統計分析講座

福山平成大学

尾崎誠・福井正康

集計と検定編

１．データの集計

1.1　質的データの集計

基礎

単純集計　　１次元分割表　→　棒グラフ（値重視），円グラフ（割合重視）

クロス集計　２次元分割表　→　積み重ね棒グラフ

例

２０人に以下のようなアンケートを取った。入力フォームを Excel で作成せよ。

質問１　あなたの性別は。

１．男性　　２．女性

質問２　あなたは学校改革案に賛成ですか。

１．はい　　２．いいえ　　３．どちらともいえない

性別	回答	性別	回答	性別	回答	性別	回答
1	1	2	3	1	2	2	1
2	1	1	1	1	2	1	2
1	2	2	2	2	1	1	1
1	2	2	3	2	1	2	3
2	1	1	1	1	3	1	1

入力されたデータを College Analysis に移し、以下の問いに答えよ。

１）回答に関する１次元分割表を描け。

２）１）の分割表を用いて棒グラフと円グラフを描け。

３）性別と回答に関する２次元分割表を描け。

４）３）の分割表を用いて積み重ね棒グラフを描け。

問題

Samples¥テキスト 9.txt を用いて以下の問いに答え、結果は文書にまとめよ。但し、地域について１：市街、２：郊外、意見１について１：賛成、２：反対、意見２について１：はい、２：いいえ、３：どちらとも（いえない）とする。

１）地域に関する１次元分割表を描け。

市街	郊外	合計

2）意見１に関する１次元分割表を描け。

賛成	反対	合計

3）意見２に関する１次元分割表を描け。

はい	いいえ	どちらとも	合計

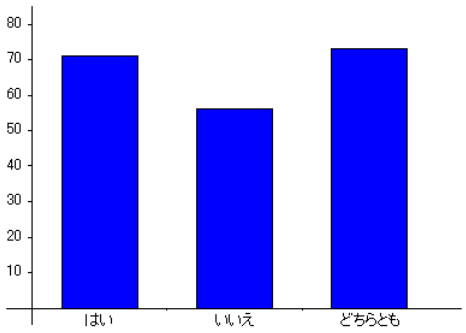
4）地域と意見１に関する２次元分割表を描け。

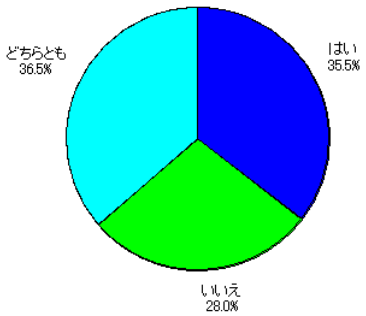
	賛成	反対	合計
市街			
郊外			
合計			

5）地域と意見２に関する２次元分割表を描け。

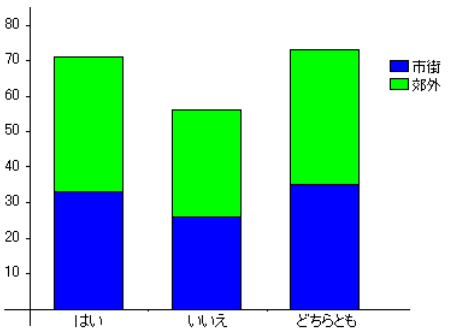
	はい	いいえ	どちらとも	合計
市街				
郊外				
合計				

6）意見２に関する棒グラフと円グラフを描け。





7）地域と意見２に関する積み重ね棒グラフを描け。



演習１（質的データの集計）

ある大学の学費と授業について以下のアンケート調査を行った。

１．あなたは男性ですか、女性ですか。

１）男性　　２）女性

２．あなたの所属する学科はどれですか。

１）経済学科　　２）経営学科

３．今の学費をどう思いますか。

１）安い　　２）妥当である　　３）高い

４．今の授業に満足していますか。

１）満足している　　２）どちらともいえない　　３）満足していない

演習 1.txt のデータを Analysis に読み込んで集計し、問題に答えよ。

問題

１）男女の数　　男性［　　］人　女性［　　］人

２）所属学科の人数　　経済学科［　　］人　経営学科［　　］人

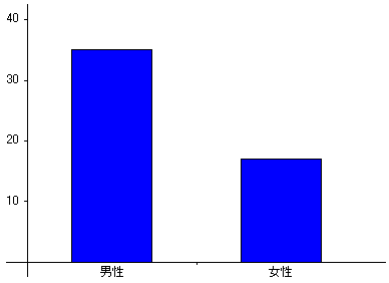
３）学費をどう思うか　　安い［　　］人　妥当［　　］人　高い［　　］人

学費をどう思っている人が最も多いか　［安い・妥当・高い］

４）授業に満足か　満足［　　］人 どちらとも［　　］人 不満足［　　］人

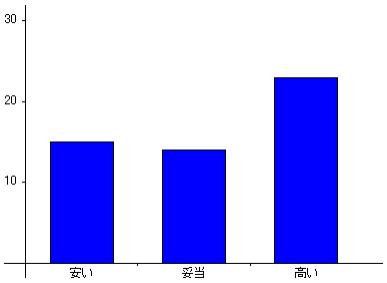
どの人が最も多いか　［満足・どちらとも・不満足］

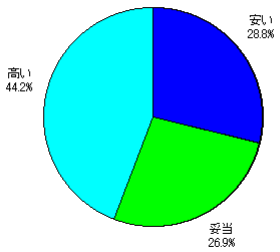
５）男女の数を以下のような棒グラフで表す。



６）この調査で男性の割合は何%か。［　　］%

７）学費をどう思うかを以下のように棒グラフと円グラフで表わす。





８）男女別の学費に対する意見を２次元分割表で表わす。

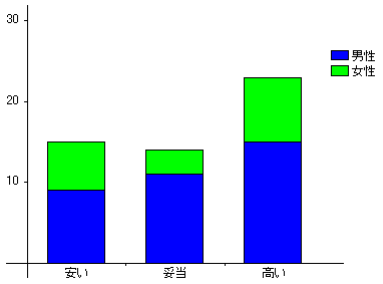
	安い	妥当	高い
男性			
女性			

９）学科別の授業に対する満足度を２次元分割表で表わす。

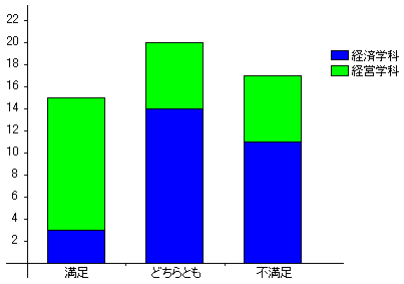
	満足	どちらとも	不満足
経済学科			
経営学科			

１０）学科別に授業に対する満足の割合が高いのはどちらか。［経済学科・経営学科］

１１）男女別の学費に対する意見の２次元分割表を積重ね棒グラフで表わし、図のような凡例を付ける。



１２）学科別の授業に対する満足度の２次元分割表を積重ね棒グラフで表わす。



1.2　量的データの集計

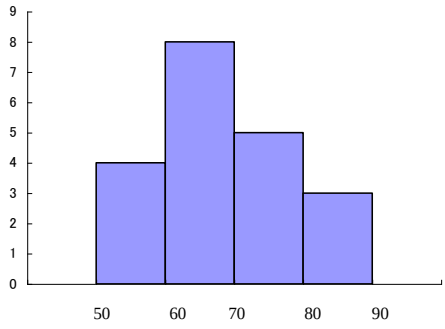
1.2.1　単純集計

度数分布表

階級	度数	相対度数(%)	累積度数	累積相対度数(%)
50 <= x < 60	4	20	4	20
60 <= x < 70	8	40	12	60
70 <= x < 80	5	25	17	85
80 <= x < 90	3	15	20	100
計	20	100		

注）各階級の幅を階級幅、各階級の中央の値を階級値という。

ヒストグラム



基本統計量（要約統計量）　【データ　3, 3, 4, 2, 8 】

分布の中心を表わす統計量（代表値）

注）基本統計量を代表値の意味で使う場合も多い

平均値（average, mean）

$$\text{平均値} = \frac{1}{5}(3+3+4+2+8) = 4$$

中央値（中間値，メジアン median）

データを小さい方から順番に並べて中間の値

2, 3, 3, 4, 8　　→　3

2, 3, 3, 4, 6, 8　　→　(3+4)/2=3.5

最頻値（モード mode）

度数分布表やヒストグラムでまとめられている場合は、最大度数の階級値

分布の拡がりを表わす統計量（散布度）

レンジ（range）

R=最大値－最小値=6

分散（variance）

$$s^2 = \frac{1}{5}\left[(3-4)^2 + (3-4)^2 + (4-4)^2 + (2-4)^2 + (8-4)^2\right] = 4.4$$

標準偏差（standard deviation）

$$s = \sqrt{\text{分散}} = 2.098$$

不偏分散

$$u^2 = \frac{1}{5-1}\left[(3-4)^2 + (3-4)^2 + (4-4)^2 + (2-4)^2 + (8-4)^2\right] = 5.5$$

標準偏差（standard deviation）

$$u = \sqrt{\text{不偏分散}} = 2.345$$

例

以下のデータ（Samples¥テキスト 1.txt）を用いて次の問いに答えよ。

学校	身長(cm)	体重(kg)	学校	身長(cm)	体重(kg)
2	169	71	1	170	62
1	175	68	1	182	75
2	170	67	2	177	70
1	179	72	1	175	70
1	176	69	1	172	62
2	174	81	2	166	58
2	173	75	2	168	60
1	181	65	2	173	58
1	179	74	2	169	59
2	178	71	2	170	73

1）身長についての基本統計量を求めよ。

2）体重についての基本統計量を求めよ。

3）身長について 5cm 毎の度数分布表を描け。

4）身長について 5cm 毎のヒストグラムを描け。

5）体重について 10kg 毎のヒストグラムを描け。

6）学校別に身長についての基本統計量を求めよ。

7）学校 1 について、身長のヒストグラムを描け。

演習２（量的データの集計）

ある中学のクラスについて英語・数学・国語の試験結果を調べた。

１．性別

１）男子　　２）女子

２．英語点数

３．数学点数

４．国語点数

演習 2.txt のデータを Analysis に読み込んで集計し、以下の問題に答えよ。

問題

１）英語、数学、国語の平均値と標準偏差を求め、下の質問に答えよ。

	英語	数学	国語
平均値			
標準偏差			

２）どの科目が最も点数が高いか

［平均・標準偏差］でみると［英語・数学・国語］の点数が高い。

どの科目が最も点数のばらつきが大きいか

［平均・標準偏差］でみると［英語・数学・国語］のばらつきが大きい。

３）男子の英語、数学、国語の平均値と標準偏差を求めよ。

	英語	数学	国語
平均値			
標準偏差			

４）女子の英語、数学、国語の平均値と標準偏差を求めよ。

	英語	数学	国語
平均値			
標準偏差			

５）男子と女子、どちらが英語の成績が良いか。

［平均・標準偏差］でみると［男子・女子］の成績が良い。

男子と女子、どちらが英語の点数のばらつきが大きいか。

［平均・標準偏差］でみると［男子・女子］のばらつきが大きい。

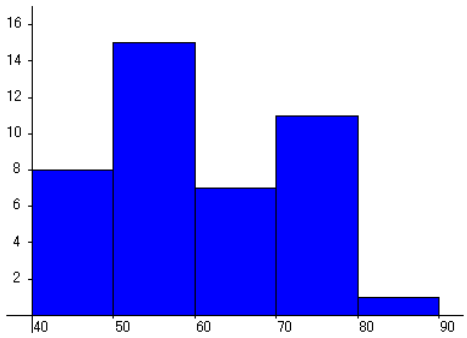
6）英語の点数について度数分布表を描き、下の質問に答えよ。

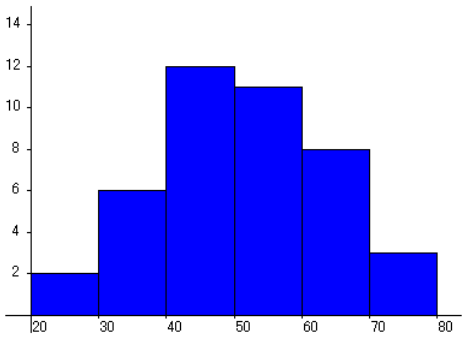
英語	度数	相対度数 %	累積度数	累積相対度数 %
40<=x<50				
50<=x<60				
60<=x<70				
70<=x<80				
80<=x<90				

度数分布表の階級幅はいくらか　　[　　] 点

この度数分布表で見た最頻値はいくらか　　[　　] 点

7）英語と数学の点数についてヒストグラムを描く。



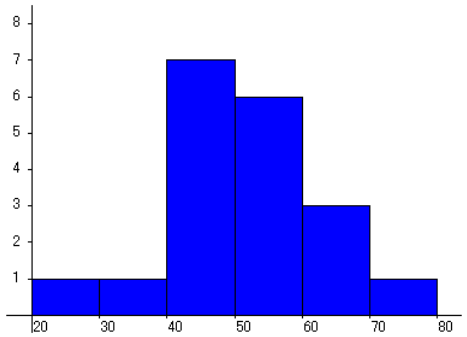


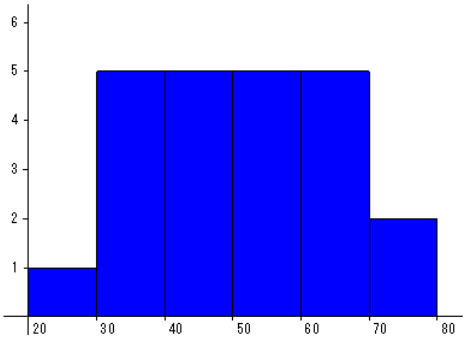
8）上の数学について、

ヒストグラムの階級幅はいくらか。　　[　　] 点

ヒストグラムで見た最頻値はいくらか　[　　] 点

9）男女の数学の点数のヒストグラムを描く。

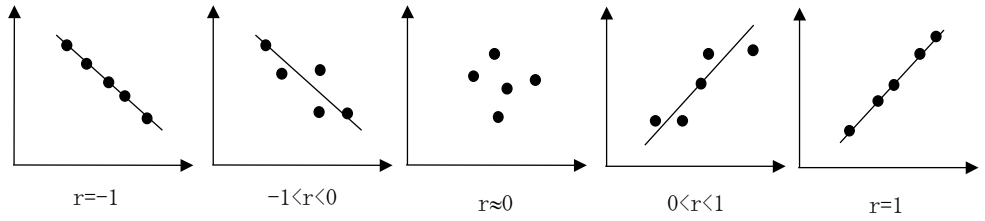




注）女子の最頻値は１つに定まらない。

1.2.2　クロス集計

散布図（分布図，相関図），回帰直線，相関係数



例

以下のデータ（Samples¥テキスト 1.txt）を用いて次の問いに答えよ。

学校	身長(cm)	体重(kg)	学校	身長(cm)	体重(kg)
2	169	71	1	170	62
1	175	68	1	182	75
2	170	67	2	177	70
1	179	72	1	175	70
1	176	69	1	172	62
2	174	81	2	166	58
2	173	75	2	168	60
1	181	65	2	173	58
1	179	74	2	169	59
2	178	71	2	170	73

1）身長と体重に関する散布図を描け（体重を縦軸)。

2）身長と体重の相関係数を求めよ。

3）身長で体重を予測する回帰式を求めよ。

問題

Samples¥テキスト 9.txt を用いて以下の問いに答え、結果は文書にまとめよ。但し、地域について 1：市街、2：郊外とする。

1）年収に関する基本統計量を求めよ。

データ数	最小値	最大値	平均値	中央値	不偏分散	標準偏差

データの拡がりをみるには上のどの指標が適切か［　　］

2）地域別の年収に関する基本統計量を求めよ。

	データ数	最小値	最大値	平均値	中央値	不偏分散	標準偏差
市街							
郊外							

市街と郊外ではどちらの年収が高いか　　［市街・郊外］

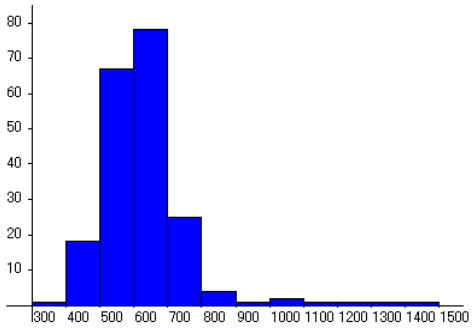
市街と郊外ではどちらの年収の拡がりが大きいか　　［市街・郊外］

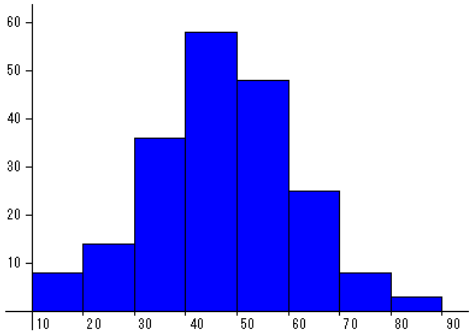
3）年収に関するヒストグラムを描け。（下図左）

このヒストグラムの階級幅はいくらか［　　］

このヒストグラムの最頻値はいくらか［　　］

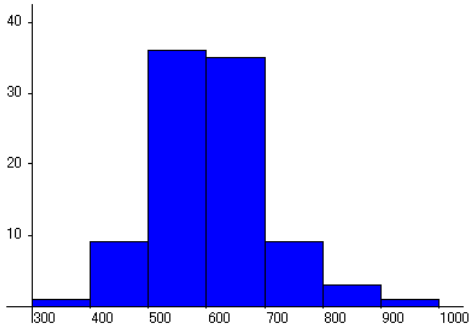
4）支出に関するヒストグラムを描け。（下図右）

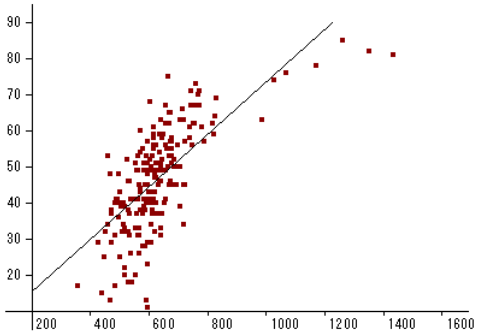




5）地域:1 の年収に関するヒストグラムを描け。（下図左）

6）年収と支出に関する散布図を描け（支出を縦軸，下図右）。





7）年収と支出に関する相関係数を求めよ。　　相関係数［　　］

8）支出を目的変数に年収を説明変数としたときの回帰式を求めよ。

支出＝［　　］×年収＋［　　］

演習３（総合演習・集計）

ある商店街で以下のアンケート（一部）を行った。

１．あなたは男性ですか女性ですか。

１）男性　　２）女性

２．よろしければあなたの年齢を教えて下さい。

［　　］歳

３．週何回位この商店街を利用（買い物他）されますか。

［　　］回

上の問題で０～１回と答えられた方にお聞きします。

４．なぜあまり買い物をされないのですか。

１）値段が高い　２）品揃えが不足　３）近くに良い店がある

４）その他［　　］

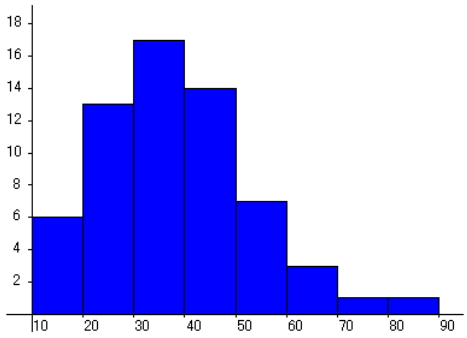
これを元に新たに２段階の年齢階級を加えて（40 歳未満と 40 歳以上）、演習 3.txt のようなデータを得た。問題に答え結果をレポートにまとめよ。

問題

１）男女別の回答数を求める。男性［　　］人 女性［　　］人 合計［　　］人

２）年齢の平均を求める。　［　　］歳

３）年齢についてのヒストグラムを描く。



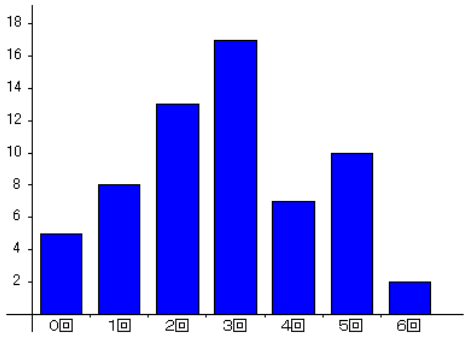
このヒストグラムの階級幅はいくらか　［　　］才

このヒストグラムでみた最頻値はいくらか　［　　］才

４）年齢階級の人数を求める。40 歳未満［　　］人　40 歳以上［　　］人

5）商店街利用回数の平均を求める。　［　　］回

6）商店街利用回数の棒グラフを描く。



7）男女別の利用回数の２次元分割表を描く。

	０回	１回	２回	３回	４回	５回	６回	合計
男性								
女性								
合計								

8）男女別の利用回数の平均を求める。男性［　　］回　女性［　　］回

男性と女性、利用回数の多いのはどちらか　［男性・女性］

9）年齢階級別の利用回数の平均を求める。

40 歳未満［　　］回　40 歳以上［　　］回

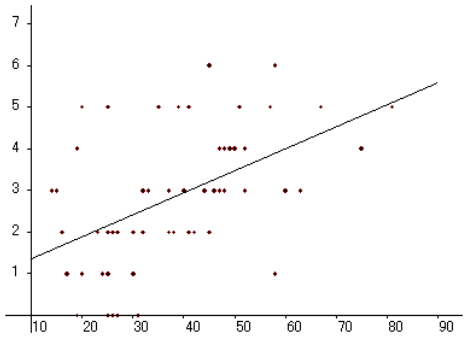
年齢階級で比較すると、利用回数の多いのはどちらか　［40 歳未満・40 歳以上］

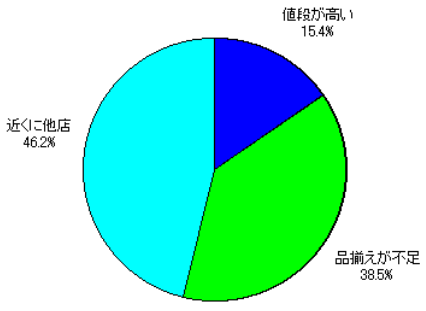
10）年齢と利用回数の相関係数を求める。ｒ＝［　　］

これは非常に強い相関か。［非常に強い・あまり強いとはいえない］

11）年齢と利用回数の散布図を描く。（下図左）

12）利用しない理由の円グラフを描く。（下図右）





1.3 欠損値の除去

例

番号	学校	国語	数学
1	1	76	82
2	2		63
3	1	62	58
4		73	74
5	2	81	
6	2	73	65
7	1		46

各集計で利用する人は？（よく使われる欠損値の除去方法）

国語の平均　1,3,4,5,6　①データ単位の除去

学校ごとの国語の平均　1,3,5,6　②（分類変数を除いて）データ単位の除去

国語と数学の相関係数　1,3,4,6　③（選択）レコード単位の除去

（分類変数以外で）１変数だけを除去する場合は、データ単位の除去

（選択した）複数変数を連動して除去する場合は、レコード単位の除去

問題

欠損値を含む Samples¥テキスト 9b.txt を用いて、以下の問いに答え、よく使われる欠損値の除去方法について、上の①、②、③のどれに一番近いか右上の［　］に答えよ。

