Quality in Primary Care 2006;14:111–19

Principles of quality improvement

Statistics and reality: part 2

Davis Balestracci MS
Harmony Consulting, Portland, USA

ABSTRACT

In Part 1, after discussing why a process-orientated view of statistics is necessary, the conclusion reached was that the best statistical analysis will encourage the art of asking better questions in response to variations in data.[1] The purpose of this article is to debunk the myth that statistics can be used to 'massage' data and prove anything. It will demonstrate the counterintuitive simplicity and power of merely 'plotting the dots'– simple time plots of process outputs. These alone usually yield far more profound questions than the most complicated (alleged) statistical analysis.

Keywords: common cause, common cause strategy, run chart, special cause, stable process, tampering, variation

Old habits die hard

Suppose that it is time for the dreaded quarterly meeting about 'how you're doing' regarding your system's three hospitals' infection rates. The summary table is shown in Table 1 (the data are fictitious). But, first, of course, one is 'obligated' to show the data through what this author has found to be the most *useless* of tools: the bar graph (Figure 1).

So ... how are you doing?

Luckily, your local statistical 'guru,' a 'Six Sigma black belt', has come to your rescue and written a report, which is handed out. Several 'significant' findings are shared:

1 'Pictures are very important. A comparative histogram was done to compare the distributions of the three hospitals' infection rates (see Figure 2). There seem to be no differences among the hospitals; however, the appearance of bell shapes suggests that we can test the normal distribution hypothesis so as to be able to perform more sophisticated statistical analyses.
2 The three data sets were statistically tested for the assumption of normality. The results (not shown) indicated that we can assume them to be normally distributed (P values of 0.889, 0.745, and 0.669,

Table 1 'Statistical' comparison of three hospitals' infection control performance

Infection rate	n	Mean	Median	Tr mean	SD	SE	Minimum	Maximum	Q1	Q3
Hospital 1	30	3.027	2.90	3.046	0.978	0.178	1.000	4.800	2.300	3.825
Hospital 2	30	3.073	3.10	3.069	0.668	0.122	1.900	4.300	2.575	3.500
Hospital 3	30	3.127	3.25	3.169	0.817	0.149	1.100	4.500	2.575	3.750

SD, standard deviation; SE, standard error of the mean; Q1, lower quartile; Q3, upper quartile; Tr, trimmed mean

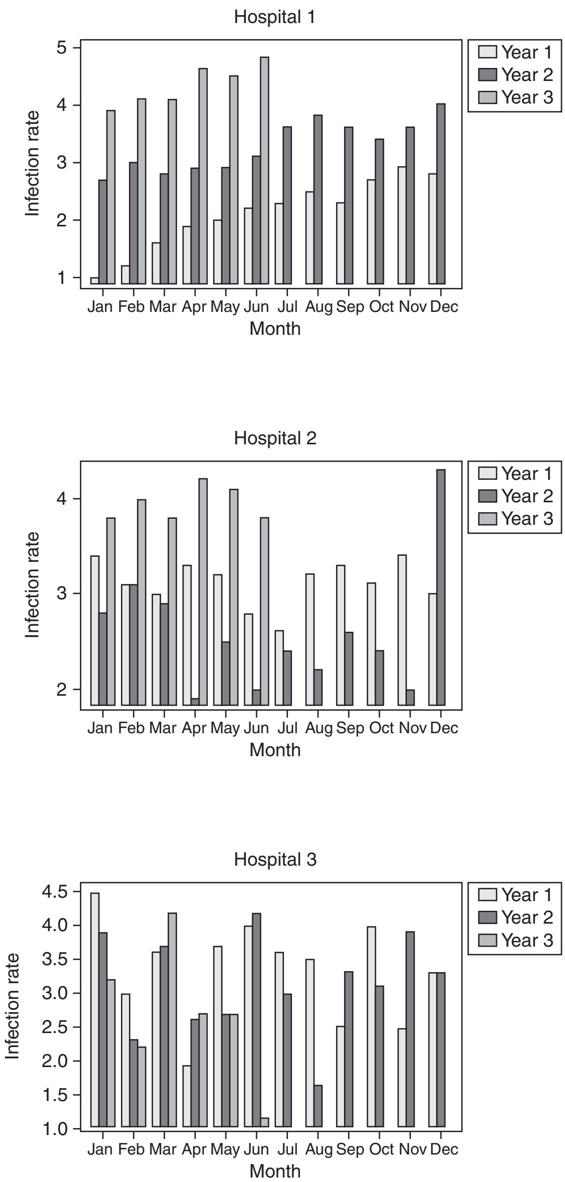

Figure 1 Charts showing three hospitals' infection control performance

respectively, all of which are > 0.05); however, we have to be cautious – *just because the data passed the test for normality does not necessarily mean that the data is normally distributed ... only that, under the null hypothesis, the data cannot be proven to be non-normal.*

