〔日本生理誌（1996）**58**, 307-320〕

シリーズ「医学統計処理の問題点」

医学統計処理の問題点

—誤用のはなしⅢ—

栗谷典量
（久留米大学医学部小児科）

t検定の問題点 2

医学関連の論文・報告書などに使われる医学統計で，誤用の多いt検定（平均値・標準偏差の扱い）について，前回に続いて2つの問題点を取り上げる．

誤用のはなしⅠでは

① 重複登録（症例とデータ数のアンバランス）

誤用のはなしⅡでは

② 非計量値（順序尺度）の計量値扱い．

③ 歪んだ分布の（正規分布でない）データの処理

の3項目を取り上げた．

今回は

④ 比較する二群のデータの分散の差が大きい．（等分散の検定で有意）

⑤ データの対応の有無を無視する．（検定法選択の誤り）

⑥ 多群間の平均値の比較法．（多重比較法の誤解）

の残り3項目中の，④⑤の2項目を取り上げる．

4．比較する2群の等分散性のチェック

t検定は，2群のデータの分散（variance，バラツキ）に大きな差がない（等分散である）ことを前提として導かれた手法である．分散に差がないことを確認するには，あらかじめF検定と呼ばれる手法で有意差がないことを確認しておく必要がある．目的とする平均値の差の検定の前に，もう一つ別の検定が必要な訳だ．F検定の結果，二つの分散に有意差がなければ，t検定が使える．もし差が有意の場合は，等分散とみなせないので次の(a)または(b)の対策をとる．

(a) 近似法を採用する．

(b) ノンパラメトリック検定を採用する．

などの方法で対処する．

(a) 近似法による法

i Aspin-Welch（Welch）の法

ii Cochran-Cox の法

（Aspin-Welch を略して Welch test と表現することが多い）

この二法に大差はないが Welch の法がよいといわれており多用される．Welch の法の計算は特に煩わしい程のことはないが，自由度に端数がつき，P値の算出が必要となる．

（t分布表には端数自由度の記載がない）

(b) ノンパラメトリック法（よくノンパラと略称される）による法

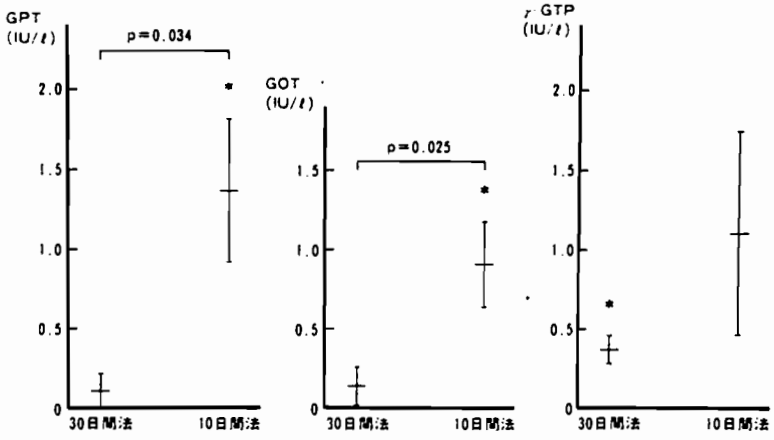
統計的仮設検定の中でも，母集団分布に正規分布など特定の分布を仮定しないで検定を行うものや，母集団分布を仮定するとしても，母数に関してではない仮設の検定をするものをノンパラメトリック「分布によらない（distribution-free test）」検定と称する．代表的手法にカイ二乗検定，フィッシャーの直接確率，ウィルコクソンの順位和検定（マンホイットニイのU検定と同じ理論の検定法）などがある．平均値のt検定が等分散でないために使えないとき，ノンパラ手法ではウィルコクソン順位和検定法が使用できる．この方法を用いるとデータの等分散性に煩わされないですむ．

医学論文中には，等分散性の確認をしないままt検定を行った解析例は多数存在する．統計処理の誤用例として取り上げると際限がない程である．この過りは常に判断や結論に悲劇的な影響を来す程のひどい誤りと言うわけではないが，統計理論的に誤りと定義されていることから無視は避けた方がよいと思う．

誤用例　4－1

肥満児に対する低カロリー食の有用性の研究論文．GPT, GOT, γGTP の減少量が検討の対象であるが，比較グラフの標準偏差の幅をみれば，10日間対30日間療法の2群の分散が等分散でないことが一見してわかる．Student's t test で処理されている．

解析対象の測定値が GPT, GOT, γGTP であるから，等分散性の検討以前に正規性の検討が必要，このデータを直接 t 検定で処理するのは問題．対数に変換して対処するとか，順位検定(ノンパラメトリック法)を考えるなどが基本．



