ЕВРОПЕЙСКИЙ
УНИВЕРСИТЕТ
В САНКТ-ПЕТЕРБУРГЕ

Лекция 9. Связи между двумя переменными

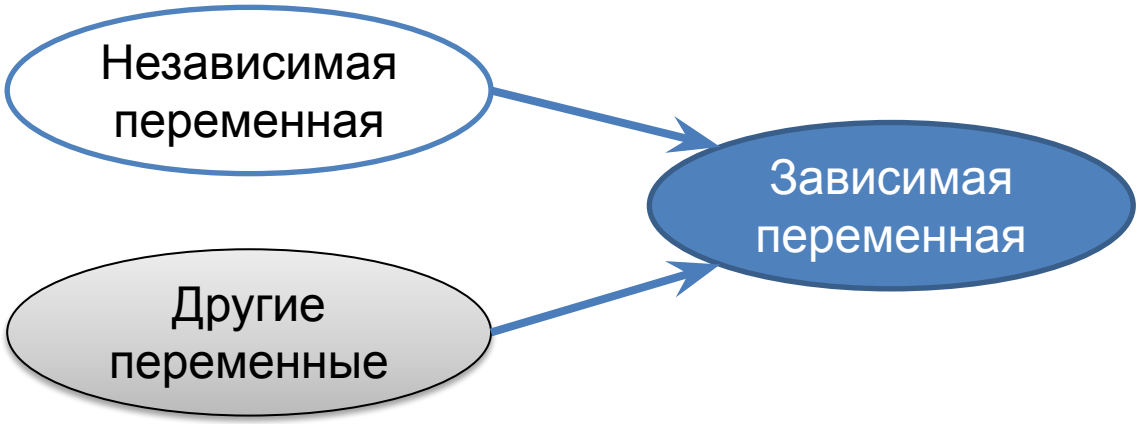
Дмитриева Арина
admitrieva@eu.spb.ru

16 ноября, 2016

Двумерные связи

- Таблицы сопряженности
- Корреляция и регрессия

Исследовательская модель



- Будет ли назначено подсудимому реальный или условный срок в зависимости от наличия детей

Таблица сопряженности

- Позволяет увидеть связи между двумя переменными
  - Номинальной и номинальной
  - Номинальной и порядковой
  - Порядковой и порядковой
- **Таблица сопряженности** (contingency table, cross-tab) – статистический метод, который отражает совместное распределение двух или больше переменных с ограниченным числом категорий

Таблица сопряженности

Вид срока (1=реальный)	Наличие иждивенцев (1=есть)		
	Нет	Есть	Total
Условный	7.849,0	2.946,0	10.795,0
Реальный	9.104,0	3.958,0	13.062,0
Total	16.953,0	6.904,0	23.857,0

Вид срока (реальный / условный) – зависимая переменная, обычно располагается по строкам
Наличие иждивенцев – независимая переменная, обычно располагается по столбцам

Таблица сопряженности, %

- В абсолютных цифрах таблица сопряженности неинформативна

Вид срока (1=реальный)	Наличие иждивенцев (1=есть)		
	Нет	Есть	Total
	%	%	%
условный	46,3	42,7	45,2
реальный	53,7	57,3	54,8
Total	100,0	100,0	100,0

Рассчитан % по столбцам: предполагается, что «наличие иждивенцев» – **независимая** переменная и показывает, какая доля из людей, имеющих иждивенцев получает реальный срок (57,3%) и какая доля из тех, у кого нет детей получает реальный срок (53,7%)

