Statistical Thinking for Forensic Practitioners

Lab on Part 9: Reporting & Testimony

In the Part 9 lecture, Dr. Kaplan discussed a 2018 study from Thompson et al. that analyzed how lay people evaluate the strength of possible reporting statements relative to one another. It is important to understand how, for example, members of a jury interpret a conclusion made by a forensic practitioner. For example, might a jury's verdict change if a forensic practitioner reports their conclusion about an important piece of evidence as a likelihood ratio (LR) as opposed to a random match probability (RMP)? Are two types of equally-probative evidence interpreted differently by the jury? Another study performed by Thompson and Newman in 2015 called *Lay Understanding of Forensic Statistics: Evaluation of Random Match Probabilities, Likelihood Ratios, and Verbal Equivalents* considered how verdicts were affected by the following parameters:

1. The nature of the forensic evidence presented (DNA or shoeprint)
2. The strength of the forensic evidence (very strong or moderate)
3. The presentation format used by the forensic practitioner to describe the strength of forensic evidence (Random Match Probability (RMP), Likelihood Ratio (LR), or Verbal Equivalents (VE)).

The study was performed via Amazon Mechanical Turk, which is an online labor pool in which individuals perform tasks for pay. In this study, 517 participants considered a hypothetical case involving a woman who was sexually assaulted in her home. The fabricated defendant in the case, Brian Kelly, became a suspect because he was spotted near the crime scene, although the victim could not identify him as her assailant. Evidence was found at the crime scene and compared to reference samples taken from the defendant. This evidence was either DNA found on a faucet or a shoeprint found on the bathroom floor (randomly chosen for each participant). The participants were given one of a series of randomly assigned conclusions made by a forensic expert regarding the type and strength of the evidence. For example, the conclusion associated with a moderate strength Random Match Probability conclusion based on recovered DNA evidence read as:

> "Based on the scientific data on the genetic characteristics of the human population, I estimate that approximately one in one hundred people has a DNA profile that is consistent with the partial DNA profile on the faucet. That means that there is one chance in one hundred of finding a consistent profile in a randomly chosen person."

The very strong conclusions replaced "one in one hundred" as "one in a million." The conclusions based on Likelihood Ratios were made in terms of how likely the evidence was if it "came from Brian Kelly than if it came from a randomly chosen person." The Verbal Equivalents conclusions stated whether the evidence provides "moderate" or "very strong" support for the hypothesis that the evidence came from Brian Kelly.

Based on this conclusion, the participants were asked to reach a verdict. We will consider verdict data from this study (note: these data were simulated based on the results given in paper, but they look qualitatively similar to the original data).

Open the `verdict.xlsx` file found on the course website. The following describes the contents of the data set. Each row represents a single participant in the study. The columns of the data set are:

- `format`: presentation format used to describe the strength of evidence.
- `type`: nature of the forensic evidence.

- `strength`: strength of the forensic evidence.
- `formatLR`: indicator column where 1 means that the Likelihood Ratio format was used to describe the strength of evidence and 0 means that a different format was used.
- `formatRMP`: indicator column where 1 means that the Random Match Probability format was used to describe the strength of evidence and 0 means that a different format was used.
- `typeDNA`: indicator column where 1 means that DNA evidence was considered while 0 means that shoeprint evidence was considered.
- `strengthVeryStrong`: an indicator column where 1 means that the strength of evidence was Very Strong while 0 means that it was Moderate.
- `verdict`: whether the participant reached a Guilty or Not Guilty verdict.

Best practice dictates that we should formulate our research questions/hypotheses before collecting/analyzing data. In this problem, we are interested in how the proportion of Guilty verdicts compare across different types of conclusions.

1. Formulate a research question and/or hypothesis that addresses how the proportion of Guilty verdicts compare between two types of conclusions. (e.g., "I hypothesize that the proportion of Guilty verdicts associated with conclusions that involve [...describe some subset of conclusions...] is [less than/greater than/different from] the proportion of Guilty verdicts associated with conclusions that involve [...describe some other subset of conclusions...].")

