2014/10/11 JCOG臨床試験セミナー中級編

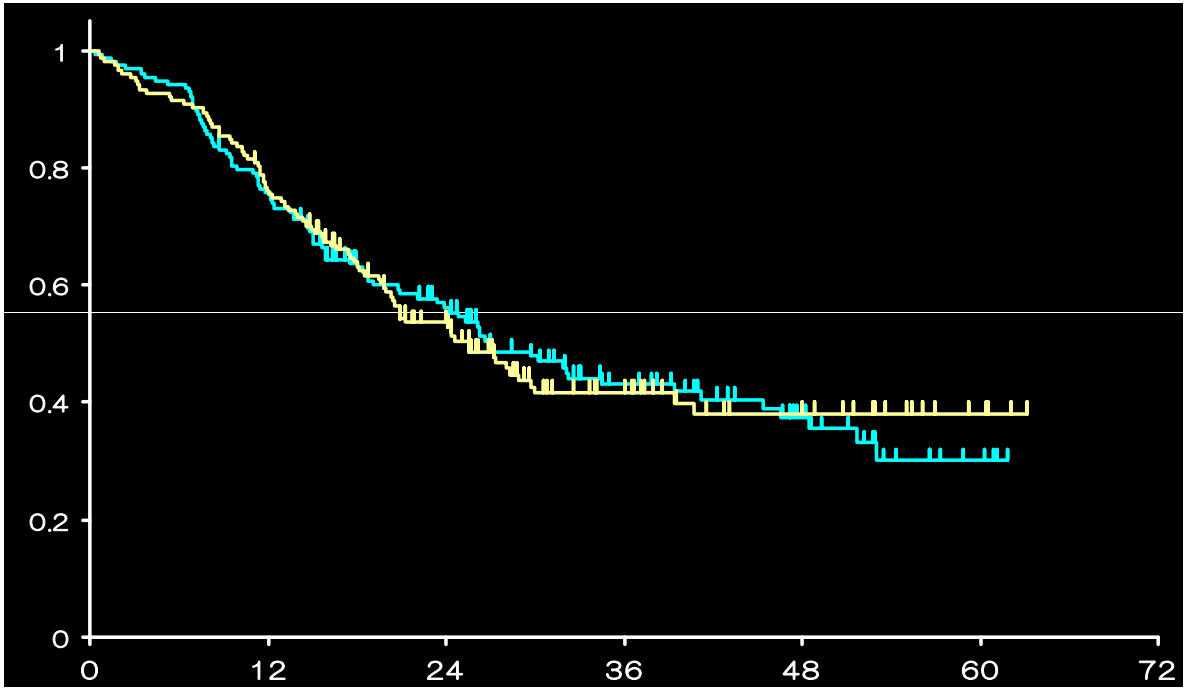
検証的試験における多重性の調整

独立行政法人 国立がん研究センター 研究支援センター 生物統計部
JCOGデータセンター
柴田大朗

Outline

- 複数の検定を行う際に多重性の調整が必要な理由
- 多重性の調整方法
  - Bonferroni法
  - Holm法
  - Hochberg法
  - その他の方法
- 多重性の調整の理解の仕方
- 多重性の調整に関する留意点

