データ分析
Data Analysis

１学期の授業

授業内容

・表計算ソフトウェア（EXCEL）（3時間）

・EXCELに関する実技試験（1時間）

・Rでのデータ処理

・Rの実技試験

評価

EXCEL試験，Rの実技試験，提出物

1学期の内容

・統計（Statistics)とは

・データ解析（Data Analysis)の重要性

・Data Analystとは

1学期の内容

・GAFA vs BATH

・Ciscoの例

https://www.cisco.com/c/ja_jp/services/overview.html

・AWSの例

https://aws.amazon.com/jp/about-aws/

・SAS 機械学習（Machine Learning)

https://www.sas.com/ja_jp/insights/analytics/machine-learning.html

代表値

- 平均値(average)
- 中央値(median)
- 最頻値(mode)
- 四分位偏差（Quartile deviation)
- 分散(variance)
- 標準偏差(standard deviation)

1：サンプリングの重要性

例題1：「1924年度エール大学卒業生の年間平均所得は25,111ドルである」(by タイム誌) このレポートの問題点を指摘しなさい.

例題2：歴史的事例から(アメリカの大統領選挙)

1936年アメリカ大統領選挙において，ルーズベルト対ランドンの選挙予想について，「リテラリー・ダイジェスト」誌は1000万枚の質問用紙を発送し225万枚を回収し，ランドン氏有利を予想した．しかし，結果はルーズベルト氏の勝利に終わった.

1：サンプリングの重要性

＊例題1の視点：サンプリングに注目する.

・誰に依頼したのか
・何人回答したか
・正確に回答しているのか（見栄を張っていないか）

＊例題2の視点：この失敗の原因は標本（サンプリング）の選択に，顕著なバイアス（偏り）があったとされた.

・電話加入者と自動車所有者からの標本だったこと
・自分に送られてきた質問用紙に自ら進んで記入し，返送する階層にある傾向があったこと

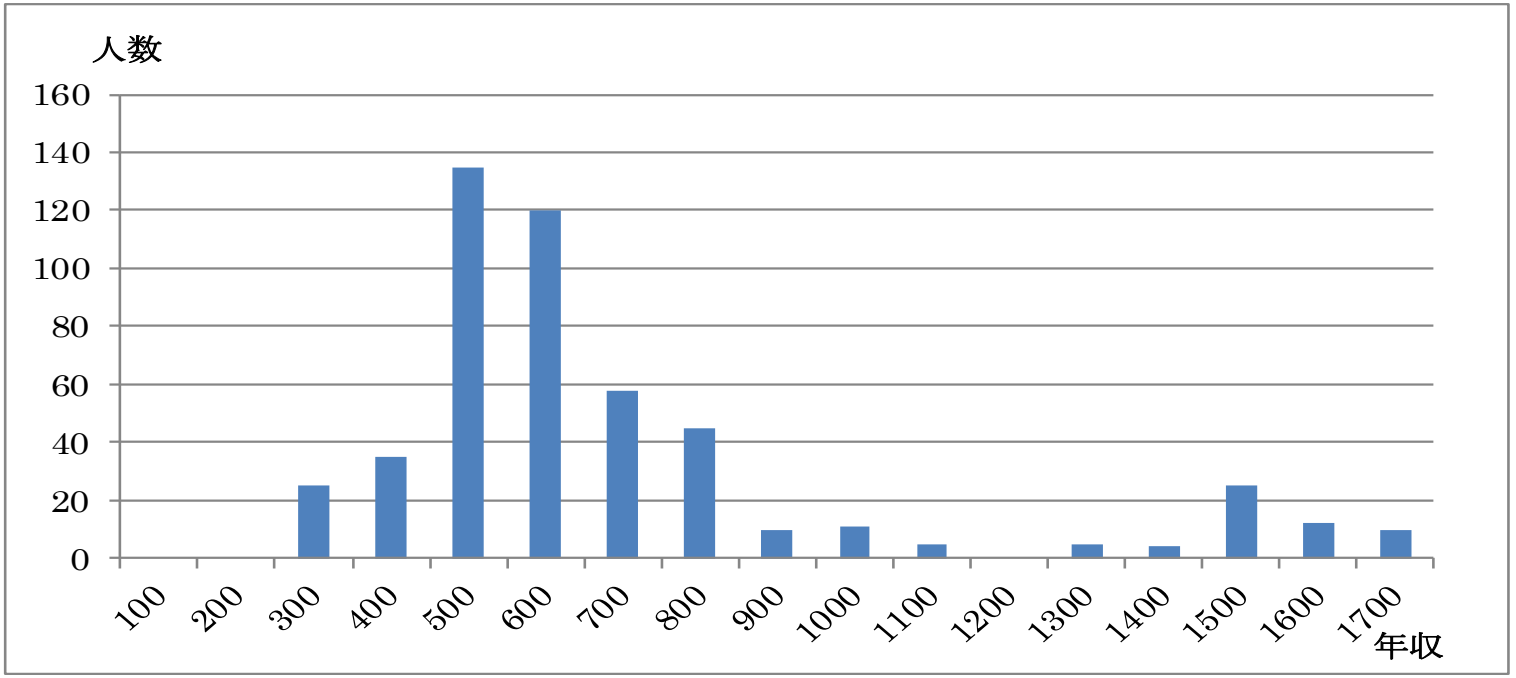
これらがランドン候補派に多いことが判明した.

