第10回　計数データと比率の解析(2)

- カテゴリデータの扱いや考え方は，あまり馴染みがないものなので，先週と今週の2回にわたってやります。つまり今回は概ね復習です。
  - 二項検定とやりたいことの分解のしかた
  - 適合度検定とvcdライブラリのgoodfit()関数の使い方
  - 比率の差の検定，コクラン＝アーミテージ検定とprop.test()とpairwise.prop.test()とprop.trend.test()の使い方→2009年の第10回課題解答例参照
  - 課題

やりたいことを分解する

- （例）MASSライブラリに含まれているsurveyというデータフレームの中の，Clapという変数は，両手を叩くときに，右手を上にする("Right")か，左手を上にする("Left")か，どちらでもない("Neither")のどれかである。左右のどちらが上になるかに差が無いかどうか調べよ
  ↓（大きく以下3つのステップに分解する）
- MASSライブラリを読みこむ
- surveyデータフレームのカテゴリ変数Clapの各カテゴリ度数を調べる
- 左の度数と右の度数に差が無いという帰無仮説を検定する

MASSライブラリを読み込む

- library(MASS)またはrequire(MASS)
  ↓
- surveyデータフレームが参照できるようになる

変数Clapのカテゴリごとの度数を求め，オブジェクトresに付値

- res ＜－ table(survey$Clap)
  または
- res ＜－ xtabs(~Clap, data=survey)
  または
- res ＜－ summary(survey$Clap)

- 度数分布を確認するにはprint(res)
- str(res)とすると，summary()の結果の場合，resが整数型ベクトルであることがわかる。xtabs()やtable()の結果はやや複雑なオブジェクト

帰無仮説

- 「左右の度数に差が無い」
  ↓（言い換えると）
- 「Neither」や欠損値(NA)を除外して，Leftの出る母比率が0.5
  ↓
- binom.test(res["Left"],res["Left"]+res["Right"],0.5)の結果をみればよい。

binom.test(res["Left"],res["Left"]+res["Right"],0.5) の結果

```
Exact binomial test

data:  res["Left"] and res["Left"] + res["Right"]
number of successes = 39, number of trials = 186, p-value =
6.002e-16
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.1535532 0.2752969
sample estimates:
probability of success
             0.2096774
```

対立仮説：母比率が0.5でない
（帰無仮説：母比率が0.5）

有意確率が6×10^{-16}なので帰無仮説は棄却される!

その95%信頼区間

$$\frac{\mathrm{res}[\text{"Left"}]}{\mathrm{res}[\text{"Left"}]+\mathrm{res}[\text{"Right"}]}$$

両手を叩くときに左手が上か右手が上かのどちらかになったとして，左手が上になる母比率の点推定量

適合度検定とvcdのgoodfit()

- （例）http://phi.med.gunma-u.ac.jp/medstat/p11.txt は，ある途上国の3つの地域（変数 GRP，離島にある P 村，首都から車で約 1 時間離れた M 村，首都 H 市）の，再生産をおえたカップル 30 組ずつが生涯に産んだ子供の数（変数 PARITY）を含む（変数名が1行目となっている）タブ区切りテキストデータである（架空のものである）。
- 一般に，生涯に産んだ子供数は完結出生力と呼ばれており，その分布は，意図的な出産抑制があるとき負の二項分布に従い，出産抑制がないとき（つまり，それまでに産んだ子供数と，次に子供を産む確率が独立なとき）ポアソン分布に従うことが知られている。P村の完結出生力がポアソン分布に従うかどうか，グラフを描いた上で適切な検定をせよ。

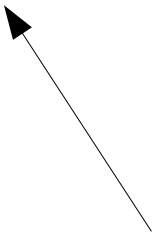
問題の分解

- データを読み込む
- P村の完結出生力ベクトルを得る
- 分布を棒グラフ（度数分布図）として描く
- 平均完結出生力を求め，ポアソン分布に従う場合の完結出生力ごとに期待される人数を得る
  - グラフに期待度数を赤い線で重ね描きする
- カイ二乗適合度検定する
（帰無仮説「ポアソン分布に従う」であることに注意）
- 上の2ステップの代わりにvcdライブラリのgoodfit()関数を使うことも可能

