Directions in Information Systems and Data Analysis: Opportunities and Challenges from a personal perspective
Outline

• Applications
• Problems & Challenges
• Solutions
  – R Framework (Business Model)
  – Reproducible Research
Applications

Analysis of Investment Strategies
Analysis of Investment Strategies

- Data sources:
  - Accounting
  - Macro economics
  - Financial markets
  - Reports, etc...

- Techniques:
  - Technical Analysis
  - Statistics (econometrics, TSA, VaR, MCMC/bootstrap simulations)
  - Data Mining (classifiers, neural networks, ...)

Reasons for studying IS

• Financial
  – Reduce risk (market, operational, human)
  – Improve return
• Monitoring / Auditing / Rating
• HRM
  – Renumeration schemes
  – Training
• ICT (alignment, usability)
Application 1

Rating of Investment Strategies and Hedge Funds
Example: market neutral IS

- Hedge funds often use a market neutral IS:
  - Define a Universe of tradable items (stocks)
  - Define a investment horizon (2 – 30 days)
  - Create a discrimination model that makes 3 piles:
    - Neutral pile (no position)
    - Long pile (buy stocks = long position)
    - Short pile (sell stocks = short position)
  - Hold short and long positions simultaneously during horizon period
Example: conclusion

- The hedge fund managers are unable to adequately demonstrate the qualities of their investment approach and the clients have no information about the risk they are actually taking by using the hedge fund investment vehicle.
Hypothesis

- real stock market time series exhibit fundamental, testable differences when compared to the Random-Walk (Fama-efficient)
Rating Procedure

• Rating is a function of the Expert's ability to discriminate between Real and Random-Walk time series

This equally applies to Experts using:

– Technical analysis
– Statistical models
– No model or technique at all
Fundamental Problem

• This procedure does not take into account exogenous factors such as:
  – Temporal properties of the market (volatility, ...)
  – Geographical properties of the market
  – IS-related restrictions imposed by senior management or by law

• Therefore the rating can only be used for a single case (we cannot compare Experts)
Solution

• Statistical model that discriminates well (low alpha and beta errors)

• The discrimination quality of the model is used as a benchmark to create a relative rating

• This is possible if the model's performance is not too sensitive to external factors (time, place, ...)
Model

- **Quasi Random-Walk (Airoldi, 2001)**


- Marco Airoldi, Correlation Structure and Fat Tails in Finance: a New Mechanism, Risk Management & Research, Intesa-Bci Bank, Milan, Italy (July 30, 2001)

Airoldi [1] formulates his model for $N$ equities $S_i$ for $i = 1, 2, ..., N$ that exhibit movements $\partial S_i = \pm s$ following a Quasi Random-Walk with ”hopping probabilities” $P_{\partial S_i}$ that may depend on previous market returns. He continues to define the states of the market:

\[
\begin{align*}
    h = 1 : P_{\partial S_i} \left( M^{(t-\Delta t)} \right) &= \frac{1}{2} + \frac{1}{2} \frac{\partial S_i^{(t)}}{s} g \left( M^{(t-\Delta t)} \right) \\
    h = 0 : P_{\partial S_i} \left( M^{(t-\Delta t)} \right) &= \frac{1}{2}
\end{align*}
\]

with $M^{(t-\Delta t)} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial S_i^{(t-\Delta t)}}{s}$, $|g(M)| \leq 1$.

Obviously, $P_{\partial S_i}$ depends on the previous market movements if $h = 1$. On the other hand, if $h = 0$ then $P_{\partial S_i}$ behaves like an ordinary Random-Walk and does not depend on market movements.
Simple Logistic Regression

- I expand on this idea and introduce the logistic relationship
  \[ f = \exp(\gamma + \delta X) \]
  where \( P(h=1) = f / (1+f) \)
  and where \( X \) is a “discriminating statistic”

- Estimation: Bias Reduced Logistic Regression:
  
  
  
Best discriminating factor?

- Based on preliminary investigation we identified the p-value of the small sample Kurtosis as the best factor (Vandervorst, Wessa, 2005).

