HITACHI
Inspire the Next

データ分析手法の理論と適用

株式会社 日立インフォメーションアカデミー

HITACHI
Inspire the Next

Contents

1. はじめに
2. データ分析手法の分類
3. データ分析手法の適用
4. 確率分布と検定
5. 実務へ向けて

1-2.　データ分析力とは何か

データ分析力は分析計画を立てる思考力と分析作業を進める作業力に分けられる。
なかでも、「手を動かす前にいかに分析計画を考えるか」の思考力に本研修では焦点を当てる。

思考力

インプット

○○を分けて比べたい

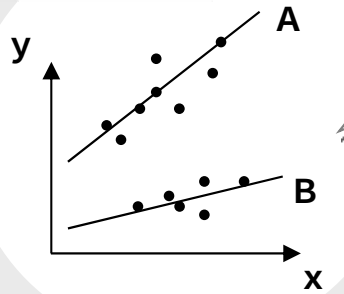
○○は××によって説明できるはずだ

行と列に目的や仮説が表現される

アウトプット

AがBより上にあるだろう

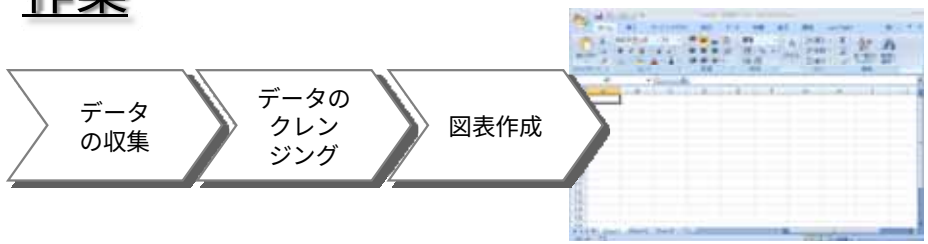
それは○○を意味するはずだ



意味合いや新たな仮説を抽出する

作業力

作業



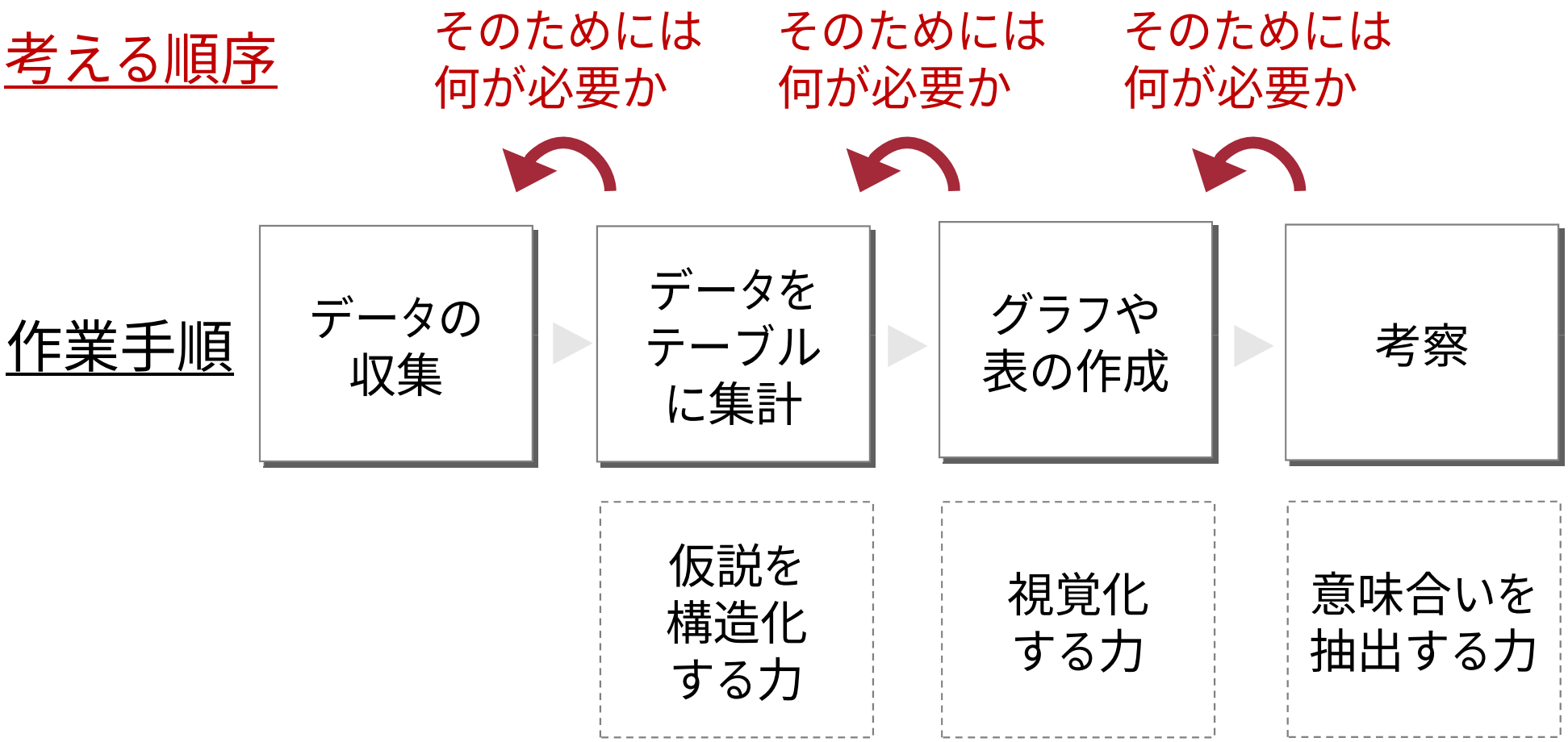
理論

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

1-3.　分析計画を立てる思考力とは何か

HITACHI
Inspire the Next

分析計画を立てる際には作業手順と逆順での思考が求められる。その思考力は仮説を構造化する力、視覚化する力、意味合いを抽出する力に分けられる。



2-1.　データの分類

区別をするために用いる尺度　(e.g. 1.男/2.女)

・カテゴリデータとも呼ばれ、数値は分類としての意味のみ持つ
・大小に意味はなく、四則演算できない
・平均値、中央値、標準偏差に意味をもたない

大小、強弱などの順位関係を表す尺度
(e.g. 1.良い/2.ふつう/3.悪い)

・数値の順位に意味を持つ
・一方、数値の間隔に意味をもたず、四則演算できない
・平均値、標準偏差に意味はなく、中央値のみが意味を持つ

任意の基準だが、一定の測定単位をもつ尺度
(e.g. 摂氏、年号、テスト)

・一般的な数値データを指す
・加減算はできるが乗除算はできない
・平均値、標準偏差に意味を持ち、様々な分析・検定手法が適用できる

一定の測定単位をもつ、かつ絶対的な原点0をもつ尺度
(e.g. 人口、売上高)

