

多変量解析 <初歩の初歩>に入門(1)

統計のきほんのきほん

なかもとけいこ
kenakamoto@nifty.com

<初歩の初歩>入門の目的と予定

- 目的
 - とりあえず数値・数式に脅えないようになる。
 - 自分がどんなデータに何をしているか分かるようになる。
 - 意外と使えるということを実感する。
- 予定
 - 分布と記述統計量(代表値, データの散らばり), 共分散と相関係数
 - 重回帰分析
 - 主成分分析と因子分析
 - クラスタ分析,
 - 判別分析とか, 気が向いたら数量化とか (この辺は未定)
- それ以上は自分で勉強してください。

まず心構え.

- 質問を歓迎します。分かっていないふりをしてもしません。
 - 今さら見栄をはって、どーする。
- 理解する努力は要求します。
 - 「イイお話」をするわけではないので、聞いているだけで賢く(なった気にはなれません。
 - 数学ちゃんとやってないから(教えてもらってないから), 分からなくても仕方ない, は, ただの言い訳。
- 最終的には自分のデータで色々と解析を試す必要があります。
 - 聞いているだけではホントには身に付かないでしょう。
 - データ自体に対する直感的な理解と統計解析は相互作用します。なので, 「自分のデータ」を使うことが重要です。

今回の目的

- 全体のイントロダクション
- 測定とはなにかをおぼろげに理解する
- 代表値と散布度を理解する
- 共分散と相関係数を理解する

参考書籍の一例

- 平均・順位・偏差値 吉村功 (岩波ジュニア新書)
 - 統計解析のはなし
 - 多変量解析の話 大村平 (日科技連)
 - 多変量解析のはなし 有馬哲夫・石村貞夫 (東京図書)
 - 複雑さに挑む科学 柳井晴夫・岩坪秀一 (ブルーバックス)
 - 初歩からの多変量統計 三土修平(日本評論社)
 - 誰も教えてくれなかった因子分析--数式が絶対に出てこない因子分析入門 松尾 太加志・中村知靖(北大路書房)
-
- 統計解析のための線形代数
 - 線形代数30講
 - 高校の時の教科書!

注意!

- 統計は魔法ではありません。
 - 生データに含まれていない情報が出てくるはずはない。
 - 色々こった(高度な)統計解析をする前に,
 - データ収集の手続き(実験, 調査のロジック, コーディングの妥当性, 信頼性),
 - 記述統計量の算出,
 - 分布や散布図の傾向
 - などを見極めておきましょう。不適切な解析をしても, 意味のない結果しか得られません。
 - もし, 解析方法を変えるたびに大きく結論が変わるなら, データに安定性が無いか, 解析の前提を見極められていない可能性が高いです。 **たまたま良い結果が出てても, たぶん幻です。**

統計は何のために？

- 記述統計
 - データの傾向を要約して上手く表す。
- 推測統計
 - 手元にあるデータ(=母集団からのサンプル)から、結論を一般化して良いかどうかを見極める。
- 多変量解析
 - 複数の変数(変量)の関係を整理して分かりやすくする。
 - たくさんの変数をより少ない変数にまとめる。
 - ある変数を他の変数で予測・説明する、等

多変量解析の前に・・・

- そもそもなぜ統計解析をする必要があるのでしょうか？
- たとえば、下のようなデータがあったとしましょう。
 - 高校生40人の性別・身長・体重

性別	身長	体重	足のサイズ
1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16
17	18	19	20
21	22	23	24
25	26	27	28
29	30	31	32
33	34	35	36
37	38	39	40

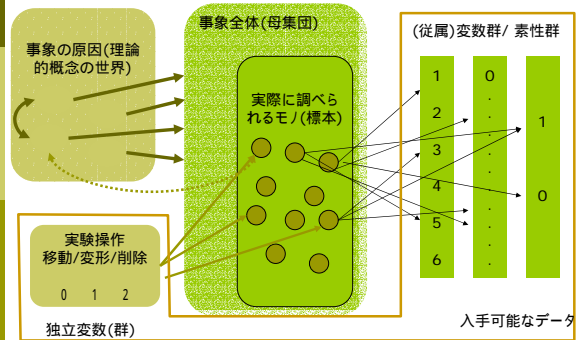
- だいたいどんな感じのデータなの？
- 高校生で男女に身長の高さの差があるの？
- 身長と体重と足のサイズの関係は？

