Estimating Effectiveness of Online Geographically-based Advertising Campaigns

IMAN SAID SAIF AL-HASANI

Estimating Effectiveness of Online Geographically-based Advertising Campaigns

Iman Al-Hasani

A Thesis presented for the degree of
Doctor of Philosophy

Statistics and Probability Group
Department of Mathematical Sciences
University of Durham
England

August 2021

Dedicated to

my late father; the light of my life's journey

my mother; the pillar of my strength

my sisters and brothers; the hand when loss

Abstract

The effectiveness of online geographically-based advertising campaigns is estimated experimentally using a randomised experimental approach called a geo-experiment. In these experiments, a region of interest is partitioned into geographical-targeting areas called geos. The experiments are conducted in two distinct time periods where in the first time period there is no difference in advertising campaigns between geos, whereas during the second time period the campaigns for some selected geos are modified. The main concern is, which geos should be assigned to the treatment condition to serve the modified advertising campaigns during the second time period? It is a simple question with a not so simple answer in reality, especially in the presence of unobserved heterogeneity structure within geos. The issue therefore is to design a robust advertising campaigns which permits estimation of the effectiveness of the campaigns using geo-experiments. In this thesis, a conceptual model of geo-experiments is presented to improve our understanding of the potential impact of hidden heterogeneity on estimating the effectiveness of advertising campaigns. A theoretical framework based on theory of maximum likelihood estimation of misspecified model and Kullback-Leibler divergence is developed to study the implications of unobserved heterogeneity for inferences about estimated effects for geo-experiments. An important part of the framework is a proxy model linking the fitted model, with homogeneity structure within geos, and the assumed truth which includes unobserved heterogeneity. The theoretical framework plays a key role in approximating the behaviour of the estimated fitted model parameters. This saves having to do expensive Monte Carlo simulation all the time. The accuracy of the theoretical approximation is investigated for different campaign design strategies across different truth instances. The results reveal the advantage of design strategies based on unobserved covariates, such as social-grades, in reducing the variability of the approximation error and that designs based on spatial proximity may achieve some of the same benefit. Nonetheless, for the more complex truth instances investigated, none of the design strategies considered succeeds in avoiding bias due to unobserved heterogeneity.

Declaration

The work in this thesis is based on research carried out at the Probability and Statistics: Statistics, the Department of Mathematical Sciences, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

My Ph.D. research journey has been filled with hardships, lessons, and special moments that ultimately led me to complete this thesis. This achievement would not have been accomplished without assistance, love, support, and encouragement from several great individuals - my thanks and appreciation to all of them for being part of this long, worthwhile journey and making this piece of work possible.

The first person who deserves my deepest gratitude is my supervisor Peter Craig, who generously shared with me his immense knowledge and expertise. I appreciate his accepting to supervise me when my previous supervisor retired. He succeeded in embracing the topic and creating positive changes in the trajectory of the research. He valued me as an individual and showed me care and compassion to work better. He has given me the freedom to wander intellectually while paying attention to not diverge from the research topic. The door to his office was always open whenever I had a question about my research or ran into technical or administrative troubles. He taught me how to transform my mistakes into skills, allowing me to grow as a research scholar. He was very supportive when I had to return to my job in Oman and continue my research from there. He has been a great mentor to me and his generous guidance has brought me a long way into the computational and applied statistics field.

I would also like to thank Michael Goldstein for his suggestion to search on estimation under misspecification. Thank you also to David Wooff and the digital marketing consulting company for proposing the research topic of measuring the impact of paid search advertising using geo-experiment. Thank you to the company for offering a week' s training to understand the online advertising campaign process and how digital data are gathered from different resources. Thank you to Amin Jamalzadeh at the company for providing various datasets, and sharing his knowledge of digital marketing.

Very special thanks to Sarah Heaps and Jochen Einbeck- my viva examiners, for their very helpful and valuable comments and suggestions. I am also so thankful

to Jochen Einbeck for being so supportive during the difficult period of supervision transformation; and also, for his acceptance and encouragement when I requested to host the Research Students' Conference in Probability and Statistics, RSC 2017. I am grateful to Samuel Jackson for making the conference a huge success. I should also thank all maths office members at Durham University for their help and support.

I am grateful to the people who have contributed to my academic development. Many thanks to Eleanor Loughlin, Thomai Tsiftsi, and James Edwards, for allowing me the opportunity to be a statistics tutor in Maths and Stats Lab, enabling me to test my abilities in providing advice and guidance to students from different backgrounds in the university. A special thanks to Eleanor for her confidence in me, allowing me to present my Ph.D. experience to new postgraduates in the Science faculty induction event, in particular in the 'Reality of Postgraduate Research'. I am very thankful to all tutors in the Maths and Stats Lab, especially since our team won the Above and Beyond Team Award at the Student Employee of the Year Awards.

I would like to thank Sultan Qaboos University (SQU) for providing me with financial support throughout my doctorate scholarship, especially the Statistics department at SQU, for their confidence in me and their support during the writing stage. I should also thank Atsu Dorvlo for proofreading sections of chapters 1, 4, 5, 6 and 7 for spelling and grammatical errors. In addition, thank you to Ronald Wesonga for reviewing the thesis and for his valuable comments.

My deep and sincere gratitude to my mother Fatimah for the unfailing faith, continuous encouragement and loving support in my life, not just my years of study. I am grateful to my sisters and brothers for providing me moral and emotional support along the way. Special thanks to my nephews and my nieces for helping me to break out of my serious mood, providing much laughter and joy along the way. Thanks to my family for always believing in me, even when I didn't believe in myself.

I owe thanks to a number of people for their valuable support and motivation, which drive me to give my best. Anum, Asma, Fatimah, Fatimah Al Ghamdi, Behnaz,

Shatha, Bushra, Dhuha, Faten, Hana, Amal and Manal- thanks to one and all. Special thanks to a little friend Laura for making me laugh and bringing joy always. I am indebted to Sharifa and her daughters who opened their house to me during my last days in Durham. Thanks to everyone who has shown me love and kindness during my stay in Durham and I am sorry if I have missed listing you here.

Thank you to Hugh Kearns for daily inspirational and motivational writing tweets. Thanks to M. Imran for saving my time by preparing thesis latex template in 2001-06-18 with no copyright. Thank you to Hussain, Hermes and Jonathan for valuable discussions and suggestions. Big thanks to a coffee lover Huda Al Amri, for the encouragement and coffee during writing time spent in SQU.

My acknowledgement will never be complete without the special mention of myself! The journey was long, difficult and full of obstacles and I would have never arrived to this stage without the assistance I have received from others and myself. I thank myself for accepting all challenges that have come my way, keeping believing and never giving up! (*Praise be to Allah*)

Contents

List of Figures

List of Tables

Introduction

Geographically-based online advertising campaigns are digital customised and targeted marketing campaigns delivering advertisements to consumers based on their geographic locations. The alternative name for such advertising campaigns is geo-targeting campaigns as they are called in Google Adwords, which is one of the most well-known search engine marketing platforms. The name is derived from the geo-targeting feature in Google Adwords which allows advertisers to show their advertisements to people in certain geographic locations. Geo-targeting campaigns have been considered effective digital geo-marketing practice for reaching the right consumers and help to improve the conversion rate in sales.

The effectiveness of geo-targeting campaigns is estimated experimentally using a randomised experimental approach called geo-experiments. The approach is a patent published in Google Patents by Vaver and J. R. Koehler 2011. The basic concept of geo-experiments is as simple as it is intriguing: that is, a geographic region of interest is partitioned into a set of smaller non-overlapping areas, called "geos". These geos are randomly assigned to either a control or treatment group and geo targeting are then used to serve advertisements accordingly. The experiments are conducted into two-time periods. During the first period, there is no difference in the advertising campaigns between control and treatment geos, while during the second time period the advertising campaign is changed for the treatment group. The outcomes of the experiments measure the differences in user behaviours in the treatment geos relative to the control geos across the-two time periods with respect to the corresponding differential in ad-spend, which is the amount of money spent on advertising campaigns. In other words, the results of the experiments come in

the form of return on ad-spend (ROAS), which explains the incremental impact that the ad spend has on the consumer behaviour, (Vaver and J. Koehler 2011). ROAS is now the most widely used and informative measure of advertising performance at Google AdWords.

The outcomes of geo-experiments are observed at the geo level. Therefore, the geos are required to be selected in a manner that satisfies two essential points. First, it must be possible to serve advertisements to a geo according to its condition: treatment and control. Second, it must be possible to track ad-spend and behaviours of interest at the geo level. Many advertisers, however, fail to satisfy these requirement due to geographic design's problems and behavioural tracking challenges. For example,

- The generation of geos for geo-experiments is not straightforward. AdWords provides advertisers various set of geos in different locations around the world to choose where they desire their advertisements to appear. Geos include countries, cities, regions, provinces, counties, postal codes, TV regions and Nielsen designated market areas (DMAs). It is not obvious how geos are determined by Google but they are available on Google AdWords website and can be obtained easily. Google, however, updates these locations continuously, based on unknown criteria. Geos are not associated with demographics or socio-economics except average household income which has been added for DMAs. However DMAs are only available in the US. Therefore, the frame of the available geos are not well-defined for designing the sampling process for geo-experiments.

- Designing geo-experiments with a balanced treatment assignment of geos would be a challenge. Geos do not have the same possibility of receiving advertisements due to considerable heterogeneity among them. For instance, geos vary in their location, size, population and demographic and socio-economic characteristics. The design of the experiment, i.e. how to decide which geos receive the modified campaign during the second time period, is a concern to many marketers. A primary challenge is that the available geographical

areas on AdWords platform are not associated with sufficient demographic or socio-economic covariates although these covariates are likely to affect the probability of making conversion. The population distribution of some covariates such as age, gender and income for some areas might be known but it is likely that there are other important unknown and unobserved covariates. The challenge therefore is to design the advertising campaign in a robust way which permits estimation of the effectiveness of the modified campaign.

- Google AdWords manipulates sophisticated tools for tracking ad-spend and user behaviours. User behaviours might be, for example, online search, clicks, online or offline conversion, website visits, clicks, or any behaviour that is directly attributable to the advertisements. However, it has been considered hard to track these metrics because they represent micro-data. For instance, users might be recorded that they are seeing ads in one geo but are actually staying in another. They could be anonymous and having options to hide or change their IP address or turn off their location. In this case the ad-spend or user behaviours is associated to a wrong or unknown geo. Similarly, they might be seeing advertisements in one geo but do conversion in another unit. Here the ad spend for the same users is determined at different locations. In addition users might use multiple devices and hence many cookies are associated to a single user at either one geo or multiple geos. Cookies might be applied to track users but users might switch off cookies to avoid tracking and evade annoying irrelevant advertisements. Thus, AdWords might be able to collect detailed information about unique search queries, but for most of the queries there are insufficient or dirty data. AdWords, thus, tends to provide advertisers grouped or aggregated data by some contexts such as targeted-locations, (Fain and Pedersen 2006; Rutz and Bucklin 2011).

- AdWords aggregated data are used to compute performance metrics of advertising campaigns, which are used to infer user behaviours. The aggregated data, however, have some statistical issues; for instance, there might be a correlation between aggregated groups, that may lead to incorrect inferences about the

correlation between individual level. Also, the estimated performance metrics that are derived from aggregated data might be biased towards aggregated groups. The impact of aggregated data on individual level has been widely investigated by several authors from different fields such as social, physical and health sciences because they have to rely on published aggregate level data to infer at the individual level, (Clark and Avery 1976; Moulton 1990; Rutz and Bucklin 2011).

Despite the above mentioned challenges, geo-experiments are now a conventional approach for measuring the effectiveness of online advertising campaign at Google. Studies relating to geo-experiments have been relatively few (Blake et al. 2015; Valli et al. 2017; Vaver and J. Koehler 2012; Ye et al. 2016), and as far as we know there is no study addressing either generation of geos or treating messiness in tracking ad-spend. The geo-experiments in these papers were conducted over the 210 DMAs in US. In these geos, however, users are not restricted to administrative boundaries and share the same programme offered by television and radio and might receive the same internet content. DMAs are well-defined geos and broadly used in targeting by many marketing platforms. Vaver and J. Koehler reported that ad-tracking accuracy is a concern and they suggested that location and size of the geos can be used to alleviate this issue. However, they did not discuss the generation of geos, what practical size the geos should be and what are their general feasible features.

Designing involved in implementing experimental treatments were addressed very briefly in (Blake et al. 2015; Vaver and J. Koehler 2011, 2012), where Vaver and J. Koehler emphasised that random assignment of geos to control and treatment condition is an important component of a successful experiment and suggested to constrain this random assignment across one or more characteristics such as size or demographic variable. However, as mentioned earlier the availability of such variables is limited. Ye et al. (2016) employed a two-step approach in selecting experimental treatments: first matched geos in pairs on trend key metrics to reduce variability between control and treatment geos, then stratified users in control and treatment geos based on the national buyer segment distribution in geos to reduce

variability in user level behaviours. However, this design strategy is based on pre-observational data which are related to specific geos, users and time and hence their results may be questionable.

In this thesis, we focus on the design of geo-experiments, i.e. deciding which geos receive the change in the advertising campaigns during the second time period. This includes processing the generation of geos for geo-experiments and handling unobserved heterogeneity within geos in estimating treatment effects with the aim of designing advertising campaigns in a robust way which permits estimation of the effectiveness of the modified campaign. In particular, this thesis provides extended statistical methodology for the geo-experiments in six respects:

1. It proposes geos allocation algorithm for linking AdWords geos information to background information such as population and social-grades, which might be available from Government sources. The research focuses on AdWords geos in UK and links them to local authority areas, which can be considered well-defined areas because their detailed characteristics such as census data are expected to be available at the office for national statistics.

2. It applies different advertising campaign design strategies to allocate part of geos to serve the modified advertising campaign during the second time period. Complete randomisation and matched-pair designs, which is a special case of randomised block design are applied. The pairs are matched using population, social grades, realistic expected search rates and nearest neighbour algorithm, using dissimilarity measures between social-grades, dissimilarity measures between population and social grades and distances between geographical coordinates.

3. It proposes a conceptual model of geo-experiments for measuring the effectiveness of geographically-based online advertising campaigns, where the behaviours of interest are a two-stage behavioural process: online search and online purchase. The effectiveness of the advertising campaign comes in the form of the differential in the purchasing behaviour in the treatment spatial units relative to the control.

4. It develops a theoretical framework for correcting the potential wrong inferences in estimated effects of advertising due to unobserved heterogeneity structure within geos. Logit-linear regression models are used to fit purchase behavioural data. The maximum likelihood estimation technique is used to estimate the effectiveness of advertising campaigns. The applied statistical models used for estimation, however, is known to be misspecified when the presence of unobserved covariates is taken into account. Multiple truth models are thus proposed to measure misspecification of applied models. Kullback-Leibler divergence is used to measure the distance between the applied models and truth via a proxy model. An important part of the framework is a proxy model linking applied model and the truth. The proxy model makes possible the application of standard results in the literature on the maximum likelihood estimation for misspecified models (Chow 1984; White 1982) .

5. It proposes a conceptual approach to quantifying the error associated with estimated applied model parameters in relation to specific truth parameters. As a measure of effectiveness of the modified advertising campaign, the approach uses the hypothetical differential measure that measures the difference between the expected total number of sales if the modified campaign is served in all geos and the expected total number of sales if it is served in none. The error between estimated applied model parameters and the truth is then the difference between the sales differential measure computed using the estimated applied model parameters and the sales differential measure computed using the truth parameters.

6. It sets an evaluation method for measuring the performance of different campaign design strategies. The suggested design strategies in item 2 above are meta-designs and hence performance measures are needed to compare competencies of different campaign design strategies in estimating campaign effects in relation to a specific truth.

The proposed conceptual model of geo-experiments improves our understanding of the potential impact of hidden heterogeneity on estimating the effectiveness of

advertising campaign. The findings shows some evidence of the benefit of using social-grades based campaign design in reducing the error between estimated applied model parameters and the truth. The conceptual model, however, is not adequate enough to prevent a large bias or random error when estimating the hypothetical sales differential measure.

The theoretical behaviour of estimated applied model based on the theory of the maximum likelihood estimation for misspecified models is investigated through Monte Carlo experiments, by assessing whether the theoretical distribution of estimates are suitable to describe their empirical sampling distribution. The results indicate the key role of the theoretical approximation to the sampling distribution of applied model parameters. This saves having to do expensive Monte Carlo simulation all the time.

The conceptual model counts only two behavioural process search and conversion although the real world behaviour structure is a multi-level. The model ignores ad-spend variable because our knowledge of ad-spend or its distribution is very narrow. The model also ignores the statistical issue of aggregated data and does not take into account seasonality factors, which is important in measuring campaign effects. Notwithstanding these limitation, the proposed conceptual model is a good start point for understanding the estimation of advertising campaign under misspecification when geo-experiment is applied. It is also important for showing the potential effect of including covariates related to unobserved heterogeneity in designing the sampling process of the experiments on campaign effects. We expect that thr reader of this thesis will share in our opinion that the proposed conceptual model of geo-experiments provides insights into the estimation of advertising campaign effect. Therefore, we close the introduction with a preview of the coming chapters in this thesis.

Chapter 1 provides a brief primer on Google AdWords campaigns, including the history, concept of AdWords, structure of AdWords campaigns and AdWords performance metrics. It also presents briefly some estimation approaches that have been proposed in marketing literature for measuring the effectiveness of online advertising

campaigns as well as geo-experiment approach.

Chapter 2 presents an example of a real world application of the geo-experiments. The application was conducted for a retail company in UK which aims to determine whether bidding on their brand terms drives incremental sales. It shows the challenge in designing the advertising campaign and how data from tracking users can be messy and incorrectly tracked.

Chapter 3 links UK AdWords target geos to UK local authority areas to create a well-defined sampling frame for treatment geos. The linking algorithms is based on the the shortest great-circle distances between target geos and local authority areas which depend mainly on the longitude and latitude measures. The sources of local authority areas include Office of National Statistics and UK Data Service, which provide access to census micro-data. The linking process returns subset of local authority areas associated with some micro-data such as population and social-grades characteristics. The term spatial units is used to refer to this subset of geographic locations. The research focused on spatial units in England and Wales.

Chapter 4 introduces a conceptual model of geo-experiments to understand the key tenets of user behaviours. It sets out the potential user behaviours that may result from geo-experiments and outlines two behaviours: online search and purchase as two major entities of the conceptual model. It lays out list of assumptions to enable the two-stage conceptual model measuring the impact of advertising campaign. It assumes that the effectiveness of the campaign is attributable to the act of converting online searches to purchases. It proposes logit-linear regression models that are required to fit search and purchase metrics under two conditions: homogeneity and unobserved heterogeneity with spatial units. The models with the first condition are what we called applied models and the second condition are the truth models. It also proposes the hypothetical measure of campaigns effects that are required to quantify the error between estimated applied model parameters and truth parameters. At the end of this chapter, five primitive elements are discussed: investigation of realistic searches and purchases, identification of spatial units, specification of truth parameters, identification of population-strata and specification of campaign

design strategies. The primitives are basic components of theoretical and simulations approaches that are developed in chapter 5 and chapter 6 for estimating the effectiveness of advertising campaigns.

Chapter 5 develops a theoretical framework for studying the consequences of applied model misspecification and how the applied model parameter estimates are affected. It studies the asymptotic behaviour of the estimated applied model parameters. It reviews briefly the Kullback-Leibler divergence criteria and the results obtained by White 1982 about consistency property and asymptotic normality under misspecification. It proposes a proxy structure of the applied model to make possible the application of standard results in the literature on maximum likelihood estimation for misspecified models. It gives the theoretical approximation of the asymptotic distribution of estimates for purchase model by considering a single stage stochastic process, which is the case where there is no search process. It extends the asymptotic theory for a two-stage model.

Chapter 6 tests the applicability of the theory by assessing its performance in a variety of contexts in comparison to Monte Carlo simulations. It suggests a set of interesting truth instances to explore the merits of different campaign design strategy and investigate under which circumstances the theoretical framework provide adverse inferences. The investigation followed the assumption that the search process is known. It reports that most used truth instances corroborates the ability of theory to describe the sampling distribution of the estimates of the applied model parameters. The violation of the theoretical distributions are only found in few cases that relate to using large level of heterogeneity within spatial units or large number of search.

Chapter 7 assesses the performance of a specified advertising campaign design strategy for a specified truth instance. The assessments takes into consideration contributions of two sources of variability: the approximation error and the sampling error. The approximation error refers to the difference between the proxy model hypothetical measure of campaign effect and the true hypothetical measure of campaign effect, and the sampling error refers the difference between estimated applied model

hypothetical measure of campaign effect and the proxy model hypothetical measure of campaign effect. It shows that hidden covariates have the potential to inference incorrect estimation about the campaign effect.

Finally the thesis is closed with some appropriate recommendation related to realistic behavioural structure improving the conceptual model and future studies needed on measuring online advertising campaigns using geo-experiments.

Chapter 1

Online Advertising Campaigns

The purpose of this chapter is to provide a brief overview of Online advertising campaigns. It begins with a section on background information regarding Google AdWords, including the history, concept of AdWords, structure of AdWords campaigns, AdWords performance metrics. The following section gives examples of inadequacy in AdWords performance metrics. The end of the chapter presents some estimation approaches that have been proposed in marketing literature.

1.1 A Brief Primer on AdWords

Advertising is a 'prominent feature of economic life', as noted by Bagwell in 2007 (Goldfarb 2014). The arrival of the internet has boosted what Bagwell stated in 2007 and has led economists to employ online advertising to carry out digital commerce transactions. The first online advertisement appeared in October 1994, by HotWired, now called Wired.com, as a banner advertisement (Weller and Calcott 2012). People interact hugely with online search engines in diverse ways, starting from browsing and ending with payment transactions. Web search engines function as navigational tools that transfer people in their searching journey from one webpage to another. Consequently, technology search engine companies such as Google and its partner websites like AOL, Amazon and Yahoo! are constantly trying to improve the retrieval aspects of their services. Through technology and innovation, the

search companies were able to find a technique to improve web retrieval, which they called sponsored search advertising (Jansen and Mullen 2008). Sponsored search advertising, or paid search advertising is an online service, where advertisers pay web search engines for traffic to their websites deriving from the search engine results. This service has significantly turned some of the prominent web search engines into digital commerce companies, whose revenues depend heavily on advertisements (Weller and Calcott 2012).

Based on the history of sponsored search auctions described by Jansen and Mullen 2008, sponsored advertisements (referred to as 'ads' from now on) were presented in 1998 by GoTo.com, known as the Overture Service, which was then acquired by its major client Yahoo! in 2003. In 2005, the Overture Service was changed to Yahoo! Search Marketing. In 1998 Google was founded by Larry Page and Sergey Brin and it adopted the sponsored ads model in 2002, (Fain and Pedersen 2006). At the same time, it modulated to an online advertising service that allows advertisers to position their ads in Google search result pages, called Google AdWord (Weller and Calcott 2012). On 24 July 2018, it changed into a new branded platform service called Google Ads. Google Ads is a cross-platform that connects all features of the Google AdWords and its display network, from search ads to display ads to video ads, in a more seamless way. In this thesis we focus on Adwords search ads that appear on Google search results pages.

1.1.1 What is AdWords Ads

Nowadays, it is extremely difficult to place a search query term within the Google search box and get outcomes without ads or sponsored results. For example, a Google search results page for the query of a "laptop", shown in **Figure 1.1** displays sponsored ads in different formats such as images and texts. The displayed images in the upper area, and the links that appear under an "Ads" label and "sponsored" label, are called sponsored search ads or paid search ads. On Google AdWords, these results are also known as AdWords ads. The remaining results on the search result page do not belong to Google AdWords and are called unpaid, natural or organic search results .

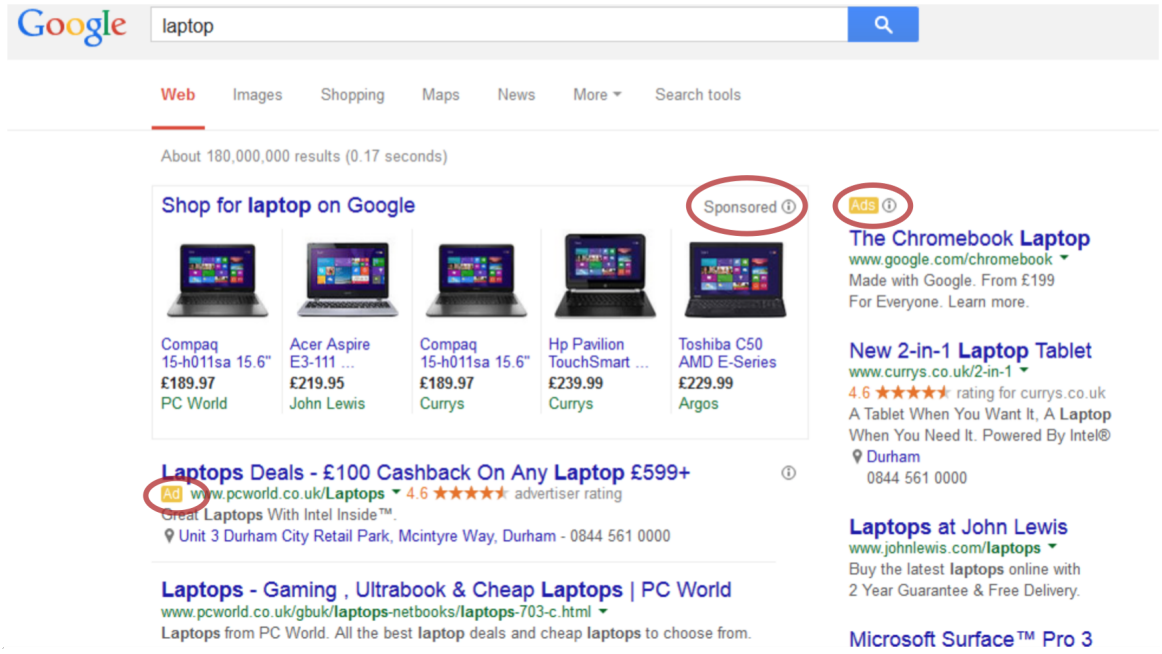


Figure 1.1: AdWords Ads appearance on Google search results page as shown by red circles.

Google AdWords provides advertisers with a mechanism to get their ads to appear on Google search results pages and to get searchers to visit their websites. The mechanism of online advertising and AdWords described here was reported in *AdWords Fundamentals: Exam Study Guide* 2015. The most common model is pay-per-click (PPC) where advertisers pay Google when a searcher clicks their ad's link. There are other models that exist, such as pay-for-impression, pay-per-action and pay-per-call. The cost per click (CPC) is determined by Google through an algorithm called a 'quality score', which depends on multiple factors including the relevance of keywords to ads, relevance of ad text, quality of advertisers' websites and the relevance of historical performance. The click through rate (CTR) for the relevant historical performance is a metric that measures the number of clicks advertisers receive on their ads per the number of times ads appear. At the same time, keywords, which are defined as selected words or phrases that are used to match ads with search terms entered by consumers are also one of the most appropriate metrics when regarding historical performance. In principle, strong quality scores are an indication to Google that ads are relevant and helpful to consumers. Hence, ads with strong quality scores are rewarded with higher ranking on the search result pages and lower CPC. There are other important factors besides quality scores that

are used to determine the ads position, including the maximum bid amount and the context of a user's searches, such as the user's location, devices and time of search. The maximum bids- also called a bid cost - is the amount that advertisers are willing to pay for their ads to appear. An ad's position and how it is displayed, differ by device type, depending on a chosen device targeting. For example, ads appearing on a computer desktop or laptop are displayed differently than those appearing on mobiles or tablets. AdWords obtains the device category from information contained in the browser's user agent string. The user's search location is determined by different mechanisms [1] such as country-specific versions of Google search, user's location preferences in search settings, geographic-location of IP addresses and world wide web consortium geographic-location application programming interface (W3C Geolocation API), which allows users' browsers to use various signals, like visible wifi networks or a GPS, to determine location. It is thus essential to the advertisers to know where to show their ads and when. In other words they need to know who their users are to reach the goal of their advertising campaigns.

AdWords offers various ways of targeting, including audience and content targeting. In audience targeting, advertisers focus on audience features such as their physical locations or geographical location, demographic characteristics such as age and gender, their own devices, search history, conversion history and affinity for TV advertising campaigns. In content targeting, advertisers focus on matching the actual content of their advertisements including topics and keywords with the content of the search query. Advertisers are thus required to propose an advertising model, a maximum bid, a list of keywords, a targeting criteria and ads layout to create their advertising campaign on AdWords.

1.1.2 AdWords Advertising Campaign

AdWords campaigns are advertising campaigns in an Adwords account in which the account can be thought of as a control and management room. Advertisers can

[1] Robert Love, who works at Google, wrote an answer in Quora in Dec 2013 to the question: how does Google Search determine my location? https://www.quora.com/How-does-Google-Search-determine-my-location.

organize their campaigns and arrange a delivery plan of products or services that they offer to users. In an Adwords account, advertisers can create more than one campaign and plan each campaign separately. Each campaign is usually composed of one or more ad groups, depending on the goal of the campaign, device segmentation and product type. The setting at each campaign level includes determining a bid cost, choosing lists of keywords and deciding targeting strategies. Here we describe briefly how to set up and manage an AdWords campaign in terms of bidding mechanisms, keywords selection and targeting criterion. The description of these three components is an overview of their main specifics presented in the *AdWords Fundamentals: Exam Study Guide* 2015.

Bidding

The placement of the advertisement is crucial to achieving the advertising goal. Advertisers have to display their ad prominently on the search results page, for example, within the top four of the the ads results page. The ad position is based on the amount of money that advertisers are willing to spend to display their ad, and on their quality score, i.e. the ad's relevance in terms of keywords, quality of their website and advertising performance (Blake et al. 2015). Advertisers, therefore, should offer higher bids to receive a higher position. This helps advertisers to gain more clicks on their advertisements, which are more likely to be converted to a purchase.

AdWords offers a variety of bidding options, such as CPC and cost per thousand impressions (CPM). CPC is mainly manual bidding, where the advertisers tell Google the maximum amount of money that they are willing to spend per click. This is the default bidding method where advertisers control the bidding amount. There are other CPC types where Google fixes or adjusts the bid, which are known as automated bidding or enhanced CPC. With CPM bidding, advertisers pay Google for each set of a thousand views for their ads.

The choice bidding method varies, based on the goal of the advertising campaigns. CPC bidding is best suited for advertisers whose aim is to increase traffic to their website or increase sales, whereas for advertisers who want to raise brand awareness,

CPM would be better, as has been recommended in AdWords Fundamental Guides. The other type of bidding is cost per action (CPA), also known as cost per conversion, which is available to advertisers who are interested in optimising the conversion rate of their advertising campaigns. The action can be, for example, a sale or click.

Keywords

AdWords allows advertisers to place ads that directly match search queries entered by consumers. AdWords ads a link to a landing page, which is a page on the advertiser's website that showcases a product that is directly attributable to the search terms. The search terms that lead to displaying relevant ads are called keywords.

Keywords are phrases or words chosen by advertisers to bid on during the advertising campaigns. There are six different characteristics assigned to the keywords: brand, generic, brand generic, manufacturer, manufacturer generic and manufacturer product. The brand keyword includes only a retailer company name, such as *Argos* or *Amazon*. The generic keyword is a general search term of a product without a company or a manufacturer name, such as *buy laptop*. The brand generic keyword includes a company name and a generic phrase, such as *Amazon laptop*. The manufacturer keyword includes only a manufacturer name, such as *Toshiba*. The manufacturer generic keyword includes the name of a manufacturer and a product, such as *Toshiba laptop*. The manufacturer product includes the manufacturer's names with a detailed description of a product, such as *Toshiba laptop windows 7*.

AdWords matches the selected keywords with the search terms entered by users, using five primary matching processes: exact match, phrase match, broad match, broad match modifier and negative keywords. In exact match, the keyword matches the search query exactly, which is defined in AdWords settings as a phrase within brackets. For example, *office chair* specifies bids on exact match keywords and so the ads appear if the entered search term matches exactly with the allocated phrase. In phrase match, the search query is not required to agree exactly with the allocated keyword phrase but must contain the keyword terms in order. This means that the search query could have words before or after the selected keywords. The bidding on

phrase match is identified by entering the keywords within quotations, like "office chair". The search term "luxury office chair", would correspond to this type. Both the exact match and the phrase match do not consider misspelling and plurals as matches. The broad match presents all possible combinations of the keyword phrase. It is specified in the bidding criteria that there should be no character around the phrase, for example office chair. This type is recommended to be employed with negative keywords, to exclude consumers who do not exactly match the aim of the ad campaigns. For instance, the broad match criteria for the phrase *office chair* will consider a search term such as *garden chair* as a match and so advertisers could select *garden*, for example, as a negative keyword to ensure the ad appears only to those who are looking for an *office chair*. In broad match modifier a "+" sign may be added in front of a word that is required to appear in the search, for example, +*office chair*. There should be no space between the word and the + sign.

Targeting

An AdWords advertising campaign would not be efficient if it did not appear to the right customers or in appropriate locations. AdWords offers different ways of targeting: audience targeting and content targeting, as discussed above. We would like to focus here on a location targeting service, which is an example of audience targeting.

AdWords applies sophisticated tools to determine internet users' locations, based on the identification features that are attributable to their computer desktop, mobile phone or tablet, including device location. Advertisers can thence target users by either their physical locations or locations they are interested in. The physical location is determined by IP address, GPS, WiFi router or Google's mobile ID location database. The location of interest can be detected using country domain, such as (.uk) or (.fr), including the name of the location in the search query, or by using the search within an area on the online map.

AdWords location targeting could be estimated by past data from users' online search history or physical locations if users enable location settings in their devices. Targeting users' locations by tracking both the personal locations and the historical

information is known as behavioural targeting (Lavrakas 2010). The identification of physical locations does not always reflect users' locations accurately. Incorrect identifications might occur due to some technical issues related to the server or privacy option chosen by users to disable detection of their locations. The details of the technical issues are beyond the scope of this research study. Due to the challenges mentioned, it is recommended to use the physical locations along with the targeting of geographic locations.

AdWords designates various types of geographic location to assist advertisers in choosing where they desire their ads to appear. The service includes countries, cities or areas within a country, a radius around a location, or location groups. The selection of the ads targeting locations depends mainly on the business type and its aim. For example, a business that serves an entire country, selling to several countries or delivering worldwide, will target ads towards a country or multiple countries. Whereas, if a business serves certain areas within a country then there is no need to target the entire region and so instead, it assigns ads to selected areas or cities. A local business that serves people located within a certain distance from its location will assign ads to a radius around a location. The areas within the target location are selective, where advertisers can choose interesting spots and exclude unwanted ones. This technique can also be used in a business with multiple locations, typically known as location extensions, to set up location groups, where each group is a list of target areas or cities within the business locations. The location groups can also by target people who are in places such as airports, universities or central commercial areas. Recently, demographic information such as average household income has been added to Google AdWords, but this information is available as interval data and for the United States only. Demographics are used to compose groups of locations that are categorised by average household income. The Google AdWords Help page presents a detailed description on how to set ads target locations for each available technique.

Location targeting types vary by country world-wide and include countries, cities, regions, provinces, counties, postal codes, TV regions and Nielsen DMAs. Nielsen

DMAs refers to designated market areas (DMAs) which are televison market areas or market areas where local television viewing is measured by The Nielsen Company. People in these areas share the same programme offered by television and radio and might receive the same internet content. There are 210 Nielson DMAs and they are only available in the United States. These areas are not restricted to administrative boundaries, so that areas with low population density can be grouped to form one market area. Conversely, very large areas can be divided into sub-areas which allow people who live on the edge of an area to receive media content from their adjacent area rather than their actual areal affiliation. This feature affords standardized behaviours and attitudes of individuals in each DMA, which appeals to advertising platforms to utilize them as geographic targeting areas (Vaver and J. Koehler 2011).

1.1.3 Performance Metrics of AdWords Campaigns

The success of AdWords campaigns depends on the goals and objectives of the campaigns. For example, the campaign goals can be brand awareness, return on investment, sales and conversions. AdWords provides a conversion tracking tool to track consumer behaviour. Google and its partners' platforms and the advertisers' websites use cookies to track consumer activities on their webpages. A cookie is a small text file placed on a hard drive of a user's device to store and transmit information to the server about the websites visited from the user's browser. With tracking behaviour data, AdWords Analytics generates different performance reports including certain statistics or metrics that are relevant to the campaigns' goals. The advertisers, for example, can receive metrics about the number of visits to their sites, known as websites' visits, the number of visitors whose clicks end up as purchase or not, as well as sales and revenue.

An AdWords performance report includes metrics, that estimate outcomes as a result of users clicks on ads. The metrics are aggregated estimates on ad group level, keywords or location, where each aggregated metric determines the value of the campaign in general or a certain component of the campaign. For example, quality score, which is provided usually in term of CTR, ad relevance to keywords and

landing page experience, measures the performance of ads, keywords and landing pages. There are also cost per view (CPV) and ad rank that measures the ad positions. The performance of sales or conversion can be measured by several metrics such as CTR, CPC, CPA, conversion rate (CR) and return on investment (ROI). The ratio between CPC and CPA calculates the CR, where the total conversion is the revenue, also known as total sales. ROI is the ratio between the net profit and cost of investment, given that the net profit is the difference between revenue and cost. According to Google AdWords, the cost of investment is defined as "the cost of goods sold, which is for physical products; the cost is equal to the manufacturing cost of all the items advertisers sold, plus their advertising costs". Another important metric is the average cost charged for a click, also known as the average cost per click (avg.CPC). It is the ratio of the total cost of ads clicks to the total number of clicks. In Google analytics, these performance metrics are not only for ads conversion but can also be customised based on other interesting aggregated features such as locations target and impressions.

1.2 Inadequacy of AdWords Metrics

From a business perspective, these metrics help to measure the effectiveness of advertising campaigns but they do not provide advertisers with an integral picture of their campaigns impact. One example is CTR which measures the number of clicks on a given ad with how often it is displayed. Given that ads are displayed at different positions on the search result page, how would the list of ranked results influences users' clicking decision process. Joachims et al. 2017 studied how users scan the results page using eyetracking - an experimental study of eye movement on search results page - and how their scanning behaviour relates to the clicking decisions. They found that users' clicking decisions are biased to the highly ranked results even if these results are not relevant, and biased to the overall quality of the ranking results. Therefore, clicks are not easy to measure and need to be attributable to the order of the results in terms of their relevance and quality. Additionally, a primary concern to advertisers, as pointed out by Vaver and J. Koehler 2011, is the potential of more clicks on an organic link, when the occurrence of AdWords ads

coincides with the occurrence of an organic link. In this case, an organic link has a competitive effect on the paid or PPC link, and hence CPC flatten the picture of the AdWords impact. Thus, it is necessary to observe the clicks in the absence of a PPC link to estimate the incremental impact of PPC clicks on the total clicks, and so the CPC could then be replaced by cost per incremental click CPIC, as suggested by Vaver and J. Koehler. The incremental clicks express the change in the paid clicks at different ad-spend over the same time period, (D. X. Chan et al. 2011). Presuming that for a certain time period, an AdWords campaign is conducted, the differential of ad-spend is then resulted from making an intervention on the existing campaign at the same time period, which is equivalent to running a new campaign. The intervention could be turning off the existing campaign or eliminating brand keywords. Given that there is a change in the ad-spend, the outcomes of the total clicks and paid clicks from the two campaigns are compared. Measurement of the causal impact of the intervention over a time period is usually carried out by a comparison experimental study. However, many advertisers have a concern about the cost of conducting distinct campaigns and the adverse impact the intervention might cause to the revenue. Therefore, they conduct campaigns with a one level of ad-spend and predict the clicks for any given level of spend through a statistical model. For example, a Bayesian statistical model for paid and organic clicks as a function of the search ad-spend and organic impressions was used by D. X. Chan et al. to estimate the incremental ad clicks (IAC). Vaver and J. Koehler estimated IAC as well, but through an experimental approach, called geo-experiments. The approach is a geography-based advertising experiment and registered as a patent in Google Inc. The result of geo-experiment is in the form of return on ad-spend (ROAS), which is the incremental impact that the ad-spend had on a response metric. In their paper, they considered clicks as a response metric and so ROAS for clicks was found. They then estimated IAC by computing the ratio of CPC to CPIC. Since September 2014, ROAS[2] has appeared as an AdWords metric in Google Analytics. It is computed as a ratio of the revenue to the amount spent on AdWords ads.

[2]https://plus.google.com/+GoogleAnalytics/posts/jGarHQe3MLx

On the other hand, Vaver and J. Koehler estimated ROAS using a two-stage linear regression model. Furthermore, since July 2018, AdWords has provided a target ROAS bidding where a maximum cost-per-click (max.CPC) bids is automatically set, to maximise conversion value to achieve a target ROAS. It is a special automated bidding strategy, called smart bidding, that uses a machine learning process to optimize conversions.

AdWords metrics are evaluated to get some perspectives on the advertising campaigns to see if they met the intended aim or not. However, one additional problem in AdWords metrics is the result from aggregated data, despite the fact that Google search engine manipulates sophisticated tools for individual-level tracking. In principle there is sufficient data of individual behaviour to estimate AdWords metrics. Consider, for example, CR and CTR, both represent conditional probabilities, where CR is a probability of an action given a click and CTR is a probability of a click given an impression. Fain and Pedersen 2006 pointed out that with sufficient AdWords data these probabilities can be estimated. However, AdWords Reporting and Google Analytics Data have detailed information about unique search queries, and for most of the queries there are insufficient data. Adwords, thus, tends to provide advertisers grouped or aggregated data by contexts such as ad level, keywords or targeted-locations, (Fain and Pedersen 2006; Rutz and Bucklin 2011). AdWords aggregated data are used to compute AdWords metrics, which are used to infer individual behaviour. The aggregated data has some statistical issues; for instance, there may be a correlation between aggregated groups, that may lead to incorrect inferences about the correlation between individual level. Also, the estimated metrics that are derived from aggregated data might be biased to aggregated groups. The impact of aggregated data on individual level has been widely investigated by several authors from different fields such as social, physical and health sciences because they have to rely on published aggregate level data to infer at the individual level, (Clark and Avery 1976; Moulton 1990; Rutz and Bucklin 2011). According to Rutz and Bucklin, the nature of the available paid search data are aggregated but the questions asked are: do advertisers rely only on aggregated data to measure the performance of their AdWords campaigns? and How does Google process individual

level data and Which data does Google share with advertisers?

It is difficult to understand how Google collects and processes individual personal and behaviour data because this represents its business model. It is also not clear which data Google shares with advertisers. Google knows a lot about individuals that use its services, despite the fact that it provides users with tools to manage their privacy and security of their Google account. According to Google Play Services, the personal identifier in the Google personal collected data is reset with a unique advertising ID. This indicates that the tracked personal data are collected and then anonymised with the advertising IDs. With these IDs, users could opt out of personalised ads or what is called interest-based ads within Google Play apps. According to Google privacy, opting out of interest-based ads does not prevent Google from collecting users' data, but it stops Google associating their collected data with their advertising IDs. Recently, with the General Data Protection Regulation (GDPR)[3], that was implemented in May 2018, Google is the handler of personal data. With GDPR, advertisers are required to obtain consent for tracking individuals on their website and for the use of cookies and of personal data for personalised ads for users in the EEA. Therefore, under GDPR, the advertisers and Google act as independent controllers. However, there is an exception for customer match advertising tactics, where Google acts as a processor for advertisers provided with personal data. For the interested reader, further details about AdWords ads data under GDPR can be found on Google Ads Help page[4]. According to Google advertising policies, advertisers are allowed to use the first-party data, i.e. individual data collected by advertisers, to create audiences for ads targeting, mainly for personalised ads. In addition, they are allowed to use the third-party data, i.e. individual data that advertisers obtained from other sources to segment their first-party audiences to create a re-targeting lists. Thus, we might say that individuals' behaviour data are effective in improving the tactics used for targeting likely customers, whereas aggregated data may add valuable insight to the performance AdWords campaigns.

[3](GDPR) is a regulation in EU law on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA).

[4]https://support.google.com/google-ads/answer/9028179?hl=en-GB

A further issue with AdWords metrics is that tracking mechanisms of data that are required in the metrics' computation are not transparent . For example, mechanisms employed in tracking ad-spend that are required to measure the ROAS metric are not demonstrated. The ad-spend could include costs incurred in setting up the campaign, of designing layout of ads, cost per impression or bidding cost and cost per click. The amount spent in setting up the campaign and ads layout might be deterministic but once the campaign is ready for bidding to appear, the ad-spend tracking is likely to be dirty. Little is known about the technologies that are employed by Google in tracking ad-spend or individuals behaviour, but we do know there is contamination due to limitation in tracking individuals' behaviour and in accounting changes in their behaviour.

In reality, the actual effectiveness of the advertising campaigns has been considered difficult to quantify. One of the key challenges is to design campaigns that comprise multiple interrelated variables including choice of keywords, setting of targeting, amount of bid, layout of the ad and design of advertiser's website. AdWords practitioners endeavour to fix these variables to optimise the performance of their ads. However, there are other uncontrolled external variables that affect the efficiency of the campaigns. Recall, for example, the fact that the higher the bids that the advertisers submit, the more traffic is driven to their website. However, the website traffic depends on consumer behaviour, which is affected by the changes that occur on the supply and demand process during the advertising campaign. These changes are the result of uncontrolled effects, such as advertiser or consumer behaviour, and time and calendar effects due to different browsing behaviour on different days, for example. These factors, consequently, affect the measurement of the actual return of the advertising.

1.3 Estimation Approaches for Measuring the Effectiveness of AdWords Campaigns

Estimating the effectiveness of online advertising campaigns has received a notable amount of attention, however there has been limited academic examination of both

AdWords campaigns and their effectiveness in a comprehensive manner. Despite this fact, by researching paid search advertising literature, several attempts have been made to develop online advertising effect models to estimate the effectiveness of online advertising campaigns using specific ads performance metrics. In the literature, models are often estimated by using observational data and are rarely based on experimental randomised designs data . For illustration, we consider the following examples, where models are estimated by using observational data: first, Kim et al. 2011 developed an analytical Bayesian approach that can incorporate CTR data to give inference about advertising effects by employing the Poisson-Gamma distribution. Second, Rutz and Bucklin 2011 developed an empirical two-stage consumer level model of CTR and conversion model to evaluate the effects of specific properties on ad performance, using behavioural primitives in accord with utility maximisation. Third, D. X. Chan et al. 2011 developed a Bayesian model to estimate an incremental click impact from advertising campaigns, by quantifying the impact pausing search ad-spend has on total clicks.

In observational studies, the effect of advertising campaigns is measured by observing consumer behaviour over a certain period of time while ads are displayed. The behaviour of users who are not exposed to ads, on the other hand, is estimated through a statistical model (D. Chan et al. 2010; Lewis et al. 2011; Rutz and Bucklin 2011; Vaver and J. Koehler 2011). The lack of direct observation of an unexposed group (control group) makes such studies questionable and their inferences and conclusions are unreliable - statistically- compared to randomised experiment.

In experimental studies, consumer behaviour is estimated by comparing the behaviour of two different groups called treatment/test and control groups. Consider, for example, studying an impact of offering higher bids and ask the question: is the advertising effective due to the bids increase or to other endogenous factors? The simultaneous comparison of two groups where one is before the bid's increase and the other is after the change, can suggest a reliable answer to the above question. That is because each incoming search is randomly allocated across the treatment or control condition and then the impact of change in advertising is measured by

comparing consumer attitudes towards the change only (Vaver and J. Koehler 2011). The randomised experiment, therefore, is implemented by Google AdWords in an application called AdWords Campaign Experiment (ACE). This application is a tool to help advertisers test the impact of modifications on their advertising campaigns.

The ACE application is deemed to be effective in identifying users' behaviour but as highlighted by (Vaver and J. Koehler 2011), ACE is limited to baseline searching level on the search platforms. The ACE for example, fails to detect the initial user condition (treatment or control) when purchasing is made. This is because purchase decision might involve several searches and multiple visits to the advertisers' website. A cookie experiment is another experimental approach, where each cookie is assigned to similar control or treatment group (Vaver and J. Koehler 2011). It is demonstrated as an alternative experimental method to overcome ACE drawback. In practice, however, the user may own multiple devices like office computer desktop, personal laptop and smartphone, where the desktop might be used for searching and the laptop is then used to make the purchase. This tends to resulting inconsistent ad serving, which affects the results produced by cookie experiment. Despite this, cookie experiments have been applied on Google to measure the effectiveness of display advertising (Vaver and J. Koehler 2011). ACE was also used at Google until February 2017 and after that it has been replaced by what are called campaign drafts and experiments [5]. In a campaign drafts and experiments approach, advertisers could propose multiple changes to the search and display campaigns and investigate the impact of the changes that were made on the campaigns. Based on the AdWords Help page, Advertisers could do this by creating a draft campaign in the exiting running campaign and make intended changes. In other words, they split the existing campaign into two versions: an original unchanged version and a changed version. Then they either apply draft changes back to the original campaign or use the draft to create an experiment. The experiment helps advertisers to understand the impact of changes before they apply them to the original campaign. The

[5] https://support.google.com/google-ads/answer/6318732?hl=en-GB

experiment could end earlier than the original campaign, if its performance is preferable, it could then be applied to the original campaign. The experiment could also be converted to a new campaign and the original one paused. The results obtained from campaign draft and experiment, however, can be biased, due to a biased sample assigned to the changed version of the campaign. Random sampling is essential for such experiments but due to presence of uncontrolled factors such as promotion and seasonality, applying a randomisation strategy could be a challenge.

Another randomised experimental approach that Google has successfully employed is the geo-experiments. The experiments were carried out by Vaver and J. Koehler 2011 to measure the incremental impact of advertising campaigns. In these experiments, advertisers estimates ROAS through the geographic bid feature. Geos are randomly assigned to either a treatment or control groups. Pre-ad-spend data for each geo are received into two time periods. The change in ad-spend data for each geo is determined. A linear model was used to estimate ROAS of the behaviour measure.

Vaver and J. Koehler 2011 stated that the geo-experiments are worthy of consideration because they are conceptually simple, have a systematic and effective design process, and their results are easy to interpret. At the same time, while this may be true, the implementation of the geo-experiments is not straightforward, where geos are required to be selected in a manner that satisfies two essential points. First, geos must be able to serve ads according to their condition, treatment and control. Second, geos must be able to track ad-spend and behaviours of interest at the geo level. Many advertisers, however, fail to satisfy these requirement due to geographic design problems and behavioural tracking challenges.

Geo-experiments have been mentioned in few recent studies. For instance, Blake et al. 2015 conducted a series of controlled experiments for a well-known brand eBay, by having no ads in some geos while the campaign continuing on in other geos. However they found no considerable difference in sales between treatment and control. Ye et al. 2016 reported two serious limitations to geo-experiments: high cost due to potential revenue impact from having turning ads off in some geos, and low

statistical power in detecting the difference between control and treatment groups due to small sample size of geos in each group and large noise in data. In addition, Brodersen et al. 2015 and Ye et al. 2016 pointed out that rigorous causal inferences can be obtained through implementing geo-experiments. At the same time, however, they criticised such experiment due to their requirements. For example, the target region of interest can be an entire country, particularly for a national advertising campaigns. Therefore they indicated that using the entire country prohibits the use of control geos within that country. Also, the experiments can be carried out in several countries but not necessarily at the same time. In this case, they have indicated there can be a large number of control group but the treatment group can be consists of one country or a few countries with considerable heterogeneity among them.

It appears from the aforementioned approaches that it is difficult to quantify the actual performance of advertising campaigns. It may also be noted that most of the studies have focused on modelling certain metrics, not on the processes involved in implementing campaigns. Nevertheless, there have been relatively few experimental studies focused on campaign implementation processes and geo-experiment is one of them. In this research we are interested in providing extended statistical methodology for geo-experiments, in particular, statistical methodology of randomised evaluation.

Chapter 2

Application of Geo-Experiment

This chapter presents a real world application of the geo-experiment. The experiment has been used by "Consultancy ★"[1], an online marketing consulting company in UK. The company has executed the geo-experiment for several retail companies. Their experience in performing the advertising campaigns using geo-experiments were shared with us through discussions in meetings and emails.

The original goal of this research was to analyse data driven by different online advertising geographically-based campaigns for clients of Consultancy ★, to measure the campaign's incremental impact. Thus company provided us data sets of different applications along with the description of the geographic designs of the campaigns. Looking at the materials provided, we found that the advertising campaigns, however, were designed by the company with some but not full attention to statistical principles. We observed also failures in tracking users' behaviours, which was not a surprise. The focus of the research then turned to reviewing the applied campaign designs and making suggestions for other design strategies that would permit estimation of the effectiveness of online advertising campaigns when geo-experiments are applied.

The chapter begins by looking at how the company applied the geo-experiment to

[1]The company name has been obscured to maintain confidentiality.

a retail company "B" [2] to investigate the importance of brand keywords - keywords include only a retailer company name - in the online advertising campaign. Then it gives a brief overview of the data gathered by them from AdWords. We move then to summarise the PPC revenue at the level of geo-locations during two time periods. This summary will be used in later chapters as a realistic input in a computer experiment. At the end some remarks and conclusions are drawn.

2.1 Advertising Campaign of Retail Company "B"

The experiment was conducted for a retail company "B" which aims to determine whether bidding on their brand terms "B" drives incremental sales or revenue. A geo-experiment advertising campaign was run for "B" in the two month period July-August 2012, where July represents the first time period in the experiment and August is the the second time period. For some targeting geo-locations in the second time period, brand keywords in the campaign were turned off.

2.1.1 Campaign Design

The experiments began with the identification of a set of targeting geo-locations. Google AdWords provided "Consultancy ★" with 1009 geo-locations with no geographical information. The company added geographical coordination data to each geo-location: longitude and latitude using a free geo-coder service called GPS Visualizer [3]. he company then used the point locations to visualise the geos as shown in the map in **Figure 2.1**.

To apply the geo-experiment, the company needed to partition the geo-locations into two groups: control and treatment. However, before assigning geo-locations to treatment, the company grouped geo-locations into 30 circular zones, such that 15 were in England, 10 in Scotland, 3 in Wales and 2 in northern Ireland. We understand that the company took into account the notion stated by Vaver and

[2]The retail company name has been obscured to maintain confidentiality.

[3]https://www.gpsvisualizer.com

J. Koehler that grouping geos by size prior to assignment can reduce the width of the confidence interval of the return on ad spend by at least 10%. Equivalently, grouping geos will assume that geos in each group share some similarities which help to reduce the variation between the geos.

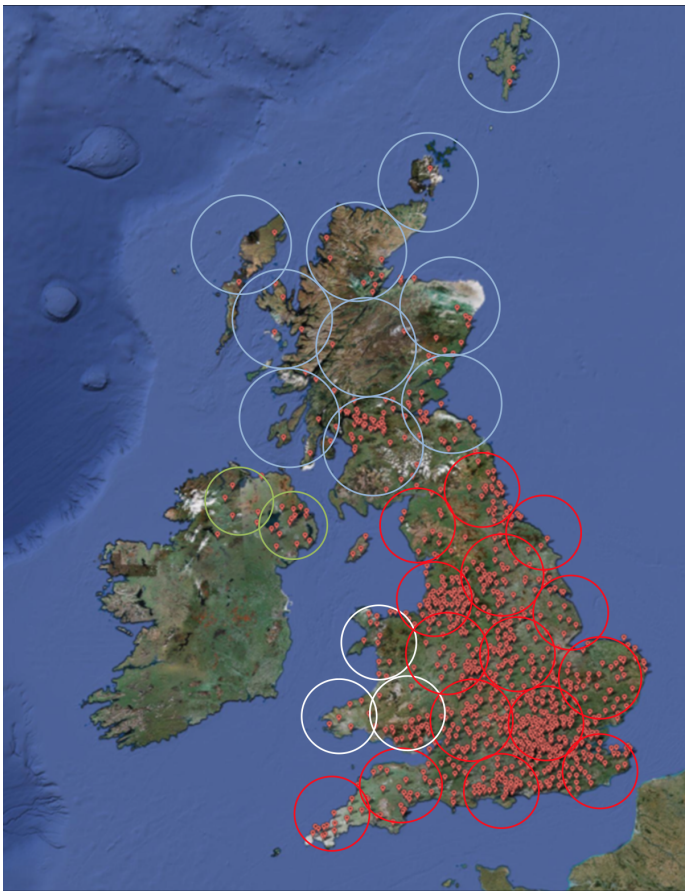


Figure 2.1: 1009 Google targeting geos used for client "B" in geo-experiment. "Consultancy ★" split the set of geos into 30 circular zones, such that 15 were in England, 10 in Scotland, 3 in Wales and 2 in northern Ireland. Zones in each country are all of the same radius. The areas of the zones in England are smaller compared to others, to allow for the density of the population. Copyright 2012 by ''Consultancy ★''.

The zones were then split into two groups: control and treatment groups, called Geo1 and Geo2, respectively, where in Geo 2 the campaign was turned off. Zones are allocated to Geo1 and Geo2 by using an algorithm employing spatial systematic sampling technique. For each country, a distance of each zone from Greenwich point were computed, see **Figure 2.2**. Based on the distance values, zones were ordered from nearest to farthest from Greenwich point. The first zone was randomly assigned to either Geo1 or Geo2 and then zones were iteratively assigned to different Geos. All geo-locations included in zones labelled by Geo1 are control geos and all geo-locations included in zones labelled by Geo2 are treatment geos.

During the first time period, in both groups Geo1 and Geo2, brand keywords were

left on, whereas during the second time period, brand keywords were turned off in Geo2. Individuals in Geo1 were supposed to see"B" PPC Ad link when they searched on Google's search engine, using "B" keyword as a search term. On the other hand individuals in Geo2 were supposed to see "B"'s natural result link when they searched on Google using "B" keyword as a search term.

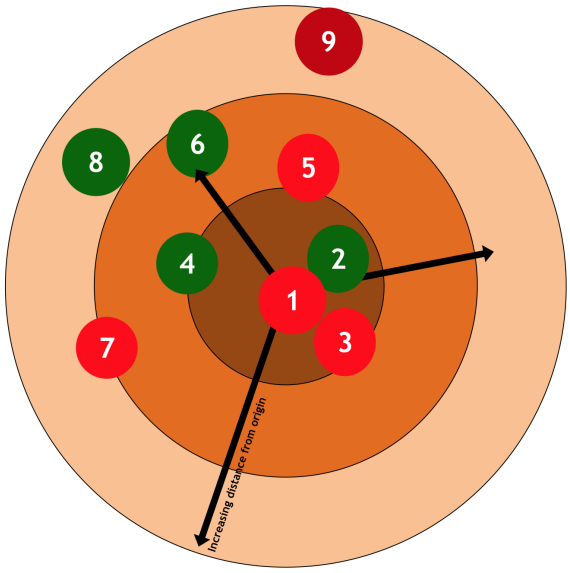


Figure 2.2: Zones ordered in ascending order, from nearest to farthest from the origin point. Zone number 1 assigned to Geo2 randomly and then iterative allocation were applied to assign zones into Geo1 or Geo 2.

The advantage of this design methodology includes eliminating the clustered selection of treatment or control grouped geos. However, this methodology did not explain to us how the distance from Greenwich point affects users behaviour. A possible explanation for this might be that the company was attempting to capture the presence of the spatial variation systematically and using Greenwich point as the origin point can support the design spatially. However, we find it difficult to see how the distance from the origin can affect the users' responses to the advertising campaign. This discussion, though, is farther away from the scope of this research.

2.1.2 Data Description

"Consultancy ★" collected data for "B"'s advertising campaign to test incrementality on Brand keywords. The observed data were provided to us in .RData file. The

data set consists of 94893 observations sorted in 28 variables. Here are the first six rows of the data frame

```
     userid        saleid      orderid   Equity.Device esc   ldate
1 8000307853257 8001446887265  95395026            NA   4 03/07/12
2 8000307853257 8001446887265  95395026            NA   4 06/07/12
3 8000375716656 8001449250050  95872291            NA   1 08/08/12
4 8000339995880 8001455570900 100475383            NA   3 19/08/12
       etst               ets esec Equity.Referrer.Domain ead
1    Natural Search    Natural Search               google.co.uk
2    Natural Search    Natural Search               google.co.uk
3    Natural Search    Natural Search               google.co.uk
4    Natural Search    Natural Search               google.co.uk
     eppckw    mt   Revenue  ekwg   Equity.Is.Brand  saledate
1                   566.1 Non Brand     No          02/08/12
2                   566.1 Non Brand     No          02/08/12
3                   573.0  Brand        Yes         08/08/12
4                   749.0  Brand        Yes         19/08/12
     tst                   geo               City   Area LCW1 LCW2 LCW3
1 Natural Search Non Brand   2             Poplar    1    1    1    1
2 Natural Search Non Brand   2             Poplar    1    0    0    0
3                SEO Brand   1          Gloucester    6    1    1    1
4                SEO Brand   2 Bishop's Stortford    1    0    1    1
   EW1 EW2 EW3   Test
1 1.0 1.0 1.0   Before
2 0.5 0.5 0.5   Before
3 0.5 0.5 0.5   During
4 0.0 0.5 1.0   During
```

The variables in this data set are arranged by "Consultancy ★" based on their own categorisation criteria in defining variables. Some variables here are simple titles from which their contents can be expected such as

```
userid : unique identifying number for a user
saleid : unique identifying number for a item
```

```
orderid: unique identifying number for a order
ldate  : date when a user lands on a retailer's web-page
saledate:date when a user make purchase decision
Revenue : total basket revenue
City: name of a targeting geo-location defined by postcode
Area: Zone code of a targeting geo-location
geo: status of a targeting geo-location (1: Geo1) and  (2: Geo2)
Test: time period, (Before: first period) and (During: second period)
```

Other variables such as (`etst, ets, tst`) describe traffic sources including direct and indirect channels. The channel could be URL, email, social media, banner, natural search or PPC. There are further details of interest in this investigation such as ad groups, keywords, matching type and referrer domain as exhibited below,

```
etst, ets, tst: different categorisations of  traffic sources
ead : equity ad group
eppckw : equity PPC keyword
ekwg  : equity keyword group, classification of keyword by type
Equity.Is.Brand : keyword type is brand or not
mt : keywords matching type type
Equity.Referrer.Domain: URLs that bring traffic to a retailer's web-page
LCW1, LCW2, LCW3, EW1, EW2, EW3: These columns define which row is
 awarded a sale or a proportion of a sale
```

The experiment was conducted into two time periods, where according to the provided data, the number of observations during the first time period "`Before`" is 38525 and during the second time period "`During`" is 26705. There are also 29663 observations assigned to a term "`Ignore`" in the `test` variable. This term indicates there is a problem in the experiment, and thus it is recommended to remove this value from the data set although it represents about 31% of the data. This change reduces the number of the observations to 65230 where about 59% of the observations belong to the first time period and 41% to the second time period.

In the experiment, the targeting geo-locations were assigned into two groups Geo1 and Geo2. The number of zones were 30 in both time periods. The data gives 14 zones assigned to control group and 16 zones assigned to treatment group. In the first time period, there were 920 geo-locations, where 491 of them allocated to Geo1 and 429 allocated to Geo2. In the second time period, however, the number of geo-locations was 446, where 423 were allocated to Geo1 and 423 belong to Geo2. The number of geo-locations utilised in the experiment are supposed to be the same over the two time periods, and the allocation to Geo1 and Geo2 are supposed to be the same across the time periods. Given that the campaign design is based on zones and there is no change in the total of treatment zones or control zones across the two time periods, we will not pay attention to the difference found in the total treatment and control geos between the two time periods. This difference can be related to the targeting criteria that are applied by Google AdWords, such as advertisers have an option to add or drop geos while the campaign is running.

In this experiment, the aim was to investigate the effectiveness of the PPC brand keywords. Therefore we need to measure the incremental revenue gained through PPC channel without bidding on brand keyword. The data contain records attributable to different traffic source, but we focus on PPC traffic channel. Thus, we filter the data by one of equity traffic source variables, say `ets`, `tst` or `etst`. Start with `ets` variable, the tracked traffic sources are summarised in Table 2.1. The table shows 11894 data values that were received from PPC source when10985 came from PPC Google UK and 909 from PPC MSN UK, where both are search engines.

If `etst` is used to summarise tracked traffic sources, there are also 11894 PPC

Affiliates	Banner	Email	Natural Search
7561	6	5219	31585
PPC	Social Media	Unlisted Referrer	
11894	1029	7936	

If `tst` is used to summarise tracked traffic sources, there are also 11894 PPC such that 10856 values were received from "`PPC Brand`" and 1038 were from other "`PPC`" as shown below,

```
Affiliates     Banner       Email      PPC         PPC Brand
  7561           6          5219      1038          10856
SEO Brand   Social Media     Unlisted Referrer     Natural Search
  20102         1029             7936                 166
Natural Search Brand Generic    Natural Search Non Brand
        1721                                9596
```

Traffic Source	Frequency
Affiliate Window	7211
Delicious	2
Email	5219
Facebook.com	1011
Linkshare	350
Natural Search	31585
PPC Google UK("B" Google UK (Google UK))	10985
PPC MSN UK("B")	909
Struq	6
Twitter.com	13
Unlisted Referrer	7936
Wikipedia	1
Yahoo! Answers	2

Table 2.1: Frequency Distribution of Equity Traffic Source

Considering the observations of "`PPC Brand`", we found 553 observations were derived from treatment geo-locations during the second time period. Having traffic through "`PPC Brand`" source during the second time period in Geo2 while pausing brand keyword in those geos during that time is questionable. The explanation of this might be related to technical issues in recording those observations. Another possible explanation is that "`PPC Brand`" classification was categorised by the company to make some keywords easily recognisable but not necessary include Brand keyword "B". The keyword might be confused with brand generic keywords. However, by filtering "`PPC Brand`" data subset using keyword type variable, i.e."`ekwg`", we found 8 terms of "`Non Brand`" category and 7 of them are in Geo1 during the

first time period, which might indicate an anomaly in the data collection operation because these terms are not expected to be in such data subset. Once again, however, this can be related to the company classification.

	Brand	Brand Generic	Non Brand
PPC Brand	10707	141	8

It is apparent that the data provided require deep understanding. However, we should bear in mind that the goal of this chapter is to show an application of geo-experiment to show the challenge in designing the advertising campaign and how data from tracking users can be messy and incorrectly tracked. Thus we leave the details of this data aside and focus on the PPC conversion channel during both time periods to examine the change in the revenue through it when the brand keyword is paused in some geo-locations during the second time period.

Consider the 11894 data values that were received from PPC channel; i.e. "`PPC Brand`" and "`PPC `". This subset data gives 29 zones and 849 geo-locations in the first time period, where 14 zones and 452 geo-locations were allocated to Geo1 and 15 zones and 397 allocated to Geo2. For the second time period, it gives 634 geo-locations in 29 zones, where 390 in 13 zones belong to Geo1 and 244 in 16 zones belong to Geo2. The distributions of neither treatment and control zones nor geos are equal across the two time periods. This can be expected because conversion behaviour through PPC in treatment or control geos or zones is not necessary to occur during both time periods. However, we are interested to see the distribution of the revenue gained from PPC channel in treatment zones or geos during the second time period in comparison with the revenue obtained from the same geos during the first time period. Thus we need to correct the distribution of the treatment and control zones and geos across the time.

To correct the distribution of the geo-locations we took the common location between the two time periods. We found 606 geos such that 373 were Geo1 distributed over 13 zones and 233 were Geo2 distributed over 15 zones. This change diminished the number of the observations to 10611, where 7987 were observed in the first time period and 2624 were observed in the second time period.

It should be noted that while correcting the distributions of geo-locations across the two time periods, we found that some geo-locations have the same city name in `City` although they are located at different zones given in variable `Area`. For example

geo	City	Area	geo	City	Area	geo	City	Area
2	Bangor	16	1	Hamilton	19	2	Newport	28
1	Bangor	29	2	Hamilton	5	1	Newport	12

In addition, we found that the data include duplicate user IDs with identical sale IDs and order IDs, for example

userid	saleid	orderid	Equity.Device	esc	ldate	etst
8000113085138	8001447952472	95651158	NA	5	11/07/12	PPC
8000113085138	8001447952472	95651158	NA	5	11/07/12	PPC
8000113085138	8001447952472	95651158	NA	5	18/07/12	PPC
8000113085138	8001447952472	95651158	NA	5	21/07/12	PPC

ets	esec	Equity.Referrer.Domain
PPC Google UK(B Google UK (Google UK))	Brand	blockedreferrer
PPC Google UK(B Google UK (Google UK))	Brand	google.co.uk
PPC Google UK(B Google UK (Google UK))	Brand	google.co.uk
PPC Google UK(B Google UK (Google UK))	Brand	google.co.uk

ead	eppckw	mt	Revenue	ekwg	Equity.Is.Brand	saledate	tst
Brand||Core B	exact	25	Brand		Yes	05/08/12	PPC Brand
Brand||Core B	exact	25	Brand		Yes	05/08/12	PPC Brand
Brand||Core B	exact	25	Brand		Yes	05/08/12	PPC Brand
Brand||Core B	exact	25	Brand		Yes	05/08/12	PPC Brand

geo	City	Area	LCW1	LCW2	LCW3	EW1	EW2	EW3	Test
1	Compton	12	1	1	1	0.5	0.5000000	0.5	Before
1	Compton	12	0	0	0	0.5	0.3333333	0.0	Before
1	Compton	12	0	0	1	0.0	0.3333333	1.0	Before
1	Compton	12	1	1	0	0.5	0.3333333	0.0	Before

The cases presented above belong to the same user ID who landed on the "B" website on different dates. In this case, we should not count the revenue more than once. Thus any duplicate User IDs with duplicate sale IDs and order IDs should be eliminated. This reduces the data to 7104 observations in 606 geo-locations such that 373 in Geo1 and 233 in Geo2. 5231 were observed in the first time period and 1873 were observed in the second time period. From the marketing point of view, one might expected a decline in the traffic to the retailer's website during the second time period and particularly in Geo2. This, however does not necessarily imply a decrease in the total PPC revenue as mentioned by Blake et al. 2015.

2.1.3 Measuring Incremental Revenue

Given the cleaned data, the revenue made by each case is the response variable of interest in this application. To measure the impact of pausing the PPC brand keyword in Geo2 locations during the second time period, the revenue gained by all individual cases in each geo-location and during each time period needs to be summed up. The distribution of aggregate revenue in each geo-location in both time periods, are illustrated in **Figure 2.3**.

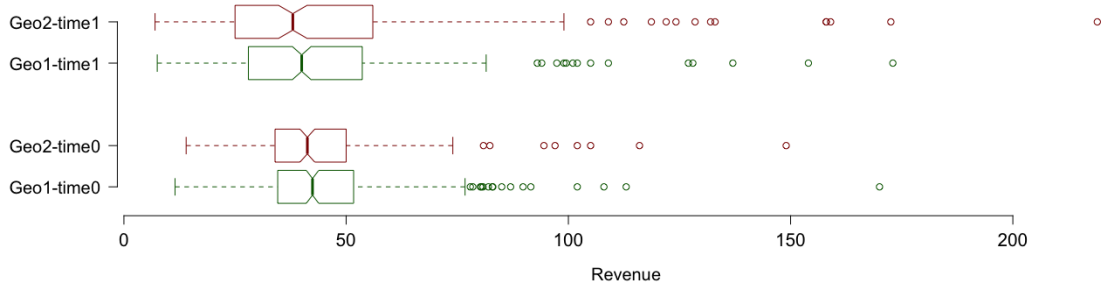


Figure 2.3: PPC Revenue Distribution in treatment geo-locations (Geo2) and control geo-locations (Geo1) during the first time period (time0) and the second time period (time1).

The figure shows that there is a slight difference in the revenue distributions between Geo1 and Geo2 in the first time period, where the revenue is slightly higher in Geo1 compared to Geo2, although all geo-locations during this time period were serving the same advertising campaign. The revenue in both Geo1 and Geo2 spread out more in the second time period compared to what observed in the first time period. Note also that the central revenue value in Geo2 during the second time period

is slightly below the central values of the other distributions. It appears there is a small impact in pausing the PPC brand keyword in Geo2 locations during the second time period. Additionally, the change in the distribution for Geo1 locations during the second time period need to be explained as well.

To measure the incremental revenue, ad spend differential values needs to be calculated to estimates the return on ad spend ROAS for PPC brand keyword. Unfortunately, the ad spend are not available in this data from B, although ad spend should be available at geo level in AdWords. Vaver and J. Koehler 2011 mentioned that the ad spend differential is zero in the control geos if the geos continue to operate at the same baseline level during the second time period. Given this is the case in this "B" 's advertising campaign where brand keywords continue ON in the control geo-locations during the second time period, the ad spend differential is then zero in the control locations. For the treatment geo-locations, however , it is difficult to approximate the ad spend differential with no ad spend information. At the same time, we expect the ad spend in those geos decreased in the second time period because the brand keywords switched OFF. In other words, users made conversions through PPC channel but without bidding on the brand keywords. Therefore the ad spend differential is expected to be negative in the treatment geos but their values are unknown. The distribution of ad spend differential is also unknown to us and beyond the area of this work. Therefore, estimation of the incremental revenue will not be discussed further.

2.2 Summary and Concluding Remarks

The purpose of this chapter was to illustrate a real world application of geo-experiment. We presented the geo-targeting approach used by the online marketing consulting company "Consultancy ★" to their client "B", provided that the company are using this approach to implement an experimental test to measure the incremental sale of online advertising campaigns for several clients.

Focusing on the design strategy that applied to "B"'s campaign, the geo-locations were set into blocks and then systematic sampling techniques based on distance from Greenwich point was applied to assign geo-locations to treatment. The company are

aware of the presence of the spatial autocorrelation, though, by the time we met them, they did not have any evidence that can show the correlation between the geo-locations. Thus they made an attempt to capture the spatial variation by measuring distance of each geo-location from the origin.

In fact, there were other attempts made by the company to control spatial variation. For example, they grouped geo-locations into 24 blocks. Then they split the UK into 10 longitudinal sections. The first block at the left in the first section is assigned randomly to Geo1 or Geo2. The assignment are then selected in a respective order from left to right in each section. **Figure 2.4** illustrates the sampling strategy used by the company.

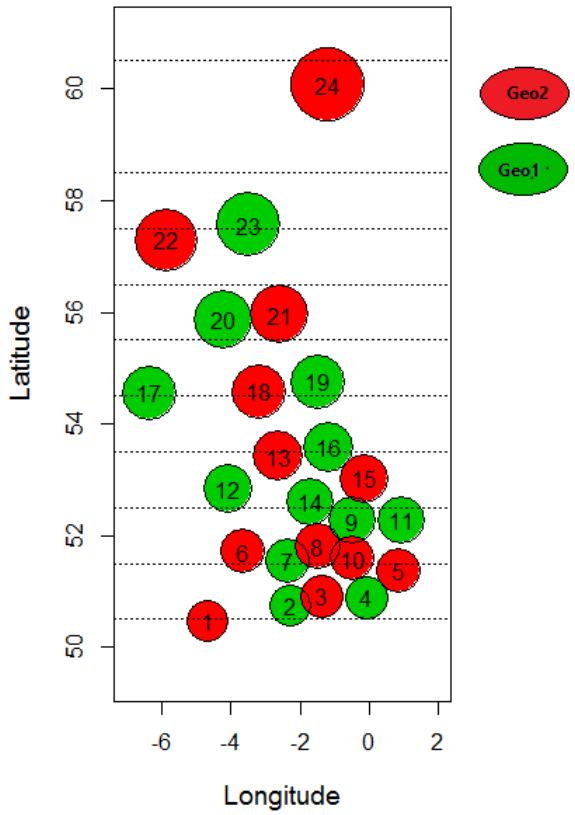


Figure 2.4: Other systematic spatial design strategy implemented by “Consultancy ★”. They grouped the geo-locations into 24 blocks. They split the UK into 10 longitudinal sections. The first block at the left in the first section is assigned randomly to Geo1 or Geo2. The assignment are then selected in a respective order from left to right in each section.

In our view, systematic spatial sampling technique ensures that geo-locations are randomly selected and evenly distributed for treatment and control across the country. However, lack of the spatial feature makes the efficiency of the mentioned strategies difficult to detect. Spatial information such as size of the geo-location, demographic information, geographic searching or purchasing are all necessary to understand how geo-locations are correlated and help to explain potential behaviours of users in particular locations.

At the end of the chapter, revenue at geo-level were discussed briefly. We observed that the variation of the revenue in the second time period is larger compared to the first time period. The high variation can be related to the distribution of the treatment geo-locations or the time factor. In this research, the design strategy is a primary focus in using geo-experiments. In the next chapter, thus, we link UK AdWords target geo-locations with government administrative areas in order to be able to draw demographic inference about them.

Chapter 3

Spatial Units

As mentioned in earlier chapters, the frame of AdWords geos are not well-defined for designing the sampling process for geo-experiments, except DMAs which are available in US. This chapter focuses on linking UK AdWords geos information to background information such as population and social-grades. The linking algorithms is based on the the shortest great-circle distances between target geos and local authority areas which depend mainly on the longitudes and latitudes.

The chapter begins by describing the distribution of the UK AdWords target cities and the local authority areas. Then, we move to outline the allocation process and the linking algorithm which treat the cities and areas as points that characterised by longitudes and latitudes. Then the algorithm is implemented and the shortest distances are calculated between the cities and areas to link cities with some demographic characteristics. Some remarks and conclusions are drawn in the final section.

3.1 UK AdWords Target Geos

AdWords provides various type of geographic targets for the UK, including countries, counties, postal codes, cities and TV regions. TV regions appears to be a good option in the UK since they are supposed to be associated with some demographic and socio-demographic information. However, there are only 15 TV regions which

is a quite small number for applying in the geo-experiments especially given the existence of a wide range of cities beside them. Thus cities might be considered as a reasonable geographic targeting option to focus on in this study. This section presents the distribution of the UK target cities.

It is not obvious how the target cities are determined by Google but they are available on Google AdWords website and can be extracted easily. Google however updates these locations continuously, based on unknown criteria. The cities presented here were provided by the Consultancy ★ and were received directly from Google.

Google provided Consultancy ★ 1015 target cities for the UK including England, Northern Ireland, Scotland, Wales, the Channel Islands and the Isle of Man. Most of the cities 812 are in England, 114 in Scotland, 59 in Wales, 23 in Northern Ireland, 4 in the Channel Islands and 3 in the Isle of Man. The Channel Islands and the Isle of Man are technically not part of the UK, but both are coded under "GB-".

Figure 3.1 depicts the point distribution of the 1015 UK geographic targets. It would be good if these target cities could be drawn as areal polygons but most of the cities are small areas where their boundary data are difficult to obtain. The UK map are plotted here based on the boundaries of the global administrative areas[1], including boundaries of counties, boroughs and districts. The points are mostly located in England.

The 1015 geos are wide range of target locations, which gives a reasonable sample size of geos for geo-experiments. However, the selection of geos for control and treatment group is not straightforward due to lack of covariates such as census data including population, income and socio-economic characteristics that needed for the sampling process. This data, though, is expected to be accessible at the Office for National Statistics (ONS) when AdWords geos are a part of the sites concerned by the government in the country. Most of these locations, however, cover small cities where their related demographic background are either not recorded for some reason or difficult to find within published data. In this case, we suggest to link these

[1] https://gadm.org/download_country_v3.html

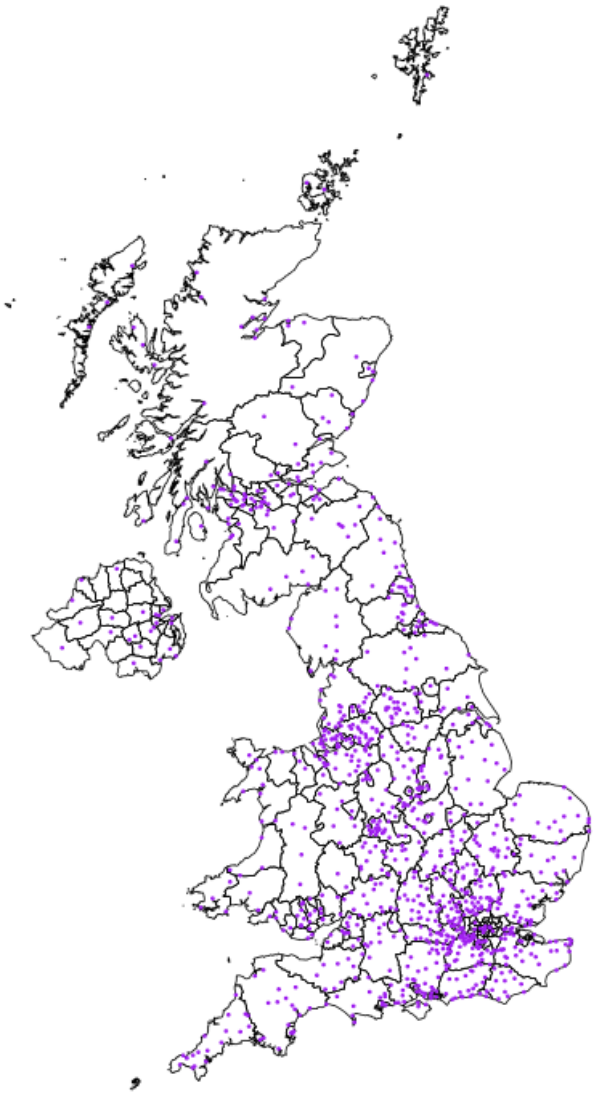


Figure 3.1: 1015 UK AdWords Target Locations

target locations with standard government areas such as administrative areas or local authority areas in order advertisers be able to draw demographic inference about the geos. In this study, we choose to use local authority area to represent the target cities, because the number of local authority areas is more suitable to represent the cities compared to the administrative areas. There are 404 local authority areas and only 192 administrative areas.

3.2 UK Local Authority Areas

Local authority areas are administrative units of local government. Thus include local government districts, council areas, unitary authorities, metropolitan districts, non-metropolitan districts and London borough. Local Authority areas are districts which are sub-divisions of administrative areas. We use these areas to describe the UK AdWords target cities demographically.

There are 404 local authority areas shown in **Figure 3.2**: 324 in England, 32 in Scotland, 22 in Wales and 26 in Northern Ireland. The boundaries of the local authority areas are obtained the UK Data Service -Census Support- website. Official statistics are expected to be available for these areas. The statistics are useful to visualize the similarity and dissimilarity between areas, which is important in order to draw conclusions about people who live in these areas.

Figure 3.2: The 404 UK Local Authority Areas

3.3 Allocation Process

The locations on the earth's surface are situated uniquely by a geographic coordinate system which is a measurement of degrees latitude against degrees longitude. Thus, let $\mathbf{G} = \{g_1, g_2, ..., g_n\}^T$ be n coordinates[2] representing the AdWords target cities within UK and $\mathbf{A} = \{a_1, a_2, ..., a_m\}^T$ be m coordinates representing the local authority. The allocation process relies on computing the shortest distance of each AdWords target city g_i from the local authority area a_j and then assign g_i to a_j that gives a minimum distance value. The shortest distance over the earth's surface is called a great-circle distance (Cassa et al. 2005) and is computed by the "Haversine" formula (3.3) . Let $d(g_i, a_j)$ denotes the distance between the two locations g_i and a_j then:

$$d(g_i, a_j) = \begin{cases} 0 & \text{if } g_i = a_j \\ d_{ij} & \text{if } g_i \neq a_j \end{cases}$$

The zero distance $d(g_i, a_j) = 0$ refers to the fact that the AdWords target city is a local authority area. The d_{ij} is the great-circle distance defined as

$$d_{ij} = 2 \cdot R \cdot atan2(\sqrt{D_{ij}}, \sqrt{1 - D_{ij}}) \tag{3.1}$$

that is

$R =$ earth's radius

$$D_{ij} = \sin^2\left(\frac{\text{lat}_j - \text{lat}_i}{2}\right) + \cos(\text{lat}_i).\cos(\text{lat}_j).\sin^2\left(\frac{\text{long}_j - \text{long}_i}{2}\right)$$

$\text{lat}_i =$ latitude value at location i in set $\mathbf{G}$

$\text{lat}_j =$ latitude value at location j in set $\mathbf{A}$

$\text{long}_i =$ longitude value at location i in set $\mathbf{G}$

$\text{long}_j =$ longitude value at location j in set $\mathbf{A}$

$$atan2(\sqrt{D_{ij}}, \sqrt{1 - D_{ij}}) = \arctan\frac{\sqrt{D_{ij}}}{\sqrt{1 - D_{ij}}}$$

[2] A GPS visualizer tool *GPS Visualizer Tool* 2007 is used to estimate the location of the cities and the areas.

For a specific AdWords target city g_i in G, the great-circle distance is measured over the vector A of local authority areas. This produces a vector $V_{d_{i.}}$ of elements $d_{i.}$. Then the area a_j corresponding to the minimum value of this vector $min(V_{d_{i.}})$ is noted to assign g_i to it.

The set of areas $\{a_j : j = 1, ..., p, \quad p \leq m\}$ that are selected by this minimum distance criterion forms a new set of areas, say $\mathbf{A}'$. $\mathbf{A}'$ represents then local authority areas that are either AdWords target cities or contain at least one of target cities. The number of local authority areas in $\mathbf{A}'$ will be less than the number of geos in $\mathbf{G}$ and less than or equal to the number of areas in $\mathbf{A}$, suggesting that it is not essential that each local authority in $\mathbf{A}$ includes AdWords target cities. If this is the case then we should either ignore the $m - p$ local authority areas that do not contain any AdWords target cities or link these areas to their nearby areas in $\mathbf{A}'$ by again calculating the shortest distance between them. In this study we use the all available local authority areas and hence the latter option is taken into consideration. The allocation process is summarized in the following algorithm.

We use the term *spatial units* to refer to the local authority areas in $\mathbf{A}'$ which is defined here as a local authority area that is an AdWords target city, a local authority area that includes at least one AdWords target city or a local authority area that does not include an AdWords target city but neighbour to a local authority area that includes at least one AdWords target city.

Algorithm 1: Allocation Process

Result: well defined geographic areas set $\mathbf{A}'$

```
1  1. Allocation of AdWords target locations ;
2  initialize a vector A';
3  for i in 1 to n do
4      initialize a vector V_{d_i.};
5      for j in 1 to m do
6          if g_i = a_j then
7              write g_i to A';
8          else
9              compute d_ij;
10             write d_ij to V_{d_i.}
11         end
12     end
13     find min(V_{d_i.});
14     find a_j corresponding to min(V_{d_i.});
15     allocate g_i to a_j;
16     write a_j to A';
17 end
18 2. Allocation of local authority areas not covered by
     AdWords target cities;
19 initialize a vector V_{d_j.};
20 for j in 1 to m do
21     if a_j ∈ A' then
22         End;
23     else
24         for r in 1 to p do
25             compute d_jr;
26             write d_jr to V_{d_j.} ;
27         end
28     end
29     find min(V_{d_j.});
30     find a_r corresponding to min(V_{d_j.});
31     allocate a_j to a_r;
32     write a_j to A';
33 end
```

3.4 UK spatial units

The above mentioned algorithm is implemented for 1015 AdWords target cities and 404 local authority areas. The first step in the allocation process is identifying latitudes and longitudes of both the cities and the areas. A GPS visualizer tool *GPS Visualizer Tool* 2007 is used to estimate the location of the cities and the areas.

The cities are entered in the GPS visualizer alongside their countries because cities from different countries might have the same name, e.g. "Newport, England" and "Newport, Wales". The GPS visualizer does not provide the geographic coordinates of all possible cities that have the same names. It shows the measurements of only one of them chosen arbitrary. In addition, the country itself could include cities with the same names at different locations such as, "Alnwick" in England, see Table 3.1. Therefore cities and areas need to be associated with countries, borough or districts to get the right coordinates. The provided AdWords target cities, however, have no information rather than city names, and so we did take care of any strange results.

City	longitude	latitude	Location
Newham, England	-1.725307	55.549163	England
Newham, London	0.029318	51.53	London
Alnwick, England	-1.70728	55.413541	Northumberland
Alnwick, London	0.037833	51.511928	Alnwick Road

Table 3.1: Example of Common city names within a country

The shortest distances between AdWords target cities and the local authority areas over the earth's surface are calculated for each country separately to overcome any common names issue that may occur. The cities in Channel Islands and the Isle of Man would be excluded from the allocation process.

3.4.1 AdWords Target Cities that are Spatial Units

The names and coordinates of the Adwords target cities in each country - England, Scotland, Wales and Northern Ireland -are compared to the names and coordinates of the local authority areas. The common names with the same coordinates at each country are considered to be spatial units. It is found 160 cities that are local authority areas where 133 in England, 8 in Scotland, 8 in Wales and 11 in Northern Ireland, as shown in **Figure 3.3**.

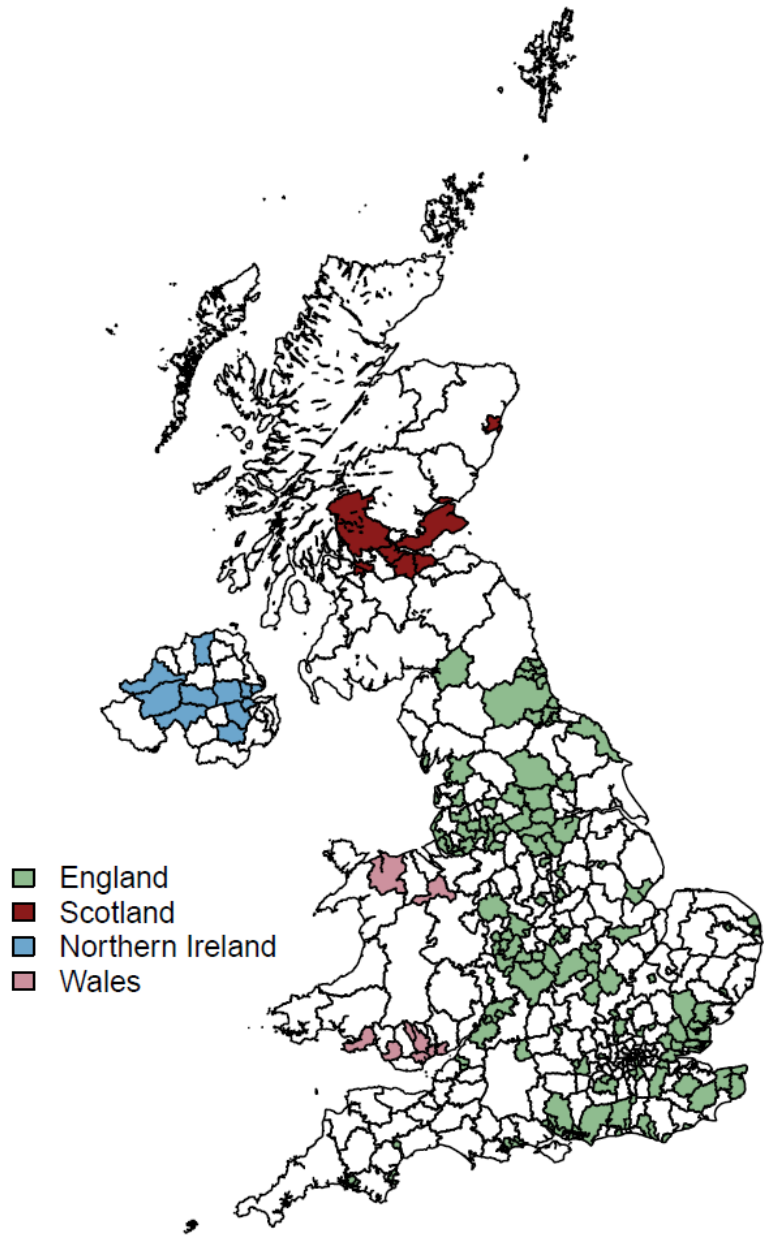


Figure 3.3: 160 AdWords Target Cities that are Local Authority

3.4.2 AdWords Target Cities Not Spatial Units

There are 848 remaining AdWords target cities need to be assigned to the 404 local authority areas. The shortest distance of the remaining cities from the areas is calculated for each country. In England there are 679 out of 812 cities required to be allocated to 324 local authority areas, whereas 106 out of 114 in Scotland, 51 out of 59 in Wales and 12 out of 23 in Northern Ireland need to be allocated in 32, 22 and 26 areas; respectively. The results of the allocation process are summarized in Table 3.2.

	Country			
	England	**Scotland**	**Wales**	**Northern Ireland**
local authority areas (LA)	324	32	22	26
AdWords target cities (cities)	812	114	59	23
cities are LAs (Set 1)	133	8	8	11
cities not LAs	679	106	51	12
Allocation Results				
LAs include at least one city (Set 2)	263	29	21	8
LAs in Set 1 and Set 2	102	6	7	2
LAs in Set 2 but not in Set 1	161	23	14	6
LAs in Set 1 but not in Set 2	31	2	1	9
LAs not in Set 1 nor Set 2	30	1	0	9
Total spatial units	**294**	**31**	**22**	**17**

Table 3.2: Distribution of the cities and areas and the results of allocation process.

