STATISTICAL INFERENCE AT WORK: STATISTICAL PROCESS CONTROL AS AN EXAMPLE

ARTHUR BAKKER
Freudenthal Institute, Utrecht University & Institute of Education, University of London
a.bakker@fi.uu.nl

PHILLIP KENT
Institute of Education, University of London
p.kent@ioe.ac.uk

JAN DERRY
Institute of Education, University of London
j.derry@ioe.ac.uk

RICHARD NOSS
Institute of Education, University of London
r.noss@ioe.ac.uk

CELIA HOYLES
Institute of Education, University of London
c.hoyles@ioe.ac.uk

ABSTRACT

To characterise statistical inference in the workplace this paper compares a prototypical type of statistical inference at work, statistical process control (SPC), with a type of statistical inference that is better known in educational settings, hypothesis testing. Although there are some similarities between the reasoning structure involved in hypothesis testing and SPC that point to key characteristics of statistical inference in general, there are also crucial differences. These come to the fore when we characterise statistical inference within what we call a "space of reasons" – a conglomerate of reasons and implications, evidence and conclusions, causes and effects.

Keywords: *Statistics education research; Context; Evidence; Hypothesis testing; Space of reasons*

1. INTRODUCTION

Statistical inference involves drawing conclusions that go beyond the data and having empirical evidence for those conclusions. These conclusions have a degree of certainty, whether or not quantified, accounting for the variability that is unavoidable when generalising beyond the immediate data to a population or a process. This is in line with Makar and Rubin's (2007) analysis that key ingredients of statistical inference are generalisations (conclusions beyond the sample data), data as evidence, and a probabilistic language. An important rationale for characterising statistical inference in

the workplace in this paper is that such a study may indicate which types of statistical reasoning students might later need as employees. Our work-based analysis can inform discussions about what students should learn in statistics education and may complement recent school-based research into informal statistical inference (Bakker, Derry, & Konold, 2006; Ben-Zvi, 2006; Pfannkuch, 2006; Rubin, Hammerman, & Konold, 2006). In particular, this workplace research points to types of statistical inference that are not typically addressed at the secondary school level and yet can be useful to employees.

Statistics textbooks traditionally make a distinction between descriptive and inferential statistics to stress that students should not too easily jump from conclusions about samples to conclusions about a population. However, the distinction also leaves important types of statistical inference unaddressed: types of statistical inference (whether formal or informal) from samples to populations or processes that are different from the well-known and commonly taught inferential techniques of hypothesis testing and confidence interval estimation. Only a few students learn these inferential techniques while research in workplaces shows that many will need to draw conclusions about a process from a sample (e.g., Noss, Bakker, Hoyles, & Kent, 2007; Smith, 1999). For example, when monitoring and improving production processes, employees typically with little formal education are routinely supposed to draw conclusions from samples about the production process for which they are responsible.

Our research in workplaces suggests that few non-graduate employees need to interpret results stemming from hypothesis testing or confidence interval estimation, and even fewer produce such results. These formal techniques are mostly not at employees' disposal, and even if they are, it is often not possible or cost-effective to use them. What they need is to draw conclusions from samples with relatively simple techniques, and generally to base decisions on an awareness of the uncertainty that comes with generalising to a population or process. In other words, learning descriptive statistics (and even exploratory data analysis) does not suffice for the majority of students, while the aforementioned inferential techniques are, as taught in current curricula, beyond the scope of the vast majority.

The goal of this paper is to characterise a type of statistical inference required in many work settings, and we do so by analysing an example of a widely used statistical technique in which statistical inferences are made: statistical process control (SPC). Because the theory of SPC has some similarities with sequential hypothesis testing (more than, say, confidence interval estimation) and because hypothesis testing is a better known type of statistical inference within the educational world, we compare SPC with hypothesis testing. The central question addressed in this paper is therefore *How is the statistical inferential reasoning ideally involved in SPC similar and different from statistical inferential reasoning involved in hypothesis testing?*

To address this question we draw on data collected in our research into the mathematical and statistical knowledge required by intermediate-level employees in various industrial sectors (Tehcno-mathematical Literacies in the Workplace Project, 2003-2007). Such employees are typically non-graduates who may be working in manufacturing as skilled operators or supervisory managers.

After discussing the key ingredients of our question – statistical inference and SPC – we describe the origin of our empirical example and illustrate characteristics of statistical inference as observed in SPC. Last, we discuss the contribution that we think this endeavour has made to the study of statistical inference at work, the limitations of our exploratory approach, as well as potential implications for workplace training and school education and research.

2. THEORETICAL BACKGROUND

2.1. STATISTICAL INFERENCE WITHIN A SPACE OF REASONS

Our experience is that within the statistics community, "statistical inference" mostly connotes formal inferential techniques. However, we use the term inference here in its general sense of drawing conclusions, including the possibly tacit reasoning processes that precede and support the explicit inference from a premise to a conclusion, a prediction, or a conjecture. The term not only includes deduction and induction, but also *ab*duction. Abduction is inference to an explanation, a method of reasoning in which a hypothesis is formed that may explain the data. For example, 8-year-old students in Paparistodemou and Meletiou-Mavrotheris' (2007) study sometimes came up with abductive conjectures that would explain the data rather than inductive conclusions from the data – contrary to the teacher's and researchers' expectations. A similar observation is reported by Zieffler, Garfield, and delMas (2007) for college students.

