

1. データの集計【第1回】

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

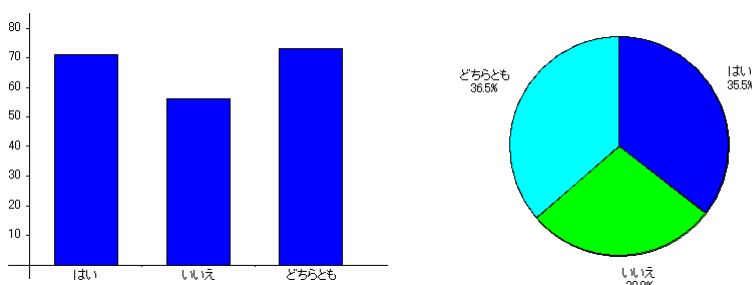
4) 地域と意見 1 に関する 2 次元分割表を描け。

	賛成	反対	合計
市街			
郊外			
合計			

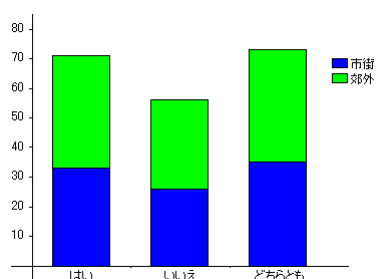
5) 地域と意見 2 に関する 2 次元分割表を描け。

	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【第2回】

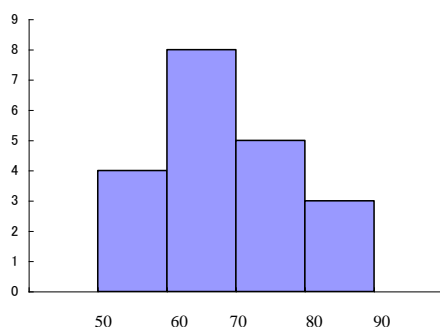
1.2.1 量的データの集計

度数分布表

階級	度数	相対度数 (%)	累積度数	累積相対 度数(%)
$50 \leq x < 60$	4	20	4	20
$60 \leq x < 70$	8	40	12	60
$70 \leq x < 80$	5	25	17	85
$80 \leq x < 90$	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

$$2, 3, 3, 4, 8 \quad \rightarrow \quad 3$$

$$2, 3, 3, 4, 6, 8 \quad \rightarrow \quad (3+4)/2=3.5$$

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

分布の広がりを表わす統計量（散布度）

レンジ（range）

$$R = \text{最大値} - \text{最小値} = 6$$

分散（variance）

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

標準偏差（standard deviation）

$$s = \sqrt{\text{分散}} = 2.098$$

不偏分散

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

標準偏差（standard deviation）

$$u = \sqrt{\text{不偏分散}} = 2.345$$

例

Samples¥テキスト 1.txt のデータ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	181	65	2	173	58
1	179	74	2	169	59
2	178	71	2	170	73

問題

1) 身長についての基本統計量を求めよ。

2) 体重について平均、中央値、標準偏差を求めよ。

平均 [] 中央値 [] 標準偏差 []

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) 学校別に体重の平均と標準偏差を求めよ。

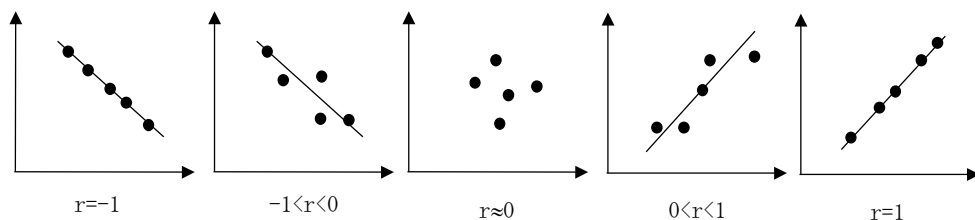
学校 1 平均 [] 標準偏差 []

学校 2 平均 [] 標準偏差 []

