健康統計学 第3回

今回は、代表値と散布度（テキスト15〜32ページ）について学習します。

テキスト

- 『やさしい保健統計学 改訂第4版』縣 俊彦著（南江堂）

今回の内容

ギリシャ文字

統計学では、数式などでギリシャ文字がよく使われます。文字と読み方を覚えておきましょう。

大文字	小文字	読み方	使用される統計量
A	α	アルファ	確率、第1種の過誤、有意水準
B	β	ベータ	第2種の過誤
Γ	γ	ガンマ	ガンマ関数
Δ	δ	デルタ	分散
E	ϵ	イプシロン	
Z	ζ	ゼータ	
H	η	イータ	
Θ	θ	シータ	定数、推定値など
I	ι	イオタ	
K	κ	カッパ	
Λ	λ	ラムダ	定数など
M	μ	ミュー	母平均
N	ν	ニュー	
Ξ	ξ	グザイ	
O	$\emptyset$	オミクロン	
Π	π	パイ	円周率
P	ρ	ロー	相関係数
Σ	σ	シグマ	分散、標準偏差
T	τ	タウ	
Υ	υ	ウプシロン	
Φ	ϕ	ファイ	空事象
X	χ	カイ	χ^2（カイ二乗）検定
Ψ	ψ	ブザイ	
Ω	ω	オメガ	全事象

代表値（average）

- データの分布などの特徴を示す数値（特性値）を「**代表値**」という。
- データ全体を**ひとつの値で代表**させる値である。

平均値（mean）

算術平均（arithmetic mean）

- 算術平均 $\bar{x}$ は、データをすべて足しあわせ、データ数で割ったもの。
- 平均値のなかで、もっとも一般的なもの。

$$\begin{aligned}\bar{x} &= \frac{1}{n}\left(x_1+x_2+\cdots+x_n\right)\\ &= \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

幾何平均（geometric mean）

- 幾何平均 Gm は、各データの値の積に対してデータ数のべき根を求めたもの。

$$\begin{aligned}Gm &= \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}\\ &= \left(x_1 \times x_2 \times \cdots \times x_n\right)^{1/n}\\ &= \left(\prod_{i=1}^{n} x_i\right)^{1/n}\end{aligned}$$

- 幾何平均の例
  - 5年間の物価上昇率が7%のとき、1年の平均上昇率は何%か？
  - 過去3年間の売上高の対前年比が120%、110%、130%のとき、平均の売上高の伸びは？

調和平均（harmonic mean）

- 調和平均 Hm は、データ数を各データの値の逆数の和で割ったもの。

$$\begin{aligned}Hm &= \frac{n}{\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}}\\ &= \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}\end{aligned}$$

- 調和平均の例
  - 山頂まで6kmの道のりを、往きは2km/hで、帰りは6km/hで歩いたとき、平均の速さはいくらか？
  - 車でドライブをして、最初の24kmは30km/h、次の24kmは40km/h、最後の24kmは60km/hで走った時、平均速度はいくらか？

中央値（median）

中央値（中位数）

- 中央値 Me は、データを大きさの順に並べたときに、中央にくる値のことである。
  - データ数が奇数のときは中央にくるデータの値になる。
  - データ数が偶数のときは中央にある2つのデータの平均の値になる。

$$Me=\begin{cases} x_m & \text{if } n \text{ odd, } m=(n+1)/2 \\ \frac{x_m+x_{m+1}}{2} & \text{if } n \text{ even, } m=n/2 \end{cases}$$

- 中央に位置するデータが複数個ある場合（「結び（tie）」があるという）、次のような式で中央値を求めることができる。

$$Me=\frac{1}{2n_M}(n_{x>M}-n_{x<M})+M$$

  - 中央にあるデータ：M
  - 値 M になるデータの個数：n_M
  - 値 M より小さいデータの個数：$n_{x<M}$
  - 値 M より大きいデータの個数：$n_{x>M}$
- 度数分布表がある場合は、階級や度数などの情報から、中央値を求めることもできる。

$$Me=l_m+\left(\frac{n}{2}-F\right)\frac{h}{f_m}$$

  - 標本数：n
  - 階級幅：h
  - m 番目の階級の下限：l_m
  - m 番目の階級の度数：f_m
  - m-1 番目までの累積度数：F

四分位数（quartile）

- ヒストグラムから考えると、四分位数はヒストグラムの面積を1/4ずつに分ける値である。
  - 中央値は、ヒストグラムの面積を半分に分ける値になる。
- データを大きさの順に並べた場合は、データの個数を4分の1ずつの部分にわける個所である。
- 小さいほうから、第1、第2、第3四分位数といい、中央値は、第2四分位数になる。
- データが n 個のあるときの第1四分位数 Q_1 と第3四分位数 Q_3 は、次のようにして求められる。
  - $n=4k+1,\ 2,\ 3$ の場合

