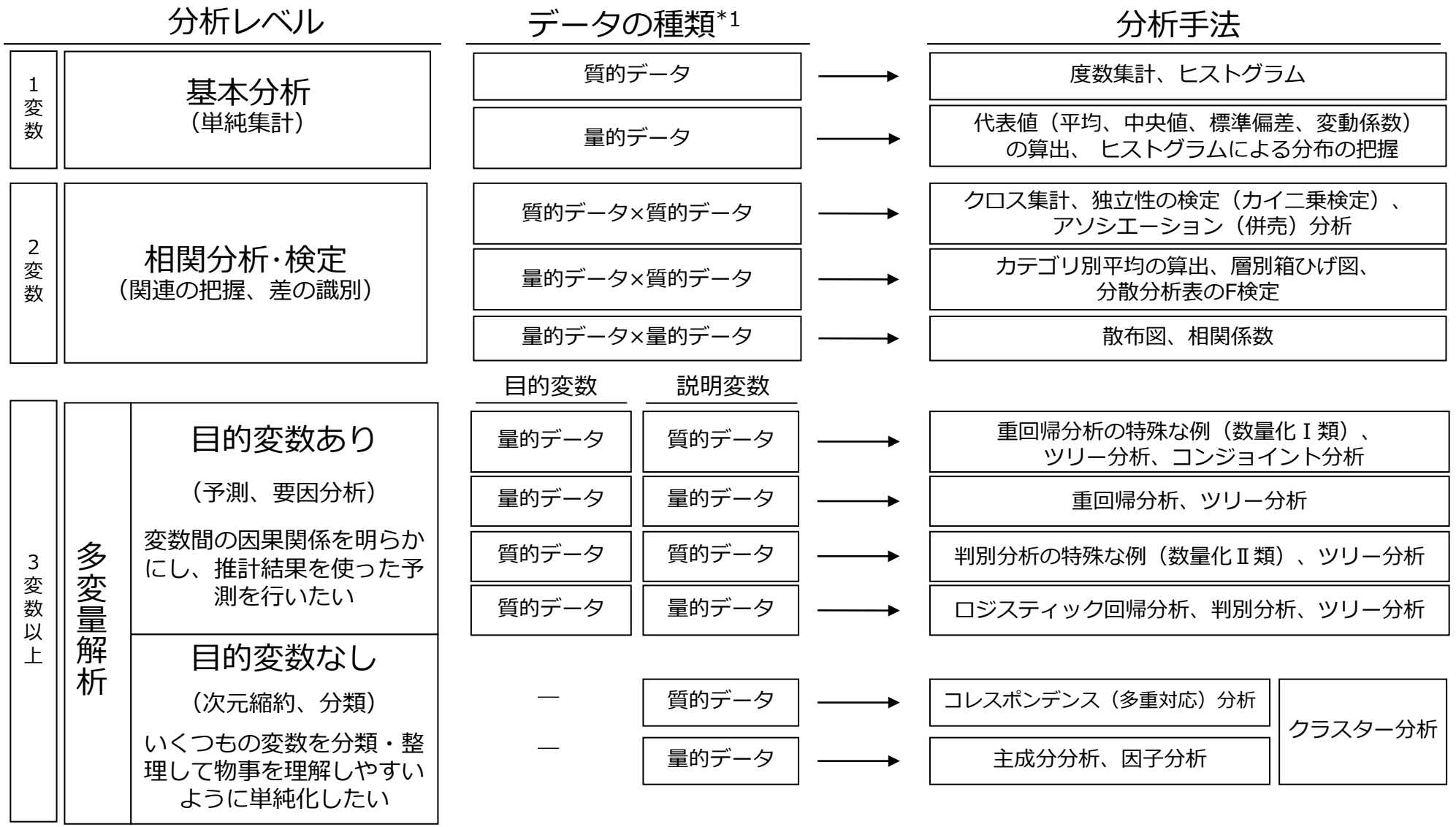
データ解析基礎研修

Ⅱ

目次

代表的な分析手法について（再掲）



*1 ： 集計、解析のためのデータの大分類。 量的データは、データ間の大小を比較したり演算を行った時に意味のあるデータ (例、年収)
質的データは、データ間の大小比較や演算が無意味なデータ(例、血液型)

Ⅴ．データの関連性を把握する

1．２変数の関連の把握

2変数の解析は、2つの項目（変数）の組み合わせで、関連性の強さを確認し、
データの傾向を把握するステップです。

１変数の解析

＜各項目の特徴を把握する＞
- 数値化（要約統計量）と視覚化（ヒストグラム）でデータの傾向（平均、ばらつき、異常値など）を把握する
- 項目ごとにデータの使用可否を判断する

２変数の解析

＜2つの項目間の関連を把握する＞
- 選択した分析手法に適切なデータであるか判断を実施
- 類似項目（関連性の高い項目）については、取捨選択もしくは合成変数作成などを実施する

多変量解析

＜3項目以上を使った分析を実施＞
- 分析の目的を達成するために、予測や要因特定、分類などを実施する
- 常識や業務知識などを元に、結果を判断する

業務やビジネスに適用できる知見へ

代表的な分析手法について（再掲）

分析レベル			データの種類*1		分析手法	
1変数	基本分析（単純集計）		質的データ	→	度数集計、ヒストグラム	
			量的データ	→	代表値（平均、中央値、標準偏差、変動係数）の算出、 ヒストグラムによる分布の把握	
2変数	相関分析･検定（関連の把握、差の識別）		質的データ×質的データ	→	クロス集計、独立性の検定（カイ二乗検定）、アソシエーション（併売）分析	
			量的データ×質的データ	→	カテゴリ別平均の算出、層別箱ひげ図、分散分析表のF検定	
			量的データ×量的データ	→	散布図、相関係数	
			目的変数	説明変数		
3変数以上	多変量解析	目的変数あり（予測、要因分析） 変数間の因果関係を明らかにし、推計結果を使った予測を行いたい	量的データ	質的データ	→	重回帰分析の特殊な例（数量化Ⅰ類）、ツリー分析、コンジョイント分析
			量的データ	量的データ	→	重回帰分析、ツリー分析
			質的データ	質的データ	→	判別分析の特殊な例（数量化Ⅱ類）、ツリー分析
			質的データ	量的データ	→	ロジスティック回帰分析、判別分析、ツリー分析
		目的変数なし（次元縮約、分類） いくつもの変数を分類・整理して物事を理解しやすいように単純化したい	—	質的データ	→	コレスポンデンス（多重対応）分析
			—	量的データ	→	主成分分析、因子分析

クラスター分析

*1 ： 集計、解析のためのデータの大分類。 量的データは、データ間の大小を比較したり演算を行った時に意味のあるデータ (例、年収)
質的データは、データ間の大小比較や演算が無意味なデータ(例、血液型)

１．２変数の関連の把握

2つの項目間の関連を把握するためには、データの測定尺度ごとに以下のようにグラフ化、数値化方法があります。

1.量的データどうしの傾向把握（散布図、相関係数）

2.質的データどうしの傾向把握（モザイク図クロス集計、カイ二乗検定）

3.量的データと質的データの傾向把握（箱ひげ図、層別散布図、カテゴリ別統計量）

2．量的データ×量的データ

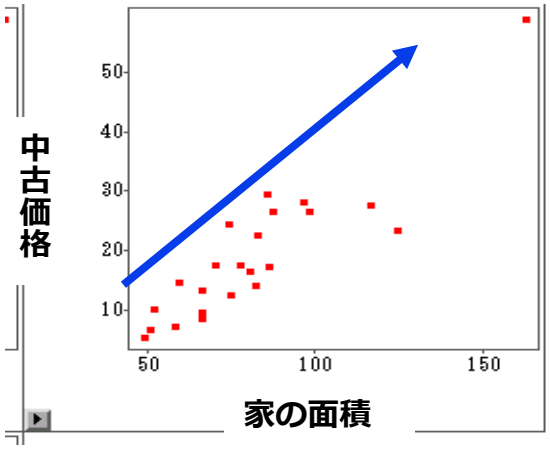
数量データどうしの関連を把握する手法には、散布図（グラフによる把握）と相関係数（数値による把握）があります

以下の中古住宅データを使って説明します

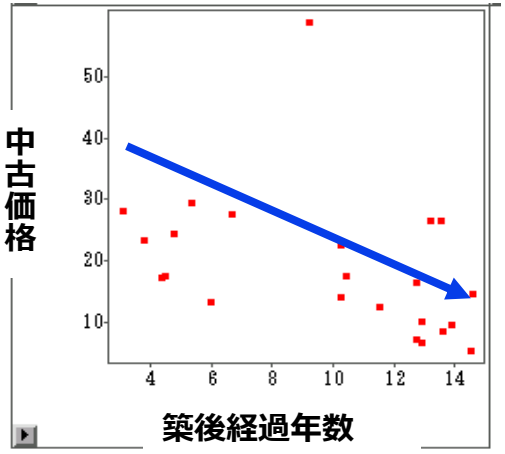
宅地面積(㎡)	住宅延べ面積(㎡)	築後経過年数(年)	東京駅からのJR電車時間(分	JR駅前からのバス時間(分	徒歩時間(分)	建物構造	中古価格(百万円)
98.4	74.2	4.8	5	15	6	1	24.8
379.8	163.7	9.3	12	0	12	1	59.5
58.6	50.5	13	16	15	2	1	7
61.5	58	12.8	16	12	1	1	7.5
99.6	66.4	14	16	13	5	1	9.8
76.2	66.2	6	16	23	1	1	13.5
115.7	59.6	14.7	16	10	4	1	14.9
165	98.6	13.6	16	14	2	1	27
215.2	87.4	13.3	16	10	7	1	27
157.8	116.9	6.7	16	13	6	1	28
212.9	96.9	3.1	16	10	5	1	28.5
137.8	82.8	10.3	19	0	20	1	23
87.2	75.1	11.6	23	5	8	1	12.9
139.6	77.9	10.5	23	10	3	2	18
172.6	125	3.8	23	15	5	2	23.7
151.9	85.6	5.4	28	0	4	2	29.8
179.5	70.1	4.5	32	5	2	2	17.8
50	48.7	14.6	37	0	3	1	5.5
105	66.5	13.7	37	4	11	1	8.7
132	51.9	13	37	0	6	1	10.3
174	82.3	10.3	37	0	18	1	14.5
176	86.1	4.4	37	0	10	1	17.6
168.7	80.8	12.8	41	5	2	2	16.8

2．量的データ×量的データ　～散布図～

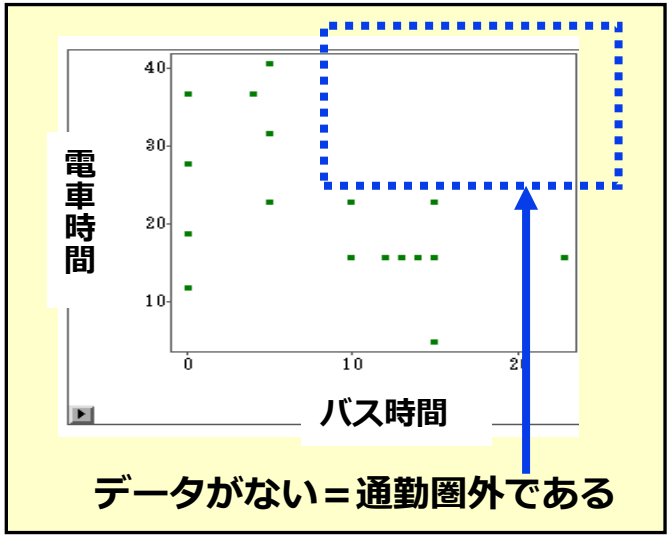
散布図とは、数量どうしの2変数の傾向を視覚的に把握する方法です

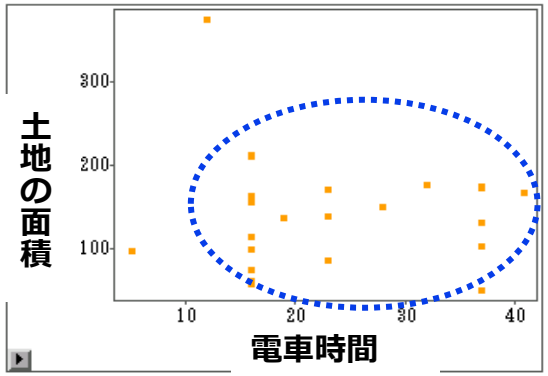


家の面積が大きいと中古価格も高い



築後経過年数が長いと中古価格は低い





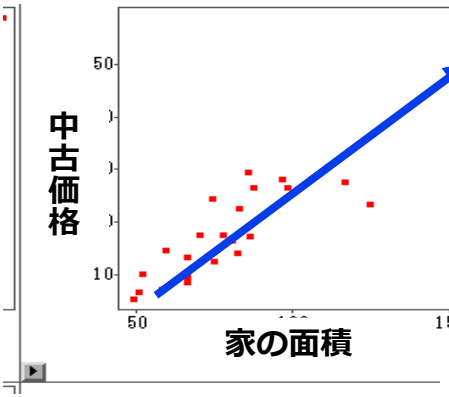
あまり関係がなさそうである

外れ値の存在は、散布図をみることによって発見しやすい

2．量的データ×量的データ　～相関係数～

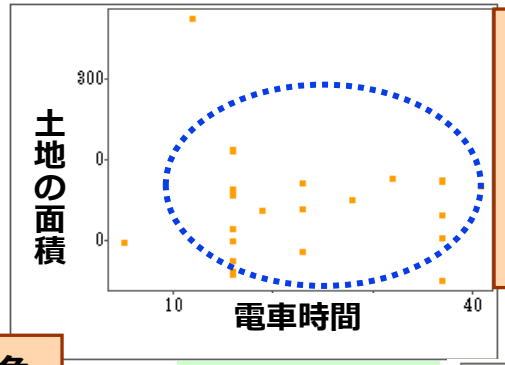
相関係数(Correlation coefficient)とは、二変数の直線的な関係の程度を表す指標です

相関係数は－１～＋１の値をとるように規格化した値
- ■ －１に近いほど、負の相関がある
- ■ 0に近いほど、相関がない
- ■ ＋１に近いほど、正の相関がある



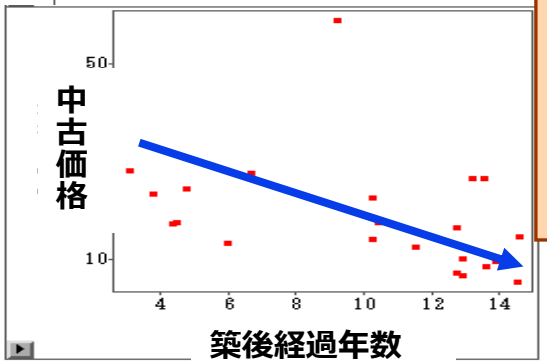
正の相関関係

R=0.8783



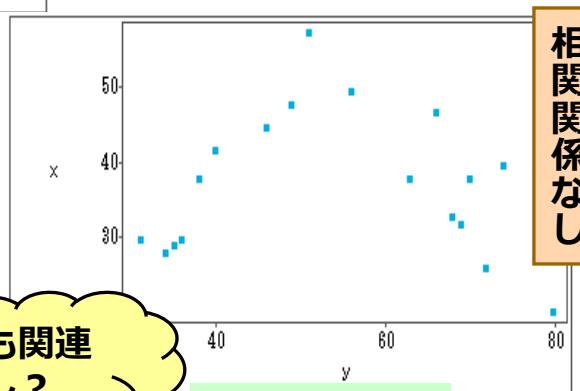
相関関係なし

R=-0.0742



負の相関関係

R=-0.3958



相関関係なし

R＝-0.0709

【参考】量的データ×量的データ　～ピアソンの積率相関係数～

＜ピアソンの積率相関係数＞　一般的に使われる相関係数

例）製造工場の過去10ヶ月の設備費と売上高の関係を調べる

（単位：億円）

No	設備費	売上高
1	13	132
2	19	183
3	18	152
4	12	118
5	15	123
6	11	105
7	11	89
8	15	123
9	15	137
10	17	162

$r_{xy} = 0.92$

設備費と売上高の相関は高い

＜計算法＞

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \bullet \sqrt{\sum (y_i - y)^2}}$$

ピアソンの積率相関係数の符号は、分子（積の項の組み合わせ）により決まります。分母の符号は、２乗の積になるため常に正です。

相関係数は、2変数の直線的な関係を表す指標で、原因と結果のような因果関係は見ていません。
通常、相関係数といえば、ピアソン(Pearson)の積率相関係数を指します。
相関係数は、このほかに順位情報を扱うスピアマンの順位相関係数・ケンドールの順位相関係数というものがあります。

2．相関係数についての補足

無相関の場合、相関係数は0、正の相関が強くなるにしたがって1に近づいていきます。

相関係数がどのような値の場合に「強い相関」「弱い相関」といえるのか明確な基準はありませんが、相関係数の大きさの評価の目安は下記になります。

相関係数	大きさの評価
-0.2 <= r <= 0.2	ほとんど相関なし
-0.4 <= r < -0.2、 0.2 < r <= 0.4	弱い相関あり
-0.7 <= r < -0.4、0.4 < r <= 0.7	中程度の相関あり
-1.0 <= r < -0.7、0.7 < r <= 1.0	強い相関あり

ピアソンの積率相関係数は直線的な関係の強さを表す指標なので、直線になった場合に１（または－１）の値をとります。直線の傾きで値が変わることはありません。

＋α考えてみよう

給与と血圧の相関関係が0.9だったとしたら、血圧をあげれば給与があがる？

!! Point !! 見せかけの相関に注意

たとえ相関係数で強い相関が出た場合でも、見せかけの相関かどうか吟味する必要があります。
給与と血圧の例でいえば、それぞれに共通して年齢との高い相関がある為、結果として給与と血圧も高い相関関係となっていると考えられます。

相関関係と因果関係

分析する際に意識すべきことのひとつに、「データとデータの関連性」があります。
相関関係があるからと言って、必ずしも因果関係があるわけではないことに注意しましょう。

相関関係	因果関係
「Aが増減すると、Bも増減」	「～だから～です」
起こる**順番は問わない**	**先に原因**が起こる
互いに**影響しないこともある**	原因が**結果に直接影響**する
因果関係があるとは**限らない**	出来事に**必ず相関関係**がある

「因果関係」は「相関関係」に含まれますが、
「相関関係」にあるものが、必ずしも「因果関係」にあるとは言えません

相関関係と因果関係

強い相関関係があると、つい因果関係に結びつけたくなりますが、**単なる偶然**や**疑似相関**の場合もあります。

疑似相関とは、「因果関係のない2つの事象であるにもかかわらず、**見えない要因が作用**して因果関係があるかのように推測されてしまうこと」をいいます。

相関関係に関する5つのパターン

パターン	説明
直接原因	一方がもう一方の出来事の直接的な原因になる（＝因果関係）
間接原因	複数の出来事に間接的に因果関係がある
相互作用	お互いに原因にも結果にもなる（＝因果関係）
疑似相関	別の出来事が複数の出来事に影響を与える
偶然の相関	まったく関係ない出来事に相関関係が見つかる

■Question■これは「因果関係」があるでしょうか？

「風が吹けば、桶屋が儲かる」

ある事象の発生により、一見すると全く関係がないと思われる場所・物事に影響が及ぶことの喩え

あてにならない期待をすることの喩え

1. 突風で砂ぼこりが立つ
2. 砂ぼこりが目に入り、視力を失う人が増える
3. 三味線を買う人が増える（※昔は三味線弾きは視覚障がい者の代表的な職業）
4. 三味線の皮の材料として猫の皮が必要になり、猫が捕獲される
5. 猫が減るとねずみが増える
6. ねずみが増えて、かじられる桶が増える
7. 桶の修繕や買い換え需要が増え、桶屋が儲かる

