第 6 章　確率地図

第 4 章や第 5 章で扱った、都道府県別の死因別死亡者数のようなデータは、イベントの発生がまれであるが、観測回数が多く繰り返されるようなデータである。そのため、イベントデータの集計期間や空間集計単位によっては、本来観測されるべきイベントが観測されず、発生数=0 となる場合がある。死因別死亡者数のような観測値の大小や空間的自己相関の有無を地域間で比較する際には、死因別死亡者数を単純比較するのではなく、死因別死亡者数を人口数で割った比率を比較する方が望ましい。死因別死亡数が人口数と比較して非常に小さい値をとる場合、死因別死亡数／人口数＝死因別死亡率のデータを地区間で比較することは容易ではない。本章では、比率値を用いて地区属性を比較する際に用いられる、確率地図や相対リスクのベイズ推定法を紹介する。これらの方法は、空間疫学やリスク分析などに用いられる[1]。また地区属性の類似性や近接性など、空間クラスタを見つける際にも有用である。

6.1 粗率

死亡や犯罪発生など、まれに発生するような空間現象を、確率値として比較したい場合には、粗率、相対危険度、ポアソン確率地図などが適用される[2]。

いま、地区 $i(=1,\ldots,N)$について、人口数を y_i、観測ケース（例えば、心疾患による死亡者数の観測値）の標本値を O_iとする。このとき、人口数に対して観測ケースが発生する粗率 r_iは次式のように求められる。

総務省統計局の「社会生活統計指標－都道府県の指標－」[3]及び「都道府県別死因の分析結果について」[4]から、2006 年の人口や心疾患による死亡者数に関するデータを用いて、都道府県別心疾患死亡者数の観測値と粗率を、それぞれ図 6.1 と図 6.2 に示す。

$$r_i = \frac{O_i}{y_i}$$

6.2 相対危険度

粗率 r_iは、対象地域全体の人口数や観測ケースの値を反映していないため、人口数に対する観測値の大小を判断することは容易でない。そこで、対象地域全体で観測ケースの発生率が均一であると仮定して、観測ケースの標本値 O_iを期待値

