Introduction to Bayesian Statistics

Alessandra Guglielmi

Politecnico di Milano
Dipartimento di Matematica
Milano, Italia
e-mail: alessandra.guglielmi@polimi.it

September 16, 2020

Bayesian learning

We all informally use probabilities to express our **INFORMATION** and **BELIEF** about UNKNOWN quantities: the probability that tomorrow morning is rainy is 0.3 , according to my knowledge (on the average weather in September and today's weather in Milano)

Based on NEW data, we update initial belief:
What if a storm will occur in the evening? I will **UPDATE** the probability of that event on the basis of some **DATA** (rain this evening!) to the value 0.7!

Bayes' Theorem provides a rational method for updating beliefs in light of new information (DATA)

Statistical Inference

Statistical inference = Process of learning about the general characteristics of a population from a subset of members of that population

Typically, numerical values of population characteristics are represented by a parameter θ, while $\boldsymbol{y}$ is the numerical description of the sample; ex: $\theta =$ is the mean of the population quantitative variable "total colhesterol", where the population refers to Italian males over 65 years; $\boldsymbol{y}$ is a vector containing values of the colhesterol level of some individuals in the sample

Before a dataset $\boldsymbol{y}$ is observed, numerical values of both the population characteristics θ AND the dataset $\boldsymbol{y}$ are uncertain

After dataset $\boldsymbol{y}$ is obtained, the information it contains can be used to update our uncertainty about the population characteristics; of course, we aim at decreasing the uncertainty

Bayesian learning

$\mathcal{Y}$ = sample space = set of of possible datasets;
$\boldsymbol{y}$: a single dataset

Θ = parameter space = set of all possible parameters values, from which we hope to identify the value that *best* represents the *true* population characteristics

this time both $\boldsymbol{Y}$ and θ are random! Obviously if they're random we need to specify the joint distribution →

Under the Bayesian approach, there are TWO random elements $(\boldsymbol{Y}, \theta)$:

- $\pi(\theta)$ prior distribution: describes **our belief** that θ represents the *true* population characteristics ← marginal distribution of θ
- $p(\boldsymbol{y}|\theta)$ likelihood: describes **our belief** that $\boldsymbol{y}$ would be the outcome if θ is the *true* parameter value ← conditional distribution of $\boldsymbol{Y}|\theta$ (:= likelihood)

Bayes' Theorem

Once we obtain $\boldsymbol{y}$, we update our beliefs about θ computing

- $\pi(\theta|\boldsymbol{y})$ posterior distribution: describes our belief that θ is the *true* value, having observed $\boldsymbol{y}$

The posterior distribution is computed via Bayes' Theorem:

$$\pi(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta)\pi(\theta)}{\int_\Theta p(\boldsymbol{y}|\theta)\pi(\theta)d\theta}$$

Do you remind Bayes's Th for a finite partition? Reverend Thomas Bayes: 1702-1761, Theorem: 1763

R Example: Ex 1 from Hoff's book.

The Bayesian approach

It is a typical scientific approach:

- the prior belief is UPDATED via observed data and yields posterior distribution
- it suggests that scientific inference is based on 2 parts: one depends on the scientist's subjective opinion and understanding of the phenomenon under study **BEFORE an EXPERIMENT** was performed, the other depends on the observed data the scientist has obtained from the experiment.

Bayesian inference

The inference is totally based on the **posterior distribution**, i.e. the conditional distribution of the parameter θ given data $\boldsymbol{Y} = \boldsymbol{y}$, where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ are the random and observed samples, respectively.

The Bayesian approach is model-based, i.e. we assume a model, i.e. the conditional distribution (likelihood) of the data and the prior

"All models are wrong, but some are useful" (G. Box) ....

Models are not the reality but may represent it, and may be useful to predict the reality ... model-checking is very important!

Parameter estimation

Computation of *summaries* of the posterior distribution $\pi(\theta|y_1, \ldots, y_n)$; for instance

- posterior mean $E[\theta|y_1, \ldots, y_n]$
- posterior variance $\mathrm{Var}[\theta|y_1, \ldots, y_n]$
- interval estimate C : $\mathbb{P}(\theta \in C|y_1, \ldots, y_n) \geq 0.95$

Computational methods to derive these estimates, or, more generally, to simulate a sample from the posterior distribution

MCMC: Markov chain Monte Carlo methods

IDEA: if we are NOT able to derive an iid sample from the *posterior distribution*, we aims at building a Markov chain with limiting distribution given by the posterior distribution itself, and then we apply the Ergodic Theorem to approximate integrals from the posterior

Prediction of new data points

Suppose we have observed the result of throwing a coin n times : $(y_1, \ldots, y_n)$ ($y_i = 1$ if the coin was H at the i-th toss)
What is the probability that the next toss will be H?

Bayesian prediction: $\mathbb{P}(Y_{n+1} = 1 | Y_1 = y_1, \ldots, Y_n = y_n)$

Posterior predictive distribution: $\mathcal{L}(Y_{new} | \boldsymbol{Y} = \boldsymbol{y})$

Prediction

What is the probability that the next toss will be H?
Likelihood: $\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}$; prior $\pi(\theta)$ (e.g. Beta); posterior $\pi(\theta|\boldsymbol{y})$

$$\mathbb{P}(Y_{n+1} = 1 | Y_1 = y_1, \ldots, Y_n = y_n) = \int_{(0,1)} \mathbb{P}(Y_{n+1} = 1|\theta)\pi(\theta|\boldsymbol{y})d\theta$$
$$= \int_{(0,1)} \theta\pi(\theta|\boldsymbol{y})d\theta = E(\theta|\boldsymbol{y}) = \frac{\alpha + \sum y_i}{\alpha + \beta + n}$$

Ex: $n = 10$, $\sum y_i = 3$, $\alpha = \beta = 1$ (θ is uniformly distr. on (0,1)):

$$\mathbb{P}(Y_{11} = 1 | Y_1 = y_1, \ldots, Y_n = y_n) = E(\theta|\boldsymbol{y})$$
$$= \frac{4}{12} \neq \frac{1}{2} = E(\theta) = \mathbb{P}(Y_{11} = 1)$$
$$= \frac{4}{12} \neq \frac{3}{10} = \bar{y}_n$$

Bayesian vs non-Bayesian approach

- FREQUENTIST approach
  - parameters are fixed at their true but unknown value
  - "objective" notion of probability
  - good large sample properties
  - estimation: maximizing the likelihood
  - confidence intervals: difficult to interpret
  - no symmetry in testing hypotheses H_0 and H_1, difficult interpretation of p-values
- BAYESIAN approach
  - parameters are r.v.s with distributions attached to them
  - subjective notion of probability (prior) combined with data
  - does not require large sample approximations (inference is exact for any n)
  - estimation: via summary statistics of the posterior distributions; their computation via simulation based approach (MCMC)
  - credible intervals: NO problems in interpreting them
  - H_0 and H_1 are symmetric

personal believes

random variables

Bayesian Hierarchical Models

Multilevel data: results of a test for students in a population of schools in US

- test/exam scores of students in different schools or universities
- failure times of items in different batches
- patients within several hospitals
- people (or items) within provinces within regions within countries

Two levels:

- groups
- units within groups

y_{ij} is the data of the i-th unit in group j: $i = 1, \ldots, n_j$, $j = 1, \ldots,$

$(\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_J)$ $\quad \boldsymbol{Y}_j = (Y_{1,j}, \ldots, Y_{n_j,j})$

particularly useful with grouped data (= multilevel data)

Bayesian Gaussian hierarchical model

with this approach we can make predictions for a new student of a school already in the sample or we could also make a prediction for a student of a new school (not contained in the observed sample)

$$
\begin{aligned}
Y_{1,j}, \ldots, Y_{n_j,j} | \theta_j &\overset{iid}{\sim} N(\theta_j, \sigma^2) && \text{within-group model} \\
\theta_1, \ldots, \theta_J | (\mu, \tau^2) &\overset{iid}{\sim} N(\mu, \tau^2) && \text{between-group model} \\
(\mu, \tau^2) &\sim \pi
\end{aligned}
$$

population of groups/group-parameters: prediction on a student coming from a new school, selected at random from the population of groups

- group-specific parameters (each θ_j is the mean of the math score in the j-th school)
- $\theta_1, \ldots, \theta_J$ are NOT independent, since we want to share information between the groups; the dependency is *mild* (exchangeability)

Bayesian Gaussian hierarchical model

The prior is completed assuming:

$$
\begin{aligned}
\frac{1}{\sigma^2} &\sim \textit{gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right) && \sigma^2 \text{ within-group variance} \\
\frac{1}{\tau^2} &\sim \textit{gamma}\left(\frac{\eta_0}{2}, \frac{\eta_0 \tilde{\sigma}_0^2}{2}\right) && \tau^2 \text{ between-group variance} \\
\mu &\sim N(\mu_0, \gamma_0^2)
\end{aligned}
$$

R Example: `Bayesian_hierarchical_primalezione.R`

Borrowing strength

$$
E\left(\theta_j | \bar{y}_j, \mu, \tau^2, \sigma^2\right) = \underbrace{\frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}}_{\text{frequentist estimator of } \theta_j} \bar{y}_j + \underbrace{\frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}}_{\text{prior mean of } \theta_j} \mu
$$

When n_j is small, i.e. group j gives little info about θ_j, i.e. the frequentist estimate $\bar{y}_j$ is poor; however the Bayesian estimate:

$$E(\theta_j | \ldots) \approx \mu$$

The Bayesian estimate is obtained borrowing strength from the other groups (through μ)

Borrowing strength

$$
E\left(\theta_j | \bar{y}_j, \mu, \tau^2, \sigma^2\right) = \underbrace{\frac{n_j/\sigma^2}{n_j/\sigma^2 + 1/\tau^2}}_{\text{frequentist estimator of } \theta_j} \bar{y}_j + \underbrace{\frac{1/\tau^2}{n_j/\sigma^2 + 1/\tau^2}}_{\text{prior mean of } \theta_j} \mu
$$

When τ^2 is large (heterogeneous groups), the Bayesian estimate:

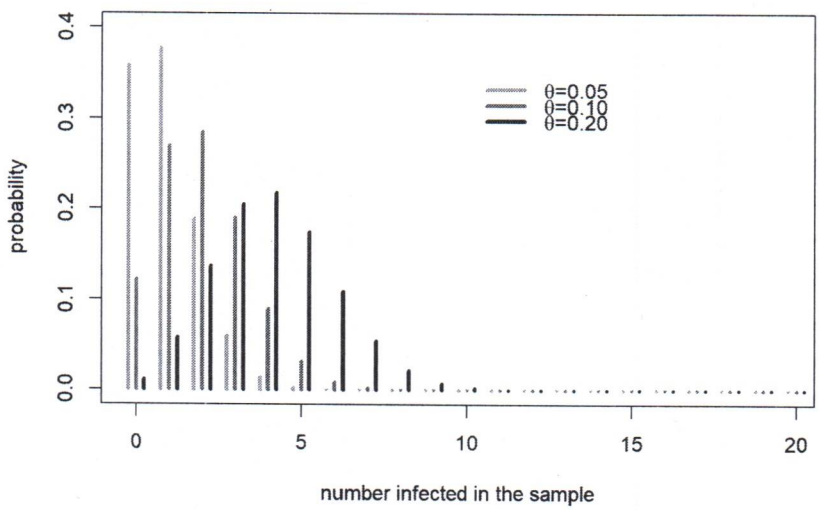
$$E(\theta_j | \ldots) \approx \bar{y}_j$$

there is less shrinkage to μ, relying more on the info in group j.

```
### ---------------------------------------------------------------------------
### Ex in Chapter 1 in Hoff(2009). A first course in Bayesian statistical methods.
### ---------------------------------------------------------------------------
### Bayesian inference on a proportion
### ---------------------------------------------------------------------------
# Suppose we are interested in the prevalence of an infectious disease theta
# in a small city. The higher the prevalence, the more public health precautious
# we would reccomend to be put into place.
# Sample: 20 individuals checked for infection
# BEFORE the sample is obtained, the number of infected individuals Y is UNKNOWN
# If theta were known, a reasonable sampling model for Y is Y|theta ~ Bin(20,theta)

### ---------------------------------------------------------------------------
### LIKELIHOOD
### ---------------------------------------------------------------------------
# What is the "true" distribution for Y ? It depends on the "true" value of theta!
# Y = number of infected individuals in the sample
n   = 20
x   = 0:n
del = .25
x11()
plot(range(x-del), c(0,.4), xlab="number infected in the sample",
     ylab="probability", type="n")

points(x-del, dbinom(x,n,.05), type="h", col=gray(.75), lwd=3)
points(x,     dbinom(x,n,.10), type="h", col=gray(.5),  lwd=3)
points(x+del, dbinom(x,n,.20), type="h", col='blue',    lwd=3)
legend(10,.35,legend=c(expression(paste(theta,"=0.05",sep="")),
                       expression(paste(theta,"=0.10",sep="")),
                       expression(paste(theta,"=0.20",sep=""))),
       lwd=c(3,3,3), col=c(gray(c(.75,.5)),'blue') ,bty="n")
```

θ = prevalence of infectious disease

Y = #infected
$Y|\theta \sim Bi(n, \theta)$



⇒ till we don't know θ we are not able to tell the distribution of Y

```
# Under these assumptions, what is the probability that there is no infected individual
# in the sample? It depends on theta!
# If theta=5%, the probability that there is no infected individual in the sample is
dbinom(0,20,0.05)
```

```
## [1] 0.3584859
```

```
dbinom(0,20,0.1)   # Se theta = 10%
```

```
## [1] 0.1215767
```

```
dbinom(0,20,0.2)   # Se theta = 20%
```

```
## [1] 0.01152922
```

```
### ---------------------------------------------------------------------------
### PRIOR DISTRIBUTION
### ---------------------------------------------------------------------------
# Other studies from various parts of the country indicate that the infection rate in
# comparable cities ranges from about 0.05 to 0.2 with an average prevalence of 0.1.
# We have to find a prior distribution on \theta consistent with this information:
# a prior pi(\theta) that assigns a substantial amount of prob to (0.05,0.2) and with
# expected value close to 0.1
# There are infinite priors consistent with these 2 conditions! But for some reasons
# (wait a couple of classes!) we consider Beta distributions.
# A priori theta is a Beta(a,b)

a = 2
b = 20
a/(a+b)  #media a priori di theta
```

```
## [1] 0.09090909
```

```
pbeta(.20,a,b) - pbeta(.05,a,b)
```

```
## [1] 0.6593258
```

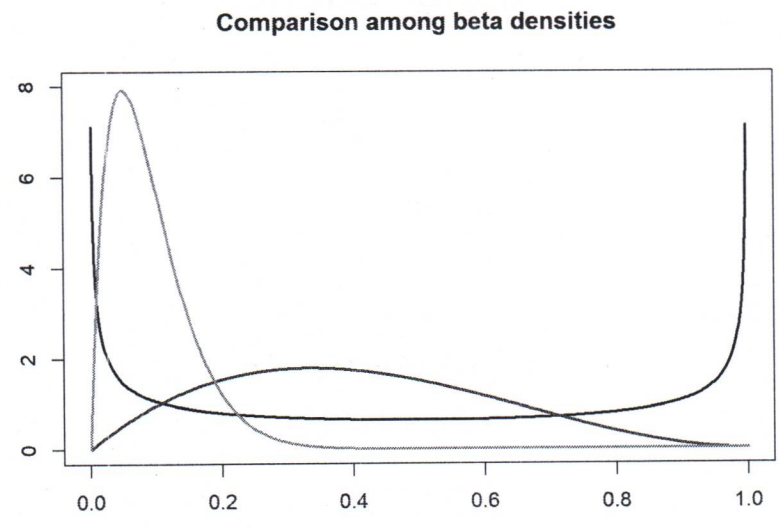
```
pbeta(.10,a,b)
```

```
## [1] 0.63527
```

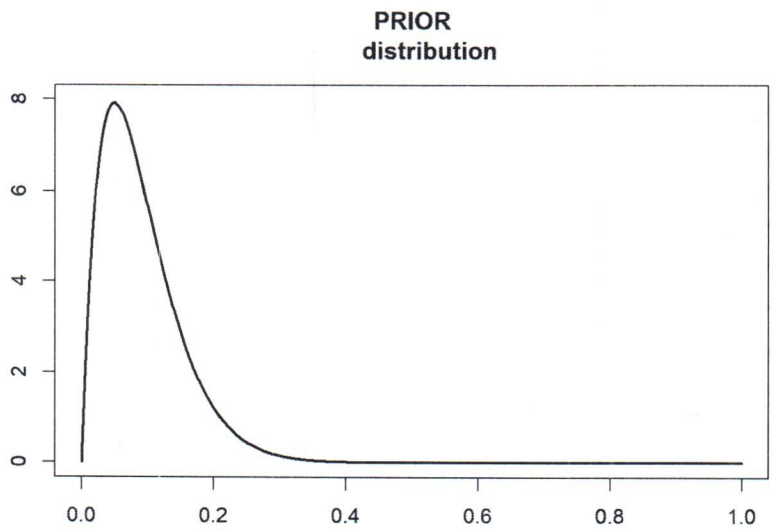
```
(a-1)/(a-1+b-1) #moda a priori di theta (se a, b>1)
```

```
## [1] 0.05
```

```
### ---------------------------------------------------------------------------
###   Comparison between Beta distributions
### ---------------------------------------------------------------------------
x11()
p=seq(0,1,length=500)
plot(p, dbeta(p,0.5,0.5), xlab=" ", ylab=" ", type="l", lwd=2, ylim=c(0,8),
     main="Comparison among beta densities")
lines(p, dbeta(p,2,3), ylim=c(0,8), col="red", lwd=2)
lines(p, dbeta(p,a,b), ylim=c(0,8), col=3, lwd=2)
```




```
X11()
### ---------------------------------------------------------------------------
### PRIOR for this EXAMPLE
### ---------------------------------------------------------------------------
p = seq(0,1,length=500)
plot(p,dbeta(p,a,b), xlab=" ", ylab=" ", type="l", ylim=c(0,8), main="PRIOR
     distribution",lwd=2)
```



← this is one possible prior consistent with the informations that we have

```
dev.off()
```

```
## png
##   2
```

```
### ---------------------------------------------------------------------------
### OBSERVED DATA
### ---------------------------------------------------------------------------
y<-0 ; n<-20 # No infected individual in the observed sample!!!

### ---------------------------------------------------------------------------
### POSTERIOR DISTRIBUTION
### ---------------------------------------------------------------------------
# a posteriori theta \sim Beta(a+y,b+n-y)
# Rappresenta la nostra "incertezza" sulla proporzione di infetti nella cittadina
# considerata, alla luce del dato y=0 (NESSUN infetto nella cittadina su 20 selezionati)
a = 2
b = 20
(a+y)/(a+b+n)        # media a posteriori
```

