Chapter 2

Categorical Data Analysis

Section 2.1 Rates

In Section 1.3, we learnt that there are two main types of variables, namely categorical variables and numerical variables. For categorical variables, there are two sub-types, namely ordinal variables and nominal variables. Ordinal variables are those whose categories come with some natural ordering. On the other hand, there is no intrinsic ordering for the nominal variables. For numerical variables, there are those that are continuous and those that are discrete. The focus of this chapter is on categorical variables and we will discuss numerical variables in the next chapter.

Much of the discussion in this section is centred around the following example.

Example 2.1.1 Suppose a patient newly diagnosed with kidney stones visits his urologist for the first time since diagnosis to discuss what are some of the best possible treatments that he should undergo. In preparation, the urologist took out some historical records of the various patients he had previously and summarised the data into a table. Part of the table is shown below.

Size of stone	Gender of patient	Treatment type (X or Y)	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

Each row of the table is a particular patient that the urologist had seen previously and the columns are the variables related to each patient. While the table only shows the first 6 cases, the data in actual fact contains 1050 observations (or data points). The four variables are

1. The size of the kidney stone. This is an ordinal categorical variable that has two categories. The kidney stones can be classified as either small or large.

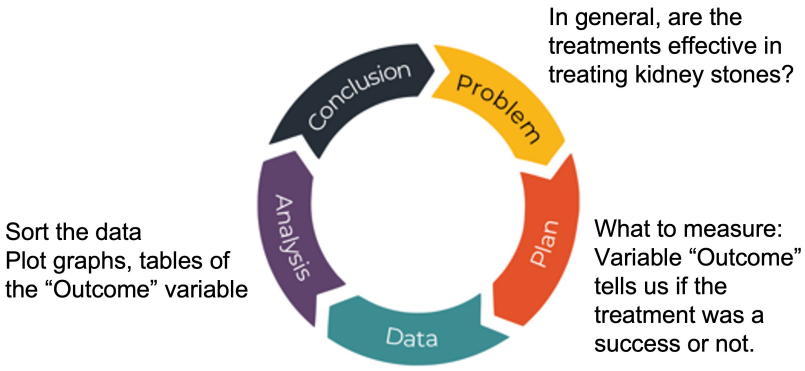
2. The gender of the patient. This is a nominal categorical variable that has two categories, male or female.

3. The treatment that the patient underwent. Again, this is a nominal categorical variable and there are two categories, namely treatment X and treatment Y.

4. The outcome of the treatment is also a nominal categorical variable. The categories are success and failure.

How should the urologist use the 1050 observations to assist in the decision for this new patient?

Before we continue, let us recall the PPDAC cycle that was introduced as the main process behind the approach to a data driven problem.



The overarching question faced by the urologist is simply how to treat his patient better. In particular, **this** new patient. What kind of insights does the data set give the urologist that will enable him to better advise his patient?

To apply the PPDAC cycle to this context, let us start with a question that we want to answer. A simple question to start with is:

- ***Question 1****: Are the treatments given to the patients successful? In other words, should this new patient receive treatment?*

Moving on from "Problem" to "Plan", we next determine what are the variables that needs to be measured and then proceed to obtain data on 1050 previous cases where the *outcome* of the treatment was recorded as either a success or failure. The PPDAC cycle is a continuous process where after looking at the data, drawing some preliminary conclusions might lead to more questions, some of which were even considered from the start. This stage of analysis involves sorting the data, tabulating and plotting graphs of the outcome variable. We may observe interesting trends and this leads us to asking more questions on those trends, leading us back to the top of the cycle again. Some of the new questions that we can ask includes

- Do males undergoing treatment X have a higher *rate*[1] of success than females?
- Does treating large kidney stones with treatment X have a higher rate of success than treatment Y?

We will now discuss some of the tables and charts that can be generated from the data that will give us useful information.

Example 2.1.2 (Analysing 1 categorical variable using a table.) Suppose out of the 1050 previous patients, there were 831 records of **success** and 219 records of **failure** after treatment was given. Thus from this simple collation, a preliminary conclusion is that we should generally recommend the new patient to go for treatment since there are more successful outcomes than failed outcomes. We can present this information on the number of success and failures in a table, together with two other columns, namely *rate* and *percentage*.

Categories of the "Outcome" variable	Count	Rate	Percentage
Success	831	rate(Success) $= \frac{831}{1050} = 0.791$	$0.791 \times 100\% = 79.1\%$
Failure	219	rate(Failure) $= \frac{219}{1050} = 0.209$	$0.209 \times 100\% = 20.9\%$
Total	1050	$\frac{1050}{1050} = 1$	$1 \times 100\% = 100\%$

The *rate* of successful treatments is simply

$$\frac{\text{Number of successful treatments}}{\text{Total number of treatments}} = \frac{831}{1050} = 0.791.$$

We can also represent this as a *percentage* of total treatments that were successful, which is 79.1%. Similarly, the rate of failed treatments is

$$\frac{\text{Number of failed treatments}}{\text{Total number of treatments}} = \frac{219}{1050} = 0.209.$$

When represented as a percentage, the percentage of failed treatments is 20.9%.

[1]The concept of rates will soon be discussed in this section.

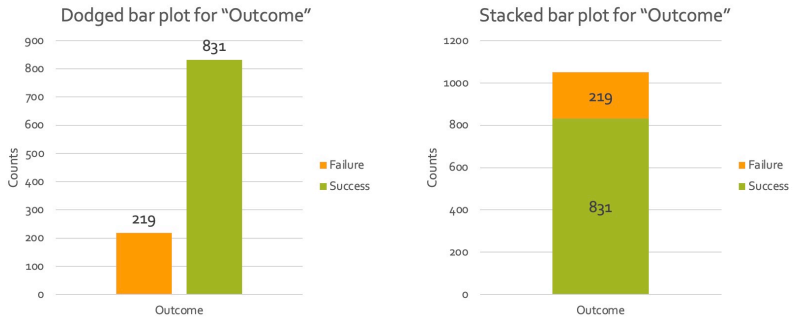
For much of this chapter, we will be using rates in our discussion of the behaviour of categorical variables. Intuitively, we can also think of rate as a fraction, proportion or a percentage. This is useful for understanding some of its properties. For example, we note that

$$0\% \leq \text{rate}(X) \leq 100\% \qquad \text{(if we think of rate as a percentage); or}$$

$$0 \leq \text{rate}(X) \leq 1 \qquad \text{(if we think of rate as a fraction).}$$

(Here, X is some variable of interest.)

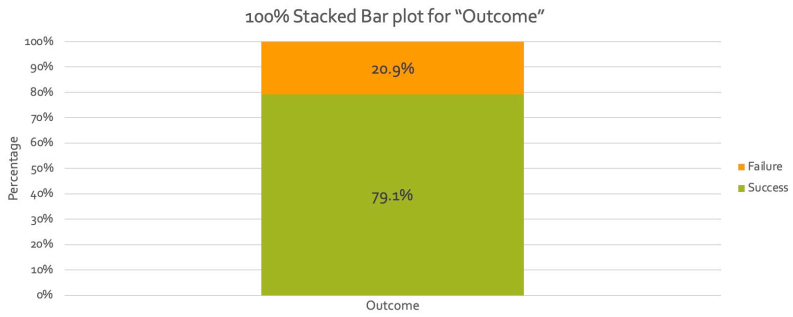
Example 2.1.3 (Analysing 1 categorical variable using a plot.) As an alternative to a table, we can also use easily available softwares to create a plot that presents the data.



Since we are interested in the variable **Outcome**, which is a categorical variable, we can illustrate the *counts* in each of the category "Success" and "Failure" in the form of a bar plot. The two bar plots above are created using Microsoft Excel.

The bar plot on the left is known as a *dodged bar plot*. The x-axis indicates the variable Outcome whereas the y-axis shows the number (that is, the count) of successes and failures in the variable Outcome. Two bars, one for success counts and the other for failure counts, are placed next to each other. Such an illustration is useful in comparing the relative numbers in the categories.

