空間の統計学(3)：確率地図と経験ベイズ推定

慶應義塾大学総合政策学部准教授
古谷　知之（Furutani Tomoyuki）

■兵庫県生まれ。2001年東京大学大学院工学系研究科博士課程修了。博士（工学）。東京大学大学院助手、慶應義塾大学環境情報学部専任講師を経て、07年4月より現職。専門分野：空間統計学、都市交通計画、観光政策。

1. はじめに

今回は、前回に引き続き、空間的自己相関に関する手法をとりあげます。第1回（2009年8月号）では主題図の作成について、第2回（同年9月号）では空間的自己相関について扱いましたが、各回とも死因別死亡者数のデータを加工することなく用いました。一般的に、各地区の死亡者数は、当該地区の人口数に比例すると考えられます。従って、各地区の死亡者数の多少や空間的自己相関の有無を検討する上では、単に死亡者数を比較するのではなく、死亡者数を人口数で割った比率（粗率）を比較する方が望ましいといえるでしょう。

そこで今回は、地区属性を比較する際に用いられる手法として、確率地図や経験ベイズ法を紹介します。これらの方法は、地区属性として比率（確率）データを比較する場合に、比率値が比較的小さい値をとる際に、特に有効であるといわれています。例えば、死亡数が人口数と比較して非常に小さい値をとる場合、死亡数／人口数の比率を地区間で比較することは非常に困難だからです。そのため、空間疫学や災害発生などのリスク分析に、しばしば用いられます。また、地区（地点）の属性値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す。

2. 演習用データの概要

演習には、総務省統計局の「社会生活統計指標－都道府県の指標－」[1]及び「都道府県別死因の分析結果について」[2]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ダウンロードして用います。

なお、筆者のHP（http://web.sfc.keio.ac.jp/~maunz/wiki/）で配布しているデータ「75data.csv」には、人口総数「POPT06」（単位：千人）、日本人人口数「POPJ06」（単位：千人）、心疾患死亡者数「HD06」（単位：人）、心疾患による標準化死亡比（SMR）の男女別データ（男性：HD06.M.SMR、女性：HD06.F.SMR）がまとめられています。

第1回の復習を兼ねて、75data.csvを都道府県境界ポリゴンデータや都道府県庁ポイントデータとマッチングしてみましょう。

```
jpn_pref <-
readShapePoly("jpn_pref.shp",IDvar="PREF_CODE")
pref_pnt <-
readShapePoints("pref_gov.shp")
hd06 <- read.table("75dat.csv",
sep=",", header=T)
ID.match <- match(jpn_pref$PREF_CODE,
hd06$PREF_CODE)
jpn_hd06 <- hd06[ID.match,]
jpn_pref_hd06 <- spCbind(jpn_pref, jpn_hd06)
summary(jpn_pref_hd06)
```

3. 確率地図

「死亡率」や「リスク発生率」のような確率値を地区（地点）間で比較したい場合は、粗率、相対危険度、ポアソン確率などが適用されます[3]。相対危険度やポアソン確率は、観測ケースの値が人口総数に比較して相対的に小さい値をとるときに用いられます。

いま、地区 i（$i=1, \cdots, N$）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）の値を O_i とします。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ます。

$$r_i = \frac{O_i}{y_i} \tag{1}$$

粗率 r_i は、対象地域全体の人口数や観測値を反映していないため、人口数に対する観測ケースの値の大小を判断することが容易ではありません。そこで、対象地域全体で観測ケース発生率が均一であると仮定して、観測値を人口数の期待値 E_i で除した相対危険度 RR_i を用いることがあります。

$$E_i = y_i \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} y_i} \tag{2}$$

$$RR_i = \frac{O_i}{E_i} \tag{3}$$

観測値の発生頻度が低い場合は、ポアソン確率を用いると、人口規模による観測ケースの過大／過小評価の影響を考慮することができます。

Rでは、spdepライブラリの**probmap( )**関数を使って、これらの値を求めることができます。

```
jpn_hd06_pm <-
probmap(jpn_pref_hd06$HD06,
jpn_pref_hd06$POPJ06/100)
summary(jpn_hd06_pm)
```

心疾患死亡者数の観測値（図1）と期待値（図2）、粗率（図3）と相対危険度（図4）及びポアソン確率（図5）を、それぞれ図示してみましょう。

```
library(classInt)
brks1 <- c(0,1500, 2500, 3500, 4500, 20000)
brks2 <- c(0,100,125,150,175,200)
brks3 <- c(0, 0.2, 0.6, 0.9, 1.0)
cols <- grey(6:2/6)
# 観測値のプロット
plot(jpn_pref,
col=cols[findInterval(jpn_pref_hd06$HD06,
brks1, all.inside=TRUE)])
legend("topleft", fill=cols,
legend=leglabs(brks1), bty="n")
```

