Занятие 2. Перекрестное использование данных семейного и индивидуального файла.

Общая рекомендация ко всем выполняемым вами заданиям:

1) Сохраняйте исходные файлы под новым именем, чтобы работать с ними.

2) СРАЗУ открывайте и сохраняйте файл аутпута. Первая команда в нем должна быть вида:

***Фамилия – номер семинара – номер задания**

3) Сохраняйте проделанную вами работу в виде кода, используя «сохранение» правильных команд в STATA (или функцию “paste” SPSS). В этом случае вы сможете дома повторить все сделанное вами в классе. Кроме того, рекомендуется прикладывать программу к вашим исследованиям.

4) В качестве отчета за семинар нужно предъявить созданные файлы данных, файл аутпута с вашей фамилией, и файл с кодом.

Исходные файлы.

Данные 24й волны

i15_s1.dta – индивидуальный файл, преобразованный на первом занятии

r15hall32.dta – домохозяйственный файл

7. Описательные характеристики количественной переменной.

7.0. Загрузите файл данных **i15_s1.dta**

7.1. Рассмотрим некоторые переменные о занятости на первом месте работы

codebook kj1 kj7 kj8_1 kj8_3 kj9 kj14 kj16 kj17 kj8 kj8_2 kj10 kj15 kj18

```
            Tabulation: Freq.   Numeric  Label
                        6,547         1  Вы сейчас работаете
                          184         2  Вы находитесь в отпуске -
                                         декретном или по у
                           37         3  Вы находитесь в любом другом
                                         оплачиваемом о
                            7         4  Вы находитесь в неоплачиваемом
                                         отпуске
                        5,706         5  Или у Вас сейчас нет работы
                            5   1.0e+08  ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
                            4   1.0e+08  ОТКАЗ ОТ ОТВЕТА
                        2,199         .

-------------------------------------------------------------------------------
kj7       Вы работали по основному месту работы в течение последних 30 дней?
-------------------------------------------------------------------------------

                  Type: Numeric (double)
                 Label: kj7

                 Range: [1,99999999]                  Units: 1
         Unique values: 5                         Missing .: 7,914/14,689

            Tabulation: Freq.   Numeric  Label
                        6,406         1  Да
                          361         2  Нет
                            5   1.0e+08  ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
                            1   1.0e+08  ОТКАЗ ОТ ОТВЕТА
                            2   1.0e+08  НЕТ ОТВЕТА
                        7,914         .
```

```
-------------------------------------------------------------------------------
kj8_1     Скажите, пожалуйста, Вы работали по основной работе дома в течение
-------------------------------------------------------------------------------

                  Type: Numeric (double)
                 Label: kj8_1

                 Range: [1,99999999]                  Units: 1
         Unique values: 5                         Missing .: 8,283/14,689

            Tabulation: Freq.   Numeric  Label
                          460         1  Да
                        5,939         2  Нет
                            3   1.0e+08  ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
                            1   1.0e+08  ОТКАЗ ОТ ОТВЕТА
                            3   1.0e+08  НЕТ ОТВЕТА
                        8,283         .

-------------------------------------------------------------------------------
kj8_3     Скажите, пожалуйста, Вы учитывали эти часы, когда называли общую п
-------------------------------------------------------------------------------

                  Type: Numeric (double)
                 Label: kj8_3

                 Range: [1,99999999]                  Units: 1
         Unique values: 5                         Missing .: 14,229/14,689

            Tabulation: Freq.   Numeric  Label
                          100         1  Да
                          343         2  Нет
                           12   1.0e+08  ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
                            2   1.0e+08  ОТКАЗ ОТ ОТВЕТА
                            3   1.0e+08  НЕТ ОТВЕТА
                       14,229         .

-------------------------------------------------------------------------------
kj9        Скажите, пожалуйста, по основному месту работы в течение последни
-------------------------------------------------------------------------------

                  Type: Numeric (double)
                 Label: kj9

                 Range: [1,99999999]                  Units: 1
         Unique values: 5                         Missing .: 7,914/14,689

            Tabulation: Freq.   Numeric  Label
                        6,206         1  Да
                          556         2  Нет
                            3   1.0e+08  ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
                            9   1.0e+08  ОТКАЗ ОТ ОТВЕТА
                            1   1.0e+08  НЕТ ОТВЕТА
                        7,914         .

-------------------------------------------------------------------------------
kj14       В настоящее время Ваше предприятие осталось должно Вам какие-то д
-------------------------------------------------------------------------------

                  Type: Numeric (double)
                 Label: kj14

                 Range: [1,99999999]                  Units: 1
         Unique values: 5                         Missing .: 8,443/14,689

            Tabulation: Freq.   Numeric  Label
                          545         1  Да
                        5,682         2  Нет
                           13   1.0e+08  ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
                            1   1.0e+08  ОТКАЗ ОТ ОТВЕТА
                            5   1.0e+08  НЕТ ОТВЕТА
                        8,443         .
```

```
-------------------------------------------------------------------------------
kj16            За сколько месяцев предприятие не доплатило Вам эти деньги?
-------------------------------------------------------------------------------

                 Type: Numeric (double)
                Label: kj16, but 18 nonmissing values are not labeled

                Range: [0,99999999]                  Units: 1
        Unique values: 21                         Missing .: 14,144/14,689

             Examples: .
                       .
                       .
                       .

-------------------------------------------------------------------------------
kj17     Вы получали в течение последних 30 дней на этом предприятии в качес
-------------------------------------------------------------------------------

                 Type: Numeric (double)
                Label: kj17

                Range: [1,99999999]                  Units: 1
        Unique values: 5                          Missing .: 8,443/14,689

           Tabulation: Freq.   Numeric  Label
                         159         1  Да
                       6,081         2  Нет
                           1   1.0e+08  ЗАТРУДНЯЮСЬ ОТВЕТИТЬ
                           1   1.0e+08  ОТКАЗ ОТ ОТВЕТА
                           4   1.0e+08  НЕТ ОТВЕТА
                       8,443         .

-------------------------------------------------------------------------------
kj8         Сколько часов Вы фактически отработали по основному месту работы
-------------------------------------------------------------------------------

                 Type: Numeric (double)
                Label: kj8, but 246 nonmissing values are not labeled

                Range: [4,99999999]                  Units: 1
        Unique values: 249                        Missing .: 8,283/14,689

             Examples: 170
                       336
                       .
                       .

-------------------------------------------------------------------------------
kj8_2      Сколько часов Вы фактически занимались этой работой дома в течени
-------------------------------------------------------------------------------

                 Type: Numeric (double)
                Label: kj8_2, but 62 nonmissing values are not labeled

                Range: [1,99999999]                  Units: 1
        Unique values: 64                         Missing .: 14,229/14,689

             Examples: .
                       .
                       .
                       .

-------------------------------------------------------------------------------
kj10      Сколько денег в течение последних 30 дней Вы получили по основному
-------------------------------------------------------------------------------

                 Type: Numeric (double)
                Label: kj10, but 435 nonmissing values are not labeled

                Range: [50,99999999]                 Units: 1
        Unique values: 438                        Missing .: 8,483/14,689
```

```
              Examples: 6000
                        30000
                        .
                        .

-------------------------------------------------------------------------------
kj15                                  Сколько всего денег Вам не доплатили?
-------------------------------------------------------------------------------

                  Type: Numeric (double)
                 Label: kj15, but 100 nonmissing values are not labeled

                 Range: [20,99999999]                 Units: 1
         Unique values: 103                       Missing .: 14,144/14,689

              Examples: .
                        .
                        .
                        .

-------------------------------------------------------------------------------
kj18       Оцените, пожалуйста, сколько стоит в рублях продукция, которую Вы
-------------------------------------------------------------------------------

                  Type: Numeric (double)
                 Label: kj18, but 51 nonmissing values are not labeled

                 Range: [15,99999998]                 Units: 1
         Unique values: 53                        Missing .: 14,530/14,689

              Examples: .
                        .
                        .
                        .
```

Задача 1: для занятых и тех, кто работал по основной работе за последние 30 дней, суммировать часы работы и часы работы дома **kj8 kj8_2**; удалить нереалистичные значения.
Задача 2: для занятых и тех, кто работал по основной работе за последние 30 дней, суммировать основной заработок, средний заработок в месяц по задержанной заработной плате, и деньги, которые можно было получить за продукцию, полученную в качестве оплаты («контрактная заработная плата») **kj10 kj15 kj18.**
Задача 3. Рассчитать ставку заработной платы и ее логарифм.

