確率と統計

中山クラス
第13週

今日の内容

- 第４回レポートの解説
- 第１０章　外れ値が相関係数に及ぼす影響
- 第５回レポート作成

第４回レポートの解説

練習問題（１）

ある国の20才男性の平均身長は平均170cmの正規分布に従うことが知られている．20人の身長データはこの国の20才男性の母集団からの無作為標本と考えて良いか？

（A） 教科書の119～122頁に記載されているt分布を用いる方法により検定せよ．t統計量に対する棄却域を求める方法と，p値を用いる方法を試みよ．但し，有意水準は5%とする．

（B） t.test関数を用いて検定を行い（A)の結果と比較せよ．

➢ 帰無仮説と対立仮説を日本語で示せ.

- 帰無仮説：20人の身長データは平均170cmの正規母集団からの無作為抽出である.（20人の身長データを無作為抽出した母集団は平均170cmの正規分布）
- 対立仮説：母集団は平均170cmの正規分布ではない.

➢ 検定統計量を文字と数式で示せ.

検定統計量は標本平均である. 母平均$\mu = 170$と分散(標準偏差)により正規化を行う. 母分散が不明であるので, 不偏分散$\hat{\sigma}^2$（標準偏差$\hat{\sigma}$）を用いる.

$$t = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

- 片側検定か両側検定かを説明せよ.

平均身長が170cmより低くても，高くても棄却されるので，両側検定となる.

- 有意水準を示せ.

$5\%(\alpha = 0.05)$

- 検定統計量の実現値と棄却域を示せ.

```
> 身長_表<-read.csv("ch5_r1.csv")
> 身長_表
   男性 身長
1     1  165
2     2  150
    ＜以下略＞
```

```
> 身長<-身長_表[,2]
> 身長
 [1] 165 150 170 168 159 170 167 178 155 159 161 162 166
＜略＞

 > t分子<-mean(身長)-170
 > t分子
 [1] -6.1
 > t分母<-sqrt(var(身長)/length(身長))
 > t分母
 [1] 1.584298
 > t統計量<-t分子/t分母
 > t統計量
 [1] -3.850286
```

検定統計量の実現値＝-3.850286

検定統計量は帰無仮説のもとで自由度$df = n - 1 = 20 - 1 = 19$のt分布に従う. 両側検定であるから, 下側確率が0.025となる検定統計量の値を求める.

```
> qt(0.025,19)
[1] -2.093024
> qt(0.975,19)
[1] 2.093024
```

これより, 棄却域は次式となる.

$t < -2.093024, \qquad 2.093024 < t$

- P値を示せ.

```
> pt(-3.850286,19)
[1] 0.0005392832
> pt(3.850286,19,lower.tail=FALSE)
[1] 0.0005392832
> 2*pt(3.850286,19,lower.tail=FALSE)
[1] 0.001078566
```

P値＝0.0005392832（片側）
　　0.001078566（両側）

- 帰無仮説を棄却するか採択するかを理由を付して述べよ.

$-3.850286 < -2.093024$　　実現値が棄却域にある

P値＝0.0005392832（片側）＜0.025

　　0.001078566（両側）　＜0.05

実現値以下，または|実現値|以上となる確率が有意水準（5%）より低い

以上の理由により，帰無仮説は棄却される.

- 解析結果を文章で述べよ.

20人の身長データは5%の有意水準で平均が170cm，分散が不偏分散で推定される正規分布に従う母集団からの無作為標本（抽出）であるとは言えない

(B) t.test関数を用いて検定を行い(A)の結果と比較せよ

```
> t.test(身長, mu=170)

        One Sample t-test

data:  身長
t = -3.8503, df = 19, p-value = 0.001079
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
 160.584 167.216
sample estimates:
mean of x
    163.9
```

上記の内容は,「検定統計量の実現値:t = -3.8503, 自由度:df = 19, 両側確率:p-value = 0.001079<0.05となり, 帰無仮説が棄却される」ことを表している. この結果は(A)の結果と同じになっている.

練習問題（２）

10人の大学生の１日の勉強時間と定期試験の得点の相関係数を無相関検定により調べる.

