GER1000 Quantitative Reasoning
Chapter: Design of Studies

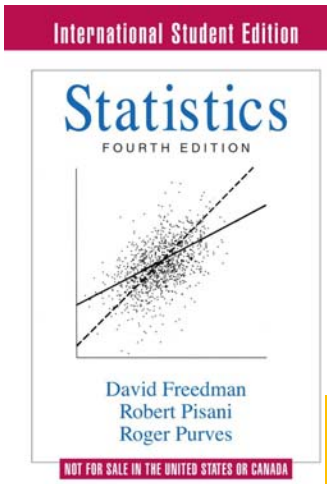
GER1000 Design of Studies
Unit 1: Introduction

[1] Welcome to Design of Studies. My name is Yap Von Bing, and I am a statistician. The mass media often carry headlines like “Sincere smiling promotes longevity”, or “Consistent, frequent TV viewing causes behaviour problems”. The aim of this chapter is to learn how to think critically about cause-and-effect relationships, also known as causal relationships, such as the two stories given here.

Cause and Effect

Karin Fischer

[2] Daily life is full of causal relationships. Turn the tap on, and water flows. Turn it off, the flow stops. Turn it on again, and the flow resumes. Indeed, we are so used to such things that if the experience is different from expected, we immediately suspect something is wrong. In our example, maybe the water supply or some parts of the piping system is faulty. If everything is in order, the turning causes the flow, and nobody will think that the flow causes the turning. Other causal relationships are harder to establish. Does watching lots of TV cause behaviour problems? Or do behaviour problems lead to watching lots of TV? Or is there some other common cause to both? Another example is in economics. Does increasing money supply cause inflation? Or does inflation increase money supply? Is something else going on?



"FPP": Chapter 1 Controlled Experiments
Chapter 2 Observational Studies

[3] You will learn two techniques for studying causal relationships: controlled experiments and observational studies. They will be presented with real examples, through which several important concepts will emerge. This will help you apply them in new situations. The lesson closely follows the first two chapters of the book Statistics by Freedman, Pisani and Purves, or in short, FPP. This book is excellent for every reader, regardless of the mathematical preparation. You are strongly encouraged to read these chapters, which are 28 pages in total. The picture shows the fourth edition, but the third edition is largely similar.

QR framework

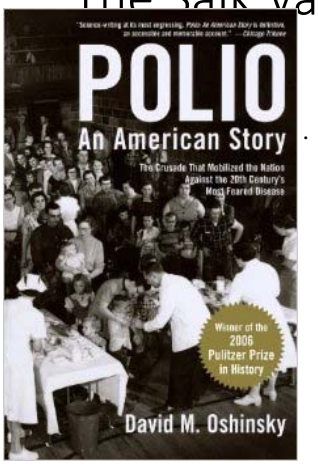
- Frame
- Specify
- Collect
- Analyse
- Communicate

[4] Chapter 0 describes in detail what we term “the quantitative reasoning framework”, which is a systematic approach for tackling a wide range of real-world issues. The current chapter fits well in the framework. As you will see, the bulk of the material is most relevant to the steps Collect, Analyse and Communicate. Specific details for Collect and Analyse will be revealed in the following lectures. The presentation style in plain language illustrates the Communicate step. The examples used do not explicitly discuss the remaining steps, Frame and Specify, but their relevance will be briefly described. Given a new problem, you are expected to examine all five steps to the best of your ability. I look forward to sharing the rest of the chapter with you.

GER1000 Design of Studies
Unit 2: The Salk Vaccine

[1] This unit introduces controlled experiments, which afford a reliable method to detect causal relationships. The ideas will be developed through a real example from public health.

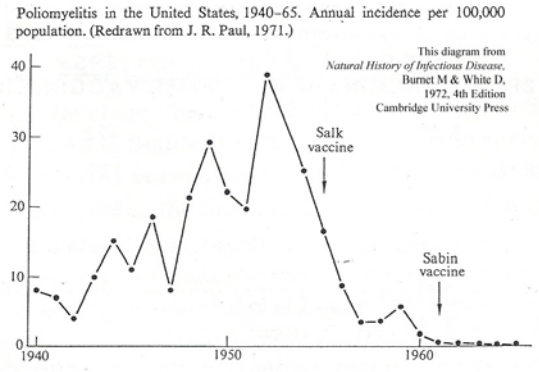
The Salk Vaccine



- In 1954, the NFIP tested the Salk vaccine in a huge experiment.

[2] Polio, also known as infantile paralysis, is an infectious disease that strikes young children, often causing permanent paralysis. In the 1950's, American scientist Jonas Salk developed a vaccine that protected monkeys from polio and was safe when injected into human subjects in the laboratory. But it was important to conduct a field trial, to test the vaccine in the general population, outside the laboratory. The National Foundation of Infantile Paralysis, NFIP in short, would investigate the vaccine in the largest public health experiment, which involved over 1 million children.

What if the vaccine was given to all children?



[3] Imagine the vaccine was given to all children in 1954. Suppose the incidence of polio in 1954 was lower than 1953. Is this convincing evidence that the vaccine works? [Pause] The answer is no, because polio incidence fluctuates much from year to year, as shown in the left half of the graph. For example, the incidence in 1953 was about one third less than that in 1952. If the incidence in 1954 turned out lower than 1953, it could be due to the vaccine, or it could be that 1954 was not an epidemic year. Hence giving the vaccine to all children makes it hard to learn about its efficacy.

Treatment vs Control

- Compare
  - Treatment group: get vaccine.
  - Control group: no vaccine.
- Ethical dilemma: harm to a few, vs benefits for many.

[4] Thus, it is necessary to compare within the year 1954. Some children are given the vaccine, forming the treatment group. The control group are deliberately not vaccinated. This is an example of a controlled experiment. It seems cruel to deny potentially life-saving treatment from some children. But experience has shown that even with extensive laboratory work on a treatment, it is often not clear whether the benefits outweigh the risks. In this case, the risks include complications from injections, and the possibility of contracting polio from the vaccine. The medical consensus to the ethical dilemma is to tolerate the injustice to a relatively small number of subjects for the benefit of the larger population.

A Solution to the Ethics Problem?

- Treatment: with parental consent.
- Control: without parental consent.

Unequal groups sizes taken care of by using

- Rate: number of polio cases for every 100,000 children.

[5] The parents had to agree, or to give consent, if their child is to be injected. By giving the vaccine to children with parental consent, but not those without parental consent, it seems that the ethical problem is solved, since no one who wants the vaccine is denied the treatment. Is there a problem with this design? [Pause] You might see that since the two groups can be unequal in size, it is not right to compare the numbers of polio cases. Instead, we will look at the number of cases for every 100,000 children, which will be called a rate. The vertical axis of the graph on slide 3 shows rate. Using rates takes care of unequal group sizes. Rate is a central idea in many quantitative problems, so you should pay close attention to it.

Do the groups have similar risks?

- Treatment: children with parental consent, higher risk.
- Control: children without parental consent, lower risk.

- Bias against vaccine: its effect will be understated.

[6] Suppose that by this design it was found that the treatment group has a higher polio rate than the control group. Do we conclude that the vaccine is in fact harmful? The answer is no, because children with consent are more at risk at the beginning, for reasons that will be explained in the next slide. Therefore the vaccine will look worse than it is. The design is biased against the vaccine. For example, suppose the vaccine does reduce polio risk moderately, but the effect is smaller than the difference in rates between the two groups. Then we will see a higher polio rate in the treatment group, leading to the wrong conclusion that the vaccine is harmful.

Why are children with consent more at risk?

Save the children

[7] Why are children with parental consent more at risk? This fact is puzzling because these parents tend to be wealthier, hence more able to protect their children, such as keeping the household hygienic. Here is the explanation. Like many infectious diseases, an exposure to polio gives immunity from future infections. Children in poor households tend to contract polio from unhygienic environment early, like in the first six months. Many such children will recover completely, with the help of their mother's immunity. This is the reason the control group is at less risk to start with. We say that the effect of socio-economic status is confounded, or mixed up, with the effect of the vaccine.

Summary

- If control and treatment groups are similar, then a difference in response can be attributed to the treatment. Otherwise, the treatment effect may be confounded with some other factor.
- Key question: “Are the groups different, aside from the treatment?”

[8] In the Salk vaccine trial, children with and without consent had different polio risks, making it difficult to gauge the vaccine’s effect if it were given only to children with consent. Indeed, knowing that children with consent were more vulnerable, we predict this design would make the vaccine look worse. The point is general. For any study involving control and treatment groups, it is important to ask: “Are the groups different, aside from the treatment?” If the answer is yes, then there are potential confounders, and one has to think about how to get around them.
In the vaccine example, we know the answer to the question. Moreover, we also know why children with consent were more vulnerable to polio. In general, such detailed knowledge may take years of hard work to attain. Yet it is important to ask this question, even if we may not be able to answer it completely at the present moment.

