

Гаврилина Александра,
магистратура «Суперкомпьютерное моделирование в науке и инженерии», группа
МСКМ191

aagavrilina@edu.hse.ru

gavrilinasanya@gmail.com

Анализ данных в Stata

Ответьте, пожалуйста, на вопросы:

1. Откройте файл salary.dta.

- a. Какова средняя заработная плата и 95%-й доверительный интервал для средней заработной платы в группе мужчин-иностранцев с заработной платой более 40 000 руб.**

`mean salary if sex==1 & salary>40000 & foreigner==1`

Средняя зарплата (mean): 132000

95-й доверительный интервал (95% Conf. Interval): (44464.89, 219535.1)

- b. Осуществите перекодирование переменной position из текстовой в численную. Задайте метки для новой численной переменной.**

`generate position_int = .`

`replace position_int = 0 if position=="lecturer"`

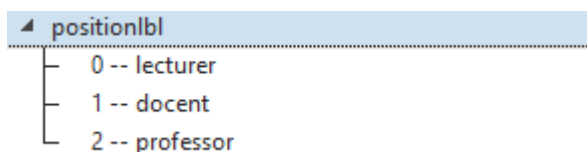
`replace position_int = 1 if position=="docent"`

`replace position_int = 2 if position=="professor"`

`label define positionlbl 0 "lecturer" 1 "docent" 2 "professor"`

`label values position_int positionlbl`

Метки:



positionlbl
0 -- lecturer
1 -- docent
2 -- professor

- c. Создайте новую переменную, которая будет отображать ранг сотрудника, посчитанный на основе значения его/её заработной платы. Ранг 1 присваивается сотруднику с самой высокой заработной платой.**

`egen salary_rank = rank(salary), field`

Фрагмент таблицы после выполнения команды:

	position	salary	foreigner	sex	position_int	salary_rank	
1	professor	250000	no	male	professor	1	
2	professor	200000	no	male	professor	2	
3	professor	200000	yes	male	professor	2	
4	professor	200000	no	male	professor	2	
5	professor	200000	yes	male	professor	2	
6	professor	180000	no	female	professor	6	
7	professor	140000	no	male	professor	7	
8	professor	140000	yes	male	professor	7	
9	professor	140000	no	female	professor	7	
10	professor	130000	no	male	professor	10	

- d. Создайте новую переменную, отображающую сумму уплачиваемого сотрудником подоходного налога в долларах (по сегодняшнему курсу ЦБ). Учитывайте, что подоходный налог для российских граждан – 13%, а для иностранцев – 30%. Укажите общую ежемесячную сумму уплачиваемого налога на всех сотрудников.

generate fee = .

replace fee = salary*0.13/64.09 if foreigner == 0

replace fee = salary*0.3/64.09 if foreigner == 1

total fee

Общая ежемесячная сумма налога в долларах: 11873.77

- e. Отсортируйте наблюдения по возрастанию значения переменной salary.

gsort salary

Фрагмент таблицы после выполнения команды:

	position	salary	foreigner	sex	position_int	salary_rank	fee
1	lecturer	20000	no	female	lecturer	50	40.56795
2	lecturer	28000	no	female	lecturer	47	56.79513
3	lecturer	28000	yes	female	lecturer	47	131.0657
4	docent	28000	yes	female	docent	47	131.0657
5	lecturer	30000	no	male	lecturer	44	60.85193
6	docent	30000	no	male	docent	44	60.85193
7	docent	30000	yes	male	docent	44	140.4275
8	docent	32000	yes	female	docent	43	149.7894
9	lecturer	34000	no	male	lecturer	41	68.96552
10	lecturer	34000	no	female	lecturer	41	68.96552
11	lecturer	40000	yes	female	lecturer	35	187.2367
12	lecturer	40000	yes	female	lecturer	35	187.2367
13	lecturer	40000	no	female	lecturer	35	81.1359
14	docent	40000	no	male	docent	35	81.1359
15	docent	40000	no	male	docent	35	81.1359
16	lecturer	40000	yes	male	lecturer	35	187.2367

f. Укажите среднюю заработную плату для профессора, доцента и преподавателя.

tabstat salary, statistics(mean) by(position)

Средняя для профессора (professor) 135700

Средняя для доцента (docent): 56777.78

Средняя для преподавателя (lecturer): 36666.67

g. Создайте новую переменную salary_cat на основе перекодирования переменной salary. Переменная salary должна остаться в базе данных. Новая переменная принимает следующие значения:

1 – заработная плата превышает 100 000 руб. в мес., метка «высокая»

2 – заработная плата от 40 000 до 100 000 руб. в мес., метка «средняя»

3 – заработная плата менее 40 000 руб. в мес., метка «низкая»

Установите метки значений переменной salary_cat. Проведите частотный анализ этой переменной и укажите число сотрудников, попавших в каждую группу.

recode salary (min/39999=1) (40000/100000=2) (100001/max=3), gen(salary_cat)

label define salarylbl 1 "High salary level" 2 "Middle salary level" 3 "Low salary level"

label values salary_cat salarylbl

tabstat salary_cat, statistics(count) by (salary_cat)

Число сотрудников:

Высокая зарплата: 10

Средняя зарплата: 28

Низкая зарплата: 12

2. Откройте файл revenue.dta.

Измените базу данных так, чтобы в ней остались три переменные:

- year;

- company;

- новая переменная, отображающая среднюю выручку каждой из компании за каждый год (без детализации по месяцам).

collapse (mean) rev, by (year company)

До

	year	month	company	rev	
1	1998	1	1	3232	
2	1998	2	1	3254	
3	1998	3	1	4632	
4	1998	4	1	8362	
5	1998	5	1	6539	
6	1998	6	1	3629	
7	1998	7	1	3828	
8	1998	8	1	3942	
9	1998	9	1	4842	
10	1998	10	1	8452	
11	1998	11	1	3629	
12	1998	12	1	3828	
13	1999	1	1	3942	
14	1999	2	1	4842	
15	1999	3	1	3828	
16	1999	4	1	3942	
17	1999	5	1	4842	
18	1999	6	1	8452	
19	1999	7	1	3629	
20	1999	8	1	3629	
21	1999	9	1	3828	
22	1999	10	1	3942	
23	1999	11	1	4842	
24	1999	12	1	8452	
25	2000	1	1	3629	
26	2000	2	1	3828	
27	2000	3	1	3828	
28	2000	4	1	3942	
29	2000	5	1	4842	
30	2000	6	1	3828	
31	2000	7	1	3942	

После

	year	company	rev	
1	1998	1	4847.417	
2	1998	2	4117.167	
3	1998	3	4281.667	
4	1999	1	4847.5	
5	1999	2	4851.417	
6	1999	3	4723.167	
7	2000	1	4351.583	
8	2000	2	4495.25	
9	2000	3	4632.083	
10	2001	1	4720.333	
11	2001	2	3982.25	
12	2001	3	6703.5	
13	2002	1	4736.917	
14	2002	2	4417.833	
15	2002	3	6897.75	
16	2003	1	4335	
17	2003	2	4415	
18	2003	3	6703.5	
19	2004	1	4626.333	
20	2004	2	5134.5	
21	2004	3	5252.417	
22	2005	1	3933.083	
23	2005	2	5084.917	
24	2005	3	4721.167	
25	2006	1	4962.545	
26	2006	2	4229.25	
27	2006	3	4707.75	
28	2006	.	3629	