環境統計学ぷらす

第1回（全5回？）
Rの基礎と仮説検定

高木　俊
shun.takagi@sci.toho-u.ac.jp

2013/10/24

今回やること

- Rの基礎
- 仮説検定
  - Fisherの正確確率検定
  - 2群の平均値の差の検定（t検定）
- 結果の表し方
  - 図と表
  - 文章中の表現

＊今後Win版を前提に話を進めます

＊次回以降もRの操作練習、統計の解説、論文での表現の3つを軸に話を進めようかと思います

Rの基礎

統計解析環境　R

- Rとは
  統計計算とグラフィックスのための言語・環境

- Rの特徴
  フリーソフト
  オープンソース（だれでも開発できる）のソフトウェア
  豊富な拡張パッケージ

- Rを使うには・・・
  R言語を覚える必要がある
  ＊R Commander（R cmdr）を使えば基本機能をGUIで使うこともできます

Rの導入

- 省略します
- 現在の最新版（多分）はR 3.0.2　(2013/09/25リリース)
- バージョン間で操作はそれほど変わらないはず（某Officeと違って）

→　使ってみましょう！

基本演算

		入力コマンド	出力結果
• +, -, *, /	加減乗除	(1+2*5)/2-0.5	5
• ^	累乗	3^2	9
• sqrt()	二乗根	sqrt(9)	3
• abs()	絶対値	abs(-8)	8
• exp()	自然対数の累乗	exp(1)	2.7182…
• log()	自然対数	log(2.718282)	1
• sin()	正弦関数	sin(pi/2)	1
• asin()	逆正弦関数	asin(1)	1.5707… (=pi/2)

オブジェクトと代入

```
a<- 2; b<- 3
a
b
a+b^2
A+b^2
a<- “enveco”
a
```

```
2
3
11
エラー:　オブジェクト ‘A’ がありません
“enveco”
```

;は同一行内にコマンドを続けて書く場合使う

オブジェクト同士の計算

大文字と小文字は区別される

オブジェクトには文字列も代入可

オブジェクトは上書きされる

```
a<- 2; b<- 3
a<- a^b
a
```

で何と出力されるか？

ベクトル

```
c(1,2,3,4,5)                1 2 3 4 5
c(1:5)                      1 2 3 4 5
c(8:3)                      8 7 6 5 4 3
rep(2,3)      #2を3回繰り返し   2 2 2
seq(2,8,3)    #2から8まで3おき   2 5 8
```

- ベクトル要素へのアクセス

```
a<- c(7,6,4,0)
a[2]                  #aの2番目   6
a[c(4,2)]     #aの4番目と2番目   0 6
```

a[a[3]]　は何と出力されるか？

ベクトル用の関数

```
a<- c(7,6,4,0,2,7,4)
sum(a)                  #和          30
mean(a)                 #平均        4.285714
sd(a)                   #標準偏差    2.627691
length(a)               #要素の数    7
max(a);min(a)           #最大･最小   7 ; 0
median(a)               #中央値      4
quantile(a)             #四分位数
  0%  25%  50%  75% 100%
 0.0  3.0  4.0  6.5  7.0
```

標準誤差 $SD/\sqrt{n}$ はどのように表すか？

行列

#行列の生成（左の列から順に埋められる）

```
mat<- matrix(1:6,nrow=2,ncol=3)
```

```
     [,1][,2][,3]
[1,]    1    3    5
[2,]    2    4    6
```

#足りない要素は繰り返し

```
matrix(1:3,nrow=2,ncol=3)
```

```
     1    3    2
     2    1    3
```

#byrow=Tで行優先で生成

```
matrix(1:6,nrow=2,ncol=3,byrow=T)
```

```
     1    2    3
     4    5    6
```

#要素へのアクセス

```
mat[,3]
```

```
5 6
```

```
mat[1,]
```

```
1 3 5
```

```
mat[2,3]
```

```
6
```

mat[2,2:1]
は何と出力されるか？

#その他行列用関数　nrow(), ncol() など

困ったときには

- Rのhelp

  help(関数名) または ?関数名 で呼び出せる

  ?mean と入力するとmean関数の説明がhtmlファイルで読める（英語）

  その関数の使い方や使い方の例が書かれている

- 役立つホームページ

  R-Tips

  http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html

  RjpWiki

  http://www.okada.jp.org/RWiki/

仮説検定

仮説検定Hypothetical testとは

- 設定した仮説が正しいと入ってよいかどうかを統計学的・確率論的に判断するための方法

仮説検定の手順

1. 対立仮説H_1および帰無仮説H_0を設定する
2. 検定統計量を設定し、データから検定統計量を計算する (もしくは事象の生起確率を直接計算する)
3. 計算した統計量の値よりも極端な値が、帰無仮説が正しいと仮定したときに得られる確率（P値）を求める
4. P値が有意水準よりも小さければ、帰無仮説を棄却する（大きければ棄却しない）

2×2分割表の検定
（Fisher's exact test・フィッシャーの正確確率検定）

		カテゴリー1		
		1	2	計
カテゴリー2	X	a	b	a+b
	Y	c	d	c+d
	計	a+c	b+d	a+b+c+d

- 2 x 2分割表(上記のような表、各セルには観察数が入る)において、カテゴリー間の関係性を見たい場合に用いる

例） 男女間で喫煙する/しないに差があるか

種Aと種Bで生/死に差があるか

実例

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

研究室名	男	女	計
鏡味研	7	1	8
西廣研	3	2	5
計	10	3	13

1. 対立仮説H_1および帰無仮説H_0を設定する

対立仮説H1

男女比は異なる

帰無仮説H0

男女比は異ならない
（鏡味研も西廣研も10:3）

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

2. 検定統計量を設定し、データから検定統計量を計算する
(もしくは事象の生起確率を直接計算する)

