統計

今泉　忠

講義目標

- 高度情報化により、問題解決のためにはデータを収集して、それをもとに考えることが必須である。
- この講義ではデータを活用するための基本力の習得を目指し、統計的データ分析の入門を取り扱う。
- 2年秋学期からの統計学Iを履修するための事前科目でもある。
- 具体的には、データの要約と因果関係の検証のためのデータ分析を内容とする。
- 講義では実際に検討し、理解を深める事が重要であるので、コンピュータソフトを用いてデータ分析の基礎をも習得する。適宜，チームレポートなどを作成する。

シラバス　キーワード

- データ分析，統計で扱うデータ，ヒストグラム，PDCA,変数の種類
- 中央値、平均、最頻値,記述統計としての要約値，四分位範囲や標準偏差
- 箱ひげ図,2つの群からのデータの比較
- 原因と結果の表現,特性要因図
- 平均値の変化,平均値を用いて比較するには
- 散布図，相関関係，相関係数，線形関係
- 因果関係，単回帰分析，特性要因図，残差

テキストなど

- テキスト　統計学基礎　（統計検定2級対応）　東京図書
  - 学習範囲は特に，第1章と第2章
- 参考となるテキスト
  - Head First データ分析　　オライリー・ジャパン
  - Head First 統計学　　オライリー・ジャパン
  - データの分析　（統計検定3級対応）　東京図書
  - AP Statistics
- URL
  - http://tnext.tama.ac.jp
  - http://stat.tama.ac.jp

PDCA（Plan->DoまたはData->Check ->Act)のサイクルで

- 問題を一般的な事柄として解決するために
  - 解決案を策定する→原因のなかで、結果に対する効果が最大となる原因を策定する
  - 改善したい結果を明確にする（目標化）。
  - 経験などを活用する→モデル（仮説）を考える
  - 「新しいアイデア②は模倣①から」
    - ①成功した事柄を真似する
      - オーソドックスな方法を適用してみる
    - ②模倣だけでは、真似で終わる。「何故真似したい」と考えたかが重要
      - 原因に関して他の要因も考えてみる

PDCAのフレーム

- Plan は仮説を考えること
  - データを集めて集計しただけでは、単なる数字の集まりであり、そこから何が読み取れるか必ずしも明らかではない
  - 統計を作成するときは、必ず、「○○について知りたい！」という目的があるはずですから、得られた結果を、その目的に合わせて上手に使うことが重要です。
  - 目的を具体化することにより、仮説へ
- Data は仮説を評価するために、データを集めること
  - 仮説に沿ったデータとしては集めない、
  - 満遍なく、十分なデータの数を
- Check は仮説をデータを用いて、検証すること
  - その道具として グラフは、結果を視覚的に表す道具としても有効
  - グラフをうまく使うことによって、自分の考えていること（仮説の妥当性）を相手に的確に伝えることができます。
- Act は結果をもとに仮説の再設定などを含めて、改善を行うこと

目標

- データに基づいて仮説を構築する
- 「何が問題か？」を考えることができ，表現できる
- データの特徴を探る
  - どのような値が多いのか
  - データはどのように分布するのか
  - たくさんのデータを少ない値で要約するには
- 2つの変数を比較する
  - 2変数間の関係を探る，連関と因果の違いを知る
  - 2つのグループの違いを表現できる
  - 関係を視覚化する
- 結果を他者に分かり易く伝えられる

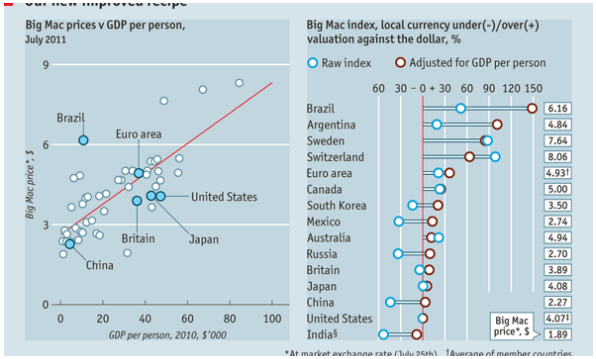
データから考える！（例題データ）

- 目的：利益拡大（減少抑制）のために顧客からの要望を吸い上げ
- 無用な長時間の管理
- 適切なコールサービスを行いたい通話時間
  - データ件数 31492件
  - 100件(ここでは、最初の100件)

length				
77	56	51	367	25
289	44	148	277	109
128	274	9	201	238
59	479	115	52	163
19	211	19	9	353
148	179	76	700	394
157	1	138	182	1
203	68	178	73	234
126	386	76	199	4
118	2631	67	325	293
104	90	102	75	415
141	30	35	103	196
290	57	80	64	23
48	89	143	121	234
3	116	951	11	126
2	225	106	9	363
372	700	55	88	85
140	40	4	1148	166
438	73	54	2	606
56	75	137	465	115

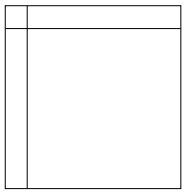
経済指標としてのハンバーガーの価格は利用できないか

- マクドナルドは世界各国で同じ商品を提供していそう
- UKのThe Economist誌が，取り上げた（以下の図は引用）



データをまとめる

- 平均余命や医療において，BMIが関係しているが示唆された
- BMI=体重（kg）/身長^2(m)
- データ行列は一般に，列方向が変数



	身長	体重	性別	BMI
	178	63	男	=
	165	62	男	
	168	69	男	
	152	41	女	
	175	71	男	
	175	61	男	
	165	62	男	
	162	48	女	
	164	52	女	
	170	55	男	
	169	69	男	
	155	48	女	
	153	44	女	
	162	49	女	
	168	69	男	
貴方の値				

課題

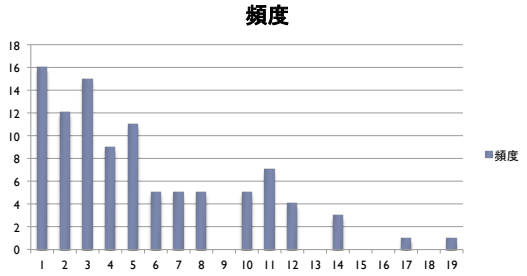
- コールセンターとはどのような目的のために設置されているかまとめなさい
- ビッグマック指数について調べなさい
  - 最新のビッグマック指数と10年前のビッグマック指数について調べなさい
  - １人当たりGDPについても調べなさい
- ボディマス指数について調べなさい
  - BMIを計算しなさい
  - 自分のBMIを求め，それから自分の肥満などについて考えなさい

データの特徴を捉える

- グラフを作成する
- 柱状グラフ
- ヒストグラム
- 円グラフ
- その他には？
  - （1）
  - （2）
  - （3）
  - （4）
- グラフの詳細についてはtnextの講義ページの資料を参照

1変数の場合　コールセンターでの対応時間

- ヒストグラムを、EXCELの関数を用いて作成するには
  - （1）柱状グラフを作成する

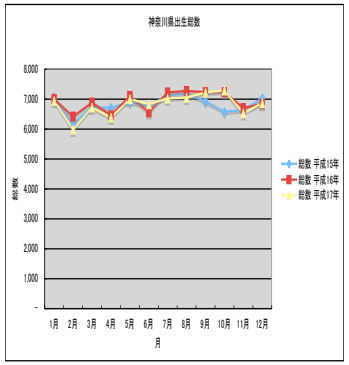


  - 区間を重ねる

データ例

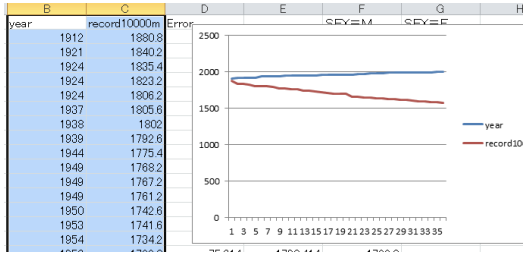
- 月別（時系列）　　グラフ

神奈川県	総数		
月	平成15年	平成16年	平成17年
1月	7,069	7,003	6,940
2月	6,174	6,411	5,981
3月	6,720	6,880	6,725
4月	6,707	6,450	6,387
5月	6,893	7,108	7,004
6月	6,767	6,558	6,824
7月	7,122	7,219	7,020
8月	7,184	7,274	7,048
9月	6,919	7,230	7,225
10月	6,574	7,234	7,286
11月	6,624	6,706	6,529
12月	7,039	6,833	6,870



時系列データ

- 年毎と収集された10000m走のデータから記録の推移を調べた
  =>年は連続と考えられるので，折れ線グラフ



変数とその型

- 各列は、各変数に対応している。
  - 変数には測定水準による変数の型と分析での役割がある。
- 変数の型
  - 名義変数、順序変数、間隔変数、比例変数
    性別　　身長、通話時間

カテゴリー変数（質的変数）　　*量的変数*

- 分析での役割
  - 結果と考えられる変数　目的変数、被説明変数、従属変数
  - 原因と考えられる変数　説明変数、独立変数

EXCELでIF文を

- 原データを変換して分析する場合がある。そのために第一歩として，EXCELでIF文などの関数を用いてことができるようになることが必要
- 資料は，tnextからダウンロードして，演習問題を実施しておくこと
  - countif(),frequency()などを利用できるようにしておくこと
- Rなどの準備すると良い
- ipadなどのタブレット端末用の統計分析ソフトも多数あるので，そのようなソフトも準備すると良い

プログラミングをしてみる

- 例 **【体型評価BMI】**
  肥満度の判定方法の一つにBMI(ボディ・マス・インデックス)指数での評価がある
  BMI＝**体重(kg)／身長(m)2**
- BMI指数の標準値は22.0である
  体重～22.0**身長(m)²** ？
- BMI 18.5未満 やせ
  18.5～25未満 標準
  25～30未満 肥満
  30以上 高度肥満
- 調べてみよう：BMI作成の目的は？

データの特徴をEXCELを用いて

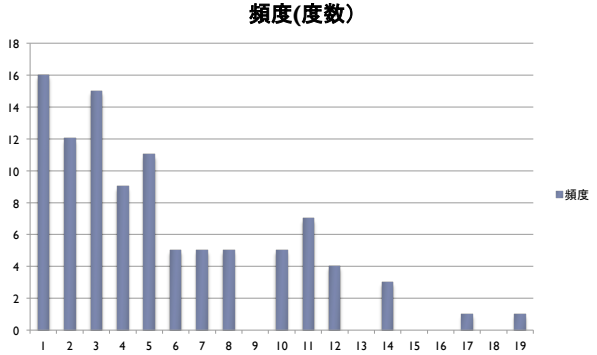
- 分布　データが大量に場合には？
  - おおよその傾向を知るには
    - ヒストグラム
      - □区間数や幅が変わると視覚的な形も変わる
      - □EXCELでは、
        - □=countif(セル範囲、条件)
        - □=frequency(セル範囲，区間）
        - □分析ツールからヒストグラム
    - データの最小値、最大値
      - □データが変わると、値も変わる
      - □EXCELでは、
        - □=min(セル範囲）、=max(セル範囲）

