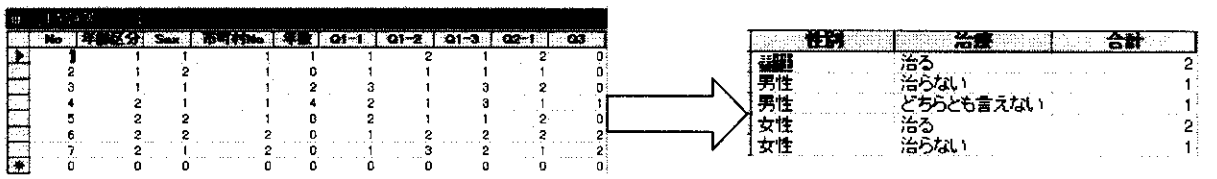
4　アンケートデータを集計する。

データ集計は、クエリーで処理を行います。
クエリーでは、グループ別の件数のカウント、合計や平均値を求めることができます。
また、クロス集計機能を使うこともできます。
また、必要に応じてデータを他のソフト（エクセルなど）に受け渡して（エクスポート）集計をすることも可能です。

（1）単純なグループ集計をする。

男女別や市町村別などに各回答の単純集計を行います。
グループ集計を行うときはグループ化するフィールドと計算（合計）を行うフィールドの２フィールド以上が必要です。

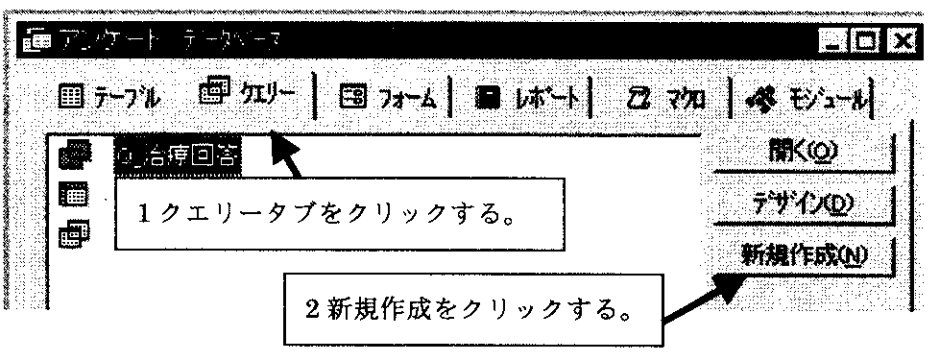
ここでは、問１の１を男女別で集計してみましょう。



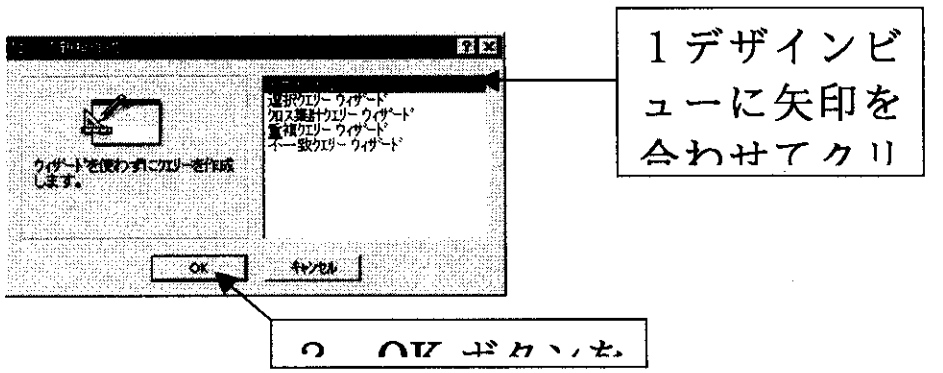
データを入力したテーブルを基にして……男女別に集計したクエリーへ

１）集計クエリーの作成手順。

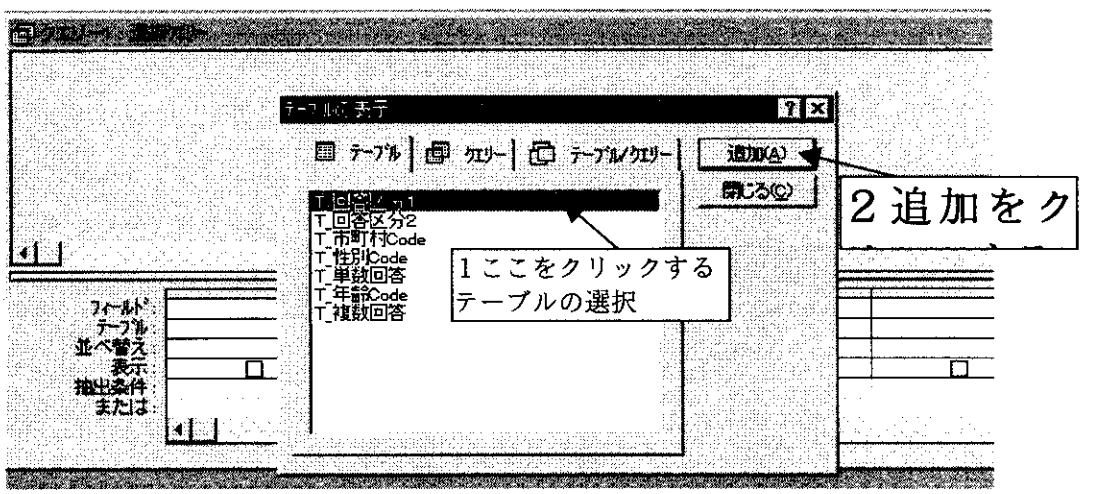
①データベースウィンドウのクエリータブを選択し新規作成ボタンを押します。



②クエリーの新規作成が表示されるので、デザインビューを選択し OK ボタンを押します。

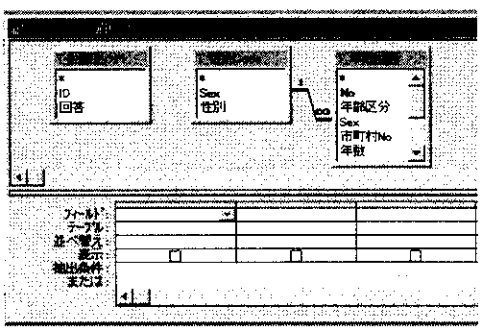


クエリーのデザインビューが表示されます。



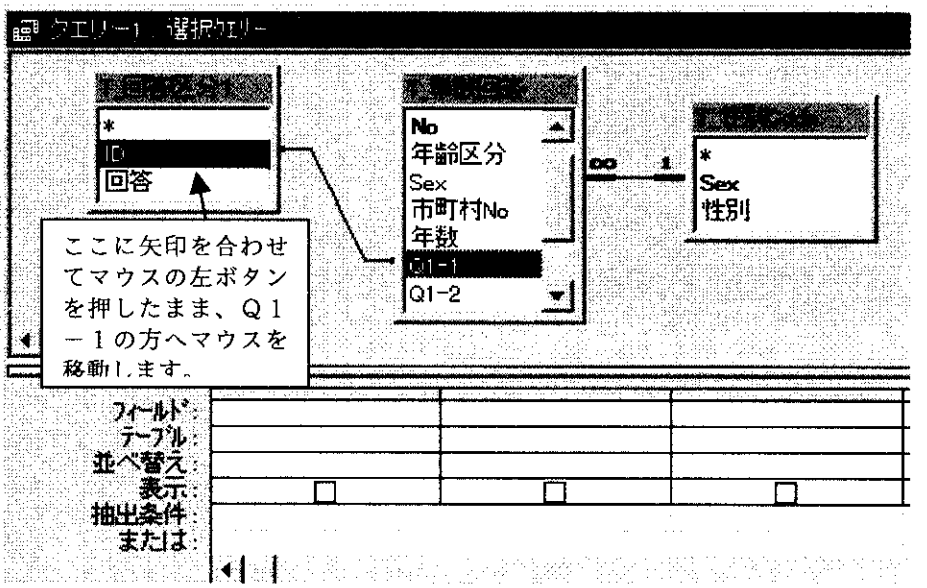
③集計元となるテーブルやクエリーを選択します。
　選択するテーブル名を指定して、追加ボタンを押します。

ここでは、以下のテーブルを選択して下さい。
