

Основы визуализации данных: графики, правила создания, типичные ошибки

Алексей Кнорре

Email: aknorre@eu.spb.ru

WWW: alexeyknorre.ru

Ссылка на презентацию: alexeyknorre.ru/courses/datavis2016/datavis-2.pdf

Приципы создания графиков

Графики должны:

- Показывать и раскрывать (reveal) данные
- Заставлять человека думать о данных, а не о том, как они показаны
- НЕ искажать данные
- Показывать много данных с минимумом использованных графических элементов (data-ink ratio)
- Показывать большие наборы данных целиком и однообразно
- Позволять читателю сравнивать разные кусочки данных

Edward Tufte. Visual Display of Quantitative Information, 2007, p.13.

The aim of good data graphics is to display
data accurately and clearly.

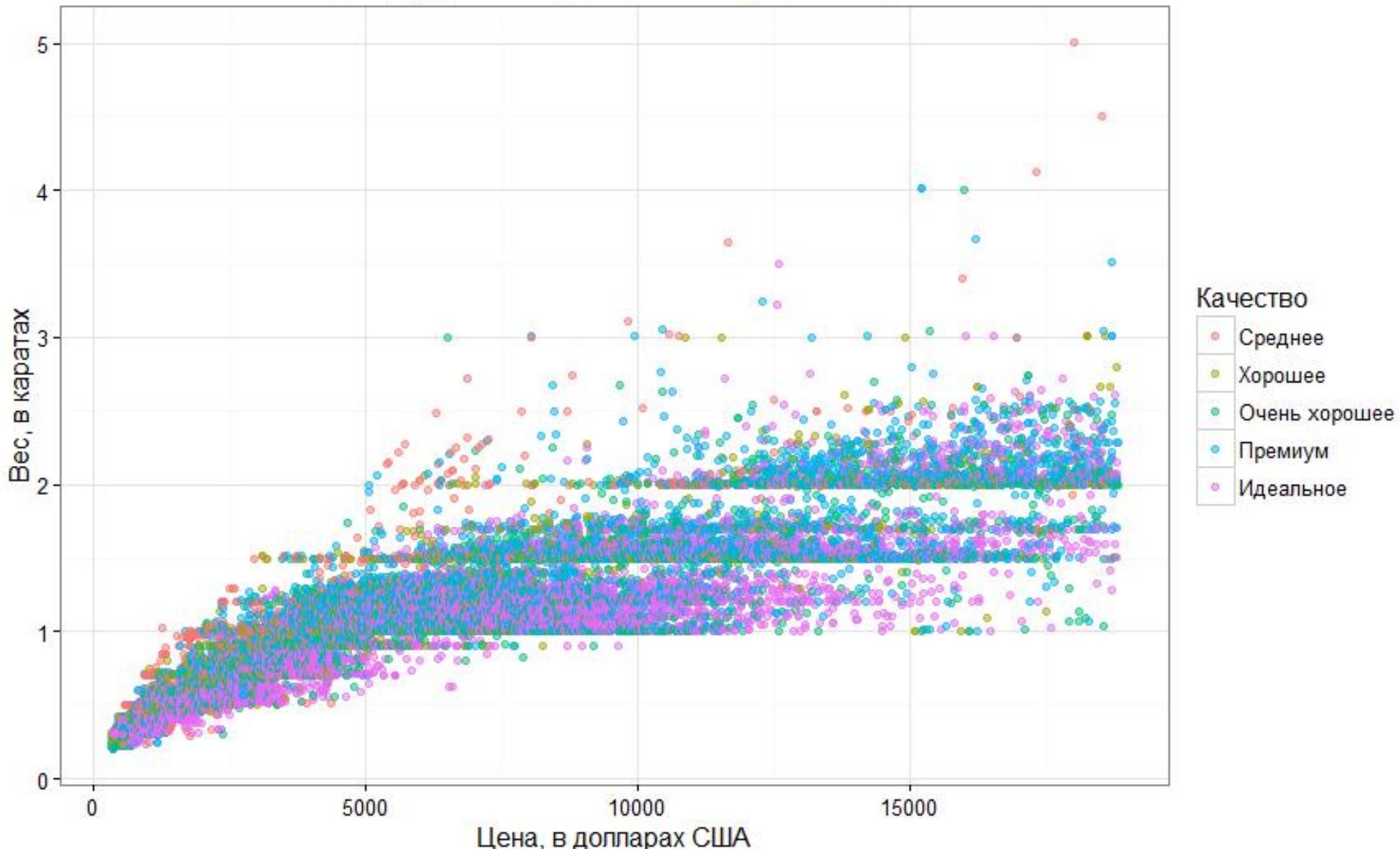
Howard Wainer,

How to Display Data Badly,
The American Statistician , Vol. 38, No. 2 (May, 1984), pp. 137-147.

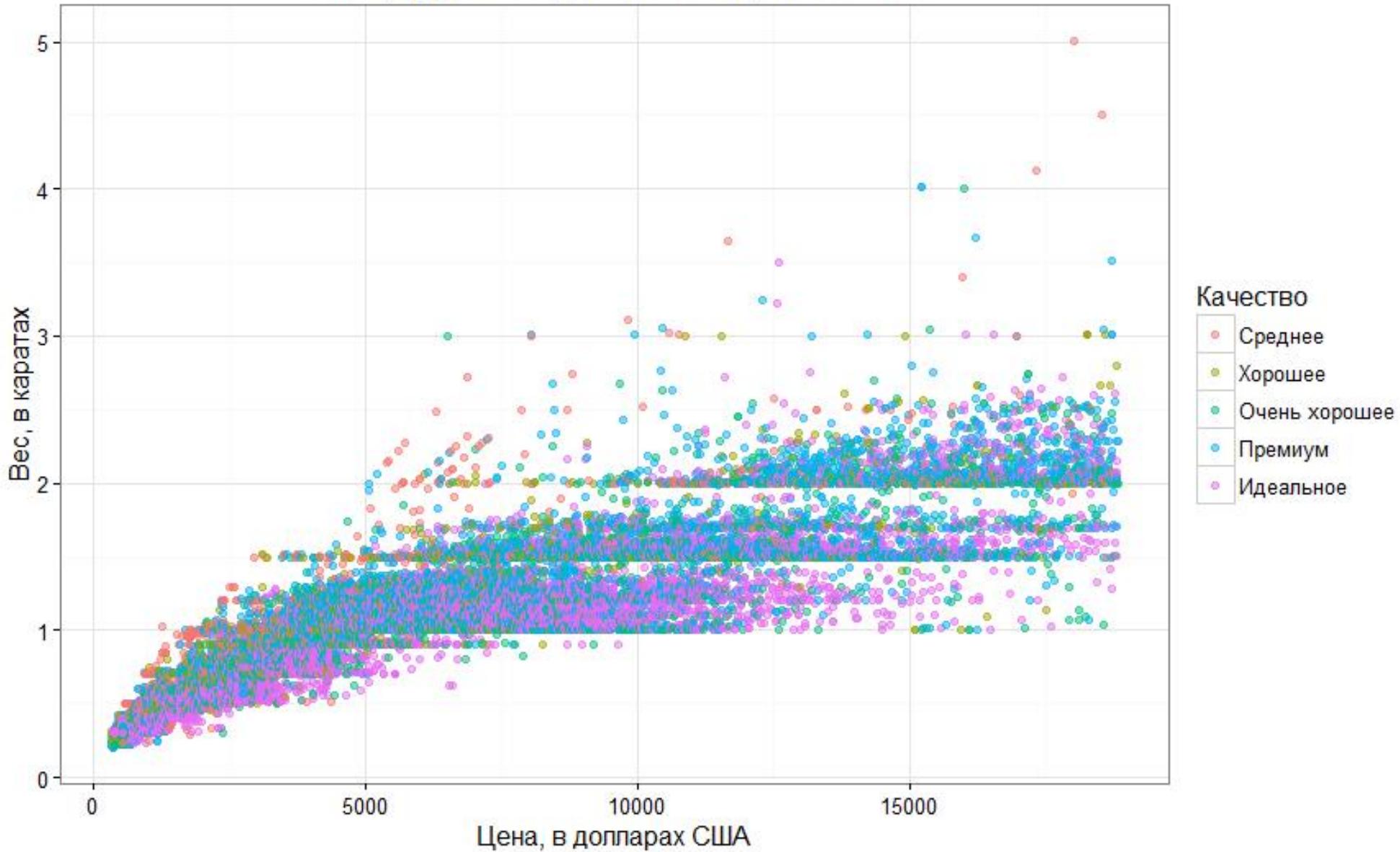
Что такое график?

- Заголовок
- Обычно двумерная плоскость
- Оси: горизонтальная (x) и вертикальная (y)
- Подписи осей (что на осях?)
- Легенда (что значат цвета и значки?)

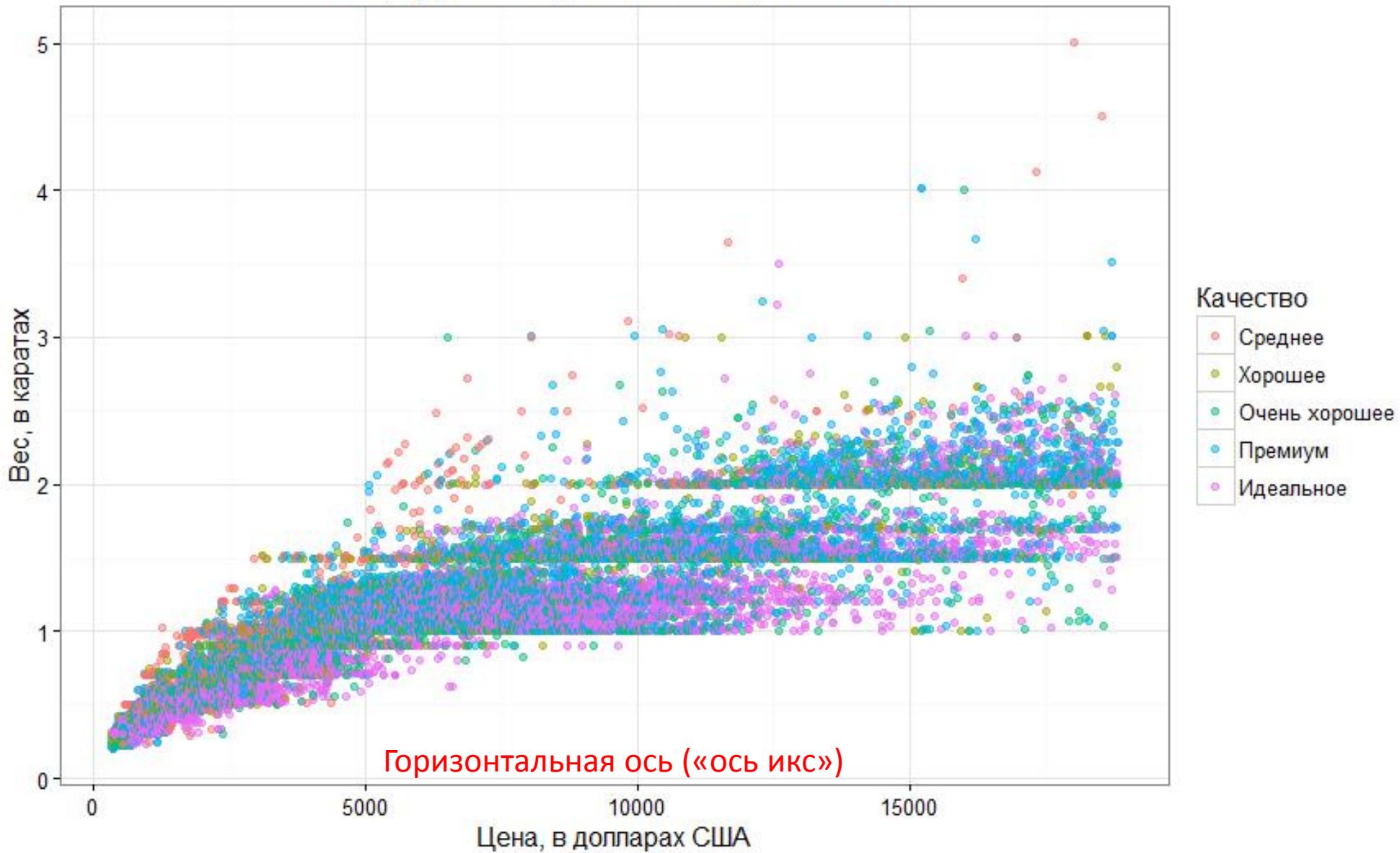
Распределение цены и веса бриллиантов



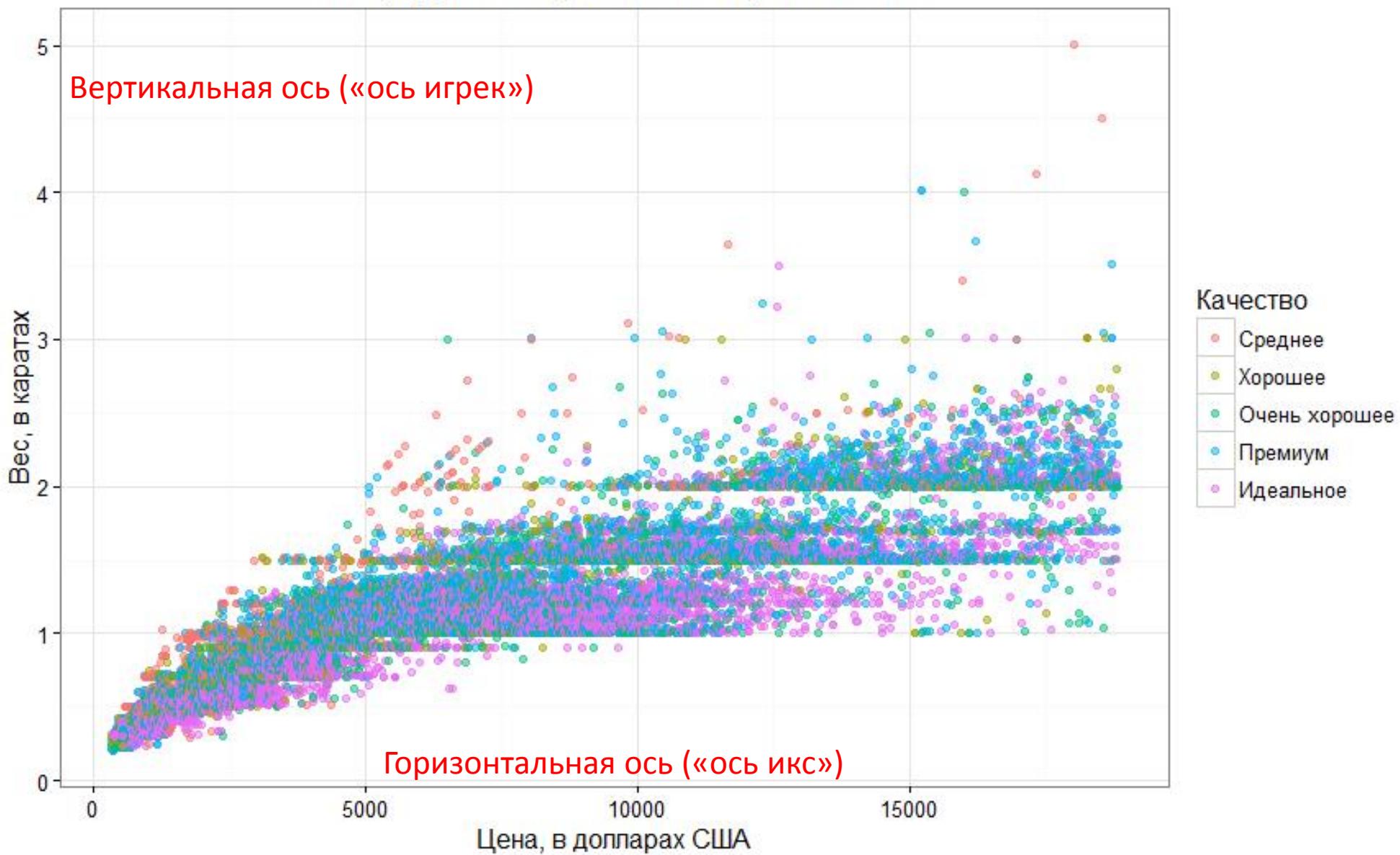
Заголовок
Распределение цены и веса бриллиантов



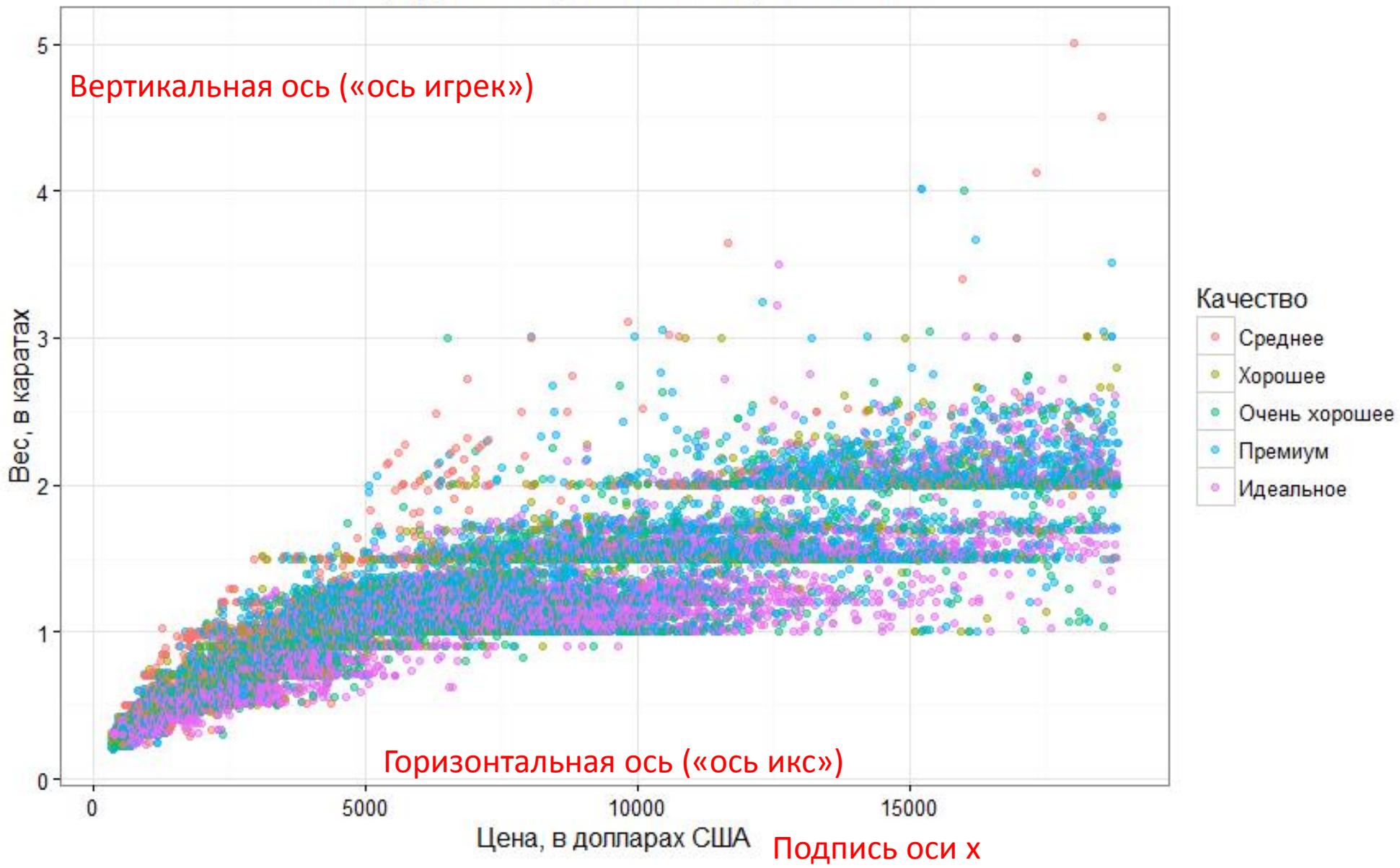
Заголовок
Распределение цены и веса бриллиантов