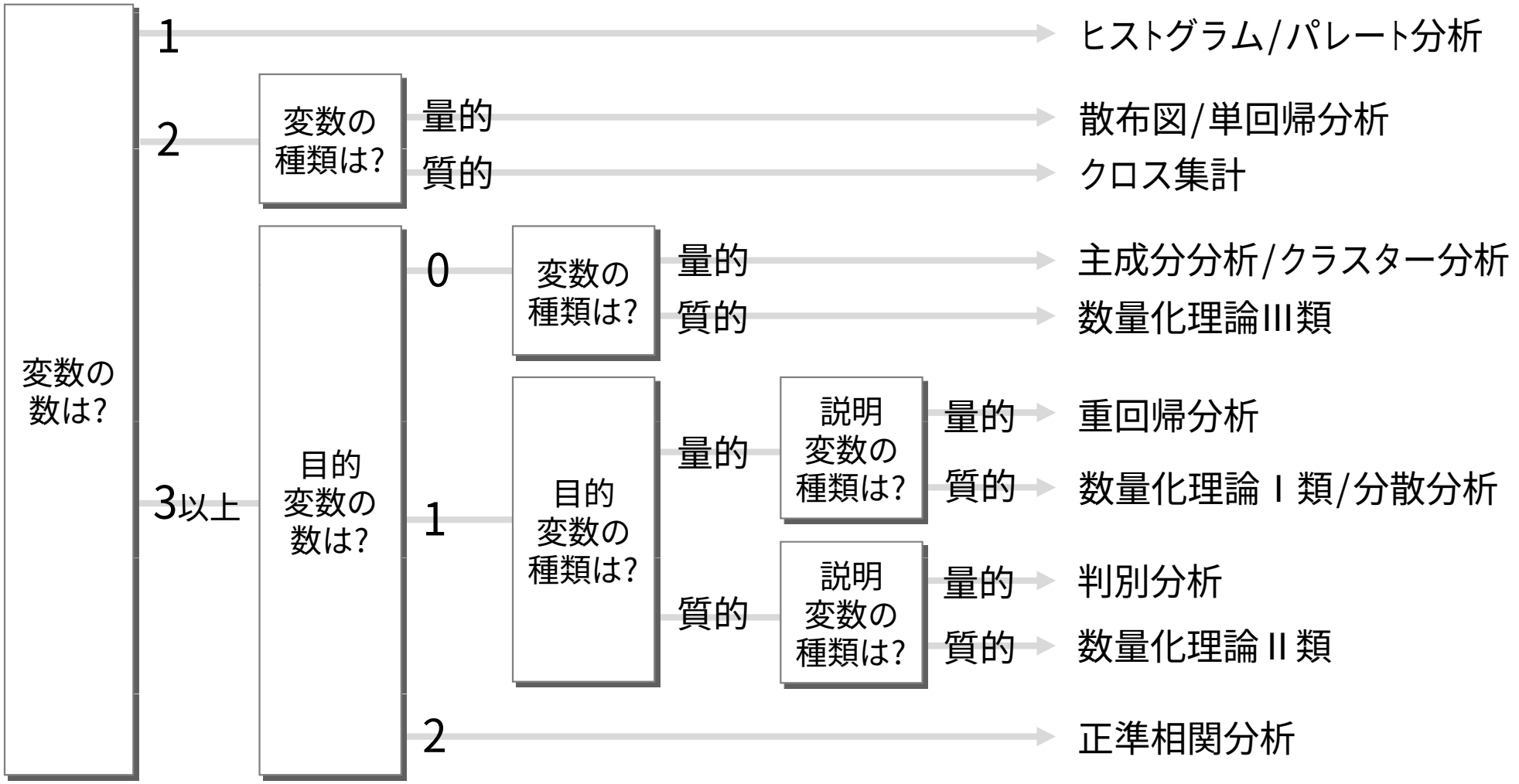
・四則演算ができる
・平均値、標準偏差に意味を持ち、様々な分析・検定手法が適用できる

2-2.　目的によるデータ分析手法の分類

HITACHI
Inspire the Next

特徴を把握したい	✓基本統計量 ✓ヒストグラム ✓時系列分析 ✓パレート分析	・品質・納期遅延の平均とばらつきがどれだけあるか知りたい ・重要顧客を特定したい ・バグの発生率が高い処理や機能を特定したい
予測したい	✓回帰分析 ✓相関分析 ✓数量化理論Ⅰ類 ✓待ち行列分析	・バグ密度が自動生成率や難度などにどの程度影響されているか推測したい ・サーバ性能の予測式をつくり、ボトルネックを決定したい
分類したい	✓判別分析 ✓クラスター分析	・顧客やユーザーを分類し、分類したグループごとの対応を検討したい
集約したい	✓主成分分析 ✓因子分析	・アンケートの結果から、満足度や問題点をはかる総合指標を抽出したい

