Workbook: Statistical Thinking for Forensic Practitioners

Sam Tyner, Hal Stern, Alicia Carriquiry

Contents

Chapter 1

Introduction

This workbook is intended to accompany the Statistical Thinking for Forensic Practitioners workshop taught by members of the Center for Statistics and Applications in Forensic Evidence (CSAFE). The slides for this workshop were constructed initially by Dr. Hal Stern of UC-Irvine and Dr. Alicia Carriquiry of Iowa State University.

When taking the workshop, please follow along with the slides handout (if given) and this workbook. The workbook contains the same material as the slides, with room for you to take notes and to fill in the missing material.

Chapter 2

Statistical Preliminaries

Briefly, this section contains a broad review of probability concepts and of statistical inference concepts, with examples from the forensic science context. We will cover probability, data collection, statistical distributions, estimation, and hypothesis testing.

2.0.1 Definitions

- **population**: ____

- **sample**: ____

- **probability**: Using knowledge about the ________________________ to make statements describing the ____________________________. Probability can loosely be thought of as a type of deductive reasoning, where we are applying general knowledge about the population of interest to make conclusions about a small part of that population.

- **statistics**: Using knowledge about the ________________________ to make statements describing the ____________________________. Statistics can loosely be thought of as a type of inductive reasoning, where we are applying knowledge about a sample to state that something *may* be true about the population generally.

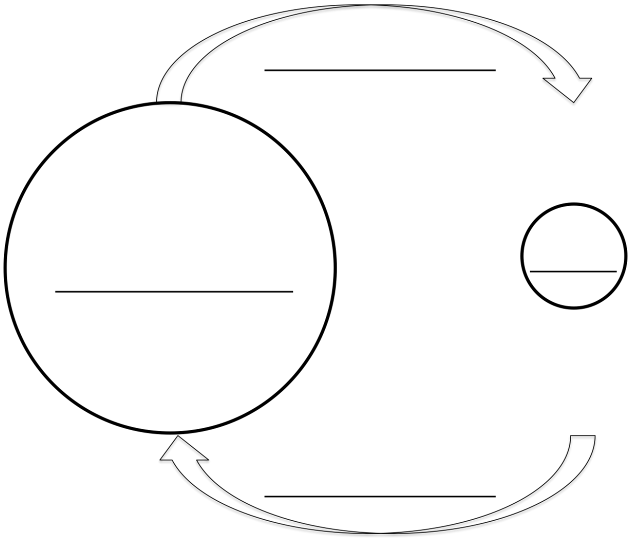

Figure 2.1: "The Big Picture"

2.0.2 Forensic Science Examples

- Suppose 100 1-pound bags of cocaine are seized on the US-Mexico border, and the FBI want to know the chemical composition of the confiscated drugs to store in their database.
    - Population: ______________________________
    - Sample: ______________________________
- A window was broken in a robbery, and the suspect who was apprehended nearby had glass fragments lodged in the soles of their shoes. Do the fragments from the suspect's shoes have the same or similar chemical composition as the broken window?
    - Population 1: ______________________________
    - Sample 1: ______________________________

    - Population 2: ______________________________
    - Sample 2: ______________________________
- A city government employee is suspected of embezzling funds from the city's coffers. Forensic accountants examine a subset of the city's transactions to determine whether embezzling occurred and how much money was lost.
    - Population: ______________________________
    - Sample: ______________________________

How do you think this pertains to pattern evidence? List some possible relevant populations and samples below.

- Population 1: ______________________________
- Sample 1: ______________________________

- Population 2: ______________________________
- Sample 2: ______________________________
- Population 3: ______________________________
- Sample 3: ______________________________

2.1 Probability

Probability concerns the *uncertainty* of outcomes. The set of all possible outcomes is called the _____________ space, and a particular outcome or set of outcomes of interest is referred to as an ____________.