研究室名	男	女	計
鏡味研	7	1	8
西廣研	3	2	5
計	10	3	13

グレーの部分（周辺度数）を固定した時、上記の比率の男女比が得られる確率

$$\frac{(8\text{人から男性7人・女性1人}) \times (5\text{人から男性3人・女性2人})}{13\text{人から男性10人・女性3人が選ばれる場合の数}}$$

$$= \frac{{}_8C_1 \times {}_5C_2}{{}_{13}C_3} = 0.28$$

Rで下の式を入れれば計算されます
choose(8,1)*choose(5,2)/choose(13,3)

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

3. 計算した統計量の値よりも極端な値が、帰無仮説が正しいと仮定したときに得られる確率（P値）を求める

	男	女	計
鏡味研	8	0	8
西廣研	2	3	5
計	10	3	13

	男	女	計
鏡味研	7	1	8
西廣研	3	2	5
計	10	3	13

	男	女	計
鏡味研	6	2	8
西廣研	4	1	5
計	10	3	13

	男	女	計
鏡味研	5	3	8
西廣研	5	0	5
計	10	3	13

すべての組み合わせを考える

$$\frac{{}_8C_0 \times {}_5C_2}{{}_{13}C_3} = 0.035$$ より男女比かたよる

$$\frac{{}_8C_1 \times {}_5C_2}{{}_{13}C_3} = 0.28$$ 観察事象

$$\frac{{}_8C_2 \times {}_5C_1}{{}_{13}C_3} = 0.49$$ より均等

$$\frac{{}_8C_3 \times {}_5C_0}{{}_{13}C_3} = 0.196$$ より男女比かたよる

観察事象よりも男女比がかたよる確率（P値）

$$0.28 + 0.035 + 0.196 = 0.511$$

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

4. P値が有意水準よりも小さければ、帰無仮説を棄却する（大きければ棄却しない）

有意水準α＝0.05 < P値＝0.511 なので、
帰無仮説を棄却できない

→対立仮説（男女比が異なる）は採用できない

→結論：男女比は異なるとはいえない
（≒男女比は異ならない）

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

Rでは下記の2行で実行

```
mat<- matrix(c(7,3,1,2),ncol=2)
fisher.test(mat)
```

Fisherの正確確率検定を行う関数

```
 Fisher's Exact Test for Count Data

data:  mat
p-value = 0.5105
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.1564982 312.8805051
sample estimates:
odds ratio
  4.091145
```

練習問題：Fisherの正確確率検定

表1. 地点A・Bにおけるオニビシとヒシの発芽および未発芽種子数

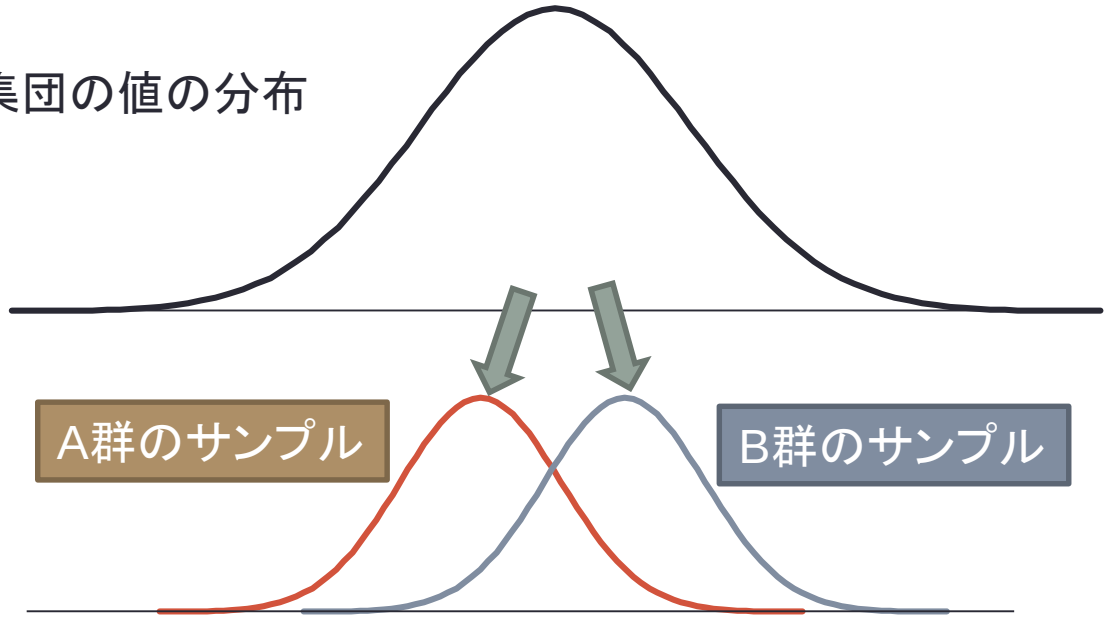
	オニビシ		ヒシ
	地点A	地点B	地点B
発芽	169	129	101
未発芽	222	158	2

