ELSEVIER

Accident Analysis and Prevention 36 (2004) 495–500

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Viewpoint

The harm done by tests of significance

Ezra Hauer*

35 Merton Street, Apt. 1706, Toronto, Ont., Canada M4S 3G4

Abstract

Three historical episodes in which the application of null hypothesis significance testing (NHST) led to the mis-interpretation of data are described. It is argued that the pervasive use of this statistical ritual impedes the accumulation of knowledge and is unfit for use.

Keywords: Significance; Statistical hypothesis; Scientific method

1. Introduction

Most university students are taught the rudiments of statistical null hypothesis significance testing (NHST for short). As a result, later in life, either as users of scientific knowledge or as its creators, they tend to regard NHST to be the hallmark of sound science, an effective safeguard against spurious findings. That the logical foundation and the scientific merit of NHST is questioned by prominent statisticians and scientists is not mentioned in text books and courses on introductory statistics; therefore it is not common knowledge. Yet, volumes have been written about the 'significance controversy' (see, e.g. books by Chow, 1996; Harlow et al., 1997). I have written about the paralyzing effect of statistical significance on road safety research a long time ago (Hauer, 1983) and did not plan to return to this topic again. However, the road safety literature is a constant reminder of the continuing real harm done by NHST. The harm is that of using sound data to reach unsound conclusions thereby giving sustenance to non-sensical beliefs. In the end, these non-sensical beliefs cause needless loss of life and limb.

In this paper, I will not dwell on the common criticisms of NHST. Therefore, it will not be necessary to explain its fine points. What I wish to demonstrate here is that NHST, as applied in research on road safety, often leads to the subversion of reason and of science. I will do so by relating three historical episodes.

2. Episode 1: the right-turn-on-red story

This is an old story (see Hauer, 1991). The practice of allowing right-turn-on-red' (or RTOR) at signalized intersections started in California in 1937 (some say that is started earlier, in New York City). For a long time, it was frowned upon by engineers in other states who had safety concerns. A major impetus for the general adoption of RTOR was the 1973 oil crisis and the "Energy Policy and Savings Act" adopted by Congress in 1975. Our story begins in 1976 when a consultant submitted a report about the safety repercussions of RTOR to the Governor and General Assembly of Virginia. The studies then extant were deemed deficient and the consultant did his own before–after study at 20 intersections with the results in Table 1.

Looking at the data in Table 1, persons without training in statistics would think that after RTOR was allowed, these intersections were somewhat less safe. However, the consultant concluded, quite correctly, that the change was not statistically significant. The Commissioner of the Virginia Department of Highways and Transportation sent the consultant's report to the Governor and in the letter of transmittal wrote: "we can discern no significant hazard to motorists or pedestrians from implementation of the general permissive rule (i.e. of RTOR). No significant increase in traffic crashes has been noted following adoption of right-turn-on-red in any state including Virginia". Obviously, there was miscommunication. In English 'significant' means 'having or likely to have considerable influence or effect'; the synonym of 'significant' is 'important'. In statistics 'not' significant' means that the data is insufficient to reject the (null) hypothesis of 'no effect'. Thus, the consultant said one

* Tel.: +1-416-978-5976; fax: +1-416-978-5054.
E-mail address: ezra.hauer@utoronto.ca (E. Hauer).

0001-4575/$ – see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/S0001-4575(03)00036-8

Table 1
The Virginia RTOR study

	Before RTOR signing	After RTOR signing
Fatal crashes	0	0
Personal injury crashes	43	60
Persons injured	69	72
Property damage crashes	265	277
Property damage (US$)	161243	170807
Total crashes	308	337

thing and the Commissioner transmitted something entirely different.

More published studies followed. One study in 1977 found that there were 19 crashes involving right turning vehicles before and 24 after allowing RTOR and "this increase in accidents in not statistically significant, and therefore it cannot be said that this increase in RTOR accidents is attributable to RTOR". And so the sequence of small studies all pointing in the same direction but with statistically not significant results continued to accumulate, till that last study which I followed was published in 1983. While 287 crashes to right turning vehicles were expected, 313 were counted. The authors concluded, once again, that there was no significant difference in vehicular crashes. Similarly for pedestrians. In one state, 74 were expected and 92 occurred; in another state 81 were expected and 87 occurred. An yet, the authors concluded ". . . that there is no statistically significant difference . . . (in) pedestrian accidents before and after RTOR. There is no reason to suspect that pedestrian accidents involving RT operations (right turns) have increased after the adoption of RTOR in either state".

