# 医療統計解析のススメ 3

# Bayesian statistics

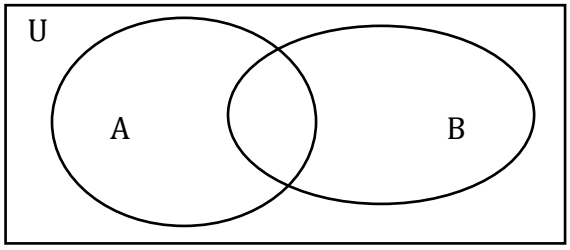
BugsXLA（エクセルで WinBUGS）

横田幸弘

# 目次

# 条件付き確率とベイズの定理

**$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$**（ベイズの定理 Bayes' theorem）




*P(A|B)：(the conditional) probability of A given B （$P_BA$ と書くことがある）

**P(A)**：事象 A が起きる確率＝$\frac{A}{U}$　、**P(B)**：事象 B が起きる確率＝$\frac{B}{U}$

**P(A∩B)**：事象 A と B が同時に起きる確率＝$\frac{(A\cap B)}{U}$

**P(A|B)**：事象 B が起きたもとで事象 A が起きる確率＝$\frac{(A\cap B)}{B}$、とすると

$\frac{(A\cap B)}{U}=\frac{B}{U}\cdot\frac{(A\cap B)}{B}$ つまり「A も B も起きる確率」＝「B が起きる確率」×「B が起きたもとで A が起きる(B の中での A の)確率」 ⇒ **「P(A∩B)=P(B)P(A|B)」(乗法定理)**

同様に A 側から考えると $\frac{(A\cap B)}{U}=\frac{A}{U}\cdot\frac{(A\cap B)}{A}$ ⇒ **「P(A∩B)=P(A)P(B|A)」**

よって恒等式 **P(A∩B)=P(A)P(B|A) =P(B)P(A|B) が成り立ち**

⇒ $P(A|B)=\frac{P(B|A)P(A)}{P(B)}$ または $P(B|A)=\frac{P(A|B)P(B)}{P(A)}$

問題）5 回に 1 回の割合で、帽子を忘れるくせのある K 君が A,B,C の 3 軒を順に訪問して家に帰った時、帽子を忘れてきたことに気づいた。2 件目の家 B に忘れてきた確率を求めよ。（早稲田の入試問題）

求めたいのは帽子をどこかで置き忘れたという確率を全体とした中で、B で忘れたという確率の割合

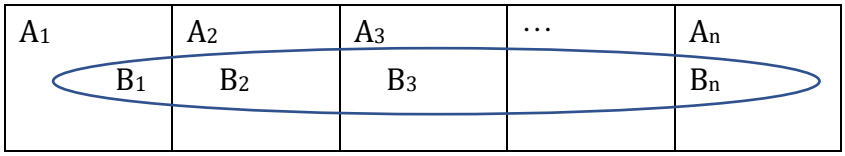
P(x)：帽子をどこかで置き忘れる確率

$1/5+4/5\times1/5+4/5\times4/5\times1/5=61/125$

A に置き忘れる事象 A，B に置き忘れる事象 B，C に置き忘れる事象 C とし確率 P(B) を求めたいが $P(B)=4/5\times1/5=4/25$

求める確率は $4/25\div61/125=20/61$

✧　ここで、ベイズ統計学 Bayesian statistics を考えるとき図として以下をイメージする。




$A=A_1+A_2+A_3+\cdots+A_n$ ,　$B=B_1+B_2+B_3+\cdots+B_n$　とする。

A を原因、B を結果と解釈し、いくつかの原因 $A_1$~$A_n$ が推測され,B という結果がでたとする。大切な点は $B_1=B\cap A_1$, $B_2=B\cap A_2$, …　$B_n= B\cap A_n$　（排反）とも表現でき、

$B= B\cap A_1+B\cap A_2+ B\cap A_3+\cdots+ B\cap A_n$　と書くことができることである。

ただし、前提として、事象 $A_1$、$A_2$、$A_3$、…$A_n$ も排反とする。

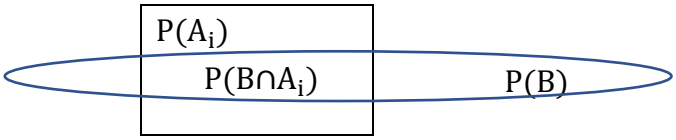
ベイズの理論では、ここで、「乗法定理が適用できる」とすることを出発点とする。

実際にイメージするのは A を全体集合 U と考え、A で割り

$P(A)=P(A_1)+P(A_2)+P(A_3)+\cdots+P(A_n)=1$ としたとき

$P(A_i)$ と同時確率 $P(B\cap A_i)$ の関係のみを考えれば（下図）、上記の条件付き確率、乗法定理が利用できる。

$$P(B\cap A_i) = P(A_i|B)\,P(B) = P(B|A_i)P(A_i) \qquad \Rightarrow \quad P(A_i|B)=\frac{P(B|A_i)P(A_i)}{P(B)}$$




そして、P(B)は上記恒等式を考慮すると、

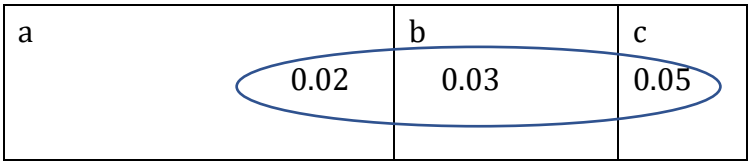
$P(B)= P(B\cap A_1)+P(B\cap A_2)+P( B\cap A_3)+\ldots+P( B\cap A_n)$

$=P(B|A_1)P(A_1)+ P(B|A_2)P(A_2) +\ldots+P(B|A_n)P(A_n)$

A を原因（$A_i$ は原因の一つ）、B を結果（データ）と解釈する。

→　$A_1$~$A_n$ の n 個の原因から、結果（データ）B が生じたと理解する。

例）機械 a,b,c で 60%、30%、10%の割合で作られている製品があり、各機械から 2%、3%、5%の不良品が出ることが経験的にわかっているとする。今、製品全体から適当に取り出した 1 個が不良品であるとき、それが機械 a で作られている確率は？




P(a)=0.6, P(b)=0.3, P(c)=0.1
P(F|a)=0.02, P(F|b)=0.03, P(F|c)=0.05　　(F:不良品である事象とする)
求めたい確率は P(a|F)

$$P(a|F)=\frac{P(F|a)P(a)}{P(F|a)P(a)+P(F|b)P(b)+P(F|c)P(c)}=\frac{0.02*0.6}{0.02*0.6+0.03*0.3+0.05*0.1)}=\frac{6}{13}$$

## 事前確率とベイズ更新

例 1）は、「身につくベイズ統計学」涌井良幸、涌井貞美著、技術評論社から引用しました。

例 1）外からは区別できない壺 1、壺 2 がある。壺 1 には赤玉 4 つと白玉 1 つの計 5 個が、壺 2 には赤玉 2 つと白玉 3 つの計 5 個がある。1 か 2 のどちらかわからない壺が 1 つあり、試しに玉を無作為に 1 個取り出しては戻すという操作を 3 回行った。すると、順に赤、赤、白の玉が出た。以上の情報からこの壺が壺 1 である確率は？

P(V1)：壺 1 から取り出される確率　　P(V2)　：壺 2 から取り出される確率
P(R)：赤玉の確率　　P(W)：白球の確率
とすると

$$P(V1|R)=\frac{P(R|V1)P(V1)}{P(R)}\quad\cdots①$$

$$P(V2|R)=\frac{P(R|V2)P(V2)}{P(R)}\quad\cdots②$$

$$P(V1|W)=\frac{P(W|V1)P(V1)}{P(W)}\quad\cdots③$$

$$P(V2|W)=\frac{P(w|V2)P(V2)}{P(W)}\quad\cdots④$$

公式を羅列

$$P(R)=P(R|V1)P(V1)+P(R|V2)P(V2)$$
$$P(W)=P(W|V1)P(V1)+P(W|V2)P(V2)$$

P(R|V1)＝4/5、P(W|V1)＝1/5、P(R|V2)＝2/5、P(W|V2)＝3/5

ここで、P(V1)、P(V2)は壺 1,2 の最初の存在確率の記載がない。(従来の確率論では解答不可能だが)ベイズの理論ではとりあえず壺 1,2 の存在確率は等しい(壺 1、または壺 2 から取り出された確率が等しい)**(理由不十分の原則)**とすることができる。

P(V1)＝P(V2)＝1/2 **(事前確率)**

1)1 回目の取り出し

①、②に代入すると

$$P(V1|R)=\frac{P(R|V1)P(V1)}{P(R|V1)P(V1)+P(R|V2)P(V2)}=\frac{\frac{4}{5}*\frac{1}{2}}{\frac{4}{5}*\frac{1}{2}+\frac{2}{5}*\frac{1}{2}}=\frac{2}{3}$$

$$P(V2|R)=\frac{P(R|V2)P(V2)}{P(R|V1)P(V1)+P(R|V2)P(V2)}=\frac{\frac{2}{5}*\frac{1}{2}}{\frac{4}{5}*\frac{1}{2}+\frac{2}{5}*\frac{1}{2}}=\frac{1}{3}$$

壺 1,2 の存在確率はそれぞれ 2/3、1/3**(事後確率)**になり、これを 2 回目の事前確率として使用する。

2)2 回目の取り出し

$$P(V1|R)=\frac{P(R|V1)P(V1)}{P(R)}\ \cdots①\qquad P(V2|R)=\frac{P(R|V2)P(V2)}{P(R)}\ \cdots②$$

$$P(R)=P(R|V1)P(V1)+P(R|V2)P(V2)$$

P(R|V1)＝4/5、P(W|V1)＝1/5、P(R|V2)＝2/5、P(W|V2)＝3/5

ここで、P(V1)＝2/3、P(V2)＝1/3 を使用する。**(ベイズ更新)**

①、②に代入すると

$$P(V1|R)=\frac{P(R|V1)P(V1)}{P(R|V1)P(V1)+P(R|V2)P(V2)}=\frac{\frac{4}{5}*\frac{2}{3}}{\frac{4}{5}*\frac{2}{3}+\frac{2}{5}*\frac{1}{3}}=\frac{4}{5}$$

$$P(V2|R)=\frac{P(R|V2)P(V2)}{P(R|V1)P(V1)+P(R|V2)P(V2)}=\frac{\frac{2}{5}*\frac{1}{3}}{\frac{4}{5}*\frac{2}{3}+\frac{2}{5}*\frac{1}{3}}=\frac{1}{5}$$

壺 1 から取り出した確率(事後確率)が 80%に更新されている。この事後確率 4/5,1/5 を 3 回目の事前確率として使用する。

3)3 回目の取り出し(白球なので次式で更新)

$$P(V1|W)=\frac{P(W|V1)P(V1)}{P(W)}\ \cdots③\qquad P(V2|W)=\frac{P(W|V2)P(V2)}{P(W)}\ \cdots④$$

$$P(W)=P(W|V1)P(V1)+P(W|V2)P(V2)$$

P(R|V1)＝4/5、P(W|V1)＝1/5、P(R|V2)＝2/5、P(W|V2)＝3/5

ここで、P(V1)＝4/5、P(V2)＝1/5 を使用する。（ベイズ更新）

③、④に代入すると

$$P(V1|W)=\frac{P(W|V1)P(V1)}{P(W|V1)P(V1)+P(W|V2)P(V2)}=\frac{\frac{1}{5}*\frac{4}{5}}{\frac{1}{5}*\frac{4}{5}+\frac{3}{5}*\frac{1}{5}}=\frac{4}{7}$$

$$P(V2|W)=\frac{P(w|V2)P(V2)}{P(W|V1)P(V1)+P(W|V2)P(V2)}=\frac{\frac{3}{5}*\frac{1}{5}}{\frac{1}{5}*\frac{4}{5}+\frac{3}{5}*\frac{1}{5}}=\frac{3}{7}$$

∴　　壺 1 から取り出した確率→4/7

データを「赤赤白」、「赤白赤」どちらの順に処理しても結果は不変（逐次合理性）

上記のベイズの定理 Bayes' theorem 式中

P(V1)、P(V2)は**事前確率　prior probability** で、P(V1)＝P(V2)＝1/2　で開始され、その後は更新された事後確率が事前確率として使用されている。

**事前確率**は事前の予想、確信などの主観確率が使用されることもある。

P(R|V1), P(R|V2), P(W|V1), P(W|V2)　は**尤度　likelihood**

**尤度**はデータの起きる確率で原因＝母数（確率変数）ごとに決まっている。ここでは P(R|V1)＝4/5、P(W|V1)＝1/5、P(R|V2)＝2/5、P(W|V2)＝3/5

**事後確率　posterior probability** はデータにより更新　Bayesian updating された確率で、次の更新時には事前確率になる。

分母の　P(R|V1)P(V1) ＋ P(R|V2)P(V2)　や

P(W|V1)P(V1) ＋ P(W|V2)P(V2)　は**周辺尤度　marginal likelihood** と呼ばれている。

## ベイズ統計学 Bayesian statistics と頻度論の統計学 frequentism

- ベイズ統計学では母数（パラメータ）を確率変数と考え、データが得られた時、母数のどの値がどれ位の確率でその生起に関与したかをみる。
  従来の統計学＝頻度論の統計学では母数は定数、定数で定められた確率分布に従ってデータの生起確率をみる。

| | 頻度論的統計 | ベイズ統計 |
|---|---|---|
| **母数** | 定数 | 確率変数 |
| **データ** | 確率変数 | 定数 |

- 母数を取り込むには：公式 $P(A_i|B)=\frac{P(B|A_i)P(A_i)}{P(B)}$ で原因に相当する A→$\theta$（母数）、
  結果に相当する B→$x$（データ）に書き換え ⇒ 母数とデータの式と考える。

離散変数の時は

$$P(\theta i|x)=\frac{P(x|\theta i)P(\theta i)}{P(x)}$$ （$x$:データ、$\theta$：パラメータ）

$P(\theta i|x)$：事後確率、$P(x|\theta i)$：尤度関数、$P(\theta i)$：事前確率、$P(x)$：周辺尤度

$$P(x) = P(x|\theta_1)P(\theta_1) + P(x|\theta_2)P(\theta_2) + \cdots + P(x|\theta_n)P(\theta_n)$$

$$\rightarrow \boldsymbol{P(\theta i|x)} = \frac{\boldsymbol{P(x|\theta i)P(\theta i)}}{\boldsymbol{P(x|\theta_1)P(\theta_1)+ P(x|\theta_2)P(\theta_2) +\cdots+P(x|\theta_n)P(\theta_n)}} = \frac{\boldsymbol{P(x|\theta i)P(\theta i)}}{\boldsymbol{\sum_{i=1}^{n} P(x|\theta_i)P(\theta_i)}}$$

連続変数の時は

$$\boldsymbol{\pi(\theta|x)} = \frac{\boldsymbol{f(x|\theta)\pi(\theta)}}{\boldsymbol{P(x)}} = \frac{\boldsymbol{f(x|\theta)\pi(\theta)}}{\boldsymbol{\int_\theta f(x|\theta)\pi(\theta)d\theta}}$$

$\pi(\theta|x)$：事後分布、$f(x|\theta)$：尤度、$\pi(\theta)$：事前分布、$\int_\theta f(x|\theta)\pi(\theta)d\theta$：周辺尤度

連続型のベイズの定理の分母（周辺尤度）に注目すると、もともと$x$は定数であり、$\theta$で積分しているので、周辺尤度は定数なので

$$\boldsymbol{\pi(\theta|x) \propto f(x|\theta)\pi(\theta)} \rightarrow$$ 事後分布∝尤度×事前分布

- ベイズ推定 Bayesian inference には、計算量の負担が大きいという問題がある。
  特に$P(x)$：周辺尤度の計算は膨大な系統に対する総和を必要とし、尤度計算にはパラメータごとの数値積分が要求される。解決法として、

①公式で簡単に計算できるモデルを作る ＝自然共役事前分布、
②PC を利用し複雑なモデルのまま計算 ＝マルコフ連鎖モンテカルロ法（MCMC）

## 自然共役事前分布 natural conjugate prior distribution

ベイズ統計の基本は『事後確率 ∝ 尤度×事前確率』という形で示される。

1.尤度関数の決定

データが従う確率分布。どのような確率分布に従うのか（二項分布か、正規分布にすべきか、ポアソン分布か等）の判断は、データが発生する仕組みを判断して自分で決める必要がある。例えば、試験問題で 10 人のうち 4 人正解だった場合、尤度関数は ${}_{10}C_4\,\theta^4(1-\theta)^6$ の二項分布{bin(10,0.4)} になる。

2.事前分布の決定

事前に何も情報がないときは、ある範囲内で母数のどの確率も等しいとする locally uniform, noninformative, flat などで表現される事前分布 prior distribution を使用する。情報 prior belief distribution があれば利用する。事前確率に尤度関数と相性の良い自然共役事前分布を用いると、事後分布と事前分布が同じタイプになる。
例えば事前に α＋β 回観測がなされ α 回が成功した場合のベータ分布 Beta(α, β)を事前分布に使用し、尤度関数として二項分布 bin(n,p)を使用すると事後分布は事前分布の種類を変えずベータ分布になる。この場合共役 conjugate 関係にあるという。

| 事前分布 | 尤度関数 | 事後分布 |
|---|---|---|
| ベータ分布 | 二項分布 | ベータ分布 |
| 正規分布 | 正規分布（分散既知） | 正規分布 |
| 逆ガンマ分布 | 正規分布（分散未知） | 逆ガンマ分布 |
| ガンマ分布 | ポアソン分布 | ガンマ分布 |

* $X|Y \sim Bin(n,Y)$, $Y \sim Beta(\alpha,\beta)$の階層モデルを考えるとき、X の周辺分布はベータ・2 項分布(beta-binominal distribution)になり、解析的に答えが出る。
* $X|Y \sim Po(Y)$, $Y \sim Ga(\alpha,\beta)$の階層モデルを考えるとき、X の周辺分布はガンマ・ポアソン分布(gamma-Poisson distribution) になり、解析的に答えが出る。

3.事後分布の評価

事後分布からサンプリング sampling を行い、パラメータ parameter を推定する。
ポアソン分布では平均 λ, 正規分布では平均 μ および分散 $\sigma_2$ などのパラメータを推定する。

# PDF3 の目的

ベイズ的解析をするとき、上記の自然共役事前分布の利用するのは手計算の範疇であり、複雑な解析ができないため、Windows 上でマルコフ連鎖モンテカルロ（Markov chain Monte Carlo; MCMC）法によりベイズ階層モデルの解析ができるフリーのソフトウェア WinBUGS（Bayesian inference Using Gibbs Sampling） がしばしば使用される。一般化線形モデルのひとつであるポアソン回帰モデルや、階層的線形モデル（Hierarchical Linear Models: HLM）などの解析が可能となる。

WinBUGS は以下からダウンロードできる。
https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/

WinBUGS を利用するには、直接 WinBUGS にコードを記載する方法や、統計ソフト R を利用する方法があるが、エクセルにアドインとして組み入れエクセル的な操作法のみで解析できる BugsXLA を利用する方法もある。

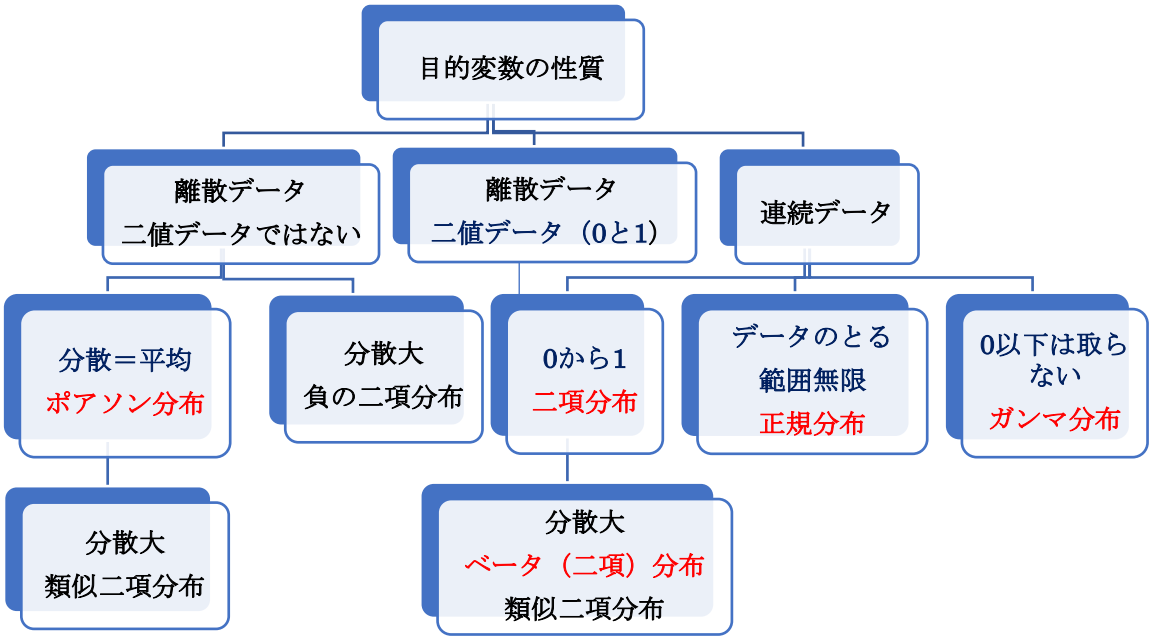
BugsXLA ではコードを書き込むなどの知識は不要なので使いやすく、2018 年 8 月現在、Web 検索した限りでは日本語で BugsXLA について解説した成書、ウェブサイトがないので、今回の PDF3 は、主に BugsXLA を紹介する目的で作成した。

BugsXLA (created by Phil Woodward)は
www.philwoodward.co.uk/bugsxla/download.html から圧縮版がダウンロードできる。その解凍には WINZIP という解凍ソフトが必要で他の解凍ソフトでは解凍できなかった。（WINZIP はアマゾンなどから少額で簡単にダウンロードできる。）
解説本は

Phil Woodward「Bayesian Analysis Made Simple An Excel GUI for WinBUGS」CRC Press であり、この本を参考にして BugsXLA を使用した解析について述べる。BugsXLA ではコード記載の必要はない。

データ（誤差）の確率分布や一般化線形モデルなどに関しての理解は頻度論の統計学 frequentism の範囲になるが、ベイズ統計学 Bayesian statistics の理解や、BugsXLA を使用するのにも必要となる。MCMC などについても文献を参考にして、次ページ以後記載してみる。

## データ(誤差)はどの様な確率分布をするか？




いくつかの確率分布

二項分布 binominal distribution ：bin(n,p)

サイコロを 30 回(n 回)振った時 1 の目(出る確率 $p = 1/6$)は何回(x 回)出るか

$$p(x) = {}_{\mathrm{n}}C_x p^x (1-p)^{n-x} \quad 、\mu = np \quad 、\sigma^2 = np(1-p)$$

⇒ 離散データのうち、二値データ(0 と 1)、連続データのうち 0 から 1 の値をとる場合

ポアソン分布 Poisson d.

一定時間内に起きる事象の回数の分布で稀な現象、発生回数の分布

(カウントデータの分布)

単位時間あたりに平均 $\lambda$ 回起こる現象が単位時間に $x$ 回起きる確率

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad 、\mu = \lambda \quad 、\sigma^2 = \lambda$$

⇒ 二値データではない離散データのうち、分散と平均がほぼ等しい値をとる場合

ただし、ポアソン分布によりモデル化されるデータでは、しばしば平均より分散が大きく、この現象を超過分散(overdispersion)といい、モデルの修正が必要。

**幾何分布** geometric d.

同じコインを投げるとき何回目($x$回目)に初めて表がでるか

$$p(x) = p(1-p)^{x-1} \quad 、\mu = \frac{1}{p} \quad 、\sigma^2 = \frac{1-p}{p^2}$$

**超幾何分布** hypergeometric d.

赤い玉 10 個(A 個)と白い玉 20 個(N-A 個)を混ぜた、計 30 個(N 個)の中から 5 個(M 個)の球を取りだすとき、赤い玉がちょうど 1 つ(X 個)である分布。

* Fisher 検定、log-rank 検定、一般化 Wilcoxon 検定の分布

$$p(x) = \frac{{}_A C_X \cdot {}_{N-A} C_{M-X}}{{}_N C_M}$$

$$\mu = \frac{A \cdot M}{N} \quad 、\sigma^2 = \frac{AM(N-A)(N-M)}{N^2(N-1)}$$

**指数分布** exponential d.

ポアソン分布に従う事象の起こる時間間隔の分布

銀行の窓口に客が到着する時間間隔(1 分あたり 0.5 人の来客がある場合は $\lambda$ =0.5)

$$f(x) = \lambda e^{-\lambda x} \quad 、\mu = \frac{1}{\lambda} \quad \sigma^2 = \frac{1}{\lambda^2}$$

**正規分布 normal distribution $N(\mu,\sigma^2)$**

確率密度関数 $f(x)$ は $f(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$、 平均$\mu$ 分散$\sigma^2$

標準正規分布 $N(0,1^2)$確率密度関数$f(x)$は $f(x)=\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

**正規分布(母分散既知)の場合**

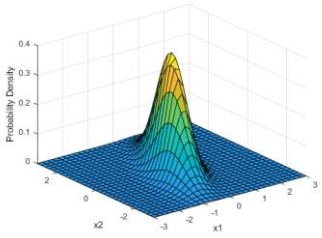
正規分布 $N(\mu,\sigma^2)$ 平均μと 分散 $\sigma^2$ を母数

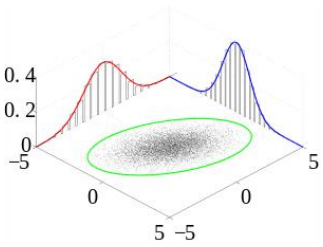
確率密度関数 $f(x)$ は $f(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

上記の正規母集団から大きさ n の標本を抽出し、標本$\bar{x}$ を得たとする。μ の事前分布が期待値 $\mu_0$ 、分散$\sigma_0^2$ の正規分布 $N(\mu_0, \sigma_0^2)$ のとき、μ の事後分布は正規分布になり、その期待値 $\mu_1$ 、分散$\sigma_1^2$ は次式となる。

$$\mu_1 = \frac{n\sigma_0^2\bar{x}+\sigma^2\mu_0}{n\sigma_0^2+\sigma^2} \quad 、\sigma_1^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

## 多変量正規分布　multivariate Normal Distribution(MVN)　$N_k(\mu, \Sigma)$




多変量の統計として共分散まで込めた多次元の正規分布も定義される。
ここで、縦ベクトル α に対して $\alpha^T$ は転置を表し横ベクトルとなる。
$X^T = (X_1, \cdots, X_k)$ , $i, j \in \{1, \cdots, k\}$ のとき平均、共分散、分散を

$$\mu_i = \mathrm{E}[X_i], \sigma_{ij} = Cov(X_i, X_j), \alpha_{ii} = Var(X_i) \text{ とし、}$$

$$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{pmatrix} \text{ とおく。}$$

ここで、$\Sigma = (\sigma_{ij})$は(分散)共分散行列である。
連続確率変数 $\boldsymbol{X}$ の $\boldsymbol{x} = (x_1, \cdots, x_k)^T$ における同時確率密度関数が

$$f_X(x|\mu, \Sigma) = \frac{1}{(2\pi)^{k/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\}$$

で与えられるとき、$\boldsymbol{X}$ は平均 $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ 、共分散行列 $\Sigma$ の多変量正規分布に従い $N_k(\mu, \Sigma)$ で表す。

特に 1 次元の場合、平均 $(\mu)$ と分散共分散行列 $\Sigma = (\sigma^2)$ は共に 1 次元の平均と分散を意味する 1 つの実数値であり、記号 $N_1((\mu), \Sigma) = N(\mu, \sigma^2)$ となる。

