第5回 Let's データ分析コンテストの規定課題の SAS プログラム

周防　節雄
兵庫県立大学・名誉教授

Exemplary SAS Programs to Solve the Compulsory Exercises for the 5th Micro Data Competition in Japan SAS Users Forum 2017

Setsuo Suoh
Professor Emeritus of the University of Hyogo

要旨

SAS ユーザー総会 2017 における「Let's データ分析　第5回ミクロデータ分析コンテスト」の規定課題について、出題者の立場から、模範解答の SAS プログラムを例示する。

キーワード　全国消費実態調査、匿名データ、新擬似ミクロデータ、集計乗率、proc freq、proc tabulate、 proc sgplot、proc univariate、SG プロシジャ、プログラミング・スタイル、ジニ係数、ロレンツ曲線、

1. はじめに

独立行政法人統計センターは、平成16年の全国消費実態調査のミクロデータから教育目的で教育用擬似ミクロデータを作成し、公的ミクロデータの利用実習用に提供してきた。SAS ユーザー総会では、2013 年から 2016 年まで、この擬似ミクロデータを使用してデータコンペを実施してきたが、2016 年度末でこのデータの提供が中止になった。そこで、今年に入って急遽「新擬似ミクロデータ」を SAS ユーザー会世話人有志が集まり、2004 年全国消費実態調査の匿名データを使って、今年のデータコンペ用の「新」擬似ミクロデータ[1]を作成した。

コンペでは、この新擬似ミクロデータを使った規定課題（付録3）と自由課題から成り、規定課題は参加者全員に課せられる。本稿では、この規定問題のためのSASプログラムを例示する。出力結果(付録2)にある図表のタイトルはプログラムの title 文と対応しているので、各プログラムの処理プロセスを追っていける。

2. 新擬似ミクロデータの概要

作成した新擬似ミクロデータと符号表（メタデータ）は、ウェブ上にアップロードしている[2]。

新擬似ミクロデータは 69,131 オブザベーション(世帯)から成る。変数としては、①世帯識別番号（変数名 No=1～69,131）、②14 個の世帯属性（変数名 X01～X14）と集計乗率（変数名 Weight）（表1）[3]、および、③203 個の収支項目（変数名 Y001～Y203）（表 2）がある。この両表の元ファイル

---

1 この新擬似ミクロデータの作成経緯とプロセスについては、本論文集に収録の高橋他(2017)を参照されたい。

2 SASデータセット http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroData.zip
CSV形式 http://mighty.gk.u-hyogo.ac.jp/confidential/Zensho2004GijiMicroDataCSV.zip

3 この他に、調査年を意味する変数 Year があるが、現在の新擬似ミクロデータでは値は全て「2004」である。これは、将来、他の調査年次の擬似ミクロデータ化も予定しているからである。

(メタデータ)は、ウェブにアップロードした符号表「全消 2004 年新擬似ミクロデータ符号表.xlsx」に収録されている。

3. 規定課題 SAS プログラム模範解答プログラムの実行手順

規定課題のために、今回「作成」したSASプログラム(付録1)は、以下の①～⑩であるが、⑥のプログラムは③のSASプログラムが自動作成した。

① set_environment.sas 環境設定:SAS を立ち上げたときに★いつも★最初に実行する。

② setALL.sas → ③④⑤が%include を使って自動的に実行される。
③createFormatNewZenshoGiji2004.sas→*自動作成*→⑥*procFormatNewZenshoGiji2004.sas*
④createtLabelNewZenshoGiji2004.sas →*自動作成*→⑪*labelNewZenshoGiji2004.txt*
⑤setLABELweight.sas→④の結果を使って、変数ラベルと 10 万世帯比 weight 変数を付加

⑥ procFormatNewZenshoGiji2004.sas → 変数 format が読み込まれる。

⑦ kitei2017.sas 規定課題の第1問～第4問を解答する。

★規定問題:第5問 ジニ係数計算とロレンツ曲線描画

⑧ make_weighted_dataX.sas
→ 集計乗率 weight 変数[4]を使って膨らませた SAS データセット"reconstruct"を作成

⑨ gini_lorenz2017 **規定課題**_sas_forum_macro.sas → ジニ係数とロレンツ曲線を表示[5]

⑩ GiniLorenz2017.sas → SAS データセット"reconstruct"を家族分類別に分割した後で
家族分類毎に、⑨を繰り返し自動実行して、ジニ係数とロレンツ曲線表示。

実行手順は以下の通り。

●初めて実行する時は、①②⑥⑦⑧⑩の順に実行する。
●同じ SAS のセッションで再実行の際は、⑦⑧⑩の順に実行する。
●SAS の新しいセッションで再実行する時は、①⑥⑦⑧⑩の順に実行する。

変数フォーマットの作成(プログラム⑥)と変数ラベル文(テキストファイル⑪)は、EXCEL ファイルの符号表をSASプログラムに読ませることで、全自動で処理をした。

4. SAS プログラミング・スタイル

本稿で示した解答例は一例であり、もちろん他のやり方もいろいろある。この解答例のプログラミング・スタイルは、ここ数年来、このユーザー総会でも一貫して採用している方式である。

SASのプログラムを書く際には、通常は人間のプログラマーが全てのSASコードを書いている。複雑なアルゴリズムをプログラム化するには、プログラマー自身でやらざるを得ないが、非常に単純だが量が多くて機械的なコード(いわゆる、"dirty work")の場合は、できれば「自動化」したい。例えば、大量の変数の proc format や label 文がそれに相当する。時間をかければもちろん出来るが、

4 集計乗率については、周防(2016)参照
5 ジニ係数とロレンツ曲線のSASプログラムについては、周防(2008)参照。そこに掲載の付録3のプログラムを規定問題用に一部書き換えて、プログラム⑨を作成した。

手作業でした場合、ケアレスミスが発生しやすい。今回の規定課題でも、proc format や label 文の作成は、エクセル形式の符号表ファイル(メタデータ)をSASプログラムに読ませることによって行った。

もう一つ重要なプログラミング・スタイルは、ひとつのプログラムで全ての処理をしないことである。つまり、全体の作業をいくつかのまとまった処理に分割してから、それぞれを個別のSASプログラムにする。SASプログラムには、メインプログラムとサブプログラムの概念が基本的にはない[6]。異なるプログラムの間は、SAS データセットで連携されている。第 3 節で示したプログラム間の連携はその例である。この方式のメリットは、バグが発生したときに、発生箇所を直ぐに発見できるので、容易に対処できることである。デメリットとしては、プログラムの数が増えて、混乱する可能性がある。その対策としては、第 3 節で示すように、%include を使って、いくつかの一連のプログラムが正しい順序で実行されるように、風呂敷に包むような形で、もうひとつ別の上位プログラムで包んでしまうのが便利である。第 3 節の②のプログラム setALL.sas がその上位プログラムに相当する。その上で、第 3 節の終わりにある三つの●で示すように、異なるシチュエーションに応じて、実行する順番をきちんと文書化しておく。この三つの実行手順を更に、それぞれ%include を使って、風呂敷で包むのも良い。

