Rで統計解析入門

(5) 2 標本 t 検定と回帰分析

準備：データ「DEP」の読み込み

1. データ「DEP」を以下からダウンロードする
   http://www.cwk.zaq.ne.jp/fkhud708/files/dep.csv
2. ダウンロードした場所を把握する ⇒ ここでは「c:/temp」とする
3. R を起動し，2. の場所に移動し，データを読み込む
4. データ「DEP」から薬剤 A と B のデータを抽出

```
> setwd("c:/temp")                    # dep.csv がある場所に移動
> DEP <- read.csv("dep.csv")          # dep.csv を読み込む
> AB <- subset(DEP, GROUP != "C")     # 薬剤 A と B のデータを抽出
> AB$GROUP <- factor(AB$GROUP)        # 薬剤の水準を 2 カテゴリに
> head(AB)
 GROUP QOL EVENT DAY PREDRUG DURATION
1     A  15     1  50      NO        1
2     A  13     1 200      NO        3
:     :   :     :   :       :        :
```

準備：架空のデータ「DEP」の変数

- GROUP：薬剤の種類（A，B，C）⇒ A と B
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- EVENT：改善の有無（1：改善あり，2：改善なし）
  ⇒ QOLの点数が 5 点以上である場合を「改善あり」とする
- DAY：観察期間（数値，単位は日）
- PREDRUG：前治療薬の有無（YES：他の治療薬を投与したことあり，
  NO：投与したことなし）
- DURATION：罹病期間（数値，単位は年）

準備：架空のデータ「DEP」（一部）

GROUP	QOL	EVENT	DAY	PREDRUG	DURATION
A	15	1	50	NO	1
A	13	1	200	NO	3
A	11	1	250	NO	2
A	11	1	300	NO	4
A	10	1	350	NO	2
A	9	1	400	NO	2
A	8	1	450	NO	4
A	8	1	550	NO	2
A	6	1	600	NO	5
A	6	1	100	NO	7
A	4	2	250	NO	4
A	3	2	500	NO	6
A	3	2	750	NO	3
A	3	2	650	NO	7
A	1	2	1000	NO	8
A	6	1	150	YES	6
A	5	1	700	YES	5
A	4	2	800	YES	7
A	2	2	900	YES	12
A	2	2	950	YES	10
B	13	1	380	NO	9
B	12	1	880	NO	5
B	11	1	940	NO	2
B	4	2	20	NO	7
B	4	2	560	NO	2
B	5	1	320	YES	11
B	5	1	940	YES	3
B	4	2	80	YES	6
B	3	2	140	YES	7
B	3	2	160	YES	13

本日のメニュー

1. 平均値の比較と 2 標本 t 検定
2. 回帰分析と 2 標本 t 検定
3. 交絡と交互作用

QOL の平均値の比較

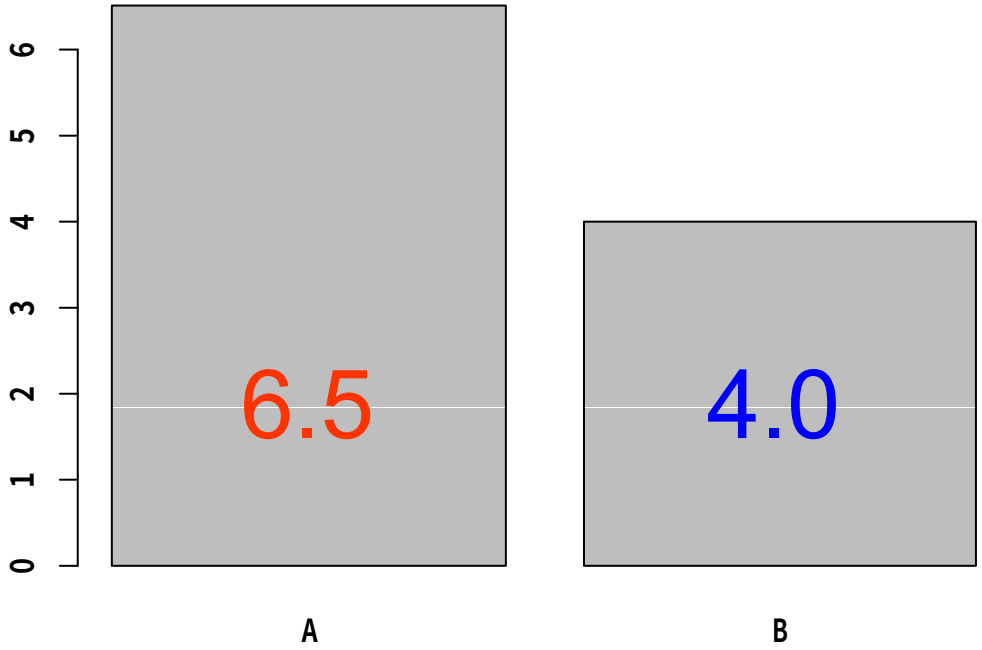
- 「薬剤 A（GROUP=A）の QOL の平均値」と
  「薬剤 B（GROUP=B）の QOL の平均値」の比較を行う
- まず，薬剤ごとに QOL の要約統計量を算出する

```
> by(AB$QOL, AB$GROUP, summary)

AB$GROUP: A
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    3.00    6.00    6.50    9.25   15.00
------------------------------------------------------------
AB$GROUP: B
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    2.00    3.00    4.00    4.25   13.00
```

QOL の平均値の比較

- 薬剤ごとに QOL に関するグラフ〔平均値の棒グラフ〕を描く

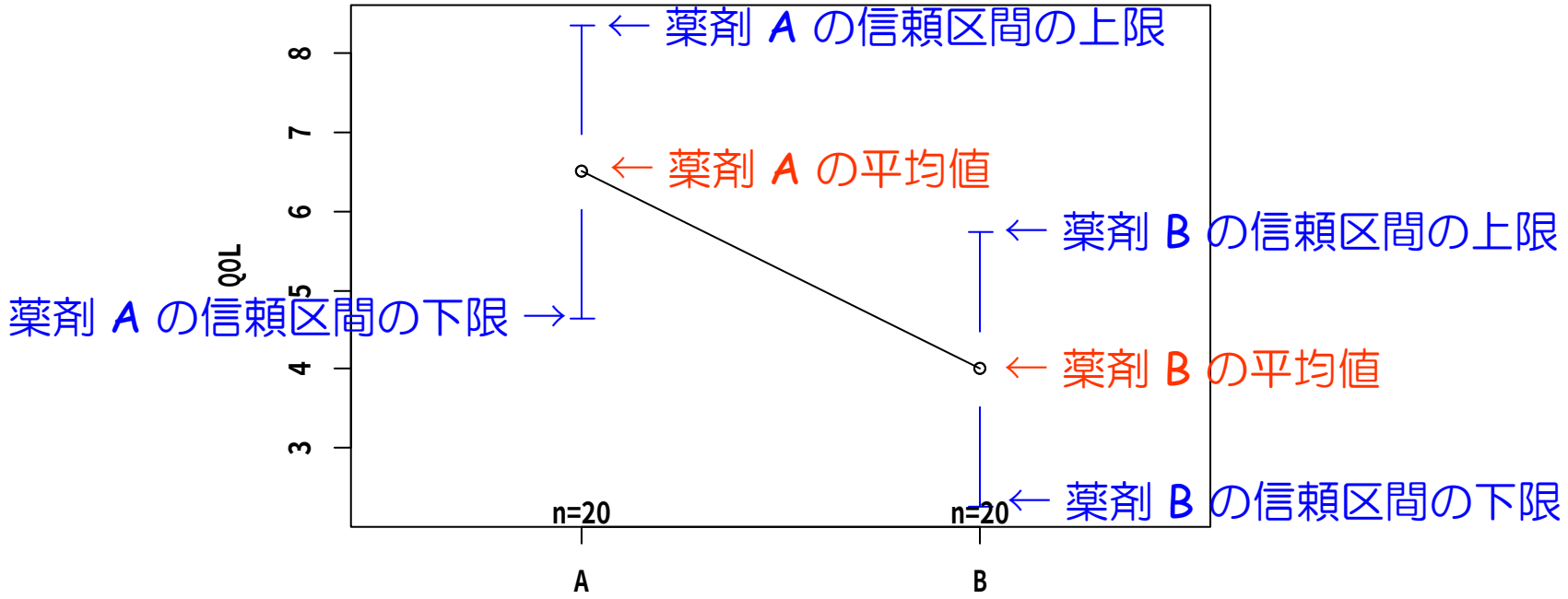
```
> MEAN <- by(AB$QOL, AB$GROUP, mean)  # 各薬剤の平均値を算出
> barplot(MEAN)                       # 平均値の棒グラフ
```



QOL の平均値の比較

- 薬剤ごとに QOL に関するグラフ〔平均・両側 95% 信頼区間〕を描く

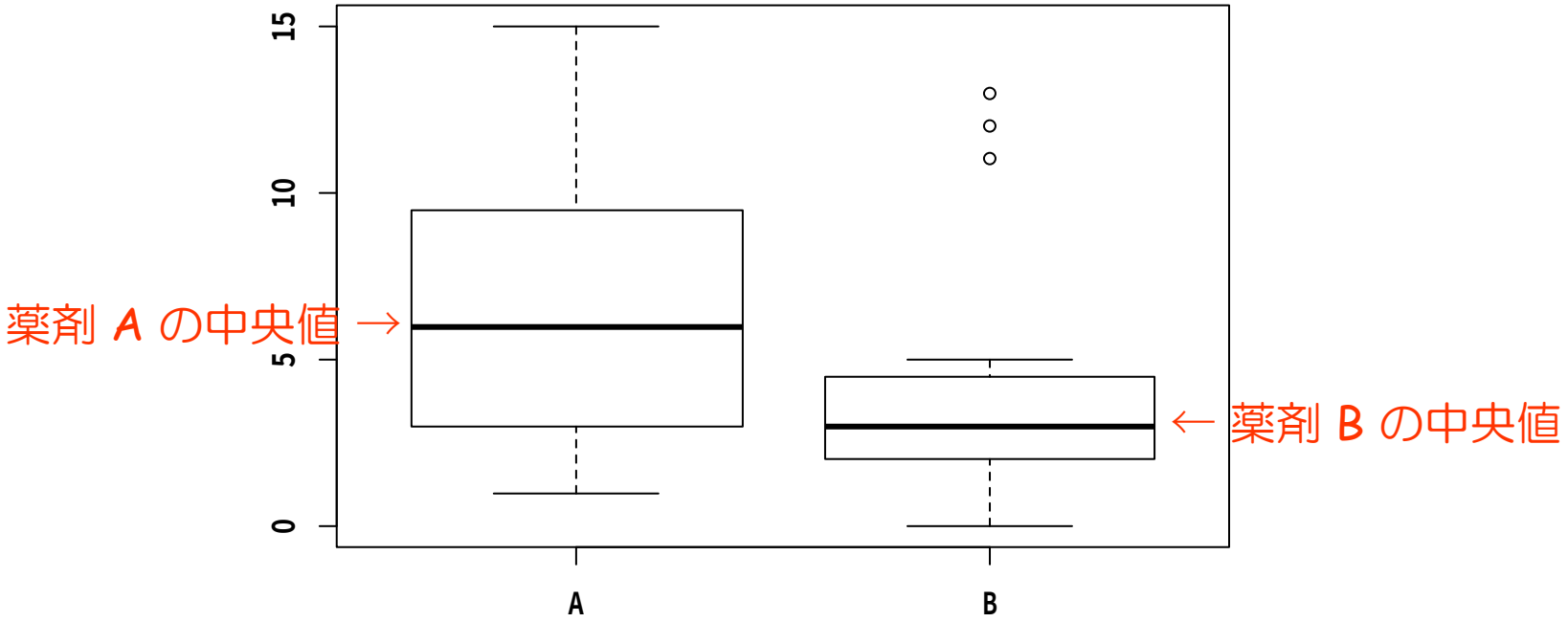
```
> install.packages("gplots")
> library(gplots)
> plotmeans(QOL ~ GROUP, data=AB)
```