Таблица сопряженности, %

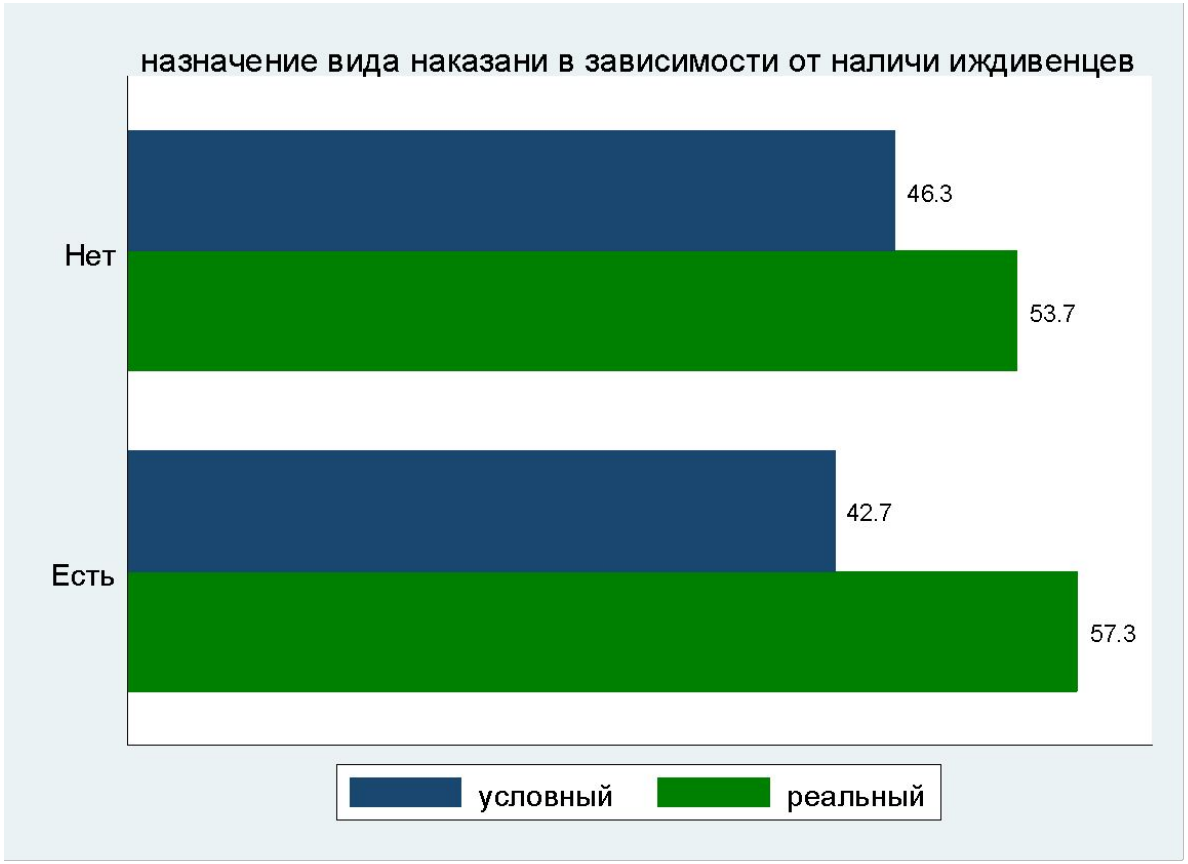
Вид срока (1=реальный)	Наличие иждивенцев (1=есть)		
	Нет	Есть	Total
	%	%	%
условный	72,7	27,3	100,0
реальный	69,7	30,3	100,0
Total	71,1	28,9	100,0

Рассчитан % по строкам: предполагается, что «вид срока» – **независимая** переменная и показывает, какая доля из людей, получивших условный срок имеет иждивенцев (27,3%) и какая доля из тех, кто получил реальный срок имеет иждивенцев (30,3%)

Таблица сопряженности между двумя порядковыми переменными

	образование					
тяжесть	высшее	среднее профессиональное	среднее	базовое	начальное / нет образовании	Total
	%	%	%	%	%	%
нетяжкое	10,0	31,0	39,1	18,4	1,5	100,0
средней тяжести	6,5	26,3	35,2	29,1	2,9	100,0
тяжкое	8,3	26,1	33,6	28,4	3,6	100,0
особо тяжкое	6,9	25,4	37,2	26,0	4,6	100,0
Total	8,4	28,3	36,8	24,1	2,5	100,0

Графическое изображение




```
catplot sentsusp dependants , percent(dependants) blabel(bar, position(outside)
format(%3.1f)) ylabel(none) yscale(r(0,60)) ytitle("") subtitle("назначение вида
наказани в зависимости от наличи иждивенцев")   asyvars bar(1, bcolor(navy)) bar(2,
bcolor(green))
```

