
Stockholms universitet

# Statistical Properties of Equity Transaction Data

Gunnar Höglund

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2011:7
Matematisk statistik
Juni 2011

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2011:7**
http://www.math.su.se

# Statistical Properties of Equity Transaction Data

Gunnar Höglund*

June 2011

**Abstract**

This Master's thesis examines arrival times between trades of shares in the Microsoft stock during the year 2000 from a descriptive statistical perspective. Instead of ignoring transactions that occur during the same second we develop tools for handling floored stochastic processes. We conclude that the arrival times in short sequences of intraday trades are iid and likely from a scalable distribution. The gamma distribution is shown to fit the arrival times well but we can not statistically prove that this is the case.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: gunnar.hoglund@gmail.com. Supervisor: Joanna Tyrcha.

# STATISTICAL PROPERTIES OF EQUITY TRANSACTION DATA

GUNNAR HÖGLUND

Abstract. This Master's thesis examines arrival times between trades of shares in the Microsoft stock during the year 2000 from a descriptive statistical perspective. Instead of ignoring transactions that occur during the same second we develop tools for handling floored stochastic processes. We conclude that the arrival times in short sequences of intraday trades are *iid* and likely from a scalable distribution. The gamma distribution is shown to fit the arrival times well but we can not statistically prove that this is the case.

## Contents

## 1. Introduction

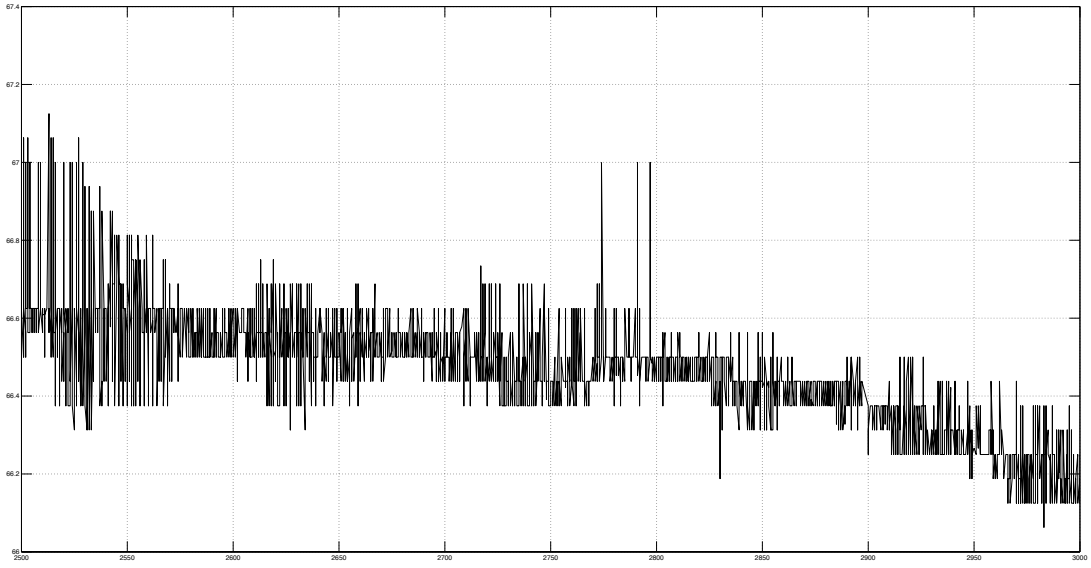
1.1. **Equity prices.** Equity prices must be one of the most well researched topics in finance. We could describe the equity price process in the following simple manner: Bid and ask prices arrive at the exchange (together with a volume) at certain time points. When the prices meet a trade occurs. A natural way to model this would be to model the bid and ask process by themselves as continuous time point processes with discrete price jumps. Another way is to only look at the actual trading times and model them in an analogous fashion. All of this has indeed been done, for example by Engle and Russell (1998) and Engle (2000) who pioneered the field ten years ago with their *Autoregressive Conditional Duration* model. Our study will begin from the same starting point Engle and Russel started at, but the approach is quite different.

In this paper we will only focus on the times between the trades and we will do this from a pure descriptive, statistical perspective. We are voluntarily ignoring loads of information when we do this, the arrival of bid and ask prices, the volume, the price change and how all of this effects the time to the next transaction.

1.2. **The data.** We will examine the times between trades in the Microsoft share on the Nasdaq stock exchange during the year 2000. The data is ordered and time stamped to the second by the exchange. We will ignore trades that take place outside regular trading hours but no other trades are removed. If two trades occur the same second the arrival time between them will be 0 seconds in our eyes - this differs from the approach taken by Engle and Russel that only consider arrival times larger than 0.

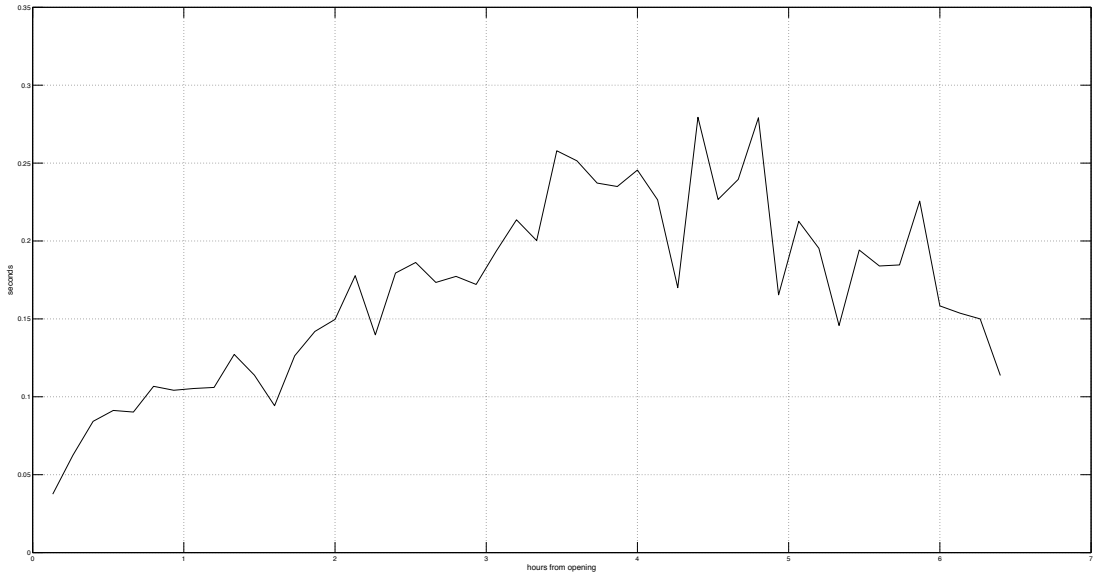
Figure 1 illustrates the trades of the Microsoft share during 500 seconds in the opening hour of April 24, 2000. The price is on the $y$-axis and the $x$-axis illustrates the number of seconds from opening. When the stock market opened after the easter holiday the

Figure 1



share price pummeled and fell close to 15% on opening. Like most other tech companies of the time, the share proved volatile intraday and the difference between the highest traded price and the lowest equaled 5 USD. During the 500 seconds depicted in the chart 4495 trades occurred, 2.71% of the total number of trades that took place during the day. The time around opening and close is usually the most intensive period both when it comes to volume (high) and the time between trades (short). Figure 2 illustrates the average arrival times between trades in the stock during 8 minute intervals the same day. Time between trades is on the $y$-axis and the time of day is on the $x$-axis. The

FIGURE 2




average arrival time the first 8 minutes this particular day was a bit below 0.05 seconds, it later rose to above 0.25 seconds around mid-day and then declined to just above 0.1 seconds before close. This non-stationarity is always an issue when analyzing intraday equity data and we will handle it somewhat differently than the researchers mentioned above.

Before we start to study the actual data we will begin with some general results that will be employed in the analysis.

### 1.3. **Notation.**

**1.** $U(0,1)$ - The uniform distribution on the unit interval.
**2.** $N$ - The natural numbers $1, 2, 3, ....$
**3.** $\{\}$ - Used both to denote *sets* and the *fractional part* of a number.

## 2. The floored renewal process

Let $(S_n)$ constitute a renewal process with arrival times $T_i$ from some continuous distribution[1]. If we round the renewal epochs down to the closest integer we receive the related *floored renewal process* $(\Sigma_n)$ with arrival times $\Delta_i = [S_i] - [S_{i-1}]$. Since $S_0 = 0$ the first arrival time $\Delta_1 = [T_1]$. For $i \in N$,

$$\Delta_{i+1} = [S_{i+1}] - [S_i] = [S_i + T_{i+1}] - [S_i] = [T_{i+1} + U_i], \tag{2.1}$$

where $U_i = S_i - [S_i]$ is the fractional part of $S_i$. $T_{i+1}$ is independent of $U_i$, but $U_{i+1}$ is not - so by flooring the renewal epochs both the independence and the equal distribution of the arrival times are lost. Fortunately the dependence between the arrival times declines with higher expected value and the arrival times are equally distributed in the limit.

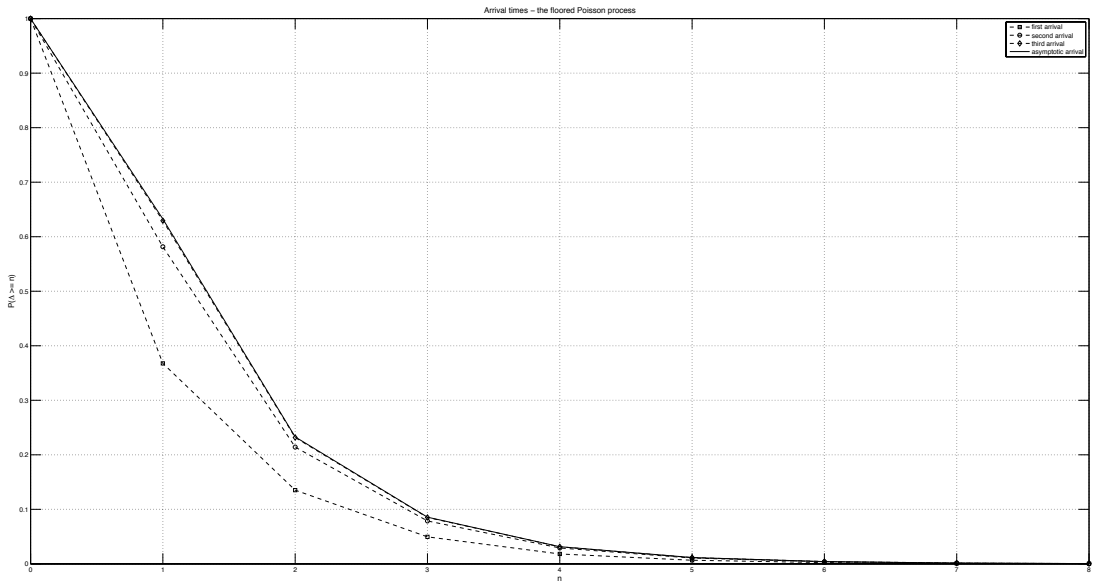
**Lemma 2.1.** *Let $U \sim U(0,1)$, then the fractional part of $S_i$ converges in distribution to $U$, $U_i \xrightarrow{d} U$ as $i \to \infty$.*

*Proof.* The result is intuitively clear - a proof can be found in Robbins (1953). □

Let $\Delta = \lim_{i\to\infty} \Delta_i$ denote the asymptotic arrival time in the floored renewal process $(\Sigma_n)$, then $\Delta \stackrel{d}{=} [T + U]$ with obvious notation.

**Example** Figure 3 depicts the tails of the distribution of arrival times in a floored Poisson process[2] ($\lambda = 1$). The first arrival is a floored exponential variable and the tail assumes the same values as its continuous analogue on the non-negative integers.

Figure 3




[1]Let $\phi(t)$ denote the characteristic function of the $T_i$'s. Formally we require that $\phi(t) = 1$ *iff* $t = 0$.

[2]The tails of arrival 1,2 and 3 are based on $10^6$ generated exponential variables, the asymptotic tail is numerically calculated.

The second arrival time displays some quite different properties compared to the first one, the 0-arrival time is not as frequent for example. The third arrival time is very hard to set apart from the asymptotic so the convergence in this case is remarkably quick. As the intensity increases the convergence naturally becomes slower and for a floored Poisson process with $\lambda = 1/31$ it takes closer to 200 arrivals (6.5 seconds).

The following lemma shows the connection between the distribution of $\Delta$ and the distribution of $T$.

