Figuring out the population attributable risk (PAR) with a bayesian approach; single and multivariatete analysis on cross-sectional study data.

University of Turku
Department of Computing
Bachelor's Thesis
Laboratory
December 2024
Peppi-Lotta Saari

UNIVERSITY OF TURKU
Department of Computing

Peppi-Lotta Saari: Figuring out the population attributable risk (PAR) with a bayesian approach; single and multivariatete analysis on cross-sectional study data.

Bachelor's Thesis, 50 p.
Laboratory
December 2024

Here be abstract

Keywords: tähän, lista, avainsanoista

Contents

1 Introduction

The primary focus of this thesis is a paper written by *Pirikahu et al. (2016)* by the title, *Bayesian methods of confidence interval construction for the population attributable risk from cross-sectional studies*. This paper proposes a fully Bayesian approach to constructing a confidence interval for the population-attributable risk (PAR). The PAR value represents the percentage of cases in a population that would not have occurred had the exposure not taken place.[1]

Confidence intervals are typically associated with Frequentist statistics. Bayesian approaches do not directly produce confidence intervals but credible intervals. We can create a credible interval with Bayesian approach and use repeated sampling to determine whether the credible interval exhibits the Frequentist properties of a confidence interval. I provide a more in-depth explanation of the differences between confidence and credible intervals in sections 2.0.3, 2.0.4, and 2.0.5.

1.1 Research question and goal

As a part of this thesis work, I will create an R package designed to construct a confidence interval for PAR based on the theoretical framework presented in the paper by *Pirikahu et al. (2016)*.

The paper by *Pirikahu et al. (2016)* provides a complete probability model for constructing a confidence interval in a single exposure scenario. I intend to implement this model within an R package and demonstrate with a workflow and

real data how the code can be used. I aim to create an easy-to-use and efficient R package that researchers can utilize. I'm focusing on thorough documentation to minimize user errors. Documentation is created by adhering to the roxygen2 structure.

Roxygen automatically generates a .Rd file from the comments in the R script without affecting code packageing and use.[2] I've chosen to use Roxygen because it allowes me to maintain documentation alongside the code.

I will evaluate the model and its expansions using simulated data. Evaluation happens by selecting specific parameters with known values and simulating data that correspond to these specified values. The *Pirikahu et al. (2016)* paper offers realistic values commonly employed in epidemiology.

In addition to the R package, another concrete outcome of this work is a data table containing evaluation results. Running evaluation with simulations is very resource-intensive. I have created a structure in the evaluation code that allows me to run in subsections and output the results in a CSV file.

1.2 Relevance

1.3 Philosophical difference between Frequentist and Bayesian approaches

1.4 Structure of this Thesis

Population-attributable risk describes the potential reduction in disease occurrences if a specific risk factor is eliminated from a population. PAR is a conditional value that can be derived with Bayes' theorem, which provides a mathematical foundation for this thesis work. In the next chapter, chapter **??**, I will begin by discussing

Bayes' theorem and its relevance to the *Pirikahu et al. (2016)* paper while providing sufficient context for understanding the underlying theory.

I will outline fundamental statistical terms crucial for later sections of this work. I will also present an overview of Bayesian inference, highlighting key characteristics and concepts such as prior and posterior distributions and likelihood. I assuming the target audience of this thesis are students of Faculty of technology and have limited prior knowledge of statistics.

In chapter 4, I will delve into the model described in by *Pirikahu et al. (2016)*, providing a detailed explanation and the necessary mathematical background. This chapter will also include descriptions of the code I created as part of this thesis, instructions on how to use it, and code for evaluating the model with simulated data. I also discuss the results of the evaluation in this chapter and provide figures and tables to illustrate the outcomes.

1.5 AI Disclaimer

I have utilized ChatGPT versions 3 and 4 to generate ideas for the structure of this thesis. All content has been composed by me from sources I have explicitly cited. While I found the AI helpful for brainstorming, I did not rely on it to generate the content. I have used Grammarly to enhance grammar and structure sentences in my text.

2 Statistical Foundations

This chapter defines foundational statistical concepts that are relevant to understandingthe approach defined by *Pirikahu et al. (2016)* and Bayesian inference.

2.0.1 Probability mass function

The probability mass function describes the chance that a discrete random variable is equal to a specific value. The sum of $f(x)$ over all x equals to 1 and $f(x) = P(X = x)$.[3]

2.0.2 Multinomial distribution

Multinomial distribution has a fixed number of trials independent from each other and samples have a fixed set of outcomes. Probability mass function of the multinomial distribution is

$$f(x_1, x_2, ..., x_k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k} \tag{2.1}$$

, where $n = x_1 + x_2 + ... + x_{k-1} + x_k$[4][5]

In Bayesian analysis, Dirichlet distribution is often used as a prior distribution to model a multinomial distribution. [5] This is considered a standard reference prior.[1]

2.0.3 Confidence intervals

Confidence interval is a frequentist term and it is based on repeated sampling theory.[6] Confidence interval is a range of numbers likely to include the unknown population parameter being estimated.[7] The confidence interval measures the uncertainty of an estimate. The interval has an upper and a lower limit. The true estimate lies within that interval some chosen percent of the time. The width of the confidence interval is determined by two factors: sample size n and standard deviation or standard error of the estimate. It is standard in health sciences to set the interval to be 95%.[8]

Confidence interval is also known as the regression coefficient, a term that describes the relationship between the predictor variable and outcome.[9] For 95% CI, the interval will contain the regression coefficient 95 out of 100 times when repeated.[6]

Confidence interval is differs from having a 95% probability that the regression coefficient is in the interval. In Frequentist statistics, parameters are not assigned probability values. [10] This hower is not the case in Bayesian statistics.

2.0.4 Credible intervals

Credible intervals describes a probability that the true value is within the chosen interval. 95% credible interval means that there is a 95% probability that the true value is in the interval. The Frequentist confidence interval is often misinterpreted this way. And that is why the credible interval is said to be more intuitive than the confidence interval.[10]

The tail method is a way of constructing the credible interval. It sets the lower and upper limits by symmetrically cutting the interval between both tails of a distribution. The limit are α / 2 and 1 $-$ α / 2 percentage points of the distribution. If the selected interval is 95%, then α is 0.05. [11]

2.0.5 Bayesian confidence intervals

Bayesian credible interval is a confidence interval if it demonstrates the Frequentist properties of a confidence interval. Whether an interval demonstrates Frequentist characteristics can be observed by calculating the overall percentage coverage for simulated data where various fixed parameters are set. [1] [11] Coverage percentage will show whether the interval is nominal. [1]

2.1 Bayesian Inference

2.1.1 Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.2}$$

$P(A|B)$ is the conditional probability [12] $P(A|B)$ is the probability that A happens given that B has happened. If A and B are independent events, then $P(A|B) = P(A)$. [12]

A person using 2.1.1 formula can place values acquired through Frequentist approaches and get a conditional probability; thus, using Bayes' theorem alone is insufficient to make a Bayesian inference. Bayesian inference is a broader philosophical and statistical approach to statistical inference.[13] In the Bayesian approach, the values of the posterior, prior, and likelihood are not set but random values of a distribution.[12] This allows us to make inferences even when the exact values are unknown and when data is limited.

