

3章 質的データの検定

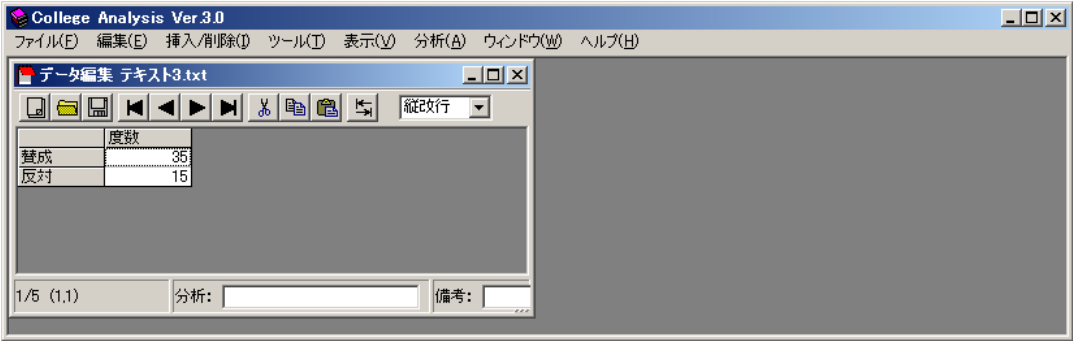
これから検定の手法を順番に詳しく見て行きましょう。最初は質的データの検定についてです。質的データの検定は構成比率の検定で、アンケートで賛成が過半数であるかとか、賛成の比率が男女で異なるかといったことを調べます。これからは最初に例題を置き、それについて **College Analysis** による回答を示します。理論のところではその検定のために使った公式を簡単に表示しますが、必要なければ無視してもらって結構です。

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。

この例は前の章の超能力の検定とほぼ同じ内容です。データはファイル¥Samples¥テキスト 3.txt を選択します。そうすると以下のような分割表になったデータが表示されます。



The screenshot shows the 'College Analysis Ver.3.0' window. The title bar includes menu items: ファイル(F), 編集(E), 挿入/削除(I), ツール(T), 表示(V), 分析(A), ウィンドウ(W), ヘルプ(H). The main window has a tab labeled 'データ編集 テキスト3.txt'. Below the tab is a toolbar with icons for file operations and a dropdown menu labeled '縦改行'. The central area displays a contingency table:

	度数
賛成	35
反対	15

At the bottom of the window, there is a status bar showing '1/5 (1,1)' and two input fields labeled '分析:' and '備考:'.

図 3.1.1 分割表データ

次にメニューで [分析－基本統計－質的データの検定－質的データ検定メニュー] を選択し、質的データの検定 **Window** を開きます。

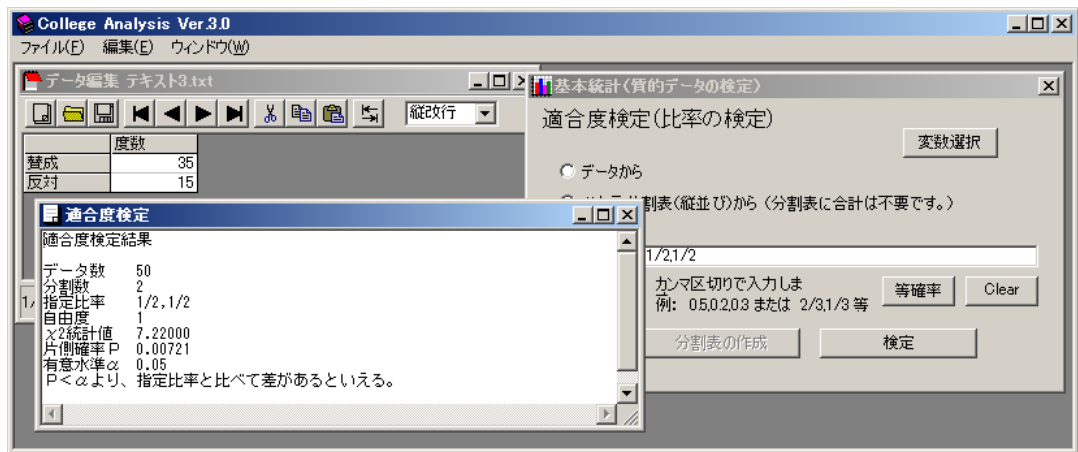


図 3.1.3 適合度検定結果画面

ここでは、いくつかのデータが表示されていますが、注意して見ておくべきは片側確率 P の値と一番下の検定結果の文章です。通常統計ソフトでは検定の確率までは表示されますが、下の文章までは表示されません。これは解釈を間違える学生が多かったことから追加した文章です。検定の評価の文章では、 $p < 0.05$ の場合「差がある」という言い方で結構ですが、 $p \geq 0.05$ の場合は「差がない」ではなく、「差があるといえない」という控えめな言い方をします。これはこのデータ数ではまだ差はみつからないというだけで、データ数を増やすと見つかるようになるかも知れないからです。この例題ではこの段階でもう有意な（有意差がある）結果になっているようです。70%が賛成して、検定の結果が 1/2 と比べて差があるということは過半数であるということです。