・T_回答区分１　・T_性別 Code　・T_単数回答



左の画面のようになりデータもとのテーブル名とフィールド名が表示されます。

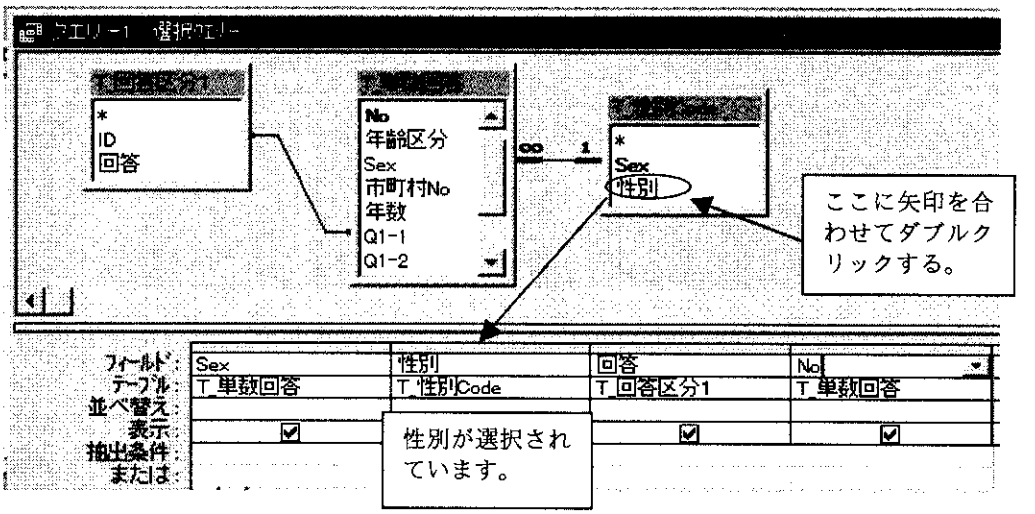
③テーブル同士を結合線で結びます。



T_回答区分１のフィールド ID と T_単数回答の Q1-1（問１）は、同じフィールド定義です。

④データの集計に使用するフィールドを選択します。

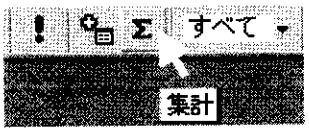
使用するフィールドにマウスを合わせてダブルクリック（左ボタンを２回続けて押す操作）すると、フィールドが選択されます。



練習８　上の図のようにフィールドを選択して下さい。

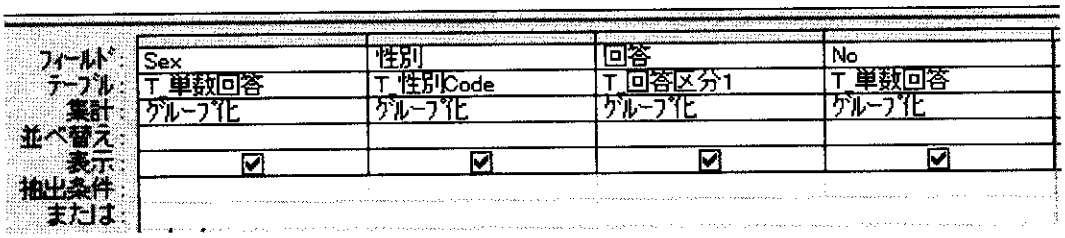
⑤集計ボタンコマンドを指定します。

コマンドバーから∑をクリックして下さい。



コマンドバーが表示されていない場合は、メニューの表示－集計をチェックして下さい。

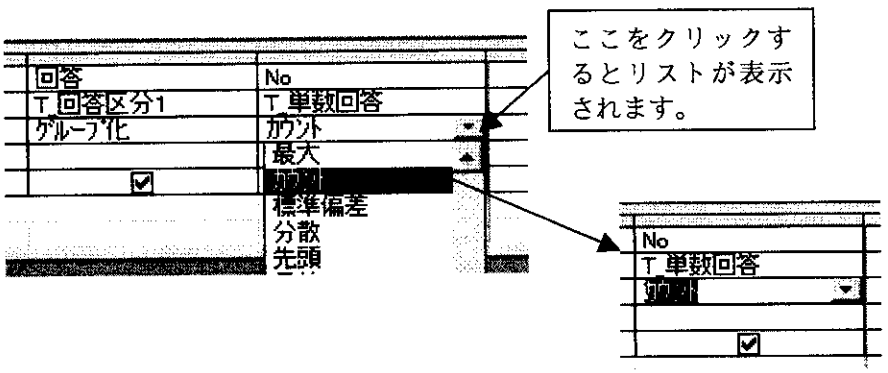
テーブルの下に集計行が表示されます。



グループ化：同じデータを１かたまりのグループにします。

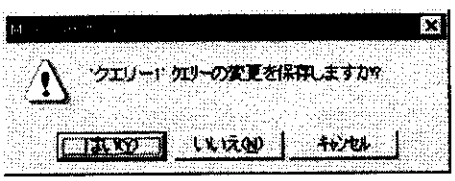
⑥回答者数をカウント（集計）します。

No フィールド（個人）が何件あるかをカウント（集計）するには、No の集計欄に「カウント」を指定します。



⑦クエリー名を名付けて保存します。

ファイル（F）の閉じるを選択すると、「クエリ１－クエリーの変更を保存しますか？」とメッセージが示されます。「はい」を選択してクエリー名をつけて OK ボタンを押してください。