**Lemma 2.2.** *For $n \in N$ holds,*

$$P(\Delta \geq n) = \int_{n-1}^{n} P(T \geq t)dt$$

*Proof.*

$$P(\Delta \geq n) = P([T+U] \geq n) = P(T+U \geq n), \text{ since } n \in N.$$

The result follows from straightforward calculations using the fact that $U$ is uniformly distributed. □

Since $P(T \geq t)$ is non-increasing in $t$,

$$P(T \geq n) \leq P(\Delta \geq n) \leq P(T \geq n-1)$$

but fortunately this does not change the expected value.

**Corollary 2.3.** *The expected value of the asymptotic arrival times in the floored renewal process equals the expected value of the arrival times in the original process*

*Proof.*

$$E\Delta = \sum_{n=1}^{\infty} P(\Delta \geq n) = \sum_{n=1}^{\infty} \int_{n-1}^{n} P(T \geq t)dt = \int_{0}^{\infty} P(T \geq t)dt = ET$$

□

One important consequence is that the sample mean of the asymptotic arrival times in a floored process is a consistent estimate of the expected value of the arrival times in the original process. It also simplifies the proof of the following Lemma.

**Lemma 2.4.** *Let $\Delta$ denote the asymptotic arrival time in the floored renewal process with underlying arrival times $T$. Then,*

$$\mathrm{Var}(\Delta) = \mathrm{Var}(T) + E\left(\{T\}(1-\{T\})\right).$$

*Proof.* Let $F$ denote the distribution of $T$. Due to Corollary 2.3

$$\mathrm{Var}(\Delta) - \mathrm{Var}(T) = \int_{0}^{1} \int_{0}^{\infty} ([t+u]^2 - t^2)dF(t)du. \tag{2.2}$$

Splitting up $t$ into its integral part $n = [t]$ and fractional part $\theta = \{t\}$ we can rewrite the integrand

$$\begin{cases} n^2 - (n+\theta)^2, & \text{for } u < 1-\theta, \\ (n+1)^2 - (n+\theta)^2, & \text{for } u \geq 1-\theta, \end{cases} = \begin{cases} -2n\theta - \theta^2, \\ 2n+1-2n\theta-\theta^2. \end{cases}$$

Hence (2.2) equals

$$\int_0^\infty \left( \int_0^{1-\theta} (-2n\theta - \theta^2) du + \int_{1-\theta}^1 (2n + 1 - 2n\theta - \theta^2) du \right) dF(t).$$

Simplifying the expression completes the proof. □

So the rounding of a renewal process increases the variance of the arrival times but only up to 0.25 since $\{T\}(1 - \{T\}) \leq 0.25$. Lemma 2.4 also gives an approximation of the variance when $\mu = ET$ is small and the distribution of $T$ is well behaved.

**Corollary 2.5.** *If $T < 1$ a.a. then*

$$\mathrm{Var}(\Delta) = \mu - \mu^2.$$

**Summary** When observing a floored process the observed distribution will differ from the underlying one but the expected value (of the asymptotic arrival times) is unchanged. The variance of the asymptotic arrival times is larger than the underlying variance - the relative difference decreases with higher variance. There is a dependence between the arrival times in the floored process - the magnitude depends on the precision loss introduced by the rounding. The arrival times are identically distributed in the limit - the speed of the convergence depends on the expected value. When analyzing rounded data as described above one needs to be aware of the consequences of the precision loss - fitting and scaling of distributions is much more intricate.

## 3. Statistical properties of the data

That the intensity of the arrival times vary during the day complicates the analysis. One common way to handle this problem of non-stationarity is to try to normalize the arrival times[3] but this, as we shall see on page 16 actually risks to distort the data in a non-trivial way. Our approach will instead be to split up every day into several shorter sequences $\hat{\Delta}_i$ and conduct the analysis on these instead. Each day will be split up into subsequences of 31, 51 and 200 observations.

Assume that we split every day into consecutive sequences of $n$ observations and let $D$ be the collection of arrival time sequences, then

$$D = (\hat{\Delta}_1, \hat{\Delta}_2, ..., \hat{\Delta}_m), \text{ where } \hat{\Delta}_i = (\Delta_1^i, \Delta_2^i, ..., \Delta_n^i)$$

are the subsequences of observed arrival times. If the number of observations a given day is not a multiple of $n$ the last observations will be discarded. This way, all sequences $\hat{\Delta}_i$ will have the same number of observations $n$ and will only contain arrival times between trades that occur the same day. Let $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ denote the sample mean and variance of the $i$:th sequence $\hat{\Delta}_i$

$$\hat{\mu}_i = \sum_j \Delta_j^i / n, \;\; \hat{\sigma}_i^2 = \sum_j (\Delta_j^i - \hat{\mu}_i)^2 / n$$

To avoid making the notation to burdensome the $n$ is left out and during the text the context decides whether for example $\hat{\Delta}_i$ is a sequence of 31, 51 or 200 observations.

[3] Engle (2000) for example *diurnally adjusts* the data by regressing the arrival times on the time of day using a piecewise linear spline and then normalizing the arrival time that occur at $t$ with the value of the spline at $t$.

Another word on notation, we will use the word length to refer to the time length of a sequence i.e. the length of $\hat{\Delta}_i$ equals $\Sigma_j \Delta_j^i$.

So what do these sequences look like and what properties do they possess?

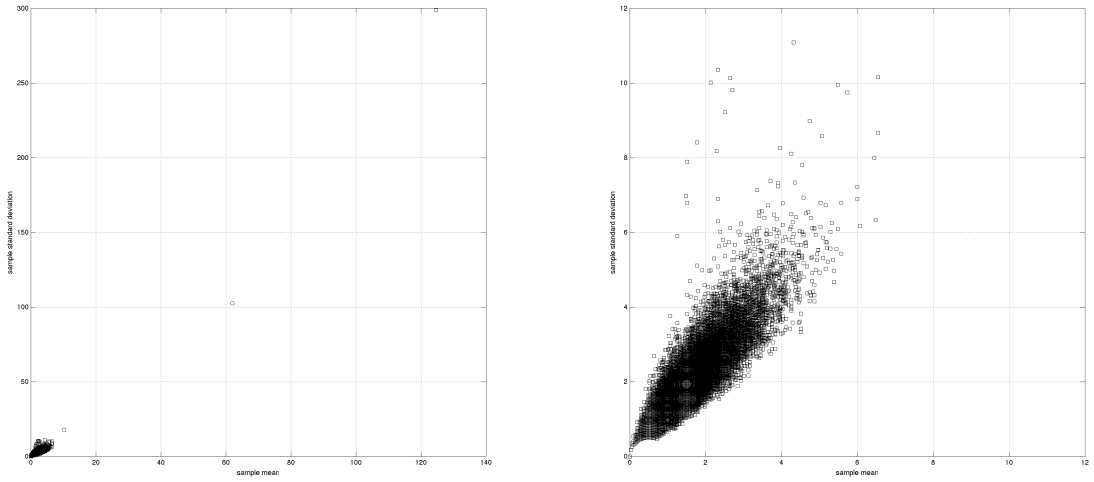
**Remark** Among the sequences with 31 observations, 1186 have a total (time) length 0. Let $\hat{\Delta}_i$ be one of these, then

$$\Sigma_j \Delta_j^i = 0, \hat{\mu}_i = 0 \text{ and } \hat{\Delta}_i = (0, 0, ..., 0).$$

If $\hat{\Delta}_i$ would have been one of the second shortest sequences instead, then $\hat{\Delta}_i$ would have consisted of 30 zeros and 1 one, $\hat{\mu}_i = 1/31$ and $\hat{\sigma}_i^2 = \hat{\mu}_i - \hat{\mu}_i^2 \approx \hat{\mu}_i$. There are 6419 sequences of length 1. The most common length of the sequences is 8 seconds (13052 sequences) and the longest observed sequence is 3866 seconds long. In theory, the possible number of variations of $\hat{\Delta}_i$ obviously increases dramatically with $\hat{\mu}_i \in \{k/31 \colon k = 0, ..., 3866\}$.

A second look at the data is provided in Figure 4, the sample means are plotted versus the sample standard deviations for $n = 31$. The chart to the left is the whole material, the chart to the right is zoomed in close to origo. Just by examining the charts we can

FIGURE 4




spot an obvious, almost linear relationship between $\hat{\mu}$ and $\hat{\sigma}$.

3.1. **Test of independence.** We know that the flooring of a renewal process in theory induces a dependence and that this effect increases with higher precision loss. This is however always a problem in the analysis of non-discrete stochastic processes where only the epochs are known and we can assume that the effect should be small unless the expected arrival time is very short.

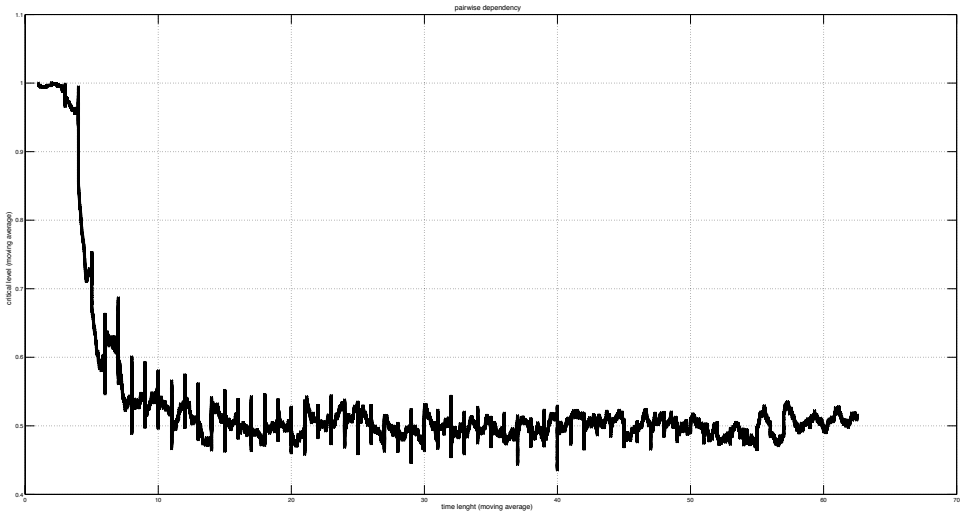
**Null hypothesis.** The following tests will test the null hypothesis that two subsequent arrival times in a short intraday sequence are independent.

To test the null hypothesis of pairwise independence of two consecutive arrival times we will create a contingency table from each $\hat{\Delta}_i$ and then conduct Fisher's exact test[4] extended to contingency tables larger than $2 \times 2$ [5]. Similarly to Martin-Löf (1974) we will use *critical level* to denote the $p$-value. The exact test is done in $R$ using $fisher.test()$.

**Sequences of 31 observations.** There are in total 313272 sequences $\hat{\Delta}_i$ with 31 observations. The shortest 1186 sequences only contain 0-arrival times and therefore are not tested - with the observed precision the dependence in these sequences is total. The last 2.4% is not tested due to time restrictions, 0.1% of the sequences are skipped due to problems with the underlying algorithm[6]. The remaining 97.5% of the sequences are tested successfully.

Result: 4.1% of the critical levels $\epsilon$ are smaller than 5%, 0.8% are smaller than 1% and 0.08% are smaller than 0.1%. To get an overview of the critical levels and see if they vary with the length of the sequences we plot the 500 moving average of the critical levels versus the average time length of the underlying sequences. Judging from the plot

FIGURE 5




the average critical level stabilizes around 50% for $\hat{\mu}_i > 0.3$ seconds. This is what to be expected if the null hypothesis is true since this would make $\epsilon$ the outcome of a $U(0,1)$ variable. The average critical level does not decrease with increasing expected value and there are therefore no reasons to expect that the longest 2.4% not tested would show a dependence. The dependence of the arrival times introduced by the flooring can not be detected by the test, on the contrary two consecutive arrival times from a low $\hat{\mu}$ sequence are considered totally independent and it is fair to assume that the precision error in these sequences is too large to effectively test the sequences.

[4]See Fisher (1934).

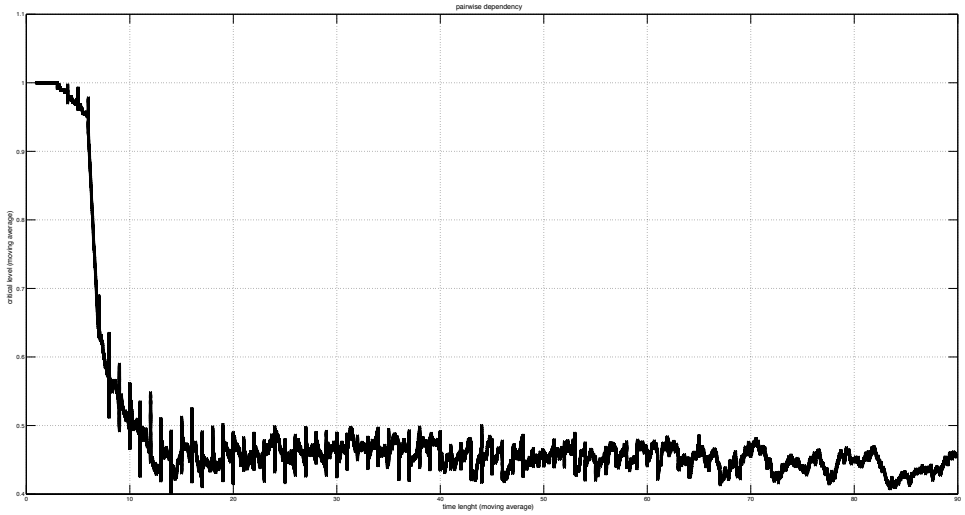
[5]See for example Martin-Löf (1970) and Martin-Löf (1974).