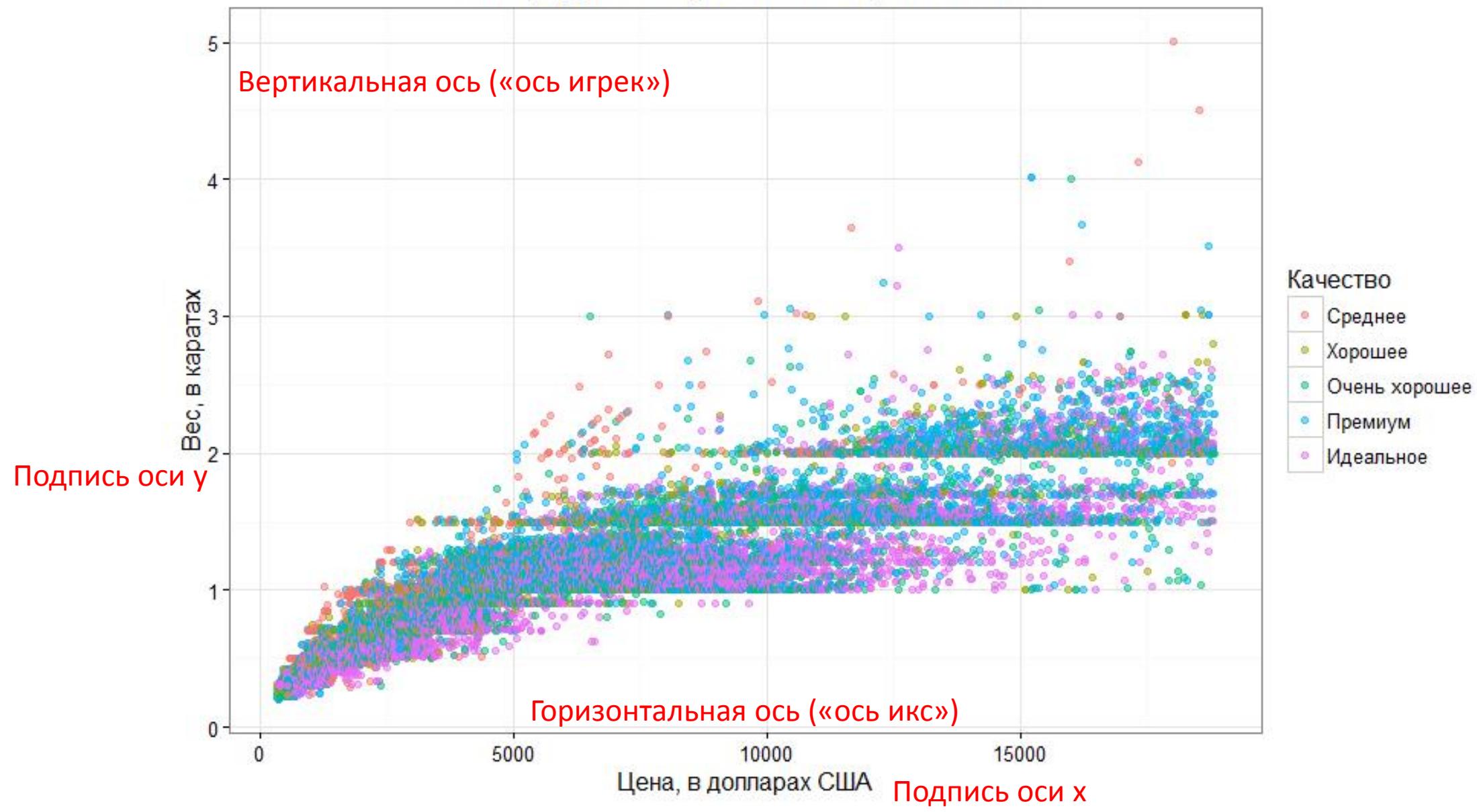
Заголовок
Распределение цены и веса бриллиантов



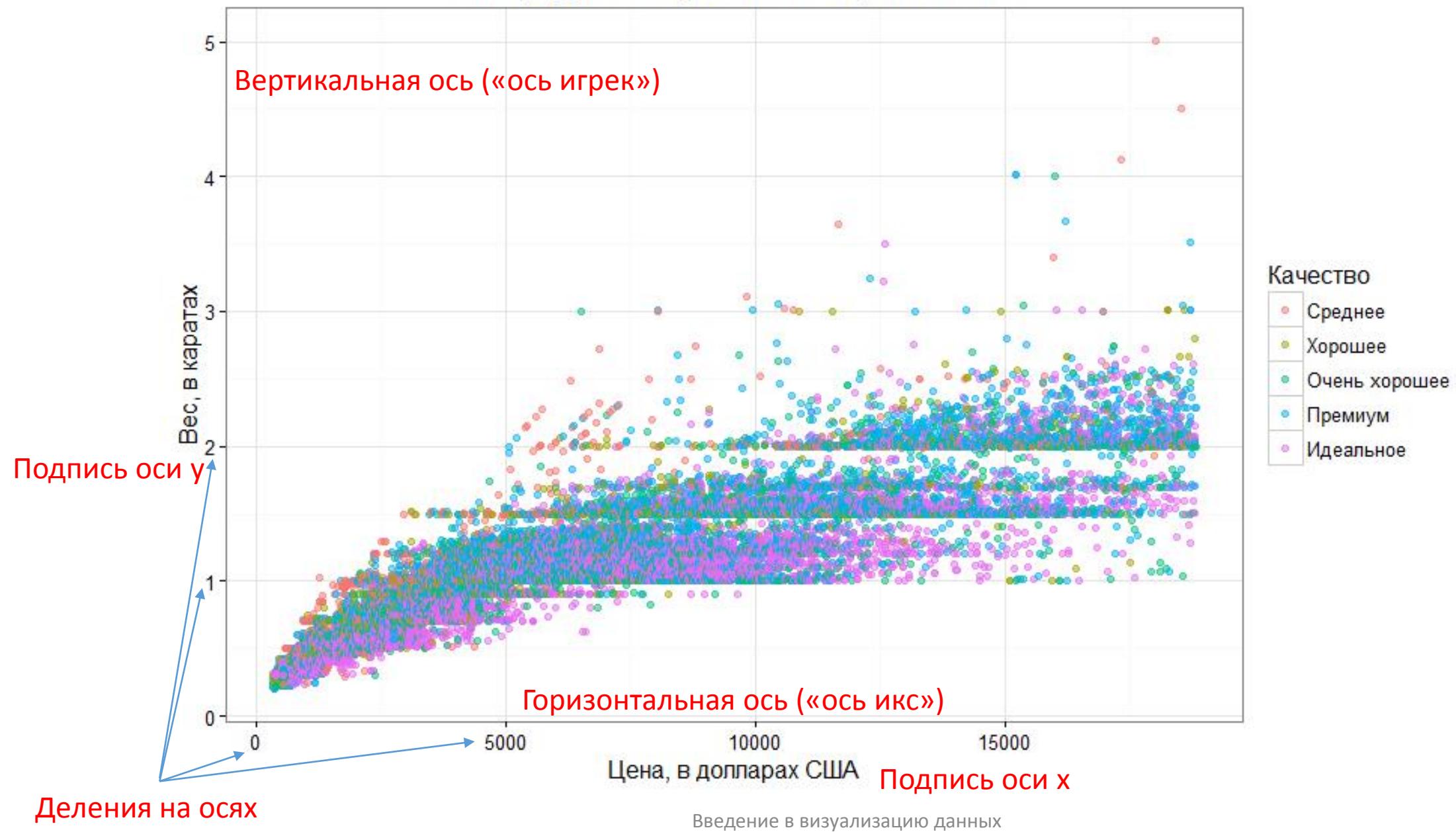
Заголовок
Распределение цены и веса бриллиантов



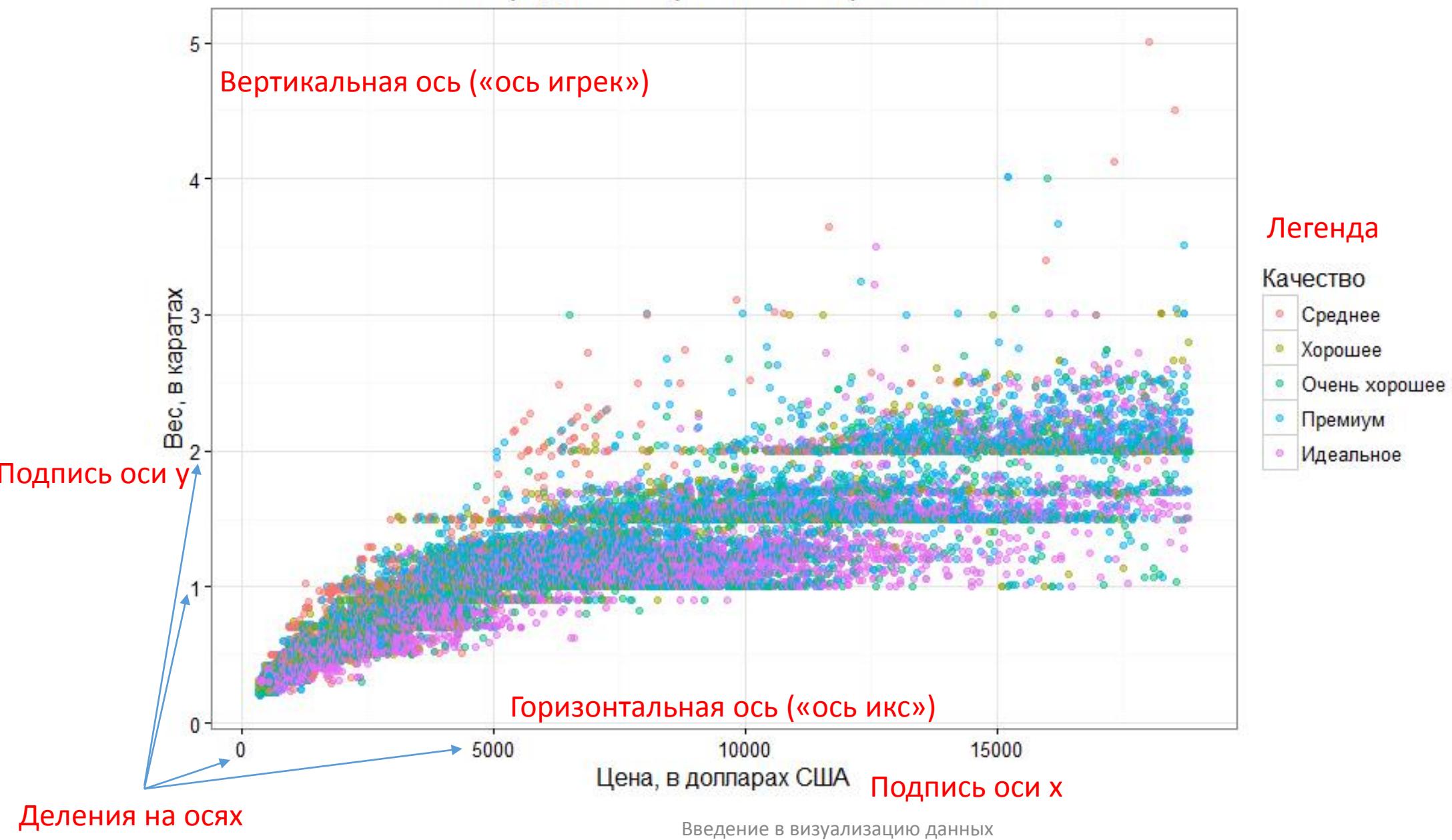
Заголовок
Распределение цены и веса бриллиантов



Заголовок
Распределение цены и веса бриллиантов



Заголовок
Распределение цены и веса бриллиантов



Основные типы графиков

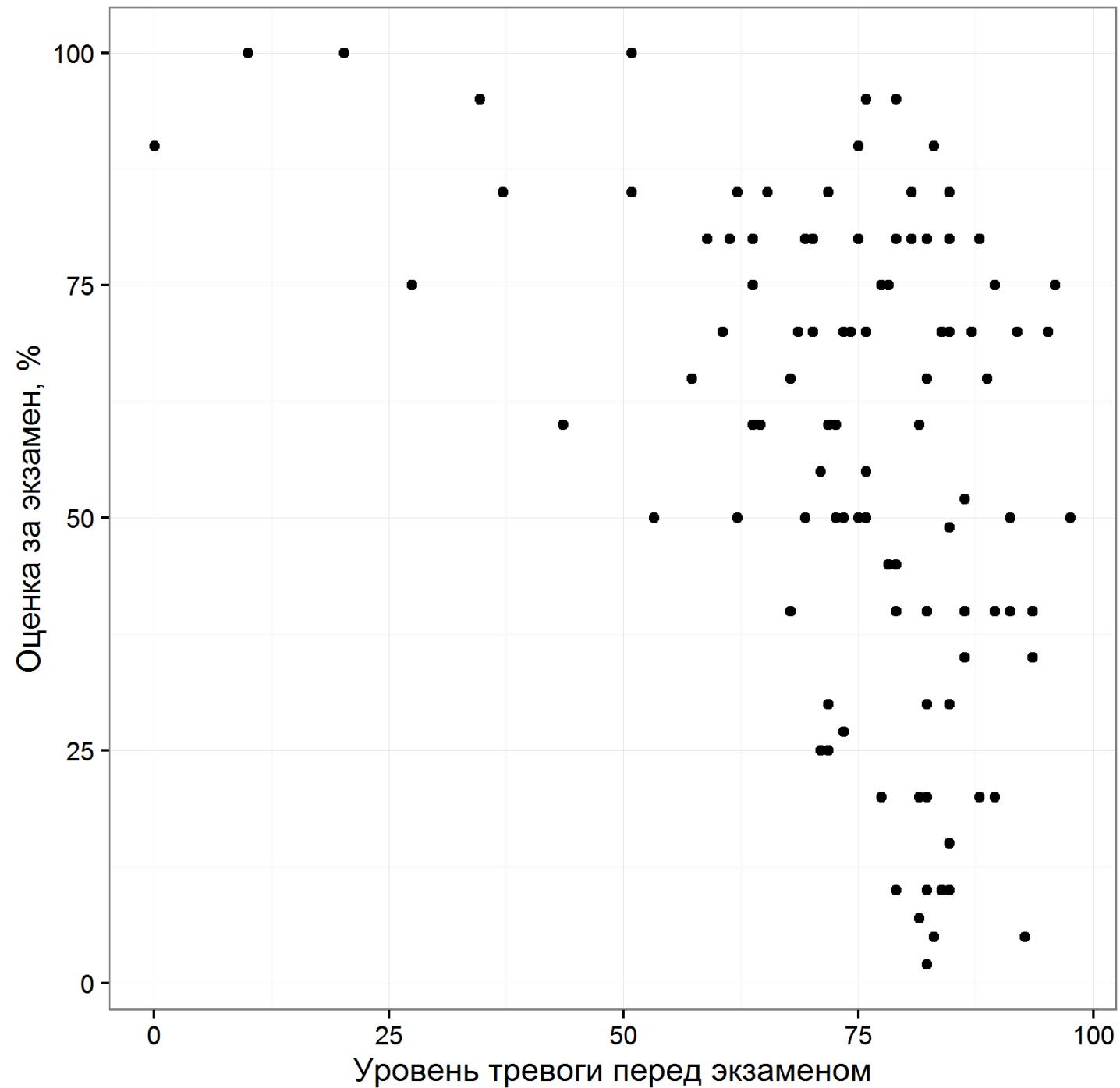
- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

