Time Series Analytics Notes
Random Samples

Recap from the Introduction:
There are two primary reasons for studying time series:

- To forecast the future
- To explain the present

A **forecast** is simply a numerical guess about a future value, together with a numerical assessment of the uncertainty of the guess.

Both forecasting and explanation depend upon having a valid statistical model for the time series. A **statistical model** is a set of **specifications**, or hypotheses, about the way that the time series behaves. The specifications can – and should! – be tested for validity before using the model to make forecasts or to offer explanations.

Statistical modeling is a three-step process:

1) Propose the model
2) Validate the model
3) Use the model

When you propose a model, you nominate one of the many time series models (e.g., random sample, random walk, autoregression, etc. – or create your own) as a candidate for the data at hand. When you validate your proposal, you test the model's specifications against the data. In step 3, you make the forecast or provide the explanation *according to the rules of your validated model*.

Random Sample Time Series

In this Topic Note, I will introduce and discuss the fundamental time series and forecasting model that underlies all other models: the Random Sample. It is very important to master the Random Sample time series model thoroughly. There are three reasons for this:

- Some real-world times series actually are Random Samples
- Many other time series models have Random Samples hidden inside them
- The rules for forecasting/explaining of time series depend upon how the Random Sample is hidden inside.

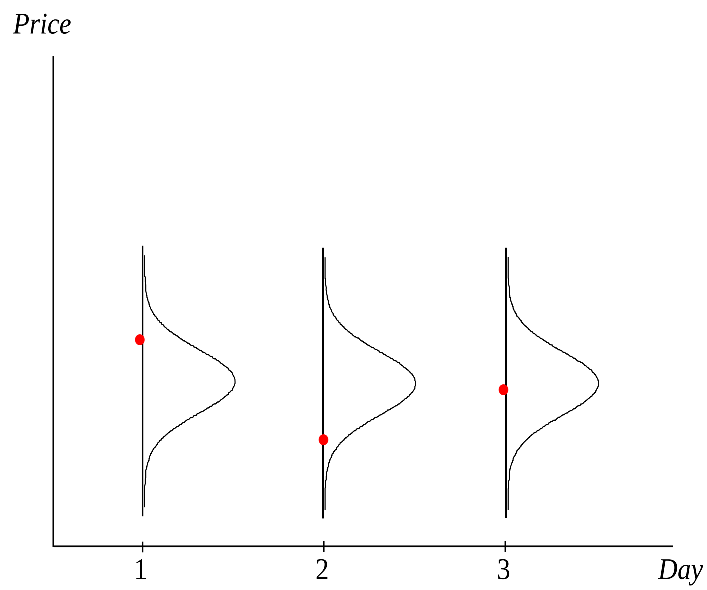
Validation of other time series usually means ferreting out how the Random Sample is hidden inside.

<u>The Model</u>: A **Random Sample** (**RS**) time series $Y_1, Y_2, \ldots, Y_t, \ldots$ satisfies two specifications:

- $Y_1, Y_2, \ldots, Y_t, \ldots$ are independent
- $Y_1, Y_2, \ldots, Y_t, \ldots$ are identically distributed

RS is a synonym for **independent and identically distributed** (**i.i.d.**) Figure 1 is a schematic that illustrates the model for a RS. At each day, there is an uncertainty distribution (population) of possible values that could be observed. The distributions are shown as normal curves.[1] At each day a draw is made at random from the distribution – the normal curve – for that day. The values that are actually seen are the red dots. The distributions are not seen – only the red dots are known.

Figure 1. A schematic to illustrate the specifications for a hypothetical Random Sample



The **independent** part of i.i.d. means that the draw on day 2 does not depend upon what happened on day 1; the draw on day 3 does not depend upon what happened on days 1 or 2; etc. In general, what happens on any day does not depend upon what happens on *any* other day or days.

The **identically distributed** part of i.i.d. means that the distributions are the *same* for every day. Each day's curve is a carbon copy of every other day's curve. The curves do not move up or down over time, and they do not widen or narrow, nor change shape or position in any way.

More broadly, in statistics, a random sample is a sequence $Y_1, Y_2, ..., Y_n, ...$ of independent and identically distributed random variables, not necessarily sequenced in time. The Random Sample Time Series is a random sample in the broader sense – only sequenced by time. The assumption that random variables are independent and identically distributed plays a big role in classical statistical theory.

[1] The distributions need not be normal. Normal curves are shown only to make the illustration specific.

Let us now turn to the three-step modeling process that I outlined: propose, validate, use.

