The Temporal Correlation between Two Random Processes

A Project
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Zizhen Wu
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Master Of Science

BARRY JAMES, KANG JAMES

July, 2012
Acknowledgements

I would like to take this opportunity to give my sincere thanks to my advisors, Dr. Barry James, and Dr. Kang James, for their support and guidance, which make it possible for me to finish the project on time. Thanks also to Dr. Bruce Peckham for being on my committee, and all three courses I took from him.

I would also like to thank our department to provide a supportive and friendly environment for my study in Duluth.
Abstract

Detecting the temporal correlation between two random processes has a wide application in many scientific research areas, such as astronomy, epidemiology, and biology. In this paper, a method based on U-statistics proposed by Bickel, James, and James is presented. Simulations are done to find empirically optimal values of parameters of that method.
# Contents

**Acknowledgements**  
1

**Abstract**  
ii

**List of Tables**  
v

**List of Figures**  
vi

1 **Introduction**  
1.1 Literature Review  
1.1.1 Method Proposed by A. Hugeback  
1.1.2 Method proposed by Bickel, James, and James  
1.1.3 Summary of the Two Methods  
1

2 **Method proposed by Barry James and Kang James**  
7

3 **Simulation**  
3.1 An Ideal Case: Two Uniform Densities  
3.2 Relax the Restriction  
3.3 $Z_e$ with Minimum Number of Data Points in Each Subinterval  
3.4 How to Choose $\delta$?  
3.5 How to choose $\epsilon$?  
3.6 Method of Using Weights  
3.7 More Simulations  
3.8 Summary for the Large Sample Case  
12  
13  
14  
15  
16  
17  
18  
19  
24
3.9 Application to the Small Sample Size

3.9.1 Choose appropriate δ value when sample size is small

3.9.2 Choose Appropriate r Value When Sample Size is Small

3.9.3 Conclusion for the Small Sample Case

4 Applications

References

Appendix A. U-statistics

Appendix B. R Code

B.1 An illustration of R Code

B.2 Function that can pick the minimum number of data points

B.3 Complete Z Function

B.4 Program of Choosing δ and r
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>parameters of the simulation</td>
<td>13</td>
</tr>
<tr>
<td>3.2</td>
<td>Summary statistics</td>
<td>13</td>
</tr>
<tr>
<td>3.3</td>
<td>Summary Statistics</td>
<td>15</td>
</tr>
<tr>
<td>3.4</td>
<td>summary statistics: controlling minimum numbers of Data Points</td>
<td>16</td>
</tr>
<tr>
<td>3.5</td>
<td>parameters of the simulation</td>
<td>16</td>
</tr>
<tr>
<td>3.6</td>
<td>summary statistics</td>
<td>17</td>
</tr>
<tr>
<td>3.7</td>
<td>parameters of the simulation</td>
<td>17</td>
</tr>
<tr>
<td>3.8</td>
<td>Summary statistics</td>
<td>18</td>
</tr>
<tr>
<td>3.9</td>
<td>Summary Statistics: with weights</td>
<td>19</td>
</tr>
<tr>
<td>3.10</td>
<td>Summary Statistics: comparison of two methods</td>
<td>20</td>
</tr>
<tr>
<td>3.11</td>
<td>Means, variances, and $p$-values under different $r$</td>
<td>22</td>
</tr>
<tr>
<td>3.12</td>
<td>parameters of the simulation</td>
<td>23</td>
</tr>
<tr>
<td>3.13</td>
<td>Summary Statistics of $Z_e$</td>
<td>24</td>
</tr>
<tr>
<td>3.14</td>
<td>$\delta$ value corresponding to $p$-value&lt;0.05</td>
<td>25</td>
</tr>
<tr>
<td>3.15</td>
<td>$\delta$ value corresponding to $p$-value&lt;0.05</td>
<td>27</td>
</tr>
<tr>
<td>3.16</td>
<td>$r$ value corresponding to $p$-value&lt;0.05</td>
<td>29</td>
</tr>
<tr>
<td>3.17</td>
<td>$\delta$ value corresponding to $p$-value&lt;0.05</td>
<td>30</td>
</tr>
</tbody>
</table>
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Histogram of $\tau_D$</td>
<td>2</td>
</tr>
<tr>
<td>2.1</td>
<td>An Illustration of Calculating $p_c$</td>
<td>9</td>
</tr>
<tr>
<td>3.1</td>
<td>Histograms and $Q-Q$ Plot</td>
<td>14</td>
</tr>
<tr>
<td>3.2</td>
<td>Histogram of linear density, histogram of $Z_c$, and $Q-Q$ Plot of $Z_c$</td>
<td>14</td>
</tr>
<tr>
<td>3.3</td>
<td>Histogram of triangular density, histogram of $Z_c$, and $Q-Q$ Plot of $Z_c$</td>
<td>16</td>
</tr>
<tr>
<td>3.4</td>
<td>Histogram and $Q-Q$-Plot of $Z_t$ using weights</td>
<td>19</td>
</tr>
<tr>
<td>3.5</td>
<td>Histogram and $Q-Q$-Plot of $Z_c$ without weights</td>
<td>20</td>
</tr>
<tr>
<td>3.6</td>
<td>Histograms of Densities</td>
<td>20</td>
</tr>
<tr>
<td>3.7</td>
<td>Means and variances</td>
<td>21</td>
</tr>
<tr>
<td>3.8</td>
<td>Two Triangular Densities</td>
<td>23</td>
</tr>
<tr>
<td>3.9</td>
<td>Means and Variances for $\delta = 33$ to 57</td>
<td>26</td>
</tr>
<tr>
<td>3.10</td>
<td>Histograms of Densities</td>
<td>27</td>
</tr>
<tr>
<td>3.11</td>
<td>Means and Variances of $Z_c$ values</td>
<td>28</td>
</tr>
<tr>
<td>3.12</td>
<td>Means and Variances for $r = 0.01$ to 0.5</td>
<td>30</td>
</tr>
<tr>
<td>3.13</td>
<td>Means and Variances for $\delta = 5$ to 100</td>
<td>31</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Given two random processes, a natural question would be: are these two random processes correlated in a local sense? To be more specific, does the occurrence of one event imply the high possibility of occurrence of another event in a short time interval around it? The detection of such a temporal correlation can be applied to many scientific problems. One application is about the detection of temporal correlation between solar flares and coronal mass ejections (A. Hugeback, 2008). A solar flare is a sudden brightening observed over the Sun’s surface or the solar limb. CME is a massive burst of solar wind and magnetic fields rising above the solar corona or being released into space. The study of such phenomena is of importance, since they can potentially generate geomagnetic storms that cause electrical disturbances on or near Earth.

1.1 Literature Review

1.1.1 Method Proposed by A. Hugeback

A. Hugeback[1] proposed a method to detect temporal correlation given the onset time of solar flares $\tau_F$ and CMEs $\tau_C$. First, a secondary measurement $\tau_D$ is calculated: $\tau_D = \tau_C - \tau_D$. Then, define $T = \{\tau_D | \tau_D \in [-W, W]\}$ for some $W$, which is small compared to the total time interval, since we want to detect the temporal correlation. In Figure-1.1 there is a bump near zero, which indicates strongly that there is a temporal correlation.
Then, two parametric models are proposed by Hugeback to fit $\tau_D$. Based on Figure 1.1, one piecewise-uniform model and one Gaussian-Uniform mixture model are fitted, respectively.

The piecewise-uniform model is of the form:

$$f(t; \gamma, \mu, \sigma) = \begin{cases} \frac{\gamma}{2W} + \frac{1-\gamma}{2\sigma}, & \text{if } t \in [\mu - \sigma, \mu + \sigma] \\ \frac{\gamma}{2W}, & \text{otherwise} \end{cases}$$

The log-likelihood function of the piecewise-uniform density is:

$$l(\gamma, \mu, \sigma | T) = |T_0| \cdot \ln\left(\frac{\gamma}{2W}\right) + |T_1| \cdot \left[\frac{\gamma}{2W} + \frac{1-\gamma}{2\sigma}\right]$$

where $|T_0|$ = number of relative time points that lie outside the bump, $|T_1|$ = number of relative time points that lie inside the bump.

The Gaussian-Uniform mixture model is of the form:

$$f(t; \gamma, \mu, \sigma) = \frac{\gamma}{2W} + \frac{1-\gamma}{I} \cdot g(t; \mu, \sigma), \ t \in [-W, W]$$

where $g(t; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t-\mu)^2}{2\sigma^2}}$ is the normal density with mean $\mu$ and variance $\sigma^2$. $I$ is the integral of $g(t; \mu, \sigma)$ over $[-W, W]$, which makes $f(t; \gamma, \mu, \sigma)$ a legitimate probability density function.
The log-likelihood function of the mixture model density is:

\[ l(\gamma, \mu, \sigma | T) = \sum_{i=1}^{n} \ln \left[ \frac{\gamma}{2W} + \frac{1-\gamma}{I} \cdot g(t_i; \mu, \sigma) \right] \]

To maximize the log-likelihood function, a numerical approach called semi-greedy grid search was used. The Gaussian-uniform model has larger likelihood than the piece-wise mixture model. Hence, the second model is utilized to model \( \tau_D \). For the details, see [1].

To test

\( H_0 : \) The two sequences are independent

versus

\( H_1 : \) The two sequences are temporally correlated,

a block-of-blocks bootstrap was used. The total length of the time spectrum for the data was 2,499,840 minutes, and the spectrum was divided into \( K=499 \) blocks of length \( L=5000 \) minutes. The procedure was: For iteration \( r = 1 \) to 499:

- For \( k = 1 \) to \( K \),
  1. Select a block \( B_1 \) at random with replacement from the set of \( K \) blocks.
  2. Condition on the number of flare events in block \( B_1 \), and select a block \( B_2 \) at random with replacement from the set of \( K \) blocks such that \( B_1 \) and \( B_2 \) contain the same number (or a similar number) of flare events. Flares that occur during high-activity periods are paired with CMEs that also occurred during periods with high flaring activity, and vice versa.
  3. Overlay the two blocks \( B_1 \) and \( B_2 \) on top of each other so that the starting points and ending points are the same.
  4. Compute the relative onset times for each CME event in block \( B_2 \) with respect to each flare event in block \( B_1 \).
- Combine all relative onset times from those \( K \) resampled blocks into a final collection.
- Fit a Gaussian-uniform mixture model and a uniform model to the relative time points.
• Compute the likelihood-ratio test statistic for the resampled data.

