実用統計

基本統計編

福井正康

はじめに

経営統計学基礎の講義では、データの集計方法と推測統計について少し理論に踏み込んで勉強しました。その際計算は、厄介でもすべて Excel を使い、何を計算しているのか分かるようにしました。ところが、統計処理には専用ソフトがあり、これらを使えば苦労した t 検定など何も考えず一発解答です。これらの統計ソフトは昔から多くの人に利用されてきました。統計処理をよく理解している人にとってこれらは必需品ですが、初心者にとってブラックボックス的に処理を進めることには、思わぬ落し穴が待ちうけています。例えば、1, 2, 3 で与えられるカテゴリデータどうしで t 検定をしても一応結果は出ます。何をやっているのか理解せずに統計ソフトを利用すると、取り返しのつかない過ちを犯すことにもなりかねません。

さて、我々は 1 年間統計を勉強してきましたので、Excel を使った練習もそろそろ卒業です。何を計算するのかもある程度頭に入ったでしょうし、この講義の後半で学ぶ多変量解析は、計算が複雑で Excel の基本機能ではもはや限界です。また Excel の分析ツールの中にも、分散分析の一部と重回帰分析位しか含まれておりません。いよいよ統計ソフトを利用するときがきました。

統計ソフトにはいろいろなものがあります。SPSS, SAS, S-PLUS, R のように世界的に評価されているものや比較的使い易い STATISTICA 等、数多くのものが開発されています。これらの単独ソフトの他にも Excel の機能を利用するために VBA で記述されたマクロ的なソフトもあります。どれを利用するかは個人の好みでしょうが、一般に上中級者用のものは非常に高価で、初心者用のものでもある程度費用がかかります。またフリーのものでも、R は文系の学生にはちょっと難しいし、他のソフトはインターフェースがもうひとつという感じがします。

そこで我々は、学生に自由に使ってもらうために、分かり易い初心者向けの統計ソフトを開発することにしました。せっかくですからその当時開発中だった OR 関係の分析ソフトに統合させ、できたものが「College Analysis」です。「分析」という大げさな名前ですので、今後より多くの分析手法を加えて充実させていかなければなりません。これはインターネット上で公開していますので、いつでも最新のものを自由に利用することができます。経営統計学基礎で学んだ例題をもう一度このソフトでやり直してみて下さい。全体的な視野が広がると確信しています。

