

2003.11.17

生物多様性論 I: 「生態学における統計学的手法の基礎」 (と勝手に改題)

全部で 3 回講義の 1

「検定」の使われかたを観察しよう

— 「検定」はそんなにエラいのか? —

<http://hosho.ees.hokudai.ac.jp/~kubo/stat/2003/>

講釈: 久保拓弥 kubo@ees.hokudai.ac.jp

この 3 回だけの統計学授業でやること

統計学が **何もわかってない** 修士課程大学院生が対象。
この善男善女どもがデータ解析に際して「統計学って、
まるっきり理解不能」と遭難する確率を減少せしめる
べく、この世界の **おおまかな地図**を提供し解説する。

- 第 1 回: 2003.11.17 (月)
「検定」の使われかたを観察しよう
— 「検定」ってそんなに**エラい**のか?
- 第 2 回: 2003.11.19 (水)
統計モデリングと推定を重視してみる
— **理解できる**統計学めざして
- 第 3 回: 2003.11.26 (水)
データにあわせる統計モデリングの例とか
— なんでも**割算すんな!**

個別的なワザより全体に共通する考えかたを— ただし内容は**偏ってるよ**

統計学とおつきあいするためには

あなた自身のココロがまえとして

- われわれが必要とするような統計学は**難しいものではない**，しかしながら勉強するのに**時間はかかる**かもしれない— 確率，という非日常的なヘンてこな抽象概念などとおつきあいするから
- 疑いぶかく— あなたも私もいつでもどこかで何かを誤解している— そして世の中は統計学の**誤用だらけ**

とりあえず

- よい教科書が必要（インターネット上にもいろいろある）
- よい統計ソフトウェアが必要（R…… 後述）
- できれば相談できる相手も

必読! 粕谷英一「統計のはなし」

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 —

- 著者: 粕谷英一 (九州大・理・生物)
- 出版: 文一総合出版
- サイズ: A5 判 / 199 ページ
- ISBN: 4-8299-2123-4
- 発行年月: 1998.3



「この【ピンク本】を読まずにすますことはできない」
(三中信宏・書評 on “BK1” <http://www.bk1.co.jp/>)

これ使いましょう: 統計ソフトウェア R

<http://www.r-project.org/>

- いろいろな OS で使える **freeware**
- 使いたい機能が充実している
- **作図**機能も強力
- S 言語による **プログラミング**可能
- よい教科書が出版されつつある
 - 「R による統計解析の基礎」 中澤港 (2003)
 - “Introductory Statistics with R” P. Dalgaard (2002)
 - “Computational Statistics” M. Crawley (2002)
 - **ネット上**のあちこち



R が変えつつある生態学のデータ解析

- 使いたい手法はたいていそろってる
- 無ければ自分で何でも簡単に作れる
- 統計学的 simulation も簡単にできる



..... となると

- データを無理やりある手法にこじつける, ということが不要になる— **データの構造にあわせた**統計モデリングを行えばよい
- 手法の前提となる**統計学の基本**(統計モデル) の理解がむしろ重要
- 単純な検定ではなく、「こういう標本のばらつきを生成したメカニズム」の**推定**のよしあしが問われる

今日のハナシ: 「検定」はそんなにエラいのか?

統計学的手法とは何か?

人間の測定誤差, 手ぬきサンプリングで生じた偏り, 自然の気まぐれが作りだした正体不明の「ばらつき」……などなどによってわけのわからなくなった観測データに「もっともらしい説明を与える」(つまり**モデリング**する) 技法

問題は, この統計学的手法の中で

生態学のデータ解析において「検定」だけがひどく重要視されちゃってる現状

この現状を把握しつつ, 私たちが少しでも楽しいデータ解析生活を送れる方法を考える(「検定って何?」というのは**あとから説明**するとして)