The bar plot on the right is known as a *stacked bar plot*. The x and y-axes are similar to the dodged bar plot but instead of two bars, we now have only one bar where the counts of failure (219) is stacked on top of the counts of success (831). Such an illustration is useful in comparing the occurrences of each category as a *percentage* or *fraction* of the total number of responses. Instead of showing the absolute numbers in each category, it is also possible to show the percentage directly in the plot itself, as seen in the figure below. However, it should be noted that the y-axis is now giving the percentage rather than the actual numbers.



Regardless of which bar plot is used, we can see that there are many more successes than failures and based on this, it is reasonable to recommend our patient to go for some form of treatment based on the information that we have at this stage.

Remark 2.1.4 In this example on treatment of kidney stones, the success of any treatment is defined as having the kidney stones removed or reduced significantly so that it does not pose any further threat to the patient. On the other hand, failure means that the stones were not able to be removed. In general, kidney stones cause little morbidity and mortality. It is useful to note that for other kinds of illness, where treatments have higher stakes, the conclusion may be different.

Now that we are rather convinced that the new patient should receive treatment, the PPDAC cycle brings us back with new questions that arise from our investigation into the data set of 1050 previous patients. It is reasonable to ask the next question as follows:

- ***Question 2**: There are two types of treatment, namely X and Y. Which treatment type is better for our new patient?*

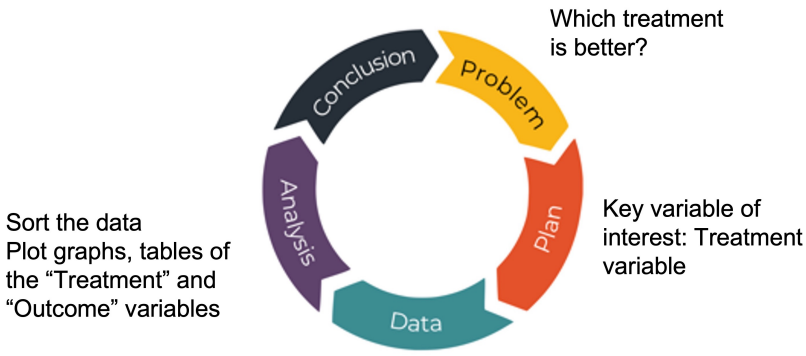
To answer this question, we can revisit the PPDAC cycle and define a new **problem** and **plan** to look at new variable(s) of interest and **analyse** the data again using plots that we have introduced previously.

1. The new problem is as stated above, namely, which treatment is better for our new patient.

2. This means that the key variable that we should look at is the *treatment type* categorical variable, which has two categories, treatment X and treatment Y.

3. This does not mean that treatment type is the only variable of interest, but rather, it should be investigated together with the outcome variable. This is because we want to know how the treatment type affects the outcome.

This leads us to our discussion of how to analyse two categorical variables.



Example 2.1.5 (Analysing 2 categorical variables using a table.) When we used a table to analyse 1 categorical variable (Outcome), the table showed only the number of successes and failures among the 1050 previously treated patients. When we introduce a second categorical variable (Treatment type), we have a 2×2 *contingency table* that will summarise the two variables across the 4 (that's why it is called 2×2) possible combinations of (Treatment, Outcome).

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

Recall that out of 1050 previous patients, 831 underwent successful treatment while the other 219 were failed treatments. The 2×2 table breaks down the 831 successful treatments according to the treatment type. As seen from the *Success* column, 542 were given treatment X while 289 were given treatment Y. Similarly, for the 219 failed treatments, 158 were given treatment X while 61 were given treatment Y.

If we look across a row instead of down a column, we could, for example, see that there were 700 previous patients given treatment X, of which 542 were successful and 158 failed. Similarly, looking at the row for treatment Y, we see that out of 350 people who underwent treatment Y, 289 of them had successful treatments while 61 did not.

Remark 2.1.6

1. It should be noted that by convention, the dependent variable *Outcome* is placed on the columns on the table while the independent variable *Treatment type* is placed on the rows.

2. The column total values for the success (831) and failures (219) columns should add up to the same values as the sum of the row total values for Treatment X (700) and Treatment Y (350), which obviously should both add up to the total number of data points in the data set which is 1050.

Discussion 2.1.7 In order to answer ***Question 2***, it will be useful to ask other related questions, for example:

1. ***Question 2a***: What proportion of the total number of patients were given treatment Y (or X)?

2. ***Question 2b***: Among those patients given treatment X, what proportion were successful?

3. ***Question 2c***: What proportion of patients were given treatment Y and had a failed treatment outcome?

To answer ***Question 2a***, we note that there were 350 previous patients who underwent treatment Y. The *proportion* of the total number of patients that underwent treatment Y is

$$\frac{350}{1050} = \frac{1}{3} = 33\frac{1}{3}\%.$$

We can also denote this as

$$\text{rate(Y)} = \tfrac{1}{3} \text{ or } 33\tfrac{1}{3}\%.$$

We have seen earlier that out of 1050 patients, there were 831 successful treatments, so we can write rate(Success) = 0.791 or 79.1%. We know that

$$\text{rate(X)} = \tfrac{700}{1050} = \tfrac{2}{3} \text{ or } 66\tfrac{2}{3}\%.$$

Outcome / Treatment	Success	Failure	Row Total
X	542	158	700
Y	289	61	**350**
Column Total	**831**	219	**1050**

Notice that in the calculations above, we have used two numbers in the margin of the table (for example, **831** and **1050**) that relate to just one of the categorical variables (Outcome) each time, we call these *marginal rates*. Similarly, rate(Y) $= \frac{350}{1050} = \frac{1}{3}$ is also a marginal rate.

How should we answer ***Question 2b***? In this case, we need to zoom in onto the patients who had undergone treatment X and figure out what proportion of them have had a successful treatment.

Referring to the table again, we see that out of 700 patients who were given treatment X, 542 of them were successfully treated. Hence the proportion of successful treatments was

$$\frac{542}{700} = 0.774 = 77.4\%.$$

This *rate* of success is computed based on only those patients who were under treatment X, which sets the *condition* for the calculation of the rate. Once such a condition is set, those patients on treatment X will be considered as the population and those on treatment Y will not be part of any consideration. Such a rate is known as a *conditional rate*, which is one that is based on a given condition.

A note on the notation used for conditional rates is that we replace the word "given" by a vertical bar so that rate(Success given treatment X) is written as

$$\text{rate(Success} \mid \text{X)}.$$

Let us consider ***Question 2c***. From the table, we can see easily that there are 61 cases where treatment Y was given but had an unsuccessful outcome.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	**61**	350
Column Total	831	219	1050

So the rate of patients who were given treatment Y **AND** had a failure was

$$\frac{61}{1050} = 0.0581 = 5.81\%.$$

This rate is known as a *joint rate* and it is not a conditional rate since we are looking at all the 1050 patients as our baseline. In other words, we are now considering patients on treatment X, as well as patients on treatment Y as the population. One should be careful with the implicit difference in the phrasing of the two statements:

- What proportion of patients were given treatment Y and had an unsuccessful outcome?

  Answer: $\text{rate(Unsuccessful and Y)} = \frac{61}{1050} = 0.0581.$

- What proportion of patients given treatment Y had an unsuccessful outcome?

  Answer: $\text{rate(Unsuccessful} \mid \text{Y)} = \frac{61}{350} = 0.174.$

The first question refers to the joint rate/proportion/percentage while the second question refers to the conditional rate/proportion/percentage.

Discussion 2.1.8 It should be clear at this point from our discussion of rates and proportions that our decision on which treatment to suggest to our new patient cannot be based simply on the absolute number of successes and failures for each treatment type. If we had based the decision on absolute numbers, we would have gone for treatment X since there were 542 success cases compared to only 289 for treatment Y.