（A） 教科書の124～127頁に記載されているt分布を用いる方法により検定せよ. t統計量に対する棄却域を求める方法と，p値を用いる方法を試みよ. 但し，有意水準は5%とする.

（B） cor.test関数を用いて検定を行い，（A)の結果と比較せよ.

➢ 帰無仮説と対立仮説を日本語で示せ.

- 帰無仮説:勉強時間と定期試験の得点の間には相関がない.
- 対立仮説:勉強時間と定期試験の得点の間には相関がある.

➢ 検定統計量を文字と数式で示せ.

標本相関係数を検定統計量とする. t分布に従うように正規化する.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

rが標本相関係数, nがサンプルサイズである.

- 片側検定か両側検定かを説明せよ.

帰無仮説では相関係数$\rho = 0$であり，相関係数が正負で絶対値が大きくなると棄却されるので，両側検定となる.

- 有意水準を示せ.

$$\alpha = 0.05$$

- 検定統計量の実現値と棄却域を示せ.

```
> 勉強試験<-read.csv("ch3_r1.csv")
> 勉強試験
  勉強時間 定期試験の得点
1        1          20
2        3          40
3       10         100
4       12          80
＜以下略＞
```

```
> 勉強<-勉強試験[,1]
> 勉強
 [1]  1  3 10 12  6  3  8  4  1  5
> 試験<-勉強試験[,2]
> 試験
 [1]  20  40 100  80  50  50  70  50  10  60

> 標本相関<-cor(勉強,試験)
> 標本相関
[1] 0.9092974

> サンプルサイズ<-length(勉強試験[,1])
> サンプルサイズ
[1] 10
```

```
> t分子<-標本相関*sqrt(サンプルサイズ-2)
> t分子
[1] 2.571881
> t分母<-sqrt(1-標本相関^2)
> t分母
[1] 0.4161469
> t統計量<-t分子/t分母
> t統計量
[1] 6.180225
```

検定統計量の実現値＝6.180225

検定統計量tは帰無仮説$(\rho = 0)$のもとで自由度$df = n - 2 = 10 - 2 = 8$のt分布に従う. qt関数を用いて有意水準$\alpha = 0.05$の両側検定の棄却域を求める.

```
> qt(0.025,8)
[1] -2.306004
> qt(0.975,8)
[1] 2.306004
> qt(0.025,8,lower.tail=FALSE)
[1] 2.306004
```

以上より, 棄却域は次のようになる.

$$t < -2.306004, \qquad 2.306004 < t$$

➢ P値を示せ.

pt関数を用いて検定統計量の実現値が−6.180225以下，または6.180225以上となる確率を求める.

```
> pt(6.180225,8,lower.tail=FALSE)
[1] 0.0001325426
> 2*pt(6.180225,8,lower.tail=FALSE)
[1] 0.0002650851
```

以上より，検定統計量の実現値が−6.180225以下，または6.180225以上となる確率は

p=0.0001325426　（片側）

2*p=0.0002650851　（両側）

- 帰無仮説を棄却するか採択するかを理由を付して述べよ

検定統計量$t = 6.180225 > 2.306004$ 棄却域にある

p値（両側）＝$0.0002650851 < 0.05$　棄却域にある

以上の理由により，帰無仮説は棄却される.

- 解析結果を文章で述べよ.

勉強時間と定期試験の得点の間には5%の有意水準で相関がないとは言えない.

5%の水準で有意な相関がある.

（B） cor.test関数を用いて検定を行い，（A)と比較せよ.

```
> cor.test(勉強,試験)

        Pearson's product-moment correlation

data:  勉強 and 試験
t = 6.1802, df = 8, p-value = 0.0002651
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6542283 0.9786369
sample estimates:
      cor
0.9092974
```

検定統計量の実現値 = 6.1802, p値（両側）= 0.0002651<0.05となり，帰無仮説が棄却される. →5%の水準で有意な相関がある.
（A)と同じ結果になった.

練習問題(4)

（A） 教科書の130～134頁に記載されているカイ二乗分布を用いる方法により検定せよ. X^2統計量に対する棄却域を求める方法と，p値を用いる方法を試みよ. 但し，有意水準は5%とする.