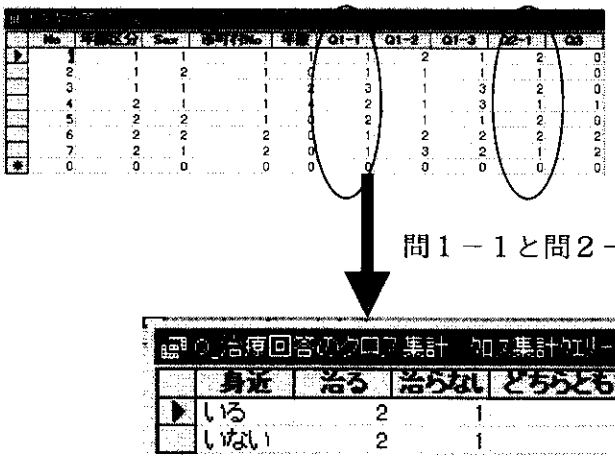


クエリーが保存されました。

クエリーを開くとその時々のデータの集計が行えます。

（2）クロス集計クエリーを作成する。

クロス集計をするには、列と行に該当するフィールドと演算（合計やカウントなど）を行うフィールド（項目）が３つ以上必要です。



データテーブルを元にして２つの項目で縦横集計

問１－１と問２－１のクロス集計結果

	身近	治る	治らない	どちらとも言えない
▶	いる	2	1	
	いない	2	1	1

クロス集計クエリーを作成するには、クロス集計ウィザードを使用する方法と、クエリーのデザインビューで定義する方法の２種類ありますが、今回はウィザードを利用してクロス集計を実行します。

１）クロス集計クエリー用のクエリーを作成する。

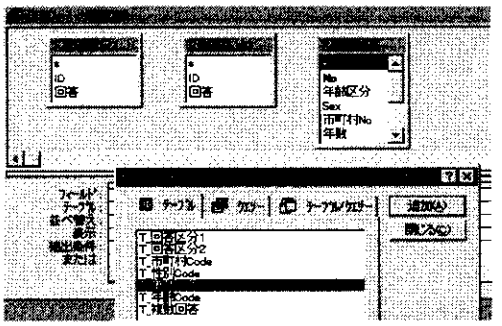
データテーブルから直接クロス集計クエリーを作成することも可能です。
しかし、コード表を利用して新たに集計を行う項目のみのクエリーをいったん作成してクロス集計を行ってみます。

手順の①～③までは、先ほどの集計クエリーと同じ操作です。

①データシートビューのクエリータブを選択し、新規作成ボタンをクリックする。

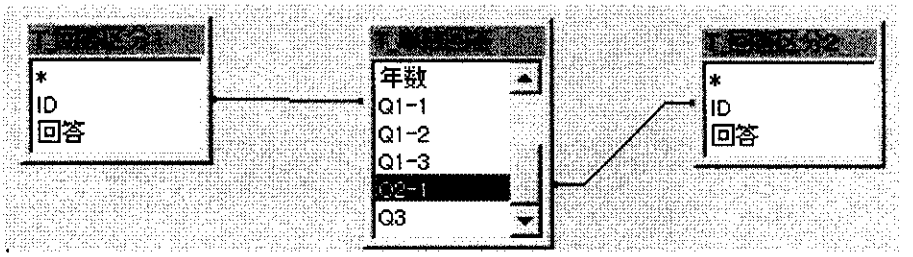
②クエリーの新規作成が表示されるので、デザインビューを選択しOKボタンを押します。

③集計元となるテーブルを選択します。



以下のテーブルを選択して下さい。
T_回答区分１
T_回答区分２
T_単数回答

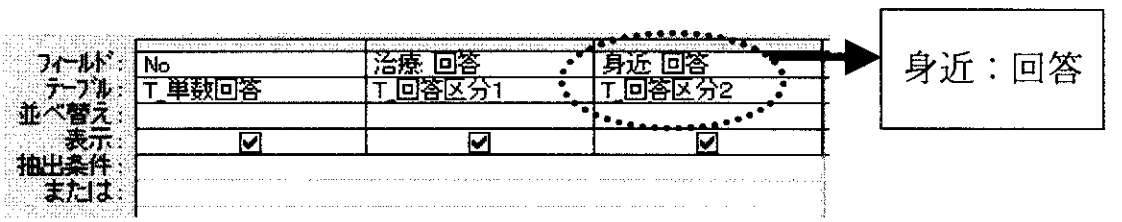
④テーブル同士を結合線で結びます。



T_単数回答のフィールド「Q1-1」と T_回答区分１のフィールド「ID」とを、T_単数回答のフィールド「Q2-1」と T_回答区分２のフィールド「ID」とを結合線で結びます。(操作が分からない場合は、Ｐ３９を参考にして下さい)

⑤出力するフィールドを選択します。

コード表である回答区分１と２からは、フィールド「回答」を、データテーブルである T_単数回答からはフィールド「No」を選択します。



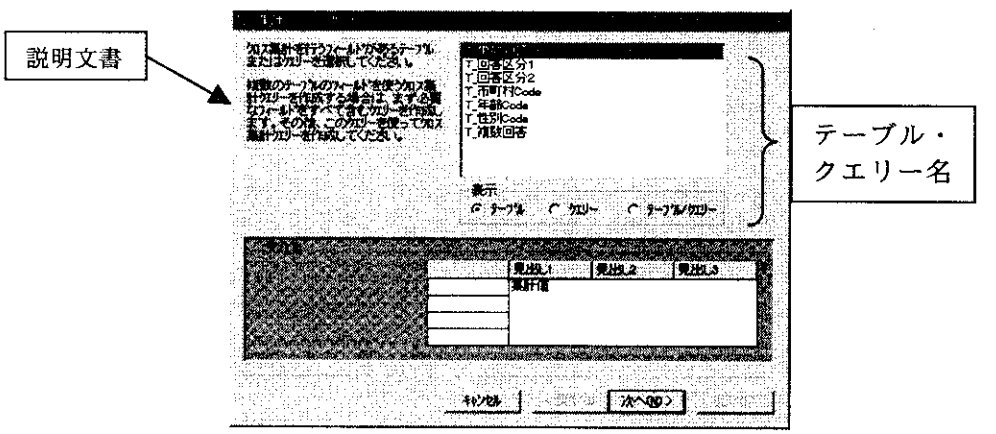
回答区分のフィールドの表示名を「治療」と「身近」に変更しますので、フィールド名の前に「身近」とキーボードで入力し、半角のコロン「：」でフィールド名とを区切って下さい。