The reason why we should look at rates rather than absolute numbers is because the number of patients undergoing each treatment is **different**, so it would not be surprising if there were more successful cases for treatment X because there were just more patients given this treatment, rather than because it is more effective. Finding the rate of success for each treatment before comparing them is a form of *normalisation*. At this stage of our analysis, when using the success rates to compare the treatment types, our conclusion is to recommend treatment Y to our patient.

- The rate of success *given* treatment X is the conditional rate we have already calculated in answering ***Question 2b***, which is

$$\text{rate}(\text{Success} \mid \text{X}) = \frac{542}{700} = 0.774 = 77.4\%.$$

  Similarly, we can calculate the rate of success *given* treatment Y, which is

$$\text{rate}(\text{Success} \mid \text{Y}) = \frac{289}{350} = 0.826 = 82.6\%.$$

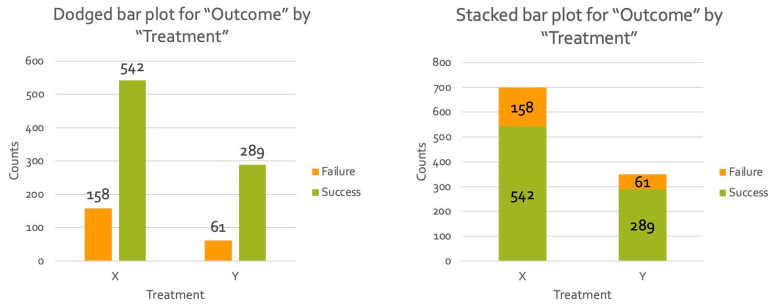
- We can also look at the conditional rates in another way. For treatment X, having the conditional rate of success to be 77.4% means that out of 100 patients who underwent treatment X, 77 of them had a successful outcome. For treatment Y, the numbers were 83 successes out of 100 patients receiving this treatment.

- As the rate of success for treatment Y is higher, we can now say that treatment Y is better than treatment X and advise the patient appropriately. Notice that we would have given the opposite advice if we were looking at absolute numbers instead of rates, which is incorrect.

- We can now add in the rates to the 2×2 contingency table given at the beginning of Example 2.1.5.

Outcome / Treatment	Success	Failure	Row Total
X	542 (77.4%)	158 (22.6%)	700 (100%)
Y	289 (82.6%)	61 (17.4%)	350 (100%)
Column Total	831(79.1%)	219 (20.9%)	1050 (100%)

Example 2.1.9 (Analysing 2 categorical variables using a plot.) In Example 2.1.3, we introduced dodged bar plots and stacked bar plots to present the data on a single variable **Outcome**. We can also use these plots to present the counts of **Outcome** broken down by **Treatment**. These were the two variables we analysed using a table in Example 2.1.5.

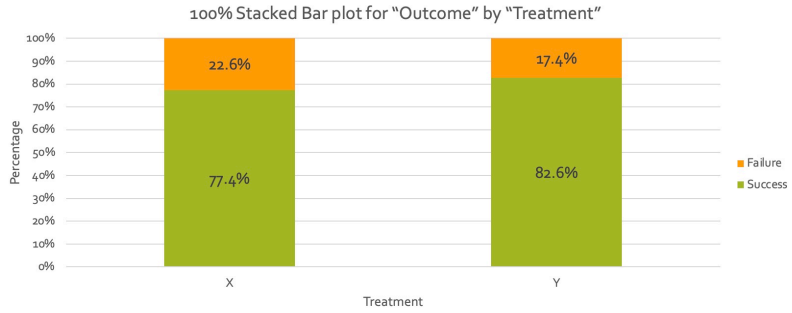


The dodged bar plot on the left shows that success and failure counts for both treatments X and Y. The numbers above each bar is the success or failure count for that particular treatment type.

The stacked bar plot on the right shows the same information but with the success and failure bars under the same treatment stacked instead of being placed side by side.

Both bar plots tell us that there are a lot more successful treatments in treatment X than in treatment Y, which may lead to the conclusion that treatment X is more effective (since the green bars for treatment X are bigger than the green bars for treatment Y). However, it is also obvious from the stacked bar plot that these two treatments have very different number of patients (represented by the height of the two bars).

Similar to our analysis using tables, we can also create the plots using rates instead of absolute numbers.



In this plot, notice that both the treatment X and treatment Y bars have been *normalised* to the same *height* (which is 100%). We are no longer comparing absolute numbers, but instead comparing the rates of success (the height of the green bars, as a proportion of the total height) between the two treatments. We can see immediately that treatment Y has a higher rate of success (taller green bar) compared to treatment X.

To summarise this section, we have discussed how we can analyse two categorical variables. This can be done either using a 2×2 contingency table, or bar plots (dodged or stacked) which makes it easier for us to observe any differences between the categories. We also introduced the concept of rates, as a means of fair comparison when group sizes are unequal. To formally discuss the relationship between two categorical variables, we will introduce the concept of *association* in the next section.

Section 2.2 Association

Definition 2.2.1 In Section 2.1, we considered the example of two different treatments for patients with kidney stones. Let's say that initially, we guessed that the treatment type involved does not affect the outcome of the treatment, meaning that we could advise our patient to undergo either treatment because the outcome would not be affected. If this was the case, then we can say that the treatment type is not related to the outcome of the treatment.

After analysing the data using rates, we found that this was not the case. There was a higher success rate observed for patients under treatment Y compared to those under treatment X. Due to the difference in success rates, we say that there is a *relationship* between the type of treatment and the outcome of the treatment.

To formalise the notion of such a *relationship*, we say that treatment type is *associated* with the outcome of the treatment. More specifically, treatment Y is *positively associated* with the success of the treatment. What this means is that treatment Y and successful treatments tend to occur together.

On the other hand, we say that treatment X is *negatively associated* with the success of the treatment. This is because we tend to see treatment X and failed treatments go hand in hand.

Remark 2.2.2

1. Note that treatment X is negatively associated with the success of the treatment does not mean that a significant proportion of patients undergoing treatment X will see the treatment fail (77.4% of them still recorded success). The negative association is stated as a **comparison** between the two treatment types X and Y, where in this case treatment Y tends to produce **more successful outcomes**.

2. We should be conscious of the choice of the word **associated** because we do not know if the outcome of the treatment was entirely **due to** the treatment type received. The data we had came from an observational study hence it might be erroneous for us to say that the type of treatment and the outcome of the treatment have a *causal* relationship. It is important to see the distinction between *association* and *causation* and for the rest of this chapter, we will be focussing on discussing associative relationships between categorical variables rather than causal relationships.

Discussion 2.2.3 So how do we identify an association between two variables? Suppose the two variables we are considering represent two characteristics in a population. Let us call these two characteristics A and B. For example, A could be *smoker* (so one categorical variable could be smoking habit, with two categories *smoker* and *non-smoker*) while B could be *male* (so the other categorical variable could be gender, with two categories *male* and *female*). The population can be a well-defined group of people. In the population, those "with A", refers to smokers, while "without A", denoted by NA refers to non-smokers. Similarly, those "with B" refers to male and those "without B", denoted by NB, refers to female.

So if the rate of A given B (proportion of smokers among males) is the same as the rate of A given NB (proportion of smokers among females), then it means that the rate of A is not affected by the presence or absence of B. Thus in this case, there is no difference in the proportion of smokers between both gender groups and we write

$$\text{rate}(\text{A} \mid \text{B}) = \text{rate}(\text{A} \mid \text{NB}).$$

However, if the rate of A given B is not the same as the rate of A given NB, then there are two possible situations.

- The first possibility is the rate of A given B is more than the rate of A given NB. This means that the presence of A **when B is present** is stronger compared to **when B is absent**. Hence we say that there is *positive association* between A and B. In this case, we write

  $$\text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB})$$

  and for the gender/smoking example, this means that there is a higher proportion of smokers among males than the proportion of smokers among females. So being male and smoking are positively associated.