GER1000 Design of Studies
Unit 3: The NFIP Study

[1] The unit “Salk Vaccine” introduces a field trial of a polio vaccine. The experiment was conducted in many US districts, involving over a million children in primary schools. It consisted of two parts. The first part, called the NFIP study, is the focus of this unit. The second part will be dealt with in the next unit.

The NFIP Study

Control	Treatment	Observed
Years 1 and 3	Year 2, with consent	Year 2, without consent

- What question will you ask?

[2] The NFIP decided to put children in years 1 and 3 in the control group, and to put year 2 children with parental consent in the treatment group. Year 2 children without consent were not in either group, but were merely observed. This is called the NFIP study, which accounted for almost one million children. What question will you ask about this design?

A reasonable question

- "Are the groups different, aside from the treatment?"
- Now try to answer the question.

[3] A reasonable question is "Are the control and treatment groups different, aside from the treatment?". [Pause] If the groups are different, then the vaccine's effect may be hard to determine, because of confounding. This point was discussed in the unit "Salk Vaccine", which you may want to review, but the point applies to any controlled experiment. After understanding why we should ask such a question, you should try to answer it in the context of the NFIP study.

Answer

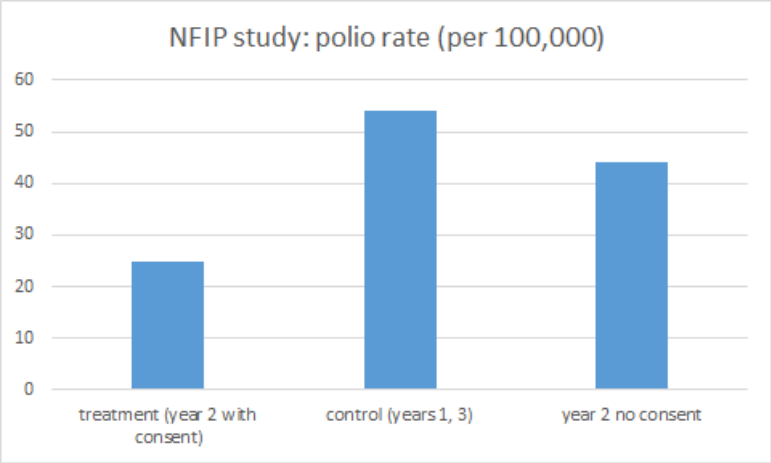
- Polio risk is smaller in control group. Groups are different, aside from treatment.
- The NFIP study is biased against the vaccine.

[4] The NFIP control and treatment groups are different, aside from the treatment. In the unit “Salk Vaccine”, we learnt that children with consent are more prone to polio. Hence, the control group, a mixture of children with and without consent, is at a lower risk than the treatment group. Hence any protective effect of the vaccine will appear smaller. In other words, the NFIP study is biased against the vaccine.

NFIP study: polio rate (per 100,000)

	Size	Rate
Y2 with consent (treatment)	225,000	25
Y1 and Y3 (control)	725,000	54
Y2 no consent	125,000	44

[5] The table summarises the results of the NFIP study. The treatment group had 225,000 year 2 children with consent, and the rate of polio is 25 per 100,000. The control group had 725,000 children in years 1 and 3, and the rate is 54 per 100,000. The investigators also looked at 125,000 year 2 children without consent. The polio rate in this third group is 44 per 100,000.



[6] The bar graph is a summary of the table in the previous slide. The bars represent the three groups, and their heights reflect the respective polio rates. Note that this plot only displays the rates, and do not indicate the size of the groups, which are shown in the table.

Interpreting the NFIP result (1)

From the data, the vaccine seems to reduce polio rate by 54 – 25 = 29 cases per 100,000. Is its actual effect likely to be smaller or larger? Why?

[7] Here is a question on interpreting the NFIP result. From the data, the vaccine seems to reduce polio rate by 54-25=29 cases per 100,000. Is its actual effect likely to be smaller or larger? Why? You should refer to the table or the plot in previous slides. Spend a couple of minutes to come up with your answer before proceeding to the next slide, where a suggested answer is provided.

A reasonable answer (1)

From the data, the vaccine seems to reduce polio rate by 54 – 25 = 29 cases per 100,000. Is its actual effect likely to be smaller or larger? Why?

The actual effect of the vaccine is likely larger than 29 per 100,000. This is because in the beginning of the study, the treatment group (children with consent) is more at risk than the control group (a mixture of children with and without consent).

[8] This is a reasonable answer. The actual effect of the vaccine is likely larger than 29 per 100,000. This is because in the beginning of the study, the treatment group (children with consent) is more at risk than the control group (a mixture of children with and without consent). There are many ways to write a reasonable answer.

Interpreting the NFIP result (2)

Why is the rate of the third group lower than the control group (44 vs 54 per 100,000)?

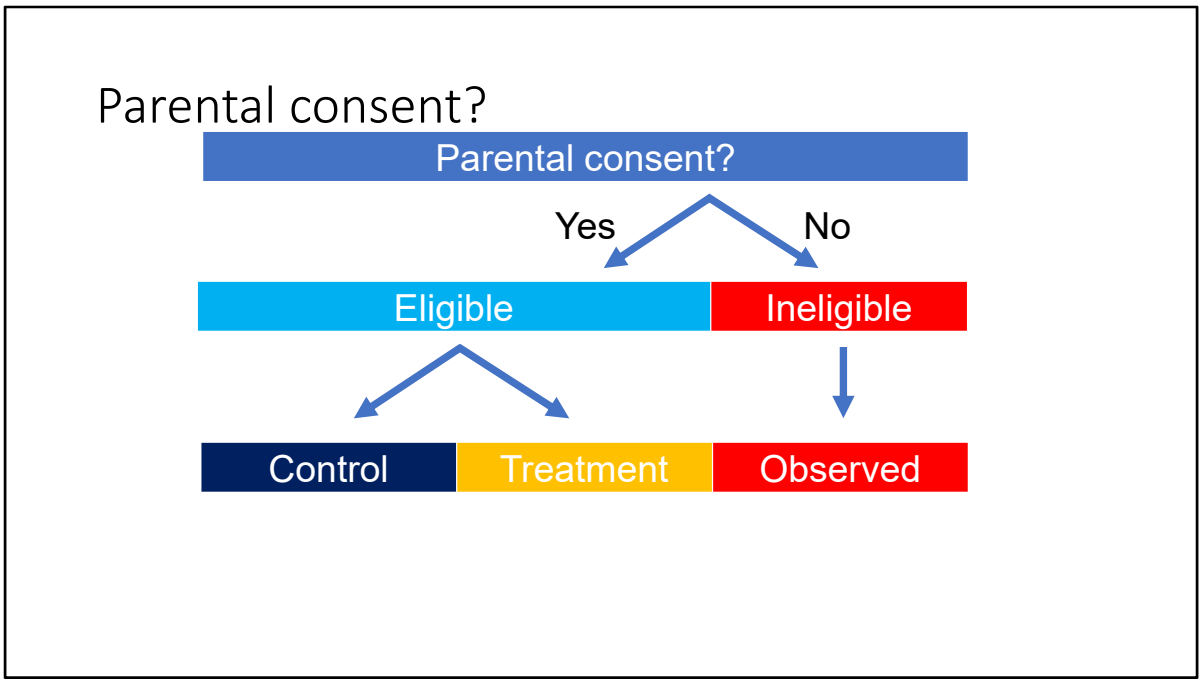
[9] This is another question on the NFIP study. Why is the rate of the third group lower than the control group (44 vs 54 per 100,000)? Like the previous question, try to answer this one before going to the next slide.

A reasonable answer (2)

Why is the rate of the third group lower than the control group (44 vs 54 per 100,000)?

Since the control group (a mixture of children with and without consent) is more at risk than the third group (children without consent), its polio rate is higher.

[10] Here is a reasonable answer. Since the control group (a mixture of children with and without consent) is more at risk than the third group (children without consent), its polio rate is higher. In both answers, we made use of the fact that children with consent are more vulnerable to polio.



[11] Generally, it makes ethical sense to put children without parental consent to the control group, but this can make the control and treatment groups different, thus biasing the experiment. The NFIP study is a concrete example, where the control group has a lower risk than the treatment group in the beginning of the trial. To get around this problem, some statisticians suggested that only children with consent were eligible for the study. This means that only children with consent should be assigned to control and treatment groups. Children without consent will only be observed, and nothing should be done to them. The assignment should be done carefully, so that the groups have similar polio risks. This way, any confounding will be minimised. The next unit discusses how to do the assignment.

