Statistics, Python and R for Data Science

E. Lee Rainwater[†‡]

5 September 2020

Mays Business School
Texas A&M University
College Station TX

[†]Grand Nagus of Mechanical Engineering, Raven Industries, Sioux Falls, SD, USA, Sol III (Earth), United Federation of Planets, Alpha Quadrant
[‡]Mays Business School, Texas A&M University, College Station, TX, USA, Sol III (Earth), United Federation of Planets, Alpha Quadrant

Abstract

Herein may be found the mystic runes that were carved in the living rock of the Cavern of Caerbannog by Olfin Bedwëre of Rheged[1], which therein make known the dark knowledge of statistics, Python, and R as pertinent to data science and the prevention of earthquakes.

1 Introduction

In preparation for an assesment of skills stemming from an inquiry into a position of employment with McKinsey & Company™, I prepared this summary as a means to review pertinent skills. I eventually withdrew myself from consideration for this position, as I did not feel it to be a good match; however, this exercise provides continual benefit in the preparation that it provides for may other opportunities.

This writing was initially devised for my own personal use; therefore, it may be found to contain much that is superfluous, irrelevant, irreverent, pompous, and deserving of many other adjectives; the enumeration of which would be of such voluminousity as to preclude them from practical inclusion in this manuscript.

2 Statistics

2.1 The Basics

In preparation to matriculate in the Master of Science in Analytics at Texas A&M University, I completed a senior-level class in Statistics at Texas A&M University—Commerce[2]. Basic concepts are summarized below.

1. *mean* - generally refers to the arithmetic mean, or colloqually, the *average*.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x \tag{1}$$

   Note that $\bar{x}$ refers to the sample mean. For the population mean, μ is substituted for $\bar{x}$, and the population size, N is substituted for the sample size, n.

2. *dispersion* - the spread or variability of the data about its central tendency. The common measures of dispersion are:

[1] One of the former kingdoms of *Hen Ogledd* ("Old North"), in the region that is now southern Scotland and northern England

[2] MATH 453 – I somehow contrived an "A" in this class despite taking it over a 5 1/2-week summer term

(a) *range* - the difference between the *greatest* and *least* observed values

(b) *interfractile range* - a location in a frequency distribution for which a given fraction or proportion of the data lie at or below it

(c) *interquartile range* - a range, measured above and below the *median*, which contains one-half of the observed data. Where Q_1 corresponds to the first (lowest) quartile, Q_2 corresponds to the second quartile (the median), and Q_3 corresponds to the third quartile, the interquartile range is given by:

$$\text{interquartile range} = Q_3 - Q_1$$

(d) *average absolute deviation* - the absolute deviation of each datum from the average is summed and divided by the sample size, n, (or population size, N):

$$\text{AAD}_{sample} = \frac{1}{n}\sum_{i=1}^{n}|x - \bar{x}| \text{ for a sample, and,} \tag{2}$$

$$\text{AAD}_{population} = \frac{1}{N}\sum_{i=1}^{N}|x - \mu| \text{ for the population} \tag{3}$$

(e) *population variance* - similar to the *average absolute deviation*, except that the square of the deviation of each datum from average is used instead of its absolute deviation:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x - \mu)^2 \tag{4}$$

Note that the units of *variance* are the units of the data *squared*. For this reason, it is more intuitive to define the *standard deviation*:

(f) *population standard deviation* - defined as the *square root* of the *population variance*:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x - \mu)^2} \tag{5}$$

For those of the Math Police, only the positive square root is considered in determining variance.

(g) *sample variance* and *sample standard deviation* - analogous to the above, except that s^2 represents the variance, and N is replaced with $n-1$ to account for the reduction in *degrees of freedom.*

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x-\bar{x})^2} \tag{6}$$

Someday, I intend to compose an *ELIF* treatise on why N is replaced with $n-1$, but *today is not that day.* Suffice to say that it was a mathematician who determined to do such, as statisticians are not generally so *anal-retentive.*

It should be noted that the *variance* and *standard deviation* are *absolute* measures of dispersion. When comparing populations with significantly different means, it is often useful to divide the standard deviation by the mean, which yields the:

(h) *coefficient of variation* - standard deviation, normalized by the arithmetic mean:

$$c_v = \frac{\sigma}{\mu} \tag{7}$$

2.2 Probability

I enjoy studying probability in about the same manner that I enjoy a good intestinal virus. However, understanding of the basic principles of probability is foundational to further understanding of statistics, so I'll get my Pepto-Bismol ready and delve into it.

