Robustness of Kendall's Tau to Outliers in Sparse Non-zero Count Data

Stacey Hancock
Department of Mathematical Sciences
Montana State University

June 11, 2004

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

Robustness of Kendall's Tau to Outliers in Sparse Non-zero Count Data

Montana State University
Statistics Master's Writing Project
Stacey Hancock
John Borkowski, Advisor
April 2004

Introduction

Studies are often concerned with detecting trends in rare event count data over time. The statistics group at Pacific Northwest National Laboratory[1] (PNNL) is conducting one such study aimed at detecting trends over time in the frequency of events that endanger airplane flights. This study, run by the National Aviation Operation Monitoring Services (NAOMS), is part of the NASA Aviation Safety Program. Researchers survey one hundred to three hundred airline pilots each month, and ask each pilot how many times a certain event, such as running out of gas or engine failure, has happened during the past sixty days. The average rate of the event (number of occurrences per pilot) for each month is calculated. For each type of event, a trend analysis is done on the data. The researchers use Kendall's Tau, a correlation coefficient, to measure trends over time (in months) in the average rate of the event.

Many of the events measured in the NAOMS study have a small probability of occurrence, say .00001 or .001 per flight hour, and thus the majority of pilots have zero occurrences. However, there are a few pilots that report one, two, or more occurrences. Occasionally, a pilot may report a large number of occurrences (more than five standard deviations from the expected number of occurrences). This pilot would be considered an outlier. Depending on which month this pilot was surveyed, this outlier may cause the Kendall's Tau test for trend to show a significant trend in the rate over time when there is none (Type I error), or show no trend when there is one (Type II error). Thus, it is essential to determine how outliers affect Kendall's Tau estimates and p-values when the probability of occurrence is small.

A previous study determined that Kendall's Tau was sufficiently robust against a substantial number of outliers (Abdullah, 1990). However, the simulated data in the study were generated from a normal distribution with mean 5 and variance 1. In order to determine the robustness of Kendall's Tau to outliers in sparse non-zero count data, the analysis described in this paper was conducted on simulated count data rather than normal variates. The simulated data are not representative of the actual data used in the NAOMS study, and should not be interpreted as so. The explanatory variable is months,

[1] Pacific Northwest National Laboratory is located in Richland, WA, and is run by Battelle for the U.S. Department of Energy.

and the response variable is the mean rate of occurrence per month. The factors of interest that could influence Kendall's Tau estimates are the number of total months in the study, the number of pilots surveyed each month, the month in which the outlier occurred, and the probability (or rate) of the rare event occurring per hour. One thousand data sets with no trend over time were generated for each factor combination. The same was done with a positive linear trend in the response over time imposed on the data.

Background

Kendall's Tau (τ) is a commonly used nonparametric correlation coefficient which is a measure of the association between two quantitative variables. Like other correlation coefficients, it can take on values between –1 and 1; 0 indicating no association between variables, and –1 or 1 indicating perfect negative or positive association between variables, respectively. It is a useful alternative to Pearson's correlation coefficient because it does not require the two variables to have a bivariate normal distribution. It only requires that the data consist of a random sample and are measured on at least an ordinal scale.

First introduced by M. G. Kendall, Kendall's Tau is based on the ranks of the observations. It measures association by examining the number of concordant and discordant pairs of observations. A pair of observations, (x_1, y_1) and (x_2, y_2), is said to be concordant if the either x_1 is greater than x_2 and y_1 is greater than y_2, or x_1 is less than x_2 and y_1 is less than y_2. They are discordant if the difference between x_1 and x_2 is not in the same direction as the difference between y_1 and y_2. An estimate of τ is calculated from the data by the following formula:

$$\hat{\tau} = \frac{S}{n(n-1)/2}$$

where n is the number of observation pairs and S is the difference between the number of concordant pairs of X and Y values, and the number of discordant pairs of X and Y values. For the case when there are tied observations (in either or both variables), see Daniel (1997).

