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

Turun yliopisto
Tietotekniikan laitos
TkK-tutkielma
Labra
Joulukuu 2024
Peppi-Lotta Saari

TURUN YLIOPISTO
Tietotekniikan laitos

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

TkK-tutkielma, 18 s., 2 liites.
Labra
Joulukuu 2024

Here be abstract

Asiasanat: tähän, lista, avainsanoista

Sisällys

1 Introduction

1.1 Baysian Mindset

1.2 Perseiving risk

1.3 The Problem

1.4 The Solution

1.4.1 R package

2 Key Consepts

This chapter will provide understanding of the Bayesian approach. This chapter will lead the reader through Bayes' theorem to Bayesian inference and data-analytics. The meaning of population attributable risk and relevant consepts to statistics in general will be explored in this chapter.

2.1 Bayes' theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

where

- $P(A|B)$ is the conditional probability [1] or posterior distribution.[2] $P(A|B)$ is the probability that A happens given that B has happened. If A and B are independent events, then $P(A|B) = P(A)$. [1]
- $P(B|A)$ is the likelihood[1][2]
- $P(A)$ is the prior probability [1] or prior distribution.[2]
- $P(B)$ is the set of unobserved data.

A person using this formula can just place values that they have acquired through frequentist approaches and get a conditional probability thus using the Bayes' theorem alone is not enough to make Bayesian inference. Bayesian inference is a broader

philosophical and statistical approach to statistical inference.[2] In the Bayesian approach the values of the posterior, prior and likelihood are no set but random values of a distribution.[1] This allows us to make inference even when the exact values are not known and when data is limited.

2.2 Key Statistics consepts

Probability mass function

Probability mass function describes the probability that a discrete random variable is equal to a certain value: $f(x) = P(X = x)$ $f(x)$ over all x equals to 1.[3]

Multinomial distribution

Properties of the multinomial distribution: fixed number of trials, trials are independent from each other, samples have a fixed set of outcomes. The Probability mass function of the multinomial distribution is:

$$f(x_1, x_2, ..., x_k) = \frac{n!}{x_1!x_2!...x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k} \tag{2.2}$$

, 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 often considered a standard reference prior.[6]

Confidence intervals

Confidence interval measures uncertainty of the estimate. The interval has a upper and lower limit and the true estimate lies with in that interval with some conffidence level. Width of the confidence interval is determined by two factors: sample size n,

and standard diviation or standard error of the estimate. Confidence level in health sciences ís usually chosen to be 95%.[7]

Confidence interval is a frequentist term. Confidence interval is based on repeated sampling theory.[8] Confidence interval is a range of numbers that is likely to include the unknown population parameter that is being estimated.[9] This is also known as the regression coefficient, a term that describes the relationship between the predictor variable and outcome.[10] For 95% CI the interval will contain the regression coefficient 95 out of 100 times when repeated.[8] This is different from having a 95% probability that the regression coefficient is in the interval. In frequentist statistics parameters aren's assigned propabilities. [11]

Credibility intervals

Credibility interval is the Bayesian equivalent of confidence interval. It is a probability that the true value is contained in the chosen interval. 95% credibility interval means that there is 95% probability that the true value is in the interval. The frequentist confidence interval is often missinterpered this way. This is why the credibility interval is more intuitive than the confidence interval.[11]

2.3 Bayesian Inference

Inference means summary or characterization[12]. It is the process of finding an appropriate model and fitting it to a set of data.[13] Statistical inference's goal is to make predictions about an unobserved set of data y based on already observed set of data x.[14][13] The Bayesian approach to describing this connection is trough probability that y happens given x. If $x = \{x_1, x_2, ..., x_n\}$ and $y = x_{n+1}$ then:

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

When the amount of observation x_i starts to grow it becomes increasingly more difficult to calculate the effect of each observation on y. The formula can be simplified to $P(y|\theta)$ where θ is a distribution where x_i are independent and identically distributed i.e. iid. This simplifies the problem because we can assume x_i depends only on θ and not on other x_j.[14] This way of expressing uncertainty via a distribution, leans on the concept on exchangeability. Exchangeability is a basic concept of statistical analysis. [13] There is no operational difference between θ 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. The Bayes' theorem becomes:

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

[15]

2.3.1 Prior

Prior distribution is the knowledge we have of the subject matter before we take into account the data.[16] Prior provides the plausibility for all parameters before the data is taken into account. [2][12] Prior distribution is the best way to summarize information and lack of it.[12]

θ can be chosen based on some previous inference obtained by Bayesian approach. Even if there is no prior distribution available a prior must be chosen and this should not be completely arbitrary. Almost always there a some domain knowledge available that will help choose the most suitable prior.[2] Choosing a prior is still often at least partly arbitrary.[12] Whether the prior is a posterior distribution or a distribution chosen by some other means, it doesn't make a difference qualitatively. Prior distribution is a distribution that describes prior beliefs of the person doing the inference.[14] The prior can have a great effect on the posterior. Bad priors can lead

to misleading posteriors and bad inference.[2] Priors can have a negligible, moderate or highly noticeable effect on the posterior.[12]

