生物統計学
2012 年 4 月 10 日

第１回　生物統計学とは？

Ａ．授業の枠組み

１．講義について

農業，農学実験においてはどんなにていねいな実験をしたとしても数値にはかなりのばらつきがある．ばらつきを考慮しないで平均だけで実験の結果を議論するならば大きな誤りを犯す可能性がある．このように生物，農業などで得られるばらつきのあるデータから適切な結論を得るための統計学が生物統計学である．この授業では実際のデータに当てはめていくことで統計的な考え方を身につけるようになるのが望ましい．したがって，自分でデータを使った計算を必ずこなし，授業に積極的に参加して，統計的考え方を育てることが必要である．

授業ではノートパソコン（授業ではエクセルおよびエクセルに付属する分析ツールの使い方を説明する）をかならず持参のこと．

授業のホームページ http://www.ipc.shimane-u.ac.jp/food/kobayasi/
に資料などをのせておく．授業用のデータ集もあらかじめダウンロードしておくこと．

２．宿題について

宿題は原則として**次回の授業の前日**（通常は，授業の翌週の月曜日）**午後１時まで**に Moodle で提出してください．Moodle は　https://moodle.cerd.shimane-u.ac.jp/moodle/　にあります．Moodle の使い方は別紙をみてください．

宿題の解答をエクセルファイルとして,Moodle 上に提出してください.提出した宿題を採点し，印刷したものを次回の授業の最初に返却します．間違えた問題を訂正して，再提出することができます．宿題の解答は宿題の締め切りが過ぎてから，Moodle で公開します．再提出の時はこの解答を参考にしてもかまいません．ただし丸写し（そのままプリントアウト）はしないこと．宿題を期限以降に出しても採点はします．その場合は提出期限からどれだけ遅れたかによって，減点をします．

３．成績について

以下の３つから成績を判定する．① 授業参加，② 宿題，③ 期末試験（ノート，参考資料，ノートパソコンの持ち込み可，通信機能の使用は不可）

Ｂ．エクセルの分析ツールの組み込み（Excel2007 の場合）

１．分析ツールが組み込まれているかを確認する

データをクリックし，データ分析があるかを確認する．データ分析がないときは，②の方法で分析ツールを組み込む．

２．分析ツールの組み込み

一番左上のボタンを押し，現れたメニューの下にあるエクセルのオプションを押す．オプションのメニューが出たら，左のフレームメニューからアドインを選ぶ．アドインできるアプリケーションのリストが並び，アクティブでないアプリケーションアドインのリストの中に分析ツールをクリックし，設定をクリックする．次に表示されたアドインのリストの中から，アドイン（組み込み）するべきプログラムにチェックを入れる．ここでは分析ツールにだけチェックを入れる．

このとき場合によってはエクセルの入ったＣＤを要求してくる場合がある．このときは自分のパソコンに付属した，あるいは自分のパソコンにインストールしたエクセルのＣＤを用意する．

