CHAPTER 2

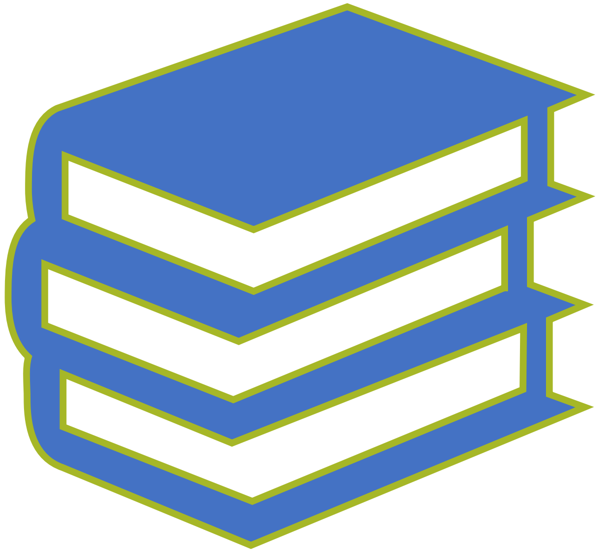
Categorical Data Analysis

Overview

1A. Rates I → 1B. Rates II → 2A. Association I

2B. Association II → 3A. Simpson's Paradox I → 3B: Simpson's Paradox II

4. Confounders



Unit 1A: Rates I

By the end of this unit you should be able to do the following:

1. Identify a categorical variable.
2. Understand and interpret tables and plots created from 1 categorical variable.

RECAP

Types of variables

Categorical variables

- **Ordinal**: Categories come with some natural ordering and numbers are often used to represent the ordering. E.g.: Happiness level
- **Nominal**: No intrinsic ordering for the variables. E.g.: Nationality

Numerical variables

- **Continuous**: One that can take on all possible numerical values in a given range or interval. E.g.: Time
- **Discrete**: One where possible values of the variable form a set of numbers with “gaps”. E.g.: Module credits

THE PROBLEM

A SNAPSHOT OF THE DATA

Size	Gender	Treatment	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

APPLYING THE PPDAC CYCLE

In general, do treatments tend to be successful?

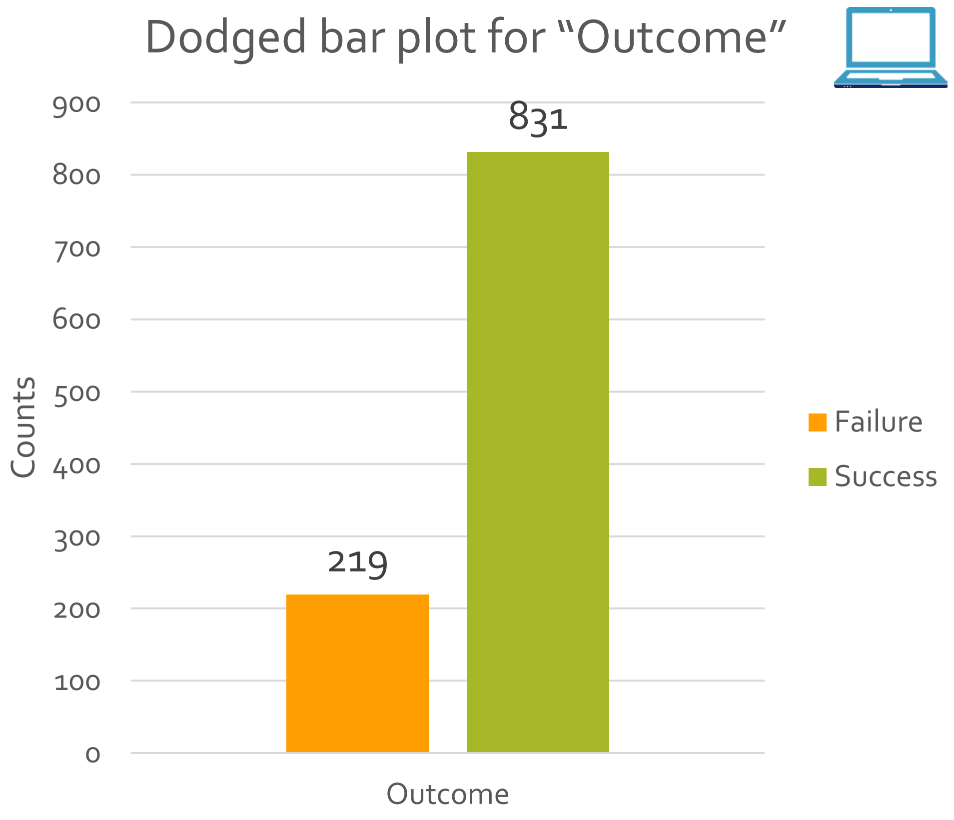
What to measure:
Variable "Outcome" tells us if the treatment was a success or not

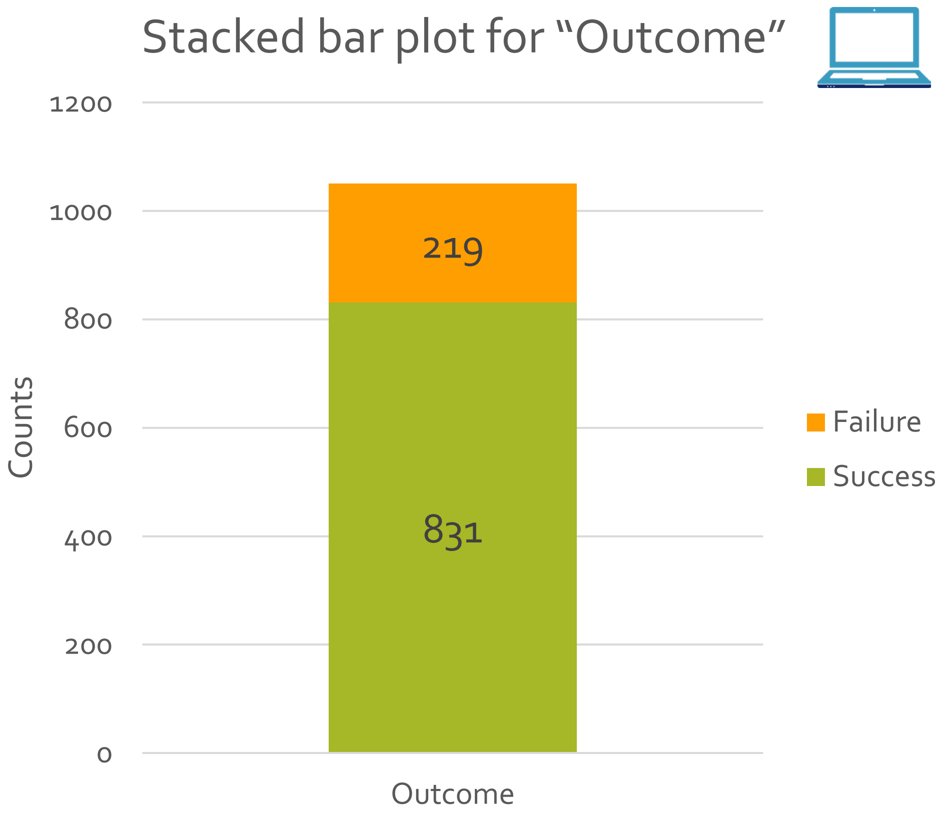
Sort the data
Plot graphs, tables of the "Outcome" variable

ANALYSING 1 CATEGORICAL VARIABLE - TABLE

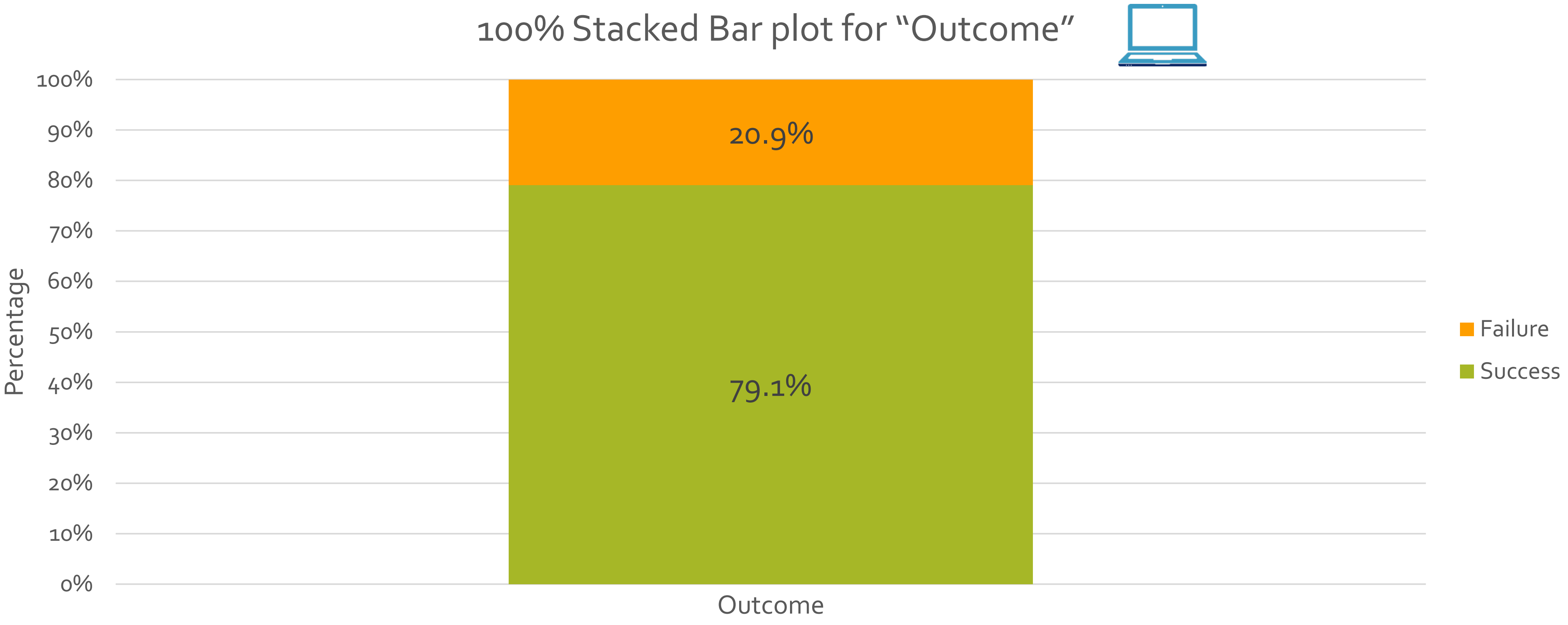
Categories of the "Outcome" variable	Count	Rate	Percentage
Success	831	$\text{rate(Success)} = \frac{831}{1050} = 0.791$	$0.791 \times 100\% = 79.1\%$
Failure	219	$\text{rate(Failure)} = \frac{219}{1050} = 0.209$	$0.209 \times 100\% = 20.9\%$
Total	1050	$\frac{1050}{1050} = 1$	$1 \times 100\% = 100\%$

