第 5 回：数値の誤りへの対処

北村　友宏

2020 年 6 月 5 日

本日の内容

1. 数値の誤りの探索（前回までの復習）

2. 数値の誤りへの対処

データの数値の誤り

- 統計データには，編集時の（作成・公表者側の）ミス等により誤った数値が観測されていることがある.
  - e.g., 他の観測値に比べ異常に大きい（小さい）
  - ミクロデータではこうしたケースが少なくない.
- 誤った数値をそのままにして分析すると，結果に大きく影響する場合がある.

$$\Downarrow$$

対処して分析する必要がある.

数値の誤りの探し方

- 記述統計やヒストグラムを確認する.
- データセットの各観測値を確認する.

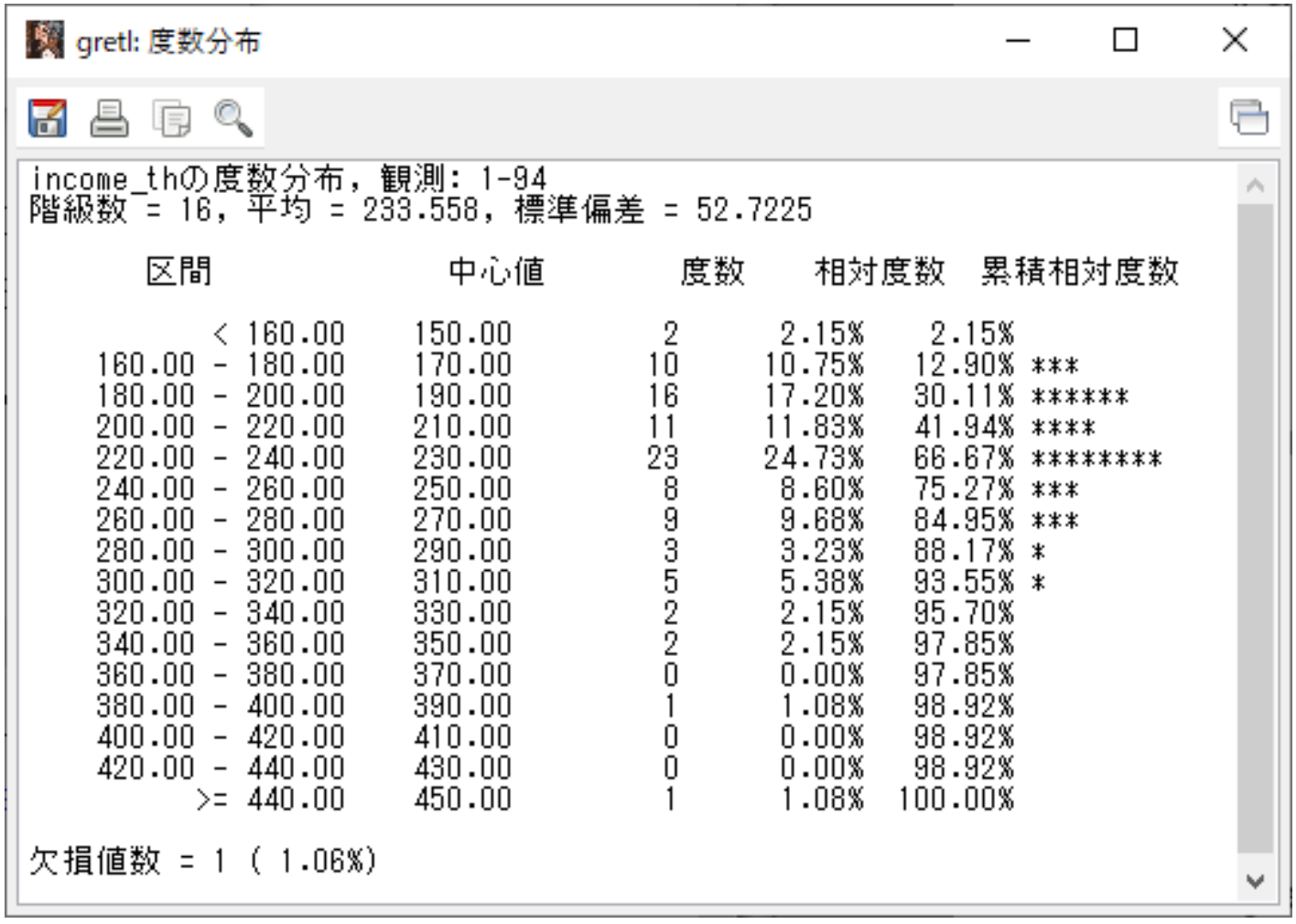
前回まで，現在分析中の全国消費実態調査の都道府県別・男女別データについて，データセットの各観測値，記述統計，ヒストグラムを確認してきた.
今回は，ヒストグラムから再び確認し，数値の誤りを探索する.

実習 1

前回の復習として，可処分所得（千円単位）の度数分布とヒストグラムを出力する.