2.2.1 Basic Concepts of Probability

Here are a few basic definitions:

- *probability* - a quantification of the chance that an event will occur
- *event* - the occurence of a possible outcome in an experiment
- *experiment* - in *probability theory*, an activity in which one or more *events* occur and are recorded

- *mutually exclusive* - in a *probability theory experiment*, an *event* is said to be *mutually exclusive* if only one event can take place at one time.

  In the example of a single coin toss, the event of the coin landing *heads* and the event of the coin landing *tails* are said to be mutually exclusive, because the coin can only occupy one of those states at a time—the coin can not land both *heads* and *tails* simultaneously.

2.2.2 Classical Probability

Classical probability includes the works of Wolfgang Amadaeus Mozart, Ludwig van Beethoven, Joseph Hayden...er, sorry, wrong subject. Classical probability defines the probability of an event's occurence as simply:

$$P = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}} \tag{8}$$

This definition assumes that each of the possible outcomes is equally likely. Classical probability is sometimes termed *a priori* probability, because the probabilities can be stated in advance without conducting the actual probability experiment.

Classical probability is a simplification (sometimes grossly so) of real-world events. For example, classical probability assumes that a coin toss will either land heads or tails, and ignores the infinitesimally small, but extant probability that the coin will land on its edge, or that some inertial property of the coin will cause it to favor one side slightly over the other.

2.2.3 Relative Frequency of Occurence

The *relative frequency of occurrence* (RFOC) approach defines probability in one of two ways:

1. the observed relative frequency of an event occurring in a large number of trials

2. the proportion of times an event occurs in stable conditions

Actual data of past occurrences are used to determine probabilities, as opposed to the use of theoretical values per *classical probability*. One characteristic of probabilities determined by *RFOC* methods is that the proportion of an event's occurrence stabilizes as the number of observations increases.

As an example, if one were to flip a fair coin four times, they might obtain three occurrences of heads and one of tails. The *calculated* probability of heads would thus be $P(H) = 0.75$, although we know that the *theoretical* probability is 0.5. However, if one were to increase the number of coin tosses to 1,000, the calculated probability would be expected to converge nearer to the expected value of 0.5.

2.2.4 Subjective Probabilities

Subjective probabilities are, as the name implies, based upon the presumably educated judgement of the person assigning the probability. Suppose that your Texas Class AA high school football team from East Fork Consolidated won nine of their last ten games against opponents from similarly-sized schools, then were scheduled to play the next game against Texas A&M. A relative frequency of occurence approach would predict that that they would have a 90% chance of outscoring the Aggies. However, your subjective knowledge of the relative strengths of the Texas Aggies and the East Fork Consolidated Horny Toads might lead you to predict that the Toads would have closer to a zero percent chance of winning.

2.3 Probability Rules

We will succinctly state some of the basic rules of probability:

2.3.1 Addition Rule for Mutually Exclusive Events

Suppose that we are interested in the probability that **A** *or* **B** will occur, such that **A** and **B** are mutually exclusive—in other words, they cannot *both* happen in the same event. In such a case, the probability is additive such that:

$$P(\mathbf{A}\, or\, \mathbf{B}) = P(\mathbf{A}) + P(\mathbf{B}) \tag{9}$$

2.3.2 Addition Rule for Non-mutually-exclusive Events

If the two events are not mutually exclusive, then both events *can* occur simultaneously. An example would be in determining the probability of drawing a six *or* a club *(but not both)* from a standard deck of playing cards. In the deck, there are four cards which are *sixes*, and thirteen cards which are *clubs*. Exactly one of the cards is both a *six* and a *club*.

To calculate the probability of drawing either a *six* or a *club* but not a six of clubs, we must calculate as per the previous addition rule, but subtract out the probability that the card is *both* a six and a club:

$$\begin{aligned} P(6 \text{ or } \clubsuit) &= P(6) + P(\clubsuit) - P(6 \text{ and } \clubsuit) \\ &= \frac{1}{4} + \frac{1}{13} - \frac{1}{4 \times 13} \\ &= \frac{9}{26} \end{aligned} \tag{10}$$

2.3.3 Probabilities under Statistical Independence

As stated before, *statistically independent* events are those for which the occurrence of one event has no effect upon the probability of occurrence of any other event. Under the assumption of statistical independence, there are three types of probabilities:

1. *marginal* – the *simple* probability of the occurrence of an event; such as that of a coin toss—each toss stands alone, and is unconnected with any other toss.