ヒストグラムを作成する

- データの出現度数を捉える。
  - 設定した区間内に、何個のデータがあるか
- パラメトリック分布と対応づける
  - 代表的なもの　正規分布、二項分布
  - 知っておくとよいもの　ガンマ分布、χ２乗分布
- Excelでは、関数FREQUENCY(データ、区間)
  - 実際には累積分布
- 関数COUNTIF（データ、条件）

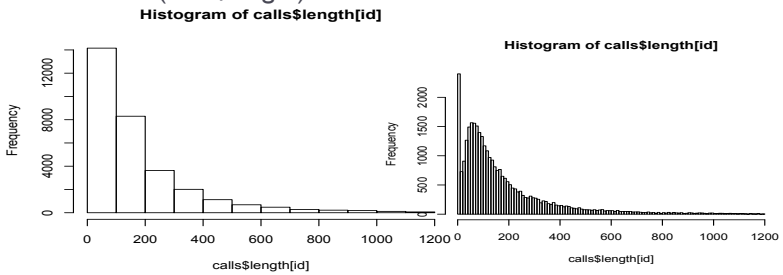
コールセンターでの対応時間

- ヒストグラムを、EXCELの関数を用いて作成する
- はじめに柱状グラフを作成する



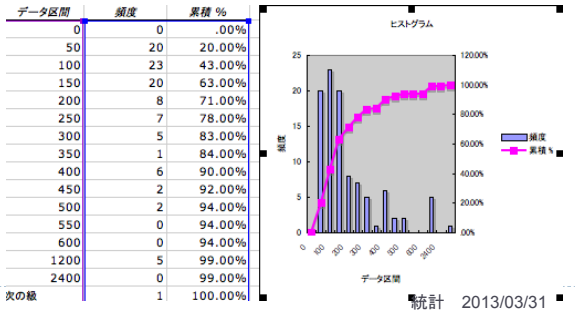
コールセンターのヒストグラム

- Rを用いて
  - hist(calls$length)



全体を捉える

- ヒストグラム
- (1)区間の幅、または、区間数を設定する
  - 区間数は普通5～20
- (2)分析ツールからヒストグラムを選択する



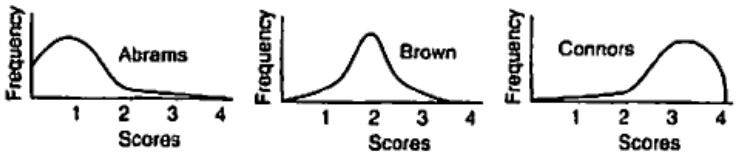
データ区間	頻度	累積 %
0	0	.00%
50	20	20.00%
100	23	43.00%
150	20	63.00%
200	8	71.00%
250	7	78.00%
300	5	83.00%
350	1	84.00%
400	6	90.00%
450	2	92.00%
500	2	94.00%
550	0	94.00%
600	0	94.00%
1200	5	99.00%
2400	0	99.00%
次の級	1	100.00%

ヒストグラム　区間数の設定

- 仮説を調べるために必要と考える区間数
- データ数 n によって、区間数を変える。
  - 経験的な階級数mの決め方
    - 平方根則
      m = SQRT(n)
  - Struges
    m = 1+(log n)/(log 2)
  - もっと大雑把に
- 適当に(四捨五入、切捨て、切上げ)整数化

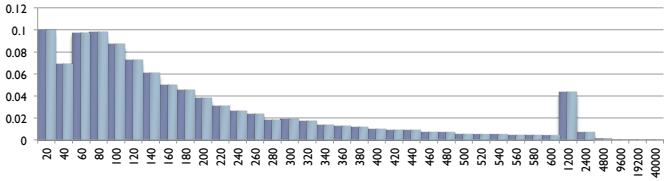
ヒストグラムの特徴

- □ 右に裾が長い
- □ 対称
- □ 左に裾が長い
- □ ベル型
- □ 一様



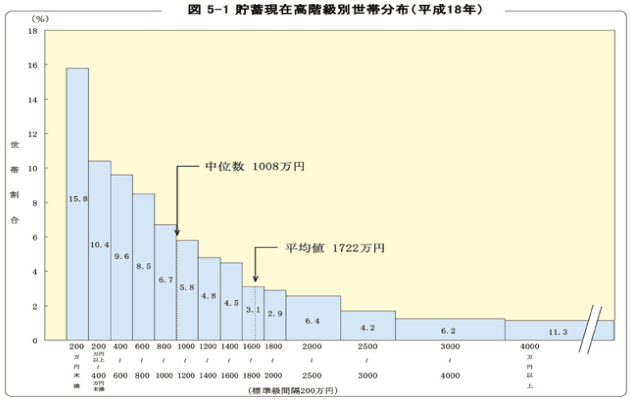
コールセンターデータのヒストグラムから

- 600秒までは区間幅は20秒，その後は
  601~1200,1201~2400,2401~4800,4801~9600,…
- 右に裾が長い（右に歪んでいる,Right-Skewed）
- 山が2つではないか　？（通話時間が短い所）
- 外れ値は？



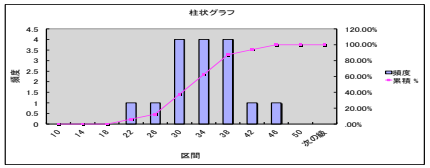
区間幅が異なるヒストグラム

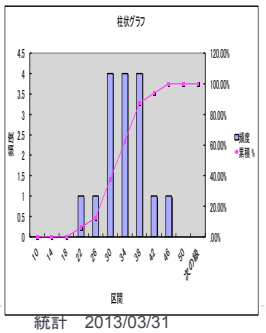
- 所得分布や貯蓄分布など
  - 面積が度数に比例するように



柱状グラフの場合

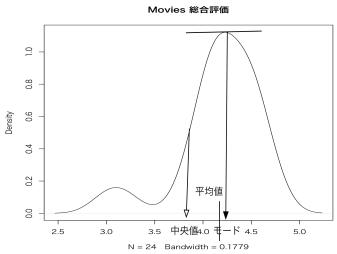
- 1つの項目を比べる場合には、頻度（度数）と%は同じ
- 複数の項目を比較する場合、目的により
  - 頻度（度数）または%
- 見せ方で受ける印象が異なることに注意





ヒストグラムから代表値へ

- ヒストグラム
  - 全体がわかる　vs　視覚的な形状でイメージが変わる
- 形状に依存しない量　n個のデータの情報を表現するために代表値を用いる。
- 代表値
  - 平均　データの値をもとに計算
  - 中央値　累積度数をもとに計算
  - モード　度数をもとに計算



代表値としての平均　データをもとに計算

- （算術）平均　値の大まかな情報

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- 幾何平均値　伸び率について情報
  - 年ごとに 5%,6%,10%,8%伸びた

$$y_G = \sqrt[n]{y_1 y_2 \cdots y_n} = \sqrt[n]{\prod_{i=1}^{n} y_i}$$

- 調和平均 往復の平均時速

$$\frac{1}{y_H} = \frac{1}{n}\left(\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_n}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{y_i}$$

順位値にも使用できるメディアンとモード

- 尺度
  - 比尺度、間隔尺度、順序尺度、名義尺度
- メディアン　並べた時に大きい方、小さい方から50%となるyの値
- モード　頻度がもっとも多いxの値

平均と中央値

- 両方共に、データの値から求められる
- 平均値：値での重心
- 中央値：累積度数での50%点となるデータ
  - データ 4,6,2,8,5　n=5
    - 平均値＝（4＋6＋2＋8＋5）／5＝3
    - 中央値＝5　＜＝（2，4，5，6，8）
    - n=7で、データの値が20，25
      - 平均値＝(4+6+2+8+5+20+25)/7=10
      - 中央値＝6　<=(2,4,5,6,8,20,25)
    - データ数が偶数の時はどうするの？

バラツキの指標

- 分散と標準偏差 $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$$s_y = \sqrt{s_y^2} \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 : MLE$$

- 範囲、四分位範囲(順位値にも使用できる)
  - Q1 累積確率で0.25, Q2は0.50,Q3は0.75

$$R = \max(y_1, y_2, \cdots, y_n) - \min(y_1, y_2, \cdots, y_n)$$

$$Q = \frac{1}{2}(Q_3 - Q_1),\ Q_2 : Median$$

- データ例 (2,2,3,4,4)と(1,1,3,5,5)
  - 偏差(2-3), (2-3),(3-3),(4-3)(4-3) と (1-3),(1-3),(3-3),(5-3),(5-3)

平均と分散

- 中央値と四分位範囲
- 平均と？ 分散(または標準偏差)
  - 偏差 平均からの偏差 $y_i - \bar{y}$
  - 正負がある。
    - 絶対偏差 $|y_i - \bar{y}|$
    - 2乗偏差 $|y_i - \bar{y}|^2$
- 2乗偏差の平均値は
  - 単位を平均値と同じに(標準偏差)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

統計量を求める

- EXCEL
- 1変量について調べる
  - 平均値 AVERAGE(範囲)
    - 標準偏差 STDEV(範囲) 分散 VAR(範囲)
  - 中央値 MEDAIN(範囲) 範囲
    - QUANTILE(範囲,0~4)
- 頻度を数える 条件を全て満足するデータの個数
  - COUNTIF(範囲1，条件1，範囲2，条件2，．．．)
    - 条件指定は “>=200” のように指定
    - 分析ツールでヒストグラム
      - 入力範囲、データ区間範囲

最小値、最大値

- データの範囲
  - その1 最大値ー最小値
  - その2 3/4Q-1/4Q(四分位範囲)
  - 3／4Q データを小さい順に並べた時、75%に対応するデータ値
  - 1／4Q データを小さい順に並べた時、25%に対応するデータ値
  - EXCELでは
    - =QUARTILE(データの範囲、計算する値)
      - 計算する値 0＝最小値、1＝第1四分位、2＝第2四分位(＝中央値)、3＝第3四分位、4＝最大値
    - ＝PERCENTILE(データの範囲、率) 率 0.75/0.25

統計量

- データのばらつきを調べるには
  - 平均と分散（標準偏差）
  - 中央値と第1四分位値と3四分位値
  - 平均値については、平均値と標準誤差
- Excel
  - =AVERAGE(Data),=VAR(Data)
  - =MEDIAN(Data),=QUARTILE(Data,種類）

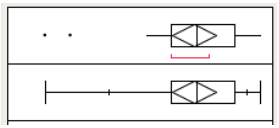
EXCELでの関数一覧

- セル　A1とかB4
- セル範囲 A2:A101　　コロン：で指定
- 統計計算の関数

名称	関数（＝ではじめること）	
最小値	min(範囲)	quartile(範囲,0)
最大値	max(範囲)	quartile(範囲,4)
平均値	average（範囲）	
分散	var(範囲)	
標準偏差	sqrt(var(範囲))	stdev(範囲)
中央値	median(範囲)	quartile(範囲,2)
Q1（第1四分位値）		quartile(範囲,1)
Q2（第2四分位値）		quartile(範囲,2)
Q3（第3四分位値）		quartile(範囲,3)

箱ひげ図　またはダイヤモンド図

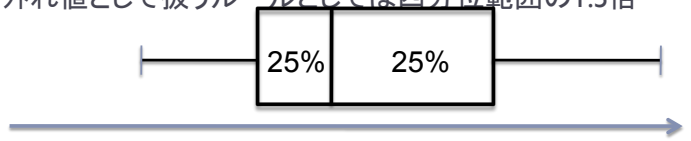
- データを要約するために、数値を用いて表す
- 最小値<->Q1<—>Q2<—>Q3<->最大値
- 平均値-αx標準偏差<—>平均値<—>平均値+αx標準偏差
  - α≒2



