〔日本生理誌（1996）**58,** 175-182〕

**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Ⅱ—

栗谷典量
（久留米大学医学部小児科）

### t検定の問題点

平均値の差の比較検定にはt検定(Student's t test)が多用される．ところが医学論文で使用されているt検定には用法の誤りが多いとされている．

t検定の誤用率が高いのは，日本の医学界特有の現象ではなく，以前，アメリカでも誤用例は少なくなかったという．かってt検定の誤用の問題がアメリカの医学会で取り上げられ指摘されたことがある．即ち，1980年 American Heart Association の学会誌 Circulation に CURRENT TOPICS として，統計の誤用に関する論文が掲載されたことがある．即ち，医学誌に掲載された論文のほぼ半分に統計手法の誤りがあると断じ，この論文中にt検定の誤用について言及した文がある．(Circulation, Vol. 61, No. 1, p 1-7, 1980)

日本の医学誌に見られるt検定の誤用の内容を分析してみると，次の6つのパターンに分類できるようだ．以下，順をおって誤用例を取り上げて検討の参考に供したい．

①症例数に対するデータ数がアンバランスである．

症例(患者)によって提供データ数に差がある．たとえば患者数10，データ数14のように，延べ患者数をもって，登録症例数14として処理(t検定)された論文がこれに該当する．(前回のテーマとして取り上げた問題)

②名義尺度・順序尺度(計数値)に適用する．

t検定は，本来，計量値(間隔尺度・比率尺度)に対し適用する検定手法である．これを症状の程度や改善度などのデータに対し0点・1点・2点・・・などの数値を振り当て，この数値をスコアーと称し計量値とみなしてt検定を適用する誤り(習慣)が少なくない．

③データ分布が正規分布から著しくかけ離れたデータに対して適用する．

t検定は，データの平均値の比較を目的とする検定法であり，本来，正規分布を前提として構成されている．従って正規分布から著しく偏ったデータ分布は，そのままではt検定は使えない．何らかの対策を考える必要がある．

データ分布の正規性の検討に際し注意を要する問題点の一つに，多くの臨床検査値がある．

臨床検査値には分布が高値側に歪んだ(歪度＋)ものが多いことに留意したい．

④比較する二群のデータの分散の差が大きい．

分散に差がないと判定するには「二つの分散の差のF検定」により有意差がなければ一応両者に差はないと判断してt検定が使える．もし差が有意であったときは次の近似法により処理する．

1．Welch (Aspin-Welch)の法
2．Cochran-Cox の法

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，自由度の計算に端数がつくためP値の算出が必要である．(t分布表には端数自由度の記載がない)

⑤データの対応の有無を無視する．

t検定には二種類ある

対応のない計量値(Student's t test(Ⅰ), Unpaired t test)

対応のある計量値(Student's t test(Ⅱ), Paired t test)

の二つで対応の有無により使い分ける.

⑥多群(3群以上)の平均値の比較に当たり，多群中の2群の組み合わせ毎に，各々をt検定で比較する.

以上のような不適当なt検定の使用例があり，これらの誤用例を6項目に整理分類した.

①データ収集法のあやまり.(重複登録など)

②計数値（順序尺，名義尺度）に適用する.

③④⑤をまとめると，対応のないt検定の適用には次の条件を満たす必要がある.

二つの母集団はそれぞれに正規分布を前提とし,

二つの分散に差がない.

二つのデータは互いに独立した母集団からの無作為標本である,

対応のあるt検定では分散の検定は必要ない.

⑥多群間の平均値の比較を2群の組み合わせ（k群では ${}_kC_2=k(k-1)/2$）毎にt検定の繰り返しで比較しているのが多いが，この方法は誤りである.

①については前回，症例数とデータ数のアンバランスの問題としで取り上げた．今回は②③について述べる.

**2．記号の平均値を求める錯覚**

平均値の計算は極めて簡単だが，データ群を代表する値として優れた統計量である．しかしデータの種類によって使えるものと，使えない平均値とがあることを知っておく必要がある．そのためには先ず，データの種類による分類法を理解しておかねばならない.