そもそも数値は何を反映してるの？

- データは散らばる/散らばってこそ価値がある。
 - 比較がデータ解析の基本
 - 相関係数、連関係数(後述)も散らばりがなければ計算できない。
- 「散らばり」を生み出すのは何か？
 - 測定対象それ自身の性質が生み出す散らばり
 - 測定誤差によって生じる散らばり

↑
統計解析によってコントロールが期待される

数値の割り当て: 実験, 調査, コーディング



よい測定とは

- 信頼性 reliability が高い
 - 正確で精度の高い測定
 - 何度やっても同じ結果がでる。
 - 誰がやっても同じ結果が出る。
 - (詳細な目盛り幅で測定できる)
- 妥当性 validity が高い
 - 本質的な特徴を捉えた測定
 - ターゲットになっている構成概念を適切に捉えられている。
 - 予測が目的なら、その予測に役に立つ。
 - そもそもターゲットになっている構成概念に理論的意義がある。

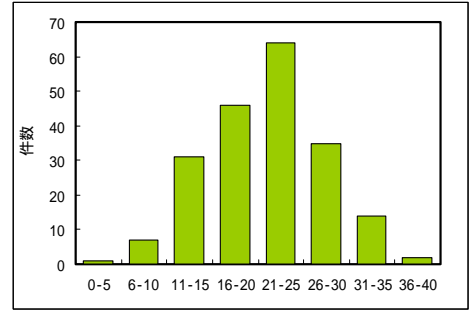
測定と数量化

- 竹谷(1991)より引用
 - ...直接間接を問わず、対象からデータを得る過程を数量化というわけである。換言すると数量化とは対象のある事象の側面における客観的に識別可能な特性に、特定の数値(または記号)的要素を対応させることである、といえる。この対応づけられた要素の集合がデータである。すなわち、データは、対象のある事象をある数量的要素で代表させているもので、事象そのものでもないし、まして対象そのものでもない。データへの過信や盲信は、こうしたデータのもつ意味を誤解することに起因していることが多い。...ここにデータのもつ限界があることを見極める必要があるし、またデータを解析したり解釈したりするとき忘れてはならない点である。データだけが一人歩きするのは、現に慎むべきだからである。
 - それでは、なぜ数量化したデータを収集しようとするのだろうか。それは、数量化されたデータがそうでないものに比べ、主観性を排除しているからである。すなわちデータの数量が個人の主観の相違を超越してある普遍性ある特性を表しているからである。大多数の人に普遍的に許容されるデータにすることによって、分析を狩野にするわけである。したがって、データ処理や分析の結果までも、ある普遍性・客観性をもった情報として意味をもつことになる。

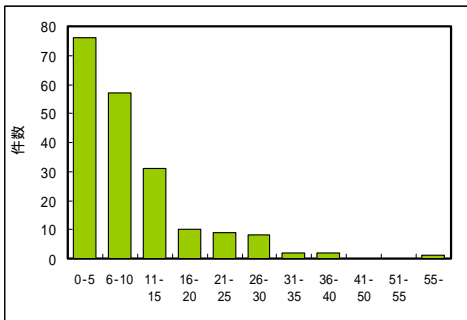
測定と統計処理

- つまり,
 - 測定 measurement とは 観察 observation の具体化のひとつであり,
 - 統計処理 statistical analysis とは 記述 description (記述的一般化)のための補助手段である.
- したがって,
 - 測定(観察)がクズなら, 統計結果(記述)もクズ.
 - Garbage In, Garbage Out.
 - 統計結果(記述)はそのままでは説明(理論, モデル)を与えない.
 - これを同一視するのは第一次同型性錯誤 First Order Isomorphism Fallacy
 - しかし, 統計結果(記述)がゴミならそれを元にした説明もゴミ.

データの散らばりの例(1)



データの散らばり(2)



データの要約

- 散らばってるデータの性質をどうやって一言で表すか?
- 一つのやり方は代表的な値と散らばり方を示すこと.
- これが記述統計の基本

記述統計 基礎の基礎

- 代表値
 - たくさんあるデータを上手く代表する値を計算したい.

最頻値 mode	平均情報量, χ^2
中央値 median	範囲, 四分領域
平均値 mean	分散, 標準偏差

- 万能の統計量はない. 分布や尺度水準に応じて使い分ける.

尺度水準って何?