## ガンマ関数とガンマ分布、逆ガンマ分布

**ガンマ関数**(本当は複素数などにも定義されるようだが、x>0 の範囲で考える)

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

これは実際に積分すると

$$\Gamma(1) = \int_0^\infty e^{-t} dt = [-e^{-t}]_0^\infty = 1$$

x>1 で部分積分すれば

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} dt = [-e^{-t} t^{x-1}]_0^\infty - \int_0^\infty (-e^{-t})(x-1) t^{x-2} dt$$

$$= 0 + (x-1) \int_0^\infty e^{-t} t^{(x-1)-1} dt$$

$$= (x-1)\,\Gamma(x-1)$$

以上から$\boldsymbol{\Gamma(x+1) = x\Gamma(x) = x!}$

ここで$\Gamma(1/2)$を求めるため　$t = x^2$として積分し、

ガウス積分の公式　　$\int_0^\infty e^{-x^2}dx = \frac{\sqrt{\pi}}{2}$ を使って

計算してみると

ガウス積分の公式

$$\int_0^\infty e^{-ax^2}dx = \frac{1}{2}\sqrt{\frac{\pi}{a}}$$

$$\int_0^\infty x^n e^{-x^2}dx = n!$$

$$\int_0^\infty x^{2n+1}e^{-ax^2}dx = \frac{n!}{2a^{n+1}}$$

$$\int_0^\infty x^{2n}e^{-ax^2}dx = \frac{1\times3\times5\cdots(2n-1)}{2^{n+1}a^n}\sqrt{\frac{\pi}{a}}$$

$$\Gamma(1/2) = \int_0^\infty e^{-t}t^{-1/2}dt = \int_0^\infty e^{-x^2}x^{-1}(2xdx) = 2\int_0^\infty e^{-x^2}dx = \sqrt{\pi}$$

ガンマ関数は階乗を一般化した関数と考えられ例えば 2.5!に相当する$\Gamma(3.5)$は

$$\Gamma(3.5) = 2.5\Gamma(2.5) = 2.5\times1.5\Gamma(1.5) = 2.5\times1.5\times0.5\Gamma(0.5)$$
$$= 2.5\times1.5\times0.5\times\sqrt{\pi} \fallingdotseq 3.32$$ と計算できる。

**ガンマ分布　gamma distribution　$Ga(\theta|a,b), Ga(a,b)$**

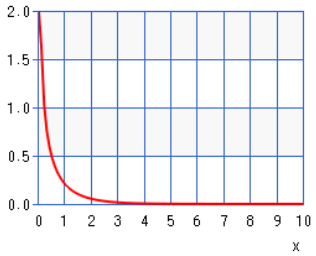
$$Ga(\theta|a,b)=\frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta} \quad 0<\theta,\ 1\leqq a$$

$$\int_0^\infty \theta^{a-1}e^{-b\theta}d\theta = \frac{\Gamma(a)}{b^a}$$

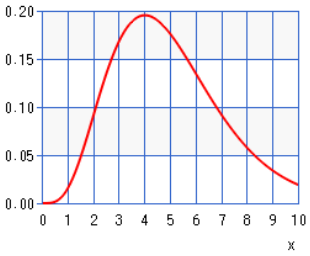
(確率密度関数→積分すると 1 になるのに必要な係数として Γ 関数使用)

$$E(\theta) = \int_0^\infty \theta\frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}d\theta = \frac{a}{b}$$

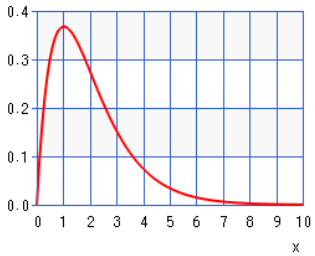
$$V(\theta) = E(\theta^2) - \big(E(\theta)\big)^2 = \frac{a}{b^2}$$




$Ga(0.5,1)$




$Ga(2,1)$




$Ga(5,1)$

$Ga(\theta|a,b)=\frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}$　の式でa = 1　の時

$Ga(\theta|a,b)=\frac{b^1}{1}e^{-b\theta} = be^{-b\theta}$　:指数分布

「同じ指数分布に独立に従う確率変数，$\theta_1$，$\theta_2$，$\cdots$，$\theta_\alpha$ の和，$\theta_1+\theta_2\cdots+\theta a$ の従う確率分布」がガンマ分布

この式で$\theta = 1$ の時

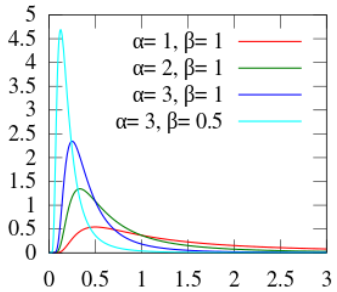
$Ga(\theta|a,b)=\frac{b^a}{\Gamma(a)}e^{-b}=\frac{b^a}{(a-1)!}e^{-b} = a\frac{b^a}{a!}e^{-b}$　　$\frac{b^a}{a!}e^{-b}$ ： ポアソン分布

ガンマ分布$Ga(\theta|a,b)= Ga(a,b)$は，期間平均 $\frac{1}{b}$ ごとに 1 回くらい起こるランダムな事象が $a$ 回起こるまでの時間の分布を表す。例えば「10 年に一度の割合でランダムに起こるイベント（ $b = 0.1$ ）が 3 回 （$a = 3$ ） 起こるまでに何年かかるか」という問題には「期待値は 3/0.1=30 年、確率分布としては $Ga(3,0.1)$ のガンマ分布が対応」

## 逆ガンマ分布 IGa

$IGa(x,\alpha,\beta)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}e^{-\beta/x}$　　$x > 0$

$E(x) = \frac{\beta}{\alpha-1}$ $\alpha > 1$　$V(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ $\alpha > 2$　　$M$（モード）$= \frac{\beta}{(\alpha+1)}$




## ベータ関数とベータ分布

### ベータ関数

$B(p,q) = \int_0^1 x^{p-1}(1-x)^{q-1}dx$　$(p,q\ >0)$

$B(p,q) = I_{p,q}$　　部分積分して

$$= \int_0^1 x^{p-1}(1-x)^{q-1}dx = \left[\frac{x^p}{p}(1-x)^{q-1}\right]_0^1 + \int_0^1 \frac{x^p}{p}(q-1)(1-x)^{q-2}dx$$

$$= \frac{q-1}{p}\int_0^1 x^p(1-x)^{q-2}dx = \frac{q-1}{p}I_{p+1,q-1} = \frac{q-1}{p}\frac{q-2}{p+1}I_{p+2,q-2}$$

$$= \frac{(q-1)(q-2)\cdots 2\cdot 1}{p(p+1)(p+2)\cdots(p+q-2)}I_{p+q-1,1} = \frac{(q-1)(q-2)\cdots 2\cdot 1}{p(p+1)(p+2)\cdots(p+q-2)(p+q-1)} = \frac{(p-1)!(q-1)!}{(p+q-1)!}$$

ベータ関数の性質

(1) $B(p,q)=B(q,p)$

(2) $B(p+1,q)=\dfrac{p}{p+q}B(p,q)$ , $B(p,q+1)=\dfrac{q}{p+q}B(p,q)$

(3) $B(p,q)=2\displaystyle\int_0^{\pi/2}(\cos x)^{2p-1}(\sin x)^{2q-1}dx$

(4) $B(1,1)=1$ , $B\left(\dfrac{1}{2},\dfrac{1}{2}\right)=\pi$

(1) $B(p,1)=\dfrac{1}{p}$

(2) $q>1 \Rightarrow B(p,q)=\dfrac{q-1}{p}B(p+1,q-1)$

(3) $p,q\in\mathbf{N} \Rightarrow B(p,q)=\dfrac{(p-1)!(q-1)!}{(p+q-1)!}$

$B(p,q)=\dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$ の関係あり

## ベータ分布 beta distribution $Be(p,q)$

$$Be(p,q)=\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}=\frac{(p+q-1)!\,x^{p-1}(1-x)^{q-1}}{(p-1)!\,(q-1)!} \quad 0<x<1, 0<p,q$$

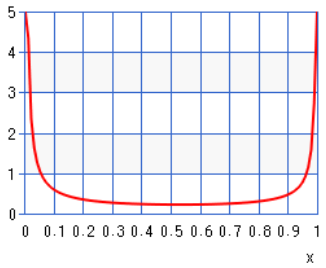
$p=q=1$ の時

$$Be(1,1)=\frac{(1+1-1)!x^{1-1}(1-x)^{1-1}}{(0)!(0)!}=1$$ （一様分布）

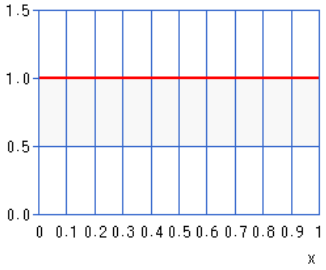
$$E(x)=\int_0^1 x\frac{x^{p-1}(1-x)^{q-1}}{B(p,q)}=\frac{p}{p+q}$$

$$V(x)=E(x^2)-\big(E(x)\big)^2=\frac{pq}{(p+q)^2(p+q+1)}$$

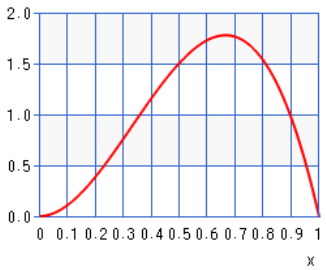
$$M(\text{モード})=\frac{p-1}{(p+q-2)}$$




$Be(0.1,0.2)$




$Be(1,1)$




$Be(3,2)$

ベータ分布$Be(p,q)$ ⇒ $p+q$ 回観測がなされ $p$ 回が成功した場合の分布

# 線形モデルについて

## 正規線形モデル: normal linear model(NLM)

{一般線形モデル: general linear model(LM)}

回帰分析と分散分析はデータを要因効果と誤差の和と考えている点で同一手法であり、[$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$　$\boldsymbol{\beta}$: 推定したいパラメータ、$\boldsymbol{X}$:デザイン行列、$\boldsymbol{\varepsilon}$: 誤差] の形で表現できる。正規線形モデル（一般線形モデル）では、① 応答変数（従属変数、目的変数）y が正規分布に従うことを仮定し，y が正規分布に従わない場合（連続変数ではない場合、例えば合否、生死などの 2 値変数、カウントデータなど）には使用できない。そして、② y の平均がパラメータの一次結合で表わされる。

誤差（残差） $\boldsymbol{\varepsilon} = \boldsymbol{X\beta} - \boldsymbol{y}$ が等分散正規分布するとして、最小二乗法を用いて残差平方和を最小にする $\boldsymbol{\beta}$ を推定する。誤差は$N(0, \sigma^2)$、多変量正規分布では $I$ を n×n 単位行列として$N_n(0, \sigma^2 I)$ に従う。単回帰分析、分散分析、共分散分析、多変量分散分析、多変量共分散分析などの統計モデルが含まれる。

単回帰モデル

$y_i = a + bx_i + \varepsilon_i \qquad i = 1, \cdots, n$

$a$: y-切片項(intercept term)　$b$: 回帰係数(regression coefficient)

$y_i$: 応答変数(response variable)　$x_i$: 説明変数(explanatory variable)

$\varepsilon_i$: 誤差項(error term)

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} a + bx_1 \\ \vdots \\ a + bx_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}$ :推定したいパラメーター

$\boldsymbol{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ :デザイン行列

$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$

重回帰モデル

$y_j = \beta_0 + \beta_1 x_{1j} + \cdots + \beta_k x_{kj} + \varepsilon_j \qquad j = 1, \cdots, n$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}$$ :推定したいパラメーター

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix}$$ :デザイン行列

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

一元配置分散分析モデル

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad i = 1, \cdots, I \qquad j = 1, \cdots, n$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1k} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2l} \\ \vdots \\ y_{j1} \\ \vdots \\ y_{jm} \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 \\ \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_j \\ \vdots \\ \mu + \alpha_j \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1k} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2l} \\ \vdots \\ \varepsilon_{j1} \\ \vdots \\ \varepsilon_{jm} \end{pmatrix}$$

このモデルではパラメータの数が $\mu, \alpha_1, \alpha_2, \cdots, \alpha_j$ の $j+1$ になり多いので、対処法としては、「式を1つ増やして、文字を1つ消す」という方法がとられる。これを制約条件と言い、一つの方法として制約条件 $\alpha_1 + \alpha_2 + \cdots + \alpha_j = 0$ を用いると

$\alpha_j = -\alpha_1 - \alpha_2 \cdots - \alpha_{j-1}$

$$\begin{pmatrix} \mu + \alpha_1 \\ \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_j \\ \vdots \\ \mu + \alpha_j \end{pmatrix} = \begin{pmatrix} \mu + \alpha_1 \\ \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \\ \vdots \\ \mu - \alpha_1 - \alpha_2 \cdots - \alpha_{j-1} \\ \vdots \\ \mu - \alpha_1 - \alpha_2 \cdots - \alpha_{j-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & \\ 1 & 0 & 1 & & \\ \vdots & \vdots & \vdots & & \\ 1 & -1 & -1 & & -1 \\ 1 & -1 & -1 & & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \\ \alpha_{j-1} \end{pmatrix}$$

より

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1k} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2l} \\ \vdots \\ y_{j1} \\ \vdots \\ y_{jm} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & \\ 1 & 0 & 1 & & \\ \vdots & \vdots & \vdots & & \\ 1 & -1 & -1 & & -1 \\ 1 & -1 & -1 & & -1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \\ \alpha_{j-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{1k} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{2l} \\ \vdots \\ \varepsilon_{j1} \\ \vdots \\ \varepsilon_{jm} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \\ \alpha_{i-1} \end{pmatrix} \text{ :推定したいパラメーター}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 0 & & 0 \\ 1 & 1 & 0 & & 0 \\ \vdots & \vdots & \vdots & & 0 \\ 1 & 1 & 0 & & 0 \\ 1 & 0 & 1 & & \\ 1 & 0 & 1 & & \\ \vdots & \vdots & \vdots & & \\ 1 & -1 & -1 & & -1 \\ 1 & -1 & -1 & & -1 \end{pmatrix} \text{ :デザイン行列}$$

$$\boldsymbol{y = X\beta + \varepsilon}$$

二元配置分散分析モデル、共分散分析モデル

① 応答変数 y が正規分布に従い、② y の平均がパラメータの一次結合で表わされるので正規線形モデルである。 $\boldsymbol{y = X\beta + \varepsilon}$ の形式であらわすことができる。

## 一般化線形モデル:generalized linear model(GLM)

一般化線形モデルでは応答変数が正規分布に従わなくても適用でき、質的変数でもよい。応答変数$y$ と説明変数$x$ との関係式は$\beta$ を偏回帰係数、$n$ を説明変数の数として、$g(y) = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_n x_n$ と表現できる。
ここで関数 $g(y)$ は リンク関数（link function） と言う。

＊誤差構造は、応答変数が従う確率分布に一致する。
応答変数が身長 → その誤差構造は正規分布
2 値 (p,1-p) のバラツキを表す分布 → 2 項分布（ベルヌーイ分布）
カウントデータ → ポアソン分布となる。
＊一般化線形モデルで用いる誤差構造は、指数型分布族（正規分布、指数分布、ガンマ分布、ポアソン分布、二項分布などの確率分布）である。

| GLM の種類 | | |
|---|---|---|
| **誤差項の分布** | **リンク関数** | **回帰分析の名称** |
| 正規分布 | 線形（Identity） | 線形回帰 |
| 二項分布 | ロジット（Logit） | ロジスティック回帰分析 |
| 二項分布 | プロビット（Probit） | プロビット回帰分析 |
| ポワソン分布 | 対数線形（Log） | ポワソン回帰分析 |
| ガンマ分布 | 対数線形（Log） | |

**ポアソン回帰モデル**
ポアソン分布 Poisson distribution ：
一定時間内に起きる事象の回数の分布で稀な現象、発生回数の分布

$p(y) = e^{-\lambda}\frac{\lambda^y}{y!}$ 、$\mu = \lambda$ 、$\sigma^2 = \lambda$ で表され、単位時間あたりに平均 $\lambda$ 回起こる現象が単位時間に $y$ 回起きる確率

ポアソン回帰分析 Poisson regression analysis ：
ある現象が一定時間内に起こった回数を数え上げたデータのことをカウントデータ count data といい、カウントデータのうち、分布の期待値（平均値）と分散が$\lambda$ に一致する場合、近似的にポアソン分布するとして、ポアソン回帰分析を行うことができる。
（⇒ 二値データではない離散データのうち、分散と平均がほぼ等しい値をとる場合）
現象として交通事故の発生件数、地震の発生件数、サッカーの得点数、馬に蹴ら

れて死亡した兵士数、一定時間内に疾患を発症した例数や疾患による死亡例数など。

リンク関数に$log$ を使用し

$log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$

$\lambda$:理論的発生件数(発生件数期待値)

$\beta_0$:定数　　$\beta_n$:偏回帰係数　　$\varepsilon$:回帰誤差

一般化線形モデルの 3 要素

①一般化線形モデルで用いる誤差構造(error structure)は、指数型分布族(正規分布、指数分布、ガンマ分布、ポアソン分布、二項分布などの確率分布)

②線形予測子(linear predictor)について

説明変数は線形的にモデルに関与し平均$\mu_i$ の関数

$g(\mu_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$ と表され、ポアソン回帰では

$log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$ となる。

③ここで、$g(\mu_i)$ (例えば $log(\lambda)$) は平均 $\mu_i$ (ここでは$\lambda$ )の関数で、線形予測子と平均の関係を規定する関数$g(\cdot)$(ここでは$log$)がリンク関数(link function )と呼ばれる。

注意すべきは、対数変換して直線回帰をしたり、分散分析の前に，等分散性をめざして「対数変換」や「平方根変換」などの「変数変換」を行ったりすることとは全く別者で混同しないことが大切。

Offset(オフセット変数):例えば地域 $a_1$ では人口 $A_1$=1500 に対し 15 の発生, $a_2$ では人口 $A_2$=3200 に対し 22 発生, $a_3$ では人口 $A_3$=5000 に対して 40 の発生があるとき、割合にしないで人口 $A_i$ を Offset と指定しておけば、データは 15,22,40 等のカウントデータのまま記載できる。

$log$ をリンク関数とし、オフセットに例えば i 地区での人口 $A_i$ を指定した時、線形予測子には $1/A_i$ の形で使用した場合は $-log$ ($A_i$)、 $A_i$×割合の形で使用した場合は $+log$ ($A_i$) として定数として表され、単純な割合の分析ではなく割合に意味を持たせることができる。

ポアソン分布の標準化残差

ここで$Y \sim Poisson(\theta)$ の時、平均と分散が同じなので、$E(Y) = Var(Y) = \theta$ $E(Y)$の推定値$\hat{\theta}$ とすると、$Y - \hat{\theta}$ は残差 residual である。ポアソン分布の標準化残差 r は

$r = \frac{Y-\hat{\theta}}{\sqrt{\hat{\theta}}} \rightarrow \sum r_i^2 = \sum \frac{(Y_i-\hat{\theta}_i)^2}{\hat{\theta}_i} \sim \chi^2(m)$ となる。

ポアソン回帰についての例 （一般化線形モデル入門 Annette J Dobson より引用）

例）一般開業医の利用度が同程度の、都市部 26 人の女性と農村部 23 人の女性の慢性病状（高血圧や関節炎など）の数。（都市群、農村群ともに 70－75 歳、同じ社会経済的地位、一般開業医への来院回数 1996 年に 3 回以下という条件）
慢性病状の数によって診察の必要性を評価するとき、2 群は同程度の必要性をもつか。

| 都市群 |
|---|
| 0 1 1 0 2 3 0 1 1 1 1 2 0 1 3 0 1 2 1 3 3 4 1 3 2 0 |
| n=26, 平均 1.423 分散 1.374 |
| 農村群 |
| 2 0 3 0 0 1 1 1 1 0 0 2 2 0 1 2 0 0 1 1 1 0 2 |
| n=23, 平均 0.913 分散 0.810 |

$Y_{jk}$: $j$ 群の $k$ 番目の女性の病状数、$j=1$:都市群、$j=2$:農村群
$k=1,\cdots,K_j$で$K_1=26, K_2=23$
$Y_{jk}$は互いに独立で、病状数の期待値を表すパラメータ $\theta_j$ を持つポアソン分布に従うと仮定する。

$H_0$: $\theta_1=\theta_2=\theta$
$H_1$: $\theta_1\neq\theta_2$

$E(Y_{jk})=\theta$; $Y_{jk}\sim Poisson(\theta)$ $H_0$が真
$E(Y_{jk})=\theta_j$; $Y_{jk}\sim Poisson(\theta_j)$ $H_0$が真でないと仮定したモデル $j=1,2$

$H_0$: が真のとき

$$f(y_{jk};\theta)=\frac{\theta^{y_{jk}}e^{-\theta}}{y_{jk}!}$$

$$log(\theta;\mathrm{y})=\sum_{j=1}^{J}\sum_{k=1}^{K_j}(y_{jk}log\theta-\theta-log\,y_{jk}!)$$

都市群、農村群を一緒にした平均 $\hat{\theta}=1.184$（最尤推定値）
すべてのデータを代入し対数尤度関数の最大値=-68.3868

$H_1$: が真のとき

$$log(\theta_1,\theta_2;\mathrm{y})=\sum_{k=1}^{K1}(y_{1k}log\theta_1-\theta_1-logy_{1k}!)+\sum_{k=1}^{K2}(y_{2k}log\theta_2-\theta_2-logy_{2k}!)$$

$\hat{\theta}_1 = 1.423, \hat{\theta}_2 = 0.913$ を代入しそれぞれの対数尤度関数の最大値計=-67.0230
$log(\theta_1, \theta_2; \mathrm{y})$ の方が$log(\theta; \mathrm{y})$よりも 1 つ多いパラメータをもつので、最大値は必ず $log(\theta_1, \theta_2; \mathrm{y})$ の方が$log(\theta; \mathrm{y})$より同じか大きくなる。この違いが統計的に有意かどうかは、対数尤度関数の標本分布を知る必要がある。
ここで

$$log(\theta_1, \theta_2; \mathrm{y}) = \sum_{k=1}^{K1}( y_{1k} log\theta_1 - \theta_1 - logy_{1k}!) + \sum_{k=1}^{K2}( y_{2k} log\theta_2 - \theta_2 - logy_{2k}!)$$

は推定されうるパラメータの最大個数（ここでは 2 個）を含み飽和モデル satuated model (=最大モデル maximal model=フルモデル full model)と呼ばれる。

$$log(\theta; \mathrm{y}) = \sum_{j=1}^{J} \sum_{k=1}^{K_j} (y_{jk} log\theta - \theta - log\, y_{jk}!)$$

は関心のあるモデルであり、フルモデルと同じ確率分布（ポアソン分布）および連結関数を持つ一般化線形モデルである。
ここで$Y \sim Poisson(\theta)$ の時、平均と分散が同じなので、$E(Y) = Var(Y) = \theta$ $E(Y)$の推定値$\hat{\theta}$ とすると、$Y - \hat{\theta}$ は残差 residual である。ポアソン分布の標準化残差 r は

$$\mathrm{r} = \frac{Y - \hat{\theta}}{\sqrt{\hat{\theta}}} \rightarrow \quad \sum r_i^2 = \sum \frac{(Y_i - \hat{\theta}_i)^2}{\hat{\theta}_i} \quad \sim \chi^2(m)$$

| Y の値 | 度数 | 標準残差$H_0$<br>$\hat{\theta}$＝1.184 ($\sqrt{\hat{\theta}} = 1.088$) | 標準残差$H_1$<br>$\hat{\theta}_1 = 1.423$ ($\sqrt{\hat{\theta}_1} = 1.193$)<br>$\hat{\theta}_2 = 0.913$($\sqrt{\hat{\theta}_2} = 0.956$) |
|---|---|---|---|
| 都市群 | | | |
| 0 | 6 | -1.088 | -1.193 |
| 1 | 10 | -0.169 | -0.355 |
| 2 | 4 | 0.750 | 0.484 |
| 3 | 5 | 1.669 | 1.322 |
| 4 | 1 | 2.589 | 2.160 |
| 農村群 | | | |
| 0 | 9 | -1.088 | -0.956 |
| 1 | 8 | -0.169 | 0.091 |
| 2 | 5 | 0.750 | 1.138 |
| 3 | 1 | 1.669 | 2.184 |

上記をカイ 2 乗適合度検定と同じ計算をすると
$H_0$ の時

$$\sum r_i^2 = 6 \times (-1.088)^2 + 10 \times (-0.169)^2 + \cdots + 1 \times (1.669)^2 = 46.759$$

自由度 23+26-1＝48 のカイ 2 乗分布に従う

$H_1$ の時

$$\sum r_i^2 = 6 \times (-1.193)^2 + 10 \times (-0.355)^2 + \cdots + 1 \times (2.184)^2 = 43.659$$

自由度 23+26-2＝47 のカイ 2 乗分布に従う

カイ 2 乗値の差は 46.759-43.659＝3.10 と小さい。このことは、2 つのパラメータを持つ$H_1$ のモデルはより単純な$H_0$ のモデルに比べてそれほど良くデータを表現するわけではない。(差がない)

→ 本来はここで尤度比検定等を行う。

**ロジスティック回帰モデル** (→ PDF p159 参照)

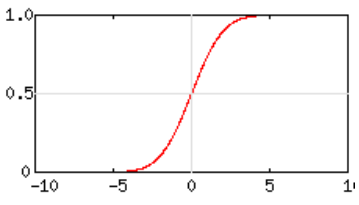
あることが起こる確率を$p$とすると、起こらない確率は$(1-p)$

$\frac{p}{1-p}$ **はオッズ(odds)**と呼ばれている。

ここで、$\boldsymbol{\lambda = \text{logit}p = log_e(\frac{p}{1-p})}$ という変換(ロジット、対数オッズ)を考える。

下図の関係になり logit$p$ がどんな値をとっても $0< p <1$ で

$p$




logit$p$

これを $p$ について解くと $\boldsymbol{p = \frac{exp(\lambda)}{1+exp(\lambda)} = \frac{1}{1+exp(-\lambda)}}$

ある集団(予後因子のない)で起きる確率$p_{x_{i=0}}$ 、ある集団(予後因子のある)で起きる確率$p_{x_{i=1}}$ とすると

$\left(\frac{p_{x_{i=1}}}{1-p_{x_{i=1}}}\right) \Big/ \left(\frac{p_{x_{i=0}}}{1-p_{x_{i=0}}}\right)$ **はオッズ比**であり相対危険度 relative risk にほぼ等しい。

ここで logitp が予後因子を表す変数$X_i$の一次式($\beta$は係数)で表現されたものをロジスティックモデルという。

$$\boldsymbol{\text{logit}p = log_e(\frac{p}{1-p}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n}$$

$$\Leftrightarrow \quad (\frac{p}{1-p}) = exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n)$$

## 線形混合モデル linear mixed model (LMM)

線形混合モデルは、サブグループごとに収集および集計されたデータに関する線形回帰モデルの拡張であり、固定効果と変量効果の 2 つの部分から構成される。固定効果の項は通常、従来の線形回帰部分であり、変量効果はそのサブグループに関連付けられる。変量効果には事前分布がある一方、固定効果にはない。同じレベルのグループ化変数を含む観測値に共通の変量効果を関連付けることで、データのグループ化に関する共分散構造を表現できる。

正規線形混合モデル normal linear mixed model ,NLMM と

一般化線形混合モデル: Generalized linear mixed model, GLMM がある。

正規線形混合モデル normal linear mixed model (NLMM)

正規線形モデル NLM を拡張した統計解析モデルであり、固定効果に加えて変量効果を考慮している。変量効果は通常正規分布を仮定される。

階層的線形モデル hierarchical linear models (HLM)

