CALCULATING KENDALL'S TAU WITH MULTIPLE MEASUREMENTS

A Thesis

by

ADAM GELLER

BS, University of the Ozarks, 2016

Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF SCIENCE

in

MATHEMATICS

Texas A&M University-Corpus Christi
Corpus Christi, Texas

May 2018

May 2018

CALCULATING KENDALL'S TAU WITH MULTIPLE MEASUREMENTS

A Thesis

by

ADAM GELLER

This thesis meets the standards for scope and quality of
Texas A&M University-Corpus Christi and is hereby approved.

Blair Sterba-Boatwright, PhD
Chair

Lei Jin, PhD
Committee Member

Jose Guardiola, PhD
Committee Member

May 2018

ABSTRACT

Relationships between time series of environmental variables are commonly calculated using non-parametric methods, such as Kendall's τ, because of "non-detects", i.e., left-censored data that falls below a measurement limit. However, these methods are not well-adapted to situations where variables have multiple contemporaneous measurements. In this thesis, we define a new method, $\tilde{\tau}$, in an attempt to calculate correlations using each of the multiple measurements instead of daily means. We investigate $\tilde{\tau}$ using two methods: simulations that approximate a null distribution for $\tilde{\tau}$ and closed form calculations for a specific special case. We also apply $\tilde{\tau}$ to an actual data set.

The results of our investigation shows that $\tilde{\tau}$ may handle certain things, such as outliers, better than current methods. However, its requirements for distributional assumptions about the data make it a less practical option for real data. Further work could explore ways to avoid the prerequisite need for distribution knowledge and could also further investigate $\tilde{\tau}$ under noise sampled from asymmetric distributions.

ACKNOWLEDGEMENTS

I would like to first thank Dr. Blair Sterba-Boatwright for being my chair and offering his guidance through each step of this thesis. I would also like to thank my committee members Dr. Lei Jin and Dr. José Guardiola for their assistance throughout my research.

DEDICATION

To Goose

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I: INTRODUCTION

Samples measuring the levels of *Vibrio vulnificus* and dissolved organic carbon (DOC) were taken from July 12, 2016 to August 6, 2016 in the water off of Looe Key, Florida. On each day, three separate samples were taken and measured for *Vibrio*=y and two were taken for DOC=x, giving six (x, y) observations for each day. These measurements were taken within the same hour and will be treated as contemporaneous events. We wish to determine if there is a correlation between DOC and *Vibrio* levels. Although DOC measurements were not censored, other nutrient measurements in the dataset were left-censored (below detection limits), making the use of Kendall's τ the best choice for measuring correlation. [1]

The current practice is to take the average of the multiple measurements from a given day to create a single data point for that day. Correlation is then calculated based on the daily averages. However, measurements can be very different within a single day, and daily averages can be greatly influenced by outlying data points, so the mean is not necessarily the best single estimate of the "true" daily value. In addition, information is lost when multiple observations are replaced by their mean.

Another possible approach would be to treat each of the six pairs of measurements from a given day as independent observations. However, examination of the data makes independence of observations on the same day unlikely. If each pair of points on each day is inappropriately treated as a full observation, the effective size of the data set is exaggerated. This can cause problems when testing for significance. As we will soon see, the test statistic for Kendall's τ relies on the sample size of the data. An increase in the sample size will cause the test statistic to decrease, reducing the power of the test.

In this thesis, we will explore a new approach to estimating correlations with multiple measurements. We will calculate τ for each pair of days, then use the average of the τ's to estimate the correlation in the data. This method will still use each data point collected on each day and retain

robustness to outliers.

We further explore τ defined in this new manner. We created software to simulate how this new approach is affected by different sized noise around randomly sample data points and also wrote code for simulations that allow us to observe null distribution of the new method. Using this code, we were able to verify our results found analytically for a particular ”all-normal” case.

CHAPTER II: REVIEW OF THE LITERATURE

2.1 Kendall's τ

Kendall's τ was first introduced in Kendall (1938). For this thesis, we will just be using Kendall's τ without accounting for ties. The formula for calculating τ without ties as given by Kendall (1962) is:

$$\tau_{obs} = \frac{\sum_{1\leq i<j\leq n}^{n} Q((x_i,y_i),(x_j,y_j))}{n(n-1)/2}$$

where Q is the concordance indicator function:

$$Q((x_i,y_i),(x_j,y_j)) = -1 \text{ if } (x_i - x_j)(y_i - y_j) < 0$$
$$Q((x_i,y_i),(x_j,y_j)) = +1 \text{ if } (x_i - x_j)(y_i - y_j) > 0$$

For testing the significance of τ we define the following quantities:

$$\text{S} = \sum_{1\leq i<j\leq n}^{n} Q((x_i,y_i),(x_j,y_j))$$
$$\text{Z} = \frac{S - sign(S)}{\sqrt{(n(n-1)(2n+5))/18}}$$

Because it is defined using ranks, τ is robust to outliers and invariant under rank-preserving transformations.

Correlations of environmental parameters are often calculated using Kendall's τ instead of Pearson's ρ [1]. This is because of the many practical benefits of using a nonparametric ordinal method for correlation when dealing with environmental measurements, such as no need to assume any distributions of the data and the previously mentioned robustness to outliers. Environmental measurements can also have non-detects (in statistical terms, left-censored data). A common method for handling non-detects is substituting a 0 or the reporting limit for all non-detects, but substituting data this way and calling it "real data" can impose an artificial pattern.