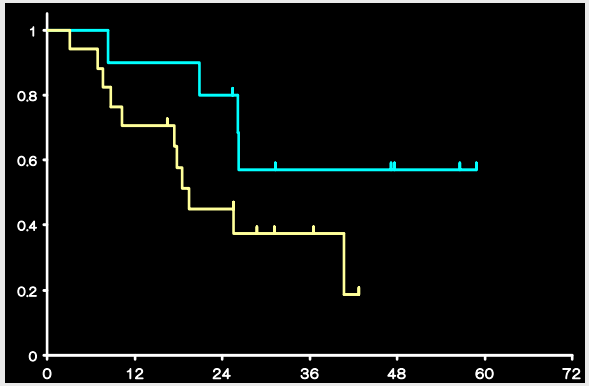
ある第Ⅲ相試験によるシミュレーション



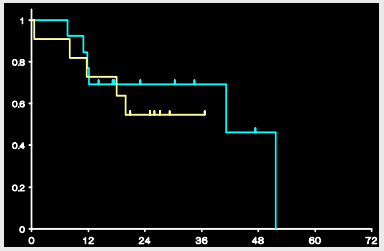
$P = 0.985$
(Logrank test)
n = 304

登録番号の下1ケタで10分割してみると・・

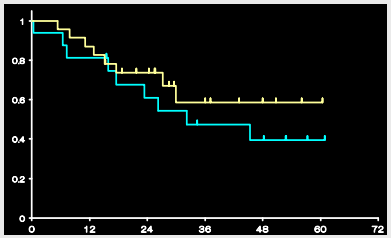
P = 0.091



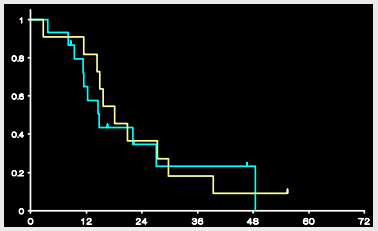
P = 0.592



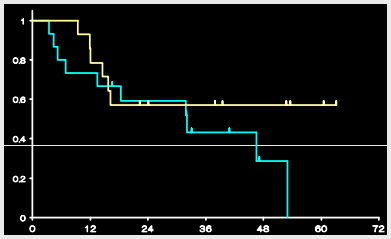
P = 0.393



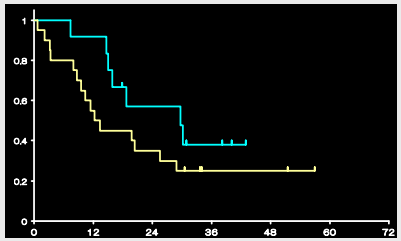
P = 0.767



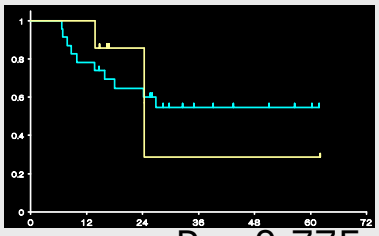
P = 0.229



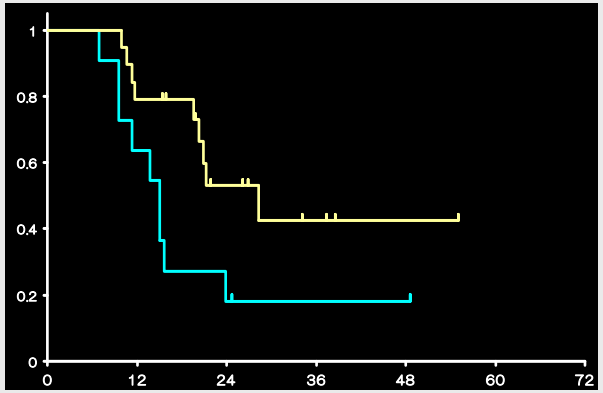
P = 0.185



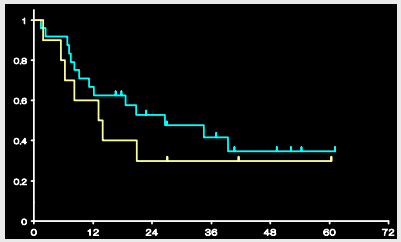
P = 0.897



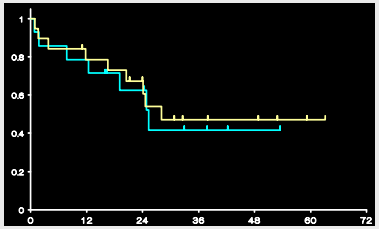
P = 0.035



P = 0.452



P = 0.775



意味のないサブグループでも沢山調べれば有意差が出る

検定の多重性 multiplicity

- 何回も比較（検定）を行うと、どこかで有意差が出る可能性が高くなること
  - 多くの検定、多くのエンドポイント、多くのサブグループ
- 検定の回数と「有意差 p < 0.05」が得られる確率
  - 帰無仮説（差がない）のもとで有意水準 5%で検定すると…

比較の回数	どこかで「有意差 p < 0.05」が出る確率（%）
1	5.0
2	9.7
3	14.3
4	18.5
5	22.6
10	40.0
20	64.1

言っていること（「第一種の過誤5%」）と実態とに大きな乖離が生じてしまう

Richard Simon：CANCER - Principles & Practice of Oncology：DeVita et al.

検証的試験における検定

- 検証的試験のprimary endpointの解析
  - 先行研究等の合理的根拠に基づき、検証に値する仮説を立てる
  - 事前に設定した仮説に対し、解析方法と有意水準（第一種の過誤：通常は5%など）を決めて行う
  - 第一種の過誤の大きさを事前に宣言することで、意思決定に伴う誤りの大きさを明らかにした手順となる
- 宣言している第一種の過誤の大きさが、実際には守られていなければ、意思決定ルールとしての性能が担保できない
- 問題が生じるのは、一つの研究の中で
  - 複数のエンドポイントに関する検証
  - ひとつのエンドポイントに関する複数回の検証

  などを、何の準備もなく行う場合。これでは、第一種の過誤の大きさが名義上の値よりも大きくなってしまう

統計的検定に伴う2つの種類の誤り

限られたデータでの判定	神のみぞ知る本当の状態：効いていない	神のみぞ知る本当の状態：効いている
効いている（P<0.05）	αエラー	OK
効いていない（NS）	OK	βエラー

- 第一種の過誤（αエラー、「あわてんぼう」のエラー）
  - 本当は効いていない新治療を、間違って効いていると判断してしまう誤り
- 第二種の過誤（βエラー、「ぼんやりもの」のエラー）
  - 本当は効いている新治療を、間違って効いていないと判断してしまう誤り
  - 「1-βエラー」を検出力と呼ぶ