Analyzing 1 categorical variable - Plot





Analysing 1 categorical variable - Plot



Conclusion

- Table and bar plots gave us the same conclusion
- 79% success
- 21% failure
- Should go for treatment

Summary

We have learned:

- Use of tables and plots to summarize a categorical variable
- Calculation of rates

Unit 1B: Rates II

By the end of this unit you should be able to do the following:

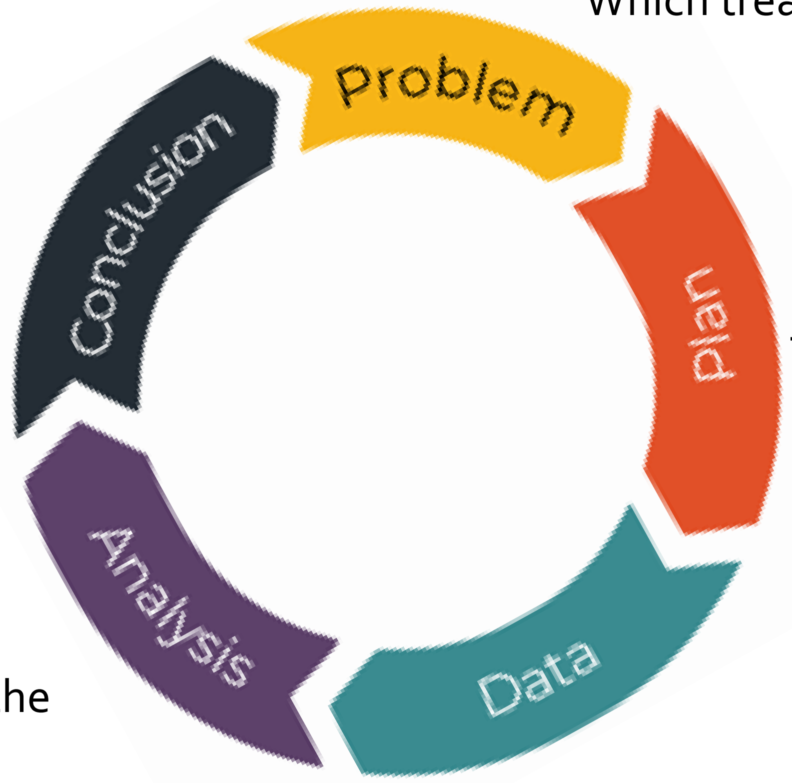
1. Understand and interpret tables and plots created from 2 categorical variables.
2. Calculate marginal, conditional and joint rates.

Size	Gender	Treatment	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

WHICH TREATMENT TO CHOOSE?

PPDAC CYCLE – A NEW QUESTION

Which treatment is better?



Key variable of interest:
Treatment variable

Sort the data
Plot graphs, tables of the "Treatment" and "Outcome" variable

2 x 2 Table

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

Marginal rates / proportions / percentages

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

- What proportion of the total number of patients underwent Treatment Y?
- $\mathrm{rate(Y)} = \frac{350}{1050} = \frac{1}{3} = 33\frac{1}{3}\%$
- What proportion of the total number of patients had a successful treatment?
- $\mathrm{rate(Success)} = \frac{831}{1050} = 0.791 = 79.1\%$
- Calculations above are called marginal rates / proportions / percentages.

Conditional rates / proportions / percentages

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

- If we focus on patients who undergone Treatment X, what proportion of them had a successful treatment?
- $\text{rate}(\text{Success given X}) = \frac{542}{700} = 0.774 = 77.4\%$
- Calculation above is known as a conditional proportion / percentage.
- An even shorter way of writing this is to use a vertical bar in place of given: $\text{rate}(\text{Success} \mid \text{X})$

Joint rates / proportions / percentages

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

- What is the proportion of patients who chose Treatment Y and had a failure?
- $\text{rate}(\text{Y and failure}) = \frac{61}{1050} = 0.0581 = 5.81\%$
- NOT a conditional rate.
- Calculation is known as a joint rate/ proportion / percentage.

Which treatment is better?

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

Treatment X has 542 successful cases.

Treatment Y has 289 successful cases.

“We should recommend Treatment X!”?

More patients choosing Treatment X as compared to Y.

Making it fair!

Compare success rate of Treatments X and Y

Given that I pick some treatment, what is the rate of success?

- $\text{rate}(\text{Success} \mid \text{X}) = \frac{542}{700} = 0.774 = 77.4\%$
- $\text{rate}(\text{Success} \mid \text{Y}) = \frac{289}{350} = 0.826 = 82.6\%$

Fair comparison

- For Treatment X, roughly 77 out of 100 patients had a successful treatment.
- For Treatment Y, roughly 83 out of 100 patients had a successful treatment.

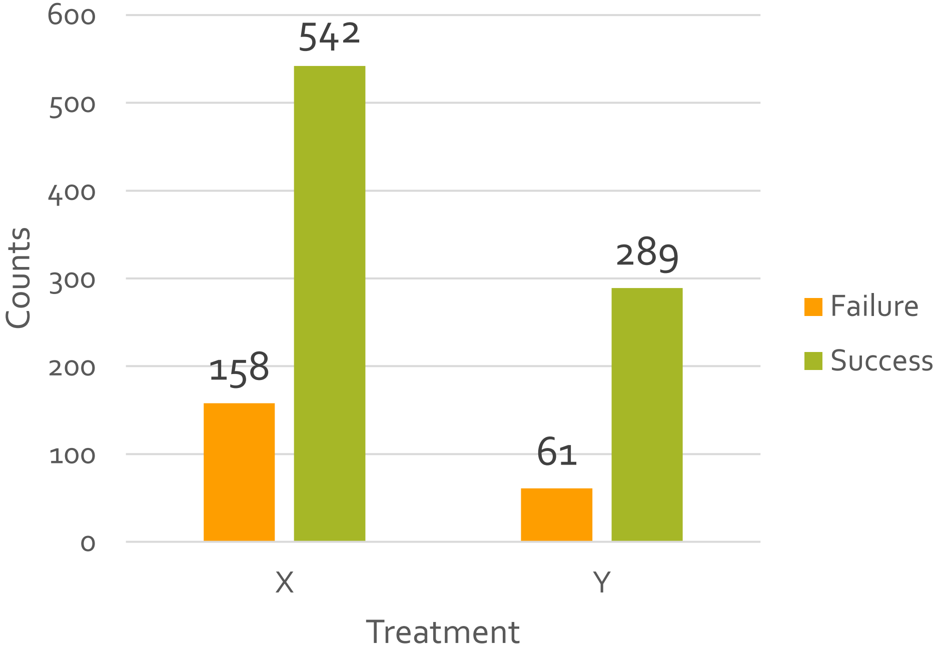
Treatment Y is better!

Table with row percentages

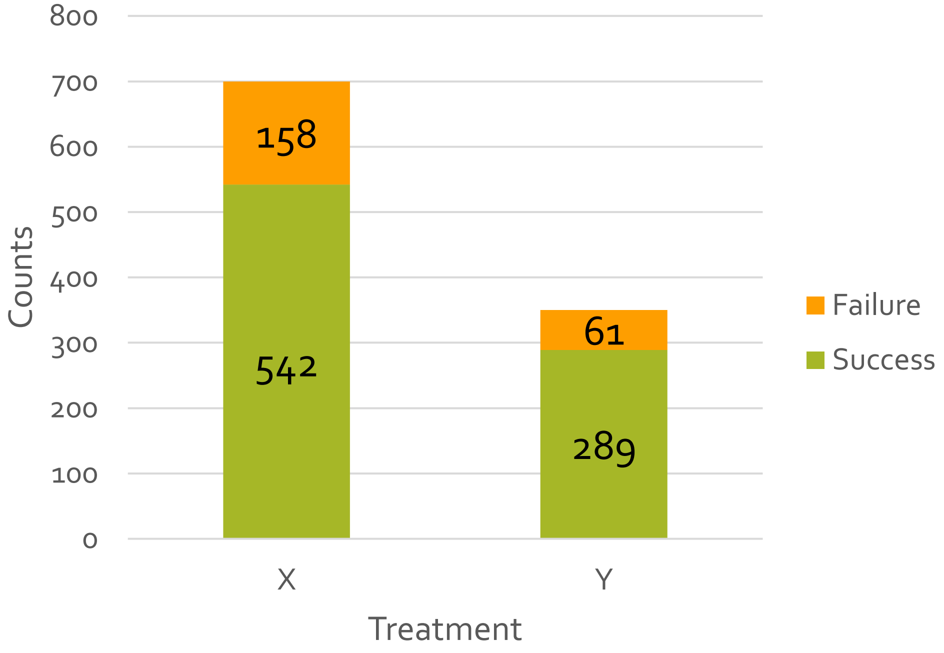
Outcome / Treatment	Success (row %)	Failure (row %)	Row Total (row %)
X	542 (77.4%)	158 (22.6%)	700 (100%)
Y	289 (82.6%)	61 (17.4%)	350 (100%)
Column Total	831 (79.1%)	219 (20.9%)	1050 (100%)

