Forensic Science and Statistics: Version 1.0.0

Jeff Holt and Joseph Swetonic

2022-01-30

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. Here we begin the development of "probability" which allows us to quantify the notion of likelihood.

Let's start by providing some additional context for our example: Suppose that the crime described takes place on an island of 1000 people. The number of these 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

A **probability** is a number between 0 and 1 that provides a measure of the likelihood that something occurs, with a larger number indicating greater likelihood.

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".

2.1.1 Sample Questions

Sprinkled thought this book are Sample Questions, which also play the role of examples. 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: $P(\text{Red}) = \frac{110}{1000} = 0.11$.

These examples are fine but do not answer the question from the previous section. In Version 1, we have a brown-haired person on the CCTV and one of the 670 brown-haired people randomly arrested. Thus the probability that the correct person is arrested is

$$\text{P(correct arrest)} = \frac{1}{670} = 0.00149$$

In Version 2 of the example we have a green-haired person on CCTV, so this time the probability that the correct person is arrested is

$$\text{P(correct arrest)} = \frac{1}{20} = 0.05$$

Dividing the two probabilities

$$\frac{0.05}{0.00149} = 33.5$$

we see that a correct arrest in Version 2 is 33.5 times as likely as in Version 1. However, even in Version 2 there is only a 0.05 probability that the police have the correct person!

2.1.2 Sample Questions

1. Suppose CCTV shows a red-haired person committed the robbery, and that a random red-haired person is arrested. What is the probability of a correct arrest in this instance?

Answer: $P(\text{correct arrest}) = \frac{1}{110} = 0.00909$.

2. Determine the number of times greater the likelihood of a correct arrest if the culprit has red hair vs having blond hair.

Answer: In the case of a blond-haired person, we have $P(\text{correct arrest}) = \frac{1}{200} = 0.005$. Therefore the ratio of probabilities is $\frac{0.00909}{0.005} = 1.81818$ so a correct arrest for a red-haired culprit is 1.81818 times as likely as for a blond-haired culprit.

2.2 Exercises

1. Put exercises here

Chapter 3

Joint Probability

3.1 Blood Types

Blood plays a role in many forensic science applications, with the “blood type” determined by a combination of two different blood group systems:

> *Blood types are inherited and represent contributions from both parents. As of 2019, a total of 41 human blood group systems are recognized by the International Society of Blood Transfusion (ISBT). The two most important blood group systems are ABO and Rh; they determine someone’s blood type (A, B, AB, and O, with +, − or null denoting RhD status) for suitability in blood transfusion.*
>
> Source: https://en.wikipedia.org/wiki/Blood_type

Suppose that we have a collection of 1000 people, with each classified based on both ABO (A, B, AB, or O) and Rh (+ or −). The number of people of each combination of ABO and Rh class (the “blood type”) is shown in the table below. (This table is based on the distribution of blood types in Canada, reported at https://www.blood.ca/en/blood/donating-blood/whats-my-blood-type)

	+	−
A	360	60
B	76	14
AB	25	5
O	390	70

Counts by ABO and Rh groups

Suppose that one of these people is selected at random and note that person’s groups. There are various possibilities for outcomes, such as the person selected has ABO group O and Rh group +. The likelihood that this combination occurs is the **joint probability** of the two classifications, ABO and Rh groupings.

There are 390 out of the 1000 people are O and +, so the probability of randomly selecting such a person is

$$\mathrm{P(O\ and\ +)} = \frac{390}{1000} = 0.39$$

Here P(O and +) stands for the probability that the person selected is ABO group O **and** Rh group +. A sample of other probabilities that come from the table are

$$\begin{aligned}
\mathrm{P(AB\ and\ -)} &= \frac{5}{1000} &= 0.005 \\
\mathrm{P(B\ and\ +)} &= \frac{76}{1000} &= 0.076 \\
\mathrm{P(B\ and\ -)} &= \frac{14}{1000} &= 0.014
\end{aligned}$$

We also can combine table entries to compute other probabilities. For instance, there are a total of $76 + 14 = 90$ people with blood group B, so the probability of randomly selecting a person with blood group B is

$$\mathrm{P(B)} = \frac{76+14}{1000} = \frac{90}{1000} = 0.09$$

As another example, the number of people with Rh group + is $360 + 76 + 25 + 390 = 851$. Thus the probability of randomly selecting a person with Rh group + is

$$\mathrm{P(+)} = \frac{360+76+25+390}{1000} = \frac{851}{1000} = 0.851$$

3.1.1 Sample Questions

1. Find the probability that a randomly selected person has blood groups A and −.

Answer: P(A and −) $= \frac{60}{1000} = 0.06$.

