数学 Ⅰ　データの分析

データと変量／ 階級値と度数分布表／ 累積度数／ ヒストグラム／
相対度数／ 最頻値／ 平均値／ 中央値／ 代表値／ 四分位数／
四分位数を求める手順／ 範囲／ 四分位範囲と四分位偏差／ 箱ひげ図／
外れ値／ 外れ値と箱ひげ図／ 偏差／ 分散と標準偏差／
☆分散と平均値の関係式／ 散布図／ 相関関係／ 相関関係と散布図／
共分散／ ☆共分散と平均値／ 共分散と相関関係／ 相関係数／
★相関係数の計算式／ ☆共分散と平均値／ 相関係数と散布図／
☆平均値の変量の関係／ ☆分散と標準偏差の変量の関係／
☆共分散の変量の関係／ ☆相関係数の変量の関係／
★直線上に分布する相関係数／ 仮説検定／ 仮説検定の考え方／

□データと変量

ある集団を構成する人や物の特性を量的に表すものを**変量**（へんりょう）といい，

調査や実験などで得られた観測値や測定値の変量の集まりを**データ**という．

データを構成する変量の個数をそのデータの**大きさ**という．

例 人の身長や体重，テストの点数，勉強時間，株価などを変量といい，
そのような変量を集めたものをデータという．
その変量の個数をデータの大きさという．

例 ある相撲部屋の力士 20 人の体重を測定した結果について

ある相撲部屋の力士 20 人（単位 kg）

145	102	167	133	99	199	173	163	121	143
123	137	151	115	156	160	182	85	219	158

それぞれの体重を変量といい，このような資料をデータという．
データの大きさは 20 人分だから **20**

□階級値と度数分布表

データを整理するために用いる区間を **階級**（かいきゅう） といい，

その区間の幅を **階級の幅**（かいきゅう） という．

とくに 階級の真ん中の値を **階級値**（かいきゅうち） という．

それぞれの階級に入っているデータの値の個数をその階級の **度数**（どすう） という．

各階級に度数を対応させたものを **度数分布**（どすうぶんぷ） という．

度数分布を表にしたものを **度数分布表**（どすうぶんぷひょう） という．

例 次のようなある相撲部屋の力士 20 人の体重を測定した結果について

ある相撲部屋の力士 20 人（単位 kg）

145	102	167	133	99	199	173	163	121	143
123	137	151	115	156	161	182	85	219	158

このデータを度数分布表にしたものが右である．
左列の体重を整理するための区間を階級という．
区間の幅はどれも 20kg
階級「100kg 以上 120kg 未満」の階級値は 110kg
階級に入っている人数を右列の度数で表す．
階級「100kg 以上 120kg 未満」に 102kg と 115kg の
2 人が入るので度数 2 が対応している．

階級（kg）	度数
80 以上～100 未満	2
100 ～120	2
120 ～140	4
140 ～160	5
160 ～180	4
180 ～200	2
200 ～220	1
合計	20

□累積度数

度数分布表において，最初の階級からその階級までの度数を合計 したものを
累積度数（るいせきどすう） という．

例 累積度数をかきこんだ表は右のようになる．

階級（kg）	度数	累積度数
80 以上～100 未満	2	2
100 ～120	2	4
120 ～140	4	8
140 ～160	5	13
160 ～180	4	17
180 ～200	2	19
200 ～220	1	20
合計	20	

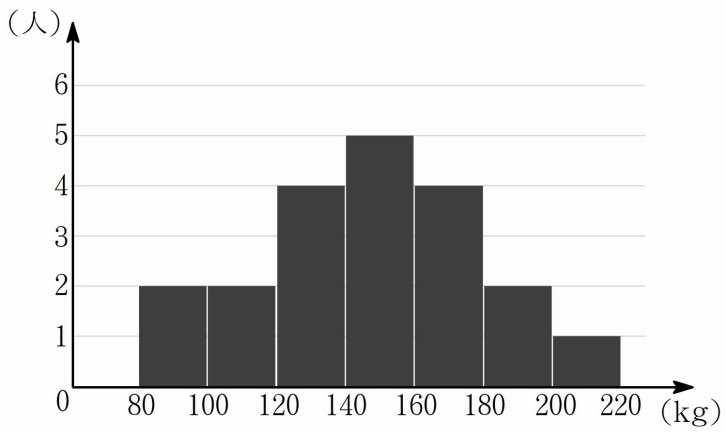
□ヒストグラム

度数分布を階級の幅を底辺，度数を高さとする長方形にしてすき間なく並べてグラフにした図を **ヒストグラム** という．

それぞれの長方形の面積は階級の度数に比例している．

例 左下の度数分布表をヒストグラムで表したものが右である．

階級（kg）	度数
80 以上～100 未満	2
100 ～120	2
120 ～140	4
140 ～160	5
160 ～180	4
180 ～200	2
200 ～220	1
計	20



□相対度数

各階級の度数を度数の合計で割った値を**相対度数**（そうたいどすう）という.

つまり　$(\text{相対度数}) = \dfrac{(\text{その階級の度数})}{(\text{度数の合計})}$

例 上の度数分布で相対度数を求めると
右表のようになる.
階級「100kg 以上 120kg 未満」の相対度数は
$\dfrac{2}{20} = 0.10$

階級（kg）	度数	相対度数
80 以上～100 未満	2	0.10
100　～120	2	0.10
120　～140	4	0.20
140　～160	5	0.25
160　～180	4	0.20
180　～200	2	0.10
200　～220	1	0.05
計	20	1.00

□最頻値

データにおいて

最も個数の多い値をそのデータの **最頻値**（さいひんち） または **モード** という．

データが度数分布表に整理されているときは度数が最も大きい階級の階級値を最頻値とする．

例 10 個のデータ

1, 1, 3, 3, 4, 5, 8, 8, 8, 9

について，8 が 3 個で最も個数が多いので 最頻値は **8**

□平均値

変量 x のデータの値の総和をデータの値の個数で割ったものを

データの平均値（へいきんち）といい $\overline{x}$ で表す.