E_iで除した相対リスクθ_iを用いることがある。都道府県別心疾患死亡者数の期待値と相対危険度を、図 6.3 と図 6.4 に示す。

$$E_i = y_i \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} y_i}$$

$$\theta_i = \frac{O_i}{E_i}$$

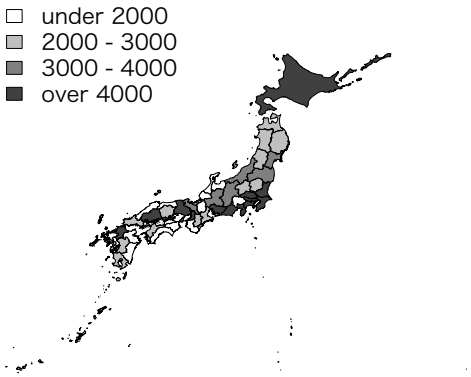


図 6.1 観測値

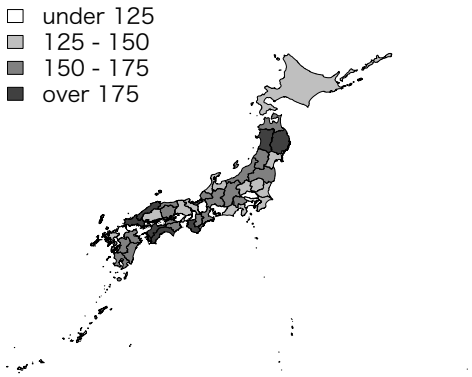


図 6.2 粗率

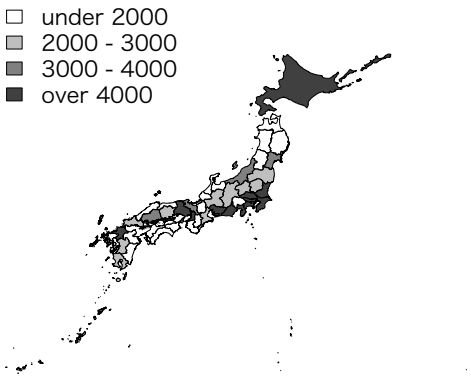


図 6.3 期待値

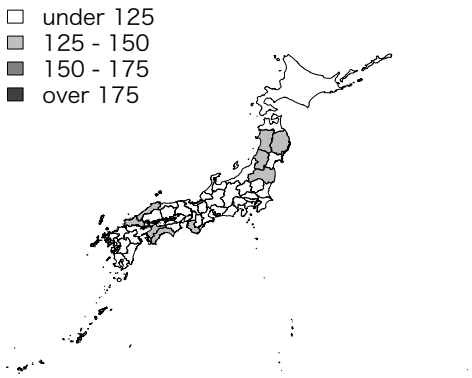


図 6.4 相対危険度

6.3 ポアソン確率地図

観測ケースの発生頻度が低いものの、非常に多く観測回数が繰り返される場合には、ポアソン確率地図を用いることで、人口規模による観測値の過大／過小評価の影響を考慮することができる。

期待値E_iが与えられたとき、観測値O_iがポアソン分布に従うとする。このとき、次式のように表すことができる。

$$O_i \sim Po(E_i)$$

期待値E_iを平均μ_iとみなしたとき、次式を用いて観測ケースの発生リスクを表現することができる。

$$p_i = \sum_{x \geq O_i} \frac{\mu_i^x \cdot \exp(-\mu_i)}{x!}$$

または、

$$p_i = \sum_{x < O_i} \frac{\mu_i^x \cdot \exp(-\mu_i)}{x!}$$

ここで、ポアソン分布

$$p(x) = \frac{\mu^x \cdot \exp(-\mu)}{x!}$$

は、平均μに応じて図 6.5 のような分布となる。

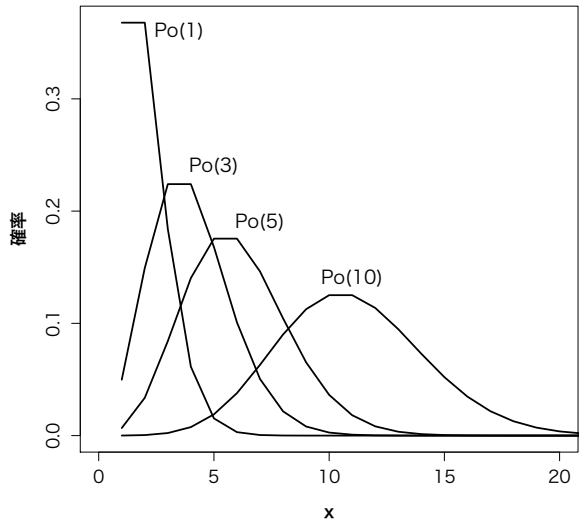


図 6.5 ポアソン分布

都道府県別心疾患死亡者数のポアソン確率地図を、図 6.6 に示す。このとき、ポアソン確率地図の結果は、確率値が 0 または 1 の近くに偏っていることを示している（図 6.7）。これは、死亡者数の期待値 E_i が非常に小さいため、過大分散（overdispersion）が生じていることによる。過大分散とは、本来想定される分散よりも、分散が過大にばらついていることを意味する。ポアソン分布は分散が期待値 E_i によって定義されるため、分散が予想された値以上にばらつく場合がある。また、外れ値などが存在する場合にも、過大分散が生じることがある。このことはポアソン確率地図を用いる場合に注意すべき点である。

