Statistical power of epidemiological studies of low dose levels of ionizing radiation and cancer

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the School of Public Health

University of Saskatchewan

Saskatoon

By

Jafar Soltani Farsani

Permission to Use

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the School of Public Health
104 Clinic Place
University of Saskatchewan
Saskatoon, Saskatchewan, S7N 2Z4, Canada
Or,
Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan, S7N 5C9, Canada

Abstract

The purpose of this research is to inspect the statistical power of studies that investigate the effects of low dose ionizing radiation on the incidence of/mortality from cancer. I use a procedure proposed in a similar study to handle a problem regarding the incidence of childhood leukemia from background ionizing radiation. I study this procedure critically and make some adjustments to make its performance better. I also propose some substitute methods to the methods proposed in the aforementioned reference in order to calculate the power. In addition, I propose other methods not used in the study mentioned above. I evaluate the efficiency of my proposed approaches using simulated data. The improved method can be applied to the National Dose Registry of Canada (NDR) to produce the power curves. The outcomes then can be used to propose the most suitable study design. Some of the previous epidemiological studies based on NDR can also be evaluated in terms of power.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors, Dr. Shahedul Khan, Dr. Cindy Feng and Dr. Gavin Cranmer-Sargison for their academic advice, and support. It has been a great opportunity for me to work under their supervision. I would like to thank the faculty and staff of the biostatistics program at the School of public health, University of Saskatchewan for providing such a nice environment. I would also like to thank my friends and fellow students in the University of Saskatchewan for the motivation they gave to me. Finally, I feel I am deeply indebted to my family who have always been supportive to me during my whole life, without any doubt, their kindness can not be paid back.

Contents

List of Tables

List of Figures

List of Abbreviations

ABCC	Atomic Bomb Causality Commission
CHD	Congenital Heart Defect
CVD	CardioVascular Disease
ERR	Excess Relative Risk
EOR	Excess Odds Ratio
LSS	Life Span Study
MCE	Monte-Carlo Error
MLE	Maximum Likelihood Estimator
RERF	Radiation Effects Research Foundation
SEER	Surveillance, Epidemiological and End Results
SIR	Standardized Incidence Ratio
SMR	Standardized Mortality Ratio
NDR	National Dose Registry
OC	Oral Contraceptive
OR	Odds Ratio
RR	Relative Risk

Chapter 1

Preliminaries

In this chapter first, I explain the motivations for carrying out this research. Then, I present the concept of statistical power (henceforth I will refer to it simply as power), and its calculation based on closed-form formulas and simulations. I also discuss factors that can potentially impact the power of study design. Since my study mainly deals with cohort and case-control studies, I then define these two study designs in a general context.

1.1 Motivations for the study

As described in Gordis (2014; page 7): "Epidemiologic reasoning is whether an association exists between exposure to a factor". Depending on the study design the association is evaluated based on some measures of association such as the risk ratio (RR) in cohort studies and the odds ratio (OR) in case-control studies (Ahrens and Pigeot. 2015; Chapters I.5 and I.6). If RR/OR falls between 0 and 1, the association is interpreted as being inverse or negative, i.e., an increase in the exposure is associated with a decrease in the outcome variable. On the other hand, a value of RR/OR greater than 1 indicates a positive/direct association between the exposure and the outcome variable. Note that 1 is the null value for RR/OR, implying that there is no association between the exposure and the outcome variable. In practice, the RR/OR is an unknown quantity and is estimated from a sample. Then we rely on statistical methods to test whether or not it is significantly different from the

null value 1. An insignificant result leads to the conclusion that there is not enough evidence to support the association between the exposure and the outcome. In general, significance is determined through a hypothesis test. It should be noted that an insignificant result might stem from several factors, including:

1. no true association between the exposure and the outcome,
2. selection and/or information bias (Gordis 2014; Chapter 15),
3. low power to detect an association (Ellis 2010; Chapter 1).

Therefore, knowledge about power is importnat. The reason is that if the epidemiologists already control the bias and are informed that the power of their study is high enough, then a non-significant result can be considered as a strong evidence of no association between the exposure and the outcome. Otherwise, if the power is low, the non-significant result should not be interpreted as the lack of association. In general, the power of a statistical test is defined as the probability of rejecting the null hypothesis when it is true (i.e., the probability of making a correct decision). In this regard, a low power indicates higher uncertainty as to whether a statistically insignificant result is due to an actual lack of difference or simply due to the sample size that is insufficient to detect the effects of the exposure.

Knowledge regarding the power of a study is crucially important: if the researcher is aware that the study is powerful enough, then an insignificant result could be strong evidence that there is truly no association (provided that other issues such as bias are also under control). By contrast, if he/she knows that the study has low power, then the insignificant association should not be interpreted as a lack of association.

In this research, the exposure and outcome of interest are exposed to low dose levels of ionizing radiation and all health outcomes of cancer (incidence or mortality), respectively. In the next section, I will briefly present different types of studies that try to detect an association between the aforementioned exposure and outcome. One type of such study

is occupational studies that aim to establish the association by studying occupations that are exposed to low dose levels of ionizing radiation, yet higher than that of the general public (Beebe et al. 1998, Richardson and Wing 1999, Sun et al. 2016). Examples of such occupations are nuclear plant workers and health workers such as dentists, nurses, etc. The National Dose Registry (NDR) of Canada (NDR website 2020) is a database which has been recording the information of such workers since 1951. Epidemiological researches in the area of ionizing radiation and cancer have been carried out based on this database. A couple of such studies are Ashmore et al. (1998), Gribbin et al. (1993). Each of these studies has covered different periods and consequently, they have different sample sizes. As a result (as will be explained in the coming sections) they have different powers to detect the association.

my purpose is to determine the power of NDR-based studies for different periods of follow-up. In this way, I will be able to relax the epidemiologist's concern regarding the power as explained above. In particular, I will answer the following questions:

1. What is the best study design in terms of power? cohort or case-control? Also, what is the optimum number of controls per case for the case-control study?

2. Have the previous studies been powerful enough to detect an effect?

1.2 About the cohort and case-control studies

Cohort and case-control studies are two very common study designs in epidemiology. Most ionizing radiation-cancer studies are of these two types. The main difference between the two designs is that the cohort study starts with individuals whose exposure status is known. They are followed up for some time to determine the status of the outcome. In comparison, the case-control studies first determine the status of the outcome and then try to find the history of exposure. More precisely the two study designs are defined in Gordis (2014) as follows:

- Cohort study. "In a cohort study, the investigator selects a group of exposed individuals and a group of non-exposed individuals and follows up both groups to compare the incidence of disease (or rate of death from disease) in the two groups. The design may include more than two groups, although only two groups are shown for diagrammatic purposes." (Gordis 2014; Chapter 9)

- Case-control study. "We begin by selecting cases (with the disease) and controls (without the disease), and then measure past exposure by interview and by review of medical or employee records or of results of chemical or biologic assays of blood, urine, or tissues." (Gordis 2014; Chapter 10)

The most commonly used measure of association in cohort studies is the Risk Ratio (RR). If both exposure and outcome are binary (i.e. two possible outcomes), then RR is defined as

$$\begin{aligned} RR &= \frac{\text{Probability of being a case in exposed group}}{\text{Probability of being a case in non-exposed group}} \\ &= \frac{a/(a+b)}{c/(c+d)}, \end{aligned} \tag{1.1}$$

where a and b are, respectively, the number of cases and non-cases in the exposed group , and c and d are those in the non-exposed group (See Table 1.1). Risk ratio can also be written as $RR = 1 + ERR$ where ERR is defined as the Excess Relative Risk.

The odds ratio is the most commonly used measure of association for case-control studies, defined as

$$\begin{aligned} OR &= \frac{\text{Odds of being exposed as a case}}{\text{Odds of being exposed as a control}} \\ &= \frac{a/c}{b/d} \end{aligned} \tag{1.2}$$

where a and c are numbers of exposed and non-exposed cases, respectively. b and d are

number of exposed and non-exposed controls, respectively (See Table 1.1). The odds ratio can be written as $OR = 1 + EOR$ where EOR is the Excess Odds Ratio.

Table 1.1: Contingency table for cohort and case-control studies

	Cases	Controls	Total
Exposed	a	b	a+b
Non-exposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

Remark 1.1. It is proven in Greenland and Thomas (1982) that when the outcome of interest is rare, the odds ratio and risk ratio are almost equal.

1.3 Power

Assume that we are testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. This test can make two types of errors (Casella and Berger 1990; Section 8.3):

1. type I error: if the true parameter belongs to Θ_0 but the test rejects it in favor of $\theta \in \Theta_0^c$, then a type I error has occurred.

2. type II error: if the true parameter belongs to Θ_0^c but the test decides not to reject the null in favor of $\theta \in \Theta_0^c$, then a type II error has occurred.

The probability of the complement of type II error is called the power of the test. Therefore, the power of a test is the probability to truly reject the null in favor of the alternative. In epidemiological terms, the power is usually interpreted as the ability to detect an effect when it truly exists. In every hypothesis test we have two competing factors that we would like to optimize: On one hand, we want to keep the probability of the type I error to a minimum.

On the other hand, we try to maximize the power of the test. As will be explained in the next section, optimizing any of these factors hurts the optimality of the other one.

In many cases, the null hypothesis has a simple form $H_0 : \theta = \theta_0$ for some θ_0. In such a case, the alternative is usually $H_1 : \theta \neq \theta_0$. Calculating the power then requires finding the distribution of the test statistic under the single value of θ_0. However, when it comes to the alternative, in practice, not all the parameters $\theta \neq \theta_0$ are considered. Rather, usually a single value $\theta_1 \neq \theta_0$ is picked. Then, to calculate the power, one way is to find the distribution of the test statistic at the null under the alternative. This means that the formula of the test statistic is calculated at $\theta = \theta_0$ but the distribution of interest is the one with samples from $\theta = \theta_1$. Another way is to take samples from alternative and check whether the test statistic under the null rejects it or not. The proportion of the rejections is then reported as power.

The value θ_1 as defined above is usually the parameter which is believed to be the true parameter or at least closer to θ_0 than the true parameter. The latter means we need to find the power for an effect size smaller than the true effect. In such a case we need a higher power to detect θ_1 and we can rest assured that with such a power we will be able to detect the true parameter too.

1.3.1 Factors that impact the power of a test

There are mainly 4 factors that can impact the power of a test (Norton and Strube 2001). Note that the first three factors as described below can directly change the power. Therefore, the designer can change any of them to reach a higher power. The fourth factor, as described below, is a kind of inherent and seems to be not adjustable. This is why most references do not refer to this factor. However, we will give an example to show that in some cases one can design the experiment in a way to obtain higher power even in case all the other three factors are fixed.

Below, I explain the impact of each factor using a plot. For simplicity, I assume that

the test statistic under both null and alternative is known and is normally distributed. The dashed area in each figure represents the power of the test.

1. The significance level. The probability of the type I error is referred to as the significance level (also, the size of the test (See Casella and Berger (1990; Section 8.3)). There is a trade-off between the significance level and power. This means increasing any of them causes the other one to increase as well. One simple explanation is that a higher significance level means that we have to reject more often (we are conservative), as a result, the chance of a correct rejection increases too, which means using threshold we get higher power.

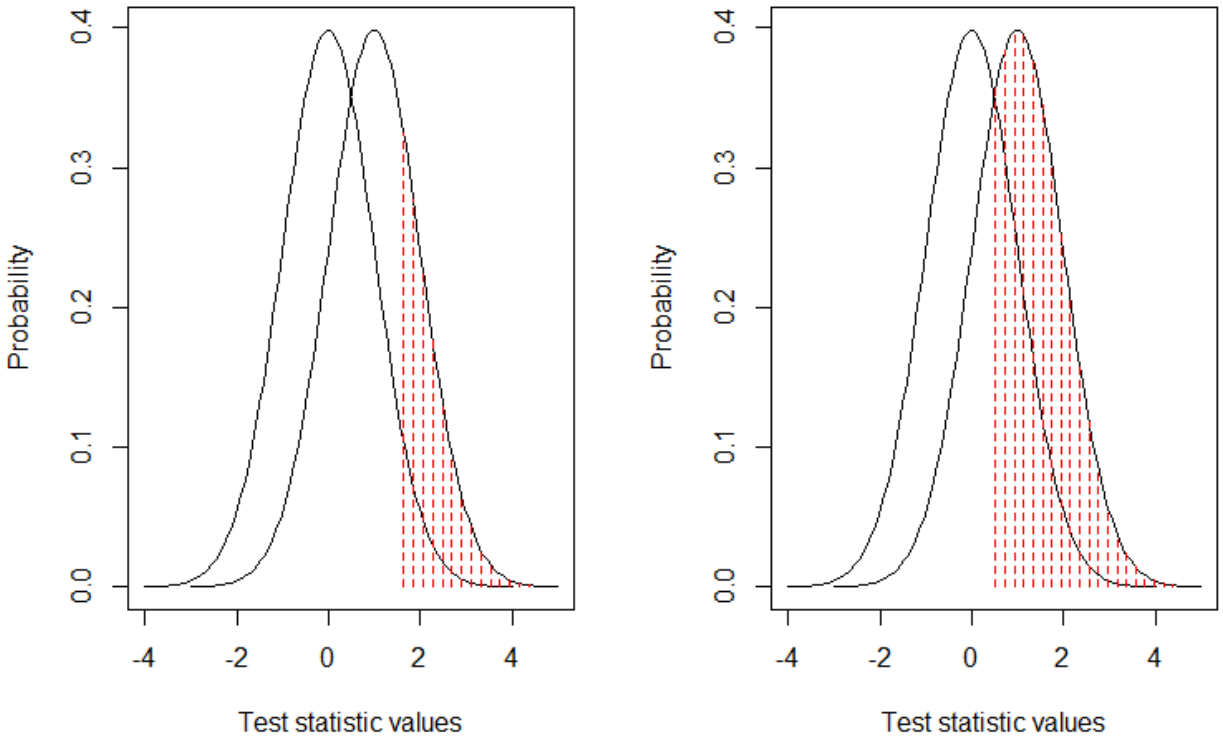

Figure 1.1: Two tests with $\alpha = 0.03$ (left) and $\alpha = 0.05$ (right), adjusted for the other factors.

Figure 1.1 represents this fact (dashed area represents the power of the test). While

the shape and position of the test statistic under both null and alternative remains the same (as a result of keeping the sample size and the effect size fixed), increasing the significance level shifts the dashed area to the left and causes the power to go up.

2. The effect size. As defined in Ellis (2010; Chapter 1): "An effect size refers to the magnitude of the result as it occurs, or would be found, in the population". It was mentioned above that to calculate the power we usually fix one parameter in the alternative space such as $\theta = \theta_1$. The closer θ_1 to θ_0, the more mixed the distribution under the null and alternative. As a result, the test statistic under the null is more likely not to reject in favor of the alternative.

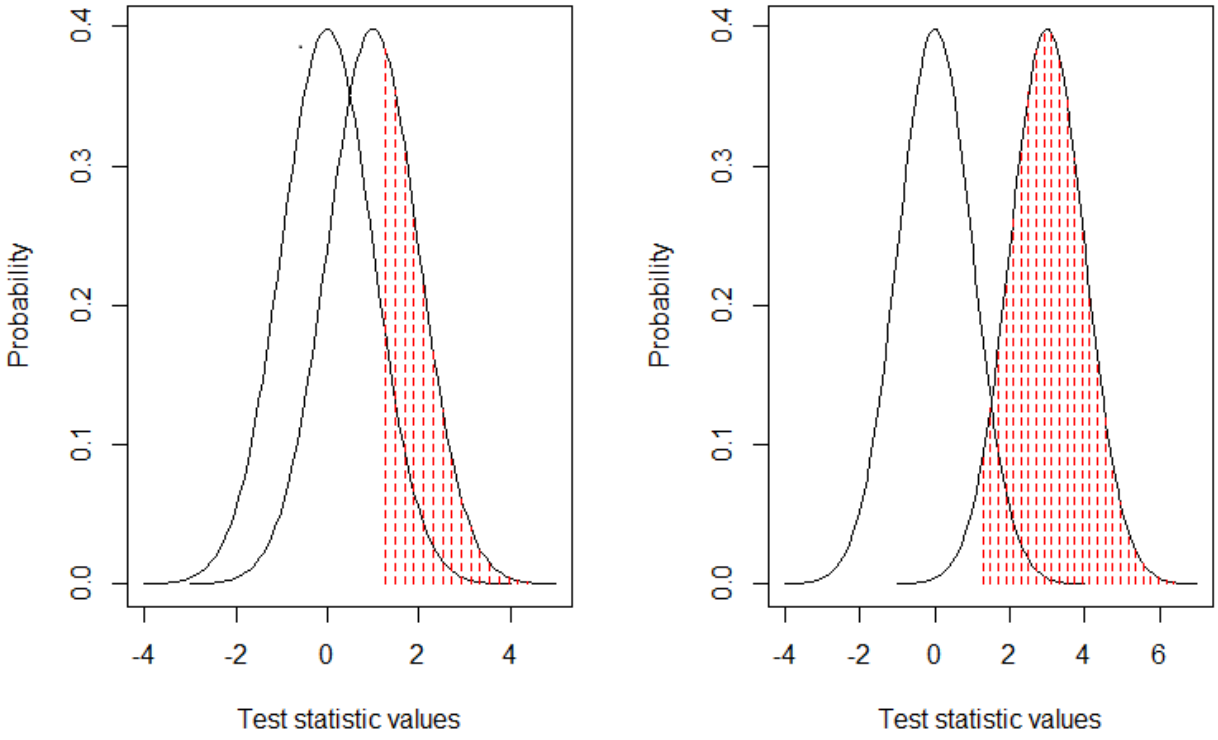

Figure 1.2: Two tests with $\theta_0 = 0$, $\theta_1 = 1$ (left) and $\theta_1 = 3$ (right), adjusted for the other factors.

Figure 1.2 shows the impact of effect size on the power. Notice that the shapes of the

test statistic and the start point of the dashed area remain the same (as a result of keeping the sample size and the significance level fixed). However, shifting the test statistic under alternative to the right in the right frame causes the power to increase.

3. The sample size. Larger sample size results in a higher power. A simple explanation is that there is a reverse relationship between the sample size and the variance of the test statistic under both null and alternative. For example, recall that the variance of $\bar{X}$ (the sample mean) which is the test statistic for the population mean is given by $\frac{\sigma}{\sqrt{n}}$ where σ is the variance of X and n is the sample size. This shows that a small sample size causes the variance of the test statistic to increase. This makes the test statistics under both null and alternative flatter and causes them to be more mixed which results in lower power. This fact has been represented in Figure 1.3. While the significance level and the effect size are the same in both tests, smaller sample size results in a flatter shape on the left.

   The effect of sample size on the power can also be justified using another argument. By definition

   $$power = P(Reject|Should\ be\ truly\ rejected).$$

   In this definition, the condition part, i.e."Should be truly rejected" refers to the real population while the event, i.e."Reject" refers to the sample. As a result, the larger the sample size, the better the true state of the population reflects on the sample. This results in more true rejections which imply higher power.

4. Variability in the data. The larger the variance of the sample, the larger the variance of the test statistic. As explained for the impact of the sample size, the large variance of the test statistic results in lower power. Again, a good example is $\bar{X}$ where its variance ($\frac{\sigma}{\sqrt{n}}$) directly depends on the variance of the sample. The impact of the sample's

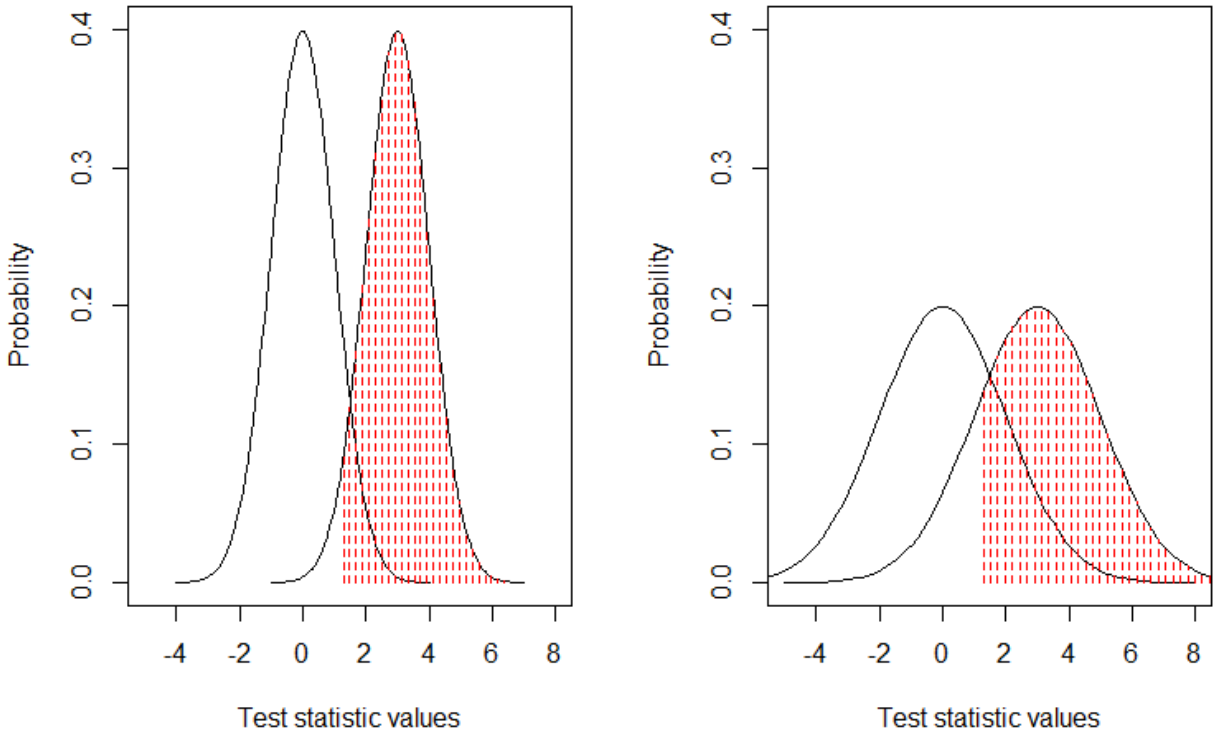

Figure 1.3: Two tests with different sizes: small (left) and large (right), adjusted for the other factors.

variance can be similarly represented using Figure 1.3.

It should be pointed out that this factor is less studied as compared to the other three factors. One main reason is that it seems to be inherent to the data and not fairly adjustable. So, it can not be deployed to improve the power of the test. However, an example is provided in Friedman et al. (2010; Chapter 8) to show that with a better study design one can improve the power while keeping the sample size, the significance level, and the effect size fixed. The example is regarding a clinical trial where the impact of a new drug on the cholesterol level is of interest. The sample is divided into the intervention and the control group. The impact of the drug can be measured by comparing the mean value of the cholesterol level in two groups. There are two possible designs. One can only measure the cholesterol level at the end of the trial and

just compare the mean values in the two groups. Another possibility is to measure it both at the beginning and the end of the trial, then calculate the differences for each individual and finally make an inference by comparing the mean of the differences. It is proven in Friedman et al. (2010; Chapter 8) that the latter trial results in lower variance and consequently higher power.

1.3.2 Why Monte-Carlo simulation to find the power?

Calculation of power using the analytical method, i.e. based on a closed-form formula, requires the distribution of the test statistic under both null and alternative to be known (Landau and Sathi 2012). In many cases, there are some tools (particularly, asymptotic tools) to deal with the distribution of the test statistic under the null. However, distribution under the alternative is usually more awkward. If any of the two distributions can not be determined analytically one needs to use sampling (Monte-Carlo simulation) to find the power (Landau and Sathi 2012). Notice that even this latter approach requires information about the distribution of samples when the true parameter is the alternative.

Apart from the case above, there are also some special complex cases which can be handled better using Mont-Carlo simulation. Two of such cases are as follows (Monte-Carlo for power 2019):

1. multiple tests: We would like to implement different tests at the same time. Such a test is of the form

$$(H_{0i} : \theta \in \Theta_{0i} \quad vs \quad H_{1i} : \theta \in \Theta_{1i} \ )_{i=1}^{n} .$$

2. multiple alternatives: We require at least one of the multiple alternative to be true. This test is of the form

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 : (\theta \in \Theta_1 \ or \ \theta \in \Theta_2 \ \cdots \ or \ \theta \in \Theta_n).$$

1.3.3 Score test statistic

The test statistic that I use for my research is the score test statistic derived from some specific likelihoods as described in Chapter 3 and 4.