例：世論操作（意識調査，選挙予想）など

→ 層化抽出法　層化ランダムサンプリング，多段抽出法 （例：R.D.D.）

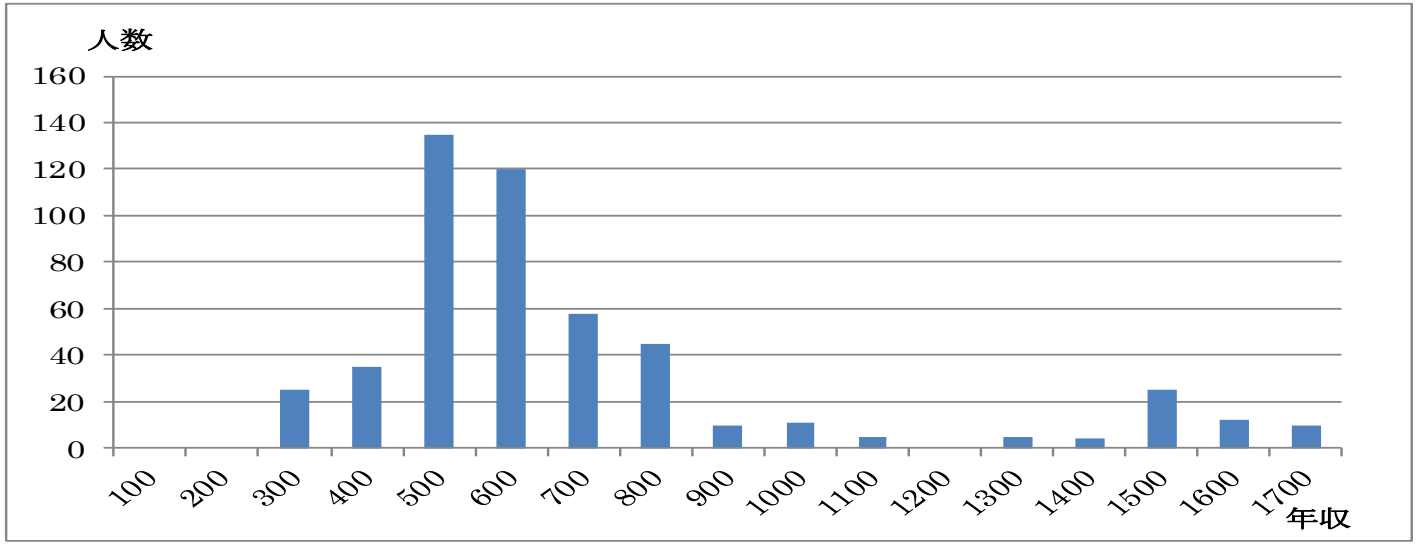
２:「平均」を単純に信用してはいけない

例題3:「新宿区内のIT企業500社の社員の年間平均所得は700万円である. その分布は下図である」



演習:このレポートの問題点を指摘しなさい.

2:「平均」を単純に信用してはいけない



集団の数値的特徴を表す場合,「平均」という概念が一般的であるが, これは必ずしもその集団を特徴づけることにはならない. かえって中央値 , 最頻値の方が適切になる。

分布が「正規分布」であるときは, この3つは同一になるが一般に通常の 分布では異なる. これらは意図的に使われることが多い.

２:「平均」を単純に信用してはいけない

(1) 代表値のいろいろ

・平均（average） ・・・相加平均（全データの総和を総数で割った値）
相乗平均，調和平均

・中央値（median） ・・・全データを並べたときの真ん中の値

・最頻値（mode） ・・・データ中に一番多い頻度で現れる値

(2) 散らばり具合を表す：分散（variance） と標準偏差（standard deviation)

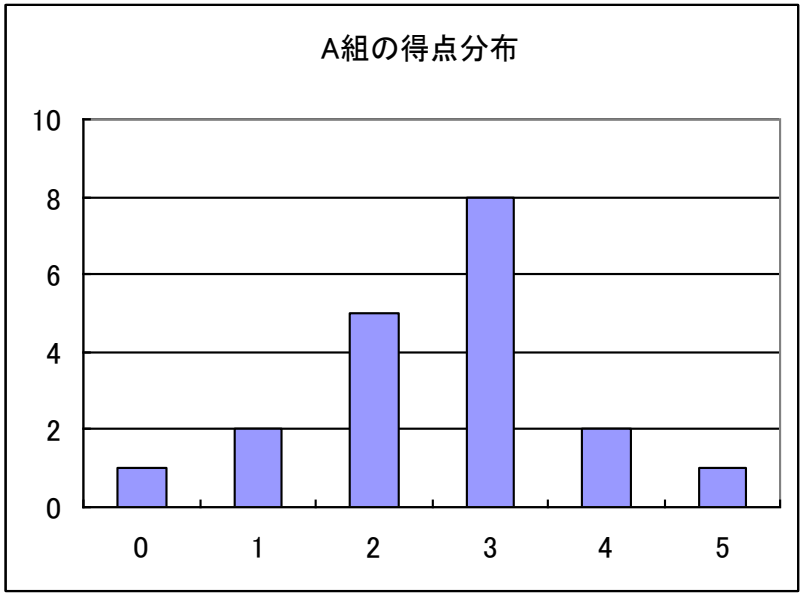
単なる平均だけでなくその分布の散らばり具合を数値で示すものが，「分散」，「標準偏差」である．いわゆる「偏差値」の計算の元になるものである．