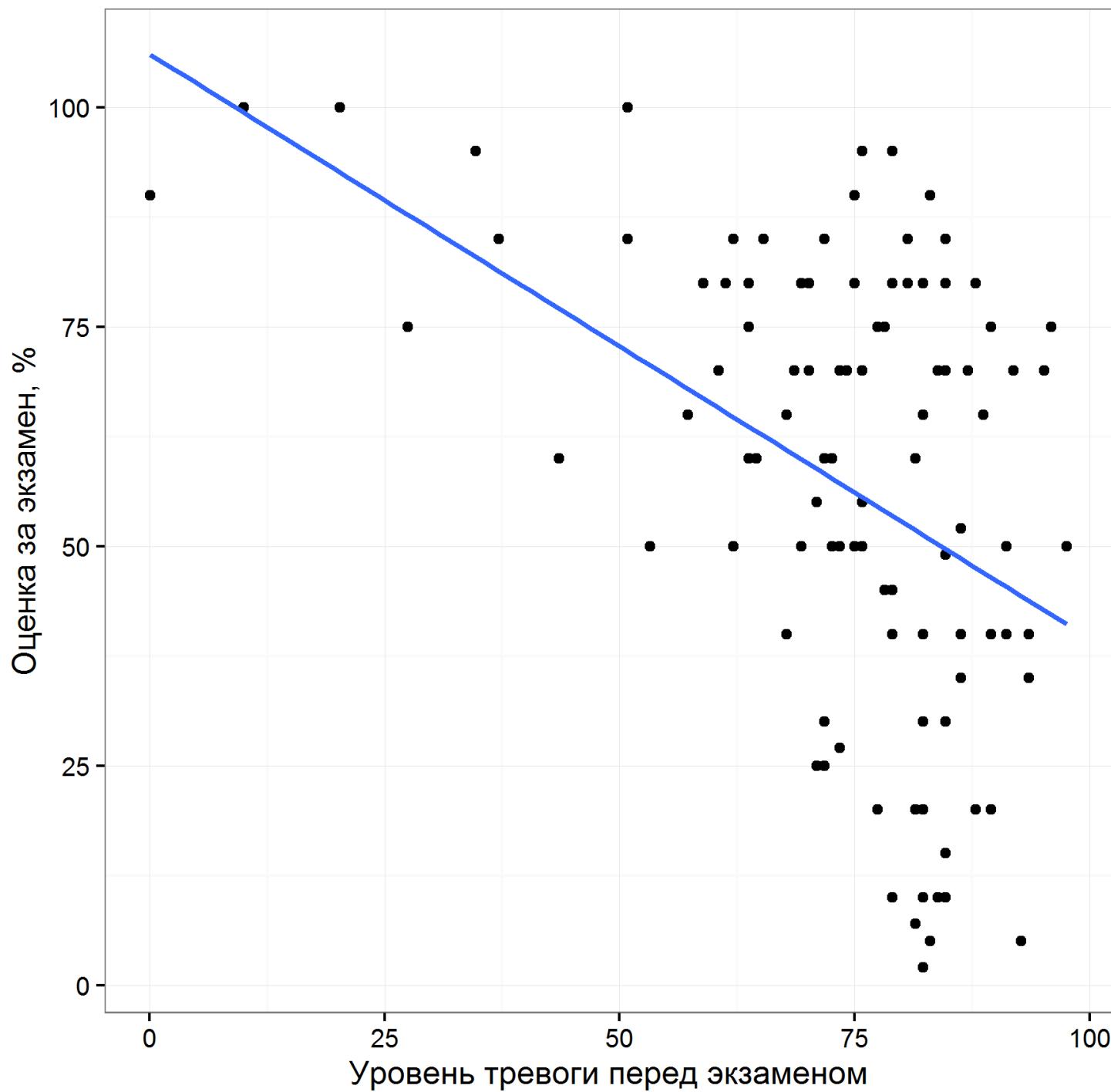
Основные типы графиков

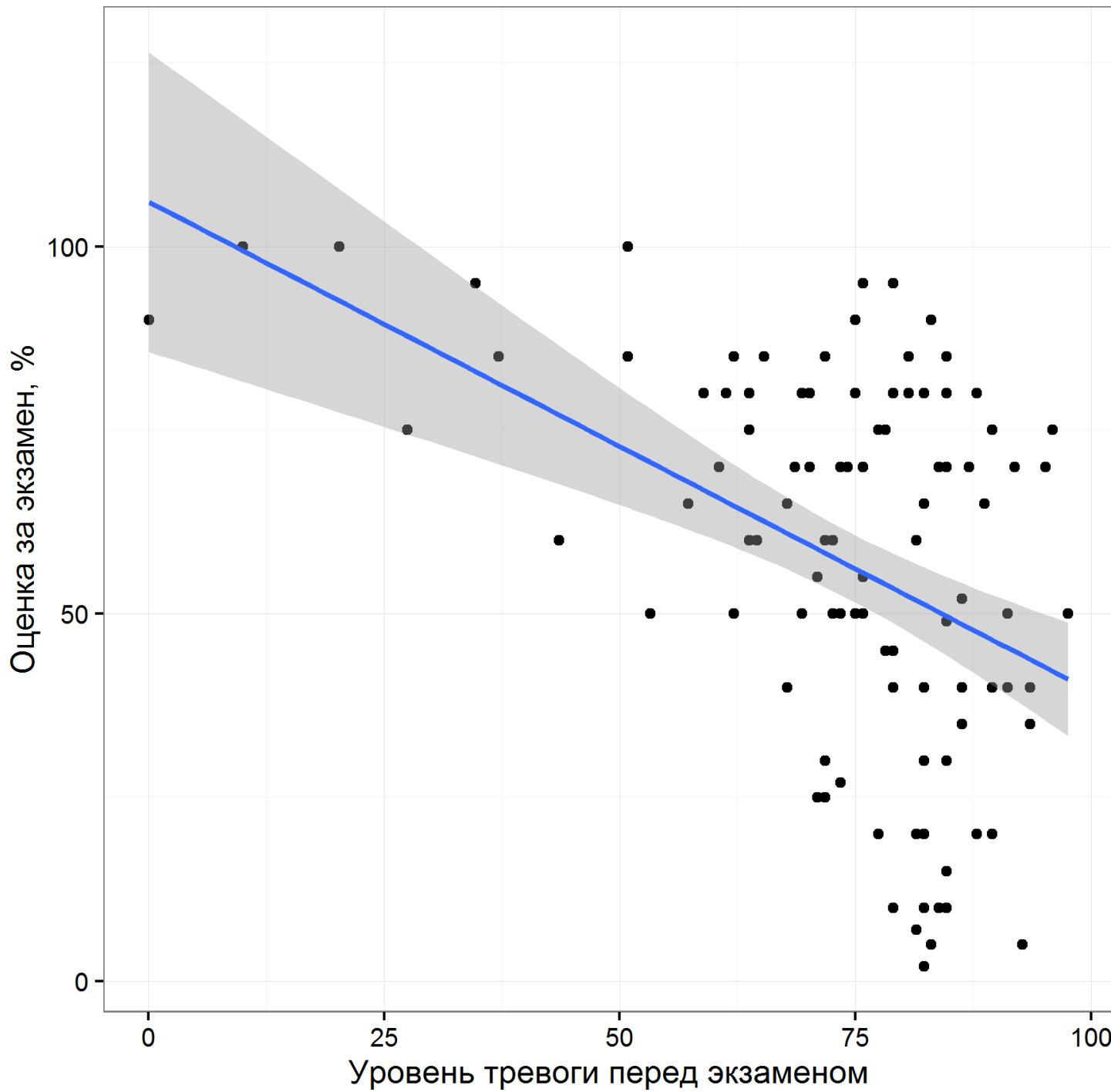
- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

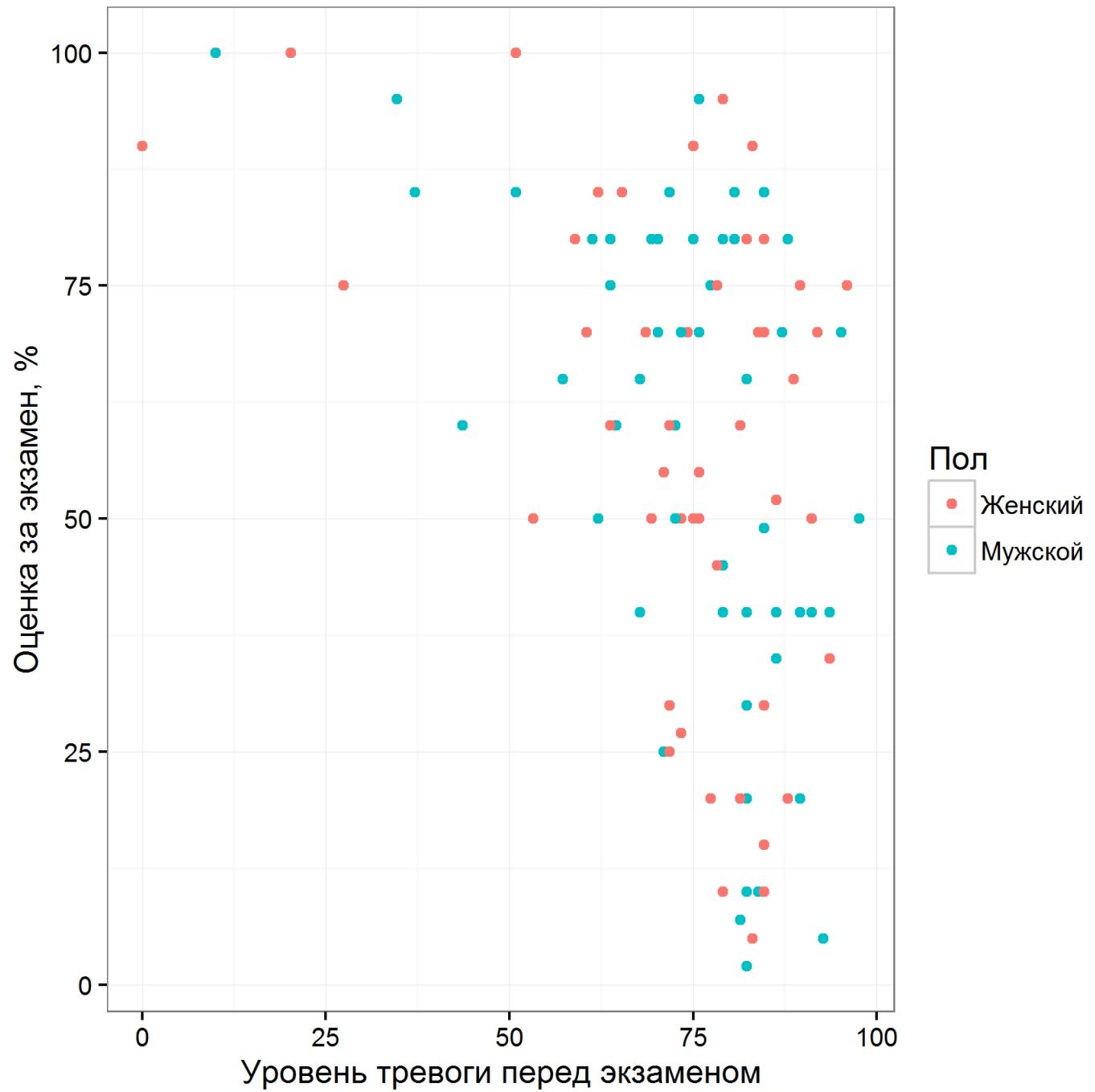
Диаграмма рассеяния: волнение и экзамены

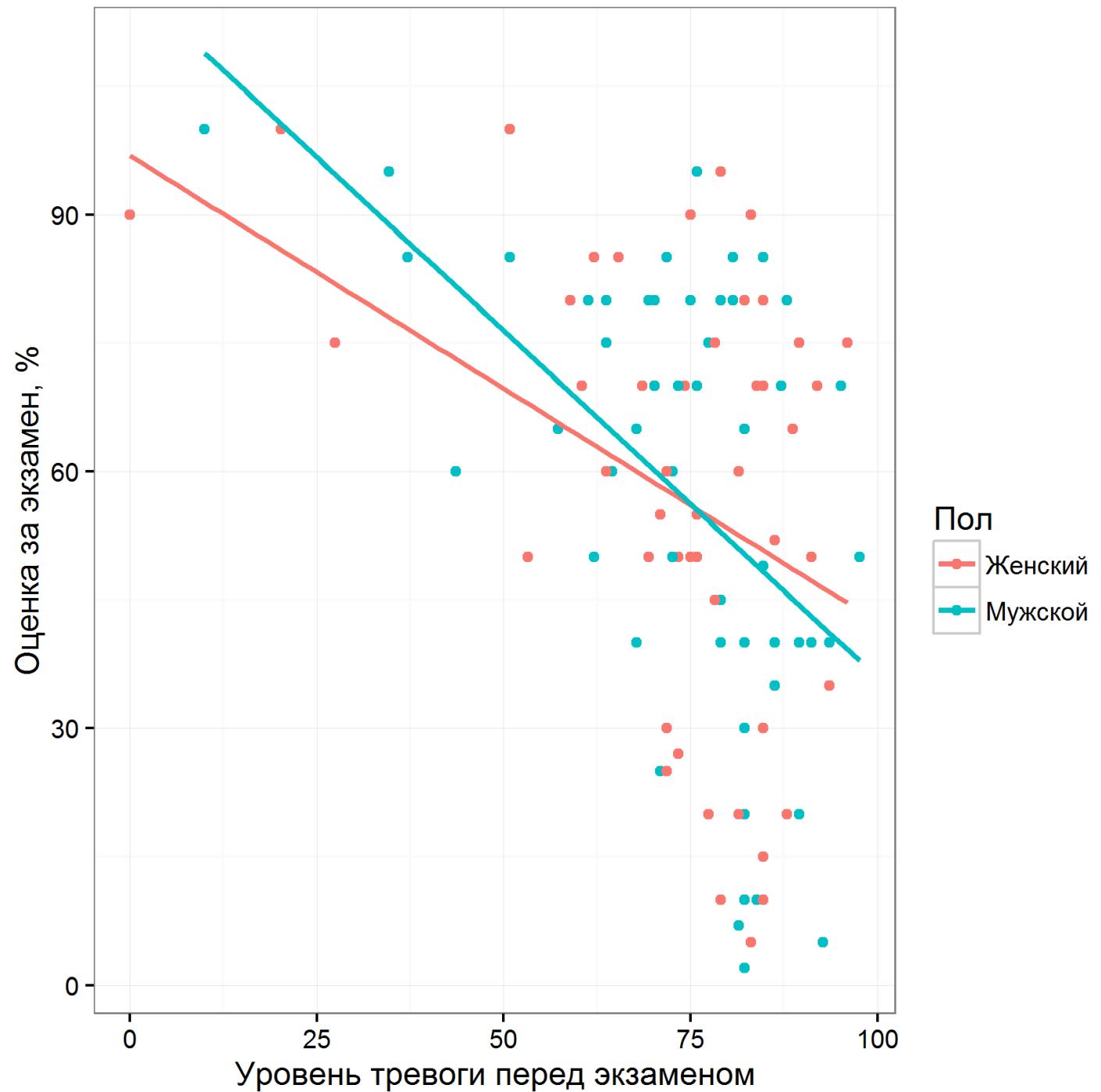
	Code	Revise	Exam	Anxiety	Gender
1	1	4	40	86.298	Male
2	2	11	65	88.716	Female
3	3	27	80	70.178	Male
4	4	53	80	61.312	Male
5	5	4	40	89.522	Male
6	6	22	70	60.506	Female
7	7	16	20	81.462	Female
8	8	21	55	75.820	Female
9	9	25	50	69.372	Female
10	10	18	40	82.268	Female
11	11	18	45	79.044	Male
12	12	16	85	80.656	Male
13	13	13	70	70.178	Male
14	14	18	50	75.014	Female
15	15	98	95	34.714	Male
16	16	1	70	95.164	Male
17	17	14	95	75.820	Male
18	18	29	95	79.044	Female
19	19	4	50	91.134	Female
20	20	23	60	64.536	Male
21	21	14	80	80.656	Male











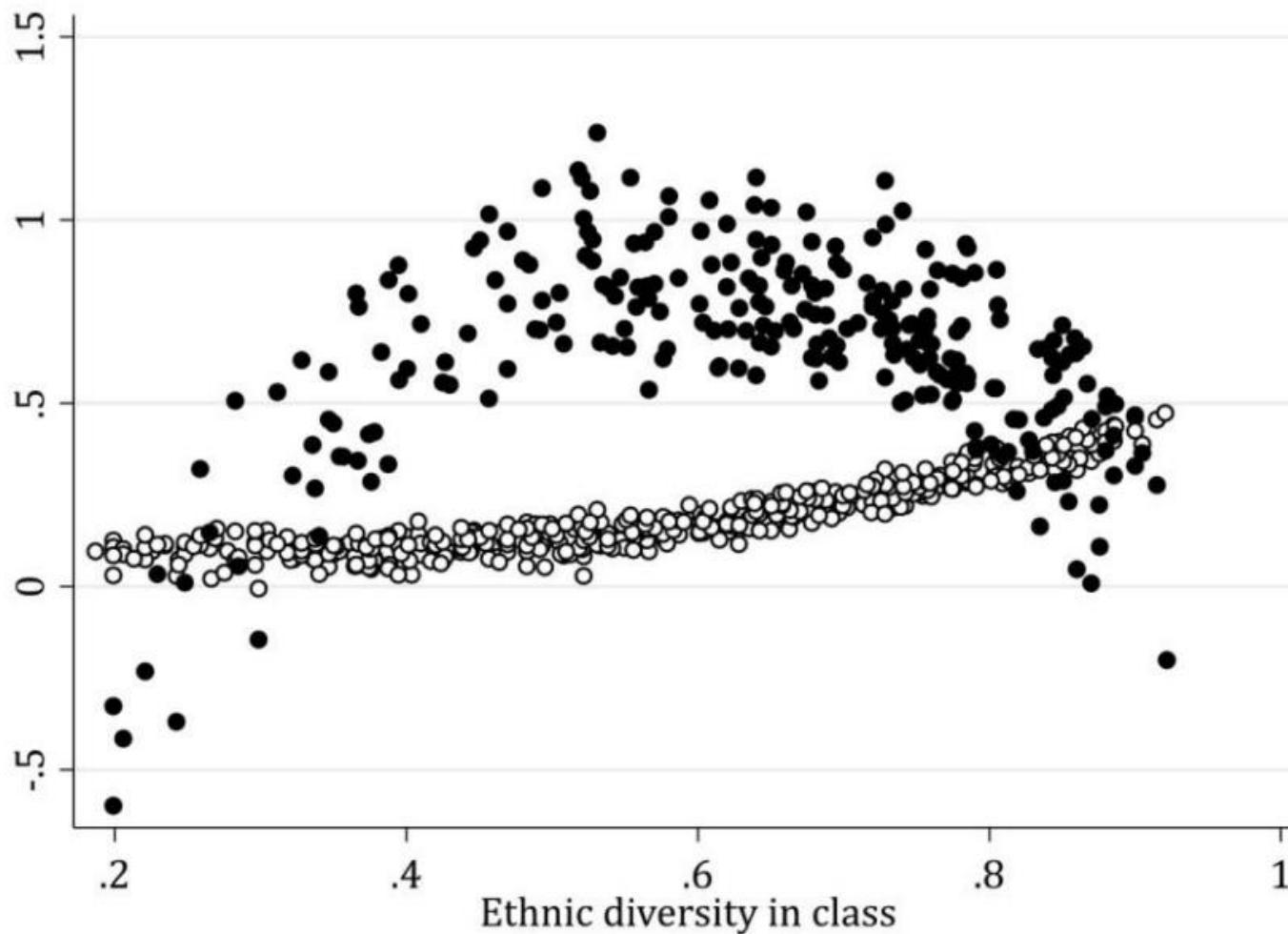


FIG. 6.—Predicted ethnic homophily scores plotted by ethnic diversity in class for native (open circles) and immigrant (filled circles) adolescents based on model 3 (natives) and model 6 (immigrants) in table 3.