１）意見１に関する１次元分割表を描け。　［　　］

意見１:1	意見１:2	合計

２）意見１と意見２に関する２次元分割表を描け。　［　　］

	意見２:1	意見２:2	意見２:3	合計
意見１:1				
意見１:2				
合計				

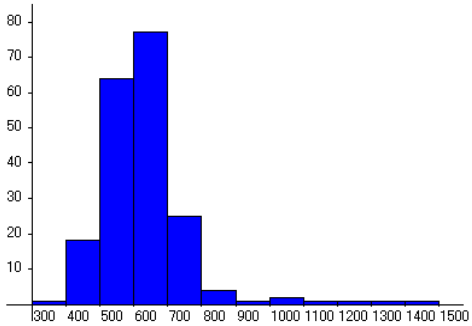
３）年収と支出に関する以下の基本統計量を求めよ。　［　　］

	最小値	最大値	平均値	中央値	標準偏差
年収					
支出					

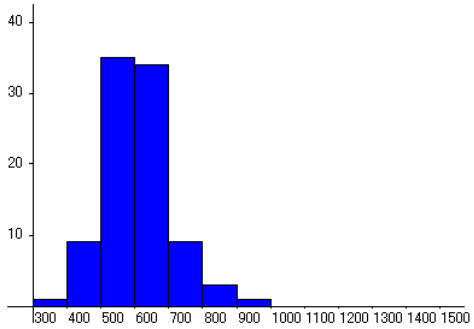
4）地域別の年収に関する基本統計量を求めよ。　［　　］

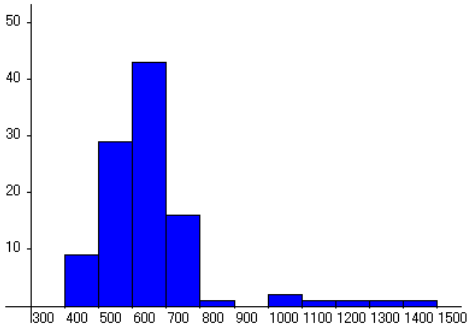
	最小値	最大値	平均値	中央値	標準偏差
地域:1					
地域:2					

5）年収に関するヒストグラムを描け。　［　　］

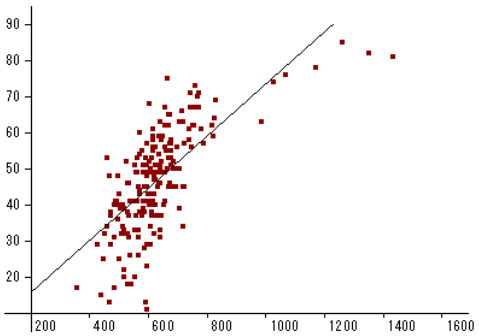


6）地域 1, 2 の年収に関するヒストグラム　［　　］





7）年収と支出に関する散布図を描け。　［　　］



8）年収と支出に関する相関係数を求めよ。　［　　］

相関係数＝［　　］

9）支出を年収で予測する回帰式を求めよ。　［　　］

支出＝［　　］×年収＋［　　］

演習４（総合演習・集計）

本屋と読書について以下のアンケート調査を行い、結果を演習 4.txt に保存した。

１．あなたは男性ですか女性ですか。

１）男性　　２）女性

２．もしよろしければあなたの年齢を教えて下さい。

［　　］歳

３．あなたは１週間にどれくらい本屋へ行きますか。

１）ほとんど行かない　２）１～２回　３）３～４回　４）５回以上

４．あなたは１週間にどれくらい本代に使いますか。

［　　］円

５．あなたがよく読む本はどれですか。いくつでも○を付けて下さい。

１）雑誌　２）コミック　３）小説　４）ノンフィクション　５）その他

６．あなたの１週間の教養・娯楽費（本代を含む）はどれくらいですか。

［　　］円

このデータを元に以下の問題に答えよ。<u>但し、判断をしなくてもよいところもある。</u>

問題

１）アンケートに答えた人数は何人ですか。［　　］人

２）そのうち男性と女性は何人ずつですか。男性［　　］人　女性［　　］人

３）年齢の平均は何歳ですか。［　　］歳

４）年齢は何歳から何歳までですか。［　　］歳～［　　］歳

５）年齢を１０歳ごとに区切って４０歳代（４０～４９）は何人ですか。［　　］人

６）年齢１０歳ごとの人数の分布を図で表示します。

［　　］を見る。＜実際に図も作る＞

７）男性の平均年齢と女性の平均年齢はそれぞれ何歳ですか。

男性［　　］歳　女性［　　］歳

８）男性と女性でどちらが年齢の拡がりが大きいですか。

［平均・標準偏差］を調べると男性［　　］歳、女性［　　］歳

なので［男性・女性］が大きい。

９）本屋に週１～２回行く人は男性と女性どちらが多いですか。（ヒント：分割表）

男性［　　］人　女性［　　］人なので、［男性・女性］が多い。

１０）本代に使う金額は平均いくらですか。［　　］円

１１）本代は最高いくら使っていますか。［　　］円

１２）男性と女性ではどちらが本代にお金を使いますか。（ヒント：分布の中心）

［平均・標準偏差］を調べると男性［　　］円、女性［　　］円

なので、［男性・女性］がたくさん使う。

１３）本屋に行く回数と本代に使うお金は関係がありますか。

平均で調べると以下のように増えているので、関係が［ある・ない］と思う。

殆ど行かない	１～２回	３～４回	５回以上

１４）本屋に行く回数で３回以上の人は過半数ですか。

（ヒント：アンケート質問の３の回答３と４）　全体［　　］人中、

３回以上の人は［　　］人なので、［過半数である・過半数でない］

１５）年齢と本代に使うお金は関係がありますか。（ヒント：量と量の関係）

［平均・標準偏差・相関係数］を調べるとその値は［　　］なので、

［あまり関係がない・強い関係がある］

１６）コミックはどの位の割合で読まれていますか。（ヒント：アンケート質問５の各回答は、読まない人０、読む人１です。自分で計算して下さい。）［　　］％

１７）小説を読む人と読まない人とで、本代に使うお金に差がありますか。

［　　］を調べると読まない人［　　］円、読む人［　　］円

１８）教養・娯楽費は平均いくら位ですか。［　　］円

１９）教養・娯楽費の最高と最低の幅はいくらですか。［　　］円

２０）教養・娯楽費と本代との関係はありますか。（ヒント：量と量の関係）

［　　］を調べるとその値は［　　］なので、

関係が［ある・ない］と思う。

２１）男性と女性とで教養・娯楽費の額に差がありますか。

［　　］を調べると男性［　　］円、女性［　　］円

２２）コミックを読む人は若い人が多いですか。

年齢の［　　］を調べると読まない人［　　］歳、読む人［　　］歳

２３）コミックを読む人と小説を読む人の関係を調べて下さい。（ヒント：分割表）

	小説を読まない	小説を読む
コミックを読まない		
コミックを読む		

コミックを読む人は小説を［読む・読まない］傾向がある。

２．確率分布と検定

2.1　確率密度関数

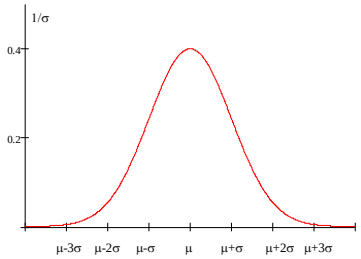
データ数を十分多く取ったヒストグラムの上端をつなぎ、全体の面積が１になるように、目盛りを付けたものを確率密度関数と呼ぶ。この確率密度関数の形で分布の名前が付けられている。

$y = f(x)$：確率密度関数

2.2　正規分布 (normal distribution) と標準正規分布

正規分布（X は平均 μ 分散 σ^2 の正規分布：$X \sim N(\mu, \sigma^2)$）

正規分布とは偶発的なデータのゆらぎによる分布（量的データの基本となる分布）



$P(\mu-\sigma \le X \le \mu+\sigma) = 0.683$　　両側　約 32%

$P(\mu-2\sigma \le X \le \mu+2\sigma) = 0.954$　両側　約 5%

$P(\mu-3\sigma \le X \le \mu+3\sigma) = 0.997$　両側　約 0.3%

概数は覚えること

よく使う正規分布の性質

１）$X \sim N(\mu, \sigma^2)$ のとき　$\boxed{X' = \dfrac{X-\mu}{\sigma} \sim N(0,1)}$

X' の分布を標準正規分布といい、統計処理では非常によく利用される。

標準正規分布の詳しい確率の値は、例えば Excel では、以下で求められる。

昔は表を使って求めていた。

$P(X \le x) = normsdist(x)$

２）$\overline{X}$ を n 個のデータの平均とすると　$\boxed{\overline{X} \sim N(\mu, \sigma^2/n)}$

１つ１つのデータに対して、平均を取るとデータの精度が上がる。

標準偏差は $\sigma/\sqrt{n}$、例えば 100 個だと $\sigma/10$ になる。

これは X の分布によらない。中心極限定理

問題（１個のデータについて）

体重の平均 10kg、標準偏差 2kg（分散 $4kg^2$）の子供 1000 人の集団がある。データは正規分布するとして以下の問いに概数（大体の値）で答えよ。

１）12kg の子供は重い方から大体何%か　［　　］%

２）14kg の子供は重い方から大体何%か　［　　］%

３）14kg の子供は重い方から大体何番目か　［　　］番目

４）8kg の子供は重い方から大体何%か　［　　］%

５）8kg の子供は重い方から大体何番目か　［　　］番目

問題（１個のデータについて）

前の問題で、子供の体重から平均の 10kg を引き、その結果を標準偏差の 2kg で割るとする。以下の問いに答えよ。

１）10kg の子供の値はいくらになるか　［　　］

２）12kg の子供の値はいくらになるか　［　　］

３）7kg の子供の値はいくらになるか　［　　］

４）この計算結果は平均［　　］、分散［　　］の正規分布になる。

５）３）について、7kg 以下となる正確な確率を求めよ。
Excel の関数 normsdist(x)を用いると　［　　］

６）２）について、12kg 以上となる正確な確率を求めよ。
Excel の関数 normsdist(x)を用いると　［　　］

７）7kg 以上、12kg 以下となる正確な確率を工夫して求めよ。
Excel の関数 normsdist(x)を用いると　［　　］

問題（データの平均について）

体重の平均 10kg、標準偏差 2kg の子供の大きな集団（母集団）がある。この中から 100 人の集団（標本）をランダムに取り出し、その平均 $\overline{X}$ を取るとする。以下の問いに答えよ。

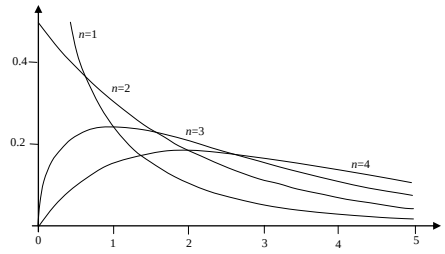
１）$\overline{X}$ の平均（標本平均の平均）はいくらか。　［　　］kg

２）$\overline{X}$ の標準偏差（標本平均の標準偏差）はいくらか　［　　］kg

３）$\overline{X}$ の値が 10.2kg の標本は重い方から大体何%か　［　　］%

2.3 標準正規分布から導かれる分布

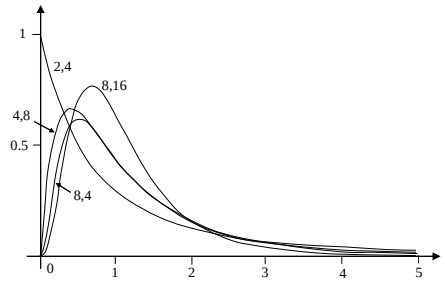
χ^2 分布



$X_i \sim N(0,1)$ で独立なとき、

$\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$ 分布（自由度 n の χ^2 分布）

F 分布

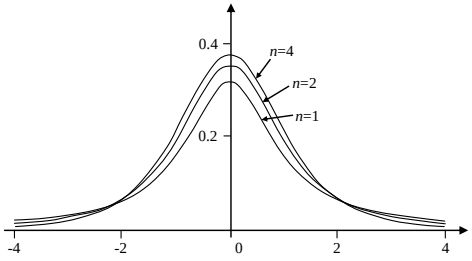


$\chi_1^2 \sim \chi_{n_1}^2$ 分布，$\chi_2^2 \sim \chi_{n_2}^2$ 分布で独立なとき、

$F = \dfrac{\chi_1^2 / n_1}{\chi_2^2 / n_2} \sim F_{n_1, n_2}$ 分布

（自由度 n_1, n_2 の F 分布）

t 分布



$X \sim N(0,1)$ 分布，$\chi^2 \sim \chi_n^2$ 分布で独立なとき、

$t = \dfrac{X}{\sqrt{\chi^2 / n}} \sim t_n$ 分布（自由度 n の t 分布）

注）$t^2 \sim F_{1,n}$ 分布

注）$n \to \infty$ で $N(0,1)$ 分布

問題　College Analysis を使って以下の値を求めよ。

1）$N(0,1)$ 分布，x 値 1.5 のときの上側確率 $p/2$　［　　］

2）$N(0,1)$ 分布，x 値 1.5 のときの両側確率 p　［　　］

3）$N(170,64)$ 分布，x 値 180 のときの上側確率 $p/2$　［　　］

4）χ_5^2 分布，χ^2 値 10 のときの上側確率 p　［　　］

5）χ_{10}^2 分布，上側確率 0.05 のときの χ^2 値　［　　］

6）$F_{8,4}$ 分布，F 値 10 のときの上側確率 p　［　　］

7）$F_{10,5}$ 分布，上側確率 0.05 のときの F 値　［　　］

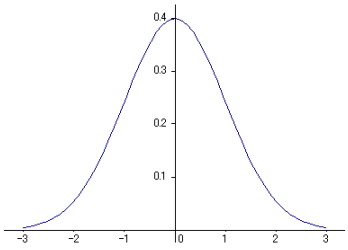
8）t_{10} 分布，t 値 2 のときの上側確率 $p/2$　［　　］

9）t_{10} 分布，t 値 2 のときの両側確率 p　［　　］

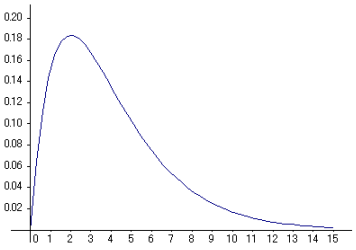
10）t_{10} 分布，両側確率 0.05 のときの t 値　［　　］

問題　College Analysis を使って以下のグラフを描け。

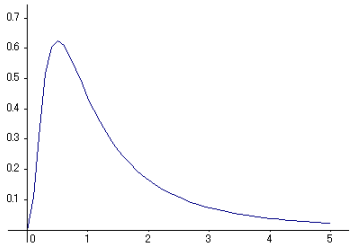
１）$N(0,1)$ 分布



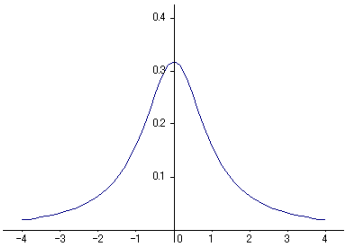
２）自由度 4 の χ^2 分布



３）自由度 8,4 の F 分布



４）自由度 1 の t 分布



補足　2 項分布について

確率pで出現する事象が、n回の試行中、r回出現する確率

$$P = {}_nC_r p^r (1-p)^{n-r}$$

不良品の出現回数などにも使われる。通常はr回以上として確率を計算する。

じゃんけんで 4 回中 3 回勝つ確率

$$P = {}_4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = 4 \times \frac{1}{16} = \frac{1}{4}$$

サイコロを 5 回振って 1 の目が 2 回出る確率

$$P = {}_5C_2 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^3 = 0.161$$

問題

１）じゃんけんで 4 回中 1 回勝つ確率　［　　］

２）じゃんけんで 6 回中 3 回勝つ確率　［　　］

３）さいころを 4 回振って 2 以下の目が 2 回出る確率　［　　］

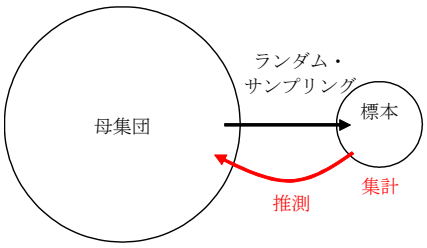
４）さいころを 4 回振ってすべて 5 以上の目となる確率　［　　］

５）平均の不良品率が 1%であるとき、100 個の製品で 3 個の不良品が発生する確率

［　　］

2.4　検定の基礎

母集団と標本

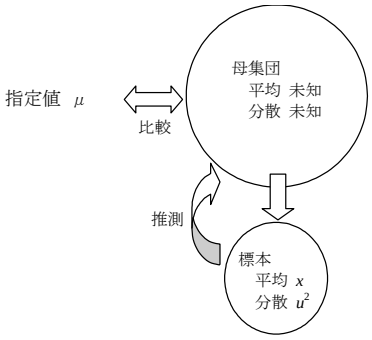


母集団：調査の対象，日本人・日本の中小企業等
（全数調査不可能な場合がある）

標本：　偏りがないように選抜（ランダムサンプリング）された実際に調査する対象

母集団の全数調査が不可能な場合、標本をとって母集団を推測する。

どんな検定があるか



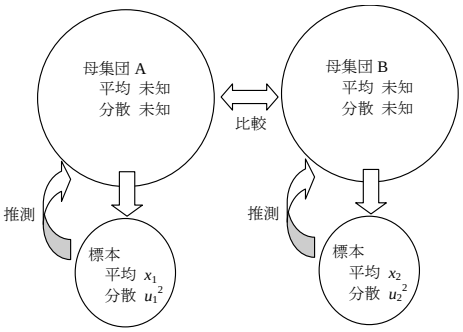
1）指定値と母集団のある指標を比較する。

量的データの比較：

標本調査世帯と全国平均との所得の比較

質的データの比較：

標本調査の結果（割合）と期待される結果（割合）との比較



2）いくつかの母集団のある指標を比較する。

量的データの比較：

２つの標本調査世帯の所得の比較
（対応がない場合）

標本店における宣伝前後の売り上げ比較
（対応がある場合）

質的データの比較：

男女間での意識調査の結果（割合）の比較（対応がない場合）

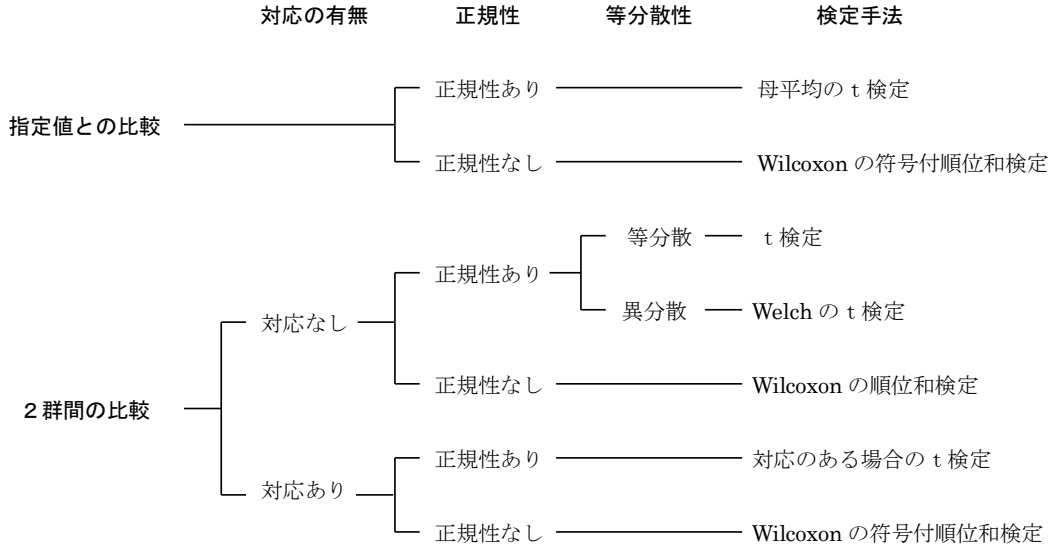
標本店における従業員教育前後の評判の変化（対応がある場合）

2.5 検定選択ツリー

質的データ

	対応の有無	検定手法
指定比率との比較		適合度検定
多群間の比較	対応なし	χ^2 検定
	対応あり	McNemar 検定

量的データ



以後、これらの検定を詳細に見て行く。

３．質的データの検定

3.1　母集団の比率と指定比率との検定

例

ある大学の学生 50 人を任意抽出し、大学改革のアンケートを行ったところ、賛成 35 反対 15 であった。学生の過半数が賛成している（賛成の比率が 1/2 と異なる）といえるか、有意水準 5%で判定せよ。

理論　適合度検定

出現比率が指定比率と比べて差がないとすると

$$\chi^2 = \frac{(n_1 - m_1)^2}{m_1} + \frac{(n_2 - m_2)^2}{m_2} + \cdots + \frac{(n_k - m_k)^2}{m_k} \sim \chi^2_{k-1} \text{分布}$$

$$\chi^2 = \frac{(|n_1 - m_1| - 1/2)^2}{m_1} + \frac{(|n_2 - m_2| - 1/2)^2}{m_2} + \cdots + \frac{(|n_k - m_k| - 1/2)^2}{m_k} \underset{n\to\infty}{\sim} \chi^2_{k-1} \text{分布}$$

（Yates の連続補正）

解答

$p_1 = p_2 = 1/2$

$\chi^2 = 7.22$

$p = 0.00721$

判定　賛成は過半数といえる。

問題１

ある工場で１年間におきた事故の件数を曜日毎に調べたところ、以下の表が得られた。事故は曜日による差があるといえるか？有意水準 5%で判定せよ。

曜日	月	火	水	木	金	計
事故件数	23	14	16	11	16	80

解答

P = ［　　］

判定　曜日による差があると［いえる・いえない］

問題２

前の問題で、月曜日は特に事故が起こっているといえるか。月曜日とその他の曜日に分けて有意水準 5%で判定せよ。

解答

P = ［　　］

判定　月曜日に事故が多く起こっていると［いえる・いえない］

3.2　対応のない２群間の比率の検定

例

ある問題についての調査で、男女別に賛成か反対かを集計したところ以下の結果を得た。賛成（または反対）の比率に男女差はあるといえるか。有意水準 5%で判定せよ。

	賛成	反対	計
男性	18	10	28
女性	12	14	26
計	30	24	54

解答

$\chi^2 = 1.1358,\quad p = 0.286542$

$p > 0.05$ より、男女差があるとはいえない。

理論（２×２分割表）

	事象１	事象２	計
要因１	a	b	$a+b$
要因２	c	d	$c+d$
計	$a+c$	$b+d$	$a+b+c+d=n$

要因間で、事象の出現比率に差がないとすると

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_1^2 \text{ 分布}$$

$$\chi^2 = \frac{n(|ad-bc|-n/2)^2}{(a+b)(c+d)(a+c)(b+d)} \sim \chi_1^2 \text{ 分布} \qquad \text{（Yates の連続補正）}$$

理論（ｍ×ｎ分割表）

	事象１	事象２	…	事象 s	計
要因１	x_{11}	x_{12}	…	x_{1s}	$x_{1.}$
要因２	x_{21}	x_{22}	…	x_{2s}	$x_{2.}$
:	:	:		:	:
要因 r	x_{r1}	x_{r2}	…	x_{rs}	$x_{r.}$
計	$x_{.1}$	$x_{.2}$	…	$x_{.s}$	n

要因間で、事象の出現比率に差がないとすると

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(x_{ij} - x_{i\cdot}x_{\cdot j}/n\right)^2}{x_{i\cdot}x_{\cdot j}/n} \sim \chi^2_{(r-1)(s-1)} \text{ 分布} \qquad 2\times 2 \text{ 表の統計量の一般形}$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\left|x_{ij} - x_{i\cdot}x_{\cdot j}/n\right| - 1/2\right)^2}{x_{i\cdot}x_{\cdot j}/n} \sim \chi^2_{(r-1)(s-1)} \text{ 分布} \qquad \text{（Yates の連続補正）}$$

問題３

ある案についてのアンケートで以下の結果を得た。男女間の回答（賛成の比率）に差があるといえるか。有意水準 5%で判定せよ。

	賛成	反対
男性	128	86
女性	107	95

確率［　　］　判定　男女間に差があると［いえる・いえない］

問題４

女性を対象とした調査で、ある化粧品の所有の有無を職業別に分類してみると、以下の結果が得られた。職業間で商品所有の割合に差があるといえるか。有意水準 5%で判定せよ。

	所有あり	所有なし	計
主婦	90	199	289
事務	32	47	79
販売・生産	53	71	124
計	175	317	492

確率［　　］　判定　職業間に差があると［いえる・いえない］

問題５

Samples¥テキスト 9.txt において、以下の問いに答えよ。

１）意見１について１次元分割表を描け。（1：はい，2：いいえ）

はい	いいえ	合計

２）意見１において、いいえは過半数といえるか。有意水準 5%で判定せよ。

P =［　　］

判定　過半数と［いえる・いえない］

3）意見２について１次元分割表を描け。（1：案１，2:案２，3:案３）

案１	案２	案３	合計

4）意見２において、回答間に差があるといえるか。有意水準 5%で判定せよ。

P＝［　　］

判定　回答間に差があると［いえる・いえない］

5）意見１の回答に地域による差があるか。有意水準 5%で判定せよ。

確率［　　］　判定　地域による差があると［いえる・いえない］。

6）上の問題で有意水準を 1%にすると結果はどう変わるか。

判定　地域による差があると［いえる・いえない］。

7）意見２の回答に地域による差があるか。有意水準 5%で判定せよ。

確率［　　］　判定　地域による差があると［いえる・いえない］。

8）意見２の回答に意見１による差があるか。有意水準 5%で判定せよ。

確率［　　］　判定　意見１による差があると［いえる・いえない］。

問題６

1）平均的な故障発生率が 1%のとき、100 個の製品で 4 個以上の故障発生が起こることは異常なことであろうか。2 項分布の正確な確率検定を用いて、有意水準 5%で判定せよ。

検定確率　［　　］　異常なことと［いえる・いえない］

2）昨年の納品 155 個のうち、クレームのあったものが 2 個、今年の納品 211 個のうち、クレームのあったものが 7 個であった。昨年と今年のクレームの比率に差があると言えるか。Fisher の正確確率検定を用いて、有意水準 5%で判定せよ。

検定確率　［　　］　差があると［いえる・いえない］

演習５（適合度検定・χ^2検定）

環境の変化と校舎の老朽化により、現在の小学校を少し離れた場所に移設しようとする案が出され、地域住民に以下のアンケートが実施された。

１．あなたの性別は。

１）男性　　２）女性

２．あなたの年齢はどれですか。

１）４０歳未満　２）４０歳～５９歳　　３）６０歳以上

３．あなたには小学生以下のお子さんかお孫さんがいますか。

１）いる　　２）いない

４．学校の移設に賛成ですか反対ですか。

１）賛成　　２）反対

５．学校移設に賛成の方、その理由を教えて下さい。(いくつでも選んで下さい。)

１）環境が良くなる　２）設備が新しくなる　３）その他

６．学校移設に反対の方、その理由を教えて下さい。(いくつでも選んで下さい。)

１）通学が不便になる　２）現校舎に魅力がある　３）その他

演習 5.txt を用いて問題に答えよ。

問題

１）男性と女性の数を求める。男性［　　］人　女性［　　］人

２）年齢階級別の人数を求める。

４０歳未満［　　］人　４０～５９歳［　　］人　６０歳以上［　　］人

３）小学生以下の子どもがいるかどうかの人数を求める。

いる［　　］人　いない［　　］人

４）賛成と反対の人数を求める。賛成［　　］人　反対［　　］人

５）男女別の賛成と反対の２次元分割表を作る。

	賛成	反対
男性		
女性		

６）子供がいるかどうか別の賛成と反対の２次元分割表を作る。

	賛成	反対
子供がいる		
子供がいない		

7）年齢層別の賛成と反対の２次元分割表を作る。

	賛成	反対
40 歳未満		
40～59 歳		
60 歳以上		

8）賛成は過半数といえるか、有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

結果　過半数と［いえる・いえない］

9）賛成の理由で「環境が良くなる」と答える人は４割より多いといえるか、有意水準 5%で判定せよ。

ヒント：答えない（0），答えた（1）順に指定比率を入力すること（0.6,0.4）

検定名［　　］　確率［　　］

結果　４割より多いと［いえる・いえない］

10）上の検定で有意水準を 1%にすると結果はどうなるか。

結果　４割より多いと［いえる・いえない］

11）反対の理由で「通学が不便になる」と答える人は３割より多いといえるか、有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

結果　３割より多いと［いえる・いえない］

12）男性と女性で賛成・反対に差があるか、有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

結果　男女間に差があると［いえる・いえない］

13）小学生以下の子供がいるかどうかで賛成･反対に差があるか、有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

結果　子供がいるかどうかで差があると［いえる・いえない］

14）年齢層によって賛成・反対に差があるか、有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

結果　年齢層間で差があると［いえる・いえない］

15）賛成・反対に影響を与えている要素はどれか。

［性別・子供または孫の有無・年齢］

3.3　対応のある母集団間の比率の検定（McNemar 検定）

例

あるキャンペーン実施の前後で、各支店の印象について客からアンケートをとり、支店毎に好印象かどうかで分類したところ、以下の結果を得た。キャンペーンは効果があったと言えるか。有意水準 5%で判定せよ。