8) 学校 1 について、身長のヒストグラムを描け。

1.2.2 相関係数と回帰式【第3回】

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

データの広がりを見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]

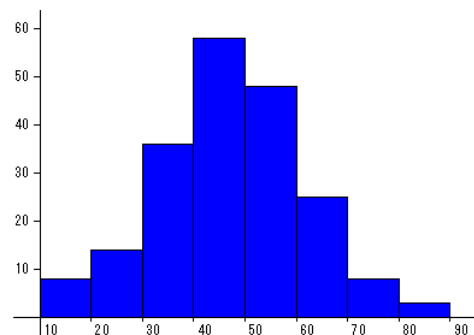
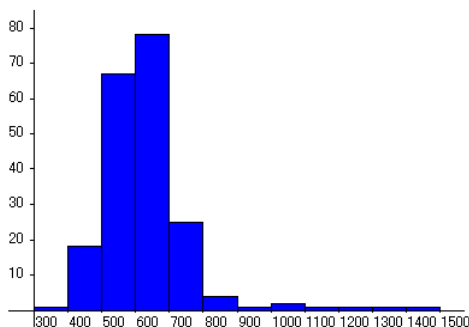
市街と郊外ではどちらの年収の拡がり大きいのか [市街・郊外]

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 []

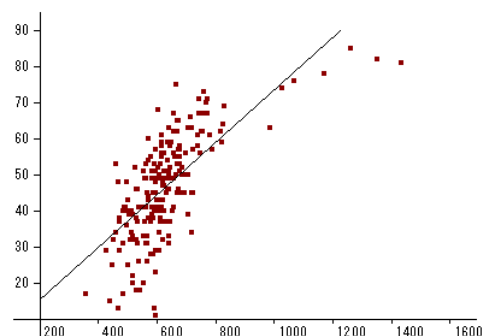
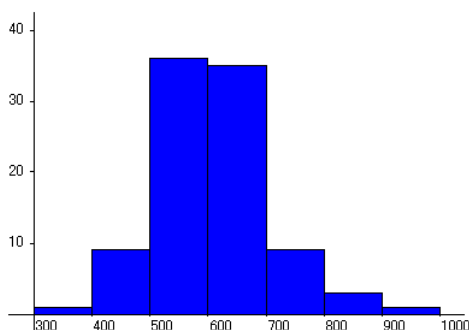
このヒストグラムの最頻値はいくらか []

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 []

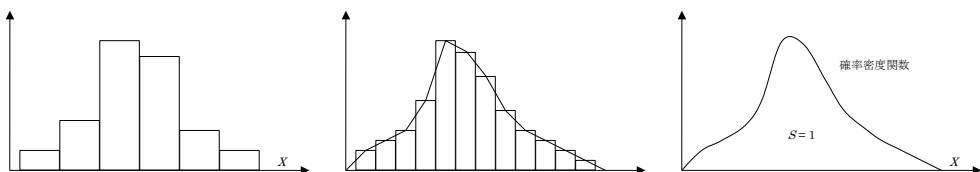
8) 支出を目的変数に年収を説明変数としたときの回帰式を求めよ。

支出 = [] × 年収 + []

2. 確率分布と検定【第4回】

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。

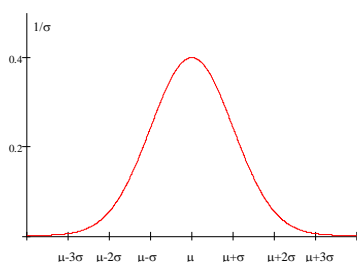


$y = f(x)$: 確率密度関数

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)



$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.683 \quad \text{両側} \quad \text{約 } 32\%$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.954 \quad \text{両側} \quad \text{約 } 5\%$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997 \quad \text{両側} \quad \text{約 } 0.3\%$$

