看護統計セミナー 2022

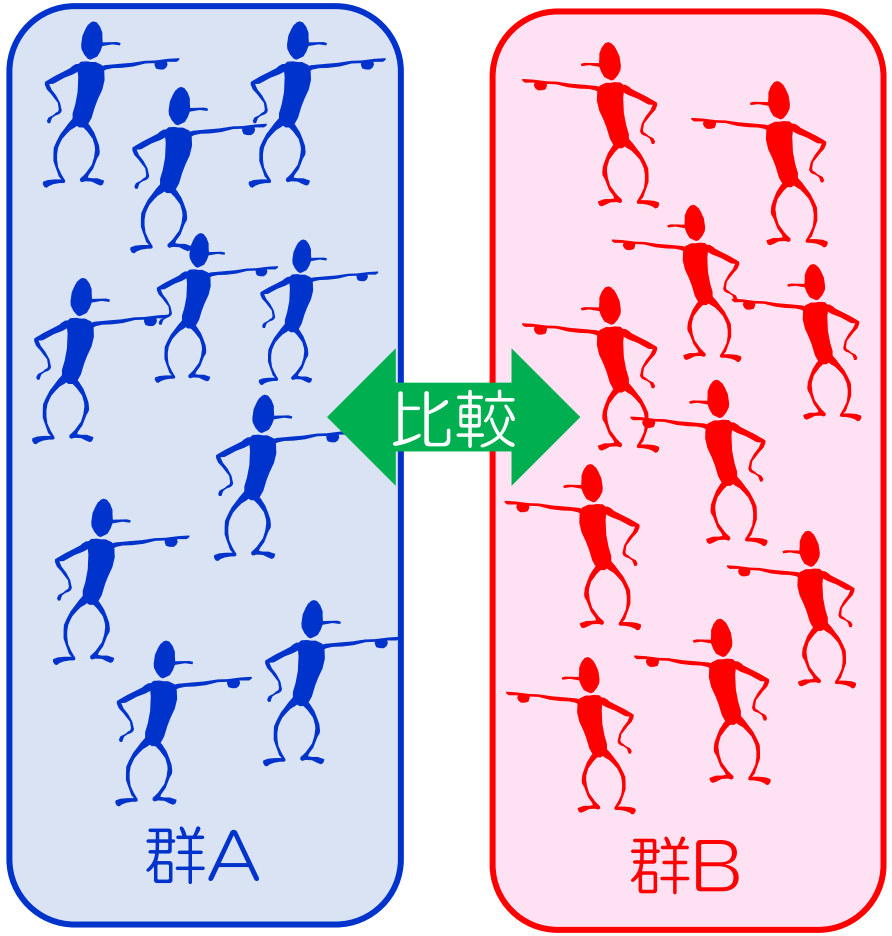
第1回目：量的データの解析

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

本日のおはなし：SPSSメニューを例に

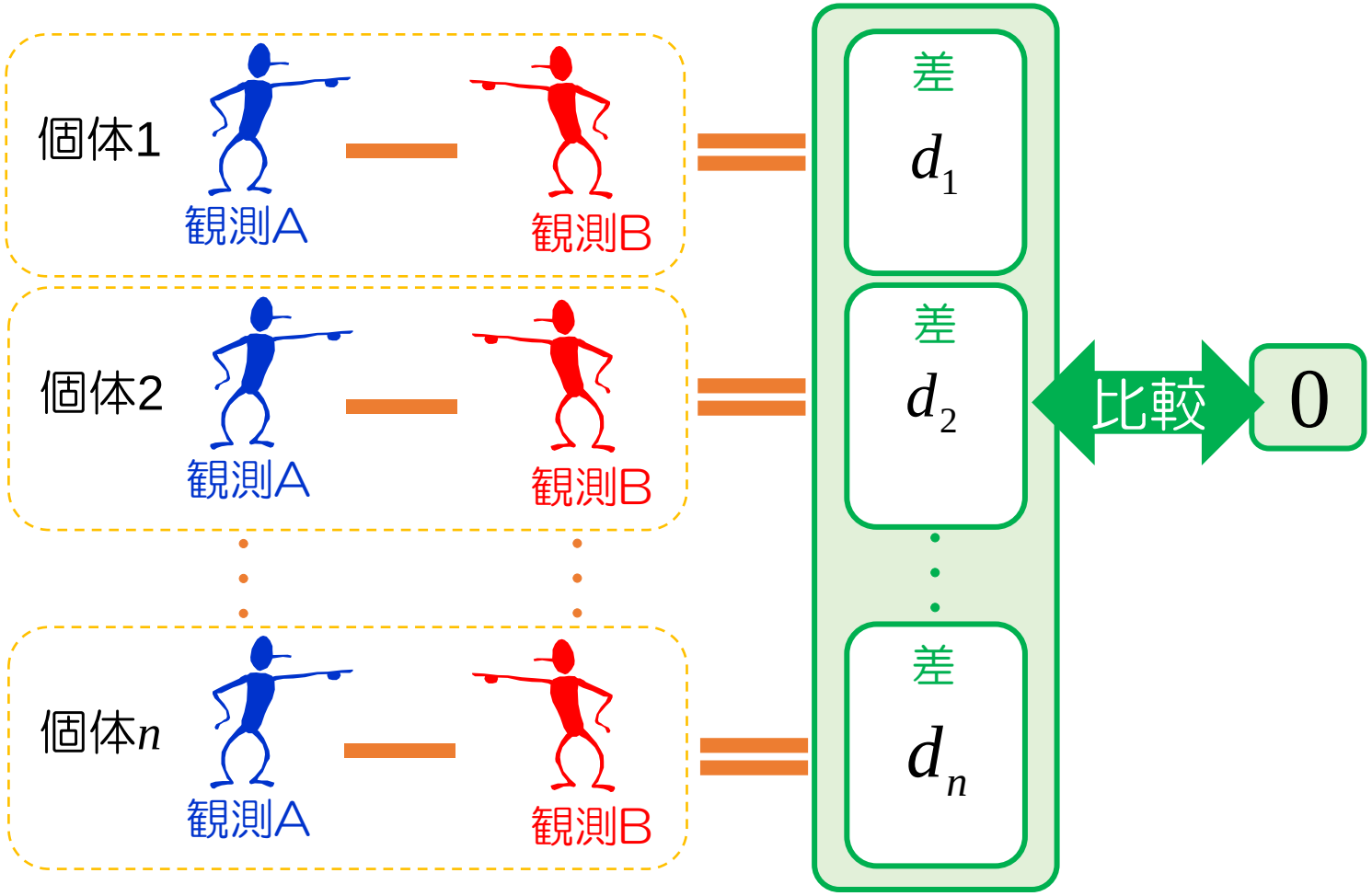
対応のあるデータと対応がない場合

独立2標本



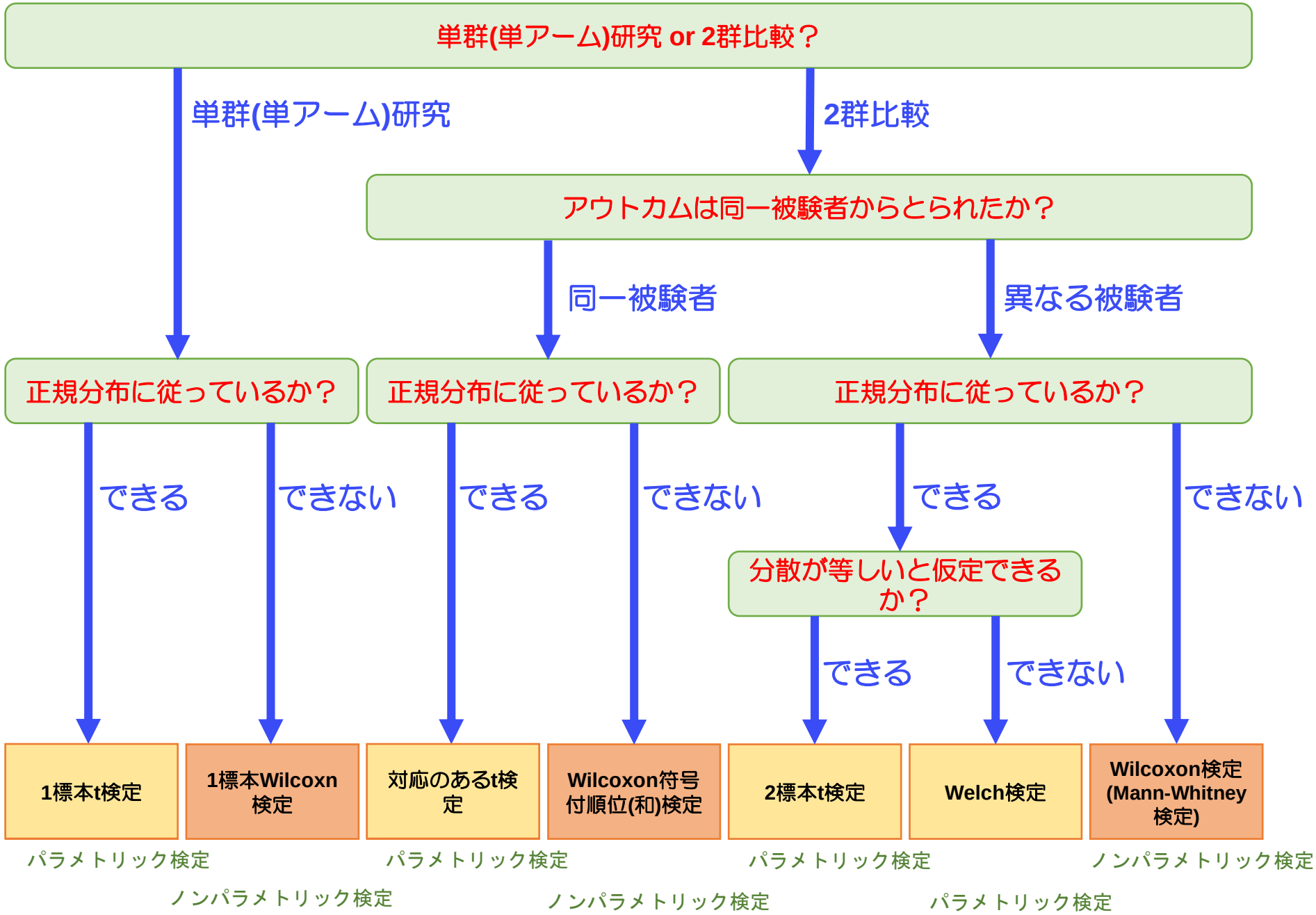
例えば，新薬と既存薬を2群に分けた被験者に投与する場合がこれにあたる．各個体が独立である．

対応のある標本



介入前と介入後のアウトカムの差を比較するような場合がこれにあたる．つまり，個体間で比較される．

教科書的な統計的検定の取捨選択



取捨選択を「検定」で選ぶのは正しいのか？

7．データ分析方法

データの欠損値については平均値で代入を行った．変数の正規性の検討にはShapiro-Wilk検定を使用したが，すべての変数において正規性が認められなかったためノンパラメトリック検定を用いた．

・被験者数は92名

中島・安東. 日本糖尿病教育・看護学会誌, 25(1), 83-92, 2021

注：欠損値を平均値で代入するのも誤り

6．解析方法

背景要因と身体症状は度数分布を求め，PSQI-Jの下位尺度と総得点の記述統計量を求め，Kolmogorov-Smirnovの正規性の検定を行った．

・被験者数は64名

浦他. 日がん看学誌, 35, 91-101, 2021

5．分析方法

質問紙調査票によって得られたデータについて，単純集計および項目別平均値および中央値の算出を行った．また，各変数が正規分布に従うかをShapiro-Wilk検定で確認した．Shapiro-Wilk検定により全項目（全変数）において正規性が確認できなかったため，手術看護経験年数と手術看護の実践経験の関係についてKruskal-Wallis検定を行い分析した．有意水準は５％とした．

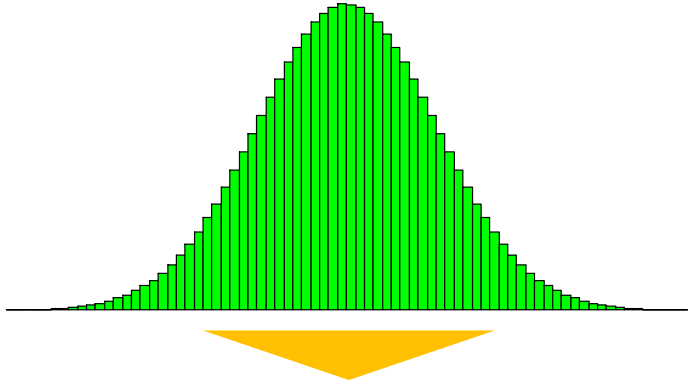
・被験者数は560名

福田他. 日看管学誌, 25(1), 108-117, 2021

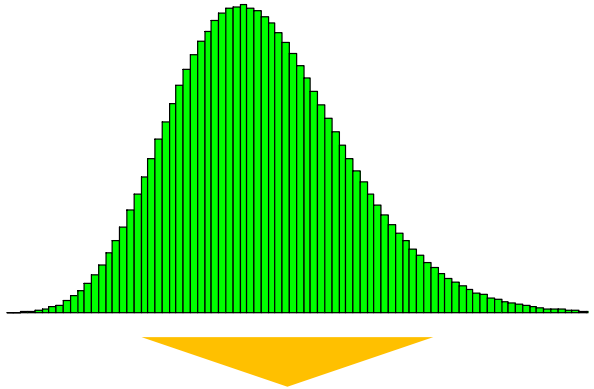
多くの看護研究論文で正規性検定に基づいて正規性を検討している

正規性検定で正規性を検討していいのか (100,000回のシミュレーション)

