多重比較とは何ぞや？

寺子屋･統計庵　其之四

杉本解析差有比数
杉本典夫

多重比較とは何ぞや？

- 多重比較は複雑怪奇!?
- 推定の原理
- 統計的仮説検定の原理
- 分散分析の原理
- 多重比較と同時信頼区間の原理
- 多重比較の種類
- 多重比較が必要か？–各種の実例

多重比較は複雑怪奇!?

検定を沢山行うと
多重比較を使えと
文句を言われる！

多重比較を使うと
有意になりにくい！

色々な多重比較法があって
どれを使えば良いのか
わからない！

まずは基本に戻り
推定と検定の原理をおさらいしましょう！

多重比較とは何ぞや？

- 多重比較は複雑怪奇!?
- 推定の原理
- 統計的仮説検定の原理
- 分散分析の原理
- 多重比較と同時信頼区間の原理
- 多重比較の種類
- 多重比較が必要か？–各種の実例

推定と検定

統計学的推論

推定
estimate

定量試験

母数を推測
↓
pH計に相当

検定
test

定性試験

母数が基準値と
等しいかどうかを
◯x式で推測
↓
リトマス試験紙に相当

検定よりも推定の方が重要

ところが研究現場や厚労省では検定が偏重されている
↑
◯x式の方が採点が楽！

推定と検定の基本原理−中心極限定理

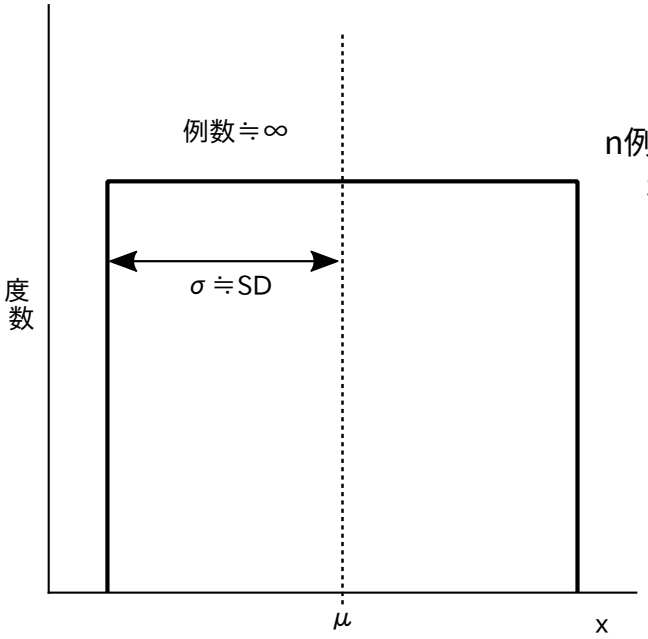


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

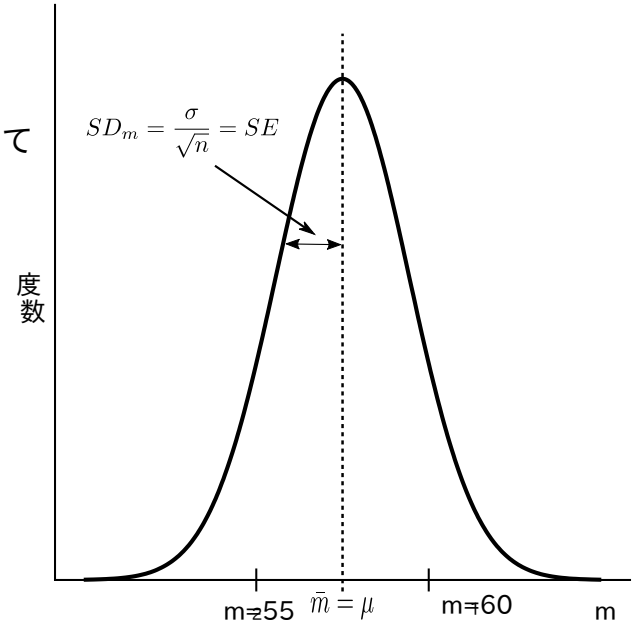


図1.4 標本平均値の分布

標本平均の分布の特徴

1. 母集団がどんな分布をしていても、漸近的に(nが多いほど)正規分布に近似する
   →中心極限定理(推測統計学の基本定理)
2. 標本平均mの平均値$\bar{m}$は母平均μと一致する
3. 標本平均の標準偏差SD_mは次のような値になる→標準誤差SE

$$SD_m = \frac{\sigma}{\sqrt{n}} \fallingdotseq \frac{SD}{\sqrt{n}} = SE$$

σ：母標準偏差　　n：標本集団の例数

SD：標本集団から求めたσの推測値

推定の原理

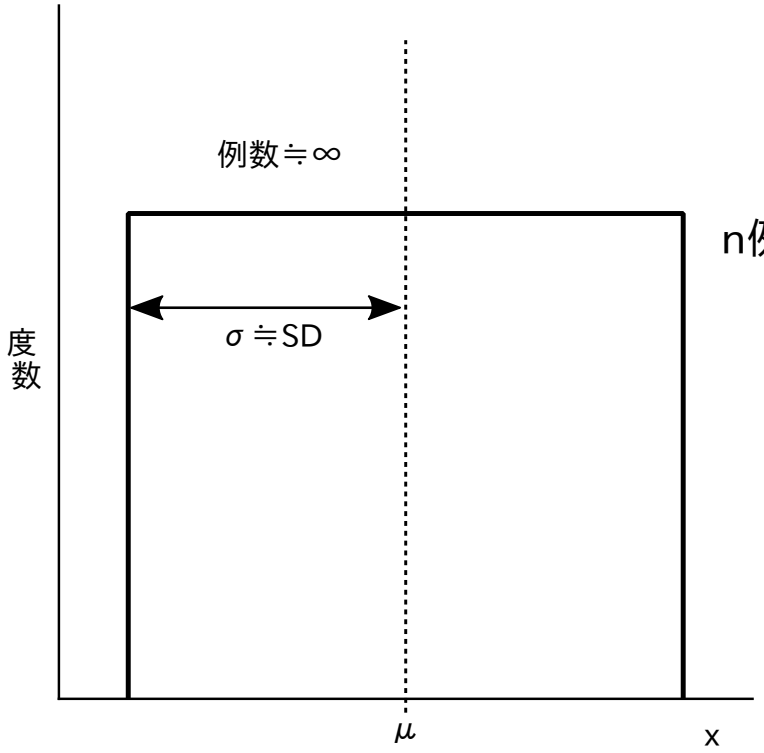


図1.3 母集団のデータ分布

n例を無作為抽出して
標本平均値mを
無限回求める

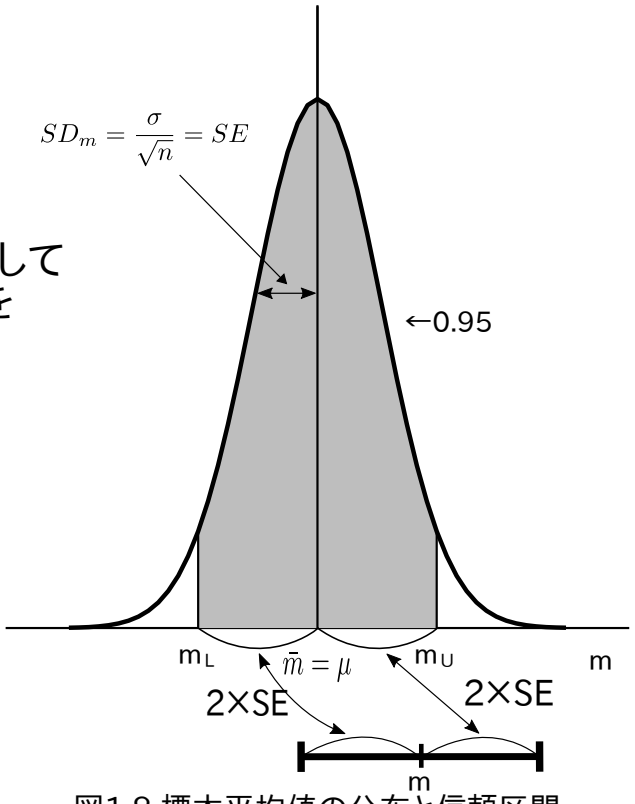


図1.8 標本平均値の分布と信頼区間

点推定(point estimation)：母数(母平均)をピンポイント(標本平均)で推測

区間推定(interval estimation)：母数をある程度の幅(信頼区間)を持たせて推測

母平均の区間推定法

標本平均mの分布は近似的に正規分布になる

↓

mの平均値は母平均μになり、mの標準偏差は標準誤差SEになる
μ±2×SEの範囲に約95%のmが含まれる

↓

ある標本平均mがμ±2×SEの範囲に含まれる確率は約95%

↓

逆にm±2×SEの範囲にμが含まれる確率も約95%

↓

95%信頼区間：$\mu \fallingdotseq m \pm 2 \times SE \rightarrow \mu_L = m - 2 \times SE \quad \mu_U = m + 2 \times SE$

95%信頼区間(95%CI)：母平均が95%の確率で含まれる区間、95%信頼限界(95%CL)