Summary

- Children with consent more prone to polio: NFIP study biased against vaccine.
- Solution: Enroll only children with consent, then assign to control and treatment groups.
- Subjects who refuse treatment should be excluded.

[12] The NFIP study was the first part of the Salk vaccine trial. The control group consisted of a mixture of children with and without parental consent, while the treatment group only had children with consent. Since children with consent were more prone to polio, the NFIP study was biased against the vaccine. To avoid this bias, in the second part of the trial, only children with consent were enrolled. These children were then assigned to the control and treatment groups. The general lesson for controlled experiments: subjects who refuse treatment should not be put into the control group. Instead enrol only subjects who agree to being assigned to control or treatment.

GER1000 Design of Studies
Unit 4: Randomised Assignment

[1] We saw that the NFIP study was biased, because children with consent and without consent had different polio risks. This unit describes an effective way to deal with this problem.

The problem of assignment

- Eligible children: those with parental consent.
- Hard: using expertise to assign eligible children to control and treatment groups with similar polio risks.

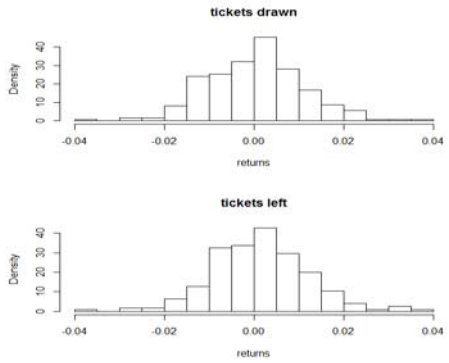
[2] After the NFIP study was done, it was decided to enrol only children with parental consent in the second part of the Salk vaccine trial. These eligible children then had to be assigned to control and treatment. It may seem reasonable that a scientist, such as Dr Salk himself, would do a good job, so that the groups had similar polio risks. But experience shows that human judgement is prone to bias. Even though the eligible children are more similar to each other than to the children without consent, there are other factors that affect polio risks. Without a sound knowledge of these factors, the task is almost impossible.

Randomised assignment

- An impartial procedure using chance, like “random draws without replacement”.
- With large number of subjects, it is very likely that the two groups are similar in all aspects.

[3] The best method of assignment uses blind chance carefully. Imagine writing the children's names on identical pieces of paper, or tickets, which are put into a large box. Mix the tickets thoroughly, then pick one out without looking. This is the first person in the treatment group. Repeat the process of mixing thoroughly and picking one from the remaining tickets in the box, until half the tickets are out of the box. Names that are picked form the treatment group, and names left in the box are the control group. At every draw, all the tickets in the box have equal chance of being picked. The mechanism is the same as a lucky draw at a dinner party, and is called “random draws without replacement”. This is an example of randomised assignment. If the number of enrolled subjects is large, by the laws of probability, the two groups tend to be similar in all respects. In particular, they will have similar polio risks.

Randomised assignment demonstration



[4] This is an illustration of the power of randomised assignment. The percentage returns of S&P companies in 2013 are written on 1,000 tickets. The returns range from -4% to 4%. A computer programme was used to simulate 500 random draws of tickets without replacement. As shown by the two histograms, the two groups have rather similar distributions. In actual experiments, computers are often used to do randomised assignment. Although computer algorithms are not truly random, they work very well. In real experiments, the control and treatment groups can have different sizes, due to other reasons. Nevertheless, if the sizes are quite large, then a randomised assignment tends to produce two very similar groups.

Random is not haphazard

- “Random” has strict meaning, different from informal usage.
- If tickets are not mixed thoroughly, the draws are not random.

[5] We often use the word “random” to mean disorderly or haphazard. For example, we might say “As I was waiting for bus, a random person starts talking to me about pineapple tarts.” Nobody will think this person was randomly picked from a group of well-defined individuals. But in the context of controlled experiment, the word “random” has a strict and rather specific meaning related to an impartial chance mechanism.
An example of haphazard assignment is as follows. If the tickets are not mixed thoroughly, the draws will likely produce tickets which enter the box last. This can cause the two groups to be quite different, if for some reason the boys’ tickets are put in after the girls. It may seem reasonable to describe any method of drawing tickets as “random”, but this is a mistake.

Randomised Controlled Double-blind Experiment

- Placebo
- Blinding subjects
- Blinding doctors

[6] We return to the Salk vaccine trial. The second part of the experiment was randomised controlled, meaning children with consent were randomly assigned to two groups. This is good, since the groups would be very similar at the start of the experiment. The treatment group will be injected with the vaccine. It seems natural to leave the control group alone, but this can cause bias. For example, the control group, knowing that they do not receive the vaccine, might adopt more precautions than the treatment group. More mysteriously, merely knowing that one gets a treatment, even if it does nothing, can produce a response. An effective way to deal with such biases is to inject the control group with a placebo, a substance with no effect, such as a salt solution, which looks identical to the vaccine. Moreover, all children were blinded: they did not know their assigned group. Then the two groups would respond the same way to the idea of treatment. Some polio cases are hard to diagnose. To prevent bias, doctors who made diagnoses were also blinded. An experiment is called double-blind if both subjects and doctors are blinded about the assignment. The randomised controlled double-blind experiment was conducted in many school districts.

Randomised Experiment Results

	Size	Rate (per 100,000)
Treatment	200,000	28
Control	200,000	71
No consent	350,000	46

[7] Does the vaccine work? Yes, because polio rate is lower in the treatment group than the control group. Indeed, we can confidently say that giving the vaccine cuts polio rate by 71-28 = 43 cases per 100,000. In other words, the vaccine roughly prevents 43 polio cases in every 100,000 children. The no consent group did not get the vaccine, like the control group. Why is the rate for the no consent group lower? [Pause] The explanation is that the no consent group consisted of poorer children, who were more resistant to polio.

Rates Put Together

Randomised Experiment		NFIP Study	
Treatment	28	Y2 with consent	25
Control	71	Y1 and Y3	54
No consent	46	Y2 no consent	44

[8] By putting the results of both the randomised experiment and the NFIP study together, we confirmed that their treatment groups were similar, but their control groups were quite different. In the NFIP study, the control group was a mixture of children with and without consent, so were more resistant to polio. As a result, the vaccine's effect is understated in the NFIP study, as 29 per 100,000. The actual effect is more like 43 per 100,000, as indicated by the randomised experiment.

The skeptic's hypothesis

- Maybe the vaccine has no effect: some children are fated to get polio. By pure luck, more of these children were assigned to the control group.
- If this is correct, the chance of seeing the difference or something more drastic is about one in a billion. Skeptic is unlikely to be right.

[9] Randomised assignment has another technical advantage, which will only be outlined here. Suppose a skeptic insists that the polio vaccine has no effect. How will he explain the result? He might say: Some children are fated to get polio, and this cannot be changed by the vaccine. The difference in rate is purely because of chance: by the luck of draw, too many of these children were assigned to the control group. This is a powerful argument, but one that can be refuted. Statisticians can calculate the chance of getting a difference as large as, or larger than the observed difference of 43 per 100,000, assuming the skeptic is right. This turns out to be absurdly small... one in a billion. The conclusion is that the skeptic's hypothesis of no effect is very unlikely. This kind of calculation is justified by the randomised assignment.

Additional points

- Randomised controlled double-blind experiment also useful for comparing new treatment to old treatment, instead of placebo.
- Conclusion may not apply to ineligible subjects.
- Ideas apply to other fields, like education.

[10] The randomised controlled double-blind experiment minimises confounding, and is the best design for comparing two treatments. Governments typically require a new therapy to be proved by several such experiments before allowing its use. For some diseases, there is an existing treatment. Then the new treatment should be compared to the old, instead of to a placebo. The new treatment should look identical to the old, and both subjects and doctors should be blinded.
The Salk vaccine was found effective by studying children with parental consent. This conclusion may not apply to the ineligible children, those without consent. This is the result of a compromise between minimising bias and generalisability. In practice, the vaccine would be offered to all children, perhaps with initial close monitoring of children without consent.
The ideas developed so far apply to interventions in other fields. For example, suppose a university wants to find out if a new academic programme is better than the existing one, by looking at the income of graduates. You can think about how to design a study to answer this question, applying the ideas developed so far.

Summary

- With numerous eligible subjects, treatment and control are likely similar.
- Ideas behind randomised controlled double-blind experiment useful for other kinds of problems.