2. Find the probability that a randomly selected person has ABO group AB.

Answer: There are $25 + 5 = 30$ people that have ABO group AB, so that P(AB) $= \frac{30}{1000} = 0.03$.

3. Find the probability that a randomly selected person has ABO group A and ABO group B and Rh group +.

Answer: The is no person that has ABO group A *and* ABO group B, so P(A and O and +) $= \frac{0}{1000} = 0$.

3.2 And vs Or

Above we saw that P(B and +) = 0.076. Suppose we instead would like to know P(B or +)? That is, we want to know the probability that a randomly selected person is either ABO group B **or** Rh group +. Note that this includes those people who are both ABO group B **and** Rh group +.

From our table we see that the total number of people satisfying one or both conditions is $360 + 76 + 24 + 390 + 14 = 864$ so that

$$\text{P(B or +)} = \frac{864}{1000} = 0.864$$

Other combinations are also possible. For instance, number of people with ABO group A blood or ABO group B blood is

$$\underbrace{(360 + 60)}_{\text{ABO group A}} + \underbrace{(76 + 14)}_{\text{ABO group B}} = 510$$

Therefore we have

$$\text{P(A or B)} = \frac{510}{1000} = 0.51$$

3.2.1 Sample Questions

1. Find the probability that a randomly selected person has ABO group AB or Rh group − blood.

Answer: The number of people who have ABO group AB **or** Rh group − is $25 + 5 + 60 + 14 + 70 = 174$ so that P(AB or −) $= \frac{174}{1000} = 0.174$

2. Find the probability that a randomly selected person has ABO group O or Rh group + blood.

Answer: The number of people who have ABO group O **or** Rh group + is $390 + 70 + 360 + 76 + 25 = 921$ so that P(O or +) $= \frac{921}{1000} = 0.921$

3. Find the probability that a randomly selected person has ABO group AB or O.

Answer: The number of people who have ABO group AB **or** ABO group O is $25 + 5 + 390 + 70 = 490$ so that P(AB or O) $= \frac{490}{1000} = 0.49$

4. Find the probability that a randomly selected person has ABO group A or AB, and also Rh group −.

Answer: We have (A or AB) and (−) = (A and −) or (AB and −). The number of people who are (A and −) is 60, and the number of people who are (AB and −) is 5, so that $P((\text{A or AB}) \text{ and } (-)) = \frac{60+5}{1000} = 0.065$.

3.3 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

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

3.4 Exercises

1. Put exercises here

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

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

The vertical bar in the notation is interpreted as *given that*, so that P(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:

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

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

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

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

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

Chapter 7

Graphing Distributions: Qualitative Variables

Note: Portions below modeled after content from *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/) Project Leader: David M. Lane, Rice University

7.1 Introduction

Suppose that we have a community of 500 people. Each is classified based on their ABO blood group, which is one of A, B, AB, or O. Below we consider graphical methods for displaying the results of the blood group classifications. This starts with tables, and then continues on to how to graph data that fall into a small number of categories.

This is an example of *qualitative data*. One characteristic of such data is that the different values do not come with any pre-established ordering. This can be contrasted with quantitative data, such as the weight of a bag of an unknown substance, which does have a natural ordering with respect to different weights.

7.2 Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 1 shows a frequency table for the results of the ABO blood group classification. It also shows the relative frequencies, which are the proportion classified in each category. For example, the relative frequency for group B is $45/500 = 0.09$.