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

$$X' = \frac{X - \mu}{\sigma} \sim N(0,1)$$

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る。

$$P(X \leq x) = \text{normsdist}(x)$$

2) \bar{X} を n 個のデータの平均とすると

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

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 σ/\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 20kg、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) 22kg の子供は重い方から大体何%か [] %
- 2) 24kg の子供は重い方から大体何%か [] %
- 3) 24kg の子供は重い方から大体何番目か [] 番目
- 4) 18kg の子供は重い方から大体何%か [] %
- 5) 18kg 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 20kg 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) 20kg の子供の値はいくらになるか []
- 2) 22kg の子供の値はいくらになるか []
- 3) 17kg 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、17kg 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、22kg 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) 17kg 以上、22kg 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 []

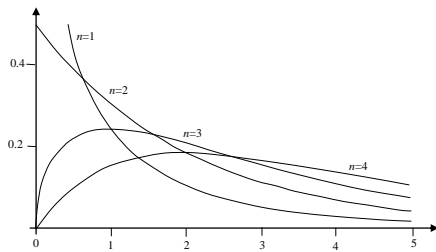
問題（データの平均について）

体重の平均 20kg、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 \bar{X} を取るとする。以下の問いに答えよ。

- 1) \bar{X} の平均（標本平均の平均）はいくらか。 [] kg
- 2) \bar{X} の標準偏差（標本平均の標準偏差）はいくらか [] kg
- 3) \bar{X} の値が 20.2kg 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【第5回】

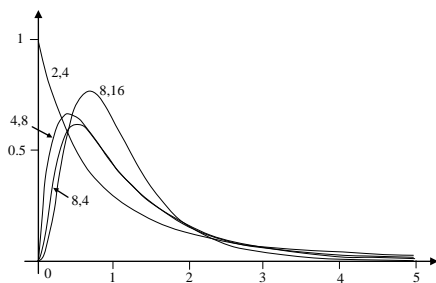
χ^2 分布



$X_i \sim N(0, 1)$ で独立なとき、

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

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、

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

(自由度 n_1, n_2 の F 分布)

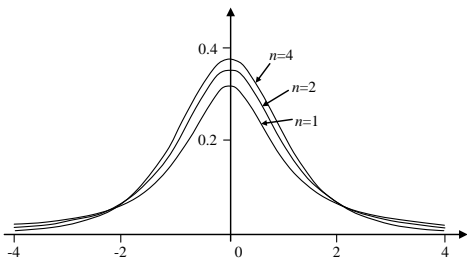
t 分布

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

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

注) $t^2 \sim F_{1, n}$ 分布

注) $n \rightarrow \infty$ 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 | [|] |

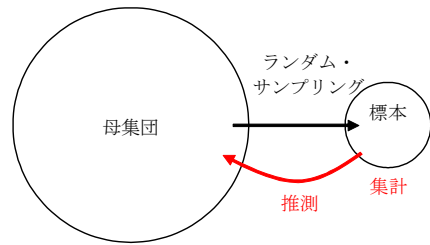
2.4 検定の考え方

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

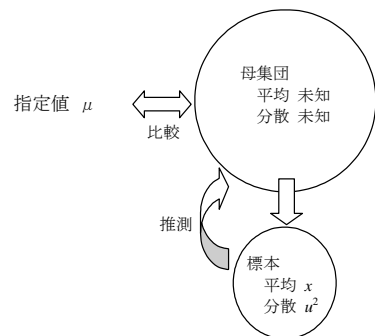
- 1) 指定値と母集団のある指標を比較する。

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)

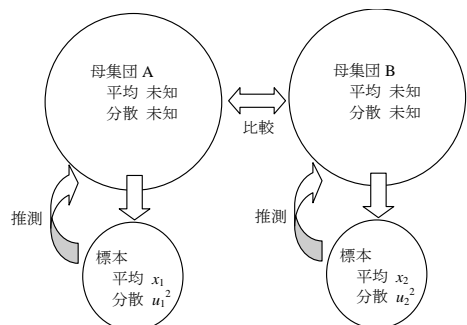
標本店における宣伝前後の売り上げ比較

(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）



3. 質的データの検定【第6回】

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

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

(Yates の連続補正)