The ratio of log-likelihood for the real data is $LR = 6$. The $p$-value of the bootstrapped significance test is $p < 0.02$.

1.1.2 Method proposed by Bickel, James, and James

Motivated by A. Hugeback’s method, a statistic

$$T_\epsilon = \sum_{i=1}^{m} \sum_{j=1}^{n} I[|X_i - Y_j| < \epsilon]$$

is defined, where $I$ is the indicator function with the property that

$$I[|X_i - Y_j| < \epsilon] = \begin{cases} 
1 & \text{if } |X_i - Y_j| < \epsilon \\
0 & \text{otherwise.} 
\end{cases}$$

Then $T_\epsilon$ is put into the framework of $U$-statistic. A test statistic is proposed under the null hypothesis $H_0$: two processes are independent.

**Theorem.** Ross (2006). Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be two sequences from two non-homogeneous Poisson processes with intensity function $\lambda_1(t)$ and $\lambda_2(t)$, respectively. Suppose the processes are observed to time $c > 0$. Given $N_i(c) = n_i$, the times of the $n_i$ occurrences of event $i$ by time $c$ can be considered as order statistics of $n_i$ i.i.d random variables having common pdf:

$$f_i(t) = \frac{\lambda_i(t)}{\int_0^c \lambda_i(t) dt}.$$ 

By this theorem, if the intensities of the two Poisson processes ($\lambda_1(t)$ and $\lambda_2(t)$) are known, the probability density functions $f_1(t)$ and $f_2(t)$ of the occurrence times can be calculated.

**Theorem.** Lehmann (1975). Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent, the $X$’s identically distributed with distribution $F$ and the $Y$’s with distribution $G$. Let

$$U = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \varphi(X_i, Y_j),$$

and assume without loss of generality that

$$m \leq n \text{ and } m/n \to \lambda \text{ as } m \text{ and } n \to \infty,$$
where \( \lambda \) may be zero. Let \( E(\varphi(X,Y)) = \theta \). Then

\[
T = \sqrt{m}(U - \theta)
\]

is asymptotically normal with mean zero and variance

\[
\sigma^2 = \sigma_{10}^2 + \lambda \sigma_{01}^2,
\]

where

\[
\sigma_{10}^2 = \text{Cov}(\varphi(X_i, Y_j), \varphi(X_i, Y_k))
\]

\[
\sigma_{01}^2 = \text{Cov}(\varphi(X_i, Y_j), \varphi(X_k, Y_j)).
\]

By this theorem,

\[
\frac{T_\epsilon}{mn} - E\left(\frac{T_\epsilon}{mn}\right)
\]

is a \( U \)-statistic. Hence, we have \( T_\epsilon \) is asymptotically normal as \( n, m \to \infty \) and \( \frac{m}{n} \to \lambda \):

\[
\sqrt{m}\left(\frac{T_\epsilon}{mn} - p_\epsilon\right) \to N(0, \sigma_{10}^2 + \lambda \sigma_{01}^2).
\]

where

\[
p_\epsilon = P[|X_i - Y_j| < \epsilon],
\]

\[
\sigma_{10}^2 = \text{Cov}(I_{|X_i - Y_j| < \epsilon}, I_{|X_i - Y_l| < \epsilon}),
\]

\[
\sigma_{01}^2 = \text{Cov}(I_{|X_i - Y_j| < \epsilon}, I_{|X_k - Y_j| < \epsilon}).
\]

Hence, if the densities \( f_1(x), f_2(y) \) are estimated, we can compute \( \sigma_{10}^2 \) and \( \sigma_{01}^2 \). Then we can test

\[
\mathbf{H}_0: \text{The two sequences are independent}
\]

versus

\[
\mathbf{H}_1: \text{The two sequences are attracted locally}
\]

by

\[
\frac{\sqrt{m}(T_\epsilon - p_\epsilon)}{\sqrt{\frac{n-1}{n} \sigma_{10}^2 + \frac{m-1}{m} \sigma_{01}^2}} \approx N(0, 1).
\]

Hence, we have the decision rule: reject \( \mathbf{H}_0 \) if \( Z_\epsilon > Z_{1-\alpha} \) at a given \( \alpha \) level.
1.1.3 Summary of the Two Methods

- The first method requires lots of computation including semi-greedy algorithm.

- The second method utilizes $U$-statistics to avoid complex computation. However the convergence rate of density estimation is of the order of $n^{-\frac{2}{5}}$, which is different from the convergence rate of $U$-statistics, thus there exists a systematic bias.

In the next section, another method proposed by Barry James and Kang James is introduced to avoid complicated computation and system bias.
Chapter 2

Method proposed by Barry James and Kang James

This method is similar to the second method, it also utilizes the $U$-statistic. The main strength is it avoids density estimation, thus the rate of convergence problem is solved.

Theorem. Ross (2007). In a homogeneous Poisson process $N(t), t \geq 0$, given that $N(t) = n$, the $n$ arrival times $S_1, S_2, \ldots, S_n$ have the same distribution as the order statistics corresponding to $n$ independent random variables uniformly distributed on the interval $(0,t)$.

In light of this theorem, if we partition the total interval into subintervals of length $\delta$, and if $\delta$ is small enough, we can approximate the non-homogeneous Poisson process by a homogeneous one. Thus, applying the theorem to the $i^{th}$ subinterval, from the left ($t = 0$) to the right ($t = \delta$) endpoint, we have the density of the arrival times

$$f_1(x) = \frac{1}{\delta}, \text{ for } 0 < x < \delta,$$

$$f_2(y) = \frac{1}{\delta}, \text{ for } 0 < y < \delta.$$

Now, let

$$S_{\epsilon_i} = \sum_{k=1}^{m_i} \sum_{l=1}^{n_i} I_{[|X_{ik} - Y_{il}| < \epsilon]}.$$
where $I$ is the indicator function with the property that

$$I_{|X_{ik} - Y_{il}| < \epsilon} = \begin{cases} 
    1 & \text{if } |X_{ik} - Y_{il}| < \epsilon \\
    0 & \text{otherwise}
\end{cases}$$

Notice that the counting function $S_{\epsilon_i}$ involves all the onset times in the $i^{th}$ subinterval. If we let

$$I_{|X_{ik} - Y_{il}| < \epsilon} = \varphi(X_{ik}, Y_{ij}),$$

we can put $S_{\epsilon_i}$ into the framework of U-statistics. By the theorem of U-statistics, we have

$$\sqrt{m_i} \left( \frac{1}{m_i} \sum_{k=1}^{m_i} \sum_{l=1}^{n_i} I_{|X_{ik} - Y_{il}| < \epsilon} - E \left[ I_{|X_{ik} - Y_{il}| < \epsilon} \right] \right)$$

is approximately normally distributed. Thus, $S_{\epsilon_i}$ is also normally distributed.

Next, let’s calculate $\text{Var} \left( S_{\epsilon_i} \right)$ and $p_\epsilon = E \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right)$.

By the variance lemma of U-statistic (see appendix I),

$$\text{Var} \left( S_{\epsilon_i} \right) = m_i n_i \text{Var} \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right) + m_i n_i (m_i - 1) \text{Cov} \left( I_{|X_{ik} - Y_{il}| < \epsilon}, I_{|X_{ik} - Y_{ij}| < \epsilon} \right)$$

$$+ m_i n_i (n_i - 1) \text{Cov} \left( I_{|X_{ik} - Y_{il}| < \epsilon}, I_{|X_{ik} - Y_{ik}| < \epsilon} \right)$$

$$= m_i n_i \text{Var} \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right) + m_i n_i (m_i + n_i - 2) \text{Cov} \left( I_{|X_{ik} - Y_{il}| < \epsilon}, I_{|X_{ik} - Y_{ij}| < \epsilon} \right) \cdot$$

The last equation is true, since $X_{ik}$ and $Y_{il}$ have the same uniform distribution on $[0, \delta]$, so

$$\text{Cov} \left( I_{|X_{ik} - Y_{il}| < \epsilon}, I_{|X_{ik} - Y_{ij}| < \epsilon} \right) = \text{Cov} \left( I_{|X_{ik} - Y_{il}| < \epsilon}, I_{|X_{ij} - Y_{ik}| < \epsilon} \right).$$

We can rewrite $\text{Var} \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right)$ as

$$\text{Var} \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right) = E \left[ \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right)^2 \right] - E \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right)^2$$

$$= 1 \cdot p_\epsilon - p_\epsilon^2$$

To calculate $p_\epsilon$:

$$p_\epsilon = E \left( I_{|X_{ik} - Y_{il}| < \epsilon} \right)$$

$$= 1 \cdot P \left( |X_{ik} - Y_{il}| < \epsilon \right) + 0 \cdot E \left( |X_{ik} - Y_{il}| \geq \epsilon \right)$$

$$= P \left( |X_{ik} - Y_{il}| < \epsilon \right)$$
We know $X_1, X_2, \ldots, X_n \sim UNIF(0, \delta)$, and $Y_1, Y_2, \ldots, Y_m \sim UNIF(0, \delta)$. So, $P(|X_{ik} - Y_{il}| < \epsilon)$ is the area of the shaded part in the Figure 2.1.