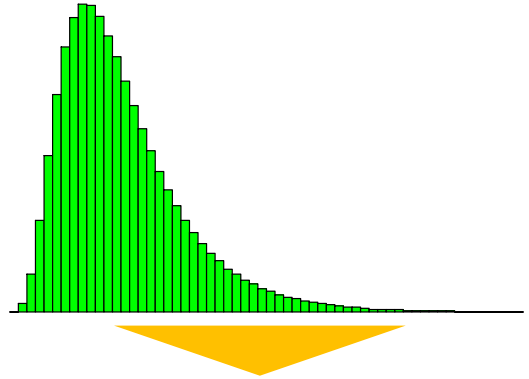
完全に正規分布に従っている場合

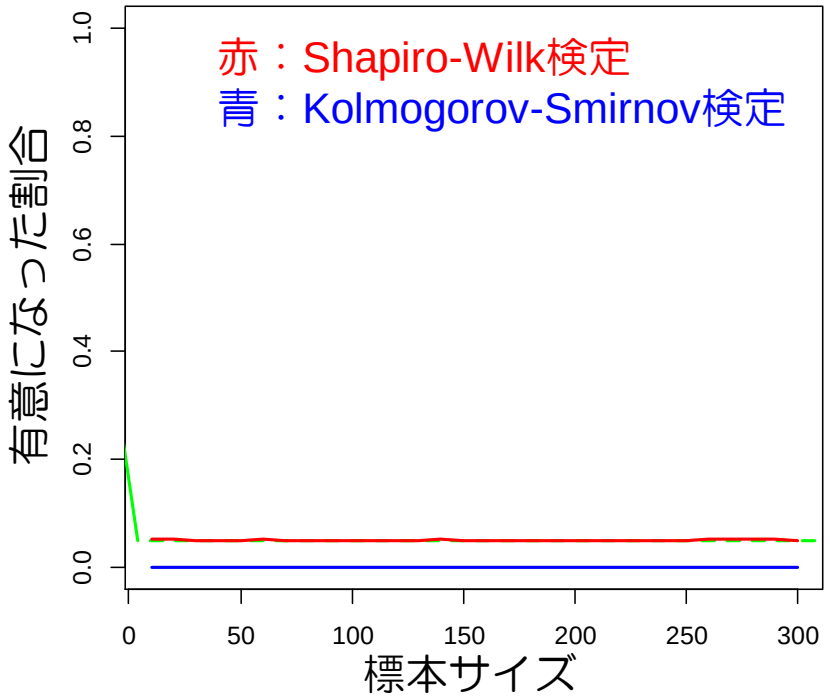


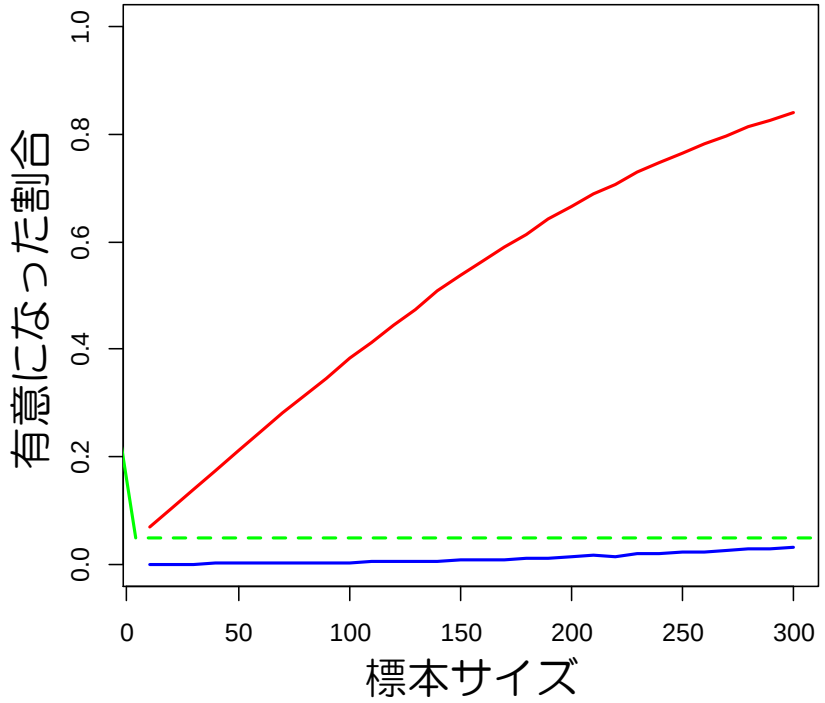
少し歪んでいる場合

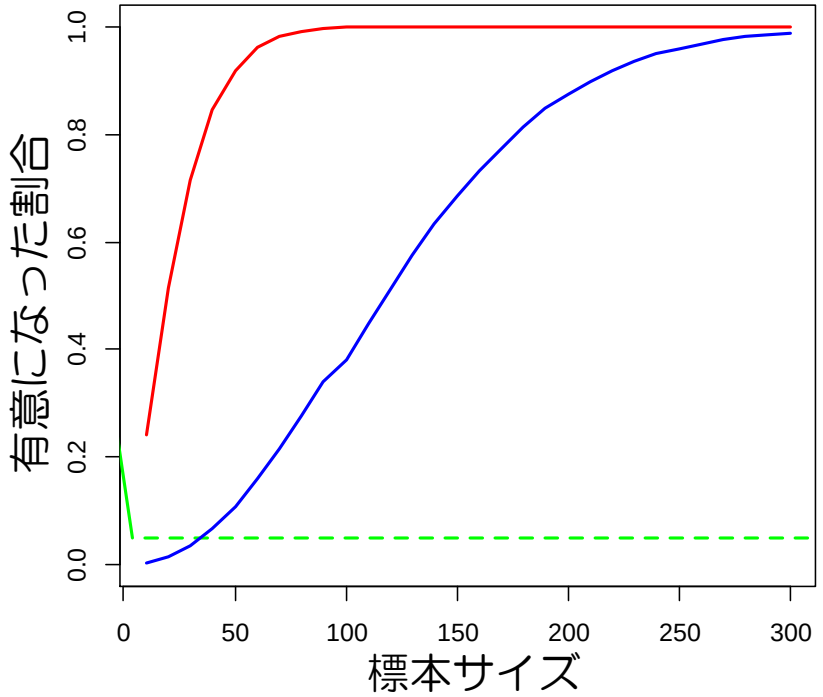


かなり歪んでいる場合







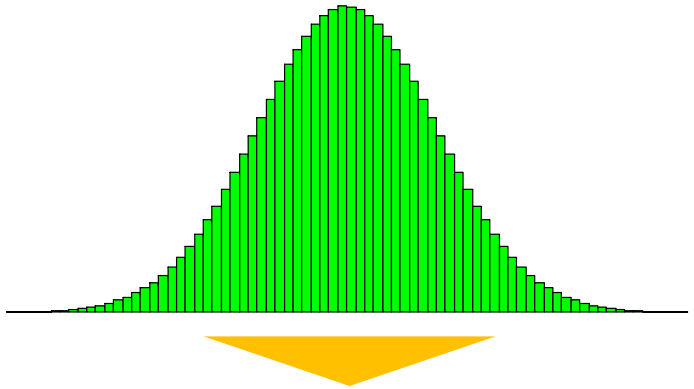


標本サイズ・検定手法に検定の取捨選択が大きく影響される

正規分布からのズレによってどれぐらい検定手法の結果に違いが生じるのか

100,000回のシミュレーション

完全に正規分布に従っている場合

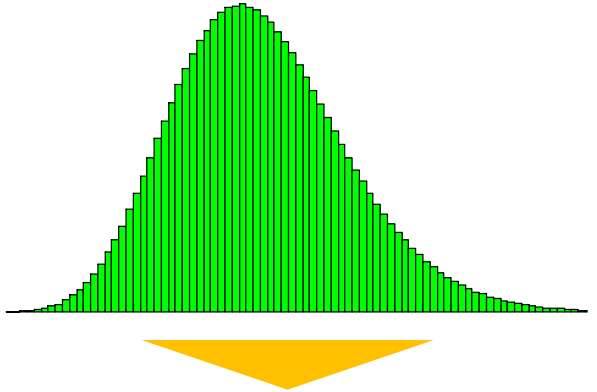


標本サイズ(各群) =50

- **2標本t検定で有意になった割合**
  0.697
- **Wilcoxn検定で有意になった割合**
  0.675

2標本t検定 ＞ Wilcoxon検定

少し歪んでいる場合

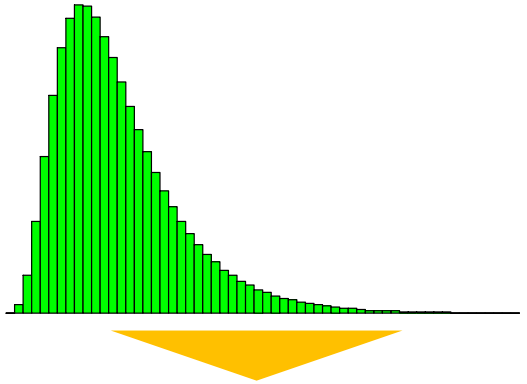


標本サイズ(各群) =50

- **2標本t検定で有意になった割合**
  0.749
- **Wilcoxn検定で有意になった割合**
  0.747

2標本t検定 ≒ Wilcoxon検定

かなり歪んでいる場合



標本サイズ(各群) =50

- **2標本t検定で有意になった割合**
  0.554
- **Wilcoxn検定で有意になった割合**
  0.714

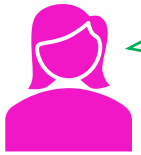
2標本t検定 ＜ Wilcoxon検定

2標本t検定のほうが有利・Shapiro-Wilkでは標本サイズが多いと有意になる可能性がある.

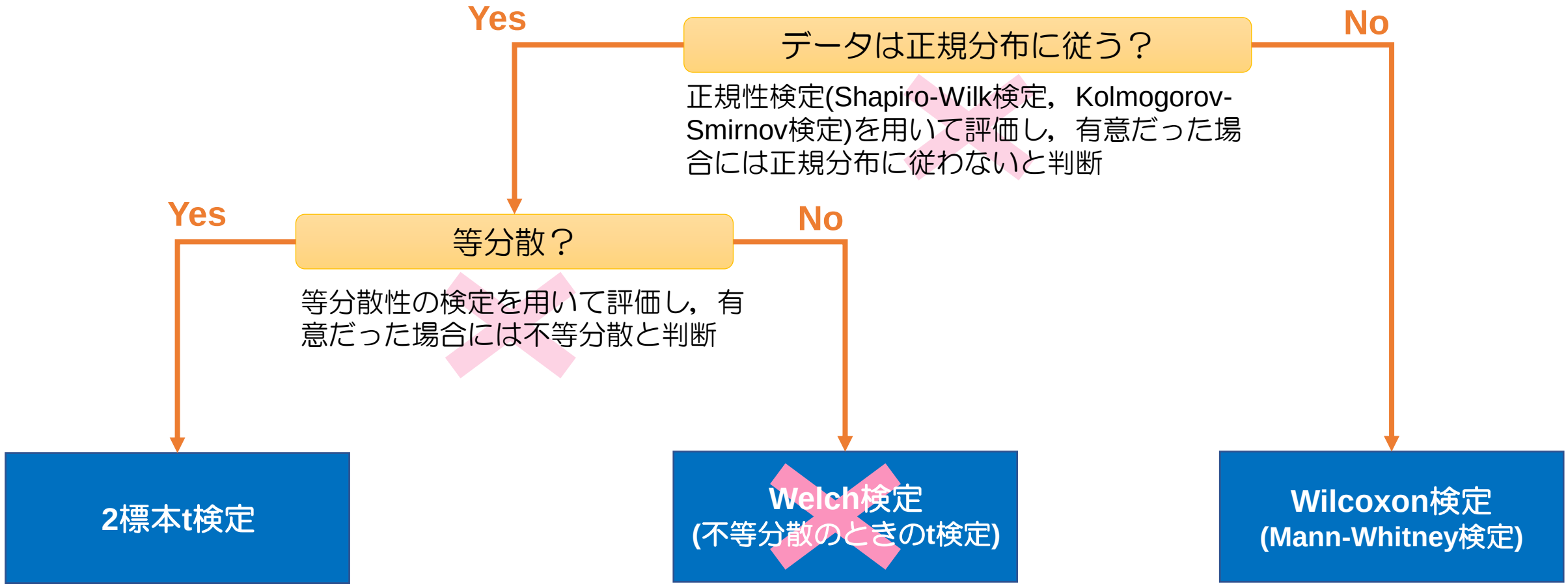
Wilcoxnn検定のほうが有利・Kolmogorov-Smirnov検定では標本サイズが多いと有意でない可能性がある.