Figure 3.4 illustrates sample of cities and their allocated areas. For example, the target city "Lowther" appears to be close to the local authority area "Allerdale" and "South Lakeland" where its distance from the two areas are 35.97761 km and 28.32715 km, respectively. Its nearest area therefore is "South Lakeland". The 679 cities are allocated to 263 areas where 102 of them are in the common set. There are then 31 spatial units out of 133 do not intersect with the 263 areas. This means

that 812 cities in England are allocated to 294 local authorities, and so 30 areas does not contain any of the provided target cities.

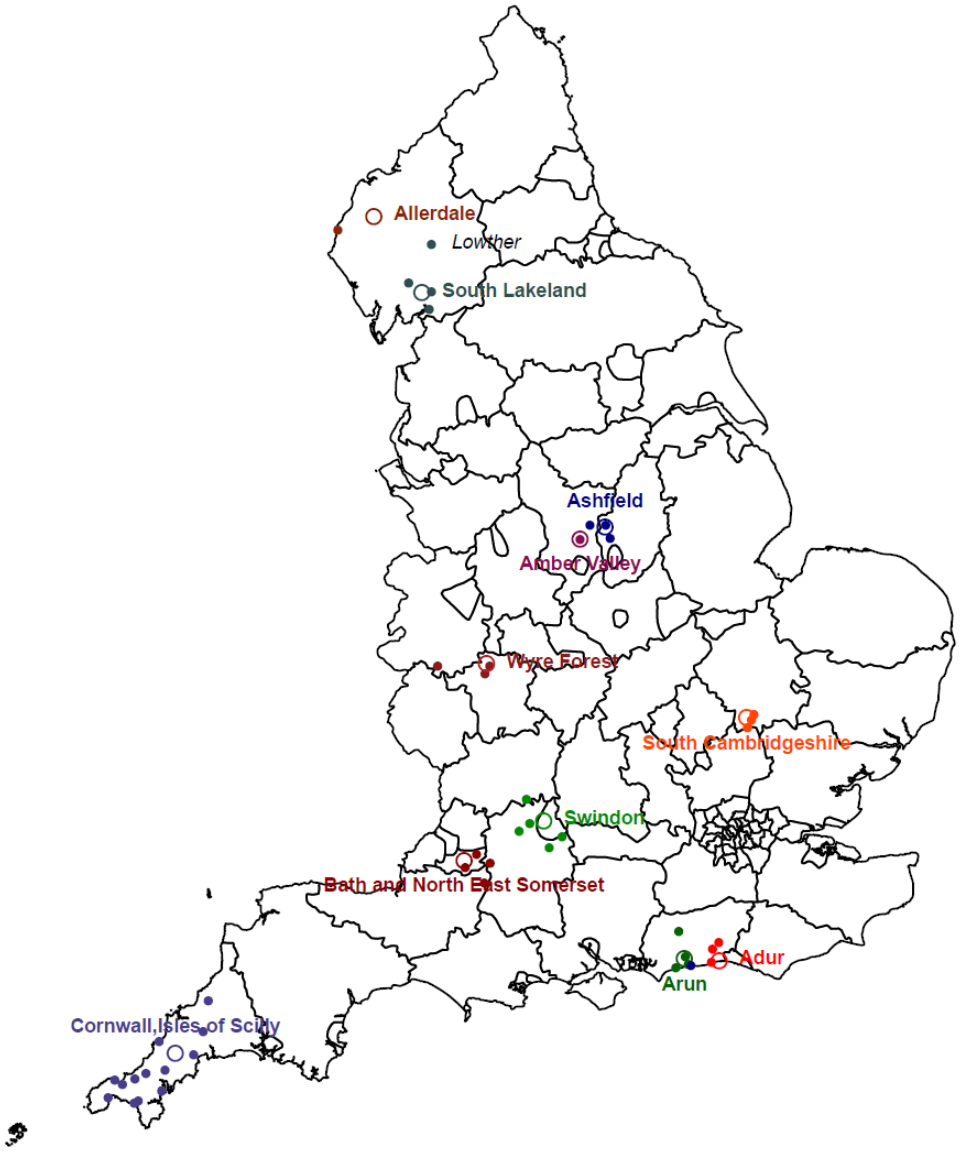


Figure 3.4: Apart of England target cities and their allocation areas

The distance computation are performed similarly for the other three countries Scotland, Wales and Northern Ireland. The 106 target cities in Scotland are allocated to 29 local authority areas, and so the 114 cities are allocated to 31 areas such that 1 areas do not contain any of the target cities. From the table above, the 51 and 12 target cities in Wales and Northern Ireland are allocated to 21 and 8 local authority areas, respectively. By excluding the common spatial units from both countries,

the 59 and 23 cities in Wales and Northern Ireland are allocated to 22 and 17 local authority areas, respectively.

3.4.3 Local Authority Areas Not Spatial Units

The second bottom row in Table 3.2 shows the numbers of local authority areas that are neither in the common spatial units nor contain one of the target city. There are 30, 1 and 9 areas in England, Scotland and Northern Ireland, and none in Wales. These areas can be ignored, treated as spatial units or allocate them to their closest neighbour spatial units. Given that Google AdWords can remove cities from the available list or add a city that belongs to a local authority that is not a part of the our defined spatial units, we think those areas should not be ignored. We decided to allocate them to their nearest spatial units using the earlier mentioned distance criteria. This will ensure sure that all our spatial units will be linked to some of the realistic campaign data that discussed in the previous chapter, i.e. all spatial units will include at least one city.

Half of the 30 areas in England are located in London, where for example "Barking and Dagenham", "Bromley", "Lewisham" and "Newham" are allocated to the spatial unit "Greenwich" as illustrated in **Figure 3.5**. "Waltham Forest" is allocated to its closest neighbour spatial units "Enfield" as it shown on the map. This allocation, however separates "Redbridge" from its closest spatial unit,"Tower Hamlets" which is an allocation area to"Hackney" as well. Combining these areas in one category might be possible but this would constrict their related information and indeed be less consistent compared to the other generated spatial units that are created by this allocation process. The "Waltham Forest" and "Enfield" are considered as one target group and "Redbridge", "Hackney" and "Tower Hamlets" as another group.

The other local authority areas in Scotland and Northern Ireland that are not spatial units are allocated similarly to their nearest units."South Ayrshire" is the only area does not contain any target city and it is allocated to "East Ayrshire". The shortest distance of the 9 remaining areas in Northern Ireland from the 17 spatial units are calculated to be contained in their nearest neighbour units.

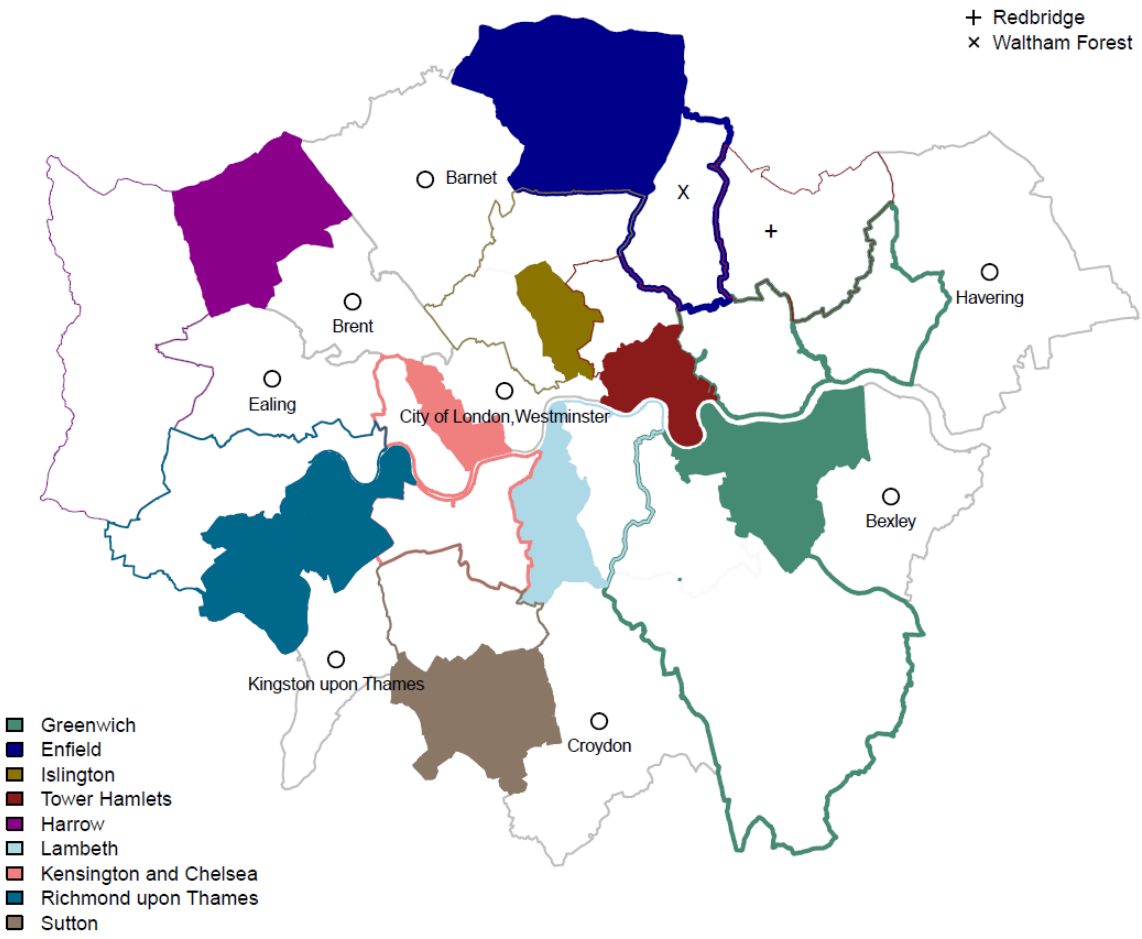


Figure 3.5: London local authority areas and their allocation spatial units

The 40 local authority areas that were not spatial units are now grouped with some spatial units that are spatial units. Their demographic characteristics will be grouped as well. In a map representation, they should be treated as a one polygon by dissolving the borders between areas within one spatial units (Lovelace and Cheshire 2014). Consider Northern Ireland, for example, "Ballymoney", "Limavady" and "Moyle" are local authority areas with no target cities and allocated to their neighbour spatial unit "Coleraine". The attribute of these areas have to be joined spatially and shown as one grouped area, Figure 3.6. The aggregated areas are shaded by the same colour as shown on the map below.

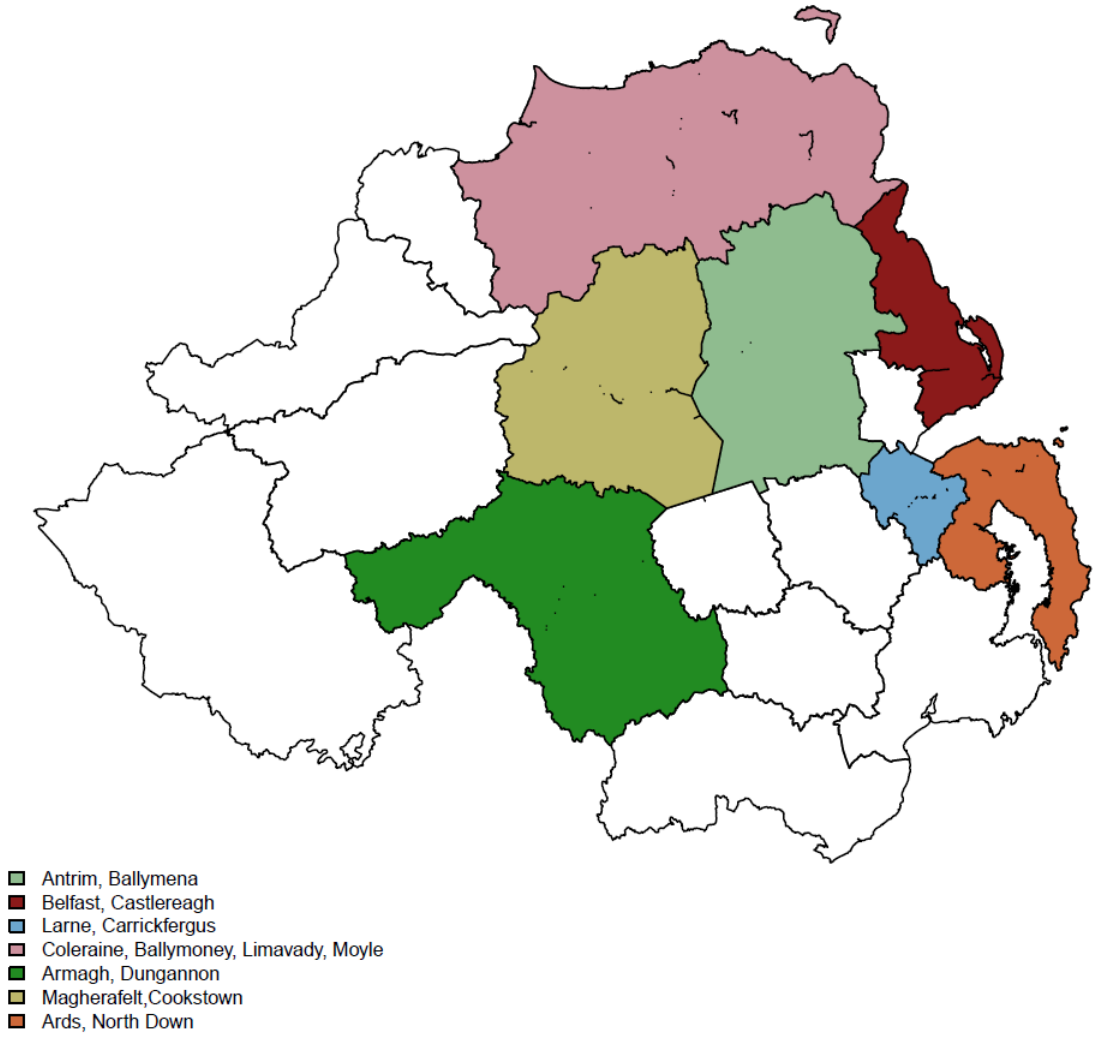


Figure 3.6: The joint polygon of grouped local authority areas in Northern Ireland,

3.5 Assigning Demographics to Spatial Units

Spatial units are formed by the local authority areas. The best source of local authority data is the neighbourhood statistics website which is expected to provide a range of data for very small areas. However, it is found that the socio-economic data at neighbourhood statistics are available individually for some local authority areas with various collection of categorised factors. Thus it is impractical and dispersal of effort and time working on gathering such limited data. The income per head is not available down to the level of local authority areas. Small area income estimates for 2007/08 is available for England and Wales on an interactive map with no recorded

data-frame where each income value needs to be recorded manually at each click on the mapped area. Gross disposable income for $1997 - 2014$ is available for local areas NUTS3[3] which form subset of local authority areas and so again there is limitation on this data at local authority level. In general local authority areas are well defined but their related covariates details are required to compile from different sources.

The UK Data Service provides access to census micro-data. The census micro-data, includes a wide range of individual and household characteristics at a high level of local authority details which is worth to consider it as the main source of the geographical and individuals covariates. The data are available for each census office in England and Wales, Scotland and Northern Ireland. Since most of the spatial units are located in England, the study will focus on published micro-data for England and Wales.

Local authority areas in micro-data are grouped local authority areas. This would present an issue in merging spatial units with micro-data. The dataset contains 265 grouped local authority areas in England and Wales whereas we have $294 + 22 = 316$ spatial units in England and Wales. The grouped local authority areas are grouped to achieve a certain local government purpose. It might be efficient to add detailed grouped areas to one group that is already defined as local authority (spatial units). For example the grouped local authority (Wear Valley and Derwentside) and (Easington and Sedgefield) can be assigned to grouped local authority Chester-lestreet and Durham which is already a spatial unit. This joint process assures that all grouped local authority are considered in the study and be a subset of the spatial units. The joint process provides a total of 255 grouped local authority areas.

[3]Nomenclature of Territorial Units for Statistics 3, Upper tier authorities or groups of lower tier authorities (unitary authorities or districts) (England)(Groups of unitary authorities in Wales, council areas in Scotland, districts in Northern Ireland)

3.6 Summary and Concluding Remarks

In this chapter, the UK AdWords target cities were linked to micro-census data of the local authority areas. The linking algorithm is based on the shortest great-circle distance which depends mainly on longitudes and latitudes of the cities and areas. The algorithm returned a subset of local authority areas that is called here spatial units. The focus in this study is on the spatial units in England and Wales, where the total number of the spatial units in these countries is 255 units.

The longitudes and latitudes were identified by the GPS visualiser tool. The tool, however, is very sensitive in regard to the place name and its location. The provided AdWords target cities have a lack in details about their exact location, and hence their coordinates could be incorrect if the name of the city is available in different locations within a country. The spatial units, therefore, would then be incorrect too. We took care of any obvious wrong results manually but there is always a room for human errors.

Some local authority areas did not contain any target cities, and so they were linked to local authority areas that are spatial units. In other words some spatial units are in form of grouped local authority areas. These spatial units will be related to aggregate characteristics. For example, “Greenwich” is a spatial unit that is a joint of three local authority areas: “Barking and Dagenham”, “Bromley”, “Lewisham” and “Newham”, Figure 3.5. The population associated to this spatial unit, for example, is the sum of the the population of the four grouped areas.

In the coming chapters, the 255 spatial units will be considered. The realistic campaign data that were discussed in the previous chapter are UK AdWords target geos based. The geos will be linked to the 255 spatial units and associated with their micro-census data.

Chapter 4

Conceptual Model of Geo-Experiments

The effectiveness of online advertising campaigns using geo-experiments relies on two primary factors: an ability to design an effective geographically-based campaign and the ability to track ad-spend and behaviours of interest. However, due to practical challenges, these two conditions have been considered tough challenges in the digital marketing domain. In this chapter, a conceptual model of geo-experiments is introduced to understand more completely the key tenets of the behaviour of interest.

The chapter begins by laying out the potential behaviours of interest that are obtained from the geo-experiments. The second part moves on to describe the conceptual model of the geo-experiment and outlines two possible behaviours of interest: online search and purchase. The third section is a description of statistical models structures that will be used to measure the effectiveness of the campaign. The models are constructed under two conditions: homogeneity and unobserved heterogeneity. By the end of third section, a hypothetical measure of the campaign effect is introduced. Then primitive data, which will be used later in this thesis when devising theoretical and simulation methods, are discussed. Some remarks and conclusions are drawn in the final section.

4.1 Potential Behaviours of Interest

Returning briefly to the geo-experiments description provided by Vaver and J. Koehler, the experiments begin with assigning each spatial unit randomly to either treatment or control conditions. The experiments contain two time periods; during the first time period all spatial units receive the same advertising campaign, and during the second, the campaign is modified for the treatment spatial units. The modification could include changing specific campaign components such as changing the keywords list or could be turning off the campaign completely. However, due to potential revenue impact from having no ads in some units while the campaign is running in other units, the advertisers prefer to keep the campaign and display modified ads.

Geo-experiments aim to measure the effectiveness of advertising campaigns in the form of return on ad-spend (ROAS), which is the incremental impact the ad-spend had on the response metric of interest. Response metric refers to individual behaviours of interest such as clicks, online or offline sales and website visits.

The potential behaviours are outlined hierarchically in **Figure 4.1**. Given an advertising campaign served in a specific spatial unit during a certain time period, individuals in that unit and during that time then do online search or do not. If they do online search, then ads may be displayed on the search result page or may not, depending on how well searchers' information and search terms used match campaign's components such as keywords, language and age. Given ads on the search result page, users then either click on the ads or click on organic links. Their click behaviours are followed by conversion behaviours including: click conversion or not click conversion. If it is a click conversion, then click converts to a desired goal such as online sale, newsletter sign-ups, or software downloads through ads channels, i.e. PPC channel. On the other hand, if users do not click on ads or click on them but do not convert their clicks to desired actions, then they could possibly make conversions through other paid channels such as displayed ads on different websites and social media channels, or through organic channels or offline (in stores). At the same time, there is a possibility that they do not make any conversion by any

means.

The depiction of behaviours in **Figure 4.1** is a hierarchical model of multiple outcomes. Advertisers might be interested in multilevel behaviours or in just one behaviour, depending on the main objective of the campaign. The geo-experiment aim to measure the incremental impact of ad-spend on behaviours of interest, however our knowledge of ad-spend or its distribution is very narrow. We have seen earlier in the introduction chapter that tracking ad-spend is a core problem in measuring effectiveness of online advertising campaigns. This problem becomes more difficult with a hierarchical behavioural structure, because both behaviours of interest and ad spend are non-deterministic measures. Therefore, it might be appropriate to neglect ad-spend, to proceed in this research study, despite the fact that ad spend is essential for evaluating the performance of advertising campaigns.

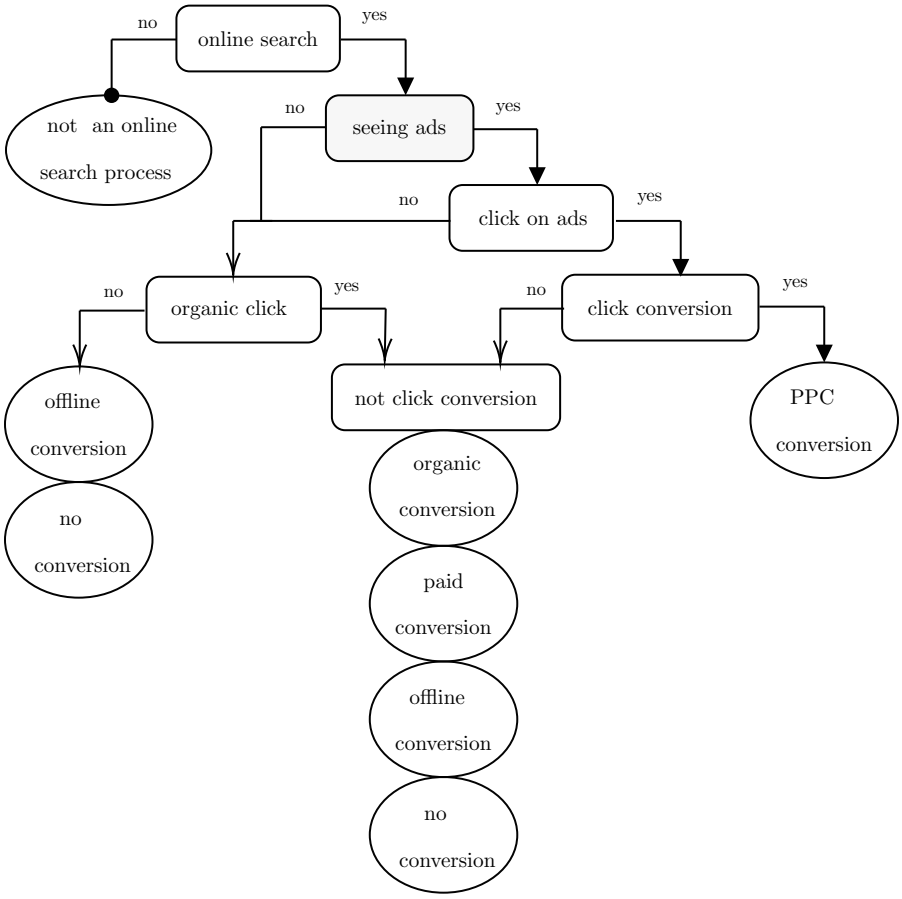


Figure 4.1: Potential Individual Behaviours on Seeing Ads on Search Engine

Measuring the effectiveness of an advertising campaign requires detailed informa-

tion about each behavioural change in each spatial unit. However, recall that the central question in this thesis asks how to assign spatial units to treatment and control groups. Thus, the first thing advertisers need to focus on is how to select or sample a set of spatial units to serve modified advertising campaigns. It is a simple question, yet no easy answer, due to limitation on information and data such as demographics and socio-economics, which are required in this case to permit the use of standard sampling techniques. Therefore, we construct a simple structure of individuals behaviours to understand how allocation of spatial units affects a given behaviour. Before moving on to characterise the simplified structure, it is necessary to make a number of assumptions to make behavioural process well-defined.

4.2 Essential Assumptions

For constructing a simple and well-defined structure of behaviours, we identify four assumptions.

1. **A Conceptual geo-experiment:** We assume advertisers run a standard advertising campaign in all spatial units during the first time period and run a proposed new advertising campaign in some spatial units and the old campaign in other units during the second time period. The experiment aims to investigate the effectiveness of the new advertising campaign relative to the old campaign.

   The basic unit of measurement is the number of sales in each spatial unit in each time period. However, the number of searches is also available on a similar basis. Therefore, the behaviours of interest are the number of searches and the number of which convert to purchases.

2. **All online searchers in each spatial unit are exposed to ads:** The behaviour of interest should be attributable to displayed ads in each spatial unit during both time periods. This depends mainly on the campaign's status in each spatial unit in both time periods. Campaign status or ads exposure does not rely only on setting target spatial units. There are other possible factors that may include some interrelated targeting parameters such as demographics,

content keywords and device types. The competitive bidding among advertisers can also play a considerable role as well in ads exposure. Therefore, we assume that all factors influencing ads exposure are under control and ensure the displaying of ads on search results. Hence, all online searchers in each spatial unit are exposed to ads relative to its campaign status.

3. **All online searchers in each spatial unit see ads:** We are aware that showing ads on search results pages does not necessarily indicate that users see the ads. Thus, we assume as long as ads are presented on the results pages this is equivalent to seeing the ads.

4. **Online conversion includes PPC, organic or paid conversions:** We believe that PPC channel is not sufficient to count sales resulting from paid search ads due to the fact that seeing ads on the results pages could impact searcher behaviour both explicitly and implicitly. We assume therefore that seeing ads can lead to complete conversion through PPC, organic or paid channels.

4.3 Conceptual Behaviours of Interest

Following the assumptions mentioned above, the behaviours of interest are summarised in **Figure 4.2** as a two-level model: online search and online purchase. From now on, we work on this two-stage behavioural process. Hence we should observe the number of searches and the number of purchases in each spatial unit during two time periods.

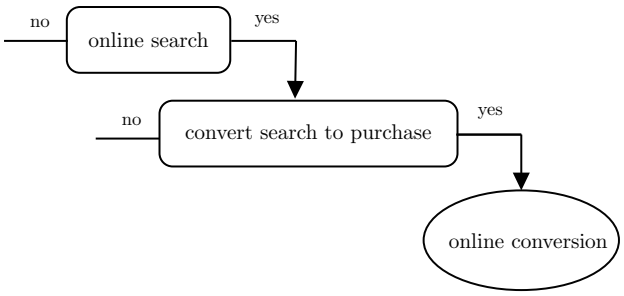


Figure 4.2: Conceptual Behaviour of Interest.

Now we continue with the original objective of the experiment, that is, to estimate the effectiveness of the advertising campaign, which is the differential in the pur-

chasing behaviour in the treatment spatial units relative to the control units.

4.4 Statistical Modelling Framework

For a spatial unit i, in a time period t we need to model the number of searches and the number of searches converted to purchases. The search-purchase process is a conditional stochastic process where purchases decision relies on a search model.

We suppose that a spatial unit is composed of a set of population-strata; that could represent a group of demographic and socio-economic characteristics. The search and purchase probabilities are likely to differ from one stratum to another within each spatial unit in a time period. Thus, the observational unit for the search-purchase process should be represented by a population-stratum k in a spatial unit i during a time period t. The sets of possible spatial units and population-strata are respectively $\{1, \ldots, I\}$ and $\{1, \ldots, K\}$. The time interval $t \in \{0, 1\}$.

For the search process, let s_{rikt} represent the absence or presence of online search for individual r in observational unit ikt, such that

$$s_{rikt} = \begin{cases} 1 & \text{if an individual } r \text{ in an observational unit } ikt \text{ searches} \\ 0 & \text{Otherwise} \end{cases}$$

Given N_{ik} individuals in a stratum k in a spatial unit i, then for $r \in \{1, \ldots, N_{ik}\}$ we view s_{rikt} as independent realisations of a random variable S_{rikt}, each has a Bernoulli distribution with probability φ_{rikt}, such that

$$P(S_{rikt} = 1) = \varphi_{rikt}$$

On a similar basis, let y_{rikt} represent the absence or presence of purchase conversion, such that

$$y_{rikt} = \begin{cases} 1 & \text{if individual } r \text{ in observational unit } ikt \text{ converts to purchase} \\ 0 & \text{Otherwise} \end{cases}$$

and given the s_{rikt}, we view y_{rikt} as independent realisations of Bernoulli random variables Y_{rikt}, where

$$P(Y_{rikt} = 1 \mid S_{rikt} = 1) = p_{rikt}$$

$$\text{and} \qquad P(Y_{rikt} = 1 \mid S_{rikt} = 0) = 0$$

For all individuals in the observational units, we could in principle observe a vector of independent pair of observations (y_{rikt}, s_{rikt}). In geo-experiments, observations however, are aggregates, i.e. the sum of all individual outcomes in each observational unit. The aggregated search and purchase are given by

$$s_{ikt} = \sum_r s_{rikt}$$
$$y_{ikt} = \sum_r y_{rikt}$$

and we assume that individuals in a spatial unit i at time t in stratum k all have the same probabilities for search and purchase but individuals in different strata within i have different probabilities; i.e. that $p_{rikt} = p_{ikt}$ and $\varphi_{rikt} = \varphi_{ikt}$.

In practice, however, population-strata, i.e. demographic and socio-economic covariates, to the best of our knowledge, are missing - unobserved - in the paid search advertising database, although these covariates are likely to affect the probability of making purchases. The distributions of some covariates are known but it is likely that there are other important unknown and unobserved covariates. This means that the heterogeneity between population-strata within spatial units is unobserved.

Therefore, what are observed are s_{it} and y_{it} that are based only on a whole spatial and a time, although the true observations are summed implicitly over some unobserved or unknown strata, i.e.

$$\begin{aligned} s_{it} &= \sum_k s_{ikt} \\ y_{it} &= \sum_k y_{ikt} \end{aligned} \tag{4.1}$$

Having defined the search and purchase observations, we now move on to discuss their statistical model structures under the two conditions: homogeneity and heterogeneity within spatial units.

4.4.1 Search Model

Assuming homogeneity within spatial units, S_{it} has a Binomial distribution with parameters N_i and φ_{it} where N_i is the size of the population in a spatial unit i

and probability φ_{it} depends on the spatial unit i and time period t. We change the notation of the probability to φ^*_{it} to distinguish between the probabilities of observations in the model assuming homogeneity and those in the model based upon heterogeneous strata. The search process is then given by

$$S_{it} \sim \text{Bin}(N_i, \varphi^*_{it}) \tag{4.2}$$

or equivalently,

$$P(s_{it}) = \binom{N_i}{s_{it}} \varphi^{*\,s_{it}}_{it} (1 - \varphi^*_{it})^{N_i - s_{it}} .$$

The probability φ^*_{it} of search can be then modelled using a logit regression model, for example

$$\text{logit}(\varphi^*_{it}) = \zeta^*_{it} = \nu^*_i + \xi^* t \,. \tag{4.3}$$

The probability of search is based on a spatial effect ν^*_i and a temporal change effect ξ^*. The model (4.3) can be formulated in a matrix notation such that

$$\zeta^*_{it} = \sum\nolimits_j x^{s*}_{it\,j} \vartheta^*_j \Longleftrightarrow \zeta^* = X^{s*} \vartheta^* \tag{4.4}$$

where X^{s*} is a design matrix of the search model and $\vartheta^* = \begin{bmatrix} \nu^*_1 & \dots \nu^*_I & \xi^* \end{bmatrix}^T$.

If we do not assume homogeneity, the aggregates s_{it} has then aggregated Binomial probability structure given by

$$S_{it} = \sum_k S_{ikt}, \quad \text{where} \quad S_{ikt} \sim \text{Bin}(N_{ik}, \varphi_{ikt}) \quad \text{are independent} \,. \tag{4.5}$$

The logit probability φ_{ikt} is modelled by

$$\zeta_{ikt} = \sum\nolimits_j x^s{}_{iktj} \vartheta_j \Longleftrightarrow \zeta = X^s \vartheta \tag{4.6}$$

where ζ_{ikt} is a search model that relies on k strata. For example

$$\text{logit}(\varphi_{ikt}) = \zeta_{ikt} = \nu_i + \tau_k + \xi t$$

where ν_i is the spatial effect ν_i, τ_k the population-stratum effect and ξ the temporal change effect.

4.4.2 Purchase Model

Following assumption of homogeneity, $Y_{it} \,|\, (S_{it} = s_{it})$ has a Binomial distribution with parameters s_{it} and p^*_{it}. The purchase process is a conditional model such that

$$Y_{it} \,|\, S_{it} \sim \text{Bin}(s_{it}, p^*_{it}), \quad \text{where} \quad S_{it} \sim \text{Bin}(N_i, \varphi^*_{it}) \tag{4.7}$$

or equivalently,

$$P(y_{it} \,|\, s_{it}) = \binom{s_{it}}{y_{it}} {p^*_{it}}^{y_{it}} (1 - p^*_{it})^{s_{it} - y_{it}} \,.$$

The probability p^*_{it} of purchasing given the searching process with probability φ^*_{it} is required to be attributable to the campaign status in each spatial unit. Let u_i represent the condition of a spatial unit i, i.e. control or treatment , then

$$u_i = \begin{cases} 0 & \text{if } i \in \text{control group} \\ 1 & \text{if } i \in \text{treatment group} \end{cases}$$

Let C_{it} be an indicator of an advertising campaign status such that

$$C_{it} = \begin{cases} 1 & \text{if } u_i = 1 \text{ and } t = 1 \\ 0 & \text{Otherwise} \end{cases} \tag{4.8}$$

where $C_{it} = 0$ indicates that a spatial unit i in time period t is serving the original advertising campaign, whereas $C_{it} = 1$ indicates that a spatial unit i in time period $t = 1$ is serving the new advertising campaign. To put it succinctly, the advertising campaign status is determined by $C_{it} = u_i t$.

The probability p^*_{it} of purchasing is modelled using a logit regression model, such that

$$\text{logit}(p^*_{it}) = \eta^*_{it} = \alpha^*_i + \beta^* t + \delta^* C_{it} \,. \tag{4.9}$$

The probability of purchase is based on a spatial effect α^*_i, the temporal change effect β^* and the campaign effect δ^*. In a matrix notation, the model (4.9) can be written as

$$\eta^*_{it} = \sum\nolimits_j x^*_{itj} \theta^*_j \iff \eta^* = X^* \theta^* \tag{4.10}$$

where X^* is a design matrix of the purchase model and $\theta^* = \begin{bmatrix}\alpha_1^* & \dots \alpha_I^* & \beta^* & \delta^*\end{bmatrix}^T$.

In this thesis, we use a term *applied model* to refer to the statistical model given by (4.9) and (4.10), and hence we refer to its parameters θ^* as the *applied model parameters*. The term applied model probability structure is then used to refer to the probability distribution (4.7).

If we do not assume homogeneity, the aggregates y_{it} have then aggregated Binomial probability structures given by

$$Y_{it} = \sum_k Y_{ikt}, \quad \text{where} \quad Y_{ikt} \,|\, S_{ikt} \sim \text{Bin}(s_{ikt}, p_{ikt}) \quad \text{are independent}\,. \tag{4.11}$$

The probability p_{ikt} of purchasing is modelled using a logit-linear regression model such that

$$\text{logit}(p_{ikt}) = \eta_{ikt} = \alpha_i + \gamma_k + \beta_k t + \delta_k C_{it}\,. \tag{4.12}$$

The probability of purchase is based on a spatial effect α_i, the population-stratum effect γ_k, the temporal change effect β_k and the campaign effect δ_k.

Using matrix notation, the model (4.12) can be written as

$$\eta_{ikt} = \sum\nolimits_j x_{ikt_j}\theta_j \iff \eta = X\theta \tag{4.13}$$

In this thesis, we refer to this model as the *truth* and to its parameter vector θ as the *truth parameter*. Hence the term truth probability structure is used to refer to the probability distribution (4.11).

The truth parameters should be estimated using maximum likelihood estimation method. There would be a need to impose a constraint on either the α_i or the γ_k when fitting the truth model in order to ensure identifiability. However, this does not arise in practice in the thesis because the truth is always specified rather than estimated and only the applied model is fitted to data. This is because the covariates of the population-stratum in each spatial unit are unobserved which means that we can not fit the truth model directly. Thus, instead we explore the consequences of using a misspecified model.

It might be worth mentioning here that there could be some benefit to spatial random effects modelling, especially when one is trying to build a predictive model for unobserved spatial units. However, such predictions are not the real interest here which is in estimating the effects of advertising campaigns. Moreover, the number of searches per spatial unit is generally quite large and so the spatial effect parameter α_i is well estimated. Spatial random effects for the time and campaign parameters might help improve the modelling of data given a particular design but would only indirectly address the issue of unobserved covariates and in particular their consequences for design of geo-experiments, the primary focus of the thesis. For these reasons, spatial random effects were not considered further in the thesis but might well be an interesting line of future enquiry.

4.5 Misspecification in Applied Model

Given the applied model probability structure (4.7) and truth probability structure (4.11), the observed (generated) search and purchase can be illustrated in **Figure 4.3**.

Heterogeneous Strata within a Spatial Unit

spatial unit	stratum	time	population	prob.search	# search	prob.purchase	# purchase
i	1	0	N_{i1}	φ_{i10}	s_{i10}	p_{i10}	y_{i10}
		1	N_{i1}	φ_{i11}	s_{i11}	p_{i11}	y_{i11}
	2	0	N_{i2}	φ_{i20}	s_{i20}	p_{i20}	y_{i20}
		1	N_{i2}	φ_{i21}	s_{i21}	p_{i21}	y_{i21}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	K	0	N_{iK}	φ_{iK0}	s_{iK0}	p_{iK0}	y_{iK0}
		1	N_{iK}	s_{iK1}	s_{iK1}	p_{iK1}	y_{iK1}

Homogeneous Strata within a Spatial Unit

spatial unit	time	population	prob.search	# search	prob.purchase	# purchase
i	0	N_i	φ^*_{i0}	s_{i0}	p^*_{i0}	y_{i0}
	1	N_i	φ^*_{i1}	s_{i1}	p^*_{i1}	y_{i1}

Figure 4.3: An illustration of the breakdown of observed data purchases given search for a spatial unit i with K population-strata in two time periods, using applied model probability structure and the truth probability structure.

Considering the unobserved covariates of the population-strata, i.e. heterogeneous strata within spatial units, the applied probability structure is misspecified. However, the applied model may be fitted to data using the logistic regression model (4.9), where parameters are often estimated using maximum likelihood estimation technique.

In reality, the underlying probability structure generating the data is unknown. However, in this thesis we pretend the truth probability structure (4.11) is the underlying structure. As a result, the applied model (4.9) is misspecified. Therefore, the presence of observed sources of heterogeneity between strata within spatial units means that statistical findings based on the applied model data may be incorrect. The question that then naturally arises: how sensitive are results to the misspecification of the applied model.

The theoretical consequences of misspecification of the applied model will be discussed in the next chapter. The effect of misspecification can be examined by measuring the distance between the applied model and truth. The distance between the two models is measured using the Kullback-Leibler divergence, which gives the amount of information lost when the probability distribution of the observations is specified incorrectly. It is therefore an essential to fit the truth.

In this research, specification of the truth parameters is taken to tackle the concern about fitting truth. Some of which are based on realistic data as we will see in section 4.7 and others of which must be specified to create an instance of the truth model. A set of interesting truth instances are created in chapter 6.

Estimation of the effectiveness of advertising campaigns under misspecification will be discussed in the forthcoming chapters through a proposed theoretical framework and a computer experiment, for a specific truth parameters and a particular campaign design.

4.6 An Overall Measure of Campaign Effect

In the truth model (4.12), the campaign effects δ_k are strata based, which could be considered as non-functional measures of the effectiveness of advertising campaigns

from a digital marketing perspective. Thus, an overall campaign effect needs to be quantified in a meaningful way for advertisers. Practically thinking, advertisers are more interested in sales they gain from running campaigns, so we need to express overall campaign effects in terms of sales.

Consider purchases $y_{it} = \sum_k y_{ikt}$, the total sales are then just $\sum_{i,t}\sum_k y_{ikt}$. Let Δ be a hypothetical differential measure that measures the difference between the expected total number of sales if the new campaign are served in all spatial units i and the expected total number of sales if it serves in none. Let c^1 and c^0 denotes the conditions of spatial units, where c^1 denotes that the new campaign is served in all spatial units, and c^0 denotes that none of the units are selected to serve it. Then the total sales with conditions c^1 and c^0, are labelled by $y_{it}^{c^1}$ and $y_{it}^{c^0}$ respectively, such that

$$\begin{aligned} Y_{it}^{c^1} &= \sum_k Y_{ikt}^{c^1}, \quad \text{where} \quad Y_{ikt}^{c^1} \,|\, S_{ikt}^{c^1} \sim \text{Bin}(s_{ikt}^{c^1}, p_{ikt}^{c^1}) \quad \text{and} \quad S_{ikt}^{c^1} \sim \text{Bin}(N_{ik}, \varphi_{ikt}^{c^1}) \\ Y_{it}^{c^0} &= \sum_k Y_{ikt}^{c^0}, \quad \text{where} \quad Y_{ikt}^{c^0} \,|\, S_{ikt}^{c^0} \sim \text{Bin}(s_{ikt}^{c^0}, p_{ikt}^{c^0}) \quad \text{and} \quad S_{ikt}^{c^0} \sim \text{Bin}(N_{ik}, \varphi_{ikt}^{c^0}) \,. \end{aligned} \tag{4.14}$$

The effectiveness of the new campaign is then measured using a differential measure Δ that is given by

$$\begin{aligned} \Delta &= \text{E}\Big[\sum_{i,t}\sum_k Y_{ikt}^{c^1}\Big] - \text{E}\Big[\sum_{i,t}\sum_k Y_{ikt}^{c^0}\Big] = \sum_{i,k,t} \text{E}[S_{ikt}^{c^1}] p_{ikt}^{c^1} - \sum_{i,k,t} \text{E}[S_{ikt}^{c^0}] p_{ikt}^{c^0} \\ &= \sum_{i,k,t} N_{ik}\varphi_{ikt}^{c^1} \frac{e^{\sum_j x_{iktj}^{c^1}\theta_j}}{1 + e^{\sum_j x_{iktj}^{c^1}\theta_j}} - \sum_{i,k,t} N_{ikt}\varphi_{ikt}^{c^0} \frac{e^{\sum_j x_{itj}^{c^0}\theta_j}}{1 + e^{\sum_j x_{iktj}^{c^0}\theta_j}} \,. \end{aligned} \tag{4.15}$$

We call Δ an overall true effect. It is a function of a specified truth parameter vector θ. In a similar way we compute an overall applied model effect, say Δ^*. By using the applied model probability structure, we have

$$\begin{aligned} Y_{it}^{c^1} &\sim \text{Bin}(s_{it}^{c^1}, p_{it}^{*c^1}), \quad \text{where} \quad S_{it}^{c^1} \sim \text{Bin}(N_i, \varphi_{it}^{*c^1}) \\ Y_{it}^{c^0} &\sim \text{Bin}(s_{it}^{c^0}, p_{it}^{*c^0}) \quad \text{where} \quad S_{it}^{c^0} \sim \text{Bin}(N_i, \varphi_{it}^{*c^0}) \,. \end{aligned} \tag{4.16}$$

where $p_{it}^{*c^0}$ and $p_{it}^{*c^1}$ are functions of a parameter vector θ^*. Hence

$$\begin{aligned}\Delta^* &= \mathrm{E}\Big[\sum_{i,t} Y_{it}^{c^1}\Big] - \mathrm{E}\Big[\sum_{i,t} Y_{it}^{c^0}\Big] = \sum_{it} \mathrm{E}[S_{it}^{c^1}]p_{it}^{*c^1} - \sum_{it} \mathrm{E}[S_{it}^{c^0}]p_{it}^{*c^0} \\ &= \sum_{i,t} N_i \varphi_{it}^{*c^1} \frac{e^{\sum_j X_{itj}^{*c^1}\theta_j^*}}{1+e^{\sum_j X_{itj}^{*c^1}\theta_j^*}} - \sum_{i,t} N_i \varphi_{it}^{*c^0} \frac{e^{\sum_j x_{itj}^{*c^0}\theta_j^*}}{1+e^{\sum_j x_{itj}^{*c^0}\theta_j^*}}\,.\end{aligned} \tag{4.17}$$

In the next chapters, the difference between the overall true effect Δ and estimates of overall applied model effect Δ^* will be computed. The differential effect between these two quantities will be used to quantify the error associated with estimated applied model parameters relative to specified truth parameters.

4.7 Primitive Data

In order to study the effectiveness of advertising campaigns using statistical models suggested above through theoretical and simulation approaches that will be discussed in the forthcoming chapters, certain primitives need to be identified in order to ensure that subsequent procedures required by each approach can be implemented. The primitives include five elements: investigation of realistic searches and purchases, identification of spatial units, specification of truth parameters, identification of population-strata and specification of campaign design strategies.

4.7.1 Investigation of Realistic Searches and Purchases

Consider the geo-experiments application, that was run by "Consultancy ★" for a retail company "B" , to check whether bidding on brand keywords can drive incremental sales. We have discussed "B"'s dataset in chapter 2, it contained 65230 observations obtained from 820 Google targeting geo-locations, in which 491 are assigned to a control group and 429 to a treatment group. The number of observations during the first time period is 38525 and during the second time period is 26705. The dataset represents individuals who convert their search to purchases through different traffic channels, with no information about actual search behaviours. With some adjustment to "B"'s dataset, though, we adopt the dataset to estimate realistic probabilities of search and purchases.

We postulate that"B"'s dataset represents individuals who search online and that

PPC[1] conversion channel is the channel of interest to observe the number of those who convert their search to purchases. Therefore, we first aggregate the 65230 observations to summarise the number of searches in each geo in each time period. Within the same dataset of 65049 observations, we aggregate the 11894 data values that were received from PPC channel to summarise the number of purchases in each geo in each time period. Both datasets are summarised by unique user IDs and the common geo-locations in the two time periods. Now we have two data sets of realistic information about search and purchases associated with geo-locations and two time periods.

Having realistic data about search and purchases in geo-locations, the probability p^*_{it} of purchasing in each spatial unit and in each time period can be estimated using a purchase applied model η^*_{it} as a fitting model. Hence, estimates, in particular, estimates of spatial effects can be used later in truth parameter specification. However, we need to draw our attention that data are based on geo-locations that are not well-defined spatial units. Thus before fitting purchase data, we need to map geo-locations to the spatial units that defined in the previous chapter, by employing the mapping distance based procedure .

4.7.2 Identification of Spatial Units

In the previous chapter, we have adjusted grouped local authority areas used in a micro-census data in England and Wales, to link spatial units to the micro-census data. Hence, micro sample data is available and associated to 255 spatial units in England and Wales. The geo-locations in the realistic search and purchases data sets are combined and mapped into 205 spatial units in England and Wales, see Appendix C. At the end, the micro sample individuals and the realistic information are merged by spatial units. The 205 spatial units are depicted on the map in **Figure 4.4** based on the company campaign design for "B". From now on we work on these 205 spatial units in England and Wales but by applying different campaign design strategies.

[1]Conversions can be through paid or organic channels, but due to errors related to data tracking and collection found in 'B"'s dataset, we chose to work on PPC channel.

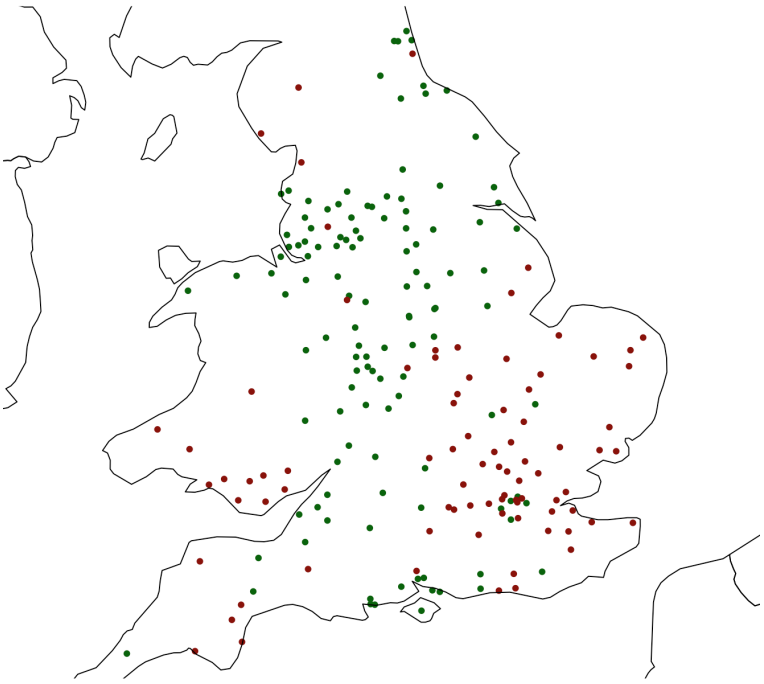


Figure 4.4: 205 spatial units in England and Wales used for "B" advertising campaign design, where red points represent spatial units in treatment group and green points represent spatial units in control group

4.7.3 Specification of Truth Parameters

Provided the realistic data of search and purchase corresponding to each spatial unit for two time periods, the probability of search φ^*_{it} and probability of purchasing p^*_{it} in a spatial unit i in a time period t can be estimated using models (4.3) and (4.9), respectively. These probabilities will be used to estimate spatial effect parameters ν^*_i in the search model and α^*_i in the purchase model.

In this study, however, the focus will be on the truth parameters of the purchase model α_i, γ_k, β_k and δ_k. Truth spatial effect α_i is assumed to be known, while other parameters need to be specified to create some truth instances. Therefore, we will fit the realistic search and purchase data using the applied model (4.9) to estimate α^*_i. The obtained estimates will be will be used to represent the truth parameter α_i. The spatial effect values are presented in Appendix F.

The other parameters γ_k, β_k and δ_k are strata based. To specify them, it may be appropriate to investigate the availability of possible covariates in the micro-census data.

4.7.4 Identification of Population-Strata

It is thought that the most important covariates that can help to create heterogeneous strata are geographical covariates such as gross disposable income per head, socio-economic status. Individual covariates such as life-style, gender, age and education can be considered useful as well. However, in the micro-census data, these covariates are either not included, have a lot of missing values or lack detailed information with categorisations. On the other hand, such data are usually not available in digital marketing databases.

However, the household reference person social grade is collected for individuals in each spatial unit is available in the micro-census data. The social grade is usually classified into six groups A (upper middle class), B (middle class), C1(lower middle class), C2(skilled working class), D(working class) and E(non working). In the micro-census data, the social grade is classified into four groups AB, C1, C2 and DE, where each group is coded by a number AB=1, C1=2, C2=3 and DE=4. There is a list of individuals in each spatial unit who do not belong to any of these categorisations. This could be an error in data collection and sampling. Those individuals is grouped into NA group. The percentage of individuals in each group is computed for each spatial unit to see the distribution of social grade in each unit. The distribution of each social grade is illustrated in **Figure 4.5**.

The AB grade is distributed about the same percentage in the most spatial units, except spatial units around London, where AB grade represents about (30-40)% of the total individuals in each unit there. For the other grades except NA, spatial around London have lower percentage compared to the other units. It seems the distribution of the social grades varies substantially between spatial units. This indicates that the social grades might have some influence on online purchasing. Therefore, it is of interest to use the social grade as a blocking covariate to create strata of within each spatial unit to investigate the impact of unobserved heterogeneity within spatial in estimating the advertising campaign effect.

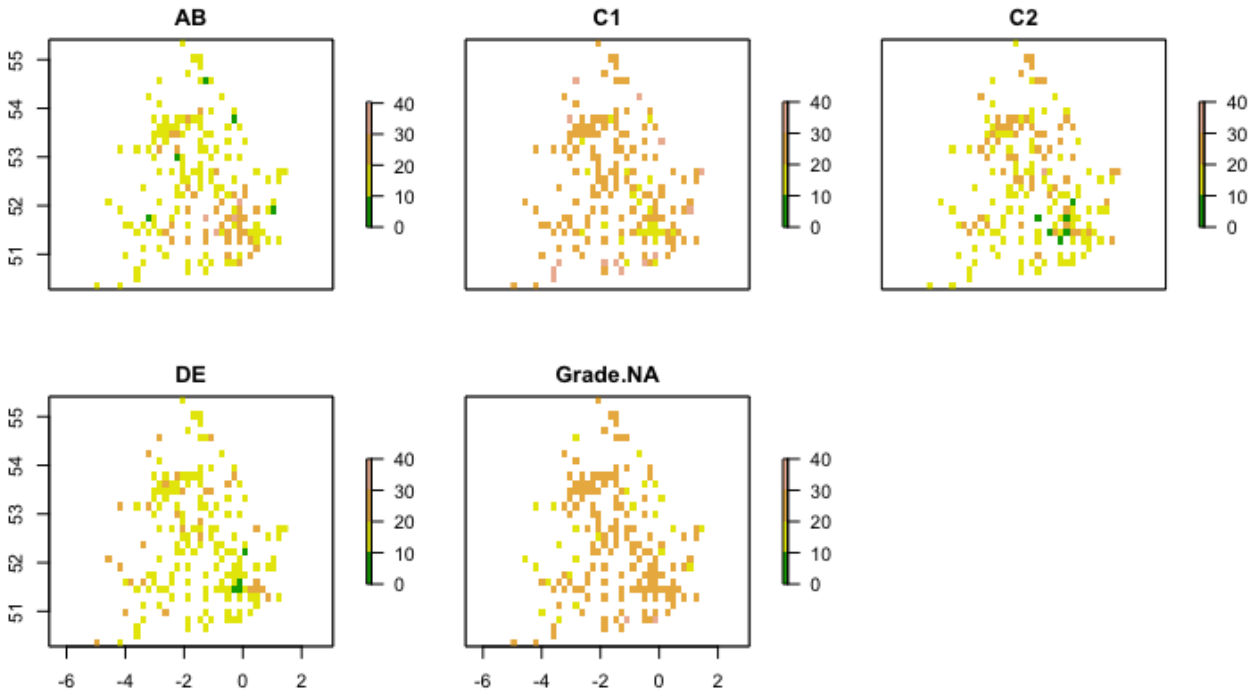


Figure 4.5: A percentage distribution of individuals in each social grade category in spatial units.

It would be good to explore the real effects of the socio-economic covariate in Google data but unfortunately this is not possible due to the limitations of the data available.

4.7.5 Specification of Campaign Design Strategies

Geo-experiment is a control comparison experiment where the control group and treatment group are compared at the same time during the advertising campaign. Given sample of spatial units, we need to decide which spatial units receive the new advertising campaign during the second time period? In this thesis we have defined the term campaign design as the strategy for allocating treatments.

Fisher 1956 proposed a method to answer such question: randomized controlled experiments, which is regarded as the gold standard trial for evaluating the effectiveness of a treatment (Akobeng 2005). In such experiments, a sample of the population of interest is randomly allocated to treatments, where randomisation gives every subject in the population an equal probability of receiving the treatments. The purpose of randomisation is to reduce any possible biases in the study. Randomisation is one of the essential experimental design principles that introduced

by Fisher 1956, 1992 and highlighted by Cochran and Cox 1957. The principle of randomisation does not use any form of restricted allocation to ensure homogeneity among the experimental units. The whole of the variation among the units is included in the experimental error, which affect the accuracy or the size of the estimate of error.

In randomised controlled trials in medicine, for example, patients are treated having identical characteristics except for the experimental treatment. Thus, random allocation of patient makes it likely that any difference in outcome can be explained only by the treatment (Akobeng 2005). In the world of marketing, randomised controlled experiments have been used for some time to investigate the true effect of a treatment in the form of user level and geo-level. For example, identifying whether sales increase after an advertising campaign is caused by the marketing campaign design or by known or unknown factors. However, the randomised controlled experiments is still a challenge in marketing experimental studies due to the lack of user or regional data that are required for balancing treatment allocation (Gordon et al. 2021).

In practice, there is a little advice on how to apply randomisation (Aufenanger 2017, 2018) although the randomisation statistical theory has a long history in field experiments (Fisher 1956, 1992; Kempthorne 1955). Despite this, the marking academic literature has shown an increase in using randomised experiments based on user-level and geo-level to understand advertising effects as were discussed in Chapter 1. In geo-experiments, spatial units are not homogeneous and are diverse in their size, population and demographic characteristics, which could impact individuals' search and conversion behaviours. Therefore, the effectiveness of the new advertising campaign depends on the characteristics of the selected spatial units. Vaver and J. Koehler 2012 recommended to constrain random assignment by one or more characteristics or demographic variables to reduce potential hidden biases. They found that grouping spatial units by size prior to treatment assignment reduce the confidence interval of the campaign effect by 10% or more. This campaign design is known as randomised block or stratified design which is can alleviate the hetero-

geneity among the units and produce reduction in the estimate of error.

In this chapter, we have identified some spatial features including: the average of the social grade and population which can be used to group the spatial units into blocks or strata prior to random assignment. Provided these two blocking covariates, the randomised block campaign design can be applied. In this thesis, we use a special case of randomised block design which is matched-pair design. In this design, spatial units are grouped into pairs according to the blocking covariate, and then within each pair, the units are randomly assigned to opposite conditions: treatment and control (Imai 2008).

The identified realistic search data can also be used to divide the spatial units into blocks. One can thought that pre-observed data such as the expected search rates can assist to understand individuals' behaviours in each spatial unit, although that outcomes change as a result of uncontrolled effects such as advertiser or consumer behaviour, and time and calendar effects due to different browsing behaviour on different days, etc. Despite this, we think the expected search rates, particularly during the first time period, will be useful in stratifying the spatial units into pairs. For all mentioned strata covariates, the matching algorithm starts by arranging the spatial units descendingly using one covariate at a time, and then sort units that have close observations into pairs.

Another possible scheme for grouping spatial units into pairs is using nearest neighbour matching, which pairing a given spatial unit with its closest unit (Rubin 1973; Stuart 2010). It is interesting to consider the fact that adjacent spatial units are more likely to share characteristics that are more similar. Using available spatial features, the closeness between the units can be expressed in terms of distances and dissimilarity measures. The spatial features we have are geographical coordinates, social grades and population. The distances between spatial units can be computed using geographical coordinates latitude and longitude points associated with each spatial units. The dissimilarity between the units are measured using social grades and population. The matching algorithm, inspired by the distance function `k2k` in the `cem` R package (Iacus et al. 2018), works by selecting an unallocated spatial

unit randomly and pairing it with a nearby unit. This procedure is repeated for the remaining spatial units until two units only remain, which have to form a pair in any way regardless if they are adjacent or not.

In addition to the matched-pair design strategy, the complete randomisation will be taken into consideration, despite the fact that the spatial units are not homogeneous. From the marketing perspective, advertisers may find it easier to target spatial units completely random to investigate the effectiveness of a specific change in their advertising campaign, especially in the absence of characteristics that required to handle the spatial heterogeneity. Also, we will take partial randomisation design, where different percentages of the spatial units are assigned randomly to the treatment group. For example, we allocate 10%, 20%, 30% and 40% of the spatial units randomly to serve the new advertising campaign during the second time period. The partial randomisation is considered because some changes in the advertising campaign may cost advertisers a lot and hence a portion of units can be selected to serve the modified campaign.

Other treatment allocation approach is the spatial design, assuming spatial dependency between the units. For this approach, spatial patterns or spatial autocorrelation is usually computed using observed spatial data to measure the similarity between nearby observations (Dutilleul 1993; Legendre et al. 2004; Van Es et al. 2007). However, due to the lack of observed spatial data, the spatial allocation is not included in this study.

In summary, the design strategies that will be employed to design campaigns are

- Complete randomised design
- Partial randomised design
- Matched-pair design, where pairs are matched using
  - population
  - social grades
  - realistic expected search rates

- nearest neighbour algorithm, using
  - dissimilarity measures between social grades
  - dissimilarity measures between population and social grades
  - distances between geographical coordinates

For matched-pair designs, the matching algorithms involve to group spatial units into pairs. This means for an odd number of spatial units, there is a one unpaired unit remaining. Given 205 spatial units in this research, pair 204 units and one unit left unpaired. The unpaired unit is not necessary be the same unit in each matched-pair strategy but for simplicity we fixed the unpaired unit in all applied design strategies. Using spatial effect estimates α_i^* resulted from realistic data, the unit associated with the minimum effect was eliminated. Henceforth, 204 spatial units will be studied instead of 205 units.

4.8 Summary and Concluding Remarks

The purpose of this chapter was to understand the complexity of behavioural structure of response variables of the geo-experiments and suggest a conceptual behavioural structure that enables measuring the impact of advertising campaign. The conceptual behaviour is a two-stage stochastic process: search process and purchasing process. The effectiveness of the campaign is assumed to be attributable to the act of converting online searches to purchases.

Logit-linear regression models were proposed to measure the effectiveness of advertising campaigns. The search and purchase data will be fitted using a statistical model called applied models, where homogeneity with spatial units is assumed. By taking unobserved heterogeneity into account, the applied models are known to be misspecified and hence the estimates result by fitting this model will be incorrect. The impact of misspecification on parameters estimation will be discussed in the forthcoming chapters through a theoretical approach and a computer experiment approach.

The sensitivity of estimation to misspecified models is quantified by measuring the

distance between the models and the truth. In this chapter a truth model was proposed based on heterogeneous strata with spatial units. A hypothetical measure of campaigns effects was also proposed to summarise the effectiveness of the campaign as an aggregate measure of strata.

At the end of the chapter, primitives data were discussed to assist in specifying truth, given that truth is unknown in practice. The primitives include also the identification of spatial units, the identification of population-strata and the specification of campaign design strategies. All these primitives form basic components for both forthcoming study approaches. In the next chapter theoretical framework is developed to study the implications of unobserved covariates for inferences about estimated effects for geo-experiments.

Chapter 5

Consequences of Misspecification of the Applied Model

The purpose of this chapter is to put forward a theoretical framework for assessing the implications of unobserved covariates for inferences about the effectiveness of paid search advertising campaigns when the geo-experiment approach is applied. The issue results from the fact that the campaign effect is estimated using a statistical model that is known to be misspecified due to the presence of unobserved covariates. The points to be raised here are how to measure misspecification of a model and how this reflects on the behaviour of its estimated parameters.

The effectiveness of advertising campaigns is based conceptually on a two-stage stochastic process: searching process S_{it} and purchasing process Y_{it}. The truth and the applied structures of Y_{it} are given in models (4.11) and (4.7), respectively. The truth is based on the assumption of heterogeneity within a spatial unit, whereas the applied model structure is based on the assumption of homogeneity within a spatial unit. The truth and applied model data are fitted using the logistic regression models represented in models (4.12) and (4.9), respectively. It was shown in the previous chapter that the applied probability structure and the applied model used to fit the data are misspecified. In this chapter we are going to study the performance of the applied model by assuming the truth has a particular structure.

For simplicity, we first study the consequences of misspecification of the applied model by considering a single stage stochastic process, which is the case where there is no search process. In other words, we assume the number of searches in each spatial units in each time period is known with value n_{ikt}. The study is then extended to a two-stage stochastic process.

The chapter is divided into two parts. The first part, consisting of the first five sections, studies the asymptotic behaviour of the estimated applied model parameters using a single stage purchase model. The first part starts with a brief description of the purchases model without search. In the following section, we provide a general overview of robustness of likelihood specification to examine the distribution theory of the maximum likelihood estimators under correct specified models and misspecified models. In the same section, we review briefly the Kullback-Leibler divergence criteria and the results obtained by (White 1982) about consistency property and asymptotic normality under misspecification. A toy model is then provided to examine the asymptotic behaviour of estimates under a model that is known to be misspecified. We propose a proxy structure of the specified applied model in the next section to make possible the application of standard results in the literature on maximum likelihood estimation for misspecified models. By the end of this part, a general theoretical approximation of the asymptotic distribution of estimates is delivered for purchase model without search process. In the second part of this chapter, the findings of the asymptotic theory are extended for a two-stage model. By the end of this chapter, some remarks and conclusions are drawn.

5.1 Purchase Model Without Search

Assume the number of searches in each spatial unit in each time period is known with value n^*_{it}, the applied model probability structure (4.7) of the two-stage process is then reduced to a single stage process such that

$$Y_{it} \sim \text{Bin}(n^*_{it}, p^*_{it}) \qquad \text{are independent.} \tag{5.1}$$

and the truth model probability structure (4.11) is reduced to

$$Y_{it} = \sum_k Y_{ikt}, \quad \text{where} \quad Y_{ikt} \sim \text{Bin}(n_{ikt}, p_{ikt}) \qquad \text{are independent,} \tag{5.2}$$

where p^*_{it} and p_{ikt} are modelled using applied and truth purchase models (4.10) and (4.13), respectively. **Figure 5.1** illustrates the truth and applied purchase model data structures when the number of searches is known in a spatial unit i during the two time points.

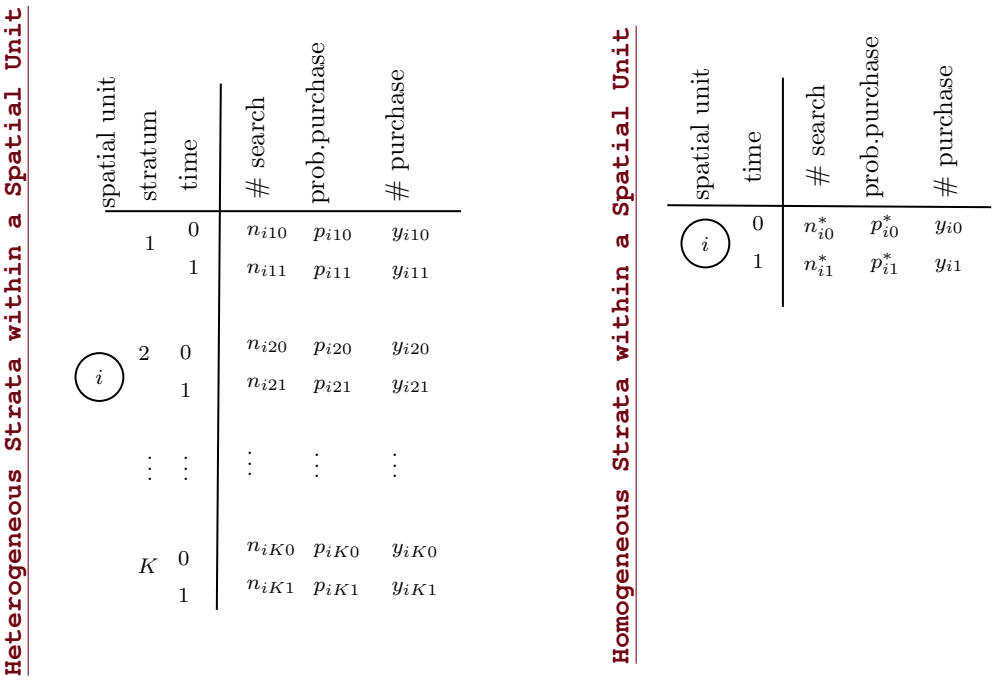


Figure 5.1: An illustration of the breakdown of observed data purchases given known number of searches for a spatial unit i with K population-strata in two time periods, using applied model probability structure and the truth probability structure.

The applied model parameter vector θ^* is estimated using the maximum likelihood estimation method. The likelihood estimates are computed directly from the likelihood function of the applied probability model. However the applied model is known to be misspecified, so how sensitive are the maximum likelihood estimates to this misspecification.

5.2 Robustness of Likelihood Specification

By drawing on the concept of likelihood under misspecification, White 1982 studied the consequences and detection of the model misspecification when the maximum

likelihood method is used. He examined the robustness of the likelihood specification and the properties of the estimators in the case of independent and identically distributed (i.i.d) observations. In section 5.2.3 we present the results related to the consistency and the asymptotic distribution of estimators under misspecification. However, before presenting White's results, we review briefly the distribution of the score function and the asymptotic distribution of consistent estimates under classical regularity conditions.

5.2.1 Classical Distribution of Score Function

Let l be the log-likelihood function of the y_{it} with applied model probability structure (5.1), the score is then the gradient of the l, i.e.

$$S(\theta^*) = \frac{\partial l}{\partial \theta^*}\,.$$

The score function might be seen intuitively as a measure of how close the parameter θ^* is to what the observed data y suggests. For a fixed value of θ^* the score $S(\theta^*)$ is called the score statistic and has a distribution with a specific mean and variance.

Assume that the applied model (5.1) is correctly specified. Then, under some regularity conditions, the expected value of the score, evaluated at the true parameter value θ_0^* of θ^*, is zero, i.e.

$$\mathrm{E}\big[S(\theta_0^*)\big] = 0\,.$$

The variance of the score function is then given by

$$\begin{aligned}\mathrm{Var}\big[S(\theta_0^*)\big] &= \mathrm{E}\big[S(\theta_0^*)S(\theta_0^*)^T\big] - \mathrm{E}\big[S(\theta_0^*)\big]\mathrm{E}\big[S(\theta_0^*)\big]^T \\ &= \mathrm{E}\big[S(\theta_0^*)S(\theta_0^*)^T\big] = \mathrm{E}\Big[\Big(\frac{\partial l}{\partial \theta^*}\Big)\Big(\frac{\partial l}{\partial \theta^*}\Big)^T\Big|_{\theta_0^*}\Big]\,.\end{aligned}$$

At the maximum, the second derivative of the log likelihood function is negative-definite. Given the derivative of the score function as

$$\frac{\partial S(\theta^*)}{\partial \theta^*} = \frac{\partial^2 l}{\partial {\theta^*}^2},$$

and define the curvature of the log-likelihood at $\hat{\theta}^*$ as $I(\hat{\theta}^*)$ where

$$I(\theta^*) = -\frac{\partial S(\theta^*)}{\partial \theta^*} = -\frac{\partial^2 l}{\partial {\theta^*}^2}\,.$$

In likelihood theory $I(\hat{\theta}^*)$ is called the observed fisher information and a large quantity of $I(\hat{\theta}^*)$ indicates a less uncertainty about θ^*. The expected value of the fisher information is given as

$$\mathcal{I}(\theta_0^*) = \mathrm{E}\big[I(\theta_0^*)\big] = -\mathrm{E}\Big[\frac{\partial S(\theta^*)}{\partial \theta^*}\Big|_{\theta_0^*}\Big] = -\mathrm{E}\Big[\frac{\partial^2 l}{\partial {\theta^*}^2}\Big|_{\theta_0^*}\Big].$$

Under some classical regularity condition, the variance of the score function can be expressed in term of the expected value of the Fisher information such that

$$\mathrm{Var}\big[S(\theta_0^*)\big] = \mathcal{I}(\theta_0^*),$$

which is equivalent to state that

$$\mathrm{E}\Big[\Big(\frac{\partial l}{\partial \theta^*}\Big)\Big(\frac{\partial l)}{\partial \theta^*}\Big)^T\Big|_{\theta_0^*}\Big] = -\mathrm{E}\Big[\frac{\partial^2 l}{\partial {\theta^*}^2}\Big|_{\theta_0^*}\Big]. \tag{5.3}$$

5.2.2 Consistency and Asymptotic Normality

In classical likelihood theory, where the applied model is assumed to be correctly specified, the goodness of the estimates can be examined through the repeated sampling properties of sample statistics which shape a basis of the statistical inferences such as bias and variance. Given that $\hat{\theta}^*$ is an estimated vector of the true parameter vector θ_0^*, the bias of the estimate is defined as

$$b(\theta_0^*) = \mathrm{E}[\hat{\theta}^*] - \theta_0^*,$$

with mean square error defined as

$$\mathrm{MSE}(\theta_0^*) = \mathrm{E}[(\hat{\theta}^* - \theta_0^*)^2] = \mathrm{Var}(\hat{\theta}^*) + (b(\theta_0^*))^2\,.$$

$\hat{\theta}^*$ is unbiased estimator if $\mathrm{E}[\hat{\theta}^*] = \theta_0^*$. The bias in the estimates is desired to be small and should not dominate the variability of the estimates (Pawitan 2001).

In the i.i.d case, as the sample size increases towards infinity, under weak regularity conditions, the estimates $\hat{\theta}^*$ converges almost surely to the true parameter vector θ_0^*; i.e. $\hat{\theta}^* \overset{a.s}{\to} \theta_0^*$.

Theorem 1 *Given that $\hat{\theta}^*$ is consistent, then under some classical regularity conditions, the distribution of $\hat{\theta}^*$ in the i.i.d case is asymptotically Normal such that*

$$\hat{\theta}^* \sim \mathrm{N}(\theta_0^*, \mathcal{I}^{-1}(\theta_0^*)),$$

where $\mathcal{I}^{-1}(\theta_0^*)$ is the inverse of the Fisher information that is used as an asymptotic approximation of the variability of $\hat{\theta}^*$ around θ_0^*.

Asymptotically, $n\text{Var}(\hat{\theta}^*)$, $n\mathcal{I}^{-1}(\theta_0^*)$ and $nI^{-1}(\hat{\theta}^*)$ all converge to the same constant matrix so that the observed information can be used to approximate the variance of $\hat{\theta}^*$ and consequently to calculate standard errors etc.

The theorem can be proved by using second-order expansion of the log-likelihood function along with the Central Limit Theorem (CLT). Here we are not interested to go through the proof but an interested reader could see the proof that was written by Pawitan 2001, when he discussed large sample results.

In the non-i.i.d case, consistency and asymptotic normality depend on the asymptotic regime, i.e. on how the experimental design changes as the sample size grows. In suitable regimes the same asymptotic results hold as stated in Theorem 1. However, in practice one wishes to use the asymptotic distribution as a useful approximation for finite sample sizes. Rather than attempt to prove consistency and asymptotic normality for the models used in this thesis, in Chapter 6 we assess the accuracy of the anticipated asymptotic distribution by Monte Carlo simulation.

5.2.3 Distance From the Truth

If the truth probability structure of y are not in the applied model, then there is no value of the applied model parameters associated to the truth. In this case the distance between the applied probability model and the truth is measured. The distance between the two models is measured by the Kullback-Leibler divergence or for short KL-divergence which gives the amount of information lost when the applied structure of observations is specified incorrectly.

Definition 1 *Assume $f(y;\theta^*)$ is a misspecified probability structure of a random variable y and $g(y)$ is the true probability structure of y. The Kullback-Leibler divergence from f to g is*

$$D_{KL}(g\|f) = \mathrm{E}_g\Big[\log\frac{g(y)}{f(y;\theta^*)}\Big]. \tag{5.4}$$

The KL-divergence measures the distance between the two probability structures. Pawitan 2001 highlighted that maximising likelihood is equivalent to minimising the KL-divergence. Let $\tilde{\theta}^*$ be a parameter vector which minimises the KL-divergence, Akaike 1998 observed that when the truth is unknown, the vector of parameter estimates $\hat{\theta}^*$ is a natural estimator for $\tilde{\theta}^*$, the parameter which minimizes the KL-divergence. This statement matters because the best possible model is not reachable in the reality because the truth is unknown, and so the maximum likelihood estimator can be applied as a proxy for fitting our estimates to the truth. We mean by the best model here is the one that fits the data exactly, i.e. when truth is included in the applied distribution. According to KL-divergence definition 1, the best model is obtained when $f(y;\theta^*) \equiv g(y)$, and so $\tilde{\theta}^*$ is the value that gives the closest model to the truth.

It is interesting to understand why maximising likelihood is asymptotically equivalent to minimising the KL-divergence (5.4). Let $y_1, y_2, ..., y_n$ be random sample, where y_i are independent and identically distributed with truth probability function $g(y_i)$. Let $f(y_i;\theta^*)$ be the applied model that is specified to fit the data with unknown θ^*. For independent and identically distributed sample, the average log-likelihood is given by

$$\frac{1}{n}\sum_i \log f(y_i;\theta^*)\,.$$

Maximising the likelihood is equivalent to maximising the log-likelihood. Given large enough n, the weak law of large number says that

$$\frac{1}{n}\sum_i \log f(y_i;\theta^*) \to \mathrm{E}_g\big[\sum_i \log f(y_i;\theta^*)\big]\,.$$

This indicates that maximising the log-likelihood is asymptotically equivalent to minimising the expectation of the negative log-likelihood with respect to the truth g, i.e. as $n \to \infty$

$$\text{maximising} \quad \frac{1}{n}\sum_i \log f(y_i;\theta^*) \equiv \text{minimising} \quad \mathrm{E}_g\big[-\sum_i \log f(y_i;\theta^*)\big]\,.$$

The KL-divergence between g and f is defined as

$$D_{KL}[g(y)\|f(y;\theta^*)] = \mathrm{E}_g\left[\log\frac{\prod_i g(y_i)}{\prod_i f(y_i;\theta^*)}\right]$$
$$= \mathrm{E}_g\big[\sum_i \log g(y_i)\big] - \mathrm{E}_g\big[\sum_i \log f(y_i;\theta^*)\big]\,.$$

But $\mathrm{E}_g\big[\sum_i \log g(y_i)\big]$ is an unknown constant c with respect to θ^*, i.e.

$$D_{KL}[g(y)\|f(y;\theta^*)] = c - \mathrm{E}_g\big[\sum_i \log f(y_i;\theta^*)\big]\,.$$

Therefore, for large enough n, finding the maximum likelihood estimate $\hat{\theta}^*$ is analogous to finding $\tilde{\theta}^*$ the value that gives the nearest model to the truth in the sense of KL-divergence. Provided that $\tilde{\theta}^*$ is the value that gives the nearest model to the truth, then for large enough n, $\hat{\theta}^*$ might be expected to converge to $\tilde{\theta}^*$.

However, when $\hat{\theta}^*$ is estimated under misspecified model, the question that is then naturally to ask in this case: is $\hat{\theta}^*$ consistent? and if so, is it asymptotically Normal? Pawitan 2001 addressed the maximum likelihood under a misspecified model and he presented an example where a misspecification can lead to a robust consistent estimates of the closest parameter $\tilde{\theta}^*$. But he also mentioned that, in general applying misspecified model yields to biased or inconsistent estimates.

5.2.4 Consistency and Asymptotic Normality Under Misspecification

White 1982 addressed consistency and asymptotic normality under misspecification. The results obtained were studied under some regularity conditions stated as assumptions A1 through A7 in his paper. Although the provided conditions A1- A7 are considered simple under which the maximum likelihood is a strongly consistent estimator for the parameter vector which minimize the KL-divergence (Diong et al. 2017) compared to the general conditions provided by Huber 1967, White's conditions are however sufficiently general to have broad applicability.

White studied consistency and asymptotic normality distribution of the maximum likelihood estimator for independent identical (i.i.d) random observations $y_1, y_2, \ldots, y_n$ that have a truth distribution function $g(y)$ and a misspecified applied distribution

function $f(y;\theta^*)$. By considering White's assumption A1-A3, the maximum likelihood estimator $\hat{\theta}^*$ of θ^*, is a consistent estimator for $\tilde{\theta}^*$, i.e.

$$\hat{\theta}^* \overset{a.s}{\to} \tilde{\theta}^* \qquad \text{as} \qquad n\to\infty\,. \tag{5.5}$$

If the applied model f includes the truth, i.e. $g(y) = f(y;\theta_0^*)$ for some θ_0^* in the domain of θ^*, the KL-divergence $D_{KL}(g\|f)$ achieves its unique minimum at $\tilde{\theta}^* = \theta_0^*$ and hence $\hat{\theta}^*$ is consistent for parameter vector θ_0^*, i.e. $\hat{\theta}^* \overset{a.s}{\to} \theta_0^*$. This result corresponds to the classical consistency of the maximum likelihood estimator presented in section 5.2.2.

White showed the estimator $\hat{\theta}^*$ is asymptotically normally distributed, provided assumptions A1-A6. He obtained a variance-covariance matrix for $\hat{\theta}^*$ through defined matrices related to the average log-likelihood. Consider the average log-likelihood

$$l(\hat{\theta}^*) = \frac{1}{n}\sum_i \log f(y_i;\theta^*)\,.$$

Assume partial derivatives of l exist, then White defined

$$\begin{aligned} A_n(\theta^*)_{jj'} &= \frac{1}{n}\sum\nolimits_i \frac{\partial^2 \log f(y_i;\theta^*)}{\partial\theta_j^*\partial\theta_{j'}^*}, \\ B_n(\theta^*)_{jj'} &= \frac{1}{n}\sum\nolimits_i \frac{\partial \log f(y_i;\theta^*)}{\partial\theta_j^*}\frac{\partial \log f(y_i;\theta^*)}{\partial\theta_{j'}^*}\,. \end{aligned}$$

He defined the following if expectations exist

$$\begin{aligned} A(\theta^*)_{jj'} &= \mathrm{E}_g\Big[\frac{\partial^2 \log f(y_i;\theta^*)}{\partial\theta_j^*\partial\theta_{j'}^*}\Big], \\ B(\theta^*)_{jj'} &= \mathrm{E}_g\Big[\frac{\partial \log f(y_i;\theta^*)}{\partial\theta_j^*}\frac{\partial \log f(y_i;\theta^*)}{\partial\theta_{j'}^*}\Big], \end{aligned}$$

where the expected value assumes y_i is sampled from the truth $g(y)$. Provided appropriate inverses exist, he defined

$$\begin{aligned} C_n(\theta^*) &= A_n^{-1}(\theta^*)B_n(\theta^*)A_n^{-1}(\theta^*), \\ C(\theta^*) &= A^{-1}(\theta^*)B(\theta^*)A^{-1}(\theta^*)\,. \end{aligned}$$

From the definition of $A_n(\theta^*)$ and $B_n(\theta^*)$ and the fact that the observations are i.i.d, we see that

$$\begin{aligned} A(\theta^*)_{jj'} &= \frac{1}{n}\sum\nolimits_i \mathrm{E}_g\Big[\frac{\partial^2 \log f(y_i;\theta^*)}{\partial\theta_j^*\partial\theta_{j'}^*}\Big] = \mathrm{E}_g[A_n(\theta^*)], \\ B(\theta^*)_{jj'} &= \frac{1}{n}\sum\nolimits_i \mathrm{E}_g\Big[\frac{\partial \log f(y_i;\theta^*)}{\partial\theta_j^*}\frac{\partial \log f(y_i;\theta^*)}{\partial\theta_{j'}^*}\Big] = \mathrm{E}_g[B_n(\theta^*)]\,. \end{aligned} \tag{5.6}$$

Given $\tilde{\theta}^*$ and that $B(\tilde{\theta}^*)$ is a nonsingular matrix, and $\tilde{\theta}^*$ is a regular point of $A(\theta^*)$, White stated the asymptotic distribution theorem of maximum likelihood estimate $\hat{\theta}^*$ in a misspecified model:

Theorem 2 Given assumptions A1-A6,

$$\sqrt{n}(\hat{\theta}^* - \tilde{\theta}^*) \xrightarrow{d} \mathrm{N}(0, C(\tilde{\theta}^*))$$

$C_n(\hat{\theta}^*) \xrightarrow{a.s} C(\tilde{\theta}^*)$.

where $C(\tilde{\theta}^*)$ is an approximate variance of the maximum likelihood estimator for θ^* from a single observation. However, the variance-covariance matrix $C(\theta^*)$ was reviewed by Chow 1984, who pointed out that the covariance matrix $B(\theta^*)$ stated by White is not correct in general. Chow examined the case for n independent, non-identically distributed observations $y_1, \ldots, y_n$. He defined

$$\mathcal{A}(\theta^*)_{jj'} = \mathrm{E}_g\Big[\frac{1}{n}\frac{\partial^2 l}{\partial\theta_j^*\partial\theta_{j'}^*}\Big],$$
$$\mathcal{B}(\theta^*) = \frac{1}{n}\mathrm{Cov}_g\Big[\frac{\partial l}{\partial\theta^*}\Big].$$

which matches White's definition of $A(\theta^*)$ in the i.i.d case.

For identifiable $\tilde{\theta}^*$, Chow defined the variance-covariance matrix of maximum likelihood estimator for θ^* as

$$\mathcal{C}(\tilde{\theta}^*) = \mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*), \tag{5.7}$$

Now consider the covariance matrix $\mathcal{B}(\tilde{\theta}^*)$, such that

$$\begin{aligned}\mathcal{B}(\tilde{\theta}^*) &= \frac{1}{n}\mathrm{Cov}_g\Big[\frac{\partial l}{\partial\theta^*}\Big|_{\tilde{\theta}^*}\Big] \\ &= \frac{1}{n}\Big[\mathrm{E}_g\Big[\Big(\frac{\partial l}{\partial\theta^*}\Big)\Big(\frac{\partial l}{\partial\theta^*}\Big)^T\Big|_{\tilde{\theta}^*}\Big] - \mathrm{E}_g\Big[\frac{\partial l}{\partial\theta^*}\Big|_{\tilde{\theta}^*}\Big]\mathrm{E}_g\Big[\Big(\frac{\partial l}{\partial\theta^*}\Big)^T\Big|_{\tilde{\theta}^*}\Big]\Big].\end{aligned}$$

But at $\tilde{\theta}^*$, from (5.4) and the definition of $\tilde{\theta}^*$ we have

$$\mathrm{E}_g\Big[\frac{\partial l}{\partial\theta^*}\Big|_{\tilde{\theta}^*}\Big] = 0\,. \tag{5.8}$$

Therefore we have

$$\begin{aligned}\mathcal{B}(\tilde{\theta}^*) = \frac{1}{n}\mathrm{E}_g\Big[\Big(\frac{\partial l}{\partial \theta^*}\Big)\Big(\frac{\partial l}{\partial \theta^*}\Big)^T\Big|_{\tilde{\theta}^*}\Big] &= \frac{1}{n}\mathrm{E}_g\Big[\Big(\sum\nolimits_i \frac{\partial \log f(y_i;\theta^*)}{\partial \theta^*}\Big)\Big(\sum\nolimits_i \frac{\partial \log f(y_i;\theta^*)}{\partial \theta^*}\Big)^T\Big|_{\tilde{\theta}^*}\Big] \\ &= \frac{1}{n}\sum_i \mathrm{E}_g\Big[\Big(\frac{\partial \log f(y_i;\theta^*)}{\partial \theta^*}\Big)\Big(\frac{\partial \log f(y_i;\theta^*)}{\partial \theta^*}\Big)^T\Big|_{\tilde{\theta}^*}\Big] \\ &+ \frac{1}{n}\sum_{i\neq i'} \mathrm{E}_g\Big[\frac{\partial \log f(y_i;\theta^*)}{\partial \theta^*}\Big|_{\tilde{\theta}^*}\Big]\mathrm{E}_g\Big[\Big(\frac{\partial \log f(y_{i'};\theta^*)}{\partial \theta^*}\Big)^T\Big|_{\tilde{\theta}^*}\Big].\end{aligned}$$

By comparing this covariance equation to White's covariance matrix $B(\theta^*)$ when evaluated at $\tilde{\theta}^*$, we see that White omitted the second part of this equation by equating the expectation of each individual term $\partial \log f(y_i;\theta^*)$ to zero. Chow pointed out this fact and stressed that the fact stated in equation (5.8) does not imply that each individual term $\partial \log f(y_i;\theta^*)$ has a zero expectation, i.e.

$$\mathrm{E}_g\Big[\frac{\partial \log f(y_i;\theta^*)}{\partial \theta^*}\Big|_{\tilde{\theta}^*}\Big] \neq 0\,. \tag{5.9}$$

But White considered i.i.d sample, and in this situation the equality sign instead of the inequality sign holds in equation (5.9). Therefore for i.i.d sample, $\mathcal{B}(\tilde{\theta}^*)$ is the same as the covariance matrix obtained by White, i.e. $B(\theta^*)$ in (5.6) when it is evaluated at $\tilde{\theta}^*$. In view of all that has been mentioned so far in this section, the variance-covariance matrix (5.7), provided by Chow holds in general situation for independent sample. Consider Chow's definition of the variance-covariance components, then define: $\mathcal{A}(\theta^*) = n\mathcal{A}(\theta^*)$ and $\mathcal{B}(\theta^*) = n\mathcal{B}(\theta^*)$. Hence

$$\begin{aligned}\mathcal{A}(\tilde{\theta}^*)_{jj'} &= \mathrm{E}_g\Big[\frac{\partial^2 l}{\partial \theta^*_j \partial \theta^*_{j'}}\Big|_{\tilde{\theta}^*}\Big], \\ \mathcal{B}(\tilde{\theta}^*) &= \mathrm{Cov}_g\Big[\frac{\partial l}{\partial \theta^*}\Big].\end{aligned} \tag{5.10}$$

Defining

$$\mathcal{C}(\tilde{\theta}^*) = \mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*), \tag{5.11}$$

and so the convergence result in Theorem 2 becomes $\sqrt{n}(\hat{\theta}^* - \tilde{\theta}^*) \xrightarrow{d} \mathrm{N}(0, n\mathcal{C}(\tilde{\theta}^*))$ giving rise to the approximate distribution for large n

$$\hat{\theta}^* \sim \mathrm{N}(\tilde{\theta}^*, \mathcal{C}(\tilde{\theta}^*))\,. \tag{5.12}$$

The models used in this thesis have independent but non-identically distributed observations and so it is anticipated that this asymptotic distribution may be useful. In Chapter 6, the accuracy of the approximation will be assessed in various scenarios.

In addition White showed that the obtained results of consistency and asymptotic normality of a misspecified model are closely related to the classical consistency and asymptotic normality provided earlier in section 5.2.2. When $g(y) = f(y, \theta_0^*)$, the KL-divergence attains its minimum at the true parameter θ_0^* as was mentioned above. Provided Theorem 3.3 stated in White's paper, then $\mathcal{A}(\theta_0^*) = -\mathcal{B}(\theta_0^*)$ so that $\mathcal{C}(\theta_0^*) = -\mathcal{A}^{-1}(\theta_0^*) = \mathcal{B}^{-1}(\theta_0^*)$, where $-\mathcal{A}(\theta_0^*)$ is Fisher's information matrix $\mathcal{I}(\theta_0^*)$. This is true because under correct specification,

$$\mathcal{B}(\theta_0^*) = \mathrm{Cov}\Big[\frac{\partial l}{\partial \theta^*}\Big|_{\theta_0^*}\Big] = \mathrm{E}\Big[\Big(\frac{\partial l}{\partial \theta^*}\Big)\Big(\frac{\partial l}{\partial \theta^*}\Big)^T\Big|_{\theta_0^*}\Big] = \mathcal{I}(\theta_0^*)\,.$$

By considering equation (5.3), then $\mathcal{A}(\theta_0^*) = -\mathcal{B}(\theta_0^*)$.

Having discussed how maximum likelihood estimates behave when fitting a misspecified model, we will now move on to discuss the theoretical derivation of the estimates of the applied model (4.9). Then we make an attempt to study the asymptotic behaviour of the estimates under two assumptions: correct specification and misspecification.

5.3 Theoretical Derivation of Maximum Likelihood Estimates of the Applied Model

The derivation of the maximum likelihood estimates $\hat{\theta}^*$ of the applied model (4.9) requires to solve the score equation $S(\theta^*) = 0$, which usually has no closed form solution. Nevertheless, it is good to show the derivation of the $\hat{\theta}^*$ for a small dimension of θ^* to illustrate the procedure of the estimation and make it easer to study the asymptotic behaviour of the estimates under misspecification.

