hence $\binom{1}{R}_N = 1$.

If $R < 0$ or $R > N$, no possible value of x_1 can satisfy $x_1 = R$; hence $\binom{1}{R}_N = 0$. □

A few other simple identities are given:

$$\binom{k}{0}_N = \binom{k}{Nk}_N = 1 \tag{2.1}$$

$$\binom{k}{1}_N = \binom{k}{Nk-1}_N = k \tag{2.2}$$

$$\binom{k}{R}_N = 0 \quad \text{if } R < 0 \text{ or } R > Nk \tag{2.3}$$

The proofs for these identities are straightforwards, similar to that of Lemma 2.2. The proof for identity (2.3) is included in the Appendix, Remark 10.1.

The number of distinct solutions $\binom{k}{R}_N$ is described through the recursive formula of Proposition 2.3 below. This formula is the central proposition of the research.

Proposition 2.3. A recursive formula and initial conditions for $\binom{k}{R}_N$ are given by

$$\binom{k}{R}_N = \sum_{j=0}^{N} \binom{k-1}{R-j}_N$$

$$\binom{0}{R}_N = \begin{cases} 1 & \text{if } R = 0 \\ 0 & \text{if } R \neq 0 \end{cases}$$

Proof. By definition, $\binom{k}{R}_N$ denotes the number of ways of obtaining R by summing k integers $x_1, x_2, \ldots, x_k$, such that $0 \le x_i \le N$. If the value of x_k is specified to be j where $0 \le j \le N$, then the remaining $k-1$ integers will add up to $R-j$. There are $\binom{k-1}{R-j}_N$ ways of doing this. Since j can take any value between 0 and N, the number of ways of summing the original k integers to R is then $\sum_{j=0}^{N} \binom{k-1}{R-j}_N$.

The proposed values of $\binom{0}{R}_N$ are consistent with Lemma 2.2: If $0 \le R \le N$, with j ranging from 1 to N, there exists exactly one value of j such that $R - j = 0$, namely $j = R$. If $R < 0$ or $R > N$, with j ranging from 1 to N, then for all values of j, $R - j \neq 0$. Hence,

$$\binom{1}{R}_N = \sum_{j=0}^{N} \binom{0}{R-j}_N = \begin{cases} 1 & \text{if } 0 \le R \le N \\ 0 & \text{if } R < 0 \text{ or } R > N \end{cases}$$ □

The identity for $\binom{1}{R}_N$ from Lemma 2.2 can also be considered as a set of initial conditions. In the biological problem described, $\binom{1}{R}_N$ has concrete physical interpretation, as shown in the respective proof; whereas $\binom{0}{R}_N$, chosen to be the primary set of initial conditions for the mathematical model, is of little physical meaning. Another simple, yet significant identity for $\binom{k}{R}_N$ is described in Proposition 2.4 below.

Proposition 2.4.

$$\sum_{R=-\infty}^{\infty} \binom{k}{R}_N = \sum_{R=0}^{Nk} \binom{k}{R}_N = (N+1)^k$$

Proof. From Identity 2.3, $\binom{k}{R}_N = 0$ if $R < 0$ or $R > Nk$. Hence,

$$\sum_{R=-\infty}^{\infty} \binom{k}{R}_N = \sum_{R=0}^{Nk} \binom{k}{R}_N$$

Clearly, $\sum_{R=-\infty}^{\infty} \binom{k}{R}_N$ denotes all possible ways of summing k integers $x_1, x_2, \ldots, x_k$, such that $0 \le x_i \le N$. This number can also be derived by multiplying the number of choices for x_1 and the number of choices for x_2 and the number of choices for x_3, and so on, which is $(N+1)(N+1)\cdots = (N+1)^k$. □

3 Analogy to Binomial Coefficients

As introduced in Definition 2.1, the notation for the number of solutions is chosen to be $\binom{k}{R}_N$ because of its analogy to binomial coefficients $\binom{k}{R}$. This analogy is described in Lemma 3.1 and Proposition 3.2 below:

Lemma 3.1.

$$\binom{k}{R}_1 = \binom{k}{R}$$

Proof. From Proposition 2.3, the recursive relation and initial conditions for $\binom{k}{R}_1$ is:

$$\binom{k}{R}_1 = \sum_{j=0}^{1} \binom{k-1}{R-j}_1 = \binom{k-1}{R}_1 + \binom{k-1}{R-1}_1$$

$$\binom{0}{R}_1 = \begin{cases} 1 & \text{if } R = 0 \\ 0 & \text{if } R \neq 0 \end{cases}$$

These are identical to the recursive relation and initial conditions for the binomial coefficients $\binom{k}{R}$. Hence, the two expressions $\binom{k}{R}_1$ and $\binom{k}{R}$ are identical. □

Proposition 3.2.

$$\binom{k}{R}_N = \binom{k}{Nk-R}_N$$

Proof. Following Definition 2.1, $\binom{k}{R}_N$ is the number of distinct solutions $\{x_1, x_2, \ldots, x_k\}$, with

$$x_1 + x_2 + \cdots + x_k = R, \quad 0 \le x_i \le N$$

This equation is equivalent to

$$(N - x_1) + (N - x_2) + \cdots + (N - x_k) = Nk - R \qquad (*)$$

Equation (*) is then equivalent to

$$y_1 + y_2 + \cdots + y_k = Nk - R, \quad 0 \le y_i \le N \tag{**}$$

Where y_i in equation (**) represents $N - x_i$ in equation (*).

The mapping $\{x_1, x_2, \ldots, x_k\} \mapsto \{y_1, y_2, \ldots, y_k\}$ is a one-to-one correspondence between solutions of Equation (*) and solutions of Equation (**). Hence,

$$\binom{k}{R}_N = \binom{k}{Nk - R}_N \qquad \square$$

Proposition 3.2 reflects the symmetry of $\binom{k}{R}_N$, as it has been alluded to in the restricted cases of identity (2.1) and identity (2.2). Taking $N = 1$ in Proposition 3.2 in particular, $\binom{k}{R}_1 = \binom{k}{k-R}_1$, resembling the familiar identity for binomial coefficients: $\binom{k}{R} = \binom{k}{k-R}$. Overall, $\binom{k}{R}_N$ can be considered as a generalization of the binomial coefficients $\binom{k}{R}$. The similarity between $\binom{k}{R}_N$ and $\binom{k}{R}$ is summarized in Table 1.

	$\binom{k}{R}_N$	$\binom{k}{R}$
Recursive formula	$\binom{k}{R}_N = \sum_{j=0}^{N} \binom{k-1}{R-j}_N$	$\binom{k}{R} = \binom{k-1}{R} + \binom{k-1}{R-1}$
Total sum	$\sum_{R=0}^{Nk} \binom{k}{R}_N = (N+1)^k$	$\sum_{R=0}^{k} \binom{k}{R} = 2^k$
Symmetry	$\binom{k}{R}_N = \binom{k}{Nk-R}_N$	$\binom{k}{R} = \binom{k}{k-R}$

Table 1: Comparison between $\binom{k}{R}_N$ and $\binom{k}{R}$

4 Other Formulae for $\binom{k}{R}_N$

Apart from the concise recursive formula derived in Proposition 2.3, two other formulae for $\binom{k}{R}_N$ are given in this section.

Proposition 4.1.

$$\binom{k}{R}_N = \sum_{j=0}^{k} (-1)^j \binom{k}{j} \binom{R + k - 1 - j(N+1)}{k-1}$$

Proof. Consider k integers x_1, x_2, …, x_k, with no restriction. The number of ways of obtaining a set of these k integers, such that their sum is R, is $\binom{R+k-1}{k-1}$, following the familiar "stars and bars" counting technique.

Again, consider k integers $x_1, x_2, \ldots, x_k$, summing to R, with restriction on one particular x_i: $x_i > N \Leftrightarrow x_i \geq N+1$. The restriction can be understood as localizing $(N+1)$ stars (among $R+k-1$ stars) between the two bars that determine x_i, ensuring that x_a is always larger than N. The number of choices for x_i (among k integers) is $\binom{k}{1}$. Hence, the number of ways of obtaining a solution under this restricted condition is $\binom{k}{1}\binom{R+k-1-(N+1)}{k-1}$.

Similarly, the number of ways of obtaining a solution, with the restriction $x_i > N$ on j terms, is $\binom{k}{j}\binom{R+k-1-j(N+1)}{k-1}$.

Applying the Inclusion-Exclusion Principle, the number of ways of obtaining a set of k integers summing to R, while each integer is restricted to be between 0 and N inclusively, is equal to:

- the number of ways of obtaining these k integers with no restriction,
- minus all the cases where one term is restricted to $x_i > N$,
- plus all the cases where two terms are restricted to $x_i > N$, which were doubly-excluded in the previous step,
- minus all the cases where three terms are restricted to $x_i > N$, which have not been excluded as a result of the two previous steps (triply-excluded in the first step and triply-included in the second step),
- and so on.

Hence,

$$\begin{aligned}\binom{k}{R}_N &= \binom{R+k-1}{k-1} - \binom{k}{1}\binom{R+k-1-(N+1)}{k-1} + \binom{k}{2}\binom{R+k-1-2(N+1)}{k-1} - \cdots \\ &= \sum_{j=0}^{k}(-1)^j\binom{k}{j}\binom{R+k-1-j(N+1)}{k-1}\end{aligned}$$

□

Compared to the existing recursive formula of Proposition 2.3, the formula given in Proposition 4.1 is less easily computable: it involves products of two binomial coefficients, which could be computationally costly when k, N, and R becomes sufficiently large.

Another formula for $\binom{k}{R}_N$ is given in Corollary 4.2. This formula was initially derived from Proposition 4.1 as an attempt at simplification. Nevertheless, by using Proposition 2.3, a concise, straightforwards proof is given.

Corollary 4.2.

$$\binom{k}{R}_N = \binom{k-1}{R}_N + \binom{k}{R-1}_N - \binom{k-1}{R-N-1}_N$$

Proof. The proof for Corollary 4.2 is given in the Appendix, Remark 10.2. □

5 Characterizing the Exact Null Distribution

The recursive formula of $\binom{k}{R}_N$ from Proposition 2.3 is used to characterize the null distribution. All common descriptive statistics of the distribution are determined in closed-form.