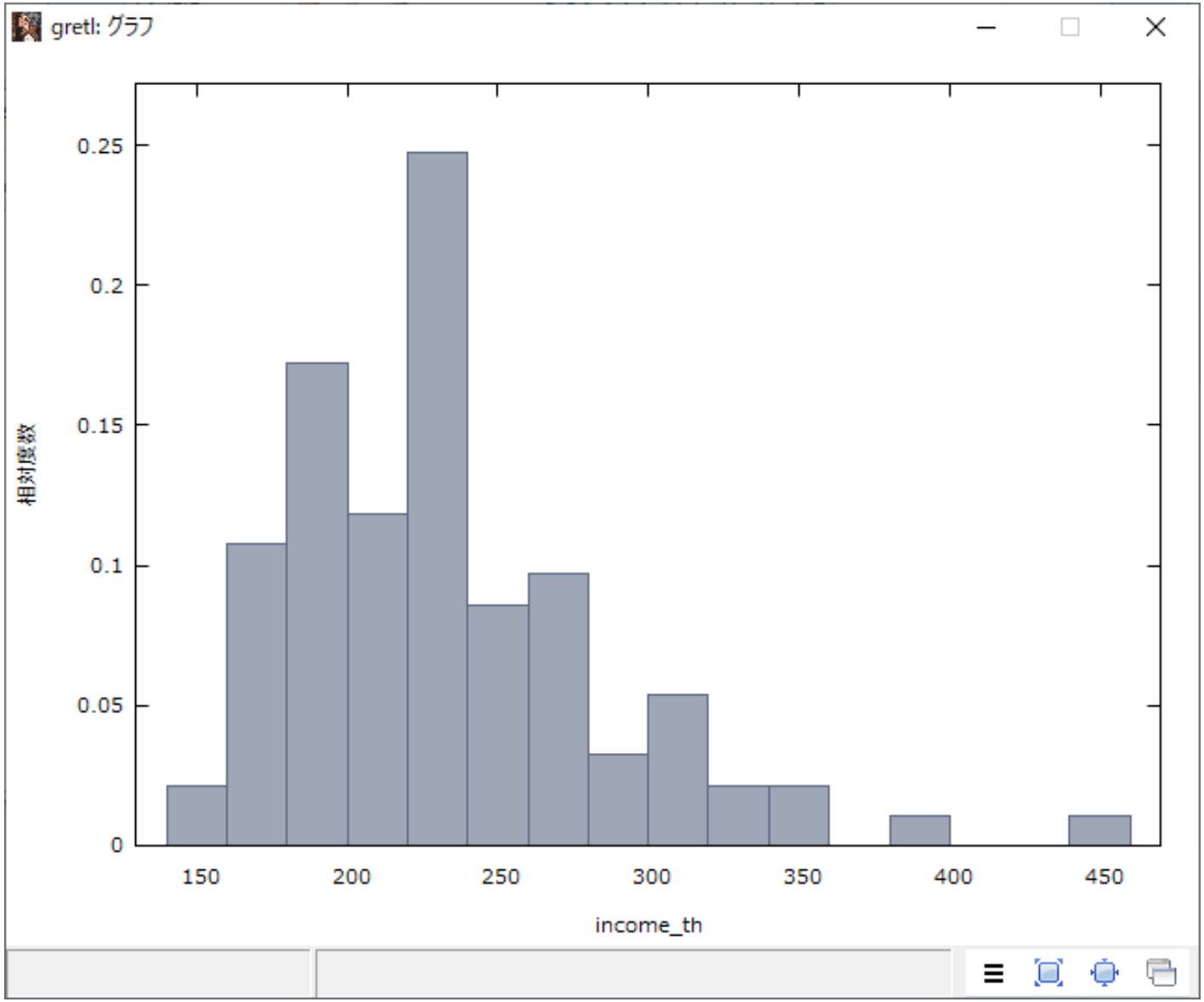
1. gretl を起動.
2. 「ファイル」→「データを開く」→「ユーザー・ファイル」と操作.
3. 消費 2009.gdt を選択し，「開く」をクリック.
4. 「income_th」を左クリックして選択し，その上で右クリック→「度数分布」と操作.

5. 「gretl: 度数分布」ダイアログボックスの，「左端の階級の下限値:」をクリックして選択し，その右の入力ボックスに 140.000 と入力.
   - 最初の階級の最小値が 140 千円（14 万円）となる.
6. 「階級幅:」の右の入力ボックスに 20.000 と入力.
   - 階級幅が 20 千円（2 万円）となる.
7. 「グラフを表示」にチェックが入っていることを確認．入っていなければクリックしてチェックを入れる.
8. 「OK」をクリックすると，千円単位の可処分所得（変数名：income_th）の度数分布とヒストグラムが表示される.



度数分布については，このような画面が表示されれば成功.

