千葉県高等学校教育研究会数学部会

統計教育のアウトライン

統計をどう教えるか

小林 健太（一橋大学）

自己紹介

氏名：小林健太

経歴：京都大学理学部　卒業

数学（確率論）を専攻

京都大学理学研究科数学・数理解析専攻　修士課程・博士後期課程　修了

数学（数値解析）を専攻

高精度計算、精度保証付き数値計算などについて研究

九州大学数理学研究員　研究員、金沢大学理工研究域　准教授、

一橋大学商学部・大学院商学研究科　准教授　を経て現職

現在は、主に有限要素法の誤差解析および精度保証付き数値計算の研究に従事

はじめに

はじめにお断りしておきますが、**統計学は専門ではありません。**ただ、

一橋大学商学部では**「ビジネス統計入門」**（学部3年生に統計学の初歩から仮説検定までを教える）

一橋大学大学院経営管理プログラム（MBAコース）では**「企業データ分析」**（MBAコースの学生に統計を用いて社会科学に関するデータ分析手法を教える）

などの授業を担当したことはあります。

また、啓林館高校数学教科書の執筆にも携わっており、数学Ⅰ**「データの分析」**および数学Ｂ**「確率分布と統計的な推測」**について、高校生にいかに分かり易く解説するか考えてきました。

内容

統計は専門外ではありますが、私自身が統計の勉強してきた際に気になったところなどを思い出しながら、以下のような内容で話をさせて頂ければと思います。

- 統計と数学の違い
- 統計の基本的な考え方
- 仮説検定とはいったい何をやっているのか
- 中心極限定理について
- 数Bの統計で少し曖昧になっている点を詳しく

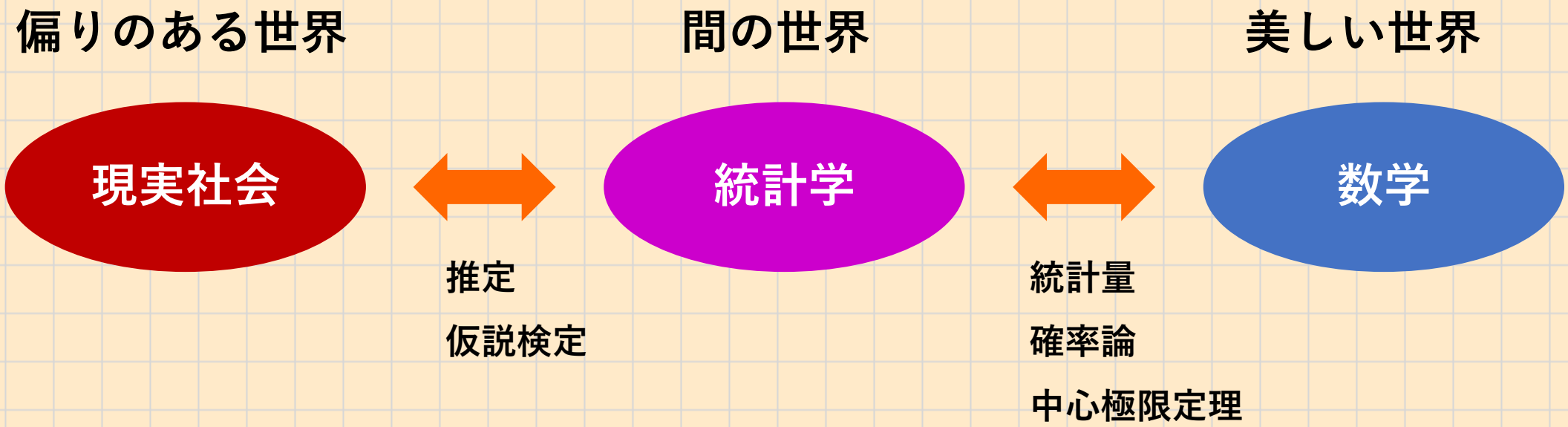
また、より広い視点から見た方が高校での統計が良く理解できるのではないかと考え、

- 回帰直線、重回帰分析、主成分分析

などについても簡単に説明したいと思います。

統計学は数学か？

統計学は数学と現実社会の間に位置する分野である。



統計学は、現実社会に対応しなければならないという要請と、数学の理論の中で扱わなければならないという要請から、制約を受ける。

その２つの要請は、ジレンマとなるケースもある。

数学からの制約

中心の統計量としては、平均値や中央値が知られている。

例：ある会社の年収　　社員A　250万円
社員B　450万円
社員C　300万円
社員D　400万円
社員E　2200万円

年収の平均：720万円、中央値：400万円

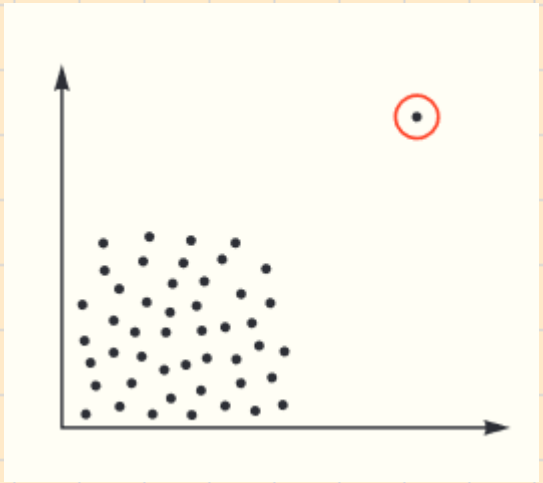
全体の様子をよく表すという観点からは、平均より中央値の方が統計量としては優れていると考えられる。