ЕВРОПЕЙСКИЙ
УНИВЕРСИТЕТ
В САНКТ-ПЕТЕРБУРГЕ

СВЯЗЬ МЕЖДУ ДВУМЯ МЕТРИЧЕСКИМИ ПЕРЕМЕННЫМИ

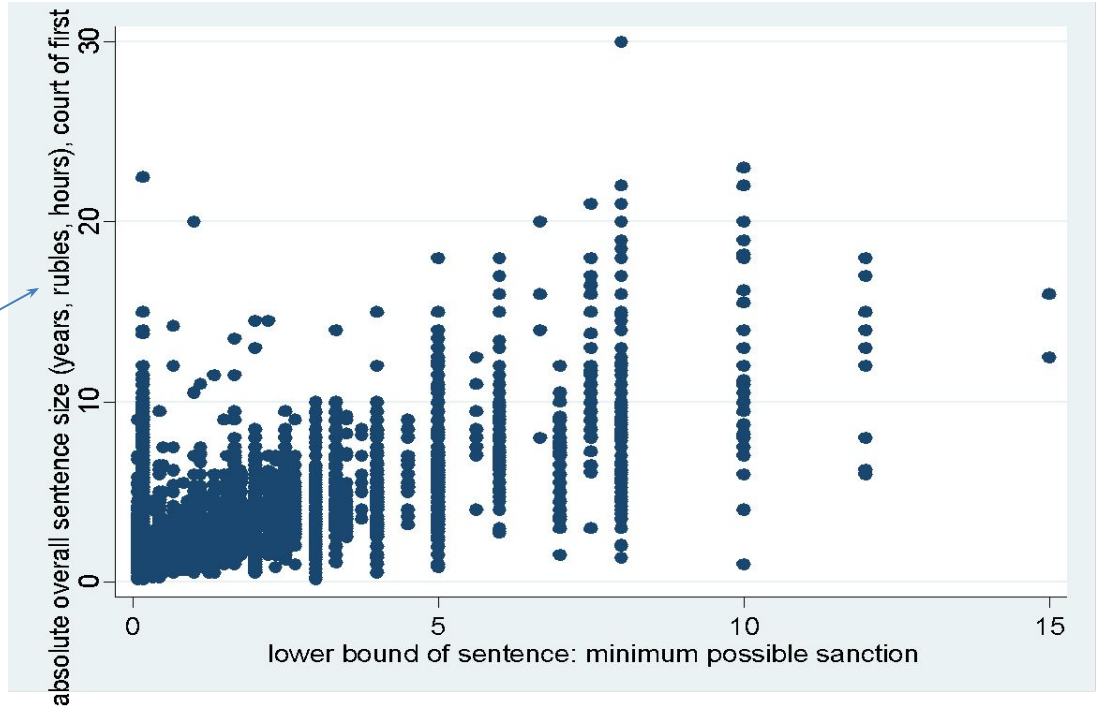
Количественный анализ данных. Тема 3. Двумерная статистика

Переменные

- Sent – размер назначенного наказания (в разных единицах: штраф – в рублях, исправительные работы или тюремное заключение – в годах и месяцах)
- Любая статья УК содержит информацию о нижней (lbound) и верхней (ubound) границе размера наказания
- Наказание может быть назначено:
  - В пределах границ
  - Ниже или выше границ

Диаграмма рассеивания (интервальные данные)

. twoway (scatter sent lbound if inprison==1)



Зависимая переменная

Независимая переменная

Корреляция между двумя переменными

- Корреляция – наличие связи между двумя переменными
  - Эта связь может быть прямой и обратной
  - Размер связи меняется от -1 до 1
  - Прямая связь: *большему* значению X соответствует *большее* значение Y
  - Обратная: *большему* значению X соответствует *меньшее* значение Y
- Гальтон: корреляция роста родителей и детей
- Наиболее известен коэффициент линейной корреляции Пирсона *r*

Коэффициент корреляции Пирсона

Наблюдение	Возраст	Размер наказания
А	31	2
Б	19	2,25
В	39	7,5
Г	19	1
Д	36	7,5
Е	32	2.08

- Каково направление и сила связи между размером наказания и возрастом?
  - Относятся ли судьи мягче к молодым подсудимым?
  - Строже, чтобы «не повадно было впредь»?

  (гипотеза исправления и наказания)

Формула для коэффициента корреляции

- Корреляция – **одно** число, которое объясняет **линейную** связь между двумя переменными
- Основная формула

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}} = \frac{\mathrm{cov}_{XY}}{\sqrt{SS_X SS_Y}}$$

- Корреляция – ковариация деленная на произведение соответствующих среднеквадратических отклонений

Характеристики корреляции

- Наклон:
  - положительная
  - отрицательная
- Сила:
  - сильная,
  - слабая,
  - совершенная
  - Отсутствие корреляции
- Нелинейная корреляция

Требования

- Линейная связь между X и Y
- X и Y являются метрическими переменными
- X и Y являются случайными величинами (выборка должна быть репрезентативна)
- X и Y распределены нормально (но при N>30 требования к распределению снижаются)

Корреляция

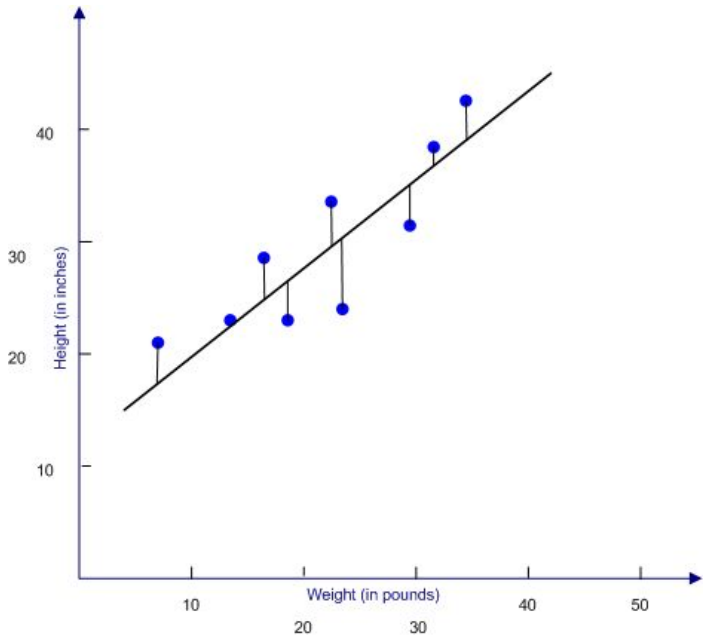
Как связаны размер наказания и количество непогашенных судимостей?

```
. cor sent  priors_count if primary_charge==15801  & dummy9==1
(obs=1669)
```

	sent	priors~t
sent	1.0000	
priors_count	0.3195	1.0000

Регрессионная линия

- Если точки на диаграмме рассеяния аппроксимируются прямой линией, то мы имеем дело с линейной регрессионной моделью