1. オニビシの発芽率は地点Aと地点Bで異なるか？
2. 地点Bにおけるオニビシとヒシの発芽率は異なるか？

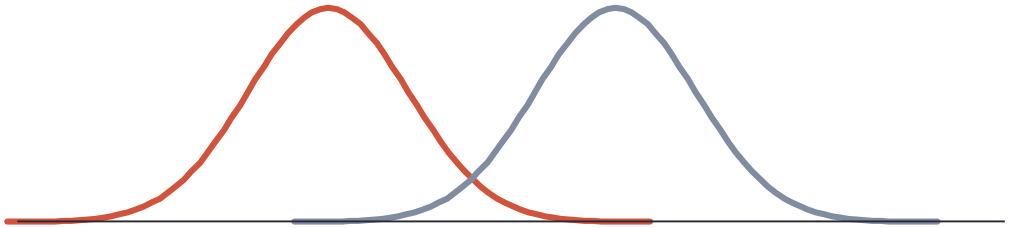
2群の平均値の差の検定
（Student's *t* test・スチューデントのt検定）

- 2群のサンプルが同じ正規母集団から得られたかどうか（平均値が同じ集団から得られたか）を検定

例） 処理Aと処理Bで成長率に差があるか
　　場所Aと場所Bで栄養塩濃度に差があるか



- *t*検定を用いることのできる前提条件



正規性 データ(の母集団)が正規分布に従うこと

等分散性 2群のデータ(の母集団)の分散が等しいこと

独立性 個々のデータは互いに独立であること

正規性が満たされない場合

→ データを変換して正規分布にする

正規分布を仮定しない解析(*U*検定などノンパラメトリック検定)を行う

Rではwilcox.test()で実行

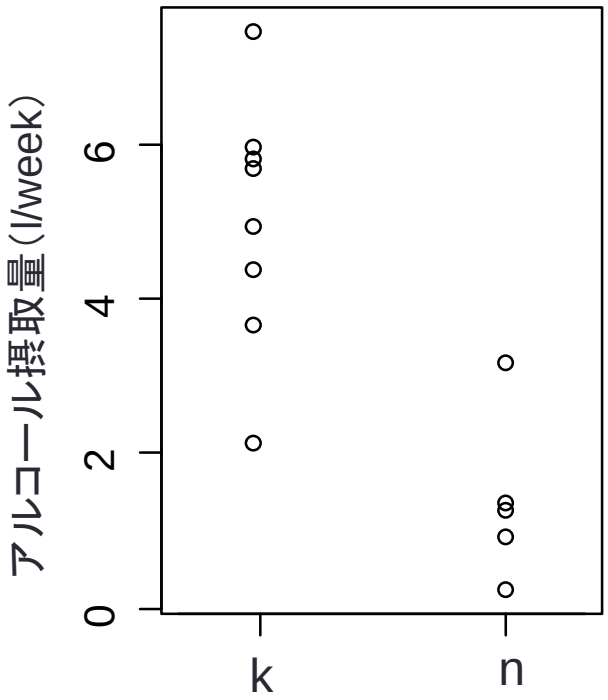
等分散性が満たされない場合

→ Welchの*t*検定を行う

実例

- 研究室間で学生のアルコール消費量は異なるか

二つの研究室（研究室kと研究室n）で1週間あたりの学生のアルコールの摂取量（リットル/週）を比較した。（架空のデータ）




```
k<- c(4.3, 3.6, 5.7, 2.1, 5.9, 5.8, 7.4, 4.9)
n<- c(1.2, 0.2, 1.3, 3.2, 0.9)
```

1. 対立仮説H_1および帰無仮説H_0を設定する

対立仮説

研究室間でアルコール摂取量は異なる

帰無仮説

研究室間でアルコール摂取量は異ならない

- 研究室間で学生のアルコール消費量は異なるか

*t*検定の前に･･･前提条件のチェック

正規性

Shapiro–Wilk 検定

帰無仮説：標本は正規母集団からサンプリングされた

```
> shapiro.test(k)

        Shapiro-Wilk normality test

data:  k
W = 0.9697, p-value = 0.8955
```

```
> shapiro.test(n)

        Shapiro-Wilk normality test

data:  n
W = 0.8789, p-value = 0.3045
```

P>0.05 正規分布でないとは言えない→正規性OK

等分散性

*F*検定

帰無仮説：2標本群は分散の等しい母集団からサンプリングされた

```
> var.test(k,n)

        F test to compare two variances

data:  k and n
F = 2.1329, num df = 7, denom df = 4, p-value = 0.4841
alternative hypothesis: true ratio of variances is not equal to 1
（略）
```

P>0.05 等分散でないとは言えない→等分散性OK

- 研究室間で学生のアルコール消費量は異なるか

2. 検定統計量を設定し、データから検定統計量を計算する

統計量tの計算

$$t = \frac{\text{平均の差}}{\text{差の標準誤差}} = \frac{\overline{y_a} - \overline{y_b}}{s.e._{diff}}$$

→ データの数と分散から計算

$$s.e._{diff} = \sqrt{\frac{(n_A - 1){s^2}_A + (n_B - 1){s^2}_B}{n_A + n_B - 2}} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$$