⑥クエリー名を名付けて保存します。
クエリー名を「Q_治療」として保存して下さい。

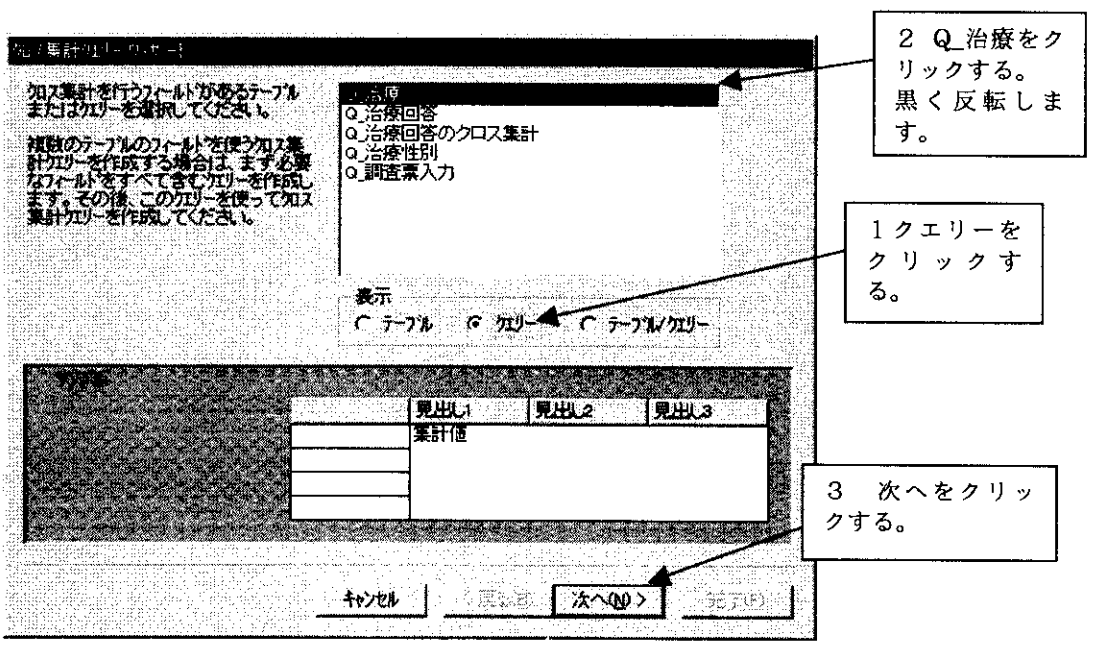
2）クロス集計クエリーを作成する手順。

手順の①～③までは、先ほどの集計クエリーと同じ操作です。
①データシートビューのクエリータブを選択し、新規作成ボタンをクリックする。

②クエリーの新規作成が表示されるので、クロス集計クエリーウィザードを選択しOKボタンを押します。
クロス集計クエリーウィザードが表示されます。

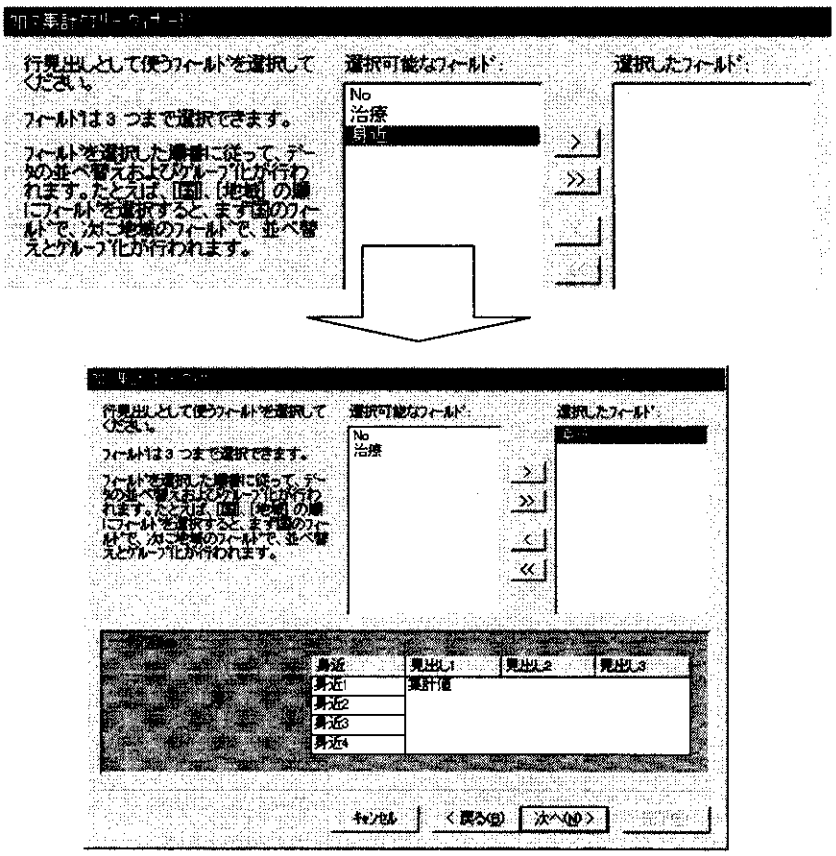


③集計元となるクエリー「Q_治療」を選択し、次へ（N）ボタンをクリック。



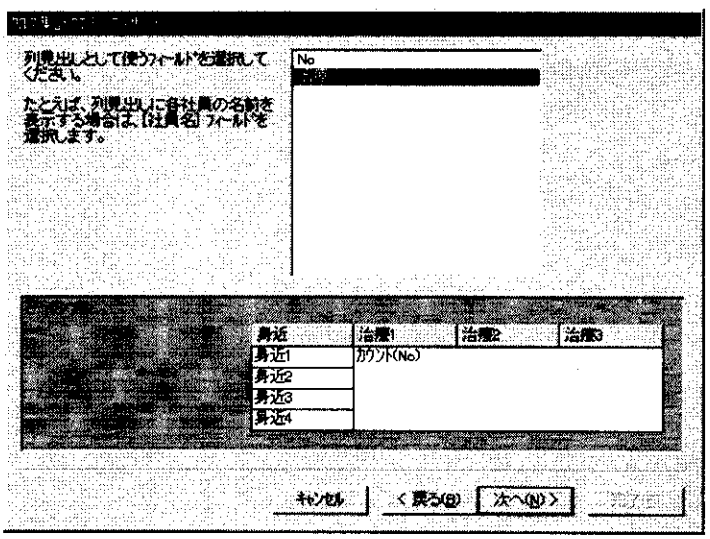
④行見出しとして使うフィールドを選択する。

身近を選択してを　＞　押して下さい。



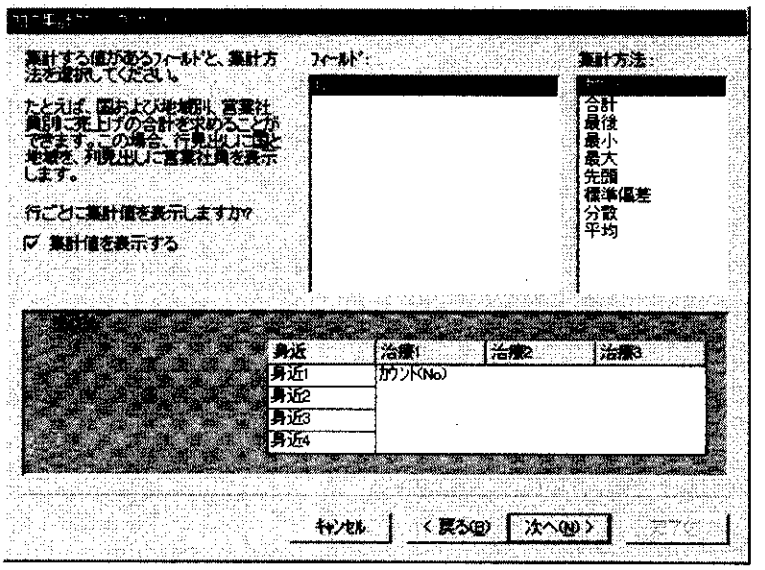
次へ進みます。

⑤列見出しを選択します。同じ要領で「治療」を選択して次へ進みます。



⑥集計するフィールドと計算方法を指定します。

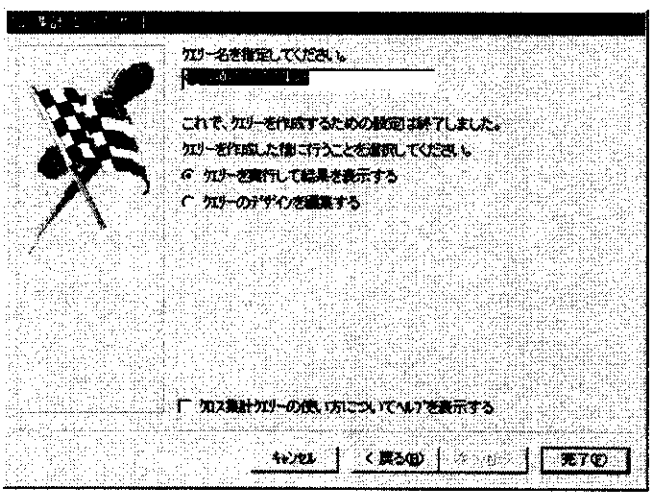
フィールド「No」と集計方法「カウント」を選択し、次へ進みます。



集計方法では、合計を選択しないで下さい。

カウントが該当する件数を数えていくのに対して、合計は該当するデータの数値を合計します。

⑦クエリー名を指定して完了（F）ボタンを押します。



⑧クロス集計の結果が表示されます！ (@_@)

	身近	集計値: No	どちらとも言えな	当らない	当る
▶	[redacted]	4	1	1	2
	いる	3		1	2

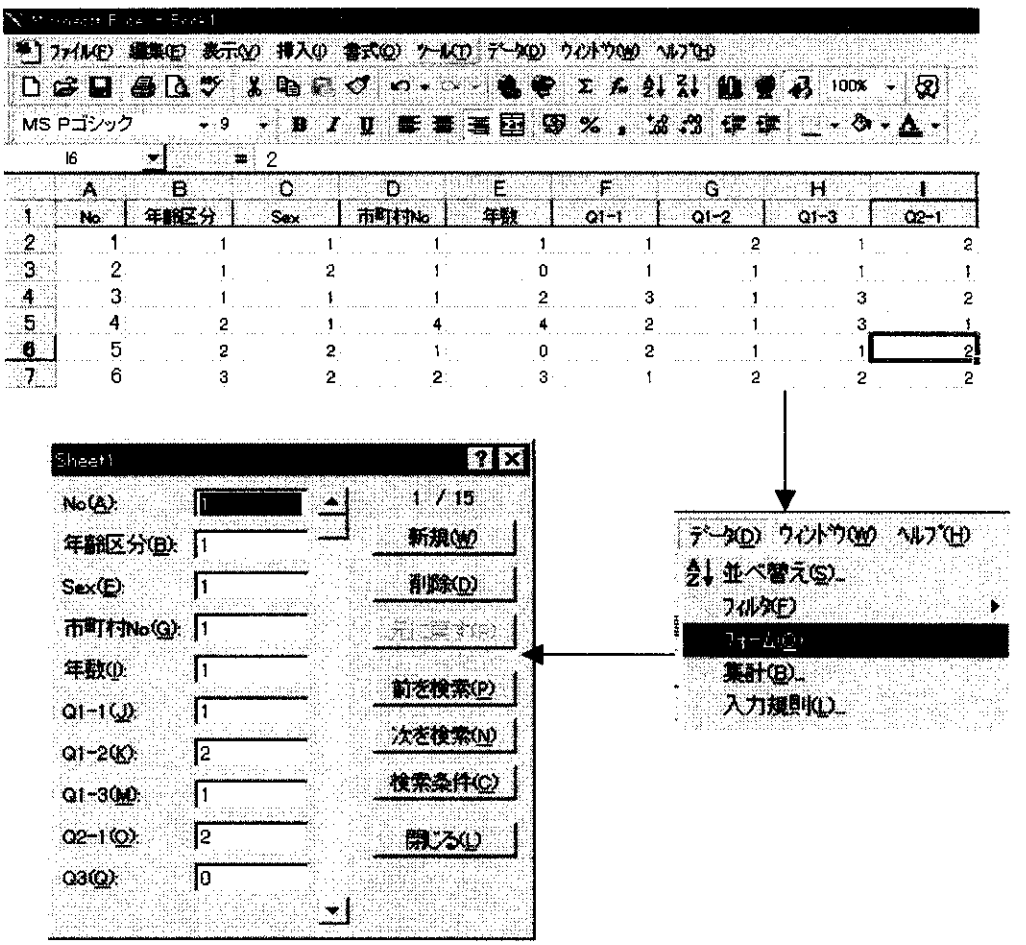
5　エクセル９７の機能紹介