$$Q_1=x_{k+1}$$
$$Q_3=x_{n-k}$$

  - $n=4k$ の場合

$$Q_1=(x_k+x_{k+1})/2$$
$$Q_3=(x_{n-k}+x_{n-k+1})/2$$

百分位数（percentile）

- 百分位数（パーセンタイル値）は、ヒストグラムの面積を1/100ずつに分ける値である。
  - 25パーセンタイル値は第1四分位数である。
  - 50パーセンタイル値は中央値（第2四分位数でもある）。
- 度数分布表がある場合は、階級や度数などからパーセンタイル値 p を求めることもできる。

$$p=l_m+\left(\frac{n\times p}{100}-F\right)\frac{h}{f_m}$$

  - 標本数：n

- 階級幅：h
- m 番目の階級の下限：l_m
- m 番目の階級の度数：f_m
- m-1 番目までの累積度数：F

最頻値（mode）

- 最頻値 Mo は、データのなかで**最も多く出てくる値**のことである。
  - 度数分布表がある場合は、もっとも度数の多い階級値を最頻値として、次の式から最頻値を求めることができる。

$$Mo = l_m + \frac{f_{m+1}}{f_{m-1} + f_{m+1}} \times h$$

    - 最大度数の階級：m
    - 階級幅：h
    - m 番目の階級の下限：l_m
    - m 番目の階級の度数：f_m
- 分布が釣り鐘形の場合は、ピアソン（Pearson）の式を用いることができる。

$$Mo = \bar{x} - 3 \times (\bar{x} - Me)$$

代表値の特性

- 平均値はすべてのデータを反映している。
  - ハズレ値（極端に小さく・大きくて飛び離れたデータ）があるとその影響を受けやすいため、ハズレ値の考慮が必要。
- 中央値（四分位数や百分位数も）は分布上の位置（中央など）を示す。
  - ハズレ値の影響を受けにくく、分布に偏りがある場合に優れている。
- 最頻値は、「データの多くはこのあたりにある」という説明をするのにわかりやすい。
  - ハズレ値の影響を受けにくい。

散布度（dispersion）

- 代表値のほかに、重要な特性値として「**散布度**」がある。
- 平均値に対して、**どれくらいデータが散らばっているか**を示す。
  - 分布の裾の広がり具合
  - 平均値への集中の度合い

標準偏差

偏差（diviation）

- 偏差 D は、各データと平均との差である。
  - ＋の偏差と－の偏差があるため、すべての偏差の合計は0になる。

$$D_i = \overline{x} - x_i$$

分散(variance)と標準偏差(standard deviation)

分析対象となる全体（母集団）の分布のバラつきの度合い求める場合には、代表的な散布度である、分散と標準偏差を用いる。

- 分散 s^2（または σ^2）は、**偏差平方和**（偏差の二乗の和）をとって、その平均を求めたものである。
  - 全データの平均からのバラツキの程度を示す。

$$\begin{aligned} s^2 &= \frac{1}{n}\sum_{i=1}^{n}(\overline{x} - x_i)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} D_i \end{aligned}$$

- 標準偏差 s は、分散の平方根を求めたものである。
  - 全データの平均からのバラツキの程度を示す（単位はデータと同じ）。

$$s = \sqrt{\frac{\sum_{i=1}^{n}(\overline{x} - x_i)^2}{n}}$$

- 標準偏差や分散の値が大きい場合はデータのバラつきが大きく、小さい場合はバラつきが小さい（データが同じ程度に揃ってる）

不偏分散（unbiased variance）と不偏標準偏差（unbiased standard diviation）

分析対象となる全体（母集団）ではなく、対象の一部分（標本）の分布のバラつきの度合い求める場合には、不偏分散と不偏標準偏差を用いる。

- 不偏分散 U^2 は、偏差平方和 (偏差の二乗の和) をとって、その平均を求めたものである。
    - 分散との違いは、分母は「標本数-1」であること。
    - データ全体についての平均値からのバラツキの程度を示す。

