Evaluating Forecast Models with an Exact Independence Test

P. Araújo Santos
Instituto Politécnico de Santarém and CEAUL
paulo.santos@esg.ipsantarem.pt

M.I. Fraga Alves
Universidade de Lisboa and CEAUL
isabel.alves@fc.ul.pt

30 de Julho de 2009

Abstract: Christoffersen (1998) shows that the problem of determining whether a given forecast model is accurate can be reduced to examine the uncondicional coverage and independence properties. A clustering of violations signals a model that does not react quickly enough to changing conditions. In the context of market risk models, clustering corresponds to large losses occurring in rapid succession and can be difficult to recover from extremal losses occurring in a short period of time. The available tests for independence, suffer from serious limitations. They do not define, define a very general or a very narrow alternative hypothesis and we do not know the exact distribution of the tests statistics. There is a need for better tests. In this work we propose definitions for tendency to clustering of violations and the opposite of clustering of violations, allowing a convenient formulation of hypothesis. With the convenient alternative hypothesis, we propose a exact test, free from the small sample bias. Additionally to these fundamental advantages, we provide evidence trough a simulation study and with real data, that our test performs better.

Keywords: backtesting, duration based test, ordinal statistics, risk management.

1 Introduction

The test we propose below can be used to evaluate any model that produces interval forecasts or even in other situations where we are interested in detecting occurrences in clusters. In applications we may want to forecast the demand for a product or service, the future price of a commodity, the state of the atmosphere for a future time and a given location, and so on. Forecast models are relevant in several areas of science, however, we will focus the application on Value at Risk (VaR), a widely used risk management forecast model. VaR is the p^{th} quantile of the return distribution and produces one-sided interval forecasts with $(1-p)\%$ confidence level. With this measure the market risk of a portfolio can be communicated with statements like "At 99% confidence level, the most the portfolio can lose over one week is 5 million euros".

In mathematical terms we consider a time series of daily log returns, $R_{t+1} = \log(V_{t+1}/V_t)$, where V_t is the value of the portfolio at time t. The corresponding one-day-ahead VaR forecasts made at time t for time $t+1$, $VaR_{t+1|t}(p)$, is defined by:

$$P[R_{t+1} \leq VaR_{t+1|t}(p)|\Omega_t] = p, \tag{1.1}$$

where Ω_t is the information set up to time-t and p is the coverage rate.

Regulatory guidelines required commercial banks to set aside capital to insure against market risk. The size of the market risk capital requirement depends on VaR forecasts and on the accuracy of the VaR model. Backtesting is a fundamental tool for evaluate the forecast model, checking whether real losses are in line with forecasts. Considering a *violation* the event that a return on the portfolio is lower than the reported VaR, we can define the hit function as

$$I_{t+1}(p) = \begin{cases} 1 & \text{if } R_{t+1} \leq VaR_{t+1|t}(p) \\ 0 & \text{if } R_{t+1} > VaR_{t+1|t}(p). \end{cases} \tag{1.2}$$

In an influential paper Christoffersen (1998) shows that the problem of determining the accuracy of a forecast model can be reduced to examine whether the hit sequence,$\{I_t\}_{t=1}^{T}$, satisfies the unconditional coverage (UC) and independence (IND) properties. UC hypothesis means that the probability of a violation must be p, i.e., $P[I_{t+1}(p) = 1] = p, \forall_t$. IND hypothesis means that past violations do not hold information about future violations. The problematic non verification of the IND hypothesis is the one that conduct to clustering of violations, because it can be even more difficult to recover from several large losses occurring in a short period of time than it would be to recover from a higher than expected number of large losses that are spread out over time. As noted by Campbell (2007) the IND property represents a more subtle yet equally important property of an accurate risk model.

Only hit sequences that satisfy both properties provides evidence of an accurate model. When both properties are valid (UC and IND), then we say that forecasts have a correct conditional coverage (CC) and more formally we write

$$P[I_{t+1}(p) = 1|\Omega_t] = p, \forall_t. \tag{1.3}$$

Christoffersen (1998) shows in Lemma 1 of his work, that condition CC (1.3) is equivalent to say that the hit sequence (1.2) is independent and identically distributed (iid) as a Bernoulli random variable with probability of success p, which can be written as $I_t(p) \overset{iid}{\sim} Bern(p)$. Checking for iid Bernoulli violations involves checking that the number of violations is correct on average and checking that the pattern of violations is consistent with the iid behavior.

In a recent paper, Berkowitz *et al.* (2009) extend and unify the existing tests by noting that the de-meaned hits $\{I_t - p\}$ form a martingale difference sequence (m.d.s.). Equations (1.2) and (1.3) imply that $E[(I_{t+1}-p)|\Omega_t]=0$ and then for any variable Z_t in the time-t information set, we must have

$$E[(I_{t+1}-p)\otimes Z_t]=0. \tag{1.4}$$

This is the motivation for tests based on the martingale property of the sequence.

The earliest proposed backtest procedures only addresses the UC property. As a practical example, until recently, central banks ignores the IND property to assess the market risk taken by financial institutions. They use a basic "traffic light"method based on the binomial distribution. With this method, only if ten or more violations occurred in the previous 250 trading days, the model is classified as inaccurate; see the appendix for details. The literature about conditional coverage is recent but extensive, various tests on IND and CC have been developed. There are many good tests for UC and CC, however, the tests available for the IND hypothesis suffers from serious limitations. Two main problems of these tests is that they do not define, define a very general or a very narrow alternative hypothesis and relying in asymptotic results for the distribution of the test statistic when usually realistic sample sizes are small. The framework we propose in section 3, has several advantages, and the main advantages are a convenient formulation of hypothesis, not relying in a asymptotic distribution but in a exact distribution, and the greater power. Other advantage is great simplicity and easy application.

The recent 2008 financial crises illustrates well the importance of testing explicitly the IND property. We apply the popular Historical Simulation (HS) VaR, widely used by financial institutions. This technique takes the VaR on a certain day to be the unconditional quantile of the past T_e daily observations. Specifically

$$VaR^{HS}_{t+1|t}(p)=\text{quantile}(\{R_s\}^{t}_{s=t-T_e+1},100p). \tag{1.5}$$

We choose $T_e=250$ and the returns from the Deutscher Aktien index (DAX) from the first trading day of 2006 until the last trading day of 2008. We are assuming a portfolio that replicates the DAX index. In Figure 1.1 we plot the returns and one-day-ahead 1% VaR. In Table 1.1, from September 29, 2008, through October 15, 2008 we present the returns, one-day-ahead 1% VaR, violations and backtest results using "traffic approach"prescribed by recent regulatory framework. We observe a first cluster of 3 violations within 13 days (between January 21 and February 6) and then a impressive cluster of five violations occurring with very short durations in only 13 consecutive trading days (between September 29 and October 15), when for 1% VaR we would expect 100 days between violations. During this short period of 13 trading days the index lost more than 20%. With this flagrant pattern of clustering the backtest result only change from yellow to red in the last violation of the year, on November 6, 2008, even with a succession of five large losses occurring in a very short period of time. A dramatic bad performance of the regulatory framework. Clustering of violations signals a model that does not react quickly enough to changing market conditions. With this two clusters during one year, we see how slowly is the popular HS method to react to changing conditions and the incapacity of a backtesting method, which ignores the IND property, to detect such a inaccurate model. We will see in section 5 how our test can detect quickly this problem. We also enhance the importance of testing explicitly the IND hypothesis, for example for diagnostic purposes. Improvement of a inaccurate model requires investigate the source of the problem and if non independence is a problem in the actual forecast model.

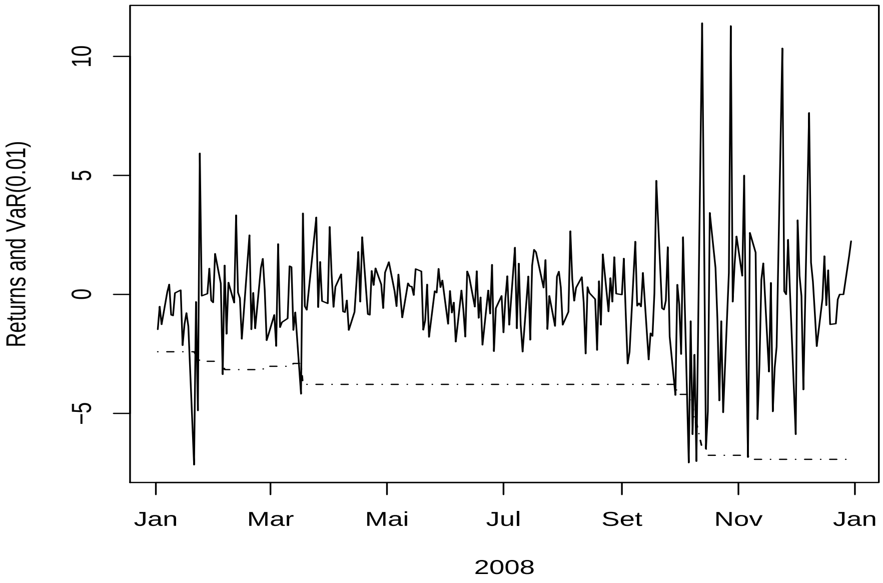

Figura 1.1: Deutscher Aktien index (DAX) returns (solid line) and one-day-ahead 1% VaR(0.01) (dotted line) based on Historical Simulation technique and on the previous 250 trading days.

Date	Index Value	R_{t+1}	$VaR_{t+1\|t}(0.01)$	$I_{t+1}(0.01)$	Backtest result
2008-09-29	5807.08	-0.0423	-0.0378	1	yellow
2008-09-30	5831.02	0.0041	-0.042	0	yellow
2008-10-01	5806.33	-0.0042	-0.042	0	yellow
2008-10-02	5660.63	-0.0251	-0.042	0	yellow
2008-10-03	5797.03	0.0241	-0.042	0	yellow
2008-10-06	5387.01	-0.0707	-0.042	1	yellow
2008-10-07	5326.63	-0.0112	-0.0456	0	yellow
2008-10-08	5013.62	-0.0588	-0.0456	1	yellow
2008-10-09	4887.00	-0.0253	-0.0539	0	yellow
2008-10-10	4544.31	-0.0701	-0.0539	1	yellow
2008-10-13	5062.45	0.114	-0.0646	0	yellow
2008-10-14	5199.19	0.027	-0.0646	0	yellow
2008-10-15	4861.63	-0.0649	-0.0646	1	yellow

Table 1.1 - Deutscher Aktien index (DAX) log returns, 1 day ahead VaR(0.01) based on Historical Simulation technique and on the previous 250 trading. Violations and regulatory framework backtesting results.