これまでは、アクセス９７でのフォームやクエリーの作成方法を紹介しましたが、実はエクセルでも簡単なデータベース機能が用意されています。

そこで、おまけのページとしていくつか便利な機能を紹介します。

（１）フォーム機能

データの入力用に既存の形式で簡易フォームが作成されます。



ファイルメニューのデータ（D）－フォームを選択するだけです。

フォームで入力をすると自動的にデータの追加（新規）や変更ができます。

（2）フィルタ機能

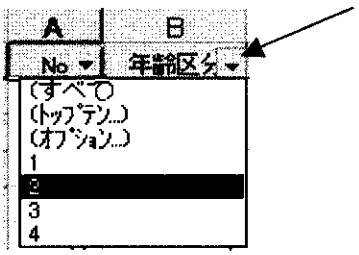
これもエクセルでのデータベース機能の一種です。
指定する条件のデータを表示する機能です。

	A	B	C	D	E	F	G	H
1	No	年齢区分	Sex	市町村No	年数	Q1-1	Q1-2	Q1-3
2	1	1	1	1	1	1	2	1
3	2	1	2	1	0	1	1	1
4	3	1	1	1	2	3	1	3
5	4	2	1	4	4	2	1	3
6	5	2	2	1	0	2	1	1
7	6	3	2	2	3	1	2	2
8	7	2	1	2	0	1	3	2
9	8	3	1	1	1	1	2	1
10	9	1	2	3	10	1	1	1
11	10	3	1	3	2	3	1	3
12	11	2	1	1	4	2	1	3
13	12	2	2	1	0	2	1	1
14	13	2	2	2	2	1	2	2
15	14	4	1	2	5	1	3	2
16	15	2	2	3	10	2	1	2

上の表で年齢区分で２だけを指定すると、下の表のように該当するデータのみが表示されます。

	A	B	C	D	E	F	G	H
1	No	年齢区分	Sex	市町村No	年数	Q1-1	Q1-2	Q1-3
5	4	2	1	4	4	2	1	3
6	5	2	2	1	0	2	1	1
8	7	2	1	2	0	1	3	2
12	11	2	1	1	4	2	1	3
13	12	2	2	1	0	2	1	1
14	13	2	2	2	2	1	2	2
16	15	2	2	3	10	2	1	2

