

基礎統計（火 1） 廣松 Shike-pri

1. 統計の取り方

1-1. 全数調査

母集団全体を調査。例・・・ 国勢調査、卒業文集のクラスで〇〇な人アンケート調査

1-2. サンプル調査

母集団の一部のみを調査。ただし、無作為に（意図的なものを排除し）サンプルを選ぶ。

トリビアの種で2000人調べればいいのかいってあるあれです。

※ しかし、意図せずして偏りが生じてしまうことがある。

例・・・インターネットによる調査（ネット環境が無い人の意見が自動的に排除されてしまう）

この結果、「統計でウソをつく」ことが有り得るんだなあー。むーん。

2. データの種類

2-1. 時系列データ

ある対象についての異なった時点におけるデータ(暦年、年次、四半期（3ヶ月）など)

2-2. 横断面 (cross-section) データ

ある属性に関して、いくつかの異なる対象についてのデータ

例・・・各国の人口（人口という属性に対するいくつかの異なる国についてのデータ）

2-3. 一次統計と二次統計

データそのものと、データが加工されたもの（物価指数とか）

3. (累積)相対度数分布

3-1. 度数とは何ぞや？

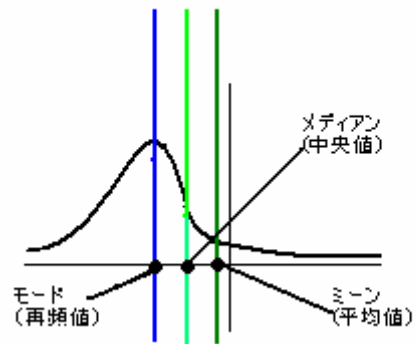
度数 (frequency) とは、ある範囲に入る観測値の個数。

f_j などとあらわす。此のとき

$\frac{f_j}{n}$ (n : 観測地の個数) のことを相対度数という。

$$\sum_{j=1}^k \frac{f_j}{n} = 1$$
 という関係があるんでやんす。

3-2. 相対度数分布（ヒストグラム）



左図が相対度数分布である。

3-3. 用語の説明

モード 再頻度。分布の峰に対応する値。

メディアン 中央値（または中位数）

ミーン 平均値

例・・・集団（1， 1， 1， 1， 2， 3， 5， 6， 7）があるとき、

モードは1（4つある）

メディアンは2

ミーンは3 となります。

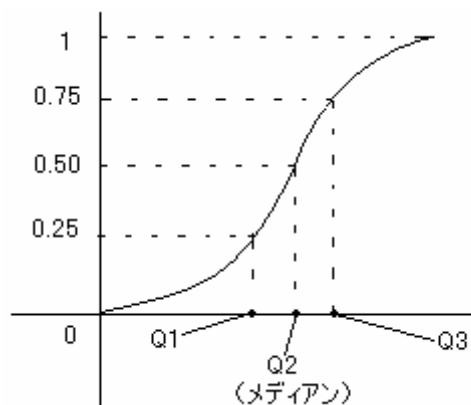
ちなみに、上の相対度数分布は「右に裾を引く分布」です。

このとき左からモード・メディアン・ミーンとなります。

左右対称のとき（正規分布のとき）は三つの値は一致します。

左に裾を引くときは左からミーン・メディアン・モードとなります。順序大事ナリ。

3-4. 累積相対度数分布



左図が累積相対度数分布である。

縦軸は相対度数の和を表す。わかって！