前＼後	好印象	悪印象
好印象	40	11
悪印象	24	10

理論（McNemar 検定）

データ＼対照データ	結果１	結果２
結果１	a	b
結果２	c	d

２つのデータによる差がないとすると

$$\chi^2 = \frac{(b-c)^2}{b+c} \sim \chi_1^2 \text{分布}$$

$$\chi^2 = \frac{(|b-c|-1)^2}{b+c} \sim \chi_1^2 \text{分布} \qquad \text{（Yates の連続補正）}$$

注）通常の分割表のまとめ方だと以下のようになる。

	結果１	結果２
データ	$a+b$	$c+d$
対照データ	$a+c$	$b+d$

解答

$\chi^2 = 4.1143,\ \ p = 0.042522$

$p < 0.05$ より、キャンペーンによる差があるといえる。

問題

ある２社は同種の製品を作っているが、この度後継の新製品が発売された。新製品の発売前後で各量販店の売上を比較したところ、以下の結果を得た。以下の問いに答えよ。新製品は売上に影響を与えたと言えるか。有意水準 5%で判定せよ。

前	1	2	2	2	1	2	1	2	1	2	1	1	2	2
後	2	1	1	2	1	1	2	1	1	2	2	2	2	1
	1	2	1	1	1	1	1	2	1	1	2	1	1	1
	2	2	1	2	2	1	2	1	1	1	1	2	1	1

1: A 社が多い　2: B 社が多い

1）このデータから２次元分割表を作れ。

	後：A 社が多い	後：B 社が多い
前：A 社が多い		
前：B 社が多い		

2）新製品は売り上げに影響を与えたと言えるか、有意水準 5%で判定せよ。
検定名［　　］　確率［　　］
売り上げに影響を与えたと［いえる・いえない］。

3）この検定は対応がない場合としても行うこともできる。その際データはどのような形であればよいと思うか。データシートの新しいページで、以下のヒントを参考に考えよ。
ヒント
分類を新製品発売前後（前:1，後:2）と A, B 社のどちらが多いか（A 社:1, B 社:2）に変更する。そうするとデータのレコード数（行数）は［　　］となり、現在の形式の行数の［　　］倍となる。

4）新しいデータを用いて２次元分割表を作れ。

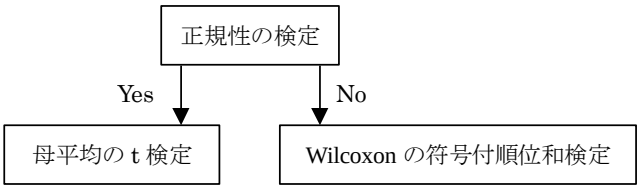
	A 社が多い	B 社が多い
［　　］		
［　　］		

5）新しいデータを用いて、新製品は売り上げに影響を与えたと言えるか有意水準 5%で判定せよ
検定名［　　］　確率［　　］
売り上げに影響を与えたと［いえる・いえない］

注）質的データの検定で正しい結果を得るためには、分割表の各セルに少なくとも 10 程度以上の値が必要である。

４．母集団と指定値との量的データの検定

4.1　検定手順



4.2　正規性の検定

視覚的方法

データ数が多い場合　　ヒストグラムによるグラフ化

データ数が少ない場合　　正規確率紙（MS-Excel でも可能）

数値的方法

データ数が多い場合

コルモゴロフースミルノフ（Kolmogorov-Smirnov 略して K-S）検定

データ数が少ない場合［今後主にこれを使用する］

シャピローウィルク（Shapiro-Wilk 略して S-W）検定　等

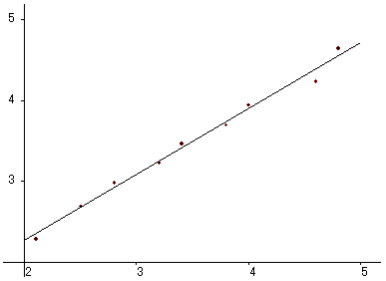
問題１

以下のデータの正規性を調べよ。

2.5，2.1，3.4，2.8，4.6，3.2，3.8，4.8，4.0

解答

データの数が少ないので、ヒストグラムは使えない。正規確率紙の方法と S-W 検定で調べる。



S-W 検定　　確率［　　］

判定　正規分布と［みなす・いえない・判定困難］

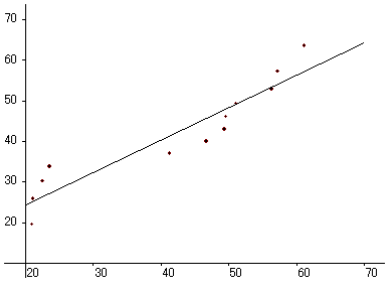
問題２

以下のデータの正規性を調べよ。

20.9，61.1，57.2，51.0，46.6，41.2，21.0，56.3，49.5，49.3，22.4，23.5

解答

正規確率紙の方法と S-W 検定で調べる。



S-W 検定　　確率［　　］

判定　正規分布と［みなす・いえない・判定困難］

4.3　母集団の平均値と指定値との比較（正規性あり）

例

ある地域のある規模の会社 11 社について１人当り売上高は以下の通りである。この地域の会社の１人当り売上高は同じ規模の会社の１人当り平均売上高 2260（万円）に比べて差があるといえるか？検定を選んで有意水準 5%で判定せよ。

2060, 2350, 1550, 1720, 1800, 1990, 1510, 1720, 2910, 1820, 2600

理論　母平均の t 検定

指定値と比べて平均に差がないとして、

$$t=\frac{\sqrt{n}(\bar{x}-\mu)}{u}\sim t_{n-1}\text{分布}$$

解答

$t=1.91469$

$p=0.08455>0.05$ より、１人当り売上高に差があるといえない。

4.4　母集団の中央値と指定値との比較（正規性なし）

例

ある地域のある規模の会社の１人当り売上高（万円）は以下の通りである。これらの会社は同じ規模の会社の中央値 2260（万円）に比べて売上高に差があるといえるか。検定を選んで有意水準 5%で判定せよ。

2060, 2064, 2072, 2005, 2602, 1987, 1824, 1720, 2035, 1890, 2025,

概要　Wilcoxon（ウィルコクソン）の符号付き順位和検定

データの順位により母集団の中央値が指定値と異なっているかどうか検定する。

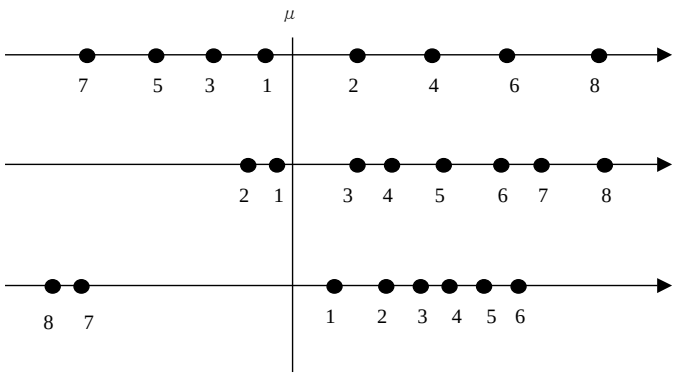


図　検定概念図

左右の順位和を求め、その小さい方を R とする。

標本数が多いとき

$$z=\frac{|R-n(n+1)/4|-1/2}{\sqrt{n(n+1)(2n+1)/24}}\ \sim\ N(0,1)\text{分布（正の部分）}\qquad\text{（Yates の連続補正）}$$

解答

$R = 8$, $p = 0.02938 < 0.05$ より、中央値に差があるといえる。

問題３

以下のデータの平均値（中央値）は 5.5 と比べて差があるといえるか。検定を選んで有意水準 5%で判定せよ。

8.4, 4.6, 5.2, 6.3, 7.2, 5.8, 6.0, 5.4, 4.9, 6.9

正規性の判定

ヒストグラムにはデータ不足。正規確率紙を描く。S-W 検定　確率［　　］

判定　正規分布と［みなす・いえない］。

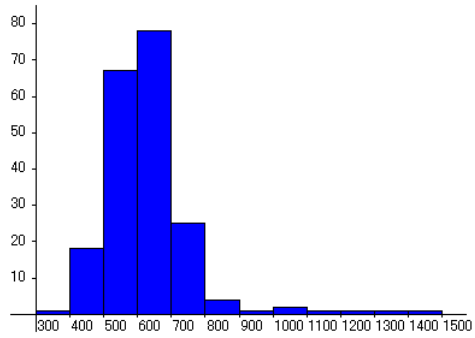
検定名［　　］　確率［　　］

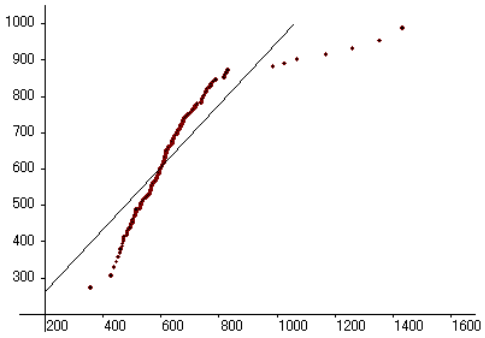
判定　5.5 と比べて差があると［いえる・いえない］。

問題４

Samples¥テキスト 9.txt のデータを用いて以下の問いに答えよ。

１）年収のデータの正規性をヒストグラム、正規確率紙、S-W 検定で調べよ。





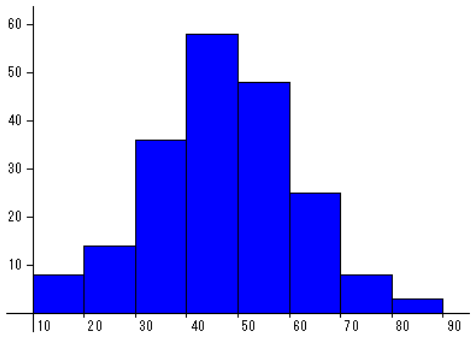
S-W 検定　　確率［　　］判定　正規分布と［みなす・いえない］。

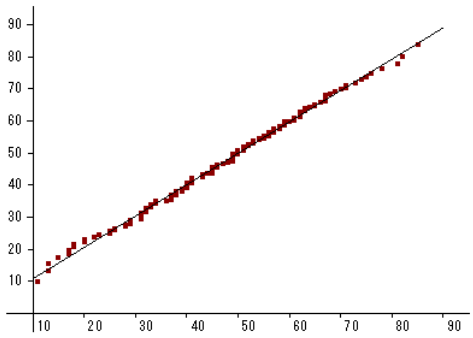
２）年収の平均値（中央値）は 610 万円と比べて差があるといえるか。分析を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　610 万円より多いと［いえる・いえない］。

3）支出のデータの正規性をヒストグラム、正規確率紙、S-W 検定で調べよ。





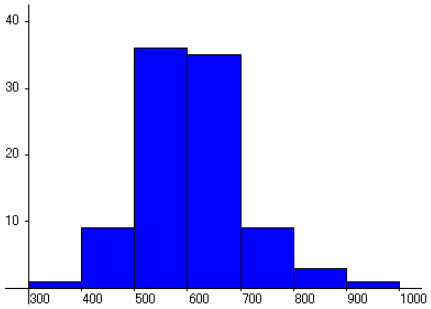
S-W 検定　　確率［　　］判定　正規分布と［みなす・いえない］。

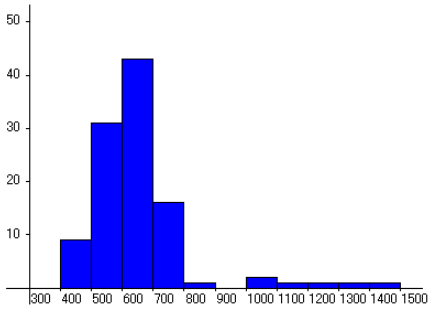
4）支出の平均値（中央値）は 44 万円と比べて差があるといえるか。分析を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　44 万円と比べて差があると［いえる・いえない］。

5）地域別に年収のデータの正規性を調べよ。





地域１　　確率［　　］判定　正規分布と［みなす・いえない］。

地域２　　確率［　　］判定　正規分布と［みなす・いえない］。

演習６

ある中学のクラスについて英語・数学・国語の試験結果を調べた。

１．性別

１）男子　　２）女子

２．英語点数

３．数学点数

４．国語点数

演習 6.txt のデータを集計し、以下の問いに答えよ。但し、正規性の判定にはヒストグラム、正規確率紙、S-W 検定の結果を総合すること。特にデータ数が少ない場合、S-W 検定の確率が 0.05 以上でも、ヒストグラムや正規確率紙で見て明らかに正規分布と異なる場合は、判定困難としておくこと。

問題

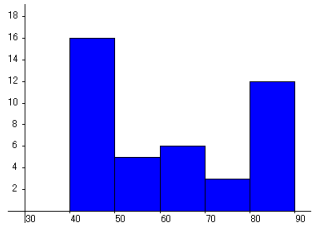
１）男女の人数について求める。　　男子［　　］人　女子［　　］人

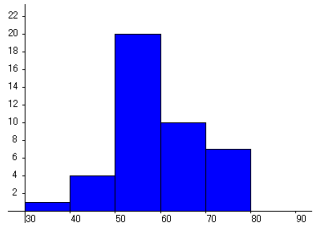
２）英語、数学、国語の平均値、中央値、標準偏差を求める。

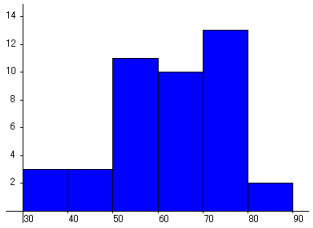
	英語	数学	国語
平均			
中央値			
標準偏差			

３）標準偏差の大きさより、英語、数学、国語のうち、最も点数の広がりが大きいのは［　　］。

４）英語、数学、国語の点数のヒストグラムを描く。





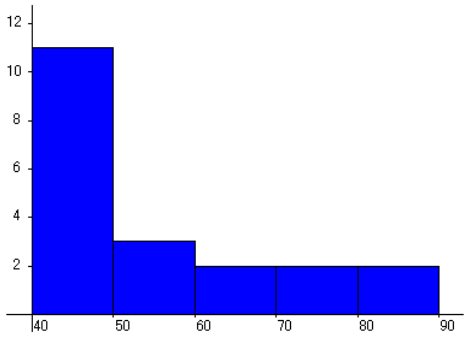


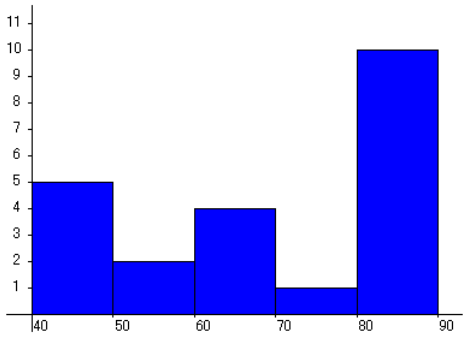
数学のヒストグラムで最頻値はいくらか。［　　］点

５）英語、数学、国語の平均値について男女別に求める。

	英語	数学	国語
男子			
女子			

6）男女の英語の点数のヒストグラムを描く。





男子の英語のヒストグラムで最頻値はいくらか。［　　］点

7）英語、数学、国語の分布について正規性を判定する。

英語　S-W 検定確率［　　］正規分布と［みなす・いえない］

数学　S-W 検定確率［　　］正規分布と［みなす・いえない］

国語　S-W 検定確率［　　］正規分布と［みなす・いえない］

8）英語の男女別の分布について正規性を判定する。

男子　S-W 検定確率［　　］正規分布と［みなす・いえない］

女子　S-W 検定確率［　　］正規分布と［みなす・いえない］

9）数学の男女別の分布について正規性を判定する。

男子　S-W 検定確率［　　］正規分布と［みなす・いえない］

女子　S-W 検定確率［　　］正規分布と［みなす・いえない］

10）国語の男女別の分布について正規性を判定する。

男子　S-W 検定確率［　　］正規分布と［みなす・いえない］

女子　S-W 検定確率［　　］正規分布と［みなす・いえない］

11）英語の平均値（中央値）は 60 点以上といえるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　60 点以上と［いえる・いえない］。

12）数学の平均値（中央値）は 50 点以上といえるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　50 点以上と［いえる・いえない］。

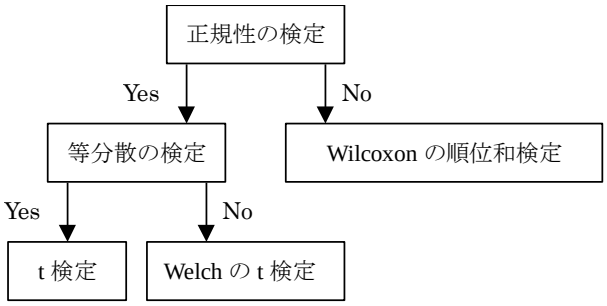
13）英語の女子の平均値（中央値）は 60 点以上といえるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　60 点以上と［いえる・いえない］。

５．２群間の量的データの検定

5.1　対応のない検定手順



5.2　対応のない２群間の分散の検定（正規性あり）

例

A機を導入した会社 18 社（１群）とＢ機を導入した会社 15 社（２群）について、機械 10 台当り１年間の故障発生件数を調べ、不偏分散を求めたら以下の結果を得た。

	平均	不偏分散
１群	10.56	10.68
２群	8.22	3.17

分布は正規分布であると仮定して、分散に差があるといえるか有意水準 5%で判定せよ。

理論　F 検定

母分散に差がないとすると

$$F = \frac{u_1^2}{u_2^2} \sim F_{n_1-1,\, n_2-1} \text{ 分布}$$

解答

$F = 3.3691$　　$p = 0.01321 < 0.05$　より、分散に差があるといえる。

5.3　対応のない２群間の平均値の検定（正規性あり・等分散）

例

ある地域の同性・同年齢の児童について、ある要因の有無による２つの集団の体重を調べたところ以下のデータを得た。２つの集団の平均値に差はあるといえるか。正規性、等分散性を仮定して、有意水準 5%で判定せよ。

	データ数	平均	不偏分散
要因なし	20	40.2	25.5
要因あり	20	36.4	16.0

理論（student の）t 検定

母平均に差がないとすると

$$t=\sqrt{\frac{n_1n_2}{n_1+n_2}}\frac{\bar{x}_1-\bar{x}_2}{\sqrt{\dfrac{(n_1-1)u_1^2+(n_2-1)u_2^2}{n_1+n_2-2}}}\sim t_{n_1+n_2-2}\text{ 分布}$$

解答

$t=2.637999$　　$p=0.01202<0.05$　より、平均に差があるといえる。

5.4　対応のない２群間の平均値の検定（正規性あり・等分散性なし）

例

A機を導入した会社 18 社（１群）とB機を導入した会社 15 社（２群）について、機械 10 台当り１年間の故障発生件数を調べ、平均と不偏分散を求めたところ以下の結果を得た。正規性があり、異分散であるとして、２群間の平均に差があるかどうか有意水準 5%で検定せよ。

	平均	不偏分散
１群	10.56	10.68
２群	8.22	3.17

理論　Welch(ウェルチ)の t 検定

母平均に差がないとすると

$c=\dfrac{u_1^2/n_1}{u_1^2/n_1+u_2^2/n_2}$　として、自由度を　$d=\dfrac{1}{\dfrac{c^2}{n_1-1}+\dfrac{(1-c)^2}{n_2-1}}$　とし、

$$t=\frac{\bar{x}_1-\bar{x}_2}{\sqrt{u_1^2/n_1+u_2^2/n_2}}\sim t_d\text{ 分布}$$

解答

$c=0.7374$　　$d=27.0931\cong 27$（自由度）（小数点以下切り捨て）

$t=2.60860$　　$p=0.01464<0.05$　より、平均に差があるといえる。

問題１

ある１人当りの売上のデータについて、2 つの地域の支店を比較したところ、以下の結果が得られた。２群間に差があるといえるか。有意水準 5%の両側検定で判定せよ。

１群　2007, 2344, 2434, 2251, 2673, 1452, 2393, 2126, 2485, 1279, 2269

２群　2579, 2899, 2258, 3086, 2998, 2829, 2408, 2287, 3020, 1989, 2136

検定名［　　］　確率［　　］

判定　母平均（母集団の中央値）に差があると［いえる・いえない］

5.5　対応のない２群間の中央値の検定（正規性なし）

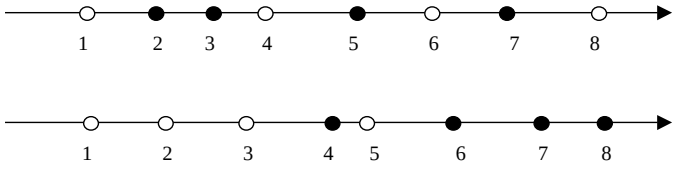
例

ある１人当りの売上のデータについて、2 つの地域の支店を比較したところ、以下の結果が得られた。２群の売上は１群のそれに比べて大きいといえるか。有意水準 5%の両側検定で判定せよ。

１群　2060, 2350, 1550, 1720, 1800, 1990, 1510, 1720, 2910, 1820, 2600

２群　1720, 2064, 2072, 2005, 2602, 1987, 1824, 2060, 2035, 1890, 2025

概要　Wilcoxon(ウィルコクソン)の順位和検定



両群のデータの小さい順に順位を付け、データ数の少ない群（$n_1 \le n_2$）の順位和をW とする。但し、同じ値にはそれらが異なると考えた場合の順位の平均値を付ける。

データ数が多い場合　両群の中央値が等しいとすると

$$z = \frac{|W - n_1(n_1+n_2+1)/2| - 1/2}{\sqrt{n_1 n_2 (n_1+n_2+1)/12}} \sim N(0,1)$$ 分布（正の部分）　（Yates の連続補正）

解答

$p = 0.34102 > 0.05$ より、中央値に差があるといえない。

問題２

以下の標本データの母平均（母集団の中央値）には差があるといえるか。検定を選んで有意水準 5%で判定せよ。

１群　112, 106, 101, 112, 102, 98, 108, 95, 101, 90, 110, 97, 95, 105, 101, 113, 114, 91

２群　98, 88, 105, 99, 96, 93, 109, 106, 103, 87, 107, 102, 97, 91

検定名［　　］　確率［　　］

判定　母平均（母集団の中央値）に差があると［いえる・いえない］

問題３

以下の標本データの母平均（母集団の中央値）には差があるといえるか。検定を選んで有意水準 5%で判定せよ。

１群　358, 469, 397, 350, 329, 446, 393, 379, 443, 348,
　　455, 332, 311, 424, 420, 354, 353, 390, 434, 430

２群　335, 387, 385, 343, 394, 351, 404, 391, 330, 363,
　　319, 334, 348, 396, 408, 403, 415, 353, 377, 399

検定名［　　］　確率［　　］

判定　母平均（母集団の中央値）に差があると［いえる・いえない］

問題４

ラットの体重増加(g)を、条件を変えた２つのグループで測定したところ、以下の結果が得られた。２群の体重増加に差は認められるか、有意水準 5%で判定せよ。

１群：7.2, 8.3, 5.4, 6.0, 7.3, 11.7, 10.5, 8.0, 9.1

２群：10.1, 13.2, 7.4, 9.1, 16.2, 14.5, 6.3, 11.2, 12.4, 7.4, 12.5, 9.1, 17.0

検定名［　　］　確率［　　］

判定　体重増加に差があると［いえる・いえない］

問題５

Samples¥テキスト 9.txt のデータを用いて以下の問いに答えよ。

１）地域別の年収に差があるか、検定を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　地域別の年収に差があると［いえる・いえない］

２）地域別の支出に差があるか、検定を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　地域別の支出に差があると［いえる・いえない］

３）意見１別の年収に差があるか、検定を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　意見１で答え方が違う人で年収に差があると［いえる・いえない］

演習６－２

ある中学のクラスについて英語・数学・国語の試験結果を調べた。

１．性別

１）男子　　２）女子

２．英語点数

３．数学点数

４．国語点数

演習 6.txt のデータを College Analysis に読み込んで集計し、各科目の平均または中央値に差があるかどうか検定せよ。特に正規性が判定困難な場合、一般の分布の検定を行って判定を下すこと。

問題

１）男女の人数を求める。

男子［　　］人　女子［　　］人

２）英語、数学、国語の平均と中央値を求める。

	英語	数学	国語
平均			
中央値			

３）英語、数学、国語の標準偏差を求める。

	英語	数学	国語
標準偏差			

４）最も成績の良かった科目は何か。　［英語・数学・国語］

最も点数のばらつきの大きかった科目は何か。　［英語・数学・国語］

５）それぞれ科目間の相関係数はいくらか。

英語と数学　［　　］

数学と国語　［　　］

国語と英語　［　　］

これらの科目間には関係があると思うか。　［ある・ない］と思う。

６）英語と数学の平均（中央値）に差があるか、（対応がないとして）有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　科目間に差があると［いえる・いえない］。

７）数学と国語の平均（中央値）に差があるか、（対応がないとして）有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　科目間に差があると［いえる・いえない］。

８）国語と英語の平均（中央値）に差があるか、(対応がないとして）有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　科目間に差があると［いえる・いえない］。

９）英語、数学、国語の平均について男女別に求める。

	英語	数学	国語
男子			
女子			

10）英語、数学、国語の中央値について男女別に求める。

	英語	数学	国語
男子			
女子			

11）男女別の英語の平均（中央値）に差があるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　男女差があると［いえる・いえない］。

12）男女別の数学の平均（中央値）に差があるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　男女差があると［いえる・いえない］。

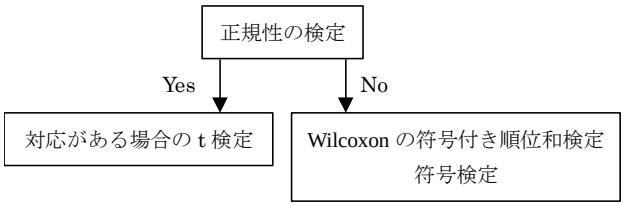
13）男女別の国語の平均（中央値）に差があるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　男女差があると［いえる・いえない］。

