



連載TOPへ

【書籍化決定！】本コンテンツの公開は2015年5月7日までとなります。

[SHARE] [Tweet](#) [シェア](#) [BI](#)

第12回 統計データ解析の地上世界と天空世界 —連載の総括として

実験医学2015年4月号

はじめに

前々回から2回にわたって、実験計画法の考え方を実際の数値データをお見せしながら説明してきました。どんな実験であっても、計画を立てるときには、この実験区配置（レイアウト）を実施したときにどのようなデータが得られることになるのかをよく吟味し、それは実験計画法が提示するスローガンに背かないことを確認する必要があります。せっかく時間と資金と人手をかけて行なう実験を無駄にしないためにも、実験をはじめ前のプランニングには十分に検討を重ねなければいけません。

いったん適切な実験計画に則ってデータが得られたならば、その先は統計データ解析の出番です。得られたデータのばらつきはどんな要因に基づくのか。実験計画を立てた時点で、私たちはデータのばらつきとその要因を説明する統計モデルを仮定します。このモデルには観測値に影響する（一つまたは複数の）処理効果および誤差効果が含まれます。

われわれが汗水流して表を作り、計算する“地上世界”では、モデルが仮定する要因に従って、データのばらつき（偏差）を要因ごとに分割し、平方和から平均平方（分散）と計算を進めれば、最後に分散比であるF値が得られます。一方、パラレルな“天空世界”では、正規分布に基づくパラメトリック統計学の理論により、帰無仮説すなわち処理効果がなかったと仮定したときに、F値がどのような確率分布（F分布）を示すのかが数学的に導かれています。実験計画法の統計解析で必ず用いられる分散分析とは、“地上世界”で数値的に求められたF値が、“天空世界”で導出されたF分布の棄却域に入るかどうかを判定する仮説検定にほかなりません。

このように、実験計画法とそれに伴う分散分析は、統計データ解析における数値計算と統計理論との関係を理解する格好の例を私たちに提示します。ともすれば私たちはデータから計算することに没頭してしまい、その背後にある論理や世界観を見失いがちです。しかし、実験データのふるまいは直感的なグラフや図表によって表されることを思い出しましょう。そのような直感（センス）があってはじめて数値化やモデル化の意味が実感をもって見えてくるでしょう。

§ 乱塊法—もう一つの実験計画法の例として

前々回と前回（第10回、第11回）に説明した完全無作為化法に基づく実験計画法は、反復実施と無作為化という2つのスローガンを組み込んで実験区を配置します。今回お話しする乱塊法（randomized block design）は、これら2つに加えてさらに局所管理というスローガンを掲げます。

まずは実際の例をお見せしましょう。実験計画法の事例と同じく、この実例もまたフ

羊土社HP会員 [English page](#)

ログインしていません

[羊土社HP会員とは？](#) [ログイン](#)

書籍検索

実験医学の定期購読

最新号がWEBでも読める！

国内送料無料

実験医学

月刊実験医学新刊

実験医学

次号予告

バックナンバー

連載一覧

掲載広告一覧

定期購読案内

詳細をみる

カートに入れる

実験医学増刊号新刊

実験医学

次号予告

バックナンバー

掲載広告一覧

定期購読案内

詳細をみる

カートに入れる

実験医学 電子バックナンバー発売中

DIGITAL ARCHIVE

新着情報 人材・セミナー案内

東京大学大学院 医学系研究科 疾患生命工学センター 分子病態医科学部 門

平成28年度 大学院生（修士・博士）募集および説明会

詳細や他の情報は[INFORMATIONコーナー](#)をご覧ください

羊土社新刊・近刊

骨ペディア

サイトカイン増殖因子キーワード

Dr.北野の0から始めるシステムバイオロジ

詳細 購入

詳細 購入

詳細 購入

>>新刊一覧へ

イリピンの国際イネ研究所（IRRI）で実際に行われた農業実験です¹⁾。この実験は、種もみの播種密度がイネの収量にどのような影響を及ぼすかを調べる目的で実施されました。種もみの播種密度は6水準（ヘクタールあたりの種もみの重量にして25 kg～150 kgの範囲）で設定し、反復数は4回です。もしもこの実験を完全無作為化法によって実施したならば、圃場をまずはじめに6水準×4反復＝24実験区に分割し、圃場全域にわたる無作為化配置をすることになったでしょう。