2.1.1 Examples

1. Footwear
    - Sample Space = All shoe sizes e.g. $\{6, 6.5, 7, 7.5, 8, 8.5, \dots\}$
    - Event = Shoe of size 9

2. Footwear
   - Sample Space = Brand of shoe e.g. { Nike, Vans, Converse, ...}
   - Event = Nike sneaker
3. Firearms
   - Sample Space = CMS (consecutive matching striae) for a pair of bullets e.g. $\{0, 1, 2, 3, 4, \dots\}$
   - Event = CMS of 10 or more

2.1.2 Interpretation

The probability of observing an event in a sample space is a number less than or equal to 1 and greater than or equal to 0 that describes the __________________ that the event will occur.

There are two primary interpretations of probability:

1. The long run __________________ of occurrence of an event.
2. The ___________________ belief of likelihood of an event occurring.

2.1.3 Basic Notation and Laws of Probability

Let an event of interest be denoted by ___________. The probability of this event occurring is then denoted _______________. Recall that the probability of an event is always between 0 and 1. When $P(Y) = 0$, the event Y will never happen. When $P(Y) = 1$, the event Y will always happen. The sum of the probabilities of all possbile outcomes in the sample space always equal to ________.

The event of interest, Y, also has a complement event, $\overline{Y}$, which is read as "not Y". The complement, $\overline{Y}$, of an event, Y, is itself an event containing all outcomes in the sample space other than that initial event of interest, Y.

$$P(Y) + P(\overline{Y}) = \underline{\qquad}$$

The above equation also gives us the following rules:

$$\begin{aligned} P(Y) &= 1 - P(\overline{Y}) \\ P(\overline{Y}) &= 1 - P(Y) \end{aligned} \tag{2.1}$$

2.1.4 Probability and Odds

The probability of an event defines the odds of the event. The odds *in favor* of an event Y are defined as the probability of Y divided by the probability of everything except Y ("not Y"):

$$O(Y) = \frac{P(Y)}{P(\overline{Y})} = \frac{P(Y)}{1 - \underline{\quad}}.$$

Conversely, the odds *against* a event Y are defined as the porbability of everything except Y ("not Y") divided by the probability of Y:

$$O(\overline{Y}) = \frac{P(\overline{Y})}{P(Y)} = \frac{1 - \underline{\quad}}{P(Y)}.$$

When we typically talk about odds, like in horse racing, the odds reported are the odds *against* the outcome of interest. Let's construct a horse race scenario using our probability notation to find the probability of a horse winning a race from the reported odds:

- Suppose you want to place a bet on a horse name Cleopatra winning the race. Odds for Cleopatra are reported as 4:1.
- Y = Cleopatra wins the race
- $\overline{Y}$ = Any horse in the race *other than* Cleopatra wins the race.
- $O(\overline{Y}) = \dfrac{P(\overline{Y})}{P(Y)} = \frac{4}{1} = 4$
- We know that $P(Y)+P(\overline{Y}) = 1$. With this information, we can determine $P(Y)$, which is the probability that Cleopatra wins the race:

$$\begin{aligned}
O(\overline{Y}) &= \frac{P(\overline{Y})}{P(Y)} = 4 \\
\Rightarrow \frac{P(\overline{Y})}{P(Y)} &= 4 \\
\Rightarrow \frac{1 - P(Y)}{P(Y)} &= 4 \qquad \textit{(See Equation 2.1)} \\
\Rightarrow \frac{1}{P(Y)} - 1 &= 4 \\
\Rightarrow \frac{1}{P(Y)} &= 5 \\
\Rightarrow P(Y) &= \frac{1}{5} = 0.2 \\
\Rightarrow P(\overline{Y}) &= 0.8
\end{aligned}$$