[11] In summary, randomised assignment refers to the use of an impartial chance mechanism to assign eligible subjects to control and treatment groups. If the number of eligible subjects is large, it is very likely that the groups are similar in all respects, hence confounding is minimised. In addition, both subjects and investigators should be blinded to the assignment, as further precautions against bias.
The randomised controlled double-blind experiment is the gold standard for proving new medical therapies. But the underlying ideas about the design apply to intervention studies in many other fields, even if the plan cannot be implemented for valid practical reasons.

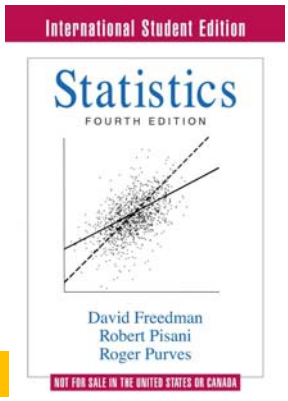
GER1000 Design of Studies
Unit 5: Non-randomised Controls

[1] We have seen that randomisation is the best way to assign eligible subjects to control and treatment groups. In this unit, we will reinforce this message by looking at more examples of non-randomised experiments.

Two kinds of controlled experiments

- Randomised: assign eligible subjects at random.
  only subjects that can
- Non-randomised: easy to assign healthy to treatment, less healthy to control.

See FPP page 10.



[2] A randomised controlled experiment enrols only eligible subjects. Patients who are too sick, for example, are excluded. The eligible subjects are randomly assigned to control and treatment. If the number of subjects is large, confounding is minimised. An experiment which does not use randomised assignment is called “non-randomised”. In the polio vaccine example, putting children with parental consent into treatment, and others into control, is non-randomised. In many medical studies, it is easy to assign healthier subjects to treatment, and less healthy ones to control. See FPP page 10 for an illustrating cartoon. The bottom line is that in non-randomised experiments, the control and treatment groups can be quite different aside from the treatment, so bias from confounding is likely.

The Portacaval Shunt

	Degree of Enthusiasm			
Design	Marked	Moderate	None	Total
No controls	24	7	1	32
Controls, but not randomised	10	3	2	15
Randomised controlled	0	1	3	4

[3] The portacaval shunt is a surgery to treat a liver disease called cirrhosis. In the table, 51 experiments are categorised two ways: the rows indicate the types of controls used, the columns indicate the nature of conclusions. For example, a "marked degree of enthusiasm" means the study concluded strongly that the surgery worked. 32 studies had treatment groups, but not control groups; of these 24, or 75%, concluded the shunt was effective. 15 studies had control groups which were not randomised, and 10, or 67%, were in favour. Finally, there were 4 randomised controlled experiments, and none of them found the shunt was effective. Based on the table, what is your conclusion about the portacaval shunt? [Pause] You should put your faith in the randomised controlled experiments, and conclude that the surgery does not help.

Three-year survival rates

	Randomised	Not randomised
Surgery	60%	60%
Controls	60%	45%

[4] The table shows the percentage of subjects who lived longer than three years after the portacaval shunt surgery. The third column comes from the 15 studies with non- randomised controls. Here, 60% of the subjects in treatment survived at least three years, compared to 45% in control. This seems to show the shunt helps. But the second column, from the 4 randomised studies, confirms that it is ineffective. Notice that the treatment group of the non-randomised studies seem similar to those in the randomised studies. However, the controls in the non-randomised studies seem less healthy. Probably, there was a tendency to assign less healthy subjects to control, which creates a bias for the surgery.

Historical controls

- Subjects from the past are compared to current subjects undergoing a new treatment.
- Confounding is a real issue.

[5] Random assignment is a lot of hard work. It is tempting to compare a treatment group to a historical control group, meaning subjects from the past, such as from medical records. For example, to find out the effect of using calculators in schools, one might compare current students with students 40 years ago, on how accurately problems are solved. The basic caveat applies: a historical control group may differ from a current treatment group in many other ways besides the treatment, so confounding is likely a real problem.

must make sure historical control groups are doing things exaclty the same way as now so only differ in 1 way

Coronary bypass surgery three-year survival

	Randomised	Historical
Surgery	88%	91%
Controls	83%	71%

[6] Coronary artery disease can be treated by a bypass surgery. In 6 randomised studies, 88% of the treatment group survived at least three years, compared to 83% in the control group. Thus, the surgery seems slightly effective. But in 9 studies with historical controls, the difference is much larger, 91% versus 71%, giving the impression that the surgery is very effective. It is easy to choose weak patients as historical controls, and this is probably what happened here.

Summary on Controlled Experiments

- Randomised assignment and double-blinding are important.
- FPP: flow chart (page 7), summary (pages 10,11).
- QR framework: Collect, Analyse, Communicate. Also Frame and Specify.

[7] In controlled experiments, the investigators decide which subjects go to control and treatment groups. The best assignment is done at random, like a lucky draw. Given enough subjects, the laws of probability virtually ensure the two groups are similar, so confounding is minimised. Double-blinding is an important precaution. Experiments with no control group and those that use historical controls are usually bad. See FPP page 7 for a flow chart on controlled experiments. For a more complete summary, please read pages 10 and 11. Note that we have paid a lot of attention to bias or confounding. This is directly relevant to the step called Collect in the QR framework, which is concerned with the collection of good data to answer the question of interest. We have compared polio rates of control and treatment groups, and made similar comparisons in other examples, which are relevant to the step Analyse. The step Communicate permeates the presentation, as mentioned in the Introduction to this chapter. Finally, the steps Frame and Specify have not been discussed, so they will be briefly touched on now. In the polio example, the research question is about the vaccine's efficacy, and the response is whether or not the child has polio. Here, as in other medical examples, we take for granted that the question and response are appropriate. More specifically, we trust that the doctors used reliable diagnoses methods. However, wherever possible, we should ask whether they are appropriate, as outlined in the steps Frame and Specify, keeping in mind that getting these steps wrong can make the whole study quite worthless. For instance, in a study on factors contributing to happiness, we might wonder about the reliability of the measurement of happiness in an individual.

1: ensure that sample to test is not biased
2: ensure splitting of groups is random
3: ensure that way to collect results is not flawed

GER1000 Design of Studies
Unit 6: Smoking and Health

[1] In previous units, we studied controlled experiments, where the investigator decides which subjects are in control and treatment groups. A randomised controlled double-blind experiment is the best method to find out whether an intervention really changes a response, for instance, a medical treatment, or an education programme.

Controlled Experiment vs Observational Study

	Controlled Experiments	Observational Studies
Assignment by …	Investigators	Subjects

- Long-term effects of smoking can only be investigated through observational studies.

[2] Many questions are impractical to answer with controlled experiments. For example, very few subjects will agree to smoke for twenty years to study the effect on health. The investigator has to resort to an observational study, where she merely records what happens to the subjects, who have chosen to smoke or not to smoke. For observational studies, the words “control” and “treatment” are used as if we are dealing with an experiment. In this example, the control group are the non-smokers and the treatment group are the smokers.

Smoking and heart disease

- A city has 100,000 residents.
  - Among 15,000 smokers, 38 have heart disease.
  - Among 85,000 non-smokers, 44 have heart disease.

	Number
Smokers (15,000)	38
Non-smokers (85,000)	44

[3] Smokers are much more likely to have heart attacks, lung cancer and other diseases than non-smokers. In other words, a much higher fraction of smokers are sick, compared to non-smokers. We say there is a strong association between smoking and disease. Here is an example that reflects Singapore roughly. Imagine a city of 100,000 residents consisting of 15,000 smokers and 85,000 non-smokers. Suppose 38 smokers and 44 non-smokers have heart disease. These numbers can be represented in a table as shown.

Association

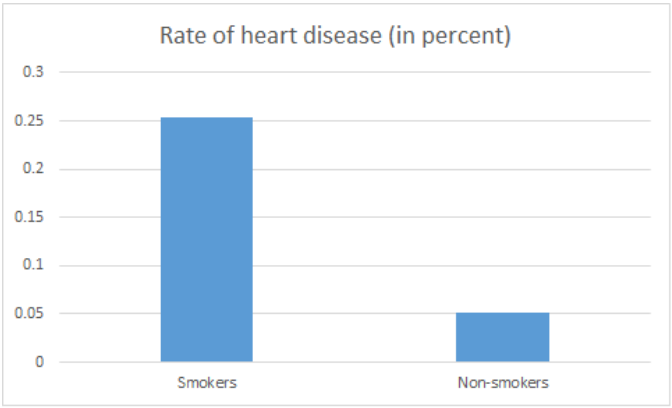
- Rate of heart disease among . . .
  - Smokers: $\frac{38}{15{,}000} \times 100\% = 0.25\%$
  - Non-smokers: $\frac{44}{85{,}000} \times 100\% = 0.05\%$

	Number	Rate
Smokers (15,000)	38	0.25%
Non-smokers (85,000)	44	0.05%

Association between smoking and heart disease

[4] The rate of heart disease among smokers is 38/15,000x100%, or 0.25% to two significant figures. This is about 5 times the rate among non-smokers, which is 0.05%. Because of this difference, we say there is an association between smoking and heart disease. The table presents both numbers and rates in the second and third columns. Since a rate can be computed from a number, and vice versa, often tables only show the rates, but not the numbers.