2.1.2 Inference

Inference means summary or characterization.[14] It is the process of finding an appropriate model and fitting it to a data set.[15] Statistical inference's goal is to make predictions about an unobserved set of data y based on an already observed

set of data x.[16][15] The Bayesian approach to describing this connection is through probability that y happens given that x has happened. If $x = \{x_1, x_2, ..., x_n\}$ and $y = x_{n+1}$ then probability $P(y|x)$ is

$$P(y|x) = P(x_{n+1}|\{x_1, x_2, ..., x_n\}) \tag{2.3}$$

When the amount of observation x_i starts to grow, it becomes increasingly difficult to calculate the effect of each observation on y.

Formula 2.3 can be simplified to $P(y|\theta)$ where θ is a distribution where x_i are independent and identically distributed, i.e., iid. x_i depends only on $P(\theta)$ and not on other x_j.[16] This way of expressing uncertainty via a distribution leans on the concept of exchangeability.[15]

There is no operational difference between $P(\theta)$ which describes belief and $P(x|\theta)$ which is a measurable quantity. They both describe uncertainty. $P(x|\theta)$ is an updated version of the prior. With prior knowlwdge added the Bayes' theorem becomes

$$P(x|\theta) = \frac{P(\theta|x)P(x)}{P(\theta)} \tag{2.4}$$

[17]

2.1.3 Bayesian Data-Analysis

Ideally, Bayesian data analysis follows a three-step process, which I will adhere to in this work. The steps are as follows

- Full probability model: Create a model that is consistent with underlying scientific knowledge of the problem and, observed and unobserved data.
- Conditioning on observed data: Calculate and interpret the posterior distribution.

- Evaluation: Assess the model fit and the implications of the posterior distribution.

[15]

2.1.4 Prior

Prior distribution is the knowledge we have of the subject matter before we take into account the data.[18] Prior contains the plausibility for all parameters before the data has been taken into account. [13][14] Prior distribution is the best way to summarize information and lack of it.[14]

$P(\theta)$ is the prior distribution. $P(\theta)$ can be chosen based on some previous inference. Even if no prior distribution is available, a prior must be selected. Almost always, some domain knowledge is available and should be used to help choose the most suitable prior.[13] Regardless of prior knowledge, choosing a prior is often at least partly arbitrary.[14]

Whether the prior is a posterior distribution or a distribution selected by some other means, it doesn't make a difference qualitatively. The prior distribution is a distribution that describes the prior beliefs of the person making the inference.[16] The prior can significantly affect the posterior. Bad priors can lead to misleading posteriors and bad inference.[13] Priors can have a negligible, moderate, or highly noticeable effect on the posterior.[14]

Calculating the posterior becomes easier when the prior is selected from the same family as the posterior. Prior is chosen so that the parametric form of the prior is the same as the posterior's. The prior distribution is conjugate to the likelihood that the prior belongs to the same distribution family as the posterior distribution. [19] The priors mainly used in this thesis are the conjugate priors.

Some other types of priors are maximum entropy prior, parametric approximations, Laplace's prior, Jeffrey's prior, empirical, hierarchical, matching priors,

reference priors, and invariant priors. [14]

The prior distribution is not the only distribution that affects the posterior. The likelihood is another factor in calculating the posterior.

2.1.5 Likelihood

Likelihood or likelihood function has meaning outside of the Bayesian approach. Here, I will define likelihood in the context of Bayesian inference. Every time likelihood is mentioned in this thesis work, this definition applies.

Likelihood function contains the information that x_i brings to the inference.[14] Likelihood is derived by forming all data sequences and removing the ones inconsistent with the observed data. It usually is the distribution function assigned to a variable or distribution of variables. [13] The information provided by x_1 about $P(\theta)$ is contained in the distribution $l(\theta|x_1)$. When a new observation x_2 is made, it needs to comply with equation

$$l_1(\theta|x_1) = cl_2(\theta|x_2) \tag{2.5}$$

[14]

Equation 2.5 is called the likelihood principle, and it is valid when the θ in both sides of the equation is the same. $P(\theta)$ includes every unknown factor of the model. [14] Bayesian inference obeys the likelihood principle. Two probability models that have the same likelihood function lead to the same inference for $P(\theta)$ given a sample of data.[15]

When sample size increases, the meaning of the likelihood increases in relation to it. [13] The prior is modified by the data x trough the likelihood function. It represents the information about $P(\theta)$ coming from the data. In 2.5 c is a constant. Multiplication by a constant leaves the likelihood function unchanged. Only the relative value of the likelihood matters, and multiplying by a constant will have no

effect on the posterior. The constant will be canceled by normalizing the right side of 2.6.[18]

2.1.6 Posterior

The simplest definition for a posterior is the probability p conditional on data.[13] Posterior is what we know of the distribution θ given the knowledge of some observed data. As outlined in paragraphs 2.1.4 and 2.1.5, we know that likelihood is the only entity modifying the prior, so we get the simplified version of Bayes rule: *posterior distribution* $\propto$ *likelihood* $\times$ *prior distribution* where,

$$P(\theta|x) \propto cP(x|\theta)P(\theta) \tag{2.6}$$

[18]

Posterior is often mathematically complex, high-dimensional and direct inference is impossible. The number of dimensions is equal to the number of parameters. The posterior function is often an integral that is not solvable analytically, and the posterior distribution can not be evaluated exactly. Sampling the posterior provides a solution for this.[20] Some leading sampling methods, like ones based on Markov chain Monte Carlo (MCMC), only produce samples of the posterior instead of a mathematical function of the posterior.[13]

2.1.7 Posterior Sampling

Posterior sampling can be used to summarize and simulate the posterior. Samples are drawn from a "bucket" where values are present in proportion to their posterior probability. The samples will have the same distribution as the actual posterior. The more samples are drawn, the more exact the distribution will be. Sampling is part of the evaluation phase of Bayesian data analysis. [13]

Summarize

Summarization is divided into three categories

- Intervals of defined boundaries: questions about the frequency of the parameters in the posterior in chosen intervals.
- Intervals of defined mass: questions about confidence and credible intervals.
- Point estimates: questions about single points in the posterior distribution.

Simulation

Simulations are done to check the model and make predictions. Simulation can be done on a prior distribution to understand it better. Sampling from a known distribution will allow testing of whether a model is working correctly. Simulation can also be used to predict possible future observations. [13]

2.2 Cross-sectional study

A cross-sectional study is a snapshot of a population at a certain point in time. Both the exposure and outcome can be observed at the same time. Individuals for observation are chosen from the population that is relevant for the study. This study method does not consider new cases of the disease that develop over a selected period of time.

Subtypes of cross-sectional studies are

- A Descriptive cross-sectional study: good for evaluating the prevalence of one or more health outcomes in a population.
- Analytical cross-sectional study: measures the prevalence of outcomes end exposures. It's challenging to figure out causal relationships based on cross-sectional study alone.

- A Repeated cross-sectional study: conducts a study multiple times on the same population at different points in time. The individuals chosen for the tests at different study instances are not the same individuals. This type of study is good for showing changes in a population over time.

[21]

2.3 Coverage

Coverage and especially coverage propability has definitions in statistics. In this work coverage refers to a conffidence intervals coverage percent. Later in chapter 4 I will evaluate Bootstrap and fully Bayesian methods of confidence interval construction. I will evaluate accuracy of constructed intervals by checking often the actual calculated par is within a given interval. The aim is to create 95% confidence interval but the coverage percent will tell what is the actual achieved accuracy.

2.4 R language

R is a language and environment for statistical computing and graphics. R is available as Free Software under the Free Software Foundation's GNU General Public License terms in source code form. R is a GNU project. R is an environment where statistical techniques are implemented. It can be extended with the usage of packages. [22]

The base R has functions to perform all major statistical tests, plotting and matrix operations. It is provided as a combination of 14 different core packages: base, compiler, datasets, grDevices, graphics, grid, methods, parallel, splines, stats, stats4, tcltk, tools, and utils. Once installed on a machine, the package can then be loaded with the library() command.