福山平成大学　福井正康

１章　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

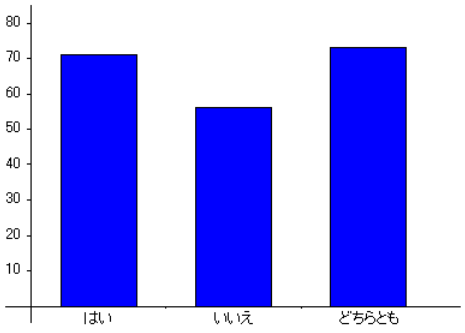
4）地域と意見１に関する２次元分割表を描け。

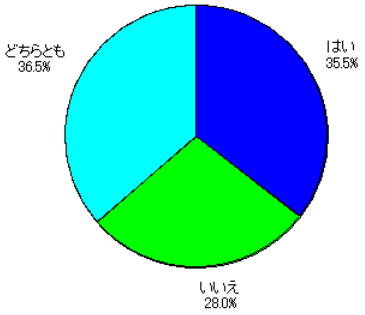
	賛成	反対	合計
市街			
郊外			
合計			

5）地域と意見２に関する２次元分割表を描け。

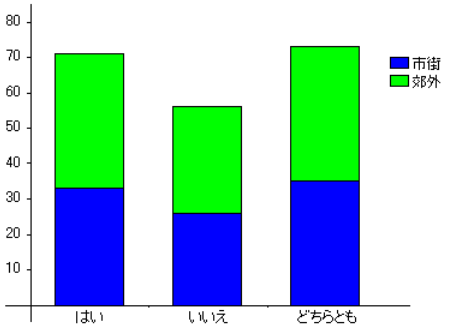
	はい	いいえ	どちらとも	合計
市街				
郊外				
合計				

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。



1.2 量的データの集計

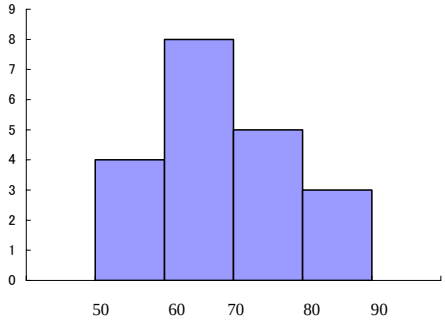
1.2.1 単純集計

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。

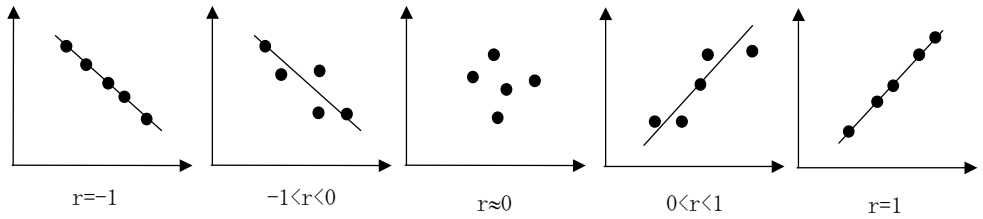
5）体重について 10kg 毎のヒストグラムを描け。

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。

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］

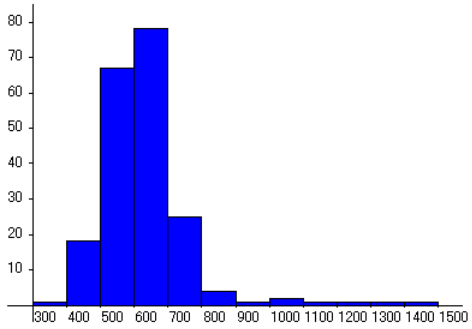
市街と郊外ではどちらの年収の拡がりが大きいか［市街・郊外］

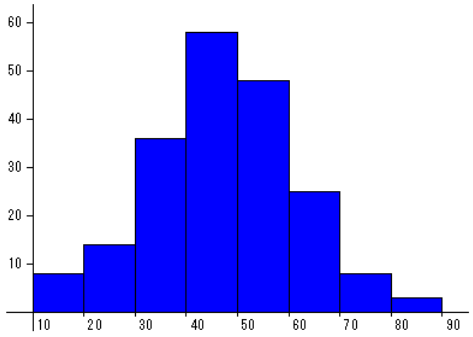
3）年収に関するヒストグラムを描け。（下図左）

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［　　］

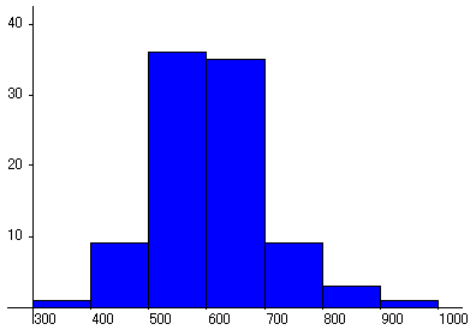
4）支出に関するヒストグラムを描け。（下図右）

