

トレーニング コース

IBM SPSS Statistics **基礎統計**

演習解答例

IBM、IBMロゴおよびibm.comは、世界の多くの国で登録されたInternational Business Machines Corporationの商標です。

他の製品名およびサービス名等は、それぞれIBMまたは各社の商標である場合があります。

現時点でのIBMの商標リストについては、www.ibm.com/legal/copytrade.shtmlをご確認ください。

この資料は研修用教材として作成したものです。

この資料は2012年9月1日現在のものであり、将来この資料の全体または一部につき変更する場合があります。

本書の内容についてお気づきの点がございましたら、下記までお知らせください。

〒103-8510 東京都中央区日本橋箱崎町19-21

TEL. 0120-105553

日本アイ・ビー・エム株式会社

演習解答例

第2章 データのエラーチェック

1. CENSUS_MISSING.savデータを読み込みます。これは、イギリスの成人を対象として1992年に実施された調査データです。このデータ内にある、次の変数について、エラーがないか調べてみましょう。

変数名	変数の説明
-----	-------

gender	回答者の性別（m=男性、f=女性）
--------	-------------------

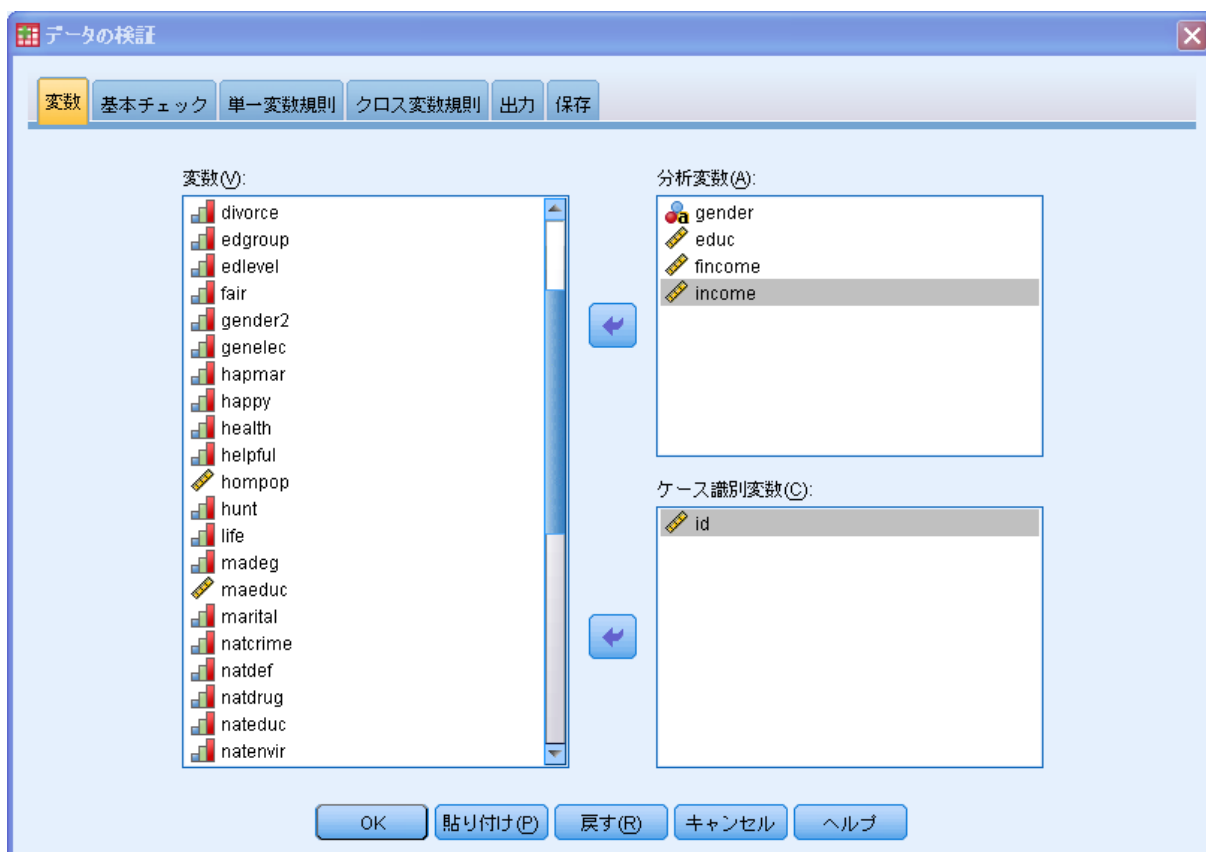
educ	教育を受けた年数
------	----------

income	回答者個人の年収（円に換算済み）
--------	------------------

fincome	回答者の世帯の年収（円に換算済み）
---------	-------------------

Data Preparationオプションの場合

- ☞ ファイルメニューの**開く**の**データ**を選択します。
CENSUS_MISSING.savを選択し、**開く**ボタンをクリックします。
- データメニューの**検証**の**データの検証**を選択します。
- 上記の変数を**分析変数**リストボックスへ移動します。
- id**を**ケース識別変数**リストボックスへ移動します。



図E.1 データの検証ダイアログボックス

- ☞ **単一変数規則**タブをクリックします。
- 規則の定義**ボタンをクリックします。
- 規則の定義**の**名前**テキストボックスに**性別のチェック**と入力します。
- 型を**文字型**に変更します。
- 有効値**のプルダウンリストから**リスト**を選択します。
- 値リストに**f**を入力、**enter**キーを押し、**m**を入力します。



図E.1 データの検証 検証規則の定義ダイアログボックス

- ☞ **新規作成**ボタンをクリックします。
規則の定義の**名前**テキストボックスに**教育年数のチェック**と入力します。
最小値に**6**を、**最大値**に**20**を入力します。
続行ボタンをクリックします。



図E.2 設定後のデータの検証ダイアログボックス

- ☞ 規則リストの**性別のチェック**の適用をチェックします。
- 分析変数リストの**educ**を選択します。
- 規則リストの**教育年数のチェック**の適用をチェックします。

データの検証

変数

基本チェック

単一変数規則

クロス変数規則

出力

保存

変数に規則を適用するには、変数を選択し、1つ以上の規則をチェックにしてください。

分析変数(A):

変数	分布	最小値	最大値	規則
gender			x	1
educ		0	20	1
fincome		425,000	12,342,0...	0
income		408,000	11,594,0...	0

規則(U):

適用	名前
☒	教育年数チェック

表示(D):

すべての変数

スキャンされたケース: 379

規則の定義(E)...

変数の分布

☒ スキャンするケースの数を制限(L) ケース(C): 5000

再スキャン(S)

スキャンするケース数を制限しても、検証されるケース数には影響しません。

OK

貼り付け(P)

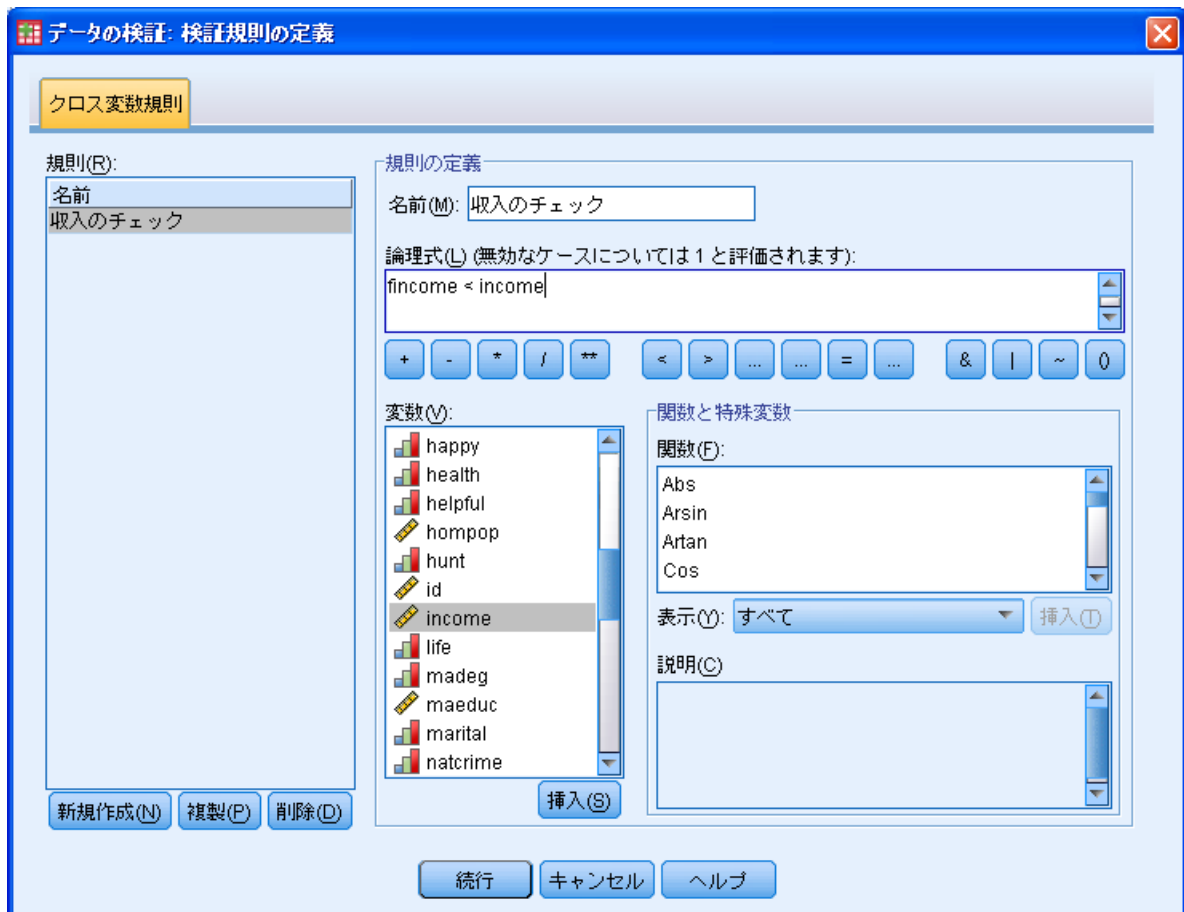
戻す(R)

キャンセル

ヘルプ

図E.3 設定後のデータの検証ダイアログボックス

- ☞ **クロス変数規則** タブをクリックします。
- 規則の定義** ボタンをクリックします。
- 規則の定義** の **名前** テキストボックスに **収入のチェック** と入力します。
- 論理式** に **fincome < income** と入力します。