データを読み込む

- タブ区切りテキストデータ（1行目は変数名）なので，read.delim()関数を用いて，読みこんだ結果のデータフレームはdatという名前に付値する
（正しく読めたかどうか，str（）関数で確認する）

```
dat<-read.delim("http://phi.med.gunma-u.ac.jp/medstat/p11.txt")
str(dat)
```



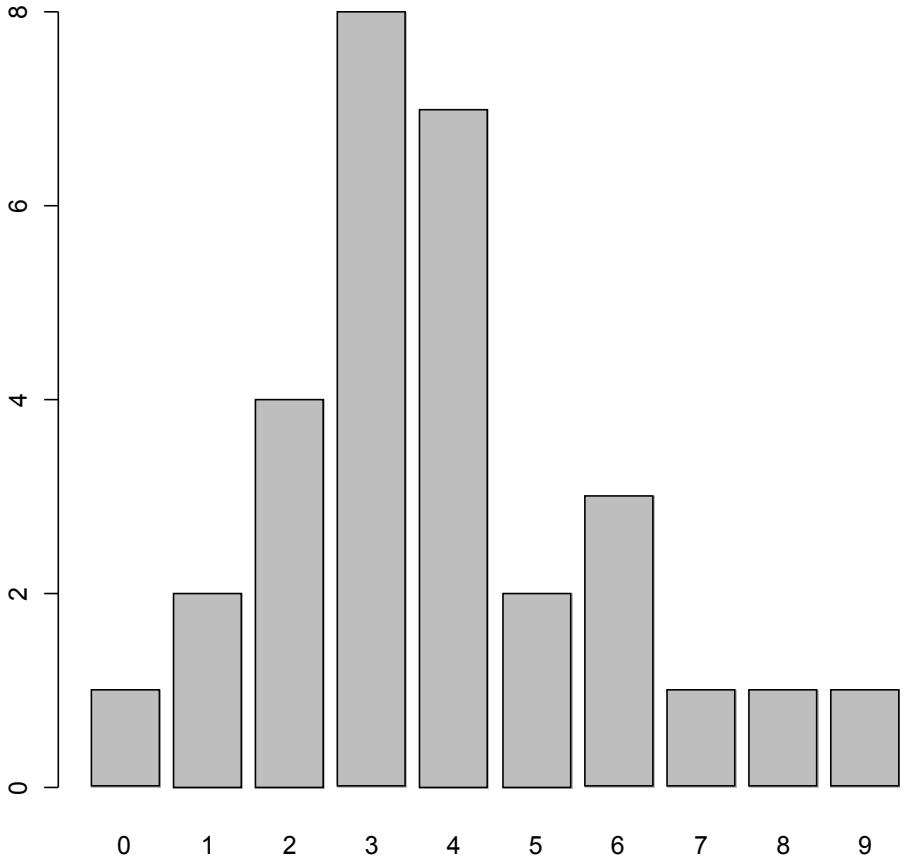
R Consoleでは1行で打つ。二重引用符閉じと括弧閉じを忘れずに！

P村の完結出生力ベクトルを得る

- 完結出生力ベクトルは**dat$PARITY**で得られる。地域を示すのは**dat$GRP=="P"**である。
- オブジェクト**CP**に**P**村の完結出生力ベクトルを付値するには，
  CP <- dat$PARITY[dat$GRP=="P"]
  とするのが最も簡単。
- 別解としては，**subset()**関数で先に**P**村だけのデータにしておくことも可能
  CP <- subset(dat,GRP=="P")$PARITY
  または
  CP <- subset(dat,GRP=="P")[,"PARITY"]

