前回のQ＆A

Q1) テストの前に 1 回目の講義から特に重要なところをサラサラと教えていただけませんか？

A1) 要点のまとめが欲しいというわけですね。基本的には，プレゼンテーションしたところが重要だと考えてください。

Q2) 「5%水準で有意でない」ということは，実際にはどういうことなのですか？　例題の遊離残留塩素濃度を例にして解答していただければ幸いです。

A2) 「有意」という考え方は重要なので復習しておきます。例題は，東京の集合住宅群と一戸建て群の間で水道水の遊離残留塩素濃度に差があるかどうかを検定するということでした。どちらが高いとか低いとかいった事前情報はないので,「集合住宅群と一戸建て群の間で水道水の遊離残留塩素濃度に差はない」を帰無仮説として両側検定をします。「有意水準を 5%にする」とは,「帰無仮説が偶然に成り立つ確率が 5%未満であれば，統計的に意味があるほど稀な現象なので帰無仮説は成り立たないとみなす」ということですから,「5%水準で有意でない」といえば,「帰無仮説が偶然に成り立つ確率が 5 %未満であれば，統計的に意味があるほど稀な現象なので帰無仮説は成り立たないとみなす，としたのに，データから計算するとその確率が 5%より大きくなってしまったので，統計的に意味があるほど稀ではなく，帰無仮説が成り立たないとはみなせない」ということになります。この例でいえば，有意水準を 5%にしたのに,「集合住宅群と一戸建て群の間で水道水の遊離残留塩素濃度に差がない」条件下で，実際に得られているデータが偶然得られる確率は 5%より大きいので,「差がない」という帰無仮説が棄却されなかったということを意味します。

Q3) カイ二乗検定は母集団の分布を仮定していないからノンパラメトリックでは？

A3) 定義からいえばその通りです。すみません。

統計学第 10 回　多群間の差を調べる〜一元配置分散分析と多重比較

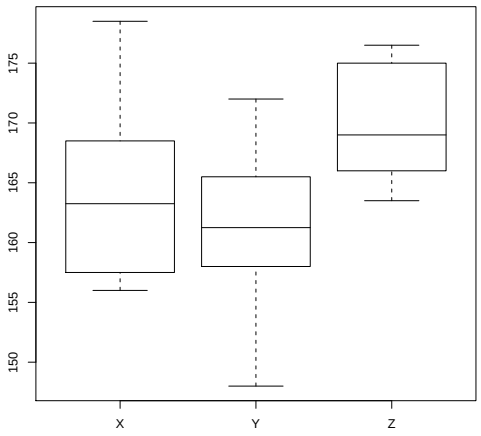
（1）多群間の比較を考える

- t 検定や順位和検定では 2 群間の差を比べた。では，3 群以上の場合はどうしたらいいだろうか？
- 単純に 2 群間の差の検定を繰り返してはいけない。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，大別して 2 つのアプローチがある。1 つは，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，というアプローチである。具体例でいえば，東京と長野と山口で年降水量の平均に差があるかどうかを見たいときに，東京と長野，長野と山口，という具合に比べるのではなくて，年降水量という変数に対して，地域という変数が有意な効果をもっているかどうか？　と立論するのである。このやり方に当たるのが一元配置分散分析やクラスカル＝ウォリス (Kruskal-Wallis) の検定（ノンパラメトリックな一元配置分散分析）である。
- もう 1 つのアプローチは，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は大きくなってしまう第 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%に抑えることである。このやり方は「多重比較法」と呼ばれる。さまざまな方法が提案されているが，中には数学的に不適切なものが歴史的に古くから使われているからというだけの理由で使われ続けている場合もあり，注意が必要である。
- これら 2 つのアプローチは別々に行うというよりも，段階を踏んで行うものと考えるのが一般的である。一元配置分散分析やクラスカル＝ウォリス (Kruskal-Wallis) の検定によって群間に何らかの差があると結論されてから，初めてどの群とどの群の差があるのかを調べるために多重比較法を使うというわけである。仮に多重比較法で有意な結果が出たとしても，一元配置分散分析の結果が有意でなければ，偶然のばらつきの効果が群間の差よりも大きいということなので，特定群間の差に意味があると考えることは解釈のし過ぎである（少なくともそのことに配慮した解釈を加えなくてはいけない）。ただし，永田・吉田「統計的多重比較法の基礎」（書誌情報は後述）が指摘するように，段階を踏んで実行すると，ここにまた検定の多重性の問題が生じるので，両方はやるべきではない，という考え方にも一理ある。つまり，厳密に考えれば，群分け変数が量的変数に与える効果があるかどうかを調べたいのか，群間で量的変数に差があるかどうかを調べたいのかによって，これら 2 つのアプローチを使い分けるべきだというのである。この点に関しては，多くの学術雑誌が現在でも「段階を踏め」式の指摘をしてくるので，思想の違いと考えるしかないし，どこかの群間にはっきりした違いがあれば，どちらの考え方をしても結果に違いは出てこないはずだから，当面は「段階を踏む」式の考え方をしておく方が無難であろう。