Assume that we have a vector of parameters θ with the corresponding likelihood function L. Then the score function is defined with

$$U(\theta) = \nabla(logL(\theta)) = (\frac{\partial logL(\theta)}{\partial \theta_i})_{i=1}^{n}.$$

The fisher information is defined with

$$I(\theta) = -E[\nabla^2 log(L(\theta))|\theta].$$

The score test statistic is then defined with

$$S(\theta) = U(\theta)^T I(\theta)^{-1} U(\theta).$$

Assume that θ is a vector of length p and assume that we want to test for $H_0 : \theta = \theta_0$. The score test statistic for this hypothesis test is then $S(\theta_0)$. In 1948, Rao proved that the score test statistic under the null has asymptotic χ_p^2 distribution (Rao 1948). In particular if θ is a single parameter then $S(\theta) = \frac{U(\theta)^2}{var(U(\theta))}$ has asymptotic χ_1^2 distribution. Equivalently $\frac{U(\theta))}{\sqrt{var(U(\theta))}}$ is asymptotically normally distributed.

Remark 1.2. At this point, we would like to highlight a point regarding terms such as "test statistic under null" and "test statistic under the alternative". Assume that we test $H_0 : \theta = \theta_0$ and we are interested in detecting effect $\theta = \theta_1$. It should be carried in mind that the test statistic under the alternative might be not necessarily the test statistic when we test $H_0 : \theta = \theta_1$. In some simple cases, it turns out to be correct. For instance, in case of testing

for the mean value of a normal population, the test statistic under the alternative is the same as the test statistic under null when the null is $H_0 : \theta = \theta_1$. However, in general, this might be not true. It seems that simple cases such as this, in addition to the vague nature of the aforementioned terms, sometimes lead to confusion. Notice that when we use a term such as"test statistic under the alternative" we are referring to the distribution of the test statistic for the null when samples come from alternative! For example, if we use the score test statistic, then by the distribution of the score test under the alternative we don't mean the distribution of $S(\theta_1)$ but we mean the distribution of $S(\theta_0)$ when samples come from θ_1.

1.4 Organization of the thesis

I proceed in Chapter 2 by presenting a literature review on both ionizing radiation-cancer studies and power studies of case-control and cohort designs. In Chapters 3 and 4 I will discuss the power model for cohort and case-control studies, respectively. The models have already been developed in Little et al. (2010). I will spot some problems regarding the model and propose solutions. I also provide the algorithms required to implement the model. In Chapter 4, I provide some additional results which can be used to improve the performance of the model. I will address some points in Little et al. (2010) which seem to require revision and propose substitute methods. Some other approaches to calculate power are also presented. Based on a literature review, I provide required information such as the break-down of the dose parameter and estimates for the ERR parameter. In Appendix A, I use some simulated data to illustrate my proposals throughout the thesis.

Chapter 2

Literature review

This chapter provides an overview of the subject. The first section presents a brief description of radiation, introduces different types of radiation, basic facts regarding radiation and different trends to study the relationship between ionizing radiation and cancer. We review the epidemiological studies of low dose levels of ionizing radiation and cancer in the second section. The final section gives a review of the statistical studies that provide a model to evaluate the power and sample size for the cohort and case-control studies.

2.1 About ionizing radiation and studies related to cancer

Ionizing radiation is a kind of radiation with high energy that can remove electrons from atoms and molecules (Canadian Nuclear Safety Commission 2020). There are two general sources of ionizing radiation (Canadian Nuclear Safety Commission 2020):

1. natural: Space and cosmic radiation, terrestrial radiation, etc.

2. artificial: X-ray, nuclear power plants, etc.

The unit used to measure the amount of ionizing radiation received for each person (usually per year) is millisievert denoted by mSv (or mSv/y for per year measurements).

Four categories of annual effective doses of ionizing radiation are as follows (Fazel et al. 2009):

1. low: Doses between 0 mSv/y and 3 mSv/y.

2. moderate: Doses between 3 mSv/y and 20 mSv/y.

3. high: Doses between 20 mSv/y and 50 mSv/y.

4. very high: Doses over 50 mSv/y.

Notice that doses below a specific threshold (usually 0.1 or 0.2 mSv/y) are recorded as 0.

Ionizing radiation-Cancer studies can be divided into two general categories:

1. The oldest and the most studied category is related to the low dose levels of ionizing radiation (presented with details in the following section).

2. A relatively new study deals with high dose levels of ionizing radiation which, for instance, studies long term radiology survivors.

In the following section, I review studies of the first type.

2.2 Epidemiological studies of low dose levels of ionizing radiation

Low dose levels of ionizing radiation-cancer studies are followed in 3 general trends:

1. studies based on the atomic bomb event.

2. other ionizing radiation-cancer studies:

   (a) based on background ionizing radiation

   (b) based on occupational ionizing radiation.

2.2.1 Studies based on atomic bomb event

Since the early days, the survivors of the atomic bomb in Hiroshima and Nagasaki have been the main subjects of cancer and ionizing radiation studies. Studies based on these subjects has provided remarkable results. Contributions based on this study continue to come out.

The first study related to the impact of ionizing radiation on the mortality and incidence of cancer dates to 1950 when the Atomic Bomb Casualty Commission (ABCC) started studying a fixed cohort of A-bomb survivors and suitable comparison subjects (as a non-exposed group). ABCC was established to investigate the health effects of atomic bombs in Hiroshima in 1947 and Nagasaki in 1948 by the US National Academy of Science and the National Research Council. The sample known as the Life Span Study (LSS) sample was later followed up by the successor of ABCC, the Radiation Effects Research Foundation (RERF) which substituted ABCC in 1975. The RERF cohort now includes the atomic bomb survivors (LSS), a cohort of those who were exposed in their mother's womb (in utter) and a cohort including the children of survivors who were conceived after paternal exposure (F1).

The results are regularly updated and published. The first results of this series were published in 1962 (Beebe et al. 1962). The study is based on the LSS cohort. One major finding of these studies was that an excess incidence of leukemia was observed while for other cancers (as a whole), the results were insignificant. Similar results persisted to show in future studies suggesting that leukemia should be the main focus of the study. This direction is still followed in ionizing radiation-cancer studies: Most studies consider leukemia separately.

The first study to publish results in this area is Beebe et al. (1962). This study is based on the follow-up of the LSS cohort from 1950 to 1958. The main results of this study are presented in Beebe et al. (1962; Table IX). This table shows that the p-value for the relative risk for leukemia is less than .01 while it is non-significant for that of other cancers.

Remark 2.1. One important point regarding this study is that the sample size has been relatively small at the time the study was carried out. Sampling which was based on sampling distance from the point where the bombs were detonated has been explained with details in the aforementioned reference. The sample size problem has also been addressed there. In the following years, more subjects were included and the sample size gradually grew. As explained before, a result of the smallness of the sample size is low power. Therefore, the non-significant p-value for "other cancer" in this study " might" reflect not the truth but it could be a result of low power to detect an effect.

The results for the LSS study for the period of 1950-1985 are provided in Shimizu et al. (1990). There are $\mathbf{75,991}$ eligible individuals included in this study (referred to as DS86) of which $\mathbf{5,936}$ individuals in the cohort died of cancer. Shimizu et al. (1990; Table II) presents the results of this study. Notice that the results of this study for leukemia alone, all cancers and all cancers except leukemia are all significant.

Remark 2.2. Table II in Shimizu et al. (1990), in particular, shows a very strong relationship between all cancers and radiation. As presented above, the very early study covering the period of 1950-1958 was unable to detect such an effect. It should be mentioned that the total number of deaths from all cancers at the time this study was carried out was $\mathbf{199}$ (compared with $\mathbf{5,936}$ deaths in the latter study). This observation "could" be a result of the difference between the powers in the two studies. This can emphasize the importance of taking the power into account before interpreting the results of any study.

Table 2 in Ozasa et al. (2019) provides the results for the cancer mortality and incidence rate based on the most recent studies of the LSS cohort. Notice that all the results show a positive ERR/Gy both for incidence and mortality for a full range of doses.

Remark 2.3. Notice that the part of Ozasa et al. (2019; Table 2) representing 'Lowest dose range with significant risk "can" give some indication of increase in the statistical power

with increase in the number of years of follow-up: For the incidence report, the evidence can clearly support this idea as the lowest effective dose has fallen from .15 Gy in 2007 to .10 Gy in 2015 and simultaneously the significance level has been reduced from .1 to .038. The results for mortality are not nearly as clear. The lowest effective dose has increased from .12 Gy in 2003 to .2 in 2012. However the fact that the significance level has been reduced from .1 to .05 should be taken into account.

2.2.2 Other ionizing radiation-cancer studies

The results of the atomic bomb studies rose the question of whether other sources of ionizing radiation are carcinogen. These studies can generally be divided into two types:

1. Natural background ionizing radiation studies
2. Occupational ionizing radiation studies

Natural background ionizing radiation studies

Background ionizing radiation is a source of radiation which has always been present in the environment (UN committee 2000). Based on United Nations report (UN committee 2000; Annex B), there are four major sources of natural ionizing radiation:

1. cosmic radiation (such as radiation emitted by the sun)
2. terrestrial radiation (such as radiation caused by the natural deposits of uranium)
3. radionuclides present in the air
4. radionuclides present in the food

Although published results in this area mostly support the harmful effects of ionizing radiation on the incidence of cancer, some studies favor a harmless effect. A review on the papers

of the first type is provided in Hendry et al. (2009) while Dobrzynski1 et al. (1998) reviews papers of the second type.

Some studies in this field are based on recruited individuals. However, major studies rely on national registries. Two important examples of such studies are the US surveillance, Epidemiological, and End Results (SEER) program (SEER 2009) and studies based on UK National Registry of Childhood Tumors (UK cancer study 1; 2, Wakeford et al. 2009) present results based on this database.

Some results on the relationship between ionizing radiation and childhood cancer in Great Britain based on the aforementioned database are presented in UK cancer study (1). The main results of this study are shown in UK cancer study (1; Table3).

Remark 2.4. Table 3 in UK cancer study (1) shows that the results for all dose categories are insignificant. This could be a reflection of truth or a result of low power. This paper has been evaluated in Little et al. (2010) in terms of power. It is shown there that the relatively small number of cases and controls results in a low power to detect the true effect. Therefore, the insignificant results of the paper are not surprising and should be expected due to low power.

Occupational ionizing radiation

Some occupations inevitably require exposure to higher-level doses of radiation than that of the general public. The most important of such occupations are related to nuclear power plants, uranium mines, health care units. Workers in each of these fields are not only exposed to the background ionizing radiation but also exposed to radiation-related to their profession. As a result, studies of mortality and morbidity of such occupations can provide some insight into the nature of the relationship between ionizing radiation and cancer. Such studies usually rely on national databases.

One of the oldest and largest of such databases is Canada's National Dose Registry (NDR

website 2020). NDR was founded in 1951 and has continually been in operation thereafter. This is the largest national database on occupational radiation. NDR contains the records of more than half a million workers of which $150,000$ are the present-day workforce (NDR website 2020).

Ashmore et al. (1998) published in 1998 is the first large scale mortality study based on the National Dose Registry of Canada. This study inspects the database for the period of 1951-1983. The study cohort consists of $256,425$ individuals. The results are presented both in terms of standardized mortality rate (SMR)which is based on comparing the death results with the general population of Canada and also in terms of Excess relative risk (ERR) which is based on comparing the exposed groups with the non-exposed group in the cohort. The results of these two measurements are contradictory! Results for the SMR show lower death risk for the cohort while ERR results prove something in contrast to that of SMR as the relative risk for the exposed groups is higher than that of the non-exposed group. This feature, less and more, shows in other occupational studies. It is widely believed that SMR is not a good measurement for occupational studies as the truth is masked by the so-called 'healthy worker effect'.

When each cancer is considered independently, the ERR results of the aforementioned study are mostly insignificant. However, for all cancers as a whole ERR is significantly non-zero. Table 7 in Ashmore et al. (1998) and Table 8 in Ashmore et al. (1998) (for males and females, respectively) presents the results of this study.

Remark 2.5. Significant results for 'all cancers' in Ashmore et al. (1998; Tables 7 and 8) could be a result of enough power due to a sufficient number of cancers as a whole while breaking down cancers into specific types reduces the number of cases and lowers the power. Therefore, insignificant results for each type of cancer independently "might" be attributable to low power.

Although NDR studies that include all types of radiation workers are few, there are more

studies solely based on nuclear workers in NDR. These types of studies started earlier with Gribbin et al. (1993) having been published in 1993. These studies are naturally based on fewer observations. The said study included only $8,977$ individuals. The results of this study even for solid cancer (i.e. all cancers as a unit) was insignificant. Zablotska et al. (2014) is a more recent study of this type published in 2013. The results of this study for solid cancer were still non-significant.

Remark 2.6. Comparison of the earlier NDR study reviewed above with the latter two studies "could" again emphasize the role of power as the first study that included a much larger dataset was able to detect a significant result for solid cancer while the last two which are based on smaller databases didn't detect any significant effect.

2.3 Power studies of the cohort and case-control designs

The results of a statistical test could turn out to be insignificant due to several reasons. One possibility is that the result truly reveals the real nature of the relationship, i.e. the exposure is not related to the outcome. However, it can also be a result of some defect in the study. Such a defect might be related to some sort of bias in the study design or due to low power which in turn can be attributed to an insufficient number of participants in the study.

There are a couple of works in the literature that try to address the latter problem of power, particularly for cohort and case-control studies. I focus on the cohort and case-control studies as the majority of studies in the area of ionizing radiation-cancer are of these types.

The results of one of the earliest studies to address the problem of power and sample size both for cohort and case-control studies were published in Schlesselman (1974). The paper deals with the simplest type of epidemiological studies where the relationship between the disease status and a binary exposure is of interest. Although the study is general, it

specifically solves the power/sample size problem for an epidemiological study that tries to answer the question of whether there is an increased risk of giving birth to a child with a congenital heart defect (CHD) among mothers who have oral contraceptive (OC) exposure three months before or after conception.
The formula provided power and sample size in the said reference is based on a simple assumption of the normal distribution for RR and OR in cohort and case-control studies respectively. The results showed that a case-control study is more powerful in detecting CHD/OC relationship.

In most case-control and cohort studies usually, more complex conditions result in the need for models that are more complicated to find the statistical power. The most natural problem that can be posed after handling discrete exposures is how to deal with continuous exposures. The simplest approach is to dichotomize the exposure to use the discrete formula. However, as noted in Lubin et al. (1990), since the risk patterns might be monotonic with increasing exposure, an explicit category of 'not exposed' might be not meaningful. Also, this approach has been proven to overestimate the required sample size (and equivalently underestimate the power)(Lubin et al. 1990).

Mckeown-eyssen and Thomas (1985) is one of the earliest papers to address the problem of continuous exposure. The paper considers case-control studies and proposes the idea of comparing the mean of exposure between cases and controls to test any effect of the exposure. Based on this idea the formula for power and sample size is provided. In particular, the following formula is given for the sample size,

$$n = \frac{2(t_\alpha - t_\beta)^2\sigma^2}{(\mu_1 - \mu_0)^2}$$

where t stands for the t-distribution, α and $1-\beta$ denote the level of the test and desired power respectively, σ^2 denotes variance of exposure and μ_0 and μ_1 is the mean of the exposure

in cases and controls. The paper also explains an apparent paradox in this formula: The formula suggests that smaller variability in exposure results in smaller sample size while epidemiological studies prove something in contrast. The model proposed by this paper has been specifically applied to cancer and diet problems.

Using a logistic regression model with an exposure x (either continuous or discrete), the probability of having the disease is given by

$$p(x;\alpha,\theta) = \frac{e^{\alpha+\theta x}}{1+e^{\alpha+\theta x}}. \tag{2.1}$$

This can provide the bases for calculating the power and sample size in a case-control study. However, in practice more complex conditions may arise which can not be handled well with this simple approach. A couple of such conditions are addressed in Lubin and Gail (1988). This reference tries to solve the power/sample size problem based on a fairly unified method which provides a general framework for all problems simultaneously. The models are specifically applied to the study of lung cancer risk associated with exposure in the home to radon-222 and its short progeny. The need to develop new power/sample size models all arose from this study but, as explained in the said paper, can be applied to any study with similar problems.

The first problem in this study came from the analysis of underground miners which showed that the standard exponential trend as in model (2.1) can not explain the true nature of relationship. Rather, it is well characterized by a linear model. They proposed the following model

$$\begin{aligned} p(x;\alpha,\theta) &= \frac{r(x)}{1+r(x)} \\ &= \frac{e^{\alpha}R(x;\alpha,\theta)}{1+e^{\alpha}R(x;\theta)} \end{aligned} \tag{2.2}$$

where $R(x) = 1 + \theta x$. Testing the null hypothesis of no association is then equivalent to $H_0 : \theta = 0$. Based on this model, formulas for power and sample size are provided.

By changing the definition of function R, they also provide the formula to handle a more complex situation that can not be dealt with using typical models. One such situation arises when one wishes to test curvilinearity in the trend. For example, there has been evidence that suggests that the increase in the relative odds of lung cancer from exposure to radon is not linear over the entire range of exposure. The model which is proposed is based on (2.2) and the following model based on which power and sample size for such a test are provided

$$R(x) = (1 + \beta x)e^{-\theta x}.$$

Test for curvilinearity is again $H_0 : \theta = 0$.

Other complex circumstances considered in this reference are:

1. Testing whether the gradient is equal to a specified non-zero value.

2. Testing for no effect after adjustment for second risk factor under a multiplicative model.

3. Testing for an additive joint association for two risk factors under a multiplicative alternative.

The first power model specifically designed for ionizing radiation-cancer studies is provided in Little et al. (2010). The model proposed tries to follow the usual traditions which are considered in epidemiological studies of ionizing radiation-cancer. This includes breaking the dataset into substrata where each substratum is a combination of age and sex. It also dichotomizes the ionizing radiation dose variable. The model has been used to detect the power of 'background ionizing radiation-childhood leukemia' studies in Great Britain. Such epidemiological studies have already been underway for several decades. However, the model

is fairly general so that it can be deployed in other settings such as 'Occupational ionizing radiation-solid cancer' studies. My research falls under the latter.

The model is obtained based on Monte-Carlo simulation. Age+sex strata are considered independently. The dose variable is broken down into a few categories. The model makes two basic assumptions:

1. The dose of ionizing radiation is assumed to have a linear effect on RR (in the cohort) and OR (in case-control). For example, in the cohort study, it is assumed that

$$\begin{aligned} RiskRatio &= \frac{\text{Risk for dose category with D}}{\text{Risk for category unexposed to radiation}} \\ &= 1 + \theta D. \end{aligned}$$

2. Fixed number of cases at each age/sex stratum. These numbers are obtained from the observed dataset.

The number of cases mentioned above serves as the number of trials in a multinomial model. To produce simulated samples (under both null and alternative), it is assumed that the number of cases at each age+sex stratum is distributed among dose categories according to a multinomial distribution the probabilities of which depend on parameter θ. More precisely, assume that we break down the i-th age+sex stratum into K dose categories. Assume also that in the observed data set, we have m_{ij} cases (deaths) in the j-th dose category of the i-th age+sex stratum and that the population proportion of the same category is p_{ij}. Then we assign a multinomial distribution to this stratum with number of trials given by

$$M_i = \sum_{j=1}^{K} m_{ij}$$

and corresponding probabilities calculated as

$$\pi_{ij} = \frac{p_{ij}(1 + \theta D_{ij})}{\sum_{j=1}^{K} p_{ij}(1 + \theta D_{ij})}.$$

The likelihood function and the score test statistic are then calculated based on the independent multinomial distributions. Using the Rao's theorem (Rao 1948), the score test statistic is assumed to have asymptotic normal distribution under the null which is $H_0 : \theta = 0$.

A sample of size $100,000$ is then randomly sampled from the alternative. The power is then calculated as the proportion of samples rejected. This procedure is repeated for every year of follow-up and power curves are drawn based on the information for all years of follow-up.

Results of this study provide two important pieces of information:

1. What is the best study design in terms of power? Figure 1 in Little et al. (2010) provides an answer to this question. As the graph shows, a cohort study with homogeneous ERR gives the largest power for any given year of follow up. Cohort study with heterogeneous ERR, ecological study and case-control study with 5 controls per case have slightly smaller power while case-control study with only one control per case has remarkably smaller power. Figure 2 in Little et al. (2010) presents the comparison between case-control studies with different numbers of control per case. It was suggested that the optimum number of controls per case is 5: This number is proposed based on the observation that fewer controls provide remarkably lower power while more controls don't significantly improve the power.

2. Some previous epidemiological studies which reported an insignificant result for the background ionizing radiation-childhood leukemia relationship were evaluated based on this power study. Many such studies were found to be underpowered. Consequently, the insignificant result of such studies can not reflect the true nature of the

relationship between ionizing radiation and childhood leukemia. Therefore, they should be interpreted with caution!

2.4 Conclusion

In this chapter, I gave a brief introduction to ionizing radiation. Sources of ionizing radiation and studies of ionizing radiation-cancer were reviewed. I also presented an overview of studies that provide a model to detect the power of cohort and case-control studies.

Chapter 3

Procedure to calculate the power for the cohort study

In this chapter, I develop the formal procedure which is used to estimate the power of the cohort design for ionizing radiation-cancer studies. The model which is used to estimate the power is called the "linear dose-effect" model which is explained with details in the following section. The procedure has been presented in Little et al. (2010) without details. I bridge all the gaps in the aforementioned reference. I spot some problems in the procedure and propose new approaches. As a result, I provide formulas that are slightly different from the formulas presented in that reference. I use simulated data to demonstrate that my formula for the power and sample size are more reliable.

I describe the algorithm which is used in Little et al. (2010) to calculate the power based on a normal-sampling approach: distribution of the score test statistic under null is assumed normal. The power is then calculated by sampling from the alternative. However, using simulated data (which is presented in Section 6.3) I realized that the score test statistic seems not only to be approximately normal under the null but it is also approximately normal under the alternative. In Corollary 5.1, I show why this seemingly unusual feature is the case (notice that Rao's theorem (Rao 1948) only states that the score test statistic is asymptotically normal under the null).

Similar results will be presented in the next chapter for the case-control studies.

Notice that for both study designs we use the score tests statistics. One could develop the model based on two other well-known test statistics, i.e. the Wald and the LRT. One

important advantage of the score statistic against the Wald statistic is that it doesn't require the point estimation for the parameter. Regarding the LRT, it is possible to reconstruct the models proposed in this chapter and the next chapter by replacing score with LRT. All the basics are the same, only the formula for the test statistic changes. Using the fact that the LRT is asymptotically normal (just as the score is), all the methods proposed with the score still work with the LRT. However, our main focus in this research is to improve th performance of the model proposed in Little et al. (2010) which is based on the score statistic.

3.1 Development of the procedure

The procedure for my study is developed based on a Monte-Carlo approach. I derive some basic information about the population under study from the given sample. I assume that any other possible sample (with the same size) would share some features with the current sample. The assumption is that after drawing samples, the total number of cases (deaths) in each stratum of age+sex (e.g., "males, age 30" or "females, age 32") will be the same as the observed value in our single sample. Variability in the samples, i.e. variation in the incidence of death from cancer, comes from the distribution of cases among dose groups at any stratum of sex+age. This variation is determined through multinomial distributions for any stratum in two steps:

1. The number of trials at each stratum is obtained from the observed sample (denoted by M_i).

2. I break down each stratum into K_i dose categories. The multinomial probabilities are then found based on a linear-dose effect model for the risk ratio as discussed below.

Figure 3.1 illustrates how I break down the dataset into age+sex strata and dose categories. Each row in this table represents one age+sex stratum. Also, Each column is corresponding to one dose category (in this table we assume $K_i = K$ for $I \leq i \leq I$). A multinomial distribution is assigned to each row (strata). In this table, pp stands for the population proportion of each cell in its row. Notice that

$$M_i = \sum_{j=1}^{K} m_{ij}, \quad \sum_{j=1}^{K} p_{ij} = 1 \quad 1 \leq i \leq I.$$

	Category 1 of dose	***Category 2 of dose***	...	***Category K of dose***
Stratum 1 of age+sex	cases=m_{11}, pp= p_{11}	cases=m_{12}, pp= p_{12}	...	cases=m_{1K}, pp= p_{1K}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Stratum I of age+sex	cases=m_{I1}, pp= p_{I1}	cases=m_{I2}, pp= p_{I2}	...	cases=m_{IK}, pp= p_{IK}

Figure 3.1: Break-down of dataset into age+sex strata and dose categories

Let $i \in I$ be the index for some stratum of age+sex, and let D_{ij} for $j \in K_i$ denote the average dose of some dose category in the i-th stratum. As mentioned above we would like to find the probability that some cases (whether incidence or death) in stratum i falls in category j of dose. To link probabilities of multinomial distributions to the excess relative risk, we use a linear dose-effect model for the risk ratio which is very common in the literature (See in particular BEIR V (1990)). Then, we will be able to derive the likelihood function

for the ERR parameter. The linear dose-effect model is defined as follows:

$$\begin{aligned}\text{Risk Ratio for category ij} &= \frac{\text{probability of being a case in the dose category } ij \text{ (mean dose=}D_{ij})}{\text{probability of being a case in the non-exposed category (Dose=0)}} \\ &= \frac{P(case - ij)}{P(case - nonexposed)} \\ &= 1 + \theta_i D_{ij}. \end{aligned} \tag{3.1}$$

This model is used to extract information regarding the probabilities of the multinomial distribution. Consequently, we can insert our desired parameters in the likelihood function. In this formula, "case-non exposed " means a case of cancer (either incidence or mortality, depending on study) int the non-exposed group. Notice that at this stage, for sake of generality, we assume that each stratum i has its specific ERR parameter θ_i. In practice, we usually assume one parameter for all strata.

