情報統計 第9-12回

2022年8月4日 神奈川工科大学



櫻井 望

国立遺伝学研究所 生命情報・DDBJセンター

昨日

- 確率分布
- t分布
- 検定の手順

- Excelで分布を描く方法
- t検定を手計算で行う

今日

● 分散分析(ANOVA)の概念を把握して、 手で計算できることを確認する

- 相関
- 多変量解析(主成分分析)のイメージ

情報統計第9回

2022年8月4日 神奈川工科大学



櫻井 望

国立遺伝学研究所 生命情報・DDBJセンター

学習の目標

- F検定 (等分散性の検定)
- 分布の仲間 カイ二乗分布、F分布
- 分散分析ANOVA(F分布を使う)

2群のt検定(独立2群)

等分散の場合

1群目:標本数 n1, 不変標本分散 s1, 標本平均 $\overline{x1}$

2群目:標本数 n2, 不変標本分散 s2, 標本平均 $\overline{x2}$

プール分散
$$s^2 = \frac{(n1-1)\times s1^2 + (n2-1)\times s2^2}{n1+n2-2}$$

検定統計量
$$t = \frac{\overline{x1} - \overline{x2}}{s\sqrt{\frac{1}{n1} + \frac{1}{n2}}}$$

自由度: n1 + n2 -2

帰無仮説: 2群の母集団の平均値は等しい

で、同様に検定できます。参考まで

2群の大検定(独立2群)

等分散が仮定できない場合 ウェルチの方法

1群目:標本数 n1, 不変標本分散 s1, 標本平均 $\overline{x1}$

2群目:標本数 n2, 不変標本分散 s2, 標本平均 $\overline{x2}$

検定統計量
$$t = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{s1^2}{n1} + \frac{s2^2}{n2}}}$$

(近似)自由度
$$v \approx \frac{\left(\frac{s1^2}{n1} + \frac{s2^2}{n2}\right)^2}{\frac{s1^4}{n1^2(n1-1)} + \frac{s2^4}{n2^2(n2-1)}}$$

帰無仮説: 2群の母集団の平均値は等しい

で、同様に検定できます。参考まで

F検定

等分散性の検定

1群目:標本数 n₁,不変標本分散 v²₁

2群目:標本数 n₂,不変標本分散 v²。

検定統計量: $F = \frac{v^2_a}{v^2_h}$ $\frac{\text{**V'}_a, \text{V'}_b \text{**}_b \text{**}_1, \text{V}_2 \text{**}_2 \text{**}_1 \text{**}_2 \text{**}_2 \text{**}_2 \text{***}_1 \text{***}_2 \text{***}_2 \text{***}_2 \text{***}_2 \text{***}_3 \text{***}_2 \text{****}_2 \text{***}_2 \text{***}_2 \text{***}_2 \text{***}_2 \text{***}_2 \text{***}_2 \text$ ※v²a, v²bは、v²1, v²2のいず

自由度: n₁ - 1, n₂ - 1

※分子と分母に対応させて、二 つ与える

帰無仮説: 2群の分散は等しい

F分布を扱うExcel関数: F.DIST, F.DIST.RTなど

例)身長データの場合

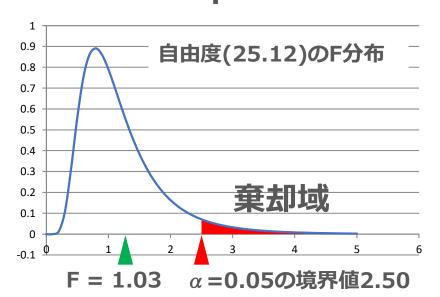
女性: $n_1 = 26$, $v_1^2 = 23.63$

男性: $n_2 = 13$, $v_2^2 = 23.02$

有意水準: 0.05とする

F = 23.63 (女性) / 23.02(男性) = 1.03

自由度(25, 12)のF分布から、F.DIST.RT関数を使って求めた右側確率pは、0.50



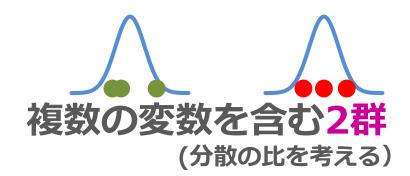
F値が棄却域の境界値より内側 $(1.03 < 2.50, p=0.50 > \alpha)$ なので、帰無仮説は棄却できず、 $(1.03 < 2.50, p=0.50 > \alpha)$ なので、帰無仮説は棄却できず、 $(1.03 < 2.50, p=0.50 > \alpha)$ なので、場無仮説は棄却できず、 $(1.03 < 2.50, p=0.50 > \alpha)$ なので、場所では、(1.03 < 2.50, p=0.50) なので、(1.03 < 2.50, p=0.50) なので、(1.03 < 2.50, p=0.50) なのでは、(1.03 < 2.50, p=0.50) なので、(1.03 < 2.50, p=0.50) ない。(1.03 < 2.50, p=0.50, p=0.50, p=0.50) ない。(1.03 < 2.50, p=0.50, p=0.50, p=0.50, p=0.50) ない。(1.03 < 2.50, p=0.50, p=0

留意すべきこと

F検定で「分散に差がある」という結論を得たのち、2群の平均値に差があるかどうかをt検定すると、「検定の多重性」の問題にあたってしまう。

近年では、等分散かどうかに関係なく適用できるウェルチの検定を最初から行うことが望ましいという考えも出てきている。

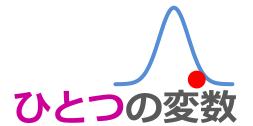
F分布



カイ二乗分布

複数の変数(分散を考える)

