INTEGRATIVE STATISTICS

Customized Statistical Analysis
and Survey Research

ERRORS IN
2020 MASSACHUSETTS UNIFORM CITATION DATA ANALYSIS REPORT

Review of an anonymous report published by Salem State University and Worcester State University, February 7, 2022[1]

March 2022

The anonymous report on Massachusetts police stops of motorists and race, released Feb. 7, 2022, relies on extensive data and expresses many insightful and useful ideas about ways to conduct this sort of research. Unfortunately, the paper's implementation of those ideas casts grave doubt upon the findings.

The report shows too many errors and inconsistencies in its use of statistical methods. Moreover, its findings too often conflict with results this reviewer has obtained in recent analyses of Department of Transportation data. Below, the nature of each type of error in the report is summarized first, and details are elaborated upon in the following section.

Summary

1. *Incorrect interpretations about probabilities*.

In multiple places in the paper, the authors use a particular method to show how probabilities differ between groups. This is the method of odds, or more specifically, the relationship between two odds.

Many statements in the paper attempt to explain these results, but they all do so incorrectly. The authors repeatedly claim that odds and probabilities are the same. Anyone reasonably familiar with gambling, let alone any statistician, will recognize that they are not.

Similarly, a ratio between two odds, say, of 0.64, does not mean that one event is 64% as probable as another. To make such a statement about probabilities requires a conversion, which the authors in all cases fail to make.

[1] Report obtained from https://www.mass.gov/doc/2020-massachusetts-uniform-citation-data-analysis-report/download.

2. *Confusion of results on group breakdowns.*

In many instances the authors attempt to evaluate the ways in which results might break down differently for different groups. For example, frequencies of drivers being stopped by police vs. not stopped by police might break down differently for Whites and Non-Whites[2]. The authors frequently report these sorts of results using a very common and basic statistical method called a Chi-Square test.

Unfortunately, in all cases, the authors confuse the main result of the test itself – the "Chi-Square statistic" – with the significance level (or "p-value") computed for it. For example, we are told erroneously that a Chi-Square result of ".000," which would show absolutely no difference in group breakdowns, somehow shows a statistically significant difference in those breakdowns. This is an impossibility.

3. *Inadequate in-depth analysis*.

The paper uses another method, logistic regression, to test the way in which the frequency of police stops may have differed during daylight as opposed to night time. Skilled use of regression could have assessed the extent to which Whether a Driver Was Stopped depended on each of multiple factors.

Instead, the authors' implementation is limited to a single factor. As a result, their analysis adds little to their existing presentation and fails to address the critical question of what differing roles may have been played by different factors.

4. *Confusion of different questions.*

The paper seems to confuse two types of questions:

a) Among Non-White drivers, what percent were stopped?

b) Among drivers stopped, what percent were Non-White?

These two are not interchangeable. An answer to one cannot serve as the answer to the other.

These four fundamental types of errors cast serious doubt on the analysis and on the conclusions reached in the report.

Elaboration

Following are specific examples and more detailed discussion of the four categories of errors in the report.

[2] The names of data categories used in the report are also used here for reference.

1. *Incorrect interpretations about probabilities*.

One example, on page 22, purports to explain odds and probabilities, but treats them as if they were the same. That is patently wrong.

Consider the following:

Rain has a 20% probability on Monday and a 40% probability on Tuesday. The probability is thus 100% higher for Tuesday.

The odds of rain, on the other hand, are:

For Monday, .2 / (1 - .2) = .2 / .8 = .25.

For Tuesday, .4 / (1 - .4) = .4 / .6 = .67.

The odds ratio, Tuesday/Monday, is thus .67/.25 = 2.68. Tuesday has odds 168% higher than Monday's, but a probability (from above) that is merely 100% higher. Odds and probabilities are not the same.

2. *Confusion of results on group breakdowns.*

The errors described above concerning Chi-Square results occur on pp. 31, 33, 34, and throughout Appendix C. This reviewer found no case in which a Chi-Square statistic was correctly reported. The extensive repetition of this type of error suggests that it is not the result of carelessness but rather reflects a misconception about, or a lack of familiarity with, statistical reporting.

In addition, the paper incorrectly states that in the context at hand a significant Chi-Square result means that

> "there is a relationship between the likelihood of a Non-White motorist being stopped during the day as compared to darkness (as well as there is a relationship between the likelihood of a White motorist being stopped during the day as compared to darkness)."

This (difficult) sentence misses the point about these tests. The existence of two such relationships does not at all mean that such a Chi-Square result will be significant. What are required are distinct relationships for each group. If each relationship is similar, so that each group is more likely to be stopped during the day than in darkness (or vice versa), then the test result will be non-significant. The authors are again attempting to explain basic statistical concepts without apparently possessing a full grasp of those concepts.

3. *Inadequate in-depth analysis*.

The paper’s logistic regressions (pp. 21, 31) could have assessed the vital questions of to what extent Whether Stopped depended on Daylight/Darkness; on White/Non-White; and on these two independent variables’ joint effect, or their statistical interaction. Analyzing the latter would have answered the very pertinent question “does the Daylight/Darkness distinction work differently for White and Non-White groups?”

Regression could also have shown to what extent the outcome could be predicted using these independent variables. A very predictive model, explaining perhaps 40% of the variability in the outcome, would merit a very different reception among readers than one explaining only 4%. The authors' actual use of regression hinges instead on their reporting of odds ratios, which, as discussed above, is incorrect. The paper's regression otherwise adds little information and misses the opportunity to shed light on key aspects of the topic.

4. *Confusion of different questions.*

The authors seem to switch numerator and denominator unexpectedly, rendering some results either confusing or inapplicable.

Table 2 is an instructive example. While the table is not clearly explained, what it shows are column percentages – not row percentages as would be expected given the text. For instance, within the Daylight column, it is the percentages for Non-White and for White that add up to 100%. This means that, for each condition (Daylight or Darkness), rather than showing how likely Non-White drivers (for example) were to be stopped, the table shows what percent of those stopped were Non-White. This confuses the handling of key questions, such as "What percent of Non-Whites (and of Whites) were stopped under each condition?"

This switching of numerator and denominator is inconsistent with the way the authors interpret the findings elsewhere. The presentation of results is inconsistent with findings that seem to treat Whether Stopped as the dependent variable and Non-White/White as an independent variable. The report seems to switch these variables' roles, confusing the results.

Conclusion

The extensive data obtained for this study, and the many constructive ideas offered about analysis, could provide the basis for an effective and trustworthy report. However, the anonymous authors' faulty implementation of statistical methods shows too many errors to afford confidence in any of their findings.

One would hope that any future revision would be prepared in collaboration with colleagues possessing greater statistical expertise, and that the paper would be submitted to a journal for peer review. While peer review is not a foolproof safeguard, it most often results in improved submissions in which readers can place greater trust.

Roland B. Stark, M.Ed., Statistician and Research Consultant
www.IntegrativeStatistics.com
Roland@IntegrativeStatistics.com