Definition 5.1. The probability of obtaining a random Rank-Sum $R = x_1 + x_2 + \cdots + x_k$ is conventionally denoted as $P_N(k, R)$. In other words, this is the probability mass function of the null distribution.

Proposition 5.2. The recursive formula and initial conditions for $P_N(k, R)$ is

$$P_N(k, R) = \frac{1}{N+1} \sum_{j=0}^{N} P_N(k-1, R-j) \tag{5.1}$$

$$P_N(0, R) = \begin{cases} 1 & \text{if } R = 0 \\ 0 & \text{if } R \neq 0 \end{cases}$$

Proof. From Lemma 2.4,

$$\sum_{R=-\infty}^{\infty} \binom{k}{R}_N = \sum_{R=0}^{Nk} \binom{k}{R}_N = (N+1)^k$$

Hence,

$$P_N(k, R) = \frac{1}{(N+1)^k} \binom{k}{R}_N$$

The initial conditions immediately follow as

$$P_N(0, R) = \frac{1}{(N+1)^0} \binom{0}{R}_N = \begin{cases} 1 & \text{if } R = 0 \\ 0 & \text{if } R \neq 0 \end{cases}$$

For the recursive relation, Proposition 2.3 is applied,

$$\binom{k}{R}_N = \sum_{j=0}^{N} \binom{k-1}{R-j}_N$$

$$\begin{aligned} P_N(k, R) &= \frac{1}{(N+1)^k} \binom{k}{R}_N \\ &= \frac{1}{(N+1)^k} \sum_{j=0}^{N} \binom{k-1}{R-j}_N \\ &= \frac{1}{(N+1)^k} (N+1)^{k-1} \sum_{j=0}^{N} \frac{1}{(N+1)^{k-1}} \binom{k-1}{R-j}_N \\ &= \frac{1}{N+1} \sum_{j=0}^{N} P_N(k-1, R-j) \end{aligned}$$

□

Similar to the case of $\binom{k}{R}_N$, a few identities for $P_N(k, R)$ are given

$$P_N(1, R) = \begin{cases} \frac{1}{N+1} & \text{if } 0 \leq R \leq N \\ 0 & \text{if } R < 0 \text{ or } R > N \end{cases} \tag{5.2}$$

$$P_N(k,0) = P_N(k,Nk) = \frac{1}{(N+1)^k} \tag{5.3}$$

$$P_N(k,1) = P_N(k,Nk-1) = \frac{k}{(N+1)^k} \tag{5.4}$$

$$P_N(k,R) = 0 \quad \text{if } R < 0 \text{ or } R > Nk \tag{5.5}$$

$$P_N(k,R) = P_N(k,Nk-R) \tag{5.6}$$

Definition 5.3. The cumulative distribution function is conventionally denoted as $F_N(k,R) \equiv \sum_{x \le R} P_N(k,x)$. For computational purposes, another function that accounts for cumulative probability is specified: $G_N(k,R) \equiv \sum_{x > R} P_N(k,x)$.

Since the study chooses high Rank-Sum R as the scenario for later data simulation, $G_N(k,R)$ is defined for convenience. Notice that by construction, $F_N(k,R) + G_N(k,R) = 1$. The recursive formulae for $F_N(k,R)$ and $G_N(k,R)$, along with their initial conditions, are given in Corollary 5.4 and Corollary 5.5 below.

Corollary 5.4. The recursive formula and initial conditions for $F_N(k,R)$ are

$$F_N(k,R) = \frac{1}{N+1} \sum_{j=0}^{N} F_N(k-1, R-j) \tag{5.7}$$

$$F_N(0,R) = \begin{cases} 0 & \text{if } R < 0 \\ 1 & \text{if } R \ge 0 \end{cases}$$

Proof. Following Definition 5.3,

$$F_N(k,R) = \sum_{x \le R} P_N(k,x) = \sum_{i=0}^{R} P_N(k,i)$$

Applying Proposition 5.2,

$$\begin{aligned} F_N(k,i) &= \sum_{i=0}^{R} P_N(k,i) \\ &= \sum_{i=0}^{R} \frac{1}{N+1} \sum_{j=0}^{N} P_N(k-1, i-j) \\ &= \frac{1}{N+1} \sum_{j=0}^{N} F_N(k-1, R-j) \end{aligned}$$

□

The initial condition is also deduced from Definition 5.3 and Proposition 5.2

$$F_N(0,R) = \sum_{i=0}^{R} P_N(0,i) = \begin{cases} 0 & \text{if } R < 0 \\ 1 & \text{if } R \ge 0 \end{cases}$$

Corollary 5.5. The recursive formula and initial conditions for $G_N(k, R)$ is

$$G_N(k, R) = \frac{1}{N+1} \sum_{j=0}^{N} G_N(k-1, R-j) \tag{5.8}$$

$$G_N(0, R) = \begin{cases} 1 & \text{if } R < 0 \\ 0 & \text{if } R \geq 0 \end{cases}$$

Proof. Similar to the proof for Lemma 5.4 □

Similarly, a few identities for $F_N k, R$ and $G_N(k, R)$ are listed:

$$F_N(k, R) = \begin{cases} 0 & \text{if } R < 0 \\ 1 & \text{if } R \geq Nk \\ \frac{1}{(N+1)^k} & \text{if } R = 0 \\ \frac{(N+1)^k - 1}{(N+1)^k} & \text{if } R = Nk - 1 \end{cases} \qquad G_N(k, R) = \begin{cases} 1 & \text{if } R < 0 \\ 0 & \text{if } R \geq Nk \\ \frac{(N+1)^k - 1}{(N+1)^k} & \text{if } R = 0 \\ \frac{1}{(N+1)^k} & \text{if } R = Nk - 1 \end{cases}$$

Definition 5.6. The mean, standard variation, skewness, and kurtosis of the distribution constructed from the probability function $P_N(k, R)$ are conventionally denoted as $\mu_N(k)$, $\sigma_N(k)$, $\gamma_N(k)$, and $\kappa_N(k)$, respectively.

The general formulae defining $\mu_N(k)$, $\sigma_N(k)$, $\gamma_N(k)$, $\kappa_N(k)$ are listed in the Appendix, Remark 10.3. The closed-form formulae for these descriptors are derived by direct substitutions of the moments (from Remark 10.4, 10.5, 10.6, and 10.7) into these defining formulae.

Lemma 5.7. *The mean $\mu_N(k)$, standard variation $\sigma_N(k)$, skewness $\gamma_N(k)$, and kurtosis $\kappa_N(k)$ of the exact null distribution are:*

$$\mu_N(k) = \frac{Nk}{2}$$

$$\sigma_N(k) = \sqrt{\frac{kN^2}{12} + \frac{kN}{6}}$$

$$\gamma_N(k) = 0$$

$$\kappa_N(k) = 3 - \frac{6}{5k} - \frac{12}{5kN(N+2)}$$

6 Comparison with the Approximate Irwin-Hall Distribution

As mentioned in Section 1, LaFramboise et al. [11] modified the Irwin-Hall distribution to approximate the null distribution of the Rank-Sum statistic. The continuous Irwin-Hall distribution gives the probability density of obtaining a sum $\mathcal{Z}$ from k real continuous random variables z_i, where each variable z_i is between 0 and 1. The scheme, the probability density

function ${}_{IH}P$, and the cumulative distribution function ${}_{IH}F$ for the Irwin-Hall distribution are, respectively [13]:

$$z_1 + z_2 + \cdots + z_k = \mathcal{Z} \mid z_i \in \mathcal{U}(0,1)$$

$${}_{IH}P = \frac{1}{2(k-1)!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(\mathcal{Z}-j)^{k-1}\,\mathrm{Sign}\,(\mathcal{Z}-j)$$

$${}_{IH}F = \frac{1}{2} + \frac{1}{2k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}(\mathcal{Z}-j)^{k}\,\mathrm{Sign}\,(\mathcal{Z}-j)$$

The Irwin-Hall distribution is chosen as an approximation because of the similarity in formulation between Irwin-Hall and the exact null distribution: for Irwin-Hall, the probability of a random sum $\mathcal{Z}$ of k continuous variables z_i is given, while for the exact null distribution, the probability of a random Rank-Sum R of k discrete variables x_i is obtained. Based on this original concept of Framboise et al. (2009) [11]: substitutions are made for $z_i = \frac{x_i}{N}$, i.e. z_i is approximately $\mathcal{U}(0,1)$ when $N >> 0$, which leads to $\mathcal{Z} = \frac{R}{N}$, followed by the re-normalization of the probability density function ${}_{IH}P$ of Irwin-Hall to PIH, these approximations of the null distribution are derived:

$$P_N(k,R) \simeq PIH_N(k,R) = \frac{1}{2N(k-1)!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\left(\frac{R}{N}-j\right)^{k-1}\mathrm{Sign}\left(\frac{R}{N}-j\right) \tag{6.1}$$

$$F_N(k,R) \simeq FIH_N(k,R) = \frac{1}{2} + \frac{1}{2k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\left(\frac{R}{N}-j\right)^{k}\mathrm{Sign}\left(\frac{R}{N}-j\right) \tag{6.2}$$

$$G_N(k,R) \simeq GIH_N(k,R) = \frac{1}{2} - \frac{1}{2k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\left(\frac{R}{N}-j\right)^{k}\mathrm{Sign}\left(\frac{R}{N}-j\right) \tag{6.3}$$

The exact null distribution and the approximate Irwin-Hall distribution are compared in Table 2 and Figure 7 (on page 29). In particular, the following differences between the two distributions are observed:

- $P_N(k,R)$ is more computable using the exact distribution. Since there is a recursive formula for the exact probability mass function (Equation 5.1), $P_N(k,R)$ can be computed easily, even for very large N, k, R. Whereas in the case of the approximate formula (Equation 6.1), binomial coefficients are involved in an alternating series, which increases the cost of the computation.
- The two distributions are of different types: the exact null distribution is discrete, while the approximate Irwin-Hall distribution is continuous. There is a large mismatch between them when N is small. Yet, as N goes to infinity, it is expected that the approximate distribution would approach the exact distribution. This is reflected in the formulae for standard deviation $\sigma_N(k)$ and kurtosis $\kappa_N(k)$:

$$\sigma_N(k) = \sqrt{\frac{kN^2}{12} + \frac{kN}{6}} \simeq \sqrt{\frac{kN^2}{12}} \mid N \to \infty$$

$$\kappa_N(k) = 3 - \frac{6}{5k} - \frac{12}{5kN(N+2)} \simeq 3 - \frac{6}{5k} \mid N \to \infty$$

	Exact null distribution	Approximate Irwin-Hall distribution
Probability mass function	Equation 5.1	Equation 6.1
Cumulative distribution function	Equation 5.7 and 5.8	Equation 6.2 and 6.3
Mean $\mu_N(k)$	$\frac{Nk}{2}$	$\frac{Nk}{2}$
Standard deviation $\sigma_N(k)$	$\sqrt{\frac{kN^2}{12} + \frac{kN}{6}}$	$\sqrt{\frac{kN^2}{12}}$
Skewness $\gamma_N(k)$	0	0
Kurtosis $\kappa_N(k)$	$3 - \frac{6}{5k} - \frac{12}{5kN(N+2)}$	$3 - \frac{6}{5k}$

Table 2: Comparison between the exact and the approximate distribution

- The two distributions give the same value for the mean $\mu_N(k) = \frac{Nk}{2}$, and skewness $\gamma_N(k) = 0$, indicating that both of them are symmetric. The two distributions have their kurtosis $\kappa_N(k)$ to be less than 3, indicating a platykurtic pattern (larger spread than the normal distribution).
- The standard deviation $\sigma_N(k)$ approximated by Irwin-Hall distribution is smaller than the real value. The kurtosis $\kappa_N(k)$ approximated by Irwin-Hall distribution is larger than the real value. Data simulations of Section 7.2 show that when using the approximate distribution, type I error is exaggerated and type II error is underestimated. Although it is often valid in other situations, the argument attributing these differences in type I and type II error to the differences in standard deviation and kurtosis seems to be inconsistent for this particular case.

7 Data Simulation

7.1 Power, Type I and Type II Errors of the Null Distribution

In this section, a statistical test for high CNV is formulated using the null distribution. The null hypothesis is stated as H_0: The Rank-Sum (the sum of all ranked CNVs) of the sample is not larger than the value randomly obtained from the null distribution, at significance level α_0. Accordingly, the alternative hypothesis is stated as H_1: The Rank-Sum of the sample is larger than the value randomly obtained from the null distribution, at significance level α_0. Since high Rank-Sum is investigated, only the right tail of the null distribution H_0 is concerned, and the real distribution of the value H_1 is assumed to shift right from the null distribution H_0, as visualized in Figure 2 on page 24. (All figures are placed at the end of the article.)

Data is simulated for the test: the raw readings of copy number variations (RCNV) are generated for each of the k DNA section, for all the N controls and the sample of interest. All RCNVs of N controls are chosen to follow the standard normal distribution $\mathcal{N}(\mu = 0, \sigma^2 = 1)$. The RCNVs of the sample are chosen to follow a shifted normal distribution $\mathcal{N}(\mu = \delta, \sigma^2 = 1)$; δ is chosen to be larger than 0 to ensure that H_1 is shifted right from H_0 (notice that δ is not the separation between H_0 and H_1).

From the RCNVs, the ranked CNVs and the Rank-Sum of the sample are obtained. The Rank-Sum is tested against the null distribution, at significance level α_0, to determine whether the null hypothesis H_0 should be rejected (positive result for the test, meaning high Rank-Sum confirmed). The procedure is repeated 200 times for each combination of the following parameters:

- number of sections k ranging from 5 to 100, in steps of 5
- number of controls N ranging from 5 to 100, in steps of 5
- significance level α_0 for the test takes values as 0.05, 0.10, and 0.20
- the parameter δ takes values 0.00, 0.25, 0.50, 0.75, and 1.00

  (many graphs below have Pos on one of their axis. Pos denotes the frequency of giving a positive result over 200 trials)

  (this list also serves as the key for abbreviations repeatedly used in graphs)

In general statistics, a widely used value for the significance level α_0 is 0.05. Yet, for medical testings, it is a common practice to set α_0 higher to increase α, decrease β, i.e. trading off a higher type I error (false positive) for a lower type II error (false negative). The rationale behind such a choice is that for medical testing, it is more favorable to give an erroneous positive result, which would be disproved following further testings or observations, rather than letting an abnormal condition go undetected. Because of this reason, three values of significance level α_0 are investigated: at 0.05, at 0.10, and at 0.20. Notice that in this context α_0 denotes the pre-determined significance level for testing, while type I error/α denotes the rate of false positive results obtained after the test.

Four surfaces of the same significance level $\alpha_0 = 0.10$ and different δ parameter are shown on Figure 3a. Three surfaces of the same parameter $\delta = 0.50$ and different significance level

α_0 are shown on Figure 3b. These surfaces are exemplary for the other combinations of parameters (i.e. the other surfaces would look very similar). These surfaces provide the overall view of the test, from which specific cross-sections are obtained and analyzed.

Several cross-sections from Figure 3a are reproduced in Figures 4 and 5, from which the properties of the test can be examined. First, at parameter $\delta = 0.00$ (no separation between the two distributions H_0 and H_1) all positive cases detected fall into type I error. From Figure 4a and Figure 4b, type I error is proportional to the numbers of controls N, and is inversely proportional to the number of sections k. The data suggests that as more sections k are investigated at once, smaller variation in the Rank-Sum is observed, which decreases the rate of false positive results α. On the other hand, as the number of controls N increases, the test takes into account more variations of the Rank-Sum, and hence increases the rate of false positive results. It is interesting that this rate of false positive result does not seem to be capped by the pre-determined significant value α_0 0.10 - yet, it is expected that as the number of repeats goes to infinity, this rate of false positive result would approach α_0.

In Figure 4c and Figure 4d, at parameter $\delta = 0.50$ (there is a separation between the two distributions) all negative cases fall into type II error, and all positive cases make up the power of the test. The power of the test is proportional to both the number of sections k and the number of controls N. The sharp correlation observed between the power and number of sections is because of the data simulation set up: all sections follow a shifted normal distribution $\mathcal{N}(\mu = \delta, \sigma^2 = 1)$. These correlations are also observed at different δ values, as evident when cross-sections of the same δ value are compared across Figures 5a, 5c, 5e, or Figures 5b, 5d, 5f.

Several projections from Figure 3b and similar surfaces are displayed in Figure 6. The projections are taken as looking down the Pos axis, viewing the plane (k, N), visualizing only the data points that satisfy the restrictions for value of Pos: blue dots and green diamonds both denote $\alpha \leq 0.05$ at $\delta = 0.00$; red squares and green diamonds both denote $\beta \leq 0.01$ or power $1 - \beta$ larger than 0.99 at $\delta = 0.50$ (green diamonds denote the overlap of both conditions on α and β). Considering the collection of blue dots and green diamonds across the sub-figures (all points satisfying $\alpha \leq 0.05$), it is reasonable that as the significance level α_0 increases, there are fewer points satisfying the given restriction on α, i.e. α increases with significance level α_0. Similar observations are made for the collection of red squares and green diamonds (all points satisfying $\beta \leq 0.01$, or power $1 - \beta$ larger than 0.99): there are more points satisfying the given restriction on β, i.e. β decreases and power increases with significance level α_0. At a given significance level α_0 and number of DNA sections k, the effective number of controls N that satisfy restrictions on both type I α and type II β errors is deduced by delineating the region of green diamonds with simples lines:

- For $\alpha_0 = 0.05$, the region is delineated with the line ℓ_1 through (60,60) and (95,40) and the line ℓ_2 through (65,85) and (100,100), which gives $\frac{3}{7}k + \frac{400}{7} \geq N \geq \frac{-4}{7}k + \frac{660}{7}$
- For $\alpha_0 = 0.10$, the region is delineated with the line ℓ_3 through (65,40) and (100,25) and the line ℓ_4 through (60,50) and (100,65), which gives $\frac{3}{8}k + \frac{55}{2} \geq N \geq \frac{-3}{7}k + \frac{475}{7}$
- There is not enough data for the case $\alpha_0 = 0.20$

Overall, the data simulation for the test using the exact null distribution shows that type I error α is inversely proportional to the number of sections k and is proportional to the number of controls N. The power $1-\beta$, on the other hand, increases when either the number

of sections k or the number of controls N increases.

7.2 Comparison with the Approximate Irwin-Hall Distribution

As mentioned in Section 6, the Irwin Hall distribution has been used as an approximation for the null distribution. The probability density function and cumulative distribution functions for this particular approximation are denoted as IH: PIH_N, FIH_N, and GIH_N, taken directly from the equations 6.1, 6.2, 6.3:

$$\mathrm{PIH}_N(k,R) = \frac{1}{2N(k-1)!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\left(\frac{R}{N}-j\right)^{k-1}\mathrm{Sign}\left(\frac{R}{N}-j\right) \tag{7.1}$$

$$\mathrm{FIH}_N(k,R) = \frac{1}{2} + \frac{1}{2k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\left(\frac{R}{N}-j\right)^{k}\mathrm{Sign}\left(\frac{R}{N}-j\right) \tag{7.2}$$

$$\mathrm{GIH}_N(k,R) = \frac{1}{2} - \frac{1}{2k!}\sum_{j=0}^{k}(-1)^j\binom{k}{j}\left(\frac{R}{N}-j\right)^{k}\mathrm{Sign}\left(\frac{R}{N}-j\right) \tag{7.3}$$

Figure 7 shows the corresponding functions for both the null distribution and the IH approximation at $k = 6$ and $N = 6$. From Figure 7a, compared to the exact probability mass function P, the approximate probability density function PIH makes an overestimation at the peak, while it underestimates the portion near the tails. This misrepresentation carries through to the pairs of cumulative distributing functions F and FIH, G and GIH in Figures 7b and 7c. By comparing the region near the right tail of GIH and G, the probability of obtaining a high Rank-Sum is underestimated by GIH.