7.2. Так как в данном случае для нас не очень важно, почему в вопросе нет ответа (человек затруднился, отказался, или по другой причине), перекодируем во всех интересующих нас переменных 99999997 99999998 99999999 в «миссинг» (точку, пропущенное значение).

recode kj1 kj7 kj8 kj8_1 kj8_2 kj8_3 kj9 kj10 kj14 kj15 kj16 kj17 kj18 (99999997 = .) (99999998 = .) (99999999 = .)

Я рекомендую также всегда перекодировать дихотомические переменные – «2» (нет) в «0».

recode kj7 kj8_1 kj8_3 kj9 kj14 kj17 (2=0)
label define kj7 0 "нет", add
label define kj8_1 0 "нет", add
label define kj8_3 0 "нет", add
label define kj9 0 "нет", add
label define kj14 0 "нет", add
label define kj17 0 "нет", add

Нельзя сделать такую команду одну для всех переменных, то есть

***label define kj7 kj8_1 kj8_3 kj9 kj14 kj17 0 "нет", add – не годится**

7.3. Посмотрим распределения номинальных переменных для взрослых

tabulate kj1 if k_adult == 1, missing
tabulate kj7 if k_adult == 1, missing
tabulate kj8_1 if k_adult == 1, missing
tabulate kj8_3 if k_adult == 1, missing
tabulate kj9 if k_adult == 1, missing
tabulate kj14 if k_adult == 1, missing
tabulate kj16 if k_adult == 1, missing
tabulate kj17 if k_adult == 1, missing

```
. tabulate kj1 if k_adult == 1, missing

                       Ваше основное |
                    занятие в настоящее |
                                время? |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
                  Вы сейчас работаете |      6,547       52.42       52.42
Вы находитесь в отпуске - декретном или |        184        1.47       53.89
Вы находитесь в любом другом оплачиваем |         37        0.30       54.19
 Вы находитесь в неоплачиваемом отпуске |          7        0.06       54.24
            Или у Вас сейчас нет работы |      5,706       45.68       99.93
                                      . |          9        0.07      100.00
---------------------------------------+-----------------------------------
                                  Total |     12,490      100.00

.
. tabulate kj7 if k_adult == 1, missing

                  Вы |
             работали |
                   по |
             основному |
                 месту |
              работы в |
               течение |
             последних |
              30 дней? |      Freq.     Percent        Cum.
----------------------+-----------------------------------
                  Нет |         361        2.89       54.18
                   Да |       6,406       51.29       51.29
                    . |       5,723       45.82      100.00
----------------------+-----------------------------------
                Total |     12,490      100.00

.
. tabulate kj8_1 if k_adult == 1, missing

              Скажите, |
            пожалуйста |
                  , Вы |
              работали |
                    по |
              основной |
                работе |
                дома в |
               течение |      Freq.     Percent        Cum.
----------------------+-----------------------------------
                  Нет |       5,939       47.55       51.23
                   Да |         460        3.68        3.68
                    . |       6,091       48.77      100.00
----------------------+-----------------------------------
                Total |     12,490      100.00

.
```

```
. tabulate kj8_3 if k_adult == 1, missing

           Скажите, |
         пожалуйста |
               , Вы |
          учитывали |
          эти часы, |
              когда |
           называли |
            общую п |      Freq.     Percent        Cum.
--------------------+-----------------------------------
                Нет |        343        2.75        3.55
                 Да |        100        0.80        0.80
                  . |     12,047       96.45      100.00
--------------------+-----------------------------------
              Total |     12,490      100.00

.
. tabulate kj9 if k_adult == 1, missing

           Скажите, |
         пожалуйста |
               , по |
          основному |
              месту |
           работы в |
            течение |
           последни |      Freq.     Percent        Cum.
--------------------+-----------------------------------
                Нет |        556        4.45       54.14
                 Да |      6,206       49.69       49.69
                  . |      5,728       45.86      100.00
--------------------+-----------------------------------
              Total |     12,490      100.00

.
. tabulate kj14 if k_adult == 1, missing

                  В |
           настоящее |
         время Ваше |
         предприяти |
         е осталось |
         должно Вам |
         какие-то д |      Freq.     Percent        Cum.
--------------------+-----------------------------------
                Нет |      5,682       45.49       49.86
                 Да |        545        4.36        4.36
                  . |      6,263       50.14      100.00
--------------------+-----------------------------------
              Total |     12,490      100.00

.
. tabulate kj16 if k_adult == 1, missing

         За сколько |
            месяцев |
         предприяти |
               е не |
           доплатило |
             Вам эти |
             деньги? |      Freq.     Percent        Cum.
--------------------+-----------------------------------
                  0 |          3        0.02        0.02
                  1 |        307        2.46        2.48
                  2 |        105        0.84        3.32
                  3 |         39        0.31        3.63
                  4 |          9        0.07        3.71
                  5 |          7        0.06        3.76
                  6 |          7        0.06        3.82
                  7 |          2        0.02        3.84
                  8 |          2        0.02        3.85
```

```
                 10 |          2        0.02        3.87
                 12 |          7        0.06        3.92
                 20 |          1        0.01        3.93
                 21 |          1        0.01        3.94
                 24 |          2        0.02        3.96
                 30 |          2        0.02        3.97
                 33 |          1        0.01        3.98
                 36 |          2        0.02        4.00
                 48 |          1        0.01        4.00
                  . |     11,990       96.00      100.00
--------------------+-----------------------------------
              Total |     12,490      100.00

.
. tabulate kj17 if k_adult == 1, missing

                 Вы |
          получали в |
             течение |
           последних |
          30 дней на |
                этом |
          предприяти |
           и в качес |      Freq.     Percent        Cum.
--------------------+-----------------------------------
                Нет |      6,081       48.69       49.96
                 Да |        159        1.27        1.27
                  . |      6,250       50.04      100.00
--------------------+-----------------------------------
              Total |     12,490      100.00
```