GPT, GOT, γGTP の1日あたりの減少値

誤用例　4－2

2群の尿蛋白を t 検定で比較されているが，分散の差は有意．

(F＝31.250　P＝0.0015＊＊)

両群の発症時検査値比較

	腎不全群	腎機能保持率	
N	3	6	
尿蛋白(g/dl)	3.1±2.5	0.7±0.4	p<0.05
TP(g/dl)	6.4±1.3	6.8±1.0	
CCr(ml/min/1.73 m^2)	70.8±27.7	81.3±73.0	
Cr(mg/dl)	0.83±0.50	0.65±0.37	
ΔCCr/year	5.3±11.0	1.2±5.9	

誤用例　4－3

心音図所見の2群比較の論文．検定は t 検定が使われている．

Q-I/RR の SD は，0.14対0.03，で等分散の検定で F＝21.8851＊＊＊，等分散でない．差の検定結果は t 検定で P＝0.03＊で有意と書かれているが，t 検定は不適当．平均値の差を Welch test でみると，df＝37.6139　t＝0.4117　P＝0.6829 ns となり，有意差は認められず，論文の表現は言い過ぎの誤り．なお，この解析は等分散の問題だけでなく，計算値すべての誤差が桁外れに大きい．プログラム・ソフトの欠陥のチェックが必要．I/RR も同様 F＝83.6159＊＊＊，等分散でない．

(本論文，表1のP値は，分散の検定と関係なく検定の計算値はすべて誤り)

ダウン群と対照群の心音図所見の関連：表1は各心音図所見について，ダウン群と対照群との比較である．(Q-I)/RRは，ダウン群で 0.16±0.14 sec, 対照群で 0.15±0.03 sec であり(P＝0.03)，(Q-II)/RR は，ダウン群で 0.57±0.08 sec，対照群で 0.61±0.07 sec であり，P＝

0.01 と有意差がみられた．さらに，(I-II)/RR は，ダウン群で 0.47±0.04 sec，対照群で 0.50±0.03 sec で，P=0.006 と有意差があり，ダウン群では同じ程度の肺高血圧では，収縮期の時間が対照群より短縮していたことになる．

II_P/II_A は，ダウン群 1.4±1.0，対照群 1.2±0.9で P=0.40 と有意差はなかった．

表1　ダウン群と対照群の心音所見の比較
II/RR と II_A/II_P に関しては，大動脈成分(II_P)の判別が心音図上不可能な例があったため，ダウン群35例，対照群30例の比較である．

	ダウン群	対照群	t 検定
Q-I/RR	0.16±0.14	0.15±0.03	P=0.03*
Q-II/RR	0.57±0.08	0.61±0.07	P=0.01*
I/RR	0.23±0.06	0.36±0.55	P=0.44
II/RR	0.04±0.02	0.05±0.02	P=0.06
I-II/RR	0.48±0.04	0.50±0.03	P=0.01*
II_P/II_A	1.40±1.00	1.20±0.90	P=0.40

数値：平均値±標準偏差

誤用例　4－4

ラットの心機能の変化の実験．16項目のｔ検定を行い12項目が等分散のＦ検定で有意差あり，分散に問題はあるが差の検定結果と結論に影響を来した項目はなく，実害は出ていないが，他の検定法を考えては？

ところが 14/16 は TS 群の SD が大きく，うち 10/14 は有意差有り．従って平均値の差だけでなく TS 群のデータはバラツキが大きくなる特性・素因があるのかも．14/16 の偏りは有意(P=0.0042**　2項検定)

Activities of Myocardial Enzyme of Heart Muscles from Rats

	Control group (n=5)	TS group (n=5)	F_0(df=4, 4) P_0	t_0(df=8) P_0 (筆者追加)
SDH(μmoles/min/mg protein)				
RVFW	3.87+0.89	15.34+5.37**	36.4056	−4.7118
			0.0021**	0.0015**
LVFW	8.55+6.50	8.97+1.07	36.9028	−0.1426
			0.0021**	0.8902 n. s.
MDH(I. U./mg protein)				
RVFW	5.98+0.92	10.38+2.93*	10.1428	−3.2037
			0.0227*	0.0125*
LVFW	6.17+0.16	5.88+1.33	69.0977	0.4841
			0.0006***	0.6413 n. s.
LDH(I. U./mg protein)				
RVFW	9.11+1.73	10.50+2.18	1.5879	−1.1168
			0.3326 n. s.	0.2965 n. s.
LVFW	9.92+1.86	9.31+1.93	1.0767	0.5089
			0.4723 n. s.	0.6246 n. s.
PFK(I. U.×10^2/mg protein)				
RVFW	4.12+1.64	6.04+1.92	1.3706	−1.7002
			0.3837 n. s.	0.1275 n. s.
LVFW	8.90+3.80	7.09+1.45	6.8680	0.9951
			0.0444*	0.3488 n. s.