しかし、統計学では中央値より平均値の方が圧倒的に用いられる機会が多い。

なぜなら、平均値は中央値より数学的に取り扱い易いから。

数学からの制約

２組のデータ間の相関の大きさを表わすのに相関係数が用いられることが多い。



しかし相関係数は外れ値の影響を受け易く、必ずしも良い指標とは言えないケースもある。

しかし数学的には良い性質を持つ。２つの変量の組 $(x_1, y_1), \cdots, (x_n, y_n)$ について以下のようなベクトルを考え、

$$\vec{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x}), \qquad \vec{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - \bar{y})$$

n 次元ユークリッド空間において $\vec{x}$ と $\vec{y}$ のなす角を θ とすると

$$r = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \cos\theta$$

という関係があり、その数学的な性質が用いられる。

ただし、それは**数学の側の事情**である。

現実社会からの制約

数学の世界は、公理の上に構築された、いわゆる美しい世界。

一方で現実社会は様々な要因が複雑に入り組んだ雑音だらけの世界。

そのような現実世界を直接数学で扱うことはできない。

例：

偏りのあるさいころを数学的に扱うことはできない。なぜなら、偏り方やその要因も様々であるし、偏り方の度合いも様々であるからである。

一方で、偏りのないさいころは数学的に取り扱うことができる。

仮説検定は、複雑で雑音だらけの偏った世界と、偏りのない数学の世界を橋渡しする、上手く考えられたロジックである。

仮説検定

さいころの例で考える。

さいころを何度か投げたところ、なんとなく１の目が多く出るような気がする。

このさいころが偏っているか調べたい。

しかし、１の目が多く出るような偏ったさいころを数学的に扱うことはできない。

そこでまず、偏りのないさいころを考え、以下の仮説を立てる。

帰無仮説 H_0：さいころには偏りがない

帰無仮説の下で実際の統計データを検証し、確率的に非常にまれなことが起きているなら

対立仮設 H_1：さいころは１の目が出易い

を採用する。

数Ⅰ「データの分析」より

右の表は，どの目が出ることも同様に確からしいさいころを100回投げて1の目が出た回数を記録することを1000回行った結果をまとめたものである。この表を用いて，次のさいころP，Qは1の目が出やすいと判断してよいか答えよ。ただし，起こる割合が5%以下であればほとんど起こり得ないと判断するものとする。

(1)　100回投げたら1の目が30回出たさいころP

(2)　100回投げたら1の目が20回出たさいころQ

1の目が出た回数	回
0 ～ 8	0
9	17
10	26
11	37
12	52
13	73
14	93
15	97
16	115
17	101
18	86
19	76
20	71
21	50
22	41
23	30
24	15
25	9
26	4
27	2
28	3
29	1
30	1
31～100	0

例えば（2）だと、表の中で20回以上1が出た割合を求めて5％と比較することになる。

なぜちょうど20回ではなく20回以上なのか？

本来は、あらかじめ5％の割合となる棄却域（上の表だと24回以上）を設定して、そこに入るかどうかを検証するのである。

棄却域の設定

稀な事が起きている、というケースを定めたものが棄却域である。

今回の場合、

1151132161...1614111213（1が多く出る）は棄却域に入れる

しかし

6636166616...2664656266（6が多く出る）

や

1234561234...1234561234（特徴的なパターンが現れる）

などは棄却域に入れない。

なぜなら、元々の関心が「1の目が出易いかどうか」だから。

もし、「特定の目が出易いかどうか」を調べたい場合には、6が多く出るようなパターンも棄却域に入れることになる。

片側検定

A社で新しい飲料を開発し，最終候補としてPとQが残った。
消費者 30 人を偏りなく選んでどちらを好むか調査したところ，
P を好む人が 24 人，Q を好む人が 6 人であった。
一般に，P の方が好まれそうだと判断してもよいか考えてみよう。

PとQを好む人は半々であると仮定する。このとき，30 人中 24 人以上がPを好むと答えることがどのくらい起こり得るかを調べてみる。
そのために，表裏が同じ割合で出る 30 枚の硬貨を同時に投げて表が出た枚数を記録することを 1000 回行った。
右の表はその結果である。この表から，24 回以上表が出る割合を求めると，

表が出た枚数	回
0～6	0
7	1
8	2
9	14
10	22
11	50
12	63
13	115
14	118
15	145
16	134
17	131
18	99
19	46
20	35
21	14
22	5
23	3
24	2
25	1
26～30	0

$$\frac{2}{1000}+\frac{1}{1000}=\frac{3}{1000}=0.003 \qquad 0.3\,\%$$

となり，ほとんど起こり得ない。
よって，P の方が好まれると判断することが妥当である。

この場合、QよりPの方が好まれるかどうかを知りたいので

帰無仮説 H_0：PとQの好まれ方は同じである

対立仮説 H_1：Pの方がQより好まれる

とし、コインの表の出た枚数をPを好むことに対応させ、棄却域は左の表で21回以上とする。

これは片側検定を行ってることになる。

片側検定の意味

無作為に消費者を選んだときに、その人がPを好む確率を p とすると

帰無仮説 H_0 ： $p = \frac{1}{2}$　　対立仮説 H_1 ： $p > \frac{1}{2}$

を検定していることになる。

帰無仮説が $p \leq \frac{1}{2}$ ではないので、帰無仮説と対立仮設で全事象をカバーしていないように思えるかもしれない。しかし今回の場合は、 PがQより好まれるかが関心事なので、PよりQを好むケースは特別視しない。

つまり、 PよりQを好む人が非常に多いという結果が出たとしても、特別なことが起こったとは考えないのである。⇒　認識しないものは存在しないのと同じ

さいころで１が多く出るか検定する場合に、２が多く出たとしても、特別視しないのと同じである。

検定の手順

検定を行う際には、まず、何をもって「稀なこと」が起きたとするかを決め、それに応じて帰無仮説の棄却域を設定する。

さいころの場合は、１の目が多く出た場合のみ「稀なこと」が起きたとして、その他の場合は特別なことが起きているとは考えない。

飲料PとQの場合は、PよりQを好む人が非常に多い場合のみ「稀なこと」が起きていると考える。

このように、何をもって「稀なこと」とするかに恣意性がある。また、その判断には、現象の機序に関する理解も関係する。

例えば、さいころを投げて、123456123456…、と目が出たとしても、それは偶然と考えるしかない。なぜなら、それぞれの目には関連が無いと考えるのが自然だからです。

