

# Package ‘BayesDA’

April 10, 2012

**Version** 2012.04-1

**Date** 2012-04-05

**Title** Functions and Datasets for the book “Bayesian Data Analysis”

**Author** Compiled by Kjetil Halvorsen

**Maintainer** Kjetil Halvorsen <kjetil1001@gmail.com>

**Depends** stats, R (>= 2.2.0)

**Suggests** MCMCpack

**LazyData** FALSE

**ZipData** no

**Description** Functions for Bayesian Data Analysis, with datasets from the book “Bayesian data Analysis (second edition)” by Gelman,Carlin, Stern and Rubin. Not all datasets yet, hopefully completed soon.

**License** GPL (>= 2)

**Repository** CRAN

**Date/Publication** 2012-04-10 11:55:13

## R topics documented:

contingency . . . . .	2
cow . . . . .	3
dilution . . . . .	4
fabric . . . . .	4
factorial . . . . .	5
fatalities . . . . .	6
football . . . . .	7
golf . . . . .	8
light . . . . .	8

meta . . . . .	9
newyork . . . . .	10
personality . . . . .	10
phones . . . . .	11
rats . . . . .	12
schiz . . . . .	13
stratified . . . . .	13

<b>Index</b>	<b>15</b>
--------------	-----------

---

contingency	<i>Contingency Table from a sample Survey</i>
-------------	---

---

### Description

Contingency table describing a survey of sources and quality of information about cancer for 1729 people.

### Usage

```
data(contingency)
```

### Format

A data frame with 32 observations on the following 6 variables.

source.1 a factor with levels N Y

source.2 a factor with levels N Y

source.3 a factor with levels N Y

source.4 a factor with levels N Y

knowledge a factor with levels Good Poor

count number of respondents with this pattern

### Details

The sources of information are: (1) news media, (2) light reading, (3) solid reading, (4) lectures. The book (page 437) is fitting Bayesian loglinear models.

### Examples

```
data(contingency)
```

---

`cow`*Data from an Experiment with Treatment Assignment Based on Co-  
variates*

---

### Description

An experiment was conducted on 50 cows to estimate the effect of a feed additive (methionine hydroxy analog) on six outcomes related to the amount of milk fat produced by each cow. Four diets (treatments) were considered, corresponding to different levels of the additive, and three variables were recorded before treatment assignment: lactation number ( seasons of lactation), age, and initial weight of cow. Multiple randomizations were calculated, and chosen that one with ‘best balane’, however that was defined.

### Usage

```
data(cow)
```

### Format

A data frame with 50 observations on the following 10 variables.

level diet, treatment

lactation lactation number, pretreatment

age age of cow, pretreatment variable

initial.weight initial weight, pretreatment

dry response

milk response

fat response

solids response

final.weight response

protein response

### Examples

```
data(cow)
summary(cow)
names(cow)
# Investigating balance on pretreatment variables:
with(cow, tapply(lactation, level, mean))
with(cow, tapply(age, level, mean))
```

---

dilution                      *Serial Dilution Assay*

---

### Description

A serial dilution assay to study concentration of cockroach allergen in homes of asthma sufferers.

### Usage

```
data(dilution)
```

### Format

The format is: List of 4 \(\$ unknowns : num [1:8, 1:10] matrix with optical densities for unknowns in each column \(\$ standards : num [1:16] vector of standard solution optical densities \(\$ dil.unknowns : num [1:8] dilution factor for unknowns \(\$ dil.standards: num [1:16] dilution factor for standard solution

### Details

Concentration of standard was 0.64

### Examples

```
data(dilution)
str(dilution)
unknowns <- dilution$unknowns
standards <- dilution$standards
dil.unknowns <- dilution$dil.unknowns
dil.standards <- dilution$dil.standards
plot(dil.standards, standards)
matplot(dil.unknowns, unknowns, type="b")
```

---

fabric                      *Numbers of Faults Found in Each of 32 Rolls of Fabric*

---

### Description

Numbers of faults found in each of 32 rolls of fabric produced in a particular factory. Also given is the length of the roll.

