医学統計セミナー2020：医学統計Q & A

下川敏雄
和歌山県立医科大学附属病院 臨床研究センター

アジェンダ

QA１　量的データにおける要約の方法 (標準偏差，標準誤差，信頼区間について)
QA.2　量的データにおける検定法の取捨選択について(t検定，Welch検定，Wilcoxon検定について)
QA.3　変化量のデータを解析したら，レフェリーから前値の影響を指摘されたのですが．．．
QA.4　カイ2乗検定とFisherの正確検定の使い分けについて
QA.5　多重比較っていったい何ですか？
QA.6　必要症例数のことですが．．．
QA.7　回帰分析(多変量解析)の変数選択について
QA.8　回帰分析(多変量解析)における2値化の問題について
QA.9　観察研究における傾向スコアって何ですか？

QA.10　名義変数と連続変数の相関を表す統計の数値はありますでしょうか？
QA.11　JMPで多変量解析を行った際，ORや信頼区間が極端に高い値になってしまうことがあります．
QA.12　多重共線性をVIF以外で検討する方法について教えてください．
QA.13　名義変数と連続変数を同時に検討する場合には，多重共線性というのは検討できますか？

QA.1：量的データにおける要約の方法

平均値を要約指標に用いた場合には，バラツキの指標に迷うことが良くあります

バラツキを何であらわしましょう？

標準偏差

$$SD = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2}$$

x_n：観測値 (n=1,2,…,N)
$\bar{x}$：平均値

標準誤差

$$SE = \frac{SD}{\sqrt{N}}$$

信頼区間

$$CI = t_{N-1}(\alpha/2)\cdot SE$$

$t_{N\text{-}1(\alpha/2)}$：自由度N-1のt分布の上側α/2パーセント点

標準偏差

標準偏差は，平均値まわりでのデータのバラツキを意味する．

各被験者の平均値からの距離を(観測値－平均値)^{2}で測る．これを偏差平方という(2乗しないと合計をとったときに0になるため)．

↓

偏差平方の平均値を計算する．これが分散(母分散)である．推測統計学では単純な平均値ではなく，偏差平方の合計値(これを偏差平方和と呼ぶ)をN-1で割る．これを不偏分散とよぶ．

↓

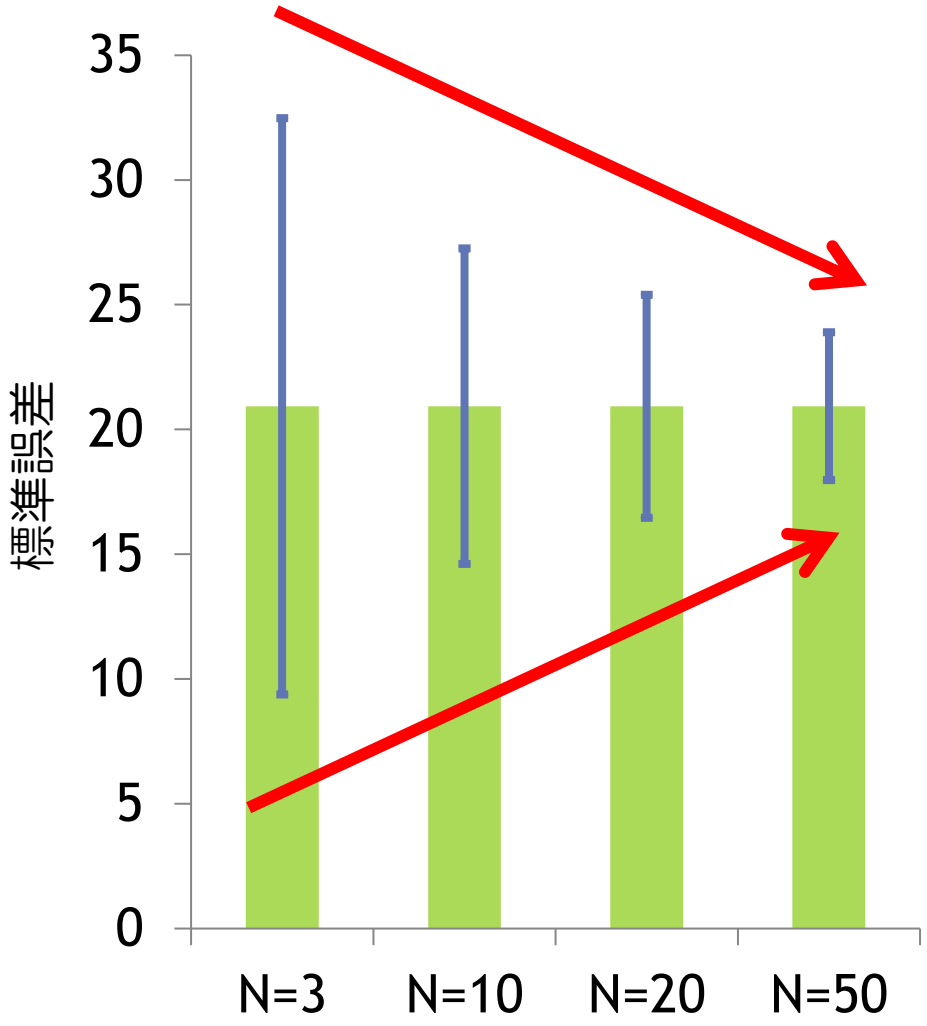
分散(あるいは不偏分散)は偏差平方和の平均を表す．ただし，偏差平方和はもとのデータを2乗しているので，平均値と単位が異なる．そのため，平方根をとることで，もとの単位に戻す．これが標準偏差である．

<バラツキの目安>

- 1SD：データが正規分布に従うとき，M±SDのなかに約68%が含まれる．
- 2SD：データが正規分布に従うとき，M±SDのなかに約95%が含まれる．

また，標本サイズが増加したからと言って，標準偏差が減少することはない．

標準誤差



平均値の信頼性(平均値のバラツキ)を表している.

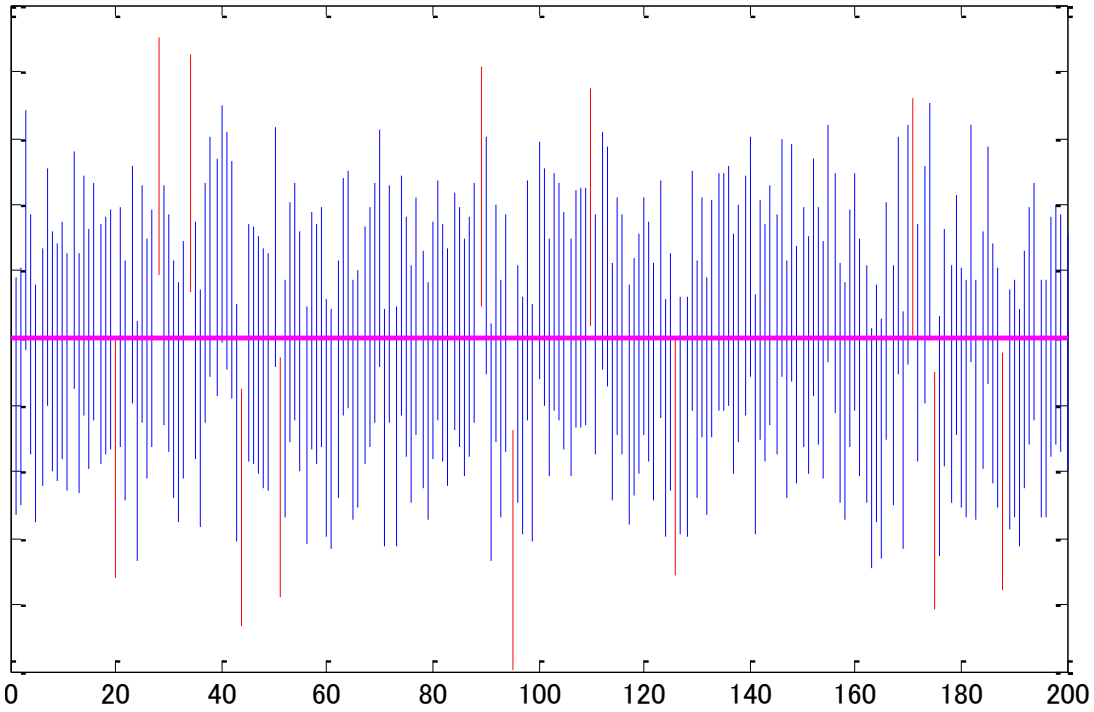
公式 : $SE = \frac{SD}{\sqrt{N}}$

データの個数(N)が増加すれば，SEは減少する.

データの個数が増えれば，平均値の信頼性が増加する．そのため，SEは減少する.

信頼区間

平均値0の正規分布から乱数を生成し，
95%信頼区間を200回生成した結果



200回のうち，188回(94%)が母平均(真の平均値)である0を含んでいる．すなわち，おおよそ95%の確率で真の平均値を含んでいる．

95%の確率で真の平均値を含む確率ではない！！

通常の平均値は，ある一つの数字で表されるため，点推定値と呼ばれる．しかし，測定されたデータの平均値と真の平均値には，乖離が生じる．信頼区間では，区間で表すことで，「〇〇～××までの範囲」で表す．そのため，区間推定値と呼ばれる．

標準偏差・標準誤差・信頼区間の使い分け

標準偏差SDの利用

- 観測値のばらつきを表すのに用いる.
- 例えば，被験者背景を表す場合には，どのような被験者が参加しているかの大まかな情報を反映させるために標準偏差を用いることが多い.

標準誤差SEの利用

- 平均値の信頼性を表すのに用いる.
- 例えば，アウトカム(応答変数)の評価などでは，平均的な傾向とその信頼性を知ることが必要になるが，このような場合には利用することが許容される.

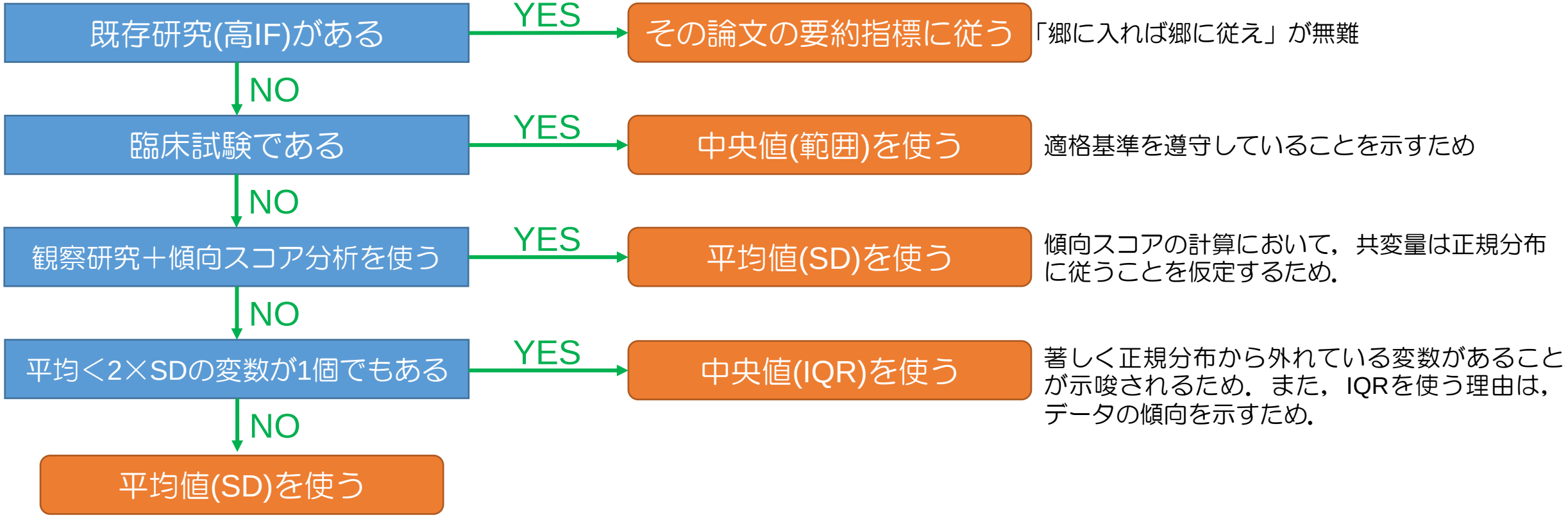
信頼区間の利用

- 母平均(真の平均)を含む範囲として計算される.
- 比較では用いるべきでない．信頼区間は，単群でのデータの精査に用いるべきである.
- 基本的には標準誤差と同じような理由で用いられることから，標準誤差あるいは信頼区間を記載するのが正しい利用法である.

ちなみに．．．

背景情報を論文に記載する場合に，平均値(標準偏差)と中央値(IQRあるいは範囲)のどちらを用いればよいでしょうか．

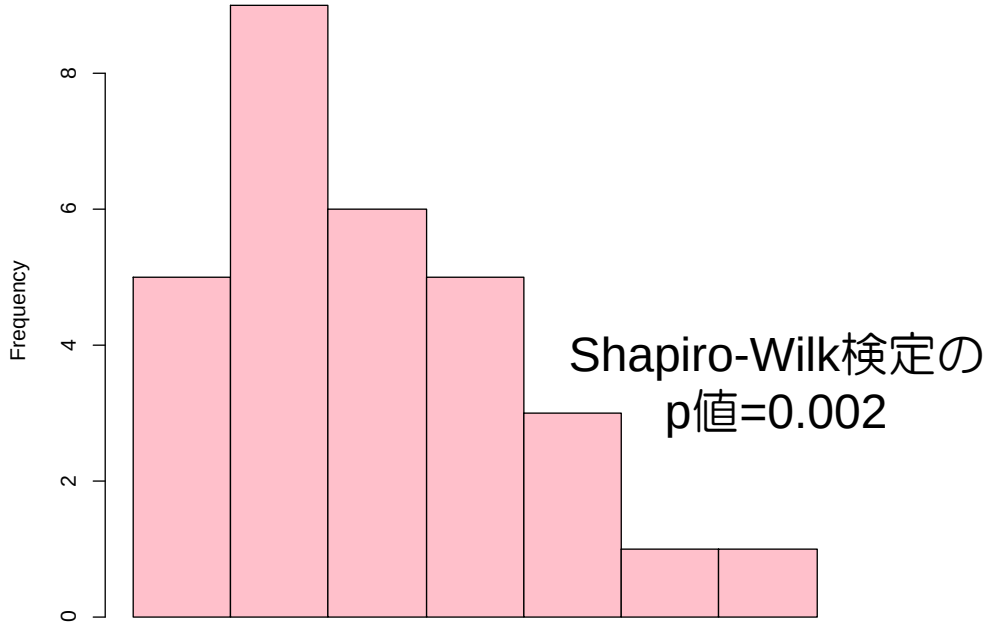
絶対にこうしないといけないという決まり事は特にありません．あくまでも個人的な考えです．



Warning!! ：「平均(IQR)」あるいは「中央値(SD)」はダメ
Tableにおける要約指標は統一されていることが望ましい(と思っています)．

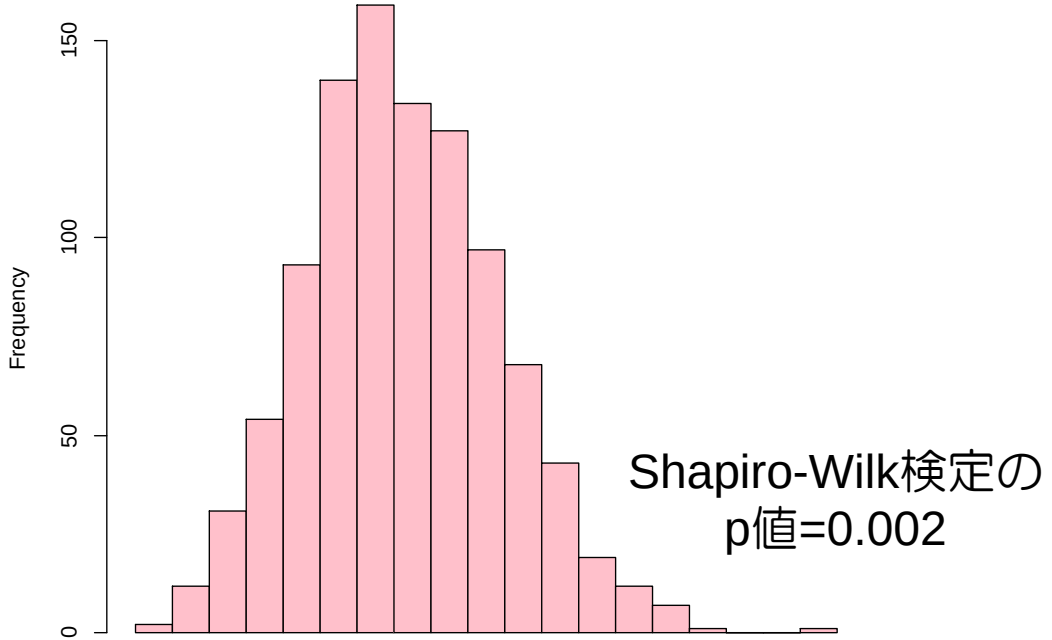
ちなみに正規分布を使って正規性は評価しないこと!! (後ほど解説)