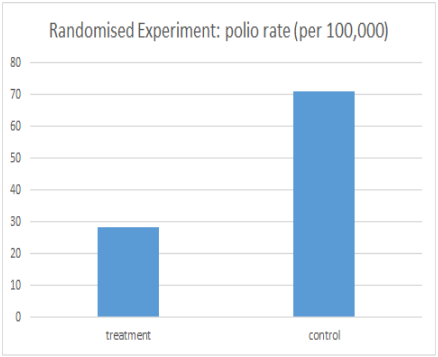
Association btw smoking & heart disease



[5] This is a bar graph. The height of a bar represents the rate of heart disease. The association between smoking and heart disease appears as bars with different heights. The bar graph is a pictorial summary of the table in the previous slide. It has less information than the table, since the numbers of smokers and non-smokers are not indicated. However it shows the association quite clearly.

Salk randomised experiment

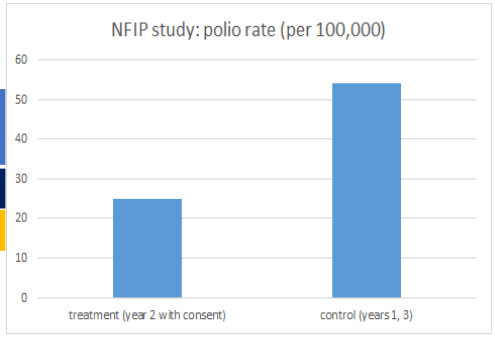
	Rate (per 100,000)
Treatment	28
Control	71



[6] The concept of association is crucial, so we look at a few examples. The table is from the Salk randomised experiment presented in unit “Randomised Assignment” slide 7. The polio rate in the treatment group is 28 per 100,000, while that in the control group is 71 per 100,000. Since the rates are different, polio is associated with the vaccine. Because the experiment is randomised and double-blind, we are confident that the association is entirely caused by the vaccine.

NFIP study

	Rate (per 100,000)
Treatment	25
Control	54



[7] Another example comes from the first part of the Salk vaccine trial, discussed in the unit “NFIP study”. Here, the polio rates in the treatment and control groups are 25 and 54 per 100,000. So there is also an association between polio and vaccine. Based on these numbers alone, we are unable to quantify the effect of the vaccine, since the treatment group are more likely to be richer, hence more at risk at the beginning of the study, as discussed in unit “Salk Vaccine”. Recall that the NFIP study is a controlled experiment, but the assignment is not randomised.

Association is not Causation

Study	Association due to
Salk randomised experiment	Vaccine only.
NFIP	Vaccine, family background, other factors.
Smoking and heart disease	Smoking? Other factors?

[8] Both the randomised experiment and the NFIP study are controlled experiments, and their results suggest that the vaccine works. The non-randomised NFIP study is harder to analyse than the randomised experiment, since the vaccine effect is mixed up with family background, and there can be other factors involved. We should be even more cautious about observational studies, where the subjects already assigned themselves to treatment and control. In general, going from association to causation requires some careful thinking, because there could be alternative explanations. For the moment, we will just be content with the slogan “Association is not causation.”, meaning an association does not always arise from a causation. Since the evidence against smoking is observational, it is valid to ask whether smoking really causes diseases. A later slide will briefly indicate how we arrive at a strong general consensus that smoking is indeed harmful.

Two-by-two table

	Heart disease	No heart disease	Row total
Smokers	38	14,962	15,000
Non-smokers	44	84,956	85,000
Column total	82	99,918	100,000

- Rate of heart disease among:
  - Smokers: $\frac{38}{15{,}000} \times 100\% = 0.25\%$
  - Non-smokers: $\frac{44}{85{,}000} \times 100\% = 0.05\%$

[9] The table classifies the residents of our fictional city into four types according to smoking status and heart health. The last column shows the row totals. For example, the number of smokers is 15,000. Similarly, the last row shows the column totals. For example, there are 38+44=82 residents with heart disease. Note that the table is just another way to present the information given earlier. It is called a 2x2 table because smoking and heart disease are either present or absent. Using the table, it is quite straightforward to calculate the rate of heart disease among smokers: 38/15,000 x 100% = 0.25%, and among non-smokers: 44/85,000 x 100% = 0.05%. These rates, which illustrate the association between smoking and heart disease, have been obtained earlier.

Flipped rates

	Heart disease (HD)	No heart disease	Row total
Smokers	38	14,962	15,000
Non-smokers	44	84,956	85,000
Column total	82	99,918	100,000

- Rate of smoking among HD: $\frac{38}{82} \times 100\% = 46.3\%$
- Rate of HD among smokers: $\frac{38}{15{,}000} \times 100\% = 0.25\%$
- Shorthand: rate(smoke | HD) = rate of smoking among HD,

  rate(HD | smoke) = rate of HD among smokers.
- In general, rate(smoke | HD) need not equal rate(HD | smoke).

[10] Let “HD” stand for “heart disease”. From the table, the rate of smoking among people with HD is 38/82 x 100%, or around 46.3%. It is very different from 0.25%, the rate of HD among smokers. This is because there are 15,000 smokers, but there are only 82 persons with heart disease. A shorthand simplifies the discussion. Denote the rate of smoking among people with HD by rate(smoke | HD), which is read “rate-smoke-given-HD”. Similarly, the rate of HD among smokers will be shortened to “rate-HD-given-smoke”. If we collect another data set on smoking and HD, the two “flipped” rates will take new values, and they will not be equal to each other, unless there are equal number of smokers and HD patients. This observation holds more generally for any 2x2 table.

Two views on association

	HD	No HD	Row total
Smokers	38	14,962	15,000
Non-smokers	44	84,956	85,000
Column total	82	99,918	100,000

- rate(smoke | no HD) = $\frac{14,962}{99,918}$ × 100% = 15.0% < 46.3% = rate(smoke | HD).
- Difference confirms association between smoking and HD, just like comparing rate(HD | smoke) = 0.25% and rate(HD | not smoke) = 0.05%.
- Both views are valid, though one may seem more natural.

[11] The rate of smokers among those without heart disease, or rate(smoke | no HD), is 14,962/99,918 x 100% = 15.0%. It is less than rate(smoke | HD), which is 46.3%, as seen in the previous slide. In other words, smokers are more common among people with heart disease than among people without heart disease. This confirms the association between smoking and heart disease, just like the comparison of the heart disease rates 0.25% and 0.05% in slide 9. When you look for evidence of association, you can use either comparison. You might prefer one that seems more natural to the problem of interest.

IQ and wealth

	Poverty	Adequacy	Row total
Low IQ (< 90)	12	48	60
High IQ (> 90)	10	170	180
Column total	22	218	240

Adult US population; numbers in millions.

(1) rate(poverty | high) = 5.6% < 20.0% = rate(poverty | low).

(2) rate(high | poverty) = 45.5% < 78.0% = rate(high | adequacy).

→ (1) and (2): equally valid evidence of association.

- rate(poverty | high) ≠ rate(high | poverty).

[12] Here is another example. The table is a rough representation of the US adult population. For example, there are 12 million people who have IQ less than 90 and live in poverty, and 60 million who have low IQ. You should check that the rate of poverty among high IQ is 5.6%, which is less than the rate of poverty among low IQ, 20.0%. So there is an association between IQ and wealth, which can also be established by the other view: the rate of high IQ among people in poverty is 45.5%, which is lower than the rate of high IQ among people in adequacy, which is 78.0%. Notice also that the two flipped rates are unequal.

Positive Association

- A, B population characteristics, with

  $0 < \text{rate}(A) < 1$ and $0 < \text{rate}(B) < 1$.

  If rate(A | B) > rate(A | not B), or rate(B | A) > rate(B | not A),

  "A and B are positively associated."

- Slides 4,5: heart disease and smoking are positively associated. "Heart disease and smoking go together."
- Positive association is not causation.