R's functionality can be divided into data interaction, analysis, and result visualization. There are three major repositories for additional R packages: CRAN, Bioconductor, and R-Forge. CRAN package, devtools, provides a convenient function for installing packages directly from a GitHub repository. [23]

3 Literature on confidence interval construction of attributable risk

I'm interested in papers that outline, compare or extend on how the population attributable risk should be calculated and how a conffidence interval can be constructed for it. I dont't care about papers that use any method to calculate PAR as a part of there analysis. I have searced for papers in databases like research gate, google scholar and pubmed with attributable AND (risk OR fraction) AND "conffidence interva". I have chosen papers that seem relevant based on title and are free to access.

Drescher and Schill give a simple history timeline on the development of the concept of attributable risk. It was first introduced by Levin in 1953, variance estimate was derived by Walter in 1975. Atributable risk has been generalized to multifactorial and polytomous risk factors by Walter in 1976, Ejigou in 1979, Walker in 1981, Deneman and Schlesselman in 1983 and Wittermore in 1982 and 1983. Alternative astimates that are based on Mantel-Haenszel estimator have been derived by Greenland in 1987 and Kuritz and Landis in 1987 and 1988. Unified approach for calculating attributable risk in general multivariate setting was given by Bruzzi et al. in 1885 and variance estimations were derived by Benichou and Gail in 1990. [24] Attributable risk seems to be discussed more often in relation to case-control studies than cross-sectional study settings. [25] The Drescher and Schill

paper outlines the history of AR but the paper is not applicaple other wise as it is about case control studies and not cross-sectional.

As many as 16 names have been used to denote population attributable risk, most popular of these include attributable risk, etiologic fraction, attributable risk percentage, fraction of etiology, attributable fraction.[26] Since the orignal introduction of "attributable proportion" there has been an expansion of terminology around this topic. Term include; attributable fraction, attributable risk, attributable risk percent, preventable fraction, prevented fraction, assigned shares, excess fraction, risk fraction, rate fraction, and etiologic fraction. [27]

The term "attributable" has a causal interpretation: PAF is the estimated fraction of all cases that would not have occurred if there had been no exposure. Similarly, we can calculate PAF for the joint effects of two or more exposures. Such a PAF is expected to be less than the sum of the PAF for each exposure because people exposed to both exposures should not be counted twice. As usual, we make the strong assumptions that there is no bias in the study design and data analysis; in particular, that the estimated effect is adjusted for all confounders. In addition, we assume that removing the exposure does not affect other risk factors[28]

3.1 Benefit of understanding Population Attributable risk

High risk ration might not indicate a serious health problem if the persons exposed to the risk factor are few. On the other hand if the risk ration is small but a large amount of population is exposed to it it could indicate a serious health impact. Atrributable fraction takes in to acount both the risk relative to those exposed and the number of people who are exposed. [29] The full impact of a risk factor depends both on the size of the risk and proportion of the population exposed to the risk

factor. [24] Attributable fraction can be used to estimate the impact of completely removing a risk factor from population. The estimation evaluates how the disease burden could be reduced by controlling the exposure to risk factors. [30] Attributable riks is not a substitute for relative risk and it is an alternative additional dimension to healt hazard appraisal.[31]

To prevent disease and injury, one must understand the impact of risk factors on a population's health. Often, one outcome has multiple risk factors, and a risk factor is associated with multiple outcomes. For example, crowded housing and poor nutrition are risk factors for contracting infectious agents, which are the direct cause of disease. Prevention can be done by encouraging healty behaviour. Risk can be targeted with tax, financial incentives, campaigns about heath, engineering and legislation. Risk factors in a population are not immutable. Improved medical care, ageing and public health like vaccination have an affect what are the risk affecting a population. [32] I a factor is found rather rarely in a population hence had low attibutable risk health administrators who are conserned with preventativve strategies don't need to focus on such risk factors.[31]

When mesuring blood pressure, a person whose bloodpressure is measured to be over 140mmHg is considered hypertensive. A large part of the population is not hypertensive but still have higher than ideal blood pressure. In these types of cases the exposure to a risk factor is not dichotomous or "yes or no"-situation but is more on a scale. It can give incomplete picture if only hypertensive members of the population are considered and a large group members with elevated blood pressure are not taken into account when making population health related decisions.[32]

PAF assumes that there is a perfect intervention which eradicates the exposure. However, complete removal of an exposure is often unrealistic; even with legal restrictions and cessation programmes, many people will continue to smoke.[28]

3.2 Population attributable fraction

Attributable fraction quatifies the porportion of cases attributable to a risk factor. [30] The number of deaths, disease and injury attributable to risk factor is quatified by applying the population attributable fraction to the total number of outcomes. [32]

Atrributable franction was originally formulated for a single dichotomous risk factor basically meaning that the risk factor is eather present or absent. This separation ignores other risk factor that may act together with the risk factor being observed. [30]

$$PAF = \frac{P(D^+) - P(D^+|E^-)}{P(D^+)} \tag{3.1}$$

[1]

This formula that Pirikahu et al. have used is representing the PAF with probabilities and was originally proposed by MacManon and Pugh.[30] $P(D^+)$ is the probability of disease in the whole population and $P(D^+|E^-)$ is the eventually hypothesized probability of disease in the unexposed population.[24]

3.3 Population attributable risk

Population-attributable risk measures the impact of completely removing a risk factor in a population.

$$PAR = P(D^+) - P(D^+|E^-) \tag{3.2}$$

where D is the disease status, and E is the exposure status. PAR is a probability distribution calculated by removing the probability distribution of disease cases, given that exposure has not happened from the probability distribution of all disease

occurrences. [1]

3.4 Confidence interval construction

Variance estimates are needed to construct conffidence intervals. [33]

Walter wrote a paper, "The Estimation and Interpretation of Attributable Risk in Health Research," in 1976. This paper outlines three study designs and how to estimate the variance of attributable risk in them. This is one of the earliest papers I could find, considering confidence interval construction for attributable risk in cross sectional study setting. The paper considers the simplest situation of dichotomous disease outcome and risk factors that can be represented in a 2x2 contingency table. The paper's third study design is cross-sectional, where N number of unstratified samples are taken from the population, where each sampled individual has the disease or not, and is exposed or not. Walter uses λ to denote attributable risk.[31]

$$\lambda = 1 - \frac{1}{(\pi_{11} + \pi_{12})(\psi - 1) + 1} \tag{3.3}$$

where

$$\psi = \frac{\pi_{11}(\pi_{21} + \pi_{22})}{(\pi_{11} + \pi_{12})\pi_{21}} \tag{3.4}$$

[31]

3.5 Adjusted Attributable risk

The way risk factor work together is not always straight forward and sum of attributable fraction could exeed 100% if calculated separately. [30] Risk factor can form causal chains where some risk factors a direct causes of disease and others

are risk factors of risk factors and indirectly affect the outcome. The sum of risk factors separately is often more than the combined mortality and burden of disease attributable to the group of risk factors. This means reducing any risk factor could lead to prevention of an outcome.[32]

