ESTIMATION OF GENOMIC COPY FREQUENCY WITH CORRELATED OBSERVATIONS

by

QIANQIAN TANG

(Under the Direction of Jaxk Reeves)

ABSTRACT

In this thesis, we compare several methods to handle correlated data related to genome frequency copies. First, we used standard Poisson Regression to analyze the data. From the results, we find that there are several problems related to over-dispersion and under-dispersion. It is easy to handle over-dispersion using the 'scale-adjustment' method. However, remedying problems related to dependence caused by correlated Poisson data are not so easily handled. We first created a statistic to help us test the null hypothesis that data are independent Poisson realizations vs. the alternative that they are positively associated. From this, we found that 225 base-pairs separation is the minimum cut-off distance needed to achieve approximate independence. We also used results from this analysis to devise a formula which yields the approximate correlation coefficient (r) between counts which are separated by 'b' base-pairs. Finally, we use our method to weight observations, and find significant improvement compared to other methods.

INDEX WORDS: Poisson Regression, Over-dispersion, Under-dispersion, Dependent Weighting Scheme

ESTIMATION OF GENOMIC COPY FREQUENCY WITH CORRELATED OBSERVATIONS

by

QIANQIAN TANG

M.S., Shaanxi Normal University, Xi'an, China, 2005

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2012

ESTIMATION OF GENOMIC COPY FREQUENCY WITH CORRELATED OBSERVATIONS

by

QIANQIAN TANG

Major Professor: Jaxk Reeves

Committee: Lily Wang
Liang Liu

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2012

DEDICATION

This paper is dedicated to my family.

ACKNOWLEDGEMENTS

First and foremost, I'd like to express my sincere appreciation to my advisor, Professor Jaxk Reeves. I couldn't have finished my research so smoothly and quickly without his insightful guidance and unlimited patience. He has always given me strong support and useful suggestions during my studies.

I would like to thank two other members in my advisory committee: Assistant Professor Lily Wang and Assistant Professor Liang Liu. The final version of this thesis has benefited from their proofreading.

I would also like to thank my family and friends for their continuous support, love and encouragement.

Thanks to the UGA Statistics Department for giving me an unforgettable and valuable graduate study.

Without all of this encouragement and help, it would be hard for me to finish this degree.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 THE WHEAT GENOME

Bread wheat, *Triticum aestivum* L., is an allohexaploid that was formed by two spontaneous hybridization events. The first event took place some 500,000 years ago between the A-genome species *T. urartu* and an unknown B-genome species to form tetraploid *T. turgidum* ssp. *dicoccoides* (AABB) (Huang *et al.* 2002). The formation of hexaploid wheat occurred some 8,500 years ago through the hybridization of a cultivated tetraploid, *T. turgidum* ssp. *dicoccum* (AABB) with the D-genome species *Ae. tauschii* (Nesbitt & Samuel 1996). As a result of the polyploidization, most genes are present in three copies on homoeologous chromosomes (McIntosh *et al.* 2003).

In recent years, great progress has been made in understanding the structure of the wheat genome. A physical map of the D genome of *Ae. tauschii* is near completion and physical mapping of individual chromosomes from hexaploid wheat is coordinated through the International Wheat Genome Sequencing Consortium (IWGSC; http://www.wheatgenome.org). The ultimate goal of the IWGSC is to produce a full sequence assembly of the hexaploid genome. Sequence analysis of 12 contigs from *Ae. tauschii* totaling 11.5 Mb and of 13 contigs from hexaploid wheat chromosome 3B totaling 18.2 Mb has provided insight into the organization of genes and repeats (Devos 2010; Choulet *et al.* 2010; Massa *et al.* 2011). Sequencing and precise annotation of 192 randomly selected BAC clones of hexaploid wheat has indicated that the wheat genome

contains around 110,000 ± 22,000 genes (JL Bennetzen and KM Devos, unpublished data). The entire data set of wheat genes will become available later this year following the completion and annotation of (1) the 5X shotgun sequence of the hexaploid wheat variety Chinese Spring (CS) generated mainly on a Roche 454 platform by a UK consortium (http://www.cerealsdb.uk.net) and (2) the 50X shotgun sequences of the CS AABBDD genome and of the AA, AABB and DD progenitor genomes generated by a team at Cold Spring Harbor Laboratories (CSHL) using mainly Illumina paired-end sequencing. There are also more than 1,000,000 ESTs (http://www.n(cbi.nlm.nih.gov/) and some 8,500 putative full-length cDNAs (http://trifldb.psc.riken.jp/index.pl) available for wheat which, together with comparative information from other species, are valuable resources for annotation of the wheat genome. The ESTs have been assembled into Unigenes (http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=4565) and in 2004, an Affymetrix microarray was generated using information from GenBank *T. aestivum* Unigene Build #38, which included 414,006 ESTs and 1,767 mRNAs, and from available *T. monococcum*, *T. turgidum* and *Ae. tauschii* ESTs. The microarray contains 61,127 probe sets representing 55,052 transcripts distributed over all 21 wheat chromosomes. A probe set consists of 11 25-mers that were designed mostly against regions in Unigene clusters that are conserved across the A, B, and D genomes.

The Affymetrix wheat gene chip has been used to study gene expression during wheat development, and in response to biotic and abiotic stresses (Crismani *et al.* 2006; Desmond *et al.* 2008; Schreiber *et al.* 2009; Winfield *et al.* 2009). Because most of the probe sets on the Affymetrix gene chip have been designed against regions that are conserved between the A, B, and D genomes of wheat, the expression profiles obtained

for an estimated 90% of the genes in those experiments are the sum of the expression profiles across the three wheat genomes (Schreiber *et al.* 2009). Akhunova *et al.* extracted some information on homoeolog-specific expression from the Affymetrix gene chip by identifying intergenomic SNPs in the probe sets that differentiated the D genome from the AB genomes. The limitations of this approach are that (1) not all three genomes can be differentiated; (2) for each genome-specific probe, expression can be measured only for the genome that has a perfect match; and (3) the expression measured for two different genomes by two different probes might be influenced by the location of each probe in the transcript (e.g. 5' located probes might show lower apparent expression due to a 3' transcript bias than 3' probes).

1.2 NON-HOMOEOLOGOUS GENE EXPRESSION IN WHEAT

The first study of differential expression in A, B, and D homoeologs in hexaploid wheat was carried out by Mochida et al. (2004). They selected 90 relatively abundant genes based on their prevalence in EST datasets and associated single nucleotide polymorphism (SNP) haplotypes with the A, B, and D homoeologs using nullisomic-tetrasomic (NT) analysis in combination with pyro-sequencing. NT lines lack one chromosome pair, and have an extra copy of a homoeologous chromosome pair. Such lines exist for each of the 21 wheat chromosomes (Sears 1954) and can be used to allocate markers to chromosomes. Relative expression in the A, B and D genomes was determined by assessing the number of ESTs with an A, B and D genome haplotype in a set of 116,232 ESTs generated from 10 tissues. Sixteen percent of the genes had similar

expression levels in the three wheat genomes, and the remaining 84% showed different expression of one homoeolog in at least one tissue. Of the genes with preferential expression, 17% were expressed in only two of the three genomes.

Homoeologous gene silencing in wheat was also demonstrated by Bottley et al. (2006). A, B and D amplicons obtained within exons of 236 single-copy genes in seedling leaves and roots were separated by single strand conformation polymorphism (SSCP) gel electrophoresis and allocated to a genome by comparing the cDNA amplicon patterns with the gDNA patterns amplified from NT lines. Absence of an amplification product from one of the three genomes was observed for 27% of the genes in leaves and 26% of the genes in root. As in Mochida et al. (2004), there appeared to be no bias towards a specific genome. If silencing occurred in the same tissue in multiple varieties, it was often the same homoeolog that was silenced (Bottley & Koebner 2008). Overall patterns of silencing, however, appeared to be variety dependent and heritable (Bottley & Koebner 2008).

The effects of allopolyploidization *per se* were examined by hybridizing RNA isolated from seedling leaves from *T. turgidum* (AABB), *Ae. tauschii* (DD) and a corresponding synthetic hexaploid to a 17K 70-mer oligoarray (Pumphrey *et al.* 2009). The oligoarray contained 35,568 features representing 17,279 potentially unique wheat genes, and expression was measured globally, that is summed over the A and B genomes in case of the tetraploid *T. turgidum* and summed over the A, B, and D genomes in case of the synthetic hexaploid. Approximately 78% of the transcripts showed differential expression between *Ae. tauschii* and *T. turgidum*. This is considerably higher than expression differences observed between the diploid relatives *Arabidopsis thaliana* and

A. arenosa and might be attributed to the fact that *T. turgidum* itself is an allotetraploid that resulted from a merger between two diploid genomes some 500,000 years ago. The number of genes that were expressed in a non-additive manner in the synthetic allohexaploid compared to the diploid and tetraploid parent was 16% (Pumphrey *et al.* 2009). A similar study including *Ae. tauschii,* an AABB synthetic tetraploid and an AABBDD synthetic hexaploid, was carried out by Akhunova et al. using intergenomic SNPs to distinguish expression of the D genome from that of the AB genomes. Parental divergence was, again, found to be an important factor contributing to differential expression of the A, B, and D genomes. The percentage of genes for which the A, B, and D genomes were expressed non-additively was 19%.

CHAPTER 2

SPECIFIC PROBLEM

2.1 THE OVERALL GOAL AND OBJECTIVES OF THE THESIS

The overall goal of the research in this area is to understand the effects of polyploidization on the sub/neofunctionalization of homoeologous gene copies in hexaploid wheat and to identify the genetic and epigenetic factors that contribute to biased expression of gene homoeologs during all stages of the polyploidization process. From a statistical perspective, the goal is to use the data on copy number most efficiently to test the null hypothesis that all three wheat homoeologs (A, B, D) are expressed at equal levels. In even simpler terms, for each gene for which we have data, we wish to estimate the mean expression count level for each homeolog and to test whether these means are significantly different from one another.

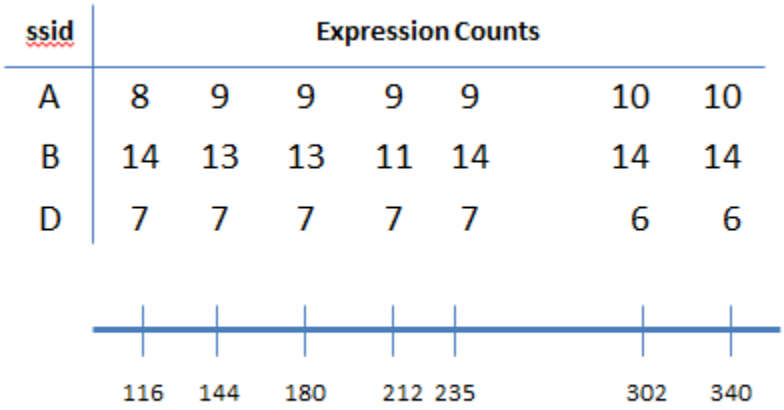
2.2 DESCRIPTION OF DATA

To have the necessary statistical power to test for differential expression of homoeologs, at least two independent measurements of expression levels are needed for each gene. Unfortunately, with technology available and currently affordable, one can't measure expression level exactly and must count copies in the amplification process in order to estimate expression level. This is an indirect measure, since the count depends on how many 'fragments' overlap sufficiently with probes designed to measure certain

signature sequences in each gene. For the data set which is the focus of this thesis, a minimum of two and up to 14 sets of signature sequences were extracted for each of 61 key genes in the wheat genome. The number of signature sequences identified per gene correlated only weakly with transcript length (r=0.26, p=0.04), indicating that some genes have higher levels of intergenomic variation than others. The expression levels for each homoeolog were calculated (1) by averaging the coverage at each base of the signature sequence in the transcript assemblies and (2) by counting the number of perfect matches against the raw 454 reads. We used method (2) in performing our analyses.

Figure 1 presents a schematic diagram of typical results. In that example (from the first of 61 key wheat genes analyzed as part of this research), there are 7 signature sequences contained in the gene, at locations which are 116, 140, …, 340 base pairs after the beginning of the gene. Signature sequences are short stretches (usually of length 25 base pairs) of nucleotide patterns that are known to occur frequently in the wheat genome. A probe to detect the signature sequence is run at each locus on each of the three homoeologs (A, B, D), and a count is made of the number of times the sequence is detected for each homoeolog at that point. From the data shown, it appears that the typical copy numbers for the three homoeologs in this gene are about {9, 13, 7}, respectively. One would like to perform a test to see if this variation is sufficient to conclude that copy number is unequal across the three homoeologs. If one assumes that the counts in each homoeolog are independent realizations from a homogeneous Poisson process, this is easy to do. However, as we shall see, such assumptions are unrealistic, leading us to the fundamental statistical problem investigated in this thesis.