The linear dose-effect model is one of the two common models used in the study of ionizing radiation and cancer, the other being the quadrative dose-effect model (BEIR V 1990). It has been proposed in (BEIR V 1990) that for low dose levels of ionizing radiation, the linear dose-effect model works well enough. So, we stick to this simpler model. Using equation (3.1) we proceed to find the probabilities of the multinomial distribution described above.

Remark 3.1. It should be mentioned that the main reference in which this procedure has been developed (Little et al. 2010) derives the formula for the probabilities of multinomial distributions using a different approach than what we present below. Actually, to derive the said formula they assume that $p_{ij}(1 + \theta_i D_{ij})$ is cancer risk where p_{ij} denotes the proportion of the population of stratum i which falls in category ij of dose. However, I prefer not to make this assumption as it seems that there is no evidence why the above value should fall between 0 and 1 (Notice that risk is a probability and should be between 0 and 1).

Using equation (3.1) we can write

$$P(case - ij) = P(\text{case-non exposed})(1 + \theta_i D_{ij}). \tag{3.2}$$

On the other hand,

$$\begin{aligned} P(case - ij) &= \frac{Number\ of\ cases\ in\ category ij}{size\ of\ ij} \\ &= \frac{M_{ij}}{p_{ij} T_i}. \end{aligned} \tag{3.3}$$

where p_{ij} is as defined above, T_i denotes the total population of stratum i and M_{ij} is the number of cases in category ij of dose. Now combining equations (3.2) and (3.3) we get

$$M_{ij} = p_{ij} T_i P(case\ non - exposed)(1 + \theta_i D_{ij}).$$

As a result total number of cases in stratum i is given by,

$$\begin{aligned} M_i &= \sum_{j=1}^{K_i} M_{ij} \\ &= \sum_{j=1}^{K_i} p_{ij} T_i P(case\ non - exposed)(1 + \theta_i D_{ij}) \end{aligned}$$

Now, we can provide π_{ij}'s, the probabilities of multinomial distribution at each stratum i. As mentioned above, π_{ij} is probability that a case in stratum i falls in category j of dose. It

is given by

$$\begin{aligned}\pi_{ij} &= \frac{M_{ij}}{M_i} \\ &= \frac{p_{ij}T_iP(case\ non-exposed)(1+\theta_iD_{ij})}{\sum_{j=1}^{K_i} p_{ij}T_iP(case-non\ exposed)(1+\theta_iD_{ij})} \\ &= \frac{p_{ij}(1+\theta_iD_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta_iD_{ij})}. \end{aligned} \tag{3.4}$$

Notice that in equation (3.4) if we set $\theta_i = 0$, then using the fact that $\sum_{j=1}^{K_i} p_{ij} = 1$, we would get $\pi_{ij} = p_{ij}$. This is natural as when there is no relationship between the exposure and outcome, the number of cases at each category is only proportional to the size (population) of the category.

Now, we are ready to provide formula for the Likelihood function. First of all notice that samples are drawn independently. Therefore, the general Likelihood function is given by $L = \prod_{i=1}^{I} L_i$ where l_i denotes the Likelihood function for stratum i. To calculate the formula for l_i notice that samples in stratum i are assumed to occur according to a multinomial distribution with M_i trials and probabilities $(\pi_{ij})_{j=1}^{K_i}$ given by equation (3.4). Therefore, if we assume that m_{ij} is the variable representing number of cases in category ij, then we will have

$$L_i(\theta_i|(m_{ij})_{j=1}^{K_i}) = \frac{M_i!}{\prod_{j=1}^{K_i} m_{ij}!} \prod_{j=1}^{K_i} \pi_{ij}^{m_{ij}}$$

Therefore,

$$\begin{aligned} L((\theta_i)_{i=1}^{I}|((m_{ij})_{j=1}^{K_i})_{i=1}^{I}) &= \prod_{i=1}^{I} L_i(\theta_i|(m_{ij})_{j=1}^{K_i}) \\ &= \left[\prod_{i=1}^{I} \frac{M_i!}{\prod_{j=1}^{K_i} m_{ij}!}\right]\left[\prod_{i=1}^{I}\prod_{j=1}^{K_i}\left(\frac{p_{ij}(1+\theta_iD_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta_iD_{ij})}\right)^{m_{ij}}\right] \end{aligned}$$

Notice that $c = \prod_{i=1}^{I} \frac{M_i!}{\prod_{j=1}^{K_i} m_{ij}!}$ does not depend on θ_i and is a fixed scalar. Now, we can rewrite

the above equation as

$$L((\theta_i)_{i=1}^{I}|((m_{ij})_{j=1}^{K_i})_{i=1}^{I}) = c\prod_{i=1}^{I}\frac{\prod_{j=1}^{K_i}(p_{ij}(1+\theta_i D_{ij}))^{m_{ij}}}{\left[\sum_{j=1}^{K_i}p_{ij}(1+\theta_i D_{ij})\right]^{\sum_{j=1}^{K_i}m_{ij}}}.$$

Finally, using the fact that $\sum_{j=1}^{K_i} m_{ij} = M_i$, we can write

$$L((\theta_i)_{i=1}^{I}|((m_{ij})_{j=1}^{K_i})_{i=1}^{I}) = c\prod_{i=1}^{I}\frac{\prod_{j=1}^{K_i}(p_{ij}(1+\theta_i D_{ij}))^{m_{ij}}}{[\sum_{j=1}^{K_i}p_{ij}(1+\theta_i D_{ij})]^{M_i}}.$$

Now, the log-Likelihood function is given by

$$\begin{aligned} l &= l((\theta_i)_{i=1}^{I}|((m_{ij})_{j=1}^{K_i})_{i=1}^{I}) \\ &= log(l((\theta_i)_{i=1}^{I}|((m_{ij})_{j=1}^{K_i})_{i=1}^{I})) \\ &= c' + \sum_{i=1}^{I}\Bigg\{\sum_{j=1}^{K_i} m_{ij} ln[p_{ij}(1+\theta_i D_{ij})] \\ &\quad - M_i ln[p_{ij}(1+\theta_i D_{ij})]\Bigg\}. \end{aligned} \tag{3.5}$$

So far we have assumed that each strata might have a different ERR parameter θ_i. Now, we assume that this parameter is the same for all strata. So, let $\theta_i = \theta$ for all $i \in I$. Carrying this fact in mind and using equation (3.5) we obtain the formula of the score test statistic,

$$\frac{dl}{d\theta} = \sum_{i=1}^{I}\left\{\sum_{j=1}^{K_i}\frac{m_{ij}D_{ij}}{1+\theta D_{ij}} - M_i\frac{\sum_{j=1}^{K_i}p_{ij}D_{ij}}{\sum_{j=1}^{K_i}p_{ij}(1+\theta D_{ij})}\right\} \tag{3.6}$$

As explained in the first chapter, the score test statistic is usually defied as $(\frac{dL}{d\theta})^2/var(\frac{dL}{d\theta})$ which has asymptotic chi-squared distribution. Since, we have only one unknown variable, the score test statistic has asymptotic chi-squared distribution with one degree of freedom under the null hypothesis. Using the fact that the square root of χ_1^2 is normal, we prefer to

work with $\frac{dL}{d\theta}/(var(\frac{dL}{d\theta}))^{1/2}$.

The goal of my research is to find the power to detect a true effect. An effect here is equivalent to a non-zero ERR value $\theta_1 \neq 0$. Therefore, our hypothesis test should be $H_0 : \theta = 0$. Then we need to check how often we truly reject H_0 in favor of $H_1 : \theta \neq 0$ when the samples come from the true value θ_1. So, we need to calculate the score statistic under $\theta = 0$. Using equation (3.6) this is given by

$$\frac{dL}{d\theta}|_{\theta=0} = \sum_{i=1}^{I} \left\{ \sum_{j=1}^{K_i} m_{ij} D_{ij} - M_i \sum_{j=1}^{K_i} p_{ij} D_{ij} \right\}. \tag{3.7}$$

As explained in the first chapter, in order to obtain a closed-form formula for the power (in particular to calculate a formula for the sample size) we need to know a theoretical distribution for the score test under the null both when samples come from $\theta = 0$ and when they come from another desired value such as θ_1 which we wish to be able to detect. As described there, under the null the standardized score test has χ^2_1 distribution (or equivalently its square root has a normal distribution) provided that sample size is large enough. However, under $\theta_1 \neq 0$ this might be not true. Actually, this is one reason why Little et al. (2010) chooses a sampling approach under θ_1 side to calculate the power.

The aforementioned reference needs to assume a normal distribution for the score test under the null when samples come from θ_1 too. In the following section, we will prove a theorem showing that in this specific study the score test under the null when samples come from θ_1 is as good normally distributed as it is when samples come from $\theta = 0$. As a result, we will be able to theoretically prove that the closed form-formula for the power is nearly as reliable as the sampling approach. Also, the sample size formula will find a stronger foundation. I use a different formula than the one provided in Little et al. (2010) to calculate the power closed-form formula and the sample size. Simulation studies that are presented in Appendix A show that my formula provides power estimates which are closer

to power estimates from the Monte-Carlo method.

We need to find the mean and the variance of score test statistic under the null when samples come from some parameter θ. Using equation (3.7) we can write

$$E_\theta[\frac{dl}{d\theta}|_{\theta=0}] = \sum_{i=1}^{I}\left\{\sum_{j=1}^{K_i} E_\theta[m_{ij}]D_{ij} - M_i\sum_{j=1}^{K_i} p_{ij}D_{ij}\right\}. \tag{3.8}$$

It should be pointed out that from now on an explicit subscript $\theta = 0$ means that the parameter is equal to 0 while a general parameter θ (such as in E_θ below) means that the parameter could be any arbitrary value (possibly non-zero). Notice that each m_{ij} is a binomial variable with $n = M_i$ trials and

$$p = \pi_{ij} = \frac{p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})}$$

success rate. Therefore, its mean is given by

$$\begin{aligned} E_\theta[m_{ij}] &= np = M_i\pi_{ij} \\ &= M_i\frac{p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})}. \end{aligned} \tag{3.9}$$

Putting equations (3.8) and (3.9) we get

$$\begin{aligned} E_\theta[\frac{dl}{d\theta}|_{\theta=0}] &= \sum_{i=1}^{I}\left\{\sum_{j=1}^{K_i} M_i\frac{p_{ij}(1+\theta D_{ij})D_{ij}}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})} - M_i\sum_{j=1}^{K_i} p_{ij}D_{ij}\right\} \\ &= \sum_{i=1}^{I} M_i\left\{\frac{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})D_{ij}}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})} - \sum_{j=1}^{K_i} p_{ij}D_{ij}\right\} \\ &= \theta\sum_{i=1}^{I} M_i\left\{\frac{\sum_{j=1}^{K_i} p_{ij}D_{ij}^2 - [\sum_{j=1}^{K_i} p_{ij}D_{ij}]^2}{1+\theta\sum_{j=1}^{K_i} p_{ij}D_{ij}}\right\}. \end{aligned} \tag{3.10}$$

Variance of the score test under the null when samples come from some θ can also be

calculated using the formula

$$var_\theta[\frac{dl}{d\theta}|_{\theta=0}] = cov(\frac{dl}{d\theta}|_{\theta=0}, \frac{dl}{d\theta}|_{\theta=0}) \\ = cov(\sum_{i=1}^{I}\sum_{j=1}^{K_i} m_{ij}D_{ij}, \sum_{i=1}^{I}\sum_{j=1}^{K_i} m_{ij}D_{ij}) \tag{3.11}$$

Notice that for i and $i' \in I$ all m_{ij} variables are independent from all $m_{i'j'}$ variables. Also, using properties of binomial variables, each m_{ij} has variance

$$var[m_{ij}] = np(1-p) = M_i\pi_{ij}(1-\pi_{ij}) \\ = M_i \frac{p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})}\left(1 - \frac{p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})}\right). \tag{3.12}$$

In addition, using properties of multinomial distribution, the covariance between two variables m_{ij} and $m_{ij'}$ is given by

$$cov(m_{ij}, m_{ij'}) = -M_i\pi_{ij}\pi_{ij'}.$$

Putting all this facts together and using equation (3.11) we can write

$$\begin{aligned} var_\theta[\frac{dl}{d\theta}|_{\theta=0}] &= \sum_{i=1}^{I} cov(\sum_{j=1}^{K_i} m_{ij}D_{ij}, \sum_{j=1}^{K_i} m_{ij}D_{ij}) \\ &= \sum_{i=1}^{I}\left\{\sum_{j=1}^{K_i}\left[D_{ij}^2 var_\theta[m_{ij}]\right] + 2\sum_{1\le j'<j\le K_i}\left[D_{ij}D_{ij'}cov(m_{ij}, m_{ij'})\right]\right\} \\ &= \sum_{i=1}^{I} M_i\Bigg\{\sum_{j=1}^{K_i}\left[D_{ij}^2\frac{p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})}\left(1 - \frac{p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})}\right)\right] \\ &\quad - 2\sum_{1\le j'<j\le K_i}\left[\frac{D_{ij}D_{ij'}p_{ij}p_{ij'}(1+\theta D_{ij})(1+\theta D_{ij'})}{\left(\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})\right)^2}\right]\Bigg\} \end{aligned}$$

$$= \sum_{i=1}^{I} M_i \left\{ \frac{\sum_{j=1}^{K_i} D_{ij}^2 p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})} - \left[\frac{\sum_{j=1}^{K_i} D_{ij} p_{ij}(1+\theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1+\theta D_{ij})} \right]^2 \right\}$$
$$= \sum_{i=1}^{I} M_i \left\{ \frac{\left[\sum_{j=1}^{K_i} D_{ij}^2 p_{ij} + \theta \sum_{j=1}^{K_i} D_{ij}^3 p_{ij}\right]\left[1+\theta \sum_{j=1}^{K_i} p_{ij} D_{ij}\right] - \left[\sum_{j=1}^{K_i} D_{ij} p_{ij} + \theta \sum_{j=1}^{K_i} D_{ij}^2 p_{ij}\right]^2}{\left[1+\theta \sum_{j=1}^{K_i} p_{ij} D_{ij}\right]^2} \right\}. \tag{3.13}$$

Now let

$$\mu_\theta = E[\frac{dl}{d\theta}|_{\theta=0}] \quad , \quad \sigma_\theta = \sqrt{var_\theta[\frac{dl}{d\theta}|_{\theta=0}]}.$$

Then,the distribution of the score test statistic under an arbitrary parameter θ is normal $Z(\mu_\theta, \sigma_\theta^2)$.

Next, we proceed to calculate a closed-form formula for power.

Remark 3.2. At this point, it should be pointed out that in the main reference (Little et al. 2010), the closed-form formula for provided for the power seems to be inappropriate. The reason is that one has to notice the difference between the variance of the score statistic under null and alternative. In many statistical studies of power the variance of test statistic under both null and alternative are equal. Therefore, in order to obtain the power they usually start from the standardized score statistic under null (i.e. score test divided by the SE under null). The results remain valid as the score statistic under the alternative is automatically standardized as well. However, equation (3.13) shows that in my study the variance (hence, SE) under null and alternative are unequal. Therefore, starting from standardized score statistic leads to inappropriate formula for power. This is what Little et al. (2010) does. In the sequel I take the a different approach and provide a different formula. In Section 6.5 I will use some simulated data to compare the power based on the closed-form formula given in Little et al. (2010) and closed-form formula proposed here. We will see that my formula provides results very close to the power result from the sampling method while results from the formula given in Little et al. (2010) are not close enough to the sampling values and

sometimes they are even completely irrelevant.

Notice that we need to calculate the distribution of score statistic under null $H_0 : \theta = 0$. Using formulas (3.13) and (3.10) we can see that

$$E_{\theta=0}[\frac{dl}{d\theta}|_{\theta=0}] = 0.$$

Also we have

$$var_{\theta=0}[\frac{dl}{d\theta}|_{\theta=0}] = \sum_{i=1}^{I} M_i \Big\{ \sum_{j=1}^{K_i} \Big[D_{ij}^2 p_{ij}(1-p_{ij}) \Big] - 2 \sum_{1 \le j' < j \le K_i} \Big[D_{ij} D_{ij'} p_{ij} p_{ij'} \Big] \Big\}. \tag{3.14}$$

Let $\sigma_0 = \sqrt{var_{\theta=0}[\frac{dl}{d\theta}|_{\theta=0}]}$, then score statistic under $\theta = 0$ is $Z(0, \sigma_0^2)$.

Remark 3.3. To calculate the power we can take samples from alternative and assume a normal distribution for the score statistic under null. The formula for the score statistic under null is

$$Z = \frac{dl}{d\theta}|_{\theta=0} \Big/ \sqrt{var_{\theta=0}[\frac{dL}{d\theta}|_{\theta=0}]}. \tag{3.15}$$

As this equation shows the variance in the denominator is the variance under null (as represented by subscript $\theta = 0$ in the variance). However, in Little et al. (2010) the variance under alternative is included which seems to be inappropriate in case one wishes to assume a standard normal distribution for the score statistic under the null. Table 1 and Table 2 in Little et al. (2010) show that the said reference assumes a standard normal distribution under null. Therefore, the variance in the normalized score statistic should be the one under the null. Using simulated data, I can show that if we use the formula proposed in Little et al.

(2010), then the distribution of the score test statistic under null is not standard normal. In Section 6.4, simulated data has been used to show that the score test under null is very well normal when we use equation (3.15) while it is not nearly as good when we use the formula proposed in Little et al. (2010).

To calculate the power assume that we have a test with significance level α. Also assume that, we would like to be able to detect an effect of size θ_1. The power of such a test is given by

$$\begin{aligned} P &= P(\text{Score test under } \theta_1 > Q_\alpha) \\ &= 1 - P(N(\mu_\theta, \sigma_\theta^2) < Q_\alpha) \end{aligned} \tag{3.16}$$

where Q_α is the $\%100(1-\alpha)$ quantile of the test statistic under null. In other words, we must have

$$\begin{aligned} 1-\alpha &= P(\text{Score test under } \theta_0 < Q_\alpha) \\ &= P(N(0, \sigma_0^2) < Q_\alpha) \\ &= P(N(0,1) < \frac{Q_\alpha}{\sigma_0}). \end{aligned}$$

Hence, $Z_{1-\alpha} = \frac{Q_\alpha}{\sigma_0}$ or equivalently,

$$Q_\alpha = \sigma_0 Z_{1-\alpha} \tag{3.17}$$

where $Z_{1-\alpha}$ is the $\%100(1-\alpha)$ quantile of the standard normal distribution. Using equations

(3.16) and (3.17) we can provide a closed-form formula for the power:

$$\begin{aligned} P &= 1 - P(N(\mu_\theta, \sigma_\theta^2) < Q_\alpha) \\ &= 1 - P(N(0,1) < \frac{\sigma_0 Z_{1-\alpha} - \mu_\theta}{\sigma_\theta}). \end{aligned} \tag{3.18}$$

Notice that, by contrast, the formula given in Little et al. (2010) is

$$P = 1 - P(N(0,1) < Z_{1-\alpha} - \frac{\mu_\theta}{\sigma_\theta}). \tag{3.19}$$

3.1.1 Formula for the sample size

Using formula (3.18) and under certain conditions, we are able to calculate the sample size in order to have tests with a desired power P. For this purpose we need to consider a single stratum. The reason is that we need to plug equation (3.13) in equation (3.18) and then solve it for the number of cases M_i, $i \in I$. However, if $I > 1$ we can not solve this equation strictly. Therefore, we need to assume $I = 1$.

Remark 3.4. Formula of the sample size is derived from the formula of the power. Since Little et al. (2010) provides a different formula for power, their formula for sample size which is based on equation (3.19) is different too. We use equation (3.18) to provide a different formula. In Section 6.6 we use some simulated data to compare the sample size derived from both formulas.

Let $I = 1$, $M_1 = M$ and $K_1 = K$. Then, from equation (3.18) we can write

$$1 - P = P(N(0,1) < \frac{\sigma_0 Z_{1-\alpha} - \mu_\theta}{\sigma_\theta}).$$

Therefore,

$$Z_{1-p} = \frac{\sigma_0 Z_{1-\alpha} - \mu_\theta}{\sigma_\theta} \tag{3.20}$$

where Z_{1-P} is the $\%100(1-P)$ quantile of the normal standard distribution. Let $Z_0 = \frac{\mu_\theta}{\sigma_\theta}$. Then, by equation (3.20) we have

$$Z_0 = (\frac{\sigma_0}{\sigma_\theta})Z_{1-\alpha} - Z_{1-P} \tag{3.21}$$

Notice that using equations (3.10) and (3.13) we can derive the following formula for Z_0,

$$Z_0 = \frac{\theta M F_1}{\sqrt{M F_2}} \tag{3.22}$$

where

$$F_1 = \frac{\sum_{j=1}^{K} p_j D_j^2 - \left(\sum_{j=1}^{K} p_j D_j\right)^2}{\sum_{j=1}^{K} p_j (1 + \theta D_{ij})} \tag{3.23}$$

and

$$F_2 = \frac{\sum_{j=1}^{K} \left\{ D_j^2 p_j (1+\theta D_j) \left(\left[\sum_{n=1}^{K} p_n (1+\theta D_n) \right] - p_j (1+\theta D_j) \right) \right\} - 2 \sum\limits_{1 \le j' < j \le K} p_j p_{j'} (1+\theta D_j)(1+\theta D_{j'})}{\left[\sum_{j=1}^{K} p_j (1+\theta D_j) \right]^2} \tag{3.24}$$

Using equations (3.21) and (3.22) we infer that

$$\theta \sqrt{M} \frac{F_1}{\sqrt{F_2}} = \frac{\sigma_0}{\sigma_\theta} Z_{1-\alpha} - Z_{1-P}$$

Therefore,

$$\theta^2 M \frac{F_1^2}{F_2} = (\frac{\sigma_0}{\sigma_\theta} Z_{1-\alpha} - Z_{1-P})^2$$

which implies

$$\begin{aligned} M &= \frac{F_2}{\theta^2 F_1^2} (\frac{\sigma_0}{\sigma_\theta} Z_{1-\alpha} - Z_{1-P})^2 \\ &= \frac{\sum_{j=1}^{K} \left\{ D_j^2 p_j (1+\theta D_j) \left(\left[\sum_{n=1}^{K} p_n (1+\theta D_n) \right] - p_j (1+\theta D_j) \right) \right\} - 2 \sum\limits_{1 \le j' < j \le K} p_j p_{j'} (1+\theta D_j)(1+\theta D_{j'})}{\theta^2 \left[\sum_{j=1}^{K} p_j D_j^2 - \left(\sum_{j=1}^{K} p_j D_j \right)^2 \right]^2} \\ &\quad \times (\frac{\sigma_0}{\sigma_\theta} Z_{1-\alpha} - Z_{1-P})^2. \end{aligned}$$

3.1.2 Sampling algorithm to calculate the power for the cohort study

Now that we have derived the formula for the score test statistic, we can design the sampling procedure to find the power for my study. The objective is to evaluate the power of the cohort study based on the years of follow up. The general algorithm is as follows:

- Step 0. Fix the effect size to be detected, i.e. fix parameter θ.
- Step 1. Fix number of years of follow-up.
- Step 2. Break down the dataset into I strata and each stratum into J cells. Strata I is defined as a combination of year (single years) and sex. Cells are defined as dose categories. Dose categories are defined according to the conventional method in the literature.
- Step 3. Obtain required information from the dataset (NDR). This includes the following information:

    1. M_i for $i \in I$: Number of cases (incidence or death) in stratum i.

2. D_{ij} for $i \in I$ and $j \in J$: Average dose in cell j of stratum i

3. p_{ij} for $i \in I$ and $j \in J$: Proportion of the population of cell ij in stratum i.

- Step 4. Sampling:

  - Step 4-1. Fix $i \in I$ and draw a sample from a multinomial distribution with M_i trials and J possible outcomes each with probability

$$\pi_{ij} = \frac{p_{ij}(1 + \theta_i D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1 + \theta_i D_{ij})}.$$

  - Step 4-2. Repeat Step 4-1 for all $i \in I$. Then using formula (3.15) calculate the value of the test statistic for this sample.