解答

$$p_1 = p_2 = 1/2$$

$$\chi^2 = 7.22$$

$$p = 0.0072$$

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

$$P = [\quad]$$

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

$$P = [\quad]$$

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

$$\chi^2 = 1.1358, \quad p = 0.2865$$

$p > 0.05$ 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

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

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

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}	$x_{1\cdot}$
要因 2	x_{21}	x_{22}	...	x_{2s}	$x_{2\cdot}$
⋮	⋮	⋮		⋮	⋮
要因 r	x_{r1}	x_{r2}	...	x_{rs}	$x_{r\cdot}$
計	$x_{\cdot 1}$	$x_{\cdot 2}$...	$x_{\cdot s}$	N

要因間で、事象の出現比率に差がないとすると

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

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

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。

$P = [\quad]$

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 [\quad] 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 [\quad] 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 [\quad] 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 [\quad] 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 [\quad] 差があると [いえる・いえない]

演習 5 (適合度検定・ χ^2 検定)【第 7 回】

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、その理由を教えてください。(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、その理由を教えてください。(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 []

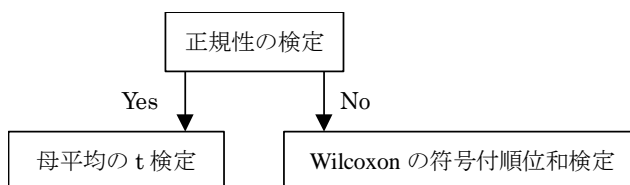
結果 年齢層間で差があると [いえる・いえない]

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]

4. 母集団と指定値との量的データの検定【第8回】

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。

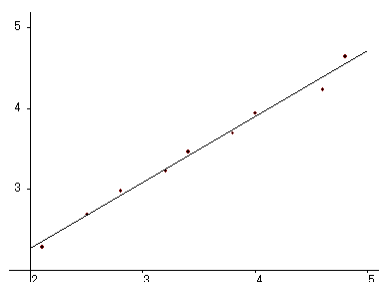
2.5, 2.1, 3.4, 2.8, 4.6, 3.2, 3.8, 4.8, 4.0

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。

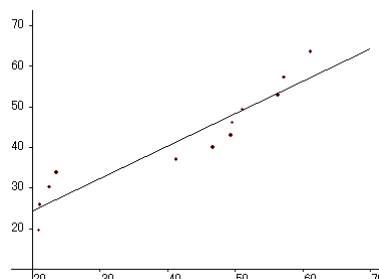
20.9, 61.1, 57.2, 51.0, 46.6, 41.2, 21.0, 56.3,
49.5, 49.3, 22.4, 23.5

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。

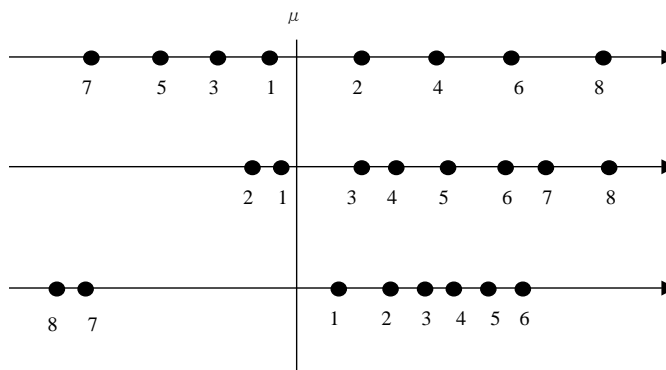


図 検定概念図

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

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

解答

$R = 8$, $p = 0.0294 < 0.05$ 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]。