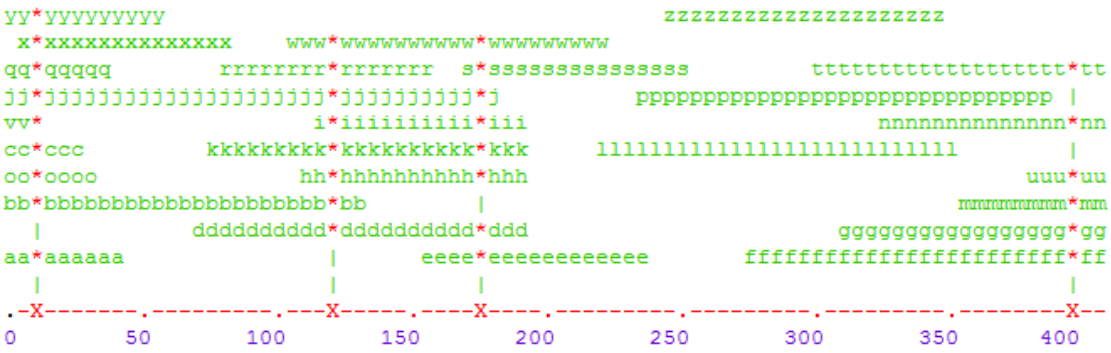
Figure 1 Schematic Diagram



At first, this seems to be a quite simple statistical exercise, but Figure 2 illustrates more clearly what happens within a homoeolog of a gene. The four 'X' at approximate positions 20, 140, 200, and 400 base pairs from the 'left' end of the gene represent signature sequence sites which are probed. The lettered sequences represent different fragments of the amplified and then shredded gene. The fragments are usually of similar, but certainly not identical lengths, and not all fragments are 'hit', even when they do contain the matching signature sequence. If a probe 'hits' a fragment, that means that the probe detected a sequence in that fragment which matched the signature sequence. In Figure 2, there are {9,8,8,6} hits, respectively for the four signature sites. From this, one might infer that the copy counts for sites in this homoeolog of the gene are distributed approximately as a Poisson random variable with mean = (9+8+8+6)/4 = 7.75. This sort of calculation might be appropriate if the counts at the sites were obtained independently of one another, as appears to be the case for the first and last of the four signature probes. However, for the two middle sites, the identical observed counts of '8' are much more than a coincidence; the two probes hit the same fragments 6 times. Even the first and

second sites, which are more separated than the middle two, share hits on two of the same fragments, so the ‘9’ and ‘8’ observed there are not quite independent counts. Unfortunately, because of limitations in how the data are collected, there is no way to know if counts at different sites include overlapping fragments. The crux of the problem in this thesis is to determine both how to estimate the true Poisson intensity parameter in such cases, and how to quantify the uncertainty in these estimates. Solving the latter problem allows us to make valid inferences concerning the hypothesis of equality of intensity across the three homeologs of a gene, the primary question of genetic interest.

Figure 2 Refined Schematic of Copy Counts Within a Homoeolog



2.3 NAIVE METHOD OF ANALYSIS

This method of analysis includes all expression counts and assumes that all the expression counts within a specific gene of a specific homeolog are independent of one another . Because of this independence assumption, we use the average of the expression

counts as estimates of the mean expression counts for specific genes and homoeologs. That is, at the most naïve level, for each of the 61 genes examined, we simply compared the mean counts over the several observations taken on each of the three (A, B, D) homoeologs and used Poisson regression techniques to test the null hypothesis that the three homoeologs had the same mean expression level. A slightly more sophisticated analysis controlled for site location (since all measurements along a specific homoeolog were taken at the same signature sequence sites), and there has been some suggestion in the literature that certain sites, especially those in G-C rich areas may tend to have larger counts than others. For the most part, however, effects due to location were either insignificant or not nearly as significant as effects due to homoeolog. The effects due to homoeolog were not at all consistent from gene-to-gene, as can be seen from the output in Table A1 of the Appendix. For each of the 61 genes, Table A1 displays the number of signature sequence sites (#SSNs), the Deviance statistic from the Chi-Squared test, and the two-tailed P-values for testing that the A-B, A-D, and B-D homoeologs had different intensities. In some cases, there were no significant differences between the mean counts in the three homoeologs, while, in others, all sorts of different orders appeared.

To perform these naïve Poisson regression analyses, we used PROC GENMOD of SAS version 9.3. Upon more careful examination of the results, we find that there are several potential violations of model assumption: (1) over-dispersion; (2) dependence within a gene; (3) dependence across homoeologs.

For a one-parameter exponential family (such as the Poisson), over-dispersion is a situation which occurs when the mean estimated from the data gives rise to a theoretical standard deviation estimate which is too small compared to the observed standard

deviation. It is a not uncommon problem encountered in Poisson regression, and is frequently 'handled' by employing a scale multiplier to adjust standard errors for this extra variability. (Within PROC GENMOD, this is most frequently done via the SCALE=D or SCALE=P options, or occasionally by generalizing the Poisson to a Negative Binomial distribution.) While there is some evidence of over-dispersion in the homoeolog data from the 61 genes, a much more common occurrence is not that the data are over-dispersed, but that they are under-dispersed. That is, although the theoretical standard deviation of a Poisson distribution is equal to the square root of the theoretical mean, there are many more gene-homoeologs sets for which the sample SD is exceeded by the square-root of the sample mean ($SD << \sqrt{\bar{X}}$) than by the converse ($SD >> \sqrt{\bar{X}}$). Although it is relatively easy to handle over-dispersion by using the scale statement in PROC GENMOD, such an approach is not recommended for under-dispersion. The reason for this is that over-dispersion typically occurs in cases where the independence assumption still seems tenable – the data are just more spread out relative to the sample mean than they should be under Poisson assumptions. For under-dispersion, on the other hand, the main culprit is *positive dependence* – successive observations tend to be more similar to one another than they should under independence assumptions. This can be seen to occur often in the gene data set of Appendix A – there is much too little variation across the values in the same homoeolog for many of the genes, especially when the signature sequence sites are relatively close to one another. However, how to handle the difficulties related to this dependence is not easy.

CHAPTER 3

APPROACHES

3.1 CREATING A STATISTIC TO TEST INDEPENCE

We want to find a statistic which can test whether the observations (counts) within the same homoeolog behave like realizations from an independent Poisson distribution. For a random variable X following a Poisson distribution with intensity parameter λ , the probability density function is given by:

$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$, x=0, 1, 2, … with mean and variance both equal to λ. Thus, it would be natural to examine the ratio of the sample variance to the sample mean ($\frac{s^2}{\bar{x}}$) as a statistic to measure relative dispersion, with ratios much greater than 1.0 indicating over-dispersion, and ratios much less than 1.0 indicating under-dispersion. A question of interest is what function of the above ratio is most useful for measuring significance. We tried estimators of the form $Z_n = a_n \times [f(\frac{s^2}{\bar{x}}) - f(1)]$ in an attempt to find an (a_n, f) pair which would have approximately a standard normal distribution even for small values of *n* , such as those which are common in our dataset ($2 \leq n \leq 7$ are typical for the dataset in question). After some trial and error, we found that $a_n = n$ and $f(x) = x^{1/4}$ behave reasonably well for a range of λ ($0.5 \leq \lambda \leq 20$). That is, we hoped that the statistic $Z = n\left[\sqrt{\frac{s}{\sqrt{\bar{x}}}} - 1\right]$ can be reasonably approximated by a standard normal distribution.

We used simulations to test this assumption. We simulated different sets of independent Poisson distributed random variables for different sizes *(n*) and different intensities (λ). The sizes ranged from 2 to 7, and λ ranged from 0.5 to 20 in increments of 0.5. We simulated 999 times for each combination of *n* and λ. For each run, the statistic $Z=n\left[\sqrt{\frac{s}{\sqrt{\bar{x}}}}-1\right]$ was calculated. In the rare case that $\bar{x}=0$, we set Z= -n, the minimum value which is achieved whenever *s*=0. Then, for each set of 999 simulations, we calculated the mean, standard deviation and different percentiles (1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 99%) for Z. These results are displayed in Tables A2 (*n=2*) to A7 (*n=7*) in the Appendix. From these results, we can see that the hoped for convergence to Standard Normal percentiles didn't quite occur. The standard deviation of the Z-statistic, as *n* increases, appears to be approaching 1, but there is a consistent bias in the mean. Of course, one can use the empirical thresholds shown in the tables (for the given values of *n* and λ) to obtain approximate P-values, but it would probably be better to search for a simple bias-correcting function $b_{n,}$ such that $Z_n= n \times [f(\frac{s^2}{\bar{x}})-f(1)- b_n]$ has approximately mean zero and standard deviation one, with the approximation improving both as λ and n increase.

The Z-statistic defined above seems to work quite well in the upper tail, but it is not really needed in that range, as there are many statistics (notably the Likelihood Deviance or the Pearson Deviance statistics) that are quite effective at detecting over-dispersion. Similarly, as the sample size, *n*, becomes large, many statistics will yield approximately normal distributions (under the null hypothesis of i.i.d. Poisson observations). However, we are really more concerned with cases which don't fall in either of the 'good' categories; we want to detect *under-dispersion* when the sample size

is *small*. As the results in Tables A2-A7 show, this is hard to do for small *n*, such as 2, or 3, especially when λ is small. Since a Poisson random variable can assume only integer values, it is hard to determine whether a sample such as {1,1,1} with n=3 is a chance occurrence from a Poisson with true λ near 1, or 3 highly dependent realizations of a single Poisson event. Thus, one observes that the statistic's lower bound ($Z = -n$) is often achieved by chance from independent Poisson samples when n=2 or n=3, but this becomes much rarer as *n* increases, and, even for n=3 or n=4 is not common unless the true λ is fairly small.

Thus, we can use the created Z-statistic (but preferably a bias-corrected version) to examine ostensibly Poisson distributed data for dependence. That is, for a given sample of *n* counts from a homoeolog of a wheat gene, we obtain the copy counts $\{X_1, X_2, \ldots, X_n\}$ at the *n* key signature sites, and calculate the sample mean ($\bar{X}$) and standard deviation (s) of these counts. We use these values to calculate the statistic $Z = n\left[\sqrt{\frac{s}{\sqrt{\bar{x}}}} - 1\right]$ (or a bias-corrected version) and check the statistic's percentile in the distribution. Roughly speaking, if the calculated statistic is in the (-1.96, 1.96) range, we suppose that the data for this homoeolog of the gene can be treated as if they were independent realizations from a constant-mean Poisson process. If the calculated statistic is in the (-n, -1.96) range, we treat the data for that homoeolog as if it is under-dispersed, while if the calculated statistic is in the (1.96, ∞) range, we treat the data for that homoeolog as if it is over-dispersed. While a separate Z-statistic can be calculated for each of the 3 homoeologs within the same gene, it wouldn't make too much sense for the dependence pattern to depend on homoeolog, so the average of the three Z-statistics for a gene is probably the best overall measure of dependence. If approximate independence can be assumed, we

proceed with the standard Poisson regression analyses. Similarly, if over-dispersion appears to have occurred, we can adjust for it in standard ways.

The most challenging aspect, and what has motivated this research, is how to proceed when the Z-statistic indicates under-dispersion. Since under-dispersion in this context is likely caused by taking measurements at signature sequence sites which are 'too close' to one another, one possibility would be to delete some of the interior counts (assuming *n* is large enough to do this) in the hope that the remaining counts would behave more 'independently'. Another possibility would be to use all the counts, but to weight them in some way such that highly dependent values don't have adverse effects on resulting inferences. The statistical pros and cons of each of these approaches are discussed later in this thesis.

3.2 ESTIMATING CORRELATION COEFFICIENT AMONG DATA

Another useful tool in testing for independence of counts is to estimate the correlation coefficient *(r*) between counts which are separated by a certain distance (in base pairs) . The correlation coefficient, *r*, for a pair of random variables $\{Y_1, Y_2\}$ is a quantity related to the covariance and is defined as:

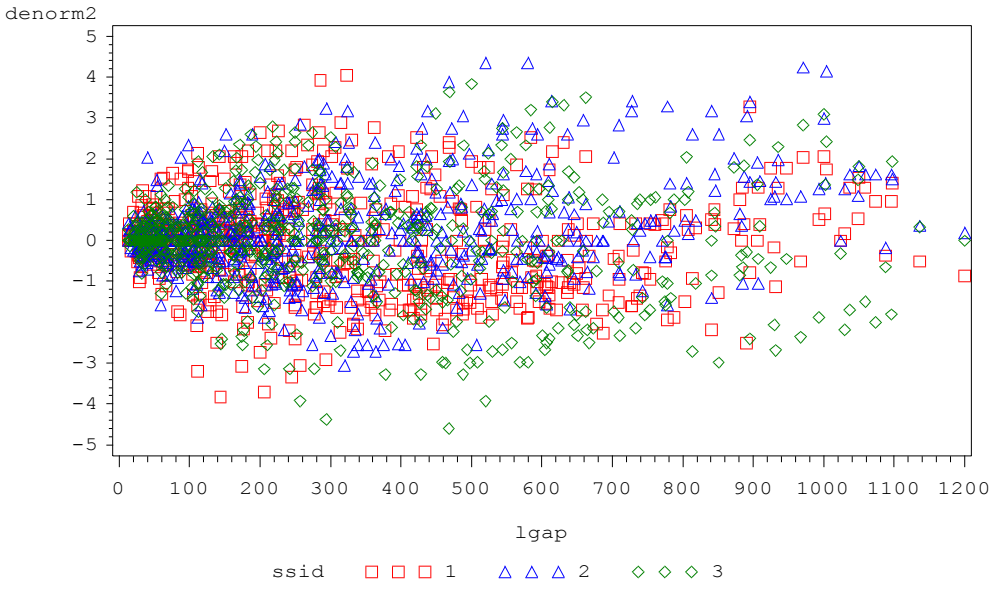
$$r = \frac{\mathrm{Cov}(Y_1,\ Y_2)}{\sigma_1 \sigma_2}$$

As we know, $0 \leq |r| \leq 1$. A value of $r = 0$ implies zero covariance and no correlation. And with the $|r|$ increasing, the correlation is more significant. If $|r| = 1$, this implies a perfect correlation, where Y_1 and Y_2 can be perfectly predicted from one another. In the context

of this problem, we can examine all *n-choose-2* pairs of counts within the same homoeolog of the same gene in the wheat genome to try to measure '*r*' for that pair.

