統計学概論

あるいは

大統計大曼荼羅

新ミレニアムの統計タントラを鳥瞰するために！

三中 信宏

minaka@affrc.go.jp
http://cse.niaes.affrc.go.jp/minaka/

租界Ｒの門前にて：統計言語「Ｒ」との極私的格闘記録
http://cse.niaes.affrc.go.jp/minaka/R/R-top.html

〒 305-8604 茨城県つくば市観音台 3-1-3
独立行政法人 農業環境技術研究所 生態系計測研究領域 上席研究員
東京大学大学院 農学生命科学研究科 教授［生態系計測学］
京都大学大学院理学研究科　連携併任教授［進化生物学］
東京農業大学大学院 農学専攻 客員教授［応用昆虫学］

確率統計的推論の認知心理的基盤

たとえ確率論や統計学についてまったく知らないあるいは関心をもたない一般の生活者たる人間であっても，日常生活を営む上で必ず「確率的もしくは統計的推論」を行なっている．自然現象を反映する数値データはある確率をともなって生じる確率変数すなわち変量と呼ばれる．変量を対象とするデータ解析・推定・検定・予測そして意思決定をおこなう学問が統計学である．ものごとの因果関係が必ずしも明らかではない，あいまいな状況のもとで，変量に関する限られた知見（データ）に基づいて未知の仮説（将来にわたる予想とか事物の関連性など）の是非を統計的に判定することは日常生活では頻繁に生じる．われわれ人間はそういう不確定状況での推論能力を自然淘汰の結果としてもちあわせている．現在の認知科学と進化心理学の研究対象となっている「帰納」とか「推論認知」とは，まさにこの推論能力－人間が生得的にもつ確率統計的な認知能力－を指している．われわれ人間は生まれながらに素朴な確率論・統計学を実践してきたのである．

しかし，人間がもつ素朴な確率統計の感覚的認知は必ずしもつねに妥当であるわけではない．認知心理学の最近の研究によると，人間は，場合によってはあるバイアスがかかった確率統計的認知を行ない，結論を誤ることがあることが明らかにされている．したがって，われわれは，必ずしも無謬ではない発見的思考法としての素朴確率統計認知が人間に

もともと備わっていることを前提として，数理統計学とりわけ生物統計学の理論の必要性ならびにその合理的な利用法を考えなければならない．

統計学の理論を長らく支えてきたのは，人間が行なう直感的判断への健全な懐疑心－すなわち経験主義の哲学－にほかならない．直感にたよっているかぎり，理論統計学が求められることはない．しかし，人間は誤りを犯すことのある生き物である．確率統計的直感もまた誤ることがある．どれくらい人間は確率統計的判断を誤るのか，その誤りを事前に防ぐにはどうすればよいのか－この問題意識は生物統計学を推進してきたそもそもの動機である．

生物統計学の生物学的なルーツ：現代的症候群からの治癒を目指して

数理統計学という数学の一分野は，とりわけ農学系・生物学系の統計学ユーザーにとっては手ごわい相手と一般にみなされている．その理由はおそらく変量の誤差構造の定量的分析という一見わかりにくいものの考え方にあると思われる．ある変量がどのような確率で値を生じるかという確率分布のモデル化を研究したドイツの数学者フリードリッヒ・ガウス（Friedrich Gauss）は，誤差のばらつきを表現するために正規分布という関数を開発した．この正規分布という確率分布は，現在もなお数理統計学の定礎の地位を保ち続けている．確かに，正規分布を前提とする数理統計学の理論体系は，推定と検定のためのさまざまなモデルと道具を生物統計学者に提供してきた．その貢献は正しく評価する必要がある．

しかし，正規分布の定礎の上にそびえ立つ理論の城を見上げる多くの農学系・生物系学習者は，数理統計学を学ぶためには正規分布に基づく理論体系を会得することが城門の通過儀礼として求められているとみなし，そして悩み続けている．その悩みのある部分は，学習者の初等的な数学的能力の欠如に起因するが，別の部分ははたして正規分布に基づく数理統計学が農学・生物学研究の現場にどれほど通用するのかという疑念に起因する．生物統計学を実践するには「正規分布を学べ」というスローガンだけでは学習者の心理的動機づけとしては不十分なのである．