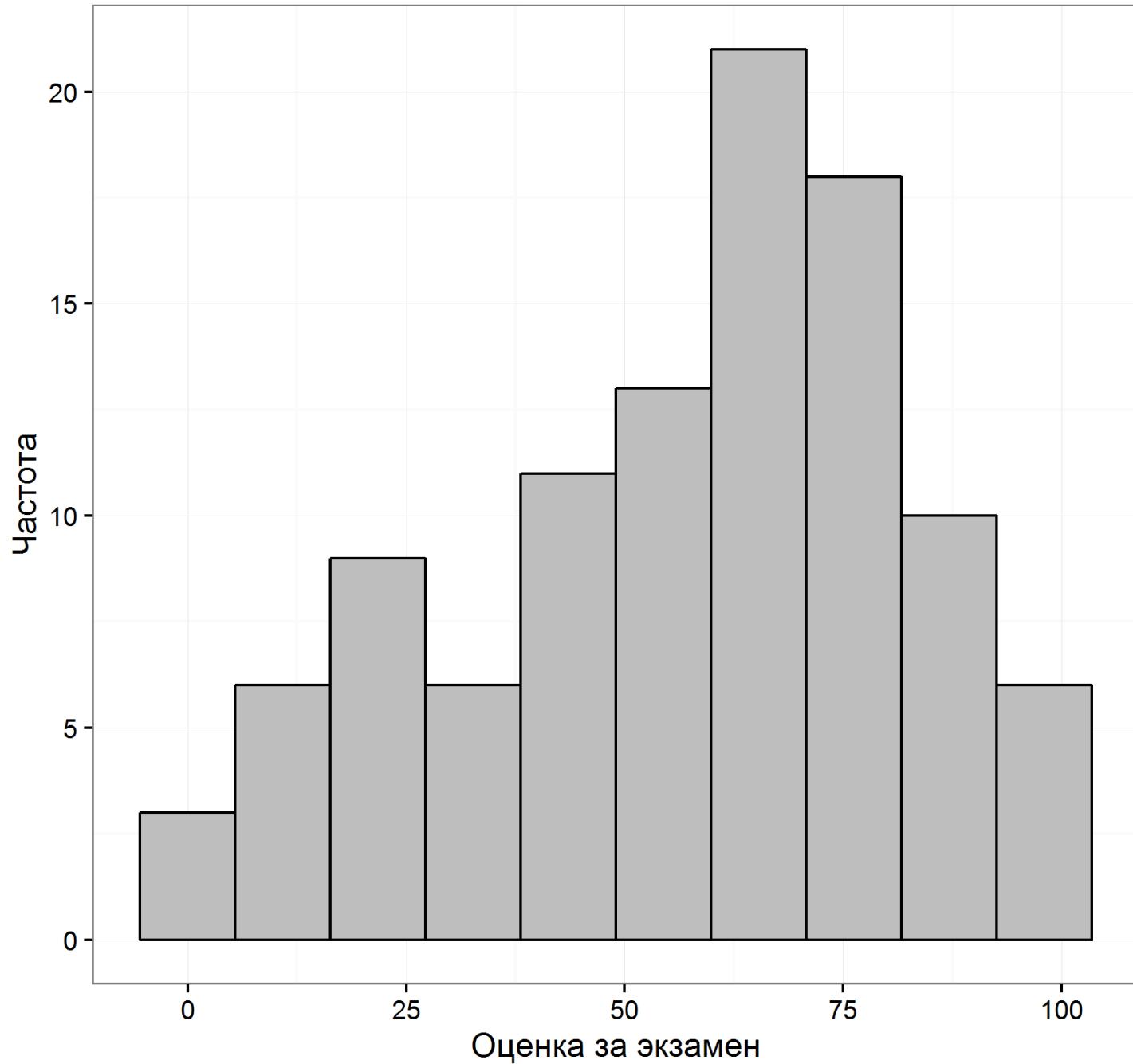
Источник: Sanne Smith, Daniel A. McFarland, Frank Van Tubergen, Ineke Maas. Ethnic Composition and Friendship Segregation: Differential Effects for Adolescent Natives and Immigrants, *American Journal of Sociology*, Volume 121, Number 4 (January 2016), pp. 1223–1272.

Основные типы графиков

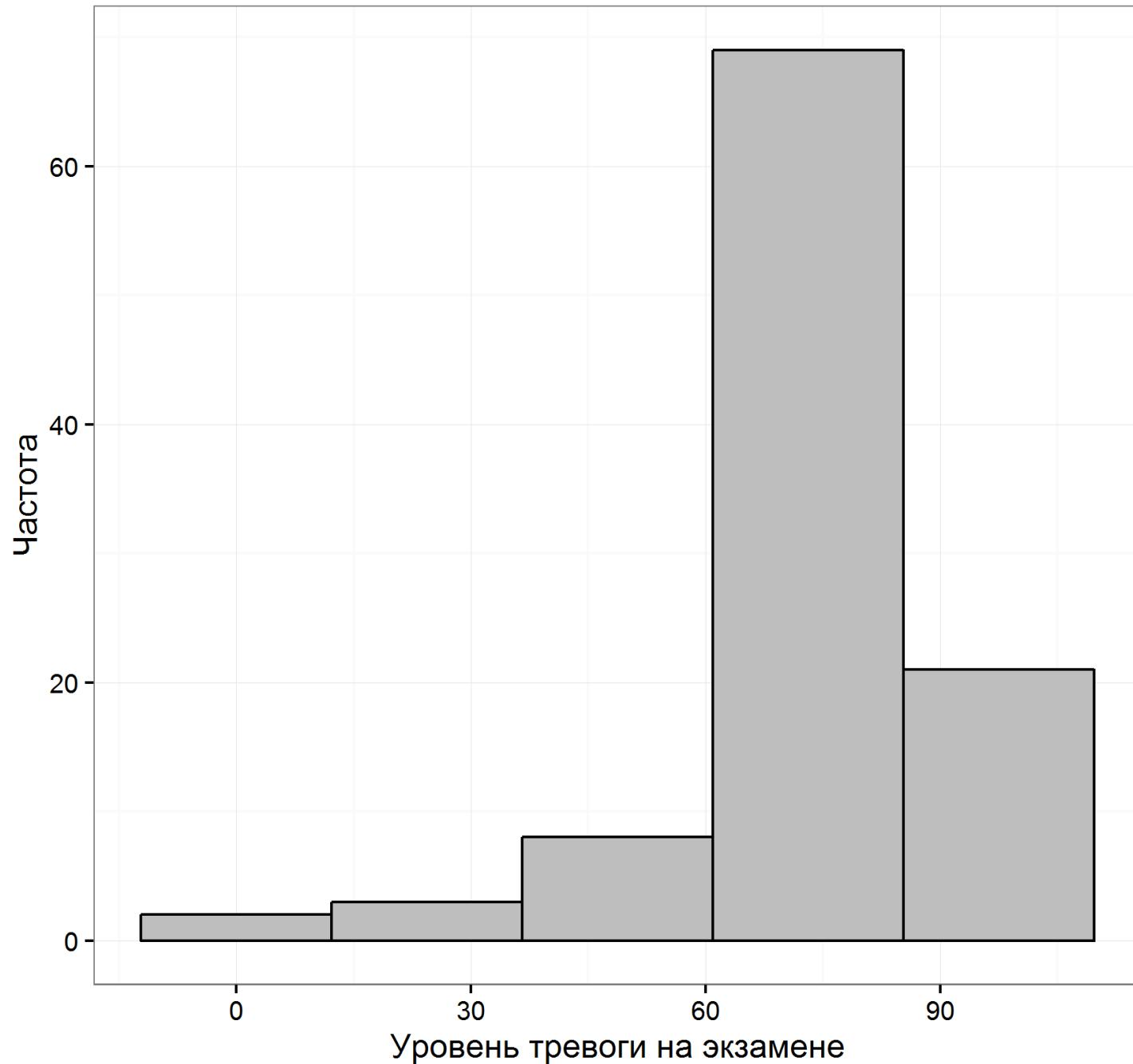
- Диаграмма рассеяния (scatterplot)
- **Диаграмма распределения / гистограмма (histogram)**
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

Гистограмма (histogram, диаграмма распределения)

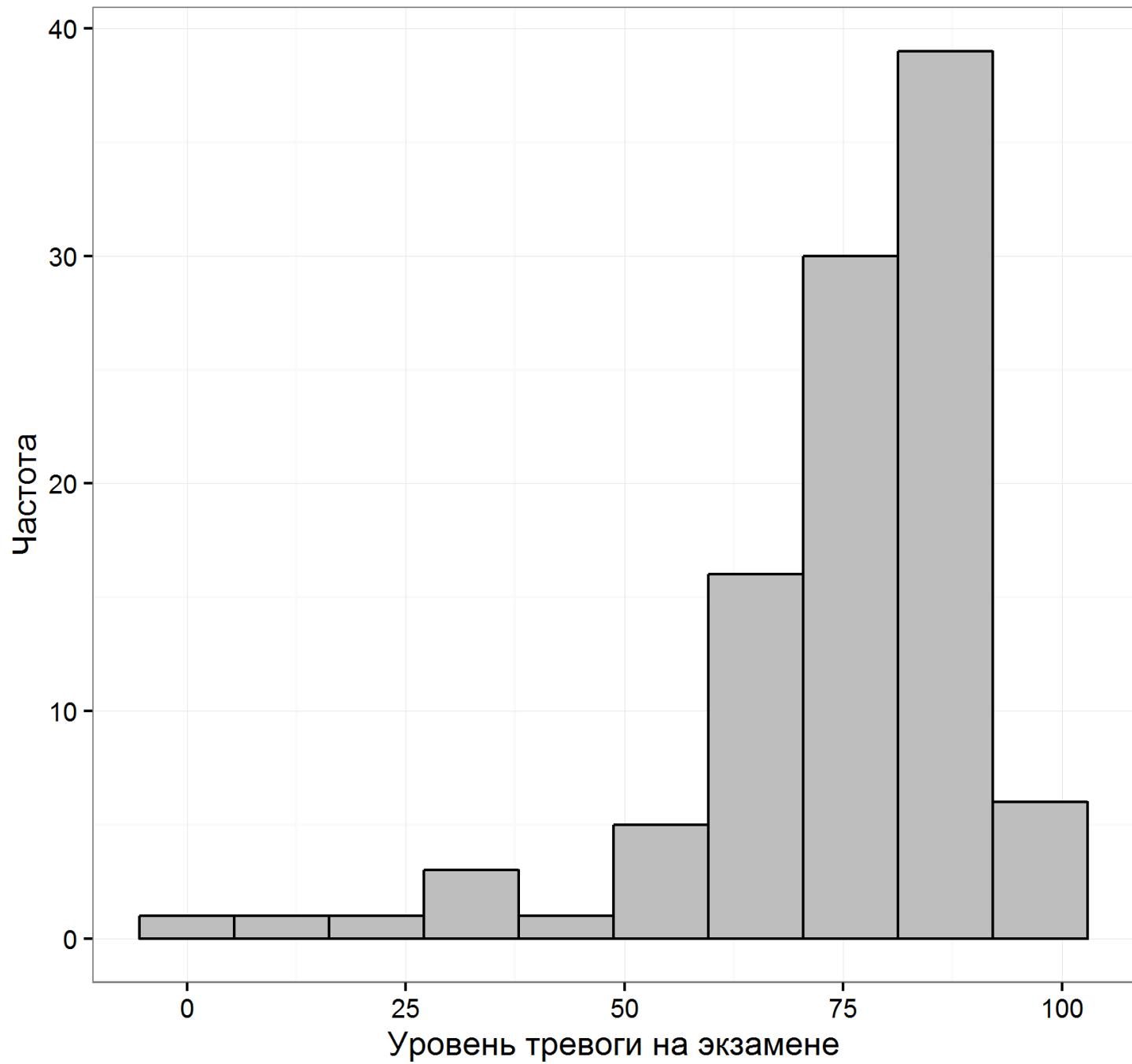
Распределение оценок за экзамен



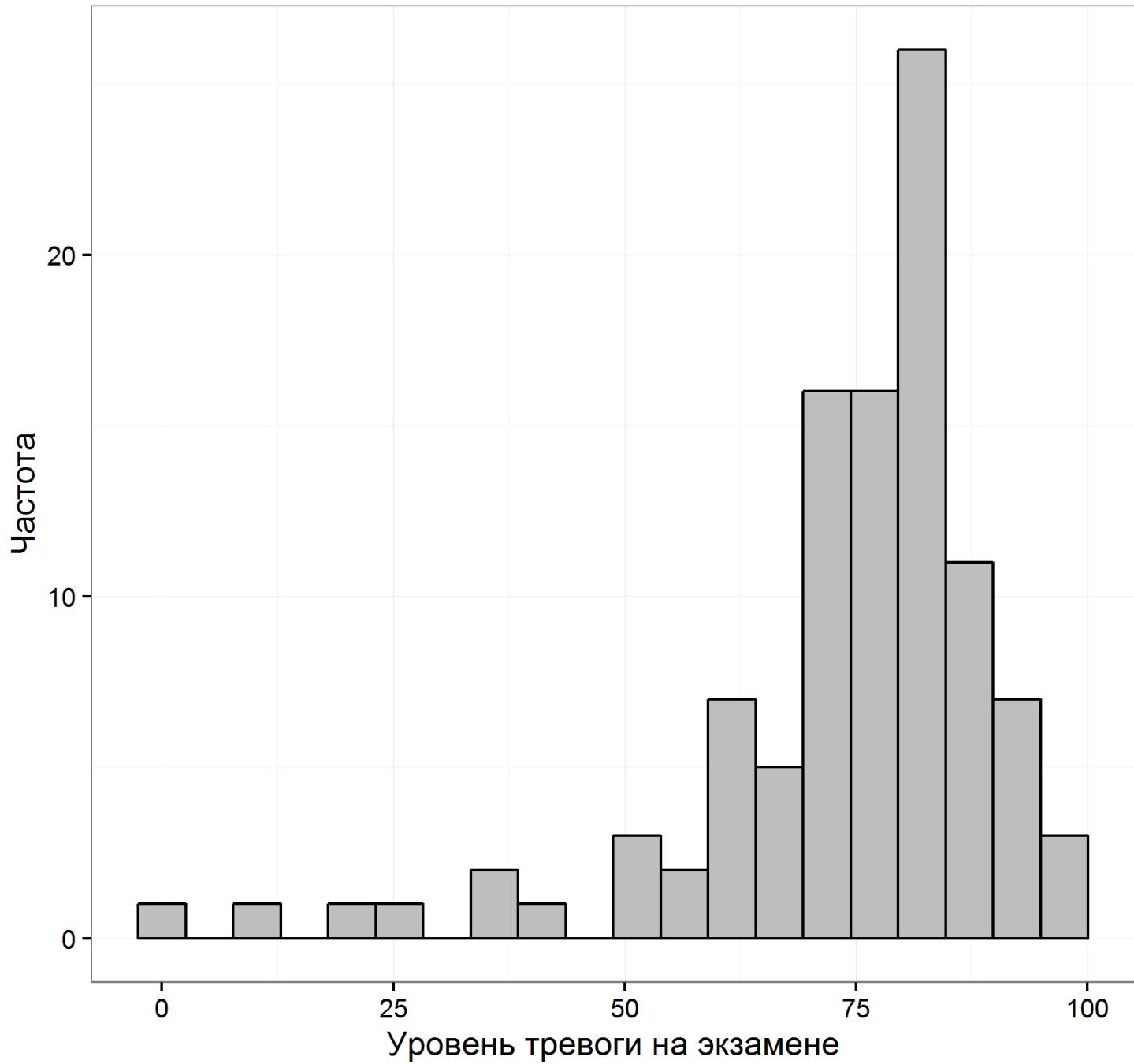
Гистограмма (5 столбиков)



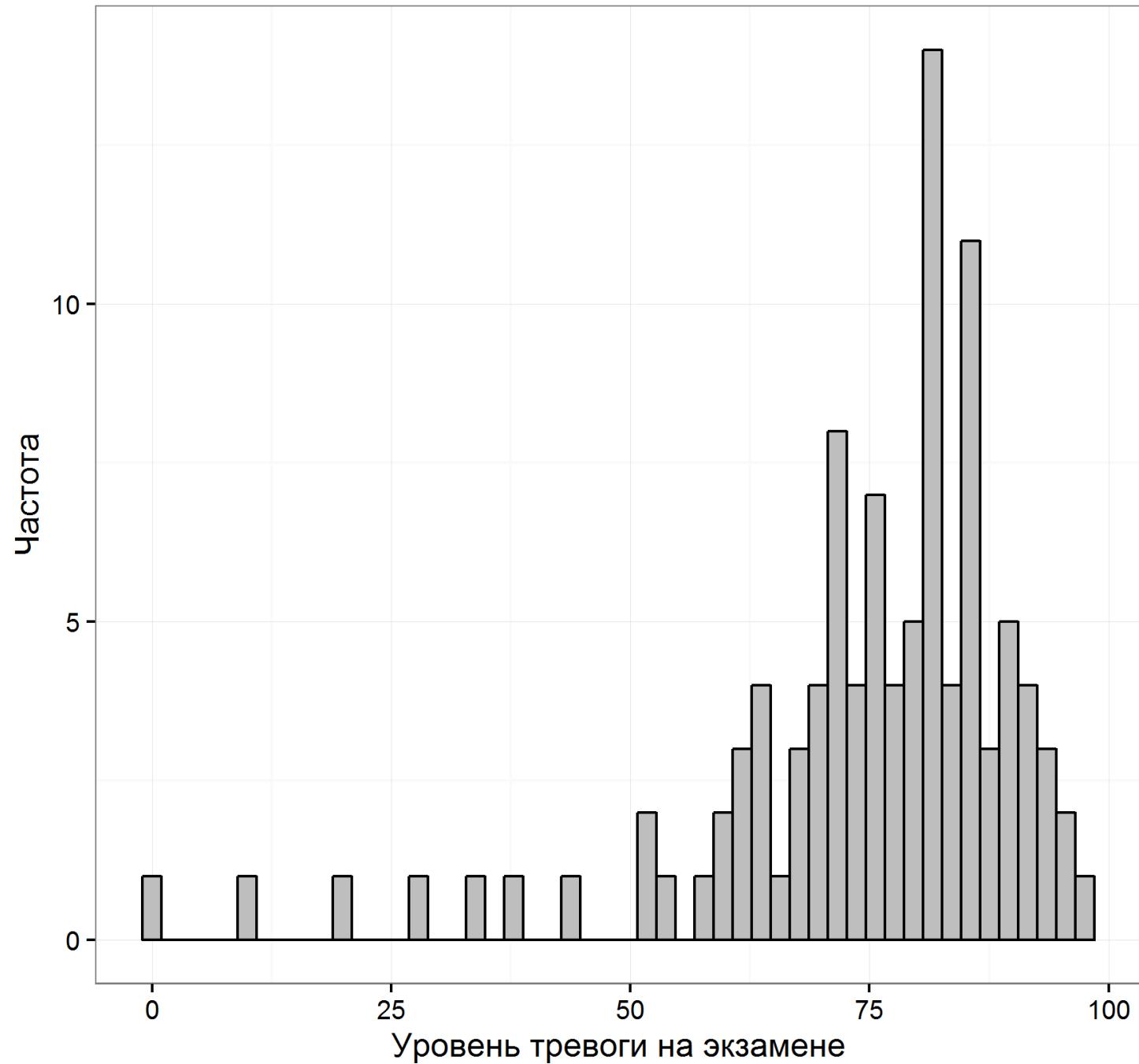
Гистограмма (10 столбиков)