The remain of our paper is organized as follows. In section 2 we give an overview of existing tests for evaluating interval forecasts. In section 3 we present our exact independence test and theoretical results. In section 4, and through simulation experiments, we compare the performance of the new test with those currently available. Section 5 presents an empirical application using daily returns from DAX index. In section 6 we conclude. In the appendix we provide a statistical table which allows easy implementation of our test.

2 Tests for Evaluating Interval Forecasts

Backtesting procedures are essential to check if the forecasts are well calibrated. In the context of financial risk models, backtesting is essential to check if real losses are in line with projected losses. Risk managers rely on them to give indications of problems with risk models and central banks rely on them to evaluate the quality of risk models used by financial institutions.

There are several backtest procedures for VaR; see the papers of Campbell (2007) and Berkowitz *et al.* (2009) for a detailed review. The earliest procedures only addresses the UC property, a well know example is the proportion of failures (POF) test proposed by Kupiec (1995).

In the last ten years a lot of research has been made and several tests have been proposed to examine both the IND and CC properties of the hit sequence $\underset{\sim}{I} = \{I_t\}_{t=1}^{T}$. An early and widely used is the Christoffersen (1998) Markov tests, which define an alternative hypothesis where the hit sequence follows a first order Markov sequence with the following switching probability matrix

$$\Pi = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} \tag{2.1}$$

where π_{ij} is the probability of an i on day $t-1$ being followed by a j on day t. The null hypothesis of this test is then $H_{0,IND} : \pi_{01} = \pi_{11}$ and the null hypothesis of the conditional coverage test is $H_{0,CC} : \pi_{01} = \pi_{11} = p$. Denotingby π_1 the common value of π_{01} and π_{11} under $H_{0,IND}$, by T_0 the number of zeros in the hit sequence $\underset{\sim}{I}$, T_1 the number of ones, $T = T_0 + T_1$ and T_{ij} the number of observations with a j following an i, the maximum likelihood estimators are $\hat{\pi}_{01} = T_{01}/T_0$, $\hat{\pi}_{11} = T_{11}/T_1$ and $\hat{\pi}_1 = T_1/T$, the log-likelihood under the alternative hypothesis is

$$\ln L(\underset{\sim}{I}, \pi_{01}, \pi_{11}) = (1-\pi_{01})^{T_0 - T_{01}} \pi_{01}^{T_{01}} (1-\pi_{11})^{T_1 - T_{11}} \pi_{11}^{T_{11}}, \tag{2.2}$$

the independence test statistic is

$$LR_{IND} = -2(\ln L(\underset{\sim}{I}, \hat{\pi}_1) - \ln L(\underset{\sim}{I}, \hat{\pi}_{01}, \hat{\pi}_{11})), \tag{2.3}$$

and the conditional coverage test statistic is

$$LR_{CC} = -2(\ln L(\underset{\sim}{I}, p) - \ln L(\underset{\sim}{I}, \hat{\pi}_{01}, \hat{\pi}_{11})). \tag{2.4}$$

The test statistics (2.3) and (2.4) are asymptotically distributed as qui-square with one degree of freedom for the IND test and two for the CC test. Notice that this tests are sensible only to violations followed immediately by other violation, ignoring all other possibilities of cluster patterns, so, the alternative is to much narrowed. If in equation (1.4) we set Z_t to be the most recent de-meaned hit we have $E[(I_{t+1} - p)(I_t - p)] = 0$, the only condition explored by the Markov tests.

If we set $Z_t = I_{t-k}$ for any $k \geq 0$, we have $E[(I_{t+1}-p)(I_{t-k}-p)] = 0$. Based in this more general condition Berkowitz *et al.* (2009) suggest the Ljung-Box statistic (LB), a joint test of whether the first m autocorrelations are zero, as a test for the accuracy of a VaR model. For test CC, we use the null hypothesis $H_0 : \gamma_1 = ... = \gamma_m = 0$, with $m \in \aleph^*$, where $\aleph^*$ denotes the integer set, and we compute the autocorrelations of $(I_{t+1} - p)$ and then

$$LB(m) = T(T+2)\sum_{k=1}^{m} \frac{\gamma_k^2}{T-k} \tag{2.5}$$

which is asymptotically distributed as qui-square with m degrees of freedom.

Considering other data in the information set such as past returns, under CC we have $E[(I_{t+1}-p)g(I_t, I_{t-1}, ..., R_t, R_{t-1}, ...)] = 0$ for any non-anticipating function $g(.)$. In the same line as Engle and Manganelli (2004), Berkowitz *et al.* (2009) consider the nth-order autoregression

$$I_t = \alpha + \sum_{k=1}^{n} \beta_{1k} I_{t-k} + \sum_{k=1}^{n} \beta_{2k} g(I_{t-k}, I_{t-k-1}, ..., R_{t-k}, R_{t-k-1}) + \varepsilon_t \tag{2.6}$$

with $n = 1$ and $g(I_{t-k}, I_{t-k-1}, ..., R_{t-k}, R_{t-k-1}) = VaR_{t-k+1|t-k}(p)$.

They estimate the logit model assuming that ε_t has a logistic distribution and test with a likelihood ratio test whether $P(I_t = 1) = 1/(1 + e^{-\alpha}) = p$ and if the coefficients β_{11} and β_{21} are statistically significant. This is the CAViaR test of Engle and Manganelli (CAViaR).

Another method for testing the martingale hypothesis consists in examine the shape of the spectral density function (e.g., Durlauf(1991)). With this method Berkowitz *et al.* (2009) proposes the Cramér-Von Mises (CVM) and the Kolmogorov-Smirnov (KS) statistics for test the accuracy of a VaR model. However, in the simulation study done by the authors, these tests exhibit less power than de CAViaR test.

In the literature, also emerged a duration-based approach with potential to capture general forms of clustering. This approach uses the durations between the violations. There are related works on testing duration dependence (e.g., Kiefer (1988), Engel and Russel (1998)). As far as we know, the first author that purpose this kind of approach for interval forecast evaluation was Danielsson and Morimoto (2000), however they used the traditional qui-square goodness of fit test which have serious limitations, especially for realistic small sample sizes.

As noted by Christoffersen and Pelletier (2004), the intuition behind duration based approach is that the clustering of violations will result in an excessive number of short durations and in very long durations. In a context of VaR this correspond to periods of market turbulence and periods of market calm. Let us define the duration between two consecutive violations as

$$D_i := t_i - t_{i-1} \tag{2.7}$$

where t_i denotes the day of violation number i. We denote a sequence of N durations by $\{D_i\}_{i=1}^N$. If the CC (1.3) hypothesis is valid then $I_t(p) \overset{iid}{\sim} Bern(p)$ and consequently the process $\{D_i\}_{i=1}^N$ has a geometric distribution with probability mass function (p.m.f.)

$$f_D(d;\pi) = (1-\pi)^{(d-1)}\pi, \quad d \in \aleph^*, \tag{2.8}$$

with $\pi = p$. We will write

$$D_i \overset{iid}{\sim} D \sim \text{Geometric}(p). \tag{2.9}$$

We will write the IND hypothesis as

$$D_i \overset{iid}{\sim} D \sim \text{Geometric}(\pi), \text{ with } 0 < \pi < 1. \tag{2.10}$$

Under the CC (2.9) hypothesis it is easy to verify that the hazard function defined as $f_h(d) = P[D_i = d]/(1 - P[D_i < d])$ should be flat and equal to p . Under the IND hypothesis (2.10), the process $\{D_i\}_{i=1}^N$ has a geometric distribution (2.8) where π can be different from p.

The exponencial distribution with probability density function

$$f_D(d;\beta) = \beta \exp(-\beta d), \tag{2.11}$$

with $d > 0$ and $\beta > 0$, is the continuous analogue of the geometric distribution. The exponencial and geometric distributions are the only ones characterized by the lack of memory property. For both, the future is independent of the past and in mathematical terms we write $P[X \geq x + y | X \geq y] = P[X \geq x]$.

Based on the exponencial distribution, Christoffersen and Pelletier (2004) suggest three tests using the duration based approach, specifying for the alternative the Weibull, the Gamma and the Exponential Autoregressive Conditional Duration model. Haas (2005) uses the discrete life time distributions instead of the continuous ones and showed trough a simulation study that tests based on discrete distribution as higher power than the continuous counterparts. In this line, Berkowitz *et al.* (2009), using the following probability mass function with non-constant probabilities of violation

$$f_D(d;p) = (1-p_1)(1-p_2)...(1-p_{d-1})p_d \quad d \in \aleph^* \tag{2.12}$$

proposed the Geometric test, specifying $p_d = ad^b$ with $b \leq 0$.

The Generalized Method of Moments (GMM) test framework proposed by Bontecamps (2006) to test for distributional assumptions was extended by Candelon *et al.* (2008) to the case of VaR forecasts accuracy, with notable improvements and several advantages. In the group of available duration-based tests they show that the proposed GMM duration-based tests are the best performers, but for testing explicitly the IND hyphothesis remains the drawbacks of not defining an alternative clustering hypothesis and relying in asymptotic results for the distribution of the test statistic. The tests consists in using the GMM framework to test (2.9).