すなわち 変量 x のデータを n 個の値 x_1, x_2, …, x_n とするとき

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

$$= \frac{1}{n}\sum_{k=1}^{n} x_k$$

つまり（平均値）$= \dfrac{\text{（データの値の総和）}}{\text{（データの値の個数）}}$

㊟ $\displaystyle\sum_{k=1}^{n} x_k = x_1 + x_2 + \cdots + x_n$（数学 B）

㊟ 10 個のデータ

1, 1, 3, 3, 4, 5, 8, 8, 8, 9

について，平均値は

$$\frac{1+1+3+3+4+5+8+8+8+9}{10} = \frac{50}{10} = \mathbf{5}$$

□中央値

データの値を小さい順に並べて中央の順位にくる値を

中央値（ちゅうおうち） または メジアン という．

1 データの値の個数が奇数のとき

中央の順位にくる値は1つ決まり，それが中央値となる．

2 データの値の個数が偶数のとき

中央の順位にくる値は2つになり，それらの平均値が中央値となる．

例 1 ある7個のデータを小さい順に並べ

7, 9, 15, **21**, 33, 44, 56

このとき，中央値は **21**

2 8個のデータを小さい順に並べ

7, 9, 15, **20**, **22**, 26, 27, 28

このとき，中央値は $\dfrac{20+22}{2}=\dfrac{42}{2}=\mathbf{21}$

代表値

データ全体の特徴を 1 つの数値で表わしたものを **代表値**（だいひょうち） という．

代表値には **平均値，中央値，最頻値** などがよく用いられる．

(補) データをみるときに、データの特徴がみえる値のこと．

(例) 例えば 500 人が受験した 100 点満のテストで，
平均点 30 点だとわかると，難しい問題のテストだとわかる．
中央値が 60 点だとわかると，60 点の人は順位は 250 位前後とわかる．
最頻値が 100 点だとわかると，満点でも単独 1 位ではないとわかる．

□四分位数

データの値を小さい順に並びかえて，4 等分される位置にくる 3 つの値を
四分位数（しぶんいすう）という．

四分位数は小さい値から順に第 1 四分位数，第 2 四分位数，第 3 四分位数といい，これらを順に Q_1，Q_2，Q_3 で表す．

□四分位数を求める手順

第1四分位数 Q_1，第2四分位数 Q_2，第3四分位数 Q_3 は次の手順で求めることができる．

1 データの値を小さい順に並びかえる

2 中央値を求め，その値が Q_2 である．

3 中央値を境にしてデータの個数が等しくなるように2つの部分に分ける．
ただし データの大きさが奇数のときは中央値を含めずに分けることにする．

4 3 で分けられた部分で最小値を含む方のデータ（下位のデータ）の中央値を求め，その値が Q_1 である．

5 3 で分けられた部分で最大値を含む方のデータ（上位のデータ）の中央値を求め，その値が Q_3 である．

1 データの値の個数が奇数のとき

$$\underbrace{\boxed{\bigcirc,\ \bigcirc,\ \cdots,\ \bigcirc}}_{\text{中央値が } Q_1},\ \bigcirc,\ \underbrace{\boxed{\bigcirc,\ \bigcirc,\ \cdots,\ \bigcirc}}_{\text{中央値が } Q_3}$$

2 データの値の個数が偶数のとき

$$\underbrace{\boxed{\bigcirc,\ \bigcirc,\ \cdots,\ \bigcirc,\ \bigcirc}}_{\text{中央値が } Q_1},\ \underbrace{\boxed{\bigcirc,\ \bigcirc,\ \bigcirc,\ \cdots,\ \bigcirc}}_{\text{中央値が } Q_3}$$

例 1 データ 7, 2, 5, 3, 6, 1, 4 について
1 データを小さい順に並べて 1, 2, 3, 4, 5, 6, 7
2 中央値を求めて $\boldsymbol{Q_2 = 4}$
3 2つの部分に分けて 1, 2, 3, 4, 5, 6, 7
4 1, 2, 3 の中央値より $\boldsymbol{Q_1 = 2}$
5 5, 6, 7 の中央値より $\boldsymbol{Q_3 = 6}$

□範囲

データの最大値と最小値の差を範囲（はんい） または レンジ という.
つまり （範囲）＝（最大値）－（最小値）

□四分位範囲と四分位偏差

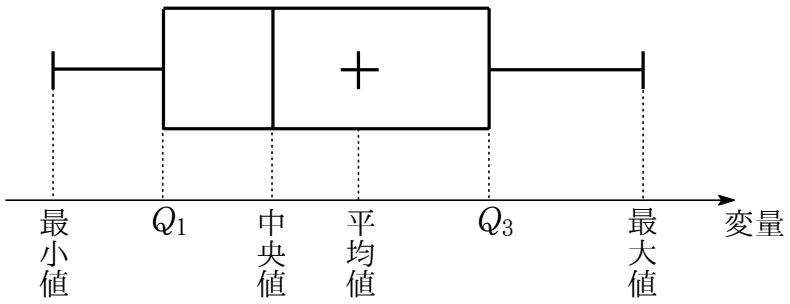
第1四分位数 Q_1，第3四分位数 Q_3 に対し

1 $Q_3 - Q_1$ を 四分位範囲（はんい） という.

2 $\dfrac{Q_3 - Q_1}{2}$ を 四分位偏差（へんさ） という.

□箱ひげ図

最小値，第1四分位数Q_1，中央値，第3四分位数Q_3，最大値，平均値の値を長方形（箱）と線（ひげ）を用いて1つの図にしたものを箱ひげ図（はこひげず）といい次のように表される．



(注) 平均値は省略することが多い．

外れ値

データの値の中に，極端に小さい値や大きい値が含まれるとき，
そのような値を **外れ値**（はずれち） という．

外れ値の基準はいろいろあるが，多くの場合は次のように判断する．

ここで，第1四分位数を Q_1，第3四分位数を Q_3 とする．

1 第1四分位数 $-1.5\times$(四分位範囲) 以下の値

つまり $Q_1-1.5\times(Q_3-Q_1)$ 以下の値は極端に小さいので外れ値とする．

2 第3四分位数 $+1.5\times$(四分位範囲) 以上の値

つまり $Q_3+1.5\times(Q_3-Q_1)$ 以上の値は極端に大きいので外れ値とする．

補 1 は 第1四分位数 $-3\times$(四分位偏差) 以下の値と表すこともできる．
2 は 第1四分位数 $+3\times$(四分位偏差) 以上の値と表すこともできる．