7.4. И описательные характеристики количественных переменных

summarize kj8 kj8_2 kj10 kj15 kj18 if k_adult ==1

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         kj8 |      5,969    174.8745    56.01043          4        480
       kj8_2 |        406    27.83498    32.46468          1        260
        kj10 |      5,982    8285.601    7443.216         50     130000
        kj15 |        447    7910.725     12370.1         20     175000
        kj18 |        145    1308.276     1952.14         15      15000
```

7.5. Посмотрим на переменные рабочего времени более внимательно

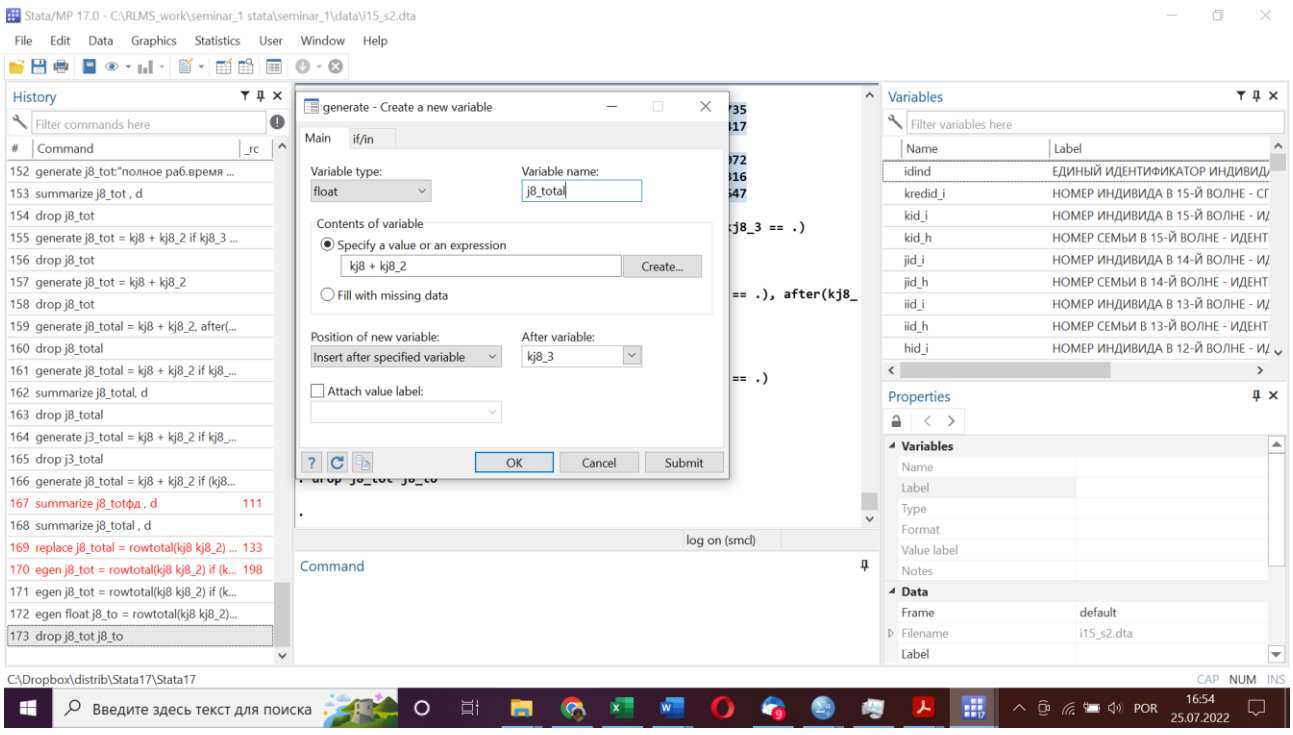
summarize kj8 kj8_2 , d

```
              Сколько часов Вы
       фактически отработали по
         основному месту работы
-------------------------------------------------------------
      Percentiles      Smallest
 1%           30              4
 5%           75              8
10%          108              8       Obs               5,969
25%          160              8       Sum of wgt.       5,969

50%          174                      Mean           174.8745
                        Largest       Std. dev.      56.01043
75%          192            420
90%          240            420       Variance       3137.169
95%          270            420       Skewness       .3250094
99%          360            480       Kurtosis       5.271761

              Сколько часов Вы
     фактически занимались этой
           работой дома в течени
-------------------------------------------------------------
```

```
      Percentiles      Smallest
 1%            1              1
 5%            2              1
10%            3              1       Obs                 406
25%            7              1       Sum of wgt.         406

50%           16                      Mean           27.83498
                     Largest          Std. dev.      32.46468
75%           40             180
90%           60             198      Variance       1053.955
95%           81             208      Skewness       2.803356
99%          160             260      Kurtosis       14.87326
```

Довольно таки сомнительно, что действительно есть люди, которые работают 480 часов за 30 дней – это 16-ти часовой рабочий день без выходных. Поэтому на совесть исследователя остается, оставить ли данные такими, или убрать слишком большие значения. Как правило, считается, что максимальная рабочая неделя – 100 часов, то есть за 30 дней это может быть около 400-420 часов. Но прежде чем убирать слишком большие значения, давайте сложим значения переменных **kj8 kj8_2** в том случае, если человек не учел работу дома, то есть если переменная **kj8_3** равна 0 или миссингу. Так как в переменной **kj8_2** всего 406 наблюдений, мы не можем использовать обычный знак «плюс» для суммирования, иначе если в одном из слагаемых миссинг, то и сумма примет пустое значение.
Давайте проверим (**after** позволяет вставить переменную в желаемое место):

generate j8_total = kj8 + kj8_2 if (kj8_3 == 0 | kj8_3 == .) , after(kj8_3)



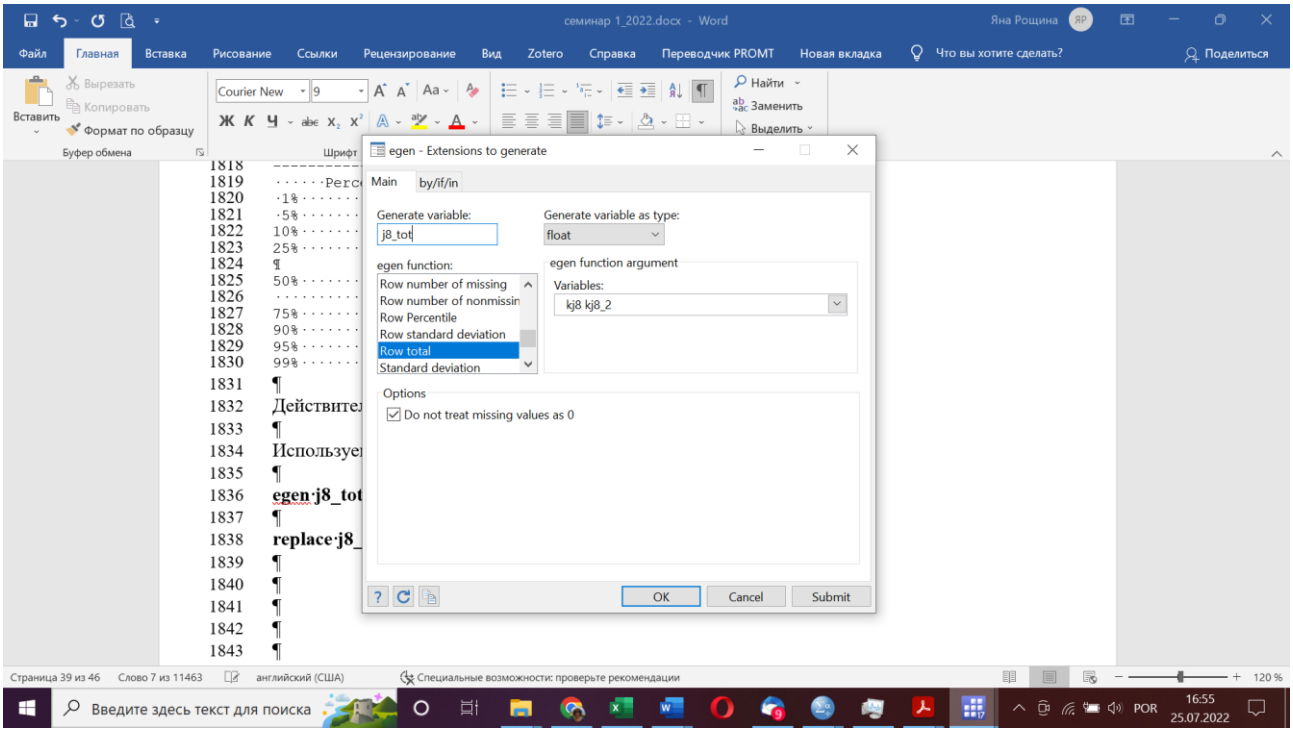
summarize j8_total , d