1. Propose: How would you know to propose the RS as a model for some data?

- Graphics: You look at a plot of the data, and the plot looks like a RS should. (What should a RS look like? Figure 2 is a typical example.)
- Theory: You may have or know a theory that suggests that the current data *should* be a RS or close to it.
- Real experience: You have had experience with similar data in the past, and the RS model has worked well for them.
- Simulated experience: You can write a simulation program that will produce data that follow a true RS model. You know that it is a true RS because you wrote the program! You can run the simulation program a number of times, examine the output of each run, and develop synthetic experience so that you can recognize key similarities with your current data.
- Trial-and-Error: You haven't a clue, and are just trying models to find something that works.

What are the key features of a RS? How does a RS manifest itself in data?

Figure 2. An example of data generated by a true RS

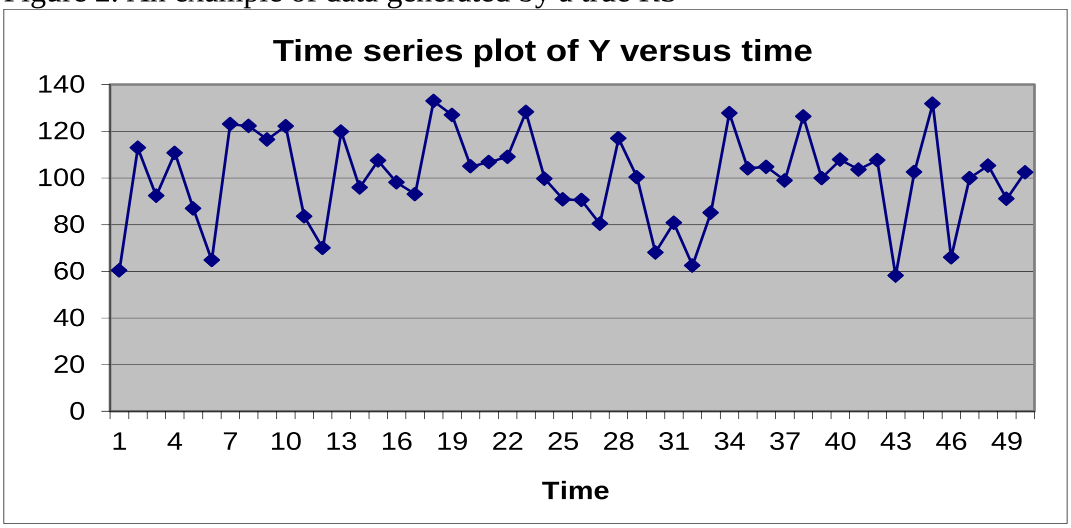


Figure 2 shows one run of a simulation of a RS in Excel. There are 50 consecutive values.
To address the question of how a RS manifests itself, let us look to the specifications. Let us look at *identically distributed* first. If the RS model is correct, then each of the 50 values is drawn from the same distribution. If the distributions are the same for each of the 50 time periods, then the means and standard deviations must be the same. If the means are all the same, we should expect the mean value of the data to be about the same, early in the period, as in the middle of the period, as late in the period. That is, we should expect the data points to cluster around the true mean of the common distribution. The data points should generally lie at about the same level across their time plot. If the standard deviations are all the same, we should expect the up-and-down variation of the data to be about the same magnitude, early in the period, as in the middle of the period, as late in the period. We would not expect to see widening or shrinking

variation across time. That is, *identical distributions* should manifest as a timeplot that looks like a uniform horizontal strip from left to right. Not exactly – but approximately.

Visually, it appears that the timeplot in Figure 2 conforms to these expectations. The data values tend to average about 100 across the timeplot, and the variation mostly lies in a fairly uniform channel from about 70 – 130.

Let us turn now to the *independent* specification of i.i.d. If the data are independent, then they should not be autocorrelated. **Autocorrelation** is a special type of correlation for time series. It measures how strongly a time series is correlated with its own past. If the current datum Y_t does not depend upon the previous datum Y_{t-1}, then $Corr(Y_t, Y_{t-1})$ should be close to zero. $Corr(Y_t, Y_{t-1})$ is called the (**lag 1**) **autocorrelation**.[2] It can be calculated from a set of time series data $Y_1, Y_2, ..., Y_t, ...$ by calculating the ordinary correlation coefficient of the pairs $(Y_1, Y_2), (Y_2, Y_3), (Y_3, Y_4) ..., (Y_{t-1}, Y_t), ...$. If the autocorrelation is large and positive, then the pairs are strongly related to each other: both members of the pairs tend to be high simultaneously or low simultaneously. Large positive autocorrelation often manifests as a "sticky", slowly meandering pattern. Figure 3a provides an example. On the other hand, if the autocorrelation is large and negative, then the pairs strongly repel each other: one member of a pair tends to be high and the other low. Large negative autocorrelation often manifests as a rapidly oscillating pattern. Figure 3b provides an example. If the autocorrelation is close to zero, there will be neither attraction nor repulsion between pair members. Figure 2 provides an example.[3]