14）対応のない２群間の量的データの比較の検定手法で、最も一般的に使えるのは［　　］検定で、逆に最も制約が多いのは［　　］検定である。しかし［　　］や等［　　］などの制約が満たされるとき、後者は最も検出力の［高い・低い］検定となり、２群間の差は見つけ［やすく・にくく］なる。

5.6　対応がある検定手順



5.7　対応がある２群間の平均値の検定（正規性あり）

例

ある商品の陳列位置を変える前と後とで売上高（千円）を規模の等しい８つの支店で比較したところ、以下の結果を得た。検定を選択して有意水準 5%で差があるかどうか判定せよ。

前	385	402	320	383	504	417	290	342
後	396	373	431	457	514	405	380	396

理論

対応する各標本の差（z_i=標本１－標本２）をとる。平均が等しいと仮定すると

$$t=\frac{\sqrt{n}\,\bar{z}}{u_z}\sim t_{n-1}\text{ 分布}$$

解答

$t=2.1499$　$p=0.068675>0.05$　より、平均に差があるとはいえない。

5.8　対応がある２群間の中央値の検定（正規性なし）

例

ある商品の陳列位置を変える前と後とで売上高（千円）を規模の等しい８つの支店で比較したところ、以下の結果を得た。検定を選択して有意水準 5%で売上高に差があるかどうか判定せよ。

前	385	402	320	383	504	417	290	342
後	396	310	342	407	514	405	380	365

概要　Wilcoxon の符号付き順位和検定

対応する各標本の差（z_i=標本１－標本２）について、z_i の正負で 2 群に分けて順位和を求め、小さい方を R とする。標本数が多いとき（少ない場合は数表を用いる）

$$z=\frac{|R-n(n+1)/4|-1/2}{\sqrt{n(n+1)(2n+1)/24}}\quad\sim\quad N(0,1)\text{ 分布（正の部分）}$$

解答

$p=0.38200>0.05$ より、２標本の中央値に差があるといえない。

注）2 群のデータの分散は大きいが、各データ間の差が同じ符号の傾向がある場合、対応のある検定が非常に有効となる。（テキスト 5.txt 7 ページ）

問題

ある小学生の集団で国語・算数・社会・理科の学力を調べたところ以下のようなデータを得た。質問に答えよ。

国語	68	58	60	63	55	69	63	79	62	74	53	75	64	77	66
算数	75	59	58	73	59	69	62	67	68	78	53	67	69	77	70
社会	66	58	50	55	57	66	54	91	57	56	65	55	80	90	63
理科	82	60	61	74	68	74	64	72	70	65	57	79	76	83	74

1）４科目の平均値と中央値を求める。

	国語	算数	社会	理科
平均値				
中央値				

2）各科目のデータの正規性を検討する。（下段にはみなす／いえないかを書き込む）

	国語	算数	社会	理科
S-W 検定確率				
正規性ありと				

3）対応があるとして以下の科目間の点数の差の正規性を検討する。（同上）

	国語－算数	国語－社会	算数－理科	社会－理科
S-W 検定確率				
正規性ありと				

２群の比較ではデータ間に１対１の対応がある場合、通常対応がある検定手法を利用するが、対応がないとして検定しても間違いではない。以下の問題は両方の方法で検定を行い、結果を比較せよ。

4）国語と算数の平均値（中央値）に差があるといえるか、有意水準 5%で判定する。

	検定名	確率	判定
対応なし			差があると［いえる・いえない］
対応あり			差があると［いえる・いえない］

5）社会と理科の平均値（中央値）に差があるといえるか、有意水準 5%で判定する。

	検定名	確率	判定
対応なし			差があると［いえる・いえない］
対応あり			差があると［いえる・いえない］

演習 7

都市圏と地方都市の小学生４年生からサンプルを選んで実施した調査で、演習 7.txt のような結果が得られた。どんなことが読み取れるか考え、質問に答えよ。

１．住んでいる地域

１）都市圏（東京・大阪）　　２）地方都市（その他）

２．性別

１）男子　　２）女子

３．どんな遊びが好きか

１）外で運動　　２）家の中でゲーム・読書他

４．体力測定（点数換算）

５．国語

６．算数

問題

１）地域・男女別に人数を求める。

	男子	女子	合計
都市圏			
地方都市			
合計			

２）地域による好きな遊びの傾向をみる。

	外で	家の中で	合計
都市圏			
地方都市			
合計			

３）家の中で遊ぶ割合は過半数といえるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　家の中で遊ぶ割合は過半数であると［いえる・いえない］

４）好きな遊びに地域差はあるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　好きな遊びに地域差はあると［いえる・いえない］

５）男女による好きな遊びの傾向をみる。

	外で	家の中で	合計
男子			
女子			
合計			

6）好きな遊びに男女差はあるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　好きな遊びに男女差はあると［いえる・いえない］

7）地域別の体力測定の平均値と中央値を求める。

	都市圏	地方都市
平均値		
中央値		

8）体力測定の結果に地域差があるといえるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　地域による差があると［いえる・いえない］。

9）男女別の体力測定の平均値と中央値を求める。

	男子	女子
平均値		
中央値		

10）体力測定の結果に男女差があるといえるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　男女による差があると［いえる・いえない］。

11）地域別の算数の結果

	都市圏	地方都市
平均値		
中央値		

12）算数の結果に地域差があるといえるか、有意水準 5%で判定する。

検定名［　　］　確率［　　］

判定　地域による差があると［いえる・いえない］。

13）国語と算数の結果に差があるといえるか、有意水準 5%で判定する。

この問題は対応がないとしても、対応があるとしても答えを求めることができる。

対応がないとした場合

検定名［　　］　確率［　　］

判定　国語と算数に差があると［いえる・いえない］。

対応があるとした場合

検定名［　　］　確率［　　］

判定　国語と算数に差があると［いえる・いえない］。

本来どちらの検定手法が良いか。対応が［ない・ある］とした場合が良い。

６．相関係数の検定と回帰分析

6.1 （Pearson の）相関係数

例

２つの商品 A, B の地域別使用率（%）のデータは以下の通りである。それぞれの商品の使用率に線形の相関が認められるか。正規性を仮定して、有意水準 5%で検定せよ。

A(%)	33	24	30	50	42	15	15	56	13	45	44	21	18	31	27	40
B(%)	20	34	50	20	58	23	12	34	26	56	42	5	25	51	19	27

理論

母相関係数を 0 と仮定して以下の性質を利用する。

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \text{ 分布}$$

解答

$r = 0.453,\ n = 16,\ t = 1.905387$

$p = 0.077 > 0.05$ より、相関があるといえない。

（相関係数が 0 と異なるといえない。この検定は正規分布以外では使えない。）

6.2 （Spearman の）順位相関係数

例

前節の問題で、それぞれの商品の使用率に相関（非線形のものも含む）が認められるか。正規性を仮定せずに、有意水準 5%で検定せよ。

理論

順位相関係数 r_s を求め、母相関係数を 0 と仮定して以下の性質を用いる。

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \sim t_{n-2} \text{ 分布}$$

解答

$r_s = 0.461,\ t = 1.945443$

$p = 0.072 > 0.05$ より、相関があるとはいえない。

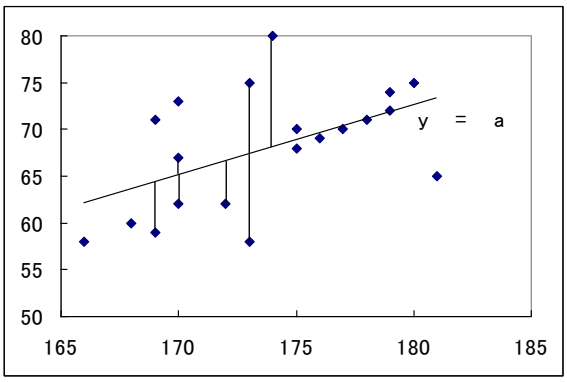
6.3 回帰分析

例

下の表のデータを用いて、身長により体重を推定する式を考える。ただし、式は１次式（体重 $=a\times$身長 $+b$）と仮定し、その有効性を検討せよ。

体重	71	68	67	72	69	80	75	65	74	71
身長	169	175	170	179	176	174	173	181	179	178
体重	62	75	70	70	62	58	60	58	59	73
身長	170	180	177	175	172	166	168	173	169	170

理論



回帰式の決定

2 変数の関係を、$y=ax+b$ の直線で表わすとすると、x を説明変数、y を目的変数と呼ぶ。データ点からこの直線へ垂直におろした線の長さの２乗が最小となるように係数 a,b を決める。

平均 $\bar{x},\bar{y}$，標準偏差 u_x,u_y，相関係数 r とすると

$$a=r\frac{u_y}{u_x},\quad b=\bar{y}-r\frac{u_y}{u_x}\bar{x}$$

回帰式の有効性の検討

重相関係数 R　目的変数の実測値と回帰式による予測値の相関係数

（説明変数が１つの場合　$R=r$）

寄与率（重決定係数）R^2　目的変数の変動のうち回帰式が説明する割合

回帰式の有効性の検定（残差が正規分布する場合のみ利用可能）

回帰式は無意味（傾きが 0）と考えられる確率で検討する。

解答

$\bar{x} = 173.7$， $\bar{y} = 67.95$

$u_x = 4.402153$， $u_y = 6.378211$， $r = 0.513047$

$a = 0.743346$， $b = -61.1692$

回帰式　　$y = 0.743346x - 61.1692$

重相関係数　　$R = 0.5130$

寄与率　　$R^2 = 0.2632$

回帰式の有効性の検定

確率　0.0207　　回帰式は有効であるといえる。

問題 1

以下の 2 変数のデータを用いて問いに答えよ。

変数1	65	86	78	83	85	89	83	80	85	93	75	85	79	80
変数2	162	210	224	179	217	230	223	204	224	197	186	189	172	185

1）２変数の Pearson の相関係数と Spearman の順位相関係数を両方を求めよ。

相関係数	順位相関係数

2）相関の検定にはどちらの相関係数を利用するか。

［相関係数・順位相関係数］

3）上で選んだ相関係数を用いて、相関の有無を有意水準 5%で判定せよ。

検定確率［　　］　相関があると［いえる・いえない］

4）変数２を目的変数、変数１を説明変数として回帰分析を行う。

回帰式　変数２＝［　　］×変数１＋［　　］

重相関係数　［　　］

寄与率　　［　　］

5）回帰分析の有効性の検定は［行える・行えない］。

検定確率［　　］

回帰式は有効であると［いえる・いえない］

問題２

Samples¥テキスト 9.txt のデータを用いて以下の問題に答えよ。

１）年収と支出についての相関係数と順位相関係数を求める。

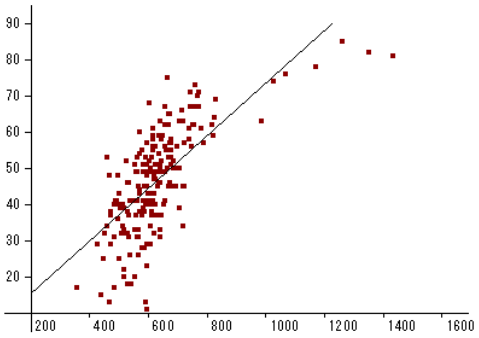
相関係数［　　］　順位相関係数［　　］

２）年収と支出に相関があるといえるか、相関係数を選んで有意水準 5%で判定する。

［相関係数・順位相関係数］で見る。

判定　確率［　　］　相関があると［いえる・いえない］

３）年収（横軸）と支出（縦軸）について以下のような散布図を描く。



４）支出を目的変数、年収を説明変数として回帰分析を行う。

回帰式　支出＝［　　］×年収＋［　　］

重相関係数　［　　］

寄与率　　［　　］

５）回帰分析の有効性の検定は［行える・行えない］。

検定確率［　　］

回帰式は有効であると［いえる・いえない］

演習８

学生（男子）の地域別の身長・体重，試験の平均点と勉強時間を比べるために、調査を行い演習 8.txt のデータを得た。以下の問題に答えよ。

１）地域（1：都市部 2：郊外）

２）身長

３）体重

４）実施した学力試験の点数（５科目の平均点）

５）１日平均の勉強時間

問題

１．都市部と郊外の調査数

都市部［　　］人　郊外［　　］人

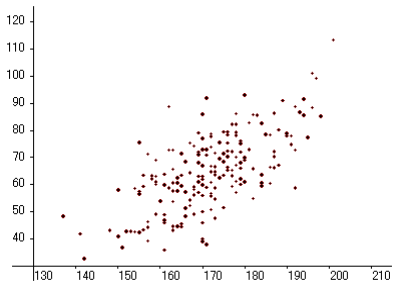
２．地域別にみた各変数の平均値と中央値を求める。

	身長	体重	点数	勉強時間
地域１平均値				
地域１中央値				
地域２平均値				
地域２中央値				

３．各項目に地域差があるといえるか、有意水準 5%で判定する。

	検定名	検定確率	判定　差があると
身長			いえる・いえない
体重			いえる・いえない
点数			いえる・いえない
勉強時間			いえる・いえない

４．身長と体重の散布図と相関係数



相関係数［　　］，順位相関係数［　　］

身長と体重には相関があるといえるか相関係数を選んで判定する。

［相関係数・順位相関係数］でみる。

検定確率［　　］　相関があると［いえる・いえない］

5．体重を身長で予測する回帰分析結果

体重＝［　　］×身長＋［　　］

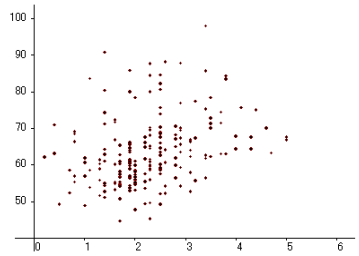
体重がどの程度身長で説明されるか示す。　寄与率　［　　］

回帰式の有効性の検定は可能か？可能な場合は有意水準 5%で判定する。

［可能・不可能］

検定確率［　　］　この回帰式は有効であると［いえる・いえない］

6．勉強時間と点数の散布図と相関係数



相関係数［　　］，順位相関係数［　　］

勉強時間と点数には相関があるといえるか相関係数を選んで判定する。

［相関係数・順位相関係数］でみる。

検定確率［　　］　相関があると［いえる・いえない］

7．点数を勉強時間で予測する回帰分析結果

点数＝［　　］×勉強時間＋［　　］

点数がどの程度勉強時間で説明されるか示す、寄与率　［　　］

回帰式の有効性の検定は可能か？可能な場合は有意水準 5%で判定する。

［可能・不可能］

検定確率［　　］　この回帰式は有効であると［いえる・いえない］

8．勉強時間を２時間未満と２時間以上に分けて、点数について調べる。

	２時間未満	２時間以上
平均値		
中央値		

9．２時間未満と２時間以上で点数に差があるといえるか、有意水準 5%で判定する。

検定名［　　］　検定確率［　　］

点数に差があると［いえる・いえない］。

７．区間推定

区間推定

標本から推測される母比率や母平均などがどの位の値の範囲に入るかを推定し、区間で表す方法。

信頼係数

推定した区間に母比率や母平均などが入る確率(%で表されることが多く、通常95%か99%)。１－信頼係数の値は検定での有意水準に相当する。

7.1　母比率の区間推定

例

ある制度についてのアンケート調査をランダムに抽出された 100 人に対して行ったところ、賛成 65 人、反対 35 人であった。母集団の賛成の比率を、信頼係数 95%（有意水準 5%に相当）で推定せよ。また、調査数 1000 人で同じ比率ではどうか。

理論

データ数n、標本比率$\hat{p}$の標本から、母比率pを信頼係数$(1-\alpha)\times 100\%$で推定する。

$z_0 = normsinv(1-\alpha/2)$として、信頼区間は以下で与えられる。

$$\hat{p} - \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0 \le p \le \hat{p} + \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} z_0$$

解答

$n = 100$、$\hat{p} = 65/100 = 0.65$、$\alpha = 0.05$

$$0.55652 \le p \le 0.74348$$

1000 人では、以下のように精度が上がる。

$$0.62044 \le p \le 0.67956$$

7.2　正規母集団の母平均と母分散の区間推定

例

ある標本データから所得について集計したところ以下の結果を得た。母集団は正規分布するとして母平均と母分散を信頼係数 95%で推定せよ。

データ数 30，平均 620，標準偏差 90

また、データ数を 100 にすると結果はどう変わるか？

理論

正規分布する母集団から得られた標本より、母平均μと母分散σ^2を信頼係数$(1-\alpha)\times 100\%$で推定する。データ数をn, 標本平均を$\bar{x}$, 不偏分散をu^2, $t_0 = tinv(\alpha, n-1)$, $x_1 = chiinv(1-\alpha/2, n-1)$, $x_2 = chiinv(\alpha/2, n-1)$として、各信頼区間は以下で与えられる。

母平均：$\bar{x} - \frac{u}{\sqrt{n}} t_0 \le \mu \le \bar{x} + \frac{u}{\sqrt{n}} t_0$

母分散：$\frac{(n-1)u^2}{x_2} \le \sigma^2 \le \frac{(n-1)u^2}{x_1}$

解答

$n = 30$, $\bar{x} = 620$, $u = 90$

$$586.4 \le \mu \le 653.6$$

$$5138 \le \sigma^2 \le 14638$$

データ数を 100 にすると、以下のように精度が向上する。

$$602.142 \le \mu \le 637.858$$

$$6244 \le \sigma^2 \le 10932$$

問題１

ある 500 人に対する調査で支持 205 人、不支持 295 人という結果を得た。母集団における支持の比率を信頼係数 95%で推定せよ。

信頼区間は［　　］≦母比率≦［　　］

問題２

正規分布を仮定して、以下の身長データ(cm)から母平均と母分散を信頼係数 95%で推定せよ。

184, 170, 164, 176, 177, 170, 171, 159, 174, 170,
165, 170, 171, 183, 175, 169, 181, 172, 171, 164

［　　］≦母平均≦［　　］

［　　］≦母分散≦［　　］

問題３

Samples¥テキスト 9.txt のデータを用いて以下の問いに答えよ。

１）ここで用いた区間推定の手法は、支出に利用できるか。［可能・不可能］
　可能な場合は支出の母平均と母分散を信頼係数 95%で推定せよ。

［　　］≦母平均≦［　　］

［　　］≦母分散≦［　　］

２）上の結果を用いて、支出の平均は 50（万円）と差があるかどうか有意水準 5%で判定したい。　信頼区間［内・外］なので、差があると［いえる・いえない］。

演習９

都市部と郊外に分けて検討事項に関する意見、月収、ある品目に対する支出を調べて演習 9.txt のデータを得た。

１．地域（1：都市部 2：郊外）

２．意見（1：支持 2：不支持）

３．月収（万円）

４．支出（万円）

以下の問いに答えよ。但し、正規性の検定は簡単のため S-W 検定の結果を元にすること。

１）地域についての回答者数を求めよ。

都市部［　　］人　　郊外［　　］人

２）意見について回答者数を求めよ。

支持［　　］人　　不支持［　　］人

３）意見で支持は過半数といえるか、有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　過半数と［いえる・いえない］

４）地域と意見についての２次元分割表を書け。

	支持	不支持
都市部		
郊外		

５）都市部と郊外とで支持率に差があるか、有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　差があると［いえる・いえない］

６）月収と支出の平均を求めよ。

月収［　　］万円　　支出［　　］万円

７）月収の平均は３０万円と比べて差があるか、検定を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　差があると［いえる・いえない］

８）地域別の月収と支出の平均を求めよ。

	月収	支出
都市部		
郊外		

9）地域別の月収について都市部と郊外とで差があるか、検定を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　差があると［いえる・いえない］

10）地域別の支出について都市部と郊外とで差があるか、検定を選んで有意水準 5%で判定せよ。

検定名［　　］　確率［　　］

判定　差があると［いえる・いえない］

11）月収と支出の相関係数を求めよ。また相関の有無を調べるにはどちらを使うか。

相関係数［　　］　順位相関係数［　　］

［相関係数・順位相関係数］で調べる。

12）支出を月収で予測する回帰式を求め、支出の変動を回帰式が説明する割合である寄与率を示せ。また、この場合は回帰式の有効性の検定は行えるか。

支出＝［　　］×月収＋［　　］

寄与率［　　］　回帰式の有効性の検定は［行える・行えない］

13）データの先頭から 3 人の回帰式の予測値はいくらか。

［　　］［　　］［　　］

14）回帰式を使って月収で支出を予測した場合、月収 30 万円の人の支出はいくらか。

予測値　支出［　　］万円

15）月収 30 万円未満とそれ以上に分けた場合、意見の支持率に差があるといえるか、有意水準 5%で判定せよ。メニュー［分析－基本統計－量から質変換］を利用する。

検定名［　　］　確率［　　］

判定　差があると［いえる・いえない］

16）意見の支持率（1 の比率）の値を信頼区間 95%で推定する。メニュー［分析－基本統計－区間推定－比率の推定］を利用する。

［　　］≦支持率≦［　　］

17）支出（正規性が必要）の母平均の値を信頼区間 95%で推定する。メニュー［分析－基本統計－区間推定－平均と分散の推定］を利用する。

［　　］万円≦母平均≦［　　］万円

８．アンケート調査

アンケート注意事項

１）アンケートは次の順序で作る。

タイトル，あいさつ文，調査団体または代表者名，

アンケート本文，謝辞

２）何を知りたいか十分検討し、アンケート対象者や項目を選ぶ。

アンケートの対象は、全数調査か、調べたい対象の中から無作為に抽出した標本とする。但し、年齢構成などで層別に抽出する場合もある。

質問に漏れがないか十分注意する。

例えば意見の男女差を知りたければ、当然性別を聞いておく必要がある。

最初に区分けのための質問、続いて具体的な意見などを聞く方が答え易い。

集計のことを頭に置いて質問項目を考える。

不必要なことはできるだけ聞かずに、アンケートをコンパクトにまとめる。

３）質問は答え易い形で書く。

数字を書かせる場合と自由記述を除いては、番号を選ぶのが無難。

例　あなたの性別は　　１）男　　２）女

集計と統計処理の簡単化のため、番号選択は１つか、いくつでもかが無難。

例　あなたの最も大切にしていることはなんですか。以下から１つだけ選んで下さい。

あなたの大切にしているものはなんですか。以下の該当するものすべてを選んで下さい。

明らかな場合を除いて、選択肢の中には「その他」の項目を設け、具体的な内容を書く欄を添える。

例　１）製造業　　２）流通業　　３）サービス業

４）その他［　　］

具体的な数字を書かせる場合は、単位を明確に。（千円はやめておくべき）

例　あなたの年収は　　＿＿＿＿万円

質問項目の右側に回答欄を設けると集計に便利であるが、利用しない人もいるので注意する。

回答者を絞って答えてもらう場合は、分かり易さを心掛ける。

例　前問で「１）はい」と答えた人のみ回答して下さい。その他の人は設問５へ進んで下さい。

４）その他

予め集計用のフォームを考えておく。（大規模でなければ Excel は有力）

あらかじめ少数の人で試し、集計までをシミュレーションしておく。

回収後、回答用紙には必ず整理番号を振っておく。

学生生活アンケート調査

この度統計学の授業において、アンケートの作成法とその集計方法を学ぶために仮想的なアンケート調査を実施することになりました。個人のプライバシー等につきましては十分な注意を払うことはもちろんですが、このアンケートをその他の目的に使用することはありません。どうかご協力をお願い致します。

福山××大学　福山太郎

質問１　あなたの性別は？

１）男性　　２）女性

質問２　あなたは自宅通学ですか？

１）自宅通学　　２）自宅通学でない

質問３　あなたの自由に使えるお金（生活費を除く）は１ヶ月におよそいくらですか？

［＿＿＿＿＿＿円］

質問４　あなたはアルバイトをしていますか？

１）している　　２）していない

前問で**１）している**と答えた人だけ回答して下さい。その他の人は**質問７**へ進んで下さい。

質問５　どれ位の頻度でアルバイトをしていますか？１つ選んで下さい。

１）週５日以上　　２）週３，４日　　３）週１，２日

４）長期休業時のみ　　５）その他［　　］

質問６　あなたのアルバイトの収入は１ヶ月におよそいくらですか？不定期にやっている人は、１ヶ月にならしてお答え下さい。

［＿＿＿＿＿＿円］

質問７　あなたの現在の悩みに当てはまるものがあればいくつでも選択して下さい。

１）特にない　　２）勉学上の問題　　３）金銭問題　　４）異性問題

５）健康上の問題　　６）就職・進路の問題

７）その他［＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿］

ご協力有難うございました。

学生生活アンケート調査報告書

福山××大学　福山太郎

福山××大学では 20XX 年 7 月 28 日に、本学統計学の授業で受講生 53 名を対象に「学生生活アンケート調査」を対面して記述させる方式で実施した。調査結果の回収数は□で回収率は□%であった。

男女別にみると男□名、女□名であり、自宅通学かどうかをみると自宅通学□名、自宅通学以外は□名であった。アルバイトをしている学生は□名、していない学生は□名で、アルバイトをしている割合は、□%であった。アルバイトをしているかどうか通学区分別に見ると、表 1 のようになった。

表 1 通学区分によるアルバイト状況

	している	していない
自宅		
自宅外		

これから通学区分によるアルバイト状況の有意差は見られなかった。また、アルバイトの頻度は、週 5 回以上□名、3～4 回□名 1～2 回□名であった。

自由に使える 1 ヶ月の金額は、平均□万円、標準偏差□万円であり、そのヒストグラムを描くと、図 1 のようになった。

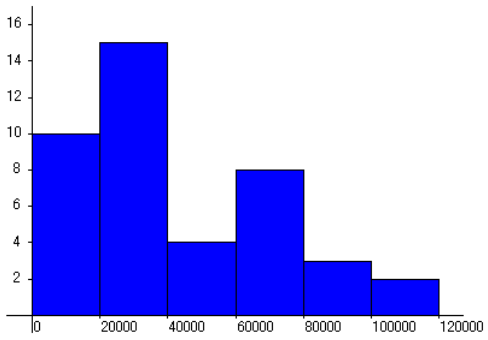


図 1　自由に使える金額

性別、通学別、アルバイト状況別の自由に使える金額の平均は表 2 のようになった。

表 2 各分類別平均（万円）

性別		通学		アルバイト	
男	女	自宅	自宅外	している	していない

これらについて差を調べたところ（[　　]検定)、アルバイトをしているかどうかで有意な差が見られたが（p< [0.05・0.01・0.001])、その他については有意な差は見られなかった。もう少しデータ数を増やして、男女間の差について検討するのも興味深い。

