



## 第1回レポートの提出

- ◆ **締切** 10月17日(金)17:00  
(10月16日(木)の授業中にも受け取ります)  
**締切を過ぎてもレポートは必ず提出すること!**
- ◆ **提出場所**: 1号館2階 庶務課(1-206)の前にある  
レポートBOX(中山謙二/確率と統計)
- ◆ **レポート内容**: レポート課題に対する解答を作成し、  
別紙で作成した表紙と共に1つにまとめて左上をス  
テーブル留めし、提出する。用紙サイズはA4とする。
- ◆ **注意**: 「PCで作成して良い」と明記したものの以外は**手  
書きでレポート用紙に記入すること。**

3

## 講義用WebPageのURL

[http://leo.ec.t.kanazawa-u.ac.jp/  
~nakayama/edu/kit\\_prob\\_sta.html](http://leo.ec.t.kanazawa-u.ac.jp/~nakayama/edu/kit_prob_sta.html)

1

## レポート表紙の書式

用紙サイズ: A4

表紙の書式を講義WebPageからダウンロードし、  
レポートの回数、クラス・名番番号、名前を記入すること。

**注意**: 指定の書式を使用しない場合は1点減点。

4

## 今日の内容

### 第1回レポート出題

講義: 教科書第2章

1つの変数の記述統計. 概要のみをスライドで説明



コンピュータ演習:

教科書の内容を自分で行う。



残り時間でレポート作成を行う。

2

## 第1回レポートの課題

- I. 以下に示す用語の意味を説明せよ。  
質的変数, 量的変数, 度数分布, 代表値, 平均, 中央値, 最頻値, 不偏分散, 標本分散, 標準偏差, 標準化, 偏差値
- II. 第2章の練習問題に関して以下の項目に答えよ。  
Rの出力を使用する場合は「R Console」の画面を  
コピー&ペーストする。記述部分は手書き。
  - (1) ヒストグラムを作成し、これから分かることを述べよ。
  - (2) 平均と標準偏差を求め、2つの大学を比較せよ。
  - (3) データの標準化を行い、2つの大学を比較せよ。

5

## レポート作成上のヒント

◆**用語の説明** 教科書の説明部分をよく読み、自分の言葉で説明すること。例があると分かりやすい。自分のノートに作成し、そのコピーを提出してもよい。

### ◆**グラフの印刷方法**

グラフ表示のウィンドウを左クリック→RGuiのプリンターアイコンを左クリック→グラフ印刷

(別法)グラフ表示ウィンドウを右クリック→印刷を左クリック→印刷

### ◆**講義WebPageにあるレポート作成例を参考**

Rの画面やグラフを印刷した場合、その意味や特徴を説明する文章を手書きで挿入する。作成例はデータが異なる。

6

## 2.3 変数の種類

- 「性別」は「男」か「女」, 「数学」, 「統計」は「好き」か「嫌い」である。

→データを構成する人を分類する・・・**質的変数**

この場合の変数は2値 ... **二値変数**

指導法も、4種類の値で分類するから質的変数

- 「心理学テスト」「統計テスト1」「統計テスト2」は点数(数値)である。

→学力のレベルを示す ... **量的変数**

変数の種類によって適用できる統計解析が変わる。

(例えば、質的変数で平均を計算することはできない)

9

## 第2章の概要

### 1つの変数の記述統計

平均, 分散, 標準偏差, etc.

### 1つの変数の要約

数値要約=データの持つ特徴を1つの数値にまとめること。

### データの視覚的表現

### データの標準化

7

## 2.4 データの視覚的表現(1)

視覚的表現→データを図や表にする。

あるカテゴリに含まれるデータの個数・・・**度数**

全てのカテゴリの度数の分布状況 ... **度数分布表**

Rによる度数の計算 `table( )`

→P. 40~41を読んで、度数分布表を作成してみよう。

**度数分布表をグラフにする・・・ヒストグラム**

→P. 41~42を読んで、ヒストグラムを作成してみよう。

10

## 2.2 本書で用いるデータの説明

統計学の力が向上するよう、4種類の指導法を考え、被験者1名に1つずつ実施した。

表2.1 (p.38)のデータに含まれる「変数」

- 被験者を区別するID, 名前と性別
- 数学と統計で好きか嫌いか
- 「心理学」という試験科目のテスト得点
- 指導の前後で実施した統計テストの点数
- その被験者が受けた指導法

8

## 2.4 データの視覚的表現(2)

**度数分布**を得るには、データの範囲をいくつかの**階級**に分け、その階級に入るデータの個数を数えて**度数**とする。

例) 教科書P. 42のヒストグラムでは、4から2ごとに階級を設定している。

→階級を細かく分けすぎると、ヒストグラムが平らになり、データの特徴がわからなくなることに注意する。

→階級数の目安を示す式

スタージェスの公式 階級数 $k$ , データ数 $n$ として

$$k \cong 1 + \log_2 n$$

11

## 2.5 平均とは

データ集合を**代表の数値**に要約する(1つの数値表現).