Figure 3a. Positive autocorrelation

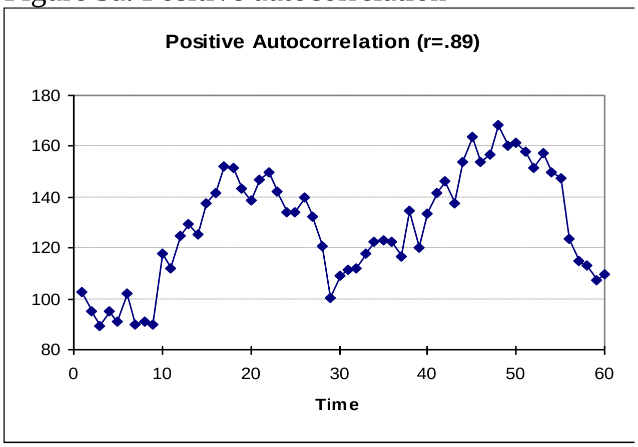
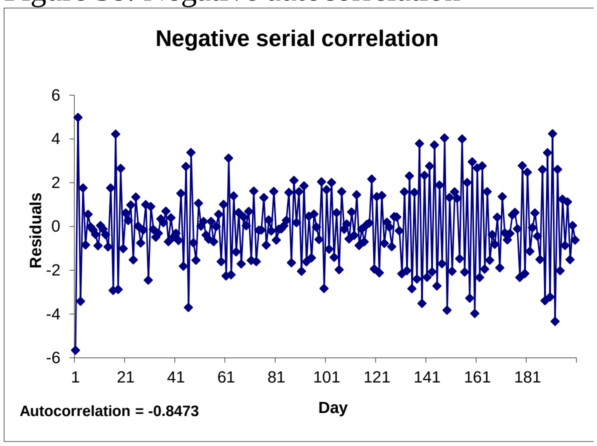


Figure 3b. Negative autocorrelation



[2] Lag 2 autocorrelation is $Corr(Y_t, Y_{t-2})$, etc.

[3] It is actually rather difficult to distinguish zero visually from negative autocorrelation. Positive autocorrelation is easier to detect visually.

2. Validate: How do you confirm that the RS is an appropriate model for some data?

First, I should point out that *validation* does not mean that we *prove* the RS model is correct. We should not really care if a model is "true" in some absolute sense, as long as it is *good enough* for our purposes. Validation tests are actually diagnostics that can establish the plausibility of a model, but not prove it.

There are three diagnostic tests for a RS model. They are based on how a RS should manifest. A RS should have:

- Approximately constant level (**L**)
- Approximately constant variability (**H**)
- Approximately zero autocorrelation (**I**)

Together, these are the **LHI** diagnostics for a RS.[4] LHI are necessary conditions to be an RS; but they are not sufficient. Nevertheless, they provide a practical means to test for RS.

L stands for Level. L refers to the fact that each of the distributions in a RS should have the same mean. This should manifest in a timeplot that is approximately level, on average, across time.

H stands for Homoscedastic – a sesquipedalian term that means "constant variability" (not an oxymoron). H refers to the fact that each of the distributions in a RS should have the same standard deviation. This should manifest in a timeplot that varies up-and-down by approximately the same amount across time.

Together, L and H constitute an approximate test of equal uncertainty distributions (i.e., identically distributed). In theory, equal means and equal standard deviations are necessary consequences of equal distributions, but are not sufficient conditions. Distributions can have the same means and standard deviations but still differ in other respects. Still, the mean and standard deviation are the two most important features of most distributions. So diagnostic tests for L and H may provide a reasonable proxy as tests for identical distributions.

I stands for Independent. "I" refers to the fact that each of the draws from the identical distributions should be independent of all other draws. This manifests in a timeplot that has zero autocorrelation. In theory, zero autocorrelation is a necessary consequence of independence, but is not a sufficient condition. Draws can be uncorrelated without being independent. Still, a diagnostic test of zero autocorrelation may provide a reasonable proxy as a test for independence.

Visual tests.

It is important to understand the intuition behind the LHI tests. Graphs can help with that understanding. Often, graphical diagnosis is all that is really required for validation. However, there are also quantitative tests for each of L, H, I.

