習うより、慣れよう！ (R による統計処理)

はじめに

私は、 R による統計処理の仕方について話したいと思います。多くの４年生は統計処理をしたことがないと思うので、まずは統計の話から始めていきたいと思います。それから、R におけるデータの扱い方、検定のかけ方、検定結果の見方等について説明していきます。

R は優れたフリーソフトウェアなので、これから研究活動を進める上で R を使えると非常に便利だと思います。まだ R を使ったことのない人にとっては、難しく感じるかもしれませんが、今日の講習会を通して、少しでも多くの人が「これから R を使ってみようかな」と思ってくれたらな…と思います。

◆内容

なぜ検定をかけるのか？

- Example 1. 北海道の人と沖縄の人では、どっちが高身長？
- Example 2. ダイエット薬の違いが体重に与える影響は？
- Example 3. 標高から気温を求めるには・・・？

実際に検定をかけてみよう！

- Case 1. カミキリムシの幼虫とゾウムシの幼虫では、どっちが重い？
- Case 2. 寄主の違いが寄生蜂の生存率に及ぼす影響は？
- Case 3. 寄主生重から寄生蜂の長翅の長さ (体サイズ) を求めるには・・・？

なぜ検定をかけるのか？

Example 1. 北海道の人と沖縄の人では、どっちが高身長？

北海道の２０代の男性１００人と沖縄の２０代の男性１００人に協力してもらい、身長を記録した。

北海道の人：１６６ｃｍ、１５８ｃｍ、１７２ｃｍ・・・・・・１７７ｃｍ
沖縄の人　：１７２ｃｍ、１６８ｃｍ、１８０ｃｍ・・・・・・１５５ｃｍ

どっちの背が高い？

ここで、北海道の１００人の平均身長が１７０ｃｍで、沖縄の１００人の平均身長が１６０ｃｍだったとしよう！

じゃあ、北海道の人の方が沖縄の人より背が高いんだねぇ！・・・・・・果たして本当にそうだろうか？

もう一度１００人選んでそれぞれの平均身長を出したら、北海道が１５０ｃｍで、沖縄が１９０ｃｍになる可能性もあるのでは？

どっちの平均身長が高いかは、北海道と沖縄の２０代の男性の全員の身長を記録し、比較する必要がある。
・・・しかし、それは、大変だぁ・・・・・・だから統計処理を行うのさ！

統計処理とは、
本当にその結果（北海道の人の方が沖縄の人より背が高いという結果）が確率的に、正しいか否かを決定付ける処理である。

ここで 2 群間の関係を明らかにする検定として、 t 検定や Mann – Whitney のU検定（Wilcoxon の順位和検定) が挙げられる。

どっちを使ったらいいの？
t 検定・・・・・・・・・・・・パラメトリック検定
Mann – Whitney のU検定・・・ノンパラメトリック検定

母集団が正規分布を仮定できる場合は、パラメトリック検定を行うことができる。
母集団の分布型を一切仮定しない場合は、ノンパラメトリック検定を行う。

Example 2.　ダイエット薬の違いが体重に与える影響は？

ダイエット薬 A と B があるとする。どちらか片方を 1 カ月間使い続けた人について以下のような結果を得られたとしよう！

A 薬を使って痩せた人　７８人、　痩せなかった人　５６人
B 薬を使って痩せた人　５５人、　痩せなかった人　１４人

分割表 (クロス集計表)

	痩せた	痩せなかった
A 薬	７８	５６
B 薬	５５	１４

どっちが効果的かをどのように判断する？
う～ん・・・A 薬で痩せた人が７８人で一番多いから A 薬が良く効くのかなぁ？

おいおいおい！確かに痩せた人は B 薬より A 薬を使った人の方が多いけど､痩せなかった人を A 薬と B 薬で比べると B 薬の方が断然少なくね？
どうやって比べようか・・・？？？