After RTOR became nearly universally used in North America, several large data sets became available and the adverse effect of RTOR could be established (Zador et al., 1982; Preusser et al., 1982).

The problem is clear. Researchers obtain real data which, while noisy, time and again point in a certain direction. However, instead of saying: "here is my estimate of the safety effect, here is its precision, and this is how what I found relates to previous findings", the data is processed by NHST, and the researcher says, correctly but pointlessly: "I cannot be sure that the safety effect is not zero". Occasionally, the researcher adds, this time incorrectly and unjustifiably, a statement to the effect that: "since the result is not statistically significant, it is best to assume the safety effect to be zero". In this manner, good data are drained of real content, the direction of empirical conclusions reversed, and ordinary human and scientific reasoning is turned on its head for the sake of a venerable ritual. As to the habit of subjecting the data from each study to the NHST separately, as if no previous knowledge existed, Edwards (1976, p. 180) notes that "it is like trying to sink a battleship by firing lead shot at it for a long time".

Table 2
Summary of results by Abboud and Bowman (2001)

Crash type	Mean crash rate 'after"	Expected crash rate 'after'	Change (%)
Treatment: addition of two foot paved shoulders			
Single-vehicle	35.66	36.83	−3.18
Same-direction	1.99	3.06	−34.97
Opposite-direction	5.23	5.63	−7.10
Fatal	1.62	1.83	−11.48
Personal injury	28.76	31.52	−8.76
Property damage only	49.79	53.60	−7.11
Treatment: addition of four-foot paved shoulders			
Single-vehicle	27.96	33.31	−16.06
Same-direction	3.32	2.71	22.51
Opposite-direction	5.66	4.58	23.58
Fatal	1.91	2.89	−33.91
Personal injury	27.10	30.64	−11.55
Property damage only	43.92	48.01	−8.52

3. Episode 2: the safety effect of paving shoulders

A more current example is in a paper reporting on the safety effect of adding two-foot wide paved shoulders on 263 miles of two-lane rural roads with 386 'before' and 478 'after' crashes and four-foot wide paved shoulders on 404 miles of two-lane rural roads with 444 'before' and 455 'after' crashes (Abboud and Bowman, 2001). The main results, given in Cash Rate (crashes/100 million vehicle-miles) as extracted from the authors' Tables 1 and 2 are in Table 2 below.

Fig. 1 shows the percentage change in crash rate for the more frequent crash types for which the findings are more reliable; Fig. 2 pertains to less frequent crash types for which the findings are less reliable.

The overall impression is what one might expect. That is, that the addition of a two-foot paved shoulder has reduced all crash types, and the addition of a four-foot paved shoulder has reduced crashes even more (except in the 'same direction' and the 'opposite direction' categories where crashes were so few that the results are simply erratic). The authors provide the following interpretation of the data:

> None of the differences, however, between the actual and expected crash rates were found to be statistically significant (p. 37).

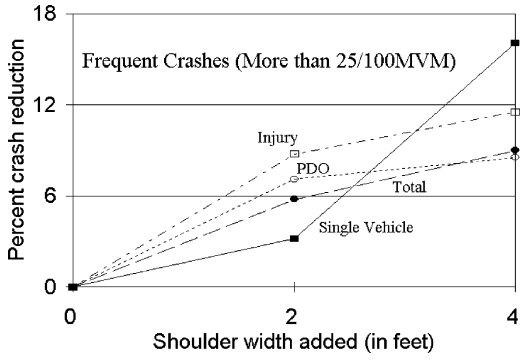

Fig. 1. Percentage reduction in frequent crashes.