データが歪むとWilcoxon検定のほうが良いが，正規性検定が判断材料となるかは疑問

教科書的な統計的検定に対するアンチテーゼ



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))

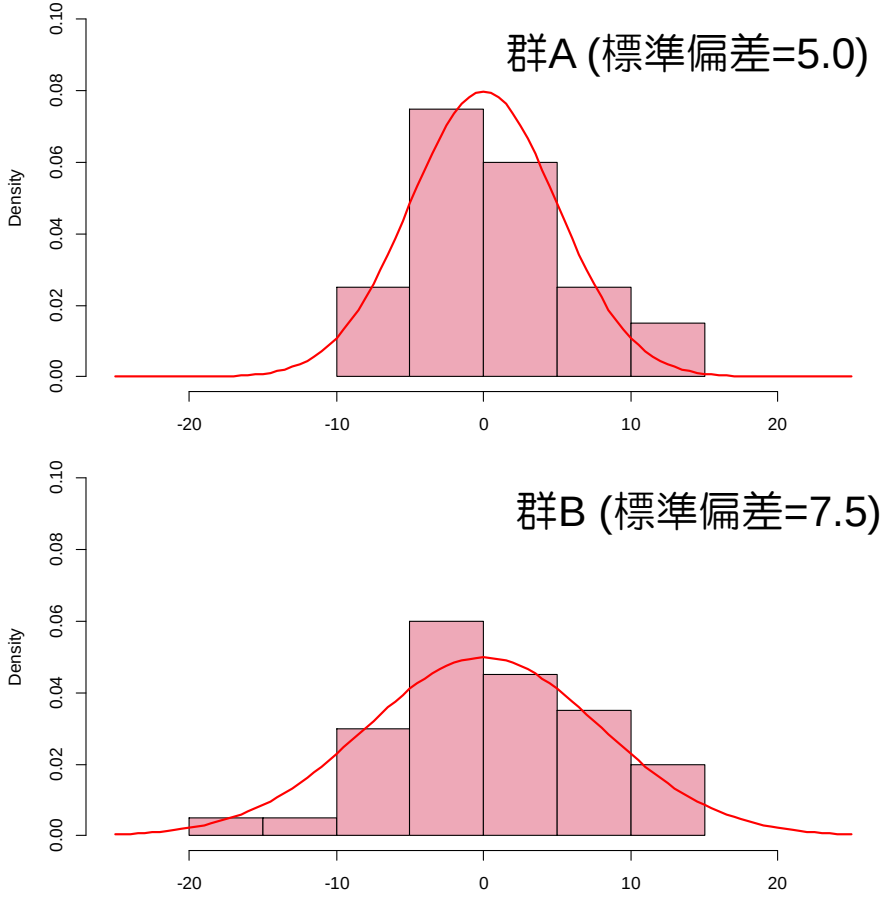


等分散性の検定を用いてt検定とWelch検定を選択することで起きえること

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．

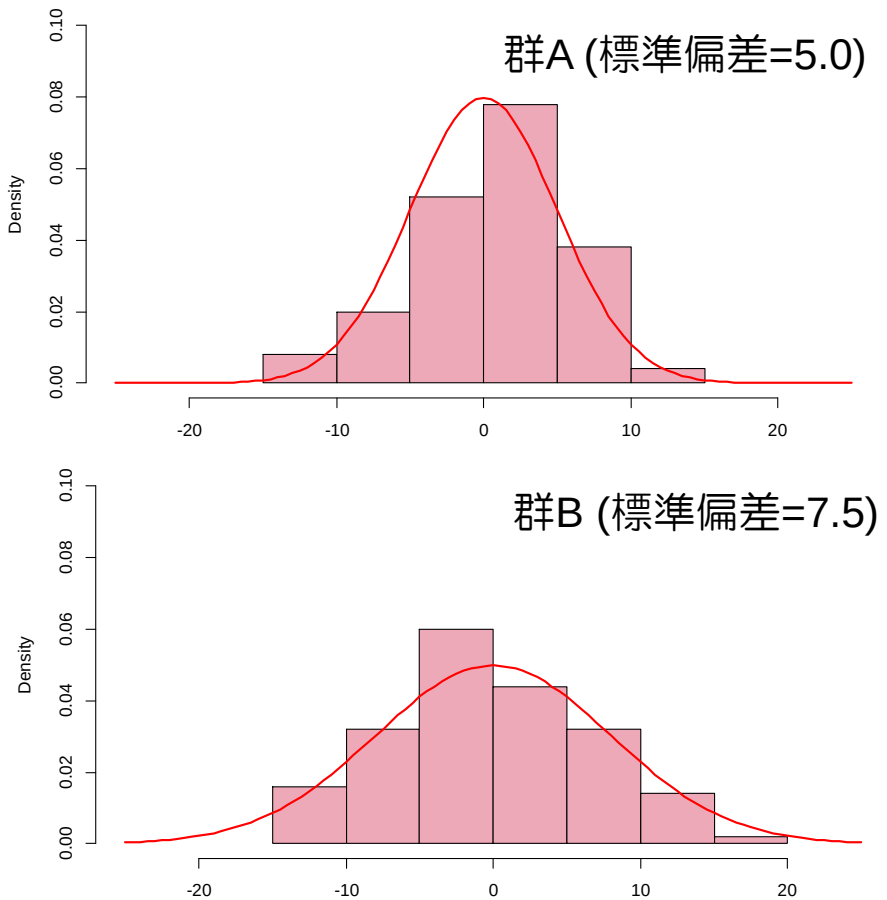
各群のサンプルサイズ = 40

等分散性の検定のp値=0.164



各群のサンプルサイズ = 60

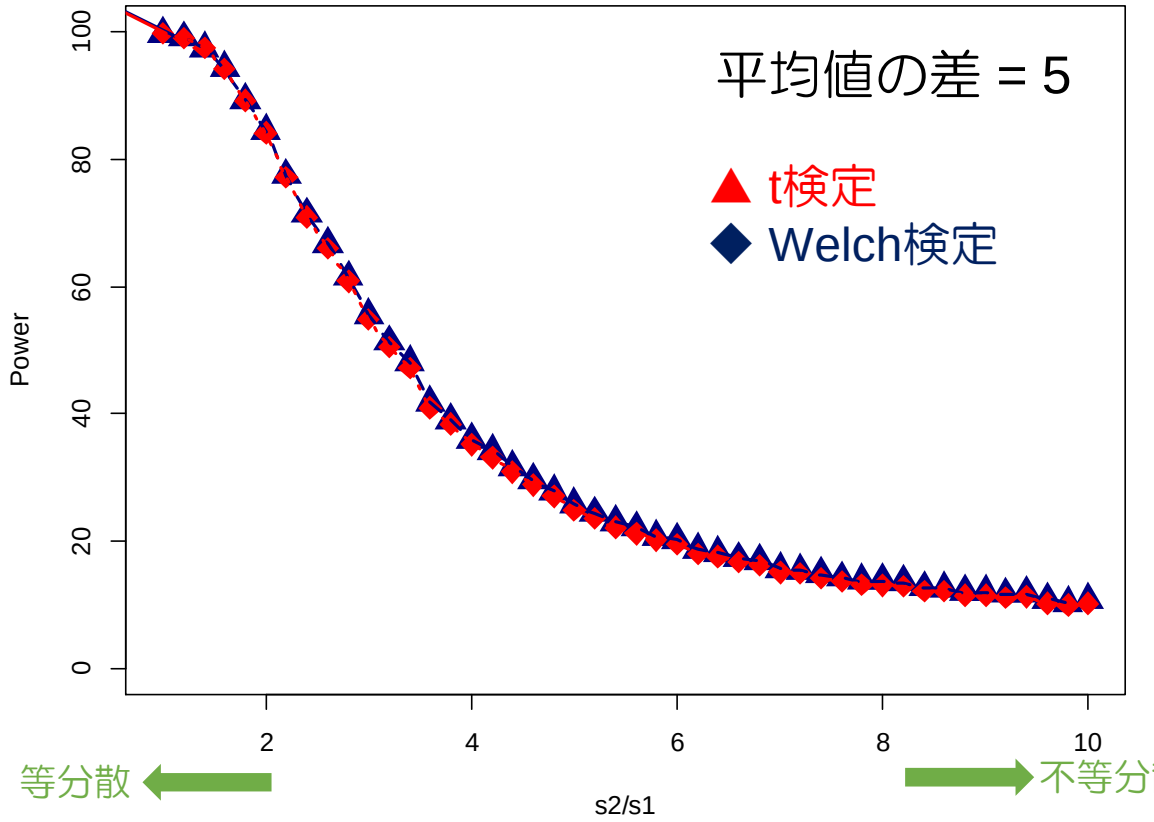
等分散性の検定のp値=0.006

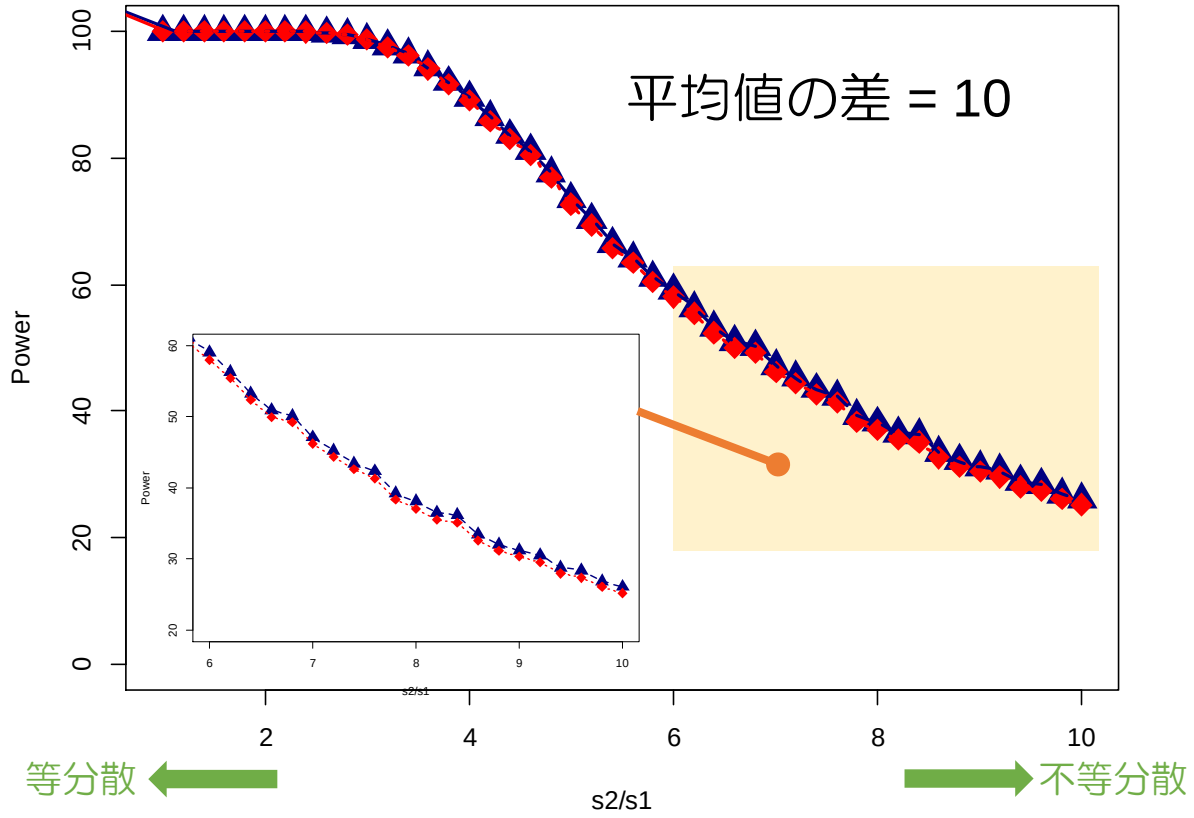


同じ母集団であっても，サンプルサイズの違いによって検定手法を変えるというのは，そもそもおかしいのでは？

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)．

Wilcoxon検定を用いるときの注意点

7. データ分析方法

データの欠損値については平均値で代入を行った．変数の正規性の検討にはShapiro-Wilk検定を使用したが，すべての変数において正規性が認められなかったためノンパラメトリック検定を用いた．

中島・安東. 日本糖尿病教育・看護学会誌, 25(1), 83-92, 2021

表２　セルフケア能力得点と下位尺度間の比較

	1型糖尿病 n=92	
	中央値（四分位範囲）	平均値±標準偏差
【セルフケア能力（合計得点）】	127.5（111.3-141.8）	124.8±21.9
知識獲得力	23.0（20.0-25.0）	22.0±3.2
自己管理の原動力	20.0（17.0-24.0）	19.6±5.0
モニタリング力	19.0（15.3-21.0）	18.3±3.7
応用・調整力	19.0（15.0-21.0）	17.6±4.3
ストレス対処力	17.0（13.0-20.0）	16.4±4.8
自分らしく自己管理する力	17.0（13.0-19.0）	16.0±4.9
サポート活用力	15.0（11.0-20.0）	14.9±6.4