最後にこの計算をするための公式を示しておきましょう。

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

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

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

(Yates の連続補正)

ここには公式が 2 種類書いてありますが、一般的な教科書などでは上の公式が説明されています。これはデータ数が十分大きい場合に成り立つ公式なので、実用上は下の公式がよく使われます。上の式から下の式への補正は **Yates** の連続補正と呼ばれています。このような分割表の検定が信用できるものになるためには分割表の 1 つのセルに 10 程度以上の度数が必要であると言われています。

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

$P = [\quad]$

判定 曜日による差があると [いえる・いえない]

問題 2

上の問題で、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。

$P = [\quad]$

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]

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。

$P = [\quad]$

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。

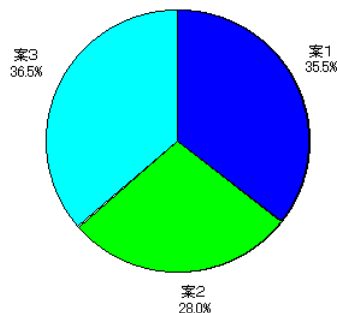
$P = [\quad]$

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。

$P = [\quad]$

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

これまでは指定した比率と母集団の比率を比較するような検定でしたが、ここからは 2 つ以上の母集団の比率の比較検定です。

1. 2×2 表の検定

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

これは男性母集団と女性母集団の賛成比率を比較する検定です。標本での比率は男性 18/28、女性 12/26 で男性の比率が高いですが、母集団ではどうでしょうか。

まずこのデータを読み込み、メニュー [分析－基本統計－質的データの検定－質的

データ検定メニュー] を選択します。

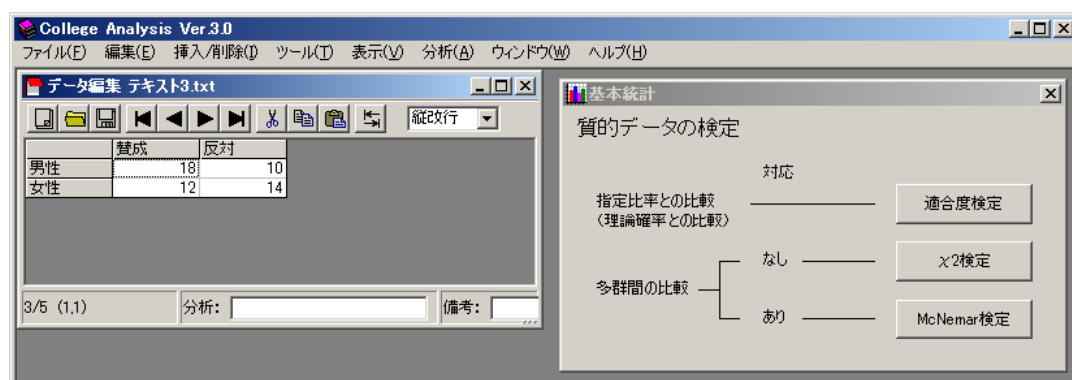


図 3.2.1 質的データの検定メニュー

この例題は男女で比べますので 2 群間の比較です。次に対応ですが、一番分かりやすいのはこの 2 つの群に同じ人がいるかどうかです。2 群は男女で分かれ、同じ人は無理なので、この場合の対応はなしです。例えば使用前と使用後の感想などというと同じ人が 2 つの群にいますので、2 群間に 1 対 1 の対応があると言います。ここでは対応なしのところをたどって、「 χ^2 検定」ボタンをクリックします。すると以下のような χ^2 検定メニューが表示されます。