The Beta distribution is convenient because in the update-phase we do not change the form of the prior distribution, we simply have to update the parameters

```
## [1] 0.04761905
```

```
(a+y-1)/(a-1+b+n-1)  # moda a posteriori
```

```
## [1] 0.025
```

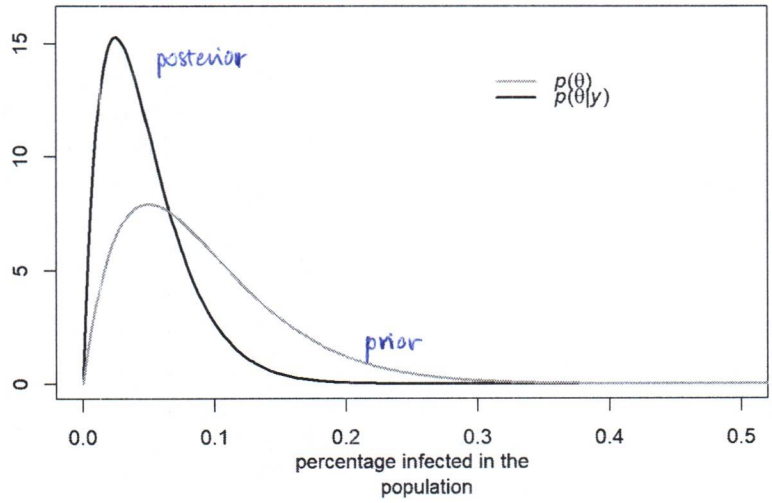
```
pbeta(.20,a+y,b+n-y) - pbeta(.05,a+y,b+n-y)
```

```
## [1] 0.3843402
```

```
pbeta(.10,a+y,b+n-y)
```

```
## [1] 0.9260956
```

```
### ---------------------------------------------------------------------------
### GRAFICO di PRIOR e POSTERIOR insieme
### ---------------------------------------------------------------------------
x11()
theta = seq(0,1,length=500)
plot(theta, dbeta(theta,a+y,b+n-y), type="l", xlab="percentage infected in the
     population", ylab="", lwd=2, ylim=c(0,16), xlim=c(0,0.5))
lines(theta, dbeta(theta,a,b),col="gray",lwd=2)
legend(.3,14,legend=c(expression(paste(italic("p"),"(",theta,")",sep="")),
                      expression(paste(italic("p"),"(",theta,"|",italic("y"),")",sep=""))),
       bty="n", lwd=c(2,2),col=c("gray","black"))
```




```
### ---------------------------------------------------------------------------
### BAYESIAN LEARNING
### ---------------------------------------------------------------------------
a/(a+b)                 # prior mean of theta
```

```
## [1] 0.09090909
```

```
(a+y)/(b+n-y)           # posterior mean of theta
```

```
## [1] 0.05
```

```
(a-1)/(a-1+b-1)         # prior mode of theta (se a, b>1)
```

```
## [1] 0.05
```

```
(a+y-1)/(a+y-1+b+n-y-1) # posterior mode
```

```
## [1] 0.025
```

```
pbeta(.20,a,b) - pbeta(.05,a,b)
```

```
## [1] 0.6593258
```

```
pbeta(.20,a+y,b+n-y) - pbeta(.05,a+y,b+n-y)
```

```
## [1] 0.3843402
```

```
pbeta(.10,a,b)
```

```
## [1] 0.63527
```

```
pbeta(.10,a+y,b+n-y)
```

```
## [1] 0.9260956
```

```r
# NOTATE che a posteriori
# E(theta|y)=(a+y)/(a+b+n)= n /(a+b+n)*(y/n) + (a+b)/(a+b+n)*(a/(a+b))

### ---------------------------------------------------------------------------
### INFERENZA CLASSICA
### ---------------------------------------------------------------------------
# STIMA puntuale di theta è y/n, ma in questo caso y=0, e quindi
# la stima puntuale di theta è uguale a 0: NON ha senso!
# STIMA INTERVALLARE si riduce ad un unico punto (0): NON ha senso!

#### Adjusted Wald interval
#    th=n/(n+4)* (y/n) + 4/(n+4)* (1/2)

a  = 2
b  = 2
th = (y+a)/(n+a+b)  # th=n/(n+4)* (y/n) + 4/(n+4)* (1/2)
th     # th corrisponde alla MEDIA a POSTERIORI di theta quando la prior è una Beta(2,2)
```

(frequentist approach)

```
## [1] 0.08333333
```

```r
th+c(-1,1)*1.96*sqrt(th*(1-th)/n) # stima intervallare
```

```
## [1] -0.03779791  0.20446458
```

```
##########################################
##    A BAYESIAN HIERARCHICAL Model     ##
##    Ex. 8.4 from P. Hoff's book       ##
##########################################
# Y_{i,j} ~ N(theta_j, sigma^2) student i in school j, i=1,...,n_j,
#                                                      j=1,...,J
# theta_j ~ N(mu,tau^2)

# Multilevel data
rm(list=ls())
Y = read.table("school.mathscore.txt")
Y
```

```
##      school mathscore
## 1         1     52.11
## 2         1     57.65
## 3         1     66.44
## 4         1     44.68
## 5         1     40.57
## 6         1     35.04
## 7         1     50.71
## 8         1     66.17
## 9         1     39.43
## 10        1     46.17
## ..
## 394      22     56.50
## 395      22     60.10
## 396      22     59.46
## 397      22     49.95
## 398      22     46.10
## 399      22     64.68
…
```

```
dim(Y)
```

```
## [1] 1993    2
```

```
head(Y)
```

```
##   school mathscore
## 1      1     52.11
## 2      1     57.65
## 3      1     66.44
## 4      1     44.68
## 5      1     40.57
## 6      1     35.04
```

```
# Scores at the math test for 1993 students in 100 US schools
# First column: index of the school of the student
# Second column: math score of the student

### ---------------------------------------------------------------------------
### Group means of the math scores
### ---------------------------------------------------------------------------
m = length(unique(Y[,1]))
n <- sv <- ybar <- rep(NA,m)
for(j in 1:m)
{
  ybar[j] <- mean(Y[Y[,1]==j,2])   # medie per gruppo (empiriche)
  sv[j]   <- var(Y[Y[,1]==j,2])    # varianze empiriche in ogni gruppo
  n[j]    <- sum(Y[,1]==j)         # numerosità in ogni gruppo (scuola)
}

ybar
```

```
##   [1] 50.81355 46.47955 48.77696 47.31632 36.58286 38.97000 40.41812 48.85000
##   [9] 49.15625 41.06833 57.94818 50.52700 49.44211 58.70778 56.43083 55.49609
##  [17] 37.92714 50.45357 43.42417 45.86250 50.70333 47.90542 51.99286 45.54600
##  [25] 44.82842 46.61300 43.44500 51.79409 46.18100 49.59250 49.21227 50.55400
##  [33] 47.38154 45.70167 55.86412 53.57250 46.08125 51.95063 46.20630 48.98440
##  [41] 56.93273 46.15071 51.39867 49.28941 45.15867 40.17619 43.92650 46.78318
##  [49] 38.16267 41.58000 64.37632 49.38200 42.54909 46.37481 43.12375 49.20750
##  [57] 43.92520 48.49250 49.75000 42.03536 45.05250 51.62077 49.65563 43.15889
##  [65] 47.87353 48.24600 65.01750 44.74684 51.86917 43.47037 46.70455 36.95000
##  [73] 53.81741 40.30941 48.85522 55.49923 44.67571 39.26308 61.70385 52.01333
##  [81] 46.88450 38.76400 44.05571 43.57042 48.88200 53.69882 57.85182 54.00810
##  [89] 53.76731 48.58667 47.84200 48.17750 52.02880 46.60750 45.01640 47.04905
##  [97] 45.75852 52.97097 51.39333 47.99167
```

```
summary(ybar)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   36.58   44.81   47.95   48.13   51.39   65.02
```

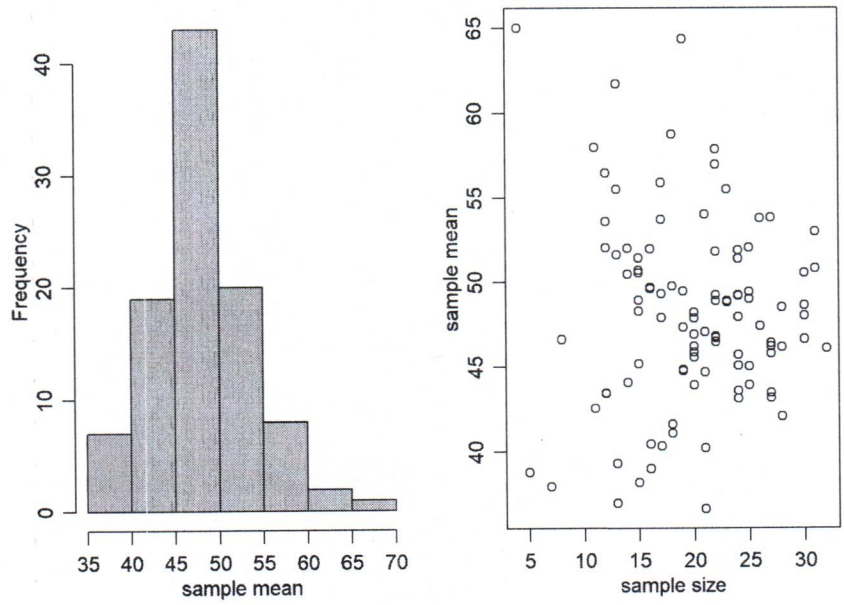
```
n
```

```
##   [1] 31 22 23 19 21 16 16 22 24 18 11 30 19 18 12 23  7 14 12 20 15 24 14 20 19
##  [26] 30 12 22 20 16 22 15 26 24 17 12 32 16 27 25 22 28 15 17 15 21 20 22 15 18
##  [51] 19 25 11 27 24 24 25 28 18 28 24 13 16 27 17 15  4 19 24 27 22 13 27 17 23
##  [76] 13 21 13 13 12 20  5 14 24 15 17 22 21 26 30 20 20 25  8 25 21 27 31 24 30
```

```
sv
```

```
##   [1] 126.56812  98.30688  58.16697 112.76908  69.51131  35.89499  36.44884
##   [8] 123.26090  57.33958  76.16034  66.14738 179.68971  74.82507  63.38191
##  [15]  74.67446  57.20942  39.31646  82.91115  94.96161  77.71746  62.49092
##  [22]  97.81670  80.36547  85.98857  86.65921  74.36839 137.08185 104.43323
##  [29]  97.55569  52.40270 115.81490  28.07778  93.76337  61.84815 122.87269
##  [36]  92.63433  74.84228  71.50259  76.25380  83.12335  81.37333  81.94130
##  [43] 144.41663  81.61739  50.05847 104.99520  60.45673  83.74836  44.25468
##  [50]  73.38935  27.27355 110.01996  91.37453 120.26890  40.07964  70.12516
##  [57] 113.62989  78.52392 161.41931 125.10754  40.30461  69.32409 129.35935
##  [64]  81.50666 157.98737  82.75414  68.09109  54.57541 121.73269  48.64489
##  [71]  94.74046  21.81058  42.58804  40.33557 108.01539  39.42647  90.67076
##  [78]  86.72092 129.69154  65.08859 100.05976 124.50308  91.21756  82.25415
##  [85]  78.25043  60.77469  77.49927  83.80724  83.70892  75.46361  90.81544
##  [92]  96.64398  99.40355  46.02965  81.30302  63.56408  84.11141 109.54818
##  [99]  51.44404  91.10613
```

```
### ---------------------------------------------------------------------------
### Grafico dei dati
### ---------------------------------------------------------------------------
x11()
par(mfrow=c(1,2), mar=c(3,3,1,1), mgp=c(1.75,.75,0))
hist(ybar, main="", xlab="sample mean")
plot(n, ybar, xlab="sample size", ylab="sample mean")
```

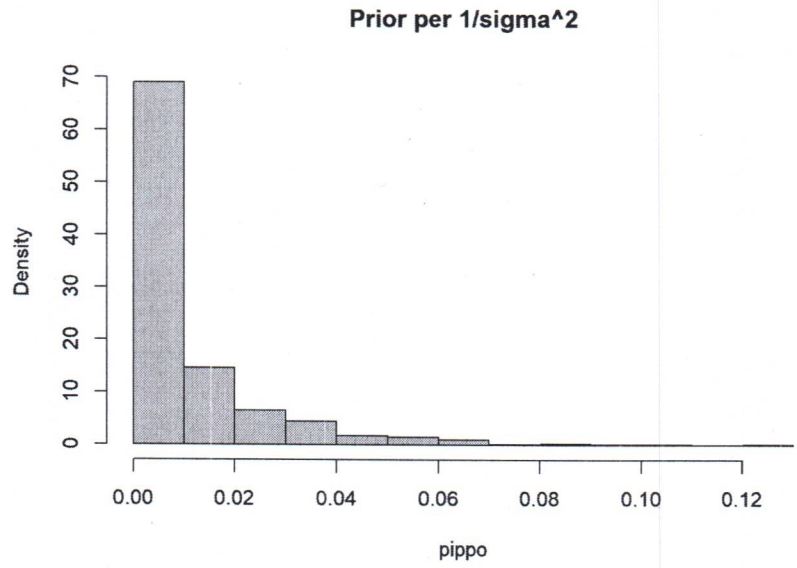



```
# LEFT:  histogram of the empirical means per school ybar_j
# RIGHT: very extreme sample averages (very small or very large)
#        tend to be associated with schools with small sample size
# Dal grafico a destra, sembra che le scuole con ybar più grande (sample average) o
# più piccolo corrispondano a numerosità piccole. Questo si spiega perchè, a parità
# di media campionaria, la varianza campionaria è \sigma^2/n_i: più è picolo n_i e
# maggiore sarà la variabilità di ybar.
```

```
### ---------------------------------------------------------------------------
### PARAMETRI
### ---------------------------------------------------------------------------
# PARAMETERS: (theta_1,...,theta_100, mu, tau^2,sigma^2)
# theta_j=mean in school j, mu= grand mean,
# tau^2 = variance between schools, sigma^2 = variance within school j (costant)
# la prior per theta_1,...,theta_100  è di tipo gerarchico - vedi i Lucidi
```

```
### ---------------------------------------------------------------------------
### PRIOR for(sigma^2, mu , tau^2)
### ---------------------------------------------------------------------------
# 1/sigma^2 \sim gamma(nu0/2, (nu0*sigma0^2)/2)
# 1/tau^2 \sim gamma(eta0/2, (eta0*tau0^2)/2)
# mu \sim N(mu0,gamma0^2)
# sigma^2: within-group variance (it is constant in this case, but a more general model
#          with sigma_j^2)
# tau^2: between-group variance
```

```
windows()
### ---------------------------------------------------------------------------
### IPERPARAMETRI
### ---------------------------------------------------------------------------
# Iperparametri assegnati sulla base di INFORMAZIONI a priori, cioè a prescindere
#   dai dati di questo specifico esperimento
par(mfrow=c(1,1))

# iperparametri per la prior di sigma^2 = within-group variance
nu0   = 1
s20   = 100
pippo = rgamma(1000, shape=nu0/2, rate=nu0/2*s20 )
hist(pippo, prob=T, main='Prior per 1/sigma^2')
```




```
mean(1/pippo)              # media di sigma^2 (+infinito!)
```

```
## [1] 47535.22
```

```
var(1/pippo)               # varianza di sigma^2 (+infinito!)
```

```
## [1] 677971351843
```

```
# iperparametri per la prior di tau^2
eta0 = 1
t20  = 100
mu0  = 50          # IMPORTANT: iperparametri di mu
g20  = 25          # IMPORTANT: iperparametri di mu
# A priori P(\mu \in (40,60))= 0.94 circa e E(\mu)=50 che è la media nazionale
# Prior informativa, ricavata da info sul test somministato a tutta la nazione
dev.off()
```

```
## png
##   2
```

we use MCMC to obtain the posterior distribution of all the parameters

(how many parameters?
100 group's specific parameters (average score in each group), then we have some extra parameters: σ^2, τ^2, μ)

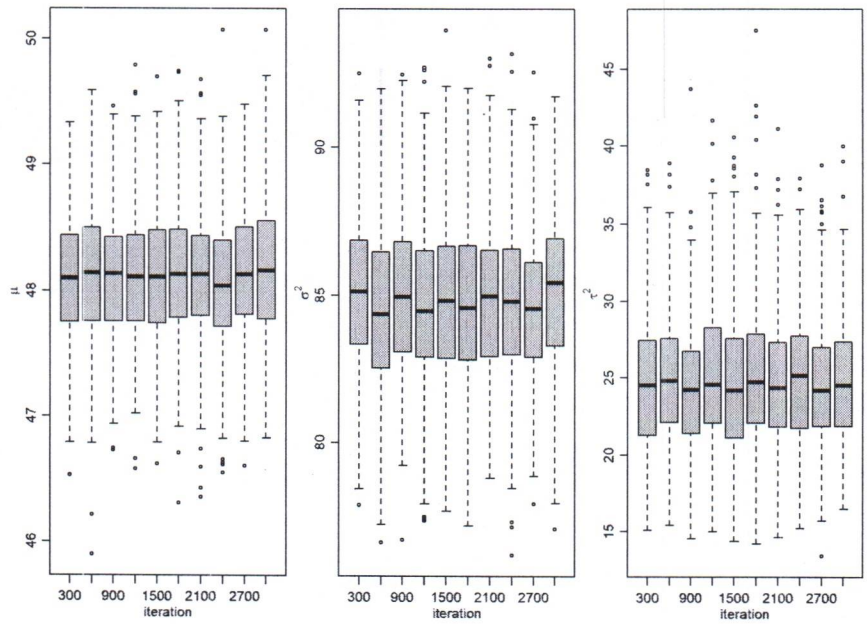
```
### ------------------------------------------------------------------------------
### MCMC analysis for school data
### ------------------------------------------------------------------------------
# PARAMETERS: (theta_1,...,theta_100, mu, tau^2,sigma^2)
# theta_j = mean in school j, mu= grand mean,
# tau^2   = variance between schools,
# sigma^2 = variance within school j (costant) starting values of the MC
theta  = ybar
sigma2 = mean(sv)
mu     = mean(theta)
tau2   = var(theta)
```

```
# setup MCMC
set.seed(1)
S     = 3000
THETA = matrix( nrow=S,ncol=m)
MST   = matrix( nrow=S,ncol=3)
```

```
# MCMC algorithm - GIBBS SAMPLER
for(s in 1:S){

  # sample new values of the thetas
  for(j in 1:m){
    vtheta   = 1/(n[j]/sigma2+1/tau2)
    etheta   = vtheta*(ybar[j]*n[j]/sigma2+mu/tau2)
    theta[j] = rnorm(1,etheta,sqrt(vtheta))
   }

  #sample new value of sigma^2
  nun    = nu0+sum(n)
  ss     = nu0*s20
  for(j in 1:m){
    ss = ss+sum((Y[Y[,1]==j,2]-theta[j])^2)
  }
  sigma2 = 1/rgamma(1,nun/2,ss/2)

  #sample a new value of mu
  vmu =  1/(m/tau2+1/g20)
  emu = vmu*(m*mean(theta)/tau2 + mu0/g20)
  mu  = rnorm(1,emu,sqrt(vmu))

  # sample a new value of tau2
  etam = eta0+m
  ss   = eta0*t20 + sum((theta-mu)^2)
  tau2 = 1/rgamma(1,etam/2,ss/2)

  #store results
  THETA[s,] = theta
  MST[s,]   = c(mu,sigma2,tau2)

}
```

```
# THETA contiene i valori simulati dei 100 theta_i, MST contiene i valori simulati di
# mu, sigma^2, tau^2

mcmc1 = list(THETA=THETA,MST=MST)

stationarity.plot<-function(x,...){
  S     = length(x)
  scan  = 1:S
  ng    = min( round(S/100),10)
  group = S*ceiling(ng*scan/S)/ng
  boxplot(x~group,...)
}
```

```
### ------------------------------------------------------------------------------
### Check CONVERGENCE of the MC
### ------------------------------------------------------------------------------
# Each graph contains 10 boxplots of 500 next iterations
# (dunque 1/10 di tutti i dati simulati) of mu,sigma^2 e tau^2
# Boxplots are similar: convergence is OK

x11()
par(mfrow=c(1,3),mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
stationarity.plot(MST[,1],xlab="iteration",ylab=expression(mu))
stationarity.plot(MST[,2],xlab="iteration",ylab=expression(sigma^2))
stationarity.plot(MST[,3],xlab="iteration",ylab=expression(tau^2))
```