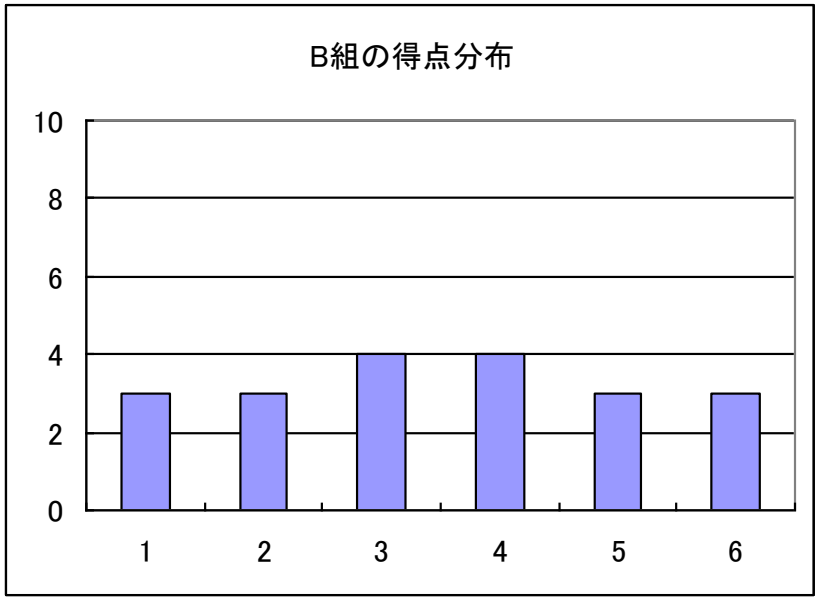
x1, x2, x3, … xnを各データ，ｍをその平均とすると，次の式で定義する．