C．生物統計学で何ができるのか？

１．実験して得たデータの解釈

実験をして得たデータの解釈をどうするか？

Ａ，Ｂ，Ｃの３種類の餌を与えたヒツジの成長を調べた．どの餌がいちばんよいか？

気温が上昇するとブドウの糖度はどうなるか？

データにはばらつきがあるので，１つや２つだけ調べてもそれが本当に正しいかは確信を持てない．ではデータをどのようにとったら確信を持ってもよいのだろうか？

★　考えてみよう

ある池に有害物質が流れ込んだ．池から調査のために魚を 5 匹捕獲し，5 匹とも基準値以下だった．この池の魚は基準値以下と保証できるか？

データの正確さはばらつき（誤差）の範囲を使って示す（統計的推定，信頼区間）

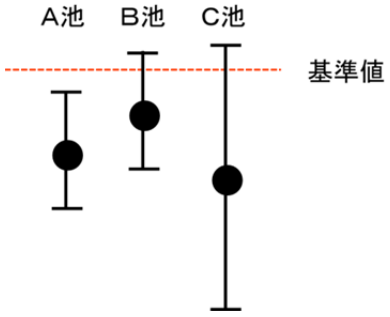
例えば，10 回，硬貨を投げて，表が何回出るかを推定してみよう．表と裏が均等な確率で出現するいびつでない硬貨なら，平均は（　　）回である．しかし，そうはいっても，ちょうど（　　）回だけ表が出る確率は（　　）％にすぎない．そこで表が出現する回数を範囲を使って示すと，より高い確率で予想できることになる．つまり・・・(二項分布で計算できる，世論調査の正しい知識　ニュートン 2008 年 8 月号より)

表が 4～6 回出る確率は・・・(　　）％程度

表が 2～8 回出る確率は・・・(　　）％程度

一般に確率 95%あるいは 99%（この確率を信頼率と呼ぶ）のもとで，平均など示したい値の範囲（この範囲を信頼区間と呼ぶ）を示すのが一般的である．すなわちこの場合 95%あるいは 99%の範囲で基準値を満たしていることを示すとよい(なお統計学では 100%保証することはできない)．

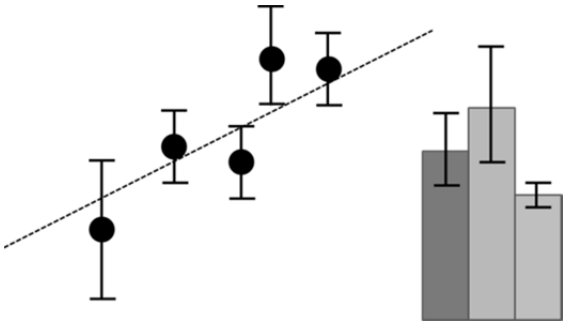
信頼区間で示すことによって，決められた確率のもとで範囲を使ってデータを示すことができる



信頼区間の例

台風の予報円　70%の確率で暴風域圏内に入る

誤差　個々のデータが持っている誤差をグラフ上では誤差線（エラーバー）で，表では±標準偏差あるいは標準誤差という形式で表す．(信頼区間を示すとは限らないので注意)



項目	処理あり	処理なし
A	23.8±0.6	18.4±0.0
B	12.3±1.4	10.8±0.7
C	46.0±2.1	61.5±4.0

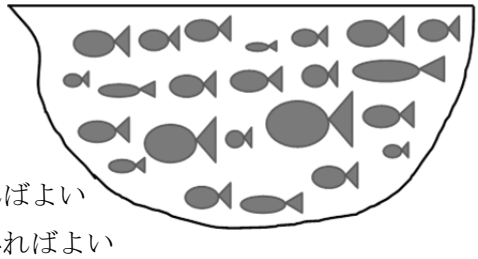
何パーセントの確率（信頼率）の信頼区間とすべきかは目的によって変わる

台風の予報円ではあまりに信頼率を高くしてもかえってわかりにくくなる

震度 6 強の地震で家が倒壊する恐れのある範囲ならどうだろうか？

自動車，宇宙船などの部品の信頼率だったらどのくらいがよいだろうか？

池の中の魚の体重を知りたい．魚は池に百匹以上はいて，しかも正確な数はわからない．魚の体重は図のようにかなりまちまちである．



1）平均体重が知りたいなら何匹調べたらよいだろうか？

（　　）内に自分の考える数を入れよ
もし池の魚が 100 匹だと分かっていれば（　　）匹調べればよい
もし池の魚が 10000 匹だと分かっていれば（　　）匹調べればよい
もし池の魚の数が不明だとすれば（　　）匹調べればよい