まだ作業があるので，「gretl: 度数分布」のウィンドウはまだ閉じない！



ヒストグラムについては，このような画面が表示されれば成功．まだウィンドウを閉じない！

出力したヒストグラムから分かること

2009 年における単身勤労世帯の 1 ヶ月間の可処分所得の都道府県別・男女別平均は,

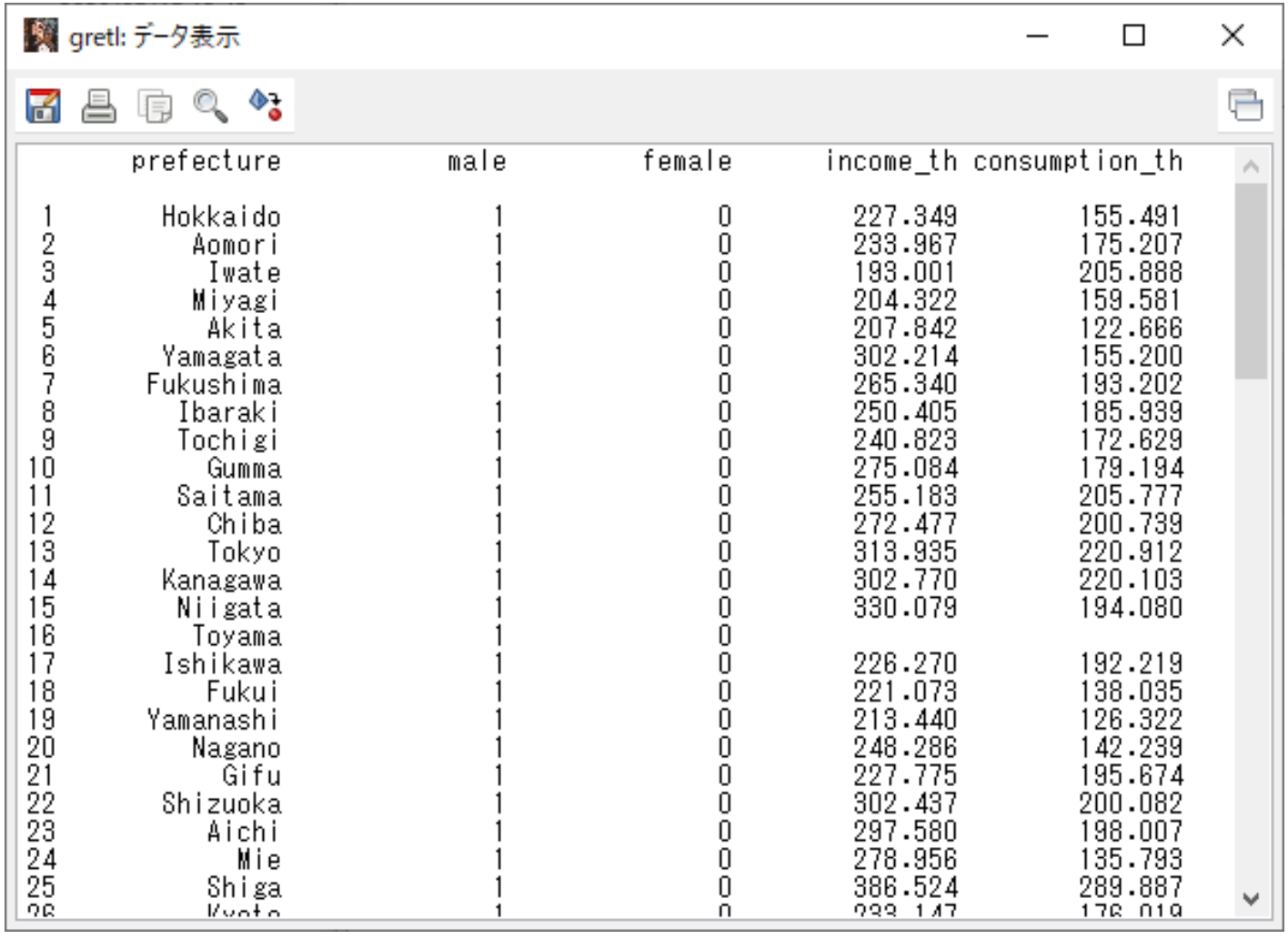
- 16 万円から 28 万円にかけて観測値が集中している.
- 22 万円から 24 万円の階級（区間）に属している観測値が最も多い（相対度数 24.73%）.
- 集中している部分から外れた観測値（38 万円〜40 万円）が存在する.
- さらに大きく外れた観測値（44 万円〜）が存在する！

⇓

大きく外れているのはどの都道府県・どの性別かを突き止める．そのために，可処分所得の観測値のリストを出力する.

実習 2

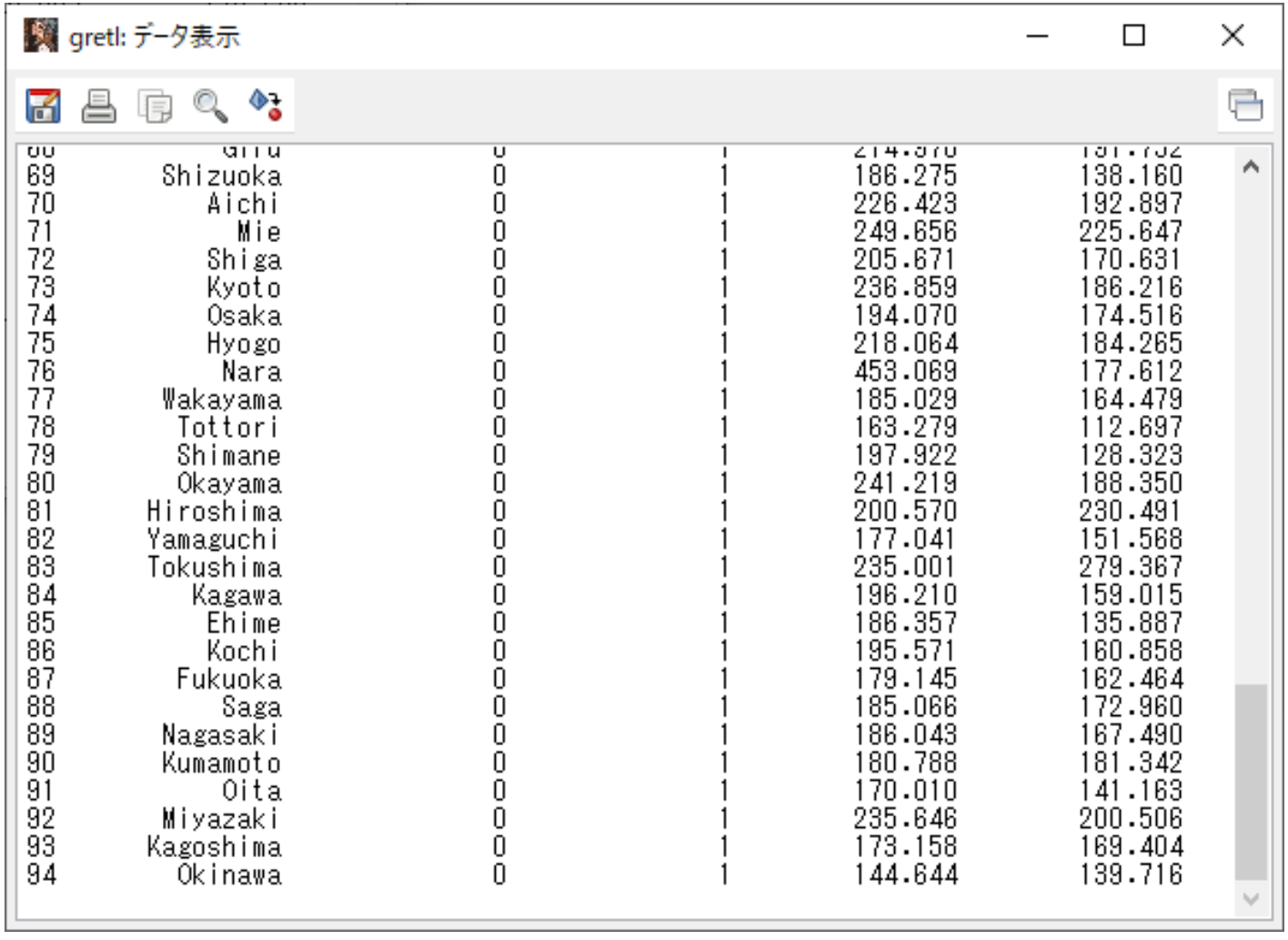
1. 度数分布とグラフのウィンドウは閉じてよい．gretl のウィンドウで，Ctrl キーを押しながら「prefecture」，「male」，「female」，「income_th」，「consumption_th」の 5 つをクリックして選択し，その上で右クリック→「データ（値）を表示」と操作すると，これら 5 つの変数の観測値リストが新規ウィンドウにて表示される．