今回の乱塊法ではどのような実験区レイアウトを用いるのですか？

乱塊法とはあらかじめ反復ごとに「ブロック」を分割し、各ブロックの中で6水準すべてを無作為化配置するという実験区の割り付けをする方法です（図1）。実験計画法では無作為化をどのように実施するかによって違いが生じます。完全無作為化法の無作為化は、実験圃場に潜む環境要因がデータにどのような影響を及ぼすかがわからない状況で、そのバイアスを回避するという目的で行われます。一方、乱塊法は、実験に用いる圃場に関して事前に背景要因の傾向性がわかっているという状況で用いられる方法です。たとえば、図2のような場合を考えてみましょう。

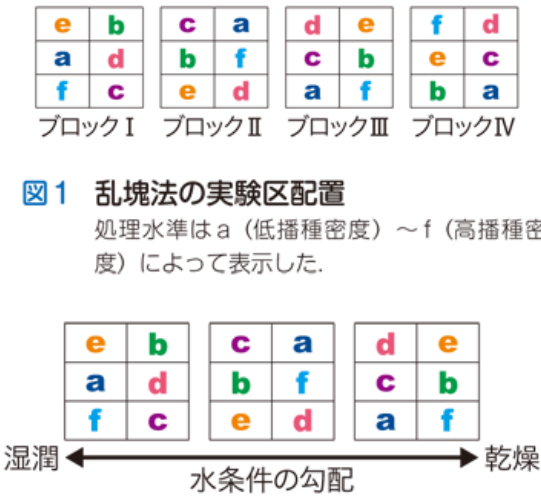


図1 乱塊法の実験区配置
処理水準はa（低播種密度）～f（高播種密度）によって表示した。

図2 乱塊法でのブロックの切り方

実験にしようとするある圃場に関して、東西方向（図の左右方向）に水条件に関する勾配があり、左側の場所は湿潤であるのに対し、右側は乾燥していることが事前にわかっていたとします。ある作物の収量に関する実験を予定しているならば、水条件の違いは得られるデータに大きな体系的バイアスをもたらすでしょう。このとき、水条件の勾配に“直交”する方向にブロックを切れば、データに影響を及ぼす可能性がある水条件を統制することができます。

図2では3つのブロックを設定し、乱塊法に基づいて6水準の無作為化配置をしました。このとき各ブロックは水条件に応じて左から「湿潤ブロック」「中間ブロック」「乾燥ブロック」と名づけられるでしょう。重要なのは、水条件に対応してブロックを切ることで、各ブロック内の環境条件をそろえた点にあります。これが第三のスローガンである局所管理です。すべての水準は水条件の異なるブロックでそれぞれ実施されるので、水準のもつ効果をよりはっきりと調べることができる。これが乱塊法の長所です。マウス実験の例では、マウスの血統、体重などがブロックごとに均一に配されるように設計することになります。

上の図1の乱塊法実験のもとで得られた数値データは表1のようになりました。このデータ表を各水準の行ごとに横方向に集計すれば、完全無作為化法の場合と同じく、処理平均を求めることができます。ところが、乱塊法ではそれに加えて列ごとに縦に集計することにより、ブロック平均も計算できます。処理平均は水準ごとの効果の大小を、そしてブロック平均はブロックごとの効果の大小を数値化しています。乱塊法の線形統計モデルを図3に示します。

学会売行き良好書情報

● 第37回 日本分子生物学会 年会 (14/12/02)

>>過去の売行き情報はこちら

実験医学 550号 突破!

アンケートに答えて
「実験医学」を
多数のご回答ありがとうございました

実験医学 @Yodosha_EM on Twitter

実験医学 jikkenigaku on Facebook

教科書・サブテキスト
をお探しの方へ

臨床医学系書籍
TOPページ(総合)

プライマリケアと救急を中心とした総合誌
レジデントノート
月刊 増刊

表1 乱塊法実験で得られた数値データ

処理 (kg seed/ha)	Rep.1	Rep.2	Rep.3	Rep.4	処理和 (T)	処理平均
25	5,113	5,398	5,307	4,678	20,496	5,124
50	5,346	5,952	4,719	4,264	20,281	5,070
75	5,272	5,713	5,483	4,749	21,217	5,304
100	5,164	4,831	4,986	4,410	19,391	4,848
125	4,804	4,848	4,432	4,748	18,832	4,708
150	5,254	4,542	4,919	4,098	18,813	4,703
ブロック和 (R)	30,953	31,284	29,846	26,947	119,030	総和 (G)
ブロック平均	5,159	5,214	4,974	4,491		4,960 総平均