【参考】QOL の中央値等の比較

- 薬剤ごとに QOL に関するグラフ〔箱ひげ図〕を描く

```
> boxplot(QOL ~ GROUP, data=AB)
```



【参考】QOL の中央値等の比較

- 薬剤ごとに QOL に関するグラフ〔箱ひげ図〕を描く

  ⇒ 外れ値を表示する

```
> boxplot(QOL ~ GROUP, data=AB)$out
[1] 13 12 11
```

QOL スコアに関する 2 標本 t 検定

- 「薬剤 A の QOL スコアの平均」と「薬剤 B の QOL スコアの平均」が等しいかどうかを検定する
  - p = 4.7%，有意水準 5% で検定すると結果は有意
  - 有意なので QOL スコアの平均は等しくない

```
> t.test(QOL ~ GROUP, data=AB, var=T)

        Two Sample t-test

data:  QOL by GROUP
t = 2.0503, df = 38, p-value = 0.04728 ← 検定結果（ p 値 = 約 4.7 %）
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.031532 4.968468
sample estimates:
mean in group A mean in group B
            6.5             4.0
```

QOL スコアに関する 2 標本 t 検定

1. 比較の枠組み ⇒薬剤 A と薬剤 B の QOL の平均を比較する
2. 比較するものの間に差がないという仮説（帰無仮説 H_0）を立てる
   ⇒ 帰無仮説 H_0 ：薬剤 A の平均 = 薬剤 B の平均
3. 帰無仮説とは裏返しの仮説（対立仮説 H_1）を立てる
   ⇒ 対立仮説 H_1 ：薬剤 A の平均 ≠ 薬剤 B の平均
4. 帰無仮説が成り立つという条件の下で，手元にあるデータ（よりも極端なこと）が起こる確率（= p 値）を計算 ⇒ p = 0.04728（4.7%）
5. 「確率が 4.7 %の珍しいデータが得られた」と考えずに
   「帰無仮説 H_0が間違っている」と考え，対立仮説 H_1 が正しいと結論
   ⇒「平均値は異なる」と解釈する
6. 「平均値は異なる」 &「薬剤 A の平均 = 6.5 ＞ 薬剤 B の平均 = 4.0」の合わせ技で「薬剤 A の平均 ＞ 薬剤 B の平均」と結論付ける

【余談】var=T って？？

```
> t.test(QOL ~ GROUP, data=AB, var=T)
```

- var=T：各薬剤のデータの分散は，薬剤間で等しいと仮定する
- var=F：各薬剤のデータの分散は，薬剤間で等しいと仮定しない
- var=F （分散は薬剤間で等しいと仮定しない）を用いる方が良いという意見が多い
- 臨床試験では var=T（分散は薬剤間で等しいと仮定する）を使用するのが慣例
  - 各薬剤の分散が異なっていたとしても，データの個数がほぼ等しい場合は var=T（分散は薬剤間で等しいと仮定する）を仮定した 2 標本 t 検定は誤った結論にならないという話がある（永田（1997））

【参考】対応のある t 検定

- 患者さんに治療を行う前の QOL の値を X，薬剤による治療を行った後の QOL の値を Y とする
- 治療前後の値（Y-X）がある値かどうかの検定を行う場合の帰無仮説：
  帰無仮説 H_0：治療前後の値（Y-X）が 0 である
- Z = Y - X に関する 1 標本 t 検定と同じ

```
> X <- round(2*rnorm(10, m=1)); Y <- round(2*rnorm(10, m=0))
> t.test(X, Y, mu=0, paired=T)

        Paired t-test
data:  X and Y
t = 4.3373, df = 9, p-value = 0.001885
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval: 1.578843 5.021157
sample estimates:
mean of the differences
                    3.3
```

【ネタフリ】QOL の平均値の比較【前治療の有無別】

- 前治療の有無別に，薬剤ごとの QOL の平均値を求める

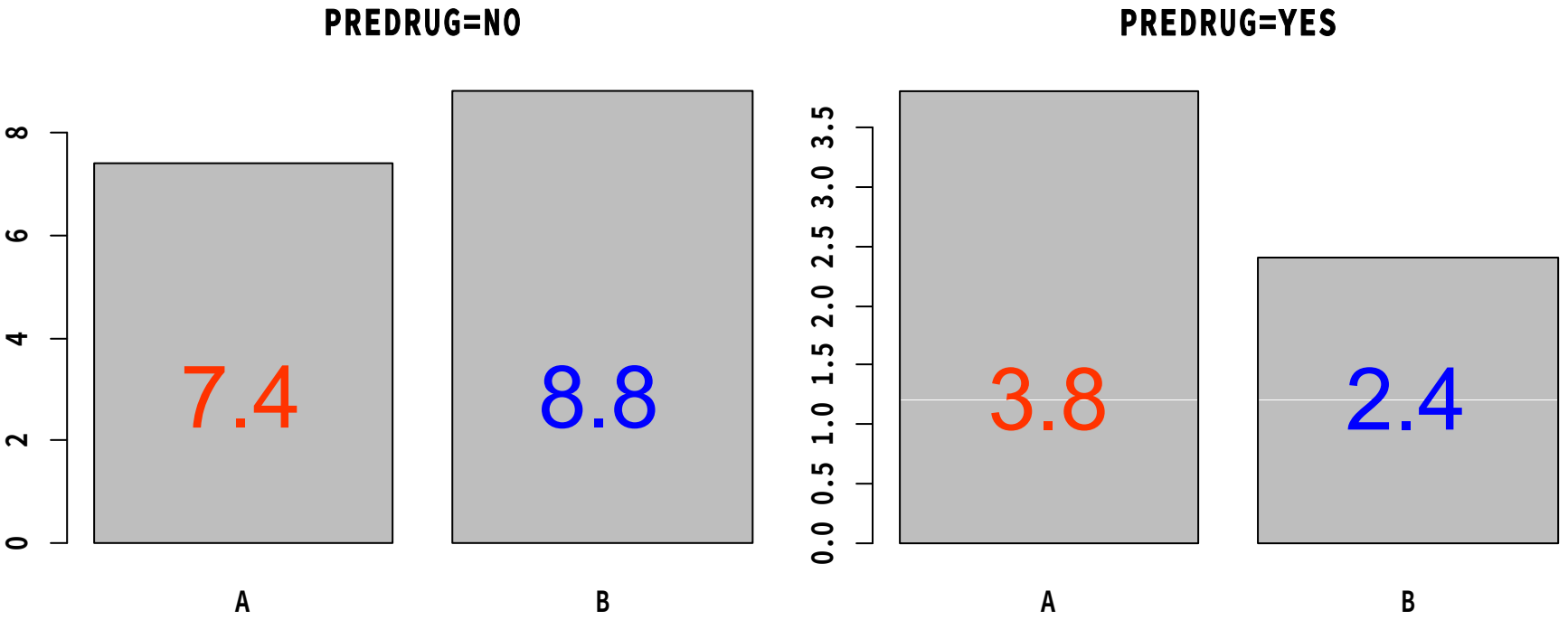
```
> MEAN2 <- tapply(AB$QOL, AB[,c("GROUP","PREDRUG")], mean)
> MEAN2

      PREDRUG
GROUP   NO YES
    A 7.4 3.8
    B 8.8 2.4
```

- 前治療なし：薬剤 A の平均 = 7.4，薬剤 B の平均 = 8.8 ⇒ B の方が高い
- 前治療あり：薬剤 A の平均 = 3.8，薬剤 B の平均 = 2.4 ⇒ A の方が高い

【ネタフリ】QOL の平均値の比較【前治療の有無別】

- 前治療の有無別に，薬剤ごとの QOL の平均値の棒グラフを描く

```
> barplot(MEAN2[,1], main="PREDRUG=NO")     # 前治療なし
> barplot(MEAN2[,2], main="PREDRUG=YES")    # 前治療あり
```

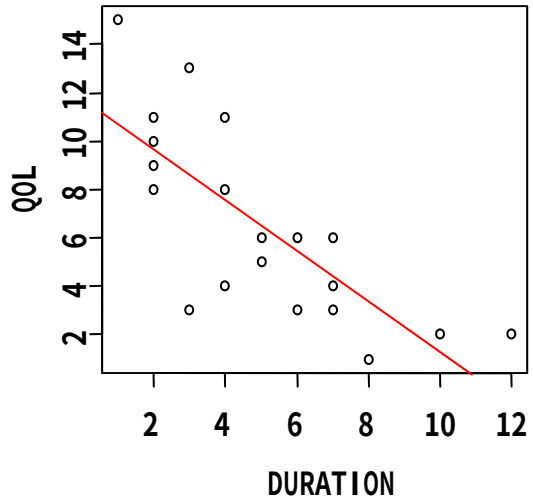


本日のメニュー

1. 平均値の比較と 2 標本 t 検定
2. 回帰分析と 2 標本 t 検定
3. 交絡と交互作用

回帰分析とは

- モデルによる解析手法
- 「回帰式」を用いて，ひとつの目的変数の値を複数の説明変数の値から推定する分析手法
- 前回，罹病期間（DURATION）と QOL がどんな関係かを調べる際に
  回帰式：QOL = 11.7 - 1.04×罹病期間（DURATION）
  を描き，罹病期間が 1 年増えた時に QOL がどう変わるかを推定した



回帰分析とは

- 回帰式の一般形：目的変数 ＝ β_0 ＋ β_1×説明変数1 ＋…＋ β_k×説明変数k
- 例えば以下のモデルより QOL を他の変数で推定することを考える
  QOL ＝ β_0 ＋ β_1×薬剤 ＋ β_2×前治療薬の有無 ＋ β_3×罹病期間
- 回帰分析を行った結果，回帰式が以下のように求まったとする
  QOL ＝ 1 ＋ 2×薬剤 ＋ 3×前治療薬の有無 ＋ 4×罹病期間
  - 薬剤：A ならば 1，B ならば 0，前治療薬の有無：なしならば 1，ありならば 0
- 「薬剤 A（1），前治療薬あり（0），罹病期間が 3 年」の人の QOL は上記の回帰式から以下のように推定される
  QOL ＝ 1 ＋ 2×1 ＋ 3×0 ＋ 4×3 ＝ 15

β_0：切片（Intercept）

QOL に関する回帰分析

- まずは以下のモデルについて回帰分析を行い回帰式を求める

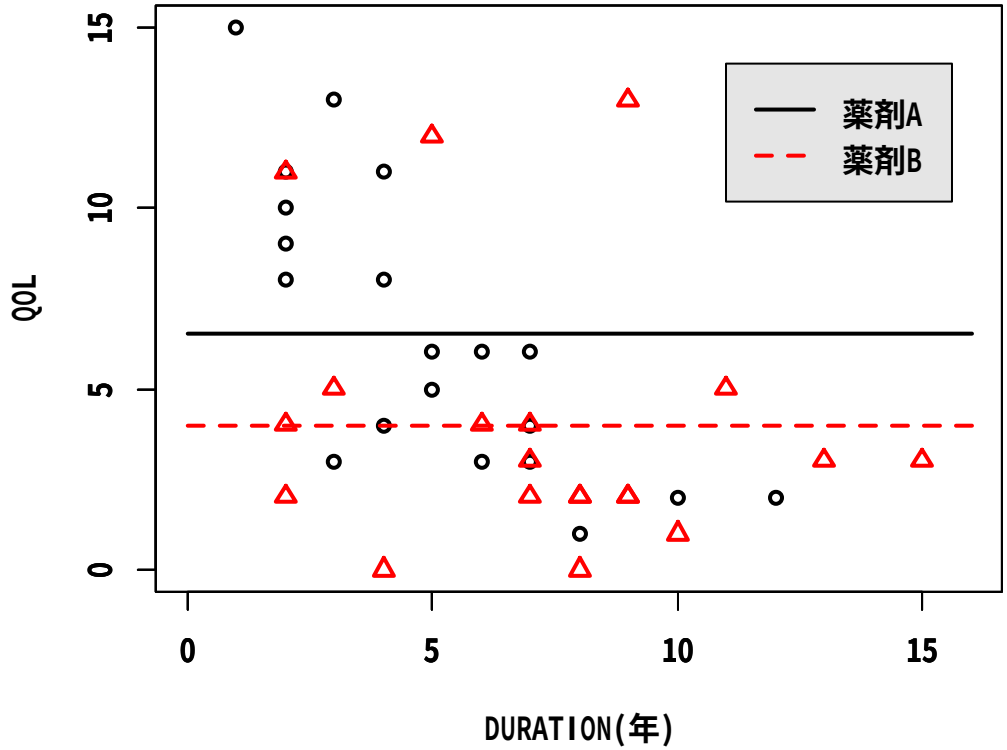
  QOL ＝ β_0 ＋ β_1×薬剤 ⇒ QOL ＝ 4.0 ＋ 2.5×薬剤 となった

```
> AB$GROUP <- relevel(AB$GROUP, ref="B") # ベースを「B」に変更
> result <- lm(QOL ~ GROUP, data=AB)     # 回帰分析
> summary(result)                        # 結果の要約を表示
 (中略)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0000     0.8622   4.639 4.07e-05 ***
GROUPA        2.5000     1.2194   2.050   0.0473 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.856 on 38 degrees of freedom
Multiple R-squared: 0.0996,     Adjusted R-squared: 0.07591
F-statistic: 4.204 on 1 and 38 DF,  p-value: 0.04728
```