分散 ｖ＝ $\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ 標準偏差 σ＝ $\sqrt{v}$

データのばらつき（分散・標準偏差）

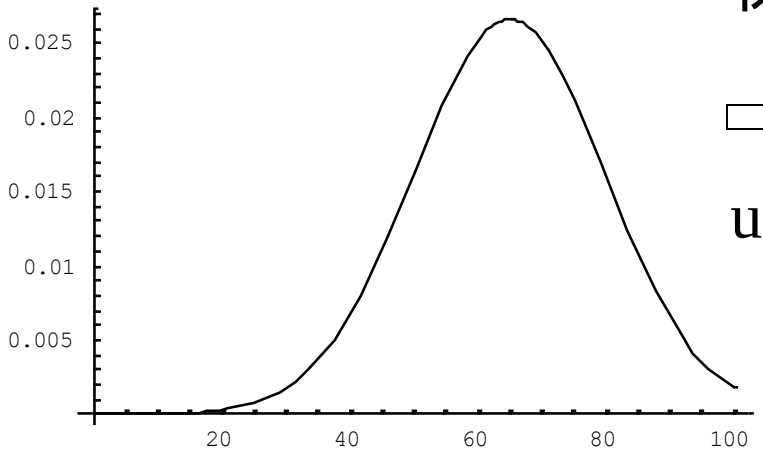
平均だけでは特徴づかない分布（平均はともに2.5点）



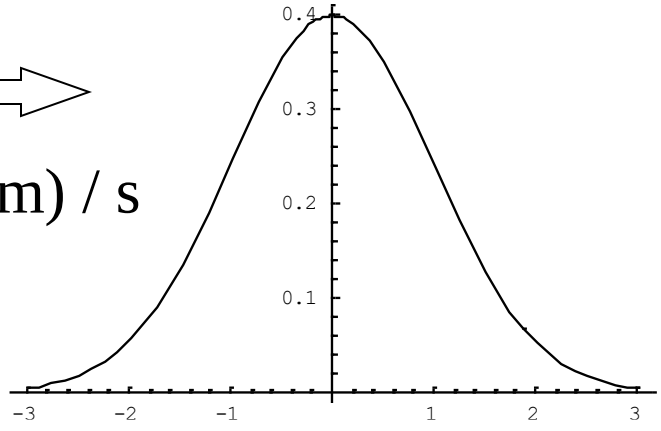


変量の標準化（正規分布の場合）

標準化



$u = (x-m) / s$



平均 m:65点　標準偏差 s:15点

変数　x

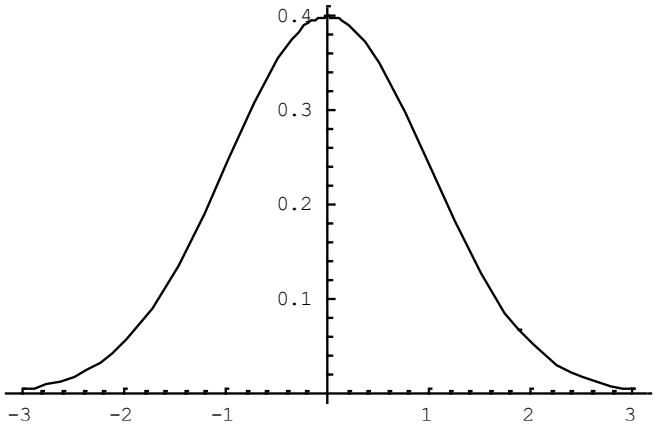
平均:0点　標準偏差1点

変数　u

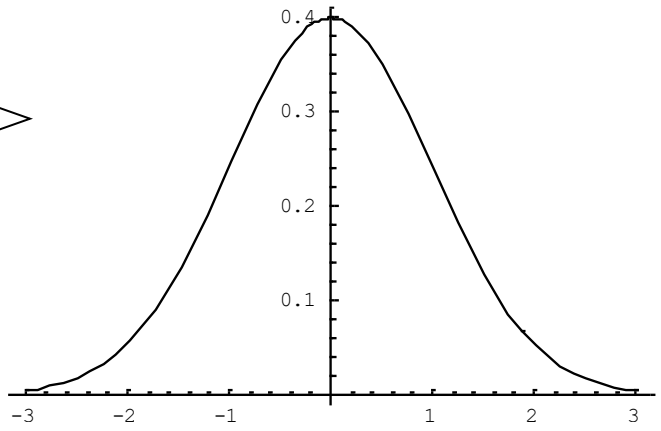
x

$u = (x-m) / s$

偏差値（正規分布の場合）



偏差値



平均：0点　標準偏差１点

変数　u

平均：50点、 偏差10点

変数　z

$u = (x-m) / s$

$z=u \times 10+50$

四分位偏差(Quartile)

テストC，テストDの第1四分位数，第2四分位数，第3四分位数をそれぞれ求める。
テストC＝{ 34, 37, 40, 45, 48, 52, 60, 65, 71, 77, 81 }
テストD={ 34, 37, 45, 50, 53, 57, 65, 66, 71, 77 }

(i) 奇数の場合（テストC:11個）
1. 最初にデータの中央値を求め，この中央値を第2四分位数とする。
2. 次に，第2四分位数をデータから取り除き，第2四分位数より小さいグループと大きいグループに分ける。
3. 小さいグループの中央値が第1四分位数，大きいグループの中央値が第3四分位数。

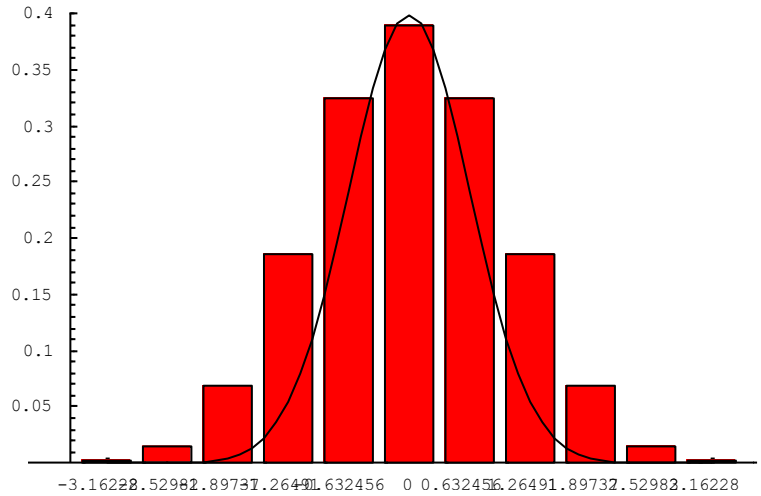
∴　第1四分位数　40　　第2四分位数　52　　第3四分位数　71

(ii) 偶数の場合（テストD:10個）
1. データの中央値を求め，この中央値が第2四分位数です。
2. 第2四分位数より小さいグループと大きいグループに分ける。
3. 小さいグループの中央値が第1四分位数，大きいグループの中央値が第3四分位数。

∴　第1四分位数　45　　第2四分位数　55　　第3四分位数　66

３：統計では正規分布(normal distribution)が重要

偶然性が高い大量データ
（身長，体重や大学入試センターテストなど）の分布．



$m-\sigma<x<m+\sigma$　約68%
$m-2\sigma<x<m+2\sigma$　約95%
$m-3\sigma<x<m+3\sigma$　約99%