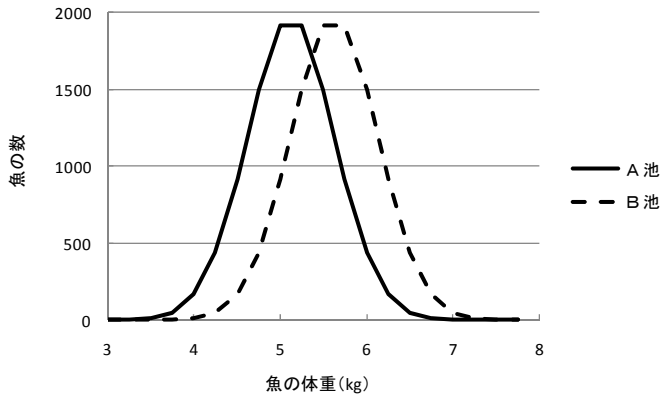
2）同じ形の池が２つあった．しかし，一方は富栄養化していて魚の体重が大きくなったようだ．この仮説を証明するには２つの池からそれぞれ何匹を調べたらよいだろうか？

（　　）内に自分の考える数を入れよ
もし池の魚が 100 匹だと分かっていれば，それぞれの池から（　　）匹調べればよい
もし池の魚が 10000 匹だと分かっていれば，それぞれの池から（　　）匹調べればよい
もし池の魚の数が不明だとすれば，それぞれの池から（　　）匹調べればよい

1）は第６回の講義で学ぶ統計的推定，2）は第７回の講義で学ぶ統計的検定と関連がある．

サンプリングを実際に体験して，データがどうなるかを知りたい人は 10000 匹の魚の体重（正規分布するデータ，仮想のデータ）で実際にサンプリング実験をやってみよう（生物統計学_授業用データ集 2012 のエクセルファイル内にある）．

A 池の魚の平均体重は 5.0kg，B 池の魚の平均体重は 5.5kg である．どちらの池の魚の平均体重のばらつきは同じで標準偏差 0.5kg である．標準偏差 0.5kg というとどのくらいばらつくだろうか？下のグラフのようになる．



２つの池から無作為にサンプリングした結果，A池の魚の体重は（　　）kg，B池の魚の体重は（　　）kg となった．教室全体ではA池の魚の方が重かった学生は（　　）名，B池の魚の方が重かった学生は（　　）名いた．

さて１回だけの実験では今回の結果がたまたま偏った結果かも知れないので，100 人の学生が 1 回ずつ 2 つの池からサンプリングした実験を 100 回行ったら，A池の魚の方が重い学生は平均 23.85 人，B池の魚が重い学生は平均 76.15 人となった（詳細はホームページに掲載）．

このようにたった１匹だけでどちらの池の魚の体重が重いかを判断しようとしたら 4 回のうち 1 回は間違った結論を出すことになる．しかし，このような判断をする人は世の中に多い．

質問：たった１つの数値だけで判断してしまった経験を下に書け．

しかし，学生がそれぞれ 5 匹ずつ，2 つの池からサンプリングして，その平均を比較したらどうなるだろうか？　A．やっぱり 3 回に 1 回はA池の魚の平均が大きい，B．A池の魚の平均が大きくなることが増える，C．A池の魚の平均が大きくなることが減る

100 人の学生が 5 匹ずつ 2 つの池からサンプリングした実験を 100 回行ったら，A池の魚の方が重い学生は平均 5.85 人，B池の魚が重い学生は平均 94.15 人となった．さらに 100 人の学生が 10 匹ずつ 2 つの池からサンプリングした実験を 100 回行ったら，A池の魚の方が重い学生は平均 1.32 人，B池の魚が重い学生は平均 98.68 人となった（詳細はホームページに掲載）．

以上のことから，１回のサンプリングで採る魚の数が多くなるほど，B池の魚が重いという結果が　（　少なくなる　・　変わらない　・　多くなる　）．つまり標本をたくさんとるほど（　正しい結果を得る可能性が高くなる　・　正しい結果を得る可能性は変わらない　・　正しい結果を得る可能性が減る　）．

ではこの魚の体重のデータは平均とばらつき（**標準偏差**という）が指定されていた．では 2 つの池の魚の平均体重の差が大きいほど，調べる魚の数（**標本数**という）を（すくなくしてもよい・変わらない・多くしなければいけない）．2 つの池の魚の体重のばらつきが大きいほど，調べる魚の数を（すくなくしてもよい・変わらない・多くしなければいけない）．