統計的検定のロジック

- 「証明したい仮説とは反対の仮説（帰無仮説）」をおいてデータを解析し、その「反対の仮説」にデータが整合しないことをもって、証明したい仮説を採用するという手続きを踏む
  - 示したいこと：医薬品候補に効果がある
  - 反対の仮説（帰無仮説）：~~医薬品候補に効果がない~~
    - 効果がないという前提でデータを見た場合につじつまが合わなくなることを示して、効果があるという結論を導く

帰無仮説を棄却

→帰無仮説のもとで実際に得られたようなデータが生じる確率が「非常に小さいこと」をもって帰無仮説を棄却

→多重性によって実際にはこの確率が非常に大きくなってしまっているのなら、検定のロジックが破綻

Outline

- 複数の検定を行う際に多重性の調整が必要な理由
- **多重性の調整方法**
  - Bonferroni法
  - Holm法
  - Hochberg法
  - その他の方法
- 多重性の調整の理解の仕方
- 多重性の調整に関する留意点

多重性への対処 - 誤った判断をしないためには

- あらかじめ宣言した解析のみを主たる判断に用いる
  - 最初に決めておいたサブグループ解析しか結論に用いない
  - *ad hoc*（後知恵）な解析はすべて“探索的”→ 別の研究／対象で確認

  ・・・など

- 多重性の調整 adjustment/correction

  Bonferroni法：α = 0.05を比較の数で割る
  - 10のサブグループ解析をするのであれば、$P < 0.005$ のものだけを“有意”と考える
  - 論文によっては、逆にp値を10倍し、0.05未満のものだけ”有意”とするケースもある（提示の仕方の違いであり、検定結果は一致）

多群比較の場合

- Bonferroni の調整

2群比較/1対比較

標準治療A ⟷ 試験治療B

P < 0.05 で有意

試験全体の α=0.05

3群比較/2対比較

試験治療B　　試験治療C

P < 0.025　　P < 0.025

標準治療A

試験全体の α=0.05

3群比較/3対比較

試験治療B ⟷ 試験治療C

P < 0.017

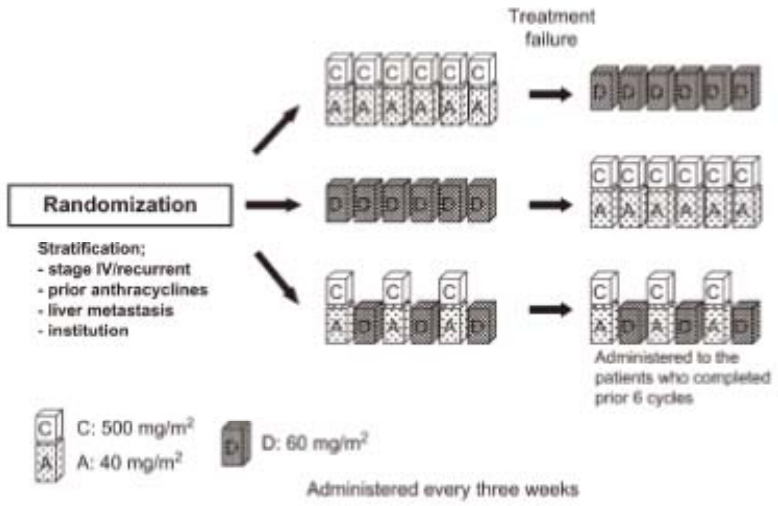
P < 0.017　　P < 0.017

標準治療A

試験全体の α=0.05

JCOG9802試験

- 乳がん対象の3群比較試験
  - AC vs. DTX
  - AC vs. AC/DTX交替療法



- 検定したい仮説が2つあるため、Bonferroni法で多重性の調整
  - 試験全体の有意水準**片側0.05**
  - 各々の仮説を**片側0.025**(=0.05/2)で検定

- 結果 [Primary endpoint TTF]
  - AC vs. DTX　　p=0.13>**0.025**
  - AC vs. AC/DTX交替療法　　p=0.14>**0.025**

多重性の調整方法－Bonferroni法の例

- 試験全体の有意水準を5%とし、4つの検定を行う場合
- 結果が仮に以下のようなものだったとすると・・・
  - 検定(a) p=0.038　< 0.05 有意！　> **0.0125**
  - 検定(b) p=0.012　< 0.05 有意！　**< 0.0125 有意！**
  - 検定(c) p=0.049　< 0.05 有意！　> **0.0125**
  - 検定(d) p=0.024　< 0.05 有意！　> **0.0125**
- この場合、「全て5%有意」と考えてはいけない
  - 試験全体の第一種の過誤は「$1-0.95^4=18.5\%$」であり、これを第一種の過誤5%で検定したとするのは誤りであるため
- Bonferroni法は、各検定のp値が0.05とではなく、**0.05を検定の回数で割った値**と比べる
  - 上の例では各検定毎に0.05ではなく、0.05/4=**0.0125**を用い検定(b)のみを統計学的に有意と結論づけることになる