```
                          j8_total
-------------------------------------------------------------
      Percentiles      Smallest
 1%           50             42
 5%           76             50
10%          108             50       Obs                 302
25%          144             50       Sum of wgt.         302
```

50%	180		Mean	182.9735
		Largest	Std. dev.	62.27417
75%	220	360		
90%	260	365	Variance	3878.072
95%	290	374	Skewness	.4382316
99%	360	400	Kurtosis	3.732647

Действительно, у нас всего 302 кейса.

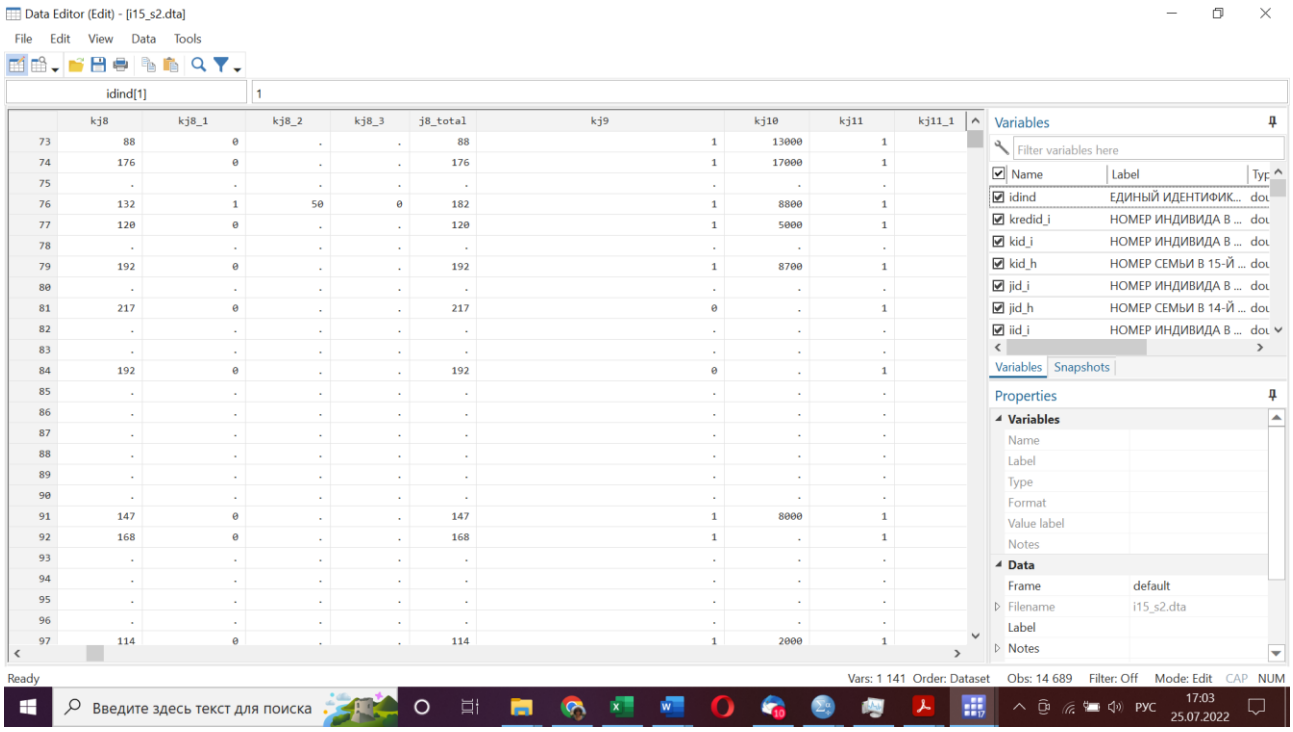
7.6. Используем теперь такой код, позволяющий суммировать «число» с «миссингом»:
egen j8_tot = rowtotal(kj8 kj8_2), missing
summarize j8_tot , d



	j8_tot			
	Percentiles	Smallest		
1%	24	1		
5%	72	1		
10%	110	2	Obs	5,992
25%	160	2	Sum of wgt.	5,992
50%	176		Mean	176.0893
		Largest	Std. dev.	57.45592
75%	192	420		
90%	240	440	Variance	3301.182
95%	280	480	Skewness	.3339783
99%	360	520	Kurtosis	5.450133

7.7. Не очень удобно, что оператором **egen** нельзя поместить переменную в нужное место; поэтому воспользуемся таким кодом и удостоверимся в таблице данных, что наша переменная посчитана верно:

replace j8_total = j8_tot



Удалим ненужную переменную и добавим лейбл новой переменной
drop j8_tot
label variable j8_total "полное рабочее время за 30 дней на первой работе"

Обратите внимание, что в описании этой переменной максимум – 520 часов, а мы говорили, что рабочее время за 30 дней вряд ли больше 400-420 часов. Перекодируем все, что больше чем 420, в 420

recode j8_total (420 / max = 420)

Сделано всего 3 изменения.
sum j8_total, d

```
       полное рабочее время за 30
            дней на первой работе
-------------------------------------------------------------
      Percentiles      Smallest
 1%           24              1
 5%           72              1
10%          110              2       Obs                5,992
25%          160              2       Sum of wgt.        5,992

50%          176                      Mean            176.0592
                        Largest       Std. dev.       57.30783
75%          192             420
90%          240             420       Variance        3284.188
95%          280             420       Skewness        .2994812
99%          360             420       Kurtosis        5.248013
```

Первое задание выполнено.

7.8. Теперь посчитаем «контрактную» заработную плату за 30 дней: деньги, которые человек получил за 30 дней, + сумму задолженности, деленную на кол-во месяцев задолженности; + сумму, которую можно выручить за проданную продукцию.

Так как среди количества месяцев, за которые предприятие должно деньги, в трех случаях принимает значение 0, а на ноль делить нельзя, не понятно, ошибка это или что-то еще. Но давайте объявим это значение миссингом в этой переменной (не удалим, вдруг мы надумаем с ним что-то еще делать):

recode kj16 (0 = .a)

Теперь у нас минимальное значение – 1.

7.9. Рассчитаем, сколько недоплатили за 1 месяц

generate kj15_16= kj15/kj16

Создадим новую переменную «контрактная ЗП на первом месте работы» в удобном для нас месте

generate Wage_1 = ., after(kj10)

7.10. Теперь рассчитаем нужное значение переменной **Wage_sup** при помощи команды

egen Wage_sup= rowtotal(kj10 kj15_16 kj18), missing

И заменим значение **Wage_1** на **Wage_sup** , удалив вспомогательные переменные

replace Wage_1 = Wage_sup
label variable Wage_1 "контрактная ЗП за 30 дней на первой работе"
drop kj15_16 Wage_sup

Посмотрим характеристики **Wage_1**

summarize Wage_1 , d

```
                          Wage_1
-------------------------------------------------------------
      Percentiles      Smallest
 1%          700       6.666667
 5%         1500             50
10%         2149             50       Obs               6,141
25%         3900             50       Sum of wgt.       6,141