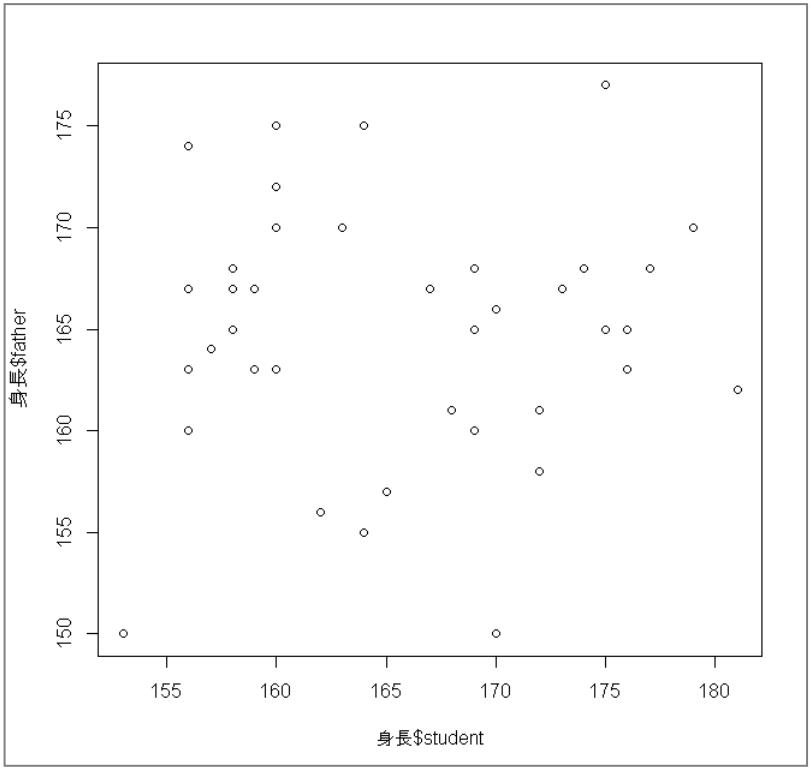
【Rによる実践】散布図（plot関数）

親子身長データheightに対して、生徒の身長と父親の身長の関係を見るために散布図を描きます。散布図を描くためにplot関数を用います。

```
>身長=read.csv("height.csv")
>plot(身長$student,身長$father)
```

plot関数の括弧内の1番目にx軸、2番目にy軸を指定する。

＜散布図＞



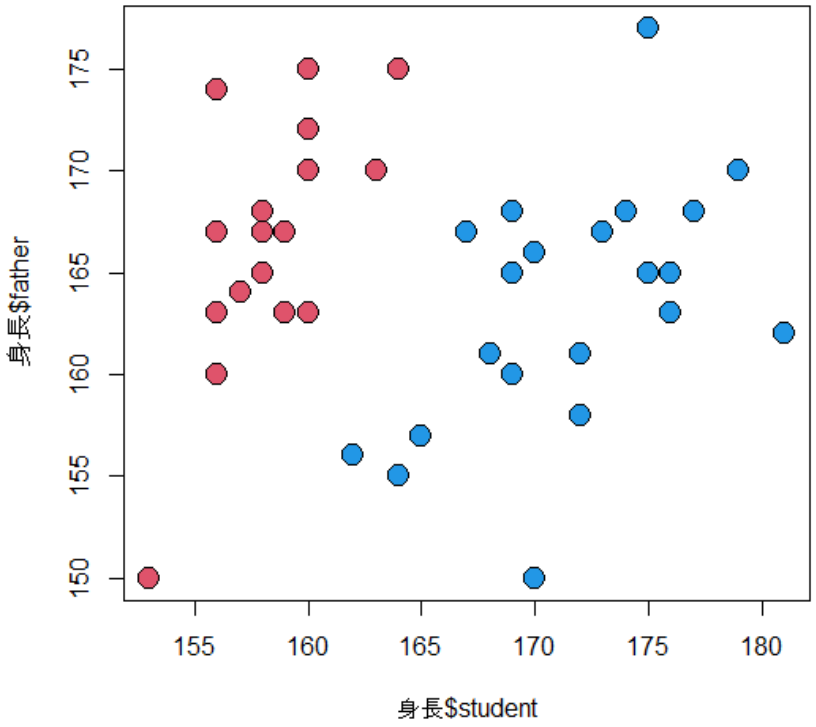
散布図（plot関数）

【Rによる実践】散布図（plot関数）

前頁で描いた散布図を以下のように色分けして表示することも可能です。

```
#性別をcharacter型からfactor型へ変換
>身長$sex=factor(身長$sex)
>plot(身長$student,身長$father,pch=21,cex=2,bg=c(2,4)[unclass(身長$sex)])
```

＜散布図＞



補足：pch=21・・・円型マークを色で塗りつぶす
cex=2・・・プロットサイズを指定
bg=c(2,4)・・・マークを2種類で塗りつぶす（赤、青）
[unclass(身長$sex)]・・・sexの水準毎に塗りつぶし

【Rによる実践】相関係数の算出（cor関数）

heightの生徒の身長、父親の身長について、ピアソンの積率相関係数を求めます。

```
>cor( 身長$student,身長$father )
```

出力結果

```
[1] 0.0227297
```

【欠損値の取扱い】
cor関数は、欠損値がある場合に欠損値を除外しないとエラーになる。オプションとして「use="pairwise"」指定をすることで、欠損値を除外した相関係数が算出される。

生徒の身長と父親の身長の相関係数は0.02程度のため、ほとんど相関がないといえます。

heightの生徒の身長、母親の身長について、ピアソンの積率相関係数を求めます。

```
>cor( 身長$student,身長$mother )
```

出力結果

```
[1] -0.08705174
```

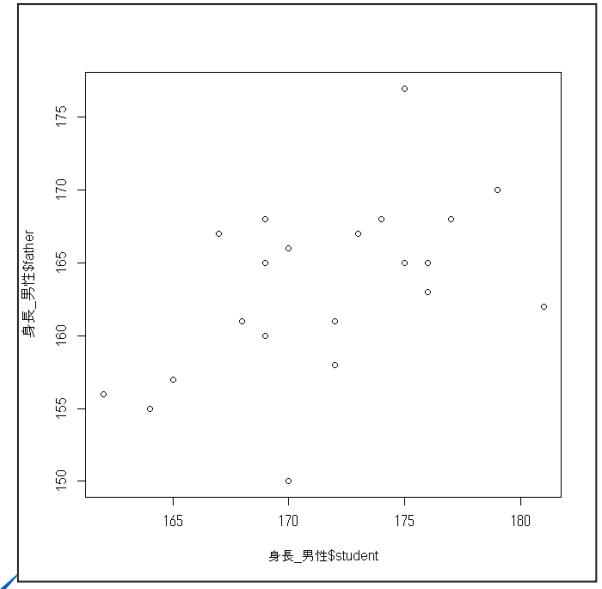
生徒の身長と母親の身長の相関係数は-0.08程度のため、ほとんど相関がないといえます。
また、生徒と母親の身長の散布図をみると相関がないことが確認できます。

【Rによる実践】相関係数の算出（cor関数）

次にheightデータを男性と女性に分けた上で、男子生徒の身長と父親の身長 及び 男子生徒の身長と母親の身長について、散布図とピアソンの積率相関係数を求めます。

＜男性のみ：生徒と父親の身長の関係＞

```
#男性のみサブセット
>身長_男性=subset(身長,sex=="M")
#散布図
>plot(身長_男性$student,身長_男性$father)
#相関係数
>cor(身長_男性$student, 身長_男性$father)
```



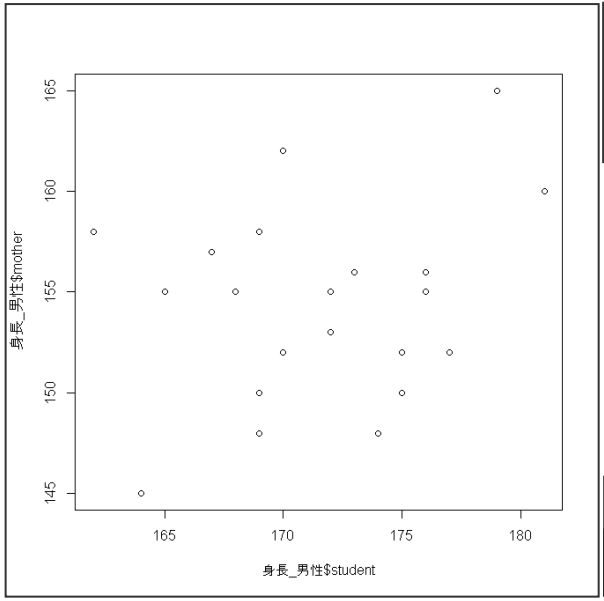
男子生徒の身長と父親の身長の関係に中程度の正の相関がみられる

出力結果

[1] 0.5141091

＜男性のみ：生徒と母親の身長の関係＞

```
#男性のみサブセット
>身長_男性=subset(身長,sex=="M")
#散布図
>plot(身長_男性$student,身長_男性$mother)
#相関係数
>cor(身長_男性$student, 身長_男性$mother)
```



男子生徒と母親の身長の関係に弱い相関がみられる

出力結果

[1] 0.2499194

【練習問題6】散布図・相関係数

野球データ（【2012年】全球団内野手.csv）を使って、翌年年俸と以下の３つの変数それぞれの関係を調べるために、散布図と相関係数を求めてください。

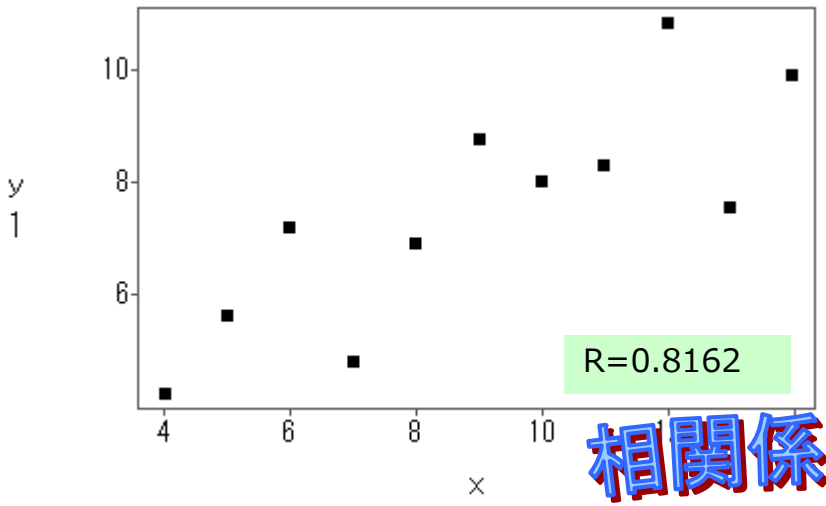
打点、　死球、　犠打

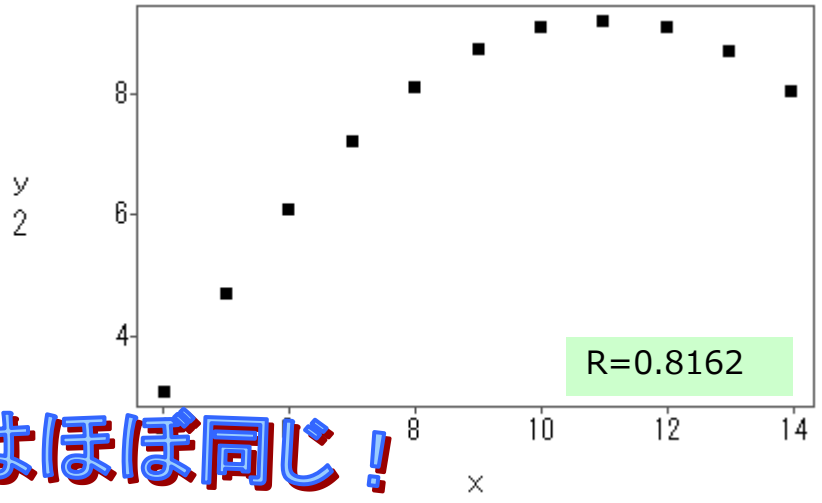
※利用データに欠損値があると相関係数が算出されないため
欠損値を含むレコードを使用しない設定をするオプション
use=“pairwise“を使用します。

例）cor(身長_男性$student, 身長_男性$father, use=“pairwise“)

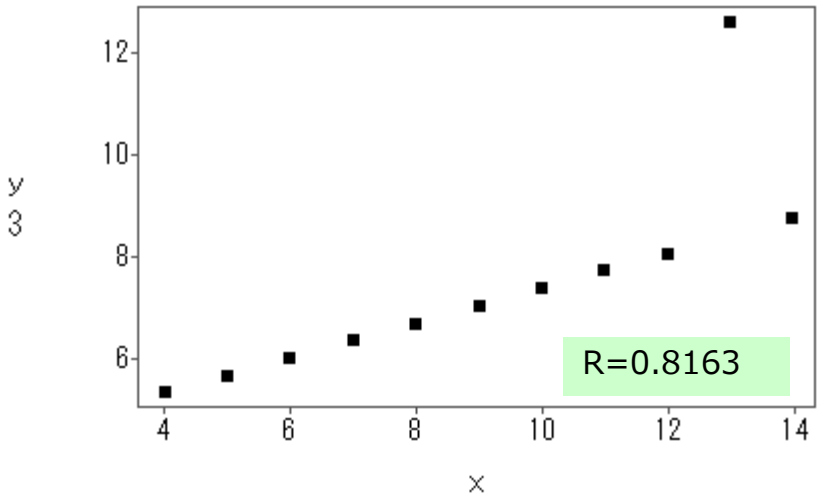
【参考】相関と外れ値

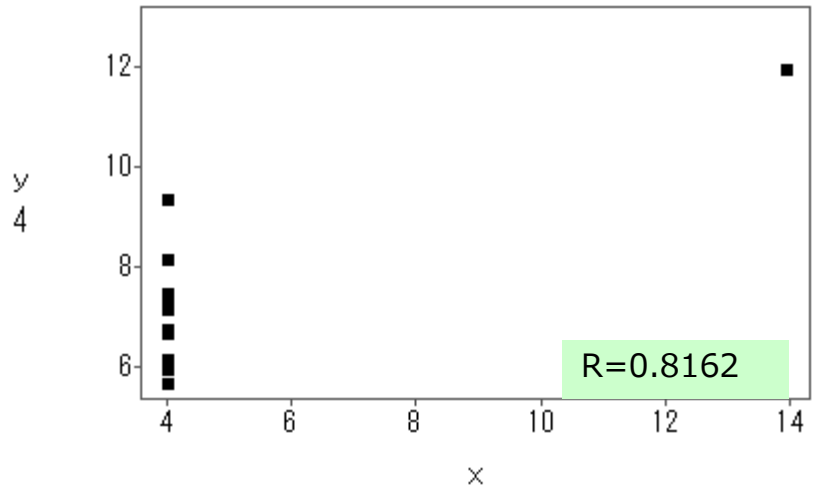
散布図を確認せずに、相関係数のみで同じ関連性の強さと判断するのは危険です。





相関係数はほぼ同じ！



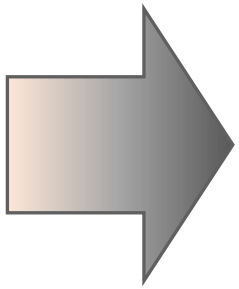


3．質的データ×質的データ　～クロス集計表～

クロス集計表は度数分布表・分割表ともいい、質的データ間の関連を数値で把握する集計表になります。

例）政権を支持する・しないかのアンケートを取りました

性別	支持する･しない
男	支持する
男	支持しない
男	支持しない
男	支持する
女	支持する
男	支持しない
男	支持する
女	支持しない
女	支持する
:	:



度数 パーセント 行のパーセント 列のパーセント	表：siji * gender		
	gender(性別)		
siji(支持する･しない)	女性	男性	合計
支持しない	84 37.67 47.19 86.60	94 42.15 52.81 74.60	178 79.82
支持する	13 5.83 28.89 13.40	32 14.35 71.11 25.40	45 20.18
合計	97 43.50	126 56.50	223 100.00

3．質的データ×質的データ　～カイ二乗検定～

一方の変数の水準によってもう一方の変数のふるまいが異なるとき、２つの変数には関連があるといいます。質的データの行の要因と列の要因に関連があるか統計的に評価するためにカイ二乗値を利用し、変数間の独立性の検定を行うものがカイ二乗検定です。

	男性	女性	行の合計
タバコを吸う	50人	50人	100人
タバコを吸わない	50人	50人	100人
列の合計	100人	100人	200人

	男性	女性	行の合計
タバコを吸う	60人	40人	100人
タバコを吸わない	40人	60人	100人
列の合計	100人	100人	200人

3．質的データ×質的データ　～カイ二乗検定～

例）３カ国（アメリカ、日本、ドイツ）の国民に好きな色を調査したデータがある。
はたして、国と好みの色は関係があるのだろうか？検定してみましょう。

観測値

要因１：国名

要因２：好きな色

	アメリカ	ドイツ	日本	行の合計
赤	4	3	2	9
白	8	6	4	18
緑	6	6	3	15
列の合計	18	15	9	42

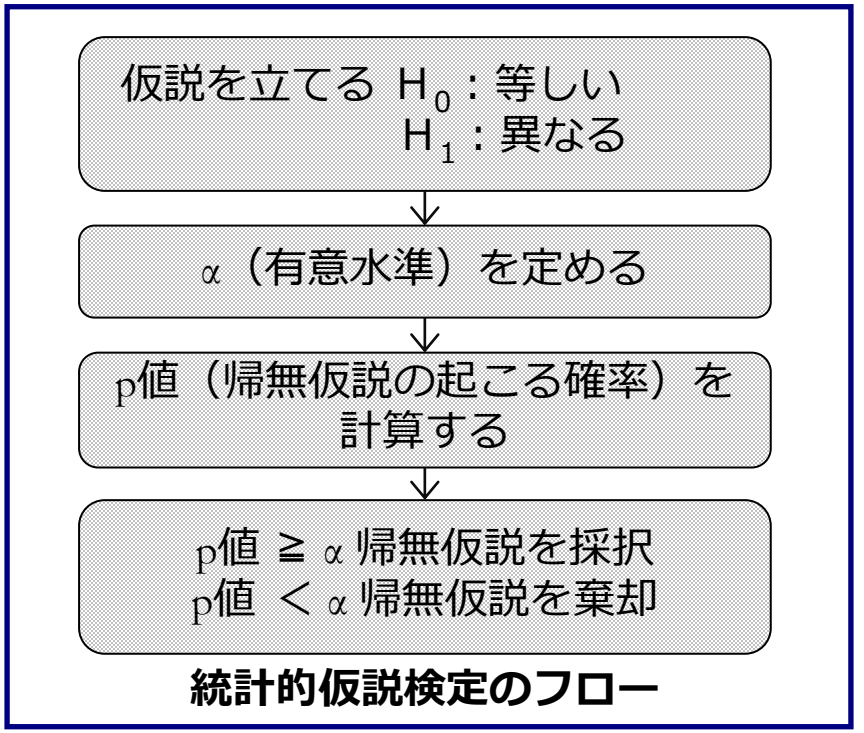
3．質的データ×質的データ　～カイ二乗検定～

カイ二乗検定とは、観測されたデータの分布が理論値の分布とほぼ同じと見なせるかどうかを統計的に判断する手法です。

クロス集計表の全体合計と行・列それぞれの合計値から期待値を求める

期待値と観測値に差があるかどうか検定をおこなう

	アメリカ	ドイツ	日本	行の合計
赤				9
白	**ここの推定値(期待値)を求める**			18
緑				15
列の合計	18	15	9	42



統計的仮説検定のフロー

3．質的データ×質的データ　～カイ二乗検定～

＜期待確率・期待値の算出＞
まず、観測値の行の合計・列の合計・総合計数から、関連が全く無い場合に行・列に期待される確率（期待確率）（$\hat{P}_i$ $\hat{P}_j$）、および関連が全く無い場合に各セルに期待される値（期待値）（E_{ij}）を求めます。

期待確率　$\hat{p}_{i\bullet} = f_{i\bullet}/n$　および　$\hat{p}_{\bullet j} = f_{\bullet j}/n$

期待値　$E_{ij} = n\,\hat{p}_{i\bullet}\,\hat{p}_{\bullet j} = f_{i\bullet}\,f_{\bullet j}/n$

期待確率と行列の合計値から期待値を求める

$f_{i\bullet}$ ＝1行の合計
$f_{\bullet j}$ ＝1列の合計
$\hat{p}_{i\bullet}$ ＝行の期待確率
$\hat{p}_{\bullet j}$ ＝列の期待確率
E_{ij} ＝1セルの期待値
n ＝行列の合計

