統計解析ソフトRを用いたデータ解析
DAY.2

下川敏雄

和歌山県立医科大学 医学部／附属病院臨床研究センター

シェーマ

日程	内容
1日目(1)	データの要約の方法(1):連続尺度の場合
1日目(2)	データの要約の方法(2):離散尺度(カテゴリカル・データ)の場合
1日目(3)	プログラミング(その1):データの要約の方法
2日目(1)	統計的推測(1):連続尺度の場合
2日目(2)	統計的推測(2):離散尺度(カテゴリカル・データ)の場合
2日目(3)	プログラミング(その2):統計的推測

DAY.2：推測統計学入門

Section.1：連続尺度の比較

仮説検定：問題の例示

いま，ある政令指定都市において，平成15年と平成30年のあいだの外国人の人口変動が調査された．この調査結果を受けて，以下のような課題が検討された．

問題1

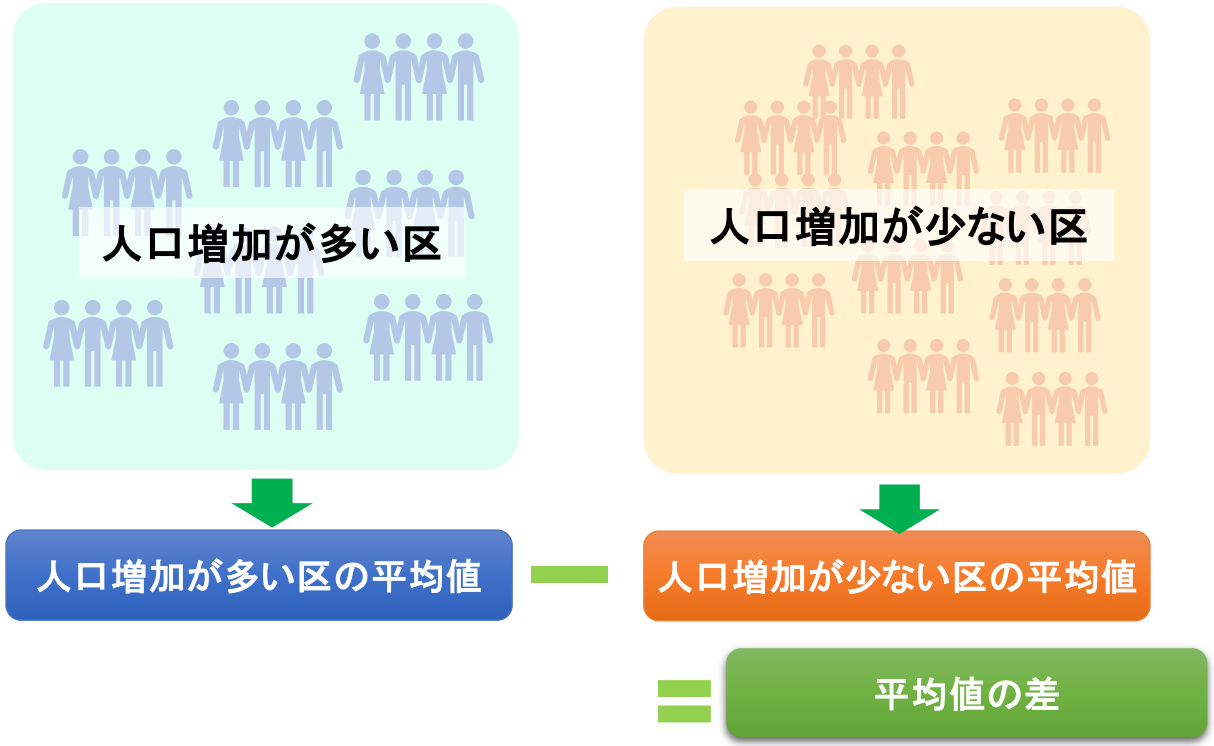
人口増加が平均よりも多い区(多)と少ない区(少)において，外国人変化率の平均値に違いがあるだろうか？

問題2

各区の人口増加量に男女差があるだろうか？

対応があるデータと対応がない(独立2標本)データ

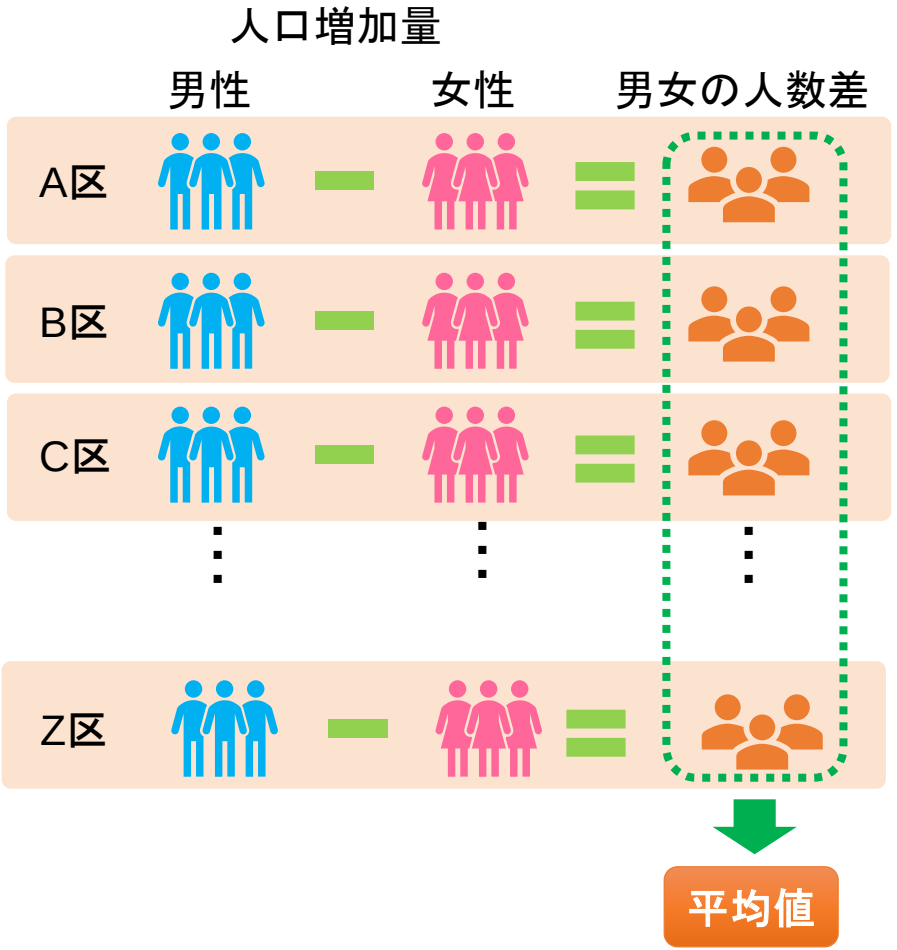
人口増加が多い区と少ない区の外国人増加率の差



人口増加が多い区と人口増加が少ない区の平均値の差が0でないということは，人口増加で分けた2種類の区のグループ間で，外国人増加率に違いが認められる．

対応のない(独立2標本)データ

外国人増加量の男女差



男女の人数差の平均値が0でないということは，人口増加に性差が認められる．

対応のあるデータ

仮説検定とは何か

問題１

人口増加が平均よりも多い区(多)と少ない区(少)において，外国人増加率の平均値に違いがあるだろうか？

人口増加が多い区

外国人増加率の平均
51.6%

平均値の差
51.6 – 27.9
= 23.7

人口増加が少ない区

外国人増加率の平均
27.9%

平均値の差=23.7%というのは，統計学的にみて，二つのグループに違いがあると判断しても良いものだろうか？

帰無仮説

人口増加が多い区と少ない区では外国人増加率は同じであるとは，

(人口増加が多い区の平均値$\mu_{多}$) ＝人口増加が少ない区の平均値$\mu_{少}$

$$\mu_{多} - \mu_{少} = 0$$

「同じである」という状況は，どのような比較でも同じである．

唯一の数値で設定できるこちら側を評価の基準とする
(ただし研究の目的と逆)

対立仮説

人口増加が多い区と少ない区では外国人増加率が異なるとは

(人口増加が多い区の平均値$\mu_{多}$) ≠人口増加が少ない区の平均値$\mu_{少}$

である．**「異なる」**という状況は**無限大に存在する**．

上の仮説が誤っていることを示すことで，逆の仮説であるこちらが正しいと判断する．

帰無仮説H_0

人口増加が多い区と少ない区で外国人増加率に違いがない

逆仮説

仮説が背反になっている

対立仮説H_1

人口増加が多い区と少ない区で外国人増加率に違いがある

違いがない = 平均値の差が0である

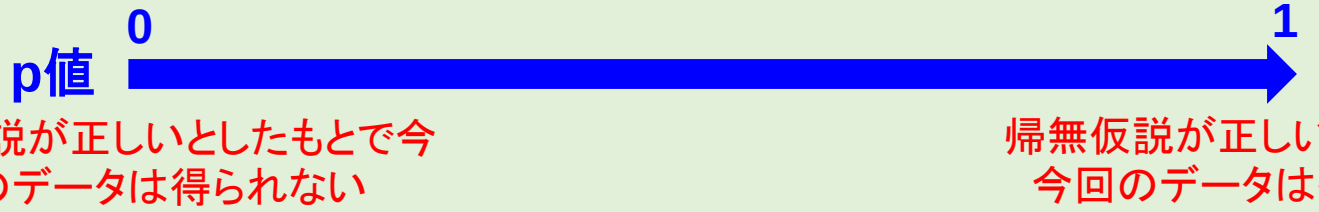
シナリオのパターンは1種類

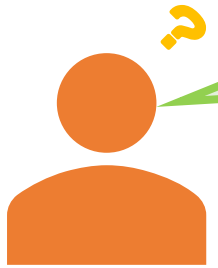
違いがある = 平均値の差にさまざまなシチュエーションがある

シナリオのパターンは無限大

帰無仮説H_0のシナリオが正しいとしたもとで，今回のデータがどれぐらいの確率で得られるかを考える．

帰無仮説H_0の確からしさを確率のようなもので表して判断するのが仮説検定である．その確からしさを表す測度をp値という．





では，p値がどの程度の値になれば，帰無仮説H_0が誤っていると判断できるのか？

p値がどれぐらい小さいと帰無仮説H_0が誤っているといえるかを決定するカットオフ値が必要である．このカットオフ値を**有意水準α**といい，一般的には**0.05 (5%)**とされている．

p値 (有意確率)
例示: p値=0.135

p値が有意水準α(一般的にはα=0.05)未満

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論
平均値に差がある

p値が有意水準α(一般的にはα=0.05)以上

例示の場合

帰無仮説H_0が間違っているとは言えない (受容)