50%         6400                      Mean           8436.435
                        Largest       Std. dev.      7620.681
75%        10000          88000
90%        16700         100000       Variance       5.81e+07
95%        21500         100000       Skewness       3.703423
99%        35000         130000       Kurtosis       32.22927
```

7.11. И рассчитаем среднюю заработную плату для разных категорий

tabstat Wage_1 [aw = k_inwgt], stat (mean med N) by (k_occup08)

```
Summary for variables: Wage_1
Group variable: k_occup08 (ПРОФЕССИОНАЛЬНАЯ ГРУППА - 15 ВОЛНА - по коду
> kj2cod08)

       k_occup08 |      Mean       p50         N
-----------------+--------------------------------
  Военнослужащие |  10579.79     10000        29
Законодатели; кр |  15839.32     10000       244
Специалисты высш |  9120.746      7000       742
Специалисты сред |  9067.157      7000       716
Служащие офисные |  6928.401      6000       229
Работники сферы  |  6004.788      5000       713
Квалифицированны |  5324.835      3000        23
Квалифицированны |  9069.105      7400       632
Квалифицированны |  8778.908      7000       650
Неквалифицирован |  4949.825      3600       339
-----------------+--------------------------------
           Total |  8489.355      6500      4317
------------------------------------------------
```

Обратите внимание, что у нас нет ни рабочего времени, ни заработка, равных 0.

7.12. Мы можем рассчитать ставку ЗП, разделив заработок на время, а также логарифм ставки ЗП (понадобится нам для модели Минцера).

gen Hwage1=Wage_1/ j8_total
label variable Hwage1 "ставка ЗП за 30 дней на первой работе"
sum Hwage1, d

```
                          Hwage1
-------------------------------------------------------------
      Percentiles      Smallest
 1%     5.555555       .0236407
 5%     10.41667       .3571429
10%     14.28571       .4444444       Obs               5,565
25%     22.72727       .8333333       Sum of wgt.       5,565

50%         37.5                      Mean           65.58148
                       Largest       Std. dev.      410.6402
75%     60.83333          4500
90%     101.1905          8500       Variance       168625.4
95%        137.5         17000       Skewness       42.29489
99%     326.0869          22000       Kurtosis       2021.906
```

7.13. Посмотрим также ставку ЗП по группам профессий.

tabstat Hwage1 [aw = k_inwgt], stat (mean med N) by (k_occup08)

```
Summary for variables: Hwage1
Group variable: k_occup08 (ПРОФЕССИОНАЛЬНАЯ ГРУППА - 15 ВОЛНА - по коду
> kj2cod08)

       k_occup08 |      Mean       p50         N
-----------------+--------------------------------
  Военнослужащие |  61.95456  52.63158        25
Законодатели; кр |  95.09192  52.08333       218
Специалисты высш |  87.66237  46.91667       678
Специалисты сред |  71.81435  41.42857       654
Служащие офисные |  43.99543  33.33333       217
Работники сферы  |  38.18263  25.46296       646
Квалифицированны |  33.24069  23.86364        18
```

```
Квалифицированны |   54.7199     43.75       581
Квалифицированны |  83.83484  38.88889       584
Неквалифицирован |  38.20211        25       302
-----------------+---------------------------
           Total |  65.24629  38.04348      3923
-------------------------------------------------
```

7.14. Расчет логарифма ставки заработной платы.

gen lg_Hwage1= ln(Hwage1)
label variable lg_Hwage1 "логарифм ставки ЗП за 30 дней на первой работе"

7.15. Создание переменной с группами возраста.

На основе переменной «возраст» **k_age** создадим порядковую переменную **age_group,** разбив на следующие интервалы: 0-2 года, 3-5лет, 6-15 лет, 16-17 лет, 18-64 года, 65 и более лет.

generate age_group = k_age, after(k_age)
recode age_group (0/2 = 1) (3/5 =2) (6/15=3) (16/17=4) (18/64=5) (65/max=6)
label variable age_group "группы возраста"
label define age_group 1 "0 - 2" 2 "3-5" 3 "6-15" 4 "16-17" 5 "18- 64" 6 ">=65", replace
label values age_group age_group

tab age_group

```
      групп |
          ы |
      возра |
        ста |      Freq.     Percent        Cum.
------------+-----------------------------------
      0 - 2 |         487        3.32        3.32
        3-5 |         483        3.29        6.60
       6-15 |       1,637       11.14       17.75
      16-17 |         442        3.01       20.76
     18- 64 |       9,544       64.97       85.73
       >=65 |       2,096       14.27      100.00
------------+-----------------------------------
      Total |      14,689      100.00
```

!!! Сохранить do – файл. Для этого удалить «плохие» команды, выделить правильные команды, и выполнить команду «перенести в do – файл», сохранить этот файл с именем «seminar1_фамилия».
!!! Сохранить ваш рабочий файл (можно с новым именем!!!)

7.16. Сохранение файла с нужными переменными

Сохраним файл под новым именем, оставив только нужные переменные. В переменной **kj60** заменим «99999996-99999999» на миссинг (.). Переменную **kj1** (есть ли работа) перекодируем в новую переменную (**employed**), при этом первые четыре категории будем считать соответствующими тому, что у человека есть работа.

save "C:\RLMS_work\seminar_1\data\i15_short.dta"
keep idind kid_h k_origsm k_inwgt psu region status popul k_adult k_child site ssu kh3 kh4 k_marst k_occup08 k_occup08_k kj72_172a kj72_173a kj1 kj7 kj8 kj8_1 kj8_2 kj8_3 kj9 kj10 kj14 kj15 kj16 kj17 kj18 j8_total Wage_1 Hwage1 lg_Hwage1 kh5 kh6 k_age age_group kj1 kj60
recode kj60 (99999997 / 99999999 = .)

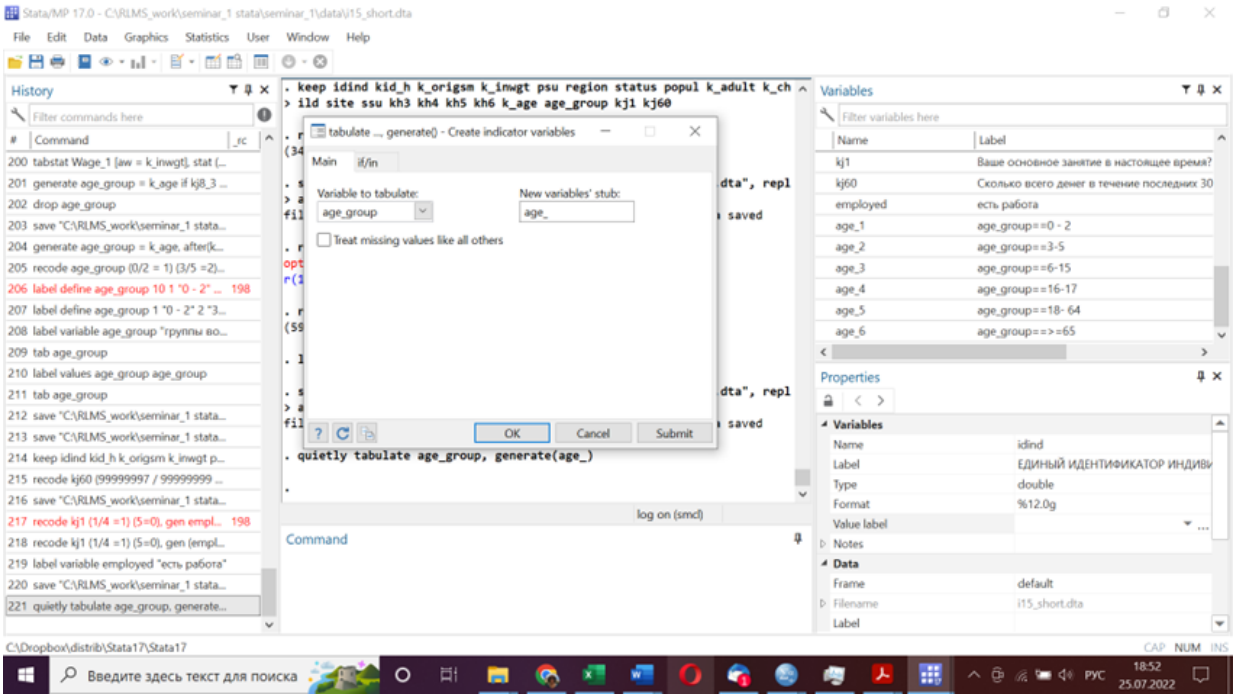
recode kj1 (1/4 =1) (5=0), gen (employed)
label variable employed "есть работа"

7.17. создадим дамми возрастных групп

quietly tabulate age_group, generate(age_)

Сохраните файл!!!!

save "C:\RLMS_work\seminar_1\data\i15_s1a.dta", replace



8. Агрегирование индивидуальных данных до уровня домохозяйства

Так как в данных РМЭЗ ВШЭ есть и файл по домохозяйствам, и файлы по каждому члену семьи, мы можем использовать для каждого индивида какие-то домохозяйственные характеристики (например, душевой доход), и наоборот, строить характеристики домохозяйства на основании свойств каждого члена семьи. Например, на основании построенных нами переменных, мы можем посчитать количество членов семьи каждой возрастной группы; количество занятых (или их долю); сумму доходов всех членов семьи. Мы используем индикатор домохозяйства **kid_h**, чтобы идентифицировать людей из одной и той же семьи.

Для этого используем такой синтаксис:

collapse (mean) employed (sum) age_1 age_2 age_3 age_4 age_5 age_6 kj60, by(kid_h)

При этом сразу получается новый файл, где кейсом является домохозяйство. Отсортируем его по идентификатору домохозяйства и сохраним его:

sort kid_h
save "C:\RLMS_work\seminar_1\data\h15_aggr.dta", replace

summarize kid_h

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       kid_h |      5,540     9587730     6221353     101001   2.41e+0
```

