エビデンスベーストヘルスケア特講 I (8) 分散分析と多重比較

中澤　港（国際保健学領域・教授）

最終更新：2013 年 6 月 5 日

1　3 群以上の位置母数の差の検定

3 群以上を比較するために，**単純に 2 群間の差の検定を繰り返すことは誤り**である。なぜなら，n 群から 2 群を抽出するやりかたは ${}_nC_2$ 通りあって，1 回あたりの第 1 種の過誤（本当は差がないのに，誤って差があると判定してしまう確率）を 5% 未満にしたとしても，3 群以上の比較全体として「少なくとも 1 組の差のある群がある」というと，全体としての第 1 種の過誤が 5% よりずっと大きくなってしまうからである。

この問題を解消するには，**多群間の比較という捉え方をやめて，群分け変数が注目している量の変数に与える効果があるかどうかという捉え方にする**のが一つの方法であり，具体的には一元配置分散分析やクラスカル＝ウォリス (Kruskal-Wallis) の検定がこれに当たる[*1]。

そうでなければ，有意水準 5% の 2 群間の検定を繰り返すことによって全体として第 1 種の過誤が大きくなってしまうことが問題なので，**第 1 種の過誤を調整することによって全体としての検定の有意水準を 5% に抑える**方法もある。このやり方は「多重比較法」と呼ばれる。

1.1　一元配置分散分析

一元配置分散分析では，データのばらつき（変動）を，群間の違いという意味のはっきりしているばらつき（群間変動）と，各データが群ごとの平均からどれくらいばらついているか（誤差）をすべての群について合計したもの（誤差変動）に分解して，前者が後者よりもどれくらい大きいかを検討することによって，群分け変数がデータの変数に与える効果があるかどうかを調べる。数学モデルとして捉えれば，量的なデータを目的変数，グループを説明変数とする線型回帰モデルといえる。グループが 2 群の場合は平均値の差を検討するための t 検定となるわけだが，3 群以上の場合は一元配置分散分析となる。

例えば，南太平洋の 3 つの村 X, Y, Z で健診をやって，成人男性の身長や体重を測ったとしよう。このとき，データは例えば次のようになる（架空のものである）[*2]。

ID 番号	村落 (VG)	身長 (cm)(HEIGHT)
1	X	161.5
2	X	167.0
（中略）		
22	Z	166.0
（中略）		
37	Y	155.5

村落によって身長に差があるかどうかを検定したいならば，HEIGHT という量的変数に対して，VG という群分け変数の効果があるかどうかを一元配置分散分析することになる。R コンソールでは以下のように入力する。

```
sp <- read.delim("http://minato.sip21c.org/grad/sample2.dat")
summary(aov(HEIGHT ~ VG, data=sp))
```

すると，次の枠内に示す「分散分析表」が得られる。

[*1] なお，分散分析は本来，その効果をみるための実験計画をした上で実施するものだから，群ごとのサンプルサイズは揃っているべきだし，効果の有無を効率よく検出するのに適したサンプルサイズが設計されているべきだが，現実には実験計画されていないデータにも適用されている。適切なサンプルサイズは，母集団の均質性，サブグループ数，母集団のパラメータ推定に求めたい正確さ，注目している現象の出現頻度，予算などで変わってくる。

[*2] http://minato.sip21c.org/grad/sample2.dat として公開してあり，R から read.delim() 関数で読み込み可能である。

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
VG           2  422.72  211.36  5.7777 0.006918 **
Residuals   34 1243.80   36.58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

右端の*の数は有意性を示す目安だが，確率そのものに注目してみるほうがよい。Sum Sq のカラムは偏差平方和を意味する。VG の Sum Sq の値 422.72 は，村ごとの平均値から総平均を引いて二乗した値を村ごとの人数で重み付けした和であり，群間変動または級間変動と呼ばれ，VG 間でのばらつきの程度を意味する。Residuals の Sum Sq の値 1243.80 は各個人の身長からその個人が属する村の平均身長を引いて二乗したものの総和であり，誤差変動と呼ばれ，村によらない（それ以外の要因がないとすれば偶然の）ばらつきの程度を意味する。Mean Sq は平均平方和と呼ばれ，偏差平方和を自由度 (Df) で割ったものである。平均平方和は分散なので，VG の Mean Sq の値 211.36 は群間分散または級間分散と呼ばれることがあり，Residuals の Mean Sq の値 36.58 は誤差分散と呼ばれることがある。F value は分散比と呼ばれ，群間分散の誤差分散に対する比である。この場合の分散比は第 1 自由度 2，第 2 自由度 34 の F 分布に従うことがわかっているので，それを使った検定の結果，分散比がこの実現値よりも偶然大きくなる確率（Pr(>F) に得られる）が得られる。この例では 0.00692 なので，VG の効果は 5% 水準で有意であり，帰無仮説は棄却される。つまり，身長は村落によって有意に異なることになる。