N=30の調査研究



ヒストグラムを見る限りでは正規分布に近い形状だと**思いますか？**

N=1000の調査研究



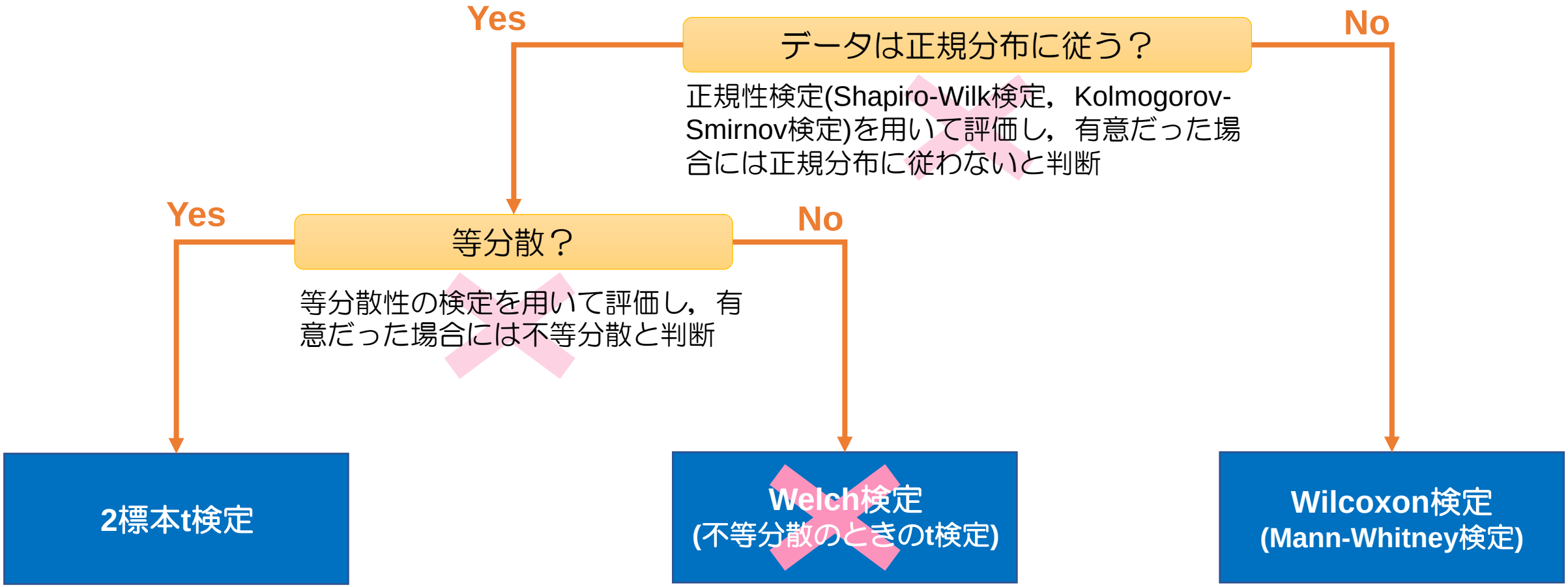
ヒストグラムを見る限りでは正規分布に近い形状だと**思いませんか？**

正規性検定を基準に検定法を選択してしまうと，「データの歪み」だけでなく，「サンプルサイズの大きさ」が影響を及ぼし，サンプルサイズが大きければ，僅かな歪みでも有意になる．そもそも，サンプルサイズで検定法を選択するというのがおかしい．

QA.2：量的データにおける検定法の取捨選択

量的データの場合にはどの検定を使えばいいですか？
(2標本t検定，Welch検定(不等分散のときのt検定)，Wilcoxon検定(Mann-Whitney検定))

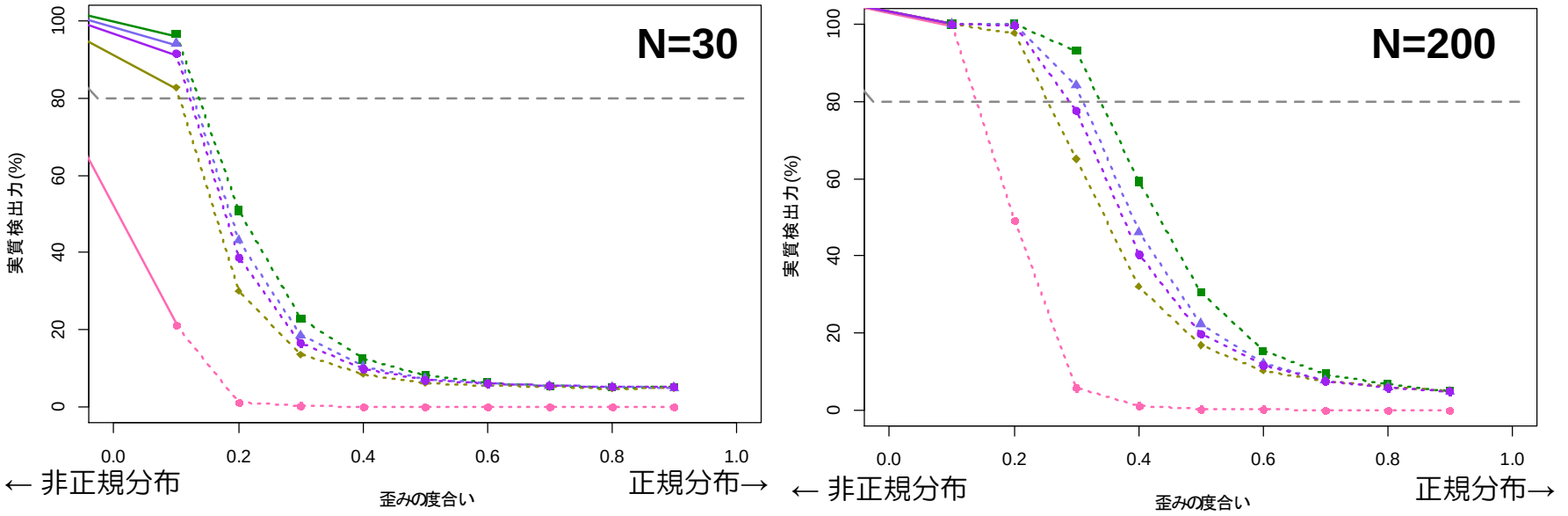
よくある仮説検定の取捨選択の誤解

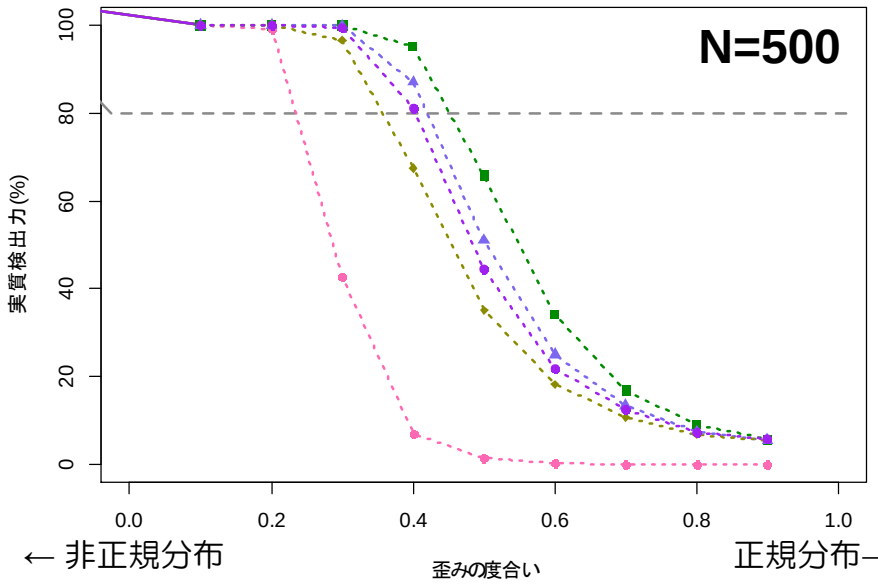


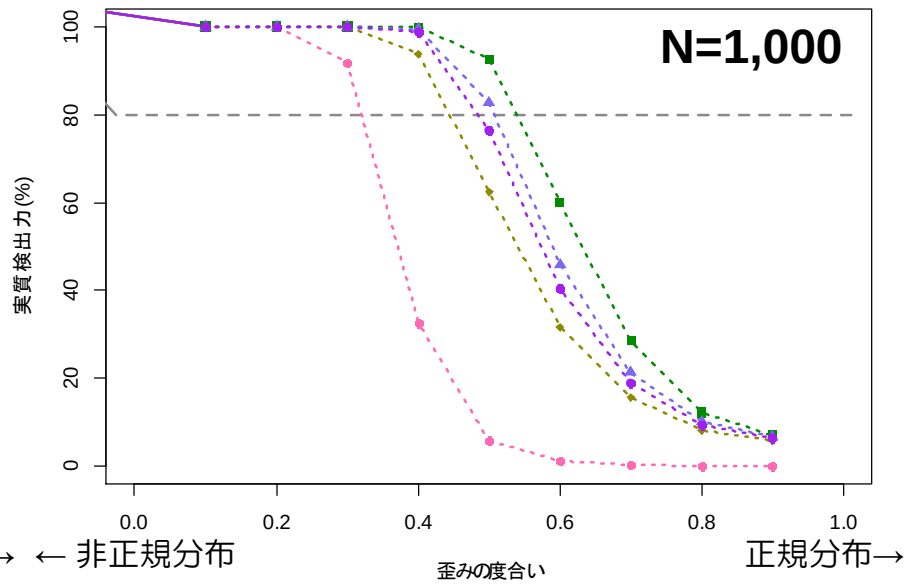
それぞれの理由について説明し，最後に一つの代替案を説明します

正規性の検定を検定の取捨選択に用いることは推奨されない

人工データのもとで10,000回の検定を行ったときに有意になった回数(実質検出力)







- Kolmogorov-Smirnov検定
- Lillie検定(修正KS検定)
- Cramer von Mises検定
- Anderson-Daring検定
- Shapiro-Wilk検定

統計パッケージに含まれていることが多いKolmogorov-Smirnov検定はなかなか有意にならない(鈍感)であり，Shapiro-Wilk検定はすぐに有意になる(敏感).

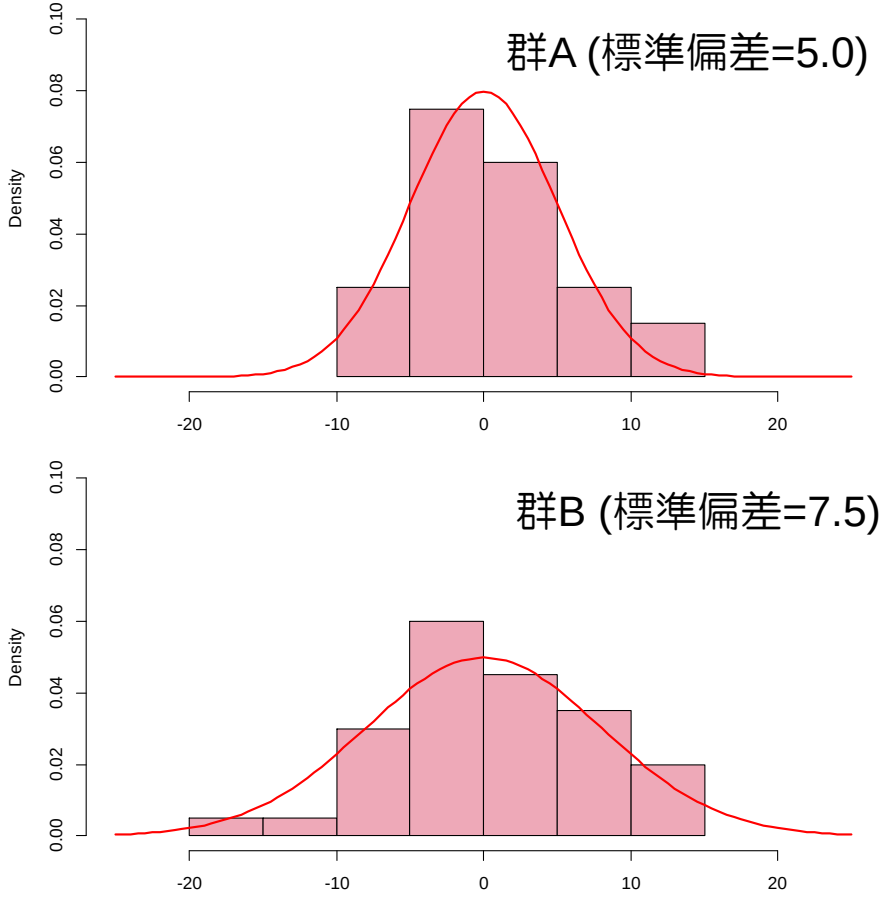
標本サイズ(N)が多いほど有意になりやすくなる．つまり，調査規模が大きくなると，いずれの正規性検定でも有意になってしまう(つまり，大規模データでの正規性検定では非正規であると判断されてしまう).

等分散性の検定を用いてt検定とWelch検定を選択するのはおかしい

統計の教科書では，等分散性の検定を用いてt検定とWelch検定を取捨選択するような説明をしているものがある．ただし，等分散性の検定による取捨選択については，先ほどの正規性と同様に，サンプルサイズが増えれば有意になりやすくなる問題がある．