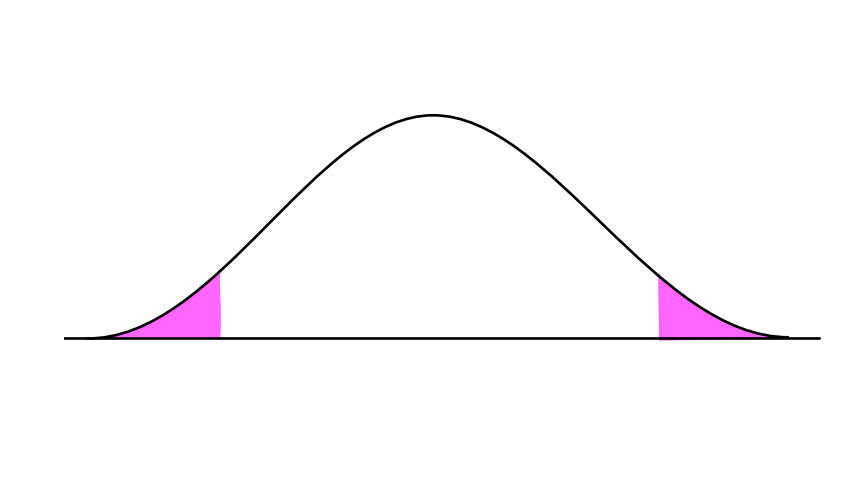
ここは、統計が実際の社会に制約を受ける部分である。

両側検定

飲料PとQの場合、PとQの嗜好に大きな差がある、というのを「稀なこと」とすると、

帰無仮説 H_0： $p = \frac{1}{2}$　　対立仮説 H_1： $p \neq \frac{1}{2}$

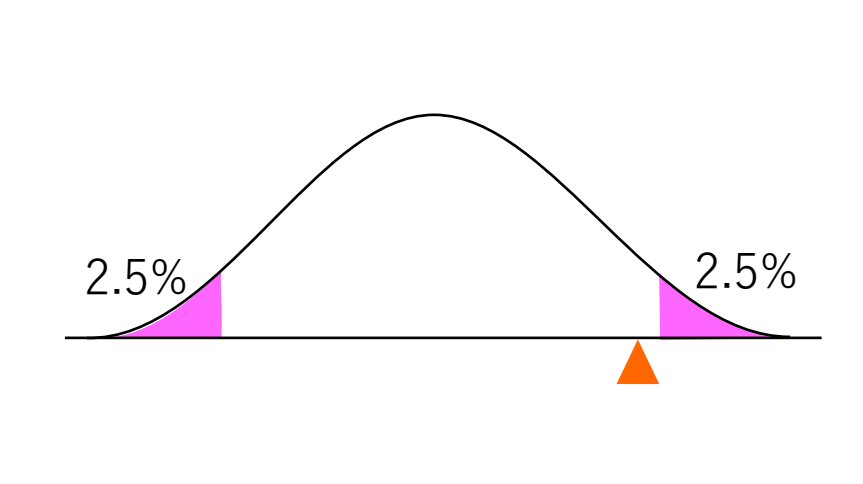
その場合、帰無仮説の棄却域は以下のように分布の両側に設定される。

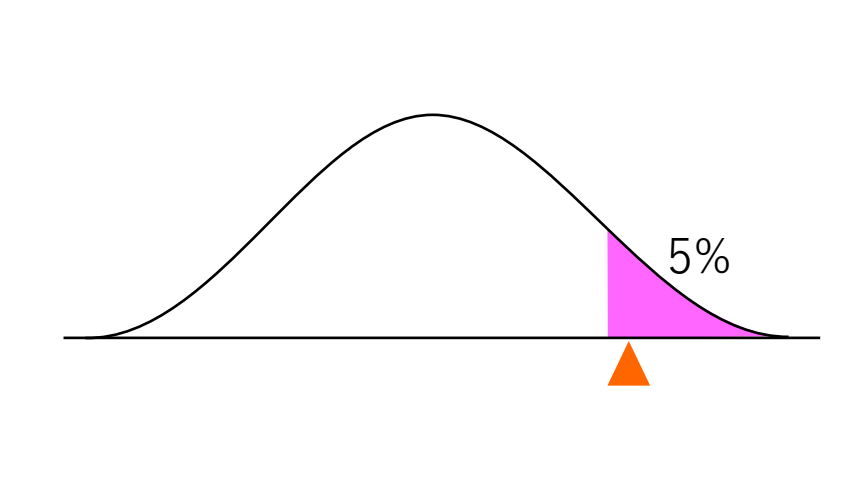


これを**両側検定**という。

片側検定と両側検定

棄却域は通常、ある決まった割合の確率になるように設定される。これを**優位水準**といい、５％がよく用いられる。５％というのも社会的な値である。





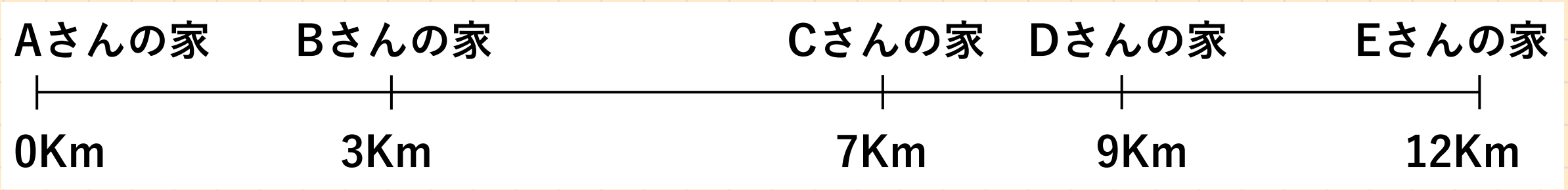
両側検定だと帰無仮説が棄却できない場合でも、片側検定では棄却できることがある。そのため、統計的に有意な結果を得るために片側検定が好んで用いられる。

しかしそもそも、検定においては「稀な場合」が何かをまず決めて棄却域を決めるべきなので、有意な結果を得るために無理に片側検定を適用するというのは許されないと考えるべきである。