2.2 Mean and Variance of τ in the Null Case

In this section we prove the following theorem:

Let (x_i, y_i) and (x_j, y_j) be an arbitrary pair of points, with i not necessarily different from j. Define $a_{ij} = sign(x_i - x_j)$, with b_{ij} defined similarly for y. The distribution of τ under the null hypothesis H_0: a_{ij} is independent of b_{ij} for all i, j, has mean 0 and variance $\frac{n(n-1)(2n+5)}{18}$. The proof is based on Kendall (1962), Chapters 4 and 5 [2]. Let $c_{ij} = a_{ij}b_{ij}$, then

$$c = \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} = 2S$$

Now

$$\sum_{l=1}^{n} a_{il} = n + 1 - 2i$$

Further,

$$\sum_{i=1}^{n} \sum_{l=1}^{n} a_{il}^2 = n(n-1)$$

for a^2 is +1 and this sum is the number of possible ways of choosing a pair from n members. It follows that

$$\begin{aligned} \sum_{i=1}^{n} \sum_{l=1}^{n} a_{il} &= \sum_{i=1}^{n} (n + 1 - 2i) \\ &= n(n+1) - 2 \sum_{i=1} i \\ &= 0 \end{aligned}$$

We also have

$$\begin{aligned} \sum_{i=1}^{n} \sum_{l=1}^{n} \sum_{t=1}^{n} a_{il} a_{it} &= \sum_{i} \sum_{l} a_{il}(n + 1 - 2i) \\ &= \sum_{i} (n + 1 - 2i)^2 \\ &= \sum (n+1)^2 - 4(n+1) \sum i + 4 \sum i^2 \\ &= n(n+1)^2 - 2n(n+1)^2 + \tfrac{2}{3} n(n+1)(2n+1) \\ &= \tfrac{1}{3} n(n^2 - 1) \end{aligned}$$

For the mean values on summation over all possible permutations, we have

$$E(c) = \sum_{i=1}^{n} \sum_{j=1}^{n} E(a_{ij} b_{ij})$$

Since a and b are independent, the mean value of any term a_{ij} or b_{ij} is zero. So

$$E(c) = 0$$

For the variance of c we have

$$E(c^2) = E[\sum_i \sum_j (a_{ij} b_{ij})]^2 = E\left(\sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} a_{ij} b_{ij} a_{kl} b_{kl} \right)$$

The quadruple sum may be broken into four pieces.

(i) Case 1: $i = j = k = l$.

Figure 2.1: Case 1

In this case τ will clearly be zero, as we are comparing a single point to itself. $a_{ii} = \text{sgn}(x_i - x_i) = 0$

(ii) Case 2: $i = k$ and $j = l$, or, $i = l$ and $j = k$.

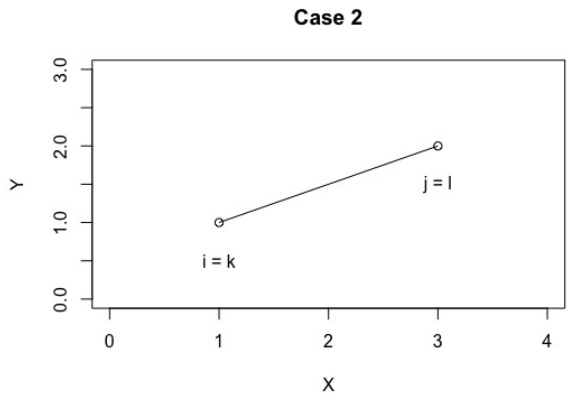

Figure 2.2: Case 2

This case corresponds to the terms involving only two (x,y) points. Here we have $a_{ij}b_{ij}a_{ij}b_{ij}$ or $a_{ij}b_{ij}a_{ji}b_{ji}$ which are equivalent to $\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 b_{ij}^2$. So for $E[\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 b_{ij}^2]$, an individual term arises from the square of $a_{ij}b_{ij}$ or the product of that term with $a_{ij}b_{ji}$. Thus the sum is twice the sum of $a_{ij}^2 b_{ij}^2$ with i,j from 1 to n. There are $n(n-1)$ terms in the sum, and the expected value is therefore

$$\begin{aligned} 2n(n-1)E(a_{12}^2 b_{12}^2) &= 2n(n-1)E(a_{12}^2)E(b_{12}^2) \\ &= \frac{2\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2}{n(n-1)} \end{aligned}$$

(iii) Case 3: Exactly one of (i,j) equals exactly one of (k,l).

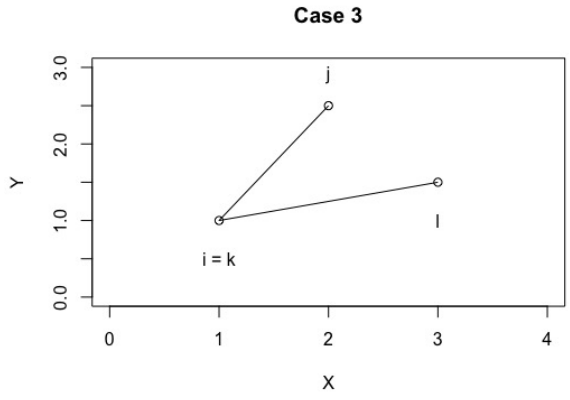

Figure 2.3: Case 3

This case corresponds to the terms involving three (x,y) points. The product $a_{ij}a_{ik}b_{ij}b_{ik}$ can arise in four ways, since if we fix i,j,k for the a's and the b terms can appear with suffixes $(ij,ik),(ji,ik),(ij,ki)$ and (ji,ki) and there are $n(n-1)(n-2)$ ways of fixing three different suffixes out of n. So

$$E[\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}(a_{ij}b_{ij}a_{ik}b_{ik})] = \frac{4}{n(n-1)(n-2)}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}a_{ij}a_{ik}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}b_{ij}b_{ik}$$

with

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}a_{ij}a_{ik} = \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}a_{ij}a_{ik} - \sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}^2$$

(iv) Case 4: Each i, j, k, l is distinct.

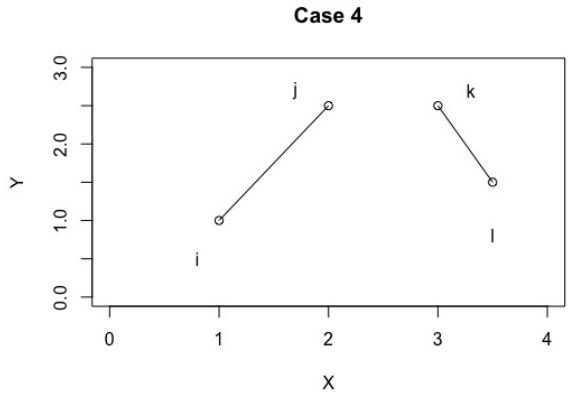

Figure 2.4: Case 4

This corresponds with terms involving four (x,y) points. We will show this sum has an expected value of 0. Since a's and b's are independent, it suffices to show $E(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}a_{ij}a_{kl}) = 0$. Let $\sum''$ be the summation over all values of where the suffixes are not equal and $\sum'$ being the sum where exactly one pair of suffixes are equal. We have:

$$\sum(a_{ij}a_{kl}) = \sum''(a_{ij}a_{kl}) + \sum'(a_{ij}a_{il}) + \sum'(a_{ij}a_{ki}) + \sum'(a_{ij}a_{jl}) + \sum'(a_{ij}a_{kj}) + \sum(a_{ij}a_{ij}) + \sum(a_{ij}a_{ji})$$