ランダム係数モデル random coefficient models

階層ベイズモデル hierarchical Bayesian model

成長曲線モデル growth curve models

混合効果モデル mixed-effect model など多数の呼び名がある。

一般化線形混合モデル: Generalized linear mixed model, GLMM

一般化線形モデル GLM を拡張した統計解析モデルであり、固定効果に加えて変量効果を考慮している。変量効果は通常正規分布を仮定される。

参考）分散分析モデルでは、

ANOVA $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ 、Two Way ANOVA $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

$\alpha_i$、$\beta_j$、$\gamma_{ij}$ を母数として扱い、これを固定効果モデル(fixed effects model)という。

これに対して、$\alpha_i$、$\beta_j$、$\gamma_{ij}$ を確率変数として扱うモデルを変量効果モデル(random effects model)といい、ANOVA では $y_{ij} = \mu + A_i + \varepsilon_{ij}$

{$A_i$は$\varepsilon_{ij}$ とは独立に$N(0, \sigma_A^2)$ に従い、$\varepsilon_{ij}$ は $N(0, \sigma_2)$ に従う。}

Two Way ANOVA では、$y_{ijk} = \mu + A_i + B_j + C_{ij} + \varepsilon_{ijk}$

{$A_i$、$B_i$、$C_i$、$\varepsilon_{ij}$ は互いに独立に$N(0, \sigma_A^2)$、 $N(0, \sigma_B^2)$ 、$N(0, \sigma_C^2)$、$N(0, \sigma_2)$に従う。}

ここで、共変量が利用可能なとき、線形回帰モデルと変量効果モデルを組み合わせたモデルを線形混合モデル linear mixed model、階層的線形モデル Hierarchical Linear Models (HLM) などという。

集団全体が類似性のあるデータを調査したサブグループ(A,B,C…)から成る場合、このようなデータを階層的データと呼び、ネストされたデータとも言う。例えばサブグループ A で 10 名、サブグループ B で 20 名…というサンプルを回収したような場合、ネストされたデータである。例として地域ごと、学校ごとで調査データ、反復測定データなどが含まれる。

このようなデータの解析に線形混合モデルが使用される。線形混合モデルは、固定効果(fixed effects)と変量効果(random effects )を含む。固定効果は、データから直接推定できるが、変量効果は直接推定できない。

従来の回帰分析、NLM などではサンプルが独立していることを仮定している。 階層的データは、サンプルが独立していないので、すべてのデータをまとめて

$$y = \beta_0 + \beta_1 x + e$$

の形で解析するのは不可。

また、サブグループ(A,B,C)ごとに回帰分析をすると、

$$y_A = \beta_{0A} + \beta_{1A} x + e_A$$
$$y_B = \beta_{0B} + \beta_{1B} x + e_B$$
$$y_C = \beta_{0C} + \beta_{1C} x + e_C$$

全体で持っている情報が失われ、変数の情報損失が起こる。

線形混合モデルは集団全体の情報も使いつつ、サブグループ($j$= A,B,C…)ごとに分析できる分析方法で、

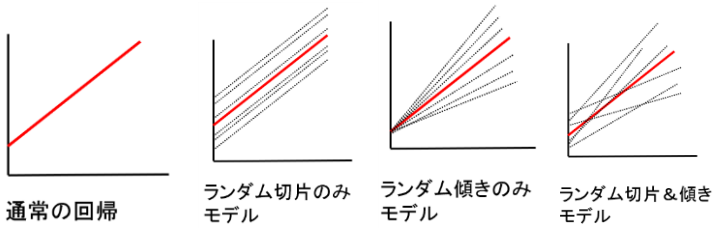
$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + e_{ij}$ と表現される。

各サブグループの切片$\beta_{0j}$ と傾き$\beta_{1j}$ について

$\beta_{0j} = \gamma_{00} + u_{0j}$ 、$\beta_{1j} = \gamma_{10} + u_{1j}$ とすると

$y_{ij} = \gamma_{00} + u_{0j} + (\gamma_{10} + u_{1j}) x_{ij} + e_{ij}$ （赤:固定効果、青:変量効果）

全てのサブグループの切片および傾きを平均したものが$\gamma_{00}$および$\gamma_{10}$によって表され，各サブグループの切片および傾きはその値からのずれ（$u_{0j}$および$u_{1j}$）を足し合わせたものとして表現される。




階層的データの分析では，サブグループの切片および傾きの分散、つまり，切片や傾きがサブグループ間でどれくらい違いが存在するのかに意味がある。

## 頻度論の統計学 frequentism とベイズ統計学 Bayesian statistics の接点

### 1）最尤推定とベイズ推定

ここで、最尤法 maximum likelihood method を復習してみると、
与えられたデータから、それが従う確率分布の母数を点推定する方法。
尤度最大化によって最尤推定値を計算する。 そのモデルで定義される「尤度」を最大化させる母数（parameter）の推定値を計算する。

→ これが最尤推定値 (maximum likelihood estimate; MLE)

幾つかのデータ$y_1, y_2, \cdots, y_n$が得られた時、データを散布図として作成すれば、
個々のデータが分布する母関数の形が推定できる。この時母関数$f(y)$とすると、
各データは同時に$f(y_1), f(y_2), \cdots f(y_n)$ を満たす（=積になる）ことになる。
ここで尤度関数 $L(\theta)$（likelihood function）、ただし $\theta$はその関数の母数とすると、

$L(\theta) = \prod_1^n f(y_1;\theta) \cdot f(y_2;\theta) \cdots \cdot f(y_n;\theta)$ と定義。

対数をとり（対数尤度関数） → $logL(\theta) = \sum_{i=1}^n logf(y_i;\theta)$

偏微分して式をたてる。 → $\frac{\partial}{\partial\theta} logL(\theta) = 0$ となる$\theta$ を求める。

例）（ポアソン Poisson 分布の場合）

散布図やヒストグラムからそのデータがポアソン分布と考えられた時、
データを$y_1, y_2, \cdots, y_n$、とすると、

ポアソン分布の確率密度関数は、$p(y) = e^{-\lambda}\frac{\lambda^y}{y!}$ であるから、

尤度関数（データの得られる確率）は、$\lambda$=$\theta$(推定値)として、

$$L(\theta) = \prod_1^n p(y_n) = \prod_{i=1}^n e^{-\theta}\frac{\theta^{y_i}}{y_i!}$$

対数尤度関数は、

$$logL(\theta) = (y_1 + y_2 + \cdots + y_n) * log(\theta) - n\theta + log(1/(y_1! * y_2! * \cdots\ y_n!)$$

この最大値を求めるため、θ で微分して、=0 とする方程式をたてると、

$$\frac{\partial}{\partial\theta} logL(\theta) = \frac{(y_1+y_2+\cdots+y_n)}{\theta} - n = 0$$

よって、$\theta = \frac{(y_1+y_2+\cdots+y_n)}{n}$ となる。

- 尤度（関数）について、パラメータ $\theta$ に従う分布の密度関数を$f(y;\theta)$とし、尤度関数を$L(\theta;y)$とすると、式の形だけをみると $L(\theta;y) = f(y;\theta)$ と同じ形になる。例えば、$f(y;\theta) = \theta y^2$ で$\theta$を固定したときの$y$ の関数であり二次関数になるが、$L(\theta;y) = \theta y^2$ では$y$を固定したときの $\theta$ の関数になり一次関数になる。

最尤推定とベイズ推定の比較

最尤推定量は頻度論の考え方に基づいた推定だが、上記の手順で$\theta$ を求めるとき、データを固定してパラメータを動かしている。→ ベイズ推定の考え方と一致

最尤推定量は事前情報を使わず、ベイズ推定は事前情報を使う。→ 違う点

また、

最尤推定では $L(\theta) = \prod_1^n f(y_1;\theta) \cdot f(y_2;\theta) \cdots \cdot f(y_n;\theta)$ を尤度関数と呼び、

ベイズ推定では$P(\theta i|x) = \frac{P(x|\theta i)P(\theta i)}{P(x)}$、$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{P(x)}$ 式中

$P(x|\theta i)$、 $f(x|\theta)$ が尤度（関数）と呼ばれる。

2）**線形モデル**: linear model(LM)に接点

⇒ ベイズ統計学では階層ベイズモデル(hierarchical Bayesian model)

パラメータを階層的に (hierarchical) にしてモデルを立てるのは自然である。データがあるパラメータに条件づけられ、$y \sim f(y|\theta)$ 、パラメータ $\theta$ が何らかのパラメータ $\lambda$ を持つ分布に条件づけられるとき、$\theta \sim g_1(\theta|\lambda)$ とすると、$\lambda$ は $\theta$を条件づけるパラメータであり，超パラメータ (hyperparameter) と呼ぶ。さらに、超パラメータに分布を考える。$\lambda \sim g_2(\lambda)$ こともできる。

階層的モデルを立てるときに重要になってくるのが交換可能性 (exchangiability) の概念である。もしあるパラメータセット $\theta = (\theta_1,\theta_2,\ldots,\theta k)$ の順番が変わっても，その同時分布が変化しないなら，そのパラメータセットは交換可能であると言われる。このような交換可能な事前分布は，$\theta$ がある事前分布 $g_1$ からのランダムなサンプルだと仮定することによって得ることができる。

（以下はベイジアン統計解析の実際　　丹後俊郎著　朝倉書店 p2 より引用）

① ある母集団から、無作為に選んだ 1 組の標本 $(y_1, y_2, \cdots, y_n)$ に対して正規分布 $N(\mu, \sigma^2)$ を考え、未知母数 $(\mu, \sigma^2)$ を推定する。(frequentism)

階層ベイズモデル(hierarchical Bayesian model)では

$$y_i \sim N(\mu, \sigma^2)$$
$$\mu \sim N(0, 100^2)$$
$$\tau = {}^1\!/_{\sigma^2} \sim Ga(0.001, 0.001)$$

② n 人の患者について、それぞれ r 回繰り返し測定した場合、(frequentism)では、repeated measures ANOVA または two-way ANOVA without replication（PDF p71 参照）で解析し、統計モデルとして、

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad (i = 1, \cdots, n\ ;\ j = 1, \cdots, r)$

さらに、$\alpha_i \sim N(0, \sigma_B^2)$ と$\alpha_i$のばらつきを推定する変量モデルも(frequentism)である。

階層ベイズモデル(hierarchical Bayesian model)では

$$y_{ij} \sim N(\mu + \alpha_i, \sigma^2) \quad \alpha_i \sim N(0, \sigma_B^2)$$

$$\mu \sim N(0, 100^2)$$

$$1/\sigma^2 \sim Ga(0.001, 0.001)$$

$$1/\sigma_B^2 \sim Uniform(0, 1000)$$

③　n 例のマウスの体重を r 回の測定時期$(x_1, x_2, \cdots, x_r)$で測定した二元配置データは(frequentism)で

$$y_{ij} = \alpha + \beta x_j + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$(i = 1, \cdots, n\ ;\ j = 1, \cdots, r)$ の場合

線形混合モデルとして

$$y_{ij} = (\mu_\alpha + \alpha_i) + (\mu_\beta + \beta_i) x_j + \varepsilon_{ij} \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

$$(\alpha_i, \beta_i) \sim N(0, \Sigma)$$

階層ベイズモデル(hierarchical Bayesian model)では

$$y_{ij} \sim N(\mu_{ij}, \sigma^2)$$

$$\mu_{ij} = \alpha_i + \beta_i x_j$$

$$\alpha_i \sim N(\mu_\alpha, \sigma_\alpha^2)$$

$$\beta_i \sim N(\mu_\beta, \sigma_\beta^2)$$

$$\mu_\alpha \sim N(0, 100^2)$$

$$\mu_\beta \sim N(0, 100^2)$$

$$1/\sigma^2 \sim Ga(0.001, 0.001)$$

$$1/\sigma_\alpha^2 \sim Uniform(0, 1000)$$

$$1/\sigma_\beta^2 \sim Uniform(0, 1000)$$

などと表現できる。

### 3）尤度比検定、逸脱度、情報量基準

#### 尤度比検定

ある母数 $\theta$ がある値をとっているとき，その母集団から取り出した標本 $X$ の値が $x$ である確率（確率密度）を $f(x;\theta)$ とする。このとき，サイズ $n$ の標本の値が

$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ であったとすると，

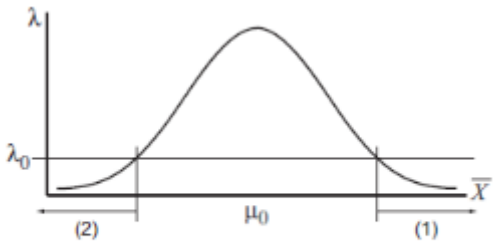
$L(\theta) = L(X,\theta) = \prod_1^n f(x_1;\theta) \cdot f(x_2;\theta) \cdots \cdot f(x_n;\theta)$ となる。

$\theta = \hat{\theta}$ を，$L(X,\theta)$ を最大にする $\theta$，すなわち $\theta$ の最尤推定値とし，尤度の最大値を $L(X, \hat{\theta})$ とする。母集団に対して、ある帰無仮説をたて、仮説が正しいとしたときの $\theta$ の最尤推定値を $\hat{\theta}'$ とし，そのときの尤度関数の値（帰無仮説が正しいとしたときの尤度の最大値）を $L(X, \hat{\theta}')$ とする。ここで，$L(X, \hat{\theta})$ のほうは，L の可能な最大の値なので，$L(X, \hat{\theta}')$ は $L(X, \hat{\theta})$ を上回ることはない。 では，これらの 2 つの尤度の比，すなわち

$$\lambda = \frac{L(X,\hat{\theta}')}{L(X,\hat{\theta})}$$

が非常に小さい，すなわち「$L(X, \hat{\theta}')$ が $L(X, \hat{\theta})$ に比べて極めて小さい」ときは，帰無仮説は間違っていると考え仮説を棄却する。（尤度比検定の考え方）

「$\lambda$ が $\lambda_0$ 以下である確率が α」であるような $\lambda_0$ を選んで，$\lambda$ が $\lambda_0$ 以下であるとき，仮説を棄却するとし，帰無仮説が正しいとき，$0 \leqq \lambda \leqq \lambda_0$ である確率が α，((1),(2)各 α/2)，この，$0 \leqq \lambda \leqq \lambda_0$ という区間が棄却域となる。




棄却域

← $\lambda$ が $\lambda_0$ 以下(1),(2)である時，仮説を棄却

例）正規分布

$$f(x,\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\}$$

尤度関数 $L(X,\mu)$ は

$$L(X,\mu) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma})^2\}$$

$\log L(X;\mu)$ を $\mu$ で微分して 0 とおき，$\mu$ の最尤推定量 $\hat{\mu}$

$\hat{\mu} = (x_1 + x_2 + \cdots + x_n)/n = \bar{X}$ 、したがって，$L(X, \hat{\mu})$ は

$$L(X, \hat{\mu}) = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - \bar{X}}{\sigma})^2\}$$

帰無仮説が正しいときは，帰無仮説で $\mu = \mu_0$ と決めたので，$\hat{\mu'} = \mu_0$ よって

$$L(X, \hat{\mu}') = (\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - \mu_0}{\sigma})^2\}$$

より，尤度比λ を求めると

$$\lambda = \frac{L(X, \hat{\mu}')}{L(X, \hat{\mu})} = \frac{(\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - \mu_0}{\sigma})^2\}}{(\frac{1}{\sqrt{2\pi\sigma^2}})^n \exp\{-\frac{1}{2}\sum_{i=1}^{n}(\frac{x_i - \bar{X}}{\sigma})^2\}} = \exp[-\frac{n}{2\sigma^2}(\bar{X} - \mu_0)^2]$$

棄却域を $0 \leqq \lambda \leqq \lambda_0$ とすると，この式から，それに対応する$\bar{X}$ の区間は，上図 (1)(2) に示される範囲になる。よって，$\bar{X}$ がこの範囲に入る確率が，有意水準 $\alpha$ になるようにすれば，その有意水準に対応する棄却域が求められる。

## 逸脱度 deviance(D)

ここで(ここでは上記式の $x$ を$y$ に置き換えた式で表現しているが特に意味はない)、b がパラメータβ の最尤推定量としてフルモデルの尤度関数$\mathrm{L}(b_{max}; y)$、関心のある尤度関数$\mathrm{L}(b; y)$とすると、尤度比$\lambda = \frac{\mathrm{L}(b_{max};y)}{\mathrm{L}(b;y)}$ で対数をとり、対数尤度関数の差$log\lambda = log\ \mathrm{L}(b_{max}; y) - log\ \mathrm{L}(b; y)$が計算できるが、$-2log\lambda$ がカイ 2 乗分布に従い、一般的に使われる統計量となり逸脱度 deviance(D)と呼ばれる。

尤度比$\lambda = \frac{\mathrm{L}(b_{max};y)}{\mathrm{L}(b;y)} = \frac{(\text{フルモデルの尤度})}{(\text{提案モデルの尤度})}$

D = $-2\ log$（尤度比 λ）$= -2[log\ (b_{max}; y) - log\ (b; y)]$

= $-2$ {$log$（フルモデルの尤度）– $log$（提案モデルの尤度）}

= $-2$(フルモデルの対数尤度 – 提案モデルの対数尤度)

自由度 df = 飽和モデルの自由度 – 提案モデルの自由度

逸脱度はカイ 2 乗分布で近似されるため、検定統計量として利用しやすい。

- フルモデル full model: 推定されうるパラメータの最大個数を含むモデル
- Null モデル:線形予測子が切片だけのモデルのこと
- 関心のあるモデル:提案モデル

最小逸脱度、フルモデルの逸脱度のこと

推定されうるパラメータの最大個数を含むモデルをフルモデル full model（飽和モデル satuated model =最大モデル maximal model)と呼ぶ。

最大逸脱度は，Null モデルの逸脱度のこと

Null モデルとは，線形予測子が切片だけのモデルのこと。

残差逸脱度 residual deviance（ ＝ 関心のあるモデルの逸脱度－最小逸脱度）

残差逸脱度が大きいほど，当てはまりが悪く，小さいほど，あてはまりがよい。

ポアソンモデルでは

$log\ (b_{max}; y) = \sum y_i log y_i - \sum y_i - \sum log y_i!$

（最小逸脱度、フルモデルの逸脱度）

$log\ (b; y) = \sum y_i log \hat{y}_i - \sum \hat{y}_i - \sum log y_i!$

（最大逸脱度，Null モデルの逸脱度）

$$D = -2[log\ (b_{max}; y) - log\ (b; y)] = -2\left[\sum y_i\ log\ (\frac{y_i}{\hat{y}_i}) - \sum (y_i - \hat{y}_i)\right]$$

$\sum y_i = \sum \hat{y}_i$ より

$D = -2 \sum y_i\ log\ (\frac{y_i}{\hat{y}_i})$ （残差逸脱度）

- 逸脱度 deviance($D$)について

逸脱度 deviance($D$) は尤度関数 $L$=$L/1$ と考えて比でなくても単独で定義できる。

parameters θ をもつ逸脱度は、

$D(\theta) = -2logL(y|\theta)$

$\hat{\theta}$ を $\theta$ の最尤推定量としたとき、winBUGS では$D(\hat{\theta})$ はパラメータの最高密度推定値(事後平均)を使用して計算した対数尤度×（-2）

$D(\hat{\theta}) = -2logL(y|\hat{\theta})$

(deviance of the posterior means 、fitted deviance)

$$\bar{D}(\theta) = E_{\theta|y} D(\theta) = E_{\theta|y}[-2logL(y|\theta)] = \frac{1}{N}\sum_{t=1}^{N}\{-2log\ L\left(y\middle|\theta^{(t)}\right)\}$$

(posterior mean of the deviance 、expected deviance)

winBUGS では事後分布の平均逸脱度であり、イタレーションの最後に計算される対数尤度の事後平均×（-2）

$pD = \bar{D}(\theta) - D\left(\hat{\theta}\right)$

(effective number of parameters、pD )

（pD：パラメータのバイアス項。複雑なモデルに対するペナルティを表す。）

## 情報量規準

**AIC（赤池の情報量規準 Akaike information criterion）**

$AIC = D(\hat{\theta}) + 2k = -2logL + 2k$（または、$-2MLL+2k$ ）

：統計モデルの予測の良さの選択規準

*L*:最大尤度、*MLL*: 最大対数尤度 *k*: 最尤推定したパラメータ数

①比較したい色々な曲線（例えば各種成長曲線、正規分布とその他の分布などに適用しモデル選択 ②次数を上げてその測定データとの適合度を高めることができるが、同種の別のデータには合わなくなる（過適合問題、Overfitting）を解決。

⇒ AIC 最小のモデルを選択

AIC は統計モデルのあてはまりの良さ（goodness of fit）ではなく，予測の良さ（goodness of prediction）を 重視するモデル選択規準

hierarchical models 以外で用いられる。

**BIC（ベイズ情報量規準 Bayesian information criterion）**は、

$BIC = D(\hat{\theta}) + klogn$

ただし、$k$ はモデルの dimension, $\theta = (\theta_1,\ldots,\theta_k)$,

$n$ はサンプルサイズ, $y = y_1,\ldots,y_n$

**DIC （ 偏差情報量基準 Deviance information criteria ）**

$DIC = \bar{D}(\theta) + pD = 2\bar{D}(\theta) - D(\hat{\theta})$

$* \; DIC = \bar{D}(\theta) + pD = D(\hat{\theta}) + 2pD$

- 低い DIC 値の方がよい
- DIC はそれ自体では情報量に乏しく、比較に用いる
- AIC を発展させたものとされるが non- hierarchical models では AIC がよい。
- BIC, AIC よりも hierarchical models には DIC が適切
- $D(\hat{\theta})$ は比較的大きいときはpD はマイナス値になることがあるが、 $\hat{\theta}$ が推定値としてよくないと解釈する。

## サンプリング（疑似乱数発生）について

例えば正規分布から1回に10個のデータをサンプリングした時、一つ一つのデータは分布内のデータではあるが値は異なり、10 個を平均すれば母平均にかなり近い値になる。1回に 10 個のサンプリングを 100 回、1000 回、さらに 10000 回行えば、より正確な母平均を推定することができる。このサンプリングは正規分布に従った乱数発生である。いくつかのデータから統計モデルを選択することは、母集団の分布を選択したことに等しく、この時点で母集団に従うデータをサンプリング（乱数発生）できる。

### パラメトリック・ブートストラップ検定

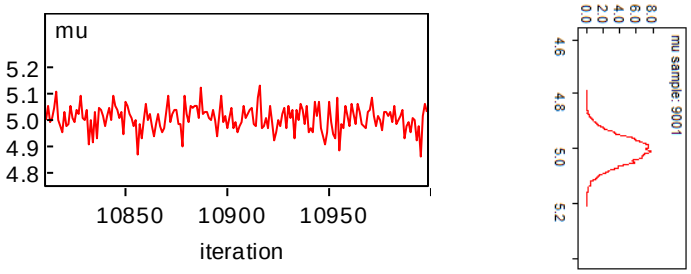
パラメトリック・ブートストラップ検定では、i.i.d(independent and identically distributed)の仮定のもと
例えば一元配置分散分析では、データ＝平均値＋効果＋誤差で表されるが、

F 比＝$\frac{\text{効果の大きさ}}{\text{誤差の大きさ}}$ で評価でき、F 比について

①手持ちのデータから F 比を計算
②推定された確率分布を用いて、①の平均値が変化しないサンプリングを 10000 回ほど繰り返し行い（シミュレーション）その都度 F 比を計算
③シミュレーションで計算された F 比のうち①で計算された F 比を上回った F 比の回数を計算し
④（上回った回数）÷（シュミレーションした回数）が p 値となる。

### 事後分布からのサンプリング

事後分布からサンプリングを行う場合、自然共役事前分布であれば解析的に答えが出る。複雑なモデルではサンプリング自体が困難であったが MCMC では、関心のある事後分布からの値の系列を（関数を確定して積分する方法ではなく）直接サンプリングできる。得られたデータは下記の通り、繰り返し Dynamic trace（左）から下図右のカーネル密度推定のチャート（密度関数）が作成されパラメータの分布がわかりやすい。



# マルコフ連鎖モンテカルロ法（MCMC）

## マルコフ連鎖（Markov chain）

https://mathtrain.jp/markovchain より引用

現在の状態 $X_t$ が与えられた時、過去のいかなる 情報$(X_0,X_1,\ldots,X_{t-1})$も、$X_{t+1}$ を予測する際には 無関係であるという性質をマルコフ性（→つまり直前の動作にのみ影響されるということ）という。

マルコフ連鎖とは、

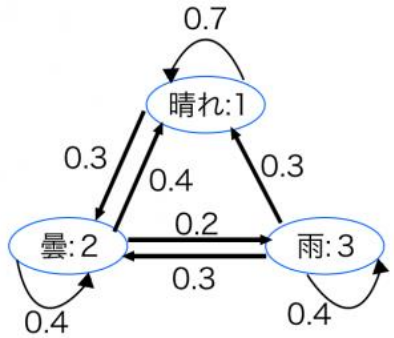
時刻 t の状態 $X_t$ だけに依存する分布で、$X_t$ を決めると次の状態、$X_{t+1}$ はそれより過去の連鎖には依存せず、直前の動作にのみ影響されるというこの列（連鎖）をマルコフ連鎖という。ここでは連鎖は時間的に一様で t に依らないとする。

時間的に一様なマルコフ連鎖は単一の**定常分布(stationary distribution)** に収束することが示され、定常分布が正確に事後確率分布になるような操作を行うのが **MCMC** 法の目的である。t が増加するにつれてサンプル点は定常分布からサンプルされるようになり、初期条件の情報は失われる。つまり定常分布は初期状態には依らない。

マルコフ連鎖の具体例として，以下のようなモデルを考える。

- 昨日以前の天気は翌日の天気に影響しない。
- 今日晴れ →翌日晴れる確率は 0.7，曇の確率は 0.3，雨の確率は 0
- 今日曇 →翌日晴れる確率は 0.4，曇の確率は 0.4，雨の確率は 0.2
- 今日雨 →翌日晴れる確率は 0.3，曇の確率は 0.3，雨の確率は 0.4

状態空間は S={晴れ，曇，雨 }で、晴れ，曇，雨の各状態をそれぞれ 1,2,3 で表すと S={1,2,3} と書ける。下図は状態遷移図という。




また，t 日目の天気を表す確率変数を $X_t$ とおく。例えば，t 日目に晴れたもとで t+1 日目も晴れる確率は 0.7 なので， $P(X_{t+1=1}|\ X_{t=1})=0.7$ となる。

状態遷移の確率を各要素に持つ行列を考える（推移確率行列）。

ij 成分に i から j に遷移する確率を入れたものを推移確率行列（または遷移確率行列，遷移行列）と言い、この例では推移確率行列 P は以下のようになる。

$$P = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.4 & 0.4 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$

- 確率が満たすべき性質より，推移確率行列の各要素は 0 以上 1 以下
- 推移確率行列のどの行も，行和は 1
- 「推移確率行列の n 乗」と「n 回の遷移の確率」が対応する。
  →n 時刻経過に対する推移確率行列を $P^{(n)}$ と書く。$P(X_{t+n=j}|X_{t=i})$ を ij 成分に持つ行列を $P^{(n)}$ とすると，$P^{(n)}=P^n$（右辺は推移確率行列の n 乗）が成立する。

推移確率行列の二乗 $P^2$ を計算してみると，

$$P^2 = \begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.4 & 0.4 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}\begin{pmatrix} 0.7 & 0.3 & 0 \\ 0.4 & 0.4 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix} = \begin{pmatrix} 0.61 & 0.33 & 0.06 \\ 0.5 & 0.34 & 0.16 \\ 0.45 & 0.33 & 0.22 \end{pmatrix}$$