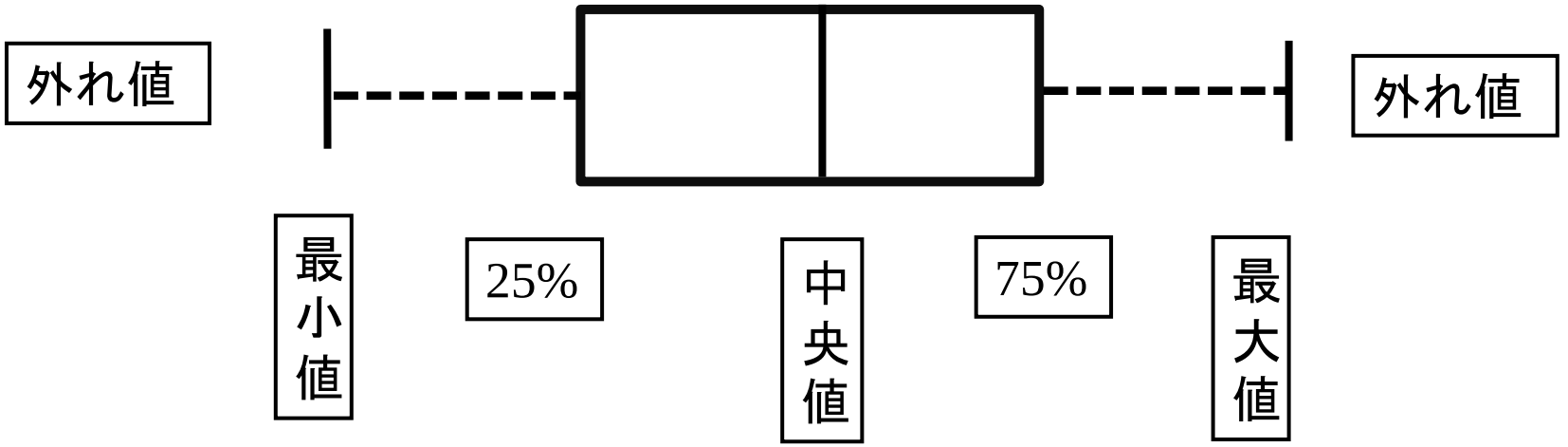
・平均＝m：グラフ中央　・標準偏差＝σ：

まとめ・・・統計のウソを見破るカギ

(1) どのような団体，研究機関の調査か（原発の例では，政府側か市民側か）
(2) 調査方法は適切か（サンプリングの適正さや回答率）
(3) 数値に信頼性があるのか
　・平均の問題点（上述）
　・特殊な要因がバイアスになっていないか・・・数値の上昇，下降
　・足りないデータはないか・・・相関関係（がんと生活習慣など）

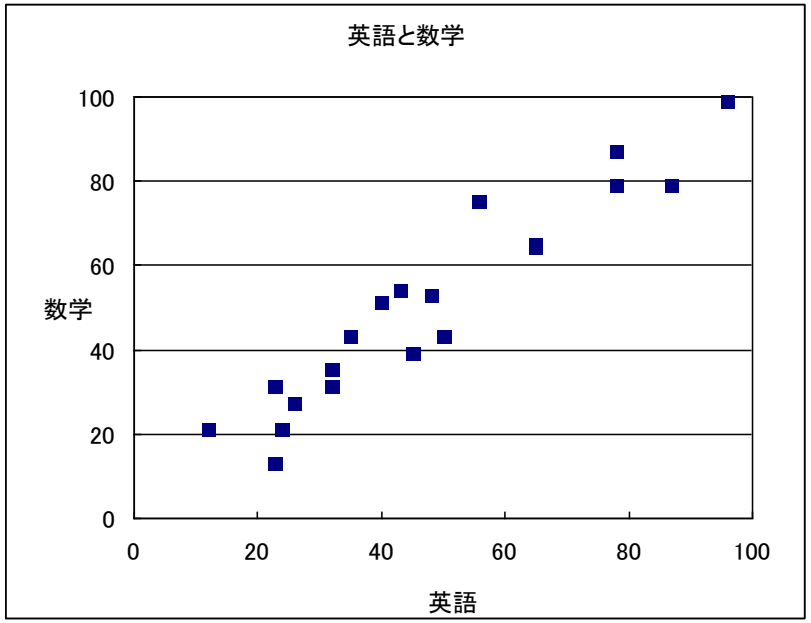
【参考書】「統計学で楽しむ」，鈴木義一郎著，ブルーバックス，講談社.
「統計でウソをつく法」，ダレル・ハフ著，ブルーバックス，講談社.
「Rによるやさしい統計学」，山田剛史，杉澤武俊，村井潤一郎著，オーム社.

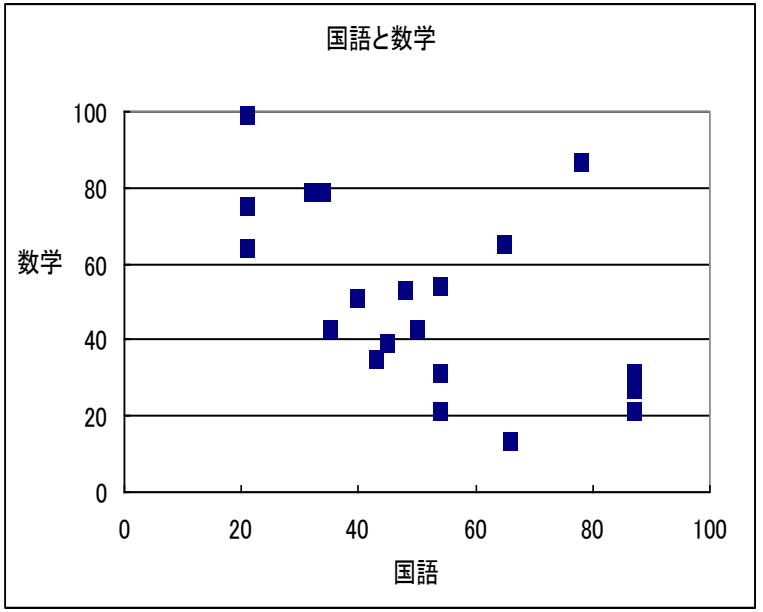
箱ひげ図



相関関係
Correlation

相関関係





このような関係をどのように数値化するか？

相関関係

no	1	2	3	・・・・	i	・・・・	n
(X,Y)	(x1,y1)	(x2,y2)	(x3,y3)	・・・・	(xi,yi)	・・・・	(xn,yn)

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

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

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

そして，$-1 \leqq r \leqq 1$が成り立つ.