図E.4 設定後のデータの検証ダイアログボックス

- ☞ **続行** ボタンをクリックします。

- ☞ **保存**タブをクリックします。
- 集計変数**リスト内の**検証規則違反**チェックボックスをクリックします。
- すべての検証規則違反を記録する指示変数を保存する**チェックボックスをクリックします。

データの検証

変数 基本チェック 単一変数規則 クロス変数規則 出力 **保存**

集計変数(S):

説明	保存	名前
空ケースの指示変数	☐	空ケース
重複 ID のグループ	☐	重複IDのグループ
不完全な ID 指示変数	☐	不完全なID
検証規則違反 (合計度数)	☒	検証規則違反

☐ 既存の集計変数を置き換える(E)

☒ **すべての検証規則違反を記録する指示変数を保存する(I)**

これらの変数は、特定のデータの値または値の組み合わせが検証規則に違反したかどうかを知らせます。

これらの変数は、データのクリーニングと調査を容易にすることができます。ただし、適用される規則の数によって、このオプションはアクティブなデータセットに多くの変数を追加する場合があります。

保存される変数の総合数： 4

OK 貼り付け(P) 戻す(R) キャンセル ヘルプ

図E.5 設定後のデータの検証ダイアログボックス

- ☞ **OK**ボタンをクリックします。

ケース報告書

ケース	検証規則違反		識別子
	単一変数 ^a	変数間	id
3		収入のチェック	3
4	性別チェック (1)		4
5	教育年数チェック (1)		5
10	性別チェック (1)		10
24	性別チェック (1)		24
42	教育年数チェック (1)		42
59	性別チェック (1)		59
134	教育年数チェック (1)		134
267	教育年数チェック (1)		267
297	教育年数チェック (1)		297
311	教育年数チェック (1)		311

a. その規則に違反した変数の数はそれぞれの規則を満たしています。

図E.6 設定後のデータの検証ダイアログボックス

検査を行った変数には、規則違反があることがわかりました。該当するケースに対して、どのように対処するかを検討します。アンケート用紙が保管されていれば、それを確認します。また、SPSSにユーザー定義の欠損値として登録したり、新しい値を割り当てたりするなどしてクリーニングします。

Base Systemの場合

- ☞ ファイルメニューの**開く**の**データ**を選択します。
- CENSUS_MISSING.sav**を選択し、**開く**ボタンをクリックします。
- 分析メニューの**報告書**の**ケースの要約**を選択します。
- 上記の変数を**変数**リストボックスへ移動します。
- ケースの制限は最初の**テキストボックスに**25**と入力します。
- 有効なケースだけを示す**の選択を解除します。
- ケース番号を示す**を選択します。



図E.7 ケースの要約ダイアログボックス

- ☞ **OK**ボタンをクリックします。

ケースの集計^a

	ケース番号	回答者の性別	教育を受けた 年数	世帯全体の年 収 (円)	回答者の年収 (円)
1	1	男性	8	1,598,000	1,122,000
2	2	女性	12	425,000	425,000
3	3	男性	8	2,125,000	3,332,000
4	4	x	8	.	.
5	5	女性	0	1,003,000	884,000
6	6	女性	13	425,000	425,000
7	7	男性	13	476,000	476,000
8	8	女性	8	2,873,000	2,057,000
9	9	男性	15	.	.
10	10	x	14	3,740,000	2,907,000
11	11	女性	12	.	.
12	12	男性	14	2,873,000	2,091,000
13	13	男性	18	5,440,000	5,440,000
14	14	男性	13	.	.
15	15	女性	18	4,454,000	4,029,000
16	16	女性	16	4,131,000	4,131,000
17	17	女性	18	11,322,000	1,666,000
18	18	男性	18	3,077,000	3,077,000
19	19	男性	14	6,851,000	4,505,000
20	20	男性	10	1,088,000	1,088,000
21	21	男性	14	6,477,000	4,114,000
22	22	男性	12	9,860,000	3,502,000
23	23	女性	18	11,016,000	3,876,000
24	24	x	13	.	.
25	25	女性	9	2,839,000	2,057,000
合計	度数	25	25	20	20

a. 最初の 25 のケースに制限されています。

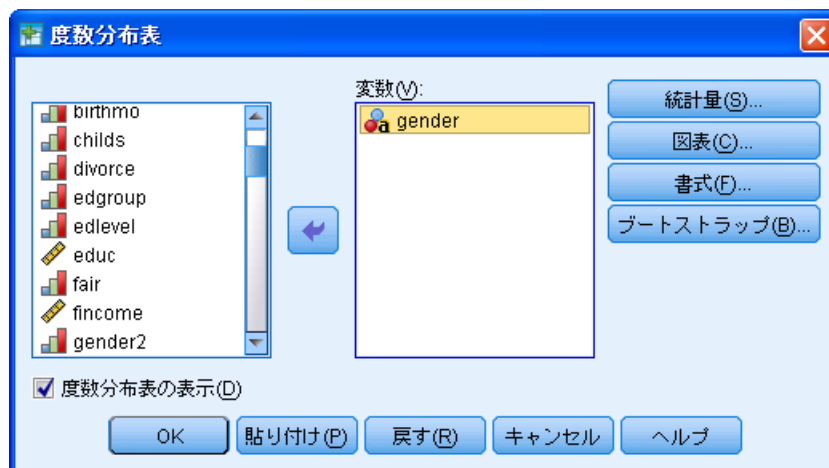
図E.8 ケースの集計テーブル

回答者の性別には、男性のm、女性のf以外のコードがあることがわかります。教育を受けた年数には0という値があります。回答者の年収と世帯全体の年収で、世帯全体の年収が下回っているケースがあります。

この表からおおよそ各変数に生じているエラーを見つけることができました。エラーデータをクリーニングするに当たってのヒントが得られました。

別の方法を使ってエラーを調べてみます。

- ☞ 分析メニューの**記述統計**の**度数分布表**を選択します。
gender（性別）を**変数**リストボックスへ移動します。



図E.9 度数分布表ダイアログボックス

- ☞ **OK** ボタンをクリックします。

回答者の性別				
	度数	パーセント	有効パーセン ト	累積パーセン ト
有効	7	1.8	1.8	1.8
女性	198	52.2	52.2	54.1
男性	170	44.9	44.9	98.9
x	4	1.1	1.1	100.0
合計	379	100.0	100.0	

図E.10 回答者の性別の度数分布表

ケースの集計表からもわかりましたが、男性のm、女性のf以外に空白のデータとxというデータがそれぞれ7ケース、全体の1.8%、4ケース、全体の1.1%あることがわかります。

これらのデータをどのように対処するかを検討します。アンケート用紙が保管されていれば、それを確認します。また、SPSSにユーザー定義の欠損値として登録したり、新しい値を割り当てたりするなどしてクリーニングします。

第3章 中心傾向と散らばり

1. CENSUS.savデータを読み込み、次の変数を見てみましょう。

agewed fincome gender genelec
 hapmar nateduc polviews wordsum

2. 上記の変数の測定の尺度は何でしょう。（測定の尺度には、名義尺度、順序尺度、量的（間隔・比率）尺度があります。）

agewed : 量的 fincome : 量的 gender : 名義 genelec : 名義

hapmar : 順序 nateduc : 順序 polviews : 順序 wordsum : 量的

第4章 データの要約

1. 2つの質的（名義尺度か順序尺度）変数を選択し、**度数分布表**の手続きを使用して、度数分布表の作成や適切なグラフと要約統計量の指定をしましょう。

☞ **ファイル**メニューの**開く**のデータを選択します。
CENSUS.savを選択し、**開く**ボタンをクリックします。
分析メニューの**記述統計**の**度数分布表**を選択します。
gender（性別）、**hapmar**（結婚の幸福度）を**変数**リストボックスに移動します。
統計量ボタンをクリックし、**中心傾向**の**中央値**と**最頻値**を選択します。
続行ボタンをクリックします。
図表ボタンをクリックし、**棒グラフ**を選択します。
続行ボタンをクリックし、**OK**ボタンをクリックします。

統計量		回答者の性別	結婚の幸福度
度数	有効	379	222
	欠損値	0	157
中央値			1.00
最頻値			1

図E.11 genderとhapmarの統計量

統計量のテーブルに中央値と最頻値が表示されます。回答者の性別の中央値や最頻値は表示されていません。これは、回答者の性別をm（男性）、f（女性）の文字型で入力しているためです。文字型のデータは、このような基本的な統計量の算出においても分析の対象外となります（回答者の性別は、名義尺度のデータです。名義尺度における中央値には意味がありませんので算出されてもその値に価値はありません）。結婚の幸福度を見てみましょう。中央値が1.00、最頻値が1となっています。

度数分布表を見ることで、最頻値や中央値を把握することができます。また、棒グラフからも同様です。

回答者の性別

		度数	パーセント	有効パーセント	累積パーセント
有効	女性	206	54.4	54.4	54.4
	男性	173	45.6	45.6	100.0
	合計	379	100.0	100.0	

結婚の幸福度

		度数	パーセント	有効パーセント	累積パーセント
有効	非常に幸せ	147	38.8	66.2	66.2
	やや幸せ	69	18.2	31.1	97.3
	あまり幸せではない	6	1.6	2.7	100.0
	合計	222	58.6	100.0	
欠損値	欠損値	157	41.4		
合計		379	100.0		

図E.12 genderとhapmarの度数分布表

- 2つの量的（スケール）変数を選択し、記述統計の手続きを使用して、要約統計量を求めましょう。同じ変数に対して、度数分布表の手続きからヒストグラムを作成しましょう。その際、度数分布表の表示は不要です。

☞ **分析**メニューの**記述統計**の**記述統計**を選択します。
aged（初婚年齢）、**fincome**（家族の収入）を**変数**リストボックスに移動します。
OKボタンをクリックします。

記述統計量

	度数	最小値	最大値	平均値	標準偏差
初婚年齢	301	16	38	22.26	4.233
世帯全体の年収（円）	351	425000	12342000	3852413.11	2874445.758
有効なケースの数（リストごと）	281				