3．質的データ×質的データ　～カイ二乗検定～

＜期待確率・期待値の算出＞
国と好きな色を調査したデータを用いて、
それぞれの期待値を計算します。

例）赤色を好むアメリカ人の期待確率、期待値を計算

$$p_{i\bullet} = f_{i\bullet} / n = 9/42 = 0.214$$
$$p_{j\bullet} = f_{j\bullet} / n = 18/42 = 0.43$$
$$E_{ij} = n \cdot p_{i\bullet} \cdot p_{\bullet j} = f_{i\bullet} \cdot f_{\bullet j} / n$$
$$= 42 \times 0.214 \times 0.43 = 9 \times 18/42 = 3.86$$

観測値

	アメリカ	ドイツ	日本	行の合計
赤	4	3	2	9
白	8	6	4	18
緑	6	6	3	15
列の合計	18	15	9	42

0.214

0.43

期待値

	アメリカ	ドイツ	日本	行の合計
赤	3.86	3.21	1.93	9
白	7.71	6.43	3.86	18
緑	6.43	5.36	3.21	15
列の合計	18	15	9	42

3．質的データ×質的データ　～カイ二乗検定～
～検定のステップ～

Step 1　独立性の検定の仮説を立てる

H_0：行変数と列変数の間に関連はない。（独立である）：帰無仮説

例）国と好みの色には関連がない。
性別とタバコを吸う・吸わないには関連がない。
性別と政権の支持には関連がない。

H_1：行変数と列変数の間に関連がある。（独立でない）：対立仮説

例）国と好みの色には関連がある。
性別とタバコを吸う・吸わないには関連がある。
性別と政権の支持には関連がある。

Step 2　有意水準αを設定する

Step 3　カイ二乗値を計算し、Ｐ値を求める

（または、有意水準αと自由度から臨界値を求める）

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \sum_j \frac{(\text{実測値} - \text{期待値})^2}{\text{期待値}}$$

Step 4　値を比較し、結論を出す

Ｐ値＞有意水準α　帰無仮説を採択

Ｐ値＜有意水準α　帰無仮説を棄却

または

臨界値＞カイ二乗値　帰無仮説を採択

臨界値＜カイ二乗値　帰無仮説を棄却

３．質的データ×質的データ　～カイ二乗検定～
～計算してみよう～

Step 2　有意水準αの設定　　α＝0.05

Step 3　カイ二乗値を計算し、

$$\chi^2 = \sum_i \sum_j \frac{\left(f_{ij} - E_{ij}\right)^2}{E_{ij}} = \sum_i \sum_j \frac{(\text{実測値} - \text{期待値})^2}{\text{期待値}} = 0.187$$

有意水準αと自由度から臨界値を求める（添付資料 χ^2分布表より）

自由度は？
自由に動ける観測値の数
自由度の計算
（Ｒ－１）×（Ｃ－１）

＜３×３の分割表＞

	アメリカ	ドイツ	日本	行の合計
赤	3.86	3.21	1.93	
白	7.71	6.43	3.86	
緑	6.43	5.36	3.21	
列の合計				

３×３の分割表の場合自由度（自由に動ける観測値）は４である。

３．質的データ×質的データ　～カイ二乗検定～
～計算してみよう～

Step３

- 臨界値
- Ｐ値

を求める

臨界値＝９．４８７
Ｐ値　＝０．９９５

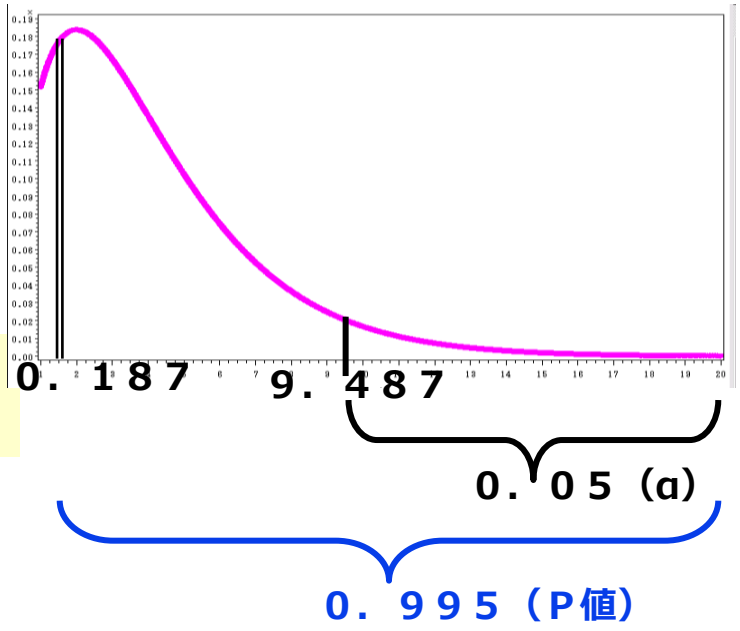
Step４

値を比較し、結論を出す

カイ二乗値＝0.187＜9.487（表より）
Ｐ値＝０．９９５＞０.０５(有意水準α)

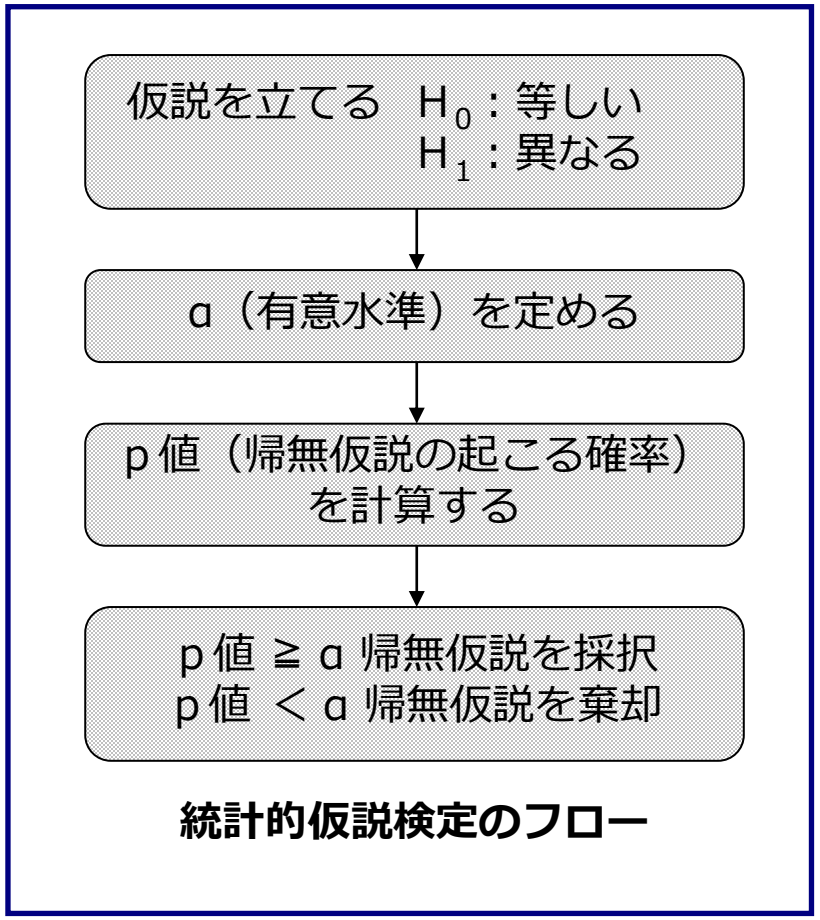
帰無仮説を棄却できず、“国と好きな色（行と列の要因）には関連あるとはいえない”という結論になります。

自由度４の χ^2 分布



3．質的データ×質的データ　～カイ二乗検定～

算出した期待値と観測値に差があるかどうか、カイ二乗値を使用して検定します



統計的仮説検定のフロー

H_0：国と好きな色の間に関連はない
H_1：国と好きな色の間に関連はある

有意水準α ＝ 0.05

p 値 ＝ 0.9959

0.9959（p 値）≧0.05（有意水準α）より、帰無仮説を採択

【結論】国と好きな色の間に関連はないといえる。

【Rによる実践】クロス集計表（table関数）

国別の好きな色のデータcolor_nationについて、国と好きな色のクロス集計表を作成します。

```
>カラー=read.csv("color_nation.csv")
>table(カラー$好きな色,カラー$国)
```

table関数の括弧内は、行変数、列変数の順で変数を指定する。

出力結果

	アメリカ	ドイツ	日本
赤	4	3	2
白	8	6	4
緑	6	6	3

【Rによる実践】カイ二乗検定（chisq.test関数）

国と好きな色の間に関連があるかどうか、chisq.test関数を使用してカイ二乗（χ^2）検定を行います。

```
>クロス集計表=table(カラー$好きな色,カラー$国)
>chisq.test(クロス集計表)
```

＜帰無仮説＞

国と好きな色の間に関連はない。

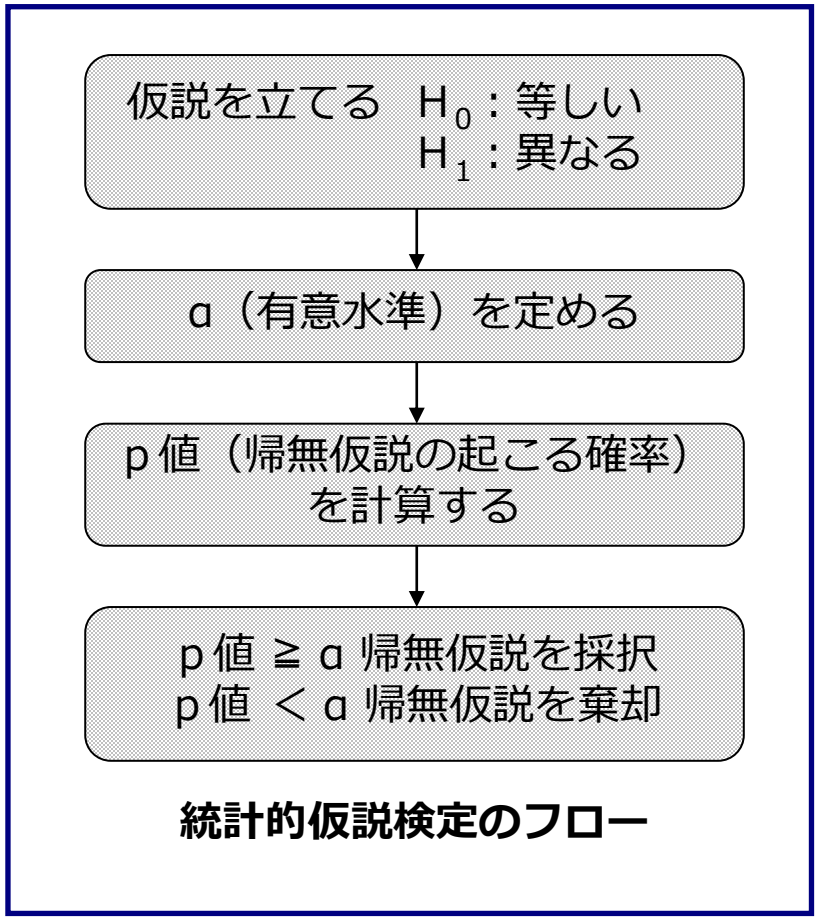
括弧内はクロス集計表を指定する

出力結果

```
  Pearson's Chi-squared test

data:  table(カラー$好きな色, カラー$国)
X-squared = 0.1867, df = 4, p-value = 0.9959

 警告メッセージ：
In chisq.test(table(カラー$好きな色, カラー$国)) :
   カイ自乗近似は不正確かもしれません
```

【Rによる実践】カイ二乗検定（chisq.test関数）

カイ二乗検定の解釈のまとめ



統計的仮説検定のフロー

H_0：国と好きな色の間に関連はない
H_1：国と好きな色の間に関連はある

有意水準α ＝ 0.05

p 値 ＝ 0.9959

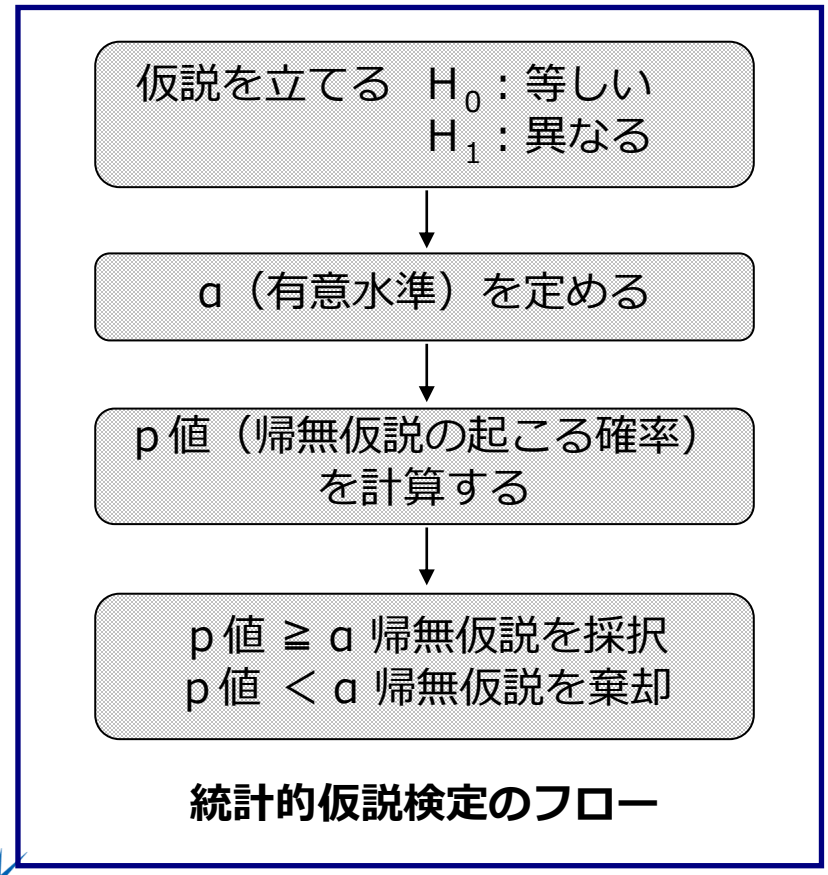
0.9959（p 値）＞ 0.05（有意水準α）より、帰無仮説を採択する

【結論】国と好きな色の間に関連はない

【練習問題7】カイ二乗検定

“甘いものと虫歯のデータ.csv“をRに読み込み、クロス集計表の作成とカイ二乗検定を実施してください。

- データ数は37件（37レコード）です



統計的仮説検定のフロー

H_0：

H_1：

有意水準α ＝ 0.05

p 値 ＝

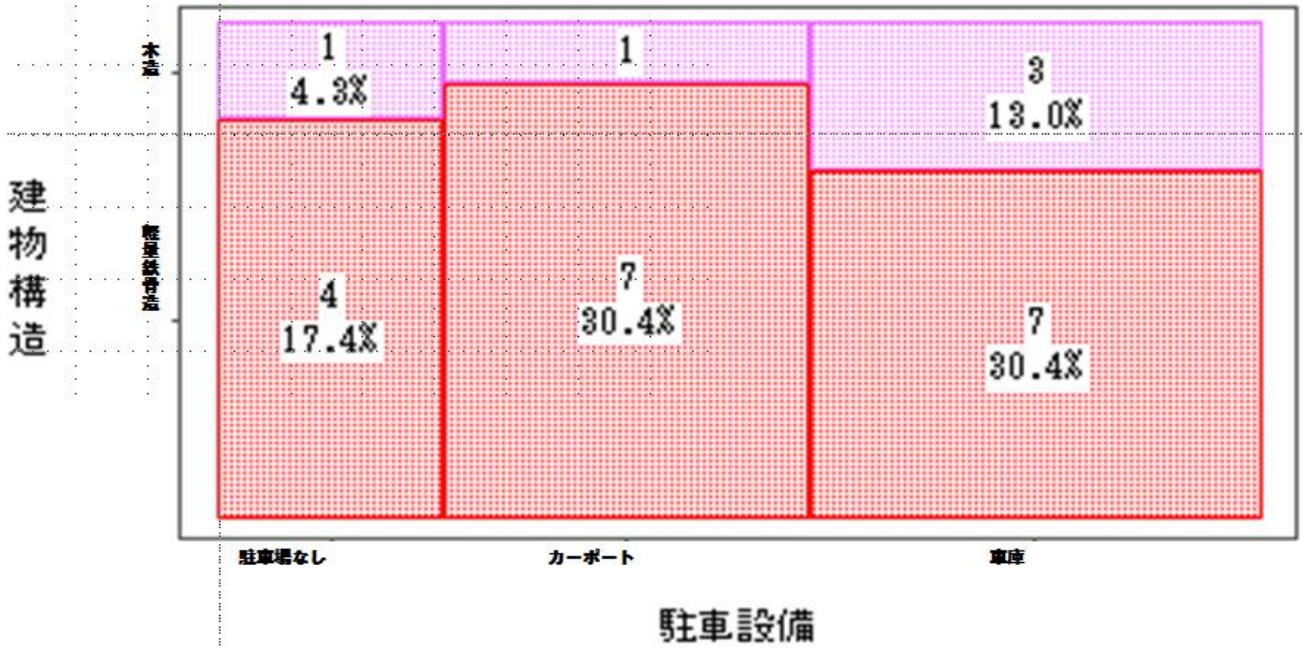
（p 値）＿＿ 0.05（有意水準α）より
帰無仮説を ＿＿＿＿

【結論】

3．質的データ×質的データ　～モザイク図～

モザイク図とは、質的データどうしの2変数の傾向を視覚的に把握する方法です。

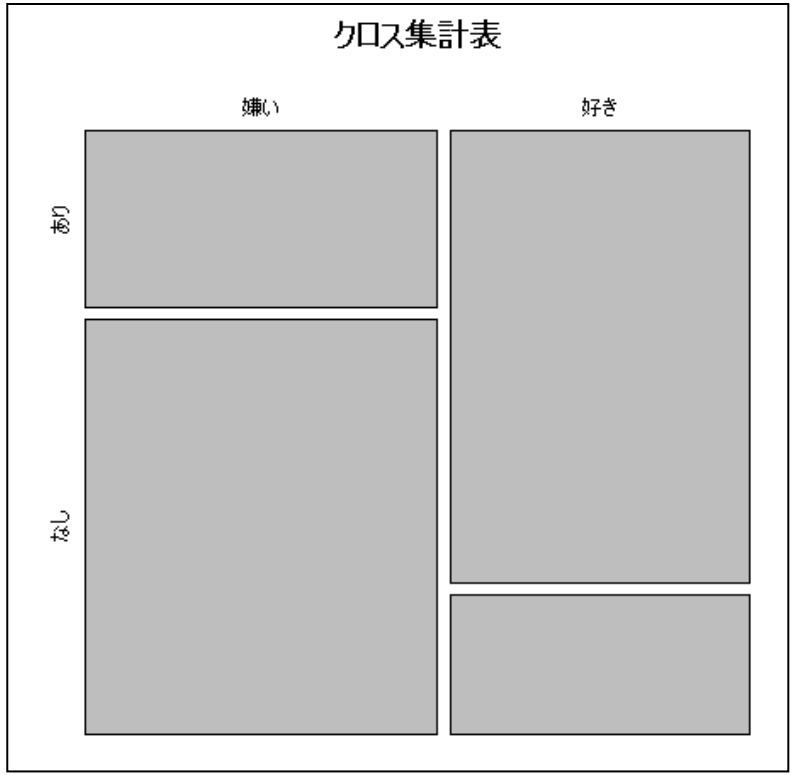
面積が割合を表しているので、二変数（カテゴリ＊カテゴリ）間の関係（割合）を視覚的に確認することができます。クロス集計表の全体のＮ数に対する、各セルの比率を、ボックスの大きさで表現します。クロス集計表の分布を可視化するのに適しています。



3．モザイク図（mosaicplot関数）

虫歯のデータを用いて、甘いものの好き嫌いと虫歯の関係を把握するためモザイク図を作成します。

```
>クロス集計表=table(虫歯$甘いもの,虫歯$虫歯)#クロス集計表の作成
>mosaicplot(クロス集計表)#モザイク図の作成
```