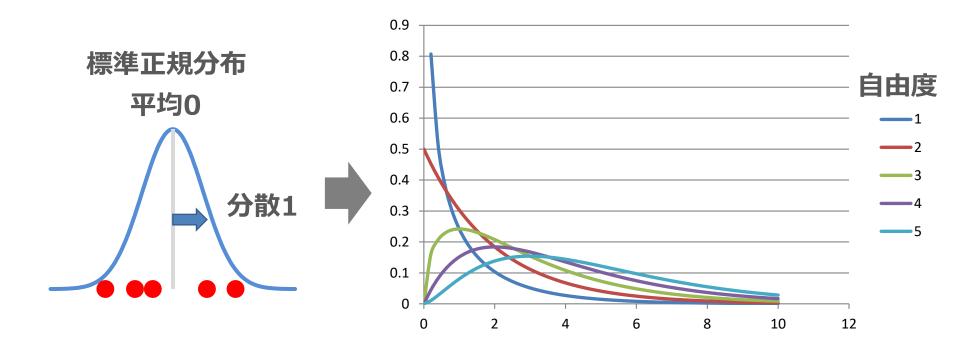
標準正規分布



by 櫻井

力 イ 二 乗 分 布

標準正規分布に従った独立した変数がいくつかあるとき、その二乗和が従う分布



カイ二乗分布の性質

正規分布N(μ, σ²)に従ったk個の変数x_iについて、偏差(平均からの差)の平方和と分散の比は、自由度kのカイ二乗分布に従う

$$\chi^{2} = \frac{\sum_{i=1}^{k} (x_{i} - \mu)^{2}}{\sigma^{2}} = \sum_{i=1}^{k} \left(\frac{x_{i} - \mu}{\sigma}\right)^{2}$$

カイ二乗検定

	ビール 好き	ビールあんまり
男性	23	12
女性	7	8

二つのカテゴリに関連があるかを調べたい

帰無仮説:

二つのカテゴリは独立である(関連がない)

有意水準:0.05

カイ二乗検定の手順

(1) 観測データから、カテゴリーごとに割合を出す

	ビール 好き	ビール あんまり	合計
男性	69	36	105 70%
女性	21	24	45 30%
合計	90 60%	60 40%	150 100%

(2) 割合から、カテゴリーが独立な場合の度数(期待度数)を出す

	ビール 好き	ビール あんまり	合計
男性	63	42	105 70%
女性	27	18	45 30%
合計	90 60%	60 40%	150 100%

カイ二乗検定の手順

(3)観測度数と期待度数の差を出す

	ビール 好き	ビール あんまり
男性	6	-6
女性	-6	6

(4) その二乗を出す

	ビール 好き	ビール あんまり
男性	36	36
女性	36	36

(5)期待度数で割る

	ビール 好き	ビール あんまり
男性	36/63 =0.57	36/42 =0.86
女性	36/27 =1.33	36/18 =2

(6) その和を求める

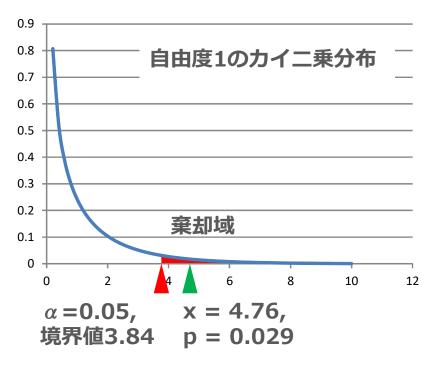
$$x = 0.57 + 0.86 + 1.33 + 2 = 4.76$$

このように求めた値xは、カイ二乗分布に近似できる。 自由度は、各カテゴリ(性別、ビールの好み)の要素数をそれぞれ n_1 , n_2 とすると、 $(n_1$ -1)* $(n_2$ -1)。 この例の場合では、(2-1)*(2-1) = 1

カイ二乗検定の手順

(フ)結論

xの値が棄却域の境界値の外側(3.84 < 4.76, p=0.029 < α)なので、帰無仮説は棄却され、「二つのカテゴリは独立ではない」と判断された。



よって、この母集団においては、 「性別とビールの好みとの間に何 かしらの関連性がある」と結論づ けられた。

カイ二乗分布を扱うExcelの関数: CHISQ.DIST, CHISQ.DIST.RT, CHISQ.INV.RTなど

カイ二乗検定の留意点

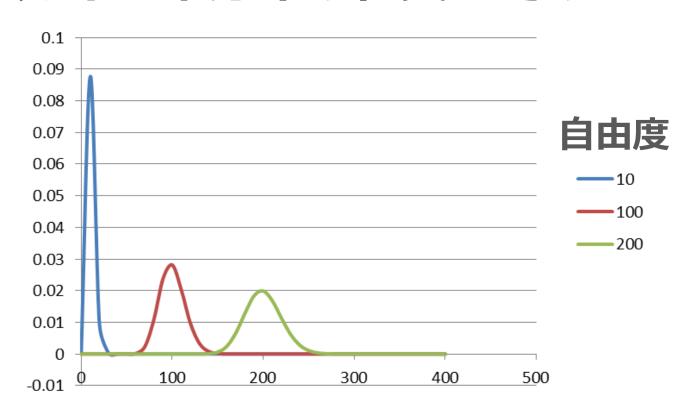
観測数が少ないとカイ二乗分布への近似ができないので、 その場合はフィッシャーの正確確率検定を行う。

目安:

期待度数が5未満のセルが、全セルの20%以上で存在する場合、近似が不正確と考えられる (コクラン・ルール)

期待度数が1未満のセルがあってはならない

カイ二乗分布の性質 その2



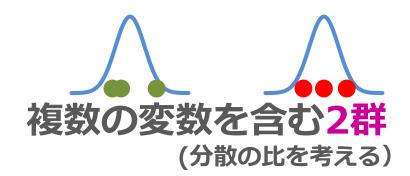
自由度kが大きくなると、

平均值:k

分散:2k

の正規分布に近づいてゆく

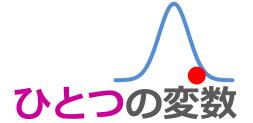
F分布



カイ二乗分布

複数の変数(分散を考える)

