

21日目： χ^2 二乗検定

さて、本日から少しノンパラメトリックをやってみようと思います。まずは χ^2 二乗検定からです。

通常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ね。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。昨日までのデータを使うと…

```
table(xx$性別, xx$学年)
```

以下のように、おそろしく(?) そっけない出力ですが…

```
> table(xx$性別, xx$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(xx$性別, xx$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することで、「とりあえず」対応できます。

```
table(xx$性別, xx$学年, xx$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。

%などが入ったものが作れないかと探してみましたが、どうもそれには自作関数が必要なようです。ここまで出してくれていたなら、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 二乗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(xx$性別, xx$学年)
```

結果は、以下のように返ってきます。

```
> chisq.test(xx$性別, xx$学年)

Pearson's Chi-squared test

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

χ^2 二乗値、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…

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なようでした。このようなデータに対して χ^2 二乗検定をやってみることを考えます。

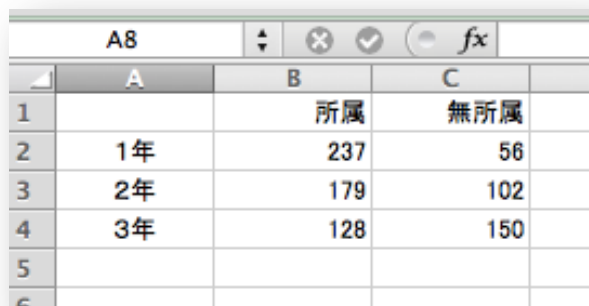
	所属	無所属
1年	237	56
2年	179	102
3年	128	150

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。以下のような感じです。

次に、コピーで 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。ここまでの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



	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			
6			

もどちらでもかまいません。（ここでは2行にわたっていますが、1行に入らなかっただけであり、改行に意味はありません）

```
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、よく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：  
In read.table(pipe("pbpaste"), header = TRUE, row.names = 1, fileEncoding = "CP932") :  
  'pbpaste' の readTableHeader で不完全な最終行が見つかりました
```

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。

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

さて、 χ^2 乗検定は、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、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。