LIARS, DAMNED LIARS AND STATISTICIANS

Arthur Fries
Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311
afries@ida.org

ABSTRACT

Nowhere is the need for insightful and accurate statistics more profound than in the arena of national policy making, yet statisticians routinely are excluded from the table during public policy debates. Why? Drawn from personal encounters, this paper presents examples of recent questionable "statistical analyses" – originally intended to support national-level decision-making within the U.S. Department of Defense, but instead having the effect of further perpetuating the widespread impression that the statistical community should not be entrusted with such a critical role.

INTRODUCTION

The all too familiar refrain goes: "There are three kinds of lies: lies, damned lies, and statistics." But it is not the "statistics" per se that ever are truly at fault, it is the individuals that produce, describe, publish and disseminate them. Nowhere is the need for insightful and accurate statistics more profound than in the arena of national policy making, yet statisticians routinely are excluded from the table during public policy debates.[1] Why?

Drawn from personal encounters, this paper documents examples of recent questionable "statistical analyses" – originally intended to support national-level decision-making, but instead having the effect of further perpetuating the widespread impression that the statistical community should not be entrusted with such a critical role. The examples motivate specific "lessons learned" and raise various "ethical questions" that readily generalize. Although most statisticians are likely to consider the associated principles of conduct to be self-evident, their application in practice obviously has not been uniformly successful. As a discipline and as a community of individuals fully capable of effectively supporting government officials and policy-makers, we can and must do better.

All of the cited examples are based on actual occurrences, but, to simplify the exposition and promote ready comprehension, the situational contexts and other details have been altered. The central observations, concerns and statistical issues, however, have been maintained. To repeat, while a specific example is introduced here as, say, a “Marine Corps Helicopter Effectiveness Study,” the underlying authentic circumstances may in fact involve, say, Army jeeps. This approach also serves to shield and protect the “guilty”, i.e., avoiding direct identification of the specific parties involved.

Five examples are presented and discussed in turn below, all dealing with high-level national defense or security issues. The examples are followed by a brief discourse of conclusions and lessons learned.

The setting for each example is similar. In response to a high-profile Department of Defense (DoD) study, an independent review was convened at government expense. The review panel comprised one or more assigned statisticians, *in toto* or as an integral subgroup of a more expansive collective of subject matter experts. The independent review panel was provided complete and unrestricted entrée to all aspects of the DoD study in question – including study data, detailed supporting data and analyses, briefings from study authors, access to study authors for follow-on questions, etc. Upon completion of their independent review, the review panel issued an assessment report to its sponsoring agency.

EXAMPLES

MARINE CORPS HELICOPTER EFFECTIVENESS STUDY

Based on observed test results on a series of developmental helicopters, the Marine Corps had determined that a new helicopter was suitable to proceed into full-rate production, despite some acknowledged unresolved performance issues. A critical component of their conclusion was a set of helicopter flight effectiveness curves that traced out aerodynamic performance across flight envelope regions.

An independent review panel of statisticians held several working sessions, interviewed the study authors, and scrutinized data sets. Its official assessment did not support the conclusions of the Marine Corps study. One cited reason was that that the flight effectiveness curves presented in the original study lacked confidence intervals, and that the observed features and reported performance could be merely manifestations of randomness.

When the review panel's assessment report was formally released, the original study authors objected to this particular cited shortcoming in their work. While it was indeed true that no confidence intervals appeared in their study *per se*, the authors had in fact constructed such curves – utilizing bootstrap procedures based on replicated observations, i.e., completely model independent. These definitively established that the reported curves were not attributable to randomness. Moreover, the study authors explicitly had briefed these exact results to the review panel and, at the review panel's request, provided paper copies of the entire briefing.

