OPINION
published: 20 August 2018
doi: 10.3389/fsoc.2018.00025

Statistical Adhockeries Are No Criteria for Legal Decisions—The Case of the Expert Medical Report on the Assessment of Urine Specimens Collected Among Athletes Having Participated to the Vancouver and Sochi Winter Olympic Games

Franco Taroni [1]*, *Alex Biedermann* [1], *Joëlle Vuille* [1] *and Silvia Bozza* [2,1]

[1] School of Criminal Justice, University of Lausanne, Lausanne, Switzerland, [2] Department of Economics, University Ca' Foscari of Venice, Venice, Italy

Keywords: legal decision making, doping in sports, statistics, forensic interpretation, Olympic Games

OPEN ACCESS

Edited by:
Athanasios Alexiou,
Novel Global Community Educational Foundation (NGCEF), Hebersham, Australia

Reviewed by:
Donald Arthur Berry,
University of Texas MD Anderson Cancer Center, United States

****Correspondence:***
Franco Taroni
Franco.Taroni@unil.ch

Specialty section:
This article was submitted to ELSI in Science and Genetics, a section of the journal Frontiers in Sociology

Received: *25 May 2018*
Accepted: *30 July 2018*
Published: *20 August 2018*

Citation:
Taroni F, Biedermann A, Vuille J and Bozza S (2018) Statistical Adhockeries Are No Criteria for Legal Decisions—The Case of the Expert Medical Report on the Assessment of Urine Specimens Collected Among Athletes Having Participated to the Vancouver and Sochi Winter Olympic Games. Front. Sociol. 3:25. doi: 10.3389/fsoc.2018.00025

Scientific literature and practice, notably expert reports, commonly involve misinterpretations of standard statistics, such as the p-value, or the calculation of so-called "3-standard deviation intervals," elements upon which decisions in medicine, physics, or legal matters are based[1]. Such instances of expert reporting reflect a misreading of the way in which scientists should assist the judiciary in assessing results coming from analytical laboratories. A recent example of such a practice are the conclusions of the international report on urine specimens collected among athletes participating in the Vancouver and Sochi Winter Olympic Games (report dated 5th October, 2017)[2,3].

Uncertainty is a complication that accompanies participants of the justice system who face inference and decision-making as core aspects of their activities. Inference relates to the use of incomplete information, as given by scientific findings, in order to reason about propositions of interest, such as whether the quantity of a given substance in some bodily fluid is larger than a legal threshold. In turn, decision-makers, notably judges, are required to make practical decisions, such as declaring whether or not an athlete has used a performance-enhancing substance. Inference and decisions of this kind abound in the legal field. Toxicology laboratories, across jurisdictional systems, are regularly asked to quantify the amount of target substances (e.g., alcohol, illegal drugs, biological markers) detected in, for example, blood samples taken from persons of interest (e.g., Karkazis and Jordan-Young, 2015).

Inference and decision require logical assistance because unaided human reasoning is liable to bias and misinterpretation. These represent causes of concern because fallacious reasoning and erroneous conclusions in legal proceedings risk endangering the fairness of the proceedings and can lead to miscarriages of justice. Statistical approaches are often used to support expert conclusions

[1] Hence the use of "adhockeries" in the title. The term adhockery was introduced by I. J. Good, and used also by de Finetti (e.g., de Finetti, 1993a,b), to denote the use of improvised measures rather than a robust and logic methodology.

[2] Report available at https://stillmed.olympic.org/media/Document%20Library/OlympicOrg/IOC/Who-We-Are/Commissions/Disciplinary-Commission/IOC-DC-Schmid/Appendix-VIII-CHUV-Report-Prof-Burnier-06-10-2017.pdf. Hereafter, the Report.

[3] For a previous example of the use of statistics in a case of alleged doping see, e.g., the Andrus Veerpalu case (Fischer and Berry, 2014)

but inferential misunderstandings regrettably plague disciplines such as forensic science and medicine when scientists report on statistical analyses conducted as part of their casework or research.

The case on which we intend to comment can be briefly summarized as follows. The International Olympic Committee requested statistical analyses on results of urine examinations performed on samples coming from the XXII Olympic Winter Games in Sochi, with the aim of identifying athletes who had used prohibited substances. Specifically, the question was "[t]o determine [. . .] if the values are within the reference values obtained from the control population at the XXII Olympic Winter Games and in agreement with data published." (Report at p. 3). A potentially doped athlete, called in the Report "true outlier," was defined as a person having a given bio-chemical parameter value—for example, urinary sodium concentration—greater than a reference mean plus three standard deviations. The reference mean and standard deviation were calculated using data from reference athletes, considered *not* doped, of the Vancouver Olympic Winter Games.