[6]The $R$ function is based on the FORTRAN subroutine FEXACT.

**Sequences of 51 observations.** There are 190368 sequences $\hat{\Delta}_i$ with 51 observations. The shortest 181 sequences only contain zero-arrival times and therefore are not tested. The longest 3% is not tested due to time restrictions.

Result: 6.3% of the critical levels are smaller than 5%, 1.6% are smaller than 1% and 0.2% are smaller than 0.1%. Judging from the plot the critical levels soon stabilize

FIGURE 6




around 0.45.

**Summary** We can not reject the null hypothesis in the case of 31 observation sequences. Traces of dependence can be detected in the 51 observation sequences - we therefore do not proceed to test the sequences of 200 observations.

3.2. **Test of homogeneity.** Continuing our process of describing the data the next step is to check whether similar sequences $\hat{\Delta}_i$ can be considered to be drawn from the same underlying distribution. We will proceed in the following fashion. Start by denoting the unique sample means by

$$\{\check{\mu}_1, ..., \check{\mu}_l\} = \{\hat{\mu}_i \colon i = 1, ..., m\}$$

and define

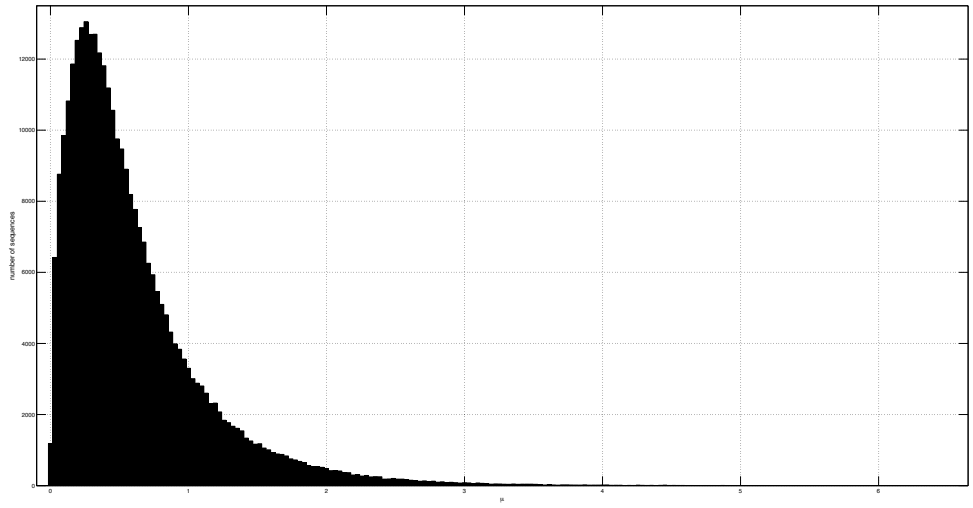
$$D_k = (\hat{\Delta}_i)_{i \colon \hat{\mu}_i = \check{\mu}_k}, \text{ for } k = 1, .., l.$$

$D_k$ is then the collection of all the sequences $\hat{\Delta}_i$ such that $\hat{\mu}_i = \check{\mu}_k$. Figure 7 illustrates the relationship between $\check{\mu}_k$ and the number of sequences in $D_k$ (for $n = 31$) excluding the 16 datasets $D_k$ that only contain one sequence each.

**Null hypothesis.** Let $T_1$ and $T_2$ be two arrival times between trades with the same expected value, then

$$H_0 \colon P(T_1 > t) = P(T_2 > t).$$

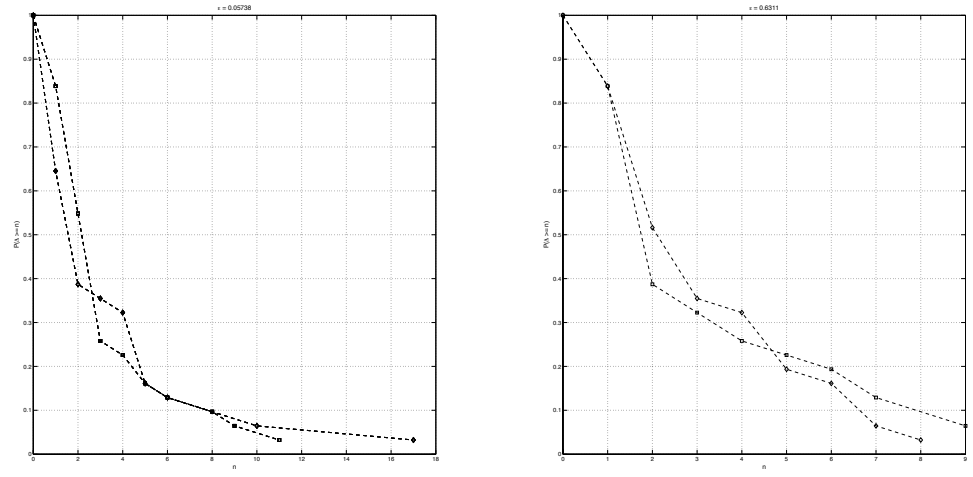
FIGURE 7




To test the null hypothesis of homogeneity we will conduct an exact test of a reductive hypothesis in the Per Martin-Löf sense[7]. This is a general formulation of Fisher's before mentioned test. The drawback of the exact tests used to be the computing power required - but unless we are dealing with very large sets of data this is not a problem anymore. The primary advantage (apart from the exactness) is the reduced need for grouping.

**Example** Figure 8 illustrates two pairs of distribution tails, all four with mean $77/31 \approx 2.48$.

FIGURE 8



[7]See Martin-Löf (1970) or Martin-Löf (1974).

Table 1 illustrates the data in the right chart above; one of the sequences ($a$) has 5 zero observations and 10 one observations, the other one ($b$) also has 5 zeros but 14 ones. Let $X$ be the underlying discrete sample space, $x \in X$ and let $t$ denote the sufficient statistic consisting of the two first rows $t(x) = (5, 10, ..., 0, 5, 14, ..., 2)$. Let $u$ denote the simpler statistic assuming values equal to the column sums, $u(t) = (10, 24, ..., 2)$. Keeping both

TABLE 1

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\Delta}_a$ | 5 | 10 | 5 | 1 | 4 | 1 | 3 | 1 | 1 | 0 |
| $\hat{\Delta}_b$ | 5 | 14 | 2 | 2 | 1 | 1 | 2 | 2 | 0 | 2 |
| Total | 10 | 24 | 7 | 3 | 5 | 2 | 5 | 3 | 1 | 2 |

row sums fixed equal to 31, the (reductive) hypothesis that $t$ can be reduced to $u$ is the hypothesis of homogeneity and is in this case rejected by the test if the number of outcomes that realize the observed value of $t(x)$ is small compared to the number of outcomes that realize the observed value of $u(t(x))$.

The exact homogeneity test on both pairs generates the critical levels (from left to right) $\epsilon_1 = 0.05738$ and $\epsilon_2 = 0.6311$. In the second case we can clearly not reject the hypothesis of equal distribution whereas in the first case we can reject it on the 10% level, but not on the 5% level.

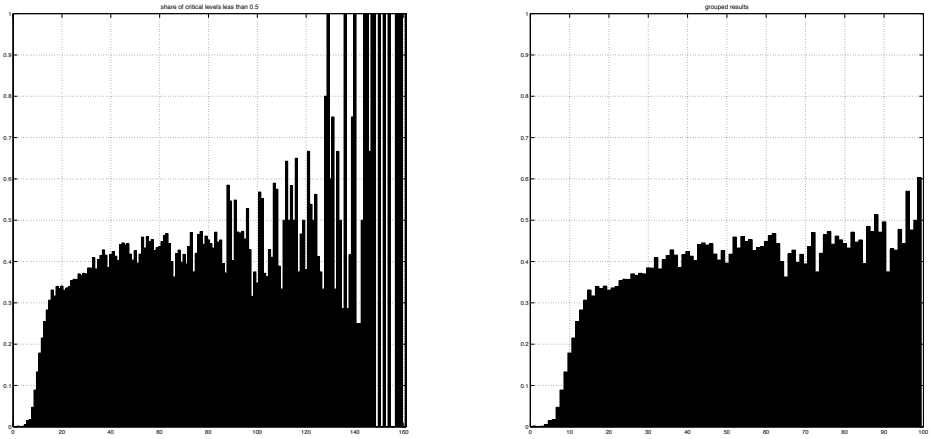
If we were to test every sequence in $D_k$ versus every other sequence with expected value $\check{\mu}_k$ we would approximately have to conduct $1.3 * 10^9$ different tests. If every test took 1/100 of a second it would take us about 150 days to get the result. Instead we will proceed in the following fashion, and test every subsequent pair of sequences in $D_k$. Let $k^* \in \{1, ..., l\}$ and assume that $D_{k^*} = (\hat{\Delta}_1^*, ..., \hat{\Delta}_h^*)$. To test if the sequences in $D_{k^*}$ are equally distributed we will test every pair of sequences $(\hat{\Delta}_i^* , \hat{\Delta}_{i+1}^*)$, $i = 1, ..., h - 1$.

**Sequences of 31 observations.** Since we test sequences with equal sample means the 16 data sets that only contain one sequence each are removed from the test. The shortest 1186 sequences $\hat{\Delta}_i$ with sample mean 0 only contain 0-arrival times and are skipped in the formal test - with the observed precision they are identical. Testing the other data sets with a total of 312086 members took 11 hours[8].

Results: 1.4% of the critical levels are smaller than 5%, 0.2% are smaller than 1% and 0.02% are smaller than 0.1%. If we only test the longest 5% of the sequences we get the following results: 4.2% of the critical levels are smaller than 5%, 0.8% are smaller than 1% and 0.1% are smaller than 0.1%. Figure 9 plots the share of critical levels smaller than 0.5 ($y$-axis) in order of increasing sample means ($x$-axis). In the chart to the right we have have grouped the critical levels from consecutive data sets with less than 100 sequences to get a better overview of the data. For example, $(\#D_{100}, ..., \#D_{105}) = (41, 47, 45, 39, 36, 45)$ and after grouping we end up with two sequences of critical levels with 133 and 120 elements each. If we plot the share of critical levels smaller than 0.1 or 0.05 for example we get a similar result. Once again it seems like the precision error in small $\check{\mu}$ sequences is too large for effective testing.

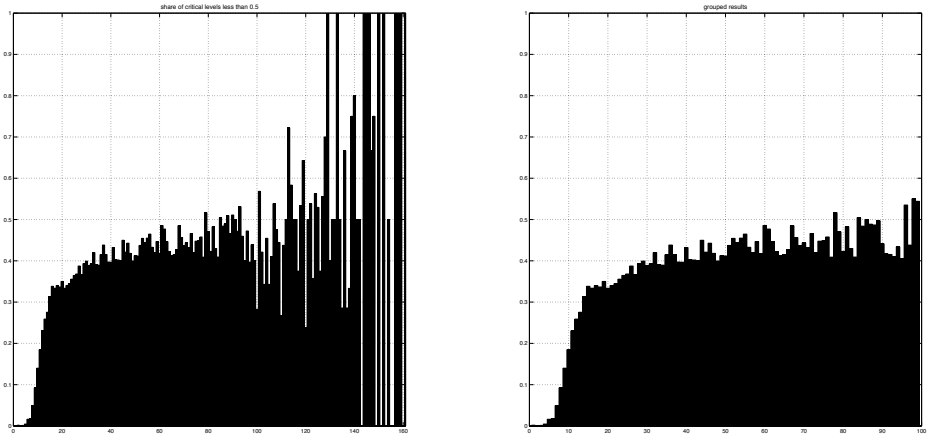
[8]We have used a computer with a 2.8 GHz Intel Core i5 processor and 4 GB 1333 MHz DDR3.

