oxera
compelling economics

Agenda—10 years

Advancing economics in business

Dealing with doping: a question of the benchmark

Originally published in July 2008. 2015 commentary by Oxera

Regulators, competition authorities, company boards and courts are increasingly making decisions based on statistical analysis of a 'panel' of both cross-sectional data (e.g. data for different companies) and time-series data (i.e. data over time). This article explores the use of panel data in the context of testing for doping in the Tour de France. Since the article's publication, Oxera has been involved in numerous cases in which panel data techniques have been used. The statistical techniques continue to evolve, and there is still a healthy debate today around the appropriateness of the different techniques in different contexts.

After a number of years of the Tour de France being marred by doping scandals, this year's doping incidents suggest that the stakes involved are so high that for some it continues to be worthwhile to illegally boost their performance, despite the risks involved. What is the case for and against a more accurate drug-testing regime, and how can this be related to performance comparisons in general?

How can we be sure that our sporting heroes have earned their status through hard work rather than through cheating? To date three riders participating in this year's Tour de France have been sent home after testing positive for the blood-booster erythropoietin (EPO). Subject to confirmation of the results, they may face an extended ban from competition. There may well be further controversies at the Beijing Olympics, and indeed British athlete Dwain Chambers has recently failed in his attempt before the High Court in the UK to overturn a ban on participating in the Olympics for previously failing a drugs test.

At first sight there are significant deterrents to doping, including the reputational and financial costs that being found guilty may imply, as well as the potential adverse health effects. The 2007 Tour de France winner, Alberto Contador, was unable to participate in this year's Tour as a result of Astana, his new team, not being invited following its involvement in last year's doping scandal. Despite the concerted efforts of the organisers to deter riders from taking illegal performance-enhancing drugs through strict sanctions, riders are still being caught, which suggests that the incentive to cheat is significant.

Given the significant damage doping has had on various sporting events, what are organisers doing to ensure that athletes compete on a level playing field?

In addition to the threat (and implementation) of tough sanctions, an effective drug-testing regime is required to monitor athletes' performance. The developments in this area have implications for making performance comparisons in general, as discussed below. First, however, this article explores developments in the tools used in drug-testing.

Since this article was published, a series of investigations have examined the culture of doping in cycling. In 2012 the US Anti-Doping Agency (USADA) published its 'Reasoned Decision' following its investigation into doping on the US Postal Service team.[1] The publication revealed the organised and systematic nature of doping by riders on the team —including seven-times Tour de France winner, Lance Armstrong—and the code of silence that existed to protect those who cheated. Lance Armstrong went on to admit doping and his role in facilitating it on US television in January 2013.

Following the USADA decision, the cycling governing body, the UCI (Union Cycliste Internationale), commissioned its own independent investigation into the causes of the pattern of doping within cycling. The Cycling Independent Reform Commission (CIRC) subsequently reported in March 2015.[2] The introduction of the Athlete Biological Passport (ABP) in 2008 was one of the first steps taken by the UCI in changing the behaviour of elite road cyclists, confirming the use of statistical analysis described in this article.

[1] USADA (2012), 'Report on proceedings under the world anti-doping code and the USADA protocol: Reasoned Decision of the United States Anti-Doping Agency on disqualification and ineligibility', 10 October.
[2] Cycling Independent Reform Commission (2015), 'Report to the President of the Union Cycliste Internationale', 9 March.

Does a failed drug test always mean a cheating athlete?

Drug tests can be broadly divided into two categories:

- the detection of any level of a particular substance;
- tests involving set cut-offs and thresholds for the level of a particular substance.

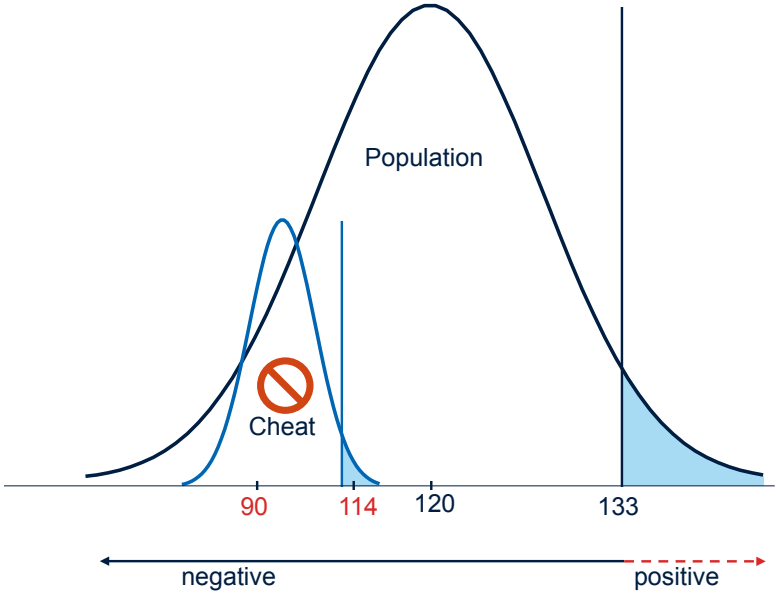
The former are less difficult to interpret as only the presence or absence of a substance are relevant for detecting cheating.[1] However, tests involving choices of certain critical values, above or below which an athlete is classified as positively testing for a drug, can be more controversial.