今日では，機能的にも操作的にもすぐれた多くの統計解析ソフトウェアが高速のパーソナル・コンピューター上で比較的容易に利用できるようになった．大量の統計計算そのものに苦労したかつての時代とは彼我の感がある．しかし，ハードウェアとソフトウェアのシンポの恩恵を受け，統計計算の負担から解放された今日の統計学ユーザーには次なる陥穽が待ち受けている．それは，得られたデータを手近にある適当な統計解析プログラムに無思慮に投げ込んでそれで満足してしまうという現代的症候群である．

いったん開発された生物統計学の手法は，数学的に磨き上げればごく一般的な数理統計学の理論となる．数学的に洗練されてしまうと，データの形式さえ適合しているかぎり，どんな統計的手法でも適用できてしまう．たとえ，その手法の前提条件が満たされていなかったとしても，統計計算はつつがなく完了し，計算結果はきれいに出力されてしまう．ユーザーはその出力をみて満足してしまう．残念なことに，この症候群はしだいに蔓延しつつあるようだ．

しかし，ある統計的手法の適用が妥当であるかどうかは，数学的にではなく，むしろ生

物学的に判断されるべきである．そのためには，ある統計理論が生まれ出てきた生物学的ルーツこそ学ぶべきである．そのときはじめてある統計的手法の適用限界がわかるだろう．その手法の生物学的ルーツを知ったあとで，現代的に洗練された数学理論と格闘しようと決心してもあるいは使用する統計解析プログラムのマニュアルをひもといてもけっして遅くはない．

生物統計学のたどってきたルーツをふりかえるとき，きわめて逆説的ながら「数学は統計学にとって必須ではない」と断言できる．われわれ統計学ユーザーにとって本当に必要なのは，日常的に取り組んでいる農学・生物学上の具体的な問題状況の把握である．生物統計学で現在用いられている多くの理論はいずれも特定の生物学的問題の解決を目指して開発されたものである．たとえば，分散分析は，当時イギリスのロザムステッド農業試験場にいたロナルド・フィッシャー（Ronald A. Fisher）が圃場データを解析するために開発した方法である．また回帰分析は，生物統計学の祖であるフランシス・ゴルトン（Francis Galton）が親子間での関連性を解決するために編み出した手法である．

世には「統計イコール数学」とか「数学は統計の基礎である」という通説がまかり通っている．この通説のせいで，多くの統計学ユーザーは統計学の理論的背景に関して思考停止してしまい，結果として上記症候群の広範な蔓延をもたらす結果となった．もうそろそろこの通説から卒業しよう．本末転倒してはいけないのだ．われわれは，統計理論の会得やソフトウェアの習熟などではなく，なによりもまず農学・生物学上の具体的問題の解決を目指していたはずだから．

統計学とは？：データの変動からの推論

データはばらつく．たとえ精密を期した工業製品であっても，製造工程でのさまざまな確率的要因の関与により，製品の特性値にはばらつきが生じる．ましてや，生物では，遺伝的変動および環境的変動の複雑な絡み合いにより，観察データの中にはばらつきが生まれる．統計学が要求されるのは，ばらつきのあるすなわち変動のあるデータからある未知のパラメーターに関する推論をしなければならない状況においてである．

データのばらつきとは，次の２段階を経てはじめて定量化できる．まずはじめに，複数データ点の平均を計算することにより数空間のなかでのデータ点のおおまかな位置付けができる．つぎに，それぞれのデータ点が計算された平均値からどれほどばらついているかを分散として数値化することにより，データ集合としてのばらつきの評価が可能になる（実験計画法の項を参照）．統計分析の出発点はこのばらつきすなわちデータの変動である．