アルバイト収入の平均は□万円、標準偏差は□万円であった。また、自由に使える金額とアルバイト収入の関係は、図 2 で与えられ、アルバイト収入がないものを除いた順位相関係数は□であった。このことからアルバイト収入と自由に使える金額には相関関係があると思われる。

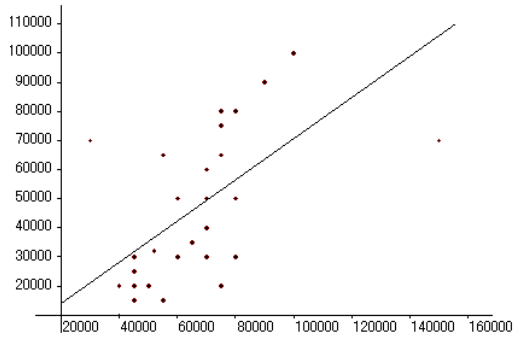


図 2　アルバイト収入（横軸）と使える金額（縦軸）の相関

自由に使える金額を目的変数、アルバイト収入を説明変数として回帰分析を行なったところ、寄与率□で、y=□x+□という結果が得られた。回帰直線は図 2 に記入している。

悩みについては「なし」が□名、項目のどれかにチェックをした学生は□名であった。全体の中で悩みの種類毎の比率は、図 3 のようになる。不況を反映してであろうか、金銭と就職の問題の比率が高いように思われる。

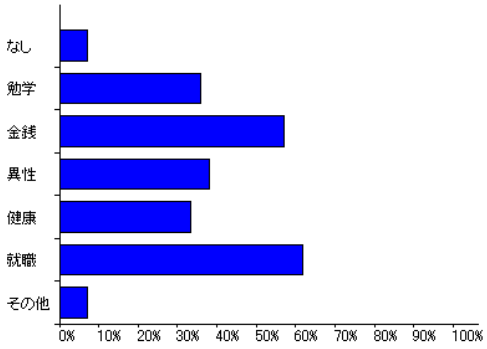


図 3 悩みの種類の割合

アンケート報告書注意事項

１）タイトル、調査団体名または代表者名及び住所等（ここまで表紙でもよい）を最初に示す。

２）アンケートの実施時期と実施方法、対象数と回収数・回収率を明記する。

３）アンケート集計結果は以下の点に注意する。

単純集計から始めて、次にクロス集計をする。

図表には番号とタイトルを付け（通し番号または章ごと)、文中で指定して説明を加える。　　例　　図１に設問３のヒストグラムを示す。

図表番号とタイトルを付ける位置として、表は上側、図は下側が多い。

必要があれば、その他を選んだ場合の内容を紹介してもよい。

質問用紙を最後に掲載するのもよい。

４）集計・検定結果の表示

集計値の桁数は、平均・標準偏差等でデータ桁数より 1 桁か 2 桁程度多く表示する。

例：171, 173, 174, ･･･ → 平均 172.7

検定の際、検定統計量の値や自由度などはあまり書かれることがないが、検定手法の名前は書く場合もある（書かない場合もある)。

検定確率値については、あまり具体的な数値を書くことはなく、n.s., $p<0.05$, $p<0.01$, $p<0.001$（n.s., *, **, *** にしてもよい）のどれかの書き方にすることが多い。

メモ

演習１０

２つの地域（関東と関西）で中学生の英語に対する意識と学力調査（時期をずらして２度、ほぼ同じレベル）を行ない、演習 10.txt のような結果を得た。

１．地域（1：関東 2：関西）
２．意識（英語は大切と思うか　1：思う 2：思わない）
３．勉強時間（１週間当り）
４．成績１
５．成績２

以下の問いに答えよ。但し、正規性の検定は S-W 検定の結果を元にすること。

１）地域について回答者数を求めよ。
　関東［　　］名　　関西［　　］名
２）意識について回答者数を求めよ。
　思う［　　］名　　思わない［　　］名
３）英語を大切と思う人は過半数といえるか、有意水準 5%で判定せよ。
　検定名［　　］　確率［　　］
　判定　過半数と［いえる・いえない］
４）地域と意識についての２次元分割表を描け。

	思う	思わない
関東		
関西		

５）関東と関西とで意識に差がみられるか、有意水準 5%で判定せよ。
　検定名［　　］　確率［　　］
　判定　差があると［いえる・いえない］
６）勉強時間と成績１の平均を求めよ。
　勉強時間［　　］　成績１［　　］
７）勉強時間は 4.5 時間と比べて差がみられるか、有意水準 5%で判定せよ。
　検定名［　　］　確率［　　］
　判定　差があると［いえる・いえない］
８）地域別の勉強時間と成績１の平均を求めよ。

	勉強時間	成績１
関東		
関西		

９）勉強時間について関東と関西とで差があるといえるか、有意水準 5%で判定せよ。
　検定名［　　］　確率［　　］
　判定　差があると［いえる・いえない］

10）成績１について関東と関西とで差があるといえるか、有意水準 5%で判定せよ。
検定名［　　］　確率［　　］
判定　差があると［いえる・いえない］

11）成績１と成績２とで差があるといえるか、有意水準 5%で判定せよ。
検定名［　　］　確率［　　］
判定　差があると［いえる・いえない］

12）成績１と勉強時間の相関係数を求めよ。
相関係数［　　］

13）成績１を勉強時間で予測する回帰式を求め、成績１の変動を回帰式が説明する割合である寄与率を示せ。
成績１＝［　　］×勉強時間＋［　　］
寄与率［　　］

14）勉強時間を 4 時間未満と 4 時間以上に分けたとき、それぞれ英語を大切に思う人の割合を示せ。（メニュー［基本統計－量から質変換］で新しい変数を作って求めます。）

	４時間未満	４時間以上
大切に思う割合		

15）上の場合、勉強時間で大切に思う意識に差があるといえるか、有意水準 5%で判定せよ。
検定名［　　］　確率［　　］
判定　差があると［いえる・いえない］

16）地域と意識別に勉強時間の平均を求めよ。（メニュー［ツール－文字列結合］で地域と意識を結合して新しい変数を作って求めます。）

関東・思う	関東・思わない	関西・思う	関西・思わない

17）関西で、英語を大切に思うかどうかの意識により勉強時間に差があるといえるか、有意水準 5%で判定せよ。
検定名［　　］　確率［　　］
判定　差があると［いえる・いえない］

多変量解析編

１．実験計画法

1.1　１元配置実験計画法

多群間の平均や中間値に差があるかどうか検討する手法であり、いくつの変数を同時に比較するかによって１元配置（１元比較ともいう)、２元配置（２元比較ともいう）などと分かれている。ここでは理解し易い１元配置についてのみ解説する。

多群間の等分散の検定には Bartlett 検定を利用する。

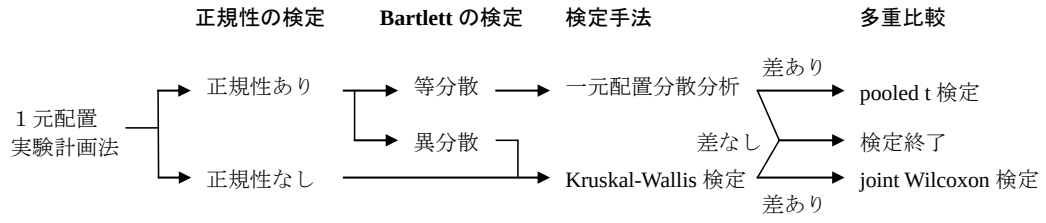


図 1.1　１元配置実験計画法の構造

例

３つの条件である商品の売上を調査したところ、Samples¥分散分析 ex.txt の結果を得た。各群に差があるといえるか、実験計画法を用いて有意水準 5%で判定せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　条件間に差があると［いえる・いえない］

差があるとするとどの条件間に差があるか。差がある条件同士を条件２＜条件３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］

結果［　　］

問題１

Samples¥分散分析 1.txt は３つの工場群の不良品率を与えたものである。各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　工場群間の不良品率に差があると［いえる・いえない］

差があるとするとどの条件間に差があるか。差がある条件同士を工場２＜工場３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］

結果［　　］

問題２

Samples¥分散分析 2.txt は４つの群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　群間に差があると［いえる・いえない］

差があるとするとどの群間に差があるか。差がある群同士を群２＜群３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］

結果［　　］

問題３

Samples¥分散分析 3.txt は３群のデータであるが、各群に差があるといえるか、実験計画法を用いて有意水準 5%で検討せよ。

正規性の検定　　正規分布と［みなす・いえない］

等分散性の検定　　検定確率［　　］　等分散と［みなす・いえない］

検定名［　　］　検定確率［　　］

判定　群間に差があると［いえる・いえない］

差があるとするとどの群間に差があるか。差がある群同士を群２＜群３（これは実際の結果とは関係ない）のように不等号で表せ。

検定名［　　］

結果［　　］

演習

ある４つの中学について英語・数学・国語の試験結果を調べた。多変量演習 1.txt のデータを読み込んで、以下の質問に答えよ。但し、検定は有意水準 5%とすること。

１．中学　　1）A 中学　2）B 中学　3）C 中学　4）D 中学

２．英語点数

３．数学点数

４．国語点数

１）数学について、各中学の平均（中央）値に差があるといえるか。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

２）数学について各中学の平均（中央）値に差があるとすると、A, B, C, D どの中学の間に差があるか調べて A<B のように不等号で表せ。（差がある場合のみ）

検定名［　　］

結果［　　］

３）国語について、各中学間の平均（中央）値に差があるといえるか。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

４）国語について、各中学間の平均（中央）値に差があるとすると、A, B, C, D どの中学の間に差があるか調べて A<B のように不等号で表せ。（差がある場合のみ）

検定名［　　］

結果［　　］

５）３教科の平均（中央）値に差があるといえるか。対応は考えないものとせよ。

検定名［　　］　検定確率［　　］

判定　平均（中央）値に差があると［いえる・いえない］。

６）３教科の平均（中央）値に差があるとすれば、どの教科の間に差があるか調べて英語＜数学のように不等号で表せ。（差がある場合のみ）

検定名［　　］　結果［　　］

1.6　対応がある場合の１元配置問題

例

３つの条件である商品の売上を調査したところ、分散分析 ex.txt の結果を得た。各データに対応があるとして差があるか検定せよ。

条件１　115，110，108，114，120，116，108，112，115，122
条件２　121，118，124，117，119，130，121，115，118，119
条件３　116，112，120，111，112，108，114，119，104，113

注）正規性の有無により、（繰り返しのない）２元配置分散分析か Friedman 検定を利用する。これは repeated measured 1 元配置分散分析、repeated measured Kruskal-Wallis 検定とも呼ばれている。

正規性の検定　　正規分布と［みなす・いえない］
検定名［　　］　検定確率［　　］
判定　群間に差があると［いえる・いえない］

1.7　２元配置分散分析

例

分散分析 4.txt は作物の品種と肥料の組み合わせによる収穫量を表したデータである。2 元配置分散分析を用いて判定せよ。

注）2 元配置分散分析では、分類データの組み合わせによる交互作用が分かる。

品種水準間
　検定確率［　　］　水準間に差があると　［いえる・いえない］
肥料水準間
　検定確率［　　］　水準間に差があると　［いえる・いえない］
交互作用
　検定確率［　　］　交互作用に差があると［いえる・いえない］

２．重回帰分析

2.1　重回帰分析

例

以下のデータ（Samples¥重回帰分析 1.txt）をもとに体重を身長と胸囲の１次関数で予測する。

体重	身長	胸囲	体重	身長	胸囲
61.0	167.0	84.0	49.5	164.7	78.0
55.5	167.5	87.0	61.0	171.0	90.0
57.0	168.4	86.0	59.5	162.6	88.0
57.0	172.0	85.0	58.4	164.8	87.0
50.0	155.3	82.0	53.5	163.3	82.0
50.0	151.4	87.0	54.0	167.6	84.0
66.5	163.0	92.0	60.0	169.2	86.0
65.0	174.0	94.0	58.8	168.0	83.0
60.5	168.0	88.0	54.0	167.4	85.2
49.5	160.4	84.9	56.0	172.0	82.0

解説

体重 ＝ b_1 身長＋b_2 胸囲＋b_0　　の形で体重を予測する。

目的変数：体重　　説明変数：身長，胸囲

係数の値は？　　→　偏回帰係数

説明変数の重要性は？　　→　標準化偏回帰係数

どの程度予測できるか？　　→　重相関係数，寄与率（決定係数）

このモデルは有効か？　　→　Ｆ検定値と確率（要残差正規性）

それぞれの係数は有効か？　　→　ｔ検定値と確率（要残差正規性）

他の変数の影響を除いた目的変数と各説明変数の相関は？　　→　偏相関係数

どの程度予測できているのか図的に見たい　　→　散布図

どの程度予測できているのかデータ毎に見たい　　→　予測値と残差

問題１

Samples¥重回帰分析 2.txt について、重回帰分析を行い、以下の問いに答えよ。

１）回帰式を求めよ。

卒業試験＝　［　　］入試点数　＋［　　］内申点数

＋［　　］勉強時間　＋［　　］出席率

＋［　　］

２）この回帰式の寄与率を求めよ。［　　］

３）この場合残差の分布は正規分布といえるか。［正規分布・正規分布でない］

変数増減法を用いて変数を自動選択する。

４）最終的な回帰式はどのようになるか。不要な変数の係数欄は空欄のままでよい。

卒業試験＝　［　　］入試点数　＋［　　］内申点数

＋［　　］勉強時間　＋［　　］出席率

＋［　　］

５）上の回帰式の寄与率を求めよ。［　　］

６）上の回帰式の寄与率はすべての変数を使った場合に比べ大きく下がっているか。

［大きく下がっている・あまり下がっていない］

７）この式を新しい予測モデルとして採用するか。

［採用する・採用しない］

８）新しい予測モデルで、データ中の最初（1 番）の学生について卒業試験の実測値，その予測値，残差（実測値と予測値の差）はいくらか。

実測値［　　］　予測値［　　］　残差［　　］

９）上と同様のモデルで、質問項目の値が入試点数 70、内申点数 3.5、勉強時間 5、出席率 70%の学生の卒業試験はいくらに予測されるか。

［　　］

演習

１）多変量演習 3.txt で、全変数を使った以下の重回帰式はどのように与えられるか。

試験成績＝［　　］×評定平均＋［　　］×模試１

［　　］×模試２＋［　　］×模試３＋［　　］

２）この重回帰式の寄与率はいくらか。［　　］

変数自動選択で偏回帰係数が有効である回帰モデルを作り、以下の問いに答えよ。

３）重回帰式はどのようになったか。説明変数に含まれないものは空欄のままにすること。

試験成績＝［　　］×評定平均＋［　　］×模試１

［　　］×模試２＋［　　］×模試３＋［　　］

４）寄与率はいくらになったか。［　　］

５）上の重回帰式を新しい予測モデルにして良いと思うか。［思う・思わない］

以後、新しいモデルで答えること。

６）データの中で最初の学生の予測試験成績はいくらか。［　　］

７）新しい重回帰式を利用すると以下の点数の学生の試験成績は何点に予測されるか。

変数名	評定平均	模試１	模試２	模試３
成績	3.5	70	73	75

予測試験成績［　　］

2.2　非線形最小 2 乗法

例

週が進むごとに以下の表（非線形最小 2 乗法 1.txt）のように変化する売上高データを、関数形　売上高 = a/(1+b*exp(-c*月)) を用いて表したい。非線形最小 2 乗法を用いてパラメータ a, b, c の値を定め、式の精度を求めよ。

週	売上高	週	売上高
1	15	9	373
2	21	10	485
3	34	11	568
4	58	12	648
5	87	13	681
6	134	14	738
7	201	15	748
8	265	16	755

解説

非線形最小 2 乗法の目的

ある変数（目的変数）の変動を、他の変数（説明変数）で予測する数式を定め、その精度を求める。

方法　計算式にパラメータを含む数式を入力し、「最小 2 乗解」ボタンを押す。

パラメータの値は？　→　計算結果の表の中に表示

式の精度は？　→　実測・予測 R，R^2（説明できる割合）

計算の安定性は？　→　収束解個数

予測の精度を数値で確かめるには？　→　「予測値と残差」ボタン

実測値と予測値の差を図で確かめるには？　→　「実測/予測の散布図」ボタン

予測曲線の形状は？（1 変数のときのみ）　→　「1 変数の予測グラフ」ボタン

問題 1

非線形最小 2 乗法 2.txt の 1 ページ目のデータを用いて、非線形最小 2 乗法の計算を行い、以下の問いに答えよ。但し、計算式は　y=a*x+b*sin(x)+c　とすること。

1）パラメータの値を求めよ。

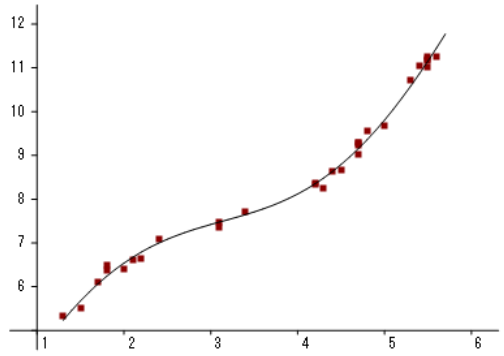
a=［　　］,　b=［　　］,　c=［　　］

２）R^2 の値を求めよ。［　　］

３）先頭の人の実測値、予測値、残差を求めよ。

実測値［　　］，予測値［　　］，残差［　　］

４）関数形を描け。



問題２

非線形最小 2 乗法 2.txt の 2 ページ目のデータを用いて、非線形最小 2 乗法の計算を行い、以下の問いに答えよ。但し、計算式は　y=a*x1+b*x2^2+c　とすること。

１）パラメータの値を求めよ。

a=［　　］，　b=［　　］，　c=［　　］

２）R^2 の値を求めよ。［　　］

３）先頭の人の実測値、予測値、残差を求めよ。

実測値［　　］，予測値［　　］，残差［　　］

2.3　局所重回帰分析

これまでの重回帰分析や非線形最小 2 乗法の予測手法は、パラメータを含んだ関数形を仮定し、最小 2 乗法によってパラメータの値を定め、予測関数を確定するものであった。しかし、局所重回帰分析は要求点を与えることによって、その近傍の点による重回帰分析の結果から直接予測値を求める方法で、関数形を必要としない予測手法である。

解説

局所重回帰分析の目的

非線形な重回帰分析を、関数形を仮定せず直接予測値を求める方法で実現する。

予測式は？　→　要求点を与えた際の線形重回帰式　「局所重回帰分析」ボタン
（偏回帰係数の値　他の要求点では使えない）

予測値は？　→　要求点を与えて求める。「局所重回帰分析」ボタン

バンド幅 p とは？　→　どの程度要求点の近傍のデータを利用するかを決める値
大きいほど遠くまで利用する。∞で通常の重回帰分析

要求点を与えた予測の数値を見るには？　→　「予測値と残差」ボタン

要求点を与えた予測の状態を見るには？　→　「実測/予測散布図」ボタン
２変量までなら回帰散布図

１個抜き交差検証について

１個抜き交差検証（LOOCV）とは？　→　要求点を各データ点にし、その点を除いて予測する検証手法

予測の精度は　→　各要求点（=データ点）の実測値と予測値の相関係数（重相関係数）及び残差２乗平均の平方根（RMSE：小さいほど良い）

各点の予測と実測の値は？　→　１個抜き交差検証内の「予測値と残差」ボタン

各点の予測と実測の状態を見るには？　→　同「散布図」

予測へのバンド幅 p の依存性は？　→　「p 依存性」ボタン（最小のところが良い値）

問題

重回帰分析 6（局所）.txt の１頁目を読み込んで以下の問いに答えよ。

１）要求点を先頭のデータ（番号 1）の位置にした際の、局所重回帰式を求めよ。
目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］

２）そのときの先頭のデータの実測値、予測値、残差、ウェイトを求めよ。
実測値［　　］予測値［　　］残差［　　］ウェイト［　　］

３）そのときの実測／予測散布図を見よ。

４）そのときの２変量回帰散布図を見よ。

５）30 番目のデータの実測値、予測値、残差、ウェイトを求めよ。
実測値［　　］予測値［　　］残差［　　］ウェイト［　　］

6）要求点を 30 番目のデータの位置にした際の、局所重回帰式を求めよ。
目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］
7）そのときの先頭のデータの実測値、予測値、残差、ウェイトを求めよ。
実測値［　　］予測値［　　］残差［　　］ウェイト［　　］
8）そのときの実測／予測散布図を見よ。
9）そのときの２変量回帰散布図を見よ。

10）要求点を（50,50）にしたときの局所重回帰式を求めよ。
目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］
11）そのときの実測／予測散布図を見よ。
12）バンド幅 p を 100 にした場合の局所回帰式を求めよ。
目的変数＝［　　］説明変数１＋［　　］説明変数２＋［　　］
13）そのときの実測／予測散布図を見よ。
14）予測値と残差のところで、ウェイトの大きさを見よ。

<u>バンド幅を元に戻して</u>
15）１個抜き交差検証をした場合の RMSE と重相関係数の値を求めよ。
RMSE［　　］　重相関係数［　　］
16）１個抜き交差検証をした場合の１番の実測値、LOOCV 予測値、残差を求めよ。
実測値［　　］LOOCV 予測値［　　］残差［　　］
17）１個抜き交差検証の予測の程度を散布図を使って見よ。
18）バンド幅 p を動かして、RMSE が最小となる p 値はどこか。p=［　　］

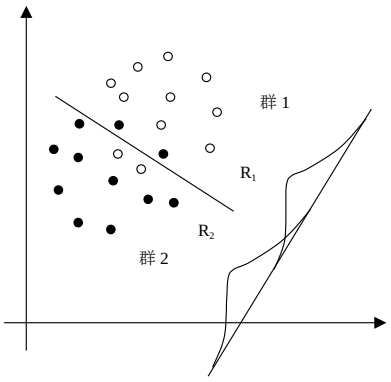
３．判別分析

例

入学試験の合否と勉強時間・模擬試験の平均点のデータを求めたところ以下のような結果を得た（Samples¥判別分析１.txt)。合否を判定するための勉強時間と平均点の１次関数を求めよ。またこの関数によってこのデータを判別し、誤判別の確率を求めよ。

合否	勉強時間	平均点	合否	勉強時間	平均点
1	5.6	70.2	2	3.8	67.4
1	5.9	74.2	2	3.8	61.3
1	4.1	72.7	2	1.7	60.6
1	5.1	84.9	2	2.7	77.2
1	5.0	93.0	2	4.3	65.9
1	3.2	80.5	2	3.3	74.4
1	4.3	62.7	2	3.5	72.1
1	4.8	85.4	2	2.1	69.7
1	3.3	84.3	2	4.3	68.7
1	5.3	64.8	2	2.0	70.5
1	5.3	60.7	2	3.6	45.9
1	5.4	74.4	2	2.8	54.6
1	3.6	85.5	2	2.5	64.4
2	3.8	47.9	2	5.2	50.7
2	3.9	70.8	2	2.2	65.7

解説



判別分析の目的

2 群（多群）を判別する最適な１次式を求める。

判別得点＝b_1 勉強時間＋b_2 平均点＋b_0

判別関数

判別関数の係数は？　→　判別関数の欄

判別関数で群を分けるのは？

→　判別の分点0(多群の場合値が最大の群)

判定に影響を与える変数は？　→　標準化係数の絶対値の大きい変数

各係数の有効性は？（要正規性・等共分散性）

→　確率の欄（係数が 0 と異なるかの検定）

誤判別の程度は？　→　誤判別確率（実測と理論）(理論値は要正規性・等共分散性)

マハラノビス距離とは　→　どの程度 2 群が離れているかを表わす指標

マハラノビス距離	1	4	9	16	25
誤判別確率	0.309	0.159	0.067	0.023	0.006

データ毎の判別関数の値と判別状況　→　判別得点

問題１

Samples¥判別分析 2.txt は、適性の有無の判定（有：1，無：2）と適性検査の結果とＳＰＩの結果を与えたデータである。判定を適性検査とＳＰＩで予測する判別分析を行い、以下の問いに答えよ。但し、事象の生起確率は各群同じ、誤判別損失は２群とも１とすること。

１）判別関数を求めよ。

判別得点＝［　　］適性検査＋［　　］ＳＰＩ＋［　　］

２）どちらの変数が判定に影響があると思われるか。［適性検査・ＳＰＩ］

３）実測値から求めた誤判別の確率は？

適性有りを無しと［　　］　適性無しを有りと［　　］

４）先頭（1 番）の人の判別得点はいくらか。［　　］

５）適性検査 50 点，ＳＰＩ 55 点の人の判別得点はいくらか、またその人の適性の有無を判定せよ。　判別得点［　　］　適性［有り・無し］

問題２

Samples¥判別分析 3.txt はあやめの種類をがくの長さと幅、花弁の長さと幅で３群に分類したデータである。あやめの群を他の変数の１次式で判別する３群以上の判別分析を行い、以下の問題に答えよ。但し、設定は前問と同じとする。

１）３つの判別得点の式を求めよ。

判別得点１＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点２＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

判別得点３＝［　　］がくの長さ＋［　　］がくの幅
＋［　　］花弁の長さ＋［　　］花弁の幅＋［　　］

２）実測値から求めた誤判別確率はいくらか。

群１を他と［　　］群２を他と［　　］群３を他と［　　］

３）先頭のデータの３つの判別得点を求めよ。

判別得点１［　　］判別得点２［　　］判別得点３［　　］

４．主成分分析

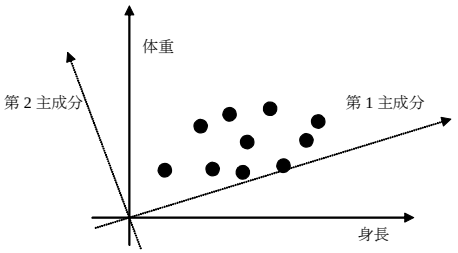
例

以下の健康診断のデータ（Samples¥主成分分析 1.txt）から、変数の１次関数として体格を表す特徴的な指標を作り、その意味を考察せよ。

身長	体重	胸囲	座高	身長	体重	胸囲	座高
148	41	72	78	139	34	71	76
160	49	77	86	149	36	67	79
159	45	80	86	142	31	66	76
153	43	76	83	150	43	77	79
151	42	77	80	139	31	68	74
140	29	64	74	161	47	78	84
158	49	78	83	140	33	67	77
137	31	66	73	152	35	73	79
149	47	82	79	145	35	70	77
160	47	74	87	156	44	78	85
151	42	73	82	147	38	73	78
157	39	68	80	147	30	65	75
157	48	80	88	151	36	74	80
144	36	68	76	141	30	67	76
139	32	68	73	148	38	70	78