β_0：切片（Intercept）

QOL に関する回帰分析：結果の解釈

- 回帰式は QOL ＝ 4.0 ＋ 2.5×薬剤 となった（薬剤：A=1, B=0）
  - 薬剤 A の回帰式：QOL ＝ 4.0 ＋ 2.5×1 ＝ 6.5
  - 薬剤 B の回帰式：QOL ＝ 4.0 ＋ 2.5×0 ＝ 4.0



※ 罹病期間をプロットする必要はありませんが，モデルに「罹病期間」を入れていないということは，「罹病期間が変動しても QOL は変わらない」という仮定を置いている，ということを示すために あえて罹病期間をプロットしています．あと，次回以降の話につなげるためでもあります．

前頁のグラフを描くプログラム

```
> # 散布図と回帰直線
> A <- function(x) 6.5+0*x
> B <- function(x) 4.0+0*x
> plot(QOL ~ DURATION, data=AB, pch=ifelse(GROUP=="A",1,2),
+      col=ifelse(GROUP=="A",1,2),
+      xlim=c(0,16), ylim=c(0,15), lwd=2, lty=1, ann=F)
> par(new=T)
> curve(A, xlim=c(0,16), ylim=c(0,15), lwd=2, col=1, lty=1, ann=F)
> par(new=T)
> curve(B, xlim=c(0,16), ylim=c(0,15), lwd=2, col=2, lty=2,
+       xlab="DURATION(年)", ylab="QOL")
> legend(11, 14, c("薬剤A ","薬剤B "), lwd=2, col=1:2, lty=1:2,
+        ncol=1, cex=1.0, bg="gray90")
```

QOL に関する回帰分析：結果の解釈

- 薬剤 A の回帰式：QOL ＝ 6.5 ⇒ 薬剤 A の QOL の平均値と一致
- 薬剤 B の回帰式：QOL ＝ 4.0 ⇒ 薬剤 B の QOL の平均値と一致
- 薬剤間のQOLの平均値の差：
  薬剤 A と薬剤 B の回帰式の引き算から，QOL の平均値の差が求まる
  ⇒ 回帰式の「薬剤の傾きの推定値（GROUPA：2.5）」を見ればよい

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0000     0.8622   4.639 4.07e-05 ***
GROUPA        2.5000     1.2194   2.050   0.0473 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```


Estimate：推定値

QOL に関する回帰分析：結果の解釈

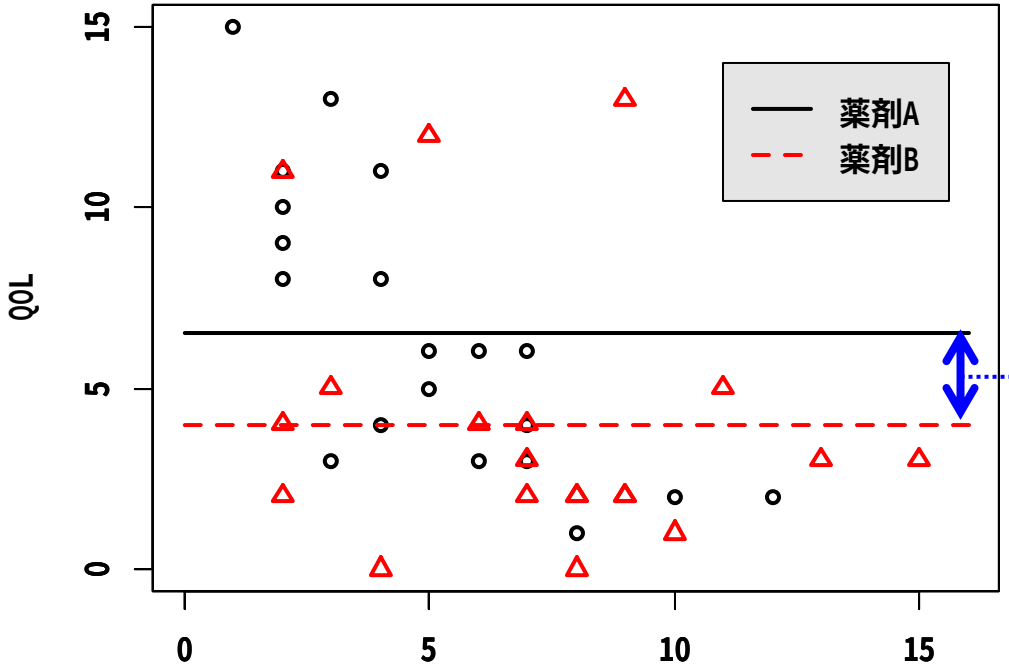
- 薬剤間の QOL の平均値の差（ GROUPAの推定値 ）は 2.5
- 薬剤間の QOL の平均値の差に対する「Pr(>|t|)」の意味：
- 「薬剤 B の QOL の平均値を 0 としたときの，
  薬剤 A のQOL の平均値が 0 かどうかの検定」の結果
  ＝「薬剤 A と薬剤 B の QOL の平均値の差が 0 かどうかの検定」
  ⇒ 結果は「Pr(>|t|)：0.0473」となっており，5% よりも小さくなっているので「帰無仮説は間違っている」と結論付ける
  ⇒ 薬剤間で QOL の平均値に差がある

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0000     0.8622   4.639 4.07e-05 ***
GROUPA        2.5000     1.2194   2.050   0.0473 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Estimate：推定値

QOL スコアに関する回帰分析 vs 2 標本 t 検定

- 回帰分析の結果：Pr(>|t|)：0.0473（実際は 0.04728）
- 薬剤 A と薬剤 B の QOL スコアの平均が等しいかどうかの
  2 標本 t 検定の結果： p-value = 0.04728
  ⇒ この場合の「回帰分析」と「2 標本 t 検定」は見てるものが同じ！



回帰分析の検定は
ここが 0 かどうかの検定
↓
この差は薬剤間の平均の差
↓
2標本t検定は平均の差
が0 かどうかの検定
↓
回帰分析と2標本t検定が一致

※ 罹病期間をプロットする必要はありません…

【参考】関数 relevel() を実行しないと・・・

- 薬剤：A ならば 0，B ならば 1 となる

  QOL ＝ β_0 ＋ β_1×薬剤 ⇒ QOL ＝ 6.5 － 2.5×薬剤 となる

```
> result <- lm(QOL ~ GROUP, data=AB)      # 回帰分析
> summary(result)                         # 結果の要約を表示
 (中略)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5000     0.8622   7.539 4.65e-09 ***
GROUPB       -2.5000     1.2194  -2.050   0.0473 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.856 on 38 degrees of freedom
Multiple R-squared: 0.0996,     Adjusted R-squared: 0.07591
F-statistic: 4.204 on 1 and 38 DF,  p-value: 0.04728
```


β_0：切片（Intercept）

【参考】切片なしのモデルで回帰分析

- QOL ＝ β_1×薬剤，というモデルでも解析可 ⇒「 -1 」をつける

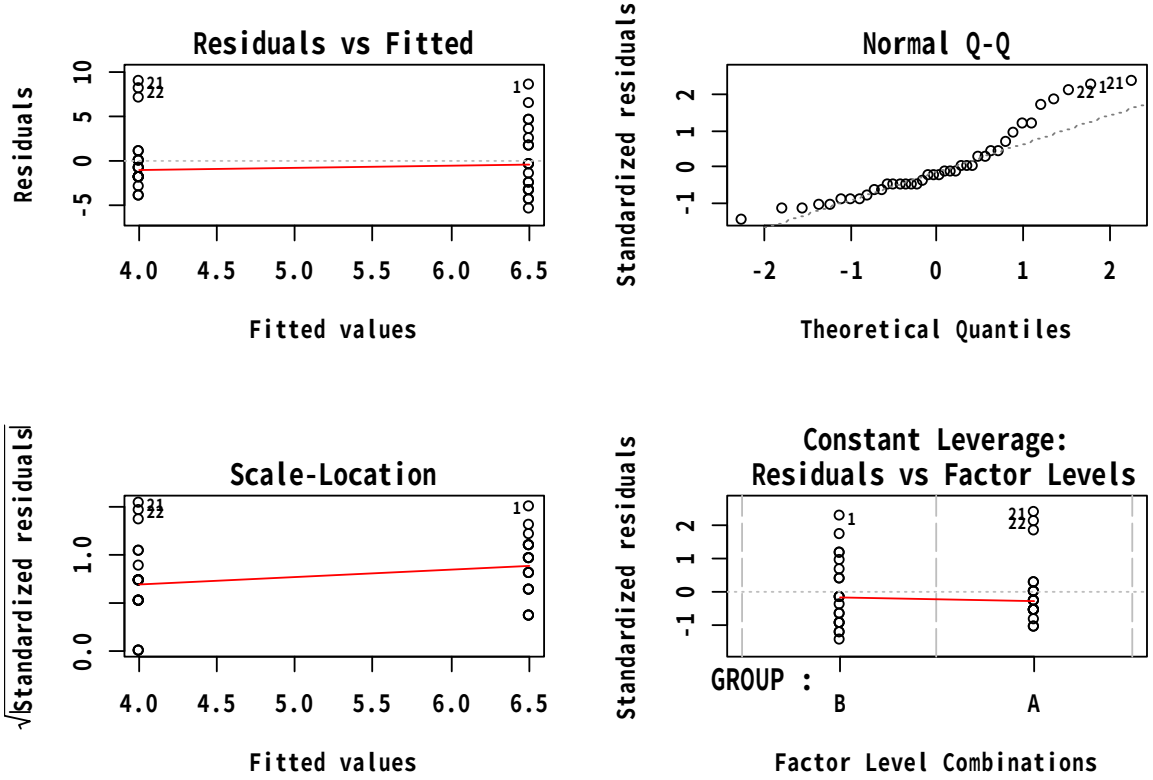
  ⇒ 薬剤 A の回帰式：QOL ＝ 6.5，薬剤 B の回帰式：QOL ＝ 4.0

```
> result <- lm(QOL ~ GROUP - 1, data=AB)      # 回帰分析（切片なし）
> summary(result)                             # 結果の要約を表示
 (中略)
Coefficients:
       Estimate Std. Error t value Pr(>|t|)
GROUPB   4.0000     0.8622   4.639 4.07e-05 ***
GROUPA   6.5000     0.8622   7.539 4.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.856 on 38 degrees of freedom
Multiple R-squared: 0.6734,     Adjusted R-squared: 0.6562
F-statistic: 39.18 on 2 and 38 DF,  p-value: 5.835e-10
```

【参考】QOL に関する回帰分析の回帰診断

```
> par(mfrow=c(2,2))
> plot(result)
```

 # グラフの画面を2×2に分割
 # 回帰診断のためのグラフ



左上：横軸が予測値，縦軸が残差（実測値ー予測値），当てはまりが悪いデータにラベルがつく
左下：横軸が予測値，縦軸が「基準化した残差の絶対値の平方根」
右上：「基準化した残差」のQQプロット，残差が正規分布に概ね従っている場合は点が直線にのる
右下：横軸が薬剤，縦軸が「基準化した残差」，モデルに対して影響が大きいデータにラベルがつく

【参考】前治療の有無別・薬剤別の要約統計量

- 前治療の有無別に，薬剤ごとの QOL の要約統計量を求める

```
> for (i in levels(AB$PREDRUG)) {
+   for (j in levels(AB$GROUP)) {
+     print( paste("PREDRUG:", i, ", GROUP:", j) )
+     print( summary(subset(AB, PREDRUG==i & GROUP==j,
+                           select=QOL)) )
+     cat("¥n")
+   }
+ }
```

【参考】前治療の有無別・薬剤別の要約統計量

```
[1] "PREDRUG: NO , GROUP: A"
      QOL
 Min.   : 1.0
 1st Qu.: 3.5
 Median : 8.0
 Mean   : 7.4
 3rd Qu.:10.5
 Max.   :15.0

[1] "PREDRUG: NO , GROUP: B"
      QOL
 Min.   : 4.0
 1st Qu.: 4.0
 Median :11.0
 Mean   : 8.8
 3rd Qu.:12.0
 Max.   :13.0
```