ABO Group	Frequency	Relative Frequency
A	210	0.42
B	45	0.09
AB	15	0.03
O	230	0.46

Table 1: Frequency Table for ABO Group Data

7.3 Pie Charts

The pie chart in Figure 7.1 depicts the ABO group data. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of items in the category – that is, the relative frequency multiplied by 100.

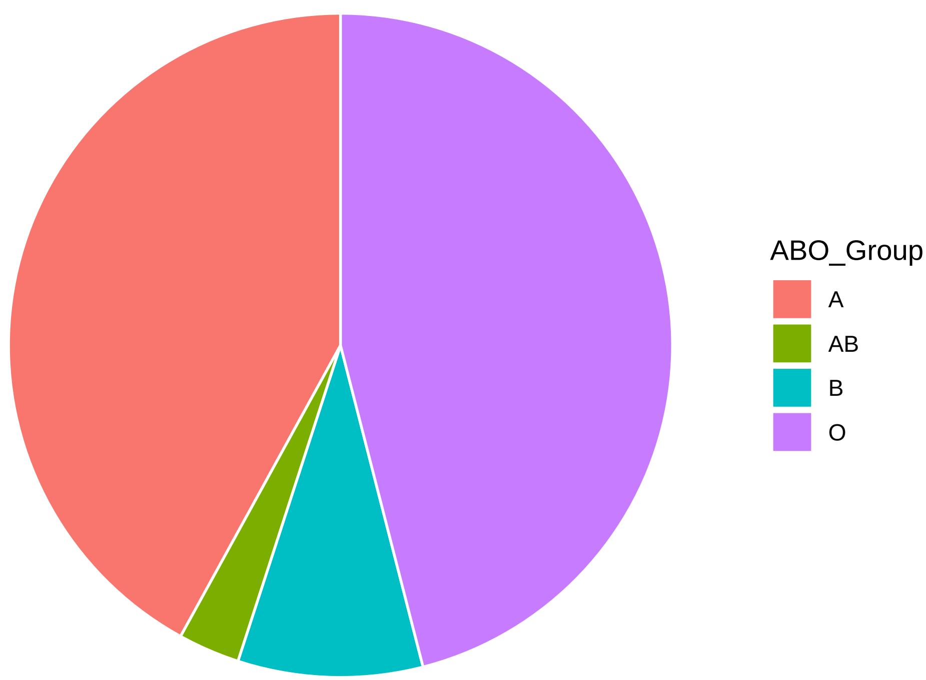

Figure 7.1: Relative Frequencies for ABO Blood Groups

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted, “The only worse design than a pie chart is several of them.”

7.4 Bar Charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the ABO frequencies is shown in Figure 7.2. Frequencies are shown on the Y-axis and the blood group is shown on the X-axis.