5. まとめ

今回のコンテストでも、集計乗率を使って、proc freq や proc tabulate をうまく使えるかが試された。集計用乗率という用語に余りなじみのないユーザーもいると思うが、公的ミクロデータの利用では必ず出てくる概念である。公的ミクロデータの分析では、初期段階で分布状況の把握のために、これらのプロシジャを使って度数分布を確認することがよく行われる。

グラフの出力も求められたが、こちらは、JMPを使う方が手軽だと感じた。筆者は、公的ミクロデータを主に利用する関係上、普段余りSASで特殊なグラフを描く必要性が少ないので、可能ならば、他のソフト、例えば EXCEL などでできる範囲内でお茶を濁すことが多い。昨年も今年も、規定課題のグラフ作成にはSGPLOTを使ってみたが、細かい微調整ができずにいて、正直なところ、自分でもまだ十分に使いこなせているとは言えない。SGPLOT は数年前から利用可能になった SG プロシジャ(Statistical Graphics Procedures)[7]に含まれるプロシジャであり、SAS/GRAPH とは別物である。上手に使えば、便利なツールになる。

参考文献

1. 周防節雄(2008)ジニ係数の計算とロレンツ曲線を描くSASプログラム、『SASユーザー総会2008論文集』、pp139-146
2. 高浪洋平(2011)SG プロシジャと GTL によるグラフの作成と ODS PDF による統合解析帳票の作成 ～TQT 試験における活用事例～、『SAS ユーザー総会2011論文集』、pp201-219
3. 周防節雄(2015)SAS プログラムで関数とサブルーティンを作成する方法、『SAS ユーザー総会2015論文集』、pp389-398
4. 周防節雄(2016)SAS ユーザー総会2016における「Let's データ分析第4回ミクロデータ分析コンテスト」の規定課題の SAS プログラム解説、『SAS ユーザー総会2016論文集』、pp330-347

6 数年前から SAS でも関数や副プログラムをユーザー自身が定義出来るようにはなった(周防(2015))。ただ、いわゆるメインプログラムという概念は SAS には馴染まない。

7 本総会でもこれまでに発表事例(高浪 2011)がいくつかある。

表１ 世帯属性と集計乗率

data=formatNewZenshoGiji2004 符号表

OBS	F1	comment	varName	V	meaning
1		世帯に関する項目の変数名及び符号		.	
2		項目名	変数名	.	符号内容
3		調査年	Year	.	全国消費実態調査の調査年(西暦)
4		レコード一連番号	No	.	
5		大都市圏の別	X01	1	3大都市圏
6				0	その他
7		世帯区分	X02	1	勤労者
8				2	勤労者以外
9				3	無職
10		世帯人員	X03	1	１人
11		（住居と生計を共にしている世帯員数）		2	２人
12				3	３人
13				4	４人
14				5	５人以上
15		就業人員	X04	0	０人
16		（就業している世帯員数）		1	１人
17				2	２人
18				3	３人以上
19		住居の構造	X05	1	木造(防火木造含む)
20				2	木造(防火木造含む)以外
21		住居の建て方	X06	1	一戸建
22				2	一戸建以外
23		住居の所有関係	X07	1	持ち家
24				2	持ち家以外
25		世帯主の性別	X08	1	男
26				2	女
27		世帯主の年齢	X09	5	24歳以下
28				6	25～29歳
29				7	30～34歳
30				8	35～39歳
31				9	40～44歳
32				10	45～49歳
33				11	50～54歳
34				12	55～59歳
35				13	60～64歳
36				14	65～69歳
37				15	70～74歳
38				16	75歳以上
39		企業区分･従業者規模	X10	1	民営･自営1～4人
40				2	民営･自営5～29人
41				3	民営･自営30～499人
42				4	民営･自営500人以上
43				5	官公
44				6	無職
45		家族分類	X11	1	単身世帯
46				2	夫婦のみ世帯
47				3	二世代世帯
48				4	二世代(ひとり親)世帯
49				5	三世代世帯
50				6	その他の世帯
51		未就学児の有無	X12	1	無
52				2	有
53		学校に通う世帯員の有無	X13	1	無
54				2	有
55		65歳以上の世帯員数	X14	0	０人
56				1	１人
57				2	２人以上
58		集計用乗率	Weight	.	

プログラム③createFormatNewZenshoGiji2004.sasで作成した

表 2 収支項目一覧

data=labelNewZenshoGiji2004 符号表

OBS	F1	varName	label	F4
1		変数名一覧		
2		変数名	項目名	
3		**〇世帯事項**		
4		Year	調査年	
5		No	レコード連番号	
6		X01	大都市圏の別	
中略				
13		X08	世帯主の性別	
14		X09	世帯主の年齢	
15		X10	企業区分･従業者規模	
16		X11	家族分類	
中略				
20		Weight	集計用乗率	
21		**〇収支項目**		
22		Y001	**年間収入**	(単位:万円)
23		Y002	**収入総額**	(単位:円)以下同様
24		Y003	実収入	
25		Y004	経常収入	
26		Y005	勤め先収入	
27		Y006	世帯主の勤め先収入	
28		Y007	世帯主の配偶者の勤め先収入	
29		Y008	他の世帯員の勤め先収入	
30		Y009	事業･内職収入	
31		Y010	農林漁業収入	
32		Y011	家賃収入	
33		Y012	他の事業収入	
34		Y013	内職収入	
35		Y014	本業以外の勤め先･事業･内職収入	
36		Y015	他の経常収入	
37		Y016	財産収入	
38		Y017	社会保障給付	
39		Y018	公的年金給付	
40		Y019	他の社会保障給付	
41		Y020	仕送り金(収入)	
42		Y021	特別収入	
43		Y022	受贈金	
44		Y023	その他の特別収入	
45		Y024	実収入以外の収入	
46		Y025	預貯金引出	
47		Y026	保険取金	
48		Y027	個人･企業年金保険取金	
49		Y028	他の保険取金	
50		Y029	有価証券売却	
51		Y030	株式売却	
52		Y031	他の有価証券売却	
53		Y032	土地家屋借入金	
54		Y033	他の借入金	
55		Y034	分割払･一括払購入借入金	
56		Y035	財産売却	
57		Y036	その他の実収入以外の収入	
58		Y037	繰入金	
59		Y038	支出総額	
60		Y039	実支出	
61		Y040	消費支出	
62		Y041	食料	
63		Y042	穀類	
64		Y043	米	
65		Y044	パン	
66		Y045	めん類	
67		Y046	他の穀類	
中略				
224		Y203	繰越金	