In search of the delimitations of what counts as statistical inference, we ask the following question: Is the calculation of the means of two samples a statistical inference? In our view, this depends on the reason why they are calculated. For example, they might be calculated to know the difference between the two means in relation to the variation of the two samples. The ratio of this difference to a measure of variation (say, SD) can help us conclude whether the difference is big enough to be likely caused by a difference between two populations from which the samples were drawn. In this case, the very fact of attending to the calculation of the means and difference arises due to specific reasons related to populations.

One way to put it is that the attention to and choice of calculations take place in the "space of reasons" within which people act and think, where "reasons" refer not only to reasons in the strict sense but also to implications, evidence, conclusions, goals, purpose, utility, and our knowledge of causes and effects. For the philosophical background of this notion, originating in the work of Wilfrid Sellars, we refer to McDowell (1996). Our intention behind using this technical terminology is not only to recognize that it is impossible to provide a complete description of any particular context in which statistical inference takes place, but also to recognize that contexts involve not just material but also ideal elements, such as reasons.

To characterise statistical inference at work, it makes sense to make a brief comparison between school and workplace settings, which give rise to different spaces of reasons. At school, contexts are often used to learn about statistics, whereas in the workplace, statistics is more likely to be used to learn about the context. Paraphrasing a famous quote by Steen (2003, p. 55) we can observe that the workplace makes sophisticated use of elementary statistics whereas in the classroom we encounter elementary use of sophisticated statistics. Seminal research into the mathematics used on the street (Nunes, Schliemann, & Carraher, 1993) or in supermarkets (Lave, 1988) led to increasing popularity of situated cognition and socio-cultural theories. In the light of such research, it is likely that context plays a different role in statistical inference learned at school than in statistical inference used in workplaces.

At work, statisticians and practitioners using statistics do not lose the context of an investigation out of sight when using statistical techniques (Wild & Pfannkuch, 1999). More generally, Noss and Hoyles (1996) have introduced the notion of "situated abstraction" to capture both the process and product of situating abstract knowledge in real-life situations such as workplaces. Situated abstractions gain their meaning not only from the mathematics from which they stem but also from the context in which they are

used. Ethnographic studies of workplace situations by the same authors, for example of nursing (Noss, Pozzi, & Hoyles, 1999), show how mathematical and contextual meanings such as those of "average" are fully integrated with the particular purposes for which they are used – in this case monitoring the blood pressure of critically ill patients.

These studies show that a dichotomy between statistics and context is problematic (see also Cobb, 1999), especially in workplace situations. One difficulty with the concept of context is that it is not well defined. It is often used to indicate the location or setting in which theoretical ideas are used, but that is a too restrictive connotation for our purpose of characterising statistical inference at work. Given this paper's focus, we will not try to define "context" but rather we will attend to the space of reasons in which statistical inference takes place. In this way, we hope to overcome the dichotomy between statistics and context that seems so deeply engrained into our thinking.

2.2. STATISTICAL PROCESS CONTROL (SPC)

The question addressed in this paper is how statistical inference involved in SPC is similar or different from hypothesis testing. We assume readers are familiar with hypothesis testing but perhaps less so with statistical process control (SPC). We therefore briefly characterize SPC (Caulcutt, 1995; Oakland, 2003) before we describe the origin of the empirical example of SPC (Section 3) and analyse it (Section 4).

SPC is a process improvement technique deployed in many industrial sectors. It is typically used in situations where variability in items produced (or services offered) has to be minimal and key performance indicators need to be very close to a specific target. Measurements of the products can be plotted on an SPC chart so as to monitor the location and variability of the production process.

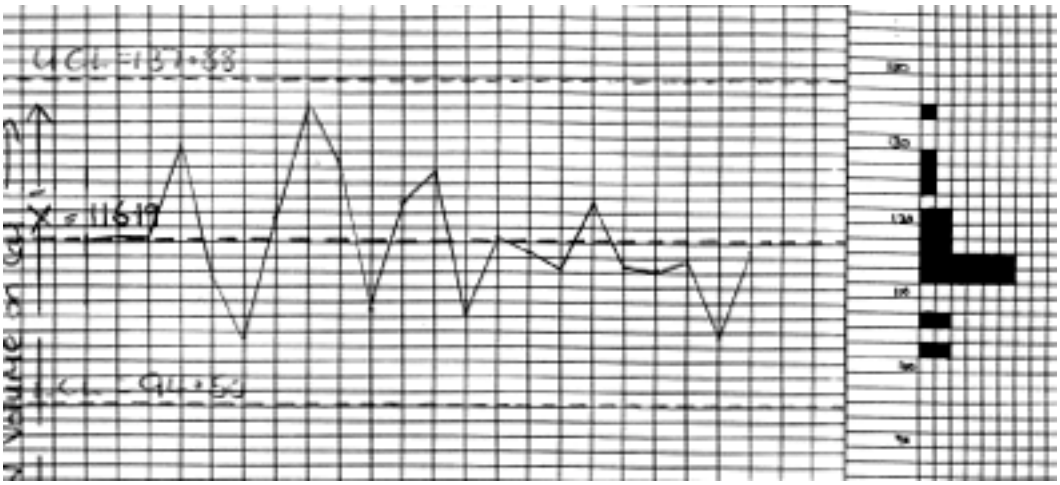

Figure 1. Part of an SPC chart on airtightness of cars

Figure 1 is part of an authentic SPC chart we collected in a car company. It shows the airtightness of a series of cars being tested as they came off the production line. On the left is the control chart with individual measurements; on the right a "sideways histogram" that is used to monitor the distribution of the measurements. The mean of this part of the process is 116.19 (in some unit of pressure that the employee we interviewed could not tell us). The dotted lines below and above the mean line are called the lower control limit (here LCL = 94.55) and the upper control limit (here UCL = 137.88),

respectively. These *control limits* are defined as follows: the lower control limit (LCL) is the mean – 3 SDs; the upper control limit (UCL) is the mean + 3 SDs. The use of these control limits is based on two assumptions:

1. The particular measure of items produced is distributed normally (this might be the means of subsamples) – this is checked with the sideways histogram,
2. data are independent and identically distributed.