データの中で全体を4等分する点の値を4分位数と呼びます。

小さい順に、 Q_1 、 Q_2 、 Q_3 で Q_2 はメディアンに等しいんだニャー。

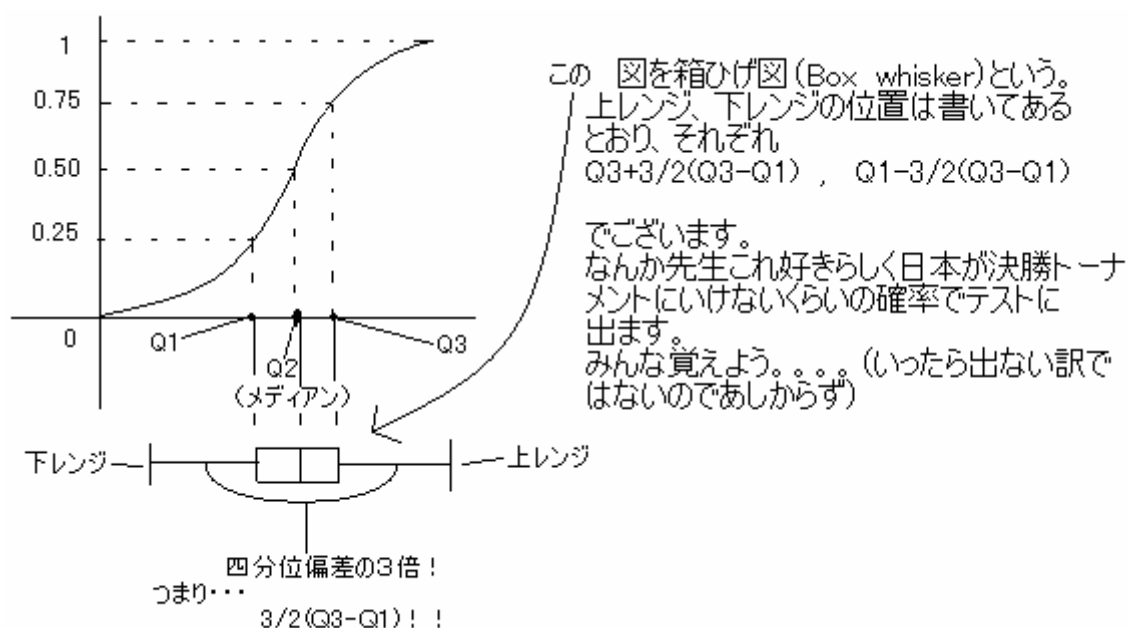
また、レンジを改良したものとして四分位偏差というものがあり、

$Q = 1/2(Q_3 - Q_1)$ として定義されます。

これは散らばりの範囲をあらわすものというくらいの解釈で・・・

3-5. 箱ヒゲ図、幹葉表示

☆箱ヒゲ図



☆幹葉表示

幹	葉	度数
0	1,3,4	3
1	0,	1
2	1,8	2
3	1	1

左図はある集団の「今までにサボった授業の
数」についての調査で、
(1,3,4,10,21,28,31)というデータが得られた時の
幹葉表示である。
リアルに31回サボった人もいるかもね・・・www

今6月18日、日曜日の午前2時です。早くこんな作業やめたいです。

でも逃げちゃだめだ、逃げちゃだめだ・・・

4・様々な(?) 尺度

4-1 位置の尺度

モード、メディアン、ミーンなどです。意味はもう説明しました。

平均値は $\bar{x}$ とかあらわしちゃいます。これを踏まえて

$$\sum_{i=1}^n (x_i - \bar{x}) = 0 \quad \text{という関係があります。}$$

4-2 ばらつきの尺度

分散っていう概念があります。定義はしたのとおりでチュウ。

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

左上のが分散です。

次に、不偏分散は↓です。不偏分散は自由度を考慮したものです。

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

また

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

としたときのSを標準偏差といいます。

4-3 その他もろもろ

他にも、ゆがみの尺度である歪度、とがり具合の尺度である尖度、積率などがあります。

こいつは分散では2乗であるところの乗数を3とか4とかKとかに変えていろいろしたりしたものです。

その意味については教科書を読んだりしてください。。。

5・2次元のデータ

今までは1次元のデータの話をしてきました。1次元のデータってのはまあ、1変数関数みたいな感じです。こっからは2次元、2変数関数みたいな感じだと思えば良いと思うよー。