The Ord´s family (Poisson, Binomial, Pascal, Hypergeometric) can be associated to some particular orthonormal polynomials whose expectation is equal to zero. The orthonormal polynomials associated to the geometric distribution with probability p are defined by the following recursive relationship, $\forall d \in \aleph^*$

$$M_{j+1}(d;p) = \frac{(1-p)(2j+1)+p(j-d+1)}{(j+1)\sqrt{1-p}} M_j(d;p) - \Big(\frac{j}{j+1}\Big) M_{j-1}(d;p) \quad (2.13)$$

for any order $j \in \aleph^*$, with $M_{-1}(d;p) = 0$ and $M_0(d;p) = 1$. If (2.9) is true, then it follows that $E[M_j(D;p)] = 0$, $\forall_j \in \aleph^*$. The CC property can be expressed as $H_{0,CC} : E[M_j(D;p)] = 0$ and the IND property can be expressed as $H_{0,IND}$: $E[M_j(D;\beta)] = 0$ with $j = \{1, ..., k\}$ and $k > 1$ denoting the number of moment conditions. The parameter β, not necessarily equal to p, can be either fixed *a priori* or estimated. The GMM test statistic for CC is

$$J_{CC}(k) = \Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M(D_i;p)\Big)^{'}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M(D_i;p)\Big) \quad (2.14)$$

and for IND is

$$J_{IND}(k) = \Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M(D_i;\hat{\beta})\Big)^{'}\Big(\frac{1}{\sqrt{N}}\sum_{i=1}^{N} M(D_i;\hat{\beta})\Big) \quad (2.15)$$

where $M(D_i;p)$ denotes a $(k,1)$ vector whose components are the orthonormal polynomials $M_j(D_i;p)$ in the CC test and $M_j(D_i;\hat{\beta})$ in the IND test, for $j = 1, ..., k$. $\hat{\beta}$ is a consistent estimator of β. The test statistics (2.14) and (2.15) follows a asymptotic qui-square distribution respectively with k and $k-1$ degrees of freedom.

In the next section we present a new method for testing explicitly the independence hypothesis and then we go through a comparative simulation study. For the CC and IND hypothesis, the Markov tests (2.3) and (2.4) is perhaps the most widely used in the literature, that is the reason why we choose the Markov independence test (2.3) for the comparative study. In the group of available duration-based tests we choose the best performers GMM tests (2.15). We also choose the CAViaR test (2.6), the best performer in the comparative simulation study done by Berkowitz *et al.* (2009).

3 A Exact Test for the Independence Property

Let $D_{1:N} \leq ... \leq D_{N:N}$ be the order statistics (o.s.'s) of durations $D_1, ..., D_N$ defined in (2.7). The motivation behind our test is as follows. When violations generated by the hit function (1.2) occur in cluster, the majority of the durations are short (the short durations between violations in the clusters) and some durations are very long (the durations between the last violation of one cluster until the first violation of the next cluster). If the majority of the durations are short then the median, $D_{[N/2]:N}$, is short. If some durations are very long, the maximum, $D_{N:N}$, is very long. Finally, with a short median and a very long maximum, the ratio $D_{N:N}/D_{[N/2]:N}$ is high.

We illustrate our motivation with an example. We choose the returns from the DAX index from January 2, 1997, through December 30, 2008, and compute durations between violations (2.7) using the VaR(0.05) and the Historical Simulation (HS) with $T_e = 250$, a forecast method that leads to violations in clusters. As noted in the previous section, under the IND hypothesis, the process $\{D_i\}_{i=1}^{N}$ is a sequence of iid random variables with geometric probability mass function (2.8). We can compute the expected values of o.s.´s, $D_{1:N} \leq ... \leq D_{N:N}$, under the independence hypothesis (2.10), using the following expression obtained by Margolin and Winokur (1967),

$$E(D_{r:N}) = \sum_{j=N-r-1}^{N} (-1)^{j-N+r-1} \left(\begin{array}{c} j-1 \\ N-r \end{array} \right) \left(\begin{array}{c} N \\ j \end{array} \right) \frac{1}{(1-(1-\pi)^j)}. \tag{3.1}$$

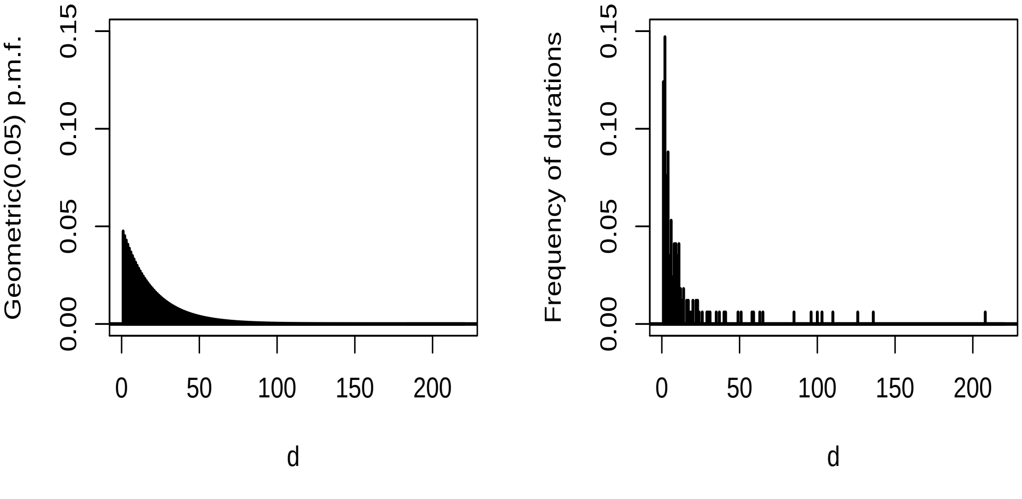

Figura 3.1: Geometric(0.05) probability mass function (left) and frequency of durations (right) between violations of VaR(0.05) for DAX index from 2 January 1997 until 30 December 2008, based on Historical Simulation technique and on the previous 250 trading days.

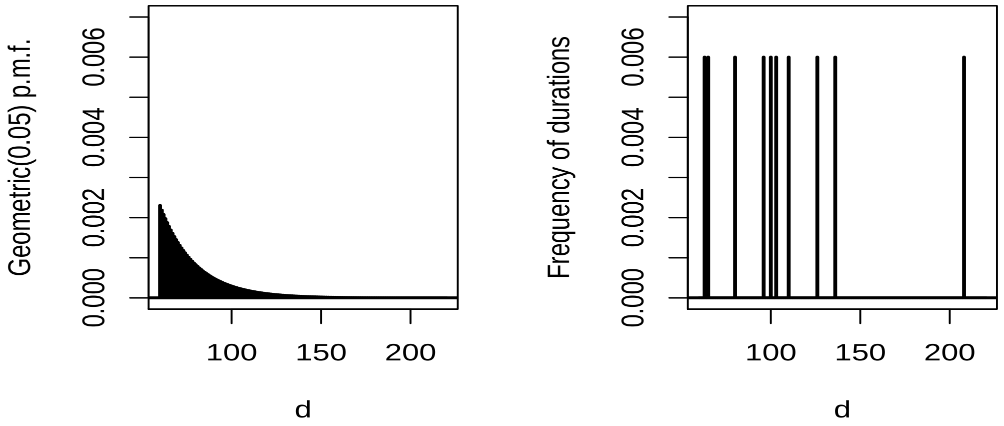

Figura 3.2: Geometric($\pi = 0.05$) probability mass function (left), for durations above 60 days, and frequency of durations (right) between violations of VaR(0.05) for DAX index from January 2, 1997, through December 30 2008, based on Historical Simulation technique and on the previous 250 trading days.

Figure 3.1 gives the plot of probability mass function of the geometric, with $\pi = 0.05$, and the frequency of durations between violations. We consider a long period and the sample size for durations (N) is 170. For short durations, the frequencies in the frequency plot are much higher than the probability masses in the geometric p.m.f. The majority of durations are short, equal or lower than 6 days and consequently the empirical median is 6, contrasting with the expected value of $D_{85:170}$, under the IND hypothesis, which is close to 14. Moreover, it is evident from the Figure 3.2 that some durations are very long, for durations above 60 days we note much higher frequencies in the frequency plot than the probability masses in the geometric p.m.f. The maximum duration, $d_{170:170}$, is 208 days, a very tranquil period of almost one year without a violation, between May 19, 2003, and March 11, 2004. This maximum duration is almost the double of the expected value of $D_{170:170}$ under the IND hypothesis, which is close to 112. In this example, where violations occur in clusters, the majority of durations are short, some durations are very long and both o.s.´s $D_{[N/2]:N}$ and $D_{N:N}$ give strong evidence against the IND hypothesis.

The Markov tests (2.4) and (2.3) are only sensible to violations followed immediately by other violations. The CAViaR test (2.6) and the GMM tests (2.14) and (2.15) do not define a alternative hypothesis related with clustering. This is a drawback of the available tests. To understand how important is a convenient definition of clustering, we use an example with 20 violations in 500 days, separated as much as possible. Figure 3.3 displays the hit sequence. We apply the GMM independence test (2.15) with $k = 3$ and $k = 5$, which do not define a alternative hypothesis related with clustering. The observed values of the test statistics are respectively 11.61 and 23.99, which clearly conducts to the rejection of the null hypothesis with

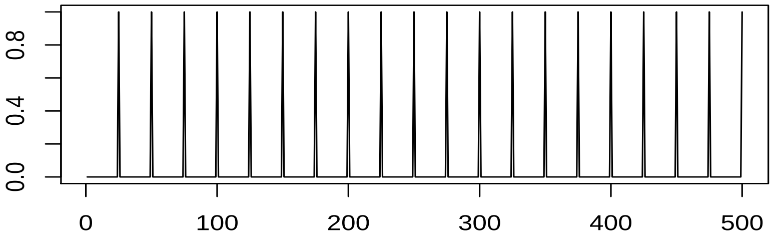

Figura 3.3: Hit sequence with 20 violations in 500 days, with homogeneous separation.

0.01 significance level. But we have the opposite of clustering! With the available tests, we can reject the independence hypothesis simply because the violations are too much separated from each other, however, too much separated is a good feature in a risk management point of view.