If the assumptions are fulfilled we can assume that about 99.7% of all measurements will be within +/- 3 SDs of the mean. (Various techniques have been suggested for approximating this standard deviation calculation in a manner accessible to the employee on the production line.) Control limits are also called *action limits*, because non-conforming points (i.e., points outside the control limits, or exhibiting certain trends) are reasons for action. The action usually involves just checking, but sometimes goes on to include adjustments of settings or even stopping the production process for detailed investigation.

The assumption about 99.7% being between the control limits only holds if the mean of the process does not change, that is if there is only *common-cause* (random) variation. However, any *special cause,* such as the supply of different sealing material or a problem with the machine or measurement system, can lead to the process drifting off target. In that case, the cause needs to be identified and addressed. There are probability-based rules for detecting potential trends in such charts. For example, the chance of seven consecutive points on either side of the mean has a probability of only $2\times(1/2)^7 = 1/64 = 0.016$, an occurrence that is therefore unlikely to have been caused by random common-cause effects (this rule can be seen as a binomial hypothesis test). In such cases, one assumes that it signals a special cause (to be identified by abductive inference). In this way, control charts can be useful to detect deviations from the normal or target situation and remove the special cause before the process creeps out of control. Other rules on unnatural patterns include the following: One point outside of control limits; a trend of six points in a row increasing or decreasing; 14 points in a row that alternate up and down.

The idea of using control limits is that they predict process variation so one can stay well within *specification limits*, limits that are imposed by law, by customer requirements, or by senior managers and engineers in the company. Even if data points are occasionally outside the control limits, they will then not surpass any of the specification limits (the spec limits are not shown in Figure 1). If the control limits are well within the spec limits, the process is said to be stable and capable. If the process is stable, control limits can be calculated in preliminary studies, after which the process only needs to be monitored. These are in short the theoretical ideas behind the industrial statistics of SPC or, in other words, *part* of the space of reasons in which SPC takes place.

3. AN EMPIRICAL EXAMPLE OF STATISTICAL INFERENCE IN SPC

3.1. ORIGIN OF THE EXAMPLE

The example that we use to compare statistical inference involved in SPC with hypothesis testing stems from the *Techno-mathematical Literacies in the Workplace* research project. The project goals were to identify the mathematical and statistical knowledge required by intermediate-level employees in financial and manufacturing sectors, and based on these to design learning opportunities that would support employees in developing such knowledge. In manufacturing companies (pharmaceuticals and packaging) we have identified SPC as an important technique that is widely used and yet difficult for employees to understand and use due to the statistical knowledge required

(Hoyles, Bakker, Kent, & Noss, 2007). Contact with a high-level manager from a car company gave us the opportunity to investigate how process improvement techniques were actually used and trained on the shopfloor in automotive manufacturing.

We spent 18 researcher days in this particular car company. First, we interviewed the manager in charge of process improvement as well as the SPC experts and their managers. Shopfloor employees explained to us their control charts on the shopfloor. We also attended and evaluated their SPC course. Moreover, we distributed a questionnaire to twelve course participants (ten responded), and carried out three one-hour follow-up interviews with participants. Some of the operators, shift leaders and course participants had had little or no formal education since they were 16. Our data collection in this company included audio recordings, workplace artefacts such as SPC charts, notes made during training courses, a questionnaire, and trainers' PowerPoint presentations. In collecting data, we made sure in our interviews that we obtained different views of the same workplace activity, from the viewpoints of shopfloor employees, shopfloor supervisors, trainers, more senior managers, and statistical consultants.

In analysing data, we triangulated interpretations of the raw data sources (audio transcripts, photographs of workplaces, artefacts such as graphs) amongst the project team. We have also carried out design-based research so as to enhance existing SPC training, but will not report on it here.

3.2. AIRTIGHTNESS OF CARS

Our empirical example of SPC stems from an area of the production process where the airtightness of cars that are almost ready to leave the factory is checked. Kevin, an operator with no formal education since he was 16, is responsible for this. We cite him to sketch part of the space of reasons in which SPC is used and to give an example of non-statistical inference in terms of causes and effects:

> If the car is too airtight you will get a problem with the windows misting up all the time; also the doors will not shut. You need to lose some sort of air otherwise the door is just so airtight you would have to run at it and give it a good push.

To check whether the airtightness is within specification, Kevin blows air into the car, which is measured in cubic feet per minute, and he reads off the pressure this causes. Mostly the measurements are fine, but occasionally they are out of specification. This is where contextual reasoning is used: "As soon as I turn the gauge on it kicks in normally at around 60-70 Pascal. If it kicks in at around 30 I know full well that we have got a big leak somewhere." (Trying to understand what the numbers meant here was challenging to us, because he kept using different units for both air speed and pressure such as cubic feet per minute, litres per second, Pascal, weight per cm^2.)

