HOKKAIDO UNIVERSITY

Title	樹木・森林生態学「よく出る」誤用統計学の基本わざ
Author(s)	久保, 拓弥
Citation	生物科学, 54(3), 188
Issue Date	2003-03
Doc URL	http://hdl.handle.net/2115/20148
Type	article (author version)
File Information	seibutsu54-3.pdf

Instructions for use

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

[特集] 資源獲得戦略としての樹木の形づくり── 「枝葉末節」から本質へ

樹木・森林生態学「よく出る」誤用統計学の基本わざ

久保拓弥

要旨: 樹木の個体・個体群・群集の問題を取りあつかう生態学のデータ解析ではさまざまな統計学的手法が駆使される．同時にこれらの手法についてよく理解されぬまま間違って用いられている事例も散見する．ここではよく普及しているごく簡単な誤用二例を紹介する．対数変換してから直線回帰することで生じる問題，そして「割り算指標」とその分母の間で「負の相関を創作」してしまう失敗である．どちらも観測データの「確率分布を見ない」解析方針に原因がある．統計学的なデータ解析では背後にある確率論的モデルを考え，それらに合致した統計学的手法を採用しなければならない．

キーワード: 確率分布，推定，統計学

樹木・森林データ解析はどのように?

2002 年生態学会大会 (仙台) 会場，この特集の起源となった「枝葉末節」シンポジウムで樹木の形作りに関する知見がつぎつぎに発表されているころ (筆者はこちらに参加してないので，ここまでの記述は空想による)，その「裏番組」のひとつとして「生態学におけるデータマイニング」(企画者: 粕谷英一・島谷健一郎) が開催されていた．こちらは生態学的なデータ解析とその統計学的手法の工夫と応用について議論する内容であった．本稿ではその議論の一部の文脈を借りつつ「枝葉末節」データ解析でよく見られる統計学「誤用」わざを見ていくことにしよう (……というのが「枝葉末節」主催者たちの依頼なので).

統計学的な解析とは「数値 (記号) のちらばりかた」を定量的に特定しようとする，つまり確率分布の (parametric もしくは nonparametric な) 推定を主目的とする数学的手法である．しかるにこの「確率分布はどうなっているのか?」といった解析の原点があまり考慮されぬまま統計学的手法が使われている身近によく見かける．そこで筆者は「データマイニング」シンポジウムでこうした一連の現象を「統計学いまふうの使われかた」として要約を試みた (かなり戯画化した表現になってるのは強い反論を期待しているため).

- **「表計算ソフト」依存症**: どんなデータも無理矢理に二次元のテイブルにする．もっぱら目視手作業で数値を操作する．計算機を使ってるつもりが計算機に使われている．
- **「チャート式」統計学**: 「○○なら→××せよ」などなどといったあまり正しくもない「統計学の公式」をよくわからぬまま丸暗記して使っている．
- **「ゆーい差」決戦主義**: なんでもいいからとにかく「ゆーい差」さえ出ればよし，とする態度．統計学的有意差すなわち生物学的有意差であるとカン違いしている．なお同シンポジウムでは「検定すべき状況でないのに検定している例が多い (モデル選択すべき場合が多々ある)」(粕谷) といった指摘もあった．

筆者は樹木の個体・集団を研究している人たちと共同で研究を進めることが多いので，あるいはこれらはそういう分野に固有な「くせ」なのかもしれない．

ともあれ，上に挙げた項目は相互に連関している．しかしここでは紙数も限られているので議論する範囲をせまく限定したい．そこでこの分野において「統計学の公式」とやらで定着している基本「誤用」わざ，その中でももっとも単純明解な事例を二つばかり紹介してみよう．

Kubo Takuya: The basics of common “mis”application in the field of tree and forest ecology
〒 060-0810 札幌市北区北 12 西 5
北海道大学大学院地球環境研究科生態科学専攻科
E-mail: kubo@ees.hokudai.ac.jp