（2）一元配置分散分析

- 一元配置分散分析は，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ものである。
- 例えば，南太平洋のある島にある 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）。

ID 番号	村落 (“vg”)	身長 (cm)(“height”)
1	X	161.5
2	X	167.0
3	X	157.5
（中略）		
22	Z	166.0
（中略）		
36	Y	156.0
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 でデータを読み込んでから，`summary(aov(height ~ vg))` とすれば（`anova(lm(height ~ vg))` でも同等），例えば次のような結果が得られる。

```
Analysis of Variance Table

Response:  height
             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
```

- このような結果の表を分散分析表という。`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。`Residual` 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 0.006918 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。

- きちんと数式で説明すると，次のようになる。X 村の N_1 人の身長が X_{11}, X_{12}, ..., X_{1N_1}，Y 村の N_2 人の身長が X_{21}, X_{22}, ..., X_{2N_2}，Z 村の N_3 人の身長が X_{31}, X_{32}, ..., X_{3N_3} だとする（総人口 $N_1+N_2+N_3=N$ 人とする）。村毎の平均身長を $\bar{X_1}$, $\bar{X_2}$, $\bar{X_3}$ と書き，全体の平均を $\bar{X_T}$ と書くことにする。このとき，総変動（総平方和）S_T は，

$$S_T = \sum_{i=1}^{3}\sum_{j=1}^{N_i}(X_{ij}-\bar{X_T})^2$$

，級間変動（群間平方和）S_A は，

$$S_A = \sum_{i=1}^{3}\sum_{j=1}^{N_i}(\bar{X_i}-\bar{X_T})^2$$

，誤差変動（級内変動，群内平方和，または誤差平方和ともいう）S_E は，

$$S_E = \sum_{i=1}^{3}\sum_{j=1}^{N_i}(X_{ij}-\bar{X_i})^2$$

となる（簡単な式変形で，このとき $S_T = S_A + S_E$ であることがわかる）。自由度は，群の効果に関して $P_A = 3-1 = 2$ で，残差の効果に関して $P_E = N-3 = 34$ である。よって，級間分散 $V_A = S_A/P_A$，誤差分散 $V_E = S_E/P_E$ と推定でき，F 統計量 $F_0 = V_A/V_E$ が，第 1 自由度 P_A，第 2 自由度 P_E の F 分布に従うことを使って検定できる[1] 。つまり，繰り返しになるが，分散分析とは，全体のばらつき S_T を，群間の違いという意味のはっきりしているばらつき S_A と，それでは説明できないばらつき，つまり誤差である S_E に分けて比べることを意味するのである。
- 念のため上の数値例の値が数式のどれに当たるかをまとめておくと，P_A が 2，P_E が 34（N が 37），S_A が 422.72，S_E が 1243.80，V_A が 211.36，V_E が 36.58，F_0 が 5.7777，p が 0.006918 である。
- なお，この例のように，群分けをするカテゴリ変数が 1 つの場合を，一元配置分散分析 (ONE-WAY ANOVA)，2 つの場合を二元配置分散分析 (TWO-WAY ANOVA)，3 つなら三元配置分散分析 (THREE-WAY ANOVA) などと呼ぶ。二元配置以上の場合は，カテゴリ変数間での交互作用による影響を調べるための交互作用項がモデルに入ってくるし，その従属変数への効果を見るために母数モデルと変量モデルの違いを区別しなくてはならない。また，量的変数による交絡がある場合は共分散分析 (ANACOVA) をすることになる。[2]

