21 日目：χ^2検定

さて，本日からは平均，相関から離れたものをやってみようと思います。まずはχ^2検定からです。データは「sam2.xls」を使い，x という名前で読み込んでいるとします。

通常のデータから計算する場合，まずはクロス表の作成から始めるでしょう。９日目に table というコマンドを使いましたが，これを使えば２変数のクロス表も作れます。カッコ内の前の変数がクロス表の縦（行）に，後ろが横（列）になります。

```
table(x$性別, x$学年)
```

以下のように，おそろしく（？）そっけない出力ですが…

```
> table(x$性別, x$学年)

   1  2  3
1 43 72 10
2 74 76 23
```

もちろんこれでは欠損値を無視してしまうので，欠損値も表に含めたいなら，exclude=NULL を入れる必要があります。

```
table(x$性別, x$学年, exclude=NULL)
```

ついで，ということで３変数のクロス表を作りたいなら，table のカッコ内に３つの変数を並べるということで，「とりあえず」対応できます。

```
table(x$性別, x$学年, x$専攻)
```

この場合，カッコ内の３つ目の変数で群分けをし，１つ目の変数がクロス表の縦（行）に，２つ目が横（列）になったクロス表を群の数だけ返してきます。「，，　= 1」のように表示がありますが，カッコ内の３つの変数とみなせばわかりやすいのではないでしょうか。

%などが入ったものが一気に作れないかと探してみましたが，どうも自作関数が必要なようです。ここまで出してくれていたら，エクセルでの加工もそれほど手間ではないと思いますが，必要ならば自分で作るか，お借りするかしましょう。

さて，先のクロス表に対してχ^2検定をやってみます。いうまでもないと思いますが，いわゆる独立性の検定（性別と学年の間に関連があるかどうか）です。

```
chisq.test(x$性別, x$学年)
```

結果は，以下のように返ってきます。

```
> chisq.test(x$性別, x$学年)

    Pearson's Chi-squared test

data:  x$性別 and x$学年
X-squared = 5.8636, df = 2, p-value = 0.0533
```

χ^2値，自由度，有意確率が並んでいます。この結果，５％水準に届きませんので，有意な偏り（関連）は認められないということになります。

もし有意になれば，どのセルに偏りが認められるのかという残差分析を行いたい場合も出てくるでしょう。しかし chisq.test はそれをやってくれません。ここからは自作が必要です…。「js-STAR」の利用も一案でしょう。R のプログラムも表示してくれます。
http://www.kisnet.or.jp/nappa/software/star/index.htm

なので，本日はこれでおしまい…としてもいいのかもしれませんが，もう少し関連することを…

たとえば，資料にある中学生を対象とした調査で，学年別に部活動に所属しているかどうかをたずねた結果があったとします。その内訳は以下のようでした。このようなデータに対してχ^2検定をやってみることを考えます。

	所属	無所属
１年	237	56
２年	179	102
３年	128	150

この表にある数値を直接Rに認識させてもよいのですが，とりあえずエクセルに入力します。右のような感じです。

	A	B	C
1		所属	無所属
2	１年	237	56
3	２年	179	102
4	３年	128	150
5			

次に，コピペで R にこのデータを認識させます。

変数名を含むデータの部分，つまり A1 から C4 の部分をドラッグで選択し，コピー。ここまでがエクセルでの作業です。

次に R に移って，次のような命令を出します。R エディタでも R コンソールでもどちらでもかまいません。

```
club <- read.table(pipe("pbpaste"), header=TRUE, row.names=1,
fileEncoding="CP932")
```

これを実行すれば，先の表が club という名前読み込まれているはずです。
なお，以下のような警告が出てくることがあります。

```
> club <- read.table(pipe("pbpaste"), header=TRUE, row.names=1, fileEncoding="CP932")
警告メッセージ:
read.table(pipe("pbpaste"), header = TRUE, row.names = 1, fileEncoding = "CP932") で:
 incomplete final line found by readTableHeader on 'pbpaste'
```

警告が出ても，ほとんどの場合はちゃんと読み込めているようですが，読み込んだ中身を表示させて確認してください。
検定の前に，すこし読み込みの命令について解説しておきます。これは Mac 用であり，Win だとちょっと違います。カッコの中は，csv ファイルを読み込んだ時と似ていると思います。読み込むファイル名が，pipe("pbpaste")というあたりが違うところです。また今回のように，１列目に行の変数名が入っている場合は，row.names=1 と入力し，１列目は行の変数名であることを伝えておきます。

では検定ですが，表自体がひとつのファイルになっている場合，そのファイルを指定するだけで計算してくれます。つまり…

```
chisq.test(club)
```

これで OK です。１%水準で有意という結果になります。

ちなみにこのやり方を知ると，「先の性別と学年の検定も，table で表を作っているのだから，その結果を読み込んで検定ができるのではないか？」と考えつくかもしれません。もちろんできます。

```
x.ta <- table(x$性別, x$学年)
chisq.test(x.ta)
```

さて，χ^2 検定は，独立性の検定をはじめ，様々な使い方があります。他の使い方も，概ねデータを操作することで可能になるようです。たとえば，このデータでは，１年生の部活所属は，所属している者 237，所属していない者 56 でした。全国調査では，１年生の部活所属率は 63%であることがわかっていたとしましょう。この中学の所属率は全国調査と比べて差があるでしょうか？　いわゆる適合度の検定ですが，このような使い方もできます。

命令は以下のようです。c(237, 56)として手持ちのデータを入れておき，全国の方の確率を p=c(63, 37)/100 として指定します。p=では確率を指定しますが，その和は１である必要があります。そのため，%表示のデータを小数表示に直すため，100 で割ってあります。

```
chisq.test(c(237, 56), p=c(63, 37)/100)
```

やってみると，結果は有意になります。つまりこの中学校の部活所属率は全国と比べて異なっている（所属率が高い）といえます。

今回は（も？）概略だけを紹介してきました。本格的に必要なら，さらに調べてください。