社会調査法　第５回

１．基本統計量

次の表に、このクラスの学生のデータを書いてみましょう。

身長は平均どのくらいだと思いますか？男性と女性でどのくらい変わるでしょう？

学生	①性別	②身長(cm)	③今アルバイトをやっているか（SA）	④アルバイトをしている場合、週の日数	⑤アルバイトをしている場合、アルバイトの種類（MA　複数回答）
1	1 女性 2 男性 3 その他	cm	1.いつもやる 2.時々やる 3.やらない	日	1　サービス業（飲食、販売など） 2　製造業（弁当製造、組み立てなど） 3　教育・福祉（家庭教師や老人ホーム） 4　その他
2	1 女性 2 男性 3 その他	cm	1.いつもやる 2.時々やる 3.やらない	日	1　サービス業（飲食、販売など） 2　製造業（弁当製造、組み立てなど） 3　教育・福祉（家庭教師や老人ホーム） 4　その他
3	1 女性 2 男性 3 その他	cm	1.いつもやる 2.時々やる 3.やらない	日	1　サービス業（飲食、販売など） 2　製造業（弁当製造、組み立てなど） 3　教育・福祉（家庭教師や老人ホーム） 4　その他
4	1 女性 2 男性 3 その他	cm	1.いつもやる 2.時々やる 3.やらない	日	1　サービス業（飲食、販売など） 2　製造業（弁当製造、組み立てなど） 3　教育・福祉（家庭教師や老人ホーム） 4　その他
5	1 女性 2 男性 3 その他	cm	1.いつもやる 2.時々やる 3.やらない	日	1　サービス業（飲食、販売など） 2　製造業（弁当製造、組み立てなど） 3　教育・福祉（家庭教師や老人ホーム） 4　その他
6	1 女性 2 男性 3 その他	cm	1.いつもやる 2.時々やる 3.やらない	日	1　サービス業（飲食、販売など） 2　製造業（弁当製造、組み立てなど） 3　教育・福祉（家庭教師や老人ホーム） 4　その他

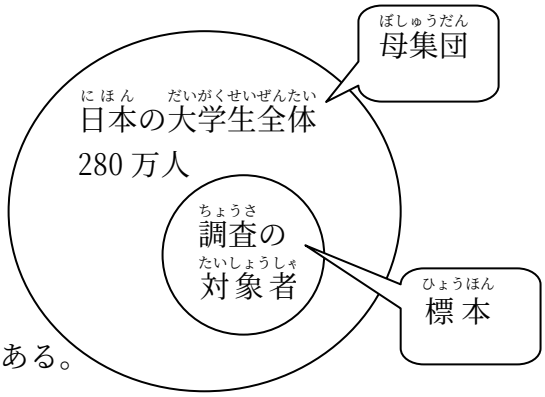
２．重要な用語

（１）標本誤差

$$\text{標本誤差} = 1.96\sqrt{\frac{N-n}{N-1}\times\frac{P(100-P)}{n}}$$

※母集団が大きいときは右でもいい（あまり変わらないから）→ $\text{標本誤差} = 2\sqrt{\frac{P(100-P)}{n}}$

n＝調査の対象者数（標本数）
P＝回答率
N＝母集団の人数



※ σ（シグマ　母集団がはっきりしている場合）を
標準偏差と、SE または S（標準誤差）という場合もある。

例：もし無作為に選んだ 80 人の大学生を対象者に、「アルバイトをしているか」聞いて、アルバイトをしている人が 60 人だった場合（75%の時）標本誤差（σ）は

$\sigma = 2\sqrt{\frac{P(100-P)}{n}}$　つまり　$\sigma = 2\sqrt{\frac{75(100-75)}{80}}$ ＝つまり 5 です。