[13] Let A and B be characteristics in a population. Assume that rate(A) is strictly between 0 and 1, meaning some people have A, and some do not have A. We assume the same about B. Suppose that rate(A | B) > rate(A | not B), meaning A is more common among people with B than among people without B, or rate(B | A) > rate(B | not A), meaning B is more common among people with A than among people without A. Then we say A and B are positively associated. Check in slides 4 and 5, that heart disease and smoking are positively associated. Intuitively, heart disease and smoking tend to go together in this population.
We do not allow rate(A) to be 0, because then everyone does not have A, and we will have trouble calculating rate (B | A). If rate(A) is 1, then by the same reasoning, we will have difficulty with rate(B | not A). By the same consideration, we also do not allow rate(B) to be 0 or 1.
A positive association is often the starting point of an investigation about whether A causes B or B causes A. Recall that association is not causation. One has to rule out confounders to work towards showing causation.

Negative Association

- A, B population characteristics with 0 < rate(A), rate(B) < 1.

  If rate(A | B) < rate(A | not B), or rate(B | A) < rate(B | not A),

"A and B are negatively associated."

- Slide 12; poverty and high IQ are negatively associated.

- If A and B are negatively associated, then "not A" and B are positively associated.

[14] Assume again that for both A and B, the rate is strictly between 0 and 1. If rate(A | B) < rate(A | not B), or rate(B | A) < rate(B | not A), then A and B are negatively associated. Check that in slide 12, poverty is negatively associated with high IQ.
Negative association is like a positive association viewed backwards. In slide 12, check that adequacy is positively associated with high IQ: the two characteristics tend to go together. This illustrates a general rule. If A and B are negatively associated, then "not A" and B are positively associated. Similarly, in slides 4 and 5, heart disease is negatively associated with non-smoking.

Is there an association?

- Hard to observe no association:

  rate(A | B) = rate(A | not B), or rate (B | A) = rate(B | not A).

- What if the rates differ very slightly? Whether this matters depends on the situation. Not an issue in module.

[15] The condition for no association is neat: rate(A | B) equals rate(A | not B), or rate(B | A) equals rate(B | not A). However, this hardly occurs in real data. The two rates can often be quite close to each other, and common sense suggests we should say “The weak association does not matter, and it is like there is no association.” This is not a simple issue: whether a small difference matters depends on the specific problem, and often subject knowledge is needed to make a sound judgment. In this module, you are not expected to decide whether a difference matters in practice.

Summary

- Observational study: subjects already assigned to groups.
- A, B population characteristics with $0 < r(A), r(B) < 1$. They are associated if

$$r(A \mid B) \neq r(A \mid \text{not } B), \text{ or } r(B \mid A) \neq r(B \mid \text{not } A).$$

A and B are ___	Condition
positively associated	(1) $r(A \mid B) > r(A \mid \text{not } B)$ or (2) $r(B \mid A) > r(B \mid \text{not } A)$
negatively associated	(1) $r(A \mid B) < r(A \mid \text{not } B)$ or (2) $r(B \mid A) < r(B \mid \text{not } A)$
not associated	(1) $r(A \mid B) = r(A \mid \text{not } B)$ or (2) $r(B \mid A) = r(B \mid \text{not } A)$

[16] The key concepts of this unit are observational study and association. In an observational study, subjects already assigned themselves to control or treatment; the investigator only determine which group each subject belongs, and the response. In contrast, an experimenter actively assigns subjects to control or treatment. Association involves comparison of rates between subpopulations. For formulae in this slide, we abbreviate “rate” to “r”. Two population characteristics A and B are associated if r(A | B) differs from r(A | not B), or r(B | A) differs from r(B | not A). An association can be positive or negative.

GER1000 Design of Studies
Unit 7: Confounding

[1] This unit will explain the importance of confounding in drawing causal conclusions from observational studies. To get the most out of this and subsequent units, you must be familiar with the concept of association as described in unit “Smoking and Health”.

Smoking and heart disease

	Rate of heart disease
Smokers (15,000)	0.25%
Non-smokers (85,000)	0.05%

Association between smoking and heart disease

- "Are smokers and non-smokers different, aside from smoking?"

[2] Recall our imaginary city where heart disease is associated with smoking: the rates of heart disease among smokers and non-smokers are different. An association suggests causation, but it is not conclusive. This question, analogous to one in the unit "Salk Vaccine", "Are smokers and non-smokers different, aside from smoking?" is relevant here.

Sex is a confounder

- Association between smoking and sex:

	Number of males	Fraction of males
Smokers (15,000)	12,500	83%
Non-smokers (85,000)	37,500	44%

- Association between sex and heart disease:

	Number	Rate
Males (50,000)	70	0.14%
Females (50,000)	12	0.02%

- Comparing smokers and non-smokers is like comparing males and females.

[3] Indeed they are different: smokers are more likely to be male. In this city, it turns out that 12,500 of smokers are male, or 83%, but only 37,500, or 44% of non-smokers are male. So there is a strong association between smoking and sex. Moreover, there is also a strong association between sex and heart disease: 0.14% for males, about seven times that for females. We say that sex is a confounder in the question of whether smoking causes heart disease.

About confounding

- A confounder is associated with both exposure and disease.
- Actual relationship between exposure and disease can be obscured by confounders.
- In an observational study, it is important to control for confounders.

[4] From observing an association between an exposure and a disease, we suspect a causal relationship. A confounder is a third variable that is associated with both exposure and disease. But if a variable is associated with only exposure or disease, then it is not a confounder. For example, if there is a gene that causes heart disease but has nothing to do with smoking, then it is not a confounder. The actual relationship between exposure and disease can be obscured by a confounder. In the example, sex exaggerates the effect of smoking. Since a higher fraction of smokers are male as opposed to female, and males have a higher rate of heart disease than females, it is not surprising that smokers have a higher rate of heart disease. In a controlled experiment, confounding can be minimised by randomised assignment, given enough subjects. In an observational study, the subjects have already assigned themselves to control and treatment, so confounding is a real problem. One way around it is to try to control for confounders. This will be explained in the next few slides.

Controlling for sex: Slicing

- Heart disease among males:

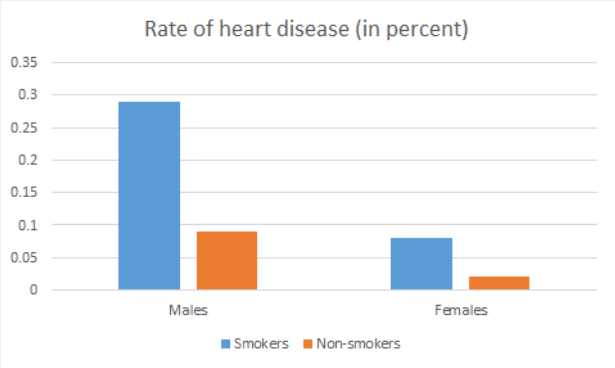
	Number	Rate
Smokers (12,500)	36	0.29%
Non-smokers (37,500)	34	0.09%

- Heart disease among females:

	Number	Rate
Smokers (2,500)	2	0.08%
Non-smokers (47,500)	10	0.02%

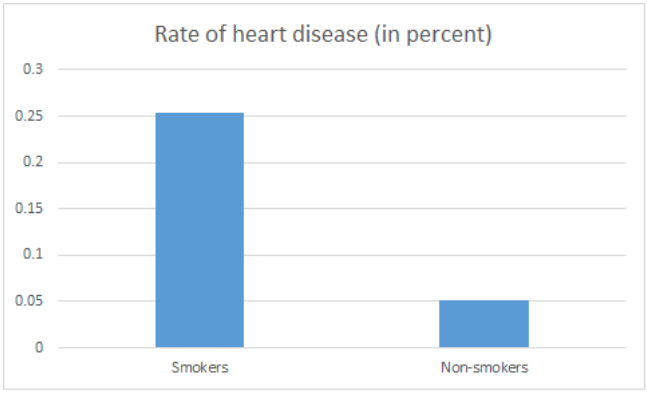
[5] To control for sex in the smoking example, we look at males and females separately. The tables show that smoking and heart disease are associated among males: 0.29% vs 0.09%, and they are also associated among females: 0.08% vs 0.02%. In both comparisons, sex is no longer a confounder, so the case against smoking is strengthened. We say that sex has been controlled for, and this method of separating males and females will be called “slicing”. Note, however, that it is not a standard term.

Slicing: graphs



[6] The bar graphs are made from the previous tables, showing only the rate of heart disease but not the number of people in each group. There is clearly an association between smoking and heart disease, in both males and females. Moreover, in both, the smokers are worse off.

In the beginning…



[7] Starting with a strong association between smoking and heart disease, we learnt that sex is a confounder. You might ask: “How would I think of sex as a confounder?”. This is a fair question. Experienced investigators tap on their vast and expertly knowledge of the subject, but even they may miss important confounders in new problems. What about novices like us? We should ask that all-important question: “Are the smokers and non-smokers different in some ways, other than smoking?” Remembering this question before data collection can prompt us to think of potential confounders, and to measure them in the planned study.