Analysing 2 categorical variables - plot

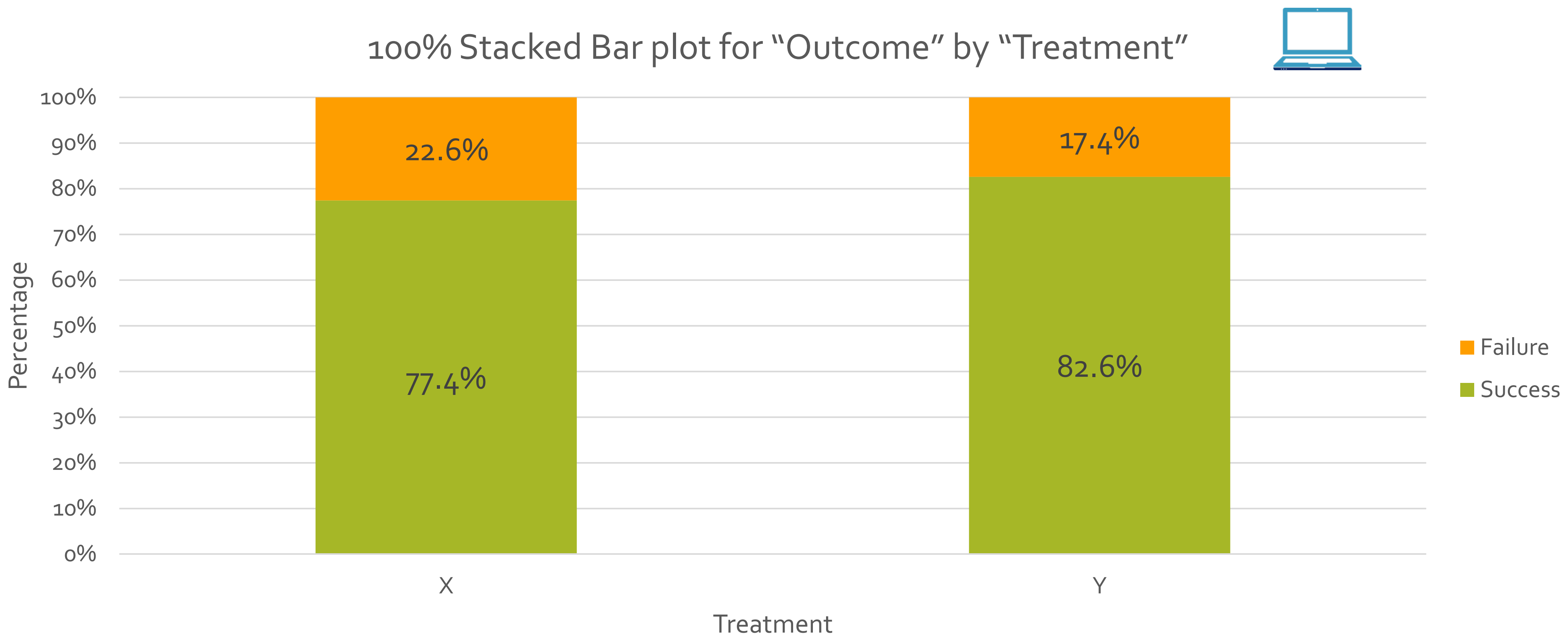
Dodged bar plot for “Outcome” by “Treatment”



Stacked bar plot for “Outcome” by “Treatment”



Analysing 2 categorical variables - plot



Summary

We have learnt how to analyse 2 categorical variables from the perspective of:

- Tables – 2x2 table
- Plots – Bar plots / 100% stacked bar plots

Unit 2A: Association I

By the end of this unit, you should be able to do the following:

1. Understand and apply association
2. Understand and apply symmetry rule

Used rates to conclude that Treatment Y is better than Treatment X.

Relationship between the type of treatment and the outcome of the treatment

"Treatment Y is positively associated to the success of the treatment."

"Treatment X is negatively associated to the success of the treatment."

Tend to see treatment Y and successful treatments go hand in hand.

Tend to see Treatment X and unsuccessful treatments go hand in hand.

Caution: Association, not causation!

Associative relationship between the 2 variables

Not sure if success of treatment is due to the treatment or not

Continuation from Unit 1

Is there an association?

Suppose we have A and B as characteristics in a population. We shall assume that some people have A, and some do not have A (labelled as NA). We assume the same about B.

Association absent	Association present	
$\text{rate}(A \mid B) = \text{rate}(A \mid NB)$	$\text{rate}(A \mid B) \neq \text{rate}(A \mid NB)$	
Rate of A is not affected by the presence or absence of B.	rate(A | B) > rate(A | NB)	rate(A | B) < rate(A | NB)
A and B are not associated.	Presence of A when B is present is stronger than when B is absent.	Presence of A when B is present is weaker than when B is absent.
	Positive association between A and B.	Negative association between A and B.

Linking back to our dataset

Checking for association between 2 variables

- Outcome of treatment
  - A: Success
  - NA: Failure
- Treatment
  - B: Treatment X
  - NB: Treatment Y

Compare

- $\text{rate}(A \mid B) = \text{rate}(\text{Success} \mid X) = 0.774$
- $\text{rate}(A \mid NB) = \text{rate}(\text{Success} \mid Y) = 0.826$

Conclusion

- rate(A | B) < rate(A | NB)
- Presence of A is weaker when B is present.
- Less successful treatments when we see Treatment X: Treatment X is negatively associated to a successful treatment.
- More successful treatments when we see Treatment Y: Treatment Y is positively associated to a successful treatment.

On Establishing Association

- Any of the following comparisons can show positive association between A and B:

rate(A | B) > rate (A | NB)
rate(B | A) > rate (B | NA)
rate(NA | NB) > rate (NA | B)
rate(NB | NA) > rate (NB | A)

- Likewise, for negative association between A and B:

rate(A | B) < rate (A | NB)
rate(B | A) < rate (B | NA)
rate(NA | NB) < rate (NA | B)
rate(NB | NA) < rate (NB | A)

- Try it – using the example in the previous slide, you can see that this relation holds true;

Eg. "If success is positively associated with treatment Y, then ..."

- "... success is negatively associated with ???"
- "... failure is positively associated with ???"
- "... failure is negatively associated with ???"

2 rules that govern rates

Suppose we have A and B as characteristics in a population. We shall assume that some people have A, and some do not have A (labelled as NA). We assume the same about B.

Symmetry rule

Basic rule on rates (to be discussed in Unit 2B: Association II)

Symmetry Rule

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

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

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

rate(A | B) > rate (A | NB) ⇔ rate(B | A) > rate(B | NA)

	B	Not B	Row Total
A	w	x	w + x
Not A	y	z	y + z
Column Total	w + y	x + z	w + x + y + z

$$\frac{w}{w+y} > \frac{x}{x+z}$$
$$w(x+z) > x(w+y)$$
$$wx + wz > xw + xy$$

$$\frac{w}{w+x} > \frac{y}{y+z}$$
$$w(y+z) > y(w+x)$$
$$wy + wz > yw + yx$$

$$wz > xy$$

$$\mathrm{rate}(A \mid B) \; > \; \mathrm{rate}(A \mid NB) \Leftrightarrow \mathrm{rate}(B \mid A) \; > \; \mathrm{rate}(B \mid NA)$$

1

2

$$\mathrm{rate}(A \mid B) \; > \; \mathrm{rate}(A \mid NB) \rightarrow \mathrm{rate}(B \mid A) \; > \; \mathrm{rate}(B \mid NA)$$

$$\mathrm{rate}(B \mid A) \; > \; \mathrm{rate}(B \mid NA) \rightarrow \mathrm{rate}(A \mid B) \; > \; \mathrm{rate}(A \mid NB)$$

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

Rate of A given B is more than rate of A given NB.

Positive association between A and B.

More likely to see A when B is present as compared to when B is absent.

Also more likely to see B when A is present as compared to when A is absent.

Rate of B given A is more than rate of B given NA.

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

Rate of B given A is more than rate of B given NA. → Positive association between B and A. → More likely to see B when A is present as compared to when A is absent. → Also more likely to see A when B is present as compared to when B is absent. → Rate of A given B is more than rate of A given NB.

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

1

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

2

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

Consequence of the symmetry rule

To identify if there is any association, check for either:

1. rate(A | B) ≠ rate(A | NB) OR

2. rate(B | A) ≠ rate(B | NA)

rate(Success | X) < rate(Success | Y):
Negative association between successful treatments and Treatment X

Check:
rate(X | Success) < rate(X | Failure)

Summary

We have learned:

- How to identify association
- Symmetry rule and its consequence on identifying association

Unit 2B: Association II

By the end of this unit, you should be able to do the following:

1. Understand and apply basic rule on rates

BASIC RULE ON RATES

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

Consequences of the basic rule on rates

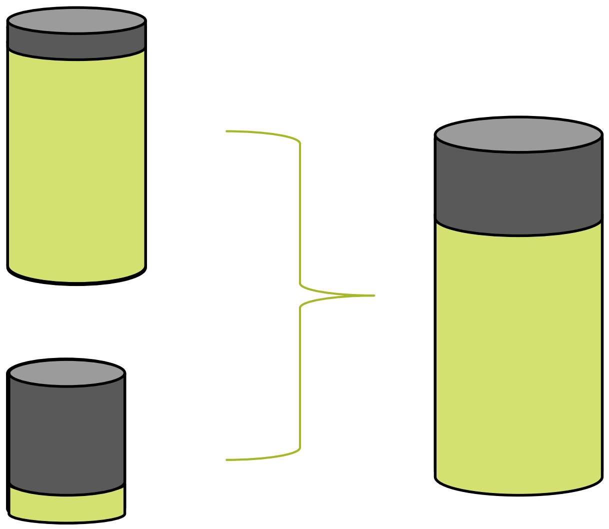
1. The closer rate(B) is to 100%, the closer rate(A) is to rate(A | B).
2. If $\mathrm{rate(B)} = 50\%$, then $\mathrm{rate(A)} = \frac{\mathrm{rate(A \mid B)+rate(A \mid NB)}}{2}$.
3. If $\mathrm{rate(A \mid B)} = \mathrm{rate(A \mid NB)}$, then $\mathrm{rate(A)} = \mathrm{rate(A \mid B)} = \mathrm{rate(A \mid NB)}$.

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

- 2 Cups of bubble tea
- Let A be the level of sweetness
  - Represented by the colour “Green” in the cup.
- Let B / NB be the cups: “Cup 1” vs. “Cup 2”

Cup 1
Size: Large cup
Sweetness: 90%

Cup 2
Size: Small cup
Sweetness: 20%



Size: Cup 1 + Cup 2
Sweetness: In between 20% to 90%, but closer to Cup 1

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

Sweetness in the final cup is between Sweetness | Cup 1 and Sweetness | Cup 2

Cup 1 takes up most of the final cup.