The estimate of Kendall's Tau can be used to conduct the hypothesis test H_0: X and Y are independent ($\tau = 0$) verses H_a: $\tau \neq 0$, or the one-sided alternatives, H_a: $\tau < 0$ or H_a: $\tau > 0$. If one of the variables is time, this test can be thought of as a test for trend over time. Critical values of the Kendall's Tau distribution, τ^*, can be found in Table A.22, Daniel (1997). For large sample sizes, the statistic

$$z = \frac{3\hat{\tau}\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

is approximately normally distributed with mean 0 and variance 1 under the null hypothesis.

Methods

Two simulations were run: one imposing a linear trend on the data, and one with no trend in the data. In each simulation, four factors were examined to determine their effect on the Kendall's Tau estimates. The factors and their levels are described in Table 1.

Table 1 – Factors

Factor Name	Factor Description	Levels
Pilots	Number of pilots surveyed per month	100 300
Monthtot	Number of total months in the survey	30 60
O_month	Month number containing outlier	Early month (10^{th} percentile) Middle month (50^{th} percentile) Late month (90^{th} percentile)
Rate	Rate of occurrence per hour	.00001 .001

The factor o_month has two codings for the three levels. While simulating data, we used the actual number of the month for the coding. Thus, for datasets with 30 months, the levels of o_month were 3, 15, and 27. For datasets with 60 months, the levels of o_month were 6, 30, and 54. For the analysis on the data, however, we used a coding that did not depend on the month total. Early month outliers were designated as 1, middle month as 2, and late month as 3. This coding was named o.month2.

The simulated data were created using SAS. For each of the 25 factor combinations, 1000 datasets were generated. For example, for the combination

pilots=300, monthtot=30, o.month2=1, and rate=.01, 1000 datasets were generated and an outlier was included in month 3 in each dataset. A Poisson distribution was assumed for the event counts and a Uniform(10, 100) distribution was assumed for the number of hours each pilot flew. In the no trend case, the mean of the Poisson distribution for each pilot was μ = rate*hours, where rate is specified by the factor level (.01 or .0001), and hours is a randomly generated value from a Uniform(10, 100) distribution. The value of the outlier was calculated by taking the next largest integer from $\mu+5*\text{sqrt}(\mu)$. Thus, outliers represented observations that were around 5 standard deviations above average. In the linear trend case, the mean for the Poisson distribution changed with time (months). The relationship between the Poisson rate and time used was Poisson rate = rate + rate*month, where rate is specified by the factor level and month is time variable (1 to 30 or 1 to 60). Thus, the mean of the Poisson distribution for each pilot was μ = Poisson rate * hours, and this mean increased with time. Outliers were calculated the same as in the no trend case.

Once the data were generated, the average rate of events per month (total events / total pilots) was calculated for each month. Two estimates of Kendall's Tau were calculated for each dataset using the average rates against time: one including the outlier and one with the outlier removed. The goal of the analysis was to examine the difference between the two Kendall's Tau estimates.

The difference between the two Kendall's Tau estimates was calculated for each dataset. Also, the 2-sided p-value for the null hypothesis that Kendall's Tau is equal to zero was calculated for each estimate. Because the sample size is large for each dataset (30 or 60 observations), the normal approximation to the Kendall's Tau distribution was used. Because the true Kendall's Tau value is zero for the no trend case, we expected that we would reject this hypothesis in 5% of the 24000 datasets with the outlier omitted if we used a significance level of 5%. When the outlier was included in the datasets, we were interested if the proportion of significant tests would differ significantly from 5%. For the linear trend case, we expected to reject this hypothesis in the datasets with the outlier removed. We were interested if the proportion of rejections for the no outlier datasets differed from the proportion of rejections for the outlier datasets.

Results

The results of the analysis are summarized in Tables 2 and 3. Table 2 gives the results with the linear trend in the data, and Table 3 gives the results with no trend in the data. The first four columns are the values of the factors that were considered. The remaining columns include the two average Kendall's Tau estimates for each factor combination, the difference between the two, and the two proportions of significant p-values for the Kendall's Tau test for trend.