3 Since the data can be assumed to be normally distributed, I proceeded with the analysis of variance (ANOVA) and generated the 95% confidence intervals (see Figure 3).
4 The *P* value of 0.897 is greater than 0.05. Therefore, we can reasonably conclude that there are no statistically significant differences among the hospitals, as further confirmed by the overlapping 95% confidence intervals.'

Let's see ... have we used all the potential jargon?

Mean ... median ... standard deviation ... trimmed mean ... quartile ... normality ... histogram ... *P* value ... analysis of variance (ANOVA) ... 95% confidence interval ... null hypothesis ... statistical significance ... Standard Error of the Mean ... *F* test ... degrees of freedom ...

Oh, by the way ... *did you know that this analysis is totally worthless?*

The good news is that you can forget most of the statistics that you have been previously taught. I can hear you all reassuring me, 'Don't worry ... we already have!'. Unfortunately, we remain a 'maths phobic' society, but there is no choice – whether or not we understand statistics, we are already using statistics!

We could think of an infection as 'undesirable variation' from an ideal state of 'no infections'. By acting on or reacting to this 'variation', you instinctively gather 'data' (hard or soft) upon which to assess the situation and take action to eradicate the variation. You have just used statistics – reacting to a perceived undesirable gap of variation from a desired state to close the gap!

However, there are two types of variation – common cause and special cause – and treating one type as the other, as is commonly done, actually makes things *worse.* As famous curmudgeonly W Edwards Deming once said, 'For every problem, there is a solution – simple, obvious ... and wrong!'.

But there is actually more good news. The statistics you need for improvement are far easier than you ever could have imagined. However, this philosophy will be initially quite counterintuitive to most of you and very counterintuitive to the people you work with,

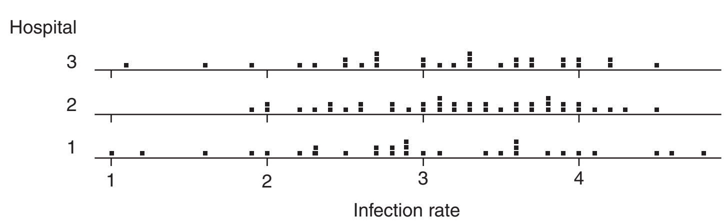

Figure 2 Histogram comparison of the three hospitals

```
Source      DF        SS       MS      F      P
Hospital     2     0.150    0.075   0.11  0.897
Error       87    60.036    0.690
Total       89    60.186
                                 Individual    95%   CIs  For  Mean

Level        N      Mean    StDev ---+----------+----------+----------+-
1           30    3.0267   0.9777 (----------------*---------------)
2           30    3.0733   0.6680   (----------------*---------------)
3           30    3.1267   0.8175    (----------------*---------------)
                                  ---+----------+----------+----------+-
Pooled StDev  =   0.8307            2.80       3.00       3.20       3.40
```

Figure 3 One-way analysis of variance

especially physicians trained in the statistics of research and clinical trials, which, unfortunately, are not appropriate for most quality improvement situations! If nothing else, it will at least make your jobs easier by freeing up a lot of time, by recognising when to walk out of time-wasting meetings! It will also help you gain the cultural respect you deserve as quality professionals because your data collections and analyses will be *simpler and more efficient ... and more effective!* The respect will also be magnified because you will be able to stop inappropriate responses to variation that would make people's jobs more complicated without adding any value to the organisation.

Back to the three hospital infection rate data

There are three questions that should become a part of every quality professional's vocabulary whenever faced with a set of data for the first time:

1 how were the data defined and collected ... and were they collected specifically for the current purpose?
2 were the systems that produced these data stable?
3 were the analyses appropriate, given the way the data were collected and the stability state of the systems?