Since $a_{ij} = -a_{ji}$, all terms after the first term to the right cancel in pairs. The term $E[\sum(a_{ij}b_{ij})] = 0$, so therefore, $0 = E(\sum'' a_{ij}a_{kl})$

For the variance of c we now have

$$E(c^2) = \frac{2}{n(n-1)}\sum a_{ij}^2 \sum b_{ij}^2 + \frac{4}{n(n-1)(n-2)}[\sum a_{ij}a_{ik} - \sum a_{ij}^2][\sum b_{ij}b_{ik} - \sum b_{ij}^2]$$

Using $\sum_{i=1}^{n}\sum_{l=1}^{n} a_{il} = 0$ and $\sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{t=1}^{n} a_{il}a_{it} = \frac{1}{3}n(n^2-1)$ we have

$$\begin{aligned} E(c^2) &= \frac{2}{n(n-1)}n(n-1)]^2 + \frac{4}{n(n-1)(n-2)}[\frac{1}{3}n(n^2-1) - n(n-1)]^2 \\ &= \frac{2n(n-1)(2n+5)}{9} \end{aligned}$$

Therefore

$$\begin{aligned} var(S) &= E(S^2) \\ &= E((\tfrac{c}{2})^2) \\ &= \tfrac{1}{4}E(c) \\ &= \frac{n(n-1)(2n+5)}{18} \end{aligned}$$

which is needed for our test statistic, Z.

The proof for τ tending to normality is beyond the scope of this thesis, but proofs can be found in [2] and [3].

2.3 Standard Statistical Tools used in the Methods Sections

Order Statistics

Nonparametric methods are often based on ordered data values called order statistics [4]. Let $X_1, X_2, ..., X_n$ be independent identically distributed observations from a continuous distribution with cumulative distribution function (cdf) $F(x)$ and probability density function (pdf) $f(x)$. Define the order statistics as $X_{(1)} < X_{(2)} < ... < X_{(n)}$. The pdf for the distribution of the r^{th} order statistic is given by

$$f_{(r)}(x) = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{r-1}[1-F(x)]^{n-r}f(x)$$

Convolution

In the following sections we will need a formula for finding the distribution for the sum or difference of two independent variables. The distribution is given by the convolution of the variables' distributions. If $Z = X + Y$, then $f_Z(z)$ is given by:

$$f_Z = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(t) f_Y(z-t) dt$$

Similarly, if $Z = X - Y$, then $f_Z(z)$ is given by:

$$f_Z = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(t) f_Y(t-z) dt$$

CHAPTER III: METHODOLOGY

3.1 Defining $\tilde{\tau}$

At time point i, let $x_i = (x_{i,1}, x_{i,2}, ..., x_{i,n_x})$ be measurements of the x variable and $y_i = (y_{i,1}, y_{i,2}, ..., y_{i,n_y})$ be measurements of the y variable. Define x_j and y_j similarly at time point $j \neq i$. Define $\tilde{\tau}_{ij}$ to be

$$\tilde{\tau}_{ij} = \frac{\sum_{s=1}^{n_x}\sum_{t=1}^{n_y}\sum_{q=1}^{n_x}\sum_{r=1}^{n_y} Q((x_{i,s},y_{i,t}),(x_{j,q},y_{j,r}))}{n_x^2 n_y^2}$$

Define $\tilde{\tau}$ to be the mean of all $\tilde{\tau}_{ij}$'s.

3.2 Simulation for Sampling Distribution of $\tilde{\tau}$ Under the Null Hypothesis

We use two simulations written in R [5] to study the sampling distribution of $\tilde{\tau}$ under the null hypothesis. The simulation assumes that the observations x_i arise as n_x random "noise" values added to a single representative x_i value. The same distribution of noise values applies to all x_i's and the x_i's in turn are a random sample from some distribution stationary in time. The same assumptions apply to y_i. The inputs required for each simulation are a distribution for x, a distribution for y, a distribution for noise in x, a distribution for noise in y, the total number n of (x,y) points, the total number of noise points n_x for x and n_y for y, and the number of simulations.

In the first simulation, the desired number of points are sampled from the distributions given for x and y, then the following is done many times. The appropriate number of noise points are sampled and added to the x and y points. τ is then calculated for each pair of (x,y) coordinates and then averaged, giving us $\tilde{\tau}$. This algorithm is done many times and $\tilde{\tau}$ is stored for each iteration. The result is a distribution of $\tilde{\tau}$ that is compared to the traditional τ before noise is added. With this simulation we can observe and compare the effects of the noise sampled from different distributions on the correlation calculated.

The second simulation runs just as the first, but samples new initial values for x and y for each iteration. This allows us to better observe the null distribution of $\tilde{\tau}$ as we can see how it behaves for

many different x and y many times. We can also use this simulation to investigate the distribution of particular cases. The results of this simulation will be used to support an analytical problem introduced in the next section.

3.3 2×2 Closed Form Case

We introduce the following notation: for any random variable W, let f_w be the pdf and F_w be the cdf.

Let x_i, $i = 1,2$, be independent random values from a distribution f_x. For each i, let $\varepsilon_{i(j)}$, $j = 1,2$, be two independent random "noise" values drawn from f_ε. Adding the "noise" gives us four data points to consider. We write these points as order statistics $x_{1(1)}$, $x_{1(2)}$, $x_{2(1)}$, and $x_{2(2)}$, where $x_{i(j)} = x_i + \varepsilon_{i(j)}$. Four comparisons are needed to calculate τ for x_1 and x_2:

$$x_{2(1)} - x_{1(1)}, x_{2(1)} - x_{1(2)}, x_{2(2)} - x_{1(1)}, x_{2(2)} - x_{1(2)}$$

The number of positive differences from these comparisons is any integer from zero to four. We want to find the probabilities of each number of positive differences. Due to the symmetry of the problem, the case of zero positive differences will have the same probability as the case for four and similarly the probabilities of one and three positive values are the same. This means there are just three cases we need to understand: the cases of four, three, and two positive differences.