注）*p<0.05　***p<0.001　注）Kruskal-Wallis 検定

正規分布に従っていないことを仮定しているため中央値(IQR)を用いるのは正しい

6. 解析方法

背景要因と身体症状は度数分布を求め，PSQI-Jの下位尺度と総得点の記述統計量を求め，Kolmogorov-Smirnovの正規性の検定を行った．

(中略)

統計解析は，睡眠データに正規性と直線性を認めずサンプル数も少ないため，Mann-WhitneyのU検定とKruskal-Wallis検定を用い，有意水準は5%とした．

浦他. 日がん看学誌, 35, 91-101, 2021

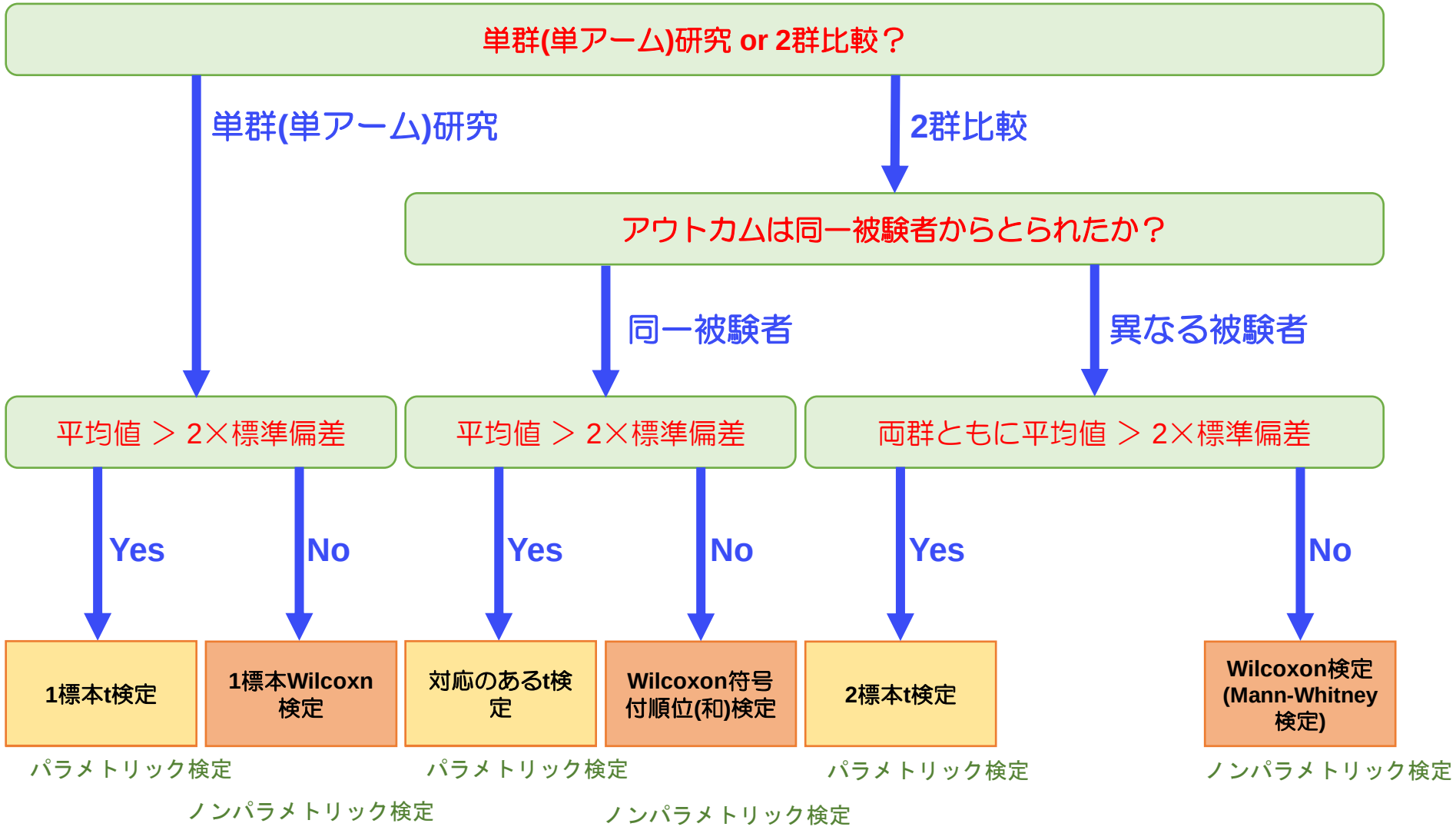
表４　身体的要因による PSQI-J の睡眠の比較

		n		PSQI 総得点	入眠の所要時間	実睡眠時間	睡眠の効率
悪心・嘔吐	なし	51	mean	4.9	20.4	6.7	92.0
			SD	2.8	16.9	1.2	10.7
	あり	13	mean	6.2	30.4	6.3	87.1
			SD	3.1	26.7	0.9	11.8
			p	0.158	0.254	0.180	0.134
倦怠感	なし	41	mean	4.9	21.7	6.7	91.5
			SD	2.8	21.5	1.2	11.7
	あり	23	mean	5.7	23.6	6.6	90.0
			SD	3.0	15.4	1.0	9.8
			p	0.341	0.215	0.571	0.270

(後略)

正規分布に従っていないことを仮定しているにも関わらず，平均値(SD)を用いるのは誤り

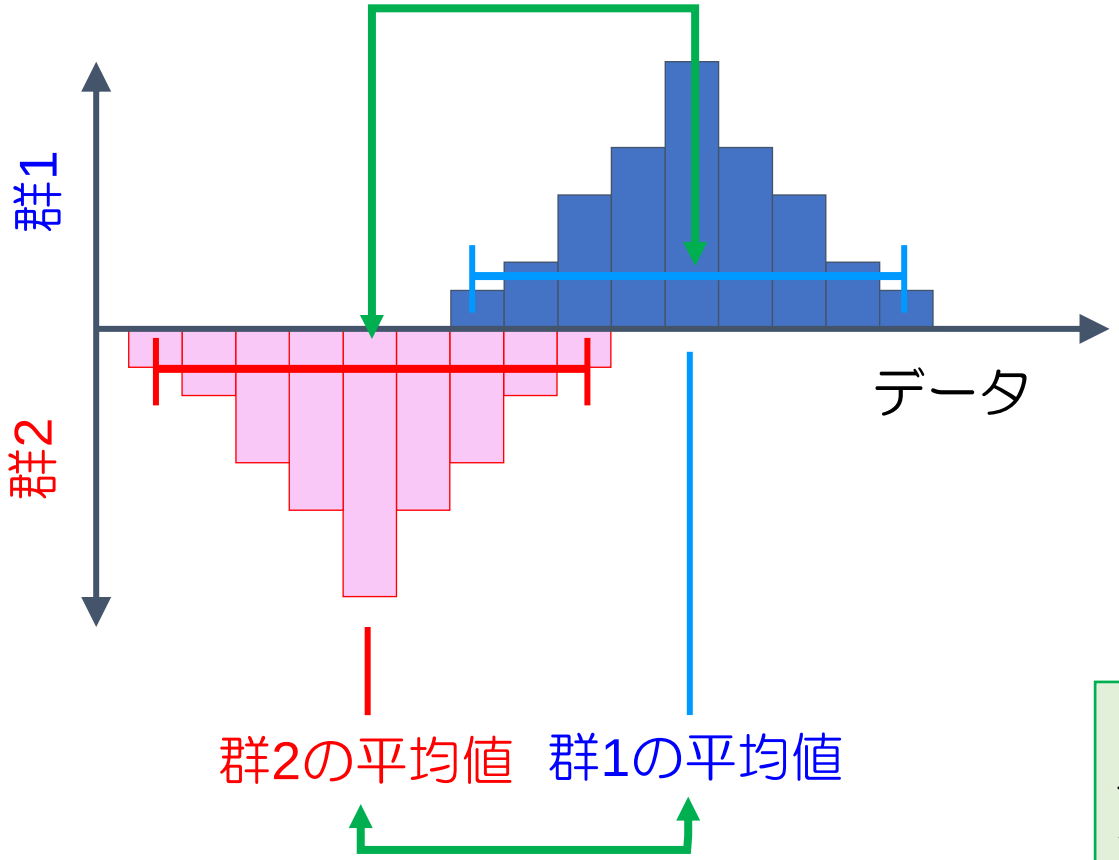
経験則に基づく統計的検定の取捨選択



正規性検定を絶対に用いてはいけないわけではない(というよりもあまり意味がないということだけ)．統計の専門家が書いた教科書で検定で統計手法を取捨選択すると書いている人はほとんどいない(多くは，非統計学者が書いたものばかり．また，**Welch検定**を推奨している教科書も同様)

パラメトリック検定とノンパラメトリック検定の意味の違い

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．

■ Wilcoxon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．

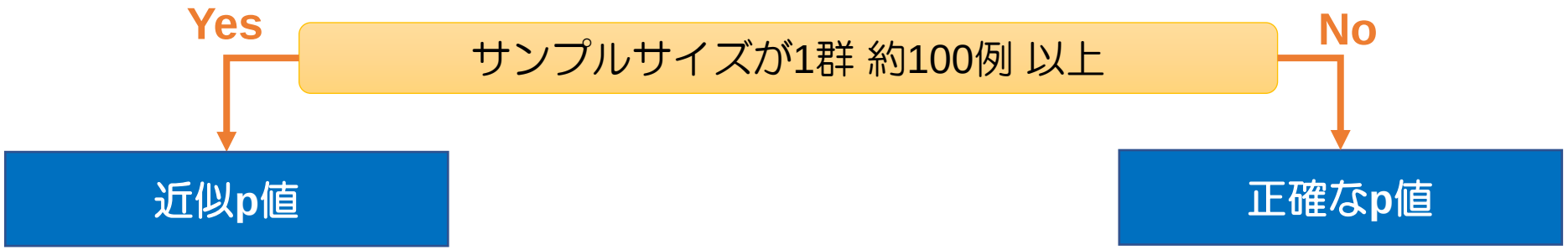
Wilcoxon (Mann-Whitney)検定のp値の計算方法には数種類存在する

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)．



素朴な疑問：背景因子をなぜ検定するのか？

表 1　対象者の属性

	合計（*n* = 202）	医師（*n* = 70）	看護師（*n* = 132）	*p* 値
性別，*n*（%）[a]				<.001
男性	65（32.2）	52（74.3）	13（9.9）	
女性	136（67.3）	18（25.7）	118（89.4）	
未記入	1（0.5）	0（0.0）	1（8.7）	
年齢（歳），*M*（*SD*）[b]	38.0（10.4）	40.8（9.1）	36.5（10.7）	.005
臨床経験年数（年），*M*（*SD*）[b]	14.7（10.0）	15.6（8.9）	14.2（10.5）	.331
小児領域臨床経験年数（年），*M*（*SD*）[b]	12.6（9.8）	12.3（9.8）	12.7（9.8）	.788
職位，*n*（%）				
ユニット系				
NICU	25（12.4）	8（11.4）	17（12.9）	
PICU	38（18.8）	10（14.3）	28（21.2）	
HCU	9（4.5）	0（0.0）	9（6.8）	
未記入	1（0.5）	0（0.0）	1（0.8）	

Note. [a] χ^2 検定，[b] スチューデントの *t* 検定.

この論文では，小児専門病院における医師と看護師の協働の態度を比較している.

- 医師と看護師の属性を比較する意図はどこにあったのか？
- 性別と年齢で有意差が認められるが，その後の解析においてこれらを(回帰分析などを用いて)調整したうえで評価しているわけではない.
- もし，性別と年齢が協働に影響を及ぼすのであれば，交絡因子になり得る.

菅原・笠原・石松. 日本看護科学会誌, 40, 47-55, 2020.