プログラム④createLabelNewZenshoGiji2004.sasで作成した

太枠で囲った変数が、規定課題で使われる。

付録1　SASプログラム一覧

① set_environment.sas

```
/* set_environment.sas */
  %let drive=F; *★外付けHDのドライブ名を指定する;
*★全消2004年新擬似ミクロデータのSASデータセット、★SASプログラム、★符号表の保存先のパスを指定;
  %let path=¥全消¥擬似ミクロ作成プロジェクト¥★新情報¥最新情報¥全消2004年新擬似ミクロデータSAS版;
  %let pathSASprogram=¥★X60s(2014-8-9)¥SAS_Forum¥2017¥SASprogram;
  %let pathFugohyo=¥全消¥擬似ミクロ作成プロジェクト¥★新情報¥最新情報¥全消2004年新擬似ミクロデータSAS版;
  %let fugohyoEXCEL=全消2004年新擬似ミクロデータ符号表.xlsx;   *★符号表エクセルファイル名を指定する;
  libname newZ "&drive:&path"; *★この行はそのままにしておく;
```

② setALL.sas

```
/* setALL.sas */  *★3つのSASプログラムを実行;
%include "&drive:&pathSASprogram¥createFormatNewZenshoGiji2004.sas" *③;
%include "&drive:&pathSASprogram¥createLabelNewZenshoGiji2004.sas";   *④;
%include "&drive:&pathSASprogram¥setLABELweight.sas";                 *⑤;
```

③ createFormatNewZenshoGiji2004.sas

```
/* createformatNewZenshoGiji2004.sas */ *★新擬似ミクロデータ2004用proc formatを自動作成する;

filename out1 "&drive:¥★X60s(2014-8-9)¥SAS_Forum¥2017¥SASprogram¥procFormatNewZenshoGiji2004.sas";

proc import datafile="&drive:&pathFugohyo¥&fugohyoEXCEL"
  out=formatNewZenshoGiji2004 (rename=(f2=comment f3=varName f4=V f5=meaning))
  replace;
  sheet='世帯事項整形'; getnames=no;
run;

proc print data=formatNewZenshoGiji2004; title "data=formatNewZenshoGiji2004 符号表"; run;

data _NULL_; set formatNewZenshoGiji2004;
  file out1; *★proc formatの自動作成★;
  length buff $ 100;
  if _N_=1 then put '/* procFormatNewZenshoGiji2004.sas */';
  if _N_<=4 then return; *最初の4オブザベーションはスキップする;
  if varName="Weight" then do; put ';' / 'run;'; return; end; *最後の集計用乗率は無視する;
  if _N_=5 then put "proc format;";
  formatname=compress(varName || 'F'); *format名の末尾が数字は反則なので、Fを末尾に付す;
  if varName NE " " then put ';' / '*' comment ';' / 'value ' formatname;
  buff=V||'="'||V||':'||meaning||'"'; buff=kcompress(buff); put buff;
run;
```

④ createtLabelNewZenshoGiji2004.sas

```
/* setLABELweight.sas */ *★全消2004年新擬似ミクロデータ:データコンペ規定課題用★データセット★設定;
*★変数ラベル付加、及び、10万世帯比weight変数を作成して、新データセット★NZ2004★を新規作成;

libname newZ "&drive:&path";

data newZ.NZ2004; set newZ.zensho2004gijimicro  end=owari;
  %include "&drive:¥&pathSASprogram¥labelNewZenshoGiji2004.txt";*変数LABEL付与;
  cnt=1; *集計乗率利用時に使用する変数;
  sum_weight+weight; *★10万比の集計乗率を計算する準備;
  if owari then call symputx("sum_weight", sum_weight); *★集計乗率の合計値をmacro変数sum_weightに設定;
run;

%put "&sum_weight"; *★macro変数sum_weightの値の確認;

data newZ.NZ2004; set newZ.NZ2004 end=owari; drop total;
  weight100000=weight*100000/&sum_weight; *★10万比の集計乗率を計算;
  total+weight100000; if owari then put "検算:weight100000の合計＝" total;
  *★LOG画面出力結果→検算:weight100000の合計＝99999.999999;
  *★10万比で計算の際は、集計乗率変数weight100000を使えば簡単;
run; *★以後、規定課題用プログラム作成にはSASデータセット「newZ.NZ2004」を使用する;
```

⑤ setLABELweight.sas

```
/* setLABELweight.sas */ *★全消2004年新擬似ミクロデータ:データコンペ規定課題用★データセット★設定;
*★変数ラベル付加、及び、10万世帯比weight変数を作成して、新データセット★NZ2004★を新規作成;

libname
newZ "&drive:¥全消¥擬似ミクロ作成プロジェクト¥★新情報¥最新情報¥全消2004年新擬似ミクロデータSAS版";

data newZ.NZ2004; set newZ.zensho2004gijimicro  end=owari;
  %include "&drive:¥&pathSASprogram¥labelNewZenshoGiji2004.txt";*変数LABEL付与;
  cnt=1; *集計乗率利用時に使用する変数;
  sum_weight+weight; *★10万比の集計乗率を計算する準備;
  if owari then call symputx("sum_weight", sum_weight); *★集計乗率の合計値をmacro変数sum_weightに設定;
run;

%put "&sum_weight"; *★macro変数sum_weightの値の確認;

data newZ.NZ2004; set newZ.NZ2004 end=owari; drop total;
  weight100000=weight*100000/&sum_weight; *★10万比の集計乗率を計算;
  total+weight100000; if owari then put "検算：weight100000の合計＝" total;
  *★LOG画面出力結果→検算：weight100000の合計＝99999.999999;
  *★10万比で計算の際は、集計乗率変数weight100000を使えば簡単;
run; *★以後、規定課題用プログラム作成にはSASデータセット「newZ.NZ2004」を使用する;
```