How were the data collected?

Table 1 represents a descriptive statistical summary for each hospital of 30 numbers that were collected monthly. At the end of each month, each hospital's computer calculates the infection rate based on incidents observed and normalises to a rate adjusting by the number of patient days. No one has looked at the calculation formula for the last 10 years.

Were the systems that produced these data stable?

This might be a new question for you. As was made clear in Part 1, everything is a process.[1] All processes occur over time. Hence, all data have a 'time order' element to them that allows one to assess the stability of the system producing the data. Otherwise, many statistical tools can become useless and put one at risk for taking inappropriate actions. Therefore, it is always a good idea as an initial analysis to plot the data in its naturally occurring time order.

Were the analyses appropriate, given the way the data were collected and stability state of the systems?

'But ... the data passed the normal distribution test. Isn't that all you need to know?' you ask.

And your local 'guru' also concluded that there were no statistically significant differences among the hospitals. Well, now consider the three simple time plots for the individual hospitals (see Figure 4).

No difference?! ...

Note that just by 'plotting the dots', you have far more insight and are able to ask more incisive questions whose answers will lead to more productive system improvements.

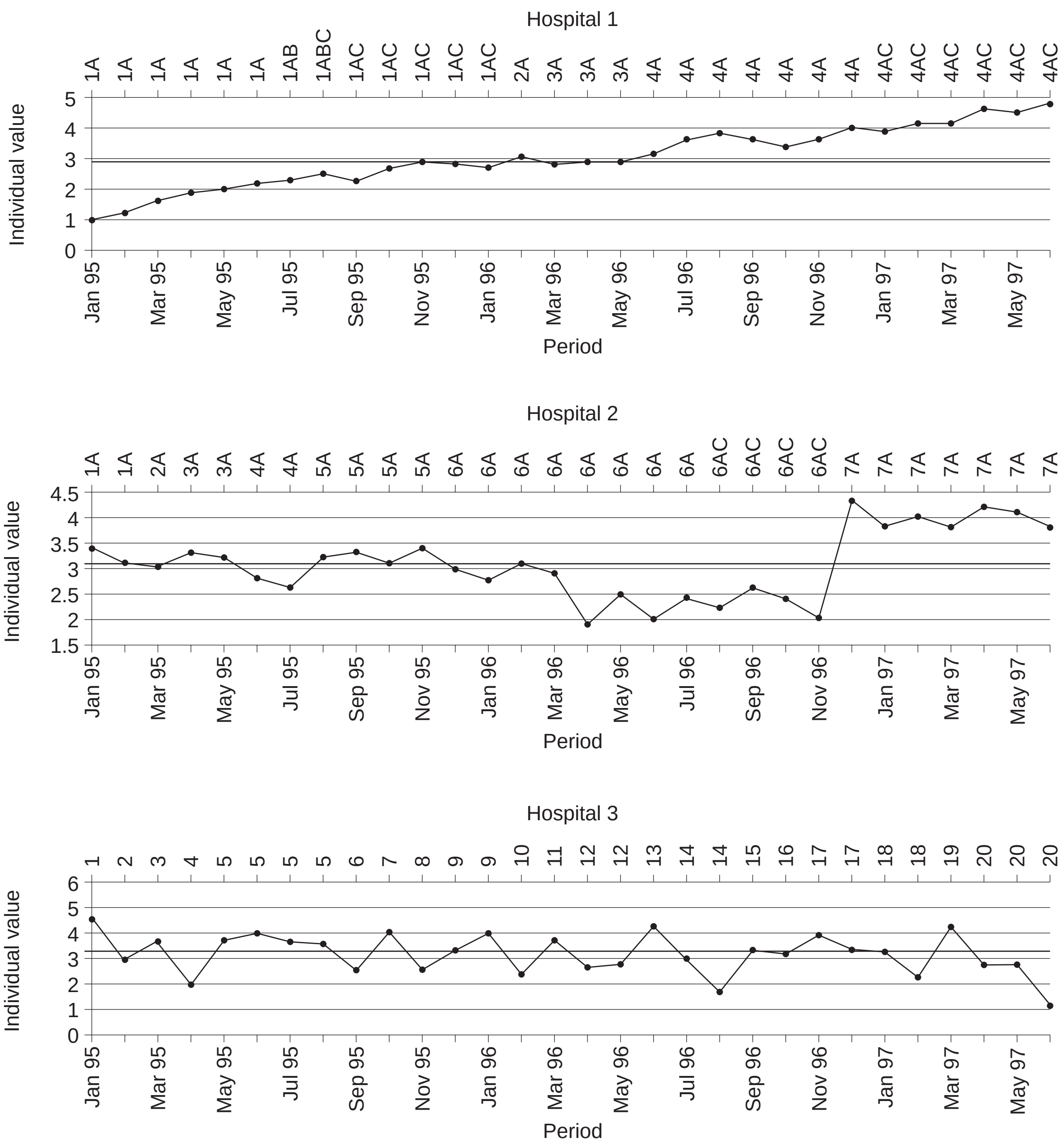

Figure 4 Time plots for the three hospitals

Compare this to having only the bar graphs, summary tables, and the 'sophisticated' statistical analyses. What questions do you ask from those? Would they even be helpful? 'Unfortunately', you are all smart people. You will, *with the best of intentions*, come up with theories and actions that could unwittingly *harm* your system. Or, worse yet, you might do nothing because 'there are no statistical differences' among the systems. Or, you might decide, 'We need more data'.

Regarding the computer-generated 'statistics', what do the 'averages' of Hospital 1 and Hospital 2 mean? I'll tell you: 'If I stick my right foot in a bucket of boiling water and my left foot in a bucket of ice water, on average, I'm pretty comfortable'. It is inappropriate to calculate averages, standard deviations, etc on unstable processes.

Also, note that you haven't calculated one statistic, yet you have just done a powerful statistical analysis! To summarise: '*plot the dots!!!*'.

More on 'plotting the dots': common and special causes