過大分散に対処するためには、対象地域全体の人口数の地域差などを考慮した方法などが提案されている。

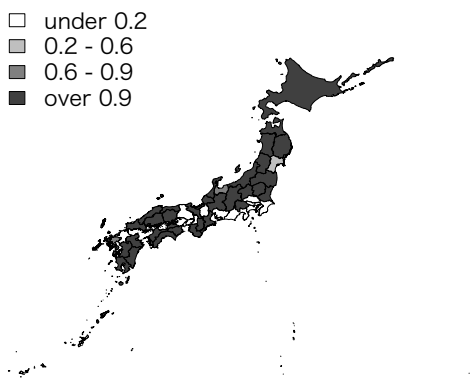


図 6.6 ポアソン確率地図

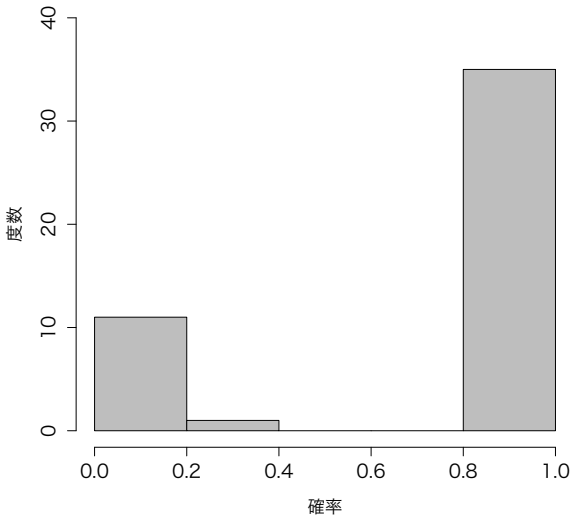


図 6.7 ポアソン確率地図のヒストグラム

R 分析例

spdep パッケージの pmap()関数を用いると、粗率や期待値、相対危険度、ポアソン確率を求めることができる。以下では、都道府県境界ポリゴンデータ（jpn_pref.shp）と都道府県別心疾患死亡者数のデータ（hd06.csv）を用いて、これらの指標を計算してみよう。

```
# spdep パッケージを使用
```

```
library(spdep)
# 都道府県境界ポリゴンデータの読み込み
jpn_pref <- readShapePoly("jpn_pref.shp",IDvar="PREF_CODE")
# 都道府県別心疾患死亡者数データの読み込み
hd06 <- read.table("hd06.csv", sep=",", header=T)
# 都道府県境界ポリゴンデータへの属性データ hd06 のマッチング
ID.match <- match(jpn_pref$PREF_CODE, hd06$PREF_CODE)
jpn_hd06 <- hd06[ID.match,]
jpn_pref_hd06 <- spCbind(jpn_pref, jpn_hd06)
# 確率地図の作成
jpn_hd06_pm <- probmap(jpn_pref_hd06$HD06,
jpn_pref_hd06$POPJ06/100) summary(jpn_hd06_pm)
```

6.4 相対リスクのベイズ推定

6.4.1 Marshall の経験ベイズ推定量

観測数が少ない場合などに、人口数の地域差を調整する方法の一つに、相対リスクの推定量 $\hat{\theta}_i$ をベイズ推定する方法がある[5][6]。

相対リスクをベイズ推定する際には、相対リスクの分布を何らかの確率分布として仮定する。その事前分布を規定する超パラメータを最尤推定法などにより推定する法と、マルコフ連鎖モンテカルロ法により事後分布を推定する方法とがある。超パラメータの分布を任意に（経験的に）与えベイズ推定する方法を経験ベイズ推定法といい、超パラメータ自体の分布を仮定して階層的に事後分布を推定する方法を、階層ベイズ推定法という。

分析対象としている観測ケースの標本値 O_i がポアソン分布に従って発生すると仮定できるようなケースであり、その期待値が E_i とする。相対リスク θ_i の最尤推定値を x_i とすると、θ_i が与えられた条件下での、x_i の平均 $E(x_i|\theta_i)$ と分散 $V(x_i|\theta_i)$ はそれぞれ以下のように表される。

$$E(x_i|\theta_i)=\theta_i$$

$$V(x_i|\theta_i)=\theta_i/E_i$$

疫学分野では、x_i を標準化死亡比（SMR）と呼ぶ。

θ_i の平均と不偏分散について事前情報を与えたとき、グローバルな経験ベイズ

推定量$\hat{\theta}_i$はその事後情報として以下の収束推定量として求めることができる。

$$\hat{\theta}_i = \hat{\mu} + \hat{C}_i\left(x_i - \hat{\mu}\right)$$

ただし、

$$\hat{\mu} = \sum_{i=1}^{N} O_i \Big/ \sum_{i=1}^{N} E_i$$

$$\hat{C}_i = \frac{s^2 - \hat{\mu}/\overline{E}}{s^2 - \hat{\mu}/\overline{E} + \hat{\mu}/E_i}$$