In this study we employ the following (sample) measure of kurtosis for the equities

\( S_i = \{x_1, x_2, \ldots x_n\} \) with \( i = 1, 2, \ldots, N; \)

\[
K_i = \left( \frac{(n-1)n}{(n-2)(n-3)(n-4)} \sum_{j=2}^{n} \left( \frac{r_j - \bar{r}_i}{s} \right)^4 \right) - \frac{3(n-2)^2}{(n-3)(n-4)}
\]

with \( s = \sqrt{\frac{1}{n-2} \sum_{j=2}^{n} (r_j - \bar{r}_i)^2} \) and \( r_j = \nabla \ln x_j \) and

\[
\bar{r}_i = \frac{1}{n-1} \sum_{j=2}^{n} r_j \ [6].
\]

The kurtosis measure \( K_i \) can be used for large and small samples. The standard error of \( K_i \) is \( s_K = \sqrt{\frac{4((n-1)^2-1)s^2}{(n-4)(n+4)}} \) with \( s^2 = \frac{6(n-1)n}{(n-3)n(n+2)} \). The test statistic is \( z = \frac{K_i}{s_K} \leftarrow N (0, 1) \ [6]. \)

as the best factor (Vandervorst, Wessa, 2005).
Figure 3. Type II error of kurtosis-based discrimination model in relationship with type I error (time series length = 100)

Figure 5. Type II error of kurtosis-based discrimination model in relationship with type I error (time series length = 500)
Figure 6. Relationship between required length and desired Type II error (criterion: Kurtosis p-value)

Figure 8. Relationship between required length and desired Type II error (criterion: Autocorrelation)

ally used in empirical research. For example, one might consider an autocorrelation-based discriminating statistic based on $X_{ijq} = \sum_{k=1}^{6} |\rho(\nabla \ln x_t, \nabla \ln x_{t-k})|$ for $t = j(q-1)+1, j(q-1)+2, \ldots, j(q-1)+j, i = 1, 2, \ldots, N$, $j = n_{min}, n_{min}+1, \ldots, n_{max}$, and $q = 1, 2, \ldots, M_{ij}$. This discriminating statistic yields a power between 8.8% (for $j = 100$) and 22.73% (for $j = 500$) when used in the logistic regression instead of the kurtosis p-value. The power
Conclusions

- “Do real stock market time series exhibit fundamental, testable differences when compared to the Random-Walk?”
  -> Yes (Kurtosis p-value works great)

- Bias-Reduced Logistic Regression = non-linear transformation of probabilities

- We can use the model as a benchmark
  => it looks like we can make “fair” comparisons:
  - it only requires re-estimation of the model parameters
  - the model's performance promises to be good over time and place (and other factors?)

- The autocorrelation-based measure requires 4 times more observations to reach the same discrimination quality
Application 2

Analysis and Parameter Estimation of trend-following Investment Strategies
“tKMACD”

Let us define the tKMACD function for an arbitrary equi-distant stock market time series $Y_t$:

$$M(Y_t) = \frac{1}{h_1} \sum_{i=1}^{h_1} \frac{\lambda_i}{\sum_{j=1}^{\lambda_i}} Y_{t-i} - \frac{1}{h_2} \sum_{i=1}^{h_2} \frac{\lambda_i}{\sum_{j=1}^{\lambda_i}} Y_{t-i}$$

for $t = h_2 + 1, h_2 + 2, ..., T$

with $-1 \leq \lambda_{1,2} \leq 1$

and $h_1 < h_2$

There are 4 parameters to be estimated: $h_1, h_2, \lambda_1, \lambda_2$

The trading rule is as follows:

- Buy shares if $M(Y_t) \geq 0$ and $M(Y_{t-1}) < 0$
- Sell shares if $M(Y_t) \leq 0$ and $M(Y_{t-1}) > 0$