L – to test visually for L, make a timeplot of the data. Examine the general trend from early to late. The trend in means should remain roughly level to be consistent with L. Examples of visually satisfactory conformance with L are given by Figures 2 and 3b. Figure 4a is an example

[4] Warning! "LHI" is a Sagerism. It is my acronym. You will not find LHI in statistics books. You will find random sample, i.i.d., homoscedasticity, autocorrelation – but not the LHI acronym.

of a violation of L (the mean level increases over time) that is easily detected by visual inspection of the timeplot.

H – to test visually for H, make a timeplot of the data. Examine the general pattern of up-and-down variation around the overall (flat) level from early to late. The pattern of variation should remain roughly the same to be consistent with H. Examples of visually satisfactory conformance with H are given by Figures 2 and 3b. Figure 4b is an example of a violation of H that is easily detected by visual inspection of the timeplot.[5] (The variability increases over time.)

Figure 4a. Violation of L

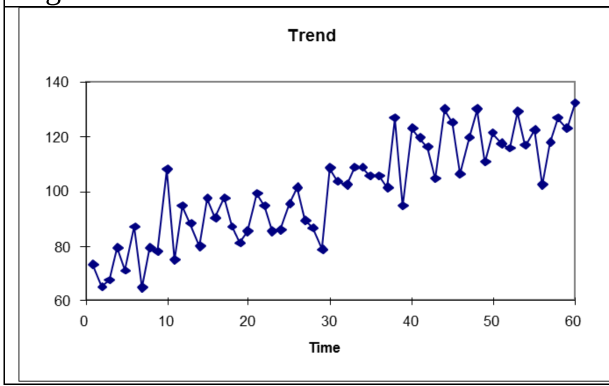
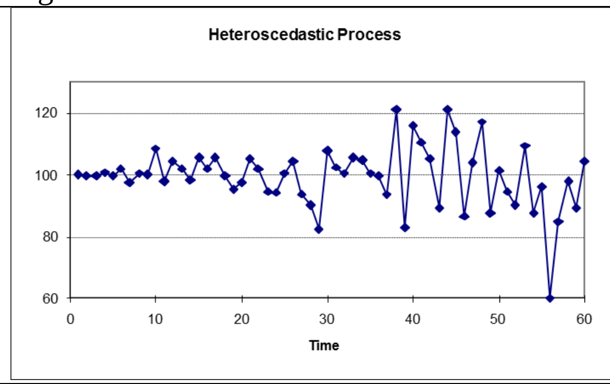


Figure 4b. Violation of H



I – to test visually for I, make a timeplot of the data. Examine the general pattern of the data for “stickiness” or oscillation. Absence of either suggests roughly zero autocorrelation. Well – actually, this is a tough one for the eyes. Although I can often tell positive autocorrelation when I see it in the timeplot (as in Figure 3a), I have great difficulty in reliably distinguishing zero autocorrelation (as in Figure 2) from negative autocorrelation, unless it is very high (as in Figure 3b). I need a quantitative test.

Quantitative test for I.

Begin by calculating the (lag 1) autocorrelation of the time series. This is the ordinary correlation of all pairs $(Y_1,Y_2),(Y_2,Y_3),(Y_3,Y_4)...,(Y_{t-1},Y_t),...$. To satisfy I, the autocorrelation should be close to zero. How close? Even if the time series truly has zero autocorrelation, the actual data – which are random draws from the same distribution time after time – will vary. So how much can we reasonably expect the autocorrelation to vary from zero? Statisticians have studied the behavior of sample autocorrelations mathematically and found that the distribution of sample autocorrelations is approximately normal, with mean 0 (if “I” is correct), and standard deviation about $1/\sqrt{n}$, where n is the number of pairs. So autocorrelations in the range $-2/\sqrt{n}$ to $+2/\sqrt{n}$ are about at the limit that is reasonably consistent with “I”. If “I” is true, there is only about a 5% chance that the sample autocorrelation will lie outside the range $-2/\sqrt{n}$ to $+2/\sqrt{n}$.

[5] Note that Figure 4a also does *not* conform with H. Although the pattern of variation is constant around the *trend* of the data, H requires that the variability be constant around the overall *level* of the data. In Figure 4a, the *level* is flat at about 100, but the variability around 100 is big at the beginning and end of the time series and small in the middle.

So we can test the hypotheses

$$H_0 : \rho_1 = 0$$
$$H_a : \rho_1 \neq 0$$

(where ρ_1 is the lag 1 autocorrelation) by rejecting the null hypothesis (that "I" is correct) if the sample autocorrelation is less than $-2/\sqrt{n}$ or greater than $+2/\sqrt{n}$. Values of the sample autocorrelation between $-2/\sqrt{n}$ to $+2/\sqrt{n}$ are judged consistent with "I". Although the cutoffs $\pm 2/\sqrt{n}$ are reasonable, you may have reasons for preferring cutoffs that are more stringent or less stringent.