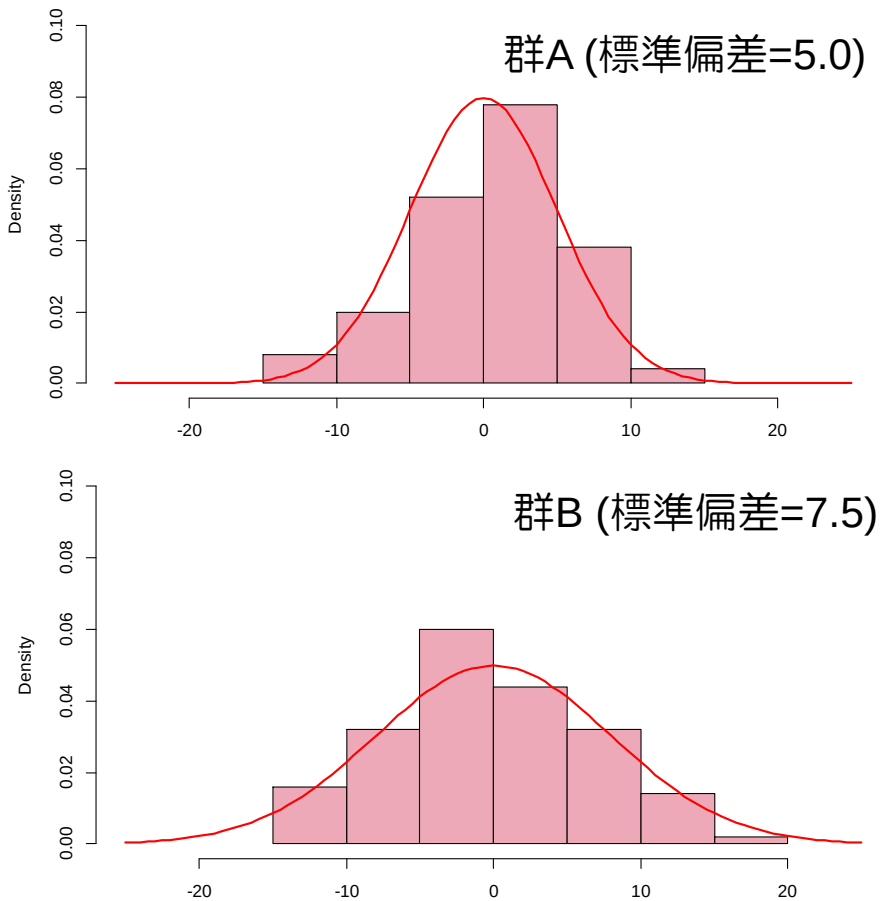
各群のサンプルサイズ = 40

等分散性の検定のp値=0.164



各群のサンプルサイズ = 60

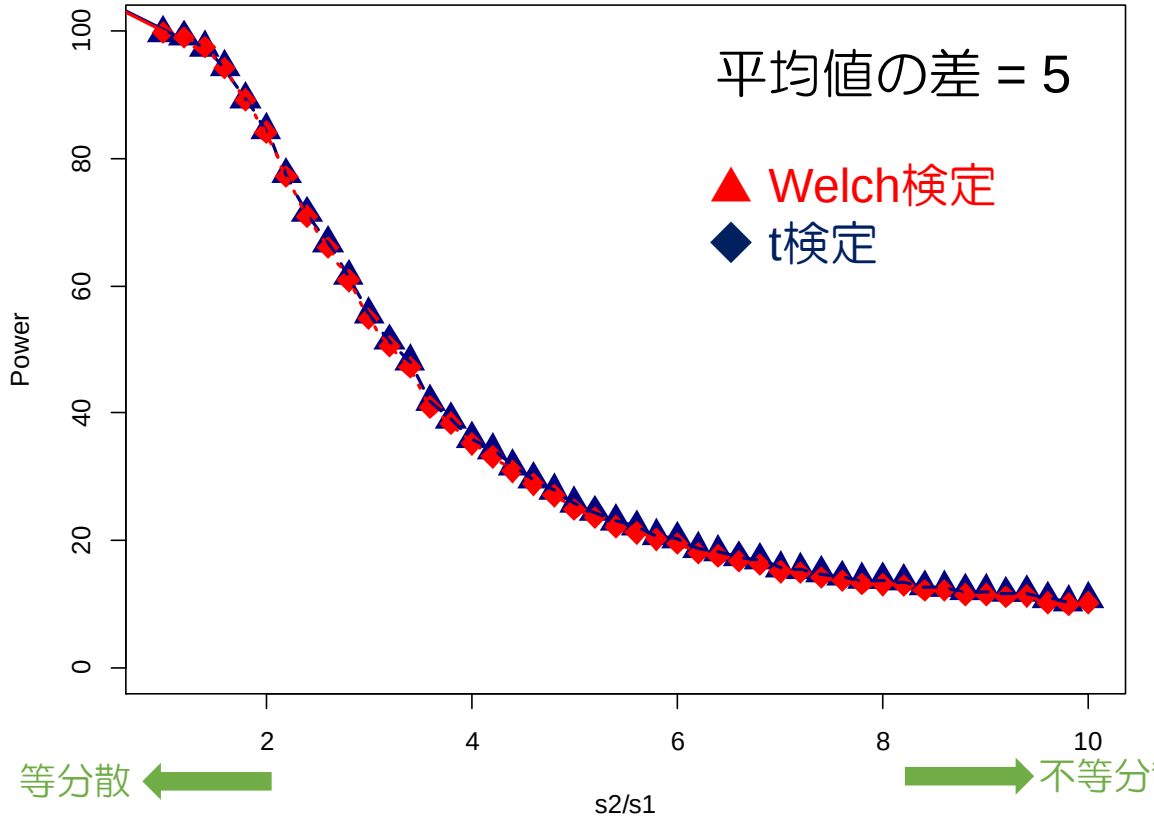
等分散性の検定のp値=0.006

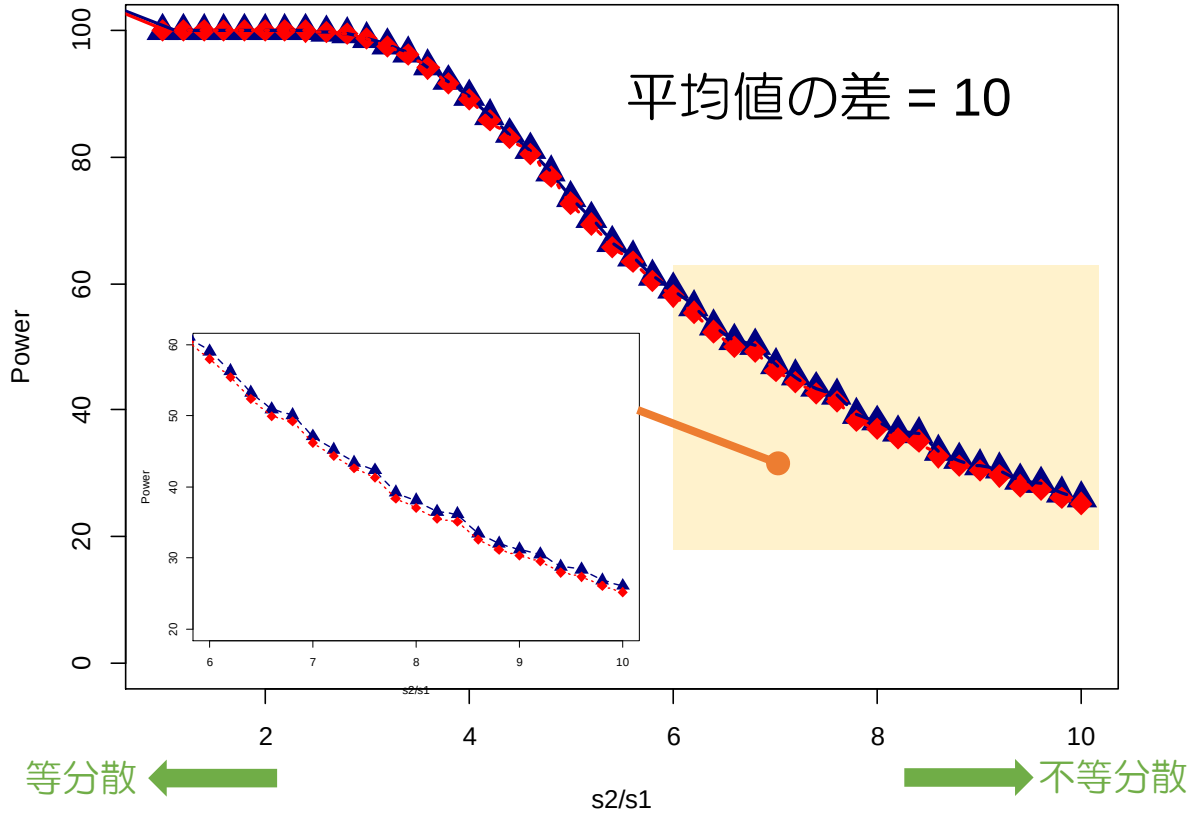


同じ母集団であっても，サンプルサイズの違いによって検定手法を変えるというのは，そもそもおかしいのでは？

不等分散のときに t 検定とWelch検定の結果に違いが出るのか？

Welch検定は，未知の不等分散のもとでの母平均の比較(Behrens-Fisher問題)に対する近似解の一つであるが，そもそもWelch検定の利用には統計学者からの悲観的な意見が多い．





上の二つのグラフは，10,000回のシミュレーションにおいて，t検定とWelch検定が有意になった割合を表している．X軸が標準偏差の比率であり，左端が等分散を表す．右側に行くほど不等分散になる．また，Y軸は有意になった割合(パーセントで表示，実質検出力という)．実質検出力が高い手法のほうが良い方法であると判断される．
不等分散であっても，t検定とWehch検定の性能はほぼ変わらない(しいていえば，僅かにt検定の方が高い)．

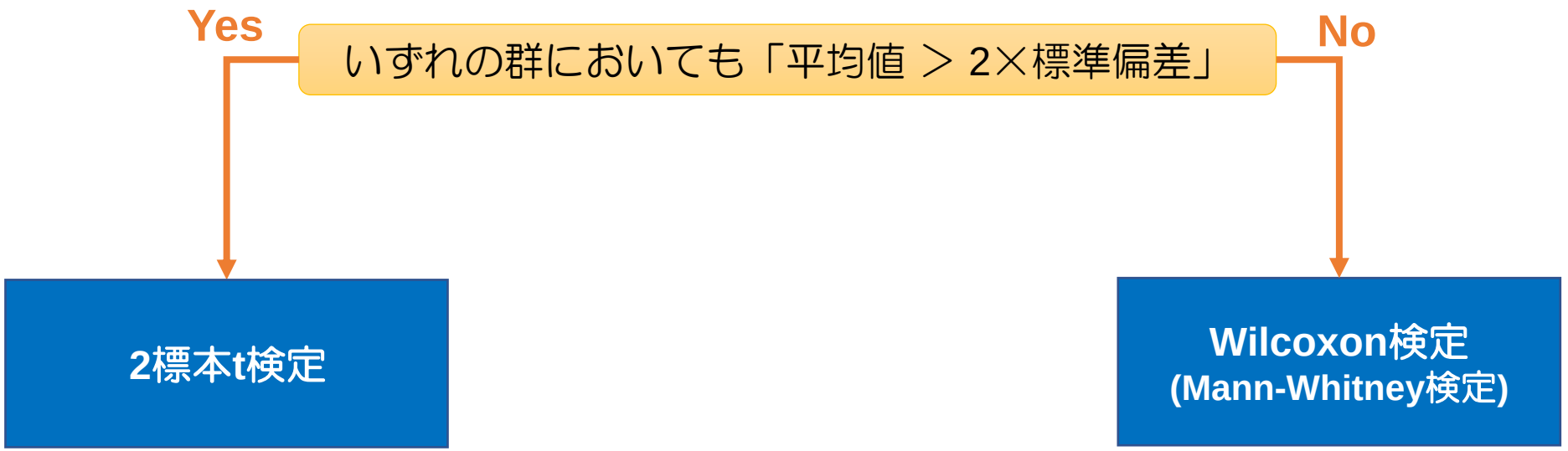
どのように考えるべきか？

丹後(2018)は，平均値と２×標準偏差を比較することで検定方法を選択することを提案している．この選択方法では，

(1) Welch検定は選択肢にない．

(2) 実際のデータ解析では，分散(標準偏差)が異なる場合には，往々にして分布形状も異なることが多い．

といった考え方に由来する．

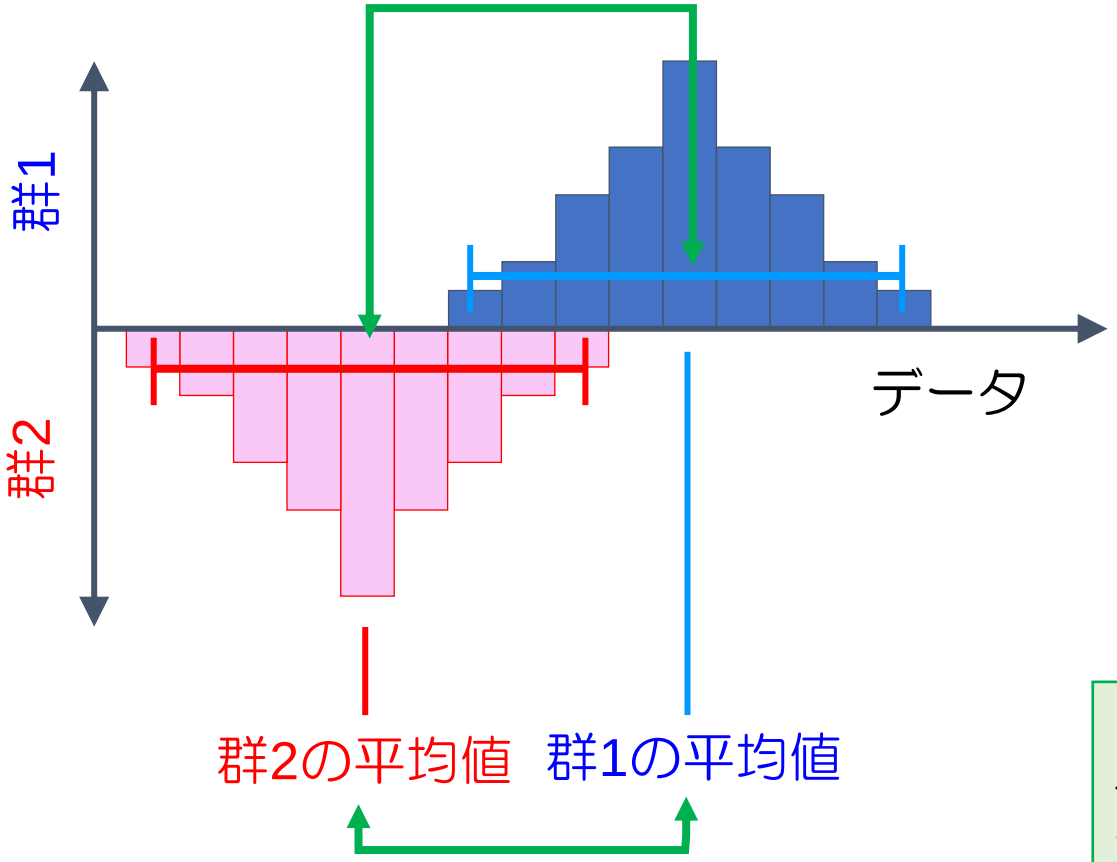


このように考えると非常にシンプル！！

丹後敏郎：新版 統計学のセンス ―デザインする視点・データを見る目―, 2018, 朝倉書店.

注意点：2標本t検定とWilcoxon検定(ノンパラメトリック検定)の違い

Wilcoxon検定では，相対的ない
い関係を比較している．



2標本t検定では平均値(それぞれの群を
代表する値)を比較している．

■ **2標本t検定で有意であるということ**

「2群間の平均値が違う」と解釈できる．

■ **Welch検定で有意であるということ**

「2群間の相対的な位置関係が違う」と解釈できる．

中央値が違うとは言っていない．ノンパラメトリッ
ク検定を用いた場合に，中央値を用いるのは，その
他に群の代表値として用いるものがないため．

上記の理由のほかにも，分散分析や回帰分析(いわゆ
る多変量解析)においても，平均値で解釈される．た
め，可能な限り2標本t検定を用いるほうが良い．

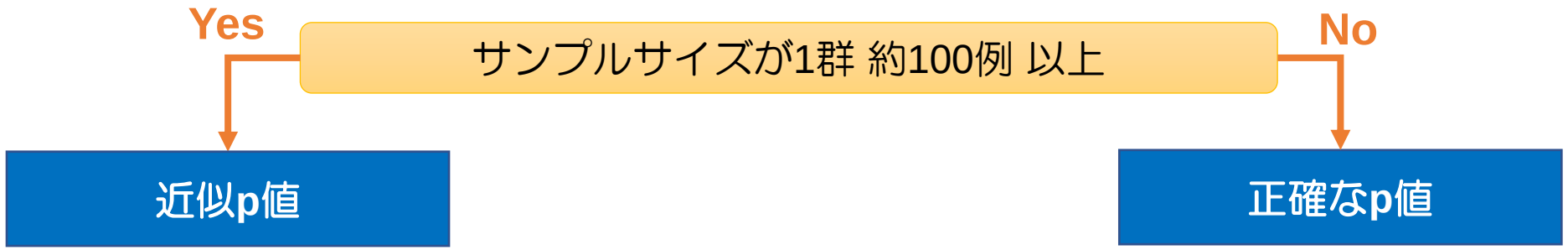
注意点：Wilcoxon (Mann-Whitney)検定のp値の計算方法には数種類存在する

t検定では検定統計量がt分布に従うことが数学的に証明できるため，p値を正しく計算できる．一方で，Wilcoxon検定はp値を計算するための計算式を導出することはできない，

そのため，Wilcoxon検定では以下のいずれかによってp値を計算する．

- 近似式を用いる場合：
  - 正規分布を用いて近似計算を行う．
  - カイ2乗分布を用いて近似計算を行う．
    - ➡ 近似式の計算の理屈は，ほぼ同じだが少しだけ値が違う
- コンピュータを用いて正確なp値(exact p-value)を計算する．

- サンプルサイズが大きければ，近似式でのp値と正確なp値はほぼ同じになるが，サンプルサイズが小さい場合には違いが生じる(近似式での近似精度が悪くなる)．
- 一方で，正確なp値はサンプルサイズが大きいと計算が膨大になり，PCがフリーズする(ときがある)．