Expect sweetness of the final cup to be nearer to the sweetness of Cup 1.

Overall rate(A) to be between rate(A | B) and rate(A | NB)

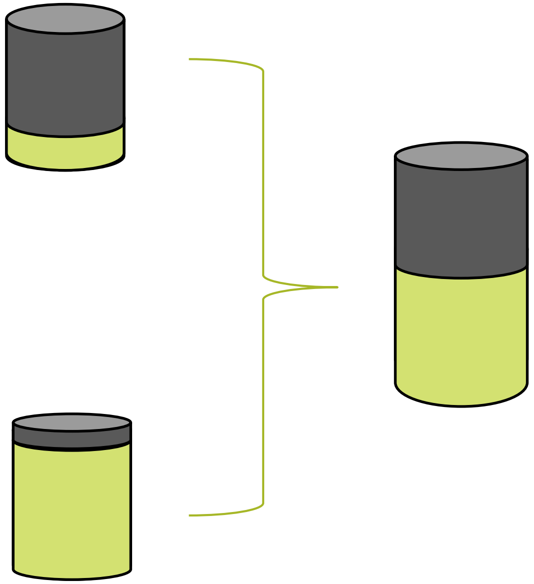
Overall rate(A) to be closer to rate(A | B) if B takes up a majority of the overall.

2. If $\text{rate}(B) = 50\%$, then

$$\text{rate}(A) = \frac{\text{rate}(A \mid B) + \text{rate}(A \mid NB)}{2}$$

Cup 1
Size: Small cup
Sweetness: 20%

Cup 2
Size: Small cup
Sweetness: 90%

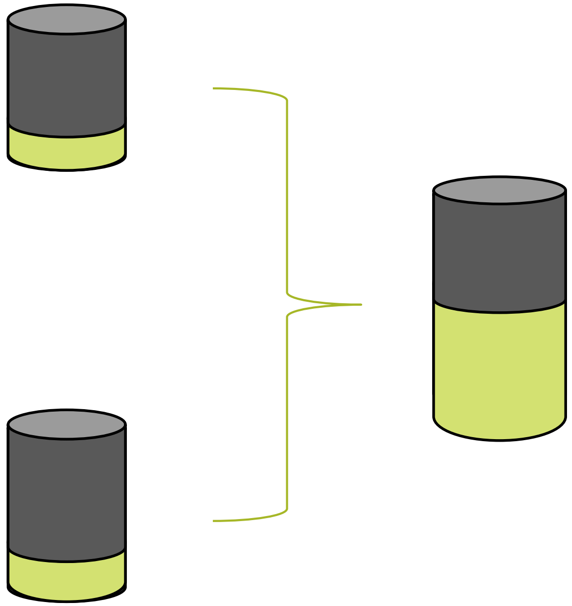


Size: Cup 1 + Cup 2
Sweetness: Exactly in between
20% to 90% = $\frac{20\%+90\%}{2} = 55\%$

3. If $\mathrm{rate}(A \mid B) = \mathrm{rate}(A \mid NB)$, then $\mathrm{rate}(A) = \mathrm{rate}(A \mid B) = \mathrm{rate}(A \mid NB)$.

Cup 1
Size: Small / Large cup
Sweetness: 20%

Cup 2
Size: Small / Large cup
Sweetness: 20%



Size: 2 Cups added together
Sweetness: Exactly 20%

Linking back to Consequences 2 and 3

If cups are of the same size, sweetness will be exactly half of the original cups.

- If Rate(B) = 50%, overall rate of A will be exactly in between the rate of A given B and the rate of A given NB.

If sweetness is the same for both cups, the sweetness of the final cup will also be the same, regardless of the sizes of the original cups.

- If rate(A | B) = rate(A | NB), then rate(A) is the same as the 2 rates.

Linking back to dataset at hand

Overall rate of successful treatments

- rate(Success) = 0.79

Groups: Treatment X and Treatment Y

- rate(Success | X) = 0.774
- rate(Success | Y) = 0.826
- rate(Success) in between the conditional rates

Overall rate of success closer to rate(Success | X)

- Treatment X takes up a majority of the treatments.
- rate(X) = $\frac{700}{1050} = 0.667 = 66\frac{2}{3}\%$
- Follows statement (1)

Summary

We have learned:

- What is the basic rule on rates
- The consequences of basic rule on rates

Unit 3A: Simpson's Paradox I

By the end of this unit you should be able to do the following:

1. Identify Simpson's paradox
2. Analysis using the slicing method

PPDAC CYCLE – A RECAP

Are the treatments are helping?

Yes. In general, there is a high rate of success.

More specifically, which treatment is better?

Treatment Y is positively associated to success rate

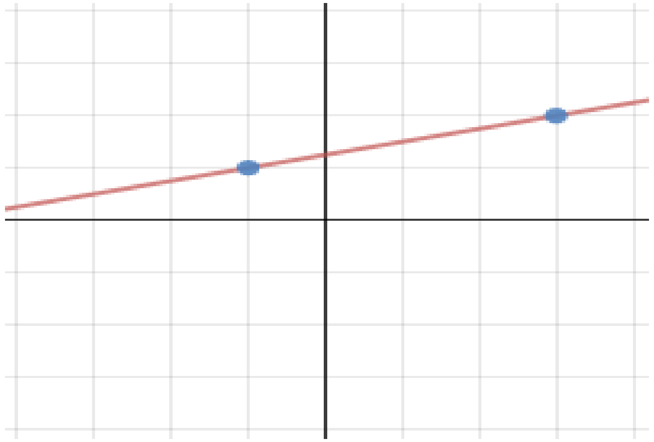
ANYTHING ELSE?

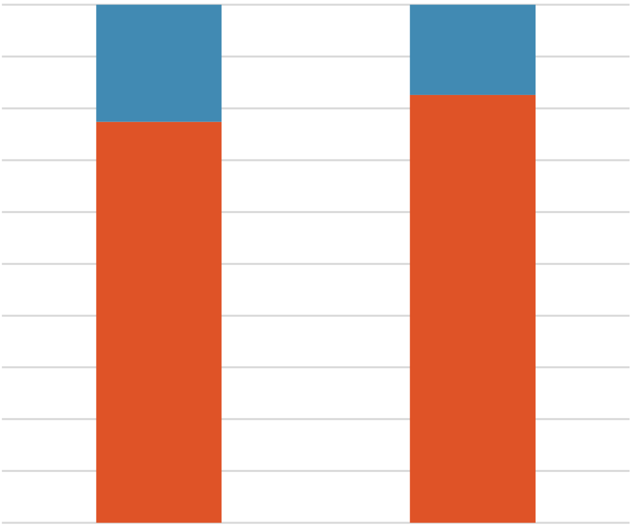
Size	Gender	Treatment	Outcome
Large	Male	X	Success
Large	Male	X	Success
Small	Male	Y	Success
Large	Male	Y	Failure
Small	Male	X	Success
Large	Male	Y	Success

WHAT ABOUT OTHER VARIABLES?

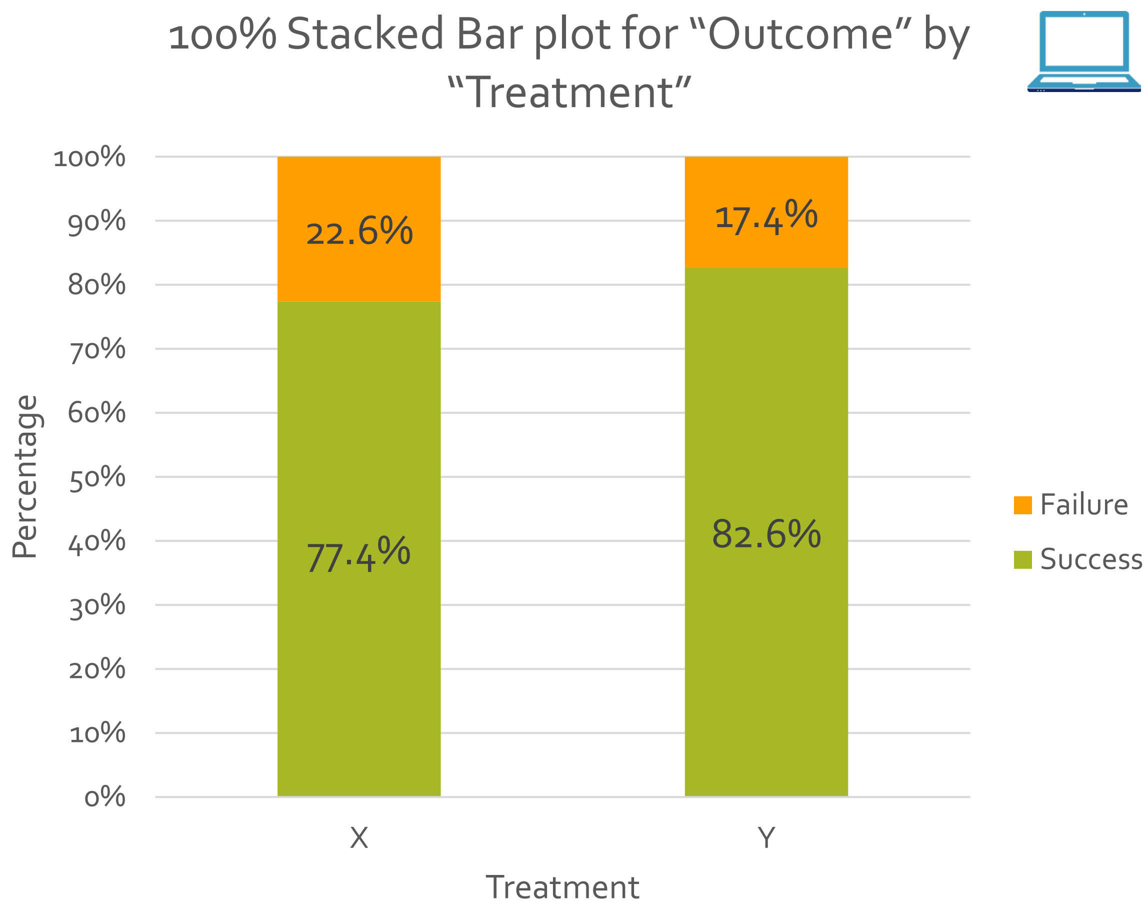
Exploring the "stone size" variable

What would be a useful visualisation?