The difference in using the IH approximation for the test, instead of the exact null distribution is investigated. First, the corresponding critical value R_{crit} is computed at a pre-determined significance level α_0 for the IH approximation. Using this critical value R_{crit}, a new significance level α_0' is obtained for the null distribution. The difference in the rate of giving positive results between using the IH approximation with significance level α_0, and in using the null distribution with significance level α_0', upon testing the simulated Rank-Sum, is visualized in the following graphs. Notice that the difference is denoted as PosIH - Pos. (The simulated Rank-Sum following the same simulation process described in Section 7.1)

Four surfaces of the same significance level α_0 0.10 and different δ parameter are shown on Figure 8a. Three surfaces of the same parameter $\delta = 0.50$ and different significance level α_0 are shown on Figure 8b. These surfaces are exemplary for the other combinations of parameters (i.e. the other surfaces would look very similar). They provide the overall view of the test, from which specific cross-sections could be obtained and analyzed.

Several cross-sections from Figure 8a are reproduced in Figures 9 and 10, from which the difference PosIH - Pos can be examined. From Figure 9a and Figure 9b, the IH approximation over-estimates type I error, and this overestimation is proportional to the number of controls N when k is sufficiently large. Interestingly, there is a sharp jump in the type I error's difference at around $k = 25$ for this particular case of $\alpha_0 = 0.10$. Then, as k increases, the difference in type I error decreases.

At parameter $\delta = 0.50$, Figure 9c and Figure 9d indicate that the power of the test is also significantly overestimated when k passes a critical value - for $\alpha_0 = 0.10$, that value is $k = 25$. The above correlations are also observed at different δ values, as evident when cross-sections of the same δ value are compared across Figures 10a, 10c, 10e, or Figures 10b, 10d, 10f. The sharp jump of the difference at $k = 25$ is consistently observed.

Cross-sections of Figure 8b are reproduced in Figure 11. The cross-sections re-confirm the existence of a critical value of k, at above which the power of the test is greatly overestimated. This critical value of k is dependent on the pre-determined significance level α_0: at $\alpha_0 = 0.05$, it is about 20; at $\alpha_0 = 0.10$, it is 25; and at $\alpha_0 = 0.20$, it is 50. Again, the sharp jump of the difference in power of the test at these critical values of k is consistently observed; however, the reason for this sudden change is not yet understood.

Overall, the comparison between testing using the exact null distribution and testing using the approximate Irwin-Hall distribution on simulated data shows that when using the approximate Irwin-Hall distribution, type I error α is exaggerated and type II error β is underestimated.

8 Future Works

The number of ways/the statistical weight $\binom{k}{R}_N$ of obtaining a random Rank-Sum R also describes the coefficient of the term x^R when the exponent of the all one polynomial [16] $(1 + x + x^2 + \cdots + x^N)^k$ is fully expanded. For example,

$$(1 + x + x^2 + x^3)^3 = x^9 + 3x^8 + 6x^7 + 10x^6 + 12x^5 + 12x^4 + 10x^3 + 6x^2 + 3x + 1$$

The corresponding values of $\binom{3}{R}_3$ when R running from 0 to 9 are:

$$\{1, 3, 6, 10, 12, 12, 10, 6, 3, 1\}$$

The proof for the general correspondence is given in Proposition 8.1 below.

Proposition 8.1. Upon expanding the polynomial $(1 + x + x^2 + \cdots + x^N)^k$, the coefficient of the term x^R is $\binom{k}{R}_N$.

Proof. The exponential expression $(1 + x + x^2 + \cdots + x^N)^k$ can be rewritten as a product $\mathcal{P}$ of k multiplicative terms: $\mathcal{P} \equiv (1 + x + x^2 + \cdots + x^N) \cdot (1 + x + x^2 + \cdots + x^N) \cdot \cdots$. Upon the initial expansion of $\mathcal{P}$ to a $(N + 1)^k$ terms summation $\mathcal{S} \equiv (x^\alpha + x^\beta + x^\gamma + \cdots)$, the exponent ι of a term x^ι in $\mathcal{S}$ is the sum of the exponents a of the inner term x^a from the first multiplicative term of $\mathcal{P}$, and the exponents b of the inner term x^b from the second multiplicative term of $\mathcal{P}$, and the exponents c of the inner term x^c from the third multiplicative term of $\mathcal{P}$, and so on. Notice that there are k terms in the sum of the exponents $a + b + c \cdots = \iota$, and each exponents a, b, c, ... is an integer between 0 and N.

Upon complete expansion, the $(N + 1)^k$ terms summation $\mathcal{S}$ is simplified to the Nk term summation $\mathcal{T} \equiv C_0 x^0 + C_1 x^1 + \cdots + C_{Nk} x^{Nk}$. The coefficient of x^R describes the number of ways of obtaining x^R, i.e., the number of ways of obtaining the exponent R by summing k other exponent a, b, c, ..., when each of these exponents is an integer between 0 and N. This construction is identical to the problem described in Definition 2.1. Therefore, $\binom{k}{R}_N$

describes the number of ways of obtaining R, thus it describes the coefficient of x^R in the expansion of polynomial $(1 + x + x^2 + \cdots + x^N)^k$. □

It has been observed that when an inner term x^a in the polynomial $(1 + x + x^2 + \cdots + x^N)^k$ is modified by a multiplicative constant C, the coefficients upon expansion are different, but the overall shape of the distribution obtained from these coefficient remains distinctive. For example, modifying $(1 + x + x^2 + x^3)^3$ to $(1 + x + x^3)^3$ or $(1 + x + 2x^2 + x^3)^3$ changes the coefficients, but it preserves the general shape of the respective distributions. Hence, it is hypothesized that there is a further generalization of $\binom{k}{R}_N$ that corresponds to a modified polynomial.

Moreover, it is very interesting that the Boltzmann distribution from Statistical Mechanics is related to $\binom{k}{R}_N$. Indeed, consider the initial set up from which the Boltzmann distribution arose [17]: given a system of k particles, each particle is allowed to have N levels of quantized energy (consecutively from 0 to N). A configuration $\mathcal{C}_R$ is defined as a general arrangement (permutation allowed) of these particles into energy levels to achieve a particular total energy R. A microstate $\mathcal{M}_R$ is defined as a specific arrangement (permutation not allowed) of these particles into energy levels to achieve a total energy R. Notice that one configuration can give rise to many different microstates. The number of ways W_R of obtaining a configuration $\mathcal{C}_R$, or the number of associated microstates $\mathcal{M}_R$ of a configuration $\mathcal{C}_R$, which is also known as the statistical weight of the configuration $\mathcal{C}_R$, is given by:

$$W_R = \frac{k!}{a_0!a_1!a_2!\ldots a_N!} = \frac{k!}{\prod_{i=1}^{N} a_i!} \tag{8.1}$$

Where a_i denotes the number of particles in energy level i^{th}.

The Boltzmann distribution is derived from equation (8.1). It describes the number a_i of k particles across N energy levels at a given total energy R such that the statistical weight W_R reaches maximum.

The number of ways to achieve a system of total energy R is defined here as Ω_R. Since many different configurations can give the same total energy, Ω_R is the sum of the weight W_R of all these configurations $\mathcal{C}_R$.

$$\Omega_R = \sum W_R = \sum_{\sum_{i=0}^{N} i a_i = R} \frac{k!}{\prod_{i=1}^{N} a_i!} \tag{8.2}$$

It is observed that

$$\binom{k}{R}_N = \Omega_R \tag{8.3}$$

This unexpected relationship can be proven by linking Ω_R back to the problem from Definition 2.1: a system consisting of k particles of quantized energy between 0 and N, having a total energy of R can be obtained from $\binom{k}{R}_N$ distinct microstates. Equation (8.3) shows that $\binom{k}{R}_N$ and the null distribution may find some applications in Statistical Mechanics, especially where the statistical weight W_R of the configuration or Ω_R of the system have only been approximated.

9 Acknowledgements

Dr. Jackson is very much appreciated for his mentoring. From Dr. Jackson's original idea this research materialized, and also thanks to his dedicated instructions this research progresses. He is as well acknowledged for his tremendous help with the editing, formatting, and revision of the article.

Dr. Linder is appreciated for his valuable advice with the statistical arguments of Section 6 and the data analysis of Section 7.

Dr. Markwardt is thanked for his valuable advice and feedback with Section 1.

Collectively, the author thanks the Department of Mathematics and Computer Science at Ohio Wesleyan University for the instruments and resources that the department has generously provided.