Some times the it is impossible or unefficiet to try and get a matched case and control groups. In some cases for example the case group might be older that the control group and the exposure is not independent of this factor. Higher age could mean longer exposure for example as is the case for smoking. One solution to take into account the different ages in the case and control groups is to standardize estimate by calculating a weighted sum where the weight indicate porportion of cases explained by the chosen atrribute[31]

$$\lambda_s = \sum_i \omega_i \lambda_i / \sum_i \omega_i \tag{3.5}$$

For age groups the weight ω_i could be chosen to be the count of samples.[31] The equation above shows how the weights could be chosen based on sample count in "case-load weighting". DiMaso et al. give a more generally weighted-sum is

$$_{adj}AF_E = \sum_{k=1}^{K} W_k \times AF_{E,k} \tag{3.6}$$

where sum of weights is equal to 1 $\sum_{k=1}^{K} W_k = 1$. This way the weights could be chosen to increase precission in "precision-weighing". Mantel–Haenszel approach adjusts AF via adjustment of the relative risk.[30]

It is also possible to decompose attributable risk λ into component representing a risk factor and the risk due to confounding with a second confounded factor. A crude estimate with out any regards of confounding is[31]

DiMaso et al. introduce modeling approach Bruzzi et al. applied logistic regression models in a way that is valid for cross-sectional studies as well as case-control

and cohort studies.Bruzzi's approach is a weighted-sum approach and consists in a weighted sum of relative risks estimates by odds ratios. For each stratum of the adjustment factors, relative risks are combined with the stratum-specific proportion of cases [30]

$$_{adj}AF_E = 1 - \sum_{k=1}^{K}\sum_{q=0}^{Q}\frac{\rho_{q,k}}{RR_{q|k}} \tag{3.7}$$

[30] First sum is taken over all adjustment strata and second sum is taken over all risk factor levels. $\rho_{q,k}$ is the proportion of cases with respect to the qth risk factor level and kth adjustment stratum, whereas $RR_{q|k}$ represents the relative risk for the qth risk factor level given the kth adjustment stratum. $\rho_{q,k}$ is replaced by the observed proportion of cases and by replacing $RR_{q|k}$ by the maximum-likelihood estimate obtained from a regression model.[30]

DiMaso et al give definitions to sequatial and average attributable fractions.

$$\lambda_{crude} = 1 - \frac{cn_2}{dn_1} \tag{3.8}$$

where, $n_1 = a+c$ is the total count of the case group and $n_2 = b+d$ is the total count of the control group. Estimate of component of attributable riks due to confounding is[31]

$$\lambda_{conf} = 1 - \frac{cn_2*}{dn_1} \tag{3.9}$$

, where $b* = \sum b_i^*$ and $b_i^* = \frac{a_i d_i}{c_i}$ and atrributable risk adjusted for confounding is[31]

$$\lambda_{adj} = \lambda_{crude} - \lambda_{conf} = \frac{c(b* - b)}{dn_1} \tag{3.10}$$

[31]

3.5.1 Confounding

Good example case for when calculating for confounding is when estimating the proportion of cases attributable to high blood pressure after adjusting for the effect of smoking which is also known to affect blood pressure.[31]

Risk factors are not always dichotomous but could be found on $(k+1)$ levels. Level 0 is baseline and all other levels are assosiated with incresed risk. Attributable risk at an exposure level i is $\hat{\lambda}_i = \frac{(a_i d - b_i c)}{n_1 d}$ and the atrributable risk for all levels

$$1 - \hat{\lambda} = 1 - \sum_{i=1}^{k} \hat{\lambda}_i = 1 - \frac{cn_2}{dn_1} \tag{3.11}$$

. Confounding over all exposure levels is same as just calculating exposure as dichotomous.[31]

Interaction of risk factor can be tested by comparing combined attributable risk λ with the product of attributable risk of factor A and factor B

$$1 - \lambda = (1 - \lambda_A)(1 - \lambda_B) \tag{3.12}$$

. If this equation doesn't hold there is synergy or antagonism betweent the risk factors and the equation for confounding should be used. [31]

Another way to write adjusted attributable risk with confoundinf is

$$AR = 1 - \sum_{j=1}^{k} \frac{P(C=j)P(D=1|E=0, C=j)}{P(D=0)} \tag{3.13}$$

where C is a stratum variable with k gategories. Cell frequesncies are $\hat{p} = (\hat{p}_{11j}, ..., \hat{p}_{22k})$ generated from a multinomial distribution $p = (p_{11j}, ..., p_{22k})$. [33]

3.5.2 Logarithmic Transformation

Attributable risk can be expressed as

$$AR = \frac{p(r-1)}{1+p(r-1)} \tag{3.14}$$

where p is the porortion exposed to the risk factor and r is the risk ratio. AR varies between 0 and 1. Maximum likelihood based interval is

$$\{\frac{\hat{p}(\hat{r}-1)exp(-u)}{1+\hat{p}(\hat{r}-1)exp(-u)}, \frac{\hat{p}(\hat{r}-1)exp(u)}{1+\hat{p}(\hat{r}-1)exp(u)}\} \tag{3.15}$$

where $u = Z_{1-\frac{1}{2}\alpha}v$ and Z is a standard normal random variable such that $P(Z \leq Z_{1-\frac{1}{2}\alpha}) = 1 - \frac{1}{2}\alpha$.

Leung and Kupper expand upon Water's work and show that logarithmic transformation (LT) method in conjunction with the monotonicity property leads to a confidence interval with much better properties than those possessed by Walter's ML based interval. Leung and Kupper start by stating a 2x2 contingency table is where $N = a + b + c + d$

Exposed	D^+ (has disease)	D^- (no disease)
E^+	a	b
E^-	c	d

for cross-sectional study design variance is

$$v^2 = \{\frac{(a+c)(c+d)}{ad-bc}\}^2\{\frac{ad(N-c)+bc^2}{Nc(a+c)(c+d)}\} \tag{3.16}$$

and $100(1-a)\%$ confidence interval is

$$\{\frac{(ad-bc)exp(-u)}{Nc+(ad-bc)exp(-u)}, \frac{(ad-bc)exp(u)}{Nc+(ad-bc)exp(u)}\} \tag{3.17}$$

where $u = Z_{1-\frac{1}{2}\alpha}v$.

If $|\hat{AR} - \frac{1}{2}| \leq \frac{\sqrt{3}}{6}$ the Logarithmic Transformation produces shorter interval than the Maximum likelihood mased approach.[29]

Transformations can improve estimating conffidence intervals when they are based on normal distribution. Logit-transformations seem to produce better results than Log-transformations. Log-tranformed intervals tent to be wider without improvement to coverage. [33]

3.5.3 Delta

The Delta method is the standard approach to variance estimation for PAR. The bootstrap method outperforms the delta method in terms of coverage and interval length. [1] Estimates of the adjusted attributable risk is often done by applying the delta method. Delta method generally tends to underestimate the standard error leading to biased conffidence intervals.