多重性の調整方法－Holm法

- **Holm法**
  - 検定をp値が小さいものから順に並べる
  - 順に（0.05ではなく）**以下の値**と比べる
  - **この値**より大きなp値があったら、それ以降の検定は**この値**との大小関係によらず「有意差なし」と結論づける

検定が4つの場合

検定をP値で並べ替え	Bonferroni法	Holm法	
1: p=0.012	* <0.0125	*<**0.0125**	**0.0125**=0.05/4×1
2: p=0.024	>0.0125	*<**0.0250**	**0.0250**=0.05/4×2
3: p=0.038	>0.0125	>**0.0375**	**0.0375**=0.05/4×3
4: p=0.049	>0.0125	(**0.0500**)	**0.0500**=0.05/4×4

もし検定が2つなら

検定をP値で並べ替え	Bonferroni法	Holm法	
1: p=0.012	* <0.025	*<**0.025**	**0.025**=0.05/2×1
2: p=0.024	* <0.025	*<**0.050**	**0.050**=0.05/2×2

多重性の調整方法－Hochberg法

- **Hochberg法**
  - 検定をp値が小さいものから順に並べる
  - 順に（0.05ではなく）**以下の値**と比べる
  - この**値**より小さなp値があったら、それ以降の検定は**この値**との大小関係によらず「有意差あり」と結論づける

検定が
4つの場合

検定をP値で並べ替え	Bonferroni法	Hochberg法	
1: p=0.012	* <0.0125	* (**0.0125**)	**0.0125**=0.05/4
2: p=0.024	>0.0125	* (**0.0167**)	**0.0167**=0.05/3
3: p=0.038	>0.0125	* (**0.0250**)	**0.0250**=0.05/2
4: p=0.049	>0.0125	* <**0.050**	**0.0500**=0.05/1

  - Bonferoni法やHolm法を用いるよりも、より多くの仮説を「有意差あり」と結論づける
  - **ただし、不適切な条件下では（仮説の間の相関の状況によっては）誤った結論に至るという欠点がある**

その他の、よく用いられる手法

- **Bonferroni法**、**Holm法**、**Hochberg法**は複数の仮説間の関係を考えず、優先順位もつけない
  - Bonferroni法は全ての仮説を同様に扱い、Holm法、Hochberg法は得られた結果（p値）に基づき仮説間の優先順位を決める

  →これが臨床的な意思決定方針と整合しない場合がある

- 仮説間の関係性や優先順位を考慮した方法
  - **階層手順（hierarchical testing procedure）**
    - 試験計画時に複数の仮説間に順序を設け、優先度の高い仮説が統計学的に有意であった場合のみ、次の仮説を順に検定
    - 有意差がでなかったら、引き続く仮説の検定は行わない
    - ひとつひとつの仮説は0.05で検定する
  - **Gatekeeping法**
    - より複雑な、複数の仮説のグループ間で優先順位を設ける方法（詳細省略）