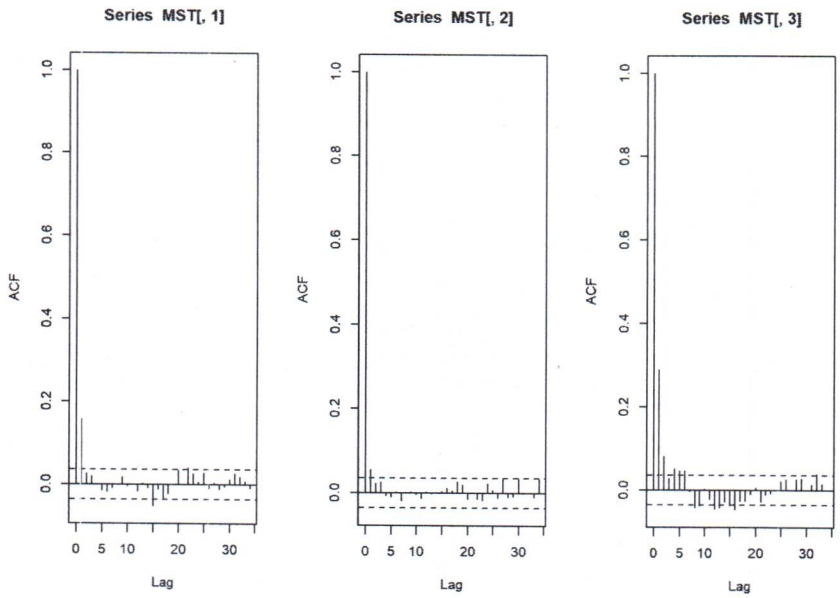

```
library(coda)
```

```
## Warning: package 'coda' was built under R version 4.0.2
```

```
effectiveSize(MST)
```

```
##     var1     var2     var3
## 2181.758 2682.883 1561.308
```

```
# effective sample size of MST are good!
# Good autocorrelation functions
par(mfrow=c(1,3))
acf(MST[,1]) -> a1
acf(MST[,2]) -> a2
acf(MST[,3]) -> a3
```




```
# Stima dell'errore Monte Carlo = posterior sd / sqrt(effectivesamplesize)
# per mu, sigma^2, tau^2
MCERR = apply(MST,2,sd)/sqrt( effectiveSize(MST) )
MCERR
```

```
##       var1       var2       var3
## 0.01144748 0.05218863 0.11075830
```

```
### ---------------------------------------------------------------------------
### POSTERIOR MEANS of  mu, sigma^2, tau^2
### ---------------------------------------------------------------------------
apply(MST,2,mean)
```

```
## [1] 48.11913 84.82484 24.81845
```

```
100*MCERR/apply(MST,2,mean)
```

```
##       var1       var2       var3
## 0.02378987 0.06152517 0.44627400
```

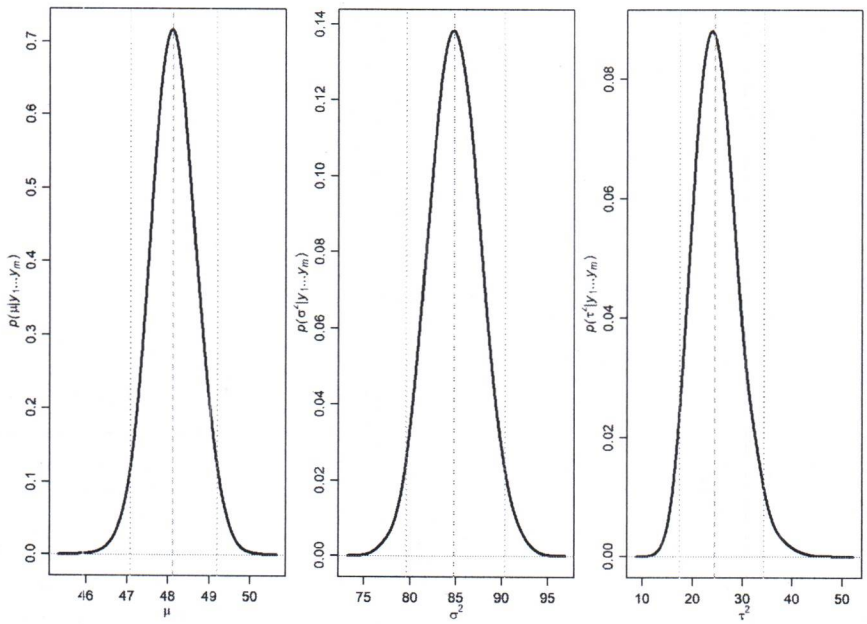
```
#sono tutti e 3 minori di 1%, cioè MCerr diviso la media a posteriori è piccolo

# effective sample size of theta_j are OK!
effectiveSize(THETA) -> esTHETA
esTHETA
```

```
##     var1     var2     var3     var4     var5     var6     var7     var8
## 3000.000 3000.000 3000.000 2835.607 2773.557 3000.000 3000.000 2823.037
##     var9    var10    var11    var12    var13    var14    var15    var16
## 3000.000 3000.000 3047.000 3000.000 3000.000 2709.323 3000.000 2823.887
##    var17    var18    var19    var20    var21    var22    var23    var24
## 3000.000 3000.000 3000.000 3000.000 3000.000 3000.000 3000.000 3000.000
##    var25    var26    var27    var28    var29    var30    var31    var32
## 3000.000 3000.000 3000.000 3000.000 3000.000 3000.000 3000.000 3000.000
##    var33    var34    var35    var36    var37    var38    var39    var40
## 3000.000 2766.578 2532.776 3000.000 3535.894 3000.000 3000.000 3022.322
##    var41    var42    var43    var44    var45    var46    var47    var48
## 2826.222 3000.000 3000.000 2831.759 3083.751 3229.801 3000.000 3000.000
##    var49    var50    var51    var52    var53    var54    var55    var56
## 2223.843 3000.000 2779.072 3000.000 2713.508 2659.269 3000.000 3000.000
##    var57    var58    var59    var60    var61    var62    var63    var64
## 3000.000 3000.000 3000.000 3000.000 3000.000 2810.154 3000.000 3328.548
##    var65    var66    var67    var68    var69    var70    var71    var72
## 3000.000 3000.000 2278.496 3000.000 3000.000 3000.000 3000.000 3023.331
##    var73    var74    var75    var76    var77    var78    var79    var80
## 2438.459 2560.733 3000.000 3000.000 3000.000 3000.000 2814.862 3231.348
##    var81    var82    var83    var84    var85    var86    var87    var88
## 3317.099 3000.000 3000.000 3000.000 3000.000 3000.000 2195.648 3000.000
##    var89    var90    var91    var92    var93    var94    var95    var96
## 2815.079 3000.000 2791.599 3000.000 3000.000 3000.000 3000.000 3000.000
##    var97    var98    var99   var100
## 2667.618 3000.000 3000.000 3000.000
```

```
# Stima dell'errore Monte Carlo dei THETA = posterior sd / sqrt(effectivesamplesize)
TMCERR = apply(THETA,2,sd)/sqrt( effectiveSize(THETA) )
TMCERR
```

```
##       var1       var2       var3       var4       var5       var6       var7
## 0.02877329 0.03422798 0.03252549 0.03644080 0.03481999 0.03858867 0.03781324
##       var8       var9      var10      var11      var12      var13      var14
## 0.03437973 0.03201497 0.03609529 0.04378003 0.02886951 0.03565356 0.03828509
##      var15      var16      var17      var18      var19      var20      var21
## 0.04284477 0.03396414 0.05257061 0.03985061 0.04261324 0.03397293 0.04057653
##      var22      var23      var24      var25      var26      var27      var28
## 0.03298834 0.04098611 0.03423903 0.03499223 0.02956241 0.04318376 0.03340499
##      var29      var30      var31      var32      var33      var34      var35
## 0.03396293 0.03843781 0.03399910 0.03912730 0.03194798 0.03345426 0.04060165
##      var36      var37      var38      var39      var40      var41      var42
## 0.04312391 0.02583338 0.03763539 0.03038907 0.03177588 0.03391712 0.03021079
##      var43      var44      var45      var46      var47      var48      var49
## 0.03952329 0.03799315 0.03816822 0.03269825 0.03543324 0.03255815 0.04626951
##      var50      var51      var52      var53      var54      var55      var56
## 0.03577592 0.03782972 0.03184833 0.04532095 0.03250366 0.03260438 0.03137780
##      var57      var58      var59      var60      var61      var62      var63
## 0.03071143 0.03035553 0.03634170 0.02971936 0.03232323 0.04318661 0.03752350
##      var64      var65      var66      var67      var68      var69      var70
## 0.02811411 0.03683646 0.03983028 0.07114928 0.03557402 0.03152138 0.03106592
##      var71      var72      var73      var74      var75      var76      var77
## 0.03377813 0.04199637 0.03430484 0.04053797 0.03258360 0.04092370 0.03368399
##      var78      var79      var80      var81      var82      var83      var84
## 0.04168852 0.04340044 0.04092854 0.03215151 0.05761592 0.04018776 0.03200330
##      var85      var86      var87      var88      var89      var90      var91
## 0.03868637 0.03702990 0.03854824 0.03402452 0.03238596 0.02905814 0.03702510
##      var92      var93      var94      var95      var96      var97      var98
## 0.03447812 0.03193285 0.04920540 0.03129507 0.03451554 0.03243099 0.02846818
##      var99     var100
## 0.03266361 0.02908111
```

```
### ---------------------------------------------------------------------------
### MARGINAL POSTERIOR densities of mu, sigma^2, tau^2
### ---------------------------------------------------------------------------
x11()
par(mfrow=c(1,3),mar=c(2.75,2.75,.5,.5),mgp=c(1.7,.7,0))
plot(density(MST[,1],adj=2),xlab=expression(mu),main="",lwd=2,
ylab=expression(paste(italic("p("),mu,"|",italic(y[1]),"...",italic(y[m]),")")))
abline( v=quantile(MST[,1],c(.025,.5,.975)),col="gray",lty=c(3,2,3) )
plot(density(MST[,2],adj=2),xlab=expression(sigma^2),main="", lwd=2,
ylab=expression(paste(italic("p("),sigma^2,"|",italic(y[1]),"...",italic(y[m]),")")))
abline( v=quantile(MST[,2],c(.025,.5,.975)),col="gray",lty=c(3,2,3) )
plot(density(MST[,3],adj=2),xlab=expression(tau^2),main="",lwd=2,
ylab=expression(paste(italic("p("),tau^2,"|",italic(y[1]),"...",italic(y[m]),")")))
abline( v=quantile(MST[,3],c(.025,.5,.975)),col="gray",lty=c(3,2,3) )
```




```
### ---------------------------------------------------------------------------
### POSTERIOR MEANS of  mu, sigma, tau
### ---------------------------------------------------------------------------
mean((MST[,1]))
```

```
## [1] 48.11913
```

```
mean(sqrt(MST[,2]))
```

```
## [1] 9.208871
```

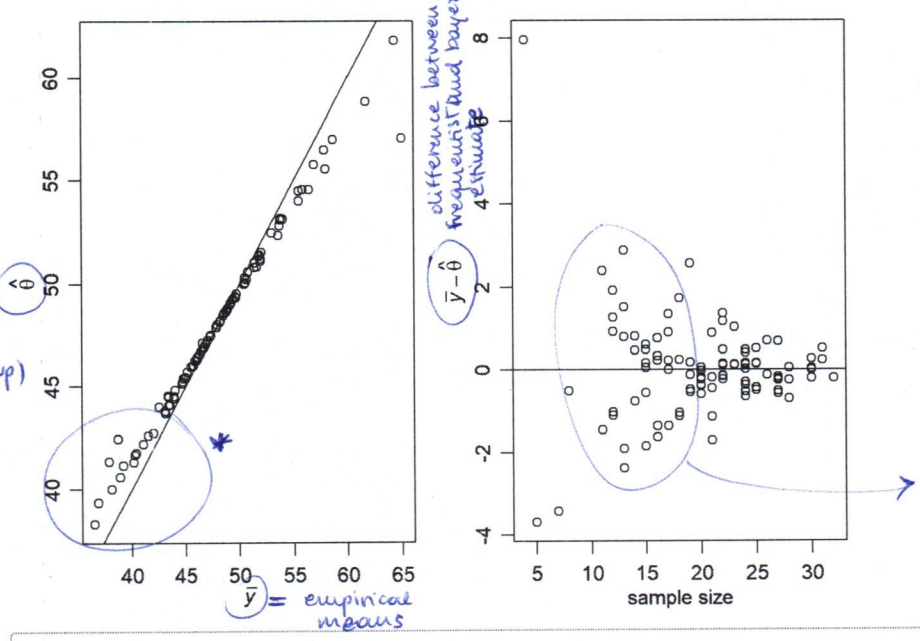
```
mean(sqrt(MST[,3]))
```

```
## [1] 4.962866
```

```
##################################################
##### SHRINKAGE effect towards grand mean mu #####
##################################################
### Vediamo come le informazioni sono state condivise tra i diversi gruppi:
### Bayesian hierarchical approach
### The Bayesian estimate of group-specific parameter theta_j (given the other
### parameters) is a convex linear combination of the frequentist group estimate ybar_j
### on one hand, and of mu (the prior mean of all theta_j); the Bayesian estimate is
### pulled a bit away from ybar_j towards mu by an amount depending on n_j.
### This is called SHRINKAGE
### BTW, mu is random, but E(mu)=50.
###  * Se n_j è GRANDE, la media empirica è una buona stima, anche da punto di vista
###    bayesiano; dunque non c'è necessità ci CHIEDERE in PRESTITO (to BORROW)
###    informazioni dal resto dei gruppi.
###  * Se n_j è piccolo, ybar_j NON va bene come stima e "correggo" in modo tale che la
###    stima bayesiana si avvicini alla (stima) di mu, la media dei gruppi.
x11()
par(mar=c(3,3,1,1),mgp=c(1.75,.75,0))
par(mfrow=c(1,2))
# theta.hat = posterior means of theta_j
theta.hat = apply(THETA,2,mean)
theta.hat
```

```
##   [1] 50.57676 46.70989 48.64741 47.44776 38.28095 40.59272 41.77588 48.74254
##   [9] 48.99676 42.19340 55.55526 50.26446 49.26732 56.98437 54.52062 54.47219
##  [17] 41.34531 49.99065 44.53789 46.24030 50.22075 47.89160 51.18006 45.94190
##  [25] 45.37965 46.81815 44.48611 51.31496 46.41368 49.37979 49.08003 50.06573
##  [33] 47.51597 45.99393 54.52298 52.31648 46.28063 51.19516 46.42103 48.84202
##  [41] 55.75782 46.40187 50.78539 49.07248 45.71934 41.32386 44.51691 46.93248
##  [49] 40.01100 42.62201 61.82637 49.23124 43.99787 46.55042 43.77808 49.08701
##  [57] 44.42030 48.44219 49.52156 42.73716 45.43251 50.82228 49.33689 43.67175
##  [65] 47.87643 48.20392 57.05386 45.23758 51.39440 44.05368 46.86159 39.32206
##  [73] 53.13924 41.66623 48.74292 53.99410 45.12978 41.17207 58.83749 51.09624
##  [81] 47.07152 42.45855 44.82787 44.10998 48.72981 52.79712 56.49483 53.11902
##  [89] 53.07082 48.56132 47.87870 48.13584 51.51001 47.13348 45.45597 47.23864
##  [97] 46.00551 52.45490 50.99465 47.99872
```

```
## LEFT
## ybar (le stime 'classiche', che sono sample average per scuola)
#  on the x-axis and theta.hat (le stime bayesiane) on the y-axis
plot(ybar,theta.hat,xlab=expression(bar(italic(y))),ylab=expression(hat(theta)))
abline(0,1)
## The slope of this line is <1, that is, high values of ybar.j correspond to
## slightly less high values of the Bayesian estimates of theta_j,
## and low values of ybar.j correspond to slightly less low values of
## the Bayesian estimates of theta_j. This is the SHRINKAGE effect

## RIGHT
## group-specific sample sizes on the x-axis, and differences
## between frequentist and Bayesian estimates on the y-axis.
plot(n,ybar-theta.hat,ylab=expression( bar(italic(y))-hat(theta) ),xlab="sample size")
abline(h=0)
```



bayesian estimate of the θ (for each group)

$\bar{y}$ = empirical means

difference between frequentist and bayesian estimate

* SHRINKAGE EFFECT:
to items corresponding to very small empirical estimates of θ we have small bayesian estimates of θ_j but less smaller than the expected
⟶ somehow we've using other informations