生態学研究で「検定」が重要視されてる現状

- この授業のもとタイトル「**統計検定**の基礎と応用」－木村さんによる命名
統計学 → 検定なのか？
- 「統計学的有意差検定の意味の無さ」 Johnson, Douglas H. 1999. The Insignificance of Statistical Significance Testing. *Journal of Wildlife Management* **63**: 763-772.
「検定」使われすぎな現状について
- 次の会話は「しゅーる」だけどたいへんリアルだ……



粕谷 “ぴんく本” 「統計のはなし」第一章「検定って、何?」より

浦井 飯山さん，村田が実験がいちおう終わったんで結果を検定したいそうなんですが，**有意差が出る検定**っていうのがあったと思うんですが，教えてください．

飯山 有意差の出る検定? — (中略) — いつでも有意差のでる検定なんてないよ．

浦井 村田，飯山さんは知らないらしい．誰か他の人に聞かなきゃだめみたいだ．

以上のやりとりから，**浦井君**は「ゆーい差決戦主義」学派の急進的な一派に属している，とわかる

そもそも「検定」って何だろう？

これを理解するためには，生態学研究における
データの流れを把握しなければならない

「流れ」は研究のスタイルによって変わる:

1. 実験計画法つかう研究 (起源: Fisher?)
 - 薬学・農学における**室内・圃場実験**とか
 - 興味のない要因を比較的簡単に統制できる
 - 実験材料についてよくわかってる
2. 探索的データ解析による研究 (起源: Tukey?)
 - **野外調査**や bioinformatics など
 - 興味のない要因，そう簡単には統制できん
 - 実験材料についてよくわかっていない

(1) 実験計画法つかう研究では

何か実験をするとする

1. 実験の目的 (どういう処理をしたいか) と実験生物を決める
2. 実験生物に関してすでにわかっている情報を使って, 事前の統計モデリング— 実験計画が決まる
3. 実験
4. 得られた観測データを事前に決めておいた解析手法で調べる— ここで最終的に検定する (処理は予測された効果を出したかどうか)

日本の (農学ではなく理学よりな) 生態学者でこれがやってるヒト, いるんだろうか? — 分散分析はやたらと使いたがるけど

(2) むしろ探索的データ解析による研究では

何か野外調査をやるとする

1. 調査の目的と調査場所・対象を決める
2. (私のみるところ) **てきとー**にデータ取り計画を決める (この段階では解析段階のことはほとんど考えていない)
3. 調査
4. データが得られてから, その解析方法・統計学的手法の適用を考える

「野外実験やってます」とかいうヒトも含めて, たいの生態学者のデータの流れは, おそらくこちら.

検定にはさまざまなものがあるんだけど.....

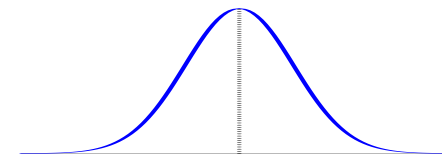
- 何と何を比較するか
 - 「真のモデル」と標本
 - 標本 vs 標本
- 「真のモデル」はどんなカタチか?
 - 離散分布 (二項分布, ポアソン分布,))
 - 連続分布 (正規分布, ガンマ分布,))
- 何を比較したいのか? (書ききれん)
 - 位置統計量 (標本平均, とか)
 - 尺度統計量 (ばらつき, つまり標本分散, とか)
 - 順序統計量 → いわゆるノンパラメトリック検定

ここから, とても簡単な例を使って「**検定って何?**」を説明してみよう

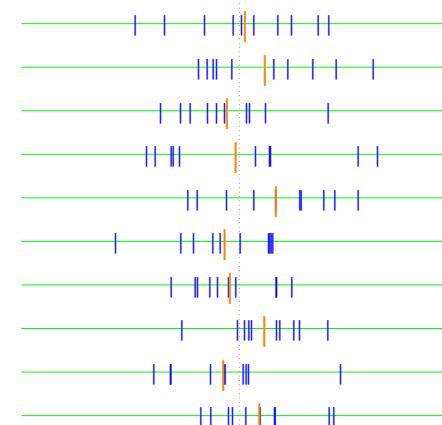
最初に，統計モデリングの検討が必要

- ここでは説明のため，非常に理想化された実験— **架空**の実験を考える (乱数生成と作図は R による)
- 架空の実験動物 X の Y 日齢目における体重 W を考える
- この動物 X については大変よくわかっており，体重 W とそのばらつきを生成する統計モデルがわかっている (つまり母集団を表現する確率分布は**既知**)
- ここではランダムサンプリングされた**標本の平均値**にだけに関心がある，とする

