Binomial model 01

Josep Fortiana 2018-10-17

This session is devoted to three case studies illustrating how a Bayesian perspective enhances the statistical treatment of real-world problems. The first two examples are adapted from the web course by Charles J. Dimaggio. The third one is a classical problem from the book by Andrew Gelman and others, *Bayesian Data Analysis - Third edition* **(BDA3).**

Case study 1: Perchlorate and Thyroid Tumors

Data for this example come from an online presentation by David Dunson at the US National Institute for Environmental Health Sciences.

Percholorate is a ground water contaminant that has been associated with thyroid tumors. In a study, 2 of 30 lab rats exposed to perchlorate developed thyroid tumors, compared to 0 of 30 control rats. What conclusions can be drawn? Given the small numbers, there are difficulties in the classical framework, for instance, because of the zero value a confidence interval collapses to infinity. What can we learn from a Bayesian approach?

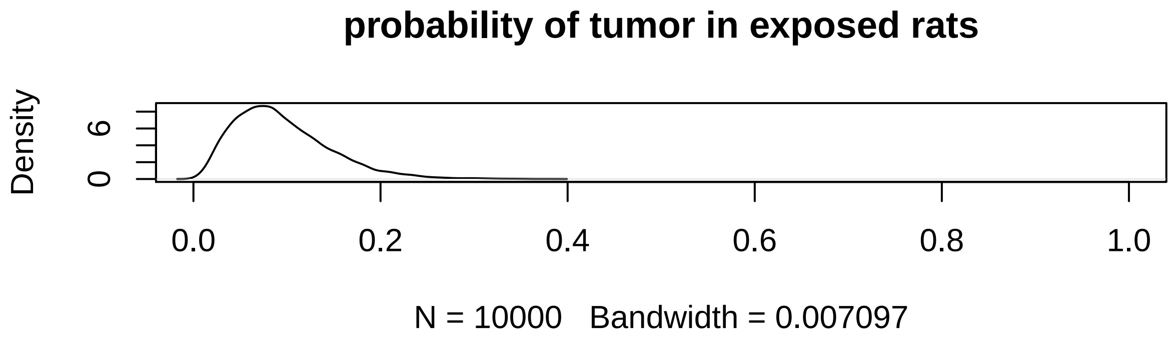
Clearly for each group, **Exposed** and **Control** rats, the likelihood (for the observed variable, the number of rats developing a tumor) is Binomial and we want to learn about possible different properties (i.e. probability distribution) of the probability parameter in both groups.

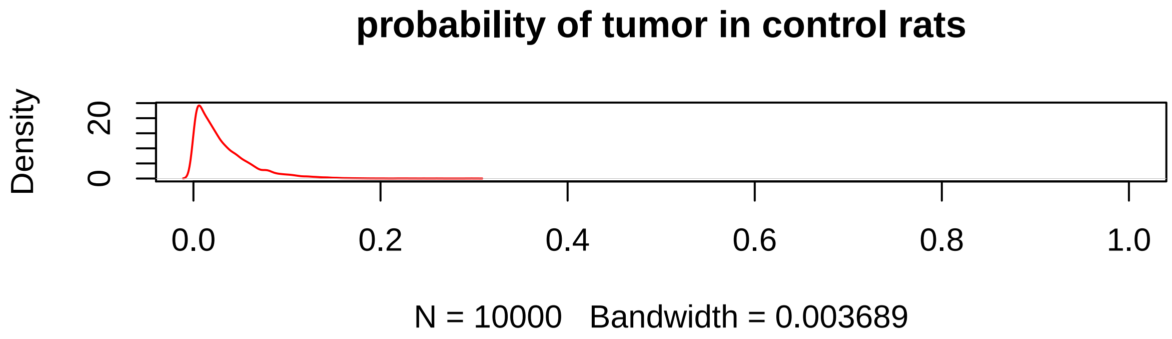
In a first run we take a $\mathrm{Unif}(0,1) = \mathrm{Beta}(1,1)$ prior as a NIP.

```
# data
y.perchlorate <- 2; n.perchlorate <- 30
y.control <- 0; n.control<- 30

# update Beta(1,1) prior for exposed and unexposed
N<-10000  # sample size for simulations
theta.perchlorate <- rbeta(N,y.perchlorate+1,n.perchlorate-y.perchlorate+1)
theta.control <- rbeta(N,y.control+1,n.control-y.control+1)

# graphically compare exposed and unexposed
old.par<-par(mfrow=c(2,1))
plot(density(theta.perchlorate), xlim=c(0,1), main="probability of tumor in exposed rats")
plot(density(theta.control),  col="red", xlim=c(0,1), main="probability of tumor in control rats")
```





```
par(old.par)

# probability that exposed have more tumors than unexposed
sum(theta.perchlorate >= theta.control)/N
```

```
## [1] 0.8868
```