であり、s^2はx_iについての重み付け不偏分散を意味する。

$$s^2 = \sum_{i=1}^{N} E_i\left(x_i - \hat{\mu}\right) \Big/ \sum_{i=1}^{N} E_i$$

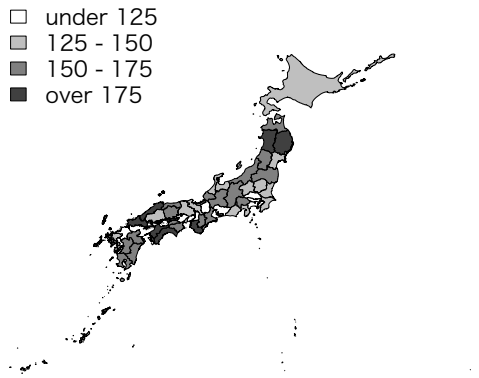
また、$\overline{E}$は期待値E_iの平均値である。

$$\overline{E} = \sum_{i=1}^{N} E_i \Big/ N$$

隣接行列を用いて、近隣地区との隣接性を考慮した経験ベイズ推定量を、ローカルは経験ベイズ推定量という。地区ij間の隣接行列の要素c_{ij}が与えられたとき、$\hat{\mu}$、$\hat{C}_i$、s^2、$\overline{E}$を、それぞれ隣接要素を考慮して地区i毎に求める。例えば、

$$\hat{\mu}_i = \sum_{j=1}^{n} c_{ij} O_j \Big/ \sum_{j=1}^{n} c_{ij} E_j$$

などとなる。



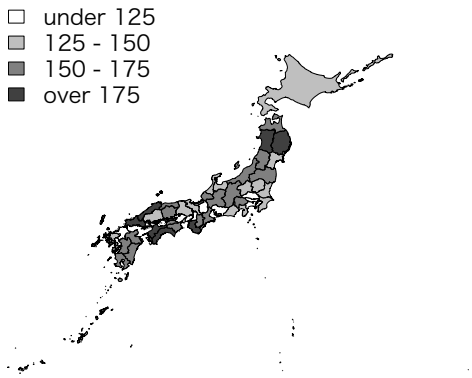
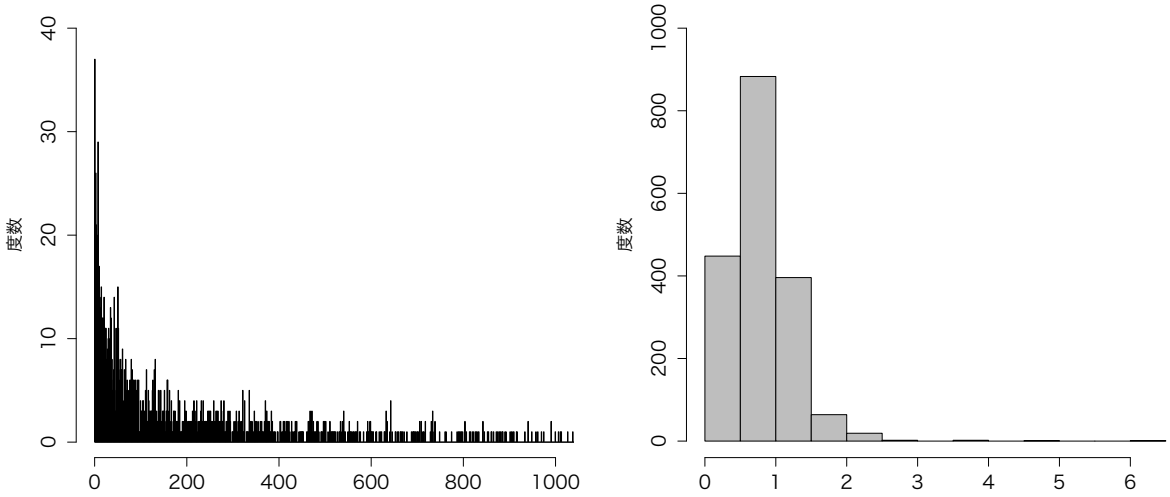