2-3.　データの種類によるデータ分析手法の分類



HITACHI
Inspire the Next

3
データ分析手法

- ✓ヒストグラムと基本統計量
- ✓時系列分析
- ✓パレート分析
- ✓単回帰分析と散布図
- ✓重回帰分析
- ✓数量化理論Ⅰ類
- ✓判別分析
- ✓クラスター分析
- ✓主成分分析

3-5-2. 重回帰分析のアウトプット

単回帰分析と同様に、目的変数(y)と説明変数(x1,x2,x3・・・)の関係式とその関係性の強さから意味合いを抽出する。

比較結果

回帰統計	
重相関 R	0.98
重決定 R2	0.96
補正 R2	0.95
標準誤差	17.35
観測数	10

CPUクロック数、HDD容量、ディスプレイサイズで
中心価格は96%説明できる
中心価格=1.73×CPUクロック数＋26.44×HDD容量
＋6.39×ディスプレイサイズー375.18

意味合い/再仮説

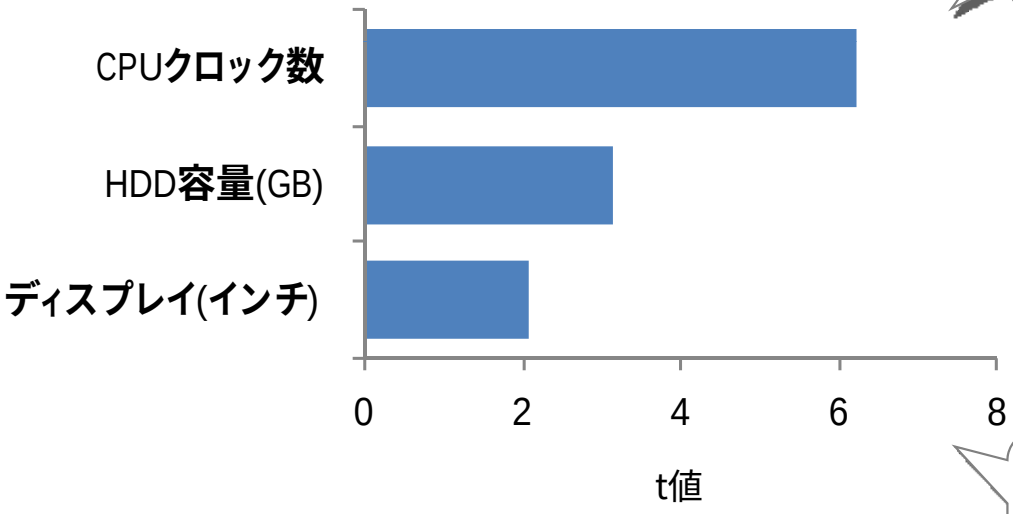
今後、CPUクロック数、HDD容量、共に最高水準の
商品を投入すると、中心価格はいくらぐらいになるか

分散分析表	自由度	変動	分散	観測された分散比	有意 F
回帰	3	49001	16334	54.2	
残差	6	1805	301		
合計	9	50807			

	係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95%	上限 95%
切片	-375.18	63.82	-5.88	1.1E-03	-531.34	-219.02	-531.34	-219.02
CPU クロック数	1.73	0.28	6.21	8.1E-04	1.05	2.41	1.05	2.41
HDD容量 (GB)	26.44	8.41	3.14	2.0E-02	5.86	47.02	5.86	47.02
ディスプレイ (インチ)	6.39	3.10	2.06	8.5E-02	-1.19	13.98	-1.19	13.98