Given the applied probability structure (5.1) of y_{it}, purchasing process, the probability function of y_{it} is

$$f(y_{it}) = \binom{n^*_{it}}{y_{it}} (p^*_{it})^{y_{it}} (1 - p^*_{it})^{n^*_{it} - y_{it}}$$

Given independent vector y of observed data of purchases y_{it}, the likelihood function is then

$$\mathcal{L} = \prod_{i,t} f(y_{it}) = \prod_{i,t} \binom{n^*_{it}}{y_{it}} (p^*_{it})^{y_{it}} (1 - \mathrm{p}^*_{it})^{n^*_{it} - y_{it}} .$$

The $\binom{n^*_{it}}{y_{it}}$ does not depend on p^*_{it} and so will not affect the derivation of estimators. In what follows, we consider the likelihood proportional to $\mathcal{L}$:

$$L = \prod_{i,t} (p^*_{it})^{y_{it}} (1 - \mathrm{p}^*_{it})^{n^*_{it} - y_{it}} .$$

The log-likelihood is

$$l = \sum_{i,t} y_{it} \log(p^*_{it}) + (n^*_{it} - y_{it}) \log(1 - p^*_{it}) . \tag{5.13}$$

Given that $\eta^*_{it} = \mathrm{logit}(p^*_{it})$, the likelihood is given by

$$\begin{aligned} L &= \prod_{i,t} \left(\frac{p^*_{it}}{1 - p^*_{it}} \right)^{y_{it}} (1 - p^*_{it})^{n^*_{it}} \\ &= \prod_{i,t} (e^{\eta^*_{it}})^{y_{it}} (1 + e^{\eta^*_{it}})^{-n^*_{it}}, \end{aligned}$$

and the log-likelihood is

$$l = \sum_{i,t} (\eta^*_{it} y_{it} - n^*_{it} \log(1 + e^{\eta^*_{it}})) . \tag{5.14}$$

On the basis of the matrix formulation of the applied model, the likelihood is given by

$$L = \prod_{i,t} \left(e^{\sum_j x^*_{it_j} \theta^*_j} \right)^{y_{it}} \left(1 + e^{\sum_j x^*_{it_j} \theta^*_j} \right)^{-n^*_{it}},$$

and the log-likelihood function is

$$l = \sum\nolimits_{i,t} y_{it} \Big(\sum\nolimits_j x^*_{it_j} \theta^*_j \Big) - n^*_{it} \log \left(1 + e^{\sum_j x^*_{it_j} \theta^*_j} \right) . \tag{5.15}$$

The log-likelihood depends on parameters vectors θ^*. The gradient of the log-likelihood is then a score vector of first partial derivatives with respect to θ^* or in specific α_i^*, β^* and δ^*. The score vector is then

$$S(\theta^*) = \frac{\partial l}{\partial \theta^*} = \begin{bmatrix} \frac{\partial l}{\partial \alpha_1^*} & \dots \frac{\partial l}{\partial \alpha_I^*} & \frac{\partial l}{\partial \beta^*} & \frac{\partial l}{\partial \delta^*} \end{bmatrix}^T .$$

The maximum likelihood estimates $\hat{\theta^*}$ is the solution of the score equation

$$S(\theta^*) = 0 .$$

Example 1 (Toy Model) Let $i \in \{1, 2\}$ such that $i = 2$ receives the new advertising campaign. The probability of purchasing p_{it} is then modelled as

$$\text{logit}(p_{it}^*) = \eta_{it}^* = \alpha_i^* + \beta^* t + \delta^* C_{it},$$

such that $\eta^* = \begin{bmatrix} \eta_{10}^* & \eta_{11}^* & \eta_{20}^* & \eta_{21}^* \end{bmatrix}^T$, where

$$\eta_{10}^* = \alpha_1^* \qquad \eta_{11}^* = \alpha_1^* + \beta^*$$
$$\eta_{20}^* = \alpha_2^* \qquad \eta_{21^*} = \alpha_2^* + \beta^* + \delta^*$$

Consider the log-likelihood equation (5.14), then

$$l = \eta_{10}^* y_{10} - n_{10}^* \log(1 + \exp(\eta_{10}^*)) + \eta_{11}^* y_{11} - n_{11}^* \log(1 + \exp(\eta_{11}^*))$$
$$+ \eta_{20}^* y_{20} - n_{20}^* \log(1 + \exp(\eta_{20}^*)) + \eta_{21}^* y_{21} - n_{21}^* \log(1 + \exp(\eta_{21}^*)),$$

substitute expressions of each element in η^*, we get

$$l = \alpha_1^* y_{10} - n_{10}^* \log(1 + \exp(\alpha_1^*)) + (\alpha_1^* + \beta^*) y_{11} - n_{11}^* \log(1 + \exp(\alpha_1 + \beta^*))$$
$$+ \alpha_2^* y_{20} - n_{20}^* \log(1 + \exp(\alpha_2^*)) + (\alpha_2^* + \beta^* + \delta^*) y_{21} - n_{21}^* \log(1 + \exp(\alpha_2^* + \beta^* + \delta^*)) .$$

The score vector, i.e. the first partial derivatives of the log-likelihood with respect to each parameter in η^* is,

$$\begin{aligned}
\frac{\partial l}{\partial \alpha_1^*} &= y_{10} - n_{10}^* \frac{\exp(\alpha_1^*)}{1 + \exp(\alpha_1^*)} + y_{11} - n_{11}^* \frac{\exp(\alpha_1^* + \beta^*)}{1 + \exp(\alpha_1^* + \beta^*)} \\
\frac{\partial l}{\partial \alpha_2^*} &= y_{20} - n_{20}^* \frac{\exp(\alpha_2^*)}{1 + \exp(\alpha_2^*)} + y_{21} - n_{21}^* \frac{\exp(\alpha_2^* + \beta^* + \delta^*)}{1 + \exp(\alpha_2^* + \beta^* + \delta^*)} \\
\frac{\partial l}{\partial \beta^*} &= y_{11} - n_{11}^* \frac{\exp(\alpha_1^* + \beta^*)}{1 + \exp(\alpha_1^* + \beta^*)} + y_{21} - n_{21}^* \frac{\exp(\alpha_2^* + \beta^* + \delta^*)}{1 + \exp(\alpha_2^* + \beta^* + \delta^*)} \\
\frac{\partial l}{\partial \delta^*} &= y_{21} - n_{21}^* \frac{\exp(\alpha_2^* + \beta^* + \delta^*)}{1 + \exp(\alpha_2^* + \beta^* + \delta^*)} .
\end{aligned} \tag{5.16}$$

The estimate vector $\hat{\theta}^*$ can be found by setting each of the equation above to zero and solving for its corresponding parameter. This results a system of four nonlinear equations with four unknown parameters. The solution to such system is not easily derived directly from (5.16).

By considering applied model $\eta^* = X^*\theta^*$ and $\eta^*_{it} = \text{logit}(p^*_{it})$ for a unit i in time t, we see there is a composition of transformations that maps a parameter vector θ^* to a probability vector p^*. Define the linear map $T : \theta^* \to \eta^*$ and the map $U : \eta^* \to p^*$, then compose transformations that maps θ^* to p^* is $(U \circ T)(\theta^*) : \theta^* \to p^*$.

If we can show that the composition $(U \circ T)$ is bijective, then

$$\frac{\partial l}{\partial \theta^*} = 0 \iff \frac{\partial l}{\partial \, \mathrm{p}^*} = 0\,.$$

The composition $(U \circ T)$ is a bijection if and only if T is a bijection and U is a bijection. Given that $\eta^*_{it} = \text{logit}(p^*_{it})$ is a component-wise bijective function, then U is a bijection. For the toy example, i.e. two spatial units, the linear map T is given by

$$T(\theta^*) = X^*\theta^* = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} \begin{bmatrix} \alpha^*_1 \\ \alpha^*_2 \\ \beta^* \\ \delta^* \end{bmatrix} = \begin{bmatrix} \alpha^*_1 \\ \alpha^*_1 + \beta^* \\ \alpha^*_2 \\ \alpha^*_2 + \beta^* + \delta^* \end{bmatrix} = \begin{bmatrix} \eta^*_{10} \\ \eta^*_{11} \\ \eta^*_{20} \\ \eta^*_{21} \end{bmatrix}.$$

If we swap the second and third rows in X^*, we get a lower triangular matrix with diagonal entries equal 1. Provided that if we swap rows in a matrix, the determinant will change its sign. The determinant of a triangular matrix is the product of the diagonal entries, hence $\det(X^*) = -1$. Since U and T are both bijections, then so is $(U \circ T)(\theta^*)$.

Consider the log-likelihood l (5.13). Its first derivative with respect to p^*_{it} is

$$\frac{\partial l}{\partial p^*_{it}} = \frac{y_{it}}{p^*_{it}} - \frac{n^*_{it} - y_{it}}{1 - p^*_{it}}\,.$$

By equating this score function to zero, we get

$$\begin{aligned} \frac{y_{it}}{p^*_{it}} - \frac{n^*_{it} - y_{it}}{1 - p^*_{it}} &= 0 \\ y_{it}(1 - p^*_{it}) &= (n^*_{it} - y_{it})p^*_{it}\,. \end{aligned}$$

The maximum likelihood estimator of p^*_{it} is then

$$\hat{p}^*_{it} = \frac{y_{it}}{n^*_{it}}\,. \tag{5.17}$$

given that the second derivative matrix of the log-likelihood function is a negative-definite. Then

$$\hat{\eta}^*_{it} = \text{logit}(\hat{p}^*_{it}) = \log(\frac{y_{it}}{n^*_{it} - y_{it}})\,.$$

and so the maximum likelihood estimates $\hat{\alpha}^*_1$, $\hat{\alpha}^*_2$, $\hat{\beta}^*$ and $\hat{\delta}^*$ are given by

$$\begin{aligned}
\hat{\alpha}^*_1 &= \text{logit}(\hat{p}^*_{10}),\\
\hat{\alpha}^*_2 &= \text{logit}(\hat{p}^*_{20}),\\
\hat{\beta}^* &= \text{logit}(\hat{p}^*_{11}) - \text{logit}(\hat{p}^*_{10})\\
\hat{\delta}^* &= \text{logit}(\hat{p}^*_{21}) - \text{logit}(\hat{p}^*_{20}) - (\text{logit}(\hat{p}^*_{11}) - \text{logit}(\hat{p}^*_{10}))\,.
\end{aligned} \tag{5.18}$$

In this toy example we have estimated the parameter vector θ^* using a composition bijection property, where there are four observed data points and all these points are used to estimate the four unknown parameters. However, it is not necessary the case for more complex model, as we see in the following example.

Example 2 Let $i \in \{1, 2, 3\}$ such that the second spatial unit receives the new advertising campaign at the second time period, then $\eta^* = \begin{bmatrix}\eta^*_{10} & \eta^*_{11} & \eta^*_{20} & \eta^*_{21} & \eta^*_{30} & \eta^*_{31}\end{bmatrix}^T$,

such that

$$\begin{aligned}
\eta^*_{10} &= \alpha^*_1 & \eta^*_{11} &= \alpha^*_1 + \beta^*\\
\eta^*_{20} &= \alpha^*_2 & \eta^*_{21} &= \alpha^*_2 + \beta^* + \delta^*\\
\eta^*_{30} &= \alpha^*_3 & \eta^*_{31} &= \alpha^*_3 + \beta^*
\end{aligned}$$

The log likelihood is given by

$$\begin{aligned}
l = {} & \eta^*_{10}y_{10} - n^*_{10}\log(1+\exp(\eta^*_{10})) + \eta^*_{11}y_{11} - n^*_{11}\log(1+\exp(\eta^*_{11}))\\
& + \eta^*_{20}y_{20} - n^*_{20}\log(1+\exp(\eta^*_{20})) + \eta^*_{21}y_{21} - n^*_{21}\log(1+\exp(\eta^*_{21}))\\
& + \eta^*_{30}y_{30} - n^*_{30}\log(1+\exp(\eta^*_{30})) + \eta^*_{31}y_{31} - n^*_{31}\log(1+\exp(\eta^*_{31}))
\end{aligned}$$

or equivalently,

$$\begin{aligned}
l = {} & \alpha^*_1 y_{10} - n^*_{10}\log(1+\exp(\alpha^*_1)) + (\alpha^*_1+\beta^*)y_{11} - n^*_{11}\log(1+\exp(\alpha^*_1+\beta^*))\\
& + \alpha^*_2 y_{20} - n^*_{20}\log(1+\exp(\alpha^*_2)) + (\alpha^*_2+\beta^*+\delta^*)y_{21} - n^*_{21}\log(1+\exp(\alpha^*_2+\beta^*+\delta^*))\\
& + \alpha^*_3 y_{30} - n^*_{30}\log(1+\exp(\alpha^*_3)) + (\alpha^*_3+\beta^*+\delta^*)y_{31} - n^*_{31}\log(1+\exp(\alpha^*_3+\beta^*+\delta^*))\,.
\end{aligned}$$

The first partial derivatives of the log-likelihood with respect to each parameter in η^* are listed below,

$$\begin{aligned}
\frac{\partial l}{\alpha_i^*} &= \sum_{t=0}^{1} y_{i0} - n_{i0}^* \frac{\exp(\eta_{i0}^*)}{1+\exp(\eta_{i0}^*)}, \qquad \text{for} \quad i = 1,2,3 \\
\frac{\partial l}{\beta^*} &= \sum_{i=0}^{3} y_{i1} - n_{i1}^* \frac{\exp(\eta_{i1}^*)}{1+\exp(\eta_{i1}^*)}, \\
\frac{\partial l}{\delta^*} &= y_{21} - n_{21}^* \frac{\exp(\eta_{21}^*)}{1+\exp(\eta_{21}^*)}\,.
\end{aligned} \tag{5.19}$$

By equating each of these equations to zero form a nonlinear system that has no analytical or closed form solution. For this model, deriving the score function in term of η^* is not a one-to-one function because we have six observed values and five parameters.

The score vector delivers a system of nonlinear equations in α_i, β and δ, that require special methods to find their corresponding solution. Indeed this would be the case if we consider more observed values, i.e. spatial units $i > 3$. These methods are iterative and have been built into available statistics software such as *R*. Iterative weighted least squares is one of the most common iterative method that are used to obtain the maximum likelihood estimates. In this research we do not need to be concerned about the algorithms of these iterative methods but the interested reader may see the text by (McCullagh and Nelder 1989) for a general description of the algorithms used in fitting generalized linear models.

Generally, for a large sample study, the maximum likelihood estimate $\hat{\theta}^*$ has no exact theoretical form which makes the theoretical asymptotic behaviour study of estimates difficult.

5.4 Asymptotic Behaviour of Toy Model

The toy model is the model that fits data for two spatial units $i \in \{1,2\}$ where each has two strata $K = 2$. Without strata, y_{it} has a probability distribution given in the applied probability structure (5.1). Given $K = 2$, the truth structure of y_{it} is $y_{it} = \sum_{k}^{2} y_{ikt}$ where y_{ikt} has a probability distribution given in the truth probability structure (5.2).

The applied toy model is given by

$$\text{logit}(p^*_{it}) = \eta^*_{it} = \alpha^*_i + \beta^* t + \delta^* C_{it}, \tag{5.20}$$

and the truth toy model is given by

$$\text{logit}(p_{ikt}) = \eta_{ikt} = \alpha_i + \gamma_k + \beta t + \delta C_{it}\,. \tag{5.21}$$

The applied toy model has been discussed in Example 1. The maximum likelihood estimates of the applied parameters $\hat{\alpha^*_1}$, $\hat{\alpha^*_2}$, $\hat{\beta^*}$ and $\hat{\delta^*}$ are found in (5.18). The question then becomes how these estimates behave in the applied model when the truth generates purchases y_{it}. Provided that the maximum likelihood mechanism involves the model to be correctly specified, the behaviour of these estimates are studied under two model structures: the applied structure and the truth structure.

5.4.1 Behaviour of the Estimates Under Correct Specification of the Applied Toy Model:

Assume that the applied probability structure (5.1) of y_{it} is correctly specified. Its expected value and its variance are then given by

$$\text{E}[Y_{it}] = n^*_{it} p^*_{it}, \qquad \text{Var}[Y_{it}] = n^*_{it} p^*_{it}(1 - p^*_{it}),$$

and so the mean and the variance of $\hat{p^*_{it}}$ are

$$\begin{aligned}
\mu_{\hat{p}^*_{it}} &= \text{E}[\hat{\text{p}}^*_{it}] = \text{E}\Big[\frac{Y_{it}}{n^*_{it}}\Big] = \frac{\text{E}[Y_{it}]}{n^*_{it}} = \frac{n^*_{it} p^*_{it}}{n^*_{it}} = p^*_{it}, \\
\sigma^2_{\hat{p}^*_{it}} &= \text{Var}[\hat{p}^*_{it}] = \text{Var}\Big[\frac{Y_{it}}{n^*_{it}}\Big] = \frac{n^*_{it} p^*_{it}(1 - p^*_{it})}{{n^*_{it}}^2} = \frac{p^*_{it}(1 - p^*_{it})}{n^*_{it}}\,.
\end{aligned}$$

If n^*_{it} is large enough, then the distribution of $\hat{p}^*_{it}$ is approximately Normal such that

$$\hat{p}^*_{it} \sim \text{N}(p^*_{it}, p^*_{it}(1 - p^*_{it})/n^*_{it})\,.$$

Using the linear approximation of the $\hat{p}^*_{it}$ around p^*_{it} gives

$$\hat{p}^*_{it} \approx p^*_{it} + z_{it} \cdot \sqrt{\frac{p^*_{it}(1 - p^*_{it})}{n^*_{it}}},$$

where $z_{it} \sim \mathrm{N}(0,1)$. Using first-order expansion of $\mathrm{logit}(\hat{p}^*_{it})$ around p^*_{it} gives

$$\mathrm{logit}(\hat{p}^*_{it}) \approx \mathrm{logit}(p^*_{it}) + z_{it} \cdot \sqrt{\frac{p^*_{it}(1-p^*_{it})}{n^*_{it}}} \cdot \frac{\partial \,\mathrm{logit}(p^*_{it})}{\partial p^*_{it}},$$

where

$$\frac{\partial \,\mathrm{logit}(p^*_{it})}{\partial p^*_{it}} = \frac{1}{p^*_{it}(1-p^*_{it})},$$

this gives

$$\mathrm{logit}(\hat{p}^*_{it}) \approx \mathrm{logit}(p^*_{it}) + z_{it} \cdot \sqrt{\frac{1}{n^*_{it}p^*_{it}(1-p^*_{it})}}\,.$$

Then the linear approximation of $\hat{\alpha}^*_1$ is

$$\hat{\alpha}^*_1 = \mathrm{logit}(\hat{p}^*_{10}) \approx \mathrm{logit}(p^*_{10}) + z \cdot \sqrt{\frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})}}\,.$$

Define $\alpha^*_1 = \mathrm{logit}(p^*_{10})$, then linear approximation of $\hat{\alpha}^*_1$ is given by

$$\hat{\alpha}^*_1 \approx \alpha^*_1 + z_{10} \cdot \sqrt{\frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})}}\,.$$

Similarly define

$$\begin{aligned}
\alpha^*_2 &= \mathrm{logit}(p^*_{20}) \\
\beta^* &= \mathrm{logit}(p^*_{11}) - \mathrm{logit}(p^*_{10}) \\
\delta^* &= \mathrm{logit}(p^*_{21}) - \mathrm{logit}(p^*_{20}) - (\mathrm{logit}(p^*_{11}) - \mathrm{logit}(p^*_{10}))\,.
\end{aligned}$$

The linear approximation of $\hat{\alpha}^*_2$ is given by

$$\hat{\alpha}^*_2 = \mathrm{logit}(\hat{p}^*_{20}) \approx \alpha^*_2 + z_{20} \cdot \sqrt{\frac{1}{n^*_{20}p^*_{20}(1-p^*_{20})}}\,.$$

The linear approximation of $\hat{\beta}^*$

$$\begin{aligned}
\hat{\beta}^* = \mathrm{logit}(\hat{p}^*_{11}) - \mathrm{logit}(\hat{p}^*_{10}) &\approx \beta^* + z_{11} \cdot \sqrt{\frac{1}{n^*_{11}p^*_{11}(1-p^*_{11})}} + z_{10} \cdot \sqrt{\frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})}} \\
&\approx \beta^* + z \cdot \sqrt{\frac{1}{n^*_{11}p^*_{11}(1-p^*_{11})} + \frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})}},
\end{aligned}$$

where z is a standard normal value. Similarly, the linear approximation of $\hat{\delta}^*$

$$\begin{aligned}\hat{\delta}^* &= \text{logit}(\hat{p}^*_{21}) - \text{logit}(\hat{p}^*_{20}) - (\text{logit}(\hat{p}^*_{11}) - \text{logit}(\hat{p}^*_{10})) \\ &\approx \delta^* + z \cdot \sqrt{\sum_{i=1}^{2}\sum_{t=0}^{1} \frac{1}{n^*_{it}p^*_{it}(1-p^*_{it})}}\,.\end{aligned}$$

In fact $\hat{\alpha}^*_1$ and $\hat{\beta}^*$ share a component of variation of $\text{logit}(p^*_{10})$ which makes them negatively correlated. Also, $\hat{\delta}^*$ shares one or more components of variation with each of the other parameters and so will be correlated with all of them. The variance-covariance matrix for $(\hat{\alpha_1}^*, \hat{\alpha_2}^*, \hat{\beta}^*, \hat{\delta}^*)$ is

$$\text{Cov}(\hat{\theta}^*) = \begin{pmatrix} \frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})} & 0 & -\frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})} & \frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})} \\ & \frac{1}{n^*_{20}p^*_{20}(1-p^*_{20})} & 0 & -\frac{1}{n^*_{20}p^*_{20}(1-p^*_{20})} \\ & & \frac{1}{n^*_{11}p^*_{11}(1-p^*_{11})} + \frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})} & -\frac{1}{n^*_{11}p^*_{11}(1-p^*_{11})} + \frac{1}{n^*_{10}p^*_{10}(1-p^*_{10})} \\ & & & \sum_{i=1}^{2}\sum_{t=0}^{1}\frac{1}{n^*_{it}p^*_{it}(1-p^*_{it})} \end{pmatrix}.$$

5.4.2 Behaviour of the Estimates Under Wrong Specification of the Applied Toy Model:

Assume that the applied model (5.20) does not include the true structure of y_{it}, then the question is how this would affect the behaviour of the estimates $\hat{\alpha}^*_1$, $\hat{\alpha}^*_2$, $\hat{\beta}^*$ and $\hat{\delta}^*$. Consider the true probability structure (5.2), then $y_{it} = y_{i1t} + y_{i2t}$, where y_{it} is a sum of two non-identically distributed, independent Binomial random variables, each with probability Binomial function, say $g_1(y_{i1t})$ and $g_2(y_{i2t})$, respectively. The convolution of $g_1(y_{i1t})$ and $g_2(y_{i2t})$ is the probability function $g(y_{it}) = g_1(y_{i1t}) \cdot g_2(y_{i2t})$ such that

$$g(y_{it} = j) = \sum_r g_1(y_{i1t} = r) \cdot g_2(y_{i2t} = j - r)$$

where j is an arbitrary value in $[0, n_{i1t} + n_{i2t}]$. This is equivalent to find the sum of all pairwise disjoint events of $y_{i1t} = j$ and $y_{i2t} = j - r$,

$$\begin{aligned}\text{P}[y_{it} = j] &= \sum_{r=max(0,j-n_{i2t})}^{min(j,n_{i1t})} \text{P}[y_{i1t} = r \cap y_{i2t} = j - r] \\ &= \sum_{r=max(0,j-n_{i2t})}^{min(j,n_{i1t})} \binom{n_{i1t}}{r}\binom{n_{i2t}}{j-r} p^r_{i1t}(1-p_{i1t})^{n_{i1t}-r} p^{j-r}_{i2t}(1-p_{i2t})^{n_{i2t}-j+r}\,.\end{aligned} \tag{5.22}$$

This distribution is cumbersome to find and difficult to compute. A number of studies have proposed approximation methods to compute the convolution of independent Binomial variables (Johnson et al. 2005; Jolayemi 1992; Eisinga et al. 2013). An example of these method is the saddlepoint approximation which was explored by (Daniels 1954) to approximate the probability distribution of variables whose exact distribution cannot be conveniently obtained. Eisinga et al. 2013 examined the saddlepoint approximation for non-identically distributed, independent Binomial variables. The implementation of the saddlepoint approximation of the distribution of the sum of non-identically distributed, independent Binomials in R was addressed by (Liu and Quertermous 2017) in the open source package called sinib. For the moment we need not to be concerned about the details of this method or its computation. Alternatively, we could assume n_{it} and n_{ikt} are very large and work with the asymptotic distribution of y_{it}. The asymptotic distribution of $y_{it} = y_{i1t} + y_{i2t}$ is

$$Y_{it} \sim \mathrm{N}(\sum_{k=1}^{2} n_{ikt}p_{ikt}, \sum_{k=1}^{2} n_{ikt}p_{ikt}(1-p_{ikt})),$$

because by CLT, $\sum_{k=1}^{2} y_{ikt}$ converges to a Normal distribution such that

$$\begin{aligned}
\mathrm{E}\Big[\sum_{k=1}^{2} Y_{ikt}\Big] &= \sum_{k=1}^{2} \mathrm{E}[\,Y_{ikt}\,] = \sum_{k=1}^{2} n_{ikt}p_{ikt}, \\
\mathrm{Var}\Big[\sum_{k=1}^{2} Y_{ikt}\Big] &= \sum_{k=1}^{2} \mathrm{Var}[\,Y_{ikt}\,] = \sum_{k=1}^{2} n_{ikt}p_{ikt}(1-p_{ikt})\,.
\end{aligned}$$

The estimate of the probability of the purchasing, $\hat{\mathrm{p}}^{*}_{it}$ under the applied mode is

$$\hat{p}^{*}_{it} = \frac{Y_{it}}{n^{*}_{it}} = \frac{1}{n^{*}_{it}} \sum_{k=1}^{2} Y_{ikt}\,.$$

The estimator $\hat{p}^{*}_{it}$ is then Normally distributed with mean $\mu_{\hat{p}^{*}_{it}}$ and variance $\sigma^{2}_{\hat{p}^{*}_{it}}$ such that

$$\mu_{\hat{p}^{*}_{it}} = \mathrm{E}\Big[\frac{1}{n^{*}_{it}} \sum_{k=1}^{2} Y_{ikt}\Big] = \frac{1}{n^{*}_{it}} \sum_{k=1}^{2} \mathrm{E}[\,Y_{ikt}\,] = \frac{1}{n^{*}_{it}} \sum_{k=1}^{2} n_{ikt}p_{ikt},$$

$$\sigma^{2}_{\hat{p}^{*}_{it}} = \mathrm{Var}\Big[\frac{1}{n^{*}_{it}} \sum_{k=1}^{2} y_{ikt}\Big] = \frac{1}{n^{*\,2}_{it}} \sum_{k=1}^{2} \mathrm{Var}[\,y_{ikt}\,] = \frac{1}{n^{*\,2}_{it}} \sum_{k=1}^{2} n_{ikt}p_{ikt}(1-p_{ikt})\,.$$

Define the total probability of purchasing p^*_{it} as

$$p^*_{it} = \sum_{k=1}^{2} \frac{n_{ikt} p_{ikt}}{n^*_{it}} \tag{5.23}$$

or equivalently,

$$\sum_{k=1}^{2} n_{ikt} p_{ikt} = n^*_{it} p^*_{it},$$

This gives

$$\mu_{\hat{p}^*_{it}} = \frac{1}{n^*_{it}} \sum_{k=1}^{2} n_{ikt} p_{ikt} = \frac{n_{it} p^*_{it}}{n^*_{it}} = p^*_{it},$$

$$\sigma_{\hat{p}^*_{it}} = \frac{1}{{n^*_{it}}^2} \sum_{k=1}^{2} n_{ikt} p_{ikt}(1 - p_{ikt}) = \frac{1}{{n^*_{it}}^2} \sum_{k=1}^{2} n_{ikt} p_{ikt} - \frac{1}{n^2_{it}} \sum_{k=1}^{2} n_{ikt} p^2_{ikt} = \frac{p^*_{it}}{n^*_{it}} - \frac{1}{{n^*_{it}}^2} \sum_{k=1}^{2} n_{ikt} p^2_{ikt}\,.$$

The asymptotic distribution of $\hat{p}^*_{it}$ is then given by

$$\hat{p}^*_{it} \sim \mathrm{N}(p^*_{it}, \frac{p^*_{it}}{n^*_{it}} - \frac{1}{{n^*_{it}}^2} \sum_{k=1}^{2} n_{ikt} p^2_{ikt}),$$

and so the linear approximation of $\hat{p}^*_{it}$ around the mean p^*_{it}

$$\hat{p}_{it}* \approx p_{it}* + z_{it} \cdot \sqrt{\frac{p_{it}*}{n^*_{it}} - \frac{1}{{n^*_{it}}^2} \sum_{k=1}^{2} n_{ikt} p^2_{ikt}},$$

and using first-order expansion of $\mathrm{logit}(\hat{p}^*_{it})$ around $\mathrm{logit}(p^*_{it})$ gives

$$\mathrm{logit}(\hat{p}^*_{it}) \approx \mathrm{logit}(p^*_{it}) + z_{it} \cdot \sqrt{\frac{p^*_{it}}{n^*_{it}} - \frac{1}{{n^*_{it}}^2} \sum_{k=1}^{2} n_{ikt} p^2_{ikt}} \cdot \frac{1}{p^*_{it}(1 - p^*_{it})}$$

$$= \mathrm{logit}(p^*_{it}) + z_{it} \cdot \sqrt{\frac{n^*_{it} p^*_{it} - \sum_{k=1}^{2} n_{ikt} p^2_{ikt}}{{n^*_{it}}^2 {p^*_{it}}^2 (1 - \mathrm{p}^*_{it})^2}}\,.$$

Given p^*_{it} in (5.23) we can express α^*_1, α^*_2, β^* and δ^* in term of p_{ikt}. Then use model (5.21) to express them in terms of $\alpha_i + \gamma_k$, β and δ. So we have

$$\alpha^*_1 = \mathrm{logit}(p^*_{10}) = \mathrm{logit}(\frac{1}{n^*_{10}} \sum_{k=1}^{2} n_{1k0} p_{1k0}),$$

$$\alpha^*_2 = \mathrm{logit}(p^*_{20}) = \mathrm{logit}(\frac{1}{n^*_{20}} \sum_{k=1}^{2} n_{2k0} p_{2k0}),$$

$$\beta^* = \mathrm{logit}(p^*_{11}) - \mathrm{logit}(p^*_{10}) = \mathrm{logit}(\frac{1}{n^*_{11}} \sum_{k=1}^{2} n_{1k1} p_{1k1}) - \mathrm{logit}(\frac{1}{n^*_{10}} \sum_{k=1}^{2} n_{1k0} p_{1k0}),$$

$$\delta^* = \mathrm{logit}(p^*_{21}) - \mathrm{logit}(p^*_{20}) - \beta^* = \mathrm{logit}(\frac{1}{n^*_{21}} \sum_{k=1}^{2} n_{2k1} p_{2k1}) - \mathrm{logit}(\frac{1}{n^*_{20}} \sum_{k=1}^{2} n_{2k0} p_{2k0}) - \beta^*\,.$$

From model (5.21), we have

$$\text{logit}(p_{ik0}) = \alpha_i + \gamma_k \qquad \text{and} \qquad p_{ik0} = \frac{\exp(\alpha_i + \gamma_k)}{1 + \exp(\alpha_i + \gamma_k)},$$

$$\text{logit}(p_{1k1}) = \alpha_1 + \gamma_k + \beta \qquad \text{and} \qquad p_{1k1} = \frac{\exp(\alpha_1 + \gamma_k + \beta)}{1 + \exp(\alpha_1 + \gamma_k + \beta)},$$

$$\text{logit}(p_{2k1}) = \alpha_2 + \gamma_k + \beta + \delta \qquad \text{and} \qquad p_{2k1} = \frac{\exp(\alpha_1 + \gamma_k + \beta + \delta)}{1 + \exp(\alpha_1 + \gamma_k + \beta + \delta)}\,.$$

Hence, under a wrong specification of the applied model, the linear approximation of the estimates $\hat{\alpha}_1^*$, $\hat{\alpha}_2^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ can be expressed as

$$\begin{aligned}
\hat{\alpha}_i^* &\approx \text{logit}(p_{i0}^*) + z_{i0} \cdot \sqrt{\frac{n_{i0}^* p_{i0}^* - \sum_{k=1}^{2} n_{ik0} p_{ik0}^2}{{n_{i0}^*}^2 {p_{i0}^*}^2 (1 - \text{p}_{i0}^*)^2}} \\
&= \text{logit}\Big(\frac{1}{n_{i0}^*} \sum_{k=1}^{2} n_{ik0} p_{ik0}\Big) + z_{i0} \cdot \sqrt{\frac{n_{i0}^* \big(\frac{1}{n_{i0}^*} \sum_{k=1}^{2} n_{ik0} p_{ik0}\big) - \sum_{k=1}^{2} n_{ik0} p_{ik0}^2}{{n_{i0}^*}^2 \big(\frac{1}{n_{i0}^*} \sum_{k=1}^{2} n_{ik0} p_{ik0}\big)^2 \big(1 - \big(\frac{1}{n_{i0}^*} \sum_{k=1}^{2} n_{ik0} p_{ik0}\big)\big)^2}} \\
&= \text{logit}\Big(\frac{1}{n_{i0}^*} \sum_{k=1}^{2} n_{ik0} p_{ik0}\Big) + z_{i0} \cdot \frac{n_{i0}^* \cdot \sqrt{\sum_{k=1}^{2} n_{ik0} p_{ik0} (1 - p_{ik0})}}{\sum_{k=1}^{2} n_{ik0} p_{ik0} (n_{i0}^* - \sum_{k=1}^{2} n_{ik0} p_{ik0})} \\
&= \alpha_i^* + z_{i0} \cdot \text{Se}(\hat{\alpha}_i^*),
\end{aligned}$$

where

$$\alpha_i^* = \text{logit}\left(\frac{1}{n_{i0}^*} \sum_{k=1}^{2} n_{ik0} \cdot \frac{\exp(\alpha_i + \gamma_k)}{1 + \exp(\alpha_i + \gamma_k)}\right).$$

and

$$\text{Se}(\hat{\alpha}_i^*) = \frac{n_{i0}^* \cdot \sqrt{\sum_{k=1}^{2} n_{ik0} \cdot \frac{\exp(\alpha_i + \gamma_k)}{1 + \exp(\alpha_i) + \gamma_k} \left(1 - \frac{\exp(\alpha_i + \gamma_k)}{1 + \exp(\alpha_i + \gamma_k)}\right)}}{\sum_{k=1}^{2} n_{ik0} \frac{\exp(\alpha_i + \gamma_k)}{1 + \exp(\alpha_i + \gamma_k)} \left(n_{i0}^* - \sum_{k=1}^{2} n_{ik0} \frac{\exp(\alpha_i + \gamma_k)}{1 + \exp(\alpha_i + \gamma_k)}\right)}$$

Similarly define

$$\beta^* = \text{logit}\left(\frac{1}{n_{11}^*} \sum_{k=1}^{2} n_{1k1} \cdot \frac{e^{\alpha_1 + \gamma_k + \beta}}{1 + e^{\alpha_1 + \gamma_k + \beta}}\right) - \text{logit}\left(\frac{1}{n_{10}^*} \sum_{k=1}^{2} n_{1k0} \cdot \frac{e^{\alpha_1 + \gamma_k}}{1 + e^{\alpha_1 + \gamma_k}}\right),$$

$$\delta^* = \text{logit}\left(\frac{1}{n_{21}^*} \sum_{k=1}^{2} n_{2k1} \cdot \frac{e^{\alpha_2 + \gamma_k + \beta + \delta}}{1 + e^{\alpha_2 + \gamma_k + \beta + \delta}}\right) - \text{logit}\left(\frac{1}{n_{20}^*} \sum_{k=1}^{2} n_{2k0} \cdot \frac{e^{\alpha_2 + \gamma_k}}{1 + e^{\alpha_2 + \gamma_k}}\right) - \hat{\beta}^*,$$

then,

$$\hat{\beta}^* \approx \beta^* + z \cdot \mathrm{Se}(\hat{\beta}^*)$$
$$\hat{\delta}^* \approx \delta^* - \hat{\beta}^* + z \cdot \mathrm{Se}(\hat{\delta}^*),$$

such that

$$\mathrm{Se}(\hat{\beta}^*) = \sqrt{\frac{\sum_{k=1}^{2} n_{1k1} \cdot \frac{e^{\alpha_1+\gamma_k+\beta}}{1+e^{\alpha_1+\gamma_k+\beta}} \left(1 - \frac{e^{\alpha_1+\gamma_k+\beta}}{1+e^{\alpha_1+\gamma_k+\beta}}\right)}{\left(\sum_{k=1}^{2} n_{1k1} \frac{e^{\alpha_1+\gamma_k+\beta}}{1+e^{\alpha_1+\gamma_k+\beta}}\right)^2 \left(1 - \frac{1}{n_{11}^*} \sum_{k=1}^{2} n_{1k1} \frac{e^{\alpha_1+\gamma_k+\beta}}{1+e^{\alpha_1+\gamma_k+\beta}}\right)^2} + \frac{\sum_{k=1}^{2} n_{1k0} \cdot \frac{e^{\alpha_1+\gamma_k}}{1+e^{\alpha_1+\gamma_k}} \left(1 - \frac{e^{\alpha_1+\gamma_k}}{1+e^{\alpha_1+\gamma_k}}\right)}{\left(\sum_{k=1}^{2} n_{1k0} \frac{e^{\alpha_1+\gamma_k}}{1+e^{\alpha_1+\gamma_k}}\right)^2 \left(1 - \frac{1}{n_{10}^*} \sum_{k=1}^{2} n_{1k0} \frac{e^{\alpha_1+\gamma_k}}{1+e^{\alpha_1+\gamma_k}}\right)^2}},$$

$$\mathrm{Se}(\hat{\delta}^*) = \sqrt{\frac{\sum_{k=1}^{2} n_{2k1} \cdot \frac{e^{\alpha_2+\gamma_k+\beta+\delta}}{1+e^{\alpha_2+\gamma_k+\beta+\delta}} \left(1 - \frac{e^{\alpha_2+\gamma_k+\beta+\delta}}{1+e^{\alpha_2+\gamma_k+\beta+\delta}}\right)}{\left(\sum_{k=1}^{2} n_{2k1} \frac{e^{\alpha_2+\gamma_k+\beta+\delta}}{1+e^{\alpha_2+\gamma_k+\beta+\delta}}\right)^2 \left(1 - \frac{1}{n_{21}^*} \sum_{k=1}^{2} n_{2k1} \frac{e^{\alpha_1+\gamma_k+\beta+\delta}}{1+e^{\alpha_1+\gamma_k+\beta+\delta}}\right)^2} + \frac{\sum_{k=1}^{2} n_{2k0} \cdot \frac{e^{\alpha_2+\gamma_k}}{1+e^{\alpha_2+\gamma_k}} \left(1 - \frac{e^{\alpha_2+\gamma_k}}{1+e^{\alpha_2+\gamma_k}}\right)}{\left(\sum_{k=1}^{2} n_{2k0} \frac{e^{\alpha_2+\gamma_k}}{1+e^{\alpha_2+\gamma_k}}\right)^2 \left(1 - \frac{1}{n_{20}^*} \sum_{k=1}^{2} n_{2k0} \frac{e^{\alpha_2+\gamma_k}}{1+e^{\alpha_2+\gamma_k}}\right)^2}}.$$

The theoretical behaviour of the estimators in this example is complicated to interpret. So far, the asymptotic Normality distribution of y_{it} has been considered to permit a smooth study of the theoretical behaviour of the estimators. The exact distribution of y_{it}, which is a sum of non-identically distributed, independent Binomials, is not workable at the moment. An alternative method for making theoretical asymptotic behaviour of maximum likelihood estimates perceptible is by using the asymptotic results of the estimates under misspecification, which were obtained by White and reviewed by Chow. But before moving on to employ White's results, it is necessary to find the nearest parameters to the truth, the parameter vector that minimises the KL-divergence function between the applied model and the truth.

5.4.3 Distance Between Applied Model and the Truth

Consider the truth model probability structure (5.2) and the applied model probability structure (5.1). Let g denotes the truth probability function and f denotes the applied probability function. For each y_{it} in g, the distance between the applied structure and the truth is

$$D_{KL}(g(y_{it}) \| f(y_{it})) = \mathrm{E}_g \Big[\log \frac{g(y_{it})}{f(y_{it})} \Big].$$

On the basis on the independence property of data values y_{it}, the KL-divergence is given by

$$D_{KL}(g\|f) = \mathrm{E}_g\Big[\log\frac{\prod_{i,t} g(y_{it})}{\prod_{i,t} f(y_{it})}\Big] = \sum_{i,t}\mathrm{E}_g\Big[\log\frac{g(y_{it})}{f(y_{it})}\Big] = \sum_{i,t} D_{KL}(g(y_{it})\|f(y_{it}))\,. \tag{5.24}$$

For the moment, we seek for simplicity, so in what follows, we drop the i,t subscript and work with $D_{KL}(g\|f)$. The structure of y_{it} is diminished to y with truth g given by

$$Y = \sum_{k=1}^{K} Y_k \quad \text{where} \quad Y_k \sim \mathrm{Bin}(n_k, p_k), \tag{5.25}$$

and applied model probability structure f given by

$$Y \sim \mathrm{Bin}(n^*, p^*) \tag{5.26}$$

Using $K = 2$, then for y in g, the truth distribution of a sum of two non-identically distributed, independent Binomials has been discussed above in (5.22) and was concluded it is difficult to derive. Therefore the derivation of the KL-divergence in this case is complex. Alternatively, we can assume n_k is large, then the distribution of each y_k in g approaches Normal such that

$$Y_k \sim \mathrm{N}(n_k p_k, n_k p_k(1-p_k)),$$

and so the distribution of y in g is a sum of two non-identically distributed, independent normals, $y = \sum_{k=1}^{2} y_k$, which is given by

$$Y \sim \mathrm{N}(\sum_{k=1}^{2} n_k p_k, \sum\nolimits_k n_k p_k(1-p_k))\,. \tag{5.27}$$

Given that $n^* = \sum_{k=1}^{2} n_k$, then the applied structure approaches Normal as well, such that

$$Y \sim \mathrm{N}(n^* p^*, n^* p^*(1-p^*))\,. \tag{5.28}$$

To compute the KL-divergence for y in g, we need first to get the $\log g(y)$ and $\log f(y)$. Given that $\mu_g = \sum_{k=1}^{2} n_k p_k$ and $\sigma_g^2 = \sum_{k=1}^{2} n_k p_k(1-p_k)$, the $\log g(y)$ is given by

$$\log g(y) = \log\Big(\frac{1}{\sqrt{2\pi}}\cdot\frac{1}{\sigma_g}\cdot\exp\big(-\frac{(y-\mu_g)^2}{2\sigma_g^2}\big)\Big) = \log\frac{1}{\sqrt{2\pi}} - \log\sigma_g - \frac{(y-\mu_g)^2}{2\sigma_g^2},$$

and given that $\mu_f = n^*p^*$ and $\sigma_g^2 = n^*p^*(1-p^*)$, the $\log f(y)$ is given by

$$\log f(y) = \log\Big(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma_f} \cdot \exp\big(-\frac{(y-\mu_f)^2}{2\sigma_f^2}\big)\Big) = \log\frac{1}{\sqrt{2\pi}} - \log\sigma_f - \frac{(y-\mu_f)^2}{2\sigma_f^2},$$

and so

$$\begin{aligned}
\log\frac{g(y)}{f(y)} &= \log(g(y)) - \log(f(y)) \\
&= \log(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma_g} \cdot \exp(-\frac{(y-\mu_g)^2}{2\sigma_g^2}) - \log(\frac{1}{\sqrt{2\pi}} \cdot \frac{1}{\sigma_f} \cdot \exp(-\frac{(y-\mu_f)^2}{2\sigma_f^2}) \\
&= \log(\frac{1}{\sigma_g}) - \log(\frac{1}{\sigma_f}) - \frac{(y-\mu_g)^2}{2\sigma_g^2} + \frac{(y-\mu_f)^2}{2\sigma_f^2} \\
&= \log(\frac{\sigma_f}{\sigma_g}) - \frac{(y-\mu_g)^2}{2\sigma_g^2} + \frac{(y-\mu_f)^2}{2\sigma_f^2}\,.
\end{aligned}$$

For y in g, the KL-divergence is then

$$\begin{aligned}
D_{KL}(g\|f) &= \mathrm{E}_g\Big[\log\frac{g(y)}{f(y)}\Big] \\
&= \mathrm{E}_g\Big[\log(\frac{\sigma_f}{\sigma_g}) - \frac{(y-\mu_g)^2}{2\sigma_g^2} + \frac{(y-\mu_f)^2}{2\sigma_f^2}\Big] \\
&= \log(\frac{\sigma_f}{\sigma_g}) - \frac{\mathrm{E}_g[(y-\mu_g)^2]}{2\sigma_g^2} + \frac{\mathrm{E}_g[(y-\mu_f)^2]}{2\sigma_f^2} \\
&= \log(\frac{\sigma_f}{\sigma_g}) - \frac{\mathrm{E}_g[(y-\mu_g)^2]}{2\sigma_g^2} + \frac{\mathrm{E}_g[(y-\mu_g)^2] + (\mu_g-\mu_f)^2 + 2(\mu_g-\mu_f)\mathrm{E}_g[(y-\mu_g)]}{2\sigma_f^2} \\
&= \log(\frac{\sigma_f}{\sigma_g}) - \frac{\sigma_g^2}{2\sigma_g^2} + \frac{\sigma_g^2 + (\mu_g-\mu_f)^2}{2\sigma_f^2} \\
&= \log(\frac{\sigma_f}{\sigma_g}) + \frac{\sigma_g^2 + (\mu_g-\mu_f)^2}{2\sigma_f^2} - \frac{1}{2}\,.
\end{aligned} \tag{5.29}$$

This means the applied structure f of y diverges from the truth g by $\log(\frac{\sigma_f}{\sigma_g}) + \frac{\sigma_g^2+(\mu_g-\mu_f)^2}{2\sigma_f^2} - \frac{1}{2}$. The KL-divergence here is a function of μ_f, σ_f, μ_g, and σ_g or in particular, of n_1, n_2, p_1, p_2 and p^*. The best value to use for p^* in the applied structure to provide a closest structure to the truth is the one that minimises the KL-divergence $D_{KL}(g\|f)$. Let $\tilde{p^*}$ be the value that minimises the KL-divergence, which can be derived by differentiating $D_{KL}(g(y)\|f(y))$ with respect to p^*. Provided the mean μ_f and the variance σ_f^2 in term of p^*, then $D_{KL}(g\|f)$ in (5.29) is given as

$$D_{KL}(g\|f) = \log\sqrt{n^*p^*(1-p^*)} - \log\sigma_g + \frac{\sigma_g^2 + (\mu_g - n^*p^*)^2}{2np^*(1-p^*)} - \frac{1}{2}\,.$$

By differentiating $D_{KL}(g(y)\|f(y))$ with respect to p^*, we get

$$\frac{\partial D_{KL}}{\partial p^*} = \frac{n^*(1-p^*) - n^*p^*}{2n^*p^*(1-p^*)} + \frac{-4n^{*2}p^*(1-p^*)(\mu_g - n^*p^*) - [\sigma_g^2 + (\mu_g - n^*p^*)^2][2n^*(1-p^*) - 2n^*p^*]}{(2n^*p^*(1-p^*))^2}$$

$$= \frac{n^* - 2np^*}{2n^*p^*(1-p^*)} + \frac{(-4n^{*2}p^* + 4n^{*2}p^{*2})(\mu_g - n^*p^*) - [\sigma_g^2 + (\mu_g - n^*p^*)^2](2n^* - 4np^*)}{(2n^*p^*(1-p^*))^2}$$

$$= \frac{\begin{array}{r}(n^* - 2n^*p^*)[2n^*p^*(1-p^*)] - 4n^{*2}\mu_g p^* + 4n^{*3}p^{*2} + 4n^{*2}\mu_g p^{*2} - 4n^{*3}p^{*3} \\ - 2n^*\sigma_g^2 + 4n^*p^*\sigma_g^2 - 2n^*(\mu_g - n^*p^*)^2 + 4n^*p^*(\mu_g - n^*p^*)^2\end{array}}{(2n^*p^*(1-p^*))^2}$$

$$= \frac{\begin{array}{r}2n^{*2}p^* - 2n^{*2}p^{*2} - 4n^{*2}p^{*2} + 4n^{*2}p^{*3} - 4n^{*2}\mu_g p^* + 4n^{*3}p^{*2} + 4n^{*2}\mu_g p^{*2} - 4n^{*3}p^{*3} \\ - 2n^*\sigma^2 + 4n^*\sigma^2 p^* - 2^*n\mu_g^2 + 4n^{*2}\mu_g p^* - 2n^{*3}p^{*2} + 4n^*\mu_g^2 p^* - 8n^{*2}\mu_g p^{*2} + 4n^{*3}p^{*3}\end{array}}{(2n^*p^*(1-p^*))^2}$$

$$= \frac{4n^{*2}p^{*3} - 6n^{*2}p^{*2} + 2n^{*3}p^{*2} - 4n^{*2}\mu_g p^{*2} + 2n^{*2}p^* + 4n^*\sigma_g^2 p^* + 4n^*\mu_g^2 p^* - 2n^*(\sigma_g^2 + \mu_g^2)}{4n^{*2}p^{*2}(1-p^*)^2}$$

$$= \frac{2n^*p^{*3} - 3n^*p^{*2} + n^{*2}p^{*2} - 2n^*\mu_g p^{*2} + n^*p^* + 2\sigma_g^2 p^* + 2\mu_g^2 p^* - (\sigma_g^2 + \mu_g^2)}{2n^*p^{*2}(1-p^*)^2}$$

$$= \frac{2n^*p^{*3} + (-3n^* + n^{*2} - 2n^*\mu_g)p^{*2} + (n^*p^* + 2\sigma_g^2 + 2\mu_g^2) - (\sigma_g^2 + \mu_g^2)}{2n^*p^{*2}(1-p^*)^2}\,.$$

To find $\tilde{p}^*$, the best value for p^*, we set $\frac{\partial D_{KL}}{\partial p^*}$ equals to zero, which gives

$$2n^*p^{*3} + (-3n^* + n^{*2} - 2n^*\mu_g)p^{*2} + (n^*p + 2\sigma_g^2 + 2\mu_g^2) - (\sigma_g^2 + \mu_g^2) = 0\,.$$

Finding the roots of this cubic equation is not straightforward, but approximations of the roots can be found numerically through root-finding algorithms such as Newton's method. However, we will not consider numerical approximations of the roots.

Considering the KL-divergence function (5.29), then the applied structure includes the truth if the KL-divergence $D_{KL}(g\|f)$ converges to zero. For this example the KL-divergence is minimised to zero by taking $\mu_f = \mu_g$ and $\sigma_f = \sigma_g$. More specifically, the KL-divergence is minimised to zero, i.e. $D_{KL}(g\|f) = 0$ if and only if $p_1 = p_2 = p^*$, by taking into consideration the assumption that n_1, n_2 and n^* are

large enough. **Figure 5.2** could elucidate that the KL-divergence meets zero when $p_1 = p_2 = p^*$, considering that $n_1 = n_2$. This asymptotic result can be readily seen on the figure when n_1 and n_2 are very large.

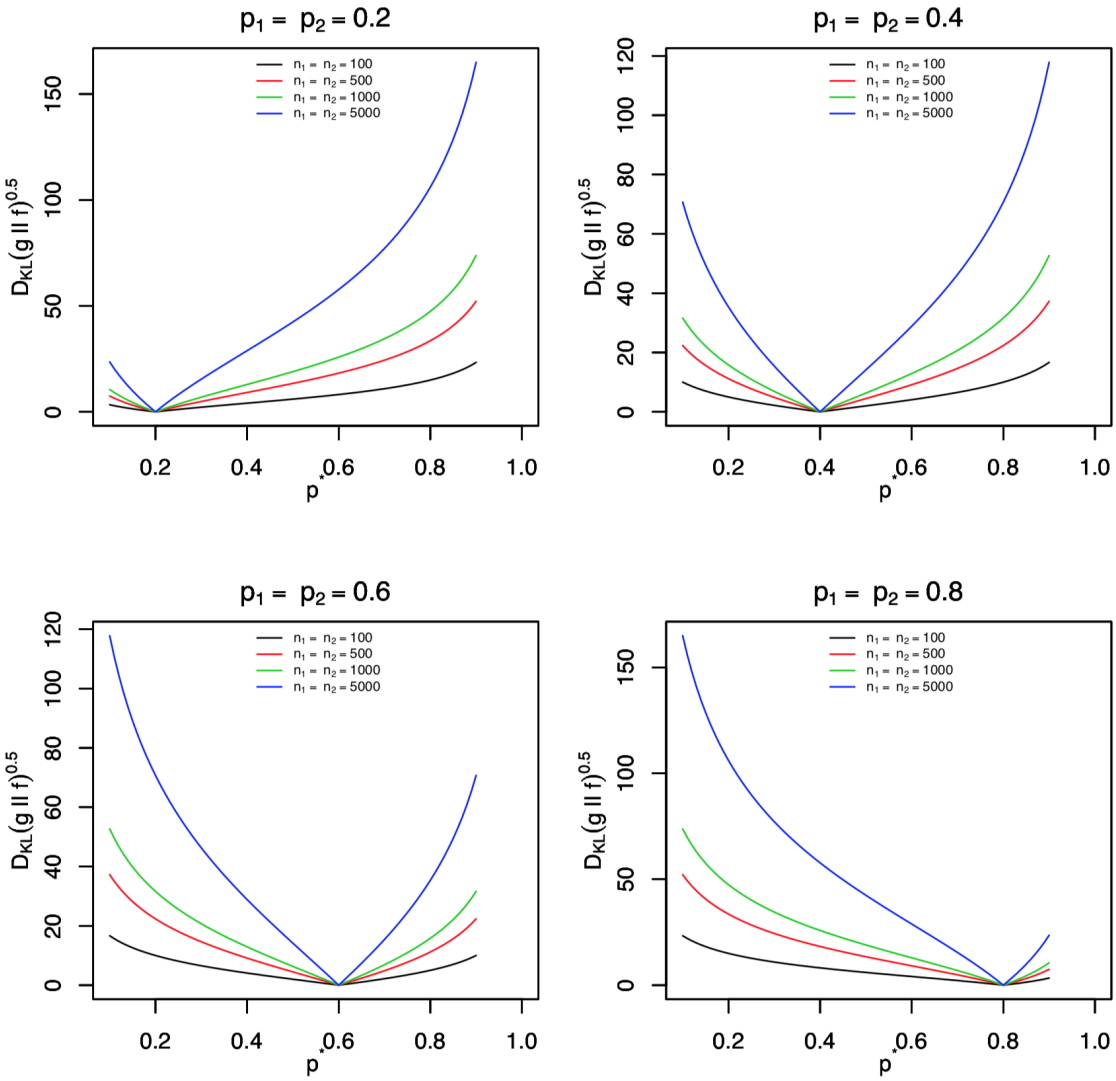
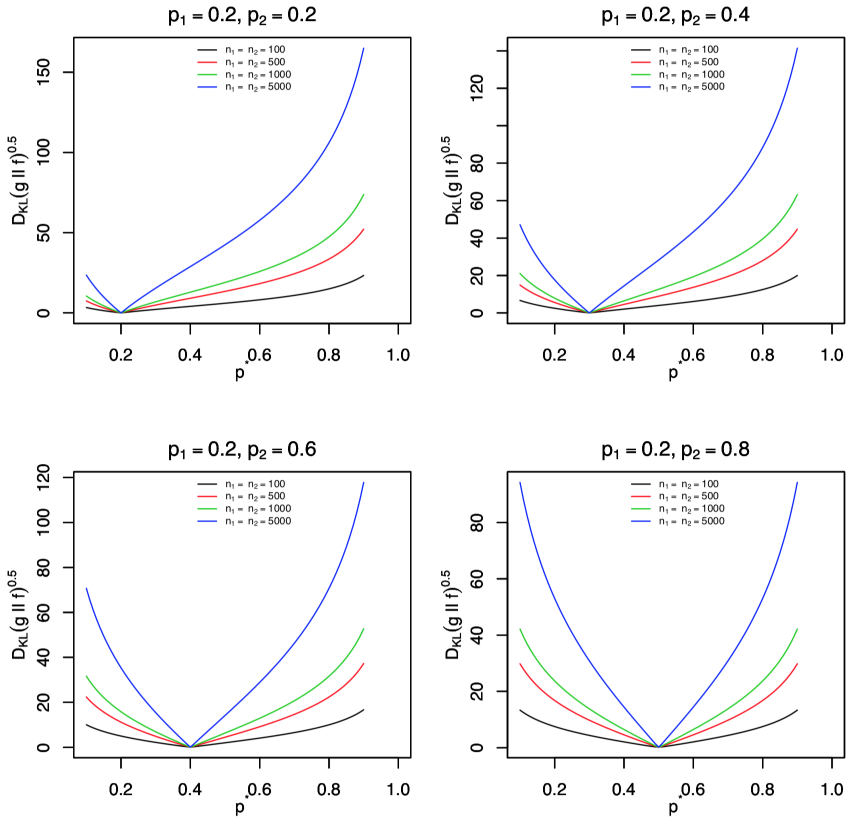


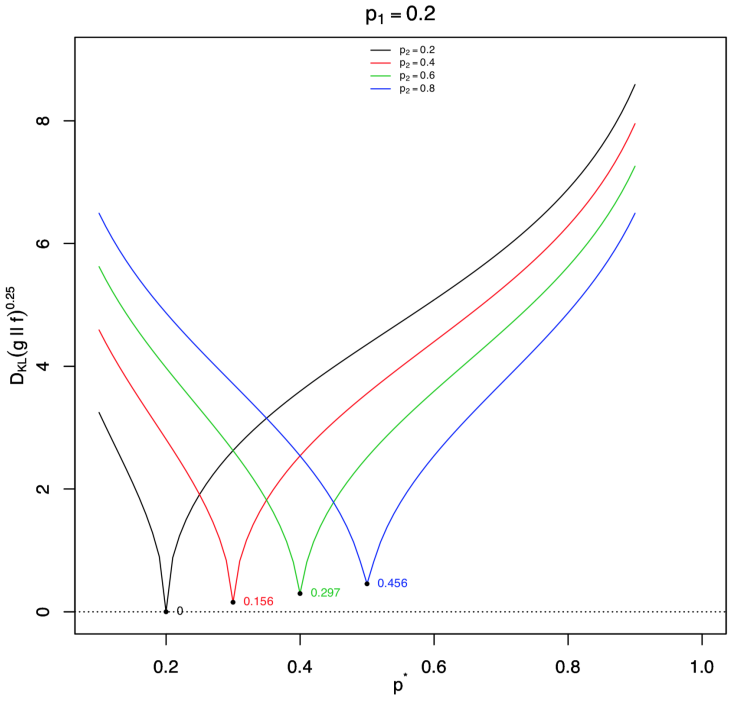
Figure 5.2: $(D_{KL}(g\|f))^{0.5} = 0$ at $p_1 = p_2 = p^*$

But when $p_1 \neq p_2$, the KL-divergence is expected to deviate from zero especially if n_1 and n_2 are large enough. Consider the example presented in **Figure 5.3b**, given that $n_1 = n_2 = 1000$, the value of $D_{KL}(g\|f)$ diverges from zero distinctly when $p_1 = 0.2$ and $p_2 \in \{0.4, 0.6, 0.8\}$. However, it appears that for $n_1 = n_2$ the best value of p^* that produces almost a zero value of the KL-divergence is $\tilde{p}^* = \frac{1}{2}(p_1 + p_2)$ as presented in **Figure 5.3**.

The below numerical patterns, however, do not support the idea of studying the theoretical asymptotic behaviour of the estimates. Additionally, by working on the assumption of the asymptotic Normal distribution, we deviate from the original structure of the truth which is a summation of k independent, non-identically distributed Binomial groups. The assumption of asymptotic normality distribution has been used to unify the structure for both truth and the applied models, to make no sharp projection in the KL-divergence computation. There is, therefore, a definite need of computing the KL-divergence with retain possession of the truth structure. We answer this need by proposing a proxy structure of the applied model, that implicates the truth. The main concept of the proxy structure is to bind applied structure to the truth, which thence permits computing the KL-divergence with no failure in addressing the truth.



(a)



(b)

Figure 5.3: $D_{KL}(g\|f) = 0$ at $\tilde{p^*} = \frac{1}{2}(p_1 + p_2)$

5.5 Proxy Structure

One possible way to think of the applied model probability structure (5.1) is that Y_{it} is a sum of independent, non-identically distributed Binomials over k strata, such that

$$Y_{it} = \sum_k Y_{ikt}, \quad \text{where} \quad Y_{ikt} \sim \text{Bin}(n_{ikt}, p^{\dagger}_{ikt}), \tag{5.30}$$

where each $p^{\dagger}_{ikt} = p^{*}_{it}$ and $\sum_k n_{ikt} = n^{*}_{it}$. By comparing this structure to the truth (5.2), the probability of purchasing here is identical between the strata, i.e. $p^{\dagger}_{ikt} = p^{*}_{it}$. This new proposed structure of the applied model is considered as a proxy for the applied structure. To retain the applied structure, the term proxy structure will be used to refer to the probability structure (5.30).

The probability $p^{\dagger}_{ikt}$ of purchasing is modelled in a similar basis of the applied but with consideration of the truth structure, i.e.

$$\text{logit}(p^{\dagger}_{ikt} = p^{*}_{it}) = \eta^{\dagger}_{ikt} = \alpha^{*}_{i} + \beta^{*} t + \delta^{*} C_{it} = \sum\nolimits_j x_{iktj} h_j \theta^{*}_{j}, \quad \text{so that} \quad \eta^{\dagger} = XH\theta^{*}\,. \tag{5.31}$$

Where X consists of k matrices $X^{(k=1)}, X^{(k=2)}, \ldots, X^{(k=K)}$ piled on top of each other and H is a transformation matrix that maps the structure of θ^{*} to the structure of θ. Also $X^{(k)}H$ is the same for any $k \in K$ so that $X^{(k)}H = X^{*}$ and $\eta^{\dagger}$ is a vector consisting of k vectors $\eta^{\dagger(k)}$ piled on top of each other, so that $\eta^{\dagger(k)} = \eta^{*}$ for all k. An illustration of the breakdown of Y_{it} using three structures: the truth, the proxy and the applied, are provided in **Figure 5.4**.

To illustrate the proxy structure, consider the following toy example.

Example 3 For the toy models, the applied model η^{*} (5.20) is given by

$$\eta^{*} = X^{*}\theta^{*} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} \begin{bmatrix} \alpha^{*}_{1} \\ \alpha^{*}_{2} \\ \beta^{*} \\ \delta^{*} \end{bmatrix},$$

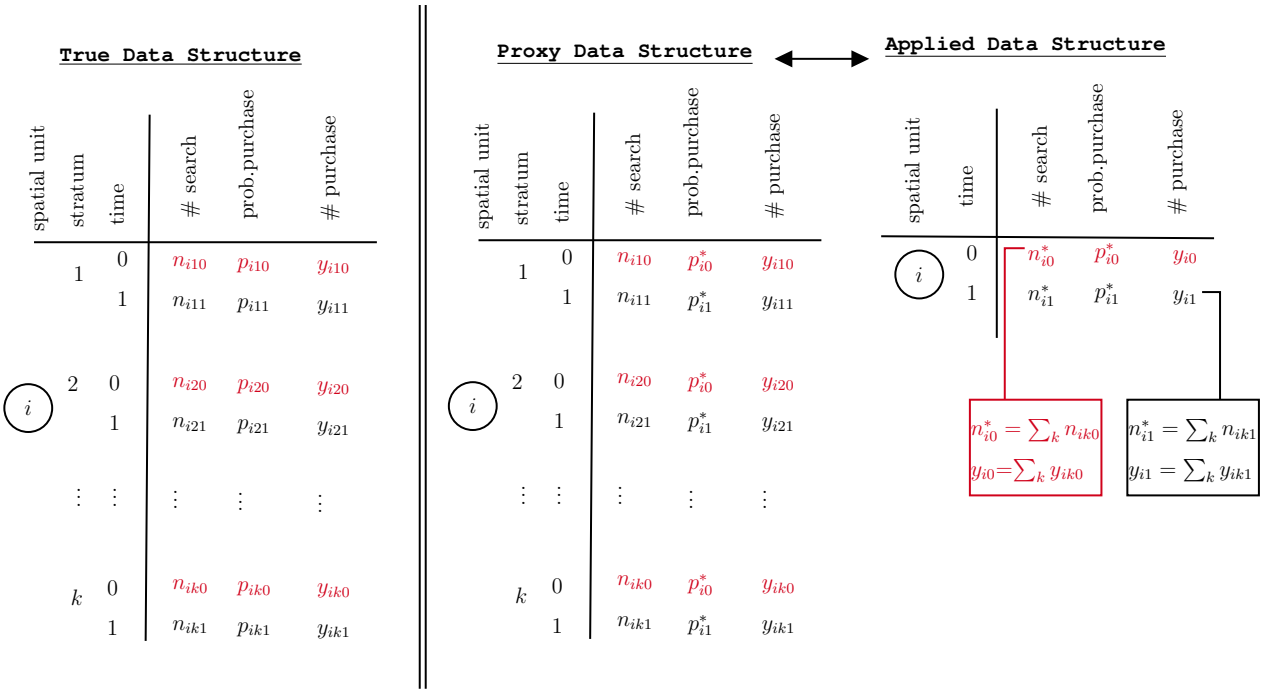


Figure 5.4: An illustration of the breakdown of Y_{it}, using applied, proxy and truth probability structure. The applied structure is characterised by n_{it} and p^*_{it}, the proxy structure is characterised by n_{ikt} and p^*_{it} and the truth structure is characterised by n_{ikt} and p_{ikt}

and the truth η (5.21) is

$$\eta = X\theta = \left(\begin{array}{cccccc} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \end{array}\right) \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \\ \beta \\ \delta \end{bmatrix},$$

We can think of X as a set of two matrices $X^{(k=1)}$ and $X^{(k=2)}$. To illustrate the breakdown of the proxy model we need to set a matrix H. Consider

$$H = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{then,} \qquad H\theta^* = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} \alpha_1^* \\ \alpha_2^* \\ \beta^* \\ \delta^* \end{bmatrix} = \begin{bmatrix} \alpha_1^* \\ \alpha_2^* \\ 0 \\ 0 \\ \beta^* \\ \delta^* \end{bmatrix}.$$

Thus, the breakdown of $\eta^\dagger$ is

$$XH\theta^* = \left(\begin{array}{cccccc} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \end{array}\right) \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} \alpha_1^* \\ \alpha_2^* \\ \beta^* \\ \delta^* \end{bmatrix},$$

such that

$$X^{(k=1)}H = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} = X^*,$$

and

$$X^{(k=2)}H = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{pmatrix} = X^*.$$

$X^{(k=1)}H = X^{(k=2)}H = X^*$, because matrix H will always include k zero rows, that cancel a non-zero kth column in each $X^{(k)}$. Therefore,

$$XH\theta^* = \left(\frac{X^{(k=1)}H}{X^{(k=2)}H}\right)\theta^* = \left(\frac{X^*\theta^*}{X^*\theta^*}\right) .$$

By drawing on the structure of toy proxy model, we can see that the proxy probability structure $\eta^\dagger$ is a vector of two applied probability structures $[\eta^{\dagger(1)} \quad \eta^{\dagger(2)}]^T$, where $\eta^{\dagger(1)} = \eta^{\dagger(2)} = \eta^*$.

Example 4 Given the true purchase model (4.12) for $i \in \{1, 2, \ldots, I\}$ and $k \in \{1, 2, \ldots, K\}$, and the applied purchase model (4.9), the H matrix is then given by

$$H = \begin{array}{c} \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_K \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \\ \delta_1 \\ \delta_2 \\ \vdots \\ \delta_K \end{array} \begin{array}{c} \begin{array}{cccccc} \alpha_1 & \alpha_2 & \ldots & \alpha_I & \beta & \delta \end{array} \\ \begin{pmatrix} 1 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & \ldots & 0 & 1 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 1 & 0 \\ 0 & 0 & \ldots & 0 & 0 & 1 \\ 0 & 0 & \ldots & 0 & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & 0 & 0 & 1 \end{pmatrix} \end{array}$$

5.5.1 Distance Between the Proxy model and Truth

Let y_{it} be in the truth probability structure g, then the distance of the proxy probability structure f from the truth g is given by

$$\begin{aligned} D_{KL}(g\|f) = \mathrm{E}_g\left[\log\frac{g(y_{it})}{f(y_{it})}\right] &= \mathrm{E}_g\Big[\log\frac{\prod_{i,k,t} g(y_{ikt})}{\prod_{i,k,t} f(y_{ikt})}\Big] \\ &= \sum_{i,k,t}\mathrm{E}_g\Big[\log\frac{g(y_{ikt})}{f(y_{ikt})}\Big] = \sum_{i,k,t} D_{KL}(g(y_{ikt})\|f(y_{ikt}))\,. \end{aligned} \tag{5.32}$$

This form of $D_{KL}(g\|f)$ is a general form of the one that has been presented in (5.24), where $D_{KL}(g(y_{ikt})\|f(y_{ikt}))$ for y_{ikt} in g is given by

$$\begin{aligned} D_{KL}(g(y_{ikt})\|f(y_{ikt})) &= \mathrm{E}_g[\log g(y_{ikt}) - \log f(y_{ikt})] \\ &= \mathrm{E}_g\Big[\log(p_{ikt}^{y_{ikt}}(1-p_{ikt})^{n_{ikt}-y_{ikt}} - \log(p_{it}^{*\,y_{ikt}}(1-p_{it}^{*})^{n_{ikt}-y_{ikt}}\Big] \\ &= \mathrm{E}_g[y_{ikt}]\log\frac{p_{ikt}}{1-p_{ikt}} + n_{ikt}\log(1-p_{ikt}) \\ &\quad - \mathrm{E}_g[y_{ikt}]\log\frac{p_{it}^{*}}{1-p_{it}^{*}} - n_{ikt}\log(1-p_{it}^{*}) \\ &= n_{ikt}p_{ikt}\log\frac{p_{ikt}}{1-p_{ikt}} + n_{ikt}\log(1-p_{ikt}) \\ &\quad - n_{ikt}p_{ikt}\log\frac{p_{it}^{*}}{1-p_{it}^{*}} - n_{ikt}\log(1-p_{it}^{*}), \end{aligned}$$

and hence,

$$\begin{aligned} D_{KL}(g\|f) = \sum\nolimits_{i,k,t}\Big[&n_{ikt}p_{ikt}\log\frac{p_{ikt}}{1-p_{ikt}} + n_{ikt}\log(1-p_{ikt}) \\ &- n_{ikt}p_{ikt}\log\frac{p_{it}^{*}}{1-p_{it}^{*}} - n_{ikt}\log(1-p_{it}^{*})\Big]\,. \end{aligned} \tag{5.33}$$

Using matrix notation of the proxy model, then

$$\begin{aligned} D_{KL}(g\|f) = \sum\nolimits_{i,k,t}\Big[&n_{ikt}p_{ikt}\sum\nolimits_j x_{iktj}\theta_j - n_{ikt}\log\big(1+\exp\big(\sum\nolimits_j x_{iktj}\theta_j\big)\big) \\ &- n_{ikt}p_{ikt}\sum\nolimits_j x_{iktj}h_j\theta_j^{*} + n_{ikt}\log\big(1+\exp\big(\sum\nolimits_j x_{iktj}h_j\theta_j^{*}\big)\big)\Big]\,. \end{aligned} \tag{5.34}$$

By taking the first derivative of the KL-divergence $D_{KL}(g\|f)$ with respect to θ_j^{*}, we get

$$\frac{\partial D_{KL}}{\partial\theta_j^{*}} = \sum_{i,k,t}\Big[-n_{ikt}p_{ikt}x_{iktj}h_j + n_{ikt}\big(1+\exp\big(\sum\nolimits_j x_{iktj}h_j\theta_j^{*}\big)\big)^{-1}x_{iktj}h_j\Big]\,. \tag{5.35}$$

Given that $\tilde{\theta}^{*}$ is the parameter vector that minimises the KL-divergence. Then $\tilde{\theta}^{*}$ is the best parameter vector to use in the proxy or applied models for θ^{*} to provide the closest model to the truth. Again, there is no need to solve equation (5.35) analytically to find

$\tilde{\theta}^*$. A numerical optimisation procedure will be used to approximate $\tilde{\theta}^*$ for real cases. However, we can derive the best estimate of θ^* when a toy model is applied as illustrated in the example below.

Example 5 Consider the truth structure g in (5.2) and the the proxy structure f in (5.30) of y_{it} for $i \in \{1,2\}$, $k \in \{1,2\}$ and $t \in \{0,1\}$. Let y_{it} in g, then by using expression (5.33), $D_{KL}(g\|f)$ can be written as

$$D_{KL}(g\|f) = \sum\nolimits_{i,k,t} \Big[n_{ikt}p_{ikt}\log p_{ikt} + n_{ikt}\log(1-p_{ikt}) - n_{ikt}p_{ikt}\log(1-p_{ikt}) \\ - n_{ikt}p_{ikt}\log p^*_{it} - n_{ikt}\log(1-p^*_{it}) + n_{ikt}p_{ikt}(1-p^*_{it}) \Big] .$$

Using the fact that $\eta^\dagger = X^{(k)}H\theta^*_j = X^*\theta^* = \eta^*$ for any k, then $\sum_j x_{iktj}h_j\theta^*_j = \sum_j x^*_{itj}\theta^*_j$. Provided that p^*_{it} is a composite function of θ^*, then the derivative of $D_{KL}(g\|f)$ with respect to θ^* can be found using chain rule as

$$\frac{\partial D_{KL}}{\partial \theta^*_j} = \sum_{i,t} \frac{\partial D_{KL}}{\partial p^*_{it}} \frac{\partial p^*_{it}}{\partial \eta^*_{it}} \cdot \frac{\partial \eta^*_{it}}{\partial \theta^*_j} .$$

For this toy model, we have seen in Example 1 there is a bijection transformations $(U \circ T)(\theta^*)$, thereby

$$\frac{\partial D_{KL}}{\partial \theta^*_j} = 0 \iff \frac{\partial D_{KL}}{\partial p^*_{it}} = 0 .$$

By differentiating $D_{KL}(g\|f)$ with respect to p^*_{it}, we get

$$\begin{aligned} \frac{\partial D_{KL}}{\partial p^*_{it}} &= \sum\nolimits_k \frac{-n_{ikt}p_{ikt}}{p^*_{it}} + \frac{n_{ikt} - n_{ikt}p_{ikt}}{1-p^*_{it}} \\ &= \sum\nolimits_k \frac{-n_{ikt}p_{ikt} + n_{ikt}p^*_{it}}{p^*_{it}(1-p^*_{it})}, \end{aligned}$$

to find $\tilde{p^*_{it}}$, which is the best value to use of p^*_{it} in the applied model, we need to set this derivative equals zero, i.e.

$$\tilde{p}^*_{it} = \frac{\sum_k n_{ikt}p_{ikt}}{\sum_k n_{ikt}} = \frac{1}{n^*_{it}} \sum\nolimits_k n_{ikt}p_{ikt},$$

because $\sum_k n_{ikt} = n^*_{it}$. Since $\eta^*_{it} = \text{logit}(p^*_{it})$, then the best parameter vector $\tilde{\theta}^*$ of θ^* is

$$\tilde{\alpha}^*_1 = \tilde{\eta}^*_{10} \qquad \tilde{\beta}^* = \tilde{\alpha}^*_1 - \tilde{\eta}^*_{11}$$
$$\tilde{\alpha}^*_2 = \tilde{\eta}^*_{20} \qquad \tilde{\delta}^* = \tilde{\alpha_2}^* + \tilde{\beta}^* - \tilde{\eta}^*_{21}$$

The small scale of this toy model has been discussed in section 5.4.3. Given its truth structure (5.25) and applied structure (5.26), the best parameter $\tilde{p}^*$ to use for p^* is then

given by

$$\tilde{p}^* = \frac{n_1 p_1 + n_2 p_2}{n^*}, \tag{5.36}$$

and when $n_1 = n_2$ then $\tilde{p}^* = \frac{1}{2}(p_1 + p_2)$ which is in agreement with the results illustrated in **Figure 5.3**.

5.5.2 Likelihood Function of Proxy Structure

Consider the proxy probability structure (5.30), then the likelihood is

$$\begin{aligned} L = \prod_{i,t}\prod_{k} f(y_{ikt}) &\equiv \prod_{i,t}\prod_{k} {p^\dagger_{ikt}}^{y_{ikt}}(1-p^\dagger_{ikt})^{n_{ikt}-y_{ikt}} \\ &= \prod_{i,t}\prod_{k} {p^*_{it}}^{y_{ikt}}(1-p^*_{it})^{n_{ikt}-y_{ikt}} \\ &= \prod_{i,t}\prod_{k} \Big(\frac{p^*_{it}}{1-p^*_{it}}\Big)^{y_{ikt}}(1-p^*_{it})^{n_{ikt}}\,. \end{aligned}$$

The log-likelihood is then

$$\begin{aligned} l &= \sum\nolimits_{i,t}\sum\nolimits_{k} y_{ikt}\log\Big(\frac{p^*_{it}}{1-p^*_{it}}\Big) + \sum\nolimits_{k} n_{ikt}\log(1-p^*_{it}) \\ &= \sum\nolimits_{i,t} y_{it}\log\Big(\frac{p^*_{it}}{1-p^*_{it}}\Big) + n^*_{it}\log(1-p^*_{it}) \\ &= \sum\nolimits_{i,t} y_{it}\big(\sum\nolimits_{j} x_{iktj}h_j\theta^*_j\big) - n^*_{it}\log\big(1+e^{\sum_j x_{iktj}h_j\theta^*_j}\big)\,. \end{aligned}$$

Provided that $\sum_j x_{iktj}h_j = \sum_j x^*_{itj}$ for all k. Thus

$$l = \sum\nolimits_{i,t} y_{it}\big(\sum\nolimits_{j} x^*_{itj}\theta^*_j\big) - n^*_{it}\log\big(1+e^{\sum_j x^*_{itj}\theta^*_j}\big)\,.$$

This is exactly the log-likelihood (5.15) of the applied probability structure. Thus, we can study the asymptotic distribution theory of the maximum likelihood estimate $\hat{\theta}^*$ using either the proxy or the applied structures because their log-likelihood functions are the same and so are their maximum likelihood estimators.

Example 6 Consider the applied structure (5.26) of the small scale of the toy model with likelihood function

$$L(p^*, y) = \binom{n^*}{y} {p^*}^{y}(1-p^*)^{n^*-y}\,.$$

The a proxy structure is given by

$$Y \equiv Y_1 + Y_2 \quad \text{where} \quad Y_1 \sim \text{Bin}(n_1, p^*_1 = p^*) \quad \text{and} \quad Y_2 \sim \text{Bin}(n_2, p^*_1 = p^*),$$

where $n^* = n_1 + n_2$. The likelihood of the proxy structure is

$$\begin{aligned} L(p^*, (y_1, y_2)) &= \binom{n_1}{y_1}\binom{n_2}{y_2} p^{*y_1+y_2}(1-p^*)^{n_1-y_1+n_2-y_2} \\ &= \binom{n_1}{y_1}\binom{n_2}{y_2} p^{*y}(1-p^*)^{n^*-y}\,. \end{aligned}$$

Therefore, $L(p^*, (y_1, y_2)) \propto L(p^*, y)$, and hence

$$\arg\max_{p^*} L(p^*, (y_1, y_2)) \equiv \arg\max_{p^*} L(p^*, y)\,.$$

As a result of this, the maximum likelihood estimator p^* can be derived from either structures: applied or proxy, and thence studying the asymptotic theoretical distribution of the estimate p^* will make no difference in using applied or proxy structure.

Consider the applied structure, the log-likelihood is

$$l \equiv y \log p^* + (n^* - y)\log(1-p^*)\,.$$

The score function is

$$\frac{\partial l}{\partial p^*} = \frac{y}{p^*(1-p^*)} - \frac{n^*}{1-p^*} = \frac{y - n^* p^*}{p^*(1-p^*)},$$

and its derivative is

$$\frac{\partial^2 l}{\partial {p^*}^2} = \frac{-y(1-2p^*)}{{p^*}^2(1-p^*)^2} - \frac{n^*}{(1-p^*)^2} = \frac{y(2p^*-1)}{{p^*}^2(1-p^*)^2} - \frac{n^*}{(1-p^*)^2}\,.$$

Let $\hat{p}^*$ be the maximum likelihood estimator of p^*. Provided the consistency property (5.5) of an estimate, then for large sample, $\hat{p}^*$ converges to $\tilde{p}^*$ expressed in (5.36). Based on the asymptotic normality results of the maximum likelihood estimates under misspecification, presented in (5.12), the asymptotic theoretical distribution of $\hat{p}^*$ is then given by

$$\hat{p}^* \overset{a.s}{\sim} \mathrm{N}(\tilde{p}^*, \mathcal{C}(\tilde{p}^*))\,.$$

By considering the truth structure (5.25) for $K = 2$ and $\tilde{p}^*$ (5.36), then

$$\mathrm{E}_g[y] = n_1 p_1 + n_2 p_2 = n^* \tilde{p}^*\,.$$

By using Chow definition of $\mathcal{A}(\tilde{p}^*)$ and $\mathcal{B}(\tilde{p}^*)$ in (5.10), we have

$$\begin{aligned} \mathcal{B}(\tilde{p}^*) = \mathrm{E}_g\left[\left(\frac{\partial l}{\partial p^*}\right)^2\Big|_{\tilde{p}^*}\right] &= \frac{\mathrm{E}_g[(y - n^*\tilde{p}^*)^2]}{{\tilde{p}^*}^2(1-\tilde{p}^*)^2} \\ &= \frac{n_1 p_1(1-p_1) + n_2 p_2(1-p_2)}{{\tilde{p}^*}^2(1-\tilde{p}^*)^2} \\ &= \frac{n_1 p_1 + n_2 p_2 - n_1 p_1^2 - n_2 p_2^2}{{\tilde{p}^*}^2(1-\tilde{p}^*)^2} = \frac{n^*\tilde{p}^* - n_1 p_1^2 - n_2 p_2^2}{{\tilde{p}^*}^2(1-\tilde{p}^*)^2}, \end{aligned}$$

and

$$\begin{aligned}\mathcal{A}(\tilde{p}^*) = \mathrm{E}_g\left[\frac{\partial^2 l}{\partial p^{*2}}\Big|_{\tilde{p}^*}\right] &= \mathrm{E}_g\left[\frac{y(2\tilde{p}^*-1)}{\tilde{p}^{*2}(1-\tilde{p}^*)^2} - \frac{n^*}{(1-\tilde{p}^*)^2}\right] \\ &= \frac{\mathrm{E}_g[y](2\tilde{p}^*-1) - n^*\tilde{p}^{*2}}{\tilde{p}^{*2}(1-\tilde{p}^*)^2} \\ &= \frac{(n_1p_1+n_2p_2)(2\tilde{p}^*-1) - n^*\tilde{p}^{*2}}{\tilde{p}^{*2}(1-\tilde{p}^*)^2} \\ &= \frac{n^*\tilde{p}^*(2\tilde{p}^*-1) - n^*\tilde{p}^{*2}}{\tilde{p}^{*2}(1-\tilde{p}^*)^2} = \frac{n^*\tilde{p}^*(\tilde{p}^*-1)}{\tilde{p}^{*2}(1-\tilde{p}^*)^2} = \frac{-n^*}{\tilde{p}^*(1-\tilde{p}^*)}\,.\end{aligned}$$

Then we have

$$\mathrm{Var}(\hat{p}^*) = \mathcal{A}^{-1}(\tilde{p}^*)\mathcal{B}(\tilde{p}^*)\mathcal{A}^{-1}(\tilde{p}^*) = \frac{n^*\tilde{p}^* - n_1p_1^2 - n_2p_2^2}{n^{*2}}\,.$$

For correct specification of the proxy model or equivalently applied model, i.e. $p_1 = p_2 = p^*$, the value of $\tilde{p}^*$ in (5.36) equals to p^*, which implies that

$$\mathrm{Var}(\hat{p}^*) = \mathcal{A}^{-1}(\tilde{p}^*)\mathcal{B}(\tilde{p}^*)\mathcal{A}^{-1}(\tilde{p}^*) = \frac{n^*p^* - n^*p^{*2}}{n^2} = \frac{p^*(1-p^*)}{n^*}\,.$$

The present behaviour is closely related to the classical maximum likelihood asymptotic normality results as mentioned in Theorem 2, where

$$\mathrm{Var}(\hat{p}^*) = \mathrm{E}\Big[\frac{\partial^2 l}{\partial p^{*2}}\Big] = \frac{p^*(1-p^*)}{n^*} = \mathcal{I}^{-1}(p^*)\,.$$

The following section will discuss the asymptotic distribution theory for a general applied structure, for any i, k, where a clear benefit of a proxy structure in estimating the nearest estimate to the truth is identified.

5.6 Asymptotic Behaviour of Estimates of General Purchase Model

Consider the truth structure (5.2), the proxy structure (5.30) and the applied structure (5.1). Given the log-likelihood (5.15) of the applied probability structure, the score function with respect to θ^* is given by

$$\frac{\partial l}{\partial \theta_j^*} = \sum_{i,t} y_{it}x_{itj}^* - n_{it}^*x_{itj}^*\frac{e^{\sum_j x_{itj}^*\theta_j^*}}{1+e^{\sum_j x_{itj}^*\theta_j^*}} = \sum_{i,t} y_{it}x_{itj}^* - n_{it}^*x_{itj}^*p_{it}^*, \tag{5.37}$$

and the derivative of the score function is given by

$$\begin{aligned}\frac{\partial^2 l}{\partial\theta_j^*\partial\theta_{j'}^*} &= \sum_{i,t} -n_{it}^*x_{itj}^*\frac{\partial l}{\partial\theta_{j'}}\left(\frac{e^{\sum_j x_{itj}^*\theta_j^*}}{1+e^{\sum_j x_{itj}^*\theta_j^*}}\right) \\ &= -\sum_{i,t} n_{it}^*x_{itj}^*\frac{e^{\sum_j x_{itj}^*\theta_j^*}}{(1+e^{\sum_j x_{itj}^*\theta_j^*})^2}x_{itj'}^* = -\sum_{i,t} n_{it}x_{itj}^*p_{it}^*(1-p_{it}^*)x_{itj'}^*\,.\end{aligned} \tag{5.38}$$

Using the proxy structure (5.30), then the closest parameter $\tilde{\theta}^*$ to the truth (5.2) can be computed numerically. Given the large sample distribution theory presented in section 5.2.4, then $\hat{\theta}^*$ converges to $\tilde{\theta}^*$. The asymptotic variance-covariance matrix (5.11) of the estimates at $\tilde{\theta}^*$ is then given by $\mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*)$, such that

$$\left[\mathcal{B}(\tilde{\theta}^*)\right]_{jj'} = \mathrm{Cov}_g\left[\frac{\partial l}{\partial \tilde{\theta}^*_j}, \frac{\partial l}{\partial \tilde{\theta}^*_{j'}}\right].$$

Given above score function (5.37), which is a function of a random variable y_{it}, in which y_{it} for different i and t are independent, then

$$\begin{aligned}\mathrm{Cov}_g\Big[\frac{\partial l}{\partial \tilde{\theta}^*_j}, \frac{\partial l}{\partial \tilde{\theta}^*_{j'}}\Big] &= \mathrm{Cov}_g\Big[\sum_{i,t} y_{it}x^*_{itj}, \sum_{i\prime,t\prime} y_{i't'}x^*_{i't'j'}\Big] \\ &= \sum_{i,t}\sum_{i',t'} \mathrm{Cov}_g[y_{it}x^*_{itj}, y_{i't'}x^*_{i't'j'}] \\ &= \sum_{i,t}\sum_{i',t'} x^*_{itj}\mathrm{Cov}_g[y_{it}, y_{i't'}]x^*_{i't'j'}\,.\end{aligned}$$

given that y_{it} and $y_{i't'}$ are independent, then $\mathrm{Cov}_g[y_{it}, y_{i't'}] = 0$ unless $i = i'$ and $t = t'$. Thus

$$\left[\mathcal{B}(\tilde{\theta}^*)\right]_{jj'} = \sum_{i,t} x_{it}{}^*_j x_{it}{}^*_{j'}\mathrm{Var}_g[y_{it}] = \sum_{i,t}\sum_k x^*_{itj}x^*_{itj'}n_{ikt}p_{ikt}(1-p_{ikt})\,. \tag{5.39}$$

The second derivative of the log-likelihood does not depend on y_{it} and so the matrix $\mathcal{A}(\tilde{\theta}^*)$ is given by

$$\mathcal{A}(\tilde{\theta}^*)_{jj'} = \mathrm{E}_g\left[\frac{\partial^2 l}{\partial\tilde{\theta}^*_j\partial\tilde{\theta}^*_{j'}}\right] = -\sum_{i,t} n^*_{it}x^*_{itj}\frac{e^{\sum_j x^*_{itj}\tilde{\theta}^*_j}}{(1+e^{\sum_j x^*_{itj}\tilde{\theta}^*_j})^2}x^*_{itj'}\,. \tag{5.40}$$

The asymptotic distribution of $\hat{\theta}^*$ is then

$$\hat{\theta}^* \sim \mathrm{N}(\tilde{\theta}^*, \mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*))\,.$$

5.7 Asymptotic Behaviour of an Overall Measure of Advertising Campaigns Effect

By taking into account that the campaign effect can be strata based, we have suggested in the previous chapter a hypothetical differential measure that measures the difference between the expected total number of sales if the new campaign is served in all spatial

units i and the expected total number of sales if it is served in none. Given the truth probability structure (5.2), the total sales given in (4.14) is reduced to

$$\begin{aligned} Y_{it}^{c^1} &= \sum_k Y_{ikt}^{c^1}, \qquad \text{where} \qquad Y_{ikt}^{c^1} \sim \text{Bin}(n_{ikt}, p_{ikt}^{c^1}), \\ Y_{it}^{c^0} &= \sum_k Y_{ikt}^{c^0}, \qquad \text{where} \qquad Y_{ikt}^{c^0} \sim \text{Bin}(n_{ikt}, p_{ikt}^{c^0}) \,. \end{aligned}$$

and so the overall true effect Δ is given by

$$\begin{aligned} \Delta &= \text{E}\Big[\sum_{i,t} Y_{it}^{c^1}\Big] - \text{E}\Big[\sum_{i,t} Y_{it}^{c^0}\Big] = \sum_{i,k,t} n_{ikt} p_{ikt}^{c^1} - \sum_{i,k,t} n_{ikt} p_{ikt}^{c^0} \\ &= \sum_{i,k,t} n_{ikt} \frac{e^{\sum_j x_{iktj}^{c^1} \theta_j}}{1 + e^{\sum_j x_{iktj}^{c^1} \theta_j}} - \sum_{i,k,t} n_{ikt} \frac{e^{\sum_j x_{itj}^{c^0} \theta_j}}{1 + e^{\sum_j x_{iktj}^{*c^0} \theta_j}} \,. \end{aligned}$$

Given the applied model probability structure (5.1), we have the total sales given in (4.16) is reduced to

$$\begin{aligned} Y_{it}^{c^1} &\sim \text{Bin}(n_{it}, p_{it}^{*c^1}), \\ Y_{it}^{c^0} &\sim \text{Bin}(n_{it}, p_{it}^{*c^0}), \end{aligned}$$

and hence the overall applied model effect is given by

$$\begin{aligned} \Delta^* &= \text{E}\Big[\sum_{i,t} Y_{it}^{c^1}\Big] - \text{E}\Big[\sum_{i,t} Y_{it}^{c^0}\Big] = \sum_{i,t} n_{it} p_{it}^{*c^1} - \sum_{it} n_{it} p_{it}^{*c^0} \\ &= \sum_{i,t} n_{it} \frac{e^{\sum_j x_{itj}^{*c^1} \theta_j^*}}{1 + e^{\sum_j x_{itj}^{*c^1} \theta_j^*}} - \sum_{i,t} n_{it} \frac{e^{\sum_j x_{itj}^{*c^0} \theta_j^*}}{1 + e^{\sum_j x_{itj}^{*c^0} \theta_j^*}} \,. \end{aligned}$$

Let $\hat{\Delta}^*$ be the estimate of the overall applied model effect. We call $\hat{\Delta}^*$ the *applied effect*. To find the asymptotic distribution of $\hat{\Delta}^*$, we need to find the nearest overall effect $\tilde{\Delta}^*$ to the truth. Given the proxy model probability structure (5.30)

$$\begin{aligned} Y_{it}^{c^1} &\sim \text{Bin}(n_{it}, \tilde{p}_{it}^{*c^1}), \\ Y_{it}^{c^0} &\sim \text{Bin}(n_{it}, \tilde{p}_{it}^{*c^0}), \end{aligned}$$

where $\tilde{p}_{it}^{*c^0}$ and $\tilde{p}_{it}^{*c^1}$ are functions of the nearest parameter vector $\tilde{\theta}^*$ to the truth, and so the overall proxy model effect, or in short the *proxy effect*, is

$$\begin{aligned} \tilde{\Delta}^* &= \sum_{i,t} n_{it} \tilde{p}_{it}^{*c^1} - \sum_{i,t} n_{it} \tilde{p}_{it}^{*c^0} \\ &= \sum_{i,t} n_{it} \frac{e^{\sum_j x_{itj}^{*c^1} \tilde{\theta}_j^*}}{1 + e^{\sum_j x_{itj}^{*c^1} \tilde{\theta}_j^*}} - \sum_{i,t} n_{it} \frac{e^{\sum_j x_{itj}^{*c^0} \tilde{\theta}_j^*}}{1 + e^{\sum_j x_{itj}^{*c^0} \tilde{\theta}_j^*}} \,. \end{aligned} \tag{5.41}$$

The maximum likelihood estimate $\hat{\theta}^*$ of θ^* has an asymptotic Normal distribution with mean $\tilde{\theta}^*$ and variance $\mathcal{C}(\hat{\theta}^*) = \mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*)$. If the derivative of Δ^* with respect to θ^* exists, then by delta method the asymptotic behaviour of the applied effect $\hat{\Delta}^*$ is

$$\hat{\Delta}^* \sim \mathrm{N}(\tilde{\Delta}^*, \widetilde{\mathrm{Var}}(\hat{\Delta}^*)), \tag{5.42}$$

where

$$\widetilde{\mathrm{Var}}(\hat{\Delta}^*)) = \left(\frac{\partial \Delta^*}{\partial \tilde{\theta}^*}\right)^T \mathcal{C}(\hat{\theta}^*) \left(\frac{\partial \Delta^*}{\partial \tilde{\theta}^*}\right),$$

given that

$$\frac{\partial \Delta^*}{\partial \theta_j^*}\Big|_{\theta_j^*=\tilde{\theta}_j^*} = \sum_{i,t} n_{it} \frac{e^{\sum_j x_{itj}^{*c^1} \tilde{\theta}_j^*}}{(1+e^{\sum_j x_{itj}^{*c^1} \tilde{\theta}_j^*})^2} x_{itj}^{*c^1} - \sum_{i,t} n_{it} \frac{e^{\sum_j x_{itj}^{*c^0} \tilde{\theta}_j^*}}{(1+e^{\sum_j x_{itj}^{*c^0} \tilde{\theta}_j^*})^2} x_{itj}^{*c^0} .$$

5.8 Kullback-Leibler Divergence for a two-stage model

In this section we extend the theoretical asymptotic distribution of the estimate $\hat{\theta}^*$ of the applied model parameter θ^* using a two-stage model, in which observations of purchases y_{it} depend on observations of search s_{it}.