分位点			モーメント	
100.0%	最大値	4.7000	平均	4.1791667
99.5%		4.7000	標準偏差	0.408581
97.5%		4.7000	平均の標準誤差	0.0834012
90.0%		4.6000	平均の上側95%信頼限界	4.3516953
75.0%	4分位点	4.5000	平均の下側95%信頼限界	4.0066381
50.0%	中央値	4.2000	N	24
25.0%	4分位点	4.0000		
10.0%		3.5000		
2.5%		3.0000		
0.5%		3.0000		
0.0%	最小値	3.0000		

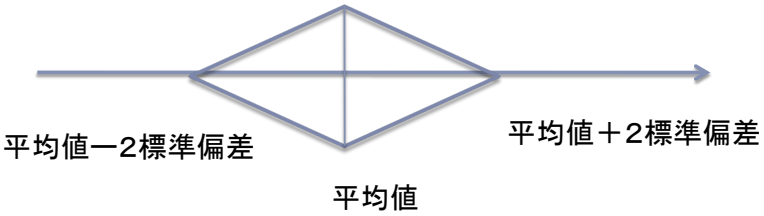
箱ひげ図

- データの累積頻度から計算する
- 五数要約
  - 箱：25%となるデータの値（Q1）、50%となるデータの値（Q2）、75%となるデータの値（Q3）
  - 四分位範囲(IQR)：Q3—Q1　箱の横幅
  - ひげ：箱の両側
    - 最小値と最大値
    - 10%となるデータの値と90%となるデータの値
  - 外れ値として扱うルールとしては四分位範囲の1.5倍



ダイヤモンド図

- データの値から計算する
- 3数要約　大まかに捉える
  - 平均値ー2標準偏差、平均値、平均値＋2標準偏差

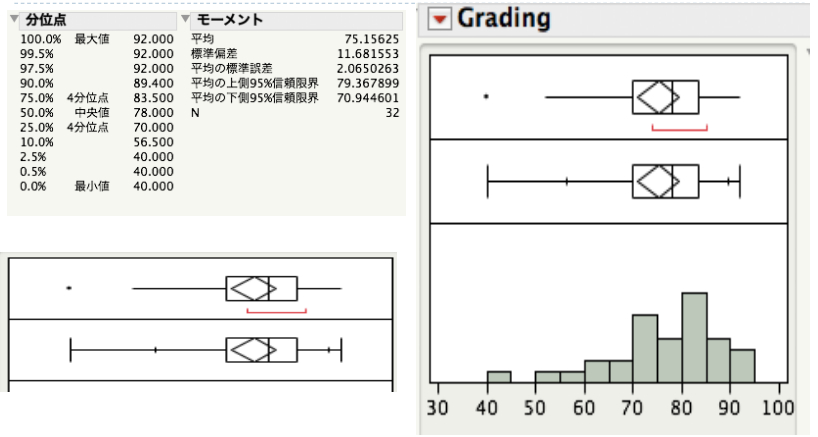


- 外れ値として扱うルールとして3標準偏差

代表値

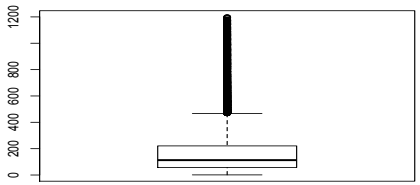
	平均(Mean)	中央値(Median)	最頻値
代表値	$ym=(y_1+y_2+\ldots+y_n)/n$	ymed 累積頻度がデータ数の半分となる値	ymod 頻度数が最大となる値
バラツキ	データ $s_y^2=sum[(y_i-ym)^2]$ 平均値 s_y^2/n	四分位範囲(IQR) ＝Q3ーQ1	
図	ダイヤモンド図 平均値ー2標準偏差、平均値、平均値＋2標準偏差 平均値ー2標準誤差、平均値、平均値＋2標準誤差	箱ひげ図 最大値、Q1、Q2、Q3、最大値	

図の表現に用いる値



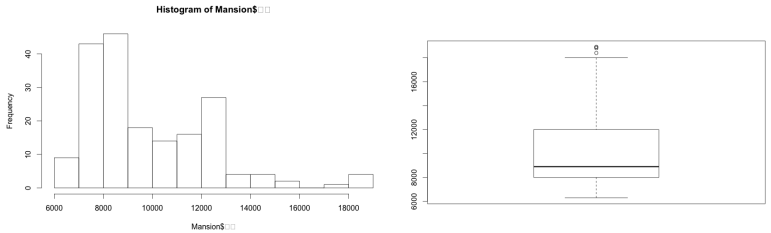
コールセンターのデータから

```
>summary(calls$length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    57.0   115.0   188.6   225.0 28740.0
>sd(calls$length)
[1] 312.7768
```



演習

- A君は，大学生になり，１人住まいをしたいと考えている。そのために，家賃について相場を調べるためにデータを収集した。
  - 収集したデータをtnextからダウンロードする
  - ヒストグラムと箱ひげ図を作成し，考察する



演習

- 各チームで，以下の場所の１つについて30物件のデータを収集してヒストグラムなどを作成しなさい
- ダウンロードしたデータとの比較を行いなさい
- 地域
  - 京王永山，多摩センター，聖蹟桜ヶ丘
- データを収集する場合に注意すべきことをまとめなさい
- 30物件と指定されているが，これは比較検討するのに十分な件数かどうか検討しなさい
- PDCAで記述しなさい
  - PDCA報告シートをtnextからダインロードすること

演習　紙コプターの飛行時間

- 各チームで紙コプターのプロトタイプをもとに，滞空時間の改善を行う
  - チーム毎に，実験セットとレポートセットを用意しなさい
- 滞空時間の改善に寄与すると考える要因を挙げなさい。
- そのなかでもっとも重要と考える要因を選びなさい
- プロトタイプ20回，改善した紙コプター20回づつの実験を行いなさい
- 実験の前に，
  - ストップウォッチの押し方や空調などの実験条件について設定すること
- PDCAで記述しなさい
  - PDCA報告シートをtnextからダインロードすること

紙コプターの実験

- 改善前の２秒の滞空時間と改善後の２秒の滞空時間は同じであるのか
- 絶対的な時間は同じ、
- 改善前の平均滞空時間が１秒で、改善後の平均滞空時間が３秒ならば、平均値からの相対的時間は異なる。
- 比較できるようにするには
  - データの標準化

標準化得点を作成するには

- **身長**
  - cmとm
    - 160cmと1.6mは同じ
- **身長と体重**
  - 160cmと170cm、 45Kgと55Kg
    差は10cmと10kg 単位も異なり比較できない

$$z_i = ay_i + b$$
$$\bar{z} = a\bar{y}, \quad s_z^2 = a^2 s_y^2$$

標準化得点(Standardized Score)

- **標準化(Z得点)** $z_i = \dfrac{y_i - \bar{y}}{s_x}$

$$\bar{z} = 0, s_z^2 = 1$$

- **偏差得点(T得点)**

$$T_i = 10z_i + 50 = 10\frac{x_i - \bar{x}}{s_x} + 50$$
$$\bar{T} = 50, s_T^2 = 10^2$$

標準化得点の平均値と分散

$$z_i = \frac{(y_i - \bar{y})}{s_y}$$
$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \bar{y})}{s_y} = \frac{1}{ns_y}\sum_{i=1}^{n}(y_i - \bar{y}) = 0$$
$$s_z^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i)^2$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left\{\frac{(y_i - \bar{y})}{s_y}\right\}^2 = \frac{1}{n-1}\frac{1}{s_y^2}\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1$$

変数変換による統計量の変化

変数変換による統計量の変化

	定数の加算 y*=y+b	定数倍 y*=ay
平均	平均＋定数	定数×平均
中央値	中央値＋定数	定数×中央値
標準偏差	変化なし	定数×標準偏差
四分位範囲	変化なし	定数×四分位範囲
範囲	変化なし	定数×範囲
変動係数	分母が変化するので，値が変わる	変化なし

演習

- 紙コプターのデータで、改善前と改善後について、
  - 平均値がどの位の値であるかを確認したい
  - 違いがあるかどうか調べたい
- どのような量を用いたグラフを作成すべきか，3つ挙げなさい
  - 地チームの滞空時間についてヒストグラムを作成しなさい
  - 滞空時間に関して外れ値とみなせるデータはあるか
- 各チームの平均について，ヒストグラムを作成して考察しなさい
  - 平均は各チームの個数分ある
  - プロトタイプは同じであるが，各チームの滞空時間は同じとみなせるか

平均　データから求めた平均は

- ある実験でn個のデータを収集した。
- このような実験をm回行った。

$$(y_1, y_2, \cdots, y_n) \to \bar{y}$$

$$1\ (y_{11}, y_{12}, \cdots, y_{1n}) \to \bar{y}_1$$
$$2\ (y_{21}, y_{22}, \cdots, y_{2n}) \to \bar{y}_2$$
$$\vdots$$
$$m\ (y_{m1}, y_{m2}, \cdots, y_{mn}) \to \bar{y}_m$$

- m個の平均のヒストグラムはどうなるの？

データでのモデル

- 個々の平均誤差
  - 同じ値 μ に誤差があると考えよう
    - 個々の平均はデータから求めている
    - データには誤差がある
- 誤差の平均の平均＝0としよう。
- (n×m個のデータの誤差の平均)

$$\bar{y}_i = \mu + \bar{e}_i$$

$$\bar{\bar{y}} = \frac{1}{m}\sum_{j=1}^{m}\bar{y}_j = \mu + \frac{1}{m}\sum_{j=1}^{m}\bar{e}_j$$

$$s_{\bar{y}}^2 = \frac{1}{m}\sum_{j=1}^{m}(\bar{y}_j - \bar{\bar{y}})^2 = \frac{1}{m}\sum_{j=1}^{m}\{(\mu + \bar{e}_j) - (\mu + \bar{\bar{e}})\}^2$$

$$= \frac{1}{m}\sum_{j=1}^{m}(\bar{e}_j - \bar{\bar{e}})^2$$

$$\bar{\bar{e}} = 0$$

$$s_{\bar{y}}^2 = \frac{1}{m}\sum_{j=1}^{m}\bar{e}_j^2$$

残差の２乗和から、標準誤差を求める

$$\sum_{i=1}^{m}\bar{e}_i^2 = \sum_{i=1}^{m}(\frac{1}{n}\sum_{k=1}^{n}e_{ik})^2 = \frac{1}{n^2}\sum_{i=1}^{m}(\sum_{j=1}^{n}\sum_{k=1}^{n}e_{ij}e_{ik})$$

$$= \frac{1}{n^2}\sum_{i=1}^{m}\sum_{k=1}^{n}e_{ik}^2 + \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}e_{ij}\sum_{k,k\neq i}^{n}e_{ik}$$

$$\approx \frac{1}{n^2}\sum_{i=1}^{m}\sum_{j=1}^{n}e_{ij}^2$$

$$s_{\bar{y}}^2 = \frac{1}{m}\frac{1}{n}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n}e_{ij}^2 = \frac{1}{m}\frac{1}{n}\sum_{i=1}^{m}s_y^2 = \frac{1}{m}\frac{1}{n}ms_y^2 = \frac{1}{n}s_y^2$$