Welchのt検定の場合 $s.e._{diff} = \sqrt{\frac{{s^2}_A}{n_A} + \frac{{s^2}_B}{n_B}}$

この架空データの場合、

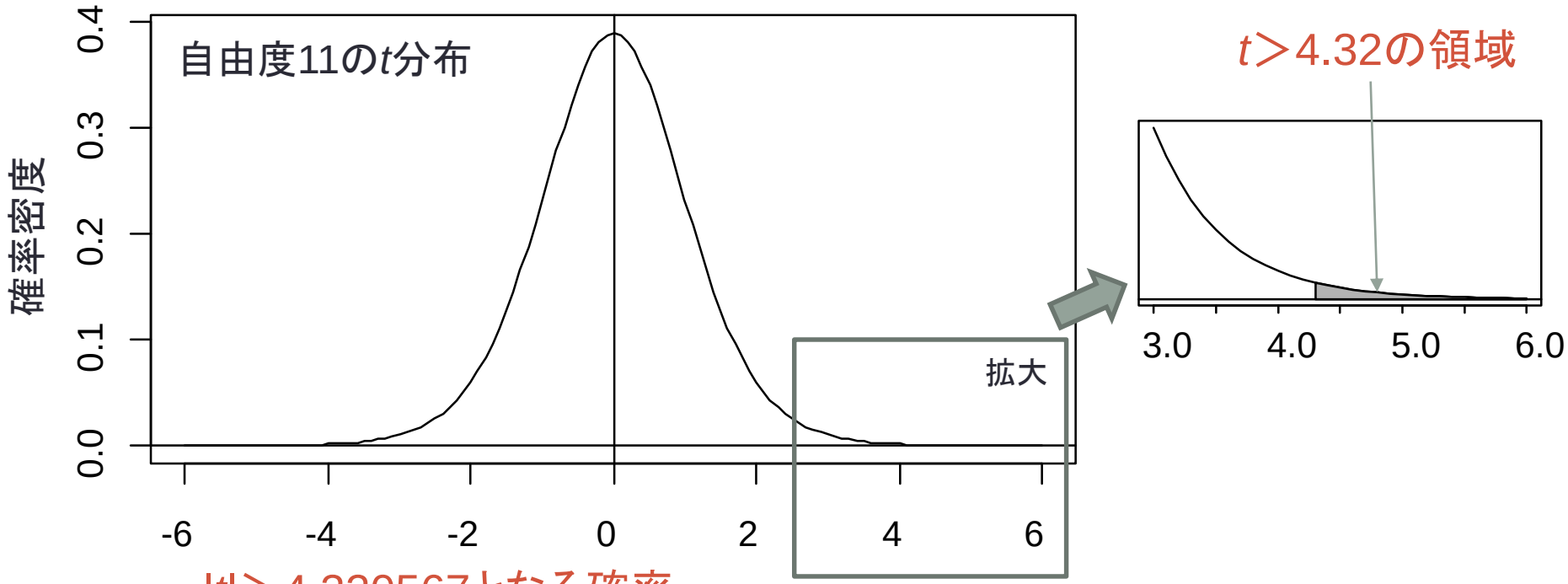
$$t = 4.320567$$

Rで下の式を入れれば計算されます

```
(mean(k)-mean(n))*sqrt(length(k)+length(n)-2)/
(sqrt((length(k)-1)*var(k)+(length(n)-1)*var(n))*sqrt(1/length(k)+1/length(n)))
```

帰無仮説が正しい時、t値は自由度$n_A + n_B - 2$のt分布に従う
このt値が極端な値であれば、帰無仮説は正しくないといえる

- 研究室間で学生のアルコール消費量は異なるか
  3. 計算した統計量の値よりも極端な値が、帰無仮説が正しいと仮定したときに得られる確率（P値）を求める
  4. P値が有意水準よりも小さければ、帰無仮説を棄却する



|t|＞4.320567となる確率

P＝0.001213521 ＜ 0.05

Rで下の式を入れれば計算されます
`(1-pt(4.320567,11))*2`

帰無仮説は棄却。研究室間でアルコール消費量は等しいとは言えない

（≒研究室kの学生は研究室nの学生よりよく飲む）

Rで*t*検定

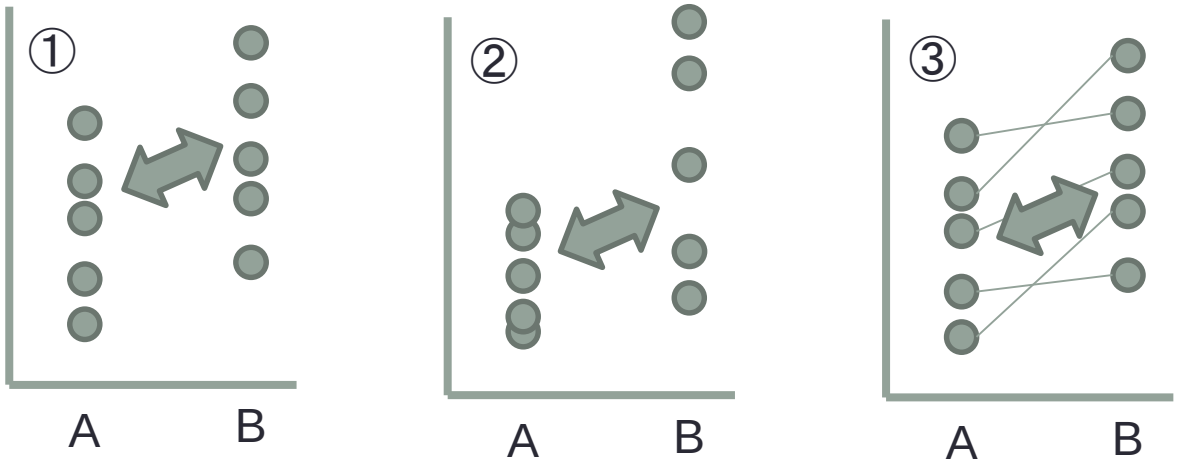
```
k<- c(4.3, 3.6, 5.7, 2.1, 5.9, 5.8, 7.4, 4.9)      #データk
n<- c(1.2, 0.2, 1.3, 3.2, 0.9)                      #データ
t.test(k,n, var.equal=T)                    #var.equal=Tで等分散仮定

 Two Sample t-test

data:  k and n
t = 4.3206, df = 11, p-value = 0.001214
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.767313 5.437687
sample estimates:
mean of x mean of y
   4.9625    1.3600
```

t.test のオプション

① `t.test(…, var.equal=T)` #Studentのt検定
`Two Sample t-test`

② `t.test(…, var.equal=F)` #Welchのt検定（デフォルト）
`Welch Two Sample t-test`
→等分散でない場合

③ `t.test(…, paired=T)` #対応のあるt検定
`Paired t-test`
→対応があるデータを比較する場合
例）同じ人の反応を処理前後で比較
同じ地点での表層と低層の比較