（3）クラスカル＝ウォリス (Kruskal-Wallis) の検定

- 一元配置の分散分析は，各群が等しい母分散をもつ正規分布に従うことを仮定して行っているパラメトリックな分析法なので，データの分布がひどく歪んでいる場合は結果がおかしくなる場合がある[3] 。そこで，多群間の差を調べるためにもノンパラメトリックな方法がある。クラスカル＝ウォリス (Kruskal-Wallis) の検定と呼ばれる方法である。R では，kruskal.test(量的変数 ~ 群分け変数) で実行できる。以下，仕組みを説明する。
- 「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。

---

[1] R では，p=1−pf(F_0,P_A,P_E) として有意確率が得られる。

[2] これらの一部は第 12 回と第 13 回に説明する。

[3] 各群の母分散が等しいかどうかを調べる検定法として，バートレット (Bartlett) の検定と呼ばれる方法がある。R では bartlett.test(量的変数 ~ 群分け変数) で実行できる。同じ目的のノンパラメトリックな方法として，Fligner-Killeen の検定という方法もある。R では，fligner.test(量的変数 ~ 群分け変数) で実行できる。また，母集団が正規分布しているかどうかを調べる方法としては，既に説明したヒストグラムや正規確率プロットなどのグラフ表示による方法の他に，シャピロ＝ウィルク (Shapiro-Wilk) の検定と呼ばれる方法もある。詳しくは説明しないが，R では shapiro.test(量的変数) で実行できる。厳密に言えば，これらの検定で等分散性と分布の正規性が確認されない限り，一元配置分散分析の結果を解釈するには注意が必要なのだが，論文や本でもそこまで考慮されずに使われていることが多い。

・ 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。

・ なお，対応のある多群間の差をノンパラメトリックな方法で調べるには，フリードマン (Friedman) の検定と呼ばれる手法を用いる。R では，`friedman.test(`量的変数 ~ 群分け変数`)` で実行できる。簡単に説明すると，まず同じ個体について群間で順位をつける（群といっても，対応がある場合だから，例えば 2005 年の予測値と 2010 年の予測値と 2025 年の予測値というように，個々の個体について順位をつけることが可能である)。次に，群ごとにこの順位の合計（順位和）を計算する。順位和の二乗和から順位和の平均の二乗を引いた値を統計量 S として，サンプル数が少ない場合は表によって（コンピュータシミュレーションによってもよい）有意確率を計算し，サンプル数が多い場合は自由度が群数より 1 少ないカイ二乗分布に従う統計量 Q を S の 12 倍を個体数と群数と「群数＋1」の積で割った値として計算して有意確率を計算する。ただし同順位がある場合は調整が必要であり，煩雑なので，通常はコンピュータソフトウェアに計算させる。

（4）多重比較

・ 仮に，上述の南太平洋の島の 3 つの村での健診の例で，一元配置分散分析か Kruskal-Wallis の検定で有意差があったときに，具体的にどの村の間に有意差があるのかを調べるには，単純に考えると，t 検定[4] や順位和検定[5] を繰り返せば良さそうである。この方法が使われている本や論文もある。しかし，3 つの村でこれをやると 3 つから 2 つを取り出す全ての組み合わせについて検定するので，3 回の比較をすることになり，個々の検定について有意水準を 5%にすると，全体としての第 1 種の過誤は明らかに 5%より大きくなる。もし村が 7 つあったら，7 つから 2 つを取り出す組み合わせは 21 通りあるので，1 つくらいは偶然によって有意差が出てしまう比較があっても全然おかしくない。したがって，先に述べた通り，t 検定の繰り返しは第 1 種の過誤が大きくなってしまって不都合である。これに似た方法として無制約 LSD（最小有意差）法や Fisher の制約つき LSD 法（一元配置分散分析を行って有意だった場合にのみ LSD 法を行うという方法）があるが，これらも第 1 種の過誤を適切に調整できない（ただし制約つきの場合は 3 群なら大丈夫）ことがわかっているので，使ってはいけない。現在では，この問題は広く知られているので，t 検定の繰り返しや LSD 法で分析しても論文は accept されない。