解説

Samples¥主成分分析 1.txt のデータから、変数の１次関数として体格を表す特徴的な指標を作る。



主成分分析の目的

複数の変数を１次関数として組み合わせて、いくつかの特徴的な量を作り出す。

各主成分の係数値は？　→　固有ベクトルの値（全体的に符号を変えてもよい）

各主成分のばらつき（分散）は？　→　各主成分の固有値

各主成分の重要性は？　→　各主成分の寄与率（変動の何%を表すか）

各主成分と各変数の関係は？　→　因子負荷量（各主成分と各変数の相関係数）

何番目の主成分まで意味があるか？　→　等固有値の検定（要正規性）

主成分が意味がある→他の主成分と値が異なる

データごとの主成分の値は？　→　主成分得点

共分散行列からと相関行列からどちらを使う　→　実用的には相関行列が一般的

問題

Samples¥主成分分析 2.txt は生徒の教科別の成績データである。相関行列をもとにするモデルを用いて以下の問いに答えよ。

1）各主成分の固有値（分散の値)、寄与率、累積寄与率を求めよ。

	第１主成分	第２主成分	第３主成分	第４主成分	第５主成分
固有値					
寄与率					
累積寄与率					

2）主成分を２つ使うとすると、第１主成分と第２主成分の関数はどのように表されるか。

第１主成分＝［　　］英語＋［　　］数学
＋［　　］国語＋［　　］理科＋［　　］社会

第２主成分＝［　　］英語＋［　　］数学
＋［　　］国語＋［　　］理科＋［　　］社会

3）これら２つの主成分で説明できるのは全体の変動の何%か。［　　］%

4）これら２つの主成分はどのように意味づけられるか。

第１主成分　意味［　　］を表す指標

第２主成分　意味［　　］を表す指標

5）先頭（1 番）の生徒の２つの主成分得点を求めよ。

第１主成分得点［　　］　第２主成分得点［　　］

6）２つの主成分の意味を考えて、この生徒にはどんな特徴があるか。

［　　］

５．因子分析

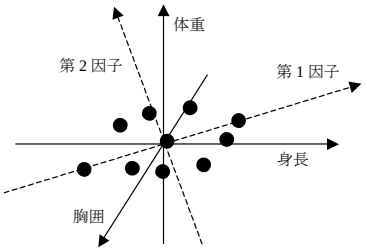
例

以下の健康診断のデータ（Samples¥因子分析 1.txt）から、変数の背後にある体格を表す共通因子を求め、その意味を考察せよ。

身長	体重	胸囲	座高	身長	体重	胸囲	座高
148	41	72	78	139	34	71	76
160	49	77	86	149	36	67	79
159	45	80	86	142	31	66	76
153	43	76	83	150	43	77	79
151	42	77	80	139	31	68	74
140	29	64	74	161	47	78	84
158	49	78	83	140	33	67	77
137	31	66	73	152	35	73	79
149	47	82	79	145	35	70	77
160	47	74	87	156	44	78	85
151	42	73	82	147	38	73	78
157	39	68	80	147	30	65	75
157	48	80	88	151	36	74	80
144	36	68	76	141	30	67	76
139	32	68	73	148	38	70	78

解説

Samples¥因子分析 1.txt のデータから、体格を表す共通因子を求める。（標準化された）各変数 z_i は因子 f_α を用いて以下の形で表されるものとする。



$$z_i \cong \sum_{\alpha=1}^{q} a_{i\alpha} f_\alpha$$

因子分析の目的

各変数の背後にある共通因子を求め、それらの１次関数として各変数が表されるように係数を求める。

各因子の係数値は？　→　因子負荷量の値（全体的に符号を変えて見てもよい）

各因子と各変数の相関係数は？　→　因子負荷量の値（因子間は無相関とした場合）

各因子の重要性は？　→　各因子の寄与率

何番目の因子まで考えるか？　→　累積寄与率が 90%程度まで（寄与率も見る）

相関行列の固有値で１より大きい固有値の数

因子が各変数の変動（分散）を説明する程度は？→　共通性の値

データごとの因子の値は？　→　因子得点

問題

Samples¥因子分析 3.txt は北海道各地の２月の気温データである。設定はデフォルトとして以下の問いに答えよ。注）江差町（えさし：南部），寿都町（すっつ：南部），小樽市（おたる：中部），留萌市（るもい：北部），天塩町（てしお：北部）

１）各都市間の相関行列の固有値を大きい順に４つ求めよ。

1	2	3	4

以後因子数を２つと決めて各質問に答えよ。

２）各因子の寄与率と累積寄与率を求めよ。

	第１因子	第２因子
寄与率		
累積寄与率		

３）因子数は２つでよいか。［よい・注意が必要］

４）各因子の因子負荷量を求めよ。

	江差	寿都	小樽	留萌	天塩
第１因子					
第２因子					

５）上の因子負荷量の値から各因子の意味を解釈せよ。

第１因子：［　　］の気温を代表する因子

第２因子：［　　］の気温を代表する因子

６）各地の気温の変動は因子によりどの程度説明されるか（共通性）。

江差	寿都	小樽	留萌	天塩

７）最初の３日間の因子の値（因子得点）を推定せよ。

	第１因子	第２因子
1		
2		
3		

８）この３日間、北海道はどのような気候だったか。

［　　］

９）このモデルは良いモデルと思うか。

［良いと思う・あまり良いと思わない］

演習

多変量演習8.txtのデータはある学校で測定した小学６年生の運動適性テストの結果である。因子分析を用いて特徴を分析し、以下の問いに答えよ。

１）各科目間の相関行列の固有値を大きい順に求めよ。

1	2	3	4	5

２）因子数を２つとして、因子分析を行い、寄与率を求めよ。

因子１	因子２

３）因子数を３つとして、因子分析を行い、寄与率を求めよ。

因子１	因子２	因子３

本来なら因子数３が良いが、解釈の問題から、以後因子数を２つと決めて各質問に答えよ。但し、バリマックス回転ありとすること

４）各因子の因子負荷量を求めよ。

	立幅跳び	腹筋	腕立伏せ	往復走	５分間走
第１因子					
第２因子					

５）この場合の各因子の意味を解釈せよ。

第１因子：[　　]を表す因子

第２因子：[　　]を表す因子

６）先頭から３人及び、特徴的な９番目の人の因子の値（因子得点）を推定せよ。

	第１因子	第２因子
1		
2		
3		
9		

７）９番目の人にはどんな特徴があるか。因子負荷量の符号に注意して考えよ。

[　　]

８）各種目の変動は因子によりどの程度説明されるか。

立幅跳び	腹筋	腕立伏せ	往復走	５分間走

９）予測値が２つの因子から予測されたことを考えると、この分析はうまくいったと思うか。　［まずまずうまくいった・うまくいっていない］

６．クラスター分析

例

表　各人の好みを 1～9 の点数で表わした表（Samples¥クラスター分析 1.txt）

	日本酒	焼酎	ビール	ウィスキー	ワイン
増川	1	2	9	6	5
西山	3	1	7	5	4
三好	5	3	4	2	2
芝田	3	6	2	8	3
尾崎	4	6	9	3	4
藤田	7	2	5	4	5
細川	7	5	4	3	2

クラスター分析の目的

１）回答の類似度で個人を分類する。　→　個体（レコード）の分類

２）回答の類似度で変数を分類する。　　→　変数の分類

クラスター分析は分類をどのように表示するか　→　デンドログラム（解答参照）

デンドログラムの縦軸は　→　要素またはクラスター間の距離（類似の程度を示す量）

要素間の距離とは

個体間について

量的データ：ユークリッド距離、標準化ユークリッド距離、マハラノビス距離等

質的 0/1 データ：類似比、一致係数、φ係数等を使ったもの

変数間について

量的データ：1-相関係数、1-相関係数|、1-順位相関係数、1-|順位相関係数|

質的データ：平均平方根一致係数、一致係数、クラメールのＶ等を使ったもの

要素間の距離を知るには　→　距離行列

クラスター構成でよく使われる方法　→　最長距離法、ウォード法

クラスター構成過程を表示するには　→　クラスター構成と距離

問題１

Samples¥クラスター分析 4.txt はある野球チームの今年度の成績である。これについてクラスター分析を行い以下の問いに答えよ。

１）ユークリッド距離及び標準化ユークリッド距離を用いた場合、山下と田中の距離はいくらか。ユークリッド距離［　　］　標準化ユークリッド距離［　　］

２）各変数の標準偏差はいくらか。

打率	安打	本塁打	打点	盗塁

３）上の結果から、距離測定法はどちらを利用すべきか。
［ユークリッド距離・標準化ユークリッド距離］　以後はこの距離を用いる。

４）クラスター構成法を最長距離法とする場合、最初にクラスターを構成するのはどの要素とどの要素でそれらの距離はいくらか。
［　　］と［　　］で距離［　　］

５）最長距離法の場合、４分類か５分類が適当と思われるが、４分類の場合、各クラスターにはどのような要素が含まれるか。
［　　］［　　］［　　］［　　］

６）最長距離法と最短距離法とでどちらの分類が理解しやすいと思われるか。
［最長距離法・最短距離法］

８）１－相関係数の距離測定法で最長距離法を用いて変数を３分類すると各クラスターに含まれる要素はどのようになるか。
［　　］［　　］［　　］

問題２

Samples¥クラスター分析 3.txt のデータを用いてクラスター分析を行い、以下の文を完成させよ。

個体の分類を行う場合、各変数の不偏分散（標準偏差）は、ほぼ等しいと思われるので、距離測定法はユークリッド距離を利用する。またクラスター構成法に最長距離法を用いると、３分類が妥当なように思われる。そのときの各クラスターに含まれる要素は以下のように与えられる。

［　　］［　　］［　　］

変数の分類では、１－相関係数を用いた距離測定法を使い、最短距離法と最長距離法で分類を行ったところ、最終的な２分類で、各クラスターに含まれる要素は、最短距離法で［　　］と［　　］、最長距離法で［　　］と［　　］となった。これらの分類はどちらとも納得できるものである。

７．正準相関分析

例

正準相関分析 1.txt のデータを用いて、複数の変数間で相関の高い特徴的な量を求める。

身長	座高	体重	胸囲
148	78	41	72
160	86	49	77
159	86	45	80
153	83	43	76
:	:	:	:
148	78	38	70

正準相関分析の目的　→　複数の変数からなる２つの群の中で特徴的な量を見出し、それらの最大の相関を求める。

どのようにして相関を考えるのか。

$y = a_1$身長$+ a_2$座高
$z = b_1$体重$+ b_2$胸囲　　正準変数の組 y と z が最大の相関を持つよう係数を選ぶ。

y と z の最大の相関とは　→　正準相関係数（変数の組によって複数ある）

係数はどのように表示されるか。　→　正準相関分析で１群係数と２群係数

正準変数 y と z の各データの値を見るには　→　正準変数値

各変数と同じ群の正準変数との関係は　→　正準負荷量（相関係数）、解釈に利用

各変数と違う群の正準変数との関係は　→　交差負荷量（相関係数）、解釈に利用

複数の正準変数の組が得られるが、他の正準変数の組同士の関係は　→　相関係数 0

問題

正準相関分析 2.txt について、文系科目（英語・国語・社会）と理系科目（数学・理科）に分け、正準相関分析を実行し、以下の問いに答えよ。但し、相関行列を用いたモデルで、第１正準変数について考えること。

１）文系科目と理系科目の正準相関係数はいくらか。［　　］

２）文系科目と理系科目の正準変数はそれぞれどのように表されるか。

文系正準変数＝［　　］英語＋［　　］国語＋［　　］社会

理系正準変数＝［　　］数学＋［　　］理科

３）各変数の正準負荷量の値はいくらか。

英語	国語	社会	数学	理科

4）各変数の交差負荷量の値はいくらか。

数学	理科	英語	国語	社会

5）各正準変数と最も相関のある同じ組の科目は何か。
文系正準変数では［英語・国語・社会］、理系正準変数では［数学・理科］

6）各正準変数と最も相関のある違う組の科目は何か。
文系正準変数へは［数学・理科］、理系正準変数へは［英語・国語・社会］

7）各科目の平均と標準偏差（不偏分散からのもの）を求め、
標準化変数＝（値－平均）／標準偏差
の式によって、英語 60、国語 72、社会 66、数学 58、理科 55 の人の標準化変数値を求めよ。

科目	英語	国語	社会	数学	理科
標準化変数値					

8）上の標準化値を利用して、この人の正準変数の値を求めよ。
文系正準変数［　　］　理系正準変数［　　］

８．数量化Ｉ類

例

以下の地域（1：都市部、2：山村部)、気候（1：温暖、2：平均的、3：寒冷)、ある商品の販売率のデータ（数量化Ｉ類 1.txt）から販売率（目的変数）を予測する式を作り、それがどの程度有効か検討する。

販売率	地域	気候
3.0	1	2
1.8	2	1
:	:	:
2.3	1	3

販売率	地域-1	地域-2	気候-1	気候-2	気候-3
3.0	1	0	0	1	0
1.8	0	1	1	0	0
:	:	:	:	:	:
2.3	1	0	0	0	1

左のようなアイテムのデータから、それぞれのアイテムが複数のカテゴリに分かれる右の形のデータを作る。

このデータをもとに以下の式で目的変数を予測する。

$$Y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{23}x_{23} + a_{00}$$

（基準化）カテゴリウェイト　→　上式の係数 a_{ij}

カテゴリウェイト、重回帰カテギリウェイト、基準化カテゴリウェイトの違いは

　　→　予測値を計算する上では同じ（予測値への影響の見易さが異なる）

予測値と実測値との相関係数　→　重相関係数

予測値は実測値をどれだけ説明しているか　→　寄与率

各アイテムの重要性は　→　相関／偏相関ボタンのウェイト範囲

予測値と実測値の散布図　→　散布図ボタン

問題

数量化Ｉ類 2.txt は店舗の売り上げを立地、人通り、競合の３段階分類データで予測しようとするものである。

１）カテゴリウェイト（定数項を０）を用いた予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３

２）重回帰カテゴリウェイト（各先頭アイテムを基準）を用いた予測式を表せ。

予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

3）基準化カテゴリウェイトを用いた（目的変数の平均値を基準）予測式を表せ。
予測売り上げ＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

4）予測式は実測値の変動を何%予測できるか。［　　］%

5）立地：２，人通り：２，競合：２の店舗の売り上げを予測せよ。［　　］

6）ウェイト範囲で見る場合、予測値に最も大きな影響を与えるアイテムは何か。
［立地・人通り・競合］

7）数量化Ｉ類と同じ分析を 0/1 データを用いた重回帰分析で行った。但し、各アイテムの第１カテゴリは係数が 0 として、変数から外した。そのときの重回帰式を示せ。
予測売り上げ＝［　　］立地２＋［　　］立地３
＋［　　］人通り２＋［　　］人通り３
＋［　　］競合２＋［　　］競合３＋［　　］

8）このことから上の重回帰分析と数量化Ｉ類は［同じ・異なる］ものと考えられる。

９．数量化Ⅱ類

例

顧客が車を購入する際、３種類の特性について検討し、a かｂの車種を購入した（数量化Ⅱ類 1.txt)。顧客がどのような選択を行うかでどちらの車を購入するか判別する式を作る。２分されたアイテムのデータから、数量化Ｉ類と同様な形のデータを作る。

群	価格	外観	性能
a	1	1	2
a	1	2	1
:	:	:	:
b	2	1	3

上のような２分されたアイテムのデータから、以下の形のデータを作る。

群	価格:1	価格:2	外観:1	外観:2	性能:1	性能:2	性能:3
a	1	0	1	0	0	1	0
a	1	0	0	1	1	0	0
:	:	:	:	:	:	:	:
b	0	1	1	0	0	0	1

そのデータからどちらの群に属するかを予測する判別関数を作る。

$$y = a_{11}x_{11} + a_{12}x_{12} + a_{21}x_{21} + a_{22}x_{22} + a_{31}x_{31} + a_{32}x_{32} + a_{33}x_{33} + a_{00}$$

（基準化）カテゴリウェイト　→　上式の係数 a_{ij}

カテゴリウェイトと基準化カテゴリウェイトの違いは

→　判別関数値を計算する上では同じ

判別方法は　→　判別得点の平均（判別の分点）で判定する。群別平均も参照する。

各アイテムの重要性は　→　相関／偏相関ボタンのウェイト範囲

問題

数量化Ⅱ類 1b.txt は店舗の成功か失敗かを立地、人通り、競合の３段階分類データで予測しようとするものである。

１）成否はある売り上げを境に成功と失敗の２つに分けたものであるが（成功：１、失敗：２)、カテゴリウェイトを用いた判別関数を表せ。

判別関数＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

2）基準化カテゴリウェイトを用いた判別関数を表せ。

判別関数＝［　　］立地１＋［　　］立地２＋［　　］立地３
＋［　　］人通り１＋［　　］人通り２＋［　　］人通り３
＋［　　］競合１＋［　　］競合２＋［　　］競合３＋［　　］

3）判別の分点はいくらか。　［　　］

4）誤判別確率はいくらか。

成功を失敗と誤判別［　　］　失敗を成功と誤判別［　　］

5）成功するためにはどちらが有利か。（ヒント　判別得点が大きな値ほど成功になる）

［立地１・立地３］　［人通り１・人通り３］　［競合１・競合３］

6）ウェイト範囲で見る場合、判別に最も影響を与えるアイテムは何か。

［立地・人通り・競合］

7）１番の店舗の判別得点はいくらか。それはどう判別されたか。

判別得点［　　］　判定［成功・失敗］

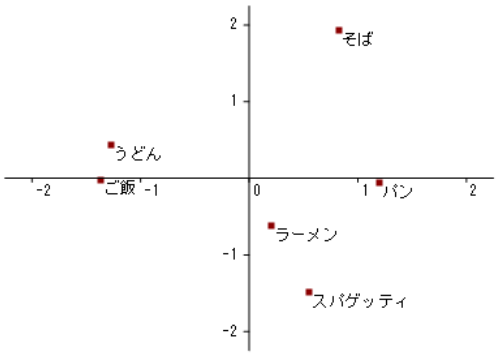
１０．数量化Ⅲ類

例

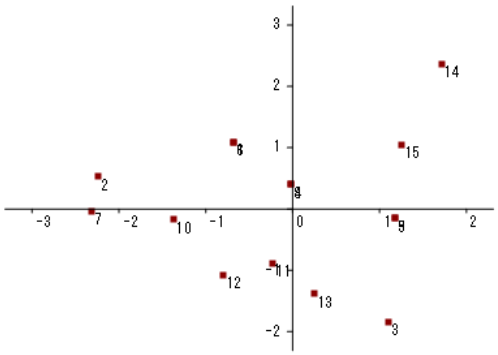
各人が以下の食品について、それぞれの好み（1：好物、0：それほどでも）を与えた（Samples¥数量化Ⅲ類 1.txt）。これから好みの特徴を表す式を求め、人と食品を分類する。

ご飯	パン	うどん	そば	ラーメン	スパ
1	0	1	1	1	0
1	0	1	0	0	0
0	1	0	0	1	1
:	:	:	:	:	:
0	1	0	1	1	0

この分割表のデータ $x_{i\lambda}$ （i：カテゴリ，λ：個体）を元に、行と列別々に類似性を見る分析が数量化Ⅲ類である。類似性は、２つの分類変数にそれぞれ、特徴的な量 u_i^a と v_λ^α を考え、それらの量が最大の相関係数を持つようにして考える。類似度はこれらの量の値の近さで見る。



カテゴリウェイトによる散布図



個体ウェイトによる散布図

問題

数量化Ⅲ類 2.txt は高校生、大学生、会社員の好きなブランドを調査した結果である。1の項目は回答者が○を付けた項目である。数量化Ⅲ類を用いて以下の問いに答えよ。
データの与え方によって、このように違った項目を混在させることもできる。

1）３次元までの寄与率と累積寄与率を求めよ。

	１次元	２次元	３次元
寄与率			
累積寄与率			

2）このうち分析に２次元まで利用するとして、カテゴリウェイトの値を求めよ。

	第１次元	第２次元
A		
B		
C		
D		
高校生		
大学生		
会社員		

3）カテゴリウェイトの散布図から以下の問いに答えよ。

高校生に最も人気のあるブランドは　　［A・B・C・D］
大学生に最も人気のあるブランドは　　［A・B・C・D］
会社員に最も人気のあるブランドは　　［A・B・C・D］

4）先頭３人の２次元までの個体ウェイトの値を求めよ。

	第１次元	第２次元
1		
2		
3		

5）個体ウェイトの散布図から高校生、大学生、会社員はグループになっているか。
［なっている・なっていない］

１１．コレスポンデンス分析

例（高橋信，Excel で学ぶコレスポンデンス分析，オーム社，2005）

各年代の学生に好きな歌手を選んでもらったところ、コレスポンデンス分析 1.txt の集計結果が得られた。それぞれの歌手はどの世代に支持されているか。コレスポンデンス分析で検討せよ。

	A	B	C	D
中学生	10	19	13	5
高校生	13	8	15	16
大学生	18	11	14	8

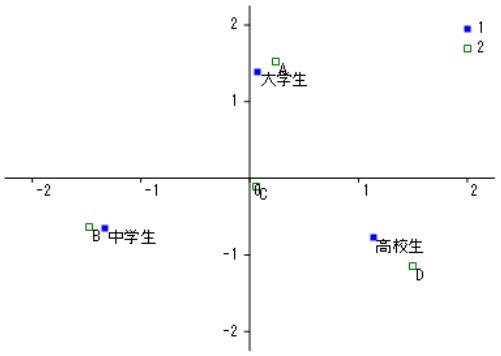
例えば、中学生に指示されている歌手はと考え、10/41, 19/38, 13/42, 5/29 と考えた結果と、A を支持する学生はと考え、10/47, 13/52, 18/51 とした結果は一致するか？
どの学生同士、どの商品同士が近いか？など、答えがすぐに見えない場合もある。

この分割表のデータを元に、行と列すべての項目について類似性を見る分析がコレスポンデンス分析である。類似性は、２つの分類変数にそれぞれ、特徴的な量 u_i^a と v_λ^α を考え、それらの量が最大の相関係数を持つようにして考える。類似度はこれらの量の値の近さで見る。2 次元分割表の２つの変数が入り混じることが特徴である。

コレスポンデンス分析結果

	群	第1成分	第2成分	重み1成分	重み2成分
▶ 固有値		0.0763	0.0183		
相関係数		0.2762	0.1352		
寄与率		0.8067	0.1933		
累積寄与率		0.8067	1.0000		
中学生	1	-1.3287	-0.6528	-0.3670	-0.0883
高校生	1	1.1333	-0.7748	0.3130	-0.1048
大学生	1	0.0690	1.3916	0.0190	0.1882
A	2	0.2373	1.5238	0.0655	0.2060
B	2	-1.4691	-0.6411	-0.4058	-0.0867
C	2	0.0596	-0.1102	0.0165	-0.0149
D	2	1.5032	-1.1547	0.4152	-0.1561

実行結果



散布図

問題

コレスポンデンス分析 2.txt は３つの地域で４つの商品の売れ筋を調べた結果である。コレスポンデンス分析を用いて以下の問いに答えよ。

１）地域と商品に関する２次元分割表を描け。（合計は不要）

	商品１	商品２	商品３	商品４
地域１				
地域２				
地域３				

２）２つの変数に付けたパラメータの相関係数、寄与率、累積寄与率を求めよ。

	第１成分	第２成分
相関係数		
寄与率		
累積寄与率		

３）２つの変数に付けたパラメータの値を求めよ。

	第１成分	第２成分
地域１		
地域２		
地域３		
商品１		
商品２		
商品３		
商品４		

４）散布図を見て地域で売れ筋の商品を選択せよ。（複数選択）

地域１　［商品１・商品２・商品３・商品４］

地域２　［商品１・商品２・商品３・商品４］

地域３　［商品１・商品２・商品３・商品４］

５）２次元分割表で見た場合、商品比率の最も高い商品はどれか。（単数選択）

地域１　［商品１・商品２・商品３・商品４］

地域２　［商品１・商品２・商品３・商品４］

地域３　［商品１・商品２・商品３・商品４］

以上から、コレスポンデンス分析の結果は、単純に比率だけから見たものとは異なる。

１２．時系列分析

時系列分析はあるデータの時間的変化を分析し、モデルを作成して今後の予測を行うことを目的とする。

変動の分解モデル

時系列データを傾向変動、季節変動、循環変動、残差に分解し、データの性質を調べると同時に予測も行う手法で、データに周期性がある場合に有効

例

Samples¥時系列分析 1.txt の売上１データを、傾向変動、季節変動、残差に分解し、来月の売上を予測せよ。

傾向変動　全体的な変化の傾向を表す変動
季節変動　一定の周期を持つ変動
循環変動　一定の周期ではない変動（ここでは考えない）
残差　　これらの変動を差し引いた残りの変動
時系列データを見る　→　「元データ」ラジオボックスを選択し、描画ボタン
傾向変動を分解する　→　傾向変動１，２で見て、よく適合するモデルを求める。
　変動の分解の表示で、元データ、傾向変動１、２、残差をチェックし、実行
周期性を見る　→　コレログラム（自己相関のグラフ）とピリオドグラムで調べる。
季節変動を分解する　→　振幅変動の種類を選択し、周期変動分解の周期を入力、これらにチェックを入れ、表示に季節変動を加え実行
再度季節変動を見る　→　再度周期性を調べ、必要なら周期変動分解の周期をカンマ区切りで追加し実行
どの程度予測があっているかの目安　→　残差標準偏差の値、R^2 値

手順

１）ここでは傾向変動の近似モデルは回帰モデルの１次式を選択する。
２）傾向変動と残差をグラフや残差標準偏差（予測の良さ）で確認する。
３）傾向変動を除いた残差から周期性をコレログラムやピリオドグラムで確認する。
４）傾向変動、季節変動、残差をグラフや残差標準偏差（予測の良さ）で確認する。
５）再度残差の周期性をコレログラムやピリオドグラムで確認する。
６）残差標準偏差の値を見て調整する。
７）データの実測値と予測値の R^2 の値を求める。（予測の良さを表す）
８）モデルの予測値を確認する。