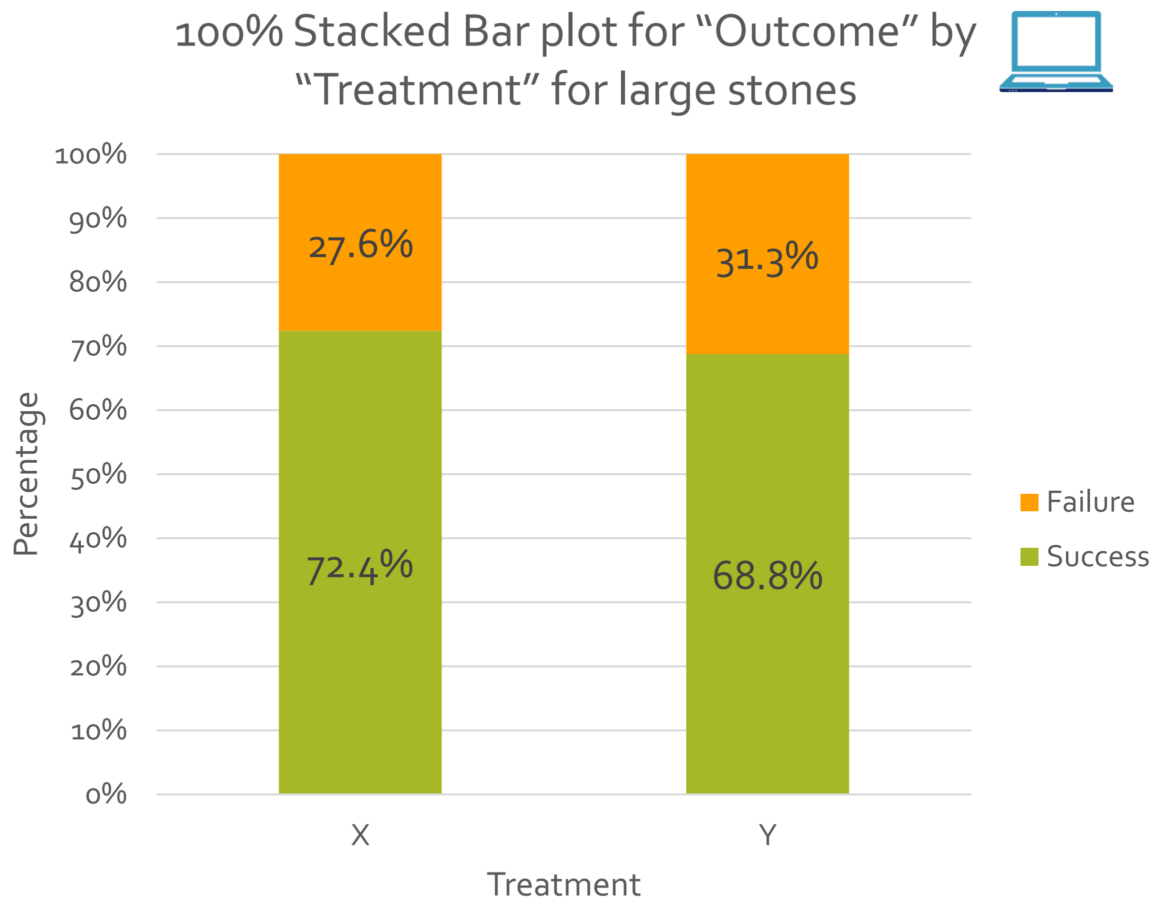
Analysing 2 categorical variables - Plot



Y Overall, Treatment Y is better

All stones	Success	Failure	Total
X	542	158	700
Y	289	61	350
Total	831	219	1050

Plot across large stones only



rate(Success | X) > rate(Success | Y)

X Across large stones, Treatment X is better

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

Exercise

X

Across large stones,
Treatment X is better

$$\text{rate}(\text{Success} \mid \text{X}) = \frac{381}{526} = 0.724$$

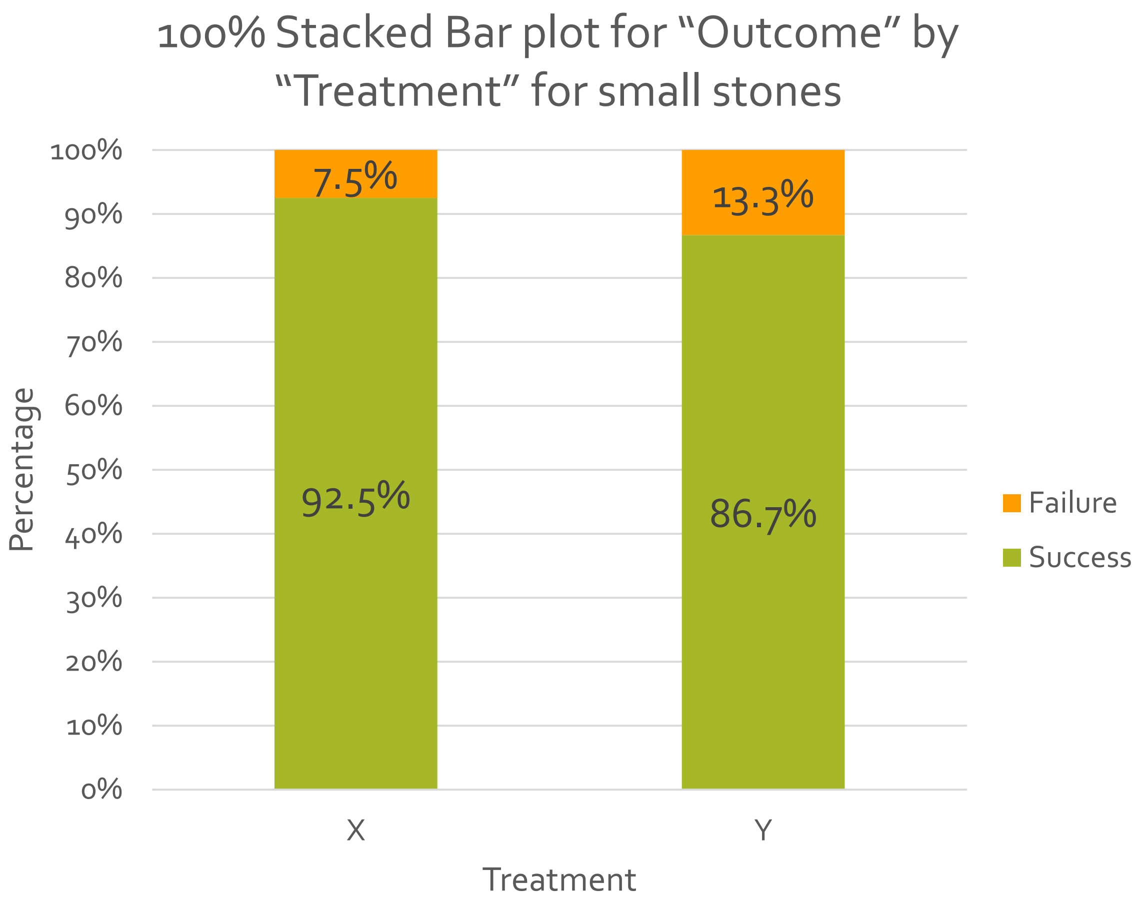
$$\text{rate}(\text{Success} \mid \text{Y}) = \frac{55}{80} = 0.688$$

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

rate(Success | X) > rate(Success | Y)

Treatment X is positively associated to success

Plot across small stones only



X

Across small stones, Treatment X is better

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

Analysing 3 categorical variables - plot

A paradox on our hands

Y — Overall,
Treatment Y is better

X — Across large stones,
Treatment X is better

X — Across small stones,
Treatment X is better

Unit 3B: Simpson’s Paradox II

By the end of this unit you should be able to do the following:

1. Explain a Simpson’s paradox

A paradox on our hands

Y

Overall,
Treatment Y is better

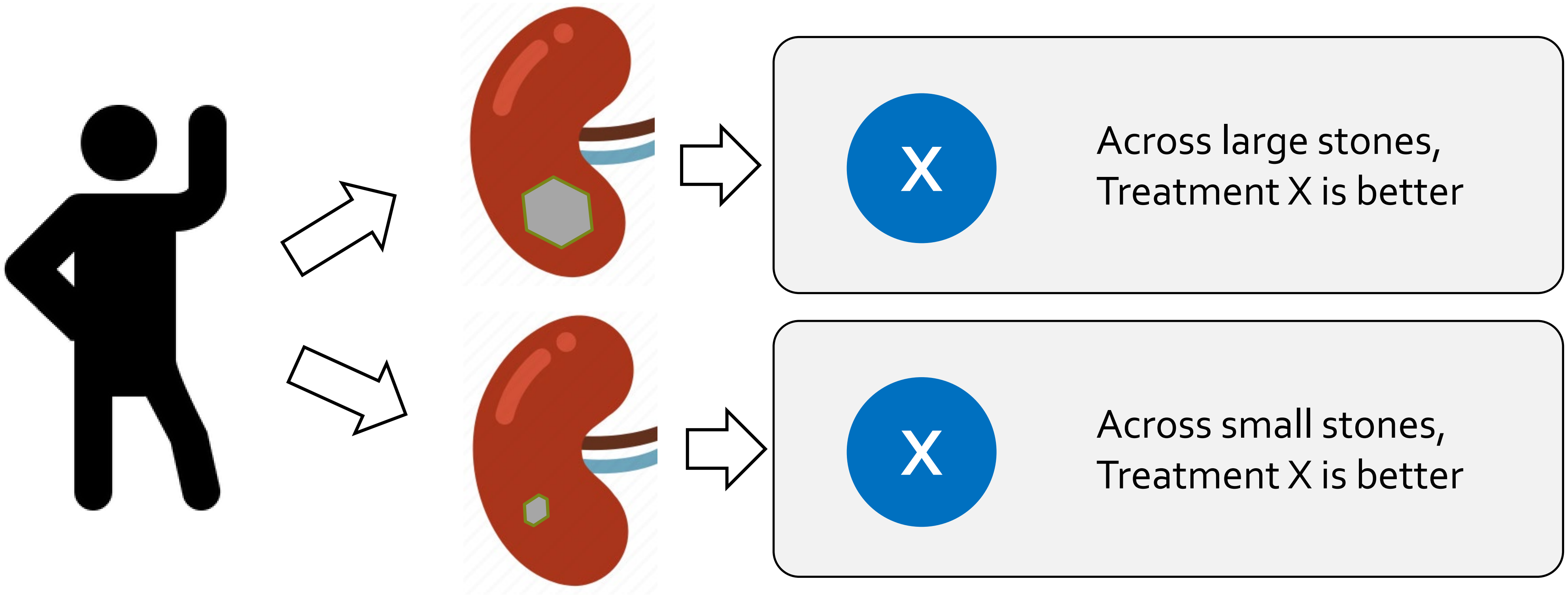
X

Across large stones,
Treatment X is better

X

Across small stones,
Treatment X is better

A paradox explained



Analysing 3 categorical variables - Table

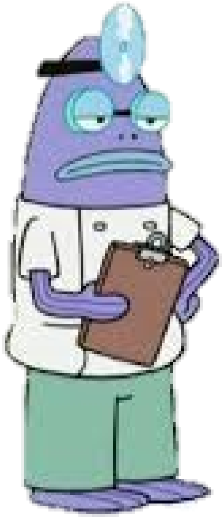
	Large stones			Small stones			Total (Large + Small)		
	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) in %
X	381	526	72.4%	161	174	92.5%	542	700	77.4%
Y	55	80	68.8%	234	270	86.7%	289	350	82.6%

Analysing 3 categorical variables - Table

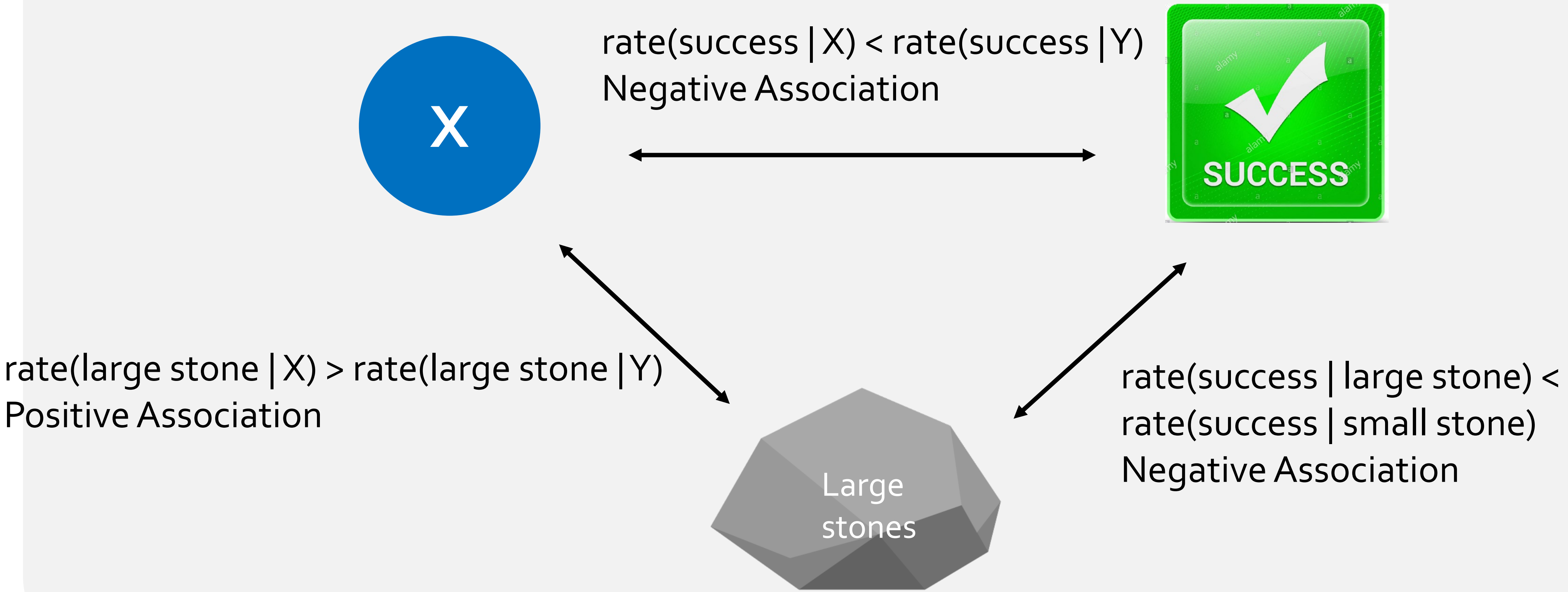
	Large stones			Small stones			Total (Large + Small)		
	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) in %
X	381	526	72.4%	161	174	92.5%	542	700	77.4%
Y	55	80	68.8%	234	270	86.7%	289	350	82.6%

Analysing 3 categorical variables - Table

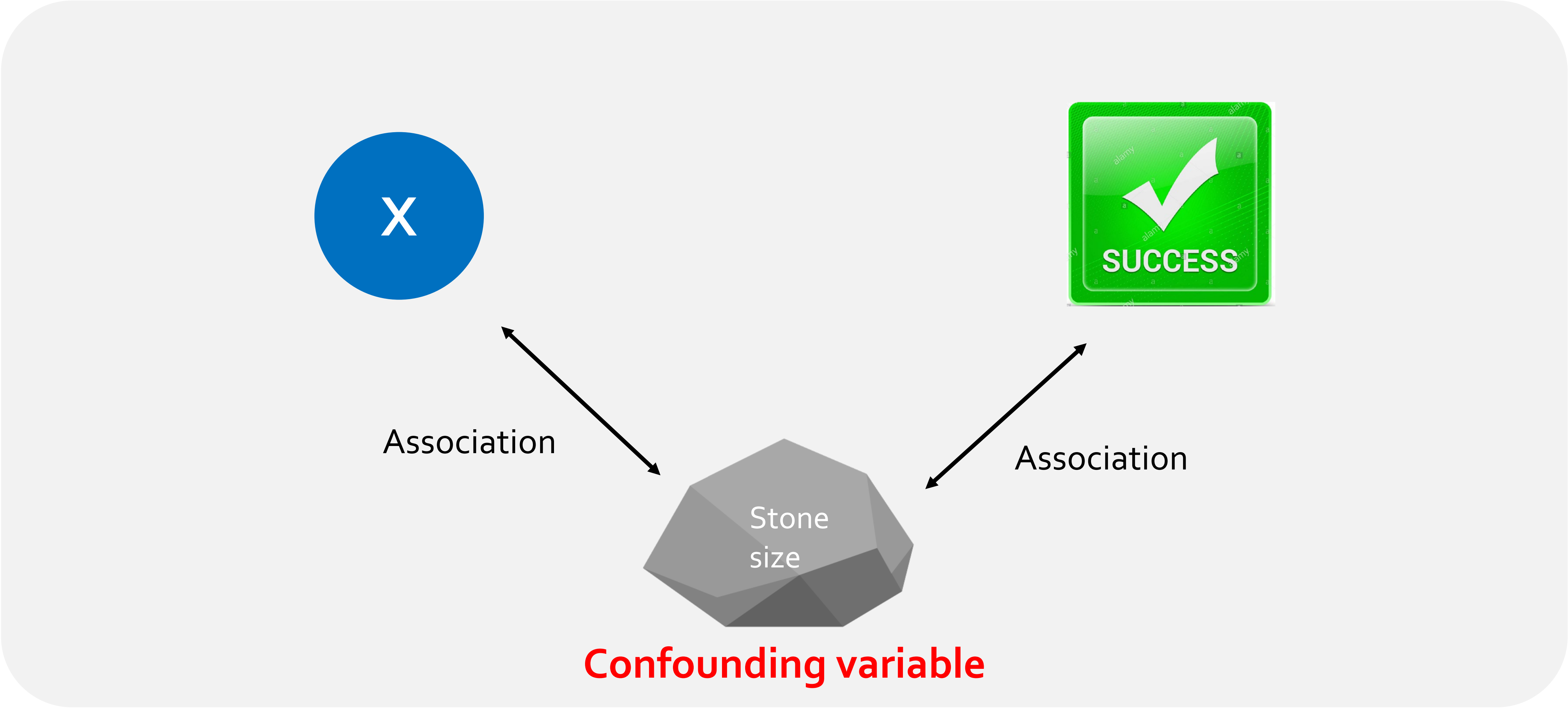
	Large stones			Small stones			Total (Large + Small)		
	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) in %	Successful treatments	Total number of treatments	rate(Success) in %
X	381	526	72.4%	161	174	92.5%	542	700	77.4%
Y	55	80	68.8%	234	270	86.7%	289	350	82.6%



ANALOGY



View from Association



Simpson's paradox $\Rightarrow$ confounder
Confounder $\not\Rightarrow$ Simpson's paradox

Summary

We learnt how to analyse 3 categorical variables from the perspective of:

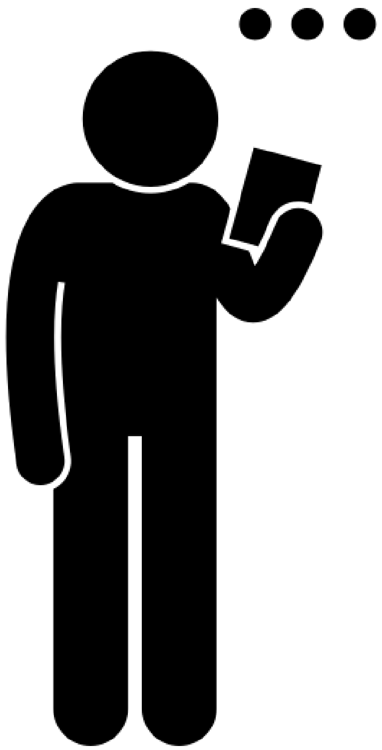
- Tables – slicing by subgroups
- Graphs – sliced bar graph