We expect that $\lambda_1^* \leq \lambda_2^*$ because $h_1 < h_2$

and $\lambda_{1,2}^* > 0$ in absence of “overshooting”

and $\lambda_{1,2}^* \in [1 - \delta, 1]$ with small $\delta$. 
“Alexander’s Filterrule”

Let us define Alexander’s Filterrule for an arbitrary equi-distant stock market time series $Y_t$ as follows:

$$
N(Y_t) = \frac{(Y_t - \min_{Y_{t \in [t-h, t-1]}} Y_t)}{\min_{Y_{t \in [t-h, t-1]}} Y_t}
$$

$$
X(Y_t) = \frac{(\max_{Y_{t \in [t-k, t-1]}} Y_t - Y_t)}{\max_{Y_{t \in [t-k, t-1]}} Y_t}
$$

for $t = 2, ..., T$

$h$ = # periods since last Sell

$k$ = # periods since last Buy

The trading rule is as follows:

- Buy shares if we don’t already own shares and $N(Y_t) \geq \delta_1$
- Sell shares if we already own shares and $X(Y_t) \geq \delta_2$

There are two parameters $(\delta_1, \delta_2)$ to be estimated but we expect that $\delta_1^* \leq \delta_2^*$. 
$$\text{par3} = \lambda_1$$
$$\text{par4} = \lambda_2$$
par3 = lambda_1
par4 = lambda_2
Expected Return (P50)

Filter P50 - Morgan Stanley

Filter P50 - S&P

par1 = delta_1
par2 = delta_2
par1 = delta_1
par2 = delta_2
Alexander’s Filterrule

Morgan Stanley

Standard & Poors
Application 3

Profit Density Forecasting
Application 4

Portfolio Analysis
Diversification = f(IS)
\[ U = 0.1 \times p50 + 0.9 \times p5 \]
\[ U = 0.5 \times p50 + 0.5 \times p5 \]
Problems & Challenges

So how can things go wrong?
Risks are not independent

The interaction between human and operational risks are often neglected

(solution: reproducibility)
ICT creates problems

- Investment parameters change
  - Horizon & timing
  - Style (large caps, markets, ...)
  - Importance of different types of analysis

  but is ICT able to cope with rapid changes?

- Algorithms change
- Features are added/changed (mostly undocumented)
- Data availability changes
- Do users make adequate use of the Information System?
- How to learn from past mistakes?
- How do we collaborate?
The black box problem

The Inf. System makes things worse

Information System

Decision by the Agent

Process

Evaluation by the Principal

Result

Assignment

Unknown factors

We cannot attribute the Result to the Individual

The reward mechanism is not always transparent

Market
Incentive Scheme

- Reproducibility and Reusability
  - Communication
  - Sociological
  - Collaboration
  - “Reliability”
  - “Innovation”
  - Responsibility
  - Psychological
  - Experimentation
Challenges

We would like to

- produce new
- maintain existing
- publish/implement

applications

- quickly & cheaply (Costs)
- with dissemination in mind (Marketing)
- with usability, scalability, security, flexibility... in mind (Infrastructure)

that generate

- reproducible
- reusable

research.
R Framework (wessa.net)

Business Model

- Open Source / Open Access

- Business & Marketing Models
  (Osterwalder 2004, Constantinides 2004)

  (DSC 2007 Auckland NZ)
  (published in Computational Statistics)
Structure

Operating System

R Language

R package

R wrapper

R Framework (wessa.net)

Meta-data

Distributed Computing & Queue Manager

Compendium Platform (www.freestatistics.org)