The observations has a truth probability structure (4.11) and the applied model probability structure (4.7). The proxy probability structure is then given by

$$\begin{aligned} S_{it} &= \sum_k S_{ikt}, \quad \text{where} \quad S_{ikt} \sim \mathrm{Bin}(N_{ik}, \varphi_{ikt}^\dagger = \varphi_{it}^*), \\ Y_{it} &= \sum_k Y_{ikt}, \quad \text{where} \quad Y_{ikt} \,|\, S_{ikt} \sim \mathrm{Bin}(s_{ikt}, p_{ikt}^\dagger = p_{it}^*) . \end{aligned} \tag{5.43}$$

The probability $p_{ikt}^\dagger$ of purchasing given searching process with probability $\varphi_{ikt}^\dagger$ is modelled in a similar basis to the applied but with consideration of the truth structure, i.e.

$$\begin{aligned} \mathrm{logit}(\varphi_{ikt}^\dagger = \varphi_{it}^*) = \zeta_{ikt}^\dagger = \nu_i^* + \xi^* t = \sum\nolimits_j x_{iktj}^s h_j^s \vartheta_j^* \iff \zeta^\dagger = X^s H^s \vartheta^*, \\ \mathrm{logit}(p_{ikt}^\dagger = p_{it}^*) = \eta_{ikt}^\dagger = \alpha_i^* + \beta_k^* t + \delta_k^* C_{it} = \sum\nolimits_j x_{iktj} h_j \theta_j^* \iff \eta^\dagger = XH\theta^* . \end{aligned} \tag{5.44}$$

Where X^s is a pile of k matrices $X^{s(k=1)}, X^{s(k=2)}, \ldots, X^{s(k=K)}$ and H^s is a transformation matrix that maps the structure of the parameter vector ϑ^* to the structure of ϑ. Furthermore, $X^{s(k)}H$ is the same for any $k \in K$ given that $X^{s(k)}H^s = X^{s*}$. Equivalently, $\zeta^\dagger$ is a pile of k probability structures $\zeta^{\dagger(k)}$, such that $\zeta^{\dagger(k)} = \zeta^*$ for all k.

Similarly as before X is a pile of k matrices $X^{(k=1)}, X^{(k=2)}, \ldots, X^{(k=K)}$ and H is a transformation matrix that maps the structure of θ^* to the structure of θ. Also $X^{(k)}H$ is the same for any $k \in K$ given that $X^{(k)}H = X^*$ and $\eta^\dagger$ is a pile of k probability structures $\eta^{\dagger(k)}$, such that $\eta^{\dagger(k)} = \eta^*$ for all k.

Let y_{ikt} and s_{ikt} have a truth joint probability function g and proxy joint probability function f , then the KL-divergence can be computed in a similar way to the case when the number of searches is known and so,

$$D_{KL}(g\|f) = \sum\nolimits_{ikt} D_{KL}(g(y_{ikt}, s_{ikt})\|f(y_{ikt}, s_{ikt})),$$

where the KL-divergence for each joint (y_{ikt}, s_{ikt}) in g is

$$D_{KL}(g(y_{ikt}, s_{ikt})\|f(y_{ikt}, s_{ikt})) = \mathrm{E}_g[\log g_{ikt}(y_{ikt}, s_{ikt}) - \log f_{ikt}(y_{ikt}, s_{ikt})],$$

where,

$$\begin{aligned}
\log g(y_{ikt}, s_{ikt}) &= \log\Big[g_{ikt}(y_{ikt} \mid s_{ikt}) g_{ikt}(s_{ikt})\Big] \\
&= \log\Big[\binom{s_{ikt}}{y_{ikt}} p_{ikt}^{y_{ikt}} (1-p_{ikt})^{s_{ikt}-y_{ikt}} \binom{N_{ik}}{s_{ikt}} \varphi_{ikt}^{s_{ikt}} (1-\varphi_{ikt})^{N_{ik}-s_{ikt}}\Big] \\
&= \log\binom{s_{ikt}}{y_{ikt}} + y_{ikt}\log\frac{p_{ikt}}{1-p_{ikt}} + s_{ikt}\log(1-p_{ikt}) \\
&\quad + \log\binom{N_{ik}}{s_{ikt}} + s_{ikt}\log\frac{\varphi_{ikt}}{1-\varphi_{ikt}} + N_{ik}\log(1-\varphi_{ikt}),
\end{aligned}$$

and

$$\begin{aligned}
\log f(y_{ikt}, s_{ikt}) &= \log\Big[f_{ikt}(y_{ikt} \mid s_{ikt}) f_{ikt}(s_{ikt})\Big] \\
&= \log\Big[\binom{s_{ikt}}{y_{ikt}} p_{it}^{*\,y_{ikt}} (1-p_{it}^*)^{s_{ikt}-y_{ikt}} \binom{N_{ik}}{s_{ikt}} \varphi_{it}^{s_{ikt}} (1-\varphi_{it})^{N_{ik}-s_{ikt}}\Big] \\
&= \log\binom{s_{ikt}}{y_{ikt}} + y_{ikt}\log\frac{p_{it}^*}{1-p_{it}^*} + s_{ikt}\log(1-p_{it}^*) \\
&\quad + \log\binom{N_{ik}}{s_{ikt}} + s_{ikt}\log\frac{\varphi_{it}^*}{1-\varphi_{it}^*} + N_{ik}\log(1-\varphi_{it}^*)\,.
\end{aligned}$$

For short we use the term $D_{KL}(g_{ikt}\|f_{ikt})$ to refer to $D_{KL}(g(y_{ikt}, s_{ikt})\|f(y_{ikt}, s_{ikt}))$. By using the law of total expectation, $D_{KL}(g_{ikt}\|f_{ikt})$ is then computed as

$$
\begin{aligned}
D_{KL}(g_{ikt}\|f_{ikt}) &= \mathrm{E}_g[y_{ikt}]\log\frac{p_{ikt}}{1-p_{ikt}} + \mathrm{E}_g[s_{ikt}]\log(1-p_{ikt}) \\
&\quad + \mathrm{E}_g[s_{ikt}]\log\frac{\varphi_{ikt}}{1-\varphi_{ikt}} + N_{ik}\log(1-\varphi_{ikt}) \\
&\quad - \mathrm{E}_g[y_{ikt}]\log\frac{p^*_{it}}{1-p^*_{it}} - \mathrm{E}_g[s_{ikt}]\log(1-p^*_{it}) \\
&\quad - \mathrm{E}_g[s_{ikt}]\log\frac{\varphi_{it}}{1-\varphi_{it}} - N_{ik}\log(1-\varphi_{it}) \\
&= \mathrm{E}_g[\mathrm{E}_g[y_{ikt}\mid s_{ikt}]]\log\frac{p_{ikt}}{1-p_{ikt}} + \mathrm{E}_g[s_{ikt}]\log(1-p_{ikt}) \\
&\quad + N_{ik}\varphi_{ikt}\log\frac{\varphi_{ikt}}{1-\varphi_{ikt}} + N_{ik}\log(1-\varphi_{ikt}) \\
&\quad - \mathrm{E}_g[\mathrm{E}_g[y_{ikt}\mid s_{ikt}]]\log\frac{p^*_{it}}{1-p^*_{it}} - \mathrm{E}_g[s_{ikt}]\log(1-p^*_{it}) \\
&\quad - N_{ik}\varphi_{ikt}\log\frac{\varphi^*_{it}}{1-\varphi^*_{it}} - N_{ik}\log(1-\varphi^*_{it}) \\
&= \mathrm{E}_g[s_{ikt}]p_{ikt}\log\frac{p_{ikt}}{1-p_{ikt}} + \mathrm{E}_g[s_{ikt}]\log(1-p_{ikt}) \\
&\quad - \mathrm{E}_g[s_{ikt}]p_{ikt}\log\frac{p^*_{it}}{1-p^*_{it}} - \mathrm{E}_g[s_{ikt}]log(1-p^*_{it}) \\
&\quad + \mathrm{E}_g[s_{ikt}]\log\frac{\varphi_{ikt}}{1-\varphi_{ikt}} + N_{ik}\log(1-\varphi_{ikt}) \\
&\quad - \mathrm{E}_g[s_{ikt}]\log\frac{\varphi^*_{it}}{1-\varphi^*_{it}} - N_{ik}\log(1-\varphi^*_{it}),
\end{aligned}
$$

and therefore,

$$
\begin{aligned}
D_{KL}(g\|f) = \sum\nolimits_{i,k,t} \Big[&\mathrm{E}_g[s_{ikt}]p_{ikt}\log\frac{p_{ikt}}{1-p_{ikt}} + \mathrm{E}_g[s_{ikt}]\log(1-p_{ikt}) \\
&- \mathrm{E}_g[s_{ikt}]p_{ikt}\log\frac{p^*_{it}}{1-p^*_{it}} - \mathrm{E}_g[s_{ikt}]\log(1-p^*_{it}) \\
&+ N_{ik}\varphi_{ikt}\log\frac{\varphi_{ikt}}{1-\varphi_{ikt}} + N_{ik}\log(1-\varphi_{ikt}) \\
&- N_{ik}\varphi_{ikt}\log\frac{\varphi^*_{it}}{1-\varphi^*_{it}} - N_{ik}\log(1-\varphi^*_{it})\Big],
\end{aligned}
$$

or

$$
\begin{aligned}
D_{KL}(g\|f) = \sum\nolimits_{i,k,t} \Big[&\mathrm{E}_g[s_{ikt}]p_{ikt}\textstyle\sum_j x_{iktj}\theta_j - \mathrm{E}_g[s_{ikt}]\log(1+\exp(\sum_j x_{iktj}\theta_j)) \\
&- \mathrm{E}_g[s_{ikt}]p_{ikt}\textstyle\sum_j x_{iktj}h_j\theta^*_j + \mathrm{E}_g[s_{ikt}]log(1+\exp(\sum_j x_{iktj}h_j\theta^*_j)) \\
&+ N_{ik}\varphi_{ikt}\textstyle\sum_j x^s_{iktj}\vartheta_j - N_{ik}\log(1+\exp(\sum_j x^s_{iktj}\vartheta_j)) \\
&- N_{ik}\varphi_{ikt}p_{ikt}\textstyle\sum_j x^s_{iktj}h^s_j\vartheta^*_j + N_{ik}\log(1+\exp(\sum_j x^s_{iktj}h^s_j\vartheta^*_j))\Big].
\end{aligned}
\tag{5.45}
$$

For a two-stage process, the KL-divergence is a summation of the KL-divergence of the purchase process and the KL-divergence of search process. Since we are interested in

measuring the effectiveness of the campaign design, which is δ_k in parameter vector θ^*, we can then consider search parameters ϑ^* as nuisance parameters although they might affect the proxy estimates of θ^*. Therefore we focus on the first part of equation (5.45), i.e. the KL-divergence of purchase process. The $D_{KL}(g\|f)$ is then

$$D_{KL}(g\|f) \equiv \sum_{i,k,t} \Big[\mathrm{E}_g[s_{ikt}] p_{ikt} \sum_j x_{iktj}\theta_j - \mathrm{E}_g[s_{ikt}] \log(1+\exp(\sum_j x_{iktj}\theta_j)) \\ - \mathrm{E}_g[s_{ikt}] p_{ikt} \sum_j x_{iktj} h_j \theta_j^* + \mathrm{E}_g[s_{ikt}] log(1+\exp(\sum_j x_{iktj} h_j \theta_j^*)) \Big] . \tag{5.46}$$

Comparing $D_{KL}(g\|f)$ (5.46) and $D_{KL}(g\|f)$ (5.34), it can be seen that the current computation of the KL-divergence is in agreement with prior computation when the search process was postulated that is known, except that the known search n_{ikt} in $D_{KL}(g\|f)$ (5.35) turns into the expected value of search which is $\mathrm{E}_g[s_{ikt}] = N_{ikt}\varphi_{ikt}$ in the current result.

To minimise the KL-divergence, take the derivative of $D_{KL}(g\|f)$ with respect to θ^*, such that

$$\frac{\partial D_{KL}}{\partial \theta_j^*} = \sum_{ikt} -\mathrm{E}_g[s_{ikt}] p_{ikt} x_{iktj} h_j + \mathrm{E}_g[s_{ikt}] \big(1+\exp(\sum_j x_{iktj} h_j \theta_j^*)\big)^{-1} x_{iktj} h_j \, .$$

A numerical optimisation procedure is required to find the best parameters vector $\tilde{\theta}^*$ that minimises the KL-divergence.

5.9 Asymptotic Behaviour of Estimates of General Purchase Model Conditioning on Search

Consider the two-stage applied model (4.9), the probability applied structure of purchasing y_{it} is given by

$$f(y_{it} \mid s_{it}) = \binom{s_{it}}{y_{it}} (p_{it}^*)^{y_{it}} (1-p_{it}^*)^{s_{it}-y_{it}},$$

given that

$$f(s_{it}) = \binom{N_i}{s_{it}} (\varphi_{it}^*)^{y_{it}} (1-\varphi_{it}^*)^{N_i - s_{it}} \, .$$

Let y be a vector of independent realisations of purchases y_{it} and s be a vector of independent realisations of searches s_{it}. Assume y and s are sampled from Binomial distributions

that are stated in model (4.7). The joint density function of a vector of independent pair of observations of (y_{it}, s_{it}) is $f(y_{it}, s_{it})$ and the contribution of (y, s) to the likelihood is then

$$\begin{aligned} L &= \prod_{i,t} f(y_{it}, s_{it}) = \prod_{i,t} f(y_{it} \mid s_{it}) f(s_{it}) \\ &= \prod_{i,t} \binom{s_{it}}{y_{it}} (\mathrm{p}^*_{it})^{y_{it}} (1 - \mathrm{p}^*_{it})^{s_{it} - y_{it}} \binom{N_i}{s_{it}} (\varphi^*_{it})^{s_{it}} (1 - \varphi^*_{it})^{N_i - s_{it}} \\ &\equiv \prod_{i,t} \left(\frac{\mathrm{p}^*_{it}}{1 - \mathrm{p}^*_{it}} \right)^{y_{it}} (1 - \mathrm{p}^*_{it})^{s_{it}} \left(\frac{\varphi^*_{it}}{1 - \varphi^*_{it}} \right)^{s_{it}} (1 - \varphi^*_{it})^{N_i} \\ &= \prod_{i,t} \left(e^{\eta^*_{it}} \right)^{y_{it}} \left(1 + e^{\eta^*_{it}} \right)^{-s_{it}} \left(e^{\zeta^*_{it}} \right)^{s_{it}} \left(1 + e^{\zeta^*_{it}} \right)^{-N_i}. \end{aligned}$$

On the basis of the matrix formulation of the model, the likelihood is given by

$$L = \prod_{i,t} \left(e^{\sum_j x^*_{itj} \theta^*_j} \right)^{y_{it}} \left(1 + e^{\sum_j x^*_{itj} \theta^*_j} \right)^{-s_{it}} \left(e^{\sum_j x^{s*}_{it\,j} \vartheta^*_j} \right)^{s_{it}} \left(1 + e^{\sum_j x^{s*}_{it\,j} \vartheta^*_j} \right)^{-N_i},$$

taking the natural log yields the log-likelihood function:

$$l = \sum\nolimits_{i,t} y_{it} \Big(\sum\nolimits_j x^*_{itj} \theta^*_j \Big) - s_{it} \log \left(1 + e^{\sum_j x^*_{itj} \theta^*_j} \right) + s_{it} \Big(\sum\nolimits_j x^{s*}_{it\,j} \vartheta^*_j \Big) - N_i \log \left(1 + e^{\sum_j x^{s*}_{it\,j} \vartheta^*_j} \right). \tag{5.47}$$

The log-likelihood of two-stage applied model expresses a sum of two log-likelihoods: purchase and search. We continue derivation of asymptotic theory based on log-likelihood (5.47) for applied model. The score function of l with respect to θ^*_j in a parameter vector θ^* is

$$\frac{\partial l}{\partial \theta^*_j} = \sum_{i,t} y_{it} x^*_{itj} - s_{it} x^*_{itj} \frac{e^{\sum_j x^*_{it} \theta^*_j}}{1 + e^{\sum_j x^*_{itj} \theta^*_j}} = \sum_{i,t} y_{it} x^*_{itj} - s_{it} x^*_{itj} p^*_{it}, \tag{5.48}$$

and with respect to ϑ^*_j in a parameter vector ϑ^*

$$\frac{\partial l}{\partial \vartheta^*_j} = \sum_{i,t} s_{it} x^{s*}_{itj} - N_i x^{s*}_{itj} \frac{e^{\sum_j x^{s*}_{itj} \vartheta^*_j}}{1 + e^{\sum_j x^{s*}_{itj} \vartheta^*_j}} = \sum_{i,t} s_{it} x^{s*}_{itj} - N_i x^{s*}_{itj} \varphi^*_{it}.$$

Then the second partial derivative matrix with respect to θ^*_j is

$$\begin{aligned} \frac{\partial^2 l}{\partial \theta^*_j \partial \theta^*_{j\prime}} &= \sum_{it} -s_{it} x^*_{itj} \frac{\partial l}{\partial \theta_{j\prime}} \left(\frac{e^{\sum_j x^*_{itj} \theta^*_j}}{1 + e^{\sum_j x^*_{itj} \theta^*_j}} \right) \\ &= -\sum_{it} s_{it} x^*_{itj} \frac{e^{\sum_j x^*_{itj} \theta^*_j}}{(1 + e^{\sum_j x^*_{itj} \theta^*_j})^2} x^*_{itj\prime} = -\sum_{it} s_{it} x^*_{itj}\, \mathrm{p}^*_{it} (1 - \mathrm{p}^*_{it}) x^*_{itj\prime}, \end{aligned} \tag{5.49}$$

and with respect to ϑ_j^* is

$$\begin{aligned}\frac{\partial^2 l}{\partial\vartheta_j^*\partial\vartheta_{j\prime}^*} &= \sum_{it} -s_{it}x_{itj}^{s*}\frac{\partial l}{\partial\vartheta_{j\prime}}\left(\frac{e^{\sum_j x_{itj}^{s*}\vartheta_j^*}}{1+e^{\sum_j x_{itj}^{s*}\vartheta_j^*}}\right)\\ &= -\sum_{it} s_{it}x_{itj}^{s*}\frac{e^{\sum_j x_{itj}^{s*}\vartheta_j^*}}{(1+e^{\sum_j x_{itj}^{s*}\vartheta_j^*})^2}x_{itj\prime}^*,\end{aligned}$$

and the two mixed partial derivatives are

$$\frac{\partial l}{\partial\theta_j^*\partial\vartheta_{j\prime}^*} = \frac{\partial l}{\partial\vartheta_j^*\partial\theta_{j\prime}^*} = 0, \quad \text{for all } (j, j\prime)\,.$$

Let $\hat{\theta}^*$ and $\hat{\vartheta}^*$ be the maximum likelihood estimators, and, $\tilde{\theta}^*$ and $\tilde{\vartheta}^*$ are the parameters vectors which minimises KL-divergence. By taking into consideration asymptotic theory, the estimates $\hat{\theta}^*$ and $\hat{\vartheta}^*$ converge to $\tilde{\theta}^*$ and $\tilde{\vartheta}^*$, respectively. The asymptotic variance-covariance of the estimates at $\tilde{\theta}^*$ and $\tilde{\vartheta}^*$ is formed by matrix $\mathcal{A}$ and matrix $\mathcal{B}$, where

$$\mathcal{A} = \left(\begin{array}{c|c}\mathcal{A}(\tilde{\theta}^*) & \mathcal{A}(\tilde{\theta}^*,\tilde{\vartheta}^*)\\ \hline \mathcal{A}(\tilde{\vartheta}^*,\tilde{\theta}^*) & \mathcal{A}(\tilde{\vartheta}^*)\end{array}\right),$$

and

$$\mathcal{B} = \left(\begin{array}{c|c}\mathcal{B}(\tilde{\theta}^*) & \mathcal{B}(\tilde{\theta}^*,\tilde{\vartheta}^*)\\ \hline \mathcal{B}^T(\tilde{\theta}^*,\tilde{\vartheta}^*) & \mathcal{B}(\tilde{\vartheta}^*)\end{array}\right),$$

such that

$$\begin{aligned}\mathcal{A}(\tilde{\theta}^*) &= \mathrm{E}_g\Big[\frac{\partial^2 l}{\partial\tilde{\theta}_j^*\partial\tilde{\theta}_{j\prime}^*}\Big], & \mathcal{A}(\tilde{\theta}^*,\tilde{\vartheta}^*) &= \mathrm{E}_g\Big[\frac{\partial l}{\partial\tilde{\theta}_j^*\partial\tilde{\vartheta}_j^*}\Big] = 0,\\ \mathcal{A}(\tilde{\vartheta}^*) &= \mathrm{E}_g\Big[\frac{\partial^2 l}{\partial\tilde{\vartheta}_j^*\partial\tilde{\vartheta}_{j\prime}^*}\Big], & \mathcal{A}(\tilde{\vartheta}^*,\tilde{\theta}^*) &= \mathrm{E}_g\Big[\frac{\partial l}{\partial\tilde{\vartheta}_j^*\partial\tilde{\theta}_j^*}\Big] = 0,\end{aligned} \tag{5.50}$$

so that $\mathcal{A}$ is block diagonal. Also

$$\mathcal{B}(\tilde{\theta}^*) = \mathrm{Cov}_g\Big[\frac{\partial l}{\partial\tilde{\theta}_j^*}\Big], \quad \mathcal{B}(\tilde{\vartheta}^*) = \mathrm{Cov}_g\Big[\frac{\partial l}{\partial\tilde{\vartheta}_j^*}\Big] \quad \text{and} \quad \mathcal{B}(\tilde{\theta}^*,\tilde{\vartheta}^*) = \mathrm{Cov}_g\Big[\frac{\partial l}{\partial\tilde{\theta}_j^*}, \frac{\partial l}{\partial\tilde{\vartheta}_{j\prime}^*}\Big]\,. \tag{5.51}$$

The variance-covariance matrix is given by matrix $\mathcal{C}$ such that

$$\mathcal{C} = \left(\begin{array}{c|c}\mathcal{C}(\hat{\theta}^*) = \mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*) & \mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*,\tilde{\vartheta}^*)\mathcal{A}^{-1}(\tilde{\vartheta}^*)\\ \hline (\mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*,\tilde{\vartheta}^*)\mathcal{A}^{-1}(\tilde{\vartheta}^*))^T & \mathcal{C}(\tilde{\vartheta}^*) = \mathcal{A}^{-1}(\tilde{\vartheta}^*)\mathcal{B}(\tilde{\vartheta}^*)\mathcal{A}^{-1}(\tilde{\vartheta}^*)\end{array}\right). \tag{5.52}$$

Because θ^* forms the central focus of this research study, the asymptotic variance-covariance matrix of the estimate $\hat{\theta}^*$ at the nearest parameters to the truth $\tilde{\theta}^*$ is then given by the

first block of matrix $\mathcal{C}(\hat{\theta}^*)$. The computations of $\mathcal{B}(\tilde{\theta}^*)$ and $\mathcal{A}(\tilde{\theta}^*)$ depend on partial derivatives equations (5.48) and (5.49), respectively. For $\mathcal{B}(\tilde{\theta}^*)$, the expectations are found by the law of total variance and the law of total expectation such that

$$\begin{aligned}
\Big[\mathcal{B}(\tilde{\theta}^*)\Big]_{jj'} = \mathrm{cov}_g\left[\frac{\partial l}{\partial \tilde{\theta}^*_j}\frac{\partial l}{\partial \tilde{\theta}^*_{j'}}\right] &= \mathrm{cov}_g\Big[\sum_{i,t} y_{it}x^*_{itj}, \sum_{i',t'} y_{i't'}x^*_{i't'j'}\Big] \\
&= \sum_{i,t}\sum_{i',t'} \mathrm{cov}_g[x^*_{itj}y_{it}, x^*_{i't'j'}y_{i't'}] \\
&= \sum_{i,t}\sum_{i',t'} x^*_{itj}x^*_{i't'j'}\mathrm{cov}_g[y_{it}, y_{i't'}] \\
&= \sum_{i,t} x^*_{itj}x^*_{itj'}\mathrm{Var}_g[y_{it}] \\
&= \sum_{i,t} x^*_{itj}x^*_{itj'}(\mathrm{E}_g[\mathrm{Var}_g[y_{it}\mid s_{it}]] + \mathrm{Var}_g[\mathrm{E}_g[y_{it}\mid s_{it}]]) \\
&= \sum_{i,t} x^*_{itj}x^*_{itj'}\Big(\sum_k p_{ikt}(1-p_{ikt})\mathrm{E}_g[s_{ikt}] + \sum_k p^2_{ikt}\mathrm{Var}_g[s_{ikt}]\Big),
\end{aligned} \tag{5.53}$$

given that $\mathrm{cov}_g[y_{it}, y_{i't'}] = 0$ unless $i = i'$ and $t = t'$. This expectations accords with previous computation of expectation in (5.39), when searches is known, but the present expectation expressed is described by $\mathrm{E}_g[s_{ikt}]$ and $\mathrm{Var}_g[s_{ikt}]$.

The expectation of $\mathcal{A}(\tilde{\theta}^*)$ (5.50) depends on the expectation of the search random variable, such that

$$\begin{aligned}
\mathcal{A}(\tilde{\theta}^*) = \mathrm{E}_g\left[\frac{\partial^2 l}{\partial\tilde{\theta}^*_j\partial\tilde{\theta}^*_{j\prime}}\right] &= -\sum_{i,t}\mathrm{E}_g[s_{it}]x^*_{itj}\frac{e^{\sum_j x^*_{itj}\tilde{\theta}^*_j}}{(1+e^{\sum_j x^*_{itj}\tilde{\theta}^*_j})^2}x^*_{itj\prime} \\
&= -\sum_{i,t}\sum_k \mathrm{E}_g[s_{ikt}]x^*_{itj}\frac{e^{\sum_j x^*_{itj}\tilde{\theta}^*_j}}{(1+e^{\sum_j x^*_{itj}\tilde{\theta}^*_j})^2}x^*_{itj\prime}\,.
\end{aligned} \tag{5.54}$$

This finding also agrees with the previous result in (5.40), but with the present $\mathcal{A}(\tilde{\theta}^*)$ expressed in term of $\mathrm{E}_g[s_{ikt}]$. Therefore, when modelling purchase process conditional on search process, the asymptotic variance of $\hat{\theta}^*$ then actually just needs to replace the known number of search n_{ikt} or $n^*_{it} = \sum_k n_{ikt}$ in the asymptotic variance formula in known searches case by expressions involving the expectation and variance of s_{ikt}. Provided $\mathcal{B}(\tilde{\theta}^*)$ (5.53) and $\mathcal{A}(\tilde{\theta}^*)$ (5.54), the asymptotic distribution of $\hat{\theta}^*$ is

$$\hat{\theta}^* \sim \mathrm{N}(\tilde{\theta}^*, \mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*))\,.$$

5.10 Asymptotic Behaviour of an Overall Measure of Advertising Campaigns Effect

The measure of overall effect of advertising campaigns has been discussed the previous chapter for a two-stage model. The true distribution of total sales is given by the probability distribution (4.14) and the overall true effect Δ is given by equation (4.15). The applied distribution of total sales is given by they probability distribution (4.16) and the overall true effect Δ^* is given by equation (4.17). Given the proxy model probability structure (5.43), the proxy probability structure of the total sales is then given by

$$\begin{aligned} Y_{it}^{c^1} &= \sum_k Y_{ikt}^{c^1}, \quad \text{where} \quad Y_{ikt}^{c^1} \,|\, S_{ikt}^{c^1} \sim \text{Bin}(s_{ikt}^{c^1}, \tilde{p}_{it}^{*c^1}) \quad \text{and} \quad S_{ikt}^{c^1} \sim \text{Bin}(N_{ik}, \tilde{\varphi}_{it}^{*c^1}) \\ Y_{it}^{c^0} &= \sum_k Y_{ikt}^{c^0}, \quad \text{where} \quad Y_{ikt}^{c^0} \,|\, S_{ikt}^{c^0} \sim \text{Bin}(s_{ikt}^{c^0}, \tilde{p}_{it}^{*c^0}) \quad \text{and} \quad S_{ikt}^{c^0} \sim \text{Bin}(N_{ik}, \tilde{\varphi}_{it}^{*c^0}) \,. \end{aligned} \tag{5.55}$$

The proxy effect is then

$$\begin{aligned} \tilde{\Delta}^* &= \text{E}\Big[\sum_{i,t} Y_{it}^{c^1}\Big] - \text{E}\Big[\sum_{i,t} Y_{it}^{c^0}\Big] = \sum_{it} \text{E}[s_{ikt}^{c^1}]\tilde{p}_{it}^{*c^1} - \sum_{it} \text{E}[s_{ikt}^{c^0}]\tilde{p}_{it}^{*c^0} \\ &= \sum_{i,t} \text{E}[s_{ikt}^{c^0}] \frac{e^{\sum_j x_{itj}^{*c^1}\tilde{\theta}_j^*}}{1 + e^{\sum_j x_{itj}^{*c^1}\tilde{\theta}_j^*}} - \sum_{i,t} \text{E}[s_{ikt}^{c^0}] \frac{e^{\sum_j x_{itj}^{*c^0}\tilde{\theta}_j^*}}{1 + e^{\sum_j x_{itj}^{*c^0}\tilde{\theta}_j^*}} \,. \end{aligned} \tag{5.56}$$

The proxy effect agrees with the previous result in (5.41), but the present $\tilde{\Delta}^*$ expressed in term of $\text{E}[s_{ikt}]$. The expectation of s_{ikt} is a function of $\tilde{\vartheta}^*$. The delta method may be used as before with the matrix $\mathcal{C}$ from (5.52) to find $\widetilde{\text{Var}}(\hat{\Delta}^*)$). However, this is not pursued further in this thesis as the focus is on consequences, for estimation of θ^*, of choice of design strategy.

5.11 Summary and Concluding Remarks

In this chapter we have developed a theoretical framework to study the implications of unobserved covariates for inferences about estimated effects of advertising campaigns that are based on geo-experiments. A toy applied model has been used to understand the theoretical behaviour of estimated effects when using a misspecified applied model. An important part of the framework is a proxy model linking applied model and the truth. The proxy model makes possible the application of standard results in the literature on the maximum likelihood estimation for misspecified models.

In reality we work on an applied probability model and usually whatever we apply or specify will be wrong but as George Box said "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful" (Box et al. 1978). In this study, we assumed truth is known, and so the presented amount of mismatch between the applied model and the truth can be measured by KL-divergence. However, the actual probability distribution of the truth is cumbersome to find, which makes the KL-divergence difficult to compute. Thus, we have proposed a proxy model, which is a proxy expression of the applied model including the truth structure to proceed the computation of the KL-divergence. We have shown that maximising the likelihood is equivalent to minimising KL-divergence. The asymptotic distribution theory of the estimates has been discussed under misspecified applied model. We have confirmed through a toy model that if the applied model is correctly specified, then the asymptotic variance of an estimate at the nearest parameter $\tilde{\theta}^*$ to the truth, is in agreement with the standard maximum likelihood inferences.

We have studied the asymptotic distribution theory of estimates under misspecification for two structures of purchase model: a one-stage model and a two-stage model. In a one-stage model, we assumed the number of search is known in each stratum in each spatial unit for each time period. On the other hand in a two-stage model, the purchase process is given by search process, i.e. the number of search is random. Given that our investigation of a asymptotic distribution theory of estimates are based on parameters θ^* in the purchase model, we have found there are similarities between the behaviours of estimates of parameters in a two-stage model and those in a one-stage model. In both model structures, the asymptotic variance of estimates $\hat{\theta}^*$ were expressed by search term, but in a two-stage model is expressed by the expected number and variance of searches instead of known fixed number of searches. According to the literature on the maximum likelihood estimation under misspecification, the asymptotic distribution theory is derived at $\tilde{\theta}^*$, parameters vector which minimises the KL-divergence. In line with θ^*, when a two-stage model is considered, the expression of KL-divergence also agreed with the one expressed by parameters of a one-stage model, except that by using a two-stage model, the KL-divergence is expressed in term of expectation and variances of searches rather than known searches as the case in a one-stage model.

The theoretical computation of KL-divergence and asymptotic variance in this chapter

are subject to parameter vector θ^* in the purchase applied model, which includes the campaign effect δ. In a two-stage process, however, there is a parameter vector ϑ^* in the search applied model and θ^* in the purchase applied model. Both parameters are a piece of the theoretical computations of the KL-divergence and the asymptotic distribution theory. The presented theoretical computation in this chapter, however, has considered ϑ^* a nuisance parameter vector. Thus, the computation was limited by θ^*. The parameter vector ϑ^* may lead to a different proxy estimate of the campaign effect but the effect of ϑ^* on estimating θ^* has not been examined further.

Given that truth model is strata based, the overall effect of advertising campaign has been quantified in term of sales using a differential measure Δ^*. Given the asymptotic Normal distribution of $\hat{\theta}^*$, the asymptotic distribution of $\hat{\Delta}^*$ has been found using delta method.

In the next chapter we will describe computational algorithms to investigate how efficient is the theoretical computation provided in this chapter. The algorithms aim to estimate the campaign effect for random generated data of purchases for a specified truth instance and a particular campaign design strategy.

Chapter 6

Demonstration of the Theoretical Asymptotic Distribution of Estimated Applied Model Parameters

The theoretical framework presented in the previous chapter provides the asymptotic distribution of the estimated applied model parameters. In this chapter, we test the applicability of the theory by assessing its performance in a variety of contexts in comparison to Monte Carlo simulations.

In the following discussions, for simplicity we shall describe the computational algorithms used to validate the theoretical asymptotic distribution of the estimates of the purchase applied model parameters assuming that search process outcomes are known. The algorithms are readily extended when used for a two-stage process: search and purchase.

The chapter starts with a specification of the theoretical basis including mainly specification of truth parameters and specification of campaign design. The following section discusses how to investigate the validity of the theoretical asymptotic distribution. The next section gives a description of the computational algorithms that are developed to find theoretical and asymptotic distributions of $\hat{\theta}^*$ and $\hat{\Delta}^*$. Computational results are

then presented for specified campaign designs using different instances of truth parameters and different number of searches. Some remarks and conclusions are drawn in the final section.

6.1 Basic Components of Finding the Asymptotic Distribution of Estimates

Given a vector estimate $\hat{\theta}^*$ of a parameter vector θ^* in the applied model, the theoretical distribution of $\hat{\theta}^*$ is then asymptotically Normal with mean $\tilde{\theta^*}$ and variance $\mathcal{C}(\hat{\theta^*}) = \mathcal{A}^{-1}(\tilde{\theta^*})\mathcal{B}(\tilde{\theta^*})\mathcal{A}^{-1}(\tilde{\theta^*})$, where the parameter vector $\tilde{\theta^*}$ is the best parameter vector to use for θ^* in the applied model to provide a closest model to the truth.

It is also important to recall that the theoretical asymptotic distribution of the applied overall effect estimate $\hat{\Delta}^*$ is Normal with mean $\tilde{\Delta}^*$ and variance $\widetilde{\text{Var}}(\hat{\Delta}^*) = \left(\frac{\partial \hat{\Delta}^*}{\partial \tilde{\theta}}\right)^T \mathcal{C}(\hat{\theta^*}) \left(\frac{\partial \hat{\Delta}^*}{\partial \tilde{\theta}}\right)$.

Finding the theoretical asymptotic distribution of $\hat{\theta}^*$ and $\hat{\Delta}^*$ is based upon finding the parameter vector $\tilde{\theta^*}$, which itself relies on two main components: a specified truth parameter vector θ_0 and a specified campaign design C_{it}. The specification of both components have been discussed at the end of chapter 4. However the specification of the truth parameters there was incomplete; in as far as only the parameters that depend on the data were discussed. There are others that need to be specified in order to create an instance of truth. In what follows, we propose a way of generating interesting instances of truth parameters.

6.1.1 Parameters for Truth

Consider a truth parameter vector $\theta_0 = \begin{bmatrix}\alpha_0 & \gamma_0 & \beta_0 & \delta_0\end{bmatrix}$, where

$$\alpha_0 = \begin{bmatrix}\alpha_1 & \dots & \alpha_{204}\end{bmatrix}$$
$$\gamma_0 = \begin{bmatrix}\gamma_1 & \gamma_2 & \gamma_3 & \gamma_4\end{bmatrix}, \qquad \beta_0 = \begin{bmatrix}\beta_1 & \beta_2 & \beta_3 & \beta_4\end{bmatrix}, \qquad \delta_0 = \begin{bmatrix}\delta_1 & \delta_2 & \delta_3 & \delta_4\end{bmatrix},$$

given that population-strata are based upon the social grades covariate, i.e. $K = 4$. Assume that the values of the vector α_0 are known and equal to the estimates of spatial effects parameters that were obtained from fitting realistic purchase data. To specify the

other parameters, we propose the following special parameter cases

$$\begin{aligned}
\delta-\text{case:}\quad & \delta_0=\begin{bmatrix}w_1c & w_2c & w_3c & w_4c\end{bmatrix},\quad \gamma_0=\beta_0=\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix},\\
\beta-\text{case:}\quad & \beta_0=\begin{bmatrix}w_1c & w_2c & w_3c & w_4c\end{bmatrix},\quad \gamma_0=\delta_0=\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix},\\
\gamma-\text{case:}\quad & \gamma_0=\begin{bmatrix}w_1c & w_2c & w_3c & w_4c\end{bmatrix},\quad \beta_0=\delta_0=\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix}.
\end{aligned}\tag{6.1}$$

Assuming that c and w_k for $k\in\{1,2,3,4\}$ are real non-zero numbers. In this truth proposal, the non-zero parameter vector in each case of the proposed truth is limited by the use of $\sum_k w_k c=0$. Specifically, we consider certain values of c which are $c=\pm 0.2$, $c=\pm 1$ and $c=\pm 5$ and we assume $w_1=-1.5$, $w_2=-0.5$, $w_3=0.5$ and $w_4=1.5$. In this way we have three differences between strata depending on c: a small difference using $c=\pm 0.2$, a moderate difference using $c=\pm 1$ and a large difference using $c=\pm 5$. Additionally, the overall difference in each truth case for each c is centred to zero on average. We are attempting to investigate the impact of different scales of change in the probability of purchasing between strata on estimating parameters of the applied model, provided that the overall change is neutral.

For each truth case, the difference between strata gets bigger or smaller depending on the sign of c but without changing the overall difference between strata. For example, for $\beta-$case and $\delta-$case, the difference between strata does not change the overall difference between the two time points and between campaigns' effects, respectively. In addition, given that the probability of purchasing are estimated using a logit-linear regression model, then a small difference between strata produces a small change between their logit values and the converse is true.

Combinations of combined truth cases will be considered as well. For example a combination of two cases such as $\beta-$case and $\delta-$case is given by

$$\begin{aligned}
\beta\delta-\text{case:}\quad & \beta_0=\begin{bmatrix}w_1c_1 & w_2c_1 & w_3c_1 & w_4c_1\end{bmatrix},\\
& \delta_0=\begin{bmatrix}w_1c_2 & w_2c_2 & w_3c_2 & w_4c_2\end{bmatrix},\quad \gamma_0=\begin{bmatrix}0 & 0 & 0 & 0\end{bmatrix},
\end{aligned}\tag{6.2}$$

or a combination of the three cases is given by

$$\begin{aligned}
\gamma\beta\delta-\text{case:}\quad & \gamma_0=\begin{bmatrix}w_1c_1 & w_2c_1 & w_3c_1 & w_4c_1\end{bmatrix},\\
& \beta_0=\begin{bmatrix}w_1c_2 & w_2c_2 & w_3c_2 & w_4c_2\end{bmatrix},\\
& \delta_0=\begin{bmatrix}w_1c_3 & w_2c_3 & w_3c_3 & w_4c_3\end{bmatrix},
\end{aligned}\tag{6.3}$$

where c_1, c_2 and c_3 are real non-zero numbers. In this study, however, the values of c_1, c_2 and c_3 are limited to one or more of the proposed c values mentioned above.

The defined truth cases with their level of differences between strata assist in demonstrating which design strategies work better. These may also demonstrate which theoretical framework may provide adverse inferences.

6.1.2 Campaign Designs

In chapter 4, we specified a number of randomisation based design strategies for choosing campaign designs. Each design strategy provides a large set of possible campaign designs. Additionally, the sets generated from different design strategies may overlap. This study shall not attempt to study all possible campaign designs to validate the theoretical asymptotic distribution of estimated applied model parameters.

In this chapter, we choose one design strategy and then consider a specific number of resulting campaign designs, because different campaign designs influence the parameters of the theoretical distribution even if the input truth parameters are unchanged. The question that then arises is, how to choose a specific number of campaign designs from a design strategy? There are different ways we might answer this question. One possible way is to look into various theoretical asymptotic distributions of $\hat{\Delta}^*$ of R randomly generated campaign designs from the chosen design strategy using a specific truth parameter θ_0. This can include checking out the behaviour of the obtained theoretical asymptotic distributions $\hat{\Delta}^*$ relative to the true effect Δ_0, given that the true effect is design independent.

We choose to use the design strategy of the matched-pairs in terms of social grades covariate. 1000 campaign designs are randomly generated from this strategy within each pair, one allocated randomly to treatment and the other to control. The theoretical asymptotic parameters $\tilde{\Delta}^*$ and $\tilde{\text{sd}}(\hat{\Delta}^*)$ are found for each generated campaign design using a truth parameter vector θ_0 that is based on $\delta-$case with extreme difference between strata using $c = 5$, i.e. the truth parameters are

$$\alpha_i = \alpha_i, \quad \gamma_k = \beta_k = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \quad \delta_k = \begin{bmatrix} -7.5 & -2.5 & 2.5 & 7.5 \end{bmatrix}.$$

The computational algorithm that is used to find the theoretical asymptotic distribution of $\hat{\Delta}^*$ is discussed in the next section. The distributions of the 1000 values of mean $\tilde{\Delta}^*$ and 1000 values of standard deviations $\tilde{\text{sd}}(\hat{\Delta}^*)$ are depicted in **Figure 6.1**.

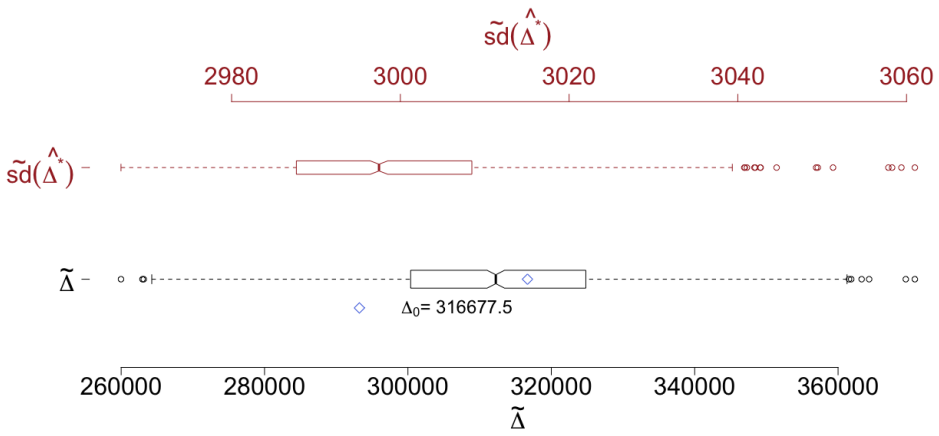


Figure 6.1: Distributions of theoretical mean $\tilde{\Delta}^*$ and $\tilde{sd}(\hat{\Delta}^*)$ of 1000 campaign designs based upon matched-pairs design strategy in terms of social grades covariate using truth parameters of $\delta-$case and $c = 5$

The points of the variability distribution $\tilde{\text{sd}}(\hat{\Delta}^*)$ can be considered close to each other, indicating that the change in the variability of $\hat{\Delta}^*$ between the campaign designs is small. Considering the distribution of $\tilde{\Delta}^*$, the median is 312258 which is about 1% below the true effect $\Delta_0 = 316677.5$. The location of $\tilde{\Delta}^*$ relative to Δ_0 varies between the campaign designs, where $\tilde{\Delta}^*$ is below, above or nearly at Δ_0. Thus it would be interesting to look into the theoretical asymptotic distributions of $\hat{\Delta}^*$ across a sample of campaign designs that provides good coverage of the overall range of $\tilde{\Delta}^*$. For simplicity we limit the investigation of the theoretical asymptotic distribution to 10 distributions; or equivalently 10 selected campaign designs that shown in **Figure 6.2** and **Figure 6.3**.

Given that the theoretical distribution of $\hat{\Delta}^*$ is approximately Normal with parameters $\tilde{\Delta}^*$ and $\tilde{\text{sd}}(\hat{\Delta}^*)$, then each campaign design produces a distribution in which about 99.3% of $\hat{\Delta}^*$ values that are not outliers lie within about 2.698 standard deviations $\tilde{\text{sd}}(\hat{\Delta}^*)$ of the mean $\tilde{\Delta}^*$, i.e. $\tilde{\Delta}^* \pm 2.698 \times \tilde{\text{sd}}(\hat{\Delta}^*)$ and about 50% of $\hat{\Delta}^*$ values lie within $\tilde{\Delta}^* \pm 0.6745 \times \tilde{\text{sd}}(\hat{\Delta}^*)$, (Lucas et al. 2014). Additionally, the median of the distribution is very similar to the mean $\tilde{\Delta}^*$. Those summaries of the distribution of $\hat{\Delta}^*$ are presented corresponding to the 10 campaign designs in **Figure 6.3**. The summaries of each distribution are presented in the figure using a horizontal line.

The corresponding designs to those presented distributions need to be extracted to be employed alongside different specification of truth parameter θ_0 to investigate the difference between a set of theoretical distributions. The distribution of treatment spatial units in

the 10 selected designs are presented in Appendix D in **Figures D.1, D.2** and **D.3**.

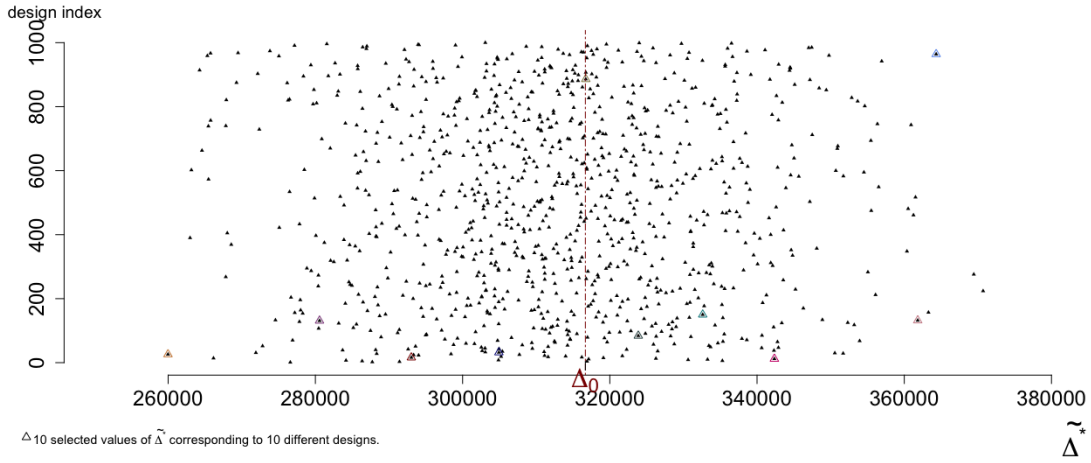


Figure 6.2: Distributions of the theoretical means $\tilde{\Delta}^*$ of 1000 campaign designs based upon matched-pairs design strategy in term of social grades covariate using truth parameters of $\delta-$case and $c = 5$

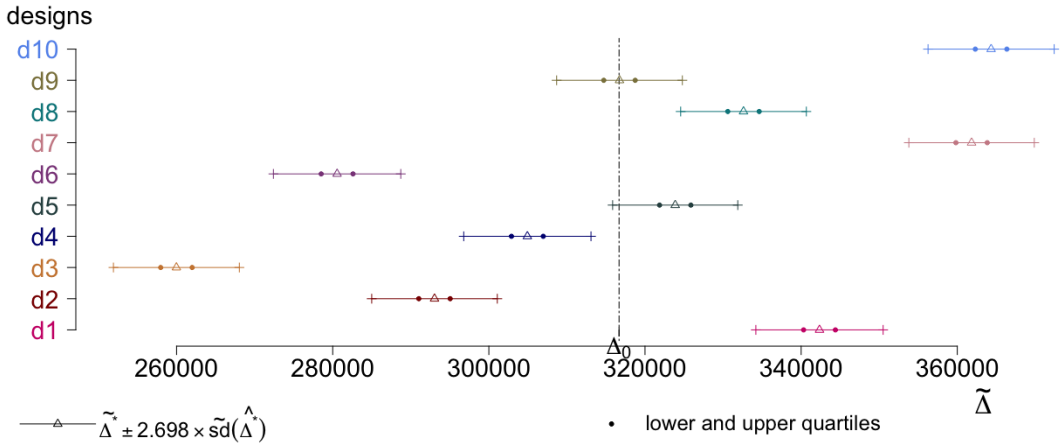


Figure 6.3: 10 randomly selected distributions of $\hat{\Delta}^*$ within 2.698 standard deviations $\tilde{sd}(\hat{\Delta}^*)$ of the mean $\tilde{\Delta}^*$.

6.2 Validation Method of the Theoretical Framework

The theoretical distribution of estimates $\hat{\theta}^*$ and $\hat{\Delta}^*$ are validated by assessing whether the theoretical distribution of both estimates are suitable to describe their empirical sampling

distribution. By using simulation procedures, a sample of size R of estimates $\hat{\theta}^*$ and $\hat{\Delta}^*$ are generated; $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_R$ and $\hat{\Delta}^*_1, \ldots, \hat{\Delta}^*_R$, where $\hat{\theta}^*_r = \begin{bmatrix} \hat{\alpha}^*_{1r} & \ldots & \hat{\alpha}^*_{Ir} & \hat{\beta}^*_r & \hat{\delta}^*_r \end{bmatrix}$, for $r \in \{1, \ldots, R\}$. Each $\hat{\theta^*}_r$ is a set of p-dimensional multivariate distribution data, where $p = I + 2$. Questions that arise here are

- how can we determine if the sampling distribution of $\hat{\theta}^*$ comes from a theoretical multivariate Normal with mean $\tilde{\theta}^*$ and variance-covariance matrix $\mathcal{C}(\tilde{\theta}^*)$.
- how can we determine if the sampling distribution of $\hat{\Delta}^*$ comes from a theoretical univariate Normal with mean $\tilde{\Delta}^*$ and variance matrix $\widetilde{\text{Var}}(\hat{\Delta}^*)$.

In other words, we need to test the following null hypothesis

$$\text{H}_0 : \hat{\theta}^* \sim \text{N}(\tilde{\theta}^*, \mathcal{C}(\tilde{\theta}^*)) \qquad \text{and} \qquad \text{H}_0 : \hat{\Delta}^* \sim \text{N}(\tilde{\Delta}^*, \widetilde{\text{Var}}(\hat{\Delta}^*)) \,.$$

We run a goodness of fit test to compare properties of the sampling distribution with the properties of the theoretical distribution. The quick check is to compute Mahalanobis distances between $\hat{\theta}^*$ and its proposed mean $\tilde{\theta}^*$ and Mahalanobis distances between $\hat{\Delta}^*$ and its proposed mean $\tilde{\Delta}^*$. Let $D^2_{\theta^*}$ denote the Mahalanobis distance of a value of $\hat{\theta}^*$ from $\tilde{\theta}^*$, then

$$D^2_{\theta^*} = (\hat{\theta}^* - \tilde{\theta}^*)^T \mathcal{C}^{-1}(\tilde{\theta}^*)(\hat{\theta}^* - \tilde{\theta}^*),$$

and let $D^2_{\Delta^*}$ denote the Mahalanobis distance of a value of $\hat{\Delta}^*$ from $\tilde{\Delta}^*$, then

$$D^2_{\Delta^*} = (\hat{\Delta}^* - \tilde{\Delta}^*)\text{Var}(\tilde{\Delta}^*)^{-1}(\hat{\Delta}^* - \tilde{\Delta}^*) \,.$$

For p-dimensional multivariate Normal data, the Mahalanobis distance $D^2_{\theta^*}$ is distributed as a Chi-squared distribution with p degrees of freedom. Then a goodness of fit test is applied to test

$$\text{H}_{0\theta^*} : D^2_{\theta^*} \sim \chi^2_p \qquad \text{and} \qquad \text{H}_{0\Delta^*} : D^2_{\Delta^*} \sim \chi^2_1 \,.$$

There exist numerous goodness-of-fit methods to test the assumption of a specific univariate distribution in the literature such as Kolmogorov–Smirnov test and Anderson-Darling test.

A computational algorithm is developed to find a theoretical and an empirical sampling distribution of $\hat{\theta}^*$ and $\hat{\Delta}^*$ for a specific campaign design and a specific truth parameters.

To find the empirical distribution, data are repeatedly generated and fitted using the applied model to produce a sample of size R for each estimate: $\hat{\theta}^*$ and $\hat{\Delta}^*$.

Given 204 spatial units, each element of the sample of $\hat{\theta}^*$ is then a $206-$dimensional multivariate distribution. The most interesting estimates are $\hat{\beta}^*$ and $\hat{\delta}^*$. Thus we investigate the sampling bivariate distribution of $\hat{\underline{\theta}}^* = \begin{bmatrix} \hat{\beta}^* & \hat{\delta}^* \end{bmatrix}$ alongside $\hat{\Delta}^*$.

6.3 Computational Algorithms

Computational algorithms are developed to validate the asymptotic distribution of the estimates $\hat{\theta}^*$ and $\hat{\Delta}^*$. The algorithms consist of two parts: theoretical asymptotic distribution of the estimates and empirical sampling asymptotic distribution of the estimates. Both parts are based on specified truth parameters and specified campaign design. The theoretical and empirical distributions will be compared in the subsequent section to demonstrate the efficiency of the theoretical framework.

The algorithms are broken down into five procedures. In the first procedure, data structures are constructed. In the second procedure, model matrices are obtained. In the third procedure, truth parameters are specified. In the fourth and fifth procedures, the theoretical and empirical distributions of the estimates are found, respectively. The following outlines the basic steps in each procedure:

procedure 1: DATA STRUCTURE (micro-census, campaign design)

function `Ground Data` (micro-census)

1. get the social grades strata; $K = 4$ for each spatial unit,
2. get the number of individuals in each stratum in a spatial unit n_{ik}[1],
3. create two time points $t = 0$ and $t = 1$ for each stratum k.

return ground data frame including four columns:

- spatial units, coded $1, \ldots, 204$,
- strata, coded $1, 2, 3, 4$,
- time periods t, coded $0, 1$,
- number of search n_{ikt},

[1]Note: we use the number of individuals - available in microcensus data - in each stratum in a spatial unit equal the number of search n_{ik} in that stratum, although this may be far away from the reality. However, micro-census data is sample based and hence n_{ik} is a subset of the population in that strata. To be close to reality, we may take a proportion of n_{ik} such as 10% and 1% of it.

where each case of the data frame represents a number of search in a stratum k in a spatial unit i during a time period t, assuming that $n_{ik} \equiv n_{ikt}$, so each unit i is repeated $2 * K$ times.

end function

function `Data Structures with No Campaign Design`(ground data)

1. add campaign status C_{it} column to ground data such that all spatial units serve the new advertising campaign in both time periods, i.e. $C_{it} = 1, \forall i, t$

return truth$^{c^1}$

2. convert C_{it} values in truth$^{c^1}$ to 0 so that all spatial units serve the old advertising campaign in both time periods, i.e. $C_{it} = 0, \forall i, t$

return truth$^{c^0}$

3. aggregate number of search in truth$^{c^1}$ by strata,

return applied$^{c^1}$

4. convert C_{it} values in applied$^{c^1}$ to 0,

return applied$^{c^0}$

end function

function `Data Structures with Campaign Design` (ground data, campaign design)

1. assign spatial units in the ground data to control or treatment based on the camping design; control: $u_i = 0$ and treatment: $u_i = 1$
2. generate campaign status $C_{it} = u_i t$
3. merge campaign status with ground data

return true data structure truth including:

- ground data,
- campaign status C_{it}: coded 0 (old ads) or 1 (new ads).

4. aggregate number of search in truth by strata,

return applied data structure applied

- aggregate ground data, i.e. no strata,
- campaign status C_{it}: coded 0 (old ads) or 1 (new ads).

end function[2]

[2]Screen-shot of subset of truth and applied data structure with and with no campaign design are presented in Appendix E in **Figure E.1** and **Figure E.2**.

end procedure 1

procedure 2: Model Matrices (data structure)

function `Model Matrices with No Campaign Design`(data structures with no campaign design)

1. create an old campaign's true model matrix X^{c^0} using the data in $\mathsf{truth}^{\mathsf{c}^0}$,
2. create a new campaign's true model matrix X^{c^1} using the data in $\mathsf{truth}^{\mathsf{c}^1}$,
3. create an old campaign's applied model matrix X^{*c^0} using the data in $\mathsf{applied}^{\mathsf{c}^0}$,
4. create a new campaign's applied model matrix X^{*c^1} using the data in $\mathsf{applied}^{\mathsf{c}^1}$,

return X^{c^0}, X^{c^1}, X^{*c^0} and X^{*c^1}

end function

function `Model Matrices with Campaign Design`(data structures with campaign design)

1. create a true model matrix X using the data in truth,
2. create an applied model matrix X^* using the data in applied,
3. create a proxy model matrix XH using applied regressors with truth structure,

return X, X^* and XH

end function

end procedure 2

procedure 3: Instances of truth parameter vector θ_0(realistic search and purchase, "B" 's advertising campaign design, K, c, (see sections 4.7.3 and 4.7.4))

function `specification of spatial effect` α_i(realistic search and purchase, B's advertising campaign design)

1. set y_{it} = realistic purchase and n_{it} = realistic search
2. create C_{it} from B's advertising campaign design
3. fit applied model: $\log(\frac{y_{it}}{n_{it}-y_{it}}) = \alpha_i^* + \beta^* t + \delta^* C_{it}$

return $\hat{\alpha_i^*}$ (The values of $\hat{\alpha_i^*}$ are presented in Appendix F)

end function

function θ_0 `based on` $\gamma-$, $\beta-$ `and` $\delta-$`cases`($\hat{\alpha_i}^*$, K, c)

1. set $\alpha_i = \hat{\alpha_i}^*$
2. create $\gamma-$, $\beta-$ and $\delta-$cases using $k = 4$ and c.

return θ_0

end function

function θ_0 `based on` $\beta\delta-$`case`($\hat{\alpha_i}^*$, K, c_1, c_2)

1. set $\alpha_i = \hat{\alpha_i}^*$
2. create $\beta\delta-$case using $k = 4$, c_1 and c_2.

return θ_0

end function

function θ_0 `based on` $\gamma\beta\delta-$`case`($\hat{\alpha_i}^*$, K, c_1, c_2, c_3)

1. set $\alpha_i = \hat{\alpha_i}^*$
2. create $\gamma\beta\delta-$case using $k = 4$, c_1, c_2 and c_3.

return θ_0

end function

end procedure 3

procedure 4: Theoretical Distribution (models matrices, θ_0, θ_0^*, n_{ikt})

function `KL-divergence Optimisation` (X, XH, θ_0, θ_0^*)

1. calculate the truth $\eta = X\theta_0$,
2. set proxy model $\eta^\dagger = XH\theta^*$ as a function of θ^*,
3. set D_{KL} function between the truth and the proxy model as a function of θ^*,
4. optimise D_{KL}, using θ_0^* as a starting parameter vector,

return $\tilde{\theta}^*$: the parameter vector that minimises D_{KL} and the approximated mean of $\hat{\theta}^*$.

end function

function `Asymptotic Variance of` $\hat{\theta}^*$(X, X^*, θ_0, n_{ikt}, $\tilde{\theta}^*$)

1. find $\mathcal{A}(\tilde{\theta}^*)$ using X^*, $n_{it} = \sum_k n_{ikt}$ and $\tilde{\theta}^*$,
2. find $\mathcal{B}(\tilde{\theta}^*)$ using X, θ_0, X^* and n_{ikt},
3. find inverse of $\mathcal{A}(\tilde{\theta}^*)$,
4. compute $\mathcal{A}^{-1}(\tilde{\theta}^*)\mathcal{B}(\tilde{\theta}^*)\mathcal{A}^{-1}(\tilde{\theta}^*)$

return $\mathcal{C}(\hat{\theta}*)$: variance of $\hat{\theta}^*$.

end function

function `Approximation of the Overall Effect` Δ^* $(X^{*c^0}, X^{*c^1}, n_{it}, \tilde{\theta}^*)$

1. find the expected total number of sales when $C_{it} = 1$, $\forall$ i and t, using X^{*c^1}, $\tilde{\theta}^*$ and n_{it},
2. find the expected total number of sales when $C_{it} = 0$, $\forall$ i and t, using X^{*c^0}, $\tilde{\theta}^*$ and n_{it},
3. find the differential measure $\tilde{\Delta}^*$; the difference between the expectations computed in steps 1 and 2.

return $\tilde{\Delta}^*$: the best value to use for the applied overall effect Δ^* and the approximated mean for $\hat{\Delta}^*$.

end function

function `Asymptotic Variance of` $\hat{\Delta}^*(\Delta^*, \tilde{\theta}^*, \mathcal{C}(\hat{\theta}^*))$

1. set the Jacobian matrix $^{\partial\Delta^*}/_{\partial\theta^*}$,
2. find Jacobian matrix at $\tilde{\theta}^*$, i.e. $^{\partial\Delta^*}/_{\partial\tilde{\theta}^*}$,
3. compute $(^{\partial\Delta^*}/_{\partial\tilde{\theta}^*})^T \, \mathcal{C}(\hat{\theta}^*) \, (^{\partial\Delta^*}/_{\partial\tilde{\theta}^*})$,

return $\widetilde{\mathsf{Var}}(\hat{\Delta}^*)$: the variance of the approximated overall effect $\hat{\Delta}^*$.

end function

end procedure 4

procedure 5: Empirical Sampling Distribution $(\eta, n_{ikt}, \Delta^*)$

1. find p_{ikt} using the truth η,
2. generate $y_{ikt} \sim \text{Bin}(n_{ikt}, p_{ikt})$,
3. aggregate y_{ikt} by strata to have applied structure y_{it} and n_{it},
4. fit y_{it} using logit-linear regression model,
5. get the regression coefficients, i.e. $\hat{\theta}^*$
6. find the estimate $\hat{\Delta}^*$ by computing Δ^* at $\hat{\theta}^*$,
7. repeat steps 2 to 6 R times,
8. get a sample of $\hat{\theta}^*$ of size R,
9. get a sample of $\hat{\Delta}^*$ of size R,
10. find the mean and variance of $\hat{\theta}^*$,
11. find the mean and variance of $\hat{\Delta}^*$,

end procedure 5

The algorithms are implemented using R. In procedure 4, the KL-divergence is optimised using Broyden-Fletcher-Goldfarb-Shanno (BFGS) method. The computational algorithm is general enough to include different truths and different campaign designs. In addition, design strategies procedure can be added in the algorithm over the data structure procedure to take in various strategies in producing campaign designs. However, to avoid any confusion with the pre-specified 10 campaign designs, the design strategies procedure are not presented here but the procedure is presented in Appendix G. The procedure will be needed in the coming chapter to evaluate performances of different design strategies.

In what follows, we investigate the empirical sampling distribution of $\hat{\underline{\theta}}^*$ and $\hat{\Delta}^*$ for different truth instances and a meta campaign design; different simulation designs, on the proposed truth instances.

6.4 Computational Results

Given a truth case with a specified c, the computational algorithm is implemented for the 10 specified campaign designs to find 10 theoretical asymptotic distributions of $\hat{\Delta}^*$. For the same input truth and designs, the algorithm is implemented to construct empirical sampling distribution to each theoretical distributions, each of size $R = 1000$. In the following discussion, we investigate whether a specific theoretical distribution describes best the sampling distribution when the applied inputs mechanism are the same.

6.4.1 Truth Parameters: $\delta-$Case, $c \in \{0.2, 1, 5\}$

Consider the truth parameter vector θ_0 based on $\delta-$case with parameters

$$\alpha_i = \alpha_i, \quad \gamma_k = \beta_k = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \quad \delta_k = \begin{bmatrix} -1.5c & -0.5c & 0.5c & 1.5c \end{bmatrix}.$$

Given $c = 5$, $\delta_k = \begin{bmatrix} -7.5 & -2.5 & 2.5 & 7.5 \end{bmatrix}$, **Figure 6.3** illustrates the theoretical distribution of $\hat{\Delta}^*$ within 2.698 units of standard deviations $\tilde{\text{sd}}(\hat{\Delta}^*)$ of the mean $\tilde{\Delta}^*$. By adding the empirical distributions of $\hat{\Delta}^*$ in line with each specified design in the figure, we get the distributions displayed in **Figure 6.4**.

Figure 6.4 shows that the empirical sampling distributions $\hat{\Delta}^*$ - that are presented in box plots - are in good agreement with the theoretical distributions for all 10 designs. This agreement provides an important indication into the key role of the theoretical approximation to the sampling distribution of applied effect $\hat{\Delta}^*$. To understand the behaviour of these distributions, we compare them with other distributions that are generated using

the same 10 selected campaign designs but using different truth parameters. It would be interesting to begin the comparison within the same truth case using different c values.

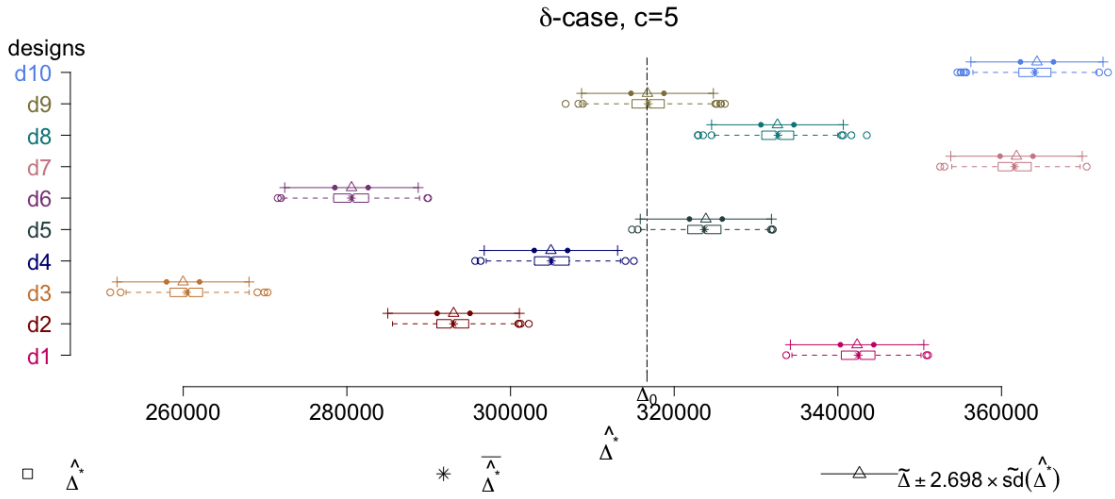


Figure 6.4: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$ for $c = 5$.

The two figures presented in **Figure 6.5** show that the empirical sampling distributions of $\hat{\Delta}^*$ are in harmony with the theoretical distributions. By comparing the distributions presented here to the ones obtained by using truth instance $c = 5$, it is noticeable that when c is smaller; there are smaller differences between the strata, values of all overall effects, i.e. Δ_0, $\tilde{\Delta}$ and $\hat{\Delta}^*$ are smaller. In other words when c is smaller, the campaign has less effect. For example for $c = 0.2$, the true effect is $\Delta_0 = 6963.728$ and for $c = 1$ is $\Delta_0 = 91576.92$ compared to $\Delta_0 = 316677.5$ when $c = 5$. Interestingly, for smaller differences between strata, locations of the centre of both theoretical and empirical distributions tend to get closer to the true overall effects. For example when $c = 0.2$, in its most presented distributions the theoretical mean $\tilde{\Delta}$ and the median of $\hat{\Delta}^*$ sample are close to Δ_0. The variability between the 10 designs appears to be low when $c = 0.2$ followed by $c = 1$ and then $c = 5$. The change in the variability within a single design across c values is not clear, maybe due to the disparity in the scales. However by focusing on the scale in the x-axis in each plot, it seems that the variation within each single design is about the same across c.

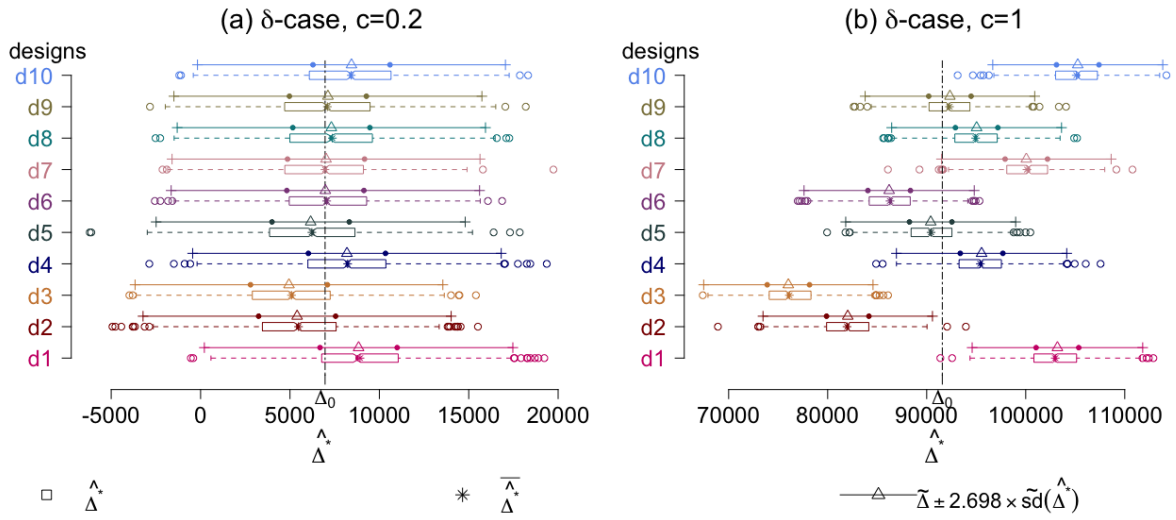


Figure 6.5: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$ for $c = 0.2$, $c = 1$.

So far the investigation of the theoretical distribution has been about $\hat{\Delta}^*$. In the following discussion we compare the theoretical distribution of $\hat{\underline{\theta}}^*$ to its empirical distribution. **Figure 6.6** and **Figure 6.7** illustrate the univariate distributions of $\hat{\beta}^*$ and $\hat{\delta}^*$, respectively.

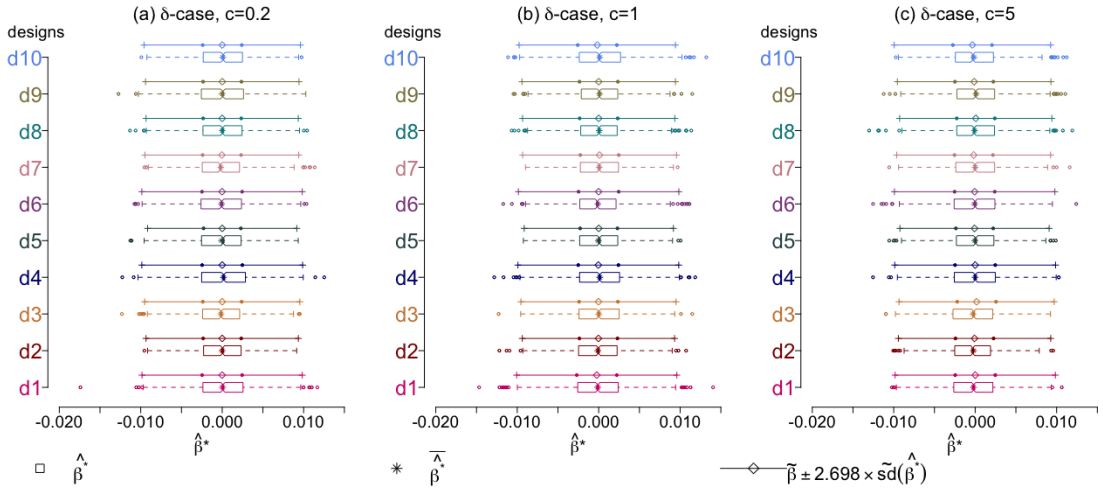


Figure 6.6: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\beta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

For $\hat{\beta}^*$, the range of the theoretical and empirical distributions is about the same across c. This may be due to the fact that $\beta_k = 0$ in $\delta-$case. The sample means $\bar{\hat{\beta}}^*$ obtained from the 10 designs - in the three truth instances - approach the theoretical mean $\tilde{\beta}^*$. The only

exception is that design $d3$ in $c = 5$ produces sample mean $\bar{\hat{\beta}}^*$ that is a bit smaller than the theoretical mean $\tilde{\beta}^*$.

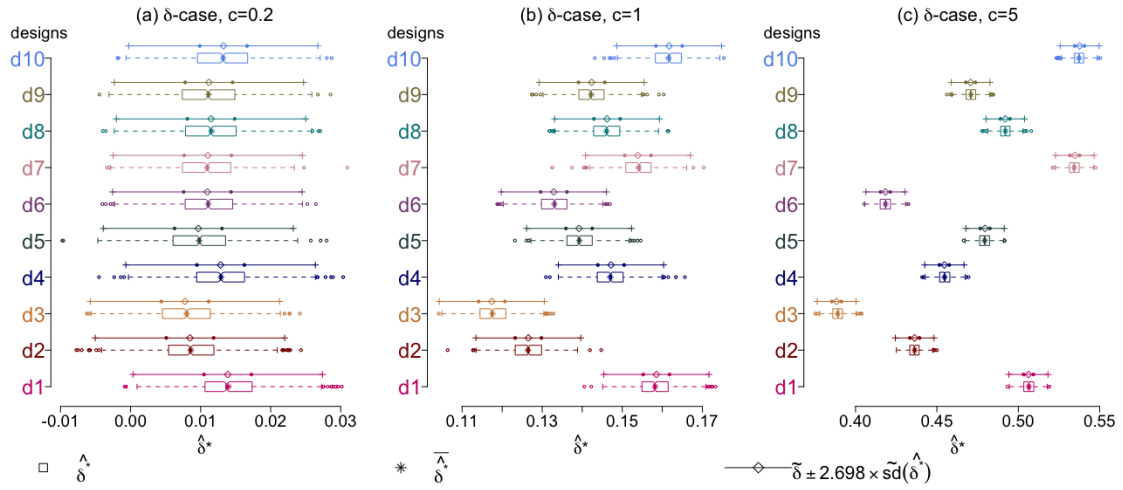


Figure 6.7: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\delta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

For $\hat{\delta}^*$, the values are positive and the range of the theoretical and empirical distributions increases as the difference between strata increases. The change in the distributions of $\hat{\delta}^*$ between designs across c looks similar to the behaviour of $\hat{\Delta}^*$ presented above. The empirical univariate distributions of both $\hat{\beta}^*$ and $\hat{\delta}^*$ agree with their theoretical distributions, indicating that theoretical approximations to the sampling distributions of both estimates work well.

It is important to note that the results presented are based on having a different campaign effect in each stratum and that the proportion of people in each stratum varies between spatial units and the allocation of units to treatment or control varies between designs. This is likely to affect the estimation of the weighted parameters. One interesting feature is that δ_k is zero on average, but $\tilde{\delta}^*$ and δ^* are non-zero and typically are positive, taking larger positive values as c increases and with more variation between designs as c increases. On the other hand, the within design variation in $\hat{\delta}^*$ remains fairly constant as c increases. However, the number of searches in each stratum is not the same and so if we calculated the weighted of those 4 numbers using the numbers of searches as weights, the average would not be zero. It is found that the weighted average based on proportion of population in each stratum is positive, about 0.027. This gives some reason for why $\tilde{\delta}^*$ and δ^* tend to be positive and getting larger as c increases. Moreover, it makes sense that $\tilde{\delta}^*$ would

vary more between designs because any difference in the weighted average of δ_k between treatment and control groups (the difference varies between designs) would be magnified as c increases.

Now consider the goodness of fit between theoretical distributions and empirical distributions. As pointed out in section 6.2, the Mahalanobis distances $D^2_{\Delta^*}$ and $D^2_{\theta^*}$ are used to measure the dissimilarity between the theoretical and empirical distributions of $\hat{\Delta}^*$ and $\hat{\underline{\theta}}^*$, respectively. We applied the Kolmogorov–Smirnov (KS) and the Anderson-Darling (AD) tests of goodness of fit to test $\mathrm{H}_{0\theta^*} : D^2_{\theta^*} \sim \chi^2_2$ and $\mathrm{H}_{0\Delta} : D^2_{\Delta^*} \sim \chi^2_1$.

The $p-$values are evaluated for both null hypotheses $\mathrm{H}_{0\theta^*}$ and $\mathrm{H}_{0\Delta^*}$ for the three truth instances and the 10 designs. The $p-$values are provided in **Figure 6.8**. By considering significance level 0.05, both tests fail to reject the two hypotheses when $c = 1$. Looking at part (a) and (c), the tests reject the two hypotheses except for a few designs. For $c = 0.2$, the tests provide significant $p-$values against both hypothesis for $d9$. In the same graph, the tests provide significant $p-$values against $\mathrm{H}_{0\Delta^*}$ for $d5$, and the KS test provides a significant $p-$value against $\mathrm{H}_{0\theta^*}$ for $d4$. In $c = 5$, the tests provide significant $p-$values against $\mathrm{H}_{0\theta^*}$ for $d2$.

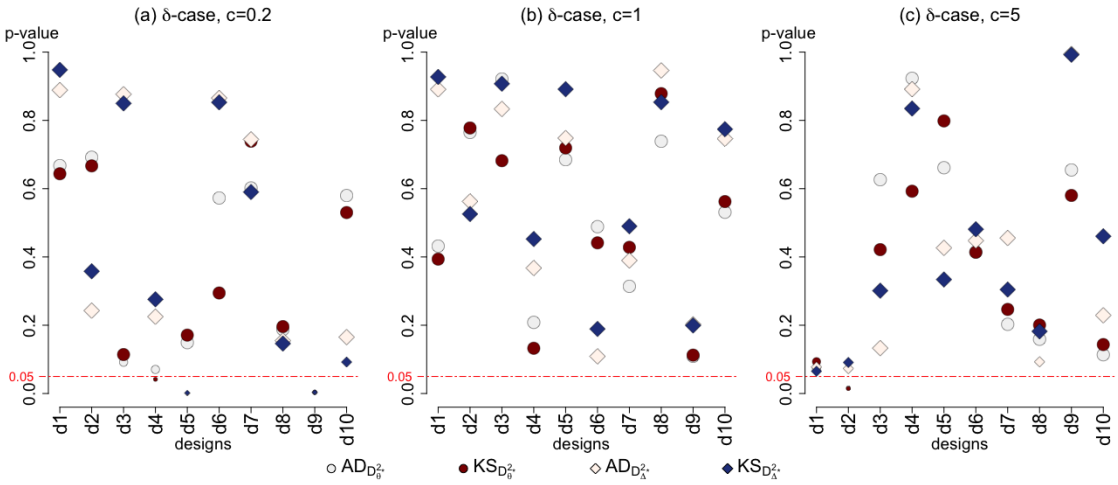


Figure 6.8: $\delta-$case: $p-$values obtained from the Kolmogorov Smirnov (KS) and Anderson Darling (AD) tests of goodness of fit to test $\mathrm{H}_{0\theta^*} : D^2_{\theta^*} \sim \chi^2_2$ and $\mathrm{H}_{0\Delta^*} : D^2_{\Delta^*} \sim \chi^2_1$ for $c = 0.2$, $c = 1$ and $c = 5$.

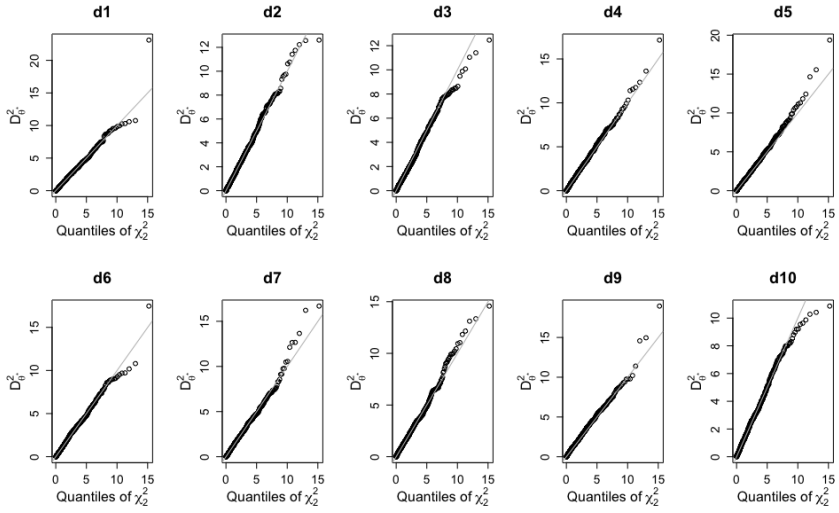
It is important to note that 120 significance tests are conducted in this truth case example, 10 designs $\times$ 3 values of $c \times$ 4 tests of goodness of fit. Interestingly, the $p-$values presented

in **Figure 6.8** are not independent. While the same 10 designs are used throughout, the 4 tests are not independent and c is moving along a scale. Thus, it is difficult to say what $p-$values should be expected. Despite this, finding some $p-$values less than 0.05 might be expected due to the fact that in conducting lots of significance tests at level 0.05, there is 5% chance to reject the null hypothesis even when it is true.

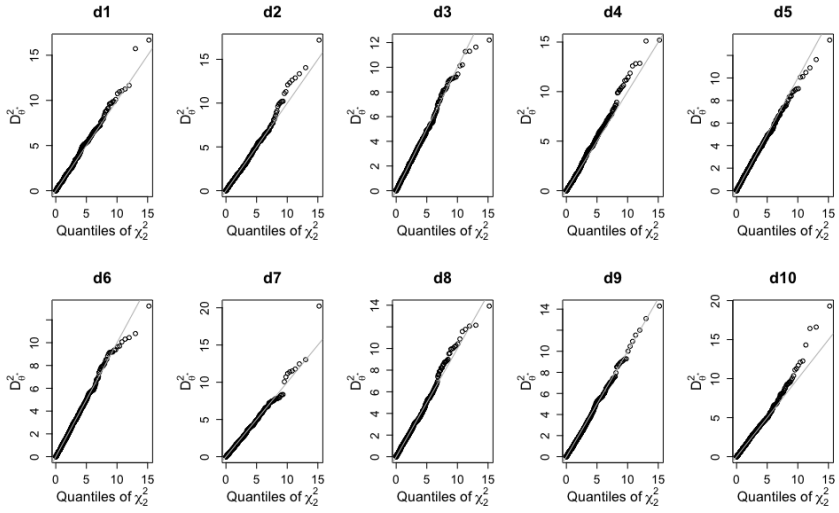
The significant cases may indicate that the sampling distribution of $\underline{\hat{\theta}}^*$ and $\hat{\Delta}^*$ are not the same as their theoretical asymptotic distributions. However, the significant evidence against the null hypotheses does not indicate how the sampling distribution differs from the theoretical one. Especially, one of the problems with goodness of fit tests is they are more likely to reject null hypotheses when the sample size is large. By taking into account the investigation of the marginal distributions $\underline{\hat{\theta}}^*$ and the distribution of $\hat{\Delta}^*$ that are presented earlier in this section, we can see that their sample means are not far from the theoretical means.

Furthermore, if we plot the Mahalanobis distances $D^2_{\theta^*}$ and $D^2_{\Delta^*}$ versus quantiles of χ^2_2 and χ^2_1, respectively, we get the following graphs shown in **Figure 6.9** and **Figure 6.10**, respectively. In both figures, we can see that the points in all graphs for the three truth instances, tend to fall along a straight line, despite extreme points on their top right tails. Taken together, these plots suggest that $\underline{\hat{\theta}}^*$ and $\hat{\Delta}^*$ are close to normal and with distributions as predicted by the theory.

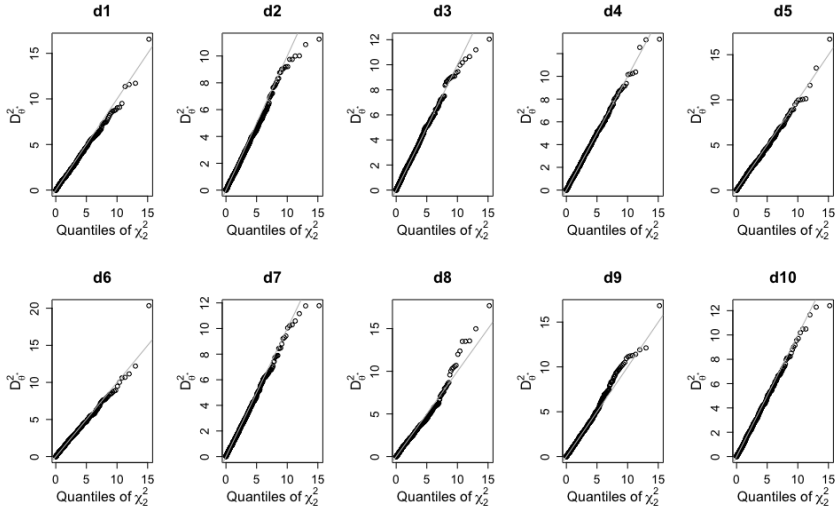
Overall, these truth instances of $\delta-$case support the view that the theoretical asymptotic distribution of $\underline{\hat{\theta}}^*$ and $\hat{\Delta}^*$ are adequate to describe the empirical distribution.



(a) c=0.2

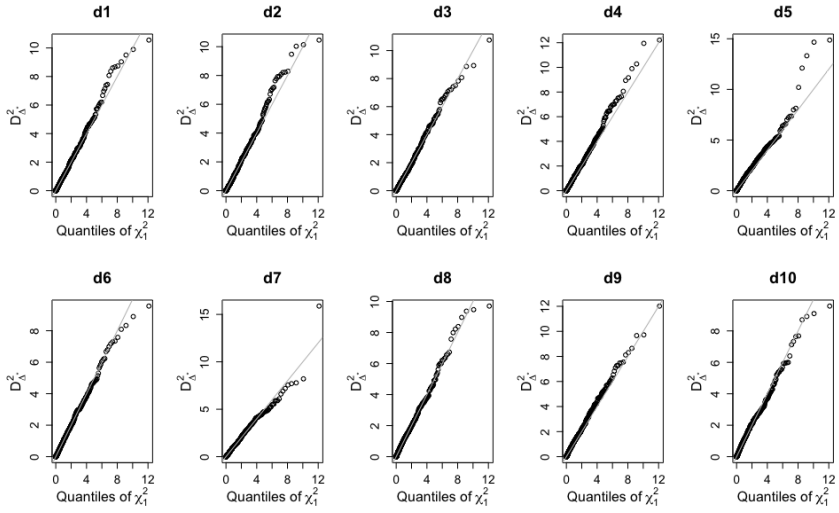


(b) c=1

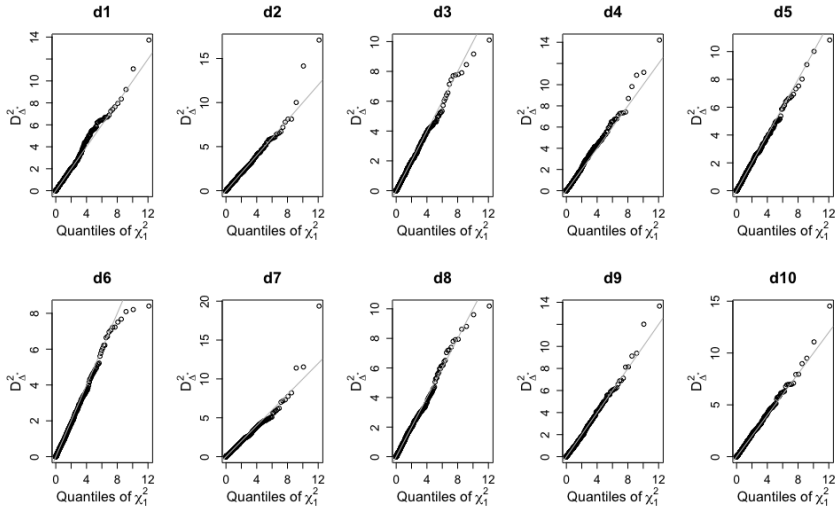


(c) c=5

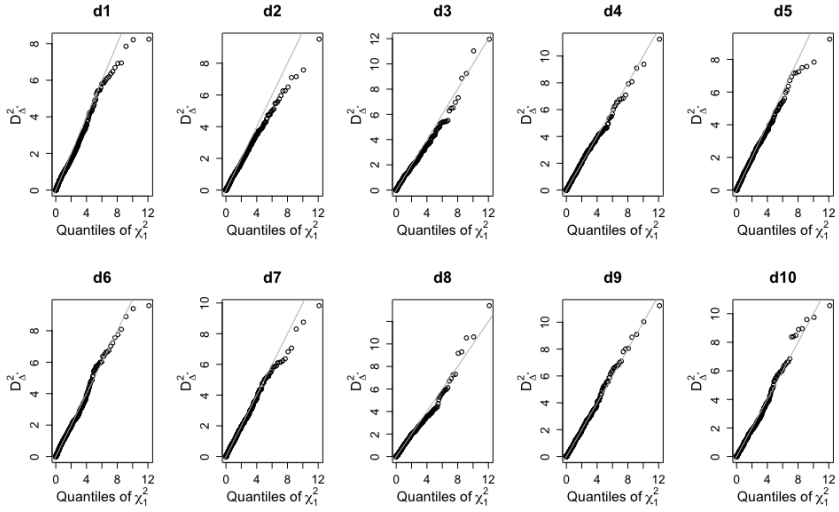
Figure 6.9: $\delta-$case: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\theta^*}$ versus quantiles of χ^2_2



(a) c=0.2



(b) c=1



(c) c=5

Figure 6.10: $\delta-$case: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ versus quantiles of χ^2_1

6.4.2 Truth Parameters: $\beta-$case, $c \in \{0.2, 1, 5\}$

Consider the truth parameter vector θ_0 based on $\beta-$case with parameters

$$\alpha_i = \alpha_i, \quad \gamma_k = \delta_k = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \quad \beta_k = \begin{bmatrix} -1.5c & -0.5c & 0.5c & 1.5c \end{bmatrix}.$$

A comparison of the empirical and theoretical distributions of $\hat{\Delta}^*$ are presented in **Figure 6.11**. The true effect $\Delta_0 = 0$ in the three instances; because δ_k is specified to be zero in this truth case. Likewise, $\delta-$case, the range of $\hat{\Delta}^*$ increases as the difference between strata gets larger. However, the values on the scale are rather smaller compared to those in $\delta-$case. The values of the proxy effect $\tilde{\Delta}^*$ vary between negative and positive unlike $\delta - case$ where all are positive. It appears from the figure that the empirical distributions are consistent with the theoretical distributions.