Гистограмма (20 столбиков)



Гистограмма (50 столбиков)



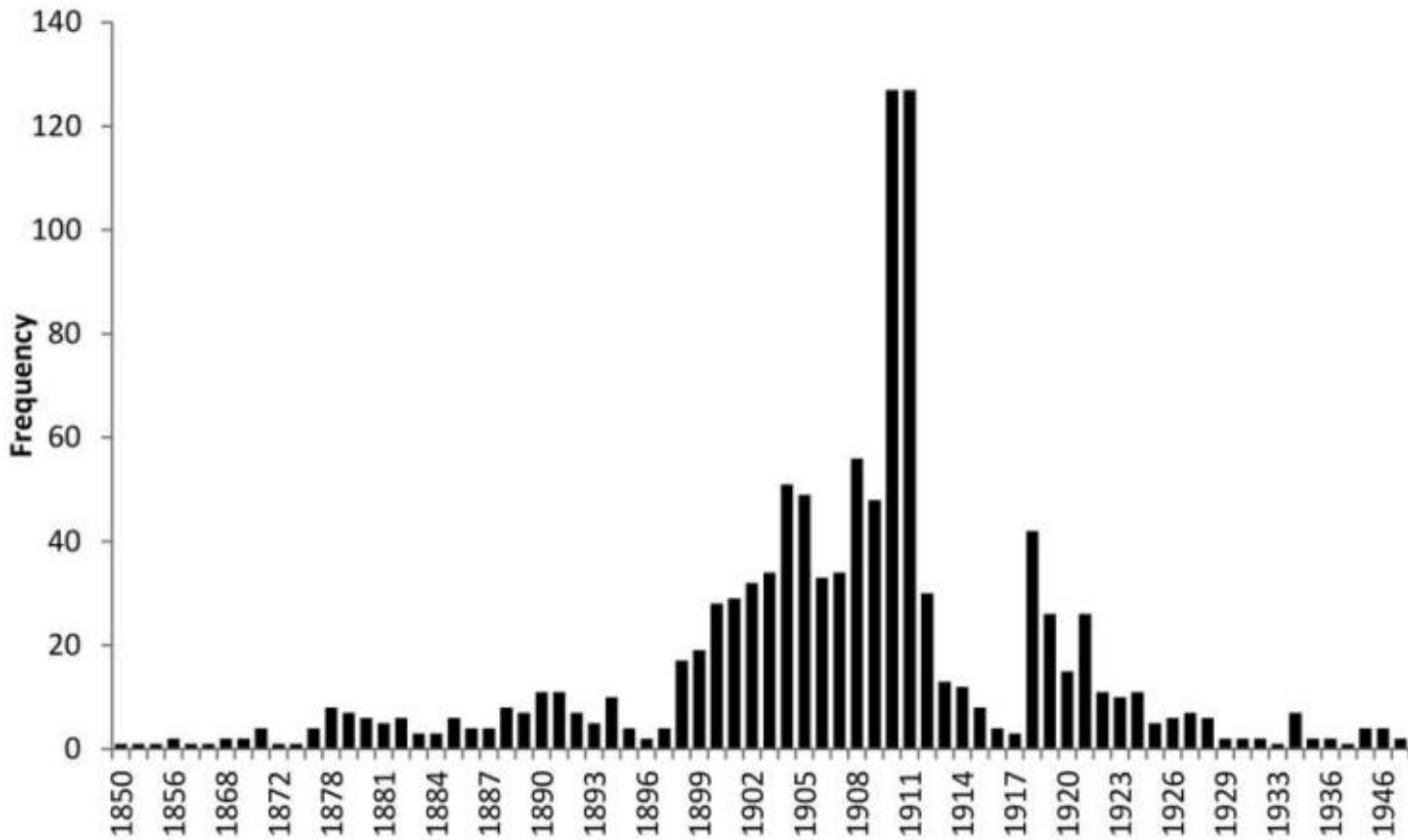


FIG. 4.—Year that an immigrant came to the United States who worked in A.M. Byers Company. Data come from the A.M. Byers files (IPCSR 6359). Frequencies broken down by ethnicity are available upon request.

Источник: Catron Peter. Made in America? Immigrant Occupational Mobility in the First Half of the Twentieth Century.
American Journal of Sociology, Volume 122, Number 2 (September 2016), p. 342.

Гистограмма и фестиваль музыки

Download Festival



Download Festival



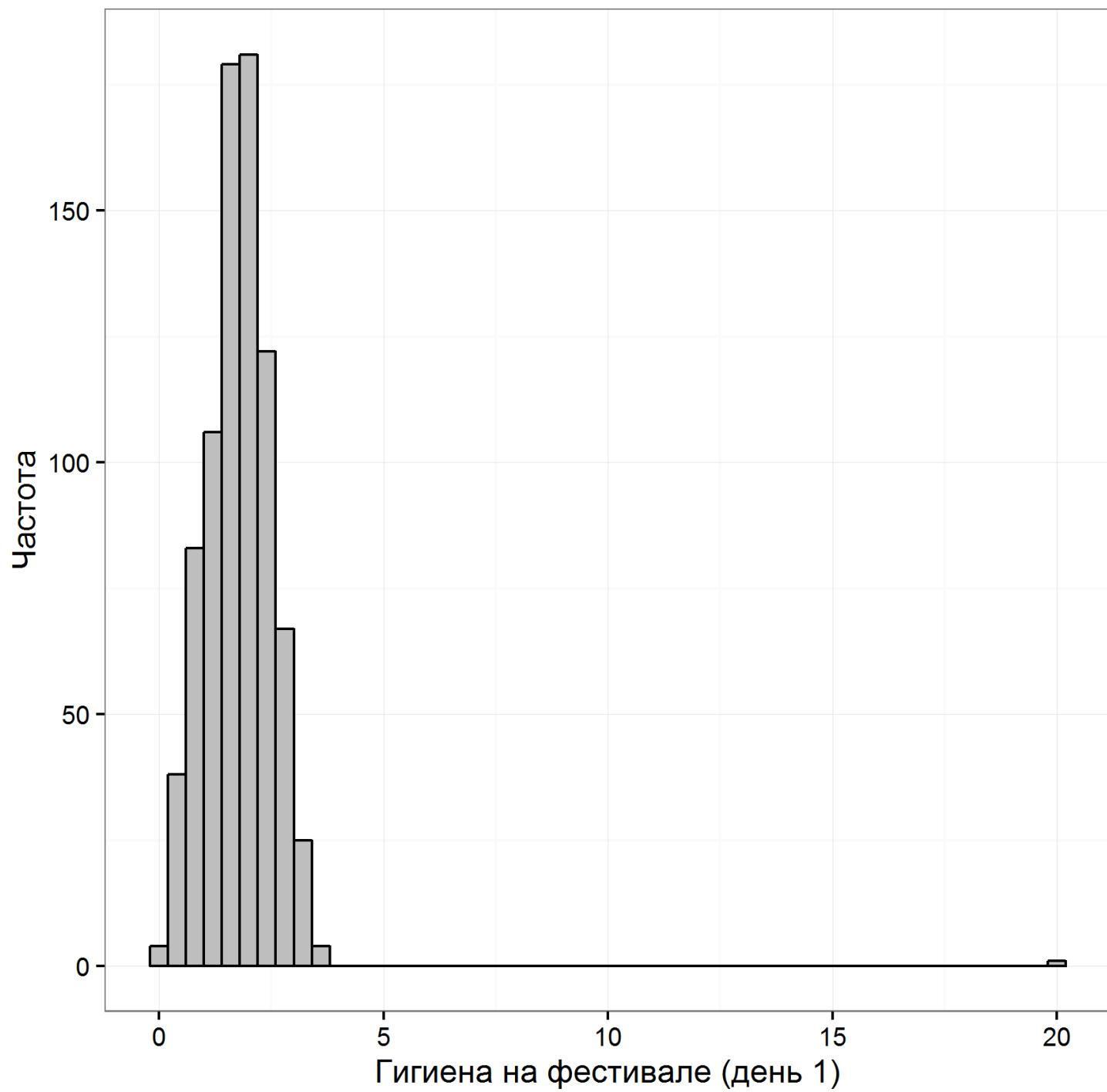
Download Festival

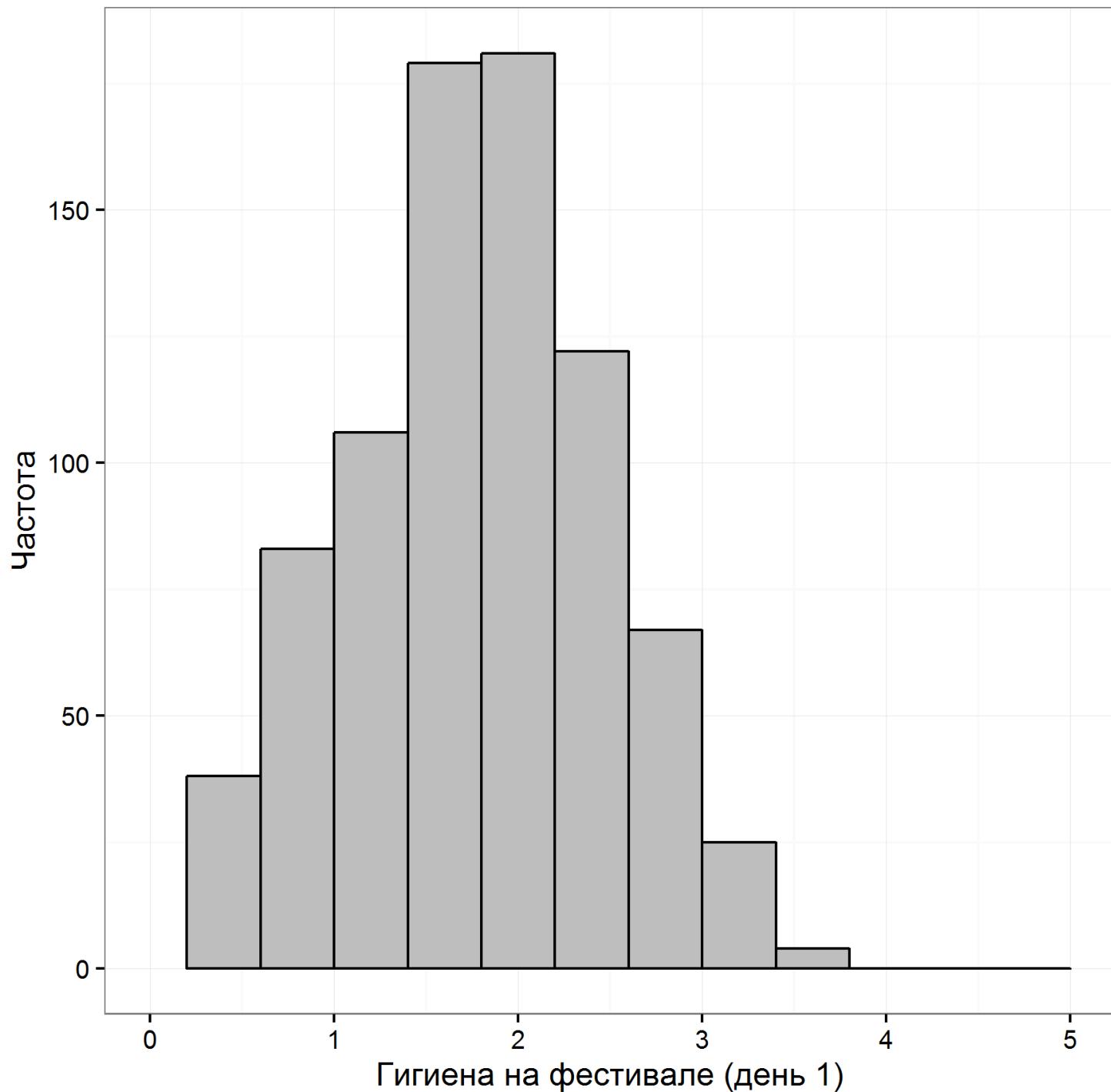


Данные: фестиваль музыки

	ticknumb	gender	day1	day2	day3
1	2111	Male	2.64	1.35	1.61
2	2229	Female	0.97	1.41	0.29
3	2338	Male	0.84	NA	NA
4	2384	Female	3.03	NA	NA
5	2401	Female	0.88	0.08	NA
6	2405	Male	0.85	NA	NA
7	2467	Female	1.56	NA	NA
8	2478	Female	3.02	NA	NA
9	2490	Male	2.29	NA	NA
10	2504	Female	1.11	0.44	0.55
11	2509	Male	2.17	NA	NA
12	2510	Female	0.82	0.20	0.47
13	2514	Male	1.41	NA	NA
14	2515	Female	1.76	1.64	1.58
15	2520	Male	1.38	0.02	NA
16	2521	Female	2.79	NA	NA
17	2529	Male	1.50	NA	NA
18	2533	Female	1.91	2.05	NA
19	2535	Female	2.32	NA	NA
20	2538	Male	2.05	NA	NA
21	2549	Male	2.17	0.70	0.76

Showing 1 to 21 of 809 entries





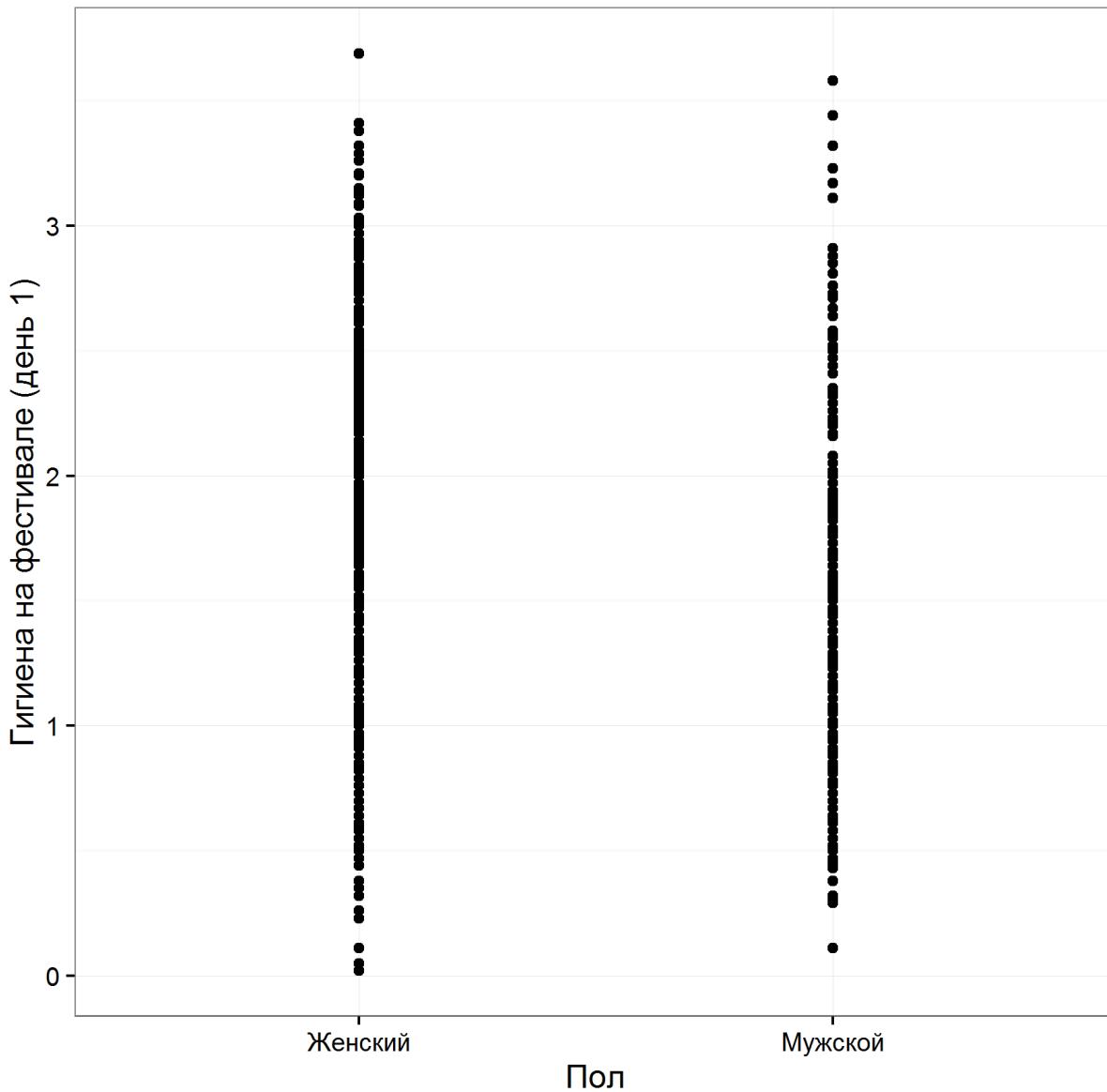
Кто грязнее на фестивале – мужчины или женщины?