平均偏差と標準偏差

現実社会では、**標準偏差**よりも**平均偏差**の方が便利なことが多い。

中央値（メディアン）は偏差の和を最小化する、という性質がある（サンプルサイズが偶数の場合は、中央値を挟む２つの値の間の任意の数が偏差の和を最小化する）。



この場合、中央値であるCさんの家に集まると、全員の移動距離の和が最小になる。

その時の移動距離の平均は平均偏差となる。

しかし統計学では、平均偏差よりも標準偏差が圧倒的に多く使用される。その理由は、標準偏差の2乗である**分散が、数学的に極めて良い性質を持つ**からである。

これは、社会とは関係無く数学的な理由による。

確率変数の独立と分散の性質

標本抽出の結果を確率変数の実現値と考える。

確率変数 X と Y が独立のとき

$$E(XY) = E(X)E(Y)$$

このとき、

$$E\big((X-\bar{X})(Y-\bar{Y})\big) = E(XY) - E(\bar{X}Y) - E(X\bar{Y}) + E(\bar{X}\bar{Y})$$
$$= E(X)E(Y) - \bar{X}E(Y) - E(X)\bar{Y} + \bar{X}\bar{Y} = 0$$

これより、 X と Y が独立のとき

$$V(X+Y) = E\big((X-\bar{X}+Y-\bar{Y})^2\big) = E\big((X-\bar{X})^2\big) + E\big((Y-\bar{Y})^2\big) = V(X) + V(Y)$$

これより、 $X_1, X_2, \cdots, X_n$ が独立のとき

$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$$

分散の性質の直感的な解釈

(X, Y) が等確率で $(x_1, y_1), \cdots, (x_n, y_n)$ を取るとき、以下のようなベクトルを考えると

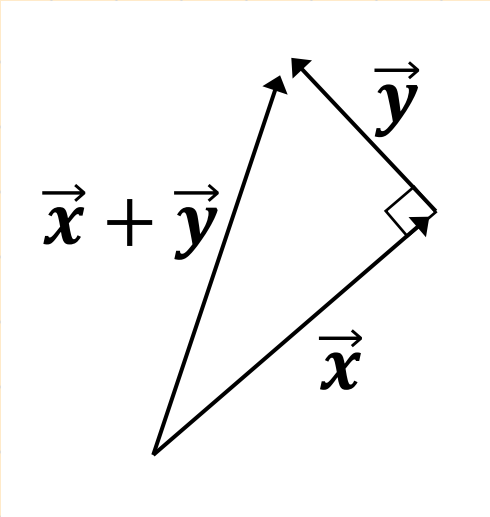
$$\vec{x} = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_n - \bar{x}), \qquad \vec{y} = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_n - \bar{y})$$

X と Y が独立のとき、 $\vec{x} \cdot \vec{y} = 0$

$$nV(X + Y) = |\vec{x} + \vec{y}|^2 = |\vec{x}|^2 + |\vec{y}|^2 = nV(X) + nV(Y)$$