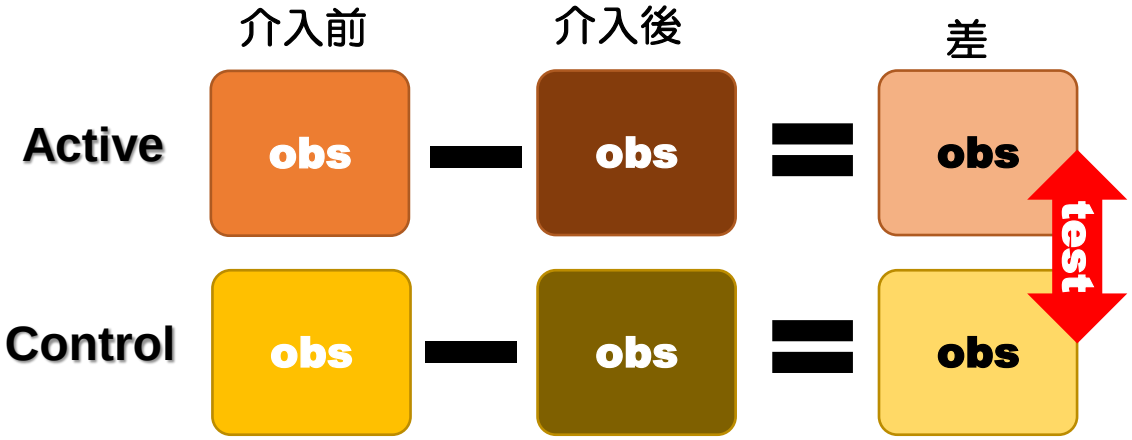


QA.3：変化量のデータを解析したら，レフェリーから前値の影響を指摘されたのですが．．．

治療前後での変化量の比較を行う研究を論文として投稿したら，治療前の値の影響を調整するようにとの意見をいただきました．どうしたらよいでしょうか？

共分散構造分析 (ANCOVA; ANalysis of COVAriance)

☐ 介入前後での変化量 or 変化率を比較する



ActiveとControlの差を評価しても，個人差(介入前による影響)を排除できない．

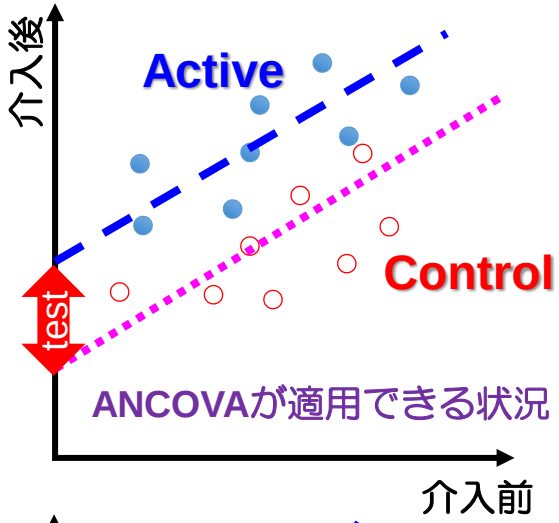
そのため，介入前の値を共変量としたもとで，ANCOVAを用いて解析する．

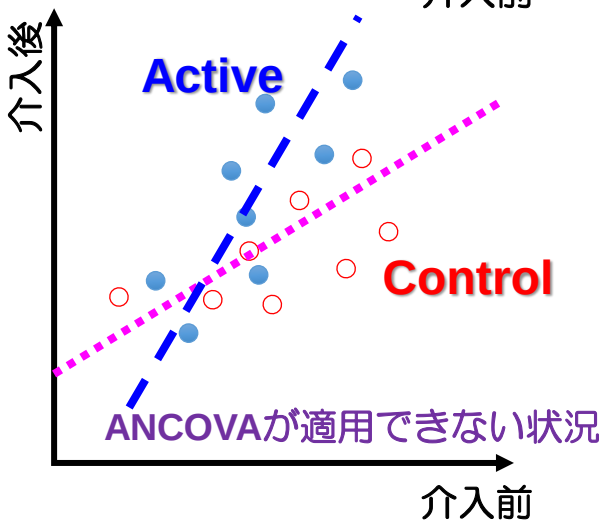
☐ 介入前の共変量の影響を排除して解析する

共変量がアウトカムに影響を及ぼす場合，群間での偏りにより，正しく評価できない．例えば，看護師のワークライフバランスに対する満足度調査において，勤務年数(10年以上/未満)で比較する研究において，被験者の勤務病棟が影響するかもしれない．そのような場合には，勤務病棟を共変量として調整したほうが良いかもしれない．

このような場合に，ANCOVAを用いて解析する．

- 共分散分析(ANCOVA：ANalysis of COVAriance)とは，2群のそれぞれに対して，説明変数Xに介入前，応答変数Yに介入前としたときの回帰直線を引き，その切片の差を比較する方法である.
- ただし，切片を比較するうえにおいて，傾きが有意に異なる場合には評価が不可能になるため，事前に傾きが異なるか否かを検定する必要がある.





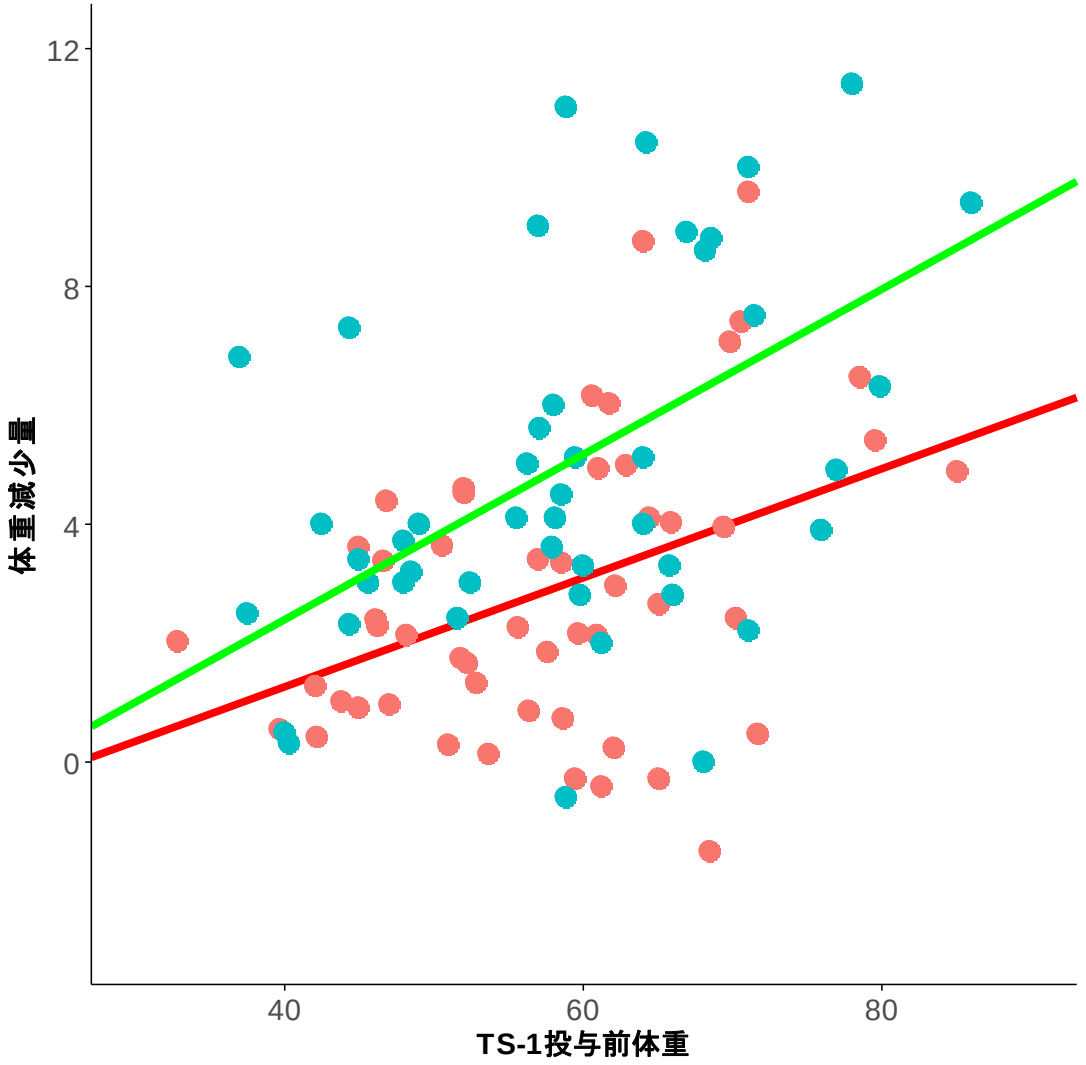
[STEP.1] Active，Controlそれぞれに回帰直線を引く.
[STEP.2] 2種類の検定を考える
[STEP.2-1] 傾きの違い(交互作用)を比較する(介入前値の違いによる介入後値の変化が同じであるか検討する)**[Test.1]**
→ 有意な場合には治療前後での比較は不可能 (左下の図)

具体的には，
「従属変数」=「群」＋「共変量」＋「群」×「共変量」
の重回帰分析を行い，「群」×「共変量」での交互作用のp値を評価する.

[STEP.2-1] 切片を比較する(介入前値の違いによる介入後値の変化が同じであるか検討する)**[Test.2]**
→ 有意な場合には変化に違いがあると解釈 (左上の図)

具体的には，
「従属変数」=「群」＋「共変量」
の重回帰分析を行い，「群」でのp値を評価する.

適用例



胃癌患者に対する術後のTS-1補助化学療法による体重減少の抑制を意図した成分栄養剤服用の有効性を検討している.

左の散布図において，●が成分栄養剤を服用してもらった患者であり，●が服用しなかった患者である.

共分散分析では，TS-1投与前の体重を共変量としたもとで，術後1年での体重減少を従属変数に用いている.

共変量×群の交互作用の評価ではp値が0.288であることから，交互作用があるとは言えない(もし，交互作用の評価で有意であれば，共分散分析が使えないとして評価を終了する).

その後の共変量と群を用いた回帰分析における群に対するｐ値は<0.001であった．すなわち，成分栄養剤を服用することで，TS-1による体重減少を抑制することができた.

おまけ：無作為化比較試験における量的アウトカムの評価

量的アウトカムを用いた無作為化比較試験において，割付調整因子による調整はどのように行えばいいでしょうか？

■ 2標本t検定と回帰分析の関係

回帰分析：アウトカム ＝ 回帰係数 ✖ 群指標

回帰係数に対する検定と2標本t検定は同じもの

■ 共変量を調整したもとでの2群間の比較 (ANCOVAでも同じですが．．．)

回帰分析：アウトカム ＝ 回帰係数 ✖ 群指標 ✚ 回帰係数 ✖ 因子1 ✚ ・・・

- 回帰係数に対する検定を用いれば調整された検定になる．
- このときの回帰係数は調整後の群間差を表す．

ちなみに，

- 2値アウトカムの場合にはロジスティック回帰 (回帰係数の指数値は調整オッズ比)
- 生存時間アウトカムの場合には比例ハザード・モデル(回帰係数の指数値は調整ハザード比)

においても同じことが言えます．

QA.4：カイ2乗検定とFisherの正確検定の使い分けについて

下表はある疾患に対する2種類の治療法の有効率を評価したときのクロス集計表である．治療法によって有効/無効に違いがあるか．

	有効	無効	計
治療A	204 **(51.1%)**	195 **(48.9%)**	399
治療B	158 **(39.4%)**	243 **(60.6%)**	401
計	362 **(45.2%)**	438 **(54.2%)**	800

帰無仮説：治療法によって有効/無効に違いがない．

対立仮説：治療法によって有効/無効に違いがある．

クロス集計表の解析において用いられる検定は，

- カイ2乗検定
- Fisherの正確検定 (Fisherの直接確率)

である．なお，今回の事例のように有効率を比較する場合には，母比率の差の検定というものがある．他方，母比率の差の検定は，正確にカイ2乗検定に一致する．

カイ2乗検定

仮説検定：帰無仮説の位置からどれだけ離れているか(どれだけ違っているか)を確率(p値)で表す

帰無仮説でのクロス集計表を作らないといけない.

	有効	無効	計
治療A	204	195	399
治療B	158	243	401
計	362	438	800

研究の結果

	有効	無効
治療A	204	195
治療B	158	243

乖離

期待度数(帰無仮説での位置)

有効	無効
$\frac{399\times362}{800}=180.5$	$\frac{399\times438}{800}=218.5$
$\frac{401\times362}{800}=181.5$	$\frac{401\times438}{800}=219.5$

各セルに対して，度数の差の2乗値を計算して，その和をとったものが帰無仮説の位置からの離れ度合いである(カイ2乗値という).

その結果，**p値は0.001**なので，有意でだった．つまり，**治療Aと治療Bの有効率に違いが認められた．ちなみに，Fisherの正確検定でのp値は0.001であり結果に大差がない.**

Fisherの正確検定

	病変あり	病変なし	計
NBI	12 **(54.5%)**	10 **(45.5%)**	22
白色光	3 **(20.0%)**	12 **(80.0%)**	15
計	15 **(40.5%)**	22 **(59.5%)**	37

	病変あり	病変なし	計
NBI	**12** 22-12 →	10	**22**
白色光	3 ↓ 15-2	12 ↓ 22-10	**15**
計	**15**	**22**	**37**

2×2クロス集計表では，周辺度数(青色の部分)を固定してしまうと，1個のセル(赤色)が決まれば，残りのセルは全て決まる．

こうなる確率 = 0.0000
a×d – b×c = -330

	あり	なし
NBI	0	22
白色光	15	0

こうなる確率 = 0.0000
a×d – b×c = -293

	あり	なし
NBI	1	21
白色光	14	1

こうなる確率 = 0.0000
a×d – b×c = -256

	あり	なし
NBI	2	20
白色光	13	2

こうなる確率 = 0.0001
a×d – b×c = -219

	あり	なし
NBI	3	19
白色光	12	3

こうなる確率 = 0.0011
a×d – b×c = -182

	あり	なし
NBI	4	18
白色光	11	4

こうなる確率 = 0.0084
a×d – b×c = -145

	あり	なし
NBI	5	17
白色光	10	5

こうなる確率 = 0.0399
a×d – b×c = -108

	あり	なし
NBI	6	16
白色光	9	6

こうなる確率 = 0.1172
a×d – b×c = -71

	あり	なし
NBI	7	15
白色光	8	7

こうなる確率 = 0.2197
a×d – b×c = -34

	あり	なし
NBI	8	14
白色光	7	8

こうなる確率 = 0.2659
a×d – b×c = 3

	あり	なし
NBI	9	13
白色光	6	9

こうなる確率 = 0.2074
a×d – b×c = 40

	あり	なし
NBI	10	12
白色光	5	10

こうなる確率 = 0.1028
a×d – b×c = 77

	あり	なし
NBI	11	11
白色光	4	11

こうなる確率 = 0.0314
a×d – b×c = 114

	あり	なし
NBI	12	10
白色光	3	12

こうなる確率 = 0.0056
a×d – b×c = 151

	あり	なし
NBI	13	9
白色光	2	13

こうなる確率 = 0.0005
a×d – b×c = 188

	あり	なし
NBI	14	8
白色光	1	14

こうなる確率 = 0.000
a×d – b×c = 225

	あり	なし
NBI	15	7
白色光	0	15

試験で得られた状況よりも極端な状況(オッズ比が大きくなる状況)のクロス集計表になる確率の総和を計算する．その結果，p値は

0.0000 + 0.0000 + 0.0000 + 0.0001+ 0.0011 + 0.0084 + 0.0314 + 0.0056 + 0.0005 + 0.0000 = 0.0471

であるので，有意水準α=0.05のもとで有意である．したがって，NBIと白色光には病変の検出に差異が認められた．

カイ2乗検定とFisherの正確検定の使い分け：まとめ

	病変あり	病変なし	計
NBI	12 (54.5%)	10 (45.5%)	22
白色光	3 (20.0%)	12 (80.0%)	15
計	15 (40.5%)	22 (59.5%)	37

カイ2乗検定

p値 = 0. 0783

有意差が認められない

Fisherの正確検定

p値 = 0.0471

有意差が認められる

カイ2乗検定では，p値を近似で計算する．そのため，
(1) 標本サイズが少ない場合 (50未満)
(2) どこかのセルの度数が小さい場合 (5未満)
には，近似精度が悪くなる．そのため，Fisherの正確検定を用いるほうが良い．一方で，標本サイズが大きくなると，Fisherの正確検定は計算できなくなる(パソコンがフリーズする)ので，カイ2乗検定を推奨する．なお，標本サイズが増えると，カイ2乗検定のp値とFisherの正確検定のp値は一致する．