Welch検定を避けたい最大の理由：非正規なのに回帰分析・・・

6. 解析方法

背景要因と身体症状は度数分布を求め，PSQI-J の下位尺度と総得点の記述統計量を求め，Kolmogorov-Smirnov の正規性の検定を行った．

(中略)

統計解析は，睡眠データに正規性と直線性を認めずサンプル数も少ないため，Mann-Whitney の U 検定と Kruskal-Wallis 検定を用い，有意水準は 5%とした．

浦他. 日がん看学誌, 35, 91-101, 2021

表 5　PSQI 総得点の影響要因に関する重回帰係数

		標準回帰係数		95% 信頼区間				
	モデル	β	p	下限	上限	VIF	R^2	調整済み R^2
1	(定数)		0.936	-2.110	2.289		0.272	0.260
	睡眠薬	0.522	0.000	2.642	6.394	1.000		
2	(定数)		0.127	-4.546	0.580		0.355	0.333
	睡眠薬	0.493	0.000	2.484	6.064	1.010		
	ステロイド薬	0.288	0.007	0.484	2.930	1.010		
3	(定数)		0.074	-4.666	0.225		0.425	0.397
	睡眠薬	0.500	0.000	2.623	6.031	1.010		
	ステロイド薬	0.272	0.008	0.446	2.778	1.013		
	味覚障害の有無	0.267	0.009	0.580	3.812	1.004		
4	(定数)		0.063	-4.618	0.129		0.468	0.432
	睡眠薬	0.485	0.000	2.542	5.858	1.015		
	ステロイド薬	0.257	0.009	0.388	2.657	1.019		
	味覚障害の有無	0.281	0.005	0.741	3.886	1.009		
	痛みの有無	0.208	0.033	0.129	3.036	1.016		

a. 従属変数 PSQI 総得点

つまり，PSQI-Jは正規分布に従っていないと述べているにも関わらず・・・

- 重回帰分析の従属変数は正規分布に従っていることが仮定されます．つまり，解析方法との間に齟齬があります．
- また，「6. 解析方法」のなかに「PSQI総得点については標準化された残差の正規性p-pプロットで正規性を確認した」とありますが，残差が正規分布に従っているのであれば，従属変数が正規分布に従っていないとおかしい．

この論文では，Kolmogorov-Smirnov検定で正規性を確認したとありますが，そうするのであれば，正規p-pプロットで正規性を確認したうえで，正規分布におおよそ従っていることを確認するだけでよかったと思われます．あるいは，そうでないのであれば，何らかの変数変換が必要だったと思います．

SPSSの出力例：2標本t検定

神経障害性疼痛患者を対象に，2 種類の除痛薬(新薬，既存薬)投与後の VAS (mm)の減少量を評価している．

「分析」→「平均の比較」→「独立したサンプルのt検定」

データ

	VAS	Group
1	31	Active
2	25	Active
3	28	Active
4	29	Active
5	23	Active
6	25	Active
7	30	Active
8	25	Active
9	29	Active
10	27	Active
11	30	Active
12	20	Active
13	20	Active
14	24	Active
15	23	Control
16	23	Control
17	20	Control
18	27	Control
19	19	Control
20	15	Control
21	25	Control
22	29	Control
23	15	Control
24	13	Control
25	28	Control
26	21	Control

グループ統計量

	Group	度数	平均値	標準偏差	平均値の標準誤差
VAS	Active	14	26.14	3.592	.960
	Control	12	21.50	5.317	1.535

Active群の平均(SD)は，26.14(3.592)，Control群の平均(SD)は21.50(5.317)なので，いずれの群も平均>2SDであり，2標本t検定で十分

独立サンプルの検定

		等分散性のための Levene の検定		2 つの母平均の差の検定							
						有意確率				差の 95% 信頼区間	
		F 値	有意確率	t 値	自由度	片側 p 値	両側 p 値	平均値の差	差の標準誤差	下限	上限
VAS	等分散を仮定する	2.164	.154	2.643	24	.007	.014	4.643	1.757	1.017	8.269
	等分散を仮定しない			2.565	18.848	.010	.019	4.643	1.810	.852	8.434

p値は0.014なので，有意水準0.05のもとで有意である．したがって，除痛薬間でVASの減少量に違いがあることがわかった．なお，このときの平均値の差[95%Cis]は，4.643[1.017，8.629]であった．因みに，有意差が認められれば，信頼区間が0をまたがない．

独立サンプルの効果サイズ

		Standardizer[a]	ポイント推定	95% 信頼区間	
				下限	上限
VAS	Cohen の d	4.466	1.040	.205	1.855
	Hedges の補正	4.612	1.007	.199	1.796
	Glass のデルタ	5.317	.873	.006	1.708

a. 効果サイズの推定に使用する分母。
Cohen の d は、プールされた標準偏差を使用します。
Hedges の補正は、プールされた標準偏差と補正係数を使用します。
Glass's のデルタは、制御グループのサンプル標準偏差を使用します。

SPSS ver.27で追加された出力(効果量(effect sizeという))．

- Cohen's d (分散を個別に利用)
- Hedges's g (共通分散を利用，2標本t検定に対応)
- Glassのデルタ (一方の群の分散を利用)

Glassのデルタは，サンプルサイズが非常に不均一な場合に利用する以外には使わない方がよい．一般的には，Cohen's dあるいはHedges's gを用いる．

Cohen's dおよびHedges's gのいずれも1.00を上回っていることから，効果量が大(高い)と解釈できる．

補足：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.

効果量の例示

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).

SPSSの出力例：Wilcoxon検定 (Mann-WhitneyのU検定)

神経障害性疼痛患者を対象に，2 種類の除痛薬(新薬，既存薬)投与後の VAS (mm)の減少量を評価している．

「分析」→「ノンパラメトリック検定」→「独立サンプル」

SPSSで選択すると，新たなメニューが表示される．「自動的にグループ間の分布を比較する」を選べば，Wilcoxon検定 (SPSSではMann-WhitneyのU検定)を行う．SPSSでは，中央値検定というノンパラメトリック検定があり，「グループ間の中央値を比較する」を選択すれば実行できる(あまりメジャーではない)．

合計数	26
Mann-Whitney の U	39.500
Wilcoxon の W	117.500
検定統計量	39.500
標準誤差	19.356
標準化された検定統計量	-2.299
漸近有意確率 (両側検定)	.022
正確な有意確率 (両側検定)	.020

Mann-Whitney検定は，Wilcoxon検定と呼ばれることが多いが，まったく同じもの(同時期に異なる著者が違うアプローチで作られた検定であるため，現在の状況に至っている)．

今回のように，標本サイズが少ない場合は，正確なp値(SPSSでは正確な有意確率)を選択する．その結果，p値は0.020なので，有意である．したがって，2群間に違いが認められる．

なお，SPSSでは，ノンパラメトリック検定における効果量が出力されないが，以下の計算式

$$\text{効果量} = \frac{\text{標準化された検定統計量}}{\sqrt{\text{標本サイズ}}}$$

- 0.10未満：効果が認められない
- 0.10～0.30：効果量小
- 0.30～0.50：効果量中
- 0.50以上：効果量大

で計算が可能である．今回の場合には，$-2.299 / \sqrt{26} = -0.431$ なので，効果量中になる．

SPSSの出力例：対応のあるt検定

助産師が 5 年間の経験で分娩介助についてどのような意識の変革を起こすかを調べるため，資格取得直後と 5 年 後に，分娩介助に関する 20 項目を自己評価してもらう研究が行われた．評価スコアの合計値を比較する (柳川他, 2019).

「分析」→「平均の比較」→「対応のあるサンプルのt検定」

	前	後
1	84	88
2	78	70
3	76	80
4	82	94
5	68	72
6	64	68
7	78	82
8	66	78
9	72	72
10	64	70
11	74	78
12	78	76
13	78	76
14	88	98
15	78	76
16	82	94
17	84	82
18	82	82
19	88	90
20	78	72

メニューにおいて，「効果サイズの推定の方法」にオプションがあるが，これは，効果量を推定する際に，相関係数による調整に関するオプション(普通はしないのでデフォルトでよい)

対応サンプルの統計量

		平均値	度数	標準偏差	平均値の標準誤差
ペア 1	前	77.10	20	7.240	1.6
	後	79.90	20	8.837	1.9

前の平均(SD)が77.10(7.240)であり，後の平均(SD)が79.90(8.837)であることから，平均＞2SDなので，対応のあるt検定で十分（正規分布の差の分布も正規分布であるため）．

対応サンプルの相関係数

		度数	相関係数	有意確率 片側 p 値	有意確率 両側 p 値
ペア 1	前 & 後	20	.762	<.001	<.001

p値は0.042なので，有意水準0.05のもとで有意である．したがって，助産師の5年間の経験で，分娩解除に関する評価スコアに変化が認められた．

対応サンプルの検定

		対応サンプルの差 平均値	標準偏差	平均値の標準誤差	差の 95% 信頼区間 下限	上限	t 値	自由度	有意確率 片側 p 値	両側 p 値
ペア 1	前 - 後	-2.800	5.745	1.285	-5.489	-.111	-2.179	19	.021	.042

対応のあるサンプルの効果サイズ

			Standardizer[a]	ポイント推定	95% 信頼区間 下限	上限
ペア 1	前 - 後	Cohen の d	5.745	-.487	-.946	-.017
		Hedges の補正	5.862	-.478	-.927	-.017

ここでは，前ー後で計算されているため，負値をとっている（スコアが増加）．効果量の絶対値が0.4を上回っているだけなので，効果量としては小さいと解釈される．

柳川 堯 他：看護・リハビリ・福祉のための統計学，近代科学社，2019.

SPSSの出力例：Wilcoxon符号付き順位和検定

助産師が 5 年間の経験で分娩介助についてどのような意識の変革を起こすかを調べるため，資格取得直後と 5 年 後に，分娩介助に関する 20 項目を自己評価してもらう研究が行われた．評価スコアの合計値を比較する (柳川他, 2019).

「分析」→「ノンパラメトリック検定」→「対応サンプル」

合計数	20
検定統計量	132.500
標準誤差	22.798
標準化された検定統計量	2.062
漸近有意確率 (両側検定)	.039

p値は0.039なので，有意水準0.05のもとで有意である．したがって，助産師の5年間の経験で，分娩解除に関する評価スコアに変化が認められた．なお，SPSSでは，exactパッケージを購入していれば，正確なp値が計算できる．

なお，SPSSでは，対応のあるデータにおいても，ノンパラメトリック検定における効果量が出力されないが，以下の計算式 (Wilcoxon検定と同じ)

$$\text{効果量} = \frac{\text{標準化された検定統計量}}{\sqrt{\text{標本サイズ}}}$$

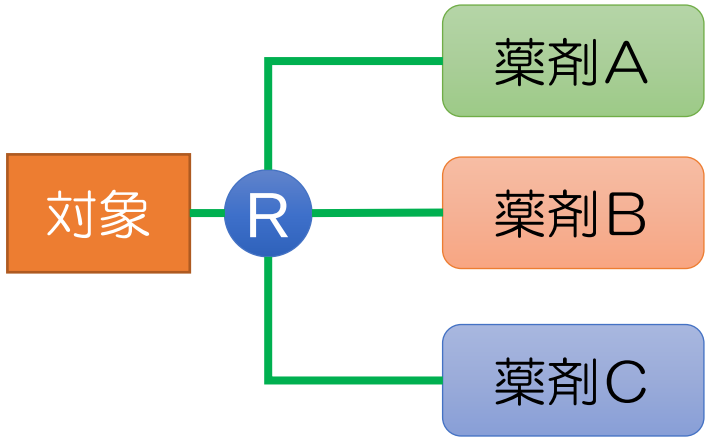
- 0.10未満：効果が認められない
- 0.10～0.30：効果量小
- 0.30～0.50：効果量中
- 0.50以上：効果量大

で計算が可能である．今回の場合には，$-2.062 / \sqrt{20} = -0.461$ なので，効果量中になる．

多標本データの比較：一元配置の分散分析

Clinical Question

神経障害性疼痛患者がランダムに割り付けられ，3種類の除痛薬(A,B,C)のいずれかが投与された.



Data structure

Ⓟ：被験者

薬剤A	ⓅⓅⓅⓅⓅⓅⓅⓅⓅⓅⓅ
薬剤B	ⓅⓅⓅⓅⓅⓅⓅⓅⓅⓅ
薬剤C	ⓅⓅⓅⓅⓅⓅⓅⓅⓅⓅⓅⓅ

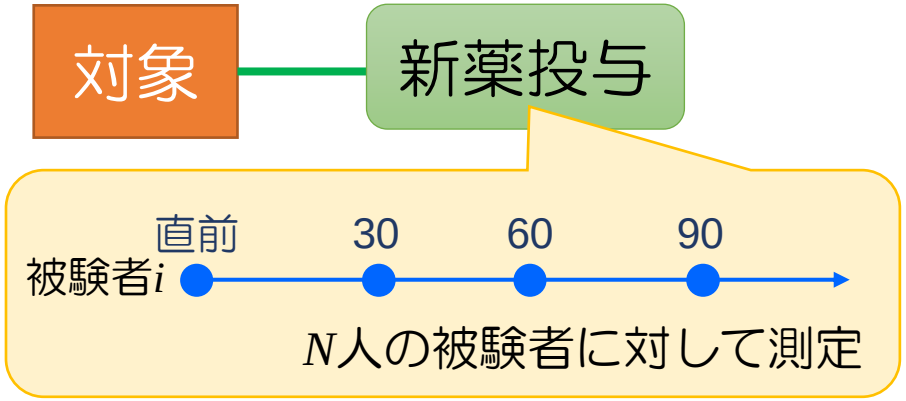
3薬のVAS減少量に違いがあるか？

- パラメトリック検定：一元配置の分散分析
- ノンパラメトリック検定：**Kruskal-Wallis検定**

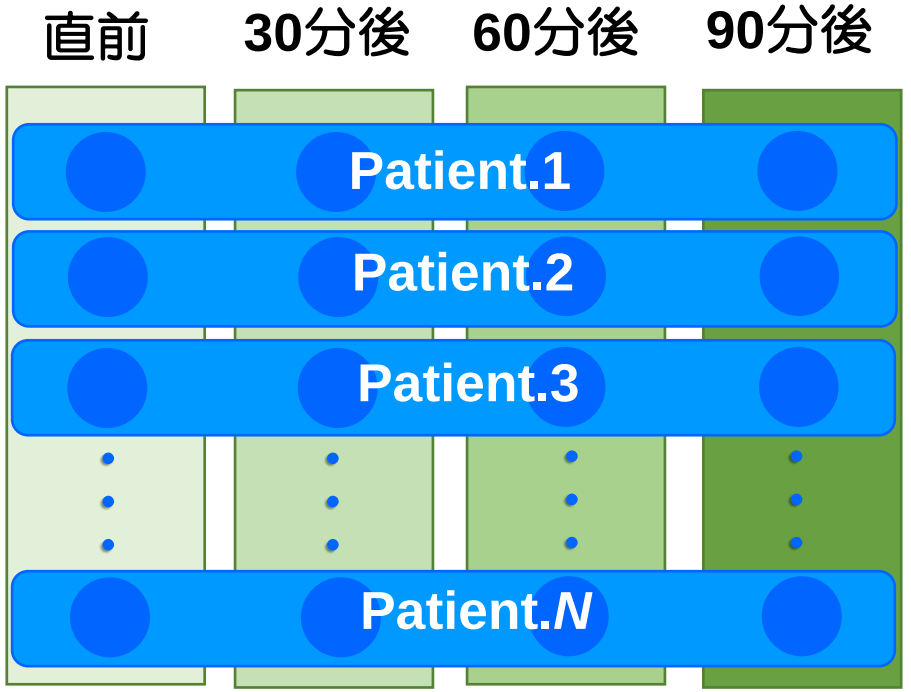
多標本データの比較：繰り返し測定の分散分析(対応のあるデータ)