Benichou and Gail first applied the delta method to compute variance estimates for attributable risk in 1989 for case-control studies. Basu and Landis adopted this for multinomial cross-sectional studies in 1995.[33]

Variance based on delta method is

$$\hat{Var}(\hat{AR}) = D(\hat{p})' \times V(\hat{p}) \times D(\hat{p}) \tag{3.18}$$

where $V(\hat{p}) = n^{-1}[diag(\hat{p}) - \hat{p}\hat{p}']$ is variance-covariance matrix of $\hat{p}$ with n being the number of observations. $D(\hat{p} = \frac{\partial g(\hat{p})}{\partial \hat{p}_{qlk}})_{q=0,1;l=0,1}$ and $g(\hat{p}) = \hat{AR}$[33]

3.5.4 Bootstrap

The bootstrap method was first introduced by Efron in 1979. It gained attencion when computing capacity has grown. The unknown atributable risk is estimated by taking samples with replacement out of underlying data. $\hat{AR}$ is computed every

time a sample is taken. This is repeated B times and caculated distrubution of the replications is taken to be the estimated distribution of the true parameter. $\hat{AR}^{*boot} = (\hat{AR}_1^{*boot}, ..., \hat{AR}_B^{*boot})$. [33]

The Bootstrap method approximates the distribution of a statistic by repeated sampling. The samples are drawn from a fitted model or from a dataset with replacement, AKA placing the sample back into the "sample bucket. Drawing from a fitted model is parametric, and drawing from a dataset with replacement is non-parametric.[1]

A contingency table has a cell for each classification. A dataset with one exposure and one outcome can be represented in a 2x2 contingency table with four classifications. The probability of selecting a classification is the same as the estimated value in a parametric model. The parametric and non-parametric bootstraps for a 2x2 contingency table are the same when the sample size is the same as the dataset size.[1]

To reduce computational burden generating B samples from whole data can be done with weighted method. The choise of the random weight vector is determined by the bootstrap version use. In nonparametric bootstrap for example random variables are taken from multinomial distribution. A variant of nonparametric bootstrap is bayesian bootstrap where random variables are taken from Dirichlet distribution. [33]

$$CI_{normal} = [\frac{1}{B}\sum_{b=1}^{B} \hat{AR}_b^{*boot} - Z_{1-\alpha} \times \hat{se}(\hat{AR}^{*boot}); \frac{1}{B}\sum_{b=1}^{B} \hat{AR}_b^{*boot} + Z_{1-\alpha} \times \hat{se}(\hat{AR}^{*boot})] \tag{3.19}$$

where $Z_{1-\alpha}$ is the $100 \times (1-\alpha)$ precentile of the standard normal distribution and

$$\hat{se}(\hat{AR}^{*boot}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{AR}_b^{*boot} - \frac{1}{B}\sum_{b=1}^{B}\hat{AR}_b^{*boot})^2} \tag{3.20}$$

[33]

3.5.5 Jack knife

The jackknife as another variant of computer intensive method was suggested by Quenouille in 1949. In this method the data set is changed systematically by leaving out every observation of the data once at a time. The original sample comprises n observations, n replications $\hat{AR}^{*jack} = (\hat{AR}_1^{*jack}, ..., \hat{AR}_n^{*jack})$. [33]

$$CI_{normal} = [\frac{1}{n}\sum_{i=1}^{n} \hat{AR}_i^{*jack} - Z_{1-\alpha} \times \hat{se}(\hat{AR}^{*jack}); \frac{1}{n}\sum_{i=1}^{n} \hat{AR}_i^{*jack} + Z_{1-\alpha} \times \hat{se}(\hat{AR}^{*jack})] \tag{3.21}$$

where $Z_{1-\alpha}$ is the $100 \times (1-\alpha)$ precentile of the standard normal distribution and

$$\hat{se}(\hat{AR}^{*jack}) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{AR}_i^{*jack} - \frac{1}{n}\sum_{i=1}^{n}\hat{AR}_i^{*jack})^2} \tag{3.22}$$

[33]

3.5.6 Comparison of methods

Lui has done a comparison of fve interval estimators of attributable risk; two interval estimators suggested by Leung and Kupper, the interval estimator using the logarithmic transformation suggested by Fleiss, the interval estimator on the basis of Wald's test statistic suggested by Walter and the interval estimator using an approach similar as that for derivation of Fieller's theorem. They compare the performances by calculating coverage probability and average legnth of resulting confidence interval. All of these methods seem to produce less than nominal results[25]

Lehnert-Batar et al. tested constructing conffidence intervals with more modern techniques; delta, jack-knife, Bootstarp and Bayesian bootstrap methods with sim-

ulated data. The bayesian and non parametric bootstraps seem to perform the best. Jack-knife produces acceptable results. Bootstrap and Jack-knife out perfor delta in almost all situations. When the sample size is small even the computationally expensive Bootstrap and Jack-knife fall short. Delta method outperforms the bootstrap and jackknife in some situations, if the logit-transformation of AR is applied. Bootstrap seems to perform better than Jack-knife in most situations.[33]

Arude attributable fractions are biased because they don't consider adjustment for potential confounders. The adjusted attributable fraction quantifies the effect removing one risk factor from the population has after controlling for other risk factors. Given two dichotomous risk factors ${}_{E}1$ and ${}_{E}2$ the crude and adjusted attributable fraction will only coincide when ${}_{E}1$ and ${}_{E}2$ are independently distributed in the population or when ${}_{E}2$ alone does not increase disease risk. The weighted-sum approach allows to control both for confounders and effect modifiers, but biased estimates occur when data is sparse. This is because risk factors are not independent and thus multi-exposed cases will be counted more than once. Adjusted attributable fraction should not be used to try partition joint risks into exposure-specific contributions. [30]

Lee et al. compare Green, Delta and Monte Carlo methods to calculating the 95% interval of the attibutable fraction. Their conclusion is that there is no significant difference between these methods.[34]

3.6 Software

Lehnert-Batar et al. used R to compare delta, jack-knife, Bootstarp and Bayesian bootstrap methods. The program is available in ISSAN which is the IMBE Statistical Software Archive Network (http://www.imbe.med.uni-erlangen.de/issan/issan.htm)[33]

There are packages called epiR, AF package, averisk and pifpaf for R that allow calculating versions of attributable fraction that Di Maso et al. outlined. [30]

The AF package is presented in a paper writen by Dahlqwist et al. The paper outlines usage of the pakage via examples done on real openly available data. It provides functions to calculate confounder-adjusted attributable fraction for cross-sectional studies, case–control studies and cohort studies. This package allows for attributable fraction estimation with binary exposures. Dahlqwist et al. name three other R packages for calculating and constructing conffidence interval for attributable fraction. These are epiR, attribrisk and paf but the writers don't consider these up-to-date at the time of writing the paper in 2015.[35]

4 The Bayesian Approach to Confidence Interval Construction for Population Attributable Risk (PAR)

In this chapter, I will outline the approach proposed in the paper titled *Bayesian Methods for Confidence Interval Construction of Population Attributable Risk from Cross-Sectional Studies* by *Pirikahu et al. (2016)*. I will propose a way to implement the mathematical model into R code, evaluate the model with simulated data and give example workflow for usinf the R package with real life data.

4.1 Mathematical Model

Table 4.1: 2 x 2 Contingency Table For n Samples

Exposed	D^+ (has disease)	D^- (no disease)	Total
E^+	a	b	a + b
E^-	c	d	c + d
Total	a + c	b + d	n

Let n denotes the total sample size, where $a + b + c + d = n$. Probability of exposute is $P(E^+) = \frac{a+b}{n}$, probability of being unexposed is $P(E^-) = \frac{c+d}{n}$ and probability of having the disease regardles of exposure status is $P(D^+) = \frac{a+c}{n}$.

