Графическое представление данных. Характеристики распределений

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

3 февраля 2012

Представление данных в форме таблицы

- Обычно данные представлены в виде таблицы, в которой строки называются наблюдениями или случаями, в столбцы – переменными
- Наблюдениями (единицами анализа) могут быть люди, страны, регионы, школы и т.д.
- В переменных содержится информация, которая нам известна об единицах анализа

Пример 1: люди

id	возраст	пол	образование
1	45	м	1
2	32	м	2
3	51	ж	2
4	34	ж	3
5	70	м	1
6	18	ж	3
7	23	ж	2
8	25	м	1
9	64	м	3
10	22	ж	2

Пример 2: страны

страна	ВВП	убийства	рождаемость
Великобритания	35,860	1.17	1.92
Германия	37,591	0.86	1.42
США	47,184	5.0	2.06
Швеция	38,947	0.99	1.67
Россия	19,840	15.0	1.54
Иран	11,467	3.0	1.89
Украина	6,658	7.0	1.27
Индия	3,586	3.4	2.65
Китай	7,536	1.12	1.54

[0] ВВП на душу населения (2010), число убийств на 100,000 населения (2010), ожидаемое число детей на одну женщину (2010).

Типы переменных (шкалы измерения)

- Количественные (quantitative) переменные:
  - Количество друзей, походов в кино и т.д. Не могут принимать отрицательные и дробные значения (counts)
  - ВВП на душу населения, коэффициент убийств: могут быть дробными, но не могут быть отрицательными (amounts, ratio variables)
  - Пропорции и проценты: меняются в определенном промежутке
  - Интервальные переменные (interval): являются непрерывными, могут быть отрицательными, ноль не имеет содержательной интерпретации (например, температура)

Качественные (qualitative, categorical) переменные

- Номинальные (nominal): поддержка той или иной партии или кандидата, принадлежность к профессиональной группе
- Порядковые (ordered): могут быть упорядочены, но расстояние между значениями неизвестно. Шкала Лайкерта: “совершенно согласен”, “согласен”, “нейтрален”, “не согласен”, “совершенно не согласен”
- Частным случаем порядковых переменных являются дихотомические (dichotomous, binary) переменные, которые могут принимать только два значения, которые часто кодируются как 0 и 1 (есть признак или нет признака). Такие переменные часто называют фиктивными (dummies)

Типы переменных и анализ данных

- Выбор метода статистического анализа данных зависит от типа переменных
- В социальных науках большая часть переменных является качественными
- На практике, различия между переменными разного типа не всегда ясны. Например, шкалу Лайкерта можно анализировать как количественную или как порядковую переменную. Образование может быть номинальной или порядковой переменной и т.д.

Таблица частотного распределения (frequency table)

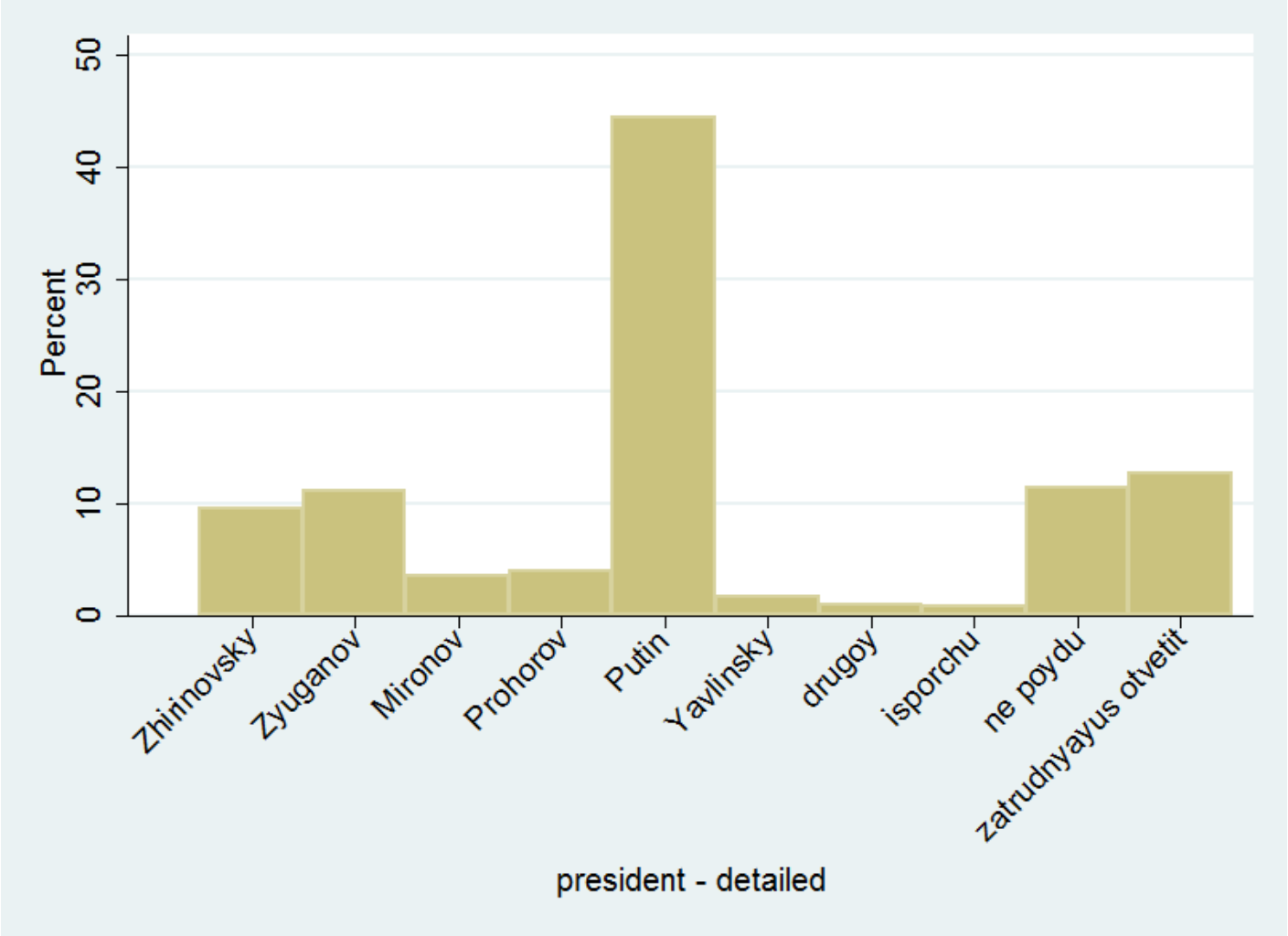
- Простой способ описать качественные переменные. Например, данные опроса ФОМ от 19 января о намерении голосовать на президентских выборах:

	n	%
Путин	1321	44.5
Зюганов	329	11.1
Жириновский	281	9.5
Прохоров	118	4.0
Миронов	106	3.6
Явлинский	50	1.7
другой кандидат	27	0.9
испорчу бюллетень	25	0.8
не пойду на выборы	336	11.3
затрудняюсь ответить	377	12.7
Всего	2970	100

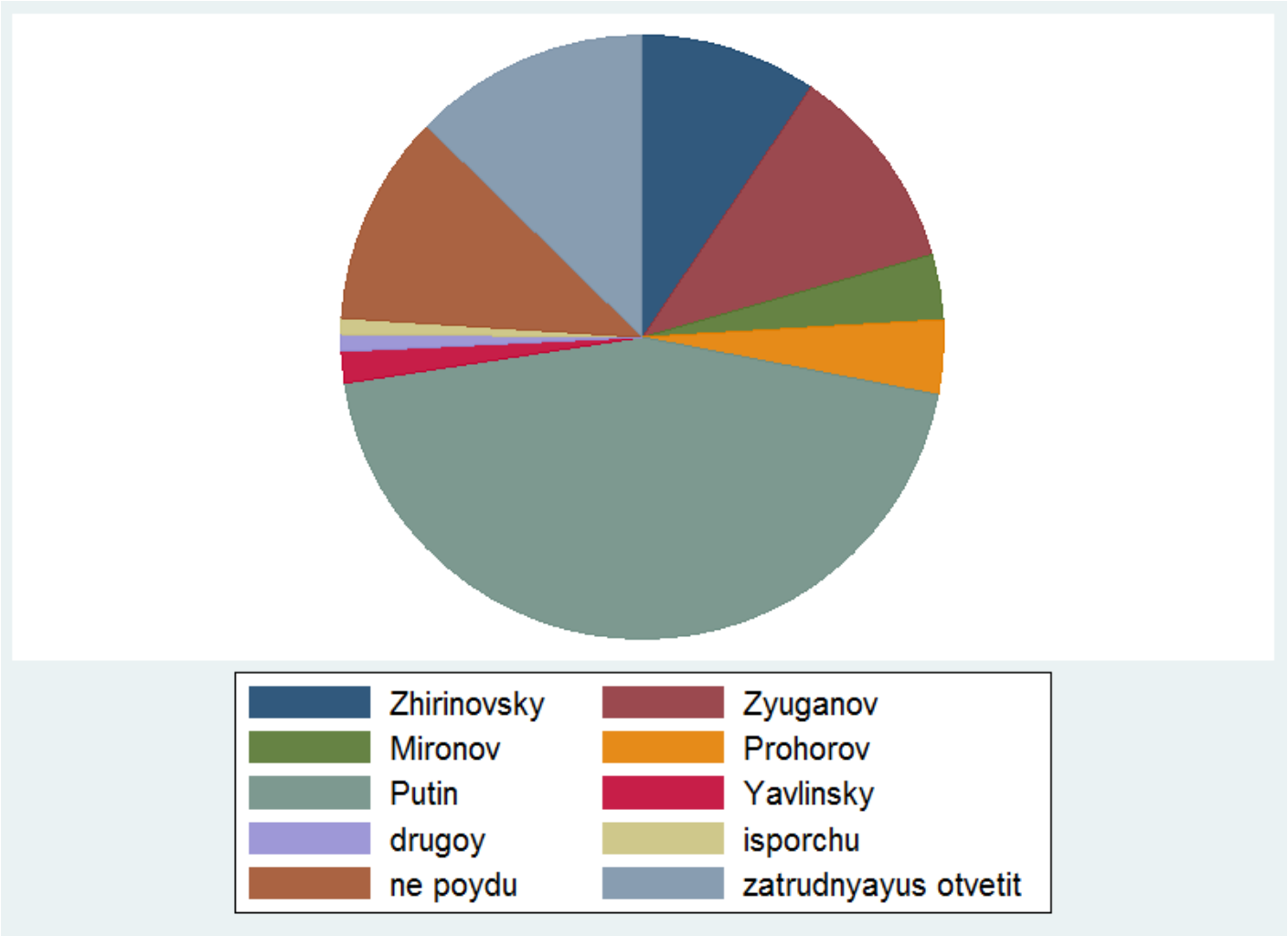
Данные из опроса студентов: образование отца

	n	%
меньше 10-11 классов	2	2.5
10-11 классов	2	2.5
начальное профессиональное	2	2.5
среднее профессиональное	6	7.6
высшее	61	77.2
ученая степень	4	5.1
нет ответа	2	2.5
всего	79	100

