1190 Sample Exam - practice questions and notes for exam

Data, Insights and Decisions (University of New South Wales)

Sample exam

QUESTION 1 30 MARKS

You have been brought in as a Data Science consultant on a court case. A chemical company has been found negligent after a chemical spill at one of their plants. All that remains in the court case is to decide on the extent of the damages for which the company is liable. One way the court has been deciding on this amount is to look at the impact the spill has had on the value of houses located near to the chemical plant where the spill occurred.

As the expert witness, you have been asked to evaluate some alternative strategies to estimate the impact on housing prices (price). Strategy A involves taking a sample of sales that occurred after the spill where the houses are classified as either being close to the plant or not. This feature was designated by a variable near that was equal to 1 if the house was deemed to be close to the chemical plant and zero otherwise. Then a regression analysis is performed using the following model (MA):

$$\square\square: \square\square\square\square\square\square = \square 0 + \square 1\square\square\square\square\square + \square\square.$$

Strategy B involves taking a sample of sales for houses near to the plant but where some sales occurred before the spill and some after. The variable after is equal to 1 if the house was sold after the spill and zero if the sale was before. Then a regression analysis is performed using the following model (MB):

$$\square\square: \square\square\square\square\square\square = \square 0 + \square 1\square\square\square\square\square + \square\square .$$

Part A. (10 Marks) Explain A and B as strategies to estimate the impact of the chemical spill and critically evaluate each of them. Is either preferable to the other?

Both regressions have a design with a structure like an experiment with a treatment and control group. This results in estimates that can be interpreted as differences in means.

Strategy A estimates the difference in means for housing near compared to houses away from the spill.

- The primary flaw in that we cannot determine if other factors contribute to price differences based on distance to the plant.
- If the chemical plant was built in a commercial area that is unattractive to homeowners, differences found using this strategy may reflect differences in the attractiveness of the near/away areas that preceded the oil spill and hence are not directly attributable to the spill.
- There is an additional concern about the representativeness of houses for sale after the spill, a sample selection problem.

Strategy B is a before and after comparison but confined to houses near the chemical plant.

- We avoid the previous problem of comparing near and away houses by using houses that were near, making them more comparable.
- The primary flaw in this strategy is a confoundment problem, which now happens over time. For example, housing market dynamics may create broad changes in the price of all houses.
- There again is a selection problem relating to the representativeness of houses put up for sale before the spill compared to those after the spill. For example, if those before were broadly representative but only poorer houses were put on the market after the spill then this would induce a downward bias of the estimated impact of the spill.

You would have to argue about the relative size of these biases so in general it is not obvious that one strategy is better than the other.

Part B. (5 Marks) Suggest an alternative regression model that is preferable to MA given that you only have data from after the spill. Does this address all your criticisms of Strategy A that you outlined in part (a)?

Adding covariates to the regression model reflecting house characteristics would help. This would control for observable differences between near and away houses that were put up for sale.

This could also help with the selection problem to the extent that observable differences explain whether houses were put up for sale or not.

Conversely if the reason the houses were put up for sale are unobservable to the analyst, then the omitted variable and selection problems remain.

Part C. (10 Marks) Using housing data models MA and MB are estimated, and the results given below. How do you interpret these results? (Note that $price$ is expressed in $1000)

$$\widehat{price} = \underset{(4.0)}{131.9} - \underset{(7.6)}{40.0}near$$
$$n = 142,\ R^2 = .165,\ standard\ errors\ in\ (.)$$

$$\widehat{price} = \underset{(5.9)}{63.7} + \underset{(9.1)}{28.3}after$$
$$n = 96,\ R^2 = .094,\ standard\ errors\ in\ (.)$$

Regression 1:

- The first regression implements Strategy A.
- As expected, houses near the chemical plant sell for prices less than those not near the plant. The estimated difference in the mean price of houses is $40,000.
- This is a large difference, implying houses in the affected areas sell for 30% less than those houses sold in the unaffected area. [30 ≅ (40/131.9) × 100]
- This difference is precisely estimated as the 95% CI of 40 ± 1.96 × 7.6 or [25.1, 54.9] does not include zero.

Regression 2:

- The second regression implements Strategy B. Here houses in the affected area sell for prices $28,300 more than those sold before the spill.
- This result is not consistent with a detrimental impact of the spill, but it is consistent with the concerns expressed in (a) that there may be biases here due to general upward movement in house prices.
- This is a large difference, implying houses sold after the spill sell for 44% more than those sold before. [44 ≅ (28.3/63.7) × 100]
- This difference is precisely estimated as the 95% CI of 28.3 ± 1.96 × 9.1 or [10.5, 46.1] does not include zero.

Part D. (5 Marks) Suppose you have sales both near and not near to the plant as well as sales before and after the spill. Suggest an alternative strategy to estimate the effect of the oil spill on housing prices that is preferable to both MA and MB?