From the contingency table stucture it is evident that a cross-sectional study in-

corporating one exposure variable and one disease variable can be characterized as a multinomial distribution with four independent possible outcomes. These outcomes can be described using the multinomial distribution as follows

$$(a, b, c, d) \sim Multinomial(n, p_{11}, p_{10}, p_{01}, p_{00}) \tag{4.1}$$

where,

- $p_{11} = P(D^+ \cap E^+) = P(D^+|E^+)P(E^+) = P(E^+|D^+)P(D^+) = \frac{a}{n}$: The probability of being exposed and having the disease.
- $p_{10} = P(D^- \cap E^+) = P(D^-|E^+)P(E^+) = P(E^+|D^-)P(D^-) = \frac{b}{n}$: The probability of being exposed and not having the disease.
- $p_{01} = P(D^+ \cap E^-) = P(D^+|E^-)P(E^-) = P(E^-|D^+)P(D^+) = \frac{c}{n}$: The probability of not being exposed and having the disease.
- $p_{00} = P(D^- \cap E^-) = P(D^-|E^-)P(E^-) = P(E^-|D^-)P(D^-) = \frac{d}{n}$: The probability of not being exposed and not having the disease.

The population-attributable risk (PAR) refers to the proportion of disease within a population that can be attributed to a specific exposure. The PAR can be calculated using the formula 3.2. By applying Bayes' theorem and by incorporating the values listed in 4.1, we obtain the maximum likelihood estimation function for PAR.

$$\begin{aligned} PAR &= P(D^+) - P(D^+|E^-) \\ &= P(D^+) - \frac{P(E^-|D^+)P(D^+)}{P(E^-)} \\ &= \frac{a+c}{n} - \frac{\frac{c}{n}}{\frac{c+d}{n}} \\ &= \frac{a+c}{n} - \frac{c}{n} \times \frac{n}{c+d} \\ &= \frac{a+c}{n} - \frac{c}{c+d} \\ &= \frac{a+c}{a+b+c+d} - \frac{c}{c+d} \end{aligned} \tag{4.2}$$

A prior distribution that estimates all the situations described in a contingency table is: $\theta = (p_{11}, p_{10}, p_{01}, p_{00})$ Observed values or samples are: $x = (a, b, c, d)$. A probability mass function denotes the likelihood in respect to p_k as

$$f(x|\theta) = \frac{n!}{a!b!c!d!} p_{11}^a p_{10}^b p_{01}^c p_{00}^d \tag{4.3}$$

The posterior distribution is:

$$p(a, b, c, d|p_{11}, p_{10}, p_{01}, p_{00}) = p(\theta|x) \propto f(x|\theta)p(\theta) \tag{4.4}$$

Due to the conjugacy relationship, the posterior can be found analytically in relation to the prior. Posterior is

$$\theta|x \ Dirichlet(a+1, b+1, c+1, d+1). \tag{4.5}$$

Representing posteriors analytically is computationally less expensive than using MCMC simulation. The confidence interval is a Frequentist concept; however, we can determine the Frequentist coverage of the credible interval through simulated data.

4.2 R Code

Implementing the model in R is quite straightforward once the underlying mathematical model is grasped. Constructing a confidence interval for a dataset involves four key steps

- Extracting the contingency table values from data
- Generating new contingency tables by simulation
- Calculating the PAR for each simulated table
- Constructing the confidence interval from the simulated PAR values

4.2.1 Extracting the Contingency Table Values from Data

4.2.1 method takes a data frame and extracts the values for a, b, c, and d, returning them in a single vector. a, b, c, and d are the count for the different categories and align with categories given in 4.1. All the function that I've created for this package, that use these category values, are expecting a vector with a, b, c, and d values in this order. I've provided this helper function so that the user can use this an trust that the values from a data set are extracted and saved to the correct order.

```
extract_abcd <- function(
    data,
    exposure_col,
    outcome_col)
{
    x_0e0d <- sum(data[[exposure_col]] == 0
        & data[[outcome_col]] == 0)
    x_0e1d <- sum(data[[exposure_col]] == 0
        & data[[outcome_col]] == 1)
```

```
    x_1e0d <- sum(data[[exposure_col]] == 1
        & data[[outcome_col]] == 0)
    x_1e1d <- sum(data[[exposure_col]] == 1
        & data[[outcome_col]] == 1)


    return(c(
    x_1e1d = x_1e1d,
    x_1e0d = x_1e0d,
    x_0e1d = x_0e1d,
    x_0e0d = x_0e0d
    ))
}
```

4.2.2 Calculate PAR

I've created a function to calculate Population Attributable Risk from contingency table cell values. 4.2.2 function accepts a vector of values and returns the corresponding PAR value. The vector must consist of the values a, b, c, and d in that precise order.

If these values are obtained using the method $extract_abcd$, the vector will be properly sequenced and can be directly passed to the 4.2.2 function.

```
calculate_par <- function(x) {
    x <- as.numeric(x)
    a <- x[1]
    b <- x[2]
    c <- x[3]
    d <- x[4]
```

```
    ...
}
```

The logic behind calculating par is derived from equation 3.2. In cases where the total number of samples n where $n = a + b + c + d$ is small and the exposure rates $a + b$ are low, the function may return zero values for c and d. Since zero cannot be a divisor, I have opted to handle this situation by returning a value of 0 if either the sum of c and d is zero.

```
if (c + d == 0) {
    return(0)
}
par <- (a + c) / (a + b + c + d) - c / (c + d)
```

The function $calculate_par$ will return a single value, that is, the PAR.

```
calculate_par <- function(x) {
...
    return(par)
}
```

4.2.3 Code for Constructing the Confidence Interval

Similar to the $calculate_par$ function, the $calculate_bayesian_ci$ method requires a vector of values as a parameter, which must be specified by the user. Additionally, this method can accept values for interval coverage, a vector for the prior distribution, and a value for the number of samples; however, these additional parameters are optional and have default settings. The function expects the x and *prior* vectors to be ordered as a, b, c, and d. The default setting for the number of samples is 10000, which is considered sufficiently large according to *Pirikahu et al. (2016)*. The

prior defaults to a vector of ones, indicating a non-informative uniform prior. The standard default value for the interval coverage is 0.95, which is commonly used.

```
calculate_bayesian_ci <- function(
    "par",
    x,
    interval = 0.95,
    prior = c(1, 1, 1, 1),
    sample_count = 10000
) {
    x <- as.numeric(x)
    a <- x[1]
    b <- x[2]
    c <- x[3]
    d <- x[4]
    n <- a + b + c + d
    prior <- as.numeric(prior)
    ...
}
```

4.2.3 is the main logic of the code. Samples of contingency tables are generated using the Dirichlet distribution. 4.2.3 is calling the *rdirichlet* function from the *MCMCpack* package to form new contingency tables and saves them to *samples* variable. *Samples* contains *sample_count* number of tables. With vector operation *apply* $calculate_par$ is applied to each table and we get "*sample_count*" of PAR values. The confidence interval is calculated using the *quantile* function. The function returns a matrix with the lower bound of the confidence interval as the first value and the upper bound as the second value.

```
calculate_bayesian_ci <- function(
```

```
...
    samples <- rdirichlet(
        sample_count,
        c(a + prior[1],
        b + prior[2],
        c + prior[3],
        d + prior[4],
        n
        )
    )
    samples <- apply(samples, 2, function(x) x * n)

    par_samples <- apply(samples, 1, calculate_par)

    # Calculate the confidence interval
    confidence_interval <- quantile(
        par_samples,
        c(
        (1 - interval) / 2,
        1 - (1 - interval) / 2
        )
    )
...
)
```

Finally, the function returns a matrix with the lower bound of the confidence interval That are extracted from the quantile function as the first value and the upper bound as the second value.

```
calculate_bayesian_ci <- function
...
    return(matrix(c(
        confidence_interval[1],
        confidence_interval[2]
    )))
```

Despite the fact that multinomial simulations are generally more efficient than MCMC simulations, conducting evaluations can be resource-intensive. When the *compiler* is loaded, the $compile_all$ function can be called and all functions within the package are converted from human-readable code to machine code, enhancing execution speed.