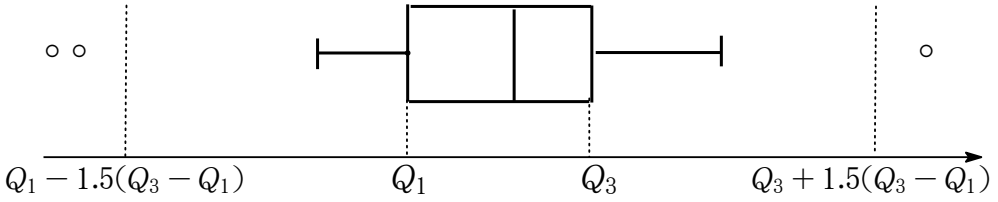
補 外れ値は規格外の正常な値ではなく，測定ミスや入力ミスによる異常な値の可能性もある．

補 外れ値が含まれている場合は，その原因を考えることが大事である．

外れ値と箱ひげ図

外れ値がある場合の箱ひげ図を下図のようにかくことある，

外れ値は○で表し，外れ値を除いた最大値と最小値でひげをかく．



（Q_1 は第 1 四分位数，Q_3 は第 3 四分位数）

㊟ 外れ値をのぞいたひげは短くなる．

偏差

変量 x についてのデータの値が n 個の値 x_1, x_2, …, x_n とする.

その変量 x の平均値を $\overline{x}$ とするとき，各値と平均値との差

$$x_1-\overline{x},\ x_2-\overline{x},\ \cdots,\ x_n-\overline{x}$$

をそれぞれ平均値からの偏差（へんさ），あるいは単に 偏差 という.

つまり

$$(\text{偏差})=(\text{変量})-(\text{平均値})=x-\overline{x}$$

補 偏差の平均値は 0 になる.

例 変量 x の 6 個のデータが 4, 5, 7, 8, 8, 10 であるとき

x の平均値は $\overline{x}=\dfrac{4+5+7+8+8+10}{6}=\dfrac{42}{6}=7$

x の偏差 $x-\overline{x}=x-7$ でありそれぞれの偏差は -3, -2, 0, 1, 1, 3

分散と標準偏差

1 偏差の 2 乗の平均値を **分散**（ぶんさん） といい s^2 と表す.

2 分散の正の平方根 を **標準偏差**（ひょうじゅんへんさ） といい s と表す.

すなわち 変量 x のデータを n 個の値 x_1, x_2, …, x_n , 平均値を $\overline{x}$ とするとき

1 $$s^2 = \frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}$$

$$= \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2$$

2 $$s = \sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}}$$

つまり （標準偏差）$=\sqrt{(分散)}$

分散や標準偏差は データの散らばりの度合を表す量 であり,

データの各値が平均値から離れるほど大きな値をとる.

話 標準偏差を英語で standard deviation という.

補 変量 x の測定単位を例えば cm とすると，分散 s^2 の単位は cm^2 となり単位が変わるが，標準偏差 s の単位は cm であり，単位が変わらない.

例 変量 x のデータが 4, 5, 7, 8, 8, 10 であるとき

平均値は $\overline{x} = \frac{4+5+7+8+8+10}{6} = \frac{42}{6} = 7$

1 分散は $s^2 = \frac{(4-7)^2+(5-7)^2+(7-7)^2+(8-7)^2+(8-7)^2+(10-7)^2}{6}$

$= \frac{9+4+0+1+1+9}{6} = \frac{24}{6} = \mathbf{4}$

2 標準偏差は $s = \sqrt{4} = \mathbf{2}$

☆分散と平均値の関係式

変量 x の分散 s^2 と平均値 $\overline{x}$，x^2 の平均値 $\overline{x^2}$ について

$$s^2 = \overline{x^2} - (\overline{x})^2$$

つまり　(x の分散) = (x^2の平均値) − (x の平均値)^2

すなわち　変量 x のデータを n 個の値 $x_1, x_2, \cdots, x_n$，平均値を $\overline{x}$ とするとき

$$s = \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} - \left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)^2$$
$$= \frac{1}{n}\sum_{k=1}^{n} {x_k}^2 - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2$$

㊟ $s^2 = \dfrac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}$

$$= \frac{({x_1}^2 + {x_2}^2 + \cdots + {x_n}^2) - 2\overline{x}(x_1 + x_2 + \cdots + x_n) + n(\overline{x})^2}{n}$$
$$= \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} - 2\overline{x} \cdot \underbrace{\frac{x_1 + x_2 + \cdots + x_n}{n}} + (\overline{x})^2$$
$$= \frac{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}{n} - 2(\overline{x})^2 + (\overline{x})^2 \quad \left(\because \frac{x_1 + x_2 + \cdots + x_n}{n} = \overline{x}\right)$$
$$= \overline{x^2} - (\overline{x})^2$$

㊟ 変量 x の 6 個のデータが 4, 5, 7, 8, 8, 10 であるとき

x の平均値は $\overline{x} = \dfrac{4+5+7+8+8+10}{6} = \dfrac{42}{6} = 7$

x^2 の平均値は $\overline{x^2} = \dfrac{4^2+5^2+7^2+8^2+8^2+10^2}{6} = \dfrac{16+25+49+64+64+100}{6}$