FIGURE 9




One drawback with the design of the test is that it assumes that the grouping in accordance with expected value is correct. Assume that the first half of the sequences in $D_k$ are identical to each other and that the other half are also identical to each other but completely different from the first half. The test would then correctly confirm that all but one pair is equal but this does not make the first and last sequence equally distributed. This specific case is of course highly unlikely but we need to do something to improve on the reliability of the results. To do this we conduct the test again but this time we permute the order of the sequences in $D_k$ randomly for each $k$ before we conduct the test. The average results for the whole body of data are identical to the precision presented above, the same holds true for the longest 5% of the sequences. Figure 10 is analogous to Figure 9 and is very similar.

FIGURE 10




The results are in strong favor of $H_0$.

**Sequences of 51 observations.** 26 of the data sets only contain one sequence each and are removed from the test. The shortest 181 sequences $\hat{\Delta}_i$ with sample mean 0 only contain 0-arrival times and are skipped in the formal test. Testing the other data sets with a total of 190161 members took 7.5 hours.

Result: 1.7% of the critical levels are smaller than 5%, 0.3% are smaller than 1% and 0.03% are smaller than 0.1%. If we only test the longest 5% of the sequences we get the following results: 4.3% of the critical levels are smaller than 5%, 1.1% are smaller than 1% and 0.16% are smaller than 0.1%.

FIGURE 11

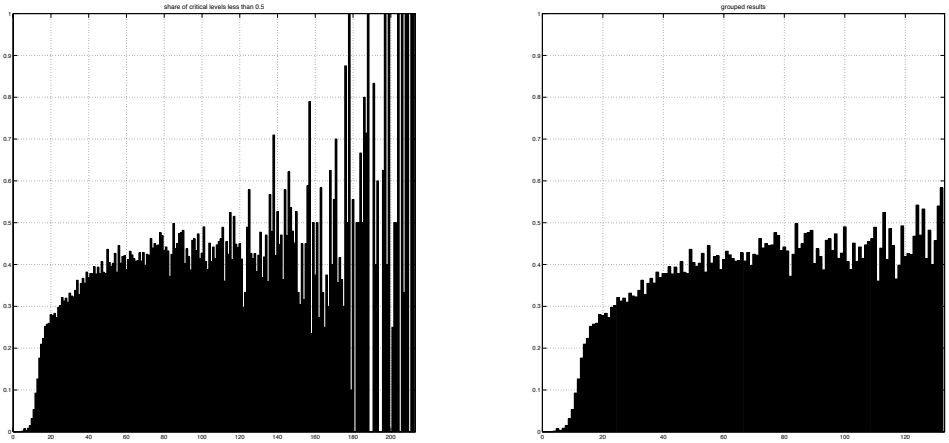



Figure 11 plots the share of critical levels smaller than 0.5 in order of increasing sample means and if we plot the share of critical levels smaller than 0.1 or 0.05 we get a similar result.

**Sequences of 200 observations.** There are 48449 sequence $\hat{\Delta}_i$ with 200 arrival times - all with at least one arrival time different from zero. The average time length of the sequences is 121 seconds. The test took less the 4 hours to perform. Result: 3.2% of the critical levels are smaller than 5%, 0.9% are smaller than 1% and 0.2% are smaller than 0.1%. If we only test the longest 5% of the sequences we get the following results: 6.1% of the critical levels are smaller than 5%, 1.7% are smaller than 1% and 0.4% are smaller than 0.1%.

**Summary** The tests strongly indicate that sequences of 31 observations with equal expected value are drawn from the same underlying distribution - the null hypothesis can not be rejected. The same thing is true for sequences of 51 observations. Sequences of 200 observations with equal expected value seem to be on the verge of homogeneity.

3.3. **The pooled data.** According to the results in the previous two subsections, arrival times in sequences $\hat{\Delta}_i$ of 31 observations can be considered independent and for each $k \in \{1, ..., l\}$, the sequences in $D_k$ can be considered observations from one underlying distribution. This is welcomed since we are no longer limited to data sets of 31 observations.

For further analysis we will pool the data from the sequences $\hat{\Delta}_i$ with equal sample mean into new sequences $\check{\Delta}_k$. The new pooled sequences will be similar to the collections $D_k$ but the elements will be observations instead of sequences. Remember that $\hat{\mu}_i = \sum_j \Delta_j^i/n$ and denote these new datasets $\check{\Delta}_k$,

$$\check{\Delta}_k = (\Delta_j^i)_{j=1,...,n}^{i:\,\hat{\mu}_i=\check{\mu}_k}.$$

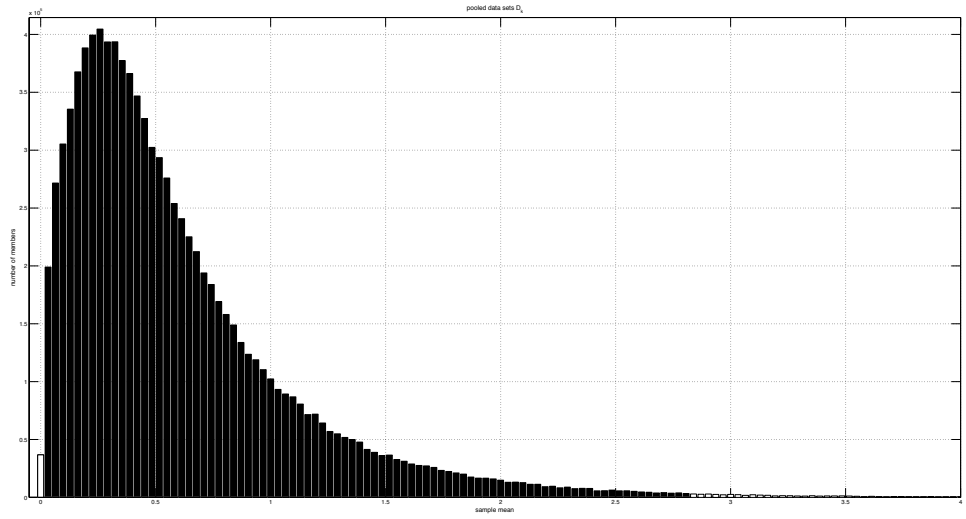
Let $\check{\sigma}_k^2$ to denote the sample variance of $\check{\Delta}_k$, the sample mean will by definition be equal to $\check{\mu}_k$.

Since we want to reduce the uncertainty of the smaller data sets we will only consider the sequences $\check{\Delta}_k$ with at least $100 \times 31$ elements and we will also remove the data sets corresponding to $\check{\mu}_k = 0$ for natural reasons. From now on we will sloppily refer to these reduced data sets as $\{\check{\Delta}_1, ..., \check{\Delta}_h\}$ (as the reduction never took place). The reduced data sets contains 99.1% of the original observations.

Figure 12 has $\check{\mu}_k$ on the $x$-axis and the number of observations in $\check{\Delta}_k$ on the $y$-axis. The transparent bars represent the removed data sets.

**Remark** The sequences $\check{\Delta}_k$ are very similar to the original sequences $\hat{\Delta}_i$. For example, 30 out of 31 elements in $\check{\Delta}_1$ is equal to zero, 1 out of 31 is equal to one. The sample means are still multiples of 1/31, $\check{\mu}_k \in \{j/31 \colon j = 1, 2, ..., 86, 88\}$ and in the case of $k = 1$, $\check{\sigma}_k^2 = \check{\mu}_k - \check{\mu}_k^2$. $\check{\Delta}_2$ consists of zeros, ones and twos.

FIGURE 12

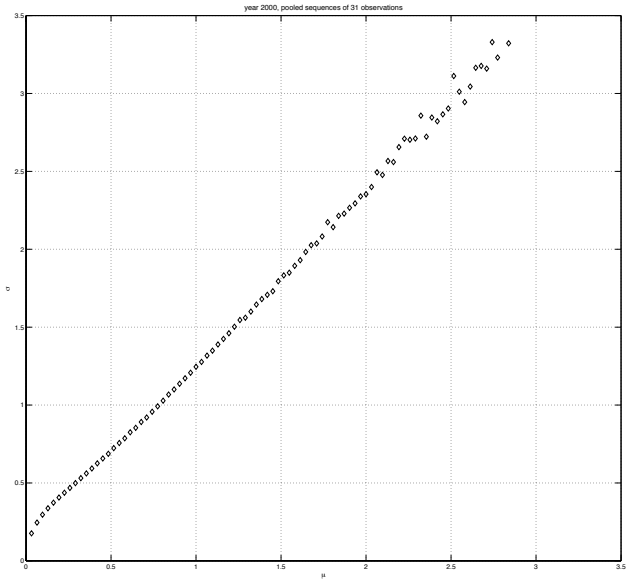



**Comparison data.** Sometimes we will need to get a feel for the precision error in the observed data - to this end we create the comparison data sets $C_k$. For $k = 1, ..., h$ let $C_k$ consist of $\#\check{\Delta}_k$ arrival times in a floored Poisson process with mean arrival time $\check{\mu}_k$.

A first look at the pooled data $\check{\Delta}_k$ is provided in Figure 13, the sample means are plotted versus the sample standard deviations. This data looks much more well behaved than the one in Figure 4 and for $\check{\mu}_k > 0.75$ there seems to be a linear relationship between

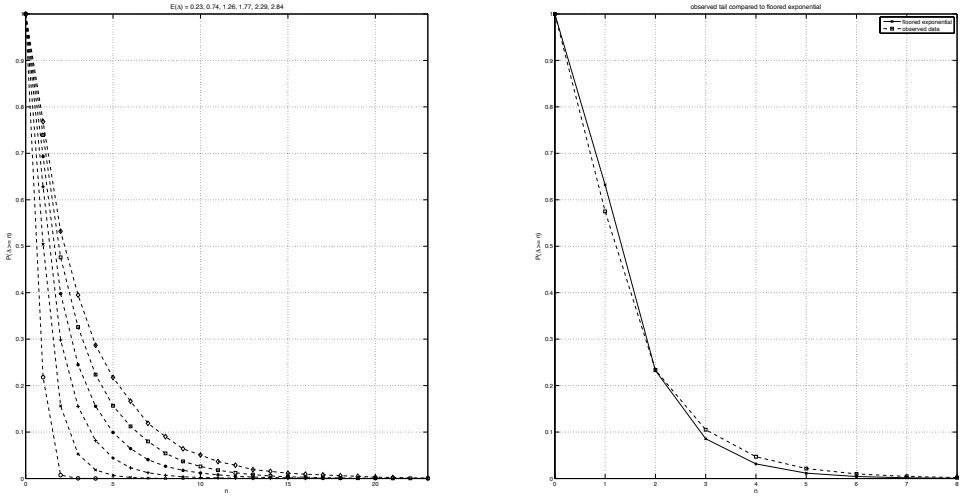
$\check{\sigma}_i$ and $\check{\mu}_k$. By Corollary 2.5, $\check{\sigma}_k \approx \sqrt{\check{\mu}_k}$ for small $\check{\mu}_k$ and this explains the non-linear behavior close to origo.

FIGURE 13




Continuing the examination of the pooled data, the chart to the left in Figure 14 illustrates the tails of the sample distributions for $\check{\Delta}_k$ with $k = 7, 23, 39, 55, 71, 87$ and $\check{\mu}_k = 0.23, 0.74, 1.26, 1.77, 2.29, 2.84$. The chart to the right is the tail of the observed data with sample mean 1 together with the tail of the asymptotic arrival time in a floored Poisson process with the same intensity.

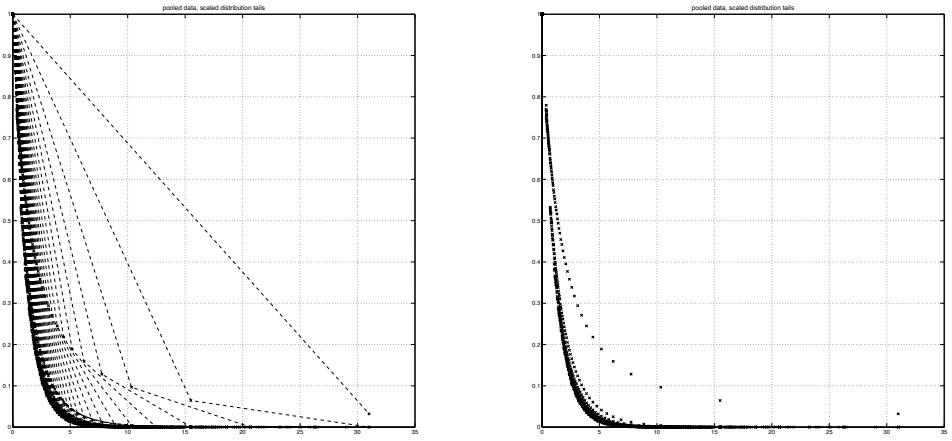
FIGURE 14



The similarities and differences between the sample tail with mean 1 and its exponential counterpart is characteristic for all means. The tail decays rapidly for small $n$, crosses the exponential tail at some point and then continues above it. In general the tails of $\check{\Delta}_k$ look well behaved and it is natural to pose the question whether they come from the same scalable distribution or not.

