【データの分析】

データを分析するための基本的な考え方を身につけよう.

☆データって何？

・データの種類（様々な解釈があるうちの一例）

データ	尺度	離散・連続	順序	数値的意味	代表値	可能な演算	具体例（要解釈）
質的データ	名義尺度	離散量	なし	区別	最頻値	なし	「評価（感想)」 「好きな食べ物」 「血液型」
	順序尺度	離散量	あり	大小順序	最頻値 中央値	なし	「評価（段階)」 「成績順位」
量的データ	間隔尺度	離散量 連続量	あり	間隔(和差) ※等間隔	最頻値 中央値 平均値	＋ －	離散量 「評価（星)」 連続量 「評価（基準とのずれ)」 「摂氏/華氏温度」
	比例尺度	連続量	あり	間隔(和差) 比率 ※等間隔 ※原点	最頻値 中央値 平均値	＋ －×÷	「評価（点数)」 「物理量」 「絶対温度」 「身長」「体重」

・離散量　中間の値がない.・連続量　中間の値が無数に存在する.
※間隔が等しいか，離散量か連続量かなどは仮定（そのデータの定義）による.
※点数は離散量で表されるが，本来は連続的な「価値」を点数という離散量のデータで便宜的に表しているという解釈において，連続量で扱える.

・表とグラフの種類

▶度数分布表　データグループ（階級）とデータの数（度数）の分布表
▶棒グラフ（質的データ　離散量　順序ありなし）　棒の間隔をあける.
▶棒グラフ（量的データ　離散量・連続量　順序あり）＝ヒストグラム　棒の間隔をあけない.
▶折れ線グラフ　棒グラフのてっぺんの真ん中を点にして折れ線で結ぶ.「変化」を示す.
▶円グラフと帯グラフ　割合を示す.
・円グラフ（扇形の角度）　単発データの「割合」を印象付ける.
・帯グラフ（帯の長さ）　並べて複数のデータの「推移」を比較する.

☆度数分布表（Frequency Table）

・階級（Class / Bin）

データのグループ

・階級値（Class Value）

各階級の中央の値

・度数（Freqency）

各階級のデータの個数

・相対度数（Relative Frequency）

各階級の度数を総数でわった値（全体に対する階級の割合）

・累積相対度数（Cumulative Relative Frequency）

その階級以下の相対度数の合計値（全体に対する位置付けの目安）

※階級の幅……階級の幅次第で，ヒストグラムの形状が変わる．（参考知識）

データの大きさ n のとき，幅 $h = \left\lceil \dfrac{\text{最大値} - \text{最小値}}{k} \right\rceil$

・平方根選択　$k = \sqrt{n}$

他にも，スタージェスの公式$k = [log_2 n + 1]$や，スコットの選択$\left(\text{標準偏差から決める}\right)$，

フリードマン＝ダイアニコスの選択（四分位範囲から決める）などがある．

・度数分布表とヒストグラムと度数分布多角形（度数分布折れ線）

☆代表値（平均値・中央値・最頻値）

代表値：データの傾向や特徴を示す数値

・平均値（Mean）

平らに均した（ならした）値
ここでは，足して割る平均（相加平均 Arithmetic Mean）を指す.

・中央値（Median）

データの度数の真ん中の値
データの大きさが奇数の時は真ん中のデータの値
偶数の時は真ん中二つのデータの平均値

・最頻値（Mode）

最も度数が多いデータの値
度数分布表から読み取るなら，最も度数の高い階級の階級値.

※平均値と中央値
分布が左に偏ると中央値も左寄りになり（中央値＜平均値　　）
分布が右に偏ると中央値も右寄りになる（　　平均値＜中央値）

例）サラリーマン 5 人の年収を比較しよう.
全員の年収を足して 5 で割ったものが平均値．年収順に並んだ真ん中の人の年収が中央値.
どちらを代表値とするかはデータの分布に依存する.
(i)

A さん	B さん	C さん	D さん	E さん
200 万	250 万	275 万	315 万	325 万

この場合，平均値 273 万，中央値 275 万でそれほど変わらない.

(ii)

A さん	B さん	C さん	D さん	E さん
200 万	250 万	275 万	315 万	1500 万

この場合，平均値 508 万，中央値 275 万となりかなり値に差がある.

(i)(ii)の違いは，外れ値の有無．(ii)は E さん一人が突出した値を持っていて残り 4 人の傾向から外れている。平均値はその影響を大きく受けている.