Данные: фестиваль музыки

	ticknumb	gender	day1	day2	day3
1	2111	Male	2.64	1.35	1.61
2	2229	Female	0.97	1.41	0.29
3	2338	Male	0.84	NA	NA
4	2384	Female	3.03	NA	NA
5	2401	Female	0.88	0.08	NA
6	2405	Male	0.85	NA	NA
7	2467	Female	1.56	NA	NA
8	2478	Female	3.02	NA	NA
9	2490	Male	2.29	NA	NA
10	2504	Female	1.11	0.44	0.55
11	2509	Male	2.17	NA	NA
12	2510	Female	0.82	0.20	0.47
13	2514	Male	1.41	NA	NA
14	2515	Female	1.76	1.64	1.58
15	2520	Male	1.38	0.02	NA
16	2521	Female	2.79	NA	NA
17	2529	Male	1.50	NA	NA
18	2533	Female	1.91	2.05	NA
19	2535	Female	2.32	NA	NA
20	2538	Male	2.05	NA	NA
21	2549	Male	2.17	0.70	0.76

Showing 1 to 21 of 809 entries

Диаграмма рассеяния не работает:



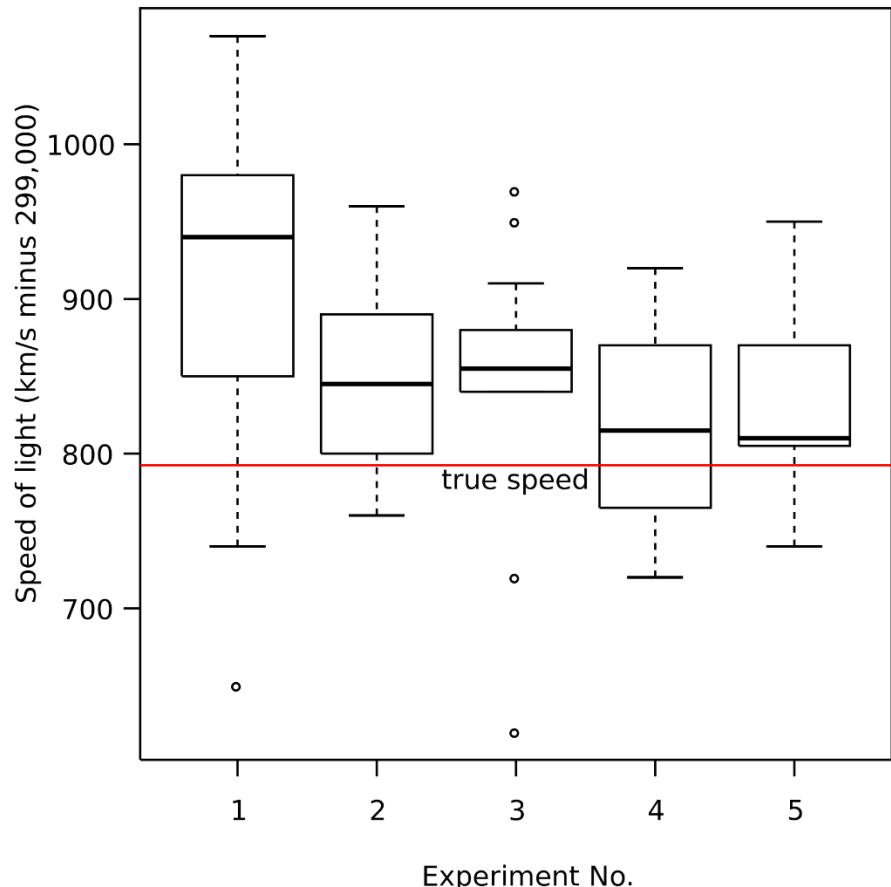
Кто грязнее на фестивале – мужчины или женщины?

Нужно сравнить среднюю нечистоплотность мужчин и женщин и понять, кто в среднем грязнее

Основные типы графиков

- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

«Ящик с усами» / ящиковая диаграмма / диаграмма размаха (boxplot, box-and-whiskers)



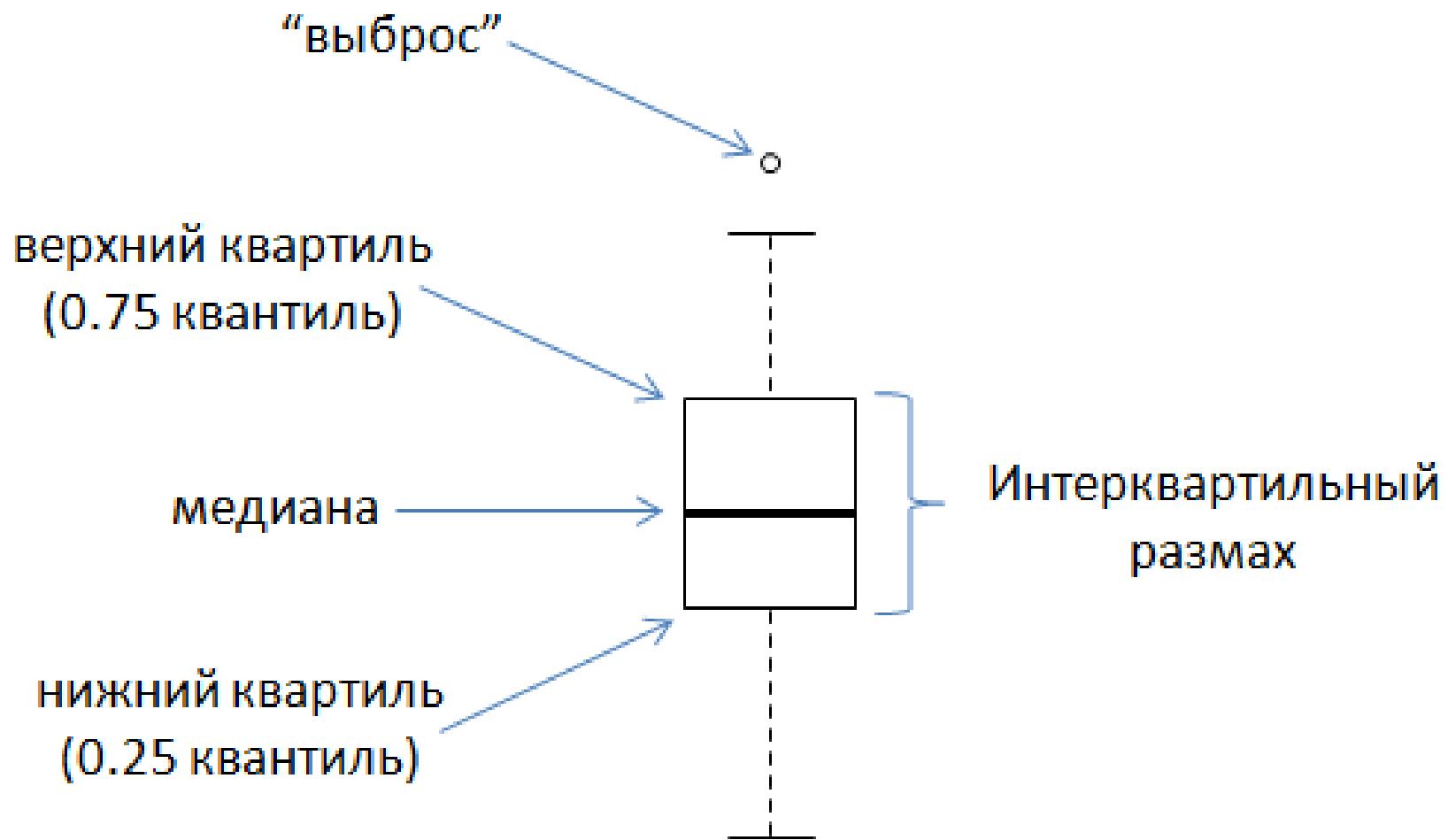
Результаты эксперимента Майкельсона-Морли

Подробнее:

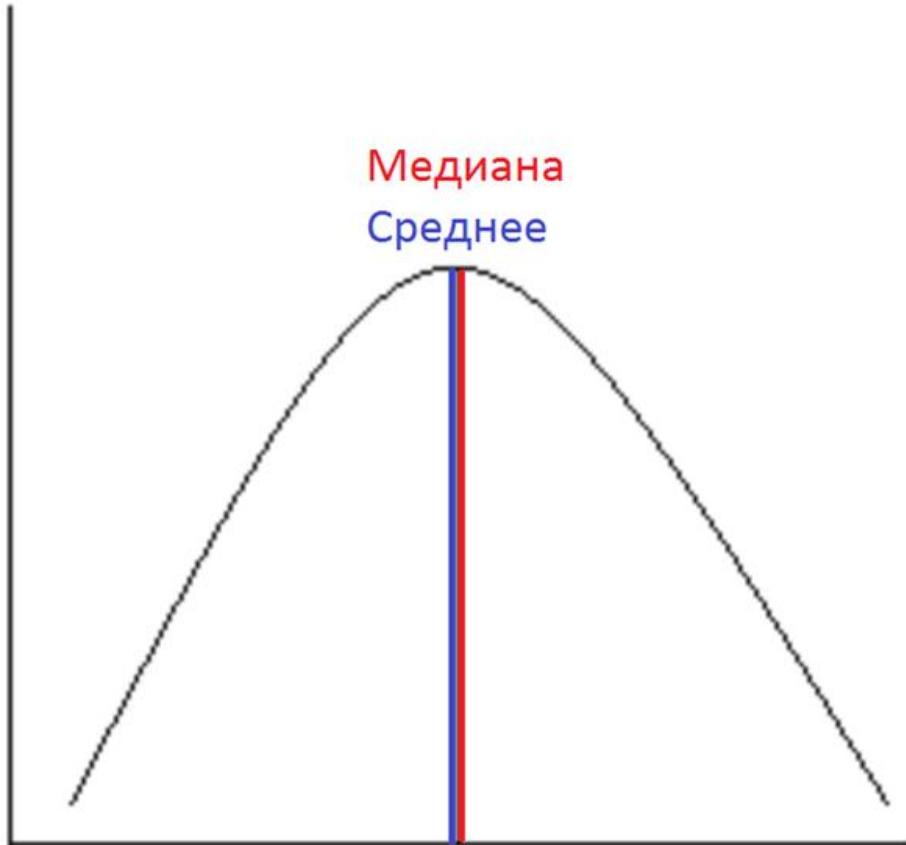
https://ru.wikipedia.org/wiki/Опыт_Майкельсона

Источник изображения: Schutz, Wikipedia, CC-BY

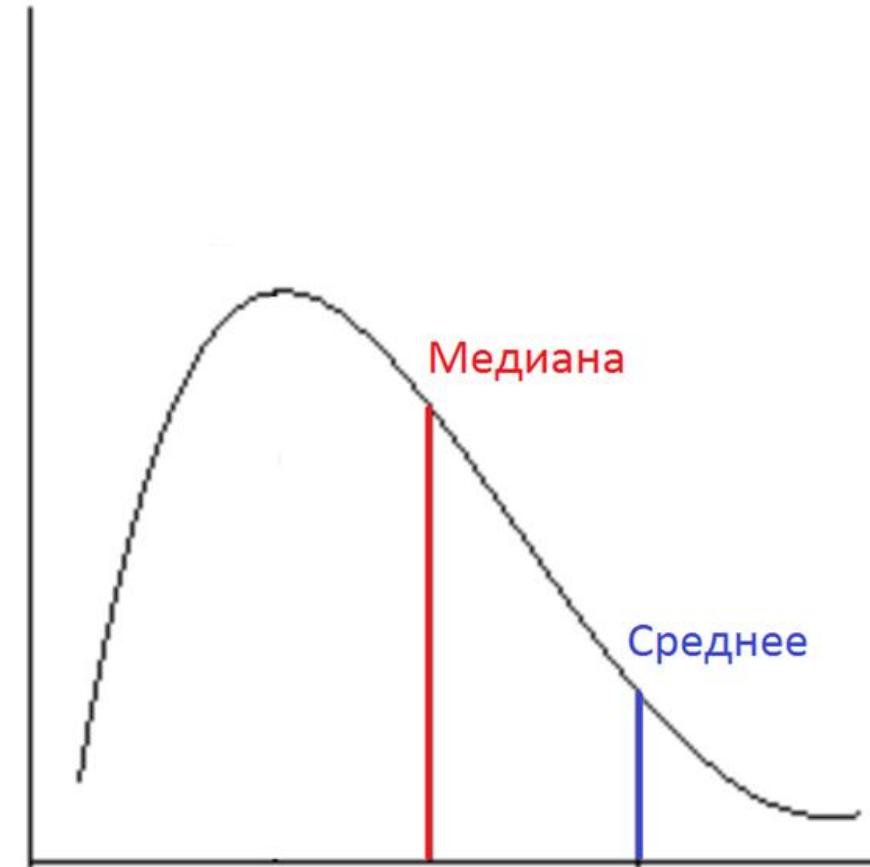
Структура ящика с усами



Медиана VS среднее



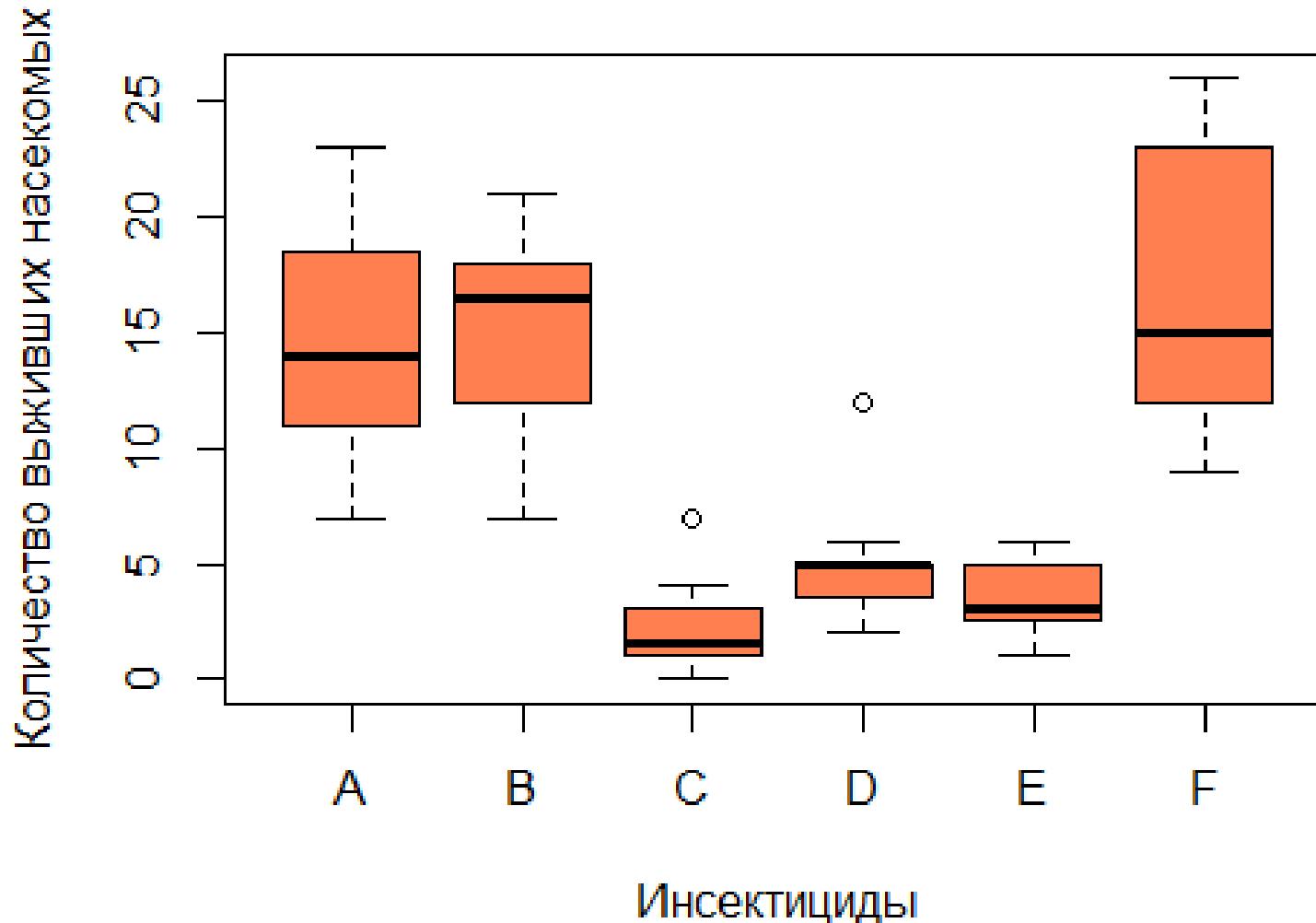
Нормальное распределение
(рост или вес человека)



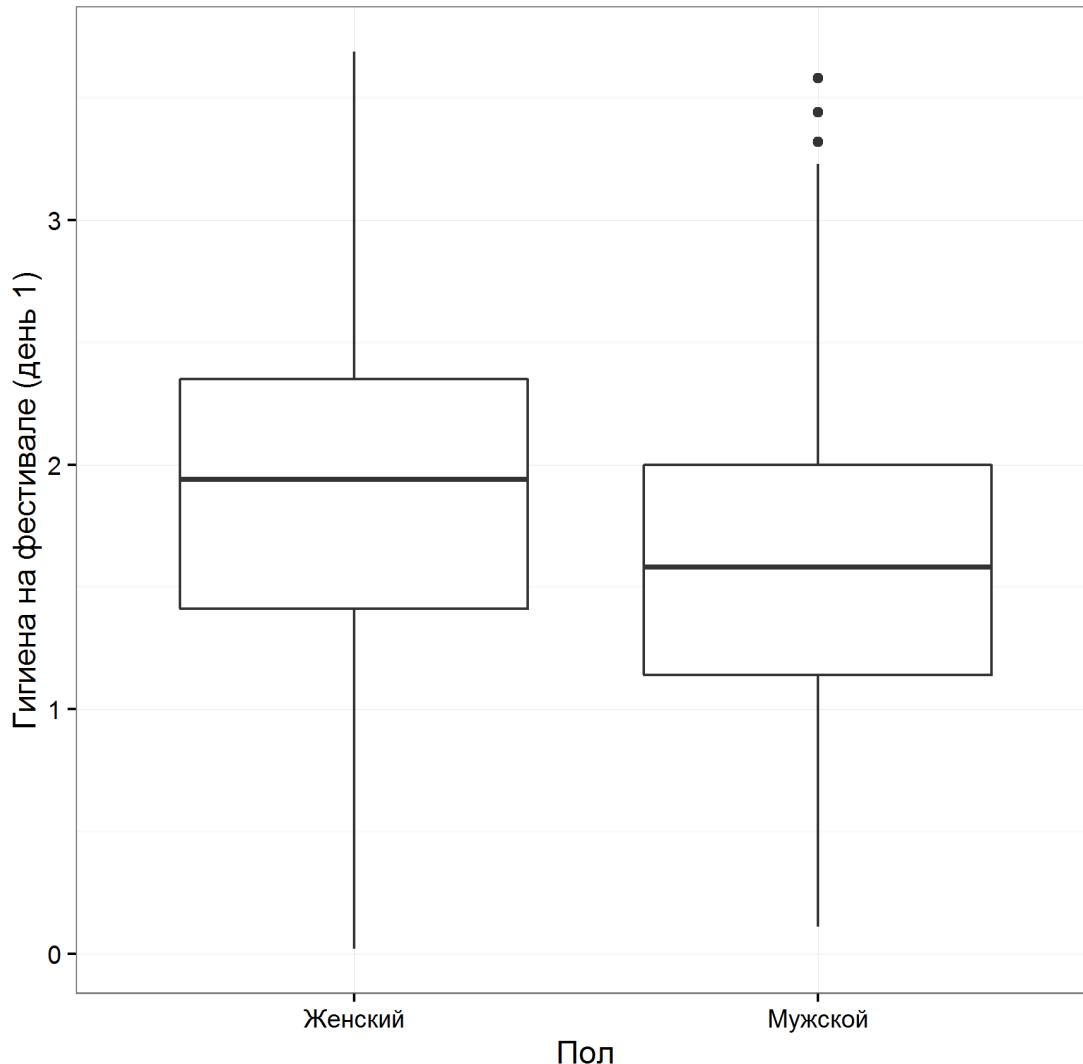
«Скошенное» распределение
(оооочень много чего в социальных науках,
например, доход)

Какой инсектицид лучше?