⑥ procFormatNewZenshoGiji2004.sas　（自動作成された）

```
/* procFormatNewZenshoGiji2004.sas */
proc format;
;
*大都市圏の別 ;
value X01F
1="1:３大都市圏"
0="0:その他"
;
*世帯区分 ;
value X02F
1="1:勤労者"
2="2:勤労者以外"
3="3:無職"
;
*世帯人員 ;
value X03F
1="1:１人"
2="2:２人"
3="3:３人"
4="4:４人"
5="5:５人以上"
;
*就業人員 ;
value X04F
0="0:０人"
1="1:１人"
2="2:２人"
3="3:３人以上"
;
*住居の構造 ;
value X05F
1="1:木造(防火木造含む)"
2="2:木造(防火木造含む)以外"
;
*住居の建て方 ;
value X06F
1="1:一戸建"
2="2:一戸建以外"
;

*住居の所有関係 ;
value X07F
1="1:持ち家"
2="2:持ち家以外"
;
*世帯主の性別 ;
value X08F
1="1:男"
2="2:女"
;
*世帯主の年齢 ;
value X09F
5="5:24歳以下"
6="6:25～29歳"
7="7:30～34歳"
8="8:35～39歳"
9="9:40～44歳"
10="10:45～49歳"
11="11:50～54歳"
12="12:55～59歳"
13="13:60～64歳"
14="14:65～69歳"
15="15:70～74歳"
16="16:75歳以上"
;
*企業区分･従業者規模 ;
value X10F
1="1:民営･自営1～4人"
2="2:民営･自営5～29人"
3="3:民営･自営30～499人"
4="4:民営･自営500人以上"
5="5:官公"
6="6:無職"
;

*家族分類 ;
value X11F
1="1:単身世帯"
2="2:夫婦のみ世帯"
3="3:二世代世帯"
4="4:二世代(ひとり親)世帯"
5="5:三世代世帯"
6="6:その他の世帯"
;
*未就学児の有無 ;
value X12F
1="1:無"
2="2:有"
;
*学校に通う世帯員の有無 ;
value X13F
1="1:無"
2="2:有"
;
*65歳以上の世帯員数 ;
value X14F
0="0:０人"
1="1:１人"
2="2:２人以上"
;
run;
```

⑦ kitei2017.sas

規定課題の問1～問4用のプログラム

```
/* kitei2017.sas */ *★set_environment.sasを実行済みか確認せよ; options nocenter;
options MPRINT;

data kitei; set newZ.NZ2004; *作業用データセット「kitei」を作成;
  rename X02=HHkubun X08=sex  X09=HHage  X11=HHkind ;*使う変数を間違わないように;
        *世帯区分  世帯主性別 世帯主年齢  世帯分類;
run;

*★規定課題1★; *世帯主の年齢階層(X09)×世帯区分(X02);

proc freq data=kitei; tables HHage*HHkubun / norow nocol nopercent format=comma7.0; label;
  format HHage   X09F.;
  format HHkubun X02F.;
  title "★表1左半分：集計乗率なし"; run;

proc freq data=kitei; tables HHage*HHkubun / norow nocol nopercent format=comma7.0; label;
  format HHage   X09F.;
  format HHkubun X02F.;
  title "★表1右半分：集計乗率あり"; weight weight100000; *★10万比乗率使用; run;

*★規定課題2★; * 世帯分類(x11)×世帯区分(X02)×世帯主の性別(X08);

title "①表2：整形前";
proc tabulate data=kitei format=comma7.0;
  class HHkind HHkubun sex;
  var cnt / weight=weight100000; *★10万比乗率使用;
  table (HHkind ALL),
        (HHkubun ALL)*(sex ALL)*cnt; *★defaultでSUMが自動指定される;
  format HHkind  X11F.;
  format HHkubun X02F.;
  format sex     X08F.;
run;

title "★①表2：整形後";
proc tabulate data=kitei format=comma7.0;
  class HHkind HHkubun sex;
  var cnt / weight=weight100000; *★10万比乗率使用;
  table (HHkind ALL="計"),
        (HHkubun="" ALL="計")*(sex="" ALL="計")*cnt=""*SUM="";
  format HHkind  X11F.;
  format HHkubun X02F.;
  format sex     X08F.;
run;

title "★①表3 weight100000";
data TB3 (keep=Y044 log10Y044 weight100000); set kitei; log10Y044=log10(Y044+1); run;

proc univariate data=TB3;
  var Y044 log10Y044; *パン支出金額;
  weight weight100000; *四分位点だけ指定できる★？平均値だけ指定できる？？？;
  label    Y044="パン支出金額(円)";
  label log10Y044="1を加えた金額の常用対数(円)"; *format指定できる？？？小数点第3位まで;
run;

title "★設問３：箱ひげ図";
proc sgplot data=TB3;
  HBOX Y044 / weight=weight100000 BOXWIDTH=0.2 meanattrs=(color=black symbol=plus);
  label Y044="パン支出";
run;

title "★設問３：ヒストグラム";
data TB3; set TB3; log10Y044=log10(sum(Y044,1)); run;
proc sgplot data=TB3;
  histogram log10Y044 / weight=weight100000 ; run;
quit;
```

（次頁に続く）

```
title "★設問４：表4 weight100000";
%macro ratio(No); if Y042 NE 0 then r&No=Y&No/Y042*100; else r&No=.; ; %mend;
                                                                    *★↑0値は欠損値に設定★;
data TB4(keep=Y042-Y046 weight weight100000 HHkubun HHkind diff r042-r046 cnt);
  set kitei;
  diff=Y042-sum(of Y043-Y046); label diff="穀類合計との差"; *データ確認,計算には不要;
  %ratio(042) %ratio(043) %ratio(044) %ratio(045) %ratio(046)
run;

proc print data=TB4 (obs=20); title "data=TB4 (obs=20)"; run;

title "★設問４：①表4：整形前";
proc tabulate data=TB4;
  var r042-r046 cnt / weight=weight100000;
  class HHkubun HHkind;
  table (HHkubun ALL) (HHkind ALL),
        cnt (r043-r046 r042)*MEAN;
run;

title "★設問４：①★表4：整形後";
proc tabulate data=TB4 out=table4;
  var r042-r046 cnt / weight=weight100000;
  class HHkubun HHkind;
  table (HHkubun ALL) (HHkind ALL),
        cnt="N"*SUM=""*F=comma7.0
       (r043="米%"*F=5.1
        r044="パン%"*F=5.1
        r045="めん類%"*F=5.1
        r046="他の麺%"*F=5.1
        r042="全体"*F=5.1)*MEAN="";
  keylabel ALL="全体";
  format HHkind  X11F.;
  format HHkubun X02F.;
run;

proc print data=table4 (obs=50); title "data=table4(1)"; run;

data table4; set table4; if HHkind NE .; run;

proc print data=table4; title "data=table4 "; run;

proc transpose data=table4 out=transTB4;
  var r043_Mean r044_Mean r045_Mean r046_Mean;
  by HHkind;
run;

proc print data=transTB4; title "data=transTB4 転置後"; run;

data transTB4; set transTB4; by HHkind;
  No+1; if No=5 then No=1; output;
run;

proc print data=transTB4; title "data=transTB4 変数No付加"; run;

proc format; *主食;
value SHUSHOKU
1="1:米"
2="2:パン"
3="3:めん類"
4="4:他の麺:"
; run;

title "★設問４：穀類の構成比の帯グラフ";
proc sgplot data=TRANSTB4;
   hbar HHkind / group=No response=COL1 stat=sum seglabel seglabelfitpolicy=NOCLIP;
   format No SHUSHOKU.;
   format COL1 F5.1;
   label COL1="穀類合計(100%)";
run;
quit;

*★設問５：ジニ係数とロレンツ曲線は別プログラムで処理する;
```