きちんとした理屈は確率変数を用いて

平均のヒストグラム

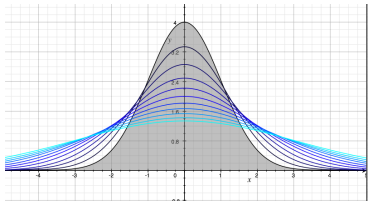
- 通話データを分割してデータセットを作成しなさい
  - 1データセットのデータレコード数として　10件，50件，100件，200件，1000件

	レコード数	データセット数
A	10	3000
B	50	600
C	100	300
D	200	150
E	1000	30

  - 平均のヒストグラムを作成して，考察しなさい

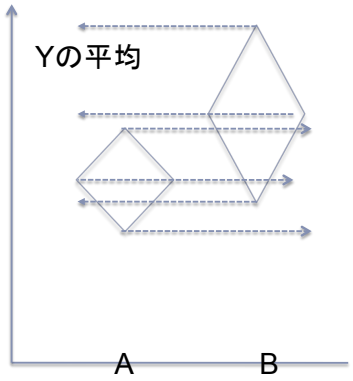
平均のヒストグラム

- 平均のヒストグラムは，どのような形
  - □　右に裾が長い
  - □　対称
  - □　左に裾が長い
  - □　ベル型
  - □　一様
- 平均のヒストグラムでのばらつきは
  - 四分位範囲
  - 標準偏差



演習

- 改善前と改善後それぞれについて、
- 平均の箱ひげ図を作成しなさい
- 平均のダイヤモンド図を作成しなさい
- 結果からわかることを書きなさい



実験前と実験後ではどう違うの？

- 実験前のデータ

$$(y_{B1}, y_{B2}, \cdots, y_{Bn_B}) \to \bar{y}_B, s_{y_B}$$

- 実験後のデータ

$$(y_{A1}, y_{A2}, \cdots, y_{An_A}) \to \bar{y}_A, s_{y_A}$$

- 改善度　$\bar{y}_B - \bar{y}_A$
  - 差が大きければ・・・　　差が大きいはどの計るの？
- 平均値の変化を調べる

和を考える

- Aさんは、1月1日から、12月30日まで、毎日
  - その日が良かったならば1、良くなかったならば−1と記録することにした。Aさんが大晦日に、この記録をまとめるとどうようになるであろうか
  - 全ての日が良くないと和は−364、全ての日が良いと和は364
  - 364日の内1日のみ良くないと和は363、364日の内1日のみ良いと和は−363
- 和＝0がもっとも有りそう
  - 何故？

$$H = \sum_{i=1}^{T} Y_i$$

$$-364 \le H \le 364,$$

$$H = -364 = (-1) + (-1) + \cdots + (-1)$$

$$H = 364 = 1 + 1 + \cdots + 1$$

和の調査と実験

- 調査：皆で記録してみよう。1年は大変なのである日数分記録してみよう。この場合、記録する日を適切に決めることが重要。
  - だれがどのように決めるか
- 実験：サイコロを振り、ある目が出たならば良い日、他の目が出たならば良くない日
  - 例　奇数ならば　良い日、偶数ならば良くない日
  - 良い日の個数と良くない日の個数は異なっていてもよいのでは！

平均の違いを調べる

- （1）パン屋さんでパンを買っている。そのパン屋さんのパンは重さの平均が200gで分散が50グラムであった。しかし、最近、すこし軽くなっているようなので、10日購入してデータを集めた。
  (190,230,200,210,195,190,205,203,192,190)
- このパンは軽くなっているのであろうか？
- 調べるために、
  - 軽くなっている（重くなっている）を調べるのか、200gから離れているのか
  - 再度パンを買うと、異なる重さのデータが得られる事になる。

データでのモデル

- 測定値は、真の平均値に誤差
- 真の平均値は同じ

$$y_{ik} = \mu_i + e_{ij}$$

- 個々の平均値

$$\mu_1 = \mu_2 = \cdots = \mu_m = \mu$$

$$\bar{y}_i = \mu + \bar{e}$$

$$s_{y_i}^2 = \frac{1}{n}\sum_{k=1}^{n}(y_{ik} - \bar{y}_i)^2 = \frac{1}{n}\sum_{k=1}^{n}\{(\mu + e_{ik}) - (\mu + \bar{e})\}^2$$

$$= \frac{1}{n}\sum_{k=1}^{n}(e_{ik} - \bar{e})^2$$

差を測るために統計量

- 仮説を立てる。平均値をμとあらわすと、
  - H0: μ（最近）＝μ（以前）
  - H1: μ（最近）＜μ（以前）　μ（最近）＞μ（以前）
- 違いを測る
  - ｜データの平均値ーμ（以前）｜
    この値が大きければ違いそう
  - データの平均値は変わるので、
    t＝｜データの平均値ーμ（以前）｜／データの平均値の標準偏差：データの平均の標準偏差を標準誤差という
  - 標準誤差＝標準偏差／データの個数

差を測るために統計量

- データに関して真の平均値μ
- データの平均値y
- データの標準偏差s

$$t = \frac{\bar{y} - \mu}{\sqrt{1/n}\hat{\sigma}_y},\ \hat{\sigma}_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- ルール
  - |t|>2ならば、データの平均値yはμとは異なりそう
  - 精密な話は統計学Iで

２グループの比較

- 2つのグループ
- 比較は平均値の差
  - 差は、単位に依存する
  - 差は分布する
- 仮説H_0:　$\mu_1 - \mu_2 = 0$
  - この仮説の元で$\bar{y}_1 - \bar{y}_2 \approx 0$　が期待される。この期待からどの程度外れるか

質的変数の変換

- 質的変数の値が3以上ある場合
  - 一元配置モデル
- 大まかに捉える
  - 値を2種類に変更してみると捉えやすくなる！
  - このどうな値と対応付けか＝＞これが仮説
- EXCELでは
- =IF(find(変換したいセル,"変換したい値のリスト",1,2)

差を調べる

- 扱っているデータには，真の平均 μ と真の分散 σ^2 は存在すると仮定しよう！
- 2つの分散は同じ値と想定しよう！
- H_0:$\mu_1=\mu_2$
  - 評価したい差 y_1-y_2
  - Z＝(Y_1-Y_2)／(差の標準偏差)
- 分散が既知　　$\sigma^2_{Y1-Y2}=\sigma^2(1/N_1+1/N_2)$
- 分散が未知

$$t=\frac{\bar{Y}_1-\bar{Y}_2}{\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}}$$

$$\hat{\sigma}^2_{\bar{Y}_1-\bar{Y}_2}=\frac{(N_1-1)\hat{\sigma}_1^2+(N_2-1)\hat{\sigma}_2^2}{N_1+N_2-2}(\frac{1}{N_1}+\frac{1}{N_2})$$

差を調べる

- 2つの群からのデータについて、分散が等しいとして、平均が等しいとみなせつかを調べる。
- サンプル数Nが小さい場合（N＜30）
- t＝（平均値の差）／（平均値の標準偏差の推定値）
- |t|を大きい程，平均が異なりそう

$$t=\frac{\bar{y}_1-\bar{y}_2}{\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\sqrt{\frac{ssq_1+ssq_2}{n_1+n_2-2}}}$$

$$ssq_i=\sum_{k=1}^{n_i}(y_{ik}-\bar{y}_i)^2$$

2変数での関連の強さを測る

- xとy　共に量的変数
  - 相関係数や共分散
- xかyの一方が質的変数，他方が量的変数
  - 相関比
- t値を計算する

統計量を求める

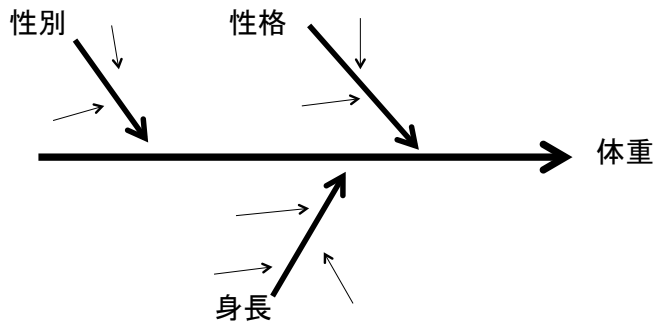
- EXCEL
- 2変数間の関連を探る
  - ピボット表を作成する
- 2変量間の関係を探る
  - 相関係数 =CORREL(範囲x，範囲y)
  - 近似直線
- 変数を比較できるようにする
  - STANDARDIZE(標準化したい範囲，平均値，標準偏差)

体重を予測する：身長がわかると体重はわかるの？

	体重	身長	ウェスト	年齢区分	BMI
1	70	170	74	1	0.24221453
2	63	178	70	1	0.19883853
3	62	170	75	1	0.21453287
4	60	167	76	1	0.21513859
5	58	176	70	2	0.18724174
6	80	182	95	1	0.24151673
7	75	175	75	1	0.24489796
8	73	173	80	2	0.24391059
9	80	178	75	1	0.25249337
10	65	166	85	1	0.23588329
11	58	170	78	2	0.20069204
12	85	178	85	1	0.26827421
13	72	176	72	1	0.23243802
14	75	168	88	2	0.26573129
15	62	173	75	2	0.20715694
16	70	165	88	2	0.25711662
17	68	170	85	2	0.23529412
18	54	170	70	2	0.18685121

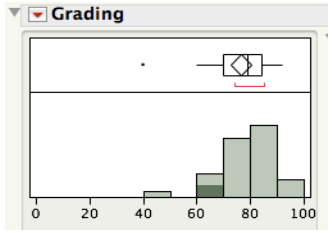
どの要因が「体重」をきめるか

- 特性要因図（Fish―Bone）



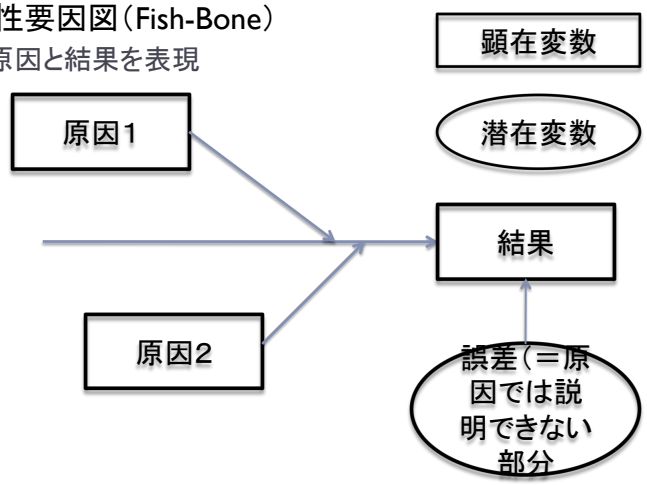
レポートページ数から成績を予測する

	Pages	Grading
1	10	80
2	8	85
3	8	78
4	6	70
5	7	68
6	12	82
7	8	76
8	15	80
9	8	70
10	10	90
11	9	82
12	9	80
13	7	74