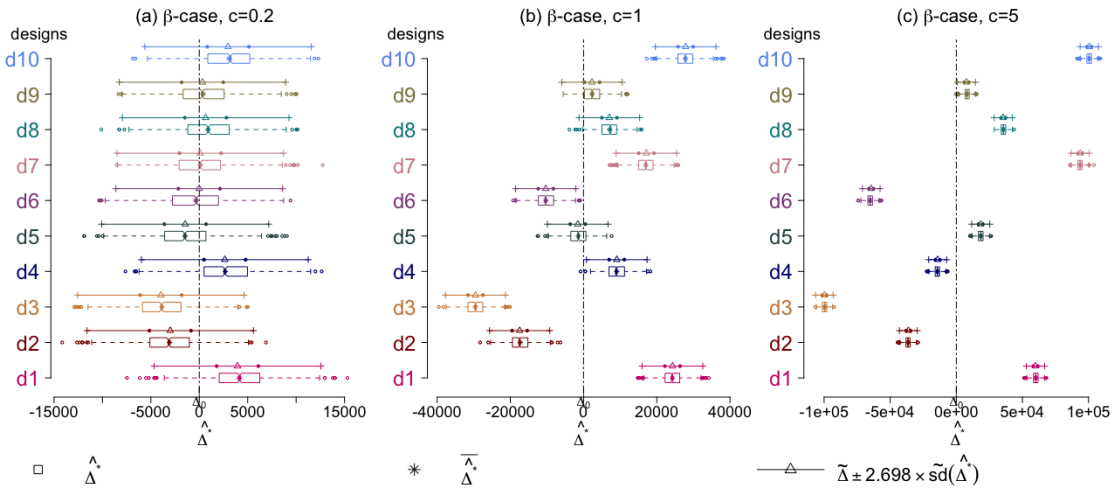


Figure 6.11: $\beta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

The marginal distributions of $\hat{\underline{\theta}}^*$ for the three instances of β-case are presented in **Figure 6.12** for $\hat{\beta}^*$ and in **Figure 6.13** for $\hat{\delta}^*$. Although $\delta_k = 0$ in this case, the change in the behaviour of $\hat{\delta}^*$ is obvious across c unlike the behaviour of $\hat{\beta}^*$ in $\delta-$case. Both figures shows that the variations between designs gets higher as the difference gets larger. Additionally, the behaviour of $\hat{\beta}^*$ in this case is in parallel with the behaviour of $\hat{\delta}^*$ in $\delta-$case. The sampling distributions of $\hat{\beta}^*$ and $\hat{\delta}^*$ appear to be consistent with their corresponding theoretical distributions across designs and c values.

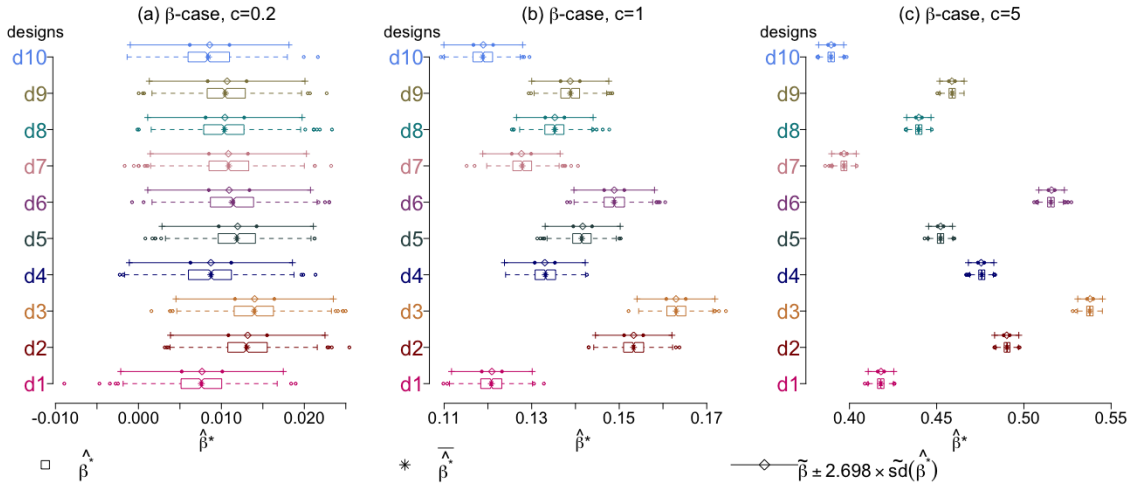


Figure 6.12: $\beta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\beta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

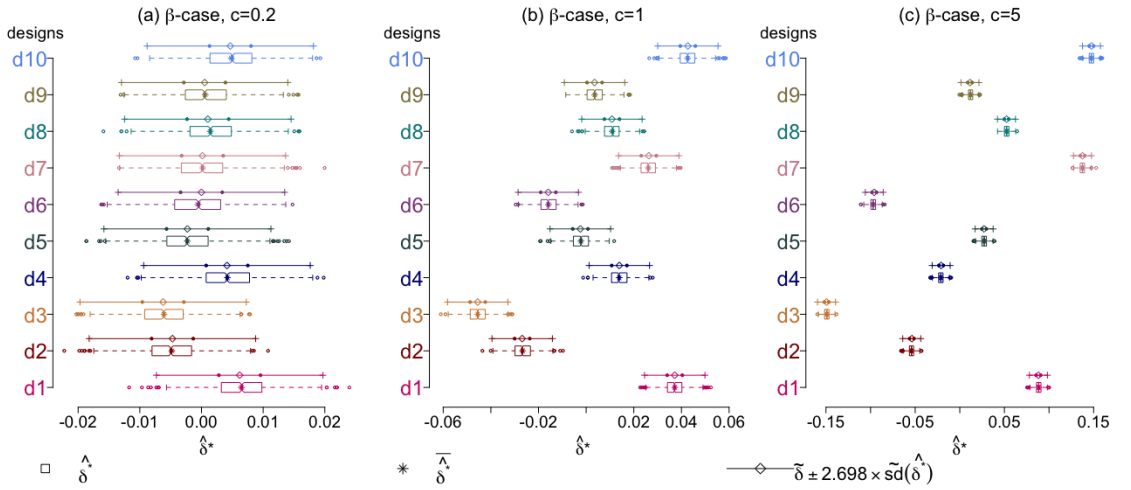


Figure 6.13: $\beta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\delta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

Figure 6.14 provides the resulting $p-$values form the AD and the KS tests. The values suggest no sufficient evidences to reject the claim that $\hat{\underline{\theta}}^*$ and $\hat{\Delta}^*$ are coming from the theoretical distribution, except for some cases in $c = 0.2$ and $c = 5$. We present the Chi-square qq-plots in Appendix H in **Figure H.2** and **Figure H.3** to assess normality of $\hat{\underline{\theta}}^*$ and $\hat{\Delta}^*$. The figures support normality claims in general except designs $d6$ in $c = 5$ provide rejection evidences, where we can see the points are not exactly on the straight line. This case requires further investigations to understand why the theoretical asymptotic normality is rejected.

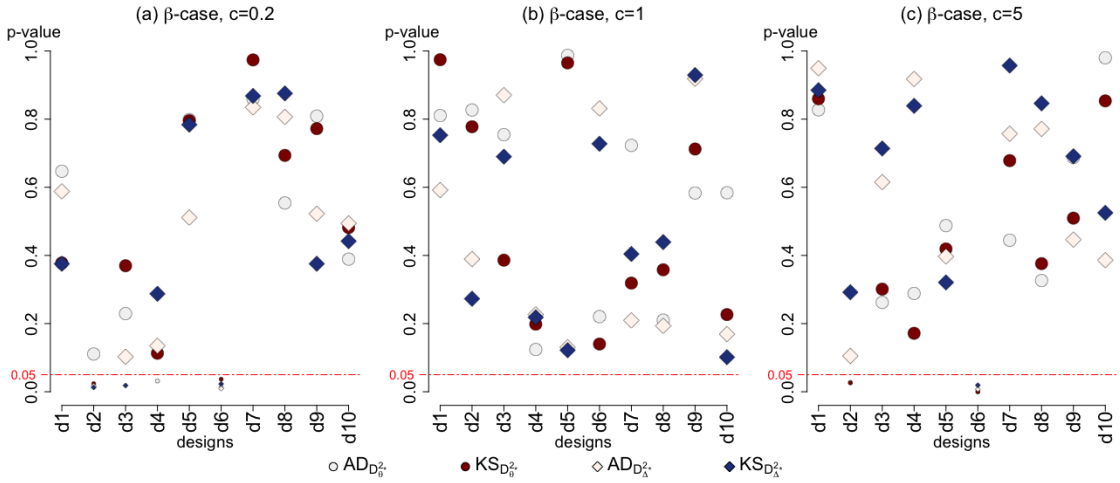


Figure 6.14: $\beta-$case: $p-$values obtained from the Kolmogorov Smirnov (KS) and Anderson Darling (AD) tests of goodness of fit to test $\mathrm{H}_{0\theta^*} : D^2_{\theta^*} \sim \chi^2_2$ and $\mathrm{H}_{0\Delta^*} : D^2_{\Delta^*} \sim \chi^2_1$ for $c = 0.2$, $c = 1$ and $c = 5$.

Given that $\hat{\Delta}^*$ is a univariate sample, a histogram can be plotted to assess graphically if the sample is normally distributed or not. Plotting a histogram gives an indication of the shape of the distribution. Consider $d6$, the sample distribution of $\hat{\Delta}^*$ is presented on histogram plots in **Figure 6.15**. From the figure it can be seen that the sample is approximately bell shaped. The other graphical method of assessing the normality is the normal Q-Q plot. On the same figure, the normal Q-Q plots are shown. The points form a line that are roughly straight, indicating that the sets of quantiles came from the same distribution shape, and hence the distribution of $\hat{\Delta}^*$ is normal in design $d6$.

Alternatively, Shapiro-Wilk's (SW) test can be used to test the normality of the underlying distribution of $\hat{\Delta}^*$. For $d6$, the resulting p-value is 0.4471, which supports the hypothesis that $\hat{\Delta}^*$ is normal. In addition, the KS can be used to test specifically if $\hat{\Delta}^* \sim \mathrm{N}(\tilde{\Delta}^*, \mathrm{Var}(\tilde{\Delta}^*))$. For $d6$, the test p-value is about 2.665×10^{-15} indicating that the sample distribution of $\hat{\Delta}^*$ is biased. The distribution's peak of $\hat{\Delta}^*$ in the histogram in **Figure 6.15** is not really far from the theoretical peak value.

To examine the dissimilarity between the theoretical and empirical distributions of $\hat{\Delta}^*$ for $d6$, we consider the standardised sampling distributions of $\hat{\Delta}^*$ with standardised scores $z_{\hat{\Delta}^*}$ such that

$$z_{\hat{\Delta}^*} = \frac{\hat{\Delta}^* - \tilde{\Delta}^*}{\tilde{\mathrm{sd}}(\hat{\Delta}^*)} \sim \mathrm{N}(0, 1)\,.$$

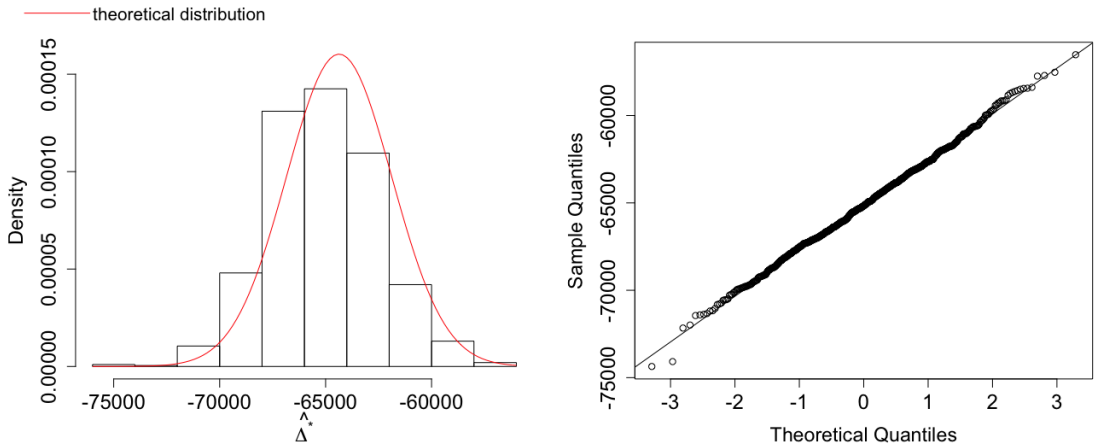


Figure 6.15: $\beta-$case: Graphical assessments of Normality of $\hat{\Delta}^*$ of design $d6$ in $c = 5$. Left: Histogram of the empirical distribution of $\hat{\Delta}^*$ with densities and overlaid theoretical asymptotic normal density curve. Right: The Normal Q-Q plot for $\hat{\Delta}^*$.

The mean and standard deviation of standardised $\hat{\Delta}^*$ is given by $\bar{z}_{\hat{\Delta}^*} = -0.30$ and $\text{sd}(z_{\hat{\beta}^*}) = 1.03$. The mean of $z_{\hat{\Delta}^*}$ deviates from the standard mean by 30% of the standard units, whereas the standard deviations is approximately 1. The standardised distributions of $\hat{\Delta}^*$ is shown in **Figure 6.16** with a standard Normal curve. The plot of the histogram indicates that $z_{\hat{\Delta}^*}$ has approximately standard Normal distribution. Equivalently, the sample $\hat{\Delta}^*$ approximately follows the theoretical asymptotic distribution.

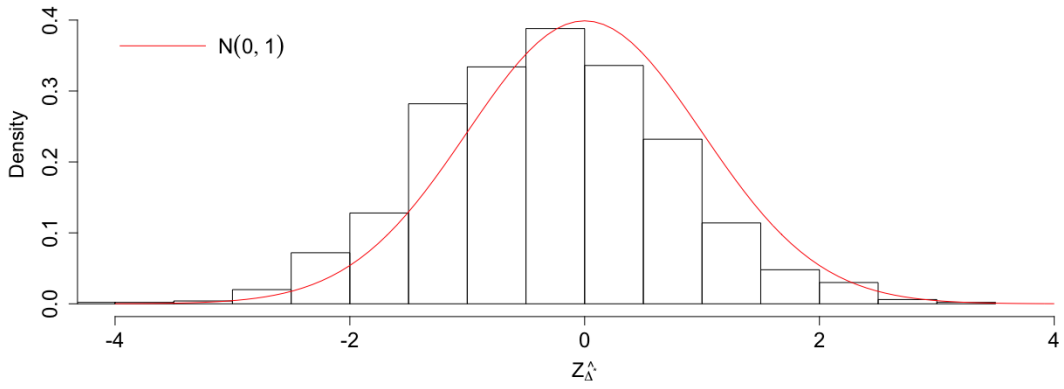


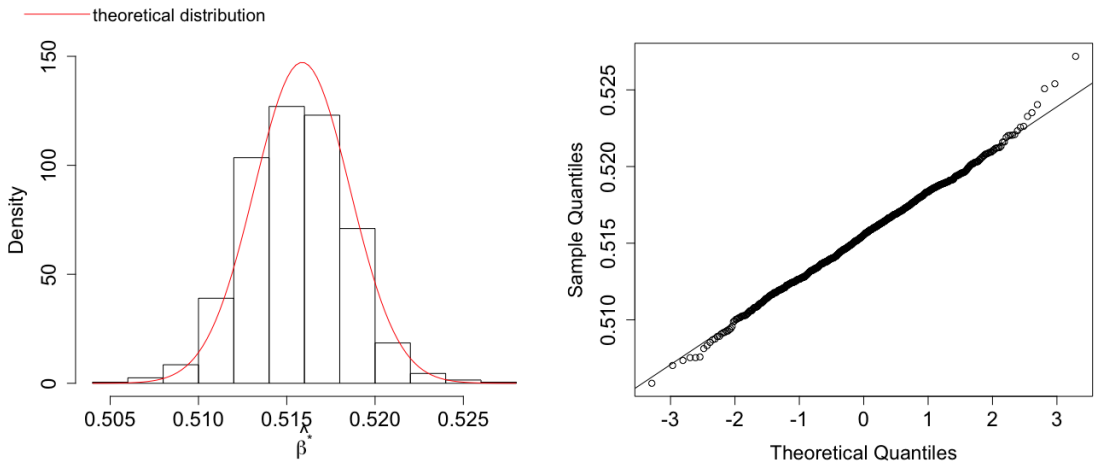
Figure 6.16: $\beta-$case: Graphical assessments of Normality of $\hat{\Delta}^*$ of design $d6$ in $c = 5$. Histogram of the standardised empirical distribution of $\hat{\Delta}^*$ with densities and overlaid standard normal density curve.

In a similar way, we assess the normality of $\hat{\underline{\theta}}^*$. The histograms and the normal Q-Q plots of $\hat{\beta}^*$ and $\hat{\delta}^*$ **Figure 6.17a** and **Figure 6.17b**, respectively, show that both samples are normally distributed shape. Additionally, SW test provide insignificant $p-$values 0.1665 and 0.4331 for $\hat{\beta}^*$ and $\hat{\delta}^*$ receptively. However, the KS test provides significant $p-$values 5.696×10^{-5} and 1.332×10^{-15} indicating that the parameters for two distributions do not match exactly the values provided by the asymptotic theory.

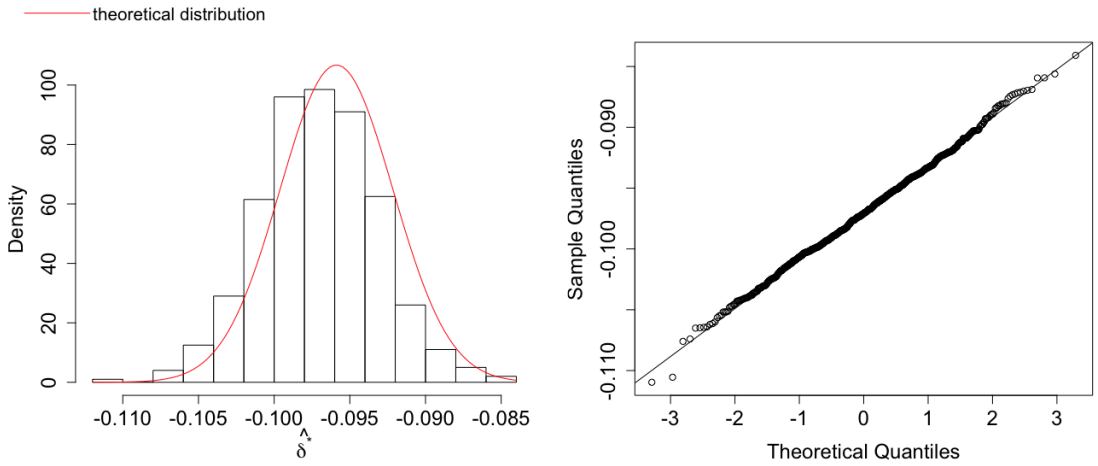
To study how far the sampling distributions of $\hat{\beta}^*$ and $\hat{\delta}^*$ are from the theoretical distributions, consider their standardised sampling distributions with standardised scores $z_{\underline{\hat{\beta}}^*}$ and $z_{\underline{\hat{\delta}}^*}$, such that

$$z_{\hat{\beta}^*} = \frac{\hat{\beta}^* - \tilde{\beta}^*}{\tilde{\text{sd}}(\hat{\beta}^*)} \sim \text{N}(0,1) \qquad \text{and} \qquad z_{\hat{\delta}^*} = \frac{\hat{\delta}^* - \tilde{\delta}^*}{\tilde{\text{sd}}(\hat{\delta}^*)} \sim \text{N}(0,1)\,.$$

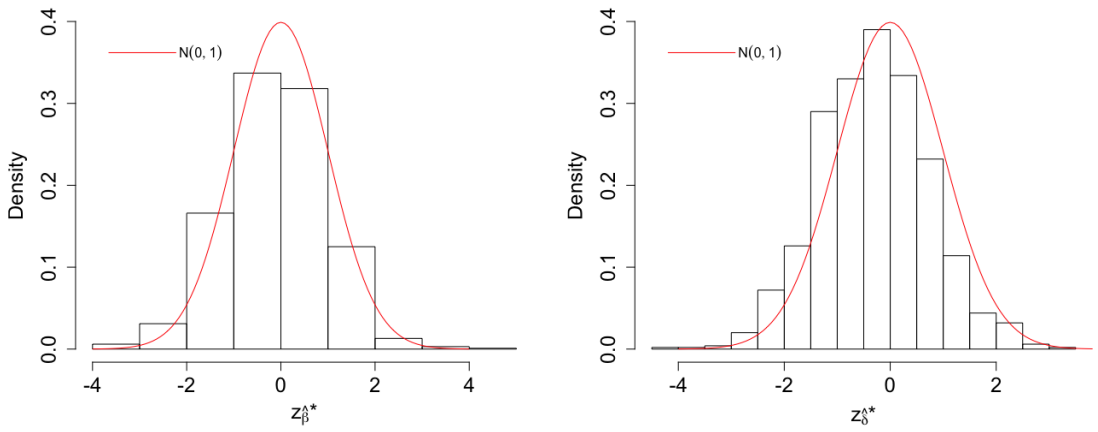
The mean and standard deviation of standardised $\hat{\beta}^*$ and $\hat{\delta}^*$ are given by $\bar{z}_{\hat{\beta}^*} = -0.1388571$, $\text{sd}(z_{\hat{\beta}^*}) = 1.045682$, and $\bar{z}_{\hat{\delta}^*} = -0.304391$, $\text{sd}(z_{\hat{\delta}^*}) = 1.034122$, respectively. The mean of $z_{\hat{\beta}^*}$ deviates from the standard mean by 13% of the standard unit and the mean of $z_{\hat{\delta}^*}$ deviates from the standard mean by 30% of the standard unit. The standard deviations of both sample estimates, however, are approximately 1. The standardised distributions of $\hat{\beta}^*$ and $\hat{\delta}^*$ are shown in **Figure 6.17c** with a standard Normal curve. The plots indicate that both $\hat{\beta}^*$ and $\hat{\delta}^*$ are approximately normally distributed with means close to $\tilde{\beta}^*$ and $\tilde{\delta}^*$ and standard deviations close to $\tilde{\text{sd}}(\hat{\beta}^*)$ and $\tilde{\text{sd}}(\hat{\delta}^*)$.



(a) Histogram of the empirical distribution of $\hat{\beta}^*$ and its Normal Q-Q plot.



(b) Histogram of the empirical distribution of $\hat{\delta}^*$ and its Normal Q-Q plot



(c) Standardised distributions of $\hat{\beta}^*$ and $\hat{\delta}^*$

Figure 6.17: $\beta-$case: Graphical assessments of Normality of $\hat{\beta}^*$ and $\hat{\delta}^*$ of $d6$ in $c = 5$.

6.4.3 Truth Parameters: $\gamma-$case, $c \in \{0.2, 1, 5\}$

Consider the truth parameter vector θ_0 based on $\gamma-$case with parameters

$$\alpha_i = \alpha_i, \quad \beta_k = \delta_k = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}, \quad \gamma_k = \begin{bmatrix} -1.5c & -0.5c & 0.5c & 1.5c \end{bmatrix}.$$

In **Figure 6.18**, the empirical and theoretical distributions of $\hat{\Delta}^*$ are provided for the three truth instances of γ-case. Given $\delta_k = 0$, the true effect Δ_0 for all instances is zero. For $c = 0.2$, the two distributions for all designs are almost identical with zero means. For $c = 1$, the behaviour of both distributions appear to be similar to those in $c = 0.2$. However there is a slight change in the empirical distributions obtained from designs $d3$, $d5$, $d7$ and $d8$. In addition, the proxy effects of $d3$ and $d5$ in particular move slightly from the centre Δ_0. These two designs show a sign that the empirical and theoretical distributions are not in agreement. For $c = 5$, the range of $\hat{\Delta}^*$ decreases compared to other instances, unlike what are seen in $c = 5$ in $\beta-$case and $\delta-$case, where the range increases as c gets larger. In this case, four designs only $d3$, $d7$, $d8$ and $d10$ show full agreement between empirical and theoretical distributions with zero means. For the other designs the theoretical mean slightly shift either to left or right of the sample mean. By comparing this figure to the distributions of $\hat{\Delta}^*$ in $\beta-$case and $\delta-$case, the impact of level of differences between strata on the variations between designs are not obvious in this truth case.

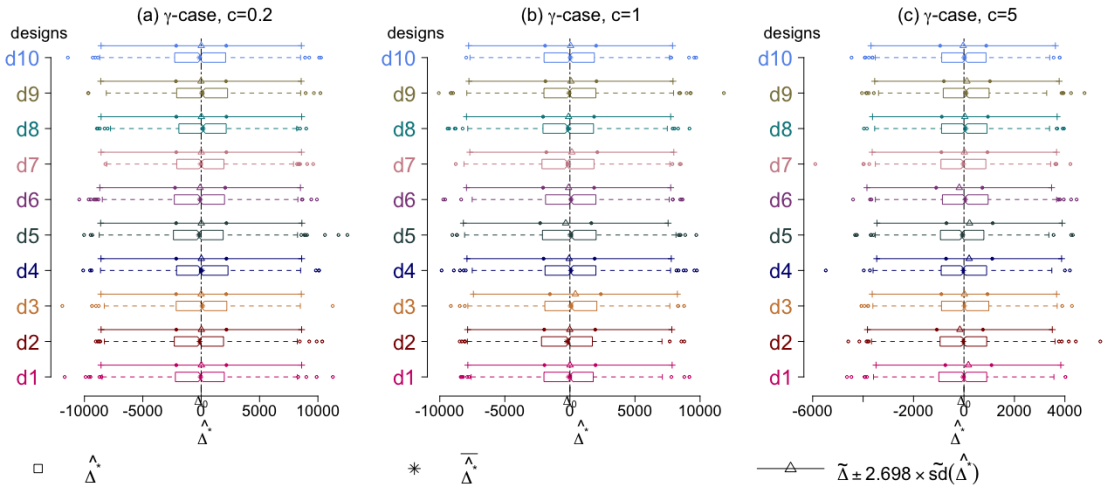


Figure 6.18: $\gamma-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

In contrast to earlier findings of the marginal distributions of $\hat{\underline{\theta}}^*$, **Figure 6.19** for $\hat{\beta}^*$ and **Figure 6.20** for $\hat{\delta}^*$ do not show a clear effect of level of difference between strata on the behaviour of the estimates. The theoretical and empirical distributions of both estimates appear to meet except for designs $d4$ and $d5$ when $c = 5$.

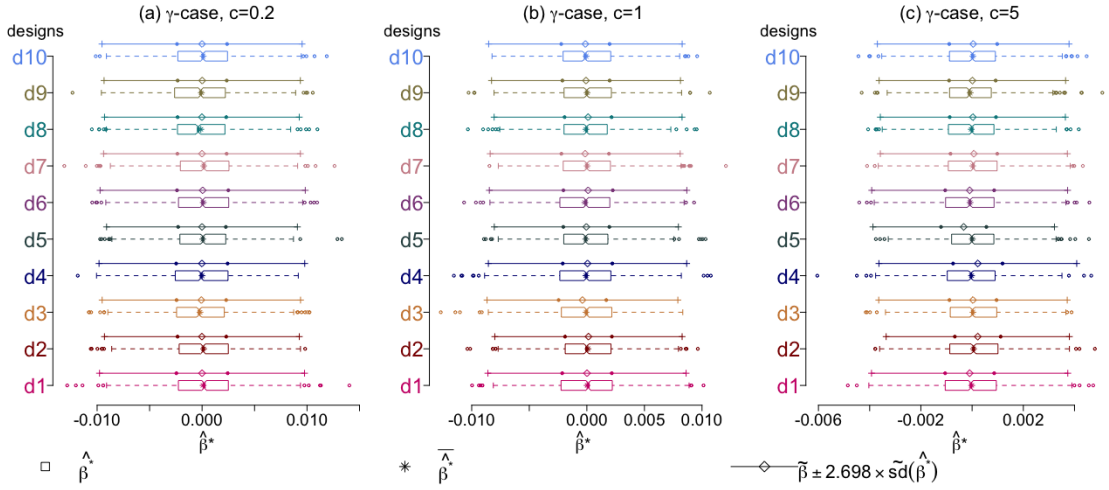


Figure 6.19: $\gamma-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\beta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

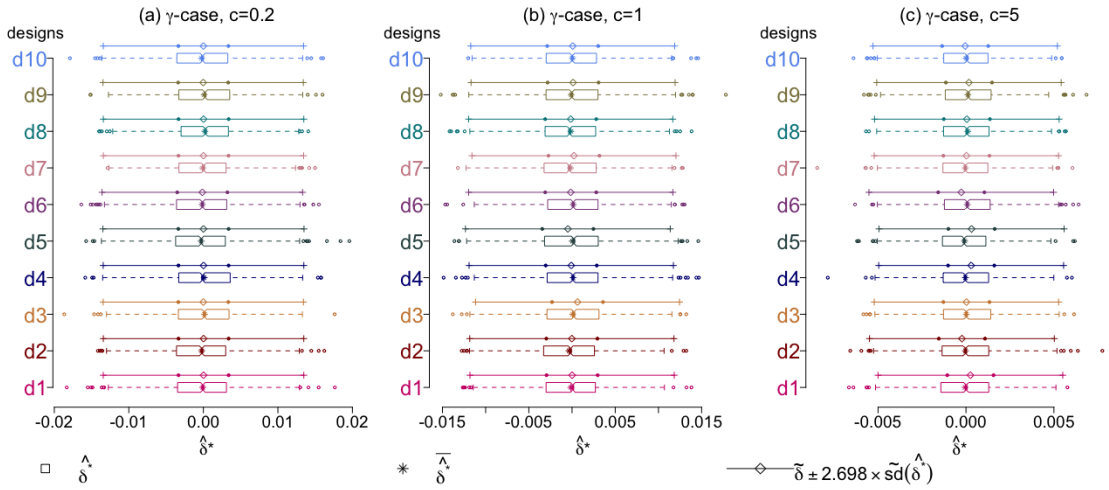


Figure 6.20: $\gamma-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\delta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$.

The $p-$values obtained from the AD and the KS tests support the idea that the empirical distribution is consistent with theoretical distributions as shown in **Figure 6.21**, except for two significant cases in $c5$ in $d4$ and $d10$ appear to counter this idea. However the

Chi-square qq-plot presented in Appendix H in **Figure H.4** and **Figure H.5** suggest that the $\hat{\underline{\theta}}^*$ and $\hat{\Delta}^*$ are close to the normal distribution.

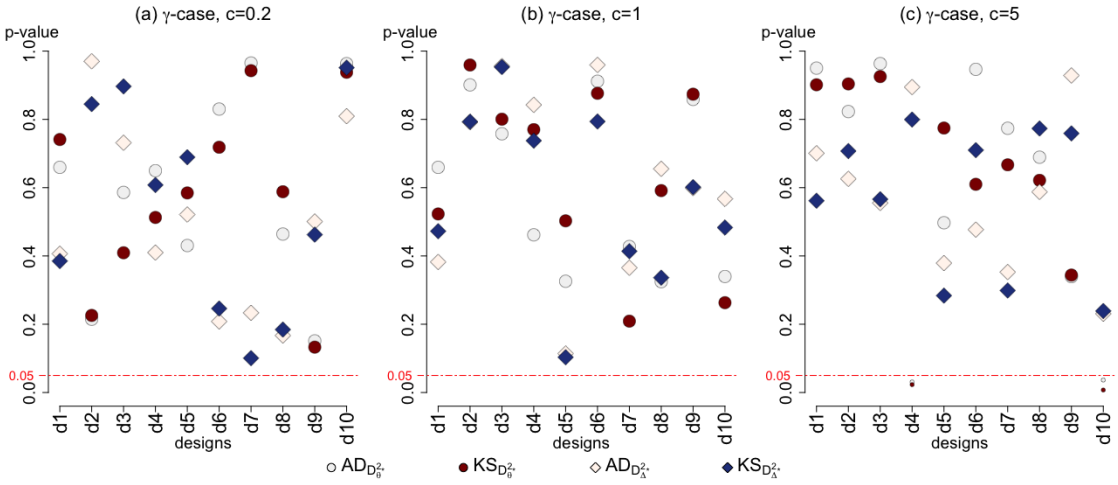


Figure 6.21: $\gamma-$case: $p-$values obtained from the Kolmogorov Smirnov (KS) and Anderson Darling (AD) tests of goodness of fit to test $\mathrm{H}_{0\theta^*} : D^2_{\theta^*} \sim \chi^2_2$ and $\mathrm{H}_{0\Delta^*} : D^2_{\Delta^*} \sim \chi^2_1$ for $c = 0.2$, $c = 1$ and $c = 5$.

6.4.4 Truth Parameters: $\delta-$case, $\beta-$case, $\gamma-$case, $c \in \{-0.2, -1, -5\}$

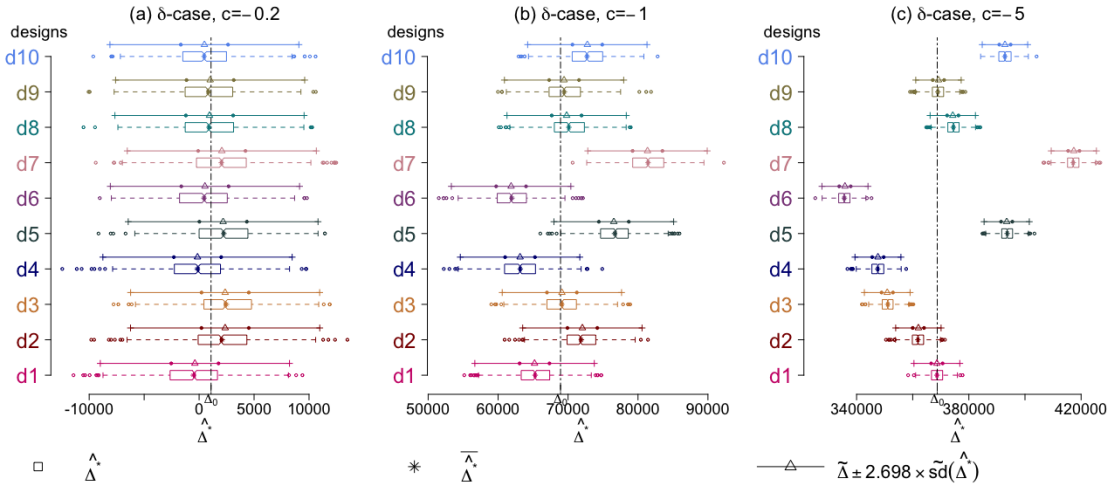
Consider the truth parameter vector θ_0 based on $\delta-$case, $\beta-$case, $\gamma-$case such that the non-zero parameter vector in each case δ_k, β_k and γ_k, respectively are attributed to negative values of c, i.e. $c \in \{-0.2, -1, -5\}$. Changing the direction of the strata's impact by using negative values of c in each truth case should not affect the earlier validated alignment between the theoretical distribution and empirical sampling distributions that were found when proposed positive c values were used. However, strata are not the same and results may not be equivalent to those obtained from using positive c. Distributions of the estimates of the applied model parameters obtained from the theory and simulation are exhibited in **Figure 6.22**, **Figure 6.23** and **Figure 6.24**, for all single truth cases across negative c.

Figure 6.22a provides the theoretical and empirical sampling distributions of the applied effect $\hat{\Delta}^*$ for $\delta-$case across negative c. From the presented distributions in this figure, it is apparent that the behaviour of all the sample distributions are aligned with their corresponding theoretical distributions. This result matches those observed earlier in **Figure 6.5** and **Figure 6.4** when positive c were used. This figure also shows that the range of $\hat{\Delta}^*$ increases as the difference between strata gets larger, which is in accord with

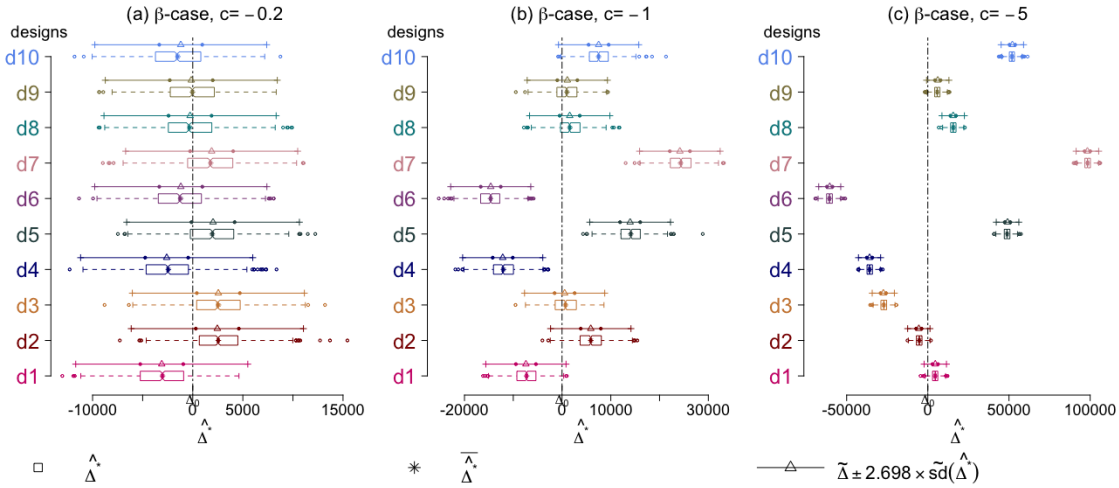
our earlier observations. However, the range of $\hat{\Delta}^*$ in $c = -0.2$ and $c = -1$ is shifted to the left compared to the distributions made by $c = 0.2$ and $c = 1$. The decrease in the overall campaign effects is observable in these two instances, where the true effect Δ_0 decreases by more than 5000 units in $c = -0.2$ compared to the value obtained by using $c = 0.2$, and decreases by more than 20000 units in $c = -1$ compared to the value obtained by using $c = 1$. Focusing on the first design in these two c values, it is possible to see that the distributions of $\hat{\Delta}^*$ are shifted sharply to the left; put them to the left of their truth instance's true effect true effect Δ_0. In comparison with distributions of $d1$ in $c = 0.2$ and $c = 1$, they are at the opposite direction relative to Δ_0. This result might be related to the impact of the strata relative information in the treatment spatial units on the stratified parameters, such as relative search proportions. When $c = -5$, on the other hand, the range of $\hat{\Delta}^*$ is shifted to the right for all designs and Δ_0 is a bit larger compared to distributions given by $c = 5$.

When truth parameters are $\beta-$case based, the theoretical and empirical sampling distributions of $\hat{\Delta}^*$ listed in **Figure 6.22b** are aligned parallel to each other, for the three negative c values. In accordance with earlier observations in using positive c, there is an increment in the range of $\hat{\Delta}^*$ as the difference between strata widens but not when truth parameters are $\gamma-$case based, see **Figure 6.22c**. By comparing the range $\hat{\Delta}^*$ in $\beta-$case here to those resulted by positive c in **Figure 6.11**, it can be observed that the negative limits of the applied effects obtained by the latter are cut here and hence the range decreases. The change in the range be seen clearly in the distributions from designs $d2$ and $d3$ where both shift to the right. Paying particular attention to the first four designs over all three sub-figures, the change in the range brings the variation between those designs down.

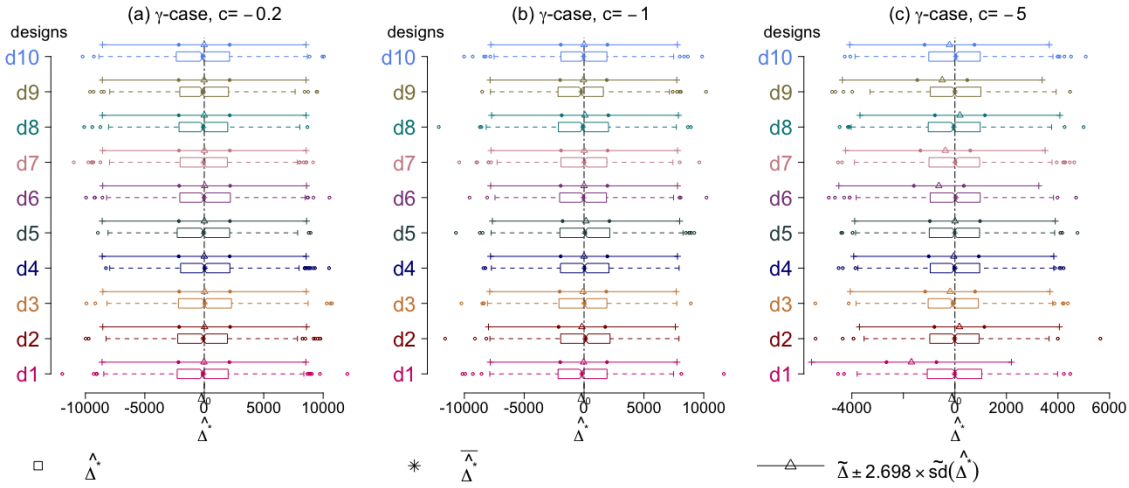
In $\gamma-$case using $c = 1$ or $c = -1$, the range and the distributions behaviour of $\hat{\Delta}^*$ for all designs do not change, and the same can be observed when using $c = \pm 0.2$. Using $c = -5$ makes no change too in the empirical sampling distributions of $\hat{\Delta}^*$ in all designs compared with those drawn by $c = 5$. However, using $c = -5$ makes an impact on the theoretical distributions of $\hat{\Delta}^*$ in almost all presented designs and clearly obvious in $d1$ as well as $d6$, $d7$, $d8$ and $d9$, where their ranges moved to the left causing mismatch with their sampling distributions.



(a) $\delta-$case

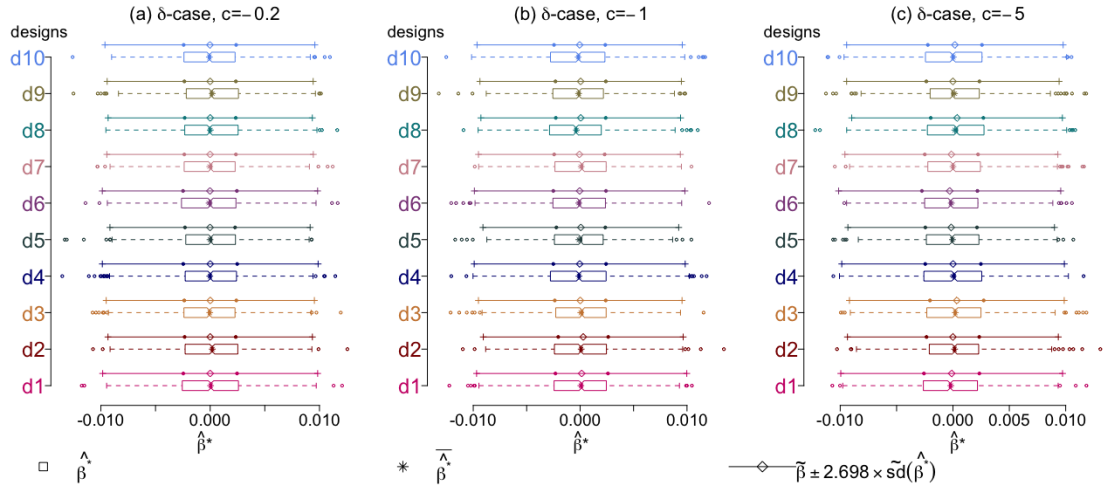


(b) $\beta-$case

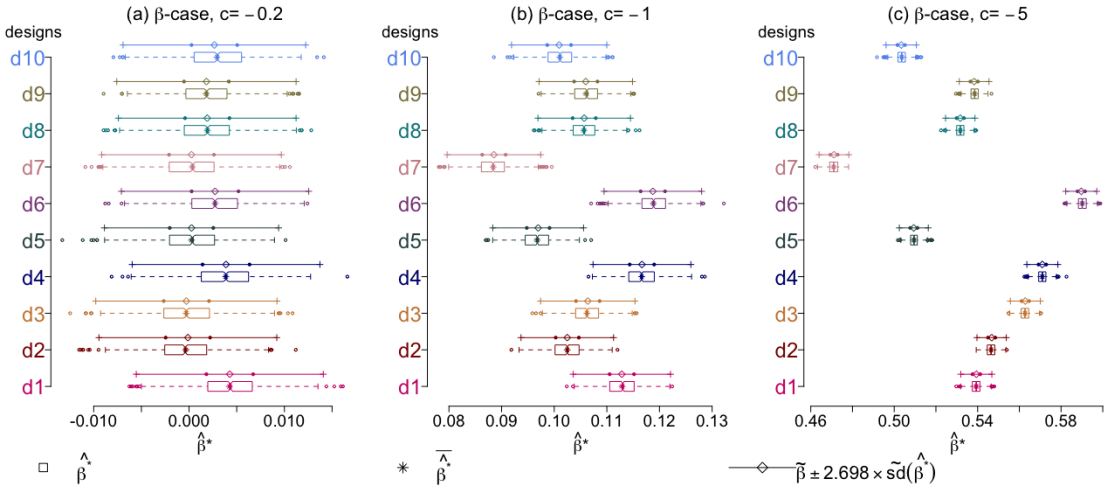


(c) $\gamma-$case

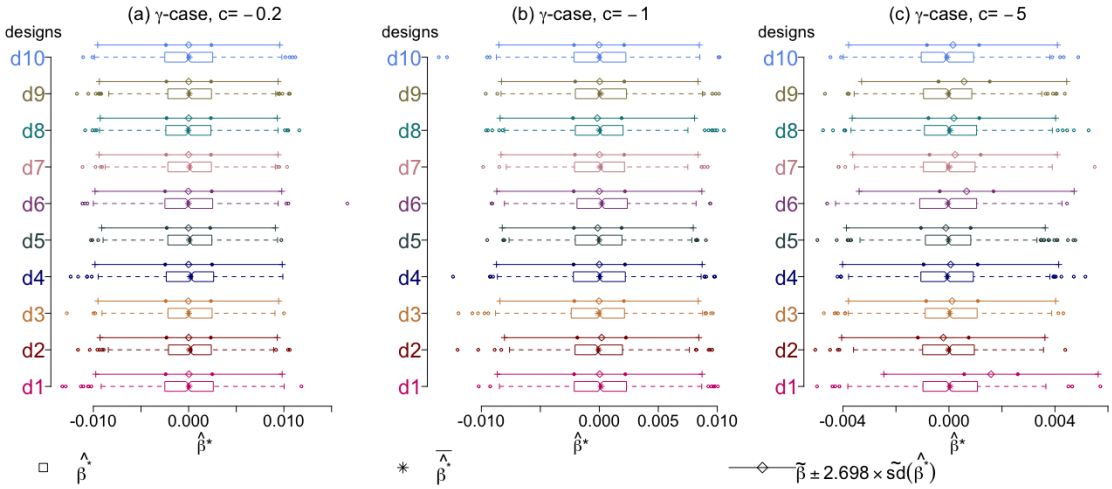
Figure 6.22: All single cases; $\delta-$case, $\beta-$case and $\gamma-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$ for $c = -0.2$, $c = -1$ and $c = -5$



(a) $\delta-$case

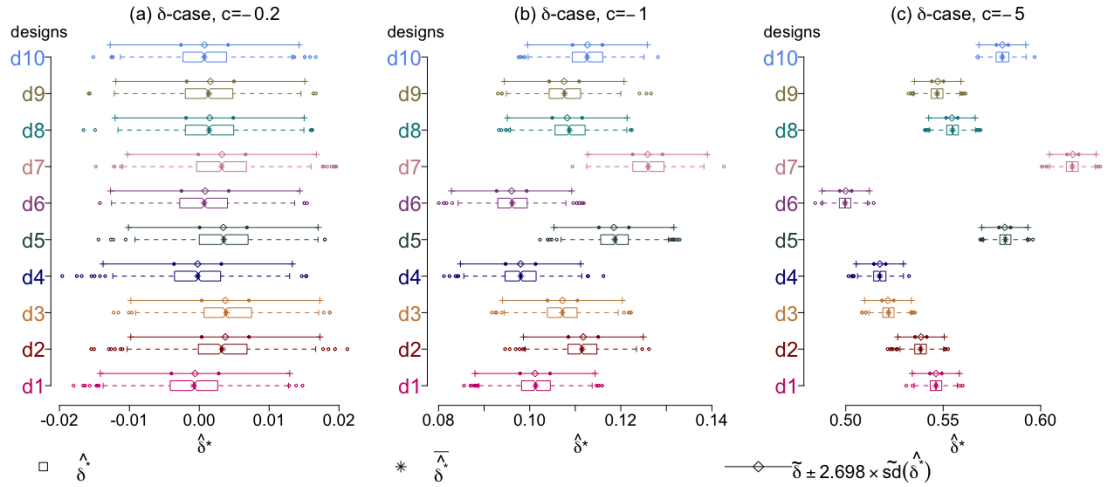


(b) $\beta-$case

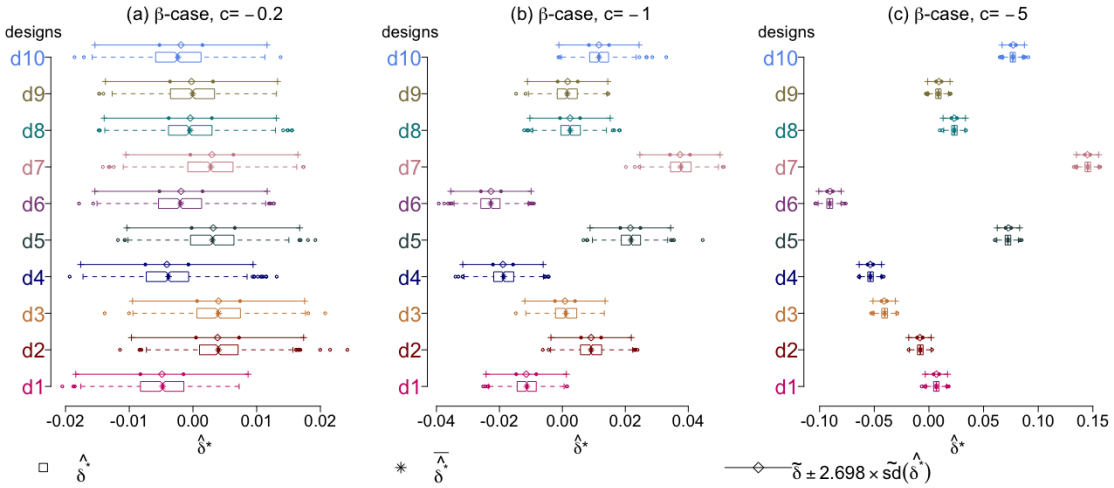


(c) $\gamma-$case

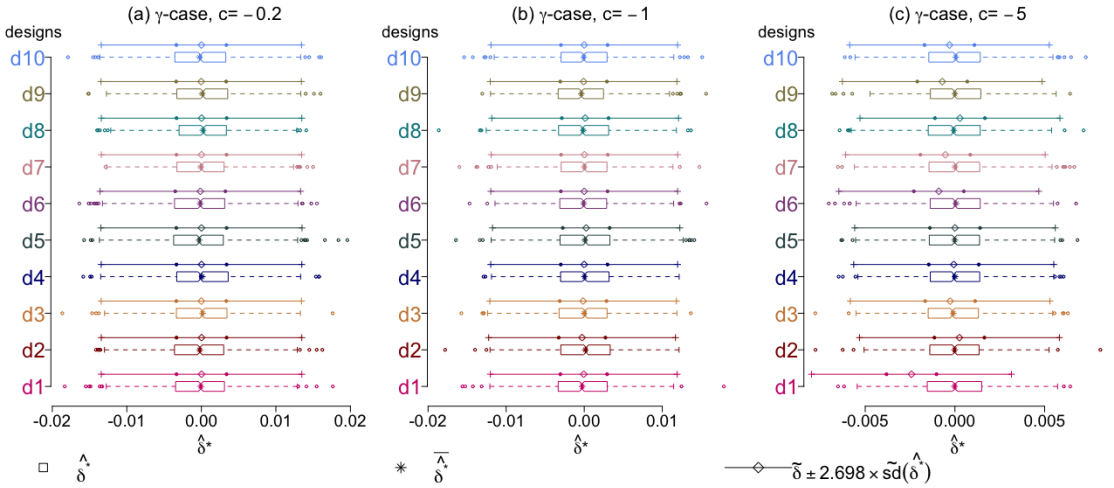
Figure 6.23: All single cases; $\delta-$case, $\beta-$case and $\gamma-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\beta}^*$ for $c = -0.2$, $c = -1$ and $c = -5$



(a) $\delta-$case



(b) $\beta-$case



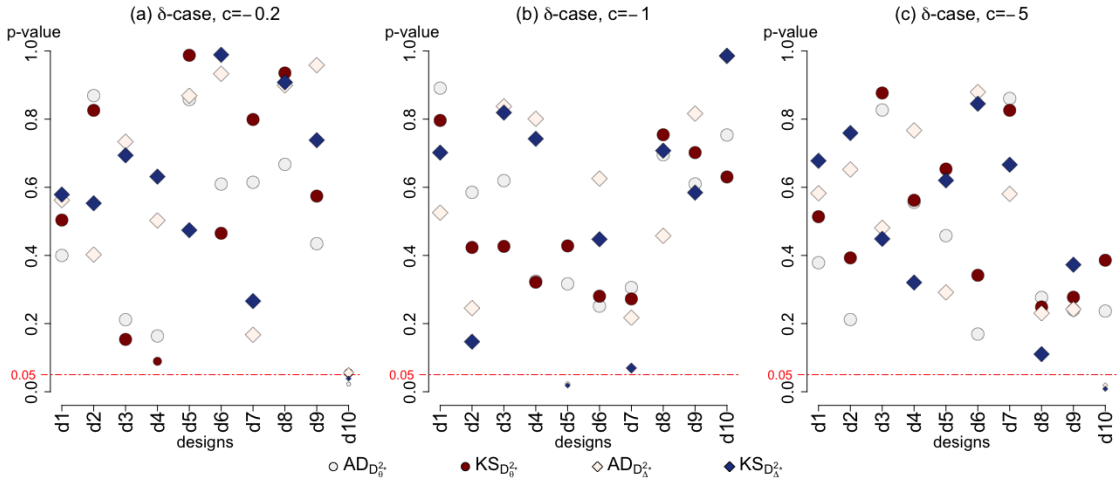
(c) $\gamma-$case

Figure 6.24: All single cases; $\delta-$case, $\beta-$case and $\gamma-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\delta}^*$ for $c = -0.2$, $c = -1$ and $c = -5$

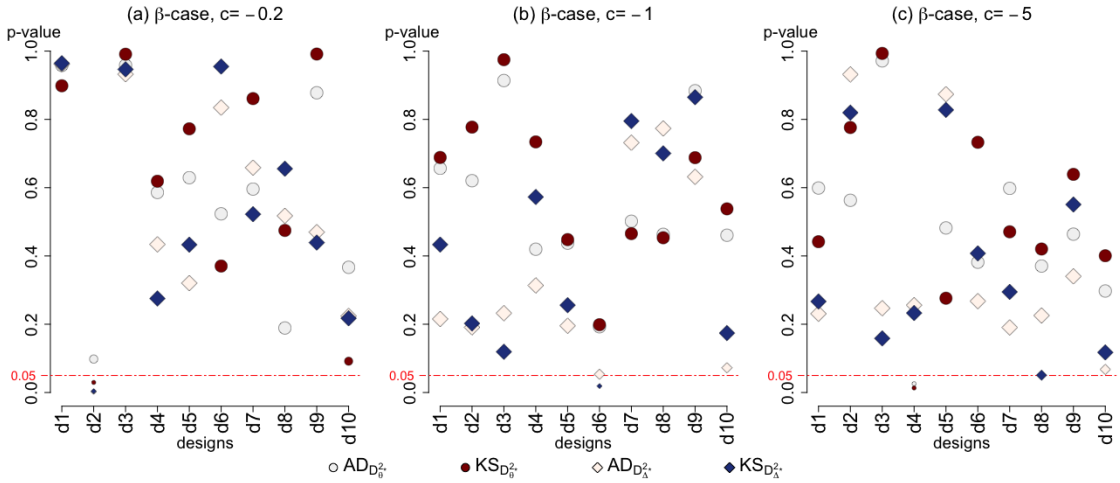
The theoretical behaviour of $\hat{\underline{\theta}}^*$; $\hat{\beta}^*$ and $\hat{\delta}^*$ in all truth cases including positive or negative c except $\gamma-$case with $c = -5$ conform to their empirical sampling distributions. The lack of agreement between the theory and the simulation in $\gamma-$case with $c = -5$ appears again very clear in $d1$ as shown in **Figure 6.23c** and **Figure 6.24c**. Not to mention, changing the sign of c makes no difference in the range of $\hat{\beta}^*$ and range $\hat{\delta}^*$ in all single truth instances, except $\gamma-$case with $c = -5$ in **Figure 6.23** and **Figure 6.24**. Notwithstanding the similarity in the resulting range between negative and positive c, the behaviours of the presented distributions in using negative c are not the same as those presented in the positive. It is not easy to explain the change in a distribution behaviour in each design, but it might be related to the impact of proportions of search and purchases in each strata in treatment spatial units. There are, however, other possible explanations, and further work needs to be done to investigate and explain the behaviour of each design. In this investigation, the aim was to assess the ability of the theory to describe the asymptotic sampling distribution of the estimates of the applied model parameters.

P-values for the AD and the KS tests for the three truth cases across negative c values are illustrated in **Figure 6.25**. Once more $\gamma-$case with $c = -5$, especially in $d1$, $d6$ and $d8$ display the most striking significant results, indicating unusual difference between the distributions behaviour of $\hat{\Delta}^*$ or $\hat{\underline{\theta}}^*$ obtained by the theory and the simulation. Focusing on this truth instance, we present the Chi-square qq-plots in **Figure 6.26** so as to assess if the applied estimates possibly come from the theoretical Normal distribution. The qq-plots of design $d1$ clearly indicate the deviation from the straight line and hence demonstrate the non theoretical normality of the estimates in this design.

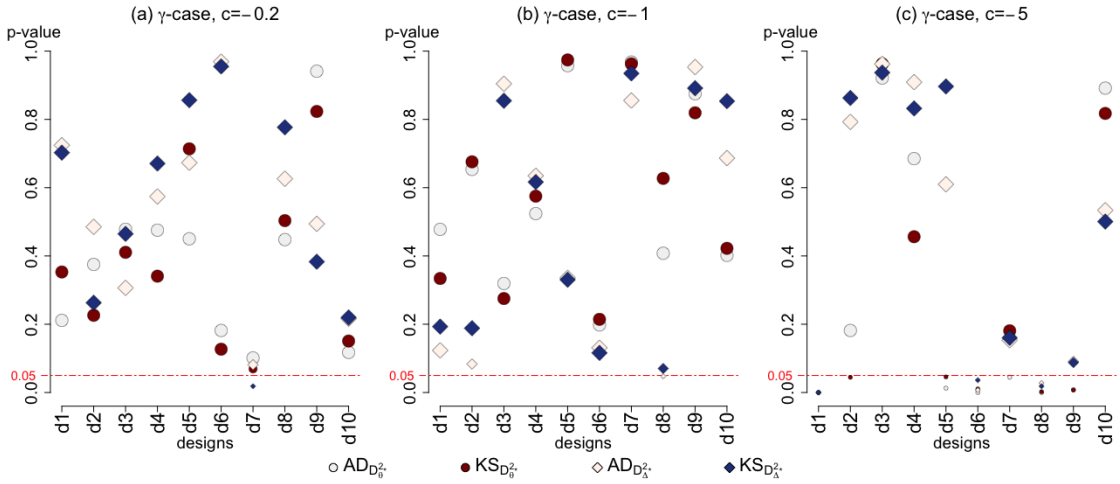
Consider $d1$ in this truth circumstance, the empirical distributions of $\hat{\beta}^*$, $\hat{\delta}^*$ and $\hat{\Delta}^*$ are presented as histogram plots in **Figure 6.28**. From the figure it can be seen that the three samples are approximately bell shaped. The red normal curve over the histograms are the theoretical distributions of the estimates. The peaks in all three curves disagree with the center of the bell-shaped histograms. On the same figure, the normal Q-Q plots are shown. The points in the three estimates form a line that are nearly straight regardless the diversion at extreme ends, indicating that the distribution of the three estimates are not far from the Normal distribution.



(a) $\delta-$case

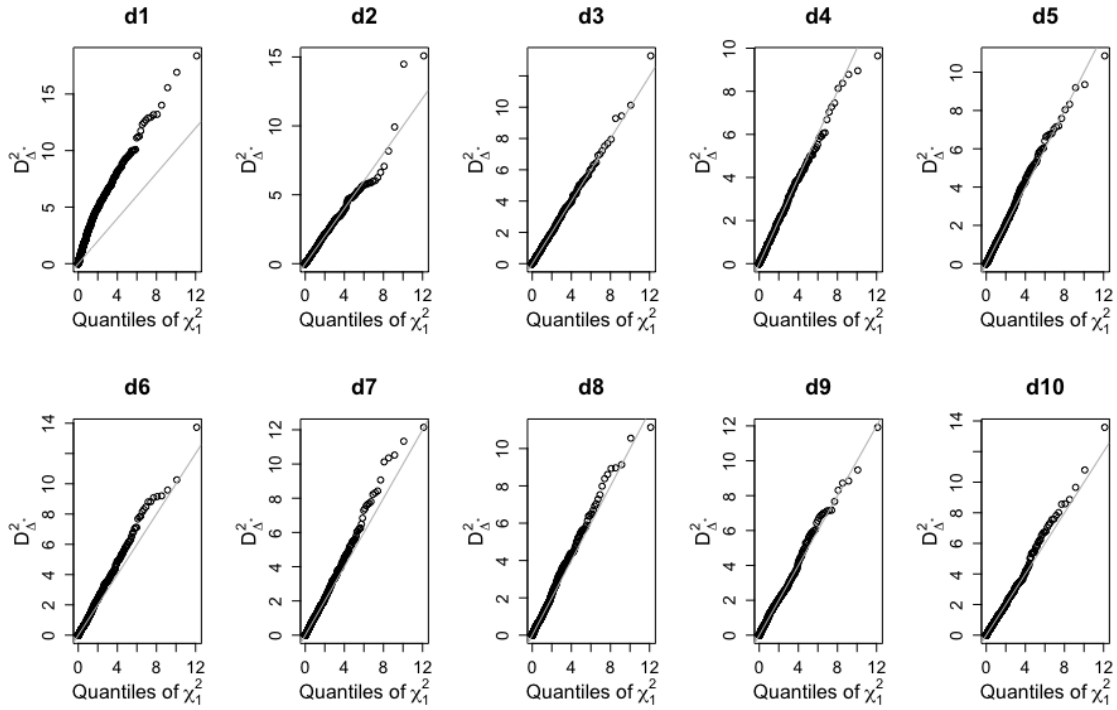


(b) $\beta-$case

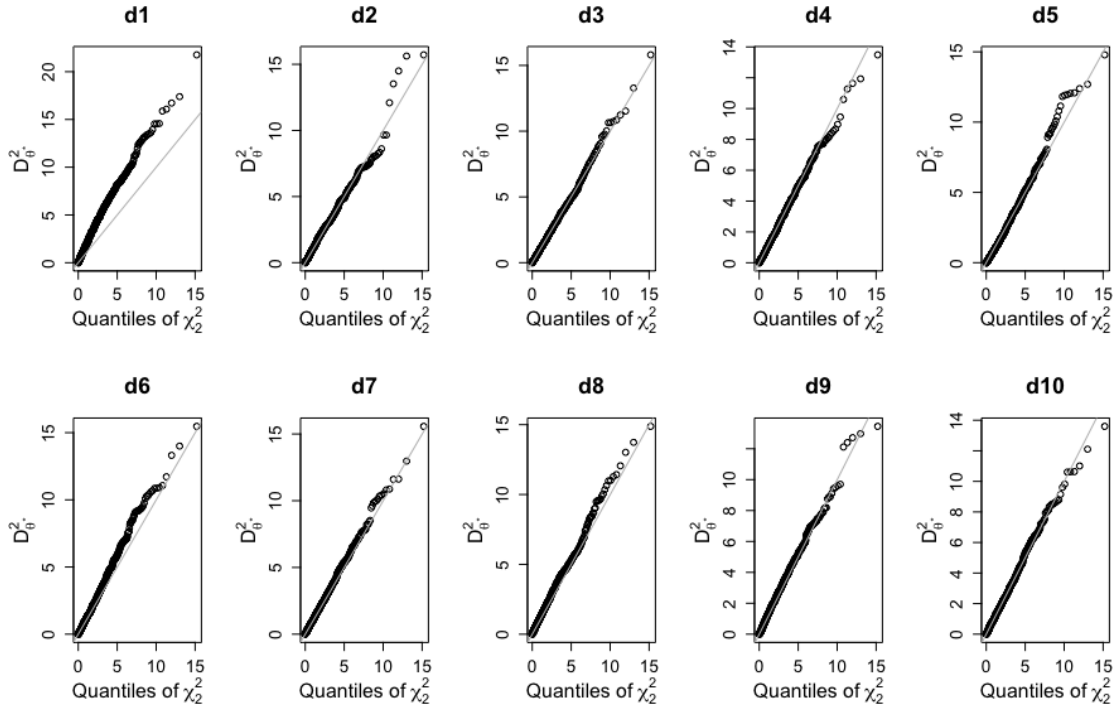


(c) $\gamma-$case

Figure 6.25: All single cases; $\delta-$case, $\beta-$case and $\gamma-$case: $p-$values obtained from the Kolmogorov Smirnov (KS) and the Anderson Darling (AD) tests of goodness of fit to test $H_{0\theta^*} : D^2_{\theta^*} \sim \chi^2_2$ and $H_{0\Delta^*} : D^2_{\Delta^*} \sim \chi^2_1$ for $c = -0.2$, $c = -1$ and $c = -5$



(a) Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ versus quantiles of χ^2_1



(b) Chi-square Q-Q plot: Mahalanobis distances $D^2_{\theta*}$ versus quantiles of χ^2_2

Figure 6.26: $\gamma-$case with $c = -5$: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ and $D^2_{\theta*}$ versus quantiles of χ^2_1 and χ^2_2, respectively.

In addition, SW test provides insignificant $p-$values 0.2499, 0.7273 and 0.2495 for testing the normality of the underlying distribution of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ respectively, suggesting that all three distributions are Normal. The KS test provides significant $p-$values less than 2.2×-16 indicating biased means in the sampling distribution of these three estimates. The standardised distributions of the three estimates are shown in **Figure 6.27** with the standard Normal curve, to examine the dissimilarity between their theoretical and empirical distributions. Again the three distributions deviate from the standard Normal. The standard deviation of the standardised $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ are about 1, in specific 1.02 0.99 and 1.02, respectively. Their means, on the other hand are not zero; 1.16, -1.03 and 1.16, respectively. The peak values of both $\hat{\delta}^*$ and $\hat{\Delta}^*$ are above the zero by about one unit of the standard deviation and $\hat{\beta}^*$ is below the zero by the about the same unit of the standard deviation.

In contrast to earlier findings, the presented assessment suggests that this truth instance shows unusual difference between the theoretical asymptotic distributions of the estimates and their empirical distribution. This disagreement could be attributed to unrealistically large difference between strata in an extreme truth case with zero true temporal effect and zero true campaign effect. The disagreement though has not been shown in $\gamma-$case in $c = 5$, so this may also be related to number of searches or purchases in the different strata.

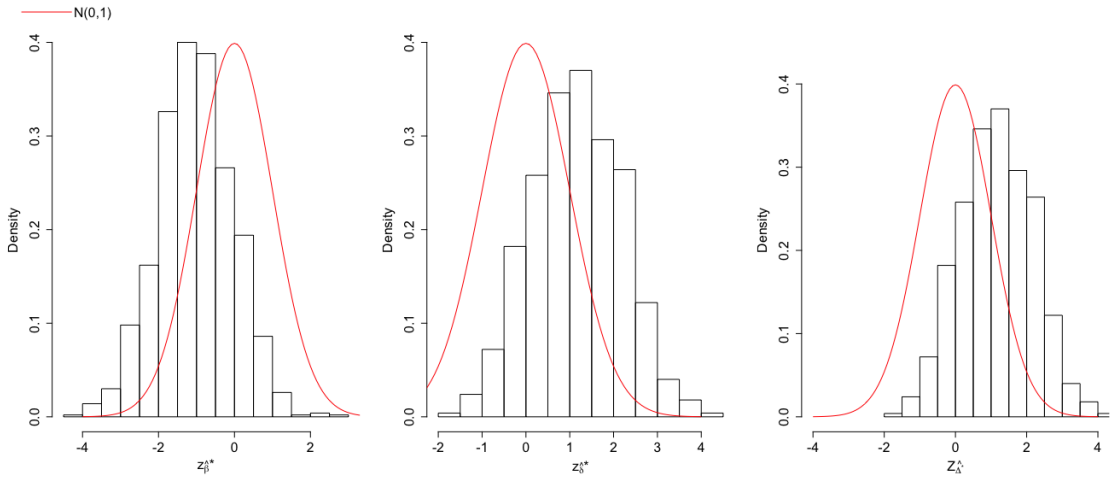
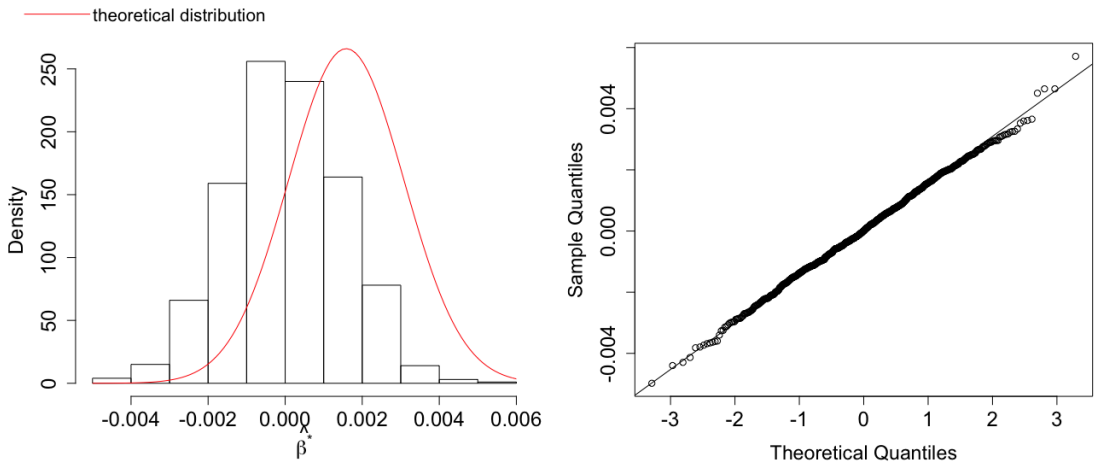
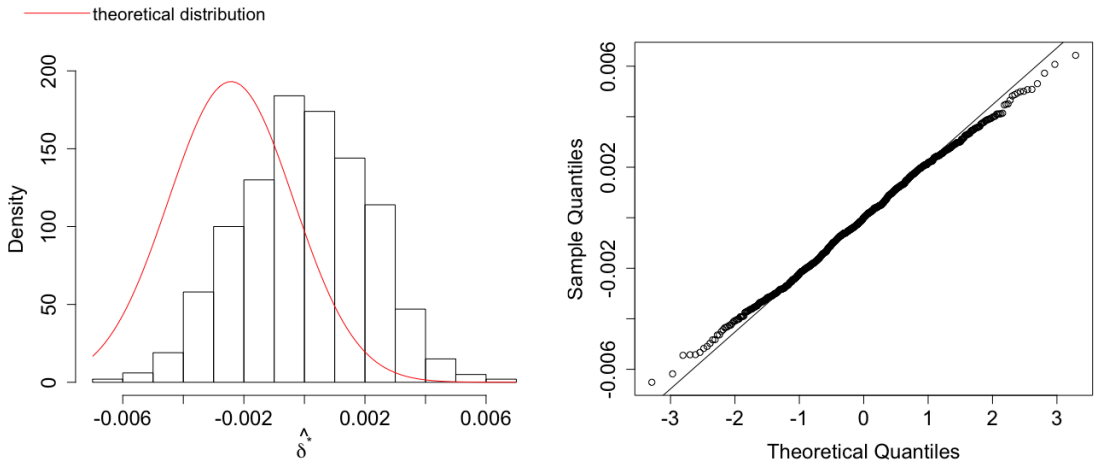


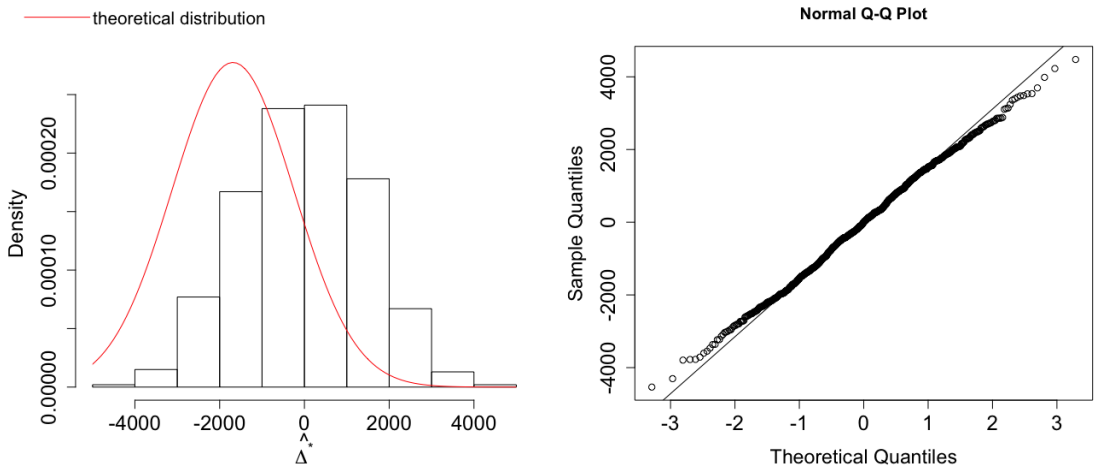
Figure 6.27: $\gamma-$case with $c = -5$: Standardised empirical distributions of $\hat{\beta}^*$, $\hat{\delta}^*$ and $\hat{\Delta}^*$ of $d1$.



(a) Histogram of the empirical distribution of $\hat{\beta}^*$ and its Normal Q-Q plot.



(b) Histogram of the empirical distribution of $\hat{\delta}^*$ and its Normal Q-Q plot.



(c) Histogram of the empirical distribution of $\hat{\Delta}^*$ and its Normal Q-Q plot.

Figure 6.28: $\gamma-$case with $c = -5$: Graphical assessments of Normality of $\hat{\beta}^*$, $\hat{\delta}^*$ and $\hat{\Delta}^*$ of $d1$.

6.4.5 Truth Parameters: $\beta\delta-$case, $c_1 = c_2 \in \{\pm 0.2, \pm 1, \pm 5\}$

Consider the truth parameter vector θ_0 based on $\beta\delta-$case with two non-zero parameter vectors β_k and δ_k. **Figure 6.29** and **Figure 6.30** show the theoretical and empirical distributions of the three estimates $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ using positive and negative c. All plots presented in the two figures appear to indicate no apparent difference between the theoretical and the empirical distributions in all 10 designs except $d10$ in $c_1 = c_2 = -1$, where the theoretical distribution of $\hat{\Delta}^*$ and $\hat{\delta}^*$ are slightly to the left of their corresponding empirical distributions.

The true effect Δ_0 is non-zero here as in $\delta-$case since $\delta_k \neq 0$. Focusing on $c_1 = c_2 = \pm 5$, the values of $\hat{\delta}^*$ and Δ_0 are less than those obtained in $\delta-$case in $c = \pm 5$. The remarkable result to emerge from these two instances that all 10 distributions of $\hat{\Delta}^*$ are to the left of the true effect indicate negative bias in $\hat{\Delta}^*$. The range of $\hat{\beta}^*$ and $\hat{\delta}^*$ are somewhat similar to the range resulting in $\beta-$case. The behaviour of the two estimates across designs are opposite to each other, whereas the behaviour of $\hat{\delta}^*$ and $\hat{\Delta}^*$ across designs gets along each other. These two behaviours match what observed in earlier truth cases.

The $p-$values form the AD and the KS tests are presented in **Figure 6.31**. The values present evidence at $d9$ in $c_1 = c_2 = 1$ and $d10$ in $c_1 = c_2 = -1$ against the claim that $\hat{\underline{\theta}}^*$ and $\hat{\Delta}^*$ are coming from the theoretical distribution. The Chi-square qq-plots are presented in Appendix H in **Figure H.6** and **Figure H.7**, respectively to assess normality of $\hat{\underline{\theta}}^*$ and $\hat{\Delta}^*$. The figures support normality claims in general except designs $d10$ in $c_1 = c_2 = -1$ provide rejection evidences, where we can see the points are not exactly on the straight line.

Focusing on $d10$ in $c_1 = c_2 = -1$, SW test provides insignificant $p-$values 0.6586, 0.8559 and 0.6229 for testing the normality of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ respectively, suggesting that the three distributions are Normal. However, KS test suggests they are not coming from the theoretical distribution where $p-$values are 7.563×10^{-13}, 3.195×10^{-4} and 3.252×10^{-12}, respectively. The histograms of their standardised distributions are shown in **Figure 6.32** overlaid standard Normal curve, to check how far these distributions are from the theoretical distributions. The peaked values in the three distributions are not far away from the standard normal peak. Thus, we can still acknowledge the importance of the theoretical framework in estimating the asymptotic distributions of the estimates.

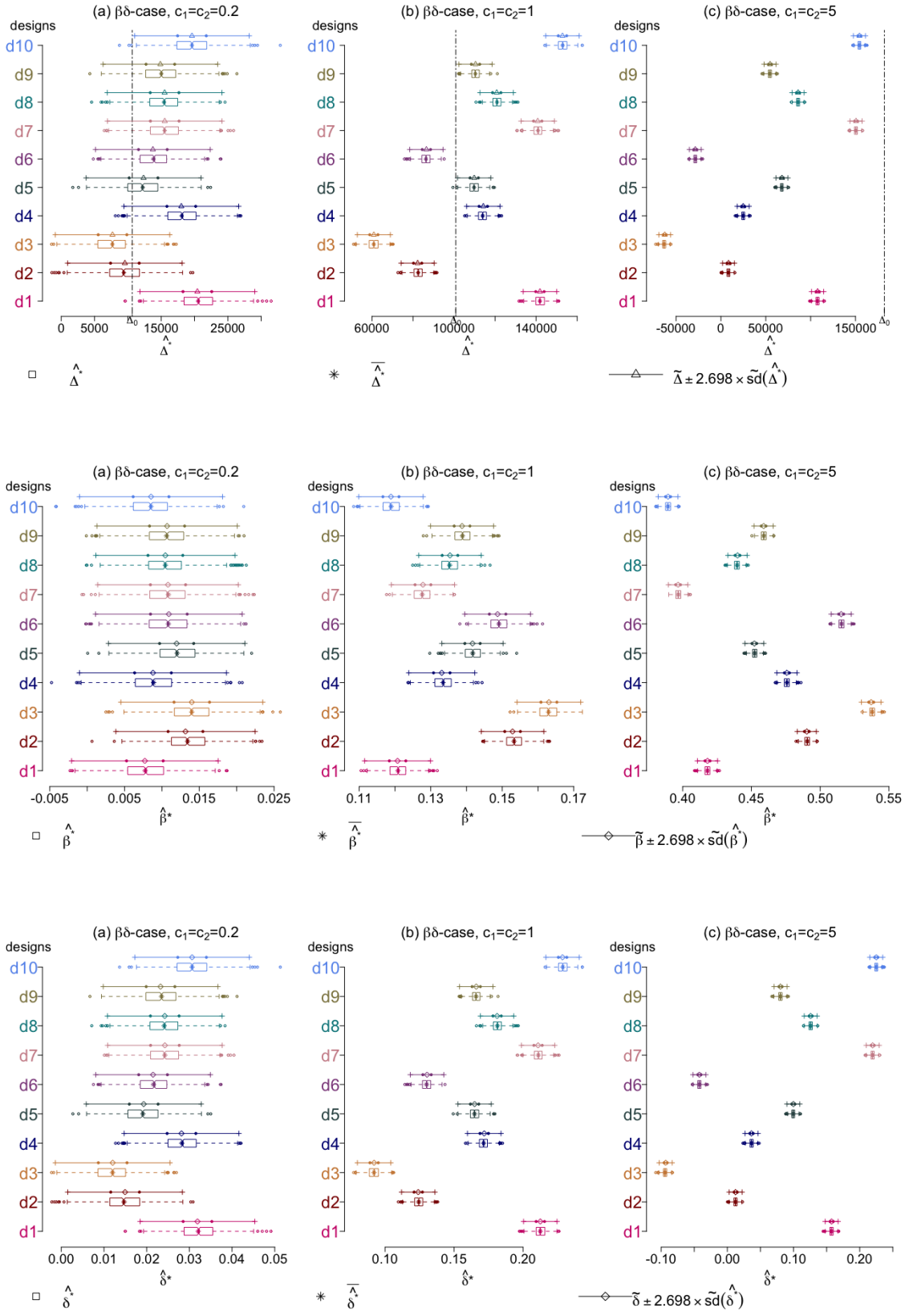


Figure 6.29: $\beta\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ for $c = 0.2$, $c = 1$ and $c = 5$

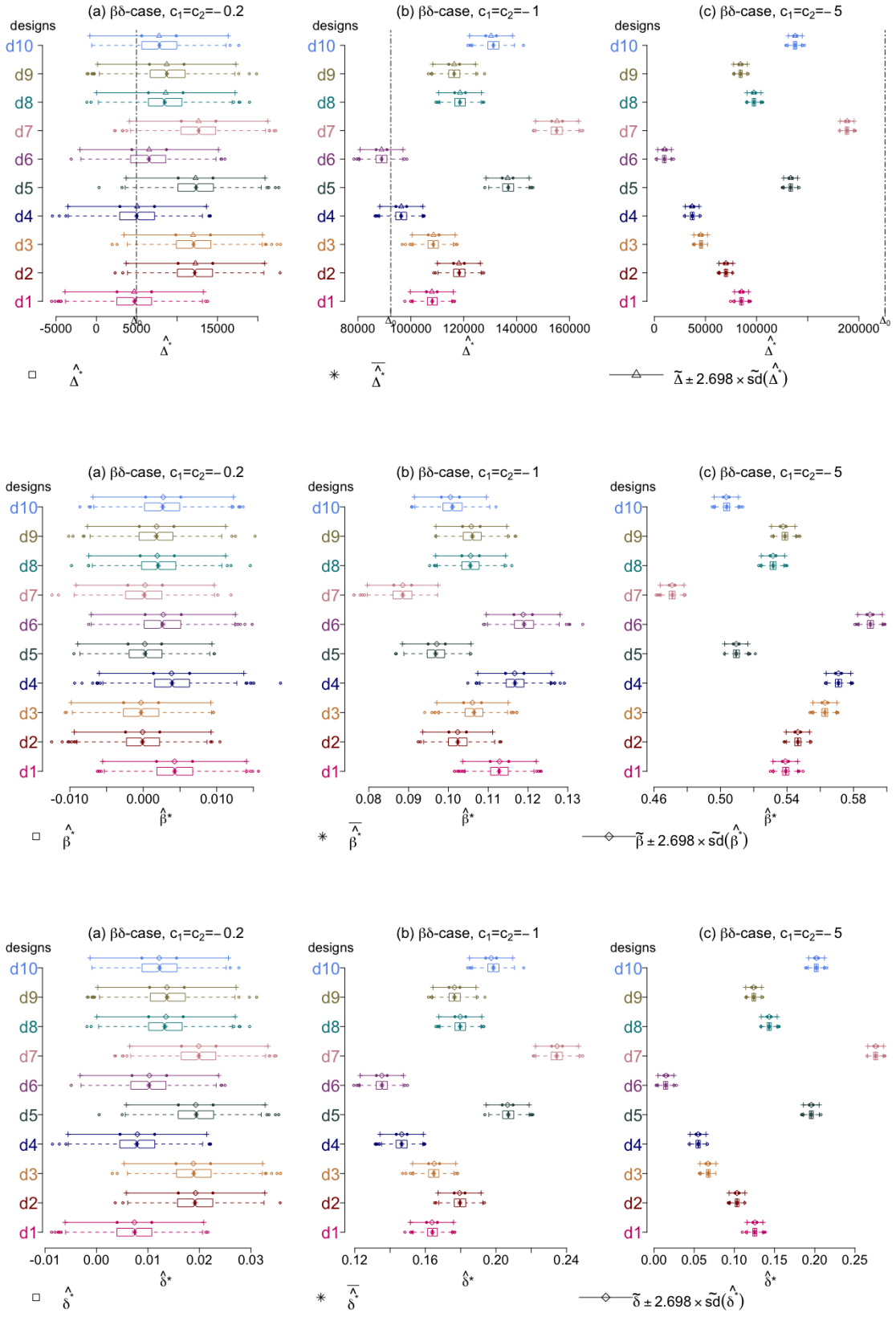


Figure 6.30: $\beta\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ for $c_1 = c_2 = -0.2$, $c_1 = c_2 = -1$ and $c_1 = c_2 = -5$

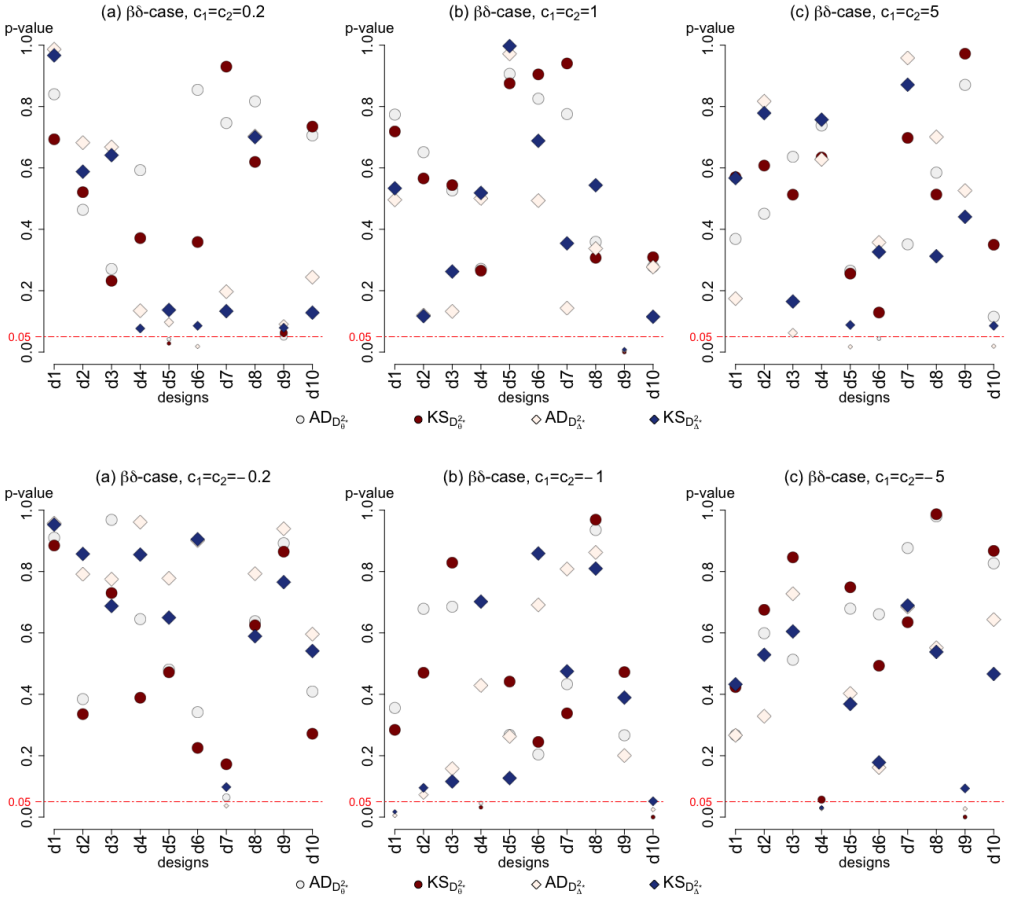


Figure 6.31: $\beta\delta-$case: $p-$values obtained from the Kolmogorov Smirnov (KS) and Anderson Darling (AD) tests of goodness of fit to test $\mathrm{H}_{0\theta^*} : D^2_{\theta^*} \sim \chi^2_2$ and $\mathrm{H}_{0\Delta^*} : D^2_{\Delta^*} \sim \chi^2_1$ for $c_1 = c_2 = \pm 0.2$, $c_1 = c_2 = \pm 1$ and $c_1 = c_2 = \pm 5$.

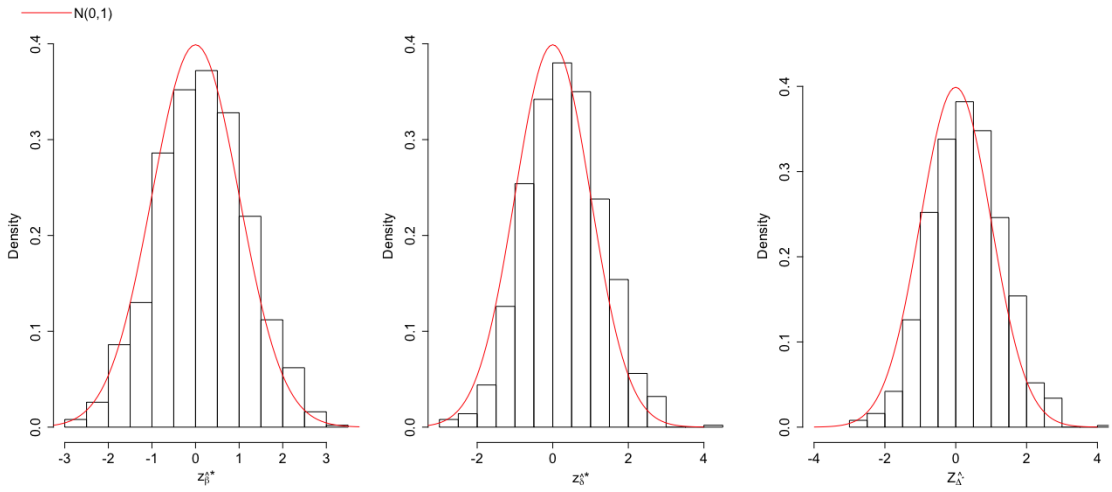


Figure 6.32: $\beta\delta-$case with $c = -1$: Standardised empirical distributions of $\hat{\beta}^*$, $\hat{\delta}^*$ and $\hat{\Delta}^*$ of $d10$.

6.4.6 Truth Parameters: $\gamma\beta\delta-$case, $c_1, c_2, c_3 \in \{1, -1\}$

Consider the truth parameter vector θ_0 based on $\gamma\beta\delta-$case where the three parameters γ_k, β_k and δ_k are non-zero. The difference between strata are considered to be based on $c_1, c_2, c_3 \in \{1, -1\}$ because we believe these values make the difference between strata plausible. Eight sets of combinations of c_1, c_2, c_3 are formed from these two values, i.e. ± 1. The theoretical and empirical distributions of the three estimates $\hat{\Delta}^*$, $\hat{\delta}^*$ and $\hat{\beta}^*$ are displayed in **Figure 6.33**, **Figure 6.34** and **Figure 6.35**.

The ranges of the three estimates vary from one truth instance to another. The true effect also varies substantially from one truth instance to another, being estimated least accurately in the cases where the coefficients c_2 and c_3 used for γ_k and δ_k have the same sign and the coefficient c_1 used for γ_k has the opposite sign. It is not clear why this should happen. **Figure 6.33** and **Figure 6.34** show that the values of $\hat{\Delta}^*$ and $\hat{\delta}^*$ are negative when c_1 and c_2 have the same sign and c_3 has the opposite sign, i.e. $c_1 = c_2 = 1, c_3 = -1$ and $c_1 = c_2 = -1, c_3 = 1$, and are otherwise positive. Again, it is not clear why this should be the case. **Figure 6.35** shows the values of $\hat{\beta}^*$ are positive when c_1 and c_2 have the same sign and negative when they have opposite signs. Again there is no obvious explanation but taken together these results show that an unobserved covariate has the potential to have unpredictable effects on parameter estimation and in particular that dependence of one parameter on a covariate can have implications for estimation of other parameters.

The non-zero bias in $\hat{\Delta}^*$ is clear in this truth case too, especially when the signs of c_2 and c_3 are different from the sign of c_1 in **Figure 6.33 (d),(e), (g)**. The behaviour of $\hat{\Delta}^*$ and $\hat{\delta}^*$ across designs are almost the same. The behaviour of $\hat{\beta}^*$ across designs on the other hand opposes the behaviour of the other two estimates.