The ‘cmpfun‘ function from the Byte Code Compiler can be utilized to compile a function into machine code. This function compiles the body of a closure and returns a new closure with the same formal parameters while replacing the original body with the compiled expression. [36]

```
    compile_all <- function() {
        calculate_bayesian_ci <-
            cmpfun(calculate_bayesian_ci)
        calculate_bootstrap_ci <-
            cmpfun(calculate_bootstrap_ci)
        calculate_par <-
            cmpfun(calculate_par)
        calculate_paf <-
            cmpfun(calculate_paf)
        extract_abcd <-
            cmpfun(extract_abcd)
```

```
    }
```

4.3 Evaluation of the Model

We will run simulation based on selected known values for parameters p, q, e and n to explore performance.

- $p = P(D^+|\ E^+)$, the probability of having the disease given exposure.
- $q = P(D^+|\ E^-)$, the probability of having the disease given no exposure.
- $e = P(E^+)$, the probability of exposure.
- n, the total number of samples.

Because exposure either has happened or not, we can deduce that $P(E^-) = 1 - P(E^+) = 1 - e$. And because a person can either have the disease or not, we can deduce that $P(D^-|E^-) = 1 - P(D^+|E^-) = 1 - q$. and $P(D^+|E^+) = 1 - P(D^-|E^+) = 1 - p$. We can use this knowledge to form the probabilities for the different categories

- $a = p_{11} \times n = P(D^+ \cap\ E^+) \times n = P(D^+|\ E^+) \times P(E+) = p \times e \times n$
- $b = p_{10} \times n = P(D^- \cap\ E^+) \times n = P(D^-|\ E^+) \times P(E+) = (1-p) \times e \times n$
- $c = p_{01} \times n = P(D^+ \cap\ E^-) \times n = P(D^+|\ E^-) \times P(E-) = q \times (1-e) \times n$
- $d = p_{00} \times n = P(D^- \cap\ E^+) \times n = P(D^-|\ E^-) \times P(E-) = (1-q) \times (1-e) \times n$

Rate of decease occurance is $P(D^+)$ and can be calculated as

$$\begin{aligned} P(D^+) &= P(D^+ \cap\ E^+) + P(D^+ \cap\ E^-) \\ &= P(D^+|E^+)P(E^+) + P(D^+|E^-)P(E^-) \\ &= p \times e + q \times (1-e) \end{aligned} \tag{4.6}$$

We need to generate 10,000 contingency tables that correspond to selected variables p, q, e, and n, using the multinomial distrubution 4.1. The parameter values for the simulation are as follows

Table 4.2: Parameters for the simulation

p	0.001	0.01	0.05	0.1	0.2	0.3	0.35	0.4	0.45	0.5
q	0.001	0.01	0.05	0.1	0.2	0.3	0.35	0.4	0.45	0.5
e	0.01	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9

We can expand this evaluation matrix to compare different sample sizes

Table 4.3: Parameters for the simulation

n	16	64	256	1024

4.3.1 Code for Evaluating the Model

Table 4.3 provides the parameters for p, q, and e. We need to reverse engineer the probabilities for a, b, c, and d based on the selected parameters. Once we have the probabilities, we can generate contingency tables and construct confidence intervals for PAR employing two different methods: the Bayesian approach proposed by *Pirikahu et al. (2016)* and bootstrap. I loop through all the different combinations of parameters and o the following steps for each combination.

subsubsectionCalculating Probabilities for a, b, c, and d From p, q, e

First we start by calculating the probabilities associated with selected p, q and e. The function *get_probabilities_2x2_table* computes and returns the probabilities for a, b, c, and d as p_11, p_10, p_01 and p_00. The function returns a list containing these probabilities in order.

```
get_probabilities_2x2_table <- function( p, q, e ) {
    p_11 <- p * e
    p_10 <- ( 1 - p ) * e
    p_01 <- q * ( 1 - e )
```

```
        p_00 <- ( 1 - q ) * ( 1 - e )


        return(list(
            p_11 = p_11,
            p_10 = p_10,
            p_01 = p_01,
            p_00 = p_00)
        )
      }
```

By utilizing the probabilities for categories and total sample size n, we can simulate contingency tables that align with the chosen parameters p, q, e and n. For each of these tables, we can compute the confidence interval (CI) using the *calculate_bayesian_ci* function.

subsubsectionSimulating Contingency Tables

While *Pirikahu et al. (2016)* gives that 10,000 simulations would be ideal. Due to resource constraints I have, I have reduced the number of simulations to a 1000. Simulated contingency tables are saved to *samples* variable.

```
    samples <- rmultinom(
        1000,
        row$n,
        c(row$p_11, row$p_10, row$p_01, row$p_00)
    )
```

subsubsectionConstructing the Confidence Interval

The confidence interval is computed for each generated contingency table in *samples*. All of these tables represent the same parameters: p, q, e, and n, and share the same PAR. PAR calculated with the Bayesian method to get a set of

confidence intervals

```
bayes_cis <- apply(samples, 2, function(sample) {
  a <- sample[1]
  b <- sample[2]
  c <- sample[3]
  d <- sample[4]
  n <- a + b + c + d
  calculate_bayesian_ci(
    "par",
    c(a, b, c, d),
    interval,
    prior,
    10000
    )
})
```

Bootstrap method is applied to the same set of samples to get a set of confidence intervals

```
boot_cis <- apply(samples, 2, function(sample) {
  a <- sample[1]
  b <- sample[2]
  c <- sample[3]
  d <- sample[4]
  n <- a + b + c + d
  calculate_bayesian_ci(
    "par",
    c(a, b, c, d),
    interval,
```

```
        10000
        )
    })
```

subsubsectionCalculating Metrics The coverage is considered nominal if the actual PAR falls within the lower and upper bounds of the interval in 95% of the simulations, or at least $1000 * 0.95 = 950$ times. Calculating the actual PAR from p, q and e has to be done so that we can calculate coverage percentage. I have given definitions for p, q, and e in 4.3. Values from 4.3 can be placed into 3.2 to get an equation to calculate PAR from the parameters.

$$\begin{aligned} PAR &= P(D^+) - P(D^+|E^-) \\ &= p * e + q * (1 - e) - q \end{aligned} \tag{4.7}$$

The second row of equation 4.7 can be directly implemented in code. We can calculate the coverage percentage by checking if the actual Par value is with in the upper and lower bounds of the confidence interval.

```
    bayes_coverage <- mean(
        bayes_cis[1, ] <= row$actual_par
        & bayes_cis[2, ] >= row$actual_par
    )
```

The mean length of interval across all simulations is computed along with the coverage percentage.

```
bayes_mean_length <- mean(
    bayes_cis[2, ] - bayes_cis[1, ]
)
```

Coverage percentage and mean interval legth are two metrics that can be used to compare different models. If the coverage percentages of all models meet the

nominal criteria, meaning they are equal to or exceed the specified interval value, the model with the narrowest mean length is considered the most effective.

I have calculated coverage precentage and mean interval lenth for both Bayesian and bootstrap methods. I have save the result in a CSV file for further analysis.

subsubsectionCSV File Output

After the steps out lined in previous section. I have generated a CSv file with the following columns

p, q, e, n, p_11, p_10, p_01, p_00, $actual\_par$, $bayes\_ci\_mean\_length$, $bayes\_ci\_coverage$, $boot\_ci\_mean\_length$ and $boot\_ci\_coverage$. The file can be found in the data folder of the created R package.

subsubsectionOptimizing the Evaluation Code

The steps I've outlined in paragraph 4.3.1 are computationally expensive. I generate a 1000 contingency tables and constructions of the conffidence interval requires 10000 simulations for Bayes and Bootstrap each. This amounts to $1000 * 10000 * 2 = 20,000,000$ simulations.