### Usage

```
data(fabric)
```

**Format**

A data frame with 32 observations on the following 2 variables.

length length of roll

faults number of faults in roll

**Details**

The book uses this for exercise 5. page 441

**Examples**

```
data(fabric)
str(fabric)
names(fabric)
# Identity link:
with(fabric, plot(faults ~ length))
# log link:
with(fabric, plot(faults ~ length, log="y"))
# Fitting poisson regression models:
mod1 <- glm(faults ~ length-1, data=fabric, family=poisson)
OK <- require(MCMCpack)
if(OK) mod2 <- MCMCpoisson(faults ~ length-1, data=fabric, b0=0, B0=0.0001)
summary(mod1)
confint(mod1)
if(OK) summary(mod2)
# The exercise is to investigate overdispersion ...
```

---

factorial

*Data From a Chemical Experiment*

---

**Description**

A factorial designed experiment from chemistry. Three experimental variables representing reactor conditions, and the response, conversion (%) from n-Heptane to acetylene.

**Usage**

```
data(factorial)
```

**Format**

A data frame with 16 observations on the following 4 variables.

temperature reactor temperature

ratio ratio of H2 to n-heptane (mole ratio)

contact Contact time (sec)

conversion the response, conversion from n-heptane to acetylene

**Details**

This data is used in an exercise on regression with many explanatory variables, page 413 of second edition. Original authors assume a quadratic functional form.

**Examples**

```
data(factorial)
summary(factorial)
# non-Bayesian analysis:
fac.mod1 <- lm(conversion ~ temperature+ratio+contact+
               I(temperature*ratio)+I(temperature*contact)+
               I(ratio*contact)+I(temperature^2)+I(ratio^2)+I(contact^2),
               data=factorial)
summary(fac.mod1)
```

---

fatalities

*Worldwide Airline Fatalities, 1976–1985*


---

**Description**

Worldwide airline fatalities, 1976–1985. Death rate is passenger deaths per 100 million passenger miles.

**Usage**

```
data(fatalities)
```

**Format**

A data frame with 10 observations on the following 4 variables.

year year

facc number of fatal accidents

pdeaths number of passenger deaths

rdeath death rate

**Details**

Source: Statistical Abstracts of the United States

**Examples**

```
data(fatalities)
summary(fatalities)
```

---

`football`*Football Point Spreads and Game Outcomes*

---

**Description**

Data on football point spreads and game outcomes (north american football) for ten seasons, 1981, 1983-1986, 1988-1992, each season are 224 games and they are strung together. Only three first seasons are used in chapter one of book.

**Usage**

```
data(football)
```

**Format**

A data frame with 2240 observations on the following 7 variables.

home home indicator

favorite favorite score

underdog underdog score

spread point spread

favorite.name a factor with levels ATL BUF CHI CIN CLE DAL DEN DET GB HOU IND KC LAA LAN MIA  
MIN NE NO NYG NYJ PHA PHX PIT SD SEA SF TB WAS

underdog.name a factor with levels ATL BUF CHI CIN CLE DAL DEN DET GB HOU IND KC LAA LAN MIA  
MIN NE NO NYG NYJ PHA PHX PIT SD SEA SF TB WAS

week a numeric vector

**Details**

Football experts provide the point spread as a measure of the difference in ability between the two teams. For example, team A might be a 3.5 favourite to team B. The implication of this is that the proposition that team A, the favourite, defeats team B, the underdog, by 4 or more points, are considered a fair bet. In other words, the probability that A wins by more than 3.5 points is 0.5. If the point spread are an integer, then the implication is that team A is as likely to win by more points than the point spread as it is to win by fewer points than the point spread (or to loose). If the win is by exactly the point spread then neither side is paid off.

**Examples**

```
data(football)
summary(football)
names(football)
# In chapter 1 only three first seasons are used:
cap1 <- football[1:672, ]
```

---

golf	<i>Number of Attempts and Successes at Golf Putts</i>
------	---

---

**Description**

Number of attempts and successes of golf putts, by distance from the hole, for a sample of professional golfers.

**Usage**

```
data(golf)
```

**Format**

A data frame with 19 observations on the following 3 variables.

distance Distance from hole in feet  
 n number of attempts  
 y number of successes

**Details**

This is used for an exercise on nonlinear modelling on page 515 in the second edition.