2. *joint* – the probability of two or more *independent* events occurring together or in sucession; *e.g.*, the probability that of two coin tosses, toss **A** and toss **B** are heads. The probability, $P(\mathbf{AB}) = P(\mathbf{A}) \times P(\mathbf{B})$.

3. *conditional* – The probability that a second event, **B** will occur given that a prior event, **A** has already occurred. This probability is expressed symbolically as $P(\mathbf{A}|\mathbf{B})$.

   This is a bit of a trick scenario—recall that the two events are *statistically independent*. Thus, the probability $P(\mathbf{A}|\mathbf{B})$ is the same as $P(\mathbf{B})$. We will later consider probabilities under conditions of *statistical dependence*, in which the occurrence of **A** *does* influence the probability of the occurence of **B**. I'm not sure why we even discuss conditional probability of independent events, except to make this point.

2.3.4 Probabilities under Statistical Dependence

To be succinct, **statistical dependence exists if the probability of some event is dependent upon, or affected by, the occurrence of some**

other event. Like in the case if *independent events*, there are three types of probabilities under statistical dependence:

1. *conditional* – eh, we'll get to that later...

2.4 Probability Distributions

2.4.1 The Binomial Distribution

The *binomial distribution* describes a distribution of discrete data, such as a coin toss. A Bernoulli process is defined as follows:

1. Each trial has only *two* possible outcomes (yes/no).
2. The probability of the outcome of any trial remains *fixted* over time.
3. The trials are *statistically independent*; that is to say that the outcome of one trial does not affect the outcome of any other trial.

The *binomial formula* is used to calculate the probability of achieving r successes in n trials:

$$P = \frac{n!}{r!(n-r)!} p^r q^{n-r} \tag{11}$$

where:

$$\begin{aligned} p &= \text{probability of success} \\ q &= 1 - p = \text{probability of failure} \\ r &= \text{number of successes desired} \\ n &= \text{number of trials undertaken} \end{aligned}$$

With a binomial distribution, it is possible to calculate the mean and standard distribution as follows:

$$\mu = n\,p \tag{12}$$

$$\sigma = \sqrt{n\,p\,q} \tag{13}$$

2.4.2 The Poisson Distribution

The *Poisson distribution* is a discrete distribution that may be used to model the probability of a discrete number of occurrences within a time period. For example, given that an event which follows a Poisson distribution occurs on average five times per hour, the Poisson distribution formula could be used to calculate the probability of that event occurring seven times in an hour.

The probability of exactly x occurrences according to a Poisson distribution is given by:

$$P(x) = \frac{\lambda^x \, e^{-x}}{x!} \tag{14}$$

where:

λ = mean number of occurrences per time interval
x = the number of occurrences for which the probability is calculated

When $n \geq 20$ and $p \leq .05$, the Poisson distribution provides a reasonable approximation of the binomial distribution.

2.4.3 The Normal (*Gaussian*) Distribution

The *normal distribution* is a *continuous* probability function that is the basis for most statistical analysis of continuous data. A few characteristics of the normal distribution are:

1. Regardless of the values of μ and σ, the total area under the normal curve equals 1.
2. $\approx$ 68% of all values in a normally distributed population lie within $\mu \pm 1\sigma$.
3. $\approx$ 95.5% of all values in a normally distributed population lie within $\mu \pm 2\sigma$.
4. $\approx$ 99.7% of all values in a normally distributed population lie within $\mu \pm 3\sigma$.

The probability density function of a normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{15}$$

The number of standard deviations from x to the mean of the distribution, μ, is given by:

$$z = \frac{x - \mu}{\sigma} \tag{16}$$

The value z is expressed as *standard units* and is often used to look up values in probability and cumulative distribution tables.

2.5 Random Sampling

For the purpose of statistical sampling, a *infinite* population refers to a population which cannot be enumerated in a reasonable period of time; thus it is *practically* infinite.

2.5.1 Simple Random Sampling

Simple Random Sampling is that in which samples are selected in such a way that:

- Each possible sample has an equal probability of being picked
- Each item in the population has an equal probability of being included in the sample

2.5.2 Systematic Sampling

Systematic random sampling selects elements from the population at a uniform interval as measured in terms of *time*, *order*, or *space*. An example would be to survey every 20th visitor to a website. Systematic sampling may be inappropriate when the measured value is in some way dependent upon the interval.