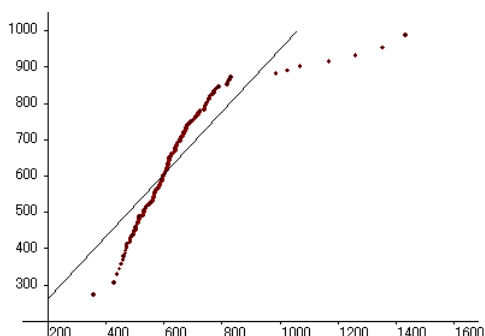
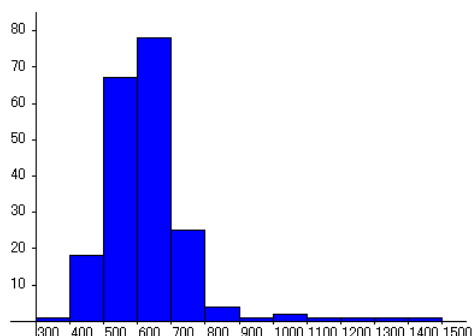
検定名 [] 確率 []

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。



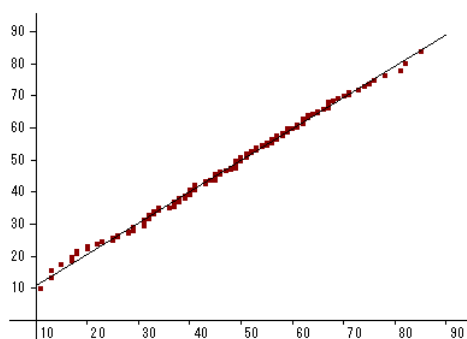
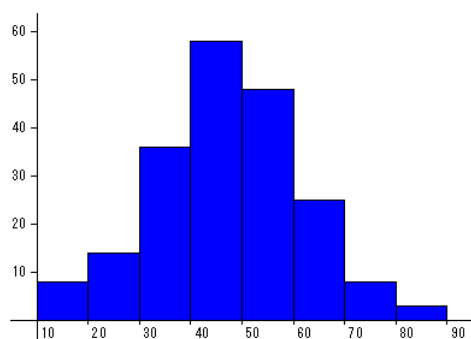
S-W 検定 確率 [] 判定 正規分布と [みなす・いえない]。

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。



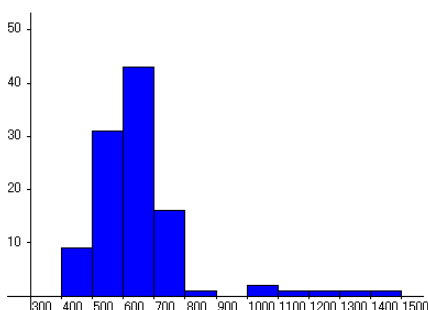
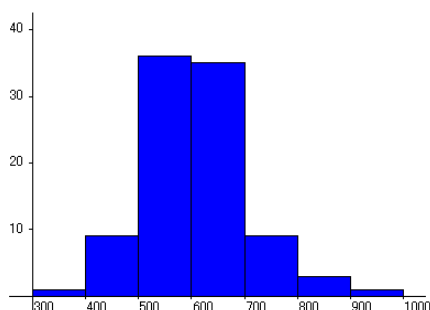
S-W 検定 確率 [] 判定 正規分布と [みなす・いえない]。

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。

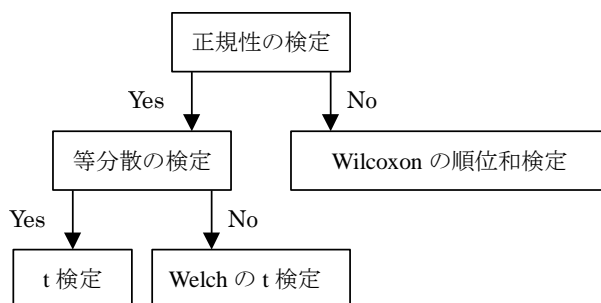


地域 1 確率 [] 判定 正規分布と [みなす・いえない]。

地域 2 確率 [] 判定 正規分布と [みなす・いえない]。

5. 2 群間の量的データの検定【第 9 回】

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

$$F = 3.3691 \quad p = 0.0132 < 0.05 \quad \text{より、分散に差があるといえる。}$$

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

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

解答

$$t = 2.6380 \quad p = 0.0120 < 0.05 \quad \text{より、平均に差があるといえる。}$$

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

$$c = \frac{u_1^2/n_1}{u_1^2/n_1 + u_2^2/n_2} \quad \text{として、自由度を} \quad d = \frac{1}{\frac{c^2}{n_1 - 1} + \frac{(1 - c)^2}{n_2 - 1}} \quad \text{とし、}$$
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{u_1^2/n_1 + u_2^2/n_2}} \sim t_d \text{ 分布}$$