μ_L：信頼区間下限　　μ_U：信頼区間上限　　95%：信頼係数

点推定と区間推定

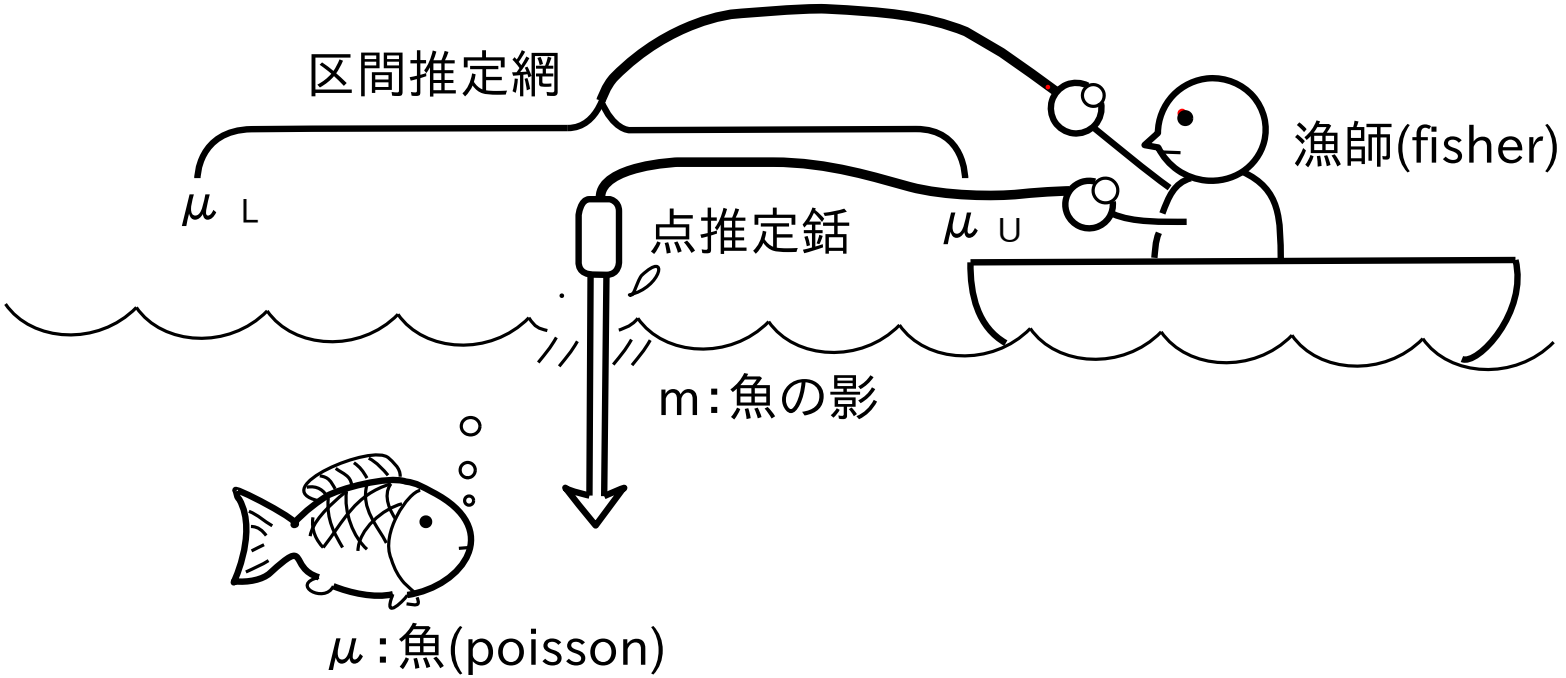


図1.9 点推定と区間推定

推定は漁師(Fisher)が
水面に映った魚(Poissson)の影mを見て魚μを捕まえるようなもの
点推定は銛で一突き、区間推定は投網を打つことに相当
普通は点推定を用い、重要な時だけ区間推定を行う

多重比較とは何ぞや？

- 多重比較は複雑怪奇!?
- 推定の原理
- 統計的仮説検定の原理
- 分散分析の原理
- 多重比較と同時信頼区間の原理
- 多重比較の種類
- 多重比較が必要か？–各種の実例

統計的仮説検定の原理

問題：日本人の平均体重は50kgか？

帰無仮説H_0:日本人の平均体重は50kgである←問題の答えは○

対立仮説H_1:日本人の平均体重は45kgまたは55kgである←問題の答えは×

統計的仮説検定の手順

1. 問題を設定する→医学的意義のある基準値μ_0=50と許容範囲(検出差)δ^*=±5を決める
2. 問題の答を○×式で設定する→帰無仮説H_0と具体的な対立仮説H_1を設定する
3. データに基づいて仮説の妥当性を判定する

検定は定性試験だから定量試験である推定結果から判定可能

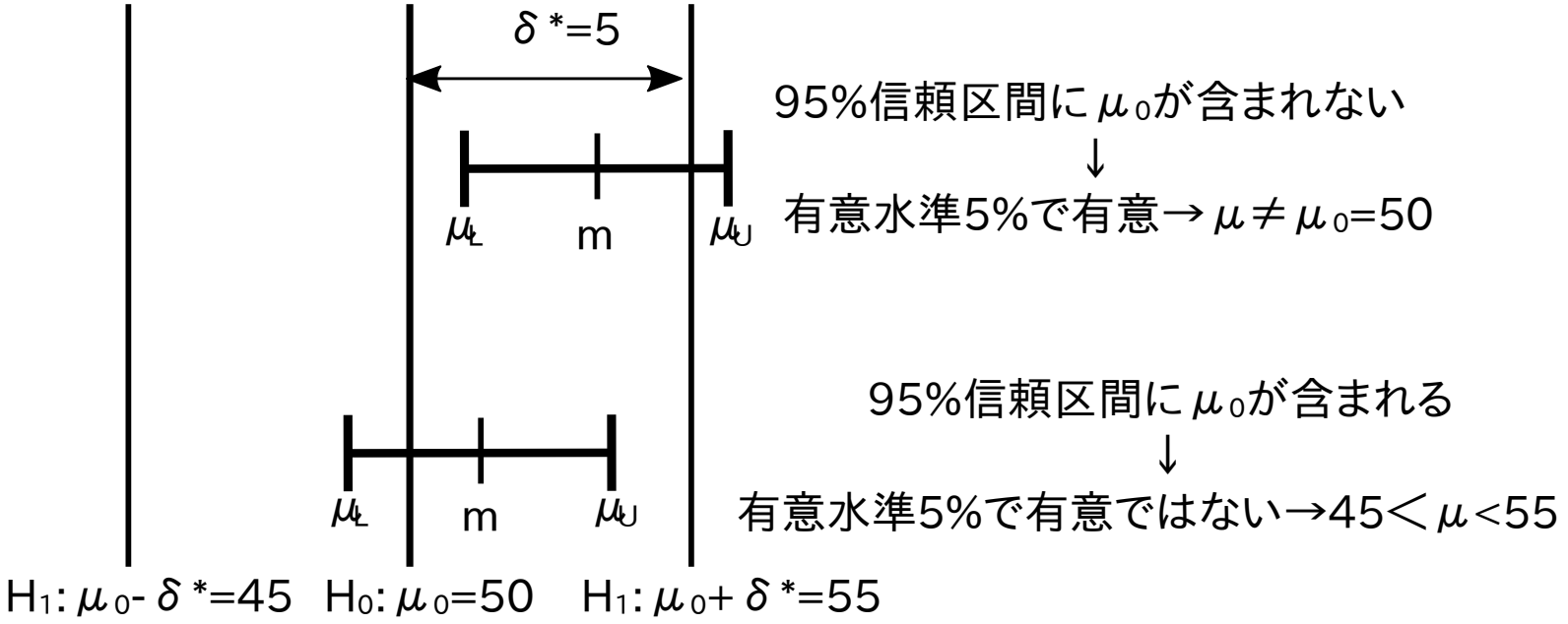


図1.13 信頼区間と統計的仮説検定

標本集団：n=100例　標本平均m=60kg　標準偏差SD=10kg　標準誤差SE=1kg

95%信頼区間: $\mu = 60 \pm 2 \times \frac{10}{\sqrt{100}} = 60 \pm 2 \rightarrow \mu = 58 \sim 62$

母平均は95%の確率で58～62kgの間にある

→母平均は95%以上の確率で50kgではない

統計学的結論

95%信頼区間に基準値が入っていない時
統計学的結論:日本人の平均体重は50kgではない←問題の答えはx

「有意水準5%で有意」または「危険率(αエラー)5%で有意」と表現する
これは「日本人の平均体重は45kgまたは55kg」の採用ではないことに注意！

95%信頼区間に基準値が入っている時
かつ信頼区間が許容範囲内に収まっている時
統計学的結論:日本人の平均体重は45kgよりも重く55kgよりも軽い←問題の答えは△

「有意水準5%で有意ではない」または「危険率5%で有意ではない」と表現する
これは「日本人の平均体重は50kgである」の採用ではないが、実質的には同じ意味
※この結論が間違っている確率がβエラー(普通はβ=0.2)←α=βにするのが理想