Thus, we will approximate the correlations between the differences of any pair of expression counts and their corresponding positions for our 61-gene data set. To do so, we first calculated the difference between any pair of expression counts (Count1 and Count2) and the difference (in base-pair units) between their corresponding positions, and named the difference of counts as '**Diff**', and the difference in positions as '**Lgap**'. Then, we let **Denorm**= $\frac{\text{Diff}}{\sqrt{\text{Count}_1+\text{Count}_2}}$ and plotted **Denorm** against **Lgap** (Figure 3). The variable **Denorm**, assuming the two counts arise from independent Poisson distributions with common intensity (λ), has a mean of zero and standard deviation of near one, and is closely related to Kruskal's G-statistic used in measuring association in 2×2 tables. Asymptotically (as λ becomes large) , under the above assumptions, 'Denorm' will follow a Standard Normal distribution, but that fact is of marginal relevance here.

Figure 3. **Denorm** vs. **Lgap**



The plot in Figure 3 displays all N=2007 pairs of **denorm** statistics that could be calculated from the data set. There are 61 genes, each with 3 homeologs, and each of these has *n-choose-2* pairs of counts that could be examined. The median *n* for the 61 genes is n=5, which would yield 1830 pairs, but the exact total is slightly greater, so that N=2007. In the figure, '**denorm**' is plotted on the Y-axis, while '**Lgap**' (the distance in base-pairs between the sites at which the two counts were made) is plotted on the X-axis. (The three homeologs are shown via the plotting symbols (1=A, 2=B, 3=D), but no interesting differences between the homoeologs is observed there.) If the count pairs were independent realizations from a common Poisson distribution, one would expect the **denorm** statistics to be distributed somewhat similarly to a N(0,1) random variable. From this plot, we can see that this is not the case at all. For small values of **Lgap**, the **denorm** statistics are much too tightly clustered around zero. As **Lgap** increases, the distribution appears to stabilize, but it appears to have a standard deviation much greater than 1. This threshold distance at which stabilization appears to occur for **Lgap** is in the 200-300 base-pairs region, but we need to measure it more accurately.

To obtain a more accurate estimate of the threshold distance, we divided **Lgap** into 20 levels (from level *a* [0-24 bp gap] to level *t* [> 1000 bp]) and calculated the number of observations (N), 25% percentile, 50% percentile, 75% percentile, mean and standard deviation (SD) of **Denorm** for each of the 20 levels . This information is shown in Table 1 below. Ideally, the SD should rise from near zero when the gap is small and then level off at 1.00 once the gap-distance such that independence can be claimed is achieved. While there does appear to be a leveling off after **Lgap**=225 base-pairs, the SD

for larger gaps seems to be much greater than 1.0, indicating that over-dispersion is confounded with under-dispersion for these data

Thus, in Table 2, we divided the data into 10 levels (from level **a** to level **j**) and re-calculated everything as in Table 1. Moreover, we calculated

$$M = \hat{\sigma} = \sqrt{\frac{\sum_{i=j}^{t} N_i \sigma_i^2}{\sum_{i=j}^{t} N_i}}$$

from which we obtained M ≈ 1.438 as the average over-dispersion factor after the **Lgap** has increased enough for independence to be viable. This agrees well with the observed SD of level *j*, 1.451, reinforcing the idea that gap=225 base pairs is the minimum distance at which independence is viable. This also suggest that we need to deflate the **denorm** statistic (since over-dispersion as well as under-dispersion are present), so that SD=1.0 is achieved at a 225 base-pair gap. Thus, we deflate the calculated SDs by using the deflated standard deviation, **def SD** $= \frac{SD}{M}$ = SD/1.438, as shown in the last column of Table 2.

Table 1 Summary of **Denorm** Based on 20 Levels

level	Lgap	N	25%-ile	50%-ile	75%-ile	Mean	SD
a	0-24	42	0	0	0.243	0.113	0.223
b	25-49	168	-0.192	0	0.268	0.055	0.422
c	50-74	138	-0.277	0	0.302	0.036	0.551
d	75-99	102	-0.333	0	0.343	0.070	0.701
e	100-124	132	-0.483	0	0.474	-0.038	0.817

f	125-149	93	-0.707	0	0.480	-0.127	1.031
g	150-174	93	-0.277	0.229	0.775	0.260	0.952
h	175-199	72	-0.895	0	0.556	-0.040	1.156
i	200-224	96	-0.832	-0.085	0.631	-0.081	1.251
j	225-249	69	-0.632	0.277	0.928	0.138	1.171
k	250-274	60	-0.971	0	1.304	0.137	1.525
l	275-300	87	-0.775	0.169	1.134	0.131	1.404
m	301-400	183	-0.943	0.000	0.775	-0.060	1.376
n	401-500	189	-1.213	-0.277	0.775	-0.134	1.503
o	501-600	138	-1	-0.209	1.029	0.029	1.527
p	601-700	111	-1.177	-0.378	0.707	-0.196	1.447
q	701-800	72	-1.283	-0.340	0.5	-0.187	1.285
r	801-900	54	-0.781	0.449	1.441	0.428	1.567
s	901-1000	36	-0.447	1.093	1.843	0.756	1.617
t	>1000	42	-0.180	0.962	1.540	0.624	1.320
TOTAL		2007					

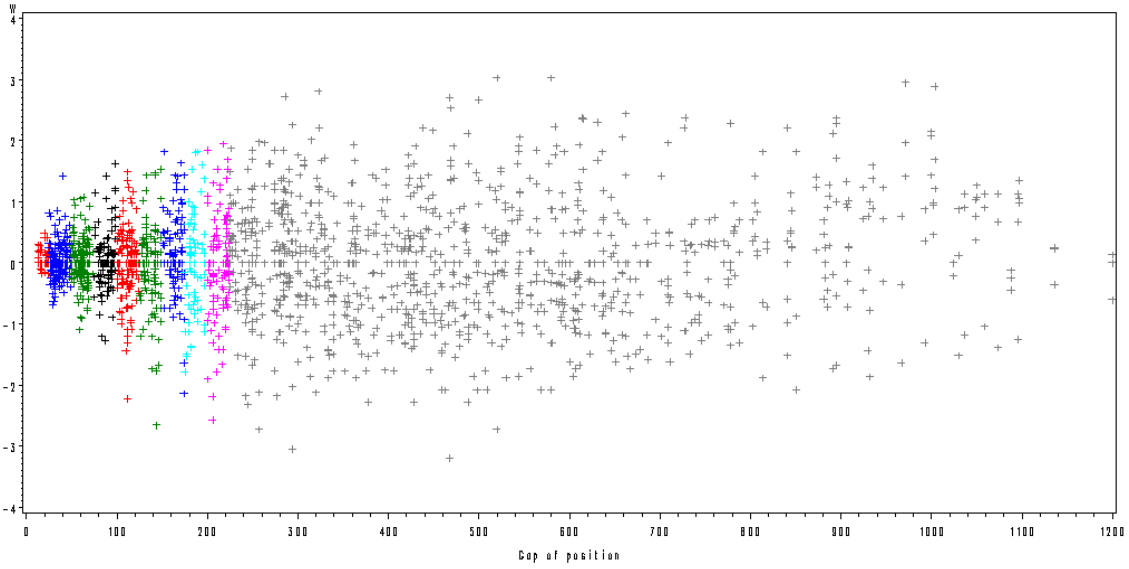
Table 2 Summary of **Denorm** Based on 10 Levels

level	lgap	N	25%-ile	50%-ile	75%-ile	Mean	SD	def SD
a	0-24	42	0	0	0.243	0.113	0.223	0.155

b	25-49	168	-0.192	0	0.268	0.055	0.422	0.294
c	50-74	138	-0.277	0	0.302	0.036	0.551	0.383
d	75-99	102	-0.333	0	0.343	0.070	0.701	0.488
e	100-124	132	-0.483	0	0.474	-0.038	0.817	0.568
f	125-149	93	-0.707	0	0.480	-0.127	1.031	0.717
g	150-174	93	-0.277	0.229	0.775	0.260	0.952	0.662
h	175-199	72	-0.895	0	0.556	-0.040	1.156	0.804
i	200-224	96	-0.832	-0.085	0.631	-0.081	1.251	0.870
j	>224	1041	-1	0	1	0.037	1.451	1.009

Thus, for all the **Denorm** values, we renormalized by calculating the statistic **W=Denorm/M** and plotted the **W** against **Lgap**. This plot is shown in Figure 4.

Figure 4 **W** vs. **Lgap**



From Figure 4, we can find there are some outliers whose W values are over 2.5. We should find very few observations whose W value are larger than 2.5, so we delete these observations (only for the purpose of finding a correlation function, not for real data analysis). As a result, we identified the 15 outliers which are listed in Table 3.

After deleting theses outliers from the data, we again divided **Lgap** into 20 levels (from level *a* to level *t*) and 10 levels (from level *a* to *j*) respectively, and calculated the number of observations (N), 25% percentile, 50% percentile, 75% percentile, mean and standard deviation (SD) of Denorm for each level (Table 4, Table 5). Moreover, we re-calculated **M** (≈ 1.373), **def SD**, and **W** to get the final results shown in Table 5. Then, we re-plotted **W** against **Lgap** (Figure 4).

Table 3 List of Outliers

geneid	ssid	Lgap	diff	denorm	Level	w
11	1	144	-55	-3.823	F	-2.659
11	1	206	-53	-3.702	I	-2.575
17	1	286	20	3.922	J	2.728
17	1	323	21	4.041	J	2.811
17	2	468	23	3.888	J	2.704
17	2	520	25	4.352	J	3.027
17	2	580	25	4.352	J	3.027
17	3	257	-20	-3.922	J	-2.728

17	3	294	-24	-4.382	J	-3.048
17	3	468	-26	-4.596	J	-3.197
17	3	520	-20	-3.922	J	-2.728
33	2	971	36	4.243	J	2.951
33	2	1004	35	4.154	J	2.889
57	3	469	25	3.647	J	2.536
57	3	500	26	3.833	J	2.666

Table 4 Summary of **Denorm** Based on 20 Levels after Deleting the Outliers

level	lgap	N	25%-ile	50%-ile	75%-ile	Mean	SD
a	0-24	42	0	0	0.243	0.113	0.223
b	25-49	168	-0.192	0	0.268	0.055	0.422
c	50-74	138	-0.277	0	0.302	0.036	0.551
d	75-99	102	-0.333	0	0.343	0.070	0.701
e	100-124	132	-0.483	0	0.474	-0.038	0.817
f	125-149	93	-0.707	0	0.480	-0.127	1.031
g	150-174	93	-0.277	0.229	0.775	0.260	0.952
h	175-199	72	-0.895	0	0.556	-0.040	1.156
i	200-224	96	-0.832	-0.085	0.631	-0.081	1.251
j	225-249	69	-0.632	0.277	0.928	0.138	1.171
k	250-274	60	-0.971	0	1.304	0.137	1.525

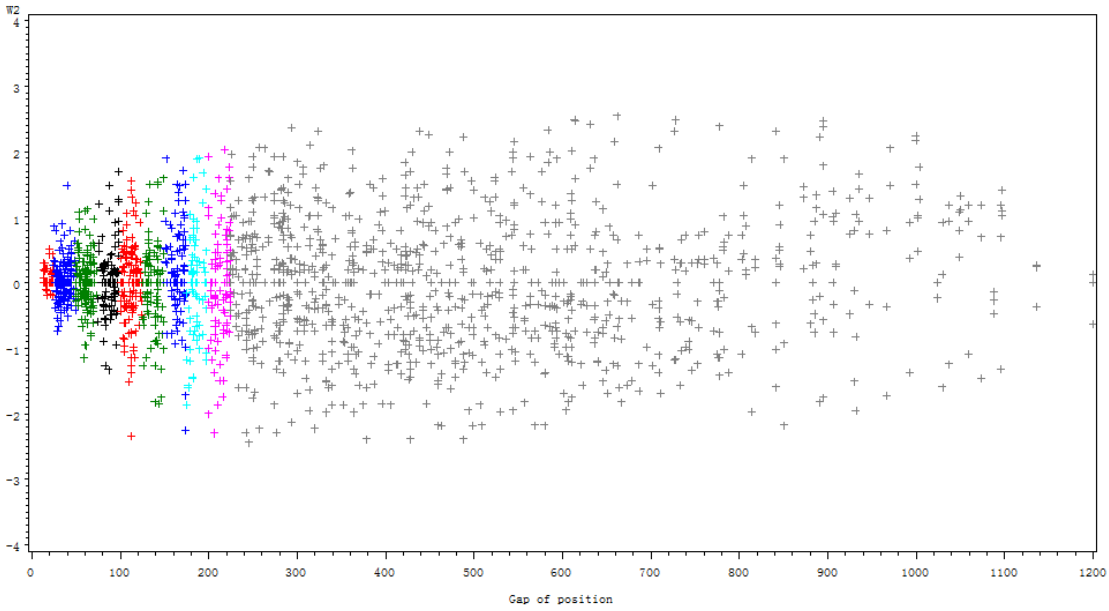
l	275-300	87	-0.775	0.169	1.134	0.131	1.404
m	301-400	183	-0.943	0.000	0.775	-0.060	1.376
n	401-500	189	-1.213	-0.277	0.775	-0.134	1.503
o	501-600	138	-1	-0.209	1.029	0.029	1.527
p	601-700	111	-1.177	-0.378	0.707	-0.196	1.447
q	701-800	72	-1.283	-0.340	0.5	-0.187	1.285
r	801-900	54	-0.781	0.449	1.441	0.428	1.567
s	901-1000	36	-0.447	1.093	1.843	0.756	1.617
t	>1000	42	-0.180	0.962	1.540	0.624	1.320