For several of the factor combinations in the no trend case, some of the datasets generated had zero counts for all pilots due to small rates of occurrence. The correlation procedure in SAS did not calculate Kendall's Tau estimates for these datasets. Thus, for the datasets with all zero counts, we assumed the Kendall's Tau estimates were zero and the p-values were non-significant when calculating the proportion of significant p-values for the Kendall's Tau test for trend. The number of datasets where at least one pilot had a nonzero count is denoted by n in the tables.

Columns 6 through 14 are the mean value of the Kendall's Tau estimates with the outlier deleted for each factor combination (dmean), the mean value of the Kendall's Tau estimates with the outlier included for each factor combination (omean), and the difference between the two mean estimates (diff). Also included are the standard deviation of the 1000 Kendall's Tau estimates with the outlier deleted (dstd), the standard deviation of the 1000 Kendall's Tau estimates with the outlier included (ostd), and the standard deviation of the 1000 differences between the two estimates (diffstd). The proportion of p-values that were below .05 was calculated for the datasets with the outlier deleted (prop.dlin for linear trend and prop.dno for no trend), and with the outlier included (prop.olin for linear trend and prop.ono for no trend). The difference between the two proportions is denoted prop.diff.

In the linear trend data, the largest difference between Kendall's Tau estimates and proportions of p-values below .05 occurs when there are 100 pilots, 30 total months, a rate of .00001, and outlier month 1 or 3. When the outlier month is 1, the outlier tends to decrease the estimate of Kendall's Tau and fails to reject the hypothesis of trend more

often than when the outlier is deleted. When the outlier month is 3, the outlier tends to increase the estimate of Kendall's Tau and rejects the hypothesis of trend more often than when the outlier is deleted. For the other factor combinations, the outlier seems to have little effect on the Kendall's Tau estimates and the proportion of p-values below .05.

Outliers have more of an effect on the Kendall's Tau estimates when there is no trend in the data. The largest differences in the estimates occur when there are 30 total months, a rate of .00001, and outlier month 1 or 3. The number of pilots does not seem to have a large effect on the estimates. The largest differences in proportions of rejections occur for 100 pilots and a rate of .00001. These differences are comparable except for the factor combination of 100 pilots, 30 month, outlier month 1, and rate .0001. For this combination, the difference in proportions is .262, much larger than any other difference.

Table 2 - Linear Trend Analysis

pilots	monthtot	o.month2	rate	n	dmean	omean	diff	dstd	ostd	diffstd	prop.dlin	prop.olin	prop.diff
100	30	1	0.00001	1000	0.459	0.414	0.045	0.135	0.140	0.011	0.925	0.882	0.043
100	30	1	0.001	1000	0.977	0.977	0.000	0.007	0.007	0.000	1.000	1.000	0.000
100	30	2	0.00001	1000	0.450	0.443	0.007	0.132	0.132	0.012	0.928	0.922	0.006
100	30	2	0.001	1000	0.977	0.977	0.000	0.007	0.007	0.001	1.000	1.000	0.000
100	30	3	0.00001	1000	0.456	0.482	-0.027	0.129	0.122	0.019	0.939	0.968	-0.029
100	30	3	0.001	1000	0.977	0.977	0.000	0.007	0.007	0.001	1.000	1.000	0.000
100	60	1	0.00001	1000	0.593	0.581	0.011	0.074	0.075	0.003	1.000	1.000	0.000
100	60	1	0.001	1000	0.986	0.986	0.000	0.003	0.003	0.000	1.000	1.000	0.000
100	60	2	0.00001	1000	0.593	0.592	0.002	0.070	0.071	0.005	1.000	1.000	0.000
100	60	2	0.001	1000	0.986	0.986	0.000	0.003	0.003	0.000	1.000	1.000	0.000
100	60	3	0.00001	1000	0.590	0.598	-0.008	0.074	0.072	0.010	1.000	1.000	0.000
100	60	3	0.001	1000	0.986	0.986	0.000	0.003	0.003	0.000	1.000	1.000	0.000
300	30	1	0.00001	1000	0.670	0.654	0.015	0.091	0.094	0.005	1.000	1.000	0.000
300	30	1	0.001	1000	0.992	0.992	0.000	0.002	0.002	0.000	1.000	1.000	0.000
300	30	2	0.00001	1000	0.663	0.660	0.003	0.092	0.093	0.008	0.999	0.999	0.000
300	30	2	0.001	1000	0.992	0.992	0.000	0.002	0.002	0.000	1.000	1.000	0.000
300	30	3	0.00001	1000	0.667	0.676	-0.009	0.092	0.089	0.010	0.999	0.999	0.000
300	30	3	0.001	1000	0.992	0.992	0.000	0.002	0.002	0.000	1.000	1.000	0.000
300	60	1	0.00001	1000	0.785	0.782	0.003	0.041	0.041	0.001	1.000	1.000	0.000
300	60	1	0.001	1000	0.995	0.995	0.000	0.001	0.001	0.000	1.000	1.000	0.000
300	60	2	0.00001	1000	0.781	0.781	0.000	0.043	0.043	0.002	1.000	1.000	0.000
300	60	2	0.001	1000	0.995	0.995	0.000	0.001	0.001	0.000	1.000	1.000	0.000
300	60	3	0.00001	1000	0.785	0.787	-0.002	0.042	0.042	0.004	1.000	1.000	0.000
300	60	3	0.001	1000	0.995	0.995	0.000	0.001	0.001	0.000	1.000	1.000	0.000