3-5-3. 重回帰分析のアウトプット

回帰式に加え、説明変数同士の目的変数への影響の強さを見ることができ、感度分析と呼ぶ。



比較結果

それぞれ8.1×10^{-4}、2.0×10^{-2}、8.5×10^{-2}の確率で中心価格の予測に使う意味がない

意味合い/再仮説

CPUクロック数、HDD容量、ディスプレイサイズの順に中心価格への影響が高い

3-5-4. インプット

重回帰分析をする際には、何と何が関係がありそうかを考える必要がある。そこで、
目的となる指標に対して、関係がある指標は何かという仮説を網羅的に挙げる。

目的

PCの中心価格が何に影響されて
決まるか知りたい
PCの中心価格を予測して
開発計画を立てたい

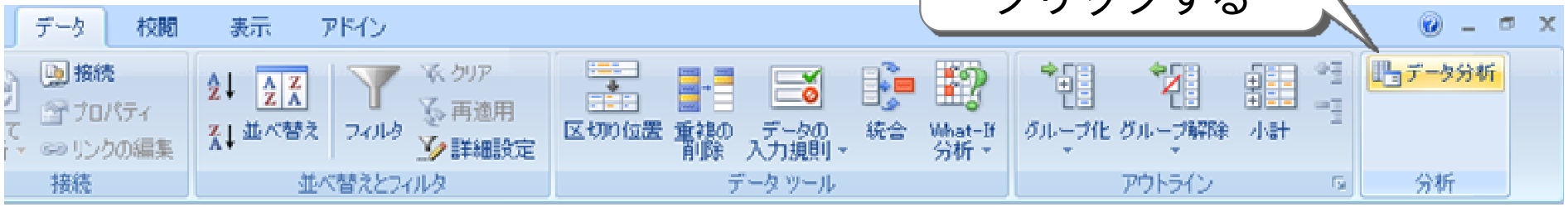
仮説

PCの中心価格はCPUクロック数と
HDD容量とディスプレイサイズで
おおむね決まっているはずだ
なぜならPCの中心価格に対し
CPUクロック数は．．．

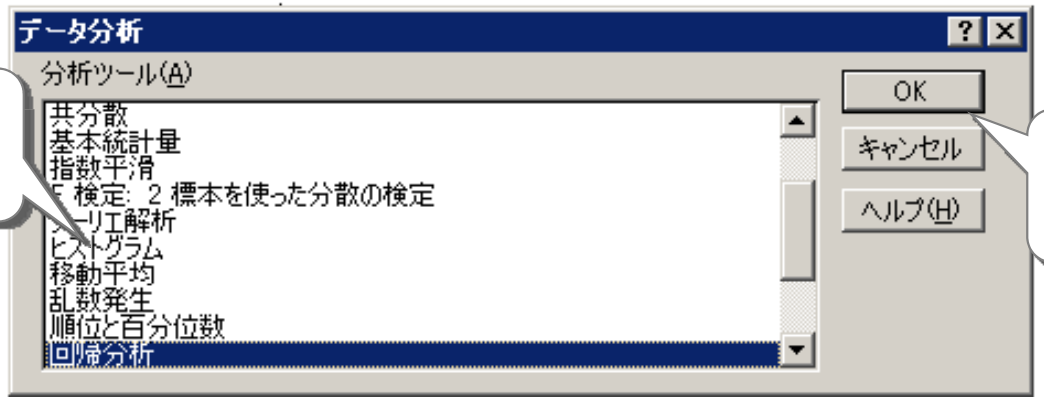
	C	D	E	
1				
2	標本No.	中心価格（千円）	CPUクロック数	HDD容量（GB）	ディスプレイ（インチ）
3	1	58.0	166	1.6	15
4	2	69.8	166	2.0	17
5	3	178.0	200	3.0	19
6	4	152.0	200	2.0	19
7	5	280.0	233	4.2	19
8	6	62.0	166	1.7	15
9	7	78.0	166	2.6	19
10	8	198.0	233	3.2	17
11	9	180.0	200	4.3	15
12	10	198.0	233	3.1	15