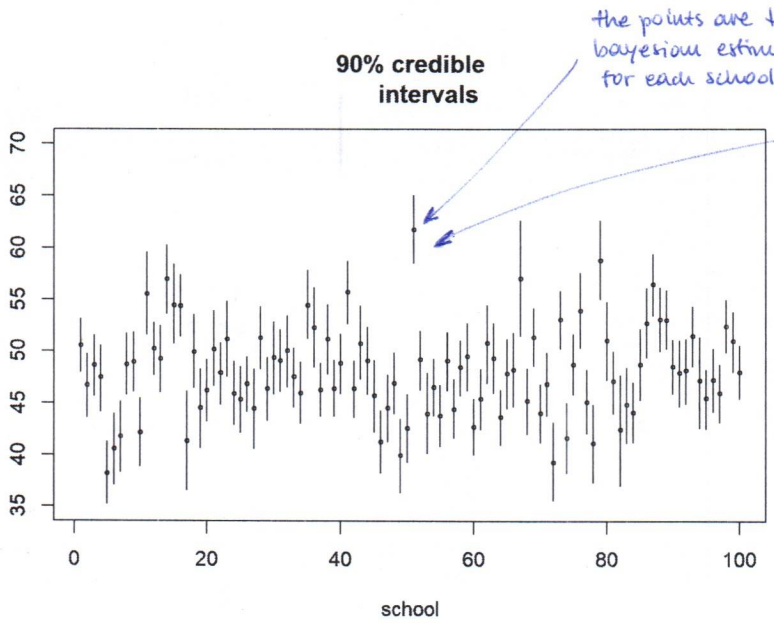
the difference between bayesian and frequentist is very large when the sample size is small

```
## Groups with low sample size get shrunk the most, whereas groups with large sample
## size hardly get shrunk at all. The larger the sample size for a group, the more
## information we have for that group and the less information we need to BORROW
## from the rest of the population.
dev.off()
```

```
## png
##   2
```

we can compare two schools comparing the posterior probability that the corresponding θ_j are ordered:
for instance if we want to order school **i** and school **j** we will compare the posterior probability that θ_i is larger than θ_j and if this probability is close to 1 (/very large) then we can say that school **i** is better than school **j**

```
windows()
### ------------------------------------------------------------------------------
### SCHOOLS COMPARISON
### ------------------------------------------------------------------------------
# posterior CIs for each theta_j, group-specific parameters
plot(1:100,theta.hat, cex=0.5,main="90% credible
     intervals",xlim=c(1,100),ylim=c(35,70),xlab='school',ylab='')
for (i in 1:100) {
  probint = quantile(THETA[,i], c(0.05, 0.95))
  lines(i*c(1,1), probint)
}
```



the points are the bayesian estimates for each school

the intervals are credible intervals ⟹ marginally, each $\theta_j \in CI_j$ with probability equal to 90% a posteriori

We can already see that some schools behave worse than others.

```
# min of theta.hat is SCHOOL 5
# max of theta.hat is SCHOOL 51
which(theta.hat==min(theta.hat))
```

```
## [1] 5
```

← School 5 seems to be the worst

```
which(theta.hat==max(theta.hat))
```

```
## [1] 51
```

```
# For each couple (theta_j, theta_l): posterior probability that  theta_j > theta_l,
# that is posterior prob that school j is BETTER than school l
# Matrix of these posterior probabilities, better

compare.rates <- function(x) {
  nc = 100
  ij = as.matrix(expand.grid(1:nc, 1:nc))
  m  = as.matrix(x[,ij[,1]] > x[,ij[,2]])
  matrix(colMeans(m), nc, nc, byrow = TRUE)
}

better=compare.rates(THETA)
better
```

```
##              [,1]        [,2]         [,3]         [,4]         [,5]
##   [1,] 0.000000000 0.057666667 0.2083333333 0.0970000000 0.0000000000
##   [2,] 0.942333333 0.000000000 0.7730000000 0.6136666667 0.0010000000
##   [3,] 0.791666667 0.227000000 0.0000000000 0.3236666667 0.0000000000
##   [4,] 0.903000000 0.386333333 0.6763333333 0.0000000000 0.0000000000
##   [5,] 1.000000000 0.999000000 1.0000000000 1.0000000000 0.0000000000
##   [6,] 1.000000000 0.986666667 0.9980000000 0.9913333333 0.2136666667
##   [7,] 1.000000000 0.959000000 0.9943333333 0.9783333333 0.0976666667
##   ..
##              [,93]        [,94]        [,95]        [,96]        [,97]
##   [1,] 0.6526666667 0.1360000000 0.0130000000 0.0830000000 0.0236666667
##   [2,] 0.9643333333 0.5540000000 0.3150000000 0.5610000000 0.3883333333
##   [3,] 0.8733333333 0.3136666667 0.0940000000 0.2860000000 0.1370000000
##   [4,] 0.9446666667 0.4553333333 0.2156666667 0.4710000000 0.2880000000
##   [5,] 1.0000000000 0.9966666667 0.9976666667 0.9996666667 0.9996666667
##   [6,] 1.0000000000 0.9686666667 0.9586666667 0.9886666667 0.9750000000
##   [7,] 1.0000000000 0.9400000000 0.9160000000 0.9776666667 0.9446666667
##   ..
##             [,98]       [,99]      [,100]
##   [1,] 0.79433333 0.5700000000 0.1183333333
##   [2,] 0.99133333 0.9503333333 0.7026666667
##   [3,] 0.94966667 0.8160000000 0.3966666667
##   [4,] 0.98133333 0.9153333333 0.5953333333
##   [5,] 1.00000000 1.0000000000 1.0000000000
##   [6,] 1.00000000 1.0000000000 0.9980000000
##   [7,] 1.00000000 0.9990000000 0.9923333333
##   ..
```

```
# Posterior probability that school 5 is better than all the other schools
better[,5]
```

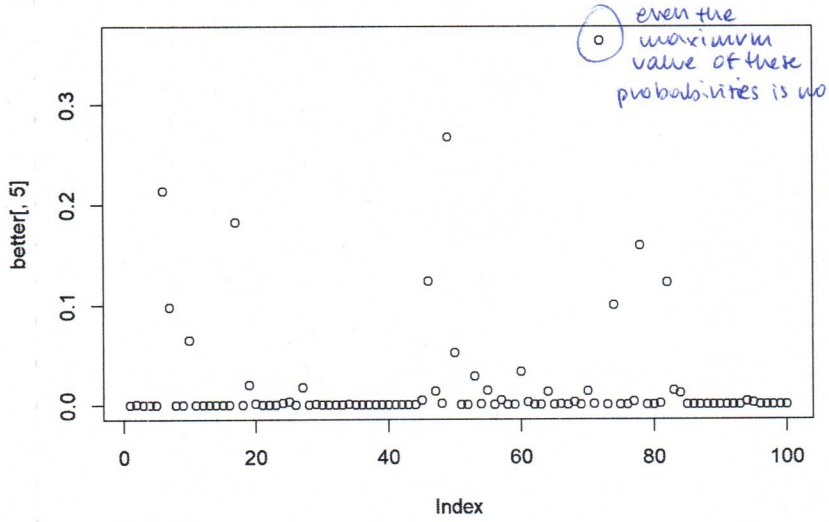
```
##   [1] 0.0000000000 0.0010000000 0.0000000000 0.0000000000 0.0000000000
##   [6] 0.2136666667 0.0976666667 0.0000000000 0.0000000000 0.0650000000
##  [11] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000
##  [16] 0.0000000000 0.1826666667 0.0000000000 0.0020000000 0.0013333333
##  [21] 0.0000000000 0.0000000000 0.0000000000 0.0020000000 0.0036666667
##  [26] 0.0003333333 0.0176666667 0.0000000000 0.0010000000 0.0000000000
##  [31] 0.0000000000 0.0000000000 0.0000000000 0.0010000000 0.0000000000
##  [36] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000
##  [41] 0.0000000000 0.0003333333 0.0000000000 0.0000000000 0.0046666667
##  [46] 0.1233333333 0.0136666667 0.0013333333 0.2680000000 0.0516666667
##  [51] 0.0000000000 0.0000000000 0.0286666667 0.0006666667 0.0146666667
##  [56] 0.0000000000 0.0043333333 0.0000000000 0.0000000000 0.0333333333
##  [61] 0.0026666667 0.0000000000 0.0000000000 0.0133333333 0.0003333333
##  [66] 0.0006666667 0.0000000000 0.0030000000 0.0000000000 0.0140000000
##  [71] 0.0006666667 0.3633333333 0.0000000000 0.1000000000 0.0000000000
##  [76] 0.0000000000 0.0036666667 0.1593333333 0.0000000000 0.0000000000
##  [81] 0.0013333333 0.1223333333 0.0143333333 0.0116666667 0.0003333333
##  [86] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0000000000
##  [91] 0.0003333333 0.0003333333 0.0000000000 0.0033333333 0.0023333333
##  [96] 0.0003333333 0.0003333333 0.0000000000 0.0000000000 0.0000000000
```

posterior probabilities that $\theta_5 > \vartheta_j \quad j = 1, \dots, 100 \setminus \{5\}$

↓

what should be the worst school

```
windows()
plot(better[,5])
```



even the maximum value of these probabilities is not very large (0.3-)

⟹ we can conclude that school 5 is worse than the others

```
# All are <= 0.37
min(better[-5,5])
```

```
## [1] 0
```

```
max(better[-5,5])
```

```
## [1] 0.3633333
```

```
better[51,5]  # probabilità a posteriori che la scuola 5 sia meglio della scuola 51
```

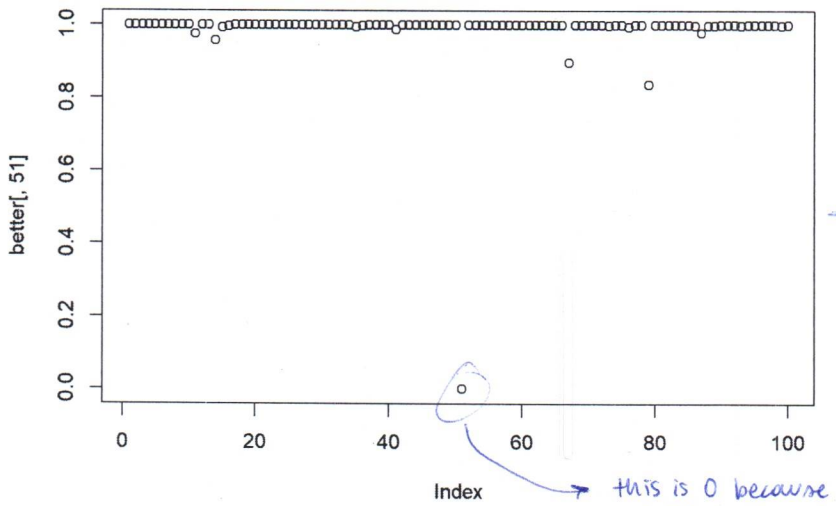
```
## [1] 0
```

```
# Posterior probability that school 51 is better than all the other schools
better[,51]
```

we do now the same for school 51 (the one we suppose is the best)

```
##   [1] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##   [8] 1.0000000 1.0000000 1.0000000 0.9756667 1.0000000 1.0000000 0.9586667
##  [15] 0.9930000 0.9973333 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [22] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [29] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 0.9946667
##  [36] 0.9996667 1.0000000 1.0000000 1.0000000 1.0000000 0.9886667 1.0000000
##  [43] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [50] 1.0000000 0.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [57] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [64] 1.0000000 1.0000000 1.0000000 1.0000000 0.8970000 1.0000000 1.0000000
##  [71] 1.0000000 1.0000000 0.9986667 1.0000000 1.0000000 0.9960000 1.0000000
##  [78] 1.0000000 0.8370000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [85] 1.0000000 0.9990000 0.9786667 0.9993333 0.9993333 1.0000000 1.0000000
##  [92] 1.0000000 0.9996667 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##  [99] 0.9996667 1.0000000
```

```
windows()
plot(better[,51])
```



posterior probability (ies) that: $\theta_{51} > \theta_j$ $j = 1, \dots, 100$

very lounge values (they go from 0.84 to 1), so we can conclude that school 51 is the best school

this is 0 because probably is the school 51

```
# All are >= 0.8398 and <= 1
min(better[-51,51])
```

```
## [1] 0.837
```

```
max(better[-51,51])
```

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

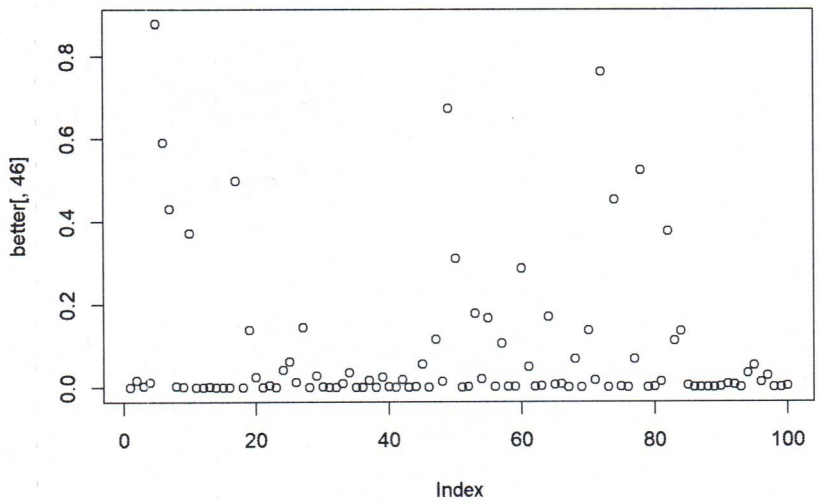
```
better[5,51]  # probabilità a posteriori che la scuola 51 sia meglio della scuola 5
```

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

```
better[,46]
```

```
##   [1] 0.0003000000 0.0180000000 0.0026666667 0.0123333333 0.8766666667
##   [6] 0.5905666667 0.4316666667 0.0026666667 0.0013333333 0.3713333333
##  [11] 0.0000000000 0.0003333333 0.0013333333 0.0000000000 0.0000000000
##  [16] 0.0000000000 0.4986666667 0.0006666667 0.1393333333 0.0260000000
##  [21] 0.0003333333 0.0056666667 0.0000000000 0.0426666667 0.0633333333
##  [26] 0.0126666667 0.1450000000 0.0000000000 0.0280000000 0.0020000000
##  [31] 0.0006666667 0.0006666667 0.0103333333 0.0363333333 0.0003333333
##  [36] 0.0000000000 0.0176666667 0.0010000000 0.0246666667 0.0023333333
##  [41] 0.0000000000 0.0196666667 0.0003333333 0.0023333333 0.0566666667
##  [46] 0.0000000000 0.1160000000 0.0150000000 0.6730000000 0.3093333333
##  [51] 0.0000000000 0.0013333333 0.1783333333 0.0203333333 0.1670000000
##  [56] 0.0013333333 0.1066666667 0.0023333333 0.0013333333 0.2860000000
##  [61] 0.0503333333 0.0013333333 0.0026666667 0.1696666667 0.0060000000
##  [66] 0.0080000000 0.0000000000 0.0696666667 0.0000000000 0.1383333333
##  [71] 0.0173333333 0.7620000000 0.0000000000 0.4530000000 0.0016666667
##  [76] 0.0000000000 0.0686666667 0.5236666667 0.0000000000 0.0013333333
##  [81] 0.0136666667 0.3756666667 0.1126666667 0.1363333333 0.0046666667
##  [86] 0.0000000000 0.0000000000 0.0000000000 0.0000000000 0.0016666667
##  [91] 0.0083333333 0.0066666667 0.0000000000 0.0346666667 0.0540000000
##  [96] 0.0123333333 0.0276666667 0.0000000000 0.0000000000 0.0040000000
```

```
windows()
plot(better[,46])
```




```
min(better[-46,46])
```

```
## [1] 0
```

```
max(better[-46,46])
```

```
## [1] 0.8766667
```

```
# According to the group sample averages, school 46 is better that school 82
ybar[c(46,82)]
```

46 is better than 82 simply confronting the average estimation (mean)

```
## [1] 40.17619 38.76400
```

```
# Posterior probability that school 46 is better than 82
better[82,46]
```

```
## [1] 0.3756667
```

(not very high ⟹ according to the posterior distribution there is no clear evidence that school 46 is better than school 82 (despite the order of the sample averages))

!

Bayesian statistics combine not only data (using the averages to compute the estimates) but use extra-informations coming from the sharing of informations between the groups.

CONDITIONAL PROBABILITY

$(\Omega, \mathcal{F}, \mathbb{P})$ probability space, $\mathcal{G}$ sub-σ-field of $\mathcal{F}$

Def. $A, B \in \mathcal{F}$, $\mathbb{P}(B) > 0 \implies \mathbb{P}(A|B) = \dfrac{\mathbb{P}(A,B)}{\mathbb{P}(B)}$

Def. (Conditional probability)
$\mathbb{P}(A|\mathcal{G})$ is a random variable s.t.:

1. $\omega \mapsto \mathbb{P}(A|\mathcal{G})(\omega)$ is $\mathcal{G}$-measurable
2. $\mathbb{P}(A|\mathcal{G})$ satisfies:

$$\mathbb{P}(A, G) = \int_G \mathbb{P}(A|\mathcal{G})(\omega)\ \mathbb{P}(d\omega) \qquad \forall G \in \mathcal{G}$$

Def. (Conditional distribution)
X random variable on $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{G}$ σ-field $\subset \mathcal{F}$.
$\mu(H, \omega)$ conditional distribution of X given $\mathcal{G}$:
$\mu: \mathcal{B}(\mathbb{R}) \times \Omega \to \mathbb{R}$ if:

1. $H \longmapsto \mu(H, \omega)$ is a probability measure (on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$)
2. $\omega \longmapsto \mu(H, \omega)$ is the conditional probability $\mathbb{P}(X \in H|\mathcal{G})$

i.e. $\mathbb{P}(\{X \in H\}, G) = \int_G \mu(H, \omega)\ \mathbb{P}(d\omega) = \int_G \mathbb{P}(X \in H|\mathcal{G})\ \mathbb{P}(d\omega)$

Remark: Typically $\mathcal{G} = \sigma(Y)$ where Y is a random variable (defined on the same probability space)
$\implies \mathcal{L}(X|Y)$ = conditional distribution of X given Y

Suppose (X, Y) random elements,
suppose that their joint distribution is:
$(X, Y) \sim f(x, y)$ w.r.t. the Lebesgue measure on $\mathbb{R}^2$

$f(x, y)$: density w.r.t. same σ-finite measure

(or counting measure if the random variables are discrete)

$\implies$ in this case means that the vector (X, Y) is ABSOLUTELY CONTINUOUS

$f(x, y) \longrightarrow f_Y(y)$ (marginal density of Y)

$\longrightarrow f_{X|Y=y}(x) = \dfrac{f(x, y)}{f_Y(y)}$ (conditional density of $\mathcal{L}(X|Y)$)

$\longrightarrow \mathbb{E}[X|Y] = \int_{\mathbb{R}} x \cdot f_{X|Y=y}(x)\, dx$

Remember: this is a random variable

What is an abs. continuous random variable? The cumulative density function (funzione di ripart.) of an absolute continuous r.v. is an integral solution.