Table 3 - No Trend Analysis

pilots	monthtot	o.month2	rate	n	dmean	omean	diff	dstd	ostd	diffstd	prop.dno	prop.ono	prop.diff
100	30	1	0.00001	809	0.007	-0.177	0.162	0.186	0.141	0.054	0.156	0.418	-0.262
100	30	1	0.001	1000	-0.003	-0.036	0.033	0.191	0.190	0.011	0.185	0.182	0.003
100	30	2	0.00001	803	-0.004	-0.010	0.006	0.182	0.131	0.039	0.148	0.053	0.095
100	30	2	0.001	1000	0.004	0.003	0.002	0.188	0.187	0.005	0.190	0.186	0.004
100	30	3	0.00001	819	-0.006	0.163	-0.150	0.173	0.132	0.050	0.128	0.181	-0.053
100	30	3	0.001	1000	0.002	0.032	-0.030	0.181	0.179	0.011	0.162	0.168	-0.006
100	60	1	0.00001	960	0.002	-0.094	0.093	0.134	0.116	0.030	0.221	0.290	-0.069
100	60	1	0.001	1000	-0.006	-0.023	0.016	0.134	0.134	0.005	0.215	0.212	0.003
100	60	2	0.00001	950	0.008	0.005	0.003	0.127	0.107	0.020	0.178	0.109	0.069
100	60	2	0.001	1000	-0.006	-0.006	0.000	0.129	0.128	0.002	0.179	0.172	0.007
100	60	3	0.00001	961	0.000	0.093	-0.089	0.131	0.112	0.030	0.205	0.275	-0.070
100	60	3	0.001	1000	0.000	0.016	-0.016	0.126	0.126	0.005	0.169	0.176	-0.007
300	30	1	0.00001	994	0.001	-0.116	0.116	0.185	0.167	0.039	0.184	0.235	-0.051
300	30	1	0.001	1000	-0.010	-0.029	0.019	0.186	0.186	0.007	0.185	0.193	-0.008
300	30	2	0.00001	996	0.005	0.000	0.005	0.187	0.169	0.022	0.199	0.150	0.049
300	30	2	0.001	1000	-0.009	-0.009	0.001	0.186	0.185	0.003	0.171	0.170	0.001
300	30	3	0.00001	990	0.003	0.110	-0.106	0.187	0.167	0.038	0.190	0.200	-0.010
300	30	3	0.001	1000	0.005	0.023	-0.018	0.181	0.180	0.006	0.171	0.169	0.002
300	60	1	0.00001	1000	-0.003	-0.060	0.057	0.136	0.129	0.014	0.210	0.237	-0.027
300	60	1	0.001	1000	-0.002	-0.012	0.009	0.132	0.132	0.003	0.184	0.182	0.002
300	60	2	0.00001	1000	0.004	0.003	0.001	0.126	0.119	0.008	0.158	0.139	0.019
300	60	2	0.001	1000	-0.003	-0.003	0.000	0.139	0.139	0.001	0.226	0.224	0.002
300	60	3	0.00001	1000	0.002	0.056	-0.054	0.127	0.121	0.014	0.186	0.200	-0.014
300	60	3	0.001	1000	0.002	0.011	-0.009	0.125	0.125	0.003	0.159	0.165	-0.006