The priors mainly used in this thesis are the conjugate priors. Prior distribution is conjugate to the likelihood if the prior belongs to the same distribution family as the posterior distribution. Prior is chosen so that the parametric form of the prior is the same as the posterior's. This is to make calculating posterior easiest. [17]

Some other types of priors are: maximum entropy prior, parametric approximations, Laplace's prior, the Jeffrey's prior, empirical, hierarchical, matching priors, reference priors, and invariant priors. [12]

In my R implementation of the approach specified in the Pirikahu paper, the prior can be specified trough a parameter when the method is called. In the evaluation phase in chapter 6 I will be try out the different priors and see how they perform with a large amount of data and when the dataset is small.

Prior distribution is not the only distribution that affects the posterior. Another factor in calculating the posterior is the likelihood.

2.3.2 Likelihood

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

Likelihood function contains the information that x_i brings to the inference. [12] Likelihood is derived by forming all data sequences and removing the ones that are inconsistent with the observed data. It usually is the distribution function assigned to a variable or distribution of variables. [2] The information provided by X_1 about θ is contained in the distribution $l(\theta|x_1)$. When a new observation x_2 is made it

needs to comply with equation:

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

[12]. This is called the likelihood principle, and it is valid when the inference is about the same θ and θ includes every unknown factor of the model. [12]. Bayesian inference obeys the likelihood principle. Simply put: two probability models that have the same likelihood function lead to the same inference for θ given a sample of data.[13] When the sample size increases the meaning of the likelihood grows with it. [2] The prior is modified by the data x trough the likelihood function. It represents the information about θ 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 θ. By normalizing the right side of 2.6 the constant will be canceled out.[16]

2.3.3 Posterior

The simplest definition for a posterior is the probability p conditional on the data.[2] Posterior is what we know of the distribution θ given the knowledge of some observed data. As outlined in paragraphs 2.3.1 and 2.3.2 we know that likelihood is the only entity modifying the prior so we get the simplified version of Bayes rule: $posterior\ distribution \propto likelihood * prior\ distribution$ where:

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

[16]

Posterir is often mathematically complex and high-dimentional and direct inference on it is not possible. The number of dimensions is equal to the number of parameters. The posterior function is often an integral that is not solvable by ana-

lytical means and the posterior distribution can not be evaluated exactly. Sampling the posterior provides a solution for this.[18]

2.3.4 Sampling

Monte-Carlo-Marcov-Chain

2.3.5 Bayesian Data-analysis

Ideally Bayesian data-analysis is a three step process:

- Full probability model: A model that is consistent with underlying scientific knowledge of the problem and, observed and unobserved data.
- Conditioning on observed data: Calculating and interpreting the posterior distribution.
- Evaluation: fit of the model and the implications of the posterior distribution.

[13]

2.4 PAR, Population attributable risk

Population attributable risk measures the impact of complete removal of a risk factor in a population.

$$PAR = P(D+) - P(D+|E-) \tag{2.7}$$

, where D is the disease status and E is the exposure status. PAR is a probability distribution that is calculated by removing the probability distribution of disease cases that occur given that an exposure has not happened from the probability distribution of all disease occurrences . [6]

2.5 R language

2.5.1 Stan

2.6 Cross-sectional study

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 population that is relevant for the study. This study method does not take into account new cases of the desease that develope over a selected period of time.

Subtypes of cross-sectional study:

- Descriptive cross-sectional study is good for evaluating the prevalence of one or more health outcomes in a population.
- Analytical cross-sectional study measures prevalence of outcomes end exposures. It's difficult to figure out causal relationships based on cross-sectional study alone.
- Repeated cross-sectional study is conducted 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 happening in a population over time.

[19]

3 The Bayesin Approach to Confidence Interval Construction for Poulation Attributable Risk (PAR)

In this chapter I will go over the approach proposed in paper *Bayesian methods of confidence interval construction for the population attributable risk from cross-sectional studies* by Pirikahu&al.

Taulukko 3.1: 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

n denotes the total sample size where $a+b+c+d = n$. From the contingecy table we can see that a cross-sectional study with one exposure variable and one desease variable can be seen as a multinomial distribution with four possible outcomes that are independet from each other. These outcomes can be described with multinomial distribution as follows:

$$(a, b, c, d) \sim Multinomial(n, p_{11}, p_{12}, p_{21}, p_{22}) \tag{3.1}$$

- $P(E+) = \frac{a+b}{n}$: The propability of the expososed in a population.