CK (I. U./mg protein)				
RVFW	16.84+1.36	20.83+4.04	8.8244	−2.0930
			0.0290*	0.0697
LVFW	19.07+0.82	22.53+6.10	55.3391	−1.2570
			0.0009***	0.2442 n. s.
m-CK (I. U./mg protein)				
RVFW	0.24+0.03	0.37+0.12	49.0000	−1.3703
			0.0012**	0.2078 n. s.
LVFW	0.31+0.05	0.31+0.10	4.0000	0.0000
			0.1040 n. s.	1.0000 n. s.
CK-MM (I. U./mg protein)				
RVFW	15.60+1.47	19.14+3.44	5.4762	−2.1160
			0.0624	0.0672
LVFW	17.05+1.08	20.03+5.19	23.0934	−1.2570
			0.0050**	0.2442 n. s.
CK-(BB+MB) (I. U./mg protein)				
RVFW	1.01+0.18	1.33+0.67	13.8549	−1.0314
			0.0130*	0.3325 n. s.
LVFW	1.71+0.37	2.20+0.90	5.9167	−1.1260
			0.0567	0.2928 n. s.

The values are given as means±SD. TS, tail suspension rats.
RVFW : right ventricular free wall, LVFW : left ventricular free wall, SDH : succinate dehydrogenase, MDH : malate dehydrogenase, LDH : lactate dehydrogenase, PFK : phosphofructokiase, CK : creatine kinase, m-CK : mitochondrial-CK.
* Significantly different from corresponding control value at $p<0.05$
** Significantly different from corresponding control value at $p<0.01$

5. データの対応の有無によって検定手法が異なる．(検定手法選択の誤り)

t 検定は，適用するデータの種類によって二種類を使い分ける．

(a) 対応のない計量値の場合：2 標本 t 検定，Student's t test (Unpaired)

(b) 対応のある計量値の場合：1 標本 t 検定，Student's t test (Paired)

対応の有無による t 検定の誤りの多くは，対応のあるデータに対して，1 標本検定を使うべきところを，対応のない 2 標本検定を使った誤りである．即ち，データの対応を無視して検定を行う人が少なくないということである．

この種の誤用が発生するのは，対応のないデータと，対応のあるデータの違い(区別)について理解ができていないからである．データに対応がない場合とある場合では，検定法を数式で表現すると違った式になる．対応がない場合の検定法を 2 標本 t 検定，対応がある場合を 1 標本 t 検定として両者は区別されている．同じデータを使っても，もちろん両者の計算結果は異なってくる．特に対応のあるデータに対して対応のない場合の検定法を使って計算すると，正しい計算では有意差があるものが，有意差がなくなるといった事態が起こりうる．(図 3)

対応の有無を一口で言うと，対応があるというのは，二つのデータに共通する条件が働いているということである．患者の治療前のデータと治療後のデータを比較して，治療効果を検討する場合などが典型的だ．同一人の右手と左手の握力，大学入試の成績と大学在学中の成績，いずれもデータ間に共通する条件が認められる対応のあるデータであり，データ数は必ず一致する．したがって同一症例から得られた複数のデータは，必ず対応のあるデータになる．対応のあるデータは，2 つ 1 組で 1 つのデータと数

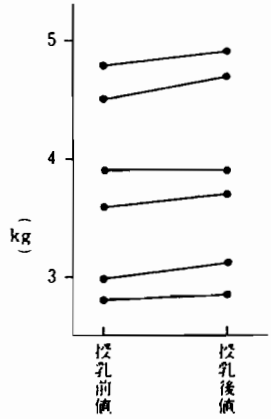
えるから，1サンプルデータと表現される．もしどちらか一方でも欠測の場合は，1症例の完成データとは見做されない．

ただし，マッチド・ペアと称するデータは同一人から得られたデータではないが，特殊な対応のあるデータの一種と解釈する．これは条件のよく似た症例の組合わせを作り，ペアに組む方法，年齢，性別，体重，重症度など，検討しようとする治療効果に影響を与えそうな因子のバランスを考えてペアを組む．このペアを比較する2群のそれぞれに振り分ける．この前作業により，比較する2群の患者構成は処理法(治療法，投与薬剤など)以外の条件はほどよく統一され，能率よく配分できることを期待した手法である．マッチド・ペアの特種型として，皮膚科でよく見られるパッチテストなどのように，皮膚の複数箇所に異なる処置をする技法もマッチド・ペアに分類される．