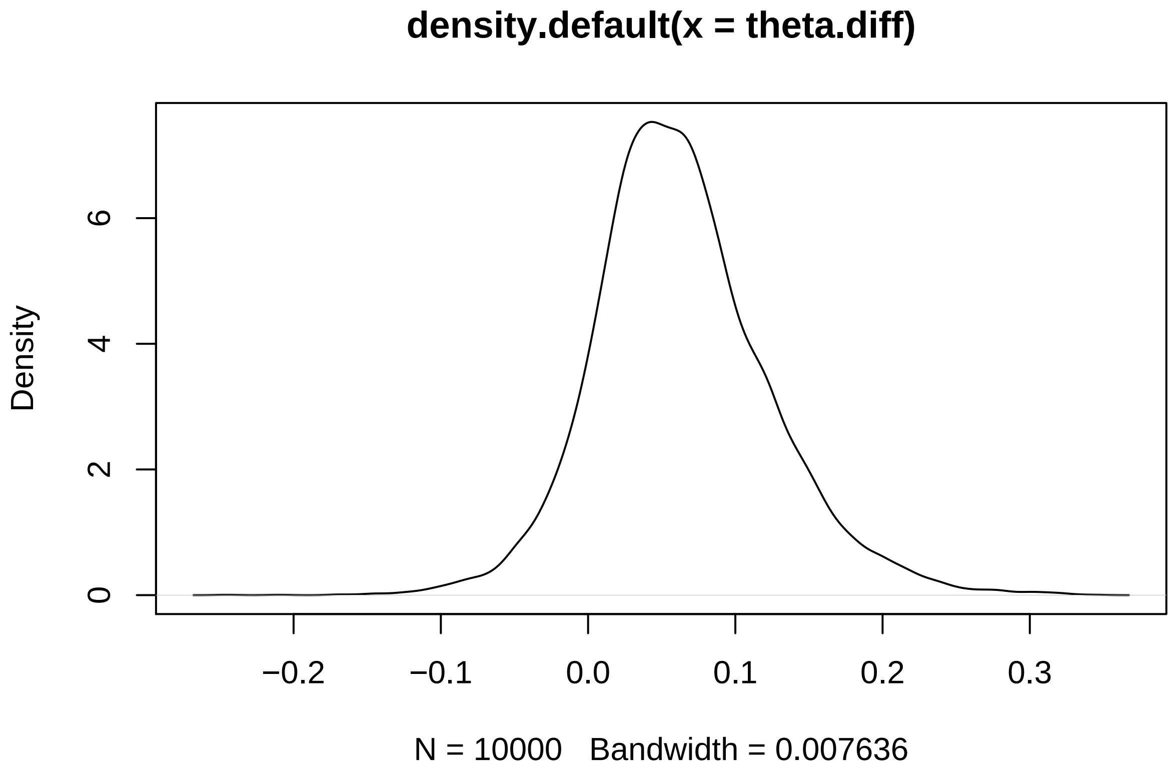
A 90% credible interval for the difference of probabilities and an estimation of the pdf of this difference. The interval includes zero, so it is consistent with the non-significance of the difference (from Fisher's exact test, see reference). But the majority of the probability lies above zero.

```
theta.diff<-theta.perchlorate-theta.control

# 90% credible interval
quantile(theta.diff, probs=c(0.05,0.95))
```

```
##          5%         95%
## -0.02396838  0.16766289
```

```
# plot differences
plot(density(theta.diff))
```