Частотное распределение можно представить в виде столбиковой диаграммы



Или круговой диаграммы



Количественные переменные

- Можно охарактеризовать средней ($\bar{x} = \frac{\sum x_i}{n}$) и стандартным отклонением ($s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}$)
- Это не всегда является хорошим описанием данных, особенно в случаях, когда распределение асимметрично
- Альтернативно, распределение можно охарактеризовать медианой, минимумом, максимумом и двумя квартилями (отсекающими 25% и 75% наблюдений)
- Правило пяти чисел Таки (Tukey): min Q1 M Q3 max

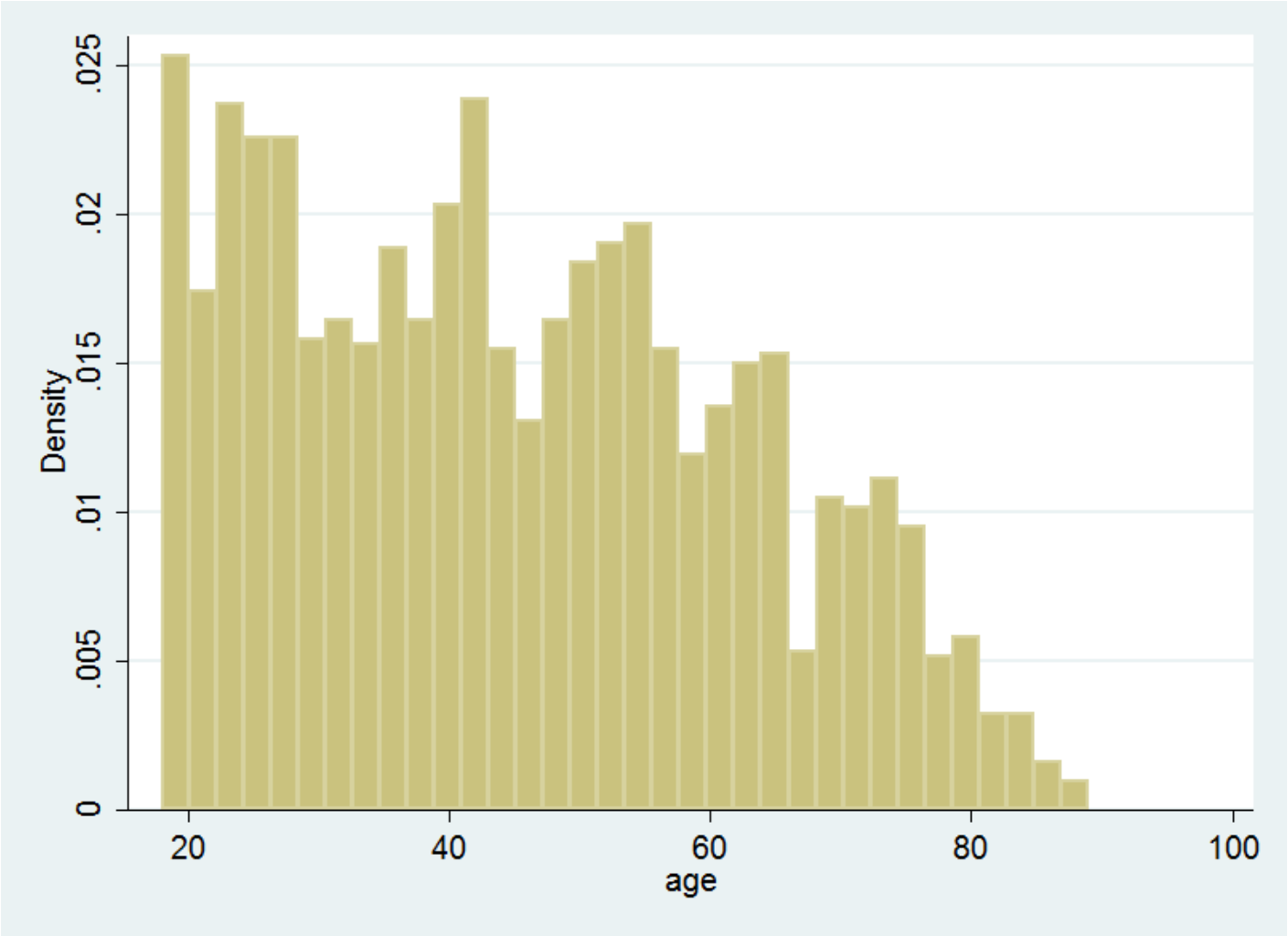
Пример 1: распределение людей по возрасту в выборке ФОМ

- Средний возраст - 45 лет, стандартное отклонение - 17.5
- Пять чисел: 18 30 44 58 89

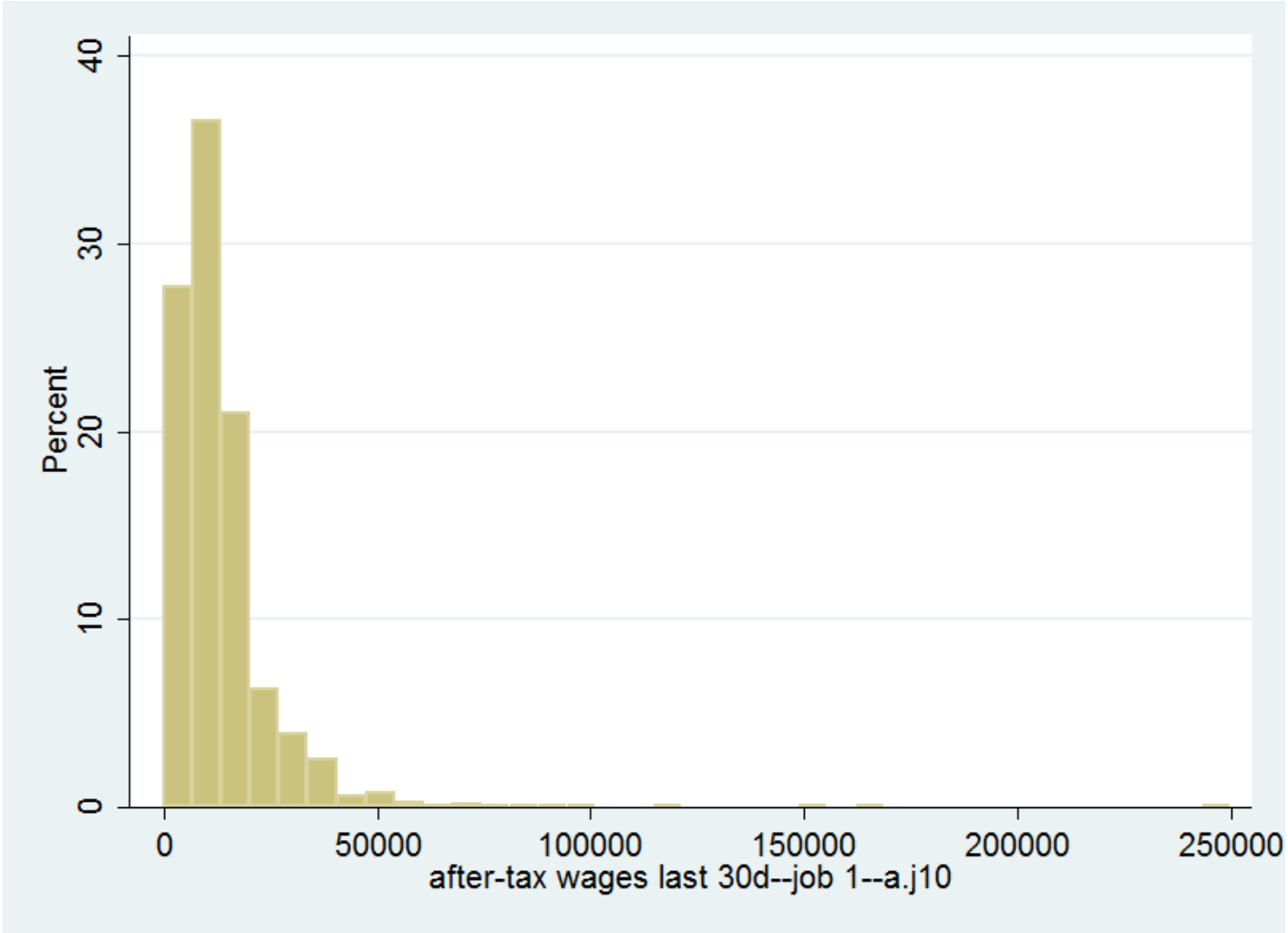
Пример 2: распределение людей по заработной плате в выборке РМЭЗ (2009)

