Statistical Hypothesis Testing Applied to US 50 Accident Data

Robert W. Byren, TESA Tech Team
Sydney Morrow, TESA Tech Team

Abstract

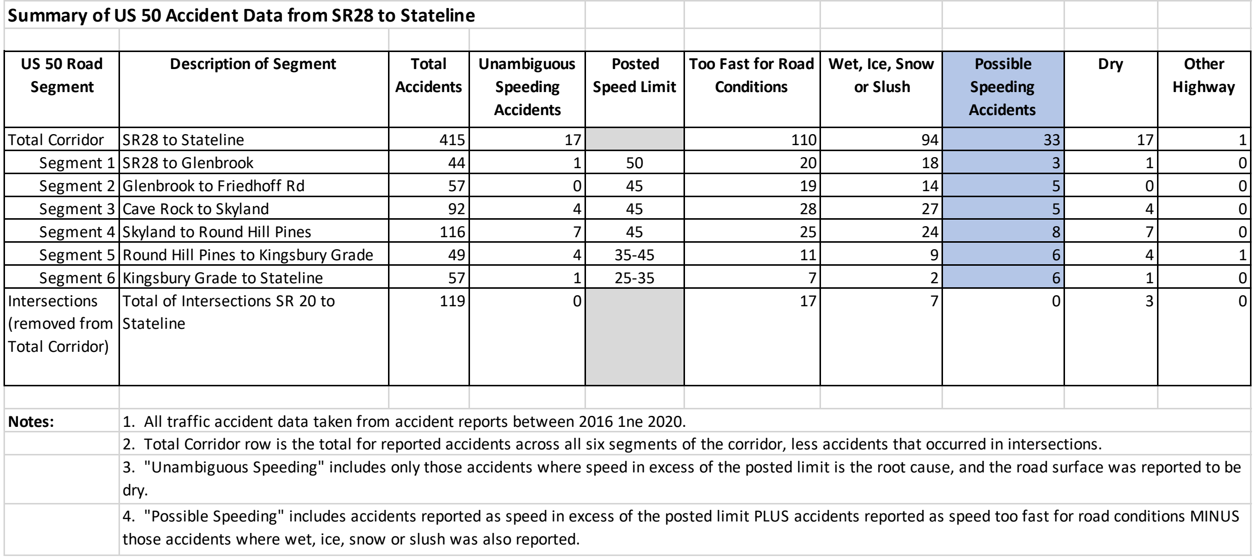
This paper tests the NDOT US 50 Tahoe East Shore Corridor Management Team's (CMP) assertion that 30% of the accidents that occur along the US 50 corridor between the SR28 intersection and Stateline are caused by excessive speed using the standard statistical hypothesis testing methodology. Our analysis is based on a sample of traffic accident data taken between 2016 and 2020. We conclude that the data for the entire corridor and individually for each segment do not support the CMP assertion that 30% of the accidents resulted from vehicles traveling above the posted speed limits, and in fact support much smaller percentages, thereby refuting NDOT's argument that so-called "road diet" measures, such as lane reduction and lane narrowing would somehow reduce the number of accidents within the corridor.

Statement of the Problem

In order to test the CMP assertion that 30% of the accidents that occur along the US 50 corridor between the SR28 intersection and Stateline are attributable to excessive speed[1], we must formulate the problem in such a way that it can be analyzed rigorously using statistical hypothesis testing techniques and applied to the reported accident data. Since the CMP assertion has been used to promote the use of speed reduction measures such as lane reduction and narrowing, it is reasonable to assume that the 30% *probability metric* applies to accidents where speed in excess of the posted limit is at fault. The accident reports, from which the accident data in Table 1 were taken[2], show two categories of excessive speed. The first category, labeled "Unambiguous Speeding," includes only those reported accidents where the root cause of the accident was speed in excess of the posted limit. The second category, labeled "Possible Speeding," also includes the unambiguous speeding accidents but adds those accidents labeled "Too Fast for Road Conditions" but where those conditions could not be correlated with " Wet, Ice, Snow, or Slush" conditions, based on the way these accidents were reported. These category definitions allow us to establish two statements of the problem, which are described precisely below, when we formulate them as *null hypotheses*.

Note that all of the parameters we used tend to favor accepting the CMP assertion, thereby giving the benefit of any doubt to the CMP team.