- データの性質によって, 数値を足したり引いたりかけたりしていいかどうかが違う.
 - 名義尺度
 - 数値自体には意味がない. 単なるラベル.
 - 例) 性別, 血液型
 - 順序尺度
 - 数値は順序性だけを持っている.
 - 例) モースの硬度計
 - 間隔尺度
 - 数値の間隔に意味有り. ただし, 原点は決まらない.
 - 例) 温度
 - 比率尺度
 - 原点が決まっている. 比がとれる.
 - 例) 長さ

↑ 可能な演算が限られる
測定/コーディングは比較的容易

↓ 色んな演算が可能
測定/コーディングは困難(というより, ほぼ不可能)

尺度水準とデータコーディング形式

- 間隔尺度, 比率尺度は基本的に普通の数値として扱っていいけど, 名義尺度, 順序尺度のコーディングには工夫が必要.
- これを覚えておかないと, 統計ソフトにデータを食わせられない.
- 具体的には補遺でおぎないます.

平均と中央値(1)

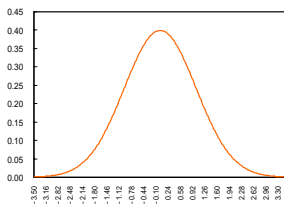
- 平均値 (Mean, M)
 - 最も頻繁に使われる代表値
 - 全ての観測値 (= 標本の測定値) を足しあわせて標本数 n で割ったもの.

$$\bar{X} = \frac{\sum_{i=1}^n x}{n}$$

- 中央値 (Median)
 - 観測値を大きいモノから順番に並べて, ちょうど真ん中に来る値

平均値と中央値(2)

- 多くの多変量解析(および頻繁に使われる検定法)では暗に平均値を求めている.
- その前提として, 個々の変数が正規分布 normal distribution していることを仮定しているから.

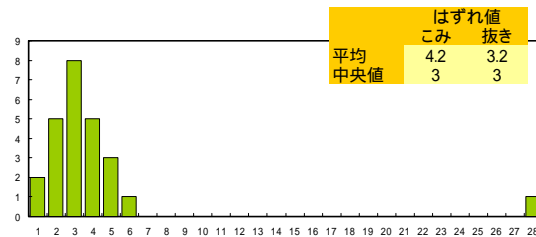


← こういう分布(確率密度関数)なので, 平均値と中央値は一致する

発見者の名をとり, ガウス分布 (Gaussian distribution)とも呼ばれる.

平均値と中央値(3)

- 平均値は分布の歪み, はずれ値に強く影響される.
- 中央値はそうでもない.



四分領域と分散・標準偏差(1)

- 四分領域 quartile range (四分位偏差 quartile deviation)
 - データを小さい値から順番に並べていって, 25%のところの値(Q1)と75%(Q3)のところの値を探す.
 - 中央値(Median = Q2)を求める.

$$Q = \frac{(Q_3 - Q_2) + (Q_2 - Q_1)}{2} = \frac{Q_3 - Q_1}{2}$$

- データの最大値と最小値の差を範囲(レンジ, range)とよび, 散らばりの指標にすることもある.
- 実際にはそんなに使われていない.

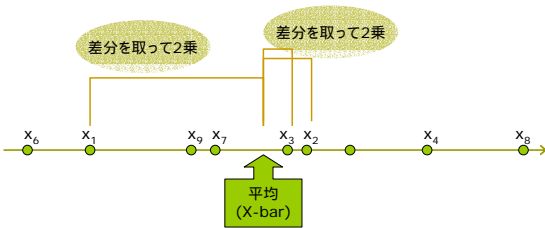
四分領域と分散・標準偏差(2)

- 分散と標準偏差
 - 各観測値と平均値との差を2乗して全て足しあわせて $n-1$ で割ったのが分散 σ^2
 - その平方根をとったのが標準偏差 standard deviation: SD
- 平均値とセットで使う

式でかくとこんなの. n でなく $(n-1)$ で割るのは, 母集団の推定値にするため

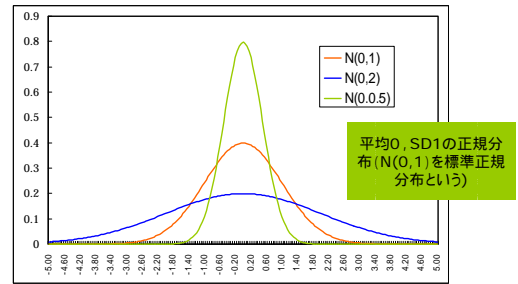
$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