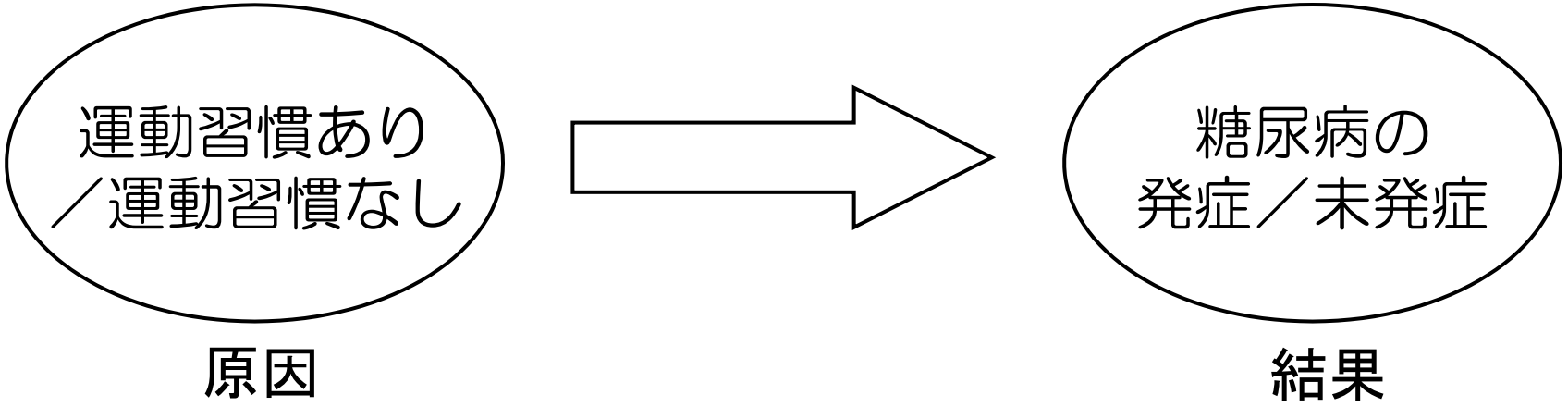
```
[1] "PREDRUG: YES , GROUP: A"
      QOL
 Min.   :2.0
 1st Qu.:2.0
 Median :4.0
 Mean   :3.8
 3rd Qu.:5.0
 Max.   :6.0

[1] "PREDRUG: YES , GROUP: B"
      QOL
 Min.   :0.0
 1st Qu.:2.0
 Median :2.0
 Mean   :2.4
 3rd Qu.:3.0
 Max.   :5.0
```

本日のメニュー

1. 平均値の比較と 2 標本 t 検定
2. 回帰分析と 2 標本 t 検定
3. 交絡と交互作用
   - 交絡と交絡因子
   - 交互作用と効果修飾因子

因果関係の例 (1)

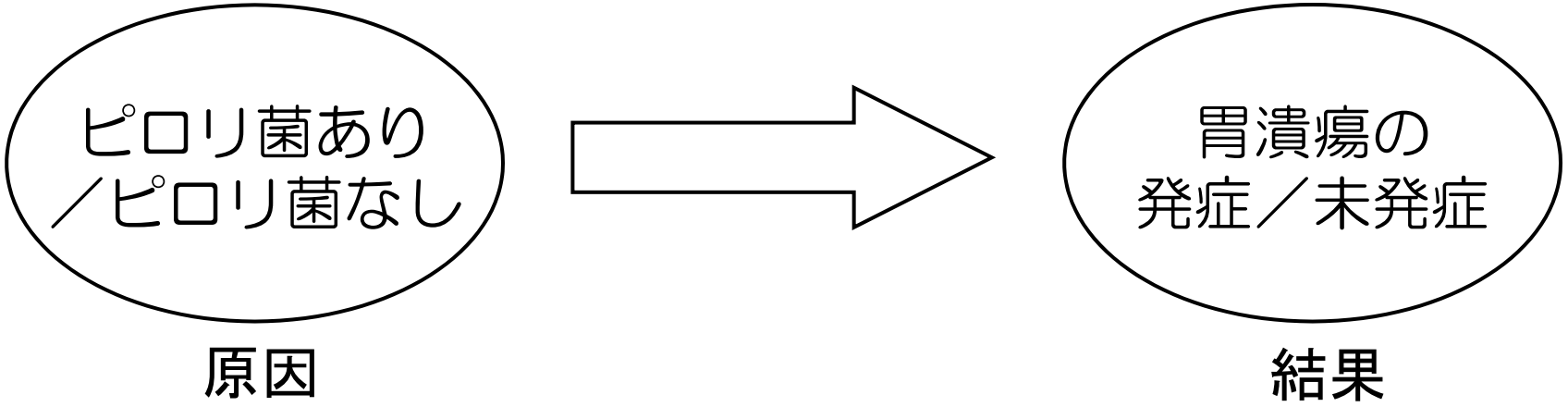
- 運動をする習慣がある ⇒ 糖尿病になりにくい
- 運動をする習慣がない ⇒ 糖尿病になりやすい



- 因果関係：原因（運動習慣の有無）と
  結果（糖尿病の発症）との関係

因果関係の例 (2)

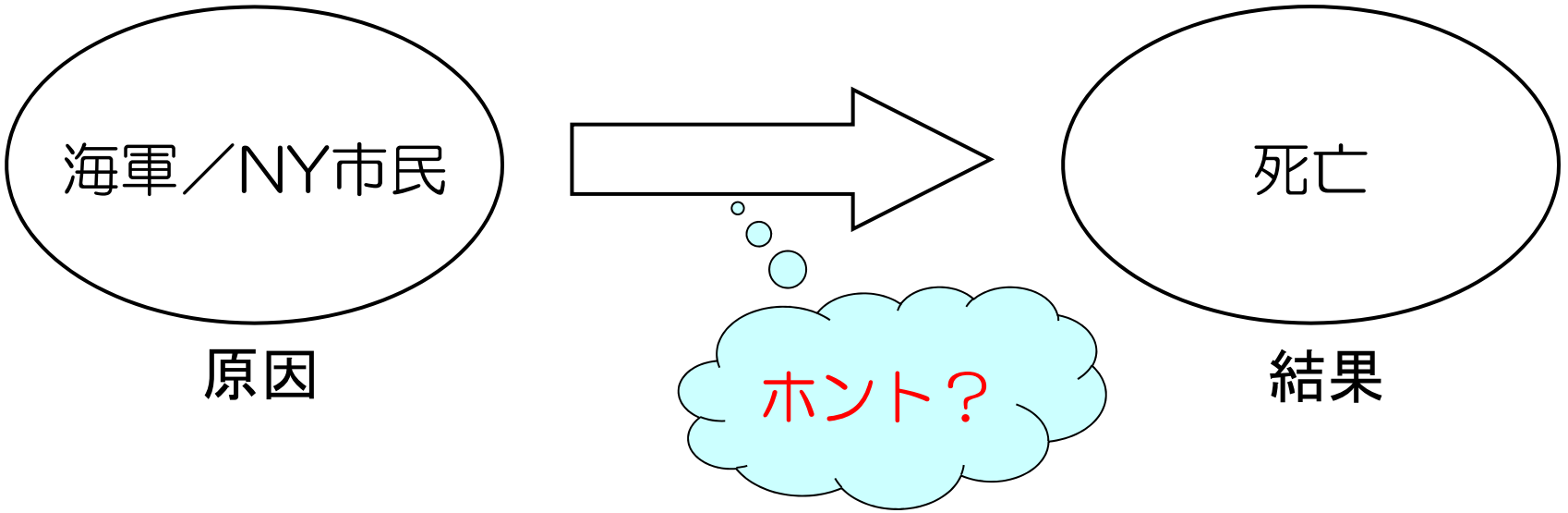
- お腹にピロリ菌がいる　⇒ 胃潰瘍になりやすい
- お腹にピロリ菌がいない ⇒ 胃潰瘍になりにくい



- 因果関係：原因（ピロリ菌の有無）と
  結果（胃潰瘍の発症）との関係

問題 (1)：米海軍のある宣伝

- 米海軍の死亡率：1000 人につき 9 人
- ニューヨーク市の死亡率 ：1000 人につき 16 人
- 実は，ニューヨーク市民よりも米海軍の方が死亡率が低いのです．海軍は安全です！皆さん海軍に入りましょう！