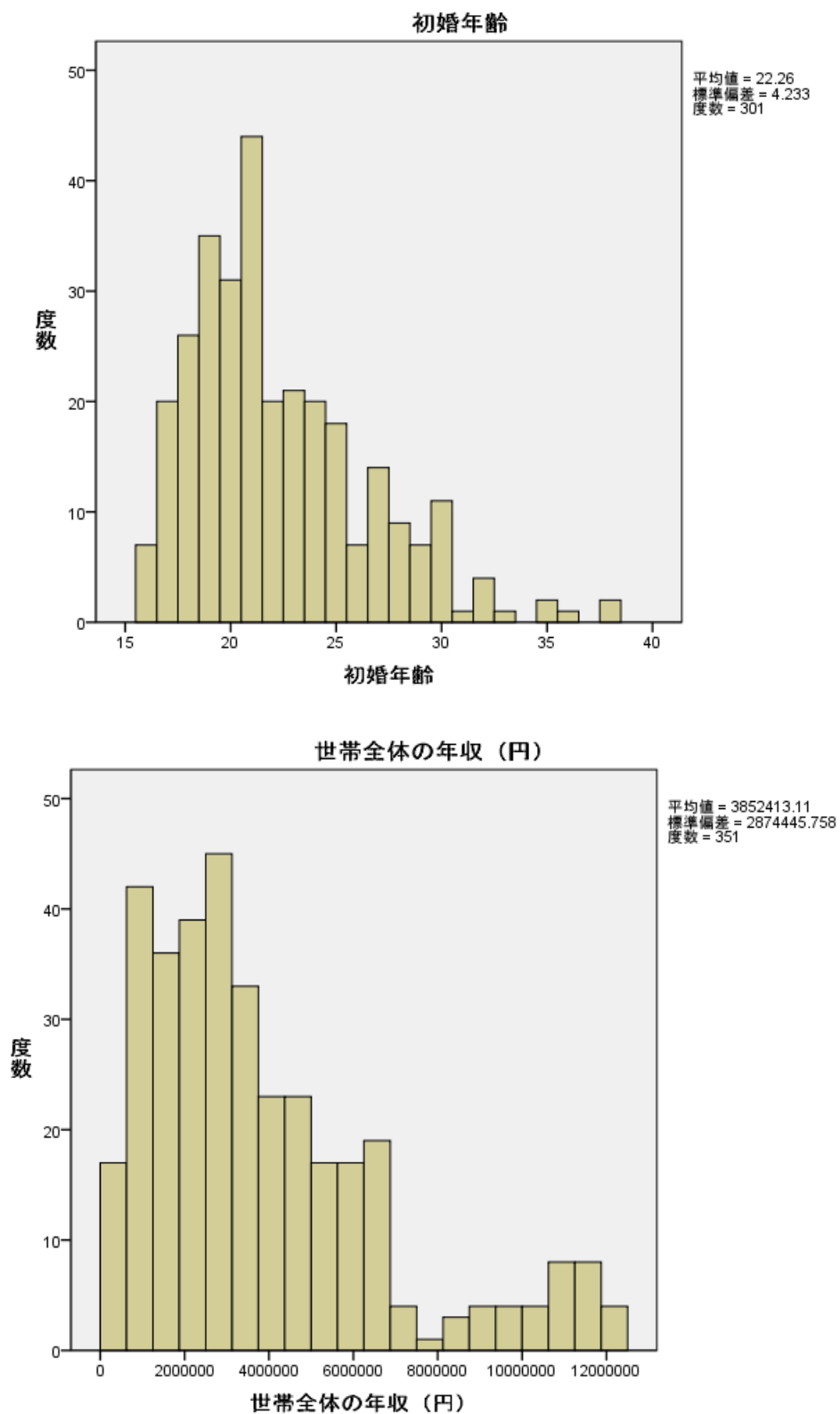
図E.13 agewedとfincomeの記述統計量

- ☞ 分析メニューの**記述統計**の**度数分布表**を選択します。
agewed（初婚年齢）、**fincome**（家族の収入）を**変数**リストボックスに移動します。
統計量ボタンをクリックし、**中心傾向**の**平均値**、**散らばり**の**標準偏差**、**最小値**、**最大値**を選択し、**続行**ボタンをクリックします。
図表ボタンをクリックし、**ヒストグラム**を選択します。
続行ボタンをクリックします。
度数分布表の表示の選択を解除し、**OK**ボタンをクリックします。

統計量

		初婚年齢	世帯全体の年収（円）
度数	有効	301	351
	欠損値	78	28
平均値		22.26	3852413.11
標準偏差		4.233	2874445.758
最小値		16	425000
最大値		38	12342000

図E.14 agewedとfincomeの記述統計量



図E.15 agewedとfincomeのヒストグラム

3. グラフや要約統計量から各変数の特徴を簡単にまとめましょう。

a) 中心傾向：

- ・ gender（回答者の性別）の最頻値は、女性です。
- ・ hapmar（結婚の幸福度）の最頻値と中央値はともに非常に幸せです。
- ・ agedwed（初婚年齢）の平均値は、22.26歳です。
- ・ fincome（家族の収入）の平均値は、3,852,413.11円です。

b) 散らばり：

- ・ agedwed（初婚年齢）の標準偏差は、4.233歳です。最小値は、16歳で最大値は、38歳です。
- ・ fincome（家族の収入）の標準偏差は、2,874,445.758円です。最小値は、425,000円で最大値は、12,342,000円です。

4. 出力をSUMMARY.spvと名前を付けて保存しましょう。

 **ファイルメニューの名前を付けて保存**を選択します。
ファイル名にSUMMARY.spvと入力し、保存ボタンをクリックします。

5. 時間に余裕がある場合は、以下について検討してみましょう。

- ・ 1日のテレビの視聴時間の平均は何時間ですか？：2.9時間
- ・ 世帯内の大人が最も少ないまた、最も多い人数は何人ですか？：1人、7人

 **分析メニューの記述統計の記述統計**を選択します。

（必要に応じて**戻す**ボタンをクリックします。）

tvhourse（1日のテレビの視聴時間）と**adults**（世帯内の大人の人数）を**変数**リストボックスへ移動します。

OKボタンをクリックします。

記述統計量

	度数	最小値	最大値	平均値	標準偏差
1日のTV視聴時間	258	0	13	2.90	2.005
世帯内の成人（18歳以上）の数	379	1	7	1.96	.786
有効なケースの数 (リストごと)	258				

図E.16 tvhourseとadultsの記述統計量

- ・ 初婚年齢では、何歳の人が最も多いですか？ : 21歳
- ・ 42歳以下の人たちは何パーセントいますか？ : 54.1%
- ・ 子供が8人以上いる人（欠損値を含む）たちは何パーセントいますか？ : 0.5%
- ・ 語彙テストでは、何点の人が最も多いですか？ : 6点
- ・ 75%タイルを上回る子供の数は？ : 3人
- ・ データの少なくとも50%は、収入が12,800円を上回りますか？ : はい



分析メニューの**記述統計**の**度数分布表**を選択します。

aged（初婚年齢）、**age**（年齢）、**childs**（子供の数）、**wordsum**（語彙テストの得点）、**income**（収入）を**変数**リストボックスへ移動します。

OKボタンをクリックします。

		初婚年齢			
		度数	パーセント	有効パーセン ト	累積パーセン ト
有効	16	7	1.8	2.3	2.3
	17	20	5.3	6.6	9.0
	18	26	6.9	8.6	17.6
	19	35	9.2	11.6	29.2
	20	31	8.2	10.3	39.5
	21	44	11.6	14.6	54.2
	22	20	5.3	6.6	60.8
	23	21	5.5	7.0	67.8
	24	20	5.3	6.6	74.4
	25	18	4.7	6.0	80.4
	26	7	1.8	2.3	82.7
	27	14	3.7	4.7	87.4
	28	9	2.4	3.0	90.4
	29	7	1.8	2.3	92.7
	30	11	2.9	3.7	96.3
	31	1	.3	.3	96.7
	32	4	1.1	1.3	98.0
	33	1	.3	.3	98.3
	35	2	.5	.7	99.0
	36	1	.3	.3	99.3
	38	2	.5	.7	100.0
	合計	301	79.4	100.0	
欠損値	該当しない	78	20.6		
合計		379	100.0		

図E.17 agewedの度数分布表

		回答者の年齢			
		度数	パーセント	有効パーセント	累積パーセント
有効	18	3	.8	.8	.8
	19	4	1.1	1.1	1.8
	20	7	1.8	1.8	3.7
	21	9	2.4	2.4	6.1
	22	6	1.6	1.6	7.7
	23	6	1.6	1.6	9.2
	24	4	1.1	1.1	10.3
	25	8	2.1	2.1	12.4
	26	8	2.1	2.1	14.5
	27	10	2.6	2.6	17.2
	28	10	2.6	2.6	19.8
	29	14	3.7	3.7	23.5
	30	8	2.1	2.1	25.6
	31	14	3.7	3.7	29.3
	32	10	2.6	2.6	31.9
	33	12	3.2	3.2	35.1
	34	9	2.4	2.4	37.5
	35	6	1.6	1.6	39.1
	36	12	3.2	3.2	42.2
	37	6	1.6	1.6	43.8
	38	11	2.9	2.9	46.7
	39	8	2.1	2.1	48.8
	40	4	1.1	1.1	49.9
	41	8	2.1	2.1	52.0
	42	8	2.1	2.1	54.1
	43	4	1.1	1.1	55.1

図E.18 ageの度数分布表（一部のみ表示）

子供の数		度数	パーセント	有効パーセン ト	累積パーセン ト
有効	0	105	27.7	27.8	27.8
	1	55	14.5	14.6	42.3
	2	101	26.6	26.7	69.0
	3	62	16.4	16.4	85.4
	4	22	5.8	5.8	91.3
	5	15	4.0	4.0	95.2
	6	9	2.4	2.4	97.6
	7	7	1.8	1.9	99.5
	8人以上	2	.5	.5	100.0
	合計	378	99.7	100.0	
欠損値	該当しない	1	.3		
合計		379	100.0		

図E.19 childsの度数分布表

語彙テストでの収集単語数		度数	パーセント	有効パーセン ト	累積パーセン ト
有効	0	1	.3	.4	.4
	1	2	.5	.8	1.2
	2	12	3.2	4.7	5.9
	3	13	3.4	5.1	11.1
	4	24	6.3	9.5	20.6
	5	39	10.3	15.4	36.0
	6	58	15.3	22.9	58.9
	7	35	9.2	13.8	72.7
	8	31	8.2	12.3	85.0
	9	22	5.8	8.7	93.7
	10	16	4.2	6.3	100.0
	合計	253	66.8	100.0	
欠損値	欠損値	126	33.2		
合計		379	100.0		