For now, I defer presenting quantitative tests for L and H.

Failure of any one of the L, H, I tests disqualifies the RS model.

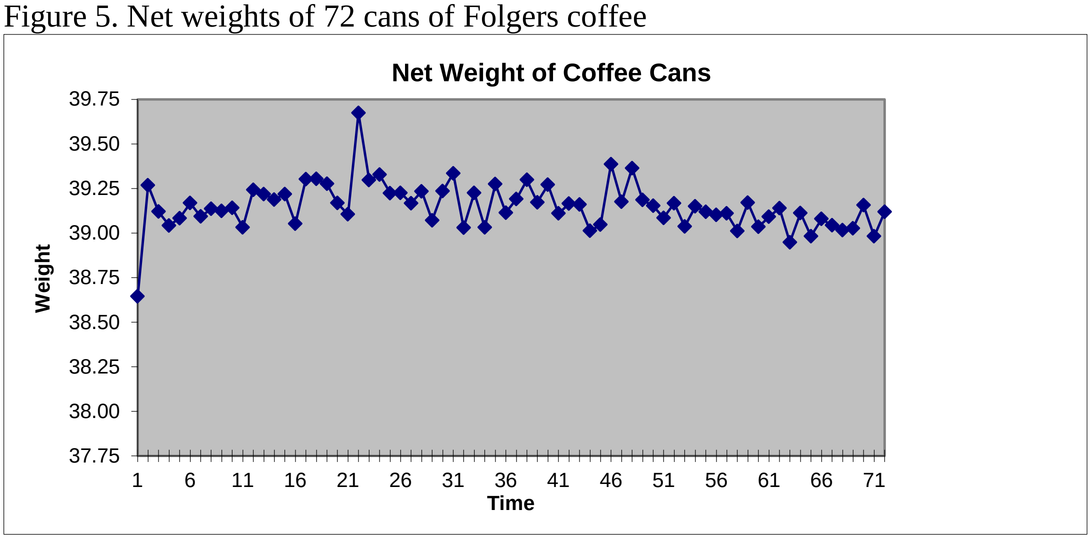
An example: Folgers coffee

One can of coffee was taken every hour for 72 consecutive hours from a production line, and the coffee was weighed. Each can was stamped "39 ounces net weight." Figure 5 shows the timeplot of the actual weight of the coffee that was in each of the 72 cans.

Does a RS model reasonably apply to these data?

Visually, it appears that the general level of the weights remains roughly constant over time – at just above 39 ounces. Visually, it appears that the up-and-down variability of the weights is roughly the same over time – following a channel from about 39.00 to about 39.30. So L and H appear to be roughly satisfied. For "I", I calculated the (lag 1) autocorrelation coefficient to be 0.0965. Since $2/\sqrt{n} = 2/\sqrt{72} = 0.2357$, then the autocorrelation is well within expected deviation from zero. We conclude that these data are consistent with the RS model.

Figure 5. Net weights of 72 cans of Folgers coffee



3. Use: It's a RS – so what?

Once you know that a time series is a RS (or at least reasonably close), then you can forecast it and apply all of the other theory that has been developed for i.i.d. data.

Let us start with forecasting. How to forecast the next value of a RS?

That is easy. Since the time series has been validated as a RS, then it is level. As long as it remains a RS, it will stay level. So you should expect the next value to be about at the same level as the past data. The past level is the mean level, best estimated by the sample mean of the past data. So the next value should be at about the mean of the past data.

Forecast of next value Y_{t+1} of RS = mean of past data.

But we should not expect the next value to be exactly equal to the sample mean. It will vary. By how much? Since the time series has been validated as a RS, then it is homoscedastic. As long as the time series remains a RS, the variation of individual values will be about the same in the future as it has been in the past. Past variation of individuals can be estimated by the standard deviation of the past data. So future variation can be estimated by the standard deviation of past data.

Average margin of error for forecast Y_{t+1} of RS = ± s. [6]

where $s = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$ is the sample standard deviation of the past individual data.

If the Y's are also normally distributed (they may not be), then $\bar{y} \pm s$ is a 68% confidence interval for the true mean of the common distributions. So we would have 68% confidence that the next value Y_{t+1} will lie within $\bar{y} \pm s$.

If the Y's are not normally distributed, an empirical estimate of the confidence that the next value Y_{t+1} will lie within $\bar{y} \pm s$ can still be given. Since the RS model has been validated, all of the past distributions are the same. As long as the time series remains a RS, future distributions will remain the same as past distributions. So the future proportion of data that lie within $\bar{y} \pm s$ should be like the past proportion of data that lie within $\bar{y} \pm s$. So tally the occurrences of past data within $\bar{y} \pm s$ and divide by n. The result is an estimate of the probability that the next value Y_{t+1} will lie within the range $\bar{y} \pm s$.