QA.5：多重比較っていったい何ですか？

「下手な鉄砲も数打てば当たる」では評価にならない

A　B　C

3種類のコレステロール治療の比較
A vs. B，A vs. C，B vs. C **(3回の比較)**

有意水準α=0.05とは，違いがないのに(H_0が正しい)のに，違いがある(H_1が正しい)と誤ってしまう確率を表している．

➡ **つまり，コレステロール治療の効果に違いがないのに，20回検定すると，1回は有意差が出てしまう(H_1と判断してしまう)．**

➡ **3回比較するということは，その可能性が増しているといえる．**

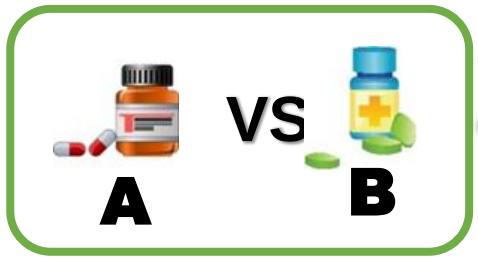
これを多重性という

多重比較が生じる場面

- 3群以上での比較の場面における対比較
- 経時データにおける時点毎の群間比較　など

多重比較を数学的に考える

有意水準とは「有意差がないのに有意差があると誤る確率(第1種の過誤)」をあらわす



検定

有意水準 α

有意差がないことを正しく判断できる確率：Pr_1

$Pr_1 = 1-\alpha$

検定

有意水準 α

有意差がないことを正しく判断できる確率：Pr_2

$Pr_2 = 1-\alpha$

検定

有意水準 α

有意差がないことを正しく判断できる確率：Pr_3

$Pr_3 = 1-\alpha$

3回の比較のすべてで，有意差がないことを正しく有意差がないと判断できる確率：Pr_0

$Pr_0 = Pr_1 \times Pr_2 \times Pr_0 = (1-\alpha)^3$

3回の比較のなかで1度でも，有意差がないのに有意差があると誤る確率

$1 - Pr_0 = 1-(1-\alpha)^3$

1回の検定の有意水準αを次のように考える：

α=0.05(1回の検定において有意差がないのに有意差があると誤る確率を0.05とする)

3回の検定を繰り返した場合，3回の検定の中で1度でも有意差がないのに有意差があると誤る確率は，

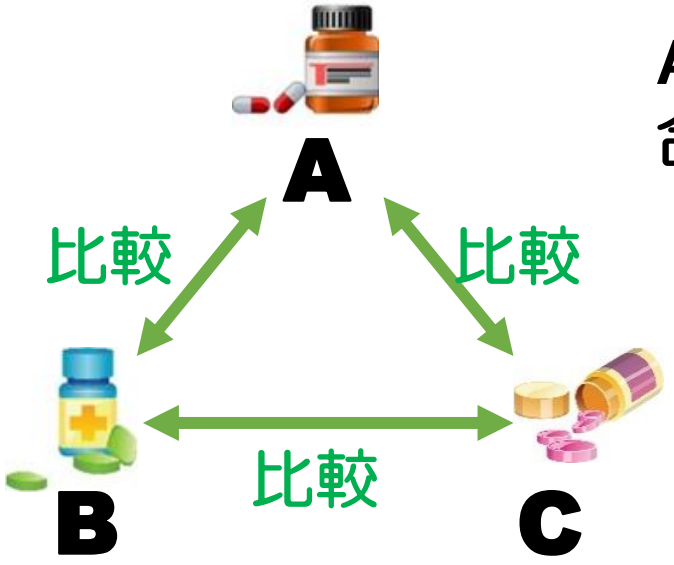
$1-(1-\alpha)^3 = 1-(1-0.05)^3 = 1-(1-0.05)^3 = 0.142$

となり，誤りの確率が増大してしまう．

多重比較の種類

多重比較の方法には，(1) p値(有意水準α)に基づく方法，(2) 正規分布(or ノンパラメトリック)に基づく方法がある

P値に基づく方法(1)：Bonfferoniの方法



A群，B群，C群の対比較の場合で全体での有意ｐがαの場合には

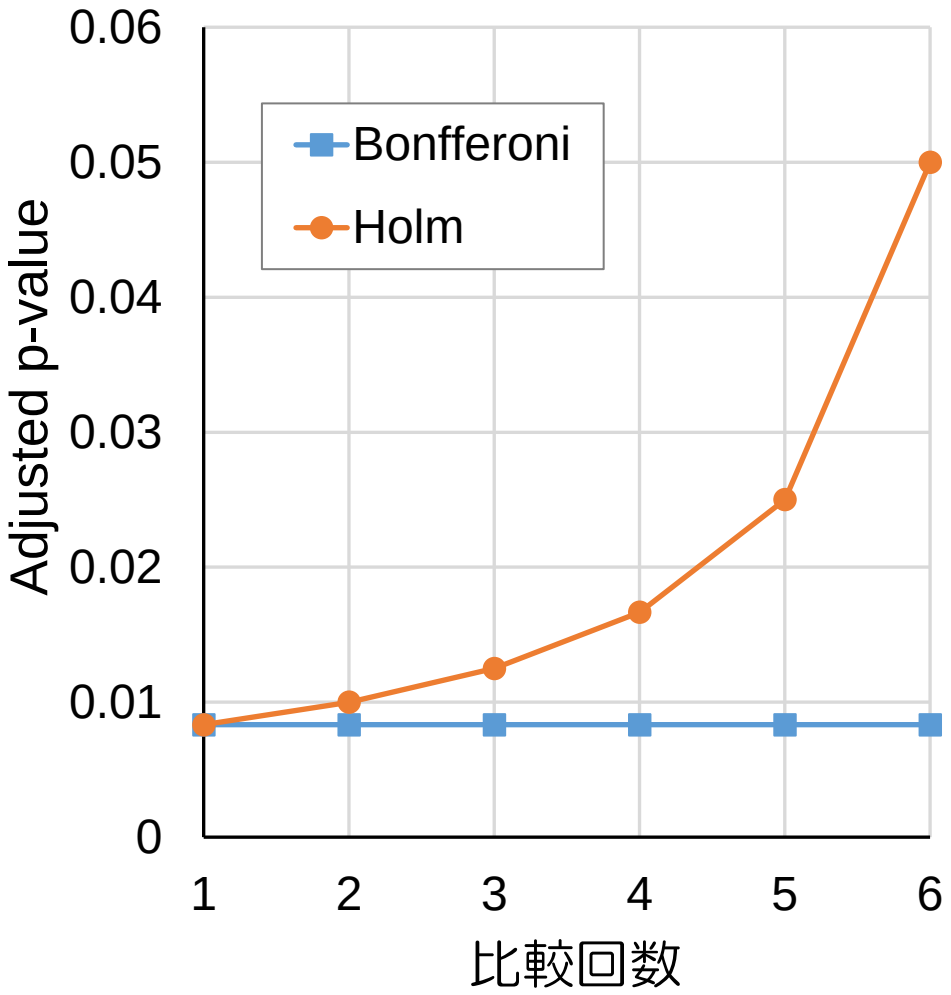
- **A群　対　B群　➡　有意水準α/3と比較(or p値を3倍)**
- **A群　対　C群　➡　有意水準α/3と比較(or p値を3倍)**
- **B群　対　C群　➡　有意水準α/3と比較(or p値を3倍)**

有意水準αを比較回数で割る(or p値を比較回数で掛ける)方法がBonfferoniの方法である．多重比較が簡単なため，最も用いられる方法の一つである．

P値に基づく方法(2)：Holmの方法

最小のp値から並べ替え，シーケンシャルに比較する方法.
いま，6回の比較のp値が次のように与えられているとする.

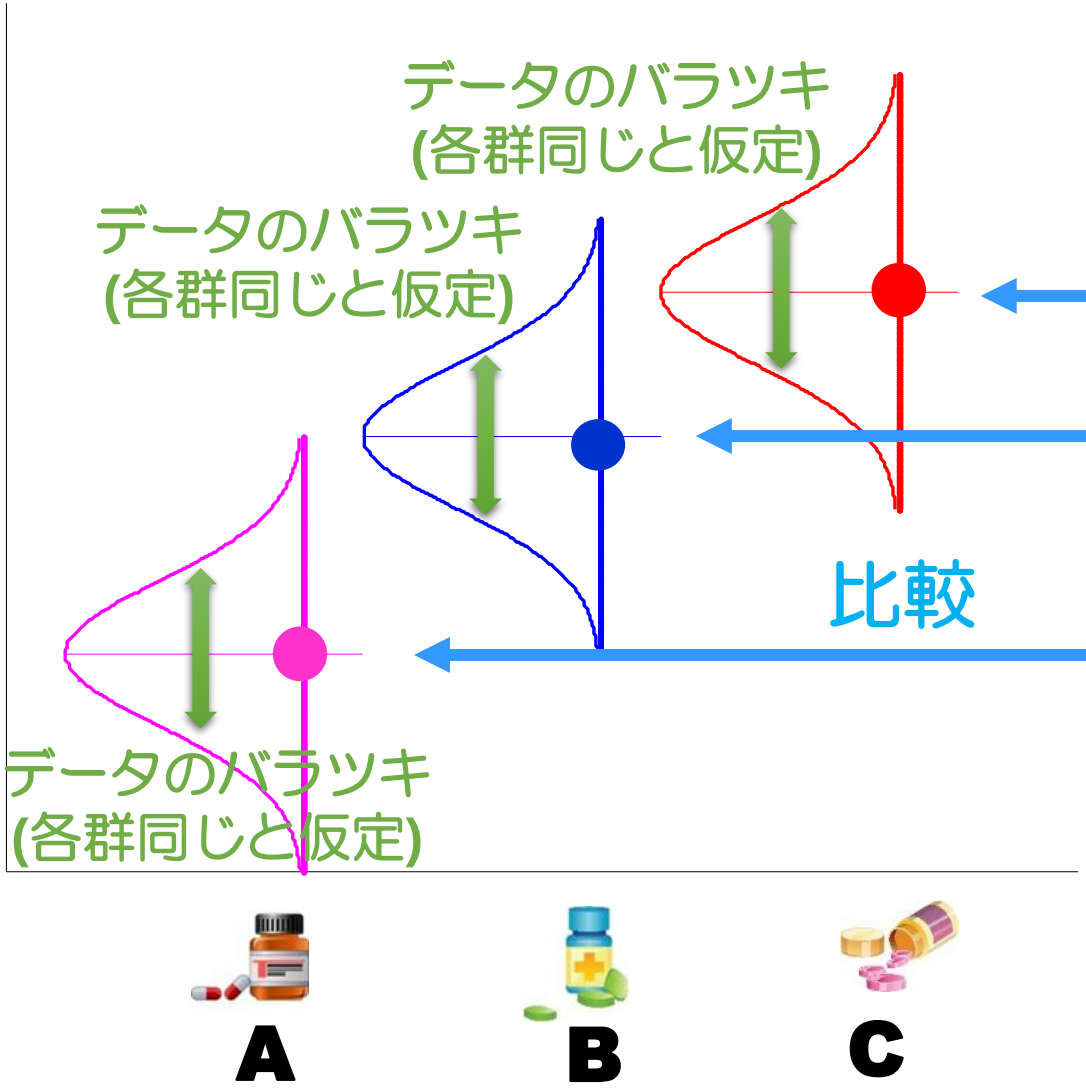
0.001　0.006　0.011　0.012　0.032　0.045



順位	p値	Bonfferoni法		Holm法	
		比較α	判定	比較α	判定
1	**0.001**	0.05/6=0.008	有意	**0.05/6=**0.008	有意
2	**0.006**	0.05/6=0.008	有意	**0.05/5=**0.010	有意
3	**0.011**	0.05/6=0.008	非有意	**0.05/4=**0.013	有意
4	**0.012**	0.05/6=0.008	非有意	**0.05/3=**0.017	有意
5	**0.032**	0.05/6=0.008	非有意	**0.05/2=**0.025	非有意
6	**0.045**	0.05/6=0.008	非有意	**0.05/1=**0.050	非有意

以降はすべて有意でない

正規分布に基づく方法(1)：Tukeyの方法（連続変数のみに利用できる）

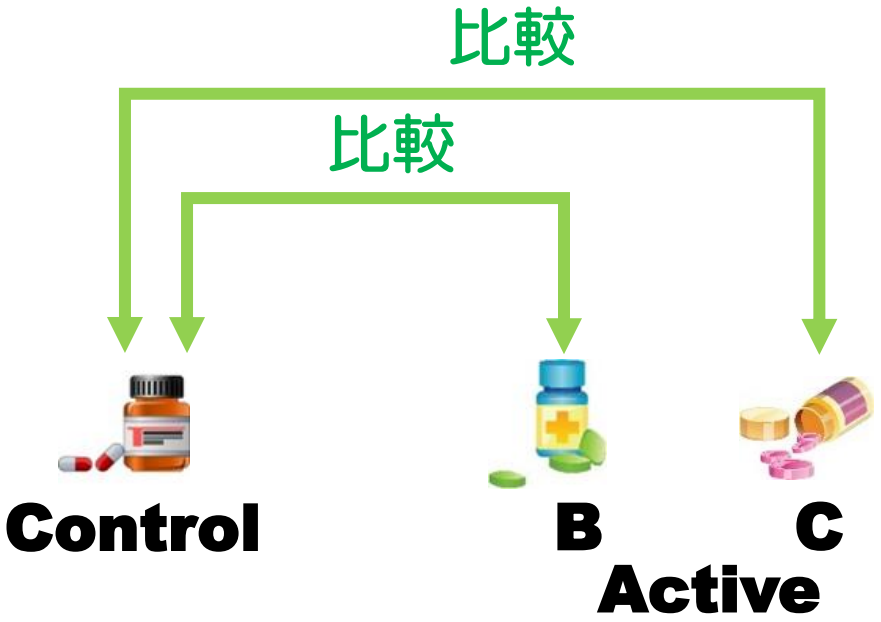


一元配置分散分析において有意だった場合には，Tukeyの方法でもいずれかの対比較において有意になっている．

データのバラツキよりも，ペアの平均値の差が大きいか否かを評価

Tukeyの方法のノンパラメトリック版には，Steel-Dwassの方法がある．

正規分布に基づく方法(2)：Dunettの方法（連続変数のみに利用できる）



Controlとの比較のみを実施する多重比較がDunnettの方法である.

Tukeyの方法のノンパラメトリック版には，Steelの方法がある.

QA.6：必要症例数のことですが．．．

〇〇研究を計画しています．どの程度の症例数が必要ですか？

■ 観察研究の場合

・仮説がある程度はっきりしている場合 (例えば，肺がんの原因としてタバコに焦点を当てている)

➡ 臨床試験と同じような流れで症例数を計算する (解析自体も臨床試験と類似).

・仮説がはっきりしない場合 (例えば，影響要因を探るなど．．．)

➡ できるだけたくさん集める (研究計画書には，見込み例数を記載する)

収集されたデータの症例数によって解析方法の幅が違う (あくまでも個人的意見).

— 30例程度未満：単変量解析ぐらい (20例を下回ると検定も厳しい...).
— 200例程度未満：サブグループ解析・多変量解析(説明変数は例数／10ぐらい).
— 200例程度以上：傾向スコア分析
— 500例程度以上：アウトカムに基づくグループ分け (ステイジ分類みたいなもの)
— 10,000例程度以上：レスポンダー探索など (とりあえず，何でもできる).

■ 臨床試験の場合

・原則として，臨床的仮説に基づく統計学的な必要症例数を計算する (パイロット試験の場合は実施可能性).

症例数設計の一例：単アーム試験による事例

論文での記述例

Phase II study of trastuzumab in combination with S-1 plus cisplatin in HER2-positive gastric cancer (HERBIS-1)

The required sample size was estimated based on a threshold RR of 35% and an expected RR of 50%, 80% power, and an alpha value of 0.1 (one-sided) using the binomial test. Given 2% of ineligible patients, the target sample size was determined to be at least 50 patients.

