今日のはなし (まずは単変量で！来週、多変量版＋α)

混乱したらここ(初心)へ戻り、**太字**の概念を理解したか確認！

[午前]

- **確率**ってそもそも何なの？
- 3つの道具：**確率変数**、**確率分布**、**期待値**
- 確率から統計へ：**母集団**と**標本**, **統計量**と**標本分布**
- **正規分布**とその性質

[午後]

- 正規分布の兄弟 (**カイ二乗分布**, **t分布**)
- 確率を導入しないとできないこと：**区間推定**と**仮説検定**
- **正規線形モデル**と**回帰係数の検定**
- 回帰係数・母回帰の区間推定：**予測区間**と**信頼区間**

区間推定と仮説検定

推測統計でやりたいことのキホン

1. 正規分布、カイ二乗分布、t分布

2. 区間推定

3. 仮説検定

正規分布まとめ

❶

$$X \sim N(\mu, \sigma^2) \ \Rightarrow \ \frac{X-\mu}{\sqrt{\sigma^2}} \sim N(0,1)$$

❷

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \ \Rightarrow \ \bar{X} = \frac{X_1 + \cdots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

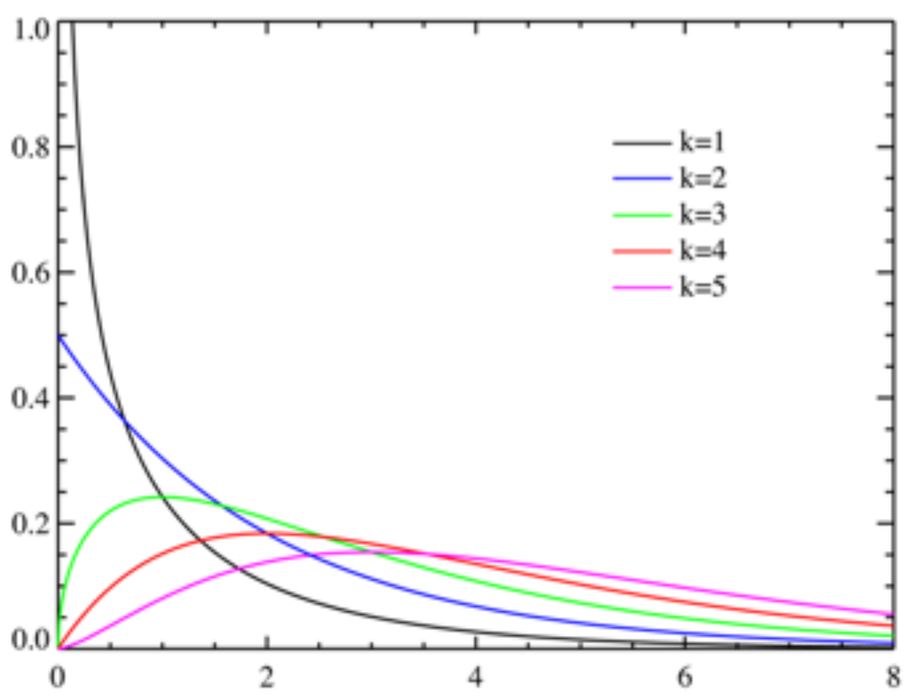
❸

$$Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

カイ二乗分布 （χ^2分布）

$X_1, X_2, \cdots, X_k \overset{\text{iid}}{\sim} N(0,1)$ に対し、

統計量 $\chi^2 := \sum_{i=1}^{k} X_i^2$ は自由度 k の χ^2分布に従う。



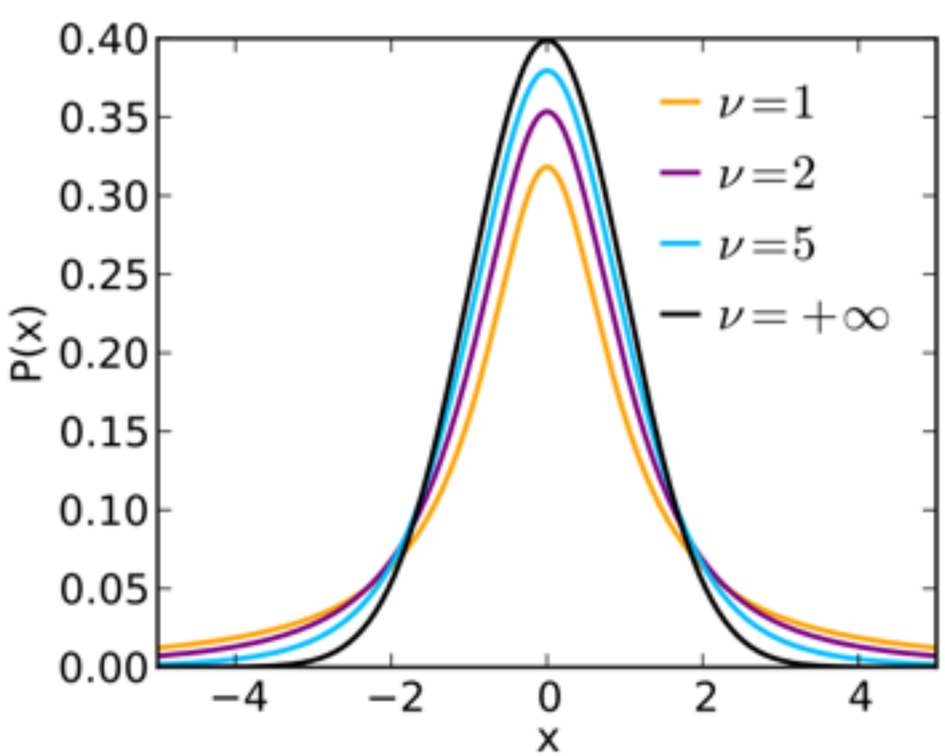
自由度 k の χ^2 分布

$$p(x) \propto x^{k/2-1} \exp\left(-\frac{x}{2}\right)$$

t分布

$X \sim N(0,1), Y \sim \chi^2(\nu)$ で X, Y が独立のとき、

統計量 $T = \dfrac{X}{\sqrt{Y/\nu}}$ は自由度 ν のt分布に従う。



自由度 ν のt分布

$$p(x) \propto \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

確率を導入しないとできないこと：**仮説検定**

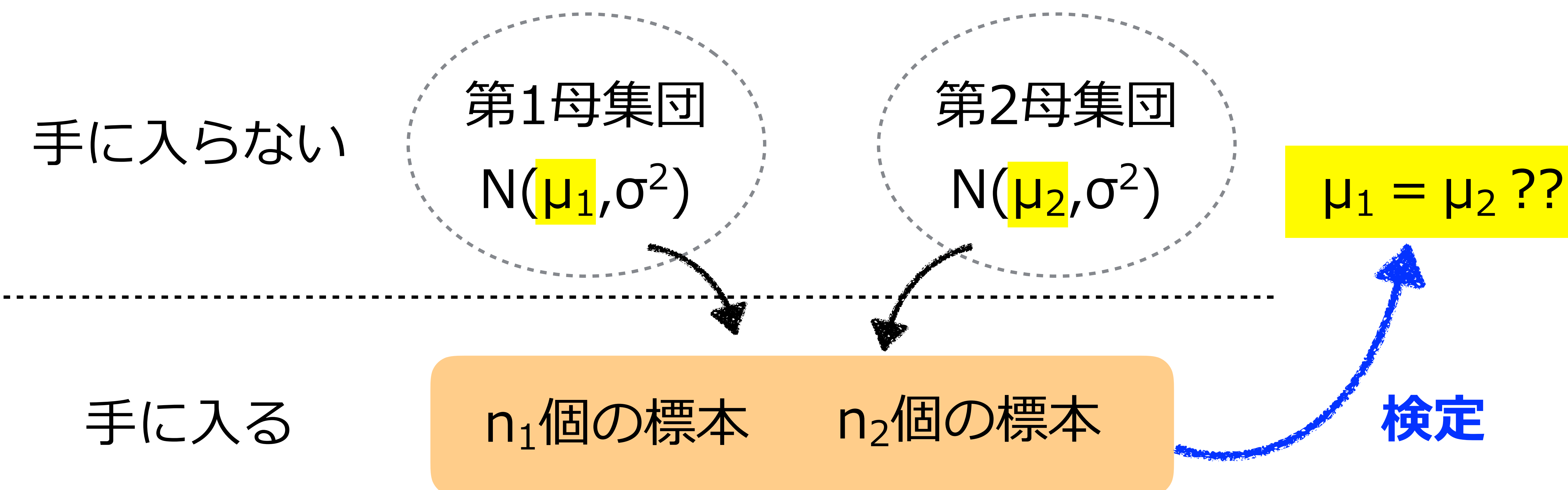
（例１）システムAとシステムBの使いやすさを、5点満点で12人の被験者に評価してもらったところ、システムBの方が平均値が高いことがわかった。システムBの方がシステムAより使いやすいと言えるだろうか。（よくあるデータのばらつきの範囲だろうか。それとも、「違いがある」と言えるだろうか。）

（例２）40人の被験者から、3月分と4月分の携帯電話の通話料のデータを集め、4月分の平均値は3月分の平均値より大きいことがわかった。この結果から、3月と4月では携帯電話の通話料に差があると言えるだろうか。

http://d.hatena.ne.jp/Zellij/20140608/p1

仮説検定の例：二つの母平均の差の検定

正規分布$N(\mu_1,\sigma^2)$から得たn_1個の標本と
正規分布$N(\mu_2,\sigma^2)$から得たn_2個の標本に基づいて
２つの母平均が異なるかどうかをテスト(検定)する。



仮説検定の戦略

1. 母集団のモデルを仮定　　システムAの評価値 $\sim N(\mu_A,\sigma^2)$
　　システムBの評価値 $\sim N(\mu_B,\sigma^2)$
2. ある**標本統計量T(検定統計量)**を設計
3. この統計量Tの**もしμ_Aとμ_Bに差がない場合の標本分布P**を導出
4. いま手元にあるデータでTの値tを計算して、**手元の実現値より極端な値が得られる確率P(T>t)**を計算。つまり、本当は差がないのに「よくあるデータのばらつきの範囲」で手元のデータが得られてしまう確率を求めてみる。

仮説検定の結論と「統計的有意性」

5. この確率が小さければ(0.01以下とか)、手元の実現値は「よくあるデータのばらつき」では起こり得ないと考える。
この場合、「μ_Aとμ_Bに差がない」と考えるのは合理的ではない！
→　**「μ_Aとμ_Bには偶然で起こりうる以上の差がある」と結論**

「統計的に有意に差がある」と言う。

検定の(トリッキーな)ロジックの構造

データから言いたいこと「差がある」を否定した状況での、ある統計量Tの標本分布を求める

→

手元のデータで統計量Tを計算し、標本分布でその値以上が得られる確率を計算する

→

確率が小さい
→「差がない」と考えるには無理があると考える

確率が大きい
→ 特になにも言えない