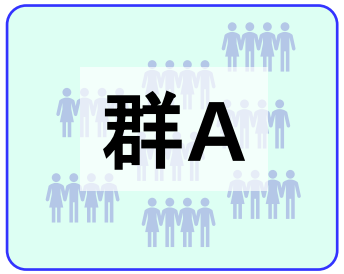
有意でない (有意差がない)

結論
平均値に差があるとは言えない

仮説検定において「差がない」とは言えないので注意

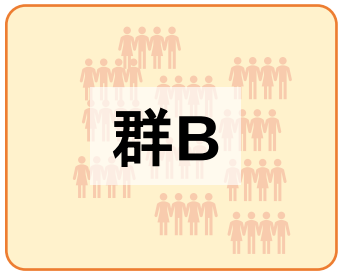
対立仮説には3種類存在する

両側対立仮説



群Aの平均 ＞ 群Bの平均

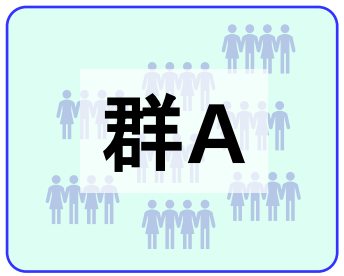
群Aの平均 ＜ 群Bの平均



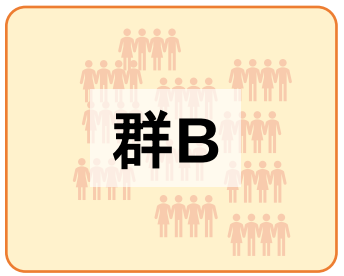
有意差がある(p値が有意水準α=0.05よりも小さい)と判定された場合には,
「群Aと群Bの平均は異なる」
と解釈する.

医学研究の特殊なシチュエーションを除くと両側対立仮説を用いるのが一般的

片側対立仮説(1)

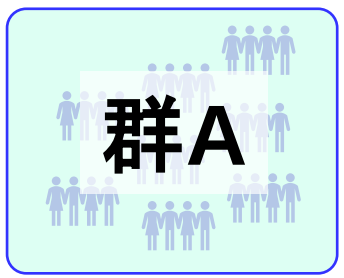


群Aの平均 ＞ 群Bの平均

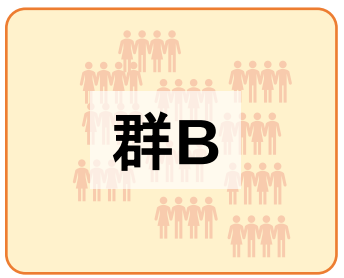


有意差がある(p値が有意水準α=0.05よりも小さい)と判定された場合には,
「群Aの平均は群Bの平均よりも大きい」
と解釈する.

片側対立仮説(2)



群Aの平均 ＜ 群Bの平均

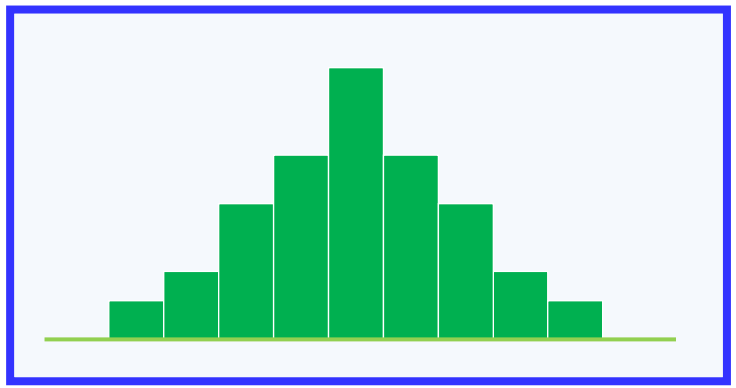


有意差がある(p値が有意水準α=0.05よりも小さい)と判定された場合には,
「群Bの平均は群Aの平均よりも大きい」
と解釈する.

パラメトリック検定とノンパラメトリック検定

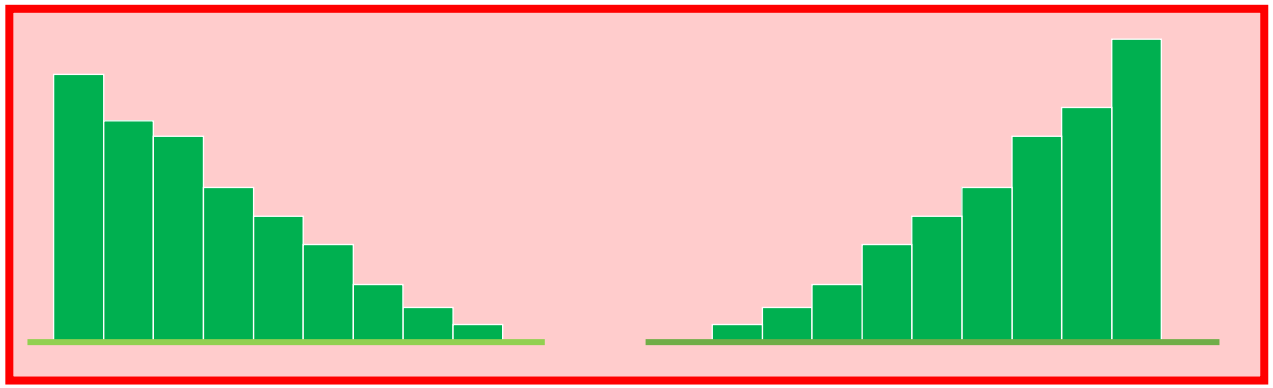
■ データは正規分布に従っているか？

データの分布を見たときに，平均値を中心に左右対称になっている(正規分布に従っている)か否かによって検定方法が異なる.



正規分布に従っている
(平均で比較することに意味がある)

パラメトリック検定

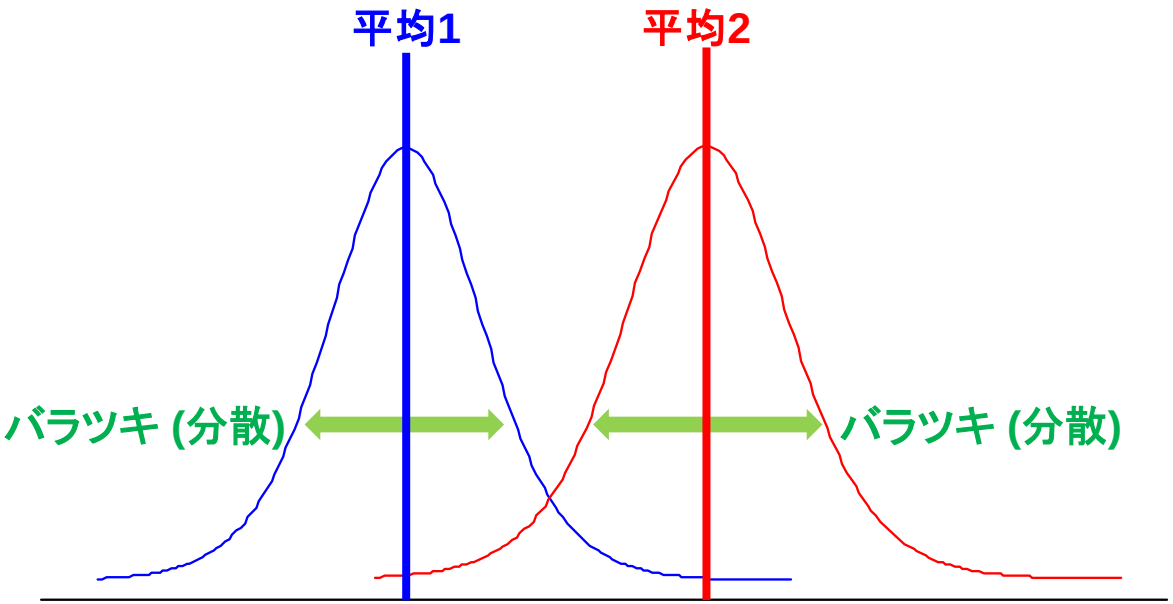


正規分布に従っていない
(平均で比較することに意味がない)

ノンパラメトリック検定

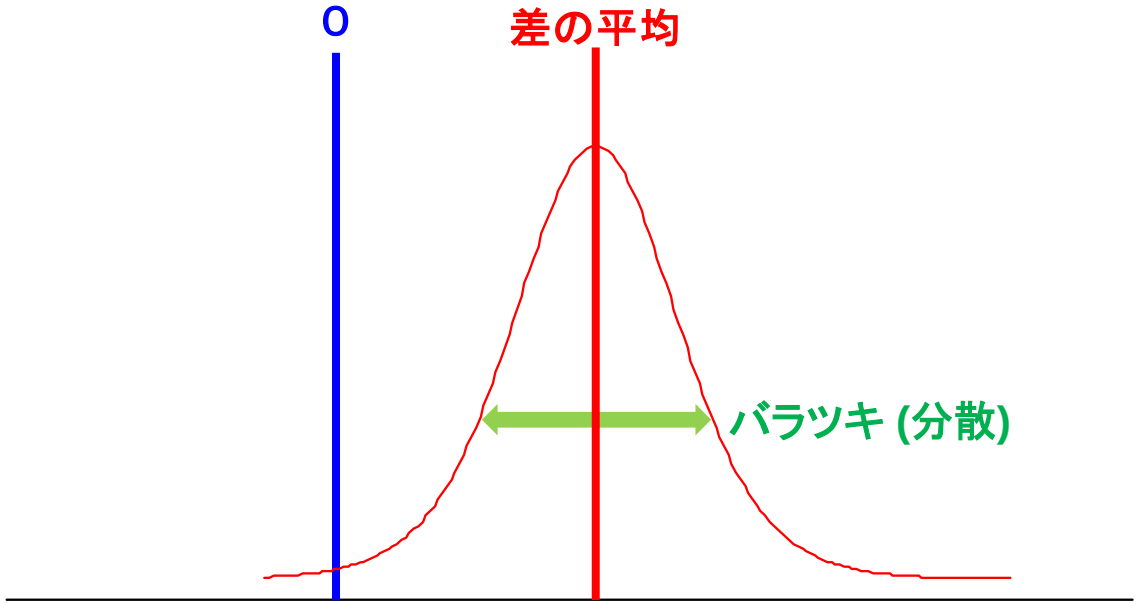
パラメトリック検定：2標本t検定と対応のあるt検定

2標本t検定



2標本t検定とは，バラツキ(分散)が2群で同じであると仮定したもとで，二つの平均値の差「平均2－平均1」が0であるかどうか(つまり平均値が等しいかどうか)を検定している.