**Scaling the arrival times.** Figure 15 depicts the tails of the distribution of the scaled arrival times - we have normalized every pooled sequence $\check{\Delta}_k$ with its mean value $\check{\mu}_k$. In the chart to the right we have excluded the dashed lines.

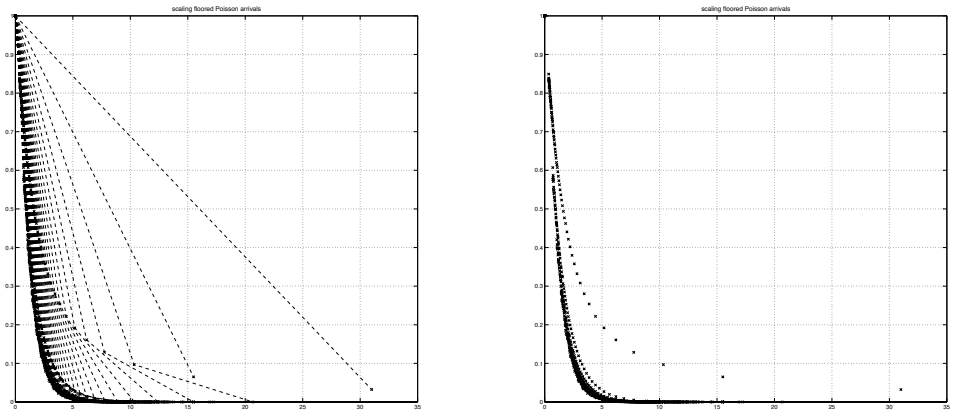
FIGURE 15




The tail from $(0, 1)$ to $(31, 1/31)$ might strike someone as odd but is the tail belonging to $\check{\Delta}_1$. The normalization maps every 0 to 0 and in this case every 1 to 31. The tails look more similar than before (c.f. Figure 14) but the results are obviously not satisfactory.

**Example** Let us conduct the same procedure on the comparison data $C_k$ to see how the arrival times in the floored Poisson process react to the scaling.

FIGURE 16



The result is very similar to how the sample data reacted and once again it is the high intensity data that reacts the worst to the scaling.

As in Section 2, let $(S_n)$ constitute a renewal process with arrival times $T_i$ and let $(\Sigma_n)$ be the floored renewal process with arrival times $\Delta_i$. Let $T_i \stackrel{d}{=} T_\lambda$ where $T_\lambda$ is scalable with parameter $\lambda$ and let $\Delta_\lambda$ be the asymptotic arrival time in the floored process. According to Lemma 2.2

$$P(\Delta_\lambda \geq n) = \int_{n-1}^{n} P(T_\lambda \geq t)dt = P(T_\lambda \geq \theta);\ n-1 < \theta < n.$$

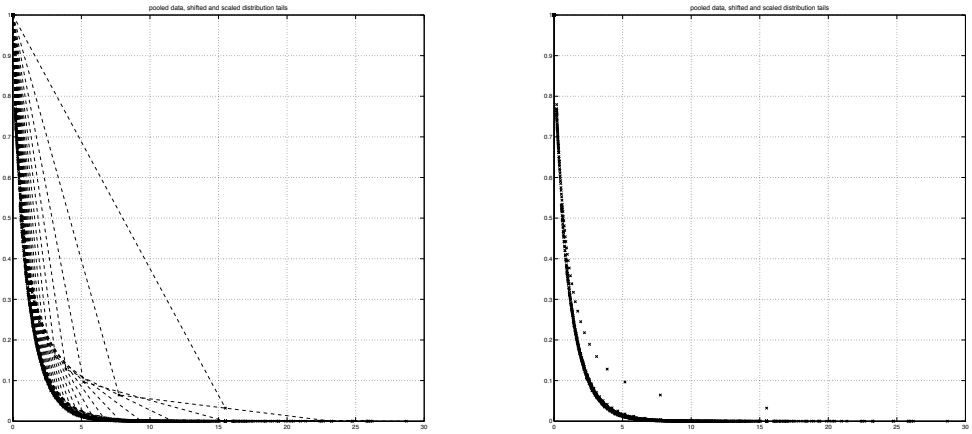
To simplify the expression, knowing nothing of the distribution of $T_\lambda$ we make the linear approximation $\theta = n - \frac{1}{2}$ rendering

$$P(\Delta_\lambda \geq n) \approx P(T_\lambda \geq n - 0.5).$$

Hence, to make the tail more similar to the underlying tail we should shift it 0.5 units to the left (except for $n = 0$) before scaling.

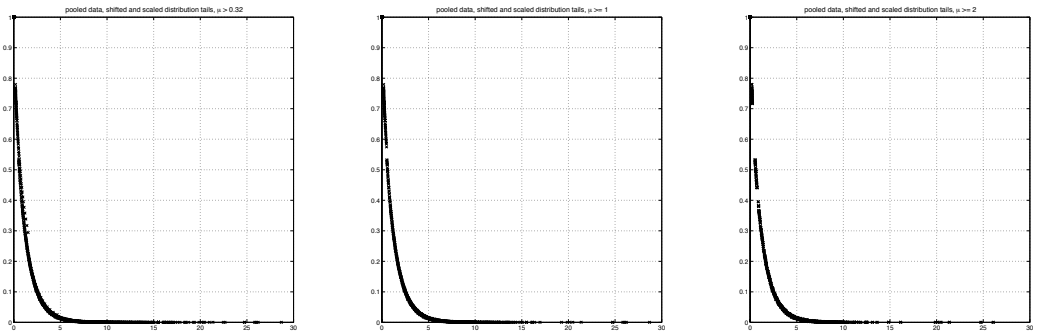
**Shifting and scaling the arrival times.** The shifted and scaled tails of the distribution of the pooled observations in $\breve{\Delta}_k$ are plotted in Figure 17.

FIGURE 17




As displayed in the chart the shifting and scaling works well except for small means. This becomes even clearer when we remove the data sets with small sample means from the plot. Figure 18 illustrates the shifted and scaled tails of $\breve{\Delta}_k$ for $\breve{\mu}_k > 0.32$, $\breve{\mu}_k \geq 1$ and $\breve{\mu}_k \geq 2$

FIGURE 18




As visualized in the right chart of Figure 14 there are similarities between the distribution of the sample data and the arrival times in a floored Poisson process. By looking at the tails in the latter distribution we might improve on our linear estimate $\theta = n - 0.5$.

Once again, let $\Delta_\lambda$ denote the asymptotic arrival time in a floored Poisson process and let $T$ denote the underlying arrival times with $\mu = 1/\lambda$. Then for $n \in N$

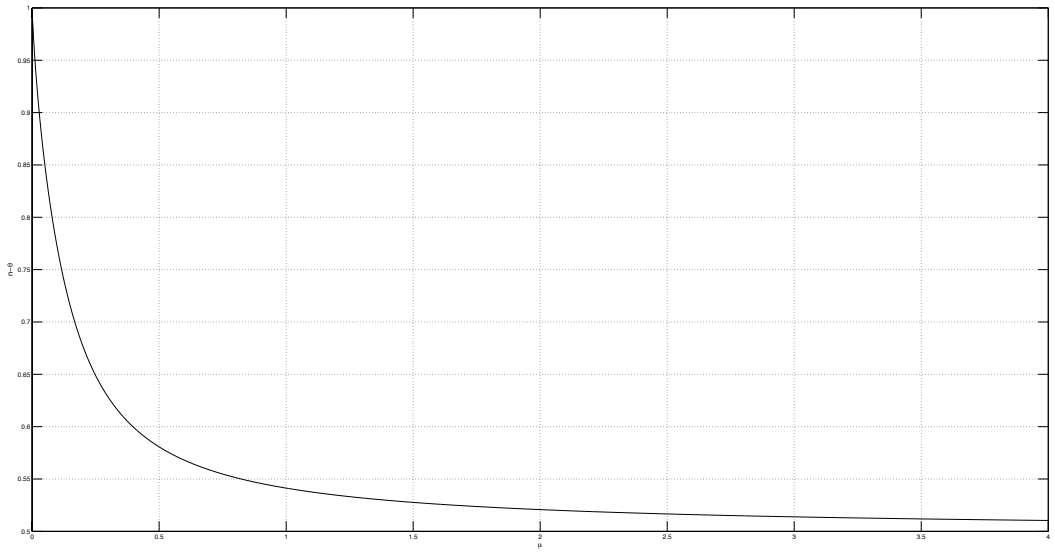
$$P(\Delta_\lambda \geq n) = e^{-\lambda n}\frac{e^\lambda - 1}{\lambda} = e^{-\lambda\theta_n(\lambda)}, \text{ where } \theta_n(\lambda) \in (n-1, n)\,.$$

Solving for $\theta_n$ gives

$$\theta_n(\lambda) = n - \frac{1}{\lambda}\log\frac{e^\lambda - 1}{\lambda}, \tag{3.1}$$

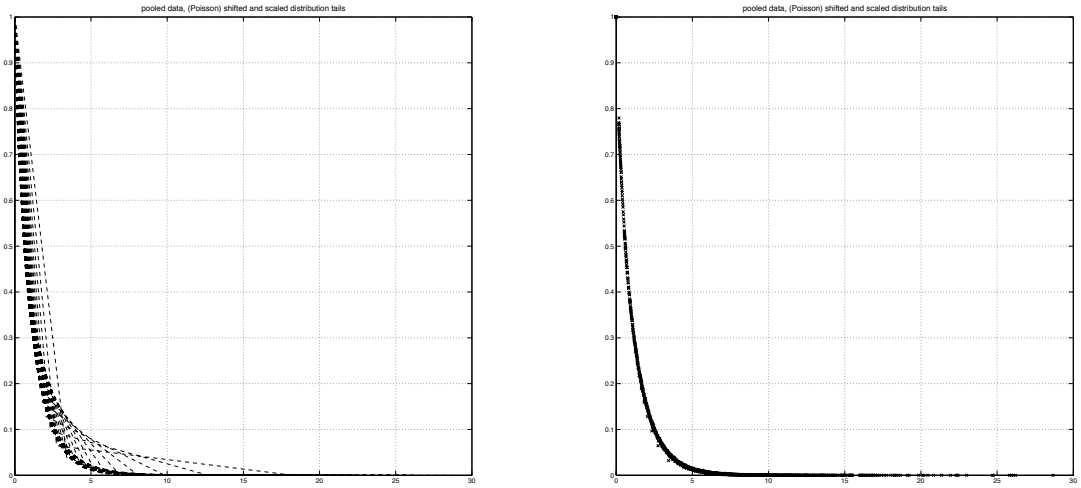
so the horizontal distance between the floored tail and the underlying tail is in this case independent of $n$. Figure 19 illustrates $n - \theta_n(\lambda)$ as a function of $\mu = 1/\lambda$.

FIGURE 19



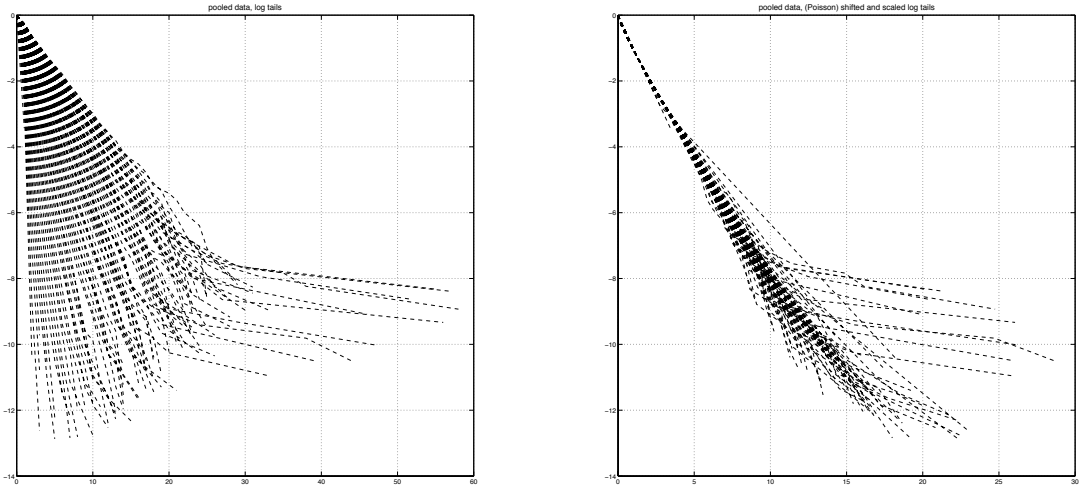
Since we will not shift the tails at 0, let $\theta_0(\lambda) = 0,\ \forall\lambda$. If we instead of shifting the sample tails 0.5 units to the left we (Poisson) shift the tails $n - \theta_n(\lambda)$ to the left before normalizing we get the following result.