10 Appendix

Remark 10.1.

$$\binom{k}{R}_N = 0 \quad \text{if } R < 0 \text{ or } R > Nk$$

Proof. For k integers x_i, with $0 \le x_i \le N$,

the maximum result would occur when each $x_i = N$, giving a sum of Nk;

the minimum result would occur when each $x_i = 0$, giving a sum of 0. □

Remark 10.2.

$$\binom{k}{R}_N = \binom{k-1}{R}_N + \binom{k}{R-1}_N - \binom{k-1}{R-N-1}_N$$

Proof. From Proposition 2.3,

$$\begin{aligned}\binom{k}{R}_N &= \sum_{j=0}^{N}\binom{k-1}{R-j}_N = \binom{k-1}{R}_N + \binom{k-1}{R-1}_N + \binom{k-1}{R-2}_N + \dots \\ &= \binom{k-1}{R}_N + \sum_{j=0}^{N}\binom{k-1}{R-1-j}_N - \binom{k-1}{R-N-1}_N \\ &= \binom{k-1}{R}_N + \binom{k}{R-1}_N - \binom{k-1}{R-N-1}_N\end{aligned}$$

□

Remark 10.3. The general defining formulae of the moments are listed in Table 3 [15]:

	Discrete distribution	Continuous distribution
First moment ${}_1\mu$	$\sum_{R=-\infty}^{\infty} RP(R)$	$\int_{R=-\infty}^{\infty} RP(R)$
Second moment ${}_2\mu$	$\sum_{R=-\infty}^{\infty} R^2P(R)$	$\int_{R=-\infty}^{\infty} R^2P(R)$
Third moment ${}_3\mu$	$\sum_{R=-\infty}^{\infty} R^3P(R)$	$\int_{R=-\infty}^{\infty} R^3P(R)$
Fourth moment ${}_4\mu$	$\sum_{R=-\infty}^{\infty} R^4P(R)$	$\int_{R=-\infty}^{\infty} R^4P(R)$

Table 3: General defining formulae of raw moments

The general defining formulae for the standard statistical descriptors are [14]:

$$\mu_N(k) = {}_1\mu_N(k)$$

$$\sigma_N(k) = \sqrt{{}_2\mu_N(k) - {}_1\mu_N^2(k)}$$

$$\gamma_N(k) = \frac{{}_3\mu_N(k) - 3\,{}_1\mu_N(k)\,{}_2\mu_N(k) + 2\,{}_1\mu_N^3(k)}{\sigma_N^3(k)}$$

$$\kappa_N(k) = \frac{{}_4\mu_N(k) - 4\,{}_3\mu_N(k)\,{}_1\mu_N(k) + 6\,{}_2\mu_N(k)\,{}_1\mu_N^2(k) - 3\,{}_1\mu_N^4(k)}{\sigma_N^4(k)}$$

In remarks 10.4, 10.5, 10.6, 10.7, the raw moments for the null distribution are computed.

Remark 10.4. The first moment for the null distribution is

$${}_1\mu_N(k) = \frac{Nk}{2}$$

Proof. From identity (5.6), page 9

$$P_N(k, R) = P_N(k, Nk - R)$$

Applying Remark 10.3,

$$\begin{aligned}
{}_1\mu_N(k) &= \sum_{R=0}^{Nk} RP_N(k, R) \\
&= \sum_{R=0}^{Nk} (Nk - R)P_N(k, Nk - R) \\
&= \sum_{R=0}^{Nk} NkP_N(k, Nk - R) - \sum_{R=0}^{Nk} RP_N(k, R) \\
&= Nk - {}_1\mu_N(k)
\end{aligned}$$

Hence,

$${}_1\mu_N(k) = \frac{Nk}{2}$$ □

Remark 10.5. The second moment for the null distribution is

$${}_2\mu_N(k) = \frac{Nk}{12}(3Nk + N + 2) \tag{10.1}$$

Proof. by induction on k:

From Remark 10.3,

$${}_2\mu_N(k) = \sum_{R=0}^{Nk} R^2 P_N(k, R)$$

Basic step $k = 1$, any N

$$\frac{N}{12}(3N + N + 2) = \frac{N(2N + 1)}{6}$$

From identity (5.2), page 8,

$$ {}_2\mu_N(1) = \sum_{R=0}^{N} R^2 P_N(1,R) = \frac{1}{N+1}\sum_{R=0}^{N} R^2 = \frac{N(2N+1)}{6} $$

Hence, when $k = 1$, with any N, the expression 10.1 holds.

Inductive step on k: assuming the expression holds at $k \le k_0$, with any N. Prove for the case $k = k_0 + 1$:

$$ \begin{aligned} {}_2\mu_N(k_0+1) &= \sum_{R=0}^{N(k_0+1)} R^2 P_N(k_0+1,R) \\ &= \sum_{R=0}^{N(k_0+1)} \left(R^2 \frac{1}{N+1}\sum_{j=0}^{N} P_N(k_0, R-j) \right) \end{aligned} $$

Substitute the following equation

$$ R^2 \frac{P_N(k_0,R-j)}{N+1} = \frac{1}{N+1}[(R-j)^2 + 2j(R-j) + j^2]P_N(k_0,R-j) $$

into the above expression for ${}_2\mu_N(k_0+1)$ gives:

$$ \begin{aligned} &{}_2\mu_N(k_0+1) \\ &= \frac{1}{N+1}\sum_{R=0}^{N(k_0+1)}\sum_{j=0}^{N}[(R-j)^2 + 2j(R-j) + j^2]P_N(k_0,R-j) \\ &= \frac{1}{N+1}\sum_{R=0}^{N(k_0+1)}\sum_{j=0}^{N}[(R-j)^2 P_N(k_0,R-j) + 2j(R-j)P_N(k_0,R-j) + j^2 P_N(k_0,R-j)] \end{aligned} $$

Notice that the double summation $\sum_{R=0}^{N(k_0+1)}\sum_{j=0}^{N}$ always includes all values from 0 to R for the term $(R-j)$. From identity (5.5), page 9,

$$ P_N(k,R) = 0 \text{ if } R < 0 \text{ or } R > Nk $$

All values larger than Nk or less than 0 of the term $(R-j)$ under the double summation $\sum_{R=0}^{N(k_0+1)}\sum_{j=0}^{N}$ are discarded. The summation $\sum_{R=0}^{N(k_0+1)}$ yields:

- ${}_2\mu_N(k_0)$ from $(R-j)^2 P_N(k_0,R-j)$,
- ${}_1\mu_N(k_0)$ from $(R-j)P_N(k_0,R-j)$,
- 1 from $P_N(k_0,R-j)$.

Similarly, the summation $\sum_{j=0}^{N}$ yields $(N+1)$, $2\frac{N(N+1)}{2}$, and $\frac{N(N+1)(2N+1)}{6}$ from 1, $2j$, and j^2. Hence,

$$ \begin{aligned} &{}_2\mu_N(k_0+1) \\ &= \frac{1}{N+1}\left((N+1)\,{}_2\mu_N(k_0) + 2\frac{N(N+1)}{2}\,{}_1\mu_N(k_0) + \frac{N(N+1)(2N+1)}{6} \right) \\ &= \frac{N(k_0+1)}{12}(3N(k_0+1) + N + 2) \end{aligned} $$

Therefore, the expression 10.1 holds for all k. □

Remark 10.6. The third moment for the null distribution is

$$ {}_3\mu_N(k) = \frac{N^2k^2}{8}(Nk + N + 2) $$

Remark 10.7. The fourth moment for the null distribution is

$$ {}_4\mu_N(k) = \frac{Nk}{240}\left[\left(N^3A_k\right) + \left(4N^2B_k\right) + (4NC_k) - 8\right] $$

with

$$ A_k = 15k^3 + 30k^2 + 5k - 2 $$
$$ B_k = 15k^2 + 5k - 2 $$
$$ C_k = 5k - 3 $$

Similar to Remark 10.5, Remark 10.6 and Remark 10.7 are proven by induction on k.

References

[1] Magnusson, K. R Psychologist [online]. Access date 2016-04-10. Available at http://rpsychologist.com/d3/NHST/

[2] Santos JL, Saus E, Smalley SV. Copy Number Polymorphism of the Salivary Amylase Gene: Implications in Human Nutrition Research. Journal of Nutrigenetics and Nutrigenomic. 2012, 5(3):117-131

[3] Faraone SV, Doyle AE, Mick E, Biederman J. Meta-analysis of the association between the 7-repeat allele of the dopamine D(4) receptor gene and attention deficit hyperactivity disorder. The American Journal of Psychiatry. 2001, 158(7):1052-7

[4] Clancy S. Copy Number Variation. Nature Education. 2008, 1(1):95

[5] Perry GH, et al. Diet and evolution of human amylase gene copy number variation. Nature Genetics. 2007, 39(10):1256-60

[6] Aitman TJ, et al. Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. Nature. 2006, 439 (7078):851-855

[7] Abyzov A, Mariani J, Palejev D, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature. 2012, 492 (7429):438-42

[8] Bruder C, Piotrowski A, Gijsbers A, et al. Phenotypically Concordant and Discordant Monozygotic Twins Display Different DNA Copy-Number-Variation Profiles. American Journal of Human Genetics. 2008, 82(3):763-771

[9] Szuhai K, Vermeer M. Microarray Techniques to Analyze Copy-Number Alterations in Genomic DNA: Array Comparative Genomic Hybridization and Single-Nucleotide Polymorphism Array. Journal of Investigative Dermatology. 2015, 135

[10] Martin C, Kirkpatrick B, Ledbetter D. Copy Number Variants, Aneuploidies, and Human Disease. Clinics in Perinatology. 2015, 42(2):227-242

[11] LaFramboise T, Winckler W, Thomas R. A flexible rank-based framework for detecting copy number aberrations from array data. Bioinformatics. 2009, 25(6):722-728

[12] Wineinger NE, Kennedy RE, Erickson SW, et al. Statistical issues in the analysis of DNA Copy Number Variations. Int J Comput Biol Drug Des. 2008, 1(4):368-95

[13] University of Alabama in Huntsville. Irwin-Hall distribution [online]. Access date 2016-04-10. Available at http://www.math.uah.edu/stat/special/IrwinHall.html

[14] University of Alabama in Huntsville. Skewness and Kurtosis [online]. Access date 2016-04-10. Available at http://www.math.uah.edu/stat/expect/Skew.html

[15] Wolfram MathWorld [online]. Access date 2016-04-10. Available at http://mathworld.wolfram.com/Moment-GeneratingFunction.html

[16] Cohen H, Frey G, Avanzi R, et al. Handbook of Elliptic and Hyperelliptic Curve Cryptography. Boca Raton (FL): CRC Press; 2006. 808 p

[17] Engel T, Reid P. Physical Chemistry. San Francisco (CA): Person; 2006. 1061 p