- The other possibility is the rate of A given B is less than the rate of A given NB. This means that the presence of A **when B is present** is weaker compared to **when B is absent**. Hence we say that there is *negative association* between A and B. In this case, we write

  $$\text{rate}(\text{A} \mid \text{B}) < \text{rate}(\text{A} \mid \text{NB})$$

and for the gender/smoking example, this means that there is a lower proportion of smokers among males than the proportion of smokers among females. So being male and smoking are negatively associated.

Example 2.2.4 Let us revisit the earlier example on two different treatments for kidney stones. Recall that the two variables were treatment outcome and treatment type. For the treatment outcome variable, let us split the patients into group A, which is the group of patients with successful outcomes and the group NA will be those with unsuccessful outcomes. For the other variable, we will also split the patients into group B for those given treatment X and group NB for those given treatment Y.

Let us revisit some conditional rates that were computed previously.

1. Rate(A | B) = rate(Success | X) = $\frac{542}{700} = 0.774$.
2. Rate (A | NB) = rate(Success | Y) = $\frac{289}{350} = 0.826$.

Since

$$\text{rate(A | B)} < \text{rate(A | NB)},$$

we can say that the presence of A when B is present is weaker than the presence of A when B is absent. Thus there are fewer successful treatments when looking at treatment X compared to treatment Y and hence success of the treatment is *negatively* associated with treatment X.

Conversely, since there are more successful treatments when looking at treatment Y compared to treatment X, we can conclude that success of the treatment is *positively* associated with treatment Y.

Section 2.3 Two rules on rates

Discussion 2.3.1 In this section, we will discuss two important rules regarding rates. Suppose we have a population with two population characteristics A and B. Among the population there are those who possess characteristic A and those who do not.

For ease of notation, we will denote those who possess characteristic A simply as "A" and those who do not as "NA". Similarly for characteristic B, those in the population who possess this characteristic will be denoted as "B" and those who do not as "NB".

(Symmetry rule)

The first rule that we will be discussing is known as the *symmetry rule*. Although there are three parts to this rule, once we can understand the first part, the second and third parts will follow naturally.

Symmetry Rule Part 1:

$$\text{rate}(\text{A} \mid \text{B}) > \text{rate}(\text{A} \mid \text{NB}) \Leftrightarrow \text{rate}(\text{B} \mid \text{A}) > \text{rate}(\text{B} \mid \text{NA}).$$

The above rule states that the rate of A given B is **more than** the rate of A given NB (call this statement 1) **if and only if** the rate of B given A is **more than** the rate of B given NA (call this statement 2). The *if and only if* here, denoted by $\Leftrightarrow$, means that statements 1 and 2 happens together, meaning that if one of the statements is true, the other one will also be true. In other words, the two statements are either *both correct* or *both incorrect*. Another way of understanding (statement 1) if and only if (statement 2) is

- If (statement 1) holds, then (statement 2) must hold; AND
- If (statement 2) holds, then (statement 1) must hold.

Suppose we know that

$$\text{rate}(\text{A} \mid \text{B}) \text{ is more than } \text{rate}(\text{A} \mid \text{NB}), \tag{1}$$

then we can safely say that

$$\text{rate}(\text{B} \mid \text{A}) \text{ is more than } \text{rate}(\text{B} \mid \text{NA}). \tag{2}$$

Why is this so? Let us try to explain this logically.

1. If rate(A | B) is more than rate(A | NB), which is (1), then this means that there is a positive association between A and B;
2. This means that we are more likely to see A when B is present, compared to when B is absent;
3. This in turn means that we are more likely to see B when A is present, compared to when A is absent;
4. Hence rate(B | A) is more than rate(B | NA), which is (2).

This is the same as saying that A and B are positively associated. Conversely, suppose we know that

$$\text{rate}(\text{B} \mid \text{A}) \text{ is more than } \text{rate}(\text{B} \mid \text{NA}), \tag{2}$$

then we can safely say that

$$\text{rate}(\text{A} \mid \text{B}) \text{ is more than } \text{rate}(\text{A} \mid \text{NB}). \tag{1}$$

The logical explanation is similar in nature.

1. If rate(B | A) is more than rate(B | NA), which is (2), then this means that there is positive association between B and A;
2. This means that we are more likely to see B when A is present, compared to when A is absent;
3. This in turn means that we are more likely to see A when B is present, compared to when B is absent;
4. Hence rate(A | B) is more than rate(A | NB), which is (1).

We have now seen Part 1 of the Symmetry Rule. Parts 2 and 3, as shown below can be similarly explained. You are encouraged to go through the logical thought process behind these two parts.

Symmetry Rule Part 1:

$$\text{rate(A | B)} > \text{rate(A | NB)} \Leftrightarrow \text{rate(B | A)} > \text{rate(B | NA)}.$$

Symmetry Rule Part 2:

$$\text{rate(A | B)} < \text{rate(A | NB)} \Leftrightarrow \text{rate(B | A)} < \text{rate(B | NA)}.$$

Symmetry Rule Part 3:

$$\text{rate(A | B)} = \text{rate(A | NB)} \Leftrightarrow \text{rate(B | A)} = \text{rate(B | NA)}.$$

Example 2.3.2 Let us revisit our kidney stones treatment example. The 2×2 contingency table below gives us the number of patients in each treatment type as well as the number of success and failure outcomes for each treatment type.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

We have earlier shown that A (representing treatment outcome) is associated with B (representing treatment type) since

$$\begin{aligned} \text{rate(A | B)} &= \text{rate(Success | X)} \\ &= 0.774 \\ &< 0.826 = \text{rate(Success | Y)} = \text{rate(A | NB)}. \end{aligned}$$

By symmetry rule part 2, we should have rate(B | A) < rate(B | NA). Let us verify that this is indeed the case.

$$\begin{aligned}
\text{rate(B | A)} &= \text{rate(X | Success)} \\
&= \frac{542}{831} \quad \text{(since there are 831 successful cases of which 542 came from treatment X)} \\
&= 0.652 \\
\text{rate(B | NA)} &= \text{rate(X | Failure)} \\
&= \frac{158}{219} \quad \text{(since there are 219 failure cases of which 158 came from treatment X)} \\
&= 0.721.
\end{aligned}$$

Since $0.652 < 0.721$, we have thus verified that rate(B | A) < rate(B | NA) as predicted by symmetry rule part 2. This also confirms that there is negative association between success of treatment (A) and treatment X (B).

Discussion 2.3.3 (Basic rule on rates.) The second rule on rates is known as the *basic rule on rates*. The main rule, as well as three consequences of the main rule are shown below.

Basic rule on rates:

The overall rate(A) will always lie between rate(A | B) and rate(A | NB).

Consequence 1:

The closer rate(B) is to 100%, the closer rate(A) is to rate(A | B).

Consequence 2:

If rate(B) = 50%, then rate(A) = $\frac{1}{2}$[rate(A |B) + rate(A | NB)].

Consequence 3:

If rate(A | B) = rate(A | NB), then rate(A) = rate(A | B) = rate(A | NB).

1. The basic rule on rates states that the overall rate(A) is always between the conditional rates of A given B and A given not B.
2. The first consequence gives us a little more indication of **where** the overall rate(A) is going to be. If rate(B) is closer to 100% (than rate(NB)), then rate(A) is going to be closer to rate(A | B) compared to rate(A | NB).

3. The second consequence specifically states that if rate(B) is **exactly** 50%, then rate(A) will be **exactly** the mid point between rate(A | B) and rate(A | NB).

4. Finally, the third consequence states that if the two conditional rates, namely rate(A | B) and rate(A | NB) are the same, then the overall rate(A) will also take the same value of the two conditional rates.

At this point, the significance of the basic rule and its consequences are not immediately apparent or intuitive. The best way towards understanding them is through an example.