図 6.8 グローバルな経験ベイズ推定量　　図 6.9 ローカルな経験ベイズ推定量

都道府県別心疾患死亡者数について、グローバルな経験ベイズ推定量とローカルな経験ベイズ推定量を計算した結果は、図 6.8 及び図 6.9 のようになる。

6.4.2 ポアソン・ガンマモデル

総務省統計局『社会生活統計指標-市区町村指標-』[7]において市区町村単位で集計された市区町村別の交通事故死亡者データ（2007 年）を用いて、観測値と相対リスクの分布をヒストグラムで示すと図 6.10 のようになる。



(a) 観測値　　(b) 相対リスク

図 6.10 市区町村別交通事故死者数の分布

この場合、期待値 E_i と相対リスク θ_i が与えられた条件つきでの観測値 O_i がポアソン分布に従い、相対リスクがガンマ分布に従うとも考えられる。このようなデータの相対リスクを推定する方法として、ポアソン・ガンマモデルが提案されている[8]。

$$O_i \sim Po(\theta_i E_i)$$

$$\theta_i \sim \Gamma(\nu, \alpha)$$

このモデルでは、相対リスク θ_i が観測値 O_i とは独立にガンマ分布に従って生成させる。ガンマ分布は平均 ν/α、分散 ν/α^2 となるような分布であり、次式の確率密度関数 $f(x)$ で表される。

$$f(x)=\frac{1}{\Gamma(\nu)}\alpha^{\nu}x^{\nu-1}\cdot\exp(-\alpha x)$$

ただし、νは形状パラメータ、αはスケールパラメータを意味する。Γ分布の確率密度関数の例を、図 6.11 に示す。

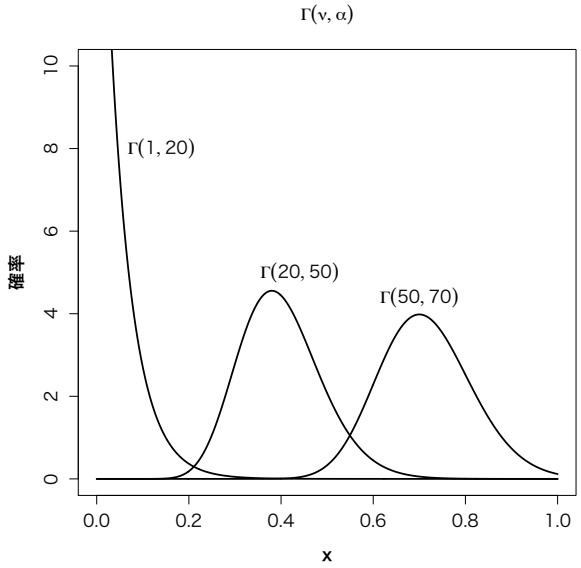


図 6.11 ガンマ分布の確率密度関数の例

このとき、観測値O_i、期待値E_i及びガンマ関数のパラメータνとαを用いて、平滑化相対リスク$(O_i+\nu)/(E_i+\alpha)$を経験ベイズ推定することができる。平滑化相対リスクを経験ベイズ推定するには、パラメータνとαの事前情報を適当に与え、次の２式を繰り返し計算することにより、各パラメータの事後情報$\hat{\nu}$と$\hat{\alpha}$を計算し、相対リスクの事後分布を求める。

$$\frac{\hat{\nu}}{\hat{\alpha}}=\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$$

$$\frac{\hat{\nu}}{\hat{\alpha}^2}=\frac{1}{N+1}\sum_{i=1}^{N}\left(1+\frac{\hat{\alpha}}{E_i}\right)\left(\hat{\theta}_i-\frac{\hat{\nu}}{\hat{\alpha}}\right)$$