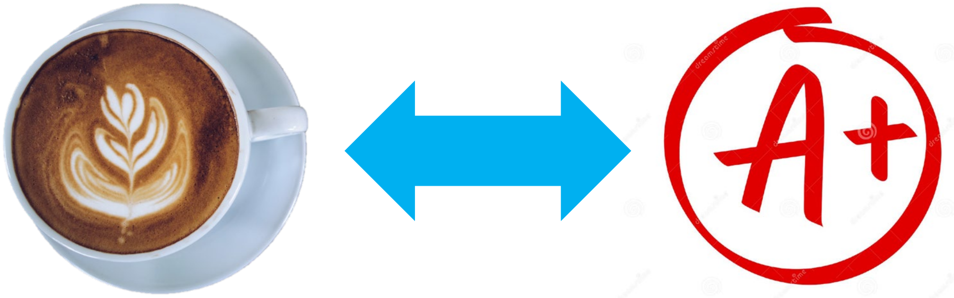
Unit 4
Confounders

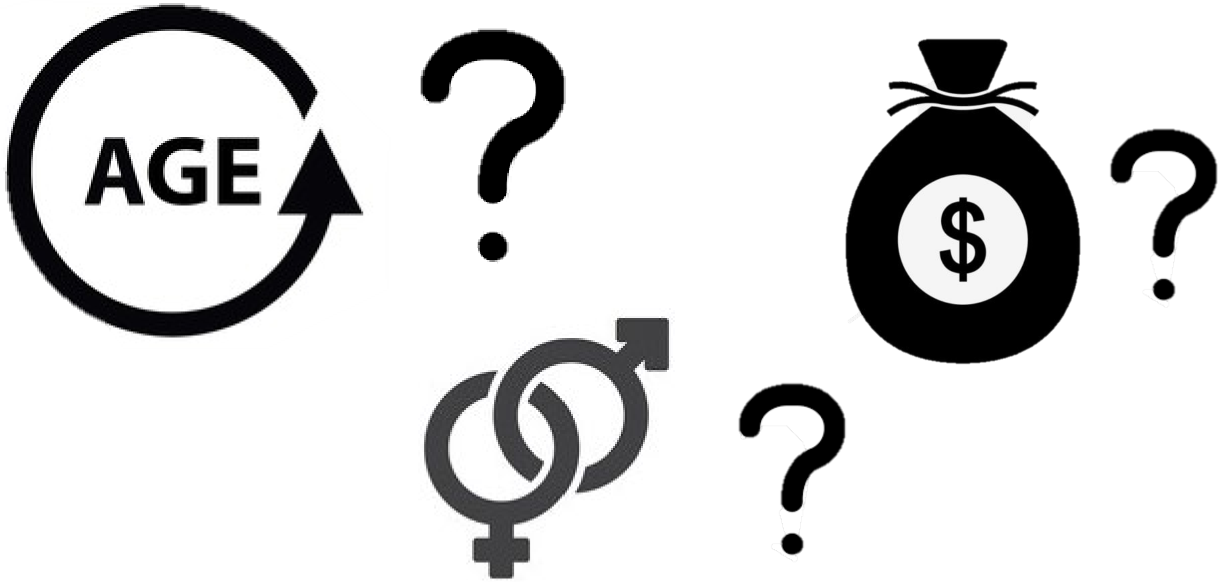
By the end of this unit you should be able to do the following:

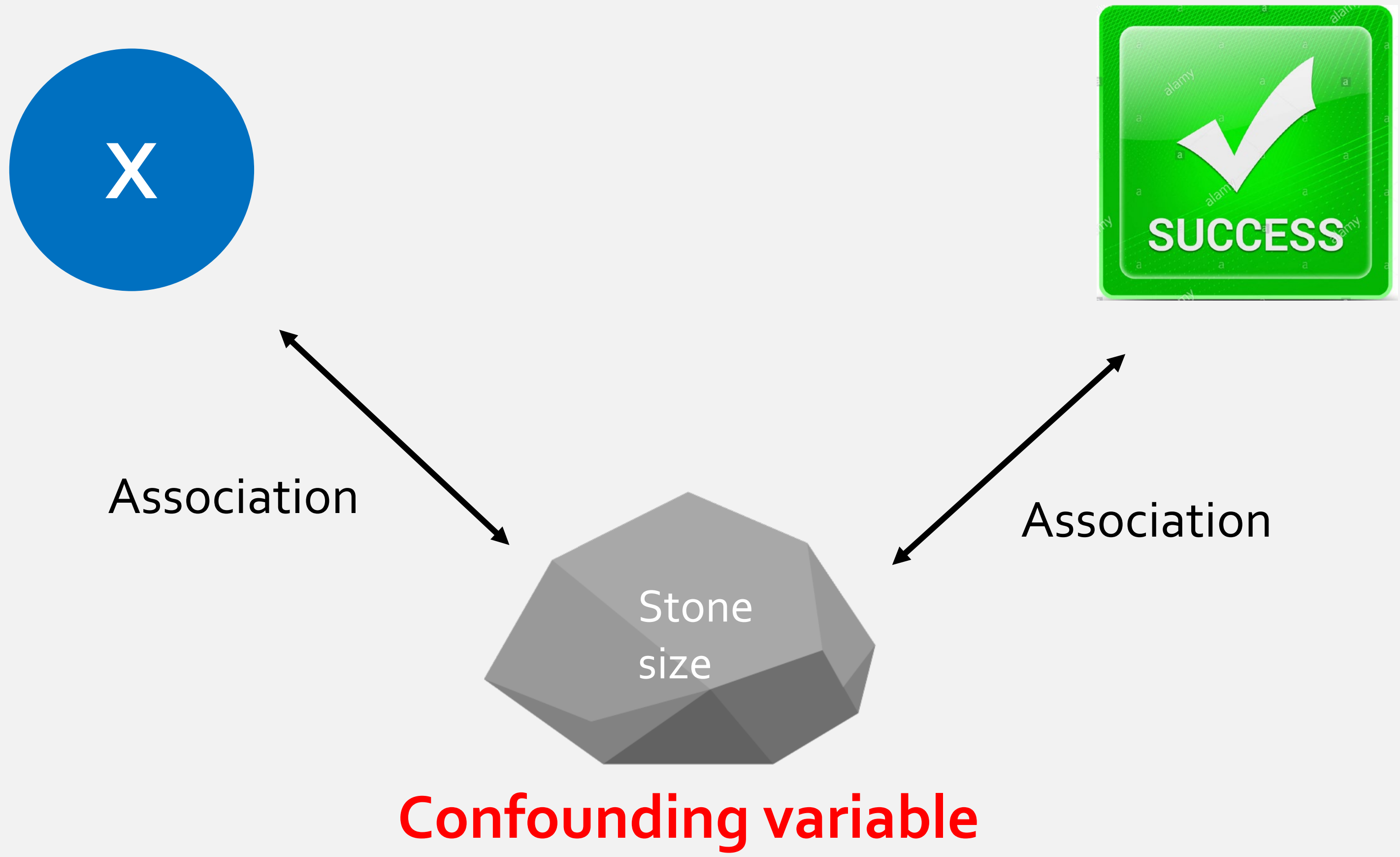
1. Define a confounder (ie. confounding variable)
2. Identify possible confounding variables in a study

Introduction









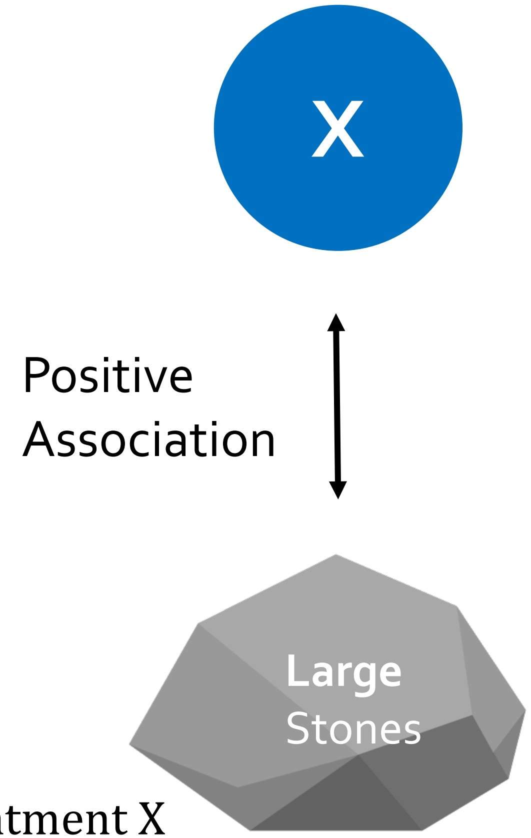
Definition:
A confounder is a third variable that is associated to both the independent and dependent variable whose relationship we are investigating

Stone size associated to treatment type

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

$$\left.\begin{array}{l} \text{rate}(\text{Large} \mid X) = \dfrac{526}{700} = 0.751 \\[2ex] \text{rate}(\text{Large} \mid Y) = \dfrac{80}{350} = 0.229 \end{array}\right\}$$