★ 仮説検定

標本 $X_1, \ldots, X_n$ から母平均に関する仮説が正しそうか検定したい。

たとえば、仮説「$\mu = 0$」かどうか検定するには？

もしこの仮説が真ならば、

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \text{ より、} \tilde{Z} = \frac{\bar{X}}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

$\tilde{Z}$ を計算して、それが $N(0,1)$ から来たっぽさを検査

ステップ１：

言いたいこと：「$\mu = 0$ 」ではない。

否定したこと：「$\mu = 0$ 」である。(帰無仮説)

帰無仮説が成り立たないとき「言いたいこと」が成り立つようにする。(言いたいこと=対立仮説)

ステップ２：

帰無仮説が成り立つと仮定して、検定統計量をきめて
その標本分布を求める。

もし帰無仮説が真ならば、

$$Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \text{ より、} \tilde{Z} = \frac{\bar{X}}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

分布もわかっているしこれが検定統計量に使えそう?

ここで問題が発生

$$\tilde{Z} = \frac{\bar{X}}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$ σ^2 も未知なので計算できず。

(仕方ないので)計算できる標本分散で代用

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

検定統計量 $T := \frac{\bar{X}}{\sqrt{\frac{\widehat{\sigma}^2}{n}}}$ の標本分布は?

使える知見1：

(1) $$Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = (\bar{X}-\mu)\frac{\sqrt{n}}{\sigma} \sim N(0,1)$$

(2) $V = (n-1)\dfrac{\widehat{\sigma^2}}{\sigma^2}$ とおけば

$$T := \frac{\bar{X}-\mu}{\sqrt{\frac{\widehat{\sigma}^2}{n}}} = \frac{\sqrt{n}}{\widehat{\sigma}^2}(\bar{X}-\mu) = \frac{\sqrt{n}}{\sigma}\frac{\bar{X}-\mu}{\sqrt{\frac{\widehat{\sigma}^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{V}{n-1}}}$$

使える知見2：

統計量 $V=(n-1)\frac{\widehat{\sigma^2}}{\sigma^2}$ に対し $V\sim\chi^2(n-1)$

つまり $\widehat{\sigma^2}=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$ なので

$$\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sigma}\right)^2\sim\chi^2(n-1)$$

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2\sim\chi^2(n)$$

証明:

http://mathtrain.jp/chinijoproof

従って、

$$T = \frac{Z}{\sqrt{\frac{V}{n-1}}} = \frac{\sqrt{n}}{\sigma}\frac{\bar{X}-\mu}{\sqrt{\frac{\widehat{\sigma^2}}{\sigma^2}}} = \frac{\sqrt{n}}{\widehat{\sigma^2}}(\bar{X}-\mu) = \frac{\bar{X}-\mu}{\sqrt{\frac{\widehat{\sigma^2}}{n}}}$$

は、自由度 n-1 のt分布に従う！

❸ $\frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

この統計量 T は母分散 σ^2 を含まない

まとめ

母平均 μ に関する仮説を検定したければ、
統計量 T を計算し、自由度n-1のt分布ぽいかを
調べればよし(分布の裾5%に入るほど稀な値か否か)。

(というかt分布はこのために生まれた)

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \qquad \widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