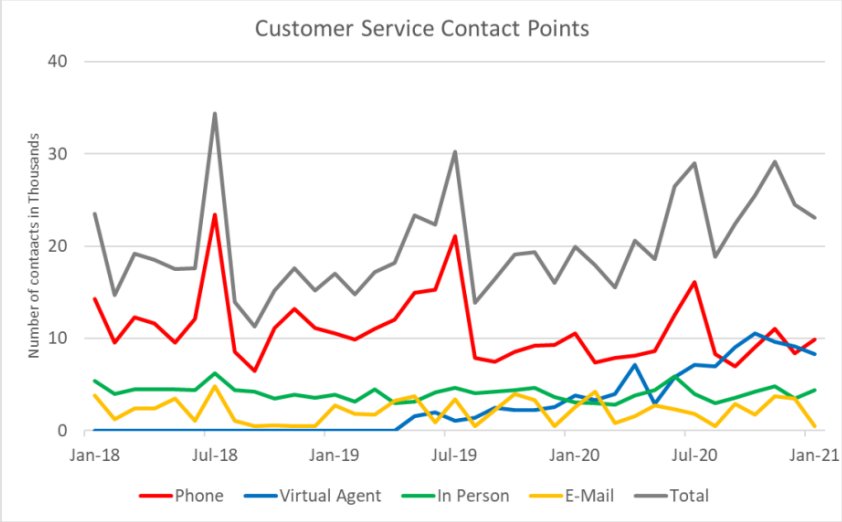
A difference in differences (DiD) approach is an appropriate way to proceed when data are available on houses sold before and after the spill and near and away.

Essentially calculate the difference in average near and away house sale prices before the spill and compare it to the difference in average near and away house prices after the spill.The resulting difference in difference estimate then helps control for many of the concerns in Strategy A and B.

Controlling for time effects and household characteristics would also help.

QUESTION 2 30 MARKS

Imagine you work for a large department store, which highly values customer service. The following chart shows how customers contact the customer service centres.

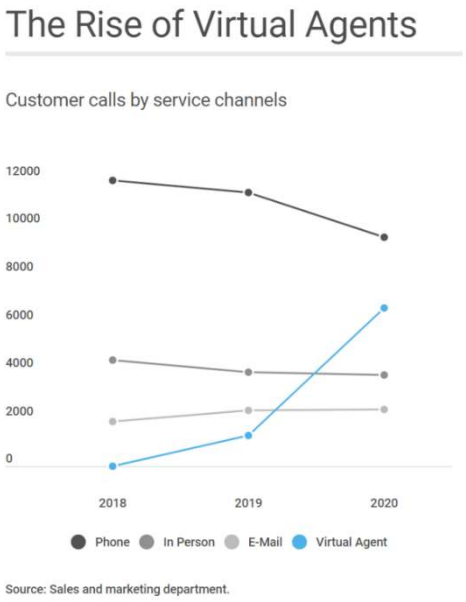


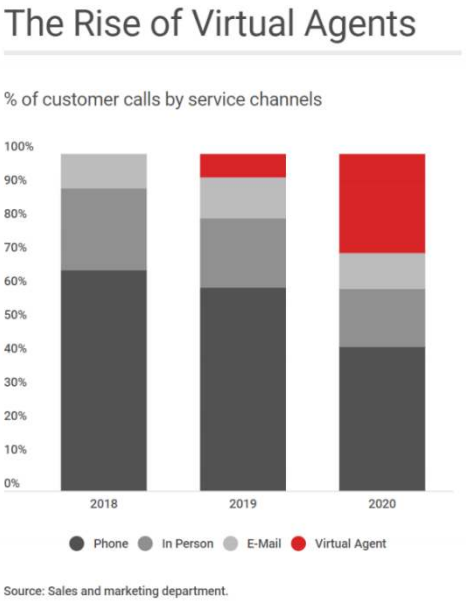
You begin to discuss the chart with your manager. Immediately, she has the following queries: "I want to see the overall trends, but it is difficult to see with all the seasonal spikes in the time series. I'd like a simpler view into the trend." You decide to create some charts to address your manager's queries.

Part A. (5 Marks) Using the two-questions four frameworks typology, identify the type of chart you would use to address the query and explain why.

This is a declarative and data-driven chart. The idea being presented is the trends, so this is data-driven. In addition, you are not exploring trends, you are just presenting a simple view of the trends, so this is declarative. Thus, the type of chart needed is an everyday data viz.

Part B. (15 Marks) Sketch two alternative charts for the query. For each chart, provide a brief explanation of your design choices. To sketch the chart, you can use any tool you want (e.g., you can use a software tool like infogram, excel, or R). Alternatively, you can sketch the chart using pencils, pens or markers on paper, then take a picture of the charts and paste them into your solutions document. You can access the underlying data "customer_service.xlsx" on Ed.





Part C. (10 Marks) Evaluate your two charts and explain which you would select to further develop to present to your manager.