⑧ make_weighted_datasetX.sas

集計乗率を考慮してオブザベーションを膨らませ、かつ、必要な変数だけから成るデータセットを作成する。

```
/* make_weighted_dataset.sas */ options nocenter;
*Don't forget to run "set_environment.sas" for the 2nd macro specification below;
*Specify the following four macro variables;
*(1)dataset name;          %let dataset= NZ2004;
*(2)path for (1)dataset;  %let libname= &drive:&path; *Leave this as it is;
*(3)target variable;      %let var_name= Y001; *年間収入(単位:万円);
*(4)weight variable;      %let weight= weight;

libname gini "&libname";

proc summary data=gini.&dataset;
   output out=minmax(keep=_stat_ &weight); var &weight; run;
proc print data=minmax; title "min and max of &weight(1)"; run;
proc transpose data=minmax out=xminmax(keep=COL2 COL3 rename=(COL2=min COL3=max)); run;
proc print data=xminmax; title "min and max of &weight(2)"; run;

data gini;
  merge gini.&dataset xminmax;
  retain xmin;
  if _n_=1 then xmin=min;
  obs_no=int(&weight/xmin*10); *★膨らませるオブザベーションの数を計算;
  total_obs+obs_no;
run;

proc print data=gini(obs=100); title "gini(1)"; run;

data _NULL_; /*data total_obs;*/
  set gini end=final;
    if final then put total_obs=; *オブザベーション数の確認;
run;

data reconstruct;
  keep x11 &var_name repeat; *家族分類別、年間収入(単位:万円)、複製個数;
  set gini;
  do repeat=1 to obs_no; output; end;
run;

proc print data=reconstruct(obs=100);
title "data=reconstruct(obs=100)"; run;
```

⑨ gini_lorenz2017 規定課題_sas_forum_macro.sas

ジニ係数を計算し、ロレンツ曲線を作図する

```
/* gini_lorenz2017規定課題_sas_forum_macro.sas */ options nocenter;
*-------------------------------------------------------------------------------;
*★"make_weighted_dataset.sas"の実行直後に、同じセッションで実行する;
*データセット名と変数名を指定して下さい。;
*SAS dataset名; %let dataset=&ds;
/**libname      ; %let libname=&drive:&path;*/
*変数名        ; %let var_name=Y001;
*-------------------------------------------------------------------------------;
*●アウトプット画面にgini係数だけを表示したい場合は、
   次のmacro変数commentに「*」を指定して下さい。
 ●計算経過も表示したい場合は半角ブランクを指定して下さい。;
%let comment= *;
*-------------------------------------------------------------------------------;
/*libname gini "&libname";*/

data original;
  keep &var_name;
  set work.&dataset;
  rename &var_name=income;
run;

/* Calculate gine coefficient */
proc sort data=original; by income; run;

&comment.proc print; title "(1)original"; &comment.run;

data income(keep=income cumm_income)
     cumm_income(keep=cumm_income rename=(cumm_income=total_income));
  set original end=last;

  if _n_=1 then do; save_income=income;
                    income=0; cumm_income=0;
                    output income;
                    income=save_income;
         end; /* Add "income=0" on top of dataset  "INCOME" */

  cumm_income+income;
  output income;
  if last then output cumm_income;
run;

&comment.proc print data=income;      title "(2)income"; &comment.run;
&comment.proc print data=cumm_income; title "(3)cumm_income"; &comment.run;

data standard_cumm_income;
  drop income;
  merge income cumm_income;
run;

&comment.proc print data=standard_cumm_income; title "(4)standard_cumm_income(1)";
&comment.run;

data standard_cumm_income;
  keep standard_cumm_income;
  set;
  retain xtotal_income;

  if _n_=1 then xtotal_income= total_income;
  standard_cumm_income=cumm_income/xtotal_income;
run;
```

注:周防(2008)の付録3のプログラムを使用した。

(次頁に続く)

```
&comment.proc print; title "(4)standard_cumm_income(2)"; &comment.run;

data trapezium;
  keep shorter standard_cumm_income; rename standard_cumm_income=longer;
  set;
  retain shorter 0;
  output;
  shorter=standard_cumm_income; /*To let "longer" to be "shorter" for next obs.*/
run;

&comment.proc print; title "(5)trapezium"; &comment.run;

proc datasets library=work; contents data=original out=no_of_obs(keep=nobs) noprint;
run; /* To count the number of observations in the dataset "original".*/

&comment.proc print data=no_of_obs; title "(6)no_of_obs"; &comment.run;

data gini;
  *keep shorter longer total_no under_bow total_under_bow bow gini;
  keep gini;
  merge  trapezium no_of_obs end=last;
  retain total_no;
  if _n_=1 then total_no=nobs;

  under_bow=(shorter+longer)*(1/total_no)/2;
  total_under_bow+under_bow;

  if last then do; bow=0.5-total_under_bow;
                        gini=bow*2;
                        output; * Make this OUTPUT valid, if you want only gini;
               end;
run;

proc print data=gini;
  title "(7)gini: variable 'gini' is the gini coefficient.";
  title2 "家族分類:&No";
  format gini F4.2;
run;

/* Draw Lorentz Curve */

data lorenz;
  keep standard_cumm_income diagonal n;
  retain total_no;
  rename standard_cumm_income=&var_name; /*income -> &var_name*/

  merge standard_cumm_income no_of_obs;

  if _n_=1 then total_no=nobs;
  n=(_n_-1)/total_no;
  diagonal=n; /* To draw a diagonal line */
run;

&comment.proc print; title "(8)lorenz"; &comment.run;

goptions reset=all htext=0.5cm vsize=8cm hsize=8cm;

symbol i=spline f=xswiss height=1;  /* i=join or spline*/

axis1 length=4.5cm label=(height=0.3cm f=simplex) value=(height=0.3cm f=simplex) offset=(0,0);
axis2 length=4.5cm label=(height=0.3cm f=simplex) value=(height=0.3cm f=simplex) offset=(0,0);

ods listing close;
ods rtf  file="Lorenz&No..rtf";
proc gplot; plot &var_name*n diagonal*n / overlay haxis=axis1 vaxis=axis2;
      title "Lorenz Curve:&No";
run; quit;
ods rtf close;
ods listing;
```