のとき、

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} = \frac{\sqrt{n}}{\widehat{\sigma}} (\bar{X} - \mu)$$

❸

$$\boxed{Z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)}$$

は $T \sim t(\underbrace{n-1}_{\text{自由度}})$

注：不偏分散の
自由度**のみ**に依存

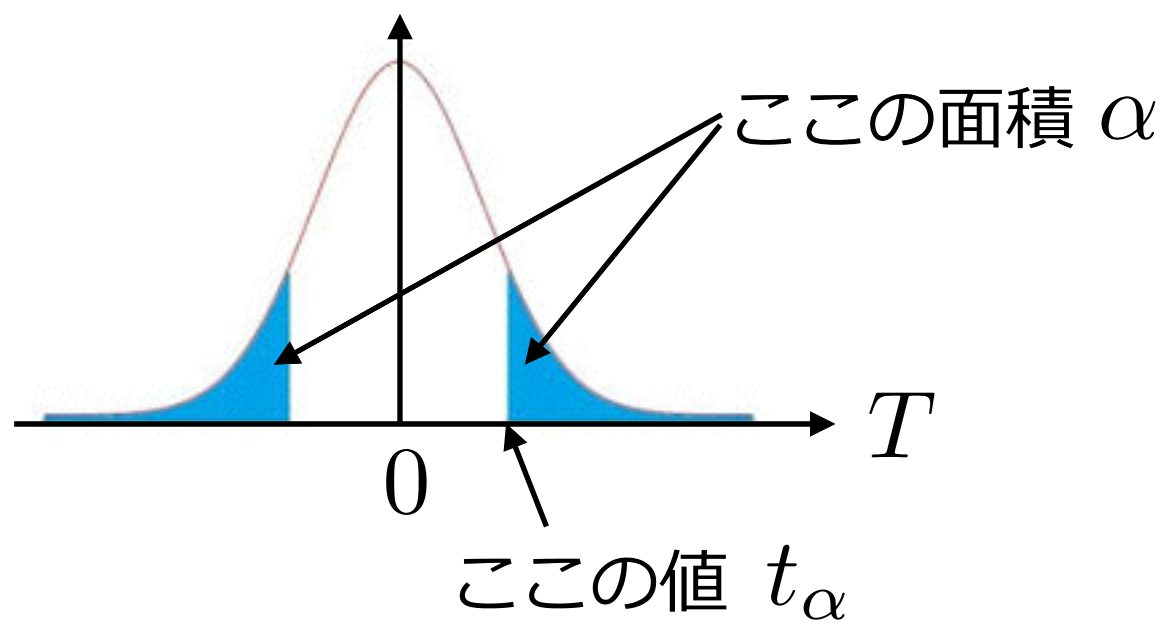
★ 検定のまとめ

まとめ

母平均 μ に関する仮説を検定したければ、
統計量 T を計算し、自由度n-1のt分布ぽいかを
調べればよし(分布の裾5%に入るほど稀な値か否か)。

5%

裾確率 $P(|T| \geqslant t_\alpha) = \alpha$



正規分布まとめ

❶

$$X \sim N(\mu, \sigma^2) \Rightarrow \frac{X-\mu}{\sqrt{\sigma^2}} \sim N(0,1)$$

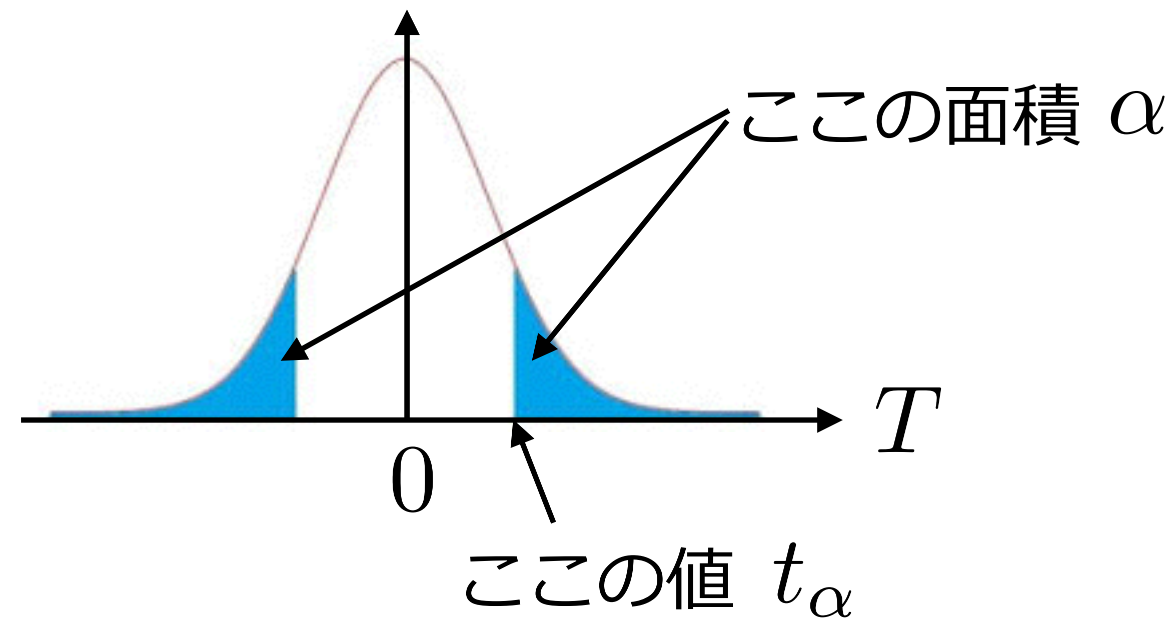
❷

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X} = \frac{X_1 + \cdots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

❸

$$Z = \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1) \quad \Rightarrow \quad \boxed{T = \frac{\bar{X}-\mu}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} \sim t(n-1)}$$

★ 検定から区間推定へ



5%

裾確率　$P(|T| \geqslant t_\alpha) = \alpha$

95%

確率 $1-\alpha$ で $|T| < t_\alpha$

$$\Leftrightarrow \frac{\left|\bar{X}-\mu\right|}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} < t_\alpha$$

$$\Leftrightarrow \bar{X} - t_\alpha \sqrt{\frac{\widehat{\sigma^2}}{n}} < \mu < \bar{X} + t_\alpha \sqrt{\frac{\widehat{\sigma^2}}{n}}$$

母平均の信頼区間

$$\bar{X} - t_\alpha \sqrt{\frac{\widehat{\sigma^2}}{n}} < \mu < \bar{X} + t_\alpha \sqrt{\frac{\widehat{\sigma^2}}{n}}$$

下限　　　　　　　　　　　　　上限

母平均
(未知)

未知量が1-αの確率でこの範囲に入ることが分かる。
なおかつ、上限・下限は簡単に計算できる量！

注： t_α の値の計算は自明ではないが数値計算できる
(昔は数値表を参照していた)

推定・検定のしかた (まとめ)

❶ 検定統計量の選定
標本統計量 Z の分布(標本分布)を解析的に求める。
あるいは分布型が分かるような統計量 Z を考える。

❷ **(両側)検定:** 検定統計量 Z を計算し、
$|Z|$の値がz_αより大きければ有意

❸ **区間推定:** $P(|Z| \geq z_\alpha)$より、確率$1-\alpha$で$|Z|<z_\alpha$となることより、関心のある母数の信頼区間を導出