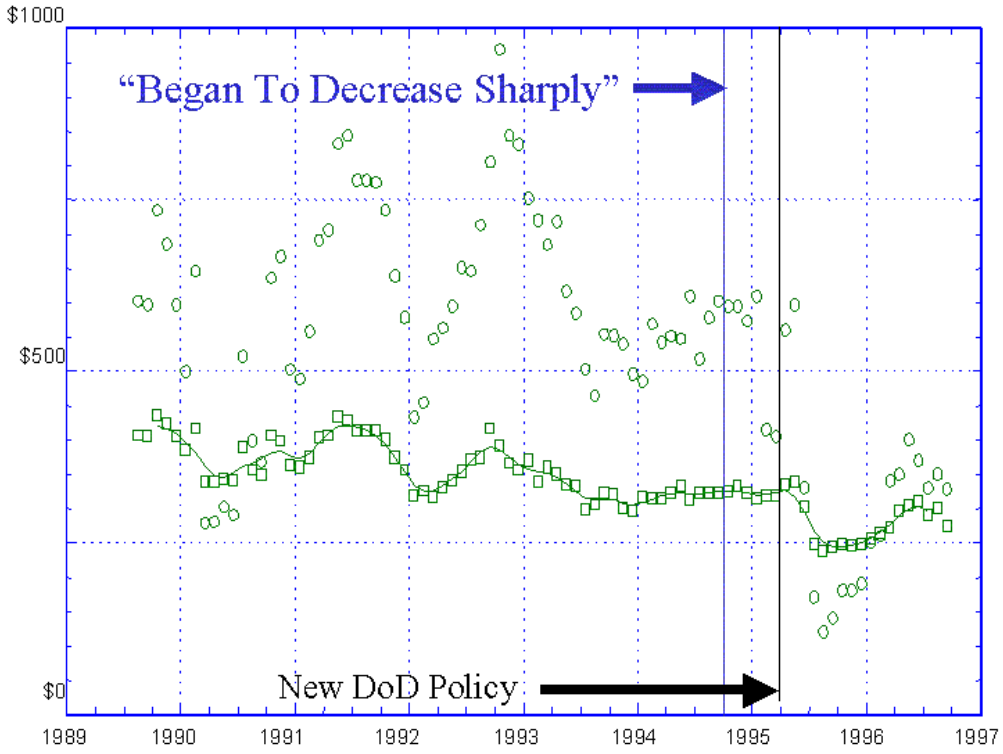
Despite the protest of the original study authors, the review panel did not rescind or even edit its assessment report. Neither did they apprise their government sponsors of the study authors' objections. Instead, the lead statistician reported back that *the review panel had been tasked to assess the original study report, and that what they had written in their assessment was technically correct and not an "error of fact"!*

DOD BUSINESS PRACTICES

In early 1995 the DoD instituted new business practices that significantly reduced procurement costs for certain classes of items, bringing expenditures down to a level commensurate with public sector commercial transactions. This conclusion was supported by a DoD study that offered a version of Figure 1 as substantiation. Here the "oval" and "rectangular"-shaped data points respectively depict per item costs for DoD purchases and for commercial non-DoD purchases. The new DoD business practices were officially implemented March 1995 (denoted by a vertical black line). Prior to that point, DoD costs generally greatly exceeded comparable public sector costs.

The group of independent reviewers declined to attribute the achieved cost reduction to the implementation of the new DoD procurement practices instituted formally in the first quarter of 1995. Instead, they argued that the costs depicted in Figure 1 actually "began to decrease sharply" in September of 1994 (denoted by a vertical blue line), a full half-year before the new DoD policy. Accordingly, they contended, something else other than the new DoD policy could just as well be the root cause for the decline in DoD procurement costs.

Now it is difficult for me, and every other individual that has been shown Figure 1, to imagine how any reasonable analysis could conclude that DoD costs started their rapid fall in the early fall of 1994. What had happened is that the review team had relied on a slightly different depiction of Figure 1 – identical in all respects except that the DoD cost data had been supplemented with a smoothing curve (similar in nature to the one displayed in Figure 1 for the public sector transaction costs). The smoothing curve was derived from a simple running average of nine neighboring points – the current point, the four preceding ones, and the four following ones. As such, the curve "anticipates" declines in future data and begins to decrease before the data actually do. This property had been explicitly noted in briefings to the independent reviewers, but it was not specifically cited in the published DoD study (although the particulars of the smoothing curve calculations were given).



DoD = "Oval" Public Sector = "Rectangle"