With the motivation presented at the beginning of the section and considering the need for a convenient definition of clustering explained with an example, we propose the following definitions:

Definition 3.1 (Tendency to clustering of violations)**.** *A hit function (1.2) has tendency to clustering of violations if the expected value of $D_{N:N}/D_{[N/2]:N}$ is higher than the expected value under the independence hypothesis (2.10).*

Definition 3.2 (Tendency to homogeneous separation)**.** *A hit function (1.2) has tendency to homogeneous separation if the expected value of $D_{N:N}/D_{[N/2]:N}$ is lower than the expected value under the independence hypothesis (2.10).*

For explicitly testing the independence hypothesis (2.10) versus tendency for clustering of violations or tendency for homogeneous separation, we propose the following *Maximum to Median Ratio* statistics and the specific results presented in propositions 3.4. and 3.6.

$$R_N^G = \frac{D_{N:N} - 1}{D_{[N/2]:N}}. \tag{3.2}$$

$$R_N^{G+} = \frac{D_{N:N}}{D_{[N/2]:N} - 1}. \tag{3.3}$$

In Remarks 3.3 and 3.4 we will explain the corrections made to the maximum and the median. We use $[N/2]$ for both N even and N odd, it is simple and a simulation study support this choice.

We will denote Y_i instead of D_i, the durations observed between two consecutive violations, when we relax to a continuous model, i.e., when we use the continuous model (2.11) and not the discrete model (2.8), under the hypothesis of independence. From now on, we denote

$$a_w = \left(\begin{array}{c} N - [N/2] - 1 \\ w \end{array} \right), \quad b_s = \left(\begin{array}{c} [N/2] - 1 \\ s \end{array} \right),$$

$$c_{w,s} = N - [N/2] - w + s \quad , \gamma_N = \frac{N!}{([N/2]-1)!(N-[N/2]-1)!} \quad \text{and}$$

$$R_N^E = \frac{Y_{N:N}}{Y_{[N/2]:N}}. \tag{3.4}$$

Proposition 3.1. *Let us consider a sequence of N durations, denoted $\{Y_i\}_{i=1}^N$, observed between two sucessive violations associated to a interval forecast method. Assuming for (2.8) the continuous analogue (2.11), under the null hypothesis of independence (2.10) the distribution function of (3.4) is*

$$F_{R_N^E}(r) = 1 - \gamma_N \sum_{w=0}^{N-[N/2]-1} \sum_{s=0}^{[N/2]-1} (-1)^{w+s} a_w b_s \big([c_{w,s}(w+1)]^{-1} - [c_{w,s}(w+1+c_{w,s}(1/r)]^{-1} \big). \tag{3.5}$$

with $1 \leq r \leq \infty$.

Proof: Under the independence hypothesis and assuming for (2.8) the continuous analogue (2.11), the density of Y_i is equal to (2.11). For the Weibull distribution with density function $f_X(x;p;\theta) = \theta p (xp)^{\theta-1} \mathrm{e}^{-(px)^\theta}$, $x > 0$, $p > 0$, $\theta > 0$, Malik and Trudel (1982) proved that the density of the ratio of the i-th and j-th o.s.'s with $i < j \leq N$, is

$$\begin{array}{rcl} f_{Z_N}(z;p;\theta) & = & \frac{\theta C_j}{(i-1)!(j-i-1)!} \sum_{w=0}^{j-i-1} \sum_{s=0}^{i-1} (-1)^{w+s} \left(\begin{array}{c} j-i-1 \\ w \end{array} \right) \\ & & \left(\begin{array}{c} i-1 \\ s \end{array} \right) z^{\theta-1} [N-j+w+1+(j-i-w+s)z^\theta]^{-2}, \end{array} \tag{3.6}$$

with $0 \leq w \leq 1$ and where $C_j = \prod_{v=1}^{j} (N-v+1)$. To obtain the ratio of the i-th and j-th o.s.'s, with $i < j \leq N$, from the (2.11) model, in (3.6) we substitute θ by 1. We also replace i and j respectively by $[N/2]$ and N to get

$$f_{Z_N}(w) = \gamma_N \sum_{w=0}^{N-[N/2]-1} \sum_{s=0}^{[N/2]-1} (-1)^{w+s} a_w b_s [w+1+c_{w,s} z]^{-2},$$

the density function of the ratio $Z_N = Y_{[n/2]:N}/Y_{N:N}$ from the (2.11) model.

The distribution function for this ratio is

$$\begin{aligned} F_{Z_N}(z) &= \int_0^z f_Z(u)du \\ &= \gamma_N \sum_{w=0}^{N-[N/2]-1} \sum_{s=0}^{[N/2]-1} (-1)^{w+s} a_w b_s \big(-[c_{w,s}(w+1+c_{w,s}u)]^{-1} \big]_0^z \\ &= \gamma_N \sum_{w=0}^{N-[N/2]-1} \sum_{s=0}^{[N/2]-1} (-1)^{w+s} a_w b_s \big([c_{w,s}(w+1)]^{-1} - [c_{w,s}(w+1+c_{w,s}z)]^{-1} \big), \end{aligned}$$

with $0 \leq z \leq 1$. For $R_N^E = Y_{N:N}/Y_{[N/2]:N} = 1/Z_N$ the distribution function is

$$\begin{aligned} F_{R_N^E}(r) &= P(Y_{[N/2]:N}/Y_{N:N} > 1/r) = 1 - F_{Z_N}(1/r) \\ &= 1 - \gamma_N \sum_{w=0}^{N-[N/2]-1} \sum_{s=0}^{[N/2]-1} (-1)^{w+s} a_w b_s \\ &\quad \big([c_{w,s}(w+1)]^{-1} - [c_{w,s}(w+1+c_{w,s}(1/r)]^{-1} \big), \end{aligned}$$

with $1 \leq r \leq \infty$. □

Proposition 3.2. *Let us consider a sequence of N durations, denoted $\{Y_i\}_{i=1}^N$, observed between two successive violations associated to a interval forecast method. Assuming for (2.8) the continuous analogue (2.11), we express the null hypothesis of independence (2.10) and the alternative hypothesis of tendency to clustering of violations, as*

$$H_0 : Y_i \overset{iid}{\sim} Y \sim \textit{Exponencial}(\beta), \ \textit{with } \beta > 0 \textit{ and } i = 1, ..., N \tag{3.7}$$

$$H_1 : E[R_N^E] > \mu_R, \tag{3.8}$$

where μ_R denotes the expected value of (3.4) when (3.7) is true. We reject (3.7) if $R_N^E > r_\alpha$ and the critical point r_α for the test with size α can be computed replacing $F_{R_N^E}(r)$ by $1-\alpha$ and solving the equation (3.5) in order to r.

Proof: $P(R_N^E > r_\alpha) = \alpha \Leftrightarrow 1 - F_{R_N^E}(r_\alpha) = \alpha$ and using the Proposition 3.1 the result follows. □

Remark 3.1. *We note that $P[I_{t+1}(p) = 1] \neq p$ does not influence the statistic (3.4), i.e., the statistic is robust with respect to p. Moreover, notice that this ratio statistic is invariant to scale-β transformations. This is a main advantage because to test explicitly the independence hypothesis, the probability of violation is unknown. As far as we know, all other independence tests available suffers from the problem of having this nuisance parameter.*

Remark 3.2. *Other main advantage of the test presented in Proposition 3.2, is knowing the exact distribution of the test statistic, established in Proposition 3.1. All the independence tests available, that we know, are based on asymptotic distributions* *and suffer from small sample bias. To aggravate the problem, with the presence of a nuisance parameter is impossible to control the size of the tests using the Monte Carlo testing approach of Dufour (2006) as other authors do for the case of joint testing both UC and IND (e.g. Christoffersen and Pelletier (2004), Candelon et al. (2008) and Berkowitz et al. (2009)); see the paper of Dufour (2006) for details. Knowing the exact distribution of the test statistic, eliminates the problem of small sample bias involved when we use a asymptotic distribution.*

With Remarks 3.1 and 3.2 we note two main advantages of the test presented in Proposition 3.2, but remains one drawback, assuming for (2.8) the continuous analogue (2.11). Proposition 3.4 eliminates this drawback, presenting a test which does not need this assumption, but first we establish a relation between the distribution functions of (3.2) and (3.4).

Proposition 3.3. *Let $D_1, ..., D_N$, be i.i.d. random variables whose common probability mass function is (2.8) with $\pi = p \in (0,1)$ and let $Y_1, ..., Y_N$, be i.i.d. exponential random variables whose common density function is (2.11) with $\beta > 0$. If we consider (3.2) and (3.4), then we have*

$$F^{\leftarrow}_{R^G_N}(1-\alpha) < F^{\leftarrow}_{R^E_N}(1-\alpha), \text{ for all } 0<p<1,\ \beta>0 \text{ and } 0<\alpha<1, \tag{3.9}$$

where $F^{\leftarrow}(t) := \inf\{x : F(x) \geq t\}$ denotes the generalized inverse function of F.