  - Step 4-3. Check whether the value from Step 4-2 is greater than $Z_{1-\alpha}$ (where α is the level of the test, usually $\alpha = .05$). If so, we mark this trial as a success.

- Step 5. Repeat Step 4 , $100,000$ times. The power is then calculated using the formula

$$power = \frac{\#\text{ of successful trials}}{100,000}$$

  where successful trial is defined in Step 4-3.

- Step 6. Return to Step 1 and choose another number of years of follow up.

- Step 7. If a different effect size is also of interest , return to Step 0 and change θ.

Chapter 4

Procedure to calculate the power for the case-control study

This chapter is the counterpart of Chapter 3 for the case-control study. This is designed, less and more, in the same way, that Chapter 3 was done. Again, a linear dose-effect model is used to estimate the power (see the following section). I explain, with all the details, the procedure which has been presented in Little et al. (2010). I bridge all the gaps and spot some problems in the procedure and provide substitute approaches. Similar to the cohort study I provide new formulas and use simulated data to prove that our formulas for the power and sample size are more reliable.

I describe the algorithm to calculate the power using a normal-sampling approach. Simulated data which appears in Section 6.3 show that the score test statistic for the case-control study is approximately normal under both null and alternative. This is a result of Corollary 5.2.

4.1 Development of the procedure

As mentioned in the first chapter, the measure of association in case-control studies is the odds ratio which compares the status of exposure in cases and controls. In our study, exposure is defined as 'being exposed to some level of radiation dose'. Therefore, in the case-control study levels of dose are of direct interest. So, we assume that there are $N_D + 1$

dose groups $0, 1, \cdots, N_D$ with associated doses

$$D_0 < D_1 < \cdots < D_{N_D}.$$

The dose categories are defined in the conventional way as used in the epidemiological studies.

For simplicity, at this stage we implement the case-control studies for each case independently. Later, we will put together cases which share similar characteristic in terms of age and sex. Let's assume that there are S cases. For each case i, we assign K_i controls. So, we have S independent case-control studies each with 1 case and K_i controls. For study $i \in S$ we define $(P_{d1i})_{d=0}^{N_D}$ to be probability that the case falls in groups d of dose. Similarly, assume that $(P_{d0i})_{d=0}^{N_D}$ represents probabilities for controls corresponding to case i. By definition

$$\text{Odds ratio} = \frac{\text{P(case exposed)/P(case non-exposed)}}{\text{P(control exposed)/P(control non-exposed)}}.$$

So, as long as we are concerned with level d of dose, the odds ratio in the case-control study i denoted by λ_d given by

$$\lambda_d = \frac{P_{d1i}/P_{01i}}{P_{d0i}/P_{00i}}. \tag{4.1}$$

Here by exposure we mean being exposed to level d of dose. At this point, two fact should be emphasized:

1. As presented in the first chapter, for studies with low incidence rate (such as cancer), odds ratio and risk ratio are approximately equal. As a result, if we apply a linear-dose effect model as proposed in BEIR V (1990), we can still apply the excess relative risk parameter θ for the case-control studies. Therefore, we must have

$$\lambda_d = \lambda_d(\theta) = 1 + \theta d. \tag{4.2}$$

2. The power procedure for the case-control study is more complex than that for the cohort study in several aspects. One reason, as evident in equation (4.1), is that we have 4 unknown parameters while we want to reduce our problem to a single parameter θ. We do this step by step. The first step is to apply equation (4.1) to write

$$P_{d1i} = \frac{P_{01i}}{P_{00i}} P_{d0i} \lambda_d. \tag{4.3}$$

Using this equation and the fact that $\sum_{d=0}^{N_D} P_{d1i} = 1$, we can write

$$\begin{aligned} \sum_{d=0}^{N_D} P_{d1i} &= \sum_{d=0}^{N_D} \frac{P_{01i}}{P_{00i}} P_{d0i} \lambda_d \\ &= \frac{P_{01i}}{P_{00i}} \sum_{d=0}^{N_D} P_{d0i} \lambda_d = 1. \end{aligned}$$

Therefore, we should have

$$\frac{P_{01i}}{P_{00i}} = \frac{1}{\sum_{d=0}^{N_D} P_{d0i} \lambda_d}. \tag{4.4}$$

Putting equations (4.3) and (4.4) together we conclude that

$$P_{d1i} = \frac{P_{d0i} \lambda_d}{\sum_{d=0}^{N_D} P_{d0i} \lambda_d}. \tag{4.5}$$

Notice that cases and controls are chosen independently (this is an ordinary case-control study not a matched one). Also, each case and each control can belong to $D+1$ dose groups. Therefore, for each $i \in I$ we have two binomial distributions:

- For case: With 1 trial and probabilities $(P_{d1i})_{i=1}^{I}$.
- For controls: With K_i trials and probabilities $(P_{d0i})_{i=1}^{I}$.

Let $(n_{d0i})_{d=0}^{N_D}$ and $(n_{d1i})_{d=0}^{N_D}$ denote the number of controls and cases in these multinomial

distributions that belong to group d of dose. So,

$$\sum_{d=0}^{N_D} n_{d1i} = 1 \quad , \quad \sum_{d=0}^{N_D} n_{d1i} = K_i. \tag{4.6}$$

Since the corresponding multinomial distribution for cases has only one trial, its probability distribution is given by

$$P(n_{01i}, \cdots, n_{d1i} | (P_{d1i})_{d=0}^{N_D}) = \prod_{d=0}^{N_D} P_{d1i}^{n_{d1i}}.$$

Similarly, for controls the corresponding probability distribution is

$$P(n_{00i}, \cdots, n_{d0i} | (P_{d0i})_{d=0}^{N_D}) = \prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}} \frac{K_i!}{\prod_{d=0}^{N_D} n_{d0i}!}.$$

Finally, since the I sets of case-controls are independent and in each case-control set, cases and controls are chosen independently, we see that the likelihood function is given by

$$\begin{aligned} L &= L(\theta, (P_{d0i})_{d=0}^{N_D}, (P_{d1i})_{d=0}^{N_D} | (n_{d0i})_{d=0}^{N_D}, (n_{d1i})_{d=0}^{N_D}) \\ &= \prod_{i=1}^{S} \left\{ \prod_{d=0}^{N_D} P_{d1i}^{n_{d1i}} \prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}} \frac{K_i!}{\prod_{d=0}^{N_D} n_{d0i}!} \right\}. \end{aligned} \tag{4.7}$$

If we replace P_{d1i} in equation (4.7) by the value given in equation (4.3) we can remove $(P_{d0i})_{d=0}^{N_D}$ and write

$$\begin{aligned} L &= L(\theta, (P_{d0i})_{d=0}^{N_D} | (n_{d0i})_{d=0}^{N_D}, (n_{d1i})_{d=0}^{N_D}) \\ &= \prod_{i=1}^{S} \left\{ \prod_{d=0}^{N_D} \left[\frac{P_{d0i}\lambda_d}{\sum_{d=0}^{N_D} P_{d0i}\lambda_d} \right]^{n_{d1i}} \prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}} \frac{K_i!}{\prod_{d=0}^{N_D} n_{d0i}!} \right\} \\ &= \left[\prod_{d=0}^{N_D} \lambda_d^{\sum_{i=1}^{S} n_{d1i}} \right] \left[\prod_{i=1}^{S} \frac{1}{\left(\sum_{d=0}^{N_D} P_{d0i}\lambda_d \right)^{\sum_{d=0}^{N_D} n_{d1i}}} \prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}+n_{d1i}} \frac{K_i!}{\prod_{d=0}^{N_D} n_{d0i}!} \right] \\ &= \left[\prod_{d=0}^{N_D} (1+\theta D_d)^{\sum_{i=1}^{S} n_{d1i}} \right] \left[\prod_{i=1}^{S} \frac{1}{\sum_{d=0}^{N_D} P_{d0i}(1+\theta D_d)} \prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}+n_{d1i}} \frac{K_i!}{\prod_{d=0}^{N_D} n_{d0i}!} \right] \end{aligned} \tag{4.8}$$

where the last equation is written according to the linear-dose effect model as given in equation (4.2) and the fact that $\sum_{d=0}^{N_D} n_{d1i} = 1$. Notice that in equation (4.8) apart from θ, we have other parameters which are $\left((P_{d0i})_{d=0}^{N_D}\right)_{i=1}^{S}$. Since we just want to test for θ, we need to remove other parameters by calculating the profile likelihood: We find MLE for all the parameters other than θ and plug all the MLEs in the likelihood function in order to obtain the profile likelihood.

To calculate MLE for $\left((P_{d0i})_{d=0}^{N_D}\right)_{i=1}^{S}$ notice that K_i , n_{d0i} and D_d are fixed. Also we treat θ as fixed at this point. Taking a glimpse at the likelihood function given in equation (4.8) reveals that this function has a nice simple form in that terms for each P_{d0i} appears as a multiplicand and there is no interaction between them for any two different values of i. As a result to maximize the likelihood function with respect to these parameters all we need is to maximize the functions for each i independently. Therefore, disregarding the fixed terms we need to maximize each of the following functions

$$f_i(p_{d0i}) = \frac{\prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}+n_{d1i}}}{\sum_{d=0}^{N_D} P_{d0i}(1+\theta D_d)}, \quad i \in I.$$

We need to solve I equations

$$\frac{\partial f_i}{\partial P_{d0i}} = 0. \tag{4.9}$$

We have,

$$\frac{\partial f_i}{\partial P_{d0i}} = \frac{(n_{d0i}+n_{d1i})P_{d0i}^{n_{d0i}+n_{d1i}-1}\prod_{d'\neq d}^{N_D} P_{d'0i}^{n_{d'0i}+n_{d'1i}}\left(\sum_{d'=0}^{N_D} P_{d'0i}(1+\theta D_{d'})\right) - (1+\theta D_d)\prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}+n_{d1i}}}{\left(\sum_{d=0}^{N_D} P_{d0i}(1+\theta D_d)\right)^2}.$$

Plugging in equation (4.9) we get

$$(1+\theta D_d)\prod_{d=0}^{N_D} P_{d0i}^{n_{d0i}+n_{d1i}} = (n_{d0i}+n_{d1i})P_{d0i}^{n_{d0i}+n_{d1i}-1}\prod_{d'\neq d}^{N_D} P_{d'0i}^{n_{d'0i}+n_{d'1i}}\left(\sum_{d'=0}^{N_D} P_{d'0i}(1+\theta D_{d'})\right).$$

Dividing both sides of this equation by

$$(1+\theta D_d)P_{d0i}^{n_{d0i}+n_{d1i}-1}\prod_{d'\neq d}^{N_D} P_{d'0i}^{n_{d'0i}+n_{d'1i}},$$

we find that

$$P_{d0i} = \frac{(n_{d0i}+n_{d1i})}{1+\theta D_d}\sum_{d'=0}^{N_D} P_{d'0i}(1+\theta D_{d'}). \tag{4.10}$$

Notice that from the multinomial distribution and equation (4.10) we can write,

$$\begin{aligned} 1 &= \sum_{d=0}^{N_D} P_{d0i} \\ &= \left(\sum_{d=0}^{N_D}\frac{(n_{d0i}+n_{d1i})}{1+\theta D_d}\right)\left(\sum_{d'=0}^{N_D} P_{d'0i}(1+\theta D_{d'})\right) \end{aligned}$$

This implies that

$$\sum_{d'=0}^{N_D} P_{d'0i}(1+\theta D_{d'}) = \frac{1}{\sum_{d=0}^{N_D}(n_{d0i}+n_{d1i})/(1+\theta D_d)}. \tag{4.11}$$

Finally, putting equations (4.10) and (4.11) we find the desired MLEs as follows

$$P_{d0i} = \frac{(n_{d0i}+n_{d1i})/(1+\theta D_d)}{\sum_{d'=0}^{N_D}(n_{d'0i}+n_{d'1i})/(1+\theta D_{d'})}. \tag{4.12}$$

Now we proceed to find the profile log-likelihood. Notice that in this procedure, all terms

that do not include parameter θ are assumed fix and we collect them under a generic term C. First, taking log from equation (4.8), we get

$$l = ln(L) = \sum_{d=0}^{N_D} \left\{ \left(\sum_{i=1}^{S} n_{d1i} \right) ln((1 + \theta D_d)) \right\} - \sum_{i=1}^{S} ln \left(\sum_{d=0}^{N_D} P_{d0i}(1 + \theta D_d) \right) + \sum_{d=0}^{N_D} (n_{d0i} + n_{d1i}) ln(P_{d0i}) + ln \left(\frac{K_i!}{\prod_{d=0}^{N_D} n_{d0i}!} \right). \tag{4.13}$$

The last term is fixed. Let's denote it by C. Also, notice that

$$\begin{aligned} \sum_{d=0}^{N_D} P_{d0i}(1 + \theta D_d) &= \sum_{d=0}^{N_D} \frac{[n_{d0i} + n_{d1i}]/(1 + \theta D_d)}{\sum_{d'=0}^{N_D} [n_{d'0i} + n_{d'1i}]/(1 + \theta D_{d'})] } \times (1 + \theta D_d) \\ &= \frac{\sum_{d=0}^{N_D} [n_{d0i} + n_{d1i}]}{\sum_{d'=0}^{N_D} [n_{d'0i} + n_{d'1i}]/(1 + \theta D_{d'})} \\ &= \frac{K_i + 1}{\sum_{d'=0}^{N_D} [n_{d'0i} + n_{d'1i}]/(1 + \theta D_{d'})}. \end{aligned}$$

Therefore,

$$ln \left(\sum_{d=0}^{N_D} P_{d0i}(1 + \theta D_d) \right) = ln(K_i + 1) - ln \left(\sum_{d'=0}^{N_D} [n_{d'0i} + n_{d'1i}]/(1 + \theta D_{d'}) \right). \tag{4.14}$$

Since $ln(K_i + 1)$ is fixed, we will collect it under the generic term C. We can also write:

$$\begin{aligned} ln(p_{d0i}) &= ln(n_{d0i} + n_{d1i}) - ln(1 + \theta D_d) \\ &\quad - ln(\sum_{d'=0}^{N_D} [n_{d'0i} + n_{d'1i}]/(1 + \theta D_{d'})). \end{aligned} \tag{4.15}$$

Again, we collect $ln(n_{d0i} + n_{d1i})$ in C. Now plug equations (4.14) and (4.15) in equation (4.13)

to get

$$
\begin{aligned}
l &= \left[\sum_{d=0}^{N_D}\left(\sum_{i=1}^{S} n_{d1i}\right) ln(1+\theta D_d)\right] + \sum_{i=1}^{S} ln\left(\sum_{d'=0}^{N_D}[n_{d'0i}+n_{d'1i}]/(1+\theta D_{d'})\right) \\
&+ \sum_{i=1}^{S}\sum_{d=0}^{N_D}(n_{d0i}+n_{d1i})\left[-ln(1+\theta D_d) - ln\left(\sum_{d'=0}^{N_D}[n_{d'0i}+n_{d'1i}/(1+\theta D_{d'})]\right) + C\right] \\
&= -\sum_{d=0}^{N_D}\left[\sum_{i=1}^{S} n_{d0i}\right] ln(1+\theta D_d) - \sum_{i=1}^{S}\sum_{d=0}^{N_D} n_{d0i} ln\left(\sum_{d=0}^{N_D}[n_{d0i}+n_{d1i}]/(1+\theta D_d)\right) + C \\
&= \sum_{d=0}^{N_D}\left[\sum_{i=1}^{S} n_{d0i}\right] ln(1+\theta D_d) - \sum_{i=1}^{S} K_i ln\left(\sum_{d=0}^{N_D}[n_{d0i}+n_{d1i}]/(1+\theta D_d)\right) + C.
\end{aligned}
$$

where the last two equations are written using equations (4.6).

Now we can calculate the score test statistic by taking derivative from equation (4.15) as follows.

$$
\begin{aligned}
S(\theta) = \frac{dl}{d\theta} &= -\sum_{d=0}^{N_D}\left[\sum_{i=1}^{S} n_{d0i}\right]\frac{1}{(1+\theta D_d)} D_d \\
&+ \sum_{i=1}^{S} K_i \frac{\sum_{d=0}^{N_D}\frac{[n_{d0i}+n_{d1i}]}{(1+\theta D_d)^2} D_d}{\sum_{d=0}^{N_D}\frac{[n_{d0i}+n_{d1i}]}{(1+\theta D_d)}}.
\end{aligned} \tag{4.16}
$$

To calculate the score test statistic under null we set $\theta = 0$ in equation (4.15). So that we will get,

$$
\begin{aligned}
S(0) &= \frac{dl}{d\theta}|_{\theta=0} \\
&= -\sum_{d=0}^{N_D}\left[\sum_{i=1}^{S} n_{d0i}\right] D_d + \sum_{i=1}^{S} K_i \frac{\sum_{d=0}^{N_D}[n_{d0i}+n_{d1i}]D_d}{\sum_{d=0}^{N_D}[n_{d0i}+n_{d1i}]} \\
&= -\sum_{d=0}^{N_D}\left[\sum_{i=1}^{S} n_{d0i}\right] D_d + \sum_{i=1}^{S} K_i \frac{\sum_{d=0}^{N_D}[n_{d0i}+n_{d1i}]D_d}{K_i+1}.
\end{aligned} \tag{4.17}
$$

Finally, if in equation (4.17) we collect all terms with n_{d0i} together and all terms with n_{d1i} together, then we will get

$$\begin{aligned} S(0) &= \sum_{i=1}^{S}\sum_{d=0}^{N_D}\left[-1+\frac{K_i}{K_i+1}\right]n_{d0i}D_d + \sum_{i=1}^{S}\frac{K_i}{K_i+1}\sum_{d=0}^{N_D}n_{d1i}D_d \\ &= -\sum_{i=1}^{S}\frac{1}{K_i+1}\sum_{d=0}^{N_D}n_{d0i}D_d + \sum_{i=1}^{S}\frac{K_i}{K_i+1}\sum_{d=0}^{N_D}n_{d1i}D_d. \end{aligned} \tag{4.18}$$

Similar to the cohort study, we will derive the formula for the variance and expectations in order to find the normal distribution of the score statistic under null and alternative. For this purpose notice that for each $1 \le i \le S$ and $0 \le d \le N_D$, n_{d0i} and n_{d1i} are binomial distributions with success rates P_{d0i} and P_{d1i} and number of trials K_i and 1 respectively. Also, equation (4.5) shows that

$$P_{d1i} = \frac{P_{d0i}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0i}(1+\theta D_d)}.$$

Recall that if $X \sim binomial(n,p)$, then

$$E[X] = np \quad , \quad var[X] = np(1-p).$$

Therefore, using equation (4.18) we can write

$$\begin{aligned} E_\theta[\frac{dl}{d\theta}|_{\theta=0}] &= -\sum_{i=1}^{S}\frac{1}{K_i+1}\sum_{d=0}^{N_D}E_\theta[n_{d0i}]D_d + \sum_{i=1}^{S}\frac{K_i}{K_i+1}\sum_{d=0}^{N_D}E_\theta[n_{d1i}]D_d \\ &= \sum_{i=1}^{S}\frac{1}{K_i+1}\sum_{d=0}^{N_D}K_i p_{d0i}D_d + \sum_{i=1}^{S}\frac{K_i}{K_i+1}\sum_{d=0}^{N_D}\frac{P_{d0i}(1+\theta D_d)}{\sum_{d=0}^{N_D}p_{d0i}(1+\theta D_d)}D_d \\ &= \sum_{i=1}^{S}\frac{K_i}{K_i+1}\sum_{d=0}^{N_D}\left[\frac{P_{d0i}(1+\theta D_d)}{\sum_{d=0}^{N_D}p_{d0i}(1+\theta D_d)} - p_{d0i}\right]D_d. \end{aligned} \tag{4.19}$$

To calculate the variance, note further that all S case-control sets are independent and so are cases and controls in each set of case-controls. Therefore, in equation (4.18), the only terms with possibly non-zero covariance are $(n_{d0i})_{d=0}^{N_D}$ among themselves and $(n_{d1i})_{d=0}^{N_D}$ among themselves, for each $1 \leq i \leq S$.

Recall that the covariance between two marginal binomial distributions X and Y that come from a multinomial distribution with n trials is $-nP_XP_Y$ where P_X and P_Y are the success rate for X and Y respectively. Using all this facts and equation (4.18) we can write

$$
\begin{aligned}
var_\theta(\frac{dl}{d\theta}|_{\theta=0}) &= cov_\theta(\frac{dl}{d\theta}|_{\theta=0}, \frac{dl}{d\theta}|_{\theta=0}) \\
&= \sum_{i=1}^{S}\left[\frac{1}{K_i+1}\right]^2\left[\sum_{d=0}^{N_D} var_\theta(n_{d0i})D_d^2 - 2\sum_{1\leq d<d'\leq N_D} D_dD_{d'}cov(n_{d0i}, n_{d'0i})\right] \\
&+ \sum_{i=1}^{S}\left[\frac{K_i}{K_i+1}\right]^2\left[\sum_{d=0}^{N_D} var_\theta(n_{d1i})D_d^2 - 2\sum_{1\leq d<d'\leq N_D} D_dD_{d'}cov(n_{d1i}, n_{d'1i})\right] \\
&= \sum_{i=1}^{S}\frac{1}{[K_i+1]^2}\left[\sum_{d=0}^{N_D} D_d^2 p_{d0i}(1-p_{d0i}) - 2\sum_{1\leq d<d'\leq N_D} D_dD_{d'}p_{d0i}p_{d'0i}\right] \\
&+ \sum_{i=1}^{S}\left[\frac{K_i}{K_i+1}\right]^2\left[\sum_{d=0}^{N_D} D_d^2 p_{d1i}(1-p_{d1i}) - 2\sum_{1\leq d<d'\leq N_D} D_dD_{d'}p_{d1i}p_{d'1i}\right] \\
&= \sum_{i=1}^{S}\frac{K_i}{[K_i+1]^2}\left[\sum_{d=0}^{N_D} D_d^2 p_{d0i}(1-p_{d0i}) - 2\sum_{1\leq d<d'\leq N_D} D_dD_{d'}p_{d0i}p_{d'0i}\right] \\
&+ \sum_{i=1}^{S}\left[\frac{K_i}{K_i+1}\right]^2\left[\sum_{d=0}^{N_D} D_d^2 \frac{P_{d0i}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0i}(1+\theta D_d)}\left(1-\frac{P_{d0i}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0i}(1+\theta D_d)}\right)\right. \\
&\qquad \left. - 2\sum_{1\leq d<d'\leq N_D} D_dD_{d'}\frac{P_{d0i}p_{d'0i}(1+\theta D_d)(1+\theta D_{d'})}{\left[\sum_{d=0}^{N_D} p_{d0i}(1+\theta D_d)\right]^2}\right].
\end{aligned}
\tag{4.20}
$$

In particular, if we assign the same number of controls to each case so that for each $1 \leq i \leq S$,

$K_i = K$ and assume that $P_{d0i} = P_{d0}$ and $P_{d1i} = P_{d1}$, then equation (4.19) can be simplified to

$$E_\theta[\frac{dl}{d\theta}|_{\theta=0}] = \frac{S\,K}{K+1}\sum_{d=0}^{N_D}\left[\frac{P_{d0}(1+\theta D_d)}{\sum\limits_{d=0}^{N_D} p_{d0}(1+\theta D_d)} - P_{d0}\right]D_d. \tag{4.21}$$

Also equation (4.20) can be written as

$$\begin{aligned} var_\theta[\frac{dl}{d\theta}|_{\theta=0}] = \frac{S\,K}{[K+1]^2}&\left[\sum_{d=0}^{N_D} D_d^2 p_{d0}(1-p_{d0}) - 2\sum_{1\le d<d'\le N_D} D_d D_{d'} p_{d0} p_{d'0}\right] \\ &+ S\left[\frac{K}{K+1}\right]^2\left[\sum_{d=0}^{N_D} D_d^2 \frac{P_{d0}(1+\theta D_d)}{\sum\limits_{d=0}^{N_D} p_{d0}(1+\theta D_d)}\left(1 - \frac{P_{d0}(1+\theta D_d)}{\sum\limits_{d=0}^{N_D} p_{d0}(1+\theta D_d)}\right)\right. \\ &\left. - 2\sum_{1\le d<d'\le N_D} D_d D_{d'} \frac{P_{d0}p_{d'0}(1+\theta D_d)(1+\theta D_{d'})}{\left[\sum\limits_{d=0}^{N_D} p_{d0}(1+\theta D_d)\right]^2}\right]. \end{aligned} \tag{4.22}$$

So far, we have considered only one general strata. However, similar to the cohort procedure we would like to break down the dataset into strata of sex and age in order to follow the convention in the epidemiological studies. So, assume that we have M independent strata (studies) and implement each study in the way that was just presented. Notice that in this case the likelihood function can be written as

$$L = \prod_{m=1}^{M} L_m,$$

where each L_m is the likelihood function of the m-th strata. Therefore, assuming that in all strata and for each case we assign K controls, the profile log-likelihood using equation (4.18)

would be

$$\frac{dl}{d\theta}|_{\theta=0} = \frac{dlog(L)}{d\theta}|_{\theta=0} = \sum_{m=1}^{M}\left[-\sum_{i=1}^{S_m}\frac{1}{K+1}\sum_{d=0}^{N_{D,m}} n_{d0im}D_{md} + \sum_{i=1}^{S_m}\frac{K}{K+1}\sum_{d=0}^{N_{D,m}} n_{d1im}D_{md}\right]. \tag{4.23}$$

Here, S_m represents total number of cases in stratum m and D_{md} represents the average dose of category d of dose in stratum m. Also, n_{d0im} (n_{d1im}) is total number of controls (cases) in the i-th case-control set in stratum m that falls in category d of dose. Notice that as explained

1. n_{d0im} is determined from a multinomial distribution with 1 trial and probabilities given by

$$P_{d1m} = \frac{P_{d0m}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0m}(1+\theta D_d)}.$$

2. n_{d1im} is obtained from a multinomial distribution with K trials and probabilities P_{d0m}.