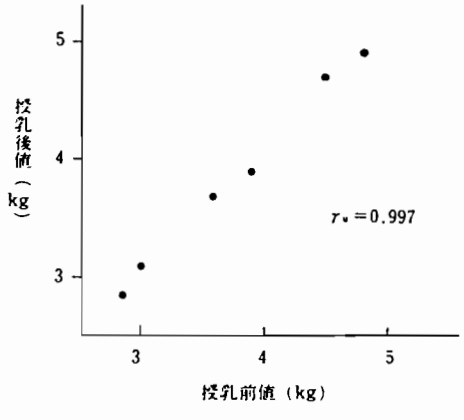
これに対して，ランダムに割り付けられた二つのグループの一方にA薬，他方にB薬を投与して二薬剤の治療効果を比較するケースなどが対応のない例になる．A群のaさんがB群の特定の患者のデータに影響を与えることも，逆に影響を受けることも考えられない．2群は独立状態にある．A群の特定患者とB群の全患者との関係はすべて等距離にあると考えるのが妥当である．こうした条件の2組のデータには対応はないとする．したがってA，B2群のサンプル数は任意でかまわない．対応のないデータをグラフ化するとき，図1のような表現は不可能で図2の形にならざるをえない．

一般に対応があるデータに対して，対応を無視して対応のない2標本検定を用いて検定すると，検出力が低下する場合がある(有意差がでにくくなることがある)．これは対応という貴重な情報を放棄したことによる．また，対応があるデータの場合，どのデータとどのデータに対応があるかが特に重要で，組み合らせが狂ったりすると正しい結論は期待できなくなり，結論も変わることがある．図1は，乳児の体重を授乳の前後で比較した例である．授乳後の体重は，授乳前の体重と授乳量で大まかに決まることはすぐに想像がつく．あえて6人の乳児のデータを使い，授乳により乳児の体重が増えるか検定を行ってみる．

	授乳前値 kg	授乳後値 kg
	2.80	2.86
	3.00	3.12
	3.60	3.71
	3.90	3.90
	4.50	4.70
	4.80	4.93
平均値	3.767	3.870　n=6
標準偏差	0.797	0.827



A



B

図1．　対応のあるグラフ

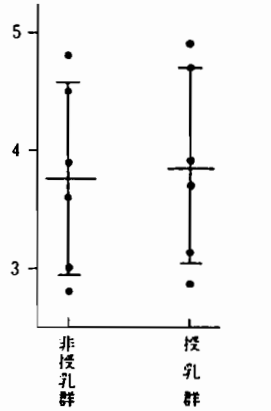
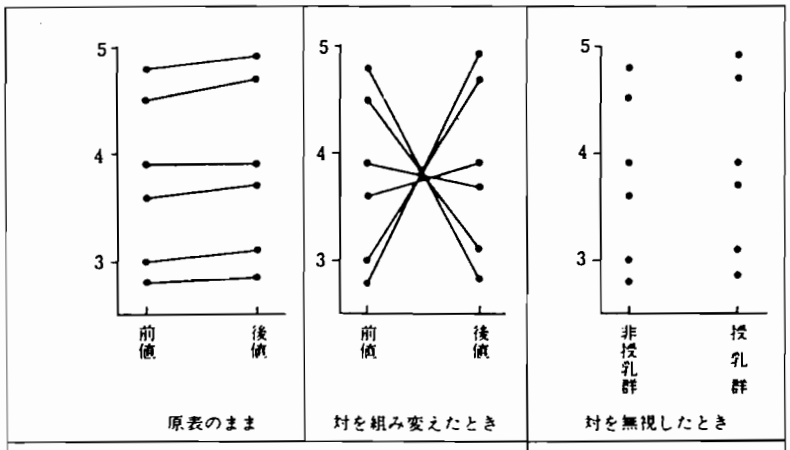


図2．　対応のないグラフ



	t 検定 I		t 検定 II
t_o	3.737	0.156	0.220
p_o	0.013^*	0.882^{NS}	0.830^{NS}
n	6	6	12

図3． 対応のあるデータの対を組み変え，無視した検定

同じ乳児の体重を2回測定したデータであるから当然対応のある検定を行う．データの対応を踏まえて Paired t test を行ったところ P＝0.013＊がえられた．「授乳により体重は有意に増加した」という結論である．(図3)

次に同じデータの対応をわざと逆順位に組み換えて同様に Paired t test を行ってみたところ，P＝0.882 が得られた．今度の結論は「授乳により体重が増加するとはいえない」である．(図3中)

授乳前後の平均値±標準偏差は全く同値であるが，2回目の検定では，データの対応の組み換えを行った点が1回目と異なる．対応のある t 検定ではペアの組み合わせの意義が大切なことを認識しておきたい．

今度は，同じデータで対応を無視した場合を見てみよう．この種の誤りは多くの医学雑誌で見られ，特に珍しいことではないが，誤用の程度はかなり重症に属する．6人の乳児の授乳前後のデータの対応を無視して対応のない検定(Unpaired t test)を行ってみると，P＝0.830 と大きくなり有意差はなくなった．(図3右)