	prefecture	male	female	income_th	consumption_th
1	Hokkaido	1	0	227.349	155.491
2	Aomori	1	0	233.967	175.207
3	Iwate	1	0	193.001	205.888
4	Miyagi	1	0	204.322	159.581
5	Akita	1	0	207.842	122.666
6	Yamagata	1	0	302.214	155.200
7	Fukushima	1	0	265.340	193.202
8	Ibaraki	1	0	250.405	185.939
9	Tochigi	1	0	240.823	172.629
10	Gumma	1	0	275.084	179.194
11	Saitama	1	0	255.183	205.777
12	Chiba	1	0	272.477	200.739
13	Tokyo	1	0	313.935	220.912
14	Kanagawa	1	0	302.770	220.103
15	Niigata	1	0	330.079	194.080
16	Toyama	1	0		
17	Ishikawa	1	0	226.270	192.219
18	Fukui	1	0	221.073	138.035
19	Yamanashi	1	0	213.440	126.322
20	Nagano	1	0	248.286	142.239
21	Gifu	1	0	227.775	195.674
22	Shizuoka	1	0	302.437	200.082
23	Aichi	1	0	297.580	198.007
24	Mie	1	0	278.956	135.793
25	Shiga	1	0	386.524	289.887
26	Kyoto	1	0	233.147	176.019

このような画面が表示されれば成功．まだウィンドウを閉じない！

下までスクロールすると，



68	Gifu	0	1	214.370	151.732
69	Shizuoka	0	1	186.275	138.160
70	Aichi	0	1	226.423	192.897
71	Mie	0	1	249.656	225.647
72	Shiga	0	1	205.671	170.631
73	Kyoto	0	1	236.859	186.216
74	Osaka	0	1	194.070	174.516
75	Hyogo	0	1	218.064	184.265
76	Nara	0	1	453.069	177.612
77	Wakayama	0	1	185.029	164.479
78	Tottori	0	1	163.279	112.697
79	Shimane	0	1	197.922	128.323
80	Okayama	0	1	241.219	188.350
81	Hiroshima	0	1	200.570	230.491
82	Yamaguchi	0	1	177.041	151.568
83	Tokushima	0	1	235.001	279.367
84	Kagawa	0	1	196.210	159.015
85	Ehime	0	1	186.357	135.887
86	Kochi	0	1	195.571	160.858
87	Fukuoka	0	1	179.145	162.464
88	Saga	0	1	185.066	172.960
89	Nagasaki	0	1	186.043	167.490
90	Kumamoto	0	1	180.788	181.342
91	Oita	0	1	170.010	141.163
92	Miyazaki	0	1	235.646	200.506
93	Kagoshima	0	1	173.158	169.404
94	Okinawa	0	1	144.644	139.716

まだウィンドウを閉じない！

出力した観測値リストから分かること

- 可処分所得が 44 万円を超えている（最大値の約 45.3 万円となっている）のは，奈良県の女性.
- 出力した観測値リストを見ると，奈良県の女性は可処分所得が消費支出額（約 17.8 万円）の 2 倍を超えており，他の都道府県と比べても可処分所得が相対的に大きい.

$$\Downarrow$$

奈良県の女性の可処分所得の数値は明らかに誤り.