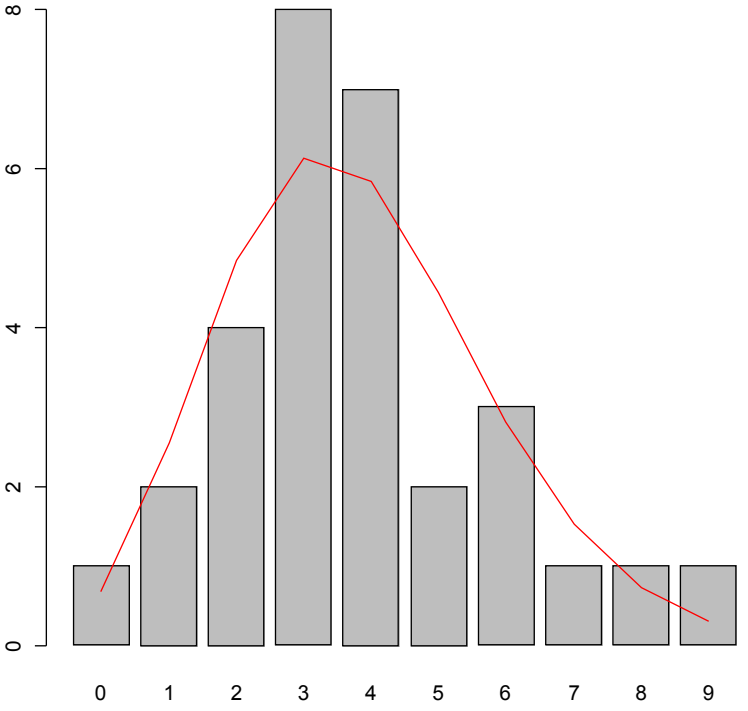
分布を棒グラフとして描く

- **CP**の度数分布を**table()**関数で求め，結果を**TCP**オブジェクトに付値する
- **TCP <- table(CP)**
  ii <- barplot(TCP)
- もちろん，**table(CP)**は**xtabs(~CP)**でも**OK**
- **ii**には，各バーの**x**座標が保存される（後でポアソン分布の線を重ね描きするのに使う）



平均完結出生力を求め，ポアソン分布に従う場合の期待度数分布を計算

- 平均完結出生力は，元データの平均値でいいので
  `MCP <- mean(CP)`
- 完結出生児数は度数分布のカテゴリ名として得られるので，
  `names(TCP)`
  数値扱いさせるために`as.integer()`で括り，
  `as.integer(names(TCP))`
- それぞれに対応する平均MCPのポアソン分布の確率は，
  `dpois(as.integer(names(TCP)),MCP)`



- 総人数`sum(TCP)`をこの割合で分配するため，
  完結出生力の期待度数分布ECPは，
  `ECP <- sum(TCP)*dpois(as.integer(names(TCP)),MCP)`
  ※本当は0人と8人以上のところが1未満なので併合する必要があるが省略
- ECPを赤い線で先の度数分布図に重ね描きさせるには，
  `lines(ii,ECP,col="red")`

観察度数TCP，期待度数ECPを使ってカイ二乗適合度検定する

- カイ二乗値**XX**は，
  XX <- sum((TCP-ECP)^2/ECP)
  で得られる。
- 自由度は完結出生力のカテゴリ数**-1**から，ポアソン分布の平均値もデータから計算したために，さらに**1**を引いて，**length(ECP)-2**となるので，**TCP**と**ECP**に差が無いという帰無仮説を検定した場合の有意確率は，
  1-pchisq(XX,length(ECP)-2)
  によって得られる。計算してみると**0.816**となる。有意水準**5%**で有意な差があるとはいえない。

vcdライブラリのgoodfit()を使う方法

- TCPの計算まではできているとして，やることは
  - vcdライブラリを読み込む
  - goodfit()を使う
- goodfit()での分布の当てはめ方法として最小二乗近似(MinChisq)か最尤法(ML)を選んで決める。
  - MinChisqだと原理は簡単だが，期待度数が1未満のときに近似が不正確になるなど問題がある（警告がでる）→カイ二乗検定
  - MLは計算時間がかかるがより正確（警告もでない）→尤度比カイ二乗検定
- コードは以下2行のみ

```
library(vcd)
summary(goodfit(TCP, type="poisson", method="MinChisq"))
```

