Advanced Methods in Epidemiology

Practical Session on Bayesian inference

Prof. dr. Steven Abrams

Master of Epidemiology | Academic year: 2022-2023

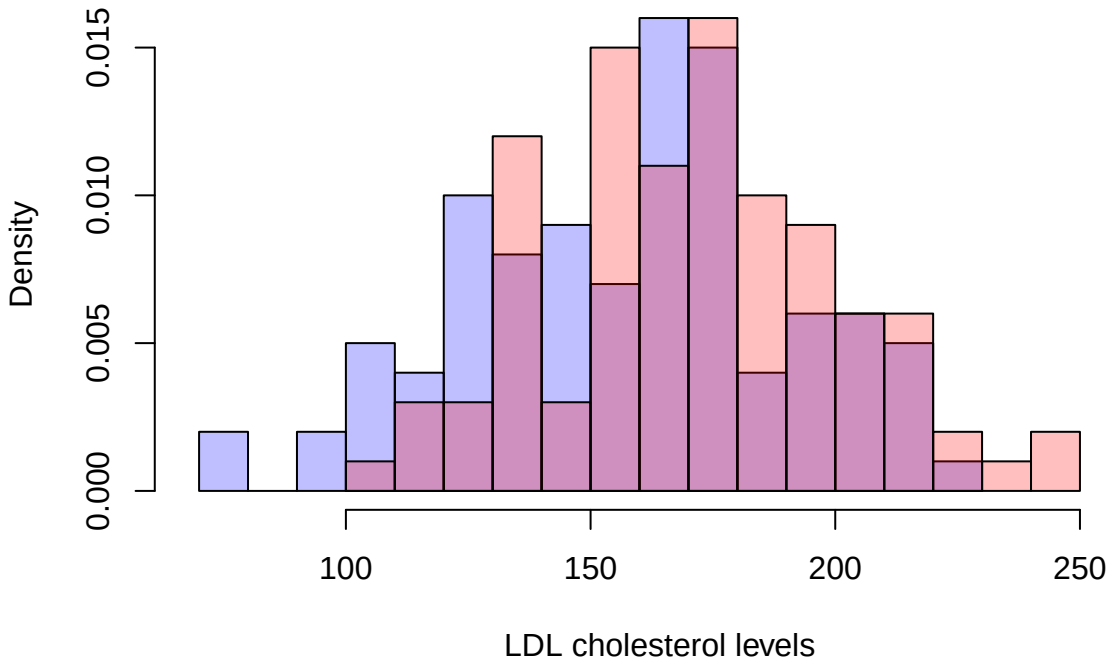
Background

The Bayesian paradigm offers an elegant framework to probabilistic statements about specific biological, clinical or other real-life processes. More specifically, Bayesian inference provides an alternative to frequentist statistics and key advantages over frequentist inference are the ability to incorporate prior knowledge with regard to the process under consideration directly into the analysis, to estimate missing values along with unknown (model) parameters (often referred to as data augmentation), and to provide direct probability statements concerning relevant statistical hypotheses. Unlike in frequentist inference, Bayesian inference does not rely on replication or asymptotics nor starting from the perspective that unknown parameters are fixed constants, but rather assume that parameters are random variables with a distribution to be estimated from data. Bayesian inference is inspired by the so-called Bayes' Theorem, proposed by the English statistician, philisopher and priest Thomas Bayes and published in 1763. Hence, the foundations of Bayesian inference were laid in the 18th century.

Key concepts in Bayesian inference

In general, a statistical data analysis confines attention to hypothesis testing or parameter estimation thereby accommodating uncertainty or variability in terms of the process that you want to characterize or describe. Uncertainty refers to the fact that typically only a random (finite) sample from the study population is available for analysis, hence, natural or instrinsic variability with regard to individual measurements of the variables of interest induce uncertainty. Next to that, measurement error, systematic errors in the experiment or data collection, and missing data induce additional variation. Uncertainty needs to be understood and quantified within a statistical analysis in order to judge the quality of our results. If there is a low level of uncertainty, results are more precise and we are more confident about our analysis. On the ohter hand, if uncertainty is high, results will be imprecise.