これも正しい検定法を選択したときと異なる結論になってしまう．検定法の選択を誤ったため危険率が違ったのは当然として，もうひとつ，見落とせない相違点がある．正しい症例数は6例であるのに，対応を無視したために12例と2倍になっている．

誤用例 5－1

学生の臨床実習前後2回に精神障害に対する態度の意識調査を行い比較したレポートである．データは5段階の評点法で32項目をアンケート方式で収集，比較は実習前群と実習後群の平均値の t 検定を行っている．実習前後のデータであるから当然，対応のある Paired t test で処理さるべきところだが，対応は無視され Unpaired t test で処理されている．さらに，データ数は前 n＝83，後 n＝41 と約半数に減っている．このデータ処理には，同一人の2回の測定データを，二例分のデータとして分割し，両群に組み込まれた例と，1回だけ測定してどちらか一方の群だけに組み込まれた症例が混在している．即ち，このレポートは，症例によりデータ登録数が一定しない解析不適確の資料である．

看護学生の実習前後の精神障害に対する態度

態度項目	実習前 (N=83) 平均得点	(SD)	実習後 (N=41) 平均得点	(SD)	t検定
精神障害の社会生活の自立性					
妄想・幻想のある人でも入院しないで社会生活ができる人も多い	3.14	(0.95)	3.37	(1.01)	
一時的に保護治療するところがあれば通院で生活できる	3.96	(0.77)	3.78	(0.84)	
精神障害者の患者の独居，伸間同士の生活は危険	2.92	(1.06)	3.27	(1.06)	
精神障害者は患者同士の会をつくることはできない	4.40	(0.82)	4.53	(0.66)	
服薬や心身のバランスなどの自己管理はほとんど望めない	3.83	(1.04)	4.00	(0.73)	
精神障害者は福祉工場のようなところでも働けない	4.40	(0.84)	4.56	(0.59)	
精神障害についての性質および原因等					
精神障害は他の病気と同様病気の一種	3.46	(1.19)	3.85	(1.09)	
精神障害者が異常行為をとるのはごく一時期だけである	3.32	(0.99)	3.34	(0.90)	
精神障害者の行動は全く理解できない	3.65	(0.94)	4.02	(0.68)	*
精神障害者は何をするのかわからないので恐ろしい	3.07	(0.88)	4.59	(0.94)	**

誤用例　5－2

第二次性徴開始前後の肥満児の成長ホルモン分泌能の調査の論文.

解析は第二次性徴開始前と後の変化の比較が目的だから，データは当然対応のあるデータ，尺度分類は全て間隔尺度であるから，Pairde t test が適切な手法だが，実際に使われたのは，Unpairde t test である．症例数も前後ですべて不揃い.

入院時のデータ

	第二次性徴		
	開始前(男/女)	開始後(男/女)	p value
暦年齢(歳)	7.1±2.5 (7/9)	13.4±2.7 (9/11)	<0.01
肥満度(%)	69.0±14.7(7/9)	62.5±21.6(9/11)	NS
身長 SD スコア(SD)	1.26±1.30(7/9)	0.11±0.98(9/11)	<0.01
GRF 負荷試験時の GH 値(g/ml)	16.8±11.0(3/7)	15.5±7.5 (7/8)	NS
夜間睡眠中 GH の平均値(ng/ml)	5.9±1.0 (7/8)	6.7±1.9 (7/8)	NS
OGTT 時のインスリン総量(mIU/ml)	297±196 (7/9)	692±490 (9/11)	<0.01
血漿 IGF-I(IU/ml)	1.47±0.60(7/9)	1.97±1.59(9/11)	NS

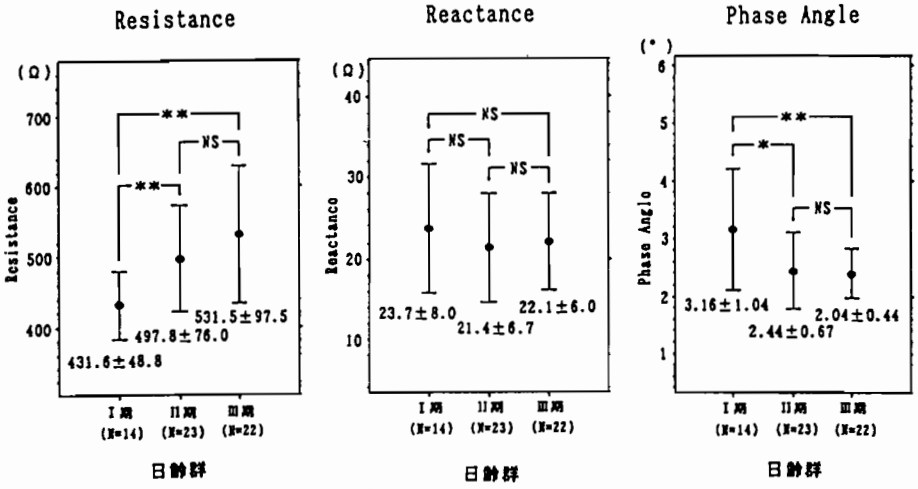
(mean±SD)

誤用例　5－3

低出生体重児23例を日齢によりⅠ群(0～1日)，Ⅱ群(2～4)，Ⅲ群(5～7)の3時点で経時的に身体計測を行い，群間比較を行った論文．使っている検定法は対応のないｔ検定．時系列データであるから当然対応のあるデータである.