図１　心疾患死亡者数（観測値、2006年）

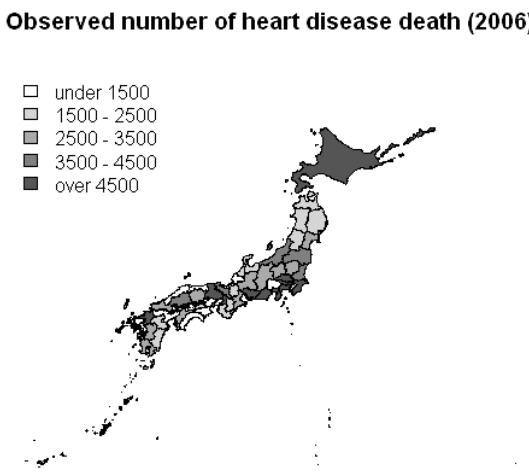
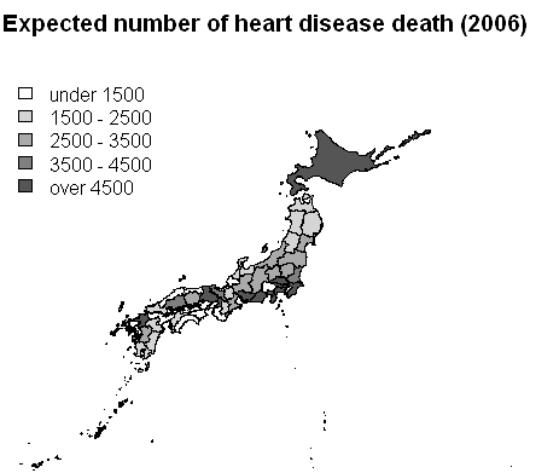
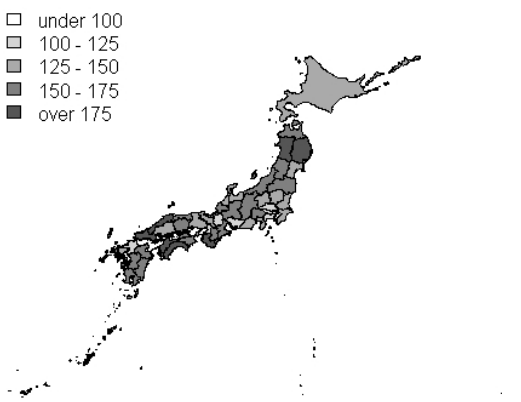


図２　心疾患死亡者数（期待値、2006年）




```
title(main="Observed number of heart
disease death (2006)")
# 期待値のプロット
plot(jpn_pref,
col=cols[findInterval(jpn_hd06_pm
$expCount, brks1, all.inside=TRUE)])
legend("topleft", fill=cols,
legend=leglabs(brks1), bty="n")
title(main="Expected number of heart
disease death (2006)")
# 粗率のプロット
plot(jpn_pref,
col=cols[findInterval(jpn_hd06_pm$raw,
brks2, all.inside=TRUE)])
legend("topleft", fill=cols,
legend=leglabs(brks2), bty="n")
title(main="Raw rate of heart disease
per 100,000 (2006)")
# 相対危険度のプロット
plot(jpn_pref,
col=cols[findInterval(jpn_hd06_pm$relRisk,
brks2, all.inside=TRUE)])
legend("topleft", fill=cols,
legend=leglabs(brks2), bty="n")
title(main="Relevant risk of heart
disease per 100,000 (2006)")
```

図３　心疾患死亡率（粗率、2006年）




```
# ポアソン確率のプロット
plot(jpn_pref,
col=cols[findInterval(jpn_hd06_pm$pmap,
brks3, all.inside=TRUE)])
legend("topleft", fill=cols,
legend=leglabs(brks3), bty="n")
title(main="Poisson probability of
heart disease (2006)")
hist(jpn_hd06_pm$pmap)
```