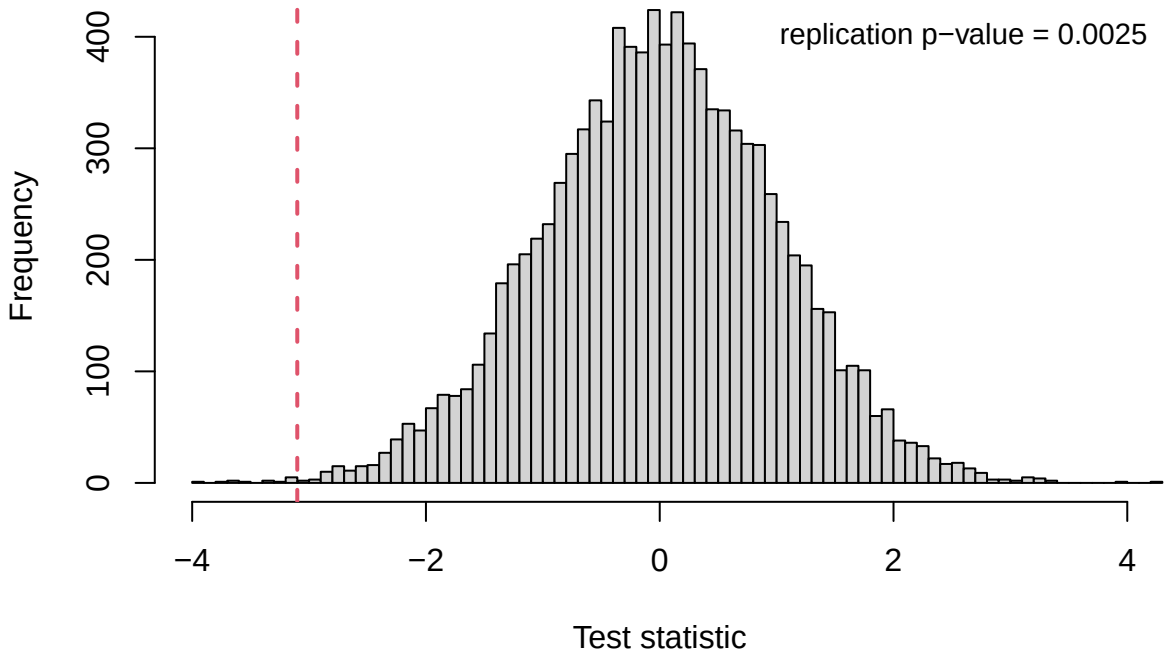
Statistical hypothesis testing is used to test specific hypotheses with regard to certain population parameters in the face of uncertainty. Assume, for example, that we want to compare an active drug to lower LDL cholesterol levels with a placebo in a drug development trial for patients with high cholesterol levels. Therefore, individuals are randomized into an active drug group and in a placebo group according to a 1:1 random assignment. Consequently, one can test whether the active drug produces a significant change in mean LDL cholesterol level (two-sided alternative hypothesis) compared to mean in the placebo group. Observed differences in mean LDL cholesterol levels between the groups can be induced by natural sampling variation only, hence, hypothesis testing compares the extent of the observed difference in means to the amount of observed uncertainty.



A frequentist approach to perform statistical hypothesis testing for this problem is to consider, for example, a parametric two-sample t-test under the assumptions of normality of the LDL cholesterol levels in both the placebo and active drug groups. Moreover, equality of variances is assumed for this test.

```
##
##  Two Sample t-test
##
## data:  y_drug and y_placebo
## t = -3.1009, df = 198, p-value = 0.002211
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22.905144  -5.097011
## sample estimates:
## mean of x mean of y
##  157.3559  171.3570
```

Frequentist approaches are based on asymptotics. Essentially this means that the interpretation of test results is based on repeating an experiment or study a large number of times (for the null hypothesis being true). For the previous example, we therefore repeat the sampling of 100 individuals in each of the groups 10,000 times under the null hypothesis and consequently calculate the so-called replication p-value.