Table 5 Summary of **Denorm** Based on 10 Levels after Deleting the Outliers

level	lgap	N	25%-ile	50%-ile	75%-ile	Mean	SD	def SD
a	0-24	42	0	0	0.243	0.113	0.223	0.162
b	25-49	168	-0.192	0	0.268	0.055	0.422	0.308
c	50-74	138	-0.277	0	0.302	0.036	0.551	0.401
d	75-99	102	-0.333	0	0.343	0.070	0.701	0.511
e	100-124	132	-0.483	0	0.474	-0.038	0.817	0.595
f	125-149	92	-0.681	0	0.487	-0.087	0.961	0.700
g	150-174	93	-0.277	0.229	0.775	0.260	0.952	0.693

h	175-199	72	-0.895	0	0.556	-0.040	1.156	0.842
i	200-224	95	-0.832	0	0.632	-0.043	1.200	0.874
j	>224	1028	-1	0	0.982	0.018	1.386	1.009

Figure 5 **W** vs. **Lgap** after Deleting the Outliers



According to results of Table 5 and Figure 5, we can see that from level '**a**' to level '**i**' the correlation of the data increases with the distance between positions increasing. However, we estimate that when distances are larger than 225 bp, the distance will not affect the **Difference** of expression counts. That is, when the distances between two signature sites are more than 225 bp apart, the expression counts appear to behave as if they are independent Poisson realizations (with an over-dispersion factor of about 1.37). This threshold distance leads us to the relatively simple piecewise linear estimate of the correlation coefficient (r) between (x_1, x_2), the counts at any two points:

$$r(x_1, x_2) = \begin{cases} 1 - \left(\frac{\text{Lgap}}{100}\right) \text{ if } 0 < \text{Lgap} < 24 \\ 0.85 - \left(\frac{\text{Lgap}}{250}\right) \text{ if } 25 < \text{Lgap} < 149 \\ 0.75 - \left(\frac{\text{Lgap}}{300}\right) \text{ if } 150 < \text{Lgap} < 224 \\ 0 \qquad\qquad \text{if } \text{Lgap} \geq 225 \end{cases}$$

This reasonably closely approximates the relationship between *r* and **Lgap** shown in Figure 5. Note that this simple piecewise linear approximation for *r* may be useful for an individual pair (x_1, x_2) of counts, but might not be consistent for a triplet (x_1, x_2, x_3) of counts, especially as **Lgap** becomes large. For example, if (x_1, x_2, x_3) were arranged sequentially such that the distance between the first two and the distance between the last two was about 112 bp, then the correlation between the first two would be about .40, as would that between the second and third, but the distance between the first and third would lead to a correlation of approximately zero, and not .16, which would be expected on theoretical grounds.

3.3 DISCARDING SOME NEAR-BY DATA

From the previous section, we have estimated that the minimum threshold distance necessary to assume independence between counts is 225 bp. Thus, one correct (but rather crude) solution to our problem is to discard those data points which are separated by a distance less than 225 bp. For example, for the gene 1, there are 7 positions originally (116, 144, 180, 212, 235, 302, 340). Thus, using the rule above, if we delete the data which are in the 5 middle positions: (144, 180, 212, 235, 302) and use only the

counts at the first (116) and last (340) positions, we will have approximately independent observations. This method suffers from two major drawbacks. First, for 25 of the 61 genes, there are not any pairs which are measured at distances more than 225 bp from one another, so no tests at all for mean equality of counts of {A, B, D} could be obtained for these genes. Secondly, for most genes, there are many non-unique ways to discard genes so as to remove those that are not separated enough, so depending upon which sites' counts are removed, different conclusions could be drawn.

After using (a particular application of) this method to reduce the original dataset (analyzed by naïve Poisson regression in Table A1) to observations which were separated by at least 225 base pairs, we again performed a Poisson Regression (with over-dispersion factors as necessary) to analyze the remaining data. In general, we found that the deviance statistic for over-dispersion became larger than before and significant differences were harder to find. This is a conservative method, so any significant results found by this method are probably 'real'. Nonetheless, the increase in untestable genes and the increase in the deviance statistics implies that we are losing much valuable information under this plan. Hence, directly discarding dependent data (i.e. counts obtained from signature sequences within 225 bp of one another) is not a recommended method for analyzing such data.

3.4 USING ALL DATA VIA WEIGHTING

Because discarding some dependent data could result in serious loss of information, we will now concentrate on procedures which use all the data. Thus, in this section, we search for an optimal weighting of the correlated counts.

First, to motivate our method, we consider a simple case where n=3. In this case, there are three positions (a, b, and c) such that a < b < c. At each position, there is a corresponding expression count (X_1 or X_2 or X_3). The three variables X_1, X_2, and X_3 are assumed to follow the Poisson distribution with mean λ. In this case, if X_1, X_2 and X_3 are independent, we can easily find the estimated mean ($\hat{\lambda}$) is equal to $(X_1+X_2+X_3)/3$ and Var ($\hat{\lambda}$) = $\hat{\lambda}$ /3. However, we want to find a general weighting which will give a minimum variance unbiased estimate of λ, even if the data are correlated. Without loss of generality, we can assume that the weights are non-negative and sum to 1. Also, we assume that we have a model which accurately estimates correlations from gap distance, so that (b-a), (c-a), and (c-b) determine r_{12}, r_{13}, and, r_{23}, respectively. So, we have:

$$\hat{\lambda} = w_1X_1 + w_2X_2 + w_3X_3 = \sum_{i=1}^{3} w_i X_i \ .$$

$$\text{Var}(\hat{\lambda}) = \text{Var}\left(\sum_{i=1}^{3} w_i X_i\right)$$

$$= \text{Var}(w_1X_1) + \text{Var}(w_2X_2) + \text{Var}(w_3X_3) + 2\,\text{Cov}(w_1X_1, w_2X_2) + 2\,\text{Cov}(w_1X_1, w_3X_3) + 2\,\text{Cov}(w_2X_2, w_3X_3)$$

$$= w_1^2\,\text{Var}(X_1) + w_2^2\,\text{Var}(X_2) + w_3^2\,\text{Var}(X_3) + 2w_1 \times w_2 \times \text{Cov}(X_1, X_2) + 2w_1 \times w_3 \times \text{Cov}(X_1, X_3) + 2w_2 \times w_3 \times \text{Cov}(X_2, X_3)$$

$$= w_1^2\,\text{Var}(X_1) + w_2^2\,\text{Var}(X_2) + w_3^2\,\text{Var}(X_3) + 2w_1 \times w_2 \times r_{12} \times \sigma_{x1} \times \sigma_{x2} + 2w_1 \times w_3 \times r_{13} \times \sigma_{x1} \times \sigma_{x3} + 2w_2 \times w_3 \times r_{23} \times \sigma_{x2} \times \sigma_{x3}$$

$$= w_1^2\,\lambda + w_2^2\,\lambda + w_3^2\,\lambda + 2\,\lambda\, w_1 \times w_2 \times r_{12} + 2\,\lambda\, w_1 \times (1-w_1-w_2) \times r_{13} + 2\,\lambda\, w_2 \times (1-w_1-w_2) \times r_{23}$$

$$= \lambda\,[w_1^{\,2} + w_2^{\,2} + w_3^{\,2} + 2\,w_1 \times w_2 \times r_{12} + 2\,w_1 \times (1\text{-}w_1\text{-}\,w_2) \times$$

$$r_{13} + 2\,w_2 \times (1\text{-}w_1\text{-}\,w_{2)} \times r_{23}]$$

Now, our purpose is to find the values of w_1, w_2 and w_3 which can minimize the Var ($\hat{\lambda}$). So, we let L= Var ($\hat{\lambda}$) and took the derivative of Var ($\hat{\lambda}$) by w_1 and w_2 respectively. This yielded:

$$\frac{\partial L}{\partial w_1} = \lambda[2w_1 - 2(1 - w_1 - w_2) + 2w_2r_{12} + 2r_{13} - 4w_1r_{13} - 2w_2r_{13} - 2w_2r_{23}$$

$$\frac{\partial L}{\partial w_2} = \lambda[2w_2 - 2(1 - w_1 - w_2) + 2w_1r_{12} + 2r_{23} - 4w_2r_{23} - 2w_1r_{13} - 2w_1r_{23}$$

Setting $\frac{\partial L}{\partial w_1} = 0$ and $\frac{\partial L}{\partial w_2} = 0$ one obtains the equation pair below.

$$\begin{bmatrix} 1+1-2r_{13} & 1+r_{12}-r_{13}-r_{23} \\ 1+r_{12}-r_{13}-r_{23} & 1+1-2r_{13} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1-r_{13} \\ 1-r_{23} \end{bmatrix}$$

which reduces to:

$$\begin{bmatrix} 2(1-r_{13}) & 1+r_{12}-r_{13}-r_{23} \\ 1+r_{12}-r_{13}-r_{23} & 2(1-r_{13}) \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1-r_{13} \\ 1-r_{23} \end{bmatrix}$$

This equation does lead to the expected estimators in common situations. For example, if all correlations are zero (all counts independent of one another), then all weights are 1/3. If the distribution is symmetric, so that $r_{12} = r_{23}$, then $w_1 = w_3$, and the

weights on the outer boundary points increase from 1/3 each when the points are well separated to ½ each when the $r_{12} = r_{23}$=1.0. (This is true if one assumes an autoregressive correlation structure, so that $r_{13} = r_{12}^2$. This is not, strictly speaking, the correlation structure which is imposed by the linear function shown in Figure 4, however.)

In general, we need to find matrix equations similar to the above for the general case. We assume there are *k* variables in this case, and find that the weights must satisfy.

$$A \begin{bmatrix} w_1 \\ \vdots \\ w_{k-1} \end{bmatrix} = \begin{bmatrix} 1 - r_{1,k} \\ \vdots \\ 1 - r_{k-1,k} \end{bmatrix}$$

Here, A is a matrix, with entries $A_{i,i}$= 2(1- $r_{i,k}$), $A_{i,j}$ = (1+$r_{i,j}$-$r_{i,k}$-$r_{j,k}$).

We use several special cases to verify our formula about weighting. In all 3 cases, n=4, and we must consistently specify the 6 pair-wise distances. For example, in the first case, demonstrated in Table 6, the observations on geneid 2000 are taken at 4 points {0, 115, 230, 345} which are equidistant from left and right neighbors, while the second case, shown in Table 8, has the 4 points at {0, 200, 300, 500}, and the third, shown in Table 10, has them at {0, 500, 600, 1100}. The 'optimal weighting' (to minimize the variance of the estimator (Var ($\hat{\lambda}$)) for these situations, assuming the true correlations as a function of lgap are as given in Section 3.2, are shown in Tables 6, 9, and 11, respectively.

Table 6 Case 1

Pair(i,j)	Pos(i)	Pos(j)	Lgap	r

(1,2)	0	115	115	0.446
(2,3)	115	230	115	0.446
(1,3)	0	230	230	0
(3,4)	230	345	115	0.446
(2,4)	115	345	230	0
(1,4)	0	345	345	0

Table 7 Weighting Results of Case 1

Position	Weight
1	0.3106
2	0.1894
3	0.1894
4	0.3106

Table 8 Case 2

Pair(i,j)	Pos(i)	Pos(j)	Lgap	r
(1,2)	0	200	200	0.246
(2,3)	200	300	100	0.496
(1,3)	0	300	300	0

(3,4)	300	500	200	0.246
(2,4)	200	500	300	0
(1,4)	0	500	500	0

Table 9 Weighting Results of Case 2

Position	Weight
1	0.2993
2	0.2007
3	0.2007
4	0.2993

Table 10 Case 3

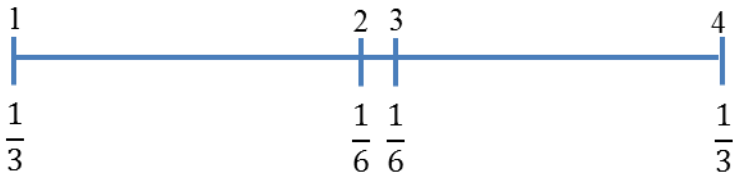
Pair(i,j)	Pos(i)	Pos(j)	Lgap	r
(1,2)	0	500	500	0
(2,3)	500	600	100	0.496
(1,3)	0	600	600	0
(3,4)	600	1100	500	0
(2,4)	500	1100	600	0
(1,4)	0	1100	1100	0

Table 11 Weighting Results of Case 3

Position	Weight
1	0.2959
2	0.2041
3	0.2041
4	0.2959

According to the results for the special cases above when n=4, we could conclude that if the two inner positions are very close relative to the outer positions (i.e. distance between the outer two positions is much greater than 225 bp, but inner positions are very close), we can approximately weight the data as shown in Figure 6.