図4　心疾患死亡率（相対危険度、2006年）

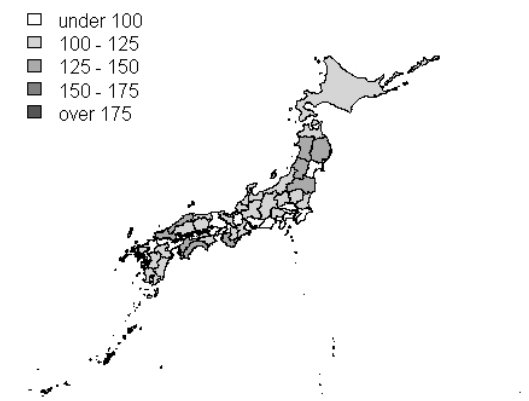


図5　心疾患死亡率（ポアソン確率、2006年）

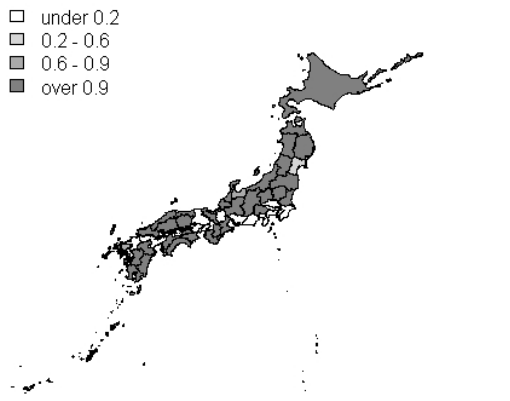
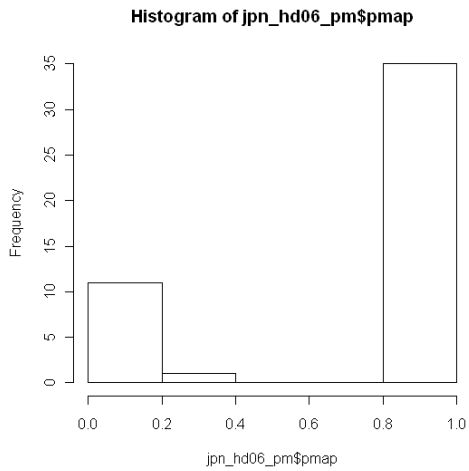


図6　ポアソン確率のヒストグラム



ポアソン確率の結果は、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ます（図6）。これは、死亡者数の期待値E_iが非常に小さいために、過分散（overdispersion）が生じていることによります。過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調整した方法を適用することが望ましいと考えられます。

4. 経験ベイズ法

人口数の地域差を調整する方法の一つに、経験ベイズ法により推定量を求める方法があります[4]。経験ベイズ推定量EBIは、次式から得られます。

$$EBI = \frac{N}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}} \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} z_i z_j}{\sum_{i=1}^{N} (z_i - \bar{z})^2} \quad (4)$$

ここで、w_{ij}は地区（地点）ij間の重み付け行列の要素であり、各変数は以下のようにして得られます。

$$z_i = \frac{p_i - b}{\sqrt{v_i}}, \qquad b = \sum_{i=1}^{N} O_i \Big/ \sum_{i=1}^{N} y_i,$$

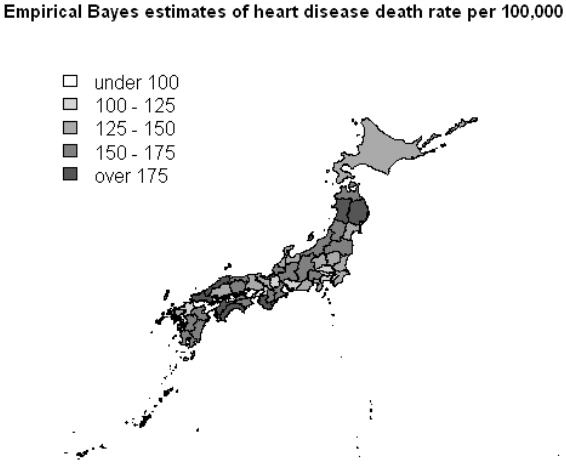
$$p_i = O_i / y_i, \qquad a = s^2 - b \Big/ \left(\sum_{i=1}^{N} y_i / N \right),$$

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

Rでは、**EBest**( )関数を使って、経験ベイズ法による推定量を計算することができます。その結果を図7に示します。

```
jpn_hd06_ebg <-
EBest(jpn_pref_hd06$HD06,
jpn_pref_hd06$POPJ06/100)
plot(jpn_pref,
col=cols[findInterval(jpn_hd06_ebg
$estmm, brks2, all.inside=TRUE)])
legend("topleft", fill=cols,
legend=leglabs(brks2), bty="n")
title(main="Empirical Bayes estimates
of heart disease death rate per
100,000 (2006)")
```