= la sua funzione di ripartizione la posso scrivere come l'integrale tra $-\infty$ e x di un'altra funzione (densità)

$\implies$ (X, Y) è vettore ass. cont. di dimensione 2 se la sua funzione di rip. congiunta si scrive come una funzione integrale (integrale doppio $\int_{-\infty}^{x}\int_{-\infty}^{y}$ densità cong.)

Properties of conditional expectation : $\mathbb{E}[X|Y]$

remember that defying $\mathbb{E}[X|Y]$ we also define the conditional probability:

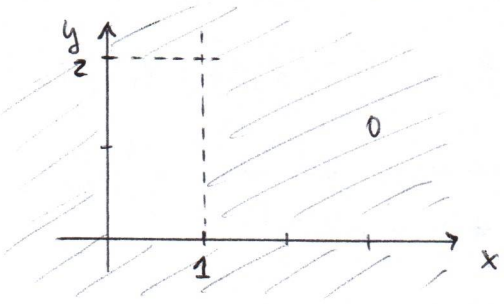
$$\mathbb{P}(X \in H | Y) = \mathbb{E}[\mathbb{1}_H(X)|Y]$$

Let **X** and **Y** be two random elements : (e.g. random variables)

1. $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$
2. $\text{Var}(X) = \text{Var}(\mathbb{E}[X|Y]) + \mathbb{E}[\text{Var}(X|Y)]$
3. $\mathbb{E}[c|Y] = c$ = constant
4. $\mathbb{E}[aX_1 + bX_2|Y] = a\,\mathbb{E}[X_1|Y] + b\,\mathbb{E}[X_2|Y]$ $\quad a,b \in \mathbb{R}$
5. $X \geq 0$ a.s. $\Longrightarrow$ $\mathbb{E}[X|Y] \geq 0$ a.s.

   almost surely (= a.e.)
6. $X_1 \geq X_2$ a.s. $\Longrightarrow$ $\mathbb{E}[X_1|Y] \geq \mathbb{E}[X_2|Y]$
7. $\mathbb{E}[X|X] = X$
8. $\mathbb{E}[X\,f(Y)|Y] = f(Y)\,\mathbb{E}[X|Y]$

Ex. 1 $(X,Y) \sim f(x,y) = \begin{cases} \frac{3}{4}\left(xy + \frac{x^2}{2}\right) & \text{if } 0<x<1,\ 0<y<2 \\ 0 & \text{otherwise} \end{cases}$



1. $f_{Y|X}$?
2. $\mathbb{E}[Y]$?

1. We need $f_X(x)$ since $f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)}$

Let $x \in (0,1)$:

$$f_X(x) = \int_0^2 \frac{3}{4}\left(xy + \frac{x^2}{2}\right) dy = \int_{-\infty}^{+\infty} f(x,y)\,dy$$

$$= [\ldots]$$

$$= \frac{3}{4}(2x + x^2)\,\mathbb{1}_{(0,1)}(x)$$

$$\Longrightarrow \boxed{f_{Y|X=x}(y)} = \frac{f(x,y)}{f_X(x)} = \frac{\frac{3}{4}\left(xy + \frac{x^2}{2}\right)}{\frac{3}{4}(2x + x^2)} = \begin{cases} \frac{y + \frac{x}{2}}{2+x} & \text{if } y \in (0,2) \\ 0 & \text{otherwise} \end{cases}$$

We can calculate also:

$$\mathbb{P}(Y<1 \mid X=\tfrac{1}{2}) = \int_{-\infty}^{1} \ldots = \int_{-\infty}^{1} \frac{y+\frac{1}{4}}{2+\frac{1}{2}}\,dy =$$

$$= \frac{2}{5}\int_0^1 \left(y + \frac{1}{4}\right) dy = \frac{2}{5}\left[\frac{y^2}{2} + \frac{y}{4}\right]_0^1 = \frac{3}{10}$$

Homework : $\mathbb{P}(Y<1 \mid X<\tfrac{1}{2})$? $\left(= \frac{\mathbb{P}(X<1/2,\ Y<1)}{\mathbb{P}(X<1/2)}\right)$ $\left[\frac{2}{7}\right]$

2. $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$

$$\mathbb{E}[Y|X] = \int_{\mathbb{R}} y \cdot f_{Y|X=x}(y)\,dy = \int_0^2 y \cdot \frac{y+\frac{x}{2}}{2+x}\,dy$$

$$= \frac{1}{2+x}\int_0^2 y\left(y+\frac{x}{2}\right)dy$$

$$= \frac{1}{2+x}\left[\frac{y^3}{3} + \frac{y^2}{2}\frac{x}{2}\right]_0^2 = [..] = \frac{1}{2+x}\left(\frac{8}{3}+x\right)$$

We can also compute the marginal density of Y and then $\mathbb{E}[Y]$ (Homework!)

$$\Longrightarrow \mathbb{E}[Y] = \mathbb{E}\left[\frac{\frac{8}{3}+X}{2+X}\right] = \mathbb{E}[h(X)]$$

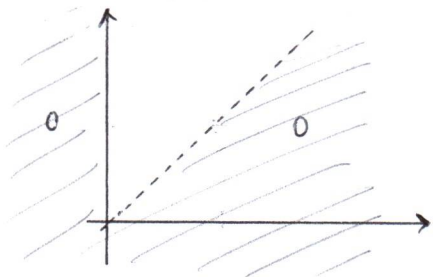
$$= \int_{\mathbb{R}} \frac{1}{2+x}\left(\frac{8}{3}+x\right) f_X(x)\,dx$$

$$= \int_{\mathbb{R}} \frac{1}{2+x}\left(\frac{8}{3}+x\right)\frac{3}{4}x(2+x)\,dx$$

$$= \int_0^1 \frac{3}{4}\left(\frac{8}{3}+x\right)x\,dx = \left[\frac{3}{4}\left(\frac{8}{3}\frac{x^2}{2} + \frac{x^3}{3}\right)\right]_0^1 = [..] = \frac{5}{4}$$

Homework: $\mathrm{Var}(Y)$? ($\Rightarrow \mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}[Y|X])$)

Ex. 2 $(X,Y) \sim f(x,y) = \begin{cases} x(y-x)e^{-y} & \text{if } 0<x<y \\ 0 & \text{otherwise} \end{cases}$



$$f_X(x) = \int_{\mathbb{R}} f(x,y)\,dy = \int_x^\infty x(y-x)e^{-y}\,dy$$

$(x>0)$

$$f_X(x) = x\int_x^\infty (y-x)e^{-y}\,dy$$

$y-x=t$

$$= x\int_0^\infty t e^{-t-x}\,dt$$

$$= x e^{-x}\boxed{\int_0^\infty t e^{-t}\,dt}$$

gamma density ($\sim$ gamma(2,1))

$$= x e^{-x}\,\mathbb{1}_{(0,+\infty)}(x) \Longrightarrow X \sim \text{gamma}(2,1)$$

$$\Longrightarrow f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{x(y-x)e^{-y}}{xe^{-x}} = \begin{cases} (y-x)e^{-(y-x)} & \text{if } \begin{cases} y>x \\ x>0 \text{ (fixed)} \end{cases} \\ 0 & \text{otherwise} \end{cases}$$

$\mathbb{E}[Y|X=x]$?

18/09

$$\mathbb{E}[Y|X=x] = \int_x^{+\infty} y \cdot f_{Y|X=x}(y)\,dy$$

$$= \int_x^{+\infty} y\left((y-x)e^{-(y-x)}\right)dy \overset{t=y-x}{=} \int_0^\infty (t+x)(te^{-t})\,dt$$

$$= \underbrace{\int_0^\infty t(te^{-t})\,dt}_{\text{expectation of the gamma density (ratio of the two parameters: gamma}(\alpha,\beta) \to \alpha/\beta)} + x\underbrace{\int_0^\infty te^{-t}\,dt}_{\text{gamma}(2,1)\text{ distribution}} = 2+x$$

$\mathbb{E}[Y]$?

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[2+X] = 2 + \mathbb{E}[X]$$

$$= 2 + \underbrace{\int_0^\infty x(xe^{-x})\,dx}_{\text{expectation of a gamma}(2,1)} = 2+2 = 4$$

Homework: derive the marginal law of Y ($\mathcal{L}(Y)$) and derive $\mathbb{E}[Y] = 4$.

$\text{Var}(Y)$?

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$$

$$\text{Var}(Y|X) = \mathbb{E}[Y^2|X] - (\mathbb{E}[Y|X])^2$$

$$= \int_x^\infty y^2((y-x)e^{-(y-x)})\,dy - (2+X)^2 \qquad y-x=t$$

$$= \int_0^\infty (x+t)^2(te^{-t})\,dt - (2+X)^2$$

$$= \int_0^\infty (t^2+x^2+2xt)(te^{-t})\,dt - (2+X)^2$$

$$= \int_0^\infty t^3e^{-t}\,dt + X^2\underbrace{\int_0^\infty te^{-t}\,dt}_{1} + 2X\underbrace{\int_0^\infty t^2e^{-t}\,dt}_{\text{expectation of gamma}(2,1) \Rightarrow = 2} - (2+X)^2$$

Trick: $\int_0^\infty t^3e^{-t}\,dt = \int_0^\infty \underbrace{t^{4-1}e^{-t}}_{\text{kernel of a gamma}(4,1)}\,dt$

but we need a normalization constant

we need $\frac{\beta^\alpha}{\Gamma(\alpha)} = \frac{1}{\Gamma(4)} = \frac{1}{3!} = \frac{1}{6}$

$$\Rightarrow \int_0^\infty t^{4-1}e^{-t}\,dt = 6\cdot\frac{1}{6}\int_0^\infty t^{4-1}e^{-t}\,dt = 6$$

$$\text{Var}(Y|X) = 6 + X^2 + 2\cdot 2X - (2+X)^2$$

$$= 6 + X^2 + 4X - 4 - X^2 - 4X = \boxed{2}$$

this time the variance is a constant but usually is a function of X

$$\Rightarrow \text{Var}(Y) = \mathbb{E}[2] + \text{Var}(2+X)$$

$$= 2 + \text{Var}(X) \leftarrow X \sim \text{Gamma}(2,1): \ \text{Var}(X) = \frac{\alpha}{\beta^2} = \frac{2}{1^2}$$

$$= 2+2 = 4$$

RADON - NIKODYM THEOREM

$(\Omega, \mathcal{F})$ measure space, μ, ν measures on $(\Omega, \mathcal{F})$.

Def. We say that ν is absolutely continuous w.r.t. μ if:

$$\mu(A) = 0 \implies \nu(A) = 0 \qquad (:= \nu \ll \mu)$$

Suppose that we have a function $f \geq 0$ $\mathcal{F}$-measurable.

Then, if $\nu(A) = \int_A f \, d\mu \implies$ ν is a measure that is absolutely continuous w.r.t. μ (we write: $\nu \ll \mu$)

Thm. Radon-Nikodym (goes the opposite)

ν, μ measures on $(\Omega, \mathcal{F})$.

ν, μ are σ-finite. (a measure is σ-finite if we can express Ω as a countable union of bounded measurable subsets so that the measure puts finite masses on these sets)

Suppose $\nu \ll \mu$.

$\implies \exists f: f \geq 0$ $\mathcal{F}$-measurable s.t. $\nu(A) = \int_A f \, d\mu$ and we denote f by $\frac{d\nu}{d\mu}$:= Radon-Nikodym density

= we can express the measure ν as an integral w.r.t. μ

Suppose to have some data $\underline{X} = (X_1, \dots, X_n) | \theta \sim P_\theta$ dominated, i.e. P_θ has a R-N density w.r.t. some fixed measure on $\mathbb{R}^n$ (Lebesgue / counting measure / ..) and we denote by $f(\underline{x}, \theta)$ this density.

In this case the likelihood of the sample $(X_1, \dots, X_n)$ coincides with $f(\underline{x}, \theta)$ but it remains a function of θ:

$$\theta \longmapsto \underbrace{L(\theta, x_1, \dots, x_n)} = f(x_1, \dots, x_n, \theta)$$

LIKELIHOOD:

- If $L(\theta_1, \underline{x}) > L(\theta_2, \underline{x})$ then it's more plausible that the observed data were generated by a mechanism where the value of θ is θ_1 instead of being generated by the same mech. with $\theta = \theta_2$.
- If $L(\theta, \underline{x}) \propto L(\theta, y)$ for two (different) observations $(\underline{x}, y)$ then the inference that we draw from $\underline{x}$ and from y is the same

BAYES THEOREM FOR DOMINATED MODELS

Consider a random sample $\underline{X}_n = (X_1, \dots, X_n) | \theta \sim P_\theta$, $\theta \in \Theta \subset \mathbb{R}^k$.
Suppose that the model is dominated (that means that there exists a density (the R-N density) : $f(\underline{x}|\theta)$ density of the joint distribution of $\underline{X}_n$ (P_θ) w.r.t. some σ-finite measure on $\mathbb{R}^n$ that we denote by $\lambda^{(n)}$).

Lebesgue / counting meas. / ..

Suppose : $\theta \sim \pi$

$\Longrightarrow$ the posterior distribution of θ given $\underline{X}_n = \underline{x}$ can be expressed as :

$$\mathbb{P}(\theta \in B \mid \underline{X}_n = \underline{x}) = \frac{\int_B f(\underline{x}|\theta)\, \pi(d\theta)}{\int_\Theta f(\underline{x}|\theta)\, \pi(d\theta)} \qquad \forall B \in \mathcal{B}(\Theta)$$

formula to compute the conditional distribution of θ given the data

proof. *

We define $\gamma :=$ joint distribution of $\underline{X}$ and θ $(\Sigma(\underline{x}, \theta))$:

$\gamma(A \times B)$ where $A \in \mathcal{B}(\mathbb{R}^n)$, $B \in \mathcal{B}(\Theta)$

$$\gamma(A \times B) := \int_B P_\theta(A)\, \pi(d\theta) = \int_B \left[\int_A f(\underline{x}|\theta)\, \lambda^{(n)} d(\underline{x})\right] \cdot \pi(d\theta)$$

somehow we want $\Sigma(\underline{x}|\theta) \cdot \Sigma(\theta)$

P_θ is dominated so we can express it as the integral of the density

$$\xrightarrow{FT} \quad \gamma(A \times B) = \int_A \left(\int_B f(\underline{x}|\theta)\, \pi(d\theta)\right) \lambda^{(n)}(d\underline{x}) \quad *$$

Now we want to define the marginal :

$$\underbrace{\mu_n(A)}_{\text{measure on } \mathcal{B}(\mathbb{R}^n)} = \gamma(A \times \Theta) = \int_A \underbrace{\left[\int_\Theta f(\underline{x}|\theta)\, \pi(d\theta)\right]}_{:= m_n(\underline{x})} \lambda^{(n)}(d\underline{x})$$

$= $ density of the measure $\mu_n(\cdot)$ w.r.t. the measure $\lambda^{(n)}(\cdot)$

$$\Longrightarrow \Sigma(\underline{x}, \theta) = \Sigma(\underline{x}|\theta)\, \Sigma(\theta) \qquad (1)$$

$$= \Sigma(\theta|\underline{x})\, \Sigma(\underline{x}) \qquad (2)$$

Bayes theorem compute this distribution equating (1) and (2) : $\longrightarrow \Sigma(\theta|\underline{x}) = \dfrac{\Sigma(\underline{x}|\theta)\, \Sigma(\theta)}{\Sigma(\underline{x})}$

(that's why we need the marginal law of $\underline{X}$ $(\mu_n(\cdot))$

We now fix B in $\mathcal{B}(\Theta)$ and we define $\gamma(A \times B)$ varying A $(A \mapsto \gamma(A \times B))$

$\Longrightarrow$ $\gamma(\cdot \times B)$ is a measure on $\mathcal{B}(\mathbb{R}^n)$

and we want to prove that :

$$\gamma(\cdot \times B) \ll \mu_n \qquad (= \gamma(\cdot \times B) \text{ is absolutely continuous w.r.t. } \mu)$$

(proof of it:)

consider A s.t. $\mu_n(A) = 0$.

$\underbrace{\gamma(A \times B) \leq \gamma(A \times \Theta)}_{\text{since } A \times B \subset A \times \Theta} \overset{\text{def.}}{=} \mu_n(A) = 0$ ← assumptions

$\Longrightarrow \gamma(A \times B) = 0$

$\Longrightarrow \left[\mu_n(A) = 0 \Longrightarrow \gamma(A \times B) = 0 \quad \forall B \in \mathcal{B}(\Theta) \right]$ ■

$\Longrightarrow$ we apply R-N theorem:

$\forall B$ fixed $\exists$ function $\pi(\underline{x}, B)$ s.t. $\gamma(A \times B) = \int_A \pi(\underline{x}, B)\, \mu_n(d\underline{x})$ (*)

$\Longrightarrow \gamma(A \times B) = \int_A \pi(\underline{x}, B)\, m_n(\underline{x})\, \lambda^{(n)}(d\underline{x})$

$= \int_A \boxed{\pi(\underline{x}, B)} \int_\Theta f(\underline{x}|\theta)\, \pi(d\theta)\, \lambda^{(n)}(d\underline{x})$

$\pi(\underline{x}, B) \longrightarrow \mathbb{P}(\theta \in B \mid \underline{X} = \underline{x})$

the very definition of conditional probability that we gave before satisfies an integral equation. In this case the equation would be:

$$\mathbb{P}(\underline{X} \in A, \theta \in B) = \int_A \overbrace{\boxed{\mathbb{P}(\theta \in B \mid \underline{X} = \underline{x})}}^{(*)} \mu_n(d\underline{x})$$

If this is the case then (*) is the def. of conditional probability.