（B） chisq.test関数を用いて検定を行い，（A)の結果と比較せよ.

- 帰無仮説と対立仮説を日本語で示せ.

帰無仮説：洋食派・和食派と甘党・辛党の間には連関がない（独立である）.

対立仮説：洋食派・和食派と甘党・辛党の間には連関がある（独立でない）.

- 検定統計量を文字と数式で示せ.

検定統計量として観測度数O_iと期待度数E_iの差を用いる. 期待度数は連関がないことを前提とした度数である. 差の二乗を期待度数で正規化している.

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

- 片側検定か両側検定かを説明せよ.

検定統計量X^2は正の値である. また, 連関がないとき観測度数は期待度数に近づくのでX^2は小さくなる. これらより,「連関がある／ない」はある基準値より大きい／小さいにより判定できるので, 片側検定となる.

- 有意水準を示せ.

$$\alpha = 0.05$$

- 検定統計量の実現値と棄却域を示せ.

```
> 和洋甘辛<-read.csv("ch3_r3.csv")
> 和洋甘辛
   洋食派か和食派か 甘党か辛党か
1            洋食        甘党
2            和食        辛党
3            和食        甘党
＜以下省略＞

> table(和洋甘辛)
                甘党か辛党か
洋食派か和食派か 甘党 辛党
          洋食    6    4
          和食    3    7
```

```
> 期待度数11<-9*10/20
> 期待度数21<-9*10/20
> 期待度数12<-11*10/20
> 期待度数22<-11*10/20
> 期待度数<-c(期待度数11,期待度数21,期待度数12,期待度数22)
> 期待度数
[1] 4.5 4.5 5.5 5.5
> 観測度数<-c(6,3,4,7)
> 観測度数
[1] 6 3 4 7

> カイ二乗要素<-(観測度数-期待度数)^2/期待度数
> カイ二乗要素
[1] 0.5000000 0.5000000 0.4090909 0.4090909
> カイ二乗<-sum(カイ二乗要素)
> カイ二乗
[1] 1.818182
```

検定統計量の実現値：$X^2 = 1.818182$

検定統計量X^2は自由度$df = (2-1)(2-1) = 1$のカイ二乗分布に従うので，有意水準：$\alpha = 0.05$に対する棄却域はqchisq関数を用いて次のようになる.

```
> qchisq(0.95,1)
[1] 3.841459
> qchisq(0.05,1,lower.tail=FALSE)
[1] 3.841459
```

棄却域：$X^2 > 3.841459$

➢ P値を示せ.

pchisq関数を用いてp値を求める.

```
> pchisq(1.818182,1,lower.tail=FALSE)
[1] 0.1775298
> 1-pchisq(1.818182,1)
[1] 0.1775298
```

p値=0.1775298

- 帰無仮説を棄却するか採択するかを理由を付して述べよ.

棄却域: $X^2 = 1.818182 < 3.841459$ 棄却されない

p値: 0.1775298>0.05 棄却されない

- 解析結果を文章で述べよ.

洋食派・和食派と甘党・辛党の間には水準5%で有意な連関がない(独立である)と言える.(有意な連関があるとは言えない).

（B） chisq.test関数を用いて検定を行い，（A)の結果と比較せよ.

```
> クロス集計表<-table(和洋甘辛)
> chisq.test(クロス集計表,correct=FALSE)

        Pearson's Chi-squared test

data:  クロス集計表
X-squared = 1.8182, df = 1, p-value = 0.1775

 警告メッセージ:
In chisq.test(クロス集計表, correct = FALSE) :
  カイ自乗近似は不正確かもしれません
```

検定統計量の実現値$X^2 = 1.8182$，p値=0.1775(>0.05)であり，帰無仮説は棄却されない．（A)と同じ結果となっている.

