Работа с данными выборочных социологических обследований в пакете SPSS

Ермолина Анна

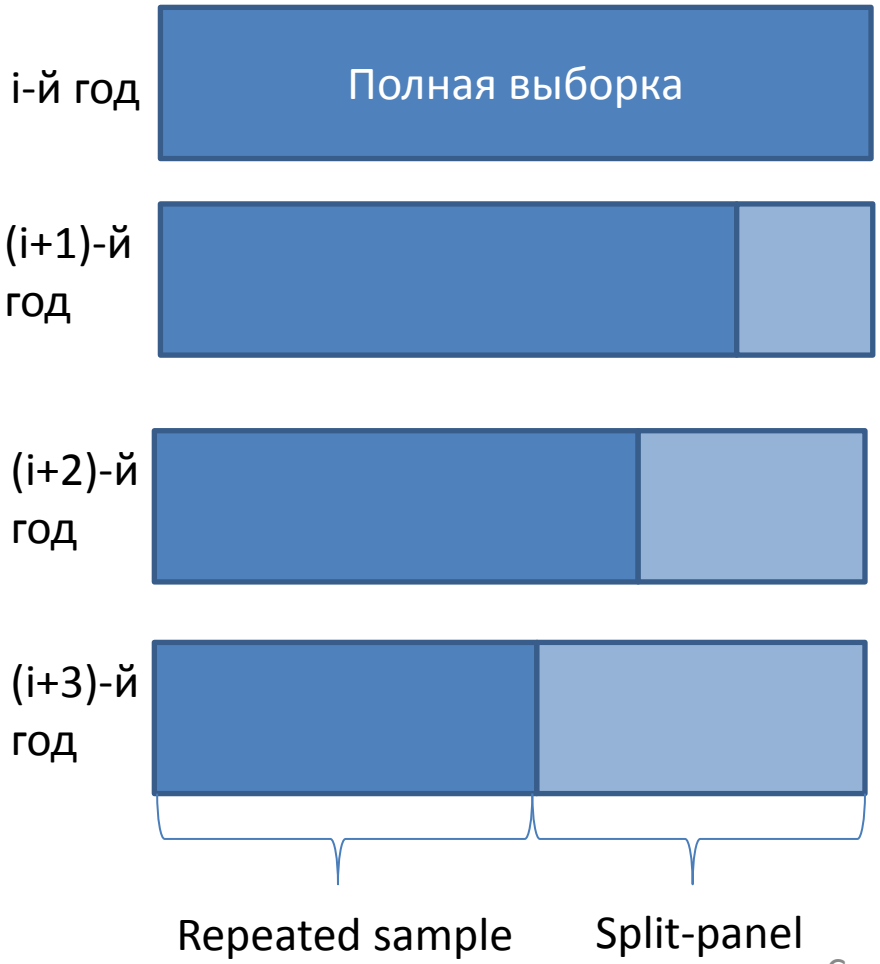
Научно-учебная группа

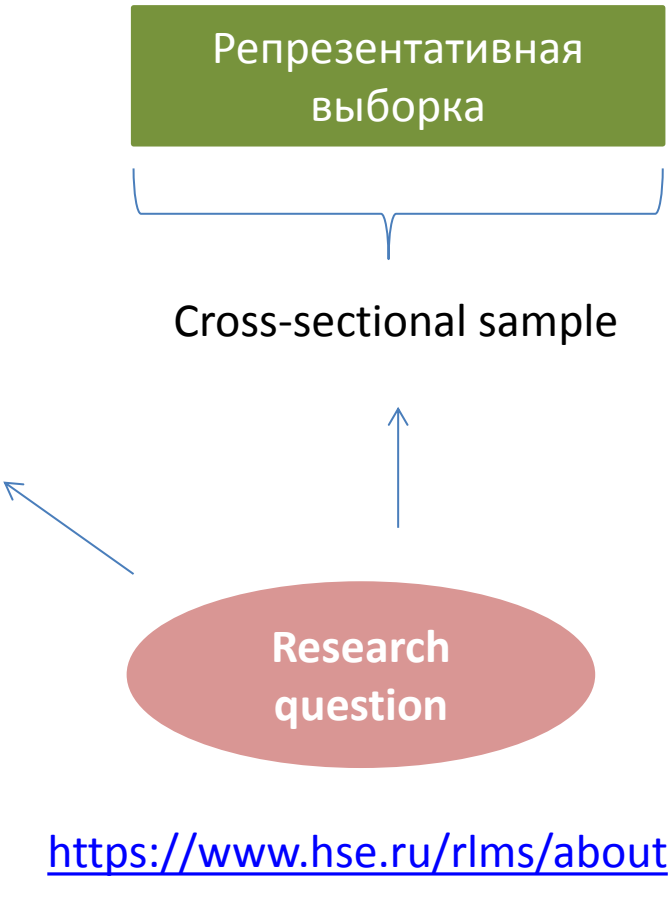
«”Дружественная” семье социальная политика, женская занятость и уровень жизни семей с детьми»

Семинар №3

11 марта 2016 г.