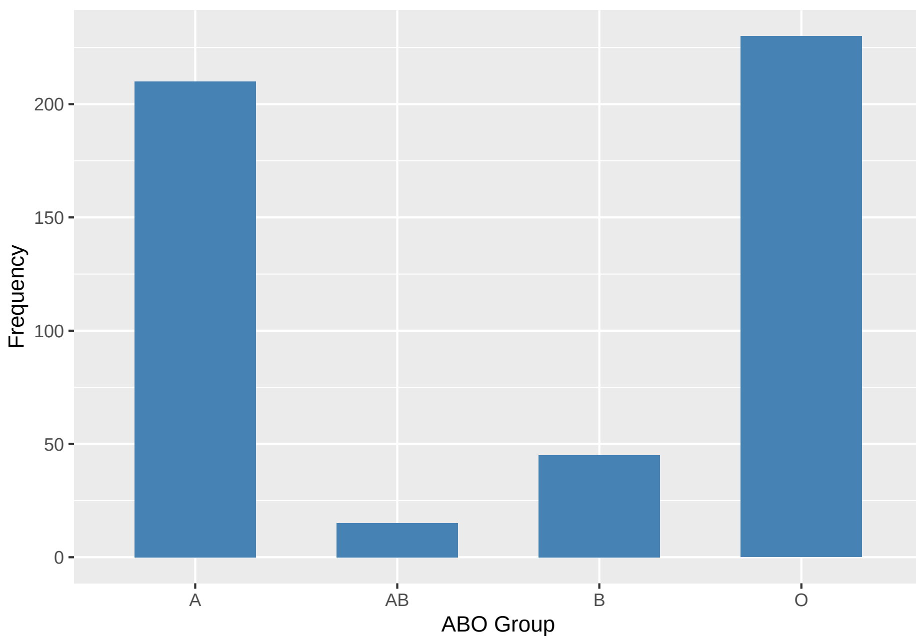

Figure 7.2: Frequencies for ABO Blood Groups

The Y-axis also can show the percentage of observations instead of the number of observations, as in Figure 7.3.

7.5 Comparing Distributions

Often we need to compare different sets of data, or different subsets within the same overall data set In this case, we are comparing the “distributions” of outcomes or responses. Bar charts are often excellent for illustrating differences between two distributions. Table 2 shows the distribution (in percentages) of ABO blood groups for those in Albania and Australia.

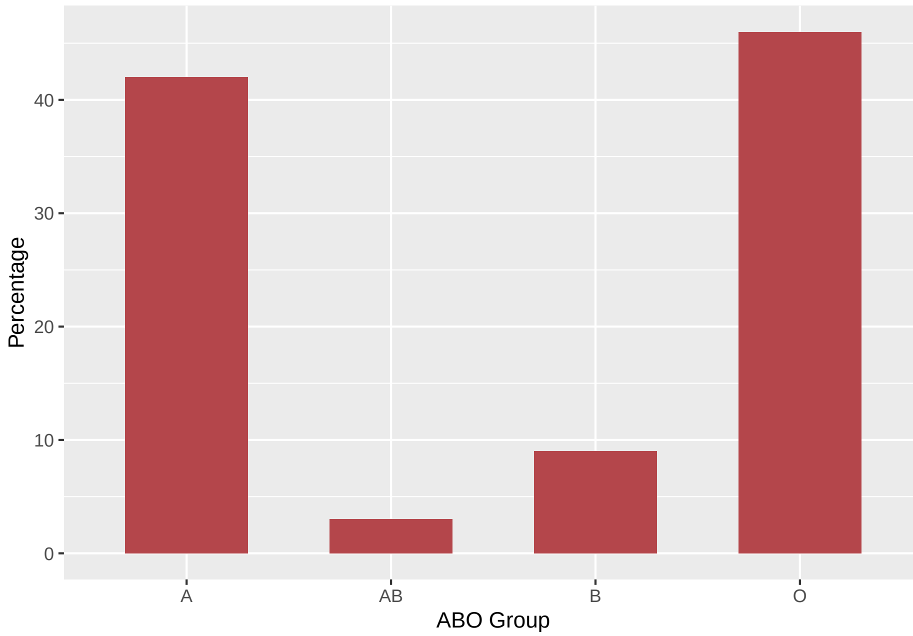

Figure 7.3: Percentages for ABO Blood Groups

ABO Group	Albania	Australia
A	36.7	38.0
B	17.1	10.0
AB	6.1	3.0
O	40.1	49.0

Table 2: ABO Blood Group Percentages

From Table 2 we see that ABO groups B and AB are more common in Albania, group O is more common in Australia, and group A is similar for both. This can be seen in the bar chart in Figure 7.4.

The bars in Figure 7.4 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels.