この論文の問題点として　1) 検定法の選択の誤り．2) 症例数の不揃い．3) 多群の平均値の比較法の誤り．の3点に誤りがある．多群比較についての問題は次号で取り上げている.

統計学的解析は two sample t test, Mann-Whitney 検定，χ^2 検定と単相関係を用いた．有意水準は５％以下とした．数値は，平均±標準差で示した．



早期新生児期の日齢での変化

Ⅰ期：日齢０～日齢１，Ⅱ期：日齢２～日齢４，Ⅲ期：日齢５～日齢７，NS：Not Significant，
*：$p<0.05$, **：$p<0.01$

誤用例　５－４

パソコンソフトの統計ソフトマニュルの誤用例である．例題を見ると手術前，中，後の経時的測定データであるから，対応のある解析が必要なことにすぐ気づく．当マニュアルに記載されている処理法は，対応を無視した，単なる例数不揃いの二元配置分散分析である．この方法は対応のないデータに用いる手法で，正しくは例数不揃いの一元配置・経時的分散分析が適用できる．本マニュアルの解説をみると，対応に対する配慮が全くなされていないことが解る．もしかすると誤って，統計手法とマッチしない例題を取り上げたミスかも知れない．

例題10-2(交互作用のある場合)

腹腔鏡下手術をうけた患者５名を対象に，気腹前，気腹中，気腹後の呼気中二酸化炭素濃度(ETCO2)の測定と，術後に検査された肺塞栓症の有無が吸気中二酸化炭素濃度に影響を与えるかどうかを検討しなさい．

※腹腔鏡下手術：二酸化炭素ガスを腹腔内に充満(気腹)して腹腔鏡を用いて行う手術法

※肺塞栓症：腹腔鏡下手術時の合併症で，肺に二酸化炭素の気泡が詰った状態．

＜準備するデータ＞

肺塞栓	気　腹	ETCO2
あり	前	21.0
あり	中	37.0
あり	中	36.0
あり	後	33.0
なし	前	26.0
なし	中	27.1
なし	中	29.6
なし	中	33.0
なし	後	27.5
なし	前	19.5
なし	中	28.7
なし	中	26.7

肺塞栓	気　腹	ETCO2
なし	後	26.8
なし	前	25
なし	中	27
なし	中	29
なし	中	28
なし	後	25
あり	前	25
あり	中	34.7
あり	中	35.6
あり	中	34.3
あり	中	32.6
あり	後	40.2

＜検定方法＞

① この例題のように2要因で分類される群に複数の測定値がある場合には，繰り返しのある二元配置分散分析法を用いて2要因の影響と交互作用について検定します．(Stat View 4.0 を用いた検定手順については第9章参照)

(結果)

分散分析表：ETCO 2

	自由度	平方和	平均平方	F 値	p 値
Embolism	1	131.747	131.747	23.082	.0001
Condition	2	275.941	137.971	24.173	＜.0001
Embolism * Condition	2	71.713	35.857	6.282	.0085
誤　差	18	102.739	5.708		

基本統計量：ETCO 2

効果：Embolism * Condition

	例　数	平均値	標準偏差	標準誤差
Yes, Pre	2	23.000	2.828	2.000
Yes, Suff	6	35.033	1.529	.624
Yes, Post	2	36.600	5.091	3.600
No, Pre	3	23.500	3.500	2.021
No, Suff	8	28.637	2.046	.723
No, Post	3	26.433	1.290	.745

→肺塞栓の有無により呼気中二酸化炭素濃度に有意な差が認められる

→気腹状態により呼気中二酸化濃度に有意な差が認められる

→肺塞栓の有無と気腹状態に交互作用が認められる

誤用例　5－5

5－1と同じ論文であるがこの例は，平均値(計量値)の検定でなく，構成比(計数値)の検定の場合である．データは順序尺度であるから Wilcoxon 1標本検定，Stuart-Maxwel 検定など適用できるが，対応を無視してカイ二乗検定が使われている．この著者は自分が処理しているデータが対応のあるデータであることを認識していないことが読み取れる．論文に書かれている分割表は3群の2時点測定で，3×2の表となっているため，対応のある解析は不能．対応を生かした表にするためには，つぎのような，3×3の表にまとめる事で表現出来る．