何が知りたいのかという目的により、使い分ける必要がある.
・年収の半分を徴収する　・年収の半額の車を購入させる

例）プロ野球チーム A と B の年棒（登録選手数は同数とする）
A の平均年棒は 5000 万　年棒の中央値は 1500 万
B の平均年棒は 3500 万　年棒の中央値は 2500 万
ここから何がわかるか。
A の方が平均値が高いのでトータルで選手にたくさん年棒を支払っていることがわかる。
B の方が中央値が高いので高い年棒をもらっている選手が多いことがわかる。
つまり、A は実績のある一部の選手が超高額の年棒をもらっており、B は一定の頑張りをそこそこ評価してもらって多くの選手がそこそこの年棒をもらっている。
自分が入団するなら、平均年棒の高い A か年棒の中央値の高い B か、どちらが良い？

☆散らばりの指標（中央値編）

最大値　最も大きいデータの値（度数ではなく値）
最小値　最も小さいデータの値
範囲　　最大値と最小値の差（データはその範囲内に収まっている）

・四分位数（Quartile）

データを度数（個数）に従って 4 等分したときの値を四分位数と言う．小さいほうから順に

Q_1：第一四分位数　Q_2：第二四分位数　Q_3：第三四分位数

と呼ぶ．なお、第二四分位数は中央値と同じことである．
※Q_1, Q_3の求め方はいくつかあるが，教科書通りにヒンジの定義を採用した．エクセルなどの表計算ソフトにおける関数とは求め方が違う．
ヒンジ：中央値の下側半分と上側半分に対してのそれぞれの中央値をQ_1, Q_3とする．

・四分位数を求めるときデータの大きさ n で手順が変わる（中央値のときと同様）
$n \equiv 0(mod.4)$

$n \equiv 1(mod.4)$

$n \equiv 2(mod.4)$

$n \equiv 3(mod.4)$

※四分位数以外にも名前の付いた分位数は多数存在する．
二分位（median）＝中央値，三分位（tertile）四分位（quartile）五分位（quintile）六分位（sextile）七分位（septile）八分位（octile）十分位（decile）十二分位（duo decile）十六分位（hexa decile）二十分位（vigintile）百分位（percentile）千分位（permille）

・箱ひげ図（Box Plot / Box and Whisker Plot）

・5 数要約（Five-Number Summary）
これまでに出そろったデータは 5 つ.
最小値
第一四分位数 Q_1（下側ヒンジ）
第二四分位数 Q_2（＝中央値）
第三四分位数 Q_3（上側ヒンジ）
最大値
この 5 つの値のことを 5 数要約と言う.

5 数要約をわかりやすく表すための図が箱ひげ図.

平均値の値を＋で書き足すこともある.

・何がどうわかりやすいのか
データの大きさの 0%,25%,50%,75%,100%の値を記すことで, 代表値の示す特徴にとどまらず, おおまかな分布がわかる. 各区間ごとに同じ度数のデータが収まっている.
・四分位範囲（IQR / interquartile range）
$Q_3 - Q_1$　中央値を挟んだ半数のデータが収まる（散らばる）範囲
・四分位偏差
$Q_3 - Q_1$　IQR の半分　中央値付近の四分の一のデータが収まる（散らばる）範囲

・度数分布との対応

☆散らばりの指標（平均値編）

中央値周りの散らばりの指標　四分位偏差
平均値周りの散らばりの指標　分散や標準偏差

・分散（variance）：$s_x{}^2$

各データの値－平均値＝偏差（deviation）
偏差には正負の符号がつく.
偏差の平均を取ると 0（正と負の値が相殺して平均値に落ち着く）
「散らばり」＝「平均値からどれくらい離れているか」
　　＝「平均値との差の絶対値」※
　　＝「平均値との差を二乗しておけば必ず正」
平均値との差の平方を偏差平方（squared deviation）と言う.
・分散の求め方
平均値を求める.
平均値との差（偏差）を求める.
偏差平方を求める.
偏差平方の平均値を求める.　→　これが分散.

※「差の絶対値の平均」を指標として用いない理由

1)そもそも、絶対値は場合分けが面倒だったり微分不可能な点があったりして扱いにくいので,
　絶対値の平均はほとんど使わない. →嘘（とまでは言わないが, 分野によってはよく使われる）

2)二乗で平均を取ると、場合分けも要らないし取り扱いが楽. →そういう面もある.

・最適化（最小化）
母集団の各要素からの絶対距離の和を最適化するものは「中央値」である.
母集団の各要素からの距離の二乗和を最適化するものは「平均値」である.