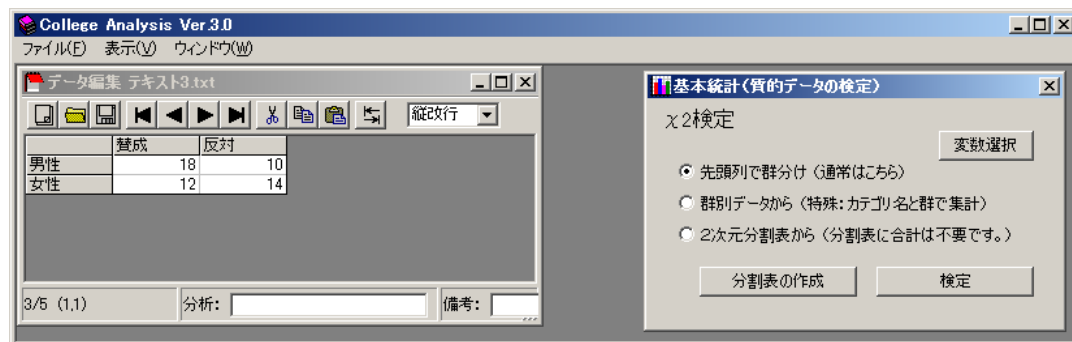


図 3.2.2 χ^2 検定メニュー

このデータは 2 次元分割表なので「2 次元分割表から」ラジオボタンを選び、変数選択ですべてを選択し、「検定」コマンドボタンをクリックします。

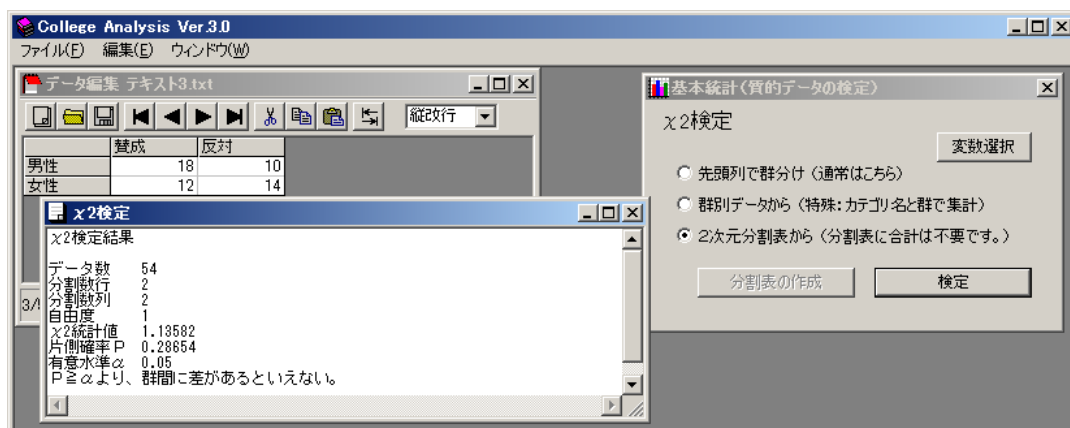


図 3.2.3 χ^2 検定結果画面

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。

以下にここで使った 2×2 分割表の χ^2 検定の公式を示しておきます。 χ^2 検定では 2×2 分割表の場合は簡単に計算できるように、一般の場合とは違った計算式が与えられています。一般の場合はちょっと複雑になりますが、それも示しておきます。

理論

2×2 分割表の場合（特別な書式）

	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

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

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

m × n 分割表の場合

	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

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

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

問題 4

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 5

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 男女間に差があると [いえる・いえない]

問題 6

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)

ここではまず対応について説明します。簡単に言うと 2 つの母集団に同じ人がいるかどうかに対応の分かり易い区別です。例えばある調査を広島県と岡山県で実施して結果を比較したとします。この調査では広島県と岡山県で調査対象が違いますので、一般に同じ人はいません。このような場合、この母集団間に対応はないと言います。同様にある調査で男性の結果と女性の結果を比較する場合、男性の集団と女性の集団には一般に同じ人はいませんので、これも母集団間に対応はありません。

しかし、例えばあるダイエット食品の使用前と使用後で体重を比較するなどは、同じ人を追跡調査するわけですから、使用前の集団に含まれる人が使用後の集団に含まれていないと意味がありません。この様な場合、母集団間に (1 対 1) の対応があると言います。同様にある試験を実施して、英語の点数と国語の点数を比較する場合、英語を受けた人はほぼ国語を受けていますので、どちらか受けていない人を除けば、英語を受験した集団に含まれる人は国語を受験した集団に含まれています。またその逆も成り立ちます。このような場合も対応がある例です。

対応がないデータと対応があるデータでは入力方式が違います。今 2 つの群で対応がある場合とない場合で比較してみましょう。まず対応があるばあいで、例えばある健康食品の使用前と使用後 (前:1, 後:2) で体調 (良い:1, 悪い) を比較する場合です。このように対応がある場合、通常同じ人を横に並べて描きます。このデータの場合の 2 次元分割表は右のようになります。