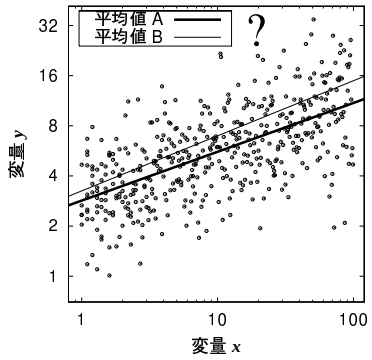

図 1　対数変換して回帰してよいのか?
タテ軸を対数変換して直線回帰を行った結果にそのまま exp つけて表示すると平均値 A のようになる．これでよいのだろうか? 図中の点々 (標本) は計算機に生成させた対数正規乱数 ($\mu(x) = 1.0 + 0.3\log(x)$, $\sigma = 0.5$) である．

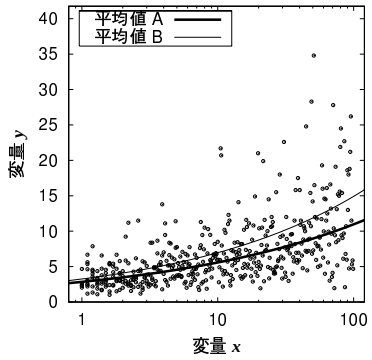

図 2　対数正規分布から得られた標本
図 1 のタテ軸を対数から常数にもどした．対数正規分布では点のちらばりに「上下非対称性」がある．点々の多いところをとる平均値 A ではなく，分散も考慮した平均値 B が「ある x に対する平均的な y」を示している．

対数タテ軸→回帰の問題

「被説明変数 (以下ではタテ軸と呼ぶ) を対数に変換して直線回帰を行う」は多用される不良「チャート式」のひとつである．これが良くない理由は，変数変換をほどこしてしまうと，もともとのタテ軸の確率分布が何か別ものになってしまうからだ．すると「タテ軸の誤差は分散の等しい正規分布」なる最小二乗法の前提が成りたたなくなる．データは等分散正規分布だと決めておきながら変数変換で自滅する問題は粕谷 (1998) 中の「回帰の悪い夢」などで例を挙げて解説されている．

次にこのような指摘に対して「いやいや，私はタテ軸の log が正規分布になってると考えている」(つまりタテ軸の数値のちらばりが対数正規分布になっていると仮定している) として正当化しようとする人がいる．これは一見もっともらしそうだけれど，自分自身で採用した対数正規分布という確率論的モデルをよく把握してないかぎり，この先に待ちかまえている「罠」にひっかかることになる (ただし，それが「間違い」につながるかどうかは推定された計算結果をどう利用するかに依存している)．

なるほど対数正規分布の確率密度関数はよく

$$\frac{1}{\sqrt{2\pi\sigma^2}y}\exp(-\frac{(\log(y)-\mu(\mathbf{x}))^2}{2\sigma^2}),$$

と表記されるから，このときに $\mu(\mathbf{x})$ を線形モデルとして定義しておけば，たしかに「タテ軸対数→直線回帰か重回帰」なる手法でパラメーターの最尤推定値が計算できる．

ところでここで注意しなければならないのは，対数から常数の世界にもどした $\exp(\hat{\mu}(\mathbf{x}))$ はタテ軸変量 y の**算術平均になっていない**ところにある (同様に $\hat{\sigma}$ も標準誤差ではない)．

具体的な例で考えてみよう．図 1 には 2 本の回帰線が描かれている (横軸も対数化されているけれど，これは今回の話とは無関係である)．平均値 A は「タテ軸対数変換→直線回帰」で推定された $\hat{\mu}(x)$ に exp をつけただけのものである．こちらが良く見かける事例であり，こういう図を描いた人は「ある x に対する平均的な y」を示しているつもりになっているのかもしれない．しかしながら，上で述べたようにこれはタテ軸 y の平均値ではない．図 2 に示しているように対数正規分布は標本のばらつきに「上下非対称性」があり，そのため「点々の一番多いところ」が分布の平均値とならない．これが「上下対称」な正規分布との違いである．

対数正規分布の平均は $\exp(\mu(\mathbf{x}) + \sigma^2/2)$ で与えられるので，タテ軸 y の平均値は図 1 の平均値 B で示されているような推定結果となる．数式と図からわかるように，平均値 B は平均値 A より常に大きい．よって「タテ軸対数変換→回帰→ $\hat{\mu}(\mathbf{x})$ に exp つけてもどす」という手法で推定した「平均値」は本当の平均値より常に過小推定になってしまう．このような回帰分析によってタテ軸 y に関する予測を導くときには平均値 B を