Here we look at (*) and since:

$\gamma(A \times B) = \mathbb{P}(\underline{X} \in A, \theta \in B)$

$\Longrightarrow \mathbb{P}(\theta \in B \mid \underline{X} = \underline{x}) = \pi(\underline{x}, B)$

$\Longrightarrow \gamma(A \times B) = \int_A \mathbb{P}(\theta \in B \mid \underline{X} = \underline{x}) \int_\Theta f(\underline{x}|\theta)\, \pi(d\theta)\, \lambda^{(n)}(d\underline{x})$ *

$\Longrightarrow$ comparing the two *, since the integrals are equal we can derive:

$$\int_B f(\underline{x}|\theta)\, \pi(d\theta) = \underbrace{\mathbb{P}(\theta \in B \mid \underline{X} = \underline{x})}_{\text{the unknown}} \int_\Theta f(\underline{x}|\theta)\, \pi(d\theta) \quad \text{a.s. w.r.t. } \mu_n$$

$$\Longrightarrow \mathbb{P}(\theta \in B \mid \underline{X}_u = \underline{x}) = \frac{\int_B f(\underline{x}|\theta)\, \pi(d\theta)}{\int_\Theta f(\underline{x}|\theta)\, \pi(d\theta)} \quad \text{a.s. w.r.t. } \mu_n$$

It's easy to prove:

$C = \{ \underline{x} : \int_\Theta f(\underline{x}|\theta)\, \pi(d\theta) = m_n(\underline{x}) = 0 \}$ is s.t. $\mu_n(C) = 0$.

($\int_C m_n(\underline{x})\, \mu_n(d\underline{x})$ but over C $m_n(\underline{x}) = 0 \Longrightarrow \int_C \ldots = 0$) ■

$\Longrightarrow$ This will be the formula we'll always use.
We only have to assume that the conditional distribution of the vector of data given ϑ $(\underline{X}|\vartheta)$ is dominated.

$\exists$ a density w.r.t. some fixed measures

- How can we write Bayes theorem if ϑ has a prior?
$\underline{\vartheta \sim \pi(\vartheta)}$ = prior of ϑ (w.r.t. some fixed measure)

equivalent to $X \sim f(x)$

$$\Longrightarrow \quad \boxed{\pi(\vartheta|\underline{x}) = \frac{f(\underline{x}|\vartheta)\,\pi(\vartheta)}{\int_\Theta f(\underline{x}|\vartheta)\,\pi(\vartheta)\,d\vartheta}}$$

$\int_\Theta f(\underline{x}|\vartheta)\,\pi(\vartheta)\,d\vartheta \longrightarrow$ marginal density of the data

- Suppose that $X_1, \dots, X_n | \vartheta \overset{iid}{\sim} f_1(\cdot|\vartheta)$ $\longrightarrow$ in this case the joint distribution of the data will be the product

$\vartheta \sim \pi$ prior probability

$\Longrightarrow$ posterior density:

$$\boxed{\pi(\vartheta|\underline{x}) = \frac{\prod_{i=1}^n f_1(x_i|\vartheta)\,\pi(\vartheta)}{\int_\Theta \prod_{i=1}^n f_1(x_i|\vartheta)\,\pi(\vartheta)\,d\vartheta}}$$

$\longrightarrow$ **BAYESIAN APPROACH**

The posterior density (assuming that this density exists) is a ratio; at the numerator we have the likelihood times the density and at the denominator there is the marginal density of the sample (= integral of the numerator saturated w.r.t. ϑ).

$\Longrightarrow$ The posterior distribution is proportional to the likelihood times the prior (we don't care too much about the denominator: if we multiply the likelihood for the prior then we only have to normalize it, but we know that the density exists).

Def. (Informal def. of conjugate prior)

Suppose $X|\theta \sim f(x|\theta)$

$\theta \sim \pi$

We have that the prior π belongs to a family $\mathcal{F}$: $\pi \in \mathcal{F}$.

Suppose moreover that $\pi(\cdot|x) \in \mathcal{F}$.

$\Longrightarrow$ the prior is called conjugate to the method

Ex. 1 Bernoulli - Beta model

We assume $X_1, \dots, X_n|\theta \overset{iid}{\sim} Be(\theta)$:

$$\mathbb{P}(X_i = 1) = \theta$$
$$\mathbb{P}(X_i = 0) = 1-\theta$$

$\Longrightarrow$ $x = (x_1, \dots, x_n)$ is s.t. $x_i \in \{0,1\}$

We assume as a prior for θ: $\theta \sim Beta(\alpha, \beta)$

Beta density: $\pi(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbb{1}_{(0,1)}(\theta)$, $\alpha, \beta > 0$

this normalizing constant is: $= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$

where $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$

gamma function

where $\Gamma(n) = (n-1)!$ if n is an integer

$\Gamma(\frac{1}{2}) = \sqrt{\pi}$

What is the posterior distribution of this model?

Bayes' theorem:

$$\pi(\theta|x) = \frac{\prod_{i=1}^n f(x_i|\theta)\,\pi(\theta)}{\int_\Theta \prod_{i=1}^n f(x_i|\theta)\,\pi(\theta)\,d\theta}$$

$$= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} \quad \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbb{1}_{(0,1)}(\theta)}{\int_0^1 (\cdot)\, d\theta}$$

$$= \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbb{1}_{(0,1)}(\theta)}{\int_0^1 (\cdot)\, d\theta}$$

$$\Longrightarrow \int_0^1 \theta^{\alpha+\sum_{i=1}^n x_i - 1} (1-\theta)^{\beta+n-\sum_{i=1}^n x_i - 1} d\theta$$

usual trick: kernel of a beta distribution with (new) parameters:

$$Beta\left(\alpha + \underbrace{\textstyle\sum_{i=1}^n x_i}_{\#\text{ success}},\ \beta + \underbrace{\left(n - \textstyle\sum_{i=1}^n x_i\right)}_{\#\text{ non success}}\right)$$

We only have to adjust the normalizing constant:

$$\Rightarrow \quad \cdot \frac{1}{B\left(\alpha + \sum_{i=1}^n x_i,\ \beta + \left(n - \sum_{i=1}^n x_i\right)\right)}$$

$$\Rightarrow \int_0^1 (\cdot)\, d\theta = B\left(\alpha + \sum_{i=1}^n x_i,\ \beta + \left(n - \sum_{i=1}^n x_i\right)\right) \cdot \underbrace{\left(\frac{1}{B(\cdot,\cdot)} \int_0^1 (\cdot)\right)}_{1}$$

$$= \frac{\Gamma\left(\alpha + \sum_{i=1}^n x_i\right)\Gamma\left(\beta + \left(n - \sum_{i=1}^n x_i\right)\right)}{\Gamma(\alpha + \beta + n)}$$

$$\Rightarrow \pi(\theta \mid \underline{x}) = \underbrace{\frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha + \sum x_i)\,\Gamma(\beta + n - \sum x_i)}\, \theta^{\alpha + \sum x_i - 1} (1-\theta)^{\beta + n - \sum x_i}}_{\text{Beta distribution}}\, \mathbb{1}_{(0,1)}(\theta)$$

$$\Rightarrow \pi(\theta \mid \underline{x}) \text{ is a } \text{Beta}\left(\alpha + \sum_{i=1}^n x_i,\ \beta + \left(n - \sum_{i=1}^n x_i\right)\right)$$

$$\Rightarrow \begin{aligned} &\text{A priori:} && \theta && \sim && \text{Beta}(\alpha, \beta) \\ &\text{A posterior:} && \theta \mid \underline{x} && \sim && \text{Beta}\left(\alpha + \sum_{i=1}^n x_i,\ \beta + n - \sum_{i=1}^n x_i\right) \end{aligned}$$

$\rightarrow$ Since the posterior is still a Beta, this is an example of conjugate prior w.r.t. the model and the update of the prior informations (= the transformation from the prior to the posterior) is "easy":
we only have to change the hyperparameters of the distribution

(Comment on the example)

23/09

$X_1, \dots, X_n \mid \theta \overset{iid}{\sim} Be(\theta)$ $\quad \theta \in [0,1]$

$\theta \sim Beta(\alpha, \beta)$ $\quad \alpha, \beta > 0$

$\Rightarrow \theta \mid \underbrace{x_1, \dots, x_n}_{\text{observed samples}} \sim Beta\left(\alpha + \sum_{i=1}^n x_i \,,\, \beta + n - \sum_{i=1}^n x_i\right)$

! **NOTE:** "hyperparameters" are the parameters of the prior distribution

Prior: α, β

Posterior: $\alpha + \sum_{i=1}^n x_i \,,\, \beta + n - \sum_{i=1}^n x_i$

$$\underbrace{\mathbb{E}[\theta \mid \underline{x}]}_{\text{posterior mean}} = \frac{\alpha + \sum_{i=1}^n x_i}{\alpha + \beta + n} = \underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{prior mean}} \cdot \frac{\alpha+\beta}{\alpha+\beta+n} + \frac{n}{\alpha+\beta+n} \cdot \underbrace{\frac{\sum_{i=1}^n x_i}{n}}_{\text{maximum likelihood estimate of } \theta}$$

$\Rightarrow \mathbb{E}[\theta \mid \underline{x}] = w_n \, \mathbb{E}_\pi[\theta] + (1 - w_n)\, \bar{x}_n$ $\Rightarrow$ the posterior mean is a convex linear combination (of $\mathbb{E}_\pi[\theta]$ and $\bar{x}_n$) with weights w_n and $(1 - w_n)$

$\mathbb{E}_\pi[\theta]$ = mean a priori of θ

However this behavior is common, it's not only for this class/example!

What happens if $n \to \infty$?

$$\frac{\alpha+\beta}{\alpha+\beta+n} \to 0, \quad \frac{n}{\alpha+\beta+n} \to 1$$

$\Rightarrow \mathbb{E}[\theta \mid \underline{x}] \approx \bar{x}_n$ (large) $\Rightarrow$ when we have enough data the prior tends to vanish (w.r.t. the data)

actually this is valid also if $\alpha+\beta$ are small w.r.t. n (whatever n)

In this case we say that the prior is NON INFORMATIVE — meaning that: a posteriori the data weights more than the prior information that we're assuming

$$Var(\theta \mid \underline{x}) = \frac{(\alpha + \sum_{i=1}^n x_i)(\beta + n - \sum_{i=1}^n x_i)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \mathbb{E}[\theta \mid \underline{x}]\, \mathbb{E}[1-\theta \mid \underline{x}] \cdot \frac{1}{\alpha+\beta+n+1}$$

$\Rightarrow Var(\theta \mid \underline{x}) \xrightarrow{n \to \infty} 0$ $\Rightarrow$ with the growing of n the posterior becomes more and more concentrated around the expectation (= more precise)

Ex. We toss a coin n times, $\theta = \mathbb{P}(T)$ $\quad (T = \text{tail} \,;\, 1 \to \text{tail})$

If we don't know anything on θ $\Rightarrow$ uniform ← somehow represent the non-informativeness of θ

(Notice: uniform = beta(1,1))

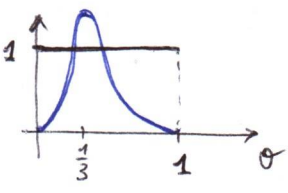
$\Rightarrow \theta \sim Beta(1,1)$

$\mathbb{E}[\theta] = \frac{1}{2}$ (prior expectation), $\quad Var(\theta) = \frac{1}{12}$

we suppose to toss 10 times → 3T

$\Rightarrow \theta \mid x_1, \dots, x_{10} \sim Beta(1+3, 1+7) = Beta(4,8)$

$\mathbb{E}[\theta \mid \underline{x}] = \frac{4}{12} = \frac{1}{3}$, $\quad Var(\theta \mid \underline{x}) = \frac{4 \cdot 8}{(12)^2 \cdot 13} = \frac{2}{117}$

$\Rightarrow$ 

— prior distr. of θ
— posterior distr. of θ $(\theta \mid \underline{x})$

$\Rightarrow$ we updated the parameters and we decreased the uncertainty around the most plausible values

3 Fundamental problems in inference:

- point estimation
- interval estimation
- hypothesis testing

BAYESIAN POINT ESTIMATION

$$\underline{X}|\theta \sim f(\underline{x}|\theta) \qquad \theta \sim \pi \qquad (\theta \in \Theta \subset \mathbb{R}^k) \quad \Longrightarrow \quad \pi(\theta|\underline{x})$$

We want an estimation of θ.

Def. Θ := states of the nature (parametric space)

$\underline{a}$:= set of all possible actions available to the statistician $(a \in A)$

$\underline{\ell(\theta, a)}$:= loss to the statistician from taking action a while the state of the nature is Θ

example: $\ell(\theta, a) = (\theta - a)^2$ → if we estimate "θ" with "a" then the loss will be $(\theta-a)^2$

θ is random $\Longrightarrow$ we consider the prior expected loss: $\mathbb{E}_\pi[\ell(\theta, a)]$

($\mathbb{E}_\pi$ refers to θ) → prior expected loss (since θ is random)

The posterior expected loss is: $\mathbb{E}_\pi[\ell(\theta, a)|\underline{x}]$

The Bayesian rule is the following:

a^* is such that:

$$\boxed{\mathbb{E}_\pi[\ell(\theta, a^*)|\underline{x}] = \min_{a \in A} \underline{\mathbb{E}_\pi[\ell(\theta, a)|\underline{x}]}}$$

→ this is a function only of a

Ex. $\underline{\ell(\theta; a) = (\theta - a)^2} \Longrightarrow$ Bayes rule gives: $\boxed{a^* = \mathbb{E}[\theta|\underline{x}]}$

BAYESIAN INTERVALS ESTIMATION

→ here we can say that the probability of θ being in our interval is ... (with the frequentist approach is not like that)

$$\underline{X}|\theta \sim f(\underline{x}|\theta) \qquad \theta \sim \pi \quad \Longrightarrow \quad \pi(\theta|\underline{x})$$

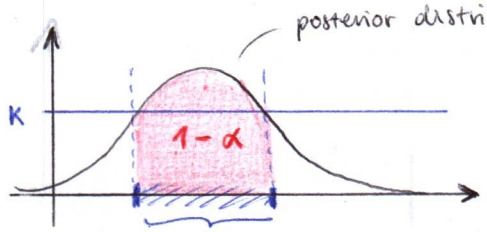
CREDIBLE REGIONS = region of Θ where the posterior distribution puts most of its mass ($(1-\alpha)\cdot 100\%$ of its mass)

Def. $\alpha \in (0,1)$: a region C of Θ is a $100\cdot(1-\alpha)\%$ posterior credibility region if:

$$\mathbb{P}(\theta \in C|\underline{x}) \geq 1-\alpha$$

Def. Let the posterior density $\pi(\theta|\underline{x})$ be unimodal. We define a region C (of Θ) to be a posterior $100\cdot(1-\alpha)\%$ highest probability density (HPD) region for θ if:

$$C = \{\theta \in \Theta : \pi(\theta|\underline{x}) \geq k\} \quad \text{with} \quad k: \ \mathbb{P}(\theta \in k|\underline{x}) = 1-\alpha$$



which k are we going to choose? We choose k in such a way that the mass under the posterior probability on the HPD support is $1-\alpha$

those are the values that define the HPD region of θ

MCMC method: interval estimate of θ? We compute the quantiles of marginal posterior densities

credible region of level 90%? → quantiles $q_{0.05}$, $q_{0.95}$ → CR

HYPOTHESIS TESTING (BAYESIAN)

$$\begin{cases} H_0: \theta \in \Theta_0 \\ H_1: \theta \in \Theta_1 \end{cases} \quad \text{where } \Theta = \Theta_1 \cup \Theta_0$$

(they're disjoint and have the same dimensionality)

$\underline{X}|\theta \sim f(\underline{x}|\theta)$

$\pi(\theta)$ prior distribution

in general they can be different

g_0 = prior density conditioning to $\theta \in \Theta_0$: $\int_{\Theta_0} g_0(\theta)\,d\theta = 1$

g_1 = prior density conditioning to $\theta \in \Theta_1$: $\int_{\Theta_1} g_1(\theta)\,d\theta = 1$

$\pi_0 := \pi(\Theta_0)$ prior probability that $\theta \in \Theta_0$ $\Longrightarrow$ $\pi_1 = 1 - \pi_0 = \pi(\Theta_1)$

Prior density: $\pi(\theta) = \pi_0\, g_0(\theta)\, \mathbb{1}_{\Theta_0}(\theta) + (1-\pi_0)\, g_1(\theta)\, \mathbb{1}_{\Theta_1}(\theta)$

- Posterior distribution?

Marginal: $m_\pi(\underline{x}) = \int_\Theta f(\underline{x}|\theta)\,\pi(\theta)\,d\theta$

$$= \int_{\Theta_0} \pi_0\, g_0(\theta)\, f(\underline{x}|\theta)\,d\theta + \int_{\Theta_1} (1-\pi_0)\, g_1(\theta)\, f(\underline{x}|\theta)\,d\theta$$

Bayes' $\Longrightarrow$

$$\pi(\theta|\underline{x}) = \begin{cases} \pi_0\, \dfrac{g_0(\theta)\, f(\underline{x}|\theta)}{m_\pi(\underline{x})} & \text{if } \theta \in \Theta_0 \\ (1-\pi_0)\, \dfrac{g_1(\theta)\, f(\underline{x}|\theta)}{m_\pi(\underline{x})} & \text{if } \theta \in \Theta_1 \end{cases}$$

Which one between Θ_0 and Θ_1 we choose? The one for which $\mathbb{P}(\theta \in \Theta_i|\underline{x})$ is higher

$\mathbb{P}(\theta \in \Theta_0|\underline{x}) > \mathbb{P}(\theta \in \Theta_1|\underline{x}) \Longrightarrow$ we chose Θ_0, otherwise Θ_1

$$\text{POSTERIOR ODDS} := \frac{\mathbb{P}(\theta \in \Theta_0|\underline{x})}{\mathbb{P}(\theta \in \Theta_1|\underline{x})} = \frac{\pi_0 \int_{\Theta_0} f(\underline{x}|\theta)\, g_0(\theta)\,d\theta}{(1-\pi_0)\int_{\Theta_1} f(\underline{x}|\theta)\, g_1(\theta)\,d\theta}$$

→ we choose Θ_0 if this ratio is $\gg 1$, otherwise Θ_1 (if $\ll 1$)

$$\frac{\text{posterior odds}}{\text{prior odds}} = \text{BAYES FACTOR} = \frac{\int_{\Theta_0} f(\underline{x}|\theta)\, g_0(\theta)\,d\theta}{\int_{\Theta_1} f(\underline{x}|\theta)\, g_1(\theta)\,d\theta} = BF_{01}$$

we want to get rid of π_0! so we analyze how much the posterior odds changes w.r.t. the prior odds

marginal distribution of the data conditioning to the set $\theta \in \Theta_0$

analogamente ma con $\theta \in \Theta_1$

$BF_{01} \gg 1 \Longrightarrow H_0$

$BF_{01} \ll 1 \Longrightarrow H_1$

"Recipe" for the BF_{01}: $BF_{01} \rightsquigarrow 2\log(BF_{01})$

If $2\log(BF_{01}) \in (0,2) \Rightarrow$ weak evidence in favor of H_0

If $2\log(BF_{01}) \in (2,5) \Rightarrow$ evidence in favor of H_0

If $2\log(BF_{01}) \in (5,10) \Rightarrow$ strong evidence in favor of H_0

If $2\log(BF_{01}) > 10 \Rightarrow$ very strong evidence in favor of H_0

If $2\log(BF_{01}) < 0$?

$$BF_{10} = \frac{1}{BF_{01}} \Rightarrow 2\log(BF_{10}) = -2\log(BF_{01})$$

(In this case the two hypothesis are totally symmetric)

PREDICTION under the BAYESIAN APPROACH

We collected $X_1, \dots, X_n$.