**Examples**

```
data(golf)
names(golf)
comment(golf)
with(golf, plot(distance, y/n))
```

---

light	<i>Simon Newcomb's Measurements for the Speed of Light</i>
-------	--

---

**Description**

Simon Newcomb's measurements (1882) to measure the speed of light. The data are recorded as deviations from 24800 nanoseconds.

**Usage**

```
data(light)
```

**Format**

The format is: atomic [1:66] 28 26 33 24 34 -44 27 16 40 -2 ... - attr(\*, "comment")= chr "Units: deviations from 24800 nanoseconds"

**Details**

The currently accepted value for the speed of light on this scale is 33.0.

**Examples**

```
data(light)
comment(light)
hist(light, breaks=40)
abline(v=33.0, col="red")
```

---

 meta

*Results of 22 Clinical Trials of beta-Blockers*


---

**Description**

Results of 22 clinical trials of beta-blockers for reducing mortality after myocardial infection. Used for meta-analysis.

**Usage**

```
data(meta)
```

**Format**

A data frame with 22 observations on the following 5 variables.

```
study id code of study
control.deaths number of deaths in control group
control.total total number of patients in control group
treated.deaths number of deaths in treatment group
treated.total total number of patients in treatment group
```

**Details**

The 22 clinical trials each consist of two groups of heart attack patients randomly allocated to receive or not receive beta-blockers ( a family of drugs that affect the central nervous system and can relax the heart muscles).

**Examples**

```
data(meta)
names(meta)
# Calculating empirical log-odds and its sampling variances:
y <- apply(meta, 1, function(x) log( (x[4]/(x[5]-x[4]))/(x[2]/(x[3]-x[2])) ) )
s2 <- apply(meta, 1, function(x) 1/(x[5]-x[4]) + 1/x[4] + 1/(x[3]-x[2]) + 1/x[2] )
cbind("Study number"=meta[,1], "empirical log odds"=y, "empirical sampling variance of y"=s2)
#if(require(meta)){
```

```
# funnel(y, sqrt(s2))
# radial(y, sqrt(s2))
#}
```

---

newyork

*Population of Municipalities in New York*

---

### Description

Population in all 804 municipalities in new York state in 1960, and two independent random samples from it.

### Usage

```
data(newyork)
```

### Format

A data frame with 804 observations on the following 2 variables.

population a numeric vector Population

code 400 if in sample 1, 300 if in sample 2, 200 if in both, and 100 if in neither

### Details

Discussed on page 265 of second edition.

### Examples

```
data(newyork)
str(newyork)
```

---

personality

*Personality Data From an Experiment in Psychology*

---

### Description

This is a tree-way array showing responses to 15 possible reactions in 23 situations for 54 persons. It is used in the book as an example for posterior predictive checks.

### Usage

```
data(personality)
```

**Format**

The format is: num [1:15, 1:23, 1:54] 0 0 0 1 0 0 1 2 0 0 ... - attr(\*, "dimnames")=List of 3 ..\$  
 response : chr [1:15] "1" "2" "3" "4" ... ..\$ situation: chr [1:23] "1" "2" "3" "4" ... ..\$ person : chr  
 [1:54] "1" "2" "3" "4" ...

**Examples**

```
data(personality)
str(personality)
# Following code adapted from file personality3.R on the book's webpage:
nsubjects <- 6
nrep <- 7

test <- function (a){
  output <- as.vector(a)>0
  glm.data.frame <- data.frame (output, response, situation, person)
  glm0 <- glm (output ~
    factor(response) + factor(situation) + factor(person),
    family=binomial(link=logit),
    data=glm.data.frame)
  pred0 <- predict.glm (glm0, type="response")
  mean (ifelse(output, (1-pred0)^2, pred0^2))
}
data <- personality[,1:nsubjects]
attrs <- attributes(data)
data <- ifelse (data>0, 1, 0)
attributes(data) <- attrs
```

---

 phones

*CBS Telephone Survey*


---

**Description**

Respondents to the CBS telephone survey classified by opinion, number of residential telephone lines, and number of adults in the household.