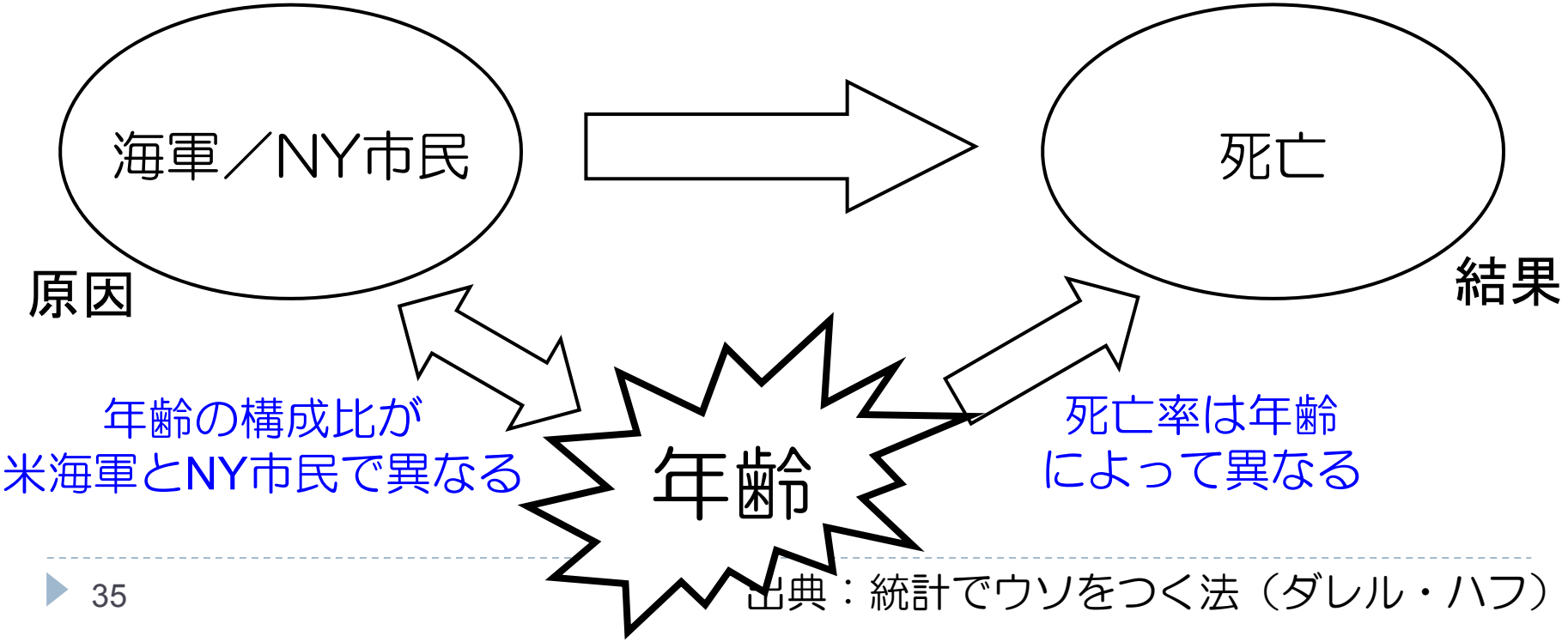


出典：統計でウソをつく法（ダレル・ハフ）

解答：両群の年齢構成が異なります

- 米海軍　：大部分が太鼓判付きの健康青年
- NY市民：赤ん坊もいればお年寄りや病気の方もいる

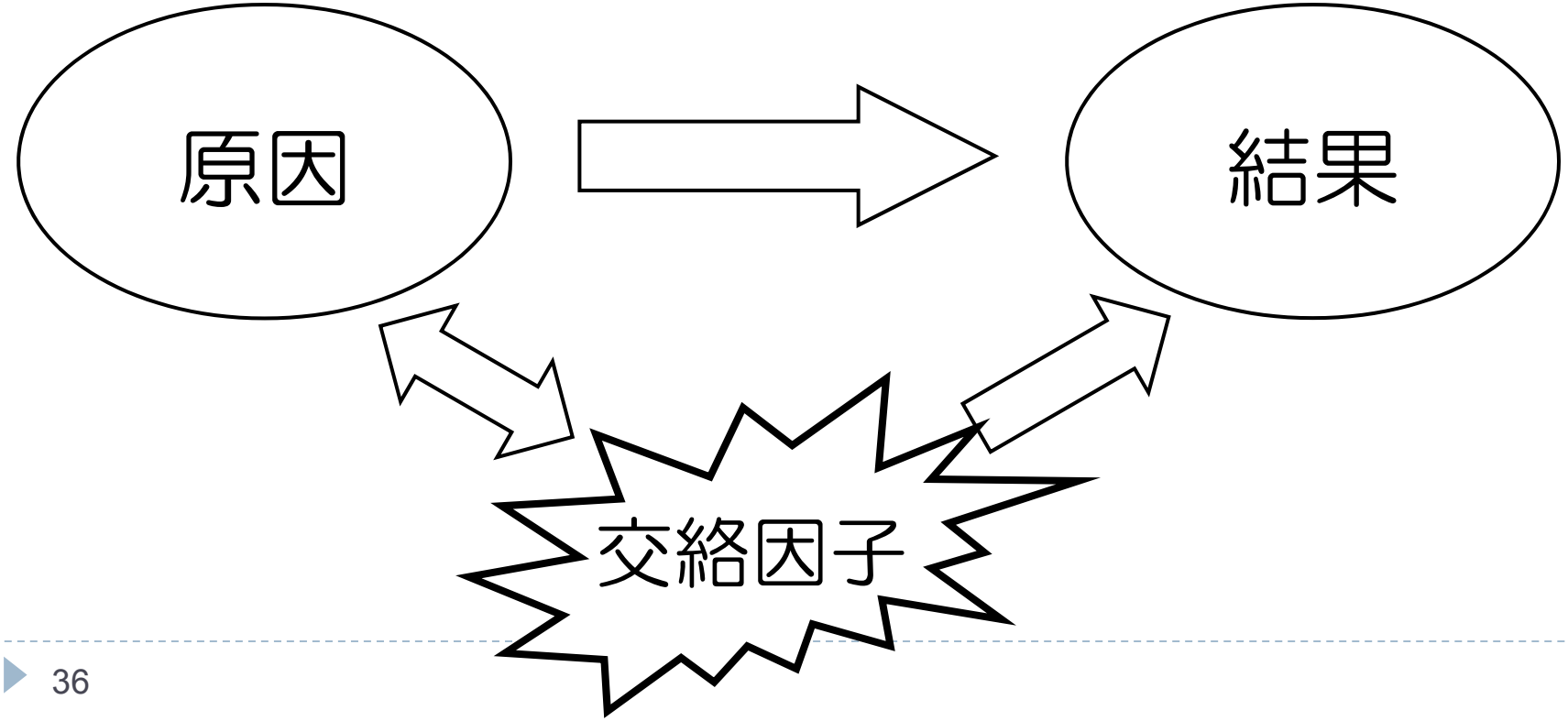
⇒ 死亡に強く影響している「年齢」という要因を無視すると
おかしな結論を招く



出典：統計でウソをつく法（ダレル・ハフ）

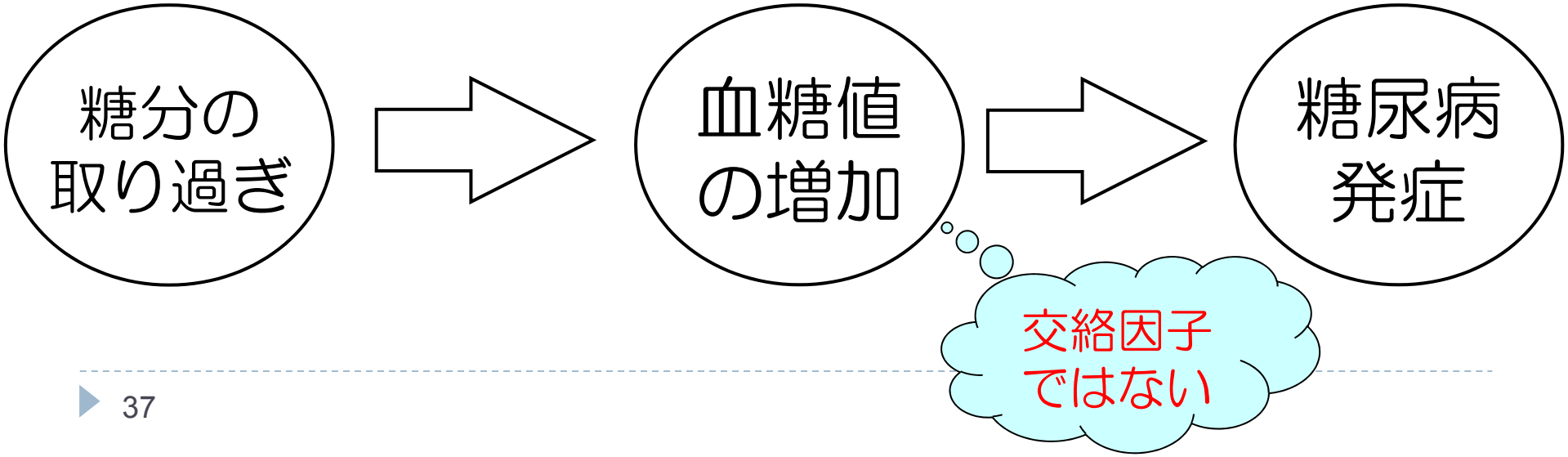
交絡と交絡因子

- 交絡：原因の結果への影響を調べる際，この 2 つの両方に影響を及ぼす因子があるため原因と結果の関係が正しく解釈できない状態
- 交絡因子：原因と結果の両方に影響を及ぼす因子
  ⇒ 原因と結果の関係（因果関係）が正しく解釈できない要因になりえる



交絡と交絡因子

- 交絡因子：原因と結果の両方に影響を及ぼす因子
  - ただし，その因子が原因と結果の中間の因子（結果の一部）である場合は，その因子は交絡因子ではない
    ⇒ 例えば，糖分の取り過ぎが糖尿病発症に影響するかどうか，を見る際，血糖値の増加は「糖分の取り過ぎ ⇒ 糖尿病発症」の関係の交絡因子ではない
    ⇒「糖分の取り過ぎ」→「血糖値の増加」→「糖尿病発症」なので「血糖値の増加」は中間の因子（交絡因子ではない）



問題 (2)：群間比較試験

- 薬剤 T と薬剤 C の群間比較試験 ⇒ 投与後の「効果の有無」を確認
- 薬剤 T を飲んでもらうか薬剤 C を飲んでもらうかは「適当に」わりふる
- 結果を見ると薬剤 T が効果がありそう

	効果あり	効果なし	計
T	80	30	110
C	70	40	110

問題 (2)：群間比較試験

	効果あり	効果なし	計
T	80	30	110
C	70	40	110

もしも男女比が均等でなかったら？

	女性	男性
T	100	10
C	10	100

男女別に結果を見てみる

	効果あり	効果なし	計
T	80	30	110
C	70	40	110

女性

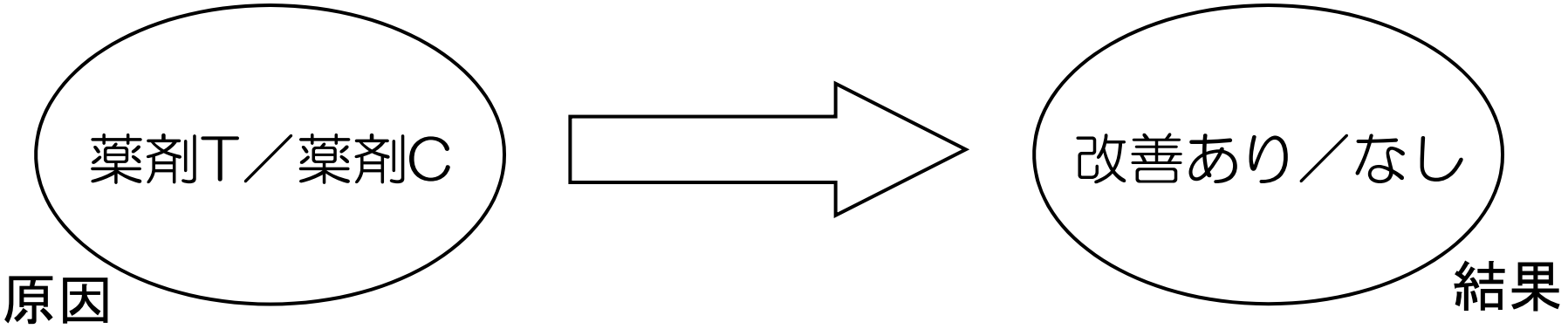
	効果あり	効果なし	計
T	75	25	100
C	8	2	10

男性

	効果あり	効果なし	計
T	5	5	10
C	62	38	100

問題 (2)：群間比較試験

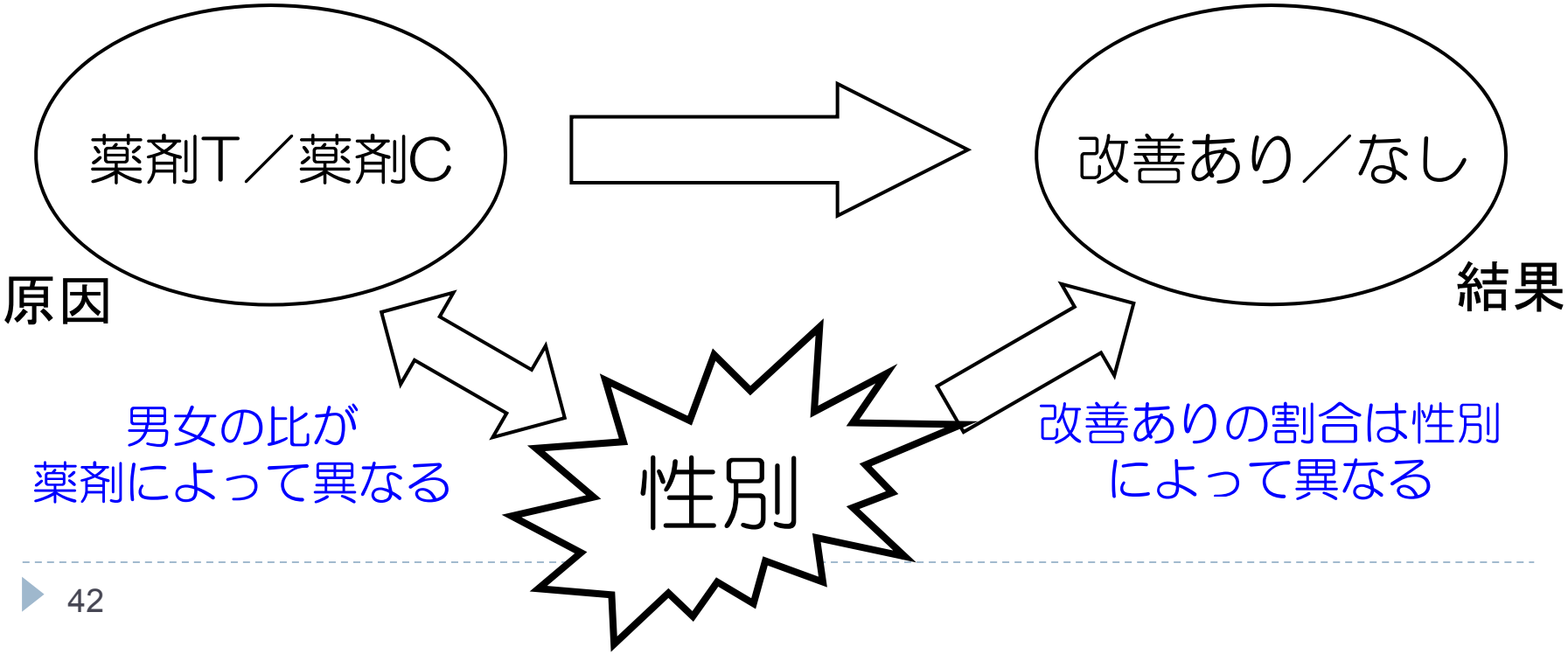
- どちらの薬剤の方が効果があるか？
- 下の因果関係の図は正しいか？
- 全体の結果と男女別の結果が異なった原因は何か？



解答：男女とも薬剤 C の方が効いている

- 男女の比が薬剤によって異なる：T は 女性：男性 = 10：1，C は 1：10
- 改善ありの割合が薬剤によって異なる
  女性：T は 75%，C は 80%，男性：T は 50%，C は 62%

⇒ 改善ありの割合に影響している「性別」という要因を無視して
（男女まとめて全体だけで）解釈をするとおかしな結論を招く



【もしも】男女の比が同じだったら? ⇒ 交絡は起きない

	効果あり	効果なし	計
T	125	75	200
C	142	58	200

まともな結果

女

	効果あり	効果なし	計
T	75	25	100
C	80	20	100

男

	効果あり	効果なし	計
T	50	50	100
C	62	38	100

ある因子が交絡因子かどうかの判定方法

興味のある因子が薬剤，「性別」が交絡因子かどうかを判定する場合...