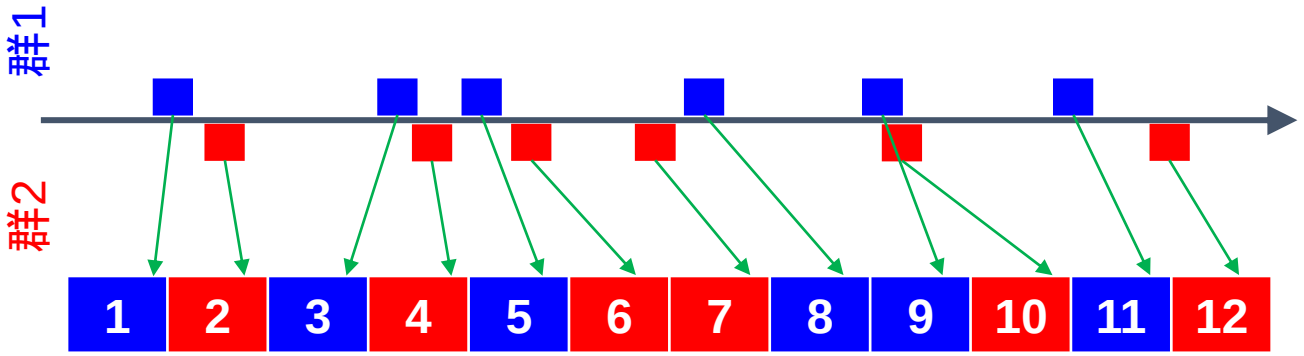
対応のあるt検定



対応のあるt検定とは，差(X-Y)の平均値が0であるかどうかを検定している．したがって，バラツキ(分散)は差(X-Y)のみに存在する.

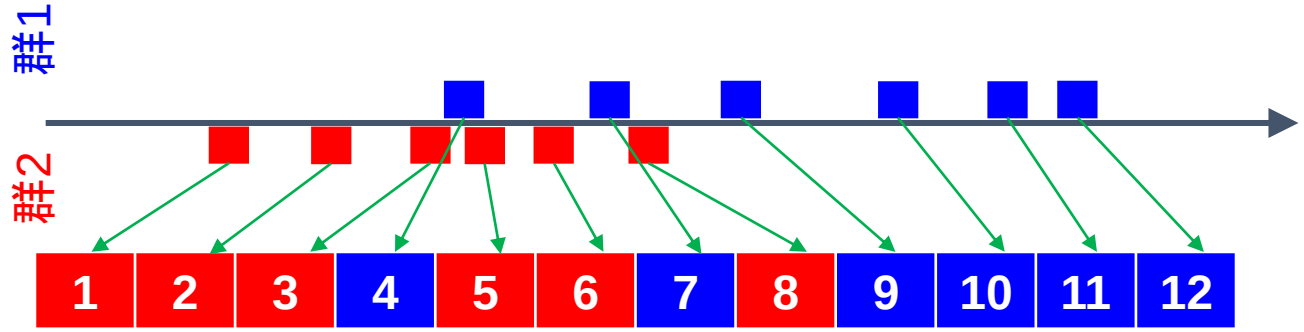
ノンパラメトリック検定の意味(1)：Wilcoxon検定

群1と群2がほぼ同じである場合(有意でない場合)



群1と群2をプールして小さい順に並べ替えたとき，それぞれの群がおおよそ交互に並んでいる.

群1と群2が異なる場合(有意である場合)

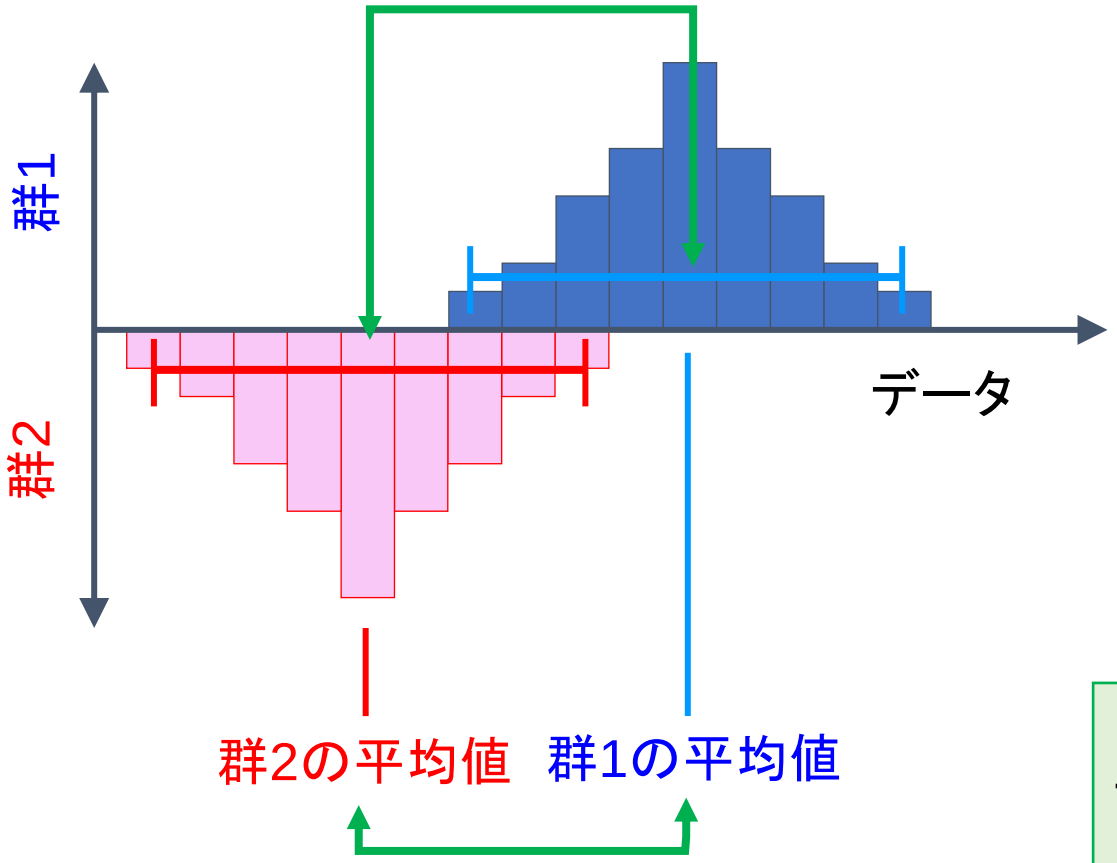


群1と群2をプールして小さい順に並べ替えたとき，群2が順位が小さい側，群1が大きい側に並んでいる.

Wilcoxon検定では，小さい順に並べ替えたときのいずれかの群の順位の和に基づいて検定を行う. Wilcoxon検定は中央値を比較しているわけではなく，分布の相対的な位置関係を評価している.

2標本t検定とWilcoxon検定(ノンパラメトリック検定)の違い

Wilcoxon検定では，相対的な位置関係を比較している．



2標本t検定では平均値(それぞれの群を代表する値)を比較している．

■ 2標本t検定で有意であるということ

「2群間の平均値が違う」と解釈できる．

■ Wilcoxon検定で有意であるということ

「2群間の相対的な位置関係が違う」と解釈できる．

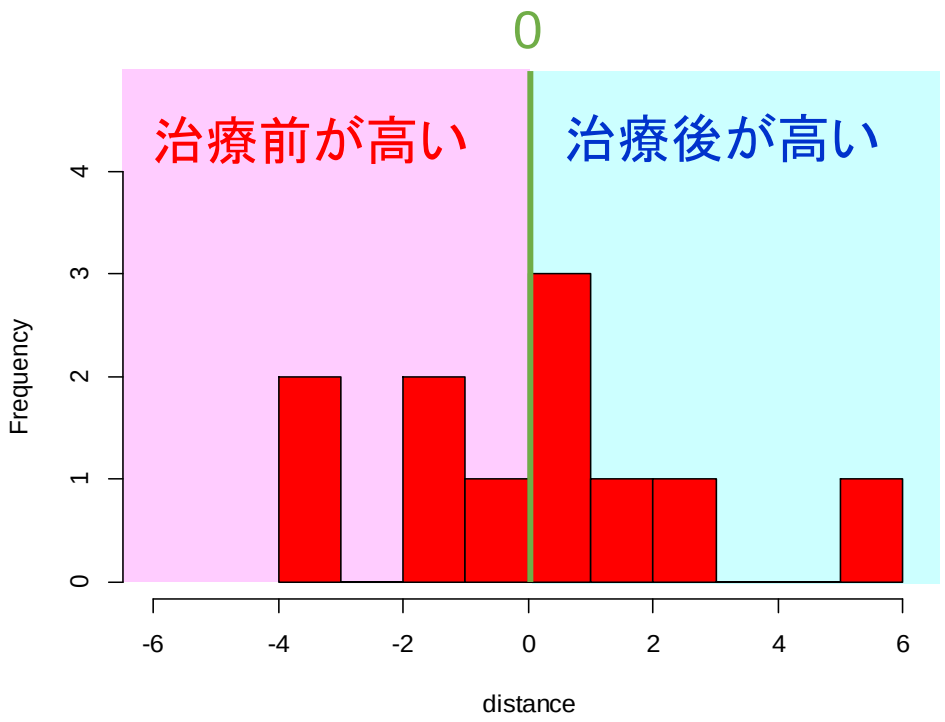
中央値が違うとは言っていない．ノンパラメトリック検定を用いた場合に，中央値を用いるのは，その他に群の代表値として用いるものがないため．