練習問題(5)

(5-1), (5-2)共にcor.test関数を用いて検定を行い，それらの結果と比較せよ.

```
> 国語社会<-read.csv("ch5_5-1.csv")
> 国語社会
  生徒 国語 社会
1    1   60   80
2    2   40   25
3    3   30   35
4    4   70   70
5    5   55   50

> 国語<-国語社会[,2]
> 社会<-国語社会[,3]
```

```
> cor.test(国語,社会)

        Pearson's product-moment correlation

data:  国語 and 社会
t = 2.6952, df = 3, p-value = 0.07408
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1590624  0.9892731
sample estimates:
      cor
0.841263
```

検定統計量の実現値t = 2.6952, 自由度$df = n - 2 = 3$, p値=0.07408>0.05であり，帰無仮説は棄却されない．5%水準で有意な相関があるとは言えない．

```
> 国語社会<-read.csv("ch5_5-2.csv")
> 国語社会
   生徒 国語 社会
1     1   60   80
2     2   40   25
3     3   30   35
4     4   70   70
5     5   55   50
6     6   60   80
7     7   40   25
8     8   30   35
9     9   70   70
10 10   55   50

> 国語<-国語社会[,2]
> 社会<-国語社会[,3]
```

```
> cor.test(国語,社会)

        Pearson's product-moment correlation

data:  国語 and 社会
t = 4.4013, df = 8, p-value = 0.002283
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4499858 0.9615658
sample estimates:
      cor
0.841263
```

検定統計量の実現値t = 4.4013, 自由度$df = n - 2 = 8$, p値=0.002283<0.05であり，帰無仮説は棄却される．従って，5%の水準で有意な相関があるといえる．

（参考）

統計的仮説検定では，母集団に対して2つの仮説を設定し，標本データから計算される検定統計量に基づき，確率に基づいて仮説を検証する．「差がない」という意味の仮説が帰無仮説で，「差がある」という意味の仮説が対立仮説である．母集団においては帰無仮説が正しいと仮定して，標本データから計算される有意確率が5%未満の場合に，帰無仮説を否定する．

帰無仮説を棄却する能力を検出力と呼ぶ．これはサンプルサイズ（n）に大きく依存する．サンプルサイズが大きくなれば検出力が高くなり，わずかな差でも「差がある」（＝帰無仮説を棄却）と判断しやすくなる．サンプルサイズが小さいと検出力が低くなり，有意な差があっても「差がない」（＝帰無仮説が棄却されない）と判断しやすくなる．

第10章
外れ値が相関係数に及ぼす影響

実際のデータから必要な情報を読み取る練習

2つの変数の関連を検討する.
散布図を利用して視覚的に表現→特徴を捉える.
2変数間の相関の強さを相関係数により要約する.

散布図から「外れ値」となっているデータを特定する.
「外れ値」を除いて散布図，相関係数を求める.
相関があるデータを絞り込む.