問題１

Samples¥時系列分析 1.txt の売上 2 について、以下の問いに答えよ。但し、ここでは周期変動２と振幅変動については考えないものとする。

１）このデータの傾向変動を１次式で推定するとどのような式になるか。
売上＝［　　］×時間＋［　　］

２）上の傾向変動を除いた場合の残差標準偏差の値はいくらか。　［　　］

３）傾向変動を除いた残差から、ピリオドグラム等を用いて季節変動の周期を求めるといくらか。　［　　］

４）上の季節変動を除いた場合の残差標準偏差の値はいくらか。　［　　］

５）データを上の傾向変動と季節変動で予測するモデルの R^2 の値はいくらか。
［　　］

６）このモデルでの１期先の予測値はいくらか。　［　　］

７）このモデルでの 5 期先の予測値はいくらか。　［　　］

問題２

Samples¥時系列分析 1.txt の売上 4 について、以下の問いに答えよ。但し、ここでは周期変動２と振幅変動については考えないものとする。

１）このデータの傾向変動は、１次近似、対数近似、べき乗近似、指数近似、多項式近似（３次まで）のうちどれが最適か？そのときの残差標準偏差の値はいくらか。
近似は［　　］,　残差標準偏差［　　］

２）傾向変動を除いたデータから、ピリオドグラム等を用いて季節変動の周期を求めるといくらか。　［　　］

３）傾向変動と季節変動を除いた残差標準偏差はいくらか。　［　　］

４）傾向変動と季節変動を除いたデータから、ピリオドグラム等を用いて再度季節変動（長周期の周期変動）の周期を求めるといくらか。　［　　］

５）上の変動をすべて除いた残差標準偏差はいくらか。　［　　］

６）データを上の傾向変動、季節変動、循環変動で予測するモデルの R^2 の値はいくらか。
［　　］

７）このモデルでの１期先の予測値はいくらか。　［　　］

８）このモデルでの 10 期先の予測値はいくらか。　［　　］

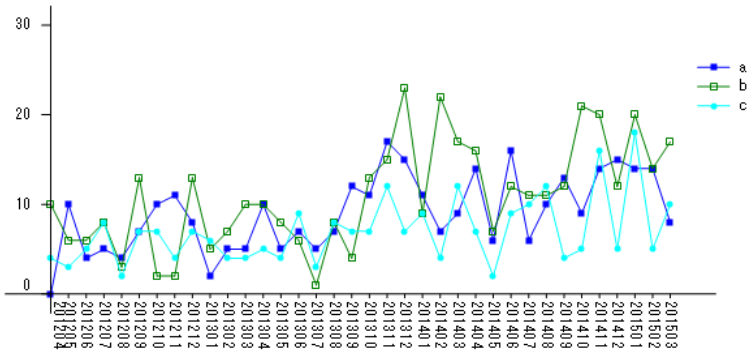
１３．パネル重回帰分析

パネル重回帰分析は複数変数の時系列データを用いて、ある１つの変数を予測する重回帰分析である。

問題

パネル重回帰分析 1.txt はある製造業の 3 つの製品の需要のデータである。パネル重回帰分析を用いて予測の可能性を以下に従って考えよ。

１）3 つの製品の時系列グラフを示せ。



以後は 3 期分のデータを使って、製品 a についての 1 期先の予測を考える。

２）これらの 3 期分のデータの中で、予測値との相関が一番大きいものはどれか。

製品［　　］の［　　］期前のデータで、相関係数［　　］

３）回帰式を求めよ。

予測値＝［　　］a1 期前＋［　　］a2 期前＋［　　］a3 期前＋…

＋［　　］c3 期前＋［　　］

４）最も影響力のある変数は何か。

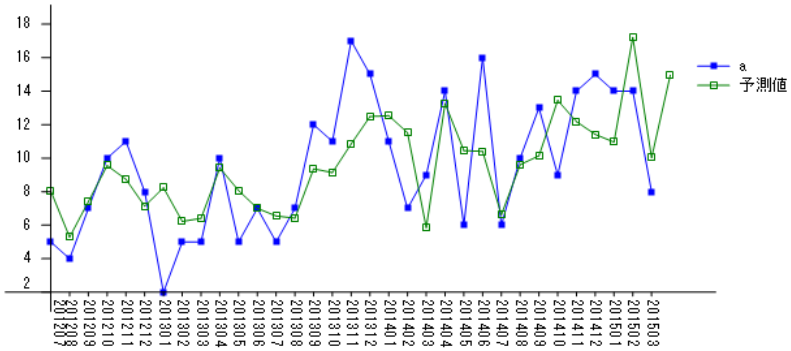
製品［　　］の［　　］期前のデータで、

標準化偏回帰係数の値［　　］、偏回帰係数の検定確率値［　　］

５）予測の寄与率はいくらか。［　　］

６）1 期先はいくらに予測されたか。［　　］

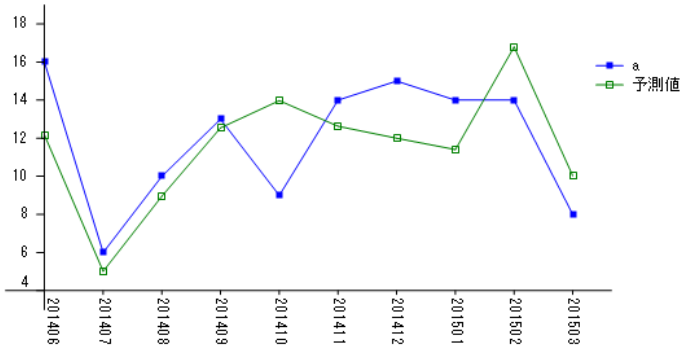
７）３）の回帰式を用いた実測値と予測値のグラフを描け。



８）これまでの予測は、予測式を求めるときに予測される数値が使われていた。これを防ぐために、予測値の値を求める際、それまでの値しか使わないようにした。そのようにした場合の予測の精度を見たい。2015 年 2 月の実測値と予測値を比較せよ。但し、交差検証には 10 期分のデータを使うことにする。

実測値	予測値	交差検証予測値

９）交差検証による実測、予測グラフを描け。



10）1 期先から 5 期先までの予測の寄与率（R^2）を交差検証の結果を元に比較せよ。

1 期先	2 期先	3 期先	4 期先	5 期先

11）4 ページ目のデータでの 2014 年 12 月の予測と 1 ページ目のデータでの交差検証の予測の値は同じか。

2014 年 12 月予測	同交差検証予測

［同じ・同じでない］

12）3 つの製品の中で過去のデータから最も影響があるものは何か。

［ 製品 a・製品 b・製品 c ］

どのような変数が過去から影響を受けやすいか、経済指標の影響はどうか、きちんと考えていくことで予測の精度は上がる。

経営科学編

１．品質管理

品質管理（Quality Control，QC）とは

広義の品質管理「品質要求事項を満たすことに焦点を合わせた品質マネジメントの一部」（JIS）

狭義の品質管理「品質保証行為の一部をなすもので、部品やシステムが決められた要求を満たしていることを、前もって確認するための行為」（JIS）

ウォルター・シューハート、エドワーズ・デミング、石川馨らによってはじめられ、1960年以降日本企業に広く普及した。

1.1　ＱＣ七つ道具

データを図にまとめたり、数値にまとめたりすることが重要

１．グラフ（折れ線グラフ、棒グラフ、円グラフ、帯グラフ、レーダーチャートなど）

２．ヒストグラム（山の形から工程の安定性、拡がりから規格からのずれなどをみる）

３．管理図（工程の安定性を見る）

４．チェックシート（確認要点事項を予め抜粋しまとめられたツール）

５．パレート図（工程改善用に問題点を原因別・損失別に並べた棒グラフ、）

６．特性要因図（問題抽出用に用いられる問題点を階層別に表した図）

７．散布図（相関）

８．層別（クロス集計）

グラフと管理図をひとまとめにして７つにすることが多い。

新ＱＣ七つ道具

連関図法，親和図法（KJ 法の別名），系統図法，アローダイアグラム法，マトリックス図法，マトリックスデータ解析法，PDPC 法

例

Samples¥品質管理 1.txt のデータを用いて品質管理の考え方を学ぶ。

１ページ目のデータは１日単位で求めた４つの工場の初期不良の個数である。

各工場の初期不良数を**「折れ線グラフ」**で見てみる。

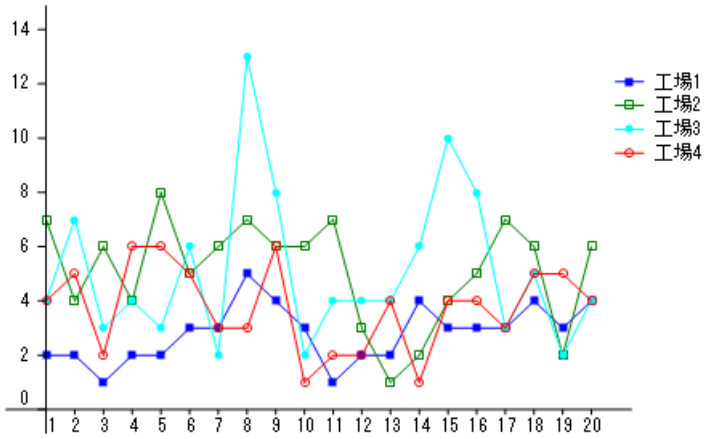


図 1　折れ線グラフ

これから工場 3 で変化が大きいと思われるが、工場 3 の重要性を調べてみる。

このために各工場の初期不良数の合計（群別データ合計から）を**「パレート図」**で見てみる。1 頁目のデータをそのまま利用する。

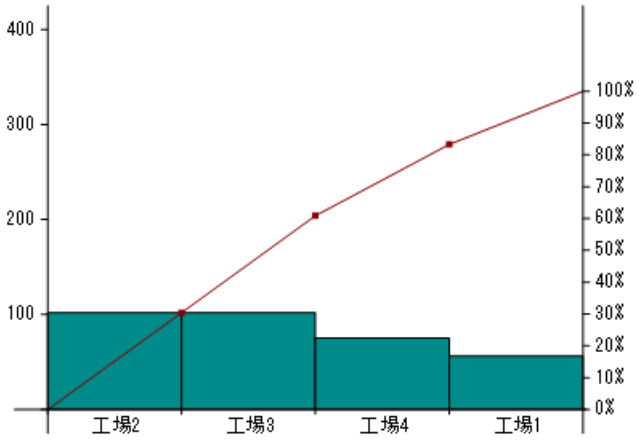


図 2　初期不良数で見たパレート図

これによると特に工場 3 の初期不良が多いとは考えにくい。しかし、工場の重要性は金額ベースでも見る必要がある。そこで、2 頁目の金額ベースの損失のデータでもパレート図を描いてみる。

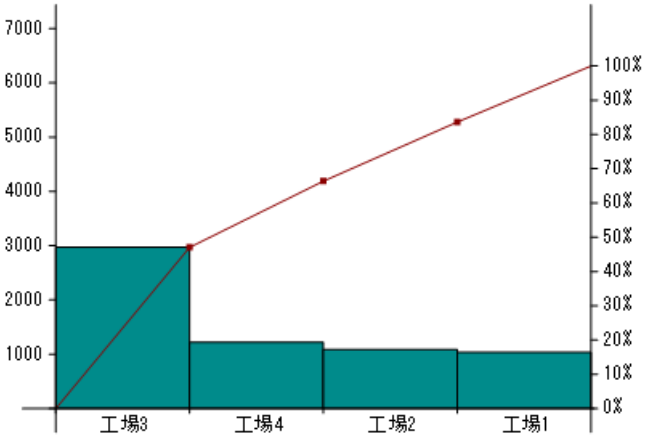


図 3　損失額で見たパレート図

件数で見た場合はあまり差が見られないが、金額ベースで見ると工場 3 の改善に取り組むことが重要であると分る。

工場 3 の初期不良数のデータ（1 頁目）は 1 つのデータであるので、**「x 管理図」**を用いて異常を調べる。

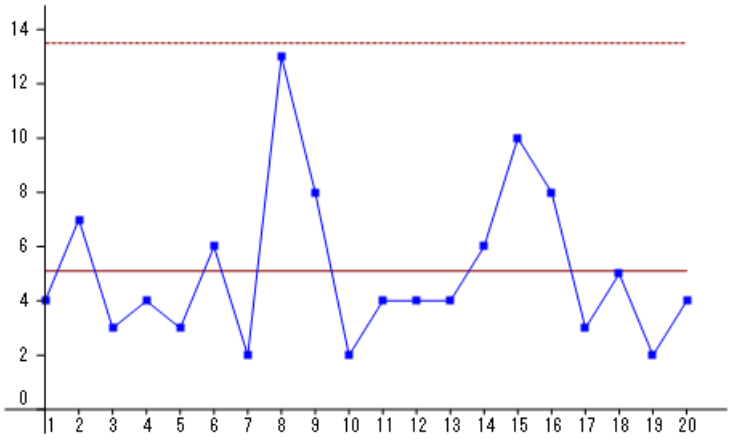


図 4　x 管理図で見たデータの異常

ここで中央の線を中心線（CL）、上の点線を管理限界線（UCL）という。下側に管理限界線（LCL）が付く場合も多い。この場合、1 回の測定でデータが 1 つであったが、複数個のデータを集める場合がある。そのときには$\bar{x}$管理図を用いる。またばらつきの異常を調べるにはR管理図が用いられる。

x 管理図で限界線近くまで拡がる場合があったので、**「ヒストグラム」**で分布の特性を見る。自然な誤差の場合には分布が正規分布に近いものになる。

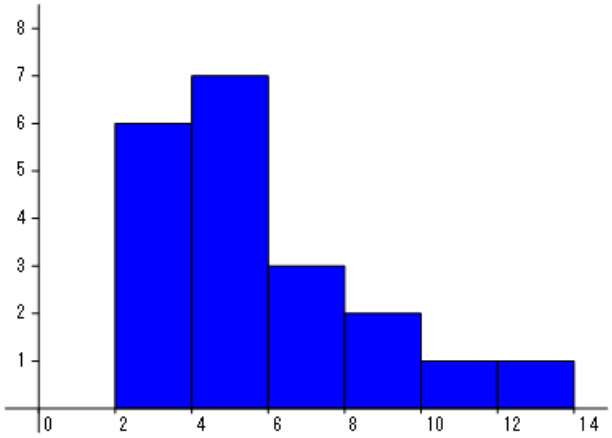


図 5 初期不良数の分布

これを見ると右方向に伸びており、正規分布とは異なる。これらのことから、異常が発生している可能性があるように思われる。

次に**「特性要因図」**によって原因の絞り込みを行いたい。

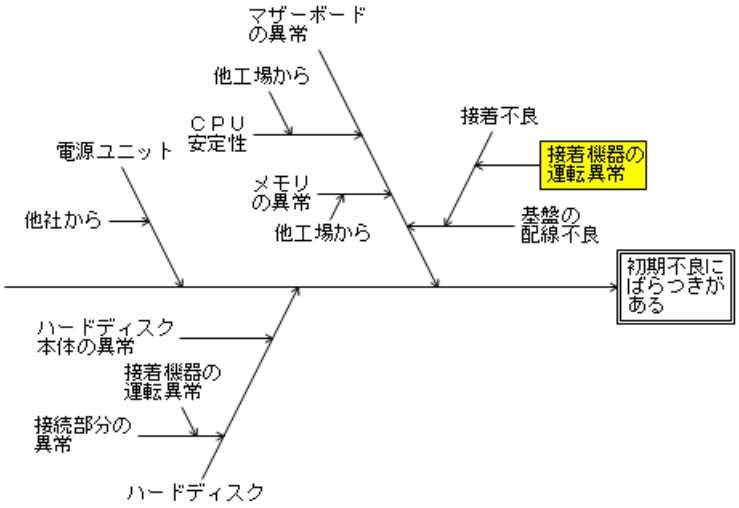


図 6 特性要因図による原因の絞り込み

初期不良にばらつきがある場合、上の特性要因図のような原因が考えられるが、今回はマザーボードの異常によるものが多かったため、その原因を考えて行くと、接着機器の運転異常が原因ではないかと疑われた。

不良品発生の状況を層別の手段で考えるために、不良品の発生に午前と午後の差はあるのか、機器による差はあるのかを**「層別」**の手法で調べてみたところ、以下の結果を得た。

不良品の発生に午前と午後の差はあるか。（層別）3 頁
　検定名　［　　］検定
　検定確率［　　］　差があると［いえる・いえない］
不良品の発生に機器による差はあるか。（層別）4 頁
　検定名　［　　］検定
　検定確率［　　］　差があると［いえる・いえない］

この結果から、機器の不良が考えられたが、異常は断続的に表れるので、何らかの外的な要因があるように思われた。異常発生と天候との関係に気付くものがおり、調べてみると以下の結果を得た。

不良品の発生に天候の影響はあるか。（層別）5 頁
　検定名　［　　］検定
　検定確率［　　］　差があると［いえる・いえない］

原因がかなり絞り込めたので、調べてみると雨漏りが原因で漏電が発生していることが分り、問題が解決された。

参考

Wikipedia
フリーソフトウェア R による統計的品質管理入門，荒木孝治，日科技連

1.2　不良品頻度に関する診断

問題

不良品発生頻度に以下の表データ（品質管理（不良率診断）.txt）が与えられた。

	過去個数	過去 Claim	今年個数	今年 Claim
A	1000	5	200	4
B	2000	7	300	5
C	3000	8	500	2

このデータを用いて以下の問いに答えよ。

1）今年の発生確率とその起きる確率である検定確率を求めよ。

	今年発生率	検定確率
A		
B		
C		

2）有意水準を 5%と 1%としたときどのように判断するか。

	5%	1%
A	正常・異常	正常・異常
B	正常・異常	正常・異常
C	正常・異常	正常・異常

1.3　抜き取り検査

問題

ロットからの抜き取り検査で、ロット内の製品数 N、ロット内製品の不良品率 p、取り出すサンプル数 n、不良品許容数 d、その際のロット合格率 Q がそれぞれ以下のように与えられるとき、問いに答えよ。

ロット内の製品数が十分多いとして考えない場合（2 項分布）

1）不良品率 0.05、サンプル数 20、不良品許容数 1 のとき、ロット合格率はいくらになるか。ロット合格率［　　］

2）ロット合格率 0.9、サンプル数 20、不良品許容数 1 のとき、ロット内製品の不良品率はいくらか。不良品率［　　］

3）ロット合格率 0.9、サンプル数 20 で、不良品率を 0.05 以下にしたいとき、不良品許容数をいくつ以下にすればよいか。
不良品許容数を［　　］以下にする。

4）ロット合格率 0.9、不良品許容数 1 で、不良品率を 0.05 以下にしたいとき、サンプル数をいくつ以上にすればよいか。
サンプル数を［　　］以上にする。

ロット内製品数を 100 にする場合（超幾何分布）

5）不良品率 0.10、サンプル数 10、不良品許容数 1 のとき、ロット合格率はいくらになるか。ロット合格率［　　］

6）ロット合格率 0.95、サンプル数 10、不良品許容数 1 のとき、ロット内製品の不良品率はいくらか。不良品率［　　］

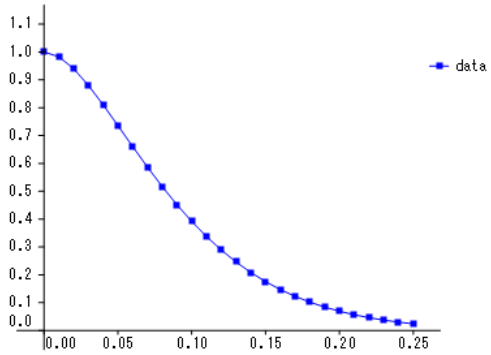
7）ロット合格率 0.95、サンプル数 10 で、不良品率を 0.1 以下にしたいとき、不良品許容数をいくつ以下にすればよいか。
不良品許容数を［　　］以下にする。

8）ロット合格率 0.95、不良品許容数 1 で、不良品率を 0.1 以下にしたいとき、サンプル数をいくつ以上にすればよいか。
サンプル数を［　　］以上にする。

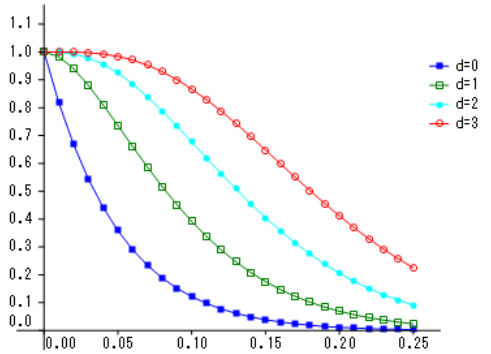
注）不良品許容数は通常合格判定個数と呼ばれます。不良品率は不適合率とも呼ばれます。

もう一度、ロット内の製品数が十分多いとする場合（2 項分布）

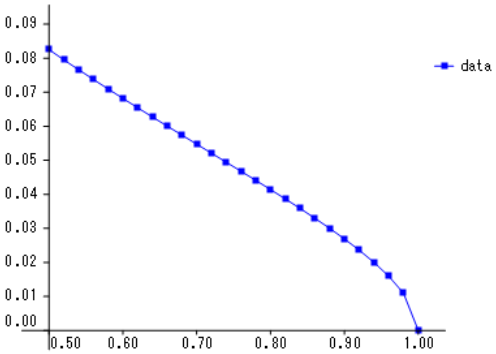
9）サンプル数 20、不良品許容数 1 として、不良品率を変化させたとき、ロット合格率はどのように変化するか。グラフで求めよ。



10）サンプル数 20、不良品許容数 0, 1, 2, 3 として、不良品率を変化させたとき、ロット合格率はどのように変化するか。グラフで求めよ。



11）サンプル数 20、不良品許容数 1 として、ロット合格率を変化させたとき、不良品率はどのように変化するか。グラフで求めよ。



２．パラメータ設計

装置からの出力は、人が制御できる制御因子と制御できない誤差因子に影響され、理想値からのずれが生じる。パラメータ設計とは、データの誤差因子の影響を抑え、理想値に近い観測値が得られるよう制御因子を調整する手法である。パラメータ設計には動特性と静特性がある。

動特性：入力に対して理想的な出力を定義する場合（例えば$y=\beta M$）

静特性：ある出力値を理想的な値（ある値、ゼロ、大きな値など）に近づける場合

用いる指標

SN 比η　　大きい値を取るほど良い。

感度S　　問題により、大きい方、目標値なった方、小さい方がよいなど様々

ここでは、動特性のパラメータ設計について考える。

2.1　パラメータ設計の理論と考え方

ここではまずゼロ点比例式の動特性パラメータ設計について考えるが、その前に SN 比と感度について、理論的な考察を加えておく。

1 つの実験では、信号水準M_j（$j=1,\cdots,p$）と誤差水準N_α（$\alpha=1,\cdots,n$）によって、表 1 のようにpn個のデータ$y_{j\alpha}$が得られる。誤差水準はできるだけ広く散らばるよう配慮されるものとする。

表 1　パラメータ設計におけるデータ

M_1			…	M_p		
N_1	…	N_n	…	N_1	…	N_n
y_{11}	…	y_{1n}	…	y_{p1}	…	y_{pn}

この実験についての、誤差水準αのゼロ比例式回帰直線を考える。実測値$y_{j\alpha}$についての推定回帰式を$Y_{j\alpha}=b_\alpha M_j$とすると、実測値との差の 2 乗和は以下となる。

$$EV_\alpha=\sum_{j=1}^{p}(y_{j\alpha}-Y_{j\alpha})^2=\sum_{j=1}^{p}(y_{j\alpha}-b_\alpha M_j)^2$$

これを最小とするには、

$$\frac{\partial}{\partial b_\alpha}EV_\alpha=-2\sum_{j=1}^{p}M_j(y_{j\alpha}-b_\alpha M_j)=0$$

として、以下を得る。

$$b_\alpha=\sum_{j=1}^{p}M_j y_{j\alpha}\Big/\sum_{j=1}^{p}M_j^2=\frac{L_\alpha}{r},\qquad \text{ここに、}\ L_\alpha=\sum_{j=1}^{p}M_j y_{j\alpha},\ \ r=\sum_{j=1}^{p}M_j^2$$

全体のゼロ比例式回帰直線については、推定回帰式を$Y_j=bM_j$とすると、実測値との差の

2 乗和は以下となる。

$$EV=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}(y_{j\alpha}-Y_j)^2=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}(y_{j\alpha}-bM_j)^2$$

これを最小とするには、

$$\frac{\partial}{\partial b}EV=-2\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_j(y_{j\alpha}-bM_j)=0$$

として、以下を得る。

$$b=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_jy_{j\alpha}\Big/\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_j^2=\frac{1}{nr}\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_jy_{j\alpha}=\frac{1}{nr}\sum_{\alpha=1}^{n}L_\alpha=\frac{1}{n}\sum_{\alpha=1}^{n}b_\alpha$$

次にデータの変動について考察する。まず、$y=0$からの全体の変動S_Tは以下となる。

$$S_T=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}y_{j\alpha}^2$$ 自由度 pn

また、$y=0$からの全体の回帰変動S_βは以下となる。

$$S_\beta=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}(bM_j)^2=nrb^2=\frac{1}{nr}\left(\sum_{\alpha=1}^{n}L_\alpha\right)^2$$ 自由度 1（b のみ）

これより、$b^2=S_\beta/nr$ となる。

誤差水準α の回帰直線の全体の回帰直線からの変動$S_{N\beta}$ は以下となる。

$$S_{N\beta}=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}(b_\alpha M_j-bM_j)^2=r\sum_{\alpha=1}^{n}(b_\alpha-b)^2=r\sum_{\alpha=1}^{n}b_\alpha^2-nrb^2=\sum_{\alpha=1}^{n}L_\alpha^2\Big/r-S_\beta$$