A single test is rarely taken as conclusive evidence for doping, and a 'B-sample' (a second test) is often involved. Yet even two independent tests of an individual may not be sufficient to conclusively prove the presence or absence of illegal substances. Given the significant consequences of wrongly testing either positive—possibly the end of an athlete's career and the reputational and financial losses associated with this—or negative—allowing the athlete an unfair competitive advantage—it is crucial to minimise the risk of getting a drug test wrong.

Traditional anti-doping tests rely on cross-sectional comparisons—i.e. a comparison between an athlete's test results at a given point in time compared with a threshold value above which the athlete is thought to have artificially improved their performance. One measure that has been proposed in testing for EPO is the stimulation index—the indexed relationship between haemoglobin and reticulocyte. This index exhibits a significantly greater variation between athletes than a given 'clean' athlete's repeated test results over time. For example, a stimulation index exceeding a score of 133 is the International Cycling Union (UCI) limit above which EPO doping is suspected. This is over five standard deviations (a measure of the dispersion of the data) above the average of the general population. The reason for this high value compared with the average is that, on any given day, an individual's values may be two standard deviations away from their long-term average value. The high variance of results between individuals added to the variation of each individual's results means that the benchmark level for a positive test has to be set high to avoid the possibility of accusing an athlete of doping when they happen to have a naturally occurring high stimulation index score.

Figure 1 depicts the distribution of stimulation index values based on a representative sample of the population of cyclists.[2] The average value is 120, and a score above 133 can be interpreted as testing positive. However, using traditional cross-sectional testing methods, an athlete with a test score of 114, for example, would not test positive and would be deemed to be competing on a level playing field with other cyclists, even if, as illustrated in Figure 1, they have really cheated.

Figure 1 Example of a doping athlete benefiting from high test thresholds



Source: Oxera.

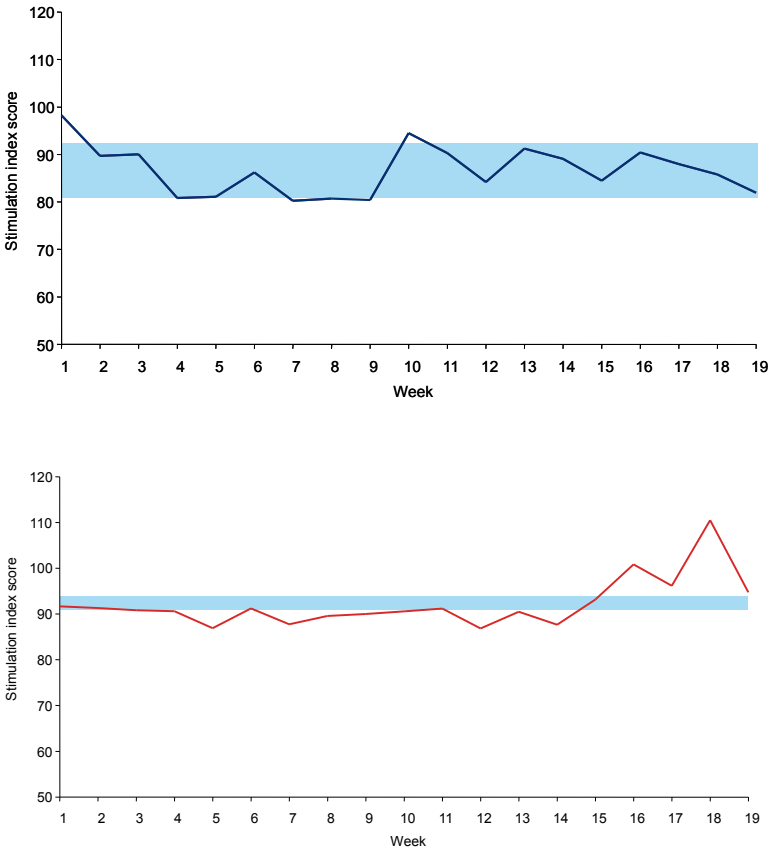
Each individual cyclist, however, has a natural variation around their mean score over time. The accuracy of the test may therefore be improved with reference to longitudinal data of an individual's score over time.

