Rを利用した統計解析入門

和田　哲
（北海道大学水産学部海洋生物科学科）

この資料では、左側にスライド、右側に（つまり、いま皆さんが読んでいる欄に）話し言葉での解説があります。スライドを眺めながら解説を読めば、まるで授業を受けているかのような感じで勉強することができるでしょう（そうだったらよいなあと願っています）。

問題：研究初心者が研究を始めるとき、統計解析の方法を真剣に検討するべきタイミングは以下のうち、どれでしょうか？

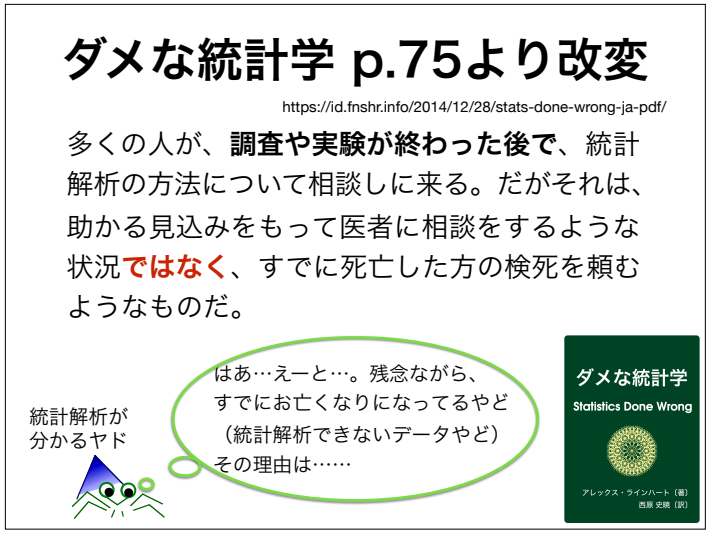
1. **調査、実験前**：**まだデータがないとき**
2. **調査、実験中**：**データをとっているとき**
3. **調査、実験後**：**データをとりおえたあと**

勉強を始める前に、いきなりですが問題です。

皆さんが、これから、統計解析が必要な研究、例えば、動物の行動や生態について調査や実験をする研究を始めるとします。「自分のデータでどのような統計解析を使えばよいのか」など、統計解析の方法について検討するべきなのは、１から３のうち、どのタイミングだと思いますか。ひとつ選んでみてください。

正解は、「調査、実験前」です。じつは統計解析は調査方法や実験方法に強く左右されます。そして**「統計解析について何も考えずに調査したり実験したりすると、統計解析できないデータをとってしまう可能性が高い」**からです。だから、調査、実験前の検討が必要です。もちろん、調査・実験の最中やその後も検討する必要はありますが。

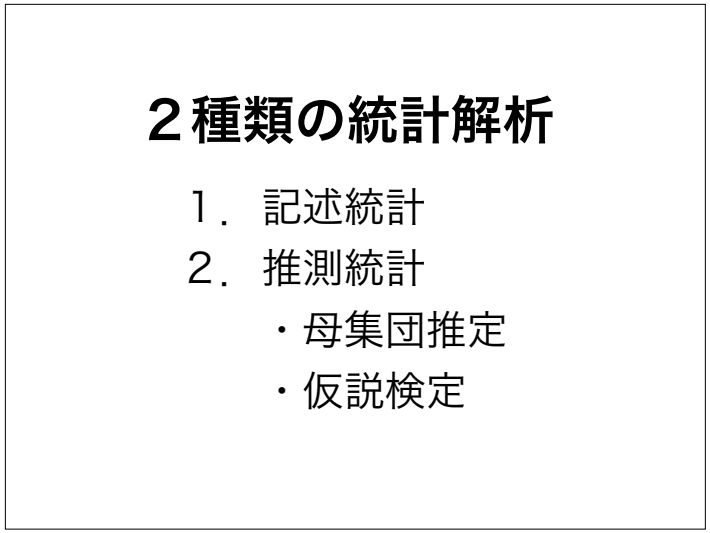
統計解析が必要な学問分野では、統計解析できないデータは「予備観察のデータ、個人的な知見のための非公式データ」としかみなされません。統計解析できないデータは「公表できないデータ」にしかならないのです。



