Forensic Science and Statistics: Version 1.0.0

Jeff Holt and Joseph Swetonic

2022-01-17

Contents

Chapter 1

Welcome

This is a draft version of a work in progress. It is NOT intended for circulation.

This book introduces statistical methods that are of use in forensic science. In some cases the methods are currently used. In others the methods are described but not generally in use, although we hope that eventually they will be!

We assume that the reader has a basic understanding of mathematics, but no prior knowledge of statistics is required. We will provide a modest description of forensic methods, but as many of these are highly nuanced we do not attempt to provide a deep treatment of forensic methods here. However, when possible we will try to provide references to more information.

There are portions of this book that borrow from the treatment of statistics given in the online materials:

> Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

This site provides a comprehensive introduction to statistics and is worth referencing when you would like additional information. We deeply appreciate their willingness to allow the use of their work.

1.1 An Introductory Example

Consider the following scenario: Suppose there is a late-night break-in at a closed convenience store. No one is present to see the perpetrator, but there is low-quality CCTV footage. Unfortunately, the *only* thing that can be discerned from the video is the color of the perpetrator's hair.

Now consider two possible versions of what comes next:

Version 1: The video shows that the perpetrator has *brown* hair. The day after the robbery, police see a person with brown hair walking down the street. The person is arrested on suspicion of committing the break-in.

Version 2: The video shows that the perpetrator has *green* hair. The day after the robbery, police see a person with green hair walking down the street. The person is arrested on suspicion of committing the break-in.

For the sake of this simplified example, in both versions we assume that the color of the hair is the **only** reason the police suspect the person who is arrested.

Question: In which version do you think it is more likely that the police have arrested the person who committed the break-in?

Answering the question requires considering the likelihood that the police have the correct person in both versions and deciding which is greater. In most communities people with brown hair outnumber people with green hair, so let's assume that is the case here. Intuitively, it seems reasonable to expect that the likelihood of a correct arrest in Version 2 is greater than that in Version 1: Brown-haired people are relatively common, so in Version 1 the likelihood of arresting the correct brown-haired person is probably low. On the other hand, green-haired people are relatively rare, so in Version 2 the likelihood of arresting the correct person is higher.

Two important points:

1) Although the likelihood of the correct arrest is greater in Version 2 than in Version 1 does not mean that the likelihood in Version 2 is high. It is possible that there are more brown-haired people than green-haired people while still having a lot of green-haird people, making the correct arrest unlikely in either case.

2) While the above discussion is interesting, the word "likelihood" is imprecise. Going forward we will develop "probability" which will allow us to quantify the notion of likelihood, so that we will be able to specify (for instance) how much more likely a correct arrest is in Version 1 than in Version 2.

Chapter 2

Introduction to Probability

2.1 The Previous Example

The example from the previous section involves the likelihood of arresting the correct person based on their hair color. Let's provide some additional context: Suppose that the crime described takes place on an island of 1000 people. The number of people with each hair color is given in the table below.

Color	Count
Brown	670
Blond	200
Red	110
Green	20

Table 1: Hair Color Counts

We start this introduction to probability with an example that purposely simplified: Suppose that there is a late-night break-in at a closed convenience store. No one is present to see the perpetrator, but there is low-quality CCTV footage. Unfortunately, the *only* thing that can be discerned from the video is the color of the perpetrator's hair.

Now consider two possible versions of what comes next:

Version 1: The video shows that the perpetrator has *brown* hair. The day after the robbery, police see a person with brown hair walking down the street. The person is arrested on suspicion of committing the break-in.

Version 2: The video shows that the perpetrator has *green* hair. The day after the robbery, police see a person with green hair walking down the street. The person is arrested on suspicion of committing the break-in.

For the sake of this simplified example, in both versions we assume that the color of the hair is the *only* reason the police suspect the person arrested.

Question: In which version do you think it is more likely that the police have arrested the person who committed the break-in?

Answering the question requires considering the likelihood that the police have the correct person in both versions and deciding which is greater.

A **probability** is a number between 0 and 1 that provides a measure of the likelihood that something occurs, with a larger number indicating increased likelihood. In probability terminology, the thing occurring is called an *event*. In our question, the event is the police arresting the correct person.

Coming back to our question, we need to decide which version has the greater probability. In most communities, people with brown hair outnumber people with green hair – let's assume that is the case here. Thus, it seems reasonable to expect that the probability of Version 2 is greater than that of Version 1: Brown-haired people are relatively common, so in Version 1 the probability of arresting the correct brown-haired person is low. On the other hand, green-haired people are fairly are, so in Version 2, the probability of arresting the correct person is higher.

Two important points:

1) Just because the probability of the correct arrest is greater in Version 2 than in Version 1 does not mean that the probability in Version 2 is high. It's possible that there are more brown-haired people than green-haired people while still having a lot of green-haird people, making both probabilities small.

2) While the above discussion is interesting, it is important to have a better sense of the actual numerical value of the probability. We start in that direction in the next section.