図E.20 wordsumの度数分布表

1938000	2	.5	.6	45.0
1972000	1	.3	.3	45.3
1989000	2	.5	.6	45.9
2006000	2	.5	.6	46.4
2023000	1	.3	.3	46.7
2040000	1	.3	.3	47.0
2057000	3	.8	.9	47.9
2074000	1	.3	.3	48.1
2091000	3	.8	.9	49.0
2125000	1	.3	.3	49.3
2142000	2	.5	.6	49.9
2176000	2	.5	.6	50.4
2193000	1	.3	.3	50.7
2210000	3	.8	.9	51.6
2227000	2	.5	.6	52.1
2278000	2	.5	.6	52.7

図21 incomeの度数分布表（一部のみ表示）

第5章 標本に基づく母集団の推定

1. 2つの量的（スケール）変数を選択し、**度数分布表**と**記述統計**の手続きを使用して、データの分布状態を評価しましょう。また、正規分布しているかどうかにも検討してみましょう。



ファイルメニューの開くのデータを選択します。

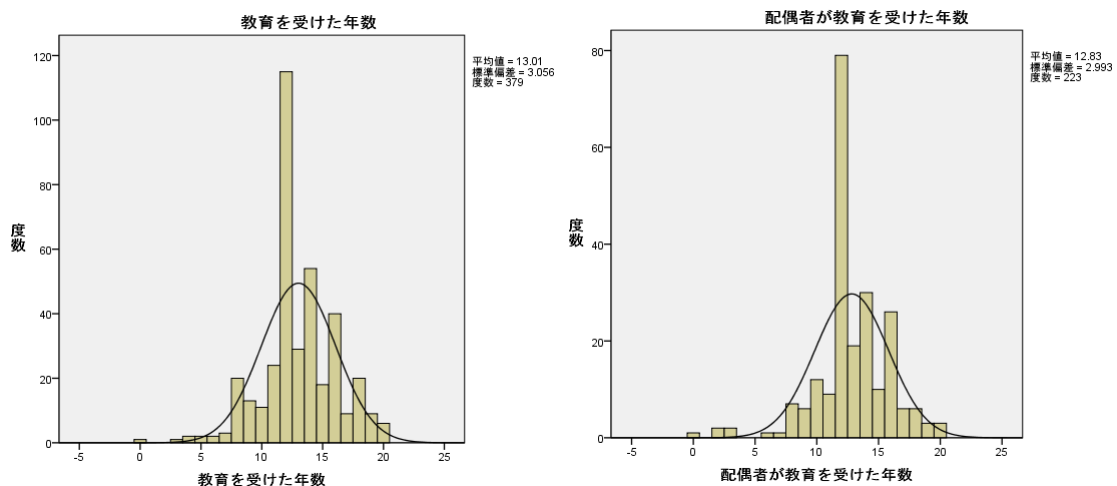
CENSUS.savを選択し、**開く**ボタンをクリックします。

分析メニューの記述統計の度数分布表を選択します。

educ（回答者の教育年数）、**speduc**（配偶者の教育年数）を**変数**リストボックスに移動します。

図表ボタンをクリックし、**ヒストグラム**を選択します。

続行ボタンをクリック、**OK**ボタンをクリックします。



図E.22 educとspeducの正規曲線付きヒストグラム

- ☞ 分析メニューの**記述統計**の**記述統計**を選択します。
educ（回答者の教育年数）、**speduc**（配偶者の教育年数）を**変数**リストボックスに移動します。
オプションボタンをクリックし、**標準誤差**を選択します。
続行ボタンをクリックし、**OK**ボタンをクリックします。

記述統計量

	度数	最小値	最大値	平均値		標準偏差
	統計量	統計量	統計量	統計量	標準誤差	統計量
教育を受けた年数	379	0	20	13.01	.157	3.056
配偶者が教育を受けた年数	223	0	20	12.83	.200	2.993
有効なケースの数 (リストごと)	223					

図E.23 educとspeducの記述統計量

ヒストグラムを見ると多くのケースが12年に集中していて、正規分布していないことがわかります。

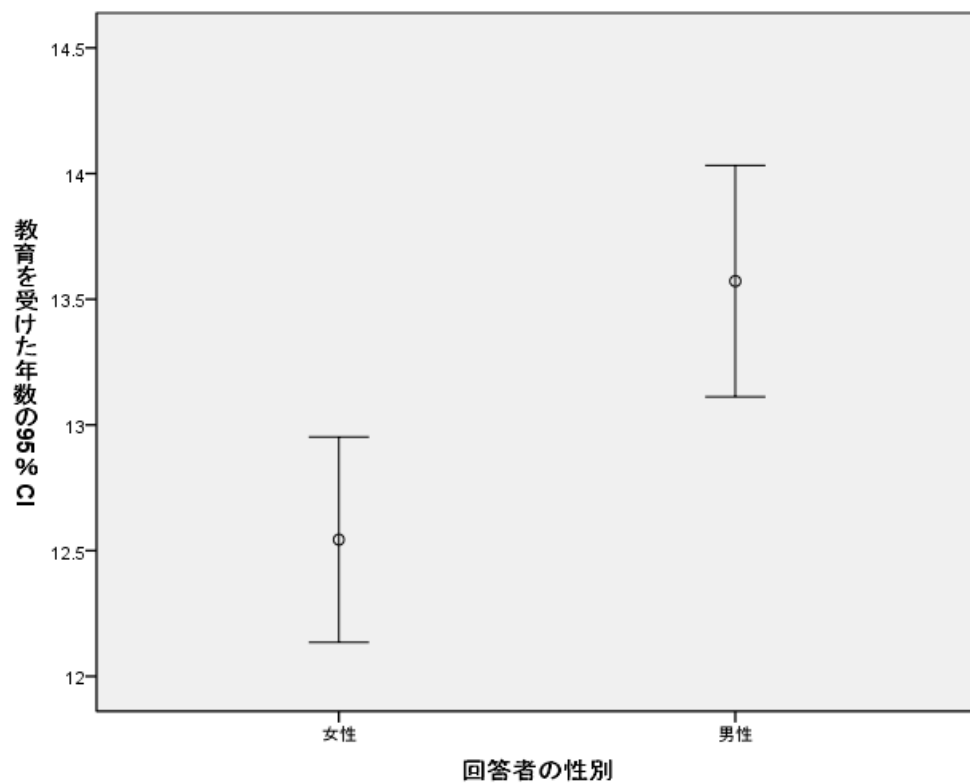
- 上記で選択した変数の標準誤差から、母集団における95%信頼区間を算出しましょう。

educ（回答者の教育年数） $0.157 \times 1.96 \pm 13.01 = 12.702 \sim 13.318$

speduc（配偶者の教育年数） $0.2 \times 1.96 \pm 12.83 = 12.438 \sim 13.222$

- 上記で選択した変数の1つと**gender**（性別）とのエラーバーを作成しましょう。男性と女性で平均値に違いはありますか？2つグループの平均値の信頼区間に違いはありますか？他の質的変数を使用してエラーバーを作成し、違いがあるか調べてみましょう。

- ☞ **グラフメニューのレガシーダイアログのエラーバー**を選択します。
定義ボタンをクリックします。
educ（回答者の教育年数）を**変数**ボックスへ、**gender**（性別）を**カテゴリ軸**ボックスへ移動します。
OKボタンをクリックします。



図E.24 性別による回答者の教育年数のエラーバー

性別による回答者の教育年数のエラーバーは、重なる部分がありません。これは、母集団において男性と女性の教育年数の平均値は、95%の確率で異なる可能性があることを示します。

第6章 データの分布

1. 量的（スケール）変数を選択し、探索的分析の手続きを使用して、ヒストグラム、幹葉図、箱ひげ図、分布の要約統計量を作成し、変数の特徴を簡単にまとめましょう。

☞ **ファイル**メニューの**開く**のデータを選択します。
CENSUS.savを選択し、**開く**ボタンをクリックします。
分析メニューの**記述統計**の**探索的**を選択します。
income（収入）を**従属変数**リストボックスに移動します。
作図ボタンをクリックし、**記述統計**の**ヒストグラム**を選択します。
続行ボタンをクリックします。
オプションボタンをクリックし、**ペアごとに除外**を選択します。
続行ボタンをクリックし、**OK**ボタンをクリックします。