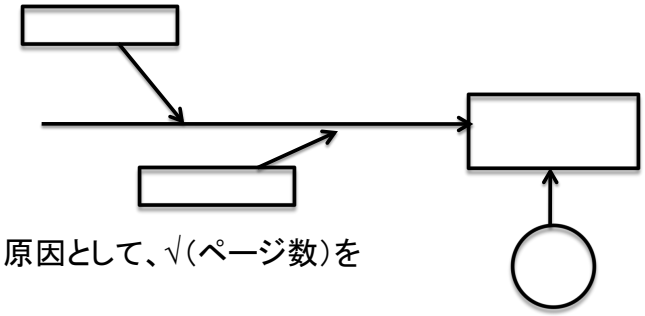
原因から結果を表現

- 特性要因図（Fish-Bone）
  - 原因と結果を表現



レポートページ数から成績を予測する

- 特性要因図を完成しなさい



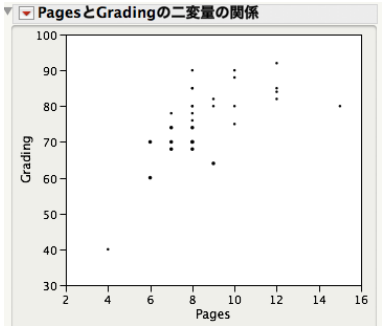
- A君は、原因として、√(ページ数)を考えた。
  - これが妥当とした場合の理由
  - これが妥当でないとした場合の理由

散布図

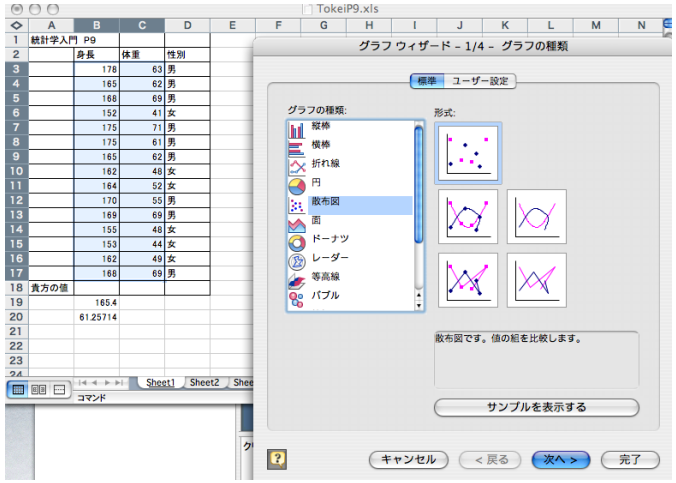
- 因果関係
  - ページ数が原因、
  - 成績が結果
- 相関関係
  - 原因系と結果形が区別できない場合
- 相関係数

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

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

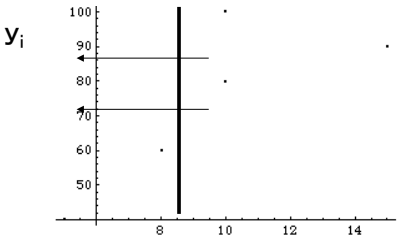


散布図を描く



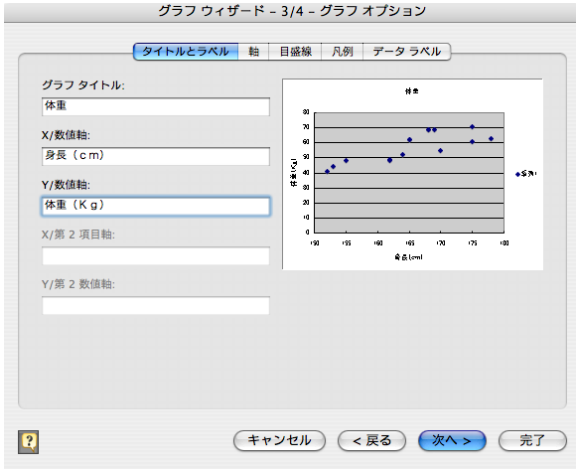
二変量散布図

- 因果が想定される場合には、x軸が原因、y軸が結果
  - (1)xの値が決まる
  - (2)その垂直延長上にyの値
  - (3)y=f(x)よりyの値は
    モデル上では1つ



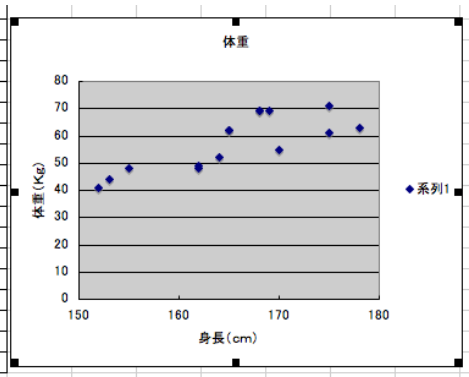
(1)

グラフー散布図（１）



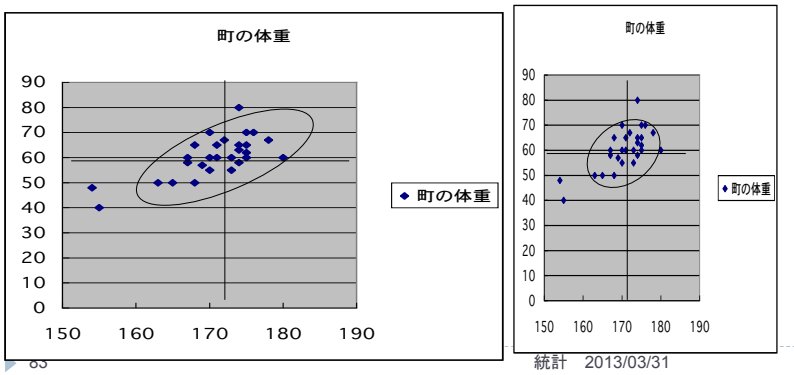
グラフー散布図（２）

	身長	体重	性別
	178	63	男
	165	62	男
	168	69	男
	152	41	女
	175	71	男
	175	61	男
	165	62	男
	162	48	女
	164	52	女
	170	55	男
	169	69	男
	155	48	女
	153	44	女
	162	49	女
	168	69	男
両方の値	165.4		
	61.25714		



２つの変量間の関係

- 体重×身長
  - 単位を変えると，図も変わるが
  - 単位を変えても，２変量間の関係は変わらないので



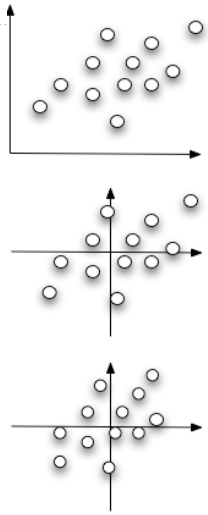
変数変換と線形関係

- 基本は積和平均

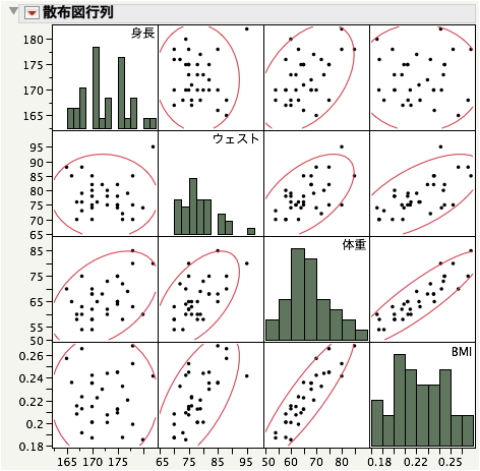
$$[x_i \quad y_i] \Leftrightarrow \frac{1}{n}\sum_{i=}^{n} x_i y_i$$

$$[x_i - \bar{x} \quad y_i - \bar{y}] \Leftrightarrow \frac{1}{n}\sum_{i=}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\left[\frac{x_i - \bar{x}}{s_x} \quad \frac{y_i - \bar{y}}{s_y}\right] \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$



散布図行列　データは身体データ



相関係数と共分散

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

$$= \frac{1}{n}(\tilde{\mathbf{x}} - \tilde{\mathbf{y}})^t(\tilde{\mathbf{x}}^t\tilde{\mathbf{x}})^{-1/2}(\tilde{\mathbf{y}}^t\tilde{\mathbf{y}})^{-1/2}(\tilde{\mathbf{x}} - \tilde{\mathbf{y}})$$

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

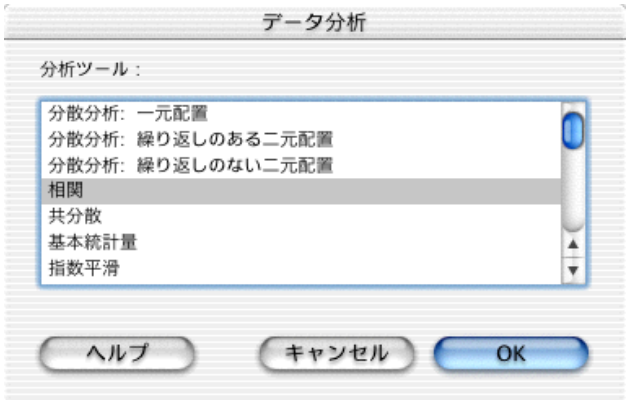
$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

相関係数の計算式

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

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

相関行列を計算する



相関係数行列

	身長	町の体重	本当の体重	今の体重	理想体重
身長	1				
町の体重	0.706809167	1			
本当の体重	0.638682378	0.977072014	1		
今の体重	0.005702898	-0.489168964	-0.548471231	1	
理想体重	0.877789601	0.876694713	0.841201539	-0.103638863	1

身体データの相関係数

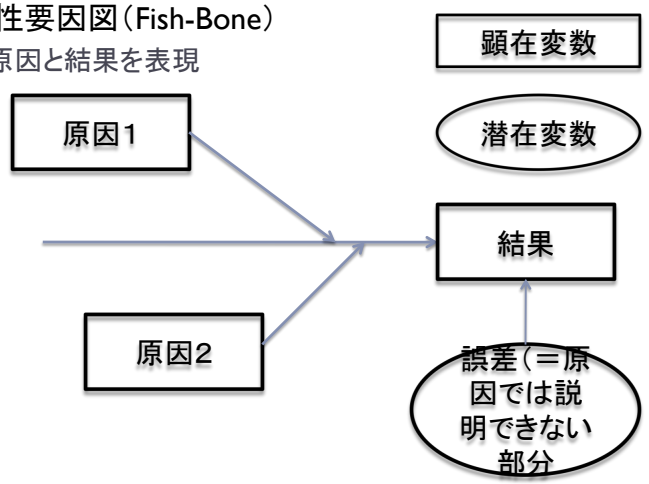
相関

	身長	ウェスト	体重	BMI
身長	1.0000	-0.0342	0.4346	-0.0096
ウェスト	-0.0342	1.0000	0.5669	0.6484
体重	0.4346	0.5669	1.0000	0.8951
BMI	-0.0096	0.6484	0.8951	1.0000