If we have some real prior evidence or belief about the putative association of perchlorate and thyroid tumors. It would be reasonable to assume that we do. Why else would we be doing this experiment in the first place? Dr. Dunson, in his presentation, suggests a $\mathrm{Beta}(0.11, 2.6)$ as representing previous experimental evidence. Let's use that prior, and combine it with our data to update the posterior.

```
theta.perchlorate <- rbeta(N,y.perchlorate+0.11,n.perchlorate-y.perchlorate+2.6)
theta.control <- rbeta(N,y.control+.11,n.control-y.control+2.6)
theta.diff<-theta.perchlorate-theta.control
quantile(theta.diff, probs=c(0.05,0.95))
```

```
##          5%        95%
## 0.00797132 0.14497070
```

Case study 2: Drug response

We are interested in the response of patients to a drug. We believe a proportion somewhere between 0.2 and 0.6 of patients will respond, with a mean of 0.4 and a standard deviation of 0.1. In a study of 20 persons, what is the probability that 15 or more will respond? Again, the problem involves a Bernoulli (or Binomial) likelihood.

We will use a Beta prior (this is the easy path, see below more involved possibilities). To translate the prior information $\mu = 0.4$, $\sigma = 0.1$ into parameters of a $\mathrm{Beta}(\alpha, \beta)$, we must isolate α, β from the formulas for the moments of a Beta distribution:

$$\mu = \frac{\alpha}{\alpha+\beta},$$

$$\sigma^2 = \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}.$$

The result is:

$$\alpha = \mu \cdot \left[\frac{\mu\,(1-\mu)}{\sigma^2} - 1\right],$$

$$\beta = (1-\mu) \cdot \left[\frac{\mu\,(1-\mu)}{\sigma^2} - 1\right].$$

In our case: $\alpha = 9.2$, $\beta = 13.8$.

```
# For a Beta(a,b) distribution, the mean and variance are
# mu<-a/(a+b)
# sigma2<-ab/((a+b)^2*(a+b+1))
#
# Solving a,b in terms of mu, sigma2
#
mu<-0.4
sigma<-0.1
sigma2<-sigma^2
G<-(mu*(1-mu))/sigma2-1
a<-mu*G
b<-(1-mu)*G
#
a; b
```

```
## [1] 9.2
```

```
## [1] 13.8
```

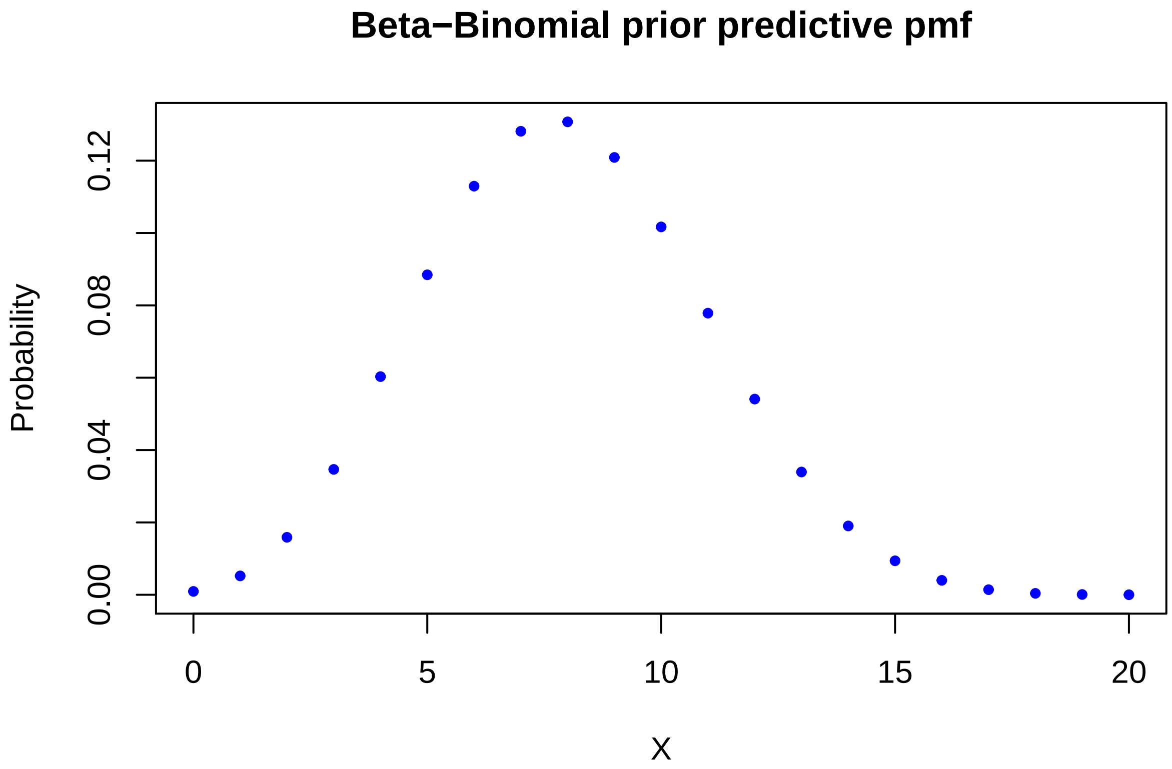
The question we are asked, the probability that 15 or more patients will respond, is previous to any experimental evidence, hence we will get the answer from the **prior predictive distribution** of the number X of patients responding to treatment out of the $n = 20$ patients participating in the study, using only prior information.

