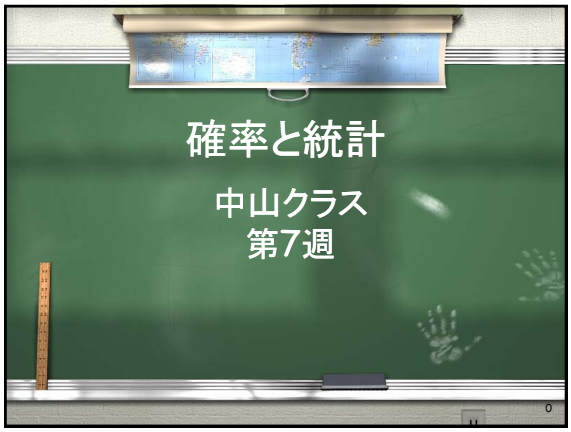


本日の内容

第2回レポート解説

第4章
4.4 推定値の精度を調べる方法

コンピュータ演習

1

第2回レポート解説

I. 以下に示す用語の意味を説明せよ.

相関 2つの量的変数どうしの関係を表す.
連関 2つの質的変数どうしの関係を表す.
共分散 2つの量的変数の相関を表す. 計算式は次式.

$$X = (x_1, x_2, \dots, x_n), \qquad Y = (y_1, y_2, \dots, y_n)$$

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$\mu_x = \{x_i\}$の平均, $\mu_y = \{y_i\}$の平均
共分散は元データの大きさによって変わるため, その大きさと相関は必ずしも比例しない.

3

相関係数
2つの量的変数の関係を表す. 次式で計算される.

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)^2 \quad S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_y)^2$$

共分散S_{xy}を標準偏差S_x, S_yで割っている(正規化)ので元データの大きさ影響を受けない. $-1 \le r_{xy} \le 1$の範囲の値をとる. $r_{xy} \approx -1, 1$の時に負または正の強い相関があり, $|r_{xy}| \ll 1$の時には相関がないと言える.

4

クロス集計表 2つの質的変数の関係を表す表.

ファイ係数 2つの質的変数の関係(連関)を数値で表したもの. 例えば,「好き」に0,「嫌い」に1を割り当て, 相関係数と同じ方法で計算され, 相関係数と同じ性質を持つ.

5

II. 第3章 練習問題, 及び, 以下の項目に対する解答を作成せよ.

(1)散布図を作成し, これから分かることを述べよ.
(2)相関係数を求め, これから分かることを述べよ.
(3)クロス集計表を求め, これから分かることを述べよ.
(4)ファイ係数を求め, これから分かることを述べよ.

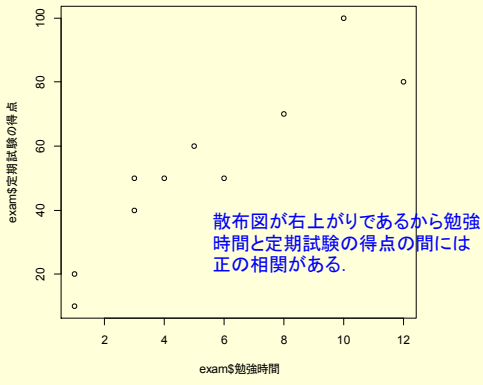
6

(1)散布図を作成し，これか分かることを述べよ

```
> exam <- read.csv("ch3_renshu-1.csv")
> exam
   勉強時間 定期試験の得点
1         1           20
2         3           40
3        10          100
4        12           80
5         6           50
6         3           50
7         8           70
8         4           50
9         1           10
10        5           60

> plot(exam$勉強時間, exam$定期試験の得点)
```

7



散布図が右上がりであるから勉強時間と定期試験の得点の間には正の相関がある.

8

(2)相関係数を求め，これから分かることを述べよ

```
> cor(exam[,1], exam[,2])
[1] 0.9092974
```

相関係数が1に近い値であるので勉強時間と定期試験の得点の間には正の強い相関がある.