Example 2.3.4 Suppose a school has two different classes (call them class Bravo and class Charlie) of students who took the same data analytics examination. We are interested in studying the passing rate of students at the entire school level and also at each individual class level. Suppose we are given the following information:

(1) The passing rate of students from Bravo is 75%.

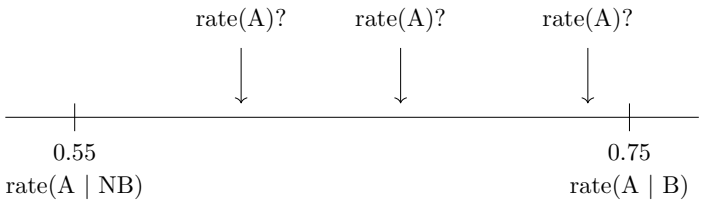
(2) The passing rate of students from Charlie is 55%.

For convenience, let us denote class Bravo as "B" and class Charlie as "NB" (not B). Similarly, denote passing the examination as "A" and not passing as "NA". So the two pieces of information we have above are:

$$\text{rate(A | B)} = 0.75 \quad \text{and} \quad \text{rate(A | NB)} = 0.55.$$

By the basic rule on rates, the overall passing rate of all students from the school, that is, rate(A) is between the two conditional rates,

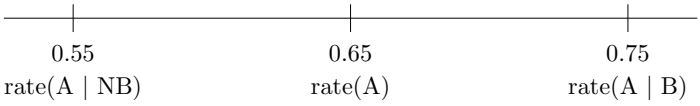
$$0.55 = \text{rate(A | NB)} \leq \text{rate(A)} \leq \text{rate(A | B)} = 0.75.$$



However, without any further information, we will not be able to determine the exact value of rate(A). What about the three consequences of the basic rule?

1. The first consequence states that the closer rate(B) is to 100%, then the closer rate(A) will be to rate(A | B). In our example, it means that if the number of students in Bravo is far more than the number of students in Charlie (thus rate(B) is closer to 100%), then the overall passing rate of the school (that is, rate(A)) will be closer to the passing rate of class Bravo (that is, rate(A | B)).

2. The second consequence states that if rate(B) = 50% (which also means that rate(NB) = 50%), then
$$\text{rate(A)} = \frac{1}{2}\left[\text{rate(A} \mid \text{B)} + \text{rate(A} \mid \text{NB)}\right].$$
That is, rate(A) will be right in between the two conditional rates. In our example, this means that if the number of students in Bravo and Charlie are exactly the same, then the overall passing rate of the school will be exactly in between 0.55 and 0.75, that is 0.65.



3. The third consequence states that if the two conditional rates rate(A | B) and rate(A | NB) are the same, then the overall rate(A) will be the same value as the two conditional rates. In our example, if the passing rates of class Bravo and class Charlie are the same, then the overall passing rate of the school will be the same as the passing rate in either class.

Example 2.3.5 Let us continue with Example 2.3.4 and validate the general rule of rates and the consequences by considering actual numbers.

1. Suppose the total number of students and the number of passes in each of the two classes are given in the table below.

	Total number of students	Number of passes	Passing rate
Bravo	600	450	$\frac{450}{600} = 0.75$
Charlie	80	44	$\frac{44}{80} = 0.55$
School	680	494	$\frac{494}{680} = 0.73$

Notice that the passing rates of both classes are what they are supposed to be, but the number of students in Bravo far exceeds the number in Charlie (so rate(B) is $\frac{450}{494}$ which is close to 100%). While the overall school passing rate is between 0.55 and 0.75 (in accordance to the general rule of rates), it is much closer to the passing rate of Bravo, as predicted by consequence 1.

2. Suppose the total number of students and the number of passes in each of the two classes are as given below instead:

	Total number of students	Number of passes	Passing rate
Bravo	600	450	$\frac{450}{600} = 0.75$
Charlie	600	330	$\frac{330}{600} = 0.55$
School	1200	780	$\frac{780}{1200} = 0.65$

   Again, the passing rates of both classes are what they are supposed to be, but in this case, the number of students in Bravo and Charlie are the same (so rate(B) = rate(NB) = 0.5). As predicted by consequence 2, the overall school passing rate will be 0.65, which is right in between the two class passing rates.

3. To illustrate consequence 3, suppose the passing rates for both classes are the same, as shown below.

	Total number of students	Number of passes	Passing rate
Bravo	600	450	$\frac{450}{600} = 0.75$
Charlie	400	300	$\frac{300}{400} = 0.75$
School	1000	750	$\frac{750}{1000} = 0.75$

   Now the two conditional rates, namely rate(A | B) and rate(A | NB) are equal. By consequence 3, rate(A) will be the same value as the two conditional rates. This is indeed the case as we see that the two classes have the same passing rate which will result in the school having the same passing rate of 0.75. It is important to note that we **do not require** rate(B) to be the same as rate(NB) for consequence 3 to hold. For our example, this means that we do not require class Bravo and Charlie to have the same number of students. As long as the two class passing rates are the same, consequence 3 will hold.

Finally, let us verify the rule on rates using our kidney stones data set.

Example 2.3.6 We have seen the following table from Example 2.3.2.

Treatment \ Outcome	Success	Failure	Row Total
X	542	158	700
Y	289	61	350
Column Total	831	219	1050

- The conditional rates of success among the two treatment types are:

$$\text{rate(Success} \mid \text{X)} = \frac{542}{700} = 0.774,$$

$$\text{rate(Success} \mid \text{Y)} = \frac{289}{350} = 0.826.$$

The overall rate of success is

$$\text{rate(Success)} = \frac{831}{1050} = 0.791,$$

which is closer to the rate of success among patients with treatment X. This agrees with the general rule on rates since there were more patients with treatment X (66.67%) compared to treatment Y (33.33%).

In the three sections of this chapter we have discussed so far, we have seen how we can use the concept of rates to investigate relationships, in particular, association between categorical variables. Very often, exact rates (overall or conditional) are unknown to us but if we can apply some general rules like the symmetry rule, basic rule or the consequences of the basic rule, we can still obtain valuable insights into the data set we have on our hands. Making the best use of limited information is an important skill when analysing data.

In the next section, we will discuss a surprising observation that can be counterintuitive to some but is very important for anyone analysing data to be aware of.

Section 2.4 Simpson's Paradox

Discussion 2.4.1 From earlier sections, when faced with the problem of advising our new kidney stones patient, we have gone through two cycles of the PPDAC process.

The first question we asked was whether having any sort of treatment was better than not having one. By comparing the rate of success (0.791) versus the rate of failure (0.209), we conclude that there are many more successful than failed treatments from past records, so the decision was to advise our new patient that he should be treated.

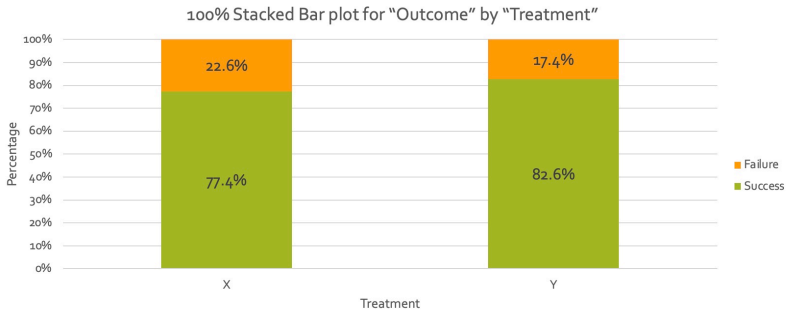
But this led us to the next question, as there are two treatment types available, which type of treatment should we recommend? This made us go back to the data and compare the success rates of those patients who were given treatment X as opposed to the success rates of those given treatment Y. Upon delving deeper into the data, we discovered that treatment Y is positively associated with success rate. This suggests that treatment Y is "better" than treatment X and perhaps we should advise our patient to undergo treatment Y.