表10.1のデータを使用

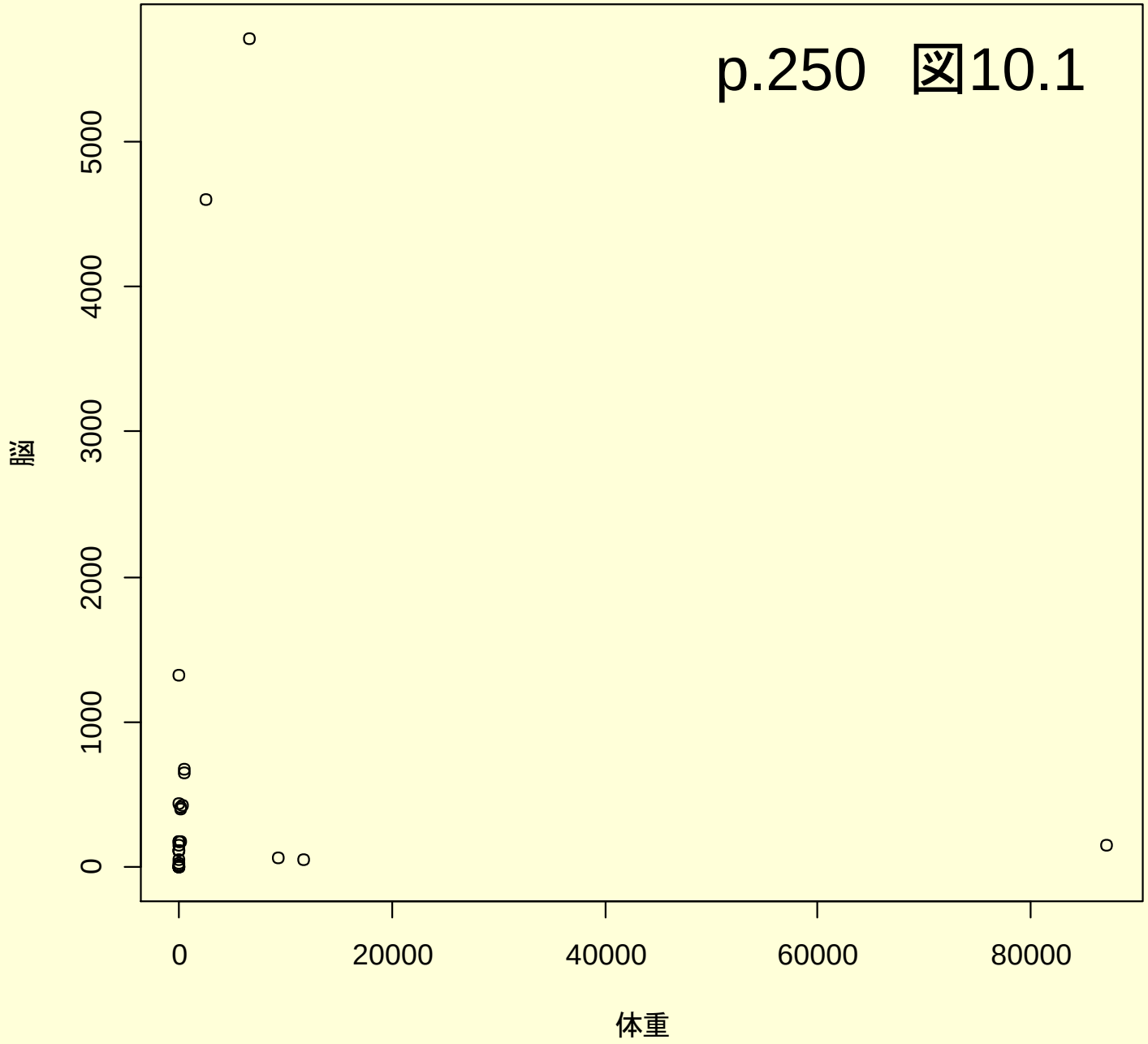
```
> 脳データ<-read.csv("ch10.csv")
> 脳データ
   動物      体重 脳の重さ
1     1     1.350      8.1
2     2   465.000    423.0
3     3    36.330    119.5
4     4    27.660    115.0
5     5     1.040      5.5
＜以下略＞
```

```
> 体重<-脳データ[,2]
```