The area of the shaded strip is $\delta^2 - (\delta - \epsilon)^2 = 2\delta\epsilon - \epsilon^2$. We have

$$p_\epsilon = P(|X_{ik} - Y_{il}| < \epsilon) = \frac{1}{\delta^2} \cdot (2\delta\epsilon - \epsilon^2) = \frac{\epsilon}{\delta} (2 - \frac{\epsilon}{\delta}).$$

Let us define $\sigma^2_{10} = \text{Cov}(I_{|X_{ik} - Y_{il}|<\epsilon}, I_{|X_{ik} - Y_{ij}|<\epsilon})$. Then,

$$\sigma^2_{10} = E(I_{|X_{ik} - Y_{il}|<\epsilon} \cdot I_{|X_{ik} - Y_{ij}|<\epsilon}) - E(I_{|X_{ik} - Y_{il}|<\epsilon}) \cdot E(I_{|X_{ik} - Y_{ij}|<\epsilon})$$

$$= E(I_{|X_{ik} - Y_{il}|<\epsilon} \cdot I_{|X_{ik} - Y_{ij}|<\epsilon}) - p_\epsilon^2$$

By the total expectation for continuous random variables, if we condition on $X_{ik}$, and set

$$I_i = I_{|X_{ik} - Y_{il}|<\epsilon}$$

$$I_j = I_{|X_{ik} - Y_{ij}|<\epsilon},$$
then,

\[
E (I_{[X_{ik}-Y_{il}]<\epsilon} \cdot I_{[X_{ik}-Y_{ij}]<\epsilon}) = \int_{0}^{\epsilon} \mathbb{E} (I_t \cdot I_j \mid x_{ik} = t) \cdot f_{X_{ik}} dt + \int_{\epsilon}^{\delta - \epsilon} \mathbb{E} (I_t \cdot I_j \mid x_{ik} = t) \cdot f_{X_{ik}} dt + \int_{\delta - \epsilon}^{\delta} \mathbb{E} (I_t \cdot I_j \mid x_{ik} = t) \cdot f_{X_{ik}} dt
\]

call it A  

call it B  

call it C

Once \(X_{ik}\) is given, \(I_t\) and \(I_j\) are independent. So, we can calculate \(A\), \(B\), and \(C\) using the fact that \(\mathbb{E} (I_t \cdot I_j) = \mathbb{E} (I_t) \cdot \mathbb{E} (I_j)\).

\[
A = \int_{0}^{\epsilon} \mathbb{E} (I_t \cdot I_j \mid x_{ik} = t) \cdot f_{X_{ik}} dt
\]

\[
= \int_{0}^{\epsilon} \mathbb{E} (I_t \mid x_{ik} = t) \cdot \mathbb{E} (I_j \mid x_{ik} = t) \cdot \frac{1}{\delta} dt
\]

\[
= \frac{1}{\delta} \int_{0}^{\epsilon} \left( \frac{t + \epsilon}{\delta} \right)^2 dt
\]

\[
= \frac{7\epsilon^3}{3\delta^3}
\]

\[
B = \int_{0}^{\epsilon} \mathbb{E} (I_t \cdot I_j \mid x_{ik} = t) \cdot f_{X_{ik}} dt
\]

\[
= \int_{\epsilon}^{\delta - \epsilon} \mathbb{E} (I_t \mid x_{ik} = t) \cdot \mathbb{E} (I_j \mid x_{ik} = t) \cdot \frac{1}{\delta} dt
\]

\[
= \int_{\epsilon}^{\delta - \epsilon} \frac{2\epsilon}{\delta} \frac{2\epsilon}{\delta} \frac{1}{\delta} dt
\]

\[
= \left( \frac{2\epsilon}{\delta} \right)^2 \left( \frac{\delta - 2\epsilon}{\delta} \right)
\]

\[
C = \int_{0}^{\epsilon} \mathbb{E} (I_t \cdot I_j \mid x_{ik} = t) \cdot f_{X_{ik}} dt
\]

\[
= \int_{\delta - \epsilon}^{\delta} \mathbb{E} (I_t \mid x_{ik} = t) \cdot \mathbb{E} (I_j \mid x_{ik} = t) \cdot \frac{1}{\delta} dt
\]

\[
= \int_{\delta - \epsilon}^{\delta} \left( \frac{\delta - t + \epsilon}{\delta} \right)^2 \cdot \frac{1}{\delta} dt
\]

\[
= \frac{1}{\delta^3} \cdot \left[ (\delta + \epsilon)^2 \cdot \epsilon + \frac{1}{3} \cdot (\delta^3 - (\delta + \epsilon)^3) - (\delta + \epsilon) \cdot (\delta^2 - (\delta - \epsilon)^2) \right]
\]

Finally, let us combine \(A\), \(B\), and \(C\) to calculate \(\mathbb{E} (I_{[X_{ik}-Y_{il}]<\epsilon} \cdot I_{[X_{ik}-Y_{ij}]<\epsilon})\), and simplify the expression. We have

\[
\sigma_{10}^2 = \frac{2}{3} \cdot \left( \frac{\epsilon}{\delta} \right)^3 - \left( \frac{\epsilon}{\delta} \right)^4
\]
Hence, we have

\[ \text{Var} \left( S_{\epsilon_i} \right) = m_i \cdot n_i \cdot p_\epsilon (1 - p_\epsilon) + m_i \cdot n_i \cdot (m_i + n_i - 2) \cdot \sigma^2_{10} \]
\[ = m_i n_i \frac{\epsilon}{\delta} (2 - \frac{\epsilon}{\delta}) \left[ 1 - \frac{\epsilon}{\delta} (2 - \frac{\epsilon}{\delta}) \right] + m_i n_i (m_i + n_i - 2) \left[ \frac{2}{3} \left( \frac{\epsilon}{\delta} \right)^3 - \left( \frac{\epsilon}{\delta} \right)^4 \right] \]

Now, let us define \( S_\epsilon = \sum_{i=1}^{K} S_{\epsilon_i} \), where \( K \) is the total number of subintervals. So, \( E(S_\epsilon) = p_\epsilon \sum m_i \cdot n_i \), and \( \text{Var}(S_\epsilon) = \sum \text{Var}(S_{\epsilon_i}) \) By the generalized central limit theorem, we propose the test statistic as follows:

\[ Z_\epsilon = \frac{S_\epsilon - E(S_\epsilon)}{\sqrt{\text{Var}(S_\epsilon)}} = \frac{\sum_{i=1}^{K} S_{\epsilon_i} - P_\epsilon \sum_{i=1}^{K} m_i \cdot n_i}{\sqrt{\sum_{i=1}^{K} \text{Var}(S_{\epsilon_i})}} \]

To test:

\[ H_0 : \text{The two sequences are independent} \]
\[ \text{versus} \quad H_1 : \text{The two sequences are attracted in a local sense,} \]

\( Z_\epsilon \) is approximately normal distributed with \( \mu = 0, \sigma^2 = 1 \) under the null hypothesis. Hence, we have the decision rule: reject \( H_0 \) if \( Z_\epsilon > Z_{1-\alpha} \) at a given \( \alpha \) level. We reject the null hypothesis when \( Z_\epsilon \) is large, since the attraction of the two sequences implies more pairs of events appear within \( \epsilon \) of each other.
Chapter 3

Simulation

Recall from the last section, we have put the counting function

\[ S_{\epsilon_i} = \sum_{k=1}^{m_i} \sum_{l=1}^{n_i} I[|X_{ik} - Y_{il}| < \epsilon] \]

into the framework of $U$-statistics. The procedure of calculating the test statistic $Z_{\epsilon}$ is following:

1. Partition the whole time interval $T$ into subintervals of length $\delta$, where $\delta < T$
2. Inside the $i^{th}$ subinterval, choose $\epsilon < \delta$, and apply the counting function $S_{\epsilon_i}$ to the onset time of the two events.
3. After calculating $S_{\epsilon_i}$ for all the subintervals, we can compute the test statistic

\[ Z_{\epsilon} = \frac{S_{\epsilon} - \mu(S_{\epsilon})}{\sqrt{\text{Var}(S_{\epsilon})}} = \frac{\sum_{i=1}^{K} S_{\epsilon_i} - P_{\epsilon} \sum_{i=1}^{K} m_i \cdot n_i}{\sqrt{\sum_{i=1}^{K} \text{Var}(S_{\epsilon_i})}} \]

Now, the question is how should we choose the values of $\delta$ and $\epsilon$ so that we can have the empirically optimal approximation to the standard normal? This question is of importance since

1. If delta is too large, then we cannot use a homogeneous Poisson to approximate the processes.
2. If delta is too small, we may not have enough sample points within each subinterval. To apply the result on $U$-statistics, we need $m_i, n_i \to \infty$. 

12
Hence, there is always a trade-off between large $\delta$ and small $\delta$. Large $\delta$ guarantees that inside each subinterval, we have enough data points for the U-statistic. On the other hand, the small delta makes sure that we can replace a non-homogeneous Poisson by a homogeneous one.

### 3.1 An Ideal Case: Two Uniform Densities

In this subsection, the simplest case is presented, i.e., two random sequences are generated from the uniform distribution. This is the ideal case, and any choice of the length of subinterval $\delta$ will achieve a good approximation to the uniform distribution, since the two sequences are uniformly distributed on the interval $[0, T]$. We use this function to test the approximation result of $Z_\epsilon$. Now, let us iterate 3,000 times to see the empirical distribution of $Z_\epsilon$ statistic. Table 3.1 summarizes the parameters of the simulation.

<table>
<thead>
<tr>
<th>time value</th>
<th>$\delta$</th>
<th>$\epsilon$</th>
<th>$m$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>5</td>
<td>0.5</td>
<td>2,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>

Table 3.1: parameters of the simulation

Table 3.2 summarizes the important statistics of the simulation, where $W$-value is the test statistics of the Shapiro-Wilk test for normality. Figure 3.1 shows the histogram and $Q-Q$ plot of empirical $Z_\epsilon$ statistic. We can see that the simulation result indicates that $z$ approximates standard normal very well. The mean and variance are close to 0 and 1, respectively. The Shapiro-Wilk test does not reject null hypothesis that sample is from normal population at $\alpha = 0.01$.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>$W$-Value</th>
<th>$p$-value of Shapiro-Wilk test</th>
</tr>
</thead>
<tbody>
<tr>
<td>two uniforms</td>
<td>-0.0031</td>
<td>0.9992</td>
<td>0.9996</td>
<td>0.854</td>
</tr>
</tbody>
</table>

Table 3.2: Summary statistics
3.2 Relax the Restriction

Since the previous simulation is under the ideal condition, we should relax the two uniform distribution restrictions for application in the real world. To test how well the approximation is when the uniform condition is violated, let us consider an extreme case, where we replace one of the uniform distributions by a linear density. Figure 3.2 contains a histogram of the linear density, a histogram of the $Z_\epsilon$, and the $Q - Q$ plot of the $Z_\epsilon$-statistics after 3,000 simulations.