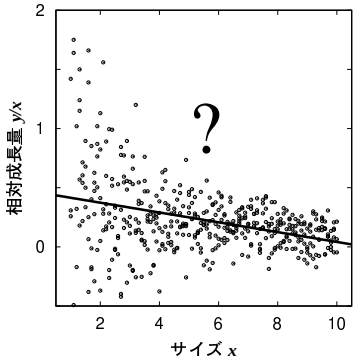

図 3 これは「負に有意にかたむいてる」?
横軸にサイズ，タテ軸に相対成長量 (成長量をサイズで割った量) をとっている．これからタテ横のあいだに「負の相関」だとかトレイドオフの傾向がある，と言うのは正しいのか?

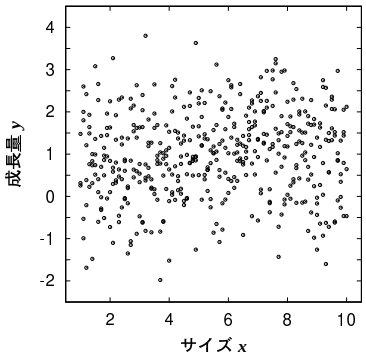

図 4 じつはタテ横無関係
図 3 のもとになった標本散布図．タテ軸は割り算値ではなくそのままで示している．図中の点々 (標本) は計算機に生成させた正規乱数 ($\mu = 1.0$, $\sigma = 1.0$)，つまり横軸とタテ軸は無相関.

使わなければならない.

図 1・2 に示した標本の算術平均値が平均値 B であるなら，平均値 A として計算された量は何なのだろうか? これは変量 y の**幾何平均値**を表している．したがって平均値 A を図示したいのであれば「何のために幾何平均を計算したのか」を明確に説明するべきだろう.

割り算指標→負のにせ相関問題

次に「もっと素朴」というしかない誤用わざをとりあげておきたい．図 3 に示しているような不適切な散布図と間違った解析をいまだに見かける．何かを横軸の変量で割り算した量がタテ軸となっている図を作っておいて，次にこれらタテ横の間に「有意な負の関係」などが検出された，という状況である.

図 3 は図 4 に示している乱数値を横軸の値 x で割り算したものにすぎない．図 4 のタテ軸「成長量」は横軸「サイズ」に対していっさい無関係であり，いかなる相関もない (ここでは「成長量」「サイズ」といった語を使っているけれど，いわゆる相対成長量だけを問題にしているのではなく，一般に x と y/x の関係についての問題を取り扱っている).

このような無相関な量を図 3 のように x で割り算してみたところで，当然ながら何か統計学的な意味あいが生じるわけではない．言い換えると，$\{1, 2, 3, \cdots\}$ という独立変数 $\mathbf{x}$ に対して c/x つまり $\{c/1, c/2, c/3, \cdots\}$ というような指標 I の集まりを解析者自身が定義しているんだから，それはただ単に I を x に反比例する関数として**作り上げた** だけであって，統計学的な「負の相関」といったたぐいのものではない．それだけである.

これに対して「横軸のサイズ x に対してタテ軸は相対成長量 y/x という生態学的に意味のある指標になっているからこのような関係を調べた」と主張する人がいる．この反論が無意味であるのは上で説明しているとおりだ．生態学的に動機づけられていようがいまいが，無相関な量の割り算値は統計学的「にせ有意」を示すにすぎない (逆に統計学的に有意差があったからといって何らかの生態学的な意味があるとは限らない).

ここでは x と y が無関係である例を使って説明した．しかし「割り算指標とその分母の間で相関・回帰分析すべからず」という原則は x と y の間になんらかの関係がある場合にも同じく適用される．データを得てまず為すべきなのは，図 4 で図示しているような観測された値そのものの間の依存性の有無を調べることであって，図 3 のごとく奇妙な指標化・規格化による「統計学的関係の創作」ではない.