Clinical Question

*N*名の神経障害性疼痛患者に対して，新薬の除痛薬を投与し，投与直前，30分後，60分後，90分後のVASの変化を調査した.



Data structure



- パラメトリック検定：繰り返し測定の分散分析
- ノンパラメトリック検定：**Friedman検定**

分散分析の概念図

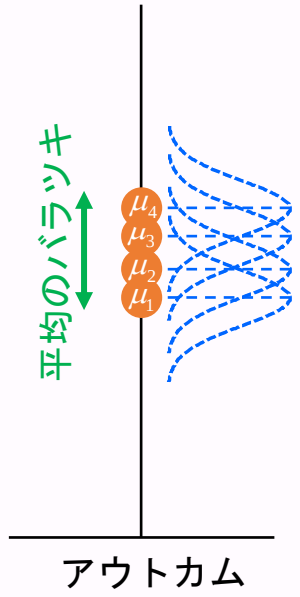
分散分析とは，3群以上の母平均に違いがあるか否かを評価する方法である.

分散分析における仮説
帰無仮説H_0 ：すべての群の母平均が等しい
対立仮説H_1 ：帰無仮説H_0ではない

なお，2群の場合の分散分析とt検定の結果は一致する.

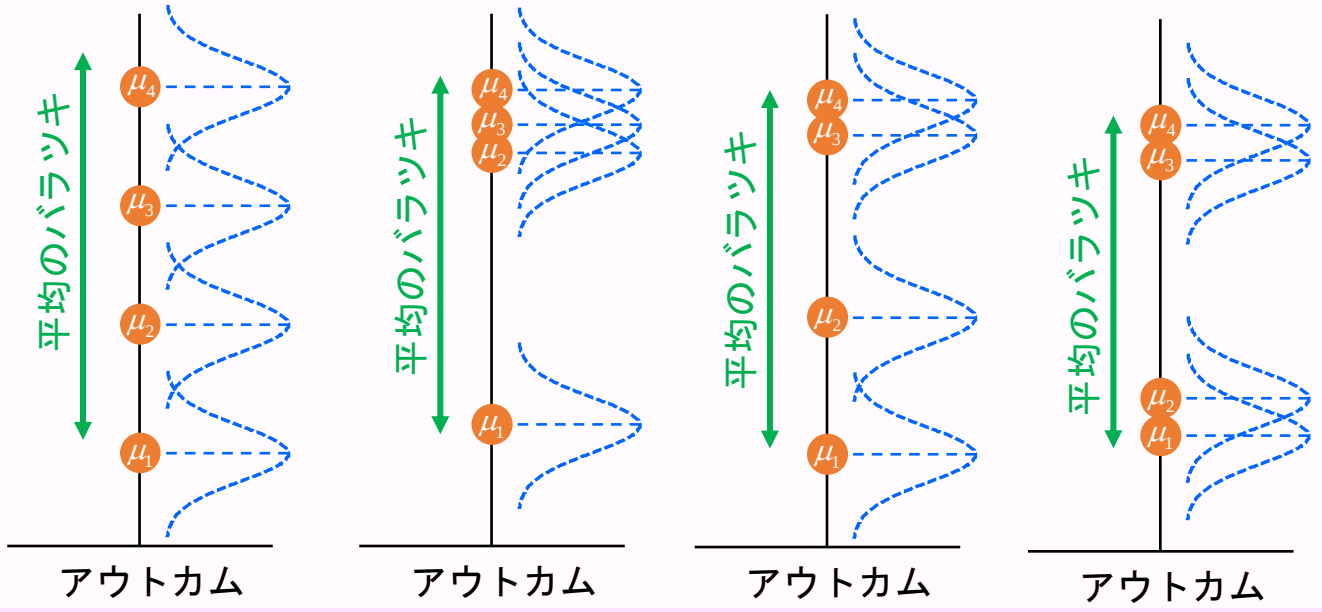
帰無仮説H_0が棄却できない(有意でない) 状況

平均値がほぼ同じ場合には，平均のバラツキ(＝分散)が小さく，すべて同じであれば，0になる.

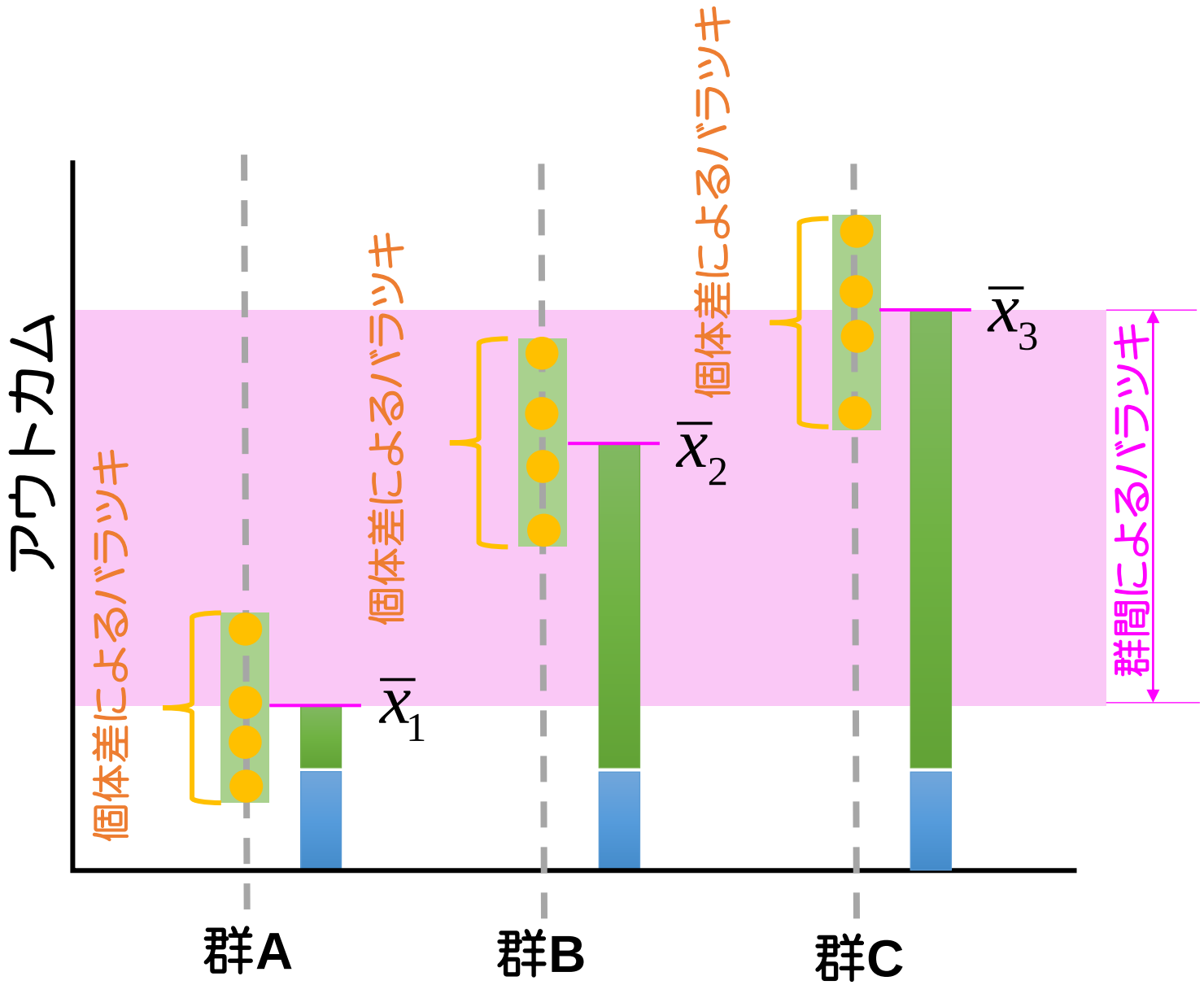


対立仮説H_1 が正しい(有意)な状況

平均値が異なる場合には，平均のバラツキ(＝分散)が大きい.



分散分析表の検定(F検定)の意味



分散分析では個体差によるバラツキ(分散)と群間のバラツキ(分散)が評価される.

□ 有意差が認められる場合

群間によるバラツキ(分散) ＞ 個体差によるバラツキ(分散)

□ 有意差が認められない場合

群間によるバラツキ(分散) ＜ 個体差によるバラツキ(分散)

分散分析ではこの関係性を応用して

$$\text{F値} = \frac{\text{群間によるバラツキ (分散)}}{\text{個体差によるバラツキ(分散)}}$$

が評価される.

3群以上の比較におけるパラメトリック検定とノンパラメトリック検定

群の数	パラメトリック検定	ノンパラメトリック検定
2群	• 2標本t検定 (独立) • 対応のあるt検定(対応)	• Wilcoxon(Mann-Whitney)検定 (独立) • Wilcoxon符号付順位和検定 (対応)
3群以上	• 一元配置の分散分析 (独立) • 反復測定分散分析 (対応)	• Kruskal-Wallis検定 (独立) • Friedman検定 (対応)

3群以上の比較における注意点

3群以上の比較においては，3群のどこかに差があることはわかるだけであり，違いがどこにあるかはわからない

A B C の比較をした結果，有意差が認められたとき，以下のいずれかである

- A B のみ有意
- A C のみ有意
- B C のみ有意
- A B と A C が有意
- A B と B C が有意
- A C と B C が有意
- A B と A C と B C が有意

Kruskal-Wallis検定の誤用

6. 分析方法

統計解析には，SPSS ver. 21 を使用した．まず記述統計において単純集計を行ったあと，対象者を「非喫煙群」「紙巻きたばこ群」「新型たばこ群」に分類し，それぞれの喫煙をすることへの認識の関係性についての統計学的有意差をχ^2検定で分析し，有意確率（p 値）5% 水準未満である場合を有意とした．新型たばこの含有成分についての認識にはクラスカル・ウォリス検定を行い，有意確率（p 値）5% 水準未満である場合を有意とした．

表 4　新型たばこの有害成分の認識

	全体	思う	思わない・わからない	p			
	人（%）	人（%）	人（%）				
新型たばこ群	92（100.0）	8（8.7）	84（91.3）	p=0.116§			
紙巻きたばこ群	48（100.0）	11（22.9）	37（77.1）		p=0.043§		p=0.034*
非喫煙群	612（100.0）	118（19.3）	494（80.7）	p=1.000§			

*χ^2検定　§クラスカル・ウォリス検定

新型たばこの含有成分についての知識を問う質問に関しては，「新型たばこ群」では，「思う群」8 人（8.7%），「思わない・わからない群」84 人（91.3%）であり，「紙巻きたばこ群」では，「思う群」11 人（22.9%），「思わない・わからない群」37 人（77.1%），「非喫煙群」では，「思う群」118 人（19.3%），「思わない・わからない群」494人（ 80.7%）であった．「新型たばこ群」「 紙巻きたばこ群」「非喫煙群」をχ2 検定を用いて検定した結果，有意差が認められた（p = 0.034）．クラスカル・ウォリス検定を用いて検定した結果，「新型たばこ群」と「 非喫煙群」の間で有意差が認められた（p = 0.043）．

横田・原田・櫻井. 日本産業看護学会誌, 8, 18-26, 2021.