   We will consider the hypothesis that proportion of Guilty verdicts associated with Verbal Equivalents conclusions is less than the proportion of Guilty verdicts associated with Random Match Probability conclusions. Your hypothesis/research question might differ from this one, which means your answers will likely differ for the next few questions. If you would like more direct feedback on your answers to this assignment you can send your work to csafelearning@iastate.edu.

We are interested in analyzing how the `verdict` variable changes with the `format`, `type`, and `strength` variables. A powerful tool for quantitatively summarizing data in Excel is called a Pivot Table.

2. Create a new pivot table by navigating to `Insert` then `Pivot Table`. Highlight the data set range of cells if this is not automatically done for you. Click `OK`.
3. To add variables to the pivot table, click the checkboxes next to `format`, `type`, and `strength`. These should populate the `Axis (Categories)` box in the bottom left corner of the toolbar. Click and drag `verdict` to the `Values` box in the bottom right corner of the toolbar. Then, click and drag `verdict` to the `Columns` box above the `Values` box. You should now see the table populated with verdict counts (i.e., Guilty and Not Guilty as columns) for each combination of format, type, and strength (as rows).
4. Insert a `100% Stacked Columns` chart summarizing the verdict counts across the 3 variables. What does this plot indicate about the research question your formulated in Question 1?

   The Guilty verdict bar heights for the `RMP` group do appear to be slightly taller than those of to the `VE` group. It's difficult to say from the plot whether they are "significantly" taller, so we can perform a formal hypothesis test to answer this question. This plot would be more useful if the hypothesis of interest was more specific (e.g., RMP + Moderate + DNA has a higher proportion of Guilty verdicts than RMP + Moderate + Shoeprint).

5. Perform a hypothesis test at the $\alpha = 0.10$ level to answer the research question/hypothesis posed above (note the difference from $\alpha = 0.05$ tests you've performed on recent assignments). Your answer should include null and alternative hypotheses for the population parameters of interest, a test statistic, a determination of a critical value or p-value (or both), and a conclusion. (Note: Refer back to the material covered in Video 5 of the Part 6 slides. The Pivot Table contains all of the information needed to start the hypothesis test.)

Let π_{RMP} and π_{VE} be the proportion of the Guilty verdicts associated with the RMP and VE conclusions, respectively. Of the 177 participants who received an RMP conclusion, 33 of them rendered a Guilty verdict meaning $\hat{\pi}_{RMP} = .186$. Similarly, $\hat{\pi}_{VE} = 23/171 = .135$. The hypotheses are $H_0 : \pi_{RMP} = \pi_{VE}$ vs. $H_a : \pi_{RMP} > \pi_{VE}$. The estimate for the common population proportion is then $\hat{p} = \frac{33+23}{177+171} = .161$. Then the standard error is $SE_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})}\sqrt{1/n_1 + 1/n_2} = .040$. This means the z-statistic is $z = \frac{\hat{\pi}_{RMP} - \hat{\pi}_{VE}}{SE_{\hat{p}}} = 1.318$. Comparing this to the one-sided critical value $z_{1-\alpha} = z_{.9} = 1.282$ means that we reject the notion that the proportion of Guilty verdicts associated with RMP conclusions is equal to that of the VE conclusions. The p-value associated with this test is .094 which is significant at the $\alpha = 0.10$ level.

6. Regardless of the statistical conclusions reached from your hypothesis test, comment on the *practical* significance of the results as it pertains to your research question/hypothesis. Posit an explanation of why the results are/are not practically significant (i.e., can you think of a logical justification for why the results turned out this way?).

   The fact that the proportion of Guilty verdicts associated with RMP conclusions is 5% higher than that of VE conclusions in these data is arguably practically significant. This might indicate that lay people put more weight on the RMP conclusions, perhaps because the RMP conclusions are more precise and/or more easily interpreted than the VE conclusions.

7. What are aspects of the study design that might affect how applicable these conclusions are to the study's overall goals?