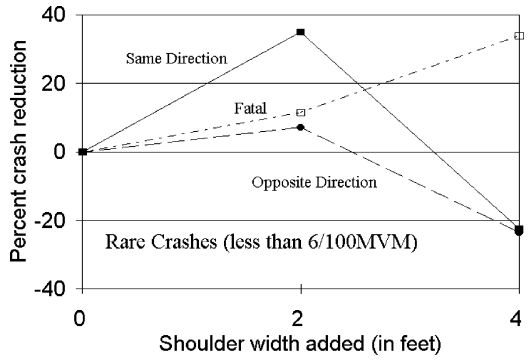

Fig. 2. Percentage reduction in rare crashes.

Once again common sense and statistical ritual point in opposite directions. The figures show that, e.g. after a two-foot paved shoulder has been added, the crash rate has declined for all crash types and all severities. Therefore, ordinary reasoning would lead to the conclusions that paving shoulders has reduced crashes. And yet, because of the paucity of the data, none of these reductions proved statistically significant. But quasi-science wins again; and so, in their Conclusion section the authors write:

> The study could not discern any statistically significant differences in either crash rate or severity rate between two- and four-foot shoulder installations. Unless (other) benefits . . . are considered important to practitioners, this study does not show the increased construction cost of four-foot shoulders on state routes to be justified by an increase in traffic safety (p. 37).

The authors are right in saying that none of the differences were statistically significant. However, the authors are entirely wrong to spin this into meaning that shoulder paving cannot be justified by its safety effect. They did not do any cost-effect calculation. They seemed to have assumed that because the estimates in the rightmost column of Table 1 are not statistically significant, it is good form to take them to be zero. This makes no sense. It is the estimates in the rightmost column of Table 2, not zero, which represent the most likely safety effect of shoulder paving when based on this study. The absence of statistical significance does not mean and should never be taken to mean that 0 is the most likely estimate.

Just as with RTOR, when articles of this kind are published in the professional literature, they are taken seriously and may affect real decisions. In this case, because money could be saved if shoulders do not need to be paved, these research results were already considered by a committee of the Association of State Highway and Transportation Officials. The danger is that the use of the NHST ritual may lead to incorrect guidance for practice and thereby to unjustified loss of life and limb.

It can be argued that these two episodes represent a mis-application of the NHST. The mis-application being that the statements about absence of statistical significance were taken to mean that a measure or intervention had no effect. Unfortunately, neither the readers of the professional literature nor many contributors to it are clear about this distinction. Therefore, if what has been presented are instances of mis-application, one must conclude, at least, that NHST is given to common mis-application in the hands of many users. Advocates of NHST could perhaps argue that what is needed are better educated users. This is a vain hope. As will be shown next, not only readers and contributors to learned journals are given to mis-application of the NHST, prominent statisticians suffer from the same affliction.

4. Episode 3: speed limit increases

Balkin and Ord (2002) published an influential study about the effect on fatal crashes of speed limit increases in the interstate highway system. Using structural modeling of time series data they estimated for each state what change in fatal crashes could be attributed to the speed limit increases implemented on interstate highways in 1987 and in 1995. Some of their results are reproduced in Table 3.

Counting the number of non-zero entries in each column of the table the authors say (p. 8):

> "We can summarize the findings as follows:
>
> - 19 of 40 states experienced a significant increase in fatal crashes along with the first (1987) speed-limit increases on rural interstates.
> - 10 of 36 states experienced a significant increase in fatal crashes along with second (1995) speed-limit increases on rural interstates."

It is obvious that 0.0 is not the best estimate of the change in fatal crashes in all these instances. Why the authors decided to enter 0.0 can perhaps be understood from the numerical example by which they explain their method. In their paper there is a graph of the monthly time series of fatal crashes from 1975 to 1998 for rural interstates in Arizona and, referring to this graph, the authors say (p. 6) that:

> "We see a significant increase in the level around 1987 but none around 1995. . . . Statistically it is estimated that

Table 3
Predicted percentage increase in the number of fatal crashes attributed to the speed-limit increases on rural interstates (from Balkin and Ord, p. 10, Table 3)