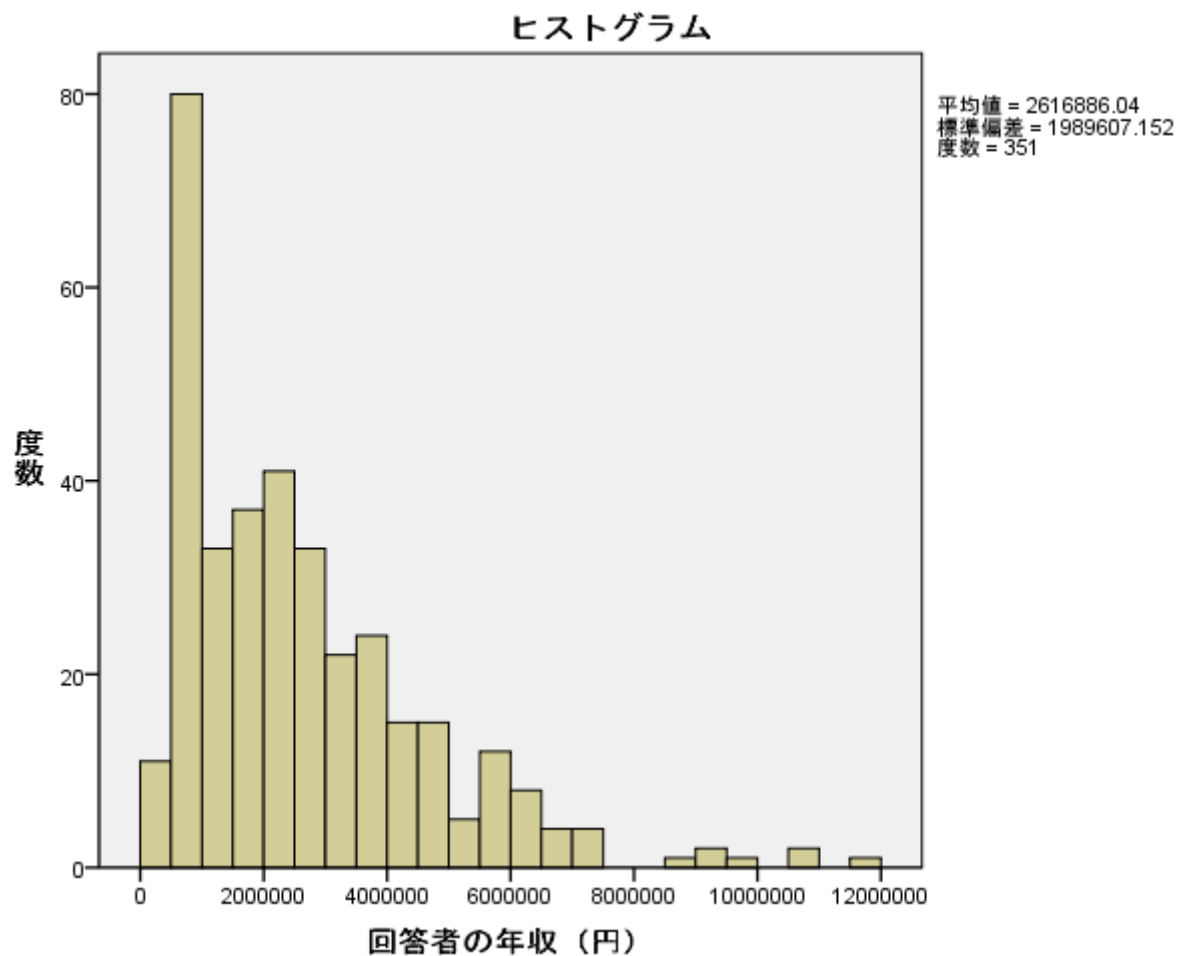
- a) 記述統計
- b) 5%トリム平均
- c) 95%信頼区間

記述統計

			統計量	標準誤差
回答者の年収（円）	平均値		2616886.04	106197.373
	平均値の 95% 信頼区間	下限	2408020.76	
		上限	2825751.32	
	5%トリム平均		2428034.82	
	中央値		2176000.00	
	分散		3.959E12	
	標準偏差		1989607.152	
	最小値		408000	
	最大値		11594000	
	範囲		11186000	
	4分位範囲		2669000	
	歪度		1.466	.130
	尖度		2.785	.260

図E.25 incomeの記述統計

a) ヒストグラムと幹葉図



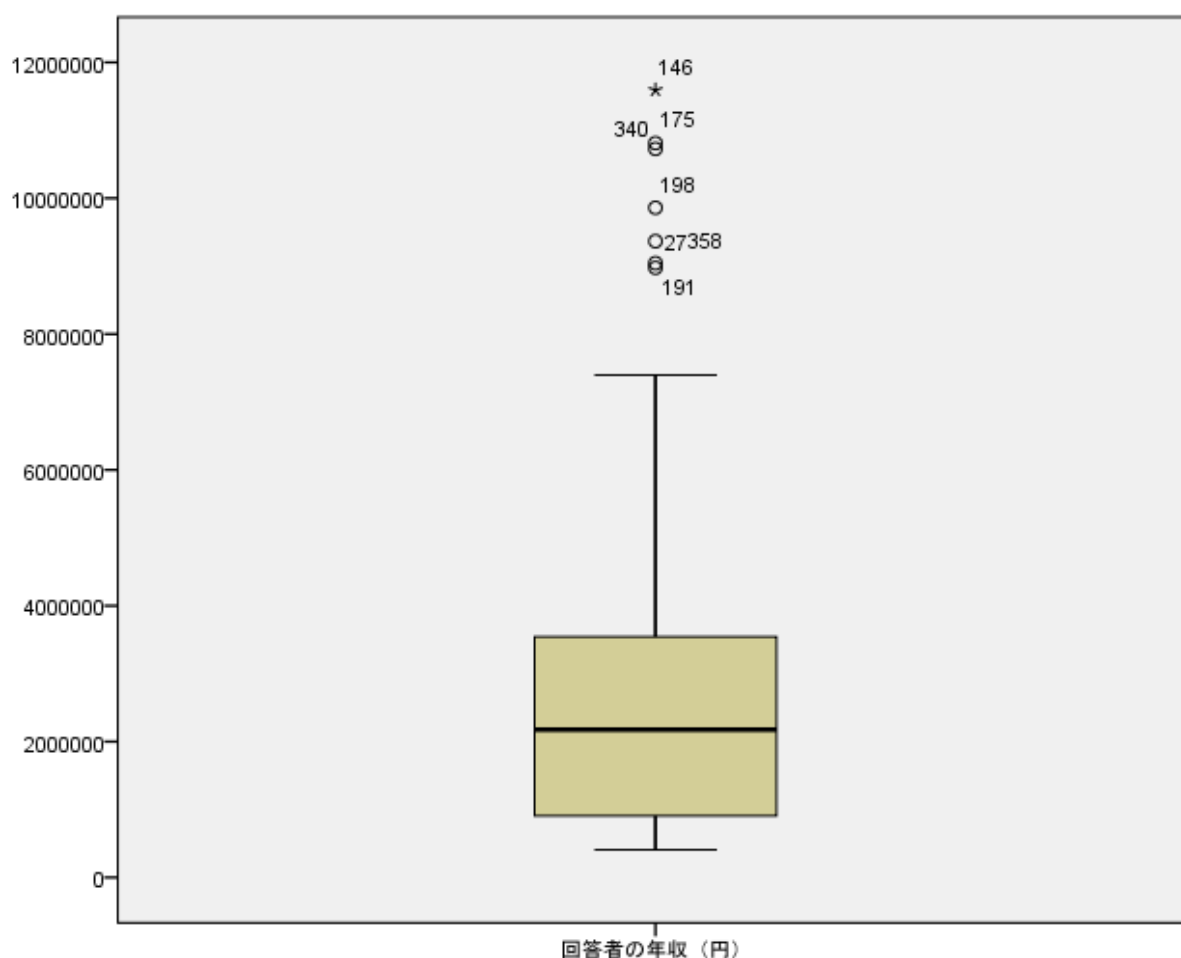
回答者の年収（円） Stem-and-Leaf Plot

```
Frequency      Stem & Leaf
 11.00         0 . 4444444444
 80.00         0 . 555555555555555556666666666677777777777777778888888888888888888999
 33.00         1 . 00000000111111111123333333334444
 37.00         1 . 5666666666677777777788889999999999
 41.00         2 . 00000000000111112222222222333334444444
 33.00         2 . 555666666677777777788888899999
 22.00         3 . 0000111122222333444444
 24.00         3 . 555555555566677888999999
 15.00         4 . 000011111223444
 15.00         4 . 555556666777779
   5.00         5 . 03444
 12.00         5 . 555556666788
   8.00         6 . 01123344
   4.00         6 . 6899
   4.00         7 . 2233
   7.00 Extremes    (>=8976000)

Stem width:    1000000
Each leaf:     1 case(s)
```

図E.26 incomeのヒストグラムと幹葉図

b) 箱ひげ図



図E.27 incomeの箱ひげ図

回答者の収入は、408,000円から11,594,000円の範囲内にあり、平均値は2,616,886.04円です。5%トリム平均は、2,428,034.82円で平均値よりも若干低くなっています。他の高額収入者がいる可能性があると考えられる。中央値と平均値を比較すると中央値が2,176,000円で平均値よりも低いことから、分布は左に歪んでいることが想定できる。平均値の95%信頼区間が、2,408,020.76から2,825,751.32円です。ヒストグラムや幹葉図を見ると、分布が値の低い方に歪んでいることがわかる。箱ひげ図を見ると収入が高い方に分布がのびていることがわかる。

収入の箱ひげ図から、収入が高い方に外れ値や極値があることがわかる。

第7章 質的データ間の関係性：カイ 2 乗検定

この演習では、**クロス集計表**の手続きを使用して2つの質的データ間の関係を探索します。

1. 興味のある質的変数を選択し、**gender**（性別）との関係を調べてみましょう。
意味のある仮説をたて、カイ 2 乗検定が可能な変数を選択してください。

edgroup（教育水準）と**gender**（性別）の関係を調べてみます。

2. 帰無仮説と対立仮説を書きとめておきましょう。

帰無仮説 (H_0) : 教育水準と性別は独立である（関係がない）。

対立仮説 (H_1) : 教育水準と性別は独立ではない（関係がある）。

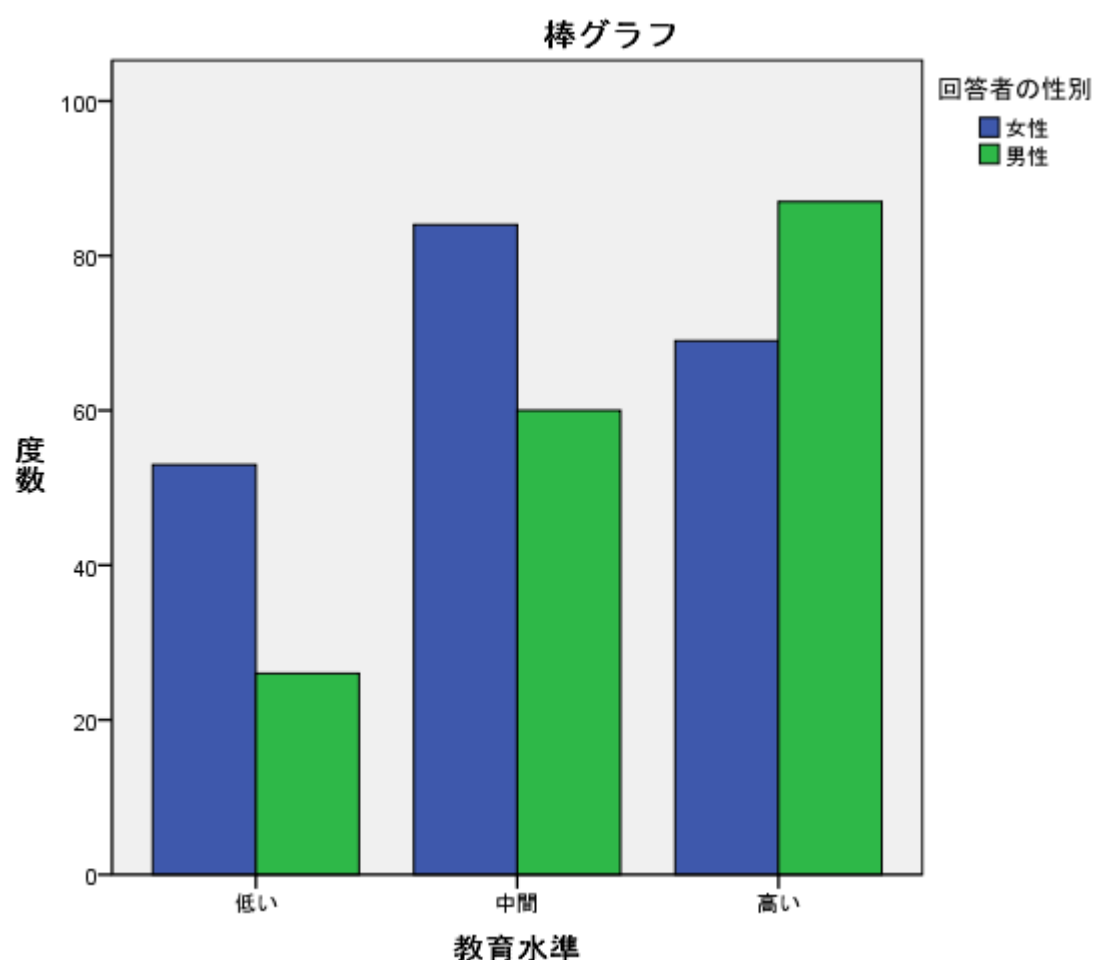
3. これら2つの変数でクロス集計表を作成してみましょう。各セルに度数と適切なパーセンテージを追加し、クラスタ棒グラフの作成を選択しましょう。