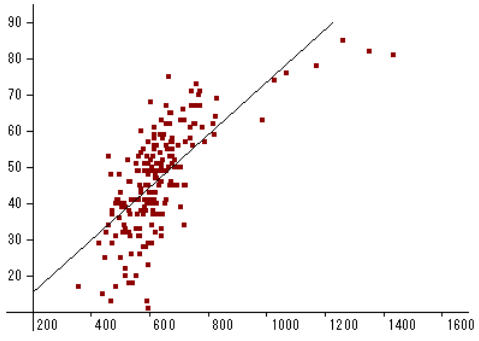




5）地域:1 の年収に関するヒストグラムを描け。（下図左）

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＋［　　］

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				

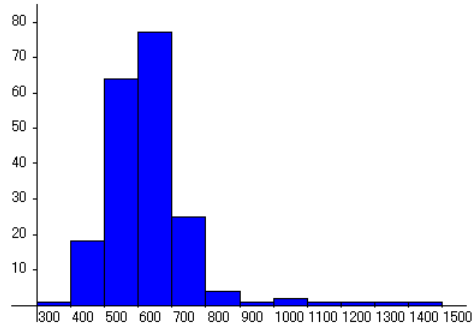
３）年収と支出に関する以下の基本統計量を求めよ。　［　　］

	最小値	最大値	平均値	中央値	標準偏差
年収					
支出					

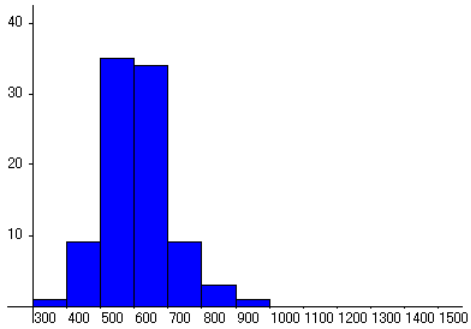
４）地域別の年収に関する基本統計量を求めよ。　［　　］

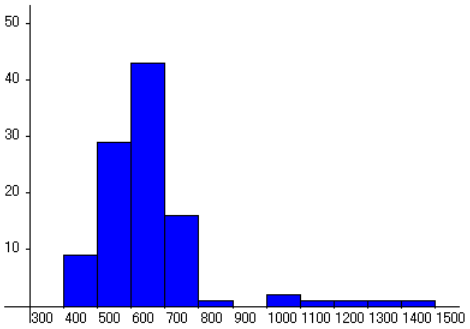
	最小値	最大値	平均値	中央値	標準偏差
地域:1					
地域:2					

5）年収に関するヒストグラムを描け。　　［　　］

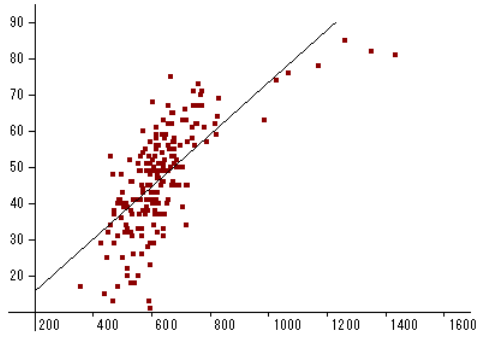


6）地域 1, 2 の年収に関するヒストグラム　　［　　］





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＝［　　］

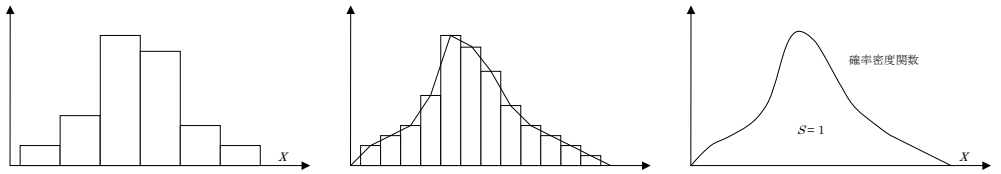
9）支出を年収で予測する回帰式を求めよ。　　［　　］

支出＝［　　］×年収＋［　　］

２章　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。

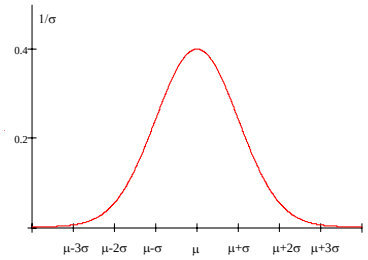


$y = f(x)$：確率密度関数

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。

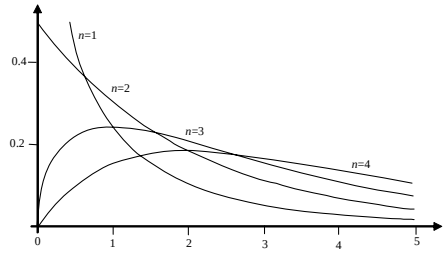
１）$\overline{X}$ の平均（標本平均の平均）はいくらか。　［　　］kg