標準正規分布



by 櫻井

F分布とカイ二乗分布の関係

自由度k₁のカイ二乗分布 χ^2 ₁ 自由度k₂のカイ二乗分布 χ^2 ₂

があるとき、次の値Fは、自由度(k₁, k₂)のF分 布に従う

$$F = \frac{\chi^2_1/k_1}{\chi^2_2/k_2}$$



F分布の活用

正規分布Ν(μ1,σ21)に従った母集団から得た標本、 標本数:n₁、不偏標本分散:v²₁

正規分布Ν(μ₂,σ²₂)に従った母集団から得た標本、 標本数: n_2 、不偏標本分散: v^2_2

があるとき、

$$F = \frac{\chi^2_1/k_1}{\chi^2_2/k_2} = \frac{v^2_1/\sigma^2_1}{v^2_2/\sigma^2_2}$$

二つの母集団の分散σ²₁とσ²₂が等しいと仮定できる場合は、

$$F = \frac{v^2}{v^2}$$



 $F = \frac{v^2_1}{v^2_2}$ これをF検定で利用している!

F分布の活用

カイ二乗分布の性質

$$\chi^2 = rac{\sum_{i=1}^k (x_i - \mu)^2}{\sigma^2}$$
 自由度k

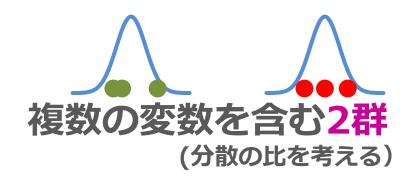
この式を変形すると、

不偏標本分散v2になっている!

$$\chi^{2} = \frac{\mathbf{k} \times \frac{\sum_{i=1}^{k} (x_{i} - \mu)^{2}}{\mathbf{k}}}{\sigma^{2}} = \frac{\mathbf{k} \times v^{2}}{\sigma^{2}}$$

したがって、
$$\frac{\chi^2}{k} = \frac{k imes v^2}{\sigma^2} imes rac{1}{k} = rac{v^2}{\sigma^2}$$

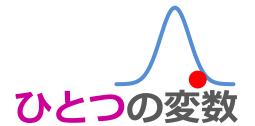
F分布



カイ二乗分布

複数の変数(分散を考える)

標準正規分布



by 櫻井

分散分析

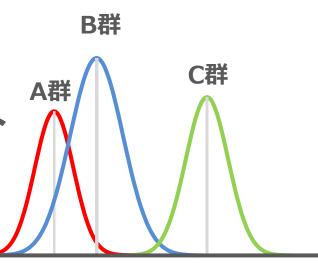
Analysis of Variance

ANOVA

- ✓ 3つ以上の群があるとき、
- ✓ 群の母平均に差があるかど うかを、
- ✓ 分散 (F分布) を使って、

検定する方法

例)1組、2組、3組で、テスト の平均点に差があるか?



帰無仮説:

A群、B群、C群の母平均は等しい

対立仮説:

A群、B群、C群の母平均の中に、 異なる値がある

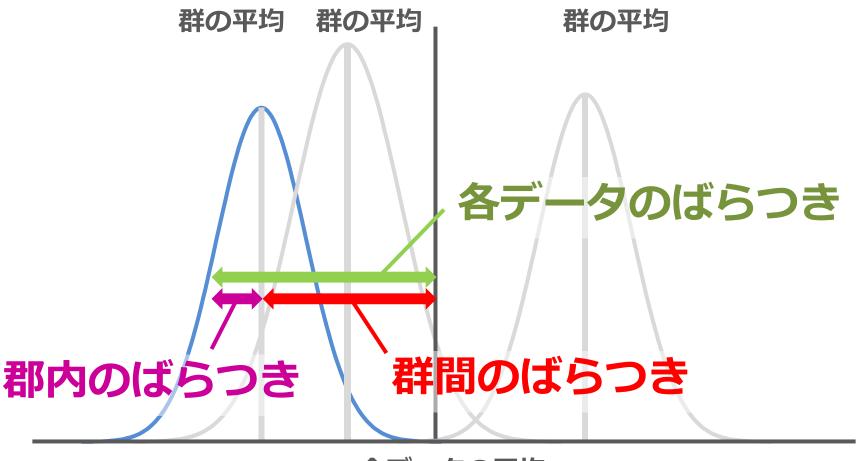


どれが異なっているかまではわからない!

帰無仮説が棄却されたときは、解釈に注意が必要

分散分析のイメージ

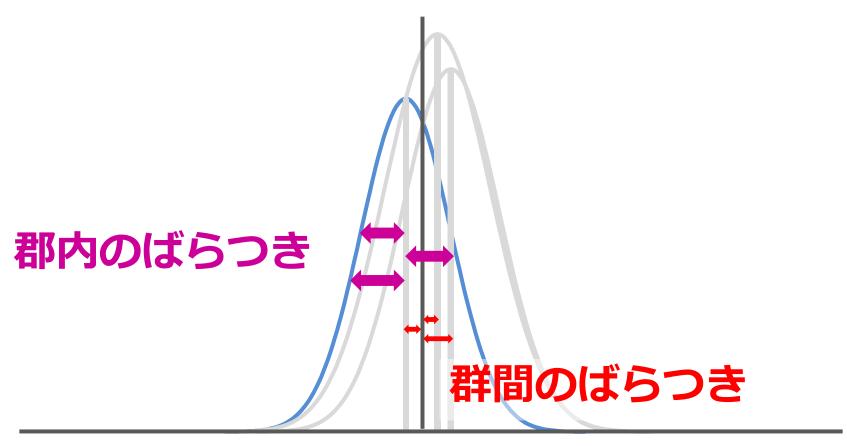
データのばらつきを、群間のばらつきと、偶然により起こる郡内のばらつきに分けて考える