具体的にはこんな感じで。

2次元データというのは

(x_1, y_1) , (x_2, y_2) , (x_3, y_3) , $\dots$, (x_n, y_n)

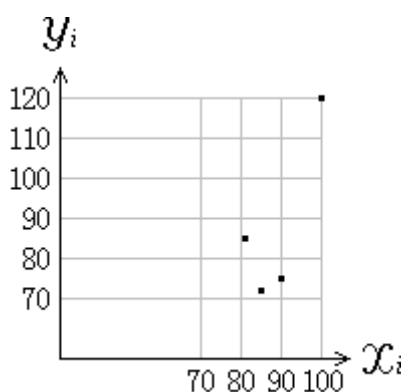
のように、1つの観測対象につき2つの観測値を持つデータです。

具体的には、下の表のようなデータです。

学生の名前	数学の点数	英語の点数
高橋	90 点	75 点
北村	81 点	85 点
野瀬	85 点	72 点
斉藤	100 点	120 点

この表を例とすると、 i を自然数として、 x_i は数学の点数、 y_i は英語の点数ということになります。

そして、 x_i 、 y_i をそれぞれ横軸、縦軸に取った、下のような散布図を作ることができます。

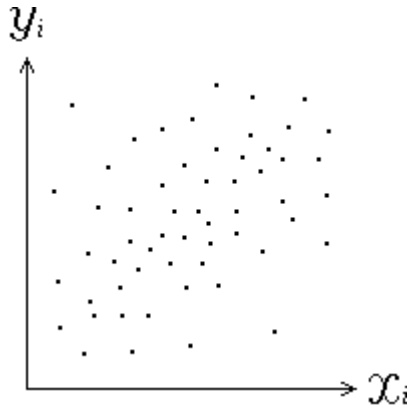


今は観測対象が4人しかいないので図中に点が4つしかありませんが、観測対象が多くなると x_i と y_i の間の相関関係が見られることがあります。

相関関係とは下のようなものです。

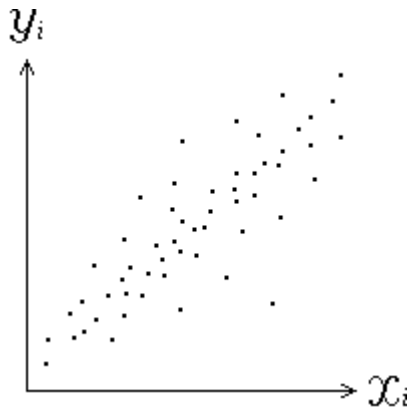
① 弱い正の相関関係

x_i が大きければ y_i も大きい傾向にあり、全体的にバラつきが大きい。



② 強い正の相関関係

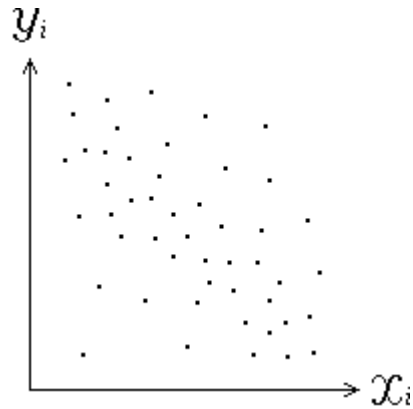
x_i が大きければ y_i も大きい傾向にあり、全体的にバラつきが小さい。（ほぼ一直線上に点が並ぶ。）