From the summary statistics, we can see that the mean and variance are very close to 0 and 1. However, the Shapiro-Wilk test rejects the normality. So, the simulation
result is not as good as the previous case. Next, Let us try some method to improve the normality.

The possible reasons to explain it:

1. The triangle density has extreme behavior in the beginning, i.e., there are very few data points in the first several subintervals. Hence, the assumption of applying central limit theorem is violated; the sample size \( n \) is not large enough to guarantee the convergence.

2. Since the triangle density varies dramatically, the approximation by a homogeneous Poisson process maybe not be good.

How can we improve the normality? One possible solution to solve above problem is to control the minimum number of data points in each subinterval, so that the assumption of a large sample would not be violated.

### 3.3 \( \ell \) with Minimum Number of Data Points in Each Subinterval

As we have seen in the last section, the violation of normality may come from lack of enough data points in each subinterval. A straightforward method to fix that problem is to specify the minimum number of data points in each subinterval for these two processes.

The simulation will use one uniform density and one triangular density. The first graph in Figure 3.3 displays the histogram of the triangular density, which is much more mild than the linear density. Table 3.4 summarizes the important statistics, the mean=\(-0.81\) is far way from 0, and variance=0.939 is also not satisfying.

There is a systematic bias to the left after other simulations, hence the minimum data points method apparently can not improve the approximation results.
In this section, the same triangular density as in last section (see Figure-5) and uniform density were used in the simulation. Table 3.5 summarizes the arguments of the simulation. The histograms, means, and variances of $Z_\epsilon$ statistic under different $\delta$ values are compared. Each simulation contains 3,000 sample points.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>$\epsilon$</th>
<th>$m$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>varies 0.5</td>
<td>2,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>

Table 3.5: parameters of the simulation

Table 3.6 summarizes the means, variances, and $p$-values of the Shapiro-Wilk normality test for different $\delta$ values.
Table 3.6: summary statistics

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>Mean</th>
<th>Variance</th>
<th>W-Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-0.0037</td>
<td>1.002</td>
<td>0.9992</td>
<td>0.598</td>
</tr>
<tr>
<td>5</td>
<td>-0.0363</td>
<td>0.9961</td>
<td>0.9993</td>
<td>0.743</td>
</tr>
<tr>
<td>10</td>
<td>-0.0225</td>
<td>1.02</td>
<td>0.9995</td>
<td>0.071</td>
</tr>
<tr>
<td>15</td>
<td>0.0120</td>
<td>0.9824</td>
<td>0.9995</td>
<td>0.928</td>
</tr>
<tr>
<td>20</td>
<td>0.0095</td>
<td>1.0404</td>
<td>0.9990</td>
<td>0.347</td>
</tr>
<tr>
<td>30</td>
<td>0.0018</td>
<td>0.9853</td>
<td>0.9996</td>
<td>0.259</td>
</tr>
<tr>
<td>50</td>
<td>-0.0172</td>
<td>1.0426</td>
<td>0.9993</td>
<td>0.7304</td>
</tr>
</tbody>
</table>

From the simulation result, the different values of $\delta$ achieve a reasonably good approximation in terms of means, variances, and normality tests. However, the smaller $\delta$ seems to give us comparatively better results. In this case, the optimal $\delta$ should make the number of subintervals more than 200. A large number of subintervals can apparently improve the normality in terms of the variance.

### 3.5 How to choose $\epsilon$?

The problem of finding an empirical optimal $\epsilon$ is equivalent to finding the ratio $\frac{\epsilon}{\delta}$. Hence, the optimal ratio $r$ is studied. The method is similar to the one of finding optimal $\delta$. Here the same densities as in the previous simulation are used with $\delta = 5$. The simulation strategy is similar to the previous case; seven different values of $r$ are used in the simulations. Then, the means, variances, and $p$-values of the Shapiro-Wilk normality test are compared to decide the optimal $r$ value. Table 3.7 summarizes the simulation parameters.

<table>
<thead>
<tr>
<th>time value</th>
<th>$\delta$</th>
<th>$\epsilon$</th>
<th>$m$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,000</td>
<td>5</td>
<td>varies</td>
<td>2,000</td>
<td>2,000</td>
</tr>
</tbody>
</table>

Table 3.7: parameters of the simulation

The simulation contains $r = 0.01, r = 0.05, r = 0.10, r = 0.15, r = 0.20, r = 0.30,$ and $r = 0.50$. Table 3.8 summarizes the means, variances, and $p$-values of the normality
Again, we can see that all the choices of \( r \) give a reasonably good approximation in terms of mean, variance, and the normality test. We can see that in this case \( r = 0.1 \) gives us a comparatively better approximation.

### 3.6 Method of Using Weights

In each subinterval, the number of events is different. Some subintervals contain more data points, hence has a better approximation. Some intervals with fewer data points may have a problem of convergence. Hence, we may put more weight on the subinterval with more data points. We can choose weight simply as
\[
w_i = \frac{m_i + n_i}{M + N}.
\]

We still use the same triangular density and uniform density as in the last section. Figure 3.4 shows the histogram and Q-Q plot of the \( Z_e \) statistic using the weighted method.

Table 3.9 summarize the mean, variance, and \( p \)-value of the normality test of the simulation.

The result is very good. Let us compare it with the method without weights. For the same sequence, let us apply the fixed \( \delta \) method without weights. Figure 3.5 shows the histogram and Q-Q plot of the \( Z_e \) statistic using the unweighted method.

Table 3.10 compares the means, the variances, and the \( p \)-values of the normality test for the weighted method and the unweighted method, respectively.

<table>
<thead>
<tr>
<th>( r )</th>
<th>Mean</th>
<th>Variance</th>
<th>( W )-Value</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>.01</td>
<td>-0.0142</td>
<td>0.9861</td>
<td>0.9994</td>
<td>0.4157</td>
</tr>
<tr>
<td>.05</td>
<td>-0.0250</td>
<td>0.9806</td>
<td>0.9996</td>
<td>0.7864</td>
</tr>
<tr>
<td>.10</td>
<td>-0.0014</td>
<td>1.0076</td>
<td>0.9992</td>
<td>0.2368</td>
</tr>
<tr>
<td>.15</td>
<td>-0.0172</td>
<td>0.9901</td>
<td>0.9993</td>
<td>0.3181</td>
</tr>
<tr>
<td>.20</td>
<td>-0.0178</td>
<td>1.0266</td>
<td>0.9996</td>
<td>0.87</td>
</tr>
<tr>
<td>.30</td>
<td>0.00467</td>
<td>0.9837</td>
<td>0.9996</td>
<td>0.259</td>
</tr>
<tr>
<td>.50</td>
<td>-0.0075</td>
<td>0.9672</td>
<td>0.9995</td>
<td>0.6659</td>
</tr>
</tbody>
</table>

Table 3.8: Summary statistics
From the last several simulations, the approximation is not very sensitive to the choice of $\delta$ if the sample size is large (more than 10,000 data points for each process).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Mean  & Variance & $W$-Value & $p$-value \\
\hline
Zweight & -0.0024 & 1.031 & 0.9994 & 0.4757 \\
\hline
\end{tabular}
\caption{Summary Statistics: with weights}
\end{table}

From the above comparison, both methods work well. The method with weights can apparently improve the mean and variance slightly. Hence, the weighted method perhaps should be considered, although it is more complicated.

### 3.7 More Simulations

From the last several simulations, the approximation is not very sensitive to the choice of $\delta$ and $r$ if the sample size is large (more than 10,000 data points for each process). The next simulation uses a triangular density and a Beta(2,5) density with number of data points $m = n = 20,000$. Figure 3.6 is the histogram of the densities.

The first simulation fixes $r = 0.1$ and lets $\delta$ be 5 to 54 with increments of 1. For each $\delta$ value, 4,000 $Z_\ell$’s are calculated. The $p$-value of the normality tests all exceed 0.05 except for $\delta = 24$, which indicates that normality is not a problem for large sample size. Figure 3.7 displays the means and variances under different $\delta$ values. The means decrease as $\delta$ increases, and the variances increase as $\delta$ increases.

From the graph, we can see that the choices of $\delta$ from 5 to 12 achieve a good approximation to the standard normal.
Figure 3.5: Histogram and Q-Q-Plot of \( Z_\epsilon \) without weights

<table>
<thead>
<tr>
<th>Method</th>
<th>Mean</th>
<th>Variance</th>
<th>( W )-Value</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zweight</td>
<td>-0.0024</td>
<td>1.031</td>
<td>0.9994</td>
<td>0.4757</td>
</tr>
<tr>
<td>Zfix</td>
<td>-0.0081</td>
<td>1.031</td>
<td>0.9995</td>
<td>0.6861</td>
</tr>
</tbody>
</table>

Table 3.10: Summary Statistics: comparison of two methods

Figure 3.6: Histograms of Densities
Figure 3.7: Means and variances
From the last simulation, let us fix $\delta = 7$ and let $r$ be 0.01 to 0.49 with increments of 0.02. Table 3.11 summarizes the means, variances, and $p$-values of $Z_\epsilon$ under different $r$ values. From Table 3.11 all choices of $r$ achieve a comparatively good approximation to the standard normal except $r = 0.23$, $r = 0.33$, and $r = 0.41$. Hence, the approximation is not very sensitive to the choice of $r$ for $\delta = 7$.