From last session we know that this r.v. follows a Beta-Binomial$(20, 9.2, 13.8)$ distribution. Thus, in principle we could use this distribution, which can be found in the `extraDistr` package. Thus we can plot the pmf and compute the exact probability we are asked:

```
#install.package("extraDistr")
require(extraDistr)
```

```
## Loading required package: extraDistr
```

```
X<-0:20
plot(X,dbbinom(X, size=20, alpha = 9.2, beta = 13.8),pch=19,cex=0.6,col="blue"
                 ,main="Beta-Binomial prior predictive pmf",ylab="Probability")
```



```
P15.exact<-1-pbbinom(14, size=20, alpha = 9.2, beta = 13.8)
round(P15.exact,3)
```

```
## [1] 0.015
```

An alternative approach, both more general in scope (only rarely will we know an explicit pdf for the prior predictive) and more according to a computational statistics mindset is to sample from this distribution by simulation and infer the desired properties from the resulting sample.

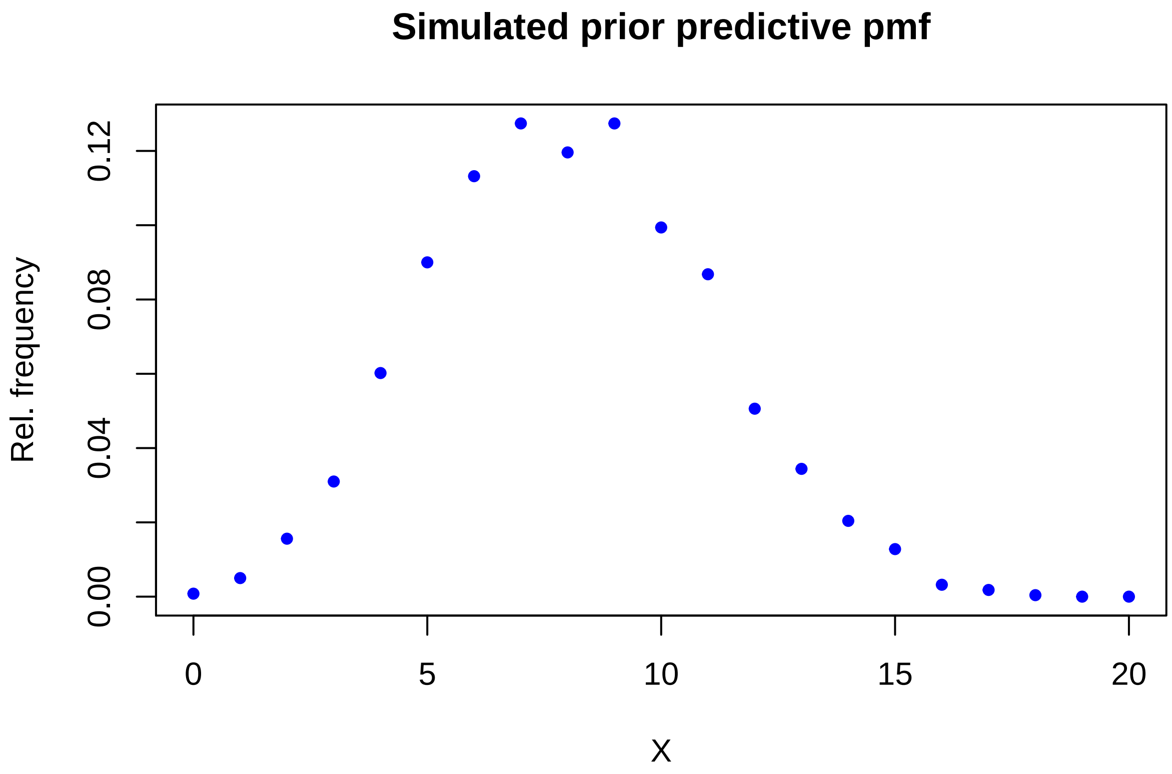
We generate a number N of realizations of the experiment. In each one of them θ is generated from a $\mathrm{Beta}(\alpha,\beta)=\mathrm{Beta}(9.2,13.8)$, then x is generated from a $\mathrm{B}(20,\theta)$. Note the way in which syntax takes advantage of the vectorized `theta` parameter in `rbinom()`. In the empirical pmf I use a clumsy piece of code as `table()` does not contemplate zero counts:

```
N<-5000
n<-20
n1<-15
theta<-rbeta(N,a,b)
x<-rbinom(N,n, theta)
accept<-(x>=n1)*1
(prob<-sum(accept)/N)
```

```
## [1] 0.0182
```

```
F<-rep(0,length(X))
for (i in X){
    F[i]<-sum(x==X[i])
    }
f<-F/N
```

```
plot(X,f,,pch=19,cex=0.7,col="blue",main="Simulated prior predictive pmf",ylab="Rel. frequency")
```



Now assume we have made an experiment with $n = 20$ patients and we observe that actually $x = 15$ of them respond to the drug treatment. The posterior θ distribution, including this evidence, is:

$$\mathrm{Beta}(\alpha + x, \beta + n - x) = \mathrm{Beta}(9.2 + 15, 13.8 + 20 - 15) = \mathrm{Beta}(24.2, 18.8),$$

which has $\mu = 0.56$, $\sigma = 0.075$, a larger mean than our prior (consistent with the large observed number) and a smaller standard deviation (reflecting that we have increased information about the experiment).

Exercise

Now consider further experimentation with this drug, after this first result. We test it with 40 new patients, being aware that funding will continue if we observe 25 or more successes in this trial. What is the probability of this result?

```
#
# Insert here your code to answer this question.
#
```

Case study 3: Placenta previa

Placenta previa is a problem in pregnancy where the placenta implants low on the uterus, potentially obstructing the birth canal. The data involve the proportion of births with placenta previa that are female gender. The question is whether placenta previa influences birth gender. We know that:

1. The proportion of female births in this population is 48.5%.
2. A study of 980 placenta previa births found that 437 (44.6%) of them were female.

How much evidence does this provide for the claim that the proportion of female births is less than 0.485?

Analysis using a nonconjugate prior distribution (a) A discrete prior

In the BDA3 book they consider conjugate (Beta) priors, just as in the above two Case Studies, which you can follow as a (highly recommended) exercise. Here, instead, we will use their next suggestion, a nonconjugate prior.

As a first stage, as a sort of springboard, we take a very simple, discrete prior. To this end we add a fictitious historical proportion (from another state, say) of 45.0%. We assume a priori that both historical figures are equally probable.