3-5-5. アウトプット作成の作業

1 データ分析をクリックする



2 回帰分析をクリックする



3 OKをクリックする

3-5-6. アウトプット作成の作業

HITACHI
Inspire the Next

標本No.	中心価格（千円）	CPUクロック数	HDD容量（GB）	ディスプレイ（インチ）
1	58.0	166	1.6	15
2	69.8	166	2.0	17
3	178.0	200	3.0	19
4	152.0	200	2.0	19
5	280.0	233	4.2	19
6	62.0	166	1.7	15
7	78.0	166	2.6	19
8	198.0	233	3.2	17
9	180.0	200	4.3	15
10	198.0	233	3.1	

4

入力Y範囲に中心価格のデータすべてを選択する

5

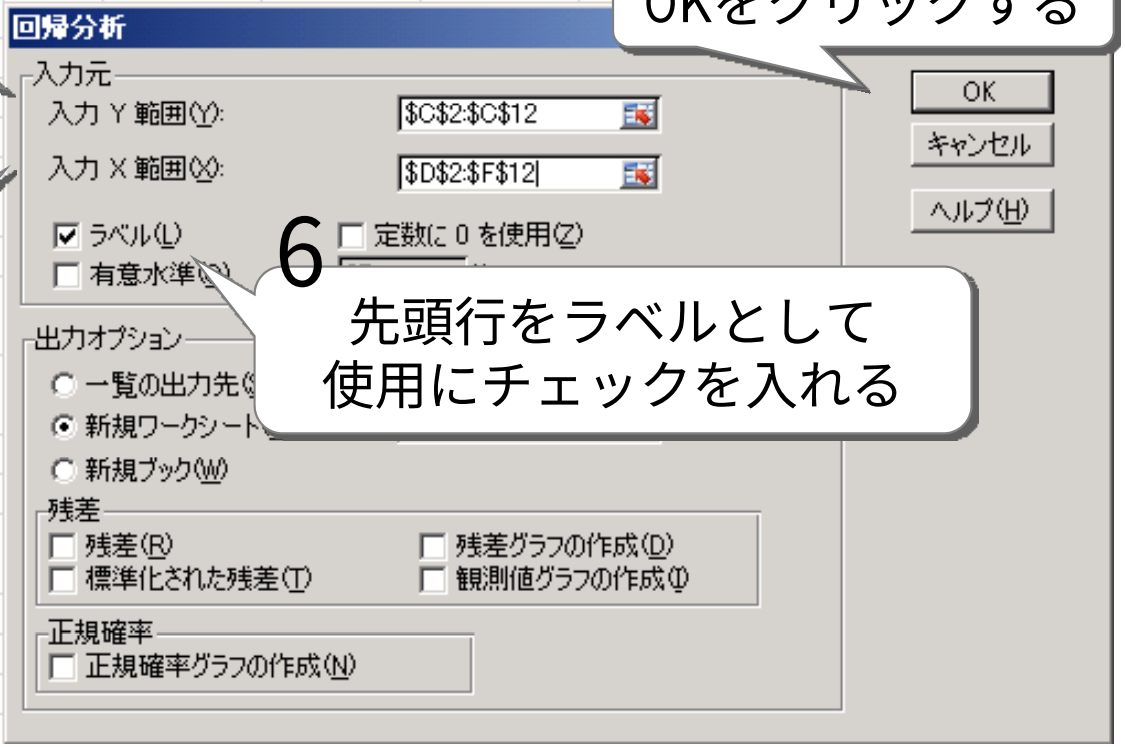
入力X範囲にCPUクロック数からディスプレイのデータすべてを選択する

6

先頭行をラベルとして使用にチェックを入れる

7

OKをクリックする



3-5-7. 重回帰分析の回帰式の理論

重回帰分析の回帰式は単回帰分析と同様、最小自乗法という方法によって決定される。

	A	B	C	D	E	F
1						
2		標本 No.	中心価格（千円）	CPU クロック数	HDD容量（GB）	ディスプレイ（インチ）
13		個数	10			
14		平均	145.4	196	2.8	17
15		分散	5080.7	775	0.8	3
16		共分散yx		1858.2	54.90	38.0
17		共分散xx			18.0	6.8
18						0.22
19		相関yx		0.94	0.84	0.30
20		相関xx			0.70	0.14
21						0.13

