Introduction to BDM for health scientists

Training Material

Eleftherios Meletis, Konstantinos Pateras, Julio Alvarez, Matt Denwood, Polychronis Kostoulas

25-27 October 2022

Bayes theorem - Predictive Values / Exercise

A rapid test has been developed to detect if a person is infected with the new SARS-CoV-2 virus. This test is fairly reliable:

- 95% of all infected individuals are detected and,
- 95% of all healthy individuals are identified as such.
- Also, it has been documented that at most one passenger out of 100 aboard on an airplane is infected.

1. What is the Sensitivity and Specificity of the test and the prevalence of the population?
2. Estimate the probability of of a person being infected given that he/she tested positive?
3. Estimate the probability of a person being healthy, given that he/she tested negative?

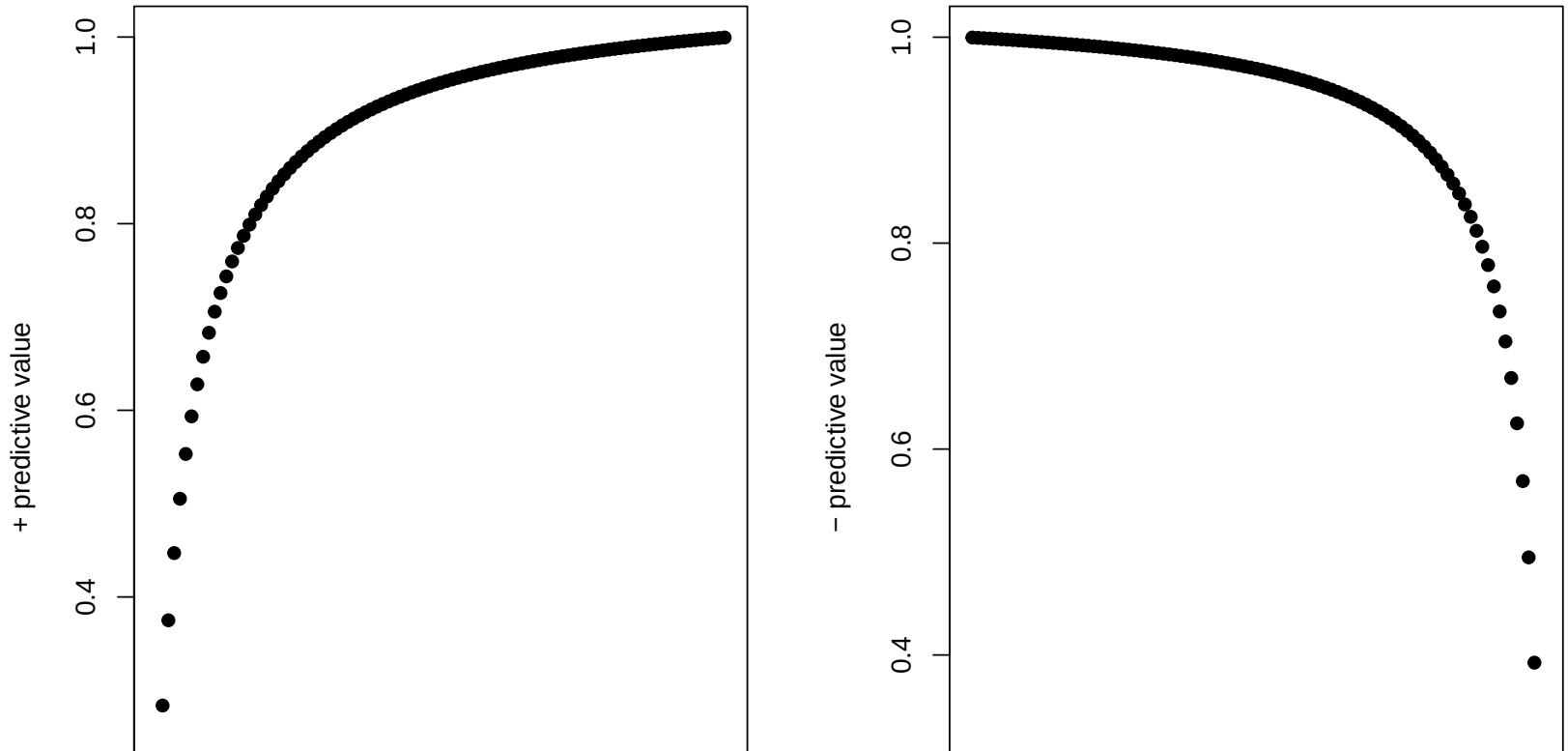
Alterinative case - What if:

- 60% of all infected individuals are detected and,
- 99% of all healthy individuals are identified as such, the Sp of the test is 100%
- The prevalence can be 100% or 0%.

1. Estimate now the predictive values of the test.
2. What are the predictive values when we have a “perfect” test?

R - package to estimate PVs

```
library(BioProbability)
p<-seq(0.01,0.99, by=0.01)
predictive.value(p,Spe=0.95,Sen=0.97,plot.it=TRUE)

## Computation of the predictive values (+ and -) from the
```



Coffee break

Introduction - Summary of pre-course work

Did everyone manage to install the necessary software (R, Rstudio, JAGS) and related R-packages?

Topics

During this school the following topics will be covered:

1. Running basic models in JAGS
2. Apparent & true prevalence estimation
3. How to choose a prior distribution?
4. Diagnostic test evaluation with Hui-Walter models

Main questions

What is JAGS and what is MCMC?

How are the apparent and true prevalence defined and why are they different?

What is a prior distribution?

How can we estimate the Sensitivity and Specificity of a diagnostic test?

- Tomorrow Matthew Denwood - Action's Vice Chair - will provide an (extremely brief) introduction to Markov Chain Monte Carlo (MCMC)

Probability distributions

- Likelihood theory is at the heart of most inferential statistics

A likelihood is the probability of observing our data given the distribution that we use to describe the data generating process.

Example

- What is the likelihood (i.e. probability) of getting 5 heads from 10 tosses of a fair coin?

We assume:

- The probability distribution describing a number of independent coin tosses is called the Binomial distribution
- In this case, we would use the parameters:
  - Number of coin tosses $= 10$
  - Probability of a head $=$ 'fair' $= 0.5$

Example - Rstudio

```
tosses <- 10
probability <- 0.5
heads <- 5
likelihood_1 <- choose(tosses, heads) * probability^heads *
                (1-probability)^(tosses-heads)
likelihood_1
```

```
## [1] 0.2460938
```

But R makes our life easier by implementing this using a function called dbinom:

```
likelihood_2 <- dbinom(heads, tosses, probability)
likelihood_2
```

```
## [1] 0.2460938
```

Maximising a likelihood

In the previous example we assumed that we knew the probability of getting a head because the coin was fair (i.e. probability of head $= 50\%$), but typically we would want to estimate this parameter based on the data.

One way to do this is via Maximum Likelihood.

Example (continued)

Let's say that we have observed 7 test positive results from 10 individuals and we want to estimate the prevalence by maximum likelihood.
We could do that by defining a function that takes our parameter as an argument, then calculates the likelihood of the data based on this parameter:

```
likelihood_fun <- function(prevalence)
  dbinom(7, 10, prevalence)
```

We can now ask the function what the likelihood is for any parameter value that we choose, for example:

```
likelihood_fun(0.8)
```

```
## [1] 0.2013266
```

```
likelihood_fun(0.5)
```

Example (continued)

We could keep doing this for lots of different parameter values until we find the highest likelihood, but it is faster and more robust to use an R function called optimise to do this for us:

```
optimise(likelihood_fun, interval=c(0, 1), maximum=TRUE)
```

```
## $maximum
## [1] 0.6999843
##
## $objective
## [1] 0.2668279
```

This tells us that the maximum likelihood for this data is 0.267, which corresponds to a parameter value of around 0.7 (or a prevalence of 70%). This is the maximum likelihood.