$$= \frac{318}{6} = 53$$

分散は $s^2 = \overline{x^2} - (\overline{x})^2 = 53 - 49 = \mathbf{4}$

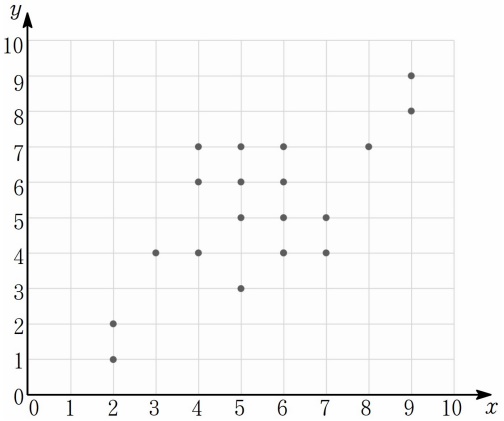
散布図

2 つの変量 x，y の値の組 (x, y) を座標とする点を平面上に取った図を
散布図（さんぷず） という．

㊟ 2 つの変量 x，y の分布が次の表のようなとき，右が散布図である．

	①	②	③	④	⑤	⑥	⑦	⑧	⑨	⑩
x	4	9	5	6	5	6	4	5	8	2
y	4	8	3	4	7	5	7	6	7	1

	⑪	⑫	⑬	⑭	⑮	⑯	⑰	⑱	⑲	⑳
x	3	9	5	6	4	6	7	6	2	7
y	4	9	5	7	6	4	5	6	2	4



相関関係

2つの変量のデータにおいて

1 一方が増えると他方が増える傾向と認められるとき

2つの変量の間に 正(せい)の相関関係(そうかんかんけい)がある または 正(せい)の相関(そうかん)があるという.

2 一方が増えると他方が減る傾向と認められるとき

2つの変量の間に 負(ふ)の相関関係(そうかんかんけい)がある または 負(ふ)の相関(そうかん)があるという.

3 どちらの傾向も認められないときは

2つの変量の間に 相関関係がない または 相関がないという.

例 1 身長と体重のデータは，身長が高いと体重が重たい傾向が認められるので正の相関がある.

2 北半球気温と南半球の気温のデータは，季節が逆なので，北半球で「春から夏」で気温が高くなると，南半球では「秋から冬」で気温は低くなる傾向が認められるので負の相関がある.

3 身長と気温のデータは，相関関係がない.

相関関係と散布図

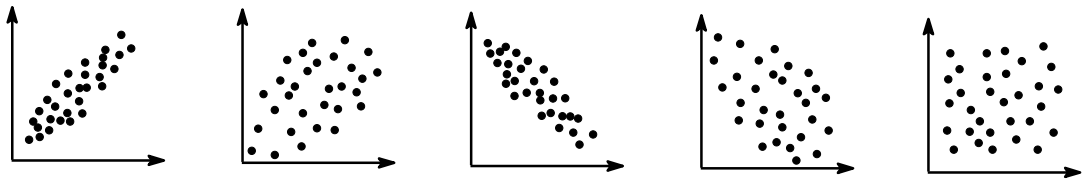
2つの変量のデータについて

1 正の相関関係があるとき，散布図は右上がりに分布する．

2 負の相関関係があるとき，散布図は右下がりに分布する．

下の5つの散布図 A，B，C，D，E において

〔図 A〕 〔図 B〕 〔図 C〕 〔図 D〕 〔図 E〕

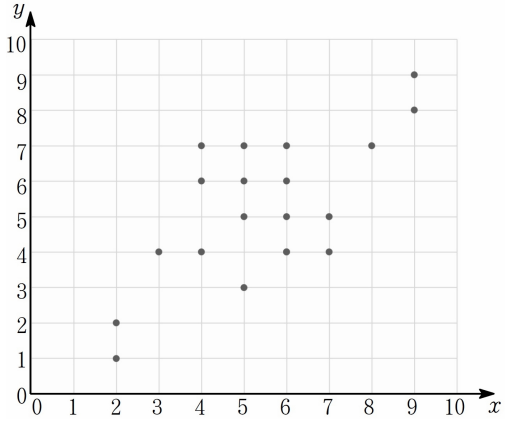


1 図 A と図 B は右上がりに分布しているので正の相関関係が認められるが
この傾向は図 A の方が著しいので
図 A にはより強い正の相関関係があるという．

2 図 C と図 D は右下がりに分布しているので負の相関関係が認められるが
この傾向は図 C の方が著しいので
図 C には より強い負の相関関係があるという．

3 図 E はどちらの傾向も認められないので 相関関係はないという．

例 右の散布図はやや強い正の相関がある．



共分散

2つの変量 x，y のデータについて

x の偏差と y の偏差の積の平均値を 共分散（きょうぶんさん） といい s_{xy} と表す.

すなわち　2つの変量 x，y のデータについて

対応する n 個の値の組 (x_1, y_1), (x_2, y_2), …, (x_n, y_n) が与えられ

それぞれの平均値を $\overline{x}$，$\overline{y}$ とするとき

$$s_{xy} = \frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n}$$

$$=\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})$$

例 2つの変量 x，y についてのデータを次の表とする.

x	8	4	2	6	10
y	4	5	6	3	2

このとき，x，y の平均値をそれぞれ $\overline{x}$，$\overline{y}$ とすると

$$\overline{x} = \frac{8+4+2+6+10}{5} = \frac{30}{5} = 6$$

$$\overline{y} = \frac{4+5+6+3+2}{5} = \frac{20}{5} = 4$$

x，y の偏差は次の表のようになる.

x	8	4	2	6	10
y	4	5	6	3	2
$x-\overline{x}$	2	−2	−4	0	4
$y-\overline{y}$	0	1	2	−1	−2

共分散を s_{xy} とすると

$$s_{xy} = \frac{2\cdot 0+(-2)\cdot 1+(-4)\cdot 2+0\cdot(-1)+4\cdot(-2)}{5} = \frac{-18}{5}$$

$$= -3.6$$

★共分散と平均値

2つの変量 x，y のデータについて

対応する n 個の値の組 $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ が与えられ

3つの平均値を $\overline{x} = \dfrac{1}{n}\sum_{k=1}^{n} x_k$，$\overline{y} = \dfrac{1}{n}\sum_{k=1}^{n} y_k$，$\overline{xy} = \dfrac{1}{n}\sum_{k=1}^{n} x_k y_k$

とするとき 共分散 s_{xy} は

$$s_{xy} = \overline{xy} - \overline{x}\cdot\overline{y}$$

つまり

$$(x \text{ と } y \text{ の共分散}) = (xy \text{ の平均値}) - \{(x \text{ の平均値}) \times (y \text{ の平均値})\}$$

㊟ $s_{xy} = \dfrac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y})$

$= \dfrac{1}{n}\sum_{k=1}^{n}(x_k y_k - \overline{y}x_k - \overline{x}y_k + \overline{x}\cdot\overline{y})$