この論文におけるKruskal-Wallis検定の利用のおかしな点

- 順序カテゴリカルデータにおいて，ノンパラメトリック検定を用いる場合があるが，今回は2値データである．
- Kruskal-Wallis検定は3群以上の比較に用いるものの，この論文のようなペアワイズな比較を行うものではない．
  (2群でのKruskal-Wallis検定はWilcoxon検定と同じなので，SPSSがエラーにならなかっただけ)

本来は，Fisherの正確検定 (orカイ2乗検定)を行ったうえで，p値を多重比較しなければならない．

正しい解析を行った結果

表 4　新型たばこの有害成分の認識

	全体	思う	思わない・わからない	p
	人 (%)	人 (%)	人 (%)	
新型たばこ群	92 (100.0)	8 (8.7)	84 (91.3)	p=0.116§
紙巻きたばこ群	48 (100.0)	11 (22.9)	37 (77.1)	p=0.043§
非喫煙群	612 (100.0)	118 (19.3)	494 (80.7)	p=1.000§

* χ^2 検定　§ クラスカル・ウォリス検定

Fisher: 0.105
カイ2乗：0.115

Fisher: 0.038*
カイ2乗：0.060

p=0.034*

Fisher: 1.000
カイ2乗：1.000

- ここでは，Fisherの正確検定とカイ2乗検定を行ったうえでBonferroniの多重性の調整を行った.
- Fisherの正確検定を用いても，結論は変わらない (カイ2乗検定では有意差が認められない).

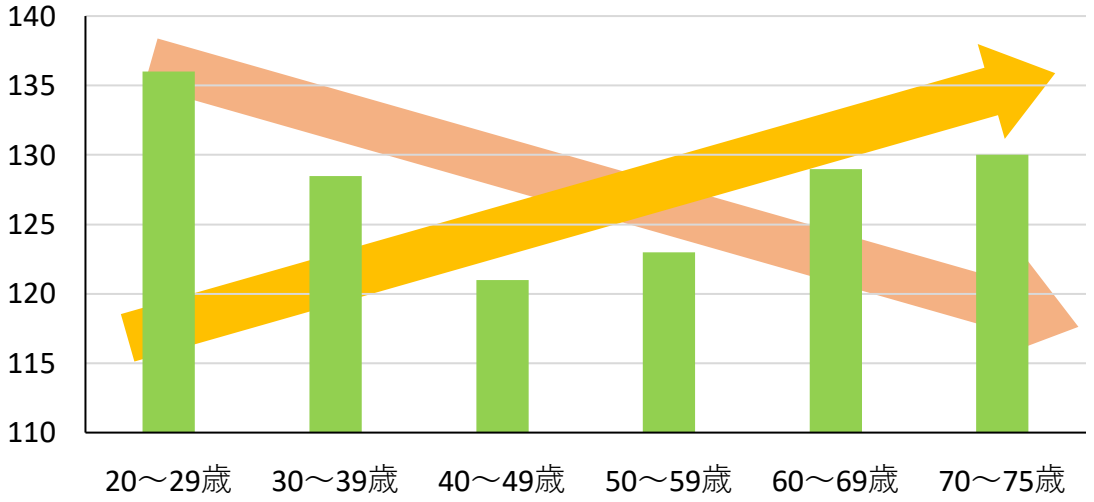
群に順序関係が存在する場合の検定

項目	内訳	人数	%	セルフケア能力 中央値（最小値一最大値）	p	検定方法
年齢 mean±SD		52.2	(13.1)		n.s.	#3
（範囲 22–75 歳）	20–29 歳	5	5.4	136 (78-144)		
	30–39 歳	12	13.0	128.5 (96-152)		
	40–49 歳	21	22.8	121 (83-157)		
	50–59 歳	21	22.8	123 (67-163)		
	60 歳 –69 歳	26	28.3	129 (76-159)		
	70 歳 –75 歳	7	7.6	130 (93-167)		

Spearmanの順位相関係数

中島・安東. 日本糖尿病教育・看護学会誌, 25(1), 83-92, 2021

Spearmanの順位相関係数は，相関関係を評価するものであり，今回のように群に順序関係がある場合には，トレンド検定(傾向検定)を用いるべき．SPSSの場合には，Jonckheere-Terpstra (ヨンクヒール・タプストラ)検定を用いるほうが良い(ちなみに，Kendallの順位相関と一致する).



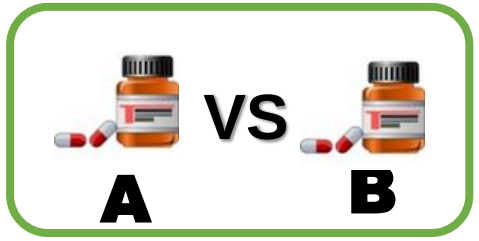
SPSSでは，
「ノンパラメトリック検定」
→「独立サンプル」
→「目的」タブ「検定のカスタマイズ」
の順に選んでいくと，Jonckheere-Terpstra検定を選択

なお，パラメトリック検定の場合には，ANOVAの対比をつくれば良い (詳しくは第3回で説明します).

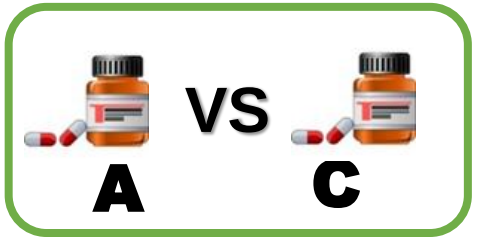
多重比較の動機

いま，3群での無作為化比較試験を検討している (A群，B群，C群)．しかしながら，各群においても用いる薬剤は同じである．

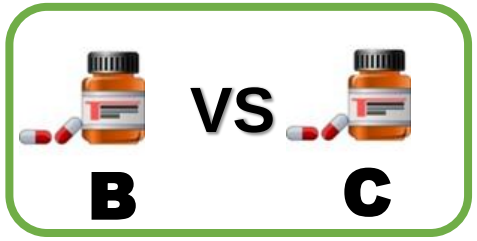
$H_{A,B}$



$H_{A,C}$



$H_{B,C}$



有意水準α=0.05とは，本来違いがない(有意差がない)にも関わらず，「有意差あり」という判断してしまう確率を5%未満にすることである．すべての比較において，有意水準α=0.05を割り振ってみる．

$H_{A,B}$ α=0.05　　$H_{A,C}$ α=0.05　　$H_{B,C}$ α=0.05

当然だが，これらの対比較のいずれもが「有意差なし」とならなければいけないが，5%の確率でそれぞれが誤ってしまう．どらかひとつの比較でも「有意差あり」と言ってしまうシチュエーションとその確率は，

(A vs B) (A vs C)は有意でない，(B vs C)は有意 = (1-0.05) × (1-0.05) × 0.05 = 0.0451

(A vs B) (B vs C)は有意でない，(A vs C)は有意 = (1-0.05) × (1-0.05) × 0.05 = 0.0451

(A vs C) (B vs C)は有意でない，(A vs B)は有意 = (1-0.05) × (1-0.05) × 0.05 = 0.0451

(A vs B)は有意でない，(A vs C) (B vs C)は有意 = (1-0.05) × 0.05 × 0.05 = 0.0024

(A vs C)は有意でない，(A vs B) (B vs C)は有意 = (1-0.05) × 0.05 × 0.05 = 0.0024

(B vs C)は有意でない，(A vs B) (A vs C)は有意 = (1-0.05) × 0.05 × 0.05 = 0.0024

なので，0.0451+0.0451+0.0451+0.0024+0.0024+0.0024 = 0.1425となる．本来の有意水準はα=0.05なので，大きくなっている(つまり，第1種の過誤が増大している)．それを補正するのが多重比較である．

多重比較には大きく分けて2種類が存在する

■ **p値 (or 有意水準)を調整する方法**

- **Bonferroni型 (SPSSにはこちらのみ実装)**
- **Holm型 (ステップダウン法，p値が最も小さな比較から順に多重比較する)**

● あらゆる検定に適用することができる (連続変数でなくても比較可能).

■ ANOVA (or Kruskal-Wallis検定)を事前に行う必要はない.

× 有意差が出にくい(ANOVA,Kruskal-Wallisの結果が有意でもすべての比較で有意差が認められない場合がある).

■ **ANOVA (or Kruskal-Wallis検定)に基づいて行う方法**

- ペアワイズ比較 (各群のすべての組み合わせを比較する場合)
- コントロール群との比較を行う場合

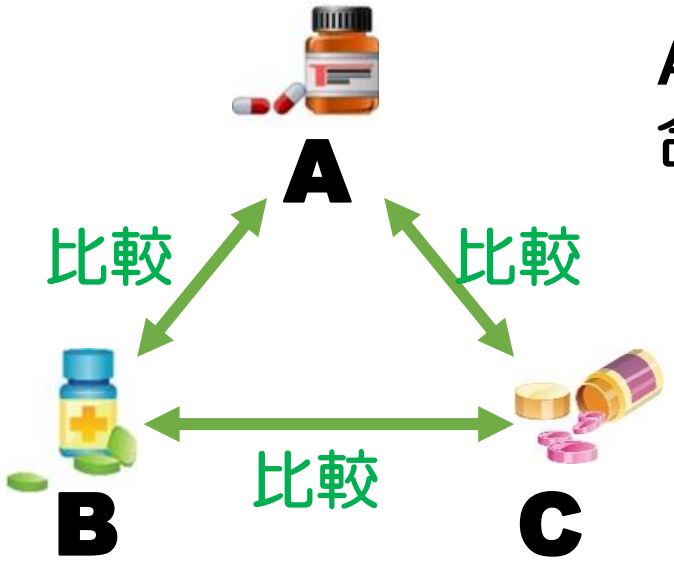
● ANOVA (or Kruskal-Wallis検定)で有意差があれば，必ずどこかの比較で有意差が認められる.

■ 必ずANOVA (or Kruskal-Wallis検定)を事前にしなければならない.

× 適応できる変数のタイプが決まっている．また，元となる多標本比較の方法が存在する.

多重比較の取捨選択について

P値に基づく方法(1)：Bonfferoniの方法

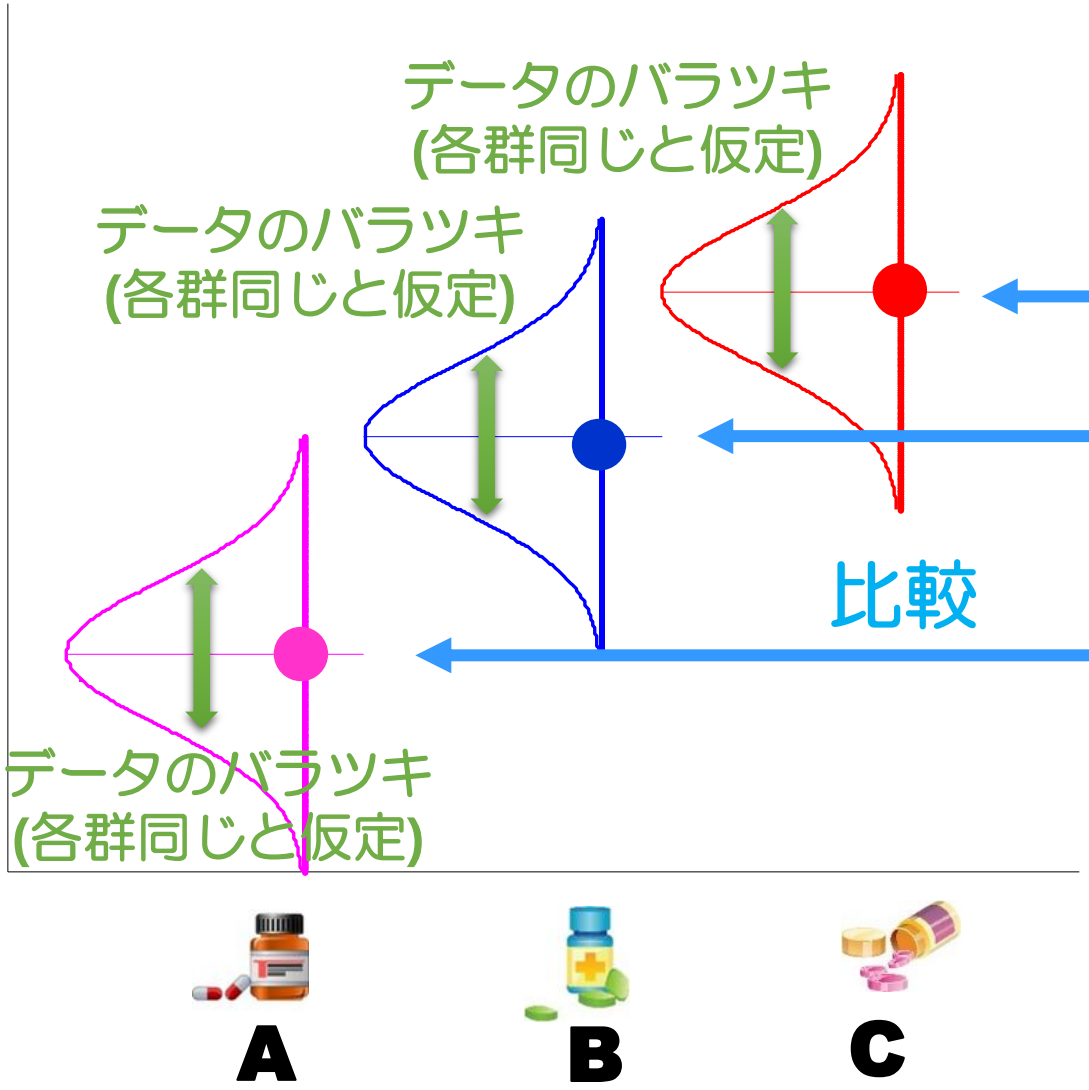


A群，B群，C群の対比較の場合で全体での有意ｐがαの場合には

- 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．

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)

