2011.6.6 中澤 港 (nminato@med.gunma-u.ac.jp)

生物統計学と臨床試験

テキスト『シンプル衛生公衆衛生学 2011』には独立した項としての記載がないが，疫学の実践には生物統計学が欠かせない。疫学ばかりでなく，他の多くの科学研究のバックグラウンドとなっている。生物統計学は英語で言うと biostatistics あるいは biometrics である。元を辿れば R.A. Fisher がロザムステッドの圃場で実践した実験計画法であり，医学分野では主として毒性試験の用量反応関係と，臨床試験がそれに当たる。

1. 参考文献

- J.L. フライス（著），KR 研究会（訳）『臨床試験のデザインと解析』株式会社アーム，2004 年
- 厚生労働省「治験」ホームページ(http://www.mhlw.go.jp/topics/bukyoku/isei/chiken/index.html)
- 改正 GCP 省令(http://law.e-gov.go.jp/htmldata/H09/H09F03601000028.html)
- 丸山由起子(2004)『CRC という仕事』メディカル・パブリケーションズ
- 佐久間昭(1964)『生物検定法－その計画と分析－』東京大学出版会
- 永田　靖(2003)『サンプルサイズの決め方』朝倉書店
- 東京大学生物測定学研究室編，岸野洋久著(2004)『実践生物統計学－分子から生態まで』朝倉書店
- ゲルト・ギーゲレンツァー(2010)『リスク・リテラシーが身につく統計的思考法』ハヤカワ文庫
- Salsburg D (2001) "The Lady Tasting Tea: How statistics revolutionized science in the twentieth century." A W.H. Freeman / Owl Book（邦訳：ザルツブルグ『統計学を拓いた異才たち』日経ビジネス人文庫，2010 年）
- 佐藤俊哉・松岡淨(2001)『何があっても割り付け通りに解析する』日本行動計量学会講演資料（著者のサイトから pdf ファイル[http://www.kbs.med.kyoto-u.ac.jp/01Sep15.pdf]をダウンロードできる）
- 佐藤俊哉・佐藤恵子(2005)『吉村先生、IDMC やってくださいよ』第 4 回東京理科大学医薬統計フォーラム講演資料（著者のサイトから pdf ファイル[http://www.kbs.med.kyoto-u.ac.jp/IDMC.pdf]をダウンロードできる

2. 実験計画

- 何らかの仮説の実験的証明には，目的とコストに見合った適切な実験計画を立てる必要がある
- 漫然と実験すると解析段階で本質的情報が欠けていることに気付いてデータ全体が無駄になることもある

2-1. Fisher の 3 原則

処理の違いに基く差が偶然である可能性が極めて低く，偶然ではありえない（有意）かどうかを判定するために必要な計画の原則として，R.A. Fisher が提唱した，以下の 3 つを Fisher の 3 原則という。

実験管理の 3 原則

反復：1 つの処理に対して少なくとも 2 回以上の繰り返しが必要である。1 度だけでは偶然のばらつきがどの程度あるのかが評価できないので，相対的に，群間の違いが偶然のばらつきに比べて，統計的に意味があるほど大きいのかということも評価できない。

無作為化：実験の順序や空間的にどの場所にどの実験群を割り当てるのかを無作為に決める必要がある。

> ある特定の処理条件を圃場の特定の場所にかためてしまうと，その場所がたまたま水はけがよかったり肥沃だったりした場合，処理効果によるものか場所の効果に よるものかの判別ができなくなる。このようなある特定の場所による効果を系統誤差(systematic error)というが，無作為に処理条件を配置させることにより，系統誤差を偶然誤差(random error)に転化させることができる。（岸野，2004）

局所管理：実験が大規模で，実験全体を無作為化するのが妥当でないとき，実験をある程度細分化してブロックを構成し，ブロック内で処理条件を無作為化し，ブロック内のバックグラウンドが均一になるよう管理すると，系統誤差の一部がブロック間変動として除去できる。

2-2. 実験計画法の発想

本当かどうか知らないが，実験計画の始まりについては，次のようなエピソードが知られている。ティーの席での話題として，ミルクティーを作るときにミルクを先にカップに入れたのか，紅茶を先にカップに入れたのかを，飲んでみれば見分けられる（これをミルクティー判別能力と呼ぶことにしよう）という女性の話が出たときに，多くの学者が化学的には差がないのだから見分けられるわけがないとか，いや見分けられるかもしれないとかいう中，実験してみれば？といったのが R.A. Fisher であった。どちらを先にして作ったのかを知らせずに，この女性にミルクティーを飲んでもらって当てさせてみれば，本当にミルクティー判別能力があるのかわかるというのだ。しかし，1 度だけ試して当たっただけでは偶然かもしれないので，何度か繰り返して試さなくてはいけない。それに，ミルクが先という場合だけで試すと，偶々片方だけ言い続けた人が全問正解してしまうことになるので，両方の条件を試さなくてはいけない。つまりは，どういう順番で何回試してみれば，得られた結果からその女性にミルクティー判別能力があるのかが判定できるような条件を考える必要がある。