母集団 (真のモデル)



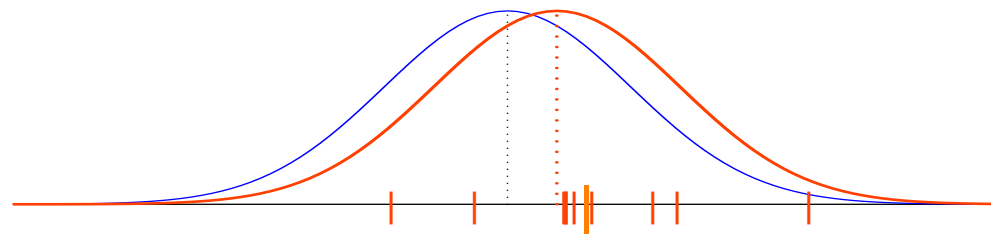
標本 (観測データ, $n = 10$)



図中の | は標本平均

実験処理をした標本のデータと 帰無仮説

実験処理: 新開発の栄養剤 Z を投与すると実験動物 X の成長が良くなるかどうか, を調べたい



Y 日齢目における体重 W が増加するか?

ここではホントに増えちゃったとしよう (上の赤線)

— しかし観測者には標本 (上の赤点) しか見えていない

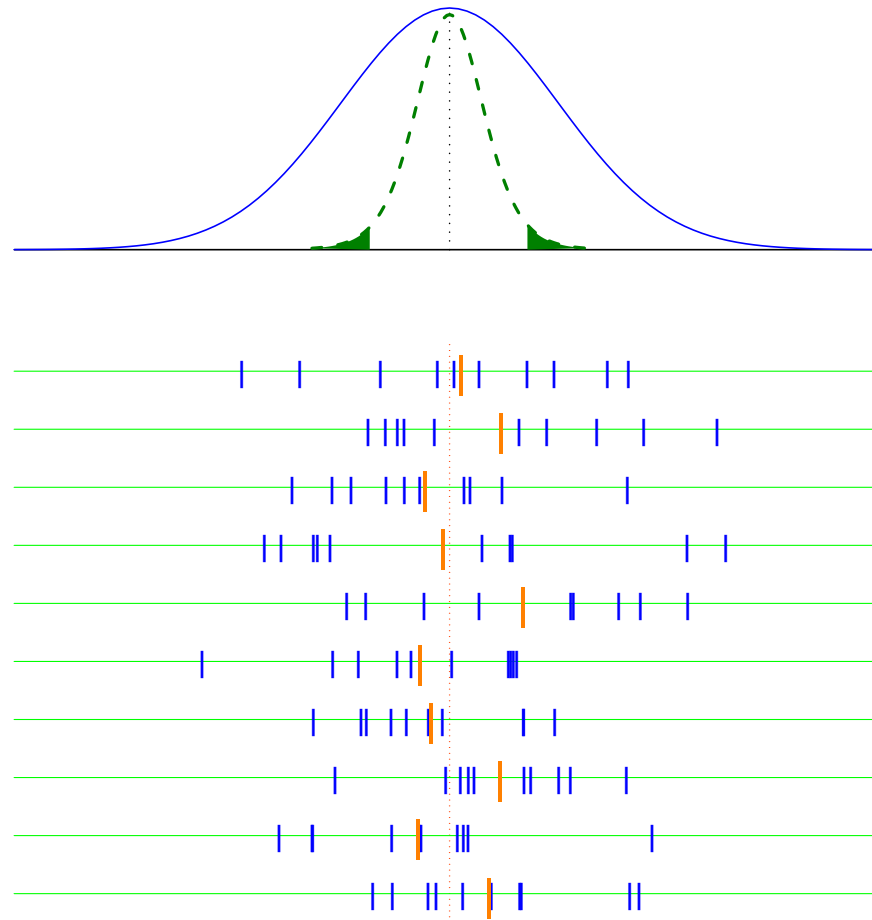
次に

帰無仮説 = 「そうはならないはず」仮説を設定する

「新開発栄養剤 Z を投与しても体重 W は増えない」というのを**帰無仮説**とする— そうなってほしくない状況を **帰無仮説**として, これを調べる (すごくひねくれている)