Proof: Let Y be an exponential random variable with density function (2.11) and denote $[Y]$ the integer part of Y and $< Y >$ the fractional part of Y. If we define $X = [Y] + 1$, then

$$\begin{aligned} P[X = x] &= P[x \leq Y + 1 < x + 1] \\ &= F_Y(x) - F_Y(x-1) \\ &= \exp(-\beta(x-1)) - \exp(-\beta x) \\ &= \big(\exp(-\beta)\big)^{(x-1)} - \Big(\big(\exp(-\beta)\big)^{(x-1)}\exp(-\beta)\Big) \\ &= \big(\exp(-\beta)\big)^{(x-1)}\big(1 - \exp(-\beta)\big) \end{aligned}$$

with $x \in \aleph^*$. Note that X, the integral part of Y plus one, is distributed as geometric with probability of success $(1 - \exp(-\beta))$. Now, for $p = (1 - \exp(-\beta))$, $D_{i:N} \overset{d}{=} X_{i:N} = [Y]_{i:N} + 1 \overset{d}{=} [Y_{i:N}] + 1$ and we have

$$R^G_N = \frac{D_{N:N} - 1}{D_{[N/2]:N}} \overset{d}{=} \frac{[Y_{N:N}]}{[Y_{[N/2]:N}] + 1} < \frac{[Y_{N:N}] + < Y_{N:N} >}{[Y_{[N/2]:N}] + < Y_{[N/2]:N} >} = \frac{Y_{N:N}}{Y_{[N/2]:N}} = R^E_N.$$

Since the distribution of R^E_N only depends on N and is the same for all $\beta > 0$, the result (3.9) follows. □

Proposition 3.4 (Independence versus Tendency to clustering of violations). *Let us consider a sequence of* N *durations (2.7), denoted* $\{D_i\}_{i=1}^{N}$*, observed between two successive violations associated to a interval forecast method. We express the null hypothesis of independence (2.10) and the alternative hypothesis of tendency to clustering of violations, as*

$$H_{0,IND} : D_i \overset{iid}{\sim} D \sim \textit{Geometric}(\pi), \text{ with } 0 < \pi < 1 \text{ and } i = 1, ..., N \qquad (3.10)$$

$$H_1 : E[R_N^G] > \mu_R, \qquad (3.11)$$

where μ_R *denotes the expected value of (3.2) when (3.10) is true. We reject (3.10) if* $R_N^G > r_\alpha$ *and the critical point* r_α *for the test with level* α *can be computed replacing* $F_{R_N^E}(r)$ *by* $1 - \alpha$ *and solving the equation (3.5) in order to* r.

Proof: By Proposition 3.3 we can write $P(R_N^G > r_\alpha) < P(R_N^E > r_\alpha) = \alpha \Leftrightarrow 1 - F_{R_N^G}(r_\alpha) < 1 - F_{R_N^E}(r_\alpha) = \alpha$. Now, using Proposition 3.1, the result follows. □

Remark 3.3. *Propositions 3.3 and 3.4 show that the critical point* r_α *obtained using equation (3.5) implies a conservative approach with* $P[\textit{Type I Error}] \leq \alpha$*, i.e., we have a test of level* α *and not of size* α*. If we can use the distribution function of (3.2), the critical value will be lower for all* $0 < p < 1$*, however we need to assume a value for* $P[I_{t+1}(p) = 1]$*. The subtraction of one to the maximum is a correction which allows us to obtain a test of level* α *using critical points computed from (3.5), but considering the true probabilistic behavior (2.10) of the durations (2.7), under the null hypothesis, and not assuming the continuous analogue. Additionally, we do not have a nuisance parameter, since the decision of the test presented in Proposition 3.4 is valid for all* $0 < P[I_{t+1}(p) = 1] < 1$.

We now present similar results for the other unilateral independence test, involving the tendency to homogeneous separation.

Proposition 3.5. *Let* $D_1, ..., D_n$*, be i.i.d. random variables whose common probability mass function is (2.8) with* $\pi = p \in (0, 1)$ *and let* $Y_1, ..., Y_n$*, be i.i.d. exponential random variables whose common density function is (2.11) with* $\beta > 0$*. If we consider (3.3) and (3.4), then we have*

$$F^{\leftarrow}_{R_N^{G+}}(1-\alpha) > F^{\leftarrow}_{R_N^E}(1-\alpha), \text{ for all } 0 < p < 1, \ \beta > 0 \text{ and } 0 < \alpha < 1, \qquad (3.12)$$

where $F^{\leftarrow}(t) := \inf\{x : F(x) \geq t\}$ *denotes the generalized inverse function of* F.

Proof: The first part of the proof is the same as in Proposition 3.4, then we write

$$R_N^{G+} = \frac{D_{N:N}}{D_{[N/2]:N} - 1} \overset{d}{=} \frac{[Y_{N:N}] + 1}{[Y_{[N/2]:N}]} > \frac{[Y_{N:N}] + < Y_{N:N} >}{[Y_{[N/2]:N}] + < Y_{[N/2]:N} >} = \frac{Y_{N:N}}{Y_{[N/2]:N}} = R_N^E.$$

Since the distribution of R_N^E only depends on N and is the same for all $\beta > 0$, the result (3.12) follows. □

Proposition 3.6 (Independence versus Tendency to homogeneous separation). *Let us consider a sequence of N durations (2.7), denoted $\{D_i\}_{i=1}^{N}$, observed between two successive violations associated to a interval forecast method. We express the null hypothesis of independence (2.10) and the alternative hypothesis of tendency to homogeneous separation, as*

$$H_0 : D_i \overset{iid}{\sim} D \sim \textit{Geometric}(\pi), \quad \textit{with } 0 < \pi < 1 \quad \textit{and } i = 1, ..., N \tag{3.13}$$

$$H_1 : E[R_N^{G+}] < \mu_{R+}, \tag{3.14}$$

where μ_{R+} denotes the expected value of (3.3) when (3.13) is true. We reject (3.13) if $R_N^{G+} < r_{1-\alpha}$ and the critical point $r_{1-\alpha}$ for the test with level α can be computed replacing $F_{R_N^E}(r)$ by α and solving the equation (3.5) in order to r.

Proof: By Proposition 3.5 we can write $P(R_N^{G+} \leq r_{1-\alpha}) < P(R_N^E < r_{1-\alpha}) = \alpha \Leftrightarrow F_{R_N^{G+}}(r_{1-\alpha}) < F_{R_N^E}(r_{1-\alpha}) = \alpha$. Now, using proposition 3.1, the result follows. □

Remark 3.4. *Subtracting one to the median is a correction which allows us to obtain a test of level α using critical points computed from (3.5), but considering the true probabilistic behavior (2.10) of the durations (2.7), under the null hypothesis, and not assuming the continuous analogue. As in the Proposition 3.4, the decision of the test presented in Proposition 3.6 is valid for all* $0 < P[I_{t+1}(p) = 1] < 1$.

To implement the test, we provide in the appendix a statistical table with $2 \leq N \leq 200$ and critical values $r_{0.95}$, $r_{0.1}$, $r_{0.05}$ and $r_{0.01}$. We use both the equation (3.5) and a simulation procedure to construct the table. Testing independence versus tendency to clustering of violations implies to calculate the assumed value by the R_N^G statistic (3.2) and reject the null hypothesis (3.10) if the assumed value is higher than the critical value r_α. Testing independence versus tendency to homogeneous separation implies calculate the assumed value by the R_N^{G+} statistic (3.3) and reject the null hypothesis (3.10) if the assumed value is lower than the critical value $r_{1-\alpha}$.

To exemplify this procedure, now, we apply our framework to the hit sequence presented in figure 3.3. With this data, the observed values for the statistics (3.2) and (3.3) are respectively $r_N^G = (50-1)/50 = 0.98$ and $r_N^{G+} = 50/(50-1) = 1.0204$. With this very small value $r_N^G = 0.98$, there is no evidence of tendency to clustering of violations and we only apply the test involving tendency to homogeneous separation presented in Proposition 3.6. Consulting the statistical table provided in the appendix, with $N = 19$ and $r_{0.95}$, we get a critical region equal to $]1; 2.91]$. The observed value clearly belongs to this region and we reject the null hypothesis with strong evidence of tendency to homogeneous separation. This is an advantage of our test. We also reject the independence hypothesis with this data, however, we reject it in favor of a tendency to homogeneous separation alternative hypothesis, clarifying that there is no tendency to clustering of violations but exactly the opposite tendency.

The available tests simply reject the independence property and do not clarify in a convenient way a alternative hypothesis of clustering, or define a very narrow alternative hypothesis. As we see in the example of the Figure 3.3, the available tests can conduct to the rejection of independence simply because the hit function has tendency to homogeneous separation.

In a risk management point of view, the problematic non verification of the independence property is the one that conducts to cluster of violations, therefore, we are interest in the application of the independence versus tendency to clustering of violations test, proposed in Proposition 3.4. In section 4 we apply this test in a simulation context, showing that the power is other fundamental advantage. In section 5, we will apply the test with a real data set.

4 Comparative Simulation Study

In the context of a Monte Carlo study and for different alternative hypothesis, we compare the power of the test we propose in proposition 3.4 (R_n^G) for independence versus clustering, with the Markov independence (2.3), the CAViaR (2.6) and the GMM independence (2.15) tests. From now on, we will denote this tests respectively by M_{IND}, $CAViaR$ and $J_{IND}(k)$. We use the R language and the fGarch package.

Following other authors (e.g. Christofferson (1998), Christofferson and Pelletier (2004), Haas (2005), Candelon *et al.* (2008) and Berkowitz *et al.* (2009)) we consider a GARCH specification for the returns process. Additionally, we use a APARCH model which nests some of the GARCH models with leverage effect (for a recent review of models with leverage see Rodriguez and Ruiz (2009)). Specifically we assume two models for the alternative hypothesis:

- Gaussian GARCH(1,1) model (Bollerslev (1986)),

  $$\begin{aligned} R_{t+1} &= \sigma_{t+1} z_{t+1} \\ \sigma_{t+1}^2 &= w + \alpha z_t + \beta \sigma_t^2 \end{aligned} \tag{4.1}$$

  where the innovation z_{t+1}´s are drawn independently from a standard normal distribution. As Christofferson (1998), we choose the parametrization $w = 0.05$, $\alpha = 0.1$ and $\beta = 0.85$.

- APARCH(1,1) model (Ding *et al.* (1993)),

  $$\begin{aligned} R_{t+1} &= \sigma_{t+1} \varepsilon_{t+1} \\ \sigma_{t+1}^\delta &= w + \alpha(|\varepsilon_t| - \xi \varepsilon_t)^\delta + \beta \sigma_t^\delta \end{aligned} \tag{4.2}$$

  where the innovation ε_{t+1}´s are drawn independently from a skewed Student´s t(ν) distribution with asymmetry coefficient ξ, a distribution proposed by Fernández and Steel (1998). The asymmetry coefficient (ξ) is defined such that the ratio of probability masses above and below the mean is

$P(\varepsilon_{t+1} \geq 0|\xi)/P(\varepsilon_{t+1} < 0|\xi) = \xi^2$. Giot and Laurent (2003) have shown that the skewed Student APARCH model has provided more accurate VaR forecasts than more simple ARCH-type models, for both tails. We estimate this model with the aim of achieving a realistic representation of a portfolio return distribution. More precisely, we assume a portfolio that replicates the DAX index and we use data from the first trading day of 2006 until the end of 2008, for estimation. The parametrization achieved is $w = 0.0014$, $\alpha = 0.095$, $\gamma = 1$, $\beta = 0.9$, $\delta = 0.76$, $\xi = 0.81$ and $\nu = 10$.

