Causal Inference Reading Course Summary

Jessica Chau

University of Toronto - Department of Statistical Sciences

August 2019

1 Introduction

In every corporation, there are metrics to evaluate the overall performance of the company over a period of time. Whether the organization is achieving its targets, over-performing or under-performing, they will try to understand the variables that are causing the results. If the variable's impact on the outcome can be confirmed, it will allow the user to implement actionable insights. For example, if it was found that days with weather above 25°C increases ice cream sales, it would help the ice cream company relate weather to ice cream sales and better manage the upcoming demand for ice cream. As a result, the ability to evaluate causality is a common question that statisticians and analysts are asked. Without the foundations of statistics, people tend to conclude that causality is equivalent to association. This causal inference reading course, under the supervision of Professor Linbo Wang, has provided myself with the tools to further investigate how to conclude whether a variable of interest causes a particular outcome while being self conscious of the biases that commonly occur when conducting experiments. This report discusses the learnings and applications of the readings from the following materials:

- Causal Inference in Statistics: A Primer by Judea Pearl, Madelyn Glymour, Nicholas P. Jewell
- Causal Inference Book (Chapters 1-10 & 16) by Hernán MA, Robins JM (2019)
- Foundations of Causal Inference Slides by Professor Linbo Wang

This report will summarize the high-level learnings in the materials with a focus on the application of causal inference on an organization's marketing campaign and on a coffee bean experiment. This report assumes that the reader understands causal terminology and foundations of probability & statistics. This report provides summary of a dichotomous treatment, A, and its impact on outcome, Y for simplicity.

2 Association vs. Causality

2.1 Simpson's Paradox

Simpson's Paradox is an example of why it is important to understand the difference between association and causality. Simpson's Paradox is an event within statistics that occurs where a trend is observed when the data is grouped but the trend is reversed or disappears when the data is partition into groups. For example, in Table 1, Treatment B is more effective when the results are grouped whereas Treatment A is more effective when the results are partitioned by kidney stone size.

Stone Size or Treatment	Treatment A	Treatment B
Small Stones	**93%** (81/87)	87% (234/270)
Large Stone	**73%** (192/263)	69% (55/80)
Both	78% (273/350)	**83%** (289/350)

Table 1: Example of Simpson's Paradox

In the kidney example, the severity of the patient's kidney is a confounding variable that was not considered, which impacts the treatment type that patients received. As a result, Treatment A has larger cases of large stone cases whereas Treatment B has more cases with small stones and generally larger stones receive the better treatment, A. Through this example, it reminds myself to never rely solely on averages and understand that summary statistics should be the starting point of data analysis and not the only tool.

2.2 Causality

Causality is difficult to prove because we want to know the outcome of the individual under both no treatment and under treatment. Generally, the data will only be available for one of the options. For example, a researcher is interested in understanding whether individuals have a lower chance of mortality after a heart transplant, however it is only possible to know one outcome for each individual. As a result, for each individual in the study, only one of the counterfactual outcomes (the one that corresponds to the treatment value that the individual receives) is actually factual. In section 5 and 6, there will be two applications of causality, one study would have the treatment and non-treatment applied to per participant whereas in the other study, each participant can only have one treatment.

Generally, there are many people who confuse causality and association. It is important to understand the difference and the implications that causation and association have on the output variable. Hernán and Robins provides definitions of causal effect and association which is discussed below.

The average causal effect of treatment (variable), A, on outcome, Y, as:

$$P(Y^{a=1} = 1) - P(Y^{a=0} = 1) \neq 0$$

Treatment A and outcome Y are associated (or dependent) when:

$$P(Y = 1|A = 1) - P(Y = 1|A = 0) \neq 0$$

Although the definitions of causality and correlation may seem similar, they have different implications. In Professor Wang's slides, he provides an example that highlights the differences between association and causality:

"We observe that there tends to be higher ice cream sales on days when crimes are also higher. Does this mean that ice cream sales are causing more crimes to occur? Obviously no, the two variables are associated which is why we see a trend between ice cream sales and crimes. The true causal factor is likely temperature. Higher temperatures causes more ice cream to be sold (ice cream is a way for people to cool down) and more crimes (generally people are more agitated in hotter weather)."

When trying to prove causality, one method is to conduct a randomized controlled experiment. Generally, if the experiment controls for all confounders, association and causation would be the same. It is generally impossible to control for all confounders and instead we are given an observational study and asked to infer causality.