- 何故、相関係数を求めるのか
  - どんな関係を想定しているのか
  - 共分散もあるが

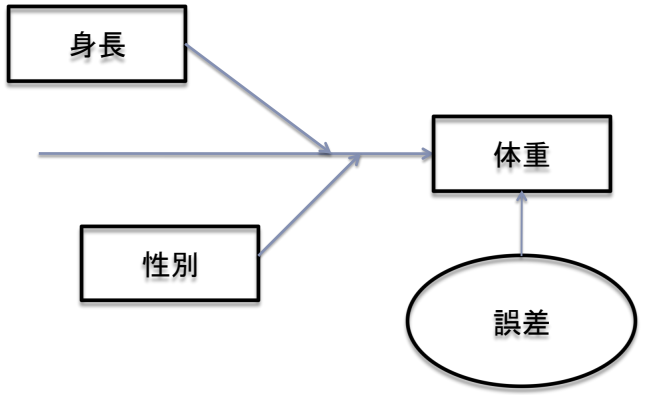
因果関係を考える　原因から結果を表現

- 特性要因図(Fish-Bone)
  - 原因と結果を表現



原因から結果を表現

- 特性要因図(Fish-Bone)
  - 原因と結果を表現



変数について

- 変数には3種類
  - 量的変数（数値）
  - 質的変数（順序）
  - 質的変数（分類）
- 説明変数（x）と応答変数（被説明変数，目的変数）（Y）
  （独立変数と従属変数）x

原因が先（因果関係）

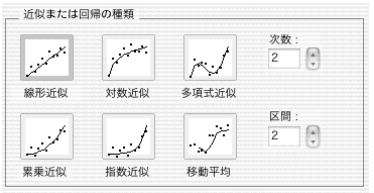
- 原因系はxで表記
- 結果系はYで表記

$$Y = f(x_1, x_2, \cdots, x_p) + e$$

- 仮説の定義式
- Yとxの種類により表現が異なる。

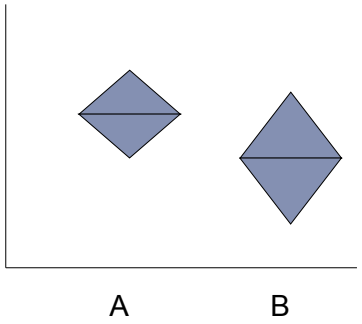
関係を示す線を引く

- 散布図の点のいずれかを左クリック
- グラフ／近似曲線の追加をクリック
- 線形近似をクリック
  - オプションをクリック
  - グラフに数式を表示グラフにＲ２乗値を表示する。



一元配置

- x軸は原因（カテゴリー）、y軸は結果
  - センターは平均
  - 両端は平均値＋／－2標準誤差
- 2つのグループに違いがあれば、平均は菱形の外



Ｙ：数値　vs　ｘ：数値

- 散布図を描く
- 目的
  - xの値がわかった場合にYを説明する、予測する
- チェックポイント
  - 関数関係はないか
- TIPS
  - xが時系列の場合、平滑化なども有効

Ｙ：カテゴリ vs　ｘ：数値

- 目的
  - xがわかった時、Yの分類や程度を説明する、予測する
- 方法
  - Yを数値化して散布図（Yが順序ならば、それを考慮すること）
  - 分割曲線を引く
- TIPS
  - ロジスティック回帰を考える

Ｙ：カテゴリvs ｘ：カテゴリ

- 目的
  - xのカテゴリがわかったとき、Yのカテゴリを予測する、パターンを比較する
- 方法
  - クロス表の作成
  - xでの%和が100となるようにする
- EXCELではピボット表
  - 列和、行和、総合計のいずれで割るかを決めること
  - 関連度係数
    - χ 2乗、γ 係数、ϕ 係数、ソマーズのDなど
- 対応分析も可能

クロス表を扱う

飲酒についての観測度数

	男性	女性
Yes	1630	1684
No	5550	8232

観測度数

	男性	女性
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}

	男性	女性	行計
Yes	$e_{11}=n\times p_{1\cdot}\times p_{\cdot 1}$ 1391.8	$e_{12}=n\times p_{1\cdot}\times p_{\cdot 2}$ 1922.2	$e_{1\cdot}=n\times p_{1\cdot}$
No	$e_{21}=n\times p_{2\cdot}\times p_{\cdot 1}$ 5788.2	$e_{22}=n\times p_{2\cdot}\times p_{\cdot 2}$ 79993.8	$e_{2\cdot}=n\times p_{2\cdot}$
列計	$e_{\cdot 1}=n\times p_{\cdot 1}$	$e_{\cdot 2}=n\times p_{\cdot 2}$	n

クロス表を扱う

- 各セルでの独立からの離れ具合
- $(n_{ij} - e_{ij})^2$
- 独立とした場合のデータは分散に考慮して，
- $\left(\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \right)^2$
- 全てのセルを合わせると，

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}}} \right)^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- この値は自由度（行数−1）（列数−1）＝(n-1)(m-1)の $\chi 2$ 分布に従う．独立でないほど，分割表（クロス表）から求めた値は大きくなる．

Y：数値 vs x：カテゴリ

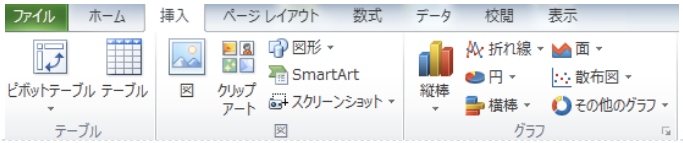
- 目的
  - カテゴリ毎でYへの反応を予測できるか
  - カテゴリ毎でYへの反応には差があるか
- 方法
  - カテゴリを数値化して散布図
  - EXCELではピボット表を作成
  - カテゴリ毎に、第1～3四分位値、平均値、標準偏差を求める
- TIPS
  - 標準誤差は平均値についての標準偏差

グループ毎に平均を求めるには

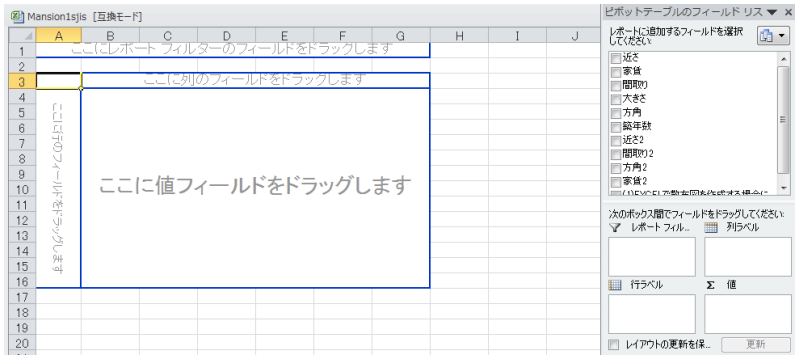
- ここではEXCELのピボットテーブルを使用しよう
- マンションの家賃について間取り毎に集計するなどのように変数×変数について集計する．
- （1）対象となるデータテーブルを選択する
- （2）ピボットテーブルを選択する
- （3）行項目，列項目，集計項目を選択する
- （4）集計項目のフィールドを設定する

ピボットテーブルの選択

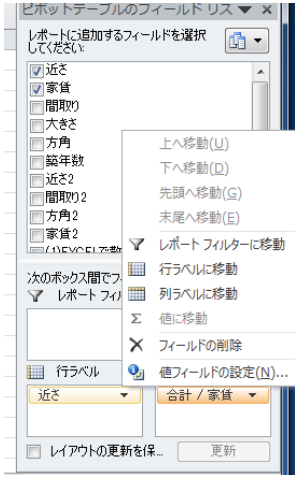
	近さ	家賃	間取り	大きさ	方角	築年数
2	B	68000	1K	19	西	12
3	B	68000	1K	19	南	12
4	B	69000	1K	19	北西	14
5	B	70000	1K	19	南	14
6	B	72000	1K	15	南	9
7	B	77000	1K	20	南	14
8	B	77000	1K	20	西	14
9	A	78000	1K	21	東	15
10	B	79000	1K	22	南	10
11	B	79000	1K	22	南	10



ピボットの選択

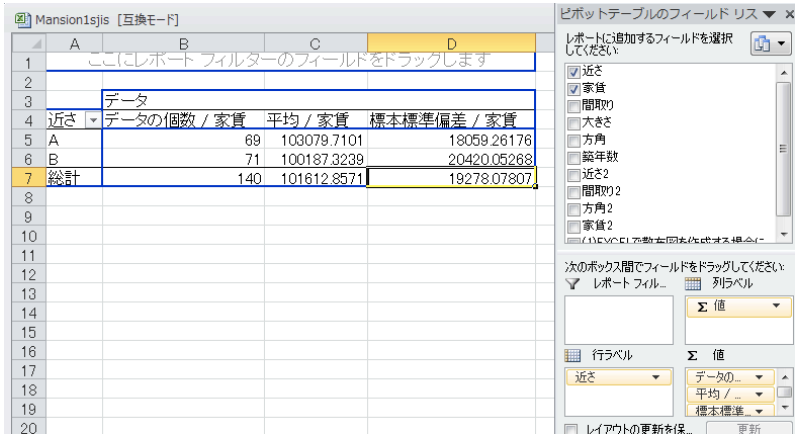


ピボットで集計項目を選択



家賃を選択して，
　合計をフィールドの設定でデータの個数へ，
家賃を選択して集計へ，
　このフィールドを平均へ
家賃を選択して集計へ，
　このフィールドを標準偏差へ

作成されたピボットテーブル



2変量間の関係の強さを測る

- 図では散布図、値では相関係数
- データの数（大きさ）nを反映していない。
  - n=10の時の相関係数r=0.8とn=1000の時のr=0.6では
- t値を使う　利用できる状況については統計学Iで

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

$$|t| > 2.0 \sim 3.0$$

量的変数と質的変数間の関係を測る

- 質的変数が2値の場合　その平均と分散は

$$x_i = \begin{cases} 1 \\ 0 \end{cases}, n_1 = \sum_{(x_i=1)} x_i, n_0 = \sum_{(x_i=0)} x_i$$

$$p = \frac{n_1}{n}, q = \frac{n_0}{n} = 1 - p$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n} = p$$

$$s_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \frac{1}{n}\{n_1 - n(\frac{n_1}{n})^2\}$$

$$= \frac{n_1 n_0}{n^2} = pq$$

量的変数と質的変数間の関係を測る

- 質的変数が2値の場合　相関係数のようなものは

$$Let \quad \bar{y}_1 = \frac{1}{n_1}\sum_{(x_i=1)} y_i, \bar{y}_0 = \frac{1}{n_0}\sum_{(x_i=0)} y_i,$$

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \bar{y}) \qquad (\because \bar{x}\sum(y_i - \bar{y}) = 0)$$

$$\frac{1}{n}\left[\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\left[n_1\bar{y}_1 - n_1\bar{y}\right] = p(\bar{y}_1 - \bar{y})$$

$$= p\left[\bar{y}_1 - \frac{1}{n}\left(n_1\bar{y}_1 + n_0\bar{y}_0\right)\right] = p(q\bar{y}_1 - q\bar{y}_0) = pq(\bar{y}_1 - \bar{y}_0)$$

$$r_{xy} = \frac{pq(\bar{y}_1 - \bar{y}_0)}{\sqrt{pq}s_y} = \frac{\sqrt{pq}(\bar{y}_1 - \bar{y}_0)}{s_y}$$