<table>
<thead>
<tr>
<th>$r$</th>
<th>Mean</th>
<th>Variance</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.0365045</td>
<td>0.9936027</td>
<td>0.7864272</td>
</tr>
<tr>
<td>0.03</td>
<td>-0.0169537</td>
<td>0.973802</td>
<td>0.2973941</td>
</tr>
<tr>
<td>0.05</td>
<td>-0.0104338</td>
<td>0.9746030</td>
<td>0.0262203</td>
</tr>
<tr>
<td>0.07</td>
<td>0.0154001</td>
<td>0.9632385</td>
<td>0.1268875</td>
</tr>
<tr>
<td>0.09</td>
<td>0.0021763</td>
<td>0.9950009</td>
<td>0.0601502</td>
</tr>
<tr>
<td>0.11</td>
<td>0.0046223</td>
<td>1.0328550</td>
<td>0.3440567</td>
</tr>
<tr>
<td>0.13</td>
<td>0.0149593</td>
<td>1.0384356</td>
<td>0.8660473</td>
</tr>
<tr>
<td>0.15</td>
<td>-0.0235090</td>
<td>0.9704262</td>
<td>0.4878831</td>
</tr>
<tr>
<td>0.17</td>
<td>0.0121975</td>
<td>1.0224707</td>
<td>0.5874498</td>
</tr>
<tr>
<td>0.19</td>
<td>0.0025625</td>
<td>0.9792202</td>
<td>0.7716585</td>
</tr>
<tr>
<td>0.21</td>
<td>0.0018382</td>
<td>0.9847646</td>
<td>0.6201994</td>
</tr>
<tr>
<td>0.23</td>
<td>0.0040124</td>
<td>0.9898493</td>
<td>0.0361313</td>
</tr>
<tr>
<td>0.25</td>
<td>0.0027337</td>
<td>1.0388430</td>
<td>0.4153030</td>
</tr>
<tr>
<td>0.27</td>
<td>-0.0138465</td>
<td>0.9986625</td>
<td>0.3224556</td>
</tr>
<tr>
<td>0.29</td>
<td>0.0050482</td>
<td>0.9929240</td>
<td>0.7326557</td>
</tr>
<tr>
<td>0.31</td>
<td>-0.0050682</td>
<td>0.9818393</td>
<td>0.8686029</td>
</tr>
<tr>
<td>0.33</td>
<td>0.0255269</td>
<td>1.0161704</td>
<td>0.0145982</td>
</tr>
<tr>
<td>0.35</td>
<td>0.0233395</td>
<td>0.9925798</td>
<td>0.6554060</td>
</tr>
<tr>
<td>0.37</td>
<td>0.0041284</td>
<td>0.9575313</td>
<td>0.9062076</td>
</tr>
<tr>
<td>0.39</td>
<td>-0.0017839</td>
<td>0.9552272</td>
<td>0.5028486</td>
</tr>
<tr>
<td>0.41</td>
<td>0.0072140</td>
<td>0.9994010</td>
<td>0.0459737</td>
</tr>
<tr>
<td>0.43</td>
<td>0.0233608</td>
<td>1.0421065</td>
<td>0.4254984</td>
</tr>
<tr>
<td>0.45</td>
<td>-0.0116703</td>
<td>0.9689418</td>
<td>0.8641576</td>
</tr>
<tr>
<td>0.47</td>
<td>-0.0265845</td>
<td>1.0251019</td>
<td>0.8496497</td>
</tr>
<tr>
<td>0.49</td>
<td>0.0006285</td>
<td>0.9920500</td>
<td>0.8674392</td>
</tr>
</tbody>
</table>

Table 3.11: Means, variances, and $p$-values under different $r$

Let us do one more simulation to check how good the approximation is for $\delta = 5$ and $r = 0.1$. Let us consider a more extreme case and let the two sequences be generated from the two different triangular densities; Figure 3.8 displays the histograms of these
two densities. Table 3.12 summarizes the parameters of the simulation.

Figure 3.8: Two Triangular Densities

<table>
<thead>
<tr>
<th>time</th>
<th>δ</th>
<th>ϵ</th>
<th>m</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>1,000</td>
<td>5</td>
<td>varies 3,000</td>
<td>4,000</td>
</tr>
</tbody>
</table>

Table 3.12: parameters of the simulation
Table 3.13 summarizes the means, variances, and $p$-values of the simulation using the weighted method and the unweighted method. Both these methods give an excellent approximation to the standard normal. The two methods almost produce the identical results. The weighted method is slightly better in terms of the mean.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Variance</th>
<th>W-Value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zweight</td>
<td>-0.0068</td>
<td>0.9949</td>
<td>0.9996</td>
<td>0.8988</td>
</tr>
<tr>
<td>Zfix</td>
<td>-0.0099</td>
<td>0.9910</td>
<td>0.9995</td>
<td>0.6338</td>
</tr>
</tbody>
</table>

Table 3.13: Summary Statistics of $Z_c$

### 3.8 Summary for the Large Sample Case

Based on the simulation results, we can conclude that

- The proper $\delta$ is the one such that the number of subintervals is more than 200.
- The approximation is not very sensitive to the $r$ value
- The unweighted method and the weighted method would produce similar results. The second one is slightly better.

### 3.9 Application to the Small Sample Size

From the previous discussion, the weighted method works well for large sample size (where each process contains 20,000 data points). The approximation is not very sensitive to the choice of $\delta$ and $\epsilon$, since there are enough data points to guarantee the convergence of both $U$-statistic and central limit theory. However, it is not clear whether this method still can produce a good approximation to standard normal if the sample size is small with each process containing less than 500 data points.

#### 3.9.1 Choose appropriate $\delta$ value when sample size is small

In this subsection, simulation is applied to the processes with small sample size. The simulation strategy is

- Take fixed $r = 0.1$, simulate 4,000 $z$ value using $\delta$ from 5 to 100.
- Compare the mean, variance, and $p$-value of the Shapiro-Wilk normality test.
• Choose a range of $\delta$ that produces a comparative good approximation to the standard normal.

Let the two densities be the linear density and a uniform density on the interval $[0, 1000]$ and let the number of data points be $m = 300, n = 250$. Let us examine the $p$-value of the normality test first. Table 3.14 lists all the $\delta$ values that make the $p$-value less than 0.05. From $\delta = 33$ to 57, only two values ($\delta=41, 44$) produce $z$ values that have $p$-value $<0.05$.

<table>
<thead>
<tr>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>14</th>
<th>17</th>
<th>18</th>
<th>23</th>
<th>26</th>
<th>27</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>44</td>
<td>58</td>
<td>60</td>
<td>63</td>
<td>68</td>
<td>70</td>
<td>72</td>
<td>74</td>
<td>77</td>
<td>78</td>
</tr>
<tr>
<td>79</td>
<td>80</td>
<td>81</td>
<td>85</td>
<td>92</td>
<td>93</td>
<td>94</td>
<td>97</td>
<td>99</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.14: $\delta$ value corresponding to $p$-value $<0.05$

Next, let us examine the means and variances of the $z$ generated by $\delta =33$ to 57. Figure 3.9 shows that the means and variances also achieve a better approximation for $\delta =33$ to 57. The points inside the red box are consistently closer to 0 and 1 for means and variances, respectively.
So, we can see that the ideal choice of $\delta$ is from 33 to 57.

Let us do another simulation to check whether such a choice is sensitive to different densities. The next simulation uses a triangular density and a Beta(2,5) density with number of data points $m = 300, n = 250$. Figure 3.10 is the histogram of the densities.
Figure 3.10: Histograms of Densities

We still simulate 4,000 $Z_\epsilon$ values using $\delta$ from 5 to 100 and $r = 0.1$. Table 3.15 summarizes the $\delta$ values that have $p$-value $< 0.05$.

<table>
<thead>
<tr>
<th>$\delta$</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>9</th>
<th>15</th>
<th>27</th>
<th>31</th>
<th>32</th>
<th>34</th>
<th>40</th>
<th>42</th>
<th>44</th>
<th>47</th>
<th>56</th>
<th>59</th>
<th>60</th>
<th>62</th>
<th>65</th>
<th>67</th>
<th>68</th>
<th>70</th>
<th>71</th>
<th>76</th>
<th>84</th>
<th>88</th>
<th>94</th>
<th>96</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.15: $\delta$ value corresponding to $p$-value $< 0.05$

From Table 3.15, for $\delta = 10$ to 30, only two values ($\delta = 15, 27$) produce $Z_\epsilon$ values that have $p$-value $< 0.05$.

Similar to the previous simulation, we then check the means and variances of the $z$ values. Figure 3.11 displays the means and variances; there exist some obvious trends for both means and variances. The means decrease from 0 as $\delta$ increases, the variances increase from 1 to 1.3 as $\delta$ increases. So, the most appropriate $\delta$ value is less than 40 based on Figure 3.11. The red lines highlight the means and variances corresponding
to $\delta$ from 10 to 30, which are picked based on $p$-values.

So the choice of $\delta$ is very sensitive to the different densities when the sample size is small.

**3.9.2 Choose Appropriate $r$ Value When Sample Size is Small**

The procedure of choosing the appropriate $r$ is:

- For different densities, fix one appropriate $\delta$ based last subsection, simulate 4,000 $Z_\epsilon$ values using $r$ from 0.01 to 0.5 with increment 0.01. In the process of calculating the variance of $z$, we assume $\delta - \epsilon > \epsilon$, so we have $\frac{\epsilon}{\delta} < 0.5$.

- Compare mean, variance, and $p$-value of Shapiro-Wilk normality test.

- Choose a range of $r$ that produces a comparatively good approximation to the standard normal.
The first simulation uses the triangular density and Beta(2,5) density. From the last section, the appropriate value of $\delta$ is from 10 to 30. Hence, let $\delta = 20$ for the simulation. Table 3.16 lists all the $r$ values that have $p$-value less than 0.05.

| 0.01 | 0.02 | 0.03 | 0.04 | 0.09 | 0.11 | 0.15 |
| 0.17 | 0.18 | 0.19 | 0.21 | 0.22 | 0.25 | 0.26 |
| 0.29 | 0.31 | 0.34 | 0.37 | 0.39 | 0.43 | 0.50 |

Table 3.16: $r$ value corresponding to $p$-value<0.05
Figure 3.12 shows the means and variances for $r = 0.01$ to 0.5. Based on $p$-value, means and variances, the most appropriate $r$ is around 0.4 to 0.5.

The next natural question would be after choosing the optimal $r$ value, can we redo the previous simulation to choose a $\delta$ value that optimizes the approximation based on the current $r$ value? Hence, another simulation is done for fixed $r = 0.41$ and $\delta = 5$ to 100 as in the previous simulation.

Figure 3.13 lists the means and variances for $\delta = 5$ to 100 with $r = 0.42$. Table 3.17 lists all the $\delta$ values that correspond to $p$-value less than 0.05. The number of such $\delta$ is significantly less than the case when $r = 0.1$. The means and variances are also closer to 0 and 1 for all choice of $\delta$. Hence, the approximation is significantly improved for $r = 0.42$ compared to $r = 0.1$.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>20</th>
<th>38</th>
<th>57</th>
<th>62</th>
<th>68</th>
<th>74</th>
<th>76</th>
</tr>
</thead>
</table>

Table 3.17: $\delta$ value corresponding to $p$-value < 0.05
3.9.3 Conclusion for the Small Sample Case

The choice of optimal $r$ and $\delta$ is more complicated than the large sample case. The main differences are:

- The approximation is not as good as the large sample case, which is expected, since we need a large number of data points to guarantee the convergence.