(Kurokawa et al., British Journal of Cancer, 110, 1163-68, 2014)

試験デザイン：Phase II Study (単アーム)

主要評価項目：奏効割合 (RR: Response Rate)

閾値奏効割合：35%,　　**期待奏効割合**：50%

臨床的判断

第1種の過誤(type I error, 有意水準)：α = 0.1

1ー第2種の過誤(1-type II error, 検出力)：1-β = 0.8 (80%)

統計的判断
必要症例数 = 50

臨床的仮説における留意点

閾値(帰無仮説)には根拠が必要(とくに単群試験では重要)

無作為化比較試験では，群間差がないことが帰無仮説になるので，とくに留意することはない．ただし，単群試験では，ヒストリカル・コントロールは，「本試験の結果を何と比較するのか」が明確にならないため，**根拠ある仮説が必要**

HERBIS-1では，ToGA試験(Y-J Bang, et al. *The Lancet*, 28, 687 - 697, 2010.)の結果を参考にしている．ToGA試験では，化学療法群(5-FU /CDDP あるいはCapecitabine /CDDP)での**奏効率が35%**であった．

閾値とは，臨床試験の評価対象となる数値である．本試験の場合には
「既存の化学療法の奏効率35%を上回るか否か」
を仮説検定で明らかにする．

期待値(対立仮説)には実現可能性及び臨床的有用性に基づいて設定

ToGA試験では，(1) Trastuzumab併用群の**奏効率が47%**であること，(2) HER2による選択基準(IHC 3+ or FISH positive + IHC 2+)，(3) 本邦の状況を鑑みて，期待奏効率を50%と設定している．

期待値とは，臨床試験の結果得られる効果に関して規定する．それは，
「検定したときに有意になるエンドポイントの最小値が期待値」
となる．したがって，エンドポイントの**目標値**になる．

第1種の過誤と第2種の過誤

	検定で評価	標本サイズで規定
判　断	帰無仮説H_0 (真) [対立仮説H_1 (偽)]	帰無仮説H_0 (偽) [対立仮説H_1 (真)]
H_0を棄却 [H_1を受容]	第1種の過誤 α (有意水準)	検出力 1−β
H_0を受容 [H_1を棄却]	1−α	第2種の過誤 β

- αとβがともに小さい検定が望ましいが，標本サイズを固定したもとで両方の確率を同時に小さくするのは不可能である．例えば，αを小さくすればβが大きくなる．
- ただし，標本サイズを大きくすれば，βは小さくなる (つまり，症例数を設定するとは，ある有意水準αのもとで，第2種の過誤βを許容範囲まで下げるように計画することを意味する)．

臨床試験のデザインのための3ステップ

STEP.1：主要評価項目(主たるアウトカム)を考える

- 本研究におけるクリニカルクエスチョンに応えられるものか(**妥当性**)？
- 当該分野においてAgreeされるものか(**信頼性**)？
- エンドポイントを取得できるか(**実施可能性**)？
  ー例えば，抗がん剤治療において標的病変がなければ奏効割合が測定できない等

STEP.2：仮説を決定する

- 無作為化比較試験の場合には非劣勢・優越性・同等性を検討する.
- 単アーム試験の場合には閾値を文献等から決定する.
- 本プロトコル治療における主要評価項目での有用性を検討する.

— 量的アウトカムの場合：群間差(各群の平均)および共通の分散(標準偏差)を設定する.
— 2値アウトカムの場合：各群の関心のある事象(例：奏効例)の割合を設定する.
(比較試験の場合にはORでも可)
— 生存時間の場合：各群のMST (OR 年次生存割合)，登録期間，フォローアップ期間
(比較試験の場合にはHRでも可)

STEP.3：有意水準・検出力を決定する

単群試験：$\alpha = 0.05$ (or 0.10) [片側対立仮説]，$1-\beta \geqq 0.80$

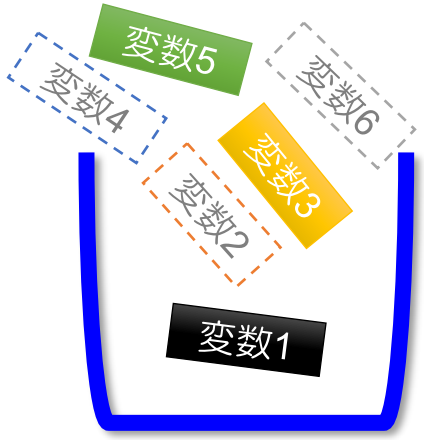
第III相試験(RCT)[優越性]：$\alpha = 0.05$ [両側対立仮説]，$1-\beta \geqq 0.80$ (or 0.90)

第III相試験(RCT)[非劣性]：$\alpha = 0.025$ (or 0.50) [片側対立仮説]，$1-\beta \geqq 0.80$ (or 0.90)

QA.7：多変量解析の変数選択について

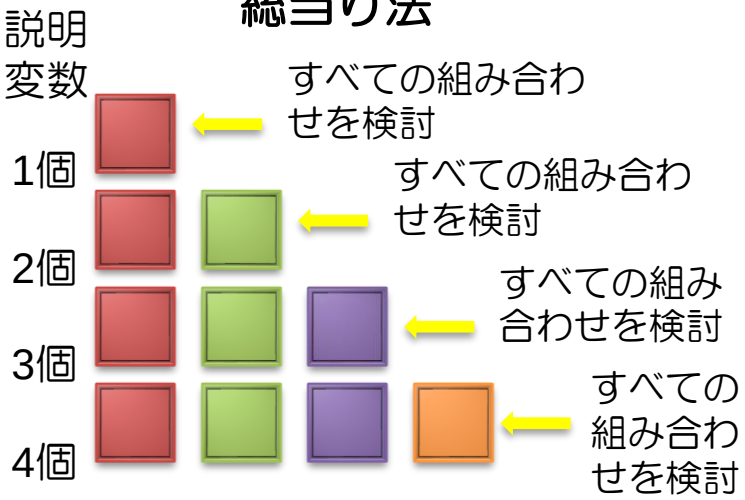
統計モデル

多変量解析では，たくさんの説明変数(共変量)を入れるほど情報がたくさんになるので，良い統計モデルになるということ？



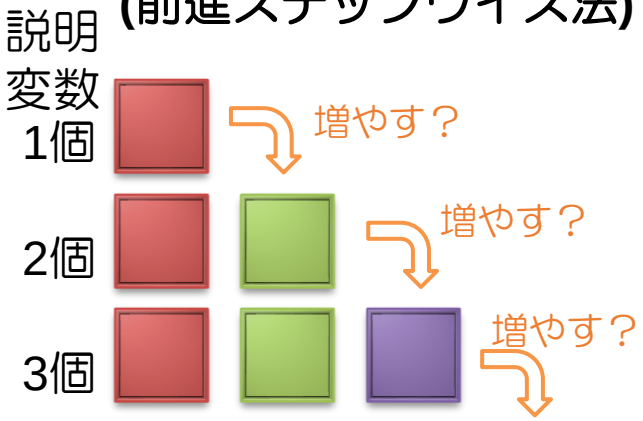
必ずしもそうではない．説明変数が多いほど不必要な変数は単なるノイズでしかないわけだし，また，多重共線性の問題などがある．そのため，必要なものだけで統計モデルをつくることが推奨されることが多い．そのための方法が変数選択である．

総当り法



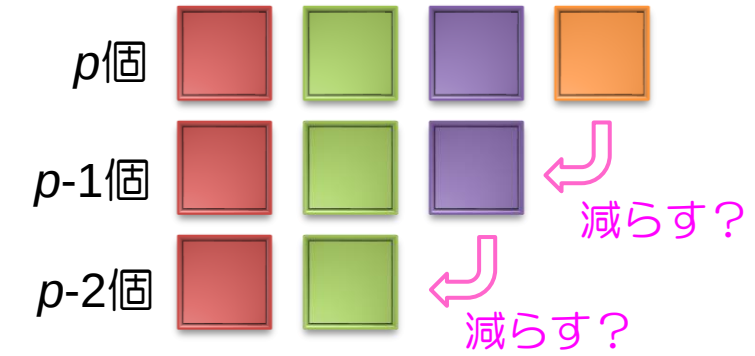
説明変数が1個,2個,...のそれぞれのパターンを計算しその中から最適なものを選ぶ

変数増加法 (前進ステップワイズ法)



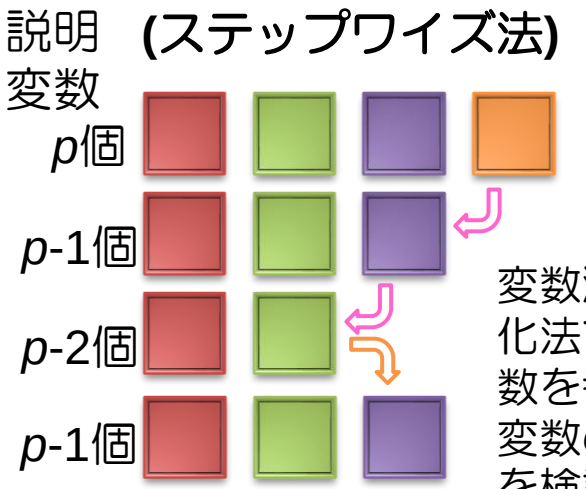
説明変数が1個の場合からスタートして，変数を追加したほうが良ければ増やし，そうでなければ変数の追加をしない．

変数減少法 (後退ステップワイズ法)



全ての説明変数からスタートして，変数を減らしても影響がなければ減らし，そうでなければ変数の削除をしない．

変数増減法 (ステップワイズ法)



変数減少法からスタートするが，変数増化法では変数が削除された場合にその変数を考慮しなかったが，変数増減法では変数の削除と削除した変数の追加の両方を検討しながら各ステップを進める．

■ 評価基準

評価基準には(1) 検定を用いる場合，(2) 情報量規準を用いる場合がある．

— 検定を用いる場合は，標本サイズに依存するため，推奨されない．

— 情報量規準には，AIC(赤池の情報量規準)，あるいはBIC(Bayes流情報量規準)がある．どちらでも良いが，BICのほうが選択される変数が少ないことが多い．

変数選択の要件

(1) 評価したい要因は変数選択に強制的に加える

ランダム化比較試験の結果を評価する場合，治療群を表す共変量を含まなければ意味をもたない．つまり，このような場合には，背景因子などの他の共変量を調整したうえで治療群(評価変数)を調べることに意義がある．

(2) 変数増加法の落とし穴

標本サイズが小さい場合に，変数増加法を用いて変数選択を行う場合，結果の解釈が困難なモデルを選択することがしばしばある．また，本当は必要な共変量が取り込まれる前に変数選択が終了する場合がある．

(3) 多数の共変量(項目)がある場合の留意点

医学系研究では，多数の調査項目(共変量)を評価に用いることは少なくない．このような場合には，全ての共変量を用いて変数選択を行うのではなく，事前スクリーニングを行うことが推奨される．事前スクリーニングでは，共変量毎に単変量解析(1個の共変量による回帰モデルを推定する)を実施し，その回帰係数に対する検定(回帰係数が0であるか否かを評価する検定)のp値や回帰係数(オッズ比，ハザード比)を用いる．

(4) 欠測が多い共変量(項目)には注意が必要である

多変量解析では，共変量のなかで1個でも欠測があれば，その被験者を削除しなければならない．そのため，欠測が多い共変量をモデルに含めると，多くの被験者を削除することになる．また，観測方法が煩雑な場合には，欠測が多くなる傾向にある．そのため，このような共変量は，予め変数選択の候補から覗いておくことが望ましい．

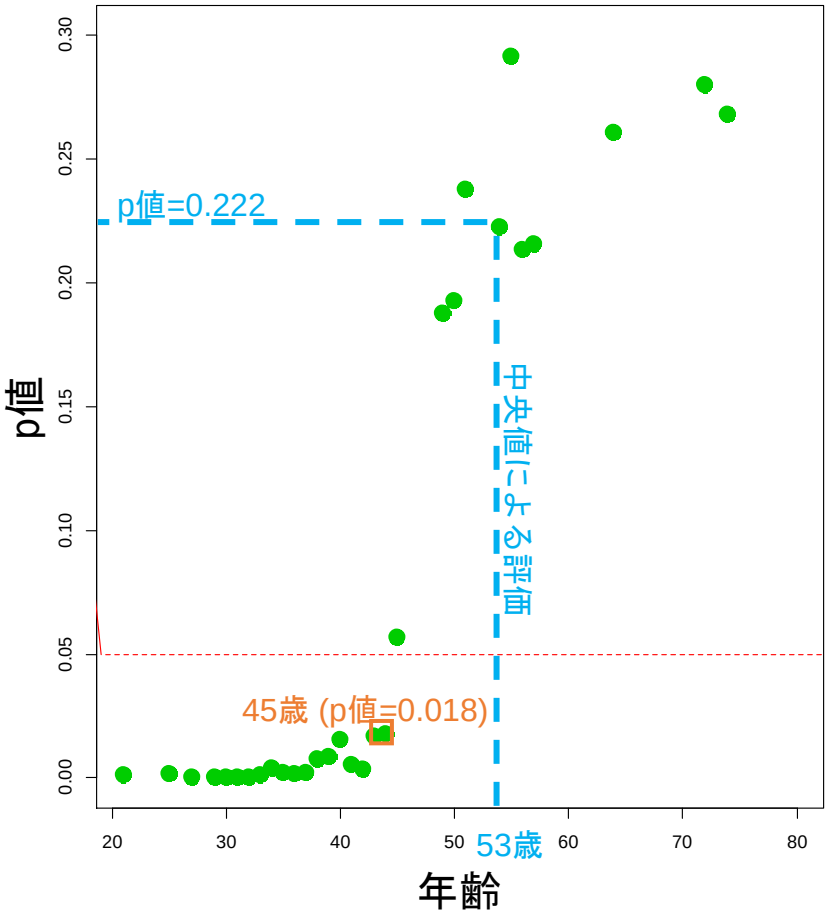
(5) 可能であれば総当たり法を用いる

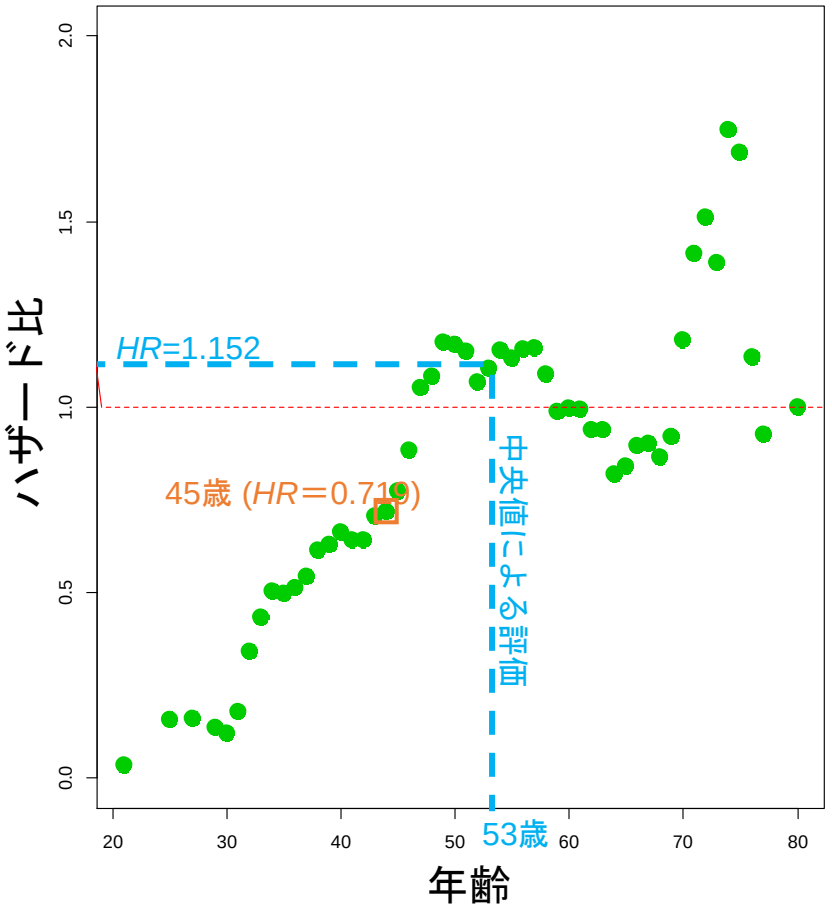
変数増加法や変数減少法が必ずしも最適なモデルに到達するとは限らない．最適なモデルを選択できる唯一の方法は，すべての候補モデルを評価する総当たり法のみである．共変量の数が10個の場合，候補となるモデルの数は1,023個である．そのため，臨床的知見あるいは，事前スクリーニングなどを用いて変数選択に用いる共変量を可能な限り少なくし，そのもとで，総当たり法によって変数選択を実施することが考えられる．

QA.8：多変量解析の2値化の問題について

多変量解析などでは，量的変数を2値化することがよくあります．

以下のグラフは，ドイツ乳がん研究グループによって実施された無作為化比較試験のデータ(Schumacher et al., 1994)において，年齢のカットオフ値を変化させながら，ハザード比およびログランク検定を用いてp値を計算したものである．





変量の2値化において用いられることが多い中央値(53歳以上，53歳未満)をカットオフ値にした場合，p値は0.222である(有意でなかった)．一方で，45歳(45歳以上，45歳未満)をカットオフ値にした場合，p値は0.018である(有意だった)．

Schumacher, M., et al. : Randomized 2×2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. Journal of Clinical Oncology, 12, 2086–2093, 1994.