Figure 6 Idealized Weights by Position



While the schematic above illustrates that a reasonable weighting will be put on the count values by the Var ($\hat{\lambda}$) minimization calculations, it also illustrates two flaws with this method of approach. First, it assumes that correlations can be accurately estimated from distance by some function similar to that which we estimated in section 3.2. As noted at the end of that section, this becomes problematic when there are adjacent pairs which are intermediately correlated with each other. A more serious problem is that finding the

optimal weights to minimize Var ($\hat{\lambda}$) does not really solve our major problem. It will give a slightly more precise estimator of $\hat{\lambda}$ than will the simple average, but in most cases, the overall point estimates will be very similar. What is more important than the optimal weighting is the equivalent sample size. For example, in the schematic example shown in Figure 7, it is clear that the two middle counts (which are probably identical) should be averaged together and counted as one, so that the n=4 in this case would have an effective sample size of n=3.

If one could calculate these effective sample sizes from the optimal weightings, then the hypothesis testing problem would be easily solved, since one could adjust the naïve tests (using all data as if it were independent) both for mean estimates for the three homoeologs in a gene (by choosing the weights which minimize Var ($\hat{\lambda}$)), but, more importantly, for effective sample size. This latter factor is much more important in making correct inferences, as the naïve method has much smaller standard errors than are correct. Fortunately, this is very easy, as effective sample size (for the optimally weighted procedure) is simply given by n_eff = ($\hat{\lambda}$)/ Var ($\hat{\lambda}$). Table 12 below displays how the effective sample sizes for the three estimators (naïve, conservative, and weighted) behave for the 61 wheat geneome genes of this study. As expected, the effective sample size of the weighted method falls between the other two, although frequently much closer to conservative than to naïve, because of the high degree of dependence in this data set. In general, the procedure could be quite useful for using all of the data in an automated way.

Table 12 List of Effective Sample Sizes by Three Methods for 61 Genes

geneid	deviance	naïve	Conservative	weighted
1	0.0595	7	1	2.00
2	0.2801	4	1	1.63
3	1.1418	5	3	3.25
4	0.8252	9	5	6.05
5	1.6368	3	2	2.13
6	0.0115	2	1	1.17
7	0.3135	3	1	1.19
8	0.0145	2	1	1.13
9	0.7733	5	2	2.15
10	0.6558	12	4	4.76
11	3.5811	5	2	2.88
12	1.88	5	2	2.94
13	1.4158	2	2	3.08
14	0.0969	2	1	1.86
15	0	2	1	1.23
16	6.0397	3	2	2.40
17	6.3875	6	3	3.49
18	2.3729	3	2	2.11
19	1.1781	4	2	2.60
20	4.3829	2	2	2.00
21	0.0765	4	1	1.66
22	1.3996	3	2	2.56
23	1.3334	6	2	2.52
24	2.0056	10	4	5.08
25	0.1194	2	1	1.66
26	1.4808	5	2	2.94
27	1.9557	3	2	2.00
28	0.1337	3	1	1.51
29	1.9692	14	4	5.38
30	1.7472	4	2	2.49
31	4.393	2	2	2.00
32	2.7021	3	3	3.00
33	4.3779	4	2	2.54
34	0.3046	3	1	1.86
35	0.1565	2	1	1.25
36	1.6177	2	2	2.00
37	0.0115	2	1	1.32
38	1.2504	2	1	1.40
39	0.6908	2	1	1.53
40	0.4812	5	3	3.10
41	1.0504	3	2	2.31
42	0.0372	2	1	1.12
43	0.3987	6	2	2.76
44	1.6129	8	3	4.44

45	1.6802	5	2	2.65
46	0.0292	2	1	1.19
47	0.5469	9	3	4.09
48	0.2974	2	1	1.29
49	0.1214	2	1	1.56
50	0.8844	4	2	2.51
51	0.7573	5	2	2.92
52	0.3351	6	3	3.44
53	1.1689	2	1	1.97
54	1.5758	6	2	2.70
55	0.0956	2	1	1.18
56	0.0898	4	1	1.48
57	2.3151	12	4	5.90
58	0.2014	2	1	1.48
59	1.4419	5	2	3.10
60	0.1262	3	1	1.31
61	0.8986	2	1	1.26

CHAPTER 4

CONCLUSION

In this thesis, we compare several methods to handle correlated Poisson count data. First, we attempted a naive method to analyze data, that is, we directly used standard Poisson regression (correcting for over-dispersion) to analyze the data, and to test the null hypothesis (of equality of intensity parameter over the A, B, and D homoeologs) separately for each gene in the wheat data set. Examining the results, we find that there are several problems: (1) over-dispersion; (2) dependence within a gene; (3) dependence across homoeologs. The over-dispersion difficulty is easily handled by standard methods, but under-dispersion caused by dependent counts within the same homoeolog is more challenging.

As a first approach, we created a 'Z-Statistic' based on the ratio of the sample variance to the sample mean in order to test whether the counts are independent. While this method could detect severe dependence, it lacked power, especially for small values of n, the number of signature sites within a gene. So, next, we tried to determine a threshold separation distance such that counts within this distance would not both be used. We found that using 225 base pairs as the minimum gap-distance would yield counts within a homoeolog which were approximately statistically independent of one another, so that standard Poisson regression results would hold. However, when these results were applied to the data set, it caused 25 of the 61 genes to be untestable and caused serious loss of power in others.

Next, we estimated the correlation (r) between counts of pairs of observations as a function of the gap (in bp) between the pairs. We did this indirectly, by comparing the standard deviation of the **Denorm**= $\frac{\text{Diff}}{\sqrt{\text{Count}_1+\text{Count}_2}}$ statistic to 1.00, as discussed in Section 3.2. This allows us to use all data points, rather than disallowing those within 225 bp of one another. In Section 3.4, we show that the set of weights $\{w_i\}$ which minimizes Var $(\hat{\lambda})$ = Var $(\sum_{i=1}^{n} w_i X_i)$ is completely determined by the correlation structure, and that the effective sample size from using the weighted estimator is given by $(\hat{\lambda})$/ Var $(\hat{\lambda})$. Using the estimators of (r) derived in Section 3.2 as if they are exact in the effective sample size estimator, we find, not surprisingly, that the effective sample size under the optimal weighting always lies between that of of the naïve estimator (n) and the conservative estimator, although in many cases not much different than that given by the conservative estimator. This weighted method can be used for any gene data set of the type discussed in the thesis, although it will be most useful in cases that are mid-way between total independence and high dependence due to overlap of fragments.

REFERENCES

[1] Bottley, A., and R. M. D. Koebner, 2008 Variation for homoeologous gene silencing in hexaploid wheat. *Plant Journal* **56**: 297-302.

[2] Bottley, A., G. M. Xia, and R. M. D. Koebner, 2006 Homoeologous gene silencing in hexaploid wheat. *Plant Journal* **47**: 897-906.

[3] Choulet, F., T. Wicker, C. Rustenholz, E. Paux, J. Salse *et al.* 2010 Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* **22**: 1686-1701.

[4] Crismani, W., U. Baumann, T. Sutton, N. Shirley, T. Webster *et al.* 2006 Microarray expression analysis of meiosis and microsporogenesis in hexaploid bread wheat. *BMC Genomics* **7**: 267-273.

[5] Desmond, O. J., J. M. Manners, P. M. Schenk, D. J. Maclean, and K. Kazan, 2008 Gene expression analysis of the wheat response to infection by *Fusarium pseudograminearum*. *Physiological and Molecular Plant Pathology* **73**: 40-47.

[6] Devos K.M. Grass genome organization and evolution. Current Opinion in Plant Biology . 2010.

[7] Huang, S., A. Sirikhachornkit, X. Su, J. Faris, B. S. Gill *et al.* 2002 Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the *Triticum/Aegilops* complex and the evolutionary history of polyploid wheat. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 8133-8138.

[8] Massa A.N., Wanjugi H., Deal K.R., O'Brien K.O., You F.M., Maiti R., Chan A., Yu Y.Q., Luo M.C., Anderson O.A., Rabinowicz P.D., Dvorak J. & Devos K.M. Evolution of gene space in grass genomes suggests a biological role for variation in genome size. Molecular Biology and Evolution . 2011.

[9] McIntosh, R. A., Y. Yamazaki, K. M. Devos, J. Dubcovsky, W. J. Rogers *et al.* 2003 Catalogue of gene symbols for wheat, pp. 1-34 in *Proc 10th Int Wheat Genet Symp, Volume 4*, S.I.M.I., Rome.

[10] Mochida, K., Y. Yamazaki, and Y. Ogihara, 2004 Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Molecular Genetics and Genomics* **270**: 371-377.

[11] Nesbitt, M., and D. Samuel, 1996 From staple crop to extinction? The archaeology and history of hulled wheats, pp. 41-100 in *Hulled Wheats. Promoting the conservation and use of underutilized and neglected crops. 4. Proc. of the First International Workshop on Hulled Wheats*, edited by S. Padulosi, K. Hammer, and J. Heller. Castelvecchio Pascoli, Tuscany, Italy.

[12] Pumphrey, M., J. F. Bai, D. Laudencia-Chingcuanco, O. Anderson, and B. S. Gill, 2009 Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. *Genetics* **181**: 1147-1157.

[13] Schreiber, A. W., T. Sutton, R. A. Caldo, E. Kalashyan, B. Lovell *et al.* 2009 Comparative transcriptomics in the Triticeae. *BMC Genomics* **10**: 285-301.

[14] Sears, E. R., 1954 The aneuploids of common wheat. *Missouri Agricultural Experiment Station Research Bulletin* **572**: 1-59.

[15] Winfield, M. O., C. G. Lu, I. D. Wilson, J. A. Coghill, and K. J. Edwards, 2009 Cold- and light-induced changes in the transcriptome of wheat leading to phase transition from vegetative to reproductive growth. *Bmc Plant Biology* **9**: 55-68.

APPENDIX

Table A1 – Results of Naïve Poisson Regression Analyses

GeneID	#SSNs	Deviance	P A-B	P A-D	P B-D
1	7	0.0595	0.0258	0.108	0.0002
2	4	0.2801	0.7181	0.0246	0.0097
3	5	1.1418	0.4099	0.6118	0.7505
4	9	0.8252	0.1034	0.9045	0.1312
5	3	1.6368	<.0001	<.0001	0.1193
6	2	0.0115	0.6702	0.0101	0.0037
7	3	0.3135	0.0001	<.0001	0.4551
8	2	0.0145	0.0677	0.4513	0.2649
9	5	0.7733	0.0154	<.0001	<.0001
10	12	0.6558	0.8912	0.0001	0.0002
11	5	3.5811	<.0001	0.1071	<.0001
12	5	1.88	0.0009	<.0001	0.4212
13	2	1.4158	0.7596	0.067	0.0328
14	2	0.0969	0.6952	0.0136	0.0052
15	2	0	0.1618	0.0285	0.3744
16	3	6.0397	0.8273	0.0105	0.0187
17	6	6.3875	0.2709	0.0015	<.0001
18	3	2.3729	0.6233	<.0001	<.0001
19	4	1.1781	<.0001	0.0066	<.0001
20	2	4.3829	0.3996	<.0001	<.0001
21	4	0.0765	0.0056	0.065	0.3407
22	3	1.3996	0.0049	0.0003	0.4113
23	6	1.3334	<.0001	<.0001	0.0002
24	10	2.0056	<.0001	0.9334	<.0001
25	2	0.1194	0.3996	0.4346	0.1112
26	5	1.4808	<.0001	<.0001	<.0001
27	3	1.9557	0.0042	<.0001	0.0312
28	3	0.1337	0.0003	0.0108	0.183
29	14	1.9692	0.3877	<.0001	<.0001
30	4	1.7472	0.9207	0.3453	0.2972
31	2	4.393	0.0653	0.0824	0.0007
32	3	2.7021	0.3179	0.0556	0.3549
33	4	4.3779	0.1476	0.0335	0.4942
34	3	0.3046	0.009	0.1486	<.0001
35	2	0.1565	0.0221	0.003	0.2917
36	2	1.6177	0.0479	0.0129	0.5878