自由度 $n-1$　束縛条件　$\sum_{\alpha=1}^{n}(b_\alpha-b)=0$　1 個

各点の誤差水準α の回帰直線からの変動S_eは以下となる。

$$S_e=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}(y_{j\alpha}-b_\alpha M_j)^2=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}[y_{j\alpha}-(b_\alpha-b)M_j-bM_j]^2$$

$$=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}y_{j\alpha}^2+\sum_{j=1}^{p}\sum_{\alpha=1}^{n}(b_\beta-b)^2M_j^2+\sum_{j=1}^{p}\sum_{\alpha=1}^{n}b^2M_j^2-2\sum_{j=1}^{p}\sum_{\alpha=1}^{n}y_{j\alpha}b_\alpha M_j$$

$$=S_T+S_{NB}+S_\beta-2\sum_{\alpha=1}^{n}L_\alpha^2\Big/r^2=S_T+S_{NB}+S_\beta-2(S_{NB}+S_\beta)=S_T-S_{NB}-S_\beta$$

自由度 $pn-n$　束縛条件　$\sum_{j=1}^{p}M_j(y_{j\alpha}-b_\alpha M_j)=0$　n 個

各点の誤差水準α の回帰直線からの不偏分散V_eは、変動を自由度で割って以下となる。

$$V_e=S_e/(np-n)$$

これが不偏分散となることは補遺で詳しく説明する。

ここまでの議論で、全変動S_Tは、全体の回帰変動S_β、全体の回帰直線からの誤差水準α の回帰変動$S_{N\beta}$、各点の誤差水準α の回帰直線からの変動S_eの和で以下のように表される

ことが分かった。

$$S_T = S_\beta + S_{N\beta} + S_e$$

各点の全体の回帰直線からの変動S_Nは以下となる。

$$S_N = \sum_{j=1}^{p}\sum_{\alpha=1}^{n}(y_{j\alpha} - bM_j)^2 = \sum_{j=1}^{p}\sum_{\alpha=1}^{n} y_{j\alpha}^2 + nb^2\sum_{j=1}^{p} M_j^2 - 2b\sum_{j=1}^{p}\sum_{\alpha=1}^{n} y_{j\alpha}M_j$$

$$= S_T + nrb^2 - 2nrb^2 = S_T - S_\beta = S_{N\beta} + S_e$$

自由度 $pn-1$　束縛条件　$\displaystyle\sum_{j=1}^{p}\sum_{\alpha=1}^{n} M_j(y_{j\alpha} - bM_j) = 0$

各点の全体の回帰直線からの不偏分散V_Nは、変動を自由度で割って以下となる。

$$V_N = S_N/(pn-1)$$

これらを使って、SN 比η　と感度Sを定義する。SN 比は、測定誤差の分散σ^2に対する有効な信号の変化の大きさβ^2の比を用いて、また感度Sはβ^2の値を用いて以下のように定義される。

SN 比：$\eta = 10\log_{10}\dfrac{\beta^2}{\sigma^2}$,　　感度：$S = 10\log_{10}\beta^2$

実際の計算ではβ^2とσ^2の値は不明であるので、これらの不偏推定量を用いて置き換える。

SN 比：$\eta = 10\log_{10}\left[\dfrac{(S_\beta - V_e)/nr}{V_N}\right]$,　　感度：$S = 10\log_{10}\left[(S_\beta - V_e)/nr\right]$

一般に SN 比は大きな値ほど、有効な信号を誤差の中から拾いやすくなり、良好な結果である。また、感度は対象により、大きな値がよい場合、小さな値がよい場合、目標値がよい場合など様々であるが、感度があまり変化しない制御因子を用いて SN 比を上げることを考えることもある。

最後に、β^2の不偏推定量を求めておく。

$$y_{j\alpha} = \beta_\alpha M_j + \varepsilon_{j\alpha},\quad E[\varepsilon_{j\alpha}] = 0,\quad E[\varepsilon_{j\alpha}\varepsilon_{j'\alpha'}] = \delta_{jj'}\delta_{\alpha\alpha'}V[\varepsilon]$$

とすると、

$$b_\alpha = \frac{1}{r}\sum_{j=1}^{p} M_j y_{j\alpha} = \frac{1}{r}\sum_{j=1}^{p} M_j(\beta_\alpha M_j + \varepsilon_{j\alpha}) = \beta_\alpha + \frac{1}{r}\sum_{j=1}^{p} M_j\varepsilon_{j\alpha}$$

より

$$S_e=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}(\beta_\alpha M_j+\varepsilon_{j\alpha}-b_\alpha M_j)^2=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\left[(\beta_\alpha-b_\alpha)M_j+\varepsilon_{j\alpha}\right]^2$$

$$=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\left[-\frac{1}{r}M_j\sum_{j'=1}^{p}M_{j'}\varepsilon_{j'\alpha}+\varepsilon_{j\alpha}\right]^2$$

$$=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\left[\frac{1}{r^2}M_j^2\left(\sum_{j'=1}^{p}M_{j'}\varepsilon_{j'\alpha}\right)^2-\frac{2}{r}M_j\varepsilon_{j\alpha}\sum_{j'=1}^{p}M_{j'}\varepsilon_{j'\alpha}+\varepsilon_{j\alpha}^2\right]$$

$$=-\frac{1}{r}\sum_{\alpha=1}^{n}\sum_{j=1}^{p}\sum_{j'=1}^{p}M_jM_{j'}\varepsilon_{j\alpha}\varepsilon_{j'\alpha}+\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\varepsilon_{j\alpha}^2$$

となり、

$$E\left[S_e\right]=E\left[-\frac{1}{r}\sum_{\alpha=1}^{n}\sum_{j=1}^{p}\sum_{j'=1}^{p}M_jM_{j'}\varepsilon_{j\alpha}\varepsilon_{j'\alpha}+\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\varepsilon_{j\alpha}^2\right]=(np-n)V(\varepsilon)\qquad\text{(A1)}$$

また、

$$b=\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_jy_{j\alpha}\Big/\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_j^2=\frac{1}{nr}\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_jy_{j\alpha}$$

より、

$$b^2=\frac{1}{n^2r^2}\left(\sum_{j=1}^{p}\sum_{\alpha=1}^{n}M_j(\beta_\alpha\mathrm{M}_j+\varepsilon_{j\alpha})\right)^2$$

$$=\frac{1}{n^2r^2}\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\sum_{j'=1}^{p}\sum_{\alpha'=1}^{n}M_jM_{j'}(\beta_\alpha\beta_{\alpha'}M_jM_{j'}+2\beta_{\alpha'}\varepsilon_{j\alpha}M_{j'}+\varepsilon_{j\alpha}\varepsilon_{j'\alpha'})$$

$$=\frac{1}{n^2r^2}\left[\beta^2n^2r^2+2\beta nr\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\varepsilon_{j\alpha}+\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\sum_{j'=1}^{p}\sum_{\alpha'=1}^{n}M_jM_{j'}\varepsilon_{j\alpha}\varepsilon_{j'\alpha'}\right]$$

であるから、

$$E\left[b^2\right]=\frac{1}{n^2r^2}E\left[\beta^2n^2r^2+2\beta nr\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\varepsilon_{j\alpha}+\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\sum_{j'=1}^{p}\sum_{\alpha'=1}^{n}M_jM_{j'}\varepsilon_{j\alpha}\varepsilon_{j'\alpha'}\right]$$

$$=\beta^2+\frac{1}{n^2r^2}\sum_{j=1}^{p}\sum_{\alpha=1}^{n}\sum_{j'=1}^{p}\sum_{\alpha'=1}^{n}M_jM_{j'}\delta_{jj'}\delta_{\alpha\alpha'}V(\varepsilon)$$

$$=\beta^2+\frac{1}{nr}V(\varepsilon)$$

よって、以下となる。

$$\beta^2=E[b^2]-\frac{1}{nr}V(\varepsilon)\qquad\text{(A2)}$$

(A2)と(A1)、及び $b^2=S_\beta/nr$ の関係から、

$$\beta^2=\frac{1}{nr}E\left[S_\beta\right]-\frac{1}{nr}E\left[S_e/(np-n)\right]=E\left[(S_\beta-V_e)/nr\right]\qquad\text{(A3)}$$

すなわち、β^2 の不偏推定量は $(S_\beta-V_e)/nr$ である。

同様の考え方でσ^2の不偏推定量が$V_N = S_N/(pn-1)$であることも示すことができる。

次に我々はSN比を最大にする制御因子の最適設定について考える。制御因子A, B, …について直交表を作ると、他の制御因子の影響をならした、1つの制御因子の影響を調べることができるようになる。表2に直交表を加えたデータを示す。

表2　パラメータ設計におけるデータ

	A	B	…	M_1			…	M_p			SN比	感度
				N_1	…	N_n	…	N_1	…	Nn		
1	1	1	…	y_{111}	…	y_{11n}	…	y_{1p1}	…	y_{1pn}	η_1	S_1
2	1	1	…	y_{211}	…	y_{21n}	…	y_{2p1}	…	y_{2pn}	η_2	S_2
:	:	:	:	:	:	:	:	:	:	:	:	:
D	2	2	…	y_{d11}	…	y_{d1n}	…	y_{dp1}	…	y_{dpn}	η_d	S_d

ここにSN比と感度は上で述べた方法で求めて加えてあるものとする。直交表は、各制御因子の同じ番号の行を見ると、他の制御因子について、すべての番号が同じ数だけ入っているという特徴を持つ。

例えば制御因子Aがkになる行について、SN比及び感度の平均を取ったものをそれぞれ$\overline{\eta}_{A=k}$、$\overline{S}_{A=k}$と書くとすると、SN比の補助表は表3のようになる。感度の補助表も同様である。

表3　SN比の補助表

制御因子	水準1	…	水準r
A	$\overline{\eta}_{A=1}$	…	$\overline{\eta}_{A=r}$
B	$\overline{\eta}_{B=1}$	:	$\overline{\eta}_{B=r}$
:	:	…	:

ここに水準の少ない制御因子の場合、その部分は空欄にしておく。

この補助表のSN比の中で、制御因子ごとの水準値の最も大きな水準を並べたものを最適条件といい、例えばA1B2C1D3…などと表す。我々のプログラムでは制御因子名は省略して番号だけで表している。この最適な水準のSN比を合計したものをSN比の最適値という。感度についてもSN比の最適条件を用いて最適値を定義する。

これに対して現実の制御因子の設定を比較条件または現状条件という。この条件を用いてSN比を合計したものをSN比の比較値または現状値という。感度についても同様である。最適値と比較値の差は、今後の改善の可能性として検討すべき値である。

ここで述べた水準値は理論的な推測値である。この値が妥当なものかどうか、追実験をして検証しておかなければならない。また、現実的に考えて最適な制御条件が最良のものであるとは限らない。その際は、できるだけSN比の値を落とさず、感度で制御因子の調整を行うこともある。

2.2　プログラムの利用法

パラメータ設計のデータ（パラメータ設計 1.txt）は図 1 のように、左の直交表の部分と右の実験結果の部分に分けられる。

図 1　パラメータ設計のデータ

ここでは、制御因子が A～H の 8 種類、信号因子が 3 種類、誤差因子が 2 種類である。信号因子と誤差因子の部分の変数名には信号因子の数値が与えられている。

メニュー［分析－OR－パラメータ設計］を選択すると図 2 のような分析メニューが表示される。

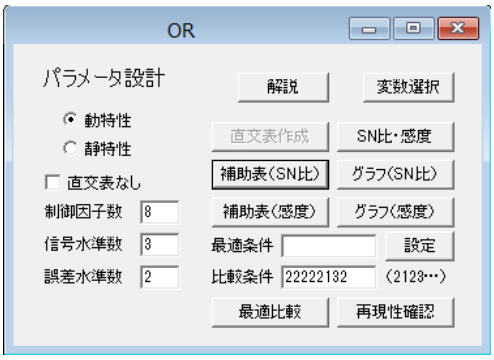


図 2　パラメータ設計分析メニュー

パラメータ設定には動特性と静特性の 2 種類あるが、今回は動特性のみについて紹介する。まずメニュー中にある、「制御因子数」、「信号水準数」、「誤差水準数」の値を入力する。この例題の場合、デフォルトの数値がそのまま利用できる。次に「変数選択」ボタンですべての変数を選択する。現在の例では直交表が付いているが、単純に SN 比と感度のみを求める場合には、直交表を省略したデータを用いることもできる。その際には「直交表なし」チェックボックスにチェックを入れておく。

「SN 比・感度」ボタンをクリックすると、図 3 の計算結果が表示される。

図 3　各実験の SN 比・感度

ここでは各実験に対して、単純に SN 比と感度を求めて表示している。

直交表を使った SN 比の補助表は「補助表（SN 比)」ボタンをクリックすることで図 4 のように与えられる。感度の補助表については「補助表（感度)」ボタンをクリックして得られる。

補助表（SN比）

	水準1	水準2	水準3	MAX-MIN
A	27.494	23.169		4.325
B	23.656	22.961	29.378	6.418
C	28.332	25.959	21.704	6.628
D	22.358	26.353	27.284	4.926
E	26.649	24.186	25.160	2.463
F	25.040	27.077	23.878	3.199
G	26.675	24.008	25.313	2.667
H	23.489	26.686	25.820	3.197

図 4　補助表（SN 比）

ここで制御因子 A は 2 水準であるから、空白が 1 つできている。

補助表をグラフにした図は「グラフ（SN 比)」ボタンをクリックして表示される。描画結果を図 5 に示す。

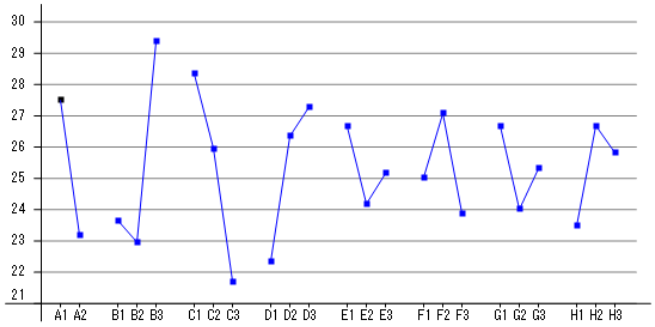


図 5　補助表のグラフ（SN 比）

図 4 と図 5 に対する感度の補助表とグラフは、それぞれ図 6 と図 7 で与えられる。

補助表（感度）

	水準1	水準2	水準3	MAX-MIN
A	-4.515	-3.975		0.540
B	-3.458	-5.041	-4.236	1.583
C	-4.337	-3.729	-4.669	0.940
D	-6.000	-4.117	-2.618	3.381
E	-3.963	-4.762	-4.011	0.799
F	-3.309	-4.239	-5.187	1.878
G	-4.045	-4.400	-4.290	0.355
H	-5.052	-3.968	-3.716	1.336

図 6　補助表（感度）

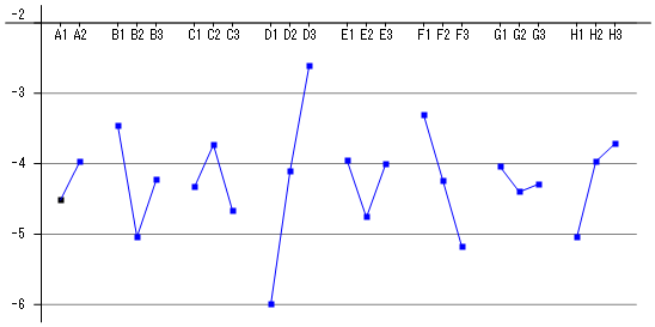


図 7　補助表のグラフ（感度）

SN 比の補助表やグラフを使った最適条件は「設定」ボタンをクリックすることでメニュー上の最適条件の部分に図 8 のように表示される。

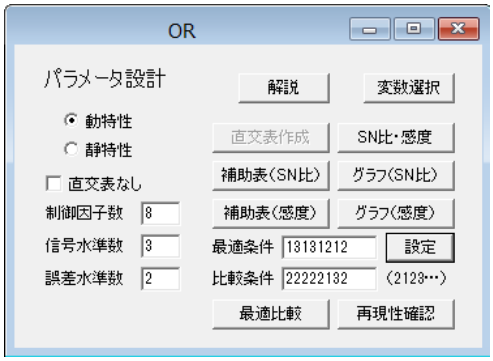


図 8　SN 比の最適条件の設定

比較条件で現在の実験から得られるデータの値を求めることができるが、最適条件との比較も可能である。これらの数値は「最適比較」ボタンで得ることができる。表示結果を図 9 に示す。最適条件の制御因子の組み合わせを変えることで結果を手動で訂正することもできる。

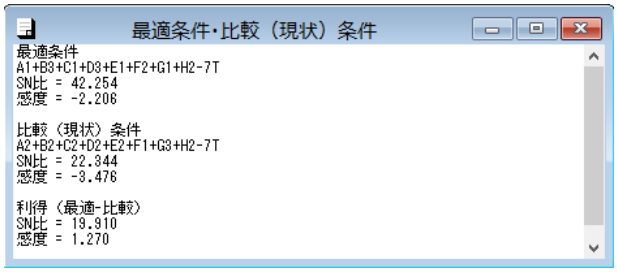


図 9　最適条件と比較条件

これらの最適条件と比較条件を実験で再現し、結果を得て、それをデータに追加する。当然その部分の直交表は空白になっているが、そのまま「再現性確認」ボタンをクリックすると図 10 に示す再現性確認表が得られる。

	SN比推定値	感度推定値	SN比実験値	感度実験値
▶ 最適条件	42.254	-2.206	44.369	-0.962
比較条件	22.344	-3.476	26.177	-4.059
利得	19.910	1.270	18.192	3.097

図 10　再現性確認表

数学編

１．グラフ

問題１　１変数関数グラフ

$-3 \le x \le 3$ の領域で以下のグラフを描き、x 切片、極値、変曲点を求めよ。また $x=1$ での接線を求めよ。

１） $y = x^2 + x - 3$

２） $y = x \sin x$

３） $y = \dfrac{e^x - e^{-x}}{2} = \sinh x$

問題２　２変数関数グラフ

$-3 \le x \le 3$，$-3 \le y \le 3$ の領域で以下のグラフを描き、極値を求めよ。また $x=1, y=1$ での接平面を求めよ。

１） $z = \sin x + \cos y$

２） $z = \dfrac{-1}{\sqrt{x^2 + y^2}}$

３） $z = x \sin(x + y)$

問題３　2 次元パラメータ表示関数

１） $-3 \le x \le 3$，$-3 \le y \le 3$ の領域において、$0 \le u \le 2\pi$ の範囲で以下の関数を描け。

$$\begin{cases} x = 3\cos u \\ y = 2\sin u \end{cases}$$

２）このグラフの描画過程をアニメーションせよ。

３） $u=1$ における接線を求めよ。

４）上のグラフは、$\dfrac{x^2}{9} + \dfrac{y^2}{4} - 1 = 0$ とも表されるが、陰関数表示を用いて描画せよ。

問題４　3 次元パラメータ表示関数

１）サンプルの球を描画せよ。

２）簡易動画を試せ。

３）他のサンプルも試せ。

２．方程式ソルバー

問題１　以下の方程式の解を実数の範囲で求めよ。

１）$x^2-2x-5=0$

２）$\begin{cases} x+y-3=0 \\ 3x-2y-7=0 \end{cases}$

３）$\begin{cases} \sin(x+y)+x-1=0 \\ y-x^2-1=0 \end{cases}$

問題２　以下の方程式の解を複素数の範囲で求めよ。

１）$x^2+x+1=0$

２）$\begin{cases} x+iy^2-1=0 \\ y-x^2-i=0 \end{cases}$

３．行列計算

問題１

行列 $\mathbf{A}, \mathbf{B}$ が以下のように与えられているとき、次の値を求めよ。

$$\mathbf{A}=\begin{pmatrix}1 & 1\\ 3 & 5\end{pmatrix},\quad \mathbf{B}=\begin{pmatrix}3 & 0\\ -2 & 1\end{pmatrix},\quad \mathbf{C}=\begin{pmatrix}1 & 1 & 2\\ 1 & 1 & -2\\ 2 & -2 & 3\end{pmatrix}$$

１）${}^t\mathbf{A}$

２）$\mathbf{A}+2\mathbf{B}$

３）$\mathbf{BA}$

４）$\mathrm{tr}\,\mathbf{C}$

５）$|\mathbf{C}|$

６）$\mathbf{C}^{-1}$

７）$\mathbf{C}$ の固有値

８）$\mathbf{C}$ の固有ベクトル（固有値大きい順）

問題２

以下の連立方程式を行列表示し、その解を求めよ。

１）
$$\begin{aligned} x_1+2x_2-x_3&=1\\ x_1+x_2+2x_3&=2\\ x_1-x_2&=1\end{aligned}$$

$$\begin{pmatrix}\quad & \quad & \quad\\ \quad & \quad & \quad\\ \quad & \quad & \quad\end{pmatrix}\begin{pmatrix}\quad\\ \quad\\ \quad\end{pmatrix}=\begin{pmatrix}\quad\\ \quad\\ \quad\end{pmatrix}$$

$$\mathbf{Ax}=\mathbf{b}$$

２）解 $\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix}=\begin{pmatrix}\quad\end{pmatrix}$

$$\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$$

４．定積分

問題１　直交座標での積分

以下の定積分を実行せよ。また、領域のグラフを描け。

１）$\int_0^{\pi}(\sin x+1)dx$

２）上の非積分関数の曲線の長さ

３）上の非積分関数を x 軸周りに回転させた図形の体積

４）$\int_0^{\pi}\int_0^{\pi}(\sin x+\cos y+1)dxdy$

５）上の非積分関数の面積

６）$\int_0^{\pi}\int_0^{y}(\sin x+\cos y+1)dxdy$

問題２　2 次元パラメータ表示積分

１）$\begin{cases}\cos u\\ \sin u\end{cases}\quad 0\le u\le\pi$　の原点と曲線を結ぶ面積

２）上の関数の曲線の長さ

３）x 軸周りに回転させた図形の体積

現在、数学についての機能を拡張しています。

参考文献

College Analysis のプログラム及びこの資料を作るに当たり、以下の本の理論とデータを参考にさせていただきました。著者の方々に心より感謝致します。

1) 宮沢健一編, 産業連関分析入門, 日本経済新聞社, 1991.
2) 椹木義一他編, 参加型システムズ・アプローチ, 日刊工業新聞社, 1981.
3) 藤田恒夫, 原田雅顕, 決定分析入門, 共立出版, 1989.
4) 利根薫, ゲーム感覚意思決定法－ＡＨＰ入門－, 日科技連出版, 1986.
5) 三根久, オペレーションズ・リサーチ上下, 朝倉書店, 1974, 1976.
6) 丹後俊郎, 古川俊之監修, 医学への統計学, 朝倉書店, 1983.
7) 篠崎壽夫・松下祐輔編, 工学のための応用数値算法入門（上下）, コロナ社, 1986.
8) 脇本和昌・垂水共之・田中豊編, パソコン統計解析ハンドブックⅠ基礎統計編, 共立出版, 1984.
9) 森雅夫, 森戸晋, 鈴木久敏, 山本芳嗣, オペレーションズリサーチⅠ,Ⅱ, 朝倉書店, 1991.
10) 柳川堯, 新統計学シリーズ９ ノンパラメトリック法, 培風館, 1982.
11) 白旗慎吾編, パソコン統計解析ハンドブックⅣ ノンパラメトリック編, 共立出版, 1987.
12) 刀根薫, 経営効率性の測定と改善 －包絡分析法 DEA によるー, 日科技連出版社, 1993.
13) 丹後俊郎, 新版医学への統計学, 朝倉書店, 1993.
14) 河口至商, 多変量解析入門Ⅰ,Ⅱ, 森北出版, 1978.
15) 田中豊・垂水共之編, Windows 版 統計解析ハンドブック 多変量解析, 共立出版社, 1995.
16) 田中豊・脇本和昌, 多変量統計解析法, 現代数学社, 1983.
17) 木下栄蔵, わかりやすい意思決定論入門, 近代科学社, 1996.
18) 高橋玲子他, 上田太一郎監修, Excel で学ぶ時系列分析と予測, オーム社, 2006.
19) 北川源四郎, 時系列解析入門, 岩波書店, 2005.
20) 豊田秀樹, 共分散構造分析［入門編］－構造方程式モデリングー, 朝倉書店, 1998.
21) 小塩真司, はじめての共分散構造分析 Amos によるパス解析, 東京図書, 2008.
22) 高橋信, Excel で学ぶコレスポンデンス分析, オーム社, 2005.
23) 荒木孝治編著, フリーソフトウェア R による統計的品質管理入門, 日科技連, 2005.
24) 勝呂隆男, 適正在庫の考え方・求め方, 日刊工業新聞社, 2003.
25) 白井豊, フラクタルで描く魅惑的な画像の世界, ゆたか創造舎, 2009.
26) 臼田昭司他, カオスとフラクタル―Excel で体験, オーム社, 1999.
27) 山口昌哉, カオスとフラクタル―非線形の不思議（ブルーバックス), 講談社, 1996.

28) 芹沢浩, 複素数とフラクタル, 東京図書, 1995.
29) 森典彦他, ラフ集合と感性, 海文堂出版, 2004.
30) 福澤英弘・小川康, 不確実性分析実践講座, ファーストプレス, 2009.
31) 土金達男, シミュレーションによるシステムダイナミックス入門, 東京電機大学出版局, 2005.
32) 高橋大輔, 理工系の基礎数学８ 数値計算, 岩波書店, 1996.
33) 川村清, 岩波基礎物理シリーズ３ 電磁気学, 岩波書店, 1994.
34) 岡崎進・吉井範行, コンピュータ・シミュレーションの基礎［第 2 版］, 化学同人, 2011.
35) Leslie F. Greengard, The Rapid Evaluation of Potential Fields in Particle Systems, MIT Press.
36) 豊田秀樹, マルコフ連鎖モンテカルロ法（統計ライブラリ）, 朝倉書店, 2008.
37) 四辻哲章, 計算機シミュレーションのための確率分布乱数生成法, プレアデス出版, 2010.
38) 永田靖, 棟近雅彦, 多変量解析法入門, サイエンス社, 2001.
39) 入門 パラメータ設計, 中野惠司他, 日科技連出版社, 2008.

注）このプリントの解答について

これまでは解答の数字の桁数を出力されるすべての桁数で書いていましたが、以後は原則として以下のように表示することに致しました。

平均や標準偏差については、データの桁数から 1 桁増やす。

検定確率は利用し易さを考えて、小数点以下 4 桁にする。

相関係数はよく使われるように、小数点以下 3 桁にする。

回帰係数などは有効桁数より 1 つ増やす。

予測値はデータの桁数から 1 桁増やす。