③ 負の相関関係

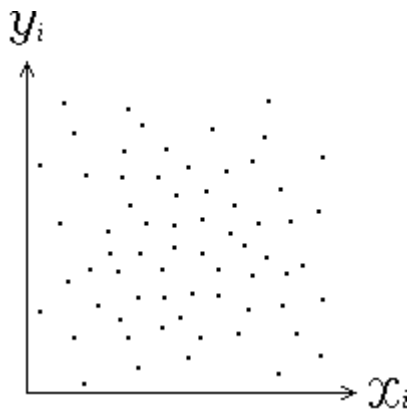
x_i が大きければ y_i は小さい傾向にある。

負の相関関係に関しては、あまり強弱の区別はしないみたいです。



④ 相関関係無し

x_i の大小と y_i の大小は無関係。



そこでこのような相関関係を調べる作業をしていきます。

1次元データの時には、データのバラつきを表す量として分散 S^2 を考えました。

2次元データでは、1次元の分散に相当する量として **共分散 C_{xy}** を考えます。

共分散 C_{xy} の定義式は分散 S^2 の定義式と似ていて、

$$C_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

分散は $(1/n) \sum (x_i - \bar{x})^2$ でしたから、2乗の代わりに $(y_i - \bar{y})$ が付いてると考えればいいのです。

ただ、それゆえ共分散は負の値を取ることがあります。

共分散が正のときは正の相関関係があり、負のときは負の相関関係があることも覚えておきましょう。

共分散で相関の正負を知ることはできますが、相関の強弱を知ることはできません。ですが、**相関係数**という値を求めれば、相関の強弱を知ることができます。

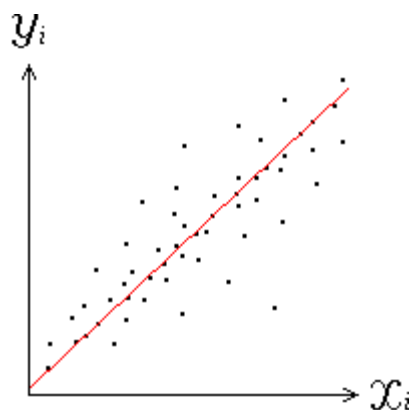
相関係数 r_{xy} の定義式は以下のとおりです。（ S_x 、 S_y は x 、 y の標準偏差のこと）

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

$C_{xy}/S_x S_y$ の値は必ず -1 以上 1 以下になりますので、 $C_{xy}/S_x S_y$ の絶対値の大小で相関の強弱を知ることができます。

6・回帰分析（だんだん難しい）

回帰直線について説明します。もしも散布図を下の図のような一本の直線で表すとしたらどのような直線が適当かを分析します。



まず、その直線を $y=a+bx$ と置きます。

b が直線の傾き、 a が定数項であることに注意してください。（中学、高校の教科書と逆です。）

散布図を直線で表そうとしても、散布図上の全ての点が $y_i=a+bx_i$ を満たす（直線 $y=a+bx$ 上に乗る）ことは普通は起こらないので、その誤差（**残差**といいます）を e_i とします。つまり、 $y_i=a+bx_i+e_i$ と置くわけです。（残差は d_i と表すこともある）

そして、残差が最小、すなわち、 $e_1 + e_2 + \dots + e_n$ が最小になるような a, b の値を求めると、(求め方は微分とか使っちゃいます。興味ある人は教科書で。でも難しい。)

$$\hat{b} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \hat{a} = \bar{y} - \hat{b} \bar{x}$$

($\hat{b}, \hat{a}$ は b ハット、 a ハットといいます。残差を最小にするような a, b のことです。)

回帰直線には次のような性質があります。

① 回帰直線は $(\bar{x}, \bar{y})$ を通る。

② $\hat{y}_i = \hat{a} + \hat{b}x_i \rightarrow$ 理論値という。 $y_i - \hat{y}_i = e_i \rightarrow$ 残差 (y_i は観測値)

③ 関係式色々。

$$\sum y_i = \sum \hat{y}_i \quad \bar{y} = \bar{\hat{y}} \quad \sum (y_i - \bar{y})^2 = \sum e_i^2 + \sum (\hat{y}_i - \bar{\hat{y}})^2$$

証明は略します。。。

7・確率