Figure 1. Dollar Transaction Costs – DoD & Public Sector

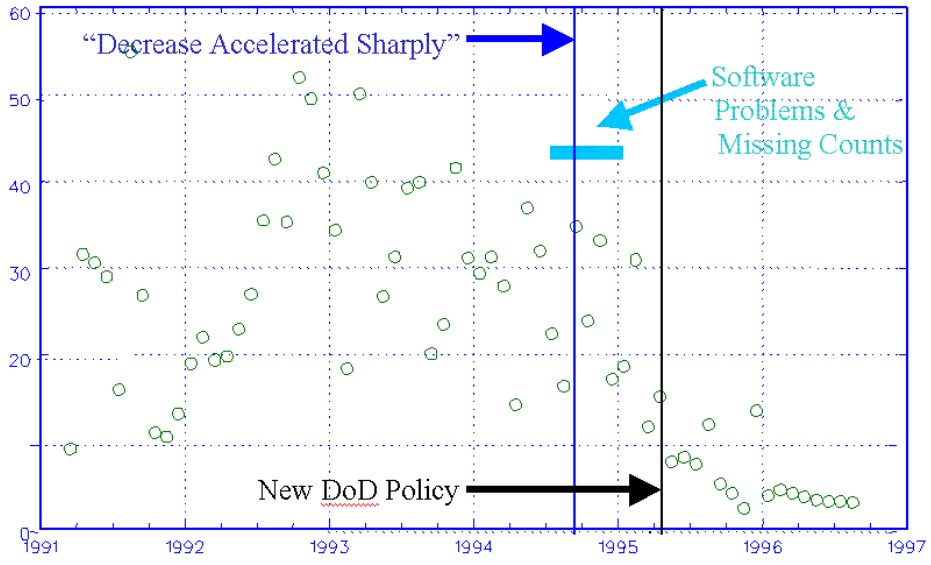


Figure 2. Counts of Monthly DoD Purchases

One aspect of the new DoD procurement policy was the shift away from many small purchases to fewer larger transactions reflecting volume discounts. The original DoD study pointed to the relatively low purchase counts post-March 1995 in Figure 2, and argued that, in tandem with Figure 1, this logically supported the contention that the new policy was effective. The independent assessment team, however, reached an entirely different conclusion. They argued instead that the purchase counts actually began to decrease well beforehand and, further, that the "decrease accelerated sharply" in September 1994 (depicted by the blue vertical line). While this observation correctly characterized Figure 2 as presented in the original DoD study, it ignored additional information made available to the review panel. The counts depicted in Figure 2 had been extracted from an automatic purchase tracking system. Although unknown to the DoD study team at the time of their original report, they subsequently learned that the computerized tracking system had experienced software problems during the latter half of 1994 (depicted by the light blue shading in Figure 2) resulting in an undercounting of the number of recorded transactions. This fact, as well as another set of independently derived confirmatory data, had been briefed explicitly to the assessment panel.

Despite the DoD study team's "reminders," the independent reviewers and their lead statistician declined to revise the assessment report. Neither did they apprise their government sponsors of any objections. To the consternation of the DoD researchers, attempts to appeal to the prestigious national committee of statisticians that furnished the assessment's team lead statistician also proved to be counterproductive. The committee's chair downplayed the significance of the divergent conclusions, referring to them as *merely "differences in data interpretation."*

AIR FORCE COST-BENEFIT COMBAT STUDY

The Air Force conducted a cost-benefit combat study in which it assessed the relative effectiveness of recent air campaigns in terms of successful mission sorties per thousands dollars of support and execution costs. An independent review panel scrutinized the Air Force's study and issued its own report. Its main conclusion was that the Air Force results are suspect because the Air Force may have selectively excluded data from its analyses:

> The Air Force study states that multiple air operations, with no kills, were not included because *"Further examination has shown that these operations were not, in themselves, major operations."* What does it mean to be a 'major' operation? Are actions deemed to be 'major' *ex ante*, or are they deemed *ex post* if they are observed to be sufficiently successful?

Here the reviewers extracted a quote from the Air Force study (in italics) to suggest that the Air Force analysis may have been biased. They did not independently review the descriptions of specific operations to determine whether any may have been improperly excluded from the calculations. Nor did they ask the Air Force analysts directly as to what the rules had been for inclusion in the computations.

One standard that the Air Force analysts employed was to not count logistical operations towards target kill effectiveness measures, since the mission did not include engaging targets. On the other hand, any associated tactical operation would be penalized the dollar cost of its supporting logistical operations. For example, Operation Desert Shield entailed the massive deployment of U.S. forces to Saudi Arabia as a prelude to the Gulf War against Iraq. The Air Force study did not score it as a "0-kill operation" and did not incorporate it into "benefit" analyses quantifying average mission accomplishment. The costs of Operation Desert Shield, however, were combined with those of the ensuing Operation Desert Storm, the actual combat portion of the Gulf War, in determining the "cost" of the latter operation. The original Air Force study indirectly alluded to this procedure [*emphasis added*]:

> Further examination has shown that these operations were not, in themselves, major operations. *Rather they established the international cooperation, increased deployed assets, and provided training that was later employed effectively in the follow-on major operations.*

This complete Air Force study citation provided explicit logical rationale for the exclusion of specific operations. The independent assessment, however, chose not to report these. When confronted with these facts, the lead statistician from the review team stated that their *assessment report was "accurate and fair."*