- 脳<-脳データ[,3]

```
> plot(体重,脳)
```

```
> cor(体重,脳)
[1] -0.005341163
```


p.250 図10.1
脳
0
1000
2000
3000
4000
5000
0
20000
40000
60000
80000
体重

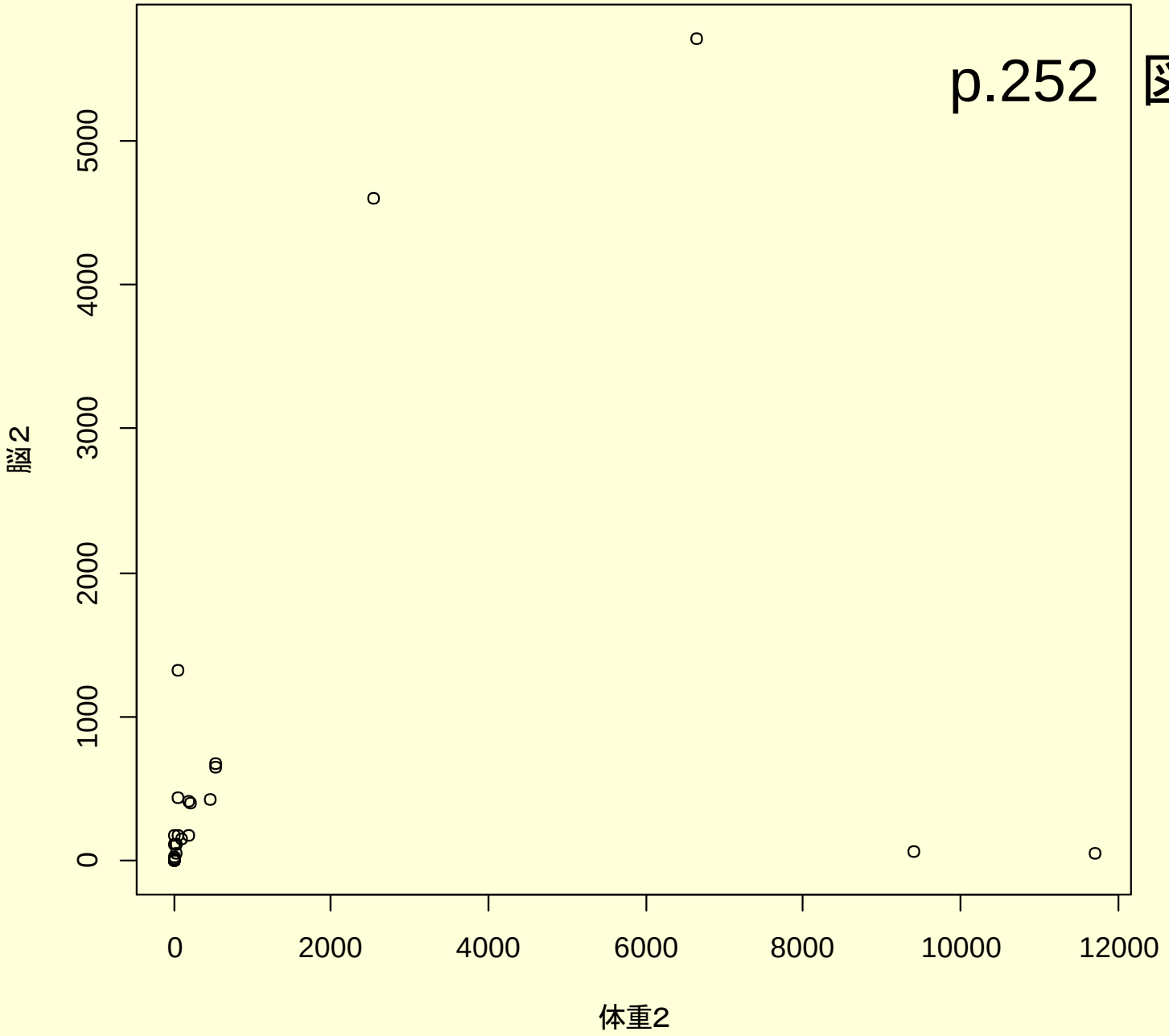
```
> 条件1<-体重<80000
> 脳データ2<-脳データ[条件1,]

> 体重2<-脳データ2[,2]
> 脳2<-脳データ2[,3]

> plot(体重2,脳2)

> cor(体重2,脳2)
[1] 0.3082425
```