Kevin has learned the probability-based rules on trends and patterns, but does not know the probabilistic origin of them. Applying the seven-in-a-row-above-or-below-the-mean rule, he faced particular cases where data points were too high. Abductive reasoning was used to explain the data running high: The sealing material from a particular batch turned out to be different—a special cause—but there was no reason to stop the process. The data points were not outside the control limits and the cause was found. If customers complained, the story would have been different, of course. Having sealing material leading to a slightly different air pressure can be seen as a constraint, and such a constraint can be framed as a reason that is dominant over others (in many situations one would try and bring the process back to the target line again).

Like hypothesis testing, SPC leaves room for two types of errors. The first is that non-conforming data points or trends are observed in the control chart whereas nothing is

wrong with the process. Purely by chance this happens regularly: Even if the probability of each individual unnatural pattern is smaller than, say, 0.05, the probability that at least one of about ten such rules finds a pattern is much bigger. The second type of error is that the data points do not show anything special whereas there is something wrong. Kevin gave us one example:

> My main failure at the moment is on estates [station wagons] on the left hand rear wheel arch. (…) In two of the cases the car was not out of spec but I was going round doing my checks to make sure there were no unusual leak paths and we found that they had missed some sealing. So although the car was in spec I raised an AP [Assigned Person, who is responsible for doing the investigation] and still got all the investigations done because there was an unusual leak path.

Thorough knowledge of the work process is required to interpret the implications of certain observations, for example how to measure and what to do about the problem. For example, certain leaks are allowed whereas others are not:

> You just have to go round [the car] and check [the air] is all coming out the normal places, like your door handles. Every door handle has got a massive air leak on, so you allow for that. At the bottom of your windows they let a lot of air out, but then when you start going to your wheel arches and underneath the car, there are certain places where maybe a plug is missing or you have a robot sealer skip.

Kevin knows that the most likely cause for a robot to skip a seal is when colleagues have switched it off and then back again; the robot then always skips a seal. If this is detected then a judgement has to be made whether the car should be sent back for an extra seal; this judgement is most likely made on both contextual and statistical grounds.

From such examples we were convinced that Kevin knew the process very well: He knew what to look for, what might cause it, what implications it has, what to do about the problem and so on. In terms of the space of reasons involved, he was aware of many reasons and conclusions, causes and effects that were linked by—we think—the right inferential relationships. However, when statistical issues were involved he felt less comfortable, for example when interpreting the control limits he and his colleagues received from a central office:

> What [the office] actually said to us is that it should have been 120 [weight per] cm squared +/- 3 sigma. It is very, very complicated because they gave us a lot of specs [actually control limits] to work on and it did not mean a lot to anybody in this company. We asked the questions to all the different people and no one could give us a definite, here's what you work off, this this this. So what we did, we went to [an SPC trainer] and said, "here is all our data for this year."

This quote also illustrates the *division of labour* and *knowledge* that is omnipresent in companies, and also a relevant feature of how people inhabit a space of reasons. Nobody knows everything that is relevant to producing a car. Each person is aware of a part, and his or her awareness is layered: Some reasons (in the wide sense) may be known well whereas others might be known in fragmented ways, only implicitly, or not at all.

Another example of the division of labour relates to the "assigned person" (AP) above. As soon as Kevin "raised an AP" he had done his job, and the problem was not his responsibility anymore. A third example is that Kevin and his colleagues fill in their charts, but know that it is the job of the SPC department (who are responsible for training and technical support) to calculate the mean and control limits from their data. This example illustrates that employees need not know all the statistical reasons behind SPC but do need to know something about the division of labour itself, and some of the statistical reasons to be able to communicate with others (team members, managers, suppliers) about their data and correctly fill in and interpret the control charts.

We were curious to what extent the SPC chart made sense to Kevin. When one of us (Res.) asked if the sideways histogram (to the right in Figure 1) helped him, he said:

> K.: It does and it doesn't because it just gives us an idea of where we are working. I mean that [histogram] is just little boxes to colour in for me [he laughs].
>
> Res.: It's not just supposed to be for little boxes to colour in. What could it tell you?
>
> K.: It tells you where about you are working. If you are running high or running low, but I concentrate on this [time series to the left] more because this tells me more than that [histogram]. That [histogram] will just give us an idea. Obviously you are supposed to get the peak in the middle behind the average line but that will tell us if we are running on average. I can look at the chart but to look at the histogram, I mean the obvious reason is that it will tell us if we are running just above the average or just below the average. (…) Really that should be in a nice spike right in the middle, right down the average line [he is probably describing the expected bell shape with the mode near the target line].

From such episodes we concluded that Kevin had a functional understanding of average (in relation to a target), variation (should be within certain limits) and distribution (roughly bell shaped) in relation to the mechanisms underlying the process. Such concepts are core in understanding and applying SPC. However, he kept calling control limits "spec limits," a phenomenon we have observed many times (Hoyles et al., 2007). As stated before, understanding control limits requires some understanding of standard deviations and the basics of the probabilistic rules in relation to the normal distribution. Crucially, despite their name, control limits are derived directly from the data, whereas spec limits are imposed externally. The trainers told us that this lack of understanding sometimes caused problems in communication between different groups in the company.