Table 1. Summary of US 50 Accident Data from SR28 to Stateline



Summary of US 50 Accident Data from SR28 to Stateline

US 50 Road Segment	Description of Segment	Total Accidents	Unambiguous Speeding Accidents	Posted Speed Limit	Too Fast for Road Conditions	Wet, Ice, Snow or Slush	Possible Speeding Accidents	Dry	Other Highway
Total Corridor	SR28 to Stateline	415	17		110	94	33	17	1
Segment 1	SR28 to Glenbrook	44	1	50	20	18	3	1	0
Segment 2	Glenbrook to Friedhoff Rd	57	0	45	19	14	5	0	0
Segment 3	Cave Rock to Skyland	92	4	45	28	27	5	4	0
Segment 4	Skyland to Round Hill Pines	116	7	45	25	24	8	7	0
Segment 5	Round Hill Pines to Kingsbury Grade	49	4	35-45	11	9	6	4	1
Segment 6	Kingsbury Grade to Stateline	57	1	25-35	7	2	6	1	0
Intersections (removed from Total Corridor)	Total of Intersections SR 20 to Stateline	119	0		17	7	0	3	0

Notes:

1. All traffic accident data taken from accident reports between 2016 1ne 2020.
2. Total Corridor row is the total for reported accidents across all six segments of the corridor, less accidents that occurred in intersections.
3. "Unambiguous Speeding" includes only those accidents where speed in excess of the posted limit is the root cause, and the road surface was reported to be dry.
4. "Possible Speeding" includes accidents reported as speed in excess of the posted limit PLUS accidents reported as speed too fast for road conditions MINUS those accidents where wet, ice, snow or slush was also reported.

Bernoulli Random Variables

A *Bernoulli random variable* is one that takes on the value of 1 or 0 with probabilities p and 1-p, respectively. A classic Bernoulli random variable problem is determining whether a coin flip is "fair" in the statistical sense, where 1 denotes heads and 0 denotes tales. The assertion would be "the coin is fair," that is, heads does come up in 50% of the flipping events. This assertion can be tested by flipping the coin a number of times and recording the number of times heads come up, which is a random variable -- the probability of heads.

We can approach the US 50 accident problem the same way, where the number of accidents satisfying the CMP assertion is treated as a Bernoulli random variable. Each accident can have a value of either 1, corresponding to an accident that is attributable to excessive speed; or 0, corresponding to an accident attributable to some other cause. This is a special case of a *binomial distribution*, where n = 1 and p is the probability that an accident falls into the former category, unambiguous speed-over-limit.

$$p_X(k) \ = \ P\{X = k\} \ = \ \frac{n!}{k!\,(n-k)!} p^k (1-p)^{n-k}, for\ k \in \ (0,1)\ and\ n = 1$$

$$p_X(0) \ = \ 1 - p$$

$$p_X(1) = p$$

The mean of this distribution is

$$E(X) \ = \ np$$

The variance of this distribution is

$$Var(X) = np(1-p)$$

This formulation tends to be unwieldy when large samples (like the 414 reported accidents in our analysis) due to the magnitude of the factorials involved. With such large numbers, it is appropriate to approximate the binomial distribution with a normal distribution. According to Hald[3], the normal approximation is good when the following inequality holds:

$$np(1-p) > 9$$

In this case, np(1-p) = 414*0.3*(1-0.3) = 86.9, which is quite adequate.

Hypotheses

A precise statement of an assertion is called a *null hypothesis* in engineering statistics. In this case, two null hypotheses can be formulated which we will denote as $H_0(1)$ and $H_0(2)$. These are stated below with their alternative hypotheses $H_A(1)$ and $H_A(2)$, respectively:

First Hypothesis set:

$H_0(1)$ = "30% of all accidents along the segment of the US 50 East Shore corridor between SR28 and Stateline are unambiguously related to speed in excess of the posted limit."

$H_A(1)$ = "Less than 30% of all accidents along this segment are unambiguously related to speed-in-excess."

Second Hypothesis set:

$H_0(2)$ = "30% of all accidents along the segment of the US 50 East Shore corridor between SR28 and Stateline are possibly related to speed in excess of the posted limit. These include all accidents that cannot be ruled out based on wet, ice, snow or sleet conditions."

