

データ解析の基礎 ーデータの分類とまとめ方ー

高木 廣文
東邦大学看護学部
国際保健看護学研究室

e-mail: halwin@med.toho-u.ac.jp
http://homepage2.nifty.com/halwin/takagi.html

統計学と統計について

● 統計学 statistics とは何か？

- ・ 統計： 統計をとる（？）
- ・ 統計学： 統計学を使う（？）

2

「統計をとる」とは？

- ・ アンケート調査で学生のアルバイト実施を調べる。
- ・ ある病院の診療科別外来患者数を調べる。

⇒データを収集する（データをとる）。
集計をする。人数を数える。等々。

3

統計学とは？

- ・ 数理統計学
- ・ 生物統計学
- ・ 経済統計学
- ・ 看護統計学

- データに内在する傾向を明らかにするための科学的方法論を与える

4

統計学の対象は何か？

集団



個人

集団がもつ各項目や特性などの傾向についてデータから検討するための方法を提供

5

統計学の立場

● 統計学の特徴

- ・ 現象の数量化（データ化）
 - ・ 各種検査値
 - ・ 臨床的な症状や患者の性格特性
 - ・ QOLなど
- ・ 「質」を「量」として把握
 - ・ 客観性を高める。
「再現性」、 「比較可能性」

科学的な研究に不可欠

6

統計学的なものの捉え方

現象を統計学ではどのように把握するか？

- (1) 具体性, 現実性
現実的な具体的な現象のみを扱う。
- (2) 操作性
具体的に扱うために数字で表現する。
- (3) 変動性
対象を常に変動するものとする。
- (4) 傾向性
変動性の中に傾向性が存在する。

7

統計学を使う目的

統計学を目的により大きく2つに分けて考えることがある。

- (1) 特定の事象の「記述」
「記述統計学」 descriptive statistics
- (2) 調査結果や研究結果の「一般化」
「推測統計学」 inferential statistics
検定, 推定

8

記述統計学の方法

データをまとめて, ある特性を示す。

- 図や表を用いて示す方法
- ひとつの数値で示す方法

9

● データを図示する手法

「円グラフ」, 「帯グラフ」, 「棒グラフ」,
「ヒストグラム」, 「折れ線グラフ」,
「幹葉表示 stem-and leaf display」,
「箱ヒゲ図 box and whisker plot」,
「相関図(散布図)」など

10

● データを要約するための指標

「代表値」— 平均値, 中央値, 最頻値
「散布度」— 分散, 標準偏差, 変動係数
「相関係数」, 「割合」, 「クロス表」 など

11

推測統計学の方法

データから得た結果を一般化, 普遍化する。

- 推定 estimation
データ (標本) から一般集団 (母集団) の特性を求める:
⇒ 母平均値, 母比率の信頼区間など
- 検定 test
データから一般集団の特性に関する仮説を検証
⇒ 独立性の検定, 無相関の検定など

12

データ解析について

- ◆従来の多くの研究:
統計的検定の多用 — **確証的解析**
confirmative analysis
- 記述的方法を多用, データに依存した解析:
探索的データ解析 exploratory data analysis

幅広い知識や教養, また洞察力・直観力も極めて重要。方法の選択や結果の解釈を正しく行うには, 統計学に関する正確な知識も必要

13

データとは何か

- 例 1. A 子さんの身長は160cmです。
- ・ A 子さんの身長の測定結果。
 - ・ A 子さんの身長の「**データ** data」。

- 例 2. 学校で行われる身体計測
- ・ あるクラス30人の身長の一覧表。
 - ・ そのクラスの「**身長**の**データ**」。

14

ラベリングについて

- 例 3. A 子さんの血液型はA型です。
- ・ 血液型のデータを示している点は, 身長の場合と全く同じである。
 - ・ 血液型は, ある特定の反応の有無により A, B, O, AB の4タイプに分類
- ⇒ **標識付け**
(ラベリング labeling, ラベル付け)

15

データの定義

個体のある特性について測定を行い, 適当なラベルを付けたもの。
もしくは, その全体, およびそれらをまとめたもの

16

データの分類

- 1) 身長や体重のデータ
 - ・ ある「物差し」を用いて測定: 数値で表現
 - ・ データを足したり引いたりできる。⇒ **量的データ** quantitative data
- 2) 血液型のデータ
 - ・ ある特性に名称をつけたもの。
 - ・ それぞれを足したりすることは不可能。⇒ **質的データ** qualitative data

17

量的データの分類

- 例 1. A 子さんの体重は50Kg,
K 子さんの体重は60Kgです。
- (1) K 子さんはA 子さんより10Kg体重が重い。
 - ・ データの「**差**」の計算ができる。
 - (2) K 子さんはA 子さんの1.2倍の体重がある。
 - ・ データの「**比**」の計算ができる。
- ⇒ 「**比尺度** ratio scale」によるデータ

18

間隔尺度

例2. 一日の最高気温20°C, 最低気温10°C.