分類は先にⅠ計数値(How many に対する答え)と，Ⅱ計量値(How much に対する答え)に分類し，次に計数値を名義尺度(nominal scale)と順序尺度(ordinal scale)の二つに分類する．計量値は間隔尺度(interval scale)と比率尺度(ratio scale)の二つに分類する．この分け方をstevens の分類尺度といい統計手法の採択はこの分類法に基づいて選ばれる,医学データでは，間隔尺度と比率尺度は特に分離せず計量値として処理でき,適用できる統計手法にも差はない.

| Stevens の尺度分類 | | | 適用できる統計手法 |
|---|---|---|---|
| Ⅰ 計数値 (digital) | 名義尺度 | 性別,疾患名,菌種の種類 | カイ二乗検定 |
| | 順序尺度 | 症状の程度，効果の程度 | 順位和検定，順位相関 |
| Ⅱ 計量値 (analog) | 間隔尺度<br>比率尺度 | 気温，体温，<br>体重，身長，血圧，時間 | 平均値，標準偏差<br>t検定，相関と回帰 |

間隔尺度・比率尺度では適用する統計手法の区別はない.

データの性質から考えて，名義尺度に対し番号を振り当て，その番号の平均値を求めても意味がないことは,直感的にも理解できる．ここで最も注意したいのは順序尺度に対して平均値を求める誤りである．臨床症状の程度(症状なし，軽い症状，重症，・・・)や，治療効果の程度(治癒，著効，軽度改善，・・・)を表現するのに,患者の状態をスコアーと称するデータに置き換えて統計処理にかかるとき,あたかも計量値であるかのような錯覚に陥り，平均値，SD，t検定，ピアソンの相関係数と何のためらいもなく誤用に走りだす．平均値に意味がある尺度は計量値だけで，名義尺度，順序尺度，には使えないことの認識は失われている．特に甚だしい誤用例をひとつ紹介しておく．即ち，名義尺度の数量化と称し，ねむけ1，下痢または嘔吐2，めまい2，GOT上昇の程度により3～5点など，副作用の種類の表を作りこれに適宜点数を振り当て，その合計点を副作用スコアーとし，t検定を行うという比較原理である.

日頃遭遇するこの種の誤用例のパターンとして

| 症状の程度 | ：スコアー |
|---|---|
| 症状 なし | ： 0点, |
| 軽い 症状 | ： 1点 |
| 重い 症状 | ： 2点 |
| 死 亡 | ： 3点 |

があげられる．この場合スコアーを，0点，1点，2点，,, と点数としないで

－，±，＋ 或いは A，B，C などの記号で表現してもいっこうに差し支えないのである．もしこの場合，記号を使用していたら記号の平均値±標準偏差などの算出の発想は思い付

かないだろう．それを１，２，３点としたために一見，計量値らしくなる．もし計量値であれば０，１，２，３のそれぞれの数値間の間隔はすべて等間隔でなければならず，症状なしと軽い症状の差１点と，重い症状と死亡の差１点とは同等でなければならないということになる．また，軽い症状３人は死亡１人に相当することになる．平均値±標準偏差，相関係数，ｔ検定が適用できるのは各データの等間隔性が保証されてはじめて有効で，その保証がない限り平均値には意味がない．当然ｔ検定も使えない．

上記のスコアーの評価基準は，以前，薬効評価レポートの副作用の解析で使用された実例からの引用である．（原文の判定基準は，副作用なし：０点，軽度の副作用：１点，重い副作用：２点，死亡：３点となっている．）

順序尺度には原則的に順位検定法の選択が第一であることに誤りはない．ｔ検定は最適手法ではないが，医学分野では，ときに計数値を計量値扱いにして処理する習慣もある．教育・心理分野などでもみられることがある．IQ などのデータは厳密にいえば順序尺度であるがｔ検定が使用されことがあるのも事実である．然し，副作用なし：０点，死亡：３点のような処理は許されない．副作用のため１０人中１人死亡した薬剤群（平均スコアー　0.3±0.9487）と，10人中４人にねむけが発生した薬剤群（0.4±0.5164）より死亡１人の方が副作用スコアーは低いことになり，安全性判定がおかしなことにも成りかねず，「統計で嘘をつく法」のテクニックとして利用できることになる．