1. 薬剤別で平均や割合などの要約統計量を求める（全体の結果）＆
   薬剤別・性別（男女別）で要約統計量を求める（層別の結果）
   ⇒ 以下の条件を両方とも満たす場合，「性別」は交絡因子
   - 薬剤間で性別の分布（男女比）が異なる（問題 2 の様な場合）
   - 全体の結果と層別の結果が異なる（問題 2 の様な場合）

ある因子が交絡因子かどうかの判定方法

興味のある因子が薬剤，「性別」が交絡因子かどうかを判定する場合...

2. 以下のモデルで回帰分析し，薬剤の効果（薬剤に関する傾き β_1 ）が変わる場合，「性別」は交絡因子
   - 「薬剤のみ」のモデル　：$QOL = \beta_0 + \beta_1 \times$薬剤
   - 「薬剤＋性別」のモデル：$QOL = \beta_0 + \beta_1 \times$薬剤 $+ \beta_2 \times$性別

交絡がない例：データセット AB_DUMMY

- GROUP：薬剤の種類（A，B）
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- GENDER：性別（1：男性，2：女性）

```
> set.seed(777)                                                    # 乱数のシード
> GROUP  <- c( rep("A",50), rep("B",50) )                          # 薬剤
> GENDER <- 1+rbinom(100, 1, 0.5)                                  # 性別（1:男，2:女）
> QOL    <- ifelse(GROUP=="A", 2.0+2.0*rnorm(50, sd=1),
+                              1.0+0.5*rnorm(50, sd=1))
> AB_DUMMY <- data.frame(QOL=round(QOL), GROUP =GROUP, GENDER=factor(GENDER))
> head(AB_DUMMY, n=3)
  QOL GROUP GENDER
1   5     A      2
2   0     A      1
3   7     A      1
```

交絡がない例：データセット AB_DUMMY

```
> table(AB_DUMMY$GROUP, AB_DUMMY$GENDER)                    # 頻度集計

    1  2
  A 23 27
  B 31 19

> tapply(AB_DUMMY$QOL, AB_DUMMY$GROUP, mean)                # 全体の結果
   A    B
1.90 1.04

> tapply(AB_DUMMY$QOL, AB_DUMMY[,c("GENDER","GROUP")], mean)   # 層別の結果
      GROUP
GENDER        A        B
     1 2.000000 1.000000
     2 1.814815 1.105263
```

交絡がない例：データセット AB_DUMMY

	QOL の平均値	例数
A	1.90	50
B	1.04	50

男性

	平均値	例数
A	2.00	23
B	1.00	31

女性

	平均値	例数
A	1.81	27
B	1.10	19

- 「薬剤 A の男女比　1：1」 ≠「薬剤 B の男女比　3：2」
- 「男性の QOL の平均値の差」　「女性の QOL の平均値の差」
  ⇒ 交絡は起きていなさそう ⇒ 一応，回帰分析でも確かめる

交絡がない例：データセット AB_DUMMY

```
> result <- lm(QOL ~ GROUP, data=AB_DUMMY)                    # 薬剤のみのモデル
> summary(result)                                             # 結果の要約を表示
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9000     0.2032   9.348 3.15e-15 ***
GROUPB       -0.8600     0.2874  -2.992  0.00351 **

> result <- lm(QOL ~ GROUP+GENDER, data=AB_DUMMY)             # 薬剤＋性別のモデル
> summary(result)                                             # 結果の要約を表示
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9237     0.2586   7.439 4.09e-11 ***
GROUPB       -0.8670     0.2927  -2.962  0.00384 **
GENDER2      -0.0438     0.2936  -0.149  0.88172
```

- 薬剤のみのモデル ：群間差 = -0.860
- 薬剤+性別のモデル：群間差 = -0.867

  ⇒傾きがほとんど変わらないので交絡は起きていなさそう

交絡がある例：データセット AB_DUMMY

- GROUP：薬剤の種類（A，B）
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- GENDER：性別（1：男性，2：女性）

```
> set.seed(777)
> GROUP  <- c( rep("A",50), rep("B",50) )                  # 薬剤
> GENDER <- ifelse(GROUP=="A", ceiling(0.8+runif(50)),
+                              ceiling(0.4+runif(50)))
> QOL    <- ifelse(GROUP=="A", 3.5+2.0*GENDER+2.0*rnorm(50, sd=2),
+                              1.0+2.5*GENDER+2.0*rnorm(50, sd=2))
> AB_DUMMY <- data.frame(QOL=round(QOL), GROUP =GROUP, GENDER=factor(GENDER))
> head(AB_DUMMY)
  QOL GROUP GENDER
1  14     A      2
2   3     A      2
3  17     A      2
```

交絡がある例：データセット AB_DUMMY

```
> table(AB_DUMMY$GROUP, AB_DUMMY$GENDER)                    # 頻度集計

    1  2
  A  7 43
  B 31 19

> tapply(AB_DUMMY$QOL, AB_DUMMY$GROUP, mean)                # 全体の結果
   A    B
6.94 5.22

> tapply(AB_DUMMY$QOL, AB_DUMMY[,c("GENDER","GROUP")], mean) # 層別の結果
      GROUP
GENDER        A        B
     1 4.714286 3.774194
     2 7.302326 7.578947
```

交絡がある例：データセット AB_DUMMY

	QOL の平均値	例数
A	6.94	50
B	5.22	50

男性

	平均値	例数
A	4.71	7
B	3.77	31

女性

	平均値	例数
A	7.30	43
B	7.57	19

- 「薬剤 A の男女比　1：6 」 ≠「薬剤 B の男女比　3：2 」
- 「男性の QOL の平均値の差」≠ 「女性の QOL の平均値の差」
  ⇒ 交絡が起きているっぽい ⇒ 一応，回帰分析でも確かめる

交絡がある例：データセット AB_DUMMY

```
> result <- lm(QOL ~ GROUP, data=AB_DUMMY)                    # 薬剤のみのモデル
> summary(result)                                             # 結果の要約を表示
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9400     0.5649  12.286   <2e-16 ***
GROUPB       -1.7200     0.7988  -2.153   0.0338 *

> result <- lm(QOL ~ GROUP+GENDER, data=AB_DUMMY)             # 薬剤＋性別のモデル
> summary(result)                                             # 結果の要約を表示
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.02180    0.92844   4.332 3.61e-05 ***
GROUPB      -0.09124    0.86108  -0.106 0.915837
GENDER2      3.39326    0.88700   3.826 0.000231 ***
```

- 薬剤のみのモデル ：群間差 = -1.720
- 薬剤+性別のモデル：群間差 = -0.091

  ⇒傾きが変わっているので交絡が起きている

データセット「AB」の場合

- GROUP：薬剤の種類（A，B，C）⇒ A と B
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- PREDRUG：前治療薬の有無（YES：他の治療薬を投与したことあり，
  NO：投与したことなし）

```
> setwd("c:/temp")                          # dep.csv がある場所に移動
> DEP <- read.csv("dep.csv")                # dep.csv を読み込む
> AB <- subset(DEP, GROUP != "C")           # 薬剤 A と B のデータを抽出
> AB$GROUP <- factor(AB$GROUP)              # 薬剤の水準を 2 カテゴリに
> AB$GROUP <- relevel(AB$GROUP, ref="B")    # カテゴリのベースを「B」に変更
> head(AB, n=5)
  GROUP QOL EVENT DAY PREDRUG DURATION
1     A  15     1  50      NO        1
2     A  13     1 200      NO        3
3     A  11     1 250      NO        2
4     A  11     1 300      NO        4
5     A  10     1 350      NO        2
```

データセット「AB」の場合

```
> table(AB$GROUP, AB$PREDRUG)                          # 頻度集計

   NO YES
B   5  15
A  15   5

> tapply(AB$QOL, AB$GROUP, mean)                       # 全体の結果
  B   A
4.0 6.5

> tapply(AB$QOL, AB[,c("PREDRUG","GROUP")], mean)      # 層別の結果
       GROUP
PREDRUG   B   A
    NO  8.8 7.4
    YES 2.4 3.8
```

データセット「AB」の場合

	QOL の平均値	例数
A	6.5	20
B	4.0	20

前治療なし

	平均値	例数
A	7.4	15
B	8.8	5

前治療あり

	平均値	例数
A	3.8	5
B	2.4	15

- 「薬剤 A のなし：あり　3：1」　≠「薬剤 B のなし：あり　1：3」
- 「前治療なしの QOL の平均値の差」≠「前治療ありの QOL の平均値の差」
  ⇒ 交絡が起きているっぽい ⇒ 一応，回帰分析でも確かめる

データセット「AB」の場合

```
> result <- lm(QOL ~ GROUP, data=AB)                    # 薬剤のみのモデル
> summary(result)                                       # 結果の要約を表示
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.0000     0.8622   4.639 4.07e-05 ***
GROUPA        2.5000     1.2194   2.050   0.0473 *

> result <- lm(QOL ~ GROUP+PREDRUG, data=AB)            # 薬剤＋前治療の有無のモデル
> summary(result)                                       # 結果の要約を表示
Coefficients:
              Estimate  Std. Error t value Pr(>|t|)
(Intercept)  7.750e+00   1.129e+00   6.863 4.32e-08 ***
GROUPA       8.654e-16※  1.166e+00   0.000 1.000000
PREDRUGYES  -5.000e+00   1.166e+00  -4.287 0.000124 ***
```

- 薬剤のみのモデル　　：群間差 = 2.500
- 薬剤+前治療の有無のモデル：群間差 = ほぼ 0

  ⇒傾きが変わっているので交絡が起きている

※ 8.654e-16 = 8.654×10^{-16}　　0

本日のメニュー

1. 平均値の比較と 2 標本 t 検定
2. 回帰分析と 2 標本 t 検定
3. 交絡と交互作用
   - 交絡と交絡因子
   - 交互作用と効果修飾因子

交互作用とは

- 交互作用：複数の変数の組み合わせにより生じる作用のこと
- 交互作用がある：2 つの要因（例えば「薬剤×性別」）が互いに影響を及ぼし合っている状態のこと

  ⇒「薬剤×性別」を，「薬剤」と「性別」との交互作用を表すこととし交互作用項と呼ぶことにする

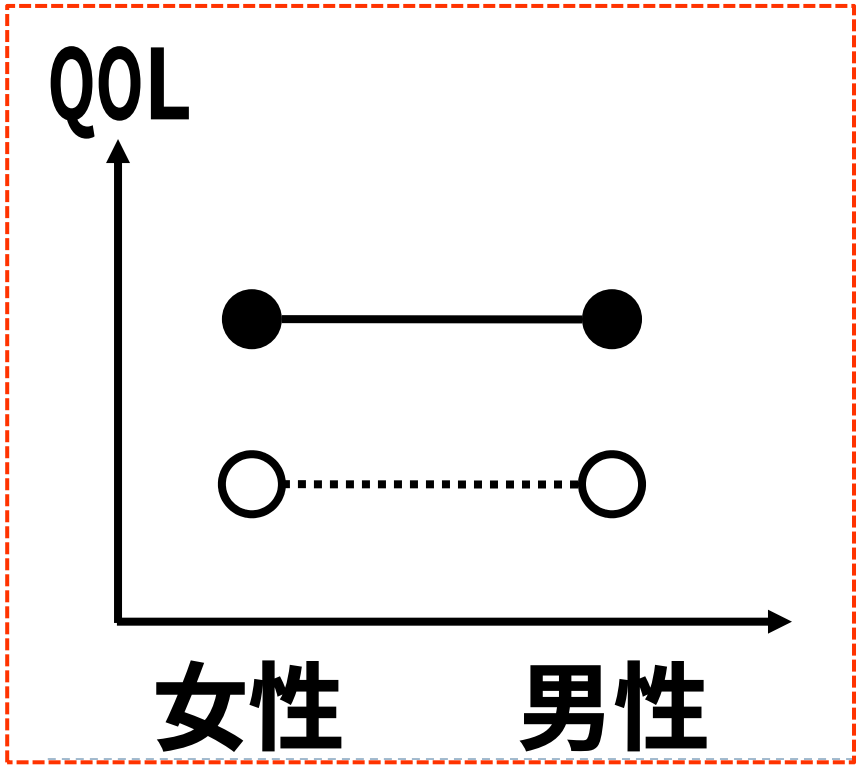
  ⇒「薬剤×性別」の交互作用がある場合，この要因である「性別」を効果修飾因子と呼ぶ