Thus, our discrete prior is:

$$\begin{array}{lll} \theta_1 = 0.485 & \text{with probability} & \pi_1 \equiv P(\theta_1) = 0.5, \\ \theta_2 = 0.45 & \text{with probability} & \pi_2 \equiv P(\theta_2) = 0.5. \end{array}$$

Likelihood is the usual binomial one:

$$f(x|\theta) = \theta^x \cdot (1-\theta)^{(n-x)}.$$

In actual computations it may be safer to use its logarithm, *Log-likelihood,* to avoid products of very large and very small quantities.

```
# Prior information
theta.1<-0.485
theta.2<-0.45
Prior.1<-0.5
Prior.2<-0.5
LogPrior.1<-log(Prior.1)
LogPrior.2<-log(Prior.2)
# Observed data
n<-980
x<-437
theta.obs<-x/n
round(theta.obs,3)
```

```
## [1] 0.446
```

Likelihood, joint pdf and their logarithms. Bayes formula needs the x marginal in the denominator to give the posterior pdf, but we can circumvent this step by normalizing to one the *non-normalized* posterior pdf. Observe that, consistently with the data, the posterior probability of 0.45 is much higher than its prior value:

```
Lik<-function(x,theta,n){theta^x*(1-theta)^(n-x)}
LogLik<-function(x,theta,n){x*log(theta)+(n-x)*log(1-theta)}
LogPost.1<-LogLik(x,theta.1,n)+LogPrior.1
LogPost.2<-LogLik(x,theta.2,n)+LogPrior.2
Post.1<-exp(LogPost.1)
Post.2<-exp(LogPost.2)
S<-(Post.1+Post.2)
Post.1<-Post.1/S
Post.2<-Post.2/S
round(Post.1,3)
```

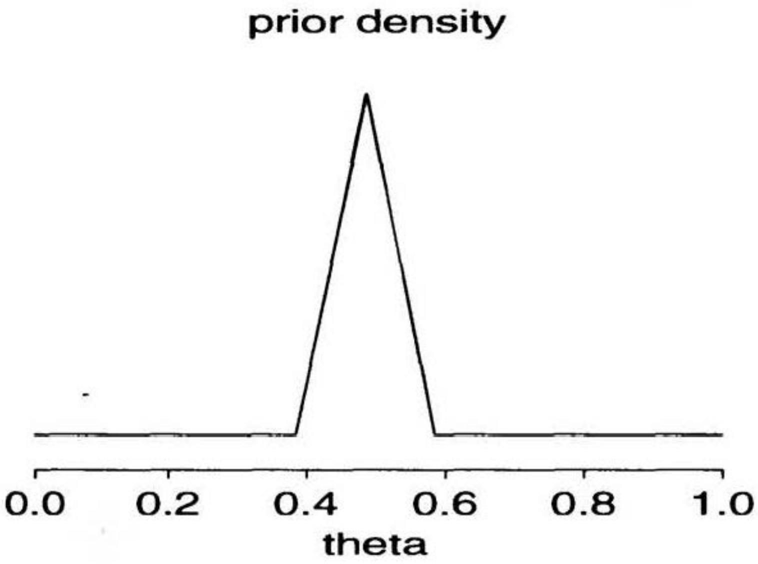


Figure 1: alt text

```
## [1] 0.049
```

```
round(Post.2,3)
```

```
## [1] 0.951
```

Analysis using a nonconjugate prior distribution (b) A triangular pdf

In the BDA3 book they consider first conjugate (Beta) priors, just the same procedure as in the two previous Case Studies, which you can follow as a (highly recommended) exercise. Here, instead, we will use their next suggestion, namely the triangular pdf in the figure:

This pdf has the following specifications:

1. It is centered at $m = 0.485$
2. 40% of the probability mass is outside the interval [m-c,m+c]=[0.385,0.585], c=0.100.

Exercise

From the information above, check that the equation of this pdf is:

$$h(t) = \begin{cases} a, & \text{for} \quad 0 < t < m - c \quad \text{and } m + c < t < 1, \\ A + b \cdot t/c, & \text{for} \quad t \in [m - c, c], \\ B - b \cdot t/c & \text{for} \quad t \in [c, m + c], \end{cases} \quad , \quad t \in [0, 1]$$

where $a = 0.50$, $b = 5.00$ is the height of the central triangle ($a + b$ is the peak ordinate), $A = a + b - (b \cdot m)/c$, $B < -a + b + (b \cdot m)/c$.