Подгонка линии

Метод наименьших квадратов

Подгонка линии

- Метод наименьших квадратов

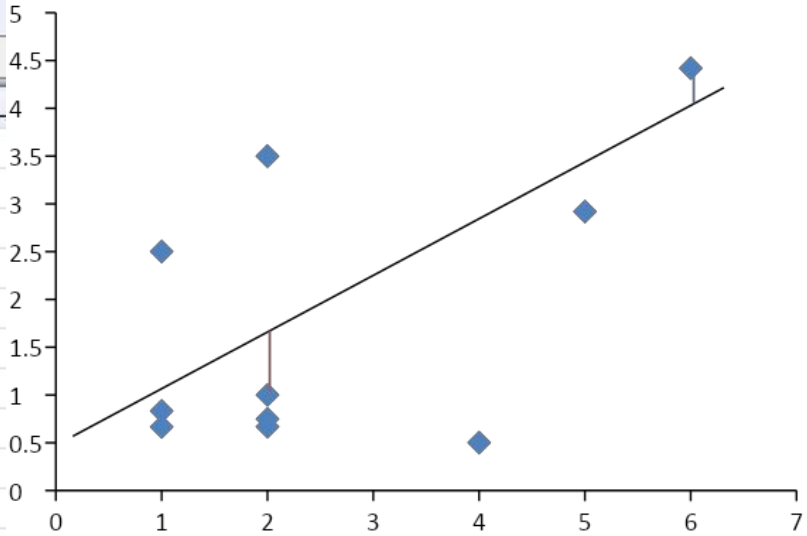
Data Editor (Browse) - [training]

File Edit View Data Tools

lbound[1] .167

Snapshots

	region	priors_count	sent	trial_
103	Krasnoyarskiy krai	2	.75	
104	Samara region	2	1	
105	Jewish Autonomous region	1	2.5	
106	Perm krai	6	4.42	
107	Murmansk region	1	.833	
108	Moscow	2	3.5	
109	Khanty-Mansi okrug	1	.667	
110	Kaliningrad region	5	2.92	
111	Udmurtia Republic	2	.667	
112	Sverdlovsk region	4	.5	
113	Saint Petersburg	1	1	



Регрессионный анализ

- Базовая модель линейной регрессии:

$$\hat{Y}_i = a + b\ X_i$$

  - a - точка пересечения с осью Y (значение Y, когда X равен 0)
  - b - наклон регрессионной линии (изменение Y в ответ на изменение X на 1 единицу), коэффициент регрессии (математически: тангенс угла, образуемого регрессионной линией и осью X)

Регрессионная модель

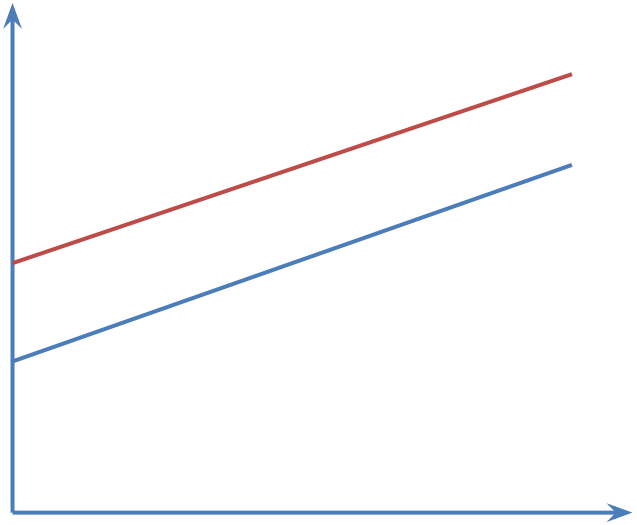
- Метод наименьших квадратов:
- Регрессионный коэффициент:

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

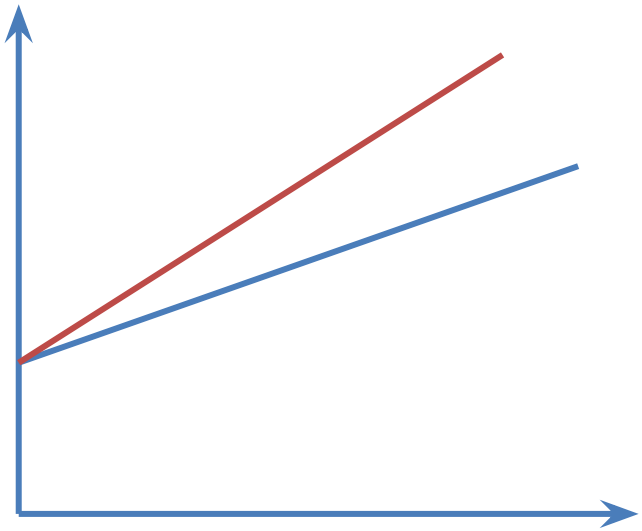
- Пересечение с осью ординат:

$$a = \bar{y} - b\bar{x}$$

- Регрессионная линия всегда проходит через точку $(\bar{x}, \bar{y})$
- Связь между коэффициентом регрессии и коэффициентом корреляции $b = r\frac{s_y}{s_x}$



$$a_1 \neq a_2$$
$$b_1 = b_2$$



$$a_1 = a_2$$
$$b_1 \neq b_2$$

- Предсказанная линия

$$\hat{y} = a + bx$$

- предсказанное значение для x_i

$$\hat{y}_i = a + bx_i$$

- Остатки:

$$e_i = y_i - \hat{y}_i$$

- Сумма квадратов остатков $RSS = \sum {e_i}^2$

Зависимая переменная

Независимая переменная

```
. reg sent  priors_count
```

Source	SS	df	MS
Model	54.9485087	1	54.9485087
Residual	950.200939	1236	.768770986
Total	1005.14945	1237	.812570289

Number of obs	= 1238
F(1, 1236)	= 71.48
Prob > F	= 0.0000
R-squared	= 0.0547
Adj R-squared	= 0.0539
Root MSE	= .8768

sent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
priors_count	.156422	.018502	8.45	0.000	.1201232	.1927207
_cons	.922182	.0435278	21.19	0.000	.8367855	1.007579

Коэффициенты модели

Регрессионное уравнение

- Регрессионное уравнение зависимости размера наказания от числа предыдущих судимостей

$$\hat{y} = 0{,}91 + 0{,}15x$$

- *Какой срок дадут человеку с 3 судимостями?*
- *Человеку с 3 судимостями дадут больше срок, чем человеку с 2 судимостями?*
- *Насколько?*
- *Сколько дадут человеку без судимостей?*

Сумма квадратов

Для проверки качества модели рассчитывают ряд стастик:

- $TSS = \sum(y_i - \bar{y})^2$ – общая сумма квадратов отклонений зависимой переменной от ее среднего
- $RSS = \sum(\hat{y}_i - \bar{y})^2$ – объясненная регрессией сумма квадратов отклонений
- ESS –сумма квадратов остатков $\sum {e_i}^2 = \sum(y_i - \hat{y}_i)^2$

$$TSS = ESS + RSS$$

R-квадрат

- Какую долю разброса данных объясняет модель линейной регрессии?

$$R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Интерпретация

- Корреляция не значит каузация (причинно-следственная связь)
  - X влияет на Y
  - Y влияет на X
  - Z влияет на X и Y
- Экстремальные значения могут сильно повлиять на построение модели

ЕВРОПЕЙСКИЙ
УНИВЕРСИТЕТ
В САНКТ-ПЕТЕРБУРГЕ

СВЯЗЬ МЕЖДУ МЕТРИЧЕСКОЙ И КАТЕГОРИАЛЬНОЙ ПЕРЕМЕННОЙ

Дисперсионный анализ

- Дисперсионный анализ позволяет ответить на вопрос, влияет ли интересующая нас номинальная переменная **(фактор)** на количественную переменную **(отклик)**
- Сравниваются средние переменной отклика для каждой группы (фактора)