EZR で sample2.dat **を** sp **というデータフレームに読み込むには，[ファイル] の [データのインポート] から [テキストファイル，クリップボードまたは URL から] と進んで，[データフレーム名を入力:] のところに** sp **と打ち，[インターネット URL] の右側のラジオボタンをチェックし，フィールド区切りを [タブ] として [OK] をクリックして表示されるダイアログに** http://minato.sip21c.org/grad/sample2.dat **と入力して [OK] する。**
ANOVA を実行するには，[統計解析] の [連続変数の解析] で [三群以上の間の平均値の比較（一元配置分散分析 one-way ANOVA)] を選び，「目的変数」として HEIGHT **を，「比較する群」として** VG **を選び，[OK] をクリックすればよい。エラーバー付きの棒グラフが自動的に描かれ，アウトプットウィンドウには分散分析表に続いて，村ごとの平均値と標準偏差の一覧表が表示される。右端の p 値は一元配置分散分析における** VG **の効果の検定結果を再掲したものになっている。**

古典的な統計解析では，各群の母分散が等しいことを確認しないと一元配置分散分析の前提となる仮定が満たされない。母分散が等しいという帰無仮説を検定するには，バートレット (Bartlett) の検定と呼ばれる方法がある。R では，量的変数を Y，群分け変数を C とすると，bartlett.test(Y~C) で実行できる[*3]。この結果得られる p 値は 0.5785 なので，母分散が等しいという帰無仮説は有意水準 5% で棄却されない。これを確認できると，安心して一元配置分散分析が実行できる。

EZR では，メニューバーの「統計解析」から「連続変数の解析」の「三群以上の等分散性の検定 (Bartlett 検定)」を選び，「目的変数」として HEIGHT**，「グループ」として** VG **を選んで [OK] する。**

しかし，このような 2 段階の検定は，検定の多重性の問題を起こす可能性がある。群馬大学の青木繁伸教授や三重大学の奥村晴彦教授の数値実験によると，等分散であるかどうかにかかわらず，2 群の平均値の差の Welch の方法を多群に拡張した方法を用いるのが最適である。R では oneway.test() で実行できる。上記，村落の身長への効果をみる例では，oneway.test(HEIGHT ~ VG, data=sp) と打てば，Welch の拡張による一元配置分散分析ができて，以下の結果が得られる。

```
> oneway.test(HEIGHT ~ VG, data=sp)

One-way analysis of means (not assuming equal variances)

data:  HEIGHT and VG
F = 7.5163, num df = 2.00, denom df = 18.77, p-value = 0.004002
```

残念ながら EZR でも Rcmdr でもメニューにないので，スクリプトウィンドウで aov **の部分を** oneway.test **に書き直して「実行」するしかない。**

[*3] もちろん，これらがデータフレーム dat に含まれる変数ならば，bartlett.test(Y~C, data=dat) とする。

1.2 クラスカル＝ウォリス (Kruskal-Wallis) の検定と Fligner-Killeen の検定

多群間の差を調べるためのノンパラメトリックな方法としては，クラスカル＝ウォリス (Kruskal-Wallis) の検定が有名である。R では，量的変数を Y，群分け変数を C とすると，`kruskal.test(Y~C)` で実行できる。以下，Kruskal-Wallis の検定の仕組みを箇条書きで説明する。

- 「少なくともどれか 1 組の群間で大小の差がある」という対立仮説に対する「すべての群の間で大小の差がない」という帰無仮説を検定する。
- まず 2 群の比較の場合の順位和検定と同じく，すべてのデータを込みにして小さい方から順に順位をつける（同順位がある場合は平均順位を与える）。
- 次に，各群ごとに順位を足し合わせて，順位和 $R_i (i = 1, 2, ..., k; k$ は群の数) を求める。
- 各群のオブザーベーションの数をそれぞれ n_i とし，全オブザーベーション数を N としたとき，各群について統計量 B_i を $B_i = n_i\{R_i/n_i - (N+1)/2\}^2$ として計算し，

  $$B = \sum_{i=1}^{k} B_i$$

  として B を求め，$H = 12 \cdot B/\{N(N+1)\}$ として H を求める。同順位を含むときは，すべての同順位の値について，その個数に個数の 2 乗から 1 を引いた値を掛けたものを計算し，その総和を A として，

  $$H' = \frac{H}{1 - \frac{A}{N(N^2-1)}}$$

  により H を補正した値 H' を求める。
- H または H' から表を使って（データ数が少なければ並べかえ検定によって）有意確率を求めるのが普通だが，$k \geq 4$ で各群のオブザーベーション数が最低でも 4 以上か，または $k = 3$ で各群のオブザーベーション数が最低でも 5 以上なら，H や H' が自由度 $k - 1$ のカイ二乗分布に従うものとして検定できる。