$= \dfrac{1}{n}\sum_{k=1}^{n} x_k y_k - \overline{y}\cdot\dfrac{1}{n}\sum_{k=1}^{n} x_k - \overline{x}\cdot\dfrac{1}{n}\sum_{k=1}^{n} y_k + \dfrac{1}{n}\sum_{k=1}^{n}\overline{x}\cdot\overline{y}$

$= \overline{xy} - \overline{y}\cdot\overline{x} - \overline{x}\cdot\overline{y} + \dfrac{1}{n}\cdot n\cdot\overline{x}\cdot\overline{y}$

$= \overline{xy} - \overline{x}\cdot\overline{y}$

㊧ 2つの変量 x，y についてのデータを次の表とする．

x	8	4	2	6	10
y	4	5	6	3	2

このとき，x，y の平均値をそれぞれ $\overline{x}$，$\overline{y}$ とすると

$\overline{x} = \dfrac{8+4+2+6+10}{5} = \dfrac{30}{5} = 6$

$\overline{y} = \dfrac{4+5+6+3+2}{5} = \dfrac{20}{5} = 4$

積 xy について次の表のようになる．

x	8	4	2	6	10
y	4	5	6	3	2
xy	32	20	12	18	20

積 xy の平均を $\overline{xy}$ とすると

$\overline{xy} = \dfrac{32+20+12+18+20}{5} = \dfrac{102}{5} = 20.4$

共分散を s_{xy} とすると

$s_{xy} = \overline{xy} - \overline{x}\cdot\overline{y} = 20.4 - 6\cdot 4 = 20.4 - 24$

$= -3.6$

共分散と相関関係

2つの変量 x，y に対して

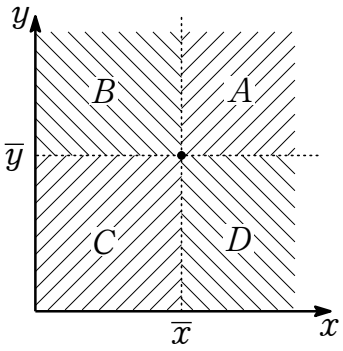
1 共分散が正のとき，x と y の間には正の相関関係があると考えられる．

2 共分散が負のとき，x と y の間には負の相関関係があると考えられる．

3 共分散が 0 のとき，x と y の間には直線的な相関関係はないと考えられる．

(考) 散布図において

$A=\{(x,\ y)\,|\,x>\overline{x},\ y>\overline{y}\}$
$B=\{(x,\ y)\,|\,x<\overline{x},\ y>\overline{y}\}$
$C=\{(x,\ y)\,|\,x<\overline{x},\ y<\overline{y}\}$
$D=\{(x,\ y)\,|\,x>\overline{x},\ y<\overline{y}\}$



となる領域を考えると

$(x_k,\ y_k)\in A$ とすると $x_k-\overline{x}>0,\ y_k-\overline{y}>0$
$(x_k,\ y_k)\in B$ とすると $x_k-\overline{x}<0,\ y_k-\overline{y}>0$
$(x_k,\ y_k)\in C$ とすると $x_k-\overline{x}<0,\ y_k-\overline{y}<0$
$(x_k,\ y_k)\in D$ とすると $x_k-\overline{x}>0,\ y_k-\overline{y}<0$

これより

$(x_k,\ y_k)\in A\cup C$ ならば $(x_k-\overline{x})(y_k-\overline{y})>0$
$(x_k,\ y_k)\in B\cup D$ ならば $(x_k-\overline{x})(y_k-\overline{y})<0$

ここで

共分散 $s_{xy}=\dfrac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n}$

であることから

1 $A\cup C$ に点 $(x_k,\ y_k)$ が多く集まると，s_{xy} の符号は正になる傾向にあり，
右上がりの形がみられる．

2 $B\cup D$ に点 $(x_k,\ y_k)$ が多く集まると，s_{xy} の符号は負になる傾向にあり，
右下がりの形がみられる．

相関係数

共分散を標準偏差の積で割った値を **相関係数**（そうかんけいすう） といい r_{xy} と表す.

すなわち 2 つの変量 x，y のデータの値について

それぞれの標準偏差を s_x，s_y，共分散を s_{xy} とするとき

$$r_{xy}=\frac{s_{xy}}{s_x s_y} \quad \text{ただし } s_x \neq 0 \text{ かつ } s_y \neq 0$$

つまり $\text{（相関係数）}=\dfrac{\text{（共分散）}}{\text{（標準偏差の積）}}$

例 2 つの変量 x，y についてのデータを次の表とする.

x	8	4	2	6	10
y	4	5	6	3	2

このとき，x，y の平均値をそれぞれ $\overline{x}$，$\overline{y}$，標準偏差を s_x，s_y とすると

$$\overline{x}=\frac{8+4+2+6+10}{5}=\frac{30}{5}=6$$

$$\overline{y}=\frac{4+5+6+3+2}{5}=\frac{20}{5}=4$$

x，y の偏差は次の表のようになる.

x	8	4	2	6	10
y	4	5	6	3	2
$x-\overline{x}$	2	-2	-4	0	4
$y-\overline{y}$	0	1	2	-1	-2

x の分散は ${s_x}^2=\dfrac{2^2\times 2+4^2\times 2}{5}=\dfrac{40}{5}=8$

y の分散は ${s_y}^2=\dfrac{1^2\times 2+2^2\times 2}{5}=\dfrac{10}{5}=2$

ゆえに $s_x=2\sqrt{2}$，$s_y=\sqrt{2}$

共分散を s_{xy} とすると

$$s_{xy}=\frac{2\cdot 0+(-2)\cdot 1+(-4)\cdot 2+0\cdot(-1)+4\cdot(-2)}{5}=\frac{-18}{5}=-3.6$$