We want to predict $X_{n+1}, \dots, X_{n+m}$.

Example: we collected $X_1, \dots, X_n$ where $X_i = 1$ if the patient is infected, we want to predict the next m patients

We compute the PREDICTIVE DISTRIBUTION: $\mathcal{L}(X_{n+1}, X_{n+2}, \dots, X_{n+m} \mid X_1, \dots, X_n)$

We assume: $X_1, \dots, X_n, X_{n+1}, \dots, X_{n+m} \mid \vartheta \overset{iid}{\sim} f(\cdot\mid\vartheta)$, $\vartheta \sim \pi$

Attention to what this means!

Conditionally does not mean that marginally are $\perp\!\!\!\perp$. If they were $\perp\!\!\!\perp$ marginally we would not learn anything by collecting data.

$X_1, \dots, X_n \mid \vartheta \overset{iid}{\sim} f_1(\cdot\mid\vartheta)$, $\vartheta \sim \pi$ $\Rightarrow$ predictive density of X_{n+1} given $X_1 = x_1, \dots, X_n = x_n$.

This is called POSTERIOR PREDICTIVE DENSITY of X_{n+1} given $X_1 = x_1, \dots, X_n = x_n$

$(:= \mathcal{L}(X_{n+1} \mid X_1, \dots, X_2))$

$$m_{X_{n+1}\mid X_1=x_1,\dots,X_n=x_n}(x) = \frac{m_{X_1,\dots,X_{n+1}}(x_1,\dots,x_{n+1})}{m_{X_1,\dots,X_n}(x_1,\dots,x_n)}$$

$$= \frac{\int_\Theta \prod_{i=1}^n f_1(x_i\mid\vartheta)\, f_1(x\mid\vartheta)\, \pi(d\vartheta)}{\int_\Theta \prod_{i=1}^n f_1(x_i\mid\vartheta)\, \pi(d\vartheta)}$$

this is the (integral of the) joint distribution of all the random variables and ϑ and then we're marginalizing w.r.t. ϑ (*)

$$= \int_\Theta f_1(x\mid\vartheta)\, \pi(d\vartheta \mid X_1 = x_1, \dots, X_n = x_n)$$

(*) $X_1 \mid \vartheta \sim f_1(x\mid\vartheta)$, $\vartheta \sim \pi$

the marginal of X_1: $m_{X_1}(x) = \int_\Theta \mathcal{L}(x_1\mid\vartheta)\, \mathcal{L}(d\vartheta)$

Ex. $X_1, \dots, X_n \mid \vartheta \overset{iid}{\sim} Be(\vartheta)$, $\vartheta = \mathbb{P}(T)$, $\vartheta \sim Beta(1,1)$ (uniform)

We toss $n = 10$ times and $\sum_{i=1}^{10} x_i = 3$. $\mathbb{P}(X_{11} = 1 \mid X_1 = x_1, \dots, X_n = x_n)$?

$$\mathbb{P}(X_{11} = 1 \mid X_1 = x_1, \dots, X_n = x_n) = \int_0^1 \mathbb{P}(X_{11} = 1 \mid \vartheta)\, \pi(\vartheta \mid \underline{x})\, d\vartheta$$

(where $\vartheta \mid \underline{x} \sim Beta(1+3, 1+7) = Beta(4,8)$)

$\mathbb{P}(X_{11} = 1 \mid \vartheta) = \vartheta$ (since $X_i \mid \vartheta \sim Be(\vartheta)$)

$$\Rightarrow \int_0^1 \mathbb{P}(X_{11}=1\mid\theta)\,\pi(\theta\mid \underline{x})\,d\theta = \int_0^1 \theta\cdot\pi(\theta\mid\underline{x})\,d\theta = \mathbb{E}[\theta\mid\underline{x}] = \frac{4}{12} = \frac{1}{3}$$

$(\mathbb{E}[\theta\mid x_1,\dots,x_{10}])$ $\quad \mathbb{P}(X_{11}=1\mid \underline{X}=\underline{x})$

$$\underset{\text{(prior predictive)}}{\mathbb{P}(X_{11}=1)} = \int_0^1 \mathbb{P}(X_{11}=1\mid\theta)\,\pi(\theta)\,d\theta = \int_0^1 \theta\,\pi(\theta)\,d\theta = \mathbb{E}_\pi[\theta] = \frac{1}{2}$$

$$\Rightarrow \mathbb{P}(X_{11}=1) \neq \mathbb{P}(X_{11}=1\mid X_{10}=x_{10},\dots,X_1=x_1)$$

$\left(\frac{1}{2}\right)$ $\qquad\qquad$ $\left(\frac{1}{3}\right)$

25/09

Remark: suppose to have two dataset $\underline{X}_1, \underline{X}_2$.

What is better? (Available together or one after the another)

- compute at once the posterior $\pi(\vartheta\mid \underline{X}_1, \underline{X}_2)$
- compute first the posterior using $\underline{X}_1$, then use it as prior and use data $\underline{X}_2$ to compute the second posterior?

It's the same! (at inference level)

$$\pi(\vartheta\mid \underline{X}_1,\underline{X}_2) = \frac{\mathcal{L}(\underline{X}_1,\underline{X}_2\mid\vartheta)\,\pi(\vartheta)}{\mathcal{L}(\underline{X}_1,\underline{X}_2)} = \frac{\mathcal{L}(\underline{X}_1\mid\vartheta)\,\mathcal{L}(\underline{X}_2\mid\vartheta)\,\pi(\vartheta)}{\mathcal{L}(\underline{X}_1)\cdot\mathcal{L}(\underline{X}_2\mid\underline{X}_1)}$$

assuming $\underline{X}_1, \underline{X}_2\mid\vartheta$ are independents $\Rightarrow$ they're exchangable

$$= \frac{\pi(\vartheta\mid\underline{X}_1)\cdot\mathcal{L}(\underline{X}_2\mid\vartheta)}{\mathcal{L}(\underline{X}_2\mid\underline{X}_1)}$$

here we have the posterior distribution using only $\underline{X}_2$ but we have as a prior we've using the posterior that we got using only $\underline{X}_1$.

$\Rightarrow$ the posterior using $\underline{X}_1, \underline{X}_2$ together is the same as the posterior using $\underline{X}_1$ and then $\underline{X}_2$

Ex. Suppose athlets are continuously tested for doping.

outcome of the test $\in \{\text{positive, negative}\} = \{E^+, E^-\}$

We can have:
- false positive: athled tested positive even if clean
- false negative: athled tested negative even if it's doped

H_D: "the athlet is doped", H_D^c: "the athlet is not"

given informations:

$\pi(H_D) = 0.03$

$\mathbb{P}(E^+\mid H_D^c) = 0.1 \Rightarrow 1-\mathbb{P}(E^+\mid H_D^c) = 0.9$ SPECIFICITY

$\mathbb{P}(E^-\mid H_D) = 0.05 \Rightarrow 1-\mathbb{P}(E^-\mid H_D) = 0.95$ SENSITIVITY

Suppose we get E^+: $\boxed{\mathbb{P}(H_D\mid E^+)?}$ $= \mathbb{P}(\text{true positive})?$

$$\mathbb{P}(H_D\mid E^+) = \frac{\mathbb{P}(E^+\mid H_D)\,\mathbb{P}(H_D)}{\mathbb{P}(E^+\mid H_D)\,\mathbb{P}(H_D) + \mathbb{P}(E^+\mid H_D^c)\,\mathbb{P}(H_D^c)}$$

$$= \frac{0.95\cdot 0.03}{0.95\cdot 0.03 + 0.1\cdot 0.97} \simeq \underline{0.227} < \frac{1}{2}$$

$\Rightarrow$ even if he's positive to doping the probability of being doped is still low

$\pi(H_D) = 0.03 \longrightarrow \mathbb{P}(H_D|E^+) = 0.227$

a priori — a posteriori

We suppose that someone guarantee for the tested guy (saying he didn't do it) (we update the a priori ($:= \gamma$))

$\pi_\gamma(H_D) = 0.001 \xrightarrow{\text{Bayes}} \pi_\gamma(H_D|E^+) \simeq 0.009 \Rightarrow$ according to the prior we have different posterior

Suppose we collect more data about the same athlet: $\longrightarrow$ E_2^+ = the athlet was tested positive also the 2nd time

using prior: $\pi(H_D) = 0.227^* \Rightarrow \pi(H_D|E_2^+) \simeq 0.736$

$\pi_\gamma(H_D) = 0.009^* \Rightarrow \pi_\gamma(H_D|E_2^+) \simeq 0.079$

* this come from the first posterior (that we now use as new prior)

(This is equivalent to use both data at once)

$E_3^+ \longrightarrow \pi(H_D|E_3^+) = 0,96$, $\pi_\gamma(H_D|E_3^+) = 0.45$

$E_4^+ \longrightarrow \pi(H_D|E_4^+) = 0.996$, $\pi_\gamma(H_D|E_4^+) = 0.88$

$E_5^+ \longrightarrow \pi(H_D|E_5^+) = \underline{0.9995}$, $\pi_\gamma(H_D|E_5^+) = \underline{0.99}$

If the data size increase, even if the priors are $\neq$ the posterior values will merge

! **Remark:** Suppose to have $\underline{X}_1, \underline{X}_2$ that: $\underline{X}_1|\theta \perp\!\!\!\perp \underline{X}_2|\theta$.

This does not imply that $\underline{X}_1$ and $\underline{X}_2$ are (marginally) independent.

(Also neither "$\Leftarrow$" is valid)

Counterexample: $X_1, X_2|\theta \overset{iid}{\sim} Be(\theta)$

$\theta \sim Beta(\alpha, \beta) \quad \alpha, \beta > 0$

We want to prove $X_1 \not\perp\!\!\!\perp X_2$.

If X_1 and X_2 were $\perp\!\!\!\perp \Rightarrow Cov(X_1, X_2) = 0$

$Cov(X_1, X_2) = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$

$\mathbb{E}[X_i] = \mathbb{E}[\mathbb{E}[X_i|\theta]] = \mathbb{E}[\theta] \left(= \frac{\alpha}{\alpha+\beta}\right) \quad (X_i|\theta \sim Be(\theta))$

$\mathbb{E}[X_1X_2] = \mathbb{E}[\mathbb{E}[X_1X_2|\theta]]$

$\downarrow \mathbb{E}[\mathbb{E}[X_1|\theta]\,\mathbb{E}[X_2|\theta]]$ — conditioning on θ X_1 and X_2 are $\perp\!\!\!\perp$

$\downarrow \mathbb{E}[\theta^2]$

$\Rightarrow Cov(X_1, X_2) = \mathbb{E}[\theta^2] - (\mathbb{E}[\theta])^2 = Var(\theta) > 0$

$Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

NORMAL NORMAL MODEL

We assume $X_1, \dots, X_n|\mu \overset{iid}{\sim} N(\mu, \sigma_0^2)$ (σ_0^2 known), $\Theta = \mathbb{R}$

$\mu \sim N(\mu_0, \tau^2)$ ($\mu_0 \in \mathbb{R}$ fixed, $\tau^2 > 0$ fixed)

In this case, what will be the distribution for the posterior?

likelihood:

$$f(\underline{x}|\mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma_0^2}}$$

$$\downarrow = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n e^{-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}}$$

note: $\sum_{i=1}^n (x_i-\mu)^2 = \sum_{i=1}^n (x_i \pm \bar{x} - \mu)^2$

$$\Longrightarrow \quad f(\underline{x}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n e^{-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^n x_i^2 - n(\bar{x})^2\right)} \; e^{-\frac{n}{2\sigma_0^2}(\mu-\bar{x})^2}$$

$$\pi(\mu|x_1,\dots,x_n) = \frac{\left[\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n e^{-(\sum_{i=1}^n x_i^2 - n(\bar{x})^2)/2\sigma_0^2} \; e^{-n/2\sigma_0^2\cdot(\mu-\bar{x})^2}\right]\left[\frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\mu-\mu_0)^2}{2\tau^2}}\right]}{\int_{\mathbb{R}}\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)^n e^{-(\sum_{i=1}^n x_i^2 - n(\bar{x})^2)/2\sigma_0^2} \; e^{-n/2\sigma_0^2\cdot(\mu-\bar{x})^2} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\mu-\mu_0)^2}{2\tau^2}} d\mu}$$

posterior

$$= \frac{e^{-\frac{1}{2}\left(\frac{n}{\sigma_0^2}(\mu-\bar{x})^2 + \frac{1}{\tau^2}(\mu-\mu_0)^2\right)}}{\int_{\mathbb{R}} (\text{same}\uparrow)\, d\mu}$$

! $d_1(z-c_1)^2 + d_2(z-c_2)^2 = (d_1+d_2)(z-c)^2 + \frac{d_1 d_2}{d_1+d_2}(c_1-c_2)^2$!

Useful for the written test

$$\downarrow$$

$$c = \frac{d_1 c_1 + d_2 c_2}{d_1 + d_2}$$

proof.

$$= d_1 z^2 + d_1 c_1^2 - d_1 c_1 z\cdot 2 + d_2 z^2 + d_2 c_2^2 - 2 d_2 c_2 z$$

$$= (d_1+d_2) z^2 - 2z(d_1 c_1 + d_2 c_2) + d_1 c_1^2 + d_2 c_2^2$$

$$= (d_1+d_2)\left[z^2 - \frac{2z(d_1 c_1 + d_2 c_2)}{d_1+d_2} + \frac{d_1 c_1^2 + d_2 c_2^2}{d_1+d_2}\right]$$

$$= (d_1+d_2)\left[z^2 - 2zc + \frac{d_1 c_1^2 + d_2 c_2^2}{d_1+d_2} + c^2 - c^2\right]$$

$$= (d_1+d_2)\left[(z-c)^2 + \frac{d_1 c_1^2 + d_2 c_2^2}{d_1+d_2} - \frac{(d_1 c_1 + d_2 c_2)^2}{(d_1+d_2)^2}\right]$$

$$= [\dots]$$

■

$$\Longrightarrow \quad \frac{n}{\sigma_0^2}(\mu-\bar{x})^2 + \frac{1}{\tau^2}(\mu-\mu_0)^2 = \left(\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}\right)(\mu-\mu_n)^2 + \frac{\frac{n}{\sigma_0^2}\cdot\frac{1}{\tau^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}}(\bar{x}-\mu_0)^2$$

$$\Longrightarrow \quad \mu_n = \frac{\frac{n}{\sigma_0^2}\bar{x} + \frac{1}{\tau^2}\mu_0}{\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}} = \frac{\frac{n\tau^2\bar{x} + \sigma_0^2\mu_0}{\sigma_0^2\tau^2}}{\frac{n\tau^2+\sigma_0^2}{\sigma_0^2\tau^2}}$$

$$= \frac{n\tau^2\bar{x} + \sigma_0^2\mu_0}{n\tau^2+\sigma_0^2}$$

$$\frac{\frac{n}{\sigma_0^2}\cdot\frac{1}{\tau^2}}{\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}} \rightarrow \frac{n}{n\tau^2+\sigma_0^2}$$

$$\Longrightarrow \quad \pi(\mu|x_1,\dots,x_n) = \frac{e^{-\frac{1}{2}\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}\right)(\mu-\mu_n)^2} \; e^{-\frac{1}{2}\frac{n}{n\tau^2+\sigma_0^2}(\bar{x}-\mu_0)^2}}{\int_{\mathbb{R}} \underbrace{e^{-\frac{1}{2}\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}\right)(\mu-\mu_n)^2}}_{\text{kernel of a gaussian}} \; e^{-\frac{1}{2}(\dots)} d\mu}$$

$$= \frac{e^{-\frac{1}{2}\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}\right)(\mu-\mu_n)^2}}{\sqrt{2\pi\cdot\left(\frac{1}{\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}}\right)}} = \frac{e^{-\frac{1}{2}\frac{(\mu-\mu_n)^2}{\sigma_0^2\tau^2/(n\tau^2+\sigma_0^2)}}}{\sqrt{2\pi\cdot\frac{\sigma_0^2\tau^2}{n\tau^2+\sigma_0^2}}}$$

$$\Longrightarrow \quad \boxed{\sim N\left(\mu_n, \frac{\sigma_0^2}{n\tau^2+\sigma_0^2}\tau^2\right)}$$

the posterior is still a gaussian, we only need to update the parameters

Prior

$\mathbb{E}[\mu] = \mu_0$

$Var(\mu) = \tau^2$

Posterior

(*) $\mathbb{E}[\mu|\underline{x}] = \mu_n = \frac{n\tau^2}{n\tau^2+\sigma_0^2}\bar{x} + \frac{\sigma_0^2}{n\tau^2+\sigma_0^2}\mu_0$

(**) $Var(\mu|\underline{x}) = \frac{\sigma_0^2}{n\tau^2+\sigma_0^2}\tau^2 < \tau^2$ ($\frac{\sigma_0^2}{n\tau^2+\sigma_0^2} < 1$)

Marginal density:

$$m(x_1,\dots,x_n) = \int_{\mathbb{R}} e^{-\frac{n}{\sigma_0^2}(\mu-\bar{x})^2 - \frac{1}{2\tau^2}(\mu-\mu_0)^2}\, d\mu$$

$$= e^{\frac{n(\bar{x}-\mu_0)^2}{2(n\tau^2+\sigma_0^2)}} \int_{\mathbb{R}} e^{-\frac{1}{2}\left(\frac{n}{\sigma_0^2}+\frac{1}{\tau^2}\right)(\mu-\mu_n)^2}\, d\mu$$

$$= \sqrt{2\pi\,\frac{\sigma_0^2\tau^2}{n\tau^2+\sigma_0^2}}\; e^{-\frac{n}{2(n\tau^2+\sigma_0^2)}(\bar{x}-\mu_0)^2}$$

if $n=1 \Longrightarrow$ $m(x) = \sqrt{2\pi\,\frac{\sigma_0^2\tau^2}{\sigma_0^2+\tau^2}}\; e^{-\frac{1}{2(\tau^2+\sigma_0^2)}(x-\mu_0)^2}$

(*) $\mathbb{E}[\mu|\underline{x}]$ is a convex linear combination of $\bar{x}$ (maximum likelihood estimator for μ) and μ_0 (prior mean). If n is large, the most weighted will be $\bar{x} \Longrightarrow$ when the sample size increase the bayesian estimate will be driven by the frequentist estimate.