分散と標準偏差－模式図



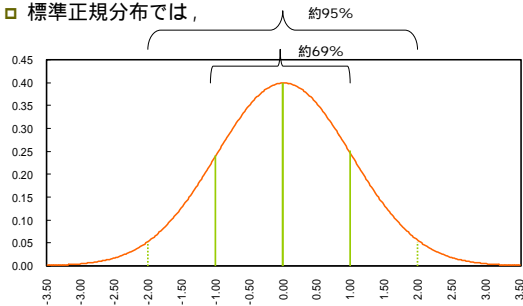
四分領域と分散・標準偏差(3)

- 平均は同じでも分散(標準偏差が違くと...)



正規分布とSDと標準化(1)

- 標準正規分布では、



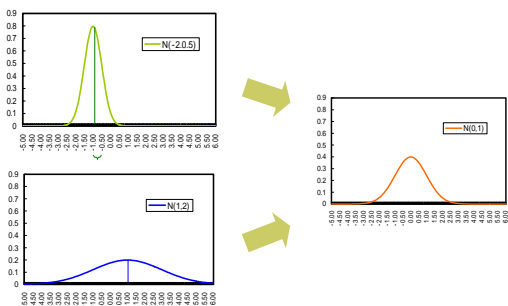
正規分布とSDと標準化(2)

- 標準化とは、

- 正規分布している(と仮定できる)変数について、各値から平均値を引き、SDで割るという操作のこと
- 要するに平均を0とし、SDを単位=1とした分布 = 標準正規分布にしている。

$$z_i = \frac{x_i - \bar{x}}{s}$$

正規分布とSDと標準化(3)



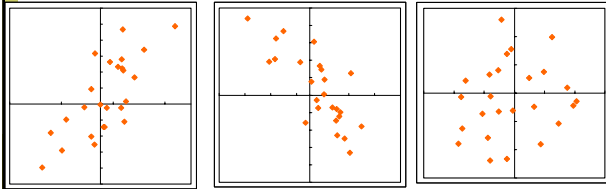
共分散(1)

- 2つの変数がどれくらい連動して散らばっているかを表す統計量を共分散という。

- 共変動

- 一つの変数 x が大きくなると他の変数 y も大きくなる傾向がある (正の関係がある; 身長と体重)
- 一つの変数 x が大きくなると他の変数 y は小さくなる傾向がある (負の関係がある; 成人期以降の年齢と体力)
- 一つの変数 x が大きくなっても、他の変数 y が大きくなるとも小さくならない(無関係である)

共分散(2)



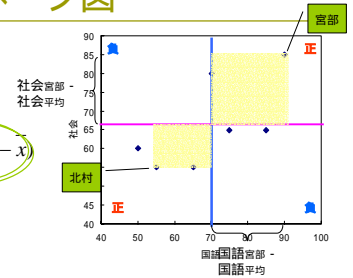
正の関係

負の関係

無関係

共分散のイメージ図

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



	東野	宮部	井上	我孫子	池坂	北村	島田	平均
国語	70	90	85	65	75	55	50	70.0
社会	80	85	65	55	65	55	60	66.4

相関係数(1)

□ 身長と体重の共分散

- 身長をセンチで表すと …… 38.83
- インチで表すと …… 15.28
- フィートで表すと …… 1.27

- 単位によって、値が変わってしまう！
- (その方が都合が良い場合もあるが) まずい場合も多い.
- 何とかならないでしょうか？

- ここで、標準偏差を使った標準化を行ってみましょう.
- これが、世に言う“Pearsonの相関係数 r”です.

相関係数(2)

□ 式で書くとこんなのです.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- -1.0 r 1.0
- 相関の大きさの目安 (特に強い根拠があるわけではない)
 - 0.0 $|r|$ 0.2 …… ほとんど相関なし
 - 0.2 $< |r|$ 0.4 …… 弱い相関あり
 - 0.4 $< |r|$ 0.7 …… 比較的強い相関あり
 - 0.7 $< |r|$ 1.0 …… 強い相関あり

相関係数(3)

□ 相関の強さ 散布図の傾斜の強さ



□ 相関の強さ(弱さ)は1本の直線から点が全体としてずれている程度.

- → XでYをどれくらい予測/説明できるか c.f. 決定係数 R^2



相関と回帰

□ 予測という観点からみると (単)回帰分析.

