【統計一口メモ　第 23話】

Bonferroni（ボンフェローニ）検定、Holm（ホルム）検定とは？

名古屋市立大学大学院医学研究科 非常勤講師　薬学博士　松本一彦

前話（第 22 話）では多重比較検定の中でもパラメトリックの Tukey（チューキー）検定とノンパラメトリックの Steel-Dwass（スチール・ドワス）検定をあげました。今回は論文にも多く見られる Bonferroni 検定と、ほとんどお目にかからない Holm 検定をとりあげます。

なお、Bonferroni 検定も Holm 検定も総当たりの Tukey 検定とは異なり、必要な群間を選択できるという特徴を持っています。ただ、前回の Tukey 検定との比較のために、ここでは、敢えて総当たりで解析しています。

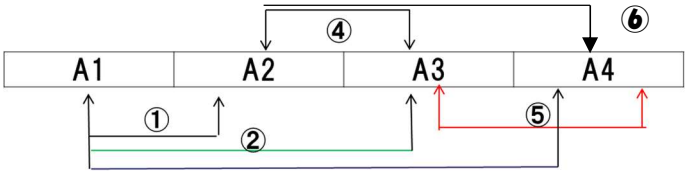
例題：4種類の方法（A1～A4)を比較するためにある特性値についてデータをランダムサンプリングした。どの方法の間に有意差があるだろうか？

永田靖，吉田道弘著「統計的多重比較法の基礎」サイエンティスト社より引用[1)]

群	n_i	1	2	3	4	5	6	7
A1	7	10.7	9.7	8.5	9.4	8.8	8.4	10.6
A2	7	8.1	8.3	8.7	6.9	5.7	9.5	6.7
A3	7	7.9	7.5	7.4	9.2	5.7	8.3	9.7
A4	7	6.2	7.1	5.5	4.7	6.3	6.9	7.5

§1. Bonferroni 検定

検定の回数に注目して危険率を調整する方法です。帰無仮説の個数（対比較する数；k個）を数えて、有意水準を調整し（α/k），棄却域を決めることで多重性の調整を行っています。言い換えると Tukey 検定が群間比較を総当たりで行うことに比べ、本法は任意な群間比較ができるということです。ただ、今回は Tukey 検定との比較も兼ねているので、とりあえず総当たりとします。



対比較は Tukey検定の場合と同じ6回とします。

1 回当たりの検定の有意確率が α^* である検定をk回繰り返した場合、全体の α は $\alpha = k \times \alpha^*$ を越えることはありません。つまり $\alpha^* = \alpha/k$とすればk回検定したときの全体の危険率を α 以下に抑えられます。上の例では$\alpha^* = \alpha/6$ で $\alpha = 0.05$ が 0.05/0.008 と厳しくなります。

t統計量を求めるために、平均と分散を求めます。

				データ					合計	平均	分散
群	n_i	1	2	3	4	5	6	7	T_i	$X_{\cdot i}$	V_i
A1	7	10.7	9.7	8.5	9.4	8.8	8.4	10.6	66.1	9.44	0.8962
A2	7	8.1	8.3	8.7	6.9	5.7	9.5	6.7	53.9	7.70	1.7333
A3	7	7.9	7.5	7.4	9.2	5.7	8.3	9.7	55.7	7.96	1.7195
A4	7	6.2	7.1	5.5	4.7	6.3	6.9	7.5	44.2	6.31	0.9414

残差自由度Φと誤差分散 V_E を求めます。

$$\phi = \Sigma n_i - 4 = 24$$

$$V_E = \frac{\Sigma (n_i-1)V_i}{\phi} = \frac{6x0.8962+6x1.733+6x1.720+6x0.9414}{24} = 1.3230$$

t統計量を求めます。A1 群と A2 群間のt値は次のような値になります。

$$|t_{12}| = \frac{|x_{\cdot 1} - x_{\cdot 2}|}{\sqrt{V_E (1/n_1+1/n_2)}} = \frac{|9.443-7.700|}{\sqrt{1.323(1/7+1/7)}} = 2.835$$

ここで留意しなければならないのは、このt統計量は Student のt検定とは異なり分母の誤差分散は2群間ではなく全群の分散V_Eを用いているということです。（※ボクのつぶやき：統計ソフトを使うと検定結果はp値が出てしまうので実務的には気にならないなー）。