この表 X_{11}～X_{33} に適当な数字を当てはめると，30数種類の組み合わせが可能で，P値もそれなりに変化する．論文では有意に変化したと記載されているが，対応を生かした検定結果を見ないことには，結論の書きようがない．

対応のある分割表(3×3)

		実習後 勤務したい	どちらでも	したくない	計
実習前	勤務したい	X_{11}	X_{12}	X_{13}	5
	どちらでも	X_{21}	X_{22}	X_{23}	17
	したくない	X_{31}	X_{33}	X_{33}	19
	計	10	24	7	41

実習後の精神科領域への就業意識

	実習前 (N=41)	実習後 (N=41)	
勤務したい	5	10	$x^2=8.40$
どちらでもよい	17	24	p=0.01
勤務したくない	19	7	

実習後，看護学生の精神科領域への就業意識は有意の増加を認めた．学生の就業意識は，実習での体験の影響を受けることがわかった．

誤用例　5－6

この誤用例もｔ検定の誤りではないが，対応の有無に対する重篤な誤用例として取り上げた．皮膚科，尋常性乾癬に対するビタミンＤ３軟膏の４濃度の用量設定の予備臨床試験．比較デザインは一被験者毎に，同程度の皮疹４カ所に，４濃度の軟膏を塗布して実施した有用度の比較検定である．評価は，極めて有用，有用，やや有用，どちらともいえない，有用でない，の５段階で判定された順序尺度データである．解析は，症例数を４倍増し(１症例当たりの４データを４症例分に分割)にして，H検定を用いて比較している．このデータは代表的な対応データである．H検定は独立多群の(対応のない)データに適用する代表的順序検定法である．従って，この試験デザインでは，絶対に使うことのない検定法である．この解析ではデータの対応は完全に無視されている．なお４軟膏の個々の多重比較(次号で紹介の予定)に使われている Tukey の多重比較部分の誤用の程度は重篤である．

有　用　度

薬剤群	極めて有用	有用	やや有用	どちらともいえない	有用でない	合計	有用率* (%)	Kruskal-Wallis 検定	Tukey の多重比較検定
基剤	2 (4.9)	8 (19.5)	14 (34.1)	13 (31.7)	4 (9.8)	41	24.4		
25 μg/g	10 (24.4)	15 (36.6)	11 (56.8)	5 (12.2)	0	41	61.0	p<0.01	p<0.01
50 μg/g	16 (39.0)	17 (41.5)	5 (12.2)	3 (7.3)	0	41	80.5		p<0.01
100 μg/g	18 (43.9)	17 (41.5)	4 (9.8)	1 (2.4)	1 (2.4)	41	85.4		

* 有用率：「有用」以上/解析症例数，( )：%

今回の試験では基剤，25 μg/g，50 μg/g および 100 μg/g の４つの試験薬剤を同一患者の皮疹を４カ所に分けた部位へ塗布し，同時比較する方法で試験を行った．

誤用例　5－7

誤用例５―４と同じパソコンの統計ソフトマニュアルの例題である．ｔ検定ではないが，データの対応の無視例である．患者30例に薬物を投与し，１，２，４，８週間後の４時点で連続評価．効果を１：悪化，２：不変，３：やや有効，４：有効，５：著効，の５段階で判定し，“治療期間によって有効率が上昇するが”を検定するのを目的とした例題．本マニュアルの検定方法によると，測定回数は１患者当たり４回であることから述べデータ数30×４＝120を症例数として，投与期間と効果判定結果の関連を相関関係(スピアマン順位相関)を使って判断資料としている．この判断法でも可能かも知れないが，対応データのメリットを放棄しているため著しく効率の悪い方法となっている．

この解析の第一の問題点は１症例当たりの４回の対応のある経時的データが分割されて４症例分のデータに化けてしまい，４回の測定値の連なりの関係が消失してしまっていることである．せっかく４回の連続測

定により入手した貴重な情報(縦断データ)を分割し，対応のない情報(横断データ)に惜しげもなく価値(質)を低下させている点が残念である．こうした大胆で無造作な処理法は“松阪牛の特上肉をミンチにしてしゃぶしゃぶで食べる”ようなやり方である．ミンチにするのがもったいないのは，図1の資料を，図2の解析法で処理していることを指していることは言うまでもない．そもそも，こうした事態に陥るのは，このデータの入力法に考慮が足りないことも原因のひとつである．データ入力時に対応を生かして操作しておくことで避けることも不可能ではない誤用である．この例題のようなデータの入力形式を採ると，データは入力の時点で対応の情報は失なわれてしまう．松阪肉は既にこの時点でミンチに変身する．即ち，このデータ形式(構造)では，症例毎の各時点での評価の連なり(推移)は断ち切られてしまう．つぎの対応を生かしたデータの入力パターンを示しておく．