練習問題：*t*検定

表2. 各調査地点における7月と9月のChl-a蛍光値

Site id	7月	9月
1	5307	1205
2	3932	1340
3	4875	2179
4	3051	1217
5	3552	1902
6	607	1535
7	2098	1388
8	1376	2001
9	522	2733
10	4687	1871

（※データ改変しています）

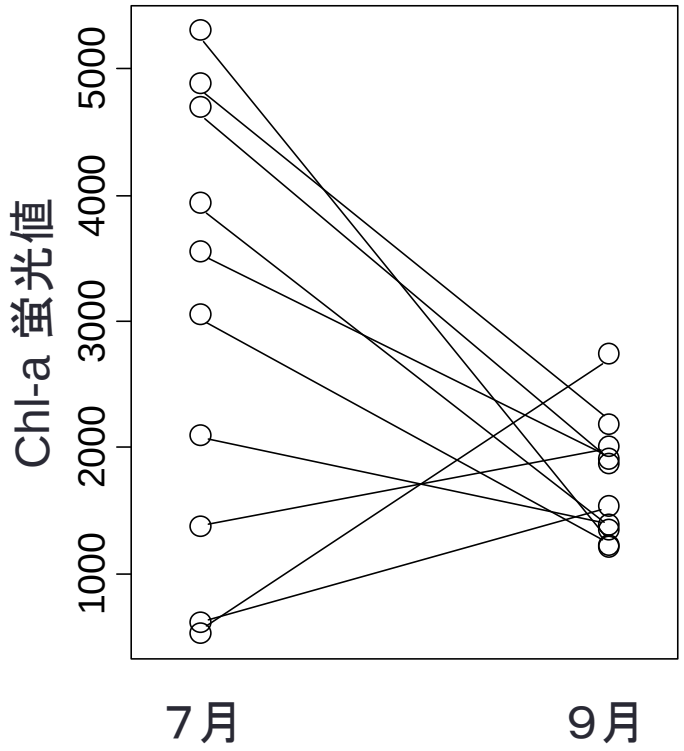


図1. 7月と9月のChl-a蛍光値

7月と9月のChl-a蛍光値の比較を行いたい

1. どのタイプの*t*検定が良いか
2. Studentの*t*検定、Welchの*t*検定、対応のある*t*検定それぞれの結果は？

結果の表現

図と表

- 表のことを図と呼ばないこと！

図

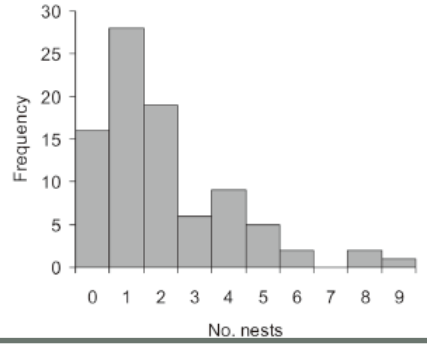


図 2. 5 × 5 km メッシュ内の営巣数の頻度分布。
Fig. 2. Frequency distribution of the number of nests in 5 × 5 km cells.

脚注は下に書く

表

表 1. 関東地方の 5 × 5 km メッシュ内の営巣数を予測するためのポアソン回帰モデル。赤池の情報量基準による複数モデル推測によりモデルの推定値を求めた。
Table 1. Results of Poisson regression to predict the number of nests in 5 × 5 km cells in and around the Kanto district, central Japan. Multimodel inference based on Akaike's information criterion was used to obtain parameter estimates.

Variable (km^2)	Polynominal	Coefficient	SE	Σw^i
Intercept		-2.4555	0.8832	
Area of flatland	First order	0.0526	0.0502	0.67
Area of flatland	Second order	0.0005	0.0013	0.45
Area of urban land	First order	3.5409	1.1743	0.98
Area of urban land	Second order	-3.0022	0.9326	0.98
Area of forest	First order	0.0817	0.0578	0.82
Area of forest	Second order	-0.0009	0.0024	0.45
Area of open land < 200 m from forest edge	First order	0.0495	0.0586	0.59
Area of open land < 200 m from forest edge	Second order	0.0008	0.0026	0.45