他の組み合わせも計算し、下記の表が求められます。

		$\|t_{ij}\|$ の値			
	A1	A2	A3	A4	
A1		2.835	2.417	5.089	*
A2			0.418	2.254	
A3				2.672	

		p値			
	A1	A2	A3	A4	
A1		0.0549	0.1417	0.0002	**
A2			4.0767	0.2015	
A3				0.0799	

永田・吉田本[1]には、「棄却限界値はt表より、$t(\phi_E, (\alpha/2)/k) = t(24\ 0.025/6) = 2.875$
t_{ij} の値が 2．875 以上であれば A_i と A_j の母平均に差あるといえる」。とありますが、今日ではt表を用いることはなく、エクセル関数の TINV を使います。その時は TINV(0.05/6，24)とします。なお、TINV 関数では α が 0.05 になっていることに留意してください。ここでは、A1 vs A4 間でのみ有意差がみられました。

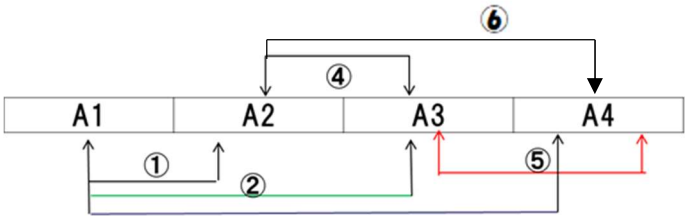
§2. Holm 検定

Holm 検定は、Bonferroni 検定の改良版で、より有意差が出やすい方法です。

Holm 検定という単語を初めて聞いたという方も少なくないと思います。お手持ちの参考書でも Holm 検定が載っている本はまれです。でも、OECD の TG210 統計ガイダンス[2]にも載っているぐらい国際的には知られている手法です。では、なぜ、論文でも見かけることが少ないのでしょうか？それは、どんなに優れている手法でも、市販統計ソフトに取り入れられない限り、繁用されないから

です。みなさんがお使いの JMP、SPSS、PRISM にも搭載されていません。でも、SAS を使う統計家の間ではよく知られている手法なので Bonferroni 検定では満足しない研究者の間で使われています。ちなみに、Pharmaco Basic には登載しています。

総当たりの組み合わせを再掲します。



Bonferroni 検定はα/kで多重性を調整していましたが、Holm 検定は、この対比較数kを使って

$$\alpha_1 = \alpha/k、\ \alpha_2 = \alpha/(k-1),\ \alpha_3 = \alpha/(k-2),\text{----}\ \alpha_k = \alpha$$

でαに重みを付けていきます。

例えば、4群の場合で対比較数を6とした場合、$\alpha_1 = \alpha/6$、$\alpha_2 = \alpha/5$-----$\alpha_6 = \alpha/1$ となり、それぞれのp値を小さい順にならべて検定していきます。

有意差が認められない段階で，それ以降の検定には進まず、解析が終了となります。この閉鎖下降手順は Williams 検定と同じです。なお、Holm 検定は、Bonferroni 検定よりも検出力が高くなります。

Tukey、Bonfferoni および Holm 検定を比較してみると下記の表のようになります。なお、エクセル関数で計算させる際、「t検定」の自由度を全群対象として 24 にしています（エクセル関数TDIST（t値、自由度“24”、両側検定“2”）。

水準番号	t値	t検定	Tukey	Bonferroni		Holm	
A1 vs A4	5.09	0.0000	0.0002	0.0001x6	0.0002	0.0001x6	0.0002
A1 vs A2	2.84	0.0090	0.0423	0.0090x6	0.0543	0.0090x5	0.0045
A3 vs A4	2.67	0.0134	0.0601	0.0134x6	0.0804	0.0134x4	0.0536
A1 vs A3	2.42	0.0235	0.1027	0.0235x6	0.1407		
A2 vs A4	2.25	0.0339	0.1319	0.0339x6	0.2033		
A2 vs A3	0.42	0.6782	0.6794	0.6782x6	1.0000		

本来ならば Bonferroni 検定は総当たりではなく、任意の群間を対比較します。Tukey と同じ総当たりをすることは検出力を下げることになります。

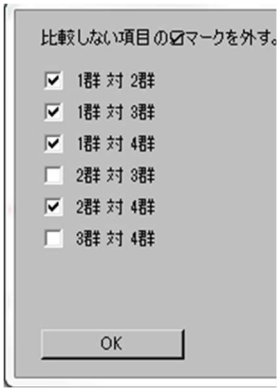
§3. Pharmaco Basic[3]で Bonferroni と Holm 検定をやってみる