If such data were available, it might show that the athlete's natural level is 90, for example, and that the standard deviation over time around this average is 8. This result would suggest that a score of 114 is highly likely to be caused by some external event—there would be a less than 1 in 100 chance that such a test score is due to natural variation. Equipped with this additional information, the medic may suspect that the athlete is using prohibited substances.

Figure 2 overleaf further illustrates how data collected at regular intervals may make measurement more accurate and thus help to increase the chances of identifying doping offenders. The first graph shows a 'clean' athlete with a naturally high variation in stimulation scores over time. The confidence intervals give little reason to suspect the athlete of cheating—the upper and lower confidence interval bounds based on the athlete's own historical variation encompass most of the time series. In contrast, the athlete represented by the second graph clearly shows unusually high values towards the end of the stimulation index relative to historical values. Even though the value is well below the critical value of 133 for the population overall, there is a high statistical probability that the athlete is in fact using prohibited substances.

This simplified representation of the use of data over time represents a significant departure from the traditional testing regime. The systematic testing of individuals and recording the natural variation in their results has the advantage of allowing the earlier detection of doping, and the detection

Figure 2 Using time-series data to detect a cheat: a flat or mountain stage?



Source: Oxera.

of smaller quantities, and thus provides greater confidence than traditional doping tests. Illegal performance enhancers are becoming increasingly difficult to detect, so physiological parameters rather than substances themselves are monitored to detect doping. Provided that the substance acts through biological parameters on which information is collected, this method may also help in detecting new substances. This approach has been advocated by the Agency for Sports Ethics.[3]

One of the prerequisites for using this approach is the regular collection of consistent data. The ongoing disagreements between the UCI and the Tour de France organiser, Amaury Sport Organisation, over unresolved issues relating to doping have meant that the 2008 Tour is outside the jurisdiction of the UCI, and is instead run by the French Cycling Federation. But why does it matter who performs the tests? The UCI has been building up biological passports—a database of each rider's physiology to make drug-testing more accurate. However, it has been reported that the French Anti-doping Agency (FADA) has not had access to the biological passport data and has had to rush to collect pre-race tests on the expected favourites.[4] This suggests that there may be significant political and institutional, as well as practical, barriers to the collection of high-quality data over time.

Traditional drug tests looked for illegal substances in blood or urine by comparing their levels with those of control subjects. However, the natural variation in the chemicals in human blood makes it hard to differentiate between a high reading due to natural variation and a high reading due to doping.

As this basic form of testing became more prevalent in sport, new drugs and methods were developed—although it can take several years for effective detection methods to be designed.

The ABP was developed to establish whether an athlete was manipulating their blood, regardless of the substance or method used. 2008, when this article was written, was the first year in which the ABP was used to supplement traditional drug testing. Taking advantage of the fact that human physiology stays broadly constant over time:

> each athlete should become his own reference, meaning that individual limits should be applied instead of population limits, and one could use the athlete's previous measurements as basal levels.[1]

CIRC identified the introduction of the ABP as bringing about a major change in the sport. Prior to the ABP, only three riders had been convicted of blood doping. In the first three years of the ABP 26 riders tested positive for the presence of EPO, and in the vast majority of cases it was the abnormal blood profile which led the authorities to conduct more targeted anti-doping testing for specific substances.[2]

[1] Zorzoli, M. (2011), 'Biological passport parameters', *Journal of Human Sport and Exercise*, **6**:2, pp. 205–17.
[2] Zorzoli, M. and Rossi, F. (2012), 'Case studies on ESA-doping as revealed by the Biological Passport', *Drug Testing and Analysis*, **4**:11, pp. 854–8.

While appealing from a statistical perspective, using longitudinal data may give athletes a slightly different incentive to artificially in crease their average and the variation in their results, so as to reduce the risk of being caught doping in the future. One way around this might be to make comparisons with a control group of individuals who do not have an incentive to artificially change their test results (i.e. non-athletes).

What have drug-testing in sports and regulatory target-setting in common?

The above developments in the sporting world—where athletes seek to outperform one another—have implications for measuring performance in the commercial world. In these settings, those undertaking performance assessments often conduct statistical and other analyses of business

units' performance, and assess the extent to which laggard units need to improve in order to catch up to an identified benchmark. But how confident can the assessors be in identifying benchmark performance and in identifying the 'true' winner?

In the same way that the UCI and FADA have to err on the side of caution when setting the benchmark and deal with inaccuracies inherent in the traditional testing system, when comparing firms' or business units' performance, assessors need to take into account the uncertainty around the benchmark arising from natural variation in the data, which may not be due to inefficiency. Just as the indicators for athletes have a natural variance over time, the data recorded by companies' accounting systems is often subject to measurement error, different interpretations of accounting rules and rounding errors, all of which add noise to the data-generating process. This makes it more difficult for those undertaking performance assessments to distinguish genuine differences in performance from the natural variation in reporting over time.

Performance comparisons and benchmarking are often used to set targets to incentivise business units to improve their productivity; however, there is a trade-off between the strength of the incentive and the potential for restricting a company or business unit from investing and innovating by setting too harsh a target. In the regulatory setting, for example, where benchmarking is often used, there is the risk of a legal challenge or competition authority referral for not allowing the firm sufficient revenues such that their ability to finance their functions is affected.

One way to deal with this trade-off is to increase the certainty over where the benchmark lies. Undertaking analysis that not only looks at performance across firms or business units but also at what has been achieved historically, and the variation around historical performance, can improve the confidence and accuracy of the analysis setting. Targets that acknowledge that there may be some variation (induced by natural variation or measurement error) are likely to be more robust and thus credible.

One example where regulators have used data over time to monitor the performance of a company is the UK Office of Rail Regulation's Network Rail Monitor, which reports key performance indicators for the network infrastructure operator on a quarterly basis, allowing for early identification of significant changes in performance while also acting as an incentive for the operator to maintain and improve its performance.

Analysis of longitudinal data: a panacea?

While on the face of it using data over time as well as cross-sectionally improves the accuracy of the estimate and gives extra information about historical patterns in performance, there could be downsides. Frequent blood tests for athletes add to agency costs and may be disruptive for the athlete, while annual (or even quarterly) reporting by firms or business units potentially introduces data consistency issues for efficiency assessments (e.g. changes in accounting practice). However, the benefits of increased confidence and mapping of performance gains over time could outweigh the costs.

This may be the case particularly where the implications of getting the wrong answer—whether in testing athletes for drugs, or testing companies' efficiency performance—could be significant. Many sports have been tainted by controversy from doping scandals, inflicting significant damage on their credibility and integrity. The use of longitudinal data may improve accuracy in testing, allowing the governing bodies to set more accurate targets, increasing the likelihood of getting caught, and thereby reducing the incentive to cheat. After a number of years of the Tour de France being marred by doping scandals, this year's incidents suggest that the stakes involved are so high that for some athletes it continues to be worthwhile to illegally boost their performance, despite the risks. While the use of longitudinal data may come at a cost to athletes and the governing bodies, an even tougher drug-testing regime could help reduce the incentives to use performance-enhancing drugs.

Both Ofwat (the economic regulator of the water industry in England and Wales) and Ofgem (the energy regulator for Great Britain) now make use of panel data (i.e. data both across companies and over time) to set performance targets for the companies they regulate.

Following feedback from the UK Competition Commission (now the Competition and Markets Authority) for the 2014 periodic review, Ofwat adopted a panel data approach to modelling companies' costs. Ofgem also saw the benefits of using panel data in the context of setting targets for gas distribution networks (GDNs) (RIIO-GD1):

> we have used a panel data approach, which makes better use of the data by considering the information provided by each year of data, rather than the information provided by the average alone. Such approach increases the degrees of freedom of the model and hence the robustness of the estimates. Given the small number of comparators in our sample (eight GDNs) any improvement in the model's degrees of freedom is important for the accuracy of the estimates. Finally, our panel approach isolates year-specific effects rather than estimating a single intercept.[1]

We are also seeing panel data techniques being used to make decisions in contentious situations such as those in courts and boardrooms.

[1] Ofgem (2012), 'RIIO-GD1: Final Proposals – Supporting document – Cost efficiency', 168/12, 17 December.

[1] Green, G.A. (2006), 'Doping Control for the Team Physician: A Review of Drug Testing Procedures in Sport', *American Journal of Sports Medicine*, **34**.

[2] Numerical example based on http://web.archive.org/web/20080708183151/http://www.agencyforsportsethics.org/Programs.html.

[3] Agency for Sports Ethics (http://web.archive.org/web/20080708183151/http://www.agencyforsportsethics.org/Programs.html).

[4] *New York Times* (2008), 'Tour de France Preview', 5 July.