7.6 Exercises

1. Put exercises here

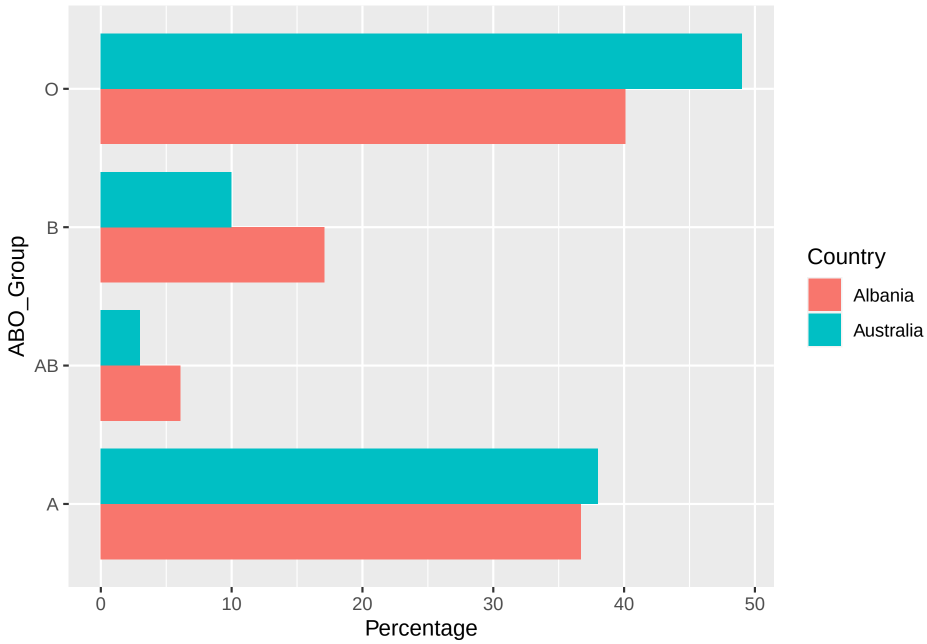

Figure 7.4: Percentages for ABO Blood Groups by Country

7.7 Code Appendix

```
library(dplyr)
library(ggplot2)
library(tidyr)

data <- data.frame(ABO_Group = c("A", "B", "AB", "O"), value = c(210, 45, 15, 230)) %>%
  mutate(prop = value / sum(value) * 100)

# Figure 7.1 -----
ggplot(data = data, aes(x = "", y = prop, fill = ABO_Group)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y", start = 0) +
  theme_void()

# Figure 7.2 -----
ggplot(data, aes(ABO_Group, value)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(x = "ABO Group", y = "Frequency")

# Figure 7.3 -----
```

```
ggplot(data, aes(ABO_Group, prop)) +
  geom_bar(stat = "identity", fill = "#B4464B") +
  labs(x = "ABO Group", y = "Percentage")

# Figure 7.4 -----
data <- data.frame(ABO_Group = c("A", "B", "AB", "O"),
                   Albania = c(36.7, 17.1, 6.1, 40.1),
                   Australia = c(38, 10, 3, 49)) %>%
  pivot_longer(!ABO_Group, names_to = "Country", values_to = "Percentage")

ggplot(data, aes(ABO_Group, Percentage, fill = Country)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.8) +
  labs(x = "ABO Group")
  coord_flip()
```

Chapter 8

Graphing Distributions: Histograms

Note: Portions below modeled after content from *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/) Project Leader: David M. Lane, Rice University

8.1 Introduction

The "distribution" of a set of data consists of the set of possible data values and the frequency that the values occur. A **histogram** is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of data values, when it is not practical to list all values.

We begin with an example consisting of the weights of 371 bags containing a substance suspected of being narcotics. The weights of the bags range from 215 to 270 grams. Below are the weights for 10 bags. (It is not practical to show them all.)

$$240.91, 234.66, 246.84, 244.24, 253.39, 245.07, 228.75, 237.29, 255.69, 254.64$$

One way to get a sense of the distribution of the weights is through the use of a frequency table, where we organize the data into groups of similar weights and the count the number in each group. The results for this data set are shown in Table 1.

Interval Lower Limit	Interval Upper Limit	Class Frequency
215	220	4
220	225	8
225	230	31
230	235	74
235	240	88
240	245	75
245	250	57
250	255	25
255	260	8
260	265	1

Table 1: Grouped Frequency Distribution of Weights

To create this table, the range of weights was broken into *class intervals*. The first interval is from 215 to 220, the second from 220 to 225, and so on. Next, the number of weights falling into each interval was counted to obtain the class frequencies. There are four weights
in the first interval, eight in the second, etc.

Class intervals of width 5 provide enough detail about the distribution to be revealing without making the graph too "choppy." How one goes about choosing the widths (called "bin widths") of class intervals is discussed later in this section.