よって 相関係数を r_{xy} として

$$r_{xy}=\frac{s_{xy}}{s_x s_y}=\frac{-3.6}{2\sqrt{2}\cdot\sqrt{2}}=-0.9$$

★相関係数の計算式

2つの変量 x，y のデータの値についての相関係数 r_{xy} は

$$r_{xy}=\frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{\sqrt{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}\sqrt{(y_1-\overline{y})^2+(y_2-\overline{y})^2+\cdots+(y_n-\overline{y})^2}}$$

$$=\frac{\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\sum_{k=1}^{n}(x_k-\overline{x})^2}\sqrt{\sum_{k=1}^{n}(y_k-\overline{y})^2}}$$

つまり

$$(x\text{ と }y\text{ の相関係数})=\frac{(x\text{ と }y\text{ の偏差の積の和})}{(x\text{ の偏差の 2 乗の和の平方根})\times(y\text{ の偏差の 2 乗の和の平方根})}$$

㊟ 相関係数の定義より

$$r_{xy}=\frac{\frac{(x_1-\overline{x})(y_1-\overline{y})+(x_2-\overline{x})(y_2-\overline{y})+\cdots+(x_n-\overline{x})(y_n-\overline{y})}{n}}{\sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}}\sqrt{\frac{(y_1-\overline{y})^2+(y_2-\overline{y})^2+\cdots+(y_n-\overline{y})^2}{n}}}$$

$$=\frac{\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})^2}\sqrt{\frac{1}{n}\sum_{k=1}^{n}(y_k-\overline{y})^2}}$$

これらの分母と分子を n 倍しただけである．

㊧ 2つの変量 x，y についてのデータの相関係数 r_{xy} について

x	8	4	2	6	10
y	4	5	6	3	2
$x-\overline{x}$	2	-2	-4	0	4
$y-\overline{y}$	0	1	2	-1	-2

$$r_{xy}=\frac{2\cdot0+(-2)\cdot1+(-4)\cdot2+0\cdot(-1)+4\cdot(-2)}{\sqrt{2^2+(-2)^2+(-4)^2+0^2+4^2}\sqrt{0^2+1^2+2^2+(-1)^2+(-2)^2}}$$

$$=\frac{-18}{\sqrt{40}\sqrt{10}}=\frac{-18}{20}=-0.9$$

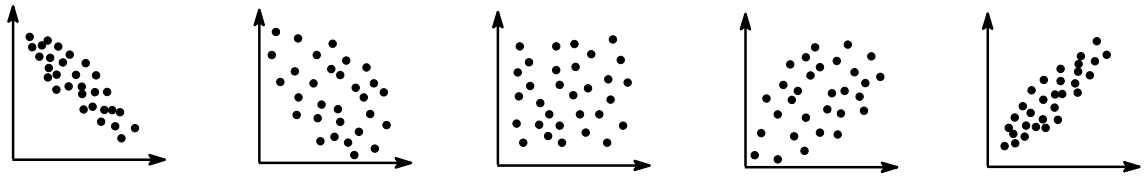
相関係数と散布図

相関係数 r_{xy} について

1 $-1 \leqq r_{xy} \leqq 1$

2 正の相関関係が強いほど r_{xy} の値は 1 に近づく.

3 負の相関関係が強いほど r_{xy} の値は -1 に近づく.



強い ← 負の相関関係 → 弱い　　弱い ← 正の相関関係 → 強い

$r_{xy} \fallingdotseq -1$　　$r_{xy} \fallingdotseq 0$　　$r_{xy} \fallingdotseq 1$

補 すべての点が傾き正の直線上に存在するならば $r_{xy} = 1$

すべての点が傾き負の直線上に存在するならば $r_{xy} = -1$

すべての点が横軸または縦軸に平行な直線上に存在するならば r_{xy} は計算できない.

☆平均値の変量の関係

2つの変量 x，y の平均値をそれぞれ $\overline{x}$，$\overline{y}$ とする．

a，b を実数の定数とするとき，次が成り立つ．

1 $\overline{ax+b}=a\overline{x}+b$

2 $\overline{x+y}=\overline{x}+\overline{y}$

3 $\overline{x-y}=\overline{x}-\overline{y}$

考 1 変量 x のデータを n 個の値 x_1，x_2，…，x_n とするとき

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n} \quad \cdots\cdots①$$

このもとで，$ax+b$ の平均値は

$$\begin{aligned}\overline{ax+b}&=\frac{(ax_1+b)+(ax_2+b)+\cdots+(ax_n+b)}{n}\\&=\frac{a(x_1+x_2+\cdots+x_n)+\overbrace{b+b+\cdots+b}^{n\text{個}}}{n}\\&=a\cdot\frac{x_1+x_2+\cdots+x_n}{n}+\frac{bn}{n}\\&=a\quad\overline{x}\quad+\quad b\quad(\because ①)\end{aligned}$$

2 変量 x のデータを n 個の値 x_1，x_2，…，x_n

変量 y のデータを n 個の値 y_1，y_2，…，y_n

とするとき

$$\overline{x}=\frac{x_1+x_2+\cdots+x_n}{n},\ \overline{y}=\frac{y_1+y_2+\cdots+y_n}{n} \quad \cdots\cdots②$$

このもとで，$x+y$ の平均値は

$$\begin{aligned}\overline{x+y}&=\frac{(x_1+y_1)+(x_2+y_2)+\cdots+(x_n+y_n)}{n}\\&=\frac{(x_1+x_2+\cdots+x_n)+(y_1+y_2+\cdots+y_n)}{n}\\&=\frac{x_1+x_2+\cdots+x_n}{n}+\frac{y_1+y_2+\cdots+y_n}{n}\\&=\quad\overline{x}\quad+\quad\overline{y}\quad(\because ②)\end{aligned}$$

3 ② のもとで，$x-y$ の平均値は

$$\begin{aligned}\overline{x-y}&=\frac{(x_1-y_1)+(x_2-y_2)+\cdots+(x_n-y_n)}{n}\\&=\frac{(x_1+x_2+\cdots+x_n)-(y_1+y_2+\cdots+y_n)}{n}\\&=\frac{x_1+x_2+\cdots+x_n}{n}-\frac{y_1+y_2+\cdots+y_n}{n}\\&=\quad\overline{x}\quad-\quad\overline{y}\quad(\because ②)\end{aligned}$$