We start with the case given by Figure 3.1, where all four comparisons are positive. This occurs if the smallest point associated with x_2 is larger than the largest point from x_1. What we must find is $P(x_{1(2)} < x_{2(1)})$. Since x_1 and x_2 are independent of each other, the probability is given by:

$$\begin{aligned} &\int_{-\infty}^{\infty} \int_{-\infty}^{x_{2(1)}} f_{x_{1(2)}}(x_{(1)}) f_{x_{2(1)}}(x_{(2)}) dx_{(1)} dx_{(2)} \\ &= \int_{-\infty}^{\infty} F_{x_{1(2)}}(x_{(2)}) f_{x_{2(1)}}(x_{(2)}) dx_{(2)} \end{aligned}$$

where $F_{x_{i(j)}}(x)$ and $f_{x_{i,(j)}}(x)$ are the respective cdf and pdf.

Next is the case of three positive differences, seen in Figure 3.2. This occurs if the points are arranged so that $x_{1(1)} < x_{2(1)} < x_{1(2)} < x_{2(2)}$, so all four points will be considered when finding the probability. We start by expanding the inequality to

$$x_1 + \varepsilon_{1(1)} < x_2 + \varepsilon_{2(1)} < x_1 + \varepsilon_{1(2)} < x_2 + \varepsilon_{2(2)}$$

Let $\Delta x = x_2 - x_1$ which has a pdf $f_{\Delta x}$, found by the convolution of f_{x_2} and $-f_{x_1}$. Then

$$(\varepsilon_{1(1)} < \Delta x + \varepsilon_{2(1)}) \cap (\varepsilon_{2(1)} < \varepsilon_{1(2)} - \Delta x) \cap (\varepsilon_{1(2)} < \Delta x + \varepsilon_{2(2)})$$

This gives us the integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon_{2(2)}+\Delta x}\int_{-\infty}^{\varepsilon_{1(2)}-\Delta x}\int_{-\infty}^{\varepsilon_{2(1)}+\Delta x} f_{\varepsilon_{1(1)}}(\varepsilon_{1(1)}) f_{\varepsilon_{2(1)}}(\varepsilon_{2(1)}) f_{\varepsilon_{1(2)}}(\varepsilon_{1(2)}) f_{\varepsilon_{2(2)}}(\varepsilon_{2(2)}) f_{\Delta x}(\Delta x)$$
$$d\varepsilon_{1(1)} d\varepsilon_{2(1)} d\varepsilon_{1(2)} d\varepsilon_{2(2)} d\Delta x$$

The third case of two positive differences can occur in two ways. Either both x_1 points are between the two x_2 points or both x_2 points are between the two x_1 points. Without loss of generality, let us consider $x_{1(1)} < x_{2(1)} < x_{2(2)} < x_{1(2)}$, given by Figure 3.3. There is a bit more difficulty in solving this particular case. This is due to $x_{2(1)}$ not being independent from $x_{2(2)}$. This means we can not find the probability in the same way as before. What we will do is transform our variables. Let

$$(x_1 + \varepsilon_{1(1)}, x_1 + \varepsilon_{1(2)}) = \bar{x}_1 \pm \Delta\varepsilon_1$$

where

$$\bar{x}_1 = \frac{x_1 + \varepsilon_{1(1)} + x_1 + \varepsilon_{1(2)}}{2} = x_1 + \bar{\varepsilon}_1$$
$$\Delta\varepsilon_1 = \frac{\varepsilon_{1(2)} - \varepsilon_{1(1)}}{2} \text{ and } \bar{\varepsilon}_1 = \frac{\varepsilon_{1(2)} + \varepsilon_{1(1)}}{2}$$

Define $\bar{x}_2 \pm \Delta\varepsilon_2$ similarly. By our previous assumption, $\Delta\varepsilon_2 < \Delta\varepsilon_1$. Finally, define $\Delta\bar{x} = \bar{x}_2 - \bar{x}_1$. This gives the inequality

$$\Delta\varepsilon_2 - \Delta\varepsilon_1 < \Delta\bar{x} < \Delta\varepsilon_2 + \Delta\varepsilon_1$$

The distribution of $\Delta\varepsilon$'s can be found by the convolution of $f_{\varepsilon_{i(2)}}$ and $f_{\varepsilon_{i(1)}}$. Call this pdf $f_{\Delta\varepsilon}$ We must next find distributions for $\Delta\bar{x}$. To do this, we derive a distribution for $\bar{x}$. We do this by convolving the pdf for x, f_x, and the pdf of ε, f_ε. The result is the pdf $f_{\bar{x}}$. Then by convolving $f_{\bar{x}}$ with $-f_{\bar{x}}$ we can derive a distribution for $\Delta\bar{x}$ with pdf $f_{\Delta\bar{x}}$. Now we are left with the integral:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{\Delta\varepsilon_2-\Delta\varepsilon_1}^{\Delta\varepsilon_1-\Delta\varepsilon_2} f_{\Delta\bar{x}}(t)dtd\Delta\varepsilon_1 d\Delta\varepsilon_2$$

The probabilities for y can be found in the same way.

In our 2×2 closed case, there are just seven possible τ's than can appear. The combinations of positive y and positive x differences possible and the resulting possible τ's are shown by the table below.

		Number of Positive y				
		4	3	2	1	0
	4	1	0.5	0	-0.5	-1
Number of	3	0.5	0.25	0	-0.25	-0.5
Positive x	2	0	0	0	0	0
	1	-0.5	-0.25	0	0.25	0.5
	0	-1	-0.5	0	0.5	1