**Usage**

```
data(phones)
```

**Format**

A data frame with 27 observations on the following 7 variables.

adults number of adults in household

preference a factor with levels Bush Dukakis No opinion/other

lines.1 number with one tele line

lines.2 number with 2 tele lines  
 lines.3 number with 3 tele lines  
 lines.4 number with 4 tele lines  
 lines.Q number with unknown tele lines

### Details

This is used in exercises pages 242–243 in the book.

### Examples

```
data(phones)
summary(phones)
```

---

rats	<i>Tumor Incidence in Historical Control Groups and Current Group of Rats</i>
------	---

---

### Description

In the evaluation of drugs for possible clinical application, studies are routinely performed on rodents. For a particular study drawn from the statistical literature, suppose the immediate aim is to estimate theta, the probability of tumor in a population of female laboratory rats of type 'F344' that receive a zero dose of the drug — a control group. This gives data from historical control groups, and one current experimental control group.

### Usage

```
data(rats)
```

### Format

A data frame with 71 observations on the following 3 variables.

y number of rats with tumors  
 N number of rats in experiment  
 Current a factor with levels 0 1

### Examples

```
data(rats)
summary(rats)
# moment estimate of (alfa, beta) in beta distribution is (1.4, 8.6)
with(subset(rats, Current=="0"), hist( y/N, freq=FALSE))
plot(function(x) dbeta(x, 1.4, 8.6), from=0, to=1, col="red", add=TRUE)
# plotting posterior in same plot:
plot(function(x) dbeta(x, 5.4, 18.6), from=0, to=1, col="blue", add=TRUE)
```

---

schiz *Data on Response Times for Scizophrenics and non-Schizophrenics*

---

**Description**

Response times (in milliseconds) for 11 non-schizophrenic and 6 schizophrenic individuals.

**Usage**

```
data(schiz)
```

**Format**

The format is: num [1:30, 1:17] 312 272 350 286 268 328 298 356 292 308 ... - attr(\*, "dim-names")=List of 2 ..\$: NULL ..\$: chr [1:17] "nonsch1" "nonsch2" "nonsch3" "nonsch4" ...

A numerical matrix with individuals as columns.

**Details**

Psychological theory from the last half century and before suggests a model in which schizophrenics suffer from an attentional deficit on some trials, as well as a general motor reflex retardation.

**Examples**

```
data(schiz)
str(schiz)
# Making figure 18.1 in the book:
opar <- par(no.readonly=TRUE)
par( mar=c(2.0, 1,1,1))
par(mfrow=c(5,4))
for (i in 1:11) hist( log(schiz[,i]), main="", xlab="", ylab="", xlim=c(5.4, 7.5))
par( mfg=c(4, 1))
for (i in 1:6) hist( log(schiz[,11+i]), main="", xlab="", ylab="", xlim=c(5.4, 7.5))
par(opar)
```

---

stratified *Results of CBS News Survey of 1447 Adults in the United States*

---

**Description**

Results of CBS News survey of 1447 adults in the United States, divided into 16 strata. The sampling is assumed to be proportional, so that the population proportions  $N_j/N$ , are approximately equal to the sampling proportions,  $n_j/n$ .

**Usage**

```
data(stratified)
```

**Format**

A data frame with 16 observations on the following 5 variables.

region a character vector

bush proportio declaring to vote for Bush

dukakis proportion declaring to vote for Dukakis

other proportion declaring to vote for other

proportion sample proportion

**Examples**

```
data(stratified)
str(stratified)
```

# Index

## \*Topic **datasets**

- contingency, 2
- cow, 3
- dilution, 4
- fabric, 4
- factorial, 5
- fatalities, 6
- football, 7
- golf, 8
- light, 8
- meta, 9
- newyork, 10
- personality, 10
- phones, 11
- rats, 12
- schiz, 13
- stratified, 13

  

- contingency, 2
- cow, 3

  

- dilution, 4

  

- fabric, 4
- factorial, 5
- fatalities, 6
- football, 7

  

- golf, 8

  

- light, 8

  

- meta, 9

  

- newyork, 10

  

- personality, 10
- phones, 11

  

- rats, 12

  

- schiz, 13
- stratified, 13