Is age a confounder?

- Is age associated with smoking?
- Is age associated with diseases?

[8] For smoking and diseases, is age also a confounder? There are two things to check. First: Is age associated with smoking? Second: Is age associated with diseases? Please pause for a few seconds to think. [Pause] It is clear that age is associated with diseases: elderly people are more likely to be sick, compared to younger people. People of different ages are likely to have different smoking habits. For instance, older people are probably heavier smokers. Thus, age is a confounder, which should be controlled for. An effective method is slicing, which was applied to control for sex before. This means separating the data sets into smaller ones according to age groups, then compare smokers and non-smokers within each group separately.

Controlling for sex and age

	Males	Females
21-30 years	Smokers vs Non	Smokers vs Non
31-40 years	Smokers vs Non	…
41-50 years	…	…
51-60 years	…	…
61-70 years	…	…

- The 10 separate comparisons between smokers and non-smokers control for sex and age.

[9] Controlling for sex and age together means comparing the health of smokers and non-smokers within groups like males of age 21 to 30 years, females of age 21 to 30 years, males of age 31 to 40 years, and so on, as indicated in the table. Notice that this is nothing but the slicing method applied to two factors simultaneously. If in most of the smaller comparisons, the smokers are still worse off than non-smokers, the case against smoking is further strengthened. This is indeed the case in many studies, further strengthening the case against smoking. In many studies of diseases, age and sex are confounders.

How do we know smoking is harmful?

Consistent results from

- Many animal experiments
- Many carefully controlled observational studies on humans

[10] This is roughly how we arrive at the consensus that smoking is bad for health, without relying on randomised experiments on humans. There is a large body of observational studies on humans, which show consistent association between smoking and diseases, even after careful controlling of many known confounders. Chemicals in cigarettes have been studied in extensive animal experiments, that clearly demonstrate harm. The vast collection of consistent results from numerous studies, both observational and experimental, persuades many governments to issue health warnings against smoking. This is a remarkable achievement, which also illustrates the difficulty of establishing causal relationships without randomised experiments.

Techniques for controlling a confounder

- Slicing is basic, but can be cumbersome.
- Statistical techniques could do the job, but there are issues.
- The Hormone Replacement Therapy.

[11] The method of slicing, which breaks the data into smaller, more homogeneous groups, is a good way to control for confounders. But it can quickly become cumbersome as the number of confounders increases, since the number of comparisons will become large quickly, and there may not be enough data. Many statistical techniques, such as regression, are widely used to control for confounders. Such methods are essentially more sophisticated versions of slicing. But like slicing, statistical techniques will not account for confounders that are not measured, and they may not control measured confounders adequately. Solid subject matter knowledge and a healthy dose of skepticism are needed to interpret the results of such analyses properly.

In the 1990's, hormone replacement therapy was prescribed to many menopausal women. It was believed to be beneficial, based on the statistical analysis of a large number of observational studies. But two randomised controlled experiments, including the large Women's Health Initiative involving nearly 70,000 women, found the therapy increased risks of stroke, heart attacks and breast cancer. This is a stark and humbling reminder that nature does not give out its secrets easily: observational studies can lead even experts astray. You are strongly encouraged to read the short discussion on page 918 of the PDF document "Stat_WHI", which is available in the IVLE.

Summary

- "Association is not causation".
- Observational studies prone to confounding. Knowledge and thinking useful for spotting potential confounders.
- Slicing is effective for controlling confounders.

[12] An association between two things often makes us wonder whether one causes the other, or believe this must be the explanation. The lesson here is that we should step back and think, because the association can be due to confounders: Association is not causation. When reasoning about a problem, we might think of one thing as "exposure" and the second thing as "disease". We need to be aware that the labelling is a convenience that should not stop us from asking whether the second thing might cause the first. On this point, you might want to revisit slide 2 of unit "Introduction".
Observational studies are prone to confounding. Coming up with potential confounders can be challenging. Subject knowledge, critical and imaginative thinking are helpful. A confounder can be controlled by studying the association in smaller data sets which are more homogeneous in the confounder. This method, called slicing, seems crude, but is very effective. The next two units elaborate on these essential ideas.

GER1000 Design of Studies
Unit 8: Yule-Simpson Paradox

[1] This unit is about an interesting numerical phenomenon known as the Yule-Simpson paradox. We will use a real example to illustrate the paradox, then describe how it arises. Then we will relate the paradox to confounding, slicing, and will introduce another way of controlling for confounding. However, before showing you the example, we need to understand a powerful fact on rates.

Cashless Dining in University

Group	Number	U	Rate
E (employees)	1000	700	rate(U | E) = 70%
S (students)	4000	3600	rate(U | S) = 90%
Total	**5000**	**4300**	**rate(U) = 86%**

rate(U | E) = 700/1000 x 100% = 70%
rate(U | S) = 3600/4000 x 100% = 90%

rate(U) = 4300/5000 x 100% = 86%

[2] Suppose a university employs 1000 people and serves 4000 students. The campus canteens offer an app for cashless dining, which attracts 700 regular users among the employees and 3600 regular users among the students. The table shows the rates of usage ("U") for the two groups, as well as the rate for the groups combined. The symbols rate(U | E), rate (U | S) and rate(U) were introduced in a previous unit "Smoking and Health". Note that rate(U) is calculated by dividing the total number of U in both groups by the combined group size.
The table shows that the overall rate is between the two group rates. For some of you, this might be obvious. For the rest of you, it is worthwhile to pay attention. This is a special case of a general rule.

Basic Rule on Rates

- In a population, let A and B be characteristics. The overall rate of B, rate(B), always lies between the two group rates:
- rate(B | A) and rate(B | not A)

- Let A = “is an employee”, or E, B = “is a regular user” or U.
  - “not A” means “is not an employee”, i.e., “is a student” or S.
  - Rule says rate(U) is between rate(U | E) and rate(U | S).

[3] To see the example as an instance of the general rule, we let A be “is an employee” or E, and let B be “is a regular user” or U. So “not A” means “is not an employee”, i.e., “is a student”, or S. Then the general rule says rate(U) is between rate(U | E) and rate(U | S), exactly as found in the previous slide.
It is important to be sufficiently familiar with the rule so that you can apply it correctly in new situations. It is helpful to confirm it in other examples, say by changing some of the numbers in the second and third columns of the table in the previous slide, or in new situations. Some of you are capable of using algebra to demonstrate the rule. You are encouraged to try, but it is not necessary to be able to do so in this module.

Elaborations on Rule

1. The rule assumes that rate(A) is strictly between 0% and 100%.
2. Rule holds even if we do not know the value of any of the rates.
3. If rate(B | A) = rate(B | not A), then rate(B) is also equal to the value.
4. The closer rate(A) is to 100%, the closer rate(B) is to rate(B | A).
5. If rate(A) = 50%, then rate(B) is exactly halfway between rate(B | A) and rate(B | not A).
   - Analogy: mixing two salt solutions.

50% employee & 50% students using means 50% total using

[4] We make some elaborations on the Basic Rule.
(1) The condition 0% < rate(A) < 100% means that in the population, some people have A, and some people do not have A. If it is false, for example, if everybody has A, then it is hard to make sense of rate(B | not A). We shall be making this assumption automatically below.
(2) In the example, we are able to calculate every rate from the table. But the rule holds even if we do not know the value of any of the rates.
(3) In the special case when there is no association between A and B, i.e., rate(B | A) = rate(B | not A), rate(B) must be equal to that value.
(4) From the table, there are four times as many students as employees, and rate(U) is closer to the students' rate than the employees' rate. This reflects the general fact. When rate(A) is closer to 100% than 0%, it is more than 50%, meaning more people have A than not.
(5) This result is not surprising, given the rule in (4).
(6) An analogous fact may be familiar from your experience. There are two bottles of salt solutions, the first one saltier than the second one. Empty both bottles into a large bowl and mix thoroughly. Then the resulting solution has a saltiness which is between the two initial solutions. If the first solution has a larger volume than the second solution, then the resulting saltiness will be closer to that of the first solution. In fact, salt concentrations obey the same mathematics as rates, though you are not expected to know the details for this module.