相対リスクθ_iの最尤推定値x_iの事前分布がガンマ分布に従うとすると、その経験ベイズ推定量は、次式の収束推定量として求めることもできる。

$$\hat{\theta}_i=\frac{E_i}{E_i+\alpha}x_i+\frac{\alpha}{E_i+\alpha}\frac{\nu}{\alpha}$$

R 分析例

交通事故データ（data84.csv）を用いて、相対リスク及びポアソン・ガンマモデルによる相対リスクの経験ベイズ推定量を計算してみよう。spdep パッケージの EBest()関数を用いて Marshall の経験ベイズ推定量を、DCluster パッケージの empbaysmooth()関数を用いてポアソン･ガンマモデルによる経験ベイズ推定量を、それぞれ求める。

```
# spdep パッケージを使用
library(spdep)
# DCluster パッケージを使用
library(DCluster)
# データ読み込み
data84 <- read.table("data84.csv", sep=",", header=TRUE)
# 交通事故発生件数（data84$TrAcc）と人口数（data84$Pop）
# から粗率を求める
r <- sum(data84$TrAcc)/sum(data84$Pop)
# 期待値
data84$TAExpected <- data84$Pop * r
# 相対リスク
data84$TARR <- data84$TrAcc / data84$TAExpected
# Marshall の経験ベイズ推定量
data84$EB <- EBest(data84$TrAcc, data84$TAExpected)
# ポアソン・ガンマモデルによる経験ベイズ推定量
eb <- empbaysmooth(data84$TrAcc,data84$TAExpected)
data84$EBPG <- eb$smthrr
# ヒストグラムの図示
hist(data84$TARR, col="grey", ylim=c(0,1000), main="",
ylab="度数", xlab="", cex.axis=1.3, cex.lab=1.2)
```