１変量データ・多変量データの別を問わず，われわれが統計理論を用いるときの出発点はデータの変動である．観察されたデータの値がばらつくとき，その原因は処置した実験処理の結果だろうか，それとも偶然誤差に起因したのだろうか．複数の実験処理を組合せたとき，それらの要因の間にはどのような関連があるのだろうか．統計学的な推定・検定とは，これらの問いに答えるための方法である．ある被検集団の平均値（パラメーター）の値を複数の無作為標本のデータ値から推定（点推定または区間推定）したり，あるいは平均値のパラメーターの大きさに関する仮説を検定することを通して，われわれは未知のパラメーターに関する推論を行なうことができる．統計学的な推論とは，データに照らし

て不適当な仮説を棄却することによって進められる．

これらの統計学的な疑問に答えるには，まずはじめにデータの変動というあいまいな現象をモデル化したり定量化したりする必要がある．上述のガウスの正規分布関数はそのための強力な武器の１つである．しかし現実には正規分布に正確に従うデータはない．正規分布（あるいは他のパラメトリック確率分布）からのずれが小さいときは，近似的にもしくは変数変換によって，正規分布ベースの推定・検定方法のようなパラメトリックな標準的統計手法を利用するのが常道である．しかし，そのずれが大き過ぎるときには，検出力は多少落ちてもノンパラメトリックな統計手法を用いるべきだろう．また，最近ではブーツストラップなど新たなコンピューター集約型の統計手法を駆使して経験的に確率分布を生成するというやり方も広く利用されるようになってきた．ベイズ統計学の利用もモデル選択や意思決定の場面では重要である．生物統計学の現場の事情に合わせて，既存の統計学の理論を鍛え直していく試みは今後も続けられていくだろう．

統計学ユーザーの心得

１変量統計学・多変量解析のいかんを問わず，そこで用いられる数学は言葉である．統計学者が数式を多用するのは，それが便利な言葉であるからにほかならない．しかし，統計学ユーザーはその学問的慣習に必ずしもなじむ必要はない．統計学の哲学的基盤は経験主義であり，その認知的ルーツはわれわれ自身がもっている素朴確率統計推論である．したがって，現在利用されている統計理論の根幹はすべて直感的に理解できるし，それをまず目指すべきである．

自分のデータを統計解析するとき，あるいは他人に頼まれて統計コンサルティングをするとき，ユーザーがあらゆる統計理論に通暁することは現在では不可能である．おそらくほとんどの農学系・生物系ユーザーは，自らの限られた統計学の知識を酷使して問題解決にあたっているという方がむしろ事実に近いだろう．事態をさらに悪くしているのは，統計学の世界があまりに広すぎるため，数理統計学に一生を捧げている専門の統計学者以外，この世界のどこにどのような統計手法があるのか，それらの手法の間の相互関係はどうなっているのかについてまったく闇の中という現実である．

とりわけ，統計学をはじめて学ぶ者にとって，いま学んでいる手法が統計界の中のどこに位置しているのかをまったく知らされないまま，数式やソフトをいじらされるというのは，教育上のみならず精神衛生上もよいはずがない．この点で統計学ユーザーに望みたいのは，統計学の世界の鳥瞰である．できるだけ広く生物統計学の視野をながめてみようということである．自分の抱えている問題解決にとって，いま使っている統計手法ははたして適切なのか，他にももっと使える方法があるのではないか－この素朴な知的好奇心こそ，蔓延する無思考症候群を予防し，主体的かつ積極的な統計学ユーザーへの道を拓くのである．

これから統計研修に出家されるみなさんに（引導渡しと憑きもの落とし）

私が農林水産省技術会議主催の統計研修の講師を引き受けてもう 15 年以上が過ぎました．その間，研修生（農水省および都道府県の農水関係研究員）から受けた質問 (感想) のうち，最も多かったのが「教わった統計手法がバラバラで相互の関連づけができない」というものでした．本来ならば，「統計学総論」のような講義があるべきなのでしょうが，統計学者の研究者適応度と個体数が最近顕著に低下している（絶滅危惧とも ...）農水省では，それだけの資質をもった講師を調達することが難しいのです．苦肉の策として 1993 年の統計研修最終日の質疑時間にふと思い付いたのが，統計学の世界を一枚の絵に描いてしまうという「**大統計大曼荼羅**」のアイデアでした．その場で鉛筆書きの「大曼荼羅」を質問用紙の裏側に描き，コピーして配ったのが「初版」です．意外に反響（？）があったので，気を良くして修正加筆しながらその後の統計研修で配布し，私が運営する生物統計学メーリングリスト（BIOMETRY*）でも宣伝してきました．