The empirical distributions of $\hat{\Delta}^*$ and $\hat{\delta}^*$ are in agreement with their corresponding theoretical distributions across the eights sets of the truth combinations. The empirical distributions of $\hat{\beta}^*$ are also in accord with theoretical distributions despite the slight divergence at design $d4$ in truth set $c_1 = c_2 = c_3 = 1$ in **Figure 6.35 (a)**. In this truth version at $d4$, there is some evidence that the distributions of $\hat{\beta}^*$ differ but it is a small difference compared to the variation in outcome between the different designs.

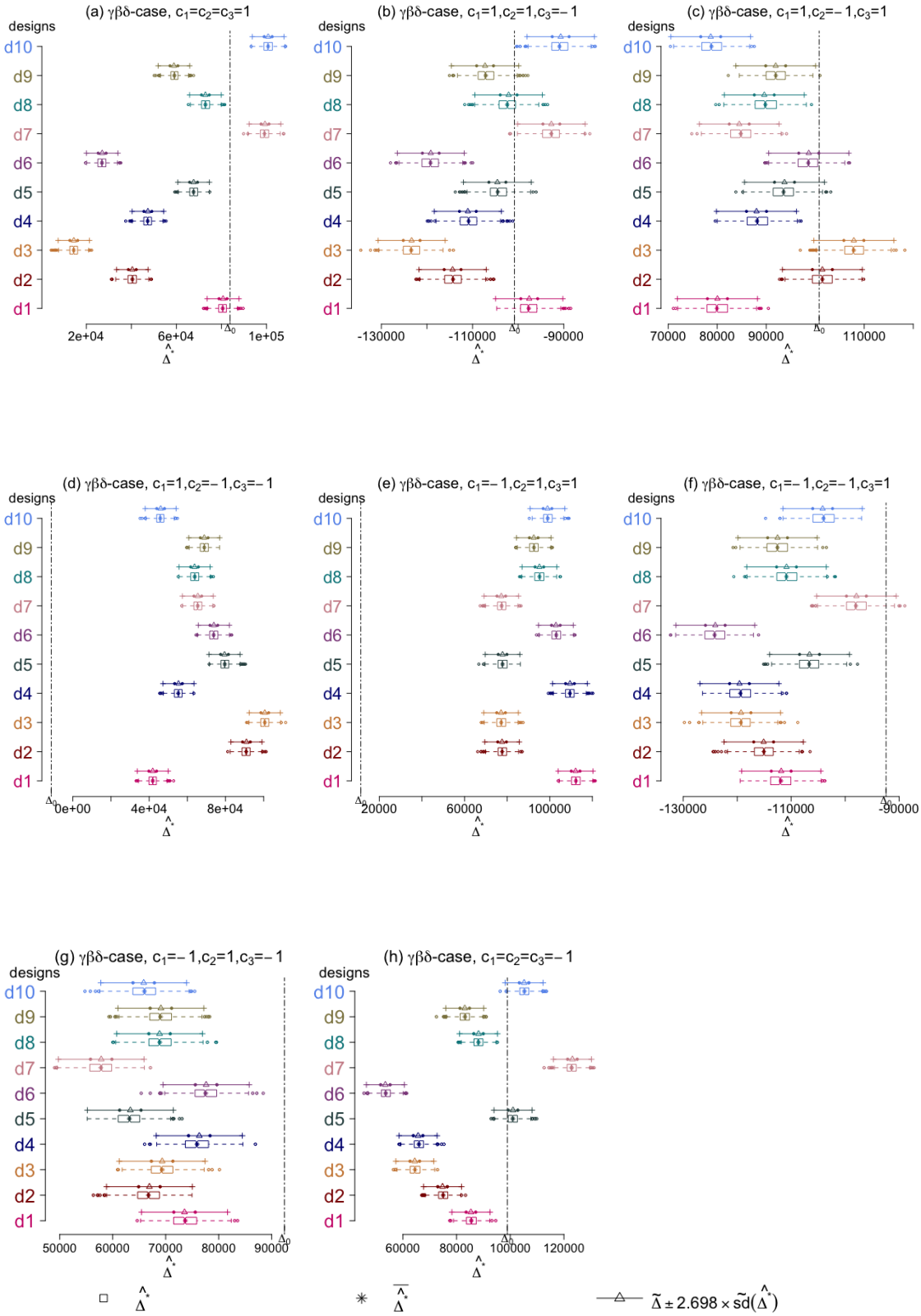


Figure 6.33: $\gamma\beta\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$ for combinations of c_1, c_2 and c_3 in the set $\{1, -1\}$

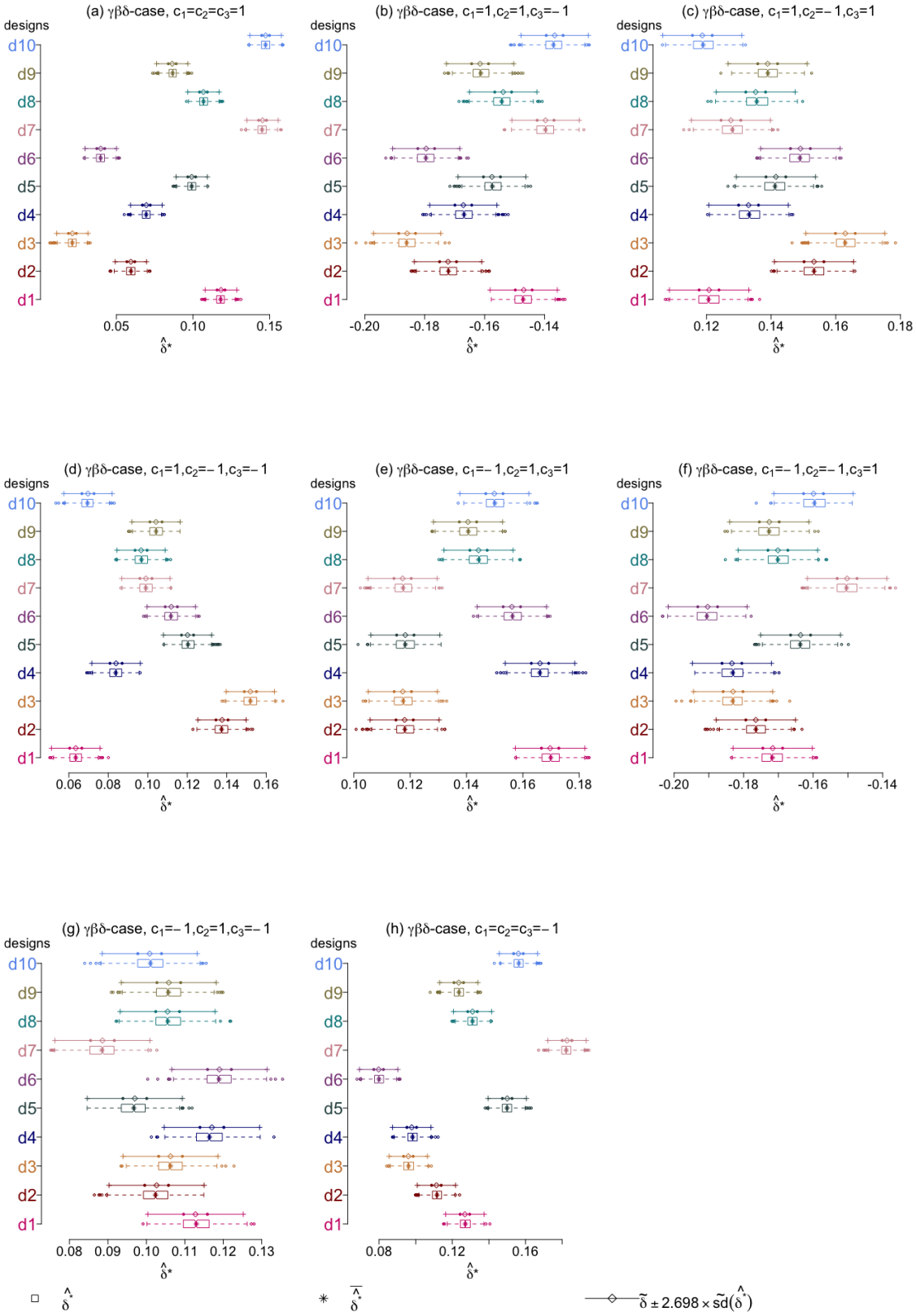


Figure 6.34: $\gamma\beta\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\delta}^*$ for combinations of c_1, c_2 and c_3 in the set $\{1, -1\}$

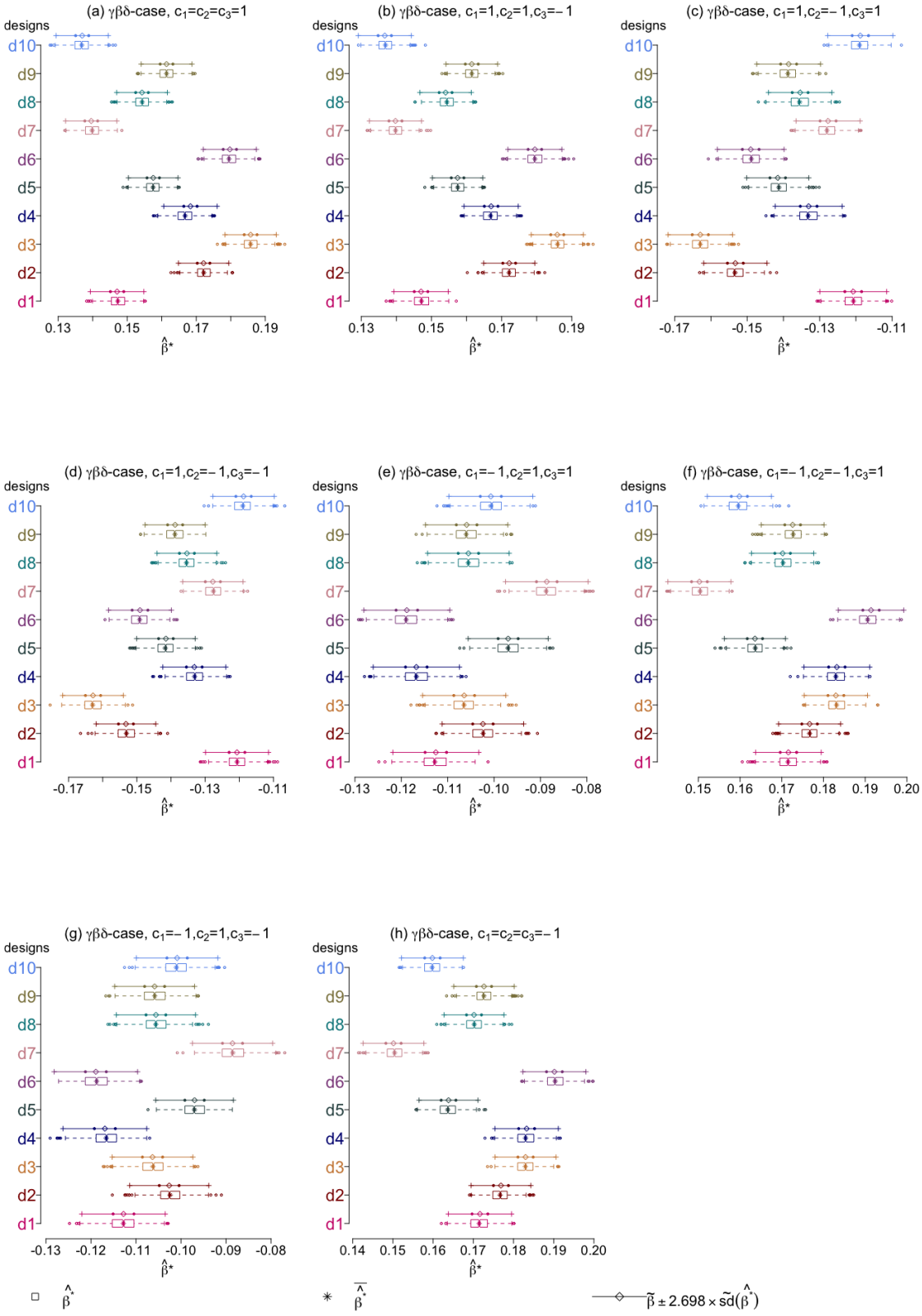


Figure 6.35: $\gamma\beta\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\beta}^*$ for combinations of c_1, c_2 and c_3 in the set $\{1, -1\}$

We have already seen how to examine if any considerable inconsistency between the theoretical and empirical distribution exists. Thus from then on it is no longer necessary to go further in the investigation by conducting hypotheses testing. Despite inadequacy mentioned in results for $\hat{\beta}^*$, the remaining results in this truth case suggest that the theoretical distributions can be used to describe the empirical distributions.

6.4.7 $10\%(n_{ik})$, $1\%(n_{ik})$, $0.1\%(n_{ik})$, $n_{ik0} \neq n_{ik1}$

All the results presented so far are calculated based on the assumption that the number of search n_{it} in each spatial unit i during time t is equal to the total number of individuals $\sum_k n_{ik}$ in all strata K in that spatial unit i that are included in micro-census data. However, by comparing the total number of micro sample individuals $\sum_{i,k} n_{ik}$ to the total available realistic search $\sum_{i,t} n_{it}$, we found that $\sum_{i,k} n_{ik} = 1435898$, $\sum_{i,t=0} n_{i0} = 12885$ and $\sum_{i,t=1} n_{i1} = 9284$. The total of the used micro search $\sum_{i,k} n_{ik}$ are more than 100 times the realistic search in both time periods. Hence, one could argue that the results presented above may not be applicable to the real world. Therefore, it would be interesting to assess the theoretical asymptotic distribution using low search rate by taking for example 10%, 1% and 0.1% of the number of micro sample individuals n_{ik} in each stratum k in spatial unit i. With 10% the total search is given by $\sum_{i,k} n_{ik} = 143582$, 1% gives $\sum_{i,k} n_{ik} = 14350$ and 0.1% gives $\sum_{i,k} n_{ik} = 1442$.

Consider for example the truth $\delta-$case with $c \in \{0.2, 1, 5\}$, the computational algorithms are implemented using the suggested proportions of n_{ik}. The obtained theoretical and empirical distributions of the applied estimates $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ are presented in **Figure 6.36**, **Figure 6.37** and **Figure 6.38**, respectively. Interestingly, the figures show that the empirical distributions of the three estimates in 10 designs across three positive c continue to be consistent with the theoretical distributions except some distributions of $\hat{\beta}^*$ and $\hat{\delta}^*$ in $c = 5$ when $0.1\%(n_{ik})$ are used.

Figure 6.36 also shows that by reducing the proportion of the used number of search, the range of the estimated applied effect $\hat{\Delta}^*$ scales down. In addition, the random variation within a single design gets bigger. The bias of $\hat{\Delta}^*$ on the other hand gets smaller in each single design. Consequently, the variation between the 10 designs gets smaller in each $\delta-$truth instance as the proportion of search decreases. The range of other estimates $\hat{\beta}^*$ and $\hat{\delta}^*$ in **Figure 6.37** and **Figure 6.38** gets wider as the number of search gets smaller due to large variability within a design.

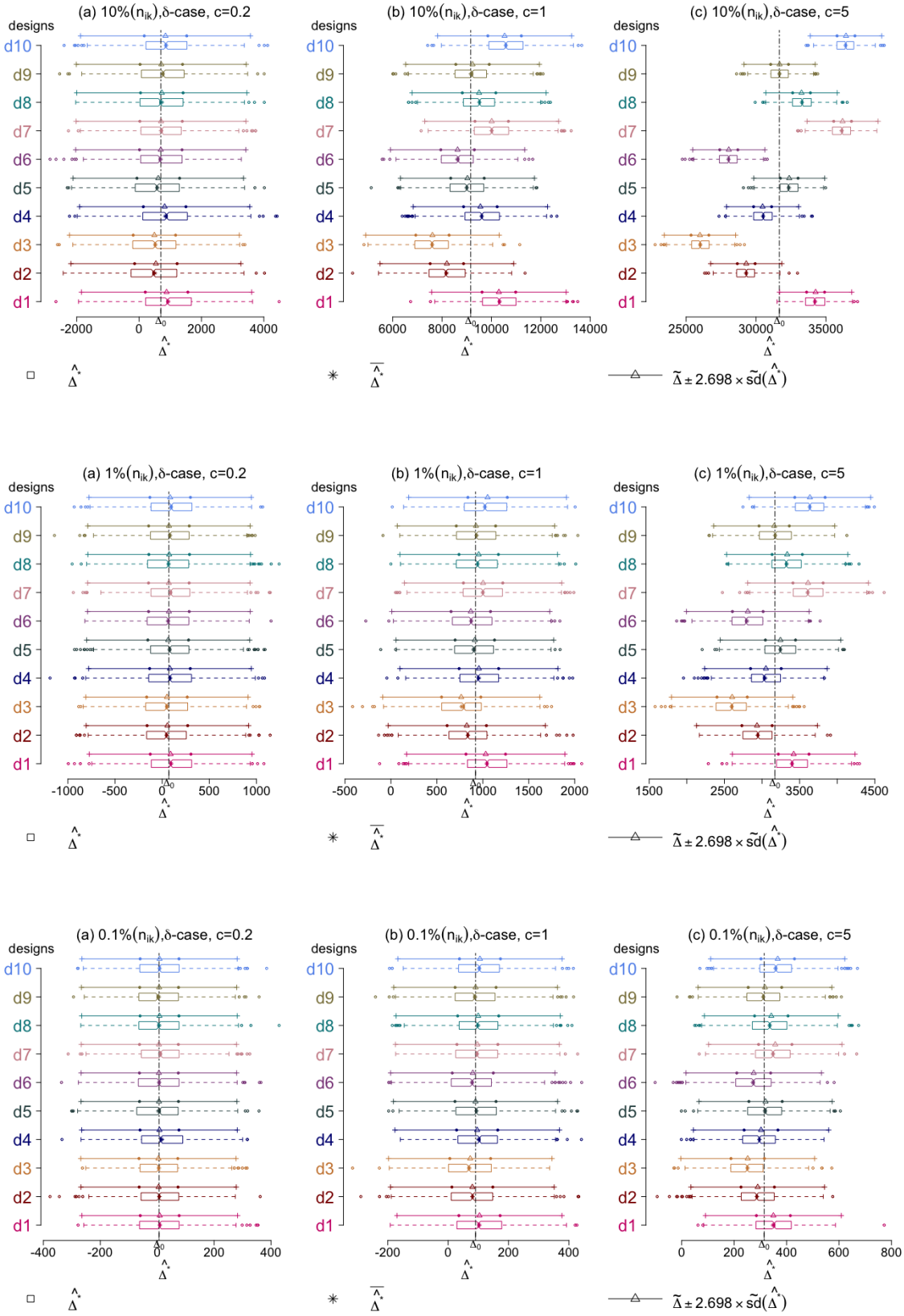


Figure 6.36: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$ for $10\%, 1\%$ and 0.1% search across $c \in \{0.2, 1, 5\}$

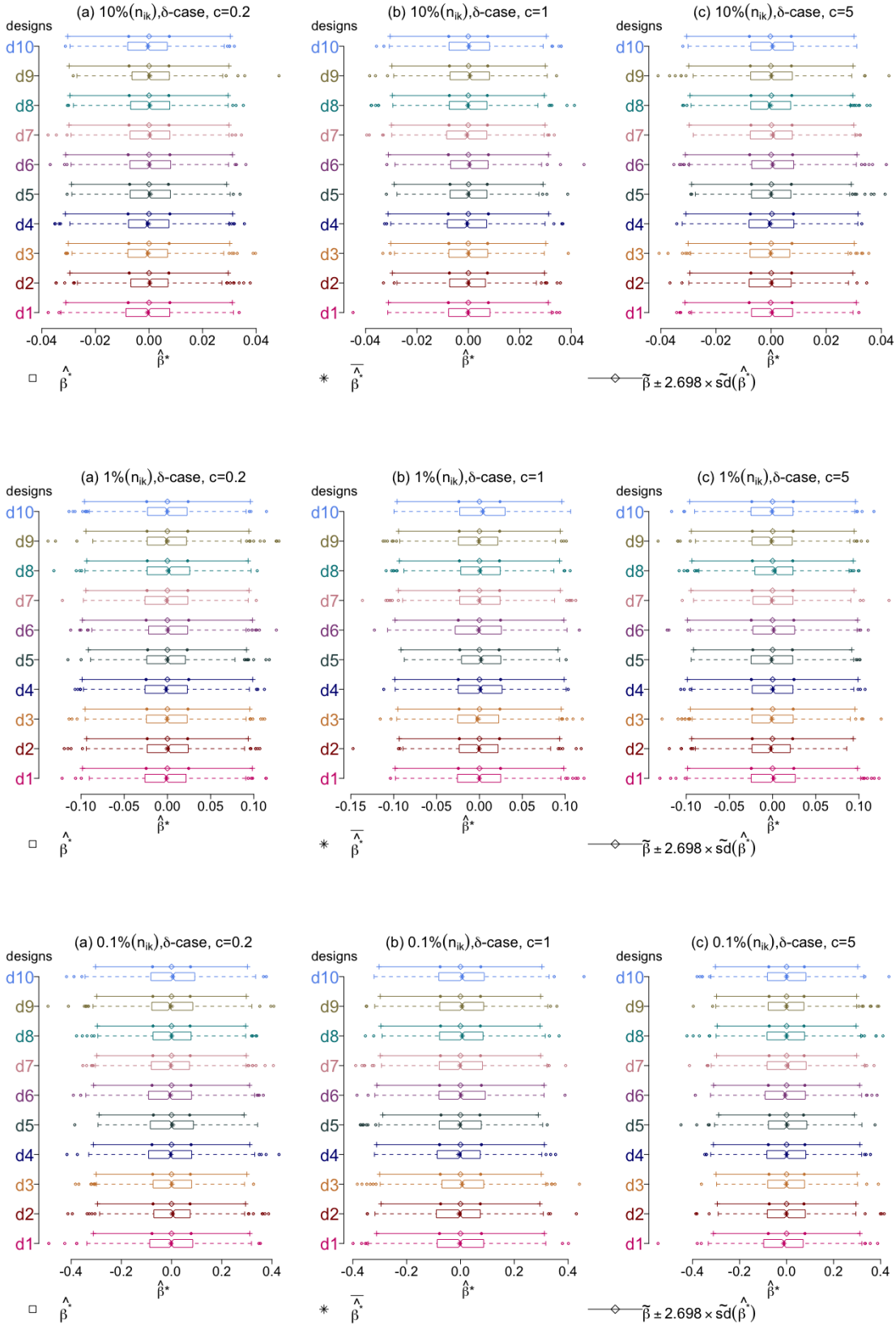


Figure 6.37: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\beta}^*$ for $10\%, 1\%$ and 0.1% search across $c \in \{0.2, 1, 5\}$

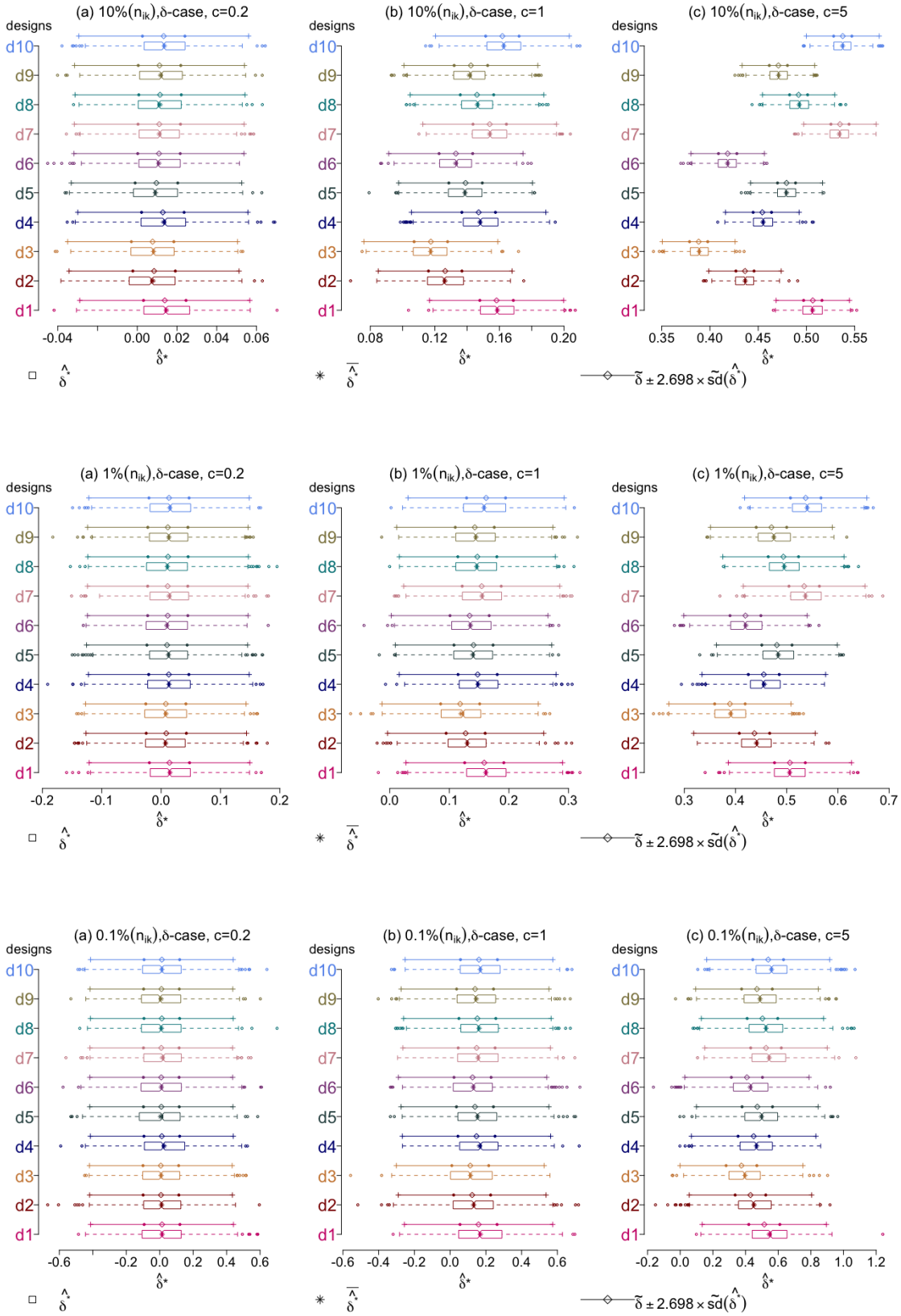


Figure 6.38: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\delta}^*$ for $10\%, 1\%$ and 0.1% search across $c \in \{0.2, 1, 5\}$

We should not run the investigation of changing the number of search to the all proposed truth cases. By using $\delta-$case, there are no notable observations resist the ability of theory to describe the sampling distribution of the estimates of the applied parameters. We believe this would continue to be valid to other truth cases. In addition, any discrepancy between the theory and simulation that appeared earlier for extreme truth instances will be expected to arise in this condition as well. In addition we examine one more interesting truth, which is $\gamma\beta\delta-$case with $c_1 = c_2 = c_3 = 1$.

In **Figure 6.39**, the theoretical and empirical distributions of the applied estimates $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ for this $\gamma\beta\delta-$case are presented for the three suggested proportions of search. The graphical findings shown in the figure support the considerable contribution of theory in describing the sampling asymptotic distribution of the estimates. This includes the distribution of $\hat{\beta}^*$ formed by $d4$ unlike earlier results obtained in this truth case.

It would be also interesting to assess the theoretical asymptotic distribution when the number of search in a stratum k in a spatial unit i in the two time periods are not similar. So far the computation is based upon the assumption that $n_{ik0} = n_{ik1}$. The realistic search is available at spatial unit level for two time periods; i.e. n_{i0} and n_{i1}. In the computation above, the search in a spatial unit i is the total micro individuals at that unit; i.e. $n_i = \sum_k n_{ik}$ which is time independent. Using realistic search, the proportion of search in each spatial unit i at each time period to the total micro individuals; i.e $\frac{n_{i0}}{n_i}$ and $\frac{n_{i1}}{n_i}$, can be used to create a difference in the number of search in each spatial unit i during two time periods. In addition, by assuming that calculated search proportions are strata independent within a spatial unit i, it then can be used to create a difference in the number of search in a stratum k in a spatial unit i between the two time periods; i.e. $n_{ik0} \neq n_{ik1}$. This change should return the total search in each time period equal to the realistic search in each time period. Calculating the total search gives $\sum_{i,k} n_{ik0} = 12872$ and $\sum_{i,k} n_{ik1} = 9288$ which are almost equal to the realistic search by taking rounding error of the proportion in each stratum in a spatial unit into account.

Using the above mentioned change in the number of search; i.e. $n_{ik0} \neq n_{ik1}$, the theoretical and empirical distributions of the applied estimates $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ for $\delta-$case with$c \in \{0.2, 1, 5\}$ and $\gamma\beta\delta-$case with $c_1 = c_2 = c_3 = 1$ are presented in **Figure 6.40** and **Figure 6.41**. The figures show that the empirical distributions can be described by the theoretical findings.

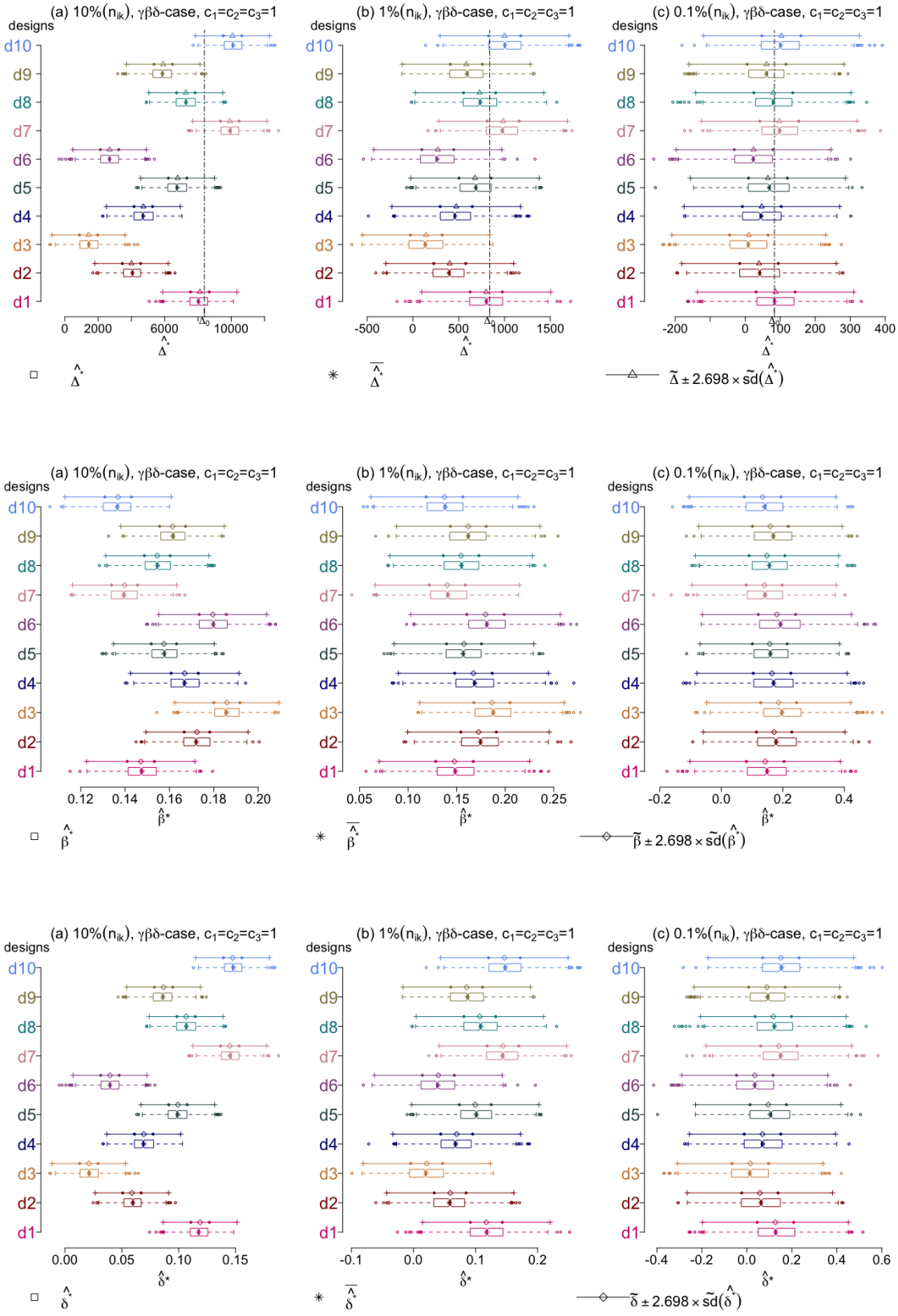


Figure 6.39: $\gamma\beta\delta-$case with $c_1 = c_2 = c_3 = 1$: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ for $10\%, 1\%$ and 0.1% search.

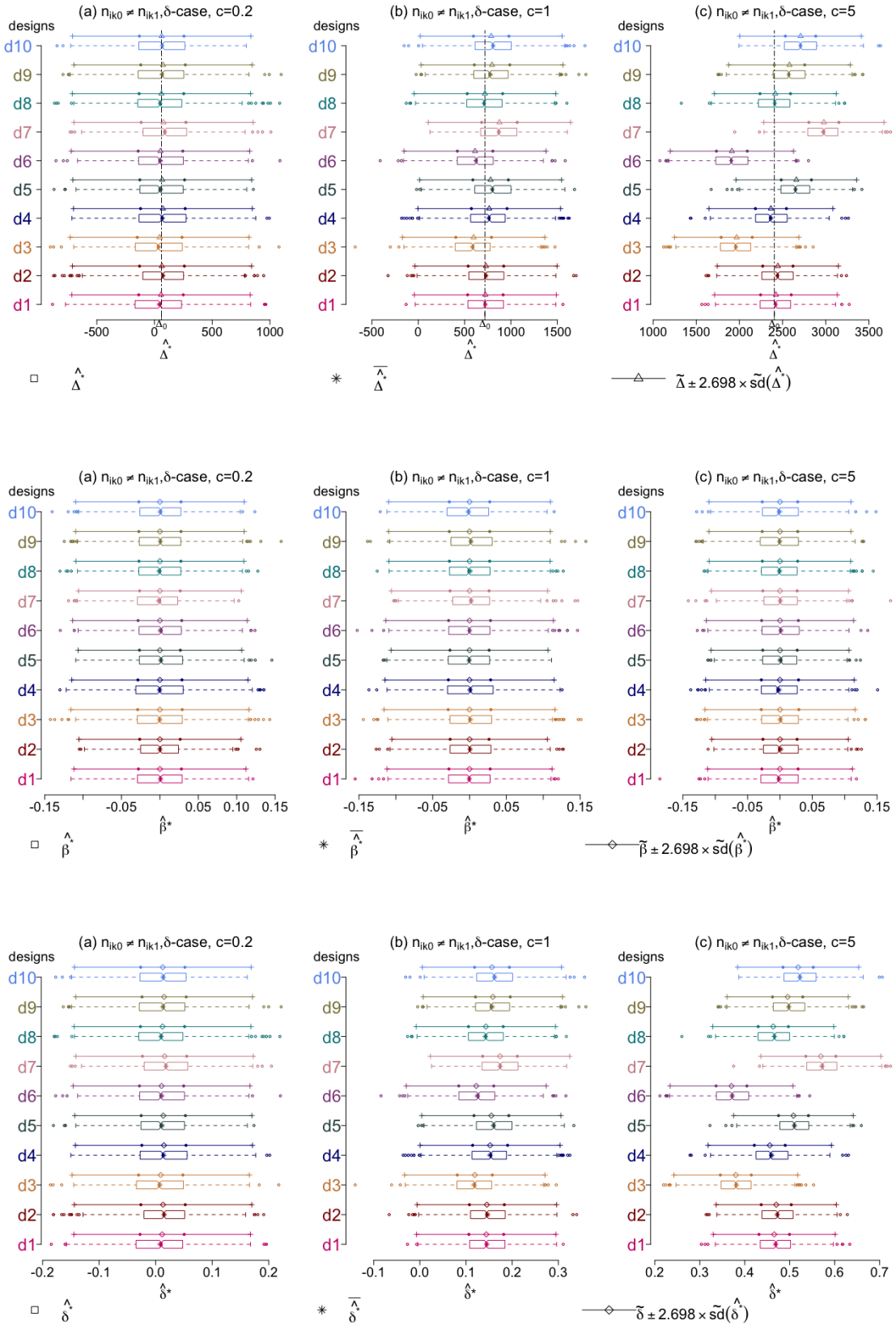


Figure 6.40: $\delta-$case: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ using different number of search over the two time periods; i.e $n_{ik0} \neq n_{ik1}$, across $c \in \{0.2, 1, 5\}$

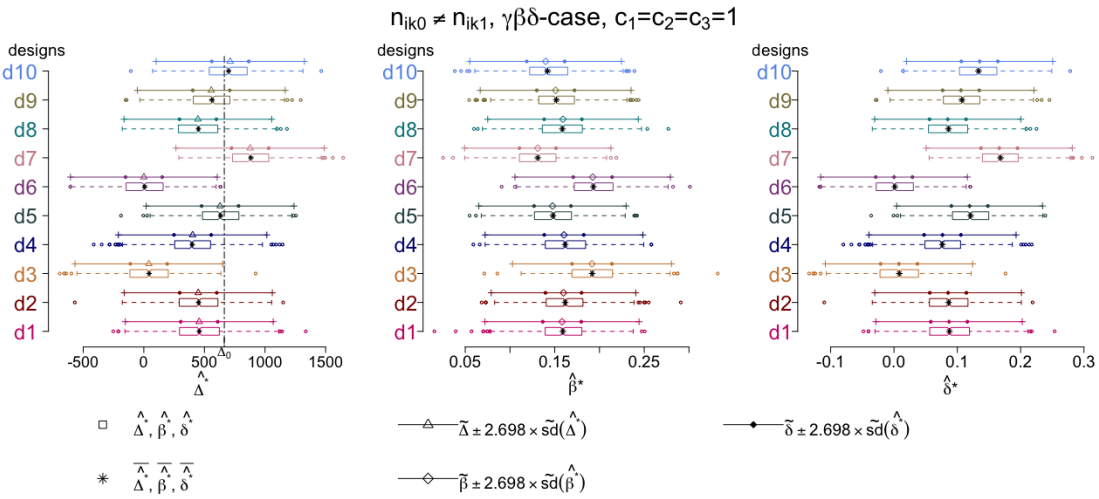


Figure 6.41: $\gamma\beta\delta-$case with $c_1 = c_2 = c_3 = 1$: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ using different number of search over the two time periods; i.e $n_{ik0} \neq n_{ik1}$.

It would also be good to check the validity of the theoretical distributions when $n_{ik0} \neq n_{ik1}$ for $\gamma-$case and $\beta-$case when $c = 1$ as well. **Figure 6.42**. The figures provide a sign that the empirical distributions can be described by the theoretical findings. The estimated applied effect $\hat{\Delta}^*$ for the presented truth cases in this figure are scaled down if they compared to their earlier results obtained by using the assumed n_{ik}; i.e. the micro number of individuals. In addition, the variation between designs is now smaller in $\beta-$case and $\beta\delta-$case, whereas there is still no obvious change between the designs in $\gamma-$case.

It would also be interesting to investigate the impact of minimising the number of search on a truth instance that showed a violation in theoretical distribution. Consider, for example, $\gamma-$case with $c = -5$ in **Figure 6.22c**. Using $1\%n_{ik}$ and $n_{ik0} \neq n_{ik1}$, the empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ are presented for this truth cases in **Figure 6.43**. Both cases show there is an agreement between the theoretical and empirical results in almost all designs as shown in the figure. Hence large number of search in each stratum at a spatial can be used to explain the earlier finding of this truth instance despite the fact that $1\%n_{ik}$ and $n_{ik0} \neq n_{ik1}$ are more close to the reality.

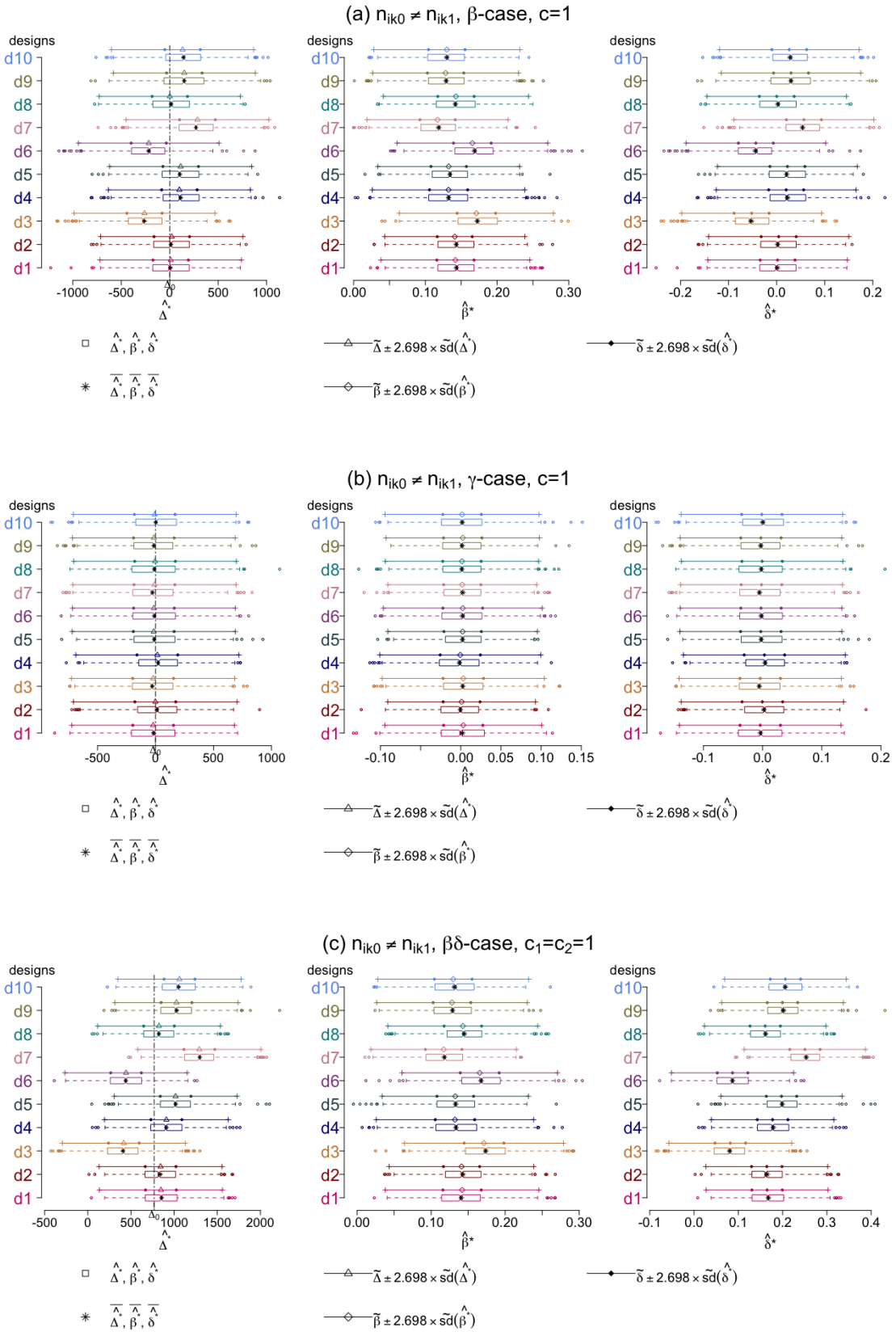


Figure 6.42: $\beta-$case and $\gamma-$case with $c = 1$ and $\beta\delta-$case with $c_1 = c_2 = 1$: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ using different number of search over the two time periods; i.e $n_{ik0} \neq n_{ik1}$.

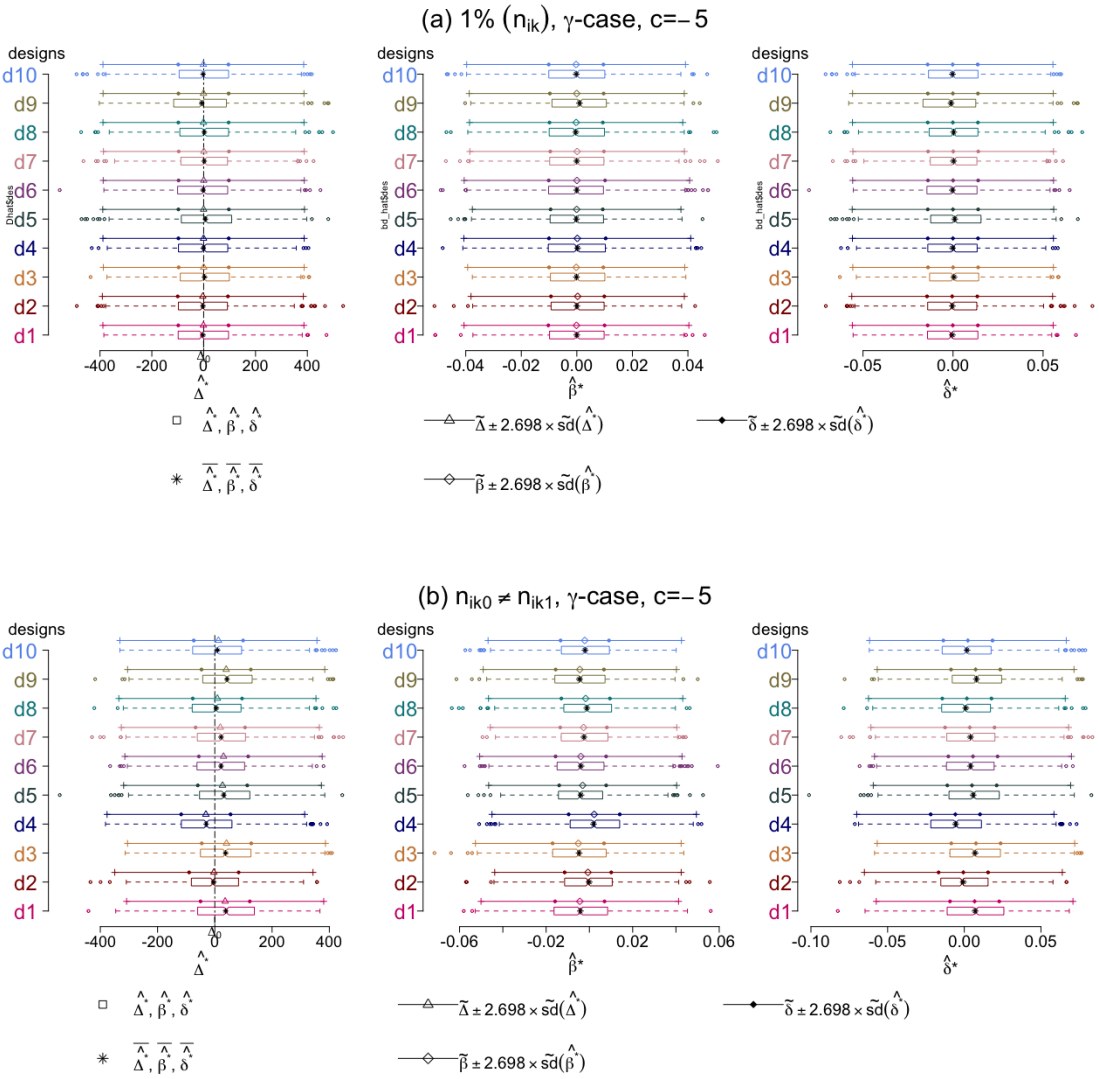


Figure 6.43: $\gamma-$case with $c = -5$: together 10 empirical sampling distributions and theoretical distributions of $\hat{\Delta}^*$, $\hat{\beta}^*$ and $\hat{\delta}^*$ using (a): 1% of the assumed search and (b) different number of search over the two time periods; i.e $n_{ik0} \neq n_{ik1}$.

6.5 Summary and Concluding Remarks

In this chapter, we discussed the validity of the theoretical asymptotic distribution of the estimated applied model parameters across different truth instances. The investigation followed the assumption that the search process is known. Different number of searches were utilised in the computation including sample micro-census individuals in each stratum in a spatial unit n_{ik}, $10\%(n_{ik})$, $1\%(n_{ik})$, $0.1\%(n_{ik})$ and realistic search. In addition, the investigations were carried out for 10 different designs in a single truth instance and the same 10 designs across different truth.

Most truth instances reported in this chapter corroborates the ability of theory to describe the sampling distribution of the estimates of the applied model parameters. Violations of the theoretical distributions were only found in few cases that related to using large difference between strata $c = 5$ or large numbers of search n_{ik}.

In this investigation there are several limitations that need to be acknowledged. The specification of truth parameters does not take into account that the overall difference between strata in a non-zero parameter vector is not centred to zero; i.e. $\sum_k w_k c \neq 0$. The values of w_k are fixed. The specification of campaign designs relied on a one design strategy. However, the developed computational algorithms are general enough to include different values of c and w_k. Future studies on these points are therefore recommended.

The computational algorithm can be extended for treating the two-stage model; purchase model conditioning on search process. The overall effects Δ_0, $\tilde{\Delta}^*$ and $\hat{\Delta}^*$ of the two-stage model were discussed in chapter 4 and chapter 5. They are in terms of the expected number of searches instead of fixed number of searches. The theoretical distributions of the estimated applied model parameters for the two-stage model were found in the previous chapter. The search parameter vector ϑ needs to be specified alongside the purchase truth parameter vector θ. This point needs more focus and therefore further research should be done to investigate the theoretical distribution of the two-stage model.

Despite these limitations, the results in this chapter indicate the key role of the theoretical approximation to the sampling distribution of applied model parameters. This saves having to do expensive Monte Carlo simulation all the time. The next chapter, therefore, moves on to discuss the performance evaluation of different design strategies.

Chapter 7

Performance Evaluation of Different Design Strategies

Assuming the plausibility of the theoretical framework for a campaign design r in a specified design strategy and a specified truth vector parameter θ_0, then $\tilde{\theta}_r^*$ is the best parameter to use for the applied model vector parameter θ^* and the proxy effect $\tilde{\Delta}_r^*$ is a rational measure of the applied effect Δ^*. Given that each design strategy is a meta-design that can be replicated R times, the question that needs to be addressed then is: which design strategy and truth instance return typically better estimates of the effectiveness of the advertising campaign, i.e. $\hat{\Delta}^*$.

This chapter begins by describing the criteria that we used to assess the performance of a specified advertising campaign design strategy for a specified truth instance. The performance evaluation criteria take into account the difference between the proxy effect and the truth, the difference between the estimate of the applied effect and the proxy effect and the difference between the estimate of the applied effect and the truth. The next two sections provide an investigation into the performance of the completely randomised design strategy and and various matched-pairs design strategies using multiple truth instances. Some remarks and conclusions are drawn in the final section.

7.1 Performance Assessments of Advertising Campaign Design Strategies

Given that a specified design strategy is repeated R times for a specified truth instance θ_0, then for design r, the estimated applied effect can be expressed as

$$\hat{\Delta}^*_{rm} = \Delta_0 + (\tilde{\Delta}^*_r - \Delta_0) + (\hat{\Delta}^*_{rm} - \tilde{\Delta}^*_r)$$

where m is a hypothetical sampling replicate within r^{th} design. The error associated with the estimated applied model parameters $\hat{\theta^*}$ in relation to the truth parameters θ_0 can be quantified by the design and sample specific total error $\hat{\Delta}^*_{rm} - \Delta_0$, which is given by a summation of the design specific approximation error $\tilde{\Delta}^*_r - \Delta_0$ and the within design sampling error $\hat{\Delta}^*_{rm} - \tilde{\Delta}^*_r$.

The quantification of the design and sample specific total error without random sampling of $\hat{\Delta}^*_{rm}$ is not straightforward. However, its error components, i.e. $\tilde{\Delta}^*_r - \Delta_0$ and $\hat{\Delta}^*_{rm} - \tilde{\Delta}^*_r$ can be used to draw inference about the design specific error. By assuming that the theoretical approximate asymptotic distribution of $\hat{\Delta}^*_{rm}$ is appropriate, then the expectation of the sampling error is given by $\mathrm{E}_r[\hat{\Delta}^*_{rm} - \tilde{\Delta}^*_r) = 0$ and its variability is $\mathrm{SD}_r(\hat{\Delta}^*_{rm} - \tilde{\Delta}^*_r) = \tilde{\mathrm{sd}}_r(\hat{\Delta}^*_{rm})$. It is desirable that the value of $\tilde{\Delta}^*_r$ should typically be near to Δ_0 to reduce the contribution of $\tilde{\Delta}^*_r - \Delta_0$, and the dispersion value $\tilde{\mathrm{sd}}_r(\hat{\Delta}^*_{rm})$ should be low to stabilize $\hat{\Delta}^*_{rm} - \tilde{\Delta}^*_r$.

For R designs sampled from the specified design strategy and the specified truth instance, a sample of $\tilde{\Delta}^*$, i.e. $\tilde{\Delta}^*_1, \tilde{\Delta}^*_2, \ldots, \tilde{\Delta}^*_R$ and a sample of $\tilde{\mathrm{sd}}(\hat{\Delta}^*)$, i.e. $\tilde{\mathrm{sd}}_1(\hat{\Delta}^*_{1m})$, $\tilde{\mathrm{sd}}_2(\hat{\Delta}^*_{2m}), \ldots, \tilde{\mathrm{sd}}_R(\hat{\Delta}^*_{Rm})$ are computed. Therefore, information is obtained about the three errors. Dropping the subscripts, the relationship between the errors is

$$\hat{\Delta}^* - \Delta_0 = (\tilde{\Delta}^* - \Delta_0) + (\hat{\Delta}^* - \tilde{\Delta}^*)$$

In this study, we use the short term *total error* to refer to $\hat{\Delta}^* - \Delta_0$, the term *approximation error* to refer to $\tilde{\Delta}^* - \Delta_0$ and the term *sampling error* to refer to $\hat{\Delta}^* - \tilde{\Delta}^*$.

From the plausible theoretical asymptotic distribution of $\hat{\Delta}^*$,

$$E[\hat{\Delta}^*] = E[E[\hat{\Delta}^*|\text{design}]] = E[\tilde{\Delta}^*],$$

and

$$\mathrm{Var}(\hat{\Delta}^* - \Delta_0) = \mathrm{Var}(\hat{\Delta}^*).$$

Using the law of total variance, $\text{Var}(\hat{\Delta}^*)$ is given by

$$\text{Var}(\hat{\Delta}^*) = E[\text{Var}(\hat{\Delta}^*|\text{design})] + \text{Var}(E[\hat{\Delta}^*|\text{design}])$$

where $\text{Var}(\hat{\Delta}^*|\text{design}) = \widetilde{\text{Var}}(\hat{\Delta}^*) = (\tilde{\text{sd}}(\hat{\Delta}^*))^2$. Hence the mean and variance of the total error $\hat{\Delta}^* - \Delta_0$ are

$$E[\hat{\Delta}^* - \Delta_0] = E[\hat{\Delta}^*] - \Delta_0 = E[\tilde{\Delta}^* - \Delta_0],$$
$$\text{Var}(\hat{\Delta}^*) = E[\widetilde{\text{Var}}(\hat{\Delta}^*|\text{design})] + \text{Var}(\tilde{\Delta}^*),$$

where $E[\widetilde{\text{Var}}(\hat{\Delta}^*|\text{design})] \approx E[(\tilde{\text{sd}}(\hat{\Delta}^*))^2]$.

Therefore, for a specified design strategy using a specified truth instance, four performance measures on the scale of Δ can be calculated using the theoretical distribution of the the campaign effectiveness estimate $\hat{\Delta}^*$ and the distribution of the approximation error $\tilde{\Delta}^* - \Delta_0$. The performance measures are $E[\hat{\Delta}^* - \Delta_0]$, $\text{sd}(\tilde{\Delta}^*)$, $E[\tilde{\text{sd}}(\hat{\Delta}^*)]$ and $\text{sd}(\hat{\Delta}^*)$, where $\text{sd}(\hat{\Delta}^*) = \sqrt{E[\widetilde{\text{Var}}(\hat{\Delta}^*)] + \text{Var}(\tilde{\Delta}^*)}$.

Also, the contributions of the two sources of variability, the approximation error $\tilde{\Delta}^* - \Delta_0$ and the sampling error $\hat{\Delta}^* - \tilde{\Delta}^*$, can easily be investigated graphically if the distribution of the approximation error is nearly symmetric about zero. From the theoretical asymptotic distribution of $\hat{\Delta}^*$, the distribution of the sampling error is approximately normal about 0. To compare the amount of variability of these two errors around zero, typical central variability measures can be used such as their interquartile ranges (IQR) or standard deviations. The advantage of using measures based on quartiles is that a standard boxplot can be used to show variability of $\tilde{\Delta}^*$ which need not be assumed to be normal. When the median of that boxplot lies at or near 0, the upper quartile (right end of box) is a good measure of the variability of $\tilde{\Delta}^*$ and the natural comparison is then to the upper quartile of the sampling error, i.e. $0.6745 \times \tilde{sd}(\hat{\Delta}^*)$. Showing boxplots of $\tilde{\Delta}^*$ and $0.6745 \times \tilde{sd}(\hat{\Delta}^*)$ on the same scale provides a convenient and immediate visual comparison of the relative sizes of the two sources of variability, as illustrated in **Figure 7.1**, by comparing the right end of the boxplot of $\tilde{\Delta}^*$ to the median of the boxplot for $0.6745 \times \tilde{sd}(\hat{\Delta}^*)$. If they lie on the same line, as illustrated in the figure, the contributions of the two sources of variability are then about the same.

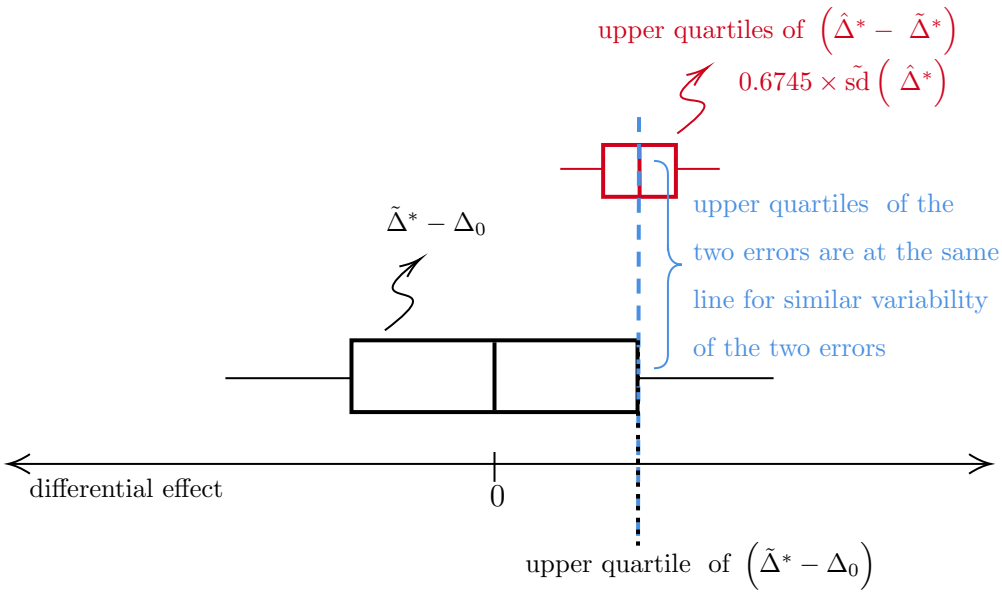


Figure 7.1: Graphical comparison of contributions of two sources of variability, variation in the sampling error $\hat{\Delta}^* - \tilde{\Delta}^*$ within designs and variation in the approximation error $\tilde{\Delta}^* - \Delta_0$ between designs, within a specified design strategy and for a specified truth instance. The label *differential effect* refers to the difference between $\tilde{\Delta}^*$ and Δ_0 which is used as a common label of two different quantities $\tilde{\Delta}^* - \Delta_0$ and $0.6745 \times \tilde{sd}(\hat{\Delta}^*)$.

When the mean bias in the approximation error is not zero, i.e. $\mathrm{E}[\tilde{\Delta}^*]$ differs noticeably from Δ_0, the two sources of variability are not directly comparable in a figure like **Figure 7.1**. Therefore, the four performance measures that are mentioned earlier in this section will be taken into consideration in a different graphical presentation introduced later in **Figure 7.5**.

The assessments are applied to the design strategies that are outlined in chapter 4: completely randomised design and matched pairs design using different matching algorithms. In the following sections, the assessments are carried out for a specified design strategy using 1000 designs and multiple truth instances to investigate how changing the truth in a certain design strategy affects the sampling error and the approximation error. In addition, the assumption made in the previous chapter that the search process is known continues here.

The assessments are performed using the number of micro sample individuals n_{ik} in a stratum k in a spatial unit i such that n_{ik} is time independent. Assessments using 1% of n_{ik} are implemented as well to be closer to reality. In addition, the realistic search

proportions in spatial unit i at a time period t are applied to each stratum k in that spatial unit; i.e. $n_{ik0} \neq n_{ik1}$. Results for the latter two schemes of the number of searches are presented in the following sections, whereas results for the first scheme are presented briefly in Appendix I.1.

7.2 Completely Randomised Design

In this section, a comparison of the two sources of variability is presented when a completely randomised design strategy is applied, using different truth instances. In addition, for the same truth instances, the variabilities are computed using partial randomisation design strategies where different percentages of spatial units are assigned to serve the new advertising campaign during the second time period. The employed percentages are 10%, 20%, 30% and 40%. The variabilities resulting from the complete randomised design strategy are compared with those obtained from partial randomised design strategies.

7.2.1 Truth Parameters: $\delta-$case, $\beta-$case, $\gamma-$case

Considering truth parameters in $\delta-$case, $\beta-$case, $\gamma-$case with $c \in \{0.2, 1, 5\}$, the approximation error $\tilde{\Delta}^* - \Delta_0$ and the upper quartiles of the sampling error $\hat{\Delta}^* - \tilde{\Delta}^*$ are shown in **Figure 7.2** and **Figure 7.3**, using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, respectively.

Focusing first on $1\%(n_{ik})$ in **Figure 7.2**, we can see that the change in the sampling error between design strategies is about the same across given truth instances. By looking to the design strategies in a single truth case across different c values, it is clear that the change in the sampling error gets narrower as the percentage of randomness increases. Comparing the variability of the sampling error associated with the complete randomised design strategy to the those resulting from partial randomised design strategies, we can see that the variability of the sampling error resulting from using the complete randomised design strategy is lower than the variability obtained by both 10% and 20% randomised design strategies but almost the same as what was obtained from using higher percentages of randomisation, i.e. 30% or 40%.

It is apparent from the figure that the range of the approximation error expands as the difference between strata increases. In addition, the change in the approximation error gets narrow as the percentage of the randomness increases or when the complete randomisation is applied. This can be observed clearly when δ truth case is used across different levels of

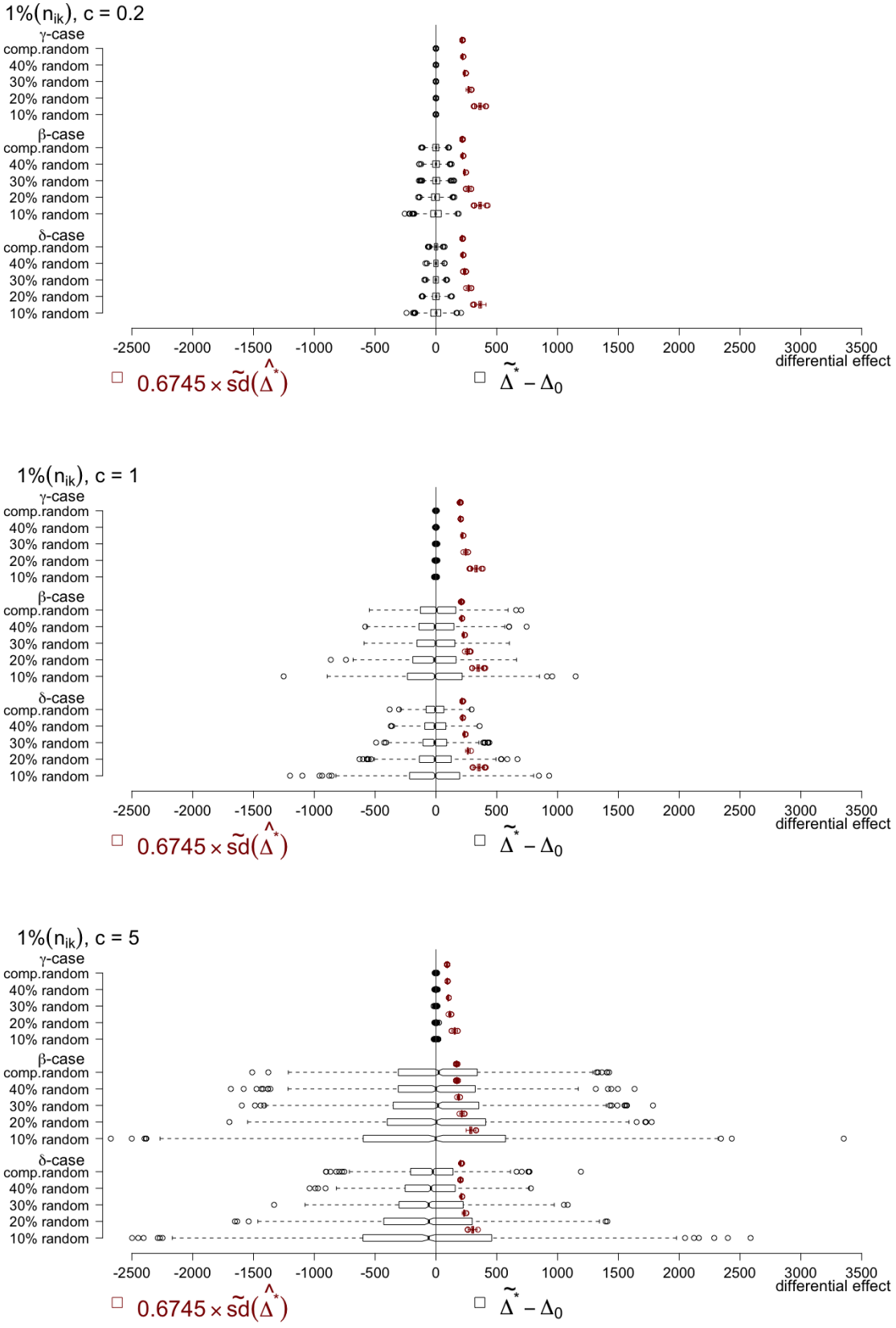


Figure 7.2: $1\%(n_{ik})$, (complete and partial randomised design strategy using truth, $\delta-$case, $\beta-$case, $\gamma-$case across $c \in \{0.2, 1, 5\}$): comparing the contributions of the two sources of variability: sampling error and approximation error. The true effect: $\Delta_{0\delta_{c=0.2}} = 70.88837$, $\Delta_{0\delta_{c=1}} = 920.7419$, $\Delta_{0\delta_{c=5}} = 3170.888$, $\Delta_{0\beta} = \Delta_{0\gamma} = 0$.

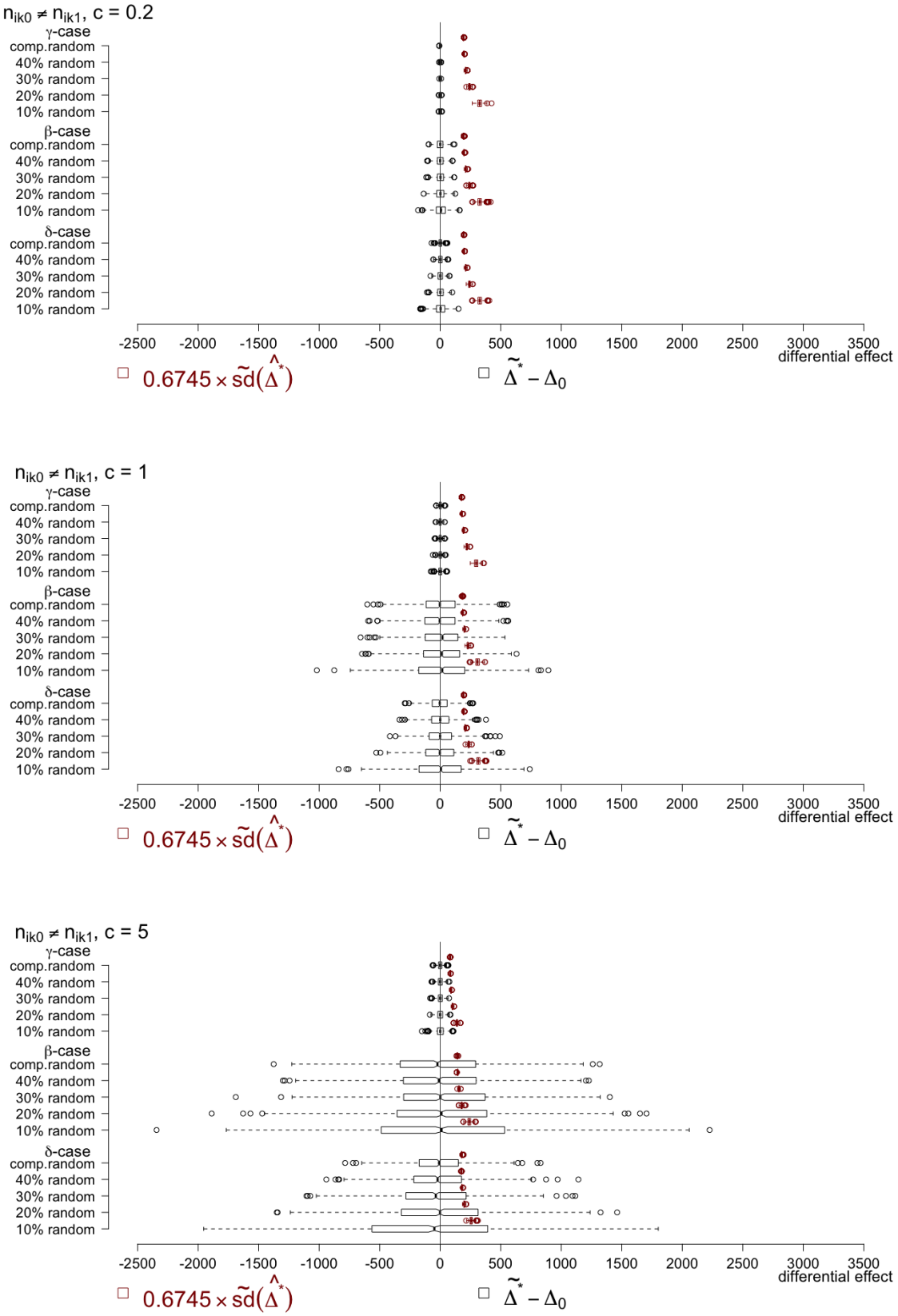


Figure 7.3: $n_{ik0} \neq n_{ik1}$, (complete and partial randomised design strategy using truth: $\delta-$case, $\beta-$case, $\gamma-$case across $c \in \{0.2, 1, 5\}$) comparing the contributions of the two sources of variability: sampling error and approximation error. The true effect: $\Delta_{0\delta_{c=0.2}} = 60.42186$, $\Delta_{0\delta_{c=1}} = 718.8298$, $\Delta_{0\delta_{c=5}} = 2404.563$, $\Delta_{0\beta} = \Delta_{0\gamma} = 0$.

difference between strata. It is hard on the other hand to notice this change consistently in the approximation error in β truth instances. Also it can be seen it is difficult to tell what is the impact of $\gamma-$case across c values on the change of approximation error over the given design strategies.

Using the proposed assessments, we can see that the contribution of the sampling error is higher than the contribution of the approximation error in using $c = 0.2$ and $c = 1$ across given truth cases and design strategies, though the difference between the errors' contributions in using $c = 1$ is slightly smaller than in $c = 0.2$. For $c = 5$, this pattern continues in $\gamma-$case but we notice the opposite in using $\beta-$case. The errors' contributions are about the same in using $\delta-$case when complete randomised design is applied. For the other design strategies in this truth case, the difference between their errors' contributions are about the same magnitude of their bias in the approximation errors.

If we turn now to **Figure 7.3** when $n_{ik0} \neq n_{ik1}$ is used, we can see that the patterns of the two sources of variability are almost the same as what was noticed above in using $1\%(n_{ik})$. In this scheme, however, the change in the approximation error obtained by $\gamma-$case with $c = 5$ is more evident compared to the earlier results.

If the whole micro sample individuals n_{ik} are employed in the computation, the standard deviations $\tilde{\text{sd}}(\hat{\Delta}^*)$ turned out to be lower than the earlier two schemes, making the sampling error to be close to zero. The contribution of the variability of both errors related to this scheme are presented in **Figure I.1** in Appendix I.1.

The findings in the three schemes of the number of search suggest that the complete randomised design strategy performs better in $\delta-$case. This result can be seen also in $\beta-$case with $c = 0.2$ and $c = 1$. This conclusion takes into consideration low magnitude of the sampling error, low magnitude of the approximation error and nearly zero bias.

7.2.2 Truth Parameters: $\beta\delta-$case

Consider truth parameters in $\beta\delta-$case with $c_1, c_2 \in \pm 1$ that gives four sets of truth combinations: $c_1 = c_2 = 1$, $c_1 = c_2 = -1$, $c_1 = -1, c_2 = 1$ and $c_1 = 1, c_2 = -1$. In **Figure 7.4**, distributions of the approximation errors and distributions of the upper quartiles of the sampling errors of complete and partial randomised design strategies are presented for the four sets of $\beta\delta-$case, using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$. The most striking observation to emerge from those four sets in the two schemes is the non-zero bias in the approximation

errors. The bias is considerably large when c_1 or c_2 is negative. This indicates that the applied model is not a good description of the truth when these truth instances are taken.

Comparing the results obtained from the four sets of combinations of the truth parameters for the two schemes in the figure below, it can be seen that truth's sets having different signs of c_1 or c_2 lead to increase in the magnitude of the approximation error as the percentage of the randomisation increases, whereas sets with the same signs of c_1 and c_2 lead to the opposite. The sets that lead to a direct relationship between the magnitude of the approximation error and the percentage of the randomization give the highest magnitude of the approximation error at the complete randomised design strategy. The interquartile range of the approximation error when the signs of c_1 and c_2 are the same is much larger than the range when the signs of c_1 and c_2 are different. The change of the sampling error between design strategies is about the same across the four sets of the truth. The magnitude of the sampling error corresponding to 10% randomisation is the highest across the four sets of the truth.

The contributions of the sampling error and the approximation error are not directly comparable in **Figure 7.4**, because of the substantial bias of $\tilde{\Delta}^*$ away from Δ_0 when c_1 and c_2 have opposite signs. However, the four performance measures: the variability $\text{sd}(\hat{\Delta}^* - \Delta_0)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^* - \Delta_0)$ and the average variability of the sampling error $\text{E}[\tilde{\text{sd}}(\hat{\Delta}^*)]$ can be considered in this case to tell more about the contributions of the variability.

In **Figure 7.5**, the four performance measures for both schemes of the number of search $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$ are presented. The results shown for the two schemes are quite similar. When the signs of c_1, c_2 are the same, the variability of the approximation error and the average variability of the sampling error have similar magnitudes though the contribution of the approximation error is a little lower for $c_1 = c_2 = -1$. The two variabilities achieve their minimum in this situation at the complete randomised design strategy. The variability of the total error is summarized by the the other two variabilities and hence it behaves the same way. When the signs of c_1, c_2 are different, there is a gap between the two variabilities where the lower contribution comes from the approximation error. The magnitude of the approximation error varies systematically between design strategies and its magnitude is much smaller than when c_1 and c_2 have the same sign.

Therefore, the total variability is almost the average theoretical variability of the estimated overall effect $\hat{\Delta}^*$. The bias is positive when the signs of c_1, c_2 are the same and negative when the signs of c_1, c_2 are different. The bias is lower in $c_1 = c_2 = 1$ than those in $c_1 = c_2 = -1$.

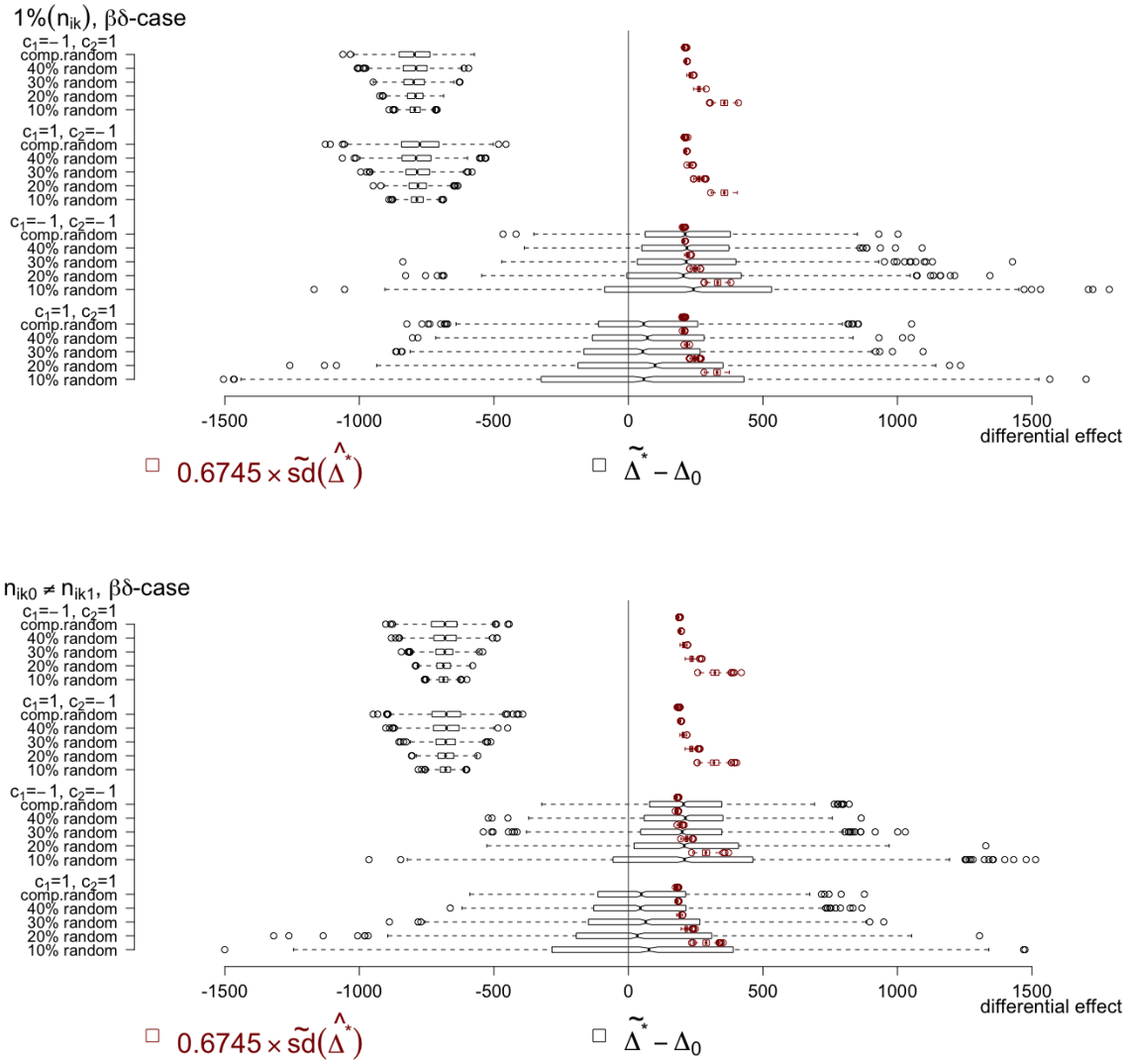


Figure 7.4: $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, (complete and partial randomised design strategy using truth, $\beta\delta-$case using combinations of $c_1, c_2 \in \pm 1$): comparing the contributions of the two sources of variability: sampling error and approximation error. Using $1\%(n_{ik})$, the true effects are: $\Delta_{0[c_1=1,c_2=1]} = 1011.249$, $\Delta_{0[c_1=-1,c_2=-1]} = 919.8866$, $\Delta_{0[c_1=1,c_2=-1]} = -118.6575$, $\Delta_{0[c_1=-1,c_2=1]} = 118.6575$. Using $n_{ik0} \neq n_{ik1}$, the true effects are: $\Delta_{0[c_1=1,c_2=1]} = 768.2345$, $\Delta_{0[c_1=-1,c_2=-1]} = 633.9853$, $\Delta_{0[c_1=1,c_2=-1]} = -26.01541$, $\Delta_{0[c_1=-1,c_2=1]} = 214.4006$.

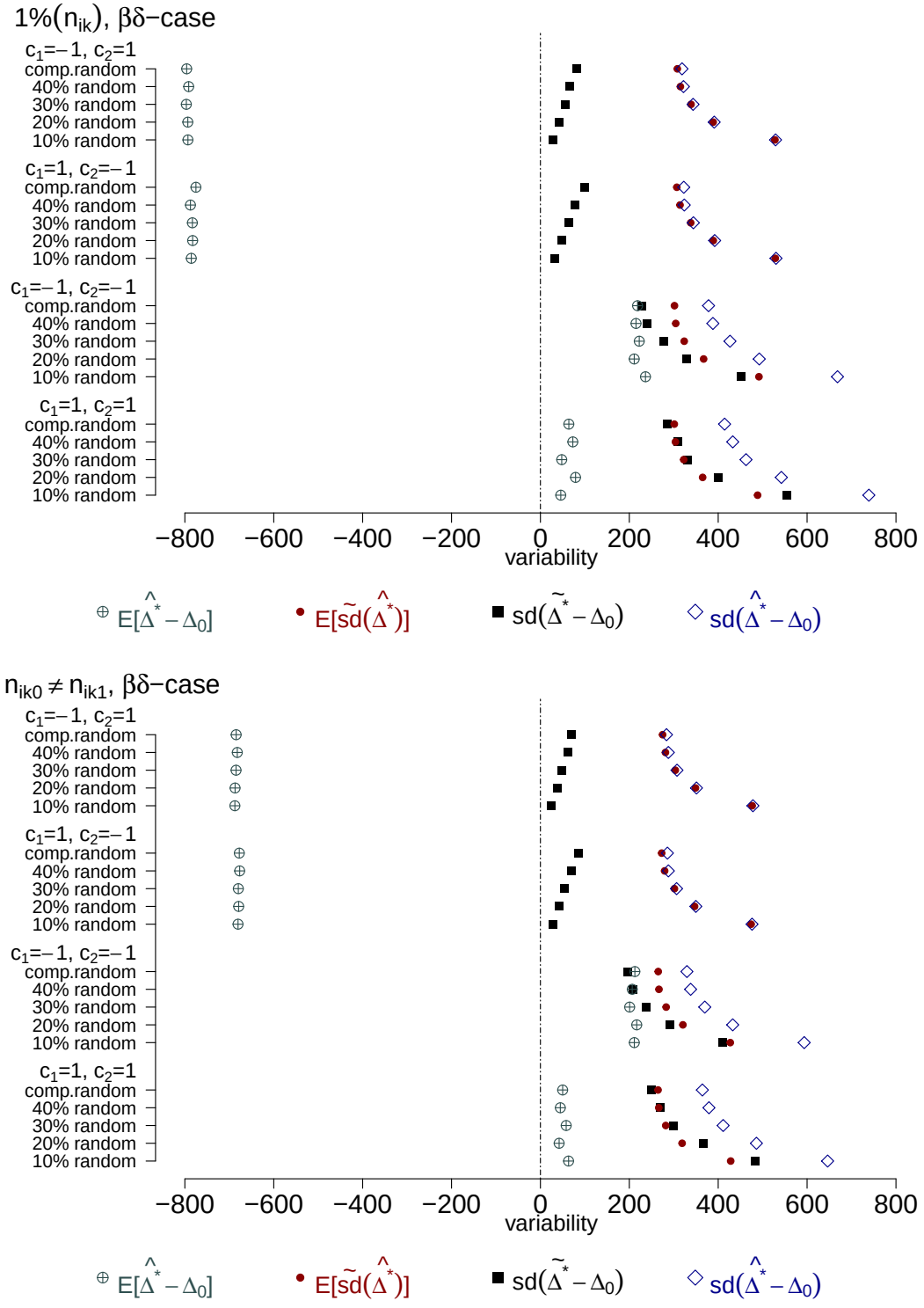


Figure 7.5: $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, (complete and partial randomised design strategy using truth, $\beta\delta-$case using combinations of $c_1, c_2 \in \pm 1$): comparing the four performance measures: the variability $\text{sd}(\hat{\Delta}^*)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ =$\text{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\text{E}[\tilde{sd}(\hat{\Delta}^*)]$.

The results presented in this figure in both schemes suggest as one would expect, that the complete randomised design strategy gives the minimum total variability in the four truth instances. However, when the signs of c_1, c_2 are different, its variability of the approximation error is slightly higher than for the other designs strategies. Also, it appears in this version of the truth that the bias is large and negative. We speculate that in some sense this version of the truth maximises the effect of the unobserved covariate on the time and advertising components of the model while limiting the overall effect of the covariate on the number of purchases, i.e. the opposite signs cancel out in the group receiving the modified campaign.

When the number of micro sample individuals n_{ik} is employed in the computation - **Figure I.2** in Appendix I.1 - the variability of the total error is attributed mainly to the approximation error. When c_1 and c_2 have similar signs, the complete randomised design strategy performs better compared to the partial randomisation strategies, whereas when c_1 and c_2 have opposite signs, the complete randomisation gives a bit higher variability of the total error. The bias in the later version of the truth is again large and negative.

7.2.3 Truth Parameters: $\gamma\beta\delta-$case

Consider truth parameters in $\gamma\beta\delta-$case with eight sets of combinations of $c_1, c_2, c_3 \in \pm 1$. In **Figure 7.6** and **Figure 7.7**, distributions of the approximation errors and distributions of the upper quartiles of the sampling errors of complete and partial randomised design strategies are presented for eight sets of $\gamma\beta\delta-$case, using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, respectively.

The eight truth versions are divided into two plots in both figures. The first plot presents the results obtained from the truth versions that have $c_1 = 1, c_2 = 1, c_3 = -1$, $c_1 = 1, c_2 = -1, c_3 = 1$, $c_1 = -1, c_2 = -1, c_3 = 1$ and $c_1 = -1, c_2 = 1, c_3 = -1$. The second plot presents the results obtained from the truth versions that have $c_1 = c_2 = c_3 = 1$, $c_1 = c_2 = c_3 = -1$, $c_1 = 1, c_2 = -1, c_3 = -1$ and $c_1 = -1, c_2 = 1, c_3 = 1$. By comparing the two plots in both schemes, the interquartile range of the approximation error is much wider in the second plot. Focusing on the minimum variability, the approximation error attains its minimum value at 10% randomisation design strategy in the first plot. On the other hand, it reaches its minimum value at the complete randomised design strategy and 40% randomisation design strategy in the second plot. The change of the sampling error between design strategies is about the same across the eight truth sets, where its

maximum magnitude is corresponding to the 10% randomisation.

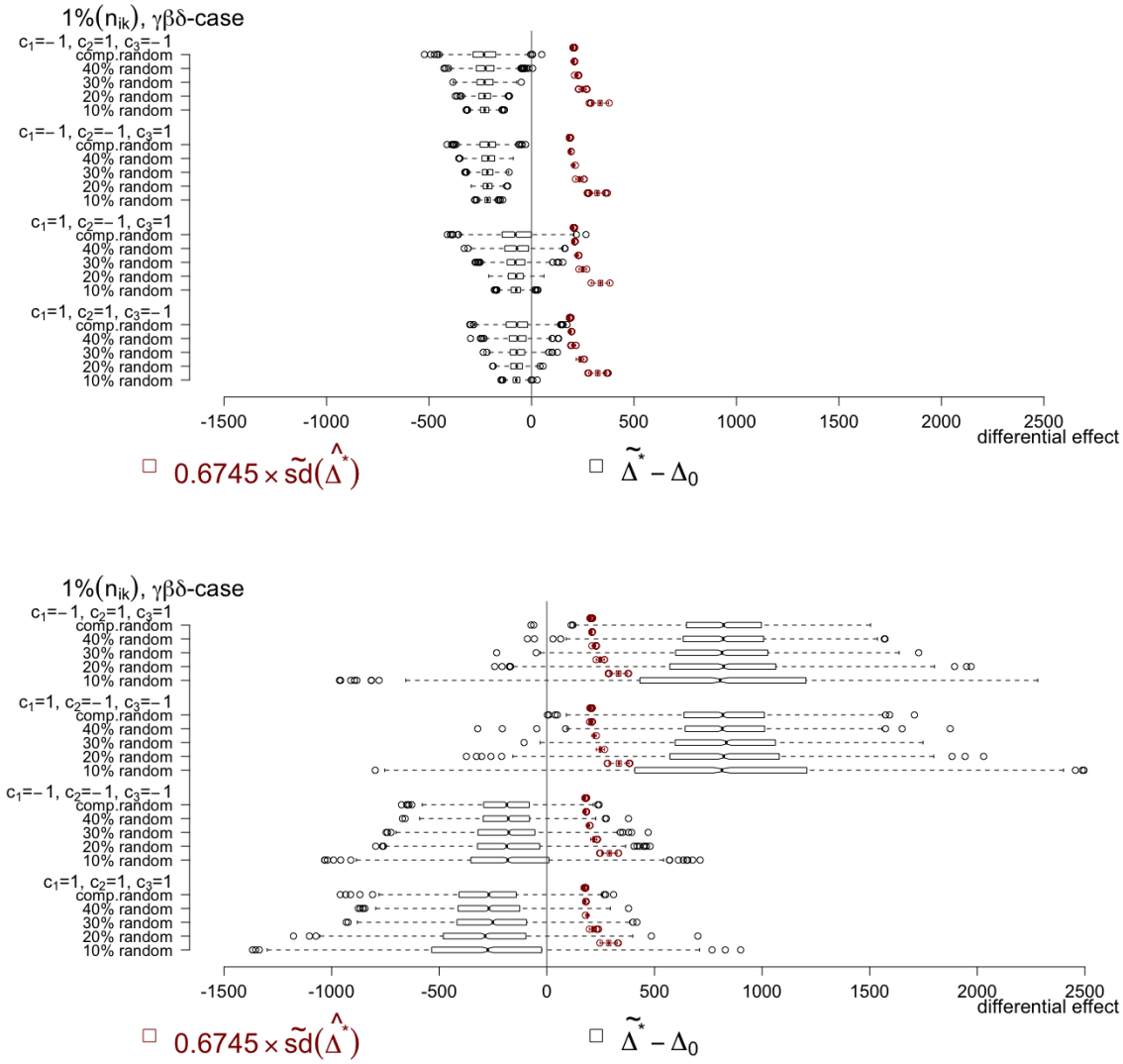


Figure 7.6: $1\%(n_{ik})$, (complete and partial randomised design strategy using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the contributions of the two sources of variability: sampling error and approximation error. The true effects are: $\Delta_{0[c_1=c_2=c_3=1]} = 838.7629$, $\Delta_{0[c_1=c_2=c_3=-1]} = 987.2001$, $\Delta_{0[c_1=1,c_2=-1,c_3=-1]} = -118.6575$, $\Delta_{0[c_1=-1,c_2=1,c_3=1]} = 118.6575$, $\Delta_{0[c_1=1,c_2=1,c_3=-1]} = -1011.249$, $\Delta_{0[c_1=1,c_2=-1,c_3=1]} = 1011.249$, $\Delta_{0[c_1=-1,c_2=-1,c_3=1]} = -919.8866$, $\Delta_{0[c_1=-1,c_2=1,c_3=-1]} = 919.8866$.