- So, the odds for Cleopatra (4:1) mean that Cleopatra has a probability of 0.2 of winning the race. Because this outcome is not very likely (it will only happen in 1 race out of 5), you win money if Cleopatra wins simply because that is not a likely outcome.
- **Betting**: Suppose you bet \$1 on Cleopatra to win the race with 4:1 odds. You will win \$4 if Cleopatra wins, otherwise you've lost \$1.
- The amount you win (\$4) is determined so that you break even in the long run.
- Suppose 5 identical races are run. In 1 of those races, Cleopatra wins, and in the other 4, Cleopatra loses. If you bet \$1 on Cleopatra in each race, you will lose that \$1 4 of 5 times. So, in order for you to break even, the designated amount you'll win when Cleopatra wins is \$4.
- This is a statistical concept known as *expected value*. Your expected value when placing the bet is \$0. We compute expected value by multiplying each possible outcome value by its probability and adding them all together:

$$\begin{aligned}
\$4 \cdot P(Y) + (-\$1) \cdot P(\overline{Y}) &= 0 \\
\$4 \cdot 0.2 + (-\$1) \cdot 0.8 &= 0 \\
\$0.8 - \$0.8 &= 0
\end{aligned}$$

2.1.5 Probability Math

Up until now, we have only considered one event, Y. Now, suppose we have another event that we are interested in, Z.

Let's consider the possibility of *either* of these two events, Y and Z, occurring. We'd write this as $Y \cup Z$, which is mathematical notation for "Y or Z occurs". There are two scenarios that arise:

1. Y and Z cannot occur together: they are _____________ _______________
2. Y and Z can occur together.

In scenario #1, computing the probability of either Y or Z happening is easy: we just add their respective probabilities together:

$$Y, Z \text{ mutually exclusive } \Rightarrow P(Y \cup Z) = P(Y) + P(Z)$$

In scenario #2, computing the probability of either Y or Z happening is more complicated because we know there is a chance that Y and Z can happen together. We'd write this as $Y \cap Z$, which is mathematical notation for "Y and Z occurs". In scenario #1, this event never occurred, so $P(Y \cap Z) = 0$ there. To compute the probability of Y or Z occurring in scenario #2, we have to consider the probability of Y, the probability of Z, and the probability of $Y \cap Z$. If we just add $P(Y) + P(Z)$ as in scenario #1, we include the event $Y \cap Z$ twice, so we have to subtract one instance of it:

$$Y, Z \text{ not mutually exclusive } \Rightarrow P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z).$$

This probability is much easier to think about when illustrated. In Figure 2.2, we consider human blood types. There are four groups: A, B, O, and AB, and there are two RH types: + and −. We first consider the blood types A and B, represented by the two non-overlapping circles. Define:

- Event Y = a person has blood type A
- Event Z = a person has blood type B
- Event $Y \cup Z$ = a person has blood type A or blood type B

These two events are *mutually exclusive* because one person cannot have both blood type A and blood type B. (The circles don't overlap in the venn diagram) So, the probability that a randomly selected person has blood type A or B is:

$$P(Y \cup Z) = \_\_\_\_ \quad + \quad \_\_\_\_$$

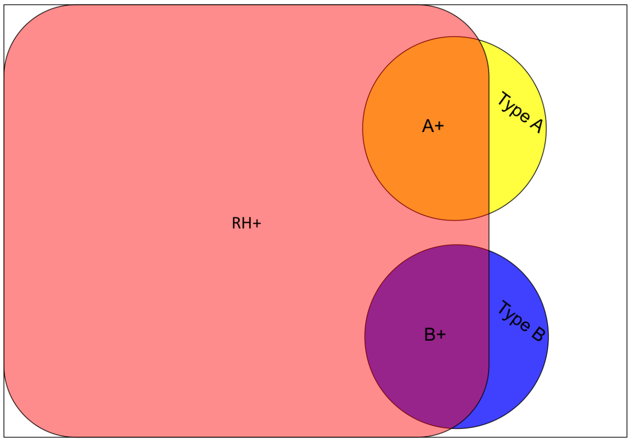

Figure 2.2: Probabilities of blood types in humans

Return to Figure 2.2 and consider two other events: a person having blood type A or having the Rh factor (RH+). We see in Figure 2.2 that someone can have both type A blood and the Rh factor (blood type A+). Define:

- Event Y = a person has blood type A
- Event Z = a person has the Rh factor

- Event $Y \cup Z$ = a person has blood type A or the Rh factor
- Event $Y \cap Z$ = a person has blood type A and the Rh factor (they have A+ blood)

So, the probabilty that someone has either type A blood or has the Rh factor is the sum of probability of having type A blood (represented by the yellow circle) and the probability of having the Rh factor (represented by the red rectangle) minus the probability of having A+ blood (represented by the orange area of overlap that is counted twice) in Figure 2.2. So, the probability that a randomly selected person has blood type A or the Rh factor is:

$$P(Y \cup Z) = \underline{\qquad} \quad + \quad \underline{\qquad} \quad - \quad \underline{\qquad}$$

2.1.6 Conditional Probability

Let's consider an event of interest Y which has probability $P(Y)$. Then, suppose we learn of another event of interest Z that has occurred. Knowing that Z has occurred already may change our opinion about the likelihood of _______ occurring. The key idea here is that the probability of an event often depends on other information, leading us to the definition of *conditional probability*:

$$P(Y|Z),$$

which is the conditional ____________________ that Y occurs given that we know Z has occurred. Return to Figure 2.2. Suppose we want to know the probability of a person having type A blood, represented by the yellow circle. But, if we already know that a person has the Rh factor, we are only interested in the part of the type A circle that overlaps with the Rh+ rectangle. Thus the probability of having type A blood is different with different knowledge. The formula for calculating conditional probability is:

$$P(Y|Z) = \frac{P(Y \cap Z)}{P(Z)} \tag{2.2}$$

Returning to the venn diagram, the value $P(Y \cap Z)$ is represented by the overlap of the type A circle and the Rh+ rectangle, and the value $P(Z)$ is represented by the Rh+ rectangle. Then, the value $P(Y|Z)$ is the ratio of the overlap (A+) to the Rh+ rectangle.

Equation 2.2 also gives us a multiplication rule for computing probabilities:

$$P(Y \cap Z) = P(Y|Z) \cdot P(Z) \tag{2.3}$$

Philosophically speaking, it can be helpful to think of *all* probabilities as conditional. It is just a question of what information is assumed to be ________________.

2.1.6.1 Examples

Death Penalty Convictions

A study of sentencing of 362 black people convicted of murder in Georgia in the 1980s found that 59 were sentenced to death (Baldus, Pulaski, and Woodworth (1983)). They also examined the race of the murder victim, either black or white, and found some disparities. In Table 2.1, DP means the defendant received the death penalty, NDP means the defendant did not receive the death penalty. The race of the victim (RV) is either black (B) or white (W).

Returning to Section 2.0.1, let's define the problem:

- **Population**: All black people convicted of murder in Georgia in the 1980s
- **Sample**: N/A (the whole population was studied)

RV	DP	NDP	Total
W	45	85	130
B	14	218	232
Total	59	303	362

Table 2.1: The results of the Baldus et al study for black defendants convicted of murder.

Using the numbers from Table 2.1, compute the following probabilities:

- $P(DP) = \text{—} = 0.$_____
- $P(DP|RV = W) = \text{—} = 0.$_____
- $P(DP|RV = B) = \text{—} = 0.$_____

Note: These numbers are selected from the study, and should not be considered a comprehensive summary of its results. There are a number of things not discussed here. The entire publication can be found online[1]

Consecutive Matching Striae

In firearms and toolmark analysis, the number of consecutive matching striae (CMS) between a crime scene sample and a lab sample is often used to help determine a match. Generally speaking, the higher the maximum number of CMS found in a pair, the more likely the two samples came from the same source. Several known match (KM) pairs and known non-match (KNM) pairs of bullets were examined, and the results are shown in Figure 2.3 (Hare, Hofmann, and Carriquiry (2017)). What is the probability of seeing two known matches (or two known non-matches) given the maximum number of CMS? Here, we condition on ____. Again, we briefly return to Section 2.0.1, let's define the problem:

- **Population**: All pairs of fired bullets from unknown sources
- **Sample**: A sample of pairs of known matches and known non-matches

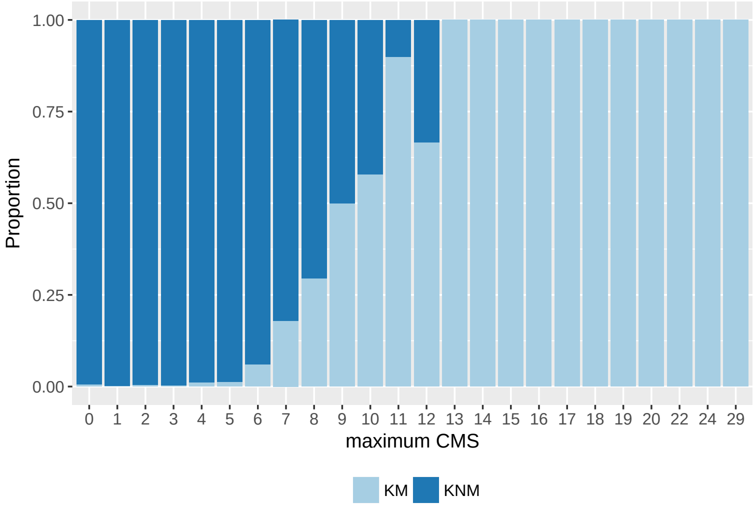

Figure 2.3: This bar chart represents the conditional probabilities of two bullets matching given the maximum number of CMS. The light blue represents known matches, while the dark blue represents known non-matches.

Generally, as seen in Figure 2.3, the probability of finding a match tends to increase with then number of maximum CMS. For ________ maximum CMS values is it much more likely that we have a ____________________________ pair.

[1] http://scholarlycommons.law.northwestern.edu/cgi/viewcontent.cgi?article=6378&context=jclc.

2.1.7 Independence

If the likelihood of one event is *not* affected by knowing whether a second has occured, then the two events are said to be ___________________________. For example, the region of the country where you live and what color car you drive are (probably) not related.

The death penalty example from the previous section demonstrates that defendants receiving the death penalty is *not* independent of the race of the victim. In other words, a black defendant found guilty of murder in Georgia in the 1980s received a different penalty according to the race of the victim.

Another example from DNA analysis relies on on independence across chromosomes. By using loci on different chromosomes, there is independence between the allele counts, allowing for simple calculation of random match probabilities.

2.1.8 Probability Math... Again

Recall the formula for

2.2 Probability to Statistical Inference

2.2.1 Collecting Data

2.2.2 Probability Distributions

2.2.2.1 Normal

2.2.2.2 Lognormal

2.2.2.3 Discrete

2.3 Statistical Inference - Estimation

2.3.1 Background

2.3.2 Point Estimation

2.3.3 Standard Errors

2.3.4 Sample Size

2.3.5 Interval Estimation

2.4 Statistical Inference - Hypothesis Testing

2.4.1 Background

2.4.2 Normal Data

2.4.3 Confidence Intervals

2.4.4 Comparing Two Means

2.4.5 Discussion

Chapter 3

Statistics for Forensic Science

3.1 Brief Review of Probability and Statistics

3.2 The Forensic Examination

3.3 Common Approaches to Assessing Forensic Evidence

3.3.1 Significance Testing / Coincidence Probability

3.3.2 Likelihood Ratio

References

Baldus, David C., Charles Pulaski, and George Woodworth. 1983. “Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience.” *Journal of Criminal Law and Criminology.*

Hare, Eric, Heike Hofmann, and Alicia Carriquiry. 2017. “Automatic Matching of Bullet Land Impressions.” *The Annals of Applied Statistics* Upcoming.