・・・こんな時は・・・比率を比べるのだ！！！

A 薬を使った人の合計は　７８ ＋ ５６ ＝ １３４
B 薬を使った人の合計は　５５ ＋ １４ ＝　６９

A 薬で痩せた人の割合は　７８ ÷ １３４ × １００ ＝ ５８．２%
B 薬で痩せた人の割合は　５５ ÷　６９ × １００ ＝ ７９．７%

B 薬の方が効きそうだ！・・・・。ほんとにそう言える？
ほんとにそうだと言いきるために検定をかけるのさ！

比率の差の検定 (比率に差があるかどうかの検定)
・・・χ^2 検定がよく用いられる。

Example 3.　標高から気温を求めるには・・・？

中島くんは富士山の頂上の気温を知りたいと思い、富士山を登り始めましたが、１０００m登ったところで力尽きてしまいました。どうしても山頂の気温を知りたい中島くんは、１０００m地点からおよそ標高５０m降りては気温を記録し、「この結果から山頂の気温を知ることができないだろうか？」と考えました・・・。

標高	５０m	１００m	１５０m	・・・・・・	１０００m
気温	２７℃	２５℃	２４℃	・・・・・・	２０℃

なんとかこのデータから富士山の頂上の気温を知ることができないかな～?
よし！まずグラフを書いてみよう！
標高が高くなると気温は下がっているなぁ・・・良い感じに相関がありそうだぁ・・・
ここに回帰直線を引けば頂上の気温を求められそうだ！！！

よし回帰分析をしよう！
う～ん、でも相関分析ってのもあったような・・・う～ん・・・う～ん・・・
標高と気温は相関があるから相関分析？？？

＜相関分析と回帰分析について＞
相関分析と回帰分析は同じ？違う？？・・・どっちだろう？？？

結論から言うと、相関分析と回帰分析は全く別物なんだ！

- 相関分析は 2 変数の間に線形関係があるかどうか、およびその強さについての分析。
  → x と y が同等の関係（x − y）
- 回帰分析は、独立変数 (説明変数) から従属変数 (目的変数) を求めるもの。
  → x が決まれば y が決まるという関係（x → y）

Example 3. について言えば、方向性を考えずに、標高と気温の間に関係があるかどうかを調べるのが相関であり、x を標高、y を気温としたとき、標高から気温を推定できないかと考える (気温から標高を推定することは考えない) のが回帰です。

すなわち、2 変数の関係を知りたいだけなら相関分析を行えばよくて、一方の変数 (x) の値から他方の変数 (y) の値を予測したいのなら回帰分析を行う！

実際に検定をかけてみよう！

まずは下準備をしよう！

・拡張子の表示

○Windows XP: フォルダ ＞ ツール ＞ フォルダオプション ＞ 表示 ＞ 詳細設定の「登録されている拡張子は表示しない」のチェックを外す。

○Windows Vista: コントロールパネル ＞ フォルダオプション ＞ 表示 ＞ 詳細設定の「登録されている拡張子は表示しない」のチェックを外す。

○Mac OS X: Finder環境設定… ＞ 詳細の「すべての拡張子を表示」をチェック。

・データの保存・Rへの読み込み

R講習会のフォルダの中にある『data.xls』を開く。

ファイル ＞ 名前を付けて保存 ＞ その他の形式 ＞ ファイルの種類を選択 ＞ テキスト (タブ区切り) (*.txt) の形式で保存。

→フォルダの中に data.txt というファイルが出てきたか確認する。

```
setwd("場所")                    # 本日用いるデータを R に取り込む
d <- read.table("data.txt" , header = T)          # d という名前にデータを入れる
d

# 時間が余った人はいろいろ試して下さい
str(d)
names(d)
summary(d)
plot(d)
```

Case 1.　カミキリムシの幼虫とゾウムシの幼虫では、どっちが重い？

Case 1. は Example 1.～3. のどれに近いだろう…?