量的変数と質的変数間の関係を測る

- 質的変数が2値の場合

$$Y_{0i} \sim (\mu_0, \sigma^2), \bar{Y}_0 = \frac{1}{n_0}\sum_{(x_i=0)} Y_{0i} \sim N(\mu_0, \frac{1}{n_0}\sigma^2),$$

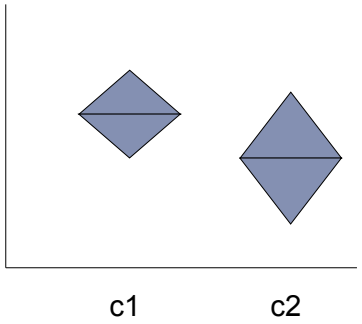
$$Y_{1i} \sim (\mu_1, \sigma^2), \bar{Y}_1 = \frac{1}{n_1}\sum_{(x_i=1)} Y_{1i} \sim N(\mu_1, \frac{1}{n_1}\sigma^2)$$

$$\bar{Y}_1 - \bar{Y}_0 \sim N(\mu_1 - \mu_0, (\frac{1}{n_0} + \frac{1}{n_1})\sigma^2)$$

$$t = \frac{\bar{y}_1 - \bar{y}_0}{s_y\sqrt{\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}} \sim t(n_0 - 1 + n_1 - 1)$$

一元配置

- x軸は原因（カテゴリー）、y軸は結果
  - センターは平均
  - 両端は平均値＋／－2標準誤差
- 2つのグループに違いがあれば、平均は菱形の外



線形モデル

- 目的変数(Y)と説明変数(x)

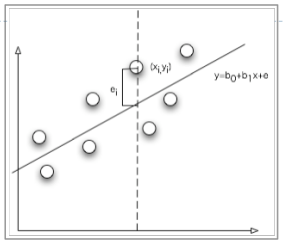
$$y_i = b_0 + b_1 x_i + e_i$$

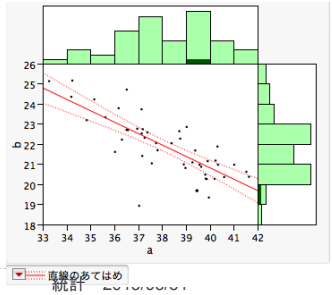
- 最小2乗解

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (b_0 + b_1 x_i)\}^2$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

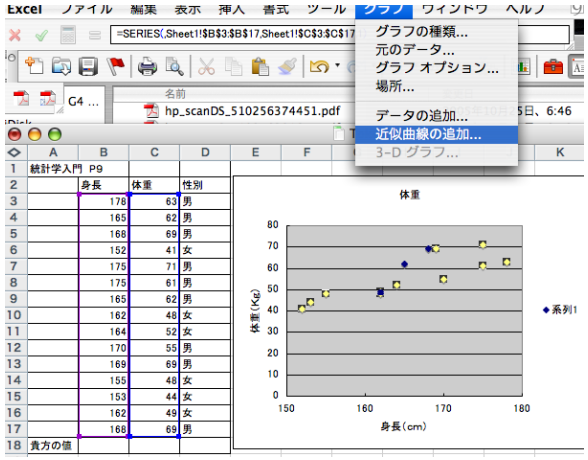
$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$





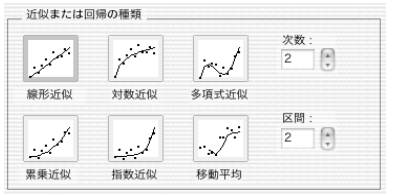
直線を当てはめる

- グラフの点を選択後
- グラフー近似曲線の追加

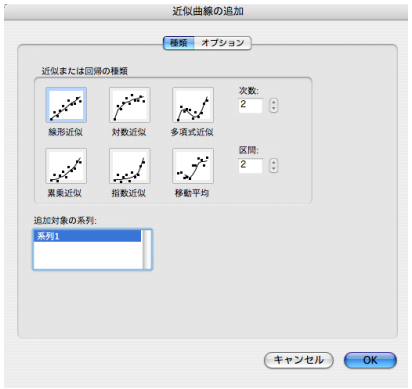


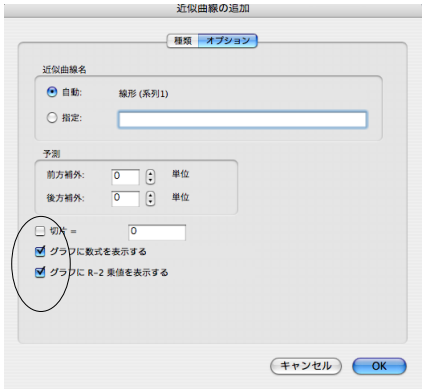
関係を示す線を引く

- 散布図の点のいずれかを左クリック
- グラフ／近似曲線の追加をクリック
- 線形近似をクリック
  - オプションをクリック
  - グラフに数式を表示グラフにR2乗値を表示する。

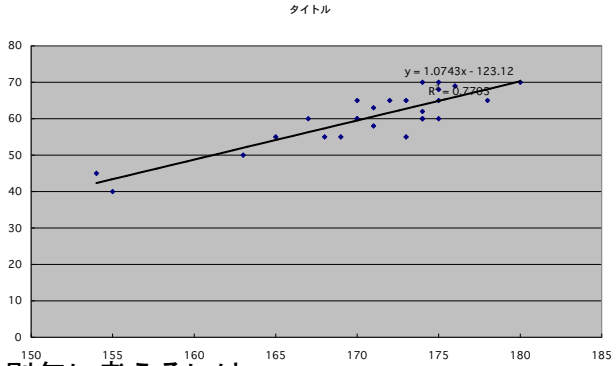


近似曲線の追加メニュー





回帰直線(EXCEL)



- 性別毎に考えるには

直線式の１つの求め方

- 次の式を最小とする(b_0, b_1)の組

$$\min_{(b_0,b_1)}\{\sum_{i=1}^{N}(y_i-b_0-b_1x_i)^2\}$$

- 求めると

$$b_0=\bar{y}-b_1\bar{x}$$

$$b_1=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

変動と決定係数

- データの変動

$$y_i=\hat{y}_i+e_i$$

$$y_i-\bar{y}=(\hat{y}_i-\bar{y})+e_i$$

$$T=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\{(\hat{y}_i-\bar{y})+e_i\}^2$$

$$if\ \ \hat{y}\ \ is\ \ the\ l.s.e.,\ \ then\ \sum_{i=1}^{n}\hat{y}_ie_i=0,\ \bar{\hat{y}}=\bar{y},\bar{e}=0$$

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2+\sum_{i=1}^{n}(e_i-\bar{e})^2$$

変動と決定係数

- 決定係数R^2

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2+\sum_{i=1}^{n}(e_i-\bar{e})^2$$

$$1=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}+\frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=R^2+(1-R^2)$$

分散分析表

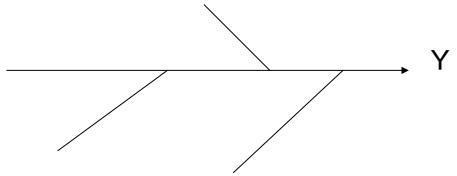
- 変動と分散を区別する！

要因	自由度	変動	分散 =変動／自由度	F 比
回帰モデル	説明変数の個数 p	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	$s_{\hat{y}}^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{p}$	$F=\frac{s_{\hat{y}}^2}{s_e^2}$
残差	（データの大きさ-1）-説明変数の個数 (n-1)-p	$\sum_{i=1}^{n}e_i^2$	$s_e^2=\frac{\sum_{i=1}^{n}e_i^2}{(n-1)-p}$	
データ	データの大きさ-1　n-1	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$s_y^2=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$	

回帰分析：分析ステップ

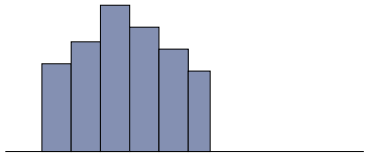
- （1）特性要因図を作成する
- （2）応答変数Yのヒストグラムを作成する
- （3）説明変数xとYとの散布図を作成する
- （4）データをスクリーングする
- （5）単回帰分析を行う
  - エラーバを作成する
- （6）（Y、x）について相関係数行列を作成する
- （7）モデルの選択をする
- （8）予測値とyとの散布図を描く

(1)特性要因図を描く



- なぜ、このような要因モデルを考えたか

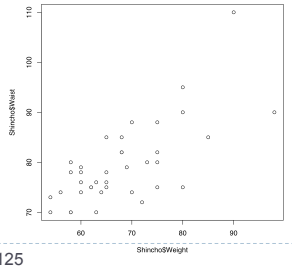
(2)Yのヒストグラムを描く

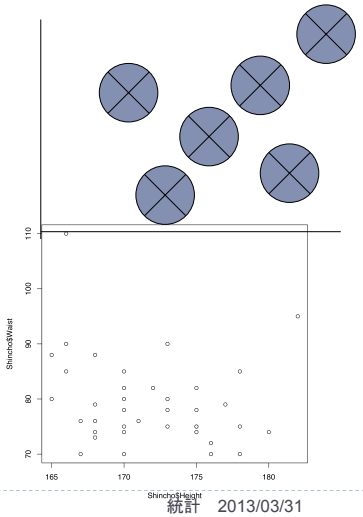


- これからわかること
  - 母集団（グループ）は1つか複数か
- 正規分布と比べてどうか

(3)Ｙとｘの散布図

- xがわかったときのYの関係を推測する
- おかしそうなデータを見つける
- 線形モデルの適合を考えられるか

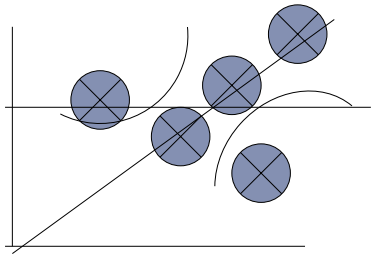


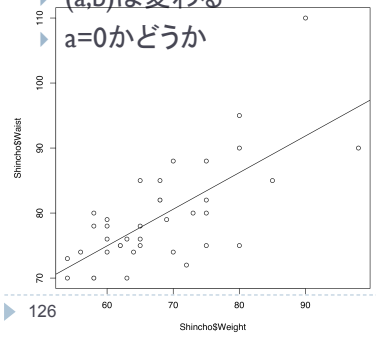


(4)単回帰分析を行う

- xがわかったときのYの関係を推測する
- Y＝ax+b+誤差
- (a,b)を求める
  - (a,b)は変わる
  - a=0かどうか

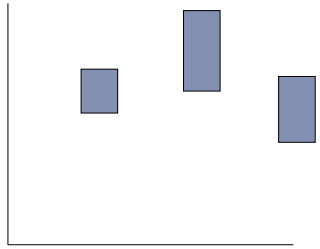
```
> plot(Shincho$Weight,Shincho$Waist)
> abline(lm(Shincho$Waist~Shincho$Weight))
```