FIGURE 20




A much more satisfactory result. To get a clearer picture of the behavior further out in the tails we plot log-tails for both the original and the (Poisson) shifted and scaled observations in $\breve{\Delta}_k$.

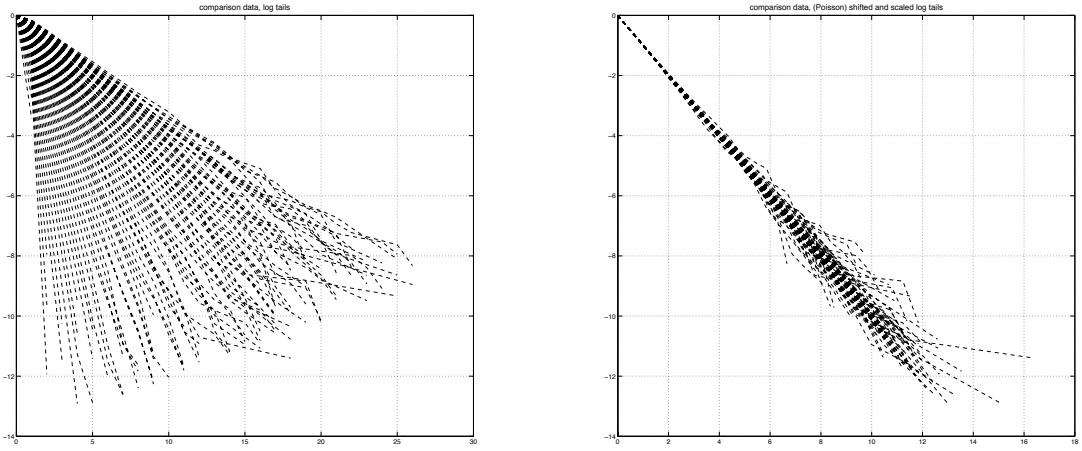
FIGURE 21




Looking at Figure 21 it is obvious that the scaling is not perfect and that we have some funny behavior further out in the tails for the sample data with lower intensity

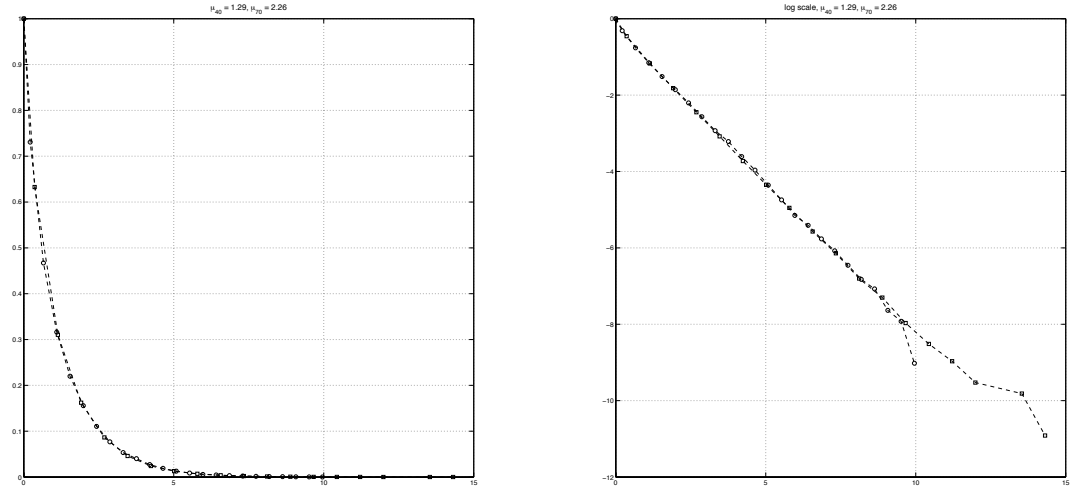
(and a lower number of observations). A comparison to the more thin tailed comparison data $C_k$, where the shifting is correct, shows on a similar behavior.

FIGURE 22




The above results indicate that the data in $\check{\Delta}_k$, $k = 1, ..., 87$ might be generated by the same underlying, scalable distribution and it is tempting to proceed with statistical testing of this, but this is not as straightforward as one may think. First of all the (Poisson) normalization of the tails is only an approximation, secondly and more importantly the data is (very) discrete and because of the normalization takes values in different, disjoint sets. Figure 23 illustrates the tails of the normalized distribution of $\check{\Delta}_{40}$ ($\square$) and $\check{\Delta}_{70}$ ($\circ$) where $\check{\mu}_{40} = 1.29$, $\check{\mu}_{70} = 2.26$, $\#\check{\Delta}_{40} = 54901$ and $\#\check{\Delta}_{70} = 8277$.

FIGURE 23



Below we have illustrated the two (non normalized) data sets in the form of a table, $\max \check{\Delta}_{40} = 19$ and $\max \check{\Delta}_{70} = 23$.

TABLE 2

| | 0 | 1 | 2 | 3 | 4 | ... | 23 |
|---|---|---|---|---|---|---|---|
| $\check{\Delta}_{40}$ | 20163 | 17700 | 8131 | 4165 | 2220 | ... | 0 |
| $\check{\Delta}_{70}$ | 2229 | 2182 | 1249 | 797 | 531 | ... | 1 |

After the (Poisson) normalization both sets will still contain the zero observation but apart from that the two sets will be disjoint. Let $V_k = \{\check{\Delta}_k\}$ denote the unique arrival times in $\check{\Delta}_k$, then the normalized observed variable in $\check{\Delta}_{40}$ will assume values in the set

$$\left\{ \frac{\theta_k(1/\check{\mu}_{40})}{\check{\mu}_{40}} : k \in V_{40} \right\}$$

while the normalized observed variable in $\check{\Delta}_{70}$ will assume values in

$$\left\{ \frac{\theta_k(1/\check{\mu}_{70})}{\check{\mu}_{70}} : k \in V_{70} \right\}.$$

Of course, some of these normalized observations could happen to coincide between the data sets but in our case, none of them does. One way around this is to somehow group the data but there would be a risk of a somewhat arbitrary grouping or a large precision loss. Another way around it could be to assume that the tails are linear between the observations and conduct a $Kolmogorov - Smirnov$ test but it would be a blunt approximation for the high intensity data and also difficult to interpret the results from a statistical point of view.

Instead we will rely on the visual analysis and claim that it is reasonable to believe that the data in $\check{\Delta}_1, ..., \check{\Delta}_h$ come from the same one parameter distribution.

**Summary** Let $T$ denote the intraday arrival time between two transactions in one stock, it is then plausible to assume that the distribution function of $T$ is of the form

$$F_\lambda(t) = F_1(\lambda t),$$

i.e. the distribution is scalable.

## 4. Distribution fitting

Let $T = (T_1, T_2, ..., T_n)$ be a short sequence of arrival times between trades of shares in a certain stock. Then the results in the previous section suggest that the $T_i$ can be considered *iid* from a scalable distribution $F_\lambda(t)$. But since the observed distribution differs from the underlying one the distribution fitting is not as straightforward as otherwise. Fortunately, by Corollary 2.3, assuming that the observed arrival times are asymptotically distributed, the sample mean of the observed arrival times is a consistent estimate of the expected value of the underlying distribution. This makes it natural to begin the search for $F_\lambda(t)$ in the $(\mu, \sigma)$-plane.

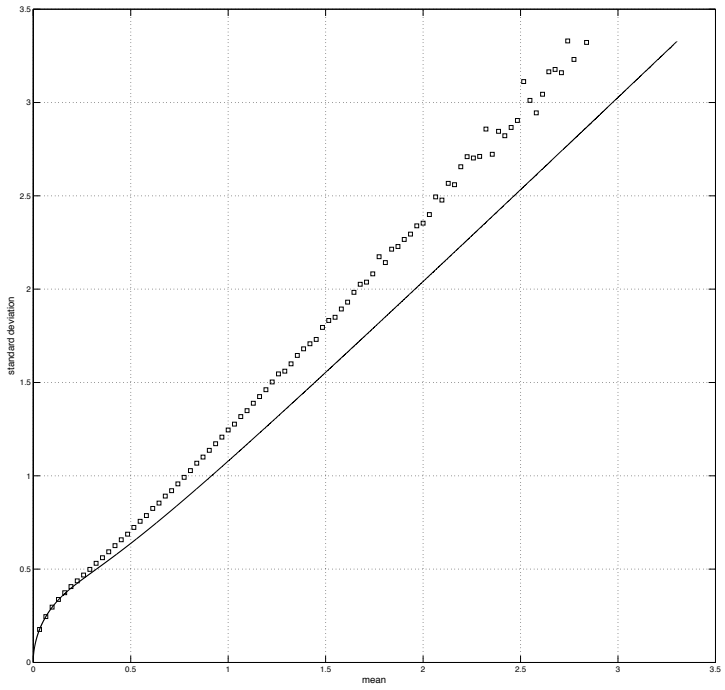
4.1. **The expected value and the standard deviation.** Since we are dealing with 87 data sets we need a fast method to reject a distribution as a candidate for $F$. One way to do this is to check if the theoretical distribution with expected value $\check{\mu}_k$ has a standard deviation close to $\check{\sigma}_k$.

**The Exponential distribution.** We start by plotting the relationship between $\mu$ and $\sigma$ for asymptotic arrival times in the floored Poisson process together with the observed $\check{\mu}$ and $\check{\sigma}$. Let $\Delta_\lambda$ denote the asymptotic arrival time in the floored Poisson process then from Lemma 2.2 and some tiresome calculations we get

$$\mathrm{Var}\Delta_\lambda = \frac{1}{\lambda}\left(1 + \frac{2e^{-\lambda}}{1-e^{-\lambda}} - \frac{1}{\lambda}\right).$$

Figure 24 has the mean on the $x$-axis and the standard deviation on the $y$-axis.

FIGURE 24




In general the fit is bad. The relatively higher standard deviation for the floored Poisson process with low mean is due to the precision loss and we see that $\mu$ gets closer and closer to $\sigma$ as the intensity (and precision error) decreases.

To be able to get a better fit we need an underlying distribution with higher coefficient of variation. The two parameter Gamma distribution is an alternative.

**The Gamma distribution.** Let $X \sim \Gamma(\alpha, \nu)$ and denote the mean and standard deviation $\mu$ and $\sigma$. Since

$$\mu = \nu/\alpha \text{ and } \sigma^2 = \nu/\alpha^2 \Rightarrow \sigma/\mu = 1/\sqrt{\nu}.$$

Since the low intensity data is the least effected by the precision errors we could make a first rough estimation of $\sigma/\mu$ from the $(\check{\mu}_k, \check{\sigma}_k)$ plot above - approximately 3/2.5. This gives an estimated shape parameter $\hat{\nu} \approx 0.69$.

A more rigorous estimate can be obtained in the following manner. Let $\Delta$ denote the asymptotic arrival time in a floored renewal process with gamma arrivals $X \sim \Gamma(\alpha, \nu)$ with $\mu = EX$ and let $\sigma_\Delta$ denote the standard deviation of $\Delta$. Then by Lemma 2.4

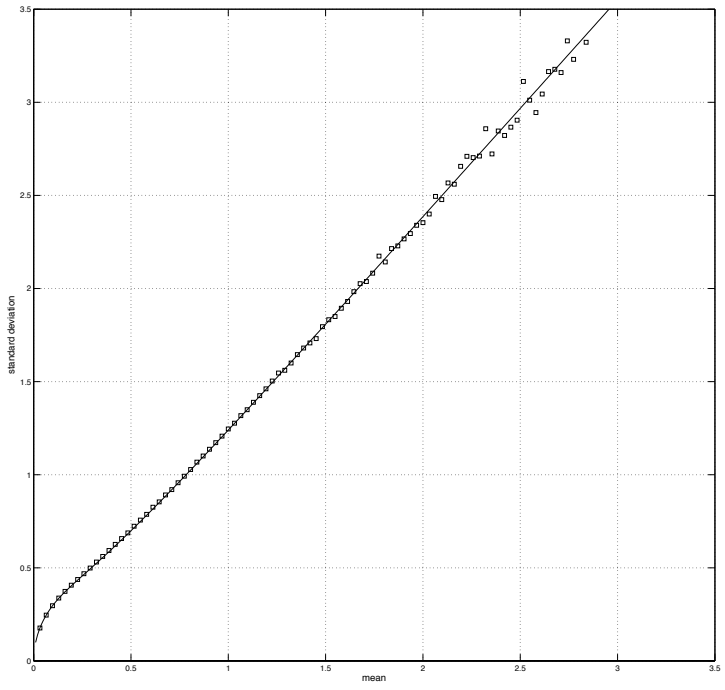
$$\sigma_\Delta = s_\nu(\mu)$$

for some continuous function $s$. By numerically minimizing

$$\sum_k |\check{\sigma}_k - s_\nu(\check{\mu}_k)| \tag{4.1}$$

over $\nu$ (implicitly estimating $\alpha$ by dividing $\nu$ with each individual sample mean) we get an estimate $\hat{\nu} \approx 0.72$. Note that Lemma 2.4 narrows down the possible values of $\nu$ that minimizes (4.1) and speeds up the search. In Figure 25 $s_{\hat{\nu}}(\mu)$ is plotted together with $(\check{\mu}_k, \check{\sigma}_k)$.