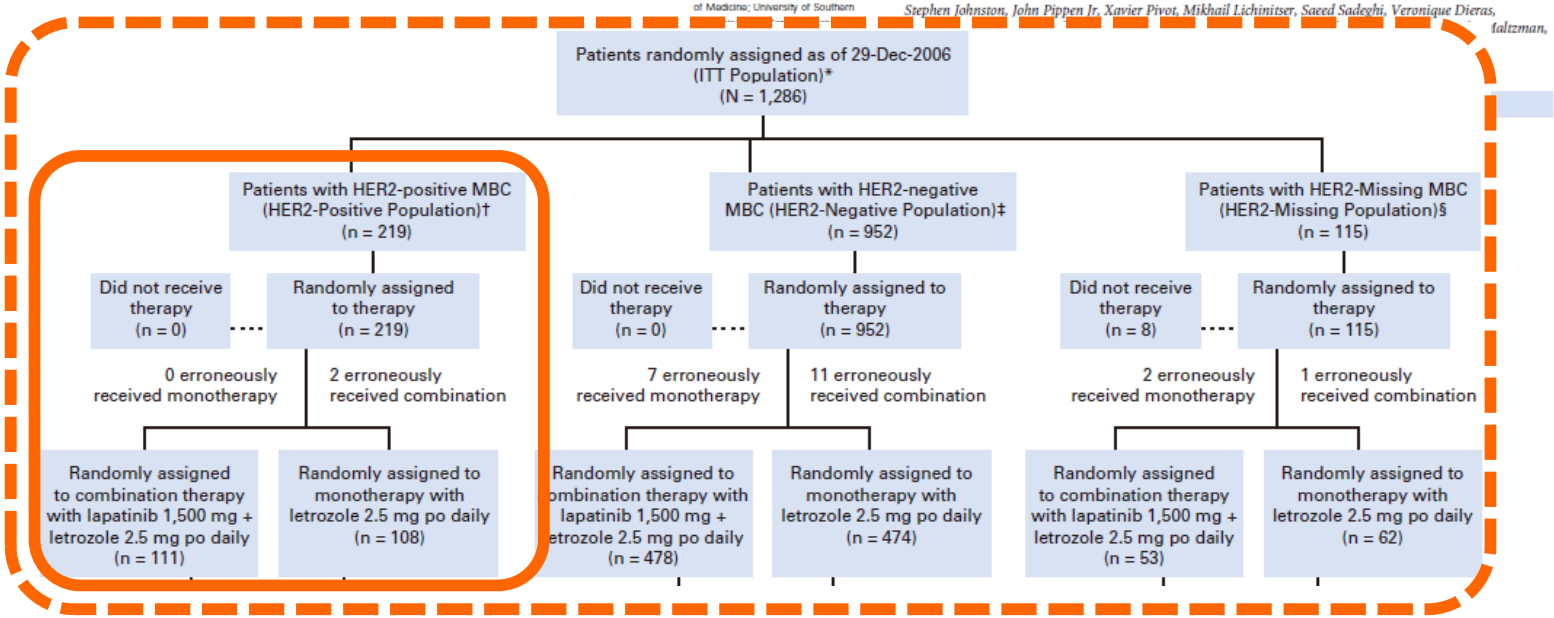
階層手順の適用例

VOLUME 27 · NUMBER 33 · NOVEMBER 20 2009
JOURNAL OF CLINICAL ONCOLOGY
ORIGINAL REPORT

Lapatinib Combined With Letrozole Versus Letrozole and Placebo As First-Line Therapy for Postmenopausal Hormone Receptor–Positive Metastatic Breast Cancer

Stephen Johnston, John Pippen Jr, Xavier Pivot, Mikhail Lichinitser, Saeed Sadeghi, Veronique Dieras, ... Valtzman,

From the Royal Marsden Hospital, London; GlaxoSmithKline, Middlesex, United Kingdom; Sammons Cancer Center, Dallas, TX; David Geffen School of Medicine; University of Southern



- 転移性乳がん対象の単剤 vs. 分子標的薬併用療法の比較
- 第一段階：HER2陽性の部分集団を対象に**0.05**で検定
- 第二段階：第一段階が有意だった時のみ、全集団を対象に**0.05**で検定

階層手順とHochberg法を組み合わせた適用事例

VOLUME 29 · NUMBER 20 · JULY 10 2011
JOURNAL OF CLINICAL ONCOLOGY ORIGINAL REPORT

Palifermin Reduces Severe Mucositis in Definitive Chemoradiotherapy of Locally Advanced Head and Neck Cancer: A Randomized, Placebo-Controlled Study

Quynh-Thu Le, Harold E. Kim, Charles J. Schneider, Györgyi Muraközy, Krzysztof Skladowski, Sabine Reinisch, Yuhchyau Chen, Michael Hickey, May Mo, Mon-Gy Chen, Dietmar Berger, Richard Lizambri, and Michael Henke

Quynh-Thu Le, Stanford University, Stanford; Michael Hickey, May Mo,

A B S T R A C T

of chemoradiotherapy for head and neck cancer

Primary endpointが0.05水準で有意であった時のみ、secondary ep.の検定を0.05で行うsecondaryの複数項目に渡る検定の多重性はHochberg法で調整

Le et al. JCO, 2011.

Table 2. Efficacy End Points

End Point	Palifermin (n = 94) No.	%	Q1	Q3	Placebo (n = 94) No.	%	Q1	Q3	Unadjusted *P*	Adjusted *P**
Primary										
Incidence of severe oral mucositis†	51	54			65	69			.041	
Secondary										
Median duration (days) of severe oral mucositis	5		0	40	26		0	50	.016	.112
Median time (days) to onset of severe oral mucositis	47		27	N/E	35		22	N/E	.026	.157
Incidence of grade ≥ 2 xerostomia at month 4†	63	67			75	80			.046	.231
Average MTS score (0-4)									.071	.285
Mean	1.66				1.86					
SD	0.73				0.65					
Median total opioid analgesic (mg IV morphine equivalent)	283.16		4.63	956.19	498.33		25.00	1513.33	.238	.684
Incidence of chemotherapy delays‡	49	52			42	45			.342	.684
Incidence of radiation breaks§	14	15			14	15			.959	.959

Abbreviations: MTS, mouth and throat soreness; N/E, not estimable; Q1, first quartile; Q3, third quartile; SD, standard deviation.

Adjustment for multiple statistical testing was performed by using a combination of the **hierarchical testing procedure** and the **Hochberg procedure** to protect the overall type I error rate of 5%.