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

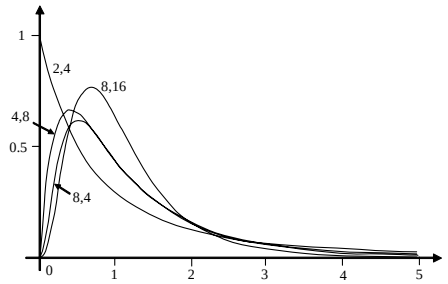
χ^2 分布



$X_i \sim N(0,1)$ 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

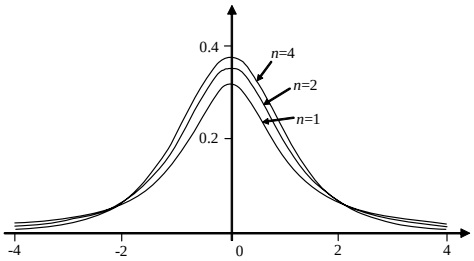


$\chi_1^2 \sim \chi_{n_1}^2$ 分布，$\chi_2^2 \sim \chi_{n_2}^2$ 分布で独立なとき、

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

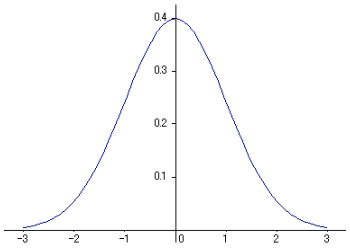
8） t_{10} 分布，t 値 2 のときの上側確率 $p/2$　［　　］

9） t_{10} 分布，t 値 2 のときの両側確率 p　［　　］

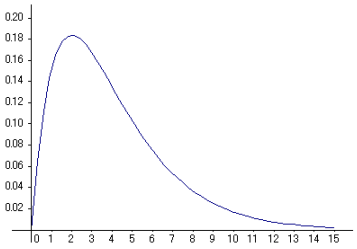
10） t_{10} 分布，両側確率 0.05 のときの t 値　［　　］

問題　College Analysis を使って以下のグラフを描け。

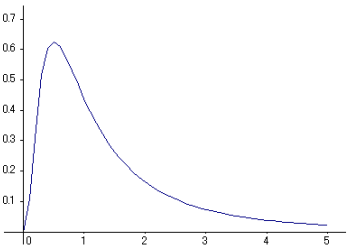
1） $N(0,1)$ 分布



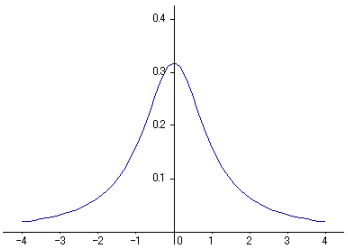
2）自由度 4 の χ^2 分布



3）自由度 8,4 の F 分布

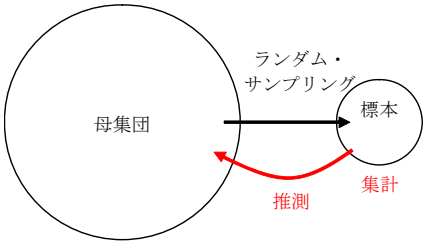


4）自由度 1 の t 分布



2.3 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。

検定とは

例　超能力を持つという人にコインの裏表を当てる実験をしてもらい、100 回の試行で 70%の正解率を得た。この人には本当に超能力があると考えられるか？

有意水準を 5%として判定せよ。20 回の試行ではどうか。

有意水準（危険率）：超能力があると判定して間違う確率

70%の正解率は確かに超能力があって起こったものか、偶然に起こったものか、判定する。

答　χ^2 検定を用いる。

試行回数 100 回

$$\chi^2 = \frac{(70-50)^2}{50} + \frac{(30-50)^2}{50} = 2\times\frac{400}{50} = 16 \quad (\sim \chi_1^2 \text{ 分布})$$

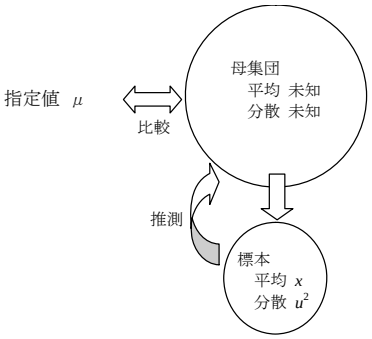
$p = 0.00006 < 0.05$ より、超能力があるといえる。

試行回数 20 回

$$\chi^2 = \frac{(14-10)^2}{10} + \frac{(6-10)^2}{10} = 2\times\frac{16}{10} = 3.2$$