・ 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，シェフェ(Scheffé) の方法，ダンカン (Duncan) 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。しかしこの中で，ダンカンの方法は，新多範囲検定などと呼ばれた時期もあったが，数学的に間違っていることがわかっているので，使ってはいけない。ボンフェローニの方法とシェフェの方法も検出力が悪いので，特別な場合を除いては使わない方が良い。せめてテューキーの HSD を使うべきである。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，適用場面が限定されている[6] 。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。

---

[4] R では `t.test(height[vg=="X"],height[vg=="Y"])` など。

[5] R では `wilcox.test(height[vg=="X"],height[vg=="Y"])` など。

[6] 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が。

- 上記いくつかの方法が良く使われている原因は，用途が限定されているダネットとウィリアムズを除けば，たんにそれらが歴史的に古く考案され，昔の統計学の教科書にも説明されているからに過ぎない。現在では，かなり広い用途をもち，ノンパラメトリックな分析にも適応可能なホルム (Holm) の方法（ボンフェローニの方法を改良して開発された方法）が第一に考慮されるべきである。その上で，全ての群間の比較をしたい場合はペリ (Peritz) の方法，対照群との比較をしたいならダネットの逐次棄却型検定（これはステップダウン法と呼ばれる方法の１つであり，既に触れたダネットの方法とは別）も考慮すればよい。とはいえ，ソフトウェアによってはこれらの方法をサポートしていない場合もあると思われ，その場合はテューキーの HSD を使うべきである（もちろん場合によっては，ダネットかウィリアムズを使い分けねばならない）。[7]
- 多重比較においては，帰無仮説が単純ではない。例えば，4 群間の差を調べるとしよう。一元配置分散分析での帰無仮説は，$\mu_1 = \mu_2 = \mu_3 = \mu_4$ である。これを包括的帰無仮説と呼び，$H_{\{1,2,3,4\}}$ と書くことにする。さて第 1 群から第 4 群までの母平均 $\mu_1 \sim \mu_4$ の間で等号関係が成り立つ場合をすべて書き上げてみると，$H_{\{1,2,3,4\}} : \mu_1 = \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2,3\}} : \mu_1 = \mu_2 = \mu_3$，$H_{\{1,2,4\}} : \mu_1 = \mu_2 = \mu_4$，$H_{\{1,3,4\}} : \mu_1 = \mu_3 = \mu_4$，$H_{\{2,3,4\}} : \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2\},\{3,4\}} : \mu_1 = \mu_2$ かつ $\mu_3 = \mu_4$，$H_{\{1,3\},\{2,4\}} : \mu_1 = \mu_3$ かつ $\mu_2 = \mu_4$，$H_{\{1,4\},\{2,3\}} : \mu_1 = \mu_4$ かつ $\mu_2 = \mu_3$，$H_{\{1,2\}} : \mu_1 = \mu_2$，$H_{\{1,3\}} : \mu_1 = \mu_3$，$H_{\{1,4\}} : \mu_1 = \mu_4$，$H_{\{2,3\}} : \mu_2 = \mu_3$，$H_{\{2,4\}} : \mu_2 = \mu_4$，$H_{\{3,4\}} : \mu_3 = \mu_4$ の 14 通りである。このうち，$H_{\{1,2,3,4\}}$ 以外のものを部分帰無仮説と呼ぶ。すべての２つの群の組み合わせについて差を調べるということは，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}, H_{\{2,3\}}, H_{\{2,4\}}, H_{\{3,4\}}\}$ が，考慮すべき部分帰無仮説の集合となる。一方，例えば第 1 群が対照群であって，他の群のそれぞれが第 1 群と差があるかどうかを調べたい場合は，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}$ が考慮すべき帰無仮説の集合となる。これらの集合をその多重比較における「帰無仮説族」と呼ぶ。
- ここで多重比較の目的を「帰無仮説族」というコトバを使って言い換えてみる。個々の帰無仮説で有意水準を 5%にしてしまうと，帰無仮説族に含まれる帰無仮説のどれか１つが誤って棄却されてしまう確率が 5%より大きくなってしまう。それではまずいので，その確率が 5%以下になるようにするために，何らかの調整を必要とするわけで，この調整をする方法が多重比較なのである。つまり，帰無仮説族の有意水準を定める（例えば 5%にする）ことが，多重比較の目的である。[8]
- R では，`pairwise.t.test(height,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 t 検定した結果を出してくれる[9]。`pairwise.wilcox.test(height,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。ボンフェローニが可能になっているのは，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，R を使って分析するのだけれども，データがきれいな正規分布をしていて，かつ古典的な方法の論文しか accept しない雑誌に対してどうしても投稿したい，という場合は，`TukeyHSD(aov(height ~ vg))` などとして，テューキーの HSD を行うことも可能だし，CRAN (http://cran.r-project.org/) から `multcomp` パッケージをインストールすることによって，`simtest(height ~ vg, type="Dunnett")` あるいは `simtest(height ~ vg, type="Williams")` としてダネットやウィリアムズの方法を使うことも可能である。
- これらの方法の中身に立ち入って説明しつくすことは不可能なので，ここではボンフェローニとホルム，テューキーの HSD だけを簡単に説明する。より詳しく知りたい場合には，永田靖・吉田道弘 (1997)「統計的多重比較法の基礎」（サイエンティスト社）を参照されたいが，この本は「基礎」とはいうものの，経験を積んだ研究者を対象として書かれており，学部学生が読むにはかなり難しい。