でも、実際には、調査や実験が終わった後で、はじめて統計解析について悩み始めて、統計解析に詳しい人に「なんとかなりませんか？」と相談する人が多いです。「ダメな統計学」という本の著者が述べているように、結論は、左に記したとおりです。「ダメな統計学」は書籍にもなっていますが、書籍のもととなった和訳の文章を、ネットからダウンロードできます。

https://id.fnshr.info/2014/12/28/stats-done-wrong-ja-pdf/

　この「ダメな統計学」には、ほかにもお役立ち情報（というか、現在の「統計学に対する典型的な誤解」など）が満載なので、ぜひ読んでみてください。

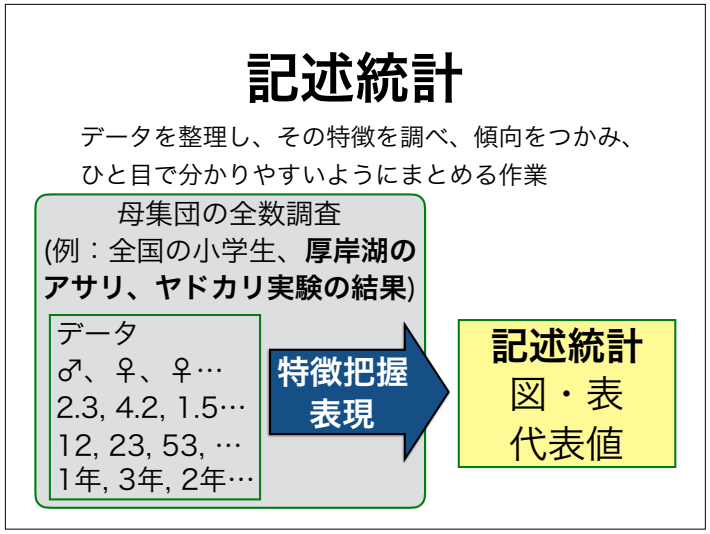


さて、まだ「自分のデータ」をとり始めていない皆さんも、「今のうちに統計解析について勉強しよう」というモチベーションが高まったことと思います。それでは、始めましょう！

　まず、一番の基本です。統計解析は記述統計と推測統計に大別できます。そして、さらに推測統計には、母集団推定と仮説検定という２つの方法があります。

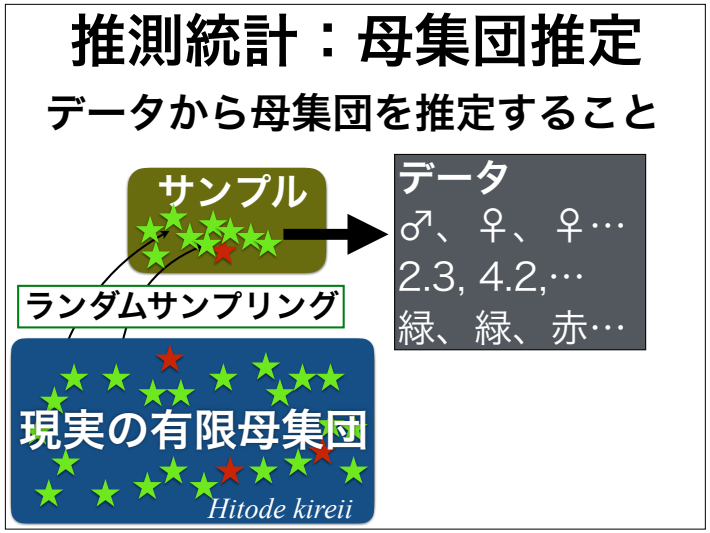
　論文を読んだことがある人や、学会発表を聞いたことがある人は、たとえば「オスとメスで比較したところ、ｔ検定で有意差が認められた」などという表現を聞いたことがあるかもしれません。これらは統計解析のうちの仮説検定の結果を説明しています。なんだかカッコいい？　でも、これだけで「意味が分かった」人はいるでしょうか。