【説明】
モザイク図とは、棒の高さと棒の幅を使って2次元の構成比を比較するためのグラフです。帯グラフ(100%積み上げ棒グラフ)では棒の高さのみで構成比を表現しますが、モザイク図では棒の高さに加えて棒の幅によって２種類の構成比を表現します。

４．量的データ×質的データ　～カテゴリ別の平均算出～

量的データと質的データの関連を数値で把握するには、カテゴリ別（層別）に統計値を算出します。

2012年度のプロ野球実績データの例

2012年年棒（万円）

球団	人数	平均	標準偏差	最小値	最大値
DeNA	36	2,891	5,837	0	35,000
オリックス	39	3,145	4,589	240	25,000
ソフトバンク	43	4,001	6,698	270	30,000
ヤクルト	38	2,868	4,103	300	20,000
ロッテ	34	3,385	4,780	240	18,000
楽天	39	2,868	3,790	0	15,000
巨人	43	4,679	9,376	240	43,000
広島	40	2,477	3,212	250	16,000
阪神	41	5,420	9,609	240	40,000
西武	38	3,851	6,554	0	28,000
中日	34	4,696	7,716	400	33,000
日本ハム	34	4,345	6,482	480	27,000

カテゴリ

数値

【Rによる実践】グループごとの基礎統計量（by関数）

野球データを使い、球団ごとの年俸の基礎統計量を算出してみましょう。

```
>base=read.csv(“【2012年】全球団内野手.csv")
>str(base)
```

str関数を使うと、読み込んだデータの内容を簡潔に表示することができる

出力結果

```
'data.frame':   359 obs. of  41 variables:
 $ リーグ          : Factor w/ 2 levels "セ・リーグ","パ・リーグ": 1 1 1 1 1 1 1 1 1 1 ...
 $ 球団            : Factor w/ 12 levels "DeNA","オリックス",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ X2012年球団順位: int  6 6 6 6 6 6 6 6 6 6 ...
 $ No              : int  0 1 3 4 5 7 8 10 23 25 ...
 $ 選手名          : Factor w/ 359 levels "Ｔ?岡田","エルドレッド",..: 134 79 17 97 184 211 154 273 303 291 ...
    :                      :          :            :
 $ 投打            : Factor w/ 5 levels "右右","右左",..: 1 3 1 2 1 2 1 1 3 2 ...
 $ 出身地          : Factor w/ 52 levels "アメリカ","アンティル",..: 33 37 5 31 42 33 31 50 42 52 ...
 $ 年俸            : num  1070 5000 35000 1300 8000 5700 2900 4000 3800 1200 ...
 $ 翌年年俸        : num  1000 3000 35000 3000 8000 5000 3000 4000 4000 1950 ...
 $ 前年度差額      : num  70 2000 0 1700 0 700 100 0 200 750 ...
 $ 増減率          : num  0.07 0.4 0 1.31 0 0.12 0.03 0 0.05 0.63 ...
```

【Rによる実践】グループごとの基礎統計量（by関数）

層別に統計量を算出するにはby関数を使用します。

```
>by(base$年俸,base$球団,summary)
```

by関数は、by(統計量を求める変数、層別に利用する変数、求めたい統計量）と指定

出力結果

```
base$球団: DeNA
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    310     610    1250    3272    3500   35000
------------------------------------------------------------
base$球団: オリックス
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    240     800    1150    3227    1875   25000
------------------------------------------------------------
base$球団: ソフトバンク
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    270     400     900    3147    2625   20000
------------------------------------------------------------
   :             :             :             :
```

【Rによる実践】グループごとの基礎統計量（by関数）

球団ごとの年俸の平均を求めてみましょう。

```
>by(base$年俸,base$球団,mean)
```

Meanは平均を表す

出力結果

```
base$球団: DeNA
[1] 3271.538
------------------------------------------------------------
base$球団: オリックス
[1] 3226.538
------------------------------------------------------------
base$球団: ソフトバンク
[1] 3147.188
------------------------------------------------------------
base$球団: ヤクルト
[1] 3092.258
------------------------------------------------------------
base$球団: ロッテ
[1] 3971.481
------------------------------------------------------------
base$球団: 楽天
[1] 2420
------------------------------------------------------------
base$球団: 巨人
[1] 5780.303
------------------------------------------------------------
   :          :
```

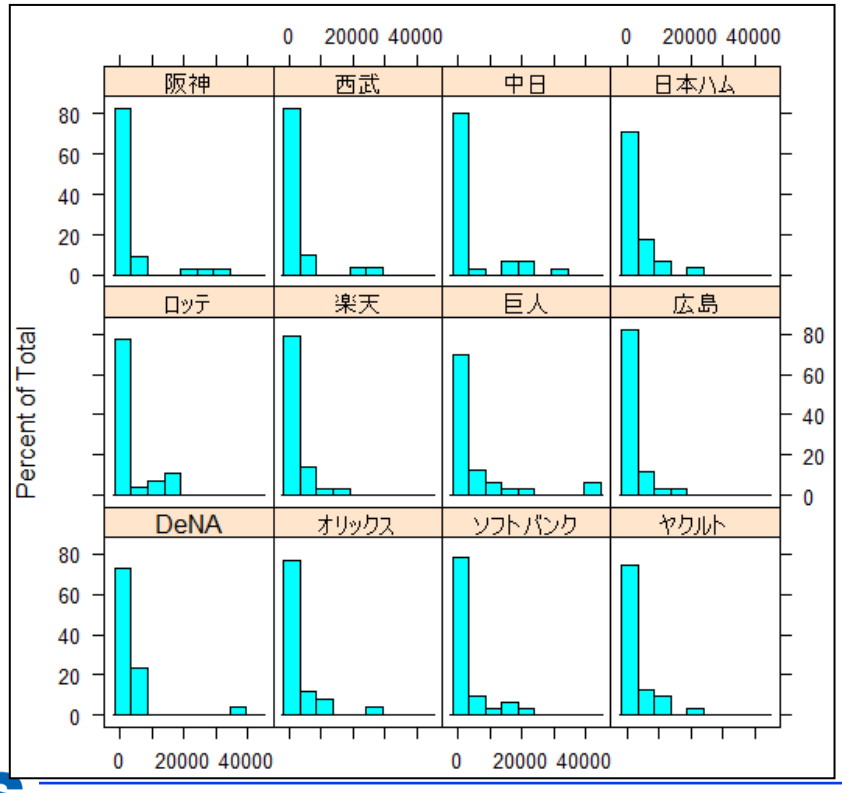
【Rによる実践】
量的データ×質的データ　～カテゴリ別の分布比較～（histogram関数）

量的データと質的データの関連、カテゴリ別（層別）ヒストグラムを書いて把握してみましょう。野球データの年俸を、球団別にヒストグラムにしてみます。

```
>library(lattice)
>histogram(~年俸|球団,data=base)
```

インストール済みのパッケージlatticeのhistogram関数を使用する為、関数を呼び出す前に指定が必要。

＜層別ヒストグラム＞



＜指定方法＞

histogram(~**列名** | **層別因子の列名** ,data=**データフレーム名**)

【Rによる実践】量的データ×質的データ　～層別箱ひげ図～

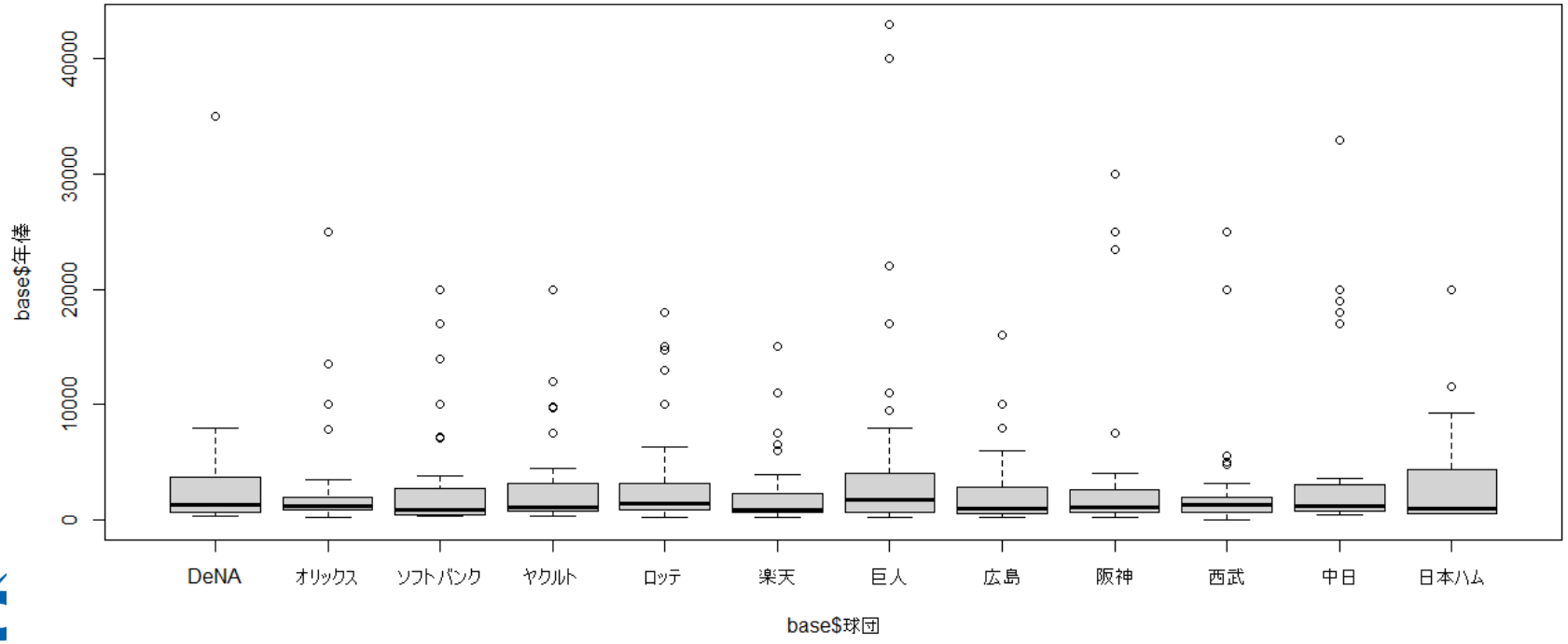
層別ヒストグラムを描いても球団ごとの特徴を把握するのは難しいので、層別箱ひげ図を書いてみましょう。

箱ひげ図の特徴

- 複数の分布の差を一目で確認することができる。
- 分布の偏りや外れ値を認識するのに有効。

```
>boxplot(base$年俸~base$球団)
```

boxplot(データ変数名~層別変数名)と指定



【練習問題8】層別箱ひげ図

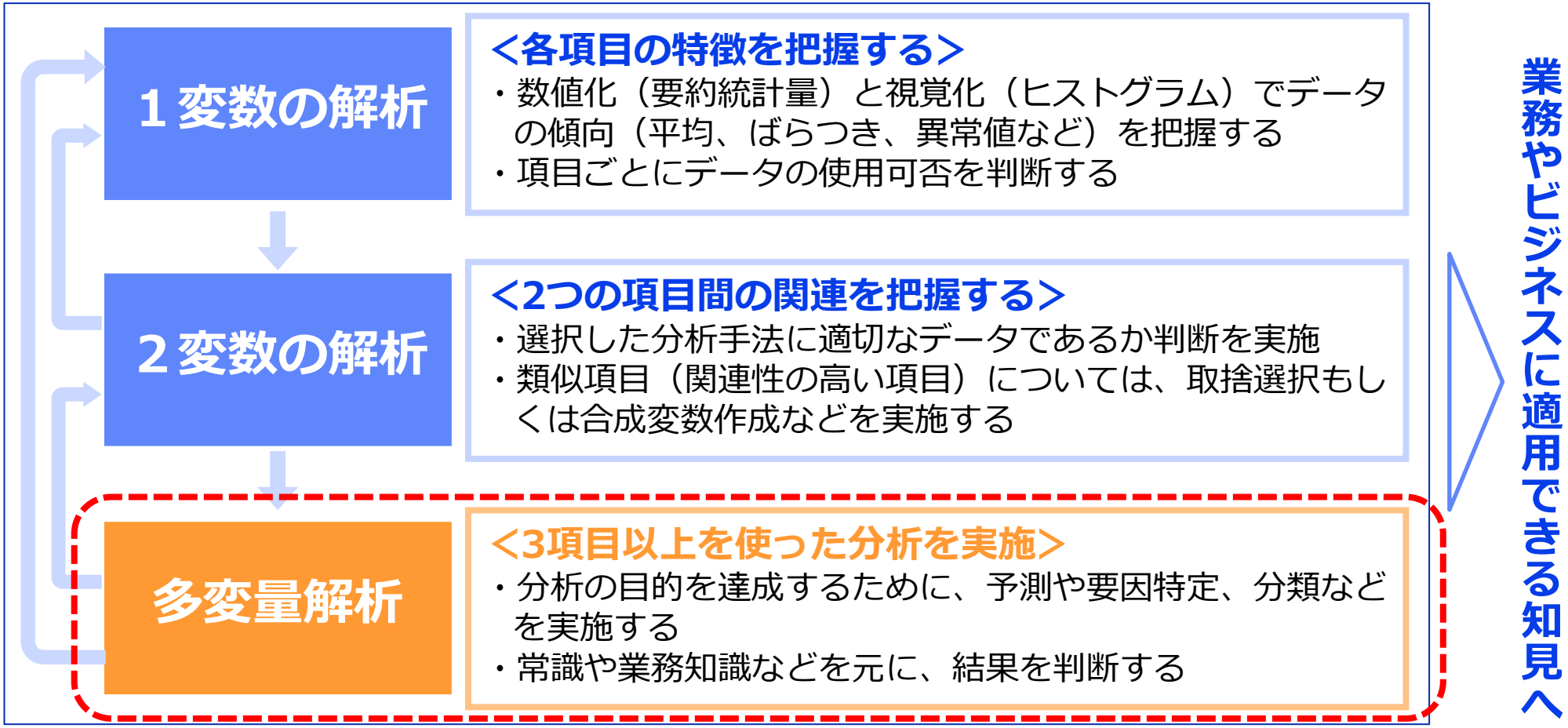
野球データ（【2012年】全球団内野手.csv ）を使って、球団別に次のそれぞれの層別箱ひげ図を作成して傾向を分析してください

安打　　本塁打　　打率

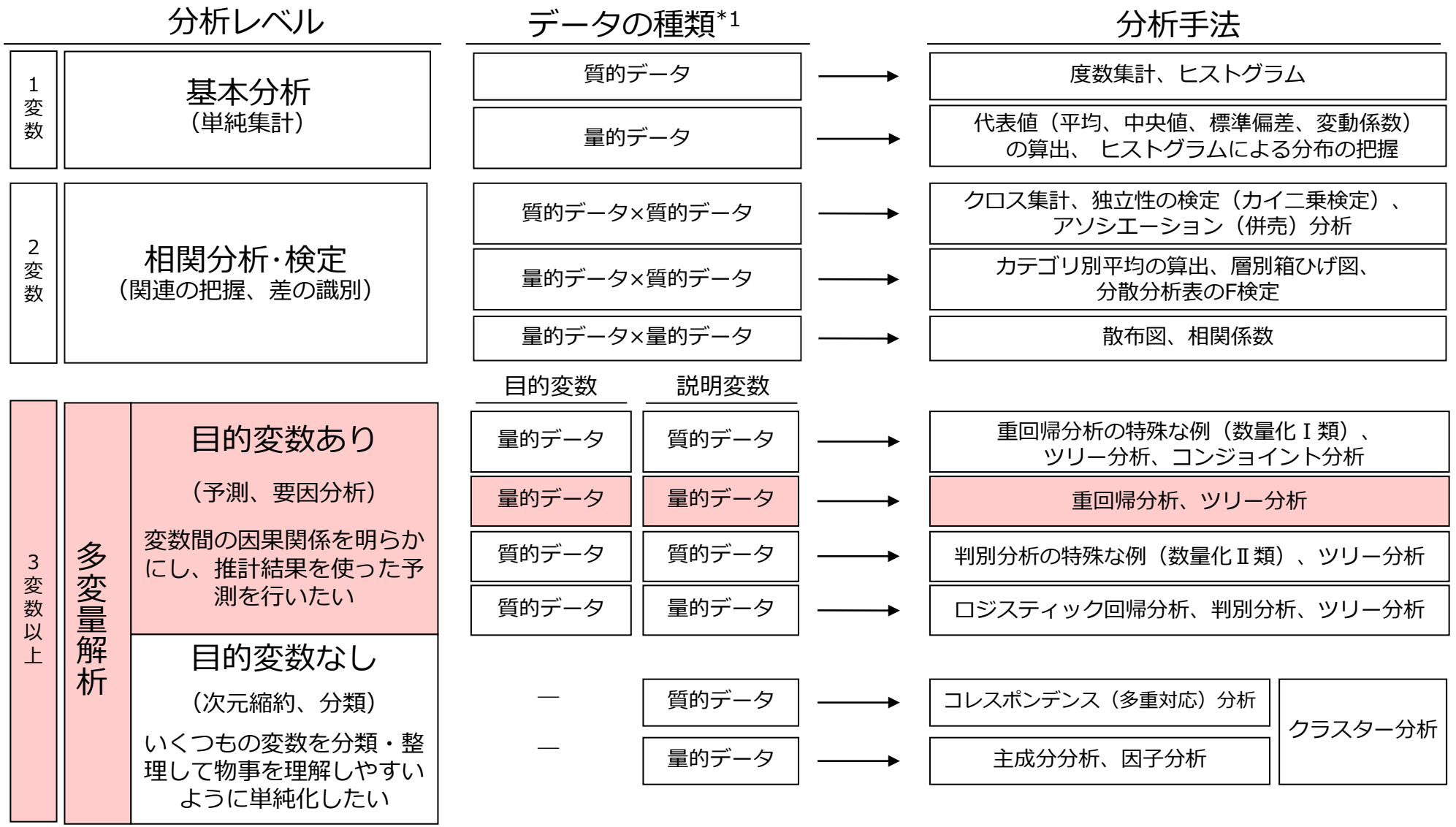
Ⅵ．多変量解析の王道
回帰分析概要

1．多変量解析へのガイダンス

多変量解析は、様々な項目（変数）の組み合わせで、複雑な現象をわかりやすく整理し、課題に対してアクションへの示唆となる分析結果を出すステップです。



代表的な分析手法について（再掲）



*1 ： 集計、解析のためのデータの大分類。 量的データは、データ間の大小を比較したり演算を行った時に意味のあるデータ (例、年収)
質的データは、データ間の大小比較や演算が無意味なデータ(例、血液型)