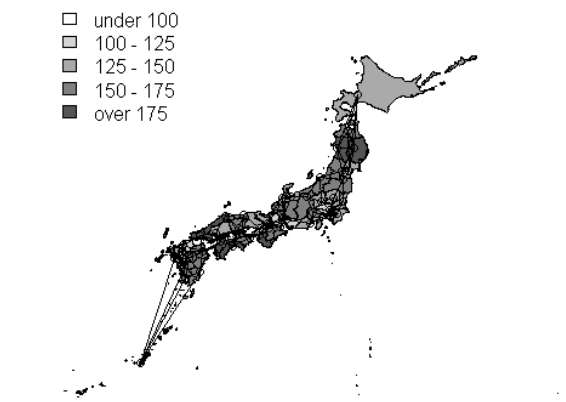
図7　心疾患死亡率（経験ベイズ推定、2006年）



また、**EBlocal**( )関数を使って、ローカルな経験ベイズ推定量を計算することができます。ローカルな経験ベイズ推定量を計算するためには、隣接行列を定義する必要があります。ここでは、**knearneigh**( )関数を使い、近隣4都道府県を隣接地区とする隣接行列を用いて、ローカルな経験ベイズ推定を行います（図8）。

```
pref_gov <-
read.table("pref_gov.txt",",",
header=T, row.names=2)
coords <- matrix(0,nrow(pref_gov),2)
coords[,1] <- pref_gov$X
coords[,2] <- pref_gov$Y
pref.knn <- knearneigh(coords, k=4)
pref.knn.nb <- knn2nb(pref.knn,
row.names=rownames(pref_gov))
jpn_hd06_ebl <-
EBlocal(jpn_pref_hd06$HD06,
jpn_pref_hd06$POPJ06/100, pref.knn.nb)
plot(jpn_pref,
col=cols[findInterval(jpn_hd06_ebl$est,
brks2, all.inside=TRUE)])
plot(pref.knn.nb, coords, add=TRUE)
legend("topleft", fill=cols,
legend=leglabs(brks2), bty="n")
title(main="Local empirical Bayes
estimates of heart disease death rate
per 100,000 (2006)", cex.main=0.9)
```

図8　心疾患死亡率（ローカルな経験ベイズ推定、2006年）



EBImoran( )関数を用いると、Moran's Iの経験ベイズ推定量を計算し、空間的自己相関の有無について繰返し検定を行うことができます[5]。

```
# 経験ベイズ推定量のMoran's I
EBImoran.mc(jpn_pref_hd06$HD06,
jpn_pref_hd06$POPJ06/100,
nb2listw(pref.knn.nb, style="B",
zero.policy=TRUE),
nsim=999, zero.policy=TRUE)
# 通常のMoran's I
moran.mc(jpn_pref_hd06$HD06/
(jpn_pref_hd06$POPJ06/100),
nb2listw(pref.knn.nb, style="B",
zero.policy=TRUE),
nsim=999, zero.policy=TRUE)
```

Moran's Iの経験ベイズ推定量の計算結果とその繰返し検定の結果は図9、通常のMoran's Iの計算結果とその繰返し検定の結果は図10のようになります。この結果からは、経験ベイズ推定量のMoran's I＝0.5923（p＝0.005）、通常のMoran's I＝0.2596（p＝0.001）となっています。

＊参考文献・URL
[1] 総務省統計局：社会生活統計指標－都道府県の指標－2009及び2008（http://www.stat.go.jp/data/ssds/5.htm）.
[2] 厚生労働省老健局老人保健課：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/）.
[3] Bivand, R.(2009)： Introduction to the North Carolina SIDS data set(revised)(http://cran.r-project.org/web/packages/spdep/vignettes/sids.pdf).
[4] Martuzzi, M. and Elliott, P.(1996)：Empirical bays estimation of small area prevalence of non-rare conditions: Statistics in Medicine, Vol. 15, pp.1867-1873.
[5] 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.

図9　Moran's Iの経験ベイズ推定量の計算結果と繰返し検定の結果

```
Monte-Carlo simulation of Empirical Bayes Index
data:cases:jpn_pref_hd06$HD06,risk population:jpn_pref_hd06$POPJ06/100
weights:nb2listw(pref.knn.nb,style="B",zero.policy=TRUE)
number of simulations+1:1000
statistic=0.5923,observed rank=995,p-value=0.005
alternative hypothesis:greater
```

図10　Moran's Iの計算結果と繰返し検定の結果

```
Monte-Carlo simulation of Moran's I
data:jpn_pref_hd06$HD06/(jpn_pref_hd06$POPJ06/100)
weights:nb2listw(pref.knn.nb,style="B",zero.policy=TRUE)
number of simulations+1:1000
statistic=0.2596,observed rank=999,p-value=0.001
alternative hypothesis:greater
```