- Средняя зарплата - 13194 рублей, стандартное отклонение - 11259
- Пять чисел: 50 6000 10000 17000 250000

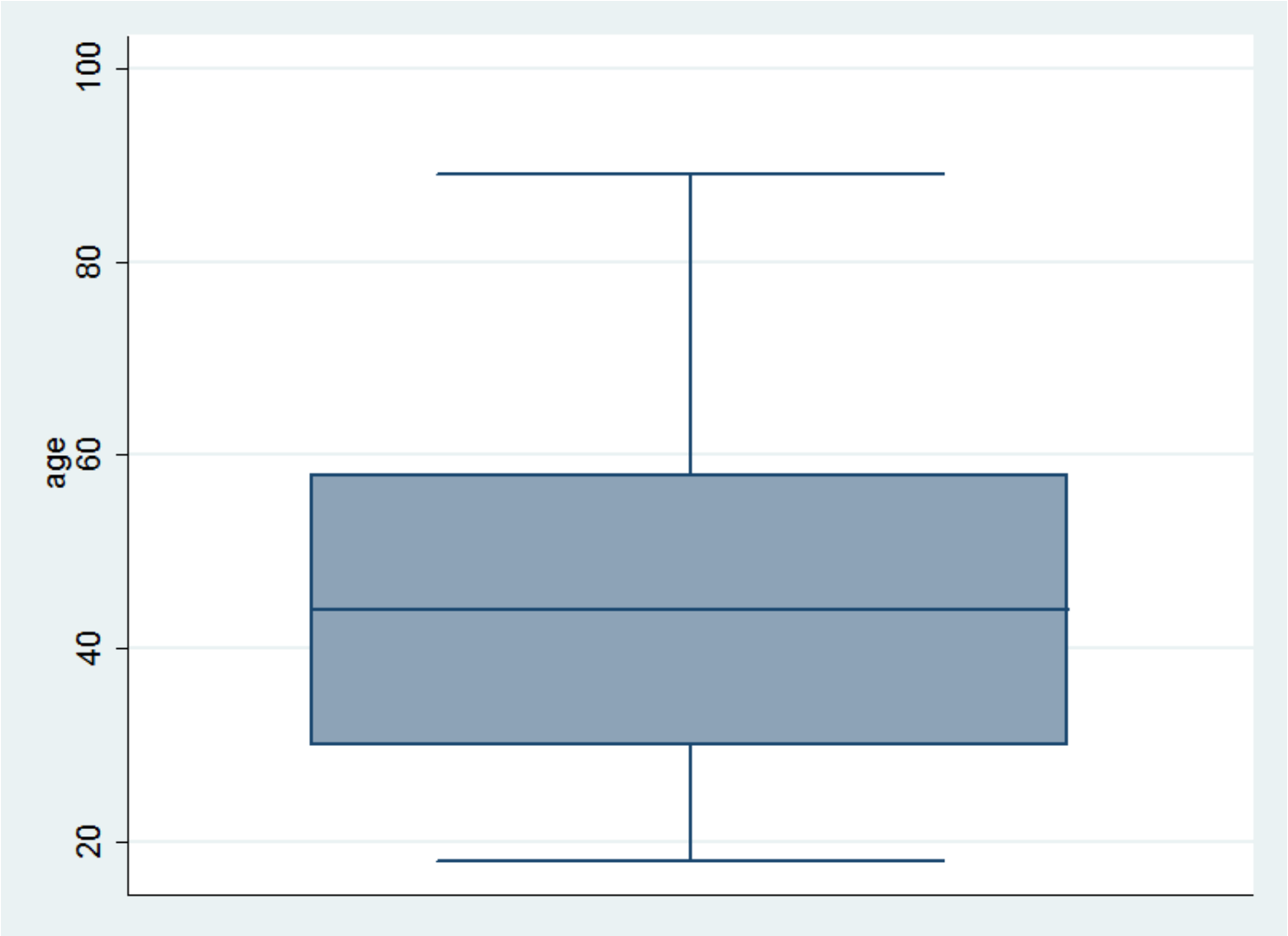
Гистограмма – возраст



Гистограмма – зарплата



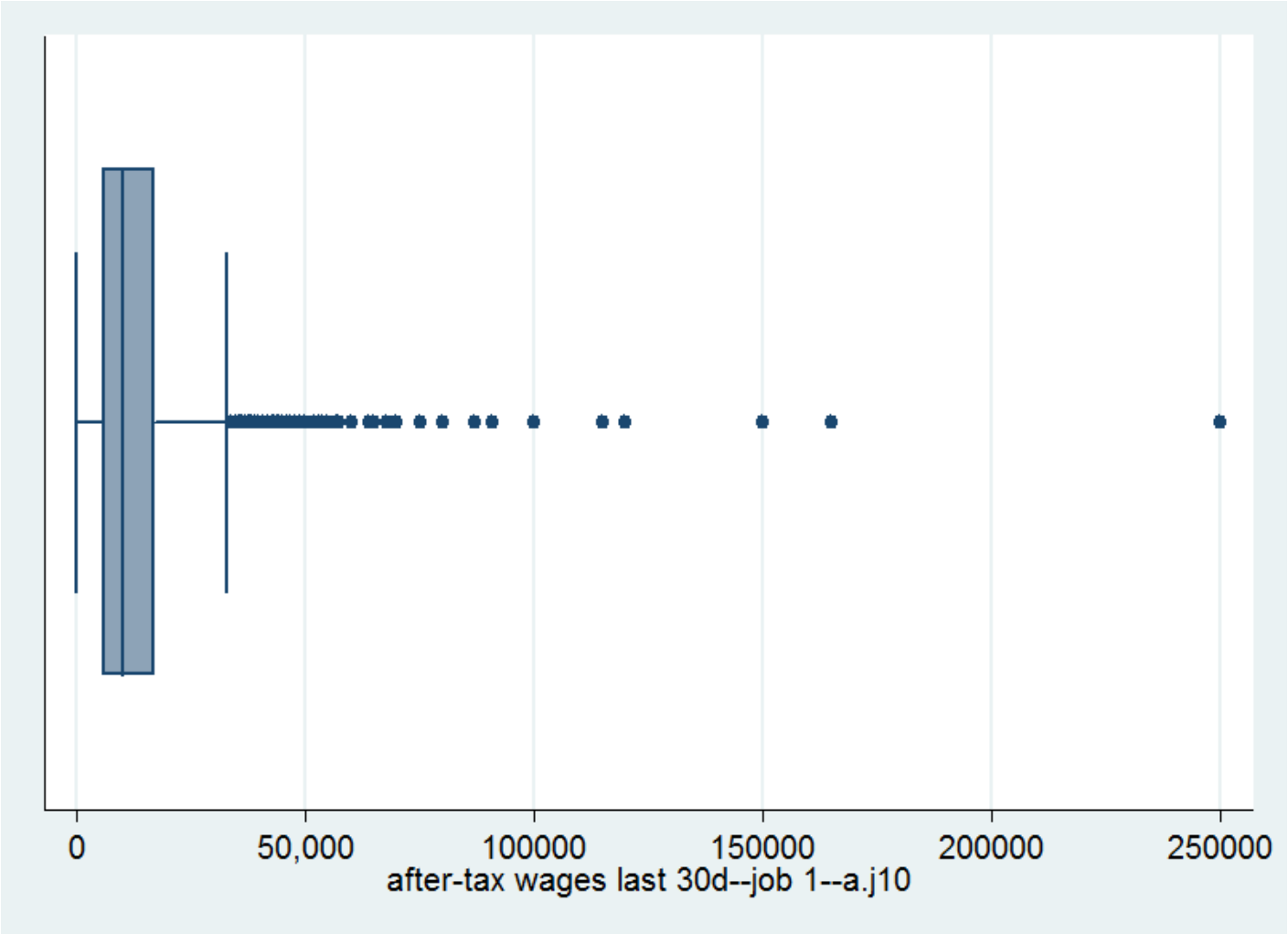
Коробчатая диаграмма (box plot) – возраст