誤用例　２－１　この誤用例はパソコンの統計ソフトのマニュアルからの引用である．

ここで取り上げた例題は重回帰分析の例でｔ検定ではないが順序尺度の計量値扱いであることと同意義の誤用である．ここで問題としているのは，目的変数の「術後循環器合併症」の配点である．軽度の合併症のスコアが１で死亡が３となっている．また説明変数の年齢，体重の２変数は計量値で問題ないが，術前状態と虚血性心疾患の既往の２変数は順序尺度で正規型の計量値とは程遠いデータである．

**例題13**

術後に発生する循環器合併症の危険因子を調査するため、患者25人の年齢，体重，術前状態（ASA分類），虚血性疾患の既往について調査したところ次のようなデータが得られた．これらの数値から術後の循環器合併症の危険を予測する式を作成し，あてはまりが良いかどうかを検定しなさい．

＜準備するデータ＞

★列挙データ形式

| 年齢 | 体重 | ASA | 虚血性心疾患 | 循環器合併症 |
|---|---|---|---|---|
| 73 | 46.5 | 2 | 3 | 3 |
| 83 | 31.5 | 2 | 1 | 2 |
| 71 | 54.0 | 2 | 0 | 1 |
| 70 | 42.0 | 2 | 0 | 1 |
| 84 | 40.5 | 2 | 0 | 2 |
| 82 | 40.0 | 2 | 1 | 1 |
| 80 | 45.0 | 2 | 0 | 1 |
| 72 | 38.0 | 2 | 0 | 1 |
| 82 | 54.0 | 3 | 1 | 2 |
| 95 | 48.0 | 3 | 1 | 1 |
| 72 | 52.0 | 3 | 1 | 2 |
| 77 | 63.0 | 2 | 1 | 1 |
| 76 | 46.5 | 4 | 2 | 2 |

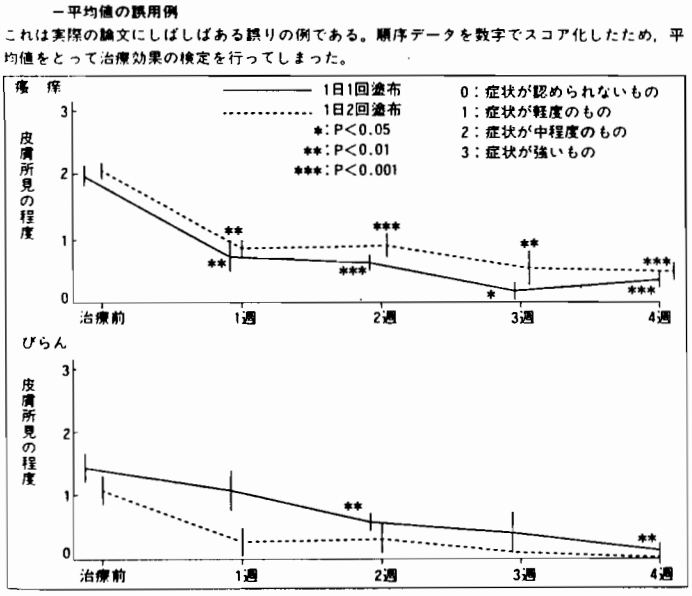
| 年齢 | 体重 | ASA | 虚血性心疾患 | 循環器合併症 |
|---|---|---|---|---|
| 73 | 44.0 | 2 | 0 | 2 |
| 79 | 47.0 | 3 | 1 | 1 |
| 77 | 64.0 | 4 | 1 | 1 |
| 84 | 70.0 | 5 | 2 | 3 |
| 89 | 47.0 | 4 | 1 | 3 |
| 79 | 49.0 | 2 | 2 | 2 |
| 82 | 54.0 | 2 | 1 | 1 |
| 74 | 55.0 | 2 | 2 | 2 |
| 76 | 45.0 | 2 | 0 | 0 |
| 83 | 35.0 | 3 | 0 | 0 |
| 72 | 38.0 | 2 | 1 | 0 |
| 88 | 41.0 | 2 | 1 | 0 |