池の中の魚をすべて調べれば正しい結論を得るはずだ．しかし，現実には魚をすべて調べるだけの時間，労力，お金を使えることはほとんどない．場合によっては破壊したり，殺したりする実験なら全部調べることが不可能である（例えば，魚の胃の中の内容物を調べるとしよう．全部，魚を調べたらその池には魚がいなくなってしまう・・・）．さらに原理的にすべて調べることが不可能な場合もある（仮説的無限母集団の概念，詳細は第３回で学ぶ）．

２．（　　）の法則

昔のヨーロッパ人は彗星は前触れもなくやってきたので不吉なことの前兆ととらえた.

★　科学の始まり　ケプラーの法則　ブラーエの集めた膨大な天文学データから

ケプラーの法則から万有引力の法則へ

ニュートンはケプラーの法則から万有引力の法則を導いた

万有引力の法則ですべての天体の運動を説明できる

多数のデータから少数の法則を導くことができた→少数の法則から**予測・発見**ができた

海王星の存在の予言，ハレー彗星が周期的に地球の近くに来ることを予言など

たくさんデータを集めれば集めるほど確実である.

保険のデータ　人間の死や事故はわからないことだらけ

統計を集めて，確率的に掛け金を決める（生命表　車の事故）

★　統計をみて考えよう.

①　日本の出生率と出生性比（女子新生児 100 人に対する男子新生児数）の右の表を見て気づいたことを書け.

	西暦	出生率	出生性比
明治38年	1905	31.2	102.7
明治39年	1906	29.6	108.7
明治40年	1907	34.0	102.7
明治41年	1908	34.7	104.6
明治42年	1909	34.9	104.1
昭和39年	1964	17.7	105.9
昭和40年	1965	18.6	105.3
昭和41年	1966	13.7	107.6
昭和42年	1967	19.4	105.3
昭和43年	1968	18.6	107.1
昭和44年	1969	18.5	107.2
昭和45年	1970	18.8	107.1

②　人の寿命はそれぞれ異なる．しかし，たくさんのデータを集めると傾向がわかる．日本人の平均余命が年々伸びていっていることを示す右の表を見て気づいたこと，さらにこのようなデータはどういうことに利用されているかを書け.

２．少数例で何かいえないか？ 近代的な統計学の登場

大数の法則とはいうけれど，たくさんのデータを集めるのは大変である

できるだけ少ない・多くのデータでできるだけ少ない・多くの法則（結果）を得たい・・・

大数の法則にかなうぐらいのデータを集めることは困難なことも多い

自動車の耐久性テスト 少ない・多くの台数を破壊試験して，しかも正確なデータがほしい

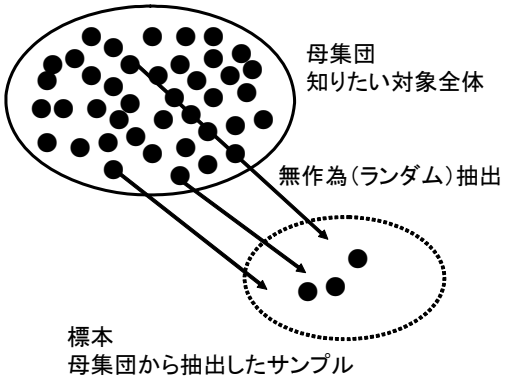
オオサンショウウオの食性を知りたい 数自体が少ない・いくらでもいる

猫のエイズの治療薬 効かないかもしれない副作用の強い薬は動物虐待かも？

それゆえできるだけ少数の・多数のデータでできるだけ少数の・多数の法則（結果）を得たい

標本（サンプル）から母集団を推定する 近代的な統計学の手法

母集団（調査対象すべて）から**ランダム（無作為）**に取ったサンプル（母集団からの代表）であれば，統計学からどの程度の精度・誤差で判断できるかがわかる．



正しく抽出した標本から母集団を推定する実例と誤差

視聴率　標本サイズと誤差　関東地区で 1455 万世帯から 600 世帯

山陰地方で 45 万世帯 200 世帯をサンプリングする

視聴率が 20%と出たら，その誤差はどのくらい？どっちの精度が高い？

視聴率の誤差はサンプルの大きさに依存し，母集団の大きさとサンプルの大きさの比はほとんど関与しない．誤差を大まかに見積もると右の表のようになる．真の視聴率と測定された視聴率との誤差は 95%の確率で以下の範囲内に収まることを右表は示す．