または，下の行は，以下でもOK

```
summary(goodfit(TCP, type="poisson", method="ML"))
```

2009年第10回の課題から

- ある小学校の6年生は4クラスあり，職員室から近い順に，1組30人，2組28人，3組29人，4組30人が在籍している。ある週の月曜日，X先生は咳をしながら授業をしていたのだが，翌日高熱を発して欠勤し，近医を受診したところ，A型インフルエンザと診断された。水曜日，1組10人，2組5人，3組4人，4組1人がインフルエンザ様症状で学校を休んだため，1組は学級閉鎖になった（注：通常，校長が学級閉鎖を決定する目安は20%の欠席である。新型インフルエンザの場合，当初は1人でも患者が出たら全県が休校となったし，その後も暫くは10%の欠席で学級閉鎖とされていたが，現在では季節性インフルエンザと同様，20%になっている）。
- この小学校6年生の4クラスにおいて，(1)欠席者の割合にはどのクラスでも差がないという帰無仮説を検定し，もし差があったなら，(2)どのクラスとどのクラスの間で差があるのかを検定せよ（検定の多重性はHolmの方法で調整すること）。また，(3)職員室から近いほどインフルエンザ罹患リスクが高く，リスクのスコアとして1組が4，2組が3，3組が2，4組が1であるという仮説が成り立つかどうか，コクラン＝アーミテージ検定せよ。なお，検定の有意水準はすべて5%とする。すべての検定に先立ち，クラスごとのインフルエンザ様疾患による欠席者割合を図示し，図と検定結果を参照して考察すること。

情報を整理して分解する

- 小学6年生4クラス，人数はc(30,28,29,30)
- 欠席人数はc(10,5,4,1)
- 罹患リスクスコアはc(4,3,2,1)
- やることは以下4点
  (1) クラス別欠席者割合の図示
  (2) 欠席者割合にクラス間で差が無いという帰無仮説の検定
  (3) 差があったとき2組ずつの多重比較
  (4) 罹患リスクスコアに応じた欠席者割合の傾向があるかどうか

生の情報をオブジェクトに付値する

- クラスごとの人数を**pop**，欠席者数を**abs**，罹患リスクを**risk**として

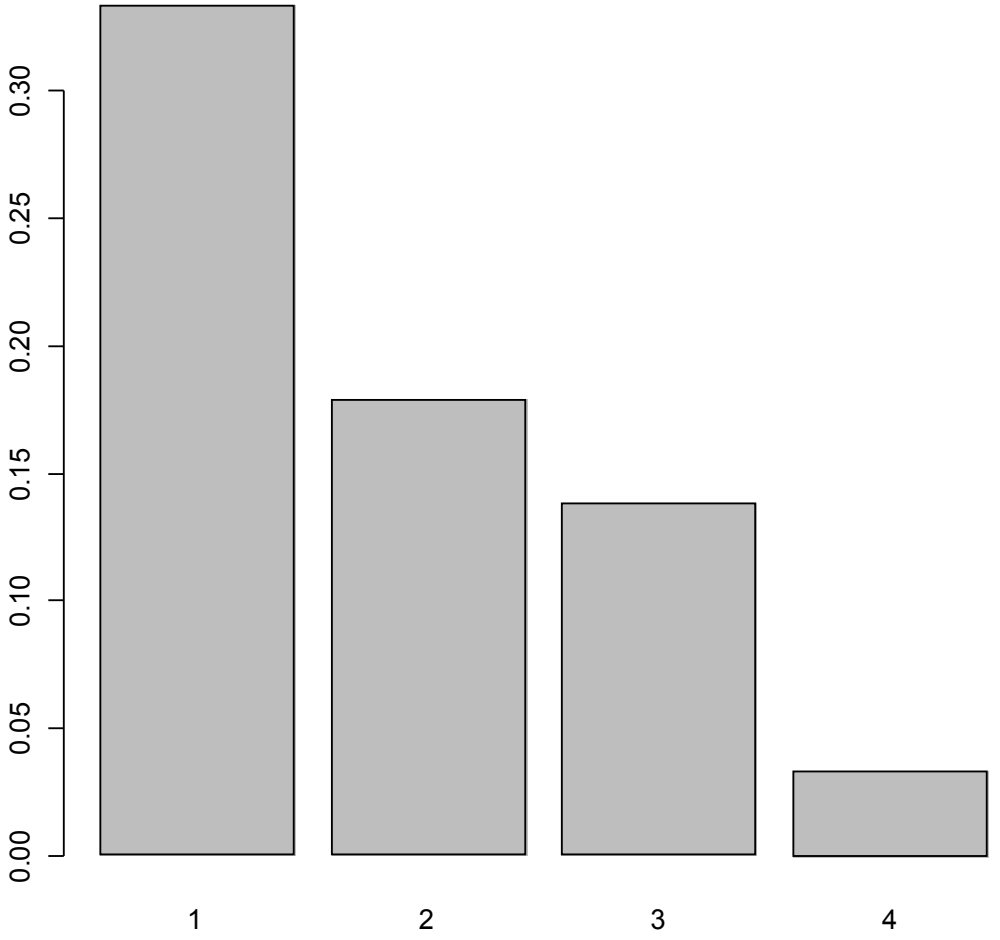
```
pop <- c(30,28,29,30)
names(pop) <- c("1","2","3","4")
abs <- c(10,5,4,1)
risk <- c(4,3,2,1)
```