- The approximation is very sensitive to the choice of $\delta$ and $r$, especially the normality.

- Different underlying densities directly affect the choice of $\delta$ and $r$. 

Figure 3.13: Means and Variances for $\delta = 5$ to 100
Chapter 4

Applications

In the chapter, two correlated sequences are artificially generated to test $Z_\epsilon$. The simulation strategy is the following:

1. Generate a sequence from a certain density.
2. Randomly sample half of the data points from 1st sequence.
3. Generate a random error term from standard normal distribution.
4. Add the random error term to the sequence from step 2 to form the second sequence.

The first sequence is generated from Beta(2,5) with 4,000 data points. Then, a normal perturbation with mean 0 and variance 1 is added to half of the points from the first sequence.

Choose $\delta = 5$, $r = 0.1$. We got $Z_\epsilon = 2.196946$. $P(Z > 2.196946) = 0.01401215$, so we reject the null hypothesis. These two random sequences are not independent.

Next, let us look at the case when sample size is small. Let the first sequence be Beta(2,5) with 400 data points. The second the sequence is generated from the first process with 200 data points. Choose $\delta = 5$, $r = 0.1$. We got $Z_\epsilon = 2.687845$. $P(Z > 2.687845) = 0.003595738$, so we reject the null hypothesis. These two random sequences are not independent.
References


Appendix A

U-statistics

Let $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$ be independent, and $X$’s identically distributed with distribution $F$ and the $Y$’s with distribution $G$.

$$U = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \varphi(X_i, Y_j)$$

Hence, $\varphi(X_i, Y_j)$ and $\varphi(X_i, Y_l)$ are dependent. So, we cannot apply central limit theorem. So, define $S = \sum a_i(X_i) + \sum b_j(Y_j)$, which can be applied to CLT and $S$ is asymptotically equivalent to $U$. Let $Z_1, Z_2, ..., Z_N$ be independently distributed, and $Z_i \sim F_i$ and $T = T(Z_1, Z_2, ..., Z_N)$ such that $E(T) = 0$. First, prove an important identity

$$E(T - S)^2 = E(T - T^*)^2 + E(T^* - S)^2$$

where $T^* = \sum_{i=1}^{N} r_i(z_i)$, and $r_i(z_i) = E(T | Z_i = z_i)$

**Proof.** $E(T - T^* + T^* - S)^2 = E(T - T^*)^2 + E(T^* - S)^2 + 2E(T - T^*)(T^* - S)$, so it is enough to show that the cross term equals to zero.

$$(T - T^*)(T^* - S) = (T - T^*)(\sum_{i=1}^{N} r_i(Z_i) - \sum_{i=1}^{N} k_i(Z_i))$$

$$= \sum_{i=1}^{N} (r_i(Z_i) - k_i(Z_i))(T - T^*)$$

34
Hence,
\[
E ((T - T^*)(T^* - S)) = \sum_{i=1}^{N} E [(r_i(Z_i) - k_i(Z_i))(T - T^*)] \text{ (by linearity of expectation)}
\]
\[
= \sum_{i=1}^{N} E [E ((r_i(Z_i) - k_i(Z_i))(T - T^*) | Z_i)] \text{ (by total expectation)}
\]

Now, we look at the single term of above sum and show that each of such term equal to zero.
\[
E ((r_i(Z_i) - k_i(Z_i))(T - T^*)| Z_i)) = (r_i(Z_i) - k_i(Z_i))E (T - T^* | Z_i)
\]

\[
E (T - T^*| Z_i) = E (T - r_i(Z_i) - \sum_{j \neq i} r_j(Z_j)| Z_i)
\]
\[
= E [(T - r_i(Z_i)) | Z_i] - E (\sum_{j \neq i} r_j(Z_j))
\]
\[
= E [T - E (T | Z_i) | Z_i] - E (\sum_{j \neq i} r_j(Z_j)) \quad \text{def of } r_i(Z_i)
\]
\[
= E (\sum_{j \neq i} r_j(Z_j)) = \sum_{j \neq i} E (r_j(Z_j)) = \sum_{j \neq i} E [E (T | Z_j)] = \sum_{j \neq i} E (T)
\]
\[
= 0
\]

Thus, \( S = T^* = \sum_{i=1}^{N} r_i(z_i) = \sum_{i=1}^{N} E(T|Z_i = z_i) \) minimize \( E(T - S)^2 \). \( \Box \)

Now, we apply above identity to the U-statistic. First, define \( E(U) = E (\varphi(X,Y)) = \theta \), and define \( \psi(X,Y) = \varphi(X,Y) - \theta \). If we let

\[
T = T(X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n) = \sqrt{m}(U - \theta)
\]
\[
= \sqrt{m}(\frac{1}{mn} \sum_{\alpha=1}^{m} \sum_{\beta=1}^{n} \phi(X_\alpha, Y_\beta) - \theta) = \sqrt{m} \sum_{\alpha=1}^{m} \sum_{\beta=1}^{n} \psi(X_\alpha, Y_\beta)
\]

It is easy to see that \( E(T) = 0 \). Hence, the "projection" of \( T \) is

\[
\sum_{i=1}^{mn} E(T|Z_i) = \sum_{\alpha=1}^{m} E(T | X_\alpha) + \sum_{\beta=1}^{n} E(T|Y_\beta)
\]
\[ E(\psi(X_\alpha, Y_\beta) \mid X_i = x_i) = \begin{cases} E(\psi(x_i, Y_\beta) & \text{if } \alpha = i \\ 0 & \text{if } \alpha \neq i \end{cases} \]

\[ E(\psi(X_\alpha, Y_\beta) \mid Y_j = y_j) = \begin{cases} E(\psi(X_\alpha, y_j) & \text{if } \beta = j \\ 0 & \text{if } \beta \neq j \end{cases} \]

Thus

\[ E(T \mid X_i = x_i) = E \left( \sum_{\alpha=1}^{m} \sum_{\beta=1}^{n} \frac{\sqrt{m}}{mn} \psi(X_\alpha, Y_\beta) \mid X_i = x_i \right) \]

\[ = \frac{\sqrt{m}}{mn} \sum_{\beta=1}^{n} \sum_{\alpha=1}^{m} E(\psi(X_\alpha, Y_\beta) \mid X_i = x_i) \]

\[ = \frac{\sqrt{m}}{mn} \sum_{\beta=1}^{n} E(\psi(x_i, Y_\beta)) \]

Similarly,

\[ E(T \mid Y_j = y_j) = \frac{\sqrt{m}}{mn} \sum_{\alpha=1}^{m} E(\psi(X_\alpha, y_j)) \]

Putting

\[ \psi_{10}(x) = E[\psi(x, Y)] \text{ and } \psi_{01}(y) = E[\psi(X, y)] \]

The desired projection is therefore

\[ \frac{\sqrt{m}}{mn} [n \sum_{i=1}^{m} \psi_{10}(X_i) + m \sum_{j=1}^{n} \psi_{01}(Y_j)] = \sqrt{m} \left[ \frac{1}{m} \sum_{i=1}^{m} \psi_{10}(X_i) + \frac{1}{n} \sum_{j=1}^{n} \psi_{01}(Y_j) \right] \]

The last expression is true, since \( X' \)'s and \( Y' \)'s are identically distributed, respectively. In summary, we have \( E(T - T^* + T^* - S)^2 = E(T - T^*)^2 + E(T^* - S)^2 \) and \( S = \sum a_i(X_i + \sum b_j(Y_j). \) Notice that \( E(T) = E(T^*) = 0, \) hence, if we let \( S = 0, \) then we get a useful special form

\[ E(T - T^*)^2 = E(T^2) - E(T^*^2) = \text{Var}(T) - \text{Var}(T^*) \]

**Theorem.** If \( T_n \) tends to \( T \) in distribution, and if \( B_n \) and \( D_n \) are r.v.s converging in probability to constant \( b \) and \( d, \) respectively. Then

\[ B_nT_n + D_n \to bT + d \text{ in law} \]
**Corollary.** If $T_n$ tends in distribution to $T$, and if $E(T_n - S_n)^2 \to 0$ as $n \to \infty$, then $S_n$ also tends to $T$ in distribution.

**Proof.** Let $R_n = S_n - R_n$. For any $\epsilon > 0$

$$P (|S_n - T_n| \geq \epsilon) \leq \frac{E(T_n - S_n)^2}{\epsilon^2} \to 0 \text{ as } n \to \infty \text{ by Chebyshev’s inequality.}$$

So, $R_n \to 0$ in probability as $n \to \infty$, then $S_n = T_n + R_n \to T$ in distribution by Slusky’s Theorem.

This theorem shows that if the mean square error of two sequences tends to 0, then these two sequences of random variable converge to the same limit distribution in law. \(\square\)

**Lemma.** Let $X_1...X_m$ and $Y_1...Y_n$ be independent. $X_i \sim F$, and $Y_i \sim G$. $\varphi$ be a function of $X_i's$ and $Y_j's$, Define $W = \sum_{i=1}^{m} \sum_{j=1}^{n} \varphi(X_i, Y_j)$, and $U_{ij} = \varphi(X_i, Y_j)$, then

$$\text{Var}(W) = mn \text{Var}(U_{ij}) + nm(m-1)\text{Cov}(U_{ij}, U_{kj}) + mn(n-1)\text{Cov}(U_{ij}, U_{il})$$

where $i \neq k, j \neq l$

**Proof.**

$$\text{Var}(W) = \sum_{i=1}^{m} \sum_{j=1}^{n} \text{Var}(U_{ij}) + \sum_{i \neq j \text{ or } k \neq l} \text{Cov}(U_{ij}, U_{kl})$$

there are three possibilities:

- if $i \neq k$ and $j \neq l$, then $U_{ij}$ and $U_{kl}$ are independent, which implies $\text{Cov}(U_{ij}, U_{kl}) = 0$

- if $i \neq k$, but $j = l$, there are $m$ choices for $i$, $(m-1)$ choices for $k$, and $n$ choices for $j$ and $l$, which implies there are total $nm(m-1)$ choices for such combination.