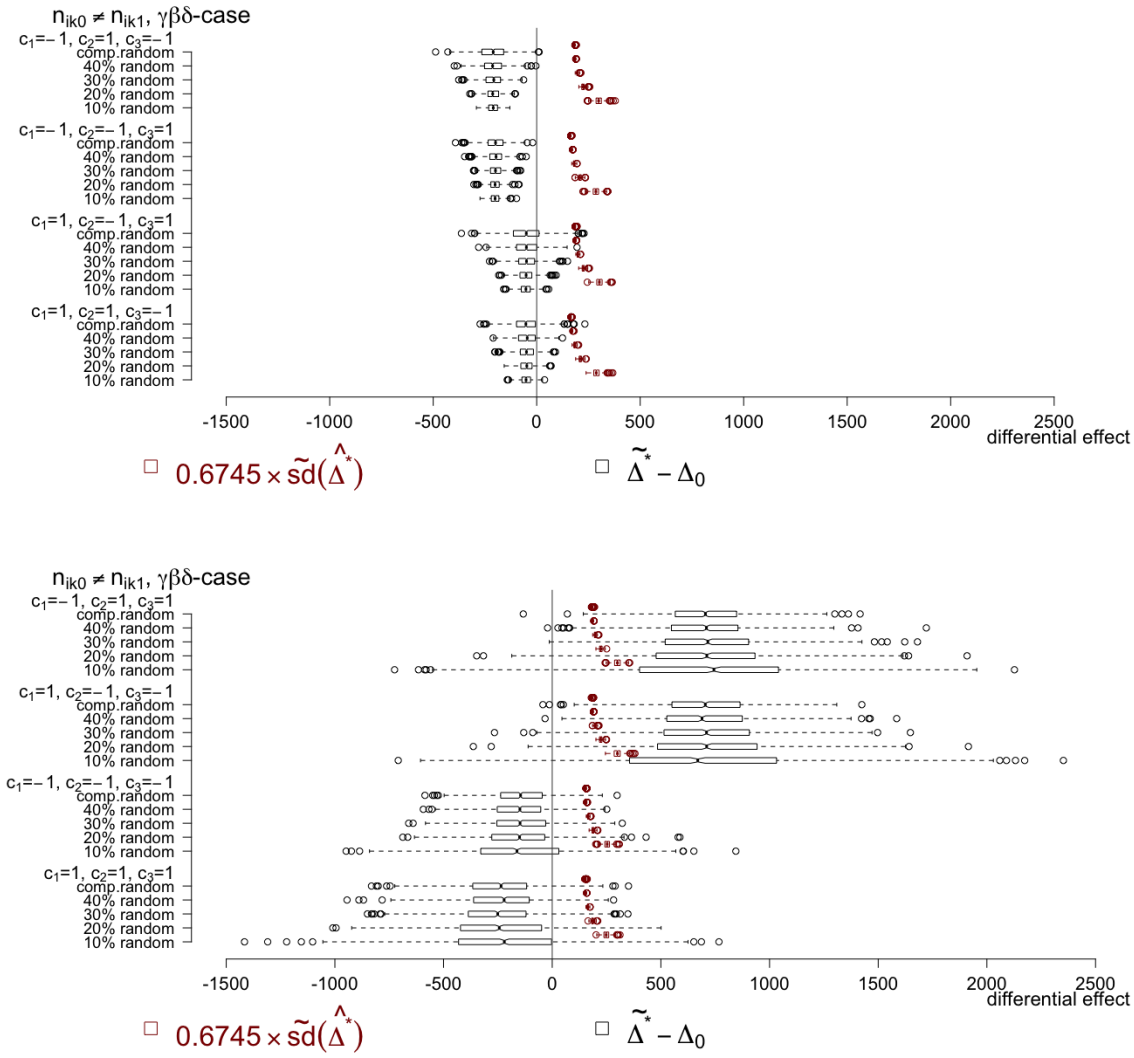


Figure 7.7: $n_{ik0} \neq n_{ik1}$, (complete and partial randomised design strategy using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the contributions of the two sources of variability: sampling error and approximation error. The true effects are: $\Delta_{0[c_1=c_2=c_3=1]} = 664.0289$, $\Delta_{0[c_1=c_2=c_3=-1]} = 747.3403$, $\Delta_{0[c_1=1,c_2=-1,c_3=-1]} = -214.4006$, $\Delta_{0[c_1=-1,c_2=1,c_3=1]} = 26.01541$, $\Delta_{0[c_1=1,c_2=1,c_3=-1]} = -768.2345$, $\Delta_{0[c_1=1,c_2=-1,c_3=1]} = 784.1475$, $\Delta_{0[c_1=-1,c_2=-1,c_3=1]} = -633.9853$, $\Delta_{0[c_1=-1,c_2=1,c_3=-1]} = 695.1372$.

The two schemes illustrate a non-zero bias in the approximation error across the eight sets of truth. Unfortunately, this signifies that the applied model is inadequate as stated in the previous truth $\beta\delta-$case.

Figure 7.8 and **Figure 7.9** depict the four performance measures, the variability and the average bias of the total error, the variability of the approximation error and the average variability of the sampling error, for the two schemes $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, respectively. The eight truth versions are presented in the two plots in both schemes in the same order as in the previous two figures. The first plot represents the truth versions when the signs of c_2 and c_3 are different, combined with $c_1 = \pm 1$. In these versions, the total variability comes mainly from the average variability of the sampling error. The second plot represents the truth versions when the signs of c_2 and c_3 are the same, combined with $c_1 = \pm 1$. In these truth instances, the change in the variabilities across the design strategies are systematic. In addition, when $c_1 = c_2 = c_3 = 1$ and $c_1 = c_2 = c_3 = -1$, the variability of the approximation error is lower than the variability sampling error, whereas when the sign of c_1 is different than the sign of c_2 and c_3, the contribution of the two variabilities is about the same. It is apparent from the figures that there is no big gain from using partial randomisation design strategies. In the eight truth instances, the figures show that the complete randomisation minimizes the total variation. The magnitude of the average bias is about the same and negative in all truth instances except when $c_1 = -1, c_2 = c_3 = 1$ and $c_1 = 1, c_2 = c_3 = -1$. The bias in these two versions is very large and positive.

When the whole micro sample of individuals n_{ik}- **Figure I.3** and **Figure I.4** in Appendix I.1 - the average variability of the sampling error is very small and the total variability is attributed mainly to the approximation error in most truth instances. In this scheme, the complete randomised design strategy still can be considered efficient compared to the partial randomisation strategies.

Overall, the investigations into the performance of the complete and partial randomised design strategies using multiple truth instances show some evidence that support the efficiency of the complete randomisation design strategy in that ability to control the the sampling error and the approximation error. This means there is no big gain from using partial design strategies. We turn now to examine the results of matched pairs designs and compare it with complete randomisation design.

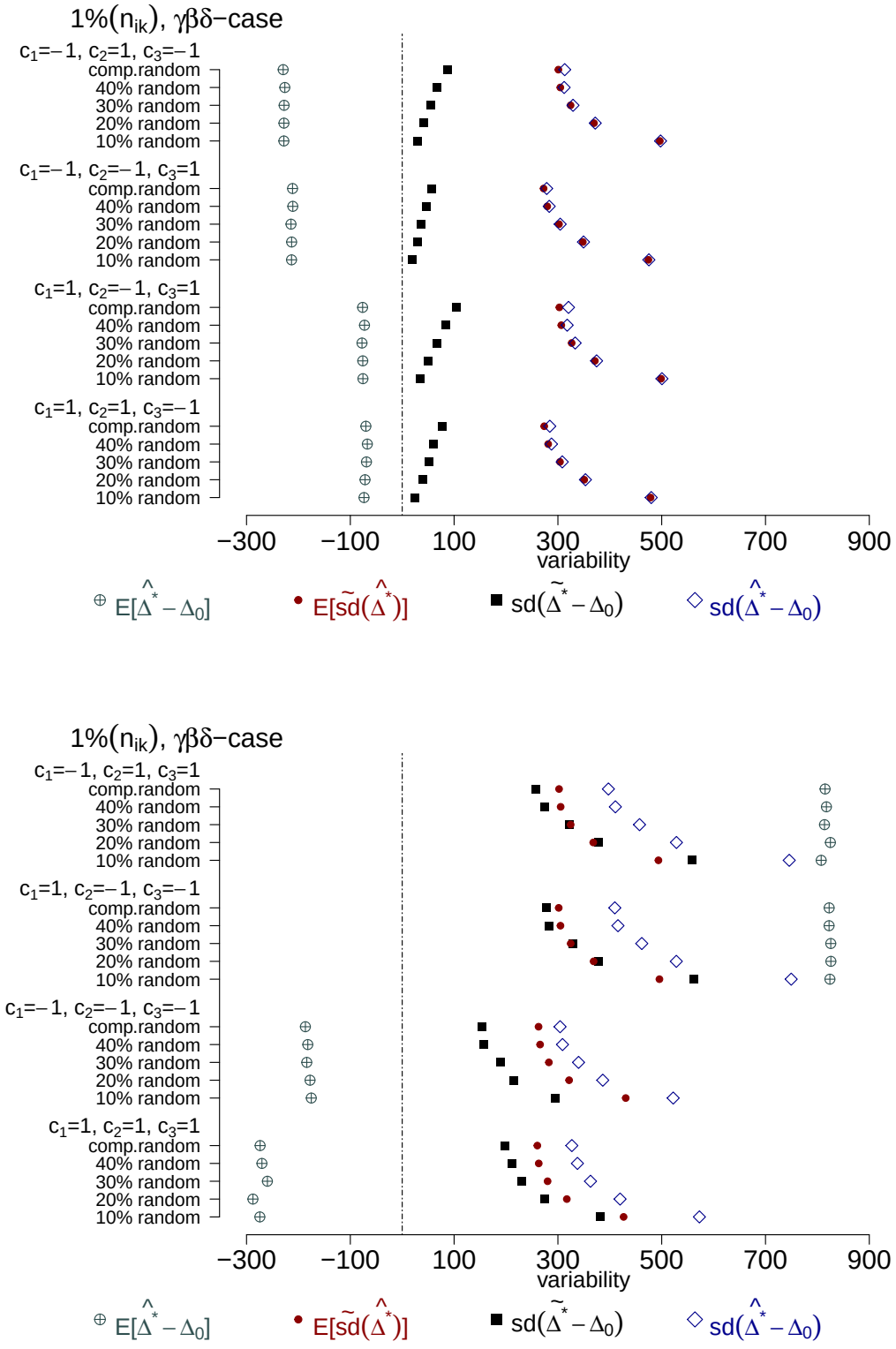


Figure 7.8: $1\%(n_{ik})$, (complete and partial randomised design strategy using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the four performance measures: the variability $\text{sd}(\hat{\Delta}^*)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ =$\text{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\text{E}[\tilde{sd}(\hat{\Delta}^*)]$.

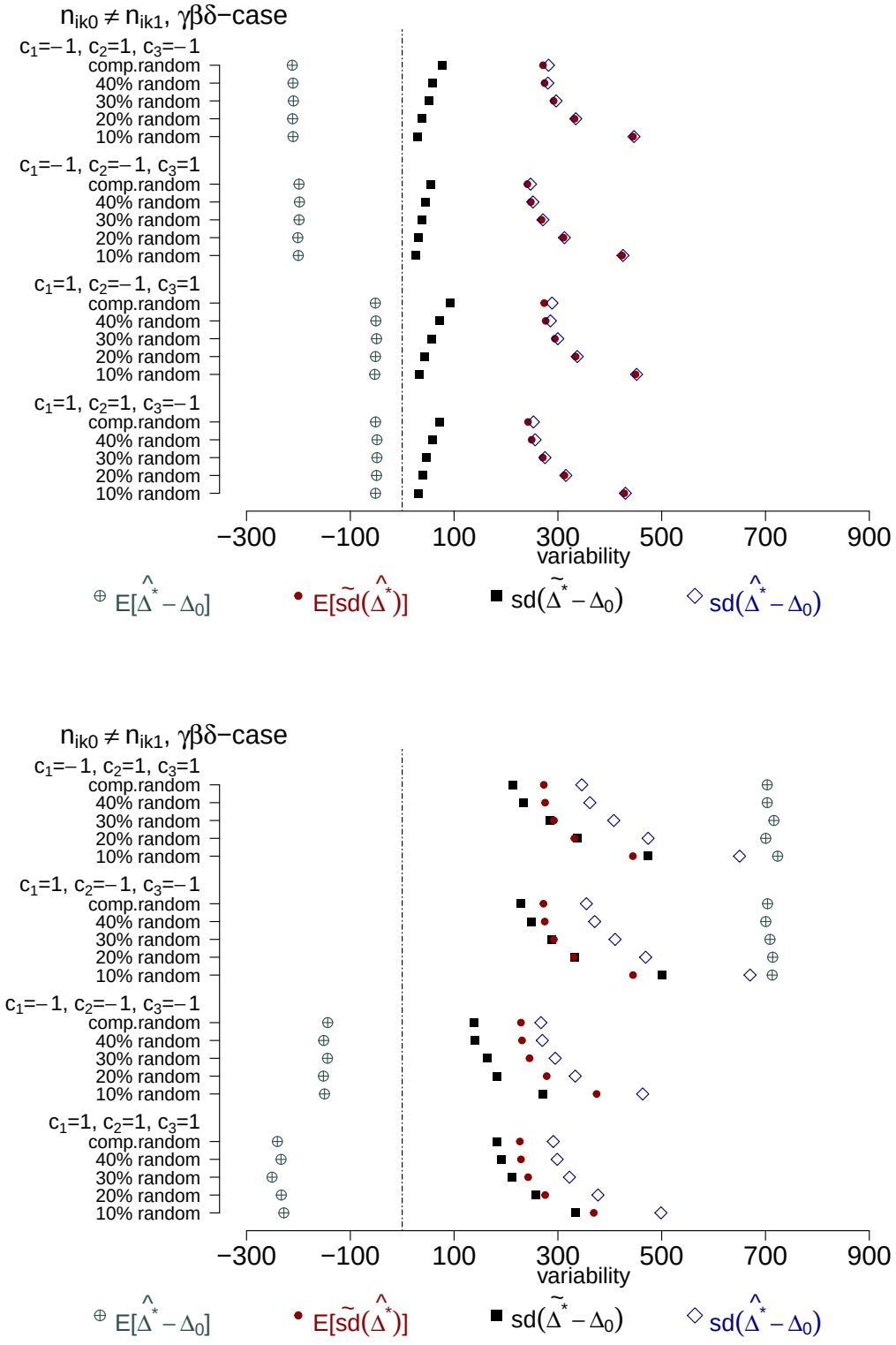


Figure 7.9: $n_{ik0} \neq n_{ik1}$, (complete and partial randomised design strategy using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the four performance measures: the variability $\text{sd}(\hat{\Delta}^*)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ =$\text{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\text{E}[\tilde{sd}(\hat{\Delta}^*)]$.

7.3 Matched-Pair Design

In this section, a comparison of the two sources of variability are presented when matched-pair design strategies are applied, using truth instances stated above. The pairs are matched using population, social grades, realistic expected search and the nearest neighbour algorithms including dissimilarity measures between social grades, dissimilarity measures between population and social grades and distances between geographical coordinates. In addition, the variabilities resulting from those strategies are compared with that obtained from the complete randomised design strategy.

7.3.1 Truth Parameters: $\delta-$case, $\beta-$case, $\gamma-$case

Consider truth parameters in $\delta-$case, $\beta-$case, $\gamma-$case with $c \in \{0.2, 1, 5\}$, the approximation error and the upper quartiles of the sampling error are displayed in **Figure 7.10** and **Figure 7.11**, using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, respectively. In both schemes, there is clear benefit of the matched-pair design strategies[1] in the reduction of the variability of the approximation error compared to the complete randomised design strategies. This can be identified clearly in $\delta-$case, $\beta-$case with $c = 1$ when pairs are matched by social grades, and the nearest neighbour algorithms especially dissimilarity measures between social grades and dissimilarity measures between population and social grades. However, matching the pairs by population or the realistic expected search has no advantage over the complete randomised design strategy. The sampling error are about the same across used design strategies and truth instances, except $\gamma-$case with $c = 5$ which provides lower sampling error. For n_{ik} scheme - **Figure I.5** in Appendix I.2, the benefit of the matched-pair design strategies is conspicuous as well and the sampling error is too low nearly zero across c and design strategies.

Using the proposed assessment criteria, we can see that the contribution of the sampling error is higher than the contribution of the approximation error across the design strategies used and for truth instances with $c = 0.2$ and $c = 1$. If we compare the variability of

[1] In the figures, short terms are used to refer to each matched-pair design strategies: pairs are matched by population ≡ *pair.pop*, social grades ≡ *pair.socgr*, realistic expected search≡ *pair.search*, dissimilarity measures between social grades ≡ *pair.d.socgr*, dissimilarity measures between population and social grades ≡ *pair.d.socgrpop*, distances between geographical coordinates ≡ *pair.d.coordinate*

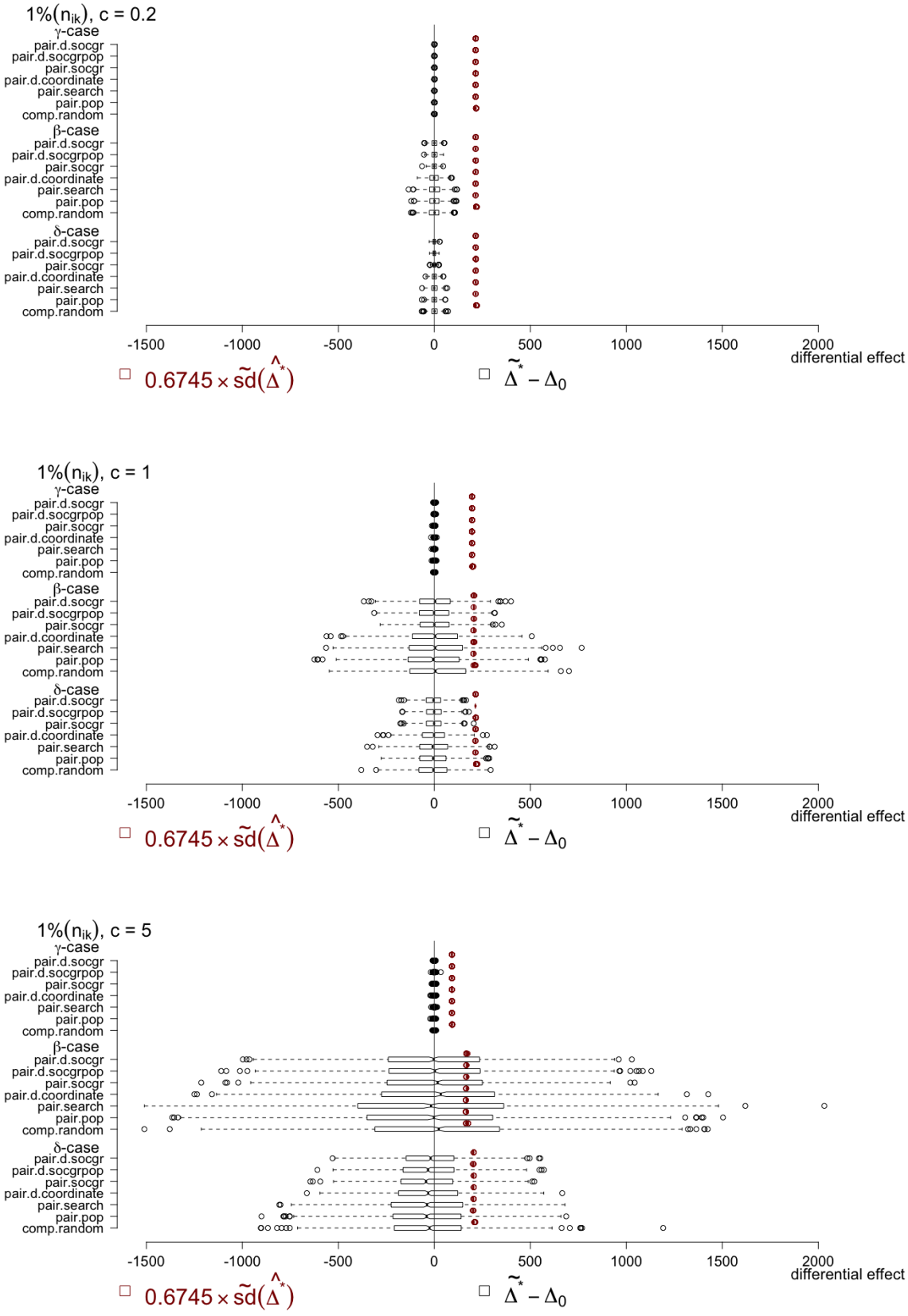


Figure 7.10: $1\%(n_{ik})$, (matched-pair design strategies using truth, $\delta-$case, $\beta-$case, $\gamma-$case across $c \in \{0.2, 1, 5\}$): comparing the contributions of the two sources of variability: sampling error and approximation error.

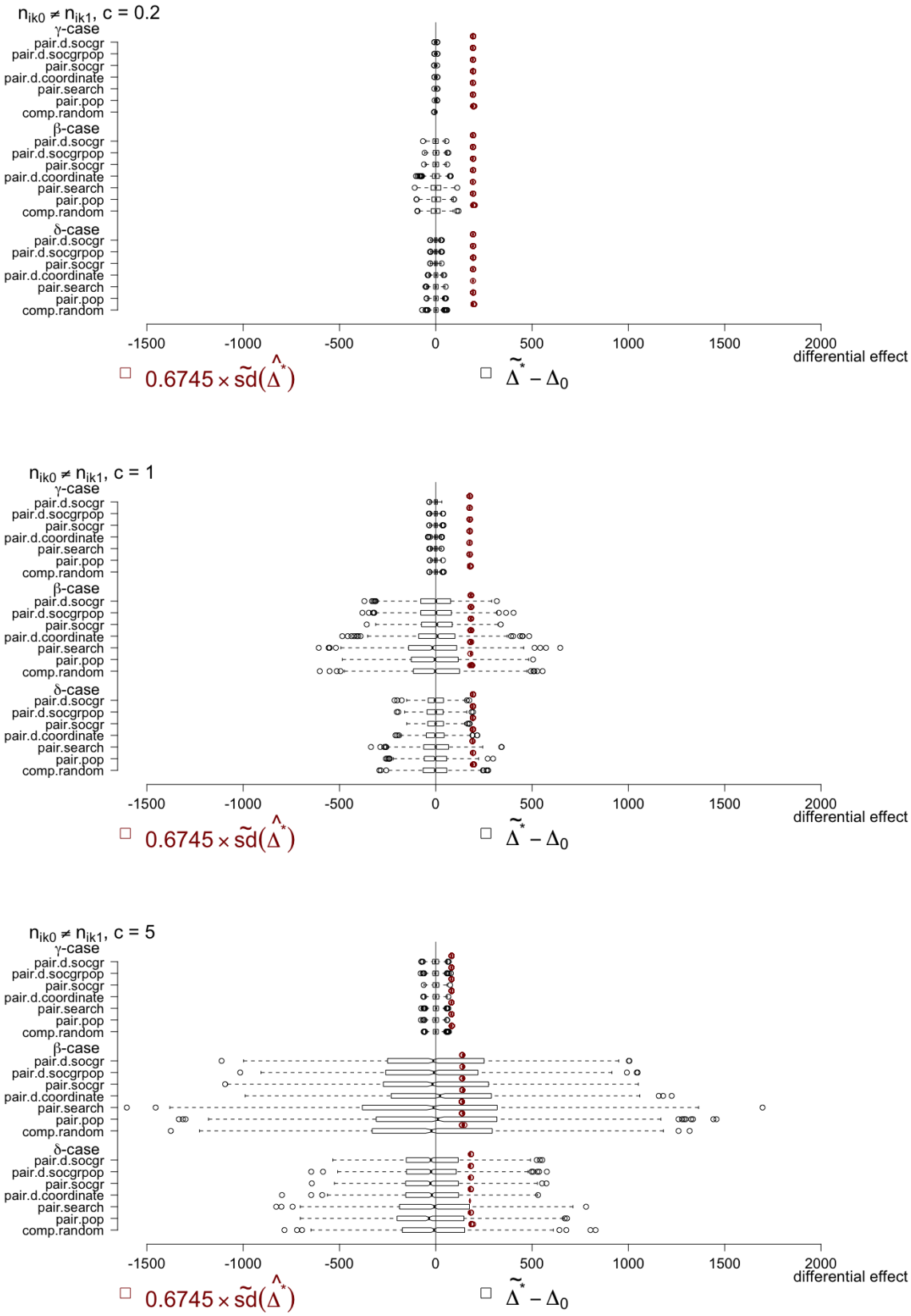


Figure 7.11: $n_{ik0} \neq n_{ik1}$:: matched-pair design strategies using truth: $\delta-$case, $\beta-$case, $\gamma-$case across $c \in \{0.2, 1, 5\}$:: comparing the contributions of the two sources of variability: sampling error and approximation error.

the approximation errors across the truth instances, we notice that the variability of the approximation error is very low in $\gamma-$case, especially in $c = 1$ and $c = 5$. This indicate a difficulty in estimating the campaign effects in these two extreme truth instances.

7.3.2 Truth Parameters: $\beta\delta-$case

Consider truth parameters in $\beta\delta-$case with four sets of truth combinations resulting from $c_1, c_2 \in \pm 1$ for both schemes of the number of search $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$. **Figure 7.12** presents the distributions of the approximation errors and distributions of the upper quartiles of the sampling errors of complete and matched-pair design strategies for the four truth sets in both schemes. The non-zero bias continues to play a role in this truth case. In the first scheme, the advantages of matching pairs by social grades over the complete randomised design strategies is apparent across the four truth sets, whereas in the second scheme this is not so obvious. The sampling error is almost identical in all truth versions but it is slightly lower in $n_{ik0} \neq n_{ik1}$.

In **Figure 7.13**, the four performance measures for both schemes $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$ are presented. The average variability of the sampling error between designs are the same across the four truth instances. The change in the total variability and the variability of the approximation error between design strategies is about the same. The variability of the approximation error is small when signs of c_1 and c_2 are different, thereby the total variability in these truth versions comes mostly from the sampling error. The magnitude of the bias in these instances is large and negative. The four measures are bunched almost together when the signs of c_1 and c_2 are the same. In these instances, design strategies relying on the social grades appear to perform better compared to other design strategies.

When the whole micro sample of individuals n_{ik}- **Figure I.6** in Appendix I.2 - the results shows the advantage of the matched-pairs design strategies over the complete random.

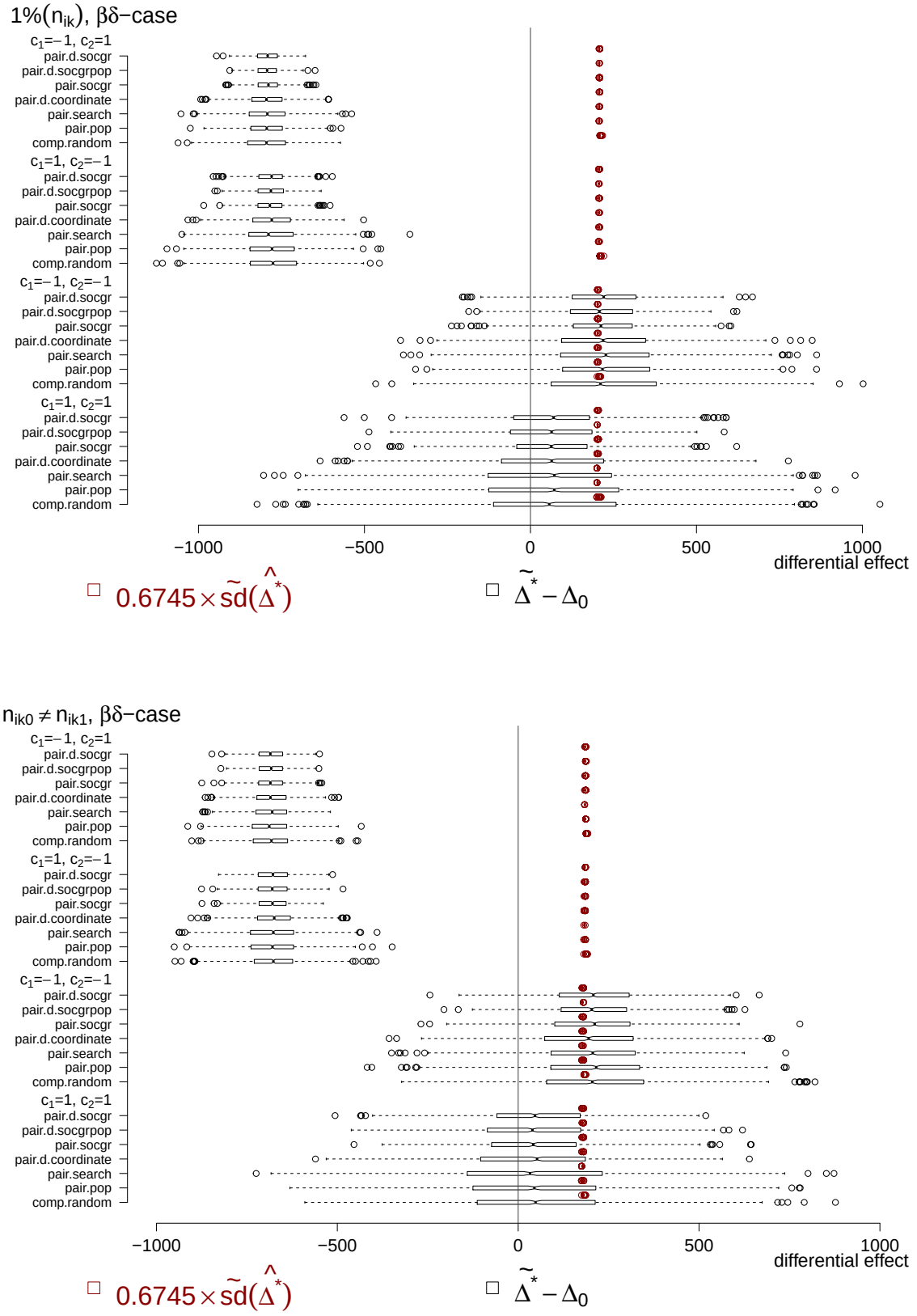


Figure 7.12: $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, (matched-pair design strategies using truth, $\beta\delta-$case using combinations of $c_1, c_2 \in \pm 1$): comparing the contributions of the two sources of variability: sampling error and approximation error.

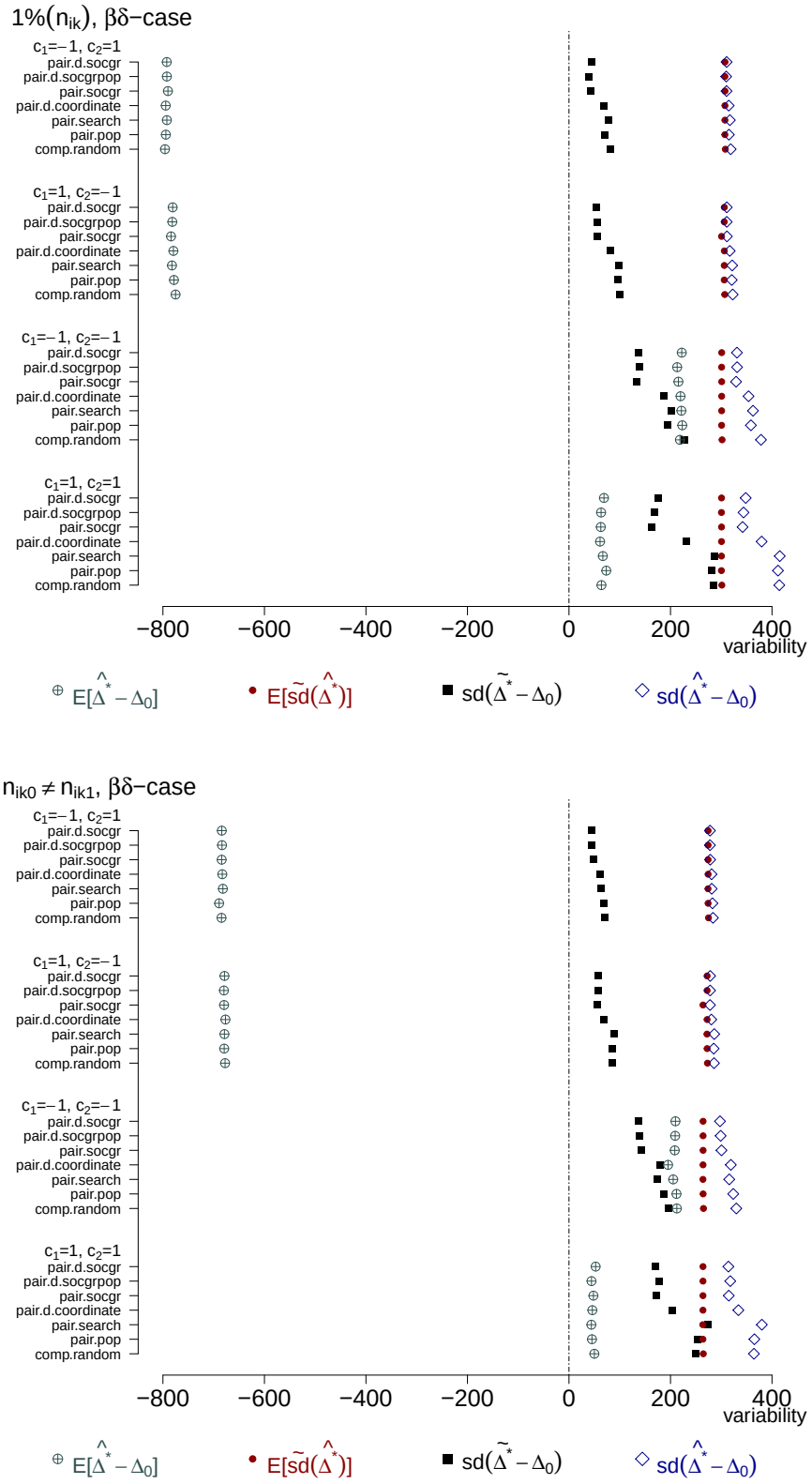
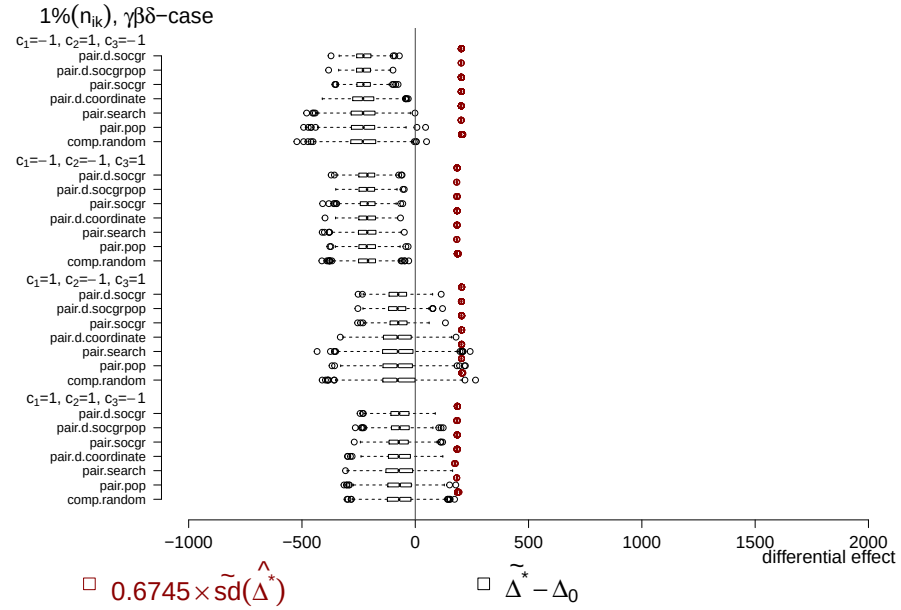


Figure 7.13: $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, (matched-pair design strategies using truth, $\beta\delta-$case using combinations of $c_1, c_2 \in \pm 1$): comparing the four performance measures: the variability $\text{sd}(\hat{\Delta}^*)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ =$\text{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\text{E}[\tilde{sd}(\hat{\Delta}^*)]$.

7.3.3 Truth Parameters: $\gamma\beta\delta-$case

Consider truth parameters in $\gamma\beta\delta-$case with eight sets of combinations of $c_1, c_2, c_3 \in \pm 1$. In **Figure 7.14** and **Figure 7.15**, distributions of the approximation errors and distributions of the upper quartiles of the sampling errors of complete and matched-pair design strategies are presented for eight sets of $\gamma\beta\delta-$case, using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, respectively. The benefit of matching pairs using social grades and spatial proximity can be seen clearly in this truth case, especially when $c_1 = -1, c_2 = 1, c_3 = -1$ and $c_1 = 1, c_2 = -1, c_3 = 1$, followed by $c_1 = -1, c_2 = 1, c_3 = 1$ and $c_1 = 1, c_2 = -1, c_3 = -1$. However, these design strategies has biased results as the other design strategies.

In **Figure 7.16** and **Figure 7.17**, the four performance measures for both schemes of the number of search $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$ are presented. Following the change in the variability of the approximation error, the advantages of using social grades can plainly be seen in the above mentioned four truth sets. Considering no substantial change in the variability of the sampling error across truth sets, the bias and variability of the approximation error are low in truth sets with $c_1 = 1, c_2 = 1, c_3 = -1$ and $c_1 = 1, c_2 = -1, c_3 = 1$. However the former set does not show any real difference between the design strategies, whereas the latter points out that the design strategies which relied on the social grades perform better. This finding shows up too when the whole micro sample of individuals n_{ik} is used, see **Figure I.7** and **Figure I.8** in Appendix I.2.



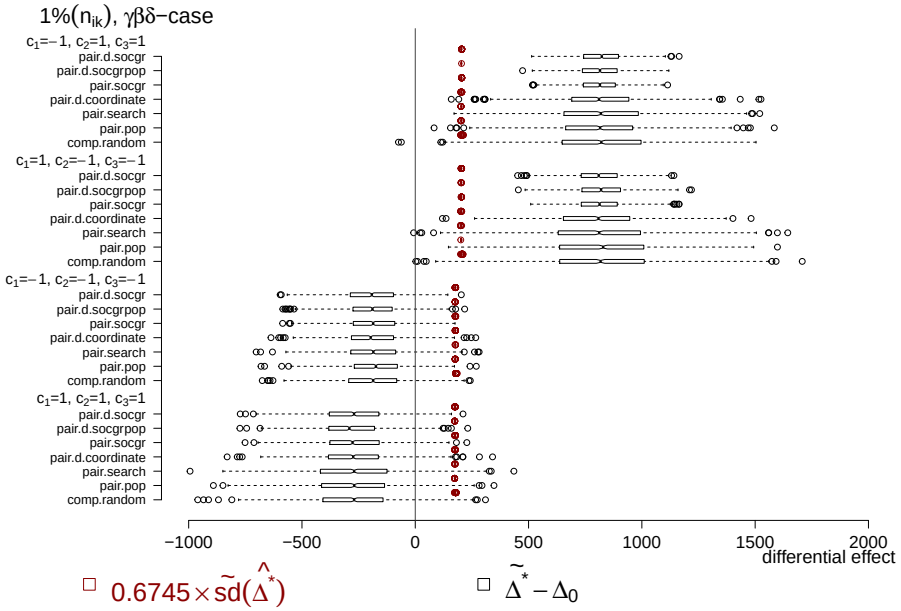


Figure 7.14: $1\%(n_{ik})$, (matched-pair design strategies using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the contributions of the two sources of variability: sampling error and approximation error.

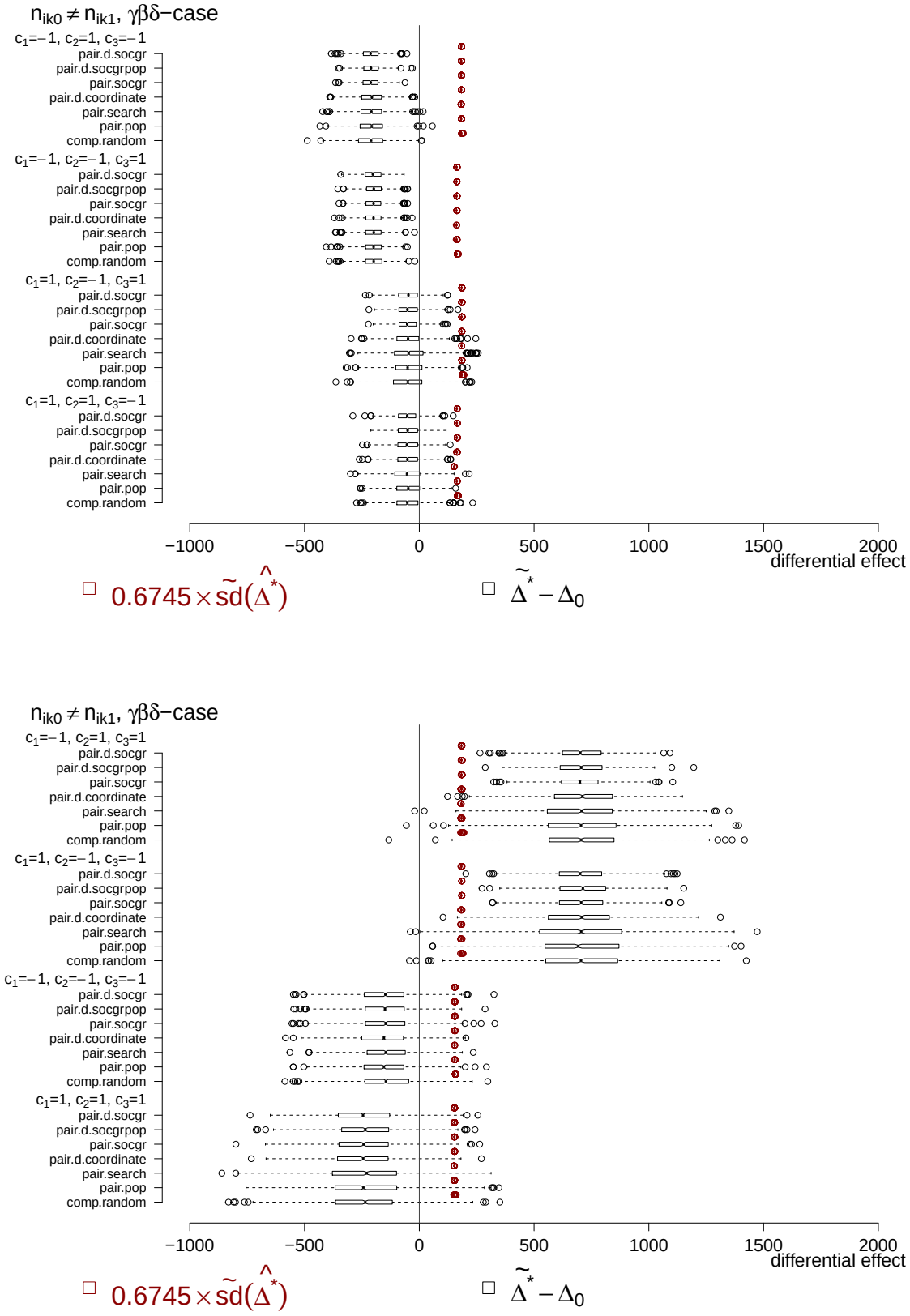


Figure 7.15: $n_{ik0} \neq n_{ik1}$, (matched-pair design strategies using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the contributions of the two sources of variability: sampling error and approximation error.

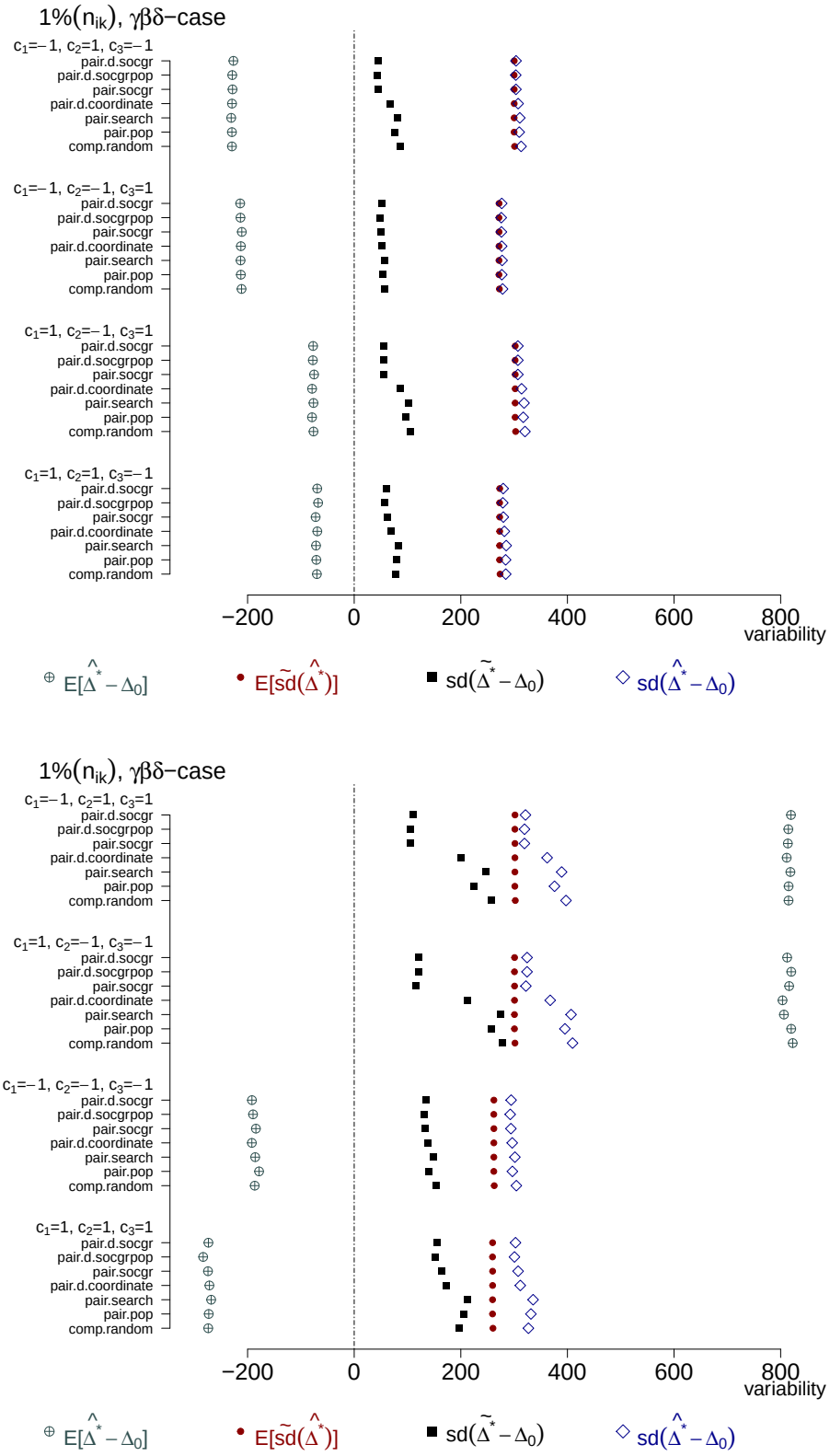


Figure 7.16: $1\%(n_{ik})$, (matched-pair design strategies using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the four performance measures: the average bias $\mathrm{E}[\hat{\Delta}^* - \Delta_0]$ =$\mathrm{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\mathrm{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\mathrm{E}[\tilde{sd}(\hat{\Delta}^*)]$.

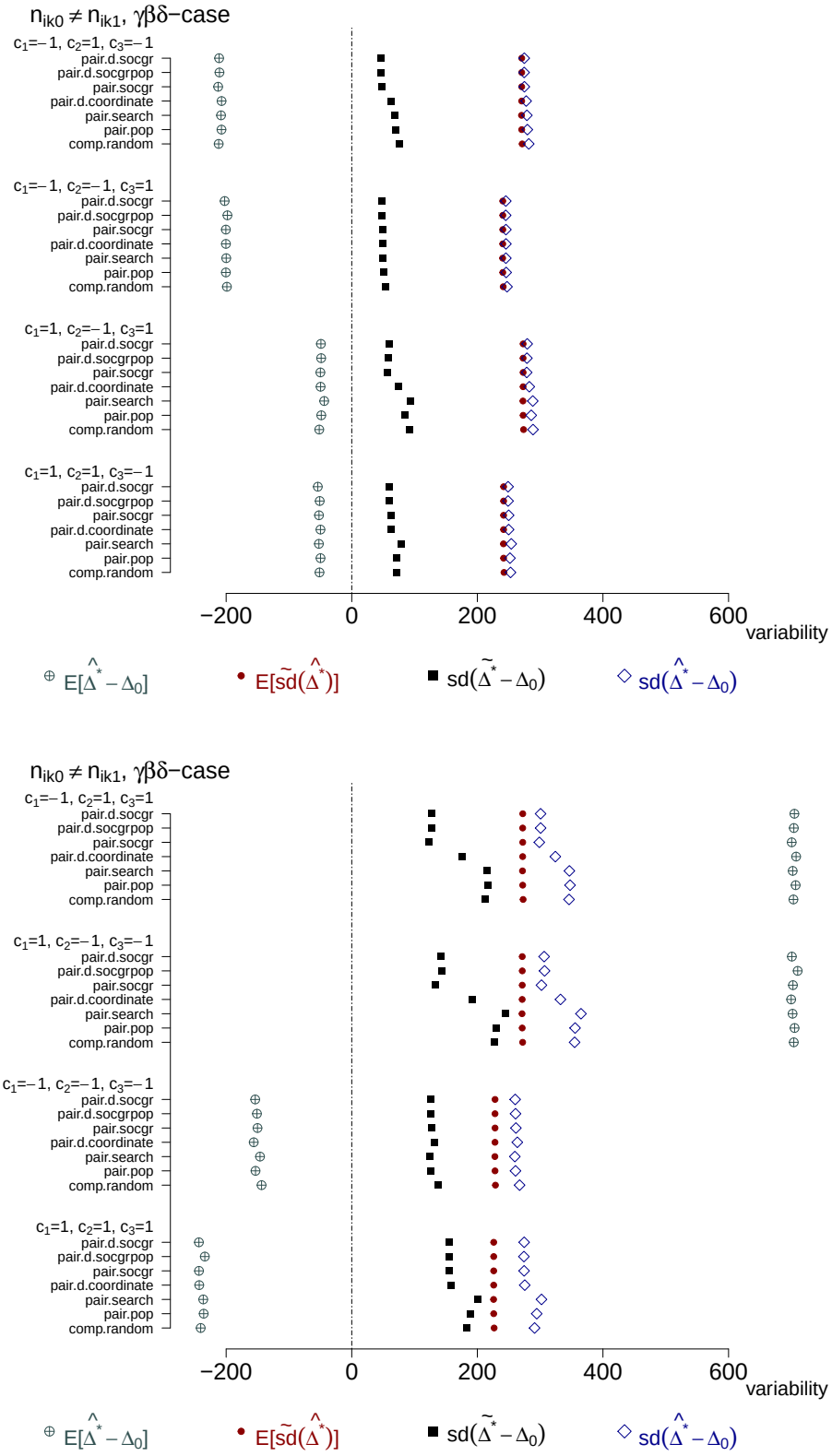


Figure 7.17: $n_{ik0} \neq n_{ik1}$, (matched-pair design strategies using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the four performance measures: the average bias $\mathrm{E}[\hat{\Delta}^* - \Delta_0]$ =$\mathrm{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\mathrm{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\mathrm{E}[\tilde{sd}(\hat{\Delta}^*)]$.

7.4 Summary and Concluding Remarks

The purpose of this chapter was to make an investigation into the performance of the design strategies that were suggested in Chapter 4. The assessments of the performance of design strategies took into consideration contributions of two sources of variability: the approximation error $\tilde{\Delta}^* - \Delta_0$ and the sampling error $\hat{\Delta}^* - \tilde{\Delta}^*$ when the mean bias in the approximation error is zero. When the mean bias of the approximation error is non-zero, four performance measures are considered: the variability and the average bias of the total error, the variability of the approximation error and the average variability of the sampling error.

The investigation into the performance of the complete and partial randomised design strategies showed there was no major advantage from using partial randomised design and hence complete randomisation is sufficient. On the other hand the investigation into the performance of the matched-pair design strategies revealed the usefulness of the social grades in reducing the variability of the approximation error.

Multiple truth instances were used in the investigations. There were many instances where matching pairs by social grades was more advantageous. The most interesting truth is $\gamma\beta\delta-$case where none of the parameters in the true model is zero. The benefits of the social grade based designs are very apparent in the instances where the differences between strata in β_k are opposite to those in δ_k and γ_k; i.e. sign of c_2 is opposite to the signs of c_1 and c_3. The non-zero bias obtained in $\gamma\beta\delta-$case and $\beta\delta-$case restricted the adequacy of the applied model.

The findings showed that spatial designs based on the nearest neighbour matching algorithms have some of the benefits of social-grade based designs. This is good because it will not be known in practice which covariate affects purchasing and spatial designs may act as a proxy for covariate-based designs. However, the bias issues for the applied model in the more complicated truth scenarios seem to suggest that there is a bigger underlying problem which is not solved even by covariate-based designs.

Conclusion

In this thesis, we addressed the estimation of the effectiveness of online geographically-based advertising campaigns when geo-experiments are applied in the presence of unobserved heterogeneity within geos. The main contribution of this research is the development of a theoretical framework based on the theory of maximum likelihood estimation of misspecified model for quantifying the error in estimated advertising effect that resulting from fitting a statistical model that known to be wrong due to the presence of unobserved covariates.

We began the thesis by understanding that measuring the effectiveness of online advertising campaigns using geo-experiments relies on spatial and temporal targeting components, interrelated elements for setting campaigns and the quality of tracked data. In this study, the focus was to investigate the potential impact of the absence of some spatial components such as demographics, social-grades and socio-economics on estimating the campaign effect. Therefore, in Chapter 3, we built a mapping algorithm that link Google AdWords target cities for the UK to some spatial characteristics using local authority areas and micro-census data. The AdWords target cities were mapped to one of the local authority areas using the shortest great-circle distance between them. The returned set were combination of local authority areas that were AdWords target cities and local authority areas that included a set of AdWords target cities. We called the returned set of areas spatial units. Although we obtained a large set of spatial units for the whole UK, the only spatial units in England and Wales were considered in this study. The algorithm helped to create a well- defined frame for advertising targeting geos in UK, which - to the best of our knowledge - has not been discussed in the literature of online advertising or geo-experiments. The previous studies were applied to the DMAs, which are designated market areas in the US. The other countries in the world, however, have no determined

marketing areas in general, but they may have a limited number of media regions such as the TV regions in the UK, which are usually not sufficient to apply geo-experiments. Thus, countries with defined governmental, administrative, or media areas can apply the linking algorithm to understand their AdWords targeting units. Alongside the benefits of the linking algorithm, however, there was a concern over the functionality of the governmental units as marketing areas.

In Chapter 4, we proposed a conceptual model of geo-experiments consisting of a two-level model: online search and online purchase, where purchase decision happens after successful search. A list of assumptions was provided to make the conceptual model possible. The number of searches and purchases were discussed under two conditions: homogeneous strata within spatial units (no strata) and heterogeneous strata within spatial units (with strata). If we assumed homogeneity, the number of searches and purchases were set to be both spatial and time dependent, i.e. s_{it} and y_{it}, but if we assumed heterogeneity, then $s_{it} = \sum_k s_{ikt}$ and $y_{it} = \sum_k y_{ikt}$; i.e. they were summed implicitly over some unobserved strata in a spatial unit. Assuming homogeneity, s_{it} and y_{it} were taken to be Binomial random variables and logit-linear regression models were used to fit them. The logit-linear regression models were used as well to fit s_{it} and y_{it} when heterogeneity was assumed. We used the term applied model to refer to the purchase model under homogeneity and the term truth to refer to it under heterogeneity. The applied model was parametrised by θ^* including spatial effect, temporal change effect and campaign effect. The truth was parametrised by θ including spatial effect, population-stratum effect, temporal change effect and campaign effect, where all parameters were strata based. The applied model was then known to be misspecified and hence the estimated applied model parameters $\hat{\theta}^*$ would be incorrect. The theory of maximum likelihood estimation under misspecification and the Kullback-Leibler divergence were therefore used in Chapter 5 to study the implications of unobserved covariates for inferences about estimated effects for geo-experiments. The Kullback-Leibler divergence was employed to measure the distance between the applied and the truth probability functions to find the nearest parameters to the truth; the parameter vector that minimises the Kullback-Leibler divergence function between the applied model and the truth.

However, when heterogeneity was assumed, the joint truth probability structure of y_{it} was cumbersome to find and difficult to compute because it was sum of k independent non-

identical Binomial random variables, which constrained the computation of the Kullback-Leibler divergence. To overcome this issue, we proposed a proxy probability structure of the applied model that linking the applied model and the truth. The proxy probability structure was similar to the truth except that the probability of purchasing in the proxy was identical between strata. The independence property of y_{it} was used then to compute the Kullback-Leibler divergence. The best parameter vector to use of θ^* in the applied model and that minimises the Kullback-Leibler divergence was denoted by $\tilde{\theta}^*$. It was shown that maximising likelihood is asymptotically equivalent to minimising the Kullback-Leibler divergence and so for large enough sample, $\hat{\theta}^*$ might be expected to converge to $\tilde{\theta}^*$. The consistency and asymptotic normality of $\hat{\theta}^*$ under misspecification were studied based on the results obtained by White 1982 and Chow 1984.

The asymptotic distribution theory of $\hat{\theta}^*$ was presented in Chapter 5 for two structures of purchase model: a one-stage model and a two-stage model. In a one-stage model, we assumed the number of searches is known in each stratum in each spatial unit for each time period. On the other hand in a two-stage model, the purchase process was given by search process, i.e. the number of searches is random. In both model structures, the asymptotic variance of $\hat{\theta}^*$ were expressed by search term, but in a two-stage model was expressed by the expected number and variance of searches instead of known fixed number of searches. According to the literature on the maximum likelihood estimation under misspecification, the asymptotic distribution theory was derived at $\tilde{\theta}^*$.

Given that the truth model was strata based, the measure of the effectiveness of the campaign might not be a functional from a business perspective. Therefore, in section 4.6 we introduced overall measures of campaign effect in terms of sales: Δ for the truth in terms of θ and Δ^* for the applied model in terms of θ^*. Given the asymptotic Normal distribution of $\hat{\theta}^*$, the asymptotic distribution of $\hat{\Delta}^*$ has been found in Chapter 5, too, using delta method. In addition the proxy effect $\tilde{\Delta}^*$ in terms of $\tilde{\theta}^*$ was introduced.

In Chapter 6, we tested the applicability of the theory by assessing its performance in a variety of contexts in comparison to Monte Carlo simulations. We built a computational algorithm to validate the theoretical asymptotic distribution of the estimates of the purchase applied model parameters assuming that search process outcomes are known. The algorithms involved list of procedures including: data structures, design matrices, truth instances and campaign designs that some of which relied on predefined data that were

discussed in section 4.7 and others of which specified in section 6.1. For example, truth spatial parameters were identified based on real data but the other truth parameters were specified in a way that generate interesting instances of truth parameters. Different truth instances were suggested that give different scales of change in the probability of purchasing between strata within spatial units, provided that the overall change is neutral. The validity of the theoretical asymptotic distribution of the estimated applied model parameters was discussed across the proposed truth instances using 10 randomly selected campaign designs based on social-grades covariate and different number of searches including sample micro-census individuals in each stratum in a spatial unit n_{ik}, $10\%(n_{ik})$, $1\%(n_{ik})$ and $0.1\%(n_{ik})$. Most truth instances reported in this chapter corroborate the ability of theory to describe the sampling distribution of the estimates of the applied model parameters. Violations of the theoretical distributions were only found in few cases that related to using large difference between strata or large numbers of search.

In Chapter 7, we addressed the question: which campaign design strategy and truth instance return typically a better estimate of the effectiveness of the advertising campaign, i.e. $\hat{\Delta}^*$. We investigated the performance of different design strategies across different truth instances. Different design strategies were suggested in Chapter 4 to allocate part of spatial units to serve the modified advertising campaign during the second time period. This includes complete randomisation and matched-pair designs, where the pairs were matched using population, social grades, realistic expected search rates and the nearest neighbour algorithm which used dissimilarity measures between social-grades, dissimilarity measures between population and social grades and distances between geographical coordinates. The assessments of designs strategies' performance took into consideration contributions of two sources of variability: the approximation error $\tilde{\Delta}^* - \Delta_0$ and the sampling error $\hat{\Delta}^* - \tilde{\Delta}^*$. The investigation into the performance of the matched-pair design strategies revealed the usefulness of the social grades and the spatial proximity in reducing the variability of the approximation error. This is good because it will not be known in practice which covariate affects purchasing and spatial designs may act as a proxy for covariate-based designs. However, for the more complex truth instances, none of the design strategies were successful in avoiding bias due to unobserved heterogeneity.

While we believe that this thesis provides a substantial contribution to the online geographically-based advertising campaigns literature, the study was limited in several ways. First, the

conceptual model of geo-experiments was not designed to measure the return on ad-spend (ROAS), which is the incremental impact the ad-spend had on the response metric, due to lack of information on it. More practicality could be given to the conceptual model if coupled with the ad-spend. The modified advertising campaigns during the second time period cause changes in the ad-spend which are important to investigate the significant changes on the response metrics. Second, the conceptual model was based on the two-level model: online search and online purchase. The actual model however could be close to the hierarchical model outlined in **Figure 4.1**. The advertising campaigns can affect multiple behaviours where each has the potential to generate differences in purchasing. Considerably more work will need to be done to investigate the effects of multiple outcomes of geo-experiments. Third, the investigation of the applicability of the theoretical distribution for estimation of θ^* was limited to a one-stage model, some specified truth parameters and a number of selected social-grades based campaign designs. There is, therefore, a definite need for assessing the two-stage model. Also, future trials should assess the impact of different campaign design strategies or truth parameters where the overall difference between strata is not centred to zero.

Despite the limitations, this study is the first attempt for addressing generation of geos, designing covariate-based campaigns, and handling unobserved heterogeneity within geos in estimating treatment effects of the geo-experiments. In addition, the conceptual model was a good start point for understanding the complexity of behavioural structure of response variables of the geo-experiments and providing insights into the estimation of advertising campaigns under misspecification.

Appendix A

Abbreviations and Symbols

Abbreviations

PPC	pay per click
CPC	cost per click
CTR	click through rate
CPM	cost per thousand impressions
CPA	cost per action
DMAs	designated market areas
CPV	cost per view
ROI	return on investment
CR	conversion rate
IAC	incremental ad clicks
ROAS	return on ad-spend
ACE	AdWords campaign experiment

Symbols

ikt	unit referred to a stratum k in a spatial unit i and time period t

it	unit referred to a spatial unit i and time period t
N_{ik}	size of the population in a stratum k in a spatial unit i
N_i	size of the population in a spatial unit i
S_{ikt}	search process in a unit ikt
S_{it}	search process in a unit it
φ_{ikt}	probability of searching in a unit ikt
φ^*_{it}	probability of searching in a unit it
ζ_{ikt}	search model, assuming heterogeneity within spatial units
ζ_{it}	search model, assuming homogeneity within spatial units
C_{it}	advertising campaign status
Y_{ikt}	purchase process in a unit ikt
Y_{it}	purchase process in a unit it
p_{ikt}	probability of purchasing in a unit ikt
p^*_{it}	probability of purchasing in a unit it
η_{ikt}	truth
η^*_{it}	applied model
X	truth design matrix
X^*	applied design matrix
θ	truth parameter
θ^*	applied model parameter
α^*_i	spatial effect in the applied model
β^*	temporal change effect in the applied model
δ^*	campaign effect in the applied model
γ_k	stratum effect γ_k
$y^{c^1}_{it}$	total number of sales if the new campaign are served in all spatial units

$y_{it}^{c^0}$	total number of sales if none of spatial units selected to serve the new campaign
g	true probability structure
f	misspecified probability structure
$D_{KL}(g\|f)$	Kullback-Leibler divergence from f to g
$\tilde{\theta}^*$	parameter that gives the nearest model to the truth
$\hat{\theta}^*$	maximum likelihood estimate
$\eta_{ikt}^{\dagger}$	proxy model for the applied structure
H	transformation matrix that maps the structure of θ^* to the structure of θ.
$\mathcal{C}(\tilde{\theta^*})$	asymptotic variance-covariance matrix of $\hat{\theta}^*$
Δ	overall true effect
Δ^*	overall applied model effect
$\tilde{\Delta}^*$	overall proxy effect
$\widetilde{\mathrm{Var}}(\hat{\Delta}^*))$	asymptotic variance-covariance matrix of the estimate of $Delta^*$

Appendix B

Spatial Units

Table B.1: spatial units: Mapping Google advertising geos with grouped local authorities

spatial unit code	spatial unit label
1	Stockton-on-Tees, Hartlepool
2	Middlesbrough
3	Redcar and Cleveland
4	County Durham
5	Darlington
8	Halton
9	Warrington
10	Blackburn with Darwen
11	Blackpool
12	Kingston upon Hull, City of
13	East Riding of Yorkshire
14	North East Lincolnshire
15	North Lincolnshire
16	York, Selby
17	Derby
18	Leicester
19	Melton, Harborough, Rutland
20	Nottingham
21	Herefordshire, County of
22	Telford and Wrekin
23	Stoke-on-Trent
24	Bath and North East Somerset
25	Bristol, City of
26	North Somerset
27	South Gloucestershire
28	Plymouth

Table B.1 – continued from previous page

spatial unit code	spatial unit label
29	Torbay
30	Bournemouth
31	Poole
32	Swindon
33	Peterborough
34	Luton
35	Southend-on-Sea
36	Thurrock
37	Medway
38	Slough, Bracknell Forest
39	West Berkshire
40	Reading
41	Windsor and Maidenhead
42	Wokingham
43	Milton Keynes
44	Brighton and Hove
45	Portsmouth
46	Southampton
47	Isle of Wight
48	Northumberland
50	Cheshire West and Chester
52	Cheshire East
54	Shropshire
56	Cornwall,Isles of Scilly
59	Wiltshire
62	Bedford
63	Central Bedfordshire
64	Aylesbury Vale
65	South Bucks, Chiltern
66	Wycombe
67	Cambridge
68	Fenland, East Cambridgeshire
69	Huntingdonshire
70	South Cambridgeshire
71	Carlisle, Allerdale
72	Copeland, Barrow-in-Furness
73	South Lakeland, Eden
74	North East Derbyshire, Amber Valley
75	Chesterfield, Bolsover
76	High Peak, Derbyshire Dales
77	South Derbyshire, Erewash
78	Mid Devon, East Devon
79	Exeter
80	Torridge, North Devon
81	West Devon, South Hams

Table B.1 – continued from previous page

spatial unit code	spatial unit label
82	Teignbridge
83	Purbeck, East Dorset, Christchurch
84	Weymouth and Portland, West Dorset, North Dorset
85	Lewes, Eastbourne
86	Rother, Hastings
87	Wealden
88	Basildon
89	Uttlesford, Braintree
90	Harlow, Brentwood
91	Rochford, Maldon, Castle Point
92	Chelmsford
93	Colchester
94	Epping Forest
95	Tendring
96	Cotswold, Cheltenham
97	Stroud, Forest of Dean
98	Tewkesbury, Gloucester
99	Basingstoke and Deane
100	Havant, East Hampshire
101	Eastleigh
102	Gosport, Fareham
103	Rushmoor, Hart
104	New Forest
105	Winchester, Test Valley
106	East Hertfordshire, Broxbourne
107	Dacorum
108	Welwyn Hatfield, Hertsmere
109	Stevenage, North Hertfordshire
110	St Albans
111	Watford, Three Rivers
112	Tunbridge Wells, Ashford
113	Canterbury
114	Gravesham, Dartford
115	Shepway, Dover
116	Maidstone
117	Tonbridge and Malling, Sevenoaks
118	Swale
119	Thanet
120	Pendle, Burnley
121	West Lancashire , Chorley
122	Wyre, Fylde
123	Rossendale, Hyndburn
124	Lancaster
125	Preston
126	South Ribble, Ribble Valley

Table B.1 – continued from previous page

spatial unit code	spatial unit label
127	Oadby and Wigston, Blaby
128	Charnwood
129	North West Leicestershire, Hinckley and Bosworth
130	West Lindsey, South Holland, Boston
131	East Lindsey
132	Lincoln
133	South Kesteven, North Kesteven
134	Breckland
135	Broadland
136	North Norfolk, Great Yarmouth
137	King's Lynn and West Norfolk
138	Norwich
139	South Norfolk
140	Kettering, Corby
141	South Northamptonshire, Daventry
142	Wellingborough, East Northamptonshire
143	Northampton
144	Richmondshire, Hambleton, Craven
145	Harrogate
146	Scarborough, Ryedale
147	Mansfield, Ashfield
148	Newark and Sherwood, Bassetlaw
149	Rushcliffe, Gedling, Broxtowe
150	Cherwell
151	Oxford
152	South Oxfordshire
153	West Oxfordshire, Vale of White Horse
154	Sedgemoor, Mendip
155	South Somerset
156	West Somerset, Taunton Deane
157	South Staffordshire, Cannock Chase
158	Staffordshire Moorlands, East Staffordshire
159	Tamworth, Lichfield
160	Newcastle-under-Lyme
161	Stafford
162	Ipswich, Babergh
163	St Edmundsbury, Mid Suffolk, Forest Heath
164	Waveney, Suffolk Coastal
165	Epsom and Ewell, Elmbridge
166	Guildford
167	Waverley, Mole Valley
168	Tandridge, Reigate and Banstead
169	Spelthorne, Runnymede
170	Woking, Surrey Heath
171	Rugby, North Warwickshire

Table B.1 – continued from previous page

spatial unit code	spatial unit label
172	Nuneaton and Bedworth
173	Stratford-on-Avon
174	Warwick
175	Worthing, Adur
176	Arun
177	Horsham, Chichester
178	Mid Sussex, Crawley
179	Wyre Forest, Bromsgrove
180	Worcester, Malvern Hills
181	Wychavon, Redditch
182	Bolton
183	Bury
184	Manchester
185	Oldham
186	Rochdale
187	Salford
188	Stockport
189	Tameside
190	Trafford
191	Wigan
192	Knowsley
193	Liverpool
194	St. Helens
195	Sefton
196	Wirral
197	Barnsley
198	Doncaster
199	Rotherham
200	Sheffield
201	Gateshead
202	Newcastle upon Tyne
203	North Tyneside
204	South Tyneside
205	Sunderland
206	Birmingham
207	Coventry
208	Dudley
209	Sandwell
210	Solihull
211	Walsall
212	Wolverhampton
213	Bradford
214	Calderdale
215	Kirklees
216	Leeds

Table B.1 – continued from previous page

spatial unit code	spatial unit label
217	Wakefield
218	City of London,Westminster
219	Barking and Dagenham
220	Barnet
221	Bexley
222	Brent
223	Bromley
224	Camden
225	Croydon
226	Ealing
227	Enfield
228	Greenwich
229	Hackney
230	Hammersmith and Fulham
231	Haringey
232	Harrow
233	Havering
234	Hillingdon
235	Hounslow
236	Islington
237	Kensington and Chelsea
238	Kingston upon Thames
239	Lambeth
240	Lewisham
241	Merton
242	Newham
243	Redbridge
244	Richmond upon Thames
245	Southwark
246	Sutton
247	Tower Hamlets
248	Waltham Forest
249	Wandsworth
250	Isle of Anglesey, Gwynedd
251	Conwy, Denbighshire
252	Flintshire
253	Wrexham
254	Ceredigion, Pembrokeshire
255	Carmarthenshire
256	Swansea
257	Neath Port Talbot
258	Bridgend
259	The Vale of Glamorgan
260	Cardiff
261	Rhondda Cynon Taf

Table B.1 – continued from previous page

spatial unit code	spatial unit label
262	Caerphilly, Blaenau Gwent, Merthyr Tydfil
263	Torfaen, Monmouthshire
264	Newport
265	Powys

Appendix C

Primitive Data Structure

Table C.1: prior search and purchases matched with adjusted micro-census population

spatial unit code	Country	channel	search.t0	purchase.t0	search.t1	purchase.t1	micro sample individuals
1	ENG	PPC	40.00	15.00	27.00	7.00	11357
2	ENG	PPC	56.00	15.00	39.00	10.00	5521
3	ENG	PPC	47.00	16.00	23.00	6.00	5497
4	ENG	PPC	59.00	24.00	43.00	8.00	20188
5	ENG	PPC	30.00	14.00	34.00	7.00	5266
8	ENG	PPC	74.00	32.00	39.00	5.00	5033
9	ENG	PPC	76.00	33.00	39.00	9.00	8140
10	ENG	PPC	27.00	7.00	33.00	10.00	5610
11	ENG	PPC	17.00	9.00	19.00	6.00	5763
12	ENG	PPC	48.00	22.00	40.00	8.00	10353
13	ENG	PPC	39.00	16.00	21.00	9.00	13869
14	ENG	PPC	37.00	12.00	24.00	5.00	6345
15	ENG	PPC	20.00	9.00	16.00	4.00	6732
16	ENG	PPC	62.00	30.00	38.00	10.00	11645
17	ENG	PPC	60.00	17.00	39.00	8.00	9789
18	ENG	PPC	113.00	38.00	96.00	9.00	13074
19	ENG	PPC	54.00	22.00	48.00	11.00	6999
20	ENG	PPC	101.00	39.00	65.00	20.00	12520
21	ENG	PPC	25.00	9.00	17.00	5.00	7543
22	ENG	PPC	31.00	6.00	22.00	2.00	6567
23	ENG	PPC	56.00	24.00	38.00	7.00	9892
24	ENG	PPC	58.00	18.00	33.00	6.00	7366
25	ENG	PPC	136.00	43.00	82.00	21.00	17471
26	ENG	PPC	47.00	15.00	27.00	8.00	8140
27	ENG	PPC	109.00	46.00	76.00	17.00	10605

Continued on next page

Table C.1 – continued from previous page

spatial unit code	Country	channel	search.t0	purchase.t0	search.t1	purchase.t1	micro sample individuals
28	ENG	PPC	57.00	21.00	54.00	7.00	10415
29	ENG	PPC	44.00	22.00	30.00	7.00	5369
30	ENG	PPC	32.00	11.00	18.00	2.00	7663
31	ENG	PPC	40.00	15.00	17.00	4.00	5974
32	ENG	PPC	36.00	8.00	15.00	2.00	8439
33	ENG	PPC	70.00	26.00	71.00	8.00	7177
34	ENG	PPC	54.00	10.00	46.00	1.00	7823
36	ENG	PPC	69.00	28.00	37.00	2.00	6151
37	ENG	PPC	109.00	45.00	79.00	5.00	10459
38	ENG	PPC	100.00	21.00	71.00	4.00	9759
39	ENG	PPC	57.00	24.00	42.00	16.00	6076
40	ENG	PPC	70.00	17.00	42.00	2.00	6273
42	ENG	PPC	53.00	7.00	53.00	3.00	6169
43	ENG	PPC	56.00	8.00	47.00	4.00	9604
44	ENG	PPC	39.00	14.00	26.00	7.00	11404
45	ENG	PPC	128.00	60.00	77.00	19.00	8349
46	ENG	PPC	93.00	32.00	56.00	18.00	9799
47	ENG	PPC	24.00	11.00	18.00	4.00	5713
48	ENG	PPC	17.00	9.00	9.00	4.00	13084
50	ENG	PPC	37.00	10.00	26.00	4.00	13452
52	ENG	PPC	60.00	24.00	43.00	11.00	15166
54	ENG	PPC	29.00	9.00	19.00	5.00	12484
56	ENG	PPC	110.00	43.00	75.00	19.00	21966
59	ENG	PPC	66.00	12.00	50.00	6.00	18780
62	ENG	PPC	38.00	4.00	22.00	1.00	6189
63	ENG	PPC	35.00	14.00	20.00	3.00	10276
64	ENG	PPC	44.00	12.00	29.00	3.00	6892
66	ENG	PPC	48.00	1.00	41.00	1.00	6810
67	ENG	PPC	29.00	9.00	20.00	3.00	5371
68	ENG	PPC	70.00	32.00	51.00	8.00	7268
69	ENG	PPC	44.00	18.00	39.00	5.00	6752
70	ENG	PPC	14.00	2.00	15.00	1.00	5920
72	ENG	PPC	65.00	35.00	45.00	10.00	5698
73	ENG	PPC	56.00	16.00	27.00	4.00	6575
74	ENG	PPC	18.00	10.00	13.00	4.00	9006
75	ENG	PPC	68.00	29.00	43.00	13.00	7316
77	ENG	PPC	109.00	35.00	68.00	11.00	8365
78	ENG	PPC	59.00	23.00	51.00	10.00	8606
79	ENG	PPC	76.00	26.00	50.00	7.00	4951
80	ENG	PPC	26.00	14.00	17.00	5.00	6419
82	ENG	PPC	19.00	6.00	17.00	1.00	5050
83	ENG	PPC	54.00	18.00	59.00	12.00	7473
86	ENG	PPC	76.00	18.00	58.00	7.00	7393
87	ENG	PPC	28.00	8.00	19.00	6.00	6073
88	ENG	PPC	28.00	6.00	18.00	1.00	6901

Continued on next page

Table C.1 – continued from previous page

spatial unit code	Country	channel	search.t0	purchase.t0	search.t1	purchase.t1	micro sample individuals
89	ENG	PPC	89.00	26.00	61.00	4.00	9073
93	ENG	PPC	37.00	18.00	39.00	3.00	7037
94	ENG	PPC	26.00	9.00	20.00	2.00	4921
95	ENG	PPC	70.00	16.00	60.00	5.00	5620
96	ENG	PPC	44.00	13.00	32.00	5.00	8228
97	ENG	PPC	48.00	17.00	47.00	7.00	7924
98	ENG	PPC	42.00	16.00	35.00	7.00	8216
99	ENG	PPC	21.00	3.00	17.00	2.00	6693
100	ENG	PPC	93.00	42.00	49.00	13.00	9529
101	ENG	PPC	32.00	12.00	13.00	2.00	5084
102	ENG	PPC	39.00	7.00	37.00	4.00	7952
103	ENG	PPC	23.00	10.00	16.00	6.00	7244
104	ENG	PPC	15.00	6.00	11.00	1.00	7292
106	ENG	PPC	45.00	9.00	28.00	3.00	9204
107	ENG	PPC	21.00	4.00	19.00	3.00	5771
108	ENG	PPC	75.00	24.00	66.00	4.00	8518
109	ENG	PPC	81.00	21.00	52.00	7.00	8399
110	ENG	PPC	25.00	5.00	15.00	2.00	5478
112	ENG	PPC	88.00	46.00	51.00	5.00	9170
114	ENG	PPC	208.00	50.00	161.00	7.00	7899
116	ENG	PPC	28.00	8.00	20.00	1.00	6167
117	ENG	PPC	55.00	13.00	40.00	5.00	9307
118	ENG	PPC	58.00	25.00	37.00	4.00	5423
119	ENG	PPC	68.00	24.00	43.00	4.00	5365
120	ENG	PPC	75.00	28.00	57.00	15.00	7006
121	ENG	PPC	42.00	10.00	25.00	5.00	8820
122	ENG	PPC	18.00	9.00	11.00	2.00	7640
123	ENG	PPC	36.00	16.00	27.00	6.00	5872
124	ENG	PPC	24.00	8.00	13.00	2.00	5759
125	ENG	PPC	52.00	18.00	49.00	11.00	5691
127	ENG	PPC	38.00	13.00	27.00	2.00	6090
128	ENG	PPC	47.00	14.00	28.00	6.00	6851
129	ENG	PPC	75.00	25.00	56.00	8.00	8036
130	ENG	PPC	70.00	24.00	51.00	10.00	6293
131	ENG	PPC	58.00	21.00	37.00	9.00	5680
132	ENG	PPC	16.00	7.00	7.00	2.00	7549
133	ENG	PPC	59.00	23.00	51.00	7.00	9837
134	ENG	PPC	20.00	6.00	12.00	2.00	5402
136	ENG	PPC	94.00	25.00	54.00	2.00	8228
137	ENG	PPC	119.00	41.00	78.00	6.00	6060
138	ENG	PPC	27.00	9.00	23.00	2.00	5584
139	ENG	PPC	8.00	3.00	9.00	1.00	5006
140	ENG	PPC	88.00	24.00	54.00	4.00	6112
142	ENG	PPC	52.00	28.00	32.00	4.00	6406
143	ENG	PPC	107.00	45.00	88.00	8.00	8362

Continued on next page

Table C.1 – continued from previous page

spatial unit code	Country	channel	search.t0	purchase.t0	search.t1	purchase.t1	micro sample individuals
145	ENG	PPC	61.00	21.00	40.00	9.00	6308
146	ENG	PPC	34.00	9.00	49.00	20.00	6633
147	ENG	PPC	125.00	50.00	88.00	22.00	9064
148	ENG	PPC	82.00	32.00	48.00	14.00	9295
149	ENG	PPC	18.00	7.00	15.00	4.00	13622
151	ENG	PPC	33.00	12.00	23.00	5.00	6482
154	ENG	PPC	66.00	25.00	49.00	11.00	9067
155	ENG	PPC	22.00	9.00	12.00	1.00	6548
156	ENG	PPC	32.00	8.00	15.00	1.00	5909
157	ENG	PPC	66.00	33.00	38.00	6.00	8415
158	ENG	PPC	81.00	36.00	51.00	17.00	8609
159	ENG	PPC	153.00	62.00	128.00	42.00	7272
160	ENG	PPC	80.00	35.00	53.00	10.00	5131
161	ENG	PPC	24.00	8.00	19.00	4.00	5417
162	ENG	PPC	134.00	66.00	90.00	7.00	8842
165	ENG	PPC	112.00	16.00	90.00	5.00	8040
166	ENG	PPC	36.00	5.00	26.00	1.00	5601
172	ENG	PPC	52.00	19.00	29.00	4.00	4994
173	ENG	PPC	15.00	4.00	7.00	1.00	4902
174	ENG	PPC	49.00	14.00	32.00	7.00	5640
175	ENG	PPC	27.00	12.00	25.00	4.00	6697
176	ENG	PPC	25.00	11.00	16.00	3.00	6176
177	ENG	PPC	30.00	12.00	28.00	8.00	9928
179	ENG	PPC	65.00	21.00	48.00	5.00	7844
180	ENG	PPC	48.00	25.00	29.00	8.00	7043
181	ENG	PPC	49.00	19.00	35.00	8.00	8123
182	ENG	PPC	84.00	32.00	52.00	6.00	10818
184	ENG	PPC	91.00	28.00	78.00	18.00	20426
185	ENG	PPC	26.00	14.00	17.00	2.00	8636
186	ENG	PPC	19.00	7.00	14.00	5.00	8269
187	ENG	PPC	85.00	25.00	47.00	11.00	9422
188	ENG	PPC	96.00	37.00	64.00	15.00	11408
189	ENG	PPC	54.00	21.00	31.00	16.00	8729
190	ENG	PPC	51.00	18.00	41.00	11.00	8993
191	ENG	PPC	37.00	20.00	37.00	7.00	12771
192	ENG	PPC	53.00	19.00	23.00	8.00	5819
193	ENG	PPC	94.00	37.00	60.00	12.00	19526
194	ENG	PPC	70.00	40.00	45.00	9.00	7110
195	ENG	PPC	64.00	18.00	66.00	16.00	11203
196	ENG	PPC	81.00	21.00	41.00	7.00	12900
197	ENG	PPC	57.00	19.00	41.00	15.00	9340
198	ENG	PPC	41.00	20.00	30.00	8.00	12156
199	ENG	PPC	74.00	27.00	51.00	22.00	10286
200	ENG	PPC	91.00	26.00	46.00	13.00	22486
201	ENG	PPC	78.00	31.00	45.00	10.00	8171

Continued on next page

Table C.1 – continued from previous page

spatial unit code	Country	channel	search.t0	purchase.t0	search.t1	purchase.t1	micro sample individuals
202	ENG	PPC	57.00	25.00	45.00	13.00	11672
203	ENG	PPC	119.00	40.00	72.00	17.00	8189
204	ENG	PPC	31.00	9.00	16.00	4.00	6039
205	ENG	PPC	152.00	63.00	120.00	4.00	11237
206	ENG	PPC	177.00	69.00	164.00	30.00	41160
207	ENG	PPC	83.00	28.00	41.00	6.00	12615
208	ENG	PPC	129.00	43.00	101.00	23.00	12563
209	ENG	PPC	76.00	23.00	62.00	18.00	12103
210	ENG	PPC	78.00	30.00	81.00	21.00	8307
211	ENG	PPC	124.00	56.00	84.00	18.00	10587
212	ENG	PPC	87.00	32.00	54.00	18.00	9967
213	ENG	PPC	198.00	75.00	159.00	46.00	19965
214	ENG	PPC	65.00	23.00	44.00	17.00	8140
215	ENG	PPC	70.00	19.00	66.00	17.00	16702
216	ENG	PPC	36.00	11.00	35.00	9.00	30701
217	ENG	PPC	67.00	30.00	52.00	13.00	13148
218	ENG	PPC	51.00	11.00	34.00	1.00	9935
222	ENG	PPC	112.00	33.00	73.00	4.00	12433
225	ENG	PPC	59.00	20.00	59.00	5.00	14114
226	ENG	PPC	65.00	25.00	74.00	3.00	13529
227	ENG	PPC	103.00	40.00	76.00	3.00	12027
228	ENG	PPC	121.00	46.00	64.00	6.00	9973
232	ENG	PPC	65.00	12.00	52.00	2.00	9504
236	ENG	PPC	171.00	53.00	123.00	7.00	8794
237	ENG	PPC	55.00	19.00	42.00	5.00	6835
238	ENG	PPC	34.00	3.00	34.00	1.00	6541
239	ENG	PPC	130.00	47.00	105.00	9.00	12454
244	ENG	PPC	38.00	3.00	20.00	3.00	7488
246	ENG	PPC	63.00	46.00	37.00	4.00	7501
247	ENG	PPC	327.00	106.00	309.00	11.00	10411
250	WLS	PPC	56.00	14.00	48.00	6.00	7945
251	WLS	PPC	58.00	25.00	46.00	12.00	8535
252	WLS	PPC	29.00	14.00	19.00	3.00	6138
253	WLS	PPC	23.00	9.00	15.00	4.00	5465
254	WLS	PPC	63.00	22.00	56.00	11.00	8205
255	WLS	PPC	14.00	1.00	9.00	1.00	7475
256	WLS	PPC	80.00	41.00	66.00	7.00	9858
257	WLS	PPC	49.00	23.00	30.00	3.00	5648
258	WLS	PPC	68.00	18.00	47.00	3.00	5655
260	WLS	PPC	65.00	20.00	37.00	4.00	14194
261	WLS	PPC	39.00	19.00	28.00	3.00	9439
262	WLS	PPC	156.00	52.00	99.00	11.00	12323
263	WLS	PPC	89.00	22.00	73.00	13.00	7278
264	WLS	PPC	69.00	25.00	38.00	3.00	5746
265	WLS	PPC	35.00	10.00	25.00	4.00	5381

Appendix D

Campaign Designs

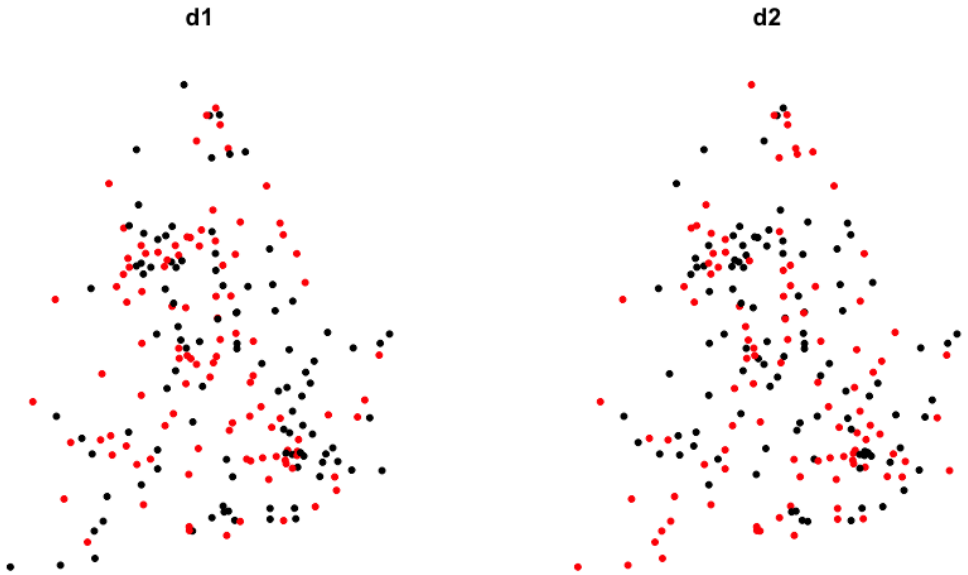


Figure D.1: campaign designs (treatment spatial units in red colour): generated by matched-pairs in term of social grades covariate and the truth parameter vector θ_0 based on $\delta-$case with$c = 5$

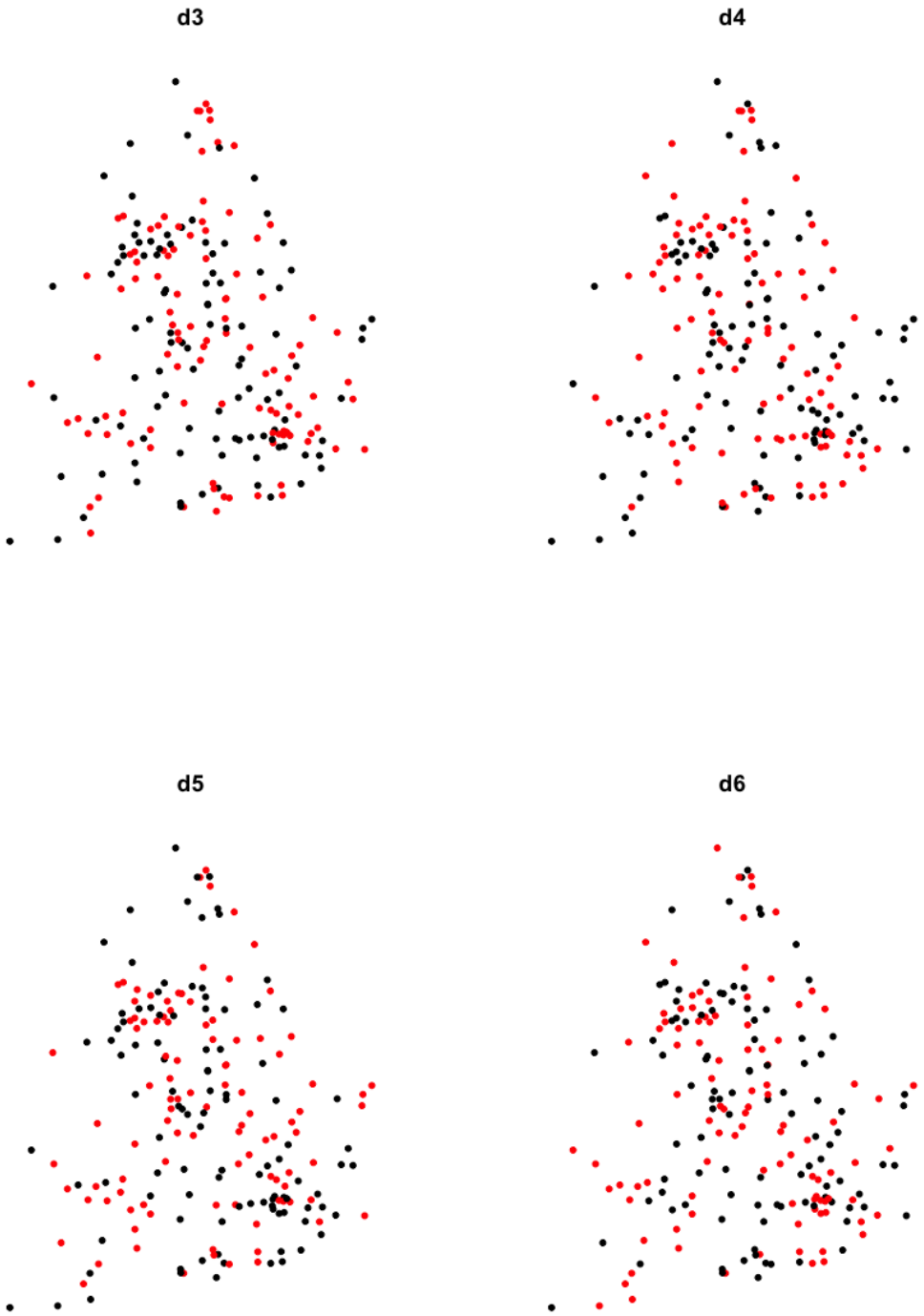


Figure D.2: campaign designs (treatment spatial units in red colour): generated by matched-pairs in term of social grades covariate and the truth parameter vector θ_0 based on $\delta-$case with$c = 5$

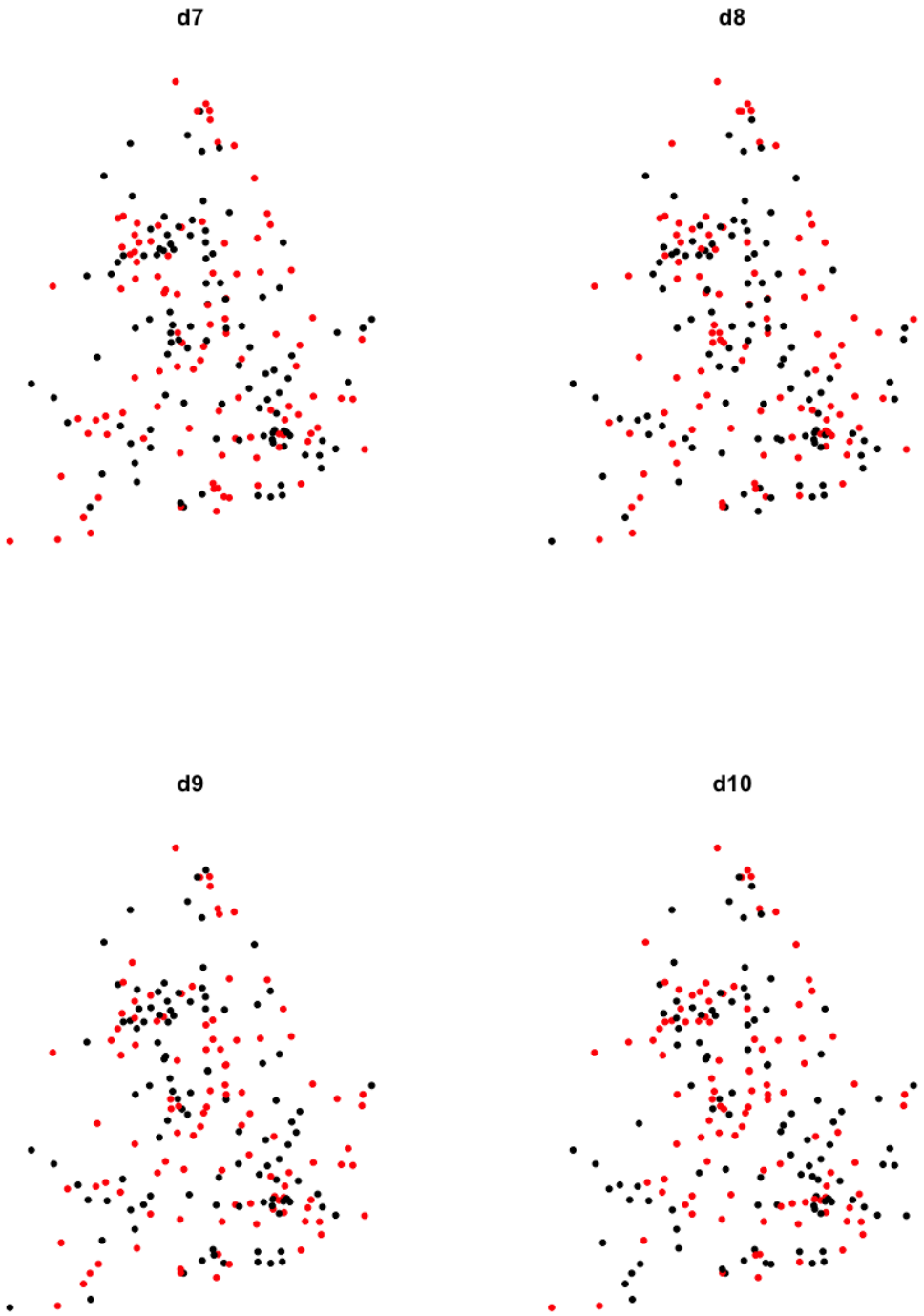


Figure D.3: campaign designs (treatment spatial units in red colour): generated by matched-pairs in term of social grades covariate and the truth parameter vector θ_0 based on $\delta-$case with$c = 5$

Appendix E

Data Structures Used in Computational Algorithms

truth

area	group	time	n	geo	ad
la_group99	1	0	170	1	0
la_group99	2	0	164	1	0
la_group99	3	0	110	1	0
la_group99	4	0	100	1	0
la_group99	1	1	170	1	0
la_group99	2	1	164	1	0
la_group99	3	1	110	1	0
la_group99	4	1	100	1	0
la_group98	1	0	141	2	0
la_group98	2	0	207	2	0
la_group98	3	0	153	2	0
la_group98	4	0	146	2	0
la_group98	1	1	141	2	1
la_group98	2	1	207	2	1
la_group98	3	1	153	2	1
la_group98	4	1	146	2	1

applied

area	time	geo	ad	n
la_group99	0	1	0	544
la_group99	1	1	0	544
la_group98	0	2	0	647
la_group98	1	2	1	647

Figure E.1: Screenshot of subset of truth and applied data structure with campaign design: area≡spatial units, group≡strata, ad≡ C_{it}, n ≡number of search, geo≡spatial unit condition where 1 =control and 2 =treatment.

truth^{c1}

area	group	time	n	geo	ad
la_group99	1	0	170	1	1
la_group99	2	0	164	1	1
la_group99	3	0	110	1	1
la_group99	4	0	100	1	1
la_group99	1	1	170	1	1
la_group99	2	1	164	1	1
la_group99	3	1	110	1	1
la_group99	4	1	100	1	1
la_group98	1	0	141	2	1
la_group98	2	0	207	2	1
la_group98	3	0	153	2	1
la_group98	4	0	146	2	1
la_group98	1	1	141	2	1
la_group98	2	1	207	2	1
la_group98	3	1	153	2	1
la_group98	4	1	146	2	1

applied^{c1}

area	time	geo	n	ad
la_group99	0	1	544	1
la_group99	1	1	544	1
la_group98	0	2	647	1
la_group98	1	2	647	1

(a) Example of truth and applied data structure with ad in all spatial units

truth^{c0}

area	group	time	n	geo	ad
la_group99	1	0	170	1	0
la_group99	2	0	164	1	0
la_group99	3	0	110	1	0
la_group99	4	0	100	1	0
la_group99	1	1	170	1	0
la_group99	2	1	164	1	0
la_group99	3	1	110	1	0
la_group99	4	1	100	1	0
la_group98	1	0	141	2	0
la_group98	2	0	207	2	0
la_group98	3	0	153	2	0
la_group98	4	0	146	2	0
la_group98	1	1	141	2	0
la_group98	2	1	207	2	0
la_group98	3	1	153	2	0
la_group98	4	1	146	2	0

applied^{c0}

area	time	geo	n	ad
la_group99	0	1	544	0
la_group99	1	1	544	0
la_group98	0	2	647	0
la_group98	1	2	647	0

(b) Example of truth and applied data structure with no ad in all spatial units

Figure E.2: Screenshot of subset of truth and applied data structure with no campaign design: area≡spatial units, group≡strata, ad≡ C_{it}, n ≡number of search, geo≡spatial unit condition where 1 =control and 2 = treatment.