Table 3.1

Figures for 2X2 Closed Form

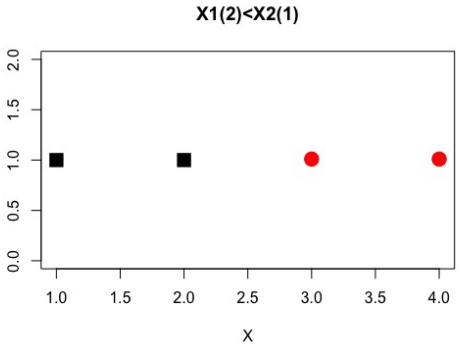

Figure 3.1: Figure for $x_{1(2)} < x_{2(1)}$

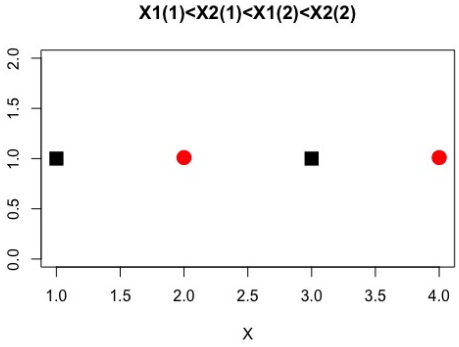

Figure 3.2: Figure for $x_{1(1)} < x_{2(1)} < x_{1(2)} < x_{2(2)}$

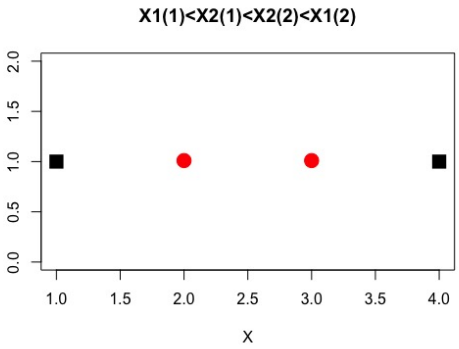

Figure 3.3: Figure for $x_{1(1)} < x_{2(1)} < x_{2(2)} < x_{1(2)}$

CHAPTER IV: RESULTS

4.1 Simulation Results

Our first function allowed us to observe how noise can affect the distribution of $\tilde{\tau}$. For our example, we choose 50 (x,y) points from $x,y \sim N(0,1)$, add two noise points to each x and three to each y, and sample each noise point for both x and y from $\varepsilon_x, \varepsilon_y \sim N(0,\sigma_\varepsilon)$. We ran each simulation 100 times. What we have shown through simulation is that a small σ_ε gives a much narrower distribution for $\tilde{\tau}$ centered around the true τ value, while an increase in σ_ε widens the distribution of $\tilde{\tau}$ and moves the center away from the true τ towards 0. This result makes sense since an increase in the size of the noise adds more randomness, pushing $\tilde{\tau}$ to 0.

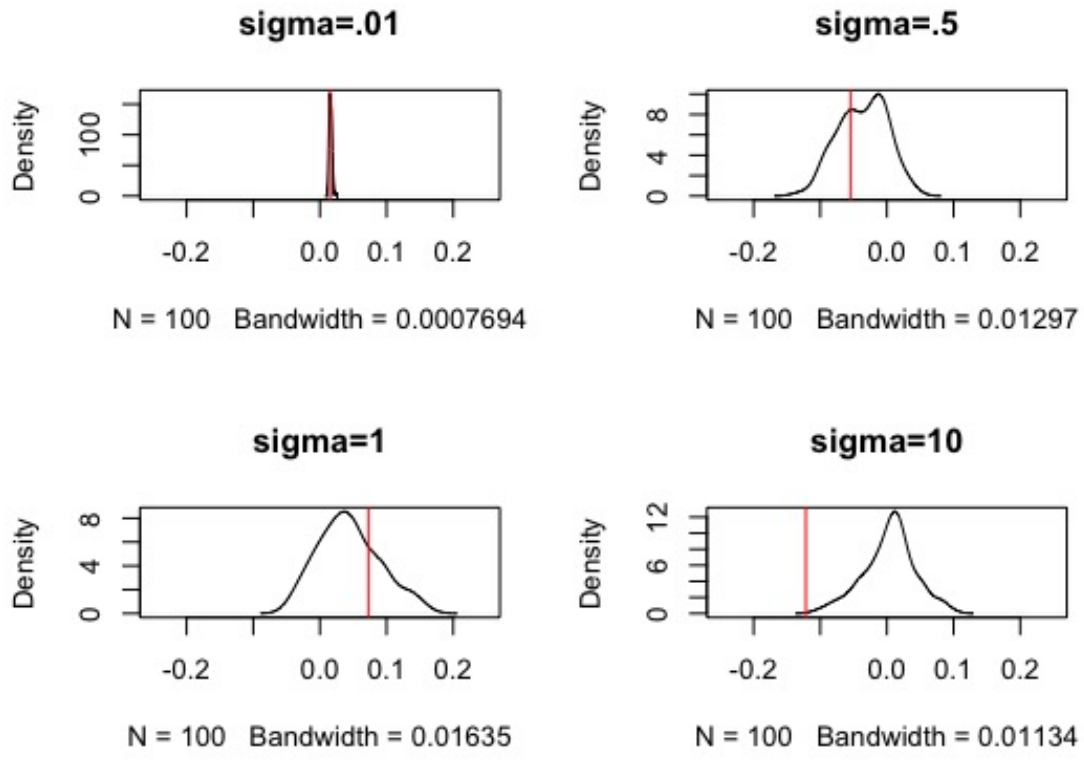

Figure 4.1: Graphs showing effect of noise with $\sigma_\varepsilon = 0.01, 0.5, 1, 10$

The simulation was also used to estimate the variation of $\tilde{\tau}$ for a given set of (x,y) points for

different values of σ_ε. Different values of σ_ε were chosen randomly between 0.05 and 20, and 500 simulations were run in the situation described above. In the plot below, σ_ε is plotted against the standard deviation of the resulting 500 values of $\tilde{\tau}$.