上の例で村落の身長への効果をみるには，R コンソールでは，`kruskal.test(HEIGHT ~ VG, data=sp)` と打てば結果が表示される。

EZR では，「統計解析」，「ノンパラメトリック検定」，「3 群以上の間の比較 (Kruskal-Wallis 検定)」と選び，「グループ」として VG を，「目的変数」として HEIGHT を選び，[OK] をクリックするだけである。
自動的に層別箱ひげ図が描かれ，アウトプットウィンドウには 3 群それぞれの中央値に続いて，クラスカル＝ウォリス検定の結果が表示される。

Fligner-Killeen の検定は，グループごとのばらつきに差が無いという帰無仮説を検定するためのノンパラメトリックな方法である。Bartlett の検定のノンパラメトリック版といえる。上の例で，身長のばらつきに村落による差が無いという帰無仮説を検定するには，R コンソールでは，`fligner.test(HEIGHT ~ VG, data=sp)` とすればよい。

EZR や Rcmdr のメニューには入っていないので，必要な場合はスクリプトウィンドウにコマンドを打ち，選択した上で「実行」ボタンをクリックする。

1.3 検定の多重性の調整を伴う対比較

多重比較の方法にはいろいろあるが，良く使われているものとして，ボンフェローニ (Bonferroni) の方法，ホルム (Holm) の方法，シェフェ (Scheffé) の方法，テューキー (Tukey) の HSD，ダネット (Dunnett) の方法，ウィリアムズ (Williams) の方法がある。ボンフェローニの方法とシェフェの方法は検出力が悪いので，特別な場合を除いては使わない方が良い。テューキーの HSD またはホルムの方法が薦められる。なお，ダネットの方法は対照群が存在する場合に対照群と他の群との比較に使われるので（そのノンパラメトリック版はスティール法である），適用場面が限定されている[*4]。ウィリアムズの方法は対照群があって他の群にも一定の傾向が仮定される場合には最高の検出力を発揮するが，ダネットの方法よりもさらに限られた場合にしか使えない。

テューキーの HSD は平均値の差の比較にしか使えないが（ノンパラメトリック版としては，スティール・ドゥワスの方法がある），ボンフェロー二の方法やホルムの方法は位置母数のノンパラメトリックな比較にも，割合の差の検定にも使える。R コン

[*4] ただし，対照群が他の群との比較のすべての場合において差があるといいたい場合は，多重比較をするのではなくて，t 検定を繰り返して使うのが正しいので，注意が必要である。もちろんそういう場合は多くはない。

ソールでは，pairwise.t.test()，pairwise.wilcox.test()，pairwise.prop.test() という関数で，ボンフェローニの方法やホルムの方法による検定の多重性の調整ができる。

例えば，上の例で，どの村落とどの村落の間で身長に差があるのかを調べたい場合，R コンソールでは，

```
pairwise.t.test(sp$HEIGHT, sp$VG, p.adjust.method="bonferroni")
```

とすれば，すべての 2 村落の組み合わせについてボンフェローニの方法で有意水準を調整した p 値が表示される[*5]。

また，pairwise.wilcox.test(sp$HEIGHT,sp$VG,p.adjust.method="bonferroni") とすれば，ボンフェローニの方法で有意水準を調整した，すべての村落間での身長の差を順位和検定した結果を出してくれる。これらの関数で，p.adjust.method を指定しなければホルムの方法になるが，明示したければ，p.adjust.method="holm"とすればよい。R でもボンフェローニが可能なのは，一番単純な方法であるという理由と，ホルムの方法に必要な計算がボンフェローニの計算を含むからだと思われる。なお，データがきれいな正規分布をしていれば，TukeyHSD(aov(HEIGHT ~ VG, data=sp)) などとして，テューキーの HSD を行ってもよい。

EZR では，一元配置分散分析メニューのオプションとして実行できる。「統計解析」「連続変数の解析」「3 群以上の平均値の比較（一元配置分散分析 one-way ANOVA)」を選んで，「目的変数」として HEIGHT を，「比較する群」として VG を選んでから，下の方の「↓ 2 組ずつの比較 (post-hoc 検定) は比較する群が 1 つの場合のみ実施される」から欲しい多重比較法の左側のボックスにチェックを入れてから「OK」ボタンをクリックする。