Since $0.751 > 0.229$,
Large stones **positively** associated to treatment X

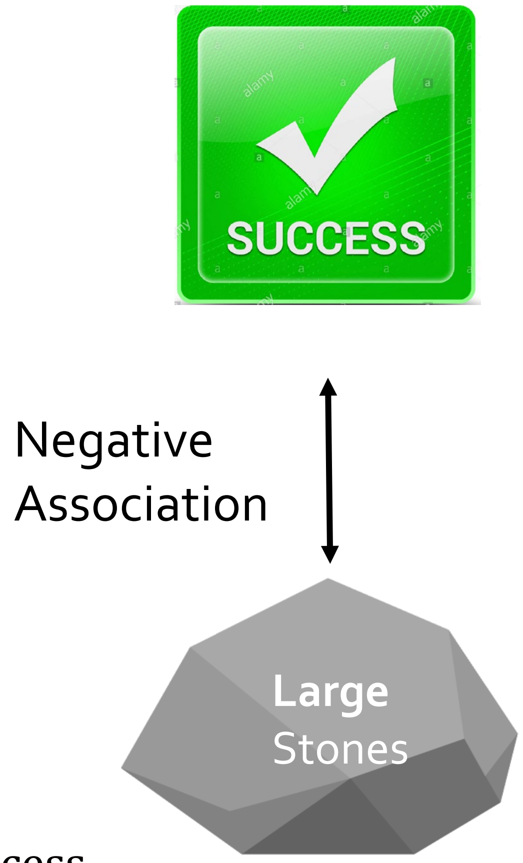


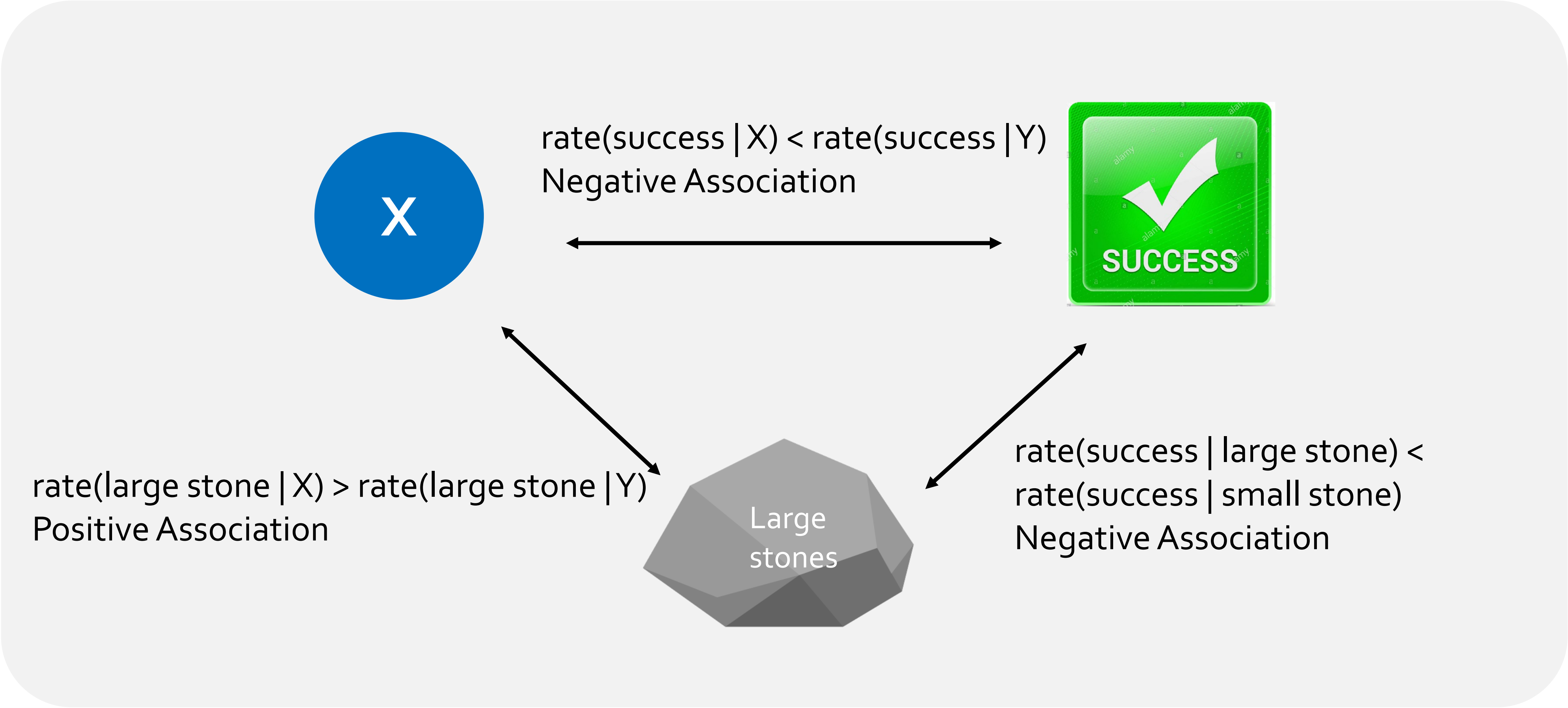
Stone size associated to success

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

$$\left.\begin{aligned}\text{rate}(\text{Success} \mid \text{Large}) &= \frac{436}{606} = 0.719 \\ \text{rate}(\text{Success} \mid \text{Small}) &= \frac{395}{444} = 0.890\end{aligned}\right\}$$

Since $0.719 < 0.890$,
Large stones **negatively** associated to success





View from Association

Recall:

Size	Treatment Type	Outcome
Large	X	Success
Large	X	Success
Small	Y	Success
Large	Y	Failure
Small	X	Success
Large	Y	Success

X

After slicing,
Treatment X is better

Size	Treatment Type	Outcome
Large	X	Success
Large	X	Success
Small	Y	Success
Large	Y	Failure
Small	X	Success
Large	Y	Success

DO WE STILL OBSERVE SIMPSONS PARADOX?

Y

No,
Treatment Y is better

We have to measure a variable in order to check if it is a confounder!

THE PROBLEM

We must measure a variable in order to check if it is a confounder

We need to collect data on lots of variables

This is not feasible
(costly, difficult to analyse)

???

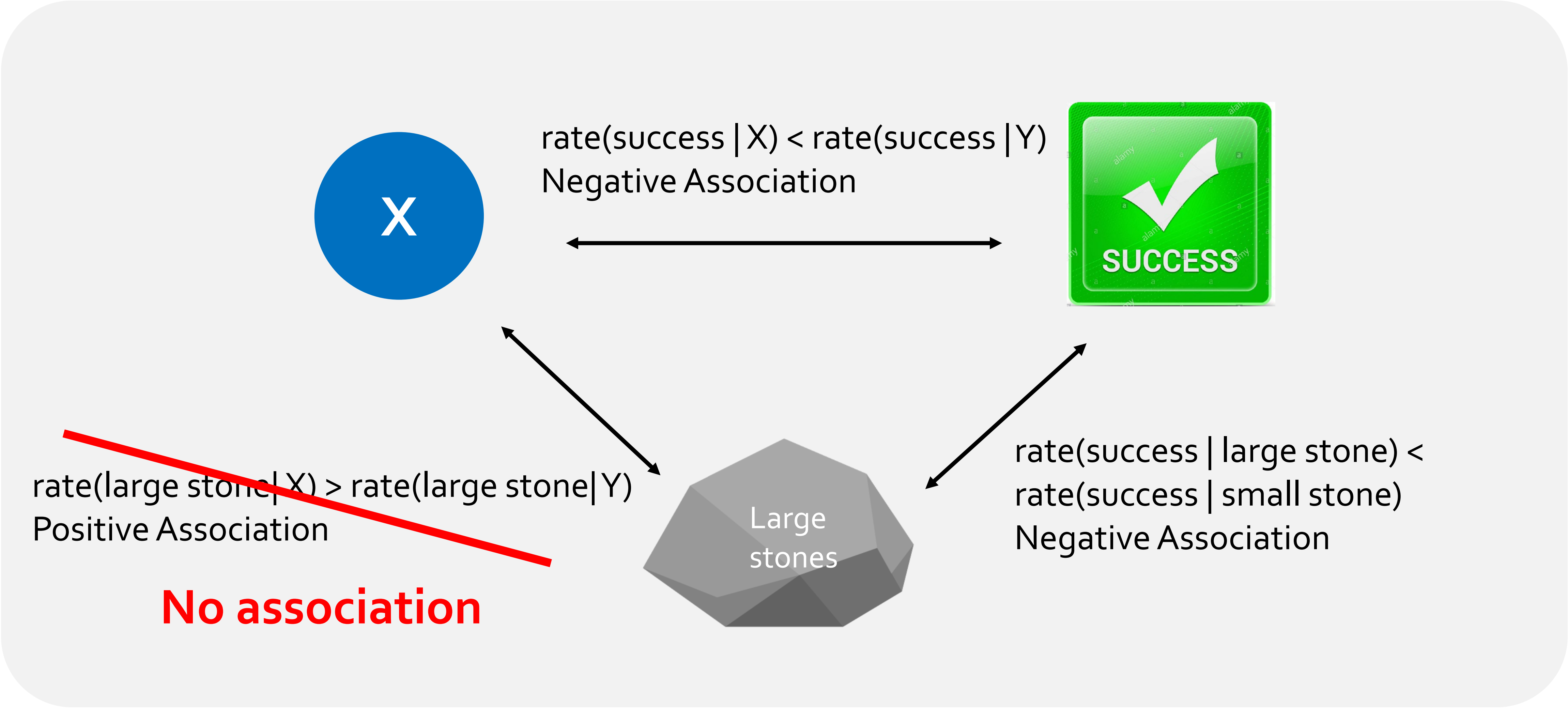
THE PROBLEM

We must measure a variable in order to check if it is a confounder

We need to collect data on lots of variables

This is not feasible
(costly, difficult to analyse)

RANDOMISATION

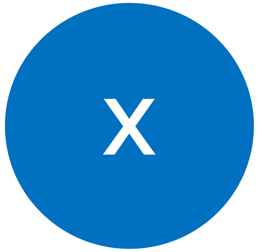


The effect of randomly assigning stone size to treatment type

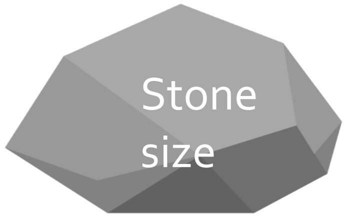
Randomisation is not always possible

Summary

Main variables



Confounding variable



(prove using association)

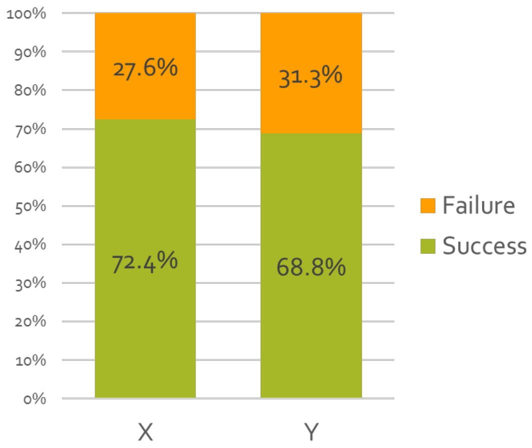
Proving Association

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

OR

$\text{rate}(B \mid A) \neq \text{rate}(B \mid NA)$

OR



Chapter 2 end

We learnt how to analyse categorical variables from the perspective of:

- Tables
- Graphs