p.252 図10.2

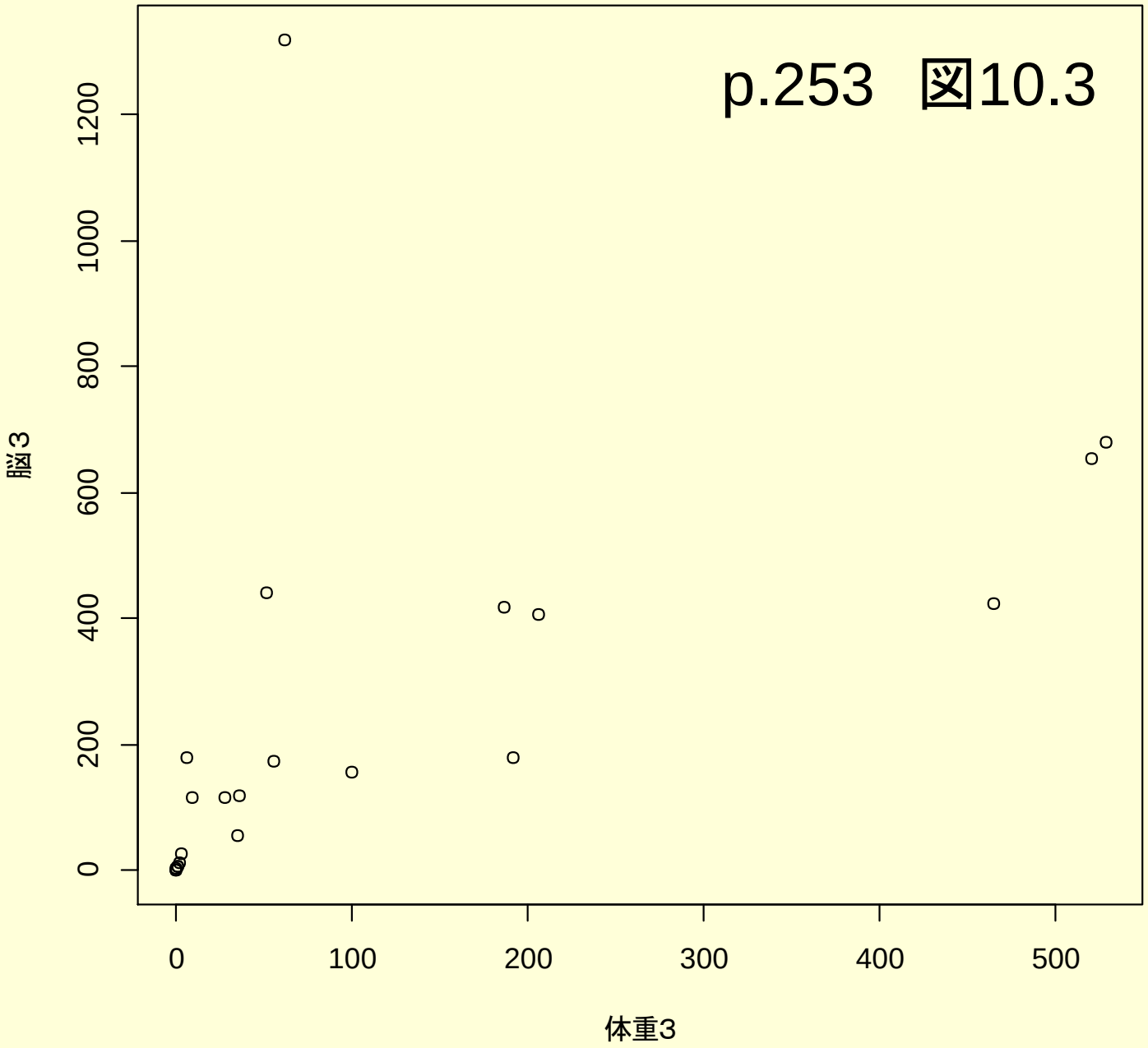

脳2
0
1000
2000
3000
4000
5000
0
2000
4000
6000
8000
10000
12000
体重2

```
> 条件2<-体重<2000
> 脳データ3<-脳データ[条件2,]

> 体重3<-脳データ3[,2]
> 脳3<-脳データ3[,3]

> plot(体重3,脳3)

> cor(体重3,脳3)
[1] 0.5423508
```