**術前状態（ASA）**
- １：正常
- ２：軽度の全身的疾患を持つ
- ３：高度の全身的疾患も持つ
- ４：生死に関わる重篤な状態
- ５：手術の有無に関係無く死亡

**虚血性心疾患の既往**
- ０：なし
- １：陳旧性心筋梗塞及び狭心症
- ２：６カ月以内の心筋梗塞
- ３：急性心筋梗塞

**術後循環器合併症**
- ０：なし
- １：軽度の合併症
- ２：重篤な合併症
- ３：死亡

誤用例　2－2　皮膚科の論文でよく見かける症状点(順序尺度)の計量値扱いの例




### 3．データ分布が歪んでいる場合

データの分布が著しく歪んでいる場合，その平均値をデータの代表値として統計処理の資料とすることは不適当な場合がある．本来 t 検定は正規分布を前提とした手法であるから当然適用できない．歪んだ分布の実例として例をあげると，国民の平均貯蓄額などがこれに該当する．平均貯蓄額は一部の極端な高額貯蓄者が全体の平均値を高値側へ引っ張り上げた分布(歪度＋)となり，平均値は中央値より必ず高値になる．平均値と中央値に差があると言うことは，分布に歪みがあり，平均値を中心とした左右対象形の正規分布からの外れを示している．この平均値は代表値としてあまり意味をなさない．このようにデータ分布に歪みがある場合，対処手段として，つぎの処置が考えられる．

1．データの分布が正規分布に従うことを前提としないノンパラメトリック検定を使用する．即ち2群比較なら Mann-Whitney's U test，多群なら Kruskal-Wallis H test などを適用する．

2．歪んだ分布に適切な変換処理を加えて，歪みを小さくした(正規分布に近づけた)後，検定を行う．変換で最も代表的な手法は対数変換である．臨床検査値に多い分布が右に尾を引く形の歪み(歪度＋)の縮小に有効であるからだ．この変換で歪んだ生データの分布を正規分布に近づけることができる．但し，データにゼロや負数が含まれる場合，対数変換は使えない．歪度が(＋)の分布では，対数変換のほかに平方根や立方根なども使われる．歪度が(－)の時は，2乗，3乗などの方法を適宜採択する．

対数変換が効果的なデータ群として，白血球，GOT，GPT などの多くの臨床検査値があげられるが，酵素系の関与のあるデータなどに対しては，先ず対数変換処理を考えるのが妥当だろう．