- if $j \neq l$, but $i = k$, then there are $m$ choices for $i$ and $k$, $n$ choices for $j$, and $(n-1)$ choices for $l$, which implies there are total $nm(n-1)$ choices for such combination.

\(\square\)
Theorem. Let $U = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \varphi(X_i, Y_j)$, and assume without loss of generality that
\[ m \leq n \text{ and } m/n \to \lambda \text{ as } m \text{ and } n \to \infty \]
where $\lambda$ may be zero. Then
\[ T = \sqrt{m}(U - \theta) \]
is asymptotically normal with mean zero and variance
\[ \sigma^2 = \sigma^2_{10} + \lambda \sigma^2_{01} \]
where
\[ \sigma^2_{10} = \text{Var}[\psi_{10}(X_i)] \text{ and } \sigma^2_{01} = \text{Var}[\psi_{01}(U_j)] \]

Proof. By the lemma proved in the beginning, the projection $T^*$ of $T$ is
\[ T^* = \sqrt{m} \left[ \frac{1}{m} \sum_{i=1}^{m} \psi_{10}(X_i) + \frac{1}{n} \sum_{j=1}^{n} \psi_{01}(Y_j) \right] \]

By the corollary of Slusky’s theorem, we need to show that
\[ E(T - T^*)^2 \to 0 \text{ as } m, n \to \infty \]

By the special form of the identity: $E(T - T^*)^2 = \text{Var}(T) - \text{Var}(T^*)$. It is enough to show that
\[ \text{Var}(T) - \text{Var}(T^*) \to 0 \text{ as } m, n \to \infty \]

Next, we define variances
\[ \sigma^2_{10} = \text{Var}[\psi_{10}(X_i)] \text{ and } \sigma^2_{01} = \text{Var}[\psi_{01}(U_j)] \]

\[
\text{Var}(T^*) = (\sqrt{m})^2 \text{Var} \left( \frac{1}{m} \sum_{i=1}^{m} \psi_{10}(X_i) + \frac{1}{n} \sum_{j=1}^{n} \psi_{01}(Y_j) \right) \\
= m \left[ \frac{1}{m^2} \text{Var} \left( \sum_{i=1}^{m} \psi_{10}(X_i) \right) + \frac{1}{n^2} \text{Var} \left( \sum_{j=1}^{n} \psi_{01}(Y_j) \right) \right] \quad \text{(by independence of } X' \text{s and } Y' \text{s)} \\
= m \left[ \frac{1}{m^2} m \text{Var} \left( \psi_{10}(X_i) \right) + \frac{1}{n^2} n \text{Var} \left( \psi_{01}(Y_j) \right) \right] \quad \text{(by identical distribution)} \\
= \text{Var} \psi(x_i) + \frac{m}{n} \text{Var}[\psi_{01}(Y_j)] \\
\to \sigma^2_{10} + \lambda \sigma^2_{01} \text{ as } m, n \to \infty. \]
On the other hand, if we define \( W = \sum \sum \varphi(X_i, Y_j) \). Then

\[
T = \frac{\sqrt{m}}{mn}(W - mn\theta).
\]

Hence, by the notation of lemma \( U_{ij} = \varphi(X_i, Y_j) - \theta = \psi(X_i, Y_j) \) and \( E(U_{ij}) = 0 \). So

\[
\text{Var} (T) = \frac{m}{m^2n^2} \text{Var} (W).
\]

\[
\text{Var} (T) = \frac{1}{mn^2} \left[ mn \text{Var} (U_{ij}) + mn(m - 1) \text{Cov} (U_{ij}, U_{kj}) + nm(m - 1) \text{Cov} (U_{ij}, U_{il}) \right]
\]

\[
= \frac{1}{n} \text{Var} (U_{ij}) + \frac{m - 1}{n} \text{Cov} (U_{ij}, U_{kj}) + \frac{n - 1}{n} \text{Cov} (U_{ij}, U_{il})
\]

\[
\rightarrow \lambda \text{Cov} (U_{ij}, U_{kj}) + \text{Cov} (U_{ij}, U_{il}) \text{ as } m, n \rightarrow \infty.
\]

But, we know

\[
\text{Cov} (U_{ij}, U_{kj}) = E(U_{ij}, U_{kj}) - E(U_{ij})E(U_{kj})
\]

\[
= E(U_{ij}, U_{kj})
\]

\[
= E(\psi(X_i, Y_j) \cdot \psi(X_k, \psi(Y_j)))
\]

\[
= E(E(\psi(X_i, Y_j) \cdot \psi(X_k, Y_j)|Y_j)
\]

\[
E(\psi(X_i, Y_j) \cdot \psi(X_k, Y_j)|Y_j = y_j) = E[\psi(X_i, y_j)] \cdot E[\psi(X_k, y_j)]
\]

\[
= \psi_{01}(y_j) \cdot \psi_{01}(y_j)
\]

\[
= \psi_{01}^2(y_j)
\]

Hence,

\[
\text{Cov} (U_{ij}, U_{kj}) = E(\psi_{01}^2(Y_j))
\]

\[
= E(\psi_{01}^2(Y_j)) - E^2(\psi_{01}(Y_j)) \text{ (since } E(\psi_{01}^2(Y_j)) = 0)\)
\]

\[
= \text{Var}[\psi_{01}(y_j)]
\]

\[
= \sigma_{01}^2
\]

Similarly, we can prove that \( \text{Cov} (U_{ij}, U_{il}) = \sigma_{10}^2 \). Hence \( \text{Var} (T) \rightarrow \sigma_{10}^2 + \lambda \cdot \sigma_{01}^2 \) as \( m, n \rightarrow \infty \). So, \( \text{Var} (T) - \text{Var} (T^*) \rightarrow 0 \implies E(T - T^*)^2 \rightarrow 0 \). Therefore, we have
proved that $T = \sqrt{m} \cdot (U - \theta)$ has the same limiting distribution as $T^*$, where

$$T^* = \sqrt{m} \cdot \left[ \frac{1}{m} \sum_{i=1}^{m} \psi_{10}(X_i) + \frac{1}{n} \sum_{j=1}^{n} \psi_{01}(Y_j) \right]$$

$$= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \psi_{10}(X_i) + \frac{\sqrt{m}}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi_{01}(Y_j) \rightarrow N(0, \sigma_{10}^2 \lambda)$$

Since, $\psi_{10}(x_i)$ are i.i.d., $\sum \psi_{10}(X_i) \rightarrow N(0, \sigma_{10}^2)$ by the central limit theorem.

Similarly, $\sum \psi_{01}(Y_j) \rightarrow N(0, \sigma_{01}^2)$.

Proof. Before proving the theorem, let us consider $Y_1, Y_2, \ldots, Y_n$ be $n$ random variables. Let $Y(1), Y(2), \ldots, Y(n)$ be the order statistics corresponding to $Y_1, Y_2, \ldots, Y_n$. If the $Y_i, i = 1, 2, \ldots, n$ are independent identically distributed continuous random variable with probability density function $f$. Then, the joint probability density function of the order statistics $Y_1, Y_2, \ldots, Y_n$ is given by

$$f(y_1, y_2, \ldots, y_n) = n! \prod_{i=1}^{n} f(y_i), \quad y_1 < y_2 < \cdots < y_n$$

To see that, consider the cumulative distribution function $Y(1), Y(2), \ldots, Y(n)$, since $(Y(1), Y(2), \ldots, Y(n))$ will equal $(y_1, y_2, \ldots, y_n)$ if $(Y_1, Y_2, \ldots, Y_n)$ is equal to any of the $n!$ permutation of $(y_1, y_2, \ldots, y_n)$. Hence, let

$$\alpha = \{ (i_1, i_2, \ldots, i_n) \mid i_j = k, j, k = 1, 2, \ldots, n \}$$

Once $(i_1, i_2, \ldots, i_n)$ is given, we have

$$F(y_1, y_2, \ldots, y_n) = F(y_1)F(y_2) \ldots F(y_n)$$

So

$$F(y_1, y_2, \ldots, y_n) = \sum_{\alpha} F(y_1)F(y_2) \ldots F(y_n)$$

$$= n!F(y_1)F(y_2) \ldots F(y_n)$$

Thus,

$$f(y_1, y_2, \ldots, y_n) = \frac{\partial F(y_1, y_2, \ldots, y_n)}{\partial y_1 \partial y_2 \ldots \partial y_n}$$

$$= n!f(y_1)f(y_2) \ldots f(y_n)$$
To obtain the conditional density of $S_1, S_2, \ldots, S_n$ given that $N(t) = n$. Note that the event that $S_1 = s_1, S_2 = s_2, \ldots, S_n = s_1, N(t) = n$ is equivalent to the waiting time $T_1 = s_1, T_2 = s_2 - s_1, T_n = t - s_n, T_{n+1} > t - s_n$. Since the waiting time of a Poisson Process of intensity rate $\lambda \sim \exp(\frac{1}{\lambda})$, $T_i \sim \exp(\frac{1}{\lambda})$ for $i = 1, 2, \ldots, n + 1$. So, the conditional probability density of $s_1, s_2, \ldots, s_n$ is

$$f(s_1, s_2, \ldots, s_n \mid n) = \frac{f(s_1, s_2, \ldots, s_n)}{P\{N(t) = n\}}$$

$$= \frac{\lambda e^{-\lambda s_1} \lambda e^{-\lambda (s_2 - s_1)} \cdots \lambda e^{-\lambda (s_n - s_{n-1})} e^{-\lambda (t - s_n)}}{e^{-\lambda t} (\lambda t)^n / n!}$$

$$= \frac{n!}{t^n}$$

So, $f(s_i) = \frac{1}{\tau}$, which proves the theorem. \qed
Appendix B

R Code

B.1 An illustration of R Code

In this subsection, a simple R function is presented to compute a $Z_\epsilon$ under the simplest case, i.e., two random sequences are generated from a uniform distribution. This is the ideal case, and any choice of the length $\delta$ of subintervals will achieve a good approximation to the uniform distribution, since the two sequences are uniformly distributed on the interval $[0, T]$. We use this function to test the approximation result of $Z_\epsilon$.