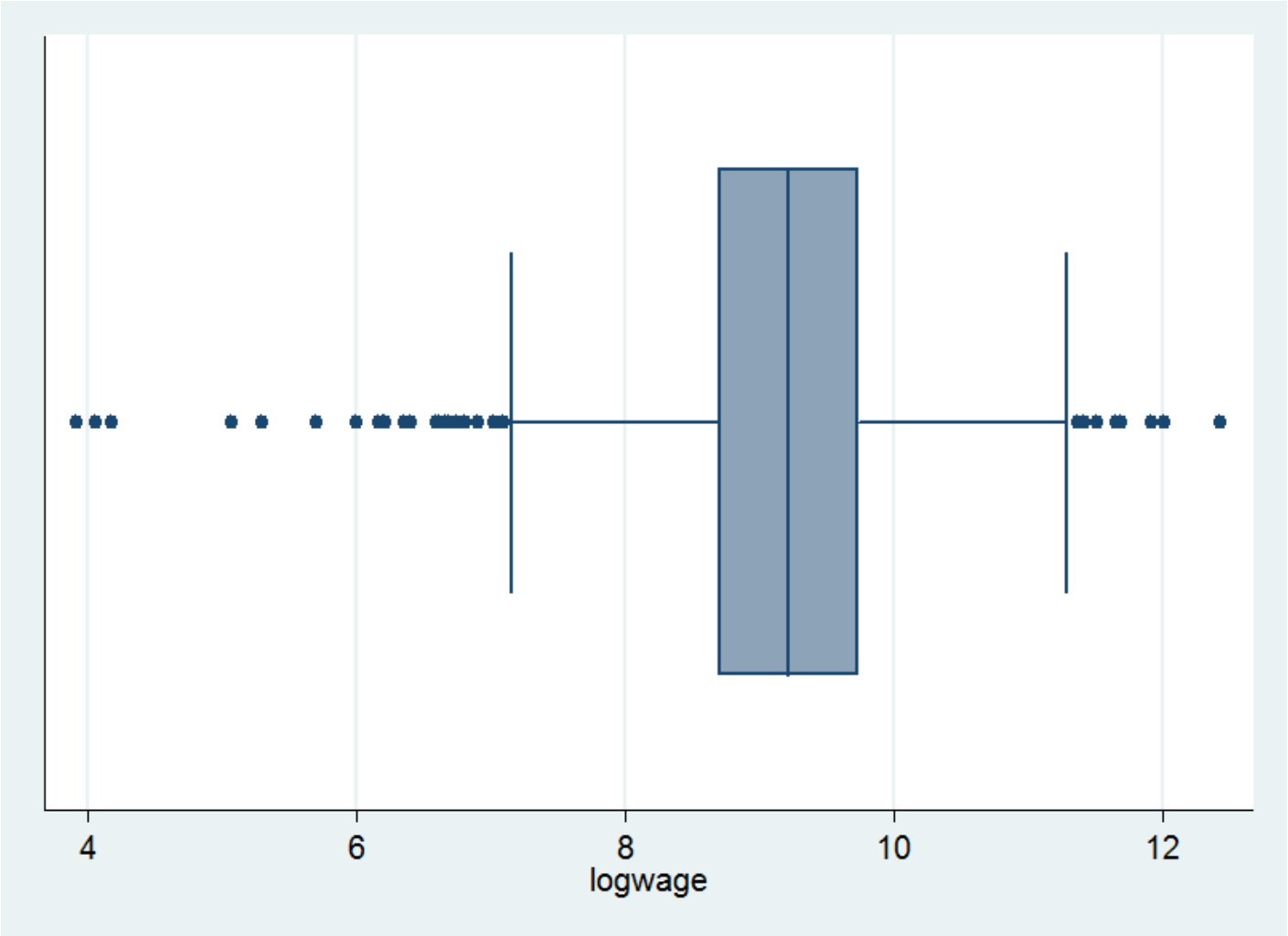
Как строится коробчатая диаграмма

- “Коробка” заключает в себя наблюдения от первого до третьего квантиля
- Линии вокруг нее заключают в себя наблюдения в следующем интервале (с каждой стороны): $1.5(Q3 - Q1)$
- Точки за пределами линий называются выбросами (outliers)

Коробчатая диаграмма (box plot) – зарплата



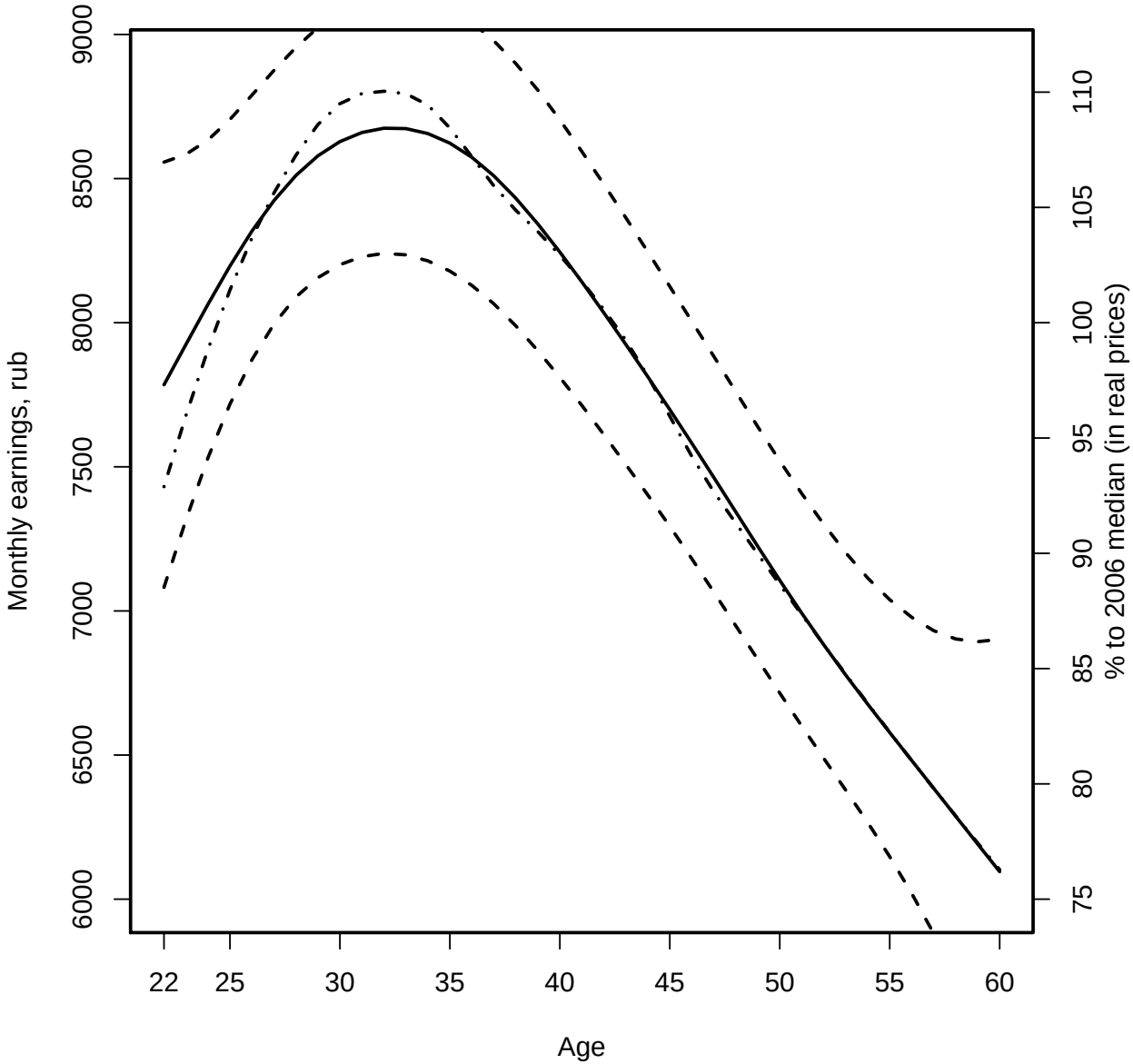
Коробчатая диаграмма – логарифм зарплаты



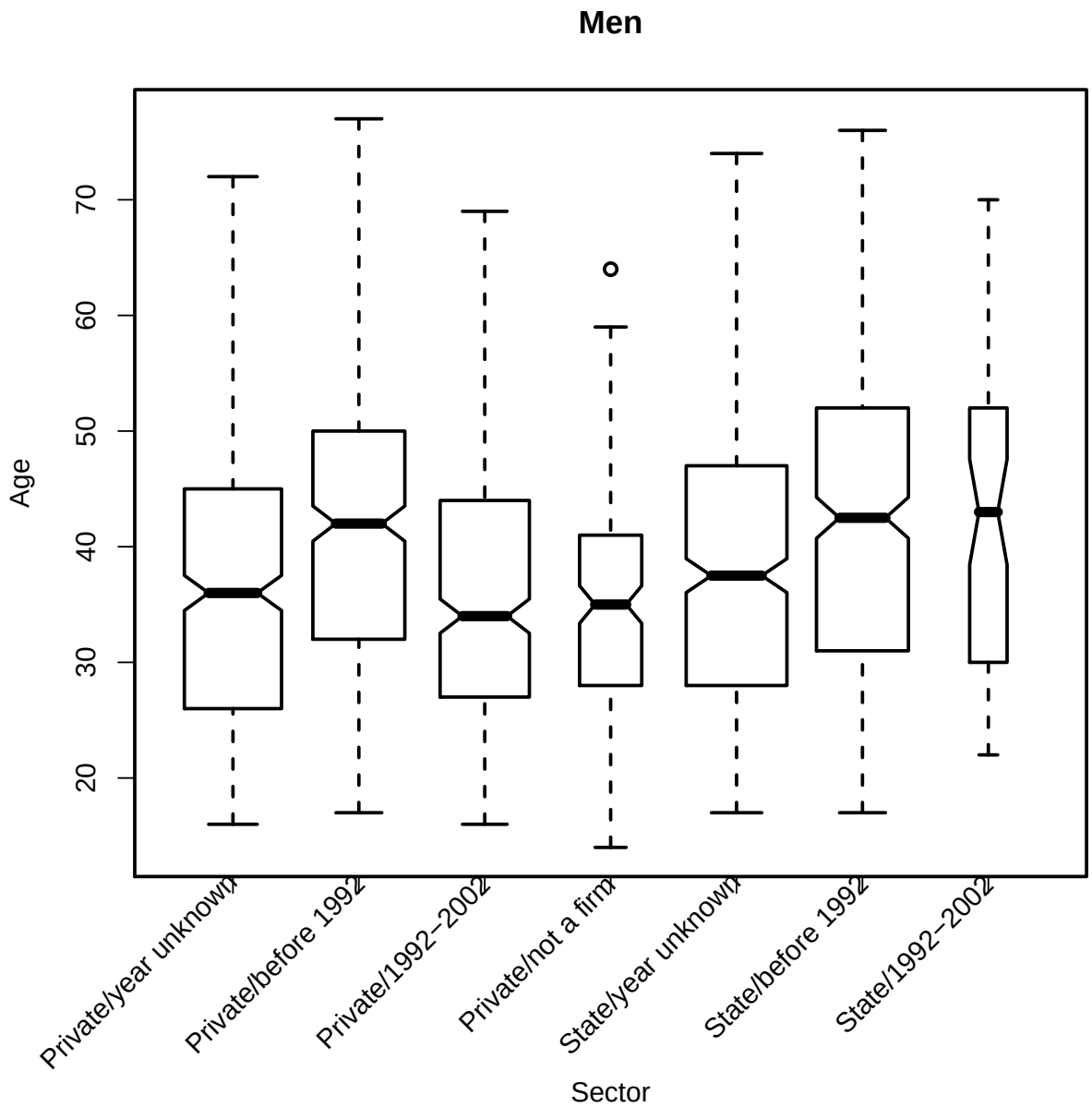
Анализ данных начинается с изучения распределений переменных