В этом файле 5540 кейсов (домохозяйств).

9. Файл данных домохозяйства и приклеивание к нему агрегированных данных по индивидам.

9.1. Откроем файл с данными домохозяйств 15 волны:

use "C:\RLMS_work\seminar_1\data\r15hall41.dta", clear

describe kredid_h kid_h jid_h iid_h hid_h gid_h fid_h eid_h did_h cid_h bid_h aid_h k_origsm k_hhwgt region psu ssu status popul site ka3 ka3_3 ka4_1 ka4_2 ka5_1 ka5_2 ka8 k_nfm , short

Посмотрим на основные переменные, большинство из которых нужно сохранять в любом файле, если вы собираетесь оставить не все переменные. Но самая важная из них – **kid_h**, так как с ее помощью вы всегда сможете приклеить недостающие переменные.
Большинство из этих переменных вы уже видели в индивидуальном файле. Правда, переменная взвешивания здесь другая – для домохозяйств, а не индивидов.
Обратите внимание также на очень важную переменную «количество человек в семье», это сконструированная переменная, учитывающая также тех людей, которые не заполнили анкету или временно отсутствуют.

```
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------
kredid_h        double  %12.0g                Идентификационная переменная 15-го
                                                раунда
kid_h           double  %12.0g                Идентификационная переменная 15-го
                                                раунда
jid_h           double  %12.0g                Идентификационная переменная 14-го
                                                раунда
iid_h           double  %12.0g                Идентификационная переменная 13-го
                                                раунда
hid_h           double  %12.0g                Идентификационная переменная 12-го
                                                раунда
gid_h           double  %12.0g                Идентификационная переменная 11-го
                                                раунда
fid_h           double  %12.0g                Идентификационная переменная 10-го
                                                раунда
eid_h           double  %12.0g                Идентификационная переменная 9-го
                                                раунда
did_h           double  %12.0g                Идентификационная переменная 8-го
                                                раунда
cid_h           double  %12.0g                Идентификационная переменная 7-го
                                                раунда
bid_h           double  %12.0g                Идентификационная переменная 6-го
                                                раунда
aid_h           double  %12.0g                Идентификационная переменная 5-го
                                                раунда
k_origsm        double  %12.0g     k_origsm   Адрес репрезентативной выборки? - В
                                                15-й волне
k_hhwgt         double  %12.0g                Постстратификационный вес
                                                домохозяйства в 15-й волне
region          double  %12.0g     region     Код региона
psu             double  %12.0g     psu        Первичная единица отбора
ssu             double  %12.0g                Вторичная единица отбора
status          double  %12.0g     status     Тип населенного пункта
popul           double  %12.0g                Численность населения
```

```
site            double  %12.0g                Номер населенного пункта
ka3             double  %12.0g                Номер семьи
ka3_3           double  %12.0g     ka3_3      Хотя бы один член семьи раньше
                                                участвовал в исследовании?
ka4_1           double  %12.0g     ka4_1      Дата проведения интервью: число
ka4_2           double  %12.0g     ka4_2      Дата проведения интервью: месяц
ka5_1           double  %12.0g     ka5_1      Интервью продолжалось: часов
ka5_2           double  %12.0g     ka5_2      Интервью продолжалось: минут
ka8             double  %12.0g     ka8        Номер члена семьи, который отвечал на
                                                вопросы семейного вопросник
k_nfm           double  %12.0g                Количество членов семьи в 15 волне
```

9.2. Сохраним файл с небольшим количеством переменных

save "C:\RLMS_work\seminar_1\data\h15_short.dta"
keep kid_h k_origsm k_hhwgt ka3 k_nfm kf14
sort kid_h

9.3. Перекодируем миссинги в переменной суммарного дохода домохозяйства **kf14**

recode kf14 (99999997/99999999 = .)

9.4. Рассчитаем душевой доход делением на количество человек в семье

gen INCOME_PC = kf14 / k_nfm if k_nfm > 0
label variable INCOME_PC "Душевой доход"

Посмотрим описательные характеристики

sum k_nfm kf14 INCOME_PC

```
    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
       k_nfm |      5,545    2.752209    1.410864          1         12
        kf14 |      5,258    16338.85    34475.89          0     944200
   INCOME_PC |      5,258    6110.313    9647.261          0   314733.3
```

Вы видите, что в файле домохозяйств на 5 кейсов больше, чем в агрегированном файле (то есть есть д\х, в которых не опрошен ни один человек, но заполнена семейная анкета).