視聴率	標本数600	標本数200
5%・95%	±1.8%	±3.1%
10%・90%	±2.4%	±4.2%
20%・80%	±3.3%	±5.7%
30%・70%	±3.7%	±6.5%
40%・60%	±4.0%	±6.9%
50%	±4.1%	±7.1%

３．正確なデータを取るために：標本の選び方は難しい

標本に要求されること

母集団を代表しているか（統計的な手法を適用するための条件）

精度が必要な程度あるか（サンプルの数が多ければ精度は高くなる）

質問　あるおじいさんがこういいました.「わたしは９０歳になるが，毎日，たばこを１箱以上すっている．わたしの妻も８８歳になるが，毎日，わたしの副流煙をたくさん吸っている．しかし，二人とも元気でぴんぴんしている．しかし，となりのおじいさんは禁煙するぞといったが，１年後に肺がんがみつかって，すぐに死んでしまった．タバコは副流煙も含めて，健康，長寿の秘訣で，禁煙こそガンの原因だ.」　この論理はどこがおかしいのでしょうか？

★　母集団を代表しているか　　インターネットを使って，世論調査したら・・・？

元気そうなウシばかりサンプリングしたら？

愛煙家の夫の副流煙を受けた妻の例

無作為（ランダム）なサンプリングであることが要請される

無作為標本（ランダムサンプル）ならば

標本数が少数であっても母集団について代表値，精度などを統計的に推定できる．

D．生物統計学・実験計画学とは

大数の法則に従って，データをたくさん取ればより正確なことがわかるとはいえ，実験を何度も繰り返すことはたいへん手間がかかる．したがって，なるべく少ないサンプルで結論を得たい．しかし，サンプルが少ないと誤差が大きくなる・・・

１．どのようにデータを取れば，精度よく，しかも少ない実験ですむのか？

不均一なところ（圃場，牛の集団）からどのように実験し，データを得るのか？

① 圃場によって地力が違うので，同じ品種でも収量が異なる．

② ヒトに栄養剤を与えた効果を知りたいが，体の大きさによって効果が違いそうだ．

不均一さの克服のためには　１．精度を知る　どの程度ばらついているのか？

２．傾向のある不均一さを偶然誤差に転化する

３．可能ならば不均一さによる誤差を除去して，精度を高める

２．生物統計学・実験計画学の３つの柱

① 誤差の定量と制御（精度がよくなる・精度がわかる）

② 実験回数を少なくできる

③ 実験のデータの変動のうち，処理による意味のある部分と偶然誤差を分けることができる．さらに処理の主効果と交互作用を検出できる．

生物統計学予定表

１．　４月１０日　生物統計学とは？
２．　４月１７日　平均と分散・データの要約
３．　４月２４日　母集団と標本，確率分布
４．　５月　１日　二項分布，ポアソン分布，正規分布
５．　５月　８日　正規分布の特徴
６．　５月１５日　統計的推定
７．　５月２２日　統計的検定
８．　５月２９日　ｔ分布とｔ検定
９．　６月　５日　カイ二乗分布・Ｆ分布
１０．　６月１２日　分散分析その１　一元配置
１１．　６月１９日　分散分析その２　二元配置
１２．　６月２６日　分散分析その３　交互作用の解析
１３．　７月　３日　相関分析
１４．　７月１６日　単回帰分析その１
１５．　７月１６日　単回帰分析その２・頻度データの分析
１６．　７月２４日　期末試験

７月１０日は学会のため，７月３１日は実験のため出張するので補講します.