$p = 0.07364 > 0.05$ より、超能力があるといえない。

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。

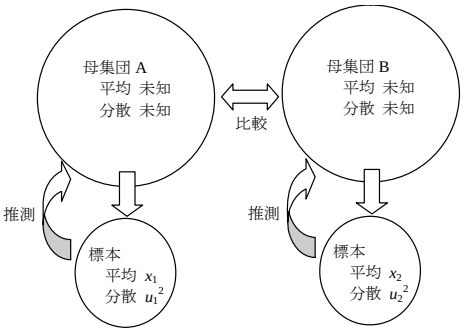
量的データの比較：

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）

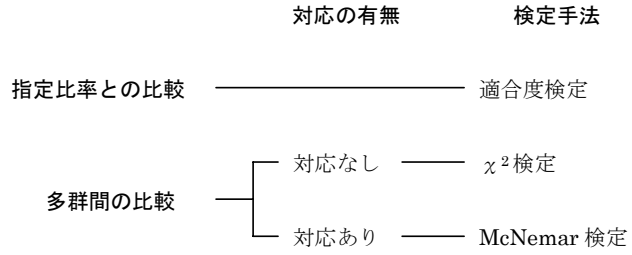
質的データの比較：

男女間での意識調査の結果（割合）の比較（対応がない場合）

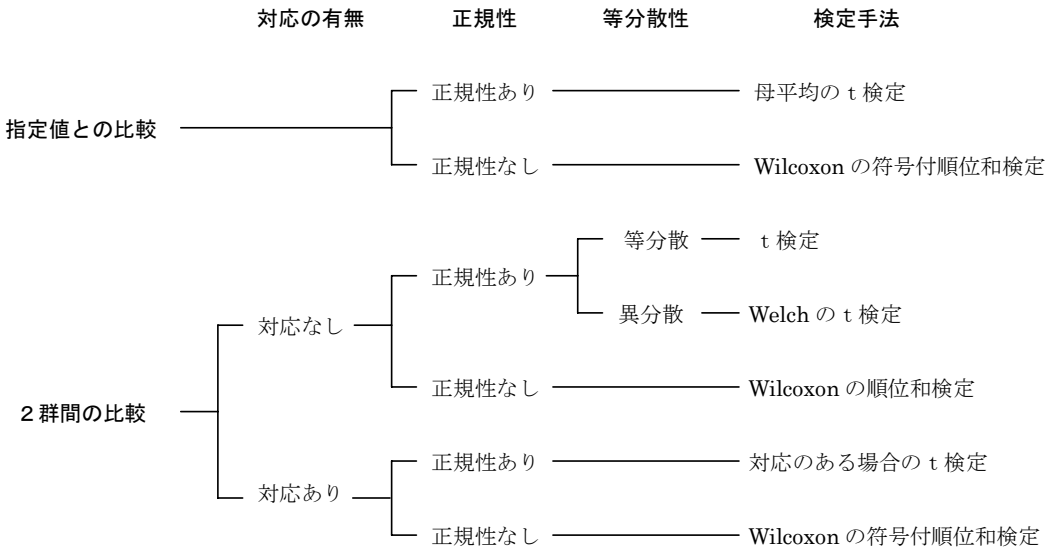
標本店における従業員教育前後の評判の変化（対応がある場合）

2.4 検定選択ツリー

質的データ



量的データ



以後、これらの検定を詳細に見て行く。

３章　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

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

$$\chi^2 = \frac{(|n_1 - m_1| - ½)^2}{m_1} + \frac{(|n_2 - m_2| - ½)^2}{m_2} + \cdots + \frac{(|n_k - m_k| - ½)^2}{m_k} \underset{n\to\infty}{\sim} \chi^2_{k-1} \text{分布}$$

（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］

問題３

Samples¥テキスト 9.txt について以下の問いに答え、結果を文書にまとめよ。

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）上の問題で Yates の補正をしない場合どうなるか。

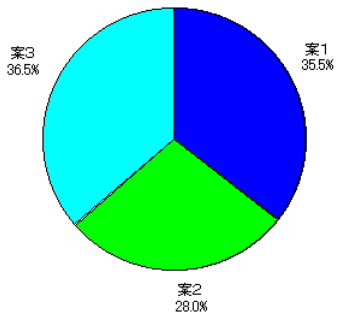
P = ［　　］

判定　過半数と［いえる・いえない］

４）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

５）意見２について以下のような円グラフを描け。



６）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］

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

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

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

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

解答

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

$p > 0.05$ より、男女差があるとはいえない。

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

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

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

問題１

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２

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３

Samples¥テキスト 9.txt において、以下の問いに答えよ。

１）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］。

２）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］。

３）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］。

４）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］。

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。

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
［　　］		
［　　］		

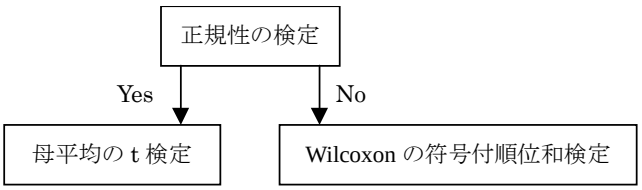
5）新しいデータを用いて、新製品は売り上げに影響を与えたと言えるか有意水準 5%で判定せよ