数値の誤りへの対処

明らかな数値の誤りへの対処方法は，以下の 3 通り．

- 明らかに誤りの観測値のみを除外する．
- トリミング（trimming）する．
  - e.g.,「90%トリミング」とすると，その変数の上位・下位 5%を除外する．
- ウィンザライズ（winsorize）する．
  - e.g.,「90%ウィンザライズ」とすると，その変数の上位 5%を全て 95%分位点の値に，下位 5%を全て 5%分位点の値に置き換える．

どの方法を使うかについては，指導教員と相談すること．

この授業では，明らかに誤りの観測値のみを除外する方法を説明する．

特定の観測値の消去方法

- gretl のメニューバーから「標本」→「基準に基づいて制限する」と操作し，「ケース（case）を選択する条件式を入力して下さい:」の下の入力ボックスに，データセットに残したい条件式を入力し，OK をクリック.
  - 入力した条件式を満たさない観測値が消去される.
  - 「ヘルプ」をクリックすると，使える演算子などを確認できる.
  - 「この制約を永続的にする」にチェックを入れなければ，消去後，データセットを上書き保存していない状態で，gretl のメニューバーから「標本」→「全範囲に戻す」と操作すると，観測値を消去する前のデータセットに戻すことができる.

実習 3

奈良県の女性の 1 ヶ月間の可処分所得が 44 万円を超えており，不自然なので，その個体をデータセットから消去したい．そのため，「可処分所得（千円単位）が 440 以下」という条件を満たす観測値を残し，それを満たさない観測値を消去する．

1. 観測値リストのウィンドウは閉じてよい．gretl のメニューバーから「標本」→「基準に基づいて制限する」と操作し，「ケース（case）を選択する条件式を入力して下さい:」の下の入力ボックスに，

   income_th<=440

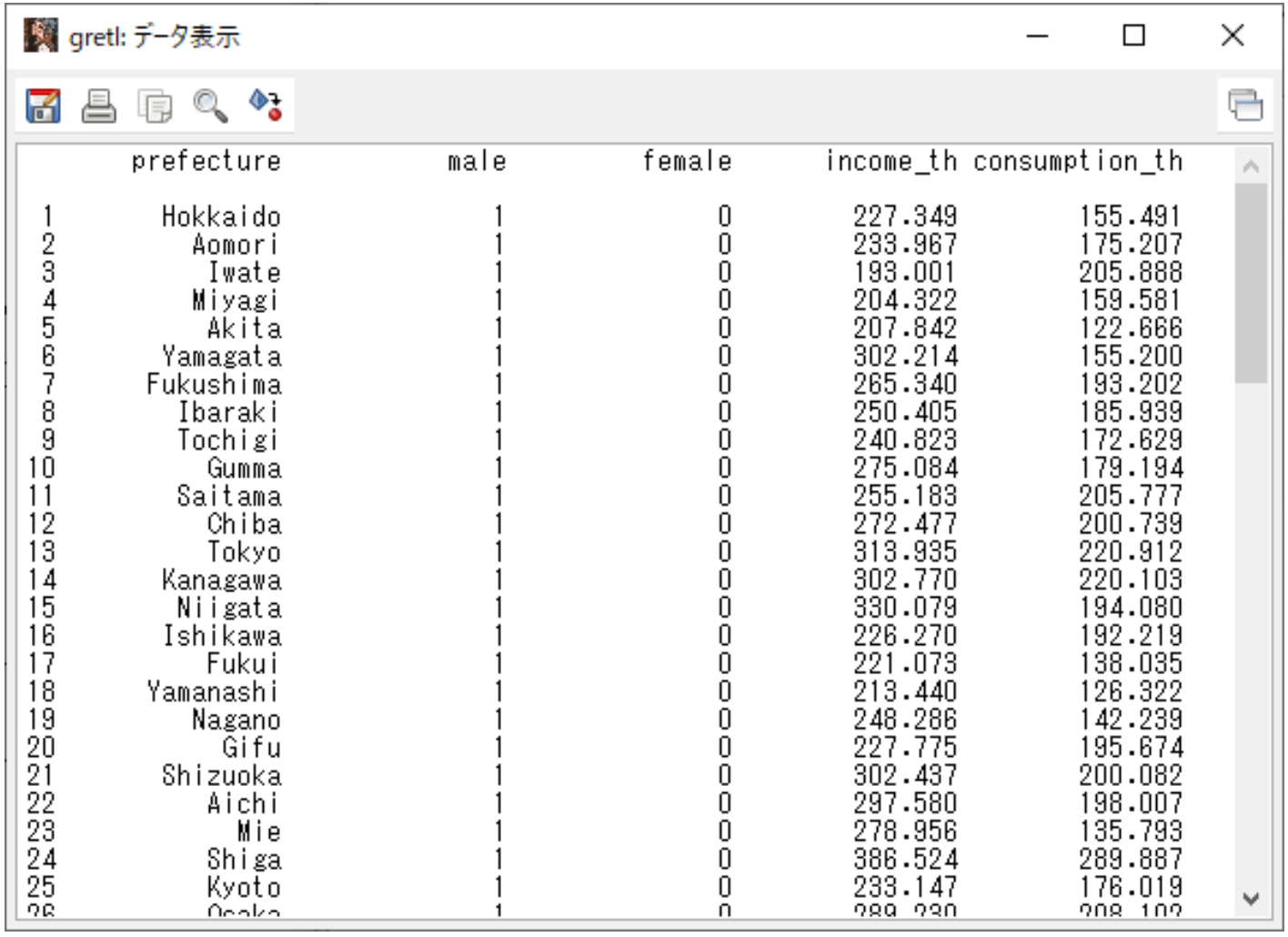
   と入力し，OK をクリック．
   - 「<=」は「以下」という意味の演算子．