Hypothesis testing and the use of p-values are subject to a lot of criticism given that they provide a poor way of quantifying uncertainty. More specifically, p-values have a distribution, meaning that p-values also come with uncertainty. Moreover, using p-values to construct a decision rule to classify effects as either significant or insignificant based on a single arbitrary threshold value (typically a significance level of 5%) have led to a lot of debate abouts its use in statistics.

As a response to this criticism, one has argues that parameter estimation is superior to formal hypothesis testing. More specifically, the difference between the active treatment and placebo group in terms of the mean LDL cholesterol level together with its corresponding 95% confidence interval could be considered for this purpose.

```
## mean of x mean of y
##  157.3559  171.3570
```

```
## [1] -22.905144  -5.097011
## attr(,"conf.level")
## [1] 0.95
```

The 95% confidence interval is used to quantify uncertainty, but they are often misinterpreted as a 95% probability that the true population parameter falls within the specific interval. However, the interpretation of a 95% confidence interval is different in the sense that if you perform the experiment an infinite number of times and calculate the 95% confidence interval for each of the experiments separately, 95% of the constructed confidence intervals will contain the true parameter. Interpretation based on a single experiment simply reduces to the true parameter being either within or outside the constructed 95% confidence interval.

Bayesian inference provides an alternative to frequentist inference, thereby giving a more natural quantification of uncertainty. Essentially, the Bayesian paradigm aims to represent beliefs with regard to the underlying probability distributions for different parameters, updating prior belief using observed data. More specifically, in contrast to frequentist inference in which parameters are assumed to have a specific (constant) value, Bayesian methods start from the assumption that unknown parameters are random variables. The less certain we are a priori, the wider our (posterior) probability distribution or vice versa. The prior belief is translated into a so-called prior distribution, which is updated in view of the data according to a likelihood function, and subsequent (Bayesian) inference is based on a posterior probability distribution capturing the combination of both prior information and data.

The theoretical foundation is based on Bayes' theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}, \tag{1}$$

where $p(\theta|y)$ is the posterior probability distribution, $p(y|\theta)$ is the data likelihood and $p(\theta)$ represents the prior distribution. The denominator is a normalizing constant which is independent of θ. Sampling from the posterior distribution is typically done using a Markov Chain Monte Carlo sampling algorithm, at least when no closed form expression for the posterior distribution is available.

Introduction to Stan and its R interface RStan

In this tutorial, we use (R)Stan to perform Bayesian inference. In general, Stan is a C// library forr Bayesian inference and modeling that relies on the No-U-Turn sampler (NUTS) proposed by Hoffman and Gelman (2012). More specifically, the NUTS algorithm is an extended version of Hamiltonian Monte Carlo to obtain posterior samples from a user-specified model and data.

The package *rstan* provides an R interface to Stan, referred to as RStan, and allows to fit different Stan model directly from within R (R Core Team, 2021). Moreover, the R package enables you to access the MCMC output, including options to perform posterior inferences and an assessment of the convergence of MCMC chains, log posterior densities and its gradients. The later mainly relates to the fact that HMC, and NUTS, are gradient-based methods to perform MCMC sampling. A general introduction to the functionality of *rstan* is including in the vignette which is available on https://cran.r-project.org/web/packages/rstan/vignettes/rstan.html

In this practical session, we will consider a Bayesian implementation of a linear regression model applied to an oncological clinical trial dataset which is described below. Here, we will make use of the so-called *rstanarm* package which provides a *glmer*-style interface to Stan. Next to this package or direct RStan implementations, several other R packages are available to perform MCMC sampling (e.g., shinystan, laplacesdemon, etc.). The latter approach allows you to perform Bayesian inference for user-specific models after specification of prior, likelihood function and corresponding posterior distribution (without normalization constant).