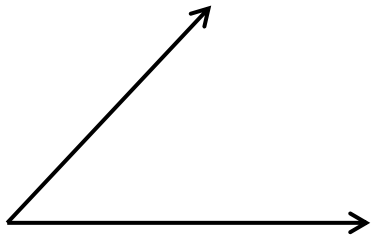
相関関係　（−1≦r≦1が成り立つ理由）

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

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})}} = \frac{X \cdot Y}{|X||Y|} = \cos\theta$$

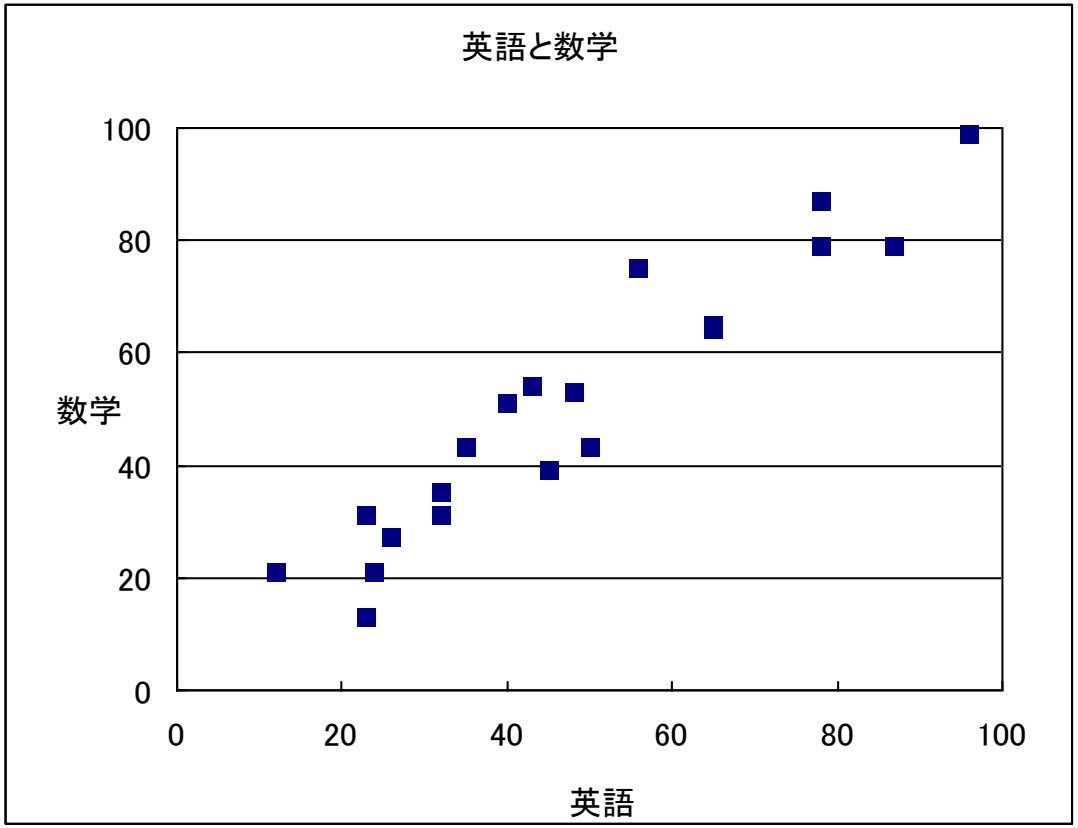
$$X = (x_1 - \bar{x}, x_2 - \bar{x}, \cdots, x_i - \bar{x} \cdots, x_n - \bar{x})$$

$$Y = (y_1 - \bar{y}, y_2 - \bar{y}, \cdots, y_i - \bar{y}, \cdots, y_n - \bar{y})$$