この結果から、ポアソン・ガンマモデルによる経験ベイズ推定量のヒストグラムは図 6.12 のようになる。

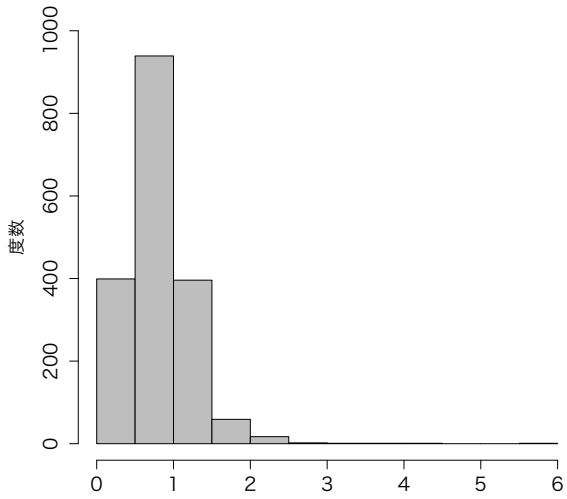


図 6.12 ポアソン・ガンマモデルによる経験ベイズ推定量のヒストグラム

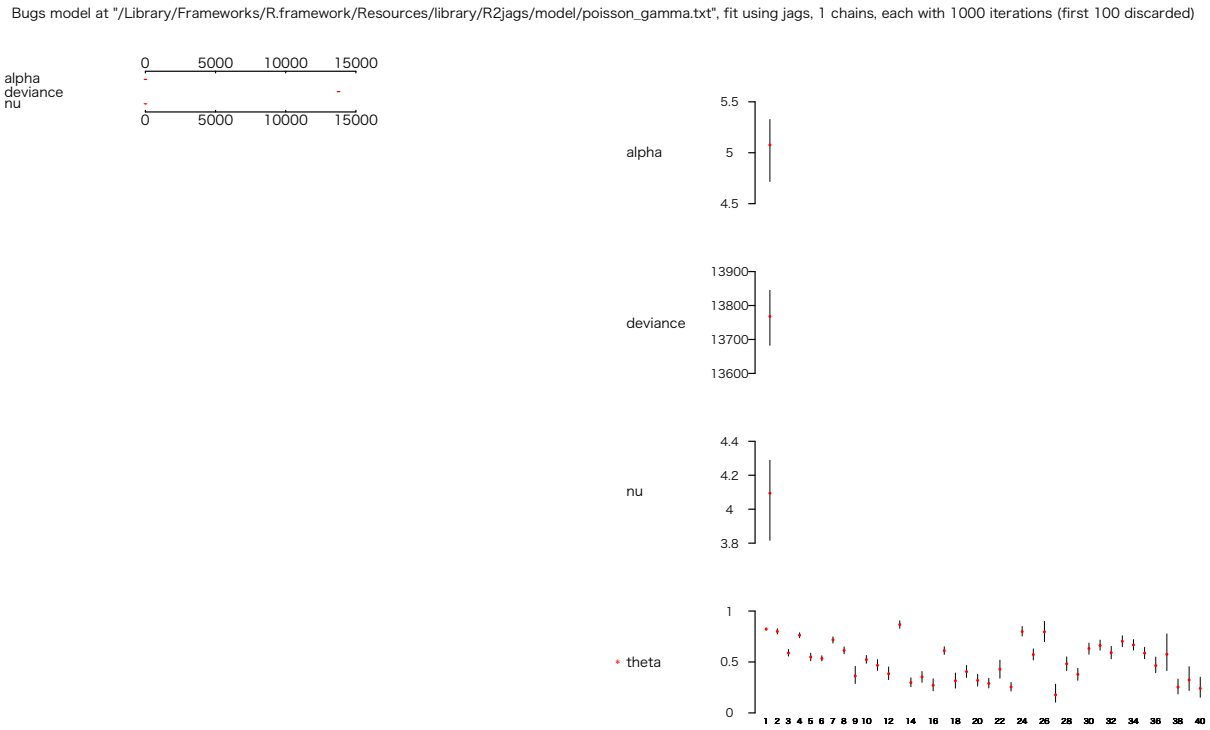


図 6.13 ポアソン・ガンマモデルの階層ベイズ推定例

また、パラメータ ν と α に、ガンマ分布や指数分布などを仮定することにより、マルコフ連鎖モンテカルロ法（MCMC）を用いて、$\hat{\theta}_i$ の事後分布を階層ベイズ推定できる。

前述の交通事故死亡者データを用いて、ポアソン・ガンマモデルによる平滑化相対リスクを階層ベイズ推定した結果を図 6.13 に示す。

ここで、階層ベイズ推定するための ν と α の階層事前情報は、それぞれ、

$$\nu \sim \Gamma(0.01, 0.01)$$

$$\alpha \sim \Gamma(0.01, 0.01)$$

とし、MCMC のシミュレーション期間を 1,000 回、稼働検査期間を 100 回、チェーン数を 1 としている[1]。

6.4.3 対数正規モデル

相対リスク θ_i に多変量対数正規分布を考慮し、相対リスクの対数 $\log\left((O_i+1/2)/E_i\right)$ を EM アルゴリズムや MCMC 法によりベイズ推定する方法も提案されている。経験ベイズ法による対数正規相対リスクの事後分布 b_i は次式のように表される。

$$b_i = \frac{\hat{\phi} + (O_i + 1/2)\hat{\sigma}^2 \log\left[O_i + 1/2/E_i\right] - \hat{\sigma}^2/2}{1 + (O_i + 1/2)\hat{\sigma}^2}$$

$$\hat{\phi} = \frac{1}{N}\sum_{i=1}^{N} b_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\left\{\hat{\sigma}^2 \sum_{i=1}^{N}\left[1 + \hat{\sigma}^2(O_i + 1/2)\right]^{-1} + \sum_{i=1}^{N}\left(b_i - \hat{\phi}\right)^2\right\}$$

EM アルゴリズムによる対数正規モデルの経験ベイズ推定結果を図 6.14 に示す。

R 分析例

6.4.2 で用いたデータを使って、対数正規モデルによる経験ベイズ推定を行う。ここでは、すでに `DCluster` パッケージを呼び出した上で、期待値が計算されているものとする。

```
# 対数正規モデルによる経験ベイズ推定
data84_ln <- lognormalEB(data84$TrAcc, data84$TAExpected)
# ヒストグラムの図示
```