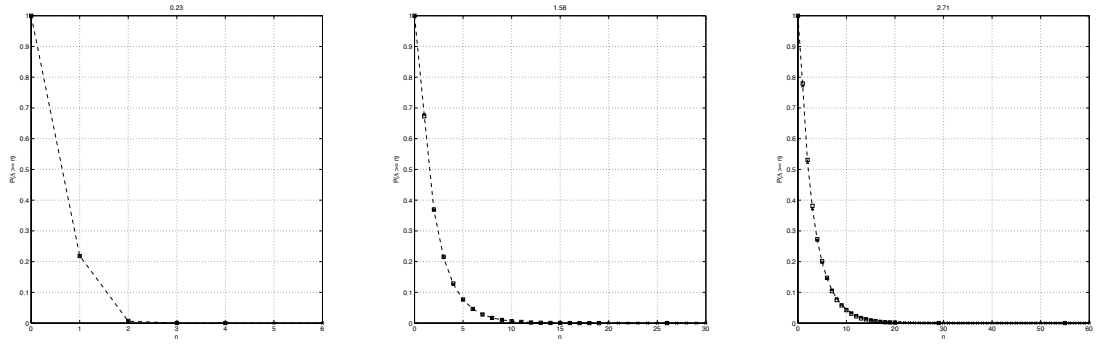
FIGURE 25




The fit is good but could it be the case that any positive two-parameter distribution could fit the $(\check{\sigma}_k, \check{\mu}_k)$ data as well? Apart from the connection to the exponential distribution we have no intuitive reasons to why the data should be gamma-distributed. With this in mind, let us continue the examination.

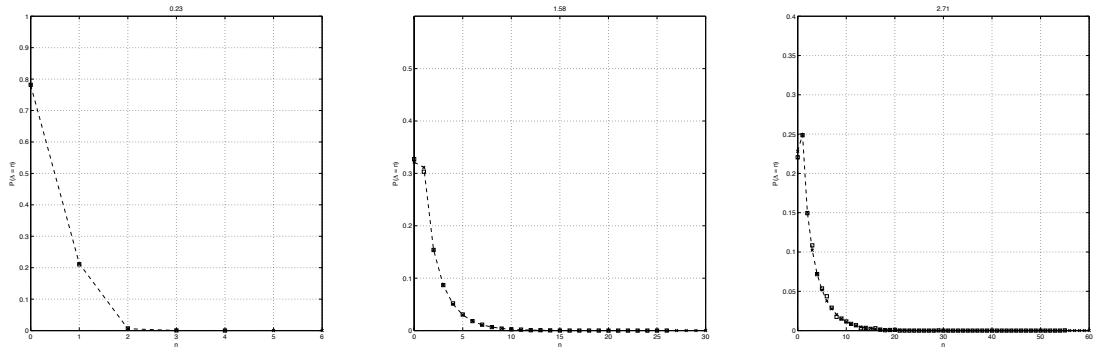
4.2. **The distribution probabilities.** Proceeding with the visual comparison of the gamma distribution and the observed data we plot the tails of the asymptotic arrival times in a floored process with gamma distributed arrivals versus the observed distribution tails of $\check{\Delta}_k$ (Figure 26). The observed data in the plots have sample mean $0.23, 1.58$ and $2.71$. Once again we estimate the shape parameter $\nu$ from the $(\check{\sigma}_k, \check{\mu}_k)$ data, $\hat{\nu} \approx 0.72$ and the scale parameter $\alpha$ from each data set by dividing $\hat{\nu}$ with the sample mean. The tails are numerically calculated using Lemma 2.2.

FIGURE 26




The theoretical tails ($\times$) seems to fit the data ($\square$) well. An similar visual inspection of the probability function is plotted in the Figure 27. The theoretical probability function

FIGURE 27




seems to fit the two first data sets well but there are some obvious discrepancies for the data with mean 2.7. Looking at the $(\check{\sigma}_k, \check{\mu}_k)$ plot above it should not come as a surprise that the choice of $\hat{\nu} \approx 0.72$ fits the low mean data sets better. We proceed by estimating the scale parameter for each individual data set.

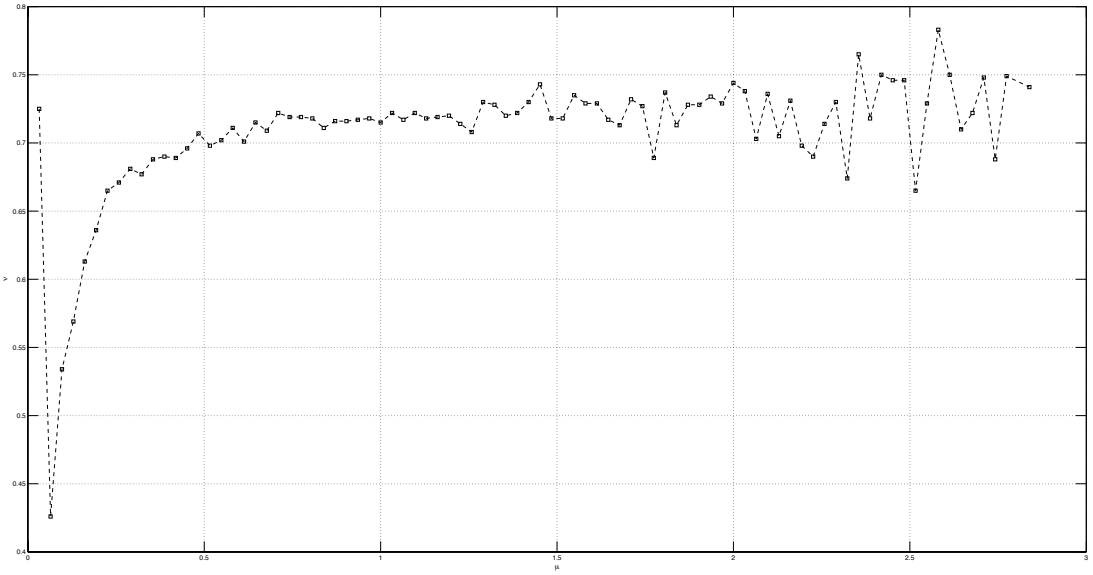
**Parameter estimation.** Assuming that the observed arrival times are asymptotically distributed we can use the results from Section 2 to estimate the parameters $\nu$ and $\alpha$ separately for each data set $\check{\Delta}_k$. Denote these estimates $\hat{\nu}_k$ and $\hat{\alpha}_k$. Since the sample

mean $\check{\mu}_k$ is a consistent estimate of the expected value in the underlying distribution we will use this to estimate the mean. Using $\hat{\alpha}_k = \hat{\nu}_k/\check{\mu}_k$ we only need to estimate $\hat{\nu}_k$ and we do this by minimizing

$$|\check{\sigma}_k - s_\nu(\check{\mu}_k)|$$

over $\nu$ for each $(\check{\mu}_k, \check{\sigma}_k)$, using the same notation and methodology as above. The resulting estimates $\hat{\nu}_k$ is plotted versus $\check{\mu}_k$ in Figure 28.

FIGURE 28




The swings in the estimate for larger sample means might be due to the smaller sample sizes (cf. Figure 12) but the dip in the graph for smaller sample means looks strange.

Let $\Delta$ and $s_\nu(\mu)$ be as above. Since the tails of the gamma distribution are well behaved in the sense that a low expected value limits the probability of outcomes larger than 1, by Corollary 2.5

$$s_\nu(\mu) \approx \sqrt{\mu - \mu^2}$$

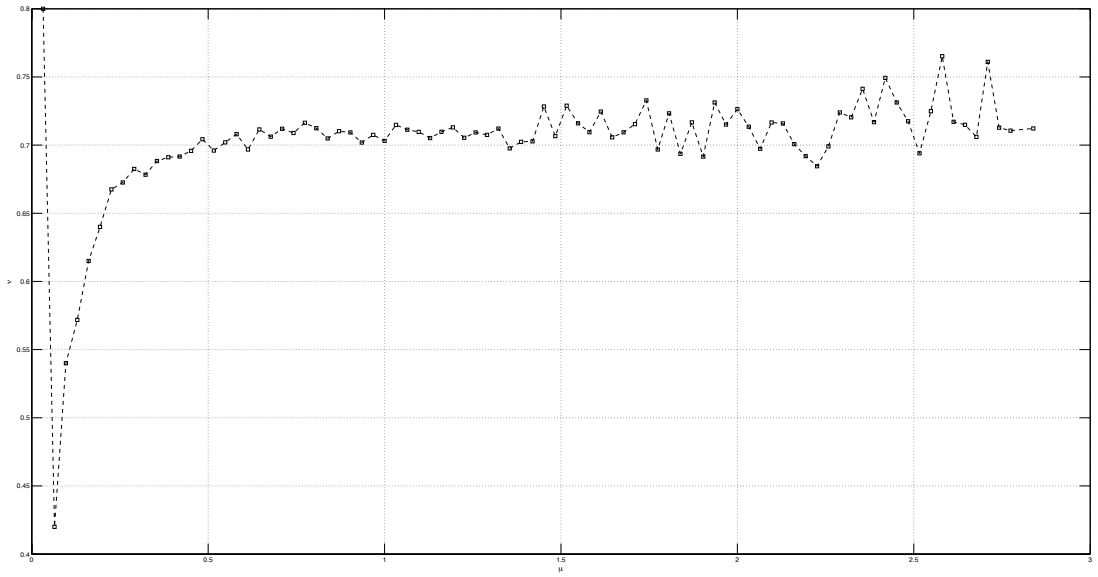
for small $\mu$. Hence, the variance is more or less independent of $\nu$ as long as $\alpha = \nu/\mu$ and $\mu$ is small and hence we have problems estimating the parameter. This will obviously distort the estimate for larger $\mu$ too.

Let $F_k^{\#}$ denote the empirical distribution function of $\check{\Delta}_k$. Instead of estimating $\nu$ by fitting the variance of the distribution to the variance of the data we estimate $\nu$ by fitting the theoretical (gamma) distribution function $F_{\alpha,\nu}$ to the empirical distribution function by minimizing

$$d_k(\alpha, \nu) = \sum_{n \in \{\check{\Delta}_k\}} |F_k^{\#}(n) - F_{\alpha,\nu}(n)|$$

over $\nu$, for each $k$, once again implicitly estimating $\alpha$ through $\hat{\alpha}_k = \hat{\nu}_k/\check{\mu}_k$. The resulting estimates plotted in Figure 29 are fairly similar (but less volatile) to the estimates obtained above.

FIGURE 29




The estimate $\hat{\nu}_1$ is especially unreliable, $d_1(\nu/\check{\mu}_1, \nu)$ is of the order $10^{-4}$ for $\nu = 0.1$ and decreases exponentially with increasing $\nu$. All the other estimates are obtained in (what numerically looks to be) a global minimum of $d_k(\nu/\check{\mu}, \nu)$ even though $d_k$ increases very slowly in $\nu$ for $k = 2, ..., 9$. One reason why the strange dip in the graph persists and why fitting a theoretical distribution to the high intensity data still is problematic is the way the data sets $\hat{\Delta}_i$ are constructed. When we from the beginning created sets with equal number of observations, we also limited the number of combinations of arrival times that make up a set with a fixed expected value. For a sequence of 31 observations, if the sample mean equals 1/31 that sequence will consist of 1 one and 30 zeros. When we later pool the sets we end up with for example $\check{\Delta}_1$ that consists of 198989 observations in total, of which $198989/31 = 6419$ are ones and the rest are zeros.

Assume that we have a 31 observation sample from a distribution with decaying tail probabilities and expected value 1/31 - we would not be surprised if that sample contained 1 one and 30 zeros, but unless we are dealing with data from a two point distribution we would be surprised if a sample of close to 200000 observations from the same distribution did not contain at least some twos and threes. When we are creating the data set $\check{\Delta}_1$ with mean 1/31 we are for example removing the potential sequence with 92 zeros and 1 three, and this might alter the empirical distribution function.