2.3 Randomized Experiments

In controlled randomized experiments, association is causation because the only difference between the treatment and control groups is the treatment variable. Algebraically this can be proven by:

$$E(Y^{a=1} = 1) - E(Y^{a=0} = 1) \text{ by definition of average causal effect}$$

$$E(Y^{a=1}|A = 1) - E(Y^{a=0}|A = 0) \text{ due to randomization, A and} Y^a \text{ is independent}$$

$$E(Y|A = 1) - E(Y|A = 0) \text{ due to consistency}$$

As mentioned in the textbook, perfect randomized experiments are unrealistic but are useful for introducing the key concepts of causal inference. In Table 2, the ? indicate the missing values of the counterfactual outcomes and randomization ensures that the missing values occurred by chance. When the individuals in the experiment are marginally randomized, it does not matter which individuals receive treatment. Exchangeability occurs when the probability of the outcome in the control group would have been the same probability as the treatment group if the individuals in the control group received the treatment.

$$P(Y^a = 1|A = 1) = P(Y^a = 1|A = 0) \text{ for all values of a}$$

	A	Y	Y^0	Y^1
Rheia	0	0	0	?
Kronos	0	1	1	?
Demeter	0	0	0	?
Hades	0	0	0	?
Hestia	1	0	?	0
Poseidon	1	0	?	0
Hera	1	0	?	0
Zeus	1	1	?	1
Artemis	0	1	1	?
Apollo	0	1	1	?
Leto	0	0	0	?
Ares	1	1	?	1
Athena	1	1	?	1
Hephaestus	1	1	?	1
Aphrodite	1	1	?	1
Cyclope	1	1	?	1
Persephone	1	1	?	1
Hermes	1	0	?	0
Hebe	1	0	?	0
Dionysus	1	0	?	0

Table 2: Table with counterfactual outcomes

2.4 Conditional Randomized Experiments

Conditionally randomized experiments occurs when randomization of the individuals in the study depends on the values of the variable, L. As a result, conditionally randomized experiments will not result in exchangeability of the treated and untreated. Table 3 displays that 65% (13 out of 20) are treated and 35% (7 out of 20) are untreated. L is a variable that determines whether treatment A would be provided which results in the conditional randomized experiment.
To verify exchangeability in conditionally randomized experiments, the following should be verified:

$$P(Y^a = 1|A = 1, L = l) = P(Y^a = 1|A = 0, L = l) \text{ or } Y^a \perp\!\!\!\perp A|L = l] \text{ for all levels of L}$$

If this can be verified, the experiment is conditionally exchangeable. To compute average causal effect in a conditional randomized experiment, there are two methods:

- Stratification - compute the average causal effect in each strata of the population
- Average Causal Effect - compute the average causal effect in the entire population, this method is used if it is not expected that there will be information on L in the future

2.5 Standardization

Standardization is a technique used to help compute the causal risk ratio, $P(Y^{a=1} = 1)/P(^{a=0}= 1)$. The numerator is the probability if all individuals in the population had been treated. The numerator $P(Y = 1|L = l, A = 1)P(L = l)$ for all L=l of the causal risk ratio is the standardized risk in the treated using the population as the standard to measure against. If we conducted an experiment with 20 individuals in the study and 8 individuals do not have a condition (L=0) while 12 have a condition (L=1), and the following information is known:

- The risk is $\frac{1}{4}$ of the outcome occurring in the sample of 8 individuals with L=0
- The risk is $\frac{2}{3}$ of the outcome occurring in the sample of 12 individuals with L=1
- The 8 individuals is 40% of the population (8/20)

	L	A	Y
Rheia	0	0	0
Kronos	0	0	1
Demeter	0	0	0
Hades	0	0	0
Hestia	0	1	0
Poseidon	0	1	0
Hera	0	1	0
Zeus	0	1	1
Artemis	1	0	1
Apollo	1	0	1
Leto	1	0	0
Ares	1	1	1
Athena	1	1	1
Hephaestus	1	1	1
Aphrodite	1	1	1
Cyclope	1	1	1
Persephone	1	1	1
Hermes	1	1	0
Hebe	1	1	0
Dionysus	1	1	0

Table 3: Study with Conditional Randomization

- The 12 individuals is 60% of the population (12/20)

If all 20 individuals in the population was treated, it will be a weighted average calculated as:

$$\frac{1}{4} \times 0.4 + \frac{2}{3} \times 0.60 = 0.5$$

3 Causal Graphical Models

When conducting causal inference, the most common way to visually depict them is called DAGs (Directed Acyclic Graphs). Researchers uses DAGs to predict independencies within the data. DAGs are not quantitative and contain less information compared to using models or mathematical notation as noted in Pearl's book. However, since information about causal relationships is qualitative, DAGs provide an intuitive understanding of causality and the information is more digestable using graphical notation.