We can simplify equation (4.23). Notice that all case-control sets are independent and in each stratum m, all cases follow similar multinomial distributions and so do controls. Therefore, by putting terms with common distribution in equation (4.23) we can conclude that

$$\frac{dl}{d\theta}|_{\theta=0} = \frac{dln(L)}{d\theta}|_{\theta=0} = \sum_{m=1}^{M}\left[-\frac{1}{K+1}\sum_{d=0}^{N_{D,m}} n_{d0m}D_{md} + \frac{K}{K+1}\sum_{d=0}^{N_{D,m}} n_{d1m}D_{md}\right]. \tag{4.24}$$

In this formula n_{d0m} and n_{d1m} are determined as follows

1. n_{d0m} is determined from a multinomial distribution with S_m trial (where S_m is number of cases in stratum m) and probabilities given by

$$P_{d1m} = \frac{P_{d0m}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0m}(1+\theta D_d)}.$$

2. n_{d1m} is obtained from a multinomial distribution with KS_m trials and probabilities P_{d0m}.

From equation (4.24) we find the mean of the score statisitc as follows

$$E_\theta[\frac{dl}{d\theta}|_{\theta=0}] = \sum_{m=1}^{M}\left[-\frac{1}{K+1}\sum_{d=0}^{N_{D,m}} Kp_{d0m}D_{md} + \frac{K}{K+1}\sum_{d=0}^{N_{D,m}} S_m K D_{md} p_{d1m}\right]$$
$$= \frac{K}{K+1}\sum_{m=1}^{M} S_m \sum_{d=0}^{N_{D,m}}\left[\frac{P_{d0m}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0m}(1+\theta D_d)} - P_{d0m}\right] D_{md}. \tag{4.25}$$

Also, using the same argument as that used to derive equation (4.20), I can propose the following formula for the variance of this latest model

$$var_\theta[\frac{dl}{d\theta}|_{\theta=0}] = \frac{K}{[K+1]^2}\sum_{m=1}^{M} S_m\left[\sum_{d=0}^{N_{D,m}} D_{md}^2 p_{d0m}(1-p_{d0m}) - 2\sum_{1\le d<d'\le N_D} D_{md}D_{md'}p_{d0m}p_{d'0m}\right]$$
$$+\left[\frac{K}{K+1}\right]^2\sum_{m=1}^{M}\left[\sum_{d=0}^{N_{D,m}} D_{md}^2\frac{P_{d0m}(1+\theta D_d)}{\sum_{d=0}^{N_{D,m}} p_{d0m}(1+\theta D_d)}\left(1-\frac{P_{d0m}(1+\theta D_d)}{\sum_{d=0}^{N_{D,m}} p_{d0m}(1+\theta D_d)}\right)\right.$$
$$\left. - 2\sum_{1\le d<d'\le N_D} D_{md}D_{md'}\frac{P_{d0m}p_{d'0m}(1+\theta D_d)(1+\theta D_{d'})}{\left[\sum_{d=0}^{N_{D,m}} p_{d0m}(1+\theta D_d)\right]^2}\right]. \tag{4.26}$$

Equations (4.24), (4.25) and (4.26) provide the framework to calculate the power for case-control studies.

Remark 4.1. In Chapter 3 some substitutes to the formulas in Little et al. (2010) were proposed in order to derive the algorithm for the cohort study. Similar substitutes can be made in providing the formulas for the case-control study. The first required substitute is similar to that explained in Remark 3.2. Actually, to calculate the power based on a normal-sampling approach (i.e. assuming normal distribution for the score statistic under the null and sampling from alternative) the following formula which is assumed to be approximately

normally distributed is used for the score test under null

$$Z = \frac{dl}{d\theta}|_{\theta=0} \Big/ \sqrt{var_\theta[\frac{dl}{d\theta}|_{\theta=0}]}. \tag{4.27}$$

So, the variance included is under the parameter which is of interest (θ). However, the same explanation as that in Remark 3.2 proves that the formula should be as follows,

$$Z = \frac{dl}{d\theta}|_{\theta=0} \Big/ \sqrt{var_{\theta=0}[\frac{dl}{d\theta}|_{\theta=0}]}. \tag{4.28}$$

I can propose a closed-form formula for the power of case-control studies by assuming a normal distribution for the score statistic under alternative. In the next Chapter I prove the fact that the score statistic is as good normally distributed under the alternative as it is under the null. The closed-form formula is as follows. It is obtained from the same argument as that for the cohort study. So, I skip the proof.

$$\begin{aligned} P &= 1 - P(N(\mu_\theta, \sigma_\theta^2) < Q_\alpha) \\ &= 1 - P(N(0,1) < \frac{\sigma_0 Z_{1-\alpha} - \mu_\theta}{\sigma_\theta}). \end{aligned} \tag{4.29}$$

where α is the level of the test and

$$\mu_\theta = E_\theta[\frac{dl}{d\theta}|_{\theta=0}] \quad , \quad \sigma_\theta = \sqrt{var_\theta[\frac{dl}{d\theta}|_{\theta=0}]}$$

are calculated from equations (4.19) and (4.20).

Remark 4.2. Notice that just like the formula for the power of the cohort study, the formula given in Little et al. (2010) for the power of case-control study needs to be changed as it doesn't take into account difference between variances of the score statistic under null and alternative. For more information see Remark 3.2. In Section 6.5, some simulated data has

been used that shows this fact.

4.1.1 Formula for the sample size

To provide a formula for the sample size, we need to assume a single stratum in order to be able to solve the problem for S. From equation (4.29) we can write

$$Z_0 = (\frac{\sigma_0}{\sigma_\theta})Z_{1-\alpha} - Z_{1-P} \tag{4.30}$$

where $Z_0 = \frac{\mu_\theta}{\sigma_\theta}$. Notice that here $\mu_\theta = E_\theta[\frac{dl}{d\theta}|_{\theta=0}]$ and $\sigma_\theta = var_\theta[\frac{dl}{d\theta}|_{\theta=0}]$ are calculated from equations (4.21) and (4.22). One can also check that equation (4.21) can be simplified to

$$E_\theta[\frac{dl}{d\theta}|_{\theta=0}] = \theta\frac{SK}{K+1}\frac{\sum_{d=0}^{N_D} p_{d0}D_d^2 - \left[\sum_{d=0}^{N_D} p_{d0}D_d\right]^2}{\sum_{d=0}^{N_D} p_{d0}(1+\theta D_d)}. \tag{4.31}$$

Therefore, we can write

$$Z_0 = \frac{\theta S H_1}{\sqrt{S}H_2}$$

where H_1 and H_2 are defined appropriately by applying equations (4.22) and (4.31). Now using a similar argument as that used in subsection 3.1.1, one can derive the following formula for the required number of cases in a case-control study with significance level α and desired power P

$$S = \left\{\frac{\sum_{d=0}^{N_D} D_d^2 p_{d0}(1-p_{d0}) - 2\sum_{1\le d<d'\le N_D} D_d D_{d'} p_{d0}p_{d'0}}{K\theta^2\left[\frac{\sum_{d=0}^{N_D} p_{d0}D_d^2 - \left[\sum_{d=0}^{N_D} p_{d0}D_d\right]^2}{\sum_{d=0}^{N_D} p_{d0}(1+\theta D_d)}\right]^2}\right.$$

$$+\frac{\sum_{d=0}^{N_D} D_d^2 \frac{P_{d0}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0}(1+\theta D_d)}\left(1-\frac{P_{d0}(1+\theta D_d)}{\sum_{d=0}^{N_D} p_{d0}(1+\theta D_d)}\right)-2\sum_{1\le d<d'\le N_D} D_d D_{d'}\frac{P_{d0}p_{d'0}(1+\theta D_d)(1+\theta D_{d'})}{\left[\sum_{d=0}^{N_D} p_{d0}(1+\theta D_d)\right]^2}}{\theta^2\left[\frac{\sum_{d=0}^{N_D} p_{d0}D_d^2-\left[\sum_{d=0}^{N_D} p_{d0}D_d\right]^2}{\sum_{d=0}^{N_D} p_{d0}(1+\theta D_d)}\right]^2}\Bigg\}$$

$$\times\left(\frac{\sigma_0}{\sigma_\theta}Z_{1-\alpha}-Z_{1-P}\right)^2. \tag{4.32}$$

Remark 4.3. Since the main reference uses a different closed-form formula, the formula provided for the sample size of case-control studies is different too. In Section 6.6 I use some simulated data to compare the required sample size based on formulas given here and in the main reference.

4.1.2 Sampling algorithm to calculate the power for the case-control study

Using the formula derived above for the score test statistic, we can implement the sampling procedure to find the power for the case-control studies. The algorithm for this procedure is as follows:

- Step 0. Fix the effect size to be detected (i.e. parameter θ) and number of controls per case (K).
- Step 1. Fix number of years of follow-up.
- Step 2. Break down the dataset into M strata where each stratum is defined based on a combination of age and sex. Then, break down each strata into N_D cells defined by dose categories.
- Step 3. Obtain required information from the dataset. This include:
    1. S_m: Number of cases in stratum m.

2. D_{md}: Average dose of individuals in category d of dose in stratum m.

3. $\left((P_{d0m})_{d=0}^{N_D}\right)_{m=1}^{M}$: Probability that a non-diseased individual in stratum m falls in category d of dose. These parameters can be calculated by dividing the total number of non-diseased individuals in dose category d in stratum m by the total number of non-diseased individuals in the stratum. Notice that since cancer is a rare disease, P_{d0m} is approximately equal to the proportion of category d in stratum m.

- Step 4. Sampling procedure:
  - Step 4-1. Fix an stratum $m \in M$:
    - Step 4-1-1. Find one sample $(n_{d0m})_{d=0}^{N_D}$ by drawing from a multinomial distribution with KS_m trials and probabilities P_{d0m}.
    - Step 4-1-2. Find one sample $(n_{d1m})_{d=0}^{N_D}$ by drawing from a multinomial distribution with S_m trials and probabilities

$$P_{d1m} = \frac{P_{d0m}(1 + \theta D_d)}{\sum_{d=0}^{N_D} P_{d0m}(1 + \theta D_d)}.$$

  - Step 4-2. Repeat Step 4-1 for all $m \in M$ in order to find a complete sample. Then calculate the value of the score test statistic for this sample using equation (4.28).
  - Step 4-3. Check whether the value from step 4-1 is greater than $Z_{1-\alpha}$ (α being the level of the test, usually $\alpha = .05$). If so, mark this trial as a success.
- Step 5. Repeat Step 4, $100,000$ times. The power is then calculated from

$$power = \frac{\#\ of\ successful\ trials}{100,000}$$

where successful trial is defined in Step 4-3.

- Step 6. Return to Step 1 and choose another number of years of follow up.
- Step 7. For a different effect size or/and different number of controls per case return to Step 0 and change θ or/and K.

Chapter 5

Further discussions on the procedures

In this chapter, we present some discussions that provide a new approach or some substitute for the methods discussed in the previous chapters. In particular, we prove the normality of the score test statistic under both null and alternative for both cohort and case-control studies. We also argue that the formula for the Monte-Carlo Error (MCE) given in Little et al. (2010) needs to be modified. We propose some possible substitutes. We discuss three possible approaches that can be used to calculate the power for the epidemiological studies. This includes normal-sampling, normal-normal and sampling-sampling approaches. It should be pointed out that Little et al. (2010) uses a normal-sampling approach to calculate power. We also argue that the procedure presented for the cohort study can be used in studies the measure of association of which is SIR (standardized incidence ratio) or SMR (standardized mortality rate).

5.1 Monte-Carlo Error

When using Monte-Carlo estimation to evaluate any quantity, one usually gets some error. The reason is that the finite number of samples is drawn and based on this samples the desired quantity is estimated. However, if one draws another set of samples and recalculates the estimated quantity based on the new samples, a different result will come out. Although in most cases this difference is slight, there is no guarantee to get the same results. For

instance, in our context we draw $100,000$ samples from alternative. Based on these samples we find an approximate distribution for the score statistic under the alternative. Now, if we draw another $100,000$ samples we will get another distribution which is (probably slightly) different from the first one. Therefore, the value that we obtain for power would be different. This fact reflects the issue of Monte-Carlo Error (MCE).

The MCE index in simulated studies is very important and should be addressed. Based on MCE one can justify the number of repetitions in the Monte-Carlo procedure. However, as noted in Koehler et al. (2009), most studies that use the Monte-Carlo procedure don't report MCE and don't justify their number of repetitions. It is reported in Koehler et al. (2009) that out of 223 reviewed papers only 8 report MCE. my main reference (Little et al. 2010) reports the MCE using the formula $\frac{1}{\sqrt{N}}$ where N is the number of repetitions in the Monte-Carlo. Notice that there is no general formula for MCE and this index should be proposed based on the quantity of interest and the Monte-Carlo procedure to estimate it. Some formulas for MCE to be used in different contexts are proposed in Koehler et al. (2009).

Here, we justify the formula proposed in Little et al. (2010) for MCE and explain why it needs adjustment. We propose other approaches to deal with MCE. As mentioned above, in my context different sample sets provide different estimated distributions for the score statistic under the alternative. Therefore, we need a measure to compare these two distributions. One possible approach is to calculate the average value of the distributions and evaluate the precision of the Monte-Carlo based on the precision of this average. In other words, we report the standard deviation of the sample means as the MCE whereby a sample we mean a full simulated data from Monte-Carlo. According to CLT, this is given by $\frac{\sigma_S}{\sqrt{N}}$ where σ_S is the standard deviation of the normalized score test statistic under the alternative and N is number of repetitions. It is assumed in Little et al. (2010) that $\sigma_S = 1$. In the sequel, we explain why this seems to be improper. The problem arises from the fact that in the aforementioned reference the variance included in the formula of normalized score statistic

is the variance under alternative (See Remark 3.3 and Remark 4.1). Since we are applying the Monte-Carlo procedure to draw from alternative, the variance of the normalized score statistic under alternative turns out to be 1. Therefore, regardless of the effect size (θ) and the year of follow-up we always get a fixed MCE which is $\frac{1}{\sqrt{N}}$. Unfortunately, the issue is not this simple! As explained in Remark 3.3, in such a case one can not assume the score statistic under the null to be standard normal. In other words, there is a trade-off between 'permanent standard normal distribution under the null' and 'fixed MCE'. If we would like to use the formula proposed in Little et al. (2010) for the normalized score statistic, then each time, depending on the alternative parameter θ and the year of follow-up, we need to adjust the variance of the score statistic under the null to obtain the correct quantile corresponding to the test significance level (α). On the other hand, if we would like to always keep the score statistic under the null standard normal, we need to use the variance under the null to normalize the score statistic (as explained in Remark 3.3 and Remark 4.1). This face is not taken into account in Little et al. (2010) and at the same time assumes a standard normal distribution for the score statistic under the null (which is evident from Table 1 and Table 2 in the said reference) and a fixed MCE.

It seems that a permanent standard normal distribution under null is more useful in my context as it relaxes the need to adjust the score statistic under the null for different values of θ and different years of follow-up. However, in such a case we get $\sigma_S = \frac{\sigma_\theta}{\sigma_0}$ where σ_θ and σ_0 are the variance of the score statistic under the alternative and null, respectively. As a result, the MCE with N repetition is given by.

$$MCE = \frac{\sigma_\theta/\sigma_0}{\sqrt{N}}. \tag{5.1}$$

Notice that this value depends on the study design, θ and years of follow-up. Assume that we fix θ. To deal with the problem of MCE, one possible solution is to pick one value of

N such that MCE for all the years of follow-up is smaller than the desired value. Another possible approach is to obtain a fixed MCE by not using a fixed number of repetitions for different years of follow-up but determine the number of repetitions independently. Using Equation 5.1, we can see that the formula for the number of repetitions for each year of follow-up is

$$N_Y = \left[\left(\frac{\sigma_\theta/\sigma_0}{MCE}\right)^2\right] + 1.$$

where $[\cdot]$ stands for the integer part. It is recommended in Little et al. (2010) that .0032 as a low MCE. In such a case, for each year of follow-up, the required number of repetition is

$$N_Y = \left[\left(\frac{\sigma_\theta/\sigma_0}{.0032}\right)^2\right] + 1.$$

In Section 6.2, simulated data has been used to compare the number of iterations that are required to attain $MCE = .0032$.

5.2 Asymptotic normality of the score test statistic under null and alternative

The well-known theorem of Rao, as presented in the first chapter, states that the score test statistic under null is asymptotically normal. This means that assuming a large sample size if one draws samples from null and plug in the score test statistics (at the null value), he or she will get an approximately normal distribution. However, when it comes to alternative, the Rao's theorem doesn't explicitly say anything. For the procedures proposed in this study, we can see that the corresponding score statistic formulas have a nice form that forces them to be asymptotically normal under both null and alternative for both cohort and case-control studies. We present the proof below.

The key tool in my approach is to rely on an indirect application of Rao's theorem rather

than using a direct application as presented in the following theorem.

Theorem 5.1. Let M be a multinomial distribution with l possible outcomes with probabilities $(p_1, \cdots, p_k)$. Let D_i for $1 \leq i \leq k$ be some real numbers with

$$0 < D_1 < D_2 < \cdots < D_k.$$

Then $\sum_{i=1}^{k} m_i D_i$ has an asymptotic normal distribution where each m_i is the binomial distribution derived from multinomial M for the i-th outcome.

Proof. We set up an artificial experiment as follows: Assume that we have n dose categories with average dose D_i in each category i for $1 \leq i \leq k$. Also, assume that we have a population which is distributed among the dose categories with proportions p_i for $1 \leq i \leq k$. Let's here denote the ERR parameter with β and assume a linear dose effect model, i.e.

$$\text{Risk ratio for group i of dose} = 1 + \beta D_i.$$

Then, using equation (3.7), we see that for such an experiment the score test statistic (assuming that we are testing for $\beta = 0$) is given by

$$\frac{dL}{d\beta}|_{\beta=0} = \sum_{i=1}^{k} m_i D_i - N \sum_{i=1}^{k} p_i D_i.$$

According to the Rao's theorem this should be asymptotically normally distributed. Considering the fact that $N \sum_{i=1}^{k} p_i D_i$ is a fixed term, we conclude that $\sum_{i=1}^{k} m_i D_i$ has asymptotic normal distribution. This proves the claim. □

Notice the difference between a direct and an indirect application of the Rao's theorem. If we were to use the Rao's theorem directly, then for example in (3.7), we had to always restrict to the case where binomial variables come from a multinomial distribution with

probabilities (P_{ij}) (i.e. $\theta = 0$) while Theorem 5.1 removes this restriction. So, we can write the following corollaries.

Corollary 5.1. The score test statistic for the cohort study is asymptotically normal under both null and alternative.

Proof. The score test statistic for the cohort study is given by

$$\frac{dL}{d\theta}|_{\theta=0} = \sum_{i=1}^{I} \left\{ \sum_{j=1}^{K_i} m_{ij} D_{ij} - M_i \sum_{j=1}^{K_i} p_{ij} D_{ij} \right\}.$$

Under the null, for a fixed $1 \leq i \leq I$, $\{m_{ij}\}_{j=1}^{K_i}$ are binomial variables that come from a multinomial distribution with M_i trials and probabilities $\{p_{ij}\}_{j=1}^{K_i}$. Under the alternative they are again binomial variables that come from a multinomial distribution with probabilities

$$\pi_{ij} = \frac{p_{ij}(1 + \theta D_{ij})}{\sum_{j=1}^{K_i} p_{ij}(1 + \theta D_{ij})}.$$

In any case according to Theorem 5.1, $\sum_{j=1}^{K_i} m_{ij} D_{ij}$ is asymptotically normal. Since multinomial distributions for $1 \leq i \leq I$ are independent and sum of independent normal variables is again normal, we conclude that the score statistic under the null and alternative is asymptotically normal. □

Corollary 5.2. The score test statistic for the case-control study is asymptotically normal under both null and alternative.

Proof. The score test statistic for the case-control study is given by

$$\frac{dl}{d\theta}|_{\theta} = \sum_{m=1}^{M} \left[-\frac{1}{K+1} \sum_{d=0}^{N_{D,m}} n_{d0m} D_{md} + \frac{K}{K+1} \sum_{d=0}^{N_{D,m}} n_{d1m} D_{md} \right].$$

Under the null for a fixed $1 \leq m \leq M$ each $(n_{d0m})_{d=0}^{N_{D,m}}$ and $(n_{d1m})_{d=0}^{N_{D,m}}$ are both binomial variables

coming from two independent multinomial distributions both with probabilities $(p_{d0m})_{d=0}^{N_{D,m}}$ and each with KS_m and S_m trials respectively. Under the alternative $(n_{d0m})_{d=0}^{N_{D,m}}$ is as before while $(n_{d1m})_{d=0}^{N_{D,m}}$ comes from a multinomial distribution with S_m trials and probabilities

$$\frac{p_{d0m}(1+\theta D_{d0m})}{\sum_{d=0}^{N_{D,m}} p_{d0m}(1+\theta D_{d0m})}.$$

In any case, according to Theorem 5.1 $\sum_{d=0}^{N_{D,m}} n_{d0m} D_{md}$ and $\sum_{d=0}^{N_{D,m}} n_{d1m} D_{md}$ are asymptotically normally distributed. Since these are independent for each m and the multinomial distributions are independent for all $1 \leq m \leq M$, we conclude that the score test statistic for the case-control study should be asymptotically normally distributed. □

5.3 Different approaches to calculate the power

Using the results from the previous section we can see that assuming a normal distribution for the score statistic under both null and alternative is reasonable in case the asymptotic assumption is met. If not, we can make no assumption on the distribution of neither of them. Therefore, we can propose the following three methods to calculate the power for both cohort and case-control studies:

1. Normal-sampling: As explained in subsection 3.1.2 and subsection 4.1.2, Little et al. (2010) uses this approach. It assumes a normal distribution for the score test under null and takes samples from the alternative to calculate the power. We would like to refer to this procedure as normal-sampling. This procedure relies on the Rao's theorem that states the normality of the score test under the null.

2. Normal-normal: As we proved in the previous section, in procedures for both cohort and case-control studies the score statistic under the alternative is as good asymptot-

ically normal as it is under the null. Therefore, if we believe that the score statistic under the null meets the asymptotic assumption, then it is very likely to meet the assumption under the alternative. This suggests that in such a case using a closed-form formula for the power is reasonable. In Chapters 3 and 4 we corrected the closed-form formulas for the power that were proposed in Little et al. (2010). Therefore, as long as the sample size is large enough we can be sure that those formulas provide reliable results. Here, we will refer to the closed-form formula method as the normal-normal procedure.

3. Sampling-sampling: What if the sample size is small? In such a case chances are that score test under none of null and alternative is normal. Then, we can first calculate the distribution of the score test under the null using a sampling procedure too! The algorithm for this procedure is explained below.

In Section 6.5, we use some simulated data and compare the results for the power of cohort and case-control studies based on all the methods explained above.

5.3.1 The sampling-sampling algorithm to calculate the power

All steps in the sampling-sampling algorithm are similar to those in algorithms described in Sections 3.1.2 and 4.1.2. However, we need to add two extra steps and adjust another step as described below:

The cohort study. We need to add two extra steps between steps 3 and 4, in the algorithm described in Section 3.1.2, to determine the distribution of score statistic under the null. Let's denote these steps by Step $\star$ and Step $\star\star$. The algorithm for these steps is as follows:

- Step $\star$. Define a vector named NL which is initially empty.
  - Step $\star$-1. Fix $i \in I$ and draw a sample from a multinomial distribution with M_i trials and J possible outcomes each with probability $(p_{ij})_{j=1}^{J}$.