よって，今日の天気をもとに二日後の天気がどうなるかの確率が求まった。n 日後の天気の確率を求めたいときは $P^n$ を計算すればよい。

定常分布 状態が収束するとき、極限を探したい。
このときの極限値 π を定常分布という。$\pi=(\pi1,\pi2,\pi3)$はベクトル，$p$は遷移確率行列。

$$P^4 = \begin{pmatrix} 0.61 & 0.33 & 0.06 \\ 0.5 & 0.34 & 0.16 \\ 0.45 & 0.33 & 0.22 \end{pmatrix}\begin{pmatrix} 0.61 & 0.33 & 0.06 \\ 0.5 & 0.34 & 0.16 \\ 0.45 & 0.33 & 0.22 \end{pmatrix} = \begin{pmatrix} 0.5641 & 0.3333 & 0.1026 \\ 0.547 & 0.3334 & 0.1196 \\ 0.5385 & 0.3333 & 0.1282 \end{pmatrix}$$

$$P^8 = \begin{pmatrix} 0.5641 & 0.3333 & 0.1026 \\ 0.547 & 0.3334 & 0.1196 \\ 0.5385 & 0.3333 & 0.1282 \end{pmatrix}\begin{pmatrix} 0.5641 & 0.3333 & 0.1026 \\ 0.547 & 0.3334 & 0.1196 \\ 0.5385 & 0.3333 & 0.1282 \end{pmatrix}$$

$$= \begin{pmatrix} 0.55577401 & 0.33333333 & 0.11089266 \\ 0.5553371 & 0.33333334 & 0.11132956 \\ 0.55511865 & 0.33333333 & 0.11154802 \end{pmatrix}$$

$\pi=(0.555, 0.333, 0.111)$に収束している

## モンテカルロ積分

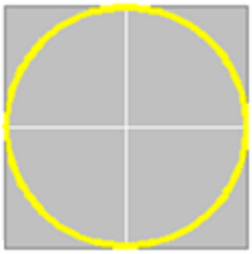
https://www.astr.tohoku.ac.jp/~chinone/Markov_Chain_Monte_Carlo より引用

モンテカルロ積分は事後確率分布 $Pr(X/D)$ からサンプル $\{Xt;\ t=1,2,...,n\}$ を取り出すことで期待値 $E[f(X)]$ を $E[f(x)] \cong \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ と近似的に評価する。

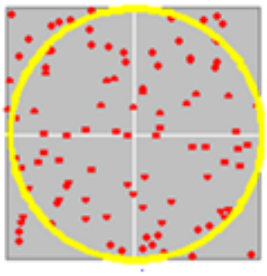
従って $f(X_i)$ の全体平均 (population mean) はサンプル平均で評価される。サンプル$\{X_t\}$が独立のとき、大数の法則 (low of large numbers) により、サンプル数 n を大きくするほど近似の精度は良くなることが保証される。ここで n は解析する人が大きさを決めることが出来る数でありデータ数ではない。一般に、事後確率分布 $Pr(X/D)$ から独立にサンプル点$\{X_t\}$を取り出すことはほとんど不可能である。しかしながら$\{X_t\}$は必ずしも独立である必要はない。$\{X_t\}$は適当な特徴を持った $Pr(X/D)$からサンプルを取り出すという過程によって生成され、これを行う一つの方法がマルコフ連鎖を用いるものである。

例として、乱数を使って円周率 π を計算する方法を考える。

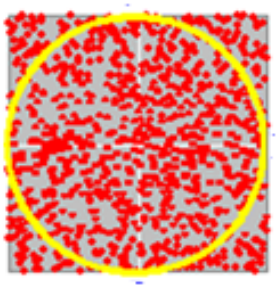
たとえば下 1)のような図形(正方形の中に円)において、



1)



2)



3)

この図 1)の正方形の面積 As は $As=4R^2$、円の面積 Ac は $Ac=\pi R^2$。

ここで、正方形の中にでたらめに点をたくさん打つと円の中に入った個数と入らなかった個数の比率は面積の比率になるので、

$Nin : Nout = Ac : (As-Ac),\ \pi = \frac{4Nin}{Nin+Nout}$

そこで、でたらめに点をうつという作業をコンピュータで、100 点くらい点を打つと図 2)その点が円の中に入っているか、いないかを分けると正方形の面積は、

$Nin = 82$ , $Nout = 18$ より、$\pi = 3.28$ となる。

次に 1000 点打ってみると図 3)。

$Nin = 785$ , $Nout = 215$ より、$\pi = 3.14$

## マルコフ連鎖モンテカルロ法（MCMC 法）

http://hoxo-m.hatenablog.com/entry/20140911/p1 より引用

マルコフ連鎖を定常分布としたサンプリングを行うこと。MCMC 法のアルゴリズムは

step1. 初期点を決める

step2. マルコフ連鎖により次のサンプリングを行う分布を決定する

※初期点近辺(今までのサンプリング)内に対象がありそうである

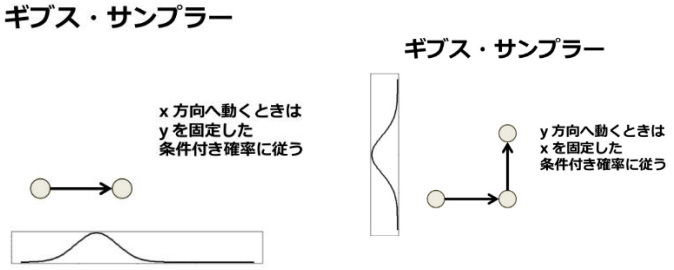
step3. 分布が収束をするまで、step2 を繰り返す

MCMC は定常分布に収束するという性質を持っており，初期値に依存しない
前述の天気の条件を満たすような推移確率を与え，マルコフ連鎖を生成することで，
ギブスサンプリングなどで定常分布に到達することが可能
最尤法では山頂にのみ関心があるが MCMC は山全体の形に興味がある

ギブス・サンプラーは、

1. y の値を固定した条件付き分布から x の次候補をサンプリングする
2. x の値を固定した条件付き分布から y の次候補をサンプリングする

ということを交互に行うシンプルな手法。アルゴリズムの性質上、条件付き分布からのサンプリングが容易にできる場合にしか適用できないが、階層ベイズモデルではこれが容易にできるので、ベイズ統計ではよく使われている。




やはりこのアルゴリズムも、次の状態が前の状態によって決まる(マルコフ連鎖)、確率を使ったアルゴリズム(モンテカルロ法)であることが分かる。
ギブズ・サンプラーは事後分布が特定の確率分布からサンプリングできる時に使える
→ WinBUGS で使える

# WinBUGS を直接使用例（サンプルプログラム）（参考）

http://cyberdoctorhiro.blogspot.com/2012/02/v-behaviorurldefaultvmlo.html より引用

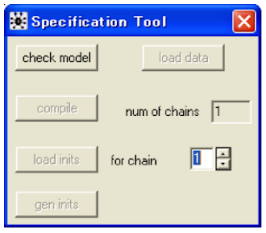
●モデルの診断

スクロールして，モデルを記述している部分を表示させる．

```
model{
for (i in 1:N){
x[ i ]~dnorm(mu, lambda)
}
mu~dnorm(0, 0.001)
sigma~dunif(0, 10)
lambda<-1/pow(sigma, 2)
}
list(N=50,
x=c(5.1,4.9,4.7,4.6,5,5.4,4.6,5,4.4,4.9,5.4,4.8,4.8,4.3,5.8,5.7,5.4,5.1,5.7,5.1,5.4,5.1,4.6
,5.1,4.8,5,5,5.2,5.2,4.7,4.8,5.4,5.2,5.5,4.9,5,5.5,4.9,4.4,5.1,5,4.5,4.4,5,5.1,4.8,5.1,4.6,5.
3,5))
```

「Model」メニューから「Specification...」を選択する．


すると，「Specification Tool」というダイアログが表示される．この時点では，まだ「check model」以外のボタンは使用不可能．




キーワード「model」を選択する．

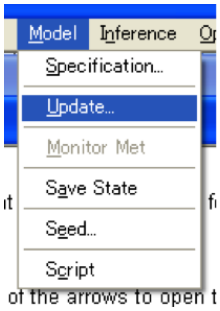
```
model{
for (i in 1:N){
x[ i ]~dnorm(mu, lambda)
}
```

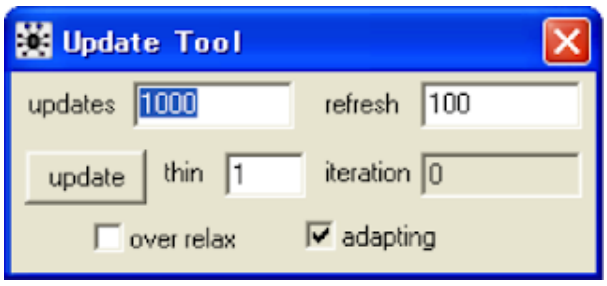
mu~dnorm(0, 0.001)
sigma~dunif(0, 10)
lambda<-1/pow(sigma, 2)

この状態で，「Specification Tool」内の「check model」をクリックする．モデルが文法的に正しければ，ウインドウの左下に「model is syntactically correct」と表示される．また「Specification Tool」内の「load data」と「compile」ボタンが使用できるようになる．

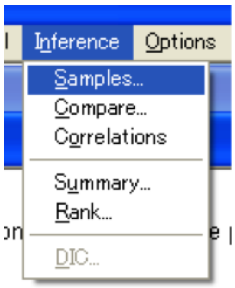
キーワード「list」を選択する．
list(N=50,
x=c(5.1,4.9,4.7,4.6,5,5.4,4.6,5,4.4,4.9,5.4,4.8,4.8,4.3,5.8,5.7,5.4,5.1,5.7,5.1,5.4,5.1,4.6,5.1,4.8,5,5,5.2,5.2,4.7,4.8,5.4,5.2,5.5,4.9,5,5.5,4.9,4.4,5.1,5,4.5,4.4,5,5.1,4.8,5.1,4.6,5.3,5))

この状態で，「Specification Tool」の「load data」をクリックすると，データが読み込まれる．データの読み込みがうまくいけば，ウィンドウの左下に「data loaded」と表示される．
●コンパイル
今回は複数のマルコフ連鎖を生成しないので，「num of chains」は「1」のままにしておく．「Specification Tool」の「compile」をクリックする．コンパイルが成功すれば，ウィンドウの左下に「model compiled」と表示される．また，「Specification Tool」の「load data」と「compile」ボタンが使用不可能になり，代わりに「load inits」と「gen inits」ボタンが使用できるようになる．
●パラメータ推定値の初期値の入力
キーワード「list」を選択し，「Specification Tool」の「gen inits」をクリックする．
うまくいけば，ウインドウの左下に「model is initialized」と表示される．また，
「Specification Tool」の「load inits」が使用不可能になる．
●MCMC の準備
「Model」メニューから「Update...」を選択し，「Update Tool」ダイアログを表示しておく．



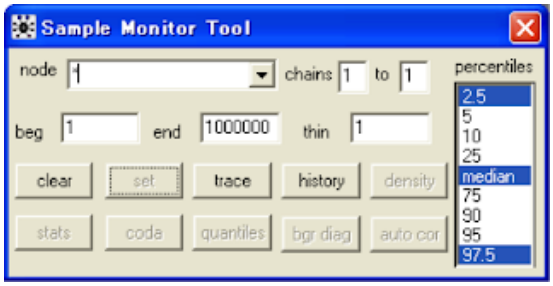
また，「Inference...」メニューから「Samples...」を選択し，「Sample Monitoring Tool」ダイアログを表示させておく．





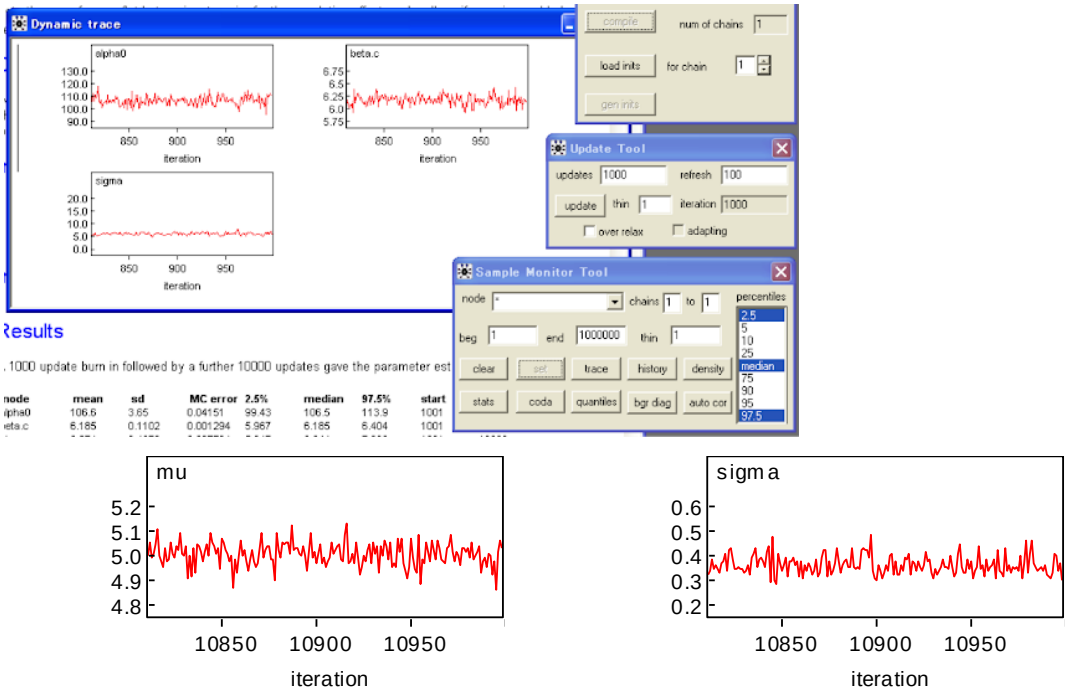
興味のある変数（ここでは alpha0, beta.c, sigma）を登録しておく．まず「Sample Monitor Tool」の「node」に「mu」と入力する．「set」ボタンが使えるようになるので，これをクリックする．すると変数名が消え，「set」ボタンも使用不可能になる．


「sigma」に関しても，同様に入力していく．「node」の右の▼をクリックすると，入力内容を確認できる．「node」にアスタリスク「*」を入力する．「clear」「trace」「history」ボタンが使用できるようになる．




●burn-in sampling

「Sample Monitor Tool」の「trace」ボタンをクリックし，「Dynamic trace」ウインドウを表示させる．ここでは burn-in として 1000 回のサンプリングをするので，「Update Tool」の「updates」は「1000」にしておく．「Update Tool」の「update」ボタンをクリックすると，サンプリングがおこなわれ，「Dynamic trace」にその状況が表示される．また，「Sample Monitor Tool」の「density」「stats」「coda」「qualities」「bgr diag」「auto cor」ボタンが使用できるようになる．







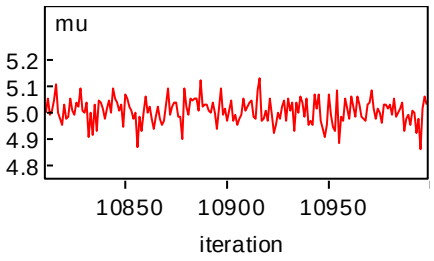
「Dynamic trace」の状況を見て，burn-in が不十分である場合には，サンプリングを適宜追加する．

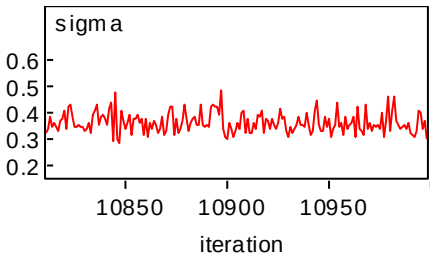
●パラメータ推定


サンプリング回数を 10,000 回にするため，「updates」に「10000」と入力する．

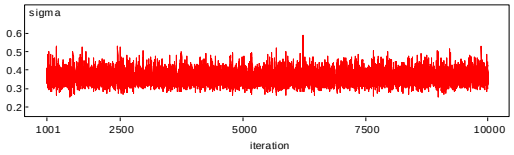
「update」ボタンを押すとサンプリングが始まり，「Dynamic trace」ウインドウに経過が表示される．

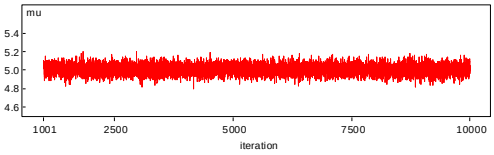
burn-in を除いた 1001～11000 回目のサンプルをもとにパラメータ推定をするため，「Sample Monitor Tool」の「beg」と「end」に「1001」と「11000」を入力する．



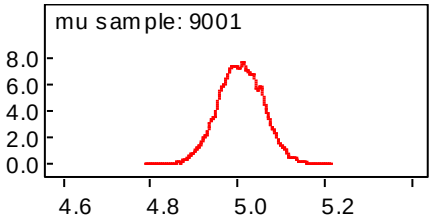



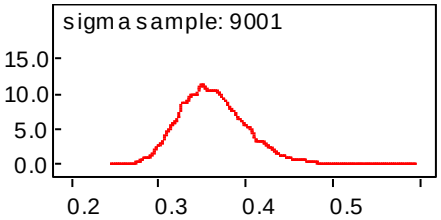

「history」ボタンをクリックして，サンプリングの安定性を確認する.




また「density」ボタンをクリックすると，パラメータ値のカーネル密度推定のチャートが表示される.




「stats」ボタンで，パラメータ推定値の平均や信用区間などが表示される。

| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| mu | 5.006 | 0.05142 | 5.486E-4 | 4.905 | 5.006 | 5.106 | 1001 | 9001 |
| sigma | 0.3622 | 0.03848 | 3.4E-4 | 0.2964 | 0.3591 | 0.4474 | 1001 | 9001 |

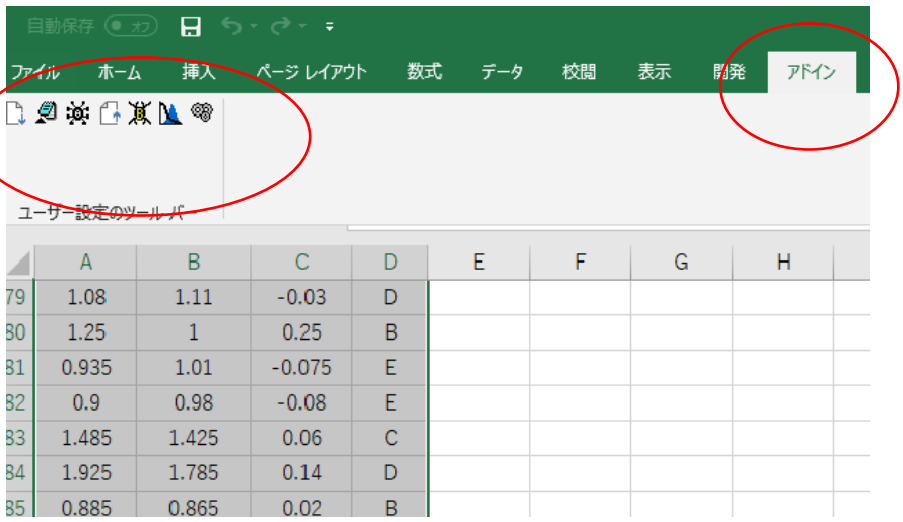
# BugsXLA

ベイズ統計学での　事後分布＝尤度×事前分布/周辺尤度　という公式の中でデータを解析するには、尤度関数、事前分布などについて指定することができれば、その計算はマルコフ連鎖モンテカルロ法（MCMC）を駆使した WinBUGS が事後分布からのサンプリングによって答えを出してくれる。

**「①事後分布内でギブスサンプリングにより場所決めをする。→②その場所に無数の点をまき、点の数からモンテカルロ法により積分代わりに周辺尤度などに必要な計算する。→③指定された尤度関数、事前分布を利用し、事後分布のパラメータ（平均、分散など）を推定する。→④それを繰り返えせば、パラメータの平均（平均、分散の平均など）は大数の法則・中心極限定理などの原理により正規分布するので推定できる。（←MCMC で何が行われているかを何となく理解するための考え方の例）」**

ここではエクセルにアドインとして組み入れエクセル的な操作法のみで解析できる BugsXLA を使用した解析例を紹介する。ポイントは 1、尤度関数、事前分布などについて指定するには統計的な知識が必要で、特に、線形回帰モデルについて知る必要がある、2、BugsXLA の操作を知る必要がある、ということである。

**アドインをクリックしたときのエクセルの状態**




| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 79 | 1.08 | 1.11 | -0.03 | D | | | | |
| 80 | 1.25 | 1 | 0.25 | B | | | | |
| 81 | 0.935 | 1.01 | -0.075 | E | | | | |
| 82 | 0.9 | 0.98 | -0.08 | E | | | | |
| 83 | 1.485 | 1.425 | 0.06 | C | | | | |
| 84 | 1.925 | 1.785 | 0.14 | D | | | | |
| 85 | 0.885 | 0.865 | 0.02 | B | | | | |

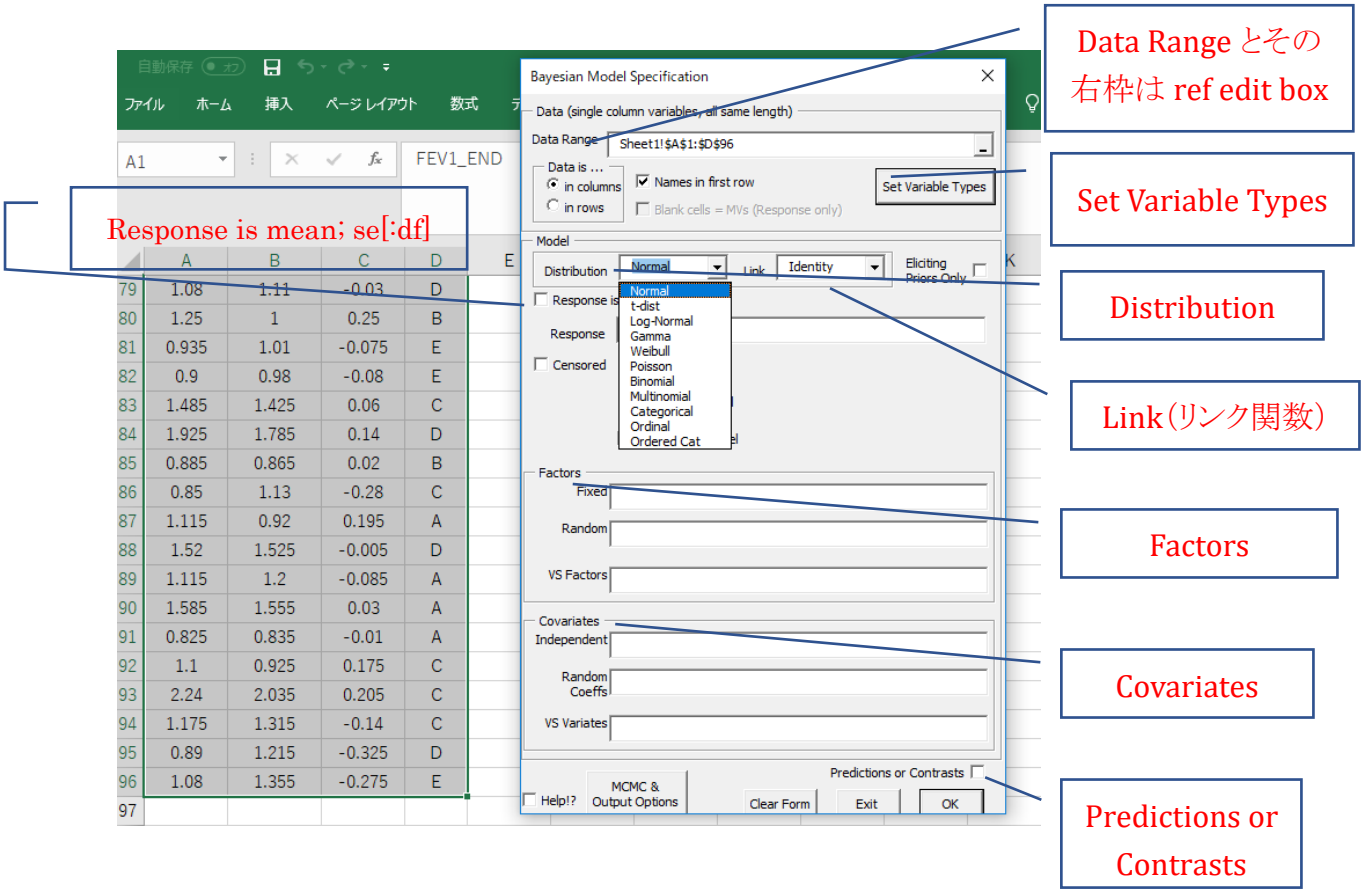
左から　Export to WinBUGS
　　Edit WinBUGS Script
　　Run WinBUGS(Script)
　　Import from WinBUGS

左から　Bayesian Model
　　Post Plots
　　BugsXLA Options

## Bayesian Model Specification

Bayesian Model を選択すると Bayesian Model Specification の図が現れる。

⇒ ここで、データ、尤度関数（Model）の選択、線形予測子（linear predictor）に関する指定を行う




### 1）データ（Data）

'Data Range'の右枠は ref edit box と呼ばれ、ここにエクセルでのデータの範囲を指定する。ただし、1 行目には変数の名前を必ず含み Names in first row、エクセルのデータは列方向 in colums に打ち込んでおく。

Set Variable Types をクリックすると、（当初、すべての変数は量的変数に指定されているので）、ここで、変数を categorical data(factors), 打ち切り例を含むデータ censored variates に再指定できる。

**2）尤度関数の選択（Model）**

① 'Distribution' を選択すると、自動的に Link（リンク関数）が選択される。
Normal, t-Distribution, Log-Normal, Gamma, Weibull, Poisson, Binominal (or Bernoulli), Multinominal, Categorical, Ordinal, Ordered Categorical がある。

② 'Eliciting Priors Only' →事前分布の確認

③ Response is mean; se[:df] は meta-analysis で使用→とりあえずチェックを入れない。

④ 'Response' エクセルに打ち込んだデータのうち、応答変数名を指定する。
ただし、

- Binominal Distribution では / を使用しエクセルでの変数を指定し例えば infected/total などと記載。
- Bernoulli observations の場合は response に AliveDead/1 の形式、例えば、変数名 relief が N,Y で記載されている時、relief/1 と記載。

3）**基本的に線形予測子（linear predictor）は'Factors' 'Covariates'** から成る。
'Factors'は'Fixed'effect または 'Random'effect、そして
'Covariates'は'Independent'または'Random Coefficients'に分類される。
'Factors' 'Covariates' で model statements を行う。

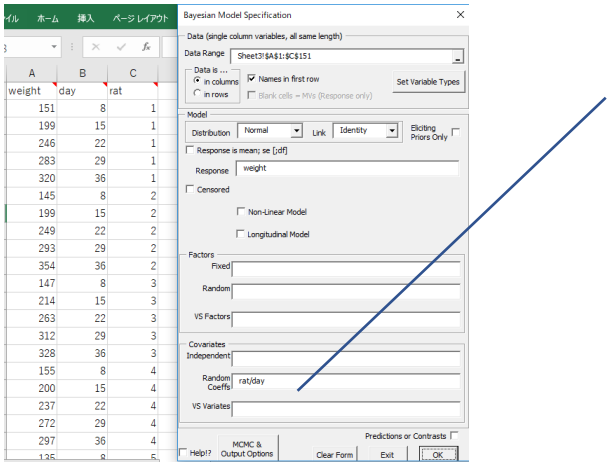
'Factors'に変数を指定する際には Set Variable Types で前もって categorical data(factors)に再指定しておく必要がある。