Then the uniform assumption is relaxed in the next several subsections, and some variations will be added to the $Z$ function defined below.

The following $Z$ function is a function of 5 arguments: time, $\delta$, $\epsilon$, $m$, and $n$. $Z$ function calculates the $Z_\epsilon$ value, which is supposed to be standard normal.

```r
> z <- function(time, delta, epsilon, m, n) {
+   a <- sort(runif(m, 0, time))
+   b <- sort(runif(n, 0, time))
+   aindex <- matrix(c(rep(NA, m)), m, 1)
+   bindex <- matrix(c(rep(NA, n)), n, 1)
+   for (j in 1:trunc(time/delta)) {
+     aindex[which(a <= j * delta & a >= (j - 1) * delta)] <- j
+   }
+   for (j in 1:trunc(time/delta)) {
+     bindex[which(b <= j * delta & b >= (j - 1) * delta)] <- j
+   }
+   aepsilon <- aindex + abs(bindex - aindex)
+   bepsilon <- bindex + abs(aindex - bindex)
+   cepsilon <- aepsilon + bepsilon
+   c <- min(cepsilon)
+   z <- c
+   return(z)
+ }
```
To see how the $Z$ function works, let us break down the whole function into several blocks and illustrate how each block works.
Program 1: Calculate $I[^{|X_{ik} - Y_{il}| < \epsilon}]$, $\sigma^2_{10}$, and $S_\epsilon$

```r
+ e <- data.frame(a, b, aindex, bindex)
+ count <- matrix(c(rep(NA, 3 * time/delta)), time/delta, 3)
+ for (j in 1:trunc(time/delta)) {
+   x <- e$a[which(e$aindex == j)]
+   y <- e$b[which(e$bindex == j)]
+   f <- merge(x, y, by = intersect(names(x), names(y)), all = T)
+   count[j, 1] <- length(abs(f$x - f$y)[which(abs(f$x - f$y) <= epsilon)])
+   count[j, 2] <- length(x)
+   count[j, 3] <- length(y)
+   }
+ pepsilon <- epsilon/delta * (2 - epsilon/delta)
+ sigma10 <- 2/3 * (epsilon/delta)^3 - (epsilon/delta)^4
+ count <- data.frame(count)
+ names(count) <- c("counts" , "m" , "n")
+ count$var <- count$m * count$n * pepsilon * (1 - pepsilon) + count$m * count$n * (count$m + count$n - 2) * sigma10
+ z <- (sum(count$counts) - pepsilon * sum(count$m * count$n))/sqrt(sum(count$var))
```

In Program 2, the first loop calculates $I[^{|X_{ik} - Y_{il}| < \epsilon}]$, $m_i$, and $n_i$. $p_\epsilon$ calculates

$$p_\epsilon = \frac{\epsilon}{\delta}(2 - \frac{\epsilon}{\delta})$$

$\sigma^2_{10}$ calculates

$$\sigma^2_{10} = \frac{2}{3} \cdot \left( \frac{\epsilon}{\delta} \right)^3 - \left( \frac{\epsilon}{\delta} \right)^4$$

count$var calculates

$$\text{Var}(S_\epsilon) = m_in_i\frac{\epsilon}{\delta}(2 - \frac{\epsilon}{\delta})[1 - \frac{\epsilon}{\delta}(2 - \frac{\epsilon}{\delta})] + m_in_i(m_i + n_i - 2)[\frac{2}{3}(\frac{\epsilon}{\delta})^3 - (\frac{\epsilon}{\delta})^4]$$

Finally, $z$ calculates

$$Z_\epsilon = \frac{S_\epsilon - E(S_\epsilon)}{\sqrt{\text{Var}(S_\epsilon)}} = \frac{\sum_{i=1}^t S_{\epsilon_i} - P_\epsilon \sum_{i=1}^t m_i \cdot n_i}{\sum_{i=1}^t \text{Var}(S_{\epsilon_i})}$$
B.2 Function that can pick the minimum number of data points
library(micEcon)
zmin<-function(time,md,epsilon,m,n,i){
    #Generate two sequences
    set.seed(i^2+106)
a<-sort(runif(m,0,time))
    set.seed(sqrt(i^3)+50*i)
b<-sort(runif(n,0,time))
    #index matrix
    index<-matrix(c(rep(NA,1000)),1000,1)
    M=0
    for (i in 1:1000){
        M<-max(b[b>M][md],a[a>M][md])
        index[i]<-M
    }
    index<-insertRow(index, 1, 0)
    index<-index[-which(apply(index,1,function(x)all(is.na(x))))]
    aindex<-matrix(c(rep(NA,m)),m,1)
    bindex<-matrix(c(rep(NA,n)),n,1)
    for (j in index[-1]){aindex[which(a<=j&a>index[which(index==j)-1])]<-which(index==j)-1}
    for (j in index[-1]){bindex[which(b<=j&b>index[which(index==j)-1])]<-which(index==j)-1}
    e<-data.frame(a,b,aindex,bindex)
    o<-length(index)-1
dif<-diff(index)
count<-matrix(c(rep(NA,5*o)),length(index)-1,5)
    for (j in 1: o){
        k=0
        x<-e$a[which(e$aindex==j)]
        y<-e$b[which(e$bindex==j)]
        k=as.numeric(loop(x,y,epsilon,k)[6])
        count[j,1]<-k
        count[j,2]<-length(x)
        count[j,3]<-length(y)
        delta<-dif[j]
        count[j,4]<-epsilon/delta*(2-epsilon/delta)
        count[j,5]<-2/3*(epsilon/delta)^3-(epsilon/delta)^4
    }
    count<-data.frame(count)
    names(count)<-c("counts","m","n","pepsilon","sigma10")
    count$var<-count$m*count$n*count$pepsilon*(1-count$pepsilon)+
        count$m*count$n*(count$m+count$n-2)*count$sigma10
    count$Es<-count$pepsilon*count$m*count$n
    #Calculate z statistic
    z<-(sum(count$counts)-sum(count$Es))/sqrt(sum(count$var))
    print(z)
}
B.3 Complete Z Function

The next program is the updated $z$ function, which can control the weight, and choose different underlying densities for random processes.
Program 3 Complete $z$ function

```r
z4 <- function(time, delta, ratio, variance, beta, alpha, m, n, i, pro1, pro2, weights) {
  a <- pro1(beta, m, time)
  b <- pro2(alpha, n, time)
  index <- seq(0, time, by = delta)
  perturb <- rnorm(length(index) - 1, 0, variance)
  for (i in 2:length(index)) {
    index[i] <- index[i] + perturb[i - 1]
  }
  aindex <- matrix(c(rep(length(index), m)), m, 1)
  bindex <- matrix(c(rep(length(index), n)), n, 1)
  for (j in index[-1]) {
    aindex[which(a <= j & a > index[which(index == j) - 1])] <- which(index == j) - 1
  }
  for (j in index[-1]) {
    bindex[which(b <= j & b > index[which(index == j) - 1])] <- which(index == j) - 1
  }
  o <- length(index) - 1
  dif <- diff(index)
  count <- matrix(c(rep(NA, 6 * o)), length(index) - 1, 6)
  for (j in 1:o) {
    k = 0
    delta <- dif[j]
    x <- a[which(aindex == j)]
    y <- b[which(bindex == j)]
    k <- as.numeric(loop(x, y, delta * ratio, k)[6])
    count[j, 1] <- k
    count[j, 2] <- length(x)
    count[j, 3] <- length(y)
    count[j, 4] <- ratio * (2 - ratio)
    count[j, 5] <- 2/3 * (ratio)^3 - (ratio)^4
    count[j, 6] <- length(y) + length(x) / (n + m)
  }
  count <- data.frame(count)
  names(count) <- c("counts", "m", "n", "pepsilon", "sigma10", "weights")
  count$var <- count$m * count$n * count$pepsilon * (1 - count$pepsilon) +
              count$m * count$n * (count$m + count$n - 2) * count$sigma10
  count$Es <- count$pepsilon * count$m * count$n
  if (weights == 1) {
    z <- (sum(count$counts * count$weights) - sum(count$Es * count$weights)) / sqrt(sum(count$weights ^ 2 * count$var))
  } else {
    z <- (sum(count$counts) - sum(count$Es)) / sqrt(sum(count$var))
  }
  print(z)
}
```
The next block of program is the code for choosing the optimal \( \delta \) and \( r \) when the sample size is large.

**Program 4** Choose optimal \( \delta \) and \( r \)

```r
# fix \( r \) to pick ideal delta value
delta<-seq(5,100,by=1)
drresultl1<-matrix(c(rep(NA,96*3)),96,3)
for (i in delta){
    z<-matrix(c(rep(NA,4000)),4000,1)
    for (k in 1:4000){
        z[k]<-z4(1000,i,0.05,0.1,0.48,20000,20000,k,trig1,beta1,1)
    }
    drresultl1[i-4,1]<-mean(z)
    drresultl1[i-4,2]<-var(z)
    drresultl1[i-4,3]<-shapiro.test(z)$p
}

# fix another \( r \) to pick ideal delta value
delta<-seq(5,54,by=1)
drresultl2<-matrix(c(rep(NA,50*3)),50,3)
for (i in delta){
    z<-matrix(c(rep(NA,4000)),4000,1)
    for (k in 1:4000){
        z[k]<-z4(1000,i,0.1,0.1,0.48,20000,20000,k,trig1,beta1,1)
    }
    drresultl2[i-4,1]<-mean(z)
    drresultl2[i-4,2]<-var(z)
    drresultl2[i-4,3]<-shapiro.test(z)$p
}

# after pick the ideal delta value, let choose optimal \( r \) value
r<-seq(0.01,0.49,by=0.02)
drresultl3<-matrix(c(rep(NA,25*3)),25,3)
for (i in r){
    z<-matrix(c(rep(NA,3000)),3000,1)
    for (k in 1:3000){
        z[k]<-z4(1000,7,i,0,0.1,0.48,20000,20000,k,trig1,beta1,1)
    }
    drresultl3[((i+0.01)*100)/2,1]<-mean(z)
    drresultl3[((i+0.01)*100)/2,2]<-var(z)
    drresultl3[((i+0.01)*100)/2,3]<-shapiro.test(z)$p
}
```