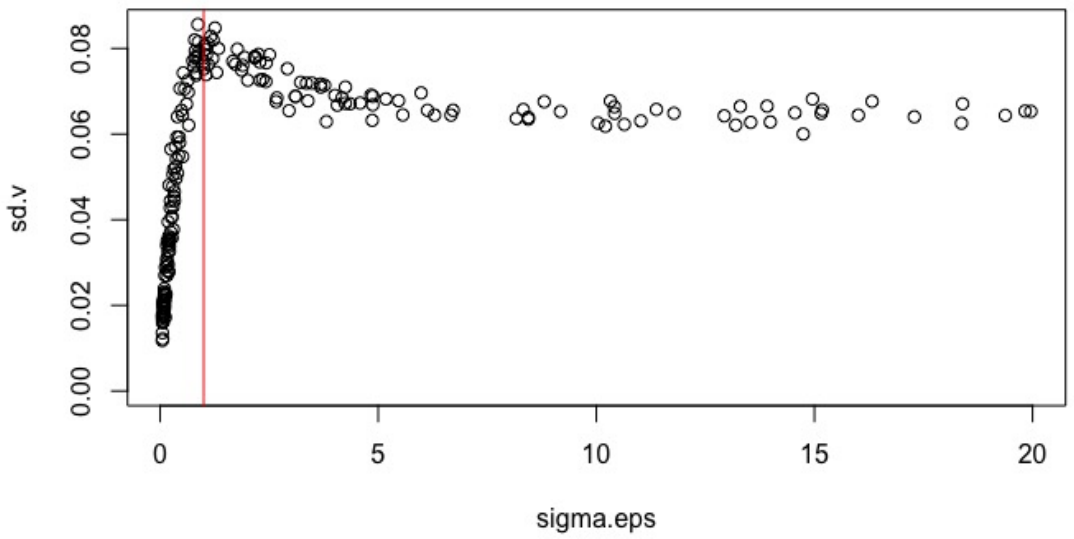

Figure 4.2: σ_ε vs. standard deviation of $\tilde{\tau}$

4.2 Integrals for 2×2 Normal Closed Case

Let us take the 2×2 closed case from the previous chapter and solve a particular case. Let $x \sim N(0,1)$ and $\varepsilon \sim N(0,0.5)$. The integrals in this section will be evaluated numerically using the trapezoid method.

Our first integral to solve is

$$\int_{-\infty}^{\infty} F_{x_{1(2)}}(x_{(2)}) f_{x_{2(1)}}(x_{(2)}) dx_{(2)}$$

We use our formula for pdf's of order statistics and plug in our normal pdf. The result of this integral is 0.3564333.

Now on to the second integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\varepsilon_{2(2)}+\Delta x}\int_{-\infty}^{\varepsilon_{1(2)}-\Delta x}\int_{-\infty}^{\varepsilon_{2(1)}+\Delta x} f_{\varepsilon_{1(1)}}(\varepsilon_{1(1)}) f_{\varepsilon_{2(1)}}(\varepsilon_{2(1)}) f_{\varepsilon_{1(2)}}(\varepsilon_{1(2)}) f_{\varepsilon_{2(2)}}(\varepsilon_{2(2)}) f_{\Delta x}(\Delta x)$$

$$d\varepsilon_{1(1)}d\varepsilon_{2(1)}d\varepsilon_{1(2)}d\varepsilon_{2(2)}d\Delta x$$

Using similar numeric techniques, we arrive to a result of 0.0812346.

Then our final integral

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{\Delta\varepsilon_2-\Delta\varepsilon_1}^{\Delta\varepsilon_1-\Delta\varepsilon_2} f_{\Delta\bar{x}}(t)dtd\Delta\varepsilon_1 d\Delta\varepsilon_2$$

is 0.0612689.

The probability distribution for the number of positive differences in x attributing to τ looks like

Number of positive differences	0	1	2	3	4
Probabilities	0.3564	0.0812	0.1225	0.0812	0.3564

Table 4.1

Using the same for y, we get

$$\tilde{\tau} = \pm 1: \qquad 2(0.3564)^2 = 0.2540$$

$$\tilde{\tau} = \pm 0.5: \qquad 4(0.3564)(0.0812) = 0.1157$$

$$\tilde{\tau} = \pm 0.25: \qquad 2(0.0812)^2 = 0.0131$$

$$\tilde{\tau} = 0: \qquad 2(0.1225) - (0.1225)^2 = 0.2299$$

To check these results we ran our second simulation 100,000 times with $f_x, f_y \sim N(0,1)$ and $f_{\varepsilon_x}, f_{\varepsilon_y} \sim N(0,0.5)$. The results from the simulation support our analytical results.

Results from 100000 simulations for 2×2 normal case						
-1	-0.5	-0.25	0	0.25	0.5	1
0.25579	0.11764	0.01381	0.22822	0.01340	0.11760	0.25354

Table 4.2

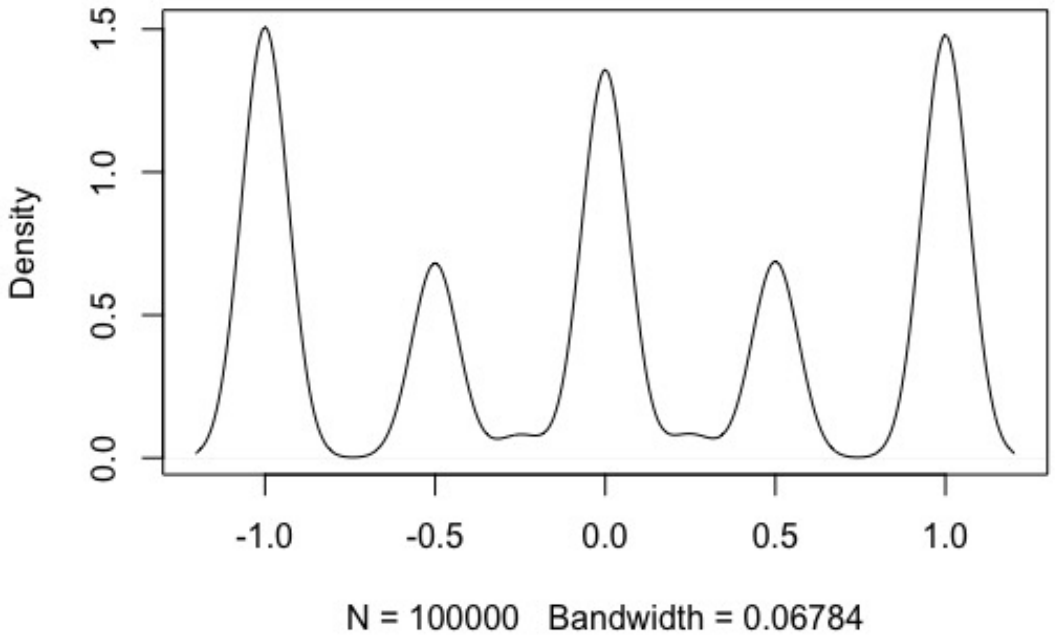

Figure 4.3: Distribution of $\tilde{\tau}$ for our particular case

We can also use our second simulation to see how an expansion of our 2×2 closed case may look. Instead of just sampling two (x, y) points, we sampled 20 and ran our simulation 1000 times. The plot of the distribution is centered at 0 as to be expected.