　いませんね。統計に詳しい人でも、上記の表現だけでは「で、どんな性差があったの？」と尋ね直すことになります。この表現は、大事なことを伝えていないからです。仮説検定だけでは、統計解析としては不十分なのです。母集団推定もしなければならないし、記述統計も必要です。「ｔ検定がなにかも分からないよ！」という声も聞こえてきそうですが、仮説検定を勉強するのは最後です。順番に勉強していきましょう。



記述統計の「記述」とは、**データを記述する**という意味です。つまり**記述統計とはデータを表現する手法**のことです。「それなら、記述統計はいつも使ってるよ。平均を計算したり、データを図示したりすることだよね」と考えた人は、きっとたくさんいると思います。正解です。

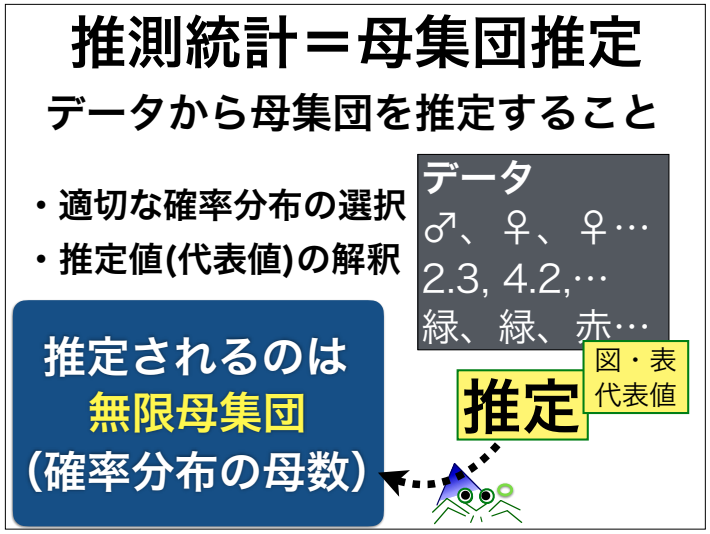
たしかに記述統計は、そのような作業のことです。でも、じつは、記述統計は、たいてい推測統計とセットになって用いられていて、そして「推測統計で不要な記述」は、すべきではありません。例えば、データの平均を計算することが不要な場合もあります。そのような場合に平均を計算して提示べきではありません。それは「**記述統計においては適切でも、推測統計においては不適切**」だからです。「そんなこと言われたって、推測統計ってなんのことか分からないよ」と、皆さんは思ったことでしょう。そこで、推測統計について説明しましょう。



母集団推定のためには、現実の母集団から**ランダムに**採集した標本（**サンプル**）が必要です。このサンプルを観察したり測定するなどして得られた文字や数値の羅列が**データ**です。もちろん、温度のように、現実の母集団から直接データをとることもあります。データに**ばらつき**がないと考えられるときは、データは１つで十分であり、母集団推定は不要です。何度データをとっても結果が同じなのですから。一方、**データが多少なりともばらつく**（つまり、母集団から複数のデータをとったとき、データごとに文字や数値が異なる可能性がある）と考えられるときは、複数のデータをとって、**そのデータのもととなる母集団はどのような集団なのか**推定することになります。これが母集団推定です。