In a histogram the class frequencies are represented by bars, which provides a graphical depiction of the data. The height of each bar corresponds to its class frequency. A histogram of the weight data is shown in Figure 8.1.

We can see from the histogram that most of the weights are between 230 and 250 grams, with fewer weights in the extremes. We can also see that the distribution is not quite symmetric, the weights extend to the right farther than to the left. This distribution is therefore said to be *skewed*.

Because the weights are measured to two decimal places, it is not likely that we get many "fence-sitters" – that is, values that land right on the border between two bins. This allows use to choose whole numbers as boundaries, which avoids a cluttered appearance. Computer software generally will use this option when feasible, and sometimes automatically labels the middle of each interval instead of the endpoints.

Histograms can be based on relative frequencies instead of actual frequencies, showing the proportion (or percentage) of values in each interval instead of the exact frequence. Our weight data is displayed as percentages in Figure 8.2. To get the percentages, we divide each frequency by the number of data values (that gives the proportion), and then multiply by 100 to convert to percentages.

We see that the shape of the histogram is unchanged, only the Y-axis is different.

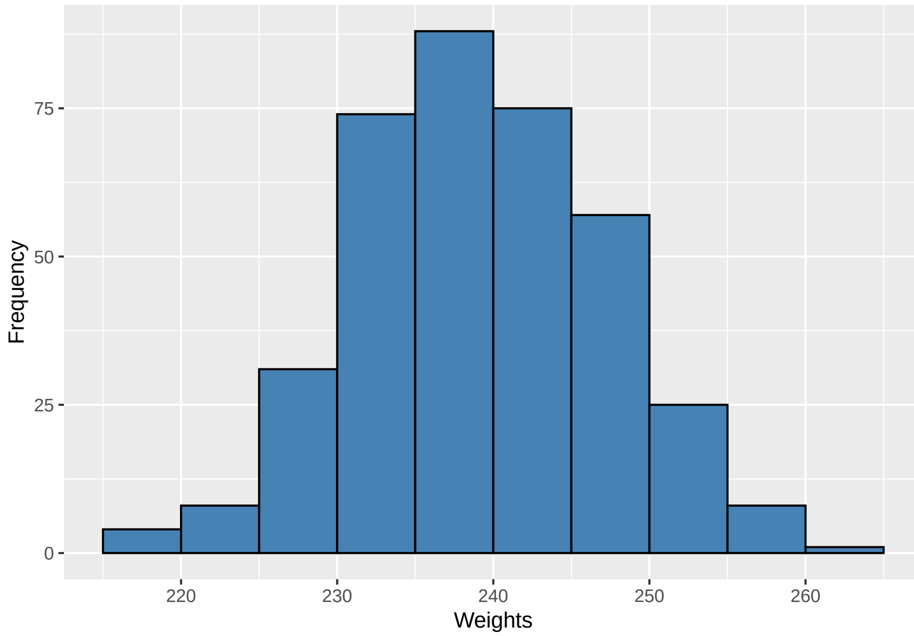

Figure 8.1: Histogram of weights

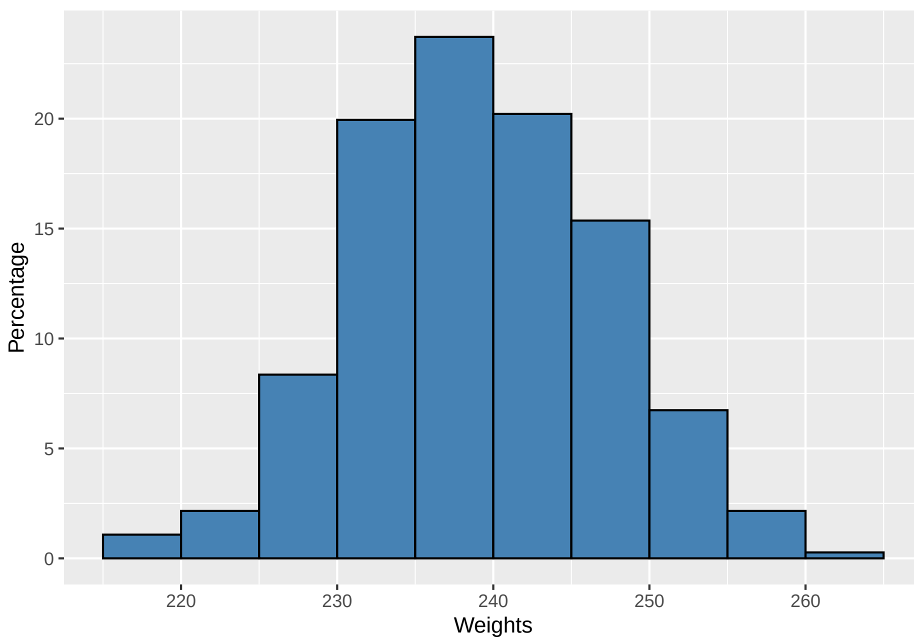

Figure 8.2: Histogram of weights

The choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the histogram. In Figure **??** we see histograms for our data using bins of width 1 (left) and width 10 (right). These both give a general sense of the distribution, but here bins of width 1 produce a graph that is too busy while bins of width 10 produce a graph that lacks detail.

Most software have defaults for bin widths that might work well for a given set of data, or might not. There are some "rules of thumb" that can help guide thed choice of bin width but these are just guidelines that should not be regarded as definitive, so do not feel bound to them. That said, here are a couple of general guidelines you can consider:

• Sturges' rule is to set the number of intervals as close as possible to $1+\log_2(N)$, where $\log_2(N)$ is the base 2 log of the number of observations N. For our data set of $N = 371$, we have $1 + \log_2(N) = 9.535$ which suggests 10 intervals, which happens to match our choice.