The values and calculations that do not require simulations are computed outside of a loop and subsequently outputted into the CSV file. These calculations include the actual PAR and the probabilities p_11, p_10, p_01, p_00. The simulations can then be executed in a loop, utilizing values from the file for each run. This approach allows me to divide the simulation into smaller subsets, enabling me to select a sufficiently large subset to run on my machine.

To further enhance the speed of the simulation, I have implemented several optimizations. For instance, I compile the functions, as demonstrated in 4.2.3. Machine code is faster to run then un compiled code. Compiling can be done by calling the $compile_all$ function.

Additionally, I utilize the *parallel* package to execute the multible samples in parallel, allowing it to leverage the available cores on my machine. The simulation

can utilize all cores if no other processes are running; otherwise, it will run on any unused cores. The variables 'start' and 'end' represent the starting and ending rows of the file that define the subset to be processed. I am enabling parallel execution with the future package as follows

```
plan(multisession)
results <- future_map(start:end, function(i) {
...
```

4.4 Comparison of Fully Bayesian Method with Bootstrap Method

Run the eval code and print some figures here fron the CSV file.

4.4.1 Different Priors

Run eval code with different priors and print some figures here.

4.4.2 Example with Real Data

4.4.3 Data

For the purpose of demonstrating the code developed in this thesis, I have utilized the dataset "Risk Factors for Cardiovascular Heart Disease" [37], curated by Kuzak Dempsy and made available on Kaggle. This dataset provides a comprehensive collection of health-related variables known to influence cardiovascular disease risk, aligning closely with risk factors identified by the Centers for Disease Control and Prevention.[38] The dataset includes the following features:

- **Age:** Age of the individual, recorded in days (integer).

- **Gender:** Gender of the individual (categorical: male or female).
- **Height:** Height in centimeters (integer).
- **Weight:** Weight in kilograms (integer).
- **ap_hi:** Systolic blood pressure reading (integer).
- **ap_lo:** Diastolic blood pressure reading (integer).
- **Cholesterol:** Cholesterol level, categorized into ordinal groups (integer).
- **Gluc:** Blood glucose level, categorized into ordinal groups (integer).
- **Smoke:** Smoking status (boolean).
- **Alco:** Alcohol consumption status (boolean).
- **Active:** Physical activity status (boolean).
- **Cardio:** Presence (1) or absence (0) of cardiovascular disease (boolean, target variable).

4.4.4 Code

Imports

```
library(MCMCpack)
library(dplyr)
library(compiler)
library(data.table)
library(devtools)
```

```
devtools::install_github("peppi-lotta/par")
library(par)
```

Read the data and calculate BMI

```
file_path <- "./data.csv"
data <- read.csv(file_path)
data <- data[sample(nrow(data), 1000), ]
data <- data %>%
    mutate(over_weight = ifelse(weight/((height/100)^2) > 24.99,
```

Calculate the population attributable risk and confidence interval

```
exposure_col <- "over_weight"
outcome_col <- "cardio"

table <- table(
    data[[exposure_col]],
    data[[outcome_col]],
    dnn = c(
    exposure_col,
    outcome_col
    )
)
print(table)

x <- extract_abcd(data, exposure_col, outcome_col)
print(x)

par <- calculate_par(x)
cat("PAR: ", par, "\n")
```

```
interval = 0.95
prior = c(1, 1, 1, 1)
sample_count = 10000


bay_ci <- calculate_bayesian_ci(
    "par",
    x,
    interval,
    prior,
    sample_count
)
cat("Confidence␣Interval:\n")
print(bay_ci)
```

Standardisation

```
unique_ages <- unique(data[["age"]])
unique_ages <- sort(unique_ages)
print(unique_ages)


data <- data %>%
    mutate(
    Age_group = case_when(
        age/365  < 20 ~ "0-19",
        age/365 >= 20 & age/365  < 30 ~ "20-29",
        age/365  >= 30 & age/365  < 40 ~ "30-39",
        age/365  >= 40 & age/365  < 50 ~ "40-49",
```

```
        age/365  >= 50 & age/365  < 60 ~ "50-59",
        age/365  >= 60 & age/365  < 70 ~ "50-59",
        age/365  >= 70 ~ "70+"
    )
    )
print(head(data))

age_groups <- unique(data$Age_group)
par <- 0
bay_lower_bound <- 0
bay_upper_bound <- 0
boot_lower_bound <- 0
boot_upper_bound <- 0

for (age_group in age_groups) {
    exposure_col <- "over_weight"
    outcome_col <- "cardio"

    age_data <- data[data$Age_group == age_group, ]
    weight <- nrow(age_data)

    x <- extract_abcd(age_data, exposure_col, outcome_col)
    par <- par + calculate_par(x) * weight

    bay_ci <- calculate_bayesian_ci(
    "par",
    x,
```

```
    interval,
    prior,
    sample_count
    )
    bay_lower_bound <- bay_lower_bound + bay_ci[1] * weight
    bay_upper_bound <- bay_upper_bound + bay_ci[2] * weight

    boot_ci <- calculate_bootstrap_ci(
    "par",
    x,
    interval,
    sample_count
    )
    boot_lower_bound <- boot_lower_bound + boot_ci[1] * weight
    boot_upper_bound <- boot_upper_bound + boot_ci[2] * weight
}

cat("par:")
print(par/nrow(data))
cat("Confidence␣Interval:\n")
cat("Bayes:\n")
print(c(bay_lower_bound/nrow(data), bay_upper_bound/nrow(data)))
cat("Bootstrap:\n")
print(c(boot_lower_bound/nrow(data), boot_upper_bound/nrow(data)))
```

5 Conclusion

5.1 Coverage of different approaces

5.2 Limitations

5.3 Future work

The formulas for PAR and PAF *Pirikahu et al. (2016)* use are considere crude and are in general biased. Two ways to improve estimation are stratification aka sorting data into groups and modeling. [30]

5.3.1 Expanding the Pirikahu Approach to a Hierarchical approach

Data pertaining to humans or animals often exhibit hierarchical structures, whether deliberately organized or otherwise, and this aspect should not be overlooked.[39] To enhance the approach introduced by *Pirikahu et al. (2016)*, it is essential to adopt a hierarchical model. This expansion will enable us to calculate the variability in the PAR across different subgroups, allowing for a more precise understanding of how subgroup-specific characteristics contribute to this variability.

Hierarchical models, commonly referred to as multilevel models, leverage the knowledge gained from previous clusters when addressing new ones. These clus-

ters may consist of individuals, groups, locations, or, in the context of this work, populations. The advantages of multilevel models include improved estimations, as well as the mitigation of the bias caused by over-sampled clusters dominating the inferences. Furthermore, these models implicitly account for variation and better preserve uncertainty, thereby avoiding unnecessary data transformation.[13]

Data can exhibit hierarchical, nested, or clustered structures. A hierarchy comprises units organized at varying levels, with groupings occurring even in random formations. The membership of these groups, and vice versa, influences the characteristics of their members.[39] Typically, datasets consist of samples that serve as the lowest-level units but can be organized into higher-level units. For instance, a dataset containing student information can have students as the lowest-level units, which can then be grouped by class, school, or district. Students from a single school form a distinct cluster.[40]