(4-1) ボンフェローニの方法とホルムの方法

- ボンフェローニの方法とは，ボンフェローニの不等式に基づく多重比較法である。きわめて単純な考え方に基づいているために，適用可能な範囲が広い。しかし，検出力が落ちてしまいがちなので，ベストな方法ではない。

---

[7] もっとも，オープンソースで多くのコンピュータで無料で使える R がホルムの方法をデフォルトとしているのに，そういう言い訳は本来通用しないと思われるが。

[8] このことからわかるように，差のなさそうな群をわざと入れておいて帰無仮説族を棄却されにくくしたり，事後的に帰無仮説を追加したりすることは，統計を悪用していることになり，やってはいけない。

[9] ただし，t 検定とは言っても，`pool.sd=F` というオプションをつけない限りは，t_0 を計算するときに全体の誤差分散を使うので，ただの t 検定の繰り返しとは違う。

- ボンフェローニの不等式とは，k 個の事象 E_i $(i = 1, 2, ..., k)$ に対して成り立つ，

$$Pr(\cup_{i=1}^{k} E_i) \leq \sum_{i=1}^{k} Pr(E_i)$$

をいう。左辺は k 個の事象 E_i のうち少なくとも 1 つが成り立つ確率を示し，右辺は各事象 E_i が成り立つ確率を加え合わせたものなので，この式が成り立つことは自明であろう（個々の事象がすべて独立な場合にのみ等号が成立する）。
- 次に，この不等式を多重比較にどうやって応用するかを示す。まず，帰無仮説族を $\{H_{01}, H_{02}, ..., H_{0k}\}$ とする。E_i を「正しい帰無仮説 H_{0i} が誤って棄却される事象」と考える。この表現をボンフェローニの不等式にあてはめれば，