参考文献

- 確率・統計入門: 小針 アキ宏 (著)
  岩波書店 (1973/05) ISBN-10: 4000051571

- 入門 数理統計学: P. G. ホーエル (著)
  培風館 (1978/01) ISBN-10: 4563008281

- 自然科学の統計学: 東京大学教養学部統計学教室 (編)
  東京大学出版会 (1992/08) ISBN-10: 4130420674

- 数理統計学―基礎から学ぶデータ解析: 鈴木 武・山田 作太郎 (著)
  内田老鶴圃 (1996/04) ISBN-10: 4753601196

正規線形モデルと回帰

Given: $x_1, x_2, \ldots, x_n$

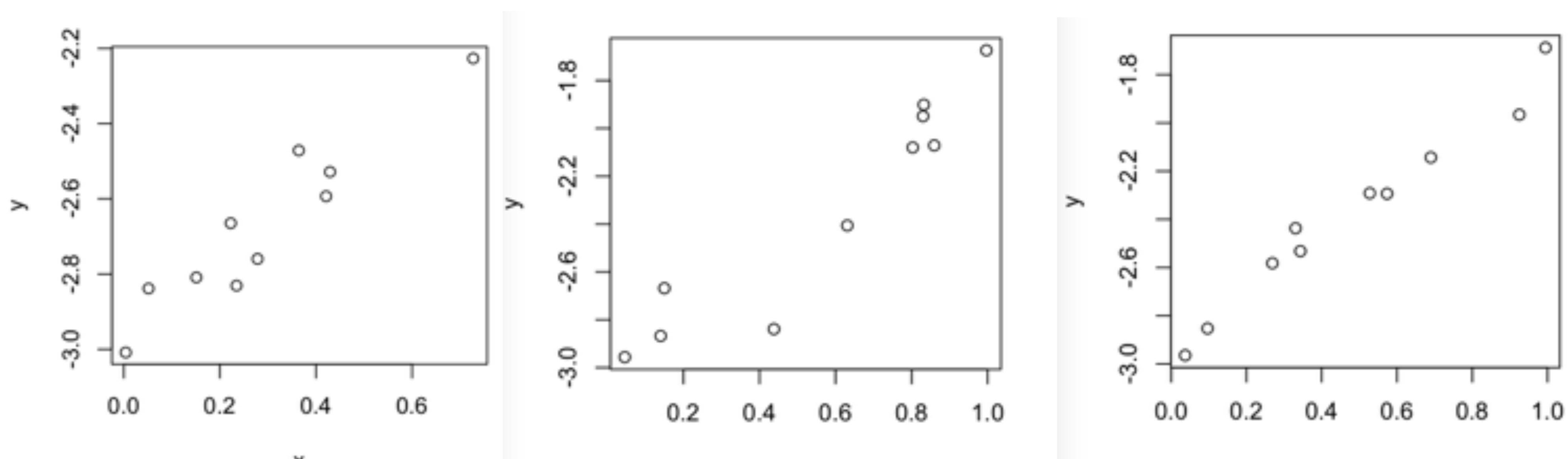
$$Y_i = a\,x_i + b + \boxed{Z_i}$$

正規乱数 $Z_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

Observe: $Y_1, Y_2, \cdots, Y_n$

Q: 観測された $x_1, x_2, \ldots, x_n$ および $Y_1, Y_2, \cdots, Y_n$ から、その背後にある未知パラメタ a, b, σ^2 をどのくらいの精度で当てられる？？

$$a = 1.2, b = -3.0, \sigma = 0.1$$



回帰分析の実例

ベネッセ教育総合研究所 東京大学共同研究
「学校教育に対する保護者の意識調査2008」

表4−6 学校外教育費（対数値）を被説明変数とする重回帰分析

	小学2年生 非標準化係数	小学5年生 非標準化係数	中学2年生 非標準化係数
説明変数			
定数	8.725	8.665	9.761
東京居住（東京=1）	0.169**	0.501***	0.166*
子どもの性別（男子=1）	−0.240***	−0.283***	−0.069+
長子ダミー（長子=1）	−0.041	0.155**	−0.001
子ども数	−0.126***	−0.144***	−0.128***
成績認知（1-5）	0.113***	0.050*	−0.047**
大学進学期待ダミー	0.313***	0.350***	0.304***
父年齢	0.000	0.000	−0.001
経済的ゆとり（1-4）	0.063+	0.140***	0.096**
母学歴（大学卒=1）	0.038	0.252***	0.107*
調整済みR^2	0.142	0.262	0.088
F値	17.684***	43.054***	13.317***
N	907	1,068	1,155

注）+p<.10 *p<.05 **p<.01 ***p<.001

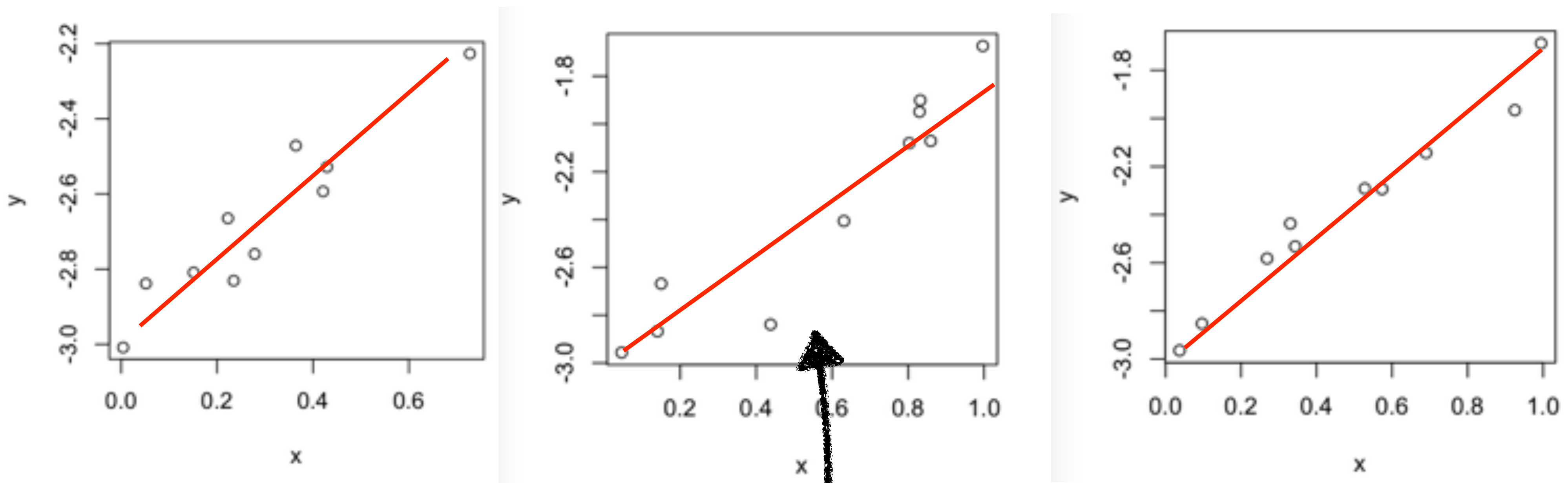
http://berd.benesse.jp/berd/center/open/report/hogosya_ishiki/2008/hon/hon4_8.html