Write a function `h()`to compute the pdf $h(t)$, plot it (verify that the integral on $[0, 1]$ is indeed equal to 1), compute its expectation and variance. *Hint: a is determined by the equation* $0.40 = p_{tails} = a \cdot (1 - 2 \cdot c)$*, then the area of the upper central isosceles triangle is* $c \cdot b = 1 - a = 0.5$.

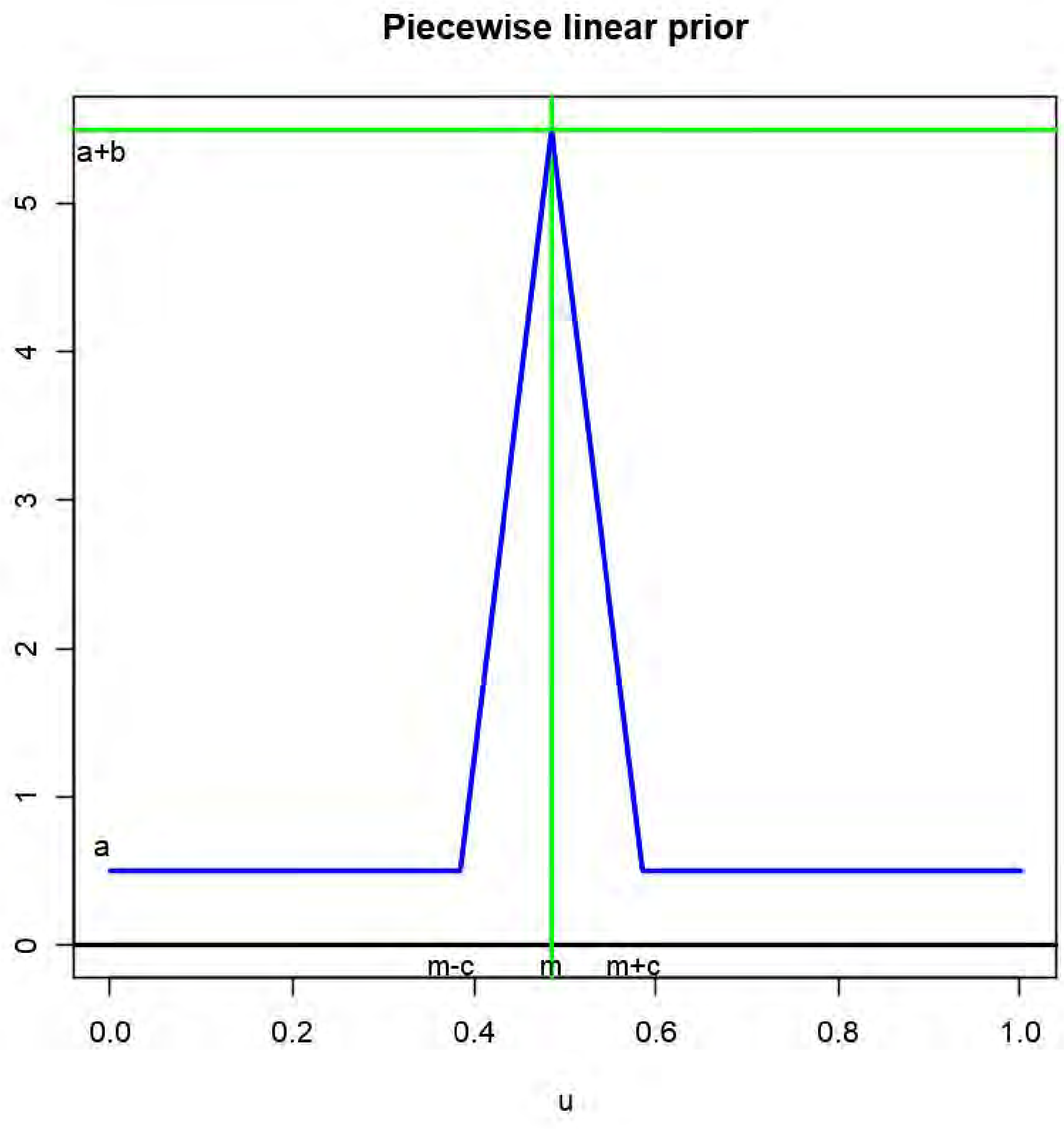


Figure 2: alt text

```
m<-0.485
c<-0.100
p.tails<-0.4
a<-0.5
b<-5.0
A<-a+b-(b*m)/c
B<-a+b+(b*m)/c
h<-function(t){
    #
    # Insert here your code.
    #
    }
```