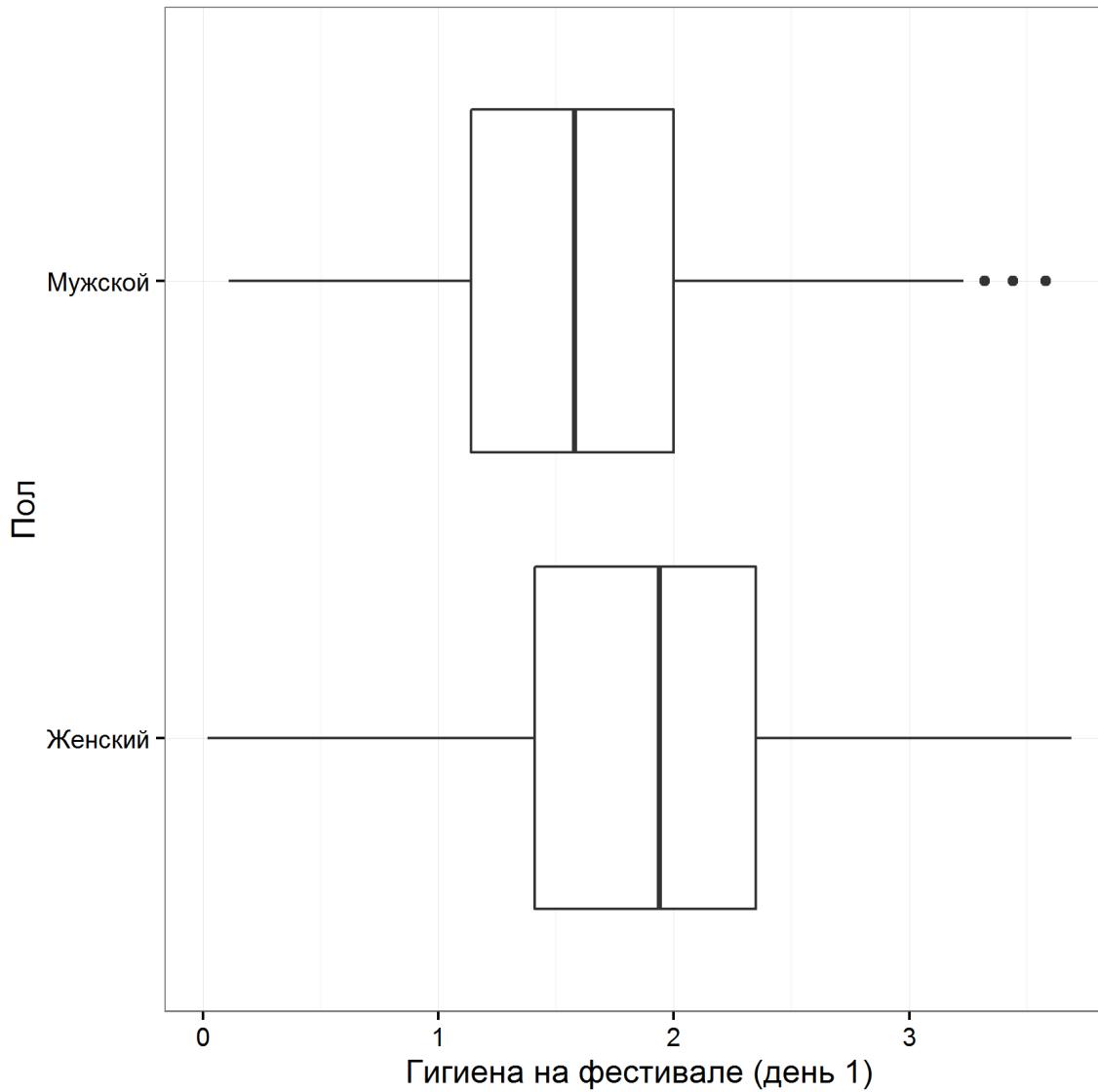
Эффективность инсектицидов



Возвращаясь к фестивалю...



Горизонтальный ящик с усами



Status, Faction Sizes, Social Influence

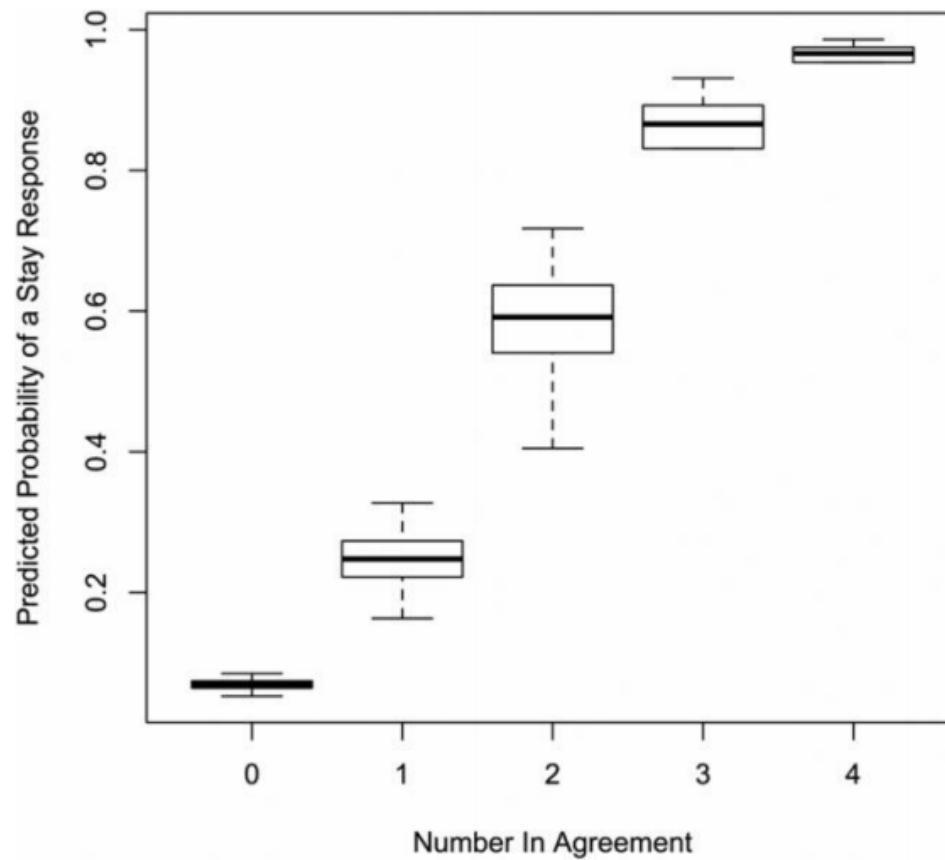


FIG. 1.—Marginal probabilities of a stay response by the number in agreement with participant's initial opinion as the expectation standing varies. Upper and lower limits reflect the 95th and 5th percentiles of the expectation standing, respectively. The upper and lower limits of the box reflect the 75th and 25th percentiles of the expectation standing, respectively. The bold line reflects the mean expectation standing. Estimates drawn from table 2, model 1.

Источник: Scott V. Savage, David Melamed. Status, Faction Sizes, and Social Influence: Testing the Theoretical Mechanism. *American Journal of Sociology*, Volume 121, Number 1 (July 2016), pp. 201–232.

Основные типы графиков

- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

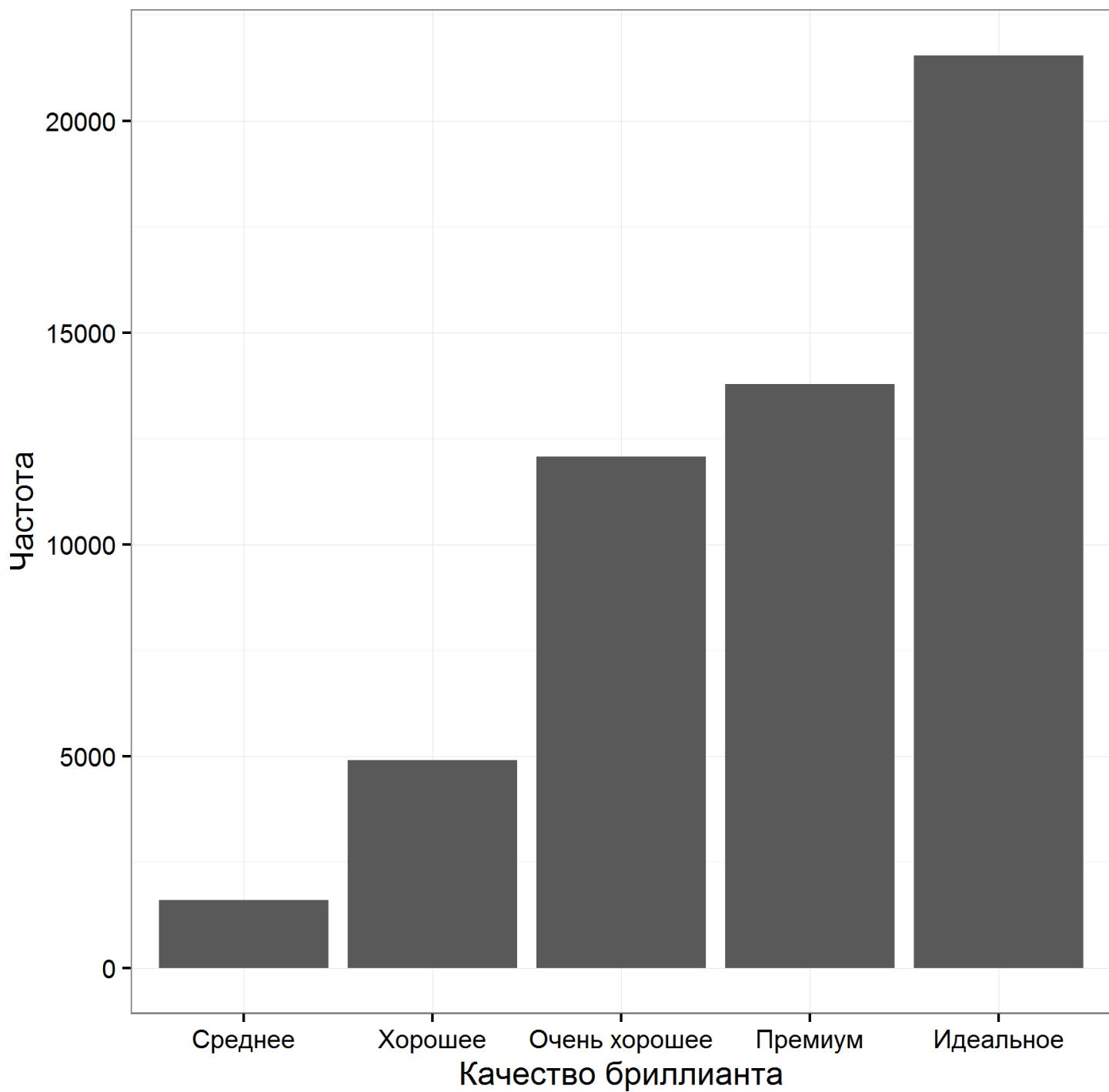
Столбиковая диаграмма (box chart) и бриллианты

