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, 14 s., 2 liites.
Labra
Joulukuu 2024

Here be abstract

Asiasanat: tähän, lista, avainsanoista

Sisällys

1 Introduction

1.1 Baysian Mindset

1.2 Perseiving risk

2 Consepts

This chapter will provide undestanding of the Bayesian approach as a whole statistical orientation. This chapter will lead the reader trough Bayes' theorem to Baysian inference and data-alytics. Calculating PAR or population attributable risk is the main focus of the Pirikahu&al. paper. The meaning of PAR and the definition Pirikahu has chosen for it will be also explored.

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 propability[1] or posterior distribution.[2] $P(A|B)$ is the propability 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 propability[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 propability thus using the Bayes' theorem alone is not enough to make Bayesian inteference. Bayesian inteference 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 Bayesian Inference

Inference means summary or characterization[3]. Statistical inference's goal is to make predictions about an unobserved set of data y based on already observed set of data x.[4][5] The Baesian 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:

$$P(y|x) = P(x_{n+1}|\{x_1, x_2, ..., x_n\}) \tag{2.2}$$

When the amount of observation x_i starts to grow it becomes increasingly more dificult 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.[4] This way of expressing uncertainty via a distribution, leans on the consept on exchangeability. Exchangeability is a basic consept of statistical analysis. [5] There is no operational difference between θ which describes belief and $P(x|\theta)$ which is a mesurable quantity. They both describe unceratainty. $P(x|\theta)$ is an updated version of the prior. The Bayes' theorem becomes:

$$P(x|\theta) = \frac{P(\theta|x)P(\theta)}{P(x)} \tag{2.3}$$

[6]

2.2.1 Prior

Prior distribution is the knowledge we have of the subject matter before we take into account the data.[7] Prior provides the plausibility for all parameters before the data is taken into account. [2][3] Prior distribution is the best way to summarize information and lack of it.[3]

θ can be chosen based on some previous inference obtained by Bayesian approach. Even if there is no prior distribution available a prior has to 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.[3] 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.[4] 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.[3]

Different types of priors:

- Maximum entropy prior
- Parametric approximations
- Empirical
- Hierarchical
- Conjugate priors
- Laplace's prior
- Invariant priors
- The Jeffrey's prior

- Reference priors
- Matching priors

[3]

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 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.

2.2.2 Likelihood

Likelihood or likelihood function has meaning outside of the Bayesian approach. Here I will give a description of likelihood in the context of Bayesian inference. Every time when likelihood is mentioned description is the version of likelihood that is meant.

Likelihood function contains the information that x_i brings to the inference. [3] 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:

$$l_1(\theta|x_1) = cl_2(\theta|x_2) \tag{2.4}$$

[3]. 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. [3]. 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 represent the information about θ coming from the data.The data affects the posterior trough the likelihood so the Bayes rule can be simplified to *posterior distribution* $\propto$ *likelihood* $*$ *prior distribution* where:

$$P(\theta|x) \propto cP(x|\theta)P(\theta) \tag{2.5}$$

In 2.4 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 affect on the posterior θ. By normalizing the right side of 2.5 the constant will be canceled out.[7]

2.2.3 Posterior

The simplest definition for a posterior is the probability p conditional on the data.[2] This is what we know of the distribution θ given the knowledge of the data.[7]

2.2.4 Sampling

Monte-Carlo-Marcov

2.2.5 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.

[5]

2.2.6 Confidence intervals

2.3 PAR, Population attributable risk

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

$$PAR = P(D+) - P(D+|E+) \tag{2.6}$$

, 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 happened from the probability distribution of all disease occurrences. In conclusion PAR is $P(D+|E-)$. From this way of representing the PAR we can see why the Bayesian approach could give the best result when calculating this value. [8]

2.3.1 Risk

2.3.2 Hazard

2.4 R language

2.4.1 Stan

2.5 Cross-sectional study

3 The Bayesin Model for Assessing Population Attributable Risk (PAR)

This chapter will give and overview of the *Bayesian methods of confidence interval construction for the population attributable risk from cross-sectional studies* -paper by Sarah Pirikahu, Geoffrey Jones, Martin L. Hazeltona and Cord Heuerb.

$$E = mc^2 \tag{3.1}$$

4 R code

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

5 Expanding the model for multivariate analysis

5.1 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] 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.

[4] 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.

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

[6] 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.

[7] 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.

[8] 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`.

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)
```