'Factors'や'Independent' Covariates に記載する際の約束事は以下：

| | |
|---|---|
| + 変数（名）を加える | － 変数（名）を引く |
| : 2 変数間に交互作用 | / A/B=A+A:B |
| * A*B=A+B+A:B | @ A*B*C@2=A*B*C－A:B:C |
| ( ) (A+B)*C=A+B+C+A:C+B:C | ^ quadratic |

polynominal models では
(X1*X2*X3*X4)@2+(X1+X2+X3+X4)^2 とすると 2 次までをすべて含む。

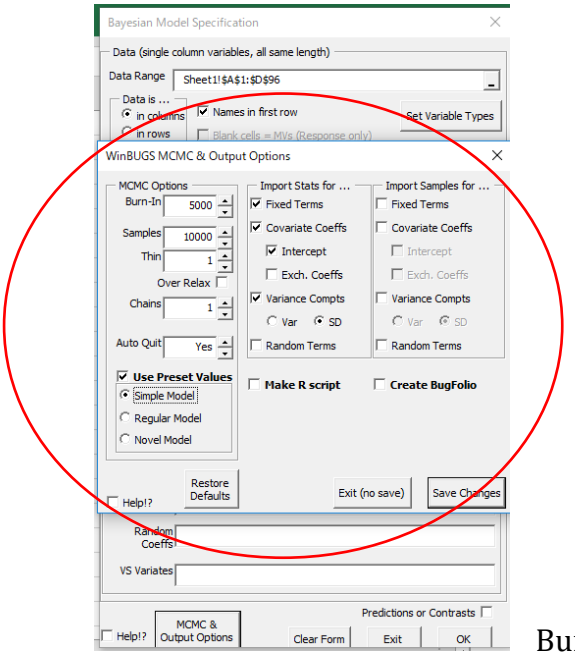
・ 'Random Coefficients'の使用例は以下
random coefficient model であるグループのそれぞれのレベルを指定したいときは、'Covariates' の 'Random Coeffs'に factor/variate の形式で記入する。




Random Coefficients の使用例
factor/variate の形式
例えば rat/day

4) **Predictions or Contrasts**
チェックを入れておき知りたい項目（たとえば項目 TRT その内訳 ABCDE）を指定すると、結果が計算される。データの平均や標準偏差などが計算される。

ここで MCMC & Output Options ボタンを押すと WinBUGS MCMC & Output Options の画面が現れるので、チェックを入れる。Save Change 後 Bayesian Model Specification の OK ボタンを押すと Prior Distributions の画面が現れる。




Burn-in, Samples, Chains などチェック

Burn-in：例えば最初の 1000～5000 個の繰り返しのサンプルの値は、定常分布に収束前のサンプルの値であり、初期値依存性が強いため捨てる。 (burn-in samples)

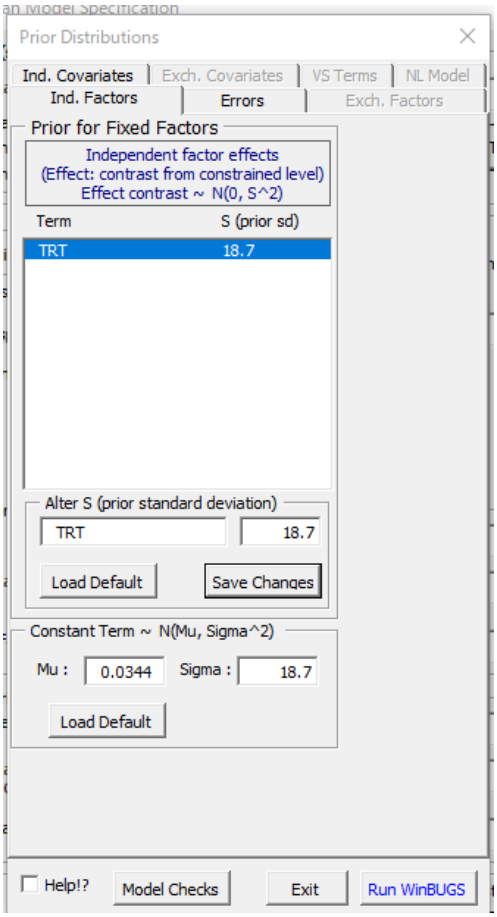
Samples：サンプリングの回数

Thin：この間隔で，ベイズ階層モデルイタレーション iteration が保存される。iteration は MCMC アルゴリズムで実際に実行され保存された繰り返し回数で、例えば iteration が 1000 で thin が 2 の場合は，1 つおき(1，3，5‥)に iteration が保存され，実際は 2000(=1000×2)回の iteration が行われ，最終的に 1000 個のデータが得られる。

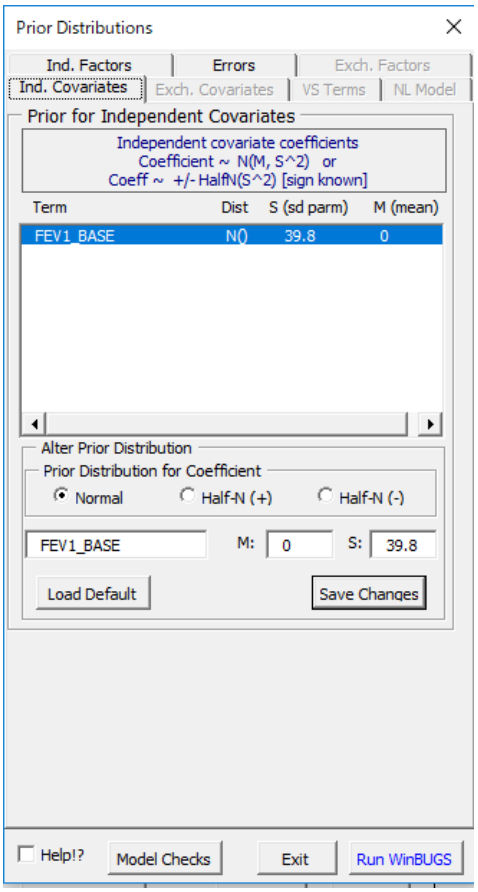
Chains：いくつかの出発点で複数の連鎖(チェーン)を設定できる。

## 6) Prior Distributions

Independent Factors page




Independent Covariates page

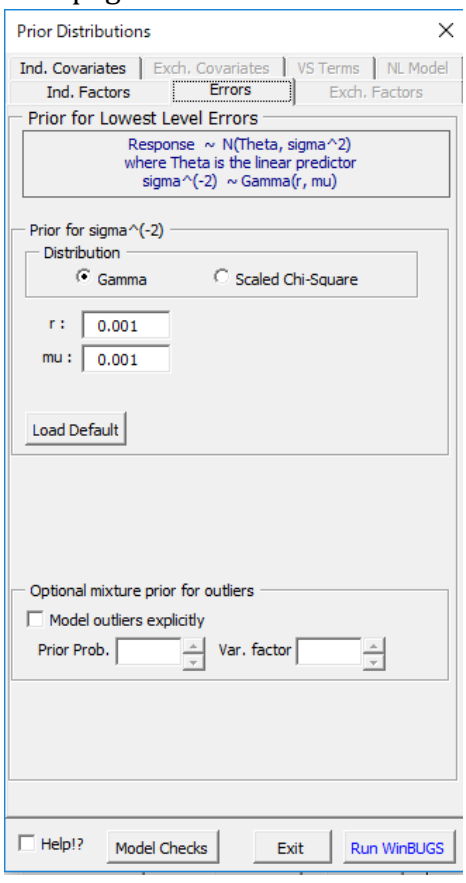



Independent Factors page では、いわゆる固定効果 fixed effects (factors)の事前分布がデフォルトで提供されているが、BugsXLA Options から確認、変更することもできる。デフォルトでの S(prior sd)は平坦な事前分布 flat prior を得るために 100 倍にな

っている。また、link function を使用すれば、informative prior も link scale になっている。

Independent Covariates page でも、いわゆる固定効果 fixed effects(covariates)の事前分布がデフォルトで提供されているが、デフォルトでは Normal Distribution となっている。Half Normal Distribution (HN+,HN-) も指定でき+ やー に違いないときに、より小さな値に対応できる。デフォルトでの S(prior sd)は平坦な事前分布 flat prior を得るために 100 倍になっている。ただし、HN+,HN-を使用すれば S(sd parameter)は通常の sd ではなく、Half Normal's scale parameter であり、link function を使用すれば、informative prior も link scale になっている。

Error' page




Prior Distributions の画面には
Independent Factor spage
Independent Covariates page
Error page がある。

BugsXLA(WinBUGS)での正規分布のパラメータは平均と分散ではなく，平均と分散の逆数である精度 precision であらわす。 $N(\mu, \sigma^2)$ ⇒ dnorm( mu , tau )
precision($\tau$, tau): $\sigma^{-2}$ 測定値や予測値の分散の逆数として定義される

$$\sigma = {}^{1}/_{\sqrt{\tau}} \rightarrow \sigma^2 = {}^{1}/_{\tau} \rightarrow \tau = {}^{1}/_{\sigma^2}$$

事前分布 prior での the lowest level residual error の例

| error distribution | error's prior |
|---|---|
| Normal | sigma^2 |
| $\tau$ = sigma^(-2) | Gamma ( r, mu) |
| t-distribution : t(Theta, sigma^2 , df) | sigma^2 , df |
| $\tau$ = sigma^(-2) | Gamma ( r, mu) |
| df^(-1) | Uniform (a, b) |

Error page での例

response ~ N(Theta, sigma^2)

where Theta is the linear predictor

sigma^(-2) ~ Gamma( r, mu)

応答変数 response が正規分布に従い、N(Theta, sigma^2)で表されたとき、

Theta は線形予測子であり、sigma^(-2)= $\tau$ はガンマ分布 Gamma( r, mu)に従う。

df^(-1)~ Uniform(a, b)つまり、自由度 df として df^(-1) は一様分布に従うなどと表されデフォルトで数値が示されている。

# BugsXLA による解析 1）

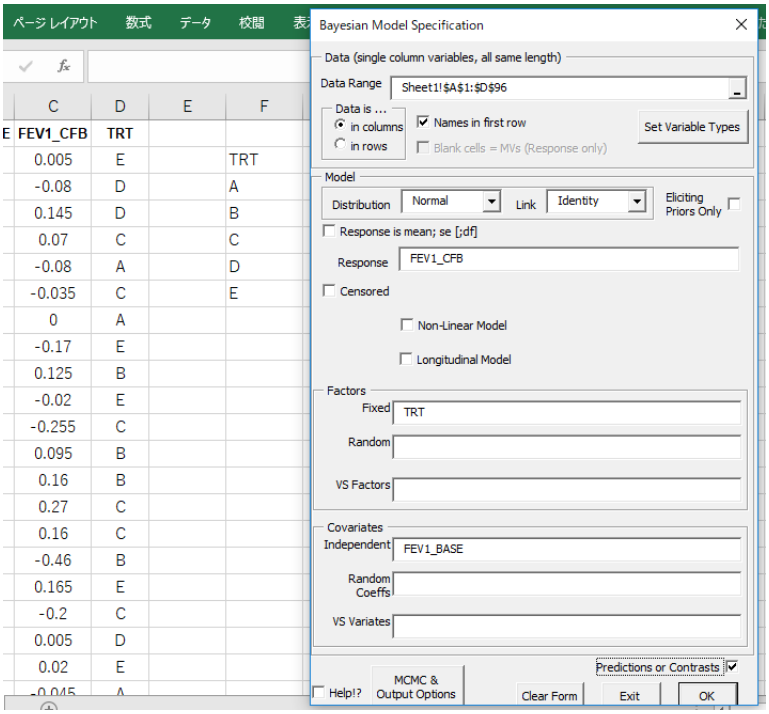
## 正規線形モデル normal linear model (NLM, LM)

Parallel Groups Clinical Study (Analysis of Covariance)

| FEV1_END | FEV1_BASE | FEV1_CFB | TRT |
|---|---|---|---|
| 1.93 | 1.925 | 0.005 | E |
| 1.92 | 2 | -0.08 | D |
| 1.595 | 1.45 | 0.145 | D |
| 1.67 | 1.6 | 0.07 | C |
| 0.495 | 0.575 | -0.08 | A |
| 0.56 | 0.595 | -0.035 | C |
| 0.5 | 0.5 | 0 | A |
| 1.47 | 1.64 | -0.17 | E |
| 1.25 | 1.125 | 0.125 | B |
| 1.22 | 1.24 | -0.02 | E |
| 1.43 | 1.685 | -0.255 | C |
| 0.975 | 0.88 | 0.095 | B |
| 1.255 | 1.095 | 0.16 | B |
| 1.7 | 1.43 | 0.27 | C |
| 1.57 | 1.41 | 0.16 | C |
| 2.16 | 2.62 | -0.46 | B |
| 1.55 | 1.385 | 0.165 | E |
| 2.045 | 2.245 | -0.2 | C |
| 1.04 | 1.035 | 0.005 | D |
| 1.07 | 1.05 | 0.02 | E |
| 1.005 | 1.05 | -0.045 | A |
| 0.715 | 0.68 | 0.035 | D |
| 0.81 | 0.72 | 0.09 | B |
| 1.19 | 1.31 | -0.12 | A |
| 1.715 | 1.65 | 0.065 | E |
| 1.67 | 1.44 | 0.23 | A |
| 1.02 | 1.155 | -0.135 | B |
| 1 | 0.96 | 0.04 | E |
| 1.515 | 1.48 | 0.035 | A |
| 2.02 | 1.71 | 0.31 | C |
| 1.76 | 1.24 | 0.52 | C |
| 2.235 | 2.04 | 0.195 | E |
| 0.745 | 0.83 | -0.085 | D |
| 2.79 | 2.41 | 0.38 | C |
| 1.665 | 1.4 | 0.265 | A |
| 2.255 | 1.965 | 0.29 | B |
| 1.645 | 1.735 | -0.09 | D |
| 0.845 | 0.85 | -0.005 | C |
| 1.22 | 0.755 | 0.465 | D |
| 0.725 | 0.7 | 0.025 | E |
| 0.745 | 0.88 | -0.135 | B |
| 1.52 | 1.27 | 0.25 | C |
| 0.945 | 1.36 | -0.415 | A |
| 0.875 | 0.79 | 0.085 | D |
| 1.34 | 1.34 | 0 | C |
| 1.28 | 1.01 | 0.27 | C |
| 0.67 | 0.625 | 0.045 | E |
| 1.625 | 1.615 | 0.01 | D |
| 0.85 | 0.825 | 0.025 | E |
| 1.13 | 1.04 | 0.09 | C |
| 1.135 | 1.105 | 0.03 | C |
| 1.545 | 1.34 | 0.205 | C |
| 1.32 | 1.22 | 0.1 | E |
| 1.85 | 1.845 | 0.005 | C |
| 1.715 | 2.25 | -0.535 | D |
| 0.785 | 0.875 | -0.09 | A |
| 0.69 | 0.65 | 0.04 | B |
| 1.91 | 1.72 | 0.19 | C |
| 1.125 | 1.225 | -0.1 | D |
| 0.665 | 0.685 | -0.02 | A |
| 1.19 | 0.97 | 0.22 | C |
| 1.15 | 0.89 | 0.26 | C |
| 1.335 | 1.52 | -0.185 | B |
| 0.795 | 0.86 | -0.065 | C |
| 1.895 | 2.08 | -0.185 | D |
| 0.45 | 0.695 | -0.245 | A |
| 1.21 | 0.975 | 0.235 | E |
| 1.17 | 1 | 0.17 | D |
| 1.55 | 1.57 | -0.02 | C |
| 2.25 | 2.08 | 0.17 | C |
| 1.135 | 0.9 | 0.235 | B |
| 0.805 | 0.82 | -0.015 | A |
| 2.155 | 2.09 | 0.065 | C |
| 1.465 | 1.16 | 0.305 | B |
| 1.76 | 1.85 | -0.09 | A |
| 2.145 | 1.98 | 0.165 | B |
| 1.165 | 1.125 | 0.04 | B |
| 1.08 | 1.11 | -0.03 | D |
| 1.25 | 1 | 0.25 | B |
| 0.935 | 1.01 | -0.075 | E |
| 0.9 | 0.98 | -0.08 | E |
| 1.485 | 1.425 | 0.06 | C |
| 1.925 | 1.785 | 0.14 | D |
| 0.885 | 0.865 | 0.02 | B |
| 0.85 | 1.13 | -0.28 | C |
| 1.115 | 0.92 | 0.195 | A |
| 1.52 | 1.525 | -0.005 | D |
| 1.115 | 1.2 | -0.085 | A |
| 1.585 | 1.555 | 0.03 | A |
| 0.825 | 0.835 | -0.01 | A |
| 1.1 | 0.925 | 0.175 | C |
| 2.24 | 2.035 | 0.205 | C |
| 1.175 | 1.315 | -0.14 | C |
| 0.89 | 1.215 | -0.325 | D |
| 1.08 | 1.355 | -0.275 | E |

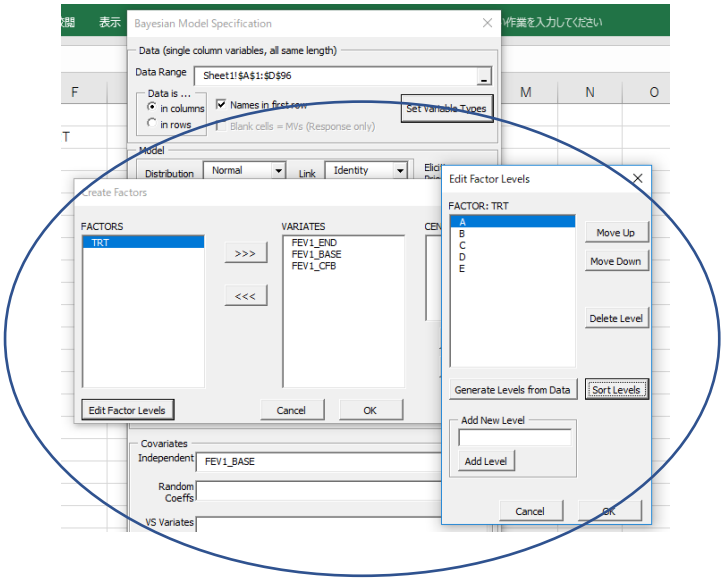
上記データは、「Bayesian Analysis Made Simple An Excel GUI for WinBUGS」の 3 章 NLM を解説した例題 Parallel Groups Clinical Study(Analysis of Covariance)である。

COPD の患者さんで、新薬の効果を知るために治療法 TRT としてプラセボ A、に対して 4 つの投与量 B,C,D,E があり、FEV1_BASE、FEV1_END を調べ、response として FEV1_CFB を調べている。FEV1_BASE が回帰係数(解析結果として、Intercept at 0、WinBUGS での名前では alpha と表現される)を伴う共変量 covariate となる。



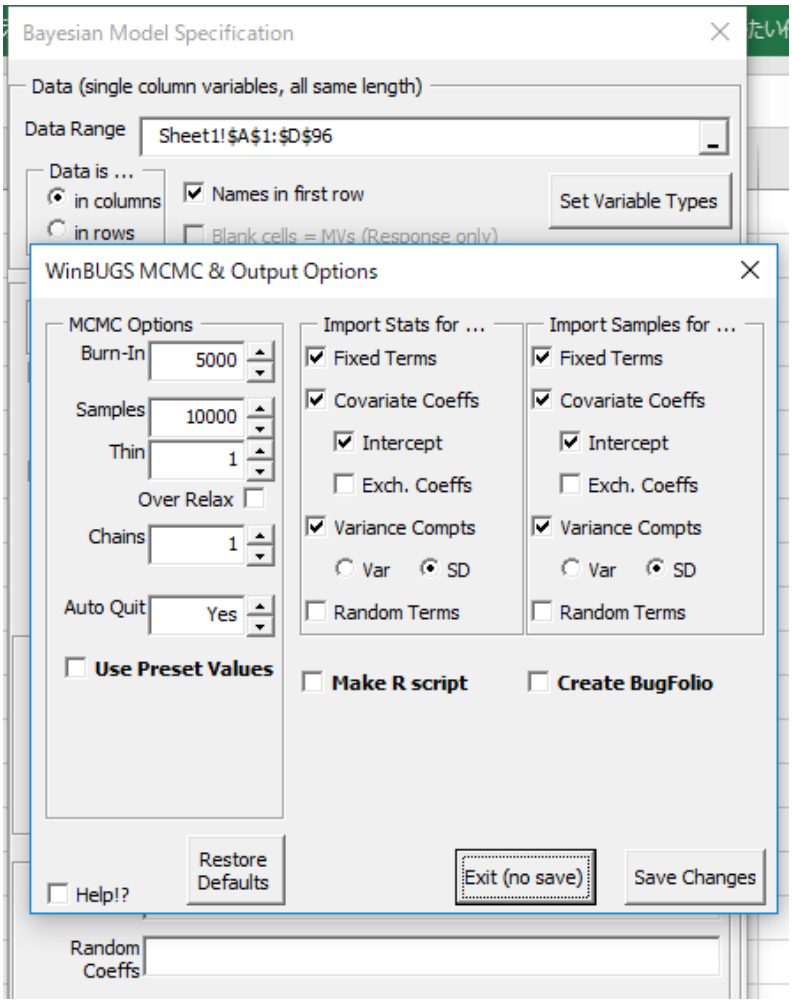

Distribution→Normal
Link→Identity

次いで、Set Variable Types をクリックし TRT を Factor に変更する。
Sort Levels を押すと並び替えが簡単にできる。




Set Variable Types

MCMC & Output Options ボタンを押すと WinBUGS MCMC & Output Options の画面が現れる

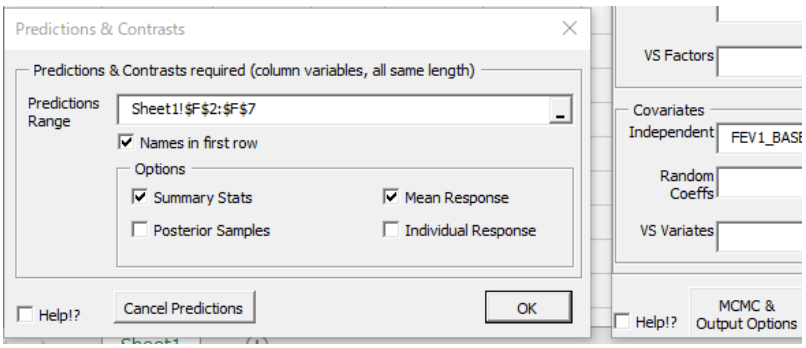



WinBUGS MCMC & Output Options

BugsXLA で Predictions or Contrasts にもチェックを入れておき知りたい項目（TRT ABCDE）を指定すると結果が計算される。

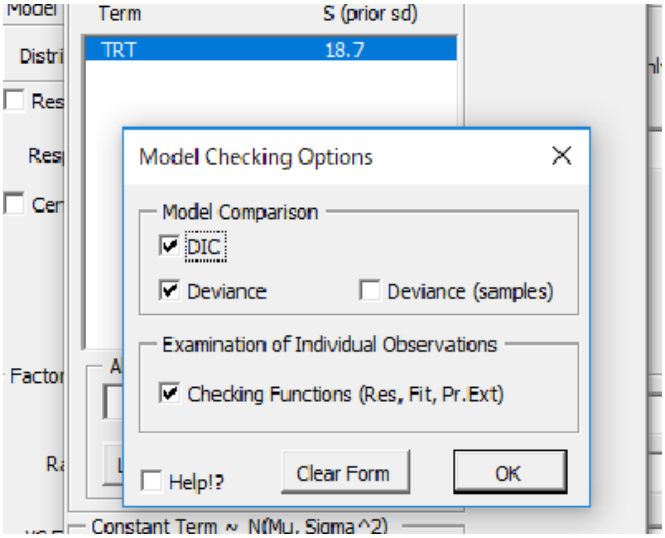
| FEV1_END | FEV1_BASE | FEV1_CFB | TRT | | |
|---|---|---|---|---|---|
| 1.93 | 1.925 | 0.005 | E | | TRT |
| 1.92 | 2 | -0.08 | D | | A |
| 1.595 | 1.45 | 0.145 | D | | B |
| 1.67 | 1.6 | 0.07 | C | | C |
| 0.495 | 0.575 | -0.08 | A | | D |
| 0.56 | 0.595 | -0.035 | C | | E |

自分でエクセルに
TRT A B C D E
を列方向に打ち込んでおき
Predictions & Contrasts
required で指定する。




Predictions & Contrasts required

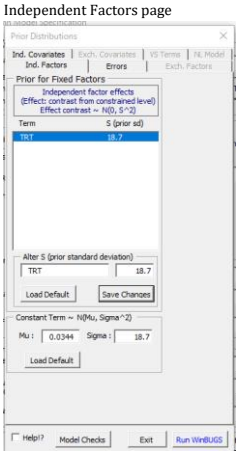
また、ここで　Model Checking Options にも必要なチェックを入れる。



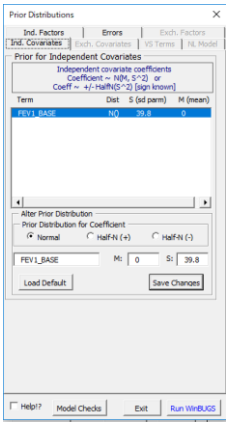

Model Checking Options

Prior Distributions の表が現れる。既にデフォルトとして数値が示されているが必要に応じて変更する。

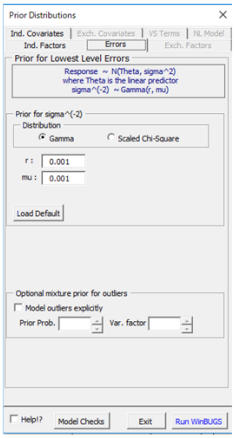
Independent Factors page



Independent Covariates page

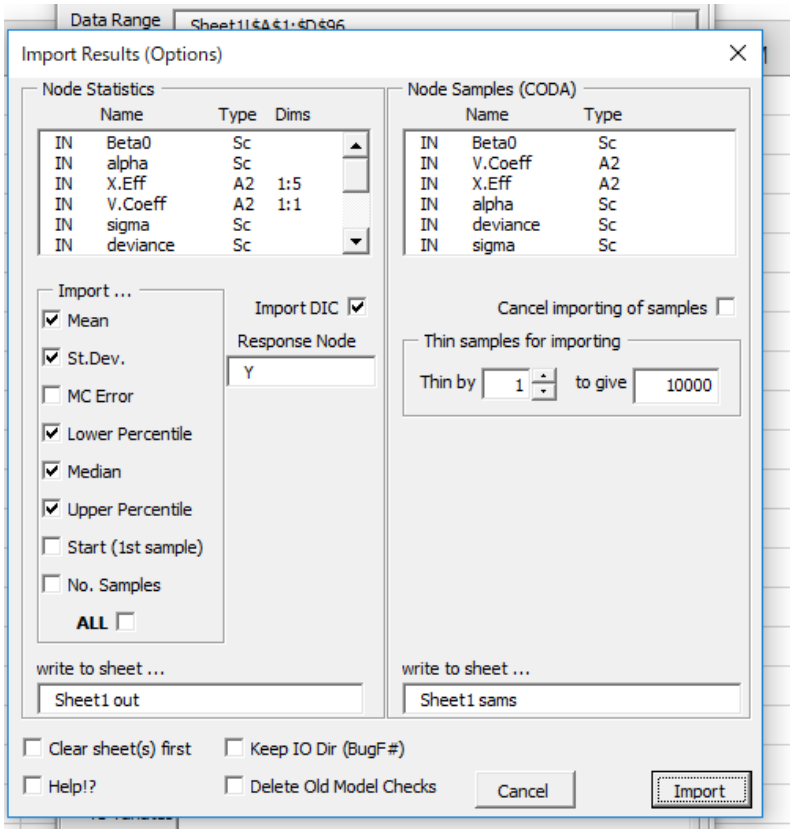


Error' page



Prior Distributions の表

Run WinBUGS のボタンをクリックすると、Imprt Results が現れるので Import ボタンを押し最終結果をエクセルに取り込む。




Imprt Results
の画面

解析結果は以下の通り

| | Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name | |
|---|---|---|---|---|---|---|---|---|---|
| | CONSTANT | -0.0400 | 0.0454 | -0.1310 | -0.0399 | 0.0487 | | Beta0 | |
| | Intercept at 0 | 0.0449 | 0.0639 | -0.0809 | 0.0441 | 0.1722 | | alpha | |
| TRT | A | 0.0000 | 0.0000 | | | | | X.Eff[1,1] | |
| TRT | B | 0.0927 | 0.0646 | -0.0334 | 0.0927 | 0.2205 | | X.Eff[1,2] | |
| TRT | C | 0.1491 | 0.0577 | 0.0368 | 0.1487 | 0.2631 | | X.Eff[1,3] | |
| TRT | D | 0.0225 | 0.0642 | -0.1038 | 0.0227 | 0.1476 | | X.Eff[1,4] | |
| TRT | E | 0.0554 | 0.0636 | -0.0703 | 0.0551 | 0.1802 | | X.Eff[1,5] | |
| | FEV1_BASE | -0.0663 | 0.0422 | -0.1501 | -0.0665 | 0.0174 | | V.Coeff[1,1] | |
| | SD(residual) | 0.1835 | 0.0140 | 0.1590 | 0.1826 | 0.2130 | | sigma | |
| | Deviance | -53.2700 | 3.9230 | -58.8400 | -53.9500 | -43.8100 | | deviance | |