See the Folgers coffee example (next) for additional uses of the RS model.

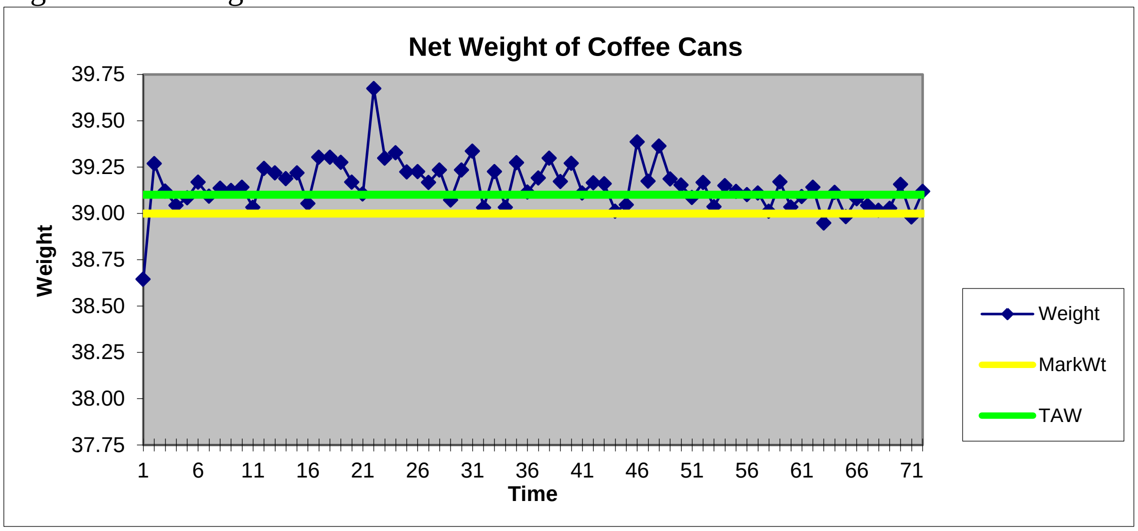
[6] This margin of error is a little too narrow. A slightly better error margin is $\pm s\sqrt{1+1/n}$.

An example: Folgers coffee (continued)

Earlier, I verified by mostly qualitative means that Folgers coffee weights satisfy LHI reasonably well and so qualify as a RS. The discussion that follows provides a forecast and also develops some additional insights that the RS validation can bring.

Figure 6 repeats Figure 5 but adds two horizontal lines.

Figure 6. Net weights of 72 cans of coffee



The yellow horizontal line is the net weight stamped on each can = 39 ounces. Nearly all cans exceed 39 ounces. (Can you guess why?) The green line is the target weight = 39.1018 ounces. This is the weight set on the filler machine. That is, the filler is set to put 39.1018 ounces of coffee into each can – somewhat more than the stamped weight. (Can you guess why?) But Figure 6 shows that the majority of cans exceed even the target weight. Does the filler need to be fixed to bring its average back down to the target weight? No – not if the deviations from the green line are just the usual random variation. Yes – if the deviations from the green line are too big to be usual random variation. Even if analysis confirms that the machine is overfilling, we might decide to live with overfilling if the costs of repair exceed the costs of overfilling.

How to test whether the overfilling is systematic and not random? We need to know if the mean of the process has drifted from 39.1018. So we can test hypotheses about the mean μ of the weight distribution:

$$H_0 : \mu = 39.1018 \quad \text{(no drift)}$$

$$H_a : \mu \neq 39.1018 \quad \text{(drift)}$$

The test is based upon the fact that if the data are a RS (i.e., i.i.d.), then the distribution of the sample mean $\bar{Y}$ is approximately normal with mean μ and standard deviation $s/\sqrt{n}$, if n is sufficiently large (rule of thumb: $n \geq 30$). So if the sample mean is more than roughly $\pm 2\, s/\sqrt{n}$ from 39.1018, then the p-value would be < 0.05, and the conclusion would be that the deviations are too big to be consistent with the process mean being 39.1018, so reject H_0.