どうしたらよいのか — 謝辞にかえて

ここで述べた両誤用わざは昔から知られているんだから，第三千年紀ともなればもはや誰も使ってないんじゃないの……という指摘があるかもしれない．そこで 2001 年の日本生態学会英文誌 *Ecological Research* に掲載された 80 数篇の投稿論文 (生態学全般の内容を掲載，複数の査読者が

掲載妥当性を調べる，そしてこの中には統計学的手法をまったく使用しない論文も多数含まれている) をごく大ざっぱに調べてみたところ— タテ軸対数変換して回帰直線を示しているもの 6 篇 (ただしこれらが「間違い」かどうかは推定結果の使いかたに依存する)，割り算指標とその分母となってる数量の間に「有意な関係を発見」しているもの 3 篇が見つかった．どうやらこれら「伝来のわざ」はまだまだ後継者不足に陥ってないようだ．

ここで紹介したような誤用わざを避けたいのであれば，結局は統計学の基本にたち返ってデータを解析するしかあるまい．まずは観測データがどういう確率分布にしたがうのかをよく調べる．よく使われる確率分布の種類とその性質はさまざまな教科書に詳しく記述されている (蓑谷 (1998) など)．最尤法によるパラメーター推定 (Pawitan (2001) など) などを使ってその確率分布を特定する (あるいは竹澤 (2001) などのように nonparametric 法でも分布を推定できる)．

対数だの割り算だのその他あれこれものごとをややこしくするだけの変数変換 (あるいは標本の「クラスわけ」などの技法) が一部に愛されている原因のひとつは，データを原型とどめぬほどひねくりまわしてでも何とかして最小二乗法・分散分析といった「等分散の正規分布」仮定する手法に帰着させんとする悪しき解析方針にあるようだ．「等分散ではない→対数変換せよ」といったたぐいの「チャート式」がそのまま通用するのかどうか，対象と目的に照らしながらよくよく考えなければならない．ついでながら「等分散ではない→nonparametric」も恒真の「定石」として使うことはできない (Kasuya, 2001)．

観測されたデータの分布が特定されれば，つづいてそれを仮定した統計学的手法を適用していくことになる．今回はそちらについては説明しない．「統計学的有意差検定の意味の無さ」(Johnson, 1999) なる愉快な題名の短い論文に多くのヒントが内包されている，と指摘しておく．さらに統計学計算プログラムの自作は「統計学ブラックボックス化」を避ける上で有効である．プログラミングに関連するリンクなどは計算生態学サイト (`http://hosho.ees.hokudai.ac.jp/~kubo/ce/`) にまとめている．

最後に，統計学的手法に関する研究者どうしの率直な議論こそは誤用まぬがれる「対策」としてもっとも重要，と指摘しておきたい．その良い例が本稿の内容である．これらはすべて他の人たちから教えられたものばかりであり，筆者自身で考えついたものは何も無い．タテ軸対数変換→直線回帰問題は京都大学の鈴木牧さんに教えていただいた．「対数平均は幾何平均」といった指摘は University of Florida の奥山利規さん (とその指導教官 Ben Bolker さん) にいただき，さらに院生が「ゆーい差」決戦主義におちいらないようにする同大学動物学部の取りくみと文献を教えていただいた．割り算指標→いんちき相関については「枝葉末節」シンポジウム主催者である環境研の竹中明夫さん・北大の甲山隆司さんたちの構成された実例に関する議論にもとづいている．変数変換の問題全般とさまざまな「統計学解析の公式」の弊害に関しては九州大の粕谷英一さんにご教示していただいたものばかりである．また北海道大の院生・研究員たちとの統計学議論もたいへん有用であった．皆さんに感謝したい (さらに何人かのかたには草稿へのコメントをいただいた— それでもなお本稿には筆者の間違いや極論も含まれているだろうから，それらに対する読者のご意見をいただけるとますます幸甚である)．

ふだんから自分自身の用いるデータ解析手法に常に疑問を持つ．それに関して他人と議論する．統計学誤用を避ける基本わざとはこのように誰もが実践できる簡単なものだ．

引用文献

Johnson D.H. 1999. J. Wildlife Management 63: 763-772.

粕谷英一 1998 「生物学を学ぶ人のための統計のはなし」．文一総合出版．

Kasuya E. 2001 Animal Behaviour 61: 1247-1249.

蓑谷千凰彦 1998 「すぐに役立つ統計分布」．東京図書．

Pawitan Y. 2001 “In all likelihood” Oxford science publications.

竹澤邦夫 2001 「みんなのためのノンパラメトリック回帰」．吉岡書店．