全データの平均

分散分析のイメージ

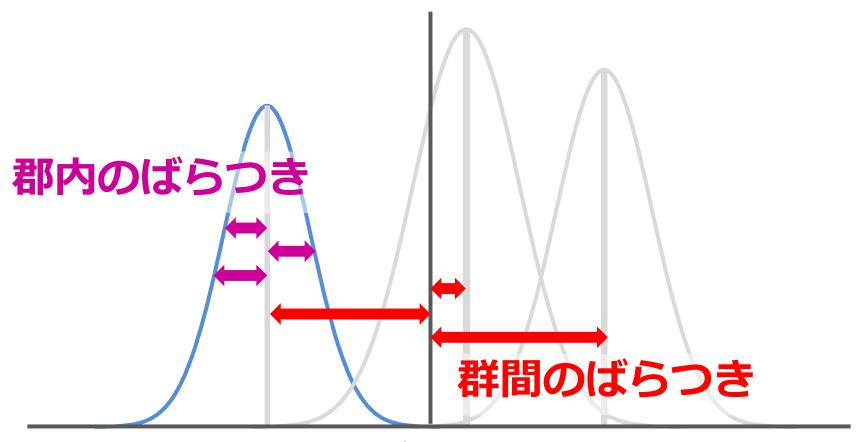
群の平均に差がなければ、
郡内のばらつき > 群間のばらつき



全データの平均

分散分析のイメージ

群の平均に差があるほど、 郡内のばらつき 〈 群間のばらつき



全データの平均

分散分析の手順

分散分析表を穴埋めしてゆく

要因	平方和 S	自由度 df	不偏標本分散 V ²	F値
群間 (因子)	S(群)	df(群) =群の数-1	V ² (群) =S(群)/df(群)	V ² (群)/V ² (残 差)
群内 (残差)	S(残差)	df(残差) =全データ数-群 の数	V ² (残差) =S(残差)/df(残差)	
全体	S(全体)	df(全体)		

分散分析の手順

例)A~Dの異なる生育環境で育てた植物の、ある成分の含量

A群	341	347	328	329	352
B群	305	317	342	322	319
C群	342	313	350	323	
D群	331	327	303	314	

エクセルファイル:200918_anova.xlsx

以下の基本情報を計算する

- ①群ごとのデータ数
- ②全データの個数
- ③群の平均値
- 4全データの平均値

以下の差(ずれ)を計算する

- ⑤全データについて、全体の平均からの差
- ⑥各群の平均について、全体の平均からの差
- ⑦郡内の各データについて、群平均からの差

差(ずれ)の二乗を計算する

- ⑧全データについて、全体の平均からの差の二乗
- ⑨各群の平均について、全体の平均からの差の二乗 群のデータ数を乗じる
- ⑩郡内の各データについて、群平均からの差の二乗

二乗和を計算する

- ⑪全データについての全体の平均からの差の二乗和
- ②各群の平均についての全体の平均からの差の二乗和
- 13郡内の各データについての群平均からの差の二乗和

分散分析表を埋める

- 14二乗和
 - ① = ① + ②となっているはず
- 15自由度

全体: ②全データ数 - 1

群間:群の個数 - 1

群内:全体の自由度 - 群間の自由度

- 16不偏標本分散(群間、群内について) 二乗和 / 自由度
- ①F値 不偏標本分散の比(群間/群内)

用語

要因:

データに影響を与えるもの

因子:

要因の中で特に母平均の差に 影響すると思われため、解析 の対象とするもの

残差:

偶然によって生じたばらつき

$p値、<math>\alpha$ のF境界値を計算する

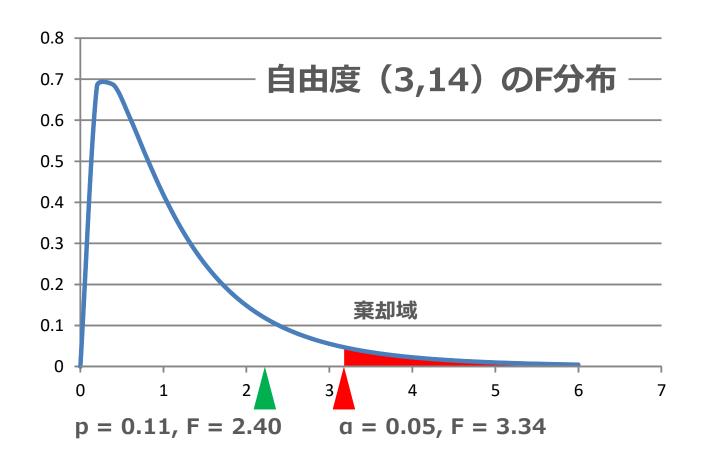
- 18 ①で求めたF値と自由度から、F.DIST.RT関数を使って、p値を計算する
- 20 F.DIST関数を用いて当該自由度のF分布を描く