9

(3)クロス集計表を求め，これから分かることを述べよ

```
> taste <- read.csv("ch3_renshu-3.csv")
> taste
   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
4          洋食       甘党
5          和食       辛党
6          洋食       辛党
7          洋食       辛党
8          和食       辛党
9          洋食       甘党
10         洋食       甘党
      ＜以下，省略＞
```

10

```
> table(taste$洋食派か和食派か,taste$甘党か辛党か)

     甘党 辛党
 洋食    6    4
 和食    3    7
```

洋食派－甘党，和食派－辛党の間に連関が認められる.
洋食派は甘党or辛党はそれほど明確ではないが，和食派には辛党が多いことが明かである.

11

(4)ファイ係数を求め，これから分かることを述べよ

```
> wayou <- ifelse(taste$洋食派か和食派か=="洋食",1,0)
> wayou
 [1] 1 0 0 1 0 1 1 0 1 1 0 1 0 1 0 0 1 1 0 0

> amakara <- ifelse(taste$甘党か辛党か=="甘党",1,0)
> amakara
 [1] 1 0 1 1 0 0 0 0 1 1 1 1 0 0 1 0 0 1 0 0

> cor(wayou, amakara)
[1] 0.3015113
```

12

ファイ係数が正の値であるので，洋食派(=1)と甘党(=1)の間，及び，和食派(=0)と辛党(=0)の間には正の連関（相関）があるが，その絶対値が大きくないので連関（相関）は強くない．

一方，
> amakara <- ifelse(taste$甘党か辛党か=="辛党",1,0)

とした場合の相関係数は-0.3015113となる．これは，洋食派(=1)と辛党(=1)の間，及び，和食派(=0)と甘党(=0)の間には負の連関（相関）があることを示している．

13

レポートの評価

A+ 5 ←期限内提出
A 4.5 ←
A- 4.25
B+ 4
B 3.75 ←遅れ提出
B- 3.5 ←
C+ 3.25
C 3

14

第4章 母集団と標本

4.4 推定値がどれくらいあてになるかを調べる方法

(1)標本抽出の方法→単純無作為抽出
(2)データの性質→確率変数
(3)確率変数のとる値→確率分布
(4)確率分布による母集団の表現→母集団分布
(5)代表的な母集団分布→正規分布
(6)Rを使って正規分布の母集団から標本抽出

15

4.4.1 標本抽出の方法－単純無作為抽出－

単純無作為抽出
母集団の中のどのデータも平等に選ばれる可能性を持っている．
→無作為標本

16

4.4.2 確率変数

例えば，「日本全国の17才男子全員の身長をxで表し，かつ，全員の身長データが分かっていない」ときに，xは確率変数である．
身長データが確定していないので確率的に扱う必要あり

例えば，「母集団に含まれる人数が10人であり，10人の身長が全て分かっている」とき，その身長をxで表しても，確率変数ではない．
全員の身長が確定しているので，確率的に扱う必要なし

単純無作為抽出によりデータが得られる場合は，身長を表すxは確率変数となる．
標本の身長データは分かるが，抽出するたびにデータが変わる（再現性がない）→確率的に扱う必要あり

17

4.4.3 確率分布

確率分布：確率変数がどのような値をどのような確率でとるかを表した分布．

サイコロの出る目	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

確率変数（サイコロの出る目）は上記の確率分布に従う．
「確率変数Xは確率分布Aに従う」

確率分布は非常に多くのデータの分布状況を表している．
サイコロを12回振ったとき「2の目は2回出る」ことは期待できない．しかし，600万回振れば，「2の目は100万回ぐらい出る」ことが期待できる．

18

ceiling(1.5)
　小数点以下を切り上げる　→　2
runif(n=10, min=0, max=6)
　0～6の範囲の一様乱数を10個発生させる.

```
> die <-ceiling(runif(n=6, min=0, max=6))
> table(die)
die
1 2 3 4 5
1 1 1 1 2
> die <-ceiling(runif(n=600, min=0, max=6))
> table(die)
die
  1  2  3  4  5  6
100 97 109 96 108 90
```

19

```
> set.seed(1)
> die <-ceiling(runif(n=6, min=0, max=6))
> table(die)
die
2 3 4 6
2 1 1 2

> set.seed(3)
> die <-ceiling(runif(n=6, min=0, max=6))
> table(die)
die
2 3 4 5
2 1 2 1
```

20

4.4.4　母集団分布

ある変数の母集団における分布を母集団分布という.