1．多変量解析へのガイダンス

3変数以上を使用する解析手法を「多変量解析」といいます。多変量解析には大きく分けて以下の2通りがあります。

【分析ターゲット（分析目的となる変数）のある解析（目的変数あり）】

予測や要因分析といった目的で分析を実施

重回帰分析、ロジスティック回帰、ツリー分析、判別分析などの手法がある

【分析ターゲット（分析目的となる変数）のない解析（目的変数なし）】

次元の縮約や分類といった目的で分析を実施

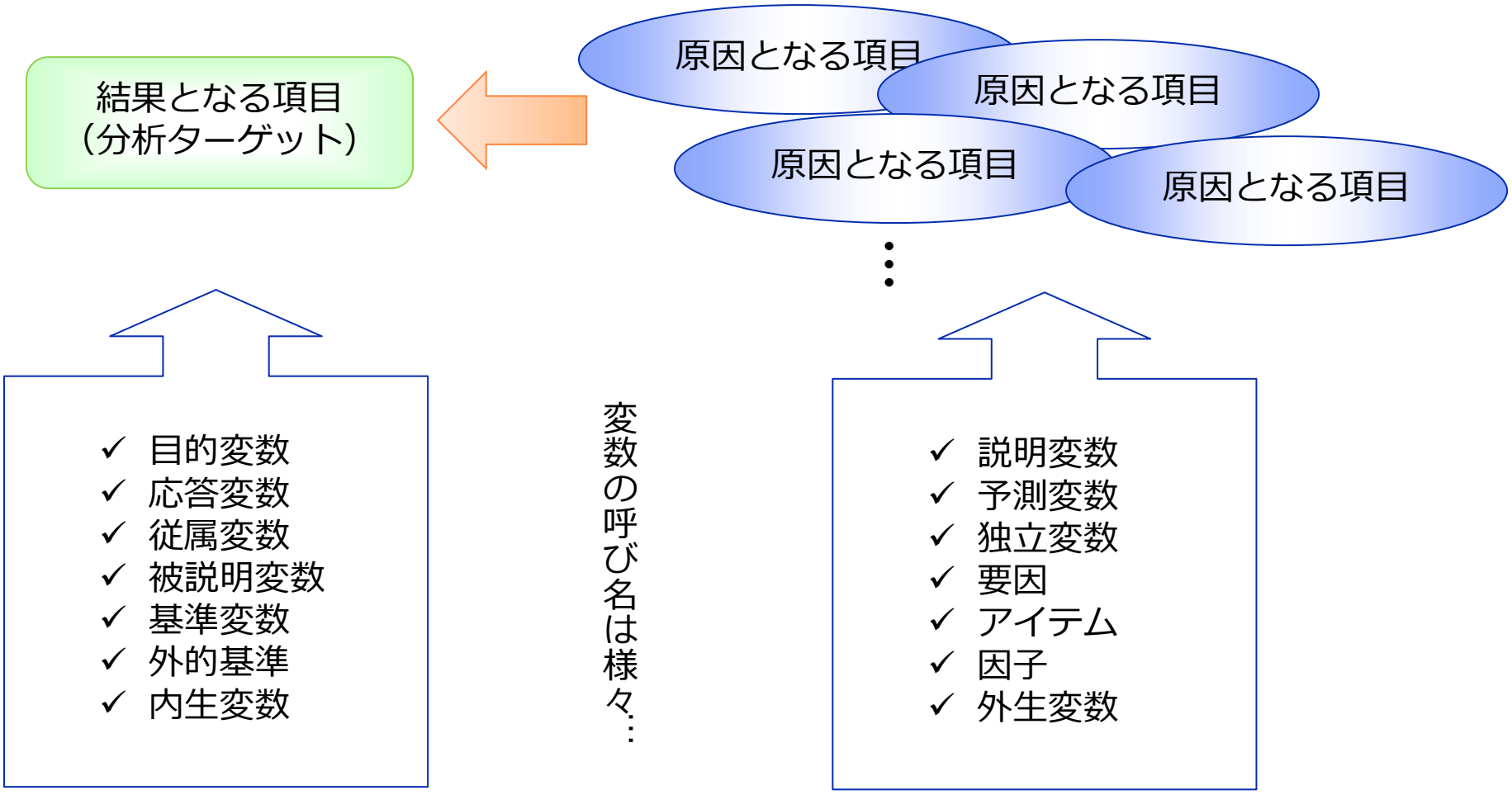
クラスター分析、主成分分析、因子分析、多重対応分析などの手法がある

<余談>　変量と変数　東洋経済新報社の『統計学辞典』より

統計集団をなす個体が"担っている"数量を抽象化して変量（ variate ）と呼ぶことが多い．数学の変数（ variable ）の概念に対応するが、個体に応じて変化し，物理的，経済的な意味をもつ量であるとの意識が強い．データは変量がとる値（ value ）である．しかし，変量とデータは変数と変数値のように混同されがちであり，うるさく区別しないほうが便利である．変量と変数も混同されがちで，本辞典内でも区別しない場合が多い。

1．多変量解析へのガイダンス

目的のある多変量解析に使う変数には、下記のように様々な呼び名があります。



2.回帰分析概要

＜回帰分析の全体像＞

回帰分析

- 説明変数が１つ → **単回帰分析**
  - モデルのあてはまりの検定 → **F検定**
  - パラメータの影響度の検定 → **t検定**
- 説明変数が２つ以上 → **重回帰分析**
  - モデルのあてはまりの検定 → **F検定**
  - 各パラメータの影響度の検定 → **t検定** → 影響度の低いパラメータを除去 → 変数選択
- 目的変数が質的データ（2値）→ **ロジスティック回帰分析** → 変数選択

※モデル：ある事象について、諸要素とそれら相互の関係を定式化して表したもの
パラメータ：・二つ以上の変数間の関数関係を間接的に表示する補助の変数
・母平均μや母分散σ^2のような母集団の未知数（母数）のこと

回帰分析（Regression Analysis）とは

ある事象の原因と結果、すなわち因果関係をモデル化する技法である。また、モデルの当てはまりが良ければ，そのモデルにより予測を行う。回帰分析には、目的変数、説明変数、共に量的データを使用する。

単回帰分析
（Simple regression analysis）

$$y = b_0 + b_1 x + \varepsilon$$

説明変数（X）が１つだけ

例）身長から体重を予測する、安打数から野球選手の年俸を予測する

重回帰分析
（Multiple regression analysis）

$$y = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_n x_{ni} + \varepsilon$$

説明変数（X）が２つ以上

例）父母の身長から子供の身長を予測する
安打数・ホームラン数・盗塁数から野球選手の年俸を予測する

y：目的変数…予測したい変数　ターゲット変数　結果となる変数
x：説明変数…原因となる変数
b_1～b_n：係数（パラメータ）…データから推定した値
b_0：切片…回帰直線がY軸と交わるところ　Xがゼロのときの Yの値

2.回帰分析概要

回帰分析の適用分野・利用例としては、以下の2つがあります
- ✓需要予測、販売予測などの予測
- ✓需要構造の分析・変数の因果関係（影響のある変数）の特定

＜予測＞

- 過去のCMと来場者数との関連を分析し、今回のCMによる予測来場者数を求めたい
- 自動販売機での飲料の売り上げは気温が高くなると多くなるが、25℃と30℃でどのぐらい変わるのか知りたい
- 入社時の評価から営業実績を予測できるか、過去10年間で採用した営業スタッフの入社時の評価を集め、現在の営業成績とどの程度関係しているのかを分析したい

＜因果関係の特定＞

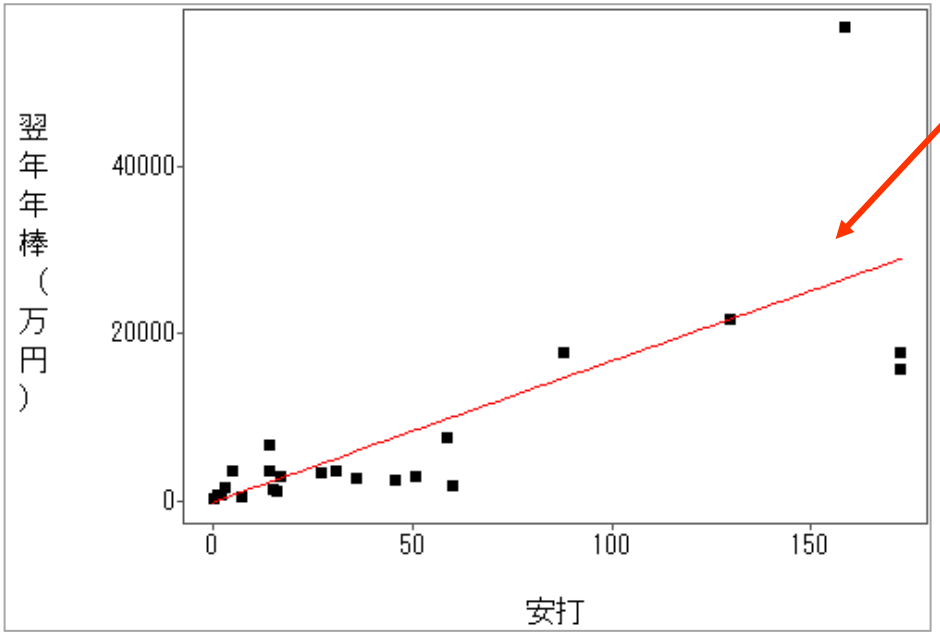
- 駅の周辺の物件について、様々な物件の条件（広さ、駅からの遠さ、階数など）がどれくらい家賃に影響するのかを調べたい
- レストランの売り上げに、グルメサイトのPVがどのぐらい影響しているのか、いくつかのグルメサイトのPV数と既存顧客へのDM送付有無・曜日・天気などと合わせて効果を検証したい

Ⅶ. 単回帰分析

1.単回帰分析

2つの量的データの関連性を y=ax+b という数式（直線）であらわしたものが単回帰分析です。データによくあてはまる直線を見つけるにはどうしたらいいでしょう。

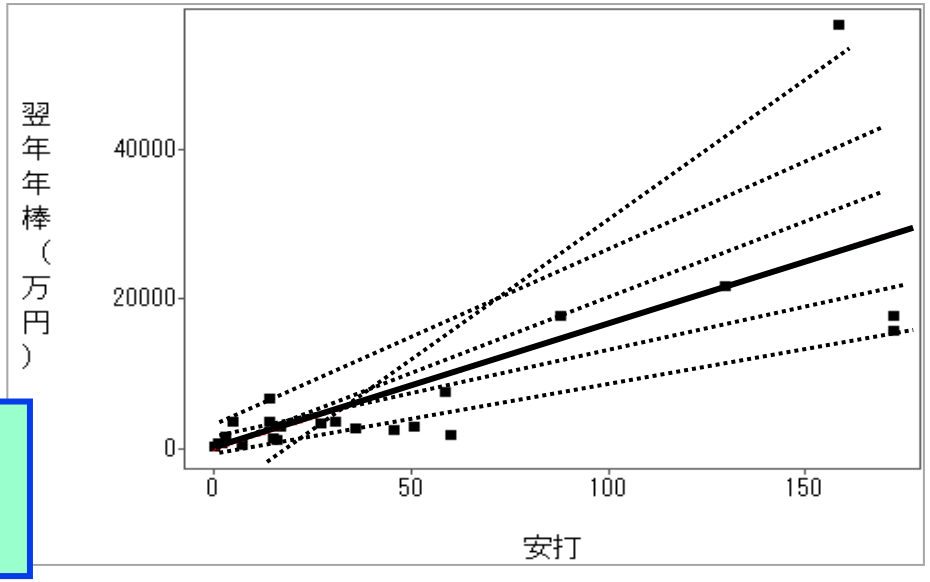
巨人軍のデータ



野手の2013年推定年俸の予測値

推定年俸=167.939×2012年の安打数－199

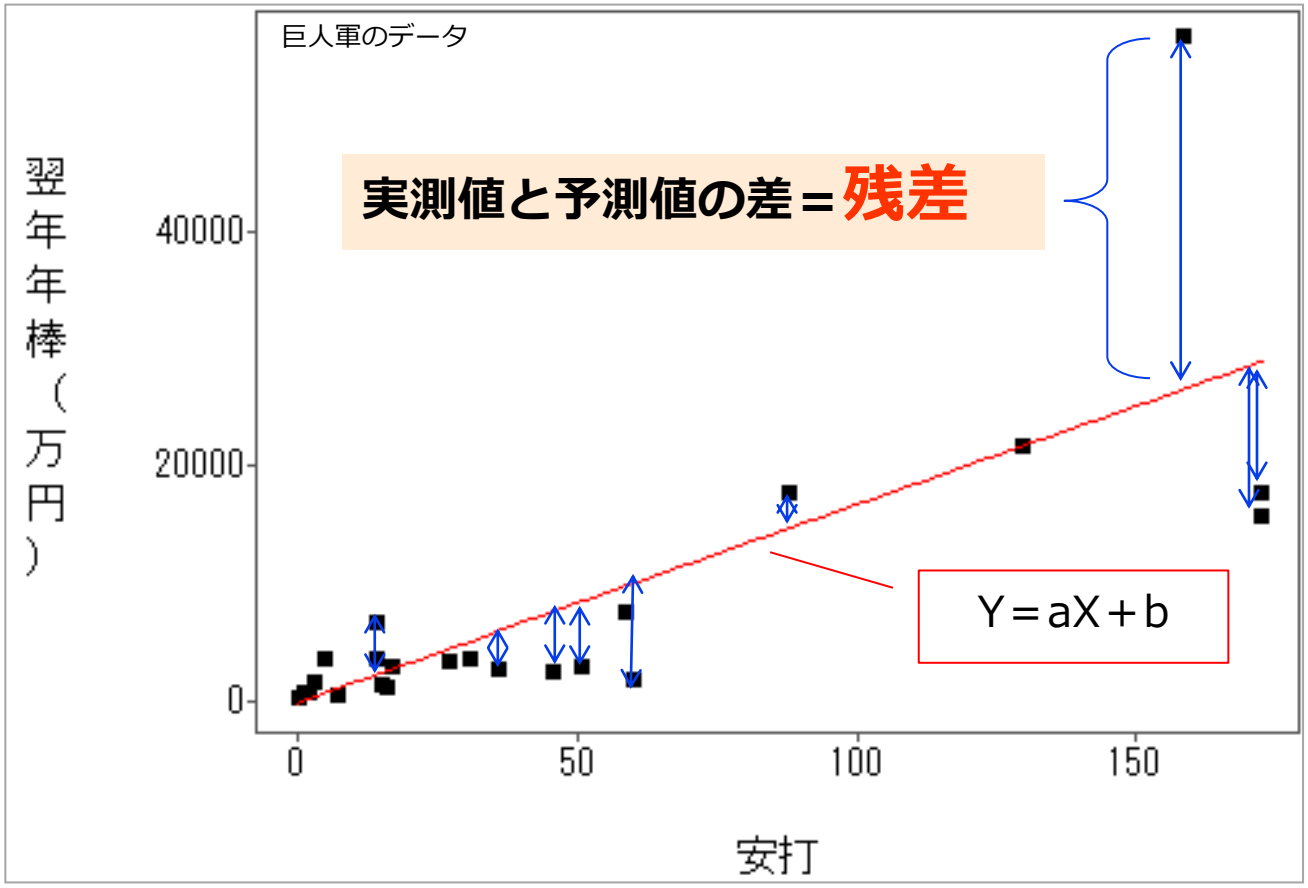
（数字の単位：万円）



このように、直線は無数に描くことができます。
では、何を基準に一番良い回帰式、回帰直線とするのでしょうか？

1.単回帰分析

回帰直線の求め方はいくつかありますが、最も基本的な方法に**最小二乗法**があります。



回帰直線（Y=aX＋b）のaとbをデータ（XとY）から決めるために実測値と予測値との差（残差）を一番小さくする方法を取る。

ただし、単純に実測値から予測値を引くと、プラスとマイナスの数値が混在し、最少を見つけるのが困難。

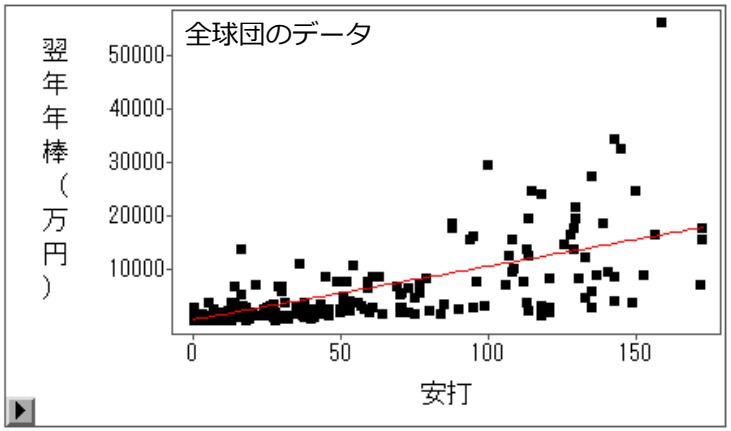
残差を二乗してすべて足したもの（二乗和）が最小になるように、aとbを求める　⇒　最小二乗法

1.単回帰分析（結果の見方）

回帰分析結果の指標で確認するところは、寄与率・F値（そのP値）・t 値（そのP値）となります。
また、推定値は予測値算出のためのモデル式を作成する時の係数になります。

モデル式

NEN_2013 = 335.610 + 100.633 ANDA



当てはめの要約

応答変数の平均	4739.4318	寄与率	0.4803
誤差の標準偏差	4939.1934	自由度調整済み寄与率	0.4783

分散分析表

変動因	自由度	平方和	平均平方	F 統計量	P 値 (F)
モデル	1	5.908E+09	5.908E+09	242.17	<.0001
誤差	262	6.392E+09	24395631.9		
全体 (C)	263	1.230E+10			

パラメータ推定値

変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)
Intercept	1	335.6104	415.3204	0.81	0.4198
ANDA	1	100.6326	6.4667	15.56	<.0001

※Interceptとは切片（Y = aX + b でいう b ）のこと。

推定値： **回帰式を作成する際の係数。**それぞれの変数に対して符号が常識と合っているかを確認する。

寄与率： 決定係数とも言い、**説明変数と目的変数との間に深い関係があるかどうかを表す指標。**ただし、ロジスティック回帰の際にはこの値は低めに計算されるので、実際の評価には使用できない。また、自由度調整済みの寄与率が算出されている場合は、そちらを参照する。寄与率の判断基準はモデルの利用目的によって変化するが、一般的に0.6以上であればそこそこの精度、0.9以上であれば充分な精度と判断してよい。

F統計量： **モデルのあてはまりを表す指標であり「母集団において、この分析で作成された式は全くあてにならない」という仮説を検証したもの。**F統計量に対するP値（有意確率）が上記の仮説が起こる確率を表す。通常5%有意水準（P値＜0.05）で判断する。

t統計量： **各変数のモデルへの影響度をはかる指標であり「母集団において、各説明変数が目的変数に影響を与えない」という仮説を検証したもの。**t統計量に対するP値（有意確率）が上記の仮説が起こる確率を表す。通常5%有意水準（P値＜0.05）で判断する。

【Rによる実践】回帰係数の算出（lm関数）

回帰係数と切片は、lm関数を用いて求めることができます。本のデータ（book_price.csv）を使用して、本のページ数と価格の関係を表す単回帰式を求めてみましょう。

```
>本 =read.csv("book_price.csv")
>summary(lm(本$価格~本$ページ数))
```

目的変数　　説明変数

lm関数は、linear model(線形モデル)を意味する。
出力結果を見やすくするため、summary関数を使用する。

出力結果

```
   Call:
lm(formula = 本$価格 ~ 本$ページ数)

Residuals:
     Min       1Q   Median       3Q      Max
-1150.47  -503.48  -292.50    29.71  1773.58

Coefficients:
             Estimate   Std. Error t value Pr(>|t|)
(Intercept) 1868.389    668.527    2.795    0.0234 *
本$ページ数      8.292      1.052    7.884 4.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1001 on 8 degrees of freedom
Multiple R-squared: 0.886,      Adjusted R-squared:
0.8717
F-statistic: 62.15 on 1 and 8 DF,  p-value: 4.851e-05
```

残差の要約統計量

切片(b0)

回帰係数(b1):傾き

寄与率

寄与率は「ここで用いられた説明変数を用いて、目的変数をどの程度説明できるか」モデルで説明できる変動の割合を示します。寄与率は、0から1の値を取ります。

パラメータ推定値の表には、b_0(切片)とb_1(回帰係数)の値や標準誤差、パラメータに対する検定結果などが表示されます。

【単回帰式】

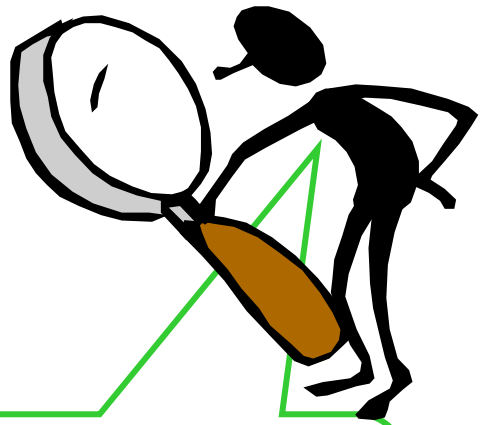
$y=1868.389+8.292x$

【Rによる実践】単回帰分析の出力

モデルのあてはまりに関する情報（分散分析表）

回帰係数に関する情報（パラメータ推定値）

```
   Call:
lm(formula = 本$価格 ~ 本$ページ数)

Residuals:
     Min       1Q   Median       3Q      Max
-1150.47  -503.48  -292.50    29.71  1773.58

Coefficients:
                Estimate  Std. Error  t value Pr(>|t|)
(Intercept) 1868.389    668.527    2.795    0.0234 *
本$ページ数       8.292       1.052    7.884 4.85e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1001 on 8 degrees of freedom
Multiple R-squared: 0.886,      Adjusted R-squared:
0.8717
F-statistic: 62.15 on 1 and 8 DF,  p-value: 4.851e-05
```