WinBUGS(BugsXLA)の結果は **95%信用区間 (credible interval)**で示される。頻度主義の考えに基づく 95%信頼区間(confidence interval)に相当する。「事後確率分布から推定し、母平均の真の値が 95%の確率で含まれる区間」の意味。

ここでは、CONSTANT（FEV1_CFB）の Mean=-0.0400, SD=0.0454
median は -0.0399、95%信用区間は -0.1310~0.0487 ということになる。
TRT のなかで、treatment C の mean 0.1491, SD 0.0577, median 0.1487, 95%信用区間は 0.0368~0.2631 と区間内に 0 を含まず有意となっている。
Covariate の FEV1_BASE の回帰係数である Intrecept at 0 は、Median で 0.0441 だが 95%信用区間内に 0 を含み有意ではない。

Predictions or Contrasts の結果

| Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name |
|---|---|---|---|---|---|---|---|
| | **Predicted Mean Response** | | | | | | |
| | TRT(A) FEV1_BASE(mean) | | | | | | |
| | -0.0403 | 0.0454 | -0.1314 | -0.0401 | 0.0484 | | Pred.Ave[1] |
| | | | | | | | |
| | TRT(B) FEV1_BASE(mean) | | | | | | |
| | 0.0524 | 0.0465 | -0.0372 | 0.0521 | 0.1437 | | Pred.Ave[2] |
| | | | | | | | |
| | TRT(C) FEV1_BASE(mean) | | | | | | |
| | 0.1088 | 0.0348 | 0.0396 | 0.1092 | 0.1766 | | Pred.Ave[3] |
| | | | | | | | |
| | TRT(D) FEV1_BASE(mean) | | | | | | |
| | -0.0179 | 0.0446 | -0.1062 | -0.0174 | 0.0681 | | Pred.Ave[4] |
| | | | | | | | |
| | TRT(E) FEV1_BASE(mean) | | | | | | |
| | 0.0150 | 0.0459 | -0.0754 | 0.0152 | 0.1048 | | Pred.Ave[5] |

| | | | |
|---|---|---|---|
| **Model** | [Sheet1!$A$1:$D$96] | | |
| Distribution | Normal | | |
| Link | Identity | | |
| Response | FEV1_CFB | | |
| Fixed | TRT | | |
| Covariates | FEV1_BASE | | |
| | | | |
| **Priors** | | | |
| CONSTANT | N(mu=0.0344, sigma=18.7) | | |
| TRT | N(mu=0, sigma=18.7) | | |
| FEV1_BASE | N(mu=0, sigma=39.8) | | |
| V(residual) | Inv-Gamma(0.001, 0.001) | | |
| | | | |
| **WinBUGS MCMC Settings** | | | |
| Burn-In: 5000 Samples: 10000 (Thin:1; Chains:1) | | | |
| Run took 16 seconds | | | |
| BugsXLA (Beta 5.0) 2011.Apr.17.(00.00) | | | |

## BugsXLA による解析 2）
## 一般化線形モデル:generalized linear model(GLM)
## ポアソン回帰　Poisson regression

データは Bayesian Analysis Made Simple から BugsXLA を使用した例題(Cntrol of Cockchafer Larvae)です。

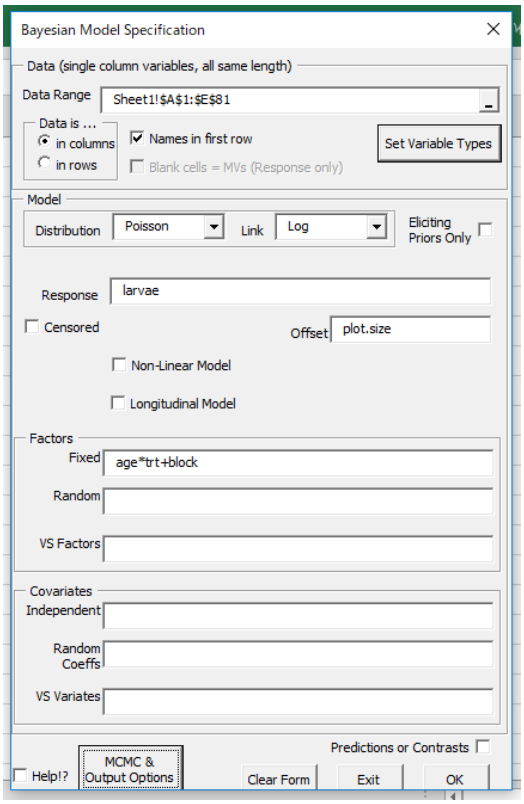
| age | block | trt | plot.size | larvae |
|---|---|---|---|---|
| a | A | 1 | 4 | 13 |
| a | A | 2 | 4 | 16 |
| a | A | 3 | 4 | 13 |
| a | A | 4 | 4 | 20 |
| a | A | 5 | 4 | 16 |
| a | B | 1 | 4 | 29 |
| a | B | 2 | 4 | 12 |
| a | B | 3 | 4 | 23 |
| a | B | 4 | 4 | 15 |
| a | B | 5 | 4 | 17 |
| a | C | 1 | 1 | 5 |
| a | C | 2 | 1 | 4 |
| a | C | 3 | 1 | 4 |
| a | C | 4 | 1 | 1 |
| a | C | 5 | 1 | 2 |
| a | D | 1 | 1 | 5 |
| a | D | 2 | 1 | 12 |
| a | D | 3 | 1 | 1 |
| a | D | 4 | 1 | 5 |
| a | D | 5 | 1 | 3 |
| a | E | 1 | 1 | 0 |
| a | E | 2 | 1 | 2 |
| a | E | 3 | 1 | 2 |
| a | E | 4 | 1 | 2 |
| a | E | 5 | 1 | 0 |
| a | F | 1 | 1 | 1 |
| a | F | 2 | 1 | 1 |
| a | F | 3 | 1 | 1 |
| a | F | 4 | 1 | 3 |
| a | F | 5 | 1 | 5 |
| a | G | 1 | 1 | 1 |
| a | G | 2 | 1 | 3 |
| a | G | 3 | 1 | 1 |
| a | G | 4 | 1 | 0 |
| a | G | 5 | 1 | 1 |
| a | H | 1 | 1 | 4 |
| a | H | 2 | 1 | 4 |
| a | H | 3 | 1 | 7 |
| a | H | 4 | 1 | 3 |
| a | H | 5 | 1 | 1 |
| b | A | 1 | 4 | 28 |
| b | A | 2 | 4 | 12 |
| b | A | 3 | 4 | 40 |
| b | A | 4 | 4 | 31 |
| b | A | 5 | 4 | 22 |
| b | B | 1 | 4 | 61 |
| b | B | 2 | 4 | 49 |
| b | B | 3 | 4 | 48 |
| b | B | 4 | 4 | 44 |
| b | B | 5 | 4 | 45 |
| b | C | 1 | 1 | 7 |
| b | C | 2 | 1 | 2 |
| b | C | 3 | 1 | 4 |
| b | C | 4 | 1 | 5 |
| b | C | 5 | 1 | 2 |
| b | D | 1 | 1 | 14 |
| b | D | 2 | 1 | 5 |
| b | D | 3 | 1 | 14 |
| b | D | 4 | 1 | 9 |
| b | D | 5 | 1 | 8 |
| b | E | 1 | 1 | 3 |
| b | E | 2 | 1 | 3 |
| b | E | 3 | 1 | 2 |
| b | E | 4 | 1 | 7 |
| b | E | 5 | 1 | 0 |
| b | F | 1 | 1 | 7 |
| b | F | 2 | 1 | 6 |
| b | F | 3 | 1 | 7 |
| b | F | 4 | 1 | 7 |
| b | F | 5 | 1 | 4 |
| b | G | 1 | 1 | 10 |
| b | G | 2 | 1 | 5 |
| b | G | 3 | 1 | 8 |
| b | G | 4 | 1 | 3 |
| b | G | 5 | 1 | 6 |
| b | H | 1 | 1 | 13 |
| b | H | 2 | 1 | 11 |
| b | H | 3 | 1 | 10 |
| b | H | 4 | 1 | 12 |
| b | H | 5 | 1 | 8 |

コフキコガネの幼虫 Cockchafer Larvae は若い木々の根に広範囲かつ致命的なダメージを与える害虫である。その制御のために 5 つの treatment, trt(1~5) を行い、その効能を調べる。土地を 8 つの block(A~H) に分け、それぞれに 5 つの trt を乱塊法で割付ける。幼虫 larvae は年齢 age により 2 つのグループ(a,b)に分け年齢も評価する。Response は trt 後にまだ生きている larvae の数（時間の関係で blockA,B ではすべての調査区域 plot を数え終え、残りの block では調査区域 plot の 1/4 の部分が数え終えている）
age, block, trt は Factors、larvae はカウントデータとして Poisson 分布に従い、plot.size が offset となる。また p55 の通り、age*trt=age+trt+age:trt であり、age:trt は交互作用(:)を意味する。

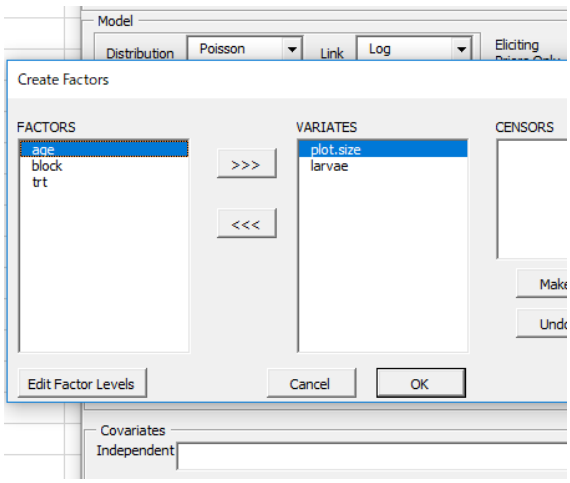
$$\eta_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \omega_k + \log(plot.size)$$
$$i = 1,2 \;\; j = 1,2,\cdots,5 \;\; k = 1,2,\cdots,8$$
$$\log(\lambda_{ijk}) = \eta_{ijk} \qquad y_{ijk} \sim \mathrm{Pois}(\lambda_{ijk})$$

$y$ は応答変数 larvae
$\alpha_i$ は Factor age の効果
$\gamma_j$ は Factor trt の効果
$(\alpha\gamma)_{ij}$ は age と trt の交互作用
$\omega_k$ は Factor block の効果

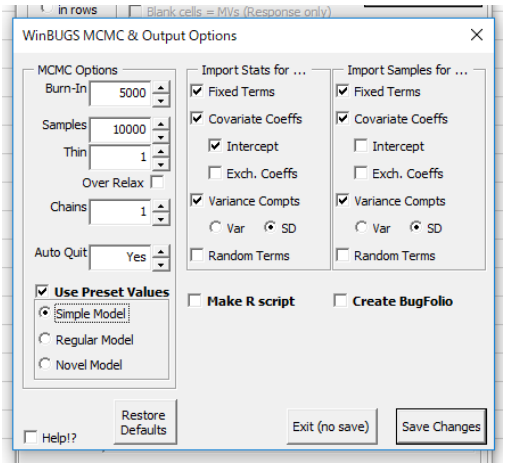
Distribution→Poisson
Link→Log



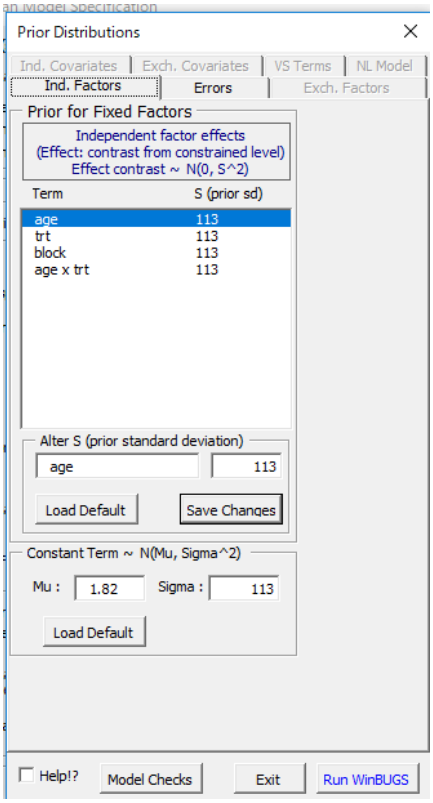

age*trt=age+trt+age:trt

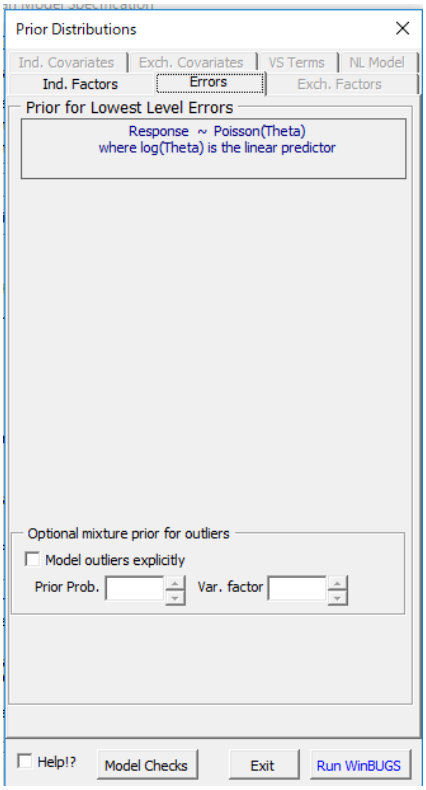


Piosson を指定すると、link 関数に Log がデフォルトで指定されている。
Factors に相当する age, block, trt を指定する。
Response は larvae、Fctors(Fixed)に age*trt+block を打ち込む。
Offset は plot.size となる。
（今回は Predictions or Contrasts は省略）
WinBUGS MCMC & Output Options をチェックし




Prior Distributions もデフォルト通り。



Run WinBUGS のボタンを押し、データをエクセルに取り込むと結果が表示される。

| | Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name | |
|---|---|---|---|---|---|---|---|---|---|
| | CONSTANT | 1.2770 | 0.1443 | 0.9837 | 1.2810 | 1.5560 | | Beta0 | |
| age | a | 0.0000 | 0.0000 | | | | | X.Eff[1,1] | |
| age | b | 0.9155 | 0.1542 | 0.6227 | 0.9124 | 1.2190 | | X.Eff[1,2] | |
| trt | 1 | 0.0000 | 0.0000 | | | | | X.Eff[2,1] | |
| trt | 2 | -0.0726 | 0.1895 | -0.4425 | -0.0753 | 0.3053 | | X.Eff[2,2] | |
| trt | 3 | -0.1047 | 0.1913 | -0.4760 | -0.1070 | 0.2734 | | X.Eff[2,3] | |
| trt | 4 | -0.1625 | 0.1973 | -0.5427 | -0.1643 | 0.2299 | | X.Eff[2,4] | |
| trt | 5 | -0.2495 | 0.2014 | -0.6597 | -0.2448 | 0.1409 | | X.Eff[2,5] | |
| block | A | 0.0000 | 0.0000 | | | | | X.Eff[3,1] | |
| block | B | 0.4878 | 0.0869 | 0.3195 | 0.4862 | 0.6566 | | X.Eff[3,2] | |
| block | C | -0.3912 | 0.1806 | -0.7604 | -0.3882 | -0.0447 | | X.Eff[3,3] | |
| block | D | 0.3620 | 0.1327 | 0.0943 | 0.3631 | 0.6212 | | X.Eff[3,4] | |
| block | E | -0.9438 | 0.2317 | -1.4170 | -0.9377 | -0.5140 | | X.Eff[3,5] | |
| block | F | -0.2384 | 0.1696 | -0.5772 | -0.2343 | 0.0838 | | X.Eff[3,6] | |
| block | G | -0.3402 | 0.1775 | -0.6983 | -0.3379 | -2.358E-4 | | X.Eff[3,7] | |
| block | H | 0.3212 | 0.1352 | 0.0563 | 0.3221 | 0.5829 | | X.Eff[3,8] | |
| age x trt | a, 1 | 0.0000 | 0.0000 | | | | | X.Eff[4,1] | |
| age x trt | b, 1 | 0.0000 | 0.0000 | | | | | X.Eff[4,2] | |
| age x trt | a, 2 | 0.0000 | 0.0000 | | | | | X.Eff[4,3] | |
| age x trt | b, 2 | -0.3619 | 0.2308 | -0.8229 | -0.3609 | 0.0929 | | X.Eff[4,4] | |
| age x trt | a, 3 | 0.0000 | 0.0000 | | | | | X.Eff[4,5] | |
| age x trt | b, 3 | 0.0283 | 0.2248 | -0.4057 | 0.0274 | 0.4677 | | X.Eff[4,6] | |
| age x trt | a, 4 | 0.0000 | 0.0000 | | | | | X.Eff[4,7] | |
| age x trt | b, 4 | -0.0353 | 0.2322 | -0.4912 | -0.0351 | 0.4176 | | X.Eff[4,8] | |
| age x trt | a, 5 | 0.0000 | 0.0000 | | | | | X.Eff[4,9] | |
| age x trt | b, 5 | -0.1654 | 0.2419 | -0.6298 | -0.1693 | 0.3213 | | X.Eff[4,10] | |
| | Deviance | 393.3000 | 5.8030 | 384.0000 | 392.7000 | 406.3000 | | deviance | |

| | | | | |
|---|---|---|---|---|
| **Model** | [Sheet1!$A$1:$E$81] | | | |
| Distribution | Poisson | | | |
| Link | Log | | | |
| Response | larvae | | | |
| Offset | plot.size | | | |
| Fixed | age*trt+block | | | |
| | | | | |
| **Priors** | | | | |
| CONSTANT | N(mu=1.82, sigma=113) | | | |
| age | N(mu=0, sigma=113) | | | |
| trt | N(mu=0, sigma=113) | | | |
| block | N(mu=0, sigma=113) | | | |
| age x trt | N(mu=0, sigma=113) | | | |
| | | | | |
| **WinBUGS MCMC Settings** | | | | |
| Burn-In: 5000 Samples: 10000 (Thin:1; Chains:1) | | | | |
| Run took 33 seconds | | | | |
| BugsXLA (Beta 5.0) 2011.Apr.17.(00.00) | | | | |

結果）

trt 、age×trt では 95% credible intervals に 0 を含み有意差をみとめなかった。

# BugsXLA による解析 3）
# 一般化線形モデル:generalized linear model(GLM)
# ラテン方格法　Latin square design

実験計画法とラテン方格法　について
実験計画法はR.A.Fisherにより
1．無作為化(randomization)：確率分布を想定できる
2．繰り返し(replication)：誤差分散の評価を可能にする
3．局所管理(local control)：偶然誤差を小さくし実験の精度の向上

$V_E$＝偶然誤差＋個体差＋温度差＋慣れの誤差＋...としたとき、この中で、大きな影響がある要因を取り上げ、その影響を除くことが重要
例えば、ある因子A（ある薬剤）を$A_1$（5μg/ml），$A_2$（10μg/ml），$A_3$（15μg/ml ）の3水準に分けて各3匹ずつのラットに一日のうち10a.m,1p.m,4p.m の3回の処理を3日間行なう。$A_1$ 、$A_2$ 、$A_3$にラットをrandom allocationしたとしても

| $A_1$ | $A_1$ | $A_1$ | $A_2$ | $A_2$ | $A_2$ | $A_3$ | $A_3$ | $A_3$ |
|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 4 | 10 | 1 | 4 | 10 | 1 | 4 |
| 一日目 | | | 二日目 | | | 三日目 | | |

このような実験では、慣れ、学習効果などの系統誤差がはいる。

**完全無作為化**　completely randomized design（＝ブロック因子なし）
全部で9回の処理の順番を無作為化

| $A_2$ | $A_3$ | $A_2$ | $A_1$ | $A_1$ | $A_3$ | $A_1$ | $A_2$ | $A_3$ |
|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 4 | 10 | 1 | 4 | 10 | 1 | 4 |
| 一日目 | | | 二日目 | | | 三日目 | | |

⇒　一元配置分散分析で解析する
この実験では一日目に$A_2$が2回処理されるなど日による影響が平等でない。

**乱塊法**　randomized block design （＝ブロック因子1個）
一日に3回の処理をブロック化し無作為化（局所管理）

| $A_2$ | $A_3$ | $A_1$ | $A_1$ | $A_3$ | $A_2$ | $A_1$ | $A_2$ | $A_3$ |
|---|---|---|---|---|---|---|---|---|
| 10 | 1 | 4 | 10 | 1 | 4 | 10 | 1 | 4 |
| 一日目 | | | 二日目 | | | 三日目 | | |

⇒　効果、ブロック（日間変動）の2要因を二元配置分散分析で解析する
このような実験でも日内変動が無視できないこともある。

**ラテン方格法**　Latin square design（＝ブロック因子2個）
各時間に処理が一回ずつになるようにする

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $A_1$ | $A_2$ | $A_3$ | $A_2$ | $A_3$ | $A_1$ | $A_3$ | $A_1$ | $A_2$ |
| 10 | 1 | 4 | 10 | 1 | 4 | 10 | 1 | 4 |
| 一日目 | | | 二日目 | | | 三日目 | | |

⇒ 要因が効果、日間変動、日内変動の3要因を三元配置分散分析で解析

ラテン方格法 Latin square design は要因数が増えると総当たり回数としての実験回数が飛躍的に増えるため、直交表などの使用が必要になり分析が難しくなるので、従来の頻度論的な解析には専門的な知識が必要。

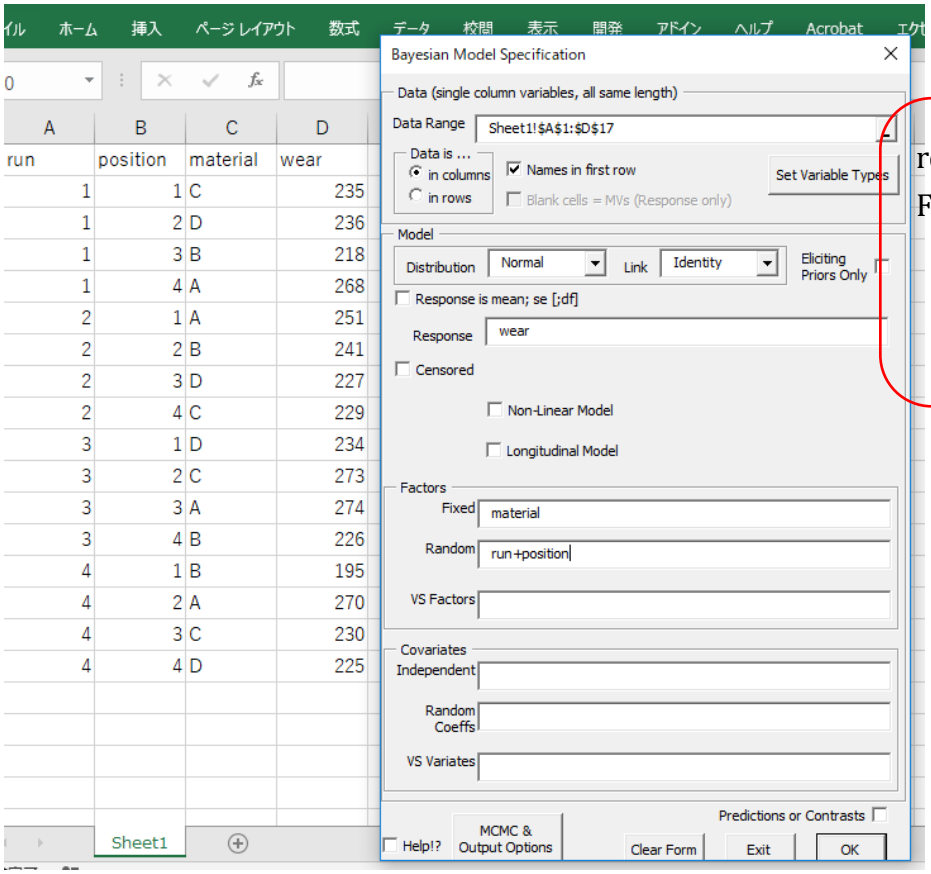
BugsXLA の解説本である Bayesian Analysis Made Simple から BugsXLA を使用した例題(Latin Square Industrial Experiment)を提示してみる。
例題）4 つの materials(A,B,C,D)で wear-testing machine　に入れ、重さの loss を 0.1mm 単位で測定した。machine には 4 つの testing positions があり、それぞれで run してテストできる。過去のデータでは、position と run の両方に系統的な誤差が存在する。ラテン方格法で解析する目的で各 material で 4 回繰り返され、計 16 回の測定が 4 回の run で行われた。materials の効果は fixed、position と run の効果は random である。

| run | position | material | wear |
|---|---|---|---|
| 1 | 1 | C | 235 |
| 1 | 2 | D | 236 |
| 1 | 3 | B | 218 |
| 1 | 4 | A | 268 |
| 2 | 1 | A | 251 |
| 2 | 2 | B | 241 |
| 2 | 3 | D | 227 |
| 2 | 4 | C | 229 |
| 3 | 1 | D | 234 |
| 3 | 2 | C | 273 |
| 3 | 3 | A | 274 |
| 3 | 4 | B | 226 |
| 4 | 1 | B | 195 |
| 4 | 2 | A | 270 |
| 4 | 3 | C | 230 |
| 4 | 4 | D | 225 |