A significance level of .05 was used for each Kendall's Tau test for trend. Thus, if there were no trend in the data, we should reject the null hypothesis 5% of the time. However, in the no trend analysis, the proportions of tests that rejected the null hypothesis where much larger than .05. This is most likely due to the normal approximation that was used to calculate the p-values. The normal approximation used in this study does not adjust the variance formula of Kendall's Tau for ties. Further study should be conducted to explore the limitations of the normal approximation in this type of data.

Linear models were fit to the linear trend data and the no trend data using diff as the response and the four factors as explanatory variables. All interactions between factors were included in the models. The analysis of variance tables for these linear models are given in the following output.

Linear Trend Data:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pilots	1	0.01695	0.01695	420.72	< 2.2e-16
monthtot	1	0.03428	0.03428	850.73	< 2.2e-16
o.month2	2	0.91073	0.45537	11301.35	< 2.2e-16
rate	1	0.06585	0.06585	1634.18	< 2.2e-16
pilots:monthtot	1	0.00769	0.00769	190.82	< 2.2e-16
pilots:o.month2	2	0.24369	0.12184	3023.92	< 2.2e-16
pilots:rate	1	0.01562	0.01562	387.75	< 2.2e-16
monthtot:o.month2	2	0.31329	0.15664	3887.61	< 2.2e-16
monthtot:rate	1	0.03277	0.03277	813.18	< 2.2e-16
o.month2:rate	2	0.90464	0.45232	11225.73	< 2.2e-16
pilots:monthtot:o.month2	2	0.06799	0.03400	843.73	< 2.2e-16
pilots:monthtot:rate	1	0.00713	0.00713	177.03	< 2.2e-16
pilots:o.month2:rate	2	0.24086	0.12043	2988.85	< 2.2e-16
monthtot:o.month2:rate	2	0.31085	0.15543	3857.37	< 2.2e-16
pilots:monthtot:o.month2:rate	2	0.06695	0.03348	830.82	< 2.2e-16
Residuals	23976	0.96607	0.00004		

No Trend Data:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pilots	1	4.792e-04	4.792e-04	0.9291	0.3351
monthtot	1	0.026	0.026	51.2434	8.402e-13
o.month2	2	53.360	26.680	51729.5526	< 2.2e-16
rate	1	0.044	0.044	84.8288	< 2.2e-16
pilots:monthtot	1	1.978e-05	1.978e-05	0.0384	0.8447
pilots:o.month2	2	1.728	0.864	1675.6662	< 2.2e-16
pilots:rate	1	9.940e-07	9.940e-07	0.0019	0.9650
monthtot:o.month2	2	4.296	2.148	4164.9456	< 2.2e-16
monthtot:rate	1	0.011	0.011	21.8251	3.003e-06
o.month2:rate	2	26.943	13.472	26120.0094	< 2.2e-16
pilots:monthtot:o.month2	2	0.036	0.018	34.4655	1.132e-15
pilots:monthtot:rate	1	1.675e-04	1.675e-04	0.3248	0.5688
pilots:o.month2:rate	2	0.807	0.403	782.2922	< 2.2e-16
monthtot:o.month2:rate	2	2.168	1.084	2101.5358	< 2.2e-16
pilots:monthtot:o.month2:rate	2	0.003	0.002	3.1494	0.0429
Residuals	23258	11.995	0.001		

The extremely small p-values are most likely due to the large sample size. However, we can still gain some information by looking at the analysis of variance tables. For example, by examining the mean square values, it is obvious that the outlier month and interactions involving the outlier month have the most influence on the differences between Kendall's Tau estimates. Rate also has a large influence on the differences.