- $P(E-) = \frac{c+d}{n}$: The propability of the expososed in a population.
- $P(D+) = \frac{a+c}{n}$, The propability of having the disease in a population.
- $p_{11} = P(D+ \cap E+) = P(D+|E+)P(E+) = P(E+|D+)P(D+) = \frac{a}{n}$: The probability of being exposed and not having the disease.
- $p_{12} = P(D- \cap E+) = P(D-|E+)P(E+) = P(E+|D-)P(D+) = \frac{b}{n}$: The probability of being exposed and not having the disease.
- $p_{21} = P(D+ \cap E-) = P(D+|E-)P(E-) = P(E-|D+)P(D+) = \frac{c}{n}$: The probability of not being exposed and having the disease.
- $p_{22} = P(D- \cap E-) = P(D-|E-)P(E-) = P(E-|D-)P(D-) = \frac{d}{n}$: The probability of not being exposed and not having the

The population attributable risk (PAR) is defined as the proportion of the disease in a population that have occured due to exposure. The PAR can be calculated with formula 2.7. With Bayes' theorem we can rewrite the formula and by add the values from the list above we get the maximum likehood estimation function for PAR.

$$\begin{aligned} PAR &= P(D+) - P(D+|E-) \\ &= P(D+) - \frac{P(E-|D+)P(D+)}{P(E-)} \\ &= \frac{a+c}{n} - \frac{\frac{c}{n}}{\frac{c+d}{n}} \\ &= \frac{a+c}{n} - \frac{n}{c} \times \frac{c+d}{n} \\ &= \frac{a+c}{n} - \frac{c+d}{c} \\ &= \frac{a+c}{a+b+c+d} - \frac{c+d}{c} \end{aligned} \tag{3.2}$$

Prior distribution that estimates all the situations dercribed in the contingency table is: $\theta = (p_{11}, p_{12}, p_{21}, p_{22})$. Observed values or samples are: $x = (a, b, c, d)$. And

the likelihood in respect to p_k is denoted by a propability mass function:

$$f(x|\theta) = \frac{n!}{a!b!c!d!} p_{11}^a p_{12}^b p_{21}^c p_{22}^d \tag{3.3}$$

Posterior distribution is: $p(a, b, c, d|p_{11}, p_{12}, p_{21}, p_{22}) = p(\theta|x) \propto f(x|\theta)p(\theta)$. Due to conjugacy relationship posterior can be found analytically to be

$$\theta|x \ Dirichlet(a1, b1, c1, d1). \tag{3.4}$$

It is computationally much less expencive to represent posteriors analytically than with MCMC simulation.

Conffidence interval is a frequentist consept but it can be constructed based on the repeated sampling of the posterior even with this approach.

4 R code

```
def hello_world():
    print("Hello, world!")
```

5 Expanding the approach for multivariate analysis

Expanding the Bayesian abroach to take into account multible variables for conffidence interval construction seems straight forward. Let's consider the equation for calculating the PAR

5.1 Dependency between variables

5.1.1 DAG

5.2 Math

5.3 Code

5.3.1 Stan

6 Evaluation

7 Comparison

7.1 Frequentist models

Description (and short math on models here) A table of results here

7.2 Machine Learning

Description (and short math on models here) A table of results here

Taulukko 7.1: Taulukon otsikko tulee taulun yläpuolelle

Taulun	elementit	erotetaan
toisistaan	et-merkillä	
soluja voi myös		jättää tyhjäksi

Taulukko 7.2: Taulukon2 otsikko tulee taulun yläpuolelle

Taulun	elementit	erotetaan
toisistaan	et-merkillä	
soluja voi myös		jättää tyhjäksi

8 Conclusion

8.1 Summary of the model and work

8.2 Implementation in R

8.3 Future work

Lähdeluettelo

[1] A. Gut, "Probability: A Graduate Course", teoksessa *Probability: A Graduate Course*, 2005. url: https://api.semanticscholar.org/CorpusID:117844972.

[2] R. Mcelreath, "Statistical Rethinking: A Bayesian Course with Examples in R and Stan", teoksessa *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*, 2015.

[3] *Probability Mass Functions*, https://online.stat.psu.edu/stat414/lesson/7/7.2, Accessed: 2024-10-14.

[4] *he Multinomial Distribution*, https://online.stat.psu.edu/stat504/book/export/html/6672, Accessed: 2024-10-14.

[5] S. Sinharay, "Continuous Probability Distributions", teoksessa *International Encyclopedia of Education (Third Edition)*, P. Peterson, E. Baker ja B. McGaw, toim., Third Edition, Oxford: Elsevier, 2010, s. 98–102, ISBN: 978-0-08-044894-7. DOI: https://doi.org/10.1016/B978-0-08-044894-7.01720-6. url: https://www.sciencedirect.com/science/article/pii/B9780080448947017206.