(1) 一日の気温の差は10°Cである。

⇒ 差の計算可能である。

(2) 最高気温は最低気温の2倍である?

⇒ × 比の計算不可。

「間隔尺度 interval scale」によるデータ

原点0のもつ意味による相違:

- ・ 負のデータの存在。
- ・ 比尺度に負のデータはない!

19

質的データの分類

例1. 血液型のデータの場合:

A子さんはA型, K子さんはAB型。

・ 差の計算: $A - AB = A(1 - B)$?

・ 比の計算: $AB / A = B$?

・ A型とB型の差や比を取ることは不可能。
どちらが大きいともいえない。

⇒ 「名義尺度 nominal scale」によるデータ

20

順序尺度

例2. あなたは寝起きはよいですか」の
ような質問項目への回答

1. 非常によい 2. よい 3. 悪い 4. 非常に悪い

・ 各カテゴリについての数値1~4の差や比は
計算できない。

・ 数値が大きくなるにつれ, 寝起きが悪く
なるという, 順序がある。

⇒ 「順序尺度 ordinal scale」によるデータ

21

データの分類再考

● データの測定尺度によるまとめ

(1) 量的データ:

(A) 比尺度によるデータ

(B) 間隔尺度によるデータ

(2) 質的データ:

(C) 順序尺度によるデータ

(D) 名義尺度によるデータ

変更可能

22

データの「質」と基本的統計手法

(1) 量的データ:

平均値, 分散, 標準偏差, 相関係数, など

(2) 質的データ:

人数, 割合, クロス表など

● 求められる基本統計量が異なる!

23

データのまとめ方

● 質的データの場合:

単純集計と度数分布表の作成

(1) カテゴリごとに人数を数える

(2) 人数から割合(%)などを求める

(3) 表や図にまとめる



度数分布表 frequency table

棒グラフ, 円グラフ, 帯グラフなど

24

度数分布表 frequency table

(例) AB0式血液型のデータ

| 血液型 | 度数 | 相対度数 |
|-----|----|-------|
| A | 17 | 42.5 |
| B | 8 | 20.0 |
| O | 12 | 30.0 |
| AB | 3 | 7.5 |
| 計 | 40 | 100.0 |

25

●用語について

度数：人数，個数，頭数，枚数など
frequency

割合：proportion

←昔は「比率」と誤って呼ばれていた。
今でも，その名残がある。

26

棒グラフ bar-graph

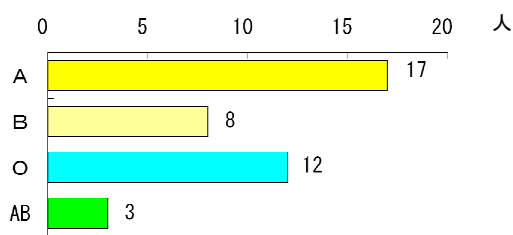


図 血液型の棒グラフ (人数)

27

円グラフ pie-graph

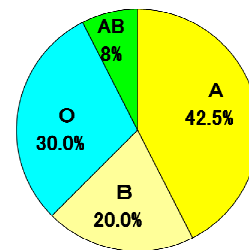


図 血液型の円グラフ

28

立体円グラフ

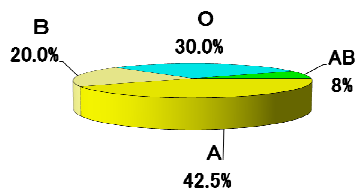


図 血液型の円グラフ

29

帯グラフ rectangular graph

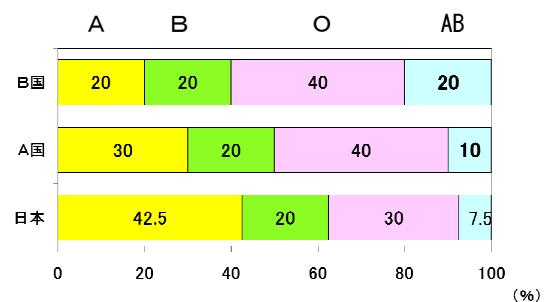


図 血液型の帯グラフ

量的データのまとめ方

- データの分布を調べる。
度数分布表の作成。

(例) 体重のデータ:

どのように人数を数えればよいのか?

