Course summary: Introduction to Bayesian Statistics

Jasper Ginn

March 2019

1 Introduction (week 1)

A Bayesian statistical analysis consists of several steps.

1. **Research question & data preview**. This is the most important step. We think about *what* it is we want to know and *how* we could represent this using data.
2. **Density of the data**. Here, we construct the likelihood of the data given our problem statement.
3. **Prior distribution for the parameters of the density of the data**. Any likelihood function has parameters. The prior distributions allow us to represent our knowledge of these parameters.
4. **Data collection**.
5. **Posterior distribution**. We construct the posterior distribution using Bayes' rule and sample it repeatedly.
6. **Inference for transformation of the parameters**. Using the samples we drew from the posterior, we can transform the parameters in a lot of ways to obtain new and interesting statistics.
7. **The posterior predictive distribution of the data**. We can use the posterior distribution to repeatedly create predictions, which gives us detailed information about our problem.

1.1 Bayes' rule

Bayesian statistics heavily rely on **Bayes' rule**, which is given by:

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \tag{1.1}$$

Where:

- $f(\theta|x)$ is the posterior density.
- $f(x|\theta)$ is the likelihood of the data.

- $f(\theta)$ is the prior distribution of the parameter θ.
- $f(x)$ is a normalizing constant to ensure that the posterior distribution is a 'proper' probability distribution.

1.2 Constructing the density of the data (likelihood)

The density of the data, or *likelihood* of the data, is the **probability/likelihood** of the data *given* the parameter θ. The likelihood of the data depends on the problem under consideration. The likelihood of the data is defined for *every data point*. That is, for one data point, we have

$$f(x_i|\theta)$$

By virtue of independence, we can take the product of the individual likelihoods. So, for n data points, we have

$$f(\vec{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

1.3 Prior distributions

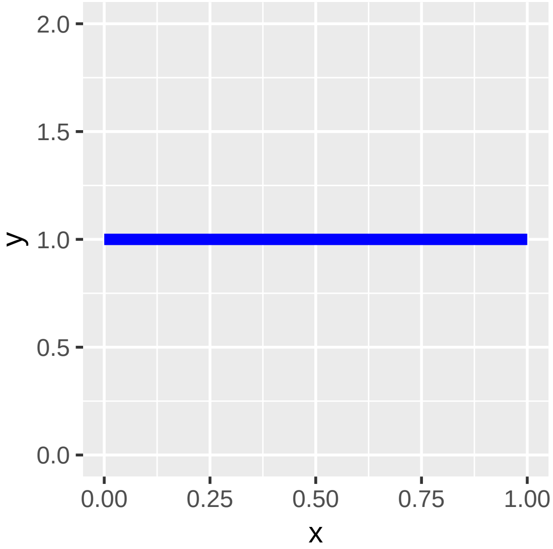
The parameter θ is not treated as a fixed but unknown parameter as we would do in frequentist statistics. Rather, we approach it as a random variable. The 'knowledge' that we have about our parameter is summarized in the *prior* $f(\theta)$. In equation 1.1, we observe that the prior is 'combined' with the likelihood of the data to form the posterior. Hence, we can see that the posterior is a mix between the evidence from observed data and the knowledge we add to the model by means of a prior.

Because of this 'mixing', it would be reasonable to expect that, if we knew a lot about the problem at hand and we could specify this in the prior, the posterior would be heavily affected by this knowledge. Conversely, if we knew nothing about the distribution of our parameters, the posterior would be affected primarily by the evidence we collected in the form of the data.

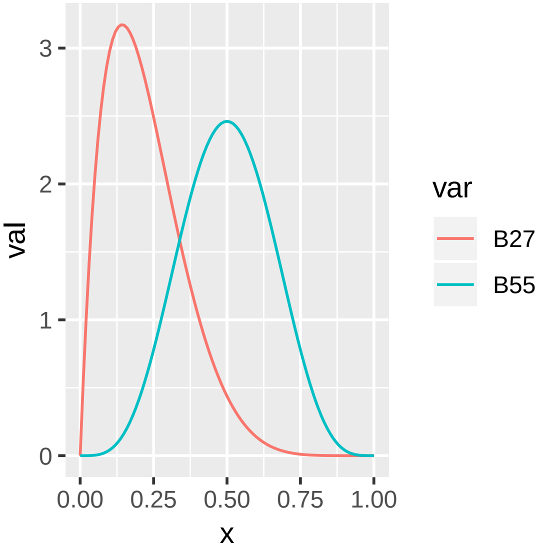
A prior can be:

1. **Uninformative**. This means that we do not know much about the form of θ, which is reflected in the prior distribution. Suppose that the prior distribution for θ is a beta distribution $\theta \sim B(\alpha, \beta)$.

Then, we could represent our lack of knowledge by using $B(1, 1)$, which evaluates to 1 for every value of $x \in (0, 1)$



2. **Informative**. This means that we have some idea about the distribution of θ prior to the analysis. We can reflect the strength of our *belief* in our knowledge in the way we construct the prior distribution. That is, in our beta distribution above, we can choose different parameters for α and β that will change the way the beta distribution looks.