The second line in the quote above ("boxes to colour in") hints at a culture in which employees have to do things but not always know why. In fact, Kevin told us: "If you ask too many questions you end up doing a deep dive issue yourself, so really you are better just dropping a couple of slight hints and letting everybody else argue over it." The picture that emerged from such interviews is that employees tend to be aware of only that small part of the space of reasons that is directly relevant to their involvement in producing cars. However, to solve non-standard problems awareness of a larger part—including statistical reasons—is required, especially during nightshifts when few engineers or managers are around.

As another illustration of the importance of knowledge and how it is divided or distributed, we mention one finding from the interviews with three participants in an SPC course. We were surprised that the trainer asked participants to estimate standard deviations and calculate control limits by hand. Our impression was that they had only a limited idea what they were doing, and were actually hindered by the calculations rather than helped in their understanding. All participants we interviewed, however, appreciated having done the calculations once, just to know that these were done by the SPC department. Where our notion of understanding was focused on the statistical concepts involved in calculating the control limits, their take on understanding was *knowing about how labour and knowledge were divided.* Apart from being happy that these calculations were not part of their own work, they were also satisfied to note that the limits were not "conjured up" by management, but calculated on *their* data. In other words, we realised we had to enhance our notion of "understanding SPC" to a much wider notion in which the ways knowledge is distributed is taken into account. This implies that it is useful for

employees to know when statistical inferences are made by others, and who these others are.

4. REFLECTIONS

To characterise statistical inference at work—the goal of this paper—we compare SPC with a form of statistical inference that is more widely known in education, hypothesis testing (4.1), and characterise the space of reasons in which SPC takes place more generally (4.2).

4.1. A COMPARISON OF HYPOTHESIS TESTING AND SPC

As we show below, the logic of SPC theory resembles that of hypothesis testing in some ways, but also differs from it in others. We start with the similarities to identify some candidate characteristics of statistical inference more generally.

1. In both types of statistical inference, the construct of interest has to be *measured*. In the airtightness example, the construct was air pressure in the car when blowing in air at a certain volume per time unit; this pressure was used as a measure of airtightness. *Samples* are used to predict some feature of a population or process.
2. Both approaches aim to detect *differences*, for example between hypothetical (expected, targeted) and the real measures of the population or process. In SPC the key idea is to detect trends in processes such as shifting means before measurements exceed specification limits.
3. The equivalent of a *null hypothesis* in SPC is "the process is stable" (there is no "significant" difference between the targeted measures and the real). This means that there is only common-cause variation: The mean and variability do not change much. The *alternative hypothesis* would then be "there is a change in the process" with a special cause, leaving unspecified what this cause might be. In this sense SPC shows more similarities to Fisher's view on hypothesis testing than that of Pearson and Neyman, because Fisher did not require alternative hypotheses to be specified in advance (for the different views on hypothesis testing see for example Batanero, 2000; Biehler, 1982; Christensen, 2005).
4. In SPC, *probability-based rules* are used such as "seven points on either side of the mean may point to a special cause." The choice of 7 seems to be based on practical effectiveness rather than theoretical significance of the 1/64 probability.
5. Possible *errors* in SPC resemble type I and II errors: Non-conforming points might be due to chance, and special causes might still not be detected by probability-based rules. In the airtightness example, Kevin noted leaks even when the measurements did not give reasons to think so.

These five points illustrate that the theory of SPC has an inferential structure that is in some respects similar to that of hypothesis testing. When comparing with Makar and Rubin's (2007) three key characteristics of statistical inference, we can observe that all three are covered. Point 1 covers the issue of generalisation beyond the immediate data; data as evidence is a key point in both SPC and hypothesis testing; and points 4 and 5 hint at a probabilistic language.

The question arises whether the other similarities add anything specific or new. The first point on measurement is characteristic of statistical investigation in general, not only of statistical inference, and can therefore be excluded as a key feature of statistical inference. Points 2 and 3 are more interesting, because they point to something that is

relevant to statistical inference but not explicitly covered in Makar and Rubin's list: The comparison of data with a model. In hypothesis testing we mostly compare the data measures with those of a hypothetical distribution; in SPC we compare data measures with those of a targeted distribution. What is underlying this is the view that data can be seen as model plus residuals or signal plus noise (Konold & Pollatsek, 2002).

Apart from the aforementioned similarities, there are also differences:

1. In the commonly used logic of hypothesis testing, the goal is to reject the null hypothesis. In SPC, however, the null hypothesis (the process being stable) is the desirable situation.
2. Hypothesis testing is focused on a single measurement of a statistic that is compared with the sampling distribution of that statistic. SPC, however, has a time dimension that is not typical of hypothesis testing. In fact, SPC could be seen as involving a series of different tests. In loose terms, we can say that SPC is focused on generalisations about a *process* rather than a population, though technically it is of course possible to frame a process as generating a population of measurements, many of which are still to come.
3. The focus of SPC is on the need for action, and the goal is to monitor special causes rather than to estimate the probability of a conditional statement which is the end result of hypothesis testing.
4. Hypothesis testing is mostly a formalised form of inductive inference. But in SPC the crucial inference to be made is to detect the special cause, which is in fact a form of abductive inference (finding an explanation for an anomaly in the data), which cannot be formalised.
5. In hypothesis testing one temporarily suppresses contextual information, whereas with SPC employees constantly use their contextual knowledge of the process to interpret data points. There may be perfectly good reasons to let the data be off target. For example, changing the process might be too expensive and not necessary for quality. In hypothesis testing, context should only be attended to before and after carrying out the statistical test, but not during the test. This means that the use of contextual information is much more "disciplined" (a term introduced by Pratt at the SRTL-5 conference) in hypothesis testing than in SPC.