- Step ⋆-2. Repeat Step ⋆-1 for all $i \in I$. Then using formula (3.15) calculate the value of the test statistic for this sample.
- Step ⋆-3. Add the value calculated in Step ⋆-2 to vector NL.

- Step ⋆⋆. Repeat Step ⋆, $100,000$ times. The vector NL now has $100,000$ components. Sort them increasingly. For a test with significance level α chose the $([(1-\alpha)\times 100,000]+1)$-th component of the sorted NL vector as the threshold and denote it by TS (here $[\cdot]$ stands for the integer part).

We also need to change Step 4-3:

- Step 4.3. Check whether the value from Step 4-2 is greater than TS. If so, mark it as a success.

The case-control study. Again, we need to add two extra steps between steps 3 and 4, in the algorithm described in subsection 4.1.2. Let's use the same notation as above to refer to these extra steps: Step ⋆ and Step ⋆⋆.

- Step ⋆. Define a vector named NL which is initially empty.
  - Step ⋆-1.
    - Step ⋆-1-1. Find one sample $(n_{d0m})_{d=0}^{N_D}$ by drawing from a multinomial distribution with KS_m trials and probabilities $(p_{d0m})_{d=0}^{N_D}$.
    - Step ⋆-1-2. Find one sample $(n_{d1m})_{d=0}^{N_D}$ by drawing from a multinomial distribution with S_m trials and probabilities $(p_{d0m})_{d=0}^{N_D}$.
  - Step ⋆-2. Repeat Step ⋆-1 for all $m \in M$. Then using formula (4.24) calculate the value of the test statistic for this sample.
  - Step ⋆-3. Add the value calculated in Step ⋆-2 to vector NL.

- Step ⋆⋆. Repeat Step ⋆, $100,000$ times. The vector NL now has $100,000$ components.

Sort them increasingly. For a test with significance level α chose the $([(1-\alpha)\times 100,000]+1)$-th component of the sorted NL vector as the threshold and denote it by TS.

Again we need to change Step 4-3:

- Step 4.3. Check whether the value from Step 4-2 is greater than TS. If so, mark it as a success.

5.4 A Bayesian approach to specifying a distribution for the power

To calculate the power, one has to pick a value for the alternative which is possibly as close to the true value as possible. One possible approach is to pick the value based on the studies already carried out. Since different studies might suggest different values (see Section 7.1), choosing a single parameter might be kind of subjective. The methods we have been discussing so far (following Little et al. (2010)) require the specification of a single estimation.

One possible approach to deal with this problem is to specify a range (rather than a single value) for the parameter and then implement a Bayesian model. Notice that in such a case we obtain a distribution for the power rather than a single estimation.

A method to find such a distribution is suggested in Lunn et al. (2012; Section 5.3): Based on a literature review find a range for the parameter of interest. Assume a normal distribution for the parameter and consider the given range as the 67% interval,i.e., mean ± 1 standard deviation. Using this information, calculate the mean and the standard deviation.

A big advantage of this approach is the inclusion of a range of possible values for θ rather than relying on a single value. Notice, however, that this approach still carries some sort of subjectivity as one might argue why we can assume a normal distribution and why the observed range is considered the 67% interval.

Remark 5.1. Some previous major studies suggest that ERR parameter falls somewhere

between .28 and .97 (see Section 7.1). We use the method explained above to find a normal distribution for the ERR parameter. The mean value of this interval is .625. Also, we can find the standard deviation based on the method proposed above.

$$.625 + \sigma = .97 \Rightarrow \sigma = .345.$$

Therefore, we can assign the following distribution to the ERR parameter

$$\theta \sim Normal(.625, .345^2).$$

Remark 5.2. (The issue of running time). An important advantage of the normal-normal approach that we proposed to calculate the power is that it remarkably reduces the running time of the Bayesian model. Notice that if we were to use a normal-sampling approach, then for any of the θ sample points, we have to draw a sample of size says $100,000$ subsequently, from the corresponding multinomial distribution! Using the normal-normal approach, all we need is the initial sample for θ as the power is calculated based on a closed-form formula. This issue will be addressed in Remark 6.1 too.

5.5 Studies whose measure of association is SIR or SMR

Standardized Mortality Ratio (SMR) and Standardized Incident Ratio (SIR) are commonly used to deal with the problem of confounding. In these methods, the dataset stratified by the confounder and the results are modified by comparison against a standard population.

There are some studies of ionizing radiation-cancer based on NDR that use this measure of association (see for example Ashmore et al. (1998)). However, it has been noted in Zielinski et al. (2008) that for occupational studies, SMR and SIR might be affected by another problem called the 'healthy worker effect'. As a result, the mortality and incidence rate

in the exposed group might seem to be lower. This might not reflect the truth. Therefore, measures of association such as risk ratio (for cohort studies) and odds ratio (for case-control studies) which use an internal non-exposed group seem to be more reliable.

Despite all the facts stated above, here we explain why the power procedure derived for the cohort study works for the SMR and SIR studies too. The main reason is that the power procedure for the cohort study already uses stratification. This helps in solving the equation for the number of cases in each stratum for SMR/SIR studies. Recall that from Equation 3.3, we could simply solve for M_{ij} the number of cases in category ij of dose. Using these values we found the probabilities of the multinomial distributions which provide the basic information for the power procedure. Notice that although this formula was calculated for the stratified dataset, one can simply check that when there are no strata the argument to derive the formula still works. For the SMR procedure, the general formula (assuming no initial strata) for the SMR (or SIR) corresponding to category j of the dose can be written as

$$\begin{aligned} SMR_j &= \frac{\#\ observed\ cases}{\#\ expected\ cases} \\ &= \frac{\frac{N_{1j}}{L_{1j}}p_1 + \frac{N_{2j}}{L_{2j}}p_2 + \cdots + \frac{N_{kj}}{L_{kj}}p_K}{\#\ cases\ in\ the\ standard\ population} \end{aligned} \tag{5.2}$$

where N_{ij} is number of cases in stratum i of the dose category j, L_{ij} is its population and p_i is population of stratum i in the standard population. To derive a power procedure, we need to obtain multinomial probabilities. For this purpose, we should find a value of N_{ij}. However, using a single equation such as 5.2 it's impossible as we have several unknown parameters.

But we are already breaking down the dataset into age+sex strata. Therefore, at each stratum we only have a single age+sex stratum. The formula for SMR at strata i for dose

group j is then

$$\begin{aligned} SMR_{ij} &= \frac{\frac{N_{ij}}{L_{ij}} p_i}{N_i} \\ &= \frac{N_{ij}/L_{ij}}{N_i/p_i} \\ &= \frac{probability\ of\ being\ a\ case\ in\ the\ dose\ category\ ij\ (Dose = D_{ij})}{probability\ of\ being\ a\ case\ in\ stratum\ i\ of\ standard\ population} \end{aligned} \tag{5.3}$$

This formula is similar to that given in (3.1). The only difference is that *probability of being a case in the non exposed category* ($Dose = 0$) is replaced by *probability of being a case in stratum i of standard population*. This doesn't affect the formula derived for the multinomial probabilities and the final procedure. So, we will get the same procedure as that for the cohort study.

5.6 Derivation of an interval estimation for ERR parameter

In this section, I prove a theorem that allows derivation of an interval estimation for the Relative risk of a general disease from the relative risks of some sub-diseases. This result enables us to give an interval estimation for the ERR parameter of the general cancer from the given ERR parameters of various cancer types. I will employ this fact in the next chapter.

Theorem 5.2. Let RR_i for $1 \leq i \leq n$ be the mortality relative risk of some disease i. Let RR denote the relative risk to die from any of the given diseases. Then we have,

$$\min\{RR_i, \quad i = 1, \cdots, n\} \leq RR \leq \max\{RR_i, \quad i = 1, \cdots, n\}. \tag{5.4}$$

Proof. Let S^e and S^u denote the population of exposed and non-exposed groups. Let N_i^e and N_i^u denote number of deaths of disease i in the exposed and non-exposed groups respectively. Similarly, assume that N^e and N^u represent the total number of deaths (of any kind) in those

groups. Then the relative risk for disease i is given by

$$\begin{aligned} RR_i &= \frac{N_i^e/S^e}{N_i^u/S^u} \\ &= \frac{S^u}{S^e} \cdot \frac{N_i^e}{N_i^u}. \end{aligned} \tag{5.5}$$

On the other hand, we can write

$$\begin{aligned} RR &= \frac{N^e/S^e}{N^u/S^u} \\ &= \frac{S^u}{S^e} \cdot \frac{N^e}{N^u} \\ &= \frac{S^u}{S^e} \cdot \frac{N_1^e + \cdots + N_n^e}{N_1^u + \cdots + N_n^u}. \end{aligned} \tag{5.6}$$

But we have

$$\min\{\frac{N_i^e}{N_i^u}, \quad i = 1, \cdots, n\} \leq \frac{N_1^e + \cdots + N_n^e}{N_1^u + \cdots + N_n^u} \leq \max\{\frac{N_i^e}{N_i^u}, \quad i = 1, \cdots, n\} \tag{5.7}$$

Putting (5.5), (5.6) and (5.7) together, we derive (5.4). □

Corollary 5.3. Let θ denote the ERR parameter for general cancer and assume that θ_i for $1 \leq i \leq n$ denotes the ERR parameter for various cancer types. Then we have

$$\min\{\theta_i, \quad i = 1, \cdots, n\} \leq \theta \leq \max\{\theta_i, \quad i = 1, \cdots, n\}.$$

Proof. We can use Theorem 5.2 to write

$$\min\{RR_i, \quad i = 1, \cdots, n\} \leq RR \leq \max\{RR_i, \quad i = 1, \cdots, n\}.$$

This together with the linear-dose effect implies that for any arbitrary mean dose D we have

$$\min\{1+\theta_i D, \quad i=1,\cdots,n\} \leq 1+\theta D \leq \max\{1+\theta_i D, \quad i=1,\cdots,n\}.$$

This is equivalent to

$$1+\min\{\theta_i, \quad i=1,\cdots,n\}D \leq 1+\theta D \leq 1+\max\{\theta_i, \quad i=1,\cdots,n\}D.$$

which implies

$$\min\{\theta_i, \quad i=1,\cdots,n\} \leq \theta \leq \max\{\theta_i, \quad i=1,\cdots,n\}.$$

□

Remark 5.3. The reason why Theorem 5.2 is stated only for mortality and not for incidence is that for the latter we might be not able to write $N^e = N_1^e + \cdots + N_n^e$ (similarly for N^u). The issue of multiple diseases (such as multiple cancers) can be well resolved for mortality but not for incidence. Multiple incidence of cancers, for instance, occurs when an individual suffers from multiple types of cancer during his/her life (either concurrently or not). When counting cases for different types of cancers, one should count it both towards all the relevant N_i^e,s (or N_i^u,s). Now the problem is whether this individual should be counted as one case of general cancer or two when calculating N^e (or N^u).

When it comes to mortality, we don't have to deal with this problem as death marks a unique event! If an individual with multiple cancers dies, the death is linked to latter cancer. Therefore, categories of death from different types of cancer remain fairly disjoint even in the presence of multiple cancers. One very unlikely possibility is dying simultaneously of several cancers. In such a case, it is more reasonable to define a new category for the cause of death including those cancer types collectively. So, categories for the cause of death remain disjoint.

For incidence, the issue is not so clear and the problem of counting multiple incidences (at least for the ionizing radiation studies) has not been referred to in the literature. Therefore, I prefer to keep Theorem 5.2 only for mortality. However, depending on the method used to count cases, it might work for incidence as well. Furthermore, if multiple diseases are infrequent, the said theorem still works "approximately" for incidence.

Chapter 6

Simulation studies

In this chapter, we use some simulated data to visually prove the claims made in the previous chapters and also to compare the different methods proposed. These includes:

1. The issue of Monte-Carlo error presented in Section 5.1.

2. Normality of score test statistic under null and alternative for cohort and case-control studies

3. The proper formula to be used for the normalized score test statistic under the null.

4. Comparison of the closed-form formulas (normal-normal) proposed in Little et al. (2010) and the one proposed in Chapters 3 and 4. We also compare both of them against two other methods: Normal-sampling and Sampling-sampling as proposed in Chapter 5.

5. Comparison of sample sizes for both cohort and case-control studies based on the formulas proposed in Little et al. (2010) and those proposed in Chapters 3 and 4.

6. A Bayesian approach to provide a distribution for the power as described in Section 5.4

6.1 simulated datasets

We will use the following datasets in our simulation studies. Each dataset is presented using four pieces of information. these includes:

1. M_i a vector the length of which equals to the number of age+sex strata in the dataset and each component of which represents the number of cases in each strata.

2. A matrix D_i the dimension of which is $n \times m$ where n is number of dose categories and m is number of age+sex strata. $(D_i)_{jk}$ is the average of the dose in the j-th category of dose in stratum k. In each column they are arranged increasingly.

3. A matrix P_i with the same dimension as D_i. $(p_i)_{jk}$ is the proportion of the population of dose category j that contributes to stratum k.

4. A number K_i which is the number of controls assigned to a case in the case-control study.

We define 6 datasets as follows:

- Dataset 1. $M_1 = (2, 5, 3, 2)$, $K_1 = 4$ and

$$D_1 = \begin{bmatrix} 1 & 1 & 2 & 1.5 \\ 1.8 & 4.3 & 3 & 2 \\ 2.9 & 7 & 4.7 & 4 \end{bmatrix} \quad , \quad P_1 = \begin{bmatrix} .5 & .6 & .5 & .4 \\ .3 & .3 & .27 & .3 \\ .2 & .1 & .23 & .3 \end{bmatrix}$$

- Dataset 2. $M_2 = (5, 7)$, $K_2 = 4$ and

$$D_2 = \begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix} \quad , \quad P_2 = \begin{bmatrix} .6 & .2 \\ .4 & .8 \end{bmatrix}$$

- Dataset 3. $M_3 = (17, 15)$, $K_3 = 4$ and

$$D_3 = \begin{bmatrix} 1 & 2 \\ 3 & 7 \end{bmatrix} \quad , \quad P_3 = \begin{bmatrix} .6 & .2 \\ .4 & .8 \end{bmatrix}$$

- Dataset 4. $M_4 = (4, 18, 3, 11)$, $K_4 = 4$ and

$$D_4 = \begin{bmatrix} 1 & 1 & 2 & 1.5 \\ 2.8 & 3.3 & 3 & 2 \\ 7.9 & 4 & 4.7 & 4 \end{bmatrix} \quad , \quad P_4 = \begin{bmatrix} .8 & .7 & .2 & .8 \\ .1 & .2 & .5 & .1 \\ .1 & .1 & .3 & .1 \end{bmatrix}$$

- Dataset 5. $M_5 = (4, 18, 3, 11)$, $K_5 = 4$ and

$$D_5 = \begin{bmatrix} 1 & 1 & 2 & 1.5 \\ 2.8 & 3.3 & 3 & 2 \\ 7.9 & 4 & 4.7 & 4 \end{bmatrix} \quad , \quad P_5 = \begin{bmatrix} .9 & .05 & .9 & .1 \\ 0 & .9 & .09 & 0 \\ .1 & .05 & .01 & .9 \end{bmatrix}$$

- Dataset 6. $M_6 = (12)$, $K_6 = 4$ and

$$D_6 = \begin{bmatrix} 1.375 \\ 2.775 \\ 4.650 \end{bmatrix} \quad , \quad P_6 = \begin{bmatrix} .2075 \\ .2925 \\ .5 \end{bmatrix}$$

6.2 Monte-Carlo Error

As explained in Section 5.1, reference (Little et al. 2010) reports a fixed number of iterations to obtain a desired MCE regardless of study design, effect size (θ) and years of follow-up. However, as shown there all this factors should be taken into account. Bellow,

we use our simulated datasets to obtain the required number of iterations for $MCE = .0032$ and $\theta = 2$. Notice that if we use the formula given in Little et al. (2010), then we require $100,000$ iterations for both study designs and all datasets.

Table 6.1: Number of iterations for $MCE = .0032$ and $\theta = 2$

Dataset #	Iterations for cohort	Iterations for case-control
1	109,906	107,456
2	51,181	60,476
3	54,797	63,369
4	160,434	147,879
5	173,475	158,312

6.3 Asymptotic normality of the score test statistic under null and alternative

We proved asymptotic normality of the score test under the null and alternative for both cohort and case-control studies in Section 5.2. Here, we use datasets defined in Section 6.1 to visually see this fact. It should be pointed out that datasets defined in the aforementioned section have been defined in pair: We manipulate one dataset in some specific way to see the impact of manipulation on the asymptotic normality of the score test statistic. We check asymptotic normality using the p-value validity test of the hypothesis test. This test relies on the fact that $F(X) \sim U$ where X is a random variable, F is its CDF and U is the uniform distribution. Notice that the flatter the p-value histogram, the better normally distributed the score test statistic. Below we evaluate each dataset by drawing four histograms: Score test statistic for cohort and case-control study under both null and alternative. All the his-

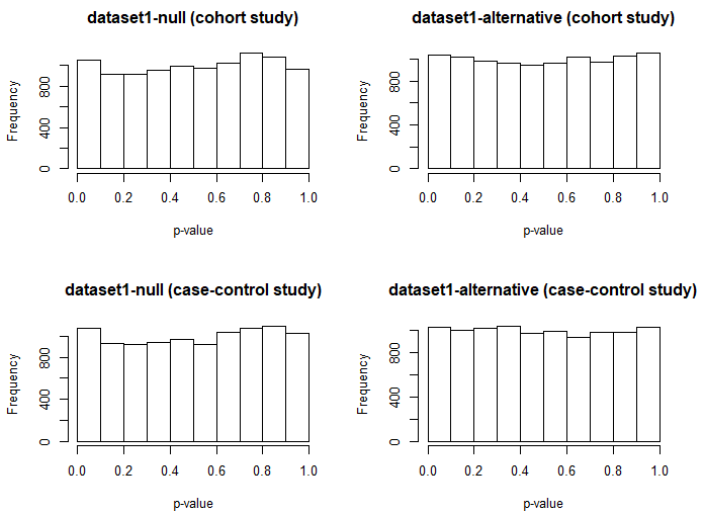

Figure 6.1: Histograms of asymptotic normality for Dataset 1

tograms are drawn based on a simulation of size $100,000$. We set $\alpha = .05$ and $\theta = 2$ (for alternative).

1. Dataset 1: Histograms for this dataset are given in Figure 6.1.

2. Dataset 6: Dataset 6 is obtained from Dataset 1 by collapsing 4 strata into 1. The result for cohort in 6.2 suggests that more strata can cause the score statistic to be better normally distributed.

3. Dataset 2: Histograms for this dataset are given in Figure 6.3.

4. Dataset 3: Dataset 3 is similar to Dataset 2. The only difference is that we increase the number of observations from 12 to 32 which results in flatter histograms (Figure 6.4).

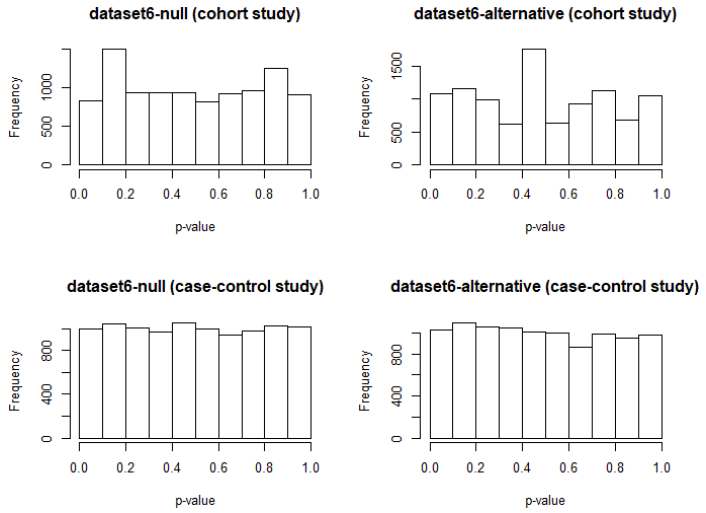

Figure 6.2: Histograms of asymptotic normality for Dataset 6

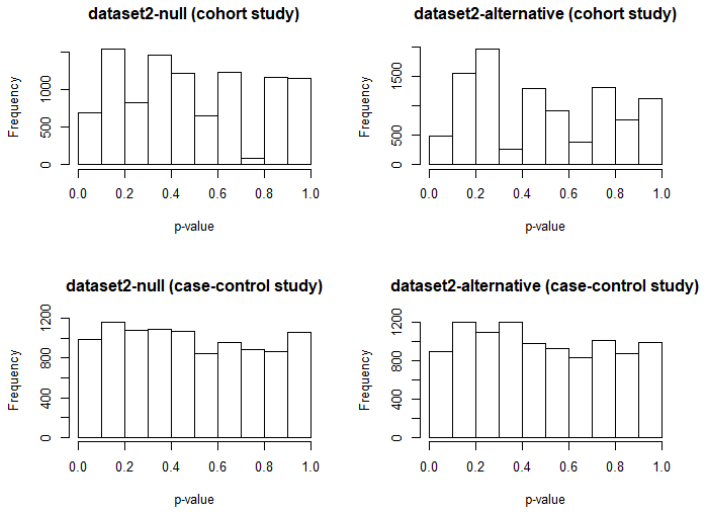

Figure 6.3: Histograms of asymptotic normality for Dataset 2

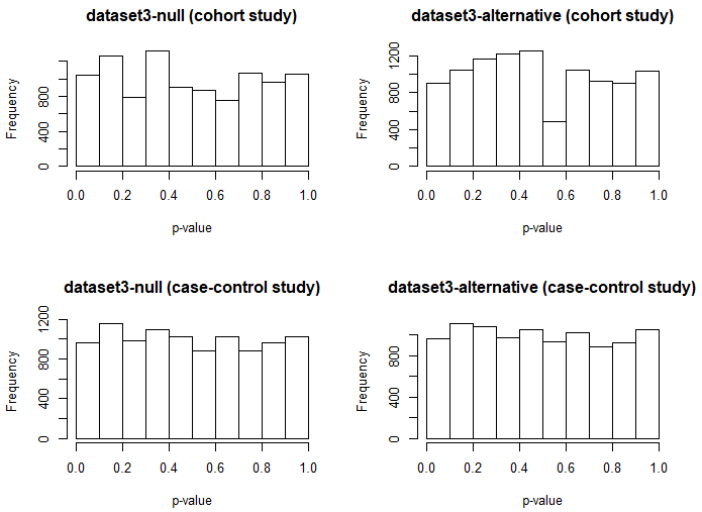

Figure 6.4: Histograms of asymptotic normality for Dataset 3

5. Dataset 4: Histograms for this dataset are given in Figure 6.5.

6. Dataset 5: Dataset 5 is obtained from Dataset 4 by making the population distribution among dose categories very heterogeneous. It seems to worsen the normality (Figure 6.6).

6.4 Formula for the score test statistic for the normal-sampling approach

As pointed out in Remark 3.3 and Remark 4.1, the formula proposed for the score test statistic for the normal-sampling approach in Little et al. (2010) needs modification as it includes the variance under alternative rather than the variance under null. Here, we specifically use Dataset 3 and check the normality of the score test statistic for the cohort study based on both approaches. As Figure 6.7 and Figure 6.8 depict, the one with null variance is much flatter!