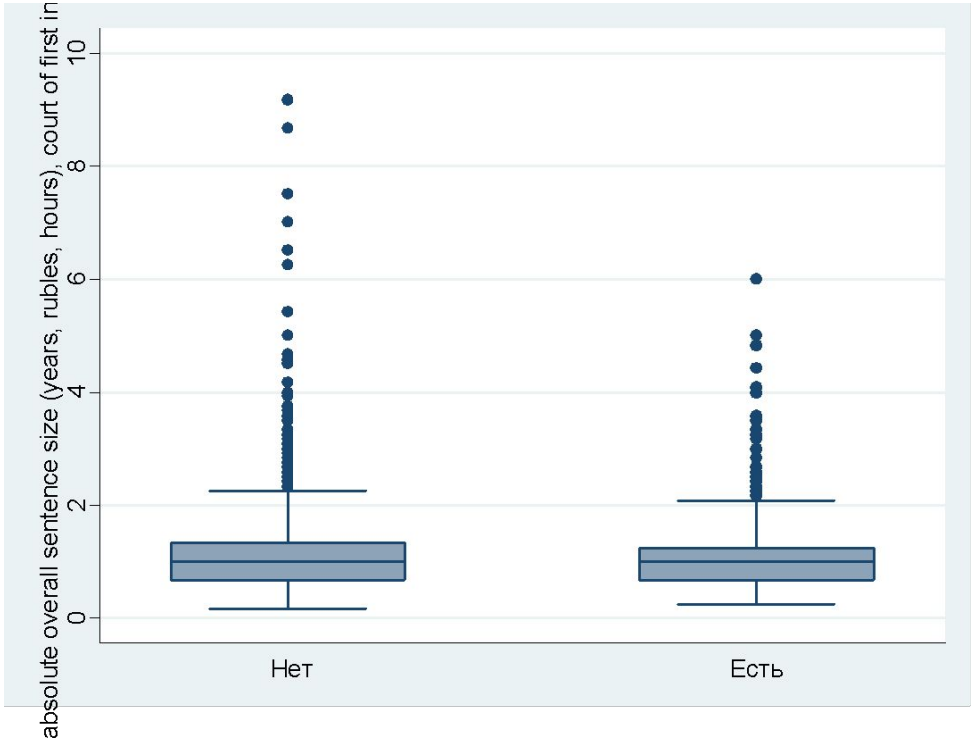
Сравнение средних

- Одинаков ли размер наказания для тех, у кого есть иждивенцы и для тех, у кого нет?

dependants	mean
Нет	1.232543
Есть	1.199237
Total	1.223907

Сравнение, используя ящичковую диаграмму

```
graph box sent , over (dependants)
```



Виды связей между переменными

Уровень измерения		Зависимая переменная		
		Номинальная	Порядковая	Интервальная (отношений)
Независимая переменная	Номинальная	Таблица сопряженности	Таблица сопряженности	Сравнение средних по двум (и более) выборкам
	Порядковая	Таблица сопряженности	Таблица сопряженности	Сравнение средних по двум (и более) выборкам
	Интервальная (отношений)			Коэффициент корреляции Пирсона Регрессионный анализ

ЕВРОПЕЙСКИЙ
УНИВЕРСИТЕТ
В САНКТ-ПЕТЕРБУРГЕ

ПРАКТИЧЕСКОЕ ЗАНЯТИЕ - 3

Пропущенные значения (missing data)

- Dealing with missing data: Key assumptions and methods for applied analysis *Marina Soley-Bori msoley@bu.edu*

Стандартизация показателей

- Z-стандартизация

$$z_{xi} = \frac{x_i - \bar{x}}{s}$$

- Позволяет сравнивать значения, измеренные в разных шкалах
- Например, при поступлении на PhD
  - Петр подал результаты сдачи IELTS = 7,5 (Mean (IELTS) = 6,02, STD = 1,2)
  - Вероника подала результаты сдачи TOEFL = 97 (mean = 85, STD = 18)

У кого английский лучше?