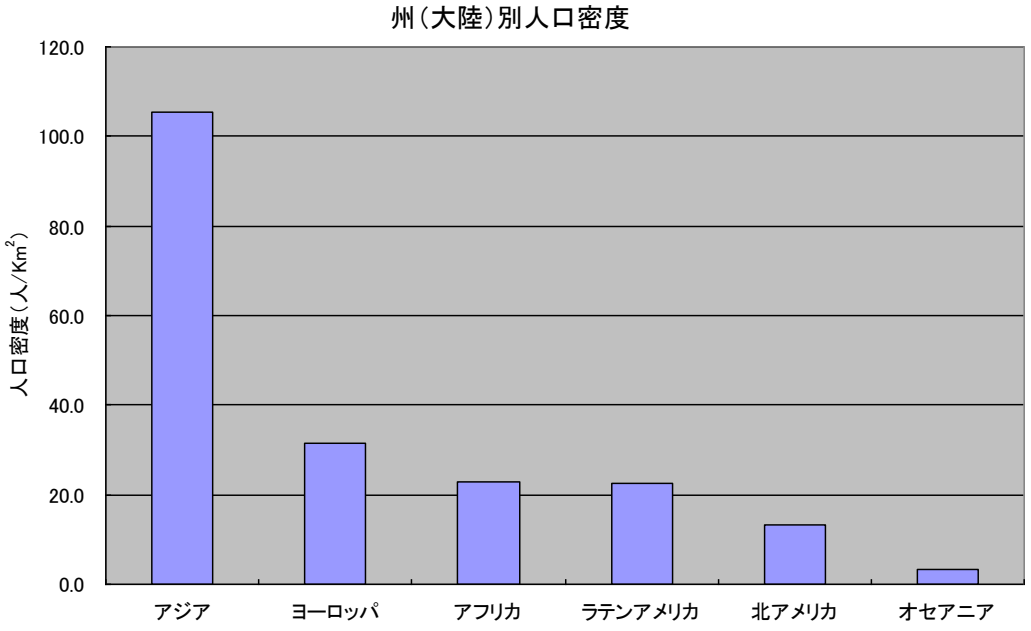
脚注は上に書く

図を使うか表を使うか

- 図と表のどちらが分かりやすいかで判断
- 図にも表にもできる情報　⇒　直感的に理解しやすい図

州（大陸）別の人口密度

州（大陸）	人口密度（人/Km^2）
アジア	105.5
ヨーロッパ	31.6
アフリカ	22.7
ラテンアメリカ	22.6
北アメリカ	13.3
オセアニア	3.3



具体的な数値が分かる

直感的に理解しやすい

表の表し方

（悪い例）

Site id	7月	9月
1	5307	1205
2	3932	1340
3	4875	2179
4	3051	1217
5	3552	1902
6	607	1535
7	2098	1388
8	1376	2001
9	522	2733
10	4687	1871

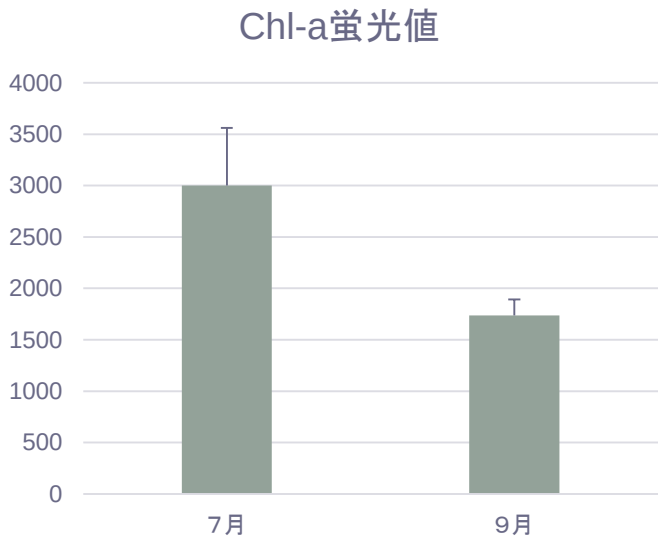
（良い例）

Site id	7月	9月
1	5307	1205
2	3932	1340
3	4875	2179
4	3051	1217
5	3552	1902
6	607	1535
7	2098	1388
8	1376	2001
9	522	2733
10	4687	1871

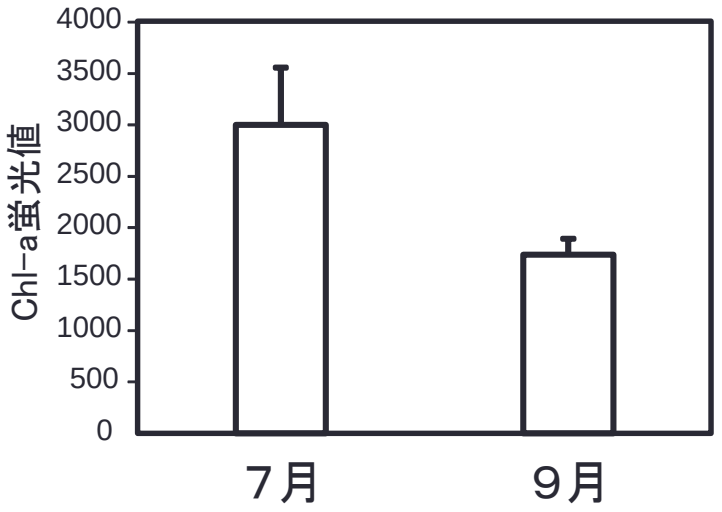
基本的に、デフォルトの表はNG。
表ツールの「罫線を引く」などで整える。
縦の線は基本的に不要。（ただし、プレゼンでは見やすさに応じて加える事も）