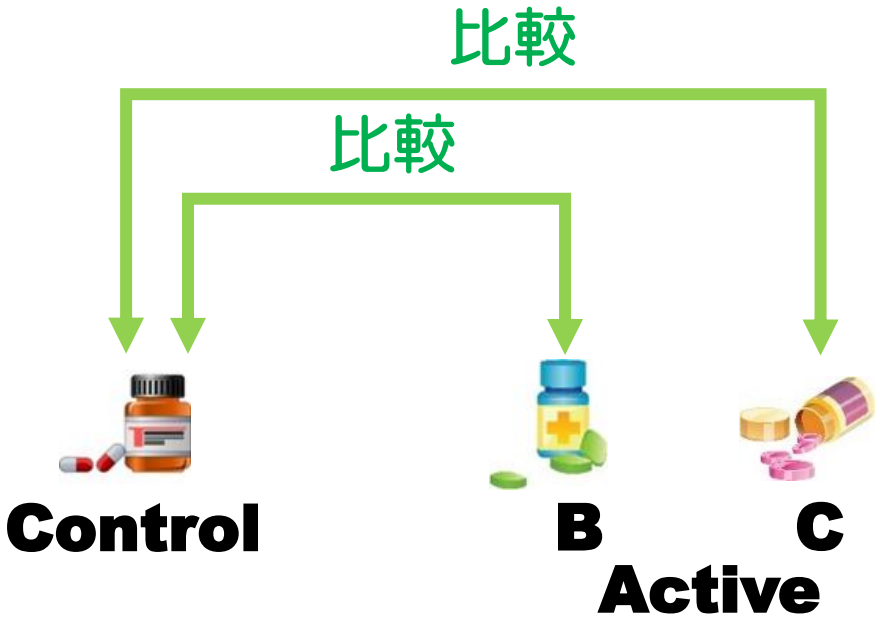


一元配置分散分析において有意だった場合には，Tukeyの方法でもいずれかの対比較において有意になっている．

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.

ちなみに，Bonferroni法で実施する場合には，次のようになる

- Control群　対　B群 ➡ 有意水準α/2と比較(or p値を2倍)
- Control群　対　C群 ➡ 有意水準α/2と比較(or p値を2倍)

Tukeyの方法のノンパラメトリック版には，Steelの方法がある.

SPSSで利用できる多重比較

手法	検定の形式
LSD	t検定によるペアワイズ比較．最も古典的だが，理論的に問題がある．
Bonfferoni	t検定によるP値を調整するタイプのペアワイズ比較，P値×比較回数によってp値を調整する．
Sidak	t検定によるP値を調整するタイプのペアワイズ比較，Bonfferoniの修正版 (比較間の独立性を仮定)
Scheffe	*n*個の各群におけるすべての対比の中で，有意なものをさがす多重比較，有意差が出にくい．
R-E-G-WのF	Tukeyの方法に基づくステップダウン型のアルゴリズムに基づく多重比較
R-E-G-WのQ	Tukeyの方法に基づくステップダウン型のアルゴリズムに基づく多重比較 (上記の改良版)
Student-Newman-Keuls	Tukeyの方法の改良版，有意差が高い組み合わせから開始する方法，否定的な意見がある．
Tukey	一元配置分散分析に対応するペアワイズ比較 (一元配置分散分析で有意ならばどこか有意になる)
Tukeyのb	Tukeyの方法とStudent-Newman-Keulsの中間的な臨界値をもつペアワイズ比較
Duncan	Student-Newman-Keulsの改良版
HochbergのGT	Tukeyに類似したペアワイズ比較
Gabriel	Tukeyに類似．ただし，標本サイズが異なる場合には，Hochbergよりも強力
Waller-Duncan	Bayes的なアプローチに基づくペアワイズ比較
Dunnett	コントロール群との比較を意図した多重比較法
TamhaneのT2	Games-Howellの改良版
DunnettのT3	DunnettのCの改良版，小標本の場合にはGames-Howellよりも良いとされる．
Games-Howell	Tukeyの方法の不等分散版
DunnetのC	Games-Howellの改良版．

SPSSの出力例：一元配置の分散分析

いま，14 名の疼痛患者が服薬した除痛薬(1,2,3)毎にグループに分け，それぞれの群での投与後の痛みの程度を測定した．

「分析」→「平均の比較」→「一元配置分散分析」

「分析」→「一般線型モデル」→「1変量」でも実行可能

「その他の検定」からTukey, Bonfferoni, Gabrielを選択

	薬剤	痛みの程度
1	3	12.40
2	1	7.69
3	3	14.00
4	1	9.69
5	3	11.60
6	1	8.89
7	1	6.94
8	1	2.13
9	1	7.26
10	1	5.87
11	2	12.90
12	3	12.20
13	1	7.20
14	3	13.90
15	1	8.18
16	2	16.60
17	3	9.41
18	3	11.20
19	2	8.35
20	1	7.24
21	1	6.81
22	2	9.81
23	1	6.67
24	1	6.98
25	1	7.07
26	3	2.40
27	2	7.84
28	2	3.84
29	2	9.42
30	1	7.00
31	1	7.00
32	1	5.00
33	1	8.00

分散分析

痛みの程度

	平方和	自由度	平均平方	F 値	有意確率
グループ間	99.517	2	49.758	6.271	.005
グループ内	238.037	30	7.935		
合計	337.554	32			

有意水準0.05のもとで有意であることから，薬剤によって，痛みの程度が異なることがわかった．

その後の検定

多重比較

従属変数: 痛みの程度

	(I) 薬剤	(J) 薬剤	平均値の差 (I-J)	標準誤差	有意確率	95% 信頼区間 下限	95% 信頼区間 上限
Tukey HSD	1	2	-2.84397	1.25472	.076	-5.9372	.2493
		3	-3.90986*	1.19693	.007	-6.8606	-.9591
	2	1	2.84397	1.25472	.076	-.2493	5.9372
		3	-1.06589	1.45785	.747	-4.6599	2.5281
	3	1	3.90986*	1.19693	.007	.9591	6.8606
		2	1.06589	1.45785	.747	-2.5281	4.6599
Bonferroni	1	2	-2.84397	1.25472	.092	-6.0256	.3377
		3	-3.90986*	1.19693	.008	-6.9450	-.8748
	2	1	2.84397	1.25472	.092	-.3377	6.0256
		3	-1.06589	1.45785	1.000	-4.7626	2.6308
	3	1	3.90986*	1.19693	.008	.8748	6.9450
		2	1.06589	1.45785	1.000	-2.6308	4.7626
Gabriel	1	2	-2.84397	1.25472	.077	-5.9271	.2391
		3	-3.90986*	1.19693	.007	-6.8703	-.9494
	2	1	2.84397	1.25472	.077	-.2391	5.9271
		3	-1.06589	1.45785	.846	-4.7411	2.6093
	3	1	3.90986*	1.19693	.007	.9494	6.8703
		2	1.06589	1.45785	.846	-2.6093	4.7411

*. 平均値の差は 0.05 水準で有意です。

p値の要約
(T: Tukey, B: Bonfferoni, G: Gabriel)

	1	2	3
2	0.076 (T) 0.092 (B) 0.077 (G)	-	-
3	0.007 (T) 0.008 (B) 0.007 (G)	0.747 (T) 1.000 (B) 0.846 (G)	-

SPSSの出力例：Kruskal-Wallis検定

いま，14 名の疼痛患者が服薬した除痛薬(1,2,3)毎にグループに分け，それぞれの群での投与後の痛みの程度を測定した．

「分析」→「ノンパラメトリック検定」→「独立サンプル」

	薬剤	痛みの程度
1	3	12.40
2	1	7.69
3	3	14.00
4	1	9.69
5	3	11.60
6	1	8.89
7	1	6.94
8	1	2.13
9	1	7.26
10	1	5.87
11	2	12.90
12	3	12.20
13	1	7.20
14	3	13.90
15	1	8.18
16	2	16.60
17	3	9.41
18	3	11.20
19	2	8.35
20	1	7.24
21	1	6.81
22	2	9.81
23	1	6.67
24	1	6.98
25	1	7.07
26	3	2.40
27	2	7.84
28	2	3.84
29	2	9.42
30	1	7.00
31	1	7.00
32	1	5.00
33	1	8.00

仮説検定の要約

	帰無仮説	検定	有意確率[a,b]	決定
1	痛みの程度 の分布は 薬剤 のカテゴリで同じです。	独立サンプルによる Kruskal-Wallis の検定	.003	帰無仮説を棄却します。

a. 有意水準は .050 です。

b. 漸近的な有意確率が表示されます。

独立サンプルによる Kruskal-Wallis の検定

痛みの程度 から 薬剤

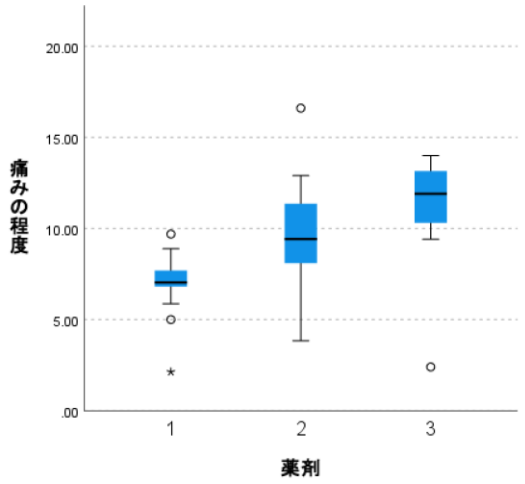
独立サンプルによる Kruskal-Wallis の検定の要約

合計数	33
検定統計量	11.680[a]
自由度	2
漸近有意確率 (両側検定)	.003

a. 検定統計量は同順位の調整が行われています。

有意水準0.05のもとで有意であることから，薬剤によって，痛みの程度が異なることがわかった．ちなみに二つは同じ意味である．

独立サンプルによる Kruskal-Wallis の検定



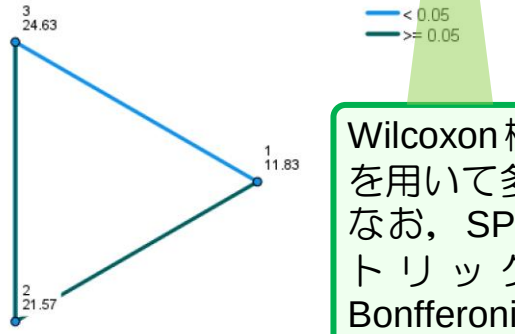
薬剤 のペアごとの比較

Sample 1-Sample 2	検定統計量	標準誤差	標準化検定統計量	有意確率	調整済み有意確率[a]
1-2	-9.738	4.307	-2.261	.024	.071
1-3	-12.792	4.108	-3.114	.002	.006
2-3	-3.054	5.004	-.610	.542	1.000

各行は、サンプル 1 とサンプル 2 の分布が同じであるという帰無仮説を検定します。
漸近的な有意確率 (両側検定) が表示されます。有意水準は .050 です。

a. Bonferroni 訂正により、複数のテストに対して、有意確率の値が調整されました。

薬剤 のペアごとの比較



Wilcoxon 検定後に Bonfferoni を用いて多重比較している．なお，SPSSではノンパラメトリックな多重比較は Bonfferoniのみ

和歌山医大において無料で利用可能な統計パッケージに実装されている多重比較

SPSS®

	p値の調整に基づく方法		ANOVA (KW)に基づく方法	
	Bonferroni型	Holm型	ペアワイズ	対照群との比較
パラメトリック	○ Bonferroni, Sidak	×	○ Tukey他多数	○ Dunnett
ノンパラメトリック	○ Bonferroni	×	×	×

jmp®

	p値の調整に基づく方法		ANOVA (KW)に基づく方法	
	Bonferroni型	Holm型	ペアワイズ	対照群との比較
パラメトリック	×	×	○ Tukey	○ Dunnett
ノンパラメトリック	×	×	○ Still-Dwass	○ Still

EZR

	p値の調整に基づく方法		ANOVA (KW)に基づく方法	
	Bonferroni型	Holm型	ペアワイズ	対照群との比較
パラメトリック	○ Bonferroni	○ Holm	○ Tukey	○ Dunnett
ノンパラメトリック	○ Bonferroni	○ Holm	○ Still-Dwass	○ Still

Thank you for your kind attention

shimokaw@wakayama-med.ac.jp

toshibow2000@gmail.com