☆分散と標準偏差の変量の関係

変量 x の分散を ${s_x}^2$，標準偏差を s_x とし，a，b を実数の定数とする．

変量 $ax+b$ の分散を ${s_{ax+b}}^2$，標準偏差を s_{ax+b} とするとき，次が成り立つ．

1 ${s_{ax+b}}^2 = a^2 {s_x}^2$

2 $s_{ax+b} = |a| s_x$

考 1 変量 x のデータを n 個の値 x_1，x_2，…，x_n とし，平均値を $\overline{x}$ とするとき

$${s_x}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2 \quad \cdots\cdots①$$

このもとで，$ax+b$ の分散は

$$\begin{aligned}
{s_{ax+b}}^2 &= \frac{1}{n}\sum_{k=1}^{n}\{(ax_k+b) - \overline{ax+b}\}^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}\{(ax_k+b) - (a\overline{x}+b)\}^2 \quad \left(\because \boxed{\text{平均値の変量の関係}}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\{a(x_k - \overline{x})\}^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}a^2(x_k - \overline{x})^2 \\
&= a^2 \cdot \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2 \\
&= a^2 \qquad {s_x}^2 \qquad (\because ①)
\end{aligned}$$

2 1 より $s_{ax+b} = \sqrt{a^2 {s_x}^2} = |a| s_x$

例 変量 x の分散が 2 のとき，変量 $2x+1$ の分散は

${s_x}^2 = 2$ とすると ${s_{2x+1}}^2 = 2^2 {s_x}^2 = 4 \cdot 2 = 8$

☆共分散の変量の関係

2つの変量 x，y の共分散を s_{xy}，a，b，c，d を実数の定数，$ac \neq 0$ とする．

2つの変量 $ax+b$，$cy+d$ の共分散を $s_{(ax+b)(cy+d)}$ とするとき，次が成り立つ．

$$s_{(ax+b)(cy+d)} = ac\, s_{xy}$$

㊟ 2つの変量 x，y のデータについて，平均値をそれぞれ $\overline{x}$，$\overline{y}$ とし，
対応する n 個の値の組 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) とするとき

$$s_{xy} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}) \quad \cdots\cdots①$$

このもとで，2つの変量 $ax+b$，$cy+d$ の共分散は

$$\begin{aligned}
s_{(ax+b)(cx+d)} &= \frac{1}{n}\sum_{k=1}^{n}\{(ax_k+b)-\overline{(ax+b)}\}\{(cy_k+d)-\overline{(cy+d)}\} \\
&= \frac{1}{n}\sum_{k=1}^{n}\{(ax_k+b)-(a\overline{x}+b)\}\{(cy_k+d)-(c\overline{y}+d)\} \\
&\qquad\qquad \left(\because \boxed{\text{平均値の変量の関係}}\right) \\
&= \frac{1}{n}\sum_{k=1}^{n}\{a(x_k-\overline{x})\}\{c(y_k-\overline{y})\} \\
&= ac\cdot\frac{1}{n}\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}) \\
&= ac \qquad s_{xy} \qquad (\because ①)
\end{aligned}$$

㊟ 2つの変量 x，y の共分散が2のとき，2つの変量 $2x+1$，$3y-1$ の共分散は

$s_{xy}=2$ とすると $s_{(2x+1)(3x-1)} = 2\cdot 3 s_{xy} = 6\cdot 2 = 12$

☆相関係数の変量の関係

2つの変量 x，y の相関係数を r_{xy}，a，b，c，d を実数の定数，$ac \neq 0$ とする．

2つの変量 $ax+b$，$cy+d$ の相関係数を $r_{(ax+b)(cy+d)}$ とするとき

次が成り立つ．

$$r_{(ax+b)(cy+d)} = \begin{cases} r_{xy} & (ac>0) \\ -r_{xy} & (ac<0) \end{cases}$$

㊟ 2つの変量 x，y の標準偏差をそれぞれ s_x，s_y とし，共分散を s_{xy} とする．
さらに，2つの変量 $ax+b$，$cy+d$ の標準偏差をそれぞれ s_{ax+b}，s_{cy+d}，
共分散を $s_{(ax+b)(cy+d)}$ とする．

$$\begin{aligned} r_{(ax+b)(cx+d)} &= \frac{s_{(ax+b)(cy+d)}}{s_{ax+b}s_{cy+d}} \\ &= \frac{ac\,s_{xy}}{|a|s_x|c|s_y} \quad \left(\because \boxed{\text{共分散の変量の関係}}, \boxed{\text{標準偏差の変量の関係}}\right) \\ &= \frac{acs_{xy}}{|ac|s_{xy}} \\ &= \begin{cases} ac>0 \text{ のとき} & \dfrac{ac\,s_{xy}}{ac\,s_xs_y} = \dfrac{s_{xy}}{s_xs_y} = r_{xy} \\ ac<0 \text{ のとき} & \dfrac{ac\,s_{xy}}{-ac\,s_xs_y} = -\dfrac{s_{xy}}{s_xs_y} = -r_{xy} \end{cases} \end{aligned}$$

㊟ 2つの変量 x，y の相関係数が 0.8 のとき，

2つの変量 $2x+1$，$3y-1$ の相関係数は 0.8（変わらない）

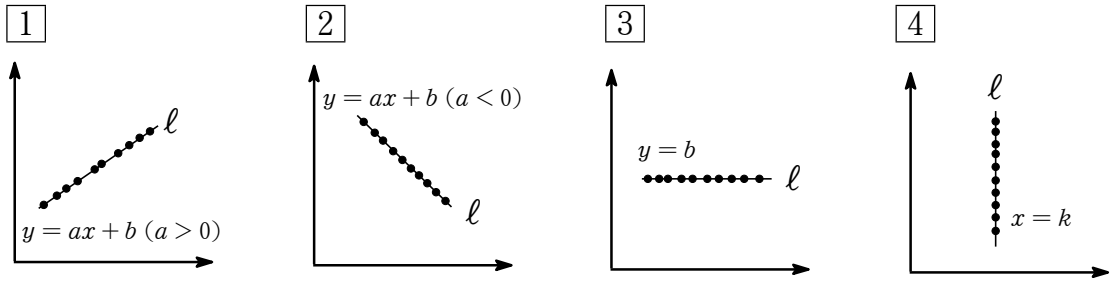
2つの変量 $-2x+1$，$3y-1$ の相関係数は -0.8（-1 倍）

★直線上に分布する相関係数

2つの変量 x，y の相関係数を r_{xy} とする．

散布図において，対応する n 個の値の組 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) のすべての点がある直線 ℓ 上に分布する場合について，次のようになる．