（1）クラス別欠席者割合を図示

- 欠席者割合はabs/popで得られるので，それをbarplot()に渡すだけでOK。つまり，

  barplot(abs/pop)

  とすると右図が得られる。



(2) 欠席者割合に差が無いという帰無仮説の検定

- もっとも単純には，**prop.test(abs,pop)**だけで**OK**。結果は下記

```
        4-sample test for equality of proportions without
        continuity correction
data:  abs out of pop
X-squared = 9.8253, df = 3, p-value = 0.02011
alternative hypothesis: two.sided
sample estimates:
    prop 1     prop 2     prop 3     prop 4
0.33333333 0.17857143 0.13793103 0.03333333
 警告メッセージ:
In prop.test(abs, pop) :  カイ自乗近似は不正確かもしれません
```

- 4組の期待値が小さすぎるためカイ二乗近似が不正確かもしれない（ので，本来はフィッシャーの正確確率検定をするべきかもしれない。その話はクロス集計のところで）が，とりあえず**p**値が**0.02**と**5%**未満なので，帰無仮説は棄却される。
- つまり，欠席者割合はクラスによって統計的に有意な差があったといえる。

(3) 2クラスずつの割合の差の多重比較

- 統計的に有意な差があったので，
  pairwise.prop.test(abs,pop)
  を実行すると，以下が表示される

```
        Pairwise comparisons using Pairwise comparison of
proportions
data:  abs out of pop
  1     2     3
2 0.888 -     -
3 0.725 0.954 -
4 0.046 0.725 0.888
P value adjustment method: holm
 警告メッセージ:
1: In prop.test(x[c(i, j)], n[c(i, j)], ...) :
   カイ自乗近似は不正確かもしれません
2: In prop.test(x[c(i, j)], n[c(i, j)], ...) :
   カイ自乗近似は不正確かもしれません
3: In prop.test(x[c(i, j)], n[c(i, j)], ...) :
   カイ自乗近似は不正確かもしれません
```

この対比（つまり1組と4組）のみ5%水準で有意差あり

4組の期待値が小さいのでFisherで多重比較すべき

(4) コクラン＝アーミテージ検定

- リスクスコアが高いほど欠席者割合が高い傾向があるかどうかを調べる
  prop.trend.test(abs,pop,risk)
- 結果は下記の通り
  Chi-squared Test for Trend in Proportions
  data: abs out of pop ,
  using scores: 4 3 2 1
  X-squared = 9.38, df = 1, p-value = 0.002194
- 有意確率が0.002なので傾向が無いという帰無仮説は棄却され，リスクスコアが高いクラスほど欠席者割合が多い傾向があると言えた。