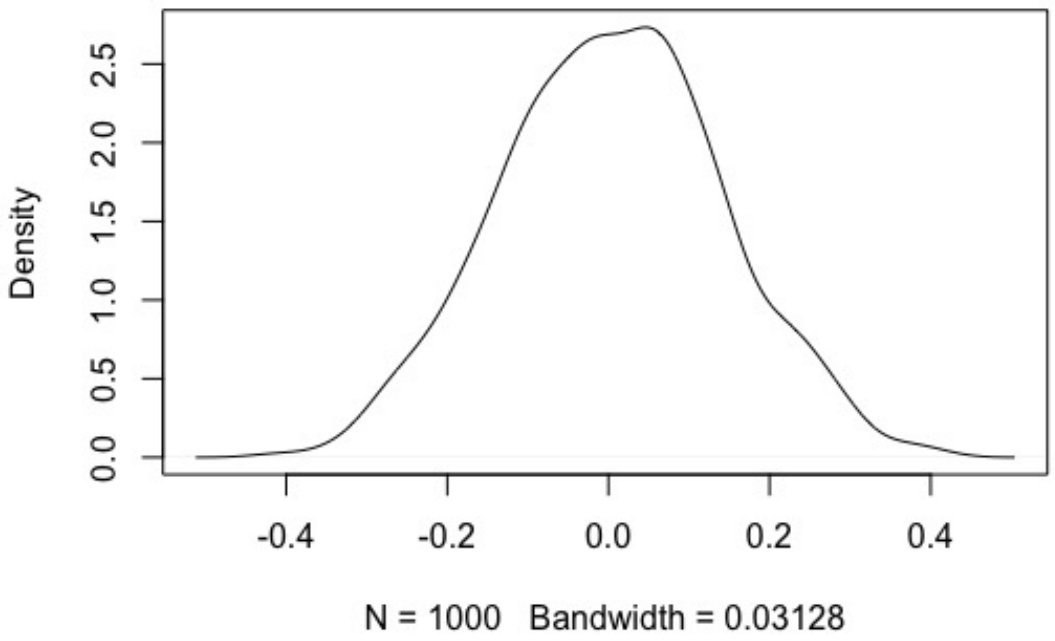

Figure 4.4: Distribution of $\tilde{\tau}$

We also found from this simulation that the standard deviation of $\tilde{\tau}$ is less than that of standard τ without noise.

	τ	$\tilde{\tau}$
sd	0.16156	0.13835

Table 4.3

4.3 Application of $\tilde{\tau}$ to Real Data

We took $\tilde{\tau}$ and applied it to the *Vibrio vulnificus* and dissolved organic carbon data from Looe Key, FL. We calculated $\tilde{\tau} = 0.12592$. Applying the method of daily averages resulted in $\tau = 0.13482$.

We then calculated the following from the original data and used it for our second simulation: *Vibrio* $\sim N(57.2, 41.8)$, DOC $\sim N(87.2, 16.4)$, $\varepsilon_{vib} \sim N(0, 39.0)$, and $\varepsilon_{DOC} \sim N(0, 5.0)$. We ran

our simulation to mimic the real data by generating three noise points for y and two for x. This resulted in the following distribution for $\tilde{\tau}$, based on 10,000 simulations.

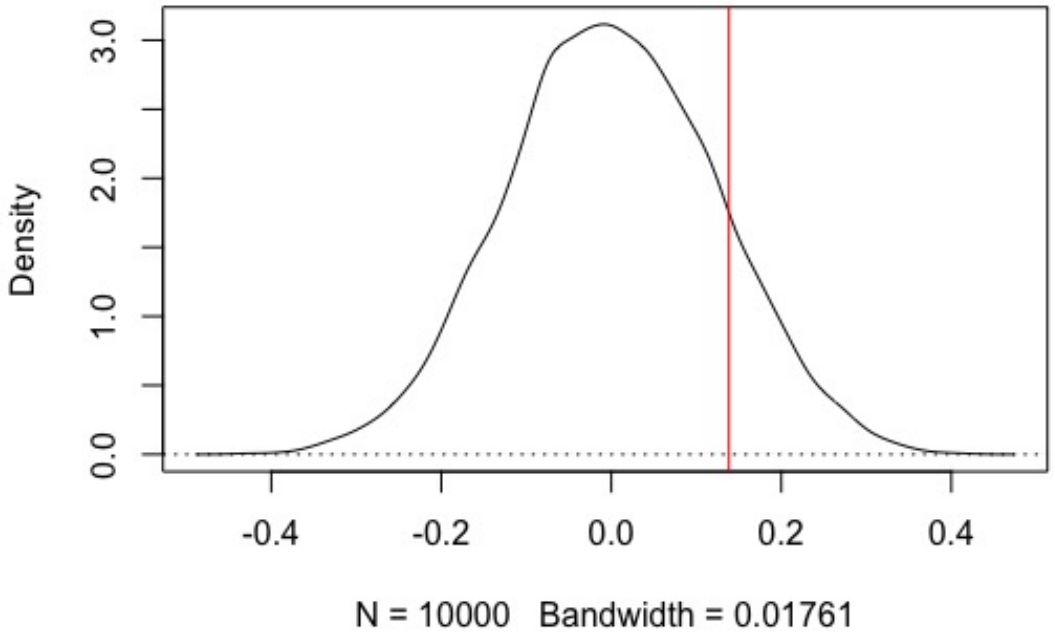

Figure 4.5: $\tilde{\tau}$ applied to real data

From the simulation, we can generate a "pseudo" p-value. To do this we let $\tilde{\tau}_i$ be the estimate of $\tilde{\tau}$ for simulation iteration $i = 1, ...10000$. Let $\hat{p} = \#(|\tilde{\tau}_i| > |\hat{\tilde{\tau}}|)$ where $\hat{\tilde{\tau}}$ is the actual sample p-value. Our two-tailed p-value for $\tilde{\tau}$ was .31. By using the `cor` function in R we were able to find a p-value for τ calculated with daily means and got a result of .38. So $\tilde{\tau}$ had a slightly stronger test, but the result still isn't powerful enough to reject the null hypothesis.

CHAPTER V: DISCUSSION AND CONCLUSIONS

5.1 Summary

From our results section, we can see that our analytical solutions for our 2×2 closed case with $x, y \sim N(0,1)$ and $\varepsilon_x, \varepsilon_y \sim N(0, 0.5)$ match our results from running our second simulation 100,000 times with the same parameters. This agreement of results confirms that both our analytical solution and our simulation are correct.