Discretizing the continuous prior

Next we construct a grid of N points $\mathbf{t} = (t_1, \ldots, t_{N-1})$ on the $[0, 1]$. On each interval $[t_i, t_{i+1}, 1 \leq i \leq N-1$, the probability from the prior pdf, the integral of $h(t)$ on it, is approximated by the rectangle area $h(t_i) \cdot \delta$, where $\delta = 1/N$. The discrete r.v. defined by both vectors: values $\mathbf{t}$ and probabilities $h(\mathbf{t}) \cdot \delta$ is a *discretization* of the continuous r.v. with pdf h, a reasonable approximation by a discrete distribution of the given continuous probability distribution. We proceed with this discrete r.v. as in the example above.

```
n<-980
x<-437
N<-1000 # number of grid intervals
t<-seq(0,1,length=N)
t<-t[-N]# remove last point (equal to 1: we want N intervals)
#
# Lik and LogLik are the Bernoulli quantities defined above
# uncomment only one of the following three lines:
# Use plain likelihood or use logs and, in this case there is the alternative
# to use the dbinom() function (which we expect has been thoughtfully and robustly designed)
#
# nnpost<-Lik(x,t,n)*h(t)
# *********
# Uncomment one of the following two lines after defining h(t)
# *********
# log.nnpost<-LogLik(x,t,n)+log(h(t))
# log.nnpost<-dbinom(x,size=n,prob=t,log=TRUE)+log(h(t))
#
# *********
# Uncomment the following line after defining h(t)
# *********
# nnpost<-exp(log.nnpost)
#
# NB. We do not bother to write delta=1/N in the probability values h(t)*delta. Furthermore,
# dbinom() is a multiple of Lik (times the combinatorial coefficient). All these constants
# will be taken care of when normalizing.
#
# Then these non normalized values, after normalized to unit sum, will be the pmf of a discretization o
# the posterior distribution.
#
# *********
# Uncomment the following two lines after defining h(t)
# *********
```

```
#dpost<-nnpost/sum(nnpost) # pmf
#ppost<-c(0,cumsum(dpost)) # cdf
```

From this pmf/cdf for a discretization of the posterior pdf we can obtain any needed posterior quantity: popsterior quantiles, in particular the posterior median, a 95% credible interval, the MAP and posterior expectation estimators of θ, etc.

Homework

Exercise 3 - Further computations with the placenta previa exercise

(See exercises 1 and 2 in the assignment sheet)

3.1

Compute the posterior pdf, and both prior and posterior predictor pmf for the placenta previa problem, using a conjugate $\operatorname{Beta}(\alpha, \beta)$ prior, for several values of the parameters, obtaining for each of them a 95% (centered) credible interval and the MAP and posterior expectation estimators of θ. In particular, reproduce Figure 2.3 and Table 2.1 in BDA3. Do the same for those α_0 and β_0 such that $\operatorname{Beta}(\alpha_0, \beta_0)$ has the same expectation and variance than the piecewise linear nonconjugate prior. Compare the posterior quantities with both priors.

3.2

Simulate the prior predictive pmf with the nonconjugate piecewise linear pdf. Do it in both possible ways: (1) Sample from h by the inverse transformation method and (2) Sample from its discretization.

Summarize the resulting simulated sample: plot the empirical pmf, compute median, mode, expectation, standard deviation. Compare the results with those obtained from the conjugate prior $\operatorname{Beta}(\alpha_0, \beta_0)$.

3.3

Simulate the posterior predictive pmf (given the observed data of 437 female births from 980 placenta previa births) with the nonconjugate piecewise linear pdf. Sample from the the discretization of the posterior pdf as in Exercise 2.

Summarize the resulting simulated sample: plot the empirical pmf, compute median, mode, expectation, standard deviation. Compare the results with those obtained from the conjugate prior $\operatorname{Beta}(\alpha_0, \beta_0)$.