Besides the simulation of the returns, it is necessary to select a method to forecast the VaR, which violate the independence property, i.e., with a tendency to form clusters of violations when applied to heterocedastic processes. As in other power studies with the same purpose, we choose the Historical Simulation method (HS) presented in (1.5) which easily generates clusters of violations. To illustrate this, in Figure 4.1 we present simulated returns and corresponding one-day-ahead 1% VaR from a observed simulation using the APARCH model(4.2). It seems that clustering is evident and the behavior of the series seems quite similar to the example of DAX index presented in Figure 1.1. ($T = 250$).

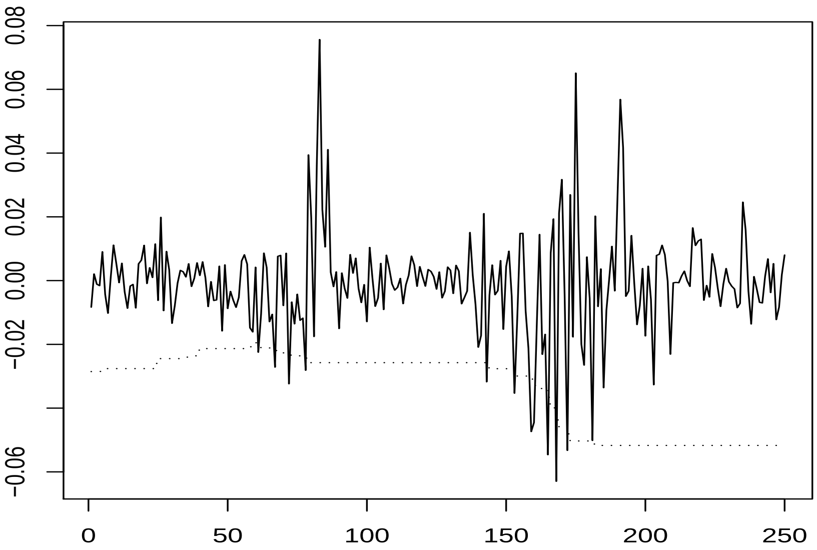

Figura 4.1: APARCH Simulated Portfolio Returns (solid line) and one-day-ahead 1% VaR(0.01) (dotted line) based on Historical Simulation technique and on the previous 250 days.

We conduct our power experiment with sample sizes ranging from 500 to 1250 days with increments of 250 days, corresponding approximately to two to five years. We set the size of the rolling window (T_e) equal to 250 days and we choose the VaR coverage rate p equal to 1% and 5%, the values typically chosen in practice.

For each sample size and coverage rate p, we simulate GARCH returns using the models (4.1) and (4.2), calculate HS VaR´s (1.5) and the various test statistics over 10,000 Monte Carlo replications. The empirical power of the tests is calculated from rejection frequencies excluding the samples with less than 2 violations based in the argument invoked by Christofferson and Pelletier (2004). It appears to be realistic that a risk management team would not start backtesting unless at least a couple of violations had occurred. In our study the frequencies of excluded samples (FES) is very small, however, for transparency we report FES in the tables.

As we noted in Remark 3.1, in practice under IND hypothesis, when we want to test explicitly the independence property, the probability $P[I_{t+1} = 1]$ is unknown, we have a nuisance parameter and is impossible to have a test of size α using a Monte Carlo approach. When we assume a probability of violation equal to p to apply the Monte Carlo testing technique, then the null hypothesis is CC and not IND. We recall that our goal is to test explicitly IND and not jointly test UC and IND. For this reason and for all test statistics except (3.2), we apply the asymptotic distributions to find critical values, noting and being aware of the strong limitations especially in the small sample cases. For the R_n^G statistic (3.2) we have a test based on a exact distribution given by the Proposition 3.1., therefore the test are free from the serious limitation of small sample bias caused by the use of an asymptotic distribution.

Table 4.1 - Power of Finite Sample Tests on 1% VaR ($\alpha = 0.1$).
Gaussian GARCH(1,1) Model.

	T=500	T=750	T=1000	T=1250
R_n^G	0.429	0.567	0.627	0.663
M_{IND}	0.154	0.205	0.211	0.239
CAViaR	0.415	0.488	0.558	0.611
$J_{IND(3)}$	0.177	0.288	0.365	0.458
$J_{IND(5)}$	0.158	0.275	0.352	0.444
FES	0.040	0.003	0.000	0.000

Table 4.2 - Power of Finite Sample Tests on 1% VaR. ($\alpha = 0.1$).
APARCH(1,1) - skewed t(10) Model.

	T=500	T=750	T=1000	T=1250
R_n^G	0.553	0.725	0.820	0.852
M_{IND}	0.232	0.324	0.362	0.425
CAViaR	0.501	0.619	0.682	0.745
$J_{IND(3)}$	0.309	0.473	0.601	0.698
$J_{IND(5)}$	0.296	0.471	0.597	0.704
FES	0.064	0.005	0.000	0.000

Notes to Tables: The results are based on 10,000 replications. For sample sizes from 500 to 1250, with increments of 250, we provide the percentage of rejection at a 10% level. We also present the percentage of excluded samples (FES).

We report the empirical rejection frequencies (power) for our test, the Markov independence, CAViaR and GMM independence tests ($k = 3$ and $k = 5$), in tables 4.1, 4.2, 4.3 and 4.4.

Results for the 1% VaR are presented in tables 4.1 and 4.2. In this case with $p = 0.01$, the coverage rate required by the regulatory framework, for all sample sizes and both models (4.1) and (4.2), our test R_n^G is more powerful than all other tests. The differences in power are, in the majority of the cases, very large. Compared with the Markov independence test, the rejection frequency of our test is for all cases more than the double and in some cases almost tree times the rejection frequency of the Markov test. Compared with the GMM independence tests, our test performs better for all cases and sometimes present more than the double of the power. The CAViaR test, although with less power than our test in all cases, presents a similar performance with $T = 500$.

Table 4.3 - Power of Finite Sample Tests on 5% VaR. ($\alpha = 0.1$).
Gaussian GARCH(1,1) Model.

	T=500	T=750	T=1000	T=1250
R_n^G	0.565	0.694	0.757	0.800
M_{IND}	0.242	0.327	0.375	0.422
CAViaR	0.514	0.570	0.623	0.679
$J_{IND(3)}$	0.441	0.639	0.752	0.835
$J_{IND(5)}$	0.372	0.559	0.675	0.773
FES	0.000	0.000	0.000	0.000

Table 4.4 - Power of Finite Sample Tests on 5% VaR. ($\alpha = 0.1$).
APARCH(1,1) - skewed t(10) Model.

	T=500	T=750	T=1000	T=1250
R_n^G	0.826	0.928	0.973	0.984
M_{IND}	0.446	0.588	0.693	0.767
CAViaR	0.718	0.783	0.847	0.882
$J_{IND(3)}$	0.755	0.914	0.971	0.989
$J_{IND(5)}$	0.698	0.887	0.953	0.983
FES	0.000	0.000	0.000	0.000

Notes to Tables: The results are based on 10,000 replications. For sample sizes from 500 to 1250, with increments of 250, we provide the percentage of rejection at a 10% level. We also present the percentage of excluded samples (FES).

Results for the 5% VaR are presented in tables 4.3 and 4.4. In this case, for all sample sizes and both models (4.1) and (4.2), our test R_n^G is more powerful than all other tests, only with two exceptions when we simulate for the the larger sample size, $T = 1250$, but with very small differences. Here the $J_{IND}(3)$ performs a bit better. The GMM tests performs quite well at larger sample sizes (5% VaR and

$T = 1000$ or $T = 1250$ but poorly at small sample sizes. This results contrasts with good results achieved for the GMM tests by Candelon *et al.* (2008) when jointly test the UC and IND hypothesis with the Monte Carlo testing technique assuming $I_t(p) \overset{iid}{\sim} Bern(\pi)$, with $\pi = p$.

In this section we show with a Monte Carlo simulation study that our exact independence test is superior than the available independence tests, in terms of power of the test. We confirm that the classic and popular Markov independence test performs much worse than all other tests under study. The results also confirm the advantage of larger backtesting sample sizes.

5 Empirical Application

In this section we complete the empirical example presented in section 1, with the application of our exact test (R_n^G) for independence versus tendency to clustering of violations, presented in Proposition 3.4. We also apply the Markov independence test M_{IND} (2.3), the CAViaR test(2.6) and the GMM independence tests $J_{IND}(k)$ (2.15) with $k = 3, 5$, to compare the empirical performance with a real data set. In the example we use the hit sequence from January 2, 2007 through December 30, 2008, generated by the DAX index log returns and one day ahead 1% VaR estimated with the Historical Simulation method, choosing a size for the rolling historical sample equal to 250, i.e., $T_e = 250$. Figure 5.1 displays the hit sequence. From this figure, the non independence of the hit sequence generated by the Historical Simulation model (1.5) is evident, due to the cluster pattern observed. In section 1 we applied the recent regulatory framework to backtesting, but only in the last violation, on November 6, 2008, the model was classified in the red zone and

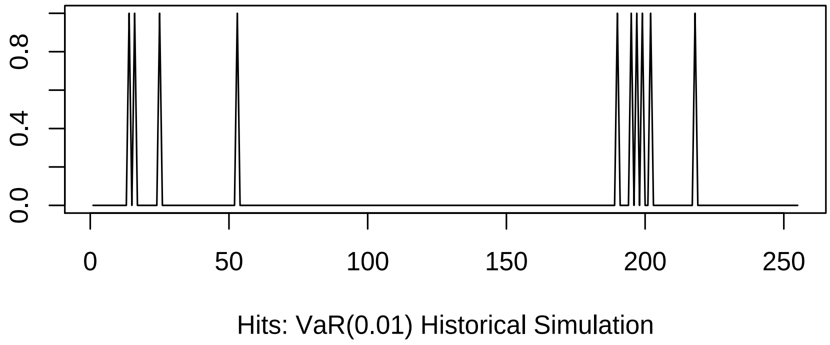

Figura 5.1: Hit sequence from January 2, 2007 through December 30, 2008, generated by the DAX index log returns and one- day-ahead 1% VaR estimated with the Historical Simulation method with $T_e = 250$.

inaccurate. A very bad performance of the recent regulatory framework.