3.1 Common Structures of DAGs

This section reviews the three basic DAG structures with their independencies.

3.1.1 DAGs - Chain

Figure 1 is a structure of a chain, which occurs when there is a sequence of three nodes and two edges and one edge is directed into and the other edge is directed out of the middle node. A chain has the following independencies:

- C and B are likely dependent
- A and B are likely dependent
- C and A are likely dependent
- C and A are likely dependent, conditional on B

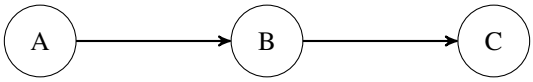

Figure 1: Structure of a Chain

3.1.2 DAGs - Fork

Figure 2 is a structure of a fork, which occurs when the middle node is the mutual parent of two other nodes. A fork has the following independencies:

- B and A are likely dependent
- B and C are likely dependent
- A and C are likely dependent
- A and C are likely independent, conditional on B

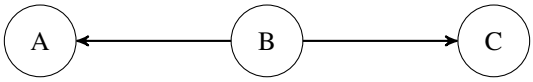

Figure 2: Structure of a Fork

3.1.3 DAGs - Collider

Figure 3 is a structure of a collider, which occurs when one node receives edges from the other two nodes. A collider has the following independencies:

- B and A are likely dependent
- B and C are likely dependent
- A and C are independent
- A and C are likely dependent, conditional on B

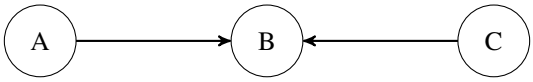

Figure 3: Structure of a Collider

3.1.4 DAGs - Conditional Independence

The three structures of DAGs provide insight into the dependencies and a quicker interpretation of the variables and their relationships to other variables. To highlight conditional independence, let's say our DAG is Figure 4, where A is aspirin and Y is heart disease. However, suppose new information is provided and the the DAG is actually Figure 5, where B is platelet, and it is known that A reduces platelets. To test whether there is association between A and Y, or equivalently, when there is information on B, does information on A help predict Y? To solve this question, we can condition on variable B, that is, restrict the study to observe individuals with low platelet (B=0). It does not matter if B is treated or not treated, because this individual will already has low platelets. In this study, by conditioning on B, we observe that A is not associated to Y.

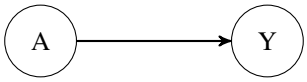

Figure 4: Example

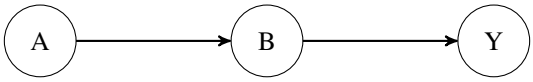

Figure 5: Example of conditional Independence

3.1.5 D-separation

A path, p, is blocked by a set of nodes Z iff

- p contains a chain of nodes $A \rightarrow B \rightarrow C$ or a fork $A \leftarrow B \rightarrow C$ such that the middle node, B, is in SZ (i.e. B is conditioned on), or
- p contains a collider $A \rightarrow B \leftarrow C$ such that the collision node, B, is not in Z, and no descendants of B is in Z

If Z blocks every path between two nodes, X and Y, then X and Y are d-separated, conditional on Z and thus are independent conditional on Z.

D-separation is an important concept because it helps to determine in any graph of any complexity, whether the nodes are d-connected (directionally-connected), which means there exists at least one unblocked path between the two nodes, or d-separated (every path between the two nodes is blocked). By understanding the relationship between two nodes, we can perform conditional independence tests as shown in section 3.1.4.

3.1.6 Discussion on DAGs

The three stuctures introduced in 3.1.2 to 3.1.4 are the foundations of causal graphs and these provide details on the variables and their relationship to each other. DAGs are helpful in providing visualization to the relationships between the variables and are easy to interpret quickly.

3.2 Causal Models - Effect Modification

This section introduces the concept of Effect Modification. Hernán and Robins defines that M is an effect modifier of A (the treatment) on Y (the outcome) when the average causal effect of A on Y varies across levels of M. This important concept says that the treatment's impact on the outcome is different for different levels of M. For example, if the study is weather's impact on ice cream sales and it is noticed that females tend to buy more ice cream when temperatures are greater than 25°C compared to males, then gender is as effect modifier.

Effect modification is an important concept because if there exists an effect modifier factor which impacts the outcome, then the average causal effect will differ between the populations with different prevalence of the effect modifier. For example, it was calculated that the probability of females purchasing ice cream when temperatures are greater than 25°C is $P(Y^{a=1} = 1|M = Females) = 6/10$ and the probability of females purchasing ice cream when temperatures are less than 25°C is $P(Y^{a=1} = 1|M = Females) = 4/10$ then the causal risk ratio for females is 0.6/0.4=1.5. If we did the same calculations and found that the male causal risk ratio was 2/3, this would illustrate that on average, females tends to purchase ice cream in warmer weather compared to males. If we had calculated the causal ratio by grouping females and males together, there would be no impact shown (due to the impact cancelling out).