図の表し方（Excel）

（悪い例）



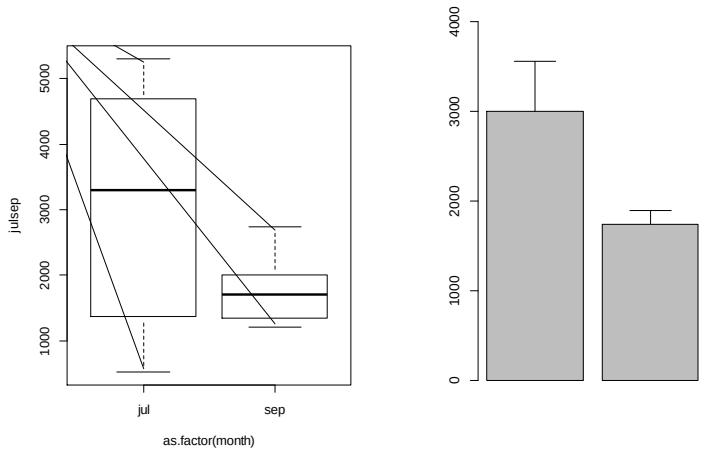
（良い例）



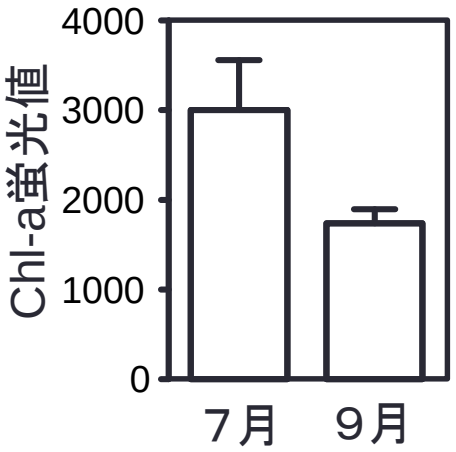
文字は大きく、線は太く、余計な情報は省く

図の表し方(R)

（悪い例）



（良い例）



*t*検定など母集団の正規分布を仮定するような場合は、箱ひげ図は普通使わない。逆に比率など正規分布していないようなデータを平均＋SEで表現するのも不適

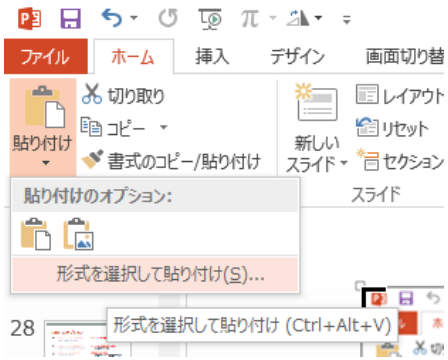
＊ただしRで棒グラフの描画は若干めんどいです（次回やるかも）

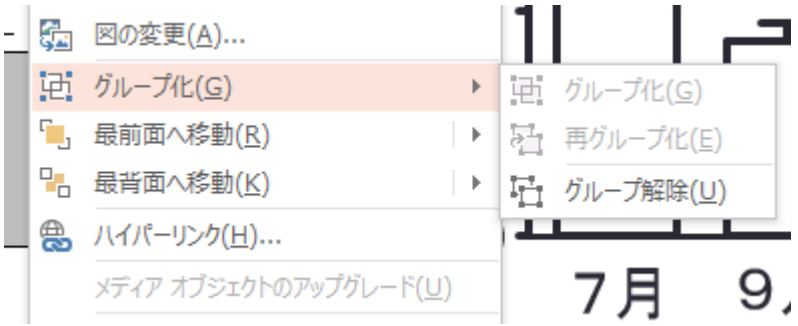
図の編集（一例）

- 図はエクセルやRで頑張るよりも。パワーポイントなどで編集する方が楽

1. ウィンドウズメタファイル形式（.wmf）で貼り付ける

2. 図を右クリック→グループ解除で各要素を分解して再編集





文章表現（方法）

（悪い例）

「7月と9月に有意差があるかt検定した」

「～でt.testを行った」

（改善例）

「7月と9月の蛍光値に差が見られるか*t*検定を行った」

「7月と9月の蛍光値の比較は*t*検定により行った」

「蛍光値に対する月の影響を見るために*t*検定を行った」

「～は*t*検定によって検定した」でも可

「～と～では分散が異なっていたため（*F*=…, P=…）、Welchの*t*検定を行った」

文章表現（結果）

（悪い例）

「７月と９月では有意差が見られた（P<0.05）」

「４月と５月ではあまり有意でなかったが（P<0.1）、５月の方が若干高かった」

（改善例）

「７月に比べ９月では有意に高い値を示した（t＝3.3, 自由度=10, P=0.038）」

「７月に比べ９月の値はおよそ1.2倍に上昇した（t_{11}＝3.3, P=0.038）」

「４月と５月ではばらつきが大きく有意な差は見られなかったが（ t_{10}＝2.56 P=0.058 ）、最大値で見ると～」

次回予告

- Rの操作
  - データの読み込み・加工
- 統計解析
  - 回帰
  - 分散分析
- 論文表現
  - 散布図
  - エラーバー付き棒グラフ
  - 分散分析表

データ募集中！

ベクトル

```
c(1,2,3,4,5)                    1 2 3 4 5
c(1:5)                          1 2 3 4 5
c(8:3)                          8 7 6 5 4 3
rep(2,3)   #2を3回繰り返し       2 2 2
seq(1,9,by=2)  #1から9まで2おき  1 3 5 7 9
seq(0,10,length=5)              0.0 2.5 5.0 10.0
                    #0から10まで5分割
#応用編
rep(1:3,2)                      1 2 3 1 2 3
rep(1:3,1:3)                    1 2 2 3 3 3
rep(1:3,rep(2,3))               1 1 2 2 3 3
```

論理演算

```
a<- c(7,6,4,0)                    T=TRUE; F=FALSE
a==4          #等号               F  F  T  F
a!=4          #不等号(≠)          T  T  F  T
a>=3          #以上(<=以下)       T  T  T  F
a<3           #未満               F  F  F  T
a!=3&a>2      #&  かつ            T  T  T  F
a!=3|a>4      #|  または          T  T  T  T
```

ベクトル要素へのアクセス

```
a<- c(7,6,4,0)
a[2]                #aの2番目                6
a[c(4,2)]           #aの4番目と2番目         0 6
a[-2]               #aの2番目以外            7 4 0
a[a>5&a!=6]         #条件式に合うもの        7
a[2]<- 9            #要素への代入
a                                            7 9 4 0
```

```
a<- c(7,6,4,0)
b<- c(2,7,8,4)
a[b>5]                で何と出力されるか
```