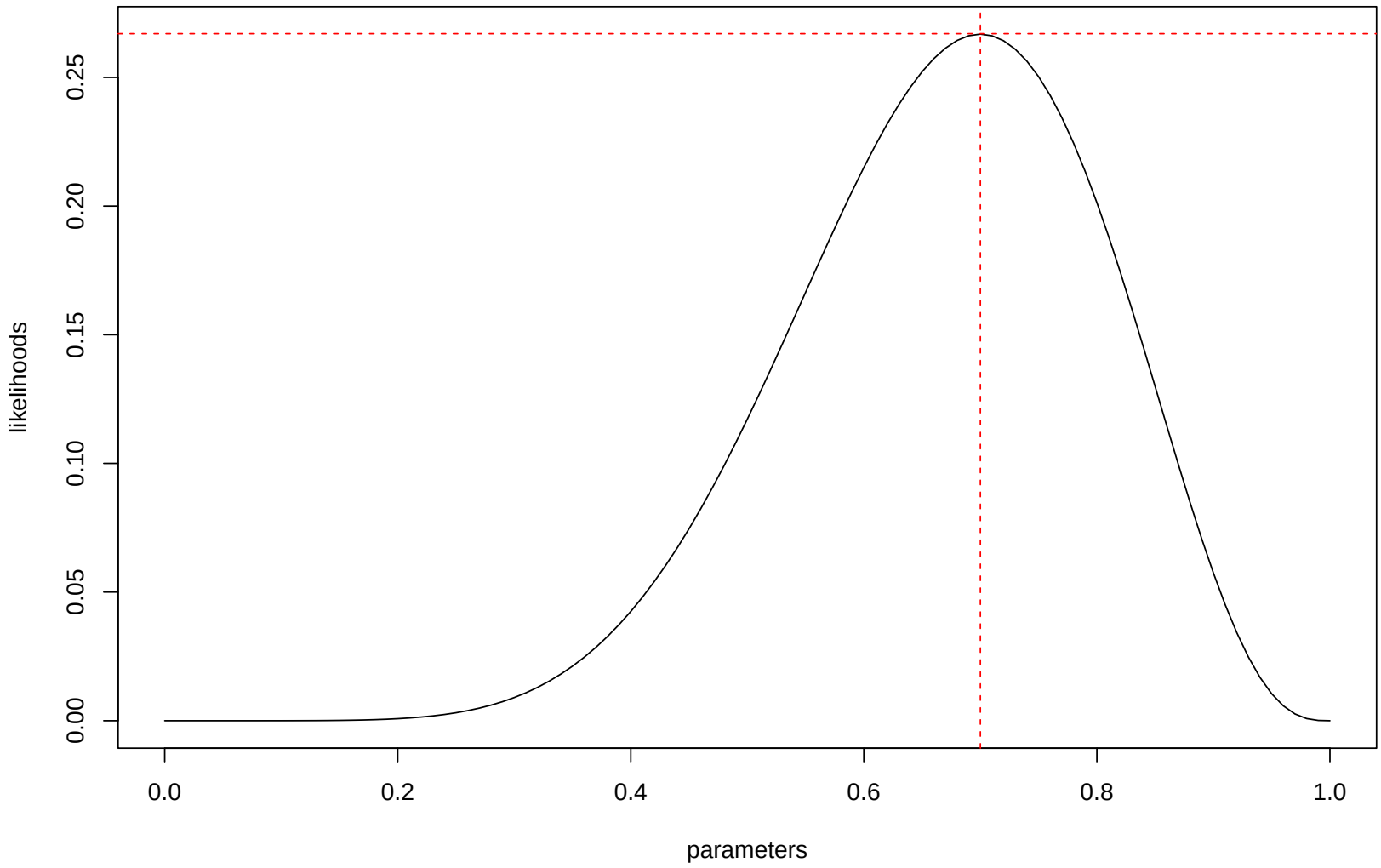
Profiling a likelihood

The parameters corresponding to the maximum likelihood give the highest probability of observing the data given the parameters, but there are other parameter values under which we could observe the data with almost as high a probability.