To calculate the average causal effect, we need to adjust for effect modification by variables M. Stratification is a method to achieve this adjustment. Going back to the ice cream example, to stratify the results, we would compute the effect measure (i.e. causal effect) for each subset - one restricted to female results and the other restricted to male results. Each of the stratum measures the average causal effect in a non-overlapping subset of the population. Another method to calibrate effect modifiers is through IP weighting.

3.3 Causal Models - Intervention

When there are two or more treatments, it is important to understand the interaction between the treatments and the implications it has on the causal effect. Joint interactions are cases when there are two or more treatment intervention. For example, in a study where customers are assigned to receive an offer (E=1) or no offer (E=0) before being assigned to receive a follow-up call (A=1) or no follow-up call (A=0). The following 4 treatment groups are:

- Offer-Call (E=1,A=1)
- Offer-No Call (E=1, A=0)
- No Offer-Call (E=0, A=1)
- No Offer-No Call (E=0, A=0)

Treatment interaction is an interesting concept because it produces several combinations of counterfactual responses, known as response type. When there is one treatment, the response type is displayed in Table 4. The explanation of the response types are:

- doomed: will develop the outcome regardless of what treatment is received
- helped: will develop the outcome only if untreated
- immune: will not develop the outcome regardless of what treatment they received
- hurt: will develop the outcome only if treated

Type	$Y^{a=0}$	$Y^{a=1}$
Doomed	1	1
Helped	1	0
Hurt	0	1
Immune	0	0

Table 4: Interaction Counterfactual Responses

When there are two dichotomous treatments, there are 16 response types. Hernán MA and Robins JM explains the response types in chapter 5. When there are additive interactions between the two dichotomous treatments (A & E), for some individuals in the population, the value of their counterfactual outcome under A=a cannot be determined without knowing E, and vice versa.

4 Causality in Observational Studies

As mentioned in the introduction, it helps to confirm whether the observational data is randomized to infer causality. An observational study can be considered conditionally randomized if the 3 conditions hold true:

- **Consistency:** Consistency holds if the observed outcome for every treated individual equals their outcome if they had received treatment and the observed outcome for every untreated individual equals her outcome if she had remains untreated
- **Exchangeability:** In an randomized experiment, exchangeability holds if the treated, had they remained untreated, would have experienced the same average outcome as the untreated did
- **Positivity:** The probability of being assigned to each treatment level must be positive

With observational studies, the data is already provided and we do not have information on whether the study is conditionally randomized which makes it difficult to prove causality. This report addresses the impact of confounders, selection bias, and measurement error when proving causality in observational studies.

4.1 Confounding

Confounding is the bias that occurs when there is a cause/variable that is shared by treatment and outcome. The simple causal diagram of confounding is shown in Figure 6 where A is treatment, Y is outcome and L is the shared cause. This depicts an association between treatment and outcome ($A \rightarrow Y$) and the backdoor path ($A \leftarrow L \rightarrow C$). Since L exists, we cannot say that the entire association of Y is caused by A.

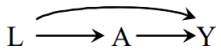

Figure 6: Confounding DAG Example

Hernán MA and Robins JM presents several other examples of confounding in Chapter 7 which has the same DAG - a presence of L that is shared by the treatment and outcome which results in a backdoor path. As a result of confounding, these are the two cases when causal effects can be identified:

- No common causes of treatment and outcome
- All backdoor paths are blocked/no unmeasured confounding

In Figure 6 the backdoor path can be blocked by conditioning on L. Thus, if the study contained data on L for all participants of the study, there is no unmeasured confounding. L is a confounder because it is needed to remove the confounding. To control for confounding, randomization is the preferred method since it randomly assigns treatment to the participants without any conditions.

Note that subject-matter knowledge of the study is needed in order to identify and measure the variables required. Causal inference assumes that we have used our knowledge to identify and measure variables that are sufficient to adjust for confounders. To adjust for the confounders, standardization and IP weighting can be used to compute the causal effect of the study.

4.2 Common Biases - Selection Bias

Selection bias occurs when there is an association with the process that individuals are selected to participate in the study and this bias can occur in randomized and observational studies. Figure 7 shows the general structure of selection bias where C is the common effect, Y is the outcome and A is the treatment.