[6] S. Pirikahu, G. Jones, M. L. Hazelton ja C. Heuer, "Bayesian methods of confidence interval construction for the population attributable risk from cross-sectional studies", *Statistics in Medicine*, vol. 35, s. 3117–3130, 2016. url: https://api.semanticscholar.org/CorpusID:3497293.

[7] L. C. Hespanhol, C. S. Vallio, L. da Cunha Menezes Costa ja B. T. Saragiotto, "Understanding and interpreting confidence and credible intervals around effect estimates.", *Brazilian journal of physical therapy*, 2019. url: https://api.semanticscholar.org/CorpusID:58581702.

[8] R. van de Schoot ja S. Depaoli, "Bayesian analyses : where to start and what to report", *The European health psychologist*, vol. 16, s. 75–84, 2014. url: https://api.semanticscholar.org/CorpusID:142633945.

[9] B. Illowsky ja S. T. Dean, "Introductory Statistics: OpenStax", 2013. url: https://api.semanticscholar.org/CorpusID:126293670.

[10] *he Multinomial Distribution*, https://statisticsbyjim.com/glossary/regression-coefficient/, Accessed: 2024-10-14.

[11] I. Fornacon-Wood, H. B. Mistry, C. Johnson-Hart, C. Faivre-Finn, J. P. O'Connor ja G. Price, "Understanding the Differences Between Bayesian and Frequentist Statistics.", *International journal of radiation oncology, biology, physics*, vol. 112 5, s. 1076–1082, 2022. url: https://api.semanticscholar.org/CorpusID:247420107.

[12] C. P. Robert, "The Bayesian choice : from decision-theoretic foundations to computational implementation", teoksessa *The Bayesian choice : from decision-theoretic foundations to computational implementation*, 2007. url: https://api.semanticscholar.org/CorpusID:50937448.

[13] A. Gelman, "Bayesian Data Analysis", teoksessa *Bayesian Data Analysis*, 2014.

[14] D. Lindley, "The 1988 Wald Memorial Lectures: The Present Position in Bayesian Statistics", *Statistical Science*, vol. 5, s. 44–65, 1990. url: https://api.semanticscholar.org/CorpusID:117797898.

[15] Y. Pawitan, ”In all likelihood : statistical modelling and inference using likelihood”, *The Mathematical Gazette*, vol. 86, s. 375–376, 2002. url: https://api.semanticscholar.org/CorpusID:117422783.

[16] G. E. P. Box ja G. C. Tiao, ”Bayesian inference in statistical analysis”, *International Statistical Review*, vol. 43, s. 242, 1973. url: https://api.semanticscholar.org/CorpusID:122028907.

[17] M. Sugiyama, ”Chapter 17 - Bayesian Inference”, teoksessa *Introduction to Statistical Machine Learning*, M. Sugiyama, toim., Boston: Morgan Kaufmann, 2016, s. 185–196, ISBN: 978-0-12-802121-7. DOI: https://doi.org/10.1016/B978-0-12-802121-7.00028-5. url: https://www.sciencedirect.com/science/article/pii/B9780128021217000285.

[18] R. van de Schoot et al., ”Bayesian statistics and modelling”, *Nature Reviews Methods Primers*, vol. 1, 2020. url: https://api.semanticscholar.org/CorpusID:268753577.

[19] X. Wang ja Z. Cheng, ”Cross-Sectional Studies: Strengths, Weaknesses, and Recommendations.”, *Chest*, vol. 158 1S, S65–S71, 2020. url: https://api.semanticscholar.org/CorpusID:220520566.

Liite A Liitedokumentti

Tähän tulee liitteeksi dokumentaatio R koodista, joka on julkaistu käyttöön.

Liitteen ohjelmakoodi 1 kuvaa matemaattisen monadirakenteen pohjalta rakentuvan Haskellin tyyppiluokan. Tyyppiluokan voi nähdä eräänlaisena abstraktina ohjelmointirajapintana (API), joka muodostaa ohjelmoijalle abstraktin ohjelmointikielen käyttöliittymän (UI).

Ohjelmalistaus 1 Tyyppiluokka 'Monad'.

```
{haskell}
class Monad m where
    ( >>= )         :: m a -> (a -> m b) -> m b
    return          :: a                 -> m a

    fail            :: String            -> m a
    (>>)            :: m a -> m b        -> m b
    m >> k          =  m >>= \_ -> k       -- default

instance Monad IO where  ...               -- omitted
```

Ensimmäisen liitteen toinen sivu. Ohjelmalistaus 2 demonstroi vielä monadin käyttöä.

Ohjelmalistaus 2 Monadin käyttöä.

```
{haskell}
main =
return "Your name:" >>=
putStr >>=
\_ -> getLine >>=
\n -> putStrLn ("Hey " ++ n)
```