これはアルバイトをしている人の割合が、母集団（日本の大学生全体280万人）70%から 80%の間（75%±5%）ということです。

±5%では、標本誤差が大きすぎるので、実際にはもっとたくさんの人を調査対象者にした方がいいでしょう。

また、正確な結果を得るには、対象者の全員に回答してもらうことが必要です。しかし現実の調査では、全員回答してもらうのは、非常に難しく、行政が行う調査でも回答率は 50%位にしかならないのが現状です。

（２）平均

身長や金額などの数値データの時のみ、集計できます。

例：もしこのクラスの学生の身長が

156cm	182cm	161cm	172cm	161cm	159cm	173cm	169cm

のとき、

（156cm＋182cm＋161cm＋172cm＋161cm＋159cm＋173cm＋169cm）÷８人で　平均は 166.6cm

（3）分散

データのばらつきを表します。（個々のデータ－平均）の2乗の総和（全部足した数）
分散が大きければ、ばらつきが大きい。（元のデータの絶対値が大きければ大きくなる）

$$\sigma^2 = \frac{\sum_{i=1}^{n}(Xi - \overline{X})^2}{n}$$

Σは「総和」と読む

$\overline{X}$は「エックスバー」または「平均」と読む

例	1	2	3	4	5	6	7	8
身長	156cm	182cm	161cm	172cm	161cm	159cm	173cm	169cm
A身長－平均(166cm	-10cm	16cm	-5cm	6cm	-5cm	-7cm	7cm	3cm
B　A×A　二乗	100cm	256cm	25cm	36cm	25cm	49cm	49cm	9cm
Bの総和	549cm							
B÷8（人数）	69	←これが分散						

（4）標準偏差（SD　standard deviation）

偏差（個々のデータ－平均値）の２乗の和（足し算したもの）を、データの数で割り、平方根（ルートのこと）をとる。
データのばらつき具合がわかる。（分散と違って、元のデータの単位となっている　たとえば身長のばらつきが上下8.3cm となる。分散では 69）

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(Xi - \overline{X})^2}{n}}$$

なぜ、わざわざ二乗して、ルートをとるのか？
→平均より大きい＋のデータ、平均より小さいマイナスのデータの、符号を無視するため

例　$\sqrt{69} = 8.3$　　だいたい 8.3cm のばらつきがあるということ

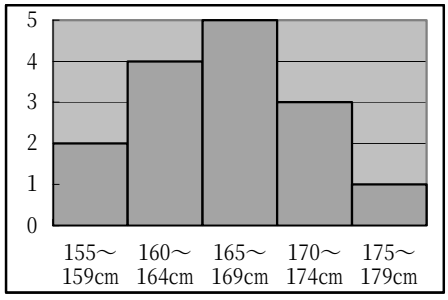
	1	2	3	4	5	6	7	8	計
身長	156cm	182cm	161cm	172cm	161cm	159cm	173cm	169cm	
A 身長－平均(166cm)	-10cm	16cm	-5cm	6cm	-5cm	-7cm	7cm	3cm	
B　A×A(二乗)	100cm	256cm	25cm	36cm	25cm	49cm	49cm	9cm	
C　Bの総和									549cm
D　分散　C÷8人									69cm
E　標準偏差　√D									8.3cm

（5）最大値　最小値　中央値　最頻値

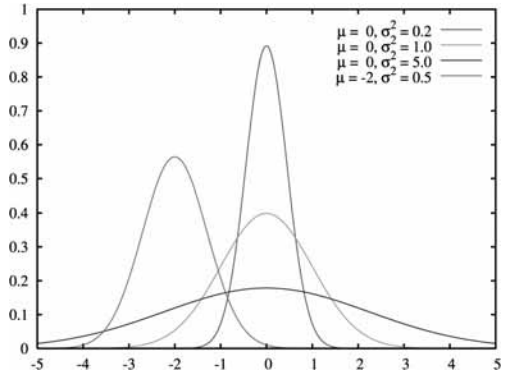
1	2	3	4	5	6	7	8
156cm	159cm	161cm	161cm	169cm	172cm	173cm	182cm
最小値		中央値	中央値				最大値
		最頻値	最頻値				