この条件を考える方法のことを実験計画法と呼び，何度繰り返さなくてはいけないかがサンプルサイズの設計に当たり，どういう順番で試すかが試験配置法に当たる。

2-3. サンプルサイズの設計

ミルクを先に入れて作ったミルクティーを 1 杯飲んだとき，まったくの山勘で答えると，ミルクが先と答えるか紅茶が先と答えるかは確率 1/2 なので，偶然当たってしまう確率が 1/2 もあることになる。2 杯飲んだときに 2 杯とも当てる確率を考えると，当たりを○，外れを×と表記すると，○○，○×，×○，××が等確率で起こるので 1/4 となる。1/4 というのはそれほど珍しいことではないので，2 杯では，どういう結果が出ようがミルクティー判別能力があるのかどうか判定不能である。では，最低何杯試したらいいのだろうか。ある水準より多く偶然当たる確率が 0.05 未満のときに，それはありえないことと判断して，偶然ではない（＝ミルクティー判別能力がある）と結論できるとすると，3 杯試して偶然で全て当たる確率は 1/8，4 杯では 1/16，5 杯では 1/32 となるので，最低 5 杯は試す必要があることになる。このとき，0.05 という判断基準（有意水準）は，裏を返せば，本当は差がないけれども間違って差があると判定してしまう確率になるので，第一種の過誤と呼ばれる。

サンプルサイズを計算するには，まず臨床試験の主要なエンドポイント（評価項目）と統計解析の方法が決まっていなくてはいけない。統計解析の方法によって（割合を比較したいのか，生存時間を比較したいのか，など）必要なサンプルサイズは変わってくる。その上で，割合を比較する場合なら(1)有意水準，(2)検出力，(3)コントロール治療での臨床イベント発生割合，(4)試験治療のイベント発生割合がコントロール治療よりどれくらい小さければ臨床的に意義があると考えられるか，その最小の値（しかしそれで計算すると必要なサンプルサイズが非実用的なほど大きくなることが多いので，(4')試験治療により期待できるイベント発生割合）が必要。

割合を比較する統計手法は Fisher の直接確率（または近似としてカイ二乗検定）なので，その式が決まっていて，例えば有意水準が片側 5%，検出力が 80%，コントロール治療でのイベント発生が 30%，試験治療により期待できるイベント発生が 15%の場合なら，各群 95 人となる。

* R では，

```
power.prop.test(p1=0.15, p2=0.3, sig.level=0.05, power=0.8, alternative="one.sided")
```

2-4. 試験配置法

試験配置にはいろいろあるが，コストや調べたい対象などによって，どういう試験配置が適しているかが変わってくる。良く用いられる配置法としては， 乱塊法（randomized block design），分割区法（split-plot design），ラテン方格法，要因試験などがある。

平行群間比較試験

もっとも単純なデザイン。インフォームドコンセントが得られた適格な患者がランダムに割り付けられ，いくつかの治療のうちの1つだけを受ける。ランダムな割り付け方としては，Fleiss は乱数表ではなくランダム置換表を単純に用いる（例えば 100 までの 整数の乱数置換表ならば，1 から 100 までの整数をランダムに並べ替えた表がいくつも提供されていて，割り付ける群数を決めて，第 1 群から順に必要なサン プルサイズまでの数字

を表の出現順に拾っていく)，あるいはランダム置換ブロック法(例えば 3 群ならばまず 1～3 をランダム置換表から拾い，次に 4～6 を 拾うとすることで，登録例数が予定に届かなかった場合にも群間のサンプルサイズがアンバランスになるのを防げる)ことを勧めている。しかし，今ではコン ピュータが簡単に使えるので，ランダム置換表を用いるメリットはそれほどない。

乱塊法

Fisher の 3 原則を完全に満たす。すべての処理組み合わせの実験を 1 回ずつ集めたもので 1 つのブロックを形成する。ブロック数が反復数になる。

分割区法

水など広い区画で管理しないと労力や費用がかかる場合，乱塊法は非現実的なので，広い区画で1次因子，その区画ごとの細かい施肥量や品種などの条件を2次因子とする分割を行う。通常，1次因子は交絡であり，2次因子間での比較が真の目的となる。

ラテン方格法

R.A. Fisher が考案した方法で，例えば 5 種類の処理を比較したいときに，効果を調整したい要因(交絡になりうる要因)が 2 つあるとしたとき，2 つの要因の 組み合わせをクロス表にした場合に，行と列のどれをとっても 1～5 の数字が一度だけ出現するように割り付け，その数字を比較したい処理の番号とするもので ある。動物実験ならば，例えば，5 種類の食餌を与えた場合のマウスの成長の差を比べるデザインであれば，効果を調整したい要因 1 が 5 匹のマウスの同腹仔， 要因 2 が産まれた順序になるし，臨床試験の場合は，通常，要因 1 が被験者個人，要因 2 が処理が施される時期である。

