医学情報処理演習第 11 回課題回答例

以下のようにして，まずデータを読み込み（1 行目），attach してから（2 行目）地域別出生児数分布表を作り（3 行目），地域別の出生児数別カップル数を別々の変数に保管する（4 行目-6 行め）ことで，分析の準備が整う。

it11-ans-2006.R(1)

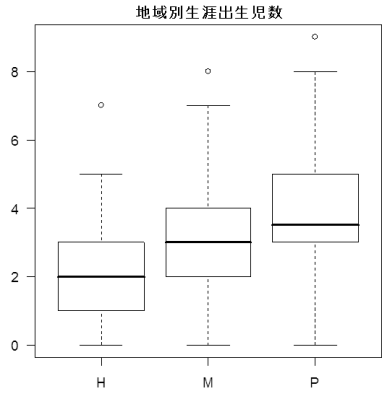
```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p11.txt")
attach(dat)
X <- table(GRP,PARITY)
HF <- X[1,]
MF <- X[2,]
PF <- X[3,]
```

次いで，グラフを描く。個別に `barplot(HF)` などをした方がよいが，より簡便には，`boxplot(PARITY~GRP)` により，1 つのグラフィック枠に 3 地域を層別して箱ヒゲ図が描かれる。外れ値があることからノンパラメトリックな比較を考え，次に Fligner-Killeen の検定を行う。$\chi^2 = 0.75, p = 0.69$ より，3 地域間でばらつきが均質であるという帰無仮説が棄却されないので，次に Kruskal-Wallis の検定を行う，$\chi^2_{\mathrm{KW}} = 11.2, p = 0.0036$ より，出生児数の分布の位置母数に 3 地域で差がないという帰無仮説は棄却される。つまり少なくともどこかの 2 地域間で差があることがわかる。

最後に，帰無仮説族 {{H と M に差がない}，{M と P に差がない}，{P と H に差がない}} を検定するため，Holm の方法で検定の多重性を調整したウィルコクソンの順位和検定を `pairwise.wilcox.test()` により行うと，調整済み有意確率が，H 市と M 村の比較で 0.0251，H 市と P 村で 0.0054，M 村と P 村で 0.4030 となるので，H 市のカップルの生涯子供数は，M 村とも P 村とも有意水準 5%で統計的有意差があるが，M 村と P 村では生涯子供数に統計的有意差はないといえた。

it11-ans-2006.R(2)

```
win.metafile("it11-ans-2006-1.emf",width=6,height=6,pointsize=14)
par(family="sans",mai=c(0.4,0.4,0.4,0.4),las=1)
boxplot(PARITY~GRP,main="地域別生涯出生児数")
dev.off()
fligner.test(PARITY~GRP)
kruskal.test(PARITY~GRP)
pairwise.wilcox.test(PARITY,GRP,exact=F)
```



なお，分布をみるには，

```
layout(1:3)
tapply(PARITY,GRP,hist,xlim=c(0,10),breaks=0:10,main="",right=F)
```

として 3 地域別々のヒストグラムを描かせるか（`right=F` として区間の右端を入れないことが重要。また，`tapply()` の 4 番目以降の引数は，3 番目の引数である関数にそのまま渡される），あるいは子供数は離散値なので，以下のように棒グラフにしてもよい。

```
layout(1:3)
barplot2 <- function(...) { barplot(table(...)) }
tapply(PARITY,GRP,barplot2)
```

最初のところをもう少し丁寧に分析するには，以下のように，地域別に棒グラフを描き，そこに既知の分布を当てはめてみるとよい。子供数の分布については，自然出生集団では Poisson 分布が，意図的な出産抑制をしている集団では負の 2 項分布（1 回につき確率 p で成功する一連のベルヌーイ試行について，成功が x 回起こるまでの失敗数の分布）が当てはまると言われているので，両方を試してみる。下枠内のように手計算もできる（H 市についてポアソン分布を当てはめた例）が，第 10 回に紹介したように `vcd` ライブラリの `goodfit()` 関数を使うと，より簡便である。

```
H <- rep(0:9,HF)
ix <- barplot(HF,main="H 市の出生児数分布と分布の当てはめ",ylim=c(0,10))
lines(ix,EH<-dpois(0:9,mean(H))*30)
print(mean(H))
print(XH <- sum((HF-EH)^2/EH))
1-pchisq(XH,8)
```

具体的には下枠内のコードで実行できる。検定結果も凡例の形でグラフに書き込んでみた。どの地域においてもポアソン分布が適合しているという帰無仮説も，負の２項分布が適合しているという帰無仮説も棄却できない結果となった。しかしパラメータは互いに違いがありそうに見えるので，この後で，最初に示したように Fligner-Killeen，Kruskal-Wallis，多重性の調整付き Wilcoxon の順位和検定，と進むのがよい。

it11-ans-2006.R(3)

```
H <- rep(0:9,HF); M <- rep(0:9,MF); P <- rep(0:9,PF)
library(vcd)
win.metafile("it11-ans-2006-2.emf",width=8,height=12,pointsize=14)
par(family="sans",mai=c(0.4,0.4,0.4,0.4),las=1,mfrow=c(3,1))
XHP <- goodfit(H,"poisson"); SXHP <- summary(XHP); TXHP <- paste("ポアソン分布 (p=",sprintf("%4.2f",SXHP[3]),")")
XHN <- goodfit(H,"nbinom"); SXHN <- summary(XHN); TXHN <- paste("負の２項分布 (p=",sprintf("%4.2f",SXHN[3]),")")
ix <- barplot(HF,main="H 市の出生児数分布と分布の当てはめ",ylim=c(0,10))
lines(ix,predict(XHP,newcount=0:9),lty=1,col="red")
lines(ix,predict(XHN,newcount=0:9),lty=2,col="blue")
legend(8,max(HF),lty=c(1,2),legend=c(TXHP,TXHN),col=c("red","blue"))
XMP <- goodfit(M,"poisson"); SXMP <- summary(XMP); TXMP <- paste("ポアソン分布 (p=",sprintf("%4.2f",SXMP[3]),")")
XMN <- goodfit(M,"nbinom"); SXMN <- summary(XMN); TXMN <- paste("負の２項分布 (p=",sprintf("%4.2f",SXMN[3]),")")
ix <- barplot(MF,main="M 村の出生児数分布と分布の当てはめ",ylim=c(0,10))
lines(ix,predict(XMP,newcount=0:9),lty=1,col="red")
lines(ix,predict(XMN,newcount=0:9),lty=2,col="blue")
legend(8,max(MF),lty=c(1,2),legend=c(TXMP,TXMN),col=c("red","blue"))
XPP <- goodfit(P,"poisson"); SXPP <- summary(XPP); TXPP <- paste("ポアソン分布 (p=",sprintf("%4.2f",SXPP[3]),")")
XPN <- goodfit(P,"nbinom"); SXPN <- summary(XPN); TXPN <- paste("負の２項分布 (p=",sprintf("%4.2f",SXPN[3]),")")
ix <- barplot(PF,main="P 村の出生児数分布と分布の当てはめ",ylim=c(0,10))
lines(ix,predict(XPP,newcount=0:9),lty=1,col="red")
lines(ix,predict(XPN,newcount=0:9),lty=2,col="blue")
legend(8,max(PF),lty=c(1,2),legend=c(TXPP,TXPN),col=c("red","blue"))
dev.off()
```