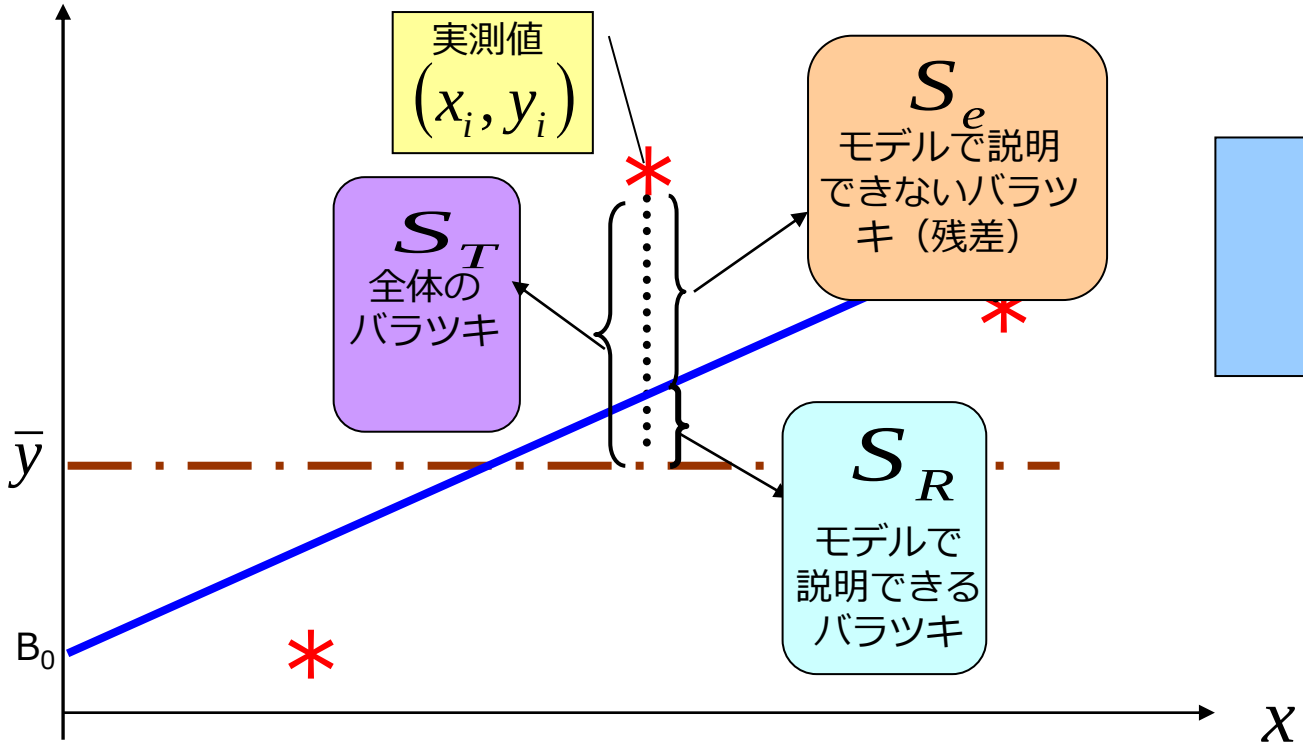


得られたモデルは、はたして良いモデルなのだろうか？
算出された統計量を見て、検討してみよう。

※本のデータ（book_price.csv）を使用した単回帰分析結果より

1.単回帰分析（回帰分析の評価）

回帰分析の当てはまりは、以下のように考えます。（単回帰の例）



モデル式：

$$\hat{y} = b_0 + b_1 x$$

・傾きが0のベースラインモデル（平均を表す）よりも目的変数を良く説明しているかどうかは、“モデルで説明できるバラツキ（S_R）”と“説明できないバラツキ（S_e）”を比較する。

・寄与率は英語でR-Squre（R^2）といい決定係数と呼ぶこともある。０～１の値をとる。回帰モデルにより説明されているバラツキの割合（何%がこのモデルで説明できるか）を示す。回帰モデルの精度を表す指標。寄与率が高いモデルほど精度が良いと判断できる。

$$R^2 = \frac{S_R}{S_T}$$

【Rによる実践】単回帰分析の出力～分散分析表～

回帰分析では、分散分析表におけるＦ検定が実行されます。回帰係数に対する検定になっており、この検定が有意になっていれば、回帰係数が0ではないと判断できます。本のデータ（book_price.csv）を使用して求めた回帰分析の結果で、Ｆ検定の結果をみていきます。

```
   Call:
lm(formula = 本$価格 ~ 本$ページ数)

Residuals:
     Min       1Q   Median       3Q      Max
-1150.47  -503.48  -292.50    29.71  1773.58

Coefficients:
              Estimate  Std. Error  t value Pr(>|t|)
(Intercept) 1868.389    668.527    2.795    0.0234 *
本$ページ数       8.292       1.052    7.884 4.85e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1001 on 8 degrees of freedom
Multiple R-squared: 0.886,      Adjusted R-squared: 0.8717
F-statistic: 62.15 on 1 and 8 DF,  p-value: 4.851e-05
```

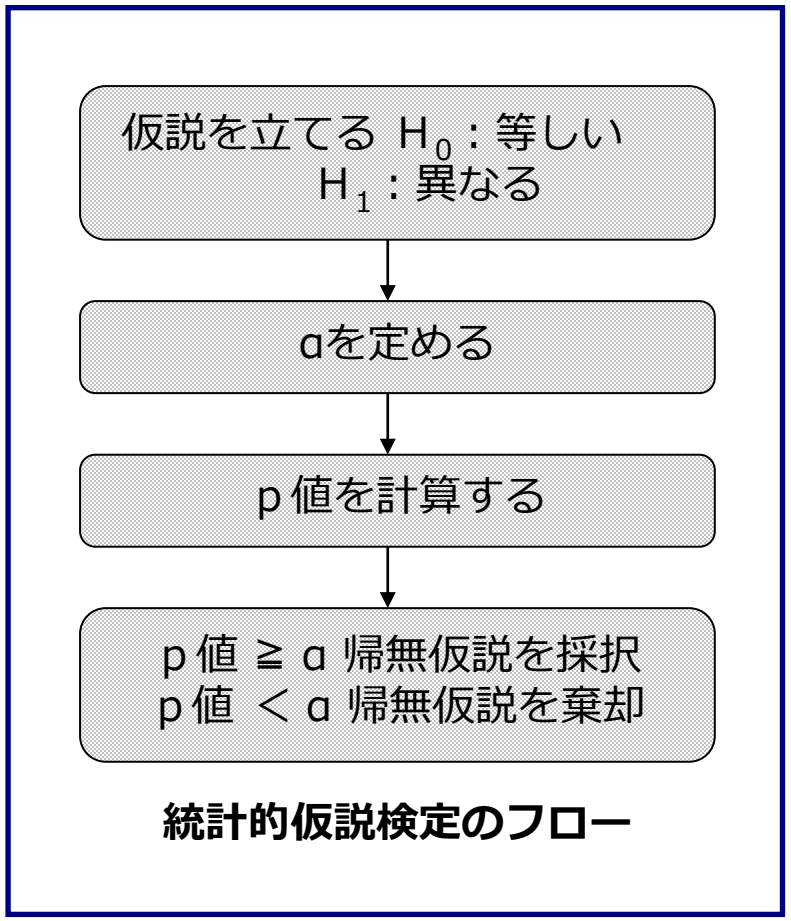
F値

p値

F-statistic（Ｆ値）　モデルで説明できるバラツキと説明できないバラツキの比＝ $\frac{S_R}{S_e}$

p-value（Ｐ値）　　Ｆ値から算出されるＰ値。
Ｆ検定（モデルの検定）をおこなっている。

【Rによる実践】単回帰分析の出力～F検定～



統計的仮説検定のフロー

F検定

H_0：回帰係数は0である

H_1：回帰係数は0ではない

有意水準α ＝ 0.05（両側検定）

p値 ＝ 4.851e-05

4.851e-05（p値）＜0.05（有意水準α）より、帰無仮説を棄却。

【結論】回帰係数は0ではないといえる。

【Rによる実践】単回帰分析の出力～パラメータ推定値～


```
   Call:
lm(formula = 価格 ~ ページ数)

Residuals:
     Min       1Q   Median       3Q      Max
-1150.47  -503.48  -292.50    29.71  1773.58

Coefficients:
             Estimate  Std. Error  t value Pr(>|t|)
(Intercept) 1868.389    668.527    2.795   0.0234 *
ページ数       8.292      1.052    7.884 4.85e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1001 on 8 degrees of freedom
Multiple R-squared: 0.886,      Adjusted R-squared: 0.8717
F-statistic: 62.15 on 1 and 8 DF,  p-value: 4.851e-05
```

Estimate （パラメータ推定値）　母集団パラメータの推定値

$$y = b_0 + b_1 x$$

Std.Error（標準誤差）　パラメータ推定値の標準誤差

t Value（ t 値）　$t\text{値} = \dfrac{\text{パラメータ推定値}}{\text{標準誤差}}$　t 検定に使われる。

Pr(>|t|)（ P 値）　t 統計量から、計算されるP値。
各変数の影響度を計る“ t 検定”に使われる。

【Rによる実践】単回帰分析の出力～ t 検定～

回帰分析では、各パラメータに対する t 検定が実行されます。本のデータ（book_price.csv）を使用して求めた回帰分析の結果で、切片に対する t 検定の結果を見ていきます。

出力結果

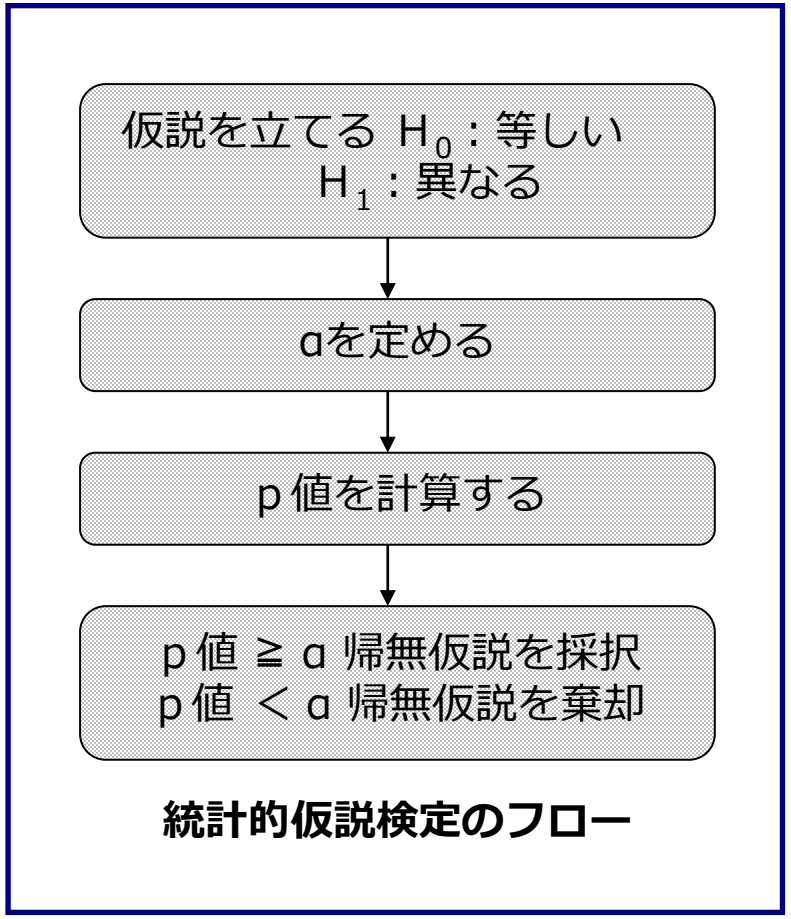
```
   Call:
lm(formula = 価格 ~ ページ数)

Residuals:
     Min       1Q   Median       3Q      Max
-1150.47  -503.48  -292.50    29.71  1773.58

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) 1868.389    668.527   2.795    0.0234 *
ページ数          8.292       1.052   7.884    4.85e-05***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1001 on 8 degrees of freedom
Multiple R-squared: 0.886,      Adjusted R-squared:
0.8717
F-statistic: 62.15 on 1 and 8 DF,  p-value: 4.851e-05
```

t値

p値

パラメータの t 検定の結果を見ます

【Rによる実践】単回帰分析の出力～ t 検定～



統計的仮説検定のフロー

個々のパラメータに対する検定
（切片の t 検定）

H_0：切片は0である

H_1：切片は0ではない

有意水準α ＝ 0.05（両側検定）

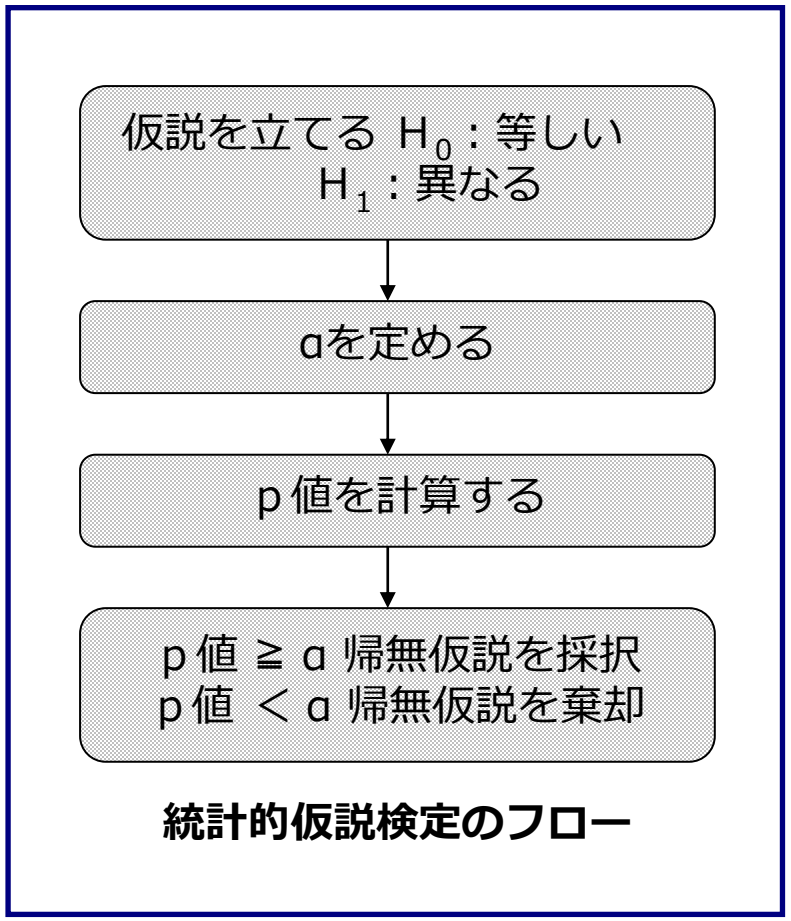
p 値 ＝ 0.0234

0.0234 （p 値）＜0.05（有意水準α）より、帰無仮説を棄却。

【結論】切片は0ではないといえる。

→切片は意味があるといえる。

【Rによる実践】単回帰分析の出力～ t 検定～



統計的仮説検定のフロー

個々のパラメータに対する検定
（ページ数の回帰係数の t 検定）

H_0：ページ数の回帰係数は0である

H_1：ページ数の回帰係数は0ではない

有意水準α ＝ 0.05（両側検定）

p 値 ＝ 4.85e-05

4.85e-05（p 値）＜0.05（有意水準α）より、帰無仮説を棄却。

【結論】ページ数の回帰係数は0ではないといえる。

→ページ数の回帰係数は意味があるといえる。

【練習問題9】単回帰分析

野球データ（【2012年】全球団内野手.csv）を使って、ヒット1本がいくらになるかを予測する回帰式を作成してください。（年俸と安打を利用）

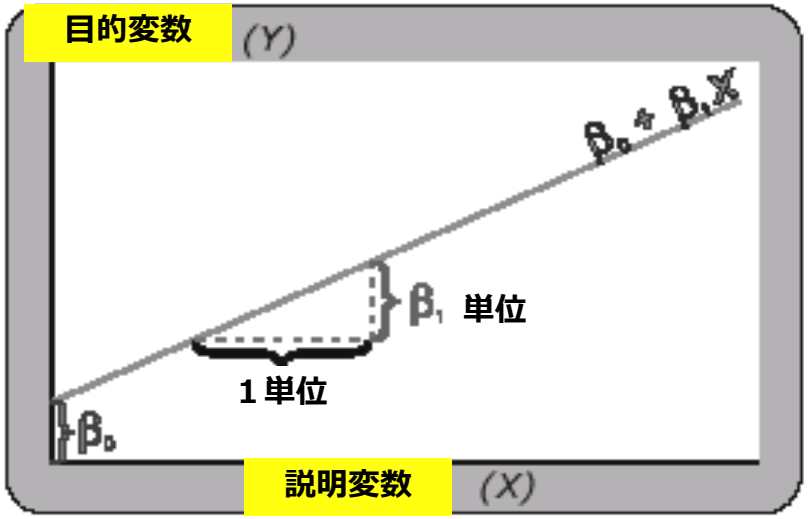
※回帰式を作成すると同時に、寄与率でモデルの説明力を判断し、パラメータに対するt検定で式の係数に意味があるかどうかも判断すること

Ⅷ. 重回帰分析

1.重回帰分析

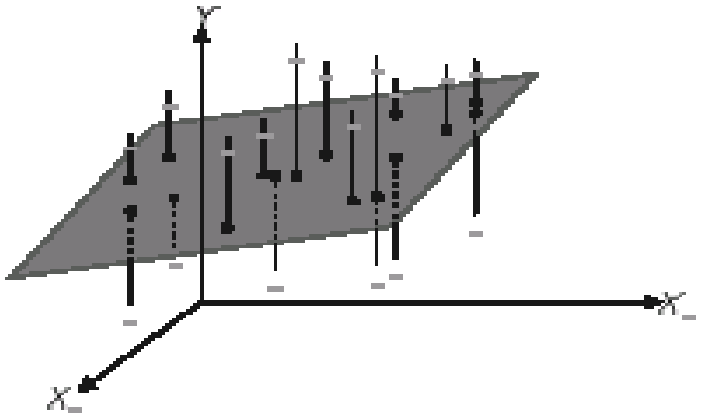
重回帰分析とは、説明変数が２つ以上の場合の回帰分析です。

＜単回帰＞



線（一次元）で２変数の関係を捉える。

＜重回帰＞



２次元、または多次元の平面、曲面で多変数の関連を捉える。

1.重回帰分析（回帰分析で説明変数を増やす際の注意点）

2013年の年俸を2012年の安打数から推計すると…

モデル式				
NEN_2013	=	335.610	+	100.633 ANDA

寄与率	0.4803

パラメータ推定値

変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)
Intercept	1	335.6104	415.3204	0.81	0.4198
ANDA	1	100.6326	6.4667	15.56	<.0001

安打数だけで説明する場合、寄与率は0.4803であったとする。そうすると、まだ約52%のデータのバラツキは説明させていないことになる。よって、打率を追加することにより、寄与率が上がることを期待して、重回帰分析を行う。

モデル式						
NEN_2013	=	1058.10	+	103.846 ANDA	-	3938.40 DARITSU

寄与率	0.4819

パラメータ推定値

変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)
Intercept	1	1058.1047	910.5006	1.16	0.2463
ANDA	1	103.8458	7.4049	14.02	<.0001
DARITSU	1	-3938.4023	4416.3666	-0.89	0.3733

打率を説明変数に加えた結果、寄与率は少しだけ上がった。
ただし、打率はパラメータとしては意味がない。

1.重回帰分析（回帰分析で説明変数を増やす際の注意点）

いろいろな変数を使用してみると…

モデル式

NEN_2013 = 1882.33 + 72.8023 ANDA + 138.348 DATEN + 9047.49 DARITSU
+ 15.1910 SHIAI - 35.6023 TOKUTEN + 178.217 HR - 25.6325 TOURUI
- 378.333 TOURUI_SOSHI - 3.0657 GIDA + 104.894 GISEIFURAI + 88.9390 FOUR_BALL
+ 363.380 KEIEN - 46.8863 DEAD_BALL - 103.164 SANSHIN + 87.5490 DOUBLE_PLAY
- 8862.42 SLG - 3028.38 OPS