（6）度数分布表（ヒストグラム）

身長	人数
155～159cm	2
160～164cm	4
165～169cm	5
170～174cm	3
175～179cm	1



（5）正規分布



Wikipedia より

（7）尺度

①比率尺度　年齢、収入など

＊間隔尺度との違いは、絶対的規準「0」を持つこと

②間隔尺度　指数、温度など　*計算できる　0がない

例　不快指数（＝0.81×気温＋0.01×湿度×（0.99×気温－14.3）＋46.3）

～55	55～60	60～65	65～70	70～75	75～80	80～85	85～
寒い	肌寒い	何も感じない	快い	暑くない	やや暑い	暑くて汗が出る	暑くてたまらない

③順序尺度

とてもよい	よい	ふつう	わるい	とてもわるい

など　＊計算できない

アルバイトをしているか
1 いつもやる　2 時々やる　3 やらない

※ただし、わかりやすさや比較のために、次のように点数化して計算する場合もある

とてもよい	よい	ふつう	わるい	とてもわるい
１０点	５点	３点	２点	０点

④名義尺度

性別など　＊計算できない

①性別
1　男性　　2　女性

名義尺度で、男性を１、女性を２などと番号を振って計算する場合「ダミー（模擬）変数」と呼ぶことがある。

（８）変数の種類

①離散変数と連続変数

例：

性別が離散変数　　男性、女性、その他　しかない

身長が連続変数　　151、151.1　151.2 など小数点以下があり連続している

②独立変数と従属変数

例：

学年によって、身長の高さが違う場合

学年が（身長などを決める）独立変数

身長が（学年によって決定する）従属変数

③説明変数と被説明変数

例：

性別で身長の平均が違うとき

性別が（身長を説明する）説明変数

身長が（性別によって説明される）被説明変数

	身長の平均
女性	161cm
男性	172cm

３．相関関係とは

（１）相関とは

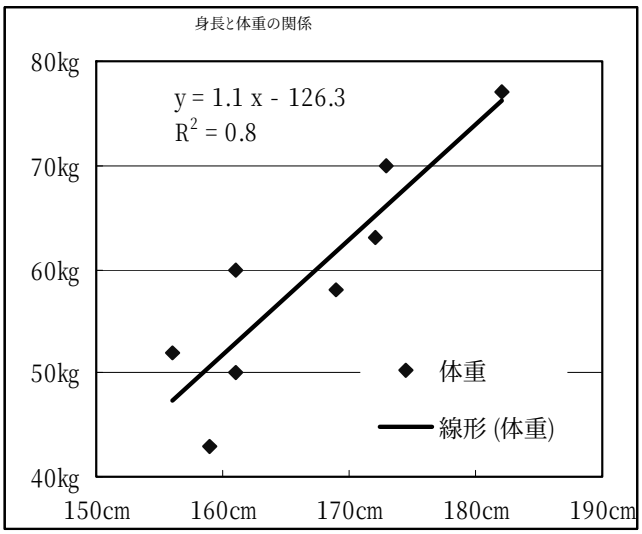
ある変数とある変数の間に、相関関係があること。

例　身長と体重

	1	2	3	4	5	6	7	8
身長	156cm	159cm	161cm	161cm	169cm	172cm	173cm	182cm
体重	52kg	43kg	60kg	50kg	58kg	63kg	70kg	77kg

相関係数	0.89

←＋の相関（高ければ重い）　0.6以上は相関がある



（２）疑似相関

相関があるように見えるが、実は他の要因と関係していること。

例：

誤り　　ジュースがたくさん売れると、水におぼれる人が増える？

正しくは　気温が上がると、ジュースが売れる、
水遊びをする人も増え、おぼれる人が増えるる

誤り？　朝ご飯を食べると、成績がよくなる？

本当は？　家庭環境で規則正しい生活をできる人は、朝ご飯を食べる人が多く、
成績もいいのかもしれない。