→ 5Kg ごとに幅を決めて人数を数える。

区間, 階級 class の設定。

31

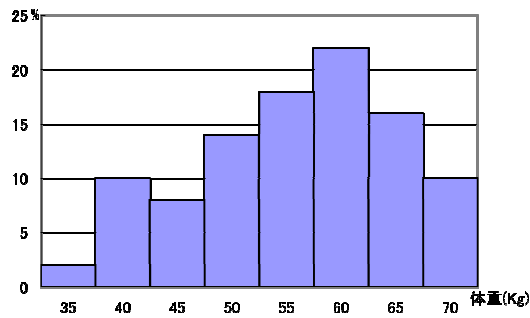
量的データの度数分布

表. 体重の度数分布表

| 区間 (Kg) | 度数 (%) | 累積度数 (%) |
|---------|------------|------------|
| 35 ~ 40 | 1 (2.0) | 1 (2.0) |
| 40 ~ 45 | 5 (10.0) | 6 (12.0) |
| 45 ~ 50 | 4 (8.0) | 10 (20.0) |
| 50 ~ 55 | 7 (14.0) | 17 (34.0) |
| 55 ~ 60 | 9 (18.0) | 26 (52.0) |
| 60 ~ 65 | 11 (22.0) | 37 (74.0) |
| 65 ~ 70 | 8 (16.0) | 45 (90.0) |
| 70 ~ 75 | 5 (10.0) | 50 (100.0) |
| 計 | 50 (100.0) | |

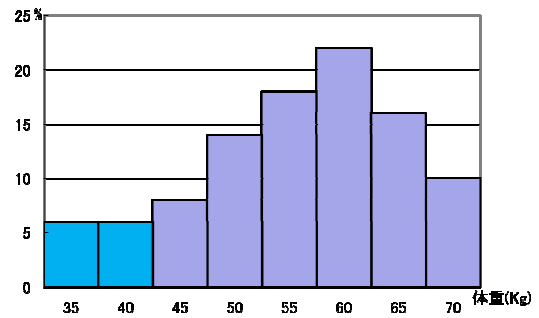
32

ヒストグラム histogram



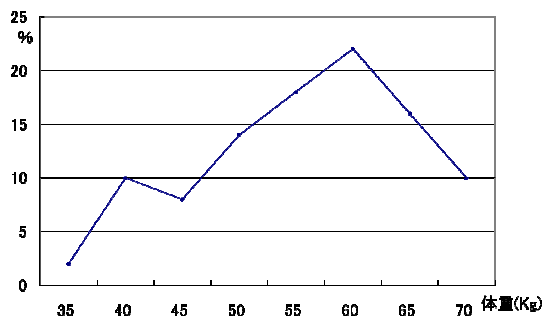
33

ヒストグラム histogram 2



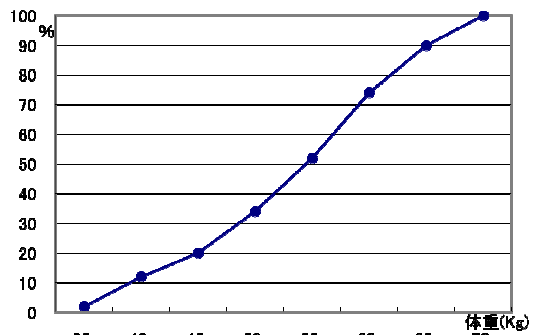
34

折れ線グラフ frequency polygon



35

累積折れ線グラフ



36

幹葉表示 stem-and-leaf display

| | |
|------------------|------------|
| 3· 9 | 39 |
| 4* 02224 | 4042424244 |
| 4· 6788 | ⋮ |
| 5* 0012444 | ⋮ |
| 5· 566777888 | ⋮ |
| 6* 01111222344 | ⋮ |
| 6· 56677789 | ⋮ |
| 7* 00004 | 7070707074 |

図. 体重の幹葉表示

37

分布の代表値

● 分布の代表値とは

代表値 : average

→ 分布を代表する値とは何か？

- (1) 分布の真中辺のデータの値
- (2) 最も多いデータの値

→ 分布の「位置の尺度」とも呼ばれる

38

分布の3つの代表値

- (1) 平均値 mean
- (2) 中央値 median
- (3) 最頻値 mode

39

平均値 mean

全データの総和を標本数で割ったもの:

$$\text{平均値} = \frac{\text{データの合計}}{\text{標本数}}$$

記号 : n 個のデータを x_1, x_2, \dots, x_n

$$\text{平均値 } \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

40

平均値の例:

例) 5人の体重のデータ : 50, 45, 60, 70, 55Kg

$$\begin{aligned} \text{平均値} &= \frac{50+45+60+70+55}{5} \\ &= 280 \div 5 = 56 \text{ (Kg)} \end{aligned}$$

41

中央値 median

データを大きさの順に並べた場合、ちょうど真ん中の順位にくるデータのもつ値。
N個のデータ大きさの順に並べる:

$$x_1 \geq x_2 \geq \dots \geq x_n$$

● Nが奇数 : 中央値 $x_{Med} = x_{\left(\frac{n+1}{2}\right)}$

● Nが偶数 : 中央値 $x_{Med} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$

42

中央値の例1:

例) 5人の体重のデータ : 50, 45, 60, 70, 55Kg

(1) まず大きさの順に並べ替える :

→ 45, 50, 55, 60, 70 (Kg)

(2) 標本数5は奇数なので,

$$(5+1)/2 = 3 \text{ 番目}$$

のデータが中央値 → 中央値=55 (Kg)

43

中央値の例2:

6人の体重のデータ : 50, 45, 65, 60, 70, 55Kg

(1) まず大きさの順に並べ替える :

→ 45, 50, 55, 60, 65, 70 (Kg)

(2) 標本数6は偶数なので, $6/2 = 3$ 番目

と4番目のデータの平均値が中央値 :

→ 中央値 = $(55+60)/2 = 57.5$ (Kg)

44

最頻値 mode

最も人数(度数)の多いデータのもつ値。
実際には、標本数が少ない場合、データが連続的なことから、各データの人数は少なくなり、どのデータが最頻値かを定めるのは困難。



度数分布表の利用

最も度数の多い区間の真中の値(級心)を最頻値とする。

45

その他の位置の尺度

● 最小値 minimum value : データ中最小の値
● 最大値 maximum value : データ中最大の値

● パーセンタイル percentile (百分位) :
大きさの順にデータを並べ、小さい方から累積して何パーセントの点にあるかを示す。

→ 5, 10, 25, 50, 75, 90, 95パーセンタイル
(第1, 2, 3四分位 quartile)

46

分布の散布度

各データは異なった値を持つので、その分布には広がりがある。
そのばらつき具合、代表値からの平均的な散らばり具合を示す。

- 1) 分散 variance, Var
- 2) 標準偏差 standard deviation, SD
- 3) 変動係数 coefficient of variation, CV
- 4) 範囲 range, R
- 5) 平均偏差 mean deviation

47

偏差について

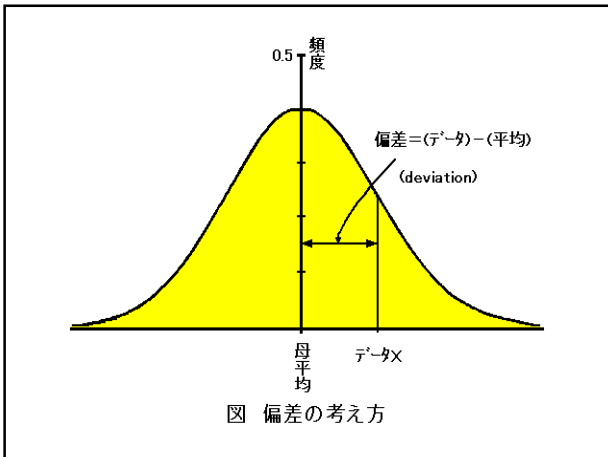
代表値とデータとの差、普通は代表値として平均値を用いる。

偏差 deviation = [データ] - 平均

(例) 身長が180cmの場合、平均身長が170cm

身長の偏差 = $180 - 170 = 10$ (cm)

48



偏差に基づく散布度

分布の散布度をどのようにして求めればよいか

●偏差の平均値は？

→ 偏差の合計は常に0。
使用不可。

偏差に正負があるので、全て正にすればよい。

→ 偏差の絶対値, 偏差の2乗(平方)

50

代表値からの平均偏差

◎代表値からの偏差の絶対値の平均値:

n 個のデータを x_1, x_2, \dots, x_n

$$\text{平均偏差} = \frac{\text{各ケースの偏差の絶対値の合計}}{\text{標本数}} = \frac{1}{n} \sum_{i=1}^n |x_i - x_{Ave}|$$