☞ **ファイル**メニューの**開く**のデータを選択します。
CENSUS.savを選択し、**開く**ボタンをクリックします。
分析メニューの**記述統計**の**クロス集計表**を選択します。
edgroup（教育水準）を**行**リストボックスに、**gender**（性別）を**列**リストボックスに移動します。
セルボタンをクリックし、**パーセンテージ**の**列**を選択します。
続行ボタンをクリックします。
クラスタ棒グラフの表示を選択します。
OKボタンをクリックします。

教育水準と回答者の性別のクロス表

			回答者の性別		合計
			女性	男性	
教育水準	低い	度数	53	26	79
		回答者の性別の %	25.7%	15.0%	20.8%
	中間	度数	84	60	144
		回答者の性別の %	40.8%	34.7%	38.0%
	高い	度数	69	87	156
		回答者の性別の %	33.5%	50.3%	41.2%
合計	度数		206	173	379
	回答者の性別の %		100.0%	100.0%	100.0%

図E.28 genderとedgroupのクロス集計表



図E.29 genderとedgroupのクラスタ棒グラフ

4. 3.で作成したクロス集計表から、カイ 2 乗検定を実行したときに 2 つの変数間に関係があると判断されるかどうかを考えてみましょう。この検定の結果は信頼できると思いますか。

データの総数が379ケース、セル数が 2×3 の6セルで（1セル平均60ケース）であることから、検定の結果はおそらく信頼できるものと思われます。

5. カイ 2 乗を指定して、クロス集計表の手続きをもう一度実行しましょう。カイ 2 乗検定の結論を導いて、書きとめてください。

☞ **ダイアログのリコールセルボタンをクリックし、クロス集計表を選択します。**
統計量ボタンをクリックし、カイ2乗を選択します。
続行ボタンをクリックします。
OKボタンをクリックします。

カイ 2 乗検定			
	値	自由度	漸近有意確率 (両側)
Pearson のカイ 2 乗	12.526 ^a	2	.002
尤度比	12.640	2	.002
有効なケースの数	379		

a. 0 セル (.0%) は期待度数が 5 未満です。最小期待度数は 36.06 です。

図E.30 カイ 2 乗検定の結果

漸近有意確率（両側）の値が0.002です。

有意確率は、母集団において教育水準と性別は独立である（関係がない）と仮定した場合、その確率は、0.002（0.2%）であるということです。つまり、これはそのようなことがほとんどないことを示します。教育水準と性別は独立である（関係がない）という仮説は0.05（5%）水準で棄却されます。教育水準と性別は統計的に有意であるといえます。

6. カイ 2 乗値に対するクロス集計表の各セルの寄与率を評価するために、セル中に適切な情報を表示しましょう。最大の寄与率を示しているのは、どのセルですか。

☞ **ダイアログのリコール** ボタンをクリックし、**クロス集計表** を選択します。
セル ボタンをクリックし、**残差の標準化** を選択します。
続行 ボタンをクリックし、**OK** ボタンをクリックします。

教育水準と回答者の性別のクロス表

			回答者の性別		合計
			女性	男性	
教育水準	低い	度数	53	26	79
		回答者の性別の %	25.7%	15.0%	20.8%
		標準化残差	1.5	-1.7	
	中間	度数	84	60	144
		回答者の性別の %	40.8%	34.7%	38.0%
		標準化残差	.6	-.7	
	高い	度数	69	87	156
		回答者の性別の %	33.5%	50.3%	41.2%
		標準化残差	-1.7	1.9	
合計	度数	206	173	379	
	回答者の性別の %	100.0%	100.0%	100.0%	

図E.31 標準化残差を追加したクロス集計表

男性の教育水準が高い人が、標準化残差が最大（1.9）のため、先の統計的に有意である結論と導くにあたって、最も寄与していると考えられます。

7. 量的データに対して、連続数のカテゴリ化を使ってグループ化を行い、カイ 2 乗検定を実行してみましょう。

☞ **ダイアログのリコールセル**ボタンをクリックし、**クロス集計表**を選択します。
戻るボタンをクリックします。
income (収入) を**行**リストボックスに、**gender** (性別) を**列**リストボックスに移動します。
セルボタンをクリックし、**パーセンテージの列**を選択します。
続行ボタンをクリックし、**OK**ボタンをクリックします。

表示されたクロス集計表は、非常に大きく冗長的なものとなっています。場合によってはこのような表も役立ちますが、一般的にクロス集計表は、質的データ同士の関係を調べることに適しています。収入や年齢などの量的なデータ同士や量的データとの質的データの組み合わせによる関係を調べるには不適切です。ここでは、収入をグループ化して性別との関係を調べてみます。

収入をいくつかのグループに分割します。特に決まりはありませんが、分析を行うことを考慮して各グループがほぼ同数になるようにグループ化します。

☞ **変換**メニューの**連続変数のカテゴリ化**を選択します。
income (収入) を**ビン分割する変数**リストボックスに移動し、**続行**ボタンをクリックします。
スキャンされた変数のリストの**income**を選択します。
ビン分割する変数テキストボックスに、**grpincome**を入力します。
分割点の作成ボタンをクリックします。
スキャンされたケースに基づく、等しいパーセンタイルを選択します。
分割点の数テキストボックスに**3**を入力し、**適用**ボタンをクリックします。
ラベルの作成ボタンをクリックします。
OKボタンをクリックします。
ビン指定により、変数が1個作成されます。のメッセージが表示されるので、**OK**ボタンをクリックします。

データエディターの一番右側の列に、新しく **grpincome** 変数が作成されます。この変数を利用して、性別と収入のグループとの関係を調べます。

以降の手順は、1~6.を参照してください。

回答者の年収（円）（ビン分割済み）と回答者の性別のクロス表

			回答者の性別		合計
			女性	男性	
回答者の年収（円）（ビン分割済み）	<= 884000	度数	63	25	88
		回答者の性別の %	33.2%	15.5%	25.1%
		標準化残差	2.2	-2.4	
	884001 - 2176000	度数	55	34	89
		回答者の性別の %	28.9%	21.1%	25.4%
		標準化残差	1.0	-1.1	
	2176001 - 3553000	度数	43	44	87
		回答者の性別の %	22.6%	27.3%	24.8%
		標準化残差	-.6	.6	
	3553001+	度数	29	58	87
		回答者の性別の %	15.3%	36.0%	24.8%
		標準化残差	-2.6	2.9	
合計	度数	190	161	351	
	回答者の性別の %	100.0%	100.0%	100.0%	

図E.32 genderとgrpincomeのクロス集計表

カイ 2 乗検定

	値	自由度	漸近有意確率 (両側)
Pearson のカイ 2 乗	28.843 ^a	3	.000
尤度比	29.430	3	.000
有効なケースの数	351		

a. 0 セル (.0%) は期待度数が 5 未満です。最小期待度数は 39.91 です。

図E.33 genderとgrpincomeのカイ 2 乗検定の結果

カイ 2 乗検定の結果、性別と年収グループは統計的に有意であることがわかります。寄与率が高いグループは、男性の収入が最も高いグループ（標準化残差：2.9）です。

8. クロス集計表の手続きで制御変数として3つ目の質的変数を指定し、特に注意すべき交互作用効果を検討してみましょう。

☞ **ダイアログのリコール** ボタンをクリックし、**クロス集計表** を選択します。
health (健康状態) を層変数リストボックスへ移動し、**OK** ボタンをクリックします。

カイ 2 乗検定				
健康状態		値	自由度	漸近有意確率 (両側)
非常に良好	Pearson のカイ 2 乗	5.399 ^a	3	.145
	尤度比	5.487	3	.139
	有効なケースの数	82		
良好	Pearson のカイ 2 乗	11.219 ^b	3	.011
	尤度比	11.590	3	.009
	有効なケースの数	151		

a. 0 セル (.0%) は期待度数が 5 未満です。最小期待度数は 7.13 です。

b. 0 セル (.0%) は期待度数が 5 未満です。最小期待度数は 13.72 です。

図E.34 healthを層変数に追加した、カイ 2 乗検定の結果

健康状態が非常に良好では、性別と年収グループは統計的に有意な差はありません。しかし、良好では、有意な差があります。つまり、健康状態により結果に違いがあります。このことから、健康状態による交互作用が発見できたことになります。