メニューのデータ（D）のフィルタ（F）のオートフィルタをチェックするフィルタが設定されます。
条件指定したい列の項目名の横にあるボタンをクリックすると条件が表示されます。



条件を入力すると上の表のように該当するデータのみが表示されます。

～精神保健に関する調査

この調査で、精神障害についてお聞きします。知的障害（精神薄弱）、老人性痴呆疾患、アルコール依存症、シンナーや覚醒剤などの薬物依存症を除いた精神障害（精神分裂病やうつ病など）についてお答え下さい。

以下の質問について、ご記入あるいは該当する番号に○印をお付け下さい。

問１．精神障害について、あなたの考えに最も近いものはどれですか。

（１）一度精神障害になると治らない。

１．そう思う　　２．そう思わない　　３．どちらともいえない

（２）薬やカウンセリングは、精神障害の治療に有効だ。

１．そう思う　　２．そう思わない　　３．どちらともいえない

（３）精神病院に入院した人でも、信頼できる友人になれる。

１．そう思う　　２．そう思わない　　３．どちらともいえない

問２．あなたの身近に、これまでの精神障害になった人がいますか。

１．いる　　２．いない

「いる」と答えた方にお聞きします。
それはどういう関係の人ですか。該当する全てを選んで下さい。

１．身内　　２．友人　　３．近所の人　　４．職場の人
５．その他（　　）

問３．精神障害者のことで困った時に、相談できる人がいますか。

１．いる　　２．いない

問４．あなたご自身のことについてお聞きします。

（１）あなたの年齢は？　　（１．40歳未満　２．40歳～60歳未満　３．60歳以上）

（２）あなたの性別は？　　（　１ 男性　　２ 女性　）

（３）あなたのお住まいは？（１．A市　２．B市　３．C町　４．D町）

（４）あなたはホームヘルパー・ボランティアを何年間していますか？
（　　）年間

ご協力ありがとうございました。

★★ グループワーク進行表

１０：００～　オリエンテーション
　　・本日の目的
　　・時間配分
１０：１０～　進行係・発表者の決定（ジャンケン）
　　自己紹介
１０：２０～　ＫＪ法（ポストイットにでた意見を記入）

１５分間	「何故、パソコンを使いたいと思いますか。何を期待しますか。」
５分間	同じ問題をくくり、関連あるものはつなげる。
１０分間	グループで作るスローガンを１００字以内でまとめる。
	「３年後、私たちは、こうなりたい！！」

１０：５０～　休憩（１０分）
１１：００～　各グループ発表
１１：１０～　目標スローガン達成のために、今、足りないもの、必要なものは何か。
　　（ポストイットにでた意見を記入）
１１：３０　　終了

★★　準備するもの

新聞紙	４テーブル分	
模造紙	４枚	
ポストイット	８ブロック	
マジック	３色くらい	
OHPシート	４枚	パソコン入力し、スクリーンに出しても良いかもしれません。
OHP		

★★　スタッフ

各グループに書記役としてのみ、入り、小さな意見でも書き留めるようにする。また、意見がそれていたら、さりげなく進行役をフォローする役のスタッフが４名は、必要になりそうです。講師をしてない人がよいかもしれません。

グループワーク名簿

所　属	氏　名	G.No	氏　名	G.No
筑　紫保健所	内村由美子	1		
粕　屋保健所	松本　絵里子	2	吉岡　雅夫	3
朝　倉保健所	莫　美紀	2	後藤　都	1
糸　島保健所	中山　雅彦	4	近藤　くみ子	1
遠　賀保健所	岩本　治也	3	藤原　哲治	4
鞍　手保健所	犬丸　陽子	2	髙島　洋子	2
嘉　穂保健所	大空　仁	3	稲田　清美	1
田　川保健所	占部　秀晴	2	川崎　弥也	4
久留米保健所	田中　博文	3	吉松　綾子	4
三潴支所	松下　隆志	3	大宜見　健二	2
浮羽支所	真子　剛幸	1	西田　郁江	4
八　女保健所	宮本　幸二	4	坂田　郁子	2
山　門保健所	上田　修	3	原田　優美子	1
京　築保健所	田中　浩二	4	野田　利絵	3
築上支所	柴田　和典	1	清永　のり子	2

平成１１年度
保健情報処理研修会

統計の基礎知識
～　分割表を中心として　～

目　次

平成１２年３月８日
福岡県保健環境研究所
情報管理課　篠原志郎

平成１２年３月８日

統計の基礎知識

～　分割表を中心として　～

保健環境研究所　篠原志郎

１．はじめに

アンケート調査データ、あるいは簡単な実験データを用いた統計処理法にカイ二乗（χ^2）検定法がある。方法は同じでありながら独立性の検定、適合度の検定、一様性の検定などと適用内容によって呼び分けられ、最も良く使われる統計手法の一つである。例えば、食中毒の原因調査のように、ある食物摂取の有無と症状の有無との関係を調べるのに、ある項目（この場合食物摂取の有無）と別の項目（この場合症状の有無）のクロス表を作り、その表の各数が偶然に現れた数であるのかどうかを調べる。検定法では、帰無仮説が否定される場合に積極的な意味を持つようになっている。いくつかの例題を通して理解していくことにしよう。

2. 独立性の検定（クロス表の検定）

症状のあり、なしを１つの変数、スープを飲んだ、飲まなかったを別の変数としてクロス表（この場合２×２分割表）を作成する。食中毒関係でいうマスターテーブルである。

表１　観測値

A ＼ B	ｽｰﾌﾟを飲んだ(＋)	ｽｰﾌﾟを飲まなかった	計
症状あり(＋	36	12	48
症状なし(－	16	20	36
計	52	32	84

表２　期待値

A ＼ B	ｽｰﾌﾟを飲んだ(＋)	ｽｰﾌﾟを飲まなかった	計
症状あり(＋	29.7	18.3	48
症状なし(－	22.3	13.7	36
計	52	32	84

期待値は周辺度数（この場合、48,36,52,32）が固定したものとして求められる。A(+)B(+)は 36 である。その期待値は 29.7=48/84×52 である。周辺度数が一定なので、表２のように、その他の期待値 18.3, 22.3, 13.7 は自動的に決まってしまう。

帰無仮説：「症状の有無とスープ飲用とは無関係である」即ち、項目Ａと項目Ｂとは独立である(Independent)という仮説である。この仮説の下では、次の計算値 χ^2_c は自由度１の χ^2 分布に従う。

$$\chi_c^2=\sum\frac{(\text{観測値}-\text{期待値})^2}{\text{期待値}}=\frac{(36-29.7)^2}{29.7}+\frac{(12-18.3)^2}{18.3}+\frac{(16-22.3)^2}{22.3}+\frac{(20-13.7)^2}{13.7}=8.18>6.63=\chi_1^2(0.01)$$

帰無仮説の下では表１のような観測値の現れ方は 1%以下の確率となることが分かる。これは、きわめて希なケースが（１%以下で）起きたといえるが、一般に、仮説を棄却する解釈がとられる。即ち、「スープ飲用と症状とは関係がある」と結論づけられる。