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章 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

例

以下のデータの正規性を調べよ。

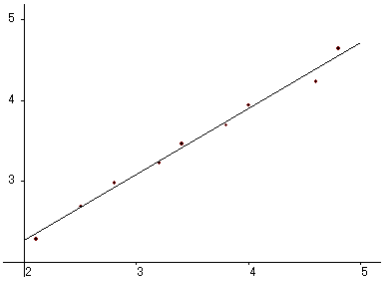
2.5，2.1，3.4，2.8，4.6，3.2，3.8，4.8，4.0

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

以下のデータの正規性を調べよ。

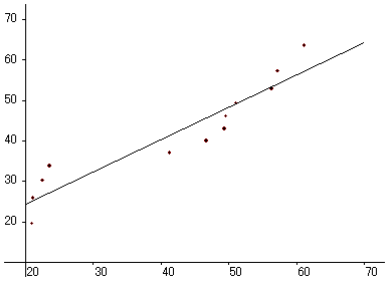
20.9，61.1，57.2，51.0，46.6，41.2，21.0，56.3，49.5，49.3，22.4，23.5

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［　　］

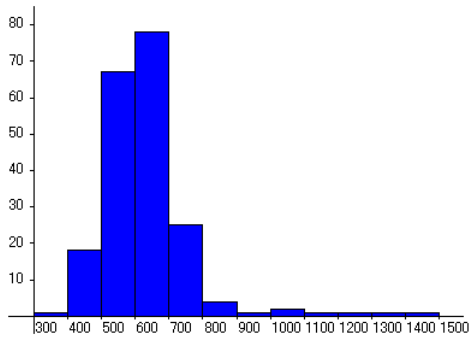
判定　正規分布と［みなす・いえない・判定困難］

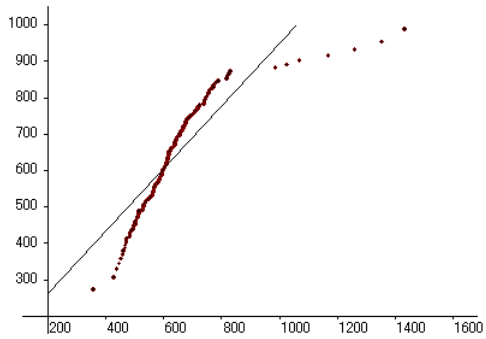


問題

Samples¥テキスト 9.txt のデータについて以下の問いに答え、結果をレポートに記せ。

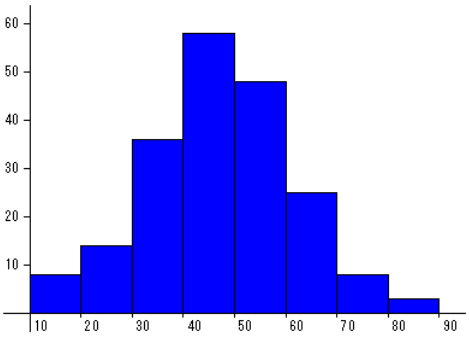
１）年収のデータの正規性をヒストグラム、正規確率紙、S-W 検定で調べよ。

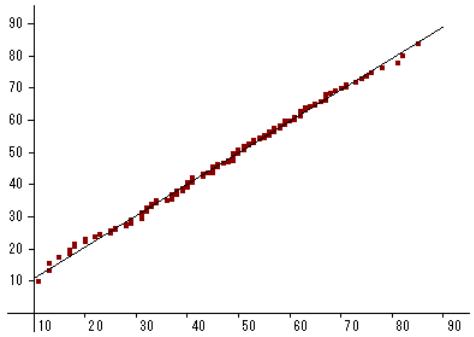




S-W 検定　　確率［　　］判定　正規分布と［みなす・いえない・判定困難］。

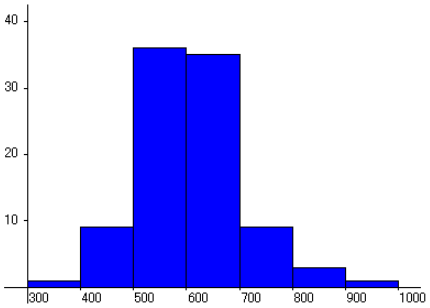
２）支出のデータの正規性をヒストグラム、正規確率紙、S-W 検定で調べよ。

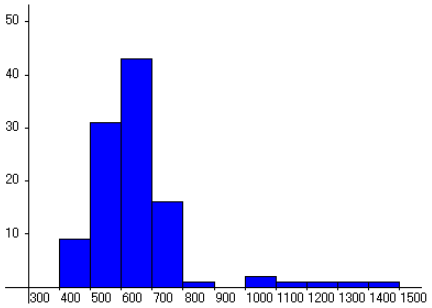




S-W 検定　　確率［　　］判定　正規分布と［みなす・いえない・判定困難］。

３）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・判定困難］。

地域２　　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。

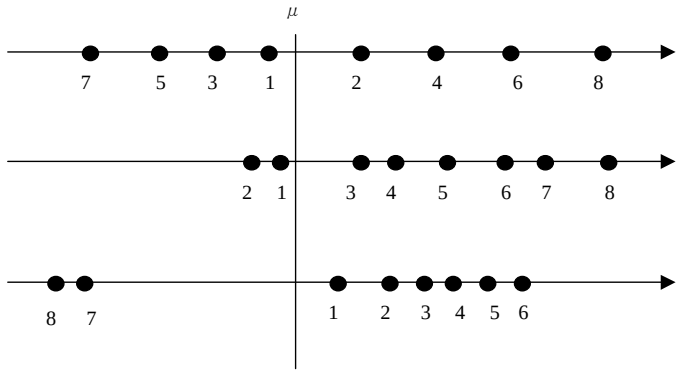


図　検定概念図

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

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

解答