Treating Kidney Stone

	Treatment X			Treatment Y		
Size	Number	Success	Rate	Number	Success	Rate
Small	87	81	93%	270	234	87%
Large	263	192	73%	80	55	69%
Total	**350**	**273**	**78%**	**350**	**289**	**83%**

[5] Now we will use a real example to describe the Yule-Simpson paradox. Charig and co-workers compared the success rates of two treatments for kidney stone, X and Y. Each treatment was given to 350 patients, who were further divided into two groups according to the size of the stones. The table shows some data from their publication. See the Wikipedia entry “Simpson’s paradox” for details.
The overall success rates for X and Y (in red) suggest that Y is a better method. However, in the previous two rows, X has higher success rate than Y (green for small, purple for large). Somehow, when the two groups are combined, we get the opposite conclusion, that Y is better. This phenomenon, where a consistent relationship within small groups is reversed upon combining the groups, seems to be first noticed by Yule, and expanded significantly by Simpson.

View of Paradox via Association

	Treatment X			Treatment Y		
Size	Number	Success	Rate	Number	Success	Rate
Small	87	81	93%	270	234	87%
Large	263	192	73%	80	55	69%
Total	**350**	**273**	**78%**	**350**	**289**	**83%**

[6] Here is another way to describe the paradox, using the terms “positive association” and “negative association”, which were introduced in the unit Association. There is a positive association between X and success among patients with small stones, because rate(success | X) = 93% > rate(success | Y) = 87%. This is also true among patients with large stones. But when the two groups are combined, X and success have a negative association.

Explanation of Paradox

	Treatment X			Treatment Y		
Size	Number	Success	Rate	Number	Success	Rate
Small	87	81	93%	270	234	87%
Large	263	192	73%	80	55	69%
Total	**350**	**273**	**78%**	**350**	**289**	**83%**

[7] We can explain the paradox completely using the Basic Rule on Rates. Let us focus on treatment X. Given the two success rates of 93% and 73% (in orange), the rule tells us that the overall success rate must be between them. Indeed, 78% is between 93% and 73%, Moreover, it is much closer to 73%, because there were a lot more patients with large stones than small stones (263 vs 87). The story is similar with Y, but in the opposite direction: a lot more people with small stones received Y. So the overall success rate of Y is closer to 87% than 69%, to the extent of exceeding the overall success rate of X. This is how we get the puzzling appearance of X being worse than Y in the combined group.
Another view: the overall comparison is somewhat like comparing the 263 patients with large stones who received X with the 270 patients with small stones who received Y. Because small stones have higher success rates, this unfairly favours Y over X.

small stones are easier to treat so if X has more small treatment the rate might be more fair

thus must look out for the percentage of each category and how many ppl are in there
- since overall rate will be influenced by the indiv population number

View of Paradox via Confounding

	Treatment X			Treatment Y		
Size	Number	Success	Rate	Number	Success	Rate
Small	87	81	93%	270	234	87%
Large	263	192	73%	80	55	69%
Total	**350**	**273**	**78%**	**350**	**289**	**83%**

[8] The paradox can also be understood through the concept of confounding. To start, there is an association between treatment and success, because overall, Y has a higher success rate than X: 83% vs 78%. Note that in talking about this association, we combine the large and small stone sizes. We will show that stone size is a confounder, meaning it is associated with both treatment and success.

(1) Association with Treatment

Treatment	Number	Small	Rate
X	350	87	25%
Y	350	270	77%

rate(small | X) = 87/350 x 100% ≈ 25%

rate(small | Y) = 270/350 x 100% ≈ 77%

	Treatment X		Treatment Y	
Size	Number	Success	Number	Success
Small	87	81	270	234
Large	263	192	80	55
Total	350	273	350	289

[9] To show the association between stone size and treatment, we will calculate rate(small | X) and rate(small | Y). We use a relevant table derived from the original table, which is reproduced in the bottom right corner, except for two columns. Since 350 patients received X, and 87 of them had small stones, these numbers go to the first row of the table. Similarly, 350 patients received Y, and 270 of them had small stones; these go to the second column. Notice that the new table does not distinguish between patients who had success or not. The two rates show that small stone size is positively associated with Y.

(2) Association with Success

Size	Number	Success	Rate
Small	357	315	88%
Large	343	247	72%

rate(success | small) = 315/357 x 100% ≈ 88%

rate(success | large) = 247/343 x 100% ≈ 72%

	Treatment X		Treatment Y	
Size	Number	Success	Number	Success
Small	87	81	270	234
Large	263	192	80	55
Total	**350**	**273**	**350**	**289**

[10] To show the association between stone size and success, we will calculate rate(success | large) and rate(success | small). Like before, we use a relevant table derived from the original one. There are 87+270=357 patients with small stones, of whom 81+234=315 were successes; these numbers go to the first row of the table. Similarly, there are 263+80=343 patients with large stones, of whom 192+55=247 were successes; these numbers go to the second row. Notice again that the new table does not distinguish between patients who received X or Y. The two rates show that small stone size is positively associated with success.

Controlling for Confounding

	Treatment X			Treatment Y		
Size	Number	Success	Rate	Number	Success	Rate
Small	87	81	93%	270	234	87%
Large	263	192	73%	80	55	69%
Total	**350**	**273**	**78%**	**350**	**289**	**83%**

[11] The previous two slides show that stone size is a confounder. To control it by slicing, we divide the data to two groups according to stone size. But that is exactly in the first two rows of the original table. The two comparisons suggest that X is better than Y. Thus, controlling for the confounding presents us with a different interpretation of the data. In this example, the Yule-Simpson paradox can be explained by confounding. However, the reverse may not be true: If there is confounding, we may not observe the paradox.

paradox appears sometimes while controlling confounding

Yule-Simpson Paradox

- Let A and B be characteristics in a population, which consists of several subgroups. Suppose in each subgroup,
- rate(B | A) > rate(B | not A)
- When the subgroups are combined, it may happen that rate(B | A) ≤ rate(B | not A).

- In other words, if A and B are positively associated in every subgroup, when the subgroups are combined, A and B may be negatively associated, or not associated.

[12] Here is a general statement of the Yule-Simpson Paradox, first in terms of the rate symbols, then in terms of association.

Elaborations on the Yule-Simpson Paradox

- Suppose in most groups, rate(B | A) > rate(B | not A), but there are a few with rate(B | A) ≤ rate(B | not A). If the combined rates have rate(B | A) ≤ rate(B | not A), we also view it as a paradox. See FPP Chapter 2 Section 4 "Sex Bias in Graduate Admissions" for an interesting real study.

- The Yule-Simpson paradox also refers to the opposite case, where within each group, rate (B | A) < rate(B | not A), but when combined, rate(B | A) ≥ rate(B | not A).

[13] We make some elaborations on the Yule-Simpson Paradox.
For a real example of the more general paradox described here, you may read FPP Chapter 2 Section 4 "Sex Bias in Graduate Admissions". The explanation of the general paradox is analogous to that for two subgroups described before. It hinges on an extension of the basic fact, i.e., the overall rate of B must be somewhere within the range of the group-specific rates of B, whatever the number of groups there are in the population.
(2) We have discussed the paradox assuming rate(B | A) > rate(B | not A) in most or all subgroups, i.e., a positive association between A and B. But the opposite case of negative association can also occur, and for completeness, we state it here.

Summary

- Let A and B be characteristics in a population.
- Basic Rule on Rates: The overall rate of B, rate(B), always lies between rate(B | A) and rate(B | not A). The closer rate(A) is to 100%, the closer rate(B) is to rate(B | A).
- Yule-Simpson Paradox: Suppose the population consists of several subgroups, and in each subgroup, rate(B | A) > rate(B | not A). When the subgroups are combined, it may happen that rate(B | A) ≤ rate(B | not A).

[14] To wrap up this unit, here are the main learning points. You should try to apply them in new situations.

Summary on Observational Studies

- Confounding is main issue. Confounder is associated with both exposure and response.
- Ask “Are the groups different, aside from the exposure?”
- Controlling for confounders: slicing method, statistical methods.
- QR framework: Collect, Analyse. Others relevant too.

[15] In an observational study, the subjects are already assigned to control and treatment groups, hence can be quite different. So confounding is an important issue. Recall that a confounder must be associated with both exposure and response.

Spotting a confounder is challenging, but asking this question can help. “Are the groups different, aside from the exposure?” Potential confounders should be measured in the study.

A confounder can be controlled for effectively using the slicing method, by studying smaller data sets which are more homogeneous in the confounder. More sophisticated statistical methods based on this idea are available, but their correctness is not guaranteed.

For a more complete summary of observational studies, and an overview of the chapter “Design of Studies”, please read pages 27 and 28 of FPP. As mentioned before, the term “slicing” is not used in the book.

In terms of the QR framework, the detailed study of confounders speaks directly to the step Collect; while the slicing method and the adjusted admission rate belong to the step Analyse. For the relevance of the others, please see the last slide of the unit “Nonrandomised Controls”.