「めったに起こらん場合の確率」を計算する

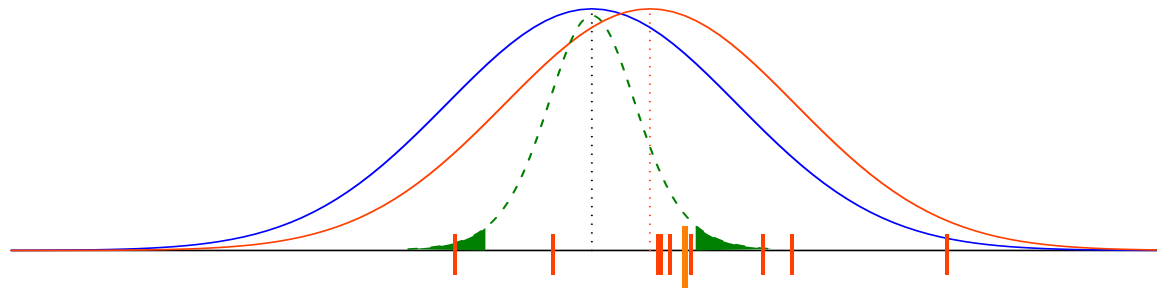
- 帰無仮説が正しいと仮定
- その仮定のもとで、標本平均がどれくらいばらつくか、という **また別の** 確率分布を計算する (詳細略)
- 「めったに発生しない標本平均」 — **棄却域** なるものを両端に設定できる



危険率 $\alpha = 0.05$ とすることが多い (いわゆる 5% 水準 — 5% に特別な意味ナシ)

統計学的検定における二つの過誤 (errors)

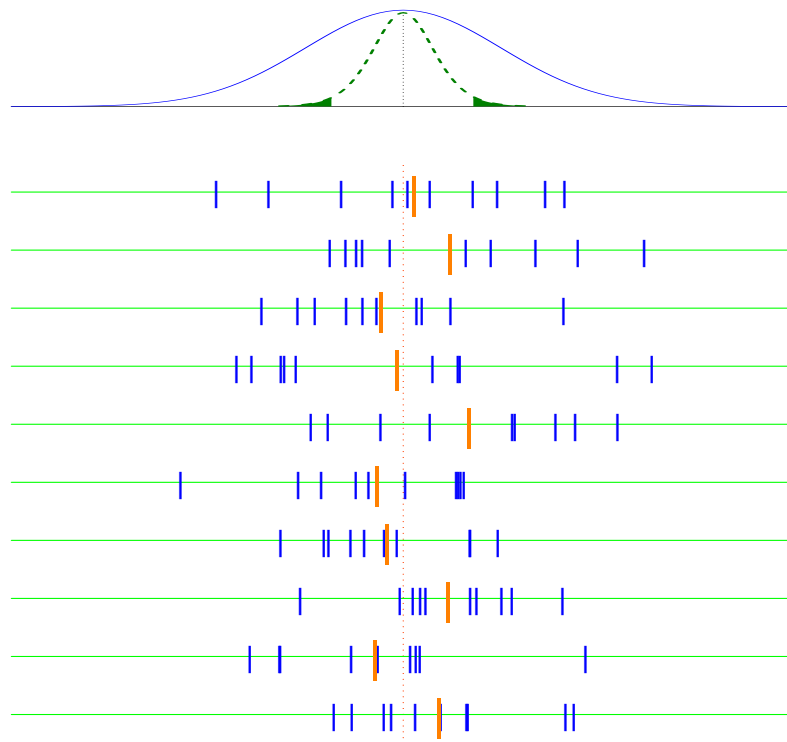
棄却域と実験データの標本平均を比較してみる



	標本平均が 棄却域内	標本平均が 棄却域外
帰無仮説が正しい	帰無仮説を誤棄却 — 第一種の過誤	帰無仮説は棄却でき きん, と結論
帰無仮説が正しくない	帰無仮説を棄却	帰無仮説を誤採用 — 第二種の過誤