Appendix F

Spatial Effects Values used to Specify α_i

Table F.1: specification of truth spatial effects α_i using estimate $\hat{\alpha_i^*}$ obtained by fitting realistic search and purchase.

i	$\alpha_i = \hat{\alpha_i^*}$	i	$\alpha_i = \hat{\alpha_i^*}$	i	$\alpha_i = \hat{\alpha_i^*}$	i	$\alpha_i = \hat{\alpha_i^*}$
la_group1	-0.46	la_group23	-0.45	la_group45	-0.23	la_group77	-0.81
la_group2	-0.78	la_group24	-0.81	la_group46	-0.45	la_group78	-0.56
la_group3	-0.58	la_group25	-0.65	la_group47	-0.31	la_group79	-0.55
la_group4	-0.52	la_group26	-0.57	la_group48	0.53	la_group80	0.35
la_group5	-0.38	la_group27	-0.40	la_group50	-1.00	la_group82	-0.86
la_group8	-0.50	la_group28	-0.47	la_group52	-0.40	la_group83	-0.69
la_group9	-0.34	la_group29	0.13	la_group54	-0.64	la_group86	-0.97
la_group10	-0.58	la_group30	-0.83	la_group56	-0.43	la_group87	-0.61
la_group11	0.01	la_group31	-0.51	la_group59	-1.44	la_group88	-1.29
la_group12	-0.37	la_group32	-1.24	la_group62	-2.19	la_group89	-0.92
la_group13	-0.11	la_group33	-0.51	la_group63	-0.35	la_group93	-0.29
la_group14	-0.71	la_group34	-1.59	la_group64	-0.89	la_group94	-0.63
la_group15	-0.29	la_group36	-0.50	la_group67	-0.88	la_group95	-1.12
la_group16	-0.16	la_group37	-0.49	la_group68	-0.15	la_group96	-0.91
la_group17	-0.85	la_group38	-1.31	la_group69	-0.35	la_group97	-0.78
la_group18	-0.68	la_group39	-0.11	la_group70	-1.58	la_group98	-0.57
la_group19	-0.10	la_group40	-1.17	la_group72	0.22	la_group99	-1.39
la_group20	-0.35	la_group42	-1.71	la_group73	-0.77	la_group100	-0.24
la_group21	-0.44	la_group43	-1.52	la_group74	0.08	la_group101	-0.44
la_group22	-1.48	la_group44	-0.22	la_group75	-0.25	la_group102	-1.49

Table F.1 – continued from previous page

i	$\alpha_i = \hat{\alpha_i^*}$	i	$\alpha_i = \hat{\alpha_i^*}$	i	$\alpha_i = \hat{\alpha_i^*}$	i	$\alpha_i = \hat{\alpha_i^*}$
la_group103	-0.10	la_group145	-0.62	la_group191	-0.23	la_group236	-1.11
la_group104	-0.74	la_group146	-0.24	la_group192	-0.41	la_group237	-0.85
la_group106	-1.20	la_group147	-0.41	la_group193	-0.52	la_group238	-2.24
la_group107	-0.99	la_group148	-0.37	la_group194	-0.04	la_group239	-0.61
la_group108	-0.84	la_group149	-0.40	la_group195	-0.72	la_group244	-1.96
la_group109	-0.87	la_group151	-0.57	la_group196	-1.01	la_group246	0.25
la_group110	-1.13	la_group154	-0.52	la_group197	-0.37	la_group247	-0.91
la_group112	-0.05	la_group155	-0.44	la_group198	-0.16	la_group250	-1.16
la_group114	-1.21	la_group156	-1.25	la_group199	-0.18	la_group251	-0.31
la_group116	-0.99	la_group157	-0.28	la_group200	-0.72	la_group252	-0.35
la_group117	-0.97	la_group158	-0.15	la_group201	-0.46	la_group253	-0.40
la_group118	-0.33	la_group159	-0.24	la_group202	-0.24	la_group254	-0.35
la_group119	-0.62	la_group160	-0.14	la_group203	-0.62	la_group255	-1.95
la_group120	-0.46	la_group161	-0.67	la_group204	-0.75	la_group256	-0.11
la_group121	-1.02	la_group162	-0.20	la_group205	-0.57	la_group257	-0.22
la_group122	-0.25	la_group165	-1.67	la_group206	-0.59	la_group258	-1.03
la_group123	-0.35	la_group166	-1.79	la_group207	-0.77	la_group260	-0.75
la_group124	-0.57	la_group172	-0.49	la_group208	-0.64	la_group261	-0.16
la_group125	-0.61	la_group173	-1.04	la_group209	-0.58	la_group262	-0.65
la_group127	-0.71	la_group174	-0.81	la_group210	-0.43	la_group263	-0.75
la_group128	-0.78	la_group175	-0.17	la_group211	-0.34	la_group264	-0.61
la_group129	-0.83	la_group176	-0.41	la_group212	-0.36	la_group265	-0.69
la_group130	-0.41	la_group177	-0.33	la_group213	-0.39		
la_group131	-0.27	la_group179	-0.95	la_group214	-0.29		
la_group132	-0.25	la_group180	-0.04	la_group215	-0.72		
la_group133	-0.69	la_group181	-0.49	la_group216	-0.63		
la_group134	-0.65	la_group182	-0.48	la_group217	-0.29		
la_group136	-1.09	la_group184	-0.70	la_group218	-1.37		
la_group137	-0.69	la_group185	-0.27	la_group222	-0.93		
la_group138	-0.71	la_group186	-0.29	la_group225	-0.70		
la_group139	-0.50	la_group187	-0.76	la_group226	-0.71		
la_group140	-0.97	la_group188	-0.48	la_group227	-0.64		
la_group142	0.04	la_group189	-0.02	la_group228	-0.73		
la_group143	-0.41	la_group190	-0.50	la_group232	-1.51		

Appendix G

Campaign Design Procedure Using Different Design Strategies

procedure DESIGN STRATEGIES(spatial units, population, social grades, search rates, coordinates)

function COMPLETELY RANDOM (spatial units)

sample spatial units randomly to allocate in control or treatment group

return comp.random

end function

function MATCHED-PAIRS BASED ON COVARIATES (spatial units, population,social grades, search rates)

1. sort spatial units by their number of individuals

2. group spatial units into pairs

3. within each pair assign each spatial units randomly to treatment or control.

return pair.pop

4. sort spatial units by their weighted mean of social grades

5. repeat steps 2 and 3

return pair.socgr

6. sort spatial units by their expected searches

7. repeat steps 2 and 3

return pair.search

end function

function MATCHED-PAIRS BASED ON NEAREST NEIGHBOUR (spatial units, social grades, population, coordinates)

1. compute pairwise distance between spatial units using social grades

2. set a matrix showing distance between each pair of spatial units

3. select a spatial unit randomly and pair it with its a nearby spatial unit

4. remove the selected pairs from the distance matrix

5. repeat steps 3 and 4

6. pair the remaining unused two spatial units

7. within each pair assign each spatial units randomly to treatment or control.

return pair.d.socgr

8. compute pairwise distance between spatial units using social grades and population

9. repeat steps 2 to 7

return pair.d. socgrpop

10. compute pairwise distance between spatial units using geographic coordinates

11. repeat steps 2 to 7

return pair.d.coordinates

end function

end procedure

Appendix H

Chi-square Q-Q plot

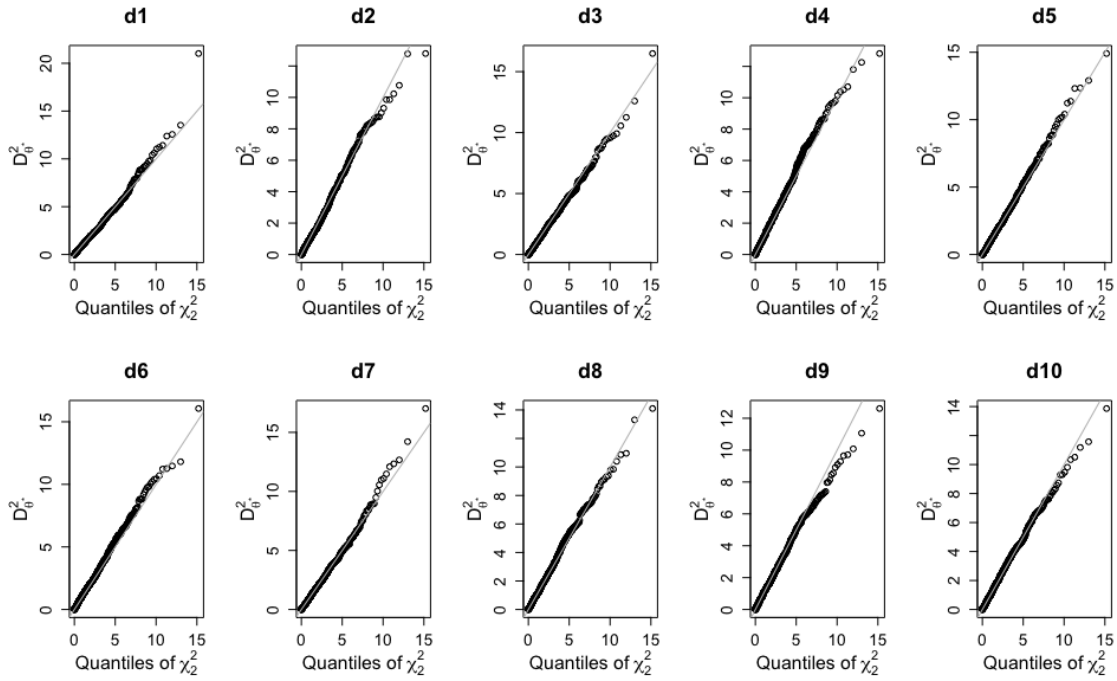
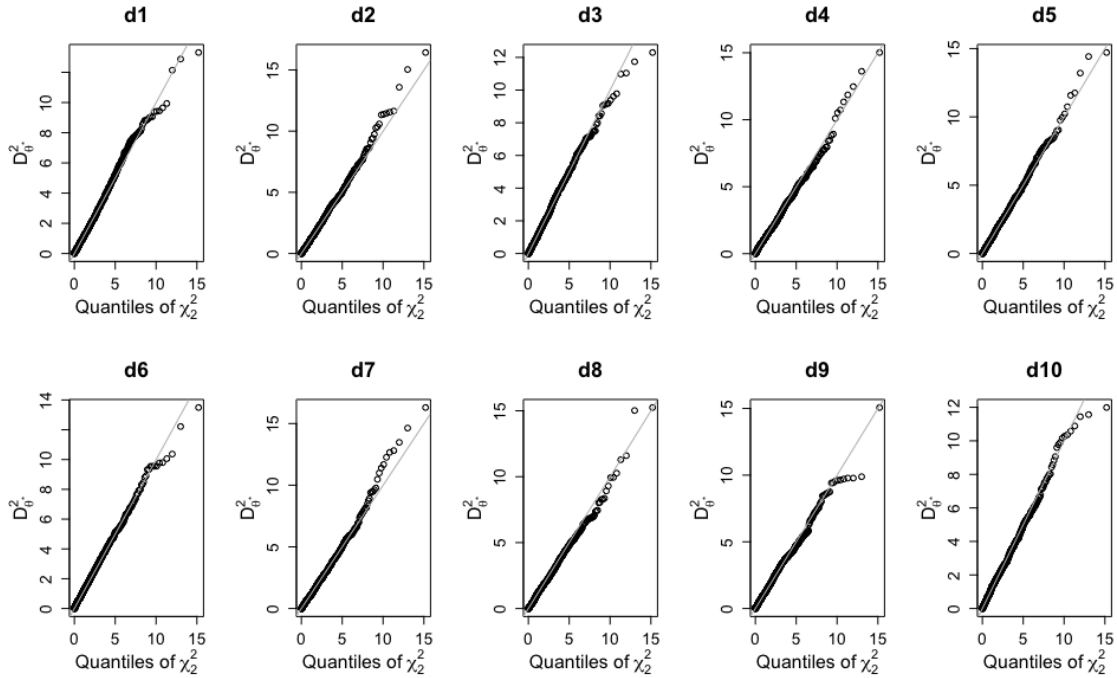
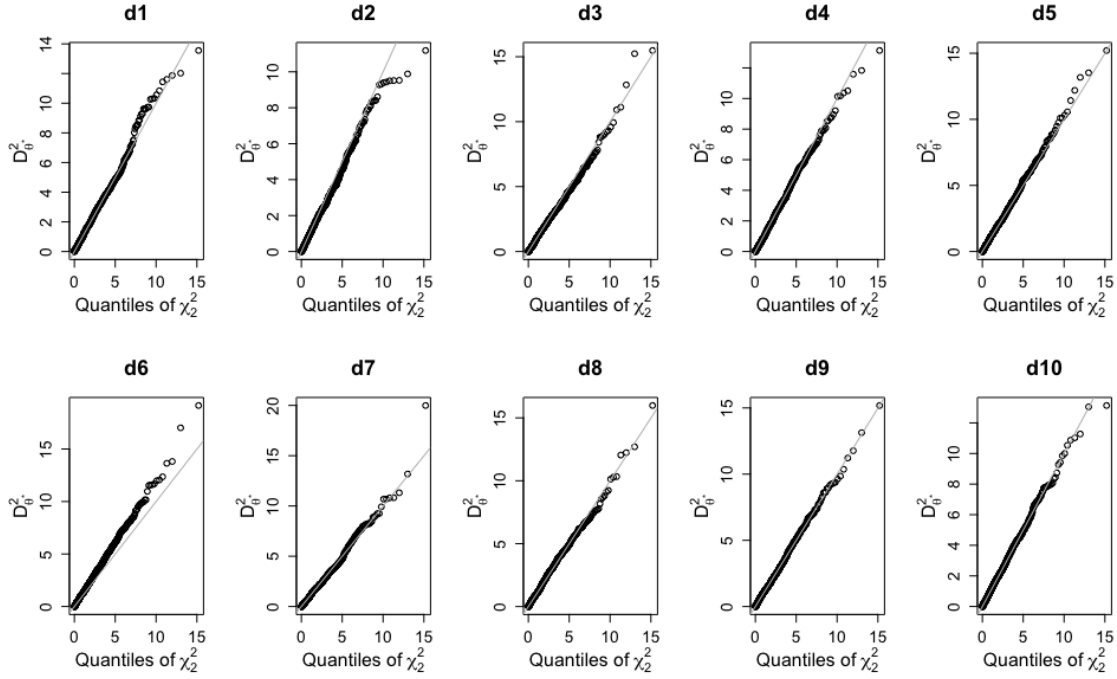


Figure H.1: $\beta-$case with $c = 0.2$: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\theta*}$ versus quantiles of χ^2_2

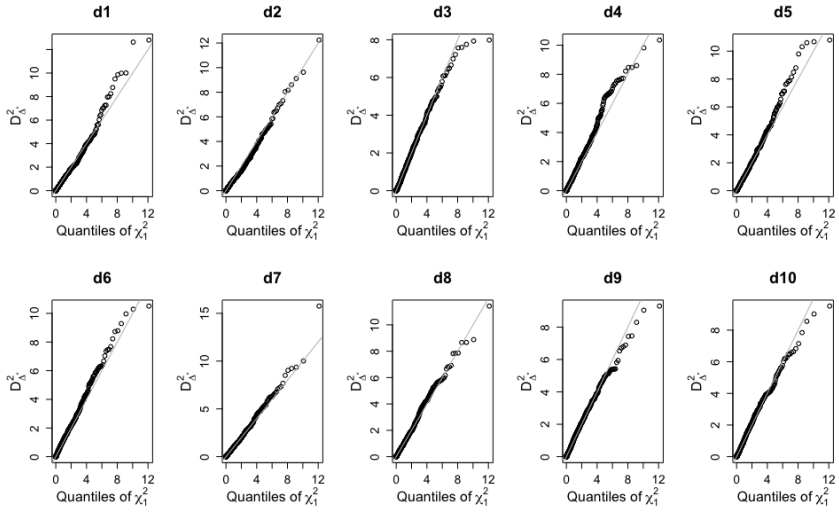


(a) c=1

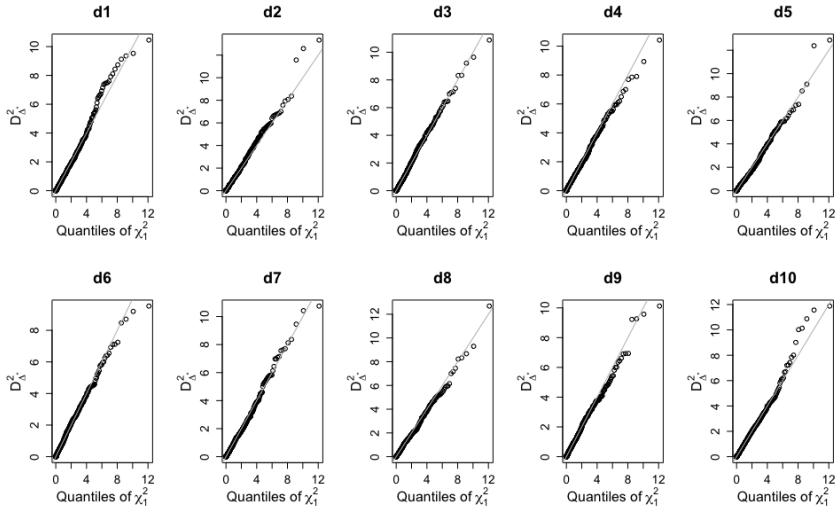


(b) c=5

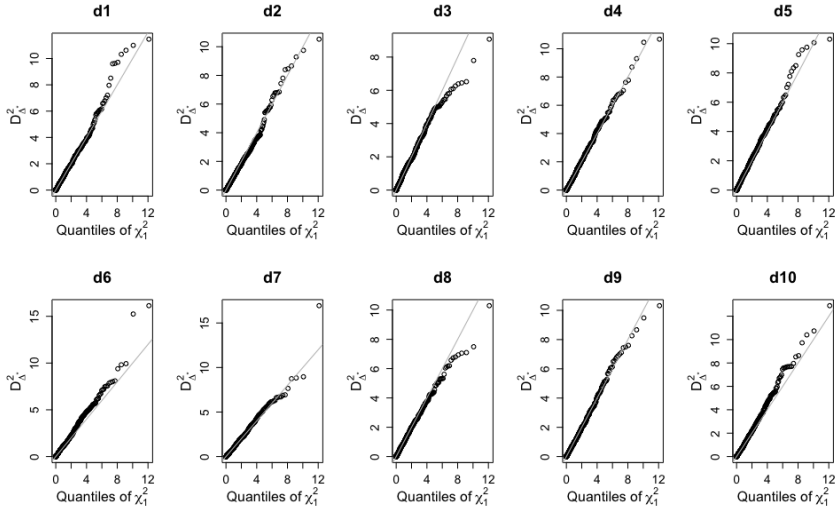
Figure H.2: $\beta-$case with $c = 1$ and $c = 5$: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\theta^*}$ versus quantiles of χ^2_2



(a) c=0.2

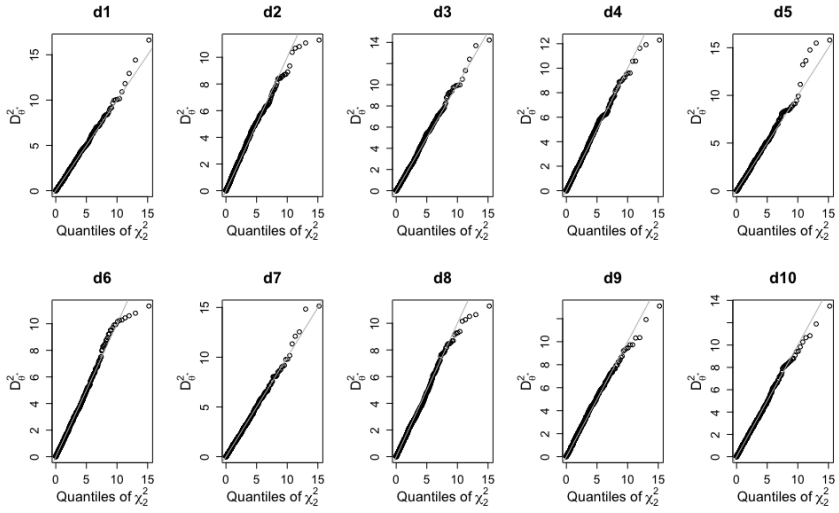


(b) c=1

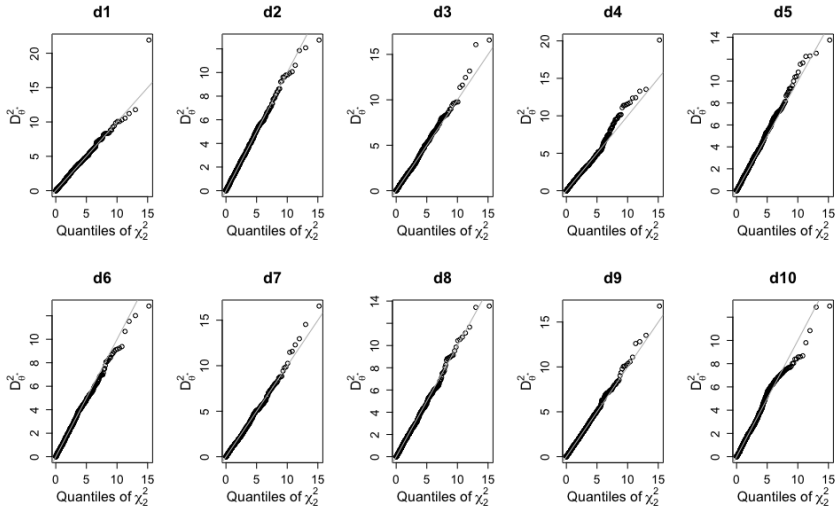


(c) c=5

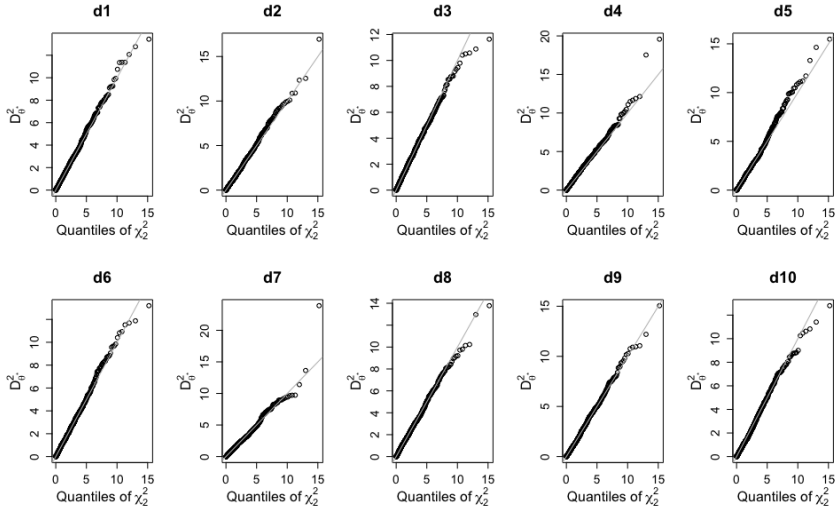
Figure H.3: $\beta-$case: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ versus quantiles of χ^2_1



(a) c=0.2

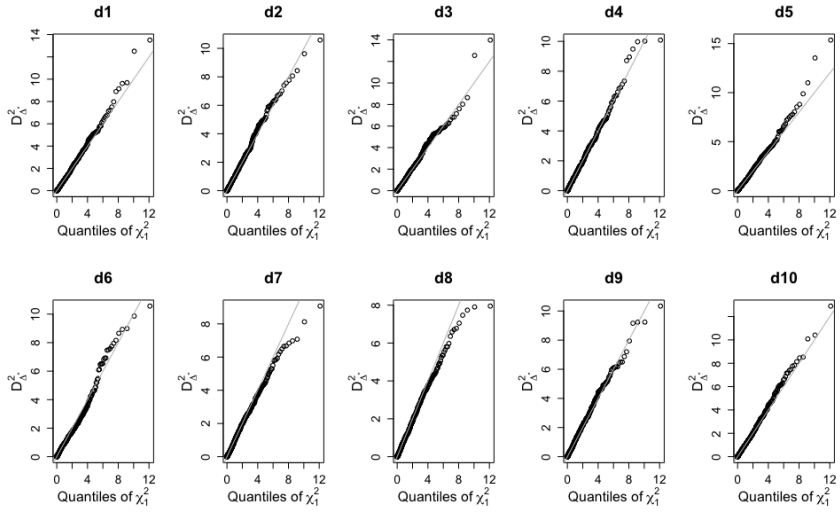


(b) c=1

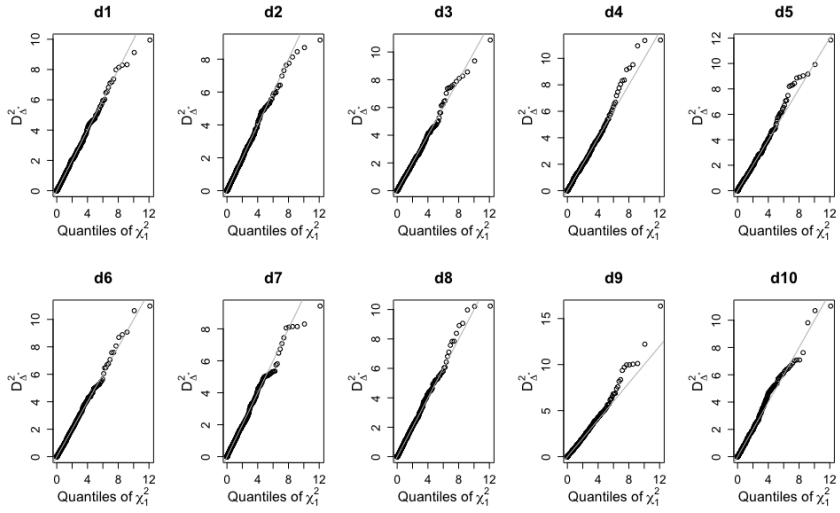


(c) c=5

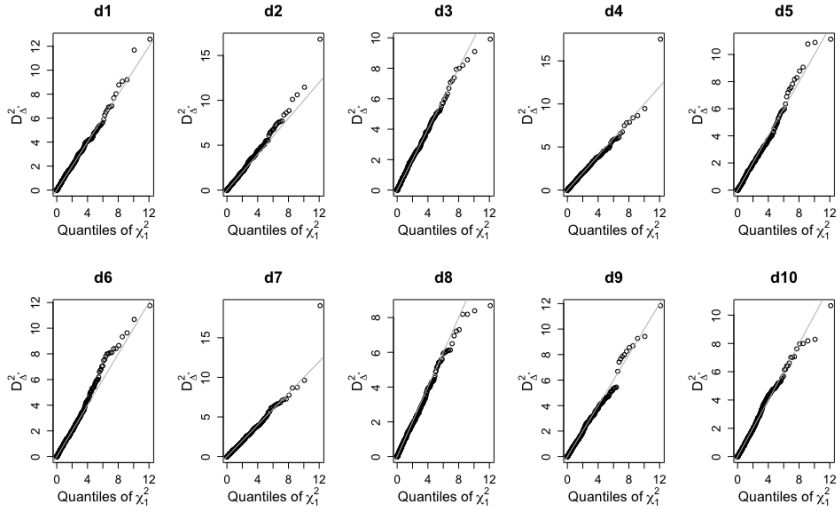
Figure H.4: $\gamma-$case: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\theta^*}$ versus quantiles of χ^2_2



(a) c=0.2

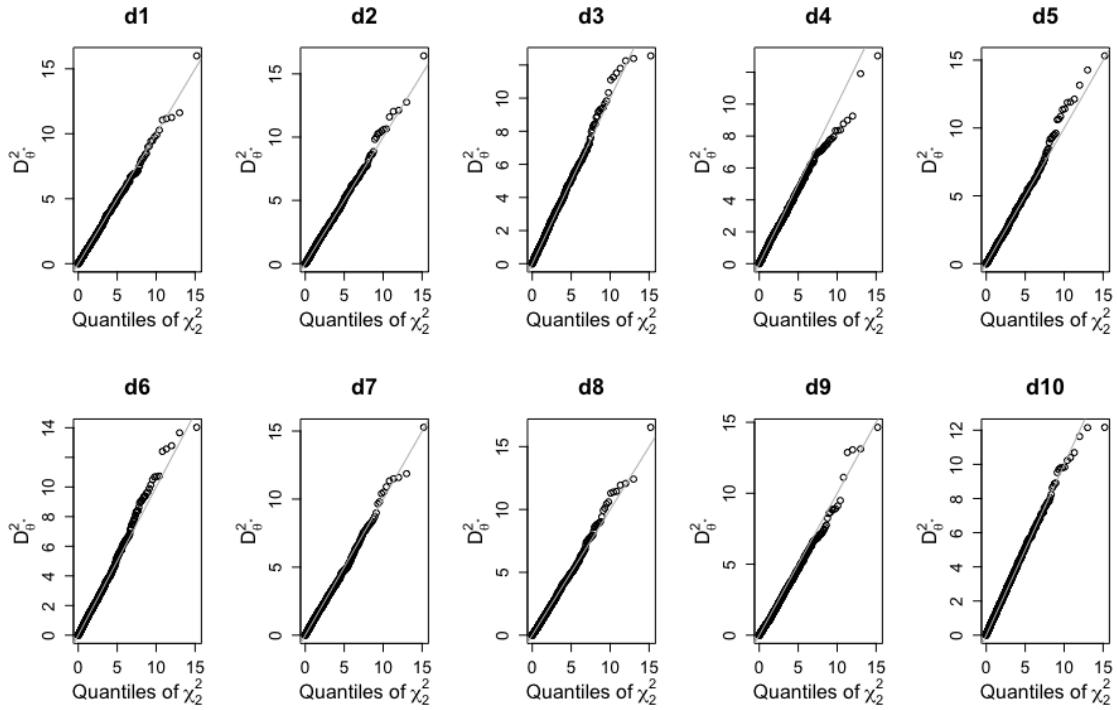


(b) c=1

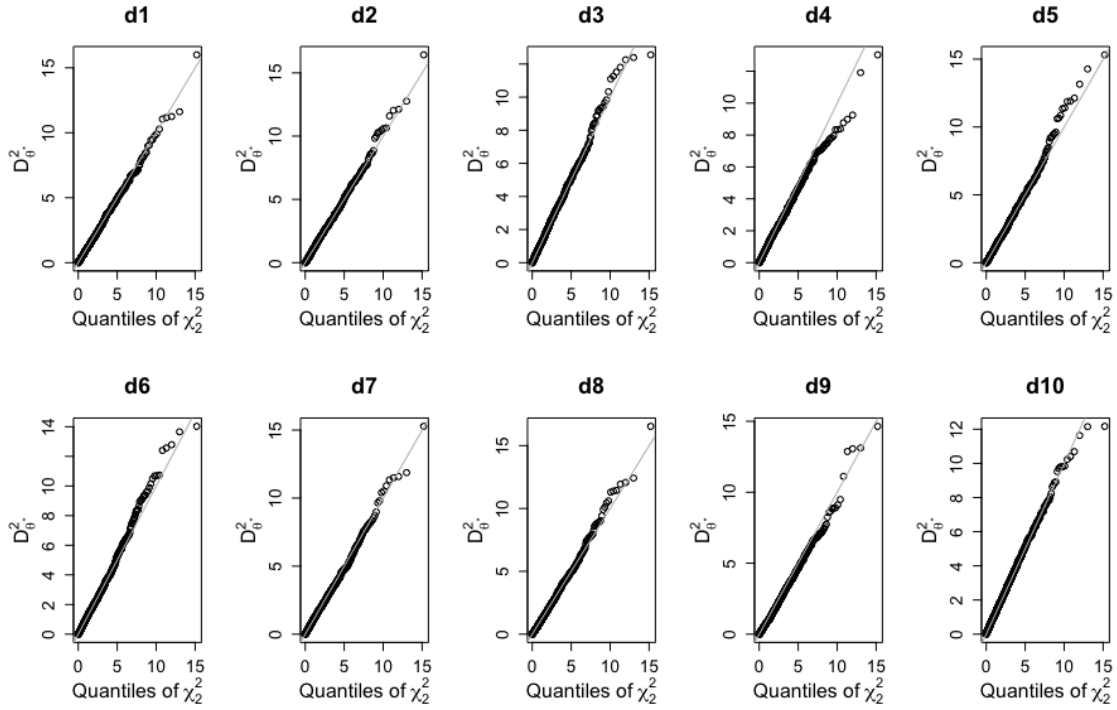


(c) c=5

Figure H.5: $\gamma-$case: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ versus quantiles of χ^2_1

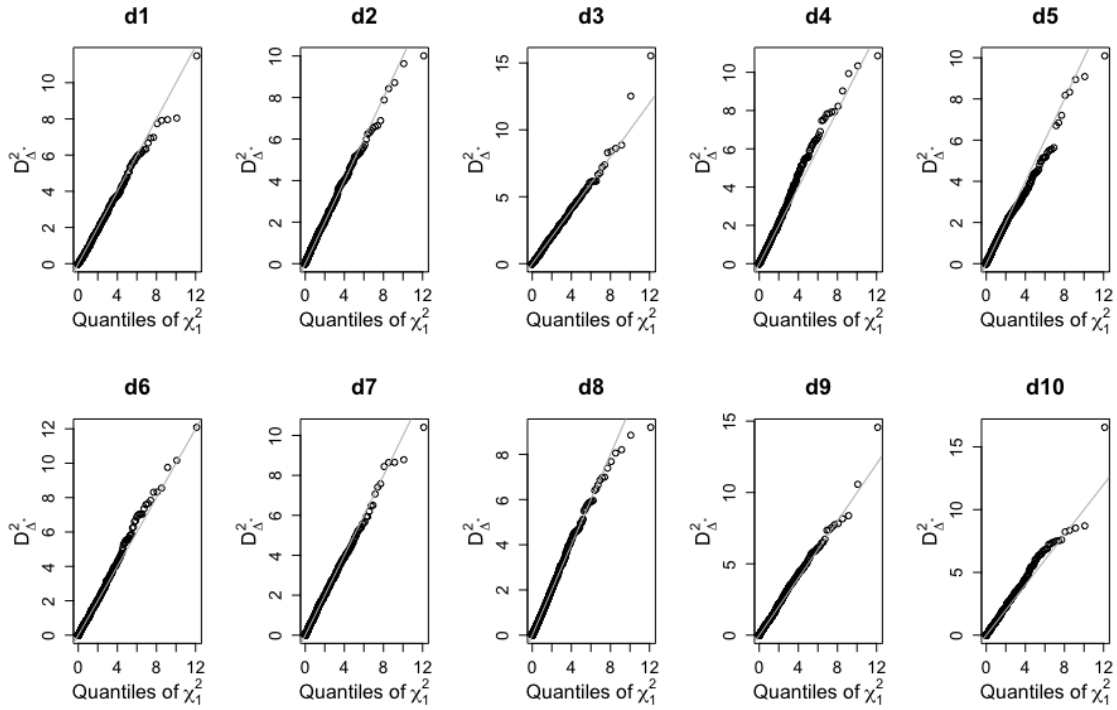


(a) Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ versus quantiles of χ^2_1

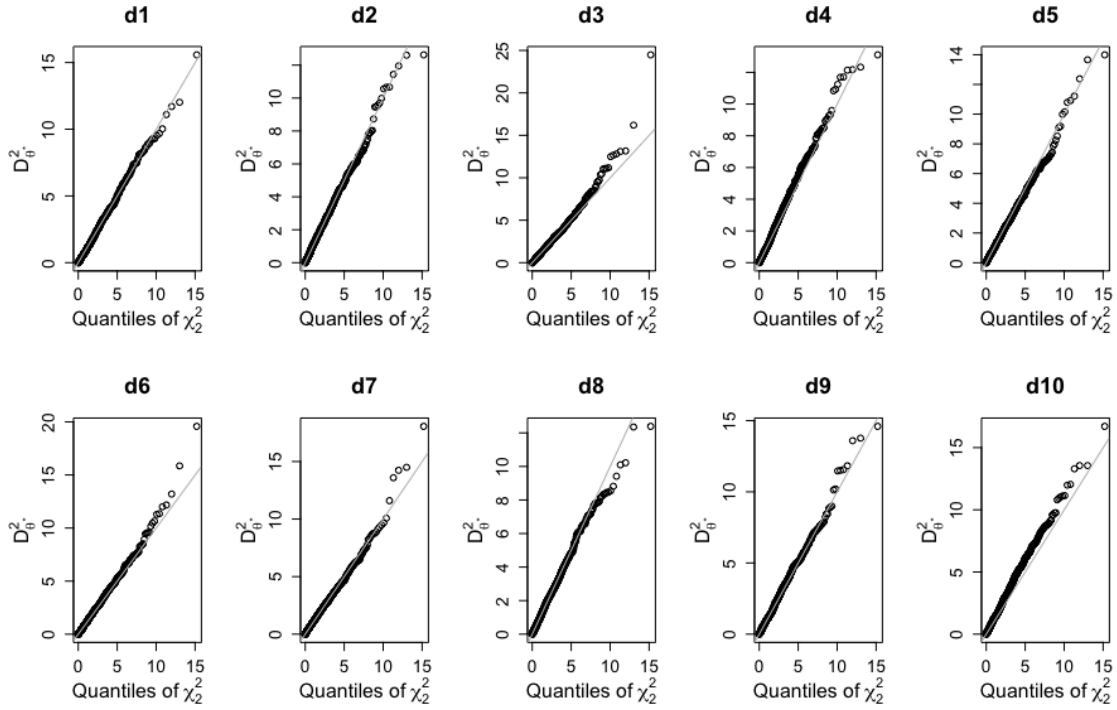


(b) Chi-square Q-Q plot: Mahalanobis distances $D^2_{\theta*}$ versus quantiles of χ^2_2

Figure H.6: $\beta\delta-$case with $c_1 = c_2 = 1$: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ and $D^2_{\theta*}$ versus quantiles of χ^2_1 and χ^2_2, respectively.



(a) Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ versus quantiles of χ^2_1



(b) Chi-square Q-Q plot: Mahalanobis distances $D^2_{\theta*}$ versus quantiles of χ^2_2

Figure H.7: $\beta\delta-$case with $c_1 = c_2 = -1$: Chi-square Q-Q plot: Mahalanobis distances $D^2_{\Delta^*}$ and $D^2_{\theta*}$ versus quantiles of χ^2_1 and χ^2_2, respectively.

Appendix I

Performance of Design Strategies Using the Whole Micro Sample Individuals n_{ik}

I.1 Complete and Partial Random Design

The figures presented in this section investigate the performance of the complete and partial randomised design strategies across different truth instances, when the whole micro sample individuals n_{ik} are utilised in the computation. The investigation includes a comparison between the contributions of the two sources of variability: the approximation error $\tilde{\Delta}^* - \Delta_0$ and the sampling error $\hat{\Delta}^* - \tilde{\Delta}^*$

Considering truth parameters in $\delta-$case, $\beta-$case and $\gamma-$case with $c \in \{0.2, 1, 5\}$, **Figure I.1** shows that the standard deviations $\tilde{\text{sd}}(\hat{\Delta}^*)$ turn out to be lower than the values obtained from using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, making the sampling error corresponding to the given design

strategies to be too low and close to zero. Focusing on $\delta-$ case across c values, we can see that the complete randomised design strategy meets the lowest typical magnitude of the approximation error with nearly zero sampling error and zero bias.

Considering truth parameters in $\beta\delta-$case with four sets of combinations of $c_1, c_2 \in \pm 1$, **Figure I.2a** shows a non-zero bias in the approximation error for the complete and partial randomisation design strategies. When the sign of c_1 and c_2 are different, the magnitude of the approximation error is minimum at 10% randomisation design strategy, whereas the other way round happens in the bottom two sets where the sign of c_1 and c_2 are the same. These results are in consistent with what was observed in using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$. The sampling error tended to be too low in this scheme and very close to zero.

Figure I.2b illustrates the four performance measures: the variability $\text{sd}(\hat{\Delta}^*)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\text{E}[\tilde{\text{sd}}(\hat{\Delta}^*)]$. The variability of the sampling error is lower than the variability of the approximation error across the four truth instances and design strategies except 10% randomisation design strategy in $c_1 = -1, c_2 = 1$. When c_1 and c_2 have same signs, the variability of the approximation error is higher than for the case when their signs are different. The bias is positive when the signs of c_1, c_2 are the same and negative when the signs of c_1, c_2 are different. The variability of the total error achieve their minimum in this situation

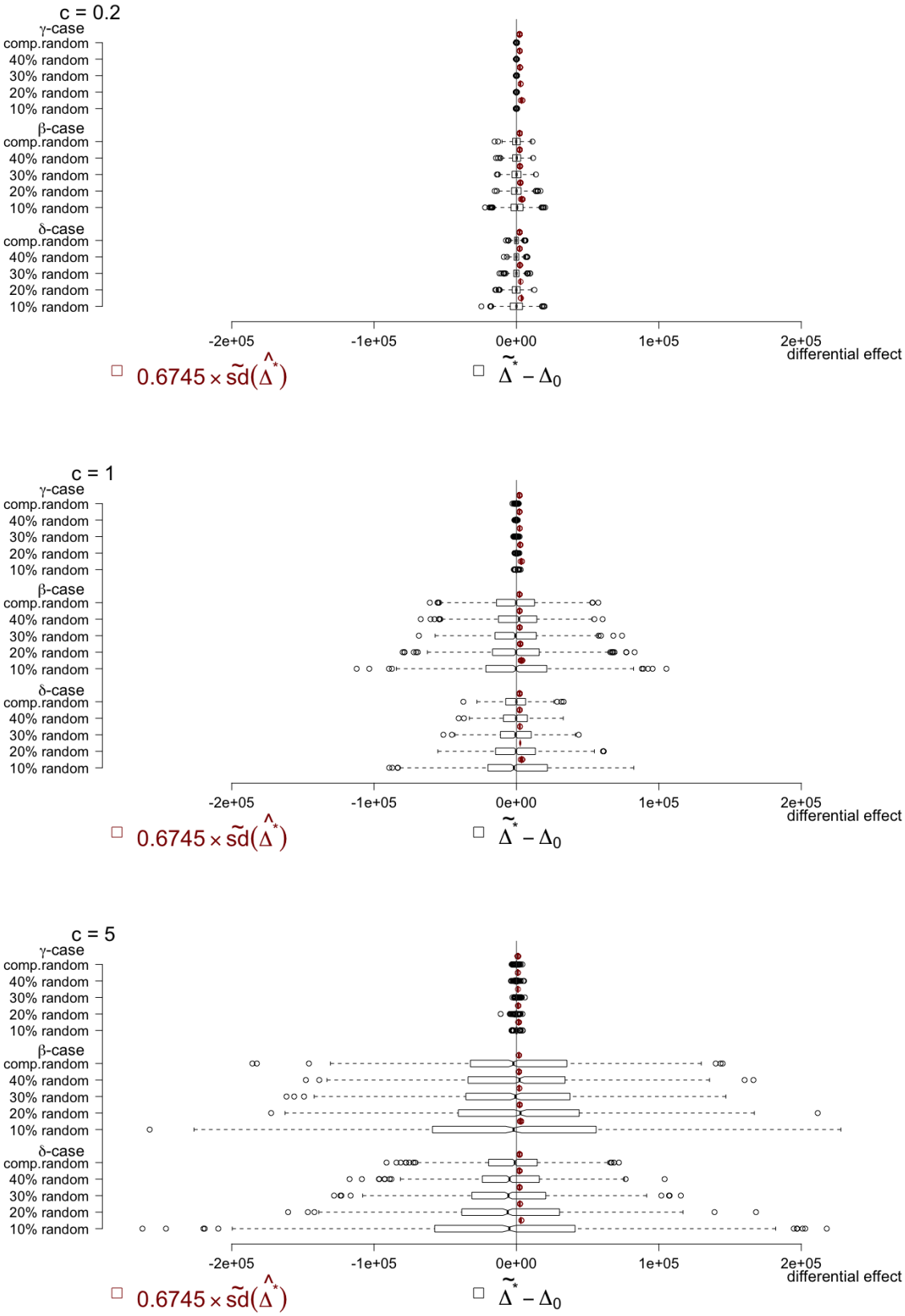
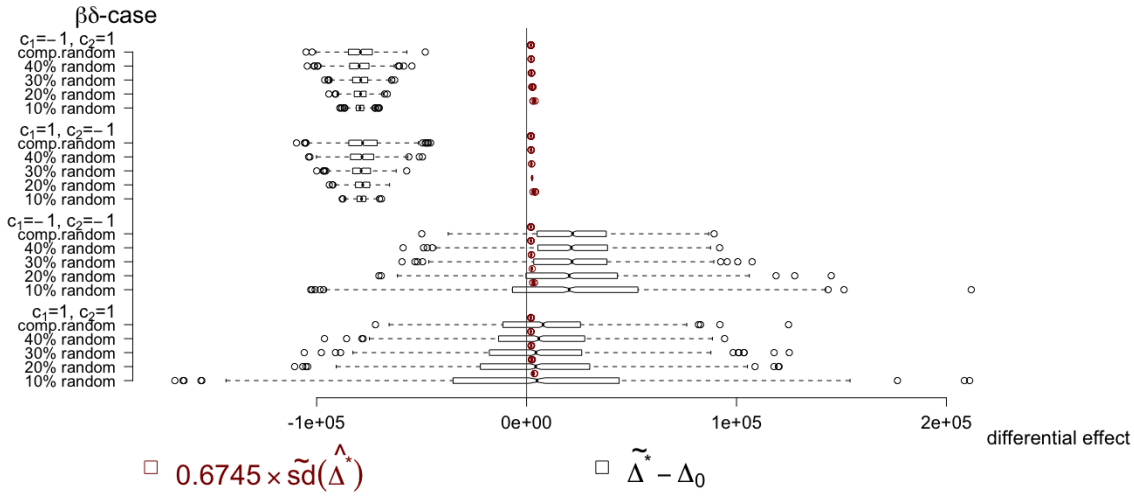
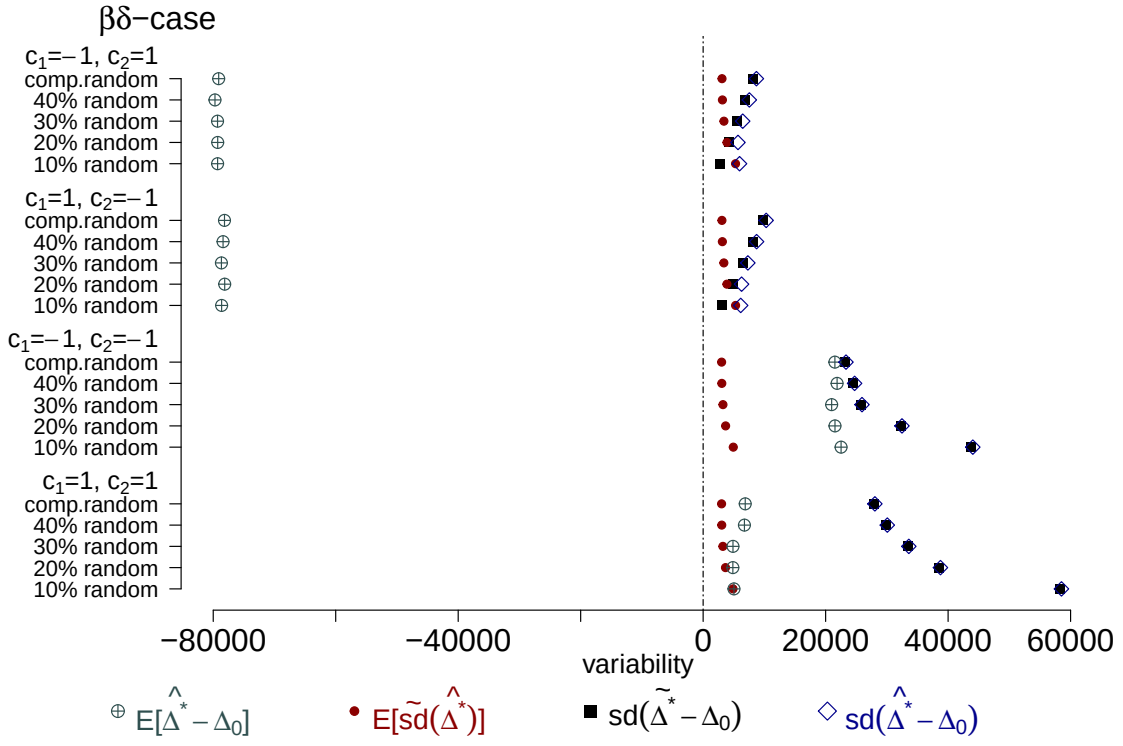


Figure I.1: n_{ik}, (complete and partial randomised design strategy using truth, $\delta-$case, $\beta-$case, $\gamma-$case across $c \in \{0.2, 1, 5\}$): comparing the contributions of the two sources of variability: sampling error and approximation error. The true effect: $\Delta_{0\delta_{c=0.2}} = 6963.728$, $\Delta_{0\delta_{c=1}} = 91576.92$, $\Delta_{0\delta_{c=5}} = 316677.5$, $\Delta_{0\beta} = \Delta_{0\gamma} = 0$.

at the complete randomised design strategy.



(a) comparing the contributions of the two sources of variability: sampling error and approximation error.



(b) Comparing the four performance measures: the average bias $\mathrm{E}[\hat{\Delta}^* - \Delta_0]$ =$\mathrm{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\mathrm{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\mathrm{E}[\tilde{sd}(\hat{\Delta}^*)]$.

Figure I.2: n_{ik}, performance evaluation of complete and partial randomised design strategy using truth, $\beta\delta-$case with combinations of $c_1, c_2 \in \pm 1$. The true effects are: $\Delta_{0[c_1=1,c_2=1]} = 100812.8$, $\Delta_{0[c_1=-1,c_2=-1]} = 92423.04$, $\Delta_{0[c_1=1,c_2=-1]} = -11325.6$, $\Delta_{0[c_1=-1,c_2=1]} = 11325.6$.

Considering truth parameters in $\gamma\beta\delta-$case with eight sets of combinations of $c_1, c_2, c_3 \in \pm 1$, **Figure I.3** shows a non-zero bias in the approximation error for the complete and partial randomisation design strategies. For the upper four sets of the truth instances in the figure, the magnitude of the approximation error becomes the largest when the complete randomisation are applied, whereas the other way round happens in the bottom four sets. These results are consistent with what was observed in using $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$. The sampling error tended to be too low in this scheme and very close to zero.

Figure I.4 illustrates the four performance measures: the variability $\text{sd}(\hat{\Delta}^*)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\text{E}[\tilde{\text{sd}}(\hat{\Delta}^*)]$. The average variability of the sampling error is small across the eight truth instances and design strategies. The total variability is attributed mainly to the approximation error in most truth instances. However, the magnitude of the total variability appears to be much larger in the second plot in the figure. In these truth version, complete randomised design strategy appears to provide the lowest variability. However, in the first plot in the figure, the complete randomisation appears having a bit higher variation than the partial randomisation design. The magnitude of the average bias is negative in all truth instances except when $c_1 = -1, c_2 = c_3 = 1$ and $c_1 = 1, c_2 = c_3 = -1$,.

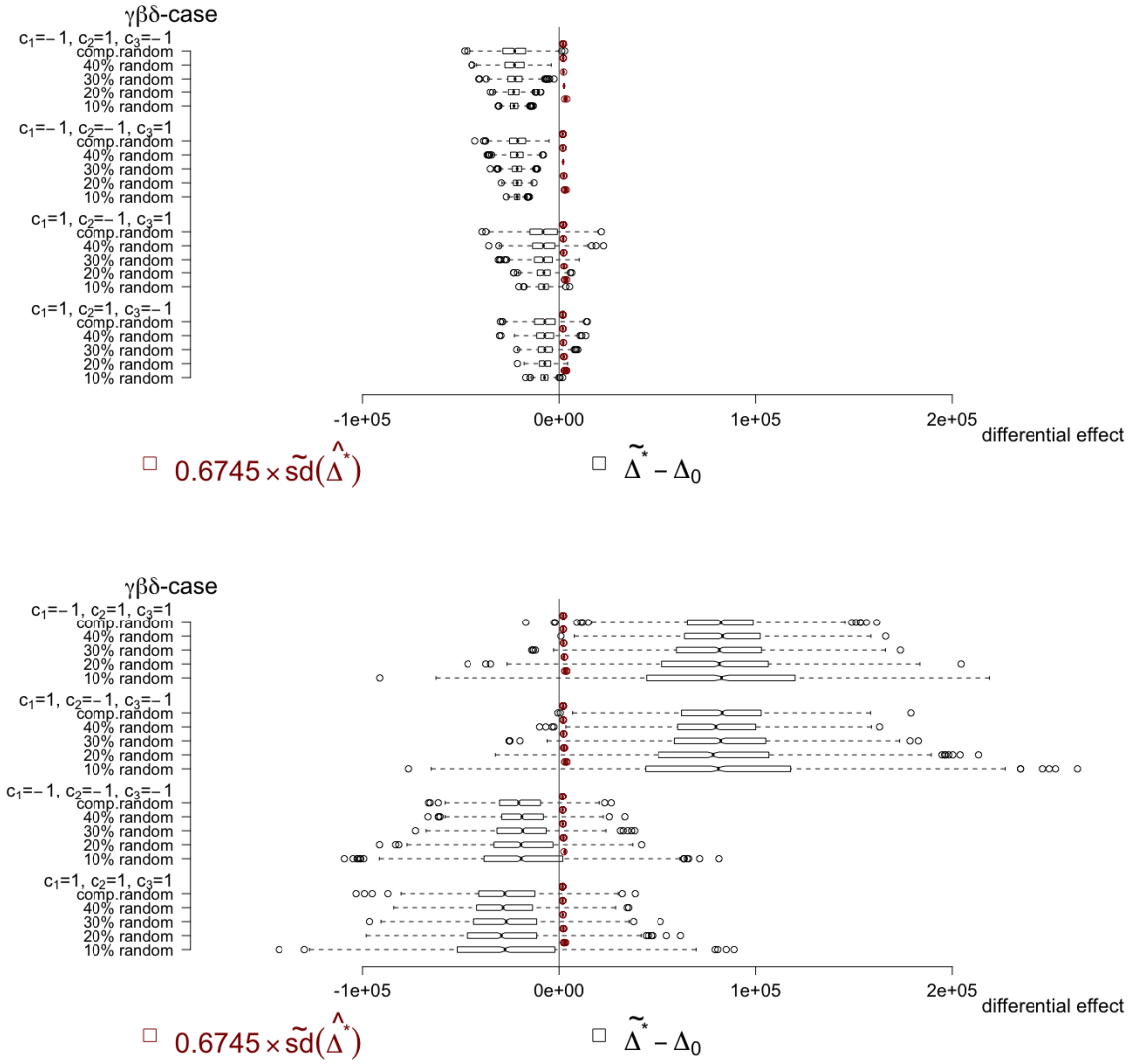


Figure I.3: n_{ik}, (complete and partial randomised design strategy using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the contributions of the two sources of variability: sampling error and approximation error. The true effects are: $\Delta_{0[c_1=c_2=c_3=1]} = 83837.85$, $\Delta_{0[c_1=c_2=c_3=-1]} = 98900.54$, $\Delta_{0[c_1=1,c_2=-1,c_3=-1]} = -11325.6$, $\Delta_{0[c_1=-1,c_2=1,c_3=1]} = 11325.6$, $\Delta_{0[c_1=1,c_2=1,c_3=-1]} = -100812.8$, $\Delta_{0[c_1=1,c_2=-1,c_3=1]} = 100812.8$, $\Delta_{0[c_1=-1,c_2=-1,c_3=1]} = -92423.04$, $\Delta_{0[c_1=-1,c_2=1,c_3=-1]} = 92423.04$.

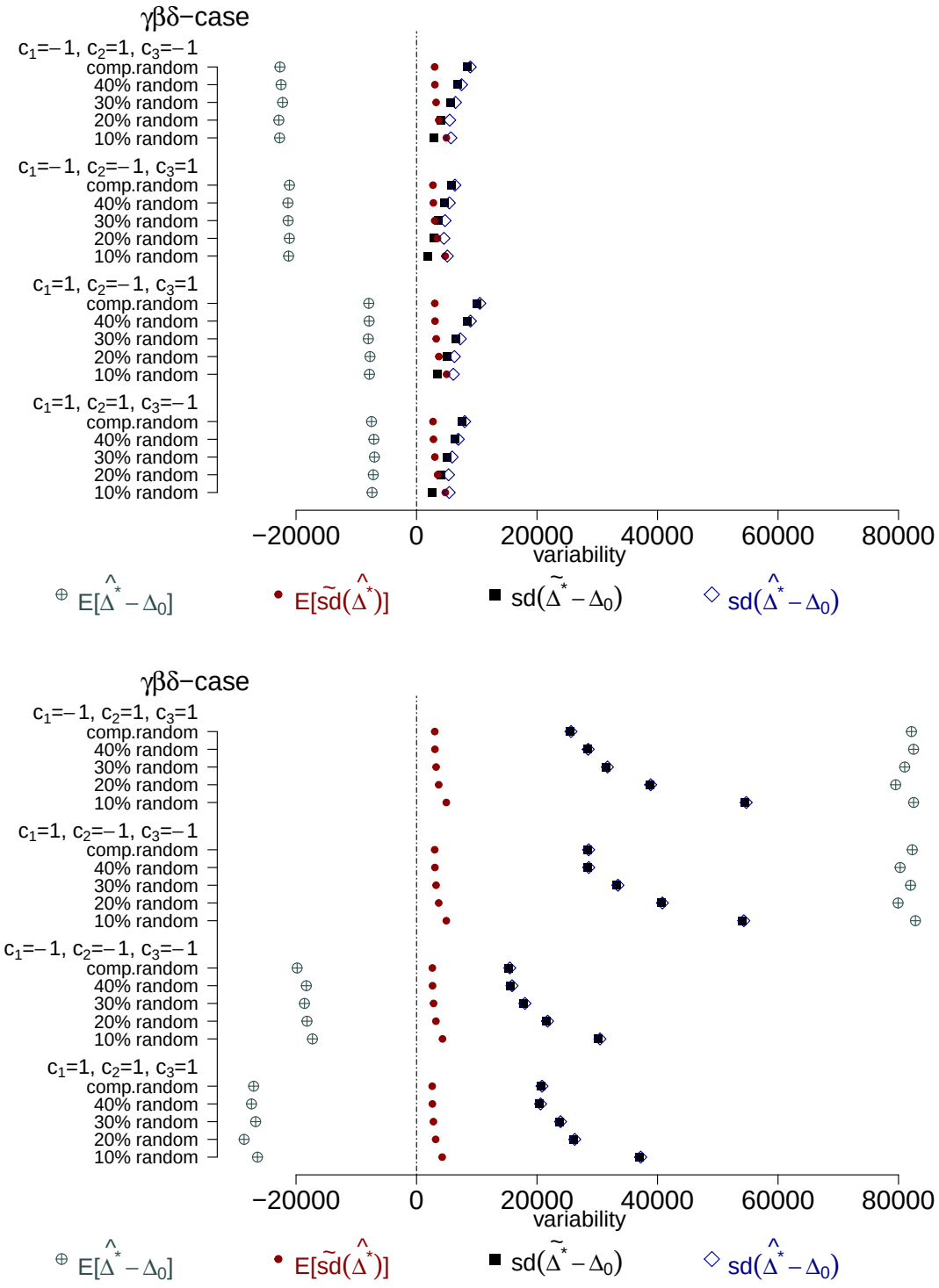


Figure I.4: n_{ik}, (complete and partial randomised design strategy using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$:): comparing the four performance measures: the variability $\text{sd}(\hat{\Delta}^*)$ and the average bias $\text{E}[\hat{\Delta}^* - \Delta_0]$ =$\text{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\text{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\text{E}[\tilde{sd}(\hat{\Delta}^*)]$.

I.2 Matched-Pair Design

The figures presented in this section investigate the performance of the matched-pair design strategies and the complete randomisation design strategy across truth instances mentioned above, when the whole micro sample individuals n_{ik} are employed in the computation.

Considering truth parameters in $\delta-$case, $\beta-$case, $\gamma-$case with $c \in \{0.2, 1, 5\}$, **Figure I.5** shows there is a considerable gain from using the matched-pair design strategies, as the variability of the approximation is lower than that obtained by the complete randomised design strategy. The advantage of matching pairs by social grades can be seen clearly in $\delta-$case, $\beta-$case in $c = 1$. Unlike the other two schemes $1\%(n_{ik})$ and $n_{ik0} \neq n_{ik1}$, the contribution of the sampling error is lower than the contribution of the approximation error. The existence of the non-zero bias continue occur in $c = 5$.

Considering truth parameters in $\beta\delta-$case with four sets of combinations of $c_1, c_2 \in \pm 1$. In this scheme, the average variability of the sampling is much lower then the approximation error. The large number of search reduces the variability of sampling error and makes the contribution of the approximation error as the total error. The benefit of matching pairs using the social grades is clear in this scheme too as shown in **Figure I.6a**. The low variability of the total error is obtained from the design strategies that relied on the social grades in the four truth instances.

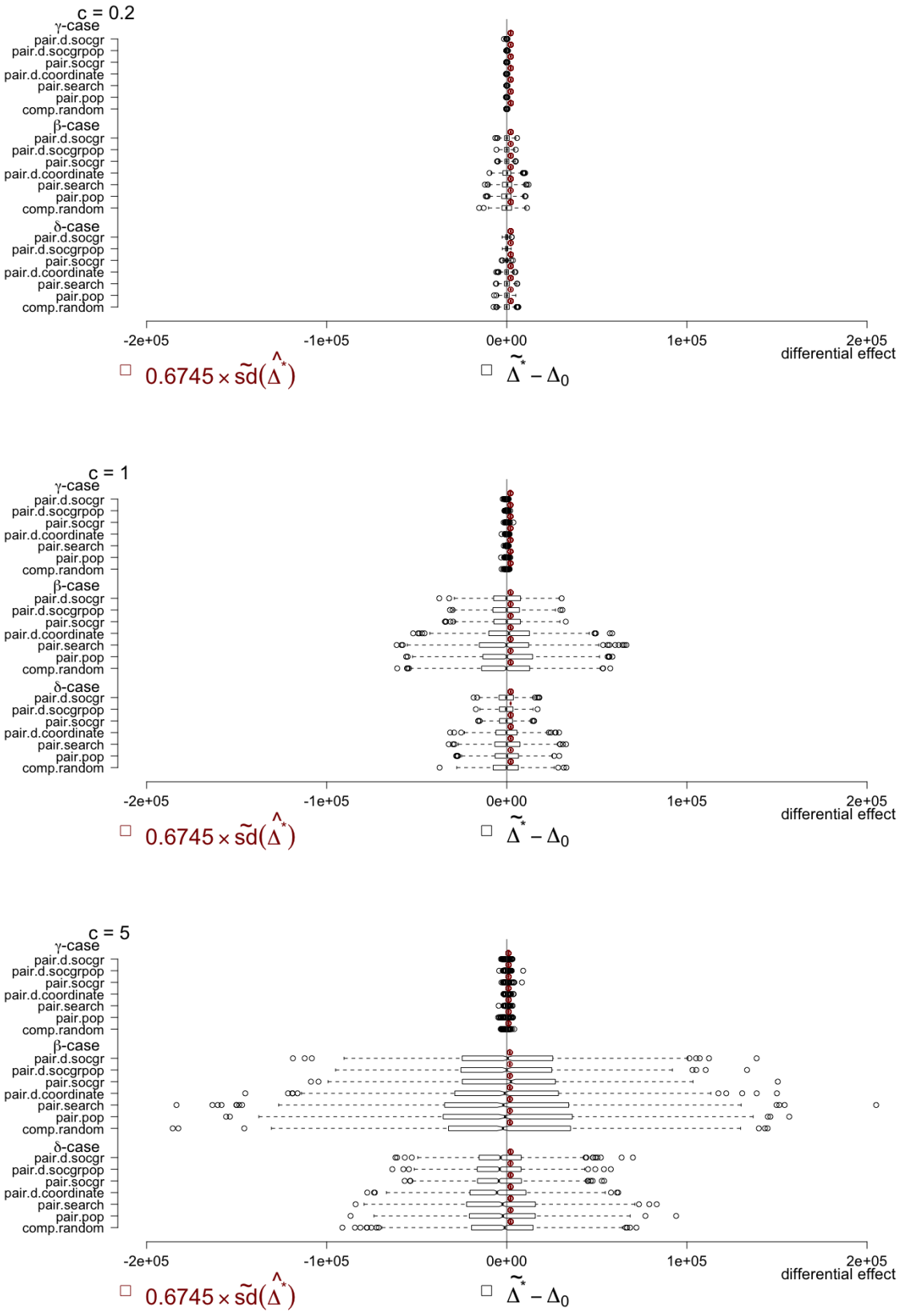
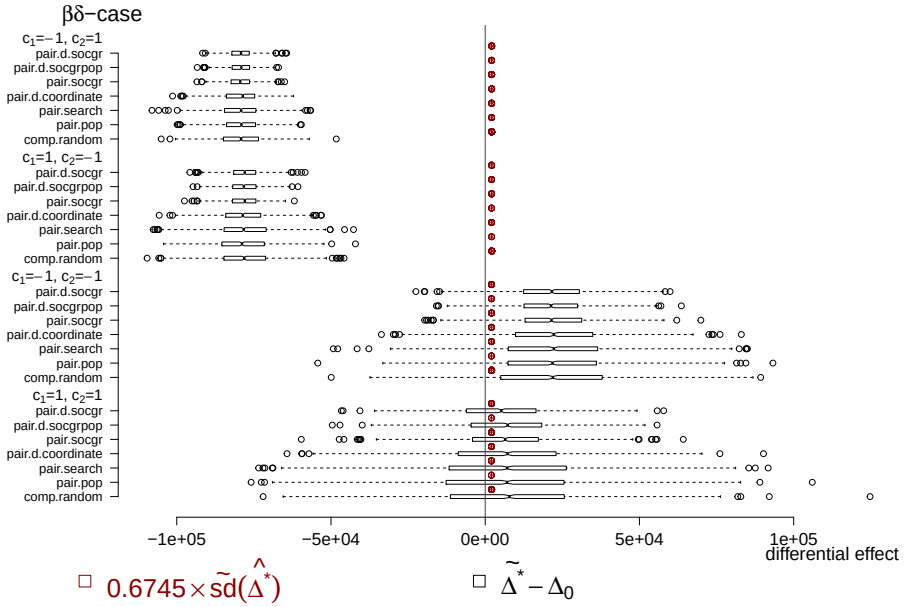
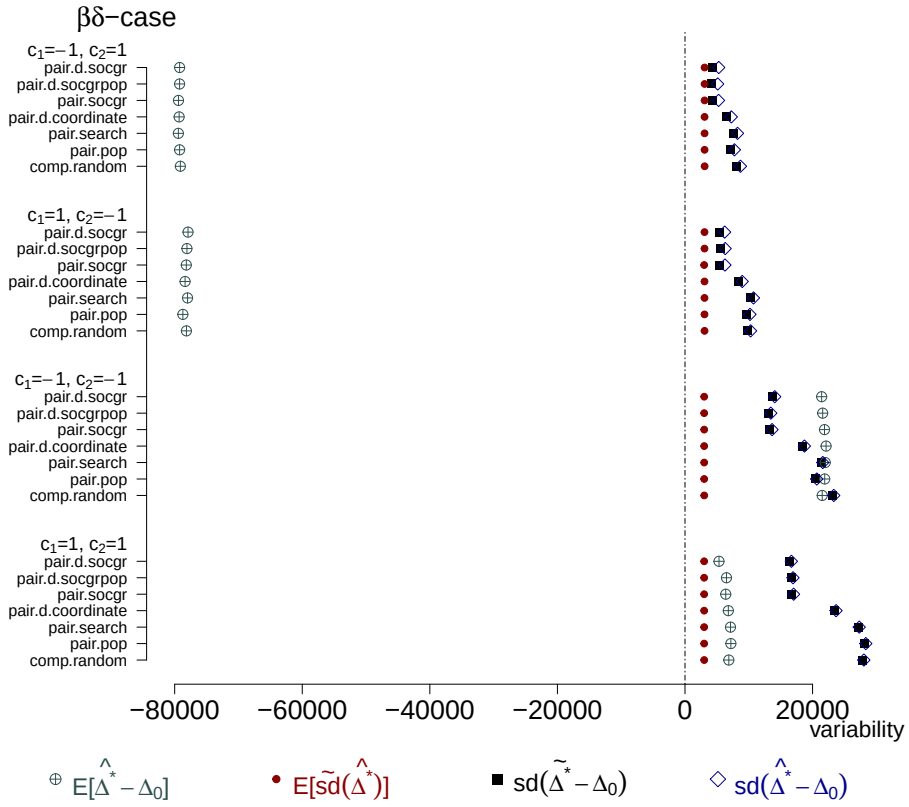


Figure I.5: n_{ik}, (matched-pair design strategies using truth, $\delta-$case, $\beta-$case, $\gamma-$case across $c \in \{0.2, 1, 5\}$), comparing the contributions of the two sources of variability: sampling error and approximation error.



(a) Comparing the contributions of the two sources of variability: sampling error and approximation error



(b) Comparing the four performance measures: the average bias $\mathrm{E}[\hat{\Delta}^* - \Delta_0]$ =$\mathrm{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\mathrm{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\mathrm{E}[\tilde{sd}(\hat{\Delta}^*)]$.

Figure I.6: n_{ik}, performance evaluation of matched-pairs randomised design strategy using truth: $\beta\delta-$case using combinations of $c_1, c_2 \in \pm 1$.

Considering truth parameters in $\gamma\beta\delta-$case with eight sets of combinations of $c_1, c_2, c_3 \in \pm 1$, **Figure I.7** shows a reduction in the variability of the approximation error when matching pairs by the social grade, especially when $c_1 = -1, c_2 = 1, c_3 = -1$, $c_1 = 1, c_2 = -1, c_3 = 1$, $c_1 = -1, c_2 = 1, c_3 = 1$ and $c_1 = 1, c_2 = -1, c_3 = -1$. The bias in the first two sets of those four instances is lower than the bias in the other two set. The expected variability of the sampling error is about the same across the applied design strategies in the eight truth instances.

The four performance measures are shown in **Figure ??**. The figure shows that the minimum total variability is attributed to the social grades and spatial proximity design strategies. However, the results are biased due to unobserved heterogeneity.

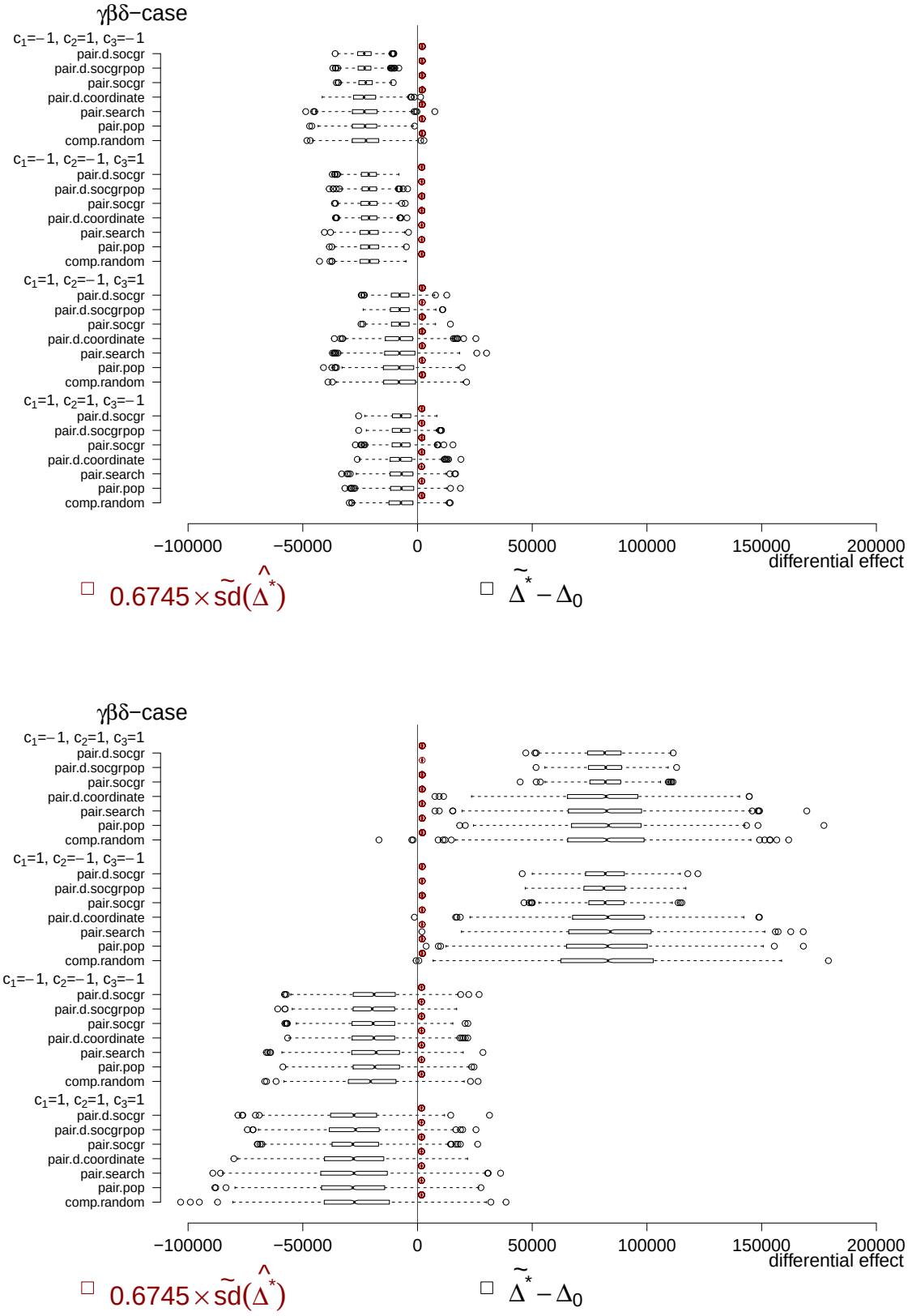


Figure I.7: n_{ik}:: matched-pair design strategies using truth: $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$:: comparing the contributions of the two sources of variability: sampling error and approximation error.

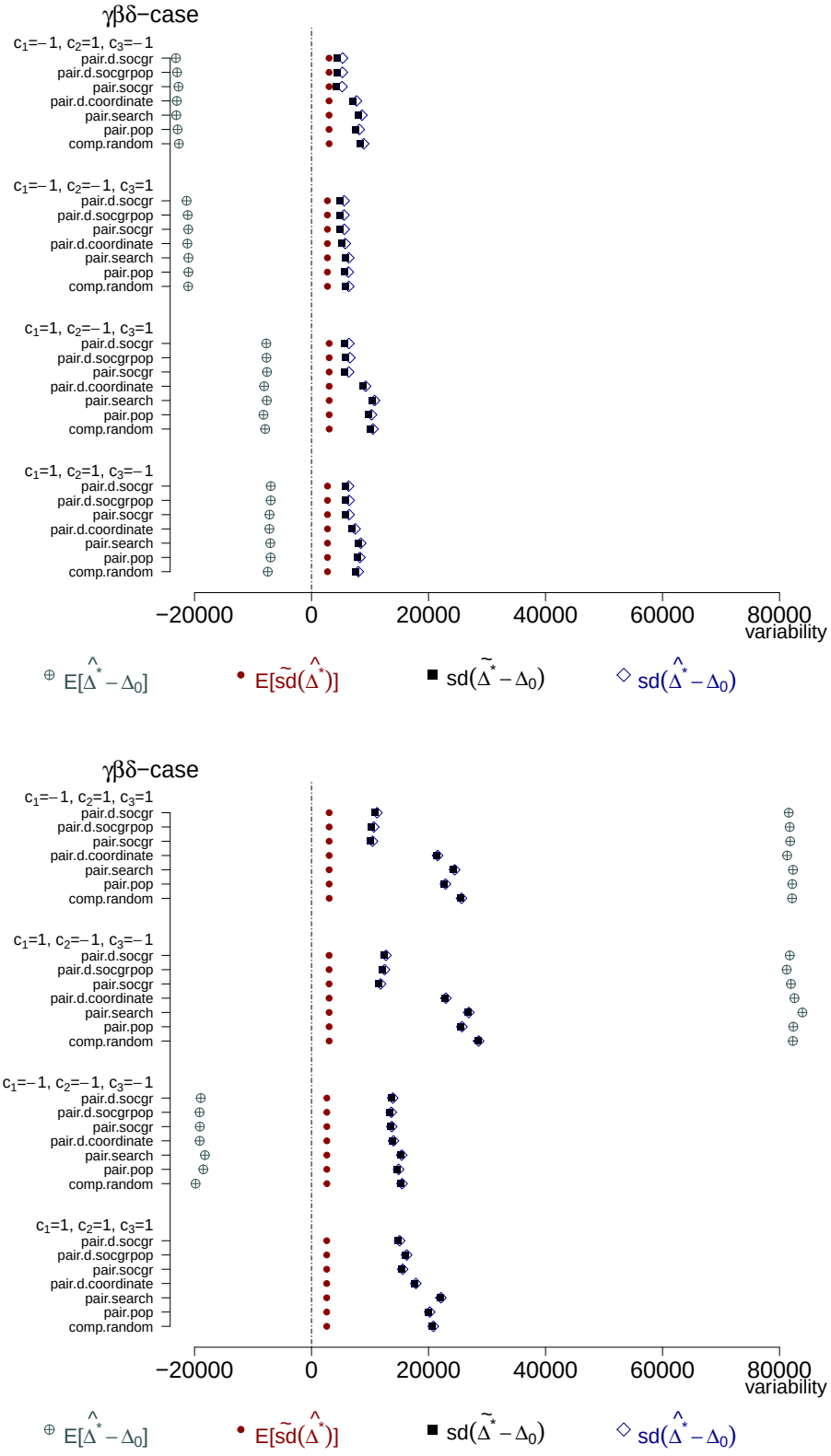


Figure I.8: n_{ik}, (matched-pair design strategies using truth, $\gamma\beta\delta-$case using combinations of $c_1, c_2, c_3 \in \pm 1$): comparing the four performance measures: the average bias $\mathrm{E}[\hat{\Delta}^* - \Delta_0] = \mathrm{E}[\tilde{\Delta}^* - \Delta_0]$ of the total error, the variability of the approximation error $\mathrm{sd}(\tilde{\Delta}^*)$ and the average variability of the sampling error $\mathrm{E}[\tilde{sd}(\hat{\Delta}^*)]$.

Bibliography

AdWords Fundamentals: Exam Study Guide. (2015). Google. Retrieved from https://www.impawa.com/documents/adwords-fundamentals.pdf

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer, New York, NY.

Akobeng, A. K. (2005). Understanding Randomised Controlled Trials. *Archives of Disease in Childhood*, *90*(8), 840–844.

Aufenanger, T. (2017). *Machine Learning to Improve Experimental Design* (tech. rep. No. 16/2017). Friedrich-Alexander University Discussion Papers in Economics. Retrieved from https://www.econstor.eu/bitstream/10419/169116/1/898624746.pdf

Aufenanger, T. (2018). *Treatment Allocation for Linear Models* (tech. rep. No. 14/2017). Friedrich-Alexander University Discussion Papers in Economics. Retrieved from https://www.econstor.eu/bitstream/10419/179521/1/14-2017-2.pdf

Blake, T., Nosko, C. & Tadelis, S. (2015). Consumer Heterogeneity and Paid Search Effectiveness: A Large-Scale Field Experiment. *Econometrica*, *83*(1), 155–174.

Box, G. E., Hunter, W. H. & Hunter, S. (1978). *Statistics for Experimenters*. New York: John Wiley and Sons.

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., Scott, S. L. et al. (2015). Inferring Causal Impact Using Bayesian Structural Time-Series Models. *The Annals of Applied Statistics*, *9*(1), 247–274.

Cassa, C. A., Iancu, K., Olson, K. L. & Mandl, K. D. (2005). A Software Tool for Creating Simulated Outbreaks to benchmark Surveillance Systems. *BMC Medical Informatics and Decision Making*, *5*(1), 22.

Chan, D. X., Yuan, Y., Koehler, J. & Kumar, D. (2011). Incrementalclicks: The Impact of Search Advertising. *Journal of Advertising Research*, *51*(4), 643–647.

Chan, D., Ge, R., Gershony, O., Hesterberg, T. & Lambert, D. (2010). Evaluating Online Ad Campaigns in a Pipeline: Causal Models at Scale. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 7–16). ACM.

Chow, G. C. (1984). Maximum-likelihood Estimation of Misspecified Models. *Economic Modelling*, *1*(2), 134–138.

Clark, W. A. & Avery, K. L. (1976). The Effects of Data Aggregation in Statistical Analysis. *Geographical Analysis*, *8*(4), 428–438.

Cochran, W. & Cox, G. (1957). *Experimental Designs*. New York: John Willey and Sons.

Daniels, H. E. (1954). Saddlepoint Approximations in Statistics. *The Annals of Mathematical Statistics*, 631–650.

Diong, M. L., Chaumette, E. & Vincent, F. (2017). On the Efficiency of Maximum-Likelihood Estimators of Misspecified Models. In *2017 25th European Signal Processing Conference (EUSIPCO)* (pp. 1455–1459). IEEE.

Dutilleul, P. (1993). Spatial Heterogeneity and the Design of Ecological Field Experiments. *Ecology*, *74*(6), 1646–1658.

Eisinga, R., Te Grotenhuis, M. & Pelzer, B. (2013). Saddlepoint Approximations for the Sum of Independent Non-Identically Distributed Binomial Random Variables. *Statistica Neerlandica*, *67*(2), 190–201.

Fain, D. C. & Pedersen, J. O. (2006). Sponsored Search: A Brief History. *Bulletin of the American Society for Information Science and Technology*, *32*(2), 12–13.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1992). The Arrangement of Field Experiments. In *Breakthroughs in Statistics* (pp. 82–91). Springer, New York, NY.

Goldfarb, A. (2014). What is Different about Online Advertising? *Review of Industrial Organization*, *44*(2), 115–129.

Gordon, B. R., Jerath, K., Katona, Z., Narayanan, S., Shin, J. & Wilbur, K. C. (2021). Inefficiencies in Digital Advertising Markets. *Journal of Marketing*, 0022242920913236.

GPS Visualizer Tool. (2007). Retrieved from https://www.gpsvisualizer.com/

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, *1*, pp. 221–233). University of California Press.

Iacus, S., King, G., Porro, G., MatchIt, I., S., Amelia & Iacus, M. (2018). Package 'cem'. *URL: https://cran.r-project.org/web/packages/cem/cem.pdf*.

Imai, K. (2008). Variance Identification and Efficiency Analysis in Randomized Experiments under the Matched-Pair Design. *Statistics in Medicine*, *27*(24), 4857–4873.

Jansen, B. J. & Mullen, T. (2008). Sponsored Search: An Overview of the Concept, History, and Technology. *International Journal of Electronic Business*, *6*(2), 114–131.

Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. (2017). Accurately Interpreting Clickthrough Data as Implicit Feedback. In *ACM SIGIR Forum* (Vol. 51, *1*, pp. 4–11). Acm.

Johnson, N. L., Kemp, A. W. & Kotz, S. (2005). *Univariate Discrete Distributions*. New York: John Wiley and Sons.

Jolayemi, J. K. (1992). A Unified Approximation Scheme for the Convolution of Independent Binomial Variables. *Applied Mathematics and Computation*, *49*(2-3), 269–297.

Kempthorne, O. (1955). The Randomization Theory of Experimental Inference. *Journal of the American Statistical Association*, *50*(271), 946–967.

Kim, C., Kwon, K. & Chang, W. (2011). How to Measure the Effectiveness of Online Advertising in Online Marketplaces. *Expert Systems with Applications*, *38*(4), 4234–4243.

Lavrakas, P. J. (2010). An Evaluation of Methods Used to Assess the Effectiveness of Advertising on the Internet. *Interactive Advertising Bureau Research Papers*. Retrieved from https://www.iab.com/wp-content/uploads/2015/07/Evaluation_of_Internet_Ad_Effectiveness_Research_Methods.pdf

Legendre, P., Dale, M. R., Fortin, M.-J., Casgrain, P. & Gurevitch, J. (2004). Effects of Spatial Structures on the Results of Field Experiments. *Ecology*, *85*(12), 3202–3214.

Lewis, R. A., Rao, J. M. & Reiley, D. H. (2011). Here, There, and Everywhere: Correlated Online Behaviors can Lead to Overestimates of the Effects of Advertising. In *Proceedings of the 20th International Conference on World wide web* (pp. 157–166). ACM.

Liu, B. & Quertermous, T. (2017). Approximating the Sum of Independent Non-Identical Binomial Random Variables. *ArXiv Preprint arXiv:1712.01410*.

Lovelace, R. & Cheshire, J. (2014). Introduction to Visualising Spatial Data in R. EloGeo. Retrieved from http://eprints.ncrm.ac.uk/3295/4/intro_to_R.pdf

Lucas, C., Muraleedharan, G. & Soares, C. G. (2014). Outliers Identification in a Wave Hindcast Dataset Used for Regional Frequency Analysis. In *Maritime Technology and Engineering* (pp. 1331–1342). CRC Press.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

Moulton, B. R. (1990). An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *The Review of Economics and Statistics*, 334–338.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Rubin, D. B. (1973). Matching to Remove Bias in Observational Studies. *Biometrics*, 159–183.

Rutz, O. J. & Bucklin, R. E. (2011). From Generic to Branded: A Model of Spillover in Paid Search Advertising. *Journal of Marketing Research*, *48*(1), 87–102.

Stuart, E. A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *25*(1), 1.

Valli, V., Stahl, F. & Feit, E. (2017). Field Experiments. *Handbook of Market Research. Springer International Publishing*, 1–29.

Van Es, H., Gomes, C., Sellmann, M. & Van Es, C. (2007). Spatially-Balanced Complete Block Designs for Field Experiments. *Geoderma*, *140*(4), 346–352.

Vaver, J. & Koehler, J. R. (2011). Performing Geography-Based Advertising Experiments. US Patent App. 12/777,035. Google Patents.

Vaver, J. & Koehler, J. (2011). *Measuring Ad Effectiveness Using Geo Experiments*. Technical Report, Google Inc. Retrieved from http://research.google.com/pubs/pub38355.html

Vaver, J. & Koehler, J. (2012). *Periodic Measurement of Advertising Effectiveness using Multiple-Test-Period Geo Experiments*. Technical Report, Google Inc. Retrieved from http://research.google.com/pubs/pub38356.html

Weller, B. & Calcott, L. (2012). *The definitive guide to google adwords: Create versatile and powerful marketing and advertising campaigns*. New York: Apress.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica: Journal of the Econometric Society*, 1–25.

Ye, Q., Malik, S., Chen, J. & Zhu, H. (2016). The Seasonality of Paid Search Effectiveness from a Long Running Field Test. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (pp. 515–530).