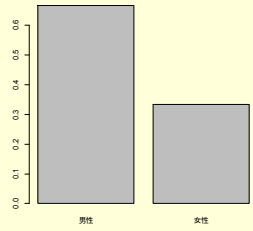
無作為抽出により得られた1つの標本データに関する確率分布は母集団分布を同じになる.

母集団分布は母集団からどのような値のデータが抽出されやすいか示した，標本の個々のデータに関する確率分布である.

21

性別	男性	女性
比率	2/3	1/3

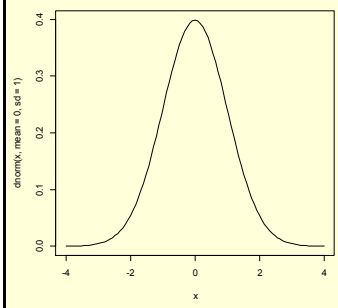
```
> barplot(c(2/3, 1/3), names.arg=c("男性","女性"))
```



22

4.4.5　正規分布

```
> curve(dnorm(x, mean=0, sd=1), from=-4, to=4)
```

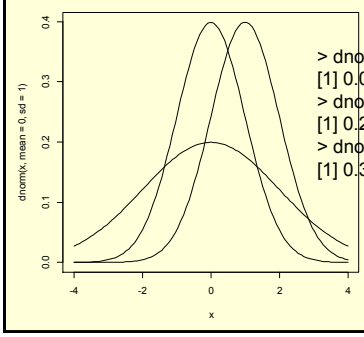


正規分布は
平均μ
分散σ^2（標準偏差σ）
で一意に決まる.

確率変数Xが正規分布$N(\mu, \sigma^2)$に従う
$X \sim N(\mu, \sigma^2)$

23

```
> curve(dnorm(x, mean=0, sd=1), from=-4, to=4)
> curve(dnorm(x, mean=1, sd=1), add=TRUE)
> curve(dnorm(x, mean=0, sd=2), add=TRUE)
```




```
> dnorm(2, mean=0, sd=1)
[1] 0.05399097
> dnorm(1, mean=0, sd=1)
[1] 0.2419707
> dnorm(0.5, mean=0, sd=1)
[1] 0.3520653
```

24

4.4.6 正規分布について少し詳しく

標準正規分布 $N(0,1)$

離散変数：サイコロの目のように，整数などとびとびの値をとる変数
確率分布：棒グラフ
$x = a$となる確率：$x = a$に対する棒グラフの高さ

連続変数：実数など連続的な値をとる変数
確率分布$N(\mu, \sigma^2)$：確率密度を表す．
xが$a \leq x \leq b$の値をとる確率：　面積で与えられる．

25

確率密度関数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

xが$a \leq x \leq b$の範囲の値をとる確率：
$a \leq x \leq b$における$f(x)$の面積

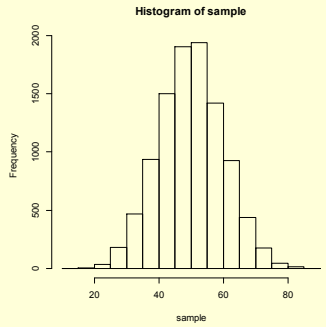
26

4.4.7 正規母集団から単純無作為抽出を行う

```
> rnorm(n=5, mean=50, sd=10)
[1] 38.47868 51.95783 50.30124 50.85418 61.16610
```

27

```
> sample <- rnorm(n=10000, mean=50, sd=10)
> hist(sample)
```



28

次回の予定

第4章　母集団と標本
　4．5　標本分布
　4．6　標本平均以外の標本分布

第3回レポート出題
　第4章
　　用語説明
　　練習問題と考察
　締め切り：2週間後

29