2.5.3 Stratified Sampling

Stratified sampling divides the population into relatively homogeneous groups, called *strata*. As an example, suppose that it is desired to randomly sample students at a particular university in which only 20% of the student body is female. The researcher desires to sample equal proportions of both male and female students. In such a case, the researcher would divide the roster of students into seperate lists of males and females, and randomly sample the same percentage of each.

2.6 Regression Analysis

2.6.1 Linear Regression

Assumptions of Linear Regression:

1. The mean of the response, $E(Y_i)$, at each value of the predictor, x_i, is a linear function of the x_i.
2. The errors, ϵ_i, are independent.
3. The errors, ϵ_i, at each value of the predictor, x_i, are normally distributed.
4. The errors, ϵ_i, at each value of the predictor, x_i, have equal variance (σ^2).

3 Time Series Analysis

3.1 Introduction to Time Series

A *univariate time series* is a sequence of measurements of one variable observed with respect to time. Typically, the observations (measurements) are made at regular intervals. One difference between a time series and a linear regression is that the data in a time series are not necessarily independent from one another—an observation in a time series may in fact be dependent upon the value at a previous point in time. Furthermore, the *ordering* of the values are important—to re-order the values would significantly change the meaning of the data.

Much of our attention will be focused upon models that relate the present value of the time series to a series of past values and past errors. These are termed *Autoregressive Integrated Moving Average* (**ARIMA**) models.

3.2 Characteristics of a Time Series

When examining a time series, we consider the following:

- **trend**—on average, do the measurements tend to increas or decrease over time?
- **seasonality**—is there a regularly repeating series of highs and lows?

- **variance**—is the variance constant over time, or non-constant?
- **long-run cycles**—is there an extended period unrelated to seasonality factors?
- **outliers**—in time series, outliers are far away from the mean
- **abrupt changes**—are there any sudden changes to either the level of the series or its variance?

A time series, Y_t is *(weakly) stationary* if it has a mean, variance, and a correlation structure that are all constant with respect to time. In other words:

1. $E(Y_t) = \mu$ is constant for all t (where μ represents the population mean). Thus, the *mean* is neither increasing or decreasing with time;
2. $Var(Y_t) = \sigma^2$ for all t. Thus, the *variance* is neither increasing nor decreasing with time; and,
3. $Corr(Y_t, Y_{t-s}) = \rho(s)$ for all s and for some function, p, which is termed the *autocorrelation function*, ACF Thus, our *correlation* is neither increasing nor decreasing with time.

A time series is declared to be *nonstationary* if any of the above three criteria are not true. For the sake of our current study, the word *stationary* shall refer to a time series that is at least weakly stationary. Formally, there is a kind of time series model which is termed *strictly stationary*; it is that for which all joint distributions are invariant to shifts in time. We will defer the study of *strictly stationary* time series until later (if ever).

The correlation structure of a time series is studied by examining:

- Plots of Y_t against Y_{t-s}, where s represents a number of lags in the time series (*e.g.*, $s = 1$ refers to the observation immediately prior to the one of interest; $s = 2$ refers to the 2nd observation prior to the one of interest, *etc.* These plots are called *lag plots*.
- The *autocorrelation function* $\rho(s) = Corr(Y_t, Y_{t-s})$, where $s = 1, 2, \ldots$
- *Probably some other stuff goes here.*

3.2.1 The Random Walk

Mathematically, a *random walk* time series is defined as:

$$Y_t = Y_{t-1} + a_t \tag{17}$$

where:

$$\begin{aligned} Y_t &= \text{the value, } Y \text{ at time } t \\ Y_{t-1} &= \text{the value, } Y \text{ at time } t-1 \text{(the point immediately prior to } Y_t) \\ a_t &= \text{the error in the value } Y \text{ at point } t \end{aligned}$$

Since a_t is simply a *white noise* value with a mean of zero, a constant variance, and no correlation to any other values of a, it can be said that **a data point in a random walk time series is nothing more than the value of the previous point with white noise added.** Why is important that we study the *random walk time series*? I'll revise this section and let you know, once I have figured that out for myself.