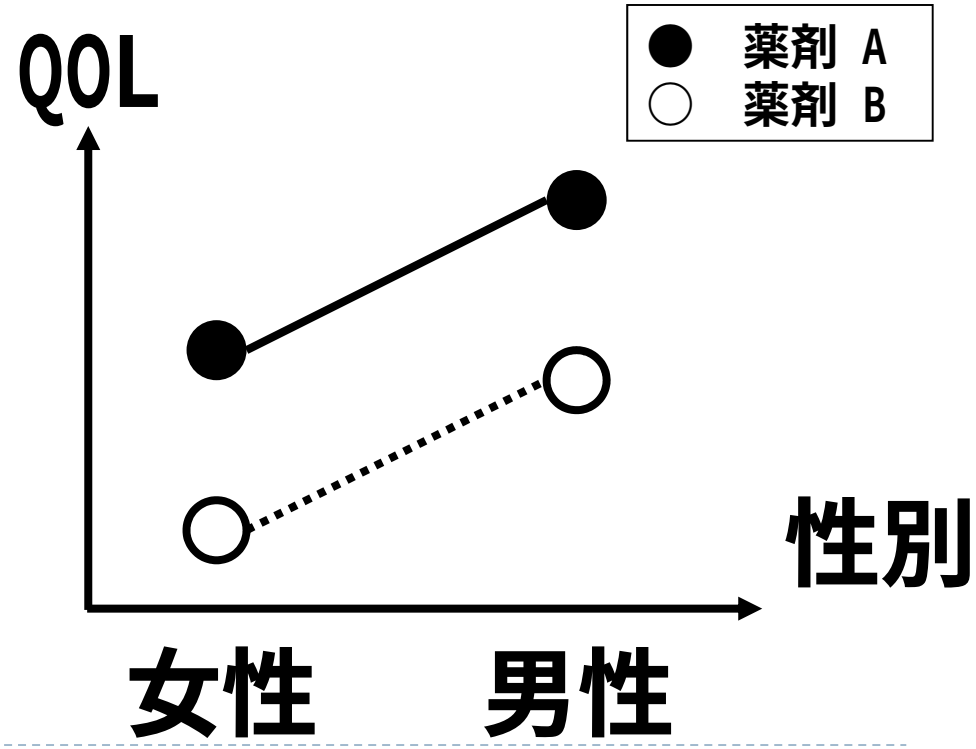
※本資料では「統計的交互作用」を「交互作用」と表現します
ロスマンの疫学等で出てくる「生物学的交互作用」は扱いません

交互作用がない状態（●，○：QOL の平均値）

- 左下の図は以下の特徴がある
  - 「薬剤×性別」の交互作用がない
  - 性別が QOL に影響を及ぼしていない

⇒ 女性も男性も，薬剤間の平均値の差は同じ

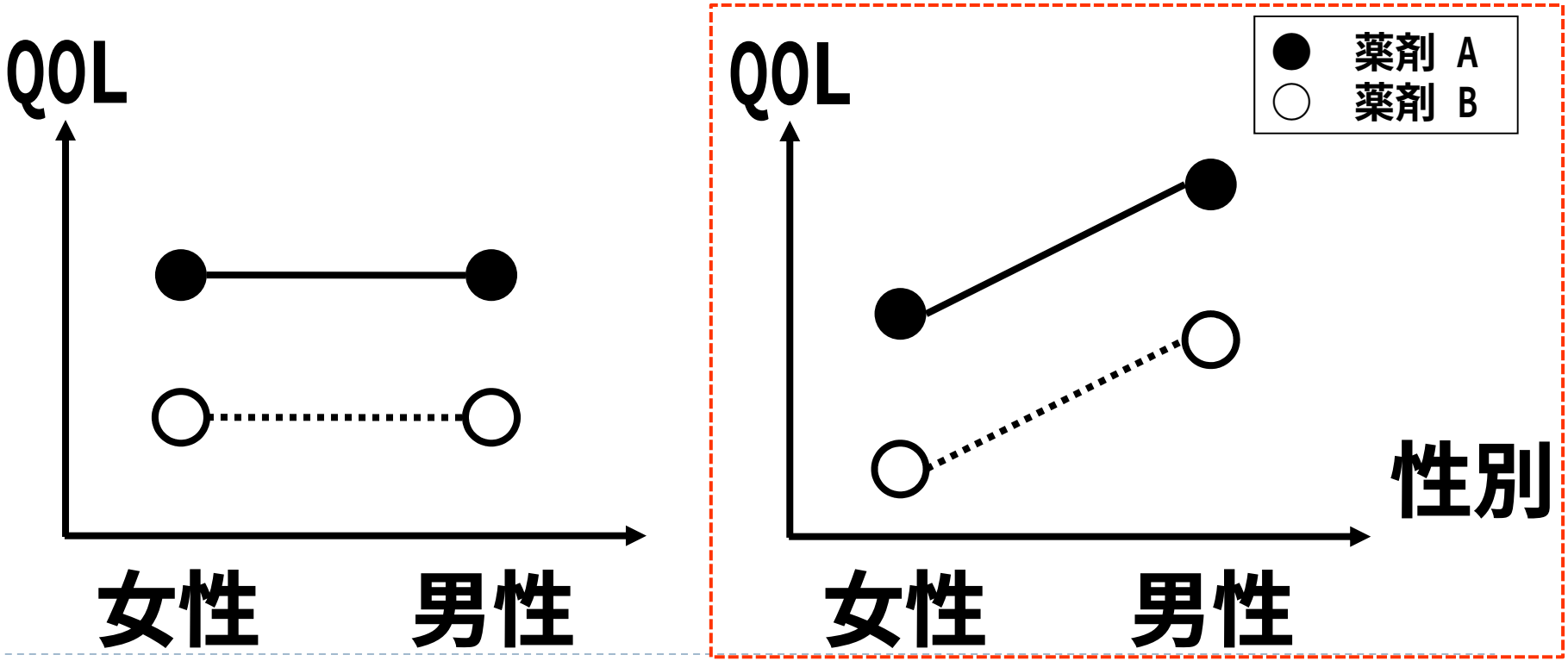




交互作用がない状態（●，○：QOL の平均値）

- 右下の図は以下の特徴がある
  - 「薬剤×性別」の交互作用がない
  - 性別が QOL に影響を及ぼしている

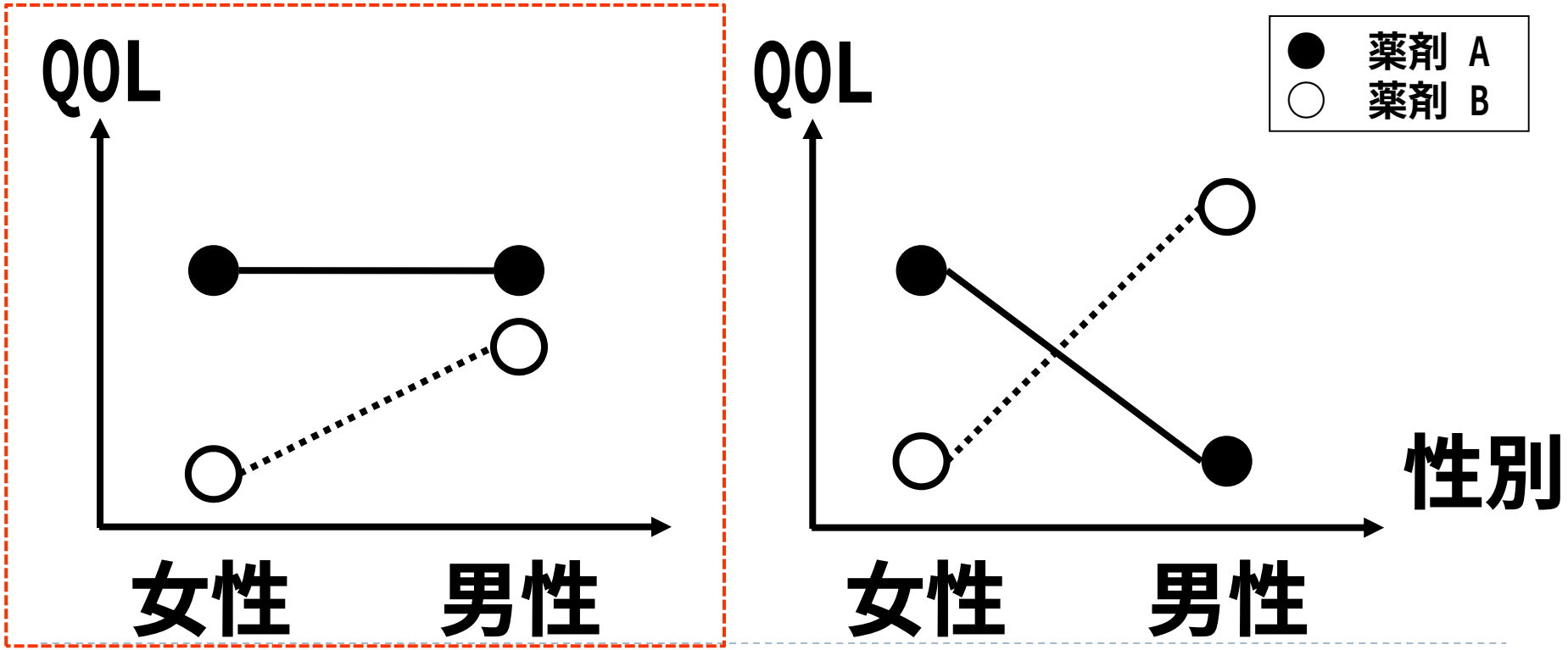
⇒ 女性も男性も，薬剤間の平均値の差は同じ



交互作用がある状態（●，○：QOL の平均値）

- 左下の図は以下の特徴がある ⇒ 量的な交互作用と呼ぶ
  - 「薬剤×性別」の交互作用がある
  - 女性も男性も，薬剤 *A* の平均値の方が高い

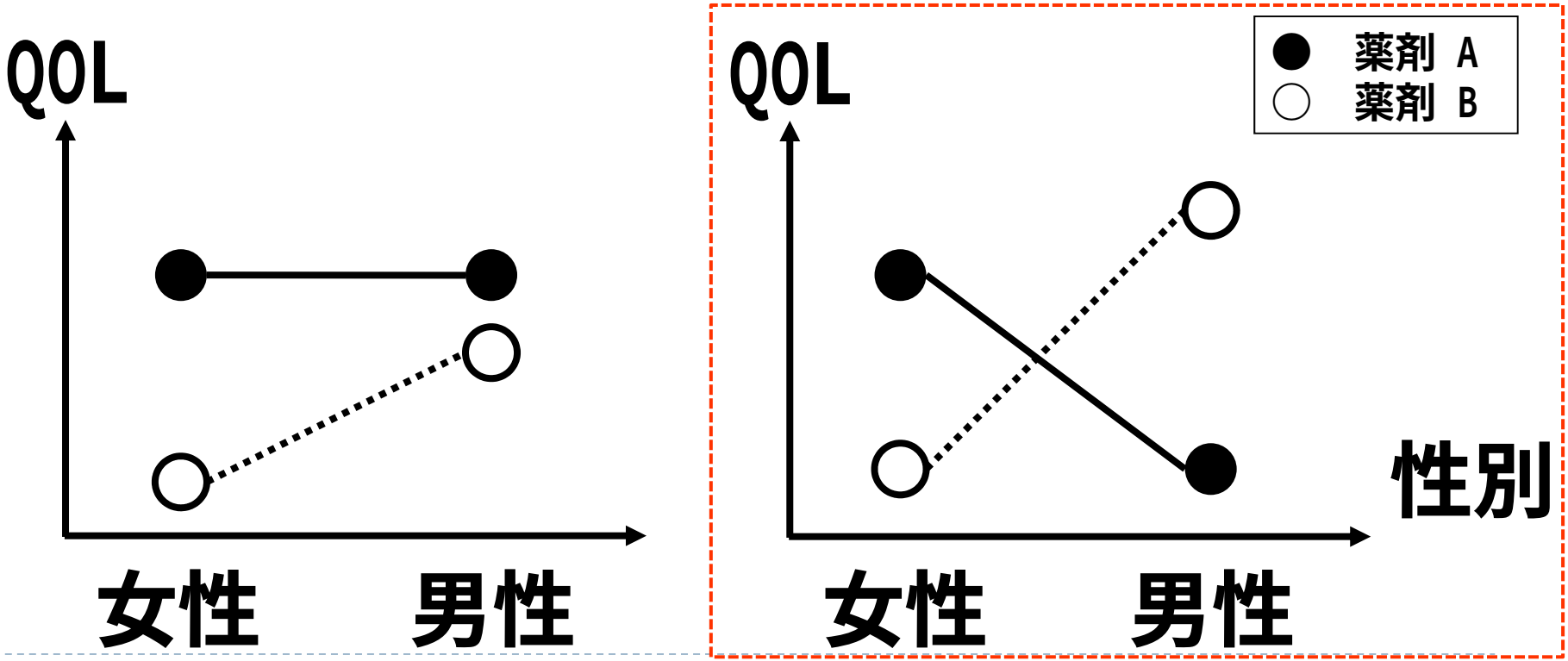
  ⇒ 性別によって薬剤間の平均値の差が異なる



交互作用がある状態（●，○：QOL の平均値）

- 右下の図は以下の特徴がある ⇒ 質的な交互作用と呼ぶ
  - 「薬剤×性別」の交互作用がある
  - 女性：薬剤 A の方が高い，男性：薬剤 B の方が高い

  ⇒ 性別によって薬剤間の平均値の差が異なる



交互作用がある状態

- 前頁の図はいずれも「薬剤×性別」の交互作用がある状態
  ⇒性別によって薬剤間の平均値の差が異なる
  ⇒ 「薬剤」と「性別」が互いに影響を及ぼし合っているため