$H_A(2)$ = "Less than 30% of all accidents along this segment are possibly related to speed-in-excess."

Random Sample

We can consider the set of reported accidents a "random sample" if it is representative of the entire process (all accidents along the corridor or segment) and no data have been removed solely because they either favor or disfavor the assertion. Data can be removed, however, if there is a problem in the data collection and this brings up a "hanging chad" problem, which is not addressed in this paper. In our analysis, we did not remove any reported accident data, except those representing intersections, none of which were attributed to excessive speed. We believe including those data would have biased the analysis against the CMP assertion. Again, this is giving the benefit of the doubt to the CMP team.

Confidence Level

Confidence level is a statistical term with a precise meaning. A 95% confidence level means that there is only a 5% chance of concluding that an assertion is true if it isn't. This is referred to as a *Type-I error*. We will use the rather lax 95% metric, since the down-side of a Type-I error is relatively minor (no one will be killed if we draw the wrong conclusion). The metric α is the probability that a Type-1 error could occur, i.e., that the alternative hypothesis is actually true.

$$\alpha \; = \; 1 - 0.95 \; = \; 0.05$$

Definition of Statistic

We can use the N accident reports as random samples from the process with results denoted $X_1,X_2,\cdots,X_N$. We then compute a statistic, Y, such that Y is the total number of sample outcomes that meet the null hypothesis:

$$Y = \sum\nolimits_{i=1}^{N} X_i$$

Y has as binomial distribution, where the probability of having exactly k sample outcomes that meet the hypothesis is:

$$P_Y(k) = \frac{N!}{k!\,(N-k)!} P^k (1-P)^{N-k}$$

Left Sided Test

What we have set up is a "left-sided test," wherein the "alternate hypothesis" asserts that the true P_Y is less than the value P_0 claimed by the null hypothesis, i.e., that the number of accidents unambiguously attributable to "speed in excess" is less than 30%.

Under the null hypothesis, the expected number of positive outcomes in N trials is NP_0. We will *reject* the null hypothesis if $Y \leq \theta_Z$, i.e. Y is less than or equal to a threshold θ_Z that is sufficiently larger than NP_0.

The probability of Type-1 error, using a threshold θ, is the probability of obtaining $Y > \theta$ under the alternate hypothesis, and is given by

$$\boxed{\alpha(\theta) \; = \; P\{\theta < X \leq N\} = \sum\nolimits_{k=\theta}^{N} \frac{N!}{k!\,(N-k)!} P_0^k (1-P_0)^{N-k}}$$

For the left-sided test, we choose threshold θz as the smallest θ such that $\alpha(\theta) \leq \alpha_Z$.

As stated above, our sample size is sufficiently large that we can use a *continuous* normal distribution with a *correction for continuity* to approximate the *discontinuous* binomial distribution with acceptable accuracy, where[4]:

$$\alpha(\theta) \ = P\{\theta \leq X \leq N\} \cong \int_{\theta-1/2}^{N+1/2} \frac{1}{\sqrt{2\pi NP(1-P)}} \ e^{-(z-NP)^2/2NP(1-P)} dz$$

We need to convert this to the form of the Excel function, NORM.DIST:

NORM.DIST(x,mean,standard_dev,cumulative)

where the functional form is:

$$f(x,\mu,\sigma,TRUE) \ = \ \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} dx$$

with the following substitutions, from before:

$$\mu = NP$$

$$\sigma \ = \ \sqrt{VarX} = \sqrt{NP(1-P)}$$

we obtain:

$$\alpha(\theta) = \ P\{\theta \leq X \leq N\} \cong \int_{-\infty}^{N+1/2} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(z-\mu)^2}{2\sigma^2}} dz - \int_{-\infty}^{\theta-1/2} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(z-\mu)^2}{2\sigma^2}} dz$$

$$\boxed{\alpha(\theta) \ \cong f(N+0.5, NP, \sqrt{NP(1-P)}, TRUE) - f(\theta-0.5), NP, \sqrt{NP(1-P)}, TRUE)}$$

For the left-sided test, we choose threshold θz as the largest θ such that $\alpha(\theta) \leq \alpha_Z$.