p.253 図10.3
脳3
0
200
400
600
800
1000
1200
0
100
200
300
400
500
体重3

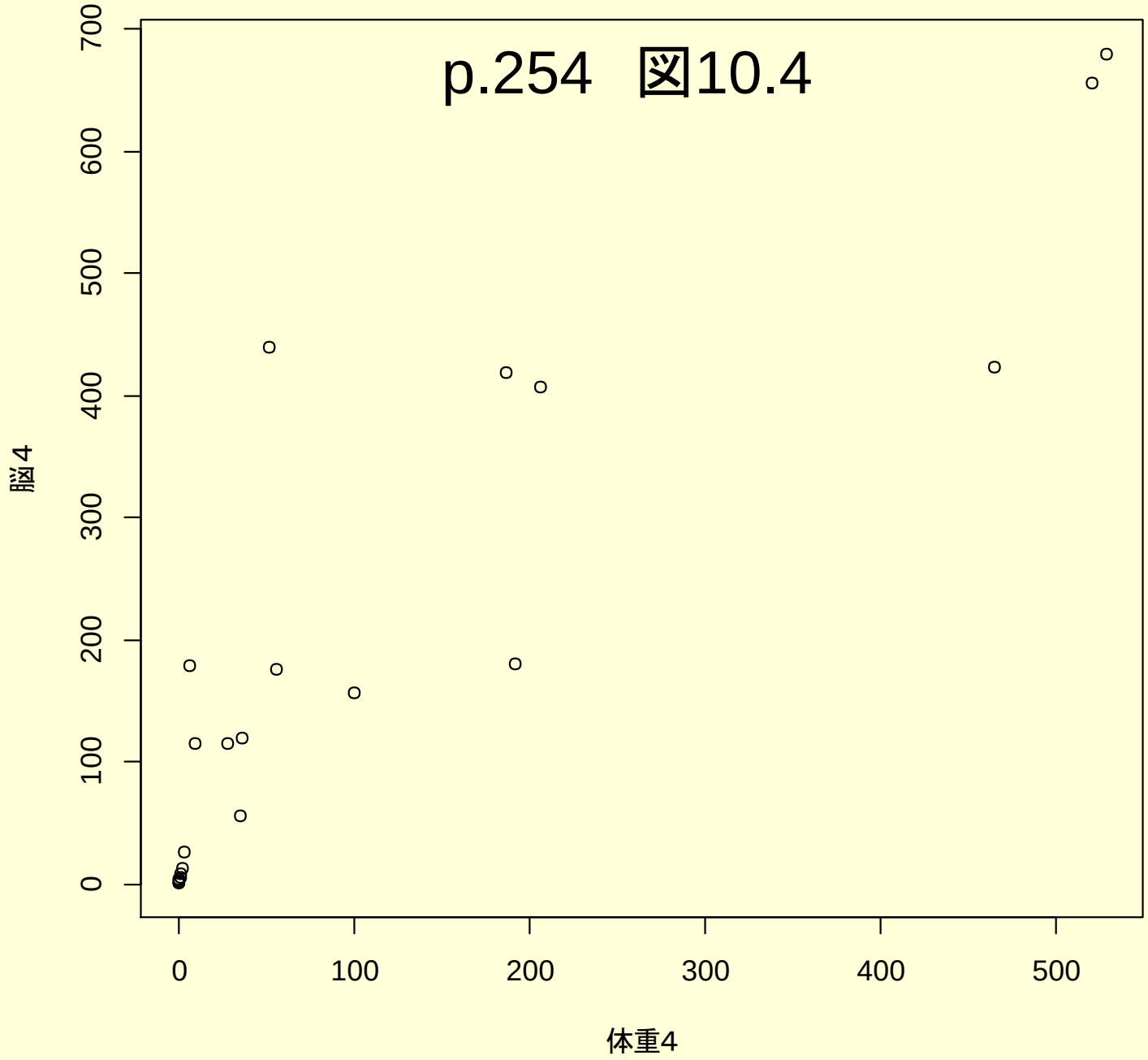
```
> 条件3<-(脳データ[,2]<2000)&(脳データ[,3]<1000)

> 脳データ4<-脳データ[条件3,]

> 体重4<-脳データ4[,2]
> 脳4<-脳データ4[,3]

> plot(体重4,脳4)

> cor(体重4,脳4)
[1] 0.8822336
```


p.254　図10.4
脳4
体重4
0
100
200
300
400
500
600
700

今後の予定

◆7月17日（火）　金曜日の授業
◆第5回レポート締め切り
　金曜クラス　7月24日（火）17:00
　火曜クラス　7月24日（火）17:00
◆模擬問題演習，回帰分析
　金曜クラス　7月20日（金）1限
　火曜クラス　7月24日（火）2限
◆達成度確認試験
　金曜クラス　7月27日（金）1限
　火曜クラス　7月31日（火）2限
◆再試験日程　8月1日（水）or 2日（木）16:30～
◆試験解答，自己点検授業，成績通知
　金曜クラス　8月3日（金）1限
　火曜クラス　8月7日（火）2限