Are we done with our analysis? Is there some lingering doubt in our minds that we may be providing wrong advice to our patient? If we are convinced that treatment Y is better, should we **always** send kidney stone patients for treatment Y? If not always, then when do we do so? What should our decision be based on? These are again questions that prompts us to go back to our data and see if more information can be obtained from it.

Size of stone	Gender of patient	Treatment type (X or Y)	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

The table above from Example 2.1.1 shows there are two other variables that we have not used in our analysis thus far, namely the size of the kidney stone and also the gender of the patient. Would these variables be an important factor in our consideration? How should we go about analysing them? Let us begin by exploring the stone size variable.

Example 2.4.2 (Analysing 3 categorical variables using a plot.) In Example 2.1.9, we used a stacked bar plot for "Outcome" by "Treatment" to compare the success rates for treatments X and Y.

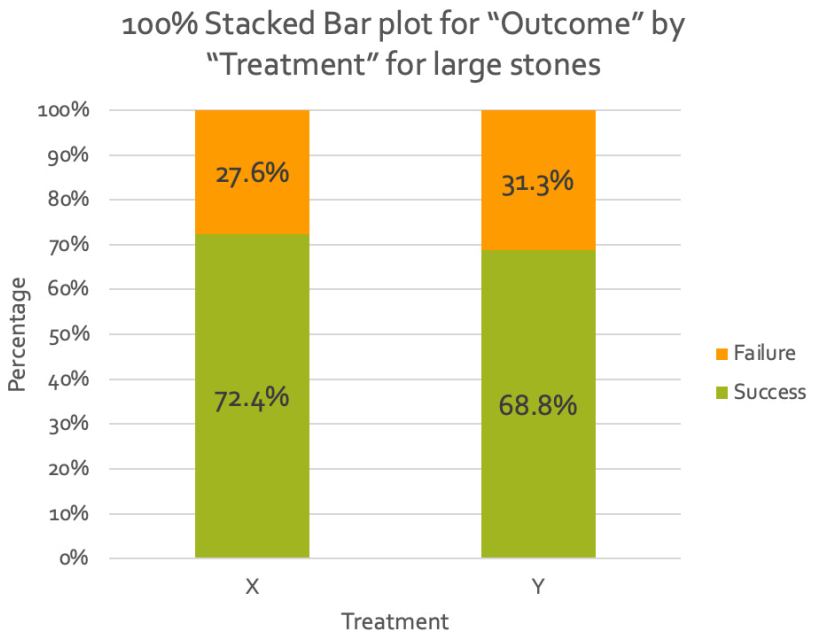


From the stacked bar plot, we have concluded that treatment Y is positively associated with success. We have not taken stone sizes into consideration thus far and the plot was made based on simply counting the number of successes and failures across all stone sizes. In other words, this plot gave us the **overall success rates** of treatments X and Y.

Let us now separate the data by considering the categorical variable of "stone size" which has two categories, namely *large stones* and *small stones*.

Large stones	Success	Failure	Total
X	381	145	526
Y	55	25	80
Total	436	170	606

The table above shows the outcome of treatments given to patients with large stones. For example, out of 526 treatment X patients with large kidney stones, 381 had a successful outcome and 145 were unsuccessful. Similarly, out of 80 treatment Y patients with large kidney stones, 55 were successful while 25 were not. We can present these information using a stacked bar plot like before, as shown below[2].



How do the two different treatments compare? Although the margin of difference is not very big, there is no doubt that treatment X has a *higher* success rate of 72.4% compared to treatment Y, which has a success rate of 68.8%. This means that, for treating large kidney stones,

$$\frac{381}{526} = 0.724 = \text{rate}(\text{Success} \mid \text{X}) > \text{rate}(\text{Success} \mid \text{Y}) = 0.688 = \frac{55}{80},$$

and thus treatment X is positively associated with success for treating large stones. This observation is surprising, since we have already concluded that **for all stone sizes combined together**,

$$\text{rate}(\text{Success} \mid \text{X}) < \text{rate}(\text{Success} \mid \text{Y}),$$

that is, treatment X is negatively associated with success if we do not segregate by stone size.

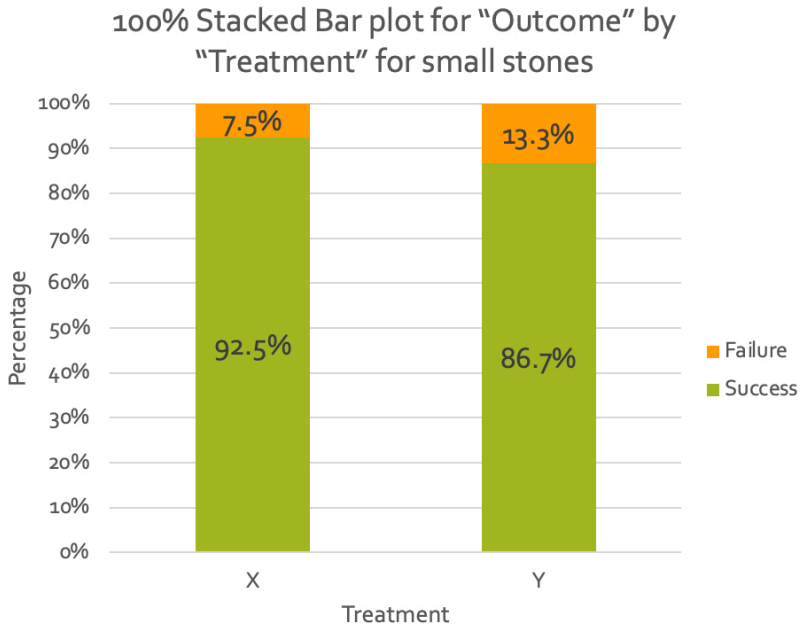
[2]Notice that for the bar plot for treatment Y, the two percentages do not add up to 100%. This is due to rounding off in Excel, where the success percentage is in fact 68.75% and the failure percentage is 31.25%.

Why are we observing a different behaviour for large kidney stones as opposed to what we saw earlier when all kidney stone sizes are combined?

Let us consider the data for small kidney stones.

Small stones	Success	Failure	Total
X	161	13	174
Y	234	36	270
Total	395	49	444

The table above shows the outcome of treatments given to patients with small stones. For example, out of 174 treatment X patients with small kidney stones, 161 had a successful outcome and 13 were unsuccessful. Similarly, out of 270 treatment Y patients with small kidney stones, 234 were successful while 36 were not. Let us again present these data using a stacked bar plot.

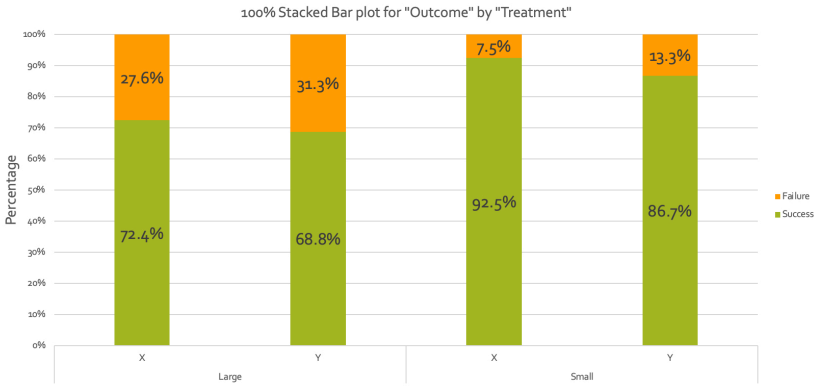


The margin of difference between the two treatment types is again not very big, but again we see that treatment X has a *higher* success rate of 92.5% compared to treatment Y, which has a success rate of 86.7%. This means that, for treating smaller kidney stones,

$$\frac{161}{174} = 0.925 = \text{rate}(\text{Success} \mid \text{X}) > \text{rate}(\text{Success} \mid \text{Y}) = \frac{234}{270} = 0.867,$$

so again treatment X is positively associated with success for treating small stones, which is the opposite from what we had when the data was combined and not segregated by stone size.