• The Rice rule suggests that the number of intervals should be the integer nearest to $2\sqrt[3]{N}$, twice the cube root of the number of data values. For $N = 371$ this formula gives $2\sqrt[3]{371} = 14.371$, which is somewhat more that we chose.

The above rules are just a guide, you are advised to experiment with different numbers of intervals and choose the histogram that you feel best conveys the shape of the distribution.

8.2 Exercises

1. Put exercises here

Chapter 9

Graphing Distributions: Box Plots

Note: Portions below modeled after content from *Online Statistics Education: A Multimedia Course of Study* (http://onlinestatbook.com/) Project Leader: David M. Lane, Rice University

9.1 Introduction

In this chapter we discuss the *box plot* which provides a useful way to grapically display information about the distribution of a set of data, identify outliers, and compare distributions.

To get us started, suppose that we have small pieces of glass from two different sources: a broken window at the scene of a burglary and the trunk of a car belonging to a suspect. To compare the glass from the two sources, a trace element present in glass can be measured. The amount present varies within a sheet of glass so measurements are taken from numerous pieces found at the crime scene and in the car trunk. (The specifics of the element and measurement units are not important for this discussion.) The measurements are given below:

Crime Scene: 61.4 64.3 68.0 68.9 67.7 66.5 66.8 59.1 68.0 71.6 60.1 62.0 68.2 62.8 63.5

Car Trunk: 67.9 68.9 69.5 67.6 73.4 64.5 73.0 63.7 66.9 68.2 71.8 64.8 63.3

To construct the box plot, we start by finding the 25th, 50th, and 75th percentiles for each of our data sets. Recall that the 25th percentile is the quantity such that 25% of the data values are less than this quantity, and similarly for the other percentiles. The percentiles for each of our sets of data are shown in Table 1.

Percentile	Car Trunk	Crime Scene
25^{th}	64.80	63.65
50^{th}	67.90	66.50
75^{th}	69.50	68.00

Table 1: Percentiles

Figure 9.1 shows how these percentiles are incorporated into box plots. The lower and upper limits of the box (called "hinges") extend from the 25th to 75th percentile and a line between those at the 50th percentile (which is the median).