3. 毒性試験

化学物質などについて、人や生物に好ましくない作用の有無またはその強さの程度を調べるための試験。

試験は、評価する毒性の項目(一般毒性、特殊毒性)、使う生物の種類(哺乳動物、魚など)と形態(全体、組織、細胞など)、曝露経路(経口、吸入、経皮など)、曝露期間(長期、短期など)によって様々な種類がある。目的によって、適切な試験方法を選定する必要がある。**(http://www.eic.or.jp/ecoterm/window.php?ecoterm=%C6%C7%C0%AD%BB%EE%B8%B3)**より，「暴露」を「曝露」に修正

とくに，用量反応関係(量-反応関係ともいう。dose-response relationship)については多くの方法が開発されてきた。LOAEL(最低毒性量)，NOAEL(無毒性量)，NOEL(無反応量)を求めるこ とや，ロジット分析とかプロビット分析によって LD50(半数致死量)あるいは ED50(半数影響量)を求めることが多い。横軸に用量，縦軸に反応割合をとってプロットするのは基本であり，通常はシグモイド(S 字状)曲線になる。

4. 臨床試験

4.1 臨床試験とは

ヒトに対する実験(侵襲あり)を臨床試験という。新薬とか新しい治療法は，モデル動物で効果があるだけではダメで，どうしても，ヒトに効くか，ヒトに有害作用がないかを確認する必要があり，臨床試験は必須。無駄になってはいけないので，科学的かつ倫理的に考え抜かれた計画に従って行われねばならない。様々なガイドラインがあって，それに沿って計画する必要がある。

4.2 臨床試験の段階

前段階(非臨床試験)：細胞，組織，動物を使った実験(主として毒性試験)＝安全性確認
第 I 相試験：健康な成人のボランティアを対象として，薬物動態や最大許容量を調べる
第 II 相試験：比較的少数の患者を対象として，有効性，安全性，用量反応関係を調べる
第 III 相試験：数百から数千の患者を対象として，「薬の候補」の有効性を，科学的に検証する目的で行う。RCT(無作為化比較試験)として二重マスク化(二重盲検)で行われる
第 IV 相試験：市販後に大勢の患者が実際に服用した結果，新薬がどういう特徴をもっているか，副作用はでないかを調べる(市販後臨床試験)

4.3 倫理的要求

ヘルシンキ宣言が大原則。しかしこれだけでは具体的にどうすればいいのか曖昧。
具体的には，日米欧による International Conference on Harmonization (ICH)という会議により，いくつものガイドラインが公表されている。

- 医薬品の臨床試験の実施の基準に関する省令(GCP)
- 臨床試験のための統計的原則について（統計ガイドライン）
- 臨床試験における対照群の選択とそれに関連する諸問題（対照群に関するガイドライン）

4.4 臨床試験の手順

試験実施計画書の作成：実験なので当然。
試験実施計画書に沿った試験の実施

- 計画書は必ず守る
- 倫理的問題：有害作用に苦しむ患者に同じ治療を続けられるか？→この場合は計画書からの逸脱が正当化される（省令 GCP 第 46 条）

データ解析：薬の候補を使う，使わないはランダムに割り付けられるが，倫理的な理由から割付が守られない場合があるのが問題。その場合の扱いとして，ITT (intention to treat / intent to treat)は重要。原則としては，守った人だけ（計画書に適合した対象集団＝バイアスがかかっている）を使った解析と，最初の割付通りに解析する（ITT；ただし，ランダム割り付け後に実験参加不適格であることが判明したとか，1 度も薬を飲まなかったとか，ランダム割り付け後のデータが一切ない人 については，解析から除外してもいい場合があり，その場合は，残りの「最大の解析対象集団」について，割付通りに解析する）のと，両方やって，一致した結果が得られれば OK。

4.5 治療効果の判定と説明

治療効果の判定と説明には，いくつかの違ったやり方がある。有名なのは以下の3つ。どれも正しいのだが，誤解の受けやすさは異なる。

＊相対リスク減少率（1－リスク比）

＊絶対リスク減少率（リスク差，超過危険，寄与危険）

＊要治療数(NNT＝絶対リスク減少率の逆数)

（例）コレステロール低下薬の効果

5 年間プラバスタチンを服用した 1000 人のうち 32 人が冠動脈疾患で死亡，偽薬を飲んでいた 1000 人のうち 41 人が死亡。新聞報道は，「プラバスタチンを飲むと死亡リスクが 22％低下した」

⇒一般市民の多くは，プラバスタチンを飲むと 1000 人の高コレステロール血症患者のうち 220 人が心臓発作を免れると誤解した

(Quiz)　3つの判定指標を計算してみてください（正解は3分考えた後で画面に出すのでメモしてください）

世論を動かし研究費を得るには，効果は大きく見えた方がいいために，相対リスク減少率が使われることが多い，という指摘がある。確かに 22％の減少！　の方が絶対リスク減少率や NNT より大きな効果に感じる。

しかし一方では，新型インフルエンザの致命割合が 0.01%なのか 0.02%なのかということを，「2 倍の違い」というか「0.01%の違い」というか，0.01%でも1億人中だと1万人の違いになることを意識したら，一概にどちらが適切とは言えない。

リスクを適切に伝え，相互理解に至る「リスクコミュニケーション」は，公衆衛生学的に重要。