37	2	0.0115	0.3862	0.3186	0.8946
38	2	1.2504	0.0002	0.0233	0.1093
39	2	0.6908	<.0001	<.0001	0.02
40	5	0.4812	<.0001	<.0001	0.0231
41	3	1.0504	0.2401	0.2781	0.0265
42	2	0.0372	0.739	0.0283	0.0138
43	6	0.3987	0.0032	0.4618	0.0249
44	8	1.6129	0.0015	0.0134	0.4636
45	5	1.6802	0.135	0.0007	<.0001
46	2	0.0292	0.5558	0.6804	0.319
47	9	0.5469	0.0003	0.0079	0.3258
48	2	0.2974	0.1239	0.0002	0.0121
49	2	0.1214	1	0.7682	0.7682
50	4	0.8844	<.0001	0.1254	0.0012
51	5	0.7573	1	0.6328	0.6328
52	6	0.3351	0.3946	0.2617	0.7855
53	2	1.1689	0.0163	0.2304	0.2061
54	6	1.5758	<.0001	0.0002	<.0001
55	2	0.0956	0.109	0.02	0.4164
56	4	0.0898	<.0001	<.0001	0.0657
57	12	2.3151	<.0001	0.4326	<.0001
58	2	0.2014	0.0074	0.0083	<.0001
59	5	1.4419	0.0174	<.0001	<.0001
60	3	0.1262	0.3973	0.0535	0.2658
61	2	0.8986	<.0001	<.0001	1

Table A2 - Distribution of Z-statistic (n=2)

λ	n	mean	sd	P1	P5	P10	P25	P50	P75	P90	P95	P99
0.5	2	-0.9035	1.0268	-2	-2	-2	-2	-0.48033	0	0.3784	0.3784	0.6321
1	2	-0.5966	0.9847	-2	-2	-2	-2	0	0	0.3784	0.6321	0.8284
1.5	2	-0.4922	0.9314	-2	-2	-2	-0.7704	-0.1928	0.1297	0.5558	0.6321	0.8284
2	2	-0.4715	0.9130	-2	-2	-2	-0.7704	-0.1928	0.3166	0.5558	0.6321	0.9907
2.5	2	-0.4252	0.8605	-2	-2	-2	-0.7704	-0.3182	0.3166	0.5558	0.7494	0.9907
3	2	-0.4447	0.8628	-2	-2	-2	-0.7704	-0.3182	0.1297	0.6321	0.7494	1.0551
3.5	2	-0.3881	0.7996	-2	-2	-2	-0.7704	-0.3182	0.1297	0.5558	0.7549	1.1302
4	2	-0.4139	0.8267	-2	-2	-2	-0.7704	-0.3182	0.1491	0.5558	0.7494	1.0551
4.5	2	-0.3841	0.8438	-2	-2	-2	-0.8453	-0.3182	0.2494	0.5820	0.7867	1.1811
5	2	-0.4149	0.7983	-2	-2	-2	-0.8453	-0.3182	0.1297	0.5558	0.7549	1.0551
5.5	2	-0.4131	0.7988	-2	-2	-2	-0.9018	-0.4095	0.1491	0.5820	0.7549	1.1598
6	2	-0.4062	0.8251	-2	-2	-2	-0.9018	-0.3182	0.1491	0.5820	0.8284	1.1811
6.5	2	-0.4302	0.8026	-2	-2	-2	-0.9018	-0.3818	0.1420	0.5326	0.7549	1.1598
7	2	-0.4160	0.7881	-2	-2	-2	-0.9467	-0.3818	0.1491	0.5326	0.7549	1.1772
7.5	2	-0.3809	0.7617	-2	-2	-1.0657	-0.9018	-0.3408	0.1297	0.6060	0.8284	1.1598
8	2	-0.4353	0.7681	-2	-2	-2	-0.9467	-0.3408	0.1420	0.4719	0.6750	0.9907
8.5	2	-0.3573	0.7684	-2	-2	-1.0867	-0.9018	-0.2940	0.1491	0.6060	0.8028	1.2761
9	2	-0.4177	0.7776	-2	-2	-1.1382	-0.9837	-0.3818	0.1420	0.5345	0.7464	1.1811
9.5	2	-0.3739	0.7437	-2	-2	-1.1056	-0.9018	-0.3408	0.1420	0.5326	0.7494	1.2667
10	2	-0.3970	0.7747	-2	-2	-1.1382	-0.9837	-0.3818	0.1420	0.5345	0.8284	1.2928
10.5	2	-0.3644	0.7551	-2	-2	-1.1226	-0.9467	-0.3408	0.1420	0.5345	0.8574	1.1772
11	2	-0.3775	0.7445	-2	-2	-1.1226	-0.7704	-0.3818	0.1420	0.6060	0.8008	1.1623
11.5	2	-0.4082	0.7532	-2	-2	-1.1382	-1.0150	-0.3818	0.1297	0.5051	0.7464	1.2532
12	2	-0.3814	0.7313	-2	-2	-1.1382	-0.7914	-0.3435	0.1297	0.5345	0.6833	1.2249
12.5	2	-0.3509	0.7469	-2	-2	-1.1382	-0.7914	-0.3435	0.1695	0.5855	0.8574	1.2761
13	2	-0.3585	0.7413	-2	-2	-1.1524	-0.7914	-0.3182	0.1964	0.5558	0.7676	1.0682
13.5	2	-0.3536	0.7315	-2	-2	-1.1382	-0.7914	-0.3182	0.1695	0.5558	0.7398	1.1463
14	2	-0.4135	0.7505	-2	-2	-1.1655	-0.8287	-0.3889	0.1297	0.5428	0.7494	1.0561

14.5	2	-0.3624	0.7479	-2	-2	-1.1655	-0.8108	-0.3182	0.1455	0.5855	0.8284	1.3130
15	2	-0.4086	0.7634	-2	-2	-1.1891	-0.8608	-0.3435	0.0933	0.5558	0.7842	1.2249
15.5	2	-0.4184	0.7558	-2	-2	-1.1891	-0.8753	-0.4095	0.0933	0.5428	0.7494	1.2928
16	2	-0.3862	0.7415	-2	-2	-1.1777	-0.8287	-0.3435	0.1455	0.5428	0.7676	1.1975
16.5	2	-0.3372	0.7348	-2	-2	-1.1777	-0.8287	-0.3435	0.2078	0.6544	0.8497	1.2249
17	2	-0.3464	0.7398	-2	-2	-1.1777	-0.8287	-0.3182	0.1755	0.5855	0.8284	1.1838
17.5	2	-0.3761	0.7485	-2	-2	-1.1997	-0.8453	-0.3670	0.1455	0.5820	0.8497	1.1846
18	2	-0.3845	0.7202	-2	-2	-1.1997	-0.8453	-0.3670	0.1174	0.5034	0.7549	1.1846
18.5	2	-0.3743	0.7671	-2	-2	-1.2096	-0.8608	-0.3670	0.1964	0.5820	0.8112	1.3331
19	2	-0.3502	0.7386	-2	-2	-1.2096	-0.8453	-0.3182	0.1755	0.6054	0.7842	1.1407
19.5	2	-0.3775	0.7331	-2	-2	-1.2096	-0.8608	-0.3889	0.1297	0.5611	0.8112	1.1407
20	2	-0.3818	0.7300	-2	-2	-1.2096	-0.8608	-0.3435	0.1455	0.5149	0.7272	1.1168

Table A3 - Distribution of Z-statistic (n=3)

λ	n	mean	sd	P1	P5	P10	P25	P50	P75	P90	P95	P99
0.5	3	-0.7087	1.2741	-3	-3	-3	-0.8787	0	0	0.5676	0.5676	0.9482
1	3	-0.3538	0.9917	-3	-3	-1.1556	-0.4773	0	0	0.5676	0.8095	1.2426
1.5	3	-0.3083	0.9295	-3	-3	-1.1556	-0.8787	0	0.32005	0.5676	0.9482	1.2426
2	3	-0.2685	0.8620	-3	-1.4201	-1.1556	-0.7205	0	0.32005	0.5676	0.8095	1.2426
2.5	3	-0.2720	0.8453	-3	-1.3527	-1.2162	-0.7657	-0.1628	0.3200	0.7078	0.8870	1.2183
3	3	-0.2814	0.8679	-3	-1.4491	-1.2162	-0.7205	-0.2082	0.3200	0.7078	0.8966	1.3520
3.5	3	-0.2536	0.8245	-3	-1.4201	-1.3130	-0.7205	-0.1946	0.3200	0.7078	0.9482	1.2183
4	3	-0.3153	0.8416	-3	-1.4491	-1.3527	-0.7657	-0.2559	0.2633	0.7078	0.9482	1.3284
4.5	3	-0.3028	0.7789	-3	-1.5	-1.3130	-0.8067	-0.2559	0.2237	0.6742	0.8411	1.2183
5	3	-0.2799	0.8124	-3	-1.5226	-1.3527	-0.7469	-0.2082	0.2986	0.6887	0.9089	1.2426
5.5	3	-0.3089	0.8036	-3	-1.5631	-1.4201	-0.8067	-0.2892	0.2380	0.6835	0.9482	1.4129
6	3	-0.2806	0.7812	-3	-1.5226	-1.3130	-0.7718	-0.2715	0.2633	0.6742	0.9674	1.4093
6.5	3	-0.2810	0.7840	-3	-1.5814	-1.2162	-0.7469	-0.2602	0.2633	0.6649	0.9482	1.2862
7	3	-0.3221	0.7947	-3	-1.6148	-1.3130	-0.7718	-0.3063	0.1945	0.6887	0.9180	1.4093
7.5	3	-0.3149	0.7919	-3	-1.5814	-1.2162	-0.8067	-0.2715	0.1945	0.6438	0.9216	1.4264
8	3	-0.3168	0.7624	-1.7836	-1.6301	-1.2679	-0.7798	-0.3167	0.2131	0.6649	0.9193	1.3520
8.5	3	-0.2731	0.7863	-3	-1.6301	-1.2162	-0.7798	-0.2559	0.2633	0.7242	0.9646	1.3815
9	3	-0.2637	0.7449	-1.7576	-1.6148	-1.2162	-0.7718	-0.2559	0.2633	0.6771	0.9631	1.3140
9.5	3	-0.3194	0.8019	-3	-1.6584	-1.3130	-0.8320	-0.2559	0.2237	0.6565	0.8958	1.3200
10	3	-0.1973	0.7725	-1.7836	-1.6148	-1.2162	-0.7205	-0.1827	0.3638	0.8010	0.9631	1.5069
10.5	3	-0.3028	0.7803	-3	-1.6584	-1.2912	-0.8320	-0.2559	0.2300	0.7039	0.8958	1.3977
11	3	-0.3008	0.7580	-1.8071	-1.6958	-1.3130	-0.7798	-0.2772	0.2165	0.6343	0.8870	1.3583
11.5	3	-0.2757	0.7784	-3	-1.4756	-1.2798	-0.7469	-0.2728	0.2300	0.6986	0.9900	1.4216
12	3	-0.2704	0.7658	-1.8352	-1.5	-1.2679	-0.7755	-0.2559	0.2380	0.6617	0.9754	1.5211
12.5	3	-0.2703	0.7601	-1.8352	-1.4756	-1.2798	-0.7755	-0.2559	0.2553	0.7206	0.9303	1.3310
13	3	-0.3011	0.7977	-1.8828	-1.7576	-1.3527	-0.8442	-0.2559	0.2633	0.7078	0.9042	1.5009
13.5	3	-0.2778	0.7890	-1.8661	-1.7072	-1.3234	-0.7882	-0.2635	0.2557	0.7280	1.0183	1.3749
14	3	-0.2653	0.7893	-1.8828	-1.4756	-1.2679	-0.7798	-0.2831	0.2870	0.7637	1.0511	1.4078

14.5	3	-0.2653	0.7744	-1.8718	-1.5435	-1.3234	-0.7882	-0.2175	0.2946	0.7280	0.9162	1.4529
15	3	-0.2786	0.7797	-1.8828	-1.5226	-1.3432	-0.8127	-0.2356	0.2633	0.7021	0.9557	1.3301
15.5	3	-0.2840	0.7393	-1.8718	-1.5	-1.2318	-0.7968	-0.2515	0.2337	0.6343	0.8738	1.3694
16	3	-0.3026	0.7920	-3	-1.5435	-1.2912	-0.7968	-0.2964	0.2568	0.6638	0.9370	1.3623
16.5	3	-0.3136	0.7662	-1.9175	-1.8071	-1.4201	-0.7755	-0.2515	0.2110	0.6327	0.8313	1.3668
17	3	-0.3124	0.7670	-1.9309	-1.5814	-1.3007	-0.8244	-0.2635	0.2277	0.6455	0.9036	1.3171
17.5	3	-0.2907	0.7713	-1.9265	-1.5435	-1.2814	-0.7763	-0.2515	0.2553	0.6796	0.8885	1.3061
18	3	-0.2666	0.7480	-1.9175	-1.5226	-1.2162	-0.7679	-0.2439	0.2695	0.6835	0.8781	1.2998
18.5	3	-0.2654	0.7608	-1.9393	-1.5226	-1.2318	-0.7588	-0.2559	0.3021	0.6438	0.9280	1.4093
19	3	-0.2671	0.7846	-1.9309	-1.5631	-1.3964	-0.8067	-0.2281	0.2957	0.6981	1.0155	1.3472
19.5	3	-0.2619	0.7644	-1.9514	-1.5435	-1.2393	-0.7679	-0.2344	0.2568	0.7242	0.9719	1.4054
20	3	-0.2978	0.7804	-1.9553	-1.6148	-1.2748	-0.8351	-0.2715	0.2271	0.6917	0.9686	1.4398