To check if the observed data is generated by an underlying gamma distribution we will conduct a $\chi^2$ test of goodness of fit[9]. This is a case where we will use an approximative test because of the sheer number of observations that limits the use of an exact test. But because of the vast number of observations in $\check{\Delta}_k$ we are pessimistic about the outcome, quoting Martin-Löf (1974): *"with large sets of data* [...]*no matter what model we try, we are sure to find significant deviations which force us to reject it"*. But for the sake of rigor let us start with the test.

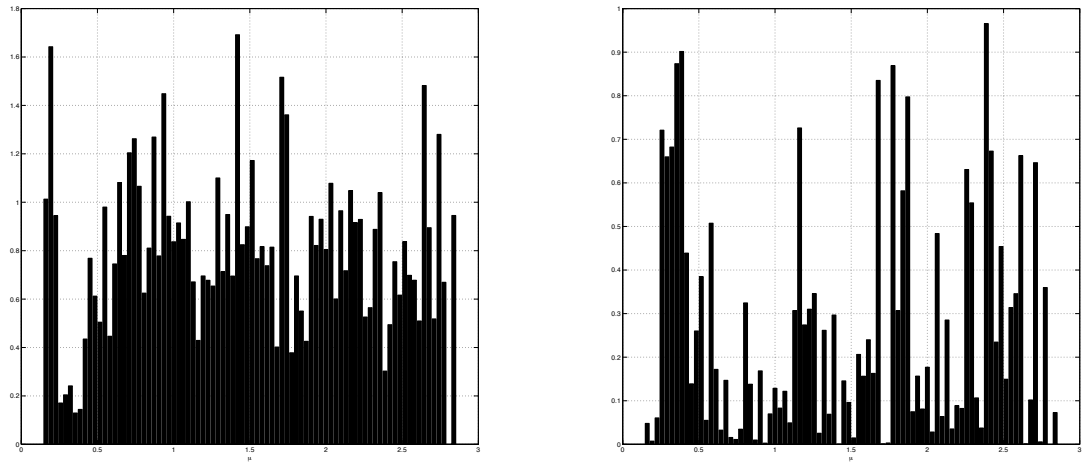
**$\chi^2$ test of goodness of fit.**

$H_0$ : The data in $\check{\Delta}_1, ..., \check{\Delta}_h$ comes from an underlying gamma distribution.

The first four datasets $\check{\Delta}_1, ..., \check{\Delta}_4$ contain too few unique observations to be tested. For the rest of the data sets we use the parameter estimations from minimizing the distance between the empirical - and the theoretical distribution function. The theoretical probability function is calculated[10] using Lemma 2.2. Since we're dealing with discrete data sets the grouping is trivial - if one group is expected to contain less than 5 observations we have merged it with every other group further out in the tail.

Both charts in Figure 30 have the sample mean on the $x$-axis, the left one has the $\chi^2$-statistic divided by the 95% quantile on the $y$-axis, the right one has the $p$-value on the $y$-axis.

FIGURE 30



Results: 77% of the $\chi^2$-statistics are smaller than the 95% quantile, 90% are smaller than the 99% quantile. If we did individual tests of each data set we would reject $H_0$ on a 5% level only 23 times out of 100. However, since $H_0$ concerns all of $\check{\Delta}_1, ..., \check{\Delta}_h$ the results must be evaluated on that basis and the test rejects the null hypothesis. If we were to ignore the data sets with smallest sample mean we would increase the number of "small" $\chi^2$-statistics but not enough to accept the null hypothesis. There are no simple relationship between the sample mean and the goodness of fit.

[9]See Cramér (1946).

[10]Numerical integration with step size $10^{-4}$ using *trapz()* in *MATLAB* .

Since the $\chi^2$-test rejected the null hypothesis we calculate the canonical redundancy[11] to measure the size of the discrepancy between the model of gamma renewals and the data.

**Example** Let us go back to the example on page 11 and the two sequences $(a)$ and $(b)$, both with mean $77/31 \approx 2.48$ . Table 1 is reprinted in Table 3 below.

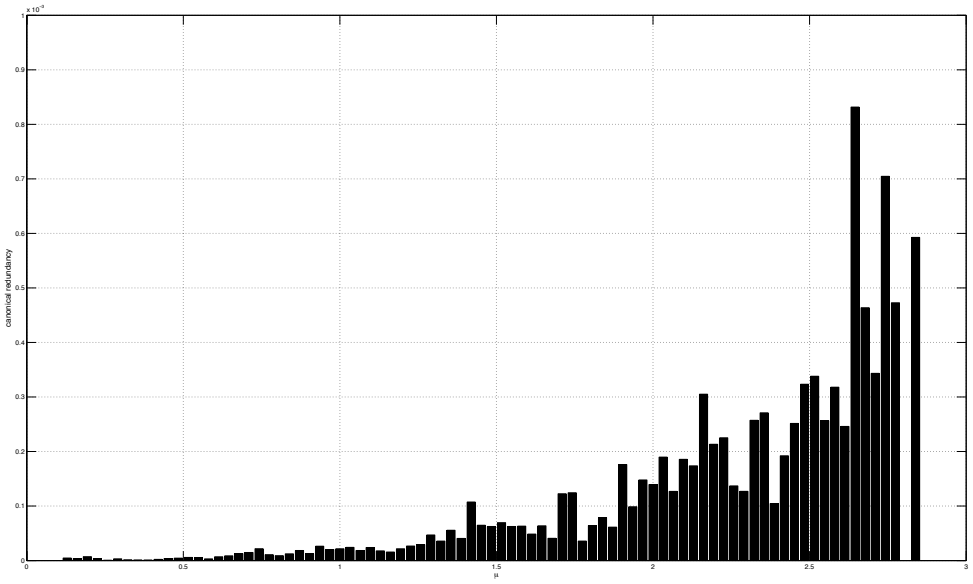
TABLE 3

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\Delta}_a$ | 5 | 10 | 5 | 1 | 4 | 1 | 3 | 1 | 1 | 0 |
| $\hat{\Delta}_b$ | 5 | 14 | 2 | 2 | 1 | 1 | 2 | 2 | 0 | 2 |
| Total | 10 | 24 | 7 | 3 | 5 | 2 | 5 | 3 | 1 | 2 |

The canonical redundancy $R$ of the reductive hypothesis of equal distribution equals 0.04 which on Martin-Löf's redundancy scale is a fit in between very bad and bad.

**Canonical redundancy.** The canonical redundancy is calculated for all data sets $\check{\Delta}_1, ..., \check{\Delta}_h$. The result is plotted versus $\check{\mu}_k$ in Figure 31.

FIGURE 31




The model describes the vast majority of the data sets with sample mean less than 2 seconds very well ($R < 10^{-4}$). All but three datasets show a redundancy of less than $5 * 10^{-4}$ which is mid in between good ($R = 10^{-3}$) and very good on the redundancy scale. All datasets have a good fit.

[11]A short note on the canonical redundancy introduced by Martin-Löf (1974) can be found in the Appendix.

**Other distributions.** The Gamma distribution seems to fit the data fairly well and arrival times between trades can probably be successfully modeled using it. But just because a distribution fits the data does not mean that there does not exist another distribution that fits the data better. We are reminded of William Feller's warnings[12] of the logistic distribution - the logistic distribution was fitted to practically all growth processes including human population, development of railroads and the height *and* weight of plants based on a belief in a *law of logistic growth.* The two parameter Weibull distribution could for example be calibrated to fit the $(\check{\sigma}_k, \check{\mu}_k)$ data in Figure 13 but the fit to the actual data is worse than the Gamma distribution.

**Summary** We can not statistically prove that the underlying data is gamma distributed but using the redundancy measure we can show that the distribution fits the observed data well. Even though we are having trouble estimating the shape parameter it is not unlikely that the data is generated by a gamma distribution with one shape parameter approximately equal to 0.72. To test this we would need larger sets of data of higher precision.

## 5. Conclusion

The main problem when analyzing intraday transaction data is not the large amount of data, the main problem is the precision loss. Estimating parameters and fitting distributions is not as straightforward as otherwise. Despite this we conclude that the arrival times in a short sequence of intraday trades are *iid* and very likely to be from a scalable distribution. The Gamma distribution fits the data well but we can not be sure that there are no other distribution that would make a better fit.

[12] See Feller (1970), pages 52-53.

## 6. References


**1.** Cramér H. (1946), Mathematical Methods of Statistics. Princeton University Press.

**2.** Engle, R.F. (2000), The Econometrics of Ultra-High Frequency Data. Econometrica, 68, 1-22.

**3.** Engle, R.F. and Russell J.R. (1998), Autoregressive Conditional Duration: a New Model for Irregularly Spaced Transaction Data. Econometrica, 66, 1127-1162.

**4.** Feller W. (1970), An Introduction to Probability Theory and Its applications, Volume II. Second Edition. John Wiley & Sons, Inc.

**5.** Fisher, R.A. (1934), Statistical Methods for Research Workers. Fifth Edition. Oliver & Boyd.

**6.** Khinchin, A.I. (1957), Mathematical Foundations of Information Theory. Dover Publications, Inc.

**7.** Martin-Löf, P. (1970), Statistiska modeller. Anteckningar från seminarier läsåret 1969-70 utarbetade av R. Sundberg. Institutionen för matematisk statistik, Stockholms universitet.

**8.** Martin-Löf, P. (1974), The Notion of Redundancy and Its Use as a Quantitative Measure of the Discrepancy between a Statistical Hypothesis and a Set of Observational Data. Scandinavian Journal of Statistics 1: 3-18.

**9.** Robbins, H (1953), On the Equidistribution of Sums of Independent Random Variables. Proceedings of the American Mathematical Society. Vol. 4, No. 5, 786-799

## 7. Appendix

Martin-Löf (1974) introduces the redundancy measure (from information theory) as a way to quantitatively measure the difference between a data set and a proposed model. The rationale behind the measure is that even if we are forced to reject a model on statistical grounds the model could still be good enough to be applied in the real world case. We will use an example to illustrate how the canonical redundancy is calculated.

**Example** Let $X_1$ and $X_2$ be two independent, discrete random variables with possible outcomes in $\{0, 1\}$. Assume that we have observed $n_i$ outcomes of $X_i$ and let $\nu_i$ denote the number of outcomes equal to 0 as illustrated in Table 4. Given that $P(X_1 = 0) = p$ and $q = 1 - p$, the probability of the observed outcome of $X_1$ is equal to

$$P_{X_1}(p; \nu_1, n_1) = p^{\nu_1} q^{n_1 - \nu_1} = e^{\nu_1 \log p + (n_1 - \nu_1) \log q}.$$

By maximizing the likelihood of the observed outcome we get the estimation $\hat{p}_1 = \nu_1/n_1$ and

$$P_{X_1}(\hat{p}_1; \nu_1, n_1) = e^{n_1(\hat{p}_1 \log \hat{p}_1 + \hat{q}_1 \log \hat{q}_1)} = e^{-n_1 H(\hat{p}_1, \hat{q}_1)},$$

where $H(\cdot)$ is the entropy as defined by Shannon[13]. Using a similar notation

$$P_{X_2}(\hat{p}_2; \nu_2, n_2) = e^{-n_2 H(\hat{p}_2, \hat{q}_2)},$$

and hence the maximum likelihood of the joint outcome of $X_1$ and $X_2$ equals

$$P_{X_1, X_2}(\hat{p}_1, \hat{p}_2) = e^{-n_1 H(\hat{p}_1, \hat{q}_1) - n_2 H(\hat{p}_2, \hat{q}_2)},$$

using a simplified notation. If we instead maximize the likelihood of the observed outcomes under the reductive hypothesis of equal distribution of $X_1$ and $X_2$ we get

$$\hat{p}_0 = \frac{\nu_1 + \nu_2}{n_1 + n_2} \text{ and } P_{X_1, X_2}(\hat{p}_0) = e^{-(n_1 + n_2) H(\hat{p}_0, \hat{q}_0)}.$$

The canonical redundancy $R$ is defined as

$$R = 1 - \frac{\hat{H}}{H_0}$$

where $\hat{H} = n_1 H(\hat{p}_1, \hat{q}_1) + n_2 H(\hat{p}_2, \hat{q}_2)$ and $H_0 = (n_1 + n_2) H(\hat{p}_0, \hat{q}_0)$. The smaller $R$ is the better the fit of the reductive hypothesis.

Table 4

| | 0 | 1 |
|---|---|---|
| $X_1$ | $\nu_1$ | $n_1 - \nu_1$ |
| $X_2$ | $\nu_2$ | $n_2 - \nu_2$ |

[13] See for example Khinchin (1957) for an easily accessible treatment of the entropy concept from an information theoretic point of view.