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

- 不偏標準偏差 U は、分散の平方根を求めたもの
    - 全データの平均からのバラツキの程度を示す (単位はデータと同じ) 。

$$U=\sqrt{\frac{\sum_{i=1}^{n}\left(\bar{x}-x_i\right)^2}{n-1}}$$

標準偏差の和

n 組の資料（データ）があるとき、資料全体の標準偏差は次のようになる。

$$s_T=\sqrt{\frac{\sum_{i=1}^{n}N_i\left(V_i+D_i^2\right)}{\sum_{i=1}^{n}N_i}}$$

- s_T …全体の標準偏差
- N_i …組目の資料の標本数
- V_i …組目の資料の分散
- D_i …組目の資料の偏差

範囲（range）

- 範囲 R は、データの最大値 x_{max} と最小値 x_{min} との差で、データ全体の範囲を示す。
- ハズレ値の影響を受けやすい

$$R=x_{max}-x_{min}$$

四分位偏差（quartile deviation）

- 四分位偏差はデータの変動の目安に利用される散布度で、代表値として中央値を用いたときに使われることがある。
- ハズレ値やデータ数に影響されにくい値である。

四分位偏差=(第3四分位数-第1四分位数）/2

平均偏差（mean deviation）

- 平均偏差 M_{dev} は、偏差の絶対値を平均したもので、データと平均値とのずれの程度を示す。

$$M_{dev} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$$
$$= \frac{1}{n}\sum_{i=1}^{n}|D_i|$$

変異係数（coefficient of variance）

- 変異係数（変動係数） Cv は、標準偏差を平均で割ったもので、平均値に対する標準偏差の割合を示す（%表示）。
- 変異係数は相対的な散布度（割合を示す無名数で単位はない）で、平均値や標準偏差の異なる複数の種類のデータを比較するときに用いる。

$$Cv = s/\overline{x}$$

- 2つの系列（データの集まり）を比較するとき、次のような場合は相対的散布度が有利になる。
  - 双方の単位が同じで、平均がほぼ等しい
  - 双方の単位は同じだが、平均が違う
  - 双方の単位が違う

歪度と尖度

教科書には載っていませんが、散布度に関連する、分布の特徴を表すための重要な指標について説明します。

歪度（skewness）

- 歪度（わいど）は、分布の左右対称性の違いを表す。

$$Sk=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3\Big/s^3$$

  - データの個数：n
  - データ全体の平均値：$\bar{x}$
  - データ全体の標準偏差：s

- 歪度 Sk の値によって分布の左右対称性がわかる。
  - $Sk=0$ の場合、データは左右対称に分布
  - $Sk>0$ の場合、左に偏った分布
  - $Sk<0$ の場合、右に偏った分布
  - ただし、$Sk=0$ だからといって、常に分布の形が左右対称とは限らない

尖度（kurtosis）

- 尖度（せんど）は、分布の形が先がとがっているか偏平かを表す。

$$Ku=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4\Big/s^4$$

  - データの個数：n
  - データ全体の平均値：$\bar{x}$
  - データ全体の標準偏差：s

- 尖度 Ku の値によって分布の尖り具合がわかる。
  - Ku が大きいほど、尖った形の分布
  - Ku が小さいほど、偏平な形の分布

Excelで代表値と散布度を計算

数式の入力

Excelでは、セルに「**数式**」を入力することで、計算ができます。数式を入力するときの基本的なルールは、次のとおりです。

- 最初は「=」ではじめる
- カッコ「()」を使って計算する順番を指定できる
- 四則演算が使える（半角で入力）

演算	数学での記号	Excelでの記号	計算式の例	表示される結果
足し算	＋	+	=1+2	3
引き算	−	-	=2-3	-1
掛け算	×	*	=4*5	20
割り算	÷	/	=1/2	0.5
べき乗	ˆ	^	=2^3	8

数式の入力例

たとえば、身長と体重のデータから人の肥満度をはかる指標である、BMI（ボディマス指数）を計算する場合を考えてみましょう。

```
BMI＝体重（kg）÷身長（m）の2乗
```

身長のデータがB2～B11セルに、体重のデータがC2～C11セルに入力されており、それらから求めたBMIをD2～D11セルに表示させるには、次のように操作します。

1. D2セルに次の計算式を入力する

   ```
   =C2/((B2/100)^2)　（「/100」としているのは、身長がcm単位のため）
   ```

2. 「Enter」キーを押すと、計算結果が表示される
3. D2セルの計算結果を、D3～D11セルへコピーする

平方根、n 乗根の計算

- 正の平方根（√）を計算するには、**SQRT**関数を利用します。

  SQRT（平方根を計算する）

  - 書式：SQRT(数値)
  - 引数：平方根を求める数値
  - 例：A12セルの数値の平方根を計算する

    ```
    =SQRT(A12)
    ```

- n乗根を計算する関数はないため、べき乗（^）を利用する
  - 「n乗根の計算」は、「1/n のべき乗の計算」と同じ意味になることを利用する

    - 例：A12セルの数値の4乗根 $\sqrt[4]{A12}$ を計算

```
=A12^(1/4)
```

代表値を計算

平均値

- 算術平均は、**AVERAGE**関数を利用します。

AVERAGE (平均値を計算する)