With our first simulation, we observed how the size of the noise added to original (x, y) points affects $\tilde{\tau}$. As the variation in the noise increases, the distribution of $\tilde{\tau}$ widens and moves its center towards 0. It is also notable that we observed the largest standard deviation for $\tilde{\tau}$ when the noise variance was closest to the variance of the original points.

When we applied $\tilde{\tau}$ to the real data from Looe Key, we found that the correlation found with $\tilde{\tau}$ was smaller than when calculated with daily averages. The p-value for $\tilde{\tau}$ was more powerful than what was obtained with the method of daily averages, but the result of the hypothesis test was the same.

Since $\tilde{\tau}$ varies with σ_ε, we can't use Kendall's assumption that all configurations are equally likely, and we have to know distributional information about both x and ε.

Future works could further explore how $\tilde{\tau}$ behaves when data is sampled from skewed distributions or what effects skewed noise could have. Future efforts could also be made to explore why the standard deviation of $\tilde{\tau}$ is at its greatest when the standard deviation of the noise is close to that of the data. Also, a way to avoid the need for distributional knowledge of our data is needed in order to make $\tilde{\tau}$ a more applicable option in the field.

REFERENCES

[1] D. R. Helsel. *Statistics for Censored Environmetal Data Using Minicab and R*. Wiley Series in Statistics in Practice. John Wiley & Sons, Inc., 2 edition, 2012.

[2] Maurice G. Kendall. *Rank Correlation Methods*. Hafner Publishing Company, 3^{rd} edition, 1962.

[3] H. Silverstone. A note on the cumulants of kendall's s-distribution. *Biometrika*, 37(3/4):231–235, 1950.

[4] Ajit C. Tamhane and Dorothy D. Dunlop. *Statistics and Data Analysis from Elementary to Intermediate*. Prentice-Hall, Inc., 1 edition, 2000.

[5] R Core Team. R: A language and environment for statistical computing, 2016.

APPENDIX A: R CODE

6.1 Sample Distributions

```
 % The following are the functions for the distributions for x, y,
 % ep to be used for the simulations
normal.0.10<-function(n){rnorm(n,0,10)}
normal.0.01<-function(n){rnorm(n,0,.01)}
normal.0.0.05<-function(n){rnorm(n,0,.5)}
normal.0.1<-function(n){rnorm(n,0,1)}
normal.1.1<-function(n){rnorm(n,1,1)}
```

6.2 First Simulation

```
tau.prob.func<-function(fx, fy, fepsx, fepsy, npts, nxnoise,
      nynoise, iterations) {
  % tau.prob.func has inputs of a distribution for x, a distribution for y, a
  % distribution for the noise in x, a distribution for the noise in y, total
  % number of (x, y) points, total number of noise points for x,
  % total number of noise points for y, and total number of iterations.
  % This function samples x and y points once, then adds different
  % randomly generated noise to the points in each iteration. Tilde
  % tau is calculated and stored for each iteration.

  X=fx(npts)
  Y=fy(npts)
  tau.without.noise <- cor(X, Y, method="kendall")
  tau<-numeric()
```

```
a<-numeric()
for(n in 1:(iterations)){

  % generate noise
  epsx=fepsx(npts*nxnoise)
  epsy=fepsy(npts*nynoise)

  % add the noise
  x= X+epsx
  y= Y+epsy
  x.mx=matrix(x, npts, nxnoise)
  y.mx=matrix(y, npts, nynoise)

  tau.with.noise.one.pair <- matrix(NA, npts, npts)
  for(i in 1:((npts)-1)){
    for(j in (i+1):npts){
      nxpos=0
      nypos=0

      for(k in 1:nxnoise){
        for(l in 1:nxnoise){
          nxpos=nxpos+sign(x.mx[i,k]-x.mx[j,l])

        }
      }
      for(k in 1:nynoise){
        for(l in 1:nynoise){
```

```
            nypos=nypos+sign(y.mx[i,k]-y.mx[j,l])

          }
        }
        % concordance
        tau.with.noise.one.pair[i, j] <- nxpos*nypos/nxnoise^2/nynoise^2

      }
    }


    tau[n] <- mean(tau.with.noise.one.pair, na.rm = T)

  }


  plot(density(tau), xlim=c(-.25, .25), main = "sigma=10")
  abline(v=tau.without.noise, col=2)

  return(list(tau.without.noise=tau.without.noise, tau=tau,
    mean.tau=mean(tau)))

}
```

```
many_tau.prob.func<-function(fx, fy, fepsx, fepsy, npts, nxnoise, nynoise,
iterations) {
  % many_tau.prob.func takes the same inputs as tau.prob.func.
  % In this function x and y values are resampled for each iteration,
  % noise is added to the values and tilde tau is calculated and recorded
  % each iteration.

  tau.without.noise <- numeric(iterations)
  tilde.tau<-numeric(iterations)
  for(n in 1:iterations){
    X=fx(npts)
    Y=fy(npts)
    tau.without.noise[n] <- cor(X, Y, method="kendall")
    % generate noise
    epsx=fepsx(npts*nxnoise)
    epsy=fepsy(npts*nynoise)

    %add the noise
    x= X+epsx
    y= Y+epsy
    x.mx=matrix(x, npts, nxnoise)
    y.mx=matrix(y, npts, nynoise)


    tau.with.noise.one.pair <- matrix(NA, npts, npts)
```

```
for(i in 1:((npts)-1)){
  for(j in (i+1):npts){
    nxpos=0
    nypos=0

    for(k in 1:nxnoise){
      for(l in 1:nxnoise){
        nxpos=nxpos+sign(x.mx[i,k]-x.mx[j,l])

      }
    }
    for(k in 1:nynoise){
      for(l in 1:nynoise){

        nypos=nypos+sign(y.mx[i,k]-y.mx[j,l])

      }
    }
    %concordance
    tau.with.noise.one.pair[i, j] <- nxpos*nypos/nxnoise^2/nynoise^2

  }
}



tilde.tau[n] <- mean(tau.with.noise.one.pair, na.rm = T)
```

```
  }

  return(list(tau.without.noise=tau.without.noise, tilde.tau=tilde.tau,
              mean.tilde.tau=mean(tilde.tau)))
}
```