Oncological data application

Tumors or neoplasms are abnormal masses of tissue that can be solid or fluid-filled and do not necessarily pose a health threat. Malignant tumors are cancerous and can grow and spread. In a multi-center randomized controlled trial (RCT), researchers are interested in studying the effect of a new oral cancer treatment (involving a newly identified compound with good performance in mice) on a continuous tumor size measurement as compared to the standard treatment. Lung cancer patients with malignant solid tumors (carcinoma formed from epithelial cells) are randomized into two treatment groups (standard and new treatment), balancing groups according to gender, and combined chemotherapy treatment. On top of that, age at lung cancer diagnosis is observed for the patients. Informed consent was obtained for all recruited individuals.

The data obtained from the RCT can be found in the file 'oncology_ss.txt'. The endpoint of interest in this trial is "Rel_reduction", defined as the relative tumor size reduction (x 100) between the start of the study (week 0) and the end of the study (week 24).

The following variables were reported on the study participants:

- "Subject_index": ID for the participant
- "Treatment": Type of treatment (1 = standard, 2 = new)
- "Gender": Gender of the individual (1 = male, 2 = female)
- "Chemo": Combined chemotherapy (1 = yes, 2 = no)
- "Age": Age at lung cancer diagnosis
- "Center": Identifier for the health center
- "Time": Time of tumor size measurement (measurement at week ((Time-1) x 2))

Please consider the following questions:

1. Read in the 'oncology_ss.txt' dataset in R and explore the data to familiarize yourself with variables and corresponding distributions.

```
##   Subject_index Treatment Gender Chemo      Age Center Rel_reduction
## 1           254         1      1     1 65.06413      1     37.857143
## 2          6024         1      1     1 71.91413      2     13.768116
## 3          8238         1      1     1 39.64061      3     42.142857
## 4          3298         1      1     1 38.76399      4     13.475177
## 5          1479         1      1     1 39.07769      5     53.900709
## 6          6859         1      1     1 50.64677      6      2.919708
```

2. Regress the relative tumor size reduction against treatment, gender, chemotherapy and age (use the *lm* function in R to perform linear regression). Interpret the corresponding results.

3. Extend the previous model to check for effect modification of the treatment effect by gender (include an interaction effect between treatment and gender in the model). Interpret these results.

4. Consider now a Bayesian linear regression approach. In order to do so, we use the function *stan_glm* from the *rstanarm* package. The most important arguments are listed below:

- formula: the formula specification is similar to the formulation for the *glm* function in R.
- data: dataset to be used for the analysis.
- family: as a default a Gaussian distribution is specified for the endpoint.
- prior: prior distribution specification for the regression coefficients. The default specification is a normal prior is used. However, alternative prior distributions can be specified as well (see *rstanarm* documentation). If we want to use a flat uniform prior (non-informative improper prior), we set this argument to NULL.
- prior_intercept: prior for the intercept - this can be set to normal, student_t, or Cauchy for the respective distributions. Again, if we want to specify a flat uniform prior (improper prior) we set this argument to NULL.
- prior_aux: prior fo auxiliary parameters such as the error standard deviation for the Gaussion family.
- algorithm: this argument specifies the MCMC algorithm used to produce chains of values from the posterior distributions for the model parameters. The default algorithm is sampling for Markov Chain Monte Carlo (MCMC).
- iter : the number of iterations for the MCMC method, default value is 2000 iterations.
- chains : the number of MCMC chains that is constructed, default value is 4.
- warmup : also known as the burnin, i.e., the number of iterations used for adaptation. The burn-in should not be used for inference and is therefore discarded.The default implies that half of the specified number of iterations is considered as burn-in.