NAVY CIGARETTE SMOKING CESSATION STUDY

A group of Navy analysts had been assigned the task of assessing the degree of success of various programs aimed at assisting enlisting men curtail or cease altogether their cigarette smoking. Unfortunately, personal data on individual smoking frequencies had not been chronicled systematically. The analysts augmented the available data by examining indirect indicators of smoking, *i.e.*, surrogate variables, including official records of sick days and hospital visits. One Navy initiative that was assessed was whether increasing the cost of cigarettes charged to sailors aboard ship lead to less smoking. In their final study, the analysts noted that data limitations precluded them from computing direct estimates of the price elasticity of demand (*i.e.*, percent reduction in total consumption per percent increase in cigarette cost). However, they were able to estimate price elasticities for the cigarette usage indicators. Table 1 presents their study findings. The estimated elasticities are negative, synonymous with increases in cigarette prices being associated with fewer sick days and fewer hospital visits.

The Navy study focused on the "Best Estimates *e*" – describing how they were calculated and providing graphical depictions (with confidence intervals). It also drew a clear distinction between the price elasticities of the indirect indicators (that had been estimated) and the price elasticity of the demand for cigarettes (that had not been estimated):

> Assuming equal errors in the two variable, 10,000 bootstrapped replications of the data yield a best estimate of $e = -0.63$, a 95th percentile confidence interval of $-0.86 < e < -0.50$, and a standard least squares regression coefficient of $R = -0.71$." …
> The table summarizes the estimates of price elasticity as determined from the two indirect measures of cigarette use. While it is apparent that the 'elasticities' of the indirect usage indicators developed above are each related to the price elasticity of demand for cigarettes, those relationships are imprecisely understood. [*emphasis added*]

These careful characterizations were not faithfully duplicated in the subsequent descriptions thereof offered by a selected group of independent statistician reviewers. Their rendition appears in Table 2. Note that the Navy "Best Estimates *e*" are nowhere reported. Further, the presented values are incorrectly asserted to be estimates for the demand for cigarettes, exactly what the Navy study had insisted they were not estimating. The independent assessment continued to criticize the Navy study's "estimates of demands," essentially for the very reasons that had led the Navy analysts to abandon estimation of demand.

When informed of these discrepancies, the statisticians argued: "There had been some uncertainty on their part as to which set of estimates to report (*i.e.*, *e* or R)," "They were accustomed to regression-based estimates of elasticity", and "These are not 'errors of fact'." The Navy analysts countered that if the statisticians were uncertain, all they had to do is request clarification. They further noted that the estimates *e* also had been derived from regression analyses. And they continued to press for a corrected assessment report.

Rather than admit any errors, the statisticians offered an "addendum for clarification." They did not withdraw or correct the copies of their report that had already been published and distributed (*e.g.*, to the government sponsor that paid for their services). Instead, they published new report copies and inserted, with limited amplifying explanation, a solitary page after the References section to "reproduce the original Navy study table." This "reproduction" appears here in Table 3. Note that the sign of the "Best Estimate" for Hospital Visits erroneously is missing a negative sign, and that the estimate is mistakenly labeled as "*c*" vice "*e*."

Summary of Price Elasticity Estimates for Indicators of Cigarette Usage

Usage Indicator	Best Estimate	95% Confidence Interval	Regression Coefficient
Hospital Visits	-0.63	$-0.86 < e < -0.50$	-0.71
Sick Days	-0.60	$-0.98 < e < -0.29$	-0.51

Table 1. Navy Study Table [*emphasis added*]

Estimated Price Elasticities of Demand in the Navy Study

Data Set	Estimated Price Elasticity of Demand
Hospital Visits	-0.71
Sick Days	-0.51

SOURCE: Navy Study

Table 2. Independent Assessment Table [*emphasis added*]

Summary of Price Elasticity Estimates for Indicators of Cigarette Usage

Usage Indicator	Best Estimate	95% Confidence Interval	Regression Coefficient
Hospital Visits	0.63	$-0.86 < c < -0.50$	-0.71
Sick Days	-0.60	$-0.98 < e < -0.29$	-0.51

Table 3. Addendum to Independent Assessment

MULTI-SERVICE MEDICAL STUDY

Two military services conducted studies of how best to treat a particular health malady prevalent among combat veterans. Service A deduced "Treatment Regimen 1 was much more cost effective than Treatment Regimen 2," while Service B determined that "Treatment Regimens 1 and 2 were approximately equally cost effective." The inexact conclusion espoused by Service B's study was accompanied by an explicit acknowledgement that all available pertinent data were imprecise.