2値化のとり方によっては，解釈が逆転することがあるので注意が必要です

45歳(＜45/≧45)の場合

高年齢群(45歳以上)のほうが低年齢群(45歳未満)よりも死亡リスクが低い **(HR=0.719)**

解釈が逆

53歳(＜53/≧53)の場合

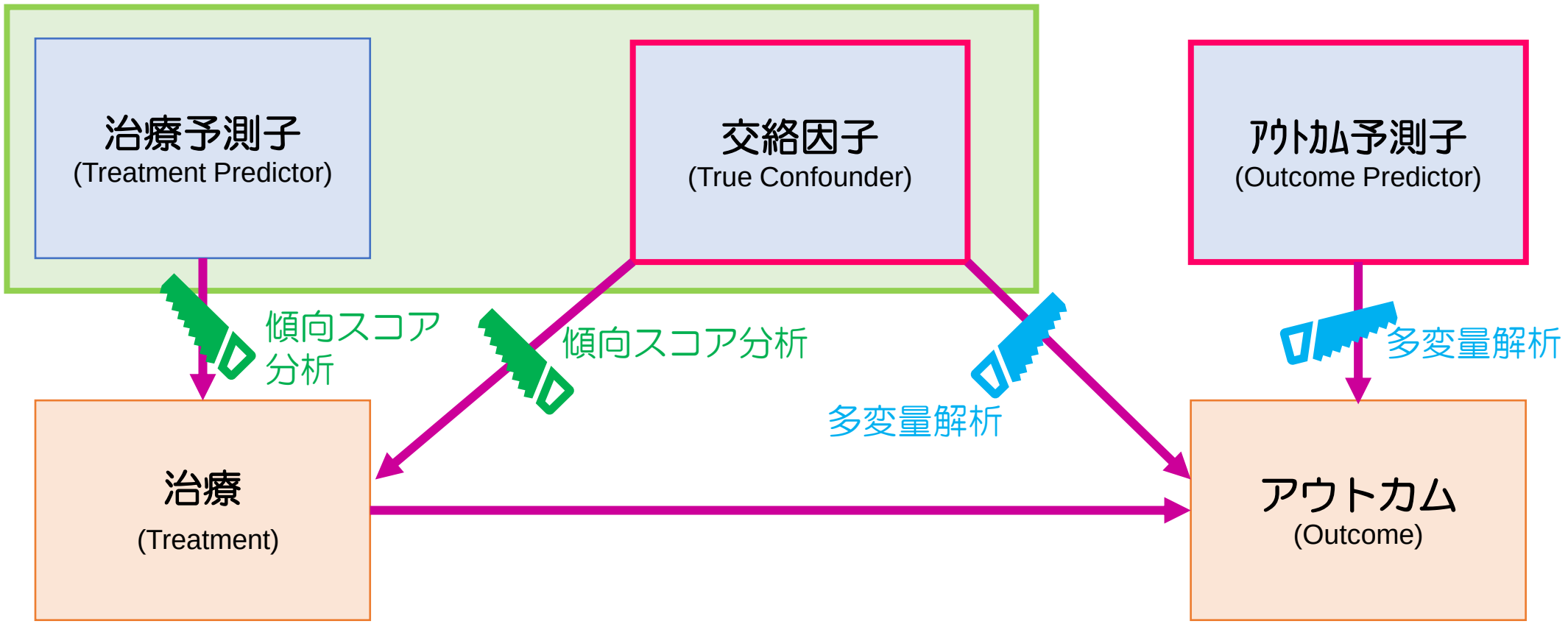
高年齢群(53歳以上)のほうが低年齢群(53歳未満)よりも死亡リスクが低い **(HR=1.152)**

カットオフ値の選定に関するRule of Thumb

(1)中央値を用いる
(2) 臨床的にリーズナブルなカットオフ値を選定する

臨床的にリーズナブルとは，今回の事例の場合には，例えば疫学的調査などから得られた乳癌患者の平均年齢を用いたり，あるいは，閉経年齢を用いることを意味する．いずれにしても，カットオフ値選定におけるゴールド・スタンダードは存在しないため，試行的に決定する必要がある．

QA.9：観察研究における傾向スコアって何ですか？

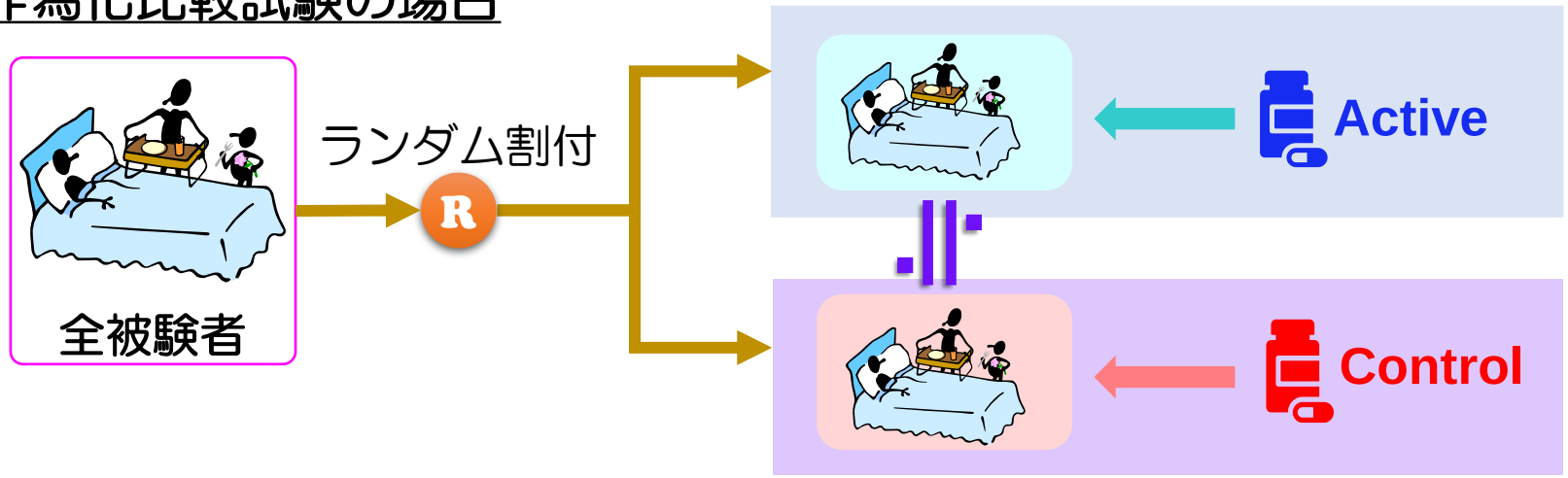


傾向スコア分析：観察研究における治療選択に対する影響を排除するための統計的方法

多変量解析：治療以外のアウトカムへの影響を排除(調整)するための統計的方法

なぜ治療選択への影響を排除しなければならないか？

無作為化比較試験の場合



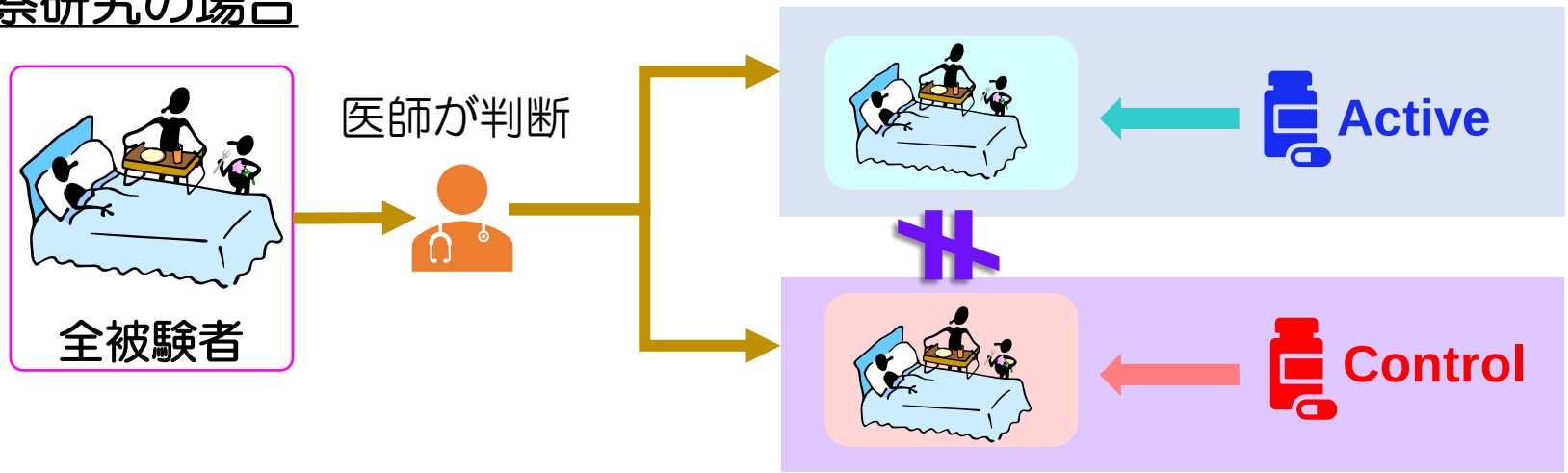
Active群

Control群

ランダム化されているので各群の被験者の分布は同じと仮定される.

→ したがって，「治療法の違い」のみがアウトカムに反映される評価される.

観察研究の場合



Active群

Control群

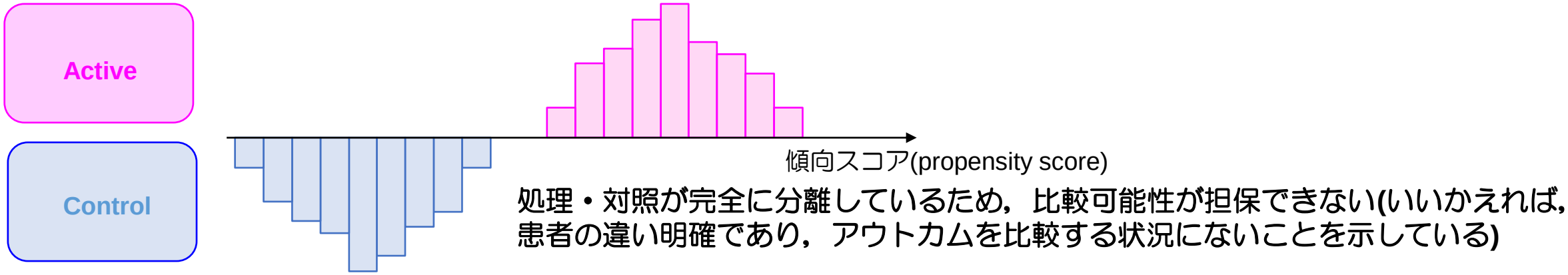
医師が治療法を判断するので，被験者の分布が異なる可能性が高い

→ したがって，「治療法の違い＋患者の状態」がアウトカムに反映される.

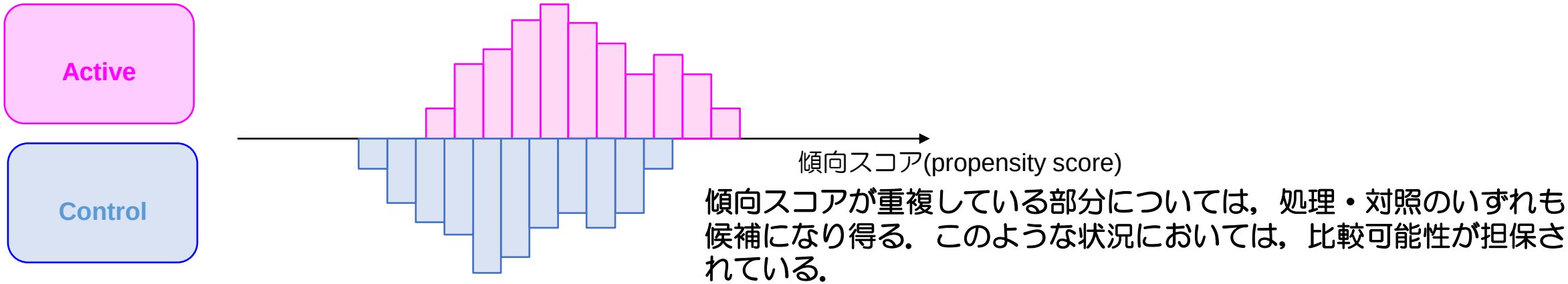
傾向スコア分析の「傾向スコア(propensity score)」とは？

ロジスティック回帰分析
- 治療法(1：Active, 0：Control)を応答変数，
- 治療法に影響を及ぼす要因(治療予測子，交絡因子)を説明変数

このとき，得られたロジスティック回帰モデルを用いて得られるActive群に対する帰属確率e_iは**傾向スコア(Propensity score)**と呼ばれる．

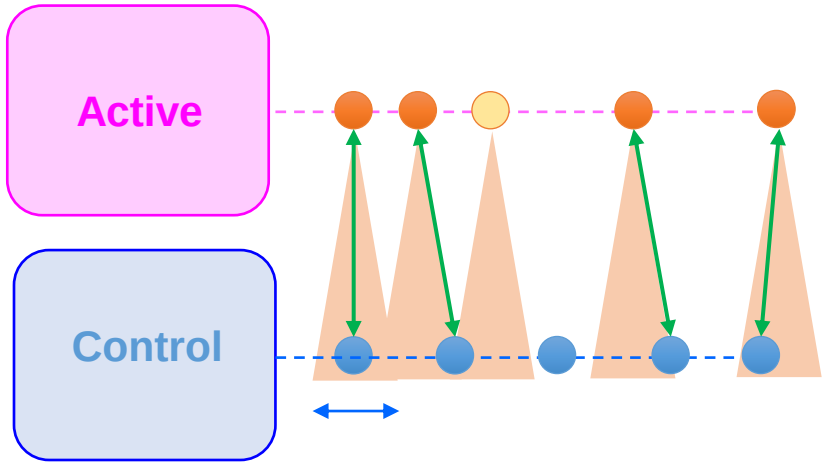


処理・対照が完全に分離しているため，比較可能性が担保できない(いいかえれば，患者の違い明確であり，アウトカムを比較する状況にないことを示している)



傾向スコアが重複している部分については，処理・対照のいずれも候補になり得る．このような状況においては，比較可能性が担保されている．

傾向スコア計算後の分析方法のとしてもちいられるもの：マッチング



キャリパーを伴う最近傍マッチング[Nearest neighbor within caliper]
マッチングさせる許容領域(キャリパー)を設定した最近傍マッチング(キャリパーを大きくするほどマッチングされる個体は増加する).
— Rosenbaum & Rubin(1985)は0.25×SDをキャリパーに設定することを推奨している(0.20を推奨する報告もあるため，0.25～020を推奨している).

Short communication

Radiosurgery alone is associated with favorable outcomes for brain metastases from small-cell lung cancer

Tyler P. Robin[a], Bernard L. Jones[a], Arya Amini[b], Matthew Koshy[c,d], Laurie E. Gaspar[a], Arthur K. Liu[a], Sameer K. Nath[a], Brian D. Kavanagh[a], D. Ross Camidge[e], Chad G. Rusthoven[a,*]

小細胞肺がん患者の脳転移に対する全脳照射療法(WBRT)と定位放射線手術(SRS)の比較

Variable	WBRT	SRS	p-value
Total patients	N=5,752	N=200	
Age			
<65	3,013 (52.4%)	102 (51.0%)	0.701
≥65	2,739 (47.6%)	98 (49.0%)	
Sex			
Male	2,926 (50.9%)	96 (48.0%)	0.425
Female	2,826 (49.1%)	104 (52.0%)	
Race/Ethnicity			
White (non-Hispanic)	5,030 (87.4%)	172 (86.0%)	
Black	476 (8.3%)	14 (7.0%)	0.008
Hispanic	116 (2.0%)	11 (5.5%)	
Other/unknown	130 (2.3%)	3 (1.5%)	
CDCC			
0	3,423 (59.5%)	125 (62.5%)	0.437
1	1,617 (28.1%)	48 (24.0%)	
2+	712 (12.4%)	27 (13.5%)	
Extracranial metastases			
No	3,069 (53.4%)	138 (69.0%)	<.001
Yes	2,683 (46.6%)	62 (31.0%)	

Supplemental table 1. Patient characteristics.
CDCC=Charlson/Deyo combined comorbidity score; WBRT=whole brain radiation therapy; SRS=stereotactic radiosurgery.