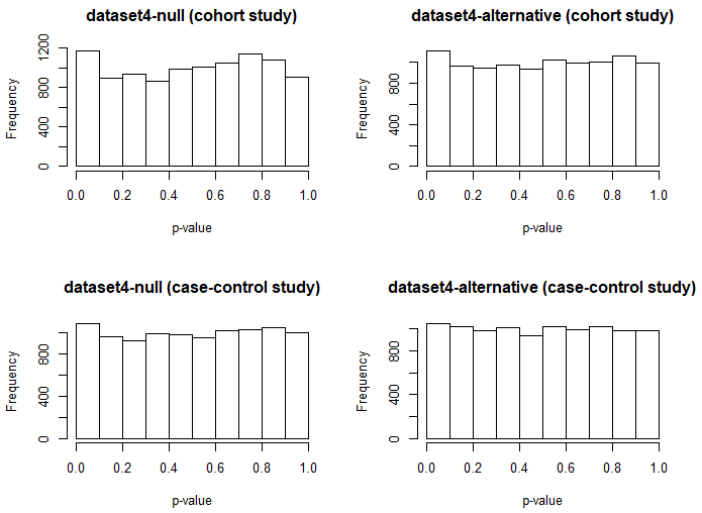

Figure 6.5: Histograms of asymptotic normality for Dataset 4

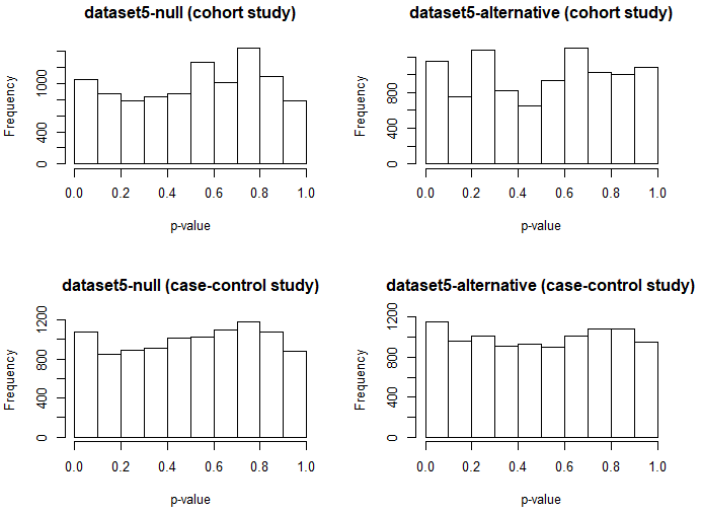

Figure 6.6: Histograms of asymptotic normality for Dataset 5

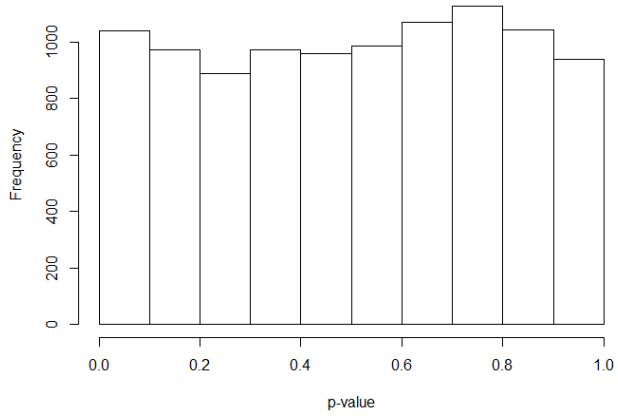

Figure 6.7: Histograms of the score test statistic with null variance (Dataset 3-cohort)

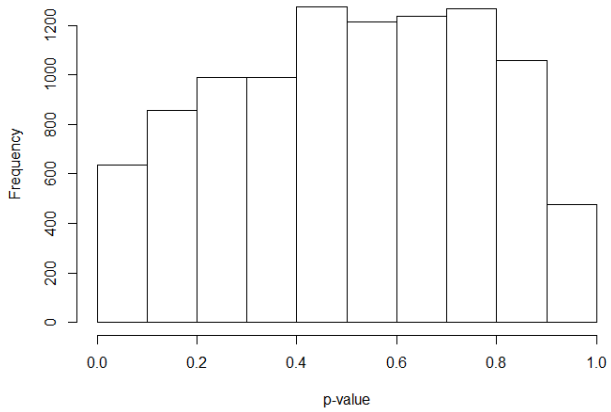

Figure 6.8: Histograms of the score test statistic with alternative variance (Dataset 3-cohort)

6.5 Different approaches to calculate the power

In Section 5.3 we proposed three different approaches to calculate the power for both cohort and case-control studies. this includes Normal-Sampling (N-S), Normal-Normal (N-N) and Sampling-Sampling (S-S). Also for the N-N approach, we explained in Remark 3.2 and Remark 4.2 that the formulas are given in Little et al. (2010) (here referred to as old formula) should be revisited. We provided substitute formulas (here referred to as new formula). In Table 6.2 below we use all these methods and apply them to Datasets 1 to 5

defined in Section 6.1 to derive the power for both cohort and case-control studies. From this table, it is clear that: 1. The values for N-N (new formula) are always fairly close to both values from N-S and S-S method. 2. The value given by N-N (old formula) is not close to those of N-S and S-S and in many cases completely irrelevant!

6.6 Sample size

As explained in Remark 3.4 and Remark 4.3 the formulas proposed in Little et al. (2010) (here referred to as old formula) for the sample size needs to be changed for both cohort and case-control studies. In Table 6.3 we use two datasets to compare the sample size for powers .8 and .9 based on the formulas proposed in Little et al. (2010) and also based on the formulas we obtained in Chapters 3 and 4 (referred to as new formula).

Recall that for the sample size the dataset shouldn't be stratified. So we improvise two datasets:

1. Dataset 6: 2 Dose categories with average group dose D=(1,2)and distribution of population among the dose groups p=(.4,.6). Also, let $\alpha = .05$ and $\theta = 2$. For the case control study, assume 5 controls per case.

2. Dataset 7: : 4 Dose categories with average group dose D=(.5,1.4,2.5,4) and distribution of population among the dose groups p=(.5,.3,.15,.05). Also, let $\alpha = .1$ and $\theta = 1$. For the case control study, assume 4 controls per case.

Table 6.2: Comparison of power for Datasets 1 to 5 using four methods

Dataset #	Method	Results for cohort	Results for case-control
1	N-N (old formula)	0.788	0.716
1	N-S	0.817	0.748
1	S-S	0.809	0.746
1	N-N (new formula)	0.819	0.746
2	N-N (old formula)	0.374	0.294
2	N-S	0.159	0.168
2	S-S	0.155	0.197
2	N-N (new formula)	0.195	0.176
3	N-N (old formula)	0.683	0.555
3	N-S	0.561	0.427
3	S-S	0.559	0.464
3	N-N (new formula)	0.501	0.418
4	N-N (old formula)	0.633	0.579
4	N-S	0.738	0.681
4	S-S	0.720	0.664
4	N-N (new formula)	0.751	0.687
5	N-N (old formula)	0.976	0.874
5	N-S	0.822	0.675
5	S-S	0.853	0.708
5	N-N (new formula)	0.816	0.671

Table 6.3: Comparison of sample size based on old and new formulas

Dataset #	Method	desired power	Size (cohort)	size (case-control)
6	old formula	.8	97	126
6	new formula	.8	108	137
6	old formula	.9	134	174
6	new formula	.9	147	187
7	old formula	.8	29	57
7	new formula	.8	35	48
7	old formula	.9	44	61
7	new formula	.9	51	53

6.7 A Bayesian approach to specifying a distribution for the power

In Section 5.4 we explained a Bayesian method to find the power to include a range of possible values for the ERR parameter rather than relying upon a single value. We illustrate the proposed approach by applying it to dataset 4 to derive a distribution for the power of the cohort study. We assume that the 67% interval for θ is $[1, 3]$.

Using Winbugs, we can see that a sample of size 5000 converges to the distribution specified for θ which is Normal(2,1). The history plot shows that convergence is attained with 5000 samples (Figure 6.9).

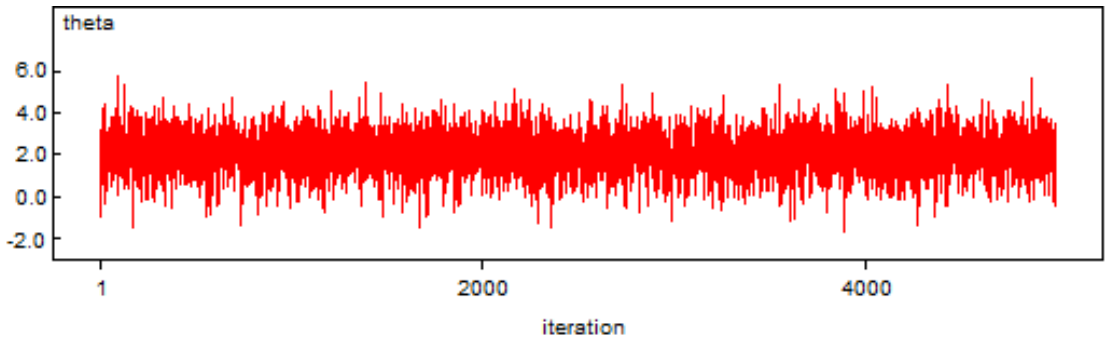

Figure 6.9: History plot for **5000** samples from θ

Figure 6.10 shows the density plot of the drawn sample.

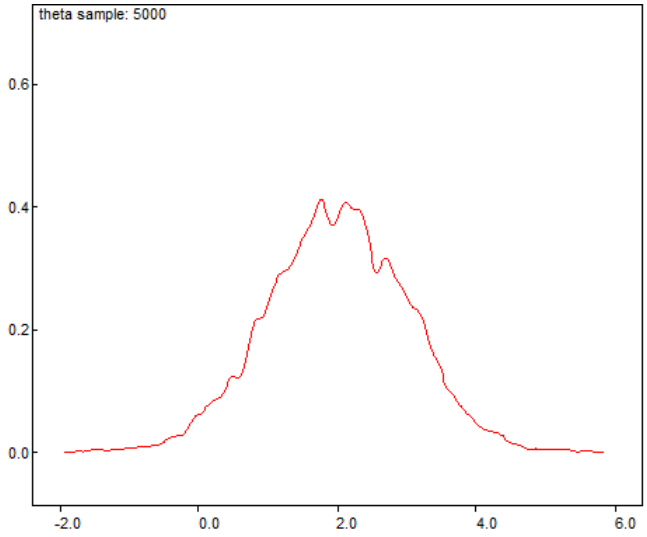

Figure 6.10: density plot for 5000 samples from θ

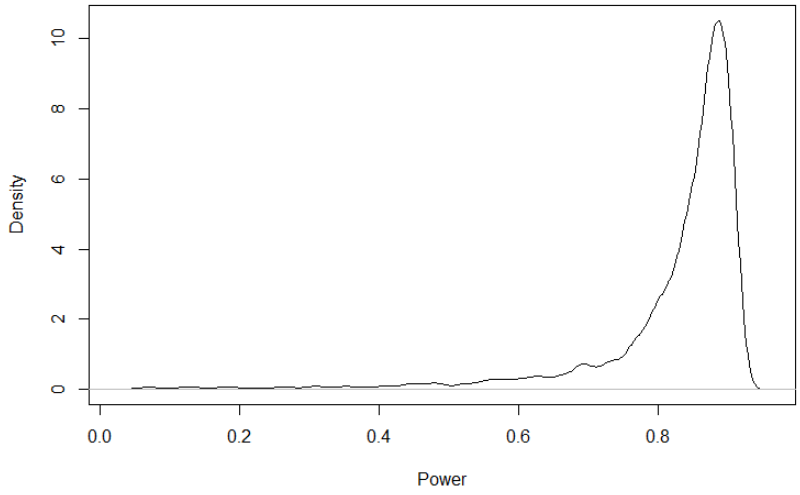

Figure 6.11: Distribution of Power for the cohort study, dataset 4

Now, for each sample point, we can apply one of the three methods explained in Section 5.3 to calculate the corresponding power. Finally, we can make inference based on the power sample we have found. The power distribution of the cohort study for dataset 4 is shown in Figure 6.11. Based on the drawn sample, we can also find that there is 77% chance for this study to have a power of .8 or higher.

Remark 6.1. In Remark 5.2, we addressed the issue of running time. Notice that for the example presented above, if we use the normal-sampling approach, then we need to sample $500,000,000$ times while using the normal-normal approach all we need is the initial sample for θ of size 5000. We found the above power distribution using both methods. The running time is dramatically different. Using the normal-normal approach the running time is only .9 seconds while using the normal-sampling approach it is 1929 seconds!

Chapter 7

Conclusion

In this thesis we considered procedures developed in Little et al. (2010) to calculate the power of studies related to low-level doses of ionizing radiation and cancer health outcomes. We explained the procedures with mathematical and statistical details and proposed several substitutes to methods proposed in the mentioned reference to improve its performance.

The improvement includes: Changing the formula for the score statistic for both cohort and case-control studies, proposing a more effective closed-form formula for the power, changing the formula for the sample size and redefining the formula for the Monte-Carlo Error. We also present a third method to calculate the power which is a sampling-sampling approach. We laid out the algorithm to implement all the methods. This could help researchers who are interested in carrying out the same kind of study in the future. We also proposed a Bayesian approach to find distribution for the power. We address the issue of running time for the Bayesian model to explain why a closed-form formula for the power is indispensable. Therefore, the closed-form formula we proposed before becomes more prominent.

We also dealt with some secondary issues including Calculation of an interval estimation for the ERR parameter and studies with SMR/SIR as the measure of association.

The R codes to implement all such methods are presented in the appendix. We illustrated our work using simulation which is presented in Chapter 6.

Future work. Notice that one goal of this research is to calculate the power of ionizing radiation-cancer studies based on the National Dose Registry of Canada. Since at this point

the data is not ready, this will be done in the future. However, the required information to analyze the data are presented in the following section.

7.1 Future studies

In this section, we provide the information required to analyze the National Dose Registry of Canada (and any similar data set). The information provided in this section together with the procedures we discussed in the previous chapters can be used to obtain the power of cohort and case-control studies for all the possible years of follow-up. Based on the analysis, it will be possible to suggest the best study design for ionizing radiation-cancer studies. One can also evaluate the power of the previous studies and suggest the minimum year of follow-up to attain a satisfactory power (widely considered to be 0.8).

7.1.1 Dose categories

The first step in analyzing the data is to break down the dose variable into a few dose categories. We will use the following dose categories which are widely used in the literature. In particular, Beebe et al. (1998) which is a well-known study comprising data from nuclear power workers from 15 countries uses these categories to evaluate the ERR parameter. Other ionizing radiation studies such as CVD studies useless and more the same categories (see for example Zielinski et al. (2009)). The categories are as follow:

$$0,\ 0- < 5,\ 5- < 10,\ 10- < 20,\ 20- < 50,\ 50- < 100,\ 100- < 150,\ 150- < 200,\ 200- < 300,$$

$$300- < 400,\ 400- < 500,\ > 500.$$

7.1.2 The ERR parameter

Beebe et al. (1998) is the most well-known study to evaluate The Excess Relative Risk parameter for ionizing radiation-cancer. This study combines information on nuclear power workers in 15 countries across the world. It includes information of over $600,000$ individuals. The ERR parameter as reported by this study is .97 Sv^{-1} for all cancers excluding leukemia and 1.93 Sv^{-1} for leukemia excluding chronic lymphocytic leukemia. Another study (Krestinina et al. 2005) that follows the residence of a highly radiation-exposed area in Russia reports the parameter for all cancers to be .92 Sv^{-1}. A smaller ERR parameter is proposed in Muirhead et al. (2009): .28 Sv^{-1} . Notice also that BEIR V (BEIR V 1990) provides the parameter for all types of cancer independently. However, cancer as a whole has not been referred to there. Table 7.1 summarizes the results of this study on the ERR parameter of various cancers. The ERR parameter for most major cancers is around 1 Sv^{-1}. Using this table and Corollary 5.3, we can see that the ERR parameter for general cancer should fall between 0 Sv^{-1} and 8.3 Sv^{-1}. Although this range is wide, it matches the estimations provided in Beebe et al. (1998) and Krestinina et al. (2005). Notice that we can argue to shrink this range as most ERR parameters are near 1 Sv^{-1} except those for leukemia, salivary, brain and thyroid cancers. If we remove these four cancer types, we will find out that the ERR parameter for all cancers except those four types falls between 0 Sv^{-1} and 2.3 Sv^{-1}. This result seems to provide a fairly narrow interval that encompasses the estimations given in Beebe et al. (1998) and Krestinina et al. (2005). It should be pointed out that the proportion of leukemia, salivary, brain and thyroid cancers in all types of cancer is 2.6%, .3% 1.7% and 3.3%, respectively (7.9% altogether). Therefore, the ERR parameter for general cancer should be closer to that of all the other cancer types than to the ERR parameter of the four types mentioned.

Based on all the results provided above, we judge that $\theta = 1\ Sv^{-1}$ is a fair choice for the

ERR parameter. Notice that in our study dose amounts are given in milli-Sievert. Therefore, we will use $\theta = .001\ mSv^{-1}$.In a Bayesian approach, as explained in Section 5.4, we assume a normal distribution for the ERR parameter with $[.28, .97]$ as the 67% interval (± 1 standard deviation).

Table 7.1: BEIR V results on ERR parameter (per Sv) for various cancer types

Cancer type	ERR parameter	Cancer type	ERR parameter
Salivary	6.5	Breast	.6
Stomach	1.19	Brain	3.44
Skin	1.5	Esophageal	1.58
Colon	1.85	Lung	1.33
Thyroid	8.3	Small intestine	0
Liver	1.26	Bone	0
Testis	0	Ovary	2.33
Uterus	0	Prostate	1.05
Urinary tract and sinus	2.3	Nasal cavity	0
Pancreas	0	Pharynx, hypopharynx and larynx	0
Leukemia	4.24-5.21		

7.1.3 What is the practical implications of this power study?

Using this power study, after analyzing the data, I will be able to provide three important pieces of information for the epidemiologists working on the NDR:

- I can determine which study design is more suitale for the ionizing radiatoin-cancer study based on the NDR, the cohort or the case-control study?
- There are already several papers in the litterature based on the NDR that try to detect an ionizing radiation-cancer relationship. Many of such studies fail to detect a significant result. I will be able to answer whether the insignificant result is due to low power or not? If I realize that the power of such studies has been low, then the

non-significant result should not be interpreted as a true lack of association between ionizing radiation and cancer.

- Finally, I can propose the minimum number of the years of follow-up to attain a satisfactory power (say 80%) based on the NDR. Therefore, for future epidemiological studies, this important factor will be taken into account.

References

Ahrens W., Pigeot I. (2015). Handbook of epidemiology. Springer. 1

Ashmore J. P.,Krewski D., Zielinski J. M.,Jiang H.,Semenciw R., and Band P. R. (1998). First Analysis of Mortality and Occupational Radiation Exposure based on the National Dose Registry of Canada. Radiation Research, 148 (6): 564–574. 3, 20, 73

Beebe G. W., Ishida M. and Jablon S. (1962). Studies of the Mortality of A-Bomb Survivors: I. Plan of Study and Mortality in theMedical Subsample (Selection I), 1950-1958. Radiation Research, 16 (3): 253–280. 16

Cardis E., Vrijheid M., Blettner M., Gilbert E., Hakama M., Hill C., Howe G., Kalor J., Muirhead C. R., Scubauer-Berigan M., Yoshimura T., Bermann F., Cowper G., Fix J., Hacker C., Heinmiller B., Marshall M., Thierry-Chef I., Utterback D., Ahn Y. O., Amoros E., Ashmore P., Auvinen A., Bae J. M., Bernar Soloano J., Biau A., Combalot E., Deboodt P., Diea Sacristan A., Eklof M., Engels H., Enghold G., Gulis G., Habib R., Holan K., Hyvonen H., Kerekes A., Kurtinaitis J., Malker H., Martuzzi M., Mastauskas A., Monnet A., Moser M., Pearce M. S., Richardson D. B., Rodrigues-Artalejo F., Rogel A., Tardy H., Telle-Lamberton M., Turai I., Usel M., Veress K. (2005). Risk of cancer after low doses of ionizing radiation: retrospective cohort study in 15 countries. British Medical Journal, 331: 77–-82. 3, 94, 95

Casella G., Berger R. L. (1990). Statistical inference. Duxbury press, Belmont. 5, 7

Dobrzynski1 L., Fornalski K. W. and Feinendegen L. E. (2015). Cancer Mortality Among People Living in Areas With Various Levels of Natural Background Radiation. Dose-response, 1–10. 19

Fazel R., Krumholz H. M., Wang Y., Ross J. S., Chen J., Ting H. H., Shah, N. D., Nasir K., Einstein A. J., and Nallamothu B. K. (2009). Exposure to Low-Dose Ionizing Radiation from Medical Imaging Procedures. N Engl J Med, (361): 849–857. 14

Friedman L. M., Furberg C. D., DeMets D. L. (2010). Fundamentals of Clinical Trials. Springer. 10, 11

S. Greenland, D. C. Thomas (1982). Cancer Mortality Among People Living in Areas With Various Levels of Natural Background Radiation. American Journal of Epidemiology, 116(3): 547-53. 5

Gordis L. (2014). Epidemiology, Fifth edition. , Elsevier Sunders. 1, 2, 3, 4

Gribbin M. A., Weeks J. L., Howe G. R., (1993). Cancer mortality (1956–1985) among male employees of Atomic Energy of Canada Limited with respect to occupational exposure to external low- linear-energy-transfer ionizing radiation. Radiat Res, 133 (3): 375–-380. 3, 21

Hendry J., Simon S. L. and Sohrabi M. (2009). Human exposure to high natural background radiation: What can it teach us about radiation risks?. Journal of Radiological Protection, doi: 10.1088/0952-4746/29/2A/S03. 19

Ellis P. D. (2010). The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results, First edition. , Cambridge University Press. 2, 8

Koehler E., Brown E., HaneuseS. J.-P. A. (2009). On the Assessment of Monte-Carlo Error in Simulation-Based Statistical Analyses. Am Stat., 63(2): 155–-162. 64

Krestinina L.Y., Preston D.L., Ostroumova E. V., Degteva M. O., Ron E., Vyushkova O. V., Startsev N. V., Kossenko M. M., Akleyev A. V. (2005). Protracted radiation exposure and cancer mortality in the Techa River cohort. Radiat Res., 164(5):602–-611. 95

S. Landau, D. StahlDaniel Stahi. (2012). Sample size and power calculations for medical studies by simulation when closed form expressions are not available. Statistical Methods in Medical Research, 22(3): 324–-345. 11

Little M.P.,Wakeford R.,Lubinc J.H. and Kendall G.M. (2010). The statistical power of epidemiological studies analyzing the relationship between exposure to ionizing radiation and cancer, with special reference to childhood leukemia and natural background radiation. Radiat Res. 174 (3): 387-–402. 13, 19, 24, 26, 28, 29, 31, 35, 38, 39, 40, 41, 45, 57, 58, 63, 64, 65, 66, 69, 70, 72, 79, 81, 82, 85, 87, 88, 93

Lubin J. H., Gail M. H. and Ershow A. G. (1990). Sample size and power for case-control studies when exposures are continuous. Statistics in Medicine, 131 (3): 552–566. 22

Lubin J. H., Gail M. H. (1988). On power and sample size for studying features of the relative odds of disease. American Journal of Epidemiology, 7: 363–376. 23

Lunn D., Jackson C., Best N., Thomas A., Spiegelhalter D. (2012). The BUGS Book: A Practical Introduction to Bayesian Analysis. CRC press. 72

Mckeown-eyssen G. E. and Thomas D. C. (1985). Sample size determination in case-control studies: the influence of the distribution of exposure. J. Chron Dis, 38 (7): 59–568. 22

Muirhead C. R., O'Hagan J. A., Haylock R. G. E., Phillipson M. A., Willcock T., Berridge G. L. C., Zhang W. (2009) Mortality and cancer incidence following occupational radiation exposure: third analysis of the National Registry for Radiation Workers. British Journal of Cancer, 100: 206–-212. 95

Ozasa K., Cullings H. M. , Ohishi W., Hida A., Grant E. J. (2019). Epidemiological studies of atomic bomb radiation at the Radiation Effects Research Foundation. International Journal of Radiation Biology, 95 (7), 879–-891. 17

D. B. Richardson, S. Wing. (1999) Greater sensitivity to ionizing radiation at older age: follow-up of workers at Oak Ridge National Laboratory through 1990. International Journal of Epidemiology, 28 (3): 428–-436. 3

Schlesselman J. J. (1974). Sample size requirements in cohort and case-control studies of disease. Journal of Epidemiology, 99 (6): 381–384. 21

Shimizu Y., kato H., and Schullt W. J. (1990). Studies of the Mortality of A-Bomb Survivors. Radiation Research, 121: 120–141. 17

Sun Z. J. , Inskip P. D., Wang J., Kwon D., Zhao Y., Zhang L., Wang Q., Fan S. (2016). Solid cancer incidence among Chinese medical diagnostic x-ray workers, 1950–1995: Estimation of radiation-related risks. International Journal of Cancer, 138 (12): 2875–288. 3

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov). SEER*Stat Database: Total U.S., 1969–2006 Counties. National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch; 2009. released April 2009, based on the November 2008 submission. 19

United Kingdom Childhood Cancer Study Investigators. The United Kingdom Childhood Cancer Study of exposure to domestic sources of ionizing radiation: 1: radon gas. Br J Cancer, 86: 1721–-1726, 2002. 19

United Kingdom Childhood Cancer Study Investigators. The United Kingdom Childhood Cancer Study of exposure to domestic sources of ionizing radiation: 2: gamma radiation. Br J Cancer., 86: 1727–-1731, 2002. 19

United Nations. Scientific Committee on the Effects of Atomic Radiation. Sources and Effects of Ionizing Radiation: Sources. 2000. 18

US National Research Council. Committee on the Biological Effects of Ionizing Radiations. Health effects of exposure to low levels of ionizing radiation. BEIR V. Washington, DC, USA: National Academy Press, 1–421. 30, 31, 46, 95

Norton B., Strube M. (2001). Understanding Statistical Power. Journal of Orthopaedic Sports Physical Therapy, 31 (6): 307–315. 6