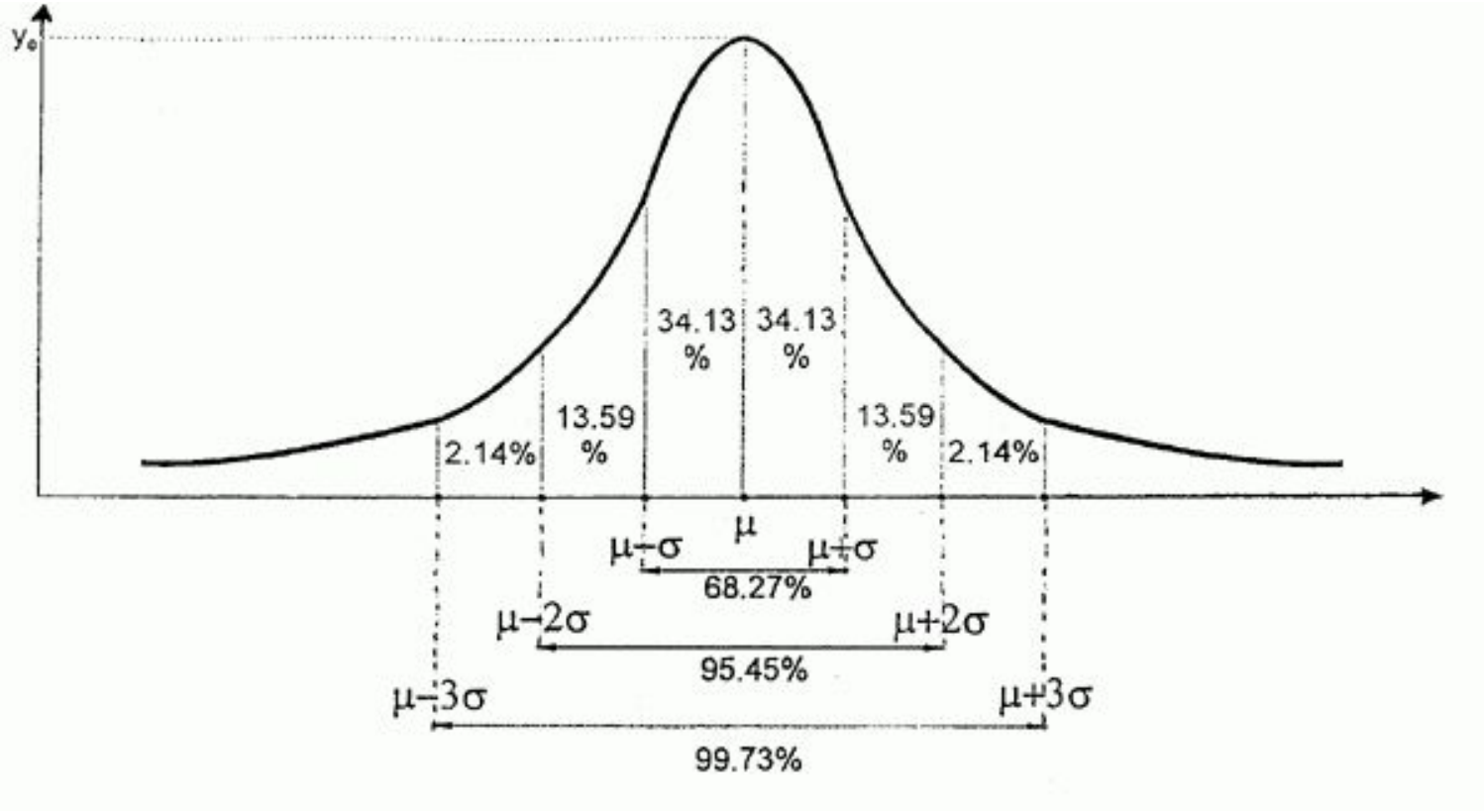
Операция стандартизации

- Стандартизация – преобразование произвольного распределения с параметрами μ, σ в нормальное с параметрами (0,1)

$$z_{xi} = \frac{x_i - \bar{x}}{s}$$

- Стандартизация – смещение распределения и изменение его формы, чтобы оно стало стандартным

Правило «трех сигм»



Создание таблиц сопряженности

- `tabout gravity education using table2.doc, append dpcomma cells (row)`
- `tabout gravity education using table2.doc, append dpcomma cells (row)`

Построение гистограмм для двух категорий

- ```
  catplot sentsusp dependants ,
  percent(dependants) blabel(bar,
  position(outside) format(%3.1f))
  ylabel(none) yscale(r(0,60)) ytitle("")
  subtitle("назначение вида наказани в
  зависимости от наличи иждивенцев") asyvars
  bar(1, bcolor(navy)) bar(2, bcolor(green))
  ```
  ```

Построение корреляций

Корреляция между двумя переменными

- `cor sent episodes`

Все парные корреляции между набором переменных

- `pwcorr sent episodes age`

Регрессионный анализ: этапы

- Построить модель (что является зависимой переменной, что независимой)
- Построить диаграмму рассеяния
- Построить описательные статистики для всех переменных, включенных в модель

Диаграмма рассеяния

Диаграмма рассеяния

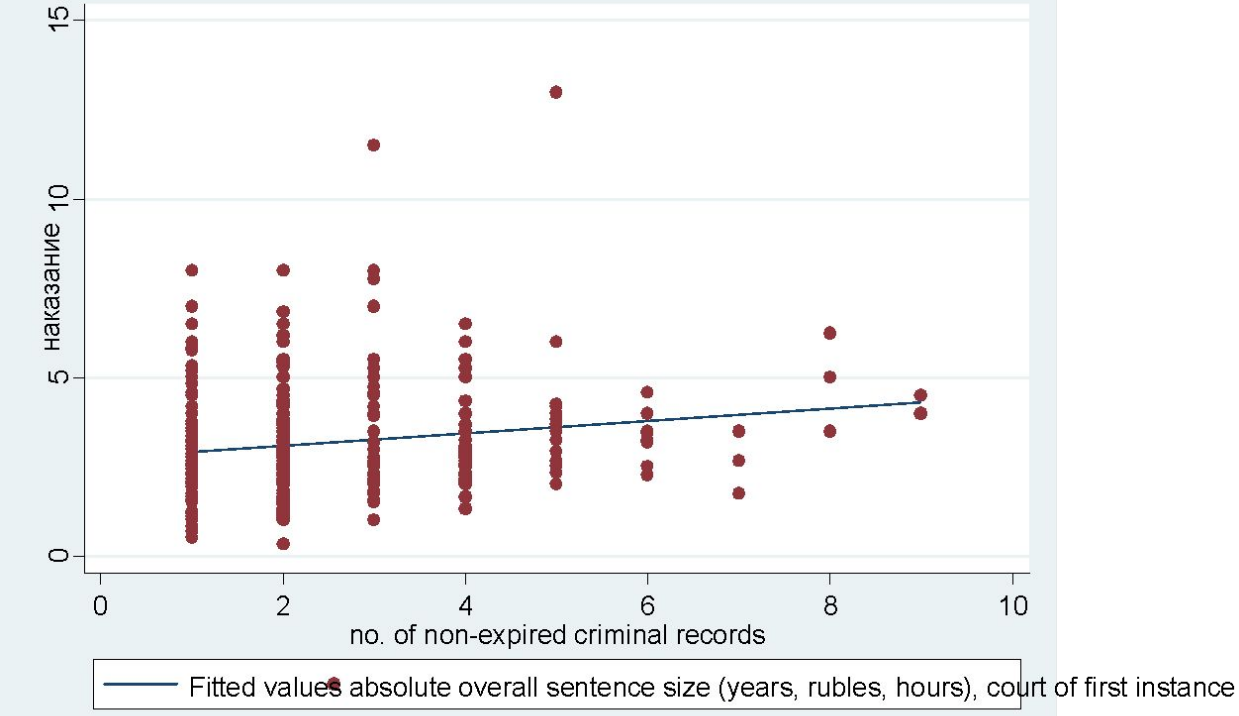
- `twoway (scatter sent priors_count)`

Регрессионная линия

- `graph twoway lfit sent priors_count`

Диаграмма рассеяния и регрессионная линия на одном графике

- `graph twoway (lfit sent priors_count) (scatter sent priors_count) ,`


15
10
5
0
наказание
0
2
4
6
8
10
no. of non-expired criminal records
Fitted values
absolute overall sentence size (years, rubles, hours), court of first instance

Зависимая переменная

Независимая переменная

Коэффициент детерминации R^2