2.2 Calculating Probabilities From Data

Let's suppose that the crime described above takes place on an island that has 1000 people. The number of people with each hair color is given in the table below.

Color	Count
Brown	670
Blond	200
Red	110
Green	20

Table 1: Hair Color Counts

Now suppose that we select a person from this population at random. To compute the **probability** that this person has brown hair, we take the number with brown hair and divide by the number of people:

$$\text{P(Brown hair)} = \frac{670}{1000} = 0.67$$

We use "P(…)" to denote the probability of something. For instance in this case P(Brown hair) = "probability of brown hair".

A probability is a number between 0 and 1, with a smaller number indicating a less likely outcome.

2.2.1 Sample Questions

Sprinkled thought this book are Sample Questions. In the online version, the solutions to these questions are hidden. Before looking at the solution (by hovering), we recommend that you try to do them yourself!

1. What is the probability that a randomly selected person has blond hair?

Answer: $P(\text{Blond}) = \frac{200}{1000} = 0.20$.

2. What is the probability that a randomly selected person has red hair?

Answer: $P(\text{Red}) = \frac{110}{1000} = 0.11$.

Chapter 3

Joint Probability

Suppose that we have a population of 1000 people, each classified by gender (Female or Male) and blood type (A, AB, B, or O). The number of people of each combination of gender and blood type is shown in the table below.

	F	M
A	175	235
AB	16	24
B	37	63
O	202	248

Gender and blood type counts

Suppose that we select one of the people at random, and note that person's gender and blood type. There are various possibilities for outcomes, such as the person selected is female with Type A blood. The chance that this outcome occurs is called the **probability** of this outcome. As 175 out of the 1000 people are female of blood type A, the probability of selecting such a person is

$$\text{P(Female and Type A)} = \frac{175}{1000} = 0.175$$

Here, P(Female and Type A) stands for the probability that the person selected is Female **and** has Type A blood. A sample of other probabilities that come directly from the table are

$$\begin{aligned} \text{P(Female and Type O)} &= \frac{202}{1000} = 0.202 \\ \text{P(Male and Type AB)} &= \frac{24}{1000} = 0.024 \\ \text{P(Male and Type A)} &= \frac{235}{1000} = 0.235 \end{aligned}$$

We also can combine table entries to compute other probabilities. For instance, there are a total of $16 + 24 = 40$ people with blood type AB, so

$$\text{P(Type AB)} = \frac{40}{1000} = 0.04$$

For another example, there are a total of $235 + 24 + 63 + 248 = 570$ males in our population, so

$$\text{P(Male)} = \frac{570}{1000} = 0.57$$

3.0.1 Sample questions here

Place some example questions and answers here.

3.1 And vs Or

Above we saw that P(Female and Type A) = 0.175. Suppose we instead would like to know P(Female or Type A)? That is, we want to know the probability that a randomly selected person is either female or has type A blood. (Note that this group includes those who are both female **and** have type A blood.) From our table we see that the total number in this group is $175+16+37+202+235 = 665$ so that

$$\text{P(Female or Type A)} = \frac{665}{1000} = 0.665$$

Other combinations are also possible. For instance, there are $175 + 235 = 310$ people with type A blood and $37 + 63 = 100$ people with type B blood, so there are a total of $310 + 100 = 410$ people that have either type A or type B blood. Therefore, we have

$$\text{P(Type A or Type B)} = \frac{410}{1000} = 0.41$$

3.1.1 Sample questions here

Place some example questions and answers here.

3.2 Probability Tables

Frequently, tables provide probabilities (or percentages) instead of counts. To make the conversion from counts to probabilities, all we do is divide each table

entry by the total. Thus, for our table, we divide each entry by 1000 to arrive at

	F	M
A	0.175	0.235
AB	0.016	0.024
B	0.037	0.063
O	0.202	0.248

Gender and blood type counts

Table 2 entries give the probability of selecting each combination. For instance

$$\text{P(Female and Type O)} = 0.202$$

We can add entries in this table to compute other probabilities.
For example, suppose we want to compute the probability that a randomly selected person has type O blood. Based on the table, this can happen if a person is female and has type O blood, or if a person is male and has type O blood. Because these two groups are disjoint, we can add the probabilities to arrive at

$$\text{P(Type O)} = \text{P(Female and Type O)} + \text{P(Male and Type O)} = 0.202 + 0.248 = 0.45$$

If instead (for instance), we want to compute P(Female) then we add the entries in the corresponding column, giving us

$$\text{Pr(Female)} = 0.175 + 0.016 + 0.037 + 0.202 = 0.43$$

Chapter 4

Conditional Probability

Now suppose that we just focus on the females in the population, which forms a subset of the population. The probability that a randomly selected female has blood type O is an example of a **conditional probability**.

There are 430 females in the population, and 202 of those have type O blood. Hence, the probability that female chosen at random has type O blood is

$$\text{P(Type O} \mid \text{Female)} = \frac{202}{430} = 0.470$$

The vertical bar in the notation is interpreted as *given that*, so that P(Type O | Female) is read as

The probability of blood type O, given the person is female.