言わずもがな、Example 1. ですね・・・ほんとにそう？

データを見て確認しましょう。

```
d1 <- d[　,4:5]　　　　# まず解析に用いる部分を抜き出す
d1
```

データからも Example 1. に近いことがわかりますね。

２群間の関係を明らかにする検定をかけよう！

では、 t 検定と Mann – Whitney のＵ検定…どっちを使ったらいいの？？？

→ 本データの母集団が正規分布を仮定できるかどうか (パラ or ノンパラ) を調べる必要がある。

→ 本データが正規分布するかどうか調べる必要がある。

●Kolmogorov – Smirnov (コロモゴロフ・スミノフ) 検定

正規性の検定である。 R では、頭文字をとって ks.test() という名前の関数が用意されている。この検定の帰無仮説は、「あるデータが、正規分布をなす」である。

p 値が有意水準より大きければ正規分布 → パラメトリック → t 検定を行うことができる！

p 値が有意水準以下なら非正規分布 → ノンパラメトリック → Mann – Whitney のＵ検定を行う！

d1 にはカミキリムシとゾウムシのデータが合わさっているので分けましょう！

```
kamikiri <- d1[d1$host.species == "kamikiri",　]
kamikiri
zou <- d1[d1$host.species == "zou",　]
zou
```

正規分布か否かを視覚的にとらえるためにヒストグラムを書きましょう！

```
par(mfrow = c(2,1))　　# 1 つのグラフィックデバイスを上下 2 つに分割
hist(kamikiri$host.weight)
hist(zou$host.weight)
```

正規性の検定をかけよう！

```
ks.test(kamikiri$host.weight, "pnorm", mean = mean(kamikiri$host.weight),
        sd = sd(kamikiri$host.weight))
ks.test(zou$host.weight, "pnorm", mean = mean(zou$host.weight),
        sd = sd(zou$host.weight))
```

タイがあるため、正しい p 値を計算することができません。
※上の警告メッセージは正確ではない。
→「正確な p 値を計算できない」ということであって、誤った p 値を計算しているわけではない！

正規性はあったかな？…なければ、ここでMann – WhitneyのU検定（Wilcoxonの順位和検定）を行う！

正規性があったら次のステップだぁ！
正規性が確認できれば、パラメトリック検定、すなわち、この場合は t 検定を行うことができる！ただ、t 検定には、Studentの t 検定とWelchの t 検定の２種類がある！
どちらにしたらいいのかは、等分散性の検定を行う必要があり、等分散であれば、Studentの t 検定、不等分散であれば、Welchの t 検定を行う。

● F 検定
等分散性の検定である。R では、var.test() という関数が用意されている。帰無仮説は「２群の母分散は等しい」である。
p 値が有意水準より大きければ２群は等分散 → Student の t 検定
p 値が有意水準以下なら２群は不等分散 → Welch の t 検定

等分散性の確認

```
var.test(kamikiri$host.weight, zou$host.weight)
```

等分散性はあっただろうか？？？
あれば、Student の t 検定、なければ、Welch の t 検定！

これでどの検定を行えばよいか決定です！！！
今回は・・・Welch の t 検定でした！

● t 検定

平均値の差の検定である（平均値に差があるかどうか）。R では、t.test() という関数が用意されている。帰無仮説は「二群の母平均は等しい」である。

```
t.test(kamikiri$host.weight, zou$host.weight, var.equal = F)　　#　等分散の場合は F を T に変える
```

みなさん検定結果はでましたか？？？

●Wilcoxon の順位和検定（Mann – Whitney のU検定)

正規性が仮定できなかった場合の２群比較の検定として、Rでは wilcox.test()（Wilcoxon の順位和検定）が用意されている。帰無仮説は「２群が同じ母集団から抽出された」である。

まとめ

正規性、等分散性の検定を行ったうえで、適切な検定方法を使用する。

① 正規性の検定：ks.test(　, “pnorm”, mean = mean(), sd = sd())
　　#　平均値 mean( )，標準偏差 sd() の正規分布か？
　　ks.test(scale()　, “pnorm”)

② 正規性がある → 等分散性の検定：var.test()

③ 正規性があり、等分散である → Student の t 検定：t.test(　, var.equal = T)

④ 正規性があり、等分散でない → Welch の t 検定：t.test(　, var.equal = F)

⑤ 正規性がない → Wilcoxon の順位和検定（Mann – Whitney のU検定)：wilcox.test()

Case 2. 寄主の違いが寄生蜂の生存率に及ぼす影響は？

Case 2. は Example 1.～3. のどれに近いだろう…?