On the date of the first violation of the five violations cluster between September 29, 2008, through October 15, 2008, our exact independence test rejects the IND hypothesis for 0.1 significance level, classifying the forecast model as inaccurate. On this date we have a sample of durations (2.7) of size 4, with the observed values of o.s.´s $d_{1:4} = 2$, $d_{2:4} = 9$, $d_{3:4} = 28$ and $d_{4:4} = 137$ days. The observed value of our test statistic (3.2) is $r_4^G = (d_{4:4} - 1)/d_{2:4} = 15.11$. Consulting the statistical table with quantiles for R_N^G, which we provide in the appendix, with $N = 4$, we get a critical region equal to $[11.69; +\infty[$ for 0.1 significance level. The observed value belongs to this region and we reject the null hypothesis. On the date of the second violation, October 6, 2008, our exact independence test rejects the IND hypothesis for 0.05 significance level. The observed value of our test statistic is $r_5^G = 27.2$ and the critical region, with $N = 5$, obtained from the statistical table in the appendix is equal to $[26.51; +\infty[$. A good performance of our test with this real data, better than the results obtained with all other tests. Table 5.1 resumes the results for all tests. Only on October 8 and October 10, 2008, the $J_{IND}(3)$ rejects the null hypothesis. The $J_{IND}(5)$ only rejects with a 0.1 significance level on October 10, 2008. The CAViaR only rejects with a 0.1 significance level on October 15, 2008. For this real data set, the performance of the Markov test is even worse than the regulatory framework, never rejecting the null hypothesis. The poor performance of CAViaR and especially of the Markov test is explained because in this hit sequence we never had a violation immediately followed by another violation.

With this real data set, the test we propose in Proposition 3.4 is clearly superior to all other tests. Furthermore, this example also illustrate the very important advantage of the exact property of our test, especially, when compared with the alternative duration based test. In this example, we apply the duration based tests with sample sizes ranging from 4 to 7. With such small sample sizes is highly questionable the application of the asymptotic results. We can not apply the Monte Carlo testing technique assuming $I_t(p) \overset{iid}{\sim} Bern(\pi)$, with $\pi = p$, because the null hypothesis involves only independence and not require $\pi = p$. Our test is exact and do not suffer from this serious limitation.

Table 5.1 - Dates of rejection of IND hypothesis

Date	$I_{t+1}(0.01)$	$\alpha = 0.05$	$\alpha = 0.1$
2008-09-29	1		R_4^G
2008-09-30	0		
2008-10-01	0		
2008-10-02	0		
2008-10-03	0		
2008-10-06	1	R_5^G	
2008-10-07	0		
2008-10-08	1		$J_{IND}(3)$
2008-10-09	0		
2008-10-10	1	$J_{IND}(3)$	$J_{IND}(5)$
2008-10-13	0		
2008-10-14	0		
2008-10-15	1		CAViaR

Notes to Table: Starting dates of rejection of IND hypothesis, applying the R_n^G, M_{IND}, CAViaR, $J_{IND}(3)$ and $J_{IND}(5)$ tests to the hit sequence from January 2, 2007 through December 30, 2008, generated by the DAX index log returns and one- day-ahead 1% VaR estimated with the Historical Simulation method with $T_e = 250$.

6 Summary

Christoffersen (1998) shows that the problem of determining whether a given forecast model is accurate can be reduced to examine the uncondicional coverage and independence properties. It is widely recognized that the independence property is equally important as the unconditional coverage property. In this work we show that the available independence tests suffer from serious limitations, namely:

- They do not define, define a very general or a very narrow alternative hypothesis for clustering.
- We do not know the exact distribution of the test statistics and relying in the asymptotic results is highly questionable with realistic small sample sizes. For explicitly testing the independence hypothesis we can not apply the Monte Carlo testing technique assuming $I_t(p) \overset{iid}{\sim} Bern(\pi)$ with $\pi = p$, because the null hypothesis not require $\pi = p$.

Therefore, there is a need for better tests to overcome this limitations. With this goal we proposed a new framework to explicitly test the independence hypothesis, with the following tree fundamental advantages:

- The tests proposed in section 3 are based in a convenient definition of tendency to clustering of violations and tendency to homogeneous separation, overcoming the first limitation.
- The tests proposed in section 3 are exact, allowing us to avoid the second limitation.
- A Monte Carlo experiment shows that our independence versus tendency to clustering of violations test, has better power properties than the available tests.

Finally, in section 5, application to real data provide empirical support to our framework. With the results achieved, we argue that the tests we propose in section 3, with propositions 3.4 and 3.6, should be included in any tool box of backtesting. Although the serious limitations, concerning the formulation of hypothesis and not knowing the exact distribution, we observed in section 4 that the CAViaR and GMM independence tests also showed good power properties in some situations, and for that reason, should not be excluded from the tool box of backtesting.

Referências

[1] Berkowitz, J., Christoffersen, P. and Pelletier, D. (2009). Evaluating Value-at-Risk models with desk-level data. *Management Science*, Published online in Articles in Advance.

[2] Berkowitz, J., Christoffersen, P. and Pelletier, D. (2007). Evaluating Value-at-Risk models with desk-level data. *Working Paper*, University of Houston.

[3] Bontemps, C. (2006). Testing distributional assumptions: A GMM approach. *Working Paper.*

[4] Bontemps, C. (2006). Moment-based tests for discrete distributions. *Working Paper.*

[5] Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31**,307-327.

[6] Candelon, B., Colletaz, G., Hurlin, C., and Tokpavi, S. (2008). Backtesting value-at-Risk: A GMM Duration-Based Test. *HAL, Working Paper.*

[7] Campbell, S.D. (2007). A review of backtesting and backtesting procedures. *Journal of Risk*, **9(2)**, 1-18.

[8] Christoffersen, P. and Pelletier, D. (2004). Backtesting Value-At-Risk: A Duration-Based Approach. *Journal of Financial Econometrics*, **2,1,**,84-108.

[9] Christoffersen, P. (1998). Evaluating Intervals Forecasts. *International Economic Review*, **39**, 841-862.

[10] Dufour, J.M. (2006). Monte Carlo tests with nuisance parameters: a general approach to finite sample inference and nonstandard asymptotics. *Journal of Econometrics*, **127(2)**,443-477.

[11] Danielsson, J. and Morimoto, Y. (2000). Forecasting Extreme Financial Risk: A Critical Analysis of Practical Methods for the Japanese Market. *Monetary and Economic Studies*, **18(2)**, 25-48.

[12] Ding, Z., Engle, RF and Granger, CWJ (1993). A long memory property of stock market return and a new model. *Journal of Empirical Finance*, **1**, 83-106.

[13] Durlauf, S.N. (1991). Spectral Based Testing of the Martingale Hypothesis. *Journal of Econometrics*, **50**, 355-376.

[14] Diethelm Wuertz, Yohan Chalabi and Michal Miklovic (2008). fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. R package version 290.76. http://www.rmetrics.org.

[15] Engel, R.F. and Manganelli (2004). CAViaR: Conditional Autoregressive Value-at-Risk by Regression Quantiles. *Journal of Business and Economics Statistics*, **22**,367-381.

[16] Engle, RF. and Russel, J. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica*, **66**, 1127-1162.

[17] Fernández, C. and Steel, M.F.j. (1998). On Bayesian modelling of fat tails and skewness. *Journal of the American Statistical Association*, **93**, 359-371.

[18] Giot, P. and Laurent, S. (2003). Value-at-risk for long and short trading positions. *Journal of Applied Econometrics*, **18**,641-664.

[19] Haas, M. (2005). Improved duration-based backtesting of Value-at-Risk. *Journal of Risk*, **8(2)**,17-36.

[20] Kupiec, P. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, **3**, 73-84.

[21] Kiefer, N. (1988). Economic Duration Data and Hazard Functions. *Journal of Economic Literature*, **26**, 646-679.

[22] Malik, R.J., Trudel, R. (1982). Probability density function of quotient of order statistics from the pareto, power and weibull distributions. *Communications in Statistics - Theory and Methods*, **11(7)**, 801 - 814.

[23] Margolin, B. H., Winokur, H.S., Jr. (1967). Exact moments of the order statistics of the geometric distribution and their relation to inverse sampling and reliability of redundant systems. *J. Amer. Statist. Assoc.*, **62**, 915 - 925.

[24] Rodriguez, M.J. and Ruiz, E. (2008). GARCH models with leverage effect: Differences and Similarities. *Universidad Carlos III de Madrid, Working Paper 09-03.*

[25] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

APPENDIX: Regulatory Framework for Backtesting VaR models

According to the regulatory framework defined in Market Risk Amend of Basel Capital Accord, the market risk capital that a bank must hold is

$$MRC_t = MAX\Big(VaR_t(0.01), S_t\frac{1}{60}\sum_{i=0}^{59} VaR_{t-i}(0.01)\Big),$$

where S_t is a multiplication factor determined by T_1, the number of violations in the previous 250 trading days, using the correspondence defined in the following table. If more than ten violations occurred in the previous 250 trading days, the backtest result is red and the risk model is classified as inaccurate.