Almost all quality experts agree that merely plotting a process' output over time is one of the most simple, elegant, and awesome tools for gaining a *deep* understanding of any situation. Before one can plot, one must ask questions, clarify objectives, contemplate action and review current use of the data. Questioning from this statistical thinking perspective leads immediately to unexpected deeper understanding of the process. This results in establishing baselines for key processes and then allows honest dialogue to determine meaningful goals and action.

A more typical process is to impose arbitrary numerical goals that are 'retrofitted' onto the process and 'enforced' by exhortation that treats *any* deviation of process performance from the goal as *unique* and needing explanation – known as a *special cause* strategy. In paraphrasing the question and looking at specific undesirable events into the context of observing a process over time:

> Is this an isolated excessive deviation ('special cause') or, when compared to the previous measurement of the same situation, does it merely reflect the effects of ongoing actions of process inputs that have *always* been present and can't be predicted ahead of time ('common cause')? Would I necessarily expect the exact same number the next time I measure? If not, then how much difference from the current or previous number is 'too much'?

It is very important to realise that just because one can explain an occurrence 'after the fact' does not mean that it was 'unique'. Thinking in terms of process, you have inputs causing variation that are *always* present and conspire in *random* ways to affect your process' output; however, they conspire to produce a predictable *range* of possible outputs. Many explanations merely point out things that have been *waiting* to happen ... and will happen again at some *random* time in the future! Also, your process 'prefers' some of these 'bogeys' to others. So, how can you collect data to find these deeper solutions to reduce the range of variation encountered by customers ('common cause' strategy)?

So, if a process fluctuates within a relatively fixed range of variation, it is said to exhibit *common cause variation* and it can be considered 'stable' and predictable – although one may not necessarily like the results. If there is evidence of variation *over and above* what seems to be inherent, the process is said to exhibit *special cause variation.* This usually occurs in one of two ways: either isolated single data points that are totally out of character in the context of the other data points or a distinct 'shift (or shifts)' in the process level due to outside interventions (intentional or unintentional) that have now become part of the everyday process inputs.

The most common error in improvement efforts is to treat common cause (inherent) variation as if it were special cause (unique) variation. This is known as *tampering* and will generally add more complexity to a process without any value. In other words, *despite the best of intentions*, the improvement effort has actually made things worse.