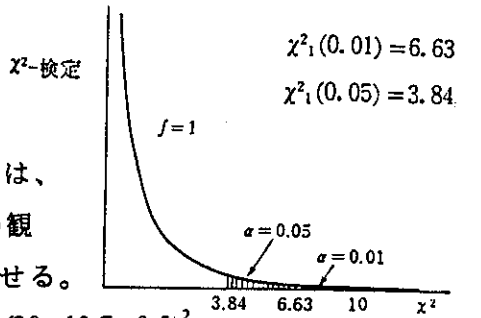


表１でマス目の 36,12,16,20 のどれかが５未満のときは、F.Yates の補正（連続修正）を施すが、それには χ^2_C の観測値と期待値の差が小さくなるように、0.5 だけ増減させる。

$$\chi_c^2=\frac{(36-29.7-0.5)^2}{29.7}+\frac{(12-18.3+0.5)^2}{18.3}+\frac{(16-22.3+0.5)^2}{22.3}+\frac{(20-13.7-0.5)^2}{13.7}=6.93>6.63=\chi_1^2(0.01)$$

マス目が５以上のケースでもこの程度の差であり、連続補正を常に行うと定めておいた方が安全である。

【演習１】

慈善パーティーの昼食後に急性の咽頭炎が発生したので、咽頭炎になった人の食べた食物と咽頭炎にならなかった人の食べた食物が調査された。下の二つの食物、卵サラダとチーズ付マカロニのうち、どちらが咽頭炎の原因として疑わしいか検定しなさい。

	食べた者				食べなかった者			
	発病者	非発病者	計	罹患率	発病者	非発病者	計	罹患率
卵サラダ	38	27	65	58%	3	18	21	14%
チーズ付マカロニ	20	14	34	59%	21	31	52	40%

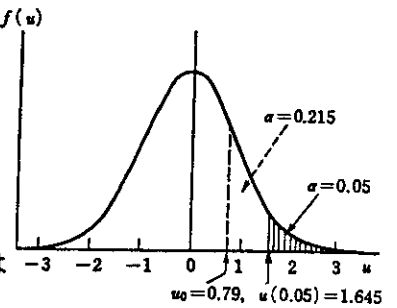
２－１．２項分布の正規分布近似

表１で A(+)B(+)の確率 p_1、A(+)B(−)の確率 p_2 の間に差が見られるだろうかを考える。

$$p_1=r_1/n_1=36/52=0.692,\quad p_2=r_2/n_2=12/32=0.375,\text{仮説}H_0:P_1=P_2=P,\quad A(+)\text{の推定値}p=\frac{r_1+r_2}{n_1+n_2}=48/84$$

$$z=\frac{r_1/n_1-r_2/n_2}{\sqrt{p(1-p)(\frac{1}{n_1}+\frac{1}{n_2})}}=\frac{0.692-0.375}{\sqrt{\frac{48}{84}(1-\frac{48}{84})(\frac{1}{52}+\frac{1}{32})}}=\frac{0.317}{\sqrt{0.571\times0.429\times0.05}}=2.86,\ \text{連続修正のとき}p_1-p_2=0.292$$

として計算する。このとき z=2.638, z＞2.576.
故に、仮説は危険率１％で棄却され、症状は
スープを飲んだ方が有意に高く発生している。



正規近似による確率の算出

２－２．直接確率計算法

R.A.Fisher の正確な確率計算法がある。表１の状態の確率は

$$F_0=\frac{\text{周辺度数の階乗の積}}{(\text{四分割表内度数の階乗 の積})(\text{総数の階乗})}=\frac{48!\times36!\times52!\times32!}{36!\times12!\times16!\times20!\times84!}$$

この状態から一方に偏るケースを計算する。この場合、「スープ飲んだこと」と「症状あり」の関係がより明らかになる方向に四分割表内の数が偏るケースをすべて計算する。即ち、下記の 13 ケースすべてについて計算し、合計した直接確率計算値 $T=\Sigma F_C$ を求める。

	B(+)	B(−)	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−	+	−
A(+)	36	12	37	11	38	10	39	9	40	8	41	7	42	6	43	5	44	4	45	3	46	2	47	1	48	0
A(−)	16	20	15	21	14	22	13	23	12	24	11	25	10	26	9	27	8	28	7	29	6	30	5	31	4	32
	0		1		2		3		4		5		6		7		8		9		10		11		12	

実際、確率を求める場合、F_{i+1} は F_i を使って求められる。いわゆる漸化式である。
以下の計算から約 T=0.00425 である。χ^2 検定と比べるには、T を２倍すればよい。

$$F_1=\frac{16\times12}{37\times21}F_0, F_2=\frac{15\times11}{38\times22}F_1, F_3=\frac{14\times10}{39\times23}F_2, F_4=\frac{13\times9}{40\times24}F_3, \cdots, F_{12}=\frac{5\times1}{48\times32}F_{11}$$

F_0を求めるには、

$$ln(F_0)=\sum_R^{48}ln(R)+\sum_S^{36}ln(S)+\sum_T^{52}ln(T)+\sum_U^{32}ln(U)-(\sum_V^{36}ln(V)+\sum_W^{12}ln(W)+\sum_X^{16}ln(X)+\sum_Y^{20}ln(Y)+\sum_Z^{84}ln(Z))$$

$$=140.6739+95.7197+156.3608+81.5580-(95.7197+19.9872+30.6719+42.3356+291.324)=-5.726, F_0=0.00326$$

$$F_1=0.2471\times0.00326=0.00081, F_2=0.19737\times0.00081=0.00016, \ldots, F_{12}=0.00326\times1.2\times10^{-16}=3.8\times10^{-19}, T=0.00425$$

【演習２】

地域的に甲状腺腫の集積が疑われたので、その地域内の２つの推計に沿って調査を行ったところ、A 水系の住民は 19 人中 4 人に、B 水系の住民では 20 人中 2 人に腺腫が認められた。両水系に沿った住民の間で腺腫の発生率に差があるといえるか？χ_2 検定法と直接確率計算法の両方で検定し比較しなさい。

x ＼ y	甲状腺腺腫（＋）	甲状腺腺腫（－）	計
A水系	4	15	19
B水系	2	18	20
計	6	33	39

２－３．McNemar の検定（同一群による２×２表の場合）

まず、次の例題を考えよう。アレルギー性疾患の患者 30 人に対し、室内塵に対する過敏性反応を２つの方法で調べた。両反応の陽性率に差があるか？

B ＼ A	A反応陽性	A反応陰性	計
B反応陽性	16	1	17
B反応陰性	3	10	13
計	19	11	30

これまでの比較対象は２つの異なる群であった。しかし、この例のように、同一群の時間的前後、ある処置前と処置後を比較したいケースがある。この場合、帰無仮説は A 反応と B 反応で異なった結果がでる確率は同じであると考える。A、B 反応が共に陽性、共に陰性には関心がない。このとき、次の簡単な式で検定できる。

$$\chi_c^2=\frac{(|3-1|-1)^2}{3+1}=0.25<3.84=\chi_1^2(0.05)$$

３．適合度の検定

ある母集団から抽出された標本が理論分布関数に適合しているかどうかを考えよう。仮説は「適合している、即ち、分布は一致している」というものである。例えば、航空郵便用封筒 100 通のランダムな抽出標本の重さ(g)を量り、次表のようにまとめた。表中の組間隔は 100 個のデータをいくつかの等間隔な階級に分け、それに含まれる個数を数える。1.795～1.825 の中央値が組代表 1.81=(1.975+1.825)/2、この幅にデータは２個含まれている。この表から平均値と標準偏差を求める。

$$\text{平均値}\ \bar{x}=\frac{1}{N}\sum f_i\times x_i, N=\sum f_i, \text{標準偏差}\ s=\sqrt{\frac{\sum f_i(x_i-\bar{x})^2}{N-1}}$$