検定結果だけから実質科学的な判断をするのは危険

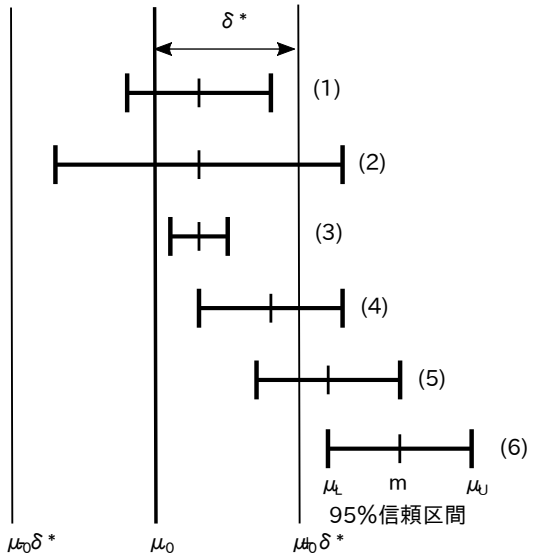


図1.7.1 検定結果と信頼区間

	検定結果	推定結果	実質科学的な判断
(1)	有意ではない	$\mu_0-\delta^*<\mu<\mu_0+\delta^*$	母平均は基準値とほぼ等しい
(2)	有意ではない	$\mu \fallingdotseq \mu_0 \sim \mu_0+\delta^*$	この結果だけでは判断できない 検出力をもっと高くする必要がある(例数を増やす)
(3)	有意	$\mu_0<\mu<\mu_0+\delta^*$	母平均は基準値と実質的に変わらない
(4)	有意	$\mu \fallingdotseq \mu_0+\delta^*$	母平均は基準値と実質的に変わらない可能性が高い
(5)	有意	$\mu \fallingdotseq \mu_0+\delta^*$	母平均は基準値よりも大きい可能性が高い
(6)	有意	$\mu_0+\delta^*<\mu$	母平均は基準値よりも大きい

※生物学的同等性試験では推定結果を重視し、検定結果は参考程度→検定廃止論

例数が多いと検定結果は科学的な判断には使えない

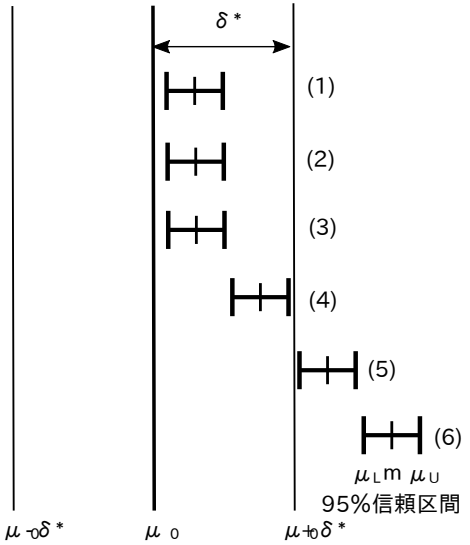


図1.7.6 例数が多い時の検定結果と信頼区間

	検定結果	推定結果	実質科学的な判断
(1)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(2)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(3)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(4)	有意	$\mu_0 < \mu < \mu_0 + \delta^*$	母平均値は基準値と実質的に変わらない
(5)	有意	$\mu_0 + \delta^* < \mu$	母平均値は基準値よりも大きい
(6)	有意	$\mu_0 + \delta^* < \mu$	母平均値は基準値よりも大きい

※生物学的同等性試験では推定結果を重視し、検定結果は参考程度 → 検定廃止論

有意確率p<0.001になっても結果の信頼性は95%

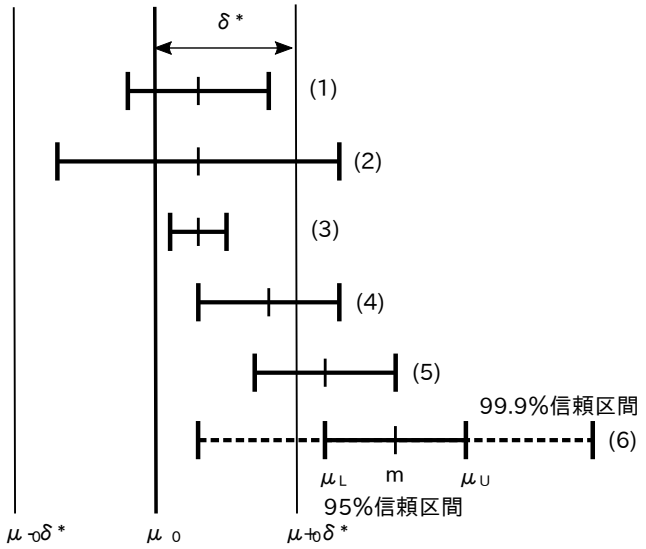


図1.7.1 検定結果と信頼区間

(1)と(2):p>0.05　(3)と(4)：p<0.05　(5)：p<0.01　(6)：p<0.001になった時

(5)を「有意水準1%で有意」と表現すると99%信頼区間が対応

(6)を「有意水準0.1%で有意」と表現すると99.9%信頼区間が対応

→信頼区間の幅が広くなって($\mu_0+\delta^*$)が入ってしまい、結論が曖昧になる

そのため通常は95%信頼区間を用いる→結果の信頼性は95%

※必要例数を計算した時のαエラーとβエラーによって結果の信頼性が決まる

多重比較とは何ぞや？

- 多重比較は複雑怪奇!?
- 推定の原理
- 統計的仮説検定の原理
- 分散分析の原理
- 多重比較と同時信頼区間の原理
- 多重比較の種類
- 多重比較が必要か？–各種の実例

多群比較のデータ

＜3群の薬剤投与後の収縮期血圧＞

群内No.	A剤投与群	B剤投与群	C剤投与群
1	116	106	108
2	128	102	100
3	129	108	108
4	137	118	114
5	140	116	110

高血圧患者15例を無作為に5例ずつの3群に分け
各群に薬剤A、B、Cを投与して投与後の収縮期血圧を測定した

多群のデータと平均値

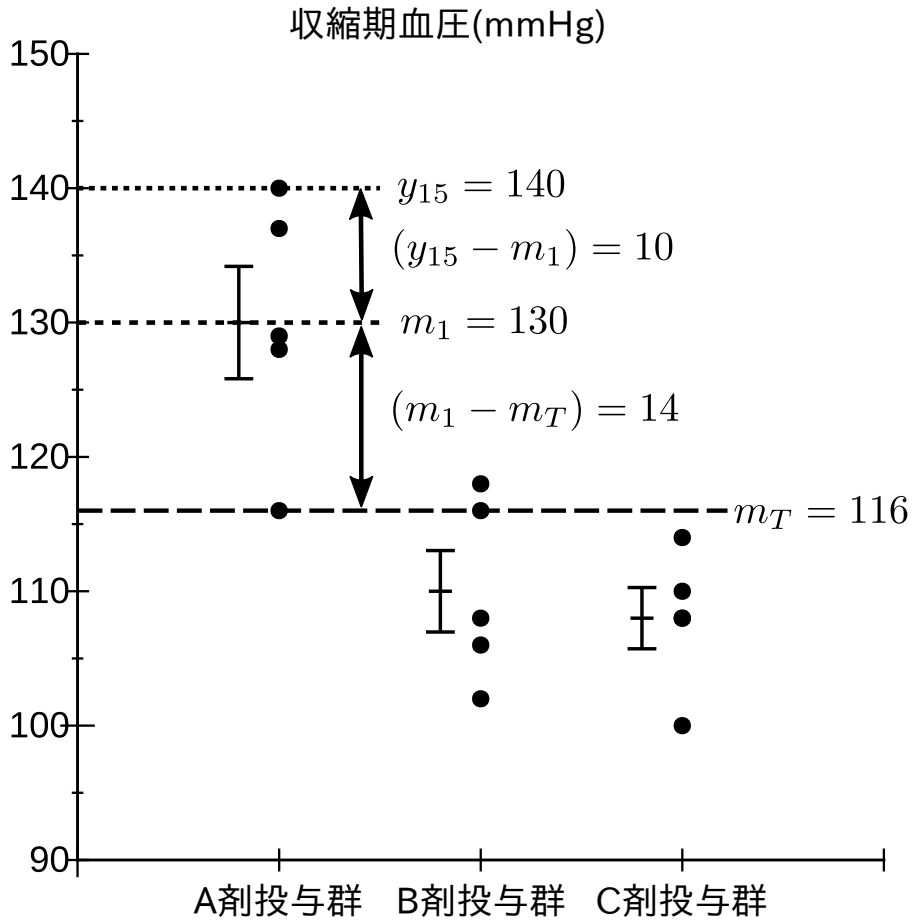


図4.18 一元配置分散分析の模式図

高血圧患者15例を無作為に5例ずつの3群に分け
各群に薬剤A、B、Cを投与して投与後の収縮期血圧を測定した

一元配置分散分析の問題と仮説設定

薬剤の種類によって降圧効果が違うか？

仮説設定に必要な事項

- 評価指標：収縮期血圧の実測値平均→一元配置分散分析を適用
- 基準値：3群の実測値平均のバラツキ(分散)＝0
- 検出差(医学的な許容範囲)：具体的な値を設定するのは困難

→具体的な対立仮設ではなく帰無仮説を否定した対立仮説を設定(有意性検定)