In general, high dispersion (variance) indicates more uncertainty about the prior distribution whereas low dispersion indicates more certainty about our prior distribution.

A prior distribution is called *conjugate* if it has the same mathematical form as the density of the data. For example, if we have a binomial likelihood and a beta prior, then we get:

$$\begin{aligned} f(\theta|x) &\propto f(x|\theta)f(\theta) \\ &\propto \left[\prod_{i=1}^{n}\binom{n}{x}\theta^{x}(1-\theta)^{n-x}\right]\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\sum x}(1-\theta)^{\sum n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{(\sum x_i)+\alpha-1}(1-\theta)^{(\sum n-x_i)+\beta-1} \end{aligned} \tag{1.3.1}$$

The resulting posterior is mathematically similar to the prior.

1.4 Constructing the posterior distribution

After we define the likleihood of the data and the priors, and we have collected our data for analysis, we can construct the posterior probability density $f(\theta|x)$, which is given by

$$f(\theta|x) = f(x|\theta)f(\theta) \tag{1.4.1}$$

Using a sampling procedure, we can then repeatedly sample the posterior distribution. This (hopefully) gives us a good approximation of the posterior.

1.5 Evaluation of the results

To evaluate the results, we compute the *Maximum a Posteriori* (MAP), which is usually the mean of the posterior distribution, but can also be the median or the mode. Furthermore, we calculate the posterior standard deviation of the sampled values and the 95% credible interval. This is analogous to the confidence interval in classical statistics with a much simpler interpretation: the 95% credible interval (CI) has a 95% probability of containing the true value of interest.

1.6 Why Bayesian statistics is called 'subjective'

The use of a prior in Bayesian statistics raises the question *which* prior should be considered best. Ultimately, this is a subjective choice, and a function of

1. (Subjective) prior knowledge.
2. Attitude with respect to the use of prior knowledge on historical data.
3. Evaluation of the ‘fit’ between study subjects and application in a previous study and current study.
4. Attitude with respect to the use of prior distributions to represent this prior knowledge.

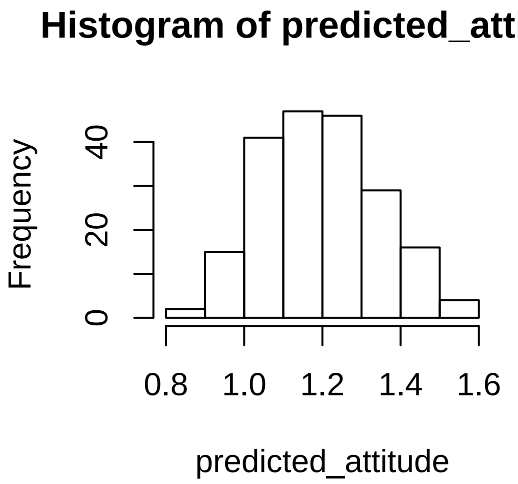
Ultimately, we should reasoned and transparent thinking to select priors of interest.

If we care more about historical results, we could choose to weight historical results more heavily, which results in *power priors*. We weight the likelihood of the historical data by some value α_0 to indicate that it should be more or less important than our believe of the distribution of the parameters. **This practice requires domain expertise**.

1.7 Transformation of posterior densities & posterior predictive distributions

Besides retrieving point estimates, we can also transform posterior distribution as we see fit after we have repeatedly sampled it to compute ratios, differences and other statistics. Such variables are treated as any other, meaning that we similarly calculate metrics like the EAP etc.

Unlike classical statistics, which focuses mainly on the computation of point estimates, bayesian statistics allows us more interesting analysis of posterior predictive results. For example, suppose that we run a linear regression on a dataset which predicts an attitude score from agreeableness and extraversion. Then a posterior predictive distribution of a person with an agreaableness score of 0.5 and an extraversion score of 2.5 would be predicted to have an attitude score that lies between 0.8 and 1.6 with the mean at an attitude score of 1.16.



Hence, posterior predictive distributions help us understand future predictions.

1.8 Practical using R and JAGS

We first need to specify the data in R. This we can do by copying it from the table

```
dat <- list(y.PE=58, n.PE=141, y.PC=40, n.PC=143)
```

Next, we specify a JAGS model. Given that we want to test the effectiveness of the treatment PE against PC, we will model the success probability θ, which indicates the probability of being cured.

Notice that JAGS doesn't require a specific order of the variables in the model file. This means that you can reference items before assignment.