つまり、確率変数の独立性はベクトルの直交、
分散の性質は三平方の定理と解釈することができる。

厳密には、$\vec{x} \cdot \vec{y} = 0$ が成り立っても X と Y が独立とは
限らないので、あくまで直感的なものだが。



中心極限定理

$X_1, X_2, \cdots, X_n$ が独立かつ同分布のとき、個々の平均を μ、 分散を σ^2 とすると、

$$S_n = X_1 + X_2 + \cdots + X_n$$

の平均は $n\mu$、分散は $n\sigma^2$ となる。そこで、

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$$

と標準化すると、 Z_n は平均 0 かつ分散 1 となる。それだけでなく、 Z_n は標準正規分布に収束する。ただし、収束の意味には注意が必要である。

収束というのは、**分布の意味で収束する**ということである。すなわち、任意 a, b に対し、

$$P(a \leq Z_n \leq b) \to P(a \leq Z \leq b) \qquad (n \to \infty)$$

ただし、 Z は**標準正規分布**に従う確率変数である。

これを**中心極限定理**という。中心極限定理は、平均と分散を持つ任意の確率変数について成り立ち、統計や確率の中で最も重要な定理の一つである。

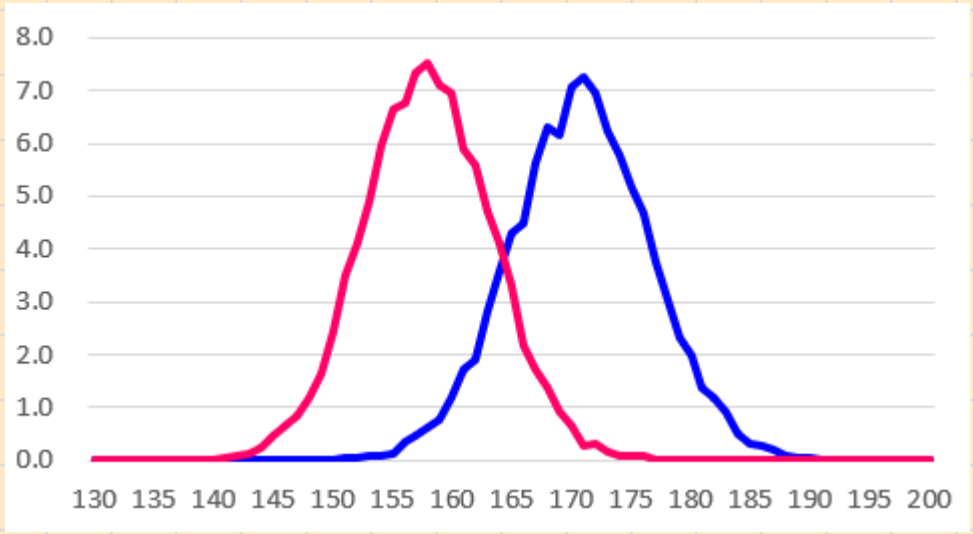
中心極限定理の影響

$X_1, X_2, \cdots, X_n$ が独立であるが、必ずしも同分布でないとしても、

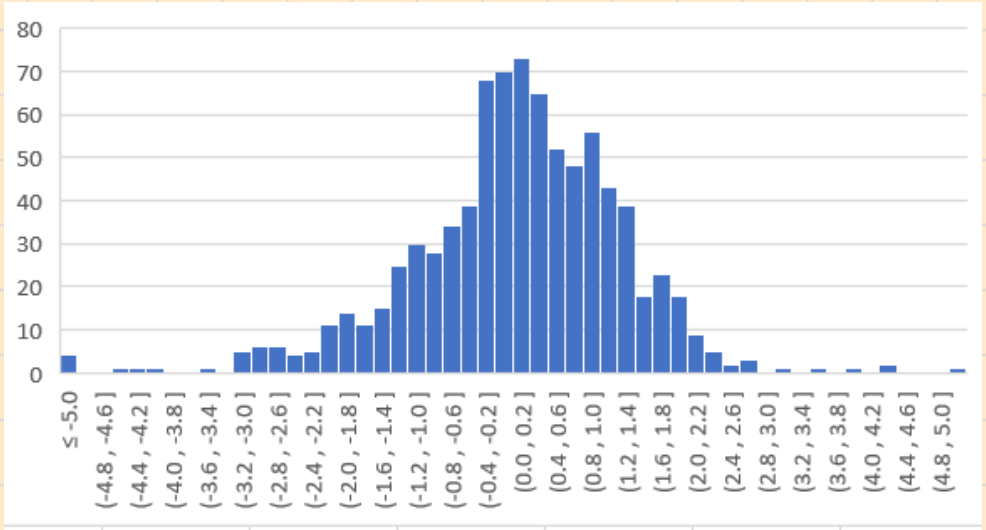
$$S_n = X_1 + X_2 + \cdots + X_n$$

を標準化したものは正規分布に近づく（ただし、個々の X_k の分散が一定値以下である等、条件はある）。

自然現象や社会現象は、様々な独立な要因の和として考えられることが多いので、正規分布に近い分布をすることが多い。



2021年度全国17歳身長分布
青-男、赤-女、1cm刻みごとの割合（%）



2020年1月7日～2023年6月12日
日経平均株価前日比（%）

中心極限定理の利用

元の分布が正規分布でなくとも、その分布を持つ母集団から独立に取り出したサンプルの和や平均は正規分布に従うと考えてよい。

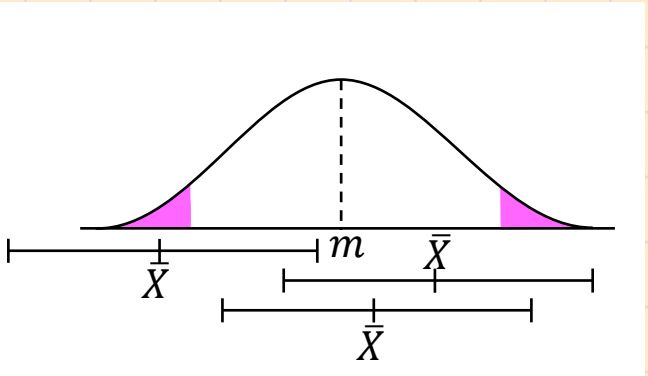
これにより、正規分布を仮定した検定を行うことができる。

中心極限定理はすごい！

例えば、偏りのないさいころを100回投げて1の目が出る回数は、$\frac{1}{6}$の確率で1、$\frac{5}{6}$の確率で0を取るような確率変数（ベルヌーイ分布に従う確率変数）を独立に100個足し合わせたものと考えることができる。

サンプルサイズが概ね30個以上であれば、その和や平均を正規分布とみなしてよいとされている（正規分布で近似することによる誤差が、棄却域の大きさ５％などに比べてかなり小さくなる）。

信頼区間と棄却域



母平均の推定

標本の大きさ n が大きいとき，標本平均を $\overline{X}$，標本の標準偏差を s とすると，母平均 m に対する信頼度 95％ の信頼区間は，

$$\left[\overline{X}-1.96\times\frac{s}{\sqrt{n}},\ \overline{X}+1.96\times\frac{s}{\sqrt{n}}\right]$$