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

問題

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

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。

正規性の判定

ヒストグラムを描く。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　610 万円より多いと［いえる・いえない］。

2）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。

正規性の判定

ヒストグラムを描く。正規確率紙を描く。S-W 検定　確率［　　］

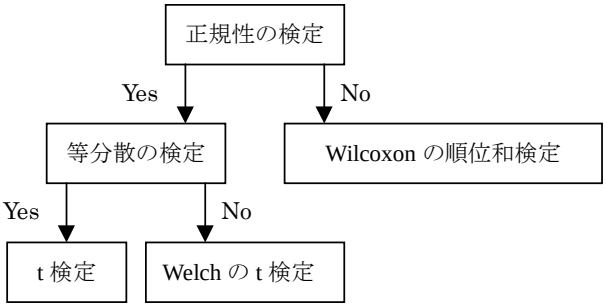
判定　正規分布と［みなす・いえない］。

検定名［　　］　確率［　　］

判定　44 万円と比べて差があると［いえる・いえない］。

５章　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。

5.5 対応のない２群間の中央値の検定（正規性なし）

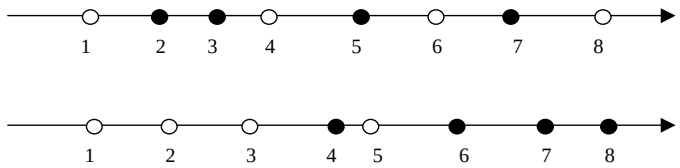
例

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

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

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

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。

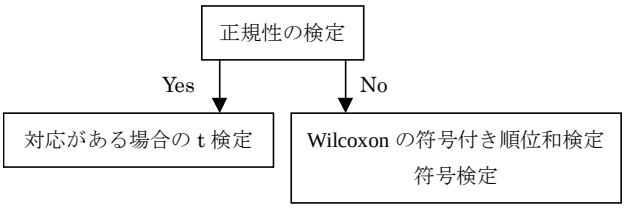
１群：7.2, 8.3, 5.4, 6.0, 7.3, 11.7, 10.5, 8.0, 9.1
２群：10.1, 13.2, 7.4, 9.1, 16.2, 14.5, 6.3, 11.2, 12.4, 7.4, 12.5, 9.1, 17.0
検定名［　　］　確率［　　］
判定　体重増加に差があると［いえる・いえない］

問題

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］

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

$t = 2.149398 \quad p = 0.068675 > 0.05$　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				