手順）

下図の通り指定して、Set Variable Types で Factors を指定する。

MCMC & Output Options では Simple Model とする。




response: wear

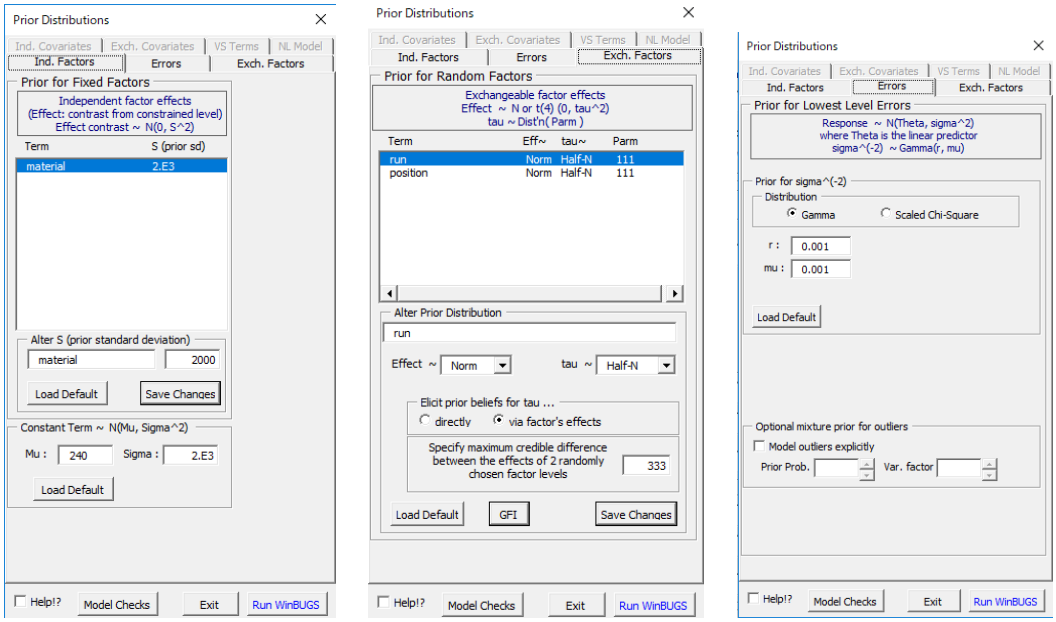
Factors

Fixed: material

Random:

run+position

Prior Distributions では下図の通り既に指定されているのでそのまま使用。



結果は下図

| | Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name | |
|---|---|---|---|---|---|---|---|---|---|
| | CONSTANT | 266.5000 | 14.8700 | 237.4000 | 266.1000 | 298.5000 | | Beta0 | |
| material | A | 0.0000 | 0.0000 | | | | | X.Eff[1,1] | |
| material | B | -45.6000 | 7.0980 | -59.8100 | -45.6000 | -30.7500 | | X.Eff[1,2] | |
| material | C | -23.9900 | 7.2850 | -39.0400 | -24.0400 | -9.4790 | | X.Eff[1,3] | |
| material | D | -35.1800 | 7.1650 | -49.6600 | -35.1800 | -20.8900 | | X.Eff[1,4] | |
| | SD(run) | 15.5400 | 14.8800 | 1.6620 | 11.4800 | 54.8900 | | sigma.Z[1] | |
| | SD(position) | 18.2100 | 14.3800 | 2.9700 | 14.3600 | 55.6200 | | sigma.Z[2] | |
| | SD(residual) | 9.4400 | 3.4090 | 5.0840 | 8.6460 | 18.2500 | | sigma | |

| **Model** | [Sheet1!$A$1:$D$17] | | | |
|---|---|---|---|---|
| Distribution | Normal | | | |
| Link | Identity | | | |
| Response | wear | | | |
| Fixed | material | | | |
| Random | run+position | | | |
| | | | | |
| **Priors** | | | | |
| CONSTANT | N(mu=240, sigma=2000) | | | |
| material | N(mu=0, sigma=2000) | | | |
| run | Norm(0,tau^2); tau ~ Half-N(sigma=111) | | | |
| position | Norm(0,tau^2); tau ~ Half-N(sigma=111) | | | |
| V(residual) | Inv-Gamma(0.001, 0.001) | | | |
| | | | | |
| **WinBUGS MCMC Settings** | | | | |
| Burn-In: 5000 Samples: 10000 (Thin:1; Chains:1) | | | | |
| Run took 13 seconds | | | | |
| BugsXLA (Beta 5.0) 2011.Apr.17.(00.00) | | | | |

事後分布の結果から、4 つの materials の摩耗率 wear rates にはその 95% credible intervals は 0 を含まないことから、明らかな違いがあることがわかる。また、run, position の分散はともに residual と比較して重要だが、小規模な実験のため、hierarchical SD parameters である SD(run), SD(position)はやや正確性が低い。

参考までに解説本には、この結果は R を使用した結果と比較し Mean は R の Estimate と近いが、St.Dev.は R の Std.Error と比較すると R の方がやや小さな値となっている。また、run と position については R では 95% credible intervals は提示されておらず、SAS と比較すると SAS(run 4.2-53, position 5.5-54, residual 5.0-17)と比べて、上限はほぼ同じだが、下限値が SAS の値の方がやや大きな値の様である。

## BugsXLA による解析 4）
## 正規線形混合モデル normal linear mixed model (NLMM, LMM)
## repeated measures design / normal hierarchical model

マルチレベル分析　multilevel analysis
階層線形モデル　hierarchical linear models →HLM（Raudenbush と Bryk）
共分散要因モデル　covariance components models
分散要因モデル　variance components models
経験的ベイズモデル　Empirical Bayes models
成長曲線モデル　growth curve models
混合効果モデル　mixed-effect model　　など多数の呼び名がある。

例）WinBUGS Examples　Volume 1 Rats: a normal hierarchical model

| | Weights $Y_{ij}$ of rat i on day $x_j$ | | | | |
|---|---|---|---|---|---|
| | $x_j$ = 8 | 15 | 22 | 29 | 36 |
| Rat 1 | 151 | 199 | 246 | 283 | 320 |
| Rat 2 | 145 | 199 | 249 | 293 | 354 |
| ……. | | | | | |
| Rat 30 | 153 | 200 | 244 | 286 | 324 |

から、データをエクセルに移し下記に整理しなおす。

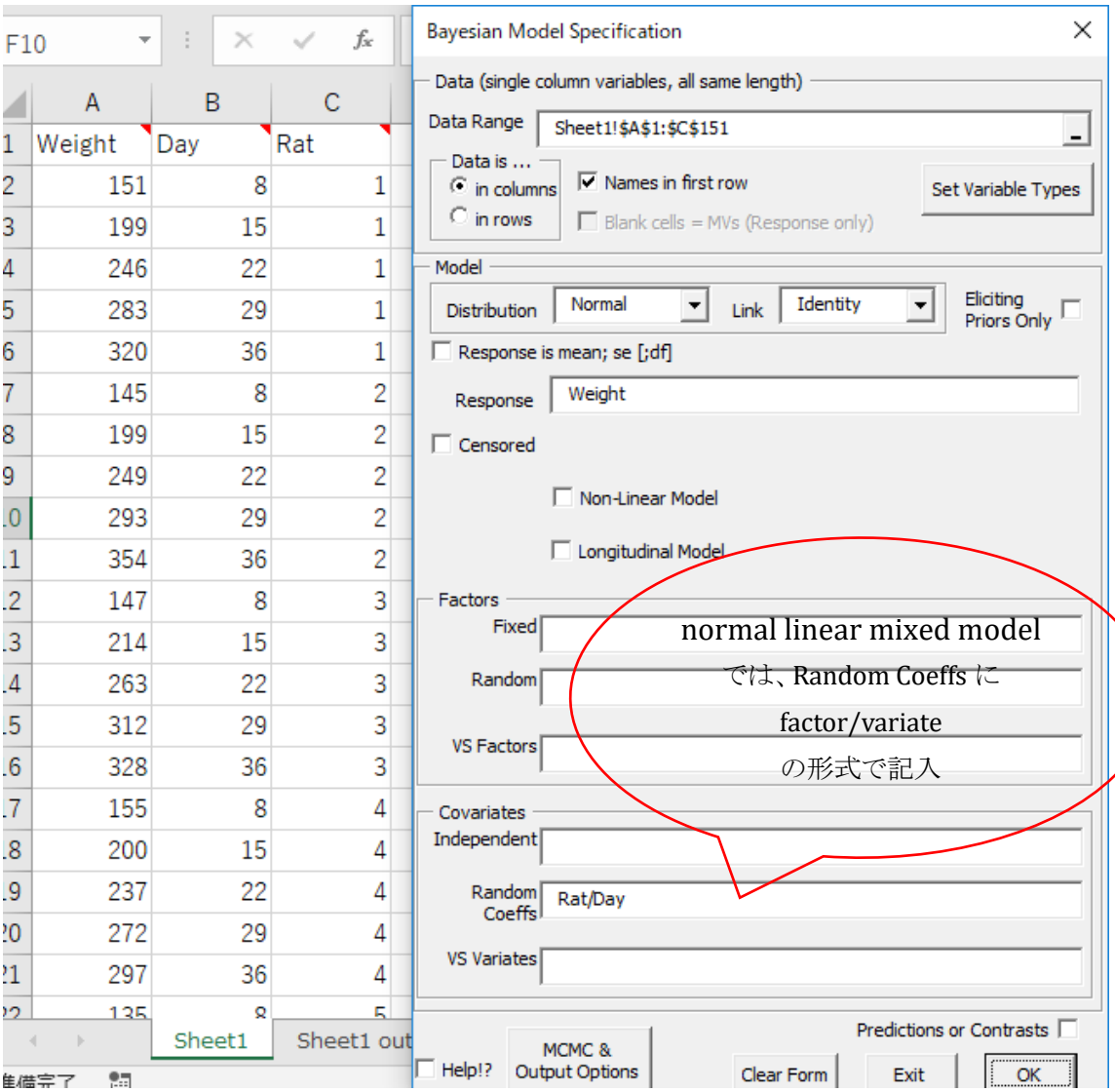
| weight | day | rat |
|---|---|---|
| 151 | 8 | 1 |
| 199 | 15 | 1 |
| 246 | 22 | 1 |
| 283 | 29 | 1 |
| 320 | 36 | 1 |
| 145 | 8 | 2 |
| 199 | 15 | 2 |
| 249 | 22 | 2 |
| 293 | 29 | 2 |
| 354 | 36 | 2 |
| 147 | 8 | 3 |
| 214 | 15 | 3 |
| 263 | 22 | 3 |
| 312 | 29 | 3 |
| 328 | 36 | 3 |
| 155 | 8 | 4 |
| … | … | … |

実際のデータは次の頁（WinBUGS から取り込み BugsXLA の形式に変換）

| Weight | Day | Rat |
|---|---|---|
| 151 | 8 | 1 |
| 199 | 15 | 1 |
| 246 | 22 | 1 |
| 283 | 29 | 1 |
| 320 | 36 | 1 |
| 145 | 8 | 2 |
| 199 | 15 | 2 |
| 249 | 22 | 2 |
| 293 | 29 | 2 |
| 354 | 36 | 2 |
| 147 | 8 | 3 |
| 214 | 15 | 3 |
| 263 | 22 | 3 |
| 312 | 29 | 3 |
| 328 | 36 | 3 |
| 155 | 8 | 4 |
| 200 | 15 | 4 |
| 237 | 22 | 4 |
| 272 | 29 | 4 |
| 297 | 36 | 4 |
| 135 | 8 | 5 |
| 188 | 15 | 5 |
| 230 | 22 | 5 |
| 280 | 29 | 5 |
| 323 | 36 | 5 |
| 159 | 8 | 6 |
| 210 | 15 | 6 |
| 252 | 22 | 6 |
| 298 | 29 | 6 |
| 331 | 36 | 6 |
| 141 | 8 | 7 |
| 189 | 15 | 7 |
| 231 | 22 | 7 |
| 275 | 29 | 7 |
| 305 | 36 | 7 |
| 159 | 8 | 8 |
| 201 | 15 | 8 |
| 248 | 22 | 8 |
| 297 | 29 | 8 |
| 338 | 36 | 8 |
| 177 | 8 | 9 |
| 236 | 15 | 9 |
| 285 | 22 | 9 |
| 350 | 29 | 9 |
| 376 | 36 | 9 |
| 134 | 8 | 10 |
| 182 | 15 | 10 |
| 220 | 22 | 10 |
| 260 | 29 | 10 |
| 296 | 36 | 10 |
| 160 | 8 | 11 |
| 208 | 15 | 11 |
| 261 | 22 | 11 |
| 313 | 29 | 11 |
| 352 | 36 | 11 |
| 143 | 8 | 12 |
| 188 | 15 | 12 |
| 220 | 22 | 12 |
| 273 | 29 | 12 |
| 314 | 36 | 12 |
| 154 | 8 | 13 |
| 200 | 15 | 13 |
| 244 | 22 | 13 |
| 289 | 29 | 13 |
| 325 | 36 | 13 |
| 171 | 8 | 14 |
| 221 | 15 | 14 |
| 270 | 22 | 14 |
| 326 | 29 | 14 |
| 358 | 36 | 14 |
| 163 | 8 | 15 |
| 216 | 15 | 15 |
| 242 | 22 | 15 |
| 281 | 29 | 15 |
| 312 | 36 | 15 |
| 160 | 8 | 16 |
| 207 | 15 | 16 |
| 248 | 22 | 16 |
| 288 | 29 | 16 |
| 324 | 36 | 16 |
| 142 | 8 | 17 |
| 187 | 15 | 17 |
| 234 | 22 | 17 |
| 280 | 29 | 17 |
| 316 | 36 | 17 |
| 156 | 8 | 18 |
| 203 | 15 | 18 |
| 243 | 22 | 18 |
| 283 | 29 | 18 |
| 317 | 36 | 18 |
| 157 | 8 | 19 |
| 212 | 15 | 19 |
| 259 | 22 | 19 |
| 307 | 29 | 19 |
| 336 | 36 | 19 |
| 152 | 8 | 20 |
| 203 | 15 | 20 |
| 246 | 22 | 20 |
| 286 | 29 | 20 |
| 321 | 36 | 20 |
| 154 | 8 | 21 |
| 205 | 15 | 21 |
| 253 | 22 | 21 |
| 298 | 29 | 21 |
| 334 | 36 | 21 |
| 139 | 8 | 22 |
| 190 | 15 | 22 |
| 225 | 22 | 22 |
| 267 | 29 | 22 |
| 302 | 36 | 22 |
| 146 | 8 | 23 |
| 191 | 15 | 23 |
| 229 | 22 | 23 |
| 272 | 29 | 23 |
| 302 | 36 | 23 |
| 157 | 8 | 24 |
| 211 | 15 | 24 |
| 250 | 22 | 24 |
| 285 | 29 | 24 |
| 323 | 36 | 24 |
| 132 | 8 | 25 |
| 185 | 15 | 25 |
| 237 | 22 | 25 |
| 286 | 29 | 25 |
| 331 | 36 | 25 |
| 160 | 8 | 26 |
| 207 | 15 | 26 |
| 257 | 22 | 26 |
| 303 | 29 | 26 |
| 345 | 36 | 26 |
| 169 | 8 | 27 |
| 216 | 15 | 27 |
| 261 | 22 | 27 |
| 295 | 29 | 27 |
| 333 | 36 | 27 |
| 157 | 8 | 28 |
| 205 | 15 | 28 |
| 248 | 22 | 28 |
| 289 | 29 | 28 |
| 316 | 36 | 28 |
| 137 | 8 | 29 |
| 180 | 15 | 29 |
| 219 | 22 | 29 |
| 258 | 29 | 29 |
| 291 | 36 | 29 |
| 153 | 8 | 30 |
| 200 | 15 | 30 |
| 244 | 22 | 30 |
| 286 | 29 | 30 |
| 324 | 36 | 30 |

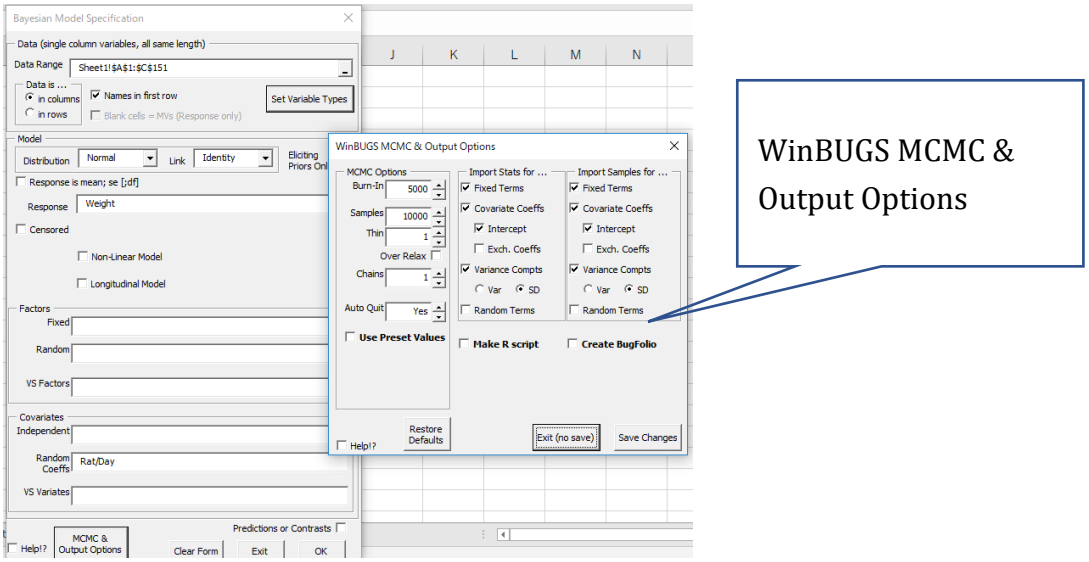
データは経時的繰り返し測定デザイン repeated measures design に属し、解析モデルは正規線形混合モデル normal linear mixed model、階層線形モデル hierarchical linear model、ランダム係数モデル random coefficient models とか成長曲線モデル growth curve model ともいわれる。

BugsXLA を起動すると Bayesian Model Specification が現れるので、一行目には必ず名前(英語で)を含めて Data Range を指定する。
データを眺めると、30 匹の young rats の成長曲線であり、その体重が 1 週ごとに 5 週間調べられている。それぞれの rat は別々の成長度であるが、階層的モデルであり、normal linear mixed model である。2 段階のモデル化を行うことで，切片や傾きについて解析できる

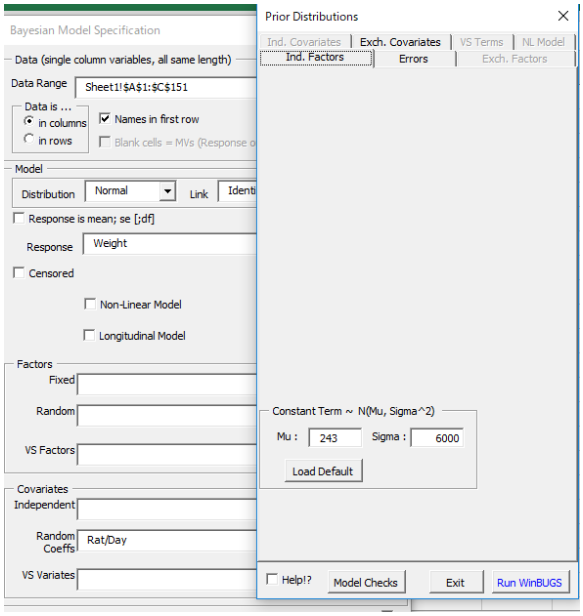


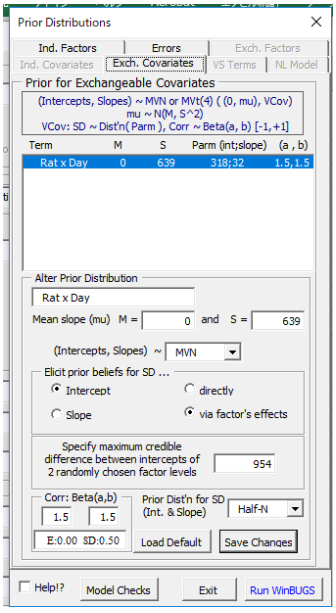
normal linear mixed model と判断した際には、BuggsXLA では、
Covariates の欄にある Random Coeffs に factor/variate の形式で記入する。
今回、Rat は Factor、Day は variate であり、Rat/Day と記入する。
Response には Weight を指定する。

ここで、MCMC & Output options のボタンを押した際、WinBUGS MCMC & Output Options が開くので、今回 Use Present Values から Simple Model を指定すると a burn-in of 5000 followed by 10,000 samples が選択される。




Bayesian Model Specification の OK を押すと、Prior Distribution の表が現れる（最初に Ind.Factors の項目）ので、Exch.Covariates の項目に変えてみると、



上表右の様に、既にいろいろ指定されている。例えば Term は Rat×Day、(Intercepts, Slopes)として MVN、その他 SD の事前分布などが指定済みになっているのでこのまま使用するが、変更も可能である。
ここで、normal linear mixed model の構造、誤差について
　y は response variable で Weight、x は回帰係数$\beta_i$を持つ covariate Day、$\alpha_i$は factor Rat の平均体重に関しての主効果を表すとしたとき、

$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \varepsilon_{ij}$$

$\alpha_i, \beta_i$ 多変量正規分布に従い、誤差は正規分布する。

$$(\alpha_i, \beta_i) \sim \mathrm{MVN}(((0, \beta_\mu), \Sigma)$$

$$\varepsilon_{ij} \quad \sim \mathrm{N}(0, \sigma^2)$$

参考）多変量正規分布（MVN）では　$\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{pmatrix}$

平均 $\mu = (\mu_1, \mu_2, \ldots, \mu_k)$ 、共分散行列 $\Sigma$ の多変量正規分布を $\mathrm{N_k}(\mu, \Sigma)$ で表す。

ここで、Model Checks で後に取り込みたい Model Comparison などにチェックを入れ、Run WinBUGS を押す。

| | | Label | Mean | St.Dev. | MC Error | 2.5% | Median | 97.5% | Start | Sample | | WinBUGS Name |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CONSTANT | 242.3000 | 2.7630 | 0.1906 | 237.2000 | 242.2000 | 248.0000 | 5001 | 10000 | | Beta0 |
| | | Intercept at 0 | 106.4000 | 2.4980 | 0.0982 | 101.5000 | 106.4000 | 111.3000 | 5001 | 10000 | | alpha |
| | (Slope mean) | Rat x Day | 6.1780 | 0.1126 | 4.603E-3 | 5.9570 | 6.1770 | 6.4040 | 5001 | 10000 | | W.MuCoeff[1] |
| | (Corr) | Rat x Day | 0.6047 | 0.1462 | 3.108E-3 | 0.2774 | 0.6207 | 0.8461 | 5001 | 10000 | | Cmvn.Beta[1] |
| | (Intercept) | SD(Rat x Day) | 14.8600 | 2.1550 | 0.0724 | 11.2800 | 14.5500 | 19.8900 | 5001 | 10000 | | sigma.WZ[1] |
| | (Slope) | SD(Rat x Day) | 0.5340 | 0.0958 | 3.550E-3 | 0.3721 | 0.5244 | 0.7473 | 5001 | 10000 | | sigma.Wcoeff[1] |
| | | SD(residual) | 6.0750 | 0.4611 | 8.382E-3 | 5.2610 | 6.0450 | 7.0520 | 5001 | 10000 | | sigma |
| | | Deviance | 965.9000 | 14.3600 | 0.2929 | 940.3000 | 965.0000 | 995.9000 | 5001 | 10000 | | deviance |
| Note: CONSTANT & Factor effects are determined at the mean of the covariate(s). | | | | | | | | | | | | |
| Var-Cov(slope-intercept) terms are also for centred data. | | | | | | | | | | | | |

WinBUGS(BugsXLA)の結果は 95%信用区間（credible interval)で示される。頻度主義の考えに基づく 95%信頼区間(confidence interval)に相当する。「事後確率分布から推定し、母平均の真の値が 95%の確率で含まれる区間」の意味。
ここでは、constant（Weight）の median は 242.2、95%信用区間は 237.2~248.0 ということになる。また、intersept at 0（切片、winbugs での alpha0）は 106.4, その 95%信用区間は 101.5~111.3 であり、slope mean（傾きの平均 W_Mu Coeff.、winbugs での beta.c）は median は 6.177、その 95%信用区間は 5.957~6.404 である。slope の SD(residual)（傾きの残差の SD、winbugs での sigma）は median6.045、その 95%信用区間は 5.261~7.052 となる。
推定された回帰式：　$y = 106.4 + 6.18x$

| | | | Y | |
|---|---|---|---|---|
| | | **Dbar** | **965.8910** | |
| | | **Dhat** | **914.3350** | |
| | | **pD** | **51.5560** | |
| | | **DIC** | **1017.4500** | |
| **Model** | [Sheet1!$A$1:$C$151] | | | |
| Distribution | Normal | | | |
| Link | Identity | | | |
| Response | Weight | | | |
| Random Coeffs | Rat/Day | | | |
| | | | | |
| **Priors** | | | | |
| CONSTANT | N(mu=243, sigma=6000) | | | |
| Rat x Day | MVN((0,mu), V); mu ~ N(mean=0, sigma=639) | | | |
| | V: sd(int) ~ Half-N(sigma=318) | | | |
| | V: sd(slope) ~ Half-N(sigma=32) | | | |
| | V: Correlation ~ Beta(1.5,1.5) [-1,+1] | | | |
| V(residual) | Inv-Gamma(0.001, 0.001) | | | |
| | | | | |
| **WinBUGS MCMC Settings** | | | | |
| Burn-In: 5000 Samples: 10000 (Thin:1; Chains:1) | | | | |
| Run took 39 seconds | | | | |
| BugsXLA (Beta 5.0) 2011.Apr.17.(00.00) | | | | |

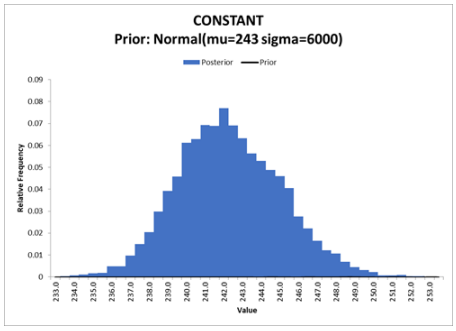
下図は WinBUGS で直接解析した例題の結果を転載した。

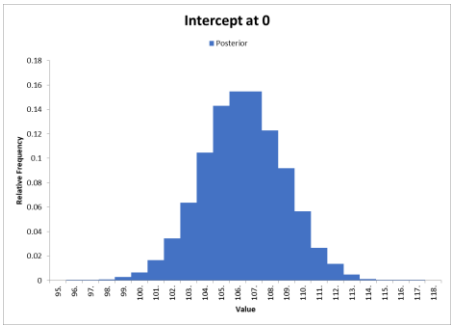
## Results

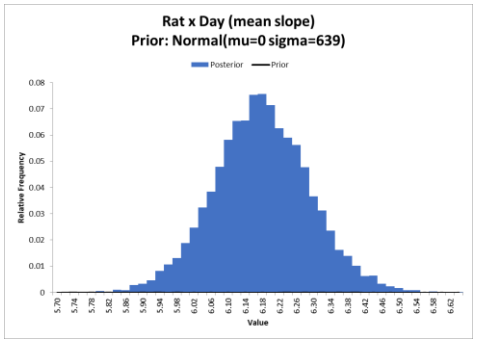
A 1000 update burn in followed by a further 10000 updates gave the parameter estimates:

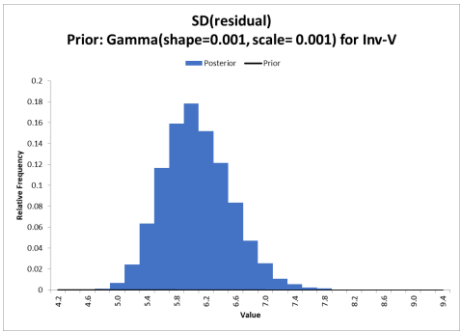
| node | mean | sd | MC error | 2.5% | median | 97.5% | start | sample |
|---|---|---|---|---|---|---|---|---|
| alpha0 | 106.6 | 3.625 | 0.03477 | 99.32 | 106.6 | 113.6 | 1001 | 10000 |
| beta.c | 6.185 | 0.1068 | 0.001354 | 5.979 | 6.184 | 6.398 | 1001 | 10000 |
| sigma | 6.082 | 0.4714 | 0.007308 | 5.248 | 6.052 | 7.093 | 1001 | 10000 |