表 3.3.1 対応のある場合のデータと集計

使用前	使用后	使用前\使用后	良い	悪い
1	1	良い	2	1
2	1	悪い	2	1

1	2
1	1
2	2
2	1

次に名前を調べずに同じ人が特定できない場合の結果はどうなるでしょうか。まず上のように横 2 列に入力することも可能です。但し、同一人物が横 1 列になっている保証はありません。このような形式を我々は群別データと呼んでいます。また別の形式もあります。使用前と使用後を数字で 1 と 2 に分け、人を区別せずに、以下のようにならべます。もちろん順番は以下のようでも構いません。このデータの場合の 2 次元分割表は右のようになります。

表 3.3.2 対応のない場合のデータと集計

前/後	体調
1	1
1	2
1	1
1	1
1	2
1	2
2	1
2	1
2	2
2	1
2	2
2	1

	体調良い	体調悪い
使用前	3	3
使用后	4	2

これを見ると前の表の 1 列と 2 列を縦にくっつけて 2 列目とし、1 列目に使用前は 1、使用後は 2 と入力していることが分かります。対応がある場合に比べてデータ数が 2 倍になっていますが、これは対応がある場合は 2 つのデータを 1 組として数えるからです。

このように対応があるデータには対応がある場合の集計法と対応を考えない集計法があり、検定も 2 つの集計法に応じて変わります。対応のある場合の 2×2 分割表の場合には McNemar という検定法を利用します。もちろん対応を考えない場合は通常の χ^2 検定を用います。

ここでは対応のあるデータの検定の例題を見てみましょう。

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

データを¥Samples¥テキスト 3.txt から開いて、質的データ検定メニューの中から「McNemar 検定」ボタンをクリックすると以下の画面が表示されます。

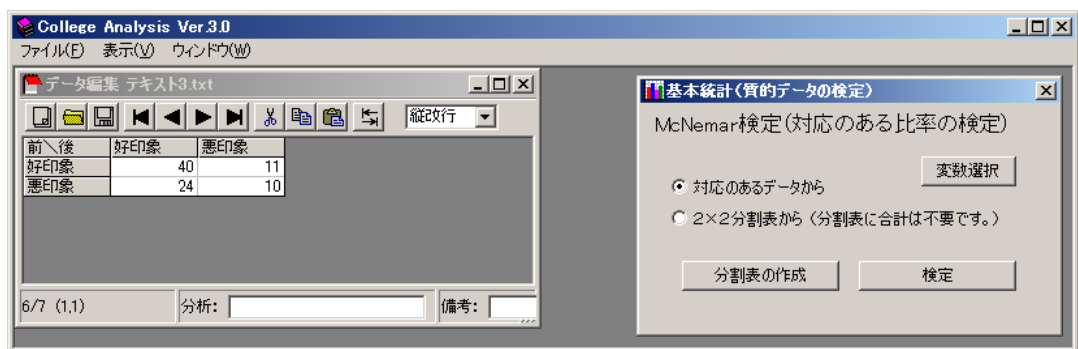


図 3.3.1 McNemar 検定画面

ここで「検定」ボタンをクリックすると以下のような結果が示されます。

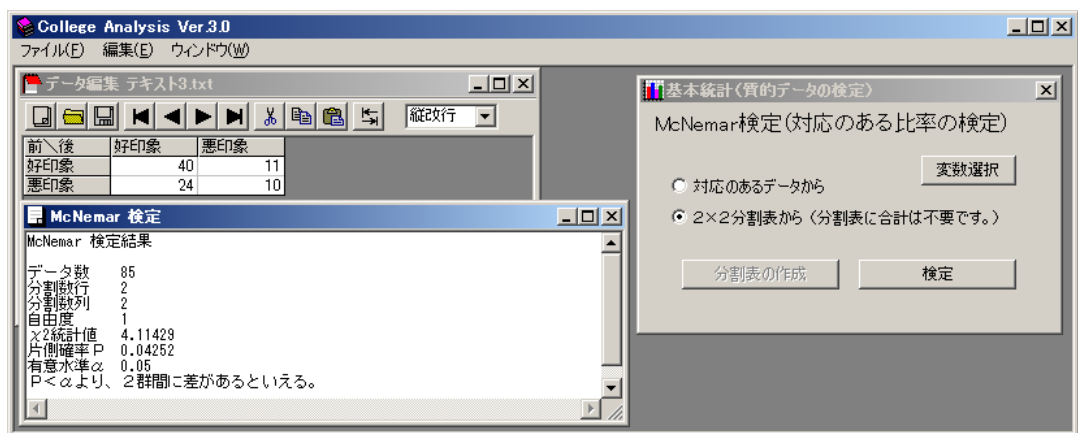


図 3.3.2 McNemar 検定結果

解答

$$\chi^2 = 4.1143, \quad p = 0.04252$$

$p < 0.05$ より、キャンペーンによる差があるといえる。

この計算は大変簡単で、以下のような公式を利用しています。

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

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

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

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	$a+c$	$b+d$

問題 7

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 [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。