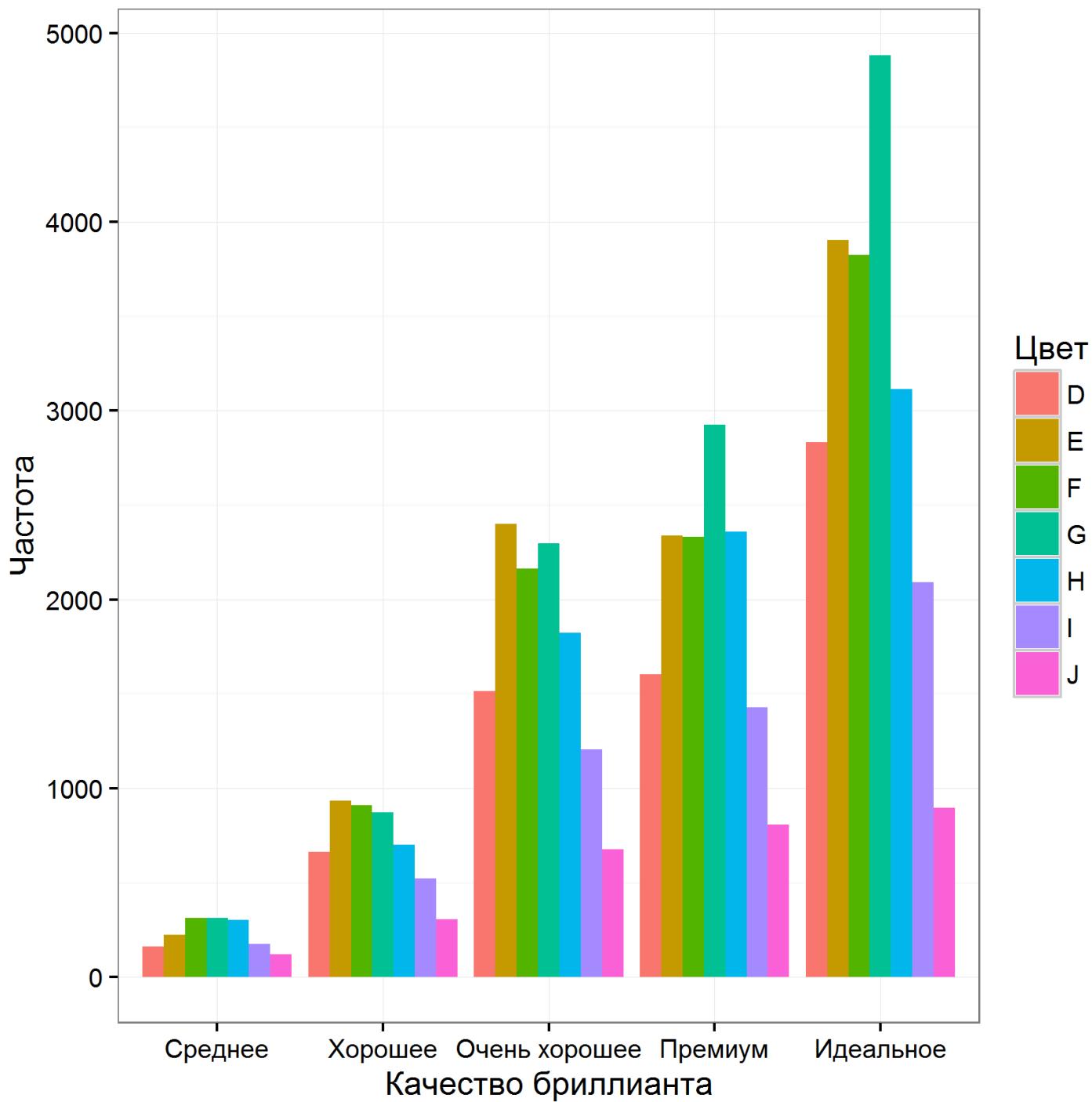


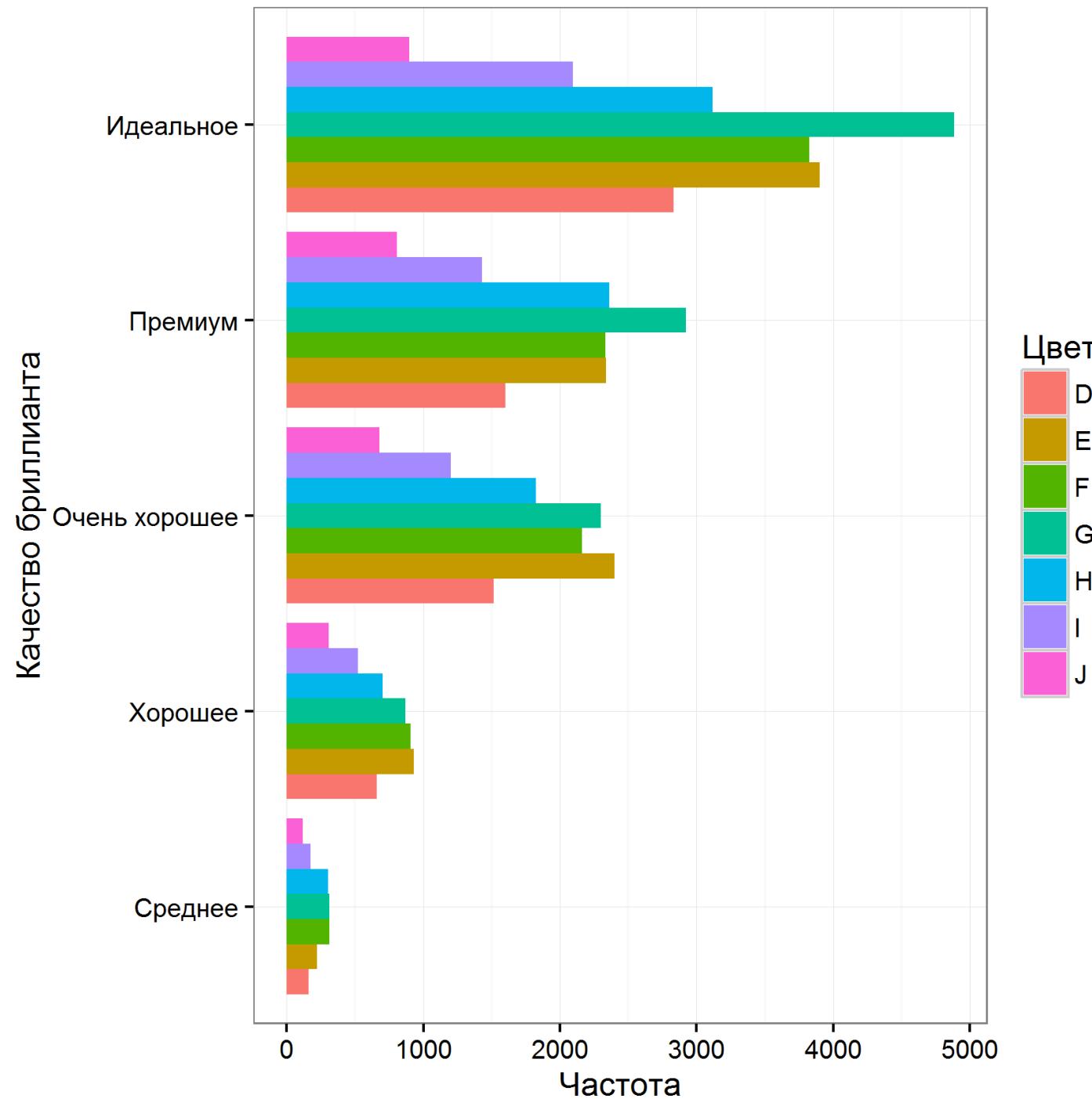
	carat	cut	color	clarity	depth	table	price
1	0.23	Ideal	E	SI2	61.5	55.0	326
2	0.21	Premium	E	SI1	59.8	61.0	326
3	0.23	Good	E	VS1	56.9	65.0	327
4	0.29	Premium	I	VS2	62.4	58.0	334
5	0.31	Good	J	SI2	63.3	58.0	335
6	0.24	Very Good	J	VVS2	62.8	57.0	336
7	0.24	Very Good	I	VVS1	62.3	57.0	336
8	0.26	Very Good	H	SI1	61.9	55.0	337
9	0.22	Fair	E	VS2	65.1	61.0	337
10	0.23	Very Good	H	VS1	59.4	61.0	338
11	0.30	Good	J	SI1	64.0	55.0	339
12	0.23	Ideal	J	VS1	62.8	56.0	340
13	0.22	Premium	F	SI1	60.4	61.0	342
14	0.31	Ideal	J	SI2	62.2	54.0	344
15	0.20	Premium	E	SI2	60.2	62.0	345
16	0.32	Premium	E	I1	60.9	58.0	345
17	0.30	Ideal	I	SI2	62.0	54.0	348
18	0.30	Good	J	SI1	63.4	54.0	351
19	0.30	Good	J	SI1	63.8	56.0	351
20	0.30	Very Good	J	SI1	62.7	59.0	351
21	0.30	Good	I	SI2	63.3	56.0	351

Showing 1 to 21 of 53,940 entries

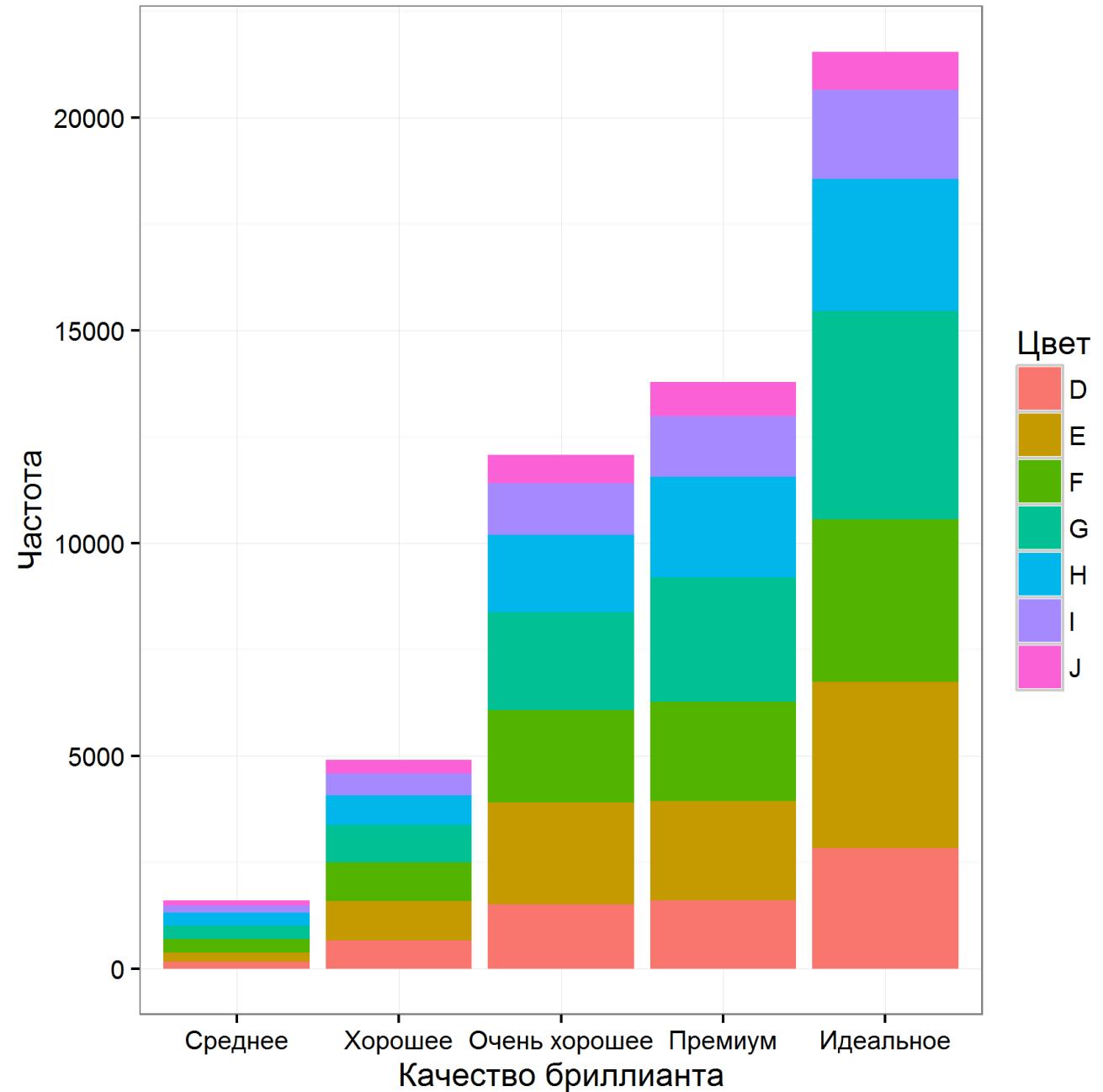


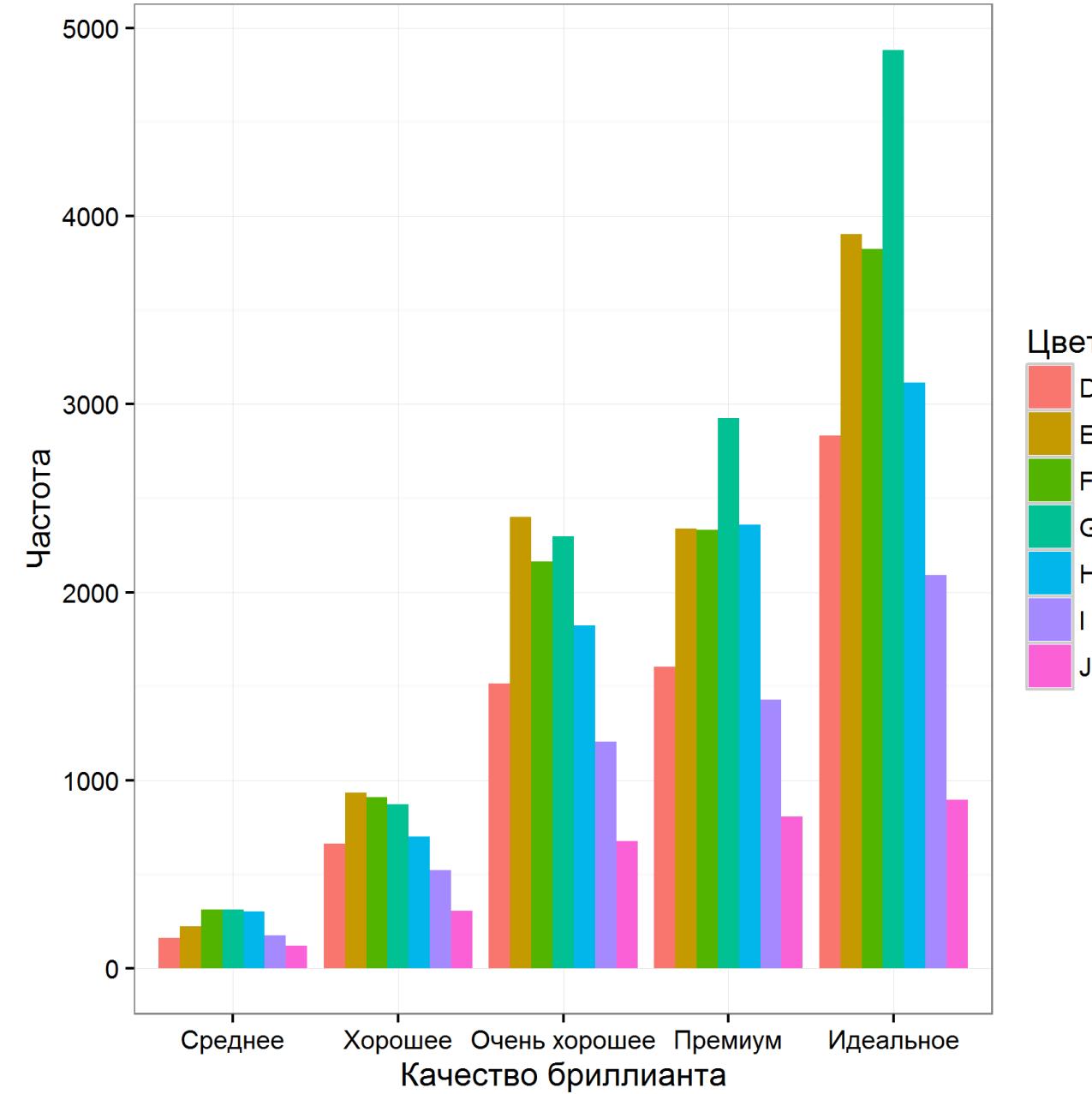
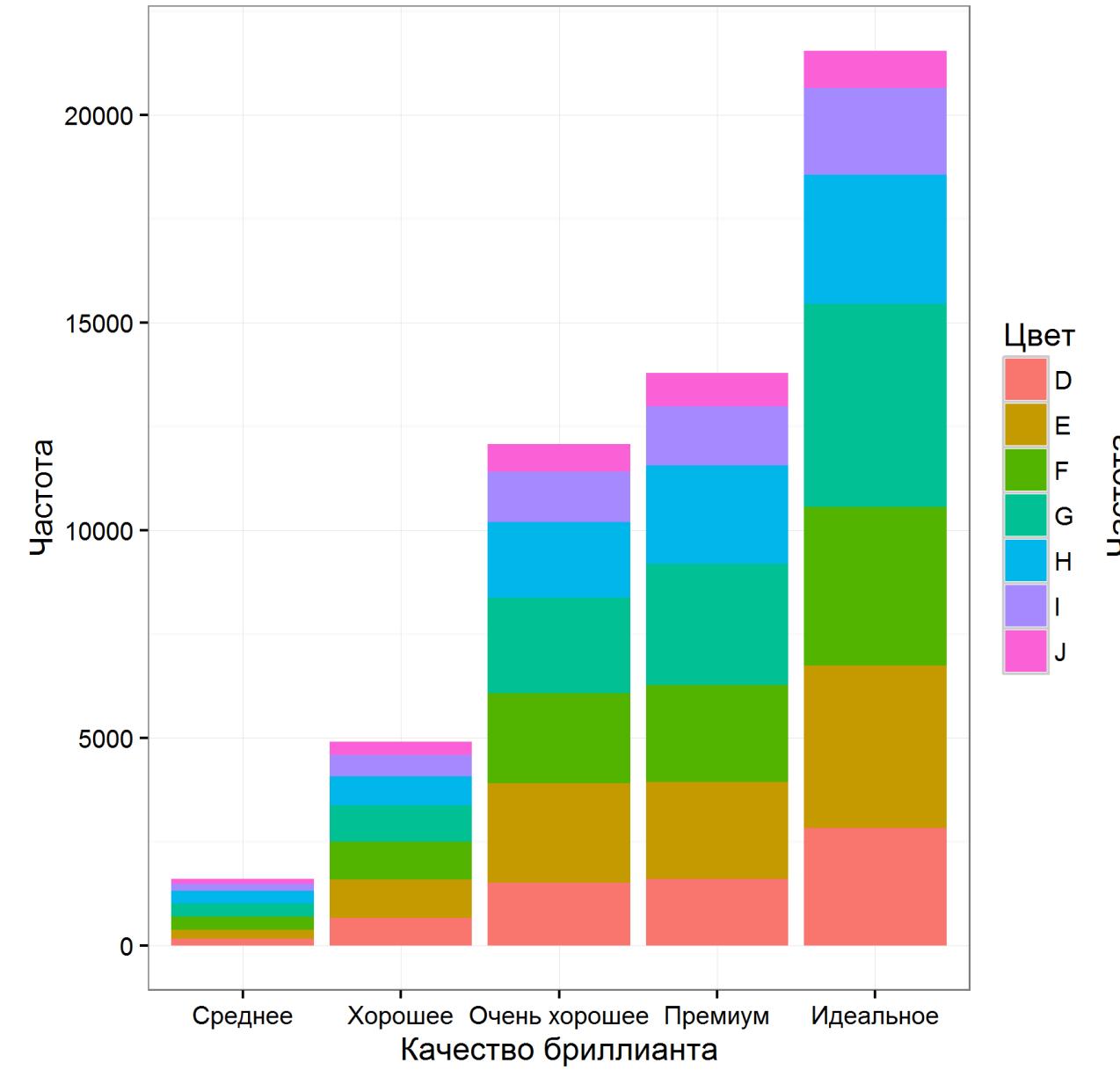




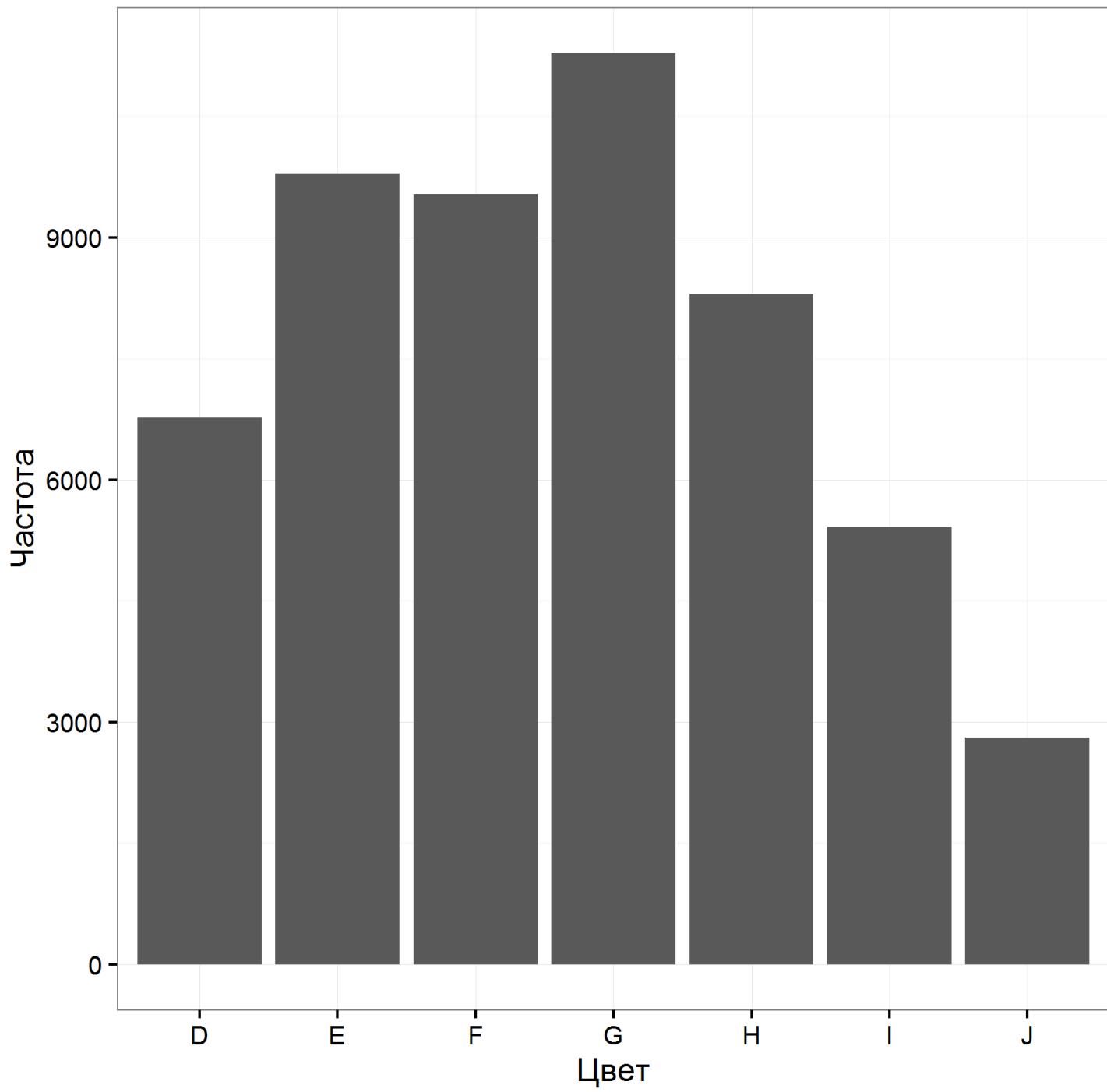