[1] ポアソン・ガンマモデルの階層ベイズ推定には、Bivand (2008) p.323[9]の BUGS コードをそのまま JAGS コードとして用い、`R2jags` パッケージにより JAGS コードを呼び出して推定した。

```
hist(data84_ln$smthrr, main="", xlab="")
```

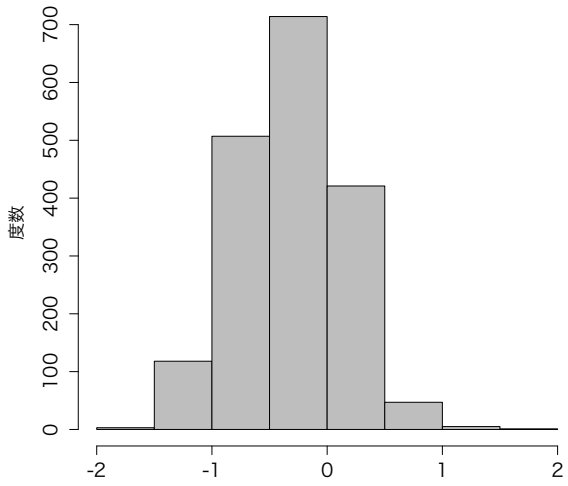


図 6.14 対数正規モデルによる経験ベイズ推定

6.5 経験ベイズ推定値の Moran's I

6.4 で示した Marshall の経験ベイズ推定量について、Moran's I 値の経験ベイズ推定量 *EBI* を次式から求めることができる。

$$EBI = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} z_i z_j}{\sum_{i=1}^{N}(z_i - \bar{z})}$$

ここで、w_{ij}は地区 ij間の重み付け行列の要素であり、5.2 節で示した方法により得られる。また、他の変数は以下のように得られる。

$$z_i = (p_i - b)/\sqrt{v_i},$$

$$p_i = O_i / y_i,$$

$$v_i = a + (b/y_i),$$

$$a = s^2 - b\Big/\left(\sum_{i=1}^{N} y_i / N\right),$$

$$b = \sum_{i=1}^{N} O_i \Big/ \sum_{i=1}^{N} y_i,$$

$$s^2 = \sum_{i=1}^{N} y_i (p_i - b)^2 \Big/ \sum_{i=1}^{N} y_i$$

*EBI*が正の値となるとき、相対危険度は空間的に自己相関するという。観測値O_iが正規分布に従わないことから、繰り返し検定などを行い、空間的自己相関の有無についての仮説検定を行う[10]。

参考文献

[1] 中谷友樹・谷村晋・二瓶直子・堀越洋一編著（2004）『保健医療のための GIS』、古今書院.

[2] Bivand, R.（2009）Introduction to the North Carolina SIDS data set (revised), http://cran.r-project.org/web/packages/spdep/vignettes/sids.pdf (2010/10/14 閲覧)

[3] 総務省統計局（2009）『社会生活統計指標－都道府県の指標－』, http://www.stat.go.jp/

[4] 厚生労働省老健局老人保健課（2009）『都道府県別死因の分析結果について』, http://www.mhlw.go.jp/topics/kaigo/shiin/gaiyo/ （2009 年 9 月現在）

[5] Marshall, R.J. (1991) Mapping disease and mortality rates using empirical bayes estimators, *Applied Statistics*, Vol. 40, No. 2, pp. 283-294.

[6] Martuzzi, M. and Elliott, P (1996) Empirical bayes estimation of small area prevalence of non-rare conditions, *Statistics in Medicine*, Vol. 15, pp. 1867-1873.

[7] 総務省統計局（2009）『社会生活統計指標－市区町村の指標－』, http://www.stat.go.jp/

[8] Clayton D.G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics*, Vol. 43, pp. 671-681.

[9] Bivand, R.S., E.J. Pebesma and V. Gomez-Rubio (2008) *Applied Spatial Data Analysis with R (Use R)*, Springer.

[10] Assuncao, R.M., and Reis, E.A. (1999) A new proposal to adjust Moran's I for population density, *Statistics in Medicine*, Vol. 18, pp.2147-2162.