よって平均値周りの散らばりの指標として分散
　　中央値周りの散らばりの指標として中央絶対偏差（MAD/median absolute deviation）
が用いられる.

理由としては，次のように説明をつけて納得しておこう.

「

」

分散＝偏差平方の平均値

例題）5 人にアンケートを行い，10 問の問いに対する○の数を調べた．○の数の分散を求めよ．

○の数	3	5	4	7	3

標本番号	データ（○の数）	偏差	偏差平方
1	3		
2	5		
3	4		
4	7		
5	3		
	平均値		**分散**

・標準偏差（standard deviation）：S_x

分散は元のデータの平方を基準にするため，元のデータと単位が揃わない．
そこをそろえる工夫をしたものが標準偏差である．

標準偏差＝$\sqrt{\text{分散}}$

（定義式まとめ）

偏差	$x_1-\bar{x}, x_2-\bar{x}, x_3-\bar{x}, \cdots, x_n-\bar{x}$
偏差平方	$(x_1-\bar{x})^2, (x_2-\bar{x})^2, (x_3-\bar{x})^2, \cdots, (x_n-\bar{x})^2$
分散	$s_x{}^2 = \dfrac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$
標準偏差	$s_x = \sqrt{\dfrac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}}$

・分散と平均値

分散＝2 乗の平均値－平均値の 2 乗

☆偏差値ってナニ？（番外編）

$$\text{偏差値（}Standard\ Score\text{）} = \frac{\text{得点} - \text{平均点}}{\text{標準偏差}} \times 10 + 50$$

☆変量の変換

・仮平均（Assumed Mean / Working Mean）

都合の良いようにデータを加工してから平均値や分散の計算をすることができる。
・計算が楽になるように
・データが見やすくなるように

$x = cu + x_0$ とおいて $u = \dfrac{x - x_0}{c}$ とし，変量をxからuに変換する．

このときxとuの平均値$\bar{x}$と$\bar{u}$および標準偏差s_xとs_uについて，以下の関係が成り立つ．

$\bar{x} = c\bar{u} + x_0$ となり， ${s_x}^2 = c^2{s_u}^2 (s_x = |c|s_u)$ つまり，分散は$\dfrac{1}{c^2}$倍，標準偏差は$\dfrac{1}{|c|}$となる．

(∵)データの大きさを n，変量 x を x_1，x_2，…，x_k，…，x_n とする．

例）次の変量 x のデータについて，仮平均の考え方を用いて，平均値と分散を求めよ．$[c = 40, x_0 = 1500]$

x	度数
1360 以上 1400 未満	1
1400 以上 1440 未満	2
1440 以上 1480 未満	3
1480 以上 1520 未満	2
1520 以上 1560 未満	3
1560 以上 1600 未満	4
1600 以上 1640 未満	3
1640 以上 1680 未満	2

☆2 変量データの分析

1 変量データについては,

代表値・5 数要約・分散（標準偏差）

を調べることを主な分析手法として学んだ.

変量が 2 つに増えると,

「各変量間に相互に何らかの相関性がないか」という新しい視点が発生する。

それを調べるために使うのが、散布図というグラフと相関係数である。

☆散布図（相関図 Scatter Plot / Scatter Diagram）と相関表（Correlation Table）

散布図）x：数学の点数 / y：理科の点数　　相関表）x：数学の点数 / y：理科の点数

・相関関係（Correlation）

正の相関	負の相関	相関なし
傾き正の直線上	傾き負の直線上	直線上に集まらない

・相関関係の強弱　直線上に乗っているほど相関関係は強い

強い負の相関関係　弱い負の相関関係　相関関係なし　弱い正の相関関係　強い正の相関関係

☆相関関係の指標

・共分散（Covariance）：s_{xy}

2 変量の平均からのずれの積（偏差積）を考えると、ずれの方向が同じとき正に、逆のとき負になる。この偏差積の平均値を共分散と言う。

・相関係数（Correlation Coefficient）：r

2 変量について、ずれの方向まで考慮した相関性を示す指標が相関係数である。具体的な計算を通して理解しよう。

「表を使って整理」

	変量	変量	偏差	偏差	偏差平方	偏差平方	偏差積
番号	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	①	①	③	③	④	④	⑦
2	①	①	③	③	④	④	⑦
…	…	…	…	…	…	…	…
n	①	①	③	③	④	④	⑦
計	①	①	0	0	④	④	⑦
平均	②	②	⑥	⑥	⑤	⑤	⑧

① xとyのデータを埋めて，合計を計算する．

② データの合計からxとyの平均値を求める．

③ xとyの各データからそれぞれの平均値を引いた値（偏差）を求める．合計は 0.