«Российский мониторинг экономического положения и здоровья населения НИУ ВШЭ» (RLMS-HSE)





https://www.hse.ru/rlms/about

Идентификационные переменные

- **aid_h, bid_h, cid_h, …, sid_h** – идентификационные номера ДХ
- **aid_i, bid_i, cid_i, …, sid_i** - идентификационные номера индивидов
- **idind** – уникальный идентификатор индивида
- **s_origsm = 1** для репрезентативной выборки
- **s_inwgt** – постстратификационный вес индивида

- Сортировка наблюдений

```
SORT CASES BY sh6 (A).
SORT CASES BY sh6 (D).
```

- Отбор наблюдений без удаления

```
USE ALL.
COMPUTE filter_$=(sh5=1)
VARIABLE LABELS filter_$ ' sh5=1 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
```

- Отбор наблюдений с удалением

```
FILTER OFF.
USE ALL.
SELECT IF (sh5=1).
EXECUTE.
```

- Снятие всех фильтров

```
FILTER OFF.
USE ALL.
EXECUTE.
```

Вычисление переменных

```
COMPUTE age1=s_int_y-sh6.
EXECUTE.

IF  (age < 16) SocioDemType=1.
EXECUTE.
IF  ((age >= 16)&((age<=54)&(sh5=2) |
(age<=59)&(sh5=1))) SocioDemType=2.
EXECUTE.
IF ((age >=55)&(sh5=2) | (age>=60)&(sh5=1))
SocioDemType=3.
EXECUTE.
```

Частотные таблицы

```
FREQUENCIES VARIABLES=age
  /NTILES=4
  /STATISTICS=MEAN MEDIAN
  /ORDER=ANALYSIS.
```

/FORMAT=NOTABLE – отключить вывод частот

/NTILES=5 – квинтили

/NTILES=10 – децили

/PERCENTILES=3.0 – процентили

SEMEAN – стандартная ошибка среднего

MINIMUM, MAXIMUM – минимум, максимум

MODE - мода

SUM - сумма

STDDEV – среднеквадратическое отклонение

VARIANCE – дисперсия

RANGE - размах

SKEWNESS, SESKEW – коэффициент асимметрии и его стандартная ошибка

KURTOSIS, SEKURT – коэффициент эксцесса и его стандартная ошибка

Дескриптивные статистики

```
DESCRIPTIVES VARIABLES=age
  /STATISTICS=MEAN STDDEV VARIANCE.
```

Задание №1

Сравнить средний и медианный возраст мужчин и женщин

Гистограмма распределения

```
COMPUTE filter_$=(sj10<99999997).
VARIABLE LABELS filter_$ 'sj10<99999997 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

GRAPH
  /HISTOGRAM(NORMAL)=sj10.
```

Таблицы сопряженности

- Описание связи двух или более номинальных (категориальных) переменных
- Непрерывные переменные можно разбить на интервалы
- Критерий независимости χ-квадрат

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

CROSSTABS

/TABLES=sh5 BY status.

Дополнительно:

/TABLES=sh5 BY status [BY …]

/FORMAT=AVALUE TABLES

/FORMAT=DVALUE NOTABLES – отключить вывод таблицы

/STATISTICS=CHISQ – критерий χ-квадрат

/CELLS=COUNT – наблюдаемое количество

ROW – процент по строке

COLUMN – процент по столбцу

TOTAL – процент общего итога

EXPECTED – ожидаемое количество

Задание №2

Проверить значимость различий уровня образования (s_diplom) мужчин и женщин

Корреляционный анализ

- Теснота линейной зависимости между переменными
- Парный, частный, множественный коэффициенты корреляции
- $-1 \leq r \leq 1$ (парный и частный), $0 \leq r \leq 1$ (множественный)

Диаграммы рассеивания

```
COMPUTE filter_$=(sj10<99999997).
VARIABLE LABELS filter_$ 'sj10<99999997 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

GRAPH
  /SCATTERPLOT=age WITH sj10.
```

```
COMPUTE filter_$=((sj10<99999997) &
(sj60<99999997)).
VARIABLE LABELS filter_$ '((sj10<99999997) &
(sj60<99999997)) (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.

CORRELATIONS
  /VARIABLES=sj10 sj60
  /PRINT=NOSIG.
```

Регрессионный анализ

- Метод исследования зависимости случайной величины Y от переменных X_j (j=1, 2, …, k), рассматриваемых как неслучайные величины
- $Y = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_j x_{ij} + \ldots + \beta_k x_{ik} + \varepsilon_i$
- $Y = X\beta + \varepsilon$
- Метод наименьших квадратов (МНК)

```
REGRESSION
  /DEPENDENT sj10
  /METHOD=ENTER sh5 age sj4.1 s_diplom
  /SAVE PRED RESID.
```

/STATISTICS COEFF OUTS R ANOVA

CHANGE – изменение R^2

CI(value) – доверительный интервал

/CRITERIA=PIN(.05) POUT(.10) – критерий отбора переменных

/NOORIGIN - включение константы в модель

/METHOD=ENTER – принудительное включение

BACKWARD - назад

FORWARD - ввод

STEPWISE - пошагово

Задание №3

Построить диаграмму рассеивания по переменным общего дохода индивида (sj60) и возраста

Построить линейную зависимость общего дохода индивида (sj60) от пола, возраста, уровня образования, статуса занятости (sj1)