Conditional probabilities arise all the time when evaluating forensic evidence. Other examples of conditional probabilities:

The probability of Type AB, given that the person is male:

$$\text{P(Type AB} \mid \text{Male)} = \frac{24}{570} = 0.042$$

The probability the person is female, given Type B blood:

$$\text{P(Female} \mid \text{Type B)} = \frac{37}{100} = 0.37$$

The probability the person is male, given Type A blood:

$$\text{P(Male} \mid \text{Type A)} = \frac{235}{410} = 0.573$$

4.0.1 Sample questions here

Place some example questions and answers here.

Chapter 5

Probability Rules

Earlier we computed the conditional probability

$$\text{P(Type O} \mid \text{Female)} = \frac{202}{430}$$

The numerator 202 came from the count of Females who have blood Type B, and the denominator 430 is the total number of Females. If we divide the numerator and denominator by 1000, then the quotient is not changed, so that

$$\text{P(Type O} \mid \text{Female)} = \frac{202/1000}{430/1000} = \frac{0.202}{0.43} = \frac{\text{P(Type O and Female)}}{\text{P(Female)}}$$

This illustrates a general property of probability: If A and B represent possible outcomes, then the probability of A given B is

$$\text{P(A} \mid \text{B)} = \frac{\text{P(A and B)}}{\text{P(B)}}$$

Multiplying on both sides of the above equation by P(B) gives

$$\text{P(A and B)} = \text{P(A} \mid \text{B)P(B)}$$

This is called the **multiplication rule**. Reversing the order of A and B above gives

$$\text{P(B and A)} = \text{P(B} \mid \text{A)P(A)}$$

In the blood type example, it is clear that P(Female and Type O) is the same as P(Type O and Female). This is true in general: P(A and B) = P(B and A)}, it follows that

$$\Pr(\text{A} \mid \text{B})\Pr(\text{B}) = \Pr(\text{B} \mid \text{A})\Pr(\text{A})$$

so that

$$\Pr(\text{A} \mid \text{B}) = \Pr(\text{B} \mid \text{A})\frac{\Pr(\text{A})}{\Pr(\text{B})}$$

This formula is called **Bayes Rule**. Applied to our earlier population, we have

$$\Pr(\text{Type O} \mid \text{Female}) = \Pr(\text{Female} \mid \text{Type O})\frac{\Pr(\text{Type O})}{\Pr(\text{Female})}$$

Two outcomes A and B are **independent** when $Pr(A \mid B) = Pr(A)$, which implies that knowing if B has occurred has no effect on the probability of A.

Chapter 6

Counterintuitive Applications

6.1 Birthday Paradox

How many people are needed in a classroom so that the probability of them sharing a birthday is at least $\frac{1}{2}$? Intuitively, one could claim that 183 people in the group would imply that the probability is greater than $\frac{1}{2}$, since $\frac{183}{365} = 0.501$. However, intuition does not correctly solve this problem.

To begin, there are some assumptions that need to be made. For simplicity, we will ignore leap years and assume that all 365 birthdays have an equal probability of occurring.

We want to compute $P(\beta)$, the probability that at least 2 people share the same birthday. Recall from the previous chapter that $P(\beta) = 1 - P(\beta')$. In this case, it is much simpler to calculate the probability that no one in the group shares the same birthday with someone else.

Let's start with the simple example involving only two people. The probability that they do not share the same birthday is

$$P(\beta') = \left(\frac{365}{365}\right)\left(\frac{364}{365}\right) = 0.997$$

$$1 - P(\beta') = 0.003$$

Extending this result to a group of five people:

$$P(\beta') = \left(\frac{365}{365}\right)\left(\frac{364}{365}\right)\left(\frac{363}{365}\right)\left(\frac{362}{365}\right)\left(\frac{361}{365}\right)$$

$$= \frac{365!}{365^5(365-5)!} = 0.973$$

$$1 - P(\beta') = 0.027$$

This formula can be applied to any number of group size, so that in the general case of n people, the probability that at least two share the same birthday is

$$P(\beta) = 1 - \frac{365!}{365^n(365-n)!}$$

Finally, to solve the original question: How many people are needed in a classroom so that the probability of them sharing a birthday is at least $\frac{1}{2}$? Iteratively increasing the group size from five, it is clear to see that at a group size of 23, $P(\beta) > \frac{1}{2}$

$$P(\beta) = 1 - P(\beta') = 1 - \frac{365!}{365^{23}(365-23)!} = 0.507$$

6.2 Monty Hall Problem

The original problem was popularized as a letter by Craig F. Whitaker sent to Marilyn vos Savant's column in Parade magazine in 1990. You can read more about the article here: https://web.archive.org/web/20130121183432/http://marilynvossavant.com/game-show-problem/

Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car, behind the others, goats. You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat. He says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice of doors?