Rでの回帰分析の出力

```
> model1 <- lm(Life.Exp ~ ., data=data.frame(state.x77))
> summary(model1)

Call:
lm(formula = Life.Exp ~ ., data = data.frame(state.x77))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.48895 -0.51232 -0.02747  0.57002  1.49447 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .  
Income      -2.180e-05  2.444e-04  -0.089   0.9293    
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269    
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *  
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .  
Area        -7.383e-08  1.668e-06  -0.044   0.9649    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared: 0.7362,     Adjusted R-squared: 0.6922 
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

$$a = 1.2, b = -3.0, \sigma = 0.1$$

観測されなかったがこうであったかもしれない標本を考慮する



標本分布を考えることができる！

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ とおくと

$$\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{S_{xx}} \qquad \hat{b} = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}} \right) y_i$$

推定問題としての「回帰」

まず**「単変量」**の場合 (説明変数が 1 こだけの回帰)

Given: $x_1, x_2, \ldots, x_n$

$$Y_i = a\,x_i + b + Z_i$$

正規乱数 $Z_i \overset{\mathrm{iid}}{\sim} N(0, \sigma^2)$

Observe: $Y_1, Y_2, \cdots, Y_n$

Q: 観測された $x_1, x_2, \ldots, x_n$ および $Y_1, Y_2, \cdots, Y_n$ から、その背後にある未知パラメタ a, b, σ^2 をどのくらいの精度で当てられる？？

平均

(2.1)式

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

平方和

(2.2)式

$$S_{xx} = \sum_{x=1}^{n}(x_i - \bar{x})^2$$

偏差積和

(2.8)式

$$S_{xy} = \sum_{x=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$
\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y} = \begin{bmatrix} \frac{\sum_{i=1}^n x_i y_i - \bar{x}\sum_{i=1}^n y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ -\frac{\bar{x}\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} + \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\sum_{i=1}^n y_i \end{bmatrix}
$$

$$
= \begin{bmatrix} \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ -\bar{x}\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{n}\sum_{i=1}^n y_i + \bar{x}\frac{\sum_{i=1}^n \bar{x} y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{bmatrix}
$$

$$
= \begin{bmatrix} \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \sum_{i=1}^n \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}\right) y_i \end{bmatrix}
$$

$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$ とおくと

比較: 教科書(4.10)(4.15)式

$$
\hat{a} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{S_{xx}} \qquad \hat{b} = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right) y_i
$$

確認

(1) $$\sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = 0$$

(2) $$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

(3) $$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i$$

(4) $$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

単回帰のまとめ

回帰係数の推定量
(最小二乗推定)

$$\hat{a} = \frac{S_{xy}}{S_{xx}} \qquad \hat{b} = \bar{y} - \hat{a}\bar{x}$$

回帰式

$$\begin{aligned} y &= \hat{a}x + \hat{b} \\ &= \hat{a}x + \hat{y} - \hat{a}\bar{x} = \hat{a}(x - \bar{x}) + \bar{y} \end{aligned}$$

残差平方和

$$\begin{aligned} \sum_{i=1}^{n} \left(y_i - (\hat{a}\, x_i + \hat{b}) \right)^2 &= S_{yy} - \hat{a} S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \\ &= S_{yy} \left(1 - \boxed{\frac{S_{xy}^2}{S_{xx} S_{yy}}} \right) \end{aligned}$$

(4.22)式　**寄与率**
(決定係数)
= 相関係数の二乗

$$R^2 = \frac{S_R}{S_{yy}} = \sum_{i=1}^{n} (\bar{y} - (a\, x_i + b))^2 \cdot \frac{1}{S_{yy}}$$

$$S_R = \frac{S_{xy}^2}{S_{xx}}$$

推定の準備1

$$\hat{a} = \frac{\boxed{S_{xy}}}{S_{xx}} = \frac{\boxed{\sum_{i=1}^{n}(x_i - \bar{x})y_i}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(a\,x_i + b + Z_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= a\underbrace{\boxed{\frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}_{=1} + b\underbrace{\boxed{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}_{=0} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})Z_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= a + \frac{\sum_{i=1}^{n}(x_i - \bar{x})Z_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