解答

$$c = 0.7374 \quad d = 27.0931 \cong 27 \quad (\text{自由度}) \quad (\text{小数点以下切り捨て})$$
$$t = 2.6086 \quad p = 0.0146 < 0.05 \quad \text{より、平均に差があるといえる。}$$

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）

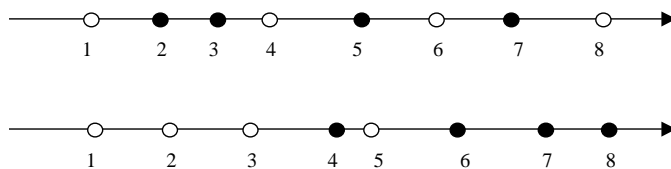
例

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 ($n_1 \leq n_2$) 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

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

解答

$p = 0.3403 > 0.05$ 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]

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。

- 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。

1) 男女の人数を求める。
男子 [] 人 女子 [] 人

- | | | | |
|-----|----|----|----|
| | 英語 | 数学 | 国語 |
| 平均 | | | |
| 中央値 | | | |

- | | | | |
|------|----|----|----|
| | 英語 | 数学 | 国語 |
| 標準偏差 | | | |

- 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]。

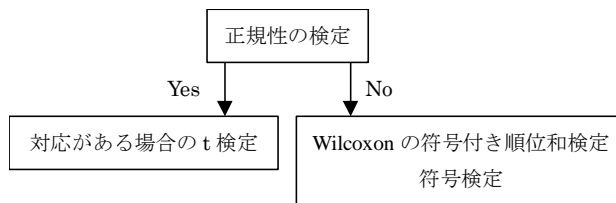
- 13) 男女別の国語の平均（中央値）に差があるか、有意水準 5%で判定する。

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【第 11 回】



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（ $z_i = (\text{標本 1} - \text{標本 2}) \div 2$ ）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

$$t = 2.1494 \quad p = 0.0687 > 0.05 \quad \text{より、平均に差があるとはいえない。}$$

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（ $z_i = (\text{標本 1} - \text{標本 2}) \div 2$ ）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）

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

解答

$p = 0.3820 > 0.05$ 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、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]

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) 外で遊ぶ人数と家の中で遊ぶ人数を求める。

外で	家の中で	合計

- 2) 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]

- 3) 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			

- 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 of 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 of 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]。

6. 相関係数の検定と回帰分析【第 12 回】

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

$$r = 0.454, \quad n = 16, \quad t = 1.9054$$

$$p = 0.0775 > 0.05 \text{ より、相関があるといえない。}$$

(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

$$r_s = 0.461, \quad t = 1.9454$$

$$p = 0.0721 > 0.05 \text{ より、相関があるとはいえない。}$$

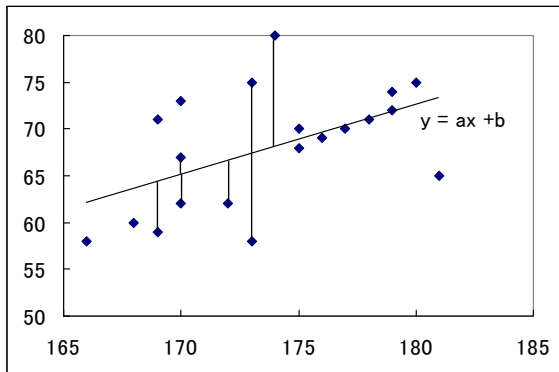
6.3 回帰分析

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、 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