Fit a Bayesian linear regression model (including treatment, gender, chemotherapy and age as covariates) with one single chain of 10,000 iterations and a burn-in of 5,000 iterations. Use flat prior distributions for all model parameters and rely on the default MCMC algorithm (NUTS - see above).

```
library(rstanarm)
bayes_lm_fit0 <- stan_glm(Rel_reduction ~ factor(Treatment) + factor(Gender) +
                            factor(Chemo) + Age, data = oncology_dat,
                          family = gaussian(link = "identity"), iter = 10000,
                          prior = NULL, prior_intercept = NULL, warmup = 5000,
                          chains = 1)
```

5. Fit now a Bayesian linear regression model with one single chain of 10,000 iterations and a burn-in of 5,000 iterations. Use the default settings for the specification of the prior distributions and rely on the default MCMC algorithm (NUTS - see above). **Assess the impact of changes in prior distribution specification in terms of the posterior distibutions.**

6. Obtain the MCMC chain and construct a trace plot for the treatment effect plotting the iteration number on the x-axis and the sampled values on the y-axis.

7. In order to visualize the posterior distribution for, for example, the age effect, we use the R package *bayesplot*. More specifically, use the function *mcmc_dens* to produce a plot for the posterior distribution of the age effect.

8. Another important aspect to check prior to performing posterior inference, is checking the convergence of the MCMC chain(s). Using the *summary*-function, you get several diagnostic metrics that can be used. Alternatively, you can rely on the *coda* package in R to perform formal checks of convergence for different chains.

9. Posterior inference can be performed using the R function *describe_posterior* within the *bayestestR* package. Download the package and use the aforementioned function.

The following elements are part of the summary output:

- CI: the Credible Interval used to quantify the uncertainty about the regression coefficients (**NOTE: in statistical literature, one typically uses CrI as abbreviation for credible interval in order to avoid confusion with the confidence interval defined above and used in the frequentist context**). There are two different methods to compute a credible interval, namely the highest density interval (HDI) which is the default, and the Equal-tailed Interval (ETI).
- pd: probability of direction, defined as the probability that the effect goes to the positive or to the negative direction, and it is considered as the best equivalent for the frequentist p-value.
- ROPE_CI: Region of Practical Equivalence. Given that Bayesian inference deals with true probabilities, it does not make sense to estimate the probability of the effect being exactly equal to zero (i.e., under the null hypothesis) given that for a continuous probability distribution such a probability equals zero. Thus, we define a small range around zero which is considered practically the same as no effect. This range is referred to as the ROPE. The Rope is considered to be [-0.1,0.1] as a default (Cohen, 1988).
- Rhat: scale reduction factor $\hat{R}$ computed for each scalar quantity of interest. The scale reduction factor represents the standard deviation of the quantity of interest from all the chains together, divided by the root mean square of the separate within-chain standard deviations. When this value is close to 1 we do not have any convergence problem with MCMC. **NOTE: calculation of $\hat{R}$ based on a single MCMC chain is meaningless.**
- ESS: Effective Sample Size capturing how many independent draws contain the same amount of information as the dependent sample obtained by the MCMC algorithm, the higher the ESS the better. The threshold used in practice is 400.

10. Change the default setting for the prior distributions (of the different effects). More specifically, change the prior distribution for the treatment effect based on an earlier (pilot) study showing that the treatment parameter can be characterized by a normal distribution with mean 13 and standard deviation 2. Study the impact on the posterior distributions for the model parameters.

11. Fit now a Bayesian linear regression model with 4 chains of 10,000 iterations with a burn-in of 5,000 iterations. Use the previous settings for the specification of the prior distributions (including the informative prior for the treatment effect) and use the default MCMC algorithm.

12. Next to the normal distribution, other distributions such as the Student t-distribution can be considered. Explore the use of alternative prior distributions. Also visualize the differences.

13. In the frequentist *lm* exercise, we consider a model with an interaction effect between treatment and gender. Extend the Bayesian linear regression approach to include treatment effect modification by gender.

14. What about clustering of observations within health centers?

References

- Cohen, J. (1988). Statistical power analysis for the behavioural sciences.
- Hoffman, Matthew D., and Andrew Gelman. 2012. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research.
- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.