⑩ GiniLorenz2017.sas

```
/* GiniLorenz2017.sas */ *★gini係数の計算とLorenz曲線の作図;

*データセット"reconstruct"を家族分類毎のデータセット(D1～D6)に分割する;
%macro divide(No); data d&No ; set reconstruct; if X11=&No; run; %mend;
%macro repeat; %do No=1 %to 6; %divide(&No); %end; %mend;
%repeat;

*★家族分類1～6毎に、gini係数の計算とLorenz曲線の作図;
%macro GiniLorenz(No);
  %let ds=d&No;
  %include "&drive:¥&pathSASprogram¥gini_lorenz2017規定課題_sas_forum_macro.sas";
%mend;

%GiniLorenz(1)
%GiniLorenz(2)
%GiniLorenz(3)
%GiniLorenz(4)
%GiniLorenz(5)
%GiniLorenz(6)

*★全ての家族分類のgini係数の計算とLorenz曲線の作図;
%let No=全体; %let ds=reconstruct;
%include "&drive:¥&pathSASprogram¥gini_lorenz2017規定課題_sas_forum_macro.sas";
```

⑪ labelNewZenshoGiji2004.txt

プログラム④createtLabelNewZenshoGiji2004.sasが作成したラベル文。
プログラム⑤setLABELweight.sasの中でラベル文として挿入される。

```
label Year ="調査年";
label X01 ="大都市圏の別";
label X02 ="世帯区分";
label X03 ="世帯人員";
label X04 ="就業人員";
label X05 ="住居の構造";
label X06 ="住居の建て方";
label X07 ="住居の所有関係";
label X08 ="世帯主の性別";
label X09 ="世帯主の年齢";
label X10 ="企業区分･従業者規模";
label X11 ="家族分類";
label X12 ="未就学児の有無";
label X13 ="学校に通う世帯員の有無";
label X14 ="65歳以上の世帯員数";
label Y001 ="年間収入";
label Y002 ="収入総額";
label Y003 ="実収入";
label Y004 ="経常収入";
                中略
label Y013 ="内職収入";
label Y014 ="本業以外の勤め先･事業･内職収入";
label Y015 ="他の経常収入";
label Y016 ="財産収入";
                中略
label Y043 ="米";
label Y044 ="パン";
label Y045 ="めん類";
label Y046 ="他の穀類";
label Y047 ="魚介類";
label Y048 ="生鮮魚介";
                中略
label Y056 ="牛乳";
label Y057 ="乳製品";
label Y058 ="卵";
                中略
label Y132 ="(特掲)ガソリン";
label Y133 ="(特掲)自動車整備費";
label Y134 ="(特掲)自動車保険料";
label Y135 ="通信";
label Y136 ="(特掲)移動電話通信料";
label Y137 ="教育";
label Y138 ="授業料等";
label Y139 ="教科書･学習参考教材";
label Y140 ="補習教育";
label Y141 ="教養娯楽";
label Y142 ="教養娯楽用耐久財";
label Y143 ="教養娯楽用品";
label Y144 ="書籍･他の印刷物";
label Y145 ="教養娯楽サービス";
label Y146 ="宿泊料";
label Y147 ="パック旅行費";
label Y148 ="月謝類";
label Y149 ="他の教養娯楽サービス";
label Y150 ="(特掲)インターネット接続料";
                中略
label Y159 ="交際費";
label Y160 ="交際費(食料)";
label Y161 ="交際費(家具･家事用品)";
label Y162 ="交際費(被服及び履物)";
                中略
label Y189 ="他の非消費支出";
label Y190 ="実支出以外の支出";
label Y191 ="預貯金";
label Y192 ="保険掛金";
label Y193 ="個人･企業年金保険掛金";
label Y194 ="他の保険掛金";
                中略
label Y202 ="その他の実支出以外の支出";
label Y203 ="繰越金";
```

付録2　規定問題解答出力結果

問1

以下の各図表の上部にあるタイトルは、プログラムで指定したtilte文

★表1左半分：集計乗率なし

表：HHage * HHkubun				
Hhage (世帯主の年齢)	HHkubun(世帯区分)			
	1:勤労者	2:勤労者以外	3:無職	合計
5:24歳以下	765	18	15	**798**
6:25～29歳	2,289	157	66	2,512
7:30～34歳	4,223	511	107	4,841
8:35～39歳	5,217	728	193	6,138
9:40～44歳	5,492	1,076	205	6,773
10:45～49歳	5,702	1,305	277	7,284
11:50～54歳	6,194	1,797	392	8,383
12:55～59歳	5,682	2,221	624	8,527
13:60～64歳	3,237	2,226	2,534	7,997
14:65～69歳	1,343	1,743	3,590	6,676
15:70～74歳	469	1,197	3,214	4,880
16:75歳以上	166	940	3,216	4,322
合計	**40,779**	13,919	14,433	**69,131**

★表1右半分：集計乗率あり

表：HHage * HHkubun				
Hhage (世帯主の年齢)	HHkubun(世帯区分)			
	1:勤労者	2:勤労者以外	3:無職	合計
5:24歳以下	**1,729**	10	21	**1,760**
6:25～29歳	4,314	151	57	4,522
7:30～34歳	6,441	480	87	7,008
8:35～39歳	7,420	824	153	8,396
9:40～44歳	7,611	1,145	247	9,004
10:45～49歳	7,739	1,479	244	9,462
11:50～54歳	8,072	2,085	508	10,665
12:55～59歳	7,697	2,725	893	11,315
13:60～64歳	4,208	2,785	4,160	11,153
14:65～69歳	1,487	2,213	6,374	10,074
15:70～74歳	459	1,652	6,124	8,235
16:75歳以上	207	1,325	6,873	8,406
合計	**57,386**	16,874	25,741	**100,000**