④ ③（偏差）の各データの二乗（偏差平方）とその合計を計算する．

⑤ ④（偏差平方）の合計からその平均値を求める．これが分散.

⑥ ⑤（分散）の平方根を計算する．これが標準偏差.

⑦ ③（偏差）の$x-\bar{x}$と$y-\bar{y}$の積（偏差積）とその合計を計算する．

⑧ ⑦（偏差積）の合計からその平均値を求める．これが共分散.

「相関係数 r を求める」

$$r=\frac{s_{xy}}{s_x s_y}=\frac{\text{⑧共分散}}{\text{⑥標準偏差}\times\text{⑥標準偏差}}=\frac{(x-\bar{x})(y-\bar{y})\text{の平均値}}{\sqrt{(x-\bar{x})^2\text{の平均値}}\times\sqrt{(y-\bar{y})^2\text{の平均値}}}$$

「相関係数と散布図の関係」

完全な負の相関関係	強い負の相関関係	弱い負の相関関係	相関関係なし	弱い正の相関関係	強い正の相関関係	完全な正の相関関係
$r=-1$	$r=-0.8$	$r=-0.4$	$r=0$	$r=0.4$	$r=0.8$	$r=1$

・共分散と平均値

共分散＝積の平均値－平均値の積

※相関関係と因果関係は違う.

例えば，朝食を食べることと学力との関係を調査して，正の相関関係があったとしても，朝食を食べれば学力が上がることにはならない．もっと他にも，食事の品目数や睡眠時間，テレビの視聴時間など複数の項目にわたって学力との相関関係を調べれば，生活習慣との相関関係が一般化され，因果関係に近いものが見えてくるかもしれない．勉強すれば学力が上がる．これが因果関係である（本当はこれすらもあやしいかもしれない）.

例題）電卓を使っても良い

ある模擬試験の点数が以下の通りであったとする.

生徒名	A	B	C	D	E	F	G	H
理科の点数x	45	62	78	90	84	65	53	67
数学の点数y	51	61	66	100	76	58	39	77

このうち, 理科と数学の点数の間に相関関係があるか調べたい. 答は小数第二位を四捨五入して答えよ. 必要なら以下の表を用いても良い。

(1)理科と数学それぞれの平均点を求めよ.

(2)理科と数学それぞれの分散を求めよ.

(3)理科と数学それぞれの標準偏差を求めよ.

(4)理科と数学の共分散を求めよ.

(5)理科と数学の相関係数を求めよ.

生徒	x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
A							
B							
C							
D							
E							
F							
G							
H							
計							
平均							

☆データの分析　総まとめ（イメージ重視）

・1 変量（度数分布表→ヒストグラム→代表値→四分位数→箱ひげ図→分散→標準偏差）

・2 変量（散布図・相関表→散布図における相関関係→共分散→相関係数）

・変量の変換

変量Xの平均値を$E(X)$，分散を$V(X)$，標準偏差を$\sigma(X)$，変量XとYの共分散を$Cov(X,Y)$とする．

$$\begin{cases} E(aX+b)=aE(X)+b \\ V(aX+b)=a^2V(X) \\ \sigma(aX+b)=|a|\,\sigma(X) \\ Cov(aX+b,cY+d)=acCov(X,Y) \end{cases}$$

$x=c_xu+x_0$　とおいて　$u=\dfrac{x-x_0}{c_x}$　とし，変量をxからuに変換する．

$y=c_yv+y_0$　とおいて　$v=\dfrac{y-y_0}{c_y}$　とし，変量をyからvに変換する．

※散らばりの指標は x_0 や y_0 の影響を受けない（どこを基準にしても散らばり具合は同じ）

標準偏差　平均からのずれの大きさ　→　$\dfrac{1}{|c_x|},\dfrac{1}{|c_y|}$ 倍

分散　　偏差平方の平均　　→　$\dfrac{1}{{c_x}^2},\dfrac{1}{{c_y}^2}$ 倍

共分散　　偏差積の平均　　→　$\dfrac{1}{c_x}\cdot\dfrac{1}{c_y}$ 倍

相関係数　$\dfrac{\textit{共分散}}{\textit{標準偏差}\times\text{標準偏差}}$　　→　$\dfrac{\frac{1}{c_x}\cdot\frac{1}{c_y}}{\frac{1}{|c_x|}\cdot\frac{1}{|c_y|}}=\pm1$ 倍