State	First percentage (1987)	Second percentage (1995)
Alabama	0.0	24.8
Arizona	41.0	0.0
. . .	. . .	. . .
Missouri	13.0	42.2
Nebraska	35.5	0.0
. . .	. . .	. . .
West Virginia	46.2	0.0
Wisconsin	24.3	0.0

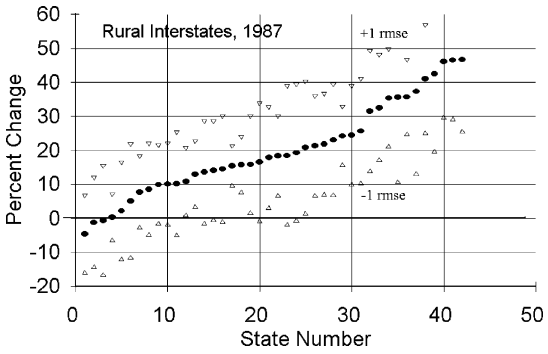

Fig. 3. Percentage change in fatal crashes after 1987 speed-limit increase.

the 1987 speed-limit increase resulted in a 41% increase in rural interstate crashes an Arizona. There is no statistical evidence that the 1995 speed-limit increase has any additional effect on the number of crashes."

That is, failure to reject the null hypothesis of zero effect at the 10% level of significance was equated with the absence of statistical evidence for an increase in the expected number of crashes. In all these cases, 0.0 was entered in the table. Thus, the table contains two kinds of entries: either estimates of percentage change when the increase was statistically significant, or 0.0 by NHST convention but unsupported by either data or prior-knowledge when the increase was not statistically significant.

The authors have generously shared with me their detailed results. Figs. 3 and 4 show their original estimates and their precision. The states are ranked in the increasing order of the estimated percentage increase in expected accidents and the arrowheads are placed one root-mean-square error around these estimates. The same data is shown as histograms in Figs. 5 and 6.

Balkin and Ord conclude (p. 1) that:

. . . The results cast doubt on the blanket claim that higher speed limits and higher fatalities are directly related. After the initial speed-limit increases in 1987, the number of fatal accidents on rural interstates increased in some states

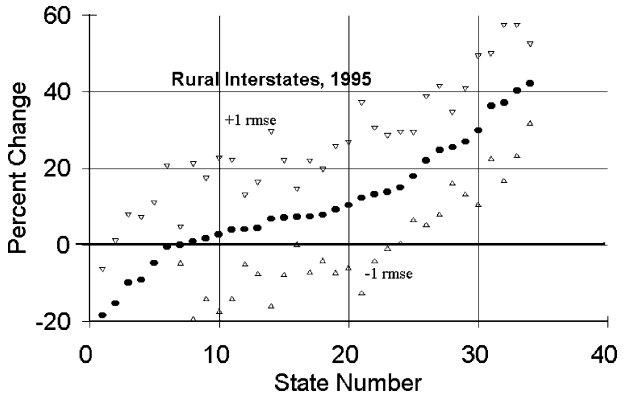

Fig. 4. Percentage change in fatal crashes after 1995 speed-limit increases.

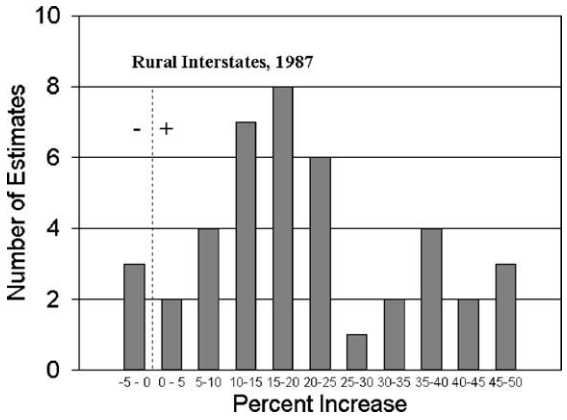

Fig. 5. Histogram of percentage change for 1987 speed-limit increases.

but not in all. The 1995 round of speed-limit increases generally showed smaller increases on rural interstates . . .

This statement is somewhat loosely worded because it speaks about the increase in "number of fatal accidents" when the intent is to speak about the increase in "expected number of fatal accidents". Disregarding this semantic imprecision, the question is whether, based on the evidence of the data, it is sensible to conclude that the expected number of fatal accidents "increased in some states but not in all"; after all, the figures show only estimates, not underlying expected values, and the estimates are of limited precision.