- Всегда смотрите, как распределены ваши переменные, перед тем как начать более сложный анализ
- Иногда простые описательные графики могут дать много информации

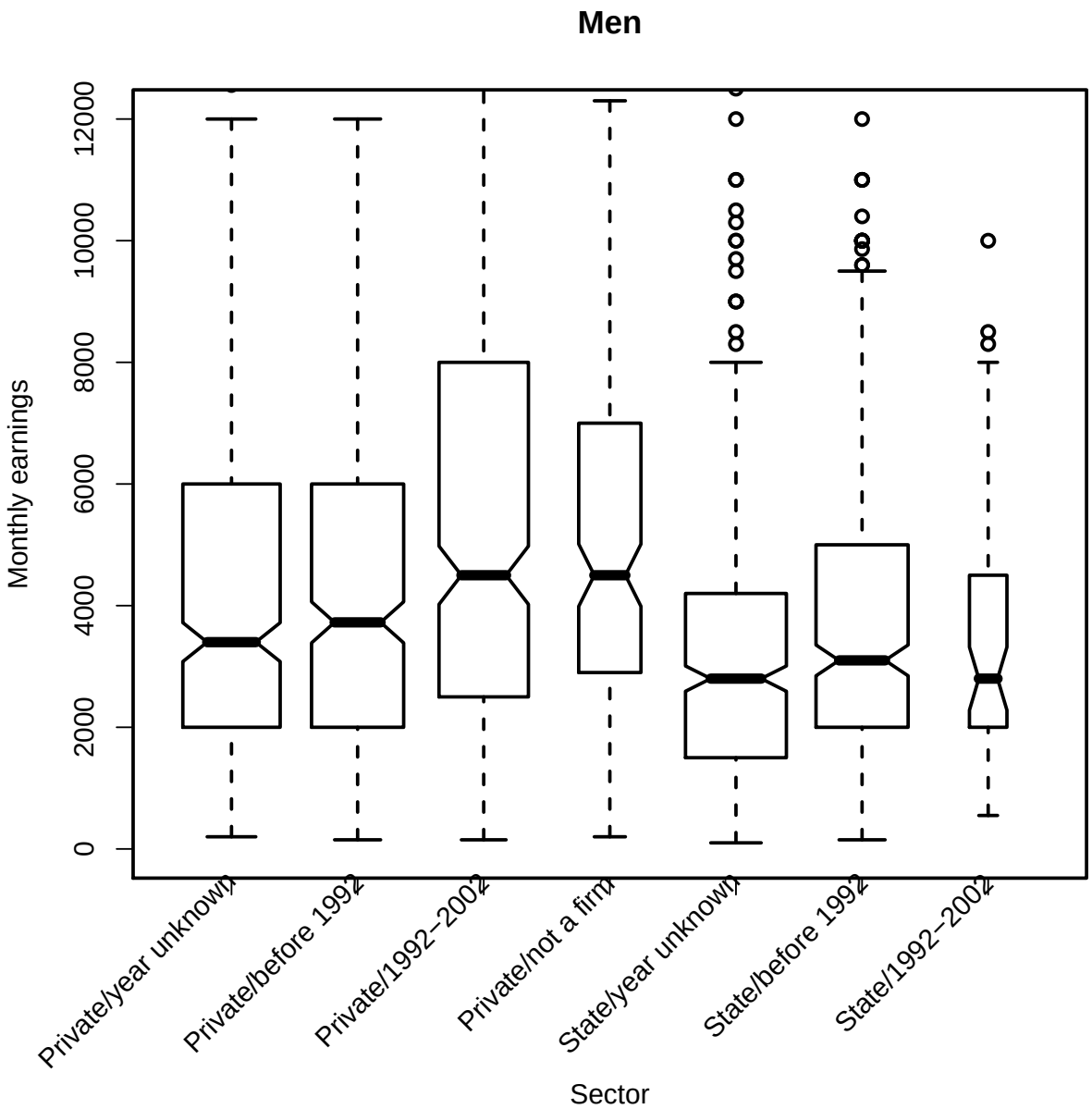
Возраст и зарплата мужчин в России (РМЭЗ 2006)



Возраст по секторам экономики



Зарплата по секторам экономики



Совместное распределение двух (и более) переменных

- В социальных науках нас чаще всего интересуют вопросы о связи нескольких переменных
- Есть ли связь между возрастом (или полом) и заработной платой?
- Верно ли, что люди с разным доходом различаются по своим политическим предпочтениям? и т.д.

Две категориальные переменные: таблица сопряженности (данные ФОМ)

кандидат/образование	среднее и ниже	техникум	высшее	всего
Жириновский	124	108	49	281
Зюганов	143	116	70	329
Миронов	31	42	33	106
Прохоров	27	45	46	118
Путин	574	521	225	1320
Всего	899	832	423	2154

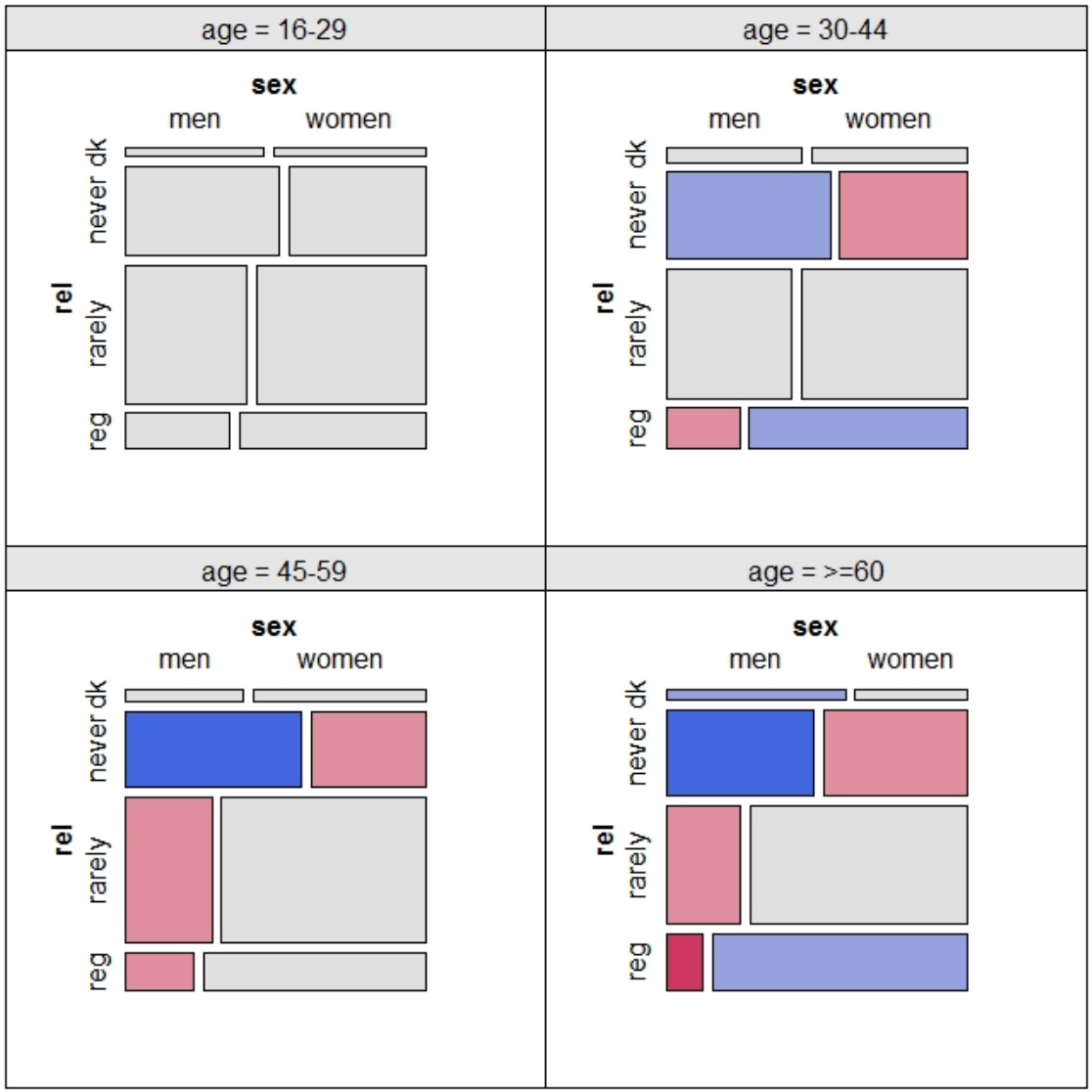
Добавим проценты по строкам

кандидат/образование	среднее и ниже	техникум	высшее	всего
Жириновский	124	108	49	281
%	44	38	17	100
Зюганов	143	116	70	329
%	43	35	21	100
Миронов	31	42	33	106
%	29	40	31	100
Прохоров	27	45	46	118
%	23	38	39	100
Путин	574	521	225	1320
%	43	40	17	100
Всего	899	832	423	2154
%	42	39	20	100