患者番号	1週間後	2週間後	4週間後	8週間後	
1	1	→ 2	→ 5	→ 3	
2	5	→ 3	→ 5	→ 5	1：悪化
3	5	→ 3	→ 4	→ 4	2：不変
4	4	→ 4	→ 3	→ 1	3：やや有効
5	1	→ 4	→ 4	→ 2	4：有効
6	2	→ 1	•	•	5：著効
7	2	•	•	•	
•	•	•			

このテキストの例題を使ってデータの対応の有無を無視した場合，活かした場合，解析結果にいかに影響するかを説明する．問題を簡単にするために，1週間後と2週間後の効果判定の推移について対応のない検定には Wilcoxon 2 sample test．対応のある検定には Wilcoxon 1 sample test (符号つき順位和検定)を使って違いを比べてみた．

次に示す3種類の検定資料には対応のない場合，例題3－3のデータを用いた．有効，著効などの度数はすべて3種類に共通であるが，検定結果の統計量はそれぞれ異なる．

① 対応のない場合

	悪　化	不　変	やや有効	有　効	著　効	順位和検定
第1週間後	3	13	6	5	3	
第2週間後	2	7	10	7	4	P＝0.1462 ns

対応のない場合の検定結果は有意ではない．

② 対応のある場合

このデータベースにはすでに対応の情報が失われているので筆者が適当に対応を組んで検定を試みた．この組み合わせ方次第で，(i) 有意差のある場合，(ii) ない場合，と結果は不規則に変化する．その組み合わせ(対応)の変化によって，統計量(P値)が変化することを観察できる．

(i)

		2週間後						
	判定	1	2	3	4	5	計	符号つき順位和検定
1週間後	1	0	3	0	0	0	3	
	2	0	4	8	1	0	13	
	3	1	0	2	3	2	6	Z＝2.5364
	4	0	0	0	3	2	5	P＝0.0112*
	5	1	0	0	0	2	3	
	計	2	7	10	7	4	30	

対応のある検定結果で有意差あり．

評価：2週間後の判定は1週間後の判定と比較し有意上昇している．

(ii)

	判定	2週間後 1	2	3	4	5	計	符号つき順位和検定
	1	1	1	0	1	0	3	
	2	1	2	5	3	2	13	
1週間後	3	0	3	3	0	0	6	Z＝1.5272
	4	0	1	1	3	0	5	P＝0.1267 ns
	5	0	0	1	0	2	3	
	計	2	7	10	7	4	30	

対応のある場合でも対応の違いによって有意差なし.

評価：2週間後の判定は1週間後の判定と比較し上昇しているとはいえない.

対応のない場合の検定①では有意差なしであるが,対応がある場合は組み合わせによって有意になったり,ならなかったり,検定の結果は一定でない.②の1，2週間後の計の欄の数値は(i)と(ii)は同じである.異なるのは,対応の組み合わせだけである.即ち,対応のある場合,組み合わせ次第で結果はどう変化するのかわからない.対応データでも平均値,頻度が重要なことは当然として,更に推移(対の組み合わせ)の情報が大切であることを認識しておきたい.対応のある縦断データは情報量が豊富であるから活用法次第で強力な解析手段になりうる.この情報は無駄にしたくないものだ.松阪牛はミンチにしないで食したい.

例題3-3（2つの要因の間に相関関係が認められるかどうか）

ある治験薬の効果を判定するため,薬物の投与を開始してから1週間後,2週間後,4週間後,8週間後に効果を判定し,次のように集計された.治療期間によって者効率が上昇するかどうかを検定しなさい.

	悪　化	不　変	やや有効	有　効	著　効
1週間後	3	13	6	5	3
2週間後	2	7	10	7	4
4週間後	2	5	6	10	7
8週間後	2	3	7	9	10

＜準備するデータ形式＞

★列挙データ形式

投与期間・効果判定が順序を示しているので,次の様に整数を割り当てる.

投与期間：1週間を「1」,2週間を「2」,4週間を「3」,8週間を「4」

効果判定：悪化を「1」,不変を「2」,やや有効を「3」,有効を「4」,著効を「5」

期間	効果
1	3
1	5
1	5
1	4
1	1
1	2
1	2
1	4

期間	効果
1	4
1	3
1	2
1	1
1	3
1	2
1	1
1	2

期間	効果
2	4
2	2
2	3
2	4
2	3
2	5
2	2
2	3

期間	効果
3	5
3	5
3	4
3	3
3	4
3	5
3	3
3	5

＜検定方法＞

順序関係のある2要因について分割表の形で集計されたデータは，次のような散布図に置き換えて考えることができます．2要因とも飛び飛びの値(離散変数)なので検定にはノンパラメトリック検定を用いることになります．このようなデータについて相関関係が認められるかどうかを検定したいときは，**スピアマン順位相関係数**(Spearmanrnakcorrelation)の検定を行います．