This illustrates how the two types of inference are subject to different norms. Hypothesis testing is supposed to be independent of specific features of the situation, and contextual "noise" is left out during the calculations, whereas SPC is *pragmatic*. Hypothesis testing has become standardised whereas SPC is used in loose ways and often in non-standard ways. For instance, we have observed SPC use that we had not anticipated based on the SPC literature; for example, the use of control charts particularly when processes were unstable, whereas the literature typically recommends using SPC for stable processes (e.g., Oakland, 2003). Of course, hypothesis testing is sometimes used in loose or non-standard ways too, for example if conditions do not apply or if p-values do not tell what people really want to know (cf., Abelson, 1995).

4.2. SPACE OF REASONS AT WORK

To further characterise statistical inference at work, we address the space of reasons involved, in particular in SPC. Let us mention a few features we think are relevant.

1. A space of reasons encompasses what can be analytically distinguished as *contextual* and *statistical* reasons. The key issue however is that a holistic view on such reasons does not prioritise any of them *a priori*. Contextual or statistical reasons are

prioritised depending on whatever is required to reach a goal, such as delivering cars that are airtight enough and not too airtight.

In workplace statistical inference *contextual* reasons can be put into the foreground where statistical reasons might point in a different direction. In the case of the sealing material leading to airtightness being off-target, contextual constraints, particularly cost of implementation balanced against resulting productivity gains, led Kevin and his colleagues not to re-centre the process to its target. Such constraints emphasise the significance of one reason over another and hence form the background (part of the space of reasons) constituting attentiveness to one concern over another and hence to the taking of certain actions rather than others.

2. A space of reasons includes reasons informing both *statements (claims, judgements, etc.)* and *actions (decisions etc.)*. Statistical tests in educational settings are mostly focused on testing the veracity of statements. Employees, as the SPC examples illustrate, are often more concerned with the right action. Of course, better knowledge can lead to better decisions, but their focus is on meeting targets and being (more) efficient and productive, and it is these goals that constitute the normative background in which one concern figures as dominant and demanding of attention over another.

3. A space of reasons can be analysed at both a collective and an individual level. At an individual level we can focus on the reasons, implications, causes, and effects that a person is responsive to. Being responsive to reasons does not entail full awareness but only that judgements are made within such a space of reasons. At the collective level, we can envision the space of reasons as being constitutive of the community, practice, activity system, or context in which the inferences are made.

4. A space of reasons is necessarily normative. Some forms of drawing conclusions are culturally more acceptable than others. If someone makes a statement, we expect him or her to believe it and to able to give evidence for it. In scientific research (whether or not using formal statistical inference) the norm of credibility or even truth is important. But in a company the most important norm might be quality of the products or services defined in relation to efficiency and appeal to customers. Such norms have a major impact on what counts as evidence, inference, or conclusion.

So what have we gained by focusing on spaces of reasons instead of context, for example? First of all, we think that by starting with a space of reasons, we can temporarily overcome the distinction between statistics and context that is so deeply engrained in our thinking. This is particularly beneficial in situations in which it is no longer clear what is statistical or contextual. For example, in some plants of the factory, SPC has become part of the shopfloor "context," and in ideal cases, employees seamlessly coordinate what can be analytically separated as statistical and contextual reasons.

Second, when studying statistical inference it makes sense to focus on reasoning, and by highlighting "reasons," which we use as short for premises, conclusions, evidence, motives, purpose, utility, our knowledge of causes and effects, and so forth, we bring something to the fore that might be ignored if we strictly interpret "context" as location or setting.

5. DISCUSSION

5.1. A FEW OBSERVATIONS

This paper's goal was to characterise statistical inference at work and point to a type of statistical inference typically neglected in school-based educational research. We did so by comparing statistical process control as a commonly used technique for process

improvement in industry with hypothesis testing. Next, we analysed the wider space of reasons in which statistical inference takes place. It turned out that there are commonalities that point to characteristics of statistical inference more generally. In addition to Makar and Rubin's (2007) three characteristics—generalisation, data as evidence, and probabilistic language—we also specified a fourth characteristic: comparing data with models, in particular, measures of the data with measures of a hypothetical or targeted distribution.

There are also differences. Firstly, SPC is pragmatic and focused on action. Secondly, there is a clearer place for abduction, whereas formal statistical inference tends to focus on induction. A third observation from our research in several companies is that inferences are mostly about *processes* rather than populations (see also Frick, 1998). In manufacturing sectors, data are mostly monitored to ensure that all items produced are within specification and to improve current production processes. When items are non-conforming or even fall out of specification, should the production process be stopped or is it more sensible to keep going? Stopping a car production line costs thousands of pounds per minute! In this setting, inferences are generally not based on formal statistical tests, but involve both statistical and contextual reasoning with a clear goal: maintaining the production process at a required accuracy or efficiency, or improving it.

The airtightness example emphasises *constraints of available resources* and the importance of the *division of labour and knowledge*. One advantage of using the notion of a space of reasons is that it helps to see human judgement as involving all such reasons including those that are beyond the more visible formal results of applying a statistical technique. Judging the consequences of having a batch of sealing material leading to a slightly different air pressure is one such example.