p値の大きさ、αに対応する境界値の大きさなどから、 検定統計量が棄却域に入ったかどうかを判断する

結論づけをする

結論

p値は0.11となり、有意水準0.05で帰無仮説は棄却されなかった。したがって、「A~Dの生育方法によって成分の平均値に差があるとは言えない」と結論付けられた。



分散分析の種類

一元配置の分散分析 one-way ANOVA



一つの因子からなるデータを分析する方法

二元配置の分散分析 two-way ANOVA

二つの因子からなるデータを分析する方法。例)薬剤の種類と投与量など。二つの要因が組み合わさる交互作用(相乗効果)を確認することもできる

多元配置の分散分析

情報統計第10回

2022年8月4日 神奈川工科大学



櫻井 望

国立遺伝学研究所 生命情報・DDBJセンター

相関

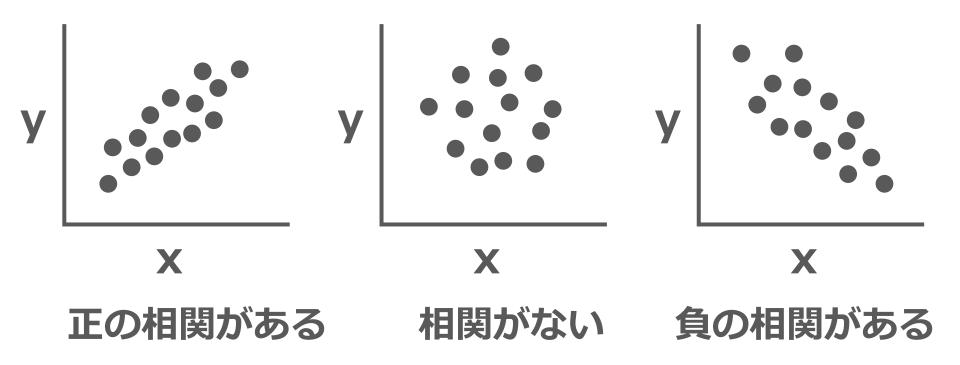
学習目標

相関のあるなしを評価できるように
なる

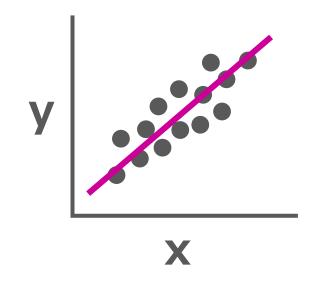
相関関係と因果関係の違いが分かる

散布図

二つの変数の間の関係性を見える化する手法



散布図の回帰曲線



エクセルのグラフ上でプ ロットを右クリックし、挿 入できる

相関係数

- 二つの変数の間の関係性の強さを数値化したもの
- -1~1の間の値をとる

0.7~1.0:強い正の相関 -1.0~-0.7:強い負の相関

0.4~0.7:中程度の正の相関 -0.7~-0.4:中程度の負の相関

0.2~0.4:弱い正の相関 -0.4~-0.2:弱い負の相関

-0.2~0.2:相関がない

● ExcelではPEARSON関数で計算できる

注意点

回帰曲線のR²値は、相関係数ではありません。

R²値は、回帰曲線への当てはまり度を示すもので、「決定係数」と呼ばれます。

Excelで、原点を通らない直線近似をした場合は、ピアソン相関係数の二乗に当たります。このため、相関係数が-1~1の値を取るのに対し、R²値は0~1の値を取ります。負の相関であっても、R²が正の値を取っているのはこのためです。

生や負の相関のあるなしや、強弱を考える場合は、必ず相関係数をもとに考えましょう。

相関関係を見るの人

都道府県別の統計

https://todo-ran.com/

携帯版 | スマホ版 | English

都道府県別統計とランキングで見る県民性 [とどラン]

都道府県別統計とランキングで見る県民性

https://todo-ran.com/

国土・インフラ トップ 社会・政治 産業・経済 文化・くらし・健康 娯楽・スポーツ 店舗分布 その他 サイトについて 作者について 引用・転載について 統計八百屋 リクエスト ① X 栄養士、管理栄養士募集 《完全無料》栄養士複数在籍、未経験歓迎など栄養士の非 公開求人をご紹介 中

都道府県別統計を比較 した都道府県ランキン グ。**1339** ランキング 掲載中

odomon@gmail.c

🗐 当サイト一番人気

都道府県 ベスト&ワースト

各都道府県の1位と47 位だけを一覧表にまと めました。県民性が一 目で分かります。

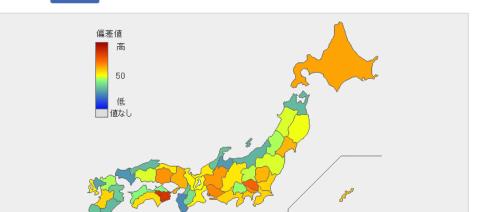
都道府県比較

東京vs大阪、埼玉vs千 葉vs神奈川など任意の 都道府県の似たとこ、 似ていないところを一 トップ

最新ランキング

2019年参議院比例代表: NHKから国民を守る党得票率 [2019年 第一位 徳島県]







📗 記事を探す

▶ 検索から探す(googleサイト内検索)

Google カスタム検索 サイト内検索

▶ カテゴリから探す

政治・経済などカテゴリ別全記事表示

▶ 新着から探す 新しい順に全記事表示

データを集めてみる

例)

神奈川県の高いランクのうち、

「しゅうまい消費量」と

「最低賃金」や「農業就業人口」との相関

- サイトでデータをコピー
- エクセルに貼り付け
- エクセルで加工(県の列で並び替え)
- 散布図を描く
- PEARSON関数で相関係数を計算する

相関係数を手で計算する

ピアソンの積率相関係数

$$r = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x}) (y_i - \overline{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \overline{y})^2}}$$

sxy: xとyの共分散

sx: xの標準偏差

sy: yの標準偏差

n: xとyのペアの数

無相関の検定

帰無仮説:

母集団の相関係数は0(無相関)である

分布: t分布

検定統計量: $t = \frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$

自由度: n-2

※|r|はrの絶対値エクセルではABS関数で計算できる

そのほかの相関係数

- ●スピアマンの順位相関係数
- ●コサイン相関係数

相関と因果

相関関係:

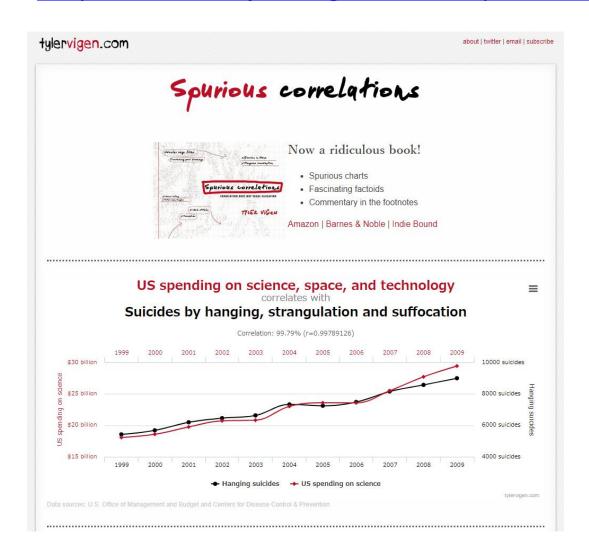
二つの事柄に関連性がある

因果関係:

二つの事柄が、原因と結果の関係である

疑似相関

https://www.tylervigen.com/spurious-correlations

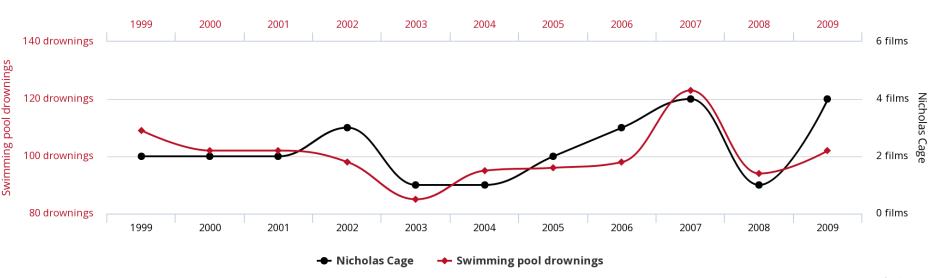


ニコラス・ケイジの映画出演本数と、 プールでおぼれた人の数に、 高い相関がある?

Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in



tylervigen.com



統計学と経済学の最新の知見を凝縮

中室牧子, 津川友介著、 ダイヤモンド社2017年

情報統計第11回

2022年8月4日 神奈川工科大学



櫻井 望

国立遺伝学研究所 生命情報・DDBJセンター

多変量解析

学習目標

主成分分析について

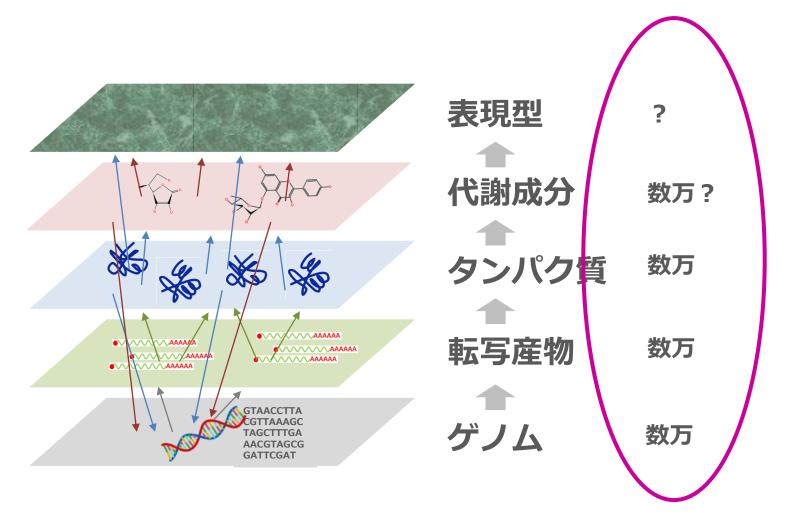
- 概念を理解する
- 結果の解釈の仕方を理解する

多変量データの例

- 大規模な疫学研究データ
- 生物等のオミクスデータ

など

生物の遺伝子情報の流れとオミクス



オミクス

それぞれの要素を一斉に検出 しようとする技術・学問

多変量解析の目的

- データを要約して解釈しやすくする
- データに含まれる潜在的な因子を見 つける
- 状況を判別したり、分類したりする
- 状況を予測する

さまざまな多変量解析

- 似ているものをグルーピングする クラスター解析
- データを要約する主成分分析
- 判別、分類、予測判別分析、PLS、PLS-DA、重回帰分析

主成分分析

主成分分析で扱うデータ

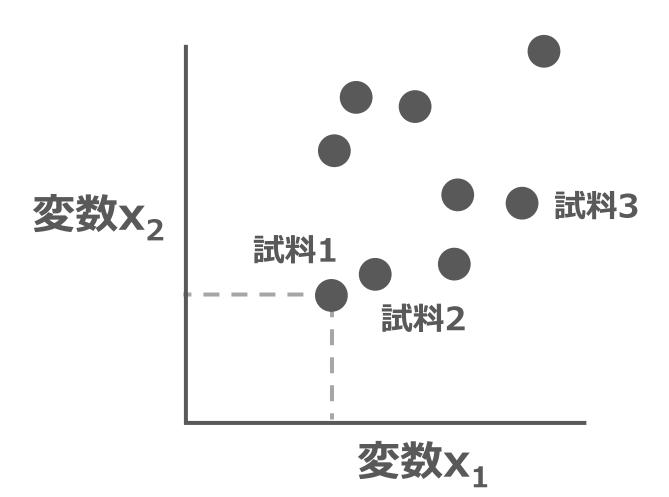
,組織ごとの生体試料など

		対象				
		1	2	3	3	n
変数	X_{I}	X_{II}	X 21	X 31		X_{nI}
	X_2	X 12	X 22	X 32		X n2
	$X_{\mathcal{J}}$	X 13	X 23	X 33		X_{B3}
	X_{m}	X_{Im}	X_{2m}	X_{3m}		X _{nm}