第一種の過誤: 違ってないのに「違う」と判断

	標本平均が 棄却域内	標本平均が 棄却域外
帰無仮説が正しい	帰無仮説を誤棄却 — 第一種の過誤	帰無仮説は棄却で きん, と結論



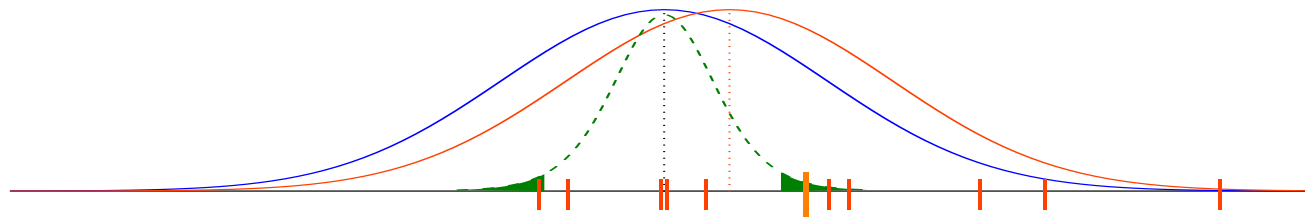
Type I Error

標本集団が 帰無仮説 (というモデル) から生成されている場合でも, たまに「外れた」統計量 (この場合は標本平均 — 左の図中の |) が得られるため

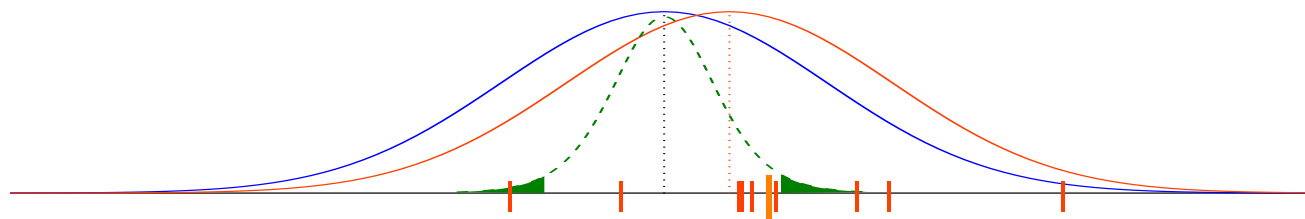
第二種の過誤: 違ってるのに「違わない」と判断

	標本平均が 棄却域内	標本平均が 棄却域外
帰無仮説が正しくない	帰無仮説を棄却	帰無仮説を誤採用 — 第二種の過誤

- 帰無仮説を棄却できてる例 (標本平均 | が 棄却域内)



- 第二種の過誤をやってる例 (標本平均 | が 棄却域外)
— Type II Error



検定の非対称性 — 第一種の過誤だけを重視する

	標本平均が 棄却域内	標本平均が 棄却域外
帰無仮説が正しい	帰無仮説を誤棄却 — 第一種の過誤	帰無仮説は棄却でき きん, と結論
帰無仮説が正しくない	帰無仮説を棄却	帰無仮説を誤採用 — 第二種の過誤

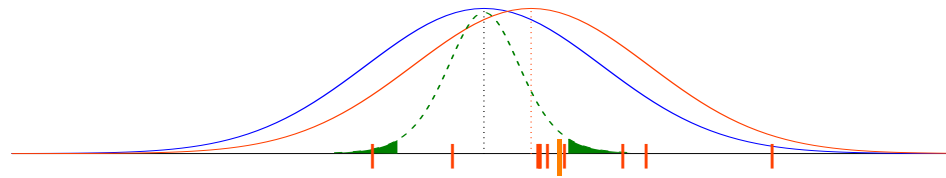
- 統計学的検定においては, この第一種の過誤だけをチェックする
- 統計学的検定では第二種の過誤は考慮されない (cf. 検定力 β)

なぜか?