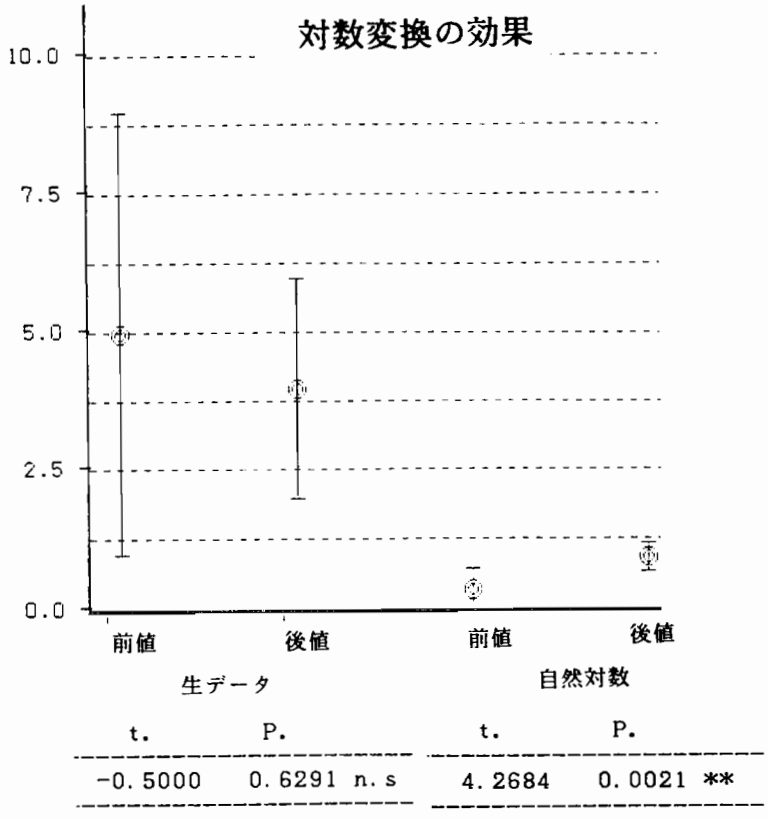
白血病患者の白血球数を題材に，対数変換処理の効能を説明する．患者数10のうち9例の治療前の白血球数が1万，残る1例が41万とする．治療後1万だった9例は2万に上昇し，41万だった1例は22万に減少した．この平均値±SD の推移を単位を万で表記し，対応のある t 検定を，生データと対数変換値の二通りで行ない結果を比較すると．

| | 治療前 | | 治療後 | 前後差 | P値 |
|---|---|---|---|---|---|
| 生データ | 5.0000±12.6491 | → | 4.0000±6.3246 | −1.0000±2.0000 | 0.6291 ns |
| 対数変換値 | 0.3714± 1.1743 | → | 0.9329±0.7583 | +0.5616±0.1316 | 0.0021 ** |

対数は自然対数値　　df＝9 検定は Paired t test

生データでは平均値の推移は1万の低下であるが有意差なし，一方対数変換値での平均値は＋0.56の上昇で，差は有意(P＝0.0021**)である．つまり，治療前後の変化量を比較すると，上昇は9/10に対し，低下は1/10である．両者の結果を比較すると，対数変換値の方が実情にマッチしていることが理解できる．(図1)

図 1




**正規性の向上の確認**

歪んだデータの分布に変換を加え，正規性の向上をチェックするポイントを述べる．

変換前後で

1. 歪度が小さくなったか．(ゼロに近づいた)
2. 尖度が3に近づいたか．
3. 平均値と中央値のズレ幅(差)が縮小したか．
4. ヒストグラムの左右のバランスは改善したか．
5. 累積頻度曲線の50％値が分布の中心に寄り，
   曲線がS字型(シグモイド)に近づいたか．
6. 変動係数(CV)は改善したか．