Rao C. R. (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation Mathematical Proceedings of the Cambridge Philosophical Society, 44 (1): 50–57. 12, 26, 28

Wakeford R., Kendall G.M., Little M. P. (2009). The proportion of childhood leukemia incidence in Great Britain that may be caused by natural background ionizing radiation. Leukemia, 23: 770–-776. 19

Zablotska L. B., Lane R. S. D., Thompson and P. A. (2014). A reanalysis of cancer mortality in Canadian nuclear workers (1956–1994) based on revised exposure and cohort data. British Journal of Cancer, 110: 214-–223. 21

Zielinski J. M., Shilnikova N. S. and Krewski D. (2008). Canadian national dose registry of radiation workers: overview of research from 1951 through 2007. International Journal of Occupational Medicine and Environmental Health, 21 (4): 269–-275. 73

Zielinski J. M., Shilnikova N. S. and Krewski D. (2009). Low dose ionizing radiation exposure and cardiovascular disease mortality: cohort study based on Canadian National Dose Registry of radiation workers. International Journal of Occupational Medicine and Environmental Health, 22 (1): 27–-33. 94

Canadian Nuclear Safety Commission, http://nuclearsafety.gc.ca/eng/resources/radiation/introduction-to-radiation/types-and-sources-of-radiation.cfm. 14

Monte-Carlo for power, https://deliveroo.engineering/2018/12/07/monte-carlo-power-analysis.html. 11

NDR website, https://www.canada.ca/en/health-canada/services/health-risks-safety/radiation/national-dose-registry.html. 3, 19, 20

Appendix

In this chapter we present the R programming codes used to implement the power models discussed throughout the thesis.

```
# CODES FOR THE PROJECT: STATISTICAL POWER OF EPIDEMIOLOGICAL STUDIES TO
#  DETEDCT THE EFFECT OF IONIZING RADIATION ON THE MORTALITY
# RATE FROM CANCER BASED ON THE NATIONAL DOSE REGISTRY OF CANADA

# ALL THE CODES ARE WRITTEN BASED ON A VECTORIZATION APPROACH AS FAR AS
#POSSIBLE. WE TRY TO AVOID WRITING LOOPS TO SPEED UP CODES
# WHICH IS NECCESSARY IN THIS STUDY CINSIDERING LARGE DIMENSIONS WE DEAL WITH
#-----------------------------------------------------------------------------------------------------------------------------------------

# COHORT STUDY

#function in this part might possibly takes all or some of the following five
# parameters:
#     1. M: a FIXED vector containing the totall number of cases (deaths) at
# each stratum of dataset.
#     2. D: a FIXED matrix the ij-th entry of which shows the average dose of
# individuals at the j-th cell in the i-th strata of dataset.
#     3. p: a FIXED matrix the ij-th entry of which represents the proportion
# of cases in the j-th cell of the i-th stratum of dataset.
```

```
#     4. v: a matrix of VARIABLES the ij-th entry of which represents observed
# number of cases (deaths) at the j-th cell in the i-th stratum.
#           of dataset. The i-th column of v (related to the i-th stratum) forms
#     a multinomial variable.
#     5.theta: The true ERR parameter.

#-------------------------------------------------------------------------------

#Non-normalized score function for cohort
score_coh_nonnrm<-function(M,D,p,v){
  score<-sum(colSums(v*D)-M*colSums(p*D))
  score
}

#TEST score_coh_nonnrm
score_coh_nonnrm(M0,D0,p0,v0) #PASSED
#-----------------------------------------

#Expectation of the non-normalized score for cohort under parameter theta.
expectation_score_coh<-function(M,D,p,theta){
  cols_pD<-colSums(p*D)
  expectation<-theta*sum(M*(colSums(p*D**2)-cols_pD**2)/(1+theta*cols_pD))
  expectation
}

#-----------------------------------------
```

```
#Variance of the non-normalized score for cohort under parameter theta.
variance_score_coh<-function(M,D,p,theta){
  cols_pD<-colSums(p*(1+theta*D))
  #print(cols_pD)
  part1<-colSums((D**2)*t((t(p*(1+theta*D))/cols_pD))*(1-t((t(p*(1+theta*D))
   /cols_pD))))
  n_col<-length(M)
  n_row<-nrow(D)
  part2<-NULL
  for(i in 1:n_col){
    new_ent<-0
    for(j in 1:(n_row-1)){
      for(k in (j+1):n_row){
          new_ent<-new_ent+(D[j,i]*D[k,i]*p[j,i]*p[k,i]*(1+theta*D[j,i])
           *(1+theta*D[k,i]))
         }
    }
    part2<-c(part2,new_ent)
  }
  part2<-part2/cols_pD**2
  #print(part2)
  variance<-sum(M*(part1-2*part2))
  variance
}
#-----------------------------------------
```

```
# Number of cases to be observed in a cohort study to acheive a desired power:
# For future studies, this function gives the required number of cases to
# be observed in order to achieve a desired power.
# only a single strata is assumed (e.g. no age/sex stratum is considered.
# The cohort is treated a single study).
# We need some prior knowledge to calculate the sample size:
#  1. First fix the number of dose categories. D which is a vector that represents
# average population dose at each category must be known.
#  2. p, The proportion of the population of interest which falls in category j
# of dose should be provided.
# In addition, we need the following information:
#  3. Specify alpha, the size of the test and pow, the desired power.
#  4.Fix theta which is the smallest effect size we wish to detect at the given power.



Num_cases_coh<-function(D,p,theta,alpha,pow){
  D_mat<-matrix(D,nrow=length(D))
  p_mat<-matrix(p,nrow=length(p))
  var_0<-variance_score_coh(c(1),D_mat,p_mat,0)
  var_theta<-variance_score_coh(c(1),D_mat,p_mat,theta)
  p1<-sum(D**2*p)+theta*sum(D**3*p)
  p2<-1+theta*sum(p*D)
  p3<-sum(D*p)+theta*sum(D**2*p)
  p4<-sqrt(var_0/var_theta)
  p5<-theta*(sum(p*D**2)-(sum(p*D))**2)
```

```
  #print(list(p1=p1,p2=p2,p3=p3,p5=p5))
  size_new_formula<-ceiling((((p1*p2)-p3**2)*(p4*qnorm(1-alpha)-qnorm(1-pow))**2)
   /p5**2)
  size_old_formula<-ceiling((((p1*p2)-p3**2)*(qnorm(1-alpha)-qnorm(1-pow))**2)/p5**2)
  list(size_new_formula=size_new_formula,size_old_formula=size_old_formula)
}

#-------------------------------------------

# The following function can be used for two different tasks. First, it can take the
# essential characteristic of the dataset including D,p, M and theta as explained in
# the beginning of this section. Also takes the following extra parameters:
#   1. alpha: Size of the test.
#   2. n: Number of simulations.
# Then returns the power.

# The function can also be used to calculate proportion of rejected samples under null
# if we change the variance of test
# statistic from var_0 to to var_theta (in such a case parameter test_var should be set
# to anything other than "n").

power_coh<-function(M,D,p,theta,alpha,n,test_var="n"){
  var_0<-variance_score_coh(M,D,p,0)
  var_theta<-variance_score_coh(M,D,p,theta)
  if(test_var!="n"){score_var<-var_0}
  else{score_var<-var_theta}
```

```
  exp_theta<-expectation_score_coh(M,D,p,theta)
  SE_ratio<-sqrt(var_0/var_theta)
  if(test_var!="n"){sim_probs<-p}
  else{sim_probs<-t(t((p*(1+theta*D)))/(colSums(p*(1+theta*D))))}
  vector_rejected<-NULL
  n_row<-nrow(D)
  n_col<-ncol(D)
  for(k in 1:n){
    one_samp<-matrix(rep(0,n_row*n_col),nrow=n_row)
    for(i in 1:n_col){
      one_samp[,i]<-rmultinom(1,M[i],sim_probs[,i])
    }
    normalized_score<-score_coh_nonnrm(M,D,p,one_samp)/sqrt(score_var)
    vector_rejected<-c(vector_rejected,normalized_score>qnorm(1-alpha))
  }
  power_sim<-mean(vector_rejected)
  if(test_var!="n"){return(power_sim)}
  power_formula_old<-(1-pnorm(qnorm(1-alpha)-exp_theta/sqrt(var_theta)))
  power_formula_new<-(1-pnorm(SE_ratio*qnorm(1-alpha)-exp_theta/sqrt(var_theta)))
  list(power_sim=power_sim,power_formula_old=power_formula_old,
   power_formula_new=power_formula_new)
}

#-----------------------------------------------

# The function below calculates the power based on a naive approach. i.e., assumes no
```

```
# theoretical distribution for the score test statiastic neither under null nor under
# alternative. The idea is to construct the distribution of score under null
# experimentally and then take samples from alternative and compare it against the
# experimental distribution under null.
# The parameters are as before. There are two new parameters though:
#   1. n1: Number of samples to be drawn from null in order to derive its experimental
#  distribution.
#   2. n2: Number of samples to be drawn from alternative to calculate the power.


power_coh_naive<-function(M,D,p,theta,alpha,n1,n2){
  score_null<-NULL
  vector_rejected<-NULL
  n_row<-nrow(D)
  n_col<-ncol(D)
  sim_probs_null<-p
  sim_probs_alter<-t(t((p*(1+theta*D)))/(colSums(p*(1+theta*D))))
  for(k in 1:n1){
    one_samp<-matrix(rep(0,n_row*n_col),nrow=n_row)
    for(i in 1:n_col){
      one_samp[,i]<-rmultinom(1,M[i],sim_probs_null[,i])
    }
    score_null<-c(score_null,score_coh_nonnrm(M,D,p,one_samp))
  }
  treshold<-sort(score_null)[floor((1-alpha)*n1)+1]
  for(k in 1:n1){
    one_samp<-matrix(rep(0,n_row*n_col),nrow=n_row)
```

```
    for(i in 1:n_col){
      one_samp[,i]<-rmultinom(1,M[i],sim_probs_alter[,i])
    }
    vector_rejected<-c(vector_rejected,score_coh_nonnrm(M,D,p,one_samp)>treshold)
  }
  mean(vector_rejected)
}

#-------------------------------------------

#TEST normality of null and alternative:
# According to Rao's theorem the score statistic under null is assymptotically normal.
# However, In their model (sample size), they assume a normal distribution for the
# score test under alternative as well. There is no theoretical approach to prove
# this claim. However, we can test normality of score under null and alternative
# using Monte-Carlo simulation. See the function below.

test_norm_coh<-function(M,D,p,theta,n,test_var="n"){
  var<-variance_score_coh(M,D,p,theta)
  if(test_var!="n"){expect<-0}
  else{expect<-expectation_score_coh(M,D,p,theta)}
  if(test_var!="n"){sim_probs<-p}
  else{sim_probs<-t(t((p*(1+theta*D)))/(colSums(p*(1+theta*D))))}
  p_values<-NULL
  score_dist<-NULL
  n_row<-nrow(D)
```

```
  n_col<-ncol(D)
  for(k in 1:n){
    one_samp<-matrix(rep(0,n_row*n_col),nrow=n_row)
    for(i in 1:n_col){
     one_samp[,i]<-rmultinom(1,M[i],sim_probs[,i])
    }
    normalized_score<-(score_coh_nonnrm(M,D,p,one_samp)-expect)/sqrt(var)
    score_dist<-c(score_dist,normalized_score)
    p_values<-c(p_values,1-pnorm(normalized_score))
  }
  list(dist=score_dist,p_values=p_values)
}

#-----------------------------------------------------------------------------------------------------------------------------

# CASE-CONTROL STUDY

#function in this part might possibly takes all or some of the following five
# parameters:
#     1. S: a FIXED vector containing the totall number of cases (deaths) at each
#  stratum of dataset.
#     2. K: number of controls per case.
#     3. D: a FIXED matrix the ij-th entry of which shows the average dose of
# individuals at the j-th cell in the i-th strata of dataset.
#     4. p: a FIXED matrix the ij-th entry of which is given by:
#     (non-diseased individuals in cell j of stratum i)
```

```
# /(total number of non-diseased individuals in stratum i)
#     5. n_0: a matrix of VARIABLES the ij-th entry of which represents number of
# controls that fall in the j-th cell of the
#        i-th stratum of dataset. The i-th column (related to the i-th stratum) forms
# a multinomial variable.
#     6. n_1: a matrix of VARIABLES the ij-th entry of which represents number of cases
# (deaths) that fall in the j-th cell of the
#        i-th stratum of dataset. The i-th column (related to the i-th stratum) forms a
#  multinomial variable.
#     7.theta: The true EOR parameter.
#-------------------------------------------------------------------------------

#Non-normalized score function for case-control
score_cascnt_nonnrm<-function(D,K,n_0,n_1){
  score<-sum((-1/(K+1))*colSums(n_0*D)+(K/(K+1))*colSums(n_1*D))
  score
}

#-----------------------------------------------

#Expectation of the non-normalized score for case-control under parameter theta.
expectation_score_cascnt<-function(S,K,p,D,theta){
  expectation<-(K/(K+1))*sum(S*colSums((t(t(p*(1+theta*D))
   /colSums(p*(1+theta*D)))-p)*D))
  expectation
}
```

```
#-----------------------------------------------

#Variance of the non-normalized score for cohort under parameter theta.
variance_score_cascnt<-function(S,K,p,D,theta){
  n_col<-length(S)
  n_row<-nrow(D)
  cols_pD<-colSums(p*(1+theta*D))
  #cat("cols_pD",cols_pD,"\n")
  part0_1<-colSums((D**2)*p*(1-p))
  #cat("part0_1",(D**2)*p*(1-p),"\n")
  part0_2<-NULL
  for(i in 1:n_col){
    new_ent<-0
    for(j in 1:(n_row-1)){
      for(k in (j+1):n_row){
        new_ent<-new_ent+(D[j,i]*D[k,i]*p[j,i]*p[k,i])
      }
    }
    part0_2<-c(part0_2,new_ent)
  }
  #cat("part0_2",part0_2,"\n")
  part1_1<-colSums((D**2)*t((t(p*(1+theta*D))/cols_pD))*(1-t((t(p*(1+theta*D))
   /cols_pD))))
  #cat("part1_1",part1_1,"\n")
  part1_2<-NULL
```

```
  for(i in 1:n_col){
    new_ent<-0
    for(j in 1:(n_row-1)){
      for(k in (j+1):n_row){
      new_ent<-new_ent+(D[j,i]*D[k,i]*p[j,i]*p[k,i]*(1+theta*D[j,i])*(1+theta*D[k,i]))
      }
    }
    part1_2<-c(part1_2,new_ent)
  }
  part1_2<-part1_2/cols_pD**2
  #cat("part1_2",part1_2,"\n")
  variance<-(K/(K+1)**2)*sum(S*(part0_1-2*part0_2))+(K/(K+1))
   **2*sum(S*(part1_1-2*part1_2))
  variance
}

#-------------------------------------------------

# Sample size for case-control studies with K control per case:
# This function gives number of cases to be included in a study with K control per
# case in order to acheive the desired power.
# Notice that unlike the cohort studies where the sample size couldn't be directly
# calculated (rather, we only got number of
# cases to be observed as study goes on), for the case-control we can calculate the
# sample size. To determine the sample size (# of cases) for future studies assuming
# only a single strata (e.g. no age/sex stratum is considered.
```

```
# The cohort is treated a single study). We need some prior knowledge to calculate
# the sample size:
#  1. First fix the number of dose categories. D which is a vector that represents
#       average population dose at each category must be known.
#  2. p, The proportion of the non-diseased population of interest which falls in
#       category j of dose should be provided.
# In addition, we need the following information:
#  3. Specify alpha, the size of the test and pow, the desired power.
#  4.Fix theta which is the smallest effect size we wish to detect at the given power.



Num_cases_cascnt<-function(K,p,D,theta,alpha,pow){
  D_mat<-matrix(D,nrow=length(D))
  p_mat<-matrix(p,nrow=length(p))
  var_0<-variance_score_cascnt(c(1),K,p_mat,D_mat,0)
  var_theta<-variance_score_cascnt(c(1),K,p_mat,D_mat,theta)
  SE_ratio<-sqrt(var_0/var_theta)
  #print(SE_ratio)
  pD<-sum(p*(1+theta*D))
  part0_1<-sum((D**2)*p*(1-p))
  part0_2<-0
    for(j in 1:(length(D)-1)){
      for(k in (j+1):length(D)){
        part0_2<-part0_2+(D[j]*D[k]*p[j]*p[k])
      }
  }
```

```
  part1_1<-sum(((D**2)*(p*(1+theta*D))/(pD))*(1-(p*(1+theta*D))/(pD)))
  sum((D**2)*p*(1-p))
  part1_2<-0
  for(j in 1:(length(D)-1)){
    for(k in (j+1):length(D)){
      part1_2<-part1_2+(D[j]*D[k]*p[j]*p[k]*(1+theta*D[j])*(1+theta*D[k]))
    }
  }
  part1_2<-part1_2/pD**2
  part2<-((sum(p*D**2)-(sum(p*D))**2)/(pD))**2
  size_new_formula<-ceiling(((SE_ratio*qnorm(1-alpha)-qnorm(1-pow))
   **2*((part0_1-2*part0_2)+K*(part1_1-2*part1_2)))/(K*theta**2*part2))
  size_old_formula<-ceiling(((qnorm(1-alpha)-qnorm(1-pow))**2*((part0_1-2*part0_2)+
  K*(part1_1-2*part1_2)))/(K*theta**2*part2))
  list(size_new_formula=size_new_formula,size_old_formula=size_old_formula)
}

#-------------------------------------------------

# The following function takes the essential characteristic of the dataset including
# S,K,p,D and theta as explained in the beggining
# of this section. Also takes the following extrta parameters:
#   1. alpha: Size of the test.
#   2. n: Number of simulations.
# Then returns the power.
```

```
power_cascnt<-function(S,K,p,D,theta,alpha,n){
  var_0<-variance_score_cascnt(S,K,p,D,0)
  var_theta<-variance_score_cascnt(S,K,p,D,theta)
  exp_theta<-expectation_score_cascnt(S,K,p,D,theta)
  SE_ratio<-sqrt(var_0/var_theta)
  sim_probs_control<-p
  sim_probs_case<-t(t((p*(1+theta*D)))/(colSums(p*(1+theta*D))))
  vector_rejected<-NULL
  n_row<-nrow(D)
  n_col<-ncol(D)
  for(k in 1:n){
    one_samp_control<-matrix(rep(0,n_row*n_col),nrow=n_row)
    one_samp_case<-matrix(rep(0,n_row*n_col),nrow=n_row)
    for(i in 1:n_col){
      one_samp_control[,i]<-rmultinom(1,K*S[i],sim_probs_control[,i])
      one_samp_case[,i]<-rmultinom(1,S[i],sim_probs_case[,i])
    }
    normalized_score<-score_cascnt_nonnrm(D,K,one_samp_control,one_samp_case)
     /sqrt(var_0)
    vector_rejected<-c(vector_rejected,normalized_score>qnorm(1-alpha))
  }
  power_sim<-mean(vector_rejected)
  power_formula_new<-(1-pnorm(SE_ratio*qnorm(1-alpha)-exp_theta/sqrt(var_theta)))
  power_formula_old<-(1-pnorm(qnorm(1-alpha)-exp_theta/sqrt(var_theta)))
  list(power_sim=power_sim,power_formula_old=power_formula_old,
   power_formula_new=power_formula_new)
```

```
}

#-----------------------------------------------

# The function below calculates the power based on a naive approach. i.e., assumes no
# theoretical distribution for the
# score test statiastic neither under null nor under alternative. The idea is to
# construct the distribution of score under null
# experimentally and then take samples from alternative and compare it against the
# experimental distribution under null.
# The parameters are as before. There are two new parameters though:
#   1. n1: Number of samples to be drawn from null in order to derive its
# experimental distribution.
#   2. n2: Number of samples to be drawn from alternative to calculate the power.

power_cascnt_naive<-function(S,K,p,D,theta,alpha,n1,n2){
  score_null<-NULL
  vector_rejected<-NULL
  n_row<-nrow(D)
  n_col<-ncol(D)
  sim_probs_control<-p
  sim_probs_case<-t(t((p*(1+theta*D)))/(colSums(p*(1+theta*D))))
  for(k in 1:n1){
    one_samp_control<-matrix(rep(0,n_row*n_col),nrow=n_row)
    one_samp_case<-matrix(rep(0,n_row*n_col),nrow=n_row)
    for(i in 1:n_col){
```

```
    one_samp_control[,i]<-rmultinom(1,K*S[i],sim_probs_control[,i])
    one_samp_case[,i]<-rmultinom(1,S[i],sim_probs_control[,i])
   }
  score_null<-c(score_null,score_cascnt_nonnrm(D,K,one_samp_control,one_samp_case))
 }
 treshold<-sort(score_null)[floor((1-alpha)*n1)+1]
 for(k in 1:n2){
  one_samp_control<-matrix(rep(0,n_row*n_col),nrow=n_row)
  one_samp_case<-matrix(rep(0,n_row*n_col),nrow=n_row)
  for(i in 1:n_col){
   one_samp_control[,i]<-rmultinom(1,K*S[i],sim_probs_control[,i])
   one_samp_case[,i]<-rmultinom(1,S[i],sim_probs_case[,i])
  }
  vector_rejected<-c(vector_rejected,score_cascnt_nonnrm(D,K,one_samp_control,
   one_samp_case)>treshold)
 }
 mean(vector_rejected)
}

#-------------------------------------------------
#TEST normality of null and alternative:
# According to Rao's theorem the score statistic under null is assymptotically normal.
# However, In their model (sample size), they assume a normal distribution for the
# score test under alternative as well.
# There is no theoretical approach to prove this claim. However, we can test normality
# of score under null and alternative
```

```
# using Monte-Carlo simulation. See the function below.
test_norm_cascnt<-function(S,K,p,D,theta,n){
  var<-variance_score_cascnt(S,K,p,D,theta)
  expect<-expectation_score_cascnt(S,K,p,D,theta)
  sim_probs_control<-p
  sim_probs_case<-t(t((p*(1+theta*D)))/(colSums(p*(1+theta*D))))
  score_dist<-NULL
  p_values<-NULL
  n_row<-nrow(D)
  n_col<-ncol(D)
  for(k in 1:n){
    one_samp_control<-matrix(rep(0,n_row*n_col),nrow=n_row)
    one_samp_case<-matrix(rep(0,n_row*n_col),nrow=n_row)
    for(i in 1:n_col){
      one_samp_control[,i]<-rmultinom(1,K*S[i],sim_probs_control[,i])
      one_samp_case[,i]<-rmultinom(1,S[i],sim_probs_case[,i])
    }
    normalized_score<-(score_cascnt_nonnrm(D,K,one_samp_control,one_samp_case
     )-expect)/sqrt(var)
    score_dist<-c(score_dist,normalized_score)
    p_values<-c(p_values,1-pnorm(normalized_score))
}
list(dist=score_dist,p_values=p_values)
}
```

```
#-----------------------------------------------

# The function below puts out the p-vlue plots for null and alternative for both
# cohort and case-control in a single image. The parameteers are as before.
# There are 3 extra parameters:
#   1. r: The dataset number to be printed on the plots and (if applicable)
#     stored image.
#   2. sim: Number of simulations. Default is 10000.
#   3. save: to store or not the image. default is "no". Any other string value
#       results in image to be stored in the directory.

plot_norm_altogether<-function(M,K,p,D,theta,r,sim=10000,save="no"){
  test_coh_null<-test_norm_coh(M,D,p,0,sim)
  test_coh_alter<-test_norm_coh(M,D,p,theta,sim)
  test_null_cascnt<-test_norm_cascnt(M,K,p,D,0,sim)
  test_alter_cascnt<-test_norm_cascnt(M,K,p,D,theta,sim)
  if(save!="no"){
    jpeg(paste("dataset",r,".jpeg"))
  }
  par(mfrow=c(2,2))
  hist(test_coh_null$p_values,main = paste("dataset",r,"-null (cohort study)"),
   xlab = "p-value",xlim=c(0,1),breaks = seq(from=0, to=1, by=.1))
  hist(test_coh_alter$p_values,main = paste("dataset",r,"-alternative (cohort study)")
   ,xlab = "p-value",xlim=c(0,1),breaks = seq(from=0, to=1, by=.1))
  hist(test_null_cascnt$p_values,main = paste("dataset",r,"-null (case-control study)")
   ,xlab = "p-value",xlim=c(0,1),breaks = seq(from=0, to=1, by=.1))
```

```
  hist(test_alter_cascnt$p_values,main = paste("dataset",r,
      "-alternative (case-control study)"),xlab = "p-value",xlim=c(0,1),
   breaks = seq(from=0, to=1, by=.1))
  if(save!="no"){
    dev.off()
  }

}
#-------------------------------------------------
```