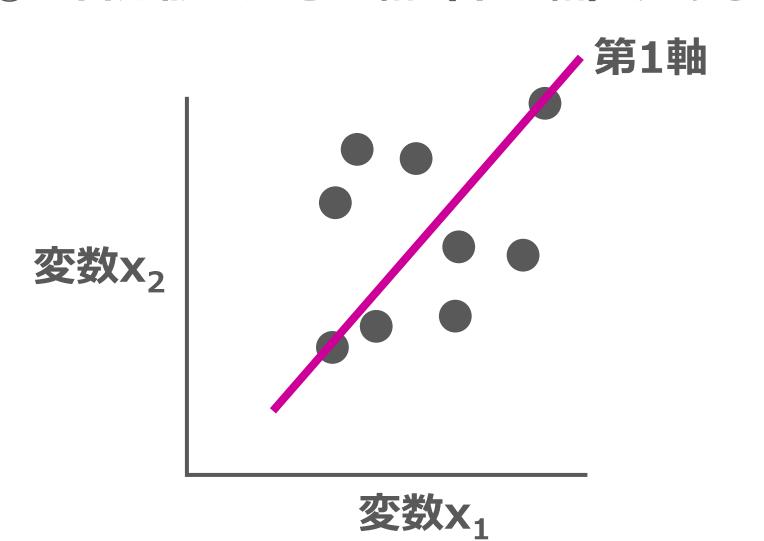
遺伝子など 説明変数, 観測変数

遺伝子発現量など

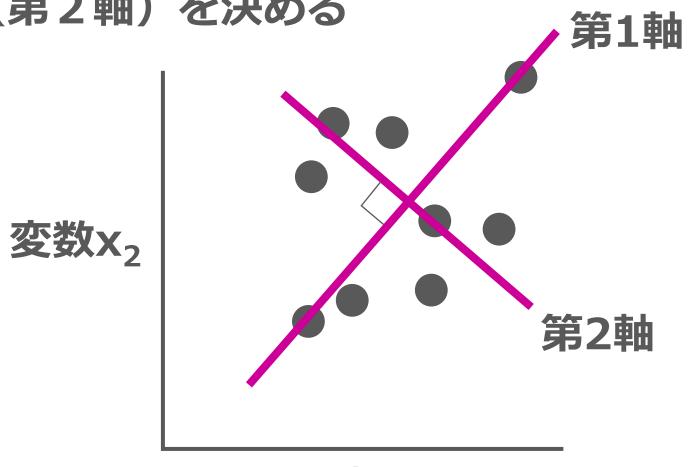
①例えば変数が2個しかないとき、2次元の散布図に、試料ごとに変数をプロットできる



②一番分散の大きい軸(第1軸)決める

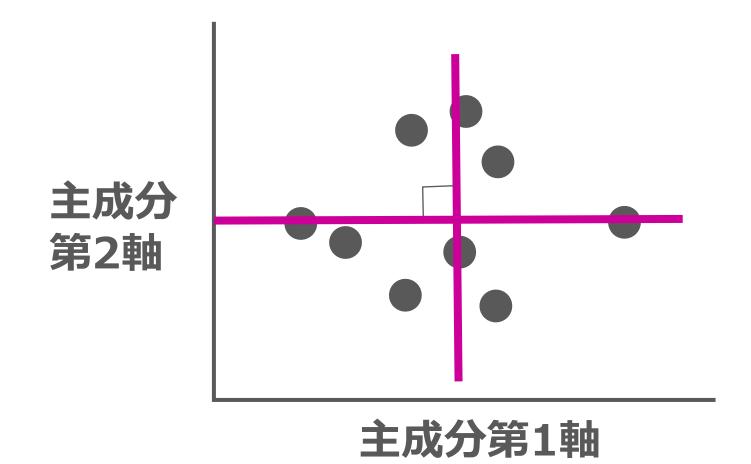


③第1軸に直角に交わり、次に分散が大きい軸(第2軸)を決める 2511

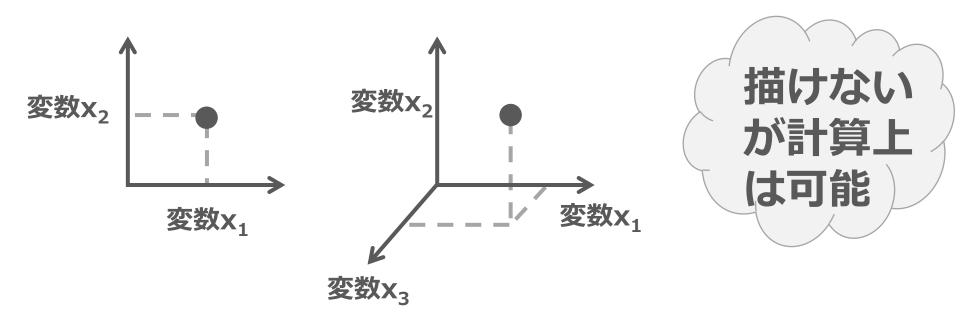


変数x₁

④第1軸がx軸、第2軸がy軸になるように、図を回転させた新たな図を作る

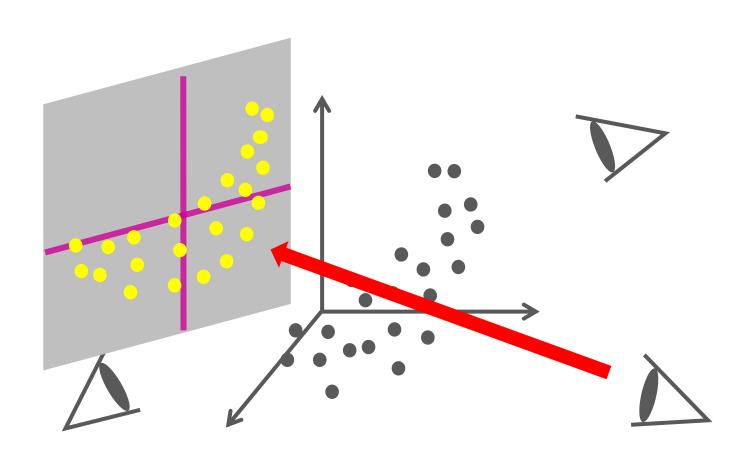


m個の変数の値をm次元の図にプロットし、 同様の計算を行うことが可能



変数2個 2次元 変数3個 3次元 変数m個 m次元

試料間の違い(特徴)が一番はっきりと見える方向から見た図が描ける



スコアプロット

主成分軸に各試料を投影しなおした図

軸に示した%は寄与率と呼び、全体の分散のうち各主成分軸が説明する分散の比率を表す。第1主成分の寄与率が最も大きい。 .

