

Forensic Science and Statistics: Version 1.0.0

Jeff Holt and Joseph Swetonic

2022-01-11

Contents

1	Introduction	5
2	Probability	7
2.1	Introduction	7
2.2	Another Subsection	7
2.3	And vs Or	9
2.4	Conditional Probability	9
3	Introduction to Probability	11
3.1	A Initial Example	11
3.2	Calculating Probabilities From Data	12
4	Joint Probability	13
4.1	And vs Or	14
4.2	Probability Tables	15
5	Conditional Probability	17
6	Probability Rules	19
7	Counterintuitive Applications	21
7.1	Birthday Paradox	21
7.2	Monty Hall Problem	22

Chapter 1

Introduction

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 they could 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 science methods, but as many of these are highly nuanced we do not attempt to provide a deep treatment of them 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 probability and 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 probability and statistics, and is worth referencing when you would like additional information. We appreciate their willingness to allow the use of their work.

Chapter 2

Probability

Probability is an important and complex field of study. Fortunately, only a few basic issues in probability theory are essential for understanding statistics at the level covered in this book. These basic issues are covered in this chapter.

2.1 Introduction

TODO: Edit the introductory section below.

“The introductory section discusses the definitions of probability. This is not as simple as it may seem. The section on basic concepts covers how to compute probabilities in a variety of simple situations. The Gambler’s Fallacy Simulation provides an opportunity to explore this fallacy by simulation. The Birthday Demonstration illustrates the probability of finding two or more people with the same birthday. The Binomial Demonstration shows the binomial distribution for different parameters. The section on base rates discusses an important but often-ignored factor in determining probabilities. It also presents Bayes’ Theorem. The Bayes’ Theorem Demonstration shows how a tree diagram and Bayes’ Theorem result in the same answer. Finally, the Monty Hall Demonstration lets you play a game with a very counterintuitive result.”

2.2 Another Subsection

In what follows we introduce the basics of probability, which quantifies the chance that something occurs. Let’s start with a specific situation to serve as an example. 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	Gender and blood type counts
AB	16	24	
B	37	63	
O	202	248	

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

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

Here, $P(\text{Female and Type A})$ 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

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

We also can combine table entries to compute other probabilities. For instance, there are a total of $235 + 24 + 63 + 248 = 570$ males in our population, so

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

There are a total of $16 + 24 = 40$ people with blood type AB, so

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

2.2.1 Sample Questions

1. What is the probability that the person selected is female?

Answer: $P(\text{Female}) = 1 - P(\text{Male}) = 1 - 0.57 = 0.43$ This is equivalent to $(175 + 16 + 37 + 202)/1000$.

2. What is the probability that the person selected is male and has type A blood and type O blood?

Answer: $P(\text{Male and A and O}) = 0$. The probability of this event occurring in the table is zero, as no men have two different types of blood.

2.3 And vs Or

Above we saw that $P(\text{Female and Type A}) = 0.175$. Suppose we instead would like to know $P(\text{Female or Type A})$? 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

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

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

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

2.3.1 Sample Questions

1. What is the probability of selecting a female or male with type AB blood?

Answer: $P((\text{Male or Female}) \text{ and AB}) = P((\text{Male and AB}) \text{ or } (\text{Female and AB})) = (16 + 24)/1000 = 0.04$

2.4 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

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

The notation $P(\text{Type O} \mid \text{Female})$ is read out loud as *the probability of blood type O, given the individual 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:

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

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

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

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

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

2.4.1 Sample Questions

1. What is the probability of blood type AB or B given the person selected is female?

Answer: $P(\text{AB or B} \mid \text{Female}) = (16 + 37)/430 = 0.123$

Chapter 3

Introduction to Probability

3.1 A Initial Example

We start this introduction to probability with an example that has been simplified in many ways: 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 rare, 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-haired people, making both probabilities small. (Consider if we replace “green” with “blond”.)
- 2) While the above discussion is interesting, it is important to have a better sense of the actual value of the probability. We start in that direction in the next section.

3.2 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	Table 1: Hair Color Counts
Brown	670	
Blond	180	
Red	110	
Green	40	

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:

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

This space will get examples and a probability table

Chapter 4

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	Gender and blood type counts
AB	16	24	
B	37	63	
O	202	248	

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

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

Here, $P(\text{Female and Type A})$ 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

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

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

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

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

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

4.0.1 Sample questions here

Place some example questions and answers here.

4.1 And vs Or

Above we saw that $P(\text{Female and Type A}) = 0.175$. Suppose we instead would like to know $P(\text{Female or Type A})$? 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

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

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

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

4.1.1 Sample questions here

Place some example questions and answers here.

4.2 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

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

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

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

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

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

Chapter 5

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

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

The vertical bar in the notation is interpreted as *given that*, so that $P(\text{Type O} \mid \text{Female})$ 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:

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

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

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

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

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

5.0.1 Sample questions here

Place some example questions and answers here.

Chapter 6

Probability Rules

Earlier we computed the conditional probability

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

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

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

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

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

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

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

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

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

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

$$\Pr(A \mid B)\Pr(B) = \Pr(B \mid A)\Pr(A)$$

so that

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

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 7

Counterintuitive Applications

7.1 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, $P(\beta) > \frac{1}{2}$

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

7.2 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?