上記の理由のほかにも，分散分析や回帰分析(いわゆる多変量解析)においても，平均値で解釈される．ため，可能な限り2標本t検定を用いるほうが良い．

ノンパラメトリック検定の意味(1)：Wilcoxon符号付き順位和検定

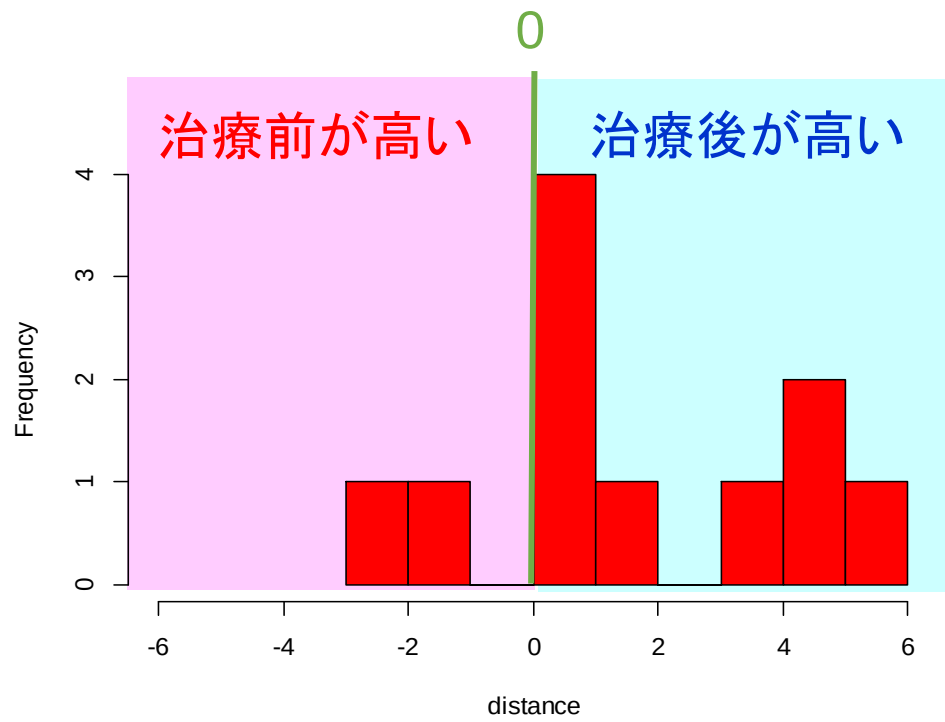
有意でない状況

No	前	後	差	No	前	後	差
1	5	4	-1	7	1	2	1
2	1	2	1	8	9	6	-3
3	7	8	1	9	4	3	-1
4	9	5	-4	10	2	2	0
5	3	9	6	11	3	6	3
6	2	4	2				



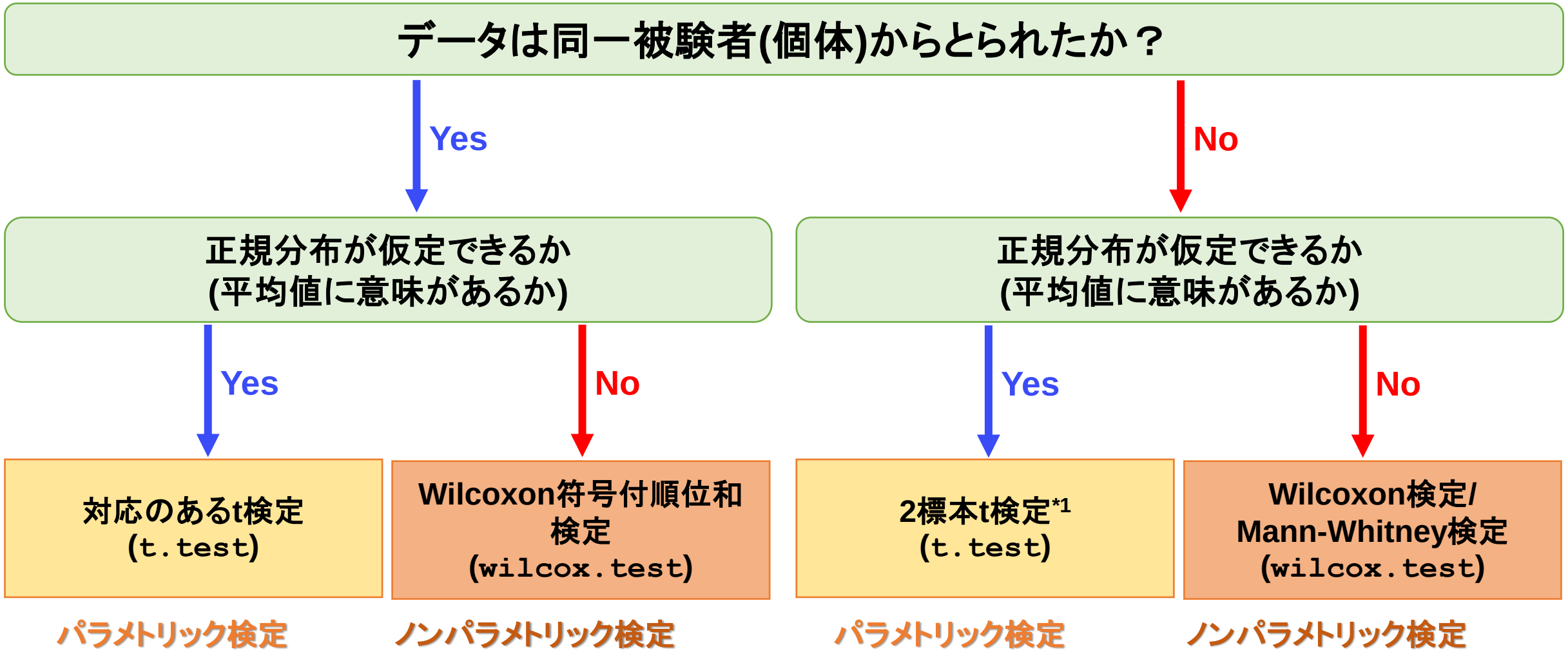
有意である状況

No	前	後	差	No	前	後	差
1	3	4	1	7	1	2	1
2	1	2	1	8	2	6	4
3	3	8	5	9	4	3	-1
4	4	5	1	10	5	2	-3
5	3	9	6	11	1	6	5
6	2	4	2				



Wilcoxon符号付順位和検定では，青色の領域の数と赤色の領域の数の偏りを検定している．

量的変数における検定方法の取捨選択



*1：平均の差の検定には，2標本t検定(等分散を仮定する検定)とWelch検定(等分散を仮定しない検定)がある．しかしながら，Welch検定は問題が多い検定なので，使わないほうが良い．

Rにおける2標本t検定

ここでは，Cドライブの`Fukuoka_Seminor`というフォルダにある`data_Cancer.csv`というCSVファイルを読み込む．これは，e-statを用いて作成した47都道府県の病院数，病床数，高齢者世帯割合，各種がんの罹患率(10万人対)である．

Input

```
>   dat <- read.csv("C:/Fukuoka_Seminor/data_Cancer.csv",fileEncoding = "cp932")
>   head(dat)
```

Output

	Pref	WE	Pop	Hosp	Bed	Old	All_Cancer	Stmach	Liver	Cervix	Prostate	Colon	Lung
1	北海道	E	5250000	9.2	997.4	40.86751	413.0396	38.47137	13.020425	15.010433	66.91269	62.68715	49.45740
2	青森県	E	1246000	6.2	807.7	49.51296	412.2754	46.41344	11.319561	14.224902	61.24145	71.85182	42.79534
3	岩手県	E	1227000	6.2	739.3	49.72934	384.3939	41.37791	11.182693	14.402174	68.11246	61.65999	39.07964
4	宮城県	E	2306000	4.9	674.9	40.26219	394.9599	50.75323	10.190467	9.324404	59.58221	58.44151	44.45808
5	秋田県	E	966000	5.4	893	55.65678	413.8067	58.59089	11.133075	12.213162	61.32791	73.62666	35.98922
6	山形県	E	1078000	5.0	802.7	54.50513	375.8901	60.10692	8.750719	10.012467	62.33251	52.89904	38.92873

`Pref`:都道府県名

`Pop`:人口

`Bed`: 100,000人当たりの病床数

`All_Cancer`: 100,000人当たりのがん罹患者数

`Liver`: 100,000人当たりの肝臓がん罹患者数

`Prostate`: 100,000人当たりの前立腺がん罹患者数

`Lung`: 100,000人当たりの肺がん罹患者数

`WE`:気象庁による東西の定義(`E`:東日本，`W`:西日本)

`Hosp`:100,000人当たりの病院数

`Old`:高齢者世帯の割合(%)

`Stmach`: 100,000人当たりの胃がん罹患者数

`Cervix`: 100,000人当たりの子宮頸がん罹患者数

`Colon`: 100,000人当たりの大腸がん罹患者数

東日本と西日本ですべてのがん罹患者数(All_Cancer)の平均値に違いがないかを比較する.

Input

```
>   t.test(dat$All_Cancer~dat$WE, var.equal=TRUE)
```

関数t.test()は，t検定全般を行うことができる検定である．ここで，

X~G　(今回の場合は，dat$All_Cancer~dat$WE)

は，Xがデータであり，Gが群を表す変数を意味する．また，

var.equal=TRUE

は，2標本t検定を選定していることを意味する(デフォルトはvar.equal=FALSEであり，お勧めできないWelch検定になっている).

Output

```
        Two Sample t-test

data:  dat$All_Cancer by dat$WE
t = -1.1715, df = 45, p-value = 0.2476
alternative hypothesis: true difference in means between group E and group W is not equal to 0
95 percent confidence interval:
 -15.723030   4.158682
sample estimates:
mean in group E mean in group W
       383.1662        388.9484
```

① p値を表している，　② 各群の平均値を表している.

帰無仮説H_0: 西日本と東日本ですべてのがんの10万人対の罹患者数の平均に違いがない.
対立仮説H_1: 西日本と東日本ですべてのがんの10万人対の罹患者数の平均に違いがある.

p値 (有意確率)
例示: p値=0.2479

p値が有意水準α(一般的にはα=0.05)未満

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論
平均値に差がある

p値が有意水準α(一般的にはα=0.05)以上

今回の場合

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論
平均値に差があるとは言えない

仮説検定において「差がない」とは言えないので注意

RにおけるWilcoxon検定

東日本と西日本ですべてのがん罹患者数(All_Cancer)の相対的な位置に違いがないかを比較する.