Education & Research

Publishing

…

…

…
Example

• We build an R module that computes Bivariate Kernel Density Estimation

# Meta Information

<table>
<thead>
<tr>
<th>Specification of Data Input</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Allow users to specify a subset of the dataset (from - to)</td>
<td>Yes</td>
</tr>
<tr>
<td>Allow users to specify the width and height of charts</td>
<td>Yes</td>
</tr>
<tr>
<td>Allow users to specify the minimum and maximum value of the y-axis</td>
<td>No</td>
</tr>
<tr>
<td>Default Chart Title</td>
<td>Bivariate Kernel Density</td>
</tr>
<tr>
<td>Default X-axis Title</td>
<td>x</td>
</tr>
<tr>
<td>Default Y-axis Title</td>
<td>y</td>
</tr>
<tr>
<td>Number of datasets</td>
<td>2 datasets</td>
</tr>
<tr>
<td>Default dataset 1 (x)</td>
<td>86.6 71.6 93.3 84.3 60.6 75.2 69.7 82 69.4 83.3</td>
</tr>
<tr>
<td>Default dataset 2 (y)</td>
<td>20 16 19.8 18.4 17.1 15.5 14.7 17.1 15.4 15.2 15</td>
</tr>
<tr>
<td>Default dataset 3 (z)</td>
<td></td>
</tr>
<tr>
<td>Number of parameters</td>
<td>7 parameters</td>
</tr>
<tr>
<td>Default value (par1) or comma-separated list</td>
<td>50</td>
</tr>
<tr>
<td>Label of par1</td>
<td>xgridsize</td>
</tr>
<tr>
<td>ToolTip of par1</td>
<td></td>
</tr>
<tr>
<td>Default value (par2) or comma-separated list</td>
<td>50</td>
</tr>
<tr>
<td>Label of par2</td>
<td>ygridsize</td>
</tr>
<tr>
<td>ToolTip of par2</td>
<td></td>
</tr>
<tr>
<td>Default value (par3) or comma-separated list</td>
<td>0</td>
</tr>
<tr>
<td>Label of par3</td>
<td>xbandwidth&lt;br&gt;(zero to use default)</td>
</tr>
<tr>
<td>ToolTip of par3</td>
<td></td>
</tr>
<tr>
<td>Default value (par4) or comma-separated list</td>
<td>0</td>
</tr>
<tr>
<td>Label of par4</td>
<td>ybandwidth&lt;br&gt;(zero to use default)</td>
</tr>
<tr>
<td>ToolTip of par4</td>
<td></td>
</tr>
<tr>
<td>Default value (par5) or comma-separated list</td>
<td>0</td>
</tr>
<tr>
<td>Label of par5</td>
<td>correlation&lt;br&gt;(zero to use actual correlation)</td>
</tr>
<tr>
<td>ToolTip of par5</td>
<td></td>
</tr>
<tr>
<td>Default value (par6) or comma-separated list</td>
<td>Y/N</td>
</tr>
<tr>
<td>Label of par6</td>
<td>display contours&lt;br&gt;(Y/N)</td>
</tr>
<tr>
<td>ToolTip of par6</td>
<td></td>
</tr>
</tbody>
</table>
R code

We include:
- type conversion
- input validation
- actual analysis
R code produces HTML output

```r
load(file = "createtable")
a <- table.start()
a <- table.row.start(a)
a <- table.element(a, "Bandwidth", 2, TRUE)
a <- table.row.end(a)
a <- table.row.start(a)
a <- table.element(a, "x axis", header = TRUE)
a <- table.element(a, par3)
a <- table.row.end(a)
a <- table.row.start(a)
a <- table.element(a, "y axis", header = TRUE)
a <- table.element(a, par4)
a <- table.row.end(a)
a <- table.row.start(a)
a <- table.element(a, "Correlation", 2, TRUE)
a <- table.row.end(a)
a <- table.row.start(a)
a <- table.element(a, "correlation used in KDE", header = TRUE)
a <- table.element(a, par5)
a <- table.row.end(a)
a <- table.row.start(a)
a <- table.element(a, "correlation(x, y)", header = TRUE)
a <- table.element(a, cor(x, y))
a <- table.row.end(a)
a <- table.end(a)
```