The 'reward' luncheon

You have been invited to a free pizza lunch in celebration of meeting a safety goal. Two years ago, your organisation had 45 undesirable 'incidents' and set a goal in the past year of reducing them by at least 25%. The December data are in, and the yearly total was: 32 incidents – a 28.9% decrease!

Various graphs were used to prove that the goal had been met.

Figure 5, the obligatory bar graph display, shows that adverse incidents in eight months of the second year were lower than those in the previous year – 'obviously' an improvement!

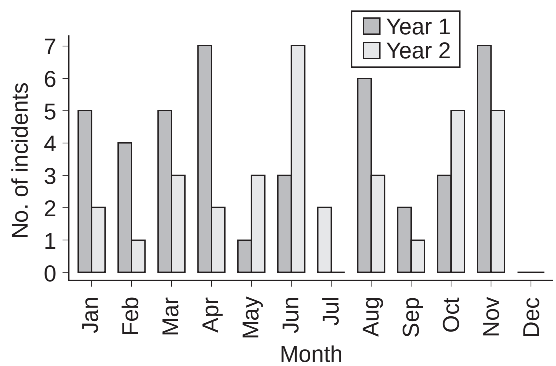

Figure 5 Undesirable incident data: two years, plotted by month

However, as Figure 6 so 'obviously' shows, the improvement was much better than originally thought! The local statistical 'guru' did a trend analysis (Thank God for Excel!), which showed a 46.2% decrease! The 'guru' also predicts 20 or fewer accidents for the next year.

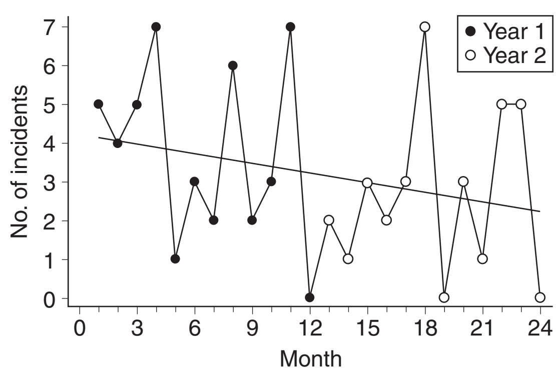

Figure 6 Trend analysis of undesirable data: 4.173 to 2.243 – 46.2% decrease!

You think of all the hard work in the monthly safety meeting where *each individual incident* is dissected and discussed to find root causes. Then there are the months where you have zero incidents and the reasons for this are discussed and implemented. It all paid off ... or did it?

Run charts

Imagine these data as 24 observations from a process. The chart in Figure 7 is known as a *run chart* – a time-ordered plot of the data with the overall data *median* drawn in as a reference. The median is the empirical 'midpoint' of a dataset, irrespective of time order. Half of the data values are literally higher than the median value and half of the data values are lower. In the present case, the median is 3 (you could sort these 24 observations from lowest to highest and the median would be the average of the 12th and 13th observations in the sorted sequence – both of which happen to be 3).

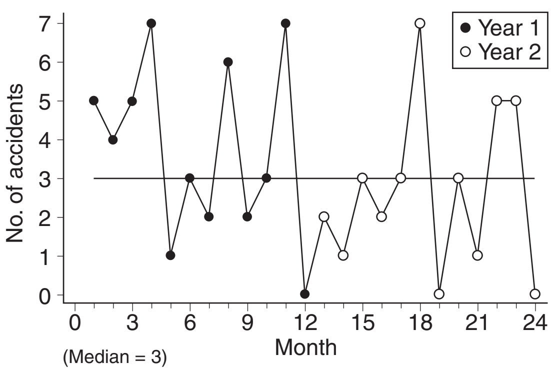

Figure 7 Run chart for accident data January 1989 to December 1990

The initial run chart of a situation always 'assumes' no change was made ('innocent until proven guilty'). A deceptively powerful set of three rules based in statistical theory can be applied to a chart like this to see whether your 'special cause' (you did intervene for the specific purpose of creating a change in the average, didn't you?) did indeed affect the level of the process.

So, the question in this case becomes, 'is the *process* that produced the 12 data points of the second year the same as the *process* that produced the data points of the first year?'. The three statistical tests, called a *runs analysis*, give no evidence of a change.[2] Thus, despite the (alleged) achievement of what was seen as an aggressive goal, there is no statistical evidence of it having been met. It just goes to show you: 'given two different numbers, one will be bigger!'.