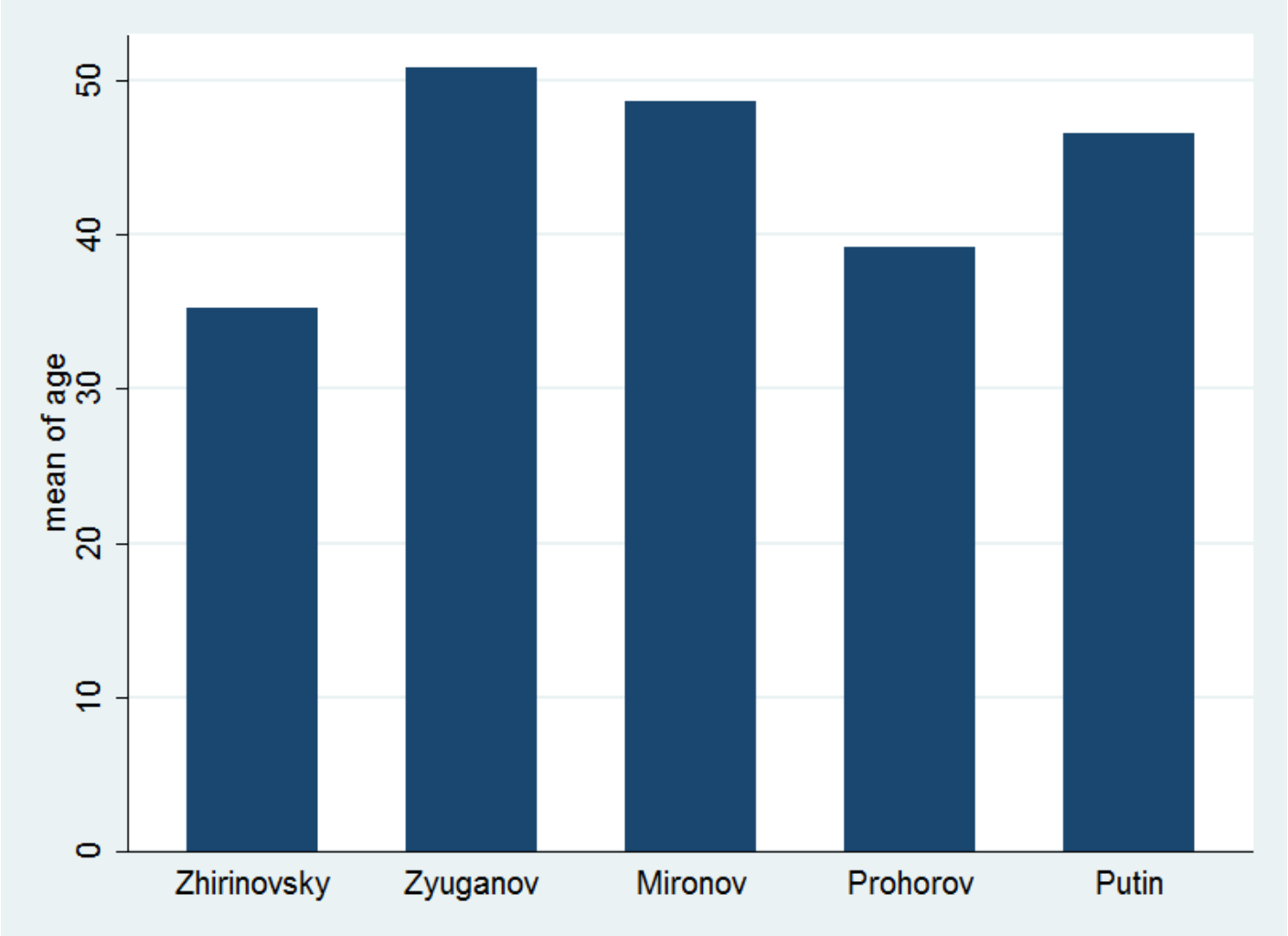
Проценты по столбцам

кандидат/образование	среднее и ниже	техникум	высшее	всего
Жириновский	124	108	49	281
%	14	13	12	13
Зюганов	143	116	70	329
%	16	14	17	15
Миронов	31	42	33	106
%	3	5	8	5
Прохоров	27	45	46	118
%	3	5	11	5
Путин	574	521	225	1320
%	64	63	53	61
Всего	899	832	423	2154
%	100	100	100	100

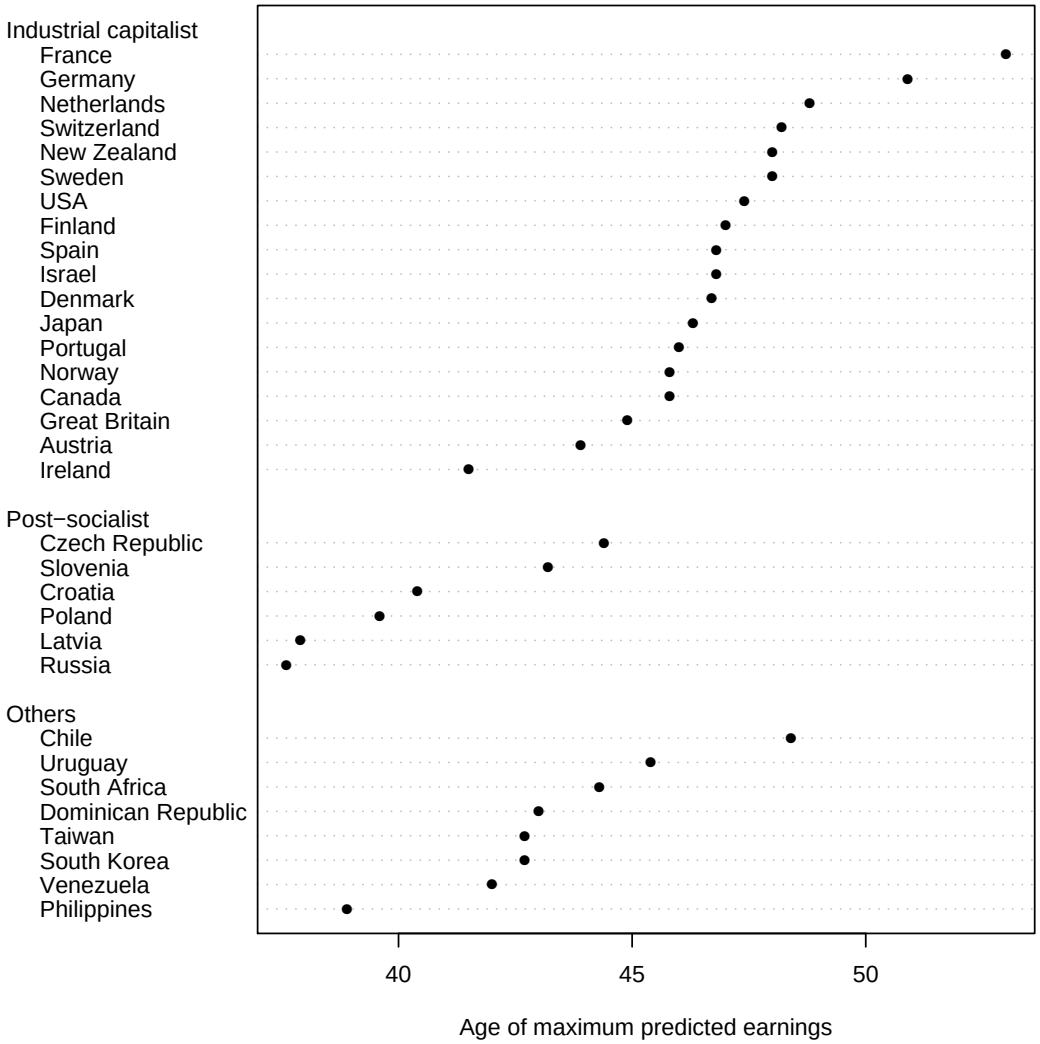
Визуализация связей в таблицах сопряженности (ESS 2008, Россия)



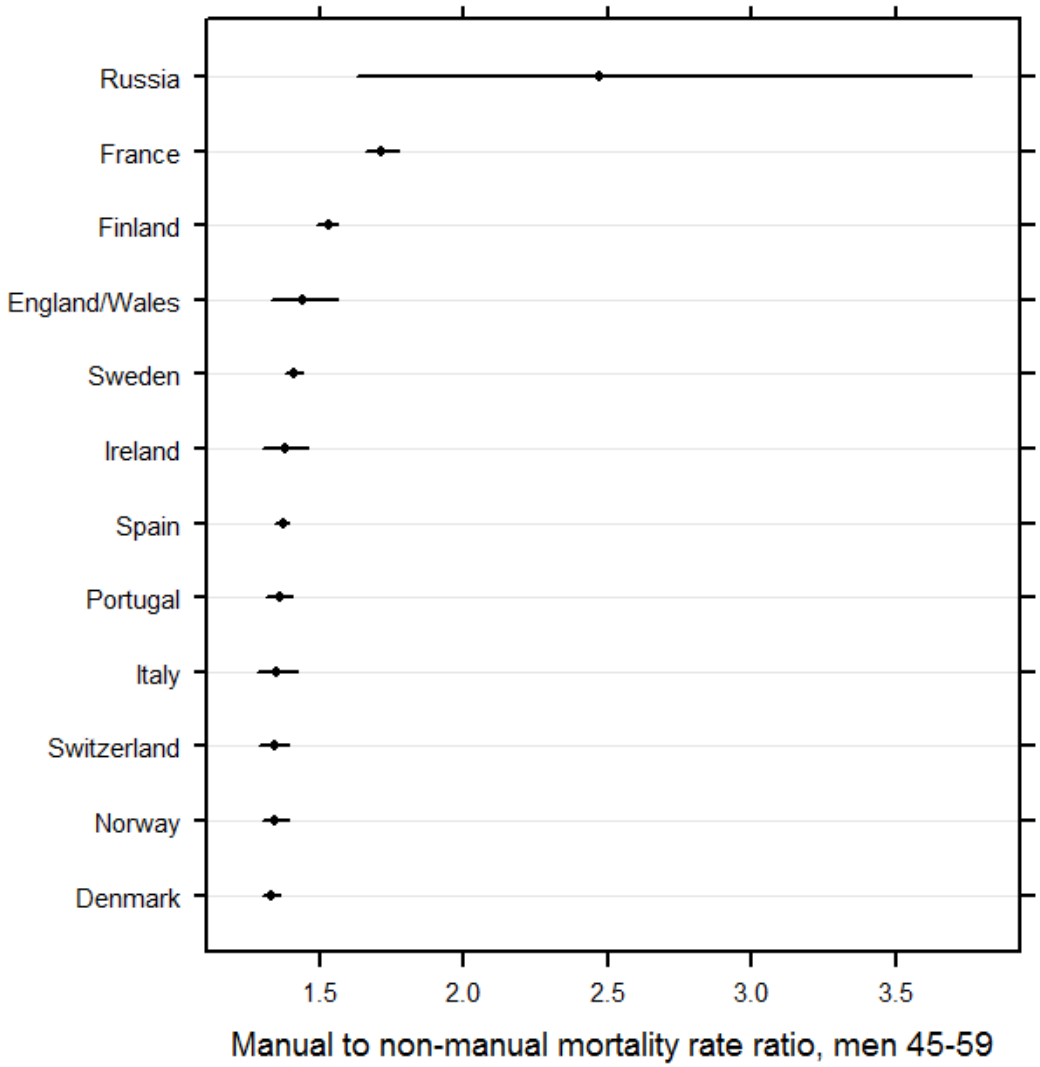
Номинальная и непрерывная переменная: столбиковая диаграмма