寄与率	0.6592

パラメータ推定値

変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)
Intercept	1	1882.3297	901.8552	2.09	0.0379
ANDA	1	72.8023	29.2227	2.49	0.0134
DATEN	1	138.3485	58.9686	2.35	0.0198
DARITSU	1	9047.4899	10172.5592	0.89	0.3747
SHIAI	1	15.1910	14.9722	1.01	0.3113
TOKUTEN	1	-35.6023	59.2151	-0.60	0.5482
HR	1	178.2173	158.1567	1.13	0.2609
TOURUI	1	-25.6325	77.4442	-0.33	0.7409
TOURUI_SOSHI	1	-378.3326	190.5798	-1.99	0.0482
GIDA	1	-3.0657	47.7171	-0.06	0.9488
GISEIFURAI	1	104.8940	263.0293	0.40	0.6904
FOUR_BALL	1	88.9390	38.4885	2.31	0.0217
KEIEN	1	363.3800	292.6051	1.24	0.2155
DEAD_BALL	1	-46.8863	128.8854	-0.36	0.7163
SANSHIN	1	-103.1636	22.3971	-4.61	<.0001
DOUBLE_PLAY	1	87.5490	122.4025	0.72	0.4751
SLG	1	-8862.4209	4966.5224	-1.78	0.0756
OPS	1	-3028.3773	8362.8899	-0.36	0.7176

寄与率は上がりましたが、係数（パラメータ）として意味がないものが多い結果となります。（p値が0.05以下が意味がある）

ある偏回帰係数は、それ以外の説明変数の値を固定した（変化させない）場合に、その説明変数が1増加すると結果がどれだけ増加/減少するかを示しています。

1.重回帰分析（回帰分析で説明変数を増やす際の注意点）

前ページでp値が0.05以下の変数だけ使用して再実行すると…

モデル式

NEN_2013 = 1182.59 + 68.9780 ANDA + 182.473 DATEN - 514.110 TOURUI_SOSHI
+ 97.9641 FOUR_BALL - 100.719 SANSHIN

寄与率	0.6432

パラメータ推定値

変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)
Intercept	1	1182.5859	368.2455	3.21	0.0015
ANDA	1	68.9780	17.5052	3.94	0.0001
DATEN	1	182.4727	33.7825	5.40	<.0001
TOURUI_SOSHI	1	-514.1098	137.3623	-3.74	0.0002
FOUR_BALL	1	97.9641	31.2409	3.14	0.0019
SANSHIN	1	-100.7192	19.1900	-5.25	<.0001

寄与率はあまり下がらず、意味のある説明変数のみとなった。

説明変数を沢山使えば説明力（寄与率）はあがります。
これは目的変数に影響及ぼす、及ぼさない関係なく寄与率は大きくなります。
しかし、意味のない説明変数を使うことは間違った予測値を算出する原因になります。
また、係数の解釈も難しくなります。（説明変数がお互いに影響しあって、常識とは反対の符号がでてくる場合がある）
重回帰分析を実施する際には、**なるべく少ない説明変数でなるべく高い説明力**のものが良いモデルとなります。

【Rによる実践】重回帰分析（lm関数）

野球データを使って、安打数と打率から年俸を予測します。モデルの精度を計るＦ検定と各パラメータのモデルへの影響度を計るｔ検定の結果をみていきます。

```
>重回帰分析 =lm(base$年俸 ~ base$安打+base$打率)
>summary(重回帰分析)
```

目的変数　　説明変数

説明変数と説明変数の間は「＋」を記述して変数を並べる。

出力結果

```
Call:
lm(formula = base$年俸 ~ base$安打 + base$打率)

Residuals:
   Min     1Q Median     3Q    Max
-12778  -1907   -490    571  40694

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1899.344   1044.029   1.819    0.070 .
base$安打     87.487      8.491  10.304  <2e-16 ***
base$打率   -5382.824   5064.041  -1.063    0.289
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5666 on 261 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared: 0.3266,    Adjusted R-squared: 0.3215
F-statistic: 63.31 on 2 and 261 DF,  p-value: < 2.2e-16
```

切片(b_0)

回帰係数(b_1):傾き

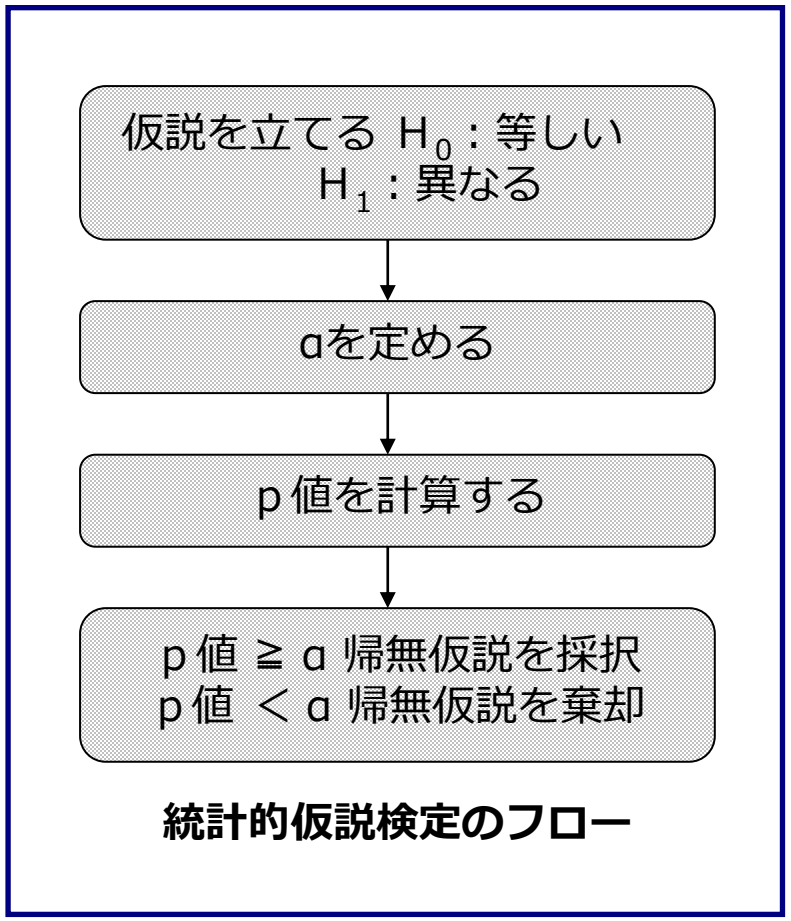
回帰係数(b_2):傾き

t 検定：p値

自由度調整済み決定係数

Ｆ検定：p値

【Rによる実践】重回帰分析の出力～F検定～



統計的仮説検定のフロー

F検定

H_0：すべての回帰係数は0である

H_1：少なくとも1つの回帰係数は0ではない

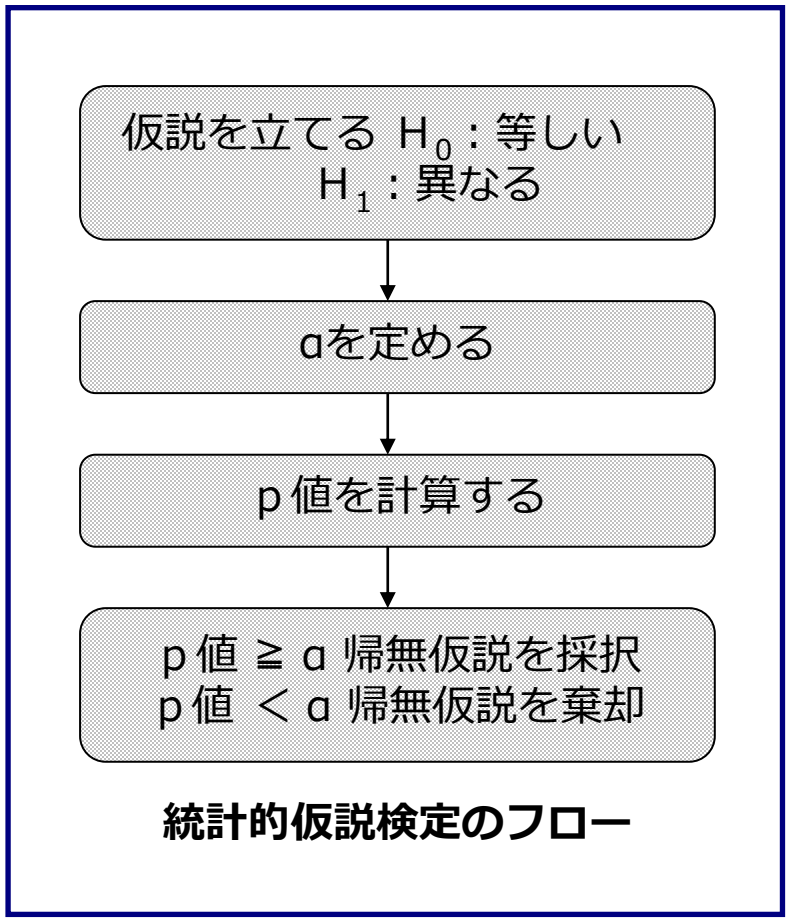
有意水準α ＝ 0.05（両側検定）

p値 ＝ 2.2e-16

2.2e-16（p値）＜0.05（有意水準α）より、帰無仮説を棄却。

【結論】少なくとも1つの回帰係数は0ではないといえる。

【Rによる実践】重回帰分析の出力～ t 検定～



統計的仮説検定のフロー

個々のパラメータに対する検定
（切片の t 検定）

H_0：切片は0である

H_1：切片は0ではない

有意水準α ＝ 0.05（両側検定）

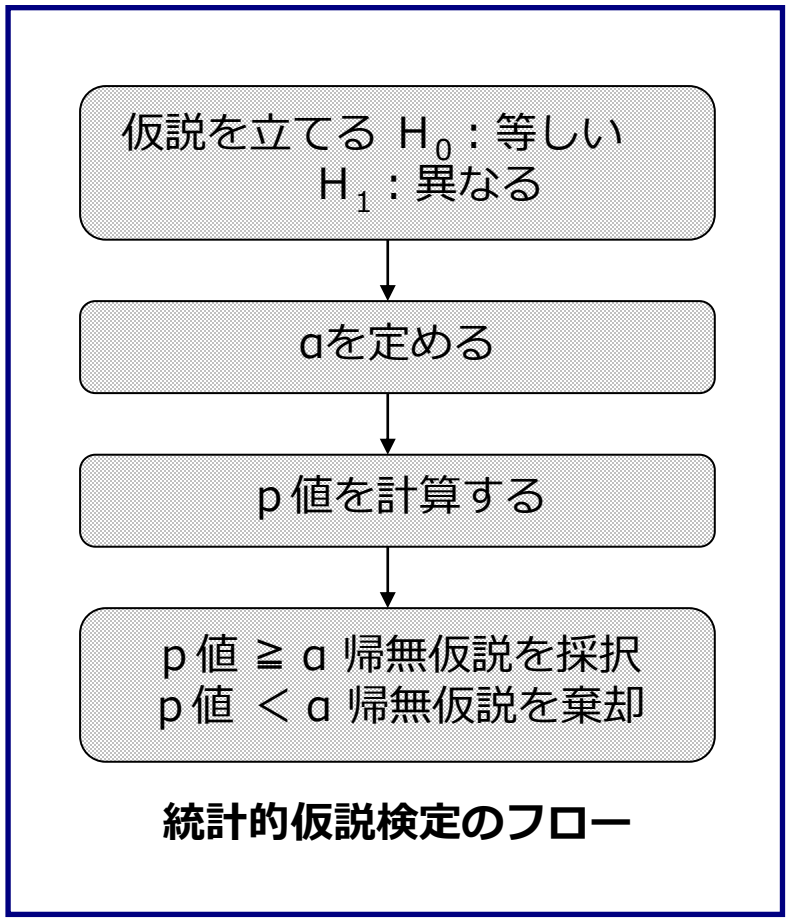
p 値 ＝ 0.070

0.070（p 値）＞0.05（有意水準α）より、帰無仮説を採択。

【結論】切片は0であるといえる。

→切片は意味がないといえる。

【Rによる実践】重回帰分析の出力～ t 検定～



統計的仮説検定のフロー

個々のパラメータに対する検定
（安打の回帰係数の t 検定）

H_0：安打の回帰係数は0である

H_1：安打の回帰係数は0ではない

有意水準α ＝ 0.05（両側検定）

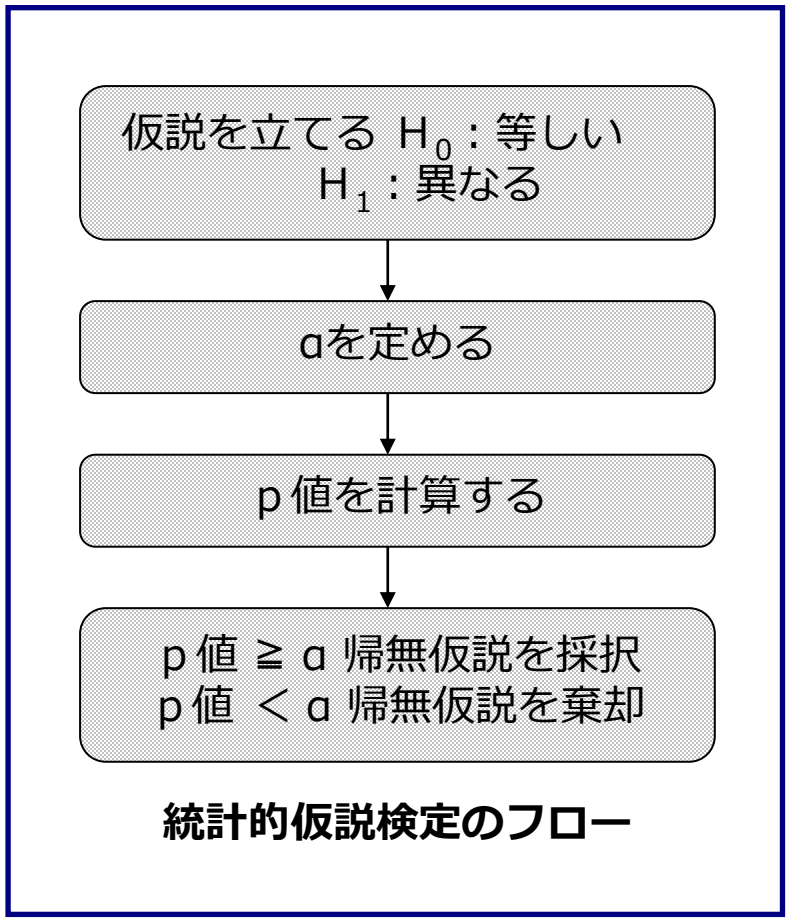
p 値 ＝ 2e-16

2e-16（ p 値）＜0.05（有意水準α）より、帰無仮説を棄却。

【結論】安打の回帰係数は0ではないといえる。

→ 安打の回帰係数は意味があるといえる。

【Rによる実践】重回帰分析の出力～ t 検定～



統計的仮説検定のフロー

個々のパラメータに対する検定
（打率の回帰係数の t 検定）

H_0：打率の回帰係数は0である

H_1：打率の回帰係数は0ではない

有意水準α ＝ 0.05（両側検定）

p 値 ＝ 0.289

0.289（ p 値）＞0.05（有意水準α）より、帰無仮説を採択。

【結論】打率の回帰係数は0であるといえる。

→ 打率の回帰係数は意味がないといえる。

１.重回帰分析（変数選択）

重回帰分析を実施する際に、有効な説明変数を選択する下記のような変数選択法があります。

総当たり法

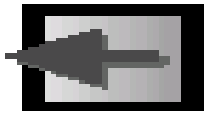
説明変数すべての組み合わせの中で、最良の回帰式を見つける方法。

変数増減法(Stepwise)

変数増加法と減少法の組み合わせ。

※要因数が非常に多い場合や探索的なモデル検討の際に有効。

変数減少法(Backward)



Ｓｔｅｐ１
全変数をモデルに含める。
Ｓｔｅｐ２
モデルの中で最も意味のない変数を除く。
Ｓｔｅｐ３
モデル中の残りの変数のＰ値がある基準以下になるまで、このプロセスを繰り返す。

変数増加法(Forward)

Ｓｔｅｐ１
全変数の中から一番良い変数を選び出す。
Ｓｔｅｐ２
最初に選択された変数を加味して、次に良い説明変数を追加する。
Ｓｔｅｐ３
Ｐ値がある基準以下の残りの変数が無くなり次第、このプロセスを止める。

【Rによる実践】変数選択（step関数）

安打と打率から年俸を予測した結果に対し、step関数を使用して変数選択し、最適なモデル選択を行います。ここでは変数増減法を指定します。

```
>step(重回帰分析, direction="both")
```

step関数はオプションとして
- 変数増加法(direction="forward")
- 変数減少法(direction="backward")
- 変数増減法(direction="both")

がある。
※step関数のデフォルトは変数減少法になっている。

出力結果

```
Start:  AIC=4566.06
base$年俸 ~ base$安打 + base$打率

            Df  Sum of Sq        RSS    AIC
- base$打率   1   36269506 8.4146e+09 4565.2
<none>                    8.3783e+09 4566.1
- base$安打   1 3407946345 1.1786e+10 4654.2

Step:  AIC=4565.2
base$年俸 ~ base$安打

            Df  Sum of Sq        RSS    AIC
<none>                    8.4146e+09 4565.2
+ base$打率   1   36269506 8.3783e+09 4566.1
- base$安打   1 4028101372 1.2443e+10 4666.5

Call:
lm(formula = base$年俸 ~ base$安打)

Coefficients:
(Intercept)     base$安打
     911.87         83.09
```

AICの値が一番小さい

最適なモデル

最後のモデルが最適なモデルとして出力される

AIC（赤池情報量基準）とは

モデルの当てはまり度を表す統計量で値が小さい程当てはまり良いとされる。
相対的な評価に用いられるため、一定の値以下であることが望ましいという基準ではない。

【Rによる実践】単回帰式と重回帰式の評価

t 検定の結果及び変数選択の結果から、 野球データについては、「打率と安打」を使った重回帰モデルより「安打」のみから「年俸」を予測する単回帰モデルの方が良さそうであると判断できます。

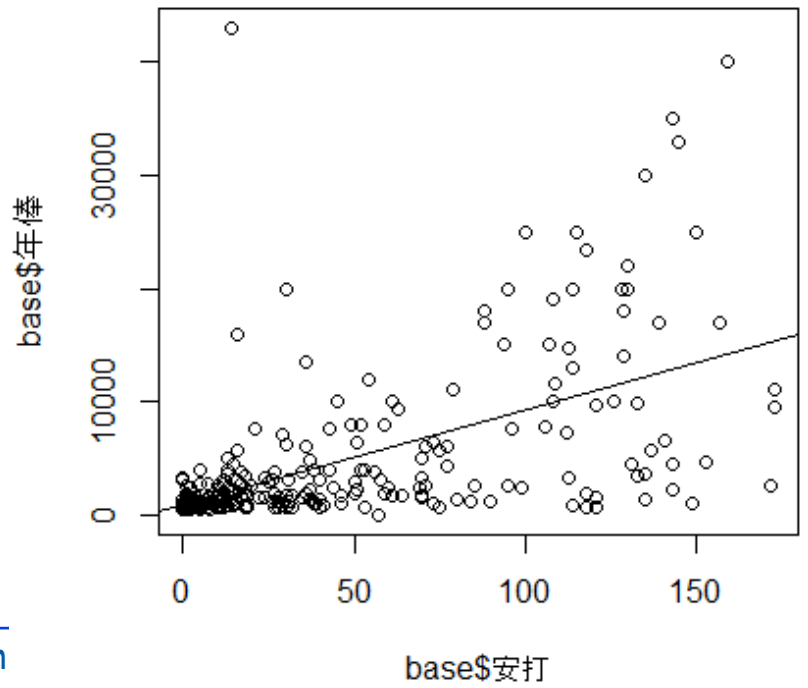
【単回帰式】
年俸＝ 911.87 ＋　83.09 × 安打

最後に回帰直線を描いてみましょう。まず、横軸に「安打」、縦軸に「年俸」の散布図を描き「単回帰結果」の結果に基づき、回帰直線をかぶせます。

```
#単回帰分析
>単回帰分析=lm(base$年俸~base$安打)

>plot(base$安打,base$年俸)
>abline(単回帰分析)
```

abline関数：
図に追記するための関数

<安打と年俸に関する散布図と回帰直線>



1.重回帰分析（説明変数の検討）

＜単回帰＞

パラメータ推定値

変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)
Intercept	1	889.7722	433.9545	2.05	0.0413
TOKUTEN	1	219.0796	16.3669	13.39	<.0001