It is useful to look at the range of parameter values that are consistent with the data, which is why R reports standard errors (and/or confidence intervals) when you run a model.

But we can also look at the full distribution of the likelihood of the data over a range of parameter values using our function above.

Example - R Session



The red dashed lines show the maximum likelihood (y axis) with corresponding parameter value (x axis), and the solid line is the likelihood of the data given the parameter value on the x axis.

Coffee break

Bayesian Statistics

In this session we'll see how we can estimate a probability of interest but in a Bayesian framework, i.e. using Bayes theorem.

Bayes' theorem

Bayes' rule

Describes the probability of an event based on prior knowledge

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{1}$$

Components

- P(A|B): Prob of event A occurring given that B is true - Posterior probability
- P(B|A): Prob of event B occurring given that A is true - Likelihood $\sim$ function of A
- P(A): Prob of event A occurring - Prior probability
- P(B): Prob of event B occurring - Marginal probability $\sim$ sum over all possible values of A

What we usually see/use

Bayes' rule

θ: parameter of interest | y: observed data

$$P(\theta|y) = \frac{P(y|\theta) * P(\theta)}{P(y)} \rightarrow P(\theta|y) \propto P(y|\theta) * P(\theta) \qquad (2)$$

Where:

- P(θ): Prior probability of parameter(s) of interest;
- P(y/θ): Likelihood of the data given the parameters value(s)
- P(θ|y): Posterior probability of parameter(s) of interest given the data and the prior

Bayesian Inference - Summary & Example

To estimate the posterior distribution $\mathrm{P}(\theta|\mathrm{y})$ we need to:

- *Specify* the **Prior distribution**: $\mathrm{P}(\theta)$
- *Define* the **Likelihood** of the data: $\mathrm{P}(\mathrm{y}|\theta)$

Example: Bayesian apparent prevalence (ap) estimation

y out of n individuals test positive. Estimate the apparent prevalence.

Parameter of interest: apϵ[0,1]

Data: n tested, y positive

- Prior distribution for ap: ap $\sim$ Beta(a,b)
- Likelihood: y $\sim$ Binomial(n,ap)

Let's write our first JAGS model

```
ap_model <-
'model {

  # Define likelihood distribution of the data
  # JAGS Binomial distribution Arguments: p, n

  y ~ dbin(ap,n)

  # Specify prior distribution for par of interest
  # Uniform (non-informative) prior distribution
  ap ~ dbeta(1,1)

  #data# n, y
  #monitor# ap
  #inits# ap
}
'
```