The adjective "informal" that is sometimes used in front of "statistical inference" does not seem to be suitable to characterise statistical inference at work. The theory of SPC is in fact formal to a certain extent: There are many books on the market on how to control production processes according to this theory. Hence, many people around the globe use SPC in somewhat similar ways. In that sense, SPC is more formal than, say, many intuitive forms of reasoning that young students display when they draw their first conclusions from data.

It could well be argued that what we emphasise as characteristics of the space of reasons in which statistical inference at work takes place applies to formal statistical inference as well. Indeed, we actually think that formal inferential techniques are used in such a way that we tend to forget that these techniques are used within a certain scientific space of reasons with particular norms and purposes, and that contextual knowledge is actually highly important in interpreting results of formal inference – despite the air of independence of contextual specificities that typically comes with significance tests. Moreover, we also face *division of labour* when using formal statistical inference: Even when we write research questions ourselves, collect the data ourselves, we use statistical software that others have programmed and most likely use statistical techniques that others have developed, before we arrive at our research conclusions. Hence we think that our analysis of statistical inference at work may shed light on statistical inference more generally.

5.2. LIMITATIONS

With the choice of SPC as a prototypical example of statistical inference commonly used in industry, we have restricted ourselves to one type of statistical inference. It is

perfectly possible that the analysis of other statistical techniques, perhaps in other workplaces, would lead to additional or other characteristics.

Another limitation of our research seems inherent to workplace studies: As outsiders it is extremely hard to gain access to companies: Time is money. Hence the companies we have studied form a convenience sample, and our time with employees, whether in interviews or training courses, was short compared to much research in school settings. This explains why we have not been able to do any workshadowing to the extent that we could observe employees actually solve problems by using SPC charts. Nor were we allowed to videotape employees or ask them to take tests to identify their level of statistical understanding. We were confined to stories told by managers, operators, and engineers, which meant that we had little opportunity to study inference in action and that we were not able to draw many generalised conclusions.

5.3. IMPLICATIONS FOR FUTURE RESEARCH AND EDUCATION

The aforementioned limitations naturally lead to implications for future research: It would be interesting to study more types of statistical inference in several workplaces, preferably in action. Another recommendation for future research is to characterise more generally what employees exactly need to know in different sectors of work. It is, for instance, hard to pinpoint to which level of formality, and in what sense, employees need to understand statistical concepts. Of course, they need to understand what the sources of variation are and what variation looks like in graphs (cf., Noss et al., 2007; Wild & Pfannkuch, 1999); they need to reason with a notion of distribution, mean versus target, spread, and measures of spread; they need to be able to interpret graphs, and so forth. Working with machines, employees need to know about causes and effects, but also how independent variables influence a dependent one. In short, employees need to know the key aspects of the *model* at stake, that is, the relationships between relevant variables and the causes and effects of changing those variables. The model need not be purely statistical or mathematical (Bakker, Hoyles, Kent, & Noss, 2006): Kevin's model of the variables of air volume and pressure were related to airtightness of cars was context specific. Yet we suggest adding the understanding of such a—possibly situated—model to Rubin et al.'s (2006) list of statistical ideas underpinning statistical inference.

Despite the advocated integration of statistical and contextual knowledge in practice, the statistics courses that we have observed and heard about in industry generally provide participants with little opportunity to connect what they have learned about statistics to their practice. Newly acquired statistical knowledge thus often stays separated from the rich contextual knowledge employees have of their work processes, instead of being perceived as an organic part of a space of reasons.

A more theoretical implication for research is the need to explore the consequences of framing educational research in terms of spaces of reasons and the tradition of inferentialism from which such philosophical constructs stem. They may provide us with a useful perspective on training in workplaces. Instead of primarily asking ourselves which statistical theory employees need to learn, we should perhaps ask, "How can employees become adept in their workplace's space of reasons? How can the reasons they attend to be enhanced by knowledge of quality improvement strategies?" In another paper we report on how we have tried this in our first design experiments (Hoyles et al., 2007).

Formulating potential implications of workplace research for school education is a tricky business. As Säljö (2003) notes, one should not make the mistake to try and copy workplace situations in school education. School is a different system with different goals than workplaces. Nor should we necessarily adapt our language: School and workplace

situations are constituted by very different spaces of reasons. The following implications therefore have to be interpreted only as a tentative list for discussion purposes:

1. Not only in workplace training but also in school settings, we should acknowledge the importance of context knowledge, real-world constraints, actions, and responsibility, and not confine the theory and learning activities to clean noise-free examples. What is required, also when teaching hypothesis testing, is to emphasise that drawing on context knowledge is disciplined. This might mean that courses that introduce formal statistical inference to students ideally also spend time afterwards on how formal statistical tests are actually used in practice, within a wider space of reasons.

2. We should pay attention to the mechanisms that cause variation, because then variation becomes easier for students to understand (cf., Wild & Pfannkuch, 1999). Yet there is a need to generalise and become familiar with statistical measures that can be applied in many other situations.

3. In our view, school-based educational research should pay attention to student understanding of processes in addition to populations because our anecdotal evidence suggests that many employees will deal with processes and not just with populations. The production of widgets at school (e.g., Konold & Lehrer, in press) might be a useful context to learn about many relevant statistical ideas.

ACKNOWLEDGEMENTS

This paper is based on an earlier version presented at the fifth research forum on Statistical Reasoning, Thinking, and Literacy (SRTL-5), 11-17 August 2007, Warwick, UK. We thank all participants for fruitful discussions about informal statistical inference, and Iddo Gal, Cliff Konold, and Rolf Biehler for their feedback on earlier versions of this paper. Kees Klaassen and Bart Ormel have also made very helpful suggestions. We further gratefully acknowledge funding by the UK Economic and Social Research Council's Teaching and Learning Research Programme, Grant number L139-25-0119.