帰無仮説　H_0：3群の収縮期血圧の平均値は全て等しい

対立仮説　H_1：3群の収縮期血圧の平均値はばらついている

一元配置分散分析の解析結果

```
=== 多群の平均値の比較 ===                                          [DANS V7.3]

群項目(要因A):群 (1:A剤投与群 2:B剤投与群 3:C剤投与群)
集計項目       :収縮期血圧 (mmHg)

群     :群別基礎統計量
-------------------------------------------------------------------------------
1      :例数=5      平均値=130          標準偏差=9.35414      標準誤差=4.1833
2      :例数=5      平均値=110          標準偏差=6.78233      標準誤差=3.03315
3      :例数=5      平均値=108          標準偏差=5.09902      標準誤差=2.28035
-------------------------------------------------------------------------------
全体   :例数=15     平均値=116          標準偏差=12.2998      標準誤差=3.1758

・一元配置分散分析(one-way layout analysis of variance)
                          分散分析表(ANOVA table)
要因              平方和     自由度     平均平方和          F値    有意確率p値
-------------------------------------------------------------------------------
群(要因A)          1480          2           740      13.9185    0.000747082***
残差                638         12       53.1667
-------------------------------------------------------------------------------
全体               2118         14
```

寄与率：$R_A^2 = \eta_A^2 = \frac{\text{要因Aの平方和}}{\text{全体の平方和}} = \frac{1480}{2118} = 0.699\,(69.9\,\%)$　η_A=0.836：相関比(相関係数に相当)

一元配置分散分析の統計学的結論と医学的結論

統計学的結論

3群の平均値はばらついている

※平均値の分散を区間推定することは可能だが、解釈が難しいので普通は行わない

統計学的結論から医学的結論を導くための検討事項

1. 約70%という寄与率は医学的に見て意義があるか？→一般には寄与率が50%以上なら意義がある
2. 130、110、108という平均値のバラツキは医学的に見て意義があるか？→医学的許容範囲と比較
3. これらの平均値のバラツキは薬剤の違いによるものか？→3群の背景因子がほぼ均等であることが必要
4. この結果をそのまま高血圧患者全体に当てはめて良いか？→通常は非無作為抽出なので当てはめられない

医学的結論

薬剤の種類によって降圧効果が異なる

→薬剤AよりもBとCの方が降圧効果は大きい

多重比較とは何ぞや？

- 多重比較は複雑怪奇!?
- 推定の原理
- 統計的仮説検定の原理
- 分散分析の原理
- 多重比較と同時信頼区間の原理
- 多重比較の種類
- 多重比較が必要か？–各種の実例

多群比較のデータ

＜3群の薬剤投与後の収縮期血圧＞

群内No.	A剤投与群	B剤投与群	C剤投与群
1	116	106	108
2	128	102	100
3	129	108	108
4	137	118	114
5	140	116	110

高血圧患者15例を無作為に5例ずつの3群に分け
各群に薬剤A、B、Cを投与して投与後の収縮期血圧を測定した

多群のデータと平均値

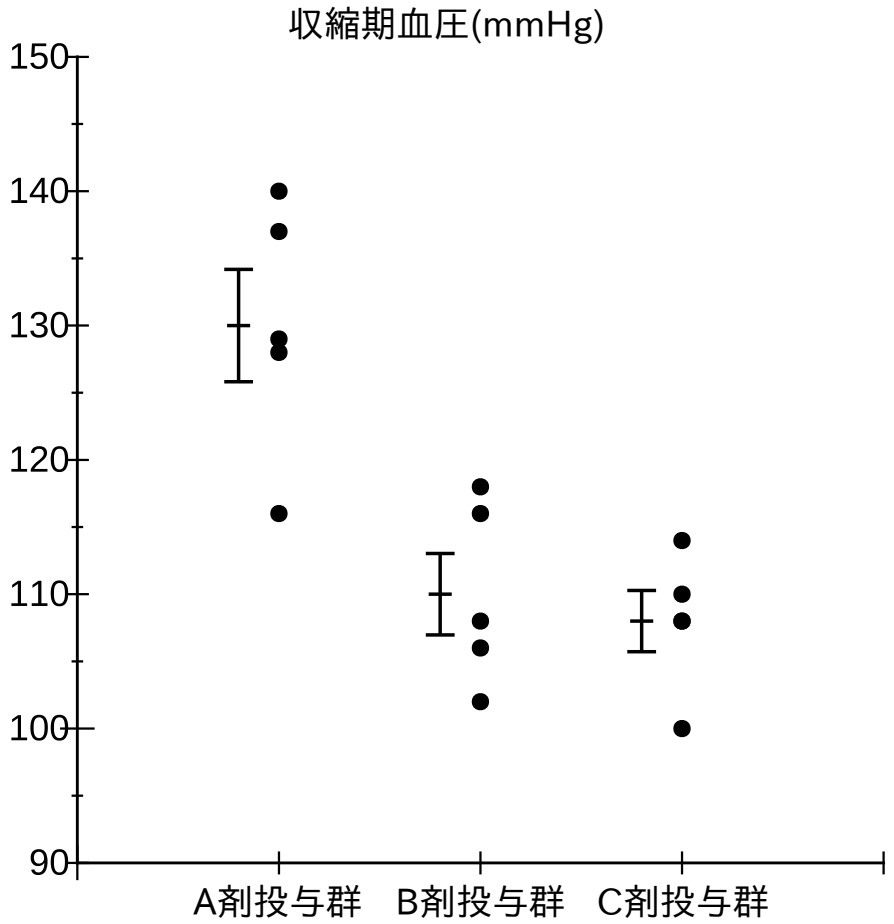


図4.19 多重比較の模式図

高血圧患者15例を無作為に5例ずつの3群に分け
各群に薬剤A、B、Cを投与して投与後の収縮期血圧を測定した

多重比較の問題

薬剤A、B、Cの降圧効果に違いがあるか？
もしあるとすればそれはどの薬剤とどの薬剤の間か？

↓

降圧効果を2薬剤ごとに3回比較し、どれか1回でも降圧効果に違いがあれば
それを採用して「降圧効果に違いがある」と結論するいいとこ取りの評価方法
分散分析とは独立した別目的の手法

↑

「有意水準5%で有意」とは「統計学的結論が間違っている危険性が5%ある」という意味
→20回比較すれば降圧効果が全て同じでも1回くらいは「有意水準5%で有意」になる
→比較回数が多いほどいいとこ取りした結論が間違っている危険性が増える
→1回ごとの有意水準にハンディキャップを持たせる必要がある

ファミリーとしての結論とファミリーとしての有意水準

多重比較の帰無仮説　H_0：3群の収縮期血圧の平均値は全て等しい

- 帰無仮説が正しい時に

  AとB、AとC、BとCの検定結果が有意水準5%で全て有意にならない確率

  ＝「3群の母平均は医学的にほぼ等しい」と結論する確率＝0.95×0.95×0.95≒0.86

- 帰無仮説が正しい時に

  「どれか1つ以上の母平均が他と異なっている」と結論する確率

  ＝1-0.86＝0.14≒0.15＝ファミリーとしての有意水準α_F←これを0.05にする必要がある

- ボンフェローニの不等式：$\alpha_F \leqq \alpha_A + \alpha_B + \alpha_C$　(ただしα_A、α_B、α_Cが互いに独立の時)

  →1回ごとの有意水準を$\alpha_F/3=0.05/3=\alpha_A=\alpha_B=\alpha_C$にする

  →1回ごとの検定は有意になりにくいが、ファミリーとして有意になる確率は5%

多重比較の仮説設定

仮説設定に必要な事項

- 評価指標：収縮期血圧の実測値平均→Tukey型多重比較を適用
- 基準値：A剤投与群の平均値またはB剤投与群の平均値
- 検出差(医学的な許容範囲)：±10

※先行研究・探索型試験等の結果に基づいて決定

帰無仮説　H_0：3群の収縮期血圧の平均値は全て等しい

対立仮説　H_1：A剤投与群とB剤投与群の平均値の差は10である
または
対立仮説　H_1：A剤投与群とC剤投与群の平均値の差は10である
または
対立仮説　H_1：B剤投与群とC剤投与群の平均値の差は10である

多重比較の解析結果

```
=== 多群の平均値の比較 ===                                          [DANS V7.3]

群項目(要因A):群 (1:A剤投与群 2:B剤投与群 3:C剤投与群)
集計項目       :収縮期血圧 (mmHg)
群     :群別基礎統計量
-------------------------------------------------------------------------------
1      :例数=5      平均値=130          標準偏差=9.35414      標準誤差=4.1833
2      :例数=5      平均値=110          標準偏差=6.78233      標準誤差=3.03315
3      :例数=5      平均値=108          標準偏差=5.09902      標準誤差=2.28035
-------------------------------------------------------------------------------
全体   :例数=15     平均値=116          標準偏差=12.2998      標準誤差=3.1758

・群(要因A)のTukey型多重比較(Tukey type multiple comparison)
群       - 群                q値         群数       自由度    有意確率p値
-------------------------------------------------------------------------
1        - 2            6.13332            3          12    0.00256936**
1        - 3            6.74665            3          12    0.00122382**
2        - 3           0.613332            3          12    0.902338
-------------------------------------------------------------------------

・Tukey型95%同時信頼区間(simultaneous confidence interval)
群       - 群          平均値の差       区間幅         下限         上限
---------------------------------------------------------------------
1        - 2                  20     12.3031      7.69693      32.3031
1        - 3                  22     12.3031      9.69693      34.3031
2        - 3                   2     12.3031     -10.3031      14.3031
---------------------------------------------------------------------
```