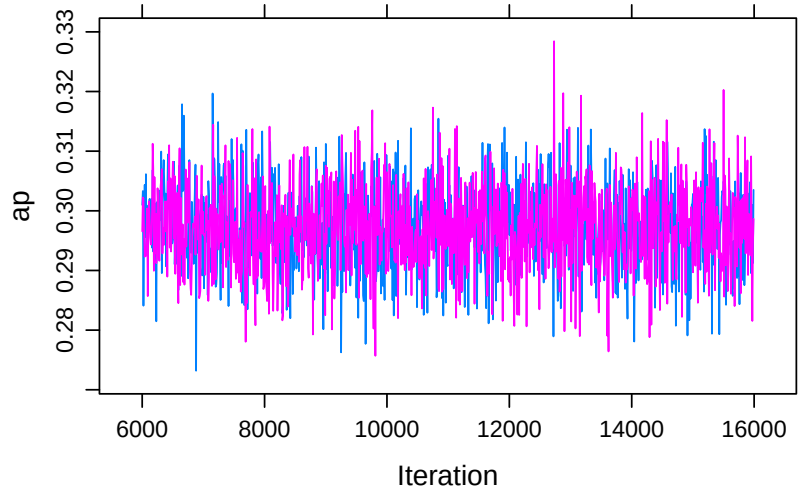
Let's run our first JAGS model

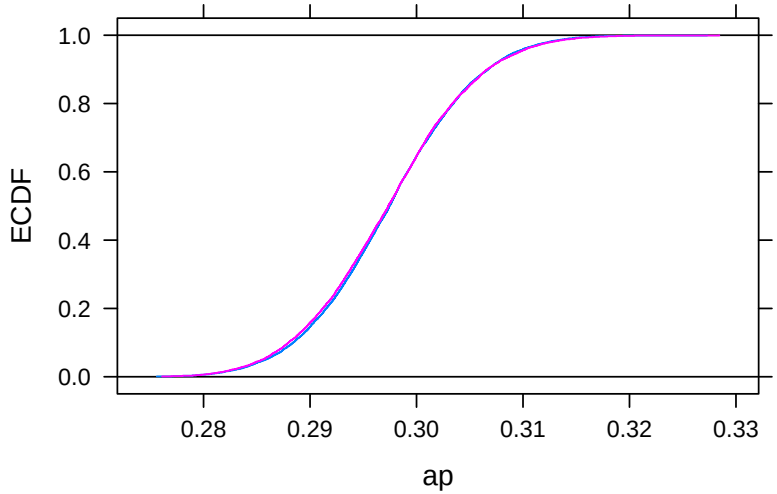
```
# Call JAGS
library(runjags)

# Provide Data
n = 4072
y = 1210

# Initial values for par of interest
ap <- list(chain1=0.05, chain2=0.95)
```

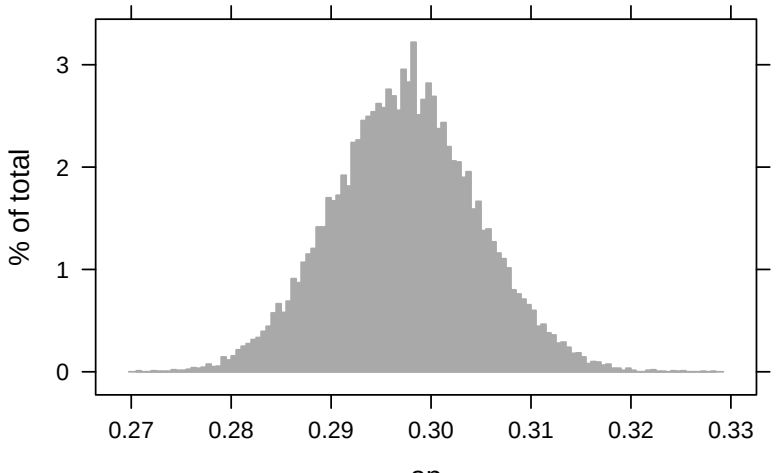
```
# Run the model
results <- run.jags(ap_model, n.chains=2,
                    burnin=5000, sample=10000)
```

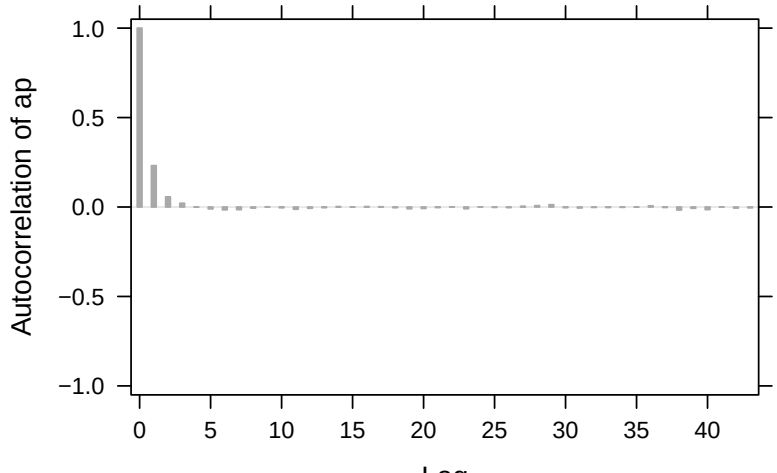
View results

```
# Plot results
plot(results)
```









Summary of Day 1

Any questions?

What we'll see tomorrow