母平均 m の95％信頼区間とは、95％の確率で m を含む区間のことである。

標本平均 $\bar{X}$ は標本抽出のたびに変わるので、信頼区間も標本抽出ごとに変わり、m を含んだり含まなかったりする。

h を定数とするとき、

$[a-h, a+h]$ が b を含む　⇔　$|a-b| \leq h$　⇔　$[b-h, b+h]$ が a を含む

であるから、$\bar{X}$ が有意水準5％の両側検定の棄却域に入ることと、95％信頼区間が母平均 m を含まないことは同じである。

帰無仮説で母平均 m を仮定していた場合は棄却される。

標本分散と不偏性

信頼区間や棄却域を求める際には、母集団の分散を知る必要がある。

高校の範囲では、母分散は標本分散で代用していた。

サンプルサイズ n が大きくなれば標本分散は母分散に近づくので問題は無いが、 n が小さい場合には誤差が大きくなる。

高校の教科書で用いられるのは

$$s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n}$$

だが、実際は不偏分散

$$s^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}$$

が用いられている。

不偏分散については、母分散を σ^2 とすると $E(\sigma^2) = E(s^2)$ が成り立つ（不偏性）。

特に高校数学では、n が大きいときに収束するなら、あまり細かいことには拘らないような記述になっている。

t 検定

$X_1, X_2, \cdots, X_n$ が独立に正規分布 $N(\mu, \sigma^2)$ に従うとき、

$$S_n = X_1 + X_2 + \cdots + X_n$$

も正規分布になる（独立な正規分布の和は正規分布になる⇒正規分布の再生性）。

S_n を不偏分散 s^2 を用いて

$$Z_n = \frac{S_n - n\mu}{\sqrt{n}s}$$

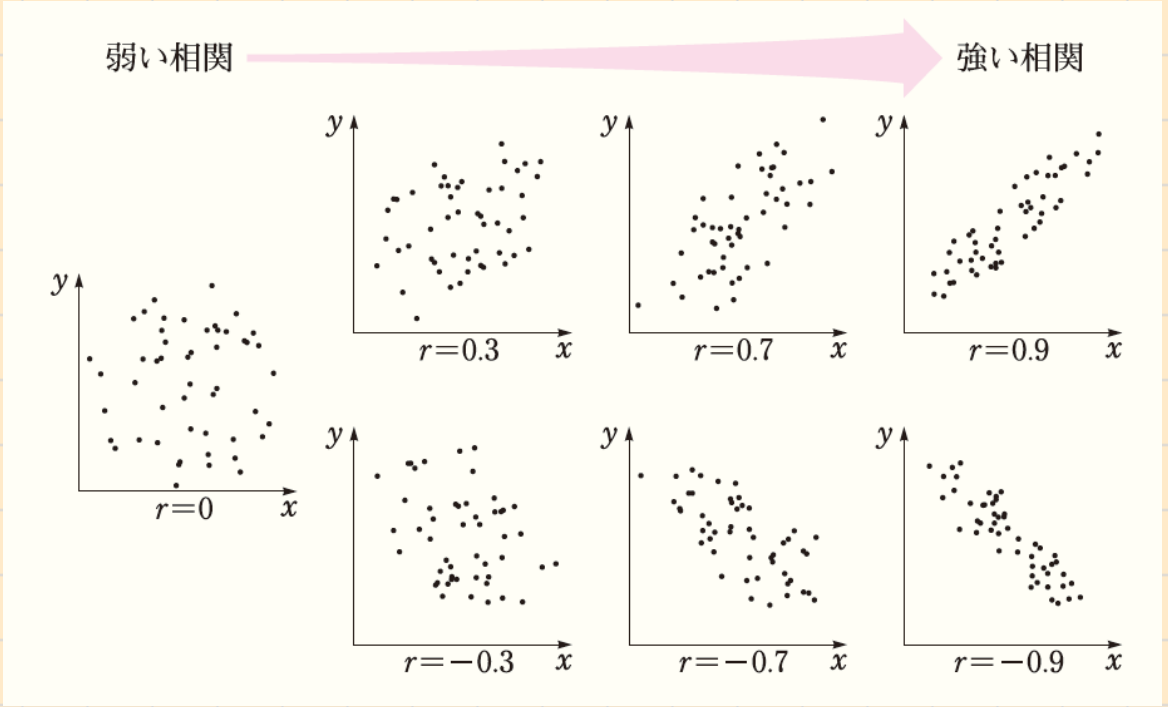
と標準化すると、分母の不偏分散も確率変数となるので、Z_n は正規分布にはならずに **t 分布**という確率分布になる。

t 分布は正規分布に比べるとやや裾の厚い分布となる。

この場合、仮説検定も、t 分布上で棄却域を考えることになる。これを t 検定といい、統計分野で最もよく用いられる検定である。

ただ、n が大きいとき（概ね30以上）のときには t 分布は正規分布とみなして扱う。

相関係数と検定



教科書には右のような説明があるが、
本当はサンプルサイズ n も重要である。

相関が無い２つの母集団から抽出した
n 組のサンプルによる相関係数を r とすると

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

は前ページで述べた t 分布になることが
知られている（自由度 $n-2$ のt 分布）。よって、t 分布を用いて仮説検定

帰無仮説 H_0：２つの変量には相関が無い

対立仮説 H_1：２つの変量には相関がある

を実行することができる。

例えば、 $n = 10$ で $r = 0.6$ となったとしても、有意水準５％では H_0 を棄却できない、すなわち、必ずしも相関があるとは言えない、ということになる。

最小二乗法

２つの変量の組 $(x_1, y_1), \cdots, (x_n, y_n)$ について、y を x で、出来るだけフィットするように表したいとする。

そこで、a, b を定数として $\hat{y}_k = ax_k + b$ とし、$\hat{y}_k$ と y_k の差が出来るだけ小さくなるような a, b を求めたい。

具体的には、

$$f(a, b) = \sum(\hat{y}_k - y_k)^2 = \sum(ax_k + b - y_k)^2$$

を最小にするような a, b を求める。そのため、f を a, b でそれぞれ偏微分し、

$$\frac{\partial f}{\partial a} = 2\sum(ax_k + b - y_k)x_k = 2a\sum x_k^2 + 2b\sum x_k - 2\sum x_k y_k = 0$$

$$\frac{\partial f}{\partial b} = 2\sum(ax_k + b - y_k) \quad = 2a\sum x_k + 2b\sum 1 - 2\sum y_k \qquad = 0$$

を解いてa, b を求める。

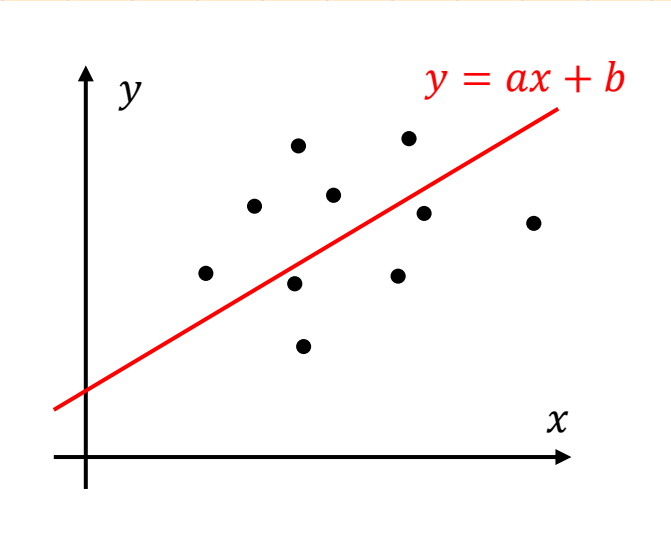
このような手順を最小二乗法という。

回帰直線

直線 $y = ax + b$ を回帰直線という。

しかし、得られた回帰直線に意味があるかどうかは
別途、仮説検定が必要である。

本来、 x と y が無関係（独立）であれば、$a = 0$
となるはずであるので、



帰無仮説 H_0 ： $a = 0$

対立仮説 H_1 ： $a \neq 0$

とし、帰無仮説が棄却されれば回帰直線には意味があると考える。

a はｔ分布に従うので、ｔ検定を用いることができる。

y 切片 b に関する検定も可能である。

何か統計量が得られたら、それらは全て仮説検定の対象となる。

重回帰

今度は３つの変量の組 $(x_1, y_1, z_1), \cdots, (x_n, y_n. z_n)$ について、z を x と y で、出来るだけフィットするように表したいとする。

そこで、 a, b, c を定数として $\hat{z}_k = ax_k + by_k + c$ とし、 $\hat{z}_k$ と z_k の差が出来るだけ小さくなるような a, b, c を求めたい。