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］

６章　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

$r = 0.453786$，$n = 16$，$t = 1.905387$

$p = 0.077476 > 0.05$ 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

$r_s = 0.461312$，$t = 1.945443$

$p = 0.072084 > 0.05$　より、相関があるとはいえない。

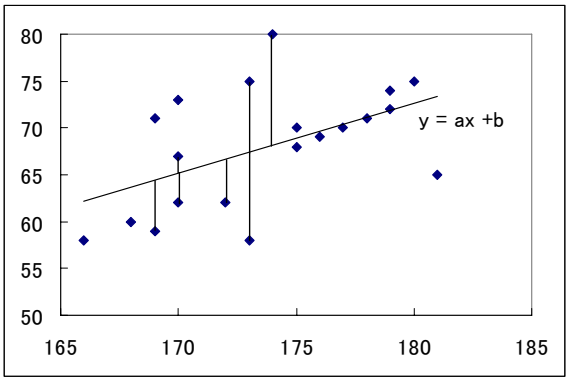
6.3 回帰分析

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

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

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。

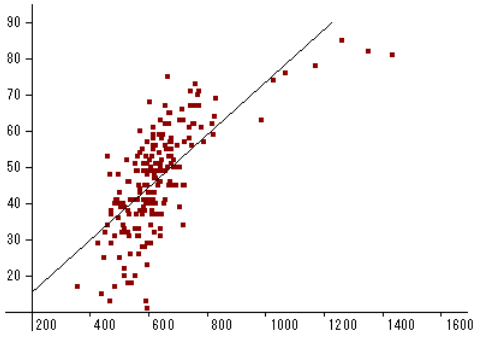
相関係数［　　］　順位相関係数［　　］

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］

７章　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

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

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

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］。

８章　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

この度情報処理論Ⅱ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。

福山平成大学　福井正康

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

福山平成大学　福井正康

福山平成大学では 20XX 年 11 月 28 日に、本学情報処理論Ⅱ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。この報告書で行なった検定については有意水準を 5%としている。

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。

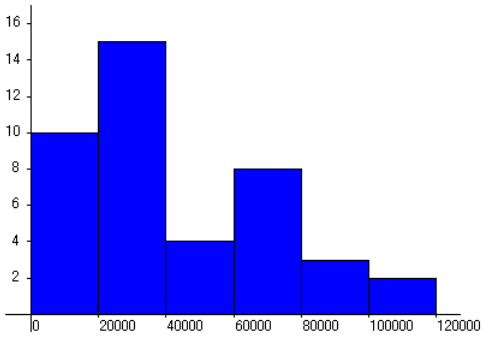


図 1　自由に使える金額

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
□	□	□	□	□	□

図 1 のヒストグラムの形から、データが正規分布していると考えにくいので、これら

の差を Wilcoxon の順位和検定で調べたところ、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=　　）、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。

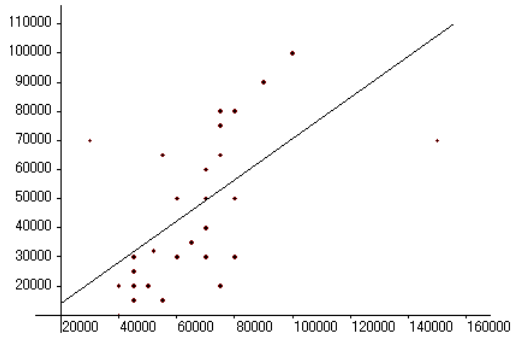


図 2　アルバイト収入（横軸）と使える金額（縦軸）の相関

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。

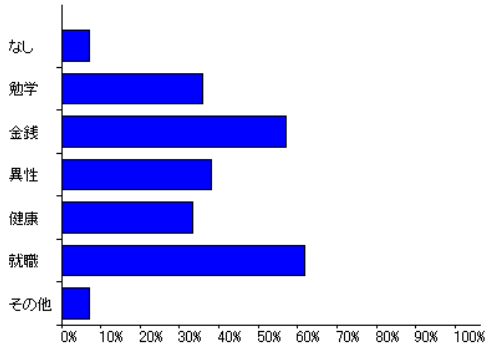


図 3 悩みの種類の割合

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（ここまで表紙にしてもよい）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、t 検定とか Wilcoxon の順位和検定とか、手法の名前は明らかにした方がよいが、t 統計量の値や自由度などは書かない。
　　有意水準・検定確率値・判定については必要に応じて流れの中で記述する。
　　検定確率値については、小数点以下 3 桁か 4 桁で表示する。