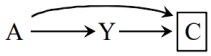

Figure 7: General Selection Bias DAG

When selection bias exists, association does not equal causation. From Figure 7, when C is conditioned on (hence the box around C), there is an open path between A and Y. There are several examples of selection bias that is demonstrated in the textbook, a few of these are listed below:

- Self-selection bias = when the study is restricted to those who volunteer to participate
- Missing data bias = individuals could have missing data because they are reluctant to provide information or because they miss study visits

To adjust for selection bias, it can be mitigated through a better experimental design but often it is unavoidable. Sometimes selection bias can be corrected by IP weighting or standardization (as mentioned in an earlier section 2.5).

4.3 Common Biases - Measurement Bias

Measurement bias exists when there is an error in the measurement of study. Measurement bias can occur on the treatment and outcome variable. Commonly, the causal diagram for measurement biases are similar to Figure 8 where * represents that measured outcome. For example, A* represents the measured treatment variable and U_A and U_Y represents the measurement error for A and Y respectively.

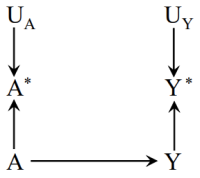

Figure 8: Common Measurement Bias DAG

In Figure 8, it can be seen that there is no backdoor path, as a result, association is causation in this setting (if no other biases exist). Generally, to correct for measurement bias, its a mixture of validating the samples and modelling assumptions. In the textbook, it mentions that the best method to reduce measurement biases would be to improve the measurement procedures.

4.4 Instrumental Variables

So far, it's been assumed that all variables needed to adjust for confounding and selection bias can be identified and correctly measured. However, if that assumption cannot be made, there would be residual bias in the causal estimates. For an instrumental variable, Z, to exist, it must meet the following conditions:

- Z is associated to A
- Z does not affect Y, except through its potential effect on A
- Z and y do not share causes
- Effect homogeneity - there are different aspects of this, one extreme version needs that A's treatment on Y has to constant across individuals.

It is important to note that condition 4 is difficult to show within the experiment, however, two approaches have been developed to bypass this condition. The first approach is to have baseline covariates in the model for instrument variable estimation. The other approach is to have an alternative condition that does not need effect homogeneity. Monotonicity is the case when effect homogeneity is not required. Monotonicity occurs when no defiers exist in the experiment, however it requires that Z was randomly assigned in the experiment.

In the next two sections, we will go over examples of applications of causality and how concept learned in this course have been applied to those studies.

5 Applications of Causality - Coffee Bean Experiment

5.1 Introduction

FreshBooks is a software company that provides their employees with coffee brewed from Starbucks coffee beans. Despite providing free coffee, they noticed that some employees were buying coffee outside of the office. These employees would go to Hale, a local Canadian coffee roaster, to purchase their coffee. Since the cost to provide Hale coffee would be similar to the cost of Starbucks coffee, Freshbooks was interested in conducting a taste test to confirm which coffee their employees would prefer.

The purpose of the experiment is to show if there exists an average causal effect of coffee beans on employee's preference in their coffee. In this experiment, the type of coffee beans is the treatment variable and the outcome variable is the preference of Hale or Starbucks. In the next section, the experimental design, the data collection method and the other variables to be considered are thoroughly discussed.

5.2 Data & Study Design

The experiment was conducted between 9am and noon for one day. All employees were invited to participate. There was enough capacity for 60 people to participate. There were 47 employees who participated out of 309 employees who were invited. In the experiment, the following characteristics of the participant and study were collected:

- Gender of participant
- Do they generally take their coffee black
- Coffee Brewing Method (Drip, French Press, Pour over)
- Coffee Bean (Hale or Starbucks)

Below is the sample data of the participants in Table 5.

	gender	usuallytakecreamer	coffee1	coffee2	cei1	cei2	outcome	binary_outcome
1	M	1	Starbucks	Hale	7.0	4.0	-3.0	0
2	F	0	Hale	Starbucks	3.0	4.0	-1.0	0
3	M	0	Hale	Starbucks	7.0	5.0	2.0	1
4	M	0	Hale	Starbucks	6.5	8.0	-1.5	0
5	M	1	Hale	Starbucks	4.0	6.0	-2.0	0
6	M	0	Starbucks	Hale	6.0	2.0	-4.0	0
7	F	0	Starbucks	Hale	6.5	7.0	0.5	1
8	M	1	Hale	Starbucks	7.0	5.0	2.0	1
9	F	1	Hale	Starbucks	7.0	1.0	6.0	1
10	M	1	Starbucks	Hale	8.0	6.0	-2.0	0
11	F	1	Hale	Starbucks	2.0	4.0	-2.0	0
12	F	1	Hale	Starbucks	7.0	1.0	6.0	1
13	F	0	Hale	Starbucks	8.0	9.0	-1.0	0
14	F	0	Starbucks	Hale	5.0	4.0	-1.0	0
15	M	0	Hale	Starbucks	3.5	6.5	-3.0	0
16	M	0	Starbucks	Hale	1.5	5.0	3.5	1
17	M	0	Starbucks	Hale	2.0	6.0	4.0	1
18	F	0	Hale	Starbucks	8.5	2.0	6.5	1