However, regardless, even among people who agree either 'yes' or 'no', there could still be as many different ways of coming to this conclusion as there are people reading this, which would result in differing proposed actions – and a *lot* of formal meeting time! What is needed is a common approach for quickly and appropriately interpreting the variation through statistical theory.

Thus, given two numbers (45 and 32), one was smaller – and it also happened to *coincidentally* meet an aggressive goal. The 'year-over-year' and 'trend' analyses were inappropriate – and very misleading.

And it suddenly hits you: The result of all the hard work in the monthly safety meetings has been no improvement over two years and a lot of unneeded new policies!

In fact, if you continue to use this strategy (treating common cause as if it were special cause), you will observe between 20 and 57 accidents the following year!

So ... does 'common cause' mean we have to live with it?

Not at all. In the case of this data, people were treating data from a stable process exhibiting common causes of variation as if there were special causes of variation. Any observed differences were due totally to chance. *Looking at individual numbers or summaries and calling any differences 'real' is a no yield strategy, as is looking at accidents individually.* Once again, treating common cause as if it were special cause: tampering. Statistics on the number of accidents do not prevent accidents.

A *common cause strategy* looks for underlying patterns producing the data – a statistical 'slicing and dicing', if you will, to try to expose process inputs that could be accounting for a significant source of the process' variation.

In the case of the adverse event data, one might ask, 'is there a *commonality* among all the high-event months ... or the low-event months ... or the months where there were zero events? Are some accidents *unique* to certain departments? Do some accidents occur in *all* departments? Does one department exhibit a disproportionate total of accidents because its safety policy enforcement process is sloppy overall?' These questions address process *patterns* that are exerting their influence *consistently* as inputs to the safety process. *Neither the monthly data points nor individual accidents should be treated uniquely in isolation.* It is only by looking at the aggregated factors contributing to *all* 77 accidents where opportunities in the underlying process inputs will be exposed.

Think of an 'accident' as: a hazardous situation that was unsuccessfully avoided!

It is a common approach to have 'incident reviews' every month and go over *every single* incident individually – in essence, 'scraping it like a piece of burnt toast' – and making a recommendation after each review. Can you see that this treats each incident as if it were a special cause?

Smart people have no trouble finding 'reasons' if they look hard enough ... after the fact. There is a high risk of treating spurious correlation as cause-and-effect, which only adds unneeded complexity to the current process, but *no value*.

This also has implications for the current issue of 'sentinel event' analysis without asking the question:

> Was this an individually *isolated event* (special cause) or is the process such that it has been *waiting to happen* because the process inputs all randomly aligned to make it happen (common cause) ... which means that it could happen again?

Summary

As quality professionals, it is important to realise that data analysis goes far beyond the routine statistical 'crunching' of numbers and useless bar graph displays. The greatest contribution to an organisation is getting people to understand and use a *process-oriented* context in analysing situations as well as principles of good, *simple, efficient* data collection, analysis and display. This cannot help but enhance the healthcare quality professionals' credibility. It will also help gain the confidence and co-operation of organisational culture during stressful transitions and investigations. It will be vital to put a stop to many of the current well-meaning but ultimately damaging *ad hoc* uses of statistics. *Whether or not people understand statistics, they are already using statistics* ... and with the *best* of intentions.

As a final summary, Box 1 shows the key lessons to keep in mind as you start looking at your organisation through a lens of statistical thinking and Box 2 and Table 2 summarise the statistical mindsets of Parts 1

Box 1 Key lessons

- Make sure that any data being used were collected *specifically* for the current purpose
- Understand *how* the numbers were calculated and how the data were *collected*
- Make sure any analysis is *appropriate* for the way the data were collected:
  - tables of raw numbers, summary 'stats', and bar graph presentations are virtually worthless
  - the 'normal' distribution is highly overrated and very rarely used in improvement
  - 'traditional' calculation of the standard deviation will typically yield an inflated estimate
- 'Plotting the dots' of an indicator over time is a powerful but simple method for studying a process
- Arbitrary numerical goals by themselves improve nothing – *your processes are currently perfectly designed to get the results they are already getting ... and will continue to get*
- Reacting to individual data points and individual 'incidents' in a 'stable' (common cause) system is, many times, a *no yield* strategy
- A 'stable' system can be 'dissected' statistically to look for hidden opportunities