具体的には、

$$f(a, b, c) = \sum(\hat{z}_k - z_k)^2 = \sum(ax_k + by_k + c - z_k)^2$$

を最小にするような a, b, c を、２変量の場合と同じように最小二乗法で求める。

ここで、 x, y を説明変数、 z を被説明変数という。

平面の式 $z = ax + by + c$ を求める手順を重回帰分析という。

説明変数が増えても重回帰分析は適用可能であるし、その重回帰の式に意味があるかどうかは仮説検定を行うことで判断することができる。

重回帰の応用

社会科学の金融や経営の分野では重回帰分析は日常的に用いられている。ただ、重回帰の式が有意であっても、**因果関係**はわからない。

少し変わったところでは、ゼミに、ファッションに興味のある学生がおり、卒業研究でファッションに関する重回帰分析を行ったことがあった。

説明変数としては、トップス（上半身に着る服、シャツ、ブラウス、Tシャツなど）とボトムス（下半身に着る服、ズボン、スカートなど）の組み合わせのフィット具合（アンケートにより0～100で数値化）を被説明変数とし、トップスとボトムスそれぞれの色（3原色RGB）、材質、装飾の多少、を説明変数として重回帰を行った。

トップスとボトムスの服の情報を入れると、良い着合わせかどうかを数値的に判断する仕組みを作りたい、というのが狙い。

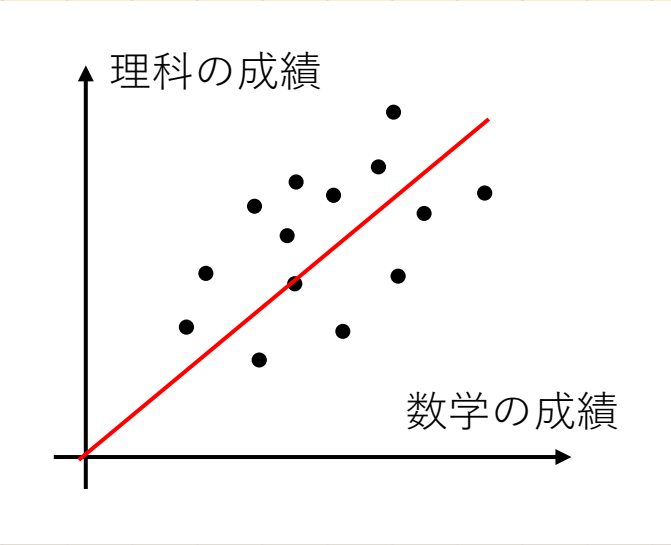
最終的に回帰式の仮説検定までは手が回らなかったが、学生のデータで少し試したところでは、有意な式を得るにはもう少しデータが必要な感じであった。

主成分分析

変量の組 $(x_{11}, x_{12}, \cdots, x_{1d}), \cdots, (x_{n1}, x_{n2}, \cdots, x_{nd})$
があったとき、この分布を最もよく表す指標

$$e_k = \alpha_1 x_{k1} + \alpha_2 x_{k2} + \cdots + \alpha_d x_{kd}$$

を作成したい（係数 $\alpha_1, \cdots, \alpha_d$ を求めたい）。



右は、数学に成績と理科の成績を組にしたものだが
（$d = 2, n = 14$）、数学の成績と理科の成績の
両方のデータを保持しなくても、

平均成績＝0.5×数学の成績＋0.5×理科の成績

で、データの分布をよく表せていることがわかる。

このように、元のデータの分布を最もよく表す量（主成分）を求めることを主成分分析という。主成分分析により、分布にあまり関係のないデータを削減し、分布に大きな影響を与えている成分のみを抽出することができる。

主成分分析２

簡単のため $d = 2$ とし、変量の組 $(x_1, y_1), \cdots, (x_n, y_n)$ を考える。ここで、あらかじめ変量変換を行って、x_k と y_k の平均は０になるようにしておくものとする。

ここで、

$$X = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ & \cdots \\ x_n & y_n \end{bmatrix}$$ とおくと、

$$X^T X = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{bmatrix} \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ & \cdots \\ x_n & y_n \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2^2 + \cdots + x_n^2 & x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \\ x_1 y_1 + x_2 y_2 + \cdots + x_n y_n & y_1^2 + y_2^2 + \cdots + y_n^2 \end{bmatrix}$$

より、$\Sigma = X^T X$ は共分散行列（の n 倍）となっている。

主成分分析３

主成分方向の単位ベクトルを $\vec{e} = \begin{pmatrix} e_x \\ e_y \end{pmatrix}$ とし、各変量 $\overrightarrow{a_k} = \begin{pmatrix} x_k \\ y_k \end{pmatrix}$ をこの単位ベクトルに正射影したものの分散を V とすると

$$nV = (\overrightarrow{a_1} \cdot \vec{e})^2 + (\overrightarrow{a_2} \cdot \vec{e})^2 + \cdots + (\overrightarrow{a_n} \cdot \vec{e})^2 = (X\vec{e})^T (X\vec{e}) = \vec{e}^T X^T X \vec{e} = \vec{e}^T \Sigma \vec{e}$$