$$Pr(\text{正しい帰無仮説のうちの少なくとも 1 つの } H_{0i} \text{ が誤って棄却される})$$

$$\leq \sum_{i=1}^{k} Pr(\text{正しい帰無仮説 } H_{0i} \text{ が誤って棄却される})$$

この 2 行目が α 以下になるためには，もっとも単純に考えれば，足しあわされる各項が α/k に等しいかより小さければよい。つまり，ボンフェローニの方法とは，有意水準 α で帰無仮説族を検定するために，個々の帰無仮説の有意水準を α/k にするものである。[10]
- 手順としてまとめると，以下の通り。
  1) 帰無仮説族を明示し，そこに含まれる帰無仮説の個数 k を求める。
  2) 帰無仮説族についての有意水準 α を定める。$\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。
  3) 帰無仮説族に含まれているそれぞれの帰無仮説に対して検定統計量 T_i $(i = 1, 2, ..., k)$ を選定する。
  4) データを取り，検定統計量 T_i を計算する。
  5) 各検定統計量 T_i について有意水準 α/k に対応する棄却限界値（通常は分布関数の $(1 - \alpha/k) \times 100\%$点）を c_i とするとき，$T_i \geq c_i$ ならば H_{0i} を棄却し，$T_i < c_i$ なら H_{0i} を保留する（採択ではない）。
- なお，R では，各々の帰無仮説の有意水準を α/k とする代わりに，各々の帰無仮説に対して得られる有意確率が k 倍されて（ただし 1 を超えるときは 1 として）表示されるので，各々の比較に対して表示される有意確率と帰無仮説族について設定したい有意水準との大小によって仮説の棄却／保留を判断してよい。
- ボンフェローニの方法では，すべての H_{0i} について有意水準を α/k としたのが良くなかったので，ホルムの方法は，そこを改良したものである。以下，ホルムの方法の手順をまとめる。
  1) 帰無仮説族を明示し，そこに含まれる帰無仮説の個数 k を求める。
  2) 帰無仮説族についての有意水準 α を定める。$\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。ここまではボンフェローニと同じ。
  3) $\alpha_1 = \alpha/k$，$\alpha_2 = \alpha/(k-1)$，...，$\alpha_k = \alpha$ を計算する。
  4) 帰無仮説族に含まれているそれぞれの帰無仮説に対して検定統計量 T_i $(i = 1, 2, ..., k)$ を選定する。
  5) データを取り，検定統計量 T_i を計算する。
  6) 各検定統計量 T_i について有意確率 P_i を求め，小さい順に並べ換える。
  7) P_i の小さいほうから順に α_i と P_i の大小を比べる。
  8) $P_i > \alpha_i$ ならばそれよりも有意確率が大きい場合の帰無仮説をすべて保留して終了する。$P_i \leq \alpha_i$ なら H_{0i} を棄却して，次に小さい P_i について比較する。$i = k$ となるまで繰り返す。
- ホルムの方法についても，R では，7) で P_i と α_i の大小を比べる代わりに $P_i' = P_i \times (k - i + 1)$ が表示されるので，値そのものを有意水準と比較すればよい。ただし，8) からすると，P_i' が有意でなかったら，P_{i+1}' が有意水準より小さくてもその仮説は保留されるべきなのだが，その点がどう表示されるのかは未確認である。

[10] ここで注意しなければいけないことは，検定すべき帰無仮説族に含まれる個々の帰無仮説は，データをとるまえに定められていなければいけないことである。データをとった後で有意になりそうな帰無仮説を k 個とってきて帰無仮説族を構成するのでは，帰無仮説族に対しての第 1 種の過誤をコントロールできないのでダメである。