Table 5: Coffee Experiment Data

In the study, these were the following factors were controlled for:

- Participants can only take their coffee black

- The coffee beans would be compared through the drip brewing method
- Participants would clean their mouth using water during the study
- Participants have a 50% chance of getting hale coffee first (and 50% chance of getting Starbucks first) which is determined themselves, by drawing out of a hat
- Participants were asked to refrain from talking to each other (to not be influenced by others' opinions)
- Participants were told that the rating scale (out of 10) is 1 for disgusting, 5 is average, and 10 is for the best coffee

There were only a few participants who were randomly selected to taste test the french press and pour over brewing methods, since the office coffee machines are drip coffee, only the drip brewing methods is included in the study. As a result, after removing participants who did not receive drip coffee, there were 18 participants in the study remaining.

5.3 Analysis

To understand whether there were any impacts on the outcome, an exploratory analysis is conducted to test the impact of gender and whether the participant generally drinks coffee without cream and sugar. Notice of the coffee taste testing was sent out 3 days in advance. We realized that there is a selection bias as only people who did not have meetings or were not away on vacation, were able to attend. In addition, certain teams had more members who joined the study than others.

The average score per participant is shown in Table 6. The average score of Hale was 5.42 whereas the score for Starbucks was 4.86. However, when assessing the binary scores, it was shown that there were more participants that enjoyed Starbucks over Hale coffee beans (10 preferred Starbucks over Hale, 56% or 10 out of 18 participants). This is shown in the bar graph in Figure 9 where the 0 represents Starbucks being favoured and the 1 represents Hale being favoured. The reason for this discrepancy is because there were participants that extremely favoured Hale beans who also provided an overall higher rating.

Coffee Beans	Hale	Starbucks
Average Score (out of 10)	5.42	4.83

Table 6: Average Scores of Coffee

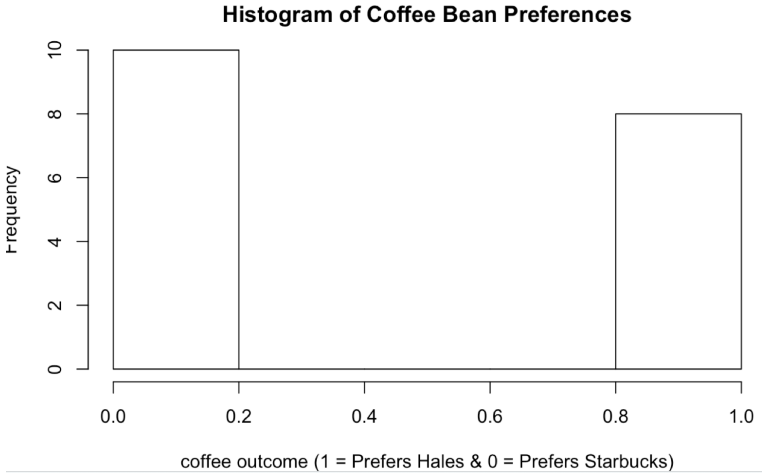

Figure 9: Comparing Coffee Preferences

It is also important to understand whether there are any effect modifiers and confounders that should be adjusted in the experiment. Gender is an effect modifier since females are known to be better at recognizing tastes compared to males according to the study conducted by the University of Copanhagen. Table 7 shows the differences in senses for females vs males and in Table 8, it shows the impact of coffee drinkers with additives (such as cream/milk/sweetener/sugar).

Coffee Beans	Hale	Starbucks	Count
Females	5.81	4.06	8
Males	5.10	5.50	10

Table 7: Gender as an Effect Modifier

Coffee Beans	Hale	Starbucks	Count
No Additives Regularly	5.50	5.05	11
Additives Regularly	5.29	4.57	7

Table 8: Additives as an Effect Modifier

5.4 Results

Table 7 shows that females tend to have a preference for Hale coffee beans (average score of 5.81) whereas males have a preference for Starbucks coffee beans (average score of 5.50). In Table 8, it appears that additives did not have not impact participants' results. Regardless of additives, participants tend to enjoy Hale coffee beans over Starbucks. That being said, when evaluating the results by popularity, generally the consensus was Starbucks. The conclusion is that based on popularity, Starbucks was favoured, but when based on the average rating scale (out of 10), Hale was preferred.