予測値 \hat{y} と 実測値 y のズレ (誤差の2乗) が最小になるように、

$$\hat{y} = ax + c$$

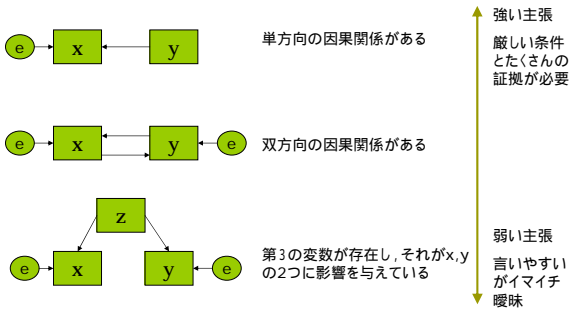
を決める.

誤差の2乗を最小にするから、

最小二乗法

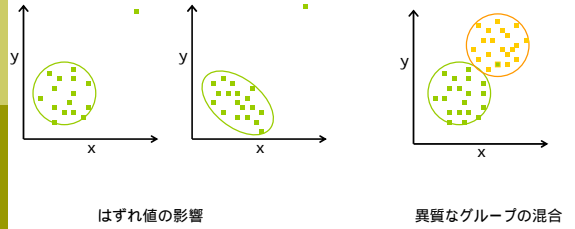
決定係数 = (重)相関係数の2乗は、 x の重み付け変量で y の分散を何%説明できているかに相当する

相関が生じる理由



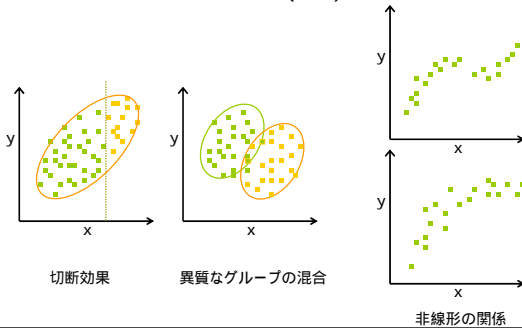
安易な計算だけではいけない(1)

- ホントは相関がないのに高い相関係数が得られるケース



安易な計算だけではいけない(2)

- ホントは関連があるのに無い(弱い)ようにみえるケース



というわけで教訓

- 自分の扱っているデータの性質をよく知りましょう。
 - 自分がやった測定(実験, 調査, コーディングなど)の性質をよく考え、尺度水準や信頼性, 妥当性を検討しましょう。
 - 統計ソフト, 表計算ソフトで数値だけを求めるのではなく, ヒストグラム(分布の検討)や散布図を書いて, 現実をよくみましょう。
- 自分の扱っている現象の性質をよく知りましょう。
 - そのような結果が得られた理由を推測できる力をつけましょう。
 - 仮説から来る要請と現象の本態を混同しないようにしましょう。

相関係数 r の仲間達

- 連関係数

- 2つの名義尺度変数の関連性(共変性)を表す
- 連関係数については, これ以降の回でまた触れます。

	無関連			関連あり			最大関連		
	好き	嫌い	合計	好き	嫌い	合計	好き	嫌い	合計
男性	16	8	24	8	16	24	24	0	24
女性	24	12	36	32	4	36	16	20	36
合計	40	20	60	40	20	60	40	20	60

男女に関係なく, 好き嫌いの比率は一定

男女で, 好き嫌いの比率が異なる

完全に偏っている

ユールの連関係数 $Q=0$

$Q=-0.88$

$Q=1$

相関係数 r の仲間達(2)

- 順位相関係数

- 2つの順序尺度変数の関係性を表す

7つの物語の意外性と面白さの順位

	A	B	C	D	E	G	H
意外性	1	2	3	4	5	6	7
面白さ	1	4	3	2	5	7	6

この順位がどれくらい似ているか?

→ Spearmanの順位相関係数を使う(式は省略します。自分で調べてください)

入門(1)のまとめ

- 今回は、下記のことを学びました。
 - なんて統計分析するの？
 - 測定とは？
 - 代表値と散らばり(散布度)
 - 平均値と中央値
 - 四分領域と分散・標準偏差
 - 共分散と相関係数

- 統計を実際に使いはじめた初心者にありがちなことに、基本統計量や分布など生データに近い部分での検討を怠って、いきなり多変量解析や検定に入る、ということがあります。これは自分のデータの性質を把握しないまま、道具に振り回されることにつながるので望ましくありません。気をつけましょう。