よって、

$$a - \hat{a} = -\frac{\sum_{i=1}^{n}(x_i - \bar{x})Z_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

推定の準備2

教科書(4.25)

$$\mathbb{E}\{\hat{a}\} = a + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mathbb{E}\{Z_i\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = a$$

$$\mathbb{E}\{(\hat{a} - a)^2\} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \mathbb{E}\left\{ \frac{\sum_{i=1}^{n}\sum_{i=j}^{n}(x_i - \bar{x})(x_j - \bar{x})Z_i Z_j}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right\}$$

$$= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \mathbb{E}\left\{Z_i^2\right\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\mathbb{E}\left\{Z_i^2\right\}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{S_{xx}}$$

回帰係数の推定量 (最小二乗推定) $\hat{a} = \frac{S_{xy}}{S_{xx}}$ は平均 a 分散 $\frac{\sigma^2}{S_{xx}}$

σ^2 の不偏推定

$$\sum_{i=1}^{n}\left(y_i-(\hat{a}\,x_i+\hat{b})\right)^2=\sum_{i=1}^{n}\left(y_i-\bar{y}-\hat{a}\,(x_i-\bar{x})\right)^2$$

$$=\sum_{i=1}^{n}\left(a\,(x_i-\bar{x})+(Z_i-\bar{Z})-\hat{a}\,(x_i-\bar{x})\right)^2$$

$$=\sum_{i=1}^{n}\left((a-\hat{a})\,(x_i-\bar{x})+(Z_i-\bar{Z})\right)^2$$

$$=\sum_{i=1}^{n}\left\{(a-\hat{a})^2(x_i-\bar{x})^2+(Z_i-\bar{Z})^2+2(a-\hat{a})(x_i-\bar{x})(Z_i-\bar{Z})\right\}$$

$$=\sum_{i=1}^{n}(a-\hat{a})^2(x_i-\bar{x})^2+\sum_{i=1}^{n}(Z_i-\bar{Z})^2+2(a-\hat{a})\boxed{\sum_{i=1}^{n}(x_i-\bar{x})(Z_i-\bar{Z})}$$

$$=\textstyle\sum_{i=1}^{n}(x_i-\bar{x})Z_i$$

準備1の結果

$$a-\hat{a}=-\frac{\sum_{i=1}^{n}(x_i-\bar{x})Z_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$ より

$$=-2\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-\bar{x})(x_j-\bar{x})Z_iZ_j}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$\mathbb{E}\left\{\sum_{i=1}^{n}\left(y_i-(\hat{a}x_i+\hat{b})\right)^2\right\}$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})^2\mathbb{E}\left\{(a-\hat{a})^2\right\}+\mathbb{E}\left\{\sum_{i=1}^{n}(Z_i-\bar{Z})^2\right\}-2\frac{\sum_{i=1}^{n}\sum_{i=j}^{n}(x_i-\bar{x})(x_j-\bar{x})\mathbb{E}\left\{Z_iZ_j\right\}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$=\sum_{i=1}^{n}(x_i-\bar{x})^2\frac{\sigma^2}{S_{xx}}+(n-1)\sigma^2-2\sigma^2=\sigma^2+(n-1)\sigma^2-2\sigma^2=(n-2)\sigma^2$$

よって、σ^2 の不偏推定量は以下で与えられる。

$$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-(\hat{a}x_i+\hat{b})\right)^2$$

教科書(4.18)

未知パラメタへの代入 $\hat{a}\to a,\hat{b}\to b$ により自由度が2減っている。

単回帰の推定量のまとめ (1)

データの背後に置く仮定	$y = ax + b + \varepsilon, \varepsilon \sim N(0, \sigma^2)$
回帰式	$y = ax + b$

① 係数の推定量

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}} = \frac{S_{xy}}{S_{xx}}$$

② 切片の推定量

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{S_{xx}}\right)y_i$$

③ 残差分散の不偏推定量

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2$$

単回帰の推定量のまとめ (2)

注意：教科書では最初から平均値を引いてあるので $\bar{x} = 0$

① 係数の推定量

$$\mathbb{E}\{\hat{a}\} = a$$

$$\mathrm{var}(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

→ 教科書(4.62)

② 切片の推定量

$$\mathbb{E}\{\hat{b}\} = b$$

→ 教科書(4.61)

$$\mathrm{var}(\hat{b}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

$$\mathrm{cov}(\hat{a}, \hat{b}) = \frac{-\sigma^2 \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

→ 教科書(4.63)

③ 残差分散の不偏推定量

$$\mathbb{E}\{\hat{\sigma}^2\} = \sigma^2$$

回帰式の推定量の分布　　教科書 (4.64)式

$$\hat{y} = \hat{a}x + \hat{b}$$

期待値

$$\mathbb{E}\{\hat{y}\} = \mathbb{E}\{\hat{a}\}x + \mathbb{E}\{\hat{b}\} = ax + b = y$$

分散

$$\begin{aligned}
\mathrm{var}\{\hat{y}\} &= \mathrm{var}\{\hat{a}x + \hat{b}\} = x^2 \cdot \mathrm{var}\{\hat{a}\} + 2x \cdot \mathrm{cov}\{\hat{a}, \hat{b}\} + \mathrm{var}\{\hat{b}\} \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right) + x^2 \cdot \frac{\sigma^2}{S_{xx}} + 2x \cdot \left(\frac{-\bar{x}\sigma^2}{S_{xx}} \right) \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} + \frac{x^2}{S_{xx}} - 2x \frac{\bar{x}}{S_{xx}} \right) \\
&= \sigma^2 \cdot \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}} \right)
\end{aligned}$$

予測値の分布