また、コメントから期待度数が5未満のセルが20%を超えておらず、最小期待度数が1以上なのでこの結果は信頼できるものです。

第8章 量的データ間の関係性：相関

この演習では、**散布図**と**相関**の手続きを使用して量的データ間の関係を探索します。

1. 興味のある2つの量的変数を選択し、意味のある仮説をたて、関係性を調べてみましょう。

educ（教育年数）とincome（収入）の関係を調べてみます。

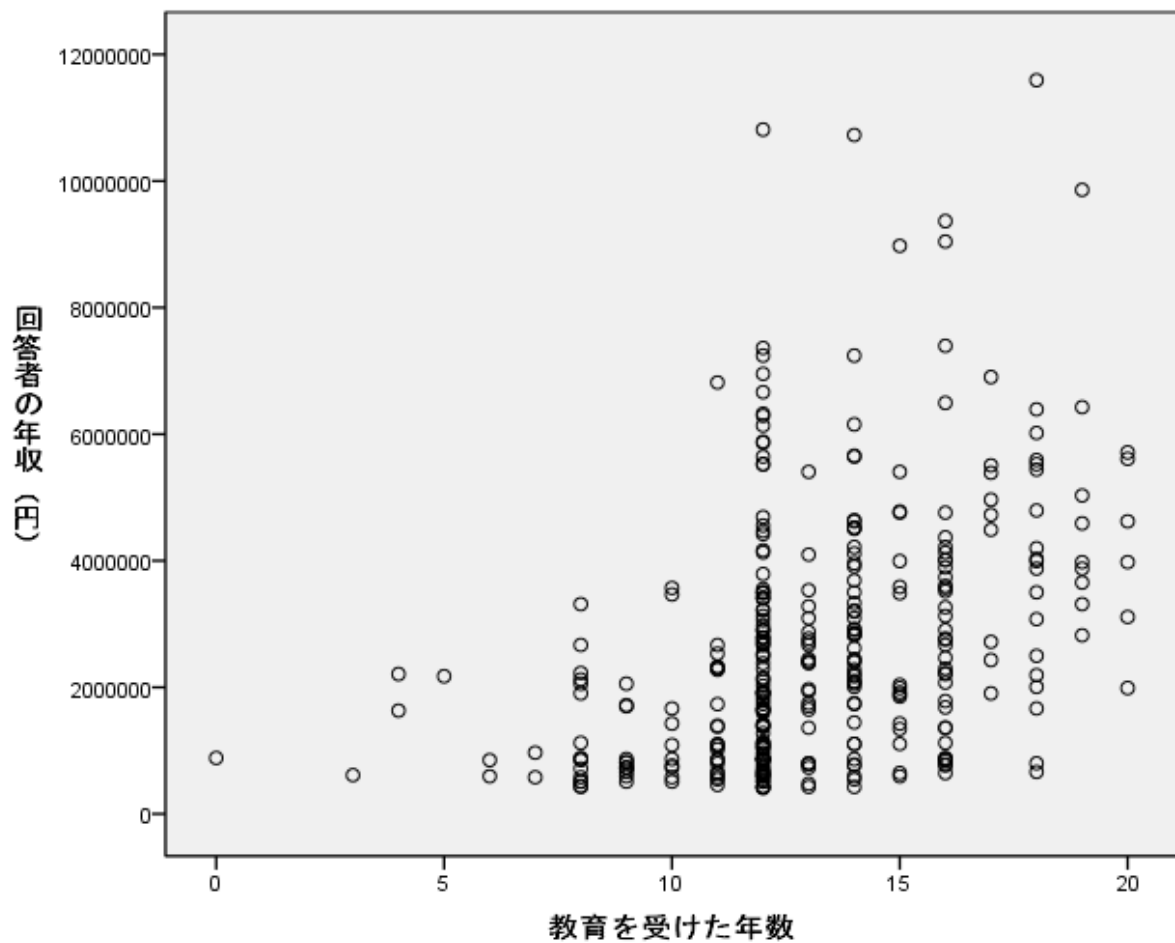
2. 帰無仮説と対立仮説を書きとめておきましょう。

帰無仮説（ H_0 ）：教育年数と収入には線型の関係がない

対立仮説（ H_1 ）：教育年数と収入には線型の関係がある

3. 2つの変数の散布図を作成しましょう。データの傾向やパターン（たとえば、線型の関係）を識別することができますか。何らかの関係が見つかった場合、それは正/負どちらのですか。また、強/弱どちらの関係ですか。

☞ **ファイルメニューの開く**のデータを選択します。
CENSUS.savを選択し、**開く**ボタンをクリックします。
グラフメニューの散布図/ドットを選択します。
単純な散布図を選択し、**定義**ボタンをクリックします。
income（収入）を**Y軸**に、**educ**（教育年数）を**X軸**に移動します。
OKボタンをクリックします。

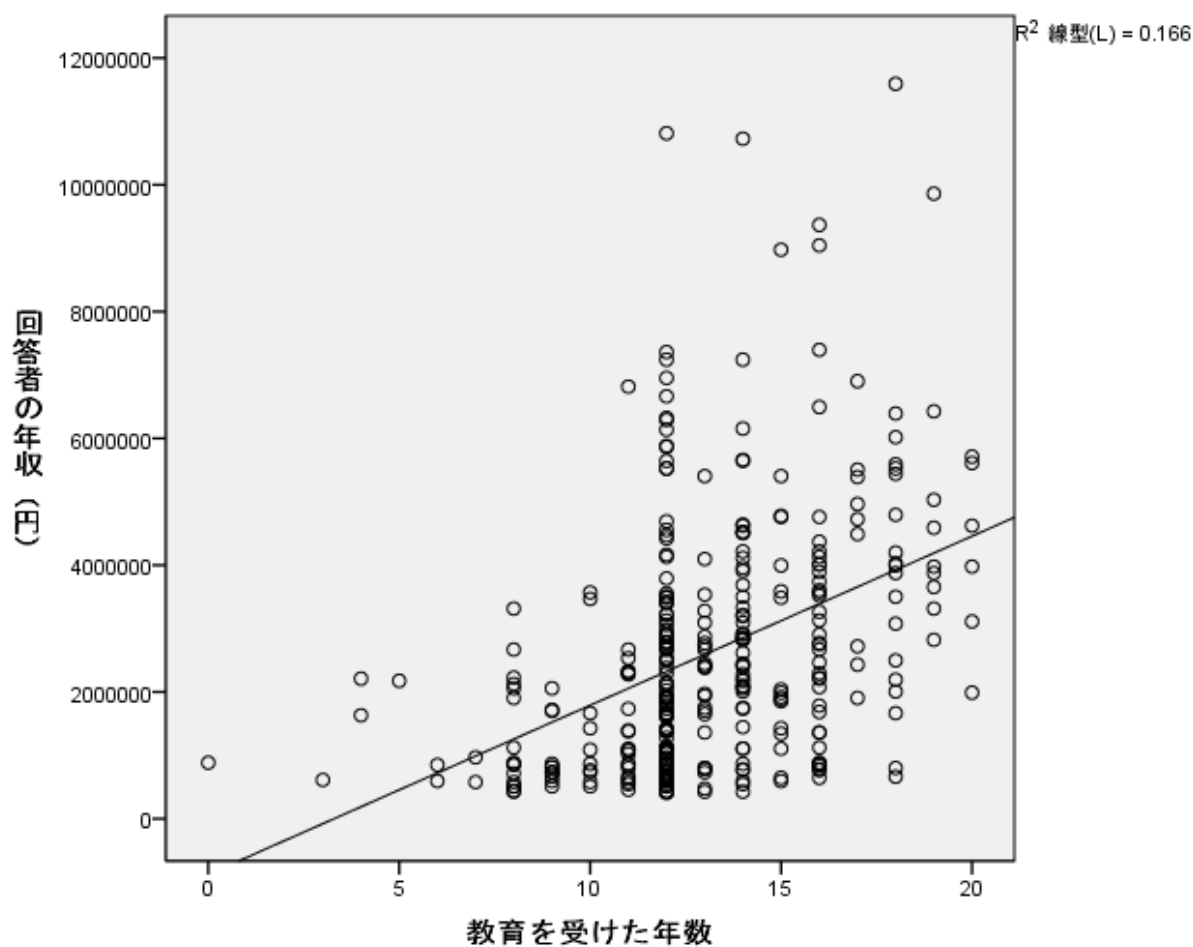


図E.35 incomeとeducの散布図

散布図から教育年数が上がると収入が上がる、右上がりの傾向、つまり変数間には正の関係があることがわかります。この散布図に直線を当てはめる想像をすると、線に乗らないケースの多くが広範囲に散らばることから、関係の強さは、弱いのではないかと思います。実際に、直線を当てはめた方がよいでしょう。

4. 散布図に回帰直線を重ね合わせてみましょう。このことから何か結論を得ることができますか。

☞ **散布図**をダブルクリックします。
ツールバーの合計での線の当てはめボタンをクリックします。



図E.36 直線を当てはめた散布図

直線は、右上がりになっています。また、想像した通り、直線上に乗らない多くのケースが広範囲に散らばっています。

5. 2変量の相関の手続きで、適切なオプションを指定し変数の関係を要約しましょう。調べた関係についてどのような結論が得られたでしょうか。有意な組み合わせはありましたか。どれが重要でしょうか。



分析メニューの**相関の2変量**を選択します。

income (収入)、**educ** (教育年数)、**age** (年齢) を**変数**リストボックスへ移動します。

OK ボタンをクリックします。

相関係数				
		回答者の年収 (円)	教育を受けた 年数	回答者の年齢
回答者の年収 (円)	Pearson の相関係数	1.000	.408**	-.005
	有意確率 (両側)		.000	.926
	N	351	351	351
教育を受けた年数	Pearson の相関係数	.408**	1.000	-.181**
	有意確率 (両側)	.000		.000
	N	351	379	379
回答者の年齢	Pearson の相関係数	-.005	-.181**	1.000
	有意確率 (両側)	.926	.000	
	N	351	379	379

** 相関係数は 1% 水準で有意 (両側) です。

図E.37 incomeとeducとageの相関行列

収入と教育年数の相関係数は、0.408となっています。弱い線型の相関といったところでしょう。また、相関係数のとなりに**が表示されています。母集団において2変数は、統計的に有意であるといえます (母集団において2変数の相関係数は、ゼロではない)。

収入と年齢の相関係数は、-0.005です。負の関係で、ほとんど線型の相関がありません。

教育年数と年齢の相関係数は、-0.181です。負の関係で、ほとんど線型の相関がありません。こちらは、相関係数のとなりに**が表示されています。母集団において2変数は、統計的に有意であるといえます。

第9章 グループ間の量的データの関係性：t 検定

この演習では、**t 検定**の手続きを使用して1つの質的変数と1つの量的変数の関係を探索します。

1. 興味のある質的変数と量的変数を1つずつ選択しましょう。平均の比較手続きを使用して、質的変数のグループごとに、量的変数の平均値と標準偏差を比較しましょう。他のサブグループと比較して、平均値が異なっているのはどのサブグループでしょうか。結論を書きとめましょう。

性別の教育年数の平均値を調べてみます。

- ☞ **ファイルメニューの開く**のデータを選択します。
CENSUS.savを選択し、**開く**ボタンをクリックします。
分析メニューの平均の比較のグループの平均を選択します。
educ（教育年数）を**従属変数**リストボックスへ、**gender**（性別）を**独立変数**リストボックスへ移動します。
OKボタンをクリックします。