**代表値**・・・分布の中心・・・**平均**(例えば)

**平均**=(データの総和)/(データ個数) **mean( )**

総和の計算 **sum( )**

→ P. 43~44を読んで, 平均を計算しよう.

注意: 関数を使わなくてもできるよう, 定義通りに計算する体験も必要であるから, P. 43の内容もやること.

12

## 2.8 分散, 標準偏差(1)

**分散** (データ(i) - 平均)<sup>2</sup>の平均

$$\{(d_1 - m)^2 + (d_2 - m)^2 + \dots + (d_n - m)^2\} / n$$

$d_i = i$ 番目のデータ,  $m =$ 平均,  $n =$ データの個数  
散らばりが大きい/小さい → 分散が大きい/小さい

**標準偏差** **sqrt(分散)**

分散はデータの2乗に対応

→ データと同じ大きさで散らばりを評価

→ 分散の平方根

15

## 2.6 平均以外の代表値

**中央値**: データを大きさ順に並べて真ん中に位置する値

→ **median( )**

**最頻値**: 最も頻繁に観測される値

Rでは関数はない → 度数分布表を計算 → 度数の最も大きいカテゴリのデータを最頻値とする.

実際によく使う代表値

**量的変数** → ほとんどは平均, ときどき中央値

**質的変数** → 最頻値(平均や中央値が計算できないため)

13

## 2.8 分散, 標準偏差(2)

分散には2種類ある. → **標本分散**と**不偏分散**

$$V = \frac{\sum_{i=1}^n (d_i - a)^2}{k} \quad \text{データの個数} = n$$

**標本分散**:  $k = n$  データ自体のばらつきを示す

**不偏分散 (var)**:  $k = n - 1$  データから母集団の値を推測

標本分散 = **var( )** \* (n-1)/n ... **var**との関係

標準偏差(不偏) = **sd()** = **sqrt(不偏分散)** ... **sd( )**の意味

標準偏差(標本) = **sqrt(標本分散)** = **sqrt(sd()<sup>2</sup> \* (n-1)/n)**

16

## 散布度

- 代表値に加えて, データの**ばらつき具合**も重要である. 代表値が同じでもばらつき具合が異なる場合がある.

- ばらつき具合を示す尺度・・・**散布度**

14

## 2.9 分散, 標準偏差以外の散布度

散布度の指標: (一般には)**分散**と**標準偏差**

それ以外の散布度の指標

**平均偏差** 平均からの偏差の**絶対値**の平均

|データ(i) - 平均|の総和 / データ個数

Rで絶対値を計算 **abs( )**

**範囲(レンジ)**

(データ中の最大値) - (データ中の最小値)

Rで最大値を計算 **max( )**

最小値を計算 **min( )**

17

## 2.10 標準化(正規化)

**標準化** 平均と標準偏差が特定の値になるように全てのデータを同じ式で変換する。

**標準得点** 変換後のデータの値

**z得点** 平均=0, 標準偏差=1となるように変換したときのデータの値

$$z\text{得点} = (\text{データの値} - \text{平均}) / \text{標準偏差}$$

\* **丸め誤差** 桁数の多い数値を最下位の桁で端数処理(四捨五入など)したときに生じる誤差。計算機で表現できる桁数が有限であるために生じる。

18

```
> read.csv("title.csv")
東京 金沢 大阪
1 1 10 100
2 2 20 200
3 3 30 300
4 4 40 400
5 5 50 500
```

全てのデータが正しく処理されている

```
> read.csv("no_title.csv")
X1 X10 X100
1 2 20 200
2 3 30 300
3 4 40 400
4 5 50 500
```

第1行目は表題として処理される  
2行目以降が数値データとして扱われている

22

## 2.11 偏差値

**偏差値** 平均50, 標準偏差10になるように標準化した標準得点。

$$\text{偏差値} = z\text{得点} \times 10 + 50$$

**使用例** 高校入試, 大学入試の模擬試験など  
9月の模試が350点, 12月の模試が400点  
順位は上がった? → 不明  
偏差値が50点→60点なら順位は上がった  
**偏差値** 全体の点数分布の変化に関わらず, 自分の順位が分かる。

19

第1行目が表題ではないことを指定する

```
> read.csv("no_title.csv", header=FALSE)
V1 V2 V3
1 1 10 100
2 2 20 200
3 3 30 300
4 4 40 400
5 5 50 500
```

仮の表題がつけられる  
全ての数値データは正しく処理される

23

## read.csv( )について

title.csv	第1行目に表題あり		
東京	金沢	大阪	
1	10	100	100
2	20	200	200
3	30	300	300
4	40	400	400
5	50	500	500

  

no_title.csv	第1行目に表題なし		
1	10	100	100
2	20	200	200
3	30	300	300
4	40	400	400
5	50	500	500

21

## 次回の予定

第3週: 10月16日(木)  
第3章 2つの変数の記述統計  
第3章の練習問題を解き, そこからわかることを第2回レポートとして出題する予定である。

24