多重性の調整方法－Holm法・Hochberg法の欠点

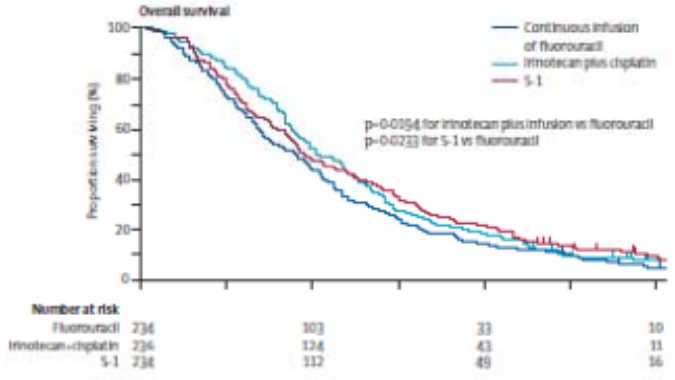
検定をP値で並べ替え	Bonferroni法	Holm法	Hochberg法
(b) p=0.012	* <0.0125	*<0.0125	* (0.0125)
(d) p=0.024	>0.0125	*<0.0250	* (0.0167)
(a) p=0.038	>0.0125	>0.0375	* (0.0250)
(c) p=0.049	>0.0125	(0.0500)	* <0.050

もし(b)のP値が違ったら

検定をP値で並べ替え	Bonferroni法	Holm法	Hochberg法
(d) p=0.024	>0.0125	>0.0125	>0.0125
(a) p=0.038	>0.0125	(0.0250)	>0.0167
(c) p=0.049	>0.0125	(0.0375)	>0.0250
(b) p=0.051	>0.0125	(0.0500)	>0.0500

- **Holm法、Hochberg法は、ある仮説を「有意」と判断するか否かが、別の仮説の状況に左右される**

JCOG9912試験



- 胃がん対象の3群比較試験

- 検定したい仮説が2つあり、
  多重性の調整が必要
  当初**Bonferroni法**を検討
  - 5-FU vs. IP(優越性)：片側0.025
  - 5-FU vs. S-1(非劣性)：片側0.025

  → 非劣性有意なら片側0.025で5-FU vs S-1(優越性) の
  検証（非劣性が証明された後の優越性検証は多重性調整不要）

  最終的に**Holm法**を採用
  - 2つの群間比較のp値の小さい方を片側**0.025**で検定し、有意ならば、もう一方の検定を片側**0.050**で行う
- 結果：
  - 5-FU vs. S-1(非劣性)　p=0.001＜**0.025**
  - 5-FU vs. IP(優越性)　p=0.055＞**0.050**（Holm法）

  5-FU vs. S-1の非劣性が片側**0.025**で統計的に有意であったので
  優越性の検証へ進む（多重性調整不要）

  → 5-FU vs. S-1(優越性)　p=0.034＞**0.025**…Not Significant

JCOG9912試験（つづき）

- 仮にIPが著しく良い成績だったら結論が変わったのか？
  - もし現実と異なり5-FU vs. IP(優越性)のp値が*0.0001*だったら

  - 5-FU vs. IP(優越性)　p=*0.0001*　<**0.025**
  - 5-FU vs. S-1(非劣性)　p=0.001　<**0.050**（Holm法）

  5-FU vs. S-1の非劣性が片側**0.05**で統計的に有意であったので優越性の検証へ進む（多重性調整不要）
  → 5-FU vs. S-1(優越性)　p=0.034　<**0.050**…有意！？

  仮想データに基づく思考実験

- ある仮説が有意であるか否かが他の仮説の状況に左右されて良いのか？
  - 良いと考えられるケースと、そうでないケースがある
  - **単に数学的に多重性が調整できていれば良いのではなく、現実の意思決定と選択した多重性の調整方法とに齟齬が無いことが必要（＝仮説間の関係を事前に検討しておく必要がある）**

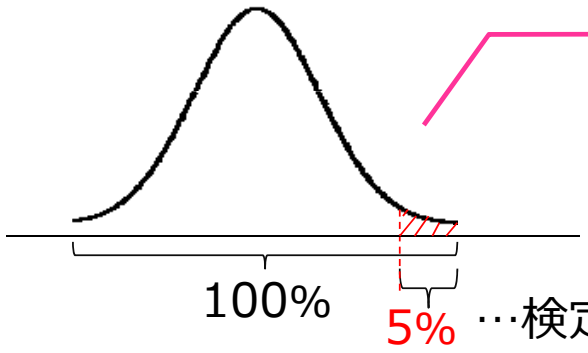
Outline

- 複数の検定を行う際に多重性の調整が必要な理由
- 多重性の調整方法
  - Bonferroni法
  - Holm法
  - Hochberg法
  - その他の方法
- **多重性の調整の理解の仕方**
- 多重性の調整に関する留意点