As an illustration, consider a measured target substance (e.g., urinary sodium concentration), where the experimental unit is a urine sample from a person under investigation. The mean values reported for the sodium concentration in urines in the reference male (female) population of athletes are 95.4 (67.39) mmol/l. The reported standard deviation values are 49.37 and 40.88 mmol/l, respectively (see Report at p. 6). According to the "three standard deviation rule," athletes with values greater than 243.51 (or 190, for women) are considered to be outliers. Thus, the measurement would be said to meet the requirement for establishing the presence of an unrealistic level of a target substance in urine *if* the measurement for the investigated sample were larger than the upper value of the bound in a reference population. It was noted in the report that "[w]ith this approach, we identify 13 samples (of 5 men and 8 women) which are definitively out of the range." (Report at p. 2).

As a preliminary, it is worth noting that such an approach for the treatment and reporting of experimental results does not address the inferential and decisional issues at stake. Instead, it is merely descriptive. This does not mean, though, that it is intrinsically wrong: scientists widely rely on effective descriptive methods of exploratory data analysis to illustrate, for example, how population data are distributed, where given sample data are located and how they spread. However, such a description does not fully address the questions of interest for the decision-maker, which are: *How can we use data (or a summary of them) on the Vancouver Olympic athletes to infer something about the value of the urinary sodium concentration in the reference population?* and: *How can we conclude that a new measurement from a given athlete is in fact an outlier (or an anomalous value) with reference to this population?* These are intrinsically inferential questions, not descriptive ones, and remain unresolved with the approach taken in the report, as we explain below.

It is commonly understood, and unquestioned, that measurements on urinary samples taken from individuals of a given population will show some variation. Stated otherwise, the results will, in some sense, distribute. Basic statistics such as the mean and the standard deviation of a quantity of interest (e.g., the sodium concentration) used to describe the reference samples from the Vancouver Olympic athletes represent indeed succinct and informative summaries. The mean provides a measure of location and the standard deviation provides a measure of dispersion (spread) of the available measurements. In this context, the "three standard deviation rule" may have some appeal. Provided that data distribute symmetrically around the mean, at least approximately, then values within one standard deviation of the mean account for about 68% of the observations, while two standard deviations account for about 95% and three standard deviations account for about 99.7% of the values. The 68-95-99.7 rule is a shorthand used to remember the approximate percentage of values that lie within a band around the mean with a width of two, four and six standard deviations, respectively. It is a rule to describe the available data (i.e., measurements from Vancouver athletes), but not to infer something about a new value coming from a new athlete, as emphasized also in Berry (2008).

However, does this rule allow one to conclude that values outside this range are necessarily "outliers"—lying at an abnormal distance from other values? Obviously, *any* set of observations contains extremes: the minimum and the maximum value are extremes. Notwithstanding, it is understandable to express concerns in situations, such as the case discussed here, where the highlighted extremes are not only the largest (or, in other cases, the smallest) observation, but are actually "*extremely* extreme": they are apparently inconsistent with the reference observations and therefore candidates for being considered "outliers." It is no accident to term these values "candidates." Several reasons can, in isolation or combined, account for extreme observations: first, natural variation, beyond the currently known bands, but also laboratory measurement or recording errors, or even intentional tampering, such as the addition of a target substance (here salt). Since these are potential accounts for the observations, it is—by definition—a matter of a *personal* judgment on the part of the scientist to decide *when* a given observation appears to be inconsistent with the remainder set of data. One way to avoid a rigid and intrinsically arbitrary threshold is to consider at least one explicit alternative account for the findings: the scientist could then provide a statistical measure called a "likelihood ratio" that represents an expression of how the measurements, whatever their value, extreme or otherwise, are capable of discriminating amongst competing propositions of interest. When no discernible alternative hypothesis can be specified (as in the case of interest) several ways of categorizing suspicious observations are available (see, e.g., Barnett and Lewis, 1994).

It is common to distinguish between frequentist (or classical) and Bayesian approaches. Statistical data analyses in the forensic and medical contexts commonly rely on a so-called "frequentist" perspective, associated with the idea that statistical conclusions could be entirely objective, with known error rates. Consider, for instance, the problem of hypothesis testing, where attempts at drawing conclusions about competing propositions often rely on a comparison between the significance level of the test and the observed significance level, i.e., *p*-value. A large majority of papers published nowadays still propose statistical treatments based on this quantity. Controversial discussion was initiated by an editorial of Basic and Applied Social Psychology (Trafimow and Marks, 2015), expressing the intention to ban from publication in their journal any paper containing procedures advocating *p*-values. This announcement has echoed widely, from general weekly science journals (e.g., Nuzzo, 2014; Leek and Peng, 2015) to specialist groups such as the International Society for Bayesian Analysis (Schmidt et al., 2015). The main concern expressed in these reactions is not the correctness or usefulness of frequentist statistical procedures, but rather the misinterpretations surrounding the use of such procedures and their consequences. There is a need to emphasize what exactly the various approaches allow scientists to draw as a conclusion, and what they do not allow them to say.