Table A4 - Distribution of Z-statistic (n=4)

λ	n	mean	sd	P1	P5	P10	P25	P50	P75	P90	P95	P99
0.5	4	-0.6278	1.4368	-4	-4	-4	-0.3856	0	0	0.7568	0.7568	1.2643
1	4	-0.3254	0.9692	-4	-1.5408	-1.2536	-0.3856	-0.2984	0.2058	0.7568	0.7568	1.2643
1.5	4	-0.2608	0.9037	-4	-1.5408	-1.3250	-0.5971	-0.2984	0.2435	0.7568	0.9437	1.5766
2	4	-0.2411	0.8536	-4	-1.5829	-1.3250	-0.8043	-0.2984	0.2983	0.7568	1.0016	1.5351
2.5	4	-0.2386	0.9085	-4	-1.6906	-1.2741	-0.8043	-0.1512	0.2983	0.8633	1.0639	1.5351
3	4	-0.2171	0.8405	-2.0841	-1.7779	-1.2536	-0.7340	-0.1670	0.4019	0.8313	1.0016	1.5579
3.5	4	-0.2231	0.8058	-1.9675	-1.5829	-1.2536	-0.7340	-0.1957	0.2983	0.7568	1.0464	1.5766
4	4	-0.2929	0.8202	-2.0301	-1.6906	-1.3488	-0.8189	-0.2436	0.2681	0.7568	1.0188	1.4950
4.5	4	-0.2723	0.8441	-2.0841	-1.5829	-1.4143	-0.8750	-0.2314	0.2824	0.7672	1.0765	1.6305
5	4	-0.2426	0.8064	-2.0841	-1.5246	-1.2741	-0.7756	-0.2058	0.2866	0.7672	1.0574	1.4400
5.5	4	-0.2225	0.8162	-2.2111	-1.6148	-1.3488	-0.7636	-0.1535	0.3568	0.7960	1.0016	1.4950
6	4	-0.2576	0.8237	-2.2111	-1.5829	-1.3575	-0.8189	-0.2329	0.3663	0.7568	1.0221	1.4804
6.5	4	-0.2864	0.8231	-2.2111	-1.6906	-1.3880	-0.8189	-0.2436	0.2883	0.7568	1.0653	1.4796
7	4	-0.2389	0.8395	-2.2200	-1.6906	-1.3525	-0.8144	-0.1853	0.3778	0.7960	1.0188	1.5034
7.5	4	-0.2141	0.7984	-2.2452	-1.5246	-1.2536	-0.8001	-0.1895	0.3639	0.7960	1.0838	1.6223
8	4	-0.2484	0.7853	-2.2111	-1.6542	-1.3108	-0.7583	-0.2283	0.2725	0.7167	1.0574	1.7130
8.5	4	-0.1980	0.8264	-2.3311	-1.6148	-1.2536	-0.7498	-0.1656	0.3731	0.8655	1.1280	1.5465
9	4	-0.2880	0.8548	-2.3311	-1.6906	-1.3382	-0.8795	-0.2436	0.2914	0.7925	1.0574	1.7338
9.5	4	-0.2546	0.8565	-2.3048	-1.7216	-1.3382	-0.8282	-0.2436	0.3434	0.8382	1.1403	1.6751
10	4	-0.2543	0.8262	-2.3311	-1.7216	-1.3880	-0.7718	-0.2246	0.2983	0.7568	1.0703	1.5410
10.5	4	-0.2472	0.8317	-2.3495	-1.6398	-1.3382	-0.8353	-0.1987	0.2983	0.8028	1.0952	1.6456
11	4	-0.2583	0.8382	-2.3555	-1.6906	-1.3250	-0.8162	-0.2170	0.3333	0.7754	1.0601	1.5465
11.5	4	-0.2064	0.8128	-2.3836	-1.5829	-1.2536	-0.7340	-0.1827	0.3667	0.7914	1.0694	1.5914
12	4	-0.2437	0.8215	-2.4380	-1.6290	-1.3540	-0.7718	-0.2314	0.2932	0.7568	1.0524	1.6118
12.5	4	-0.2200	0.8284	-2.0580	-1.6398	-1.2589	-0.7804	-0.2242	0.3395	0.8903	1.1254	1.6902
13	4	-0.2434	0.8396	-2.3495	-1.6290	-1.3250	-0.8162	-0.2314	0.2983	0.8169	1.1787	1.7022
13.5	4	-0.2484	0.8131	-2.1314	-1.5829	-1.3095	-0.7995	-0.2398	0.2923	0.8145	1.0757	1.5947
14	4	-0.2606	0.8285	-2.1634	-1.6290	-1.3760	-0.7935	-0.2487	0.2819	0.8560	1.1233	1.5481

14.5	4	-0.2500	0.8187	-2.1634	-1.5829	-1.3095	-0.8189	-0.2490	0.3309	0.8015	1.0728	1.4747
15	4	-0.2985	0.8378	-2.5140	-1.7095	-1.3760	-0.8189	-0.2658	0.2594	0.7568	1.0204	1.6226
15.5	4	-0.2146	0.8380	-2.0353	-1.6574	-1.3023	-0.7636	-0.1947	0.3555	0.8737	1.1622	1.5822
16	4	-0.2468	0.8277	-2.1634	-1.6906	-1.3760	-0.8020	-0.1996	0.3278	0.7849	1.0101	1.5957
16.5	4	-0.2697	0.8071	-2.0965	-1.6298	-1.3326	-0.8252	-0.2658	0.2766	0.7776	1.0506	1.4373
17	4	-0.2442	0.8111	-2.0965	-1.5829	-1.3311	-0.7995	-0.2043	0.3412	0.7568	1.0607	1.5381
17.5	4	-0.3106	0.8180	-2.2688	-1.7046	-1.3880	-0.8628	-0.2940	0.2406	0.7109	0.9546	1.6121
18	4	-0.2351	0.8461	-2.1314	-1.7294	-1.3382	-0.8043	-0.1866	0.3337	0.8501	1.1392	1.7149
18.5	4	-0.2788	0.8148	-2.1634	-1.6574	-1.3760	-0.8628	-0.2378	0.2983	0.7690	1.0008	1.5070
19	4	-0.2489	0.8473	-2.2200	-1.6457	-1.3382	-0.8709	-0.2137	0.3764	0.8063	1.1167	1.6347
19.5	4	-0.2752	0.8111	-2.2688	-1.6761	-1.3311	-0.8043	-0.2837	0.2947	0.7711	1.0155	1.5782
20	4	-0.2652	0.8256	-2.1833	-1.7216	-1.3964	-0.7940	-0.2281	0.2953	0.7765	1.0204	1.5823

Table A5- Distribution of Z-statistic (n=5)

λ	n	mean	sd	P1	P5	P10	P25	P50	P75	P90	P95	P99
0.5	5	-0.5499	1.5255	-5	-5	-1.4645	-0.7955	-0.1642	0	0.6607	0.9460	1.5804
1	5	-0.1882	0.8845	-5	-1.5981	-1.0516	-0.7955	-0.1642	0.3728	0.9460	1.0368	1.6192
1.5	5	-0.2312	0.8525	-2.0859	-1.5981	-1.2448	-0.7955	-0.1642	0.3728	0.7735	1.0368	1.6192
2	5	-0.2577	0.8988	-2.2545	-1.8053	-1.4645	-0.7955	-0.2529	0.3023	0.8181	1.0952	1.7134
2.5	5	-0.2612	0.9050	-2.3136	-1.8053	-1.4645	-0.7955	-0.2079	0.3608	0.7847	1.0993	1.6192
3	5	-0.1876	0.8320	-2.2749	-1.5294	-1.2448	-0.7463	-0.1642	0.3728	0.8336	1.1237	1.7134
3.5	5	-0.1965	0.8981	-2.4450	-1.5805	-1.2761	-0.7485	-0.2079	0.4143	0.9460	1.1892	1.8659
4	5	-0.2208	0.8789	-2.4151	-1.7243	-1.3444	-0.7955	-0.2079	0.3728	0.8990	1.1638	1.8245
4.5	5	-0.2131	0.9046	-2.4450	-1.7243	-1.4645	-0.7955	-0.2103	0.4347	0.9460	1.2872	1.7195
5	5	-0.2427	0.8973	-2.1883	-1.8053	-1.4645	-0.7955	-0.2151	0.3914	0.8831	1.1797	1.6756
5.5	5	-0.2490	0.8725	-2.4151	-1.7350	-1.3544	-0.7955	-0.2332	0.3426	0.8831	1.1575	1.7680
6	5	-0.2426	0.8958	-2.4450	-1.6563	-1.3819	-0.8588	-0.2158	0.3242	0.8922	1.2475	1.8166
6.5	5	-0.2614	0.9109	-2.6913	-1.8639	-1.4977	-0.7955	-0.2103	0.3575	0.8181	1.1575	1.7869
7	5	-0.2394	0.8910	-2.3446	-1.6852	-1.3831	-0.8505	-0.2228	0.3948	0.8703	1.1416	1.8389
7.5	5	-0.2509	0.8957	-2.3136	-1.8053	-1.4645	-0.8391	-0.2291	0.3690	0.8831	1.1797	1.7165
8	5	-0.2611	0.8808	-2.3136	-1.7732	-1.3819	-0.7955	-0.2529	0.3177	0.8685	1.1696	1.7372
8.5	5	-0.2385	0.8739	-2.2387	-1.6563	-1.3701	-0.8588	-0.2291	0.3638	0.9112	1.1907	1.6215
9	5	-0.1958	0.8955	-2.2973	-1.7732	-1.3588	-0.7955	-0.1865	0.4419	0.9730	1.2341	1.7134
9.5	5	-0.2506	0.8791	-2.3740	-1.7098	-1.4015	-0.8505	-0.2273	0.4112	0.8584	1.1174	1.5857
10	5	-0.2475	0.9190	-2.4772	-1.8482	-1.4953	-0.8805	-0.2176	0.3728	0.9207	1.2352	1.7949
10.5	5	-0.1948	0.8693	-2.3793	-1.7732	-1.3256	-0.7754	-0.1509	0.4251	0.8667	1.1346	1.6052
11	5	-0.2687	0.9228	-2.4151	-1.8053	-1.5019	-0.8892	-0.2179	0.3541	0.8732	1.2133	1.8516
11.5	5	-0.2189	0.8893	-2.4409	-1.7187	-1.3918	-0.7810	-0.1862	0.3852	0.9100	1.2006	1.7479
12	5	-0.2704	0.8980	-2.3446	-1.8053	-1.4953	-0.8750	-0.2463	0.3583	0.9100	1.1732	1.6942
12.5	5	-0.2321	0.8930	-2.3136	-1.7318	-1.3984	-0.8316	-0.2012	0.3678	0.9120	1.2711	1.6657
13	5	-0.2554	0.8941	-2.4801	-1.7936	-1.4322	-0.8391	-0.2021	0.3493	0.8880	1.1237	1.6004
13.5	5	-0.1683	0.8972	-2.3136	-1.6889	-1.3588	-0.7810	-0.1208	0.4128	1.0368	1.3000	1.8159
14	5	-0.1772	0.8984	-2.4801	-1.7318	-1.4021	-0.7535	-0.1545	0.4605	0.9043	1.2551	1.8016

14.5	5	-0.2197	0.9003	-2.3496	-1.7497	-1.3749	-0.8224	-0.1877	0.4083	0.9132	1.2011	1.7574
15	5	-0.2417	0.8788	-2.4377	-1.7323	-1.3469	-0.7955	-0.2486	0.3329	0.8444	1.2175	1.7860
15.5	5	-0.1889	0.9012	-2.3136	-1.7853	-1.4136	-0.7955	-0.1492	0.4549	0.9460	1.2617	1.7613
16	5	-0.2089	0.8946	-2.3915	-1.7455	-1.3304	-0.8242	-0.2191	0.4375	0.9246	1.1865	1.8389
16.5	5	-0.2867	0.9225	-2.4522	-1.8856	-1.4852	-0.9087	-0.2992	0.3993	0.8893	1.2114	1.7450
17	5	-0.2171	0.9181	-2.5809	-1.7623	-1.3867	-0.7955	-0.2323	0.4199	0.9460	1.2994	1.7840
17.5	5	-0.2448	0.9046	-2.4303	-1.6872	-1.3654	-0.8628	-0.2818	0.3710	0.9503	1.2406	1.8078
18	5	-0.2248	0.9306	-2.3752	-1.7623	-1.4916	-0.8450	-0.2135	0.4688	0.9645	1.2538	1.8501
18.5	5	-0.2135	0.8942	-2.5459	-1.7462	-1.3325	-0.7955	-0.2079	0.3874	0.9053	1.2799	1.7006
19	5	-0.1786	0.8889	-2.4291	-1.6822	-1.2996	-0.7413	-0.1522	0.4409	0.9460	1.2439	1.8516
19.5	5	-0.2266	0.8981	-2.3698	-1.7243	-1.3701	-0.8502	-0.2103	0.4143	0.9310	1.2320	1.7327
20	5	-0.2181	0.8783	-2.1621	-1.7907	-1.4158	-0.7853	-0.1970	0.3967	0.8726	1.1847	1.7099