また、BuggsXLA ではグラフ作成もできるのでいくつか作成してみた。



# BugsXLA による解析 5）
# GLM ・ GLMM
# Binominal data / meta-analysis

データは Bayesian Analysis Made Simple から BugsXLA を使用した例題 (Respiratory Tract Infectios)。

| study | trt | infected | total |
|---|---|---|---|
| 1 | T | 7 | 47 |
| 1 | C | 25 | 54 |
| 2 | T | 4 | 38 |
| 2 | C | 24 | 41 |
| 3 | T | 20 | 96 |
| 3 | C | 37 | 95 |
| 4 | T | 1 | 14 |
| 4 | C | 11 | 17 |
| 5 | T | 10 | 48 |
| 5 | C | 26 | 49 |
| 6 | T | 2 | 101 |
| 6 | C | 13 | 84 |
| 7 | T | 12 | 161 |
| 7 | C | 38 | 170 |
| 8 | T | 1 | 28 |
| 8 | C | 29 | 60 |
| 9 | T | 1 | 19 |
| 9 | C | 9 | 20 |
| 10 | T | 22 | 49 |
| 10 | C | 44 | 47 |
| 11 | T | 25 | 162 |
| 11 | C | 30 | 160 |
| 12 | T | 31 | 200 |
| 12 | C | 40 | 185 |
| 13 | T | 9 | 39 |
| 13 | C | 10 | 41 |
| 14 | T | 22 | 193 |
| 14 | C | 40 | 185 |
| 15 | T | 0 | 45 |
| 15 | C | 4 | 46 |
| 16 | T | 31 | 131 |
| 16 | C | 60 | 140 |
| 17 | T | 4 | 75 |
| 17 | C | 12 | 75 |
| 18 | T | 31 | 220 |
| 18 | C | 42 | 225 |
| 19 | T | 7 | 55 |
| 19 | C | 26 | 57 |
| 20 | T | 3 | 91 |
| 20 | C | 17 | 92 |
| 21 | T | 14 | 25 |
| 21 | C | 23 | 23 |
| 22 | T | 3 | 65 |
| 22 | C | 6 | 68 |

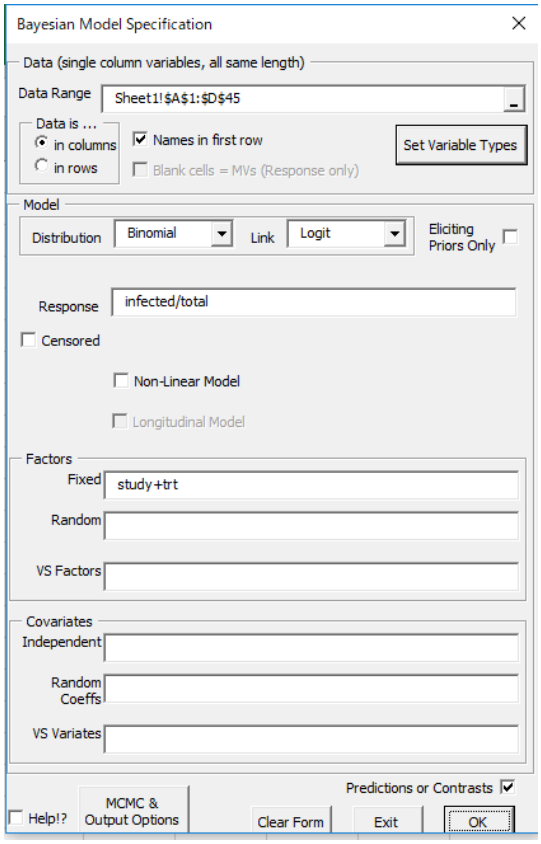
22 施設での RCT の meta-analysis の解析。ICU における呼吸器感染症に関して、消化管で選択的に除菌した場合（trt: T, treatment, C, control）、有効かどうかを調べる。

2 つの方法で解析する。

1）Generalized Linear Model (GLM)

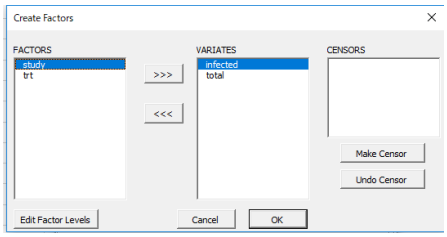
2）Generalized Linear Mixed Model (GLMM)

1）Generalized Linear Model (GLM)で解析した場合

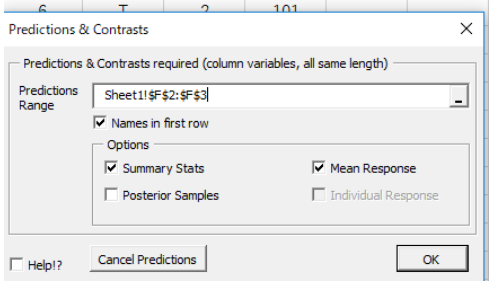



Binominal では Response に infected/total の形式で記入

GLM なので
Fixed Factors に study+trt を指定




| study | trt | infected | total | |
|---|---|---|---|---|
| 1 | T | 7 | 47 | trt |
| 1 | C | 25 | 54 | c(T, C) |
| 2 | T | 4 | 38 | |
| 2 | C | 24 | 41 | |
| 3 | T | 20 | 96 | |
| 3 | C | 37 | 95 | |
| 4 | T | 1 | 14 | |
| 4 | C | 11 | 17 | |
| 5 | T | 10 | 48 | |
| 5 | C | 26 | 49 | |



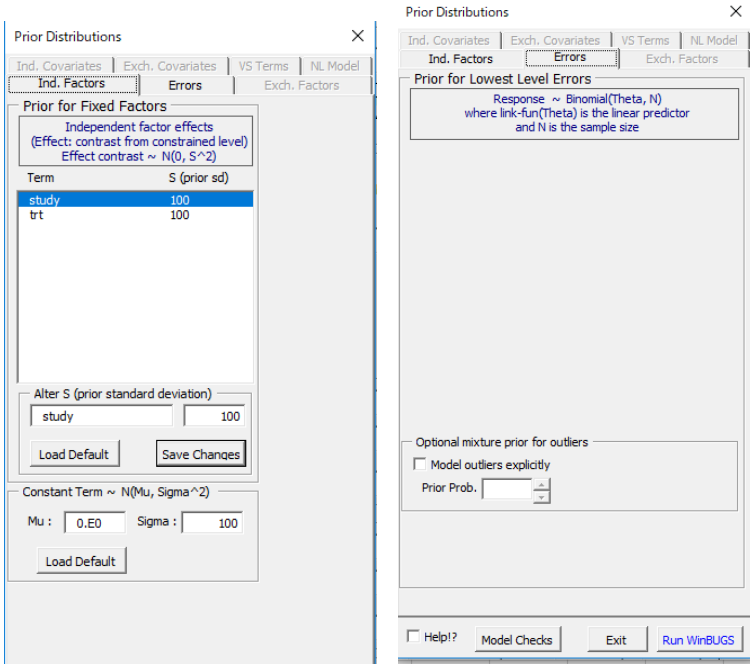

Predictions or Contrasts にチェックを入れ、あらかじめ、エクセルに contrast の指定形式

trt
c(T,C)

で打ち込んでおき、
Predictions & Contrasts の画面が出た際に指定する。

Model Distribution で Binominal を指定すると、Link Function ではデフォルトで Logit が指定されている。（その他では Probit など）
Response に infected/total の形式で記入(→ P55)

Set Variable Types で Factors を指定し、Fixed Factors に study+trt を指定する。
Predictions & Contrasts を指定する
Prior Distributions を指定して Run WinBUGS ボタンを押しエクセルに取り込む。




以下の結果が示される。

Predictions & Contrasts の結果

| Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name | |
|---|---|---|---|---|---|---|---|---|
| | Predicted Odds | | | | | | | |
| | study(1) contr[trt T / C] | | | | | | | |
| | 0.3452 | 0.0309 | 0.2881 | 0.3440 | 0.4100 | Pred.Odds[1] | | |

| Model | [Sheet1!$A$1:$D$45] | | |
|---|---|---|---|
| Distribution | Binomial | | |
| Link | Logit | | |
| Response | infected/total | | |
| Fixed | study+trt | | |
| | | | |
| Priors | | | |
| CONSTANT | N(mu=0, sigma=100) | | |
| study | N(mu=0, sigma=100) | | |
| trt | N(mu=0, sigma=100) | | |
| | | | |
| WinBUGS MCMC Settings | | | |
| Burn-In: 5000 Samples: 10000 (Thin:1; Chains:1) | | | |
| Run took 16 seconds | | | |
| BugsXLA (Beta 5.0) 2011.Apr.17.(00.00) | | | |

| | Y |
|---|---|
| Dbar | 271.2050 |
| Dhat | 248.1680 |
| pD | 23.0370 |
| DIC | 294.2420 |

| | Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name |
|---|---|---|---|---|---|---|---|---|
| | CONSTANT | -0.3245 | 0.2097 | -0.7381 | -0.3296 | 0.0930 | | Beta0 |
| study | 1 | 0.0000 | 0.0000 | | | | | X.Eff[1,1] |
| study | 2 | 0.1898 | 0.3202 | -0.4457 | 0.1921 | 0.8175 | | X.Eff[1,2] |
| study | 3 | -0.0577 | 0.2627 | -0.5744 | -0.0557 | 0.4509 | | X.Eff[1,3] |
| study | 4 | 0.2950 | 0.4355 | -0.5690 | 0.3019 | 1.1380 | | X.Eff[1,4] |
| study | 5 | 0.2856 | 0.3037 | -0.3061 | 0.2879 | 0.8750 | | X.Eff[1,5] |
| study | 6 | -1.6710 | 0.3438 | -2.3620 | -1.6660 | -1.0190 | | X.Eff[1,6] |
| study | 7 | -0.9903 | 0.2607 | -1.5060 | -0.9866 | -0.4852 | | X.Eff[1,7] |
| study | 8 | -0.0479 | 0.3136 | -0.6735 | -0.0447 | 0.5503 | | X.Eff[1,8] |
| study | 9 | -0.3208 | 0.4331 | -1.1950 | -0.3123 | 0.5067 | | X.Eff[1,9] |
| study | 10 | 1.7180 | 0.3085 | 1.1140 | 1.7180 | 2.3330 | | X.Eff[1,10] |
| study | 11 | -0.8195 | 0.2582 | -1.3340 | -0.8175 | -0.3142 | | X.Eff[1,11] |
| study | 12 | -0.7016 | 0.2452 | -1.1860 | -0.6996 | -0.2273 | | X.Eff[1,12] |
| study | 13 | -0.4103 | 0.3396 | -1.0900 | -0.4050 | 0.2571 | | X.Eff[1,13] |
| study | 14 | -0.8611 | 0.2501 | -1.3620 | -0.8568 | -0.3747 | | X.Eff[1,14] |
| study | 15 | -2.4820 | 0.5840 | -3.7500 | -2.4540 | -1.4490 | | X.Eff[1,15] |
| study | 16 | 0.1073 | 0.2466 | -0.3780 | 0.1068 | 0.5878 | | X.Eff[1,16] |
| study | 17 | -1.4070 | 0.3436 | -2.0840 | -1.3980 | -0.7501 | | X.Eff[1,17] |
| study | 18 | -0.8794 | 0.2463 | -1.3670 | -0.8763 | -0.3979 | | X.Eff[1,18] |
| study | 19 | -0.0948 | 0.2985 | -0.6799 | -0.0905 | 0.4786 | | X.Eff[1,19] |
| study | 20 | -1.3730 | 0.3184 | -2.0260 | -1.3630 | -0.7778 | | X.Eff[1,20] |
| study | 21 | 2.2010 | 0.4181 | 1.4040 | 2.1900 | 3.0350 | | X.Eff[1,21] |
| study | 22 | -1.9520 | 0.4150 | -2.8030 | -1.9330 | -1.1860 | | X.Eff[1,22] |
| trt | C | 0.0000 | 0.0000 | | | | | X.Eff[2,1] |
| trt | T | -1.0670 | 0.0894 | -1.2450 | -1.0670 | -0.8916 | | X.Eff[2,2] |

trt の結果は最後の 2 行に出ており、T (treatment) については、
95% credible interval は logit scale で示され
mean= -1.0670 SD=0.0894
median=-1.0670　95% credible interval は、-1.2450 ~-0.8916　となる。
結果を binominal response でみるには、Predictions & Contrasts の結果で、odds ratio が、BugsXLA では、contr[trt T / C] で示され、study(1) の結果が示される。この解析法では study1-22 まで odds ratio は同じとして、95% credible interval　0.34(0.29,0.41)となる。テキストには、この値は Mantel-Haenszel-Peto 法で得られた結果によく似ていると記載されている。

* Binominal data とロジスティック回帰モデルについて

二項分布 binominal distribution, bin(n,p)、例えば、サイコロを n 回振った時 1 の目は$x$回出るとき、$p(x) =_{\mathrm{n}} C_x p^x (1-p)^{n-x}$ 、$\mu = np$ 、$\sigma^2 = np(1-p)$
となる。ここで、あることが起こる確率を$p$とすると、起こらない確率は$(1-p)$となり、
$\frac{p}{1-p}$ はオッズ（odds）と呼ばれ、、$\lambda = \mathrm{logit}p = log_e(\frac{p}{1-p})$ という変換を考えれば
ロジスティック回帰モデルとなる。

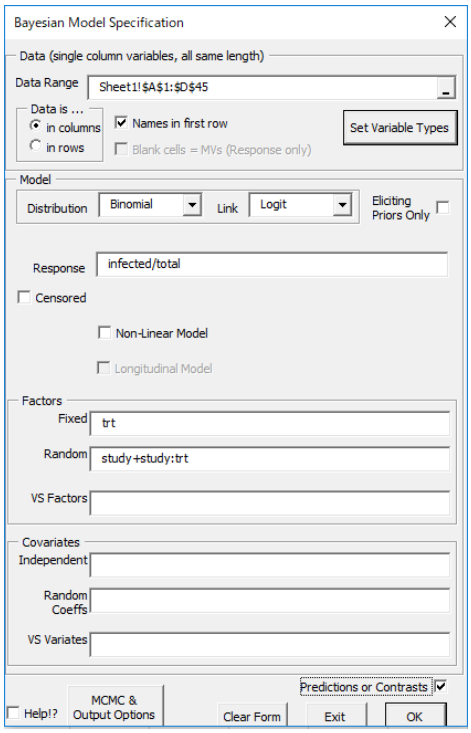
$\mathrm{logit}p = log_e(\frac{p}{1-p}) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$ と定義すれば、ロジスティック回帰モデルとなり、GLM の一種である。ある集団（予後因子のない）で起きる確 $p_{x_{i=0}}$ 、ある集団（予後因子のある）で起きる確率$p_{x_{i=1}}$ とすると
$\left(\frac{p_{x_{i=1}}}{1-p_{x_{i=1}}}\right) \Big/ \left(\frac{p_{x_{i=0}}}{1-p_{x_{i=0}}}\right)$ はオッズ比となる。

*データから得られる分散が、平均から推定される分散に比べて大きすぎるとき、過分散(overdispersion)といい、Binominal data や Poisson data などで問題になる。観測されていない要因（個体差、場所の差など）を無視して単純化して解析した場合に起こりやすいが、GLM → GLMM、つまり random coeffient model として解析する方が解決の一つの方法となる。
今回のデータでも GLM では Study を一緒にして解析している点が問題になる。
そこで、「22 施設での RCT の meta-analysis の解析。ICU における呼吸器感染症に関して、消化管で選択的に除菌した場合（trt: T, treatment, C, control）、有効かどうかを GLMM で調べる。」

再掲） ‘Factors’や’Independent’ Covariates に記載する際の約束事は以下：

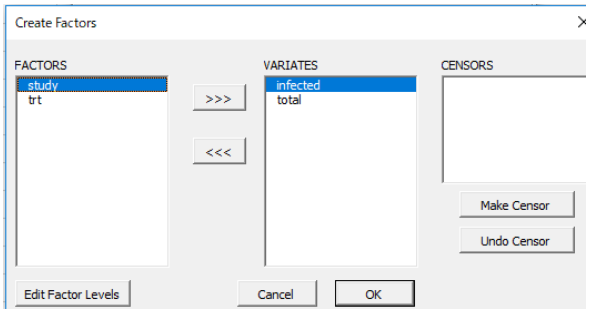
| | |
|---|---|
| + 変数（名）を加える | － 変数（名）を引く |
| : 2 変数間に交互作用 | / A/B=A+A:B |
| * A*B=A+B+A:B | @ A*B*C@2=A*B*C－A:B:C |
| ( ) (A+B)*C=A+B+C+A:C+B:C | ^ quadratic |

2）Generalized Linear Mixed Model (GLMM) で解析した場合

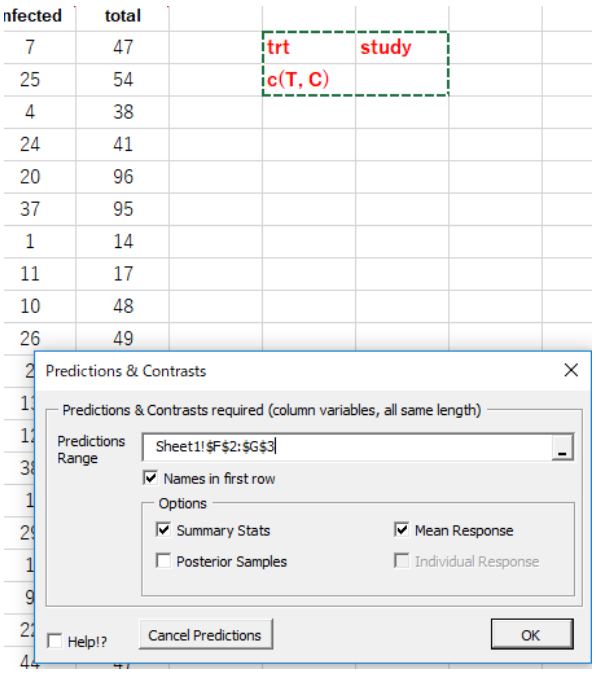



Binominal では Response に infected/total の形式で記入するのは GLM と同じ

GLMM では
Fixed Factors に trt
Random Factors に study+ study :trt を指定



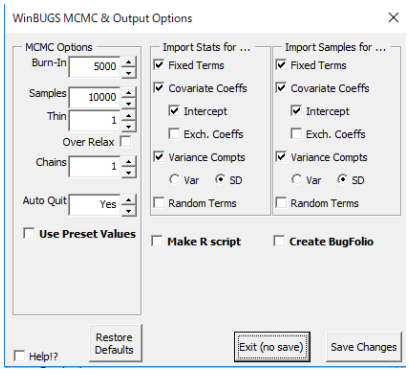

Predictions or Contrasts 画面




Predictions or Contrasts にチェックを入れ、あらかじめ、エクセルに contrast の指定形式

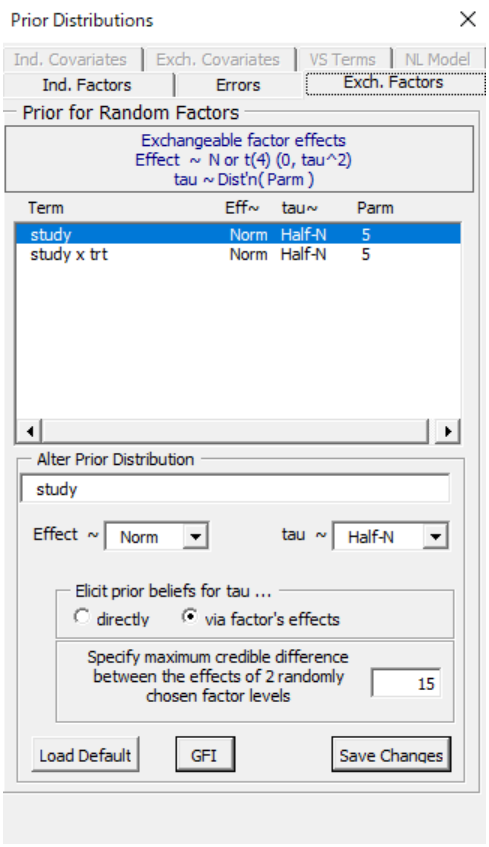
trt study
c(T,C)

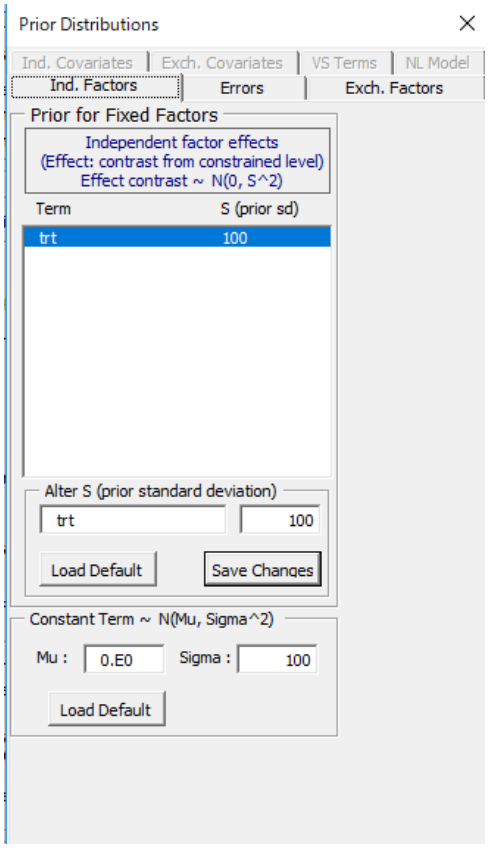
で打ち込んでおき、
Predictions & Contrasts の画面が出た際に指定する。



Prior Distributions

下図の通り、デフォルトで指定されるのでそのまま使用。





22 の study は違う研究だが、よく似ている研究で階層的またはネストされたデータになっている。その類似性は共通の分布から独立して at random に取り出した研究データの効果と考えればよい。共通の分布の SD がその効果の類似性の程度を決定づけている。また、random effect には study:trt（交互作用）が含まれ、これが、GLM で解析した場合との違いとなる。交互作用項は trt の効果の相違の部分を示している。固定効果の主効果は'population mean 母平均' treatment effect である。

Run WinBUGS ボタンを押すと以下の解析結果が示される。

| | Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name |
|---|---|---|---|---|---|---|---|---|
| | CONSTANT | -0.6269 | 0.2920 | -1.2290 | -0.6273 | -0.0507 | | Beta0 |
| trt | C | 0.0000 | 0.0000 | | | | | X.Eff[1,1] |
| trt | T | -1.3970 | 0.2281 | -1.8770 | -1.3850 | -0.9773 | | X.Eff[1,2] |
| | SD(study) | 1.0660 | 0.2334 | 0.6724 | 1.0410 | 1.5860 | | sigma.Z[1] |
| | SD(study x trt) | 0.5987 | 0.1556 | 0.3445 | 0.5798 | 0.9589 | | sigma.Z[2] |

| Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name |
|---|---|---|---|---|---|---|---|
| | Predicted Odds | | | | | | |
| | contr[trt T / C] study(dist. mean) | | | | | | |
| | 0.2538 | 0.0572 | 0.1531 | 0.2504 | 0.3763 | Pred.Odds[1] | |

| Model | [Sheet1!$A$1:$D$45] |
|---|---|
| Distribution | Binomial |
| Link | Logit |
| Response | infected/total |
| Fixed | trt |
| Random | study+study:trt |
| | |
| **Priors** | |
| CONSTANT | N(mu=0, sigma=100) |
| trt | N(mu=0, sigma=100) |
| study | Norm(0,tau^2); tau ~ Half-N(sigma=5) |
| study x trt | Norm(0,tau^2); tau ~ Half-N(sigma=5) |
| | |
| **WinBUGS MCMC Settings** | |
| Burn-In: 5000 Samples: 10000 (Thin:1; Chains:1) | |
| Run took 18 seconds | |
| BugsXLA (Beta 5.0) 2011.Apr.17.(00.00) | |

| | Y |
|---|---|
| Dbar | 217.8170 |
| Dhat | 181.1750 |
| pD | 36.6420 |
| DIC | 254.4590 |

study の結果は、Predictions or Contrasts に範囲を指定する際に、

| trt | study |
|---|---|
| c(T, C) | * |
| c(T, C) | |

と指定すると後で Forrest Plot を書くことができる。

study1-22 までそれぞれの odds ratio が示される。

| Label | Mean | St.Dev. | 2.5% | Median | 97.5% |
|---|---|---|---|---|---|
| | Predicted Odds | | | | |
| | Following all at: contr[trt T / C] | | | | |
| study(1) | 0.2253 | 0.0992 | 0.0886 | 0.2085 | 0.4694 |
| study(2) | 0.1315 | 0.0660 | 0.0411 | 0.1198 | 0.2945 |
| study(3) | 0.3947 | 0.1246 | 0.2054 | 0.3762 | 0.6849 |
| study(4) | 0.1397 | 0.0957 | 0.0262 | 0.1171 | 0.3821 |
| study(5) | 0.2481 | 0.1043 | 0.1008 | 0.2285 | 0.5024 |
| study(6) | 0.1893 | 0.1012 | 0.0547 | 0.1689 | 0.4409 |
| study(7) | 0.2852 | 0.0951 | 0.1380 | 0.2711 | 0.5096 |
| study(8) | 0.1185 | 0.0686 | 0.0288 | 0.1047 | 0.2891 |
| study(9) | 0.1727 | 0.1136 | 0.0372 | 0.1470 | 0.4572 |
| study(10) | 0.1142 | 0.0556 | 0.0370 | 0.1042 | 0.2461 |
| study(11) | 0.7058 | 0.2082 | 0.3862 | 0.6745 | 1.1860 |
| study(12) | 0.6172 | 0.1614 | 0.3637 | 0.5952 | 0.9931 |
| study(13) | 0.6791 | 0.3431 | 0.2439 | 0.6045 | 1.5490 |
| study(14) | 0.4437 | 0.1259 | 0.2469 | 0.4269 | 0.7380 |
| study(15) | 0.2457 | 0.1797 | 0.0465 | 0.2033 | 0.7080 |
| study(16) | 0.3989 | 0.1043 | 0.2324 | 0.3861 | 0.6402 |
| study(17) | 0.3130 | 0.1543 | 0.1052 | 0.2851 | 0.6942 |
| study(18) | 0.6599 | 0.1710 | 0.3860 | 0.6403 | 1.0530 |
| study(19) | 0.2014 | 0.0852 | 0.0789 | 0.1868 | 0.4055 |
| study(20) | 0.2057 | 0.1000 | 0.0692 | 0.1875 | 0.4463 |
| study(21) | 0.1437 | 0.0914 | 0.0306 | 0.1231 | 0.3693 |
| study(22) | 0.4309 | 0.2652 | 0.1215 | 0.3699 | 1.1110 |
| | | | | | |
| | contr[trt T / C] study(dist. mean) | | | | |
| | 0.2499 | 0.0523 | 0.1571 | 0.2461 | 0.3649 |

これをグラフ化すれば Forrest Plot

結果としては Odds Ratio で見るのがわかりやすい。
GLM では

| Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name | |
|---|---|---|---|---|---|---|---|---|
| | Predicted Odds | | | | | | | |
| | study(1) contr[trt T / C] | | | | | | | |
| | 0.3452 | 0.0309 | 0.2881 | 0.3440 | 0.4100 | Pred.Odds[1] | | |

| | Y |
|---|---|
| Dbar | 271.2050 |
| Dhat | 248.1680 |
| pD | 23.0370 |
| DIC | 294.2420 |

GLMM では

| Label | Mean | St.Dev. | 2.5% | Median | 97.5% | | WinBUGS Name | |
|---|---|---|---|---|---|---|---|---|
| | Predicted Odds | | | | | | | |
| | contr[trt T / C] study(dist. mean) | | | | | | | |
| | 0.2538 | 0.0572 | 0.1531 | 0.2504 | 0.3763 | Pred.Odds[1] | | |

| | Y |
|---|---|
| Dbar | 217.8170 |
| Dhat | 181.1750 |
| pD | 36.6420 |
| DIC | 254.4590 |

また、GLMM の方が GLM より、DIC が小さく望ましいと考えられる。

主な参考文献

Annette J. Dobson (2008): 一般化線形モデル入門第 2 版
（田中豊・森川敏彦・山中竹春・冨田誠　訳）共立出版
Phil Woodward(2012) : Bayesian Analysis Made Simple CRC Press
久保川達也(2017): 現代数理統計学の基礎　　共立出版
丹後俊郎(2011) : ベイジアン統計解析の実際　　朝倉書店
馬場真哉 (2017): 平均分散から始める一般化線形モデル入門　プレアデス出版
涌井良幸、涌井貞美（2016): 身につくベイズ統計学　技術評論社
インターネットから例題、解説を利用（ウェブサイト表示）