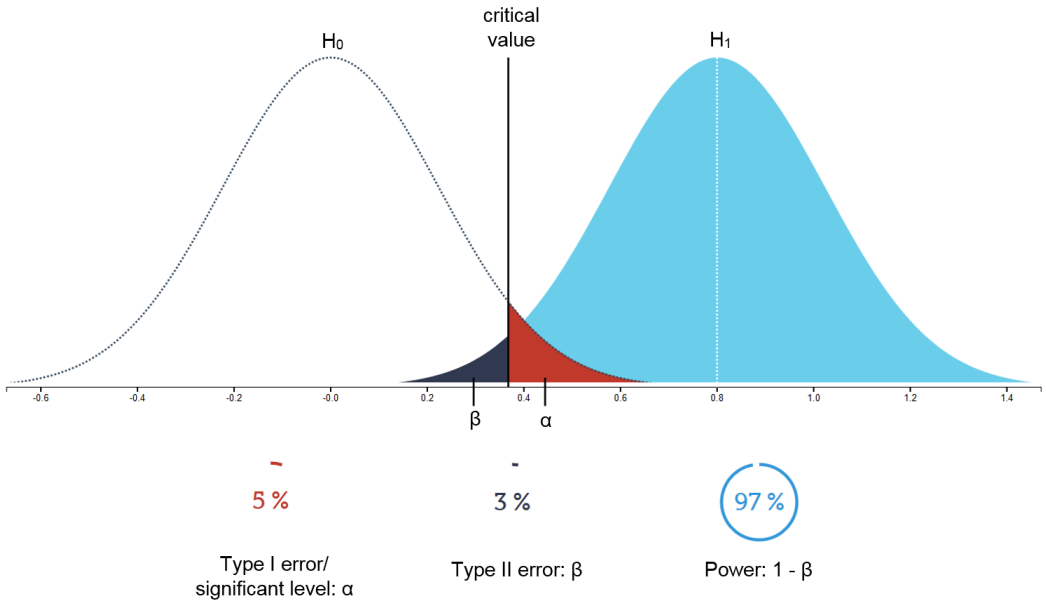
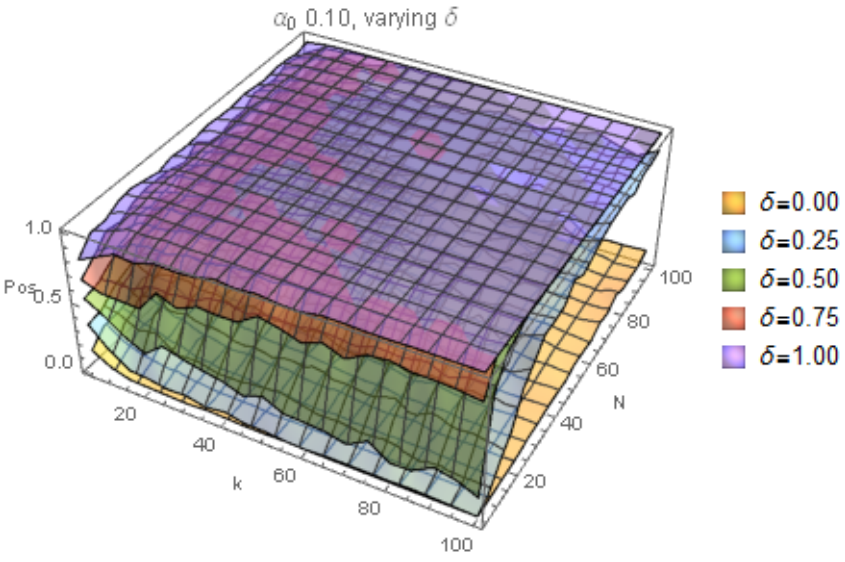
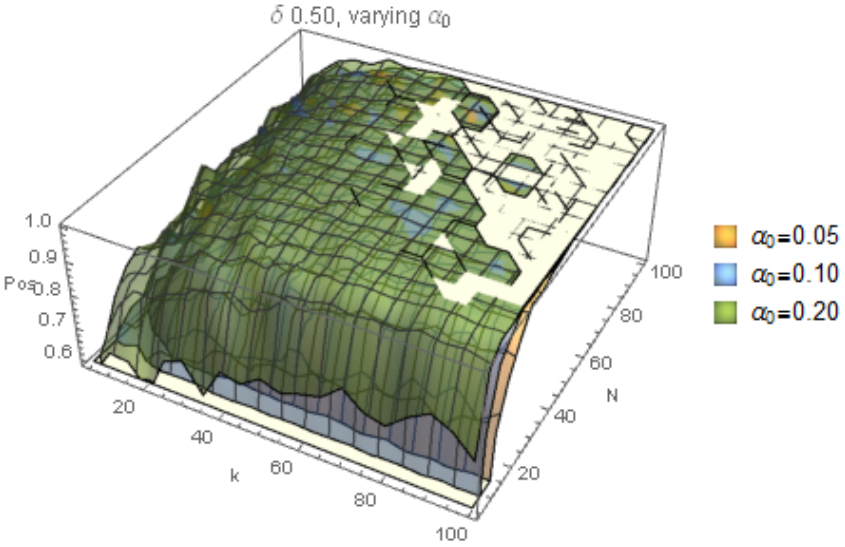


Figure 2: Visualization of H_0 and H_1 distributions for high Rank-Sum value, adapted from [1]

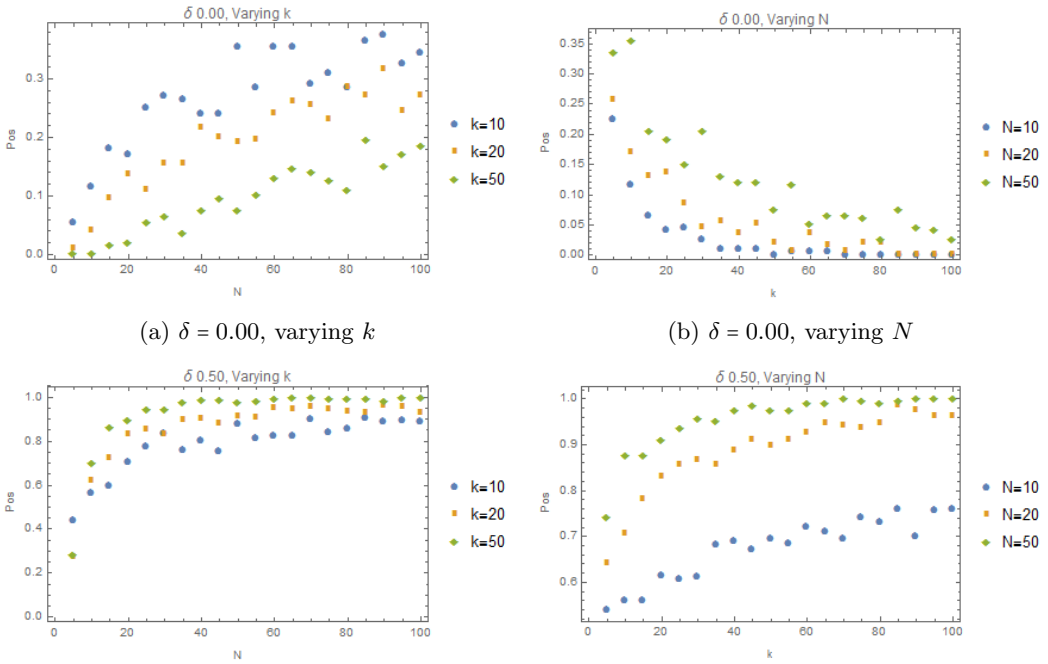


(a) α_0 = 0.10, varying δ



(b) δ = 0.50, varying α_0

Figure 3: Exemplary surfaces for Pos
Key to abbreviations: page 13



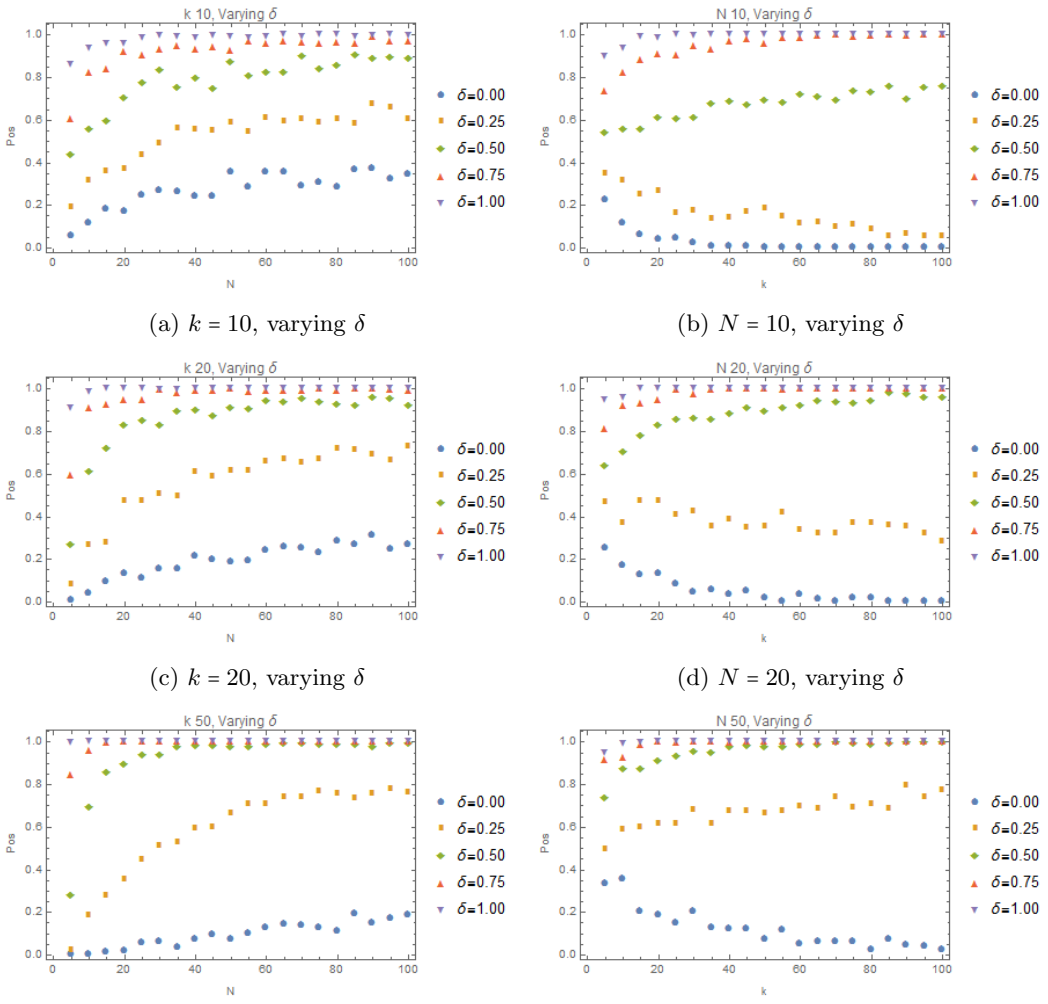
(a) $\delta = 0.00$, varying k

(b) $\delta = 0.00$, varying N

(c) $\delta = 0.50$, varying k

(d) $\delta = 0.50$, varying N

Figure 4: Cross-sections for Pos at $\alpha_0 = 0.10$
Key to abbreviations: page 13