	C	D	E	F	G
22					
23	分散共分散行列	775.41	18.04	6.80	1858.19
24		18.04	0.85	0.22	54.90
25		6.80	0.22	3.20	38.00

	C	D	E	F	G
22					
23	分散共分散行列	=D15	=E17	=F17	=D16
24		=E23	=E15	=F18	=E16
25		=F23	=F24	=F15	=F16

	C	D	E	F	G
22					
23	分散共分散行列	=分散x1	=共分散x1x2	=共分散x1x3	=共分散x1Y
24		=共分散x2x1	=分散x2	=共分散x2x3	=共分散x2Y
25		=共分散x3x1	=共分散x3x2	=分散x3	=共分散x3Y

いま解きたいのは、$y=ax_1+bx_2+cx_3+d$の
a,b,c,dの値を求めることであり、そのために
作る方程式が分散共分散行列である。

分散共分散行列は、
775.41a+18.04b+6.80c=1858.19
18.04a+0.85b+0.22c=54.90
6.80a+0.22b+3.20c=38.00
を意味し、この3つの式を解くことで、
a,b,c,dの値を求める。

3-5-8. 重回帰分析の回帰式の理論

a,b,c,dは単位行列を導くことで解ける。単位行列とは、正方行列(行と列の数が同じ行列)の対角線が1、その他が0の行列を指す。

	C	D	E	F	G
22					
23	分散共分散行列	775.41	18.04	6.80	1858.19
24		18.04	0.85	0.22	54.90
25		6.80	0.22	3.20	38.00
26					
27		1.00	0.02	0.01	2.40
28		0.00	0.43	0.06	11.67
29		0.00	0.06	3.14	21.70

行列計算のルール1

行を同じ数で割ってよい

	D	E	F	G
27	=D23/$D23	=E23/$D23	=F23/$D23	=G23/$D23

行列計算のルール2

ある行とある行の加減算をしてよい

	D	E	F	G
29	=D25-D$27*$D25	=E25-E$27*$D25	=F25-F$27*$D25	=G25-G$27*$D25

	D	E	F	G
34				
35	1.00	0.00	0.00	1.73
36	0.00	1.00	0.00	26.44
37	0.00	0.00	1.00	6.39
38				
39				-375.18

行列計算の2つのルールを繰り返し適用することで、単位行列が得られる。求まった単位行列は、
1a+0b+0c=1.73
0a+1b+0c=26.44
0a+0b+1c=6.39
を意味するため、a,b,cが求まったこととなる。

3-5-9. 重回帰分析の回帰統計の理論

相関係数や決定係数などの回帰統計は回帰式の精度を示す。回帰統計は実測のyと予測のY(Y=$ax_1+bx_2+\cdots$で求めたもの)を比べることで求められる。

	H	I	J	K	L
1					
2		予測 中心価格	予測誤差	予測誤差 の2乗	偏差平方
3		49.4	8.6	74.14	7635.26
4		72.7	−2.9	8.68	5712.34
5		170.6	7.4	54.37	1064.06
6		144.2	7.8	61.00	43.82
7		259.3	20.7	429.05	18122.54
8		52.0	10.0	99.34	6952.22
9		101.4	−23.4	547.13	4540.06
10		220.1	−22.1	486.96	2768.86
11		179.4	0.6	0.32	1198.54
12		204.6	−6.6	44.10	2768.86
13	合計		0.00	1805.11	50806.60
14	平均	145.4	0.00	180.51	5080.66
15	分散	4900.15			
16	共分散(y−Y間)	4900.15			
17	重相関係数	0.98			
18	標準誤差	17.35			
19	重決定係数	0.96			
20	補正決定係数	0.95			