The line chart seems more appropriate as it better illustrates the upwards trend of the virtual agents and the decrease of the phone (whereas e-mails and in persons are fairly consistent).

The stacked bar chart shows that virtual agents are a growing aspect of the business, proportionally, but it is more difficult to see the clear trends that are visible on the line graph.

To further develop the line chart, I would explore adding future trend line, potentially with scenarios, since this would even better highlight the trends of growth in virtual agents. In addition, I would consider moving the channels to be the graph next to the lines to limit eye travel.

When presenting the data, I would consider presenting two versions of the graph. The first would be without virtual agents. I would make the title "modest changes in customer service…" and make the comment that this does not include new technologies for handling customer service. After allowing the manager to digest that information, I would then present a version of the graph above with virtual agents accentuated in blue (greying out the other channels).

QUESTION 3 20 MARKS

Recall the Data Analytics Simulation: Strategic Decision Making that you played in Week 1's Workshop. The simulation provided first-hand experience in the benefits and challenges of making data-driven decisions.

1. **Reflect on how the data visualisations, dashboards, and filters helped you make decisions and the challenges you experienced while playing the game.**
2. **Based on your experience, identify organisational benefits and challenges from using data and modern visualization to make decisions.**

The critical reflection should include lessons learned during the game, with a focus on the value and challenges of working with data.

The dashboards, visualisations, and filters enable data driven decisions. For example, line charts quickly facilitate identifying trends (e.g., product popularity and competitor pricing patterns), which in turn facilitate discussion regarding strategy. The filters provided further functionality by providing answers for questions regarding specific market segments, which helped create a strategy for Blue. The filters also had the benefit of streamline the overload of data available through the various dashboards. After selecting a strategy, as the data was aggregated to an annual level, it was difficult to determine if increased profits resulted by chance or good decision making. Also, positive performance reinforced the acceptance and use of our strategy, which means that we would miss potential new opportunities to do even better resulting from new data and other market trends. Compounding this issue is that using our filters, it was easy to miss new developments occurring outside our area of focus. Unfortunately, we did not notice these omissions until after performance would fall. Thus, our responses were often reactionary rather than proactive, undermining a key advantage of data driven decisions.

For organisations, data driven decisions enables greater confidence in the actions, since logic and evidence underpin critical decisions. Using data as evidence can also help make arguments (e.g., for financing or resources) more compelling to decision makers. Moreover, being data-driven can help overcome biases and preconceptions. I personally know a lot of people who prefer liquid detergent to pods, and initially gravitated towards that product form. However, a quick inspection of the data showed that this hypothesis was false: Pods were more appealing to our target market. Our group found the volume of data and array of filters challenging; however, this challenge is likely to magnified for organisations. For example, each of our team members initially explored the data separately and came to different conclusions on the best strategies. For large teams in organisations, with even more data and analysis approaches could lead to debate and opinions over the best way forward. Although this is a substantial challenge, with proper structures, this type of debate is extremely valuable, as it can ultimately lead to better decision making.

QUESTION 4 20 MARKS

You are a data analyst for AppCo. AppCo produces a smartphone app that allows users to virtually try on clothes. It is funded by having sponsored links to online clothing retailers. AppCo tells its users, "top brands, all sizes, best prices". AppCo has a dashboard that is used by Board members that shows sales, revenues and a comparison of sales by

retailer. You have been asked to consider whether having a selection of the most popular sizes at a slightly lower price would increase revenue. The CEO has said, “well just use some of your magic A/B testing …”.

Part A. (5 Marks) Identify potential legal issues that may arise from A/B testing if AppCo users are unaware the experiment is taking place.

Poor design of an AB testing experiment could lead to issues under consumer law in Australia. For example, if the subjects who receive a service in group B do not receive a service which is fit for purpose, this could expose the business to legal action. There is also an important consideration in respect of AB testing to ensure that it does not breach Australia’s antidiscrimination laws.

Part B. (10 Marks) Evaluate whether A/B testing would lead to any legal consumer issues?

The “all sizes” and “best prices” might be misleading or a misrepresentation if there is no basis for the statement for some AppCo users. A business will be liable for engaging in misleading or deceptive conduct if it makes a statement about the future that later proves to be incorrect, unless the business had reasonable grounds for making the statement. Misrepresentation is associated with advertising where a business claims something about a good or a service which is untrue or misleading. However, the legislation about misrepresentation does not specify that any misrepresentation needs to be in an advertisement. As a result, it is possible for a business to mislead others by making claims which are false or misleading.

Part C. (5 Marks) Recommend steps to ensure that your organisation maintains good governance when developing analytics at AppCo.

One approach is to have a “red team” or “devil’s advocate” when decisions are made within a business. One person or one team advocating for one approach and another team or person advocating for an alternative approach. It means that a great solution to a problem can be synthesised from the advocacy of each of the two groups.