左図：QOL の違いはあるが，女性の場合も男性の場合も薬剤 *A* の平均値の方が高い（大小関係の逆転は起こっていない）
⇒ この状態を「量的な交互作用あり」の状態と呼ぶ

右図：QOL の違いがあり，かつ性別によって大小関係の逆転が起こっている
⇒ この状態を「質的な交互作用あり」の状態と呼ぶ

- 「薬剤」と「性別」の間に交互作用がない場合は，「薬剤」だけに注目して解釈，「性別」だけに注目して解釈ということが出来る
- 2 つの要因の間に交互作用がある場合は「薬剤」と「性別」の両方を考慮して結果の解釈をする必要がある

ある因子が効果修飾因子かどうかの判定方法

興味のある因子が薬剤，「性別」が効果修飾因子かどうかを判定する場合

⇒「薬剤」と「性別」の交互作用があるかどうかを判定する場合

1. 薬剤別・性別（男女別）で要約統計量を求める（層別の結果）

   ⇒ 以下の条件を満たす場合，「性別」は効果修飾因子

   - 「男性における薬剤間の平均値の差」と「女性における薬剤の平均値の差」が異なる場合

ある因子が効果修飾因子かどうかの判定方法

興味のある因子が薬剤，「性別」が効果修飾因子かどうかを判定する場合

⇒「薬剤」と「性別」の交互作用があるかどうかを判定する場合

2. 以下のモデルで回帰分析し，交互作用項の効果（傾き β_3 ）が 0 でない場合，「性別」は効果修飾因子

   - 「薬剤＋性別＋薬剤×性別」のモデル：

     $QOL = \beta_0 + \beta_1 \times$薬剤 $+ \beta_2 \times$性別 $+ \beta_3 \times$薬剤×性別

交互作用がない例：データセット AB_DUMMY

- GROUP：薬剤の種類（A，B）
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- GENDER：性別（1：男性，2：女性）

```
> set.seed(777)                                          # 乱数のシード
> GROUP  <- c( rep("A",50), rep("B",50) )                # 薬剤
> GENDER <- 1+rbinom(100, 1, 0.5)                        # 性別（1:男，2:女）
> QOL    <- ifelse(GROUP=="A", 2.0+2.0*rnorm(50, sd=1),
+                              1.0+0.5*rnorm(50, sd=1))
> AB_DUMMY <- data.frame(QOL=round(QOL), GROUP =GROUP, GENDER=factor(GENDER))
> head(AB_DUMMY, n=3)
  QOL GROUP GENDER
1   5     A      2
2   0     A      1
3   7     A      1
```

交互作用がない例：データセット AB_DUMMY

```
> table(AB_DUMMY$GROUP, AB_DUMMY$GENDER)                    # 頻度集計

    1  2
  A 23 27
  B 31 19

> tapply(AB_DUMMY$QOL, AB_DUMMY$GROUP, mean)                # 全体の結果
   A    B
1.90 1.04

> tapply(AB_DUMMY$QOL, AB_DUMMY[,c("GENDER","GROUP")], mean)   # 層別の結果
      GROUP
GENDER        A        B
     1 2.000000 1.000000
     2 1.814815 1.105263
```

交互作用がない例：データセット AB_DUMMY

	QOL の平均値	例数
A	1.90	50
B	1.04	50

男性

	平均値	例数
A	2.00	23
B	1.00	31

女性

	平均値	例数
A	1.81	27
B	1.10	19

- 「男性の QOL の平均値の差」
  ⇒ 交互作用はなさそう ⇒ 一応，
- 「女性の QOL の平均値の差」
  回帰分析でも確かめる

交互作用がない例：データセット AB_DUMMY

```
> result <- lm(QOL ~ GROUP*GENDER, data=AB_DUMMY)        # 交互作用モデル
> summary(result)                                        # 結果の要約を表示

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.0000     0.3024   6.615 2.11e-09 ***
GROUPB          -1.0000     0.3991  -2.506   0.0139 *
GENDER2         -0.1852     0.4115  -0.450   0.6537
GROUPB:GENDER2   0.2904     0.5897   0.492   0.6235
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- 薬剤×性別の傾き = 0.2904（0 かどうかの検定の p 値 = 0.6235）
  ⇒傾きは 0 なので交互作用はなさそう

交互作用がある例：データセット AB_DUMMY

- GROUP：薬剤の種類（A，B）
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- GENDER：性別（1：男性，2：女性）

```
> set.seed(777)
> GROUP  <- c( rep("A",50), rep("B",50) )                    # 薬剤
> GENDER <- 1+rbinom(100, 1, 0.5)                            # 性別（1:男，2:女）
> QOL    <- ifelse(GROUP=="A" & GENDER==1, 3.0+rnorm(50, sd=2),
+           ifelse(GROUP=="A" & GENDER==2, 2.5+rnorm(50, sd=2),
+           ifelse(GROUP=="B" & GENDER==1, 1.0+rnorm(50, sd=2),
+                                          1.5+rnorm(50, sd=2))))
> AB_DUMMY <- data.frame(QOL   =round(QOL),
+                        GROUP =GROUP,
+                        GENDER=factor(GENDER))
> head(AB_DUMMY, n=3)
  QOL GROUP GENDER
1   4     A      2
2   1     A      1
3   8     A      1
```

交互作用がある例：データセット AB_DUMMY

```
> table(AB_DUMMY$GROUP, AB_DUMMY$GENDER)                    # 頻度集計

    1  2
  A 23 27
  B 31 19

> tapply(AB_DUMMY$QOL, AB_DUMMY$GROUP, mean)                # 全体の結果
   A    B
2.94 1.02

> tapply(AB_DUMMY$QOL, AB_DUMMY[,c("GENDER","GROUP")], mean) # 層別の結果
      GROUP
GENDER        A         B
     1 3.000000 0.3548387
     2 2.888889 2.1052632
```

交互作用がある例：データセット AB_DUMMY

	QOL の平均値	例数
A	2.94	50
B	1.02	50

男性

	平均値	例数
A	3.00	23
B	0.35	31

女性

	平均値	例数
A	2.88	27
B	2.10	19

- 「男性の QOL の平均値の差」≠「女性の QOL の平均値の差」
  ⇒ 交互作用はありそう ＆ 平均値の差の大小関係は逆転していない
  ⇒ 一応，回帰分析でも確かめる（きっと交互作用項は有意のはず）

交互作用がある例：データセット AB_DUMMY

```
> result <- lm(QOL ~ GROUP*GENDER, data=AB_DUMMY)        # 交互作用モデル
> summary(result)                                        # 結果の要約を表示

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.0000     0.4612   6.505 3.50e-09 ***
GROUPB          -2.6452     0.6086  -4.346 3.45e-05 ***
GENDER2         -0.1111     0.6275  -0.177   0.8598
GROUPB:GENDER2   1.8615     0.8995   2.070   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- 薬剤×性別の傾き = 1.8615（ 0 かどうかの検定の p 値 = 0.0412）
  ⇒ 傾きは 0 ではなさそうなので交互作用はありそう
  ⇒ 層別の結果から「量的な交互作用」があると結論

データセット「AB」の場合

- GROUP：薬剤の種類（A，B，C）⇒ A と B
- QOL：QOL の点数（数値）⇒ 点数が大きい方が良い
- PREDRUG：前治療薬の有無（YES：他の治療薬を投与したことあり，
  NO：投与したことなし）

```
> setwd("c:/temp")                          # dep.csv がある場所に移動
> DEP <- read.csv("dep.csv")                # dep.csv を読み込む
> AB <- subset(DEP, GROUP != "C")           # 薬剤 A と B のデータを抽出
> AB$GROUP <- factor(AB$GROUP)              # 薬剤の水準を 2 カテゴリに
> AB$GROUP <- relevel(AB$GROUP, ref="B")    # カテゴリのベースを「B」に変更
> head(AB, n=5)
  GROUP QOL EVENT DAY PREDRUG DURATION
1     A  15     1  50      NO        1
2     A  13     1 200      NO        3
3     A  11     1 250      NO        2
4     A  11     1 300      NO        4
5     A  10     1 350      NO        2
```

データセット「AB」の場合

```
> table(AB$GROUP, AB$PREDRUG)                          # 頻度集計

   NO YES
B   5  15
A  15   5

> tapply(AB$QOL, AB$GROUP, mean)                       # 全体の結果
  B   A
4.0 6.5

> tapply(AB$QOL, AB[,c("PREDRUG","GROUP")], mean)      # 層別の結果
       GROUP
PREDRUG   B   A
    NO  8.8 7.4
    YES 2.4 3.8
```

データセット「AB」の場合

	QOL の平均値	例数
A	6.5	20
B	4.0	20

前治療なし

	平均値	例数
A	7.4	15
B	8.8	5

前治療あり

	平均値	例数
A	3.8	5
B	2.4	15

- 「前治療なしの QOL の平均値の差」≠「前治療ありの QOL の平均値の差」
  ⇒ 交互作用はありそう ＆ 平均値の差の大小関係は逆転している
  ⇒ 一応，回帰分析でも確かめる

※既に交絡があることが分かっている状況下，「交互作用」の有無の検討をすること自体あまり意味がないという話もありますが･･･，例示のためと割り切ってください

データセット「AB」の場合

```
> result <- lm(QOL ~ GROUP*PREDRUG, data=AB)                    # 交互作用モデル
> summary(result)                                               # 結果の要約を表示

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.800      1.420   6.198 3.78e-07 ***
GROUPA              -1.400      1.639  -0.854 0.398749
PREDRUGYES          -6.400      1.639  -3.904 0.000399 ***
GROUPA:PREDRUGYES    2.800      2.318   1.208 0.235019
```

- 薬剤×性別の傾き = 2.800（ 0 かどうかの検定の p 値 = 0.235）
  ⇒ 前頁の層別の結果から「質的な交互作用」がありそうで，
  交互作用項の傾きも 2.800 と大きい
  ⇒ が！検定結果は有意ではない・・・（p = 0.235）
  ⇒ 原因は「交互作用の検定は一般的に検出力が低い」ため ⇒ 検定結果だけではなく層別解析の結果と合わせて見るのが得策

※既に交絡があることが分かっている状況下，「交互作用」の有無の検討をすること自体あまり意味がないという話もありますが･･･，例示のためと割り切ってください

本日のメニュー

1. 平均値の比較と 2 標本 t 検定
2. 回帰分析と 2 標本 t 検定
3. 交絡と交互作用
   - 交絡と交絡因子
   - 交互作用と効果修飾因子

参考文献

- 統計学（白旗 慎吾著，ミネルヴァ書房）
- 統計でウソをつく法（ダレル・ハフ著，高木 秀玄翻訳，ブルーバックス）
- 統計的多重比較法の基礎（永田 靖 他著，サイエンティスト社）
- ロスマンの疫学（Kenneth J. Rothman 著，矢野 栄二 他翻訳，篠原出版新社）
- Applied Logistic Regression（David W. Hosmer 他著，Wiley）
- The R Tips 第 2 版（オーム社）
- R 流！イメージで理解する統計処理入門（カットシステム）

Rで統計解析入門

終