$$\bar{x} = 173.7, \quad \bar{y} = 67.95$$

$$u_x = 4.402153, \quad u_y = 6.378211, \quad r = 0.513$$

$$a = 0.743346, \quad b = -61.1692$$

$$\text{回帰式} \quad y = 0.7433x - 61.1692$$

$$\text{重相関係数} \quad R = 0.513$$

$$\text{寄与率} \quad R^2 = 0.263$$

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。

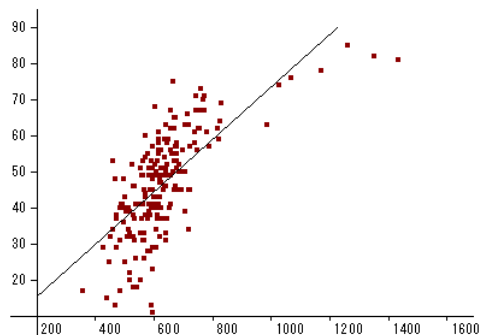
相関係数 [] 順位相関係数 []

- 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 【Skip OK】

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

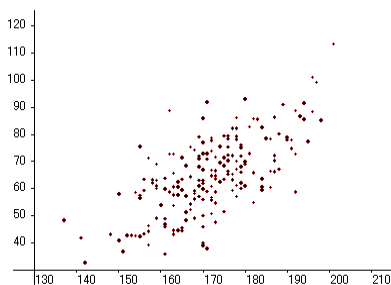
1. 都市部と郊外の調査数
都市部 [] 人 郊外 [] 人
2. 地域別にみた各変数の平均値と中央値を求める。

	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【第 13 回】

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%）。 $1 - \text{信頼係数の値}$ 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。 $z_0 = \text{normsinv}(1 - \alpha/2)$ として、信頼区間は以下で与えられる。

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

解答

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

$$0.5565 \leq p \leq 0.7435$$

1000 人では、以下のように精度が上がる。

$$0.6204 \leq p \leq 0.6796$$

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 、 $t_0 = \text{tinv}(\alpha, n - 1)$ 、 $x_1 = \text{chiinv}(1 - \alpha/2, n - 1)$ 、 $x_2 = \text{chiinv}(\alpha/2, n - 1)$ として、各信頼区間は以下で与えられる。

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

$$5138 \leq \sigma^2 \leq 14638$$
$$6244 \leq \sigma^2 \leq 10932$$

信頼区間は [] ≤母比率≤ []

$$[\quad] \leq \text{母分散} \leq [\quad]$$
$$[\quad] \leq \text{母分散} \leq [\quad]$$

演習 9【できればやってみよう】

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【第 14 回】

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。

質問に漏れないか十分注意する。

例えば意見の男女差を知りたいければ、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。

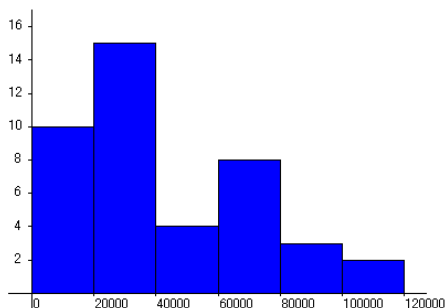


図1 自由に使える金額

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（ $p < [0.05 \cdot 0.01 \cdot 0.001]$ ）、その他については有意な差は見られなかった。

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。

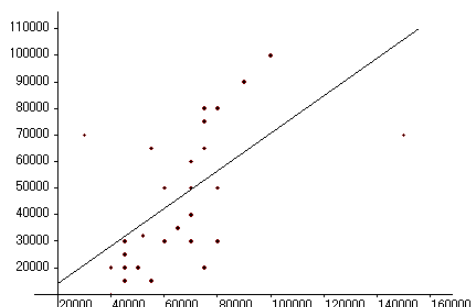


図 2 アルバイト収入（横軸）と使える金額（縦軸）の相関

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 $p < 0.001$, $p < 0.01$, $p < 0.05$, n.s. (***, **, *, n.s. にしてもよい) のどれかの書き方にする。