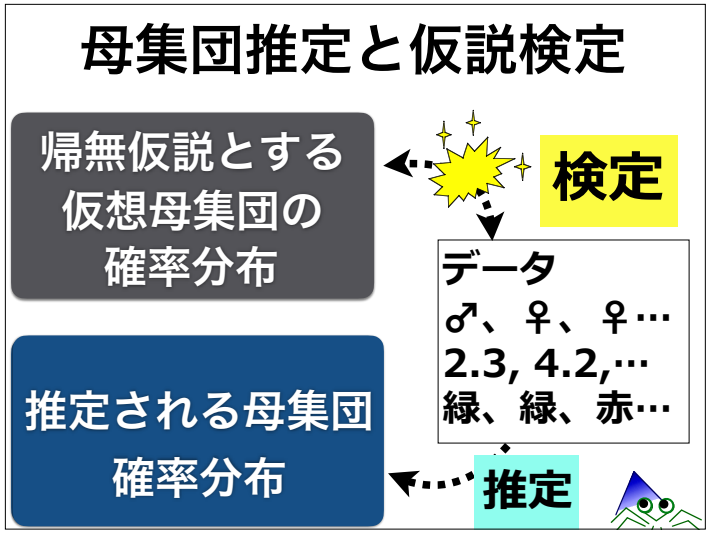
これだけでは、どのように推定すればよいのか、さっぱり分からないと思いますが、たぶん「**正確な母集団推定のために、データは多ければ多いほど良い**」ということは、なんとなく分かると思います。極端な話、母集団全部からデータをとれば、母集団推定しなくても良くなります。例えば、202X年の水産学部２年生の母集団における、性比（男女比）や、道内出身者と道外出身者の比率、カレーライスが好きな人の割合、視力、身長、通学時間などは、全員からデータをとれば、推定する必要はありません。

でも、科学の分野では、たいていそうはいきません。データはばらつく、けれど全てのデータを取り切ることは不可能、だから推定しなければならない、というわけです。さて、どうやって推定するのでしょうか？



ここで注意しなければならないのは、統計解析の基本では、**母集団推定は現実の母集団を推定するわけではない**、という点です。現実の母集団が有限母集団、つまり有限個のデータにすることができる母集団であっても、**実際に推定するのは、無限母集団、つまり無限個のデータで構成されていると想定された確率分布**です。確率分布は、母数というものによって、ある１つの分布に特定されます。そのため、母集団推定とは、この確率分布の母数を推定することになります。これが「統計解析が分からない」と多くの人が考える原因となる最初の（高い）ハードルです。はい、皆さんはきっと、私がなにを言っているのか分からないと思います。ですから、次回以降に詳しく説明していきますね。ただ、いまは「どうやら確率分布ってのが大事らしい」と思って、確率分布について勉強するぞと心の準備をしていてください

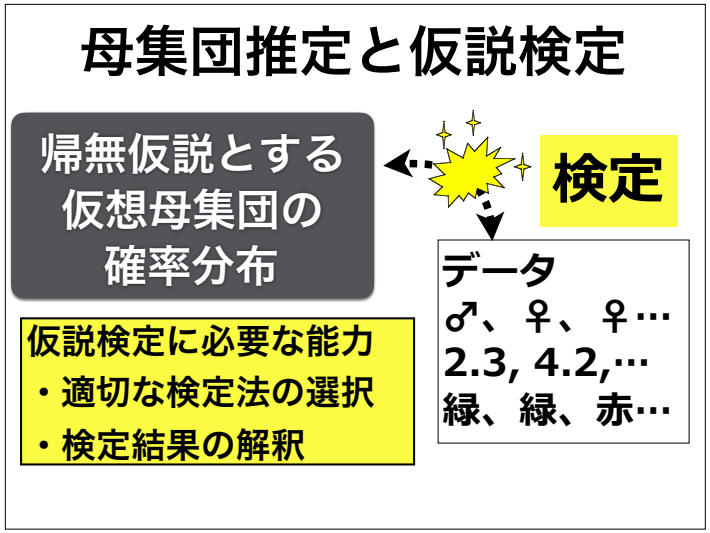
　次に、推測統計のもう１つの柱で、しかも多くの人が注目している（けれども、分かっていない人が多い）仮説検定について、おおまかに説明します。



まずは、母集団推定と仮説検定が、だいぶ違うものであることを強調させてください。ひょっとすると、少し統計学について学んだり、論文を読んだことのある皆さんは、t 検定とか分散分析（ANOVA）とかで「有意差がある」ということを言うのが統計解析だ、と思っているかもしれません。「有意差が出れば、研究は成功したってことだ」と思っている人もいるかもしれません。この t 検定とか分散分析などは仮説検定の一部です。そして、これらは母集団推定ではありません。