第2 主成分 (**%)

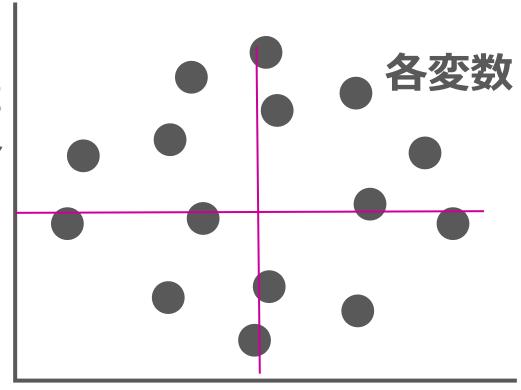
第1主成分(**%)

ローディングプロット

ローディングは、因子負荷量とも呼ばれ、各試料の主成分ス コアと、変数の間の相関係数に相当する。

(厳密には、数値の前処理の条件などいくつか制約がある)

第2主成分に 対する因子負 荷量

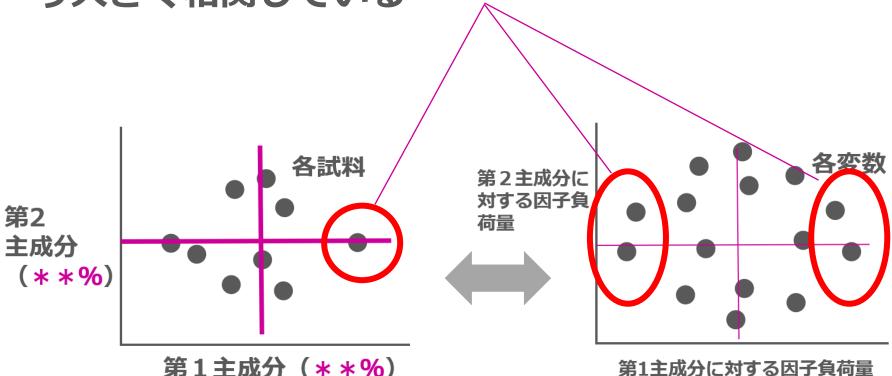


第1主成分に対する因子負荷量

こつの図をセットで見る

この試料と他の試料との違いは、これらの変数がよ

り大きく相関している



第1主成分(**%)

スコアプロット

ローディングプロット

そのほかの多質解析

さまざまな多変量解析

- 似ているものをグルーピングする クラスター解析
- データを要約する主成分分析
- 判別、分類、予測判別分析、PLS、PLS-DA、重回帰分析

PLS Partial Least Squares 部分最小二乗

PLS-DA

Partial Least Squares-Discriminant Analysis

部分最小二乗-判別分析

PLS、PLS-DAで扱うデータ

目的変数が存在する

組織ごとの生体試料など

説明変数との関連を調べたい試料の分類や、試料の特徴量など例)別途測定した、生理活性データなど

目的変数

		11 <i>5</i> 2				
		対象				
		1		2	3	n
変数	U_I	<i>y</i> 11	<i>Y 21</i>	<i>y</i> 31		y_{n1}
	<i>y</i> ₂	<i>y</i> ₁₂	<i>Y 22</i>	<i>y</i> 32		<i>Y</i> _{n2}
	•••					
	y_p	y_{Ip}	y_{2p}	y_{3p}		y_{np}
変数	X_I	X_{II}	X 21	X_{3I}		X_{nI}
	$X_{\mathcal{Z}}$	X 12	X 22	X 32		X n2
	$X_{\mathcal{J}}$	X 13	X 23	X 33		$X_{n\beta}$
	•••					
	X_{m}	X_{Im}	X_{2m}	X_{3m}		X _{nm}

遺伝子など 説明変数, 観測変数 遺伝子発現量など

PLS、PLS-DAで得られる結果

- PCAと類似したスコアプロットとローディングプロットが得られる
- 目的変数 (y) を説明変数 (x) で説明するためのモデルが構築される
- 目的変数を説明する変数重要度(VIP)が 計算される

情報統計第12回

2022年8月4日 神奈川工科大学



櫻井 望

国立遺伝学研究所 生命情報・DDBJセンター

おさらい

やったこと

- 統計的手法
- 記述統計
 - ✓平均値等の計算
 - ✓ 相関係数、回帰式
- 推測統計
 - ✓ 推定、仮説検定
- 多変量解析

- エクセル関数
- プログラミング
- Python

統計つて?

集団の状況を 数値で表したもの



目的:集団の〇〇を知りたい

統計学

- ・データを集める
- 解析する
- ・解釈する



ための方法論

結果:集団の〇〇がわかった!



結論を言う

統計的結論から、設定した目的に対する結論を導くことが最も重要。

発表会のテンプルート

表紙 1枚

- ・タイトル
- ●名前
- ・報告日など

背景と目的 1~枚

- 何に疑問を持ち、どんな目 的のためにこの課題を行っ たか?
- その疑問に至った背景

方法のページ 1~枚

どんなデータ、どんな統計 的手法を使って実施したか。

だれもが追試、検証できるように

結果のページ 1~枚

- どんな結果が得られたか
- そこから言えることは何か

結果に基づいて得られた情報 について述べる

考察のページ 1~枚

結果を総合して、目的に対してどんな結論が得られたか

最初に掲げた疑問に対する答えや、得られた結果の価値について述べる

(将来展望のページ 1~枚) もしあれば

- 今後こんなデータを集めれば…
- 今後こんな統計的手法を適用すれば…

もっとこんなことがわかるだろう、など

未来に対する夢を述べる

よいスライドの作り方



田中佐代子著、 講談社2013年

課題準備