さいわい，翌年の統計研修からはこの「統計学概論」という科目が設けられ，短い時間ですが，「統計」という広大な世界について鳥瞰的な話ができる機会に恵まれました．1994 年 12 月の動物行動学会において初めて公衆の面前で「大曼荼羅」を発表して以来，密教（タントラ）とその曼荼羅の描き方についてしばらく勉強してもみたのですが，結局最初の鉛筆書き初版の修正版をそのまま掲載することに決めました．非最節約的な書き込みを削ったり，新たな相姦関係を記入したところもあります．しかし，骨格はまったく変わりありません．なお，恥しながら，この大統計大曼荼羅は，紙上（三中 1995）のみならず WWW 上でも公開（**）されております．

いまもチベットに残るタントラの教えでは，本来の曼荼羅は彩色した砂で描かれるそうです．そして，タントラ修法の終了とともにその曼荼羅は壊されるべきものなのだそうです．私の「大曼荼羅」もまた同じ運命をたどるべきであると考えます．それは，私自身，この「大曼荼羅」を改良していく意志を私がもっているという意味です．しかし，できることなら，読者のみなさんが自分だけの「マンダラ」を描くのがベストだろうと思います．

研修生のみなさんが，今日から始まる「出家修業」の間に，自分自身のための「統計マンダラ」を描かれることを私は願っています．

参考文献

粕谷英一 1998. 生物学を学ぶ人のための統計のはなし：きみにも出せる有意差 . 文一総合出版，東京 .

田中 公明 1987. 曼荼羅イコノロジー . 平川出版社 , 東京 .

三中 信宏 1995. コンピュータ集中型統計学に基づく誤差推定（付録：大統計大曼荼羅プロトタイプ）. 日本動物行動学会ニュースレター No.27, pp.4-15.

《マンダラ・コスモロジー：チベット仏教の知恵と心の芸術》，デジタローグ発売，2000．CD-ROM for Windows / Macintosh.

*）この怪しいメーリングリストに参加ご希望の方は，http://cse.niaes.affrc.go.jp/minaka/ml/biometry-top.html をごらん下さい．2007 年 8 月現在の会員数は 990 名を越えています．

**）電子版「大統計大曼荼羅」は，http://cse.niaes.affrc.go.jp/minaka/R/R-top.jpg をご覧下さい．

「統計数学」との清く正しい交際法

道を踏み外さないために

これからの修行中，みなさんはいたるところで「統計のための数学」なるものに必ず遭遇します．おそらく，講義によっては狼藉の限りを尽くす確率分布群団に圧倒されることも，強烈な微積分アッパーカットを食らうこともあるでしょう．講義中，脳死状態に陥る受講生の率を統計研修講師は密かに「死亡率」という不穏当な言葉で呼んでいます．死亡率が数式出現率と高い正の相関を持っていることは，残念ながら事実のようです．

では，数学を「とうの昔に忘れてしまった」（きっと）多くの受講生は，統計研修から無慈悲にも見離されてしまうのでしょうか？

世には「統計イコール数学」あるいは「数学は統計の基礎である」という通説が流されています．しかし，出家の道を踏み出されたみなさんは，そのような誤った雑念にとらわれてはいけないのです．つまり，

1．統計学の本質は直面している問題状況である

2．数学なんかぜんぜん怖くない

という人生の真理をぜひ会得していただきたいと心から願っています．

「どうせ俺（私）なんか，数学オンチなんだから，わかんなくってもしかたない」なんて諦めるのは 10 年早いのです．数学はただの言葉です．言葉がわからなければ，ボディランゲージで意志疎通を図ればいいのです．えっ，統計学のボディランゲージって何って？