Input

```
>   wilcox.test(dat$All_Cancer~dat$WE)
```

関数wilcoxon.test()は, Wilcoxon検定全般を行うことができる検定である. ここで,

X~G （今回の場合は, dat$All_Cancer~dat$WE）

は, Xがデータであり, Gが群を表す変数を意味する.

Output

```
        Wilcoxon rank sum exact test

data:  dat$All_Cancer by dat$WE
W = 221, p-value = 0.2487
alternative hypothesis: true location shift is not equal to 0
```

① p値を表している

帰無仮説H_0: 西日本と東日本ですべてのがんの10万人対の罹患者数の相対的な位置に違いがない.

対立仮説H_1: 西日本と東日本ですべてのがんの10万人対の罹患者数の相対的な位置に違いがある.

p値 (有意確率)

例示: p値=0.2487

p値が有意水準α(一般的にはα=0.05)未満

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論

相対的な位置に差がある

p値が有意水準α(一般的にはα=0.05)以上

今回の場合

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論

相対的な位置に差があるとは言えない

仮説検定において「差がない」とは言えないので注意

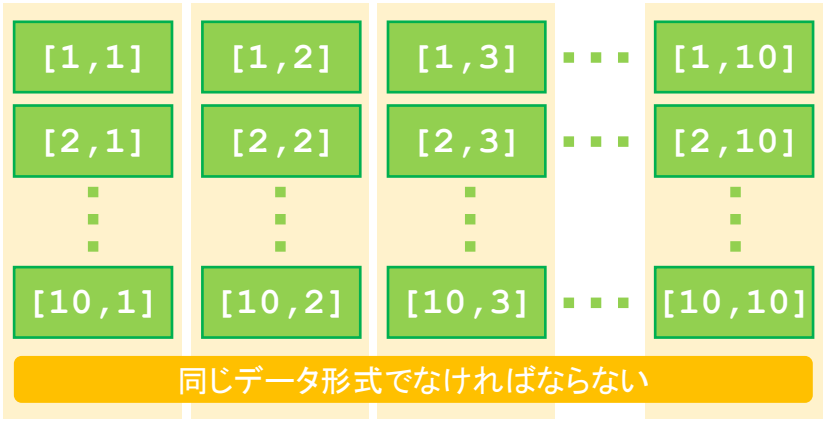
補足説明：Rでは，主に用いる形式として4種類のデータ構造がある．

ベクトル型

Example

```
> X <- c(1,2,3,4,5)
> X
[1] 1 2 3 4 5
> X[3]
[1] 3
```

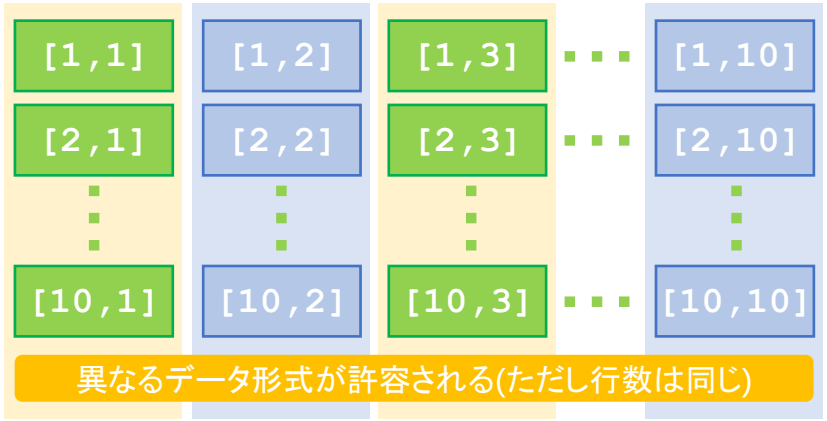
行列型



Example

```
> Y <- matrix(1:6, ncol=2)
> Y
     [,1] [,2]
[1,]    1    4
[2,]    2    5
[3,]    3    6
> Y[3,]
[1] 3 6
```

データフレーム型



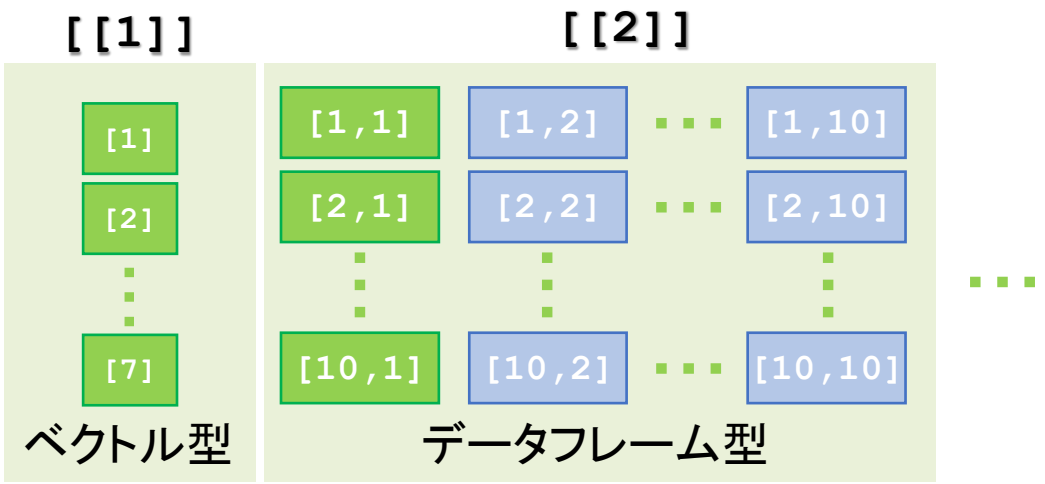
Example

```
> a <- c("M","F","M","M","F")
> b <- 1:5
> c <- c(3.2, 2.4, 12.3, 8.3, 9.6)
> DF <- data.frame(a,b,c)
> DF
  a b    c
1 M 1  3.2
2 F 2  2.4
3 M 3 12.3
4 M 4  8.3
5 F 5  9.6
> DF[3,]
  a b    c
3 M 3 12.3
```

リスト型

データフレーム型では行数が違うとエラーになる

```
> a <- c("M","F","M","M","F")
> b <- 1:5
> c <- c(8.3, 9.6)
> DF2 <- data.frame(a,b,c)
> DF2
 data.frame(a, b, c) でエラー:
   引数に異なる列数のデータフレームが含まれています: 5, 2
```



リスト型は，様々なデータ構造のものを一つに束ねることができる．

```
> Lst <- list(X,Y,DF)
> Lst
[[1]]
[1] 1 2 3 4 5

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

[[3]]
  a b    c
1 M 1  3.2
2 F 2  2.4
3 M 3 12.3
4 M 4  8.3
5 F 5  9.6

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

```
> Lst[[1]][3]
[[1] 3

> Lst[[3]][4,]
  a b   c
4 M 4 8.3
```

R における統計解析の結果の殆どがリスト型で記述される．

Input

```
>   res <- t.test(dat$All_Cancer~dat$WE, var.equal=TRUE)
>   names(res)
```

Output

```
 [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"    "null.value"
 [7] "stderr"      "alternative" "method"      "data.name"
```

このなかで，p値はp.valueのなかに含まれている．下記のように中身を見ることができる．

Input

```
>   res$p.value
```

Output

```
[1] 0.2475562
```

これを応用すると，次のようなことができる．

Rエディッターの内容

```
pval.calc <- function(X, grp){
        res <- t.test(X~grp, var.equal=TRUE)$p.value
        return(round(res,3))
}
```

Input

```
>   pval.calc(dat$All_Cancer,dat$WE)
```

Output

```
[1] 0.248
```

つまり，p値のみを表示する関数を作ることができる．

関数を作るためのメリット (関数apply()は，自作の関数でも用いることができる).

apply()

```
apply(X,dir,function,options)
```

X	:データフレーム or 行列型	dir	:計算の方向 (1:縦方向, 2:横方向)
function	:関数名	options	:functionのデータ以外のオプション(複数可)

Input

```
> apply(dat[,7:13], 2, pval.calc, dat$WE)
```

Output

```
All_Cancer      Stmach       Liver      Cervix    Prostate       Colon        Lung
     0.248       0.310       0.000       0.028       0.278       0.064       0.015
```

対応のあるt検定

各都道府県のがんの罹患状況において，胃がん(Stmach)と大腸がん(Colon)に違いがあるだろうか？

Input

```
>  t.test(dat$Stmach, dat$Colon,  paired=TRUE)
```

変数1　変数2

対応のある検定であることを宣言する

対応のある検定では変数を並べて書く

Output

```
        Paired t-test

data:  dat$Stmach and dat$Colon
t = -11.826, df = 46, p-value = 1.511e-15
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -16.72377 -11.85882
sample estimates:
mean difference
       -14.2913
```

①

②

①は，p値を表している．ここで，1.511e-15とは，1.511×10^{-15}を意味しており，とても小さな数値であるといえる．通常，p値は小数点以下3桁未満は切り捨て(あるいは四捨五入)する．

②は，二つの変数の差の平均値を表している(変数1ー変数2)．

帰無仮説H_0: 47都道府県のがんの罹患者数において，胃がんと大腸がんのあいだで差がない.

対立仮説H_1: 47都道府県のがんの罹患者数において，胃がんと大腸がんのあいだで差がある.

厳密には，各都道府県の差の平均値が0であるかどうかを検定している.

p値 (有意確率)

例示: p値<0.001

p値が有意水準α(一般的にはα=0.05)未満

今回の場合

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論

胃がんと大腸がんで差がある.

p値が有意水準α(一般的にはα=0.05)以上

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論

胃がんと大腸がんで差があるとは言えない.

したがって，47都道府県のがん罹患者数において，胃がんと大腸がんでは違いが認められる．また，各都道府県で差をとったときの平均値が-14.2913であることから，胃がんのほうが罹患者が少ない.

Wilcoxon符号付順位和検定

各都道府県のがんの罹患状況において，胃がん(Stmach)と大腸がん(Colon)に違いがあるだろうか？

Input

```
>  wilcox.test(dat$Stmach, dat$Colon,  paired=TRUE)
```

変数1　変数2

対応のある検定であることを宣言する

対応のある検定では変数を並べて書く

Output