```
. reg sent priors_count if primary_article==228& inprison==1
```

Source	SS	df	MS
Model	45.1525458	1	45.1525458
Residual	21012.3658	3872	5.42674736
Total	21057.5183	3873	5.43700447

Number of obs	= 3874
F(1, 3872)	= 8.32
Prob > F	= 0.0039
R-squared	= 0.0021
Adj R-squared	= 0.0019
Root MSE	= 2.3295

sent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
priors_count	.1156905	.0401076	2.88	0.004	.0370565	.1943245
_cons	3.350603	.0435092	77.01	0.000	3.2653	3.435907

Коэффициент b (наклон)

Константа (a)

```
. reg sent priors_count if primary_article==228& inprison==1
```

Source	SS	df	MS		Number of obs =	3874
					F(1, 3872) =	8.32
Model	(A) 45.1525458	1	45.1525458	(D)	Prob > F =	0.0039
Residual	(B) 21012.3658	3872	5.42674736	(E)	R-squared =	0.0021
					Adj R-squared =	0.0019
Total	(C) 21057.5183	3873	5.43700447	(F)	Root MSE =	2.3295

sent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
priors_count	.1156905	.0401076	2.88	0.004	.0370565	.1943245
_cons	3.350603	.0435092	77.01	0.000	3.2653	3.435907

(A) – объясненная регрессией сумма квадратов отклонений (RSS)
(B) - сумма квадратов остатков (ESS)
(C) - общая сумма квадратов отклонений зависимой переменной от ее среднего (TSS)
(D) - средняя сумма квадратов отклонений модели (RSS/k)
(E) - средняя сумма квадратов отклонений остатков (ESS/n-2)
(F) - средняя общая сумма квадратов отклонений (TSS/(n-1)

Вывод во внешний файл

- `ssc install outreg2`
- `outreg2 using regres1.doc, replace ctitle ("Модель 1") label  addtext(Other controls , NO)`

VARIABLES	(1) Модель 1
no. of non-expired criminal records	0.175***
	(0.0322)
Constant	2.723***
	(0.0774)
Observations	829
R-squared	0.035
Other controls	NO

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Описательные статистики для набора переменных

- `preserve`
- `keep(sent priors_count)`
- `outreg2 using table3.doc, replace sum(log) keep(sent priors_count)`

VARIABLES	(1) N	(2) mean	(3) sd	(4) min	(5) max
priors_count	829	2.037	1.283	1	9
sent	829	3.079	1.208	0.333	13

Отдельные описательные статистики (опция `eqkeep`)

- `outreg2 using table3.doc, replace sum(log) keep(sent priors_count) eqkeep(N mean sd)`

VARIABLES	(1) N	(2) mean	(3) sd
priors_count	829	2.037	1.283
sent	829	3.079	1.208

Средние для групп

```
bysort dependants: outreg2 using
table4.doc, replace sum(log) keep(sent
priors_count) eqkeep( mean sd )
```

VARIABLES	(1) dependants 0 mean	(2) sd	(3) dependants 1 mean	(4) sd
priors_count	2.084	1.340	1.831	0.976
sent	3.097	1.243	2.999	1.041