Output is inserted cell by cell...
<table>
<thead>
<tr>
<th>Specification of Output Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allow users to specify the type of output (html, Word, Excel, ...)</td>
</tr>
<tr>
<td>Name of page</td>
</tr>
<tr>
<td>Main Title</td>
</tr>
<tr>
<td>Caption</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Keywords</td>
</tr>
<tr>
<td>Menu Label</td>
</tr>
<tr>
<td>Menu Description</td>
</tr>
<tr>
<td>R version</td>
</tr>
<tr>
<td>--------------------</td>
</tr>
<tr>
<td>Category</td>
</tr>
<tr>
<td>Command</td>
</tr>
<tr>
<td>Send output to</td>
</tr>
<tr>
<td>Author</td>
</tr>
<tr>
<td>Submitted by*</td>
</tr>
<tr>
<td>E-mail*</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
The R Framework creates the R module when the Editor clicks the Publish link.
Descriptive Statistics

:: Free Statistics and Forecasting Software ::

:: Free Statistics Software (Calculator) - Web-enabled scientific services & applications ::

All rights reserved. The non-commercial (academic) use of this software is free of charge. The only thing that is asked in return is to cite this software when results are used in publications.

Here you find a collection of Descriptive Statistics Software modules (Calculators). The modules have been grouped in Univariate, Bivariate, and Trivariate categories. All modules can be used with any dataset that contains ungrouped observations.

Main Menu

Plot & describe series generates a simple plot of the data series and allows one to enter a description (useful for future reference when the computation is submitted to the archive).

Central Tendency

- arithmetic mean, geometric mean, harmonic mean, median, midrange, midmean, robustness of central tendency (winsorized and trimmed mean), etc...

Variability

- range, variance, standard deviation, variation, MSE, absolute deviation, interquartile difference, coefficient of quartile variation, Gini’s mean difference, Leik’s D, dispersion, diversity, qualitative variation, mean square deviation, etc...

Concentration

- entropy, exponential index, Herfindahl, variation coefficient, Gini coefficient, etc...

Moments

- general, non-centered & centered moments, trimmed moments

Skewness/Kurtosis

- Fisher 3rd centered moment, Fisher beta 1 & gamma 1, Pearson, Yule’s skew (according to 8 different quartile definitions), Beta, Gamma, small sample skewness moment, Fisher beta 2, 3, gamma 2, small sample kurtosis, etc...

Univariate Descriptive Statistics - Ungrouped Data

Bivariate Descriptive Statistics - Ungrouped Data

Correlation

- Pearson correlation, covariance, determination coefficient, scatter plot, etc...

Rank Correlation

- Spearman Rank Order Correlation (corrected and non-corrected).

Simple Regression

- general linear model, mean and variances, covariance, correlation, least squares estimation, parameters, response, significance, determination coefficient, ANOVA, residuals, autocorrelation, model selection, model performance, etc...

Bivariate Density

- computes Bivariate Kernel Density Estimates

Kendall Rank Correlation

- computes the Kendall tau Rank Correlation between two data series

Box-Cox Linearity Plot

- computes the Box-Cox Linearity Plot

Linear Regression

- computes the Simple Linear Regression model \( Y = a + bX \) and various diagnostic tools from the perspective of Explorative Data Analysis

Graphical Model Validation

- computes the Back to Back Histogram (sometimes called BiHistogram) for a bivariate dataset

Back to Back Histogram
Bivariate KDE (R Module)
Handling Requests

- Find R module (Google)
- User submits request (html form, HTTP POST)
- Webserver loads R module
- R module creates pre-processed R code
- Webserver directs request to R server (callback)
- R server invokes R engine, stores output
- Webserver gets output parses through template
- Webserver sends html reply to user
- User fetches pictures/logfiles from R server
Computational Result

Summary of computational transaction
Raw Input  
Raw Output  
Computing time 3 seconds
R Server  