Box 2 The fundamentals of variation

- Good data collection requires planning, which is equally important as the data itself:
  - the first question must be, 'What is the objective?'.
- Good data analysis requires knowing how the data were collected or will be collected. This analysis must be appropriate for the method of collection:
  - raw data say little
  - graphical methods are the first methods of choice.
- All data result from a measurement process:
  - is the measurement process agreed upon and reliable?
- Variation exists in all things, but may be hidden by:
  - excessive round-off of the measurement
  - excessive aggregation
  - using rolling averages.
- All data occur as outputs from some process and contain variation. This variation has caused and has sources that can be better identified through proper data collection.

Box 2 Continued

- There are sources of variation due to *inputs* to a process (people, methods, machines, materials, measurement and environment) and variation in *time* of these individual inputs as well as their aggregate. Both of these are reflected in the output characteristic being measured.
- The stability of the process (over time) producing the data is of great importance for prediction or taking future action.
- All data occur in time:
  – neglect of the time element may lead to invalid statistical conclusions.
- Any source of variation can be classified as either a common cause or a special cause. It is important to distinguish one from the other to take appropriate action:
  – the presence of special causes of variation may invalidate the use of certain statistical techniques.
- There is variation (uncertainty) even when an object is measured only once.

Table 2 Four common statistical traps

Trap	Problem	Comment
Trap 1: treating all observed variation in a time series data sequence as special cause	Most common form of 'tampering' – treating common cause as special cause	Given two numbers, one will be bigger! Very commonly seen in traditional monthly reports: month-to-month comparisons; year-over-year plotting and comparisons; variance reporting; comparisons to arbitrary numerical goals
Trap 2: fitting inappropriate 'trend' lines to a time series data sequence	Another form of 'tampering' – attributing a specific type of special cause (linear trend) to a set of data which contains only common cause	Typically occurs when people always use the 'trend line' option in spreadsheet software to fit a line to data with no statistical trends
	Attributing an inappropriate specific special cause (linear trend) to a data time series that contains a different kind of special cause	Improvement often takes place in 'steps,' where a stable process moves to a new level and remains stable there. However, a regression line will show statistical significance, implying that the process will continually improve over time
Trap 3: unnecessary obsession with and incorrect application of the normal distribution	A case of 'reverse' tampering – treating special cause as common cause	Ignoring the time element in a dataset and inappropriately applying enumerative techniques based on the normal distribution can cause misleading estimates and inappropriate predictions of process outputs
	Inappropriate routine testing of all data sets for normality	Mis-applying normal distribution theory and enumerative calculations to binomial- or Poisson-distributed data
Trap 4: improving processes through the use of arbitrary numerical goals and standards	Any process output has a natural, inherent capability within a common cause range. It can perform only at the level its inputs will allow	Goals are merely wishes regardless of whether they are necessary for survival or arbitrary. Data must be collected to assess a process' natural performance relative to a goal

and 2 in 'The 10 Fundamentals of Variation' and 'Four Common Statistical Traps', respectively.

REFERENCES

1 Balestracci D. Data 'sanity' – statistics and reality. *Quality in Primary Care* 2006;14:49–53.

2 Balestracci D and Barlow J. *Quality Improvement: practical applications for medical group practice* (2e). Englewood, CO: Center for Research in Ambulatory Health Care Administration (CRAHCA), 1996. Phone: +303–397–7888 to order ($8) or access through the following link: www5.mgma.com/ecom/Default.aspx?tabid=138&action=INVProductDetails&args=479 (accessed 16 March 2006).

FURTHER READING

Balestracci D. *Data 'Sanity': statistical thinking applied to everyday data.* Special Publication of the American Society for Quality Statistics Division, Summer 1998. (Write to the author for a PDF copy or download from http://deming.eng.clemson.edu/pub/den/deming-papers.htm)

ADDRESS FOR CORRESPONDENCE

Davis Balestracci, Harmony Consulting, 94 Ashley Ln., Portland, ME 04103–2789, USA. Tel: +207 899 0962; email: davis@dbharmony.com; website: www.dbharmony.com

Received 18 January 2006
Accepted 10 March 2006