Table A6 - Distribution of Z-statistic (n=6)

λ	n	mean	sd	P1	P5	P10	P25	P50	P75	P90	P95	P99
0.5	6	-0.4153	1.4443	-6	-6	-1.2284	-0.5515	-0.3256	0	0.7481	1.1352	1.7614
1	6	-0.1934	0.9886	-6	-1.5262	-1.2284	-0.7193	-0.2756	0.3653	0.7899	1.2399	1.6976
1.5	6	-0.2010	0.9705	-2.5114	-1.9876	-1.2284	-0.7193	-0.2756	0.3653	0.9846	1.3073	1.9226
2	6	-0.1984	0.9508	-2.3743	-1.7574	-1.4144	-0.8872	-0.1887	0.4794	0.9772	1.3073	1.8964
2.5	6	-0.2450	0.9456	-2.5114	-1.7574	-1.5262	-0.8726	-0.2985	0.4261	0.9498	1.2513	1.8522
3	6	-0.2000	0.9898	-2.6260	-1.9876	-1.5076	-0.8726	-0.1238	0.4721	1.0132	1.3620	2.0249
3.5	6	-0.2201	0.9204	-2.3743	-1.6971	-1.4593	-0.8330	-0.2186	0.4243	0.9672	1.3468	1.8359
4	6	-0.2195	0.9848	-2.8090	-1.8359	-1.4593	-0.8145	-0.1793	0.4261	1.0107	1.3468	1.9789
4.5	6	-0.2196	0.9383	-2.4891	-1.8359	-1.5076	-0.8177	-0.1685	0.4204	0.9220	1.2802	1.9338
5	6	-0.2032	0.9508	-2.3564	-1.7204	-1.4001	-0.8425	-0.1887	0.4204	1.0198	1.3073	2.0289
5.5	6	-0.2403	0.9632	-2.6503	-1.8514	-1.4873	-0.8781	-0.2048	0.4284	0.9793	1.2570	1.8724
6	6	-0.2225	0.9474	-2.5114	-1.8257	-1.4545	-0.8603	-0.2048	0.4365	0.9793	1.3073	1.7614
6.5	6	-0.2775	0.9218	-2.4860	-1.8378	-1.4410	-0.9049	-0.2843	0.3729	0.8743	1.1940	1.7614
7	6	-0.1824	0.9338	-2.5824	-1.7382	-1.3850	-0.7790	-0.1520	0.4589	1.0132	1.3403	1.9152
7.5	6	-0.1915	0.9473	-2.5114	-1.8248	-1.3850	-0.8096	-0.1435	0.4675	0.9580	1.3714	2.0165
8	6	-0.2575	0.9602	-2.5809	-1.8652	-1.5262	-0.9049	-0.2423	0.4067	0.9711	1.3073	1.9762
8.5	6	-0.2115	0.9555	-2.5114	-1.8744	-1.4410	-0.8603	-0.1900	0.4662	1.0007	1.2851	2.1137
9	6	-0.2753	0.9474	-2.5114	-1.8744	-1.4545	-0.9049	-0.2985	0.3781	0.9359	1.3022	1.8863
9.5	6	-0.2068	0.9733	-2.5	-1.8805	-1.4574	-0.8641	-0.2033	0.4558	1.0542	1.3890	2.0664
10	6	-0.2204	0.9373	-2.7418	-1.8164	-1.4324	-0.8177	-0.1793	0.4610	0.9498	1.2713	1.7671
10.5	6	-0.1430	0.9161	-2.4242	-1.6884	-1.3268	-0.7598	-0.0914	0.4710	1.0235	1.3309	1.7976
11	6	-0.2320	0.9254	-2.4512	-1.8953	-1.4745	-0.8429	-0.1483	0.4067	0.9168	1.2774	1.7456
11.5	6	-0.1494	0.9405	-2.3453	-1.7574	-1.3985	-0.7833	-0.1452	0.5092	1.0769	1.3324	1.9556
12	6	-0.2203	0.9647	-2.4686	-1.8993	-1.4937	-0.8573	-0.2206	0.4578	1.0257	1.3439	2.0206
12.5	6	-0.2244	0.9596	-2.5944	-1.8212	-1.4873	-0.8603	-0.2109	0.4197	1.0074	1.3453	1.9540
13	6	-0.2179	0.9335	-2.4512	-1.8396	-1.4363	-0.8498	-0.1764	0.4230	0.9708	1.3422	1.7447
13.5	6	-0.2502	0.9218	-2.4766	-1.7419	-1.4049	-0.9021	-0.2487	0.4050	0.9095	1.2506	1.8016
14	6	-0.1629	0.9606	-2.4948	-1.7738	-1.4001	-0.8221	-0.1475	0.5199	1.0688	1.3890	2.0249

14.5	6	-0.1711	0.9696	-2.4498	-1.7644	-1.4457	-0.8316	-0.1865	0.5265	1.0594	1.3680	1.9593
15	6	-0.2629	1.0035	-2.7535	-1.9876	-1.6422	-0.9392	-0.2165	0.4736	0.9498	1.3215	1.8735
15.5	6	-0.2391	0.9528	-2.6561	-1.9223	-1.5307	-0.8316	-0.2037	0.4162	0.9539	1.2426	1.8556
16	6	-0.2180	0.9689	-2.4498	-1.8396	-1.4925	-0.8433	-0.1916	0.4430	1.0632	1.3680	1.8907
16.5	6	-0.2364	0.9777	-2.6561	-1.8765	-1.5379	-0.9138	-0.1793	0.4569	0.9592	1.3073	1.8761
17	6	-0.1643	0.9944	-2.5629	-1.7574	-1.4790	-0.8433	-0.1495	0.5380	1.1073	1.4574	2.1720
17.5	6	-0.1749	0.9533	-2.3855	-1.8124	-1.3754	-0.8498	-0.1545	0.5084	1.0741	1.3664	1.9106
18	6	-0.2310	0.9688	-2.6016	-1.9512	-1.5189	-0.8433	-0.1676	0.4179	0.9804	1.2706	1.8989
18.5	6	-0.2462	0.9628	-2.5359	-1.8816	-1.4545	-0.9049	-0.2459	0.4332	0.9345	1.3039	1.9447
19	6	-0.2629	0.9901	-2.7640	-1.9876	-1.5379	-0.9209	-0.2172	0.4009	0.9725	1.3073	1.8261
19.5	6	-0.2818	0.9479	-2.7763	-1.9876	-1.4812	-0.8637	-0.2201	0.3913	0.8463	1.1413	1.7126
20	6	-0.2784	0.9876	-2.7358	-1.8884	-1.5468	-0.9489	-0.2659	0.4336	1.0003	1.2792	1.8475

Table A7 - Distribution of Z-statistic (n=7)

λ	n	mean	sd	P1	P5	P10	P25	P50	P75	P90	P95	P99
0.5	7	-0.3515	1.5074	-7	-1.6812	-1.1137	-0.6748	-0.3119	0.1415	0.9535	1.3244	2.1260
1	7	-0.1843	0.9343	-2.5274	-1.6812	-1.3206	-0.6748	-0.0993	0.4625	1.0319	1.3244	1.8183
1.5	7	-0.2477	0.9898	-2.7267	-1.6812	-1.4543	-0.8001	-0.1506	0.4625	0.9535	1.3244	1.8312
2	7	-0.2084	1.0040	-2.7267	-2.0188	-1.4606	-0.8001	-0.1506	0.4625	1.0405	1.3773	2.0300
2.5	7	-0.2251	0.9941	-2.5274	-2.0188	-1.5121	-0.9566	-0.2031	0.4905	0.9865	1.3868	2.2125
3	7	-0.1801	0.9680	-2.5274	-1.8929	-1.5121	-0.7645	-0.1935	0.5057	1.0829	1.3589	1.8142
3.5	7	-0.1273	0.9913	-2.5274	-1.8681	-1.4331	-0.7645	-0.1444	0.5220	1.1453	1.4763	2.1370
4	7	-0.2092	0.9685	-2.4833	-1.9181	-1.5222	-0.8614	-0.1796	0.5038	1.0424	1.3244	1.9077
4.5	7	-0.2488	0.9868	-2.6160	-1.9269	-1.5569	-0.9099	-0.2031	0.4233	0.9688	1.2808	2.0195
5	7	-0.1964	0.9856	-2.4027	-1.8605	-1.4623	-0.8392	-0.2031	0.4775	1.0482	1.4530	2.0897
5.5	7	-0.2772	0.9832	-2.7701	-1.8605	-1.5569	-0.9099	-0.3041	0.4293	0.9656	1.2883	1.8662
6	7	-0.2359	1.0154	-2.7589	-1.9829	-1.5765	-0.9226	-0.2031	0.4856	1.0118	1.3627	2.0655
6.5	7	-0.2297	0.9957	-2.3476	-1.8430	-1.5946	-0.9266	-0.1913	0.4407	1.0319	1.4236	2.1073
7	7	-0.1888	0.9796	-2.4027	-1.8430	-1.4382	-0.8825	-0.1560	0.4551	1.0228	1.4218	2.0715
7.5	7	-0.2676	1.0168	-2.6965	-1.9697	-1.6317	-1.0107	-0.2256	0.4317	1.0108	1.3521	2.0353
8	7	-0.2540	1.0192	-2.6572	-1.9697	-1.5870	-0.9566	-0.2298	0.4456	1.0755	1.3906	1.9452
8.5	7	-0.2355	1.0307	-2.7701	-1.9697	-1.5392	-0.9185	-0.2479	0.4153	1.1353	1.4236	2.0873
9	7	-0.2051	1.0038	-2.6253	-1.9217	-1.5109	-0.8852	-0.1820	0.5137	1.0829	1.4098	2.0279
9.5	7	-0.2962	1.0264	-2.9444	-2.0907	-1.5974	-0.9363	-0.3164	0.3753	0.9999	1.3580	2.0966
10	7	-0.2599	0.9966	-2.7701	-1.9181	-1.5529	-0.9363	-0.2247	0.4745	0.9535	1.2408	1.9539
10.5	7	-0.1879	1.0334	-2.8662	-1.9221	-1.4360	-0.8392	-0.1648	0.5045	1.0999	1.4335	1.9452
11	7	-0.2093	0.9946	-2.4630	-1.8237	-1.4722	-0.9335	-0.1946	0.4918	1.0569	1.3399	2.0939
11.5	7	-0.1767	1.0298	-2.6508	-1.8853	-1.5131	-0.8854	-0.1274	0.5562	1.1058	1.4946	1.9999
12	7	-0.2337	0.9856	-2.7188	-1.9313	-1.5131	-0.8409	-0.1684	0.4030	0.9220	1.3324	1.9655
12.5	7	-0.1951	1.0054	-2.4630	-1.8797	-1.5131	-0.9156	-0.1946	0.5009	1.1167	1.4599	1.9804
13	7	-0.2511	1.0039	-2.6926	-1.9697	-1.5541	-0.9208	-0.2376	0.4476	0.9887	1.3618	2.1301
13.5	7	-0.2267	1.0040	-2.8012	-1.8987	-1.4833	-0.8971	-0.2460	0.5029	1.0829	1.3607	1.9143
14	7	-0.2136	1.0279	-2.7240	-1.9524	-1.6388	-0.9349	-0.1895	0.5407	1.0420	1.4175	2.0622

14.5	7	-0.2211	1.0184	-2.7545	-1.9995	-1.5450	-0.8971	-0.2011	0.4667	1.0342	1.3305	2.1804
15	7	-0.1894	1.0152	-2.5442	-1.9138	-1.5345	-0.9139	-0.1576	0.5468	1.0698	1.3321	2.2447
15.5	7	-0.1908	1.0210	-2.6088	-1.9697	-1.5682	-0.8574	-0.1262	0.5177	1.0702	1.4443	2.0418
16	7	-0.2323	0.9996	-2.5947	-1.8840	-1.4838	-0.9024	-0.2404	0.4225	1.0860	1.4426	2.1829
16.5	7	-0.2336	0.9972	-2.4585	-1.9055	-1.4970	-0.8873	-0.2277	0.3730	1.0493	1.5066	2.1231
17	7	-0.2582	0.9905	-2.7559	-1.9877	-1.5322	-0.8737	-0.2259	0.3893	0.9380	1.3244	2.0914
17.5	7	-0.2536	0.9958	-2.5985	-1.9555	-1.5860	-0.9471	-0.2130	0.4424	0.9840	1.2670	1.9802
18	7	-0.1875	1.0018	-2.5874	-1.8911	-1.5127	-0.8155	-0.1495	0.4658	1.0562	1.4084	2.1969
18.5	7	-0.1703	1.0055	-2.6506	-1.8789	-1.4477	-0.8737	-0.1495	0.5714	1.0939	1.3780	1.9503
19	7	-0.2372	1.0276	-2.7737	-1.9292	-1.5802	-0.9471	-0.1861	0.4984	1.0618	1.3144	1.9154
19.5	7	-0.2172	0.9825	-2.801	-1.8773	-1.5140	-0.8591	-0.1705	0.4234	1.0088	1.3841	2.0669
20	7	-0.2328	1.0302	-2.8539	-1.9298	-1.5429	-0.9091	-0.2017	0.4302	1.0450	1.4554	2.2335