5.5 Discussion

After the experiment was over, we reflected on the study and discussed that:

- There were a few participants who did not regularly drink coffee but wanted to participate in the experiment and we cannot turn these participants away (these people were not the appropriate sample since regular coffee drinkers would be affected by the results of the experiment)
- Participants had to drink black coffee which they were not used to doing (if they added cream/sugar), this was adjusted for as shown in Table 8)
- Some participants may have lingering coffee in their month from the previous test (however, we provided water to wash away any liquids and it was random whether they received Starbucks first or Hale first)
- Randomization of the coffee prevented participants from easily knowing the sequence of the experiment
- We had capacity to allow 60 participants (in 6 groups of 10 participants) and asked earlier participants to refrain from chatting with others on the test, however it was difficult to monitor that participants did not talk to each other on the results
- Although participants were told not to chat with each other, it would have also been effective to randomize the seating of the participants too

6 Applications of Causality - Offers & Promotions Experiment

6.1 Introduction

Company A is an organization that conducts experiments to understand the effectiveness of offers and discounts on their users. For privacy, the name of the organization will not be disclosed (referred to as Company A in this report) and only the results of the tests are available in the report (The methodology cannot be provided and the results are adjusted however the interpretation of the results are the same). Company A's product is a software and generally their users pay monthly. The product has multiple functionalities, so to help new customers learn how to use the product, Company A offers a 30-day free trial.

When the new user sign up for an account with Company A, they provide their email. To incentivize customers to sign up earlier in their trial period, customers are enticed with an offer. The offer discount is 10% for the first three months of the user's subscription plan. This experiment tested two treatment variables:

- Does offering a discount increase the number of customers that sign up?
- Does offering a discount earlier in their trial (Day 1 vs Day 25) make a difference in the number of customers that sign up?

The coffee experiment was designed such that each participant of the study were able to taste test both coffee beans. Generally, all treatment cases cannot be provided to the participant similar to this study. Users either receive an offer or does not. The main question is whether an offer causes the user to sign up for a paying subscription of the product.

6.2 Data & Study Design

In this study, participants were randomly placed into a Test or Control group. The Test group receives an offer of a 10% discount on their subscription plan for 3 months, whereas the Control group does not receive any offer. This experiment was held over a few months. In the experiment the following characteristics of the participant and study collected were:

- Basic information: name, email, phone, postal code, region
- If the user was randomly assigned into the Control vs. Test group
- If the user upgraded
- What plan did the user upgrade to

To test the duration and the offer, there are three groups for this experiment:

1. Receives an offer on day 1
2. Receives an offer on day 25
3. Receives no offer

This experiment will compare offer on day 1 to day 25, offer on day 1 to no offer, and offer on day 25 to no offer. Note that there are over 100,000 samples and we will reach statistical significance.

A sample of the dataset is show in Table 9.

Identifier	Group	Offer	Upgrade	Days until upgrade	Region
1	Control	0	1	30	US
2	Control	0	1	30	Canada
3	Control	0	1	27	Mexico
4	Control	0	1	25	London
5	Control	0	1	50	US
6	Control	0	0	-	US
7	Control	0	0	-	US
8	Control	0	0	-	Australia
9	Control	0	0	-	Australia
10	Control	0	0	-	US
11	Test	1	1	1	US
12	Test	1	1	1	US
13	Test	1	1	1	US
14	Test	1	1	1	US
15	Test	1	1	25	US
16	Test	1	1	25	US
17	Test	1	1	25	Australia
18	Test	1	1	25	Australia
19	Test	1	0	-	London
20	Test	1	0	-	Canada

Table 9: Offers & Promotions Experiment Data

6.3 Analysis

Although the actual numbers cannot be presented, the results below mimic the results of the actual experiment. To eliminate the biases, we worked with the project manager to gain context on the experiment. We realized that there were certain variables to control in order to have an unbiased randomized test. Our results in the beginning is depicted in Table 10, where the results show that giving an offer earlier incentivized users to upgrade earlier (compared to offer on day 25 and no offer). However, we realized that differ countries have different upgrade rates because the product is English only, we realized that an offer would not be as effective from a non-English speaking country. The results were then broken down in the regions, as shown in Table 11 where we do see a larger discrepancy in the upgrade rates between offer on day 1 compared to offer on day 25.