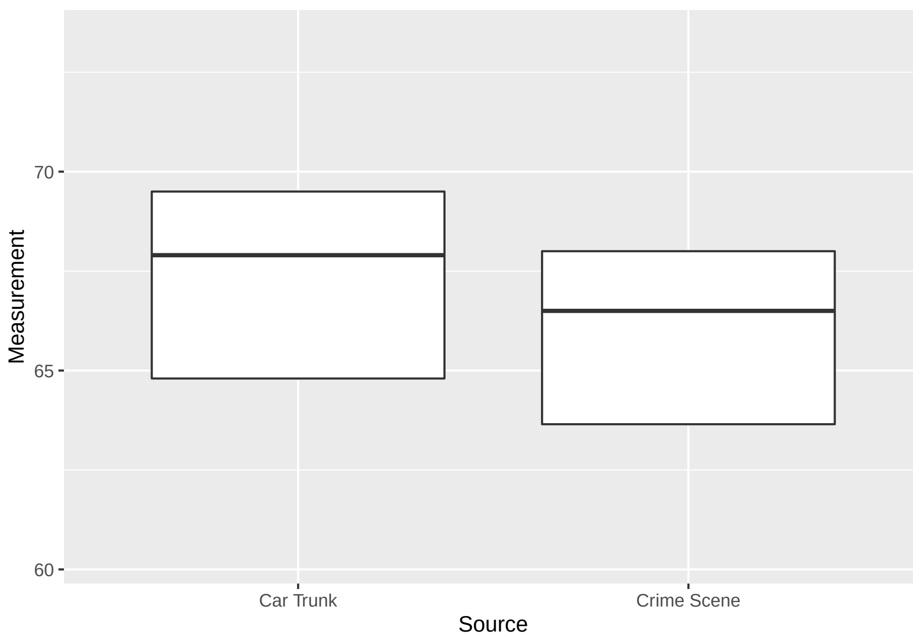

Figure 9.1: Box plots of glass measurements

Most box plots have more information than that shown in Figure 9.1 and there are a wide variety of options that can be included. We give a few examples below to provide a sense of what is possible.

Before proceeding, the terminology in Table 2 is helpful.

Table 2. Box plot terms and values for women's times.

Name Formula Value Upper Hinge 75th Percentile 20 Lower Hinge 25th Percentile 17 H-Spread Upper Hinge - Lower Hinge 3 Step 1.5 x H-Spread 4.5

Upper Inner Fence Upper Hinge + 1 Step 24.5 Lower Inner Fence Lower Hinge - 1 Step 12.5 Upper Outer Fence Upper Hinge + 2 Steps 29 Lower Outer Fence Lower Hinge - 2 Steps 8 Upper Adjacent Largest value below Upper Inner Fence 24 Lower Adjacent

Smallest value above Lower Inner Fence 14 Outside Value A value beyond an Inner Fence but not beyond an Outer Fence 29 Far Out Value A value beyond an Outer Fence None Continuing with the box plots, we put “whiskers” above and below each box to give additional information about the spread of the data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women’s data).

Figure 2. The box plots with the whiskers drawn.

Although we don’t draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small “o’s” and far out values are indicated by asterisks (*). In our data, there are no far out values and just one outside value. This outside value of 29 is for the women and is shown in Figure 3.

Figure 3. The box plots with the outside value shown.

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure 4 shows the result of adding means to our box plots.

Figure 4. The completed box plots.

Figure 4 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25th and 75th percentiles), we see that half the women’s times are between 17 and 20 seconds, whereas half the men’s times are between 19 and 25.5. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure 5 shows the box plot for the women’s data with detailed labels.

Figure 5. The box plot for the women’s data with detailed labels.

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot, and to examine these details one should create a histogram and/or a stem and leaf display.

Here are some other examples of box plots: Time to move the mouse over a target Draft lottery

Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure 6 are constructed from our data but differ from the previous box plots in several ways.

It does not mark outliers. The means are indicated by green lines rather than plus signs. The mean of all scores is indicated by a gray line. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (31 women and 16 men).

Figure 6. Box plots showing the individual scores and the means.

Each dot in Figure 6 represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to jitter the points. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittererd. Figure 7 shows what jittering looks like.

Figure 7. Box plots with the individual scores jittered.

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.