3.2.2 The AR(1) Model

A very simple ARIMA model is one in which a linear model is used to predict the value at a specific time using the value at the observation immediately preceding it. This is called an **AR(1)** model, which denotes an *autoregressive model of order* **1**. The *order* of the model refers to the *lag*, which is number of time offsets between the predicted and the predictor data points. Since, in this particular example, the value of each point is modeled from the value of the point immediately preceding it, it is said to have a lag of 1.

A typical time series plot is presented as Figure 1. From a brief examination, several characteristics of this time series are evident:

- There are no obvious outliers in the plot.
- There is no trend that is consistent over the entire time span. While there are some localized trends which soon reverse themselves, there is no obvious overall trend.
- It is difficult to visually determine whether the variance is constant.
- There is no obvious seasonality. (In this case, there is only annual data, so seasonality is not to be expected.)

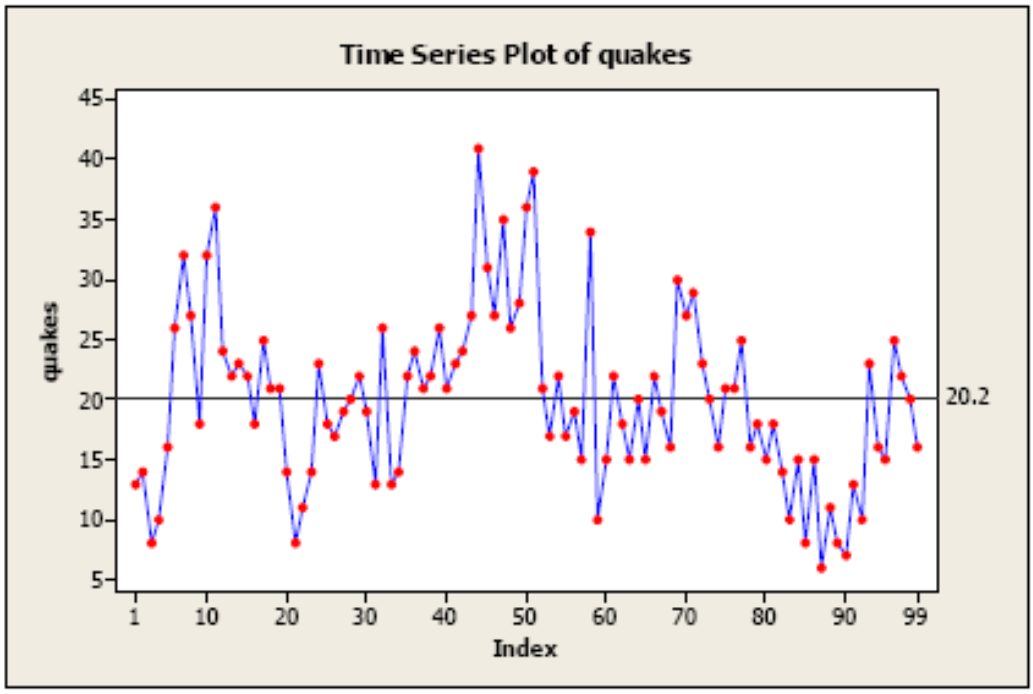

Figure 1: A typical time series plot of earthquakes. No sheep bladders were utilised to prevent them[a]. The horizontal axis represents *years* and the vertical axis represents the *number of earthquakes* in each year.