「違う」ということだけ言いたいから— ここでの例だと:

- そもそも新開発の栄養剤 Z が実験動物 X の体重 W 変化を変えてしまうことが期待されている
- 疑わしきは採用せず, という前提がある

つまり検定から言えることは.....



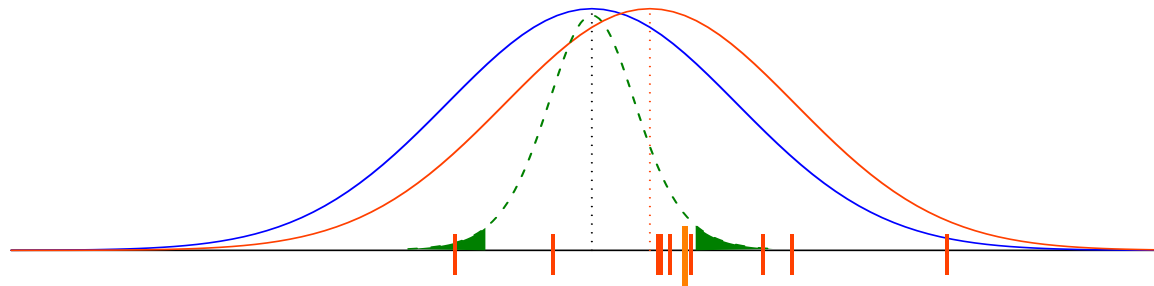
ここでは標本平均
だけで検定してる

1. 標本平均が **棄却域内**にある (p 値が危険率 α より小さい, などという)
→ 帰無仮説が棄却された, すなわち新開発栄養剤 Z によって体重 W の確率分布が変化した, と結論する
2. 標本平均が **棄却域外**にある (p 値が危険率 α より大きい, などという)
→ 帰無仮説が棄却されなかった, すなわち**何も言えない**— 第二種の過誤を無視してるから
 - 栄養剤 Z が効いてない?
 - サンプル数が足りない? (cf. 検出力)
 - 実験材料や実験方法に何か問題があった?
 - などなど (特定できん)

「君にもできるごまかし」 (in ぴんく本)

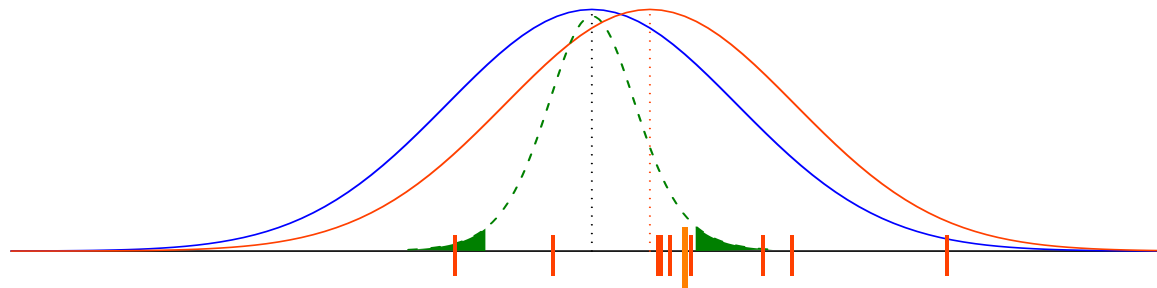
- 有意じゃないから 帰無仮説が証明された
- ぎりぎりでも有意でなければ，片側検定にする
- データを変換してパラメーターを合わせる
- (などなど.....)