(a) $k = 10$, varying δ

(b) $N = 10$, varying δ

(c) $k = 20$, varying δ

(d) $N = 20$, varying δ

(e) $k = 50$, varying δ

(f) $N = 50$, varying δ

Figure 5: Cross-sections for Pos at $\alpha_0 = 0.10$
Key to abbreviations: page 13

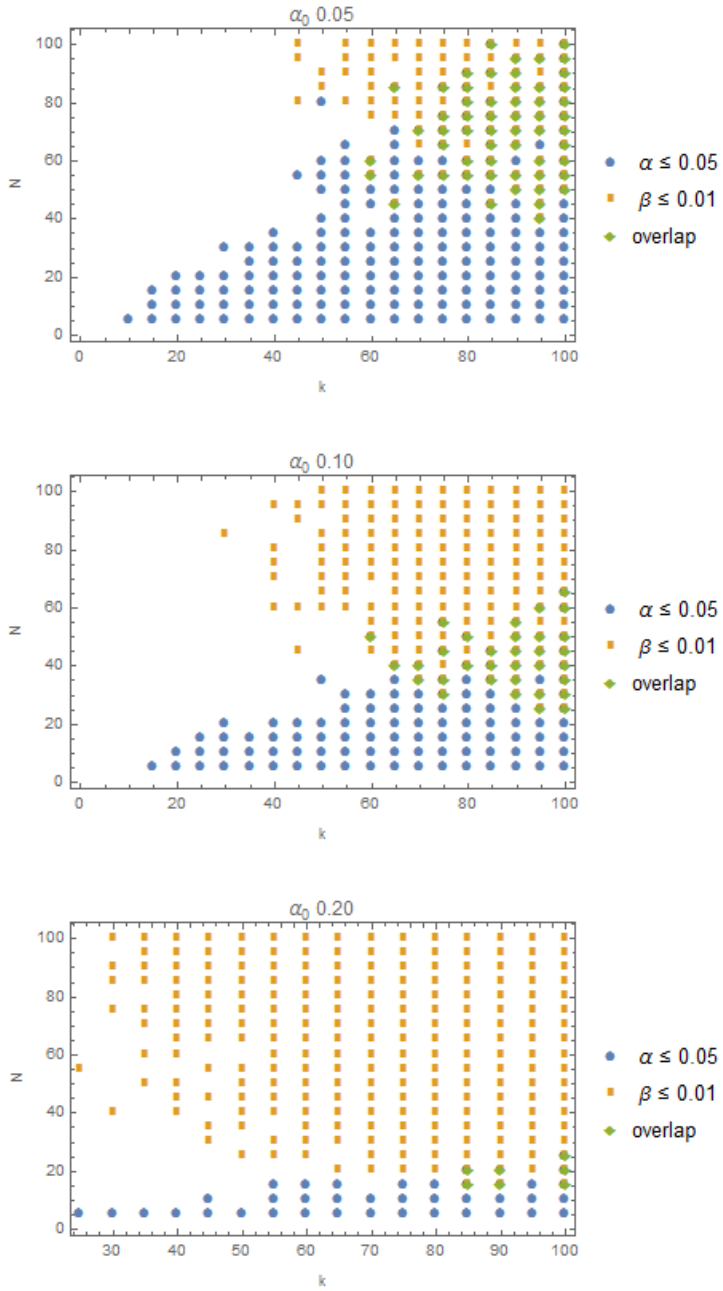
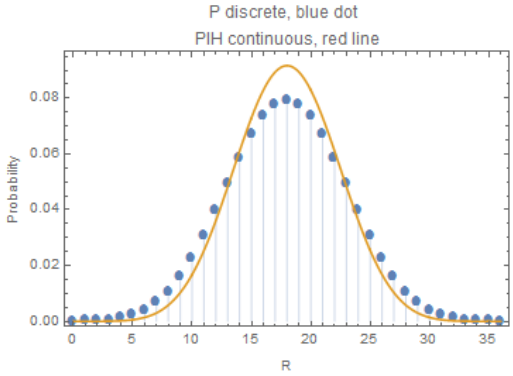
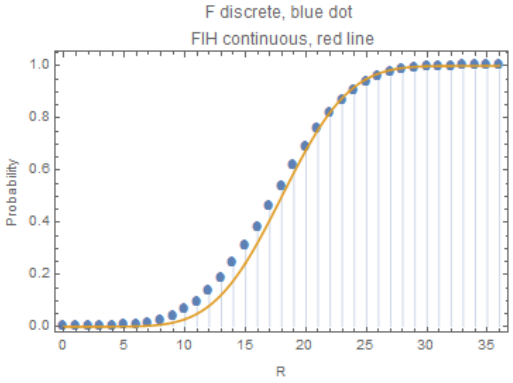


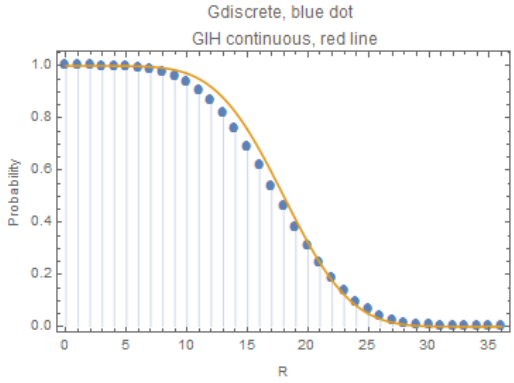
Figure 6: Projections of Figure 3b: data points that satisfy conditions on α, β. All points satisfying $\alpha \le 0.05$ at $\delta = 0$ are denoted by blue dots and green diamonds. All points satisfying $\beta \le 0.01$ or power $1-\beta$ larger than 0.99 at $\delta = 0.50$ are denoted by red squares and green diamonds (green diamonds denote the overlapping points satisfying both conditions on α and β).



(a) P and PIH

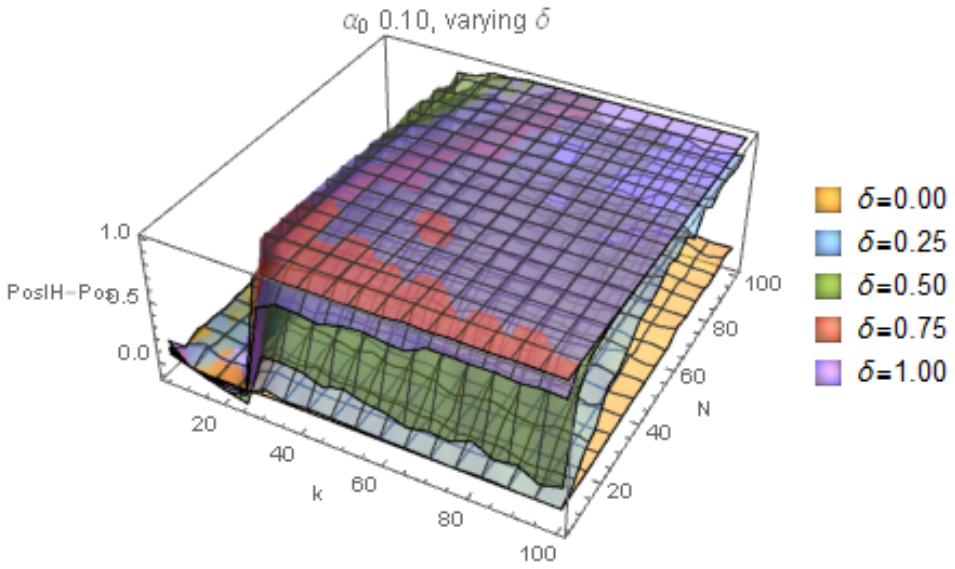


(b) F and FIH

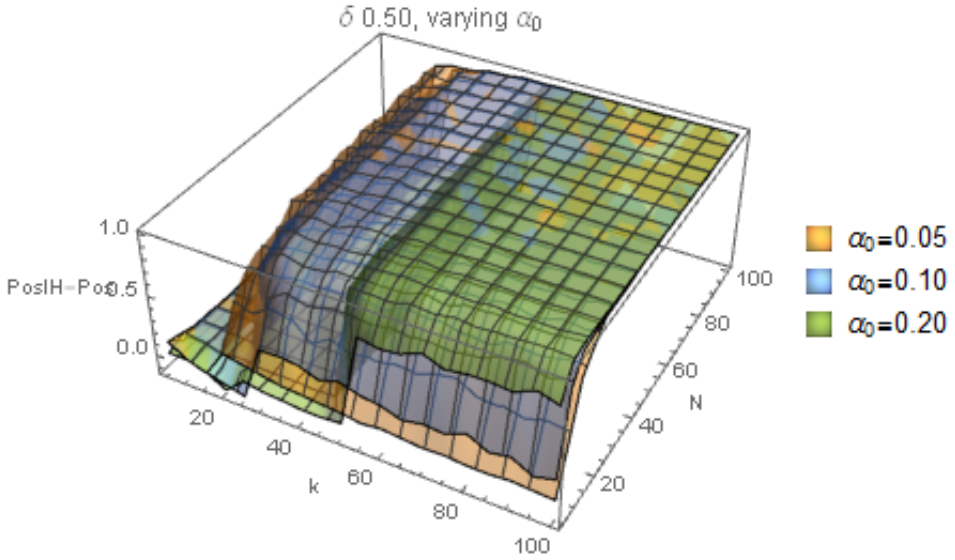


(c) G and GIH

Figure 7: the probability mass/density functions (a) and the cumulative distributing functions (b), (c) for the null distribution and the IH approximation; at the particular case $k = 6$ and $N = 6$