・ 南太平洋の３つの村の問題に戻って，ボンフェローニの方法とホルムの方法で検定した結果は，次のようになる。

誤差分散を使った t 検定，Bonferroni で調整

	X	Y
Y	0.8841	−
Z	0.1283	0.0052

t 検定の繰り返し，Bonferroni で調整

	X	Y
Y	1.0000	-
Z	0.1422	0.0026

誤差分散を使った t 検定，Holm で調整

	X	Y
Y	0.2947	-
Z	0.0855	0.0052

t 検定の繰り返し，Holm で調整

	X	Y
Y	0.3475	-
Z	0.0948	0.0026

順位和検定の繰り返し，Bonferroni で調整

	X	Y
Y	1.0000	-
Z	0.2162	0.0078

順位和検定の繰り返し，Holm で調整

	X	Y
Y	0.4865	-
Z	0.1441	0.0078

（4-2）テューキーの HSD

・ テューキーの HSD では，母集団の分布は正規分布とし，すべての群を通して母分散は等しいと仮定する。

・ データが第 1 群から第 a 群まであって，各々が n_i 個（$i = 1, 2, ..., a$）のデータからなるものとする。第 i 群の j 番目のデータを x_{ij} と書くことにすると，第 i 群の平均 $\bar{x_i}$ と分散 V_i は，

$$\bar{x_i} = \sum_{j=1}^{n_i} x_{ij}/n_i$$

$$V_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2/(n_i - 1)$$

となり，誤差自由度 P_E と誤差分散 V_E は，

$$P_E = N - a = n_1 + n_2 + ... + n_a - a$$

$$V_E = \sum_{i=1}^{a}\sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2/P_E = \sum_{i=1}^{a} (n_i - 1)V_i/P_E$$

で得られる。

・ 簡単にいえば，テューキーの HSD は，すべての群間の比較について誤差分散を使った t_0 統計量を計算し，t 分布ではなくて，ステューデント化された範囲の分布(Studentized range distribution) と呼ばれる分布の $(1-\alpha) \times 100\%$点を $\sqrt{2}$ で割った値との大小で有意水準 α の検定をする方法である。以下手順としてまとめる。

1) 帰無仮説族を明示する。テューキーの HSD の場合は，通常，

$$\{H_{\{1,2\}}, H_{\{1,3\}}, ..., H_{\{1,a\}}, H_{\{2,3\}}, ..., H_{\{a-1,a\}}\}$$

2) 有意水準 α を定める。$\alpha = 0.05$ または $\alpha = 0.01$ と定めることが多い。

3) データを取り，すべての群について $\bar{x_i}, V_i$ を計算し，P_E, V_E を計算する。

4) すべての 2 群間の組み合わせについて，検定統計量 t_{ij} を

$$t_{ij} = (\bar{x_i} - \bar{x_j})/\sqrt{V_E(1/n_i + 1/n_j)}$$

により計算する（$i, j = 1, 2, ..., a; i < j$）。

5) $|t_{ij}| \geq q(a, P_E; \alpha)/\sqrt{2}$ なら $H_{\{i,j\}}$ を棄却し，i 群と j 群の平均値には差があると判断する（比較の形からわかるように，これは両側検定である）。$|t_{ij}| < q(a, P_E; \alpha)/\sqrt{2}$ なら $H_{\{i,j\}}$ を保留する。ここで $q(a, P_E; \alpha)$ は，群数 a，自由度 P_E のステューデント化された範囲の分布の $(1-\alpha)\times 100\%$点である。つまり，$\alpha = 0.05$ ならば，$q(a, P_E, 0.05)$ は，群数 a，自由度 P_E のステューデント化された範囲の分布の 95%点である。R では，この値を与える関数は，`qtukey(0.95,`a`,`P_E`)` である。が，すべての群間比較を手計算するよりも，パッケージに計算させるのが普通である。

- 上述の例題に対する R の `TukeyHSD(aov(height ~ vg))` の出力は以下の通り。95%同時信頼区間が 0 を含まない Y 村と Z 村の身長だけが，5%水準で有意に異なる（Z−Y が正なので，Z 村の平均身長の方が Y 村の平均身長より有意に高い）と読める。

```
Tukey multiple comparisons of means
95% family-wise confidence level

Fit:  aov(formula = height ~ vg)

             diff              lwr            upr
Y-X     −2.538889       −8.3843982       3.306620
Z-X      5.850000       −0.9598123      12.659812
Z-Y      8.388889        2.3382119      14.439566
```