REFERENCES

Abelson, R. P. (1995). *Statistics as principled argument.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bakker, A., Derry, J., & Konold, C. (2006). Technology to support diagrammatic reasoning about center and variation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D4_BAKK.pdf]

Bakker, A., Hoyles, C., Kent, P., & Noss, R. (2006). Improving work processes by making the invisible visible. *Journal of Education and Work, 19*, 343-361.

Batanero, C. (2000). Controversies around the role of statistical tests in experimental research. *Mathematical Thinking and Learning, 2*, 75-97

Ben-Zvi, D. (2006). Scaffolding students' informal inference and argumentation. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics*, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D1_BENZ.pdf].

Biehler, R. (1982). *Explorative Datenanalyse – Eine Untersuchung aus der Perspective einer deskriptiv-empirischen Wissenschaftstheorie* [Exploratory data analysis - An

investigation from the perspective of a descriptive empirical scientific theory] Bielefeld, Germany: Universität Bielefeld.

Caulcutt, R. (1995). The rights and wrongs of control charts. *Applied Statistics, 44*, 279-288.

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*, 121-126.

Cobb, G. W. (1999). Discussion of "Let's use CQI in our statistics programs." *The American Statistician, 53*, 16-21.

Frick, R. W. (1998). Interpreting statistical testing: Process and propensity, not population and random sampling. *Behavior Research Methods, Instruments, and Computers, 30*, 527-535.

Hoyles, C., Bakker, A., Kent, P., & Noss, R. (2007). Attributing meanings to representations of data: The case of statistical process control. *Mathematical Thinking and Learning, 9*, 331-360.

Konold, C., & Lehrer, R. (in press). Technology and mathematics education: An essay in honor of Jim Kaput. In L. D. English (Ed.). *Handbook of international research in mathematics education* (2nd ed.). New York: Routledge.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education, 33,* 259-289.

Lave, J. (1988). *Cognition in practice: Mind, mathematics and culture in everyday life*. Cambridge, UK: Cambridge University Press.

McDowell, J. (1996). *Mind and world* (2nd ed.). Cambridge, MA: Harvard University Press.

Makar, K., & Rubin, A. (2007, August). *Beyond the bar graph: Teaching informal statistical inference in primary school.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Noss, R., Bakker, A., Hoyles, C., & Kent, P. (2007). Situating graphs as workplace knowledge. *Educational Studies in Mathematics, 65,* 367-384.

Noss, R., & Hoyles, C. (1996). *Windows on mathematical meanings: Learning cultures and computers.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Noss, R., Pozzi, S., & Hoyles, C. (1999). Touching epistemologies: Meanings of average and variation in nursing practice. *Educational Studies in Mathematics, 40,* 25-51.

Nunes, T. A., Schliemann, D., & Carraher, D. W. (1993). *Street mathematics and school mathematics*. Cambridge, UK: Cambridge University Press.

Oakland, J. S. (2003). *Statistical process control* (5th ed.). Amsterdam: Butterworth-Heinemann.

Paparistodemou, E., & Meletiou-Mavrotheris, M. (2007, August). *Enhancing reasoning about statistical inference in 8 year-old students.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

Pfannkuch, M. (2006). Informal inferential reasoning. In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International Conference on Teaching Statistics,* Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/6A2_PFAN.pdf]

Rubin, A., Hammerman, J. K. L., & Konold, C. (2006). Exploring informal inference with interactive visualization software In A. Rossman & B. Chance (Eds.), *Working cooperatively in statistics education: Proceedings of the Seventh International*

Conference on Teaching Statistics, Salvador, Brazil. [CDROM]. Voorburg, The Netherlands: International Statistical Institute.
[Online: http://www.stat.auckland.ac.nz/~iase/publications/17/2D3_RUBI.pdf]

Säljö, R. (2003). Epilogue: From transfer to boundary-crossing. In T. Tuomi-Gröhn & Y. Engeström (Eds.), *Between school and work: New perspectives on transfer and boundary-crossing* (pp. 311-321). Amsterdam: Elsevier.

Smith, J. P. (1999). Tracking the mathematics of automobile production: Are schools failing to prepare students for work? *American Educational Research Journal, 36,* 835-878.

Steen, L. A. (2003). Data, shapes, symbols: Achieving balance in school mathematics. In B. L. Madison & L. A. Steen (Eds.), *Quantitative literacy: Why literacy matters for schools and colleges* (pp. 53-74). Washington, DC: The Mathematical Assiociation of America. Retrieved October 27, 2007 from http://www.maa.org/ql/qltoc.html

Techno-mathematical Literacies Project (2003-2007). www.lkl.ac.uk/technomaths

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistics Review, 67,* 223-248.

Zieffler, A., Garfield, J., & delMas, R. (2007, August). *Studying the development of college students' informal reasoning about statistical inference.* Paper presented at the Fifth International Research Forum on Statistical Reasoning, Thinking and Literacy (SRTL-5), University of Warwick, UK.

ARTHUR BAKKER
University of London
London Knowledge Lab
23-29 Emerald Street
London WC1N 3QS
United Kingdom

Currently working at:
Utrecht University
Freudenthal Institute
PO Box 9432
3506 GK Utrecht
The Netherlands