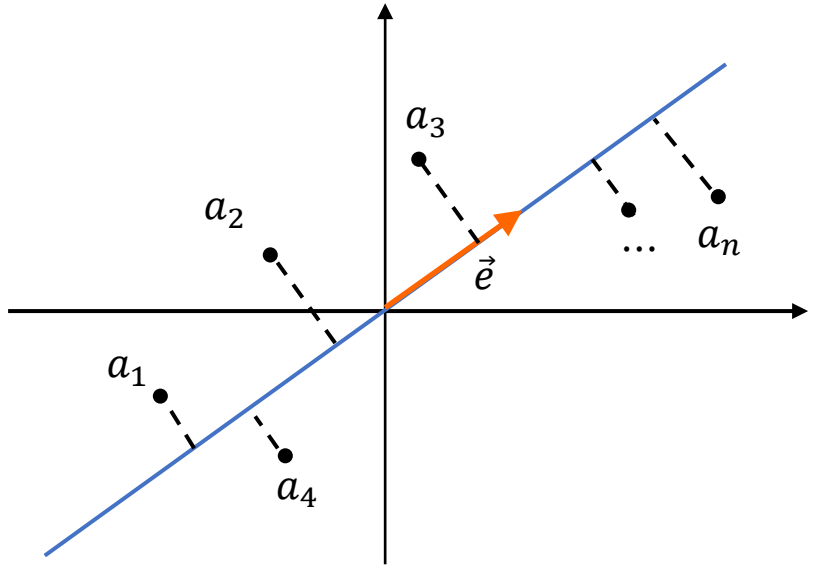
が成り立つ。このとき、 Σ の最大固有値を λ とすると、

$$|nV| = |\vec{e}^T|\,|\Sigma\vec{e}| \leq 1 \cdot \lambda = \lambda$$

ここで等号は、$\vec{e}$ が λ に対応する固有ベクトルのときに成り立つ。

つまり、共分散行列の最大固有値に対応する固有ベクトルが主成分となる。

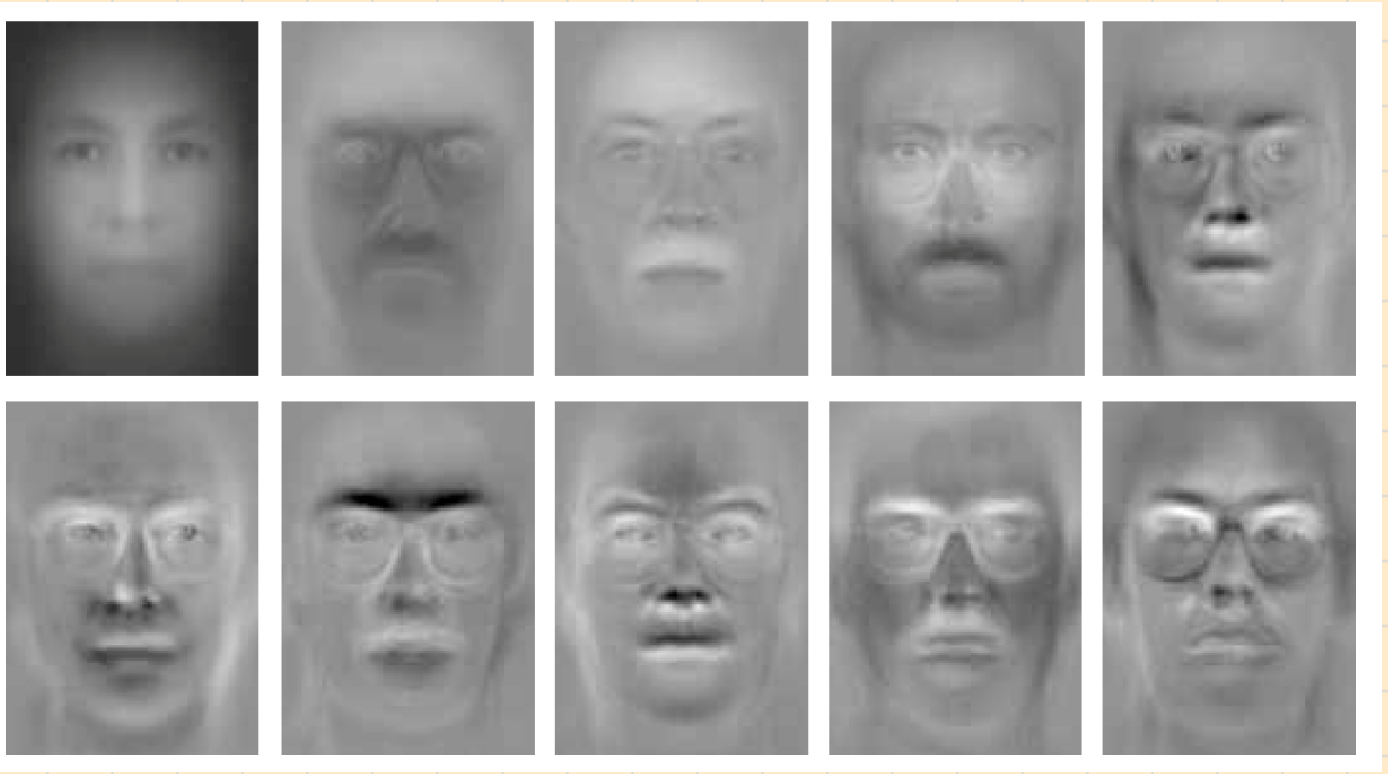
主成分が求まれば、その影響を取り除いたものから、第2、第3の主成分を順に取り出すことも可能。



主成分分析４

画像認識に主成分分析を利用した有名な研究。

顔の画像データから主成分（固有顔と名付けている）を順に取り出したもの。



M.A.Turk and A.P.Pentland,
``Face recognition using eigenfaces,''
Proc. of IEEE Conf. on Computer
Vision and Pattern Recognition, pp.586-591, 1991.

他にも、統計データを線形代数の枠組みで扱うことにより、様々な分析が可能になっている。もちろん、得られた結果に対しては検定が必須である。

今年の共通テストの問題から

なかなか面白い問題だと思う。

ただ、重さの分散を小さくするために、と書いているが、分散がどう小さくなるかには言及されていないのが惜しい。

とはいえ、共通テストのレベルは大きく超えてしまうが。

統計については、将来的には非常に重要だが、高校では過度にやる必要はないと個人的には考えている。

ただ、統計と数学の違いのような、感覚的なところは何となく理解できると良いかもしれない。

(2) (1)の確率変数Xにおいて，$m = 30.0$，$\sigma = 3.6$とした母集団から無作為にピーマンを1個ずつ抽出し，ピーマン2個を1組にしたものを袋に入れていく。このようにしてピーマン2個を1組にしたものを25袋作る。その際，1袋ずつの重さの分散を小さくするために，次の**ピーマン分類法**を考える。

ピーマン分類法

無作為に抽出したいくつかのピーマンについて，重さが30.0 g以下のときをSサイズ，30.0 gを超えるときはLサイズと分類する。そして，分類されたピーマンからSサイズとLサイズのピーマンを一つずつ選び，ピーマン2個を1組とした袋を作る。

(ii) **ピーマン分類法**で25袋作ることができる確率が0.95以上となるようなピーマンの個数を考えよう。