データ 総平均

$$x_{ij} = \mu + \alpha_i + \rho_j + \epsilon_{ij} \quad (\epsilon_{ij} \sim N(0, \sigma^2))$$

第*i*水準 第*j*ブロック

処理効果 ブロック効果 誤差項

誤差の正規性(仮定)

図3 乱塊法の線形統計モデル

完全無作為化法になかったブロック効果の項 (ρ_j) がはいってますね。

そうです。観測データ (x_{ij}) が総平均 (μ) のまわりでばらつく要因は、右辺に示されているように、処理効果 (α_i) とブロック効果 (ρ_j) そして誤差項 (ϵ_{ij}) です。誤差項は水準*i*と反復*j*に関係なく常に同一の正規分布 $N(0, \sigma^2)$ に従うと仮定します。完全無作為化法よりも複雑なこの線形統計モデルには、ブロック効果が明示的に含まれていることに注意しましょう。この統計モデルをことばで説明するならば、第*i*水準・第*j*ブロックから得られたデータ x_{ij} は、第*i*水準効果 (α_i) と第*j*ブロック効果 (ρ_j) の和に加えて誤差項 (ϵ_{ij}) を含むということになります。

この統計モデルに基づく偏差の分割は表2に示す通りです。目標はデータのもつ全偏差を処理偏差・ブロック偏差・誤差偏差の3つに分割することです。完全無作為化法で示した手順を適用すれば、続く平方和と平均平方の計算を実行し、最終的に誤差平均平方を分母として処理平均平方を分子とする*F*値、ならびに同じ誤差平均平方を分母に対してブロック平均平方を分子に置いた*F*値が計算できます。これらをまとめた分散分析表を表3に示します。この実験ではブロックは有意な効果が見出されました ($F > F_{0.01}$) が、処理に関しては有意な効果は検出できませんでした。

表2 乱塊法における偏差分割

処理 (kg seed/ha)	Rep.1	Rep.2	Rep.3	Rep.4	処理和 (T)	処理平均
25	5,113	5,398	5,307	4,678	20,496	5,124
50	5,346	5,952	4,719	4,264	20,281	5,070
75	5,272	5,713	5,483	4,749	21,217	5,304
100	5,164	4,831	4,986	4,410	19,391	4,848
125	4,804	4,848	4,432	4,748	18,832	4,708
150	5,254	4,542	4,919	4,098	18,813	4,703
ブロック和 (R)	30,953	31,284	29,846	26,947	119,030	総和 (G)
ブロック平均	5,159	5,214	4,974	4,491		4,960 総平均

全偏差 = 処理偏差 + ブロック偏差 + 誤差偏差

処理偏差

ブロック偏差

表3 乱塊法における分散分析表

変動要因	自由度	平方和	平均平方	F値	$F_{0.05}$	$F_{0.01}$
ブロック	3	1,944,361	648,120	5.86	3.29	5.42
処理	5	1,198,331	239,666	2.17 ^{ns}	2.90	4.56
誤差	15	1,658,376	110,558			
全体	23	4,801,068				

誤差に対するブロック間の分散は有意な効果が認められたが、処理間の効果は認められなかった (ns)。

乱塊法の帰無仮説は、完全無作為化法と同じく、処理効果とブロック効果をともに含まない、誤差項のみの統計モデルです。上の分散分析では処理効果とブロック効果を含む対立仮説の統計モデルに対して F 検定を実施し、要因が有意であるかどうかを F 検定しました。観察データのもとで、どのような統計モデルを当てはめるのが妥当なのかという議論は、モデル選択 (model selection) というもっと大きな問題につながっていきます。