寄与率	0.4061

＜重回帰＞

パラメータ推定値

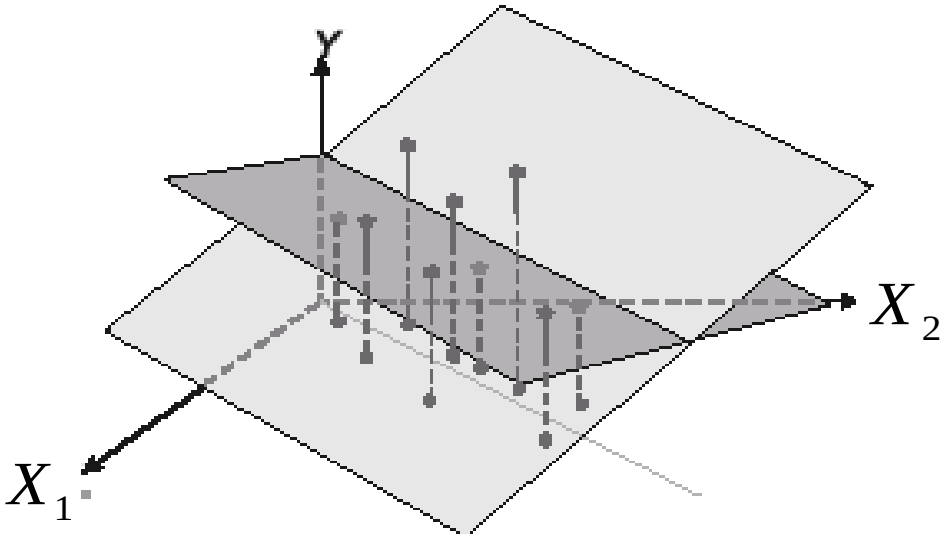
変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)
Intercept	1	336.4991	414.2738	0.81	0.4174
TOKUTEN	1	-74.6162	48.9240	-1.53	0.1284
ANDA	1	130.5738	20.6642	6.32	<.0001

寄与率	0.4849

得点のp値は、単回帰の場合“<.0001”と小さかったが、重回帰分析で安打数を加えると、寄与率は上がったが得点のp値は0.1284と大きく有意水準αが10%でも帰無仮説を棄却できない。

単回帰の場合は得点が年俸のバラツキを良く説明していたが、安打数が追加された場合、得点の説明力が小さくなったことを意味します。説明変数同士に相関があるために、このような現象が生じます。これを“多重共線性”と呼びます。

1.重回帰分析（多重共線性）



YとX$_1$、YとX$_2$の2つの平面があるとする。
この2つの変数が同じ方向性の場合、情報が重複する。すなわち共線性があるとみなす。

どうして問題なのか？

1．単回帰の場合は重要であっても、同じ方向性を持つ変数が両方入っていると、どちらも重要でないと結論を出してしまうことがある。よって、重要な変数を見過ごしてしまう可能性がある。

2．共線性は、パラメータ推定値のバラツキを増加させ、よって予測の上での誤差も増やすことになる。

3．偏回帰係数の解釈が難しくなる。

1.重回帰分析（多重共線性）

共線性を発見するのに“ＶＩＦ”という統計量があります。

パラメータ推定値

変数	自由度	推定値	標準誤差	t 統計量	p 値 (t)	トレランス	VIF
Intercept	1	336.4991	414.2738	0.81	0.4174	.	0
TOKUTEN	1	-74.6162	48.9240	-1.53	0.1284	0.0974	10.2629
ANDA	1	130.5738	20.6642	6.32	<.0001	0.0974	10.2629

分散拡大係数

ＶＩＦ（Variance Inflation Factors）

共線性による、パラメータ推定値への バラツキの影響度を示す統計量。

$$VIF = \frac{1}{1 - r^2}$$

VIFが10以上になる場合は要注意で、説明変数を除外した方がよいか検討が必要になります。ただし、共線性の判断は絶対的な基準があるわけではなく、場合によって様々に異なります。最適なモデルを判断するために、それぞれの変数や相互の関連のもつ特徴なども加味した柔軟な対応が必要です。

＜共線性の対応策＞

- ✓ 説明変数のどちらか一方だけを削除する
- ✓ 説明変数の合成変数を作成する（→合計、平均、差分）

例

$$\frac{\text{ホームラン数}}{\text{ヒット数}} = \text{ホームラン率}$$

2013年年俸－2012年年俸＝増減額

【Rによる実践】共線性の発見～“ＶＩＦ”～

野球データの安打と塁打から年俸を予測した結果に対し、vif関数を使用してVIFの値を確認してみます。まず、安打と塁打から年俸を予測する重回帰分析を実施します。

```
>安打と塁打=lm(base$年俸~base$安打+base$塁打)
>summary(安打と塁打)
```

出力結果

```
Call:
lm(formula = base$年俸 ~ base$安打 + base$塁打)

Residuals:
   Min     1Q Median     3Q    Max
-12104  -1855   -534    503  41029

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1033.89     473.03   2.186  0.02973 *
base$安打     -27.58      41.54  -0.664  0.50732
base$塁打      77.85      28.76   2.707  0.00724 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5600 on 261 degrees of freedom
  (95 observations deleted due to missingness)
Multiple R-squared: 0.3422,     Adjusted R-squared: 0.3372
F-statistic: 67.89 on 2 and 261 DF,  p-value: < 2.2e-16
```

安打の係数には意味がないという結果

【Rによる実践】共線性の発見～“ＶＩＦ”～

野球データの安打と塁打から年俸を予測した結果に対し、vif関数を使用してVIFの値を確認してみます。パッケージから”DAAG”を追加インストールし、パッケージの読み込みを行ってからvif関数を使用してください。

```
>library(DAAG)
>vif(安打と塁打)
```

出力結果

```
base$安打 base$塁打
   32.102    32.102
```

VIFが10以上になる場合は要注意で、説明変数を除外した方がよいか検討が必要になります。ただし、共線性の判断は絶対的な基準があるわけではなく、場合によって様々に異なります。最適なモデルを判断するために、それぞれの変数や相互の関連のもつ特徴なども加味した柔軟な対応が必要です。

【練習問題10】重回帰分析

野球データ（【2012年】全球団内野手.csv）以下の項目を使って、年俸を予測する最適な回帰式を作成してください。

（変数増減法を使用します）

得点、安打、本塁打、盗塁、長打率、出塁率

（参考）ロジスティック回帰分析

ロジスティック回帰分析とは？

目的変数が質的変数（通常2値）であり、その目的変数と説明変数の関連性を、ある曲線を当てはめモデル化し、確率を予測するものです。

- 2値変数とは…「知っている⇔知らない」「好き⇔嫌い」「ある⇔ない」「賛成⇔反対」など、2つのカテゴリで成り立つ変数
- ロジスティック回帰では、この2値変数を1,0と置き換えたものを目的変数とする

  →ただし、回帰分析をする場合は、目的変数が連続数量である必要がある
- そこで、まず1,0という目的変数を『1になる確率』に置き換えてみる

  →連続変数にはなったが0~1の間しかとらないのでダメ
- さらに、確率の「ロジット」を取ってみる

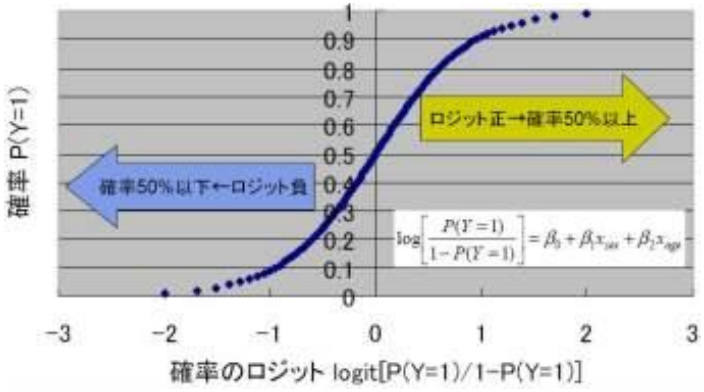
$$\text{logit}(\text{Yが１になる確率}) = \log\left(\frac{\text{Yが１になる確率}}{1-(\text{Yが１になる確率})}\right)$$

  →－∞~＋∞の値をとる目的変数になった！
- この変換した目的変数に対して回帰分析を実行する

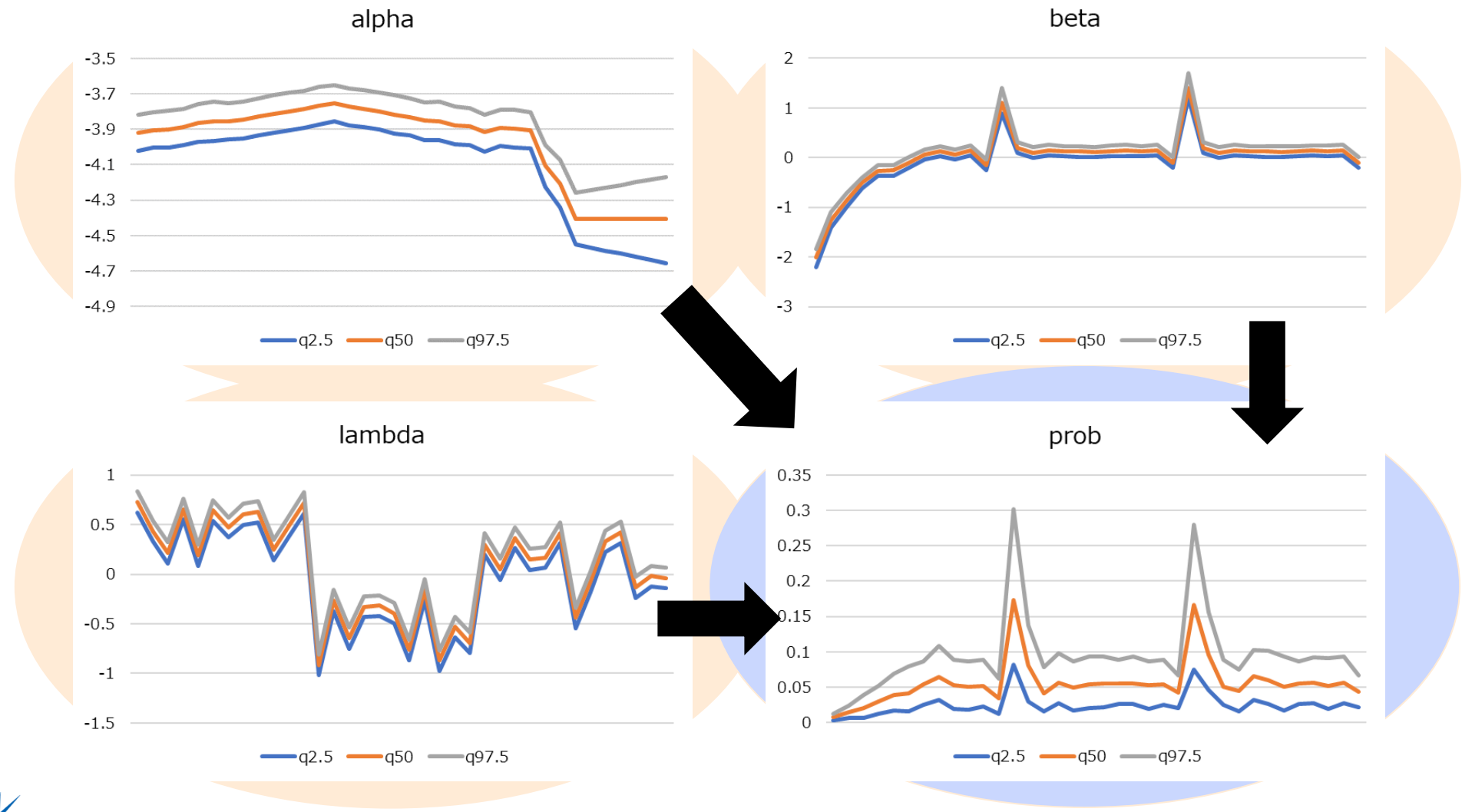
$$\text{logit}(\text{Yが１になる確率}) = \log\left(\frac{\text{Yが１になる確率}}{1-(\text{Yが１になる確率})}\right) = \beta_0 + \beta_1 x$$

  →この式を、ロジットモデル＝ロジスティック回帰モデルと呼ぶ

確率とロジットの関係



（参考）ロジスティック回帰分析(案件での活用事例)



演習問題

演習問題：二変数の関連の分析

1. businessデータに対して、国籍（NATION）と産業（INDUSTRY）のクロス集計表を出力してください。

2. 野球データ（【2012年】全球団内野手.csv）を用いて、2012年の年俸とその他の数値変数（翌年年俸から出塁率まで）との関連を調べてください。また算出した相関係数について2012年の年俸と相関の高いものを順に3つ挙げて下さい。

★参照スライド：クロス集計（P31）、相関係数（P17,18）

演習問題：単回帰分析

1.サンプルデータtelは、以下の項目を持つ電話利用者データです。電話利用者の利用料金を予測する単回帰モデルを作成してください。
予測モデルを作成するにあたり、No.1～7までの7個の変数を説明変数とし、「直近6ヶ月利用金額合計」を目的変数として、最も当てはまりの良い単回帰モデルを作成してください。

NO	項目名
1	年齢
2	直近1ヶ月_利用時間
3	直近1ヶ月_利用回数
4	直近1ヶ月ローミング利用金額割合
5	基本料金割合
6	平均利用時間
7	平均利用回数
8	直近6ヶ月利用金額合計

★参照スライド：単回帰モデル（P56、P78）

演習問題：重回帰分析

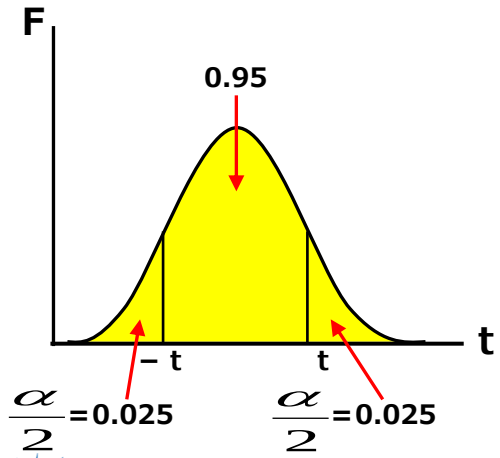
1.サンプルデータicecreamは、40人の中学生に対し、アイスクリームの一週間の消費量（Spending）、子供の数（Kids）、家庭の収入（Income）を調べた調査データです。

①収入と子供の数が消費量に与える影響を、重回帰分析を実施して予測モデルを作成してください。
②性別毎（Gender）と学年毎（Grade）に分けた場合、上記①で算出した予測式に影響があるでしょうか。

★参照スライド：重回帰モデル（P71、P77）

添付資料

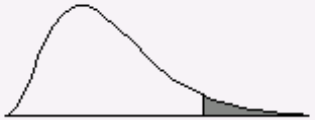
添付資料： t 分布表



df	t値			df	t値		
5	2.015	2.571	4.032	24	1.711	2.064	2.797
6	1.943	2.447	3.707	25	1.708	2.060	2.787
7	1.895	2.365	3.499	26	1.706	2.056	2.779
8	1.860	2.306	3.355	27	1.703	2.052	2.771
9	1.833	2.262	3.250	28	1.701	2.048	2.763
10	1.812	2.228	3.169	29	1.699	2.045	2.756
11	1.796	2.201	3.106	30	1.697	2.042	2.750
12	1.782	2.179	3.055	35	1.690	2.030	2.724
13	1.771	2.160	3.012	40	1.684	2.021	2.704
14	1.761	2.145	2.977	45	1.680	2.014	2.690
15	1.753	2.131	2.947	50	1.676	2.008	2.678
16	1.746	2.120	2.921	55	1.673	2.004	2.669
17	1.740	2.110	2.898	60	1.671	2.000	2.660
18	1.734	2.101	2.878	70	1.667	1.994	2.648
19	1.729	2.093	2.861	80	1.665	1.989	2.638
20	1.725	2.086	2.845	90	1.662	1.986	2.631
21	1.721	2.080	2.831	100	1.661	1.982	2.625
22	1.717	2.074	2.819	120	1.658	1.980	2.617
23	1.714	2.069	2.807	∞	1.645	1.960	2.576
α 有意水準	0.1	0.05	0.01	有意水準	0.1	0.05	0.01

添付資料：カイ二乗分布表

カイ自乗(χ^2)分布表



自由度と右側確率値に対応する χ^2 値

	右側確率値					
自由度 ν	0.995	0.975	0.05	0.025	0.01	0.005
1	0.00003927	0.0009821	3.8415	5.0239	6.6349	7.8794
2	0.01003	0.05064	5.9915	7.3778	9.2103	10.5966
3	0.07172	0.2158	7.8147	9.3484	11.3449	12.8382
4	0.2070	0.4844	9.4877	11.1433	13.2767	14.8603
5	0.4117	0.8312	11.0705	12.8325	15.0863	16.7496
6	0.6757	1.2373	12.5916	14.4494	16.8119	18.5476
7	0.9893	1.6899	14.0671	16.0128	18.4753	20.2777
8	1.3444	2.1797	15.5073	17.5345	20.0902	21.9550
9	1.7349	2.7004	16.9190	19.0228	21.6660	23.5894
10	2.1559	3.2470	18.3070	20.4832	23.2093	25.1882
11	2.6032	3.8157	19.6751	21.9200	24.7250	26.7568
12	3.0738	4.4038	21.0261	23.3367	26.2170	28.2995
13	3.5650	5.0088	22.3620	24.7356	27.6882	29.8195
14	4.0747	5.6287	23.6848	26.1189	29.1412	31.3193
15	4.6009	6.2621	24.9958	27.4884	30.5779	32.8013
16	5.1422	6.9077	26.2962	28.8454	31.9999	34.2672

添付資料：統計記号

	母集団パラメータ	サンプル統計量
平均	μ	$\overline{x}$
分散	σ^2	s^2
標準偏差	σ	s

μ ： 平均（mean）　「ミュー」と読む。
母集団の平均値。通常は未知。

σ^2 ： 分散（variance）　σは「シグマ」と読む。
母集団の分散。通常は未知。

n ： サンプル数（sample size）
任意の母集団から採ったデータ（サンプル）の数

x_i ： オブザベーション（observation）i = 1, 2, … , n
サンプルの個々の値、測定値。

$\overline{x}$ ： 平均の推定量（estimator of mean）「エックスバー」と読む。
（母集団の）平均の推定量。サンプルから計算。

S^2 ： 分散の推定量（estimator of variance）
（母集団の）分散の推定量。サンプルから計算。

df ： 自由度（degree of freedom）
母集団の平均の検定（ t 検定）を行うとき　df = n − 1

添付資料：用語説明

1．説明変数と目的変数は、以下のとおり別の表現方法があります。

【説明変数 x】説明変数、独立変数、要因

【目的変数 y】目的変数、応答変数、従属変数、基準変数

2．偏回帰係数、標準偏回帰係数とは

偏回帰係数は、他の説明変数の影響を除き、説明変数の値が1だけ変わったとき、目的変数の予測値がどれだけ変化するかを表している値である。
単位に依存しない偏回帰係数を「標準偏回帰係数」という。

※重回帰分析の目的の1つは，どの説明変数が最もよく目的変数を説明しているかを知るということである。しかし，偏回帰係数の大きさでもって判断してはならない。例えば、ある説明変数の測定単位をグラムからキログラムに変えると，その説明変数に対する偏回帰係数は 1/1,000 になるし，グラムで測られる説明変数とセンチメートルで測られる説明変数の比較はできない。
このような場合には、目的変数およびそれぞれの説明変数が、平均値=0、分散=1 に標準化（正規化）されているとしたときの偏回帰係数を求めます。これを「標準化偏回帰係数」と呼びます。