図2．は肝疾患群のGOT，GPTのデータの対数変換処理前後のグラフである．

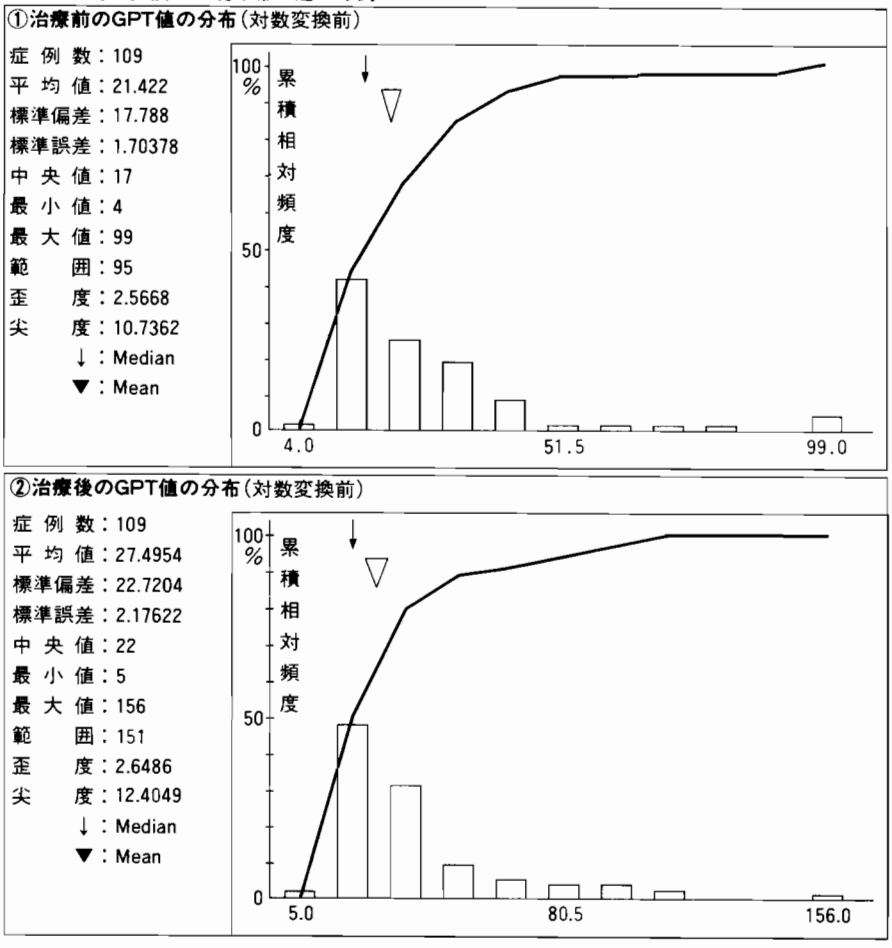
対数変換により歪度，尖度，平均値と中央値

の差，ヒストグラムのバランス，累積頻度曲線　　のが分かる．
のパターン，変動係数いずれも改善されている

| | GOT | | | GPT | | |
|---|---|---|---|---|---|---|
| | 対数変換前 | | 対数変換後 | 対数変換前 | | 対数変換後 |
| 歪　度 | 2.5668 | → | 0.3775 | 2.6486 | → | 0.3971 |
| 尖　度 | 10.7362 | → | 2.9849 | 12.4048 | → | 3.1296 |
| C　V | 83.0361 | → | 23.6540 | 82.6334 | → | 21.4359 |

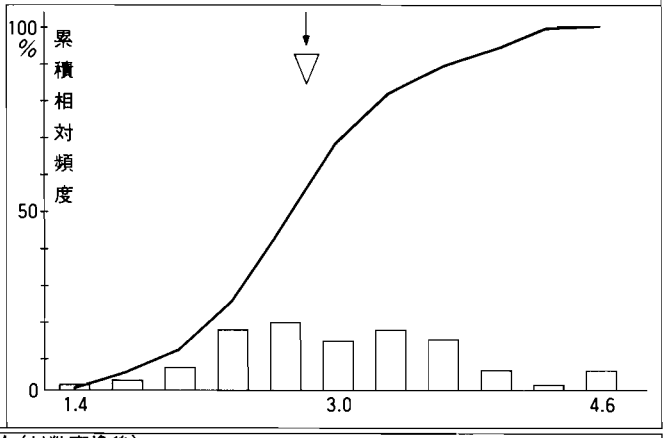
CV(変動係数 %)＝標準偏差／平均値×100

図 2　－対数変換で正規分布に近づく例



**③治療前のGPT値の分布**（対数変換後）

症 例 数：109
平 均 値：2.8255
標準偏差：0.668343
標準誤差：0.0640157
中 央 値：2.83321
最 小 値：1.38629
最 大 値：4.59512
範　　囲：3.20883
歪　　度：0.377554
尖　　度：2.98485
↓：Median
▼：Mean




**④治療後のGPT値の分布**（対数変換後）

症 例 数：109
平 均 値：3.07952
標準偏差：0.660122
標準誤差：0.0632282
中 央 値：3.09104
最 小 値：1.60944
最 大 値：5.04986
範　　囲：3.44042
歪　　度：0.397105
尖　　度：3.12962
↓：Median
▼：Mean