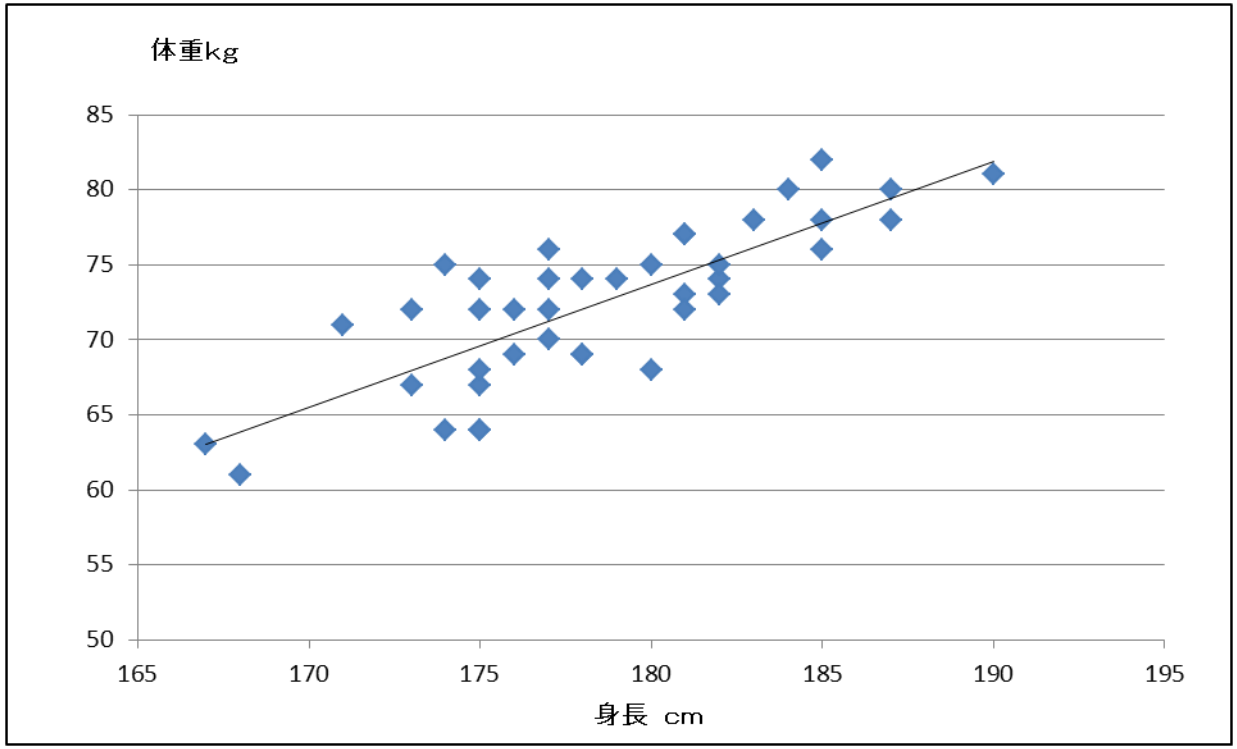
N次元ベクトルの内積

相関関係と因果関係との違いは？



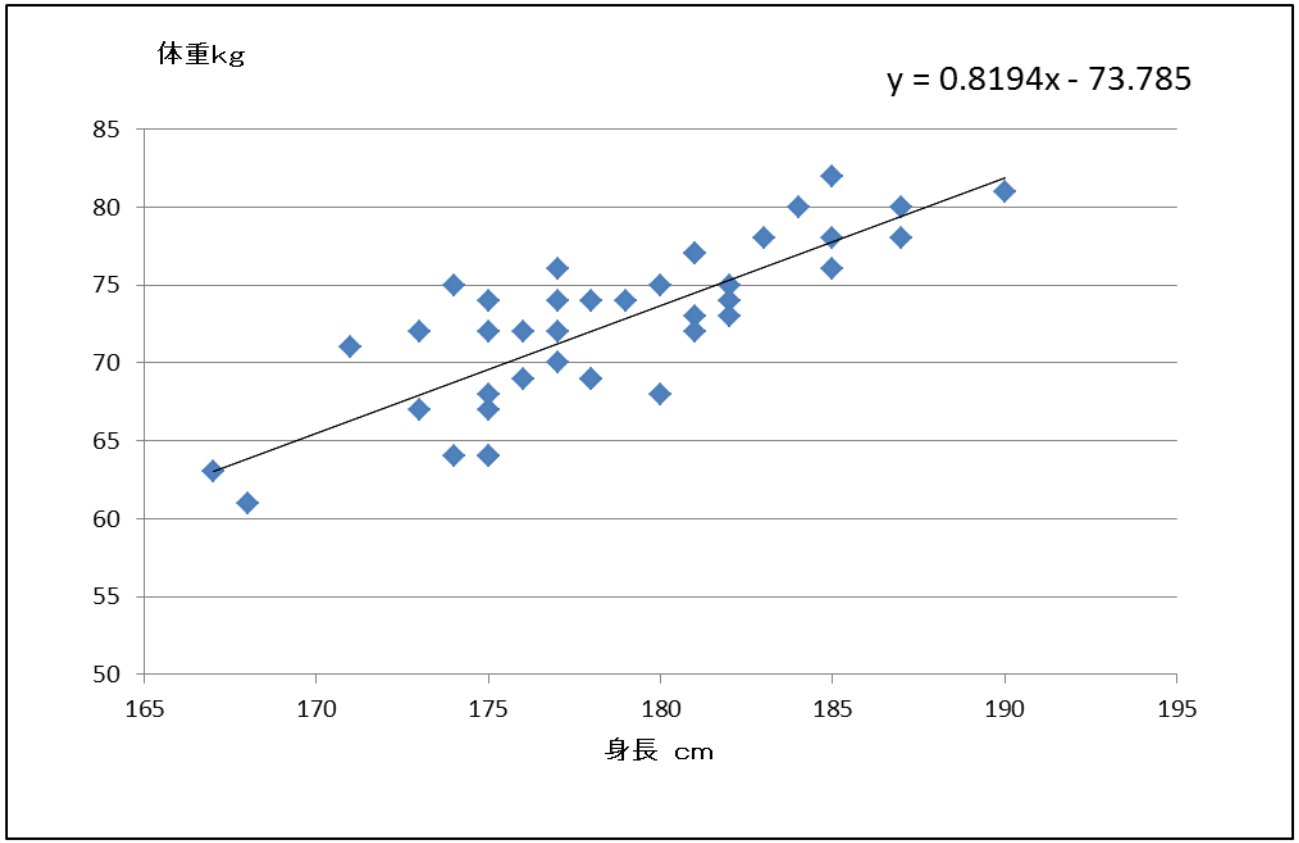
回帰分析
Regression

回帰直線



このような直線近似して予測するには？

回帰直線(線型回帰)