An independent review team of subject matter experts was convened. They judged neither of the service's studies to be plausible or persuasive. One statistician on the review panel subsequently authored an open-literature publication with main points illustrated by supporting examples. In recalling the multi-service medical study, the statistician reported "Service A concluded Treatment Regimen 1 was much more effective than Treatment Regimen 2, while Service B concluded exactly the opposite." [*emphasis added*]

The Study B analysts objected to this "incorrect characterization" of their overall conclusion, noting that they had never stated that Treatment Regimen 2 was much more effective than Treatment Regimen 1. The statistician, however, countered that his text was correct because in mathematical logic the opposite of "X" is "not X." In other words, the opposite of "Treatment Regimen 1 was much more effective than Treatment Regimen 2" is "Treatment Regimen 1 was not much more effective than Treatment Regimen 2." Apparently the statistician deemed this latter statement to be sufficiently consistent with Service B's originally reported conclusion, and was unconcerned that his particular words (*i.e.*, "Service B concluded exactly the opposite") were prone to misinterpretation.

DISCUSSION

Defense and national security debates are of critical importance. Relevant data may be limited and sparse, or, at the other extreme, voluminously intractable. Interpretational difficulties and analytical complexities may abound. The realm of statistics and our community of statisticians have much to offer – to properly frame and characterize the essential content of the available data, to focus attention on practically insightful issues, and to propose additional potentially illuminating data collection initiatives.

Despite our potential to contribute substantively, statisticians generally are excluded from defense and national security discussions – even when the primary issue is the analysis of a particular set of data or a specific statistical hypothesis is in question. Moreover, high-level decision makers rarely are inclined to suggest that statisticians should be consulted. Indeed, the exact opposite reaction is more typical. I have been at defense-oriented meetings and conferences in which the mere mention of "statisticians" (e.g., to announce an upcoming defense-related workshop sponsored by a statistical organization) spontaneously prompted condescending laughter and derisive commentary, including, for instance, "Do you *really* think that statisticians will be helpful?" What have we done to deserve such a reputation?

Judging by the examples cited above, the answer is in my opinion "plenty." It is always true that members of special panels and committees formed to address sensitive policy issues are charged with the grave responsibility of conducting impartial, objective, competent and comprehensive assessments. I argue, however, that individual statisticians granted the rare opportunity to participate in such endeavors must be held to an even higher accounting, for they represent not just themselves but also our entire discipline. If they fail to uphold exemplary standards – whether by direct act, complicity, or omission – they discredit us all.

Clearly statisticians supporting national-level panels and committees must ensure that all statistical analyses appearing in their reports are sound, whether they themselves are the architects of those analyses or not. But they should not restrict themselves to such a narrow focus. They must be cognizant of and comfortable with the basis of and derivation of overall conclusions. They must be sensitive to potential fairness and objectivity issues – including incomplete or unbalanced panel/committee membership, actual conflicts of interest or the possible perception thereof, over reliance on pre-conceived or initial impressions,

and the exclusion of inputs from and comprehensive scrutiny by independent subject matter experts. When reviewing the work of other analysts, those analysts should be provided an opportunity to comment on emerging products – especially with respect to the clarity and accuracy of the portrayals of the analysts' work.

Practicing statistical consultants learn early on that in order to be effective they must work hand-in-hand with their clients, closely and repeatedly interacting to develop a proper appreciation for the details of the problems at hand. It is my personal observation that "academic" statisticians entrusted with national policy studies all too often pursue the exact opposite tact, minimizing interactions with the very individuals whose reports they are to review. The first preference for some seems to be to just scrutinize an actual report, without any personal interchanges at all with the authors. Others occasionally may schedule *pro forma* briefings, but generally few extended two-way dialogues, especially at any detailed level, ensue.

Observers of this process might infer from such behavior that the statisticians either already have made up their mind or that they do not consider the matter at hand to be of any particular importance. Neither impression paints a flattering picture. When the statisticians' final product is itself problematic, as in the specific examples cited above, what are policy makers to conclude? Why should they even consider giving statisticians another chance?

ACKNOWLEDGEMENTS & DISCLAIMER

Portions of the authors' research were conducted at the Institute for Defense Analyses (IDA) – utilizing Professional Development funding, under various task orders supporting the Director of Operational Test and Evaluation (DOT&E) within the Office of the Secretary of Defense (OSD), and under a task order supporting the Executive Agent for the DoD Counterdrug Technology Developmental Program, under the sponsorship of the Office of the Undersecretary of Defense/Policy – Counternarcotics (OUSD(P)) in OSD.

The views expressed in this paper are solely those of the author. No official endorsement by IDA or the cited OSD offices is intended or should be inferred.

REFERENCES

1. Ellenberg, J. H. (2000), "Statistician's Significance," *Journal of the American Statistical Association*, 95, 1-8.