●統計的な扱いが難しいため、実際には、ほとんど使用されない。

51

分散 variance

◎平均値からの偏差の平均平方和:

n 個のデータを x_1, x_2, \dots, x_n

$$\text{分散 } s^2 = \frac{\text{偏差の2乗の合計}}{\text{標本数}} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

52

分散の例:

◎5人の体重のデータ: 50, 45, 60, 70, 55Kg

平均値=56Kg

$$\text{分散} = [(50-56)^2 + (45-56)^2 + (60-56)^2 + (70-56)^2 + (55-56)^2] / 5$$

$$= [36 + 121 + 16 + 196 + 1] / 5$$

$$= 370 / 5$$

$$= 74 \text{ (Kg}^2\text{)}$$

53

標準偏差 standard deviation (SD)

分散は偏差の2乗の合計から計算

→ 単位も2乗: 体重=Kg², 身長=cm², etc

●分散の平方根を計算し, 単位を戻す。

$$SD = \sqrt{\text{分散}}$$

54

標準偏差の例:

◎ 5人の体重のデータ : 50, 45, 60, 70, 55 Kg

平均値=56 (Kg)

分散=74 (Kg²)



標準偏差 $S = \sqrt{74} = 8.6$ (Kg)

55

変動係数 CV(Coefficient of Variation)

50人の身長標準偏差は5 cm, 体重標準偏差は5 kgであった。

Q. 身長と体重のばらつき具合はどちらが大きいのか, それとも等しいのか?

単位が異なるので比較できない!
単位をそろえる必要がある。

56

変動係数の定義

$$\text{変動係数} = \frac{\text{標準偏差}}{\text{平均値}} \times 100$$

平均値を100としたときの標準偏差の大きさの程度を示す。

57

変動係数の計算例:

◎ 5人の体重のデータ : 50, 45, 60, 70, 55 Kg

平均値=56 (Kg)

分散=74 (Kg²)

標準偏差= 8.6 (Kg)



$$\text{変動係数} = \frac{8.6}{56} \times 100 = 15.4$$

58

● 散布度に関するその他の話題

▼ 偏差値とは何か?

偏差は平均値からの差。

→ データの標準化 standardization

標準化:

データの平均が0, 分散が1になるようにデータを変換すること。

59

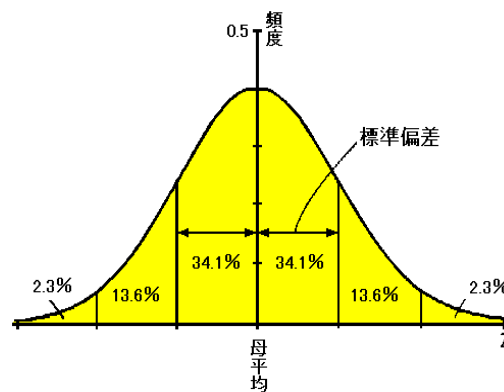


図 正規分布

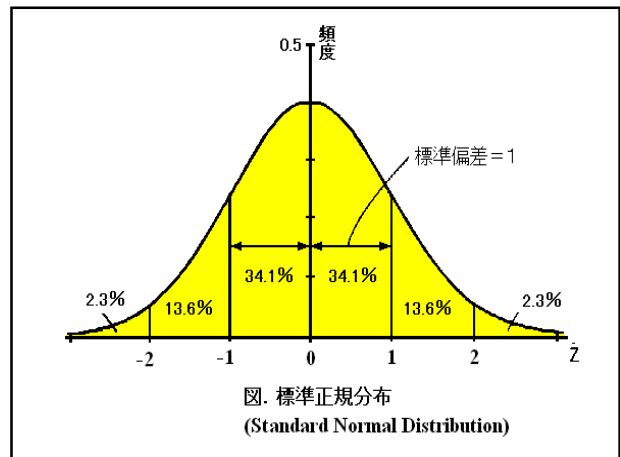
● データの標準化と偏差値

平均が μ (ミュウ), 標準偏差が σ (シグマ) の場合, あるデータ x を,

$$z = \frac{x - \mu}{\sigma} \quad (\text{平均 } 0, \text{ 分散 } 1)$$

偏差値 $T = 10 \times z + 50$
平均50, 標準偏差10

61



偏差値の計算例:

統計学の平均値が75点, 標準偏差15点の場合:
A君90点, B君60点の偏差値は?

$$\text{A君の偏差値} = \frac{90 - 75}{15} \times 10 + 50 \doteq 60$$

$$\text{B君の偏差値} = \frac{60 - 75}{15} \times 10 + 50 \doteq 40$$

63