(4.2)エラーバー

- xが属性の時
- xをダミー変数化する
- 平均の標準偏差（標準誤差）を考える
- 平均値の分布は？



(5)Ｙとｘの相関係数行列

- Yとx単回帰では相関係数rの2乗が決定係数

	Y	x1	x2
Y	1	0.8	0.6
x1	0.8	1	0.7
x2	0.6	0.7	1

- x間の相関：Yとは相関が高く、x間では相関が低いもの
- x間の相関行列の固有値をチェックする

(6)モデルを選択する

- 全てのxの組み合わせについて考える
- 理論的に除外できないxを入れる
- 決定係数R^2やF比（およびその確率）を目安にする
- 節約的なモデルを採用する
  - 決定係数は分布を仮定しない
  - F比は正規分布と、比較モデルが階層であることを想定

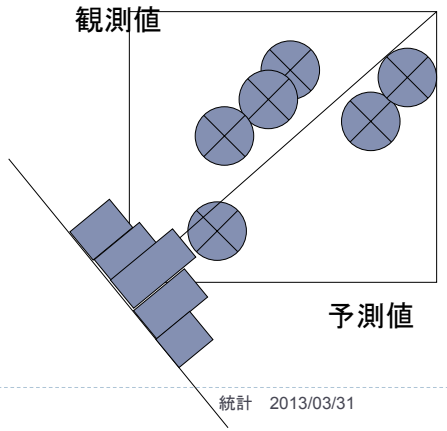
モデル	R^2	F比	AIC
A			
B			
A+B			

(7)回帰係数を評価する

- 偏回帰係数
- 標準化回帰係数（Yとxをそれぞれ標準化）
- t値＝偏回帰係数／係数の標準誤差
- 係数を個別に比較はできない
  - x間の相関を考える

(8)残差分析

- 予測値とYの散布図
- 残差に特定のパターンはないか



例　身体データ（R）

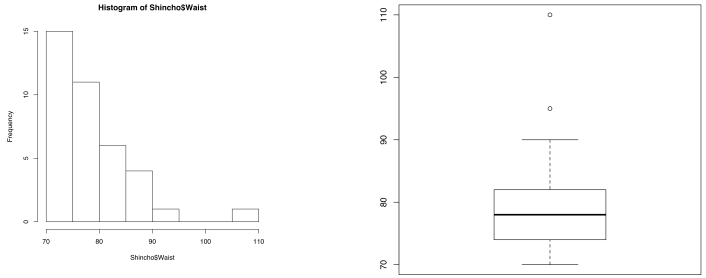
- 身長、体重、ウェストのデータ（自己申告）について、特性要因図（目的変数はウェスト）を作成する。

```
> Shincho<-read.xls(file.choose(),sheet=3)
Converting xls file to csv file... Done.
Reading csv file... Done.
> head(Shincho)
  Height Weight Waist HandSize    Age30
1    170     70    74   Middle Under30
2    178     63    70    Large Under30
3    165     75    80   Middle  Over30
4    170     62    75   Middle Under30
5    167     60    76   Middle Under30
6    176     58    70   Middle  Over30
```

例　目的変数についてのヒストグラム　R

- ヒストグラム

```
> hist(Shincho$Waist)
> boxplot(Shincho$Waist)
> summary(Shincho$Waist)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  70.00   74.25   78.00   79.42   82.00  110.00
```



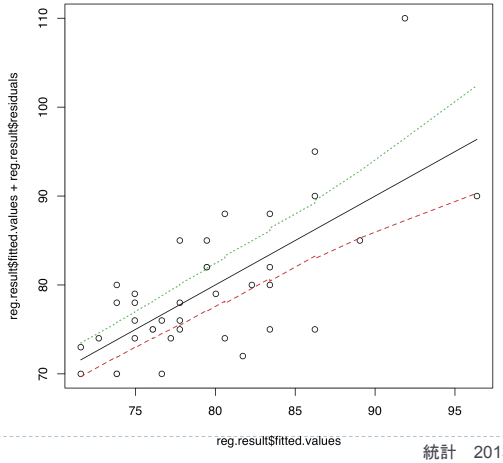
Rで回帰分析

- 関数はlm()

```
> Shincho.lmw<-lm(Shincho$Waist~Shincho$Weight)
> X<-as.matrix(Shincho[,2])
➢summary(Shincho.lmw)
➢Call:
lm(formula = Shincho$Waist ~ Shincho$Weight)
Residuals:
     Min       1Q   Median       3Q      Max
-11.2304  -3.3614  -0.9925   3.5883  18.1323
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    41.13169    6.47317   6.354 2.34e-07 ***
Shincho$Weight  0.56373    0.09431   5.977 7.45e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.75 on 36 degrees of freedom
Multiple R-squared: 0.4981,          Adjusted R-squared: 0.4842
F-statistic: 35.73 on 1 and 36 DF,  p-value: 7.448e-07
```

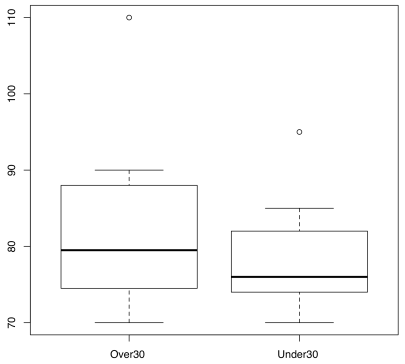

Rで回帰

- 説明変数は体重



Rで箱ひげ図

- plot(x,y)



Rで相関係数

- 相関行列はcor()

```
> cor(Shincho[,1:3])
          Height    Weight      Waist
Height  1.0000000 0.1574595 -0.2000923
Weight  0.1574595 1.0000000  0.7057750
Waist  -0.2000923 0.7057750  1.0000000
 > cov(Shincho[,1:3])
          Height     Weight     Waist
Height 20.023471   7.061878 -7.167852
Weight  7.061878 100.453058 56.628734
Waist  -7.167852  56.628734 64.088193
```


Rで回帰分析

- 結果

```
➢summary(Shincho.lm)
➢Call:
lm(formula = Shincho$Waist ~ Shincho$Height + Shincho$Weight)
Residuals:
    Min      1Q  Median      3Q     Max
-8.1543 -2.0749 -0.2670  2.5533 14.1295
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   136.47294   32.97272   4.139 0.000208 ***
Shincho$Height  -0.57095    0.19429  -2.939 0.005802 **
Shincho$Weight   0.60387    0.08674   6.962 4.28e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.222 on 35 degrees of freedom
Multiple R-squared: 0.5974,     Adjusted R-squared: 0.5744
F-statistic: 25.97 on 2 and 35 DF,  p-value: 1.215e-07
```


分散分析

- 分散分析表　ここでは説明変数毎になっていることに注意

```
> anova(Shincho.lm)
Analysis of Variance Table

Response: Shincho$Waist
             Df  Sum Sq Mean Sq F value    Pr(>F)
Shincho$Height  1   94.94   94.94   3.481   0.07048 .
Shincho$Weight  1 1321.75 1321.75  48.463 4.283e-08
***
Residuals      35  954.57   27.27
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

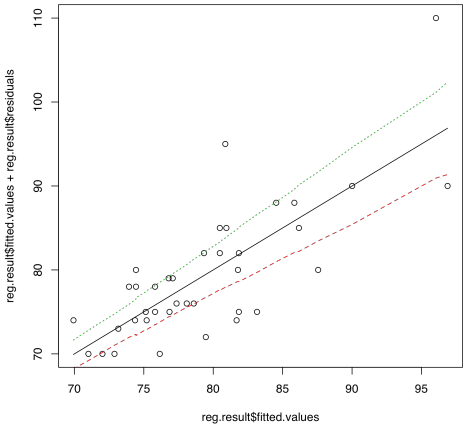

変数選択法

- Rではstep()：候補モデルの中でAIC最小のモデルを採用

```
> step(Shincho.lm)
Start:  AIC=128.5
Shincho$Waist ~ Shincho$Height + Shincho$Weight

                 Df Sum of Sq     RSS    AIC
<none>                         954.57 128.50
- Shincho$Height  1    235.52 1190.09 134.88
- Shincho$Weight  1   1321.75 2276.33 159.52

Call:
lm(formula = Shincho$Waist ~ Shincho$Height + Shincho$Weight)

Coefficients:
   (Intercept)  Shincho$Height  Shincho$Weight
      136.4729         -0.5709          0.6039
```

Rで回帰分析

- 結果

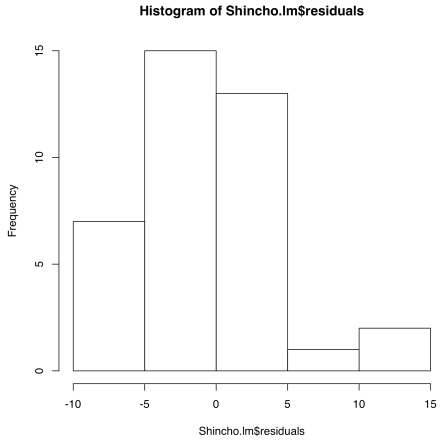


回帰分析の結果

```
> names(RegModel.1)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

coefficients   偏回帰係数
residuals      残差
fitted.values    予測値
effects,rank,qr   線形モデルの影響
```

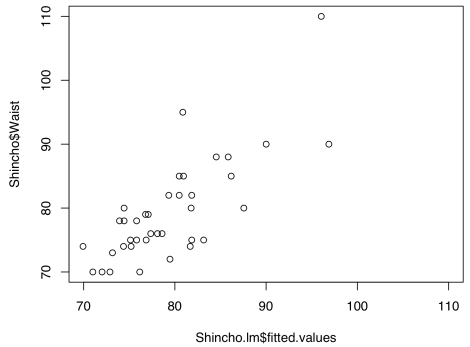

残差

- $residuals
- hist(xx$residuals)



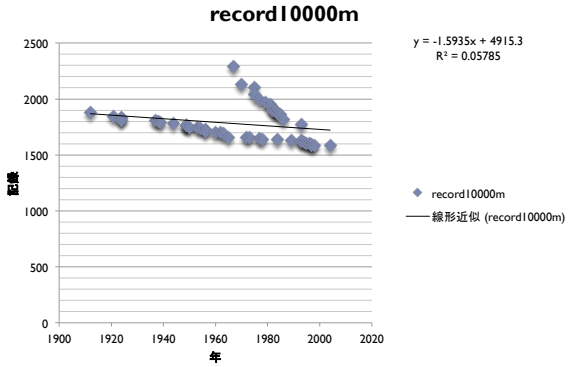
Scatter

```
plot(Shincho.lm$fitted.values,Shincho
$Waist,xlim=c(70,110),ylim=c(70,110))
>
```



不適切な例

- 10000mの男女別記録にあてはめると



演習　不適切な例

- データ　10000mの男女別記録
- どのようなモデルが想定できますか
  - 何故，その要因を取り上げましたか
- 分析して，自チームが考えたモデルについて検討してみなさい

ロジステック回帰

- y_iの値が0／1である場合

$$P(Y=1|\theta)=\frac{1}{1+e^{ax_i+b}}$$

- y=ln{p/(1-p)}=ax+b
  - 標本比率pを用いて分析
  - y_i=1 => ln{p/(1-p)}
  - y_i=0 => 0

線形回帰モデル

- 誤差は正規分布とする。

$$Y_i=\hat{y}(x_i)+e_i,\ \ e_i\sim N(0,\sigma^2)$$

$$\hat{y}(x_i)=\beta_0+\beta_1x_{i1}+\cdots+\beta_px_{ip}$$