The average of the post 1987 crash increases was 20.1%. For the three states at the far right of Figs. 3 and 4, the estimated increase was larger than 46%. Because these are estimates, and because these are the three highest of 42 estimates, they are subject to the well known regression-to-mean bias. That is, their underlying expected values are most likely less than 46%. A similar reasoning applies to the three states at the far left of Figs. 3 and 4 except that here the regression-to-mean bias is in the opposite direction. That is, because these are estimates, and because they are the lowest of 42 states, the underlying expected values are larger than the estimates, perhaps positive. In short, were the

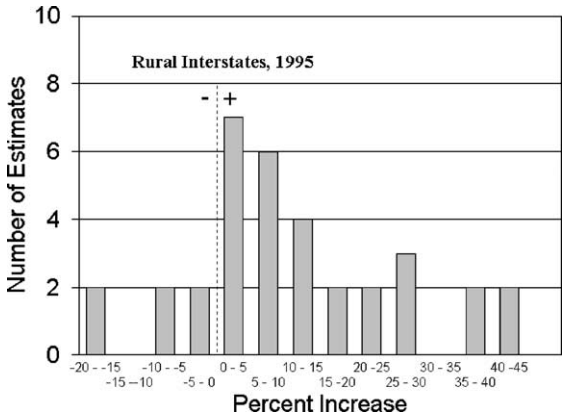

Fig. 6. Histogram of percentage change for 1995 speed-limit increases.

expected values known, their histograms would likely be more compactly concentrated around the mean than are the histograms in Figs. 5 and 6. The essential question then is how compact is the distribution of these unknown expected values.

This question can be answered. Let x_i be the estimate for state i, v_i the variance of the estimate, and m_i the underlying expected value. The sample variance of the x_i can be calculated from the histograms and its expected value is VAR$\{X\}$. Denote the variance of the m_i's by VAR$\{m\}$ and the average of the v_i's by $\bar{v}$. It can be shown (see Appendix A) that VAR$\{X\}$ = VAR$\{m\} + \bar{v}$. For the 1987 increases the sample variance of the x_i is 189. For each x_i, we have its RMSE (see triangles in Figs. 3 and 4) and $\bar{v}$ can be estimated from these as 234. Therefore, one is led to concluded that VAR$\{m\}$ close to zero. For the 1995 increases, the sample variance of the x_i is 235 and $\bar{v}$ is estimated as 255. Here too, the estimate of VAR$\{m\}$ is close to zero. Thus, neither for the 1987 nor for the 1995 speed-limit increases can there be a reasonable doubt that the expected increase of fatal crashes in all states was positive.

A less rigorous but more intuitive way to show that the data at hand is consistent with the possibility that VAR$\{m\}$ is very small is the following. If in all 42 states the expected increase post-1987 was +20.1% and if the estimates had a standard deviation of 14.7% (this being the sample mean of the RMSEs for the 42 states) one should expect to see three or four negative estimates. Three have materialized. Similarly if in all 34 states where speed-limits increased in 1995 the expected increase in expected accidents was +11.0% and the estimates had a standard deviation of 15.4% (this being the sample mean of the RMSEs for the 34 states where speed-limits increased) one should expect about eight negative estimates. Six have been observed. That is, the results obtained are perfectly consistent with the possibility that on rural interstates the speed-limit increases in 1987 and in 1995 were associated with an increase in the expected number of accidents in all states. Indeed, this is the hypothesis that is best supported by the data!

Why then do the authors say that their results "cast doubt on the blanket claim that higher speed-limits and higher fatalities are related"? The opposite would have been the more sensible conclusion to draw since the data provided extraordinarily strong evidence that following the 1987 and the 1995 speed-limit increases the expected number of accidents increased in all states. As in the previous episodes, data painted a characteristically fuzzy but reasonable reflection of reality. However, when good data is passed through the NHST filter, a negative tends to emerge; black turns to white and white to black.