§ 総括—情報可視化，統計モデリング，アブダクション

連載¹²回もあっという間でした。無味乾燥に思っていた統計学のイメージが少し変わりました！

統計学は生きているサイエンスの¹つとして、時代の風潮や傾向と無縁ではありません。今風には“データ・サイエンス”と呼ぶ方がカッコイイようですが、いわゆる“データ・ドリブン (data-driven)”な科学研究は、大量の情報と高性能コンピューターの追い風に乗って、今後さらにその影響力を増していくのかもしれませんが、その一方で、科学研究の場だけでなく、いま私たちが生きている現代社会のなかでも、より多くのデータや情報を手にすることにより、あたかも宝探しのように“金脈”が掘り当てられるというちょっと都合のいいイメージが膨らんでいるようです。

しかし、本連載ではもっと地道な案内図をみなさんに提示しました。実験や観察を通して私たちが手にするデータには、それを生み出した因果構造がどこかに埋め込まれています。定量化された数値データはそのままでは私たちには解読できません。そこで、さまざまな統計グラフィックスのダイアグラムを利用することで、データの示すふまけは誰もが理解できるように可視化することができます。万人がわかること—まさにこれが統計的データ解析の原点であるはずで。

エドワード・R・タフティ (Edward R. Tufte) は、長年にわたって、数値データをいかにして“見える”ようにできるかというデータ可視化 (data visualization) の問題に取り組んできました²⁾。昔から私たち人間は膨大かつ複雑なデータを可視化すべく試行錯誤をくり返してきました。つい最近私が翻訳したマニエル・リマ (Manuel Lima) の最新刊『THE BOOK OF TREES—系統樹大全』³⁾ もまた、多様性情報の可視化をめぐる一千年に及ぶ知的系譜を明らかにしてくれました。データや情報を視覚化するインフォグラフィックス (infographics) を単なる流行語のまま終わらせるのではなく、現場で利用できる実質的なツールとして鍛えあげることが必要とされています。統計的データ解析はこの目標を見失ってはなりません。

その一方で、データや情報は既知の知見から未知なるものへの推論、すなわちアブダクションを目指すというもう一つの目標があります。形態測定学者フレッド・L・ブックスタイン (Fred L. Bookstein) の大著⁴⁾ に力説されているように、データの統計モデリングを通じて、私たちはよりよい暫定的結論を導くことができます。ただし、このアブダクションという推論には終わりがありません。真実を前提としないアブダクションは、新たなデータが出現するたびに新たなよりよい (しかしやはり暫定的な) 結論へと移行します。

膨張し続けるデータの可視化と、はてしないアブダクションの連鎖—一身の丈サイズの統計学は情報の海とモデルの山を越える翼を私たちに与えてくれます。

おあとがよろしいようで。

文献

1. 「Statistical Procedures for Agricultural Research, Second Edition」 (Kwanchai A. Gomez & Arturo A. Gomez) , John Wiley & Sons, 1984
2. 「The Visual Display of Quantitative Information, Second Edition」 (Edward R. Tufte) , Graphic Press, 2001
3. 「The Book of Trees: Visualizing Branches of Knowledge」 (Manuel Lima) , Princeton Architectural Press, 2014
『THE BOOK OF TREES—系統樹大全：知の世界を可視化するインフォグラフィックス』 (マニエル・リマ／著 三中信宏／訳) , ビー・エヌ・エヌ新社, 2015
4. 「Measuring and Reasoning: Numerical Inference in the Sciences」 (Fred L. Bookstein) , Cambridge University Press, 2014

[SHARE] [Tweet](#) [シェア](#) [B!](#)

[Prev](#) [12](#)

[TOP](#)

「第12回 データ解析と統計的推論―連載の総括として」は、本誌[2015年4月号](#)を御覧ください

本記事の掲載号



実験医学 2015年4月号 Vol.33 No.6
グルカゴン革命 糖尿病の真の分子病態を追え！

北村忠弘／企画
定価 2,000円＋税， 2015年3月発行
[▶詳細](#) [▶購入](#)

本連載に関する質問・感想、統計に関する具体的な悩みを編集部までお寄せください！

- 下記画像中の英数字をご入力ください



[画像を変更する](#)

おすすめ書籍



[▶詳細](#) [▶購入](#)



[▶詳細](#) [▶購入](#)



[▶詳細](#) [▶購入](#)



[▶詳細](#) [▶購入](#)



[▶詳細](#) [▶購入](#)

[会社案内](#) | [採用情報](#) | [個人情報取扱い](#) | [お問い合わせ](#) | [広告掲載について](#)

(C)2015 [YODOSHA CO., LTD.](#) All Rights Reserved.