```
 Wilcoxon signed rank exact test

data:  dat$Stmach and dat$Colon
V = 11, p-value = 7.816e-13
alternative hypothesis: true location shift is not equal to 0
```

①

①は，p値を表している．ここで，7.816e-13とは，7.816×10^{-13}を意味しており，とても小さな数値であるといえる．通常，p値は小数点以下3桁未満は切り捨て(あるいは四捨五入)する．

帰無仮説H_0: 47都道府県のがんの罹患者数において，胃がんと大腸がんのあいだで差がない.

対立仮説H_1: 47都道府県のがんの罹患者数において，胃がんと大腸がんのあいだで差がある.

p値 (有意確率)

例示: p値<0.001

p値が有意水準α(一般的にはα=0.05)未満

今回の場合

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論

胃がんと大腸がんで差がある.

p値が有意水準α(一般的にはα=0.05)以上

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論

胃がんと大腸がんで差があるとは言えない.

したがって，47都道府県のがん罹患者数において，胃がんと大腸がんでは違いが認められる.

1標本検定

今回のデータは，2019年度のものである．ちなみに，2016年度のすべてのがんり患者数(万人対)は，402.028人であった．
2019年度の今回のデータ(がん罹患者数(All_Cancer))のあいだに統計学的な違いがあるだろうか？

Input
```
> mean(dat$All_Cancer)
```
Output
```
[1] 386.1188
```

2019年度の平均値だけみれば，2016年に比べて減少している．しかしながら，統計学的に変化しているかどうかを判断するには，検定を用いなければならない．今回のように，あらかじめ決まっている指標とデータを比較するような場合を1標本検定という．今回の場合には，平均値を比較するが，これを1標本t検定という．

1標本t検定での仮説

帰無仮説H_0: 2019年度のがんの10万人対の罹患者数の平均値は2016年度の402.028と同じである．
対立仮説H_1: 2019年度のがんの10万人対の罹患者数の平均値は2016年度の402.028と異なる．

1標本t検定は，2標本t検定と同様に関数`t.test()`を用いればよい．

Input

```
>  t.test(dat$All_Cancer, mu0=402.028)
```

Output

```
 One Sample t-test

data:  dat$All_Cancer
t = 155.87, df = 46, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 381.1324 391.1052
sample estimates:
mean of x
 386.1188
```

①

②

③

①は，p値を表している．ここで，`2.2e-16`とは，2.2×10^{-16}を意味しており，とても小さな数値であるといえる．通常，p値は小数点以下3桁未満は切り捨て(あるいは四捨五入)する．このとき，0.001未満の場合には「<0.001」と記載することが多い．

②は，95%信頼区間というもので，平均値③の信頼性を表すものである．この区間のなかに，あらかじめ想定した値(今回の場合には，2016年のがん罹患者数402.028)を含まなければ，有意水準α=0.05のもとで有意になる．

帰無仮説H_0: 2019年度のがんの10万人対の罹患者数の平均値は2016年度の402.028と同じである.

対立仮説H_1: 2019年度のがんの10万人対の罹患者数の平均値は2016年度の402.028と異なる.

p値が有意水準α(一般的にはα=0.05)未満

p値 (有意確率)

例示: p値<0.001

p値が有意水準α(一般的にはα=0.05)以上

今回の場合

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論

2019年度のがん罹患率は402.028とは異なる

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論

2019年度のがん罹患率は402.028とは異なるとはいえない

つまり, 2019年のがん罹患者数は, 2016年の402.028と比べて, 統計学的な違いが認められた. また, 平均値が386.1188であることから, 減少傾向が認められた.

無相関性の検定

各都道府県のがんの罹患状況において，10万人あたりの病床数(Bed)と高齢者世帯の割合(Old)に相関関係があるだろうか．つまり，高齢者世帯の割合が高い都道府県ほど10万人あたりの病床数が多いかどうかを検討する．

相関係数を計算する

Input

```
>  cor(dat$Bed, dat$Old)
```

Output

```
[1] 0.4615325
```

相関係数が0.462であった．この相関係数から，相関係数が0である(10万人当たりの病床数と高齢者世帯の割合には相関関係がない)と統計学的に判断するための方法が，**無相関性の検定**である．

cor.test()

```
cor.test(X,Y, method)
```

X　　：変数1（ベクトル型）　　Y　：変数2(ベクトル型)

method　：相関係数の種類（pearson（デフォルト），spearman，kendallから選択）

Pearsonの積率相関係数(通常の相関係数)における無相関性の検定

Input

```
> cor.test(dat$Bed, dat$Old)
```

Output

```
        Pearson's product-moment correlation

data:  dat$Bed and dat$Old
t = 3.49, df = 45, p-value = 0.001093
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2010064 0.6610816
sample estimates:
      cor
0.4615325
```

① p値を表している

② 相関係数に対する95%信頼区間を表している.

③ Pearsonの積率相関係数を表している.

Spearmanの順位相関係数における無相関性の検定

Input

```
> cor.test(dat$Bed, dat$Old, method="spearman")
```

Output

```
 Spearman's rank correlation rho

data:  dat$Bed and dat$Old
S = 10882, p-value = 0.01028
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.3708479
```

① p値を表している

② Spearmanの順位相関係数を表している.

タイのデータがあるため，正確なp値が計算できないと警告が出る (無視してかまわない)

帰無仮説H_0: 10万人あたりの病床数と高齢者世帯の割合の相関係数は0である.

対立仮説H_1: 10万人あたりの病床数と高齢者世帯の割合の相関係数は0でない.

p値 (有意確率)

例示: p値=0.001 (Peason)

p値=0.010 (Spearman)

p値が有意水準α(一般的にはα=0.05)未満

p値が有意水準α(一般的にはα=0.05)以上

今回の場合

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論

病床数と高齢者世帯の割合は関連性がある

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論

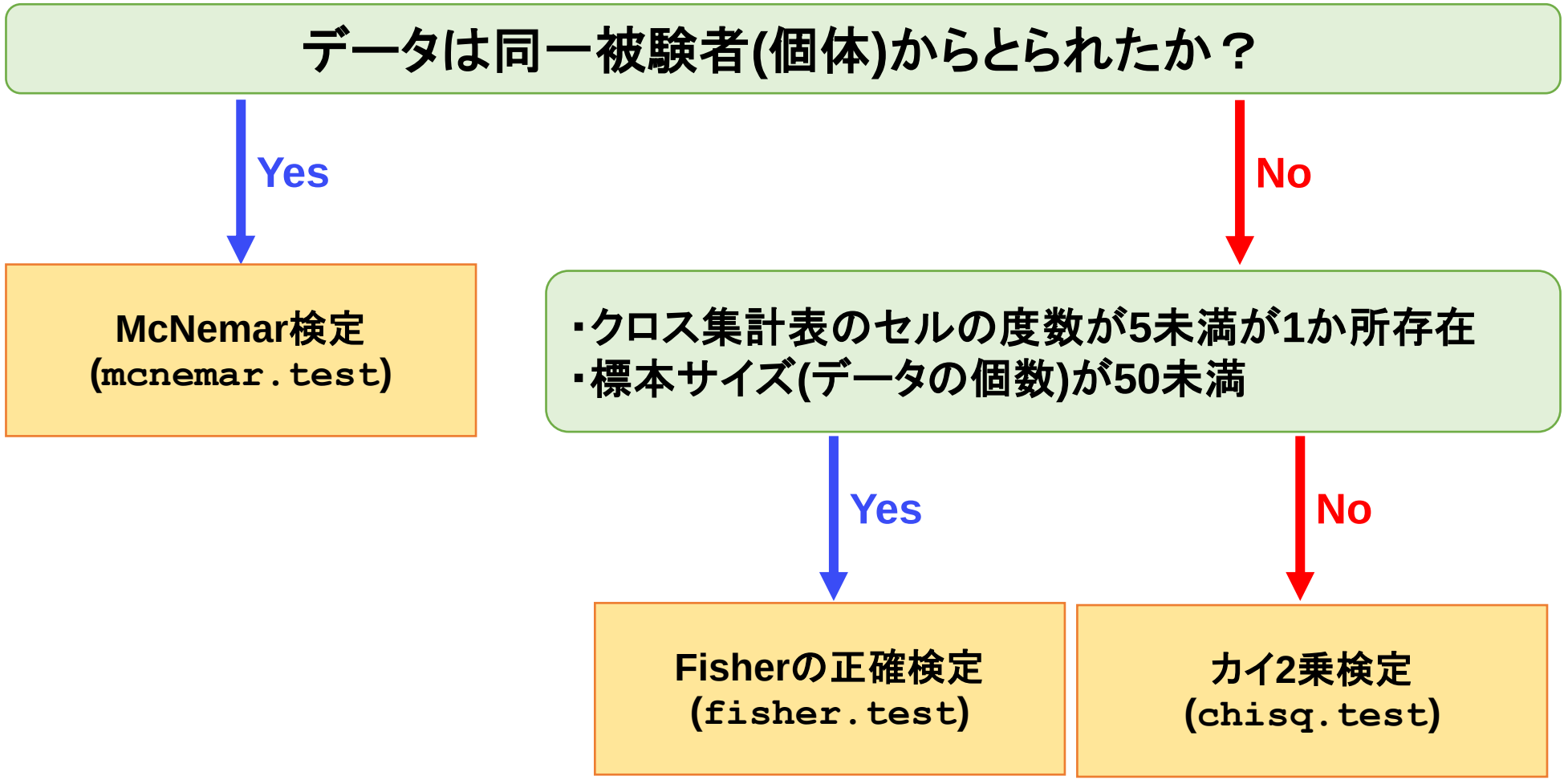
病床数と高齢者世帯の割合は関連性があるとは言えない

したがって, 47都道府県の10万人あたりの病床数と高齢者世帯の割合には関連性(相関関係)が認められる.

DAY.2：推測統計学入門

Section.2：カテゴリカル尺度の比較

質的変数における検定方法の取捨選択



クロス集計表の動機

ある都市の水道局は，取水地による水道水の満足度を調査するために，832人の住民を対象に取水地A(井戸水)と取水地B(河川)の水道水のおいしさの満足度に対して官能検査を実施した.

		変数2：満足度		
変数1：取水地		満足	不満足	計
	取水地A	291	125	416
	取水地B	270	146	416
	計	561	271	832

このときの関心は，「取水地によって水道水の満足度に違いがあるか？」である．このような判断を行うための仮設検定が**カイ2乗検定あるいはFisherの正確検定**である.

カイ2乗検定およびFisherの正確検定の意味

考え方(1)：因果関係の観点から見る場合

帰無仮説H_0：取水地の場所によって，水道水の満足度に違いがない．

対立仮説H_1：取水地の場所によって，水道水の満足度に違いがある．

原因　　結果

考え方(2)：関連性(独立性)の観点から見る場合

帰無仮説H_0：取水地の場所と水道水の満足度に関連性がない(独立である)．

対立仮説H_1：取水地の場所と水道水の満足度に関連性がある(独立でない)．

変数1　　変数2

仮説は若干異なるものの，

検定方法は同じ

カイ2乗検定の意味

● 帰無仮説H_0が正しいときのクロス集計表とデータから得られたクロス集計表の乖離度を応用する．

✖ p値を計算する際に近似を用いるため，データ数が少ない等の理由で近似精度が悪くなる可能性がある．

Fisherの正確検定の意味

● すべてのクロス集計表を考え，そのクロス集計表が得られる確率を数学的(確率的)に計算する．

✖ すべてのパターンのクロス集計表を計算するため，計算負荷がかかるため，データ数が多くなるとPCがフリーズする可能性がある．

カイ2乗検定およびFisherの正確検定の使い分け

	要因あり	要因なし	計
カテゴリA	12 (54.5%)	10 (45.5%)	22
カテゴリB	3 (20.0%)	12 (80.0%)	15
計	15 (40.5%)	22 (59.5%)	37

カイ2乗検定

p値 = 0. 0783

有意でない

Fisherの正確検定

p値 = 0.0471

有意である

カイ2乗検定では，p値を**近似で**計算する．そのため，
(1) 標本サイズが少ない場合 (50未満)
(2) どこかのセルの度数が小さい場合 (5未満)
には，近似精度が悪くなる．そのため，Fisherの正確検定を用いるほうが良い．一方で，標本サイズが大きくなると，Fisherの正確検定は計算できなくなる(パソコンがフリーズする)ので，カイ2乗検定を推奨する．なお，データの個数(標本サイズ)が増えると，カイ2乗検定のp値とFisherの正確検定のp値は一致する．

Rにおけるカイ2乗検定とFisherの正確検定

カイ2乗検定

chisq.test()