The Balkin and Ord paper has been formally discussed (pp. 13–26) by several prominent statisticians (J. Ledolter, M.D. Fotaine, T.T. Qu, K. Zimmerman, C.H. Spiegelman and A. Harvey). They comment extensively about several aspects of the statistical approach. Surprisingly, no question was raised about the use of NHST, about the appropriateness of subjecting each state separately to a NHST, or about the legitimacy of conclusions drawn in this manner. The use of NHST has received no comment.

5. Summary

The RTOR story shows how easy it is for laymen to confuse 'not significant' in the statistical sense with 'not important' in the common sense. The confusion is not merely semantic and is not confined to persons without statistical education. As is evident from the shoulder-paving episode, even the statistically sophisticated believed that a non-rejection of the 'no-effect' null hypothesis amounts to some kind of confirmation that the data show that there was no effect. Moreover, the speed-limit episode shows that even scholars who teach statistics and write text books on the subject, even they do not always distinguish between what is an estimate and what is a not-rejected (but otherwise unsupported) null hypothesis; even they, under the spell of the not-rejected zeros, read into the data the opposite of what it says. Therefore, one should not entertain the hope that a more informed use of the NHST might help avoid its many pitfalls and shortcomings. Experience shows that the ritual is so pervasively misapplied as to be simply unfit for use.

In all three episodes, the use of NHST led to conclusions which are incorrect and contrary to a straightforward interpretation of the data. The authors did not say: "it looks as if the measure (allowing RTOR, not paving shoulders, or increasing speed limits) increases accidents, but we are not sufficiently certain". The authors said that: "there is no reason to suspect that pedestrian accidents involving RT operations (right turns) have increased after the adoption of RTOR in either state" and that "... this study does not show the increased construction cost of four-foot shoulders on state routes to be justified by an increase in traffic safety" and that "the results cast doubt on the blanket claim that higher speed limits and higher fatalities are directly related". This tendency of NHST to strip useful data of their meaning, is the most disturbing aspect of its wide application. In the context of progress towards better factual knowledge it amounts to a learning disability. In the context of an aid to public decision-making and policy it leads to misapplication of resources and unnecessary loss of life and limb.

In all three episodes, the important question was about the likely effect of a measure on crashes. But, instead of asking "how many more crashes?" and the authors chose to ask "are we sufficiently sure that the effect was not zero?" This substitution of questions led to all the subsequent entanglements. These can all be avoided by not testing statistical hypotheses when the research question is about the effect of some treatment; by returning to common sense and the mainstream of science and providing estimates of effect magnitude and its standard error instead.

The notion that progress in road safety research would be faster were the habitual use of NHST abandoned evokes

strong reactions. The usual comment is that the fault is not in the procedure itself but in its frequent misapplication. Frequent misapplication, when it comes after generations of teaching and education, is in itself a serious fault. Misapplication in the hands of even the tone-givers of statistical practice should give pause to all.

6. A postscript

I sent a draft of this paper to a colleague who, after reading, commented that the paper would be more balanced if I noted that NHST could perhaps lead to correct conclusions if those who used it remembered that the application of the Neyman–Pearson procedure requires the ascertainment of the Type II error (i.e. of not rejecting the null hypothesis when some alternative hypothesis is true). Attached to his e-mail was a draft paper of his own. He argues in his paper that the kind of data aggregation which is inherent in multivariate regressions can lead to incorrect results. To illustrate this, he generated pedestrian crash data by simulation on the assumption that the probability of a crash is a function of traffic speed and volume. The output of the simulation was then used as input into a regression the aim of which is to estimate the probability of a pedestrian crash as a function of traffic volume and speed. Looking at the result he notes that "... the (regression) coefficient corresponding to traffic volume was significantly different from zero, but that corresponding to mean speed was not. This suggests that ... speed does not need to be considered in assessing pedestrian risk". His principal argument is that multivariate regression incorrectly shows that speed has no effect on crashes when in the data-generating process speed did play a role. However, the regression did show that speed is positively associated with crashes. It is only after he subjected it to an NHST that he concluded that because the parameter is not significant, it is zero! Thus, to the same message in which he advocates the correct use of NHST, he appends a paper in which no attempt is made to do so. In this case, as in the myriads other regressions performed daily, if the '*t*' statistic of a regression coefficient is not significant, the regression constant is taken to be 0 and the variable is dropped.