Conclusions

Given a sample size of 415 accidents across the entire US 50 corridor between SR28 and Stateline, to assert that 30% of these accidents are unambiguously attributable to speed in excess of the posted limit with a 95% confidence level would require 141 accidents satisfying the assertion. Since only 17 satisfy the assertion, the null hypothesis, $H_0(1)$ is therefore rejected.

Similarly, since only 33 satisfy the assertion that 30% of the accidents may be attributable to speed in excess of the posted limit, where we add accidents reported as "too fast for conditions" but where these conditions are unspecified, and with a 95% confidence level, the null hypothesis $H_0(2)$ is rejected.

We can look individually at the 6 road segments within the east shore corridor of US 50 and apply the same hypothesis testing methodology. Table 2 summarizes these results and conclusions. In all cases the null hypotheses are rejected.

Table 2: Hypothesis Testing Results and Conclusions

Unambiguous Speeding

US 50 Road Segment	Description of Segment	Total Accidents	Unambiguous Speeding Accidents	Accidents Needed to Accept Hypothesis	Conclusion
Total Corridor	SR28 to Stateline	415	17	141	Hypothesis Rejected
Segment 1	SR28 to Glenbrook	44	1	19	Hypothesis Rejected
Segment 2	Glenbrook to Friedhoff Rd	57	0	24	Hypothesis Rejected
Segment 3	Cave Rock to Skyland	92	4	36	Hypothesis Rejected
Segment 4	Skyland to Round Hill Pines	116	7	44	Hypothesis Rejected
Segment 5	Round Hill Pines to Kingsbury Grade	49	4	21	Hypothesis Rejected
Segment 6	Kingsbury Grade to Stateline	57	1	24	Hypothesis Rejected

Possible Speeding

US 50 Road Segment	Description of Segment	Total Accidents	Possible Speeding Accidents	Accidents Needed to Accept Hypothesis	Conclusion
Total Corridor	SR28 to Stateline	415	33	141	Hypothesis Rejected
Segment 1	SR28 to Glenbrook	44	3	19	Hypothesis Rejected
Segment 2	Glenbrook to Friedhoff Rd	57	5	24	Hypothesis Rejected
Segment 3	Cave Rock to Skyland	92	5	36	Hypothesis Rejected
Segment 4	Skyland to Round Hill Pines	116	8	44	Hypothesis Rejected
Segment 5	Round Hill Pines to Kingsbury Grade	49	6	21	Hypothesis Rejected
Segment 6	Kingsbury Grade to Stateline	57	6	24	Hypothesis Rejected

References

1. S. Morrow, Nevada Department of Transportation. NDOT Crash Data 2016-2020. https://ndot.maps.arcgis.com/apps/webappviewer/index.html?id=00d23dc547eb4382bef9beabe07eaefd. Accessed June, 2023.

2. Wood Rogers Consulting Firm (under contract to Nevada Department of Transportation), "Final Existing Conditions Memorandum," https://www.dot.nv.gov/home/showpublisheddocument/20142/637781886697730000, p. ix, July 27, 2021.

3. A. Hald, *Statistical Theory with Engineering Applications*, John Wiley & Sons, New York, 1952.

4. A. Bowker and G. Lieberman, *Engineering Statistics*, Prentice-Hall, Inc., Englewood Cliffs, NJ, pp. 127-128, 1972.

Author Biographies

Robert Byren: Mr. Byren is a retired electrical engineer with over 40 years professional experience in military lasers, laser radar, beam control, adaptive optics, thermal imaging, and optical metamaterials. Prior to retirement, he served as Chief Technologist for Raytheon's Space and Airborne Systems business unit with responsibility for the senior technical staff, intellectual property, innovation, and university relations. Post retirement, he led a small consulting firm in the field of high energy laser systems. Mr. Byren holds 43 US Patents and has co-authored 55 books and technical papers. He received his MSEE degree from Stanford University in 1975 and his BSEE degree from Lehigh University in 1974**.**

Sydney Morrow: Sydney Morrow is a full-time resident of the East Shore. She received her Ph.D. from the University of Texas Graduate School of Biomedical Sciences in 1986. Following a long career with the Department of Veterans Affairs, she and her husband relocated to Lake Tahoe in 2015. Sydney has been actively involved with wildfire prevention and serves as the FireWise coordinator for Glenbrook.