9.5. Количество членов семьи варьируется от 1 до 12; сделаем новую переменную **nfm_kod** (5 человек и больше – в одну градацию)

recode k_nfm (1 =1) (2=2) (3=3) (4=4) (5/max=5), gen (nfm_kod)
label define nfm_kod 1 "1" 2 "2" 3 "3" 4 "4" 5 ">=5"
label values nfm_kod nfm_kod

9.6. Посмотрим среднее значение душевого дохода по семьям с разным количеством человек, используя команду взвешивания для репрезентативности (можно было также делать отбор по условию **k_origsm == 1**)

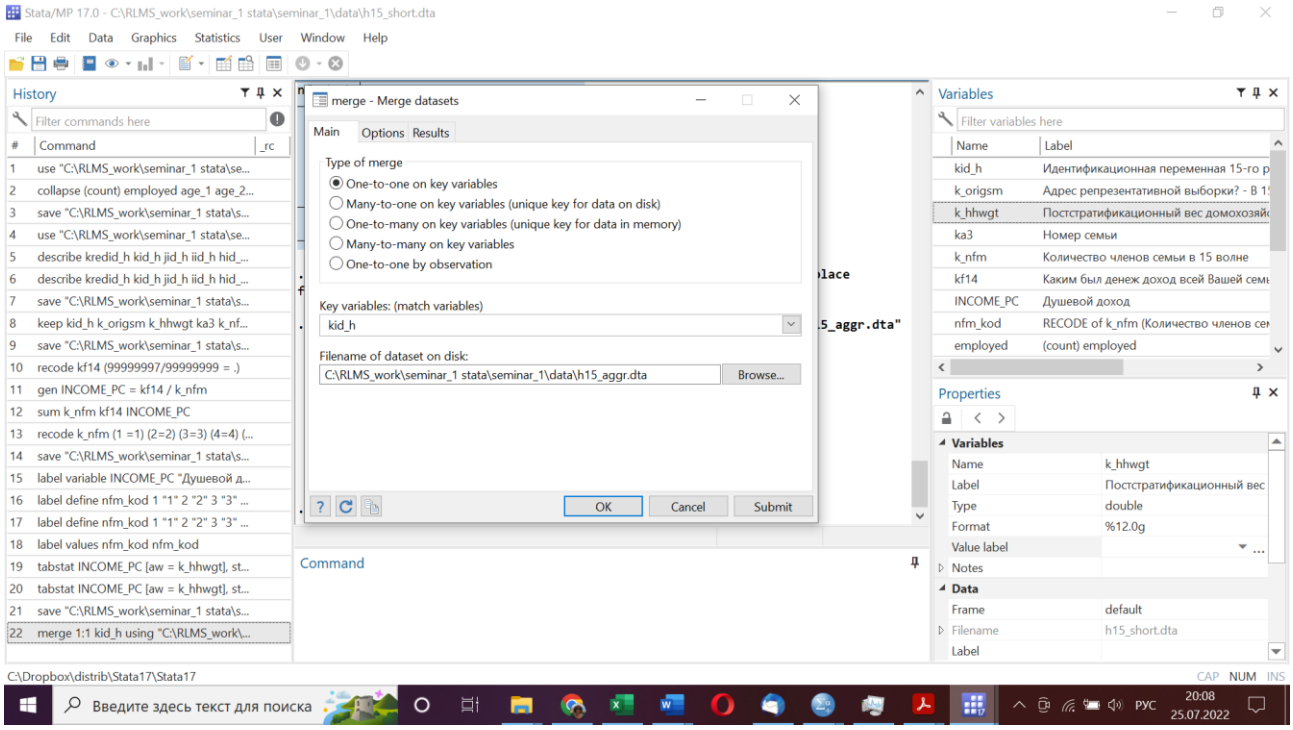
tabstat INCOME_PC [aw = k_hhwgt], stat (mean med N) by (nfm_kod)

```
Summary for variables: INCOME_PC
Group variable: nfm_kod (RECODE of k_nfm (Количество членов семьи в 15 волне))

nfm_kod |      Mean       p50         N
--------+------------------------------
      1 |  6228.863    4033.5       846
```

```
      2 |   5920.937       4500       1163
      3 |   5927.071   4666.667        908
      4 |   5962.415       4125        606
    >=5 |   5211.431       3600        380
--------+----------------------------------
  Total |   5929.812   4266.667       3903
------------------------------------------
```

Вы видите, что душевой доход выше в семьях с одним человеком, и ниже – где 5 и более человек.

9.7. Приклеим теперь данные из агрегированного файла.



Используем опцию «один-к-одному» так как у нас (по идее) должен быть один и тот же список домохозяйств. Используем «ключ» для склейки, чтобы к приклеивались данные того же самого д\х (**kid_h,** как «фамилия»).

merge 1:1 kid_h using "C:\RLMS_work\seminar_1\data\h15_aggr.dta"

```
    Result                      Number of obs
    -----------------------------------------
    Not matched                             5
        from master                         5  (_merge==1)
        from using                          0  (_merge==2)

    Matched                             5,540  (_merge==3)
    -----------------------------------------
```

Для пяти д\х не нашлось походящей строки в агрегированном файле, как мы уже говорили.

9.8. Посмотрим характеристики приклеенных переменных

summarize employed age_1 age_2 age_3 age_4 age_6 kj60

```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
```

```
    employed |       5,537    .5294218    .3962203           0           1
       age_1 |       5,540    .0879061    .2944359           0           3
       age_2 |       5,540    .0871841    .2915708           0           3
       age_3 |       5,540    .2954874    .5778119           0           5
       age_4 |       5,540    .0797834    .2795089           0           2
-------------+--------------------------------------------------------
       age_6 |       5,540    .3783394    .6173458           0           3
        kj60 |       5,540    13289.07    13487.62           0      305000
```

Переменные **age_1 age_2 age_3 age_4 age_5 age_6** – это количество членов семьи определенного возраста в домохозяйстве (хотя в подсчетах может быть ошибка, если у кого-то из членов семьи не заполнена анкета).
??? Что означают «средние» для этих переменных в аутпуте команды **summarize?**

9.9. Суммируем количество членов семьи в каждой возрастной группе

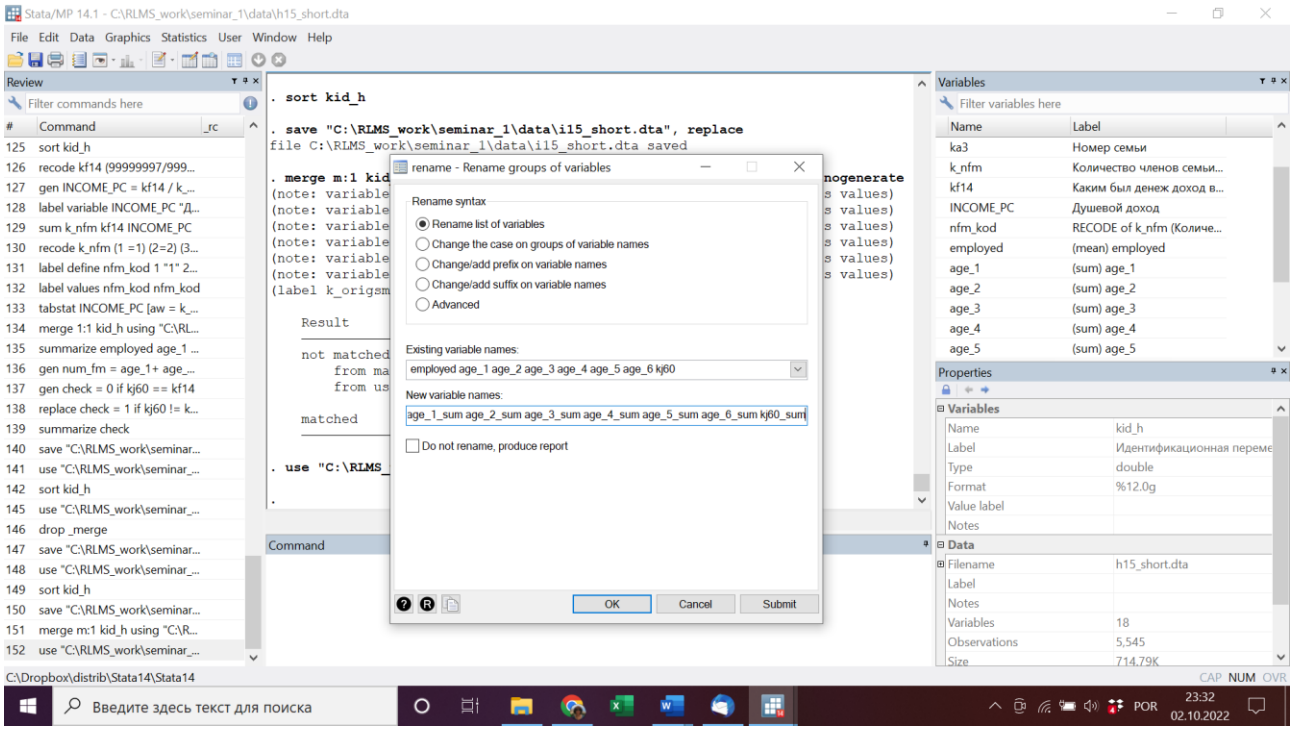
gen num_fm = age_1+ age_2+ age_3+ age_4+ age_5+ age_6