2. 「2 個の観測を落としました」というメッセージが表示されるので，「閉じる」をクリック．
3. 上書き保存する．メニューバーから「ファイル」→「データの保存」と操作すると，「データセットは現在、サブサンプルされています 全範囲に戻しますか？」というメッセージが表示されるので，「いいえ」をクリック．

実習 4

再度，観測値リストを出力する．

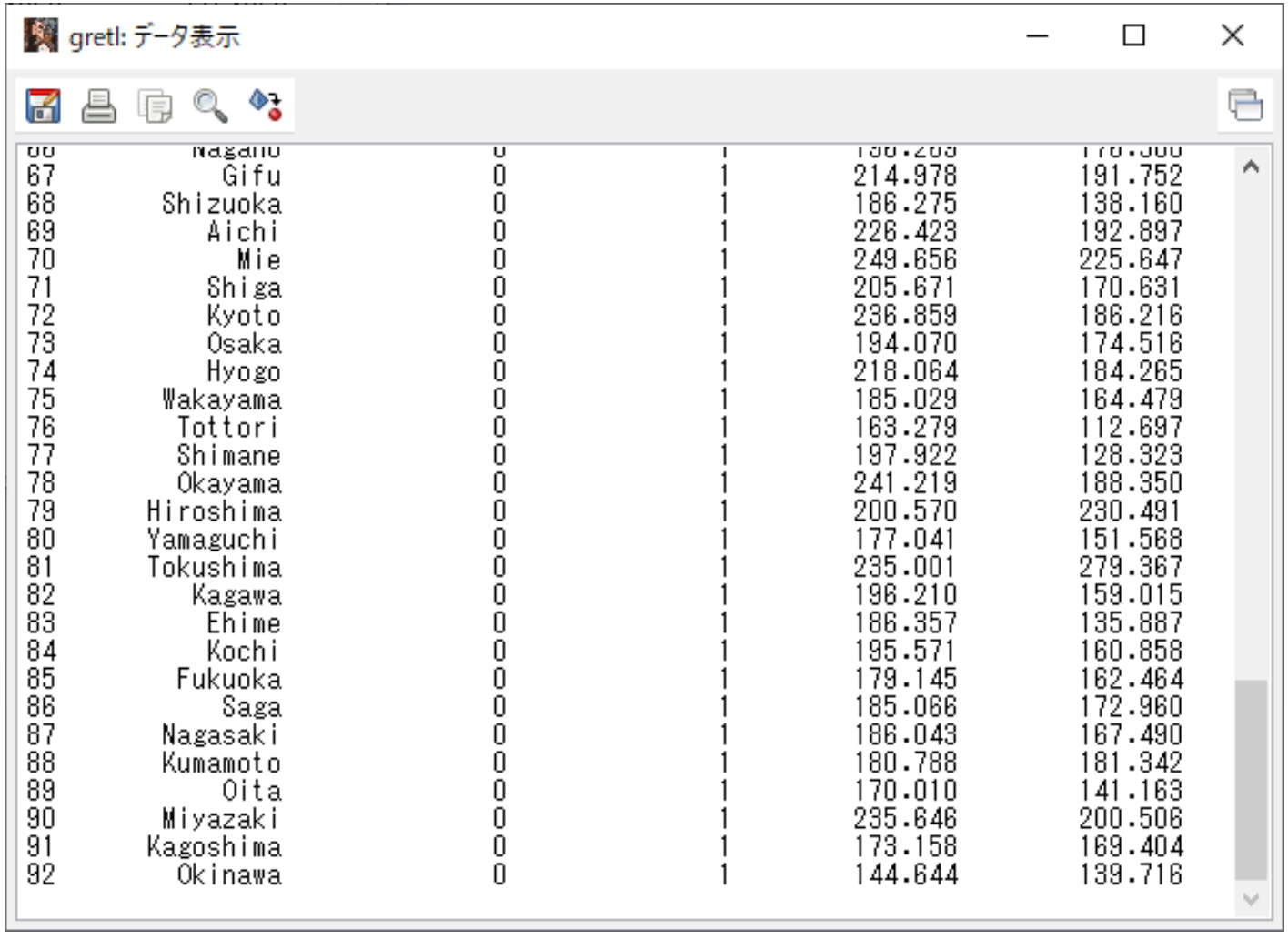
1. gretl のウィンドウで，Ctrl キーを押しながら「prefecture」，「male」，「female」，「income_th」，「consumption_th」の 5 つをクリックして選択し，その上で右クリック→「データ（値）を表示」と操作すると，これら 5 つの変数の観測値リストが新規ウィンドウにて表示される．



	prefecture	male	female	income_th	consumption_th
1	Hokkaido	1	0	227.349	155.491
2	Aomori	1	0	233.967	175.207
3	Iwate	1	0	193.001	205.888
4	Miyagi	1	0	204.322	159.581
5	Akita	1	0	207.842	122.666
6	Yamagata	1	0	302.214	155.200
7	Fukushima	1	0	265.340	193.202
8	Ibaraki	1	0	250.405	185.939
9	Tochigi	1	0	240.823	172.629
10	Gumma	1	0	275.084	179.194
11	Saitama	1	0	255.183	205.777
12	Chiba	1	0	272.477	200.739
13	Tokyo	1	0	313.935	220.912
14	Kanagawa	1	0	302.770	220.103
15	Niigata	1	0	330.079	194.080
16	Ishikawa	1	0	226.270	192.219
17	Fukui	1	0	221.073	138.035
18	Yamanashi	1	0	213.440	126.322
19	Nagano	1	0	248.286	142.239
20	Gifu	1	0	227.775	195.674
21	Shizuoka	1	0	302.437	200.082
22	Aichi	1	0	297.580	198.007
23	Mie	1	0	278.956	135.793
24	Shiga	1	0	386.524	289.887
25	Kyoto	1	0	233.147	176.019
26	Osaka	1	0	289.230	208.102