Tukey 検定は総当たりで Bonferroni と Holm 検定は任意の群間比較といわれてもイメージがわきにくいと思われるので、敢えて Pharmaco ソフトをとりあげます。

今回の例題では、4群 A1，A2、A3，A4 の多重比較検定でした。Tukey では総当たりなので6回の対比較になります。一方、実施者は A1 との比較を全群で行い、さらに A2 群と A4 群のみの4回対比較で行いたいとします。Pharmaco では Bonferroni 検定と Holm 検定は、右図のように対比較を選定します。

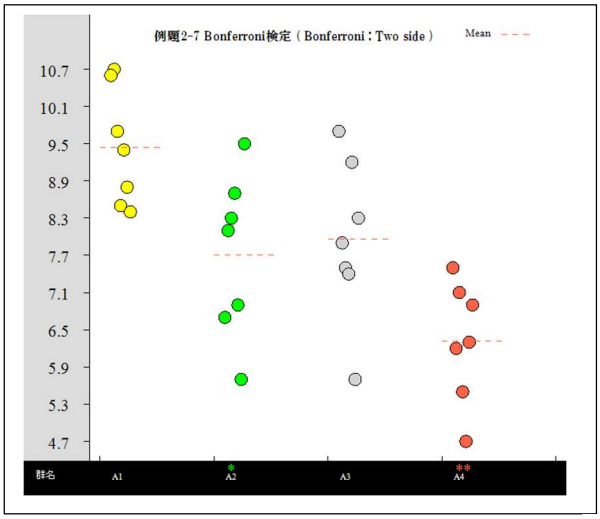
Pharmaco 入力形式

A1	A2	A3	A4
10.7	8.1	7.9	6.2
9.7	8.3	7.5	7.1
8.5	8.7	7.4	5.5
9.4	6.9	9.2	4.7
8.8	5.7	5.7	6.3
8.4	9.5	8.3	6.9
10.6	6.7	9.7	7.5



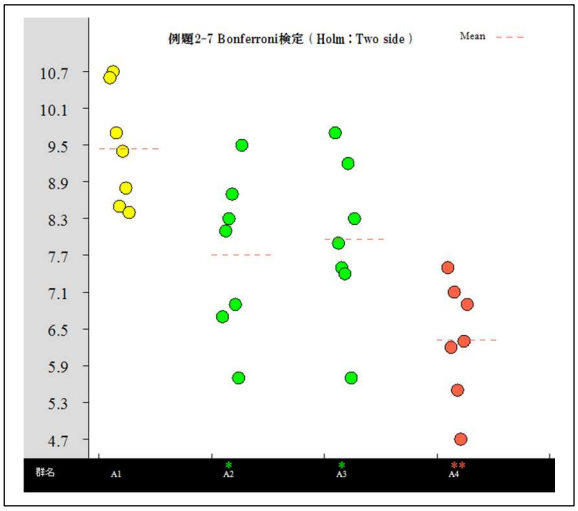
表は対照群を A1とした場合の結果です。対照群との比較は●で、他の選択群との比較は○です。

Bonferroni 検定



	統計量	P値		補正係数
● 1 x 2	2.8352	0.0366	*	K = 4
● 1 x 3	2.4169	0.0945		K = 4
● 1 x 4	5.0894	0.0001	**	K = 4
X 2 x 3	-----			
○ 2 x 4	2.2542	0.1343	-	K = 4
X 3 x 4	-----			

Holm 検定



	統計量	P値		補正係数
● 1 x 2	2.8352	0.0274	*	K = 3
● 1 x 3	2.4169	0.0472	*	K = 2
● 1 x 4	5.0894	0.0001	**	K = 4
X 2 x 3	-----			
○ 2 x 4	2.2542	0.0336	*	K = 1
X 3 x 4	-----			

Bonferroni 検定を総当たりした時の A1 vs A2 のp値は 0.0543でしたが、対比較を選択した場合は 0.0366 と有意差が付くようになりました。（※ボクのつぶやき：何で総当たりなの Bonferroni 検定をやるのかわからないなー。きっと、引用文献が Bonferroni 検定を使っていたからなんだろうなー）。

Holm 検定では Bonferroni 検定より，さらに p 値は小さくなり、有意差も２つ増えています。

参考文献

1）永田靖，吉田道弘著「統計的多重比較法の基礎」サイエンティスト社1997年

2）松本一彦、松田眞一「化学物質試験のための OECD 試験ガイドライン（魚類初期生活段階毒性試験）」
臨床評価 46(1)71-82，2016

3）Pharmaco 工房ホームページ https://pharmaco.club/