多重性の問題が生じる理由（イメージ図）

※薬に効果が無くても検定をすると20回に1回は有意差がつく

・帰無仮説の下での検定統計量の分布



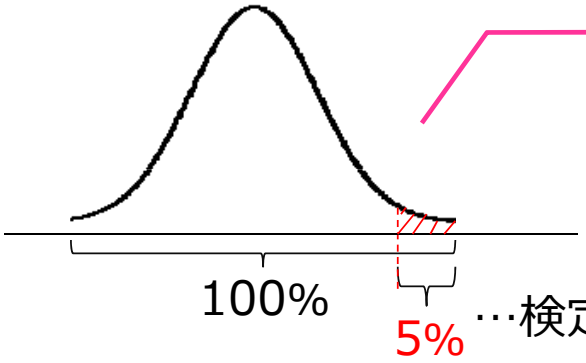
真実は「差が無い」でも5%は、誤って有意差有りと結論してしまう
（＝第一種の過誤の大きさ）

…検定で生じる第一種の過誤

多重性の問題が生じる理由（イメージ図）

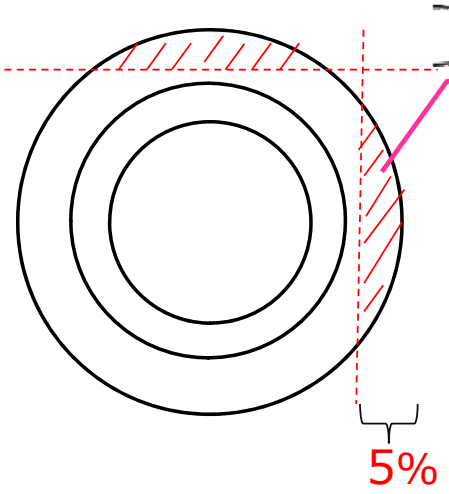
※薬に効果が無くても検定をすると20回に1回は有意差がつく

・帰無仮説の下での検定統計量の分布



…検定で生じる第一種の過誤

真実は「差が無い」でも5%は、誤って
有意差有りと結論してしまう
（＝第一種の過誤の大きさ）

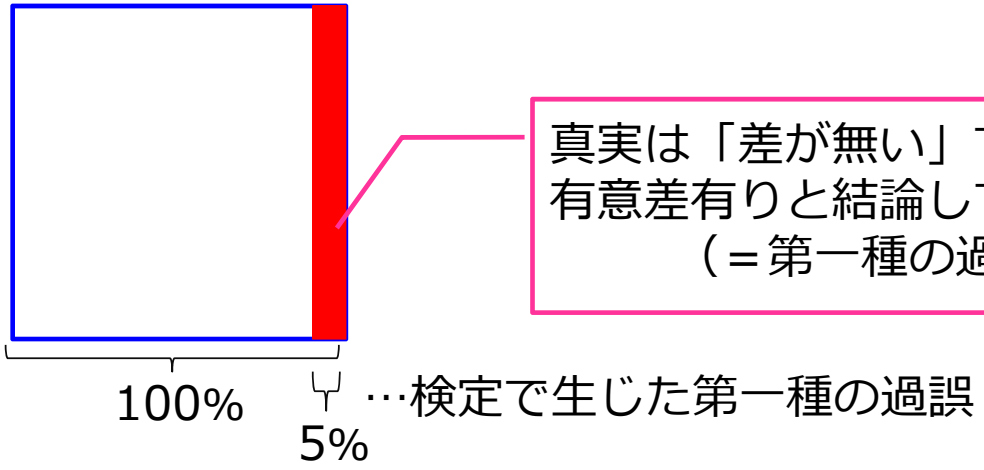


2回検定する場合、2つの検定統計量の
同時分布の斜線部分が、
検定で生じる第一種の過誤にあたる

多重性の問題が生じる理由（イメージ図）

※薬に効果が無くても検定をすると20回に1回は有意差がつく

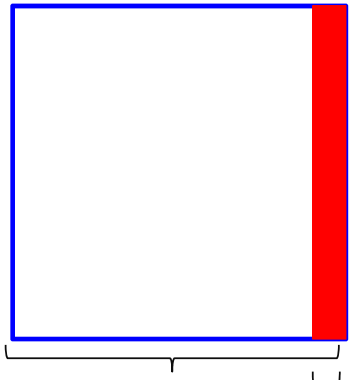
・この確率を（不正確ではあるが）面積で喩えると以下のようになる。



多重性の問題が生じる理由（イメージ図）

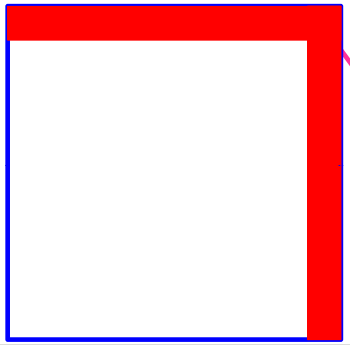
※薬に効果が無くても検定をすると20回に1回は有意差がつく

・この確率を（不正確ではあるが）面積で喩えると以下のようになる。



100%　5%　…検定で生じた第一種の過誤

・検定を2回行うと、試験全体では第一種の過誤の大きさは約2倍になる

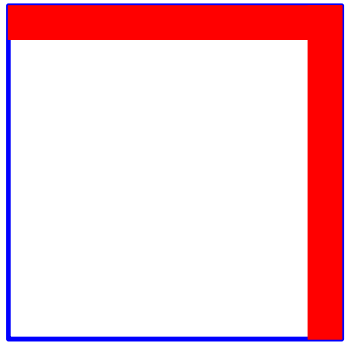


☞5%：2つめの検定で生じた第一種の過誤

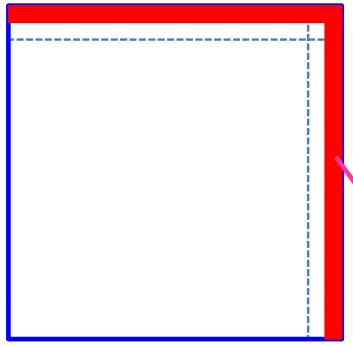
真実は「差が無い」でも5%は、誤って有意差有りと結論してしまう（＝第一種の過誤の大きさ）

Bonferroniの方法による調整

・2つの検定を行うと、この誤りの面積は5%を越える



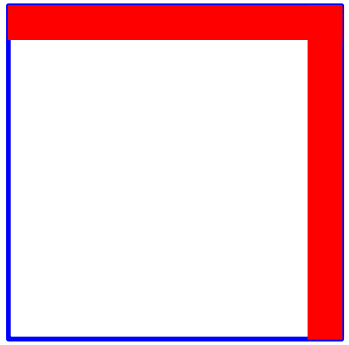
・2つの検定を行う場合でも、1つずつの有意水準を2.5%とすると（ボンフェローニ法を用いると）、この誤りの面積は5%を越えない



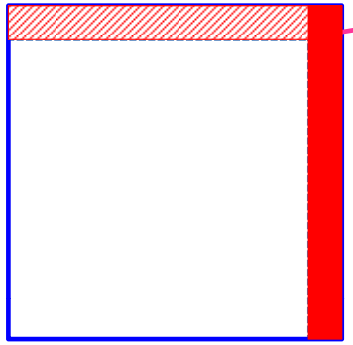
塗りつぶした部分の面積は
2.5%+2.5%-0.0575% <5%

“階層がある仮説”に対する検定の場合

・2つの検定を行うと、この誤りの面積は5%を越える



・2つの検定を行う場合でも、1つ目の検定が有意であったときに限り2つめの検定を行う場合、誤りの面積は5%を越えない