このような画面が表示されれば成功．まだウィンドウを閉じない！

下までスクロールすると，



66	Nagano	0	1	158.283	170.566
67	Gifu	0	1	214.978	191.752
68	Shizuoka	0	1	186.275	138.160
69	Aichi	0	1	226.423	192.897
70	Mie	0	1	249.656	225.647
71	Shiga	0	1	205.671	170.631
72	Kyoto	0	1	236.859	186.216
73	Osaka	0	1	194.070	174.516
74	Hyogo	0	1	218.064	184.265
75	Wakayama	0	1	185.029	164.479
76	Tottori	0	1	163.279	112.697
77	Shimane	0	1	197.922	128.323
78	Okayama	0	1	241.219	188.350
79	Hiroshima	0	1	200.570	230.491
80	Yamaguchi	0	1	177.041	151.568
81	Tokushima	0	1	235.001	279.367
82	Kagawa	0	1	196.210	159.015
83	Ehime	0	1	186.357	135.887
84	Kochi	0	1	195.571	160.858
85	Fukuoka	0	1	179.145	162.464
86	Saga	0	1	185.066	172.960
87	Nagasaki	0	1	186.043	167.490
88	Kumamoto	0	1	180.788	181.342
89	Oita	0	1	170.010	141.163
90	Miyazaki	0	1	235.646	200.506
91	Kagoshima	0	1	173.158	169.404
92	Okinawa	0	1	144.644	139.716

まだウィンドウを閉じない！

先ほどの作業による観測値リストの変化

- 女性の観測値について，兵庫県と和歌山県の間にあった奈良県が消去されている.
- 男性の観測値について，新潟県と石川県の間にあった富山県（もともと欠損値）も消去されている.
- これは，gretl では

  「全ての欠損でない値＜欠損値（.）」

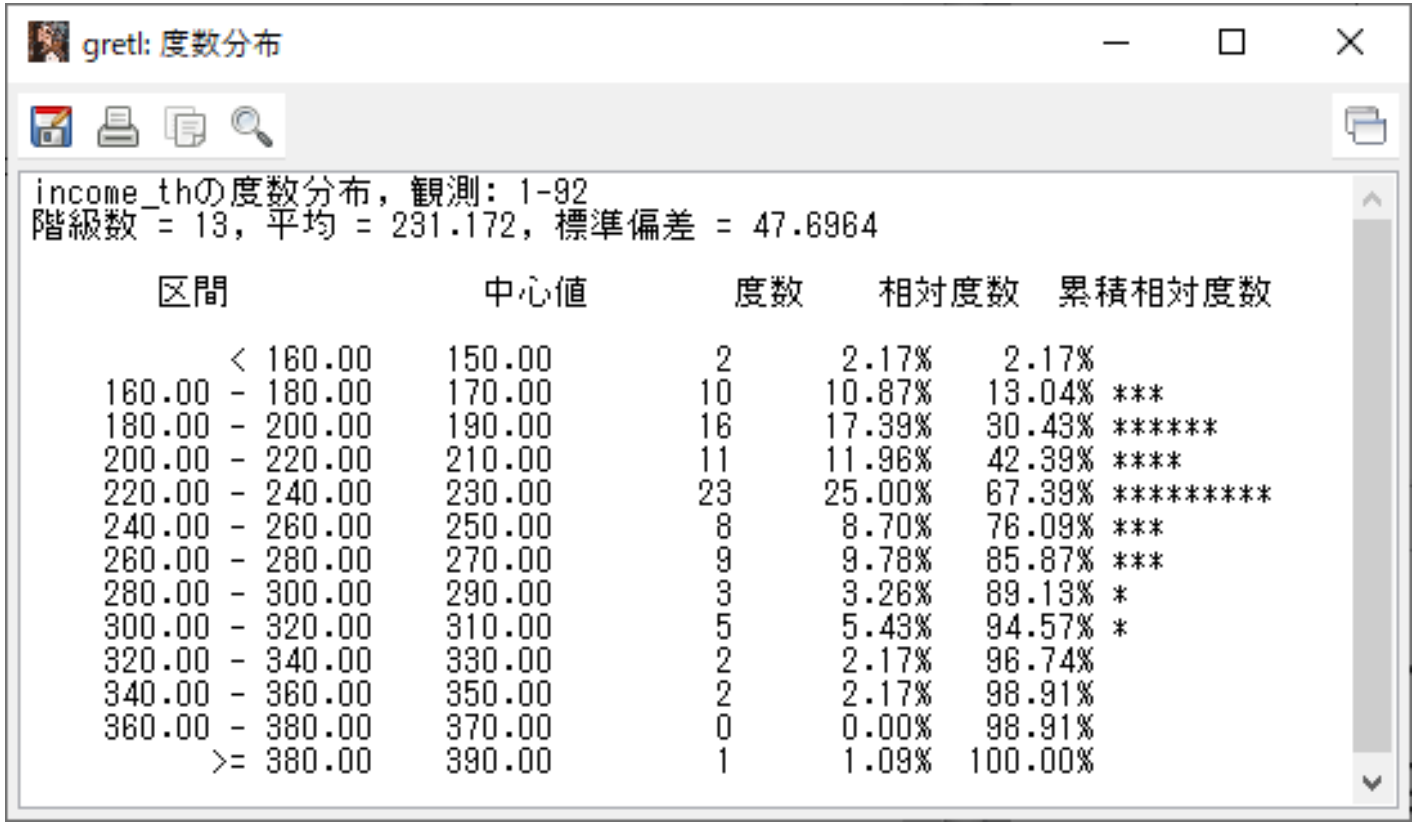
  と認識されるため.
  - 欠損値も 440 を超えている（「440 以下」という条件を満たさない）と認識.