直線の傾き、切片は最小二乗法で求める。

いろいろなデータ

データの種類

・個人に関するデータ（住所，身長など・・・）
・社会に関するデータ（GDP，販売など・・・
・自然に関するデータ（面積，地震など・・・）

データの属性

・静的（static），動的(dynamic)…時間によって変化するかしないか
・連続的（analog・アナログ）か離散的（digital・デジタル）が
・定性的（qualitative）か定量的(quantitative)か

データの表現

・表（table）…ベクトル，行列など
・グラフ（graph）…分布や関係図

データ分析の手順

データの検索・収集（Webで検索，データのダウンロード）
↓
データの整理・加工（表形式やグラフ化）
↓
データの分析　（統計処理や分析後，新たな情報の創出）
↓
発表・報告　（レポートやWebページに発表）

データ分析I(アンケート集計を例に)

【アンケートの目的】 (問題解決のための方法)

① 基礎的な統計資料を作成する。(国勢調査，人口動態調査など)
② 実態等を調査し，問題を発見する。(意識調査，世論調査，市場調査など)
③ 問題の原因や構造を解明する。(収集データを分析する)
④ 問題解決策の選択や実行可能性を探る。(新製品や新規事業のための調査など)

例:学院生の通学時間，学院食堂の人気メニュー，
学院生の女性に対する意識調査，地域と学院の関わりなど

データ分析I（アンケート集計を例に）

【手順】

① 調査項目の決定
② 調査票の作成（アンケート質問文，回答形式の決定）
③ 調査対象の決定（アンケート実施対象者の決定，サンプリング人数の決定）
④ 調査方法の決定（面接，街頭調査，郵送調査，電子メール等）
⑤ アンケート回収（回収率の計算）
⑥ 回収データの整理・加工・分析
・基本データの整理（性別人数，男女人数，回答別人数）
・・・ピボットテーブル，棒グラフ
・基礎統計処理（合計，平均，標準偏差など）
・2項目間のクロス集計　・・・ピボットテーブル，比率グラフ，3Dグラフ
・相関係数，回帰分析
・質問全体の傾向分析　・・・レーダーチャート
この他，統計的な手法である「推定・検定」という考え方を用いることも多い。
⑦ 報告書の作成

データ分析I（アンケート集計を例に）

【クロス集計】

性別	学年	系	問1	問2	問3	問4	問5
男子	1年	文系	3	2	5	1	3
男子	2年	文系	4	5	3	4	4
男子	3年	理系	1	2	1	4	1
男子	3年	理系	5	1	5	4	4
女子	3年	理系	3	4	4	3	1
男子	3年	文系	4	5	3	1	2
男子	1年	文系	3	2	5	5	4
女子	2年	理系	5	4	2	2	5
女子	3年	理系	2	4	2	2	2
男子	3年	理系	2	3	3	1	5
女子	2年	文系	1	3	1	4	5
女子	3年	文系	1	1	2	1	5

データ分析I（アンケート集計を例に）

【クロス集計】

	性別		
学年	女子	男子	総計
1年	2	10	12
2年	2	3	5
3年	5	8	13
総計	9	21	30

学年×性別

	性別		
問1	女子	男子	総計
1	3	3	6
2	2	3	5
3	2	5	7
4		8	8
5	2	2	4
総計	9	21	30

問1×性別

（参考）統計的な考え方：検定とは

統計の難しい計算はともかく，統計での有効な考え方に検定と推定がある．この基本的な考え方は，一を調べて十を知ることである．

すなわち，あるデータ（標本）が与えられた時，そこから全体集団（母集団）の様子をある幅をもって推測したり，そのデータから創り出した仮説が正しいか，正しくないかを検定したりすることを，確率や確率分布の知識を用いて行うことである．

例題5：ある運動選手の記録を測った.

{68.1, 70.4, 71.5, 67.6, 70.2, 74.5, 68.6, 70.3, 71.2, 69.6}

av=70.2　var=3.532

この選手がある薬を飲んで再度記録を測った.

{72.7,69.4,74.2,70.6,69.0,72.5}

av=71.4　var=3.5233

このとき，この薬は記録向上に「効果があった」と結論付けてもいいのだろうか

統計では次のように考えていく.

（解説）

ここでは何としても「効果があった」と結論付けたいとしよう.

しかし効果のない薬でも，標本の関係でこのようなことが起こり得る．しかし，確率的にこのような事象が起こる確率が低いことを示すことができれば，ある程度の合理的な説得力をもつことになる．この手順が仮説検定法である.