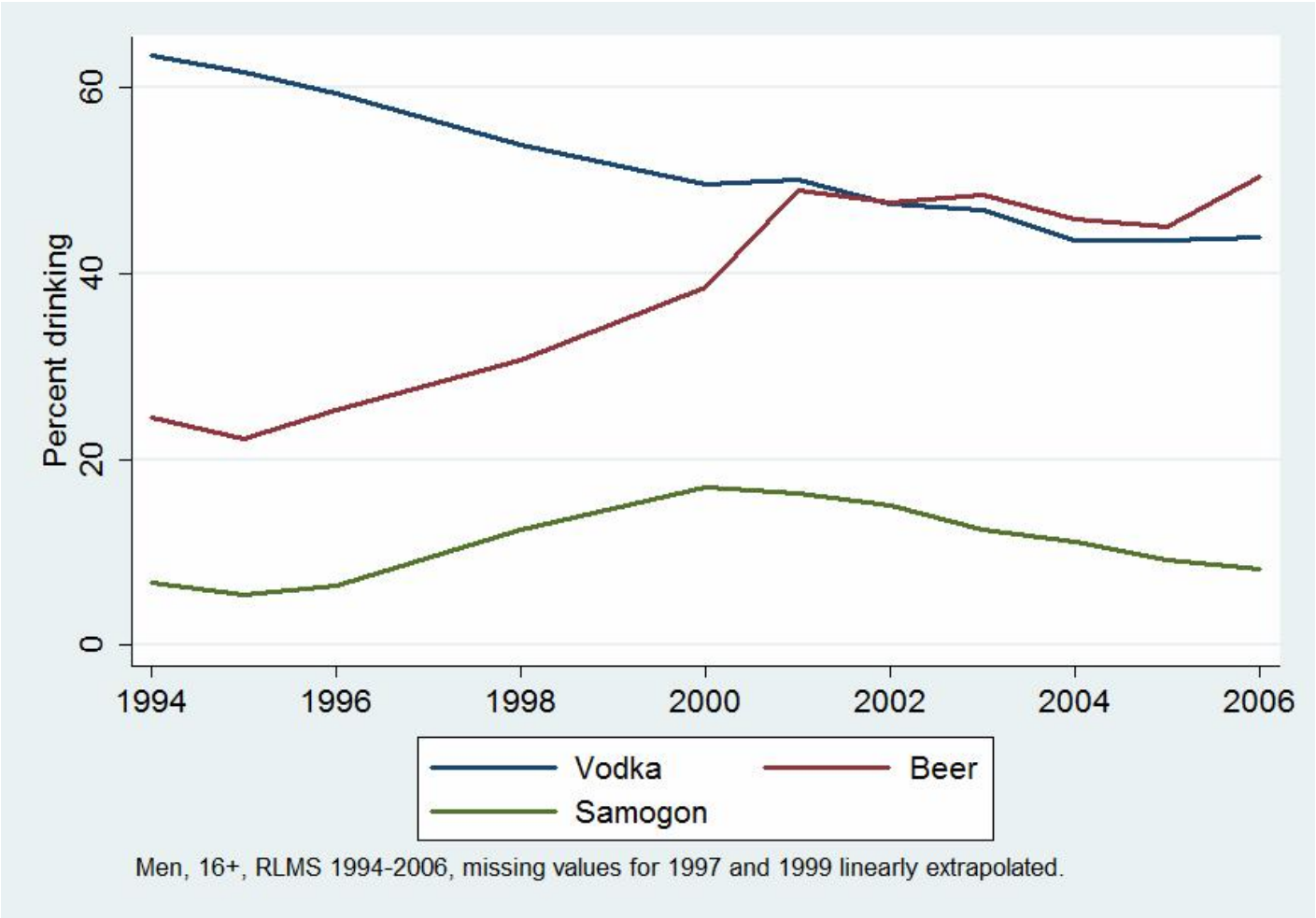
Точечная диаграмма (dot plot): возраст максимальной зарплаты мужчин по странам



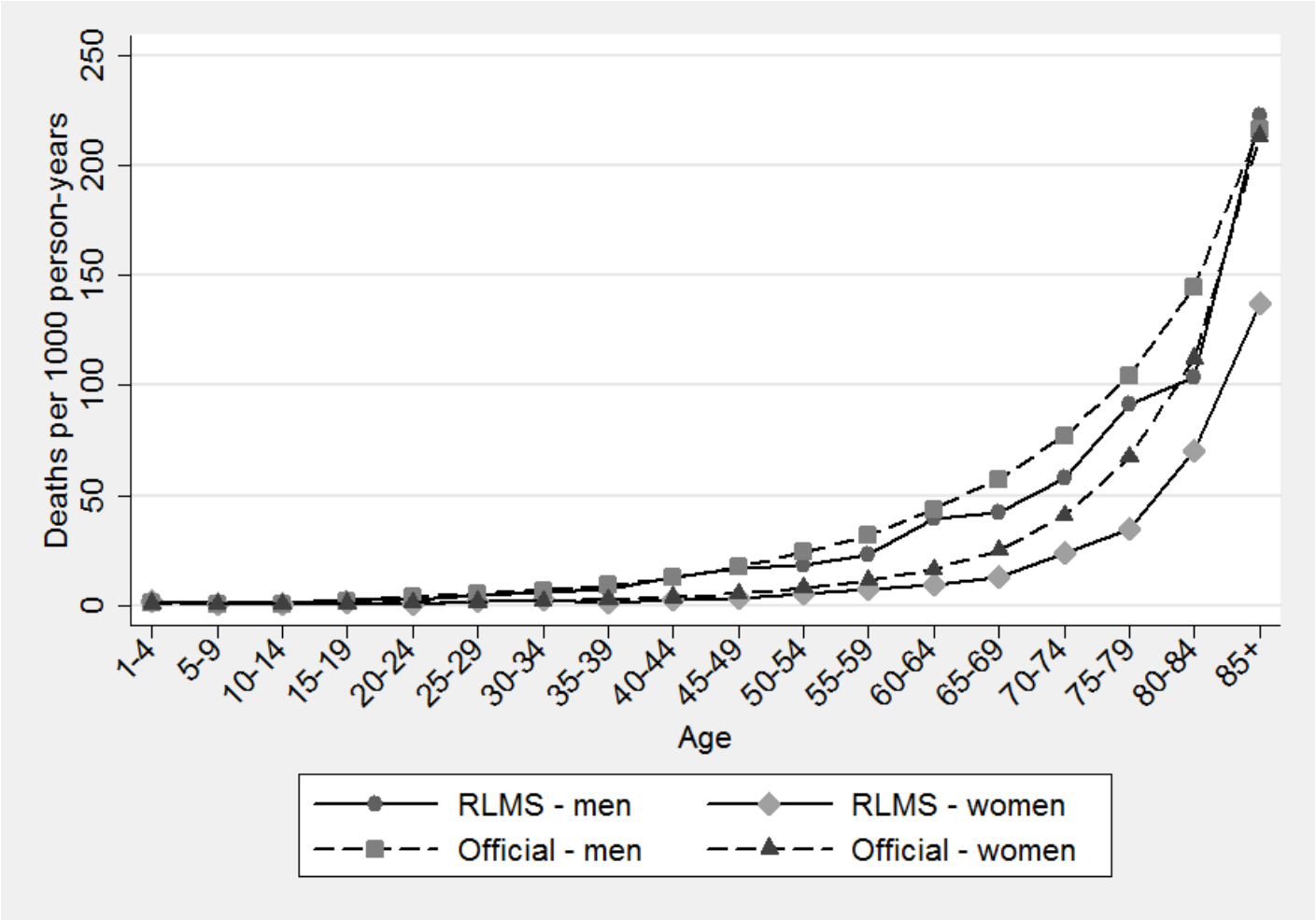
Точечная диаграмма: классовое неравенство в смертности в России и европейских странах