ただし a，b，k を実数の定数とする．



1 $\ell : y = ax + b\ (a > 0)$ ならば $r_{xy} = 1$

2 $\ell : y = ax + b\ (a < 0)$ ならば $r_{xy} = -1$

3 $\ell : y = b$ ならば r_{xy} は計算できない．

4 $\ell : x = k$ ならば r_{xy} は計算できない．

㊟ n 個の点 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) がすべて $\ell : y = ax + b$ 上に分布するならば

$$y_1 = ax_1 + b,\ y_2 = ax_2 + b,\ \cdots,\ y_n = ax_n + b$$

$a \neq 0$ とすると

y の標準偏差 $s_y = s_{ax+b} = |a| s_x$ ……①

y の偏差は $y - \overline{y} = ax + b - \overline{(ax + b)} = ax + b - (a\overline{x} + b) = a(x - \overline{x})$

このことから x と y の偏差の積は $(x - \overline{x})(y - \overline{y}) = a(x - \overline{x})^2$ となる．

つまり，共分散は x の分散を a 倍したものとなり $s_{xy} = a{s_x}^2$ ……②

①，② から相関係数は

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{a\,{s_x}^2}{s_x |a|\, s_x} = \frac{a}{|a|}$$

1 $a > 0$ ならば $r_{xy} = \dfrac{a}{a} = 1$

2 $a < 0$ ならば $r_{xy} = -\dfrac{a}{a} = -1$

3 $a = 0$ ならば $\ell : y = b$

このとき $y_1 = y_2 = \cdots = y_n$

標準偏差 $s_y = 0$ であるから相関係数 r_{xy} は計算できない．

4 n 個の点 (x_1, y_1)，(x_2, y_2)，…，(x_n, y_n) がすべて $\ell : y = k$ 上に分布するならば

$$x_1 = x_2 = \cdots = x_n = k$$

標準偏差 $s_x = 0$ であるから相関係数 r_{xy} は計算できない．

仮説検定

得られたデータをもとに，ある事柄が正しいことを判断するのに，

仮説を棄却(ききゃく)することで正しいと判断する

または

仮説を棄却しないで正しいとは判断できない

ことを決める方法を 仮説検定(かせつけんてい) という．

補 「棄却する」とは「棄(す)て去(さ)る」こと

例 ある人に「高級ワイン」と「安価のワイン」のどちらかわからないように試飲してもらい，
「高級ワイン」を答えてもらった所，10 回中 9 回を的中させた．
この人は確かにワインの味がわかって答えていると判断したい．
仮に当てずっぽうで答えたとしても，10 回中 9 回以上を的中することは起こりえる．
そこで，この人が「当てずっぽうで答えている」と仮説を立て，その仮説のもとで，
ある事象が起こり得る確率にもとづいて「この人はワインの味をわかって的中している」
かを判断する．（数学 Ⅰ では実験で行う場合が多い）
仮説が棄却されると，この人はワインの味がわかる人だと判断できる．
このような方法を仮説検定という．

話 未成年はアルコール類を飲んではいけないので，上のような例は教科書にはない．
とある本では「水道水」と「ミネラルウォーター」にした．

仮説検定の考え方

得られたデータから，主張 H_1 が正しいと判断するのに，

主張 H_1 に反する仮説の主張 H_0 を立てる．

ここで，H_0 を帰無(きむ)仮説，H_1 を対立(たいりつ)仮説という．

帰無仮説 H_0 のもとで，棄却すべき確率 p を求める．

このとき，起こりやすさの基準となる確率 α を定めておく．

この α を有意水準(ゆういすいじゅん)といい，0.05(5 %) とする場合が多い．

p と α の大小関係で，次の 2 つの場合になる．

1 $p < \alpha$ ならば

p が小さすぎるので，H_0 は棄却され，H_1 が正しいと判断できる．

2 $p > \alpha$ ならば

p が小さくなく，H_0 は棄却されず，H_1 が正しいと判断できない．

(補) $p = \alpha$ の場合は，定義があいまいで問題ではそうなることはないと考えておけばよい．

(補) 2 H_1 が正しくないと判断するわけではない．

(補) おおざっぱな説明だが，対立仮説は「本当はこうではないのかという仮説」
帰無仮説は「無かったことにしたい仮説」

(例) ある人に「高級ワイン」と「安価のワイン」のどちらかわからないように試飲してもらい，「高級ワイン」を答えてもらった所，10 回中 9 回を的中させた．
この人は確かにワインの味がわかって答えていると判断したい．そこで
帰無仮説 H_0：「高級ワイン」を選ぶ確率は $\frac{1}{2}$
（「安価のワイン」を選ぶ確率も $\frac{1}{2}$ で当てずっぽうで選んでいる）
を立てる．
対立仮説 H_1：「高級ワイン」を選ぶ確率は $\frac{1}{2}$ よりも大きい
（当てずっぽうではなく，ワインの味がわかって選んでいる）
という主張が正しいかを有意水準 5 % の仮説検定で考察する．
帰無仮説 H_0 のもとで，10 回試飲して，「高級ワイン」を 9 回以上的中する確率を p とすると

$$p = {}_{10}\mathrm{C}_9\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right) + {}_{10}\mathrm{C}_{10}\left(\frac{1}{2}\right)^{10} = \frac{{}_{10}\mathrm{C}_9 + {}_{10}\mathrm{C}_{10}}{2^{10}} = \frac{10+1}{1024} = \frac{11}{1024}$$
$$= 0.0107\cdots \text{（約 0.1 %）}$$

これは有意水準 0.05(5 %) よりも小さいので，p は小さすぎる．
これより帰無仮説 H_0 は棄却され，対立仮説 H_1 が正しいと判断できる．
よって，この人は確かにワインの味がわかって答えていると判断できる．