(こういったごまかしはすでに陳腐化してるから使うのやめましょう，という著者からの警告.....為念)



久保の座右の銘: 科学者は職業的な嘘つきである — ただし，すぐにばれる嘘をつくのは単なる間抜けだ

問題点: 「ゆーい差」なんて簡単に出せたり



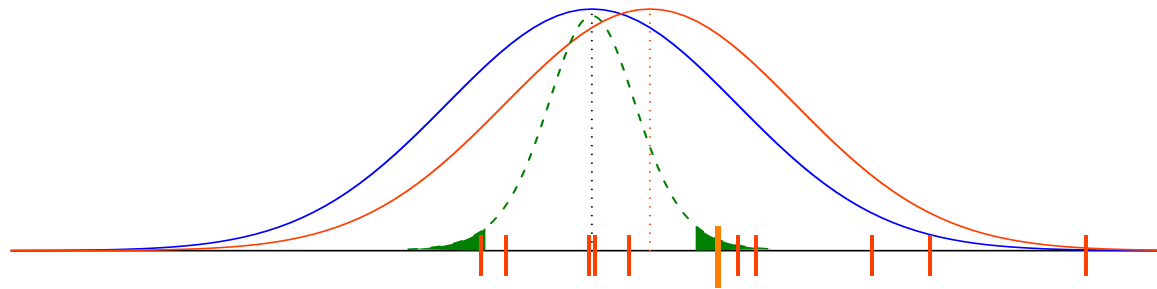
野外調査で得られたデータには、統制されてない隠れた要因が影響していたりするので、サンプル数さえ増やせば「ゆーい」になることが多い

「帰無仮説が棄却できた」というのは「十分なサンプル数とりました」と同義だったりする場合も (それはそれで重要だが)

逆に、「『ゆーい差』でてほしくない、と思ってるのでサンプル数を減らしてみました」ってのは、恣意的であると同時に、

「帰無仮説が棄却されなかった → 帰無仮説が正しい」という間違いもやってる。

問題点: 統計学的有意差だけでいいのか?



- 「統計学的有意差がある」すなわち「標本は 帰無仮説とは異なる確率分布から得られた」ということ
- で、**あなた**の知りたい生物学的有意差とは? (これは統計学で直接に扱う問題ではない)

検定で「ゆーい差」さえ出れば、それで何でもヨシ、とする立場を「ゆーい差決戦主義」という— 統計学ユーザーには「ゆーい差決戦主義者」がたくさんいる

あなたのデータ解析は検定でケリがつくのか?

もう一度データ解析の流れを考えてみよう

1. 統計モデリング (事前的な実験計画法, 事後的な探索的データ解析)
2. 推定 — これはつまり観測データを統計モデルに**あてはめる**こと
3. 推定結果の吟味
 - 検定: **第一種の過誤の危険**がある水準以下かどうかを判定— **帰無仮説**に説得力を持たせる実験準備が必要
 - その他: モデル選択 (予測力の最大化する推定になってるかの検討) など— 「**より良い推定**」を探索する道具

やってないのに**あたかも**「実験計画法やりました」てなかんぢで解析しても.....

統計モデリングや推定といった話題は, また次回以降に

本日のまとめ

- 検定とは何か? をまずはよく知ろう
- 統計学的手法ってのは**検定だけではない**
 - 実験計画・サンプリング計画
 - 統計モデリングの模索
 - パラメーターや区間の**推定**(推定量・統計量の計算)
 - **推定結果**の検討(これの一部として検定がある)
- あなたのデータ解析には**検定が必要かどうか**, よく考えよう
- 「検定を使うな」というススめでは**ない**
 - 「不適切な使いかたすんな」というススめ

次回予告

2003.11.19

生物多様性論 I: 「生態学における統計学的手法の基礎」 (と勝手に改題)

全部で 3 回講義の 2

統計モデリングと推定を重視してみる

— 理解できる統計学めざして —

“このばらつきを説明できるモデルを探せ!”

<http://hosho.ees.hokudai.ac.jp/~kubo/stat/2003/>