We also specify an additional variable that we want to collect at each iteration called RR or the relative risk of remaining sick in the PC condition versus the PE condition:

$$RR = \frac{\theta_{PC}}{\theta_{PE}}$$

```
# Specify model
mod1 <- 'model{


  # Priors (uninformative)
  theta.PE ~ dbeta(1,1)
  theta.PC ~ dbeta(1,1)

  # Likelihood of the data
  y.PE ~ dbin(theta.PE, n.PE)
  y.PC ~ dbin(theta.PC, n.PC)

  # Contrast (RR)
  RR <- theta.PC / theta.PE

}'

# Load rjags
library(rjags)
con <- textConnection(mod1)
# Create a jags model
jmod <- jags.model(file="jags_models/ex_jmod1.txt", data=dat, n.chains=2)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 2
##    Unobserved stochastic nodes: 2
##    Total graph size: 8
##
## Initializing model
```

```
close(con)
```

At this point, we have initialized the model with uninformative beta priors. Next, we specify the burn-in period and burn 1.000 samples.

```
update(jmod, n.iter = 1000)
```

Finally, we specify the parameters of interest and run the sampler to obtain a sample of the posterior distribution.

```
params <- c("theta.PE", "theta.PC", "RR")
res <- coda.samples(jmod, variable.names = params, n.iter=10000)
```

Normally, we would first inspect convergence before viewing the results, but since we have not yet done this we will skip this step. We use the `summary()` function to inspect the results.

```
summary(res)
```

```
##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 2
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##            Mean      SD  Naive SE Time-series SE
## RR       0.6919 0.11465 0.0008107      0.0010276
## theta.PC 0.2828 0.03733 0.0002640      0.0003363
## theta.PE 0.4127 0.04040 0.0002857      0.0003484
##
## 2. Quantiles for each variable:
##
##            2.5%    25%    50%    75%  97.5%
## RR       0.4907 0.6117 0.6832 0.7641 0.9376
## theta.PC 0.2136 0.2567 0.2816 0.3075 0.3586
## theta.PE 0.3345 0.3847 0.4122 0.4397 0.4921
```

We observe the following:

- The probability of recovering under treatment PC is roughly 28%, whereas the probability of recovering under treatment PE is roughly 41%.
- The relative risk of θ_{PC}/θ_{PE} is 0.68, 95% CCI = $[0.49, 0.94]$%, meaning that the risk of remaining ill under the treatment PC is $1/0.68 \approx 1.47$ higher than the treatment PE. Since the ratio is below 1, we can say that we are 95% certain that PE treatment gives a higher chance of recovery than PC treatment.

Notice that, for this model, it is also possible to derive the solution analytically. The mean for the posterior distribution is given by

$$\hat{\theta}_i = \frac{(\alpha_i + y_i)}{(\alpha_i + y_i) + (\beta_i + n_i - y_i)}$$

Which gives us

$$\begin{aligned}\theta_{PC} &= \frac{1+40}{1+40+1+143-40} \approx 0.2828 \\ \theta_{PE} &= \frac{1+58}{1+58+1+141-58} \approx 0.4126\end{aligned}$$

The RR then becomes $0.2828/0.4126 \approx 0.6854$. These results are the same as the results we obtained using sampling.

With respect to using informative hypotheses using historical data, I would argue that both datasets fit reasonably well with our current research goal. Under the assumption that the treatments PC and PE are relatively static (don't change much), and that PTSD is a stable condition (which is likely given that the impact of war is horrible in almost any situation), we could use the results from the previous studies by pooling the data together and using these to construct the prior distributions.

In the model file below, we specify the informative distributions.

```
mod2 <- 'model{

  # Priors (informative)
  theta.PE ~ dbeta(161,191)
  theta.PC ~ dbeta(126,281)
  # Likelihood of the data
  y.PE ~ dbin(theta.PE, n.PE)
  y.PC ~ dbin(theta.PC, n.PC)

  # Contrast (RR)
  RR <- theta.PC / theta.PE

}'

# Load rjags
con <- textConnection(mod2)
# Create a jags model
jmod <- jags.model(file=con, data=dat, n.chains=2)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 2
##    Unobserved stochastic nodes: 2
##    Total graph size: 11
##
## Initializing model
```

```
close(con)
# Run the model
update(jmod, n.iter = 1000)
params <- c("theta.PE", "theta.PC", "RR")
res <- coda.samples(jmod, variable.names = params, n.iter=10000)
summary(res)
```

```
##
```

```
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 2
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##            Mean      SD  Naive SE Time-series SE
## RR       0.6815 0.05618 0.0003972      0.0005144
## theta.PC 0.3020 0.01946 0.0001376      0.0001797
## theta.PE 0.4443 0.02239 0.0001583      0.0001995
##
## 2. Quantiles for each variable:
##
##            2.5%    25%    50%    75%  97.5%
## RR       0.5781 0.6425 0.6793 0.7181 0.7972
## theta.PC 0.2647 0.2887 0.3018 0.3150 0.3411
## theta.PE 0.4005 0.4293 0.4444 0.4591 0.4885
```

The posterior mean for RR is 0.68 with 95% CCI = $[0.58, 0.8]$. The CCIs are smaller and still do not include 1. Hence, the result of including prior knowledge has quite a big effect.