удалим переменную **_merge**

drop _merge

9.10. Переименуем переменные (чтобы отличать от переменных индивидуального файла)

rename (employed age_1 age_2 age_3 age_4 age_5 age_6 kj60) (employed_mean age_1_sum age_2_sum age_3_sum age_4_sum age_5_sum age_6_sum kj60_sum)



<u>9.11.</u> <u>Самостоятельное задание.</u>
В файле аупута наберите команду:
***Фамилия – номер семинара – номер задания**

Обязательно посмотрите на данные в таблице (браузер данных).
Перейдите в редактор данных и сравните переменные **num_fm** и **k_nfm** (разница может быть, если у кого-то из членов семьи не заполнена анкета).

Сравните также сумму индивидуальных доходов **kj60** и общие доходы семьи по ответу одного из членов домохозяйства **kf14** (наблюдается разница для некоторых кейсов).
Для проверки, создайте вспомогательную переменную **check**, которая равна 1, если переменные **kj60** и **kf14** не равны, и 0 в противоположном случае. Посмотрите, в какой доле д\х переменные не равны.
- Скопируйте результаты и команды из окна аутпута, и вставьте их в текстовый файл с вашими ответами на задания этого семинара. Один текстовый файл для всех выполненных заданий. Назовите файл вашей ФИО и номер группы, укажите номер семинара.

Сохраните файл данных!!!
save "C:\RLMS_work\seminar_1\data\h15_short.dta", replace

10. Приклеивание к индивидуальным данным характеристик семьи.

10.1. Откроем снова индивидуальный файл

use "C:\RLMS_work\seminar_1\data\i15_short_s1.dta"

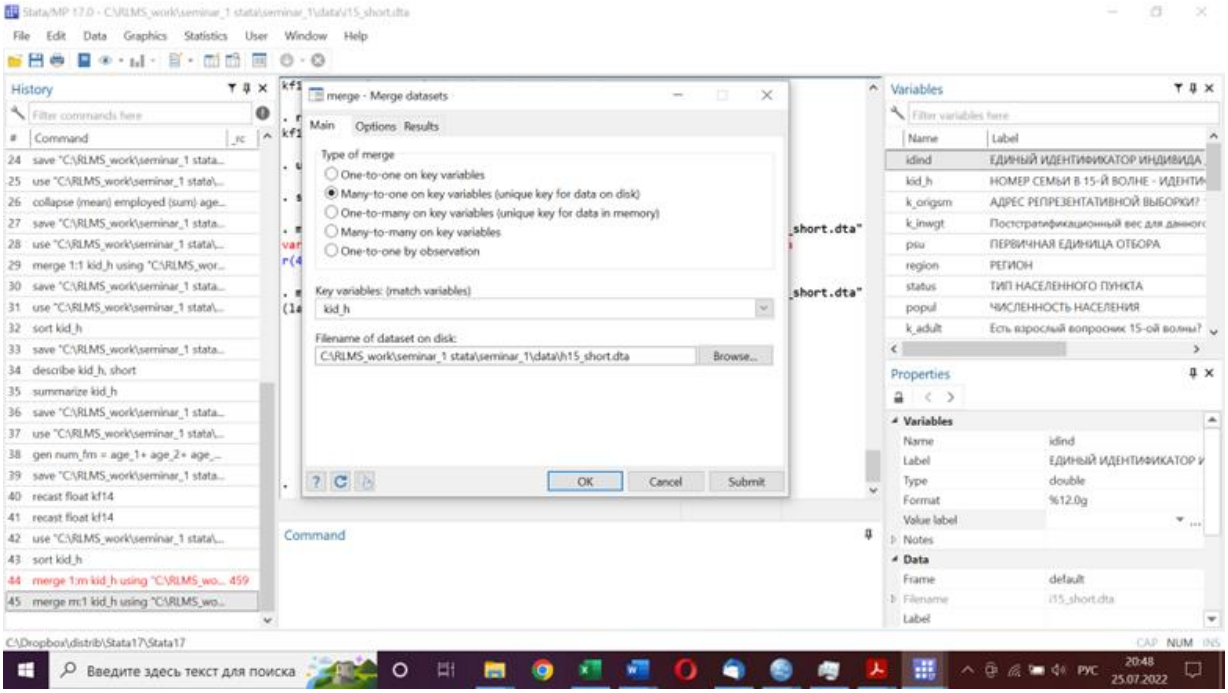
Отсортируем его по переменной идентификатор домохозяйства

sort kid_h

Теперь по этому ключу приклеим данные из домохозяйственного файла («многие к одному», так как характеристики одной и той же семьи приклеиваются ко всем индивидам из нее)
Если использовать опцию выбора переменных, которые мы хотим приклеить (список можно вставить в соответствующее окно), то получится так:

merge m:1 kid_h using "C:\RLMS_work\seminar_1\data\h15_short.dta", keepusing(k_nfm k f14 INCOME_PC nfm_kod) nogenerate

```
    Result                           # of obs.
    -----------------------------------------
    not matched                             5
        from master                         0
        from using                          5

    matched                            14,689
```



Сохраните файл
save "C:\RLMS_work\seminar_1\data\i15_short_s1.dta", replace

Сохраните do-файл и файл результатов, закройте STATA. Сохраните на флешку или перешлите себе полученные файлы данных для использования для домашнего задания.

10.2. Самостоятельное домашнее задание.
В файле аупута наберите команду:
***Фамилия – номер семинара – номер задания**

- Посмотрите характеристики приклеенных переменных при помощи команды **summarize**
- Скопируйте результаты и команды из окна аутпута, и вставьте их в текстовый файл с вашими ответами на задания этого семинара. Один текстовый файл для всех выполненных заданий. Назовите файл вашей ФИО и номер группы, укажите номер семинара.

10.3. Самостоятельное домашнее задание.
В файле аупута наберите команду:
***Фамилия – номер семинара – номер задания**

- На основе индивидуального файла создайте агрегированную переменную «доля мужчин в домохозяйстве», для этого сначала создайте переменную **male** «мужской пол» со значениями 1=мужчина, 0=женщина, на основе переменной **kh5** (со значениями 1=мужчина, 2=женщина) и приклейте ее к короткому файлу домохозяйства. Посмотрите распределение этой переменной в файле домохозяйства. Переименуйте ее в **male_sum.** Сохраните файл данных. -
- Сохраните файл кода.
- Скопируйте результаты и команды из окна аутпута, и вставьте их в текстовый файл с вашими ответами на задания этого семинара. Один текстовый файл для всех выполненных заданий. Назовите файл вашей ФИО и номер группы, укажите номер семинара.