塗りつぶした部分の面積は5%
斜線部分は生じない

※「非劣性が有意に検証されず、優越性の検定が有意になること」はあり得ないため、斜線部分が生じず、その部分の面積をカウントしなくてよいので、誤りの面積は5%を越えない

Outline

- 複数の検定を行う際に多重性の調整が必要な理由
- 多重性の調整方法
  - Bonferroni法
  - Holm法
  - Hochberg法
  - その他の方法
- 多重性の調整の理解の仕方
- **多重性の調整に関する留意点**

ICH E9「臨床試験のための統計的原則」の指摘

- 5.6有意水準と信頼水準の調整
- **多重性が存在する場合**、臨床試験データの解析に対する通常の頻度論的立場からは、**第一種の過誤の調整を必要とする**であろう。
- 多重性は、例えば**主要変数が複数ある場合**（2.2.2節参照）、**試験治療間の多重比較**、**時間に伴う繰り返し評価**、**中間解析**（4.5節参照）から生じるであろう。
- 多重性を回避する又は減じる方法は、それが利用できる場合には好ましいこともある。例えば、**変数が複数ある場合に主要な変数の中でも重要な変数を指定**すること、**多群比較の場合に試験治療間の重要な対比を選択**すること、繰り返し測定の場合に「曲線下面積」といった**要約指標を使用**すること、などである。
- 検証的解析では、このようにして**多重性を減じた後の段階でまだ残っている多重性のすべての側面について治験実施計画書において明らかにすべき**である。調整は常に考慮すべきであり、調整方法の詳細、又はなぜ調整は必要ないと考えるのかという説明は、統計解析計画書に述べるべきである。

ICH E9「臨床試験のための統計的原則(1998.11.30)」
http://www.pmda.go.jp/ich/e/e9_98_11_30.pdf

多重性の調整を用いる際の留意点

- **数学的な仮定が満たされていない場合、αが名義水準以下にコントロールできないケースがある**
  - 適用範囲が限られている手法がある

- **数学的にαがコントロールできることと、その結果の解釈が臨床的に意義のあるものであることとには乖離がある**
  - αがコントロールできても、臨床的に意義・価値のない多重性の調整方法を選択してしまっては意味がない。

- 数学的にαがコントロールでき、複数の仮説間の扱いが臨床的に妥当でも、**合理的根拠無しに検定を行う場合、結果はあてにならない**
  - 多重性の調整以前の問題：「検証」に値しない状況下での検定結果は（数学的に正しく有意であっても）意味が無い
  - αの面で損をしないから、合理的根拠の無い仮説に関して階層手順で「おまけ」の検定をつけておく・・・という方針は取るべきではないし、有意差がついても評価されない