問2

★①表2：整形後

	1:勤労者			2:勤労者以外			3:無職			計		
	1:男	2:女	計	1:男	2:女	計	1:男	2:女	計	1:男	2:女	計
家族分類												
1:単身世帯	6,503	4,207	10,710	1,030	867	1,897	2,463	5,688	8,151	9,995	10,763	20,758
2:夫婦のみ世帯	8,517	324	8,841	4,293	81	4,375	11,026	45	11,071	23,837	450	24,287
3:二世代世帯	27,267	378	27,645	6,190	51	6,242	3,264	32	3,295	36,721	461	37,182
4:二世代(ひとり親)世帯	1,543	2,421	3,963	934	430	1,364	671	1,303	1,974	3,148	4,153	7,302
5:三世代世帯	4,956	384	5,340	2,390	118	2,508	638	160	798	7,984	662	8,646
6:その他の世帯	654	234	887	413	75	488	278	174	452	1,344	482	1,826
計	49,439	7,947	57,386	15,250	1,623	16,874	18,340	7,401	25,741	83,029	16,971	100,000

問3

★①表3 weight100000

UNIVARIATE プロシジャ

変数：Y044 (パン支出金額(円))

白抜きの文字の箇所が解答に対応する。

重み変数：weight100000

モーメント (重み付き)			
N	69131	重み変数の合計	100000
平均	2284.02283	合計	228402283
標準偏差	2631.01273	分散	6922227.97
歪度	4.25528372	尖度	78.8350941
無修正	1.00E+12	修正済平方和	4.79E+11
変動係数	115.19205	平均の標準誤差	8.31999277

基本統計量 (重み付き)			
位置		ばらつき	
平均	2284.023	標準偏差	2631
中央値	1689	分散	6922228
最頻値	0	範囲	51776
		四分位範囲	2181

分位点 (重み付き)	
レベル	分位点
100% 最大値	51776
99%	10300
95%	6275
90%	4835
75% Q3	3039
50% 中央値	1689
25% Q1	858
10%	390
5%	202
1%	0
0% 最小値	0

★①表3 weight100000

UNIVARIATE プロシジャ

変数：log10Y044 (1を加えた金額の常用対数(円))

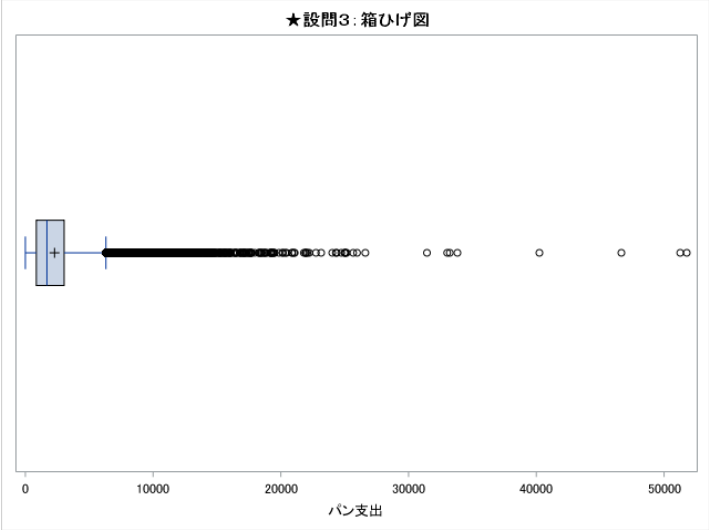
重み変数：weight100000

モーメント (重み付き)			
N	69131	重み変数の合計	100000
平均	3.13665089	合計	313665.089
標準偏差	0.7185186	分散	0.51626898
歪度	-4.6970247	尖度	48.7932748
無修正平方和	1019547.56	修正済平方和	35689.6749
変動係数	22.9071908	平均の標準誤差	0.00227216

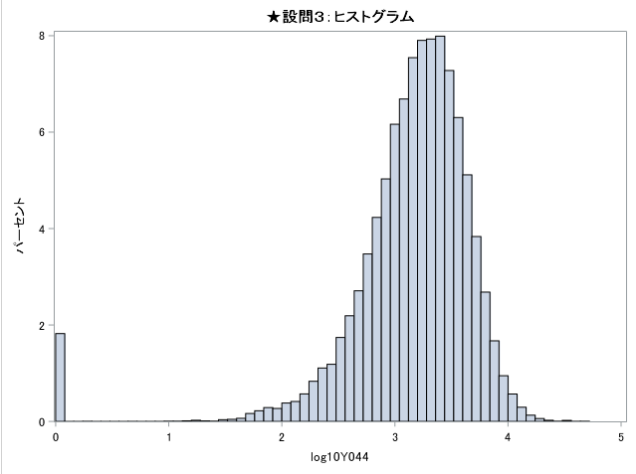
基本統計量 (重み付き)			
位置		ばらつき	
平均	3.136651	標準偏差	0.71852
中央値	3.227887	分散	0.51627
最頻値	0	範囲	4.71414
		四分位範囲	0.54888

分位点 (重み付き)	
レベル	分位点
100% 最大値	4.71414
99%	4.01288
95%	3.79768
90%	3.68449
75% Q3	3.48287
50% 中央値	3.22789
25% Q1	2.93399
10%	2.59218
5%	2.3075
1%	0
0% 最小値	0