Appendix A

Consider n distributions with means and variances (m_1, v_1), (m_2, v_2), ..., (m_j, v_j), ..., (m_n, v_n). From each of these distributions obtain one realization of a discrete random variable X that can take on values $x_1, x_2, \ldots, x_i, \ldots, x_k$. Let $p_j(x_i)$ denote the probability that a realization from the jth distribution is x_i. The relative frequency with which the realization x_i is observed in trials of this kind approaches

$$P(X = x_i) = \frac{\sum_{j=1}^{k} p_j(x_i)}{n} \tag{A.1}$$

To express $E\{X\}$ and $\mathrm{VAR}\{X\}$ as a function of (m_1, v_1), (m_2, v_2), ..., (m_j, v_j), ..., (m_n, v_n) write

$$E\{X\} = \sum_{i=1}^{k} x_i P(X = x_i) \frac{\sum_{i=1}^{k} x_i \left[\sum_{j=1}^{n} p_j(x_i)\right]}{n} = \frac{\sum_{j=1}^{n} m_j}{n} = \bar{m} \tag{A.2}$$

$$E\{X^2\} = \sum_{i=1}^{k} x_i^2 P(X = x_i) = \frac{\sum_{i=1}^{k} x_i^2 \left[\sum_{j=1}^{n} p_j(x_i)\right]}{n} = \frac{\sum_{j=1}^{k} (v_j + m_j^2)}{n} \tag{A.3}$$

From here,

$$\mathrm{VAR}\{X\} = E\{X^2\} - E^2\{X\} = \frac{\sum_{j=1}^{k} (v_j + m_j^2)}{n} - \frac{n\bar{m}^2}{n} = \frac{\sum_{j=1}^{k} v_j}{n} + \frac{\left(\sum_{j=1}^{k} m_j^2\right) - n\bar{m}^2}{n} \tag{A.4}$$

Denote the average of the variances $v_1, v_2, \ldots, v_n$ by $\mathrm{AVG}(v)$ and the variance of the $m_1, m_2, \ldots, m_n$ around their mean value by $\mathrm{VAR}\{m\}$. With this notation,

$$\mathrm{VAR}\{X\} = \mathrm{AVG}(v) + \mathrm{VAR}\{m\} \tag{A.5}$$

This interesting result shows that the histogram of unbiased parameter estimates usually has a larger variance than the histogram of the underlying (unknown) parameters; furthermore, that one can estimate the variance of the unknown underlying parameter estimates by subtracting from the sample variance of the parameter estimates the average of the estimated variances of the parameter estimates.

References

Abboud, N.K., Bowman, B.L., 2001. Evaluation of two- and four-foot shoulders on two-lane state routes. ITE J. 71 (6), 34–39.

Balkin, S., Ord, J.K., 2002. Assessing the impact of speed-limit increases on fatal interstate crashes. J. Transport. Stat. 4 (1), 1–26.

Chow, S.L., 1996. Statistical Significance. Rationale, Validity and Utility. Sage, London.

Edwards, A.W.F., 1976. Likelihood. Cambridge University Press, Cambridge.

Harlow, L.L., Mulaik, S.A., Steiger, J.H., 1997. What if There were No Significance Tests? Lawrence Erlbaum, London.

Hauer, E., 1983. Reflections on methods of statistical inference in research on the effect of safety countermeasures. Accid. Anal. Prev. 15 (4), 275–285.

Hauer, E., 1991. Should STOP YIELD? Matters of method. ITE J. 69 (9), 25–32.

Preusser, D.F., Leaf, W.A., DeBartdo, K.B., Blomberg, R.D., Leoy, M.M., 1982. The effect of right-turn-on-red on pedestrian and bicycle accidents. J. Safety Res. 13, 45–55.

Zador, P., Moshman, J., Marcus, L., 1982. Adoption of right-turn-on-red: effects on crashes at signalized intersections. Accid. Anal. Prev. 14 (3), 219–234.