Pearson Product Moment Correlation - Ungrouped Data

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Variable X</th>
<th>Variable Y</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean</strong></td>
<td>86.893442529508</td>
<td>98.111475409835</td>
</tr>
<tr>
<td><strong>Biased Variance</strong></td>
<td>109.891760279495</td>
<td>2.39249125578877</td>
</tr>
<tr>
<td><strong>Biased Standard Deviation</strong></td>
<td>10.4829270854802</td>
<td>1.54676800645370</td>
</tr>
<tr>
<td><strong>Covariance</strong></td>
<td>4.73174316939891</td>
<td></td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td>0.287034985095086</td>
<td></td>
</tr>
<tr>
<td><strong>Determination</strong></td>
<td>0.0823890826685364</td>
<td></td>
</tr>
<tr>
<td><strong>T-Test</strong></td>
<td>2.30160903402757</td>
<td></td>
</tr>
<tr>
<td>p-value (2 sided)</td>
<td>0.0249057450545596</td>
<td></td>
</tr>
<tr>
<td>p-value (1 sided)</td>
<td>0.0124528725272798</td>
<td></td>
</tr>
<tr>
<td>Degrees of Freedom</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Number of Observations</td>
<td>61</td>
<td></td>
</tr>
</tbody>
</table>

**Pearson Product Moment Correlation**

\[
r_{xy} = \frac{C(xy)}{\sqrt{V(x) \times V(y)}} = \frac{C(xy)}{s_x \times s_y}
\]

\[-1 \leq r_{xy} \leq +1\]

\[
C(xy) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

\[-\infty \leq C(xy) \leq +\infty\]

\[
V(x) = s_x^2 = C(xx) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

\[
V(y) = s_y^2 = C(yy) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]
Pictures: png and postscript

Png formaat

Postscript (herschaalbaar, exporteerbaar)
R Modules …

• are “indexable”, hence “findable”
  – hierarchy of hyperlinks
  – meta tags, titles, descriptions
  – archive of old versions
  – fast loading, pure html
• make “Business/Marketing Sense”
• are “flexible” (the underlying R code can be changed on the fly)
• have “impact”
R modules have impact

- Millions of unique users [Borghers, Wessa, 2005]

Advantages

- Distributed computing
  - Scalability
  - Robustness
  - Shared resources
- Thin client
  - No incompatibilities (unlike DIE software)
  - Client-side security
  - Easy/cheap maintenance
- Flexibility
- etc...
Reproducible Research (freestatistics.org)

Archive of Computational R Objects that support reproducibility and reusability
Compendium

• Original definition:
  An electronic collection of Data and Software that is needed to reproduce the results in a text (c.q. Article)

• New definition:
  A document with (open-access) references to archived Computations (including Data, Metadata, and Software) that allow us to reproduce, and reuse the underlying analysis

=> the compendium platform is a tool for collaboration, dissemination, and monitoring.
Blogging Results

What's next?

- Command: Description
  - Store: Store your latest output on our servers (free of charge). You can always retrieve your results, and share them with others. At any time in future you will be able to reproduce and reuse stored computations.
  - Send: Coming soon... Send a copy of this output to your e-mail address.
  - Sensitivity: Coming soon... Analyze the sensitivity of your result with respect to one (or several) parameter(s).

Last computation | Module | Delete history
--- | --- | ---
Mon, 22 Oct 2007 02:25:45 -0700 | Pearson Correlation | Command
Mon, 22 Oct 2007 02:24:42 -0700 | Start of session | -

Submit your Statistical Computation to the FreeStatistics.org Archive

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td></td>
</tr>
<tr>
<td>Keywords</td>
<td></td>
</tr>
<tr>
<td>Your Comments</td>
<td></td>
</tr>
<tr>
<td>E-mail</td>
<td></td>
</tr>
<tr>
<td>Captcha</td>
<td></td>
</tr>
</tbody>
</table>

Submit to Archive
Archive of Computational R Objects
Object properties

- Date, computing time, links, sources, etc...
- GUID
- R code
- Parameters
- Data
- Comments, keywords, tags, etc...
- R output
- HTML output
- Pictures (png, postscript)
- etc...
Objects on the web
Demo (screenshots)

\item several important stock exchanges of the U.S.A.
\item U.S.A. bonds, notes, and treasury bills
\item gold and silver
\item several well-known stock exchanges in Europe and Asia