```
chisq.test(M)
```

M　:クロス集計表

Fisherの正確検定

fisher.test()

```
fisher.test(M)
```

M　:クロス集計表

さきほどの取水地のデータで計算してみる.

	満足	不満足
取水地A	291	125
取水地B	270	146

Input

```
> M <- matrix(c(291,270,125,146), ncol=2)
> M
```

Output

```
     [,1] [,2]
[1,]  291  125
[2,]  270  146
```

カイ2乗検定

Input

```
> chisq.test(M)
```

Output

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  M
X-squared = 2.189, df = 1, p-value = 0.139
```

① p値を表している

Fisherの正確検定

Input

```
> fihser.test(M)
```

Output

```
 Fisher's Exact Test for Count Data

data:  M
p-value = 0.1389
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.9315003 1.7018446
sample estimates:
odds ratio
  1.258487
```

① p値を表している

② オッズ比③に対する95%信頼区間を表している.

帰無仮説H_0：取水地の場所によって，水道水の満足度に違いがない．

対立仮説H_1：取水地の場所によって，水道水の満足度に違いがある．

p値 (有意確率)

例示: p値=0.139

p値が有意水準α(一般的にはα=0.05)未満

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論

取水地によって満足度に違いあり

p値が有意水準α(一般的にはα=0.05)以上

今回の場合

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論

取水地によって満足度に違いがあるとは言えない

取水地によって満足度に違いがあるとは言えなかった (取水地と満足度には関連性があるとはいえなかった)

補足:オッズ比とは

	満足	非満足	合
取水地A	291 (O_{11})	125 (O_{12})	p_A=0.700
取水地B	270 (O_{21})	146 (O_{22})	p_B=0.649

要因

結果

オッズとは，ある結果が生じる比率とある結果が生じない比率との比である.

取水地A(要因あり)のオッズ: $\mathrm{odds}_A = \frac{p_A}{1-p_A}$, 取水地B(要因なし)のオッズ: $\mathrm{odds}_B = \frac{p_B}{1-p_B}$

$$\frac{0.700}{1-0.700} = 2.333$$

$$\frac{0.649}{1-0.649} = 1.849$$

オッズ比$\mathrm{odds}_A/\mathrm{odds}_B$は，結果(アウトカム)に対してどれぐらい要因が寄与しているかを表す.

➡ 要因があるとない場合に比べて〇〇倍の結果が生じるか.

オッズ比の公式

$$OR = \frac{O_{11} \times O_{22}}{O_{12} \times O_{21}}$$

事例の場合には，$OR = \frac{O_{11} \times O_{22}}{O_{12} \times O_{21}} = \frac{291 \times 146}{125 \times 270} = 1.259$ なので,

取水地Aの水道水は取水地Bに比べて1.259倍満足されているといえる.

結果が2値応答の場合のもう一つの見方

要因		結果	
		満足	不満足
	取水地A	291	125
	取水地B	270	146

取水地を要因，満足度を結果と考えると

$$\text{取水地Aの満足度の割合} = \frac{291}{416} = 0.700$$

$$\text{取水地Bの満足度の割合} = \frac{270}{416} = 0.649$$

このように考えた場合，取水地Aと取水地Bの満足度の割合に違いがあるかどうかを統計学的に判断することも考えられる．このときに用いられるのが，母比率の差の検定である．

カイ2乗検定・Fisherの正確検定における仮説

帰無仮説H_0：取水地の場所によって，水道水の満足度に違いがない．

対立仮説H_1：取水地の場所によって，水道水の満足度に違いがある．

母比率の差の検定における仮説

帰無仮説H_0：取水地Aと取水地Bの満足度の割合に違いがない．

対立仮説H_1：取水地Aと取水地Bの満足度の割合に違いがある．

Rにおける母比率の差の検定

prop.test()

```
prop.test(c(r1, r2), c(N1, N2))
```

r1 ：群1の成功回数　　r2 ：群2の成功回数

N1 ：群1のデータ数（標本サイズ）　　N2 ：群2のデータ数（標本サイズ）

	満足	不満足
取水地A	291	125
取水地B	270	146

Input

```
> prop.test(c(291,270), c(291+125,270+146))
```

Output

```
        2-sample test for equality of proportions with continuity
correction

data:  c(291, 270) out of c(291 + 125, 270 + 146)
X-squared = 2.189, df = 1, p-value = 0.139
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.01551887  0.11648041
sample estimates:
   prop 1    prop 2
0.6995192 0.6490385
```

① ②

③

① p値を表している
② 各群の割合の差に対する95%信頼区間を表している.
③ 各群の割合を表している.

帰無仮説H_0：取水地Aと取水地Bの満足度の割合に違いがない.

対立仮説H_1：取水地Aと取水地Bの満足度の割合に違いがある.

p値が有意水準α(一般的にはα=0.05)未満

p値 (有意確率)

例示: p値=0.139

p値が有意水準α(一般的にはα=0.05)以上

今回の場合

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論

取水地Aと取水地Bの満足度割合に違いがある

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

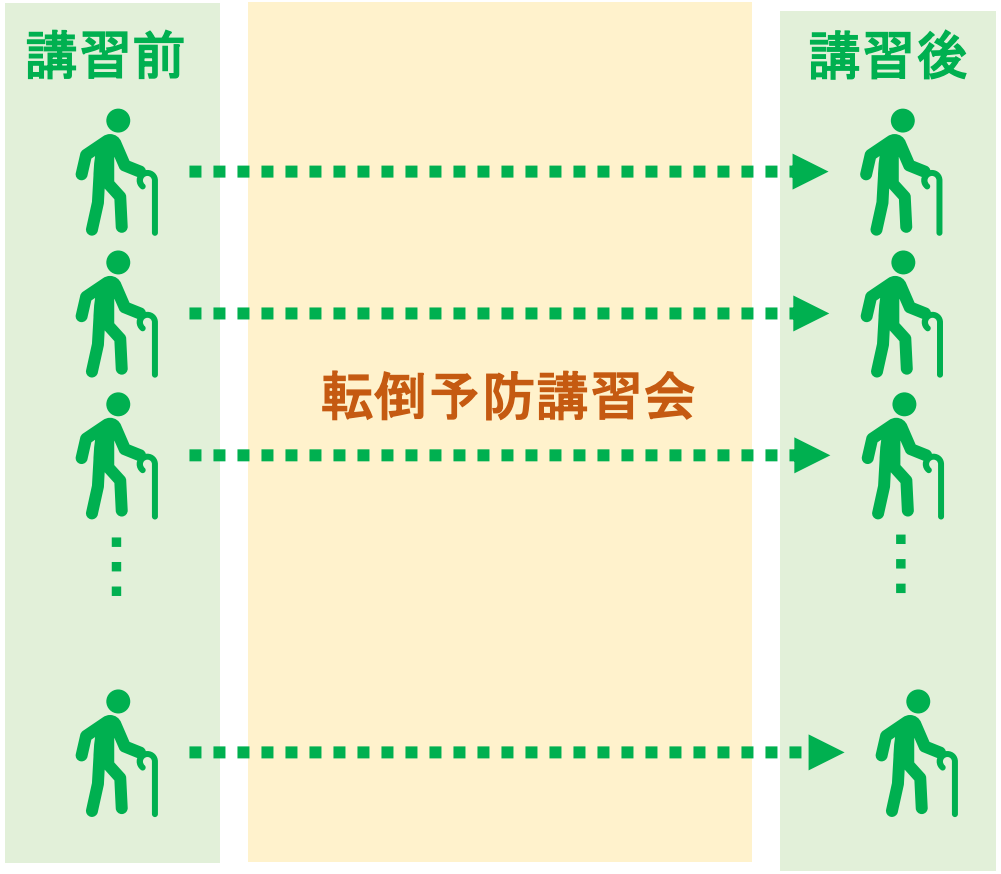
結論

取水地Aと取水地Bの満足度割合に違いがあるとはいえない

したがって，取水地Aと取水地Bの水道水の満足度には統計学的な違いは認められなかった.