[a]Bedevere, Sir, *J. of Medieval Physics*, AD 932

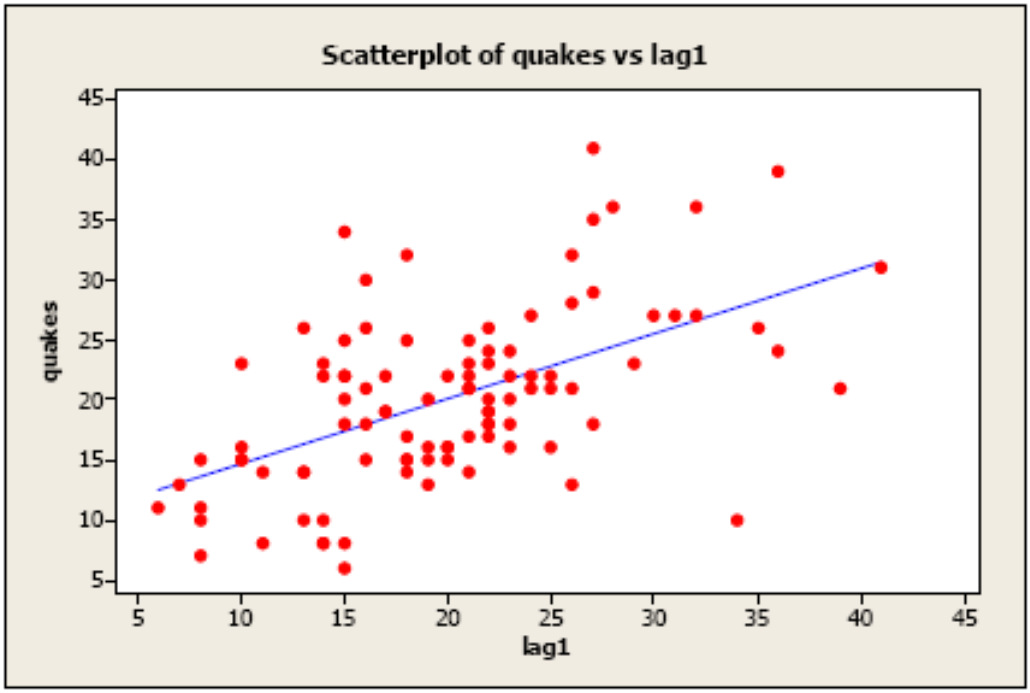

Figure 2: A *Lag 1* plot of the time series depicted in Figure 1. The horizontal axis represents the *number of earthquakes* for the x_{t-1} time point; the vertical axis represents the *number of earthquakes* for the x_t time point.

For a typical interpretation, consider $x_{t-1} = 10$ (on the axis labeled *lag1*. We can see that there were three instances where, in a year $(t-1)$ in which there were 10 earthquakes reported, there were also quakes in the following year t. In those three following years, there were 15, 16, and 23 earthquakes reported (as estimated by eyeballing the data).

We can see that there is a moderate linear relationshp between the number of earthquakes in the year $t-1$ and the year t. Thus, we suspect that this is a *first order autoregressive model*.

To assess whether an AR(1) model is appropriate, we may generate a *lag plot* as shown in Figure 2. Here, the values of the series, denoted x_t are plotted against the values of x_{t-2}. Thus, for each point in the plot, the y-axis represents the value of x_t at some time t, and the x-axis represents the value of x_{t-1}. In other words, it is a plot of the value of each point against the value of its immediate predecessor. Similarly, we could produce a lag plot of *lag 2*, or any arbitrary number of lags (limited only by the number of data points available).

Looking at Figure 2, we can see that there is a moderately strong linear relationship between a point and its lag. Thus, an AR(1) model may be useful.

An AR(1) model is defined mathematically as:

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + a_t \tag{18}$$

where:

$$\begin{aligned} \mu &= \text{the mean of } Y \\ \phi_1 &= \text{the regression coefficient for the } 1^{st} \text{ lag} \\ a_t &= \text{the error in the value } Y \text{ at point } t \end{aligned}$$

It is similar to the equation for a *random walk* time series, except that there is now a coefficient, ϕ_1, which partially determines the value of Y_t from Y_{t-1}.

Simply stated, *the deviation of Y_t from the mean, μ is determined by the deviation of Y_{t-1} from μ, multiplied by the coefficient ϕ_1, plus some* white noise. The white noise term, a_t, as was stated earlier, has the following characteristics:

- $\mu_a = 0$ (zero mean)
- ${\sigma_a}^2 = C$ (constant variance)
- its values have no correlation over time

Note that there may be more than one lag term. If it can be shown that there exists a relationship between the value, Y_t and some other lag such as Y_{t-n}, then Equation 18 is modified to include the additional lag terms. We'll address that later.

Frequency of Uncorrelated Kinetic Uncertainty Principles—**FUKUP**

4 Appendix

4.1 Code Listing

Below is the SAS code used to generate the tabulation of wine points and cost.

```
1 ODS RTF FILE="C:\Users\rainwater-e\...\hw-08\
     CABERNET.RTF";
2 PROC TABULATE DATA=&EM_IMPORT_DATA;
3 CLASS TextCluster_cluster_;
4 VAR POINTS PRICE;
5 TABLE textCluster_cluster_, (POINTS PRICE)*MEAN;
6
7 RUN;
8 ODS RTF CLOSE;
```

4.2 SAS Enterprise Miner Property Panes