問3 ②



問3 ③

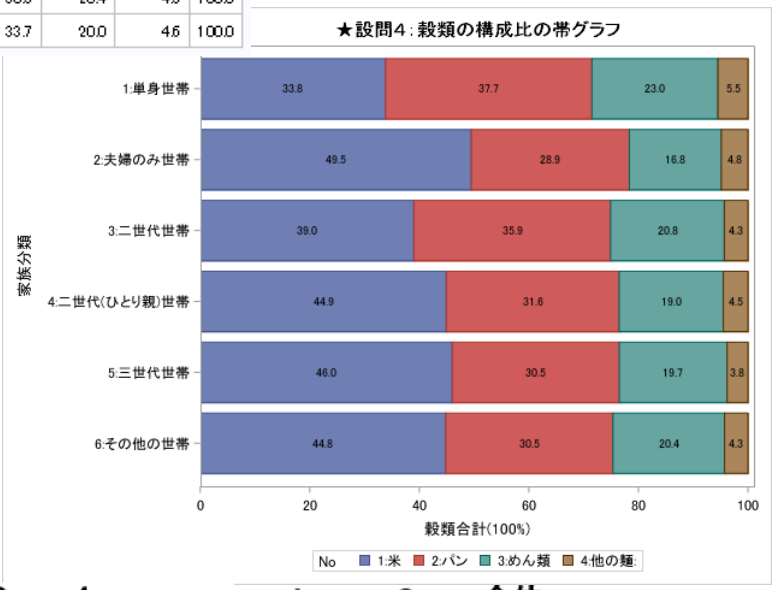


問4 ①

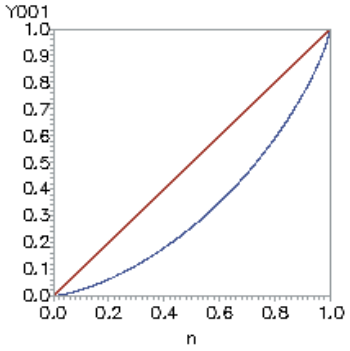
★設問4:①★表4:整形後

	N	米%	パン%	めん類%	他の麺%	全体
世帯区分						
1:勤労者	57,386	36.0	37.6	22.1	4.3	100.0
2:勤労者以外	16,874	44.9	31.0	19.4	4.7	100.0
3:無職	25,741	52.1	26.7	15.8	5.4	100.0
全体	100,000	41.6	33.7	20.0	4.6	100.0
家族分類						
1:単身世帯	20,758	33.8	37.7	23.0	5.5	100.0
2:夫婦のみ世帯	24,287	49.5	28.9	16.8	4.8	100.0
3:二世代世帯	37,182	39.0	35.9	20.8	4.3	100.0
4:二世代(ひとり親)世帯	7,302	44.9	31.6	19.0	4.5	100.0
5:三世代世帯	8,646	46.0	30.5	19.7	3.8	100.0
6:その他の世帯	1,826	44.8	30.5	20.4	4.3	100.0
全体	100,000	41.6	33.7	20.0	4.6	100.0

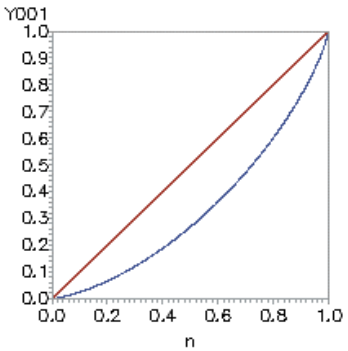
問4 ②



問5 ① Lorenz Curve:4



Lorenz Curve:全体



参考

問

(7)gini: variable 'gini' is the gini coefficient.

家族分類:1

OBS	gini
1	0.34

家族分類:2

OBS	gini
1	0.31

家族分類:3

OBS	gini
1	0.26

家族分類:4

OBS	gini
1	0.35

家族分類:5

OBS	gini
1	0.27

家族分類:6

OBS	gini
1	0.33

家族分類:全体

OBS	gini
1	0.33

付録3 2017 年 第5回 SAS ミクロデータ分析コンテスト 規定課題

1）年齢階層別・世帯区分別 クロス表

世帯主の年齢階層別・世帯区分別に集計用乗率なしのクロス集計表と、集計用乗率を使った 10 万世帯比のクロス集計表を左右に並べて、表 1 の様式で作成する。

表1

	世帯区分					世帯区分			
世帯主の年齢	勤労者	勤労者以外	無職	合計	世帯主の年齢	勤労者	勤労者以外	無職	合計
24歳以下	765			798	24歳以下	1,729			1,760
25～29歳					25～29歳				
30～34歳					30～34歳				
30～35歳					30～35歳				
40～44歳					40～44歳				
45～49歳					45～49歳				
50～54歳					50～54歳				
55～59歳					55～59歳				
60～64歳					60～64歳				
65～69歳					65～69歳				
70～74歳					70～74歳				
75歳以上					75歳以上				
合計	40,779			69,131	合計	57,386			100,000

2）世帯分類別・世帯区分別 世帯主の性別 3 重クロス表

世帯分類別・世帯区分別・世帯主の性別の 3 重クロス集計表を、集計用乗率を使って 10 万世帯比で表 2 の様式で作成する。

表 2

	勤労者世帯			勤労者以外の世帯			無職世帯			
世帯分類	男	女	計	男	女	計	男	女	計	合計
単身世帯	6,503									20,758
夫婦のみ										
二世代										
二世代(ひとり親)										
三世代										
その他										
合計	49,439									100,000

3）パン支出金額についての各種の統計グラフ

① パン支出金額(変数名:Y044)及び、パン支出金額に 1 を加えた金額の常用対数変換 log10(Y044+1) を行い、集計用乗率を使って 10 万世帯比について、それぞれの四分位統計量および平均値を表 3 に示す様式で作成する。

② パン支出金額について箱ひげ図を作成する。

③ パン支出金額に 1 を加えた金額の常用対数変換値について、ヒストグラムを作成する。

表 3

統計量	パンの支出金額(円)	1を加えた金額の常用対数
最大値	51,776	4.714
第3四分位数		
中央値		
第1四分位数		
最小値		
平均値		
件数	100,000	100,000

4) 世帯区分別および世帯分類別の穀類(米、パン、めん類、他の穀類)の支出金額についての構成比

「穀類」の支出金額(変数名:Y042)は、米(Y043)、パン(Y044)、めん類(Y045)、他の穀類(Y046)の合計金額である。

① これら 4 変数が穀類(Y042)に占める構成比(パーセント)を世帯ごとに計算した後、10 万世帯比による世帯区分別、及び、世帯分類別に構成比の平均値を、表 4 に示す様式で作成する。

表4

			構成比 (%)				
項目名	符号内容	件数	米	パン	めん類	他の穀類	全体(%)
世帯区分	勤労者世帯	57,386	36.0				100.0
	勤労者以外の世帯						
	無職世帯						
世帯分類	単身世帯	20,758	33.8				100.0
	夫婦のみ						
	二世代						
	二世代(ひとり親)						
	三世代						
	その他						
	全体	100,000	41.7				100.0

② 世帯分類別に、穀類(米、パン、めん類、他の穀類)の構成比を帯グラフで表示する。

5)世帯分類別のローレンツ曲線およびジニ係数

① 集計用乗率を考慮して、世帯分類「二世代(ひとり親)」世帯の年間収入(Y001)のローレンツ曲線を描く。

② 集計用乗率を考慮して、6つの世帯分類の年間収入(Y001)のジニ係数(小数点第2位まで)を計算して、表 5 の様式で表示する。

表 5

世帯分類	ジニ係数
単身世帯	
夫婦のみ	
二世代	
二世代(ひとり親)	
三世代	
その他	
全体	0.33