う～ん、Example 2. かな～。

```
d2 <- d[  ,c(4, 6)]          # まず解析に用いる部分を抜き出す
d2
(d2 <- na.omit(d2)           # NA (欠損値) の入った行を除く)
```

このままのデータだと分かりづらいな～。

データを分割表の形にしよう！

```
table(d2)
t(table(d2))                 # 縦横を入れ替える (やってもやらなくてもどっちでもよい)
```

やっぱり Example 2. だぁ！

このような表の形にしないとχ^2検定を実行することができないので上記の table() は重要な作業である！

●χ^2検定

比率の差の検定である。R では、chisq.test() という関数が用意されている。帰無仮説は「比率に差がない」である。

```
chisq.test(table(d2), correct = FALSE)
# または
chisq.test(t(table(d2)), correct = FALSE)
# データに１０以下の数があるときは、少数例のためにイェーツ補正を行う (correct = TURE)
```

みなさん検定結果はでましたか？？？

●フィッシャーの正確確率検定

fisher.test()

２×２分割表の２変数の間に統計学的に有意な関係があるかどうかを検討するのに用いられる。１×２分割表の場合もある。同じ状況でサンプルサイズが大きい場合には、統計量の標本分布が近似的にχ^2分布に等しくなるのでχ^2検定が用いられるが、サンプルサイズが小さい (分割表のセルの期待値に１０未満のものがある) 場合や、表中の数値の偏りが大きい場合にはこの近似は不正確である。この場合には正確確率検定が文字通りに正確である。

Case 3.　寄主生重から寄生蜂の長翅の長さ (体サイズ) を求めるには・・・？

最後は、もうお分かりですね。そう Example 3. の実践編です。

```
plot(d)                          # 最初にこの命令で全データの散布図を見ておくとよいでしょう！
d3 <- d[   ,c(5, 8)]             # まず解析に用いる部分を抜き出す
d3
(d3 <- na.omit(d3)               # NA (欠損値) の入った行を除く)
d3
plot(d3)                         # 散布図を描いてみる
plot(d3$host.weight, d3$wasp.head)          # x 軸を host.weight、y 軸を wasp.head に指定
```

うん！この 2 変数には正の相関がありそうだ！
回帰分析を行うことで寄主生重 (餌の量) から寄生蜂の長翅の長さ (体サイズ) を推定できる！

●回帰分析
R では lm() という関数が用意されている。帰無仮説は「回帰直線の傾きが 0 (つまり、y は x に依存しない) 」である。

```
model <- lm(d3$wasp.wing ~ d3$host.weight)
model                            # 回帰係数を求める
summary(model)                   # 回帰分析の結果の詳細
abline(model)                    # 回帰直線を引く
```

みなさん検定結果はでましたか？？？

●相関分析
cor.test(x, y, method = “pearson”)　　…パラメトリック
cor.test(x, y, method = “kendall”)　　…ノンパラメトリック
cor.test(x, y, method = “spearman”)　　…ノンパラメトリック
相関係数：2 変数の散らばりの程度 (相関係数の絶対値が 1 に近いほど散らばりが少ない)
p 値　　：相関があるか否か (帰無仮説「相関係数が 0 」)

最後に

今日、私が紹介した検定方法以外にも、まだまだたくさんの検定方法が存在します。みなさん忙しいとは思いますが、各自データ解析する際に最も適した検定方法を行えるよう統計学の勉強に取り組んでもらいたいです。

また、R に関しては、習うより、慣れましょう！今日 R を使ったのが初めてで、「難しいから嫌だ！」とか、「意外と簡単じゃん！」とか、様々な感想を持たれたかと思います。嫌な人に無理に R を使い続けなさいとは言いませんが、R は、今難しくても、使い続けることで必ず慣れることができるソフトだと思うので、余力があったらもう少し頑張って使ってみてはどうでしょう？

以上で、私の話は終わります。今日はありがとうございました。

「R」を使う人にオススメのサイト

・**http://www.okada.jp.org/RWiki/** （**RjpWiki**）
・**http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html** （**R-Tips**）