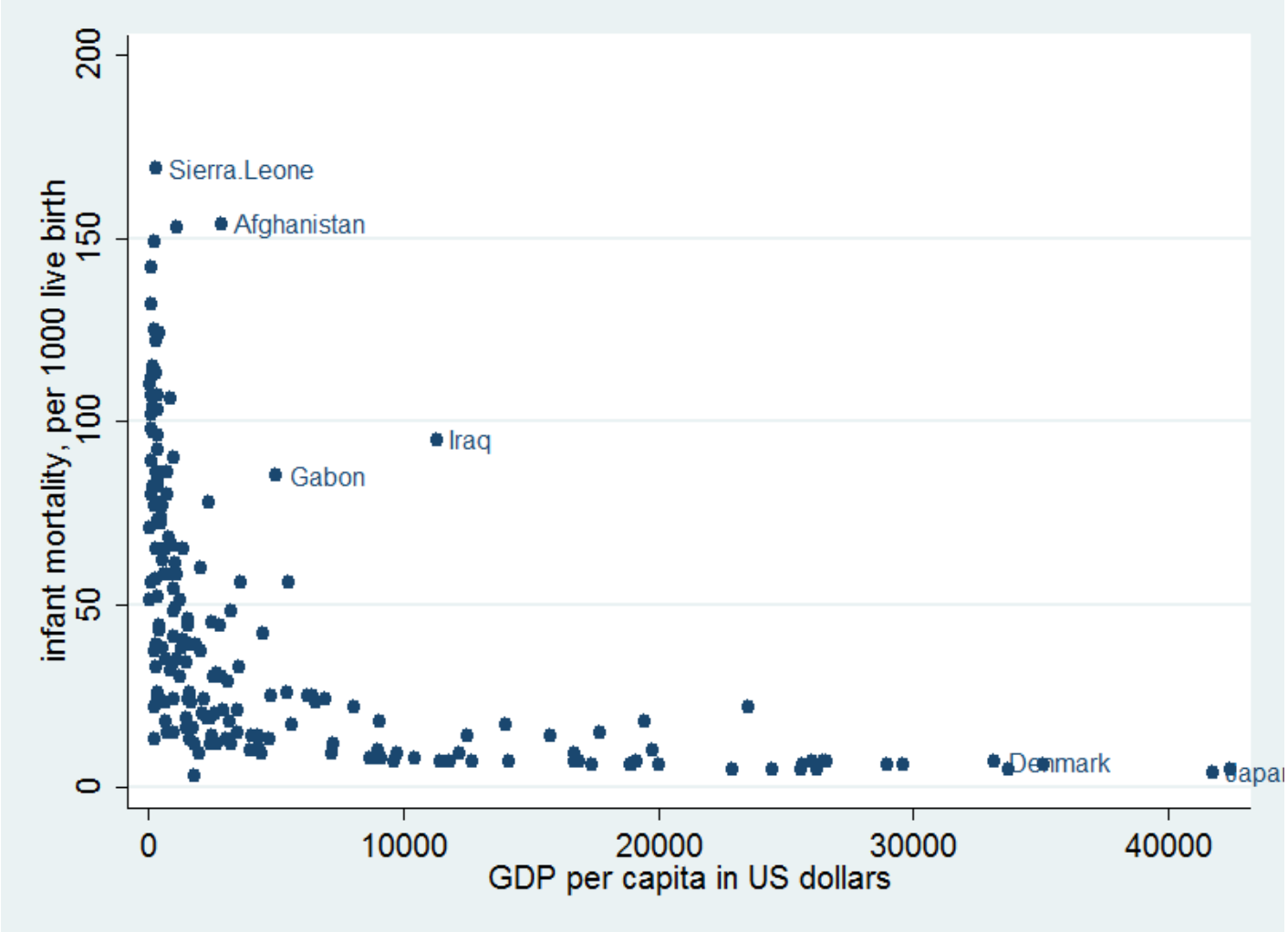
Линейная диаграмма: “пивная революция” в России



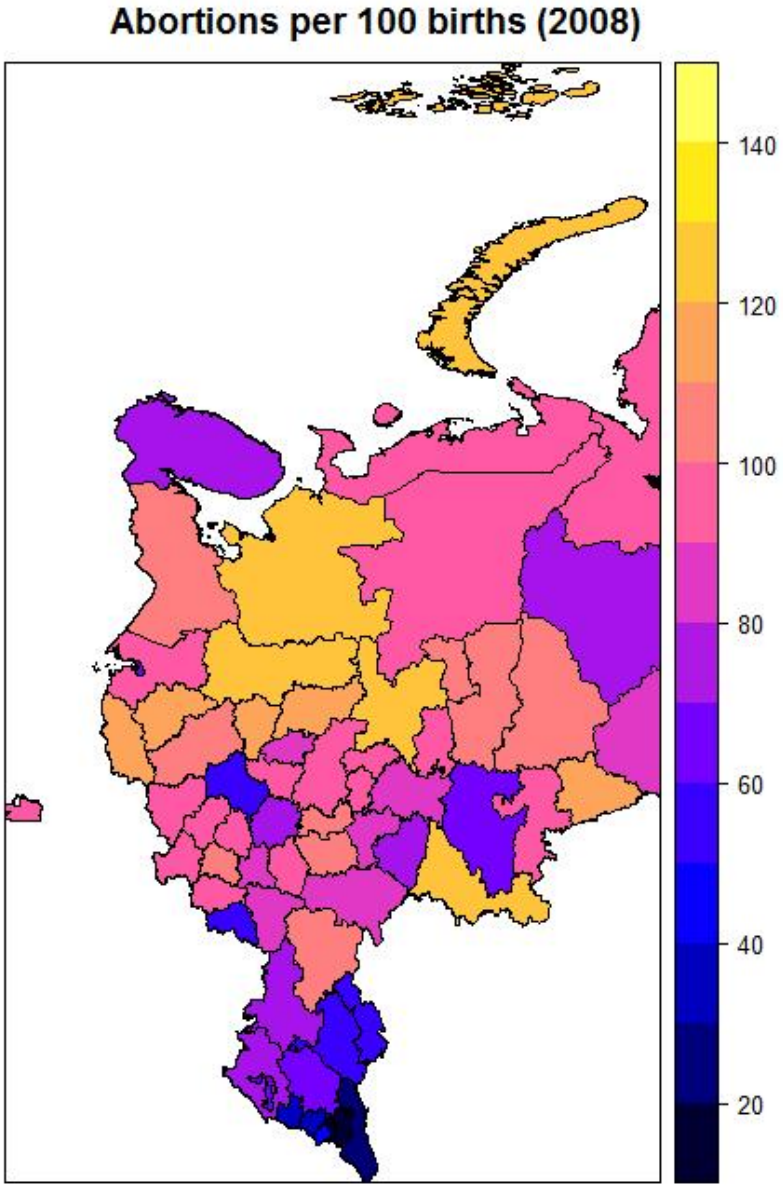
Возрастные коэффициенты смертности



Две количественные переменные: диаграммы рассеивания (scatter plots). Данные ООН, 1998



Карты: Аборты в регионах России



Сети: романтические и сексуальные связи в американской школе (Bearman, Moody, Stovel AJS 2004)

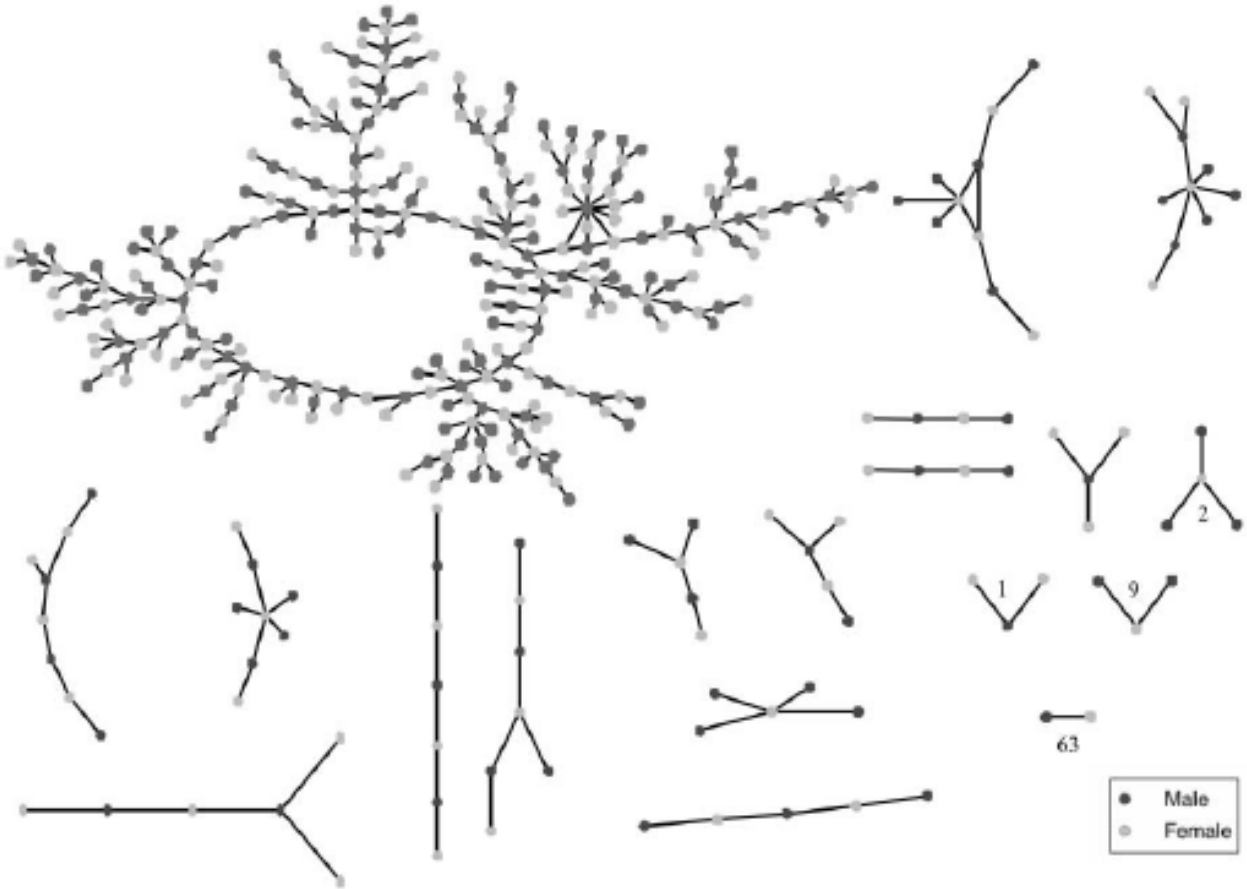


FIG. 2.—The direct relationship structure at Jefferson High

Следующее занятие

- Корреляция и линейная регрессия