対応のあるクロス集計表での検定：McNemar検定

ある都市では，65歳以上の高齢者を対象に，転倒防止のための運動講習会を3カ月にわたり開催した．この講習会では，講習会に先立って運動機能の評価(異常あり／なし)を行ったうえで，3カ月の講習会終了後にも評価を行った．



このデータでは，参加者のそれぞれにおいて，講習前の運動機能の異常の有無，講習後の運動機能の異常の有無がとられている．

このように，個々の被験者において2個のデータがとられている場合を対応のあるデータという．また，今回の場合には，以下の4パターンが存在する．

- 講習前に運動機能異常あり，講習後に運動機能異常あり
- ■講習前に運動機能異常あり，講習後に運動機能異常なし
- ■講習前に運動機能異常なし，講習後に運動機能異常あり
- 講習前に運動機能異常なし，講習後に運動機能異常なし

上記のパターンをクロス集計表で表したものを対応のあるクロス集計表という．

対応のあるクロス集計表

		講習後		合計
		異常あり	異常なし	
講習前	異常あり	40	**60**	100
	異常なし	**30**	70	100
合計		70	13	200

・講習前に運動機能異常あり，講習後に運動機能異常あり

40/200 =0.2 (20%)

■講習前に運動機能異常あり，講習後に運動機能異常なし

60/200 =0.3 (30%)

■講習前に運動機能異常なし，講習後に運動機能異常あり

30/200 =0.15 (15%)

・講習前に運動機能異常なし，講習後に運動機能異常なし

70/200 =0.35 (35%)

McNemar検定

帰無仮説H_0：講習会前後で運動機能異常の有無に変化が認められない

対立仮説H_1：講習会前後で運動機能異常の有無に変化が認められる．

McNemar検定では，■の割合(30%)と■の割合(15%)が比較される．

その結果，p値は0.002であった．したがって，歩行講習前後で運動機能に変化が認められた．また，歩行機能異常の参加者が異常なしになった割合のほうが高いことから，歩行講習の効果が認められた．

ちなみに，誤ってカイ2乗検定を用いた場合のp値が0.182なので，検定法を誤ることで，本来は有効であった歩行講習を有効でないと誤認してしまうので注意．

RにおけるMcNemar検定

ここでは，Cドライブの`Fukuoka_Seminor`というフォルダにある`data2.csv`というCSVファイルを読み込む．このデータを加工する．先ず，2018年度と2017年度を比較して上昇した市区町村にはup，そうでない市区町村にはdownとラベルとつけ，次いで，2019年度と2018年度を比較して，同様のラベルをつける．そのうえで，対応のあるクロス集計表をつくる．

Input

```
>  dat <- read.csv("C:/Fukuoka_Seminor/data2.csv",fileEncoding = "cp932")
>  H2018 <- H2019 <- rep("Down", nrow(dat))
>  H2018[dat$Y2018 > dat$Y2017] <- "Up"
>  H2019[dat$Y2019 > dat$Y2018] <- "Up"
>  M <- table(H2018,H2019)
>  M
```

Output

```
       H2019
H2018   Down Up
  Down    32 10
  Up      28  1
```

mcnemar.test()

`mcnemar.test(M)`

M　：対応のあるクロス集計表

Input

```
>  mcnemar.test(M)
```

Output

```
        McNemar's Chi-squared test with continuity correction

data:  M
McNemar's chi-squared = 7.6053, df = 1, p-value = 0.00582
```

帰無仮説H_0:2018年度と2019年度の交通事故件数の変化に違いがない.
対立仮説H_1: 2018年度と2019年度の交通事故件数の変化に違いがある.

p値が有意水準α(一般的にはα=0.05)未満

p値 (有意確率)
例示: p値=0.006

p値が有意水準α(一般的にはα=0.05)以上

今回の場合

帰無仮説H_0が間違っている (棄却)

有意である (有意差がある)

結論
2018年度と2019年度の交通事故件数の変化に違いがある

帰無仮説H_0が間違っているとは言えない (受容)

有意でない (有意差がない)

結論
2018年度と2019年度の交通事故件数の変化に違いがあるとはいえない

2018年度と2019年度で前年度比での変化の傾向に違いが認められた. 対応のあるクロス集計表より,
2019年度は2018年度に上昇傾向を示した市区町村でも減少傾向を示すところが増加している.

DAY.2：推測統計学入門

Section.3：プログラミング入門

2標本の比較において，データx，グループ変数Gが与えられたときに，各群のデータ数および平均(標準偏差)，2標本t検定でのp値を表示する関数two.sample.comp()を作る．

Rエディッターの内容

p値の計算(再掲)

```
pval.calc <- function(X, grp){
        res <- t.test(X~grp, var.equal=TRUE)$p.value
        return(round(res,3))
}
```

平均値(標準偏差)の計算

```
mean.sd.calc <- function(X, dig=3){{
        ms <- format(mean(X, na.rm=TRUE), digit=dig)
        sds <- format(sd(X, na.rm=TRUE), digit=dig)
        return(sprintf("%s(%s)", ms,sds))
}
```

本体の作成

```
two.sample.comp <- function(X,G,dig=3){
        gname <- sort(unique(G))
        X1 <- X[G==gname[1]]; X2 <- X[G==gname[2]]
        N1 <- length(X1) ; N2 <- length(X2)
        MS1 <- mean.sd.calc(X1)
        MS2 <- mean.sd.calc(X2)
        pval <- pval.calc(X,G)
        res <- data.frame(N1, Mean.SD1=MS1, N2, Mean.SD2=MS2, pvalue=pval)
        return(res)
}
```

実行すると次のように表示される

Input

```
> two.sample.comp(dat$All_Cancer, dat$WE)
```

Output

```
  N1  Mean.SD1 N2 Mean.SD2 pvalue
1 23 383(17.8) 24  389(16)  0.248
```

さらに一度に複数の変数を計算できるように改造

```
tsc.all <- function(dat,G,dig=3){
        P <- ncol(dat)
        cnames <- colnames(dat)
        res <- matrix(rep("",5*P),ncol=5)
        for (i in 1:P){
                Re <- two.sample.comp(dat[,i], G)
                for (j in 1:5){
                        res[i,j] <- Re[1,j]
                }
        }
        colnames(res) <- colnames(Re)
        rownames(res) <- cnames
        return(res)
}
```

Input

```
> tsc.all(dat[,7:13],dat$WE)
```

Output

```
           N1    Mean.SD1      N2    Mean.SD2      pvalue
All_Cancer "23" "383(17.8)"  "24" "389(16)"     "0.248"
Stmach     "23" "44.3(8.34)" "24" "41.8(8.03)" "0.31"
Liver      "23" "10.8(1.05)" "24" "13.8(1.66)" "0"
Cervix     "23" "13.1(1.91)" "24" "14.7(2.83)" "0.028"
Prostate   "23" "66.6(6.59)" "24" "68.4(4.61)" "0.278"
Colon      "23" "58.7(5.42)" "24" "56(4.04)"   "0.064"
Lung       "23" "40.9(3.12)" "24" "42.9(2.47)" "0.015"
```

対応のあるデータにおいて，データx，データyが与えられたときに，各群のデータ数および平均(標準偏差)，差の平均値および95%信頼区間，対応のあるt検定でのp値を表示する関数paired.mean.comp()を作る．

平均値(標準偏差)の計算 (使いまわし)

```
mean.sd.calc <- function(X, dig=3){{
        ms <- format(mean(X, na.rm=TRUE), digit=dig)
        sds <- format(sd(X, na.rm=TRUE), digit=dig)
        return(sprintf("%s(%s)", ms,sds))
}
```

本体の作成

```
paired.mean.comp <- function(X,Y,dig=3){
        Xsm <- mean.sd.calc(X)
        Ysm <- mean.sd.calc(Y)
        mdl <- t.test(X,Y,paired=T)
        Ms <- format(mdl$estimate, digit=dig)
        LW <- format(mdl$conf.int[1], digit=dig)
        UP <- format(mdl$conf.int[2], digit=dig)
        Dsm <- sprintf("%s[%s, %s]", Ms,LW,UP)
        pval <- round(mdl$p.value,3)
        res <- data.frame(X.mean.sd=Xsm, Y.mean.sd=Ysm, Dis.CI=Dsm, pvalue=pval)
        return(res)
}
```

実行すると次のように表示される

Input

```
> paired.mean.comp(dat$Stmach, dat$Lung)
```

Output

```
  X.mean.sd  Y.mean.sd            Dis.CI pvalue
1  43(8.19) 41.9(2.96) 1.12[-1.18, 3.42]  0.332
```

対応のあるデータにおいて，カテゴリカル・データx（縦（行）方向），カテゴリカル・データy（横（列）方向）が与えられたときに，クロス集計表(形相態度数を含む)，カイ2乗検定のp値を含む結果を表示する関数conj.chisq()を作る．

ここでは，Day.1で作成した関数Conject.table()を使いまわして作成する（付録に再掲する）．

本体の作成

```
conj.chisq <- function(X,Y){
          tbl <- table(X,Y)
          ctbl <- Conject.table(X,Y,2)
          pval <- round(chisq.test(tbl)$p.value)
          pvalue <- c(pval,"")
          res <- data.frame(ctbl,pvalue)
          return(res)
}
```

実行すると次のように表示される

Input

```
> dat <- read.csv("C:/Fukuoka_Seminor/data4.csv",fileEncoding = "cp932")

> conj.chisq(dat[,2],dat[,3])
```

Output

```
          雨          小雨          晴          曇    pvalue
昼  70 (39.11)  71 (51.82) 659 (56.62) 483 (65.45)      0
夜 109 (60.89)  66 (48.18) 505 (43.38) 255 (34.55)
```

Thank you for your kind attention

shimokaw@wakayama-med.ac.jp

toshibow2000@gmail.com

付録:Conject.table()

```
Conject.table <- function(X,Y,idx=NULL){
        dat <- data.frame(X,Y)
        tbl <- table(dat$X, dat$Y)
        pct <- prop.table(tbl, idx)*100
        N <- nrow(tbl)
        M <- ncol(tbl)
        res <- matrix(numeric(N*M), ncol=M)
        for (i in 1:N){
                for (j in 1:M){
                        res[i,j] <- sprintf("%d (%3.2f)", tbl[i,j], pct[i,j])
                }
        }
        colnames(res) <- colnames(tbl)
        rownames(res) <- rownames(tbl)
        return(res)
}
```