We can now combine the two previous plots by putting them side by side as shown below.



Notice that the first two bars from the left are for large kidney stones data while the last two bars are for small kidney stones. This type of plot is sometimes referred to as a *sliced stacked bar plot*. Such a plot can be used for comparing across three categorical variables. The three variables here are stone size, treatment outcome and treatment type.

We are now facing a *paradox*. Although treatment Y is the better treatment overall, when the stone sizes are combined and not segregated, we see that if we focus only on the large stones, or only on the small stones, treatment X is observed to have higher success rate that treatment Y. This is indeed strange!

This phenomenon is known as *Simpson's Paradox*.

Simpson's Paradox:

Simpson's Paradox is a phenomenon in which a trend appears in more than half of the groups of data but disappears or reverses when the groups are combined. Here, "disappears" means the two variables in question (say A and B) are no longer associated, that is, rate(A | B) = rate(A | NB).

We are now back to the same question which we thought we have already answered: Which treatment is better for our patient? Should we advise him to undergo treatment X or Y?

Remark 2.4.3 In the example of kidney stones, there were only two subgroups for the stone size, namely, small and large. We claim that Simpson's Paradox was observed because

the trend in **both** subgroups is different from the trend observed when the subgroups are combined.

In examples where there are more than two subgroups, we will say that Simpson's Paradox is observed as long as a majority of the individual subgroup rates shows the opposite trend to the overall rate. For example, if there are three subgroups, as long as there are at least 2 subgroups showing the opposite trend to the overall rate, we can say that Simpson's Paradox is observed.

Example 2.4.4 (Analysing 3 categorical variables using a table.) Let us put the two tables in Example 2.4.2 for both the large and small kidney stones together into one unified table.

	Large stones			Small stones			Total (Large+Small)		
	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %	Succ.	Total trt.	R(succ.) in %
X	381	**526**	**72.4**%	161	**174**	**92.5**%	542	700	77.4%
Y	55	**80**	68.8%	234	**270**	86.7%	289	350	**82.6**%

To recap, we have two different treatment types, X and Y. In the row for treatment X, we see that there were 526 large stones cases that were under treatment X, of which 381 were successful. This gives a success rate of 72.4%. Similarly, in the row for treatment Y, we see that there were 270 small stones cases that were under treatment Y, of which 234 were successful. This gives a success rate of 86.7%. The last 3 columns of the table gives the combined numbers for both stone sizes.

Recall we had initially concluded that treatment Y was the better treatment because **82.6**% of patients who were given treatment Y had a successful outcome, compared to 77.4% for treatment X. We then separated the cases according to the size of the stone, i.e., we created subgroups and this method of subgroup analysis is called *slicing*.

This is when we observed Simpson's Paradox, where the rate of success amongst small (**92.5**%) and large (**72.4**%) stones is higher for treatment X compared to treatment Y. This **reverses** the trend observed when the small and large kidney stones were combined.

Let us look at the numbers highlighted in blue in the table. A crucial observation at this point is that treatment X seems to be used to treat mostly patients with large stones as compared to small stones. Thus, by the **basic rule of rates**, we know that the overall success rate of treatment X will be closer to the large stones success rate of 72.4% than the small stones success rate of 92.5%. Indeed, we have the overall treatment X success rate to be 77.4%.

Turning our attention to the numbers highlighted in orange in the table, we observe the opposite of the above. Treatment Y seems to be used to treat patients with small stones compared to large stones. Again, by the **basic rule of rates**, we would expect the overall success rate of treatment Y to be closer to the small stones success rate of 86.7% than the

large stones success rate of 68.8%. Indeed, we have the overall treatment Y success rate to be 82.6%.

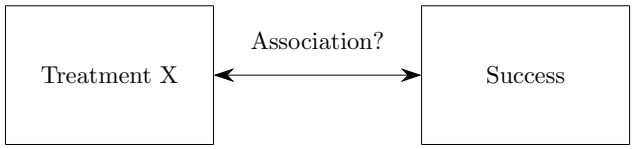
Combining these two observations, it is no wonder that we have the overall success rate of X to be lower than the overall success rate of Y.

Another very telling observation from the table is that the range of success rates for treating large stones is between 68.8% (treatment Y) and 72.4% (treatment X). Compare this with the range of success rates for treating small stones which is between 86.7% (treatment Y) and 92.5% (treatment X). This tells us that treatments for large stones have a lower rate of success compared to small stones, which is not unreasonable to believe.

In conclusion, we can explain Simpson's Paradox in the following way. Treatment X is in fact a better treatment than Y. However, because patients have been using Treatment X to treat more difficult cases (large kidney stones), this lowers the overall success rate of treatment X. It does not change the fact that in the individual subgroups, regardless of stone size, treatment X achieves a higher success rate than treatment Y. **Slicing the data** into the small and large stone subgroups will reveal that treatment X is indeed a better treatment.

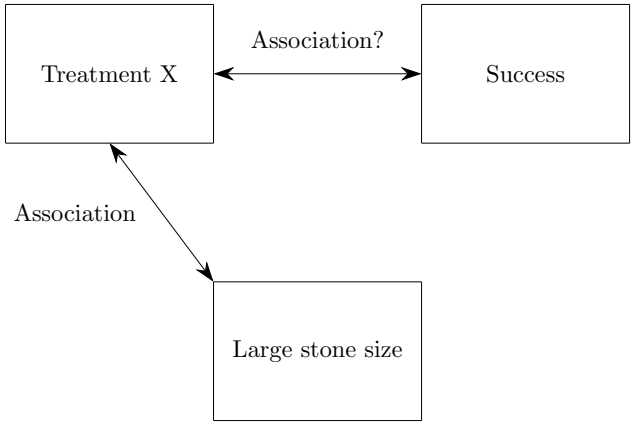
Before we conclude this section, let us recap the story so far.

- We started off with a new kidney stone patient coming to us for advice. Based on past patient records, we were convinced that the success rate of undergoing treatment is higher than the failure rate and thus conclude that the patient should undergo some form of treatment.
- We were then faced with the decision between two treatment types. Treatment X and Treatment Y. In determining which treatment type to recommend to the patient, we looked at the data on hand of past patients and investigated if there was any association (positive or negative) between Treatment X and Treatment Success.

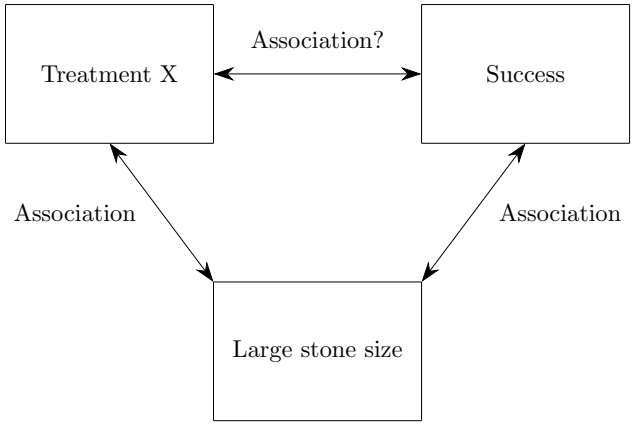


- With further analysis, we found that Treatment X was *negatively associated* with Success. This meant that we should recommend Treatment Y to our new patient. However, through another iteration of the PPDAC cycle, we wondered how another variable like stone size may affect our conclusion.
- By slicing, we segregated our data into past patients with large stone size and others with small stone size. Surprisingly for both subgroups, we found that Treatment X had a higher success rate than Treatment Y. This reversed that trend that we saw when the subgroups were combined.