$$z = \hat{y} + \epsilon, \epsilon \sim N(0, \sigma^2)$$

期待値

$$\mathbb{E}\{z\} = \mathbb{E}\{\hat{y}\} + \mathbb{E}\{\epsilon\} = y$$

分散

$$\begin{aligned} \mathrm{var}\{z\} &= \mathrm{var}\{\hat{y}\} + \mathrm{var}\{\epsilon\} \\ &= \sigma^2 \cdot \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right) \end{aligned}$$

回帰に関する推定・検定

正規分布に従う確率変数の線形和は正規分布に従う(証明なし)。

また、回帰モデルの母数の推定量(と標本分布)が得られるので

(1) 回帰係数の検定：$y = a x + b$について $a \neq 0$

帰無仮説「a=0」を検定 (a=0なら回帰に意味がない)

(2) 真の回帰係数aの区間推定

確率95%で真の回帰係数aが入る信頼区間を求める

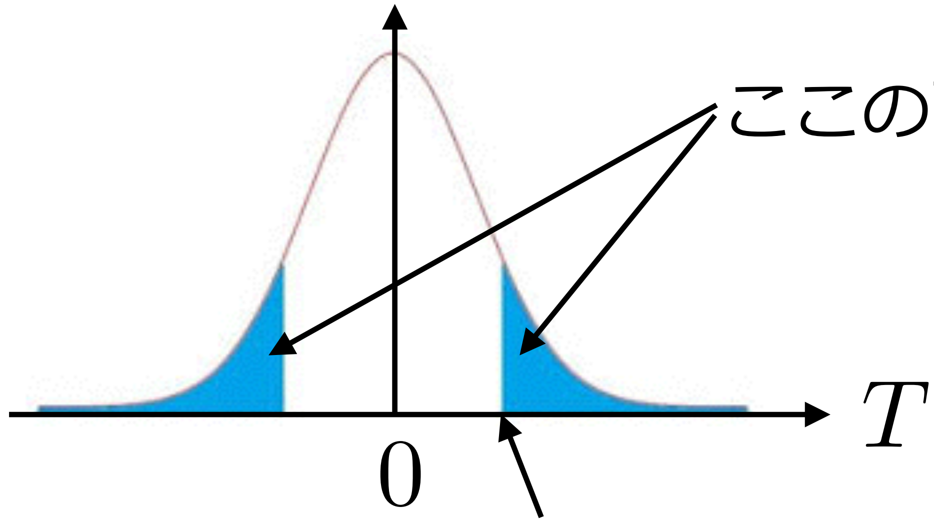
(3) 母回帰$a x + b$の区間推定

確率95%で母回帰$a x + b$が入る信頼区間を各xで求める

(4) yの予測区間の推定 (予測値の信頼区間)

確率95%でyが入る信頼区間を各xで求める

再掲：検定から区間推定へ



5%

裾確率　$P(|T| \geqslant t_\alpha) = \alpha$

95%

確率 $1-\alpha$ で $|T| < t_\alpha$

$$\Leftrightarrow \frac{\left|\bar{X}-\mu\right|}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} < t_\alpha$$

$$\Leftrightarrow \bar{X} - t_\alpha \sqrt{\frac{\widehat{\sigma^2}}{n}} < \mu < \bar{X} + t_\alpha \sqrt{\frac{\widehat{\sigma^2}}{n}}$$

回帰係数の区間推定と検定

区間推定

$$\hat{a} \sim N\left(a, \frac{\sigma^2}{S_{xx}}\right) \xrightarrow{\text{標準化}} \frac{\hat{a}-a}{\sqrt{\sigma^2/S_{xx}}} \sim N(0,1)$$

正規乱数の線形変換は
正規分布に従うので
(詳細省略)

不偏推定量を代入

$$\frac{\hat{a}-a}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t(n-2)$$

95%信頼区間 $a \pm t(n-2, 0.05)\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}$

仮説検定

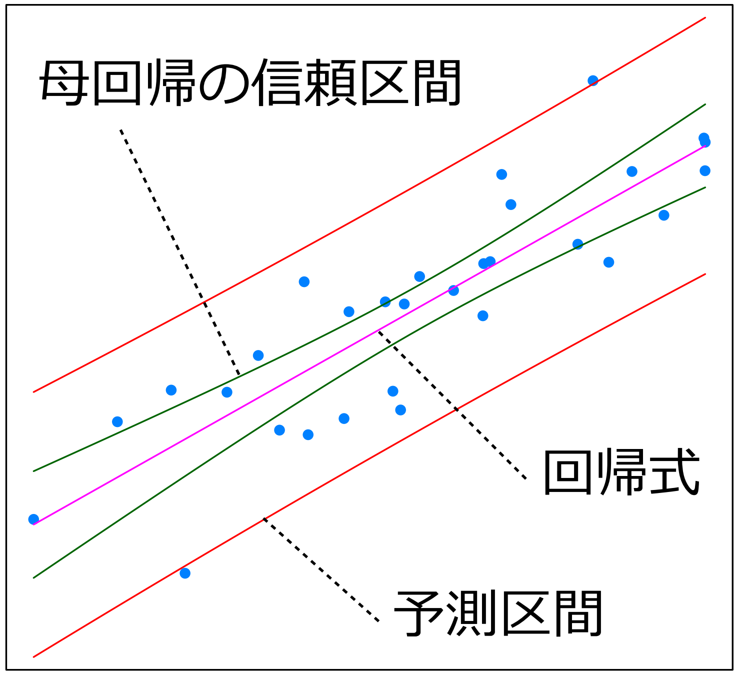
仮説 $a = 0$ を検定

もし $a = 0$ ならば $y_i = 0 \cdot x_i + b + Z_i$ であり y_i と x_i に関係がない

$$a = 0 \quad \Longrightarrow \quad t_0 := \frac{\hat{a}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t(n-2)$$

検定統計量 t_0 に対し $|t_0| \geqslant t(n-2, \alpha)$ なら有意水準 α で仮説を棄却

母回帰の推定: 信頼区間と予測区間



※回帰式は点 $(\bar{x}, \bar{y})$ を通る

回帰式 $$y = \hat{a}x + \hat{b}$$
$$= \hat{a}x + \hat{y} - \hat{a}\bar{x} = \hat{a}(x - \bar{x}) + \bar{y}$$

$$\mathbb{E}\{\hat{a}\,x + \hat{b}\} = a\,x + b$$
$$\mathbb{E}\{(\hat{a}\,x + \hat{b} - (a\,x + b))^2\} = \left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\sigma^2$$

より、点 x での母回帰の信頼率95%の区間推定(**信頼区間**)は

$$\hat{a}\,x + \hat{b} \pm t(n-2, 0.05)\sqrt{\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2}$$

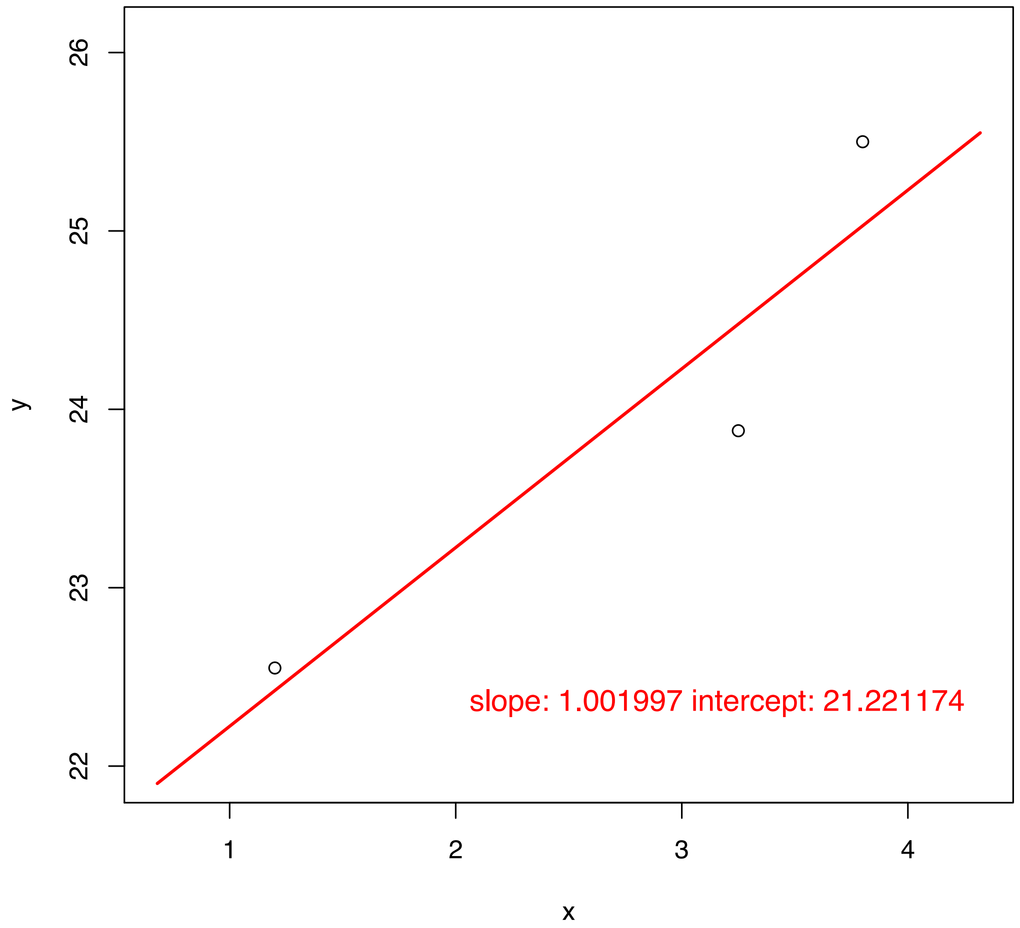
母回帰を用いて、$\hat{y} = \hat{a}x + \hat{b} + Z, Z \sim N(0, \hat{\sigma}^2)$ の関係より、点 x で目的変数 y の予測値の信頼率95%の区間推定(**予測区間**)が出来る。

$$\hat{a}\,x + \hat{b} \pm t(n-2, 0.05)\sqrt{\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)\hat{\sigma}^2}$$

例題：単回帰を計算してみよう

(1)直線の傾きと切片は?

(2) x=2.0のときのyの回帰による予測は?



x	y
3.8	25.5
1.2	22.55
3.25	23.88

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

❶〜❸を計算すること

残差分散の不偏推定量

❶ $$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a}x_i + \hat{b}))^2$$

$$\mathbb{E}\{\hat{\sigma}^2\} = \sigma^2$$

❷ $$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

σ^2 **は未知量なので**

$\hat{\sigma}^2$ **で代用**

(分布が正規分布→t分布に)

傾きの推定量

❸ $$\mathrm{var}(\hat{a}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

切片の推定量

❹ $$\mathrm{var}(\hat{b}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

(1)直線の傾きと切片は?

$$\boldsymbol{X} = \begin{bmatrix} 3.8 & 1 \\ 1.2 & 1 \\ 3.25 & 1 \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} 25.5 \\ 22.55 \\ 23.88 \end{bmatrix}$$

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 26.4 & 8.25 \\ 8.25 & 3.0 \end{bmatrix}, \boldsymbol{X}^\top \boldsymbol{y} = \begin{bmatrix} 202.0 \\ 71.9 \end{bmatrix}$$

$$\begin{bmatrix} 26.4 & 8.25 \\ 8.25 & 3.0 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 202.0 \\ 71.9 \end{bmatrix}$$ より

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} 1.0 \\ 21.2 \end{bmatrix}$$

❶ $\hat{\sigma}^2 = 0.5952517$

❷ $S_{xx} = 3.755$

❸ $\sqrt{\mathrm{var}(\hat{a})} = \sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}} = 0.3981487$

❹ $\sqrt{\mathrm{var}(\hat{b})} = \hat{\sigma}^2 \left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right) = 1.18205$

❺ $$T_a = \frac{\hat{a}}{\sqrt{\mathrm{var}(\hat{a})}} = 2.516641$$

❻ $$T_b = \frac{\hat{b}}{\sqrt{\mathrm{var}(\hat{b})}} = 17.95285$$