Let us now see what the data say about this. The sample data are $\bar{y}$ = 39.1491, s = 0.1313. So the sample mean differs from 39.1018 by $\dfrac{\bar{y} - 39.1018}{s/\sqrt{n}} = \dfrac{39.1491 - 39.1018}{0.1313/\sqrt{72}}$ = 3.0584 standard deviations. The p-value for this is $P(Z > 3.0584) + P(Z < -3.0584)$ = 0.0022. So reject H_0. In this analysis, the properties of RS are key. RS permits this test to be done. We conclude that repairs are warranted provided the cost of repair is less the cost of overfilling. Now the difference between the actual mean weight of 39.1491 and the target weight of 39.1018, namely 0.0473 ounces of coffee, does not sound like much. But this is the weight per can of coffee, and millions of these cans are produced at this plant each year. If sustained throughout the year, the accumulated value of the surplus coffee being added to the cans can easily pay the salary of an extra worker. So practical concerns support the recommendation of the statistical test: Call for repair.

Forecast: The weight of the next can is forecast to be the sample mean = 39.1491 ± 0.1313. If the can weights are normally distributed, there is about a 68% probability that the weight of the next can will lie within this range. It is not yet known whether or not the can weights are normally distributed.[7] Absent a validation of normal distribution for can weights, an empirical assessment of confidence can be given by counting: There are 56 weights in the range 39.1491 ± 0.1313. This is the proportion 56/72 = 0.7778. So the range 39.1491 ± 0.1313 has approximately 77.78% empirical confidence for containing the next can weight.

Disastrously wrong results would ensue from interchanging the standard deviations in the forecast and test. If the test had used s instead of $s/\sqrt{n}$, then the test would compute z = $\frac{\bar{y}-39.1018}{s} = \frac{39.1491-39.1018}{0.1313} = 0.3602$ and would accept H_0, and the process would continue overfilling cans. The forecast number would be the same (39.1491), but its uncertainty would be assessed an order of magnitude too small (0.1313 / $\sqrt{72}$ = 0.01547), creating overconfidence in the forecast value. If the forecast uncertainty had used $s/\sqrt{n}$ instead of s, then the forecast uncertainty would be far too small (39.1491 ± 0.0155), creating considerable overconfidence in the forecast.

Individual uncertainty vs. group uncertainty.

The preceding three paragraphs present an especially compelling case for using the correct standard deviation in probability calculations. The forecast involves uncertainty about an *individual* value: the weight of the next can. So the appropriate uncertainty measure is the uncertainty of individuals – individual can weights, of which there are 72 individual examples. We can examine these 72 individual cans to get an idea of how much individual cans have varied in the past. In a RS, they should continue to vary by about the same amount in the future. The standard deviation of individual can weights – 0.1313 ounces – is the appropriate amount by which we expect the weight of the next can to vary from the mean of its uncertainty distribution.

On the other hand, the hypothesis test involves uncertainty about a *group* statistic: the mean of the 72 cans. How much can we expect the sample mean of the group to differ from the true mean of the process? Statistical theory for Random Samples tells us that the group mean is much less variable than the individuals in the group. The standard deviation of 72-can means is 0.1313 / $\sqrt{72}$ = 0.01547 – an order of magnitude less than the variability of individual cans (0.1313).

Generally, forecasting is for individual values, not group statistics.

[7] Normality should not be taken for granted. Like other possible specifications, normality can be tested. (Not yet – be patient.)

What proportion of cans are expected to be underweight? That is, what proportion are less than 39 ounces? The proportion underweight can be estimated: First count the number of underweight cans. There are 4. This is 4/72 = 0.0556 proportion underweight. Under the specifications of the RS model, the sample proportion that are underweight has an approximate normal distribution with mean = true proportion π and standard deviation $\sqrt{\pi(1-\pi)/n}$. So we estimate the proportion underweight as 0.0556 ± $\sqrt{\pi(1-\pi)/n} \approx$ 0.0556 ± $\sqrt{\dfrac{4/72(1-4/72)}{72}}$ [8] = 0.0556 ± 0.0277. There is about 68% confidence that the true proportion π of underweight cans lies in the interval 0.0556 ± 0.0277. This, too, relies upon the specifications of a RS.

Some Consequences of Ignoring Model Validation

I emphasize the importance of validating a model. So what can go wrong if you ignore my advice? I have a few illustrations (see Figure 7).