- More importantly, we observed that Treatment X was used more often in dealing with large stones compared to Treatment Y, which was more frequently used to deal with small stones. This means that large stone size is likely to be associated with Treatment X.



- On the other hand, we also observed that patients with large stones have a lower success rate (regardless of treatment type) compared to patients with small stones. This is perfectly reasonable and thus also suggests that large stone size is likely to be associated with treatment success.



- This means that stone size is a (third) variable that was associated with the other two variables whose relationship we were initially investigating, thus affecting the conclusion of our initial study. Such a variable is called a *confounder* and they will be the focus of our discussion in the next section.

- For now, we will note that when Simpson's Paradox is observed, it implies that there is definitely a confounding variable present, that is a third variable that is associated with the two variables whose relationship we are investigating. However, the existence of a confounder does not necessarily lead to us observing Simpson's Paradox.

Section 2.5 Confounders

Discussion 2.5.1 Continuing our kidney stones patients example, we were fortunate that the data set contained information that may not seem to be important initially. Without performing further investigation into the size of the kidney stones, we could have ended up giving the wrong recommendation to our new patient.

In data collection, it is often important to collect more information on the subjects in addition to those variables that are immediately apparent to be of importance. This is because we can never be sure whether there would be some other variables that may be confounder that would influence our study of association between two variables of interest. Of course, as the owner of the study, we can ask our subjects (for example, in a survey) as many questions and collect as much data as we want, but practically, we also know that respondents do not like to see a long list of seemingly unrelated questions in surveys. There are also cost considerations if we collect more data than necessary. To design a good study, we need to strike a balance between the two.

Definition 2.5.2 A *confounder* is a third variable that is associated with both the independent and dependent variables whose relationship we are investigating. Note that we do not specify the direction (positive or negative) of association here. As long as the variable is associated in some way to the main variables, we will call it a confounder, or a *confounding variable*.

Example 2.5.3 At the end of the previous section, we explained how the variable kidney stone size is a confounding variable because it is associated with both the (independent) variable Treatment type and (dependent) variable Treatment outcome. Let us now work through the calculations to justify these associations. First, let us show that stone size is associated with treatment type.

Treatment	Large	Small	Total
X	526	174	700
Y	80	270	350
Total	606	444	1050

The table shows the number of large and small stones treated by treatments X and Y respectively. Out of 700 cases treated by treatment X, 526 were large stones and 174 were small stones. Out of 350 cases treated by treatment Y, 80 were large stones and 270 were small stones. Since

$$\text{rate}(\text{Large} \mid \text{X}) = \frac{526}{700} = 0.751 \quad \text{and} \quad \text{rate}(\text{Large} \mid \text{Y}) = \frac{80}{350} = 0.229,$$

we see that

$$0.751 = \text{rate}(\text{Large} \mid \text{X}) > \text{rate}(\text{Large} \mid \text{Y}) = 0.229,$$

and so large stones are positively associated with treatment X. This means that there is a higher proportion of large stones being treated by treatment X compared to treatment Y.

Now let us turn our attention to the association between stone size and treatment outcome.

Stone size	Success	Failure	Total
Large	436	170	606
Small	395	49	444
Total	831	219	1050

This table shows the number of success and failure outcomes for patients with large and small stones. Out of 606 large stones cases, 436 were successfully treated while 170 were not successful. Out of 444 small stones cases, 395 were successfully treated while 49 were not successful. Since

$$\text{rate}(\text{Success} \mid \text{Large}) = \frac{436}{606} = 0.719 \quad \text{and} \quad \text{rate}(\text{Success} \mid \text{Small}) = \frac{395}{444} = 0.890,$$

we see that

$$0.719 = \text{rate}(\text{Success} \mid \text{Large}) < \text{rate}(\text{Success} \mid \text{Small}) = 0.890,$$

and so large stones are negatively associated with success outcome. This means that there is a lower proportion of successful outcomes for large stones cases compared to small stones cases.

As we have now shown that stone size is associated with both the treatment type and the treatment outcome, we are convinced that stone size is a confounding variable that needs to be managed. The way to do it, as shown previously is to use *slicing*, where we segregate the data by the confounding variable. This is done by investigating the association between the dependent and independent variables for large stone cases separately from the small stone cases.

Discussion 2.5.4 We have now seen the benefits of having more information on the subjects because it allows us to identify confounding variables which would not have been possible if, for example, information of stone size was not available or collected. Thus, an

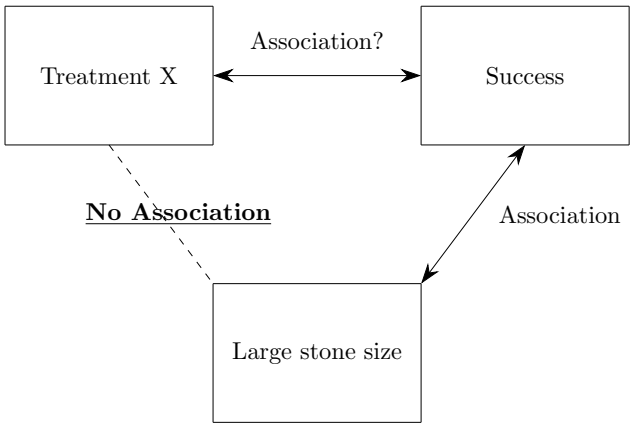
important learning point when it comes to designing a study is to measure and collect data on additional variables that we feel may be relevant in our study. Whether these additional variables turn out to be confounders or not we would have to probe further, but we will never know if we do not collect data on them in the first place.

That said, we have to come to terms with the fact that most of the time, collecting information on variables is costly in practice. Even if we do manage to collect all the information we need, the analysis can be complicated if the data needs to be sliced along many different variables.

For non-randomised designs like observational studies, it is usually the case that the two groups that we are comparing are not "identical" except for the treatment. Despite our best efforts, we can never be totally sure that every single confounder has been identified and controlled for. Thus, observational studies offer only a limited conclusion in providing evidence of *association* and not *causation*.

(Randomisation as a preferred solution to confounding.) An alternative approach to address potential confounders is to rely on a strategy that was discussed in Chapter 1: **randomised assignment**. Let us discuss how this in done in detail, using our much developed example on kidney stones treatment.

Example 2.5.5 Fundamentally, confounding variables occur due to association which is a consequence of having unequal proportion of variables in the two groups that we are trying to compare. For the kidney stones example, stone size was a confounder because patients with large stones were disproportionately allocated to treatment X instead of treatment Y. Now, if the allocation of large (and small) stone size cases to the two treatment types was done randomly, which tends to result in an equal proportion across the two groups, there would no longer be any association between stone size and treatment type. In this case, stone size would no longer be a confounder. Note that a confounding variable is associated to **both** the independent and dependent variables, so removing one of the associations is enough to remove the confounding variable.



How can we achieve randomised assignment of patients to the two treatment types? One simple way is, for example, to toss a fair coin when deciding which treatment a patient will be given. Surely, such a method of randomised assignment tends to give us equal proportions of large (and small) stone cases across the two treatment types. If we have sufficiently many patients to assign to either treatment types, the two groups of patients assigned to treatment X and treatment Y will tend to be similar in all characteristics, including stone size.

Surely this addresses the problem of confounders appropriately right? Unfortunately, randomisation is not always possible in every study. Imagine the scenario where the type of treatment given to each patient is dependent on a coin toss! Would you agree to this if you were one of the patients? Certainly not! Patients usually have the right to choose which treatment group they want to be in and this would make the assignment process non-random. Such ethical issues could very well constrain and prevent us from performing randomised assignment of our subjects. In such a situation, we have no choice but to fall back on the method of slicing for suspected confounders.

With this, we conclude Chapter 2, where we discussed, in detail how we can use rates to study the association between two (or more) **categorical variables**. We learnt about Simpson's Paradox which led us eventually to the issue of confounders and how they can be managed. In the next Chapter, we will turn our attention to the other variable type, namely **numerical variables**.