```
> x <- c(3.8, 1.2, 3.25)
> y <- c(25.5, 22.55, 23.88)
> res <- lm(y~x)
> summary(res)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3
 0.4712  0.1264 -0.5977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.2212     1.1821  17.953   0.0354 *
x             1.0020     0.3981   2.517   0.2408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7715 on 1 degrees of freedom
Multiple R-squared:  0.8636, Adjusted R-squared:  0.7273
F-statistic: 6.333 on 1 and 1 DF,  p-value: 0.2408

> coef(res)
(Intercept)           x
  21.221174    1.001997
> confint(res)
                2.5 %    97.5 %
(Intercept)  6.201800 36.240548
x           -4.056962  6.060957
> predict(res,data.frame(x=2.0),interval="confidence")
       fit      lwr      upr
1 23.22517 16.41121 30.03913
> predict(res,data.frame(x=2.0),interval="prediction")
       fit      lwr      upr
1 23.22517 11.28649 35.16385
```

4 6

3 5

$$\bar{x} = 2.75$$

$$t(1, 0.05) = 12.7062$$

$$\hat{b} \pm t(1, 0.05)\sqrt{\mathrm{var}(\hat{b})}$$

$$\hat{a} \pm t(1, 0.05)\sqrt{\mathrm{var}(\hat{a})}$$

$$2.0\hat{a} + \hat{b} \pm t(1, 0.05)\sqrt{\hat{\sigma}^2\left(\frac{1}{3} + \frac{(2.0 - \bar{x})^2}{S_{xx}}\right)}$$

$$2.0\hat{a} + \hat{b} \pm t(1, 0.05)\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{3} + \frac{(2.0 - \bar{x})^2}{S_{xx}}\right)}$$

例：単回帰を計算してみよう(教科書p.43～)

(1)直線の傾きと切片は?(例題1,p.49)

(2) x=5.0のときのyの回帰による予測は?(例題4,p.55)

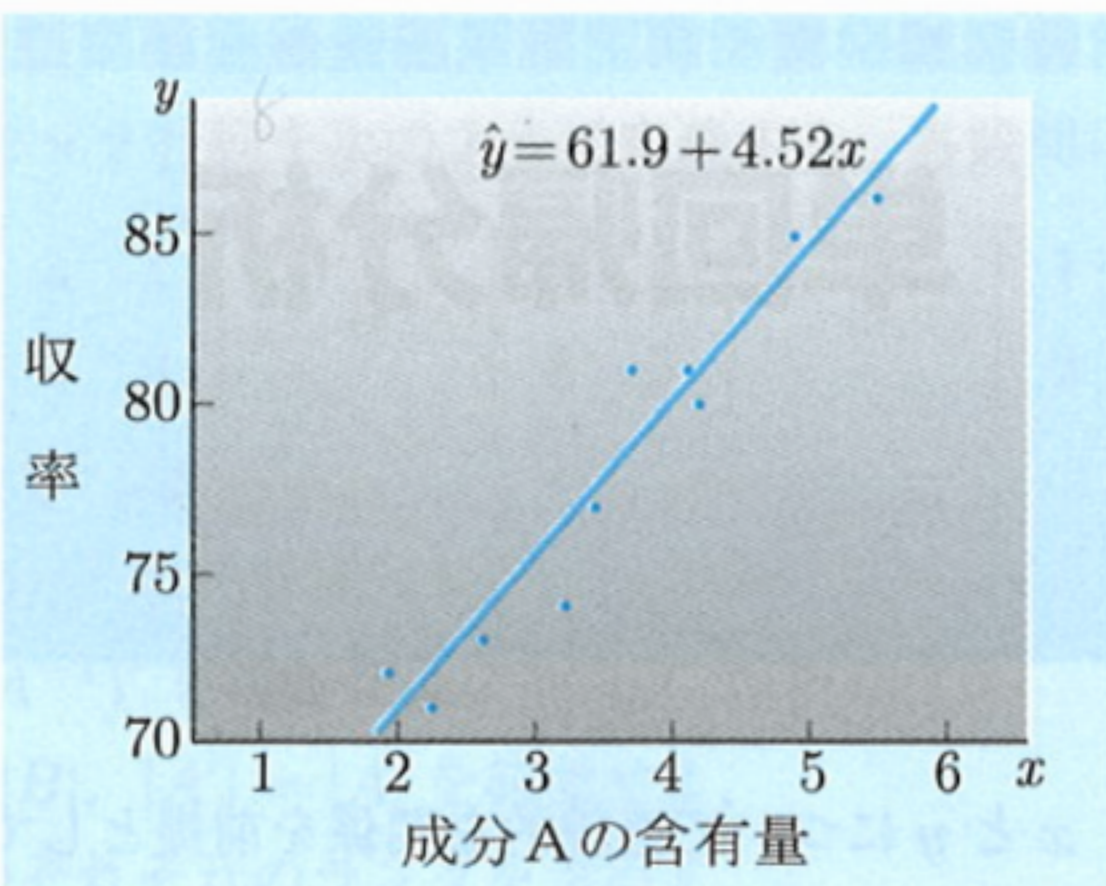


図 4.1　散布図

表 4.1　成分Aの含有量 x と収率 y のデータ

サンプル No.	含有量 x	収率 y
1	2.2	71
2	4.1	81
3	5.5	86
4	1.9	72
5	3.4	77
6	2.6	73
7	4.2	80
8	3.7	81
9	4.9	85
10	3.2	74

例：単回帰を計算してみよう(教科書p.43～)

(1)直線の傾きと切片は?(例題1,p.49)

傾き 4.52, 切片 61.9

(2) x=5.0のときのyの回帰による予測は?(例題4,p.55)

$y=4.52 \times 5.0 + 61.9 = 84.5$

(3) 傾き≠0かどうかの検定と傾きの区間推定

$T = 10.6 \geq t(8,0.05) = 2.306$ で有意である。

また、信頼率95%で、3.53 ≤ 傾き ≤ 5.51

(4) x=5.0のときの母回帰の信頼区間と予測区間

信頼率95%で $82.7 \leq 5.0\, a + b \leq 86.3$ (信頼区間)

信頼率95%で $80.7 \leq 5.0\, a + b + Z \leq 88.3$ (予測区間)