Tables 4 and 5 give the marginal means for each factor. In the linear analysis, there is a large difference between the Kendall's Tau estimates for rate .00001 and rate

.001. For the smaller rate, the Kendall's Tau estimates are much smaller than for the larger rate. Also, Kendall's Tau estimates for outlier month 1 decrease when the outlier is included. Kendall's Tau estimates for outlier month 3 increase when the outlier is included. In the no trend analysis, the most striking differences occur for outlier months 1 and 3. For outlier month 1, the Kendall's Tau estimate including the outlier is lower than the estimate without the outlier. For outlier month 3, the Kendall's Tau estimate including the outlier is higher than the estimate without the outlier.

Table 4 - Sample Means by Factor for Linear Analysis

Factor Name and Level	Variable Name			
	dmean	omean	prop.dlin	prop.olin
pilots = 100	0.75259	0.75005	0.98267	0.98100
pilots = 300	0.85938	0.85852	0.99983	0.99983
monthtot = 30	0.77267	0.76977	0.98250	0.98083
monthtot = 60	0.83930	0.83880	1.00000	1.00000
o.month2 = 1	0.80718	0.79785	0.99063	0.98525
o.month2 = 2	0.80479	0.80326	0.99088	0.99013
o.month2 = 3	0.80599	0.81175	0.99225	0.99588
rate = .00001	0.62423	0.62088	0.98250	0.98083
rate = .001	0.98774	0.98770	1.00000	1.00000

Table 5 - Sample Means by Factor for No Trend Analysis

Factor Name and Level	Variable Name			
	dmean	omean	prop.dno	prop.ono
pilots = 100	-0.00009	-0.00284	0.17800	0.20183
pilots = 300	-0.00043	-0.00043	0.18525	0.18867
monthtot = 30	-0.00040	-0.00390	0.17242	0.19208
monthtot = 60	-0.00012	-0.00111	0.19083	0.19842
o.month2 = 1	-0.00188	-0.06831	0.19250	0.24363
o.month2 = 2	-0.00002	-0.00226	0.18113	0.15038
o.month2 = 3	0.00112	0.06305	0.17125	0.19175
rate = .00001	0.00159	-0.00224	0.18025	0.20725
rate = .001	-0.00211	-0.00278	0.18300	0.18325

Conclusions

The estimate of Kendall's Tau is sufficiently robust to outliers in sparse non-zero count data when the data follow a linear trend. However, when there is no trend in the data, outliers can effect the estimate of Kendall's Tau and the p-value of the Kendall's Tau test for trend quite significantly, especially for low rates of occurrence and outliers that have high leverage.

The Kendall's Tau test for trend is more sensitive to outliers than the estimate of Kendall's Tau. For low rates and outliers with high leverage, outliers can affect the p-value significantly. In addition, although the sample size was large enough to safely use the normal approximation, the normal approximation did not perform well with or without outliers in the data. Further study should be done to determine how well the normal approximation approximates the Kendall's Tau distribution with sparse non-zero data.

References

Abdullah, Mokhtar Bin. (1990). On a robust correlation coefficient. *The Statistician*, 39(4), 455-460.

Borkowski, John J. (2003). Course Notes for Statistics 530: Nonparametrics and Resampling Methods, *Unpublished Manuscript*, Montana State University, 178-196.

Daniel, Wayne W. (1997). *Applied Nonparametric Statistics*, (2nd ed.) (pp. 365-375, 579). Wadsworth Publishing.

Farlie, D. J. G. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika*, 47(3/4), 307-323.

Fieller, E. C.,Hartley, H. O. and Pearson, E. S. (1957). Tests for rank correlation coefficients. I. *Biometrika*, 44, (3/4), 470-481.

Gideon, Rudy A. and Hollister, Robert A. (1987). A rank correlation coefficient resistant to outliers. *Journal of the American Statistical Association*, 82(398), 656-666.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2), 81-93.