標準化 z は標準正規偏差値を示し、対応する確率が下表の正規分布の確率である。

組間隔(重さg)	組代表	度数	標準正規偏差z	正規分布の確率	期待度数	χ2計算値
1.795 - 1.825	1.81	2	-2.212	0.01348	1.3	0.377
1.825 - 1.855	1.84	3	-1.635	0.05102	3.8	0.168
1.855 - 1.885	1.87	6	-1.058	0.14503	9.4	1.23
1.885 - 1.915	1.90	18	-0.481	0.31526	17	0.059
1.915 - 1.945	1.93	25	0.096	0.53824	22.3	0.327
1.945 - 1.975	1.96	18	0.673	0.74953	21.1	0.455
1.975 - 2.005	1.99	14	1.250	0.89435	14.5	0.017
2.005 - 2.035	2.02	11	1.827	0.96615	7.2	2.006
2.035 - 2.065	2.05	3	2.404	1.00000	3.4	0.047
		100			100	4.686

注)　平均値m＝1.94　自由度6のχ2(0.05)値＝12.59
　　標準偏差s＝0.052　　標準化z=(x-m)/s

χ2値＝4.686＜12.59となり、正規分布することを否定できない。適合度の検定では分布関数のパラメータを推定するので、自由度は推定したパラメータ数だけ余計に減る。

【演習３】

8人の子供がいる家族における男の子の数は２項分布になるという仮説を検定しなさい。

子供の数	度数	累積度数	2項分布	期待値	χ2値
0	2	2	0.0032	1.7	0.053
1	15	17	0.0299	14.2	0.045
2	53	70			
3	106	176			
4	150	326			
5	119	445			
6	64	509			
7	20	529			
8	3	532			
計	532				

平均値x=4.1015、男の子の生まれる確率p=0.5127

４．一様性の検定

地域Ａと地域Ｂにある小売り書店で一定期間中における書籍購読者層の年齢分布を調べた。次表の両地域で年齢構成に差があるかどうかを検定しよう。

	20歳代	30歳代	40歳代	50歳代	計
地域A	46	104	120	28	298
地域B	13	36	54	9	112
計	59	140	174	37	410

これは両地域の書籍購入年齢層が一様に分布しているかを調べるものである。一般に、２×ｒ分割表の場合、これまでの計算式は次式で表される。自由度３のχ²分布の5%

$$\chi_c^2 = \frac{410^2}{298\times 112}\left(\frac{46^2}{59}+\frac{104^2}{140}+\frac{120^2}{174}+\frac{28^2}{37}-\frac{298^2}{410}\right) = 2.389 < 7.815 = \chi_3^2(0.05)$$

7.815より小さいので、分布が一様であるという仮説は否定できない。

５．順序尺度をもつ２×ｒ分割表

非妊娠時における女性のBMI(Body Mass Index)を測定し、それぞれ肥満、正常、やせであった人が出産後に元に復帰した群(<1.0kg)と体重増加した群(≧1.1kg)の表である。

体重	肥満	正常	やせ	計
<1.0kg	12	60	43	115
増加>1.1kg	9	24	10	43
計	21	84	53	158

$$\chi_c^2 = \frac{158^2}{115\times 43}\left(\frac{9^2}{21}+\frac{24^2}{84}+\frac{10^2}{53}-\frac{43^2}{157}\right) = 4.536 < 5.991 = \chi_2^2(0.05)$$

元に復帰群と体重増加群に有意な差は認められない。つまり、BMI でみた体型からは出産後の体重増加は関連がないということである。また、それぞれを比較して、

12	60	72
9	24	33
21	84	105

$\chi_c^2 = 0.997$

12	43	55
9	10	19
21	53	74

$\chi_c^2 = 3.365$

60	43	103
24	10	34
84	53	137

$\chi_c^2 = 1.161$

$\chi_1^2(0.05) = 3.84$

いずれも 3.84 より小さく危険率５％で有意にならない。そこで、次のように整理してみる。

体重	肥満	正常	やせ	計
<1.0kg	12	60	43	115
増加>1.1kg(a	9	24	10	43
計(ni)	21	84	53	158(N)
pi=ai/ni	0.4286	0.2857	0.1887	0.2722(p)
スコアxi	1	0	-1	

やせ、正常、肥満の順に p_i が増加しているようである。そこで、肥満、正常、やせという順位にスコア Xi を与え、Pi の Xi に対する回帰式を考え、その回帰係数 b を検定しよう。回帰式は P＝a＋ｂX,各 Pi は p によって推定され、分散は $p(1-p)/n_i$ である。Pi の分散が各点で異なる場合は重みづけで回帰式を求める。以下の式より、

$$b = \frac{S_{xy}}{S_{xx}},\quad S_{xy} = \sum n_i(p_i-\bar{p})(x_i-\bar{x}) = 3.2844 + 4.4255 = 7.7099,$$

$$S_{xx} = \sum n_i(x_i-\bar{x})^2 = 74,\quad b = 7.7099/74 = 0.104,\quad S_b = \sqrt{\bar{p}(1-\bar{p})/S_{xx}} = 0.006,$$

$z = b/S_b = 0.104/0.006 = 17.33 > 1.96$(正規分布の 5%値)，　回帰式は　$P = 0.272 + 0.104X$である。

回帰係数ｂ=0 という仮説は危険率５％で棄却される。表から自明のように、やせから肥満にかけて体重増加率は増えていると考えられる。即ち、線型性が認められる。

【演習 4】

浸潤度の程度	健康改善から悪化までの変化					計
	顕著	中位	軽微	不変	悪化	
少	11	27	42	53	11	144
多（ai）	7	15	16	13	1	52
計（ni）	18	42	58	66	12	196(N)
pi=ai/ni	0.3889	0.3571	0.2759	0.197	0.0833	0.2653(p)

ハンセン病の患者を初期浸潤度の多少に従って、ある一定の処置期間における健康状態の変化を分類したデータがある。上の解析例に習って線型性を検定しなさい。

一様性の検定、各項目ごとの２×２分割表の χ^2 検定も確かめてみましょう。

11	27	38
7	15	22
18	42	60

11	42	53
7	16	23
18	58	76

11	53	64
7	13	24
18	66	84

11	11	22
7	1	8
18	12	30

$$\chi_c^2 = \frac{196^2}{144\times 52}\left(\frac{7^2}{18}+\frac{15^2}{42}+\frac{16^2}{58}+\frac{13^2}{66}+\frac{1^2}{12}-\frac{52^2}{196}\right) =$$