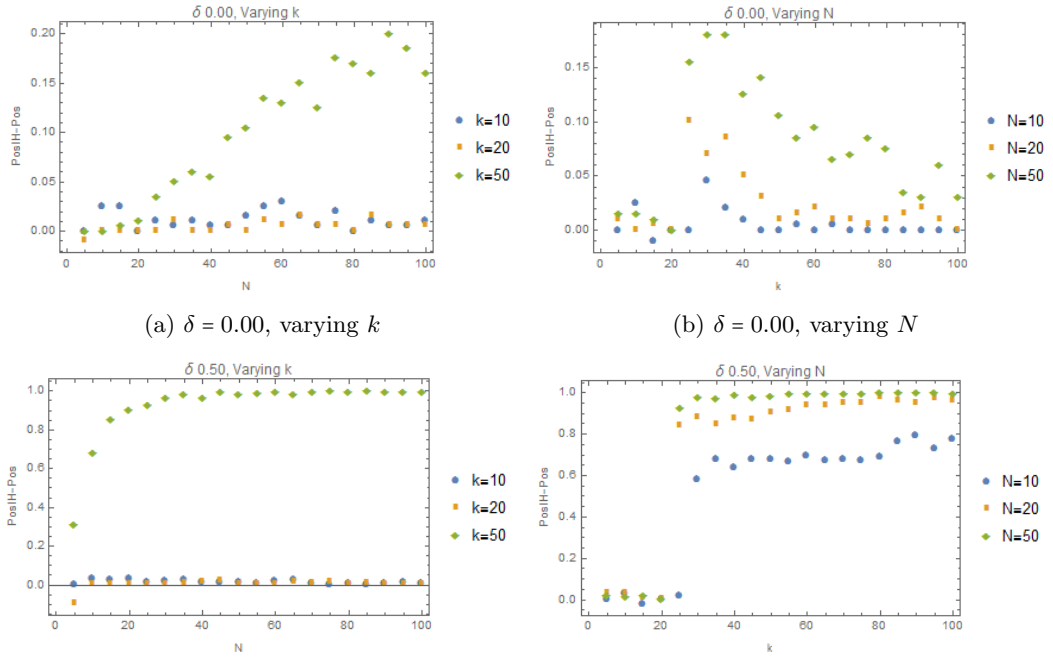


(a) α_0 = 0.10, varying δ



(b) δ = 0.50, varying α_0

Figure 8: Exemplary surfaces for PosIH-Pos
Key to abbreviations: page 13



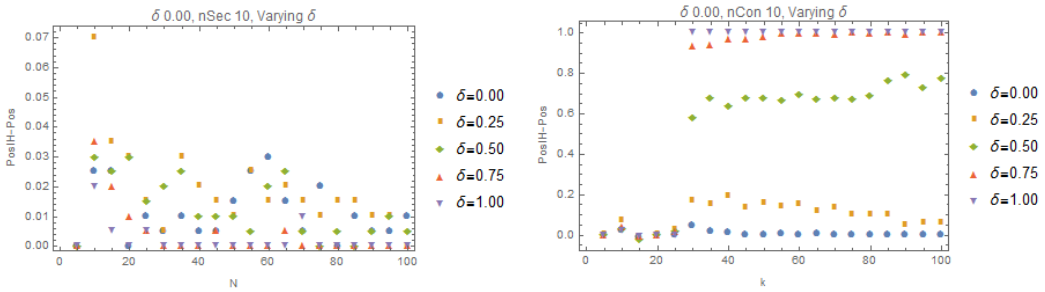
(a) $\delta = 0.00$, varying k

(b) $\delta = 0.00$, varying N

(c) $\delta = 0.50$, varying k

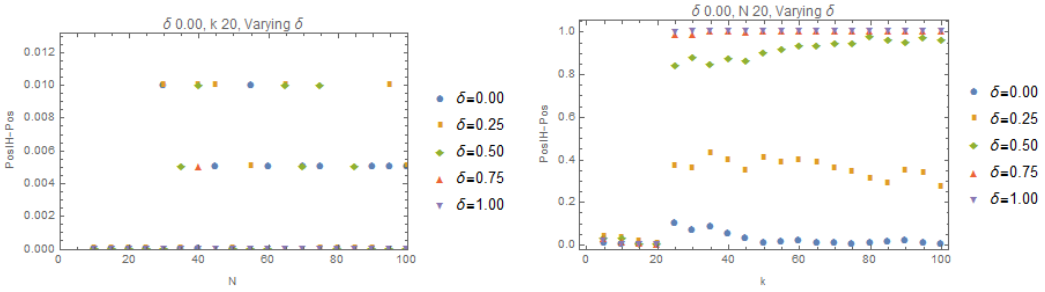
(d) $\delta = 0.50$, varying N

Figure 9: Cross-sections for the difference PosIH - Pos at $\alpha_0 = 0.10$
Key to abbreviations: page 13



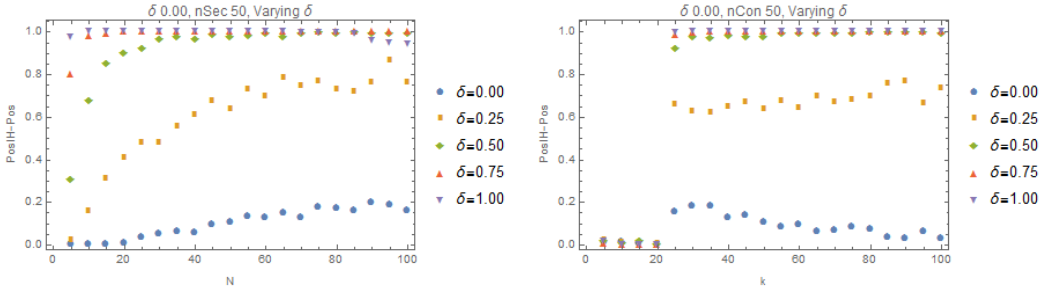
(a) $k = 10$, varying δ

(b) $N = 10$, varying δ



(c) $k = 20$, varying δ

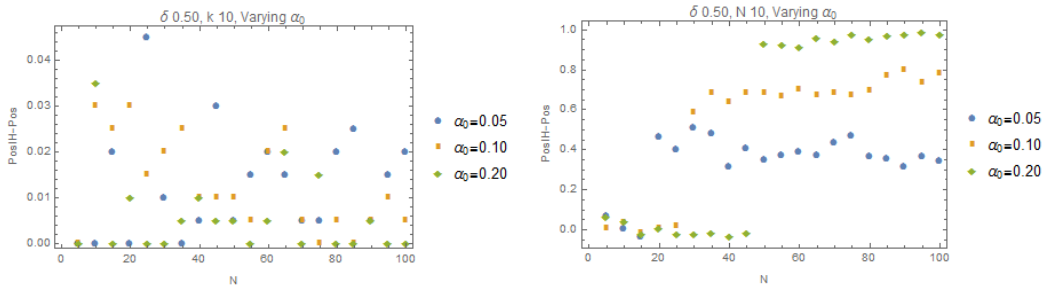
(d) $N = 20$, varying δ



(e) $k = 50$, varying δ

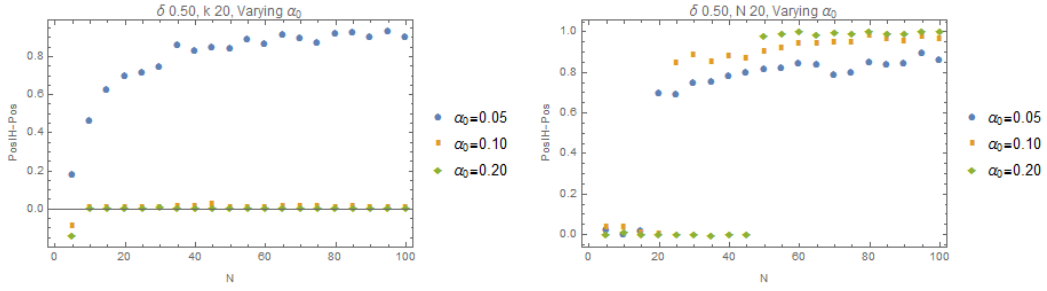
(f) $N = 50$, varying δ

Figure 10: Cross-sections for the difference PosIH - Pos at $\alpha_0 = 0.10$
Key to abbreviations: page 13



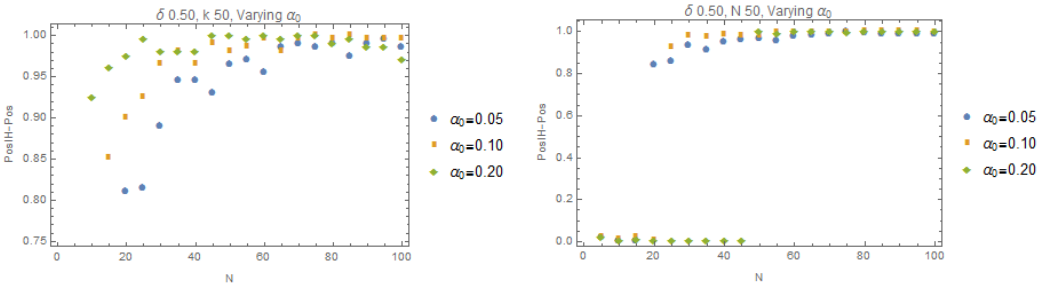
(a) $k = 10$, varying α_0

(b) $N = 10$, varying α_0



(c) $k = 20$, varying α_0

(d) $N = 20$, varying α_0



(e) $k = 50$, varying α_0

(f) $N = 50$, varying α_0

Figure 11: Cross-sections for the difference PosIH - Pos at $\delta = 0.50$
Key to abbreviations: page 13