Group	Upgrade Rate	Average Days Until Upgrade
Offer on Day 1	75.40%	20.83
Offer on Day 25	69.80%	28.09
No Offer	54.78%	36.94

Table 10: Upgrade Rate & Average Days Until Upgrade

Country	Group	Upgrade Rate	Average Days Until Upgrade	Proportion
English Speaking Countries	Offer on Day 1	76.90%	26.09	20%
English Speaking Countries	Offer on Day 25	70.30%	28.09	20%
English Speaking Countries	No Offer	52.60%	36.94	20%
Other Countries	Offer on Day 1	72.00%	24.5	15%
Other Countries	Offer on Day 25	65.40%	30.1	15%
Other Countries	No Offer	51.80%	52.4	10%

Table 11: Upgrade Rate & Average Days Until Upgrade

In order to compare offers on day 1 vs. day 25, standardization was applied to adjust for the proportion of English speaking and non-English speaking countries. The general expression applied is below, where L is the factor being standardized across, Y is the outcome and A is the treatment.

$$\text{Standardization Formula} = \sum_l E[Y \mid L = l, A = a] \times P(L = l)$$

Country	Group	Proportion	Standardized Upgrade Rate
English Speaking Countries	Offer on Day 1	57%	43.9%
English Speaking Countries	Offer on Day 25	57%	40.2%
Other Countries	Offer on Day 1	43%	30.9%
Other Countries	Offer on Day 25	43%	28.0%

Table 12: Upgrade Rate & Average Days Until Upgrade Partitioned by Country

6.4 Results

From the analysis, we learned that customers respond better to offers earlier in their trial, with an upgrade rate of 74.8% as shown in Table 13 compared to an offer sent on Day 25 of trial. In addition, the analysis of the study shows that Offers do incentivize customers to upgrade. Table 14 displays the difference in upgrade rate between a customer with an offer on Day 1 vs. without an offer. The difference in upgrade rate is large.

Group	Total Upgrade Rate
Offer on Day 1	74.8%
Offer on Day 25	68.2%

Table 13: Offer Day 1 vs Day 25 Upgrade Rate Standardized by Country

Group	Total Upgrade Rate
Offer on Day 1	74.5%
No Offer	52.3%

Table 14: Offer vs. No Offer Upgrade Rate Standardized by Country

6.5 Discussion

After the assessment of offers, there were great findings which were learned about the customers and it also allowed for more tests to be conducted in the future. There are several additional factors that we are interested in testing:

- What is the optimal offer duration (3 months, 6 months, etc.)
- Should there be a shorter offer expiry (currently the offer is available during the 30-day free trial, what if the expiry is before the trial?)
- What is the optimal magnitude of discount (what is the optimal discount that incentivizes users)
- Should we have different offers based on different subscription plans

The post mordem of the offers experiment allowed us to think about about could have been improved on the experiment:

- There is another campaign running during the same period as the offer being provided, it is difficult to control for other campaigns
- Some participants who were not suppose to receive an offer heard from their peers about the offer and demanded an offer - Company A could not refuse this

7 Reflections on Causal Inference

After the readings on causal inference, it provided me with new approaches on conducting causality testing. By learning about the concepts that have an impact on causality (e.g. confounders, selection biases, effect modifiers), it provides awareness of biases when analyzing the results of the experiment and the adjustments that need to be made for these biases. That being said, these methods are not fool-proof and require critical thinking of any assumptions made. Ideally, when we try to prove causality, it would be best if there was a randomized controlled experiment and only the treatment variable varies. Generally, it is difficult for this ideal situation to occur due to the nature of the study or limited resources available for testing.

Since the coffee and promotions experiments were conducted during this report, it was not possible to modify the experiment. Reflecting upon this course, in the future it would be ideal to have an instrument variable which is used when the treatment variable cannot be used. For the coffee experiment, an instrument variable might be the number of Starbucks and Hale cups of coffee does the participant buy per month. For the offers experiment, an instrument variable could be historical prices of similar products purchased.

Another important assumption that was mentioned in the confounding section of this report is the importance of obtaining context and knowledge of the study before performing the analysis. Without context, the biases would not have surfaced and experiments may need to be restudied without proper knowledge of the study being performed.

In conclusion, I thought that this course provided structure on testing and validating causality within experimental and observational studies. This reading course was informative and I plan on doing more readings in the future on causality with time series.

8 References

- Hernan M. A., & Robins J.M. (February 15, 2010). Causal Inference. London: Chapman & Hall
- Pearl P., Glymour M., Jewell N. P. (March 7, 2016). Causal Inference in Statistics: A Primer. Cambridge, United Kingdom: Cambridge University Press
- University of Copenhagen. (December 18, 2008). Girls Have Superior Sense of Test to Boys. Retrieved from https://www.sciencedaily.com/releases/2008/12/081216104035.htm
- Wang, Linbo. (August 5, 2019). Foundations of Causal Inference. [PowerPoint Presentation]. Toronto