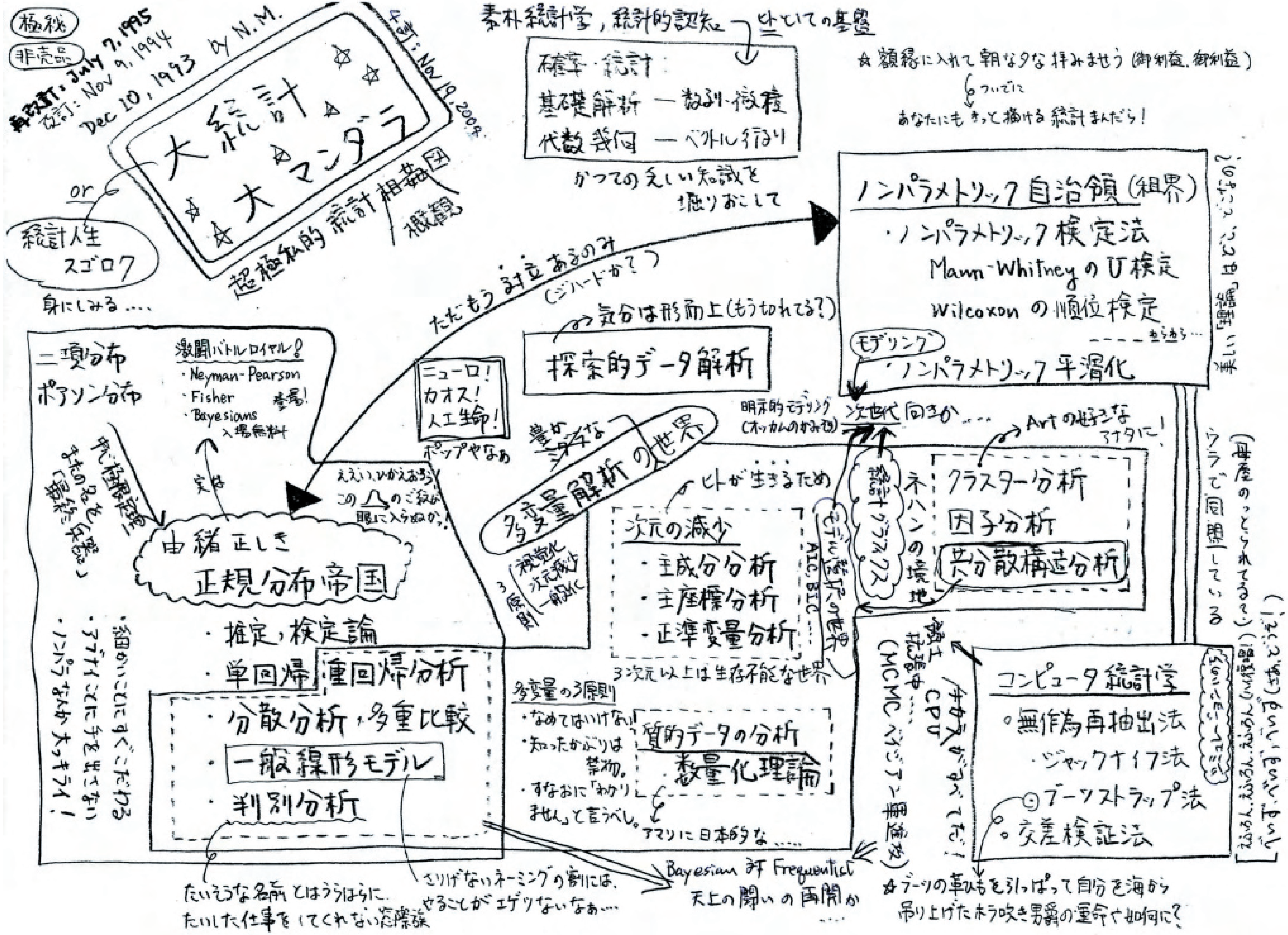
図と想像力ですよ．おそらくほとんどすべての統計手法は，適切な図を用いれば数式ゼロで説明でき，想像力をちょっとだけ働かせれば涙なしに理解できるはずです．数式いらずの統計学は絵空言ではないのです．受講生のみなさんは数式樹海で遭難しそうになったら，すかさず「要するにこの数式は何が言いたいのか？」とつぶやいてください．きっと，道が見えてくるでしょう．