Variable	WBRT	SRS	p-value
Total patients	N=1,930	N=193	
Age			
<65	960 (49.7%)	99 (51.3%)	0.680
≥65	970 (50.3%)	94 (48.7%)	
Sex			
Male	934 (48.4%)	92 (47.7%)	0.848
Female	996 (51.6%)	101 (52.3%)	
Race/Ethnicity			
Caucasian (non-Hispanic)	1,718 (89.0%)	172 (89.1%)	
Black	141 (7.3%)	14 (7.3%)	0.999
Hispanic	50 (2.6%)	5 (2.6%)	
Other/unknown	21 (1.1%)	2 (1.0%)	
CDCC			
0	1,232 (63.8%)	121 (62.7%)	0.952
1	437 (22.7%)	45 (23.3%)	
2	261 (13.5%)	27 (14.0%)	
Extracranial metastases			
No	1,344 (69.6%)	135 (69.9%)	0.929
Yes	586 (30.4%)	58 (30.1%)	

Supplemental table 2. Propensity score matched patient groups.
CDCC=Charlson/Deyo combined comorbidity score; WBRT=whole brain radiation therapy; SRS=stereotactic radiosurgery.

マッチング後の解析

マッチング後は，マッチングされたデータを用いて通常の解析と同じように行えます．

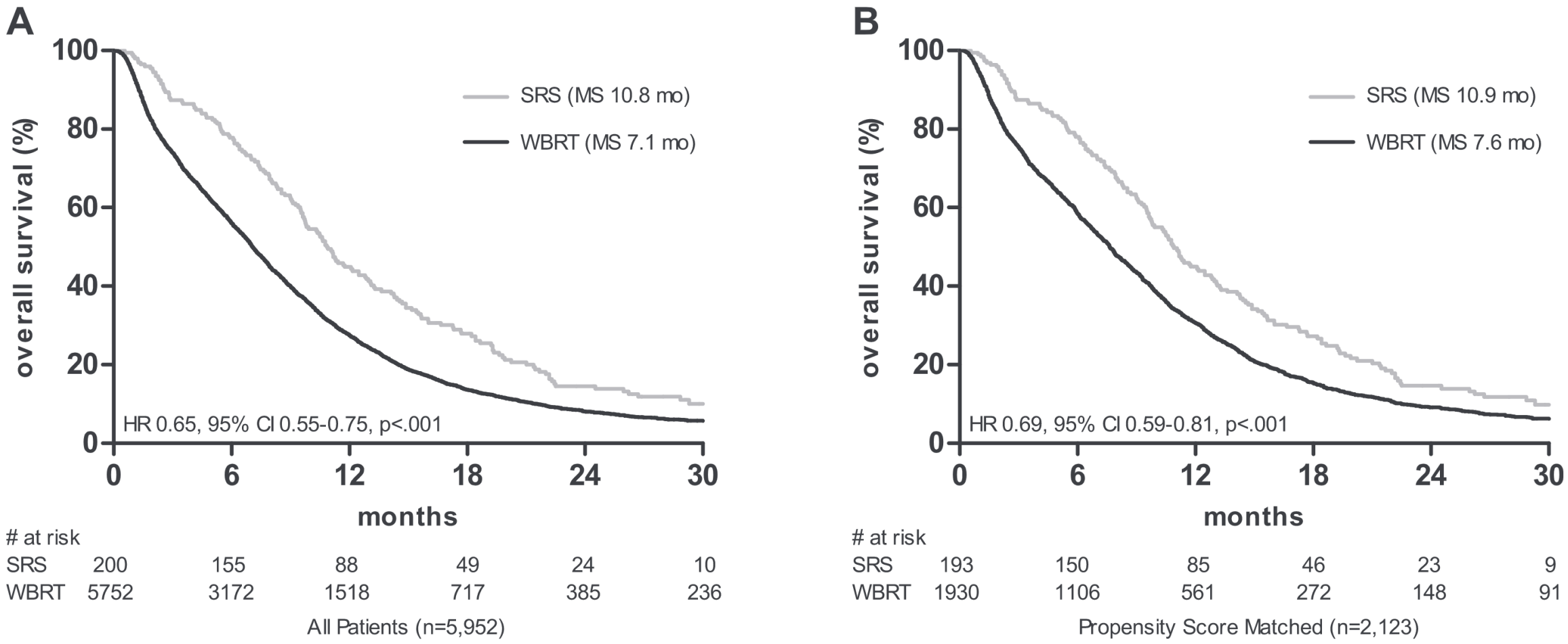


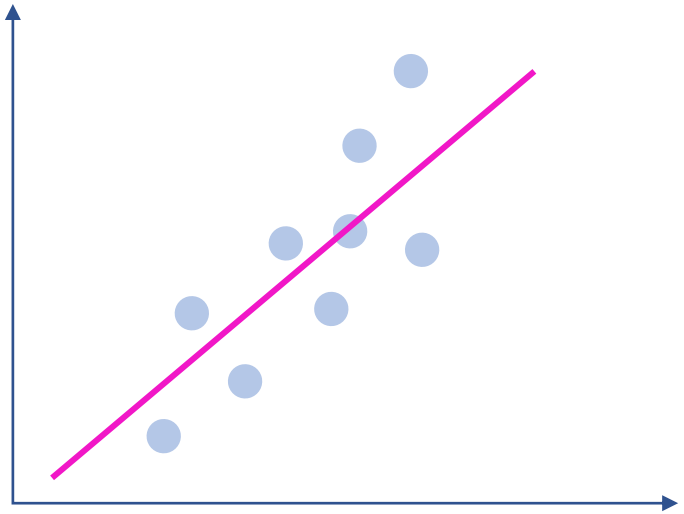
Fig. 1. Kaplan-Meier overall survival curves comparing patients receiving stereotactic radiosurgery (SRS) versus whole brain radiation therapy (WBRT) for brain metastases from small cell lung cancer. (A) All patients. (B) Propensity score-matched patients. MS, median survival.

QA.10：名義変数と連続変数の相関を表す統計の数値はありますでしょうか？

原則として尺度が異なる変数間の相関という考え方はありません.

そもそもの考え方として，相関(あるいは連関)を以下に示します.

相関(correlation)



一方の変数の値が上昇すれば，もう一方の変数の値が上昇(or 減少)するか？

連関(association)

	変数B = 1	変数B = 0
変数A = 1		
変数A = 0		

■のカテゴリの割合が■のカテゴリの割合よりも多いか？

名義変数と連続変数では，いずれの場合にも該当しません．通常の解析では，名義変数のカテゴリ毎に比較する，あるいは相関を見るなどの使い方をしたうえで，群間の違いあるいは類似度をみるものと思います.

QA.11：JMPで多変量解析を行った際，ORや信頼区間が極端に高い値になってしまうことがあります

幾つかの理由がありますが，どうしようもないところはあります．

■ 極端に高い値をとる2大理由

(1) 説明変数のカテゴリに大きな偏りがありませんか？

	応答 1	応答 0
カテゴリ 1	100	100
カテゴリ 0	100	0

各説明変数と応答変数のクロス集計表において，どこかのセルが0(あるいは0に近い数値)をとっている場合には，おかしな結果を導きます．このような場合には，当該説明変数は削除ください．

(2) 説明変数を症例数に関係なく入れていませんか？

症例数に対して極端に症例数を含めた場合には，回帰モデルが適切に推定できないため，異常な数値を計算してしまう可能性があります．このような場合には，単変量解析を用いた事前スクリーニングが必要です．あとは，変数選択も利用してください．

QA.12：多重共線性をVIF以外で検討する方法について教えてください

VIF (Variance Inflection Factor)とは

VIFは，説明変数に用いられた「任意の変数A」と「変数A以外の変数」との重相関係数r

重相関係数r ＝ A ⟷（重相関） B C ・・・ を用いて $VIF = \frac{1}{1 - r^2}$ で表される

VIFが10以上であれば多重共線性が疑われるといわれる．重相関係数だが，r^2は変数Aを応答，その他の変数を説明変数に用いたときの重回帰分析の寄与率(決定係数)と同じである

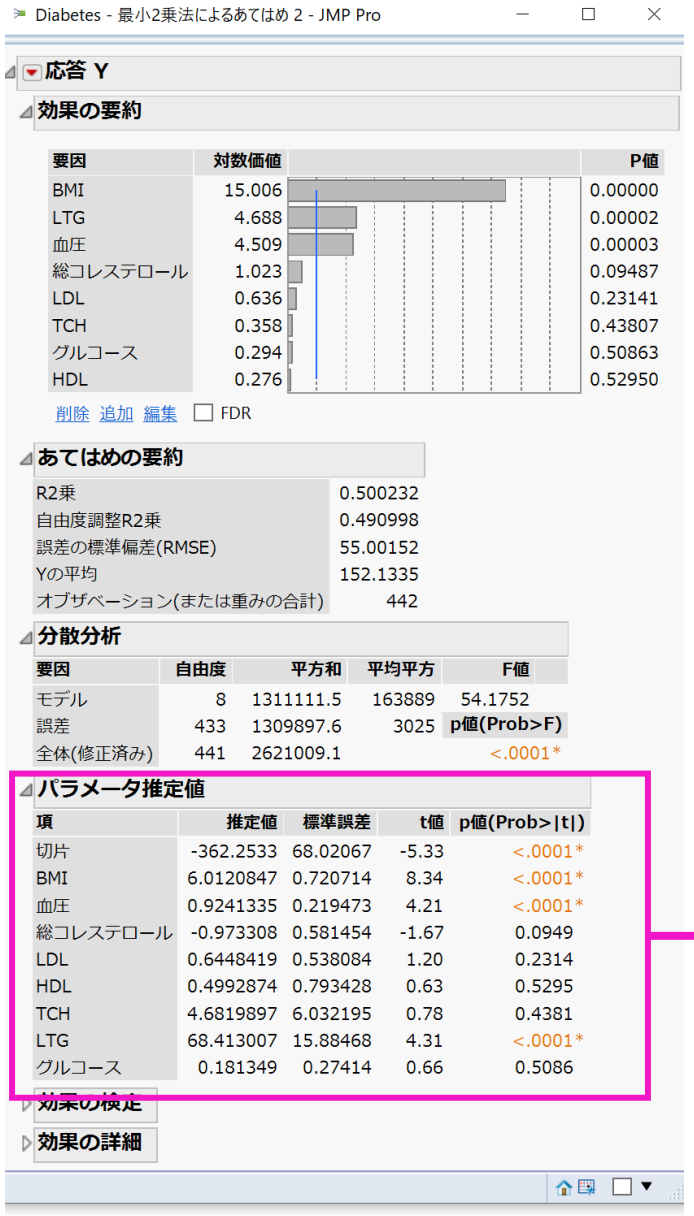
大事なのは，「VIFは説明変数間の重相関係数に基づくもの」であり，応答Yに関連する回帰分析の手法というわけではかならずしもないことである．

VIF以外での多重共線性の見方

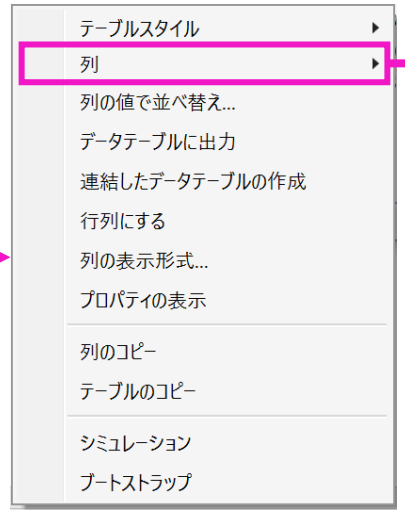
そのため，重相関係数を用いることでも可能(説明変数Aを応答にして，その他を説明変数にした重回帰の決定係数)を見ればよい(VIFの10は重相関係数の0.95とほぼ同じ意味)

ご質問ではJMPではVIFは計算できないとなっていましたが，計算できます．

(準備) 回帰分析を実行：「分析」→「モデルのあてはめ」



右クリック



選択

選択

パラメータ推定値

項	推定値	標準誤差	t値	p値(Prob>|t|)	VIF
切片	-362.2533	68.02067	-5.33	<.0001*	.
BMI	6.0120847	0.720714	8.34	<.0001*	1.4780541
血圧	0.9241335	0.219473	4.21	<.0001*	1.3433106
総コレステロール	-0.973308	0.581454	-1.67	0.0949	59.030174
LDL	0.6448419	0.538084	1.20	0.2314	39.039962
HDL	0.4992874	0.793428	0.63	0.5295	15.352656
TCH	4.6819897	6.032195	0.78	0.4381	8.8332824
LTG	68.413007	15.88468	4.31	<.0001*	10.037762
グルコース	0.181349	0.27414	0.66	0.5086	1.4479505

VIFが追加される

QA.13：名義変数と連続変数を同時に検討する場合には，多重共線性というのは検討できますか？

■ 回帰分析における名義変数の取り扱い

いま，3カテゴリの名義変数X(カテゴリをa,b,cとする)があったとき，回帰分析では，次のように扱われる．

	Xb	Xc
X=a	0	0
X=b	1	0
X=c	0	1

つまり，名義変数は「カテゴリ数－１」個の0 or 1の変数で回帰分析の説明変数に加えられる．これらの変数は，ダミー変数と呼ばれる．

したがって，VIFを計算すると，ダミー変数毎に計算される．もし，そのダミー変数の1個のVIFが高い(10以上である)ことがわかれば，そのもととなる変数を削除すればよい．

Xb	VIF = 12.0
Xc	VIF = 1.2

➡ 変数Xには共線性が疑われるので検討を行う

Thank you for your kind attention

shimokaw@wakayama-med.ac.jp

toshibow2000@gmail.com

(余談) 検定ではない評価の方法：効果量(effect size)

仮設検定におけるp値が0.05未満であるか否かということで研究のpositive/negativeの方向性が決まることについて，ASA (American Statistical Institute, 2016)が声明を発表している(日本計量生物学会が日本語版を作成).

そのなかでは，6つの主要な声明が出されている：

(1) P値はデータと特定の統計モデル（訳注: 仮説も統計モデルの要素のひとつ）が矛盾する程度をしめす指標のひとつである.
(2) P値は、調べている仮説が正しい確率や、データが偶然のみでえられた確率を測るものではない。
(3) 科学的な結論や、ビジネス、政策における決定は、P値がある値（訳注: 有意水準）を超えたかどうかにのみ基づくべきではない。
(4) 適正な推測のためには、すべてを報告する透明性が必要である。
(5) P値や統計的有意性は、効果の大きさや結果の重要性を意味しない。
(6) P値は、それだけでは統計モデルや仮説に関するエビデンスの、よい指標とはならない。

また，American Psychological Association (APA)の論文作成マニュアルでは，効果量，信頼区間が必須である旨が提示されている（2015年より日本心理学会においても論文投稿の際には効果量の記載が義務付けられた）.

Wasserstein, R.L., Lazer, N.A., Editorial: The ASA's statement on p-value: Context, process, and purpose, The American Statistician, 70, 129-133, 2016.
American Psychological Association：Publication manual of the American Psychological Association(6th ed.) , 2009.

(余談)効果量を使うとわかること：Cohen's dを参考に

比較 A

群A：平均値 = 11.52, 標準偏差 = 3.95, n=10
群B：平均値 = 7.17,　標準偏差 = 4.78, n=10

2標本t検定：検定統計量 = 2.218, p値 = 0.0396
効果量(Cohen's d) = 0.99 (効果量大)

比較 B

群A：平均値 = 5.30, 標準偏差 = 4.35, n=100
群B：平均値 = 3.80, 標準偏差 = 5.50, n=100

2標本t検定：検定統計量 = 2.136, p値 = 0.0339
効果量(Cohen's d) = 0.30 (効果量小)

つまり，同じようなp値であったとしても，効果量でみれば，比較Aのほうが群間差が顕著であることがわかる(注：p値が小さいからといって群間差の大小を比較できない).

Cohen(1992)は，効果量の目安として，

- 0.20未満：効果が認められない
- 0.20～0.50：効果量小
- 0.50～0.80：効果量中
- 0.80以上：効果量大

としている.

効果量は，標本平均，標準偏差，検定統計量などから簡単に計算できる(フリーのExcelシートなどもWeb上で落ちている).