実習 4

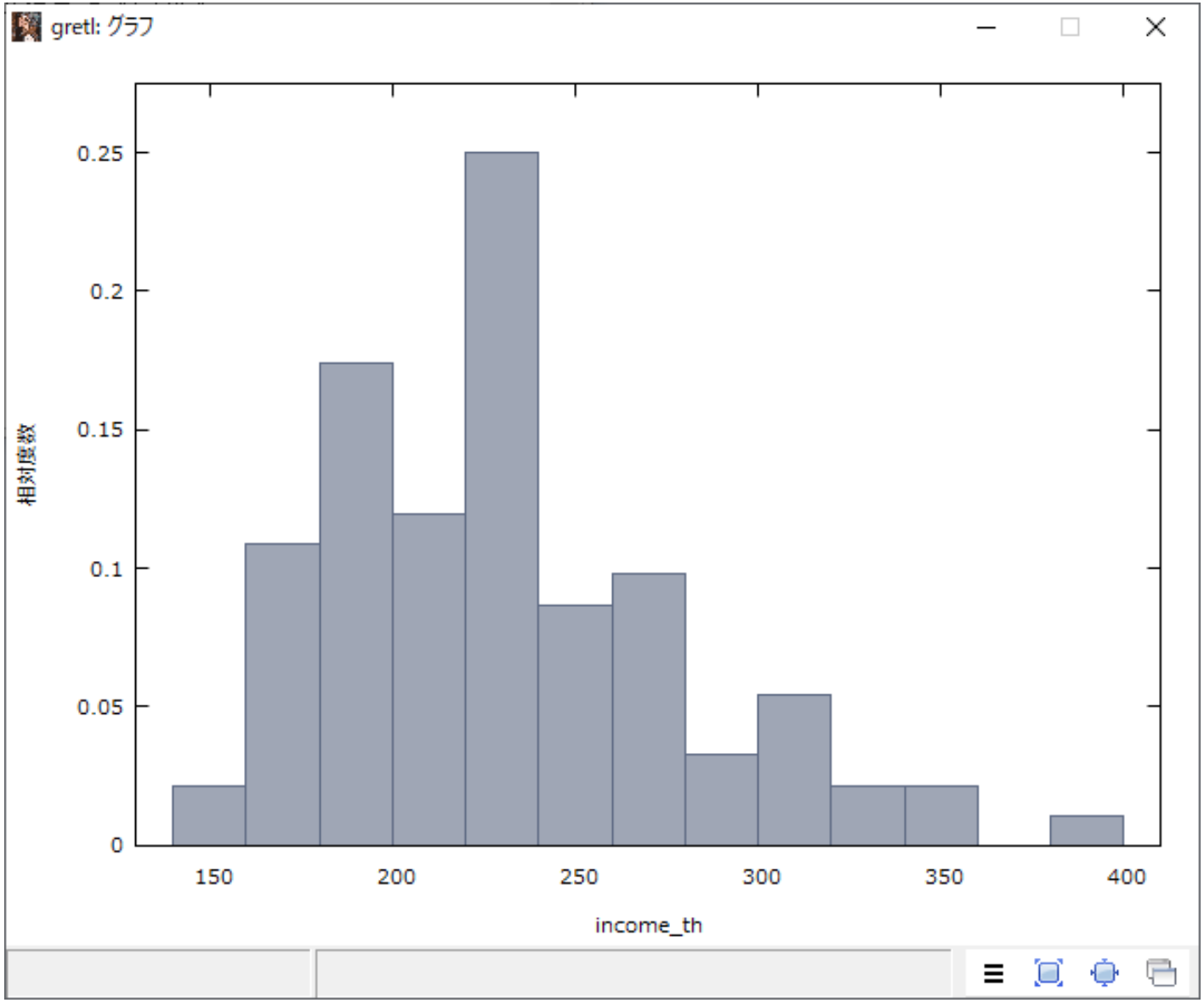
続いて，再度，ヒストグラムを確認する．

1. 観測値リストのウィンドウは閉じてよい．gretl のウィンドウで，「income_th」を左クリックして選択し，その上で右クリック→「度数分布」と操作．
2. 「gretl: 度数分布」ダイアログボックスの，「左端の階級の下限値:」をクリックして選択し，その右の入力ボックスに 140.000 と入力．
   - 最初の階級の最小値が 140 千円（14 万円）となる．
3. 「階級幅:」の右の入力ボックスに 20.000 と入力．
   - 階級幅が 20 千円（2 万円）となる．

4. 「グラフを表示」にチェックが入っていることを確認．入っていなければクリックしてチェックを入れる．
5. 「OK」をクリックすると，千円単位の可処分所得（変数名：income_th）の度数分布とヒストグラムが表示される．



度数分布については，このような画面が表示されれば成功.
まだ作業があるので，「gretl: 度数分布」のウィンドウはまだ閉じない！



ヒストグラムについては，このような画面が表示されれば成功．まだウィンドウを閉じない！

先ほどの作業によるヒストグラムの変化

集中している部分から少し外れた観測値（38 万円〜40 万円）は残っているものの，大きく外れて「44 万円〜」だった観測値は消滅した.
本日の作業はここまで.