2群の平均値の比較方法は原理的には2標本t検定と同じ

ファミリーとしての有意水準α_Fを5%にするため1回ごとの有意水準を5/3%にする

→1回ごとの有意確率p値を3倍して有意水準5%と比較する

多重比較の統計学的結論と医学的結論

統計学的結論

A剤投与群とB剤投与群の平均値は異なっていて
A剤投与群とC剤投与群の平均値は異なっているが
B剤投与群とC剤投与群の平均値は異なっているとは言えない
A剤投与群とB剤投与群の差は20であり、幅を取れば8～32の間である
A剤投与群とC剤投与群の差は22であり、幅を取れば10～34の間である
B剤投与群とC剤投与群の差は2であり、幅を取れば-10～14の間である

統計学的結論から医学的結論を導くための検討事項

1. 平均値の差20と22は医学的に見て意義があるか？→医学的許容範囲と比較
2. これらの平均値の差は薬剤の違いによるものか？→3群の背景因子がほぼ均等であることが必要
3. この結果をそのまま高血圧患者全体に当てはめて良いか？→通常は非無作為抽出なので当てはめられない

医学的結論

薬剤AよりもBとCの方が降圧効果は大きく、BとCはほぼ同じである。

多重比較の例え話ーワインとソムリエ

あるレストランのワイン貯蔵庫は管理が悪く、全体の5%のものが悪くなってしまっていた。
そのためソムリエが1本のワインをお客に出した時
それが悪くなっている危険性が5%あるので
ソムリエは20回に1回はお客に謝ることになる(危険率5%)。
ところがお客がワインを3本注文した時
3本のうちの1本でも悪くなっていればソムリエは謝らなければならないので(悪いとこ取り)
危険率が15%に増え、6〜7回に1回は謝ることになる。
そのような場合にソムリエが謝る危険性を5%に抑えるためには
貯蔵庫の管理状態を向上させて、悪いワインの割合を5/3%にする必要がある。

多重比較を適用しなければならない

キーワードは”いいとこ取り(悪いとこ取り)の結論”

3人のお客に1本ずつワインを出したのなら1人のお客に謝る確率は5%

多重比較の例え話ー名医とヤブ医者

あるところに正診率95%(誤診率5%)の医者がいた(危険率5%)。

この医者が1日に1人の患者を診断すると

20日に1回しか誤診しないので周囲から「名医」と評価される。

ところが同じ医者が1日に20人の患者を診断すると、1日に1回は誤診をすることになり

周囲から「ヤブ医者」と評価されてしまう(悪いとこ取り)。

つまり患者が多くて繁盛するほど、ヤブ医者と評価されることになる。

多重比較を適用してはいけない

医者の腕前を評価するには1回の診断に対する誤診率を指標にするべきであり

”悪いとこ取り”をした1日あたりの誤診率を指標にしてはいけない

多重比較に対応する同時信頼区間

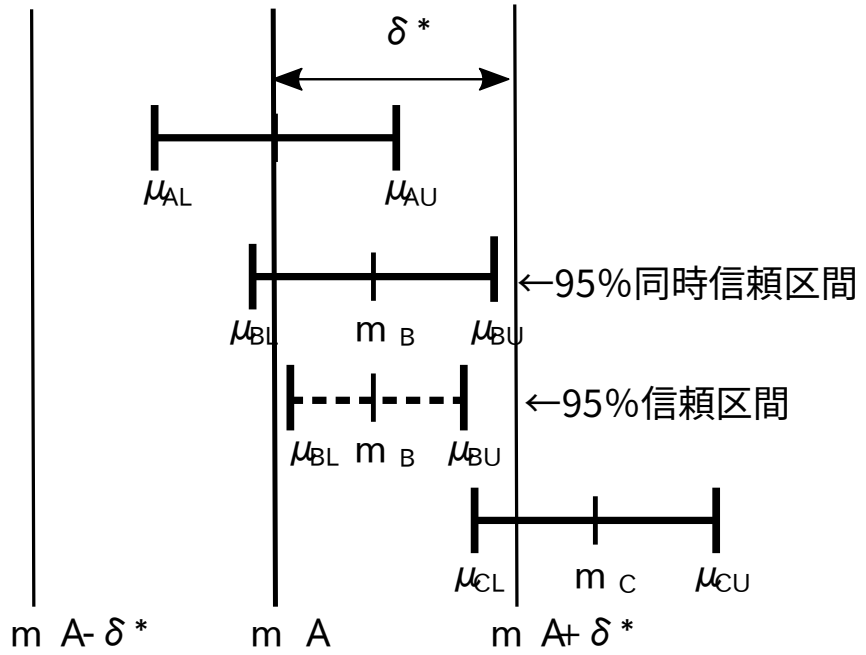


図1.16 同時信頼区間

95%同時信頼区間：3つの信頼区間に3つの母平均が同時に入っている確率が95%

普通の95%信頼区間の場合の同時確率：0.95×0.95×0.95≒0.86

95%同時信頼区間：$0.95^{(1/3)}$≒1-0.05/3≒0.98 ← 約98%信頼区間に相当

→同時信頼区間を利用して3つの母平均の関係を検討する方が実際的→検定廃止論

多重比較とは何ぞや？

- 多重比較は複雑怪奇!?
- 推定の原理
- 統計的仮説検定の原理
- 分散分析の原理
- 多重比較と同時信頼区間の原理
- 多重比較の種類
- 多重比較が必要か？–各種の実例

テューキー(Tukey)型多重比較

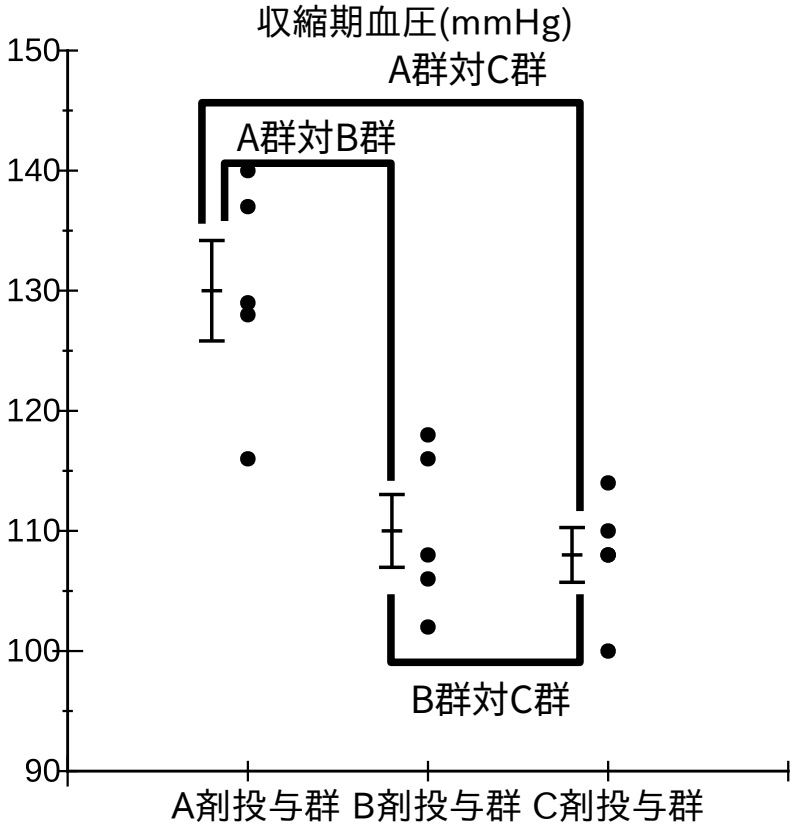


図4.18 多重比較の模式図

全ての2群を比較するリーグ戦方式の多重比較

全ての群を重複して比較するので全ての比較に相関が生じる。

→その相関性を考慮して1回ごとの有意水準を調整→実際には有意確率を調整(3回の比較ではp×2.9程度)

パラメトリック手法：テューキー法、HSD(Honestly Significant Difference)法、テューキー・クレーマー(Tukey-Kramer)法等

ノンパラメトリック手法：テューキー法、スティール・ドゥワス(Steel-Dwass)法等

ダネット(Dunnett)型多重比較

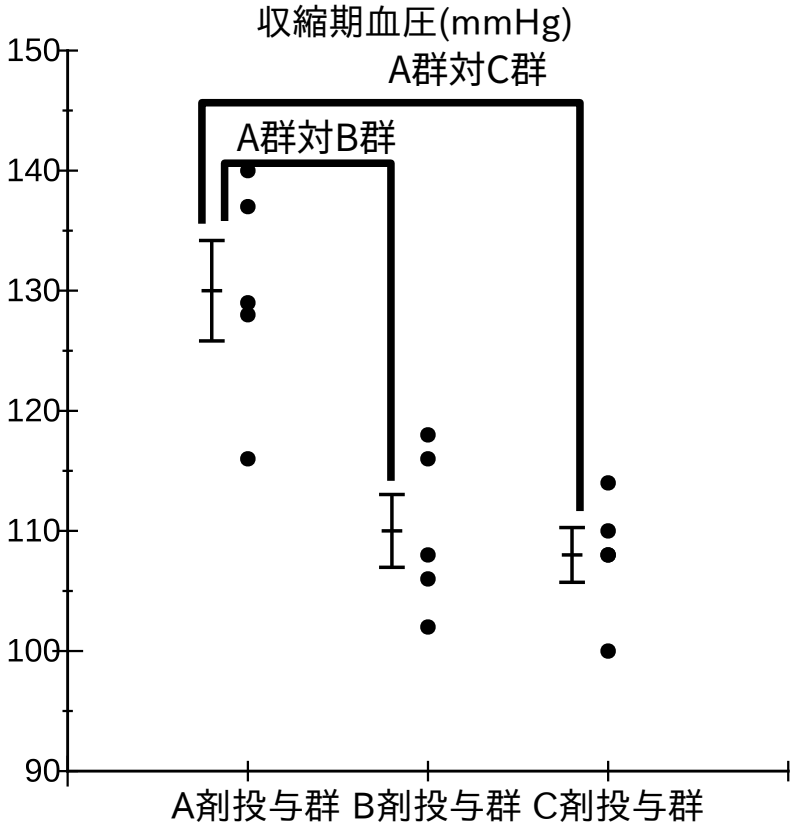


図4.18 多重比較の模式図

特定の群を対照にして、他の全ての群をこれと比較する多重比較

対照群を重複して比較するので全ての比較に相関が生じる。

→その相関性を考慮して1回ごとの有意水準を調整→実際には有意確率を調整(2回の比較ではp×1.9程度)