いずれのやり方をしても，TukeyHSD の場合だと，2 群ずつの対比較の結果として，差の推定値と 95% 同時信頼区間に加え，Tukey の方法で検定の多重性を調整した p 値が下記のように表示され，検定の有意水準が 5% だったとすると，Z と Y の差だけが有意であることがわかる。

```
> TukeyHSD(AnovaModel.3, "factor(VG)")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = HEIGHT ~ factor(VG), data = sp, na.action = na.omit)

$`factor(VG)`
         diff        lwr      upr     p adj
Y-X -2.538889 -8.3843982  3.30662 0.5423397
Z-X  5.850000 -0.9598123 12.65981 0.1038094
Z-Y  8.388889  2.3382119 14.43957 0.0048525
```

1.4 Dunnett の多重比較法

Dunnett の多重比較は，コントロールと複数の実験群の比較というデザインで用いられる。以下，簡単な例で示す。例えば，5 人ずつ 3 群にランダムに分けた高血圧患者がいて，他の条件（食事療法，運動療法など）には差をつけずに，プラセボを 1 ヶ月服用した群の収縮期血圧 (mmHg 単位) の低下が 5, 8, 3, 10, 15 で，代表的な薬を 1 ヶ月服用した群の低下は 20, 12, 30, 16, 24 で，新薬を 1 ヶ月服用した群の低下が 31, 25, 17, 40, 23 だったとしよう。このとき，プラセボ群を対照として，代表的な薬での治療及び新薬での治療に有意な血圧降下作用の差が出るかどうかを見たい（悪くなるかもしれないので両側検定で）という場合に，Dunnett の多重比較を使う。R でこのデータを bpdown というデータフレームに入力して Dunnett の多重比較をするためには，次のコードを実行する。

[*5] "bonferroni"は"bon"でも良い。また，pairwise.*系の関数では data=というオプションが使えないので，データフレーム内の変数を使いたい場合は，予めデータフレームを attach() しておくか，またはここで示したように，変数指定の際に一々，“データフレーム名 $” を付ける必要がある。

```
bpdown <- data.frame(
 medicine=factor(c(rep(1,5),rep(2,5),rep(3,5)), labels=c("プラセボ","代表薬","新薬")),
 sbpchange=c(5, 8, 3, 10, 15, 20, 12, 30, 16, 24, 31, 25, 17, 40, 23))
summary(res1 <- aov(sbpchange ~ medicine, data=bpdown))
library(multcomp)
res2 <- glht(res1, linfct = mcp(medicine = "Dunnett"))
confint(res2, level=0.95)
summary(res2)
```

つまり，基本的には，multcomp ライブラリを読み込んでから，分散分析の結果を glht() 関数に渡し，linfct オプションで，Dunnett の多重比較をするという指定を与えるだけである。

EZR では，まず「ファイル」「データのインポート」「ファイルまたはクリップボード，**URL** からテキストデータを読み込む」として，「データセット名を入力」の右側のボックスに bpdown と入力し，「データファイルの場所」として「インターネットの **URL**」の右側のラジオボタンをクリックし，「フィールドの区切り記号」を「タブ」にして「**OK**」ボタンをクリックする。表示される **URL** 入力ウィンドウに http://minato.sip21c.org/bpdown.txt と打って「**OK**」ボタンをクリックすれば，上記データを読み込むことができる。
そこで「統計解析」の「連続変数の解析」から「**3** 群以上の平均値の比較（一元配置分散分析 **one-way ANOVA)**」を選んで，「目的変数」として sbpchange，「比較する群」として medicine を選び，「**2** 組ずつの比較 **(Dunnett** の多重比較**)**」の左のチェックボックスをチェックしてから「**OK**」ボタンをクリックすればいい。
なお，このデータで処理名を示す変数 medicine の値として o.placebo，1.usual，2.newdrug のように先頭に数字付けた理由は，それがないと水準がアルファベット順になってしまい，**Dunnett** の解析において新薬群がコントロールとして扱われてしまうからである。
ノンパラメトリック検定の場合は，「統計解析」の「ノンパラメトリック検定」から「**3** 群以上の間の比較 **(Kruskal-Wallis** 検定**)**」と選び，「目的変数」を sbpchange，「グループ」を medicine にし，「**2** 組ずつの比較 **(post-hoc** 検定、**Steel** の多重比較**)**」の左のチェックボックスをチェックして「**OK**」ボタンをクリックすれば，**Steel** の多重比較が実行できる。

最近の多重比較の流行としては，対比較を繰り返して p 値を調整するのではなく，棄却される帰無仮説のうち間違って棄却される割合の期待値としての誤検出率 (FDR) を調整する方法があり（Benjamini Y and Hochberg Y, “Controlling the false discovery rate: A practical and powerful approach to multiple testing.” *J. Royal Stat. Soc. Ser. B*, 57(1): 289-300, 1995），R コンソールでは，pairwise.t.test や pairwise.wilcox.test の p.adjust.method オプションで"holm"や"bon"の代わりに"fdr"と打てば実行できる。（**残念ながら，まだ Rcmdr や EZR のメニューには含まれていない**）