ベクトル計算

#基本的に各要素に対し計算される

```
a<- c(7, 6, 4, 0)
b<- c(2, 7, 8, 4)
a+b                                  9 13 12 4
a-3                                  4 3 1 -3
a*(b-2)                              0 30 24 0
```

#T/Fは数値的には1/0 として扱われる

```
(a>5)+b                              3 8 8 4
```

ベクトル用の関数(その2)

```
a<- c(7,6,4,0,3,8,5)
sort(a)    #昇順整列                    0 3 4 5 6 7 8
order(a)   #整列した時の元の順番        4 5 3 7 2 1 6
rank(a)    #整列した時の順位            6 5 3 1 2 7 4
```

```
a<- c(7,6,4,0)
b<- c(2,7,8,4)
a[order(b)]    で何と出力されるか？
```

実例1: 分割表の検定

（Chi-squared test・カイ2乗検定）

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

	男	女	計
鏡味研	7	1	8
西廣研	3	2	5
計	10	3	13

1. 対立仮説H_1および帰無仮説H_0を設定する

対立仮説H1　男女比は異なる

帰無仮説H0　男女比は異ならない
（鏡味研も西廣研も10:3）

注:2×2分割表で少サンプルの場合は近似を用いるχ^2検定よりもFisherの正確確率検定のほうが良いとされています

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

観察値	男	女	計
鏡味研	7	1	8
西廣研	3	2	5
計	10	3	13

2. 検定統計量を設定し、データから検定統計量を計算する

統計量

$$\chi^2 = \sum_{i=1}^{n} \frac{(o-e)^2}{e}$$

o: 観察値observed
e: 期待値expected

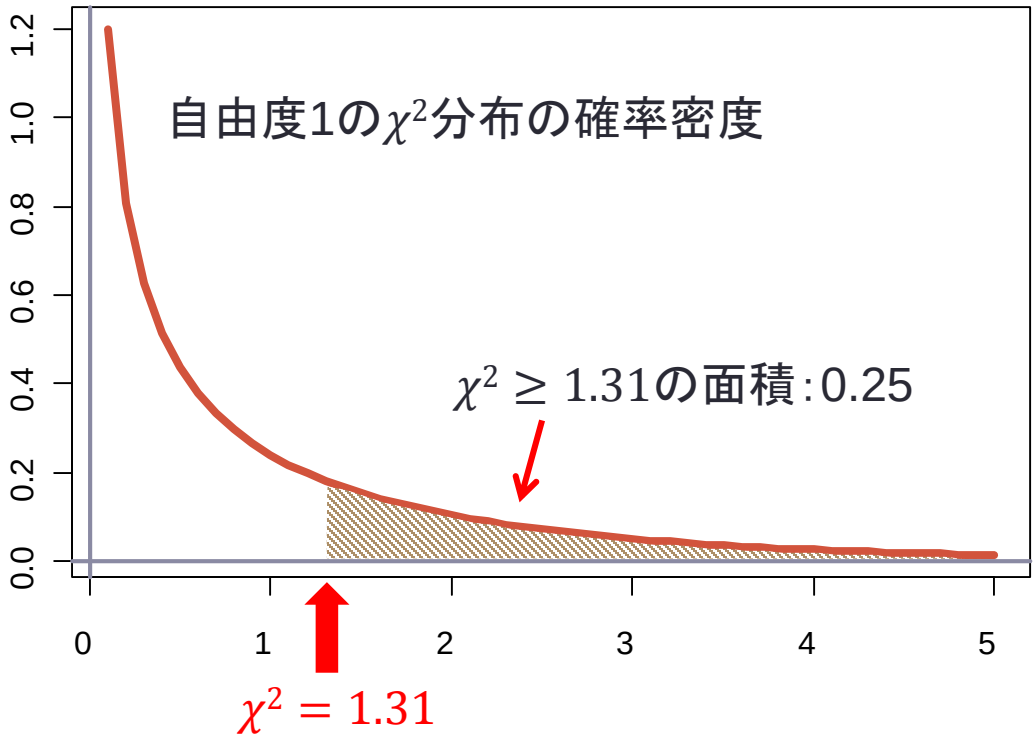
期待値	男	女	計
鏡味研	8*(10/13)=6.154	8*(3/13)=1.846	8
西廣研	5*(10/13)=3.846	5*(3/13)=1.154	5
計	10	3	13

$$\chi^2 = \frac{(6.154-7)^2}{6.154} + \frac{(1.846-1)^2}{1.846} + \frac{(3.846-3)^2}{3.846} + \frac{(1.154-2)^2}{1.154} = 1.31$$

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

3. 計算した統計量の値よりも極端な値が、帰無仮説が正しいと仮定したときに得られる確率（P値）を求める

n×m行列の分割表において、帰無仮説が正しい時のχ^2値の分布は自由度$(n-1)(m-1)$のχ^2分布に従う



P=0.25

- 鏡味研究室と西廣研究室の卒研生の男女比は異なるか？

4. P値が有意水準よりも小さければ、帰無仮説を棄却する（大きければ棄却しない）

有意水準α＝0.05の場合、P値＝0.25
なので、帰無仮説を棄却できない

→対立仮説（男女比が異なる）は採用できない

→結論：男女比は異なるとはいえない
（≒男女比は異ならない）

Rでは下記の2行で実行

```
mat<- matrix(c(7,3,1,2),ncol=2)
chisq.test(mat, correct=F)
```