パラメトリック手法：ダネット法等

ノンパラメトリック手法：ダネット法、スティール(Steel)法等

ボンフェローニ(Bonferroni)型多重比較

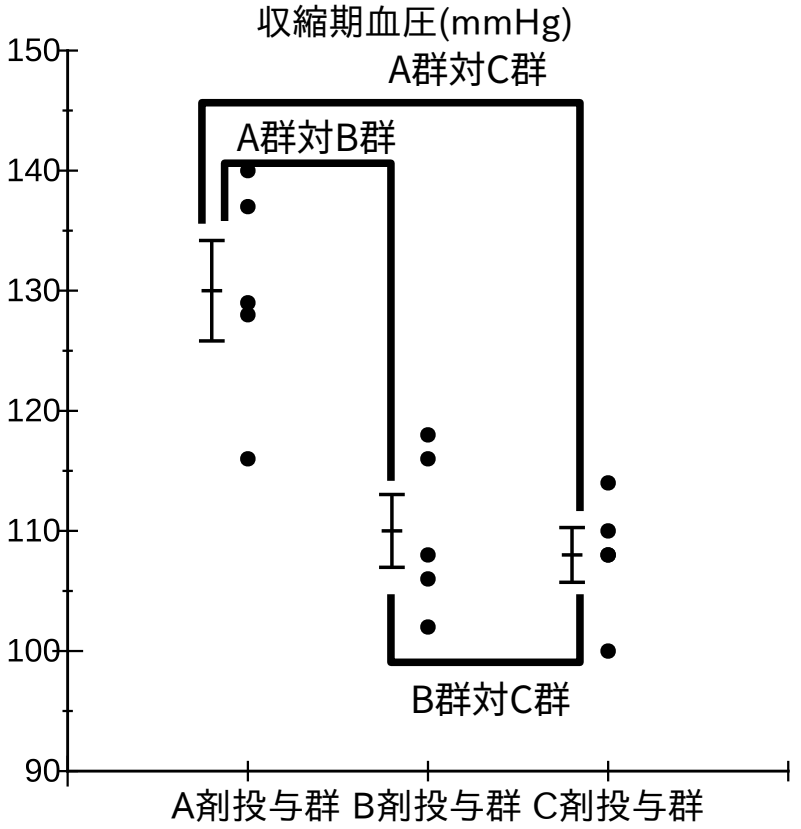


図4.18 多重比較の模式図

テューキー型またはダネット型の近似多重比較

個々の比較が独立と仮定した多重比較→個々の比較に相関があると保守的になる(有意になりにくい)

検定の種類が異なっても適用可能→多重検定(3回の検定ではp×3)

パラメトリック手法：ダン(Dunn)法、ホルム(Holm)法等

ノンパラメトリック手法：ダン法、ホルム法等

シェッフェ(Scheffe)型多重比較

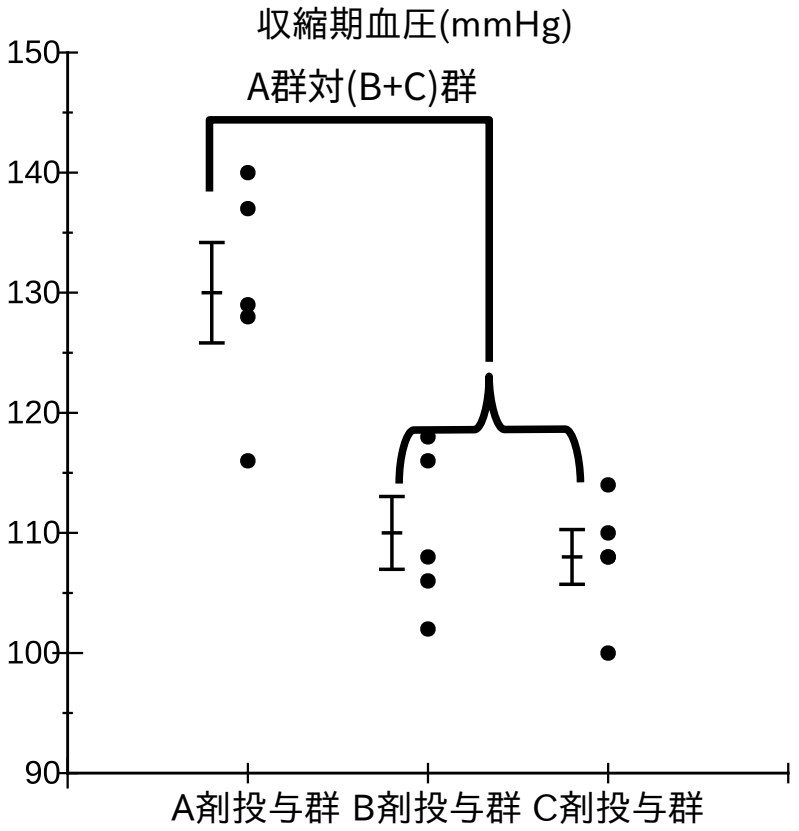


図4.18 多重比較の模式図

リーグ戦方式だけでなく、複数の群を合わせて1つの群にして比較する多重比較

分散分析と同じ原理を用い、最も一般的で保守的な多重比較(3回の比較ではp×4程度)

→通常は分散分析の後で比較を行う事後比較(post hoc comparisons)

パラメトリック手法：シェッフェ法等

ノンパラメトリック手法：シェッフェ法等

多重比較の問題点

多重比較の適用条件

- いいとこ取りした結論(ファミリーとしての結論)が科学的に有意義
- 試験の目的を1つにしぼれず、いいとこ取りした結論が必須の場合
- 個々の比較が独立または相関性がわかっている

普通は独立した3群以上を比較したい時にだけ適用する

投与前vs投与1週後、投与前vs投与2週後等の比較に多重比較を用いる

→投与前・投与1週後・投与2週後のデータは普通は相関がある

評価指標を2つ以上にしたい時→評価指標は普通は相関がある

いいとこ取りした結論が必須ではない時→ステップ法を用いる方が良い

ステップ法と多重比較

プラセボ・標準薬・新薬の3剤比較の場合

- ステップ1(絶対条件)：B(標準薬) vs A(プラセボ)

  試験の分析感度を確認するための比較。B＞Aの時だけ次のステップに進む。

- ステップ2(必要条件)：C(新薬) vs A(プラセボ)

  新薬の有効性を確認するための比較。C＞Aの時だけ次のステップに進む。

- ステップ3(十分条件)：C(新薬) vs B(標準薬)