Table - The Basel Penalty Zones

Zone	T_1	S_t
Green	0-4	3
yellow	5	3.4
	6	3.5
	7	3.65
	8	3.75
	9	3.85
Red	10 or more	4

APPENDIX: Table for the R Quantiles

$$P(R \geq r) = \gamma_N \sum_{l=0}^{N-[N/2]-1} \sum_{s=0}^{[N/2]-1} (-1)^{l+s} a_l b_s \int_1^r [2l + 1 + c_{l,s}(1/u)]^{-2} du$$

	$P(R \geq r_{\epsilon(N)}) = \epsilon$					$P(R \geq r_{\epsilon(N)}) = \epsilon$			
ϵ	0.95	0.10	0.05	0.01	ϵ	0.95	0.10	0.05	0.01
N	$r_{0.95(N)}$	$r_{0.1(N)}$	$r_{0.05(N)}$	$r_{0.01(N)}$	N	$r_{0.95(N)}$	$r_{0.1(N)}$	$r_{0.05(N)}$	$r_{0.01(N)}$
2	1.11	18.97	38.84	198.89	50	3.95	9.84	11.22	14.45
3	1.62	42.31	87.08	446.55	51	4.08	10.16	11.58	14.91
4	1.38	11.69	17.73	43.11	52	4.00	9.86	11.24	14.44
5	1.83	17.53	26.57	64.64	53	4.13	10.17	11.58	14.88
6	1.64	10.27	14.12	27.30	54	4.05	9.90	11.25	14.43
7	2.04	13.49	18.53	35.73	55	4.18	10.20	11.58	14.83
8	1.87	9.78	12.83	22.18	56	4.10	9.92	11.27	14.43
9	2.22	11.97	15.65	27.01	57	4.22	10.22	11.59	14.85
10	2.06	9.53	12.13	19.62	58	4.15	9.96	11.30	14.40
11	2.39	11.21	14.24	23.07	59	4.27	10.23	11.61	14.82
12	2.24	9.42	11.74	18.18	60	4.20	9.98	11.31	14.40
13	2.54	10.77	13.43	20.74	61	4.32	10.26	11.62	14.80
14	2.39	9.35	11.50	17.27	62	4.24	10.01	11.33	14.39
15	2.67	10.51	12.90	19.34	63	4.36	10.28	11.63	14.78
16	2.54	9.34	11.35	16.61	64	4.29	10.04	11.36	14.39
17	2.80	10.35	12.55	18.34	65	4.40	10.29	11.63	14.76
18	2.67	9.35	11.24	16.15	66	4.33	10.07	11.37	14.40
19	2.91	10.22	12.30	17.66	67	4.44	10.32	11.65	14.76
20	2.79	9.37	11.19	15.80	68	4.37	10.09	11.40	14.41
21	3.02	10.14	12.11	17.07	69	4.48	10.33	11.66	14.76
22	2.90	9.38	11.13	15.51	70	4.41	10.12	11.41	14.41
23	3.12	10.09	11.98	16.67	71	4.52	10.36	11.69	14.75
24	3.00	9.41	11.10	15.34	72	4.45	10.15	11.44	14.42
25	3.21	10.06	11.86	16.35	73	4.55	10.37	11.69	14.71
26	3.10	9.43	11.08	15.15	74	4.49	10.17	11.46	14.44
27	3.30	10.04	11.79	16.09	75	4.59	10.39	11.70	14.71
28	3.19	9.47	11.08	14.99	76	4.53	10.20	11.47	14.42
29	3.38	10.03	11.73	15.87	77	4.63	10.42	11.72	14.73
30	3.27	9.50	11.09	14.91	78	4.56	10.23	11.49	14.45
31	3.46	10.02	11.68	15.70	79	4.66	10.44	11.73	14.72
32	3.36	9.53	11.08	14.80	80	4.60	10.24	11.50	14.43
33	3.53	10.03	11.65	15.59	81	4.70	10.45	11.74	14.70
34	3.44	9.57	11.09	14.75	82	4.63	10.27	11.53	14.44
35	3.61	10.04	11.62	15.43	83	4.73	10.47	11.75	14.70
36	3.51	9.60	11.10	14.65	84	4.67	10.30	11.56	14.45
37	3.67	10.04	11.60	15.33	85	4.76	10.50	11.77	14.71
38	3.58	9.64	11.10	14.61	86	4.70	10.32	11.57	14.45
39	3.74	10.06	11.59	15.24	87	4.79	10.51	11.78	14.71
40	3.65	9.68	11.13	14.60	88	4.73	10.34	11.59	14.45
41	3.80	10.07	11.58	15.17	89	4.83	10.52	11.79	14.71
42	3.71	9.71	11.14	14.52	90	4.77	10.36	11.61	14.46
43	3.86	10.08	11.56	15.11	91	4.86	10.55	11.81	14.71
44	3.77	9.73	11.16	14.51	92	4.80	10.39	11.63	14.47
45	3.92	10.11	11.58	15.07	93	4.88	10.56	11.82	14.72
46	3.84	9.77	11.17	14.50	94	4.83	10.41	11.64	14.47
47	3.97	10.12	11.56	14.98	95	4.92	10.58	11.83	14.72
48	3.89	9.81	11.20	14.49	96	4.86	10.43	11.65	14.47
49	4.03	10.14	11.57	14.94	97	4.95	10.61	11.85	14.72

		$P(R \geq r_{\epsilon(N)}) = \epsilon$					$P(R \geq r_{\epsilon(N)}) = \epsilon$		
ϵ	0.95	0.10	0.05	0.01	ϵ	0.95	0.10	0.05	0.01
N	$r_{0.95(N)}$	$r_{0.1(N)}$	$r_{0.05(N)}$	$r_{0.01(N)}$	N	$r_{0.95(N)}$	$r_{0.1(N)}$	$r_{0.05(N)}$	$r_{0.01(N)}$
98	4.89	10.45	11.67	14.48	150	5.51	10.91	12.08	14.74
99	4.97	10.62	11.86	14.72	151	5.57	11.03	12.21	14.89
100	4.92	10.47	11.69	14.50	152	5.53	10.92	12.09	14.75
101	5.00	10.64	11.88	14.73	153	5.59	11.04	12.22	14.89
102	4.95	10.50	11.71	14.51	154	5.54	10.95	12.12	14.76
103	5.03	10.66	11.89	14.73	155	5.61	11.05	12.23	14.91
104	4.97	10.51	11.72	14.52	156	5.56	10.96	12.12	14.77
105	5.06	10.68	11.91	14.73	157	5.62	11.07	12.24	14.93
106	5.00	10.53	11.74	14.52	158	5.58	10.97	12.13	14.77
107	5.08	10.69	11.92	14.73	159	5.64	11.07	12.25	14.92
108	5.03	10.55	11.75	14.52	160	5.60	10.98	12.13	14.78
109	5.11	10.71	11.94	14.74	161	5.66	11.09	12.26	14.93
110	5.06	10.57	11.77	14.53	162	5.62	11.00	12.16	14.80
111	5.13	10.72	11.94	14.76	163	5.68	11.10	12.28	14.94
112	5.08	10.59	11.79	14.56	164	5.64	11.01	12.16	14.81
113	5.16	10.75	11.97	14.80	165	5.70	11.11	12.28	14.95
114	5.11	10.61	11.82	14.57	166	5.66	11.02	12.18	14.82
115	5.18	10.75	11.98	14.78	167	5.72	11.12	12.29	14.95
116	5.13	10.62	11.83	14.61	168	5.67	11.04	12.19	14.82
117	5.21	10.78	11.20	14.80	169	5.73	11.13	12.30	14.96
118	5.16	10.66	11.86	14.60	170	5.69	11.05	12.21	14.83
119	5.23	10.79	11.99	14.75	171	5.75	11.15	12.32	14.97
120	5.18	10.67	11.87	14.61	172	5.71	11.06	12.21	14.84
121	5.25	10.80	12.01	14.79	173	5.76	11.17	12.33	14.97
122	5.21	10.68	11.87	14.63	174	5.73	11.07	12.22	14.84
123	5.28	10.83	12.04	14.81	175	5.78	11.18	12.34	14.99
124	5.23	10.70	11.89	14.61	176	5.74	11.09	12.24	14.86
125	5.30	10.83	12.03	14.79	177	5.80	11.19	12.35	14.99
126	5.25	10.72	11.91	14.63	178	5.76	11.11	12.26	14.87
127	5.32	10.86	12.06	14.83	179	5.81	11.20	12.36	15.00
128	5.27	10.74	11.93	14.65	180	5.77	11.12	12.27	14.89
129	5.34	10.86	12.07	14.81	181	5.83	11.21	12.37	15.01
130	5.30	10.74	11.93	14.65	182	5.79	11.13	12.29	14.91
131	5.37	10.89	12.08	14.83	183	5.85	11.23	12.39	15.02
132	5.32	10.77	11.96	14.67	184	5.81	11.15	12.30	14.91
133	5.39	10.91	12.10	14.83	185	5.86	11.25	12.40	15.04
134	5.34	10.79	11.98	14.68	186	5.82	11.16	12.31	14.92
135	5.41	10.91	12.10	14.85	187	5.87	11.26	12.41	15.04
136	5.37	10.81	11.99	14.69	188	5.84	11.17	12.32	14.93
137	5.43	10.93	12.13	14.84	189	5.89	11.27	12.42	15.05
138	5.39	10.81	11.99	14.69	190	5.86	11.18	12.33	14.94
139	5.45	10.95	12.13	14.85	191	5.90	11.27	12.42	15.05
140	5.41	10.83	12.01	14.70	192	5.87	11.19	12.33	14.94
141	5.47	10.96	12.16	14.86	193	5.92	11.28	12.43	15.06
142	5.43	10.85	12.01	14.70	194	5.88	11.21	12.34	14.95
143	5.49	10.96	12.14	14.85	195	5.93	11.29	12.44	15.06
144	5.45	10.86	12.03	14.70	196	5.90	11.22	12.35	14.95
145	5.51	10.98	12.16	14.86	197	5.95	11.31	12.45	15.08
146	5.47	10.87	12.04	14.71	198	5.91	11.23	12.36	14.96
147	5.53	11.01	12.19	14.89	199	5.96	11.32	12.46	15.10
148	5.49	10.89	12.06	14.71	200	5.93	11.24	12.38	14.97
149	5.55	11.01	12.20	14.89					