	H	I
17	重相関係数	=実測Yと予測yの相関
18	標準誤差	$=\sqrt{\frac{\text{予測誤差の平方和}}{\text{個数}-\text{説明変数数}-1}}$
19	重決定係数	$=1-\frac{\text{予測誤差の平方和}}{\text{実測}Y\text{の偏差平方和}}$
20	補正決定係数	$=1-\frac{\text{予測誤差の平方和}/(\text{個数}-\text{説明変数数}-1)}{\text{実測}Y\text{の偏差平方和}/(\text{個数}-1)}$

	H	I
17	重相関係数	=CORREL(C3:C12,I3:I12)
18	標準誤差	=(K13/(C13−COUNTA(D2:F2)−1))^0.5
19	重決定係数	=1−K13/L13
20	補正決定係数	=1−I18^2/(L13/(C13−1))

HITACHI
Inspire the Next

4

確率分布と検定

- ✓様々な確率分布と検定手法
- ✓区間推定と統計的有意性
- ✓t検定(平均値の差の検定)
- ✓クロス集計とχ^2検定(独立性の検定)
- ✓F検定(等分散の検定)
- ✓シミュレーションへの確率分布の適用

4-1-4. 検定手法の分類

		名義尺度のデータ	順序尺度のデータおよび間隔/比例尺度のデータのうち正規分布しないもの	間隔･比例尺度のデータのうち正規分布といえるもの
1試料		χ^2検定	Kolmogorov-Smirnovの1試料検定	z検定/t検定
2試料	独立	χ^2検定	Mann-Whitneyの U検定	t検定
	対応	McNemaの検定	Wilcoxonの符号付き順位和検定	t検定
多試料	独立	χ^2検定	Kruskal-Wallisの検定	F検定
	対応	CochranのQ検定	Friedmannの検定	F検定

4-4-6. χ²検定(独立性の検定)

集計結果だけでは、本当に偏りがあるといっていいのか分からない。これを定量的に
評価するのが独立性の検定である。独立性の検定では集計結果と期待度数を比較する。

クロス集計

	E	F	G	H	I
105			リピート		
106			有	無	合計
107	性別	男性	26	32	58
108		女性	8	34	42
109		合計	34	66	100

期待度数

	E	F	G	H	I
111			リピート		
112			有	無	
113	性別	男性	19.7	38.3	58
114		女性	14.3	27.7	42
115			34	66	100

もし、男女によって
リピートの有無に
全く関係がないとしたら
期待される人数は何人

	G	H
113	=G$109*$I107/$I$109	=H$109*$I107/$I$109
114	=G$109*$I108/$I$109	=H$109*$I108/$I$109

	G	H
113	= リピートあり*男性 / 全人数	= リピートなし*男性 / 全人数
114	= リピートあり*女性 / 全人数	= リピートなし*女性 / 全人数

4-4-7. χ²検定の作業

	E	F	G	H
117			リピート	
118			有	無
119	性別	男性	2.000	1.030
120		女性	2.762	1.423
121		χ^2値	7.215	
122		P値	0.007	

	G	H
119	=(G107-G113)^2/G113	=(H107-H113)^2/H113
120	=(G108-G114)^2/G114	=(H108-H114)^2/H114
121	=SUM(G119:H120)	
122	=CHIDIST(H121,(COUNTA(F107:F108)-1)*(COUNTA(G106:H106)-1))	

	G
120	$=\frac{(実測値-期待度数)\text{^}2}{期待度数}$
121	$=\sum\frac{(実測値-期待度数)\text{^}2}{期待度数}$

ここでは有意水準を0.05(95%有意)とする。標本分散の確率分布を用いて、差がないといえるかを検定する。つまり、分散が有意水準より大きいかを検定すればよく、
片側検定の0.05と比較する。

すると、0.007＜0.05のため、帰無仮説「性別による再購入の有無に差がない」は
棄却され、性別によって再購入するかしないかに差があるといえる。