   One could certainly comment on/critique the particulars of the case that the participants were asked to consider (although the authors did formulate the case based on cases used in previous studies). One major issue with generalizing the conclusions drawn in this study to, for example, all lay people, is that the participants were paid workers hired through the Amazon Mechanical Turk service. There is no reason to assume that such individuals are representative of all lay people.

In the original paper, the authors analyzed the proportion of Guilty verdicts (more precisely, the proportion of Not Guilty verdicts) using a logistic regression model.

8. Why might a logistic regression model be appropriate to analyze the proportion of Guilty verdicts?

   The verdict is a binary outcome, presumably independent between different participants.

9. Propose a logistic regression model for the proportion of Guilty verdicts using the format, type, and strength of a conclusion as predictors. (Hint: recall from lecture slides on Part 7 that we use specific variables in our model to represent categorical predictors. Consider the variables in the `verdict.xlsx` file to inform how you should formulate your model.)

   We need to use dummy/indicator variables to represent the 3 categorical predictors similar to the `formatLR`, `formatRMP`, `typeDNA`, and `strengthVeryStrong` columns. Assume y_i is the verdict associated with the ith participant. We assume each y_i are independently distributed as Binomial$(1, \pi_i)$, $i = 1, ..., 571$, where

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_{LR,i} + \beta_2 x_{RMP,i} + \beta_3 x_{DNA,i} + \beta_4 x_{VS,i}}}{1 + e^{\beta_0 + \beta_1 x_{LR,i} + \beta_2 x_{RMP,i} + \beta_3 x_{DNA,i} + \beta_4 x_{VS,i}}}.$$

   where $x_{LR,i}, x_{RMP,i}, x_{DNA,i}, x_{VS,i}$ take on value 1 if the ith participant receives a conclusion in terms of an LR, in terms of an RMP, concerning DNA evidence, or that is indicated to be Very Strong, respectively, and 0 otherwise.

10. The authors were also interested in identifying whether the effect of one variable differed based on the levels of another variable. For example, does the effect that a Very Strong conclusion has on the proportion of Guilty verdicts differ if the associated conclusion is phrased in terms of a Random Match Probability as opposed to a Verbal Equivalence? Define a new variable that could be added to the model to address this example research question.

The research question can be interpreted as asking whether an interaction exists between the `strength` variable and the RMP and VE categories of the `format` variable. By how the variables in Question 9 are defined, the VE category is taken (arbitrarily) as the "default" value of the `format` variable, which is why it doesn't have an associated dummy variable. We will define an interaction variable, call it z_i, such that $z_i = x_{RMP,i} * x_{VS,i}$, which takes on value 1 only if a participant receives a Very Strong Random Match Probability conclusion (of either evidence type). So the associated regression coefficient, β_5 say, will capture the difference between a participant receiving a Very Strong RMP conclusion and one receiving a Very Strong VE conclusion.

11. A participant receives a Moderate Random Match Probability DNA conclusion. They conclude: "The evidence has little value proving Brian Kelly is guilty because a lot of other people could have left DNA." Comment on the validity of this conclusion.

    This statement is not necessarily true - it is entirely possible that the DNA evidence could have probative value for incriminating the defendant even if someone else could match. This is sometimes referred to as the "defense attorney's fallacy."

12. Two different participants receive a Very Strong Random Match Probability DNA conclusion. One concludes: "So there is a 1 in a million chance that the evidence was left by someone other than Brian Kelly - that must mean he's guilty!" The other participant concludes: "So there is a 1 in a million chance that a randomly selected person could leave DNA that matches the one at the crime scene - that must mean Brian Kelly is guilty!" Which participant is more justified in their conclusion? Explain.

    The correct interpretation of the random match probability statement is the latter. The problem with the first statement can be interpreted as reversing the conditional events "Brian is the source | Evidence" and "Evidence | Brian is the source.". In particular, the first participant is making a judgement about "Brian is Guilty | Evidence" when what the conclusion actually supports is "Evidence | Brian is the source". This makes a leap from "Brian is the source" to "Brian is guilty" which may be unfounded and, like the prosecutor's fallacy, draws a false equivalence between the two conditional statements. In either case, more than one piece of evidence should be considered before reaching a conclusion.