  新薬の優越性を確認するための比較。C＞BならC＞B＞Aと結論できる。

いいとこ取りした結論ではなく「C>B>A」という結論だけを採用する→個々のステップに多重比較は不必要

新薬開発における第1相～第3相もステップ法の一種

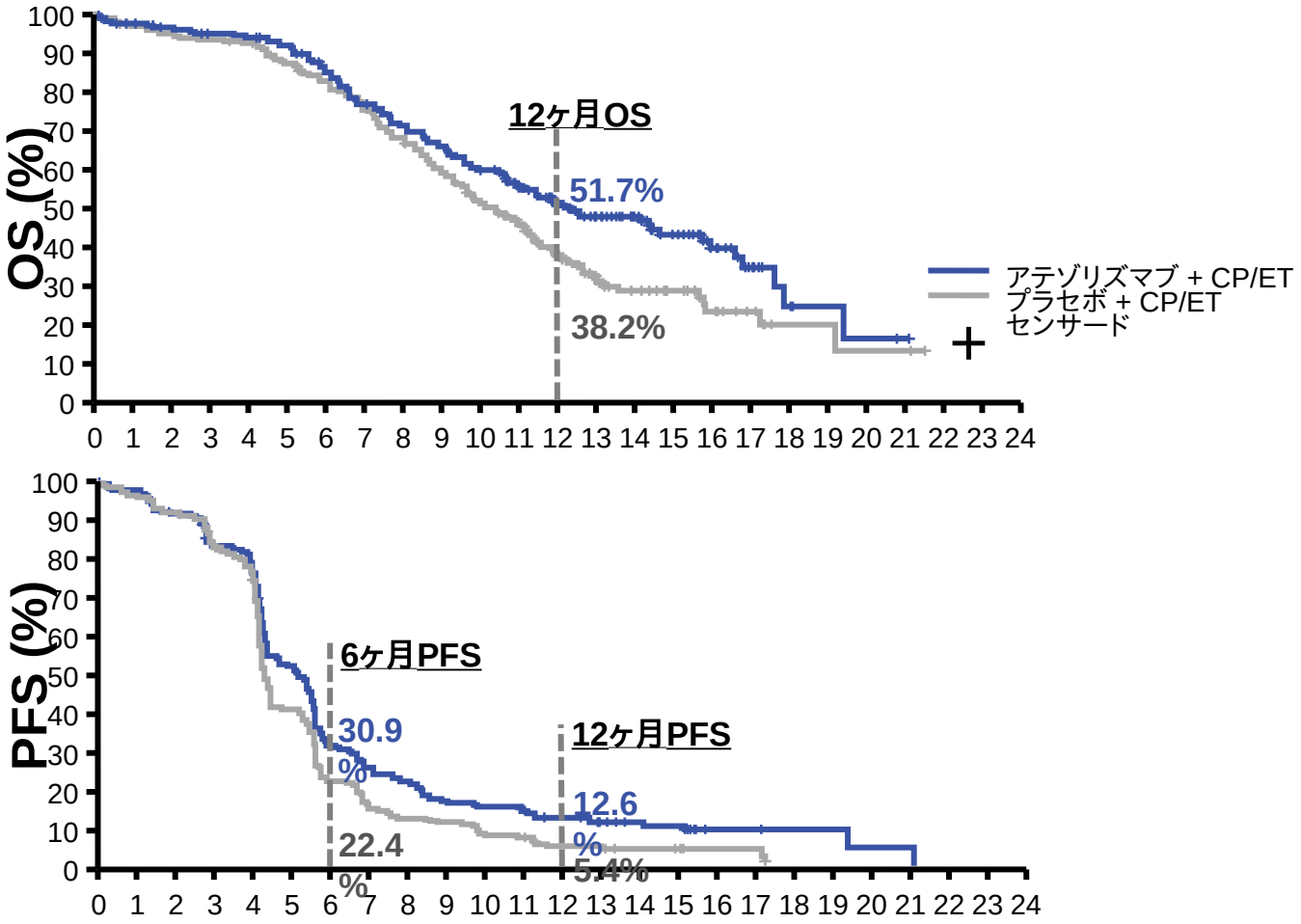
※時間と費用を節約するために1つの試験で検定だけをステップ法に従うのは不適切

→ステップ法は途中でステップを中止し、不必要な試験を実施しないためのもの

多重比較とは何ぞや？

- 多重比較は複雑怪奇!?
- 推定の原理
- 統計的仮説検定の原理
- 分散分析の原理
- 多重比較と同時信頼区間の原理
- 多重比較の種類
- 多重比較が必要か？–各種の実例

多重比較が必要か？–複数の主要評価項目



- 複数の主要評価項目は普通は相関がある
  →しかしボンフェローニー型多重検定を用いることが多い
- 複数の評価項目の結果が相反した時はどのように解釈するか？

多重比較が必要か？−複数の主要評価項目

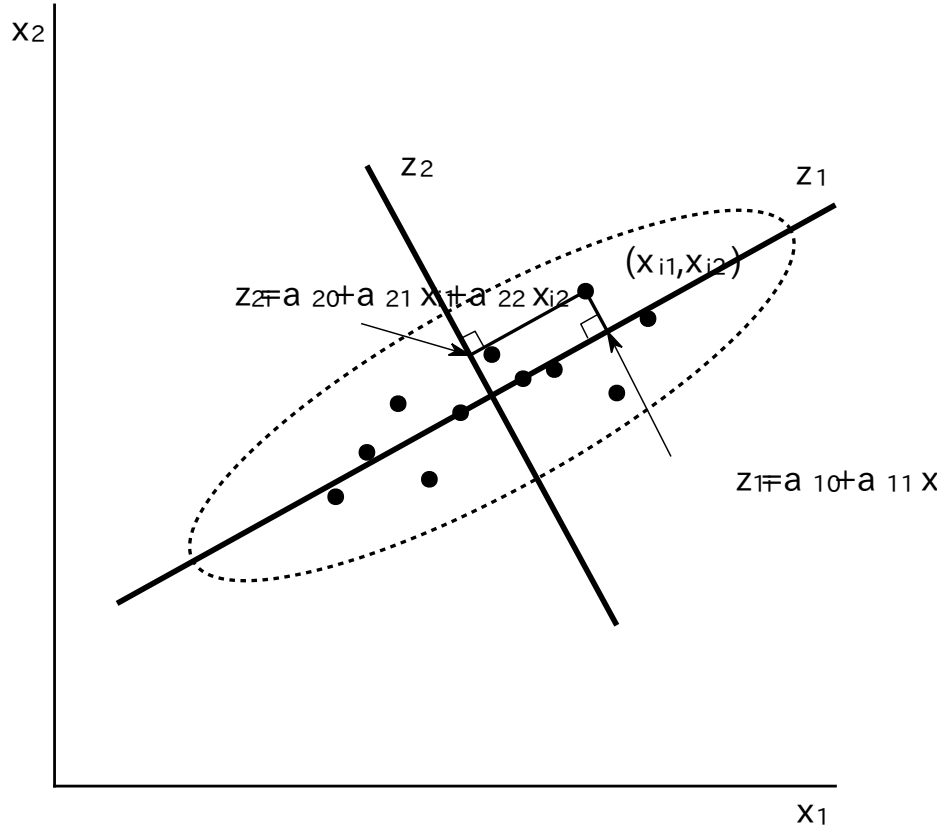


図16.1.1 主成分の幾何学的解釈

複数の評価項目を組み合わせて総合評価項目を作成する

多変量解析のひとつである主成分分析を用いる

多重比較が必要か？−中間解析と終了時解析

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.*

Efficacy End Point	BNT162b2		Placebo		Vaccine Efficacy, % (95% Credible Interval)‡	Posterior Probability (Vaccine Efficacy >30%)§
	No. of Cases	Surveillance Time (n)†	No. of Cases	Surveillance Time (n)†		
		(N=18,198)		(N=18,325)		
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	8	2.214 (1,7411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999
		(N=19,965)		(N=20,172)		
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	9	2.332 (18,559)	169	2.345 (18,708)	94.6 (89.9–97.3)	>0.9999

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.
† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.
‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.
§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

- 中間時のデータと終了時のデータは普通は相関がある
  →中間解析の結果によってはそこで試験を終了する←終了時の結果が予測できるため
- 中間解析の結果と終了時解析の結果が相反した時はどのように解釈するか？

多重比較が必要か？−背景因子の比較

Table 1. Patient demographics at baseline (PC population)

Demographics	Donepezil 5 mg/day (n = 116)	Placebo (n = 112)	p value[1]
Males Females	37 (32%) 79 (68%)	38 (34%) 74 (66%)	0.853
Age, years Mean ± SD Range	 70.1 ± 7.6 52−83	 69.4 ± 8.8 48−90	 0.521
Weight, kg Mean ± SD Range	 51.3 ± 8.4 33−70	 50.0 ± 9.3 29−73	 0.316
Severity (CDR) CDR 1 CDR 2	 79 (68%) 37 (32%)	 69 (62%) 43 (38%)	 0.305
MMSE score Mean ± SD Range	 17.8 ± 3.9 10−26	 16.6 ± 3.9 10−26	 0.035*
ADAS-J cog score Mean ± SD Range	 22.91 ± 8.49 15.0−56.7	 26.90 ± 9.84 15.0−60.0	 0.001*

[1] Data were analyzed by means of U test except for 'Sex (χ^2 test)'.
* $p < 0.15$.

- 検定を行わず、推定結果(点推定と信頼区間)を記載する論文が増えている
  →しかし同時信頼区間を使うべきかどうかが問題
- 背景因子の中には相関があるものが多い