いったいいつから数学はにくいにくい仇役になってしまったのでしょうね．数学はちょっと無愛想な友達なのです．今回の研修を通じて「愛せる数学，頼れる統計学」が少しでも実感できれば，研修は成功です．

では，みなさんの出家修業の成功を祈りつつ．

大統計大曼荼羅の正しい拝み方

研修期間中、毎朝の起床とともに、昇る朝日に向かって大統計大曼荼羅を掲げ、深く一礼します。しかる後、その日に学ぶべき科目が大統計大曼荼羅のいずこに座しているかを確認し、５分間の朝の瞑想に入ります。研修終了後は、世俗的・刹那的な飲酒・遊興・騒乱・咆哮に溺れることなく、身を律し、五穀豊穣・六根清浄・統計涅槃を念じつつ、５分間の夜の瞑想に入ります。就寝前、再び大統計大曼荼羅を静かにひもとき、一日を省みつつ一礼の上、消灯します。



極秘
非売品
再改訂: July 7, 1995
改訂: Nov 9, 1994
Dec 10, 1993 by N.M.
大統計大マンダラ
4訂: Nov 19, 2004
超極私的統計相姦図
概観
or
統計人生スゴロク
身にしみる....
素朴統計学, 統計的認知 — 生として の基盤
確率・統計
基礎解析 — 微分・積分
代数幾何 — ベクトル・行列
かつての えしい知識を 掘りおこして
☆額縁に入れて朝な夕な拝みませう(御利益、御利益)
ついでに
あなたにもきっと描ける統計まんだら!
ただもう対立あるのみ (ジハードか?)
ノンパラメトリック自治領(租界)
・ノンパラメトリック検定法
Mann-Whitney の U 検定
Wilcoxon の順位検定
......
モデリング
・ノンパラメトリック平滑化
気分は形而上(もうかれてる?)
探索的データ解析
二項分布
ポアソン分布
激闘バトルロイヤル!
・Neyman-Pearson
・Fisher 登場!
・Bayesians 入場無料
実は
中心極限定理
すべての道を 「最終兵器」
由緒正しき 正規分布帝国
ええい、ひかえおろう この印籠が眼に入らぬか。
ニューロ! カオス! 人工生命!
ポップやなあ
豊かな
多変量解析の世界
視覚化 次元縮少 一般化
3原則
ヒトが生きるため
次元の減少
・主成分分析
・主座標分析
・正準変量分析
3次元以上は生存不能な世界
AIC, BIC
モデル選択の世界
明示的モデリング (オッカムのかみそり)
次世代向きか....
Art の好きな アナタに!
統計グラフィクス
ネバーランドの境地
クラスター分析
因子分析
共分散構造分析
・推定, 検定論
・単回帰
重回帰分析
・分散分析, 多重比較
・一般線形モデル
・判別分析
細かいことにすぐこだわる
・アプナイことに手を出さない
・ノンパラなんかキッライ!
多変量の3原則
・なめてはいけない
・知ったかぶりは禁物。
・すなおに「わかりません」と言うべし。
質的データの分析
数量化理論
アマリに日本的な.....
Bayesian 対 Frequentist
天上の闘いの再開か ....
たいそうな名前とはうらはらに、たいした仕事をしてくれない窓際族
さりげないネーミングの割には、やることがエゲツないなぁ....
鋭意拡張中
(MCMC ベイジアン再興)
CPU
力ずくで何でもやってやる!
コンピュータ統計学
シミュレーションのウラ
○無作為再抽出法
・ジャックナイフ法
○ブーツストラップ法
○交差検証法
☆ブーツの革ひもを引っぱって自分を海から吊り上げたホラ吹き男爵の運命や如何に?
ウラで同盟している
(母産のっとりてるっ)
美しい「理論」はどこ、どこなの?
えふろん、えふろん、えふろん (ハイ復唱!)
べいず、べいず、べいず (さあ、どうぞ!)