Figures 7a, 7b, 7c. Consequences of violations of L,H,I

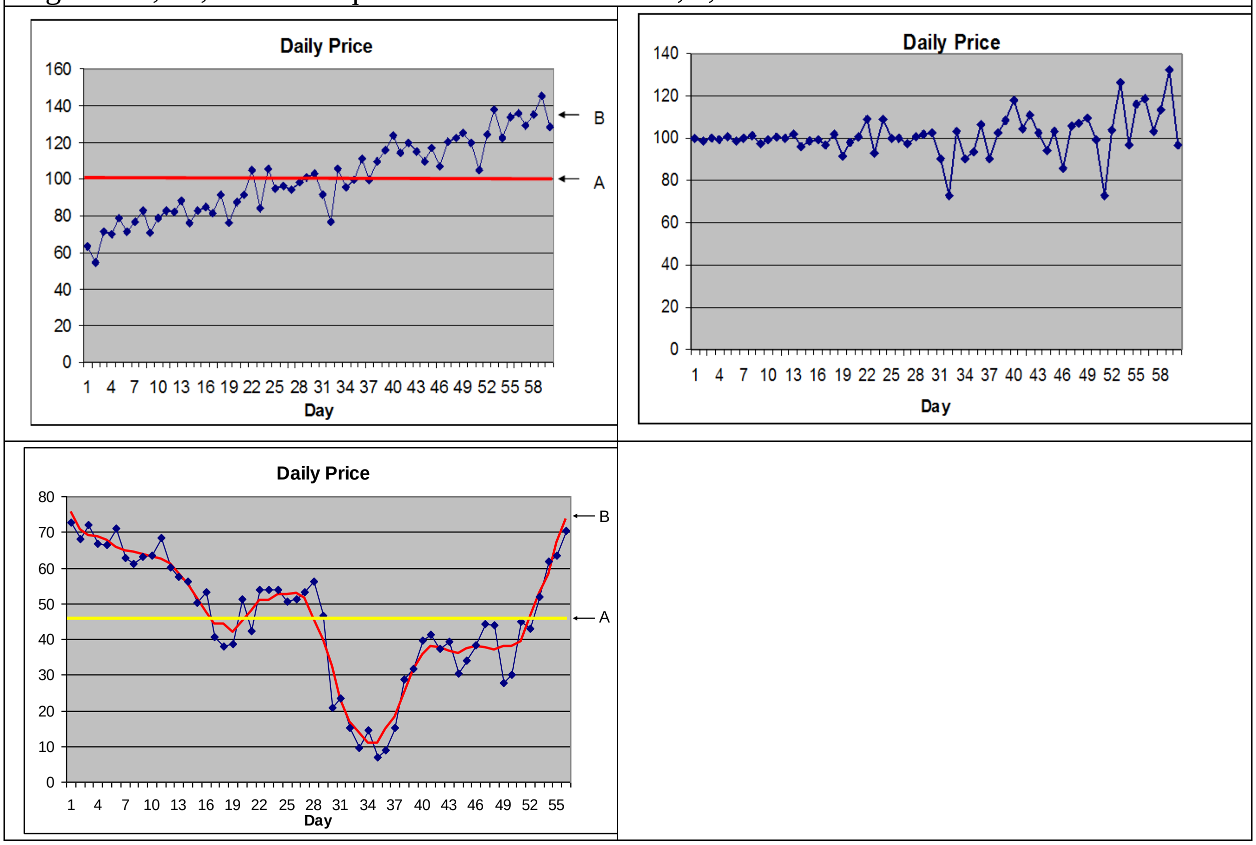


Figure 7a (upper left) repeats Figure 4a, but with the addition of a level line (red line) at 100. If you assume that the data in Figure 7a are a RS without validation, then the RS procedure of forecasting leads to predicting that the next value will be the mean of the data - at point A.

[8] The standard error uses 4/72 as an estimate of π for numerical calculation.

However, if the uptrend continues, then the next value is far more likely to occur near B. The RS estimate is biased. The uptrend violates L.

Figure 7b (upper right) repeats Figure 4b. If you assume that the data in Figure 7b are a RS without validation, then the RS procedure of forecasting leads to predicting that the next value will be around 100. There is nothing wrong with that forecast. The timeplot is level. So the next value should be expected around that level – namely, around 100. However, the problem arises in assessing the uncertainty of the forecast. The RS model has constant variability. In the RS case, a margin of error could be given for the forecast by using the standard deviation of past data. But H is violated in Figure 7b. The variability is increasing with time. Consequently, use of the standard deviation of past data, as prescribed by the RS methodology, will underestimate the uncertainty of the forecast. The margin of error will be too small. You will be overconfident in the forecast.

In Figure 7c (lower left), the “I” condition is violated. There is strong positive autocorrelation. In this case, use of the RS methodology without validation leads both to a biased forecast and to misjudging the confidence. The RS forecast is the mean level line (yellow line) at A. The RS margin of error is the standard deviation around the yellow line – this is a large number. However, the mean of the time series is actually following the meandering red line. The next value is far more likely to be at B than at A. The RS forecast is biased. Moreover, the appropriate margin of error is the variability around the red line, not around the yellow line. The average variability around the red line is far less than the average variability around the yellow line. So a much better forecast could be given with much more confidence if the true model were recognized.