報告書

教育を受けた年数

回答者の性別	平均値	度数	標準偏差
女性	12.54	206	2.974
男性	13.57	173	3.067
合計	13.01	379	3.056

図E.38 性別の教育年数の平均値と標準偏差

教育年数の平均は、女性より男性の方が高いことがわかります。

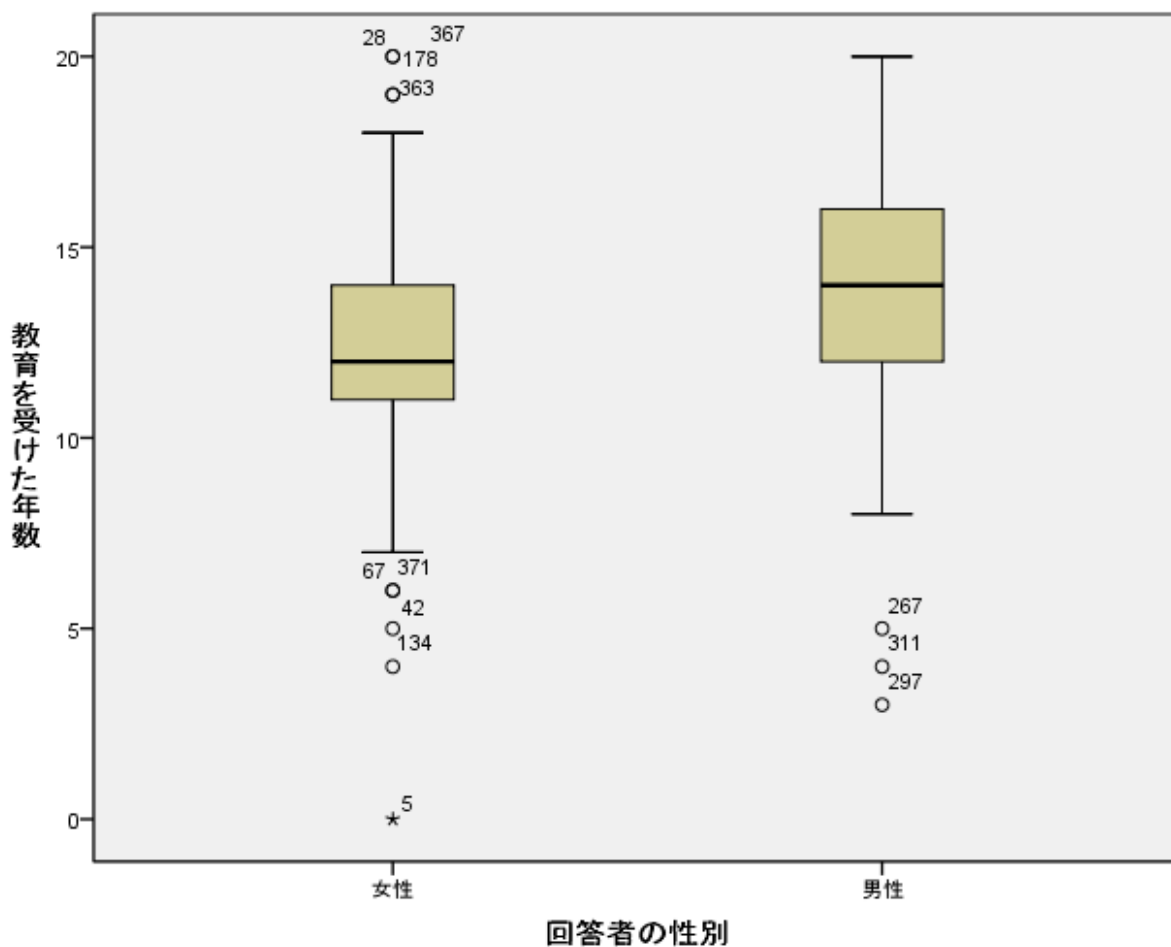
2. 1.と同じ変数を対象にして、探索的分析の手続きを使用して、サブグループ間の違いを探索してみましょう。箱ひげ図を見てみましょう。プロットは、何を示しますか。



分析メニューの**記述統計の探索的**を選択します。

educ（教育年数）を**従属変数**リストボックスへ、**gender**（性別）を**因子**リストボックスへ移動します。

OKボタンをクリックします。



図E.39 性別の教育年数の箱ひげ図

中央値は、男性の方が女性より上にあります。また、男性は下のほうに外れ値がみられますが、それ以外の分布は中央値を真ん中にして対象となっています。女性は。上下どちらにも外れ値があり、中央値が下よりとなっています。

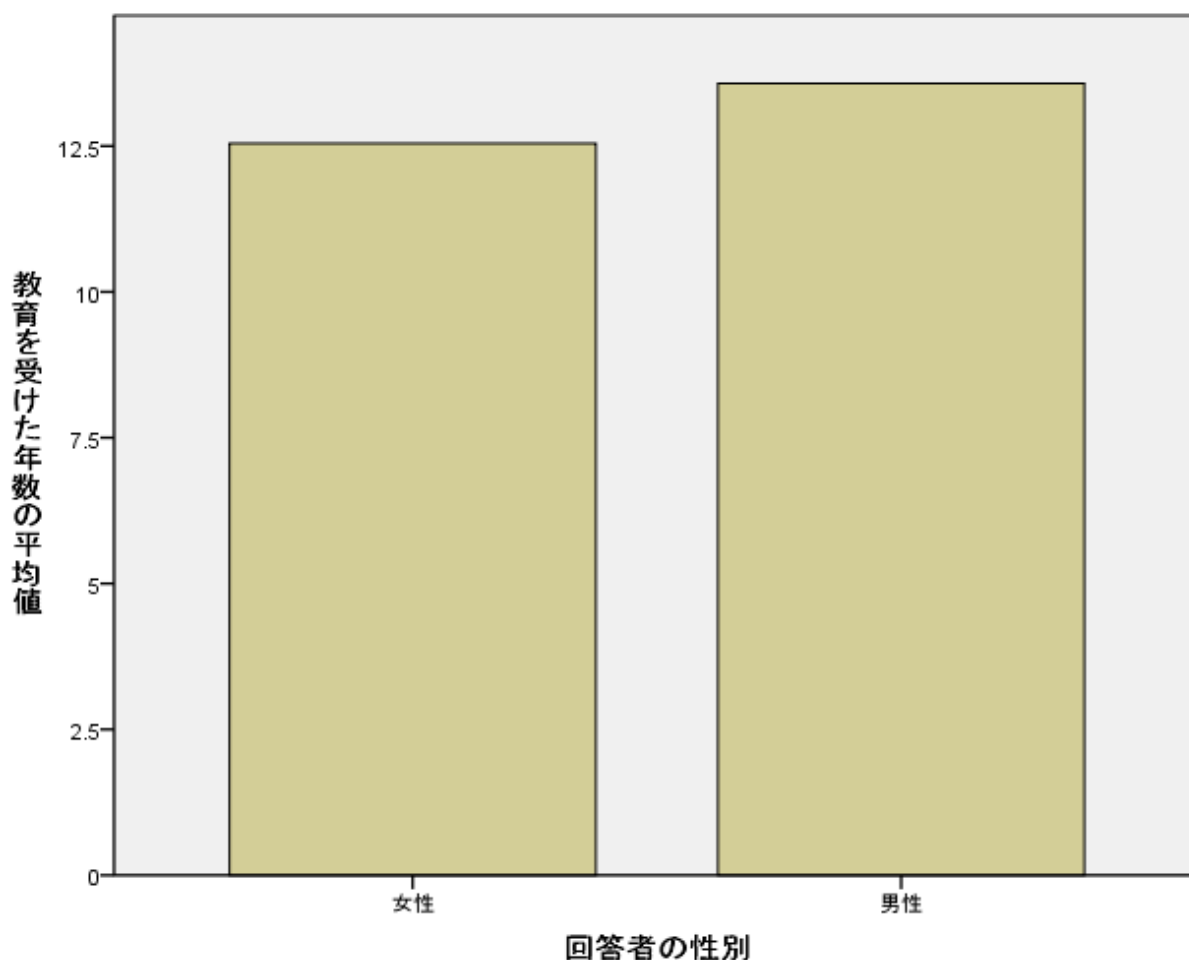
3. グラフメニューから2つの変数の単純な棒グラフを作成しましょう。棒の表現内容で、その他の統計量を選択し量的変数の平均値を使用しましょう。

☞ **グラフメニューのレガシーダイアログの棒**を選択します。

定義ボタンをクリックします。

棒の表現内容で**その他の統計量**を選択します。

educ（教育年数）を**変数**ボックスへ、**gender**（性別）を**カテゴリ軸**ボックスに移動し、**OK**ボタンをクリックします



図E.40 性別の教育年数の平均値を表わした棒グラフ

4. 質的変数に3つ以上のカテゴリがある場合、そのうち特に興味のあるものを2つ選択しましょう。もしくは、カテゴリが2つの質的変数を選択しても構いません。

帰無仮説と対立仮説を書きとめておきましょう。

帰無仮説 (H_0) : 性別において教育年数の平均値には差がない。

対立仮説 (H_1) : 性別において教育年数の平均値には差がある。

5. t 検定の手続きを使用して、選択した量的変数のグループ平均に違いがあるかどうか検定しましょう。t 検定のどの部分を参照すればよいかは、Levene検定の結果から決まることを思い出してください。分析の結果から何がわかりましたか。

☞ **分析メニューの平均の比較の独立したサンプルのT検定**を選択します。
educ (教育年数) を**検定変数**リストボックスへ、**gender** (性別) を**グループ化変数**ボックスへ移動します。
グループの定義ボタンをクリックし、**グループ1**テキストボックスに**m**と、**グループ2**テキストボックスに**f**と入力します。
続行ボタンをクリックし、**OK**ボタンをクリックします。

独立サンプルの検定									
		等分散性のための Levene の検定		2 つの母平均の差の検定					
		F 値	有意確率	t 値	自由度	有意確率 (両側)	平均値の差	差の標準誤差	差の 95% 信頼区間
教育を受けた年数	等分散を仮定する。	.822	.365	3.306	377	.001	1.029	.311	.417 1.640
	等分散を仮定しない。			3.297	361.652	.001	1.029	.312	.415 1.642

図E.41 genderとeducの独立したサンプルの t 検定の結果

等分散性のためのLeveneの検定は、性別の教育年数の分散は等しいという仮説を検定しています。ここでは、統計的に有意ではない結果となっています。そこで、2つの母平均の差の検定の結果を見ます。先の結果から、ここでは、上段の数値を利用します。有意確率から、性別による教育年数に平均値には差があることがわかります。