- 書式 : AVERAGE(数値1, 数値2, ...)
- 引数 : 数値1, 数値2, ... :平均を計算するセルの範囲
- 例:F1～F10セルまでのセルの数値の平均値を計算する

```
=AVERAGE(F1:F10)
```

- 幾何平均は、**GEOMEAN**関数を利用します。

GEOMEAN (正の数からなる配列またはセル範囲のデータの幾何平均を計算する)

- 書式 : GEOMEAN(数値1, 数値2, ...)
- 引数 : 数値1, 数値2, ... :平均を計算するセルの範囲

- 調和平均は、**HARMEAN**関数を利用します。

HARMEAN (1 組の数値の調和平均を計算する)

- 書式 : HARMEAN(数値1, 数値2, ...)
- 引数 : 数値1, 数値2, ... :平均を計算するセルの範囲

中央値

- 中央値は、**MEDIAN**関数を利用します。

MEDIAN (引数に含まれる数値の中央値を求める)

- 書式 : MEDIAN(数値1, 数値2, ...)
- 引数 : 数値1, 数値2, ... :中央値を計算するセルの範囲
- 例:F1～F10セルまでのセルの中央値を求める

```
=MEDIAN(F1:F10)
```

四分位数

- 四分位数は、**QUARTILE**関数を利用します。

QUARTILE (配列に含まれるデータから四分位数を抽出する)

- 書式 : QUARTILEE(配列, 戻り値)
- 引数 : 配列 :対象となるデータを含む配列 (セルの範囲)
- 引数 : 戻り値 :戻り値として返す四分位数の内容を指定
  - 戻り値: 0: 最小値
  - 戻り値: 1: 第1四分位数 (25%)
  - 戻り値: 2: 第2四分位数 (50%) =中央値
  - 戻り値: 3: 第3四分位数 (75%)

- 戻り値: 4: 最大値

百分位数

- 百分位数は、**PERCENTILE**関数を利用します。

> **PERCENTILE (配列に含まれるデータから百分位数 (%) を抽出する)**
>
> - 書式 : QUARTILEE(配列, 率)
> - 引数 : 配列 :対象となるデータを含む配列 (セルの範囲)
> - 引数 : 率 :0〜1の値で、目的の百分位の値 (パーセンタイル値) を指定

最頻値

- 最頻値は、**MODE**関数を利用します。

> **MODE (引数に含まれるデータのなかで最も頻繁に出現する値を求める)**
>
> - 書式 : MODE(数値1, 数値2, ...)
> - 引数 : 数値1, 数値2, ... :最頻値を計算するセルの範囲
> - 例 : F1〜F10セルまでのセルの最頻値を求める
>
> ```
> =MODE(F1:F10)
> ```

散布度を計算

分散

- 分散は、**VARP**関数を利用します。

> **VARP (引数を母集団全体と見なし、母集団の分散 (標本分散)を求める)**
>
> - 書式 : VAR(数値1, 数値2, ...)
> - 引数 : 数値1, 数値2, ... :母集団に対応するセルの値、セルの範囲

標準偏差

- 標準偏差は、**STDEVP**関数を利用します。

> **STDEVP (引数を母集団全体であると見なして、母集団の標準偏差を求める)**
>
> - 書式 : STDEVP(数値1, 数値2, ...)
> - 引数 : 数値1, 数値2, ... :母集団に対応するセルの値、セルの範囲

不偏分散

- 不偏分散は、**VAR**関数を利用します。

> **VAR (引数を正規母集団の標本と見なし、標本に基づいて母集団の分散の推定値 (不偏分散) を求める)**
>
> - 書式 : VAR(数値1, 数値2, ...)
> - 引数 : 数値1, 数値2, ... :母集団の標本に対応するセルの値、セルの範囲

不偏標準偏差

- 不偏標準偏差は、**STDEV**関数を利用します。

> **STDEV (引数を標本と見なし、標本に基づいて母集団の標準偏差の推定値を求める)**
>
> - 書式 : STDEV(数値1, 数値2, ...)

○ 引数：数値1, 数値2, … ：母集団に対応するセルの値、セルの範囲