多重比較が必要か？−背景因子の比較

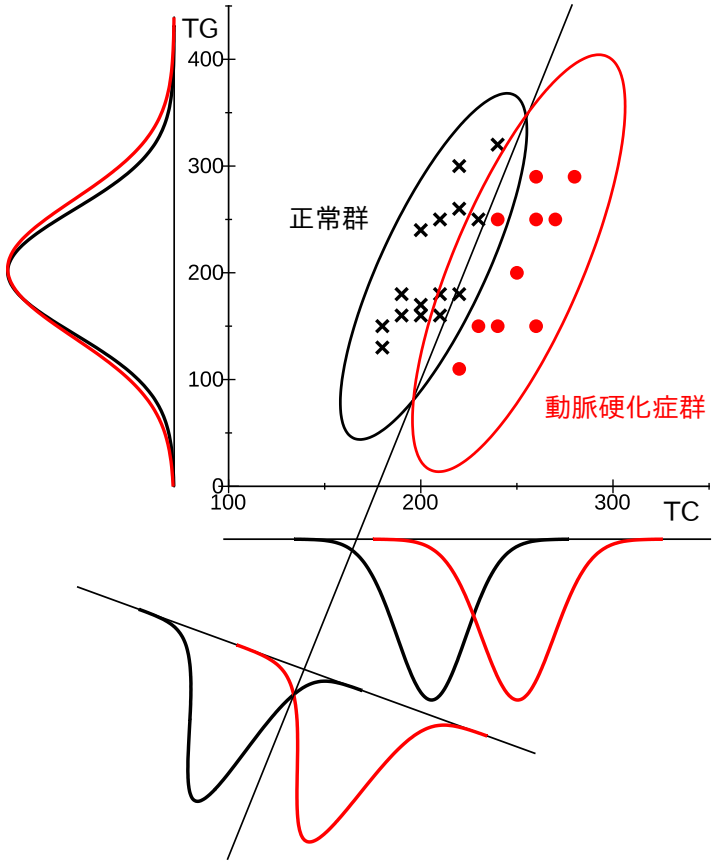
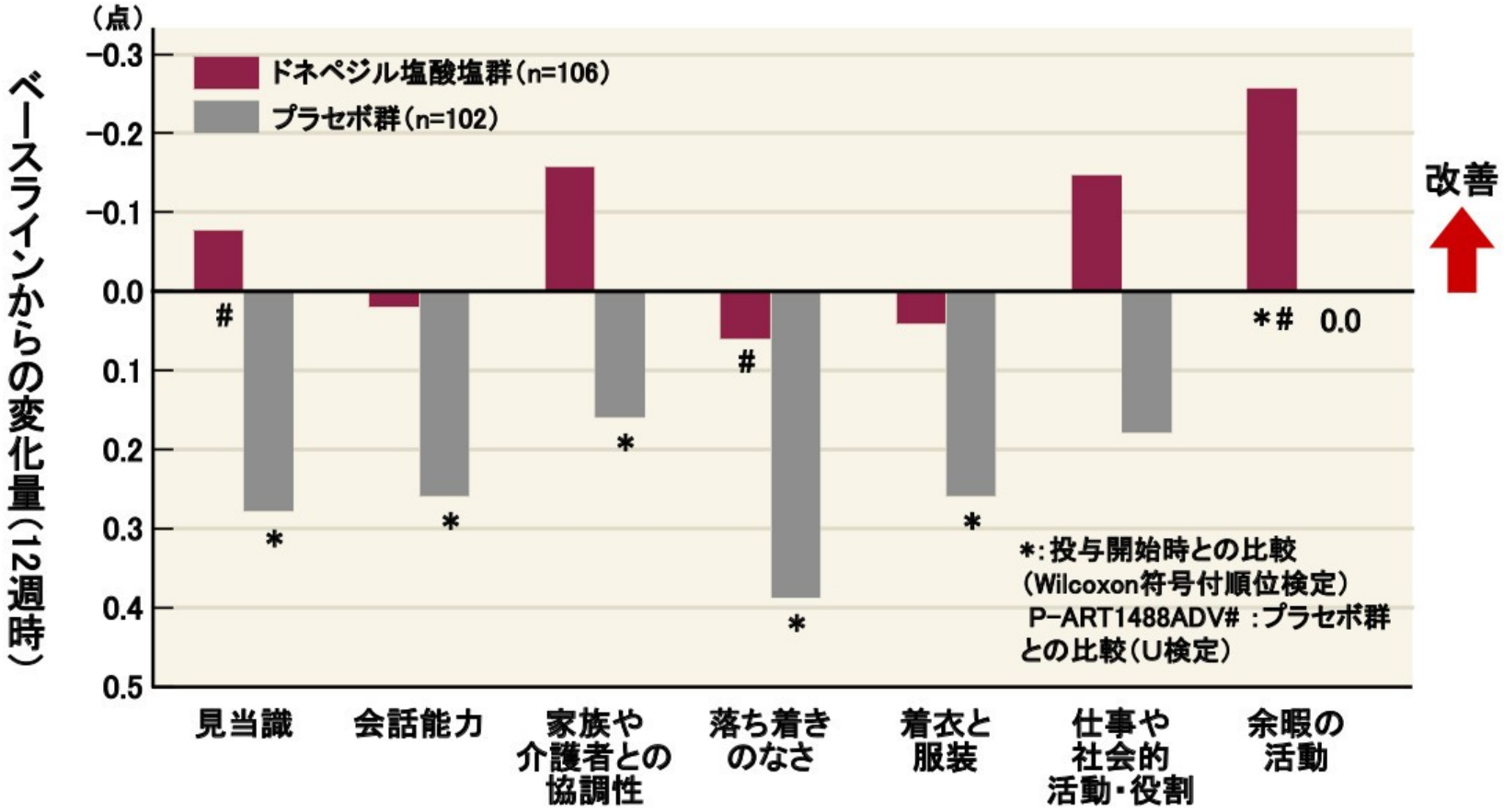


図9.1.1 TCとTGの群別散布図

背景因子間の相関を考慮して2群の重心間の距離(マハラノビスの汎距離)を定義できる

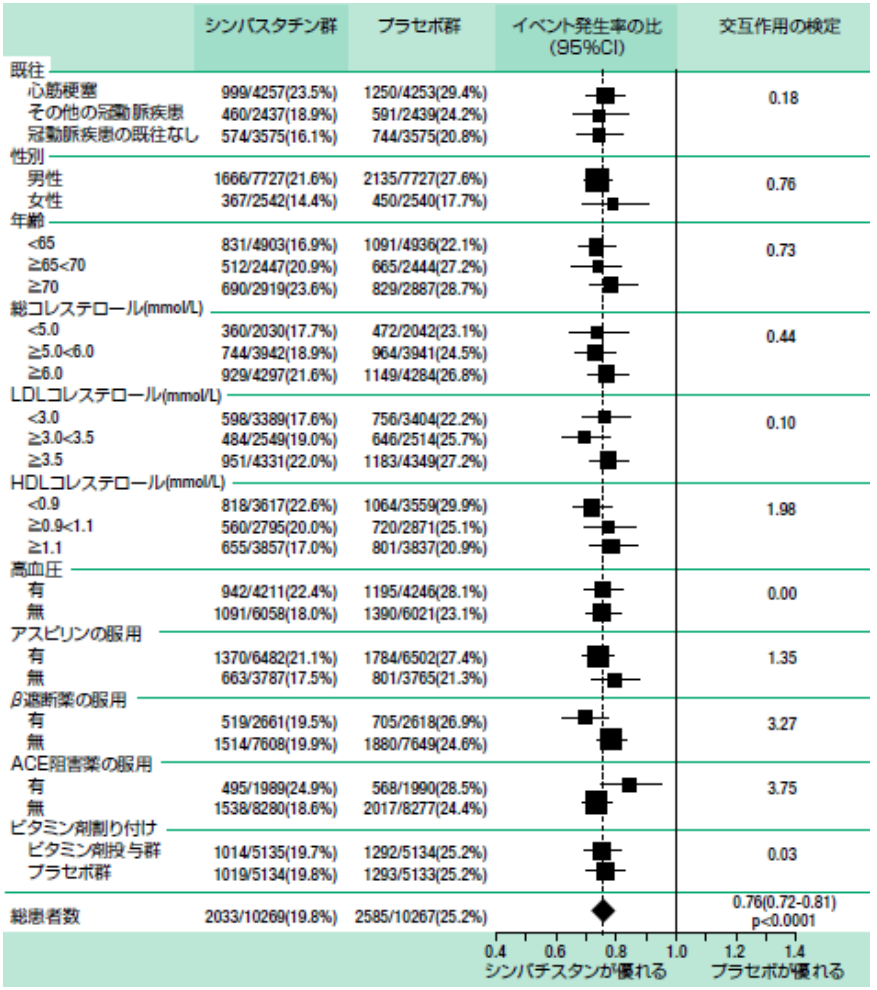
ホッテリングのT^2検定：マハラノビスの汎距離が0かどうかの検定

多重比較が必要か？−副次評価項目



- 試験は主要評価項目の科学的許容範囲に基づいて必要例数を設定する
  →副次評価項目は科学的許容範囲を検討しないので必要例数を満足しているかどうか不明
- 評価項目間に普通は相関がある

多重比較が必要か？–層別解析・サブグループ解析



- 層別解析は探索的な解析なので事前に定量的な仮説を設定するのは難しい
  →仮説を設定していないので検定は無意味なことが多い
- 層別項目間に普通は相関がある

多重比較が必要か？−多変量解析

Table 3 Summary of the multiple regression analysis

Variable	B	SE_B	β	95.0% CI for B	p-Value*
Intercept	257.95	13.29	–	231.59–284.32	< 0.001
Availability of cookies	6.15	2.76	0.210	0.67–11.62	0.028
Teacher A or B	1.80	2.75	0.06	− 3.66 to 7.27	0.514
Gender	0.961	2.94	0.03	− 4.87 to 6.79	0.744
BMI	− 0.670	0.41	− 0.163	− 1.49 to 0.15	0.107
Age	− 0.460	0.37	− 0.121	− 1.19 to 0.27	0.212

* Level of significance: $p < 0.05$.
β = standardised coefficient; B = unstandardised regression coefficient; BMI = body mass index; CI = confidence interval; SE_B = standard error of the coefficient.

・多変量解析は探索的な統計手法なので事前に仮説を設定するのは難しい
→仮説を設定していないので検定は無意味なことが多い

本日の結語

主要評価項目を1つにし

できるだけ単純な試験にし

多重比較を使わないようにしましょう！

ご清聴ありがとうございました