\end{itemize}

Figure \ref{Figure 1} shows the histograms about various statistics of the log returns of all observed time series (denoted QRW). It can be observed that many series contain more than 2500 trading days, and only a small minority of series have less than 1000 observations (variable "length"). The descriptive statistics about extreme values (c.q. range, minimum, maximum, and interquartile range) have highly skewed distributions. In addition, the variation about these statistics is substantial, indicating that the sample of index series exhibits a variety in terms of extreme returns.

\begin{figure}[h]
\centering
\includegraphics[width=8cm]{Rplot.eps}
\caption{Descriptive Statistics - Dataset}
\end{figure}

For every time series I simulated 20 Random-Walks (denoted RW) that are - by definition - known to satisfy the criteria of weak form-efficiency. Each of the 20 simulated series has the same mean, and standard deviation as the original time series. Figure \ref{Figure 1} shows that the variation of the deviation of extremes (minimum or maximum) between the simulated and original time series converges as the extreme is closer to zero. The interquartile
predicted, this will lead to the implementation of profitable
investment strategies with a time horizon that depends on
the dynamics of the market. It is also for this reason that
the fatter tails in the distribution of log returns should be more
pronounced in long time series, implying that the market is
more inefficient on the long run. This conclusion is not in-
consistent with the vast literature about short-term market
inefficiencies because I define market inefficiency in terms
of kurtosis only. The probability that the kurtosis of log
returns is significantly different from zero, increases as the
time series under investigation gets longer. In other words,
when we look at a longer price history we have a higher
probability to observe states $h = 1$ which can be related
through a logistic regression to the kurtosis p-value of log
returns.

If this model turns out to be effectively discriminating
between the states of the market ($h = 1$ and $h = 0$) then
it is possible to create fast algorithms that make a preselec-
tion of equities (from the universum of all equities under
consideration) that show a high logistic regression probability
that $h = 1$. This is of particular importance for ad-
vanced investors, such as hedge funds, employing invest-
ment strategies that involve simultaneous long and short
positions in different portfolios of equity selected from a
prospected universum. Furthermore, any model that se-
lects equity from the universum and assigns them to either
a long or short position portfolio, must have a statistical
discrimination quality that is at least as good as the quality
of the proposed model. In other words, the power of the
proposed logistic regression is a benchmark for any equity
selection algorithm when feeded with simulated and true
stock market time series.

3 Dataset

I collected 66 index time series about various important

The descriptive statistics about extreme values (e.g., range,
minimum, maximum, and interquartile range) have highly
skewed distributions. In addition, the variation about these
statistics is substantial, indicating that the sample of index
series exhibits a variety in terms of extreme returns.

| Figure 1. Descriptive Statistics - Dataset |

For every time series I simulated 20 Random Walks
denoted RW) that are - by definition - known to satisfy
the criteria of weak form-efficiency. Each of the 20 simu-
lated series has the same mean, and standard deviation as
the original time series. Figure 2 shows that the variation of
the deviation of extremes (minimum or maximum) between the
## Demo (screenshots)

**Blog & Share Statistical Computations at FreeStatistics.org - Konqueror**

- **Raw Input**: view raw input (R code)
- **Raw Output**: view raw output of R engine
- **Computing time**: 35 seconds
- **R Server**: 'Jan Tinbergen' @ 127.0.0.1/wessadotnet/public_html

### Charts produced by software:

- **Frequency**
  - Length of QFW
  - Median of QFW
  - Range of QFW
- **Frequency**
  - Min of QFW
  - Max of QFW
  - IQR of QFW
- **Frequency**
  - Min of RW
  - Max of RW
  - IQR of RW
Every archived computation can be recomputed, changed, and re-archived!

=> Archived Computation
= Collaborative Software
Examples of Compendia

http://www.wessa.net/download/tutorial.pdf
(Descriptive Statistics – Central Tendency)
http://www.wessa.net/download/tutorial1.pdf
(Time Series Analysis - Introduction)

Note: both documents are “work in progress”
Please, send corrections & suggestions to patrick@wessa.net
Queue Manager

(not freely available)