　仮説検定は、帰無仮説と呼ばれる仮想母集団の確率分布を想定して、「データがその確率分布から得られたものだとは考えにくい」ということを主張する推測統計の手法です。**統計解析の結果が「有意」であるとは、「その結果が意味の有る結果だ」という意味ではなく、「帰無仮説が正しくないように思うんだけど、、、」という意味**なのです。「〇〇検定で有意な結果が得られた」などと書かれていると、まるで「判定勝ちした！」とか「実験は成功だ！」などと感じるかもしれませんが、そういう意味ではないのです。

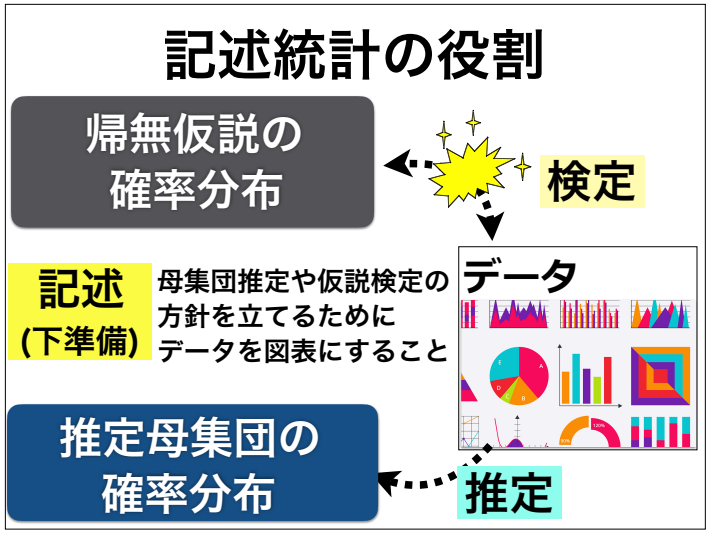
　仮説検定は、データをとってきた母集団については、なにも想定していません。そのため、t 検定とか分散分析などの結果を示す人は、同時に母集団推定の結果も示さなければ、データを見たことがない人には「その有意差って、どういうことなのか」が分からないことになります。多くの論文では、その母集団推定の結果を、文章や図で伝えています。でも、それが母集団推定の結果だと分かっていない人もたくさんいます。



仮説検定は、データをとってきた母集団については何も想定していませんが、だからといって、もとの母集団を無視した検定をおこなうことは、結果を間違える原因となります。帰無仮説の確率分布が、もとの母集団の確率分布に応じたものになっている必要があるのです。

なるほど、母集団推定と同じように確率分布をちゃんと想定すればいいってことだね、と考えた人は、正しいのですが、それでもちょっと間違えています。じつは、例えば t 検定は、たしかに t 分布という確率分布を利用した検定なのですが、実際の母集団が t 分布となることは（たぶん必ず）ないのです。なぜかというと、t 分布は仮説検定用の確率分布だからです。

このように、確率分布とひとくちに言っても、そのなかに母集団推定用の確率分布と仮説検定用の確率分布があります（統計学の入門書を読んでも、このように説明している本は見当たらないのですが、、、）。これが「統計解析が分からない」と多くの人が考える原因となる２つ目の（高い）ハードルです。いまは、なにを言っているのか分からないと思いますが、先ほどと同じように「やっぱり確率分布ってやつが重要らしい」と思っておいてください。



推測統計の概要について話したので、記述統計に戻りましょう。記述統計は推測統計を支える重要なものです。私たちは、データを手に入れたら、推測統計の方針を決めるために、データの記述をおこないます。記述統計は、推測統計の下準備になるわけですが、それだけではありません。

実際におこなった母集団推定や仮説検定の手法が適切であるかどうか、そのデータのことを知らない人にも判断できるように、データを記述する必要もあります。つまり、私たちは、データの統計解析の結果を発表するためにもデータの記述をおこなう必要もあるのです。

では、記述統計について、もっと詳しく説明していきましょう。