(**) $Var(\mu|\underline{x})$ will generally be smaller than $Var(\mu)$: thanks to data we've able to decrease the uncertainty of the random variable μ

(from the case $n=1$):

$$\left.\begin{array}{l} X_1|\mu \sim N(\mu, \sigma_0^2) \\ \mu \sim N(\mu_0, \tau^2) \end{array}\right\} \Longrightarrow X_1 \sim N(\mu_0,\ \tau^2+\sigma_0^2)$$

30/09

$$\left.\begin{array}{l} X_1,\dots,X_n|\mu \overset{iid}{\sim} N(\mu, \sigma_0^2) \\ \mu \overset{iid}{\sim} N(\mu_0, \tau^2) \end{array}\right\} \Longrightarrow \mu|x_1,\dots,x_n \sim N(\mu_n, \tau_n^2)$$

$$\begin{cases} \mu_n = \frac{n\tau^2}{n\tau^2+\sigma_0^2}\bar{x}_n + \frac{\sigma_0^2}{n\tau^2+\sigma_0^2}\mu_0 \\ \tau_n^2 = \frac{\sigma_0^2\tau^2}{n\tau^2+\sigma_0^2} \end{cases}$$

If n is large: $\mathbb{E}[\mu|\underline{x}] \sim \bar{x}_n$

If τ^2 is large $\Longrightarrow$ same behaviour ($\mathbb{E}[\mu|\underline{x}] \sim \bar{x}_n$) $\Rightarrow$ a prior with large variance will be considered as "big prior" $\Rightarrow$ a posteriori the inference will be mostly driven by data than by prior informations

Excheangeability and the Bayesian approach

Alessandra Guglielmi

Politecnico di Milano
Dipartimento di Matematica
Milano, Italia
e-mail: alessandra.guglielmi@polimi.it

30 September 2020

Notation

- $\mathbb{X}$ is a separable and complete metric space; here $\mathbb{X} = \mathbb{R}$
- $\mathscr{X}$ is the Borel σ–algebra of subsets of $\mathbb{X}$
- $(X_n)_{n\geq 1}$ is a sequence of random elements defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbb{X}^\infty, \mathscr{X}^\infty)$

$(X_n)_{n\geq 1}$ is the sequence of observations (DATA); X_n is the result of the random experiment at trial n

- $\mathbf{P}_{\mathbb{X}}$ space of all probability measures on $(\mathbb{X}, \mathscr{X})$, with the topology of weak convergence
- $\mathscr{P}_{\mathbb{X}}$ Borel σ–algebra of subsets of $\mathbf{P}_X$
- A random element $\tilde{p}$ defined on $(\Omega, \mathscr{F}, \mathbb{P})$ and taking values in $(\mathbf{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$ is a *random probability measure*

Bayes' Theorem for dominated models

$$X_i|\tilde{\theta} = \theta \overset{\text{i.i.d.}}{\sim} f_\theta(\cdot) \text{ "true" density of each observation}$$
$$\tilde{\theta} \sim \pi \text{ probability on } \Theta \text{ Euclidean space}$$

Interpretation: $\theta \mapsto \prod_{i=1}^n f_\theta(x_i)$ likelihood, $\pi(d\theta)$ prior

Then the posterior distribution of $\tilde{\theta}$, given $X_1 = x_1, \ldots, X_n = x_n$, can be computed by **Bayes' Theorem**:

$$P(\tilde{\theta} \in B | X_1 = x_1, \ldots, X_n = x_n) \overset{a.s.}{=} \frac{\int_B \prod_{i=1}^n f_\theta(x_i)\pi(d\theta)}{\int_\Theta \prod_{i=1}^n f_\theta(x_i)\pi(d\theta)}, \quad B \in \mathscr{B}(\Theta)$$

Proof: definition of conditional distribution (as the solution of an integral equation) + Radon-Nikodym Theorem

The prior distribution

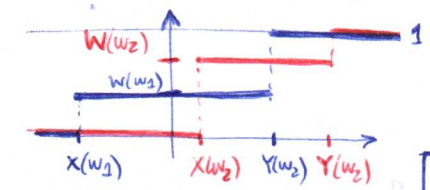
Making inference about $\tilde{\theta}$ is to learn about the unknown $\tilde{\theta}$ from the **DATA**: based on the data, explore which values of $\tilde{\theta}$ are probable, what might be plausible numbers as estimates of the different components of θ and the extent of uncertainty associated with such estimates

the distribution of $\tilde{\theta}$, **prior distribution**: it quantifies the uncertainty about $\tilde{\theta}$ *prior* to seeing data

The prior represents the subjective belief and knowledge $\mapsto$ subjective prior,

or, a conventional prior supposed to represent small or no information $\mapsto$ noninformative/vague/objective prior

* Example 2
X, Y, W r.v. $X, Y \in \mathbb{R}$, $W \in (0,1)$
$\tilde{p} := W\delta_X + (1-W)\delta_Y$
$\tilde{p}(A) = W\delta_X(A) + (1-W)\delta_Y(A)$
$w_1 \to X(w_1), Y(w_1), W(w_1)$
$w_2 \to X(w_2), Y(w_2), W(w_2)$



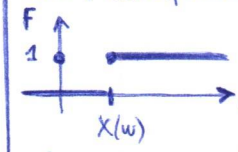
$\tilde{p}: (\Omega, \mathscr{F}) \to (\mathbb{P}_X, \mathscr{P}_X)$

$\tilde{p}(w)$ is a probability measure on $\mathbb{X}$ (in our case $\mathbb{R}$).

The simplex example of a random probability measure (r.p.m.) is "δ_X":

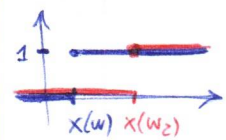
$$\delta_X(A) = \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{if } X \notin A \end{cases}$$

with X fixed random variable.
The F corresponding to this X:



because $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(X \in (-\infty, x]) = \begin{cases} 1 & \text{if } x \geq X(w) \\ 0 & \text{if } x < X(w) \end{cases}$

if we got a new X: $X(w_2)$:



They're called random probability measures because the support is random. *

Somehow it represents mathematically the notion of absence of informations about the data

Exchangeability

Suppose we toss a thumbtack and keep track of whether it come to stop with the point up ($X_i = 1$) or down ($X_i = 0$)

In the absence of any information to distinguish the tosses or to suggest that tosses occurring close together in time are more likely (or less likely) than those that are far apart in time, it is reasonable to treat the tosses symmetrically.

Under these conditions, in classical statistics we assume that data, conditionally on θ, are iid $Be(\theta)$, that is, for any $(x_1, \ldots, x_n) \in \{0,1\}^n$,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_{[0,1]} \theta^{\sum_1^n x_i}(1-\theta)^{n-\sum_1^n x_i}\pi(d\theta) \quad (1)$$

(1) does NOT depend on permutations of $(x_1, \ldots, x_n)$; e.g.

$$\mathbb{P}[X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0] = \mathbb{P}[X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1]$$

if the integral (1) is the real joint distribution then it's not important the position of 0/1s but only how many 0s and 1s there are

Exchangeable sequences

Exchangeability is a property concerning sequences of trials of a given phenomenon, all made under analogous conditions. When we assume exchangeability for the tosses of the thumbtack, this means that we treat the info obtained from any two tosses in exactly the same way as we treat the info from any other two tosses regardless of where they appear in the sequence of tosses, and so on for 3 or more tosses

Definition. *The random vector* $(X_1, X_2, \ldots, X_n)$ *is* exchangeable *if*

$$\mathcal{L}(X_1, \ldots, X_n) = \mathcal{L}(X_{\pi(1)}, \ldots, X_{\pi(n)}).$$

for any permutation π *of* $(1, \ldots, n)$.

the joint distribution of $(X_1, X_2, \ldots, X_n)$ is invariant under finite permutations

Definition. *The sequence* $(X_n)_{n\geq 1}$ *is* exchangeable *if*

$$\mathcal{L}(X_1, \ldots, X_n) = \mathcal{L}(X_{\pi(1)}, \ldots, X_{\pi(n)})$$

for any $n \geq 1$ *and permutation* π *of* $(1, \ldots, n)$.

an infinite sequence is exchangeable if the first n ($\forall n$) elements are exchang.

Exchangeable sequences

REMARK: if $(X_1, X_2, \ldots, X_n)$ is exchangeable, then

$$\mathcal{L}(X_i) = \mathcal{L}(X_1) \text{ for all } i$$
$$\mathcal{L}(X_1, X_2) = \mathcal{L}(X_i, X_j) \text{ for all } i \neq j$$
$$\mathcal{L}(X_1, X_2, X_3) = \mathcal{L}(X_i, X_j, X_k) \text{ for all distinct } i, j, k$$
$$\vdots$$

If the sequence (finite/infinite) of random variables is exchangeable then we have identity in distributions for any random variable (the marginal distr. of $\forall$ random variable is the same, the marginal distr. of $\forall$ couple of rand. variable is the same, etc. --)

- According to this definition, the order with which data are recorded is irrelevant for inferential purposes
- it is a *weak* assumption, and translates lack of (enough) information through a condition of *symmetry*
- For example, in thumbtack-tossing sequence one would have

$$\mathbb{P}[X_1 = 0, X_2 = 1, X_3 = 1, X_4 = 0] = \mathbb{P}[X_1 = 1, X_2 = 0, X_3 = 0, X_4 = 1]$$

why? suppose to have $n=2$ and $\mathcal{L}(X_1, X_2) = \mathcal{L}(X_2, X_1)$, (which means that if we get X_1 and then X_2 is the same as X_2 and then X_1 (if X_i represents somehow time)) and so if we marginalize and consider the marginal of the first component (in both $\mathcal{L}(X_1, X_2)$ and $\mathcal{L}(X_2, X_1)$) we get: $\mathcal{L}(X_1, X_2) \to \mathcal{L}(X_1)$, $\mathcal{L}(X_2, X_1) \to \mathcal{L}(X_2)$, $\Rightarrow \mathcal{L}(X_1) = \mathcal{L}(X_2)$

Exchangeable sequences - Examples

Example 1. Consider an urn that initially contains n_R red balls and $N - n_R$ blue balls. At each trial, we select a ball from the urn and then return the ball to the urn along with one new ball of the same color. In principle, this process can go on forever. Denote by X_i the random variable that takes value 1 if the outcome of the i-th draw is a red ball and 0 if blue. Then: $\{X_n\}_n$ is exchangeable.

Example 2. Consider an urn with 20 balls (14 red, 6 blue). Suppose we draw all the balls (one at a time) without replacement. Let $X_i = 1$ if the i-th ball is red and 0 otherwise. Then $(X_1, X_2, \ldots, X_{20})$ is exchangeable; however $(X_1, X_2, \ldots, X_{20})$ are NOT conditionally iid.

de Finetti's representation theorem for events

$(X_n)_{n\geq 1}$ exchangeable $\Leftrightarrow \exists\, \tilde\theta \in [0,1]$: $\tilde\theta \sim F$ (where F has the meaning of a prior distribution)*, $X_1, \dots, X_n \mid \tilde\theta \overset{iid}{\sim} Be(\tilde\theta)$

Theorem. *The sequence $(X_n)_{n\geq 1}$ of 0-1 r.v.'s is exchangeable if and only if there exists a probability measure F on $([0,1], \mathscr{B}([0,1]))$ such that*

$$\mathbb{P}[X_1 = x_1, \dots, X_n = x_n] = \int_{[0,1]} \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i} F(\mathrm{d}\theta)$$

= the joint distribution can be expressed as the integral of some parameter $\theta \in [0,1]$ in the form (←) (where $\theta \sim F(\cdot)$: θ is distributed as F)

for any $n \geq 1$ and $(x_1, \dots, x_n)$ in $\{0,1\}^n$.

Moreover, when $(X_n)_{n\geq 1}$ is exchangeable:

- $\frac{\sum_1^n X_i}{n} \overset{a.s.}{\to} \tilde\theta \sim F$ as $n \to +\infty$.
- Conditionally on $\tilde\theta$, $X_1, \dots, X_n | \tilde\theta \overset{i.i.d.}{\sim} Be(\tilde\theta)$ for all n; $\tilde\theta \sim F$.

⇒ assuming exchangeability (= invariance w.r.t. the order with which these data arrive) we can prove the existence of a random variable $\tilde\theta$ which is the success probability of each data (⇒ the data are iid Bernoulli distributed with parameter $\tilde\theta$ and $\tilde\theta$ is random, that's why F has the meaning of a prior distribution*)

⇒ simply assuming exchang. we are forced to be bayesian

Reinterpretation of the Bayesian paradigm...

... through exchangeability

✓ there is a formal equivalence between exchangeable trials of the same phenomenon (i.e. X_n's are binary) and those trials that are designated as "independent, with a fixed, but unknown, probability"

$(X_n)_{n\geq 1}$ exchang. 0-1 r.v.'s $\Leftrightarrow X_1, \dots, X_n | \tilde\theta \overset{i.i.d.}{\sim} Be(\tilde\theta)$ for all n; $\tilde\theta \sim F$ (Bernoulli trials)

✓ the infinite exchangeability is a keypoint here: finite exchangeable sequences have different representations

de Finetti's representation theorem (general case) (1933)

Again, we're saying $(X_n)_{n\geq 1}$ exchangeable $\Leftrightarrow \exists\, \tilde p$ random probability measure s.t. $X_1, \dots, X_n | \tilde p \overset{iid}{\sim} \tilde p$, $\tilde p \sim Q$

Theorem. *The sequence $(X_n)_{n\geq 1}$ is exchangeable if and only if there exists a probability measure Q on $(\mathbf{P}_{\mathbb{X}}, \mathscr{P}_{\mathbb{X}})$ such that*

$$\mathbb{P}[X_1 \in A_1, \dots, X_n \in A_n] = \int_{\mathbf{P}_{\mathbb{X}}} \prod_{i=1}^n p(A_i)\, Q(\mathrm{d}p)$$

for any $n \geq 1$ and $A_1, \dots, A_n$ in $\mathscr{X}$, where the probability Q is uniquely determined.

✏ $(X_n)_{n\geq 1}$ is exchangeable if and only if there exists a random probability measure $\tilde p$ on $(\mathbb{X}, \mathscr{X})$ such that $\tilde p \sim Q$ and

$$\mathbb{P}[X_1 \in A_1, \dots, X_n \in A_n \mid \tilde p] = \prod_{i=1}^n \tilde p(A_i)$$

for any $n \geq 1$ and $A_1, \dots, A_n$ in $\mathscr{X}$.

⇒ assuming exchangeability (simmetry on data) is equivalent to assume that those data conditionally to $\tilde p$ (r.p.m.) are iid according to $\tilde p$ and $\tilde p$ has a distribution

de Finetti's representation theorem (general case)

Q is a probability measure on $\mathbf{P}_{\mathbb{X}}$: *de Finetti measure* of $(X_n)_{n\geq 1}$

✏ If $(X_n)_{n\geq 1}$ is exchangeable, then its empirical distribution is such that

$$\frac{1}{n}\sum_{i=1}^n \delta_{X_i} \Rightarrow \tilde p \qquad \text{a.s.–}\mathbb{P}$$

the empirical distribution (empirical measure) gives the mass $1/n$ on each datum

where $\Rightarrow$ denotes *weak convergence*.

✏ Hierarchical representation: $(X_n)_{n\geq 1}$ exchangeable is equivalent to

$X_i \mid \tilde p \overset{i.i.d.}{\sim} \tilde p$ "true" distribution of each observation

$\tilde p \sim Q$ prior distribution

The Bayesian nonparametric framework is equivalent to exchangeability of $(X_n)_n$

Parametric case through the representation theorem

Parametric model: Q degenerate on a finite–dimensional subset $\mathbf{P}^*_{\mathbb{X}}$ of $\mathbf{P}_{\mathbb{X}}$, such that

$$Q(\{\mathbf{P}^*_{\mathbb{X}}\}) = Q(\{p \in \mathbf{P}_{\mathbb{X}} : \ p = p_\theta, \theta \in \Theta\}) = 1$$

and there exists a function $\tilde{\theta} : \mathbf{P}^*_{\mathbb{X}} \to \Theta$ bijective.

$\Theta \subset \mathbb{R}^p$ is called parametric space. The prior Q induces a probability on Θ:

$$\pi(B) = Q(\tilde{\theta}^{-1}(B)), B \in \mathscr{B}(\Theta).$$

In these cases:

$$X_i \,|\, \tilde{\theta} = \theta \overset{\text{i.i.d.}}{\sim} p_\theta$$
$$\tilde{\theta} \sim \pi \text{ prior distribution}$$

For instance:

$$Q(\{p \in \mathbf{P}_{\mathbb{X}} : \ p(\mathrm{d}x) = \varphi((x-\mu)/\sigma)\,\mathrm{d}x, \ (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}) = 1$$

with φ being the density function of a N(0, 1) distribution.

Parametric vs Nonparametric case

When can we assume $Q(\{\mathbf{P}^*_{\mathbb{X}}\}) = 1$, where $\mathbf{P}^*_{\mathbb{X}}$ is finite-dimensional? More clearly, when could we assume the model is parametric?

- if, from past experience in cases similar to the one analized, we believe that the parametric family approximates well the "true" distribution
- if, in addition to exchangeability, we assume different conditions for the sequence of observations. For example, if $(X_n)_{n\geq 1}$ is also spherically symmetric ($\mathcal{L}(X_1, \ldots, X_n)^T = \mathcal{L}(A(X_1, \ldots, X_n)^T)$ for any orthogonal matrix A), then $\mathbf{P}^*_{\mathbb{X}}$ is the family of Gaussian distributions with 0-mean.

Otherwise: nonparametric model

$\longrightarrow$ greater flexibility when Q has large support, possibly $\mathrm{supp}(Q) = \mathbf{P}_{\mathbb{X}}$.

Bibliography

✏ Schervish M. J. (1995). *Theory of Statistics*. Springer.

✏ Regazzini (1996). *Impostazione non parametrica di problemi di inferenza statistica bayesiana*. TR IAMI 96.21.

The main idea is that so far we assumed for instance $X_1, \ldots, X_n | \theta \sim N(\theta, \sigma_0^2)$ (where σ_0^2 is known and fixed), $\theta \sim Q$. Now we consider that the distribution of the population is not gaussian but $\tilde{p}$:

$$X_1, \ldots, X_n \,|\, \tilde{p} \overset{iid}{\sim} \tilde{p}$$
$$\tilde{p} \sim Q$$

Current approach (Nonparametric)

$\tilde{p}$: random probability measure (population distribution)

$$X_1, \ldots, X_n \,|\, \theta \overset{iid}{\sim} N(\theta, \sigma_0^2)$$
$$\theta \sim Q$$

here we're assuming a prior that is concentrated on a subfamily of distributions ($N(\cdot, \sigma_0^2)$) but we want something more general