One of the major misunderstandings found in the reporting on significance testing through a *p*-value consists in interpreting this value as the probability that the null hypothesis (e.g., as previously stated a difference between populations mean values) is true. This fallacious conclusion is also known as the fallacy of the transposed conditional. The temptation to believe that, if an observation is rare under a given hypothesis it can be regarded as evidence against that hypothesis, must be resisted[4]. Bayesian approaches avoid these intricacies by relying on the fundamental tenet of capturing, using probability, all uncertainties characterizing a problem. According to these approaches, discordancy can be assessed by means of a predictive probability to observe a value greater than the particular (suspicious) observation given the rest of the reference sample, which allows one to restrict attention to manifestly extreme (unlikely) observations (Geisser, 1998).

Despite struggles over philosophical stances regarding statistical inference and decision-making, the restriction of attention to the sole question of outliers still falls short of the fundamental problem that the case in question poses. Among the ultimately disputed questions is the issue of whether there is sufficient evidence to conclude that a given urine value is an outlier. The answer to this question cannot rely on scientific findings only, because it requires the assessment of all available information, scientific, and other, in a given case. What is more, it cannot be reduced to a descriptive (statistical) account of scientific findings, but extends to inference and decision-making, and associated decision criteria. The latter are not given by *ad-hoc* statistical thresholds, but are intimately related to the decision-maker's preferences and policy values, which are even further beyond the scientist's area of competence.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

AB acknowledges the support provided by the Swiss National Science Foundation through Grant No. BSSGI0_155809 and the University of Michigan Law School, Ann Arbor, MI (Michigan Grotius Research Fellowship).

[4] "Researchers often rely on the seeming objectivity of the $p<0.05$ criterion without realizing that theory behind the p-value is invalidated when analysis is contingent on data." (Gelman and Hennig, 2017).

REFERENCES

Barnett, V., and Lewis, T. (1994). *Outliers in Statistical Data, ed*. 3. Chichester: Wiley.

Berry, D. (2008). The science of doping. *Nature* 454, 692–693. doi: 10.1038/454692a

de Finetti, B. (1993a). "The role of probability in the different attitudes of scientific thinking," in *Probabilità e induzione – Induction and probability*, eds P. Monari and D. Cocchi (Bologna: Editrice Clueb), 491–511.

de Finetti, B. (1993b). "Bayesian statistical inference" in *Probabilità e Induzione – Induction and Probability*, eds P. Monari and D. Cocchi (Bologna: Editrice Clueb), 513–524.

Fischer, K., and Berry, D. A. (2014). Statisticians introduce science to international doping agency: the Andrus Veerpalu case. *Chance* 27, 10–16. doi: 10.1080/09332480.2014.965625

Geisser, S. (1998). "Some uses of order statistics in Bayesian analysis," in *Handbook of Statistics 17 - Order Statistics*, eds N. Balakrishnan and C.R. Rao (Amsterdam: Elsevier), 379–399.

Gelman, A., and Hennig, C. (2017). Beyond subjective and objective in statistics. *J. R. Stat. Soc. A* 180, 967–1033. doi: 10.1111/rssa.12276

Karkazis, K., and Jordan-Young, R. (2015). Debating a testosterone 'sex gap'. *Science* 348, 858–860. doi: 10.1126/science.aab1057

Leek, J. T., and Peng, R. D. (2015). Statistics: *P*-values are just the tip of the iceberg. *Nature* 520, 612. doi: 10.1038/520612a

Nuzzo, R. (2014). Scientific method: statistical errors. *Nature* 506, 150–152. doi: 10.1038/506150a

Schmidt, A., Berger, J., David, P., Kadane, J., O'Hagan, T., and Pericchi, L. (2015). Banning null hypothesis significance testing. *ISBA Bull.* 22, 5–9. Available Online at: https://bayesian.org/wp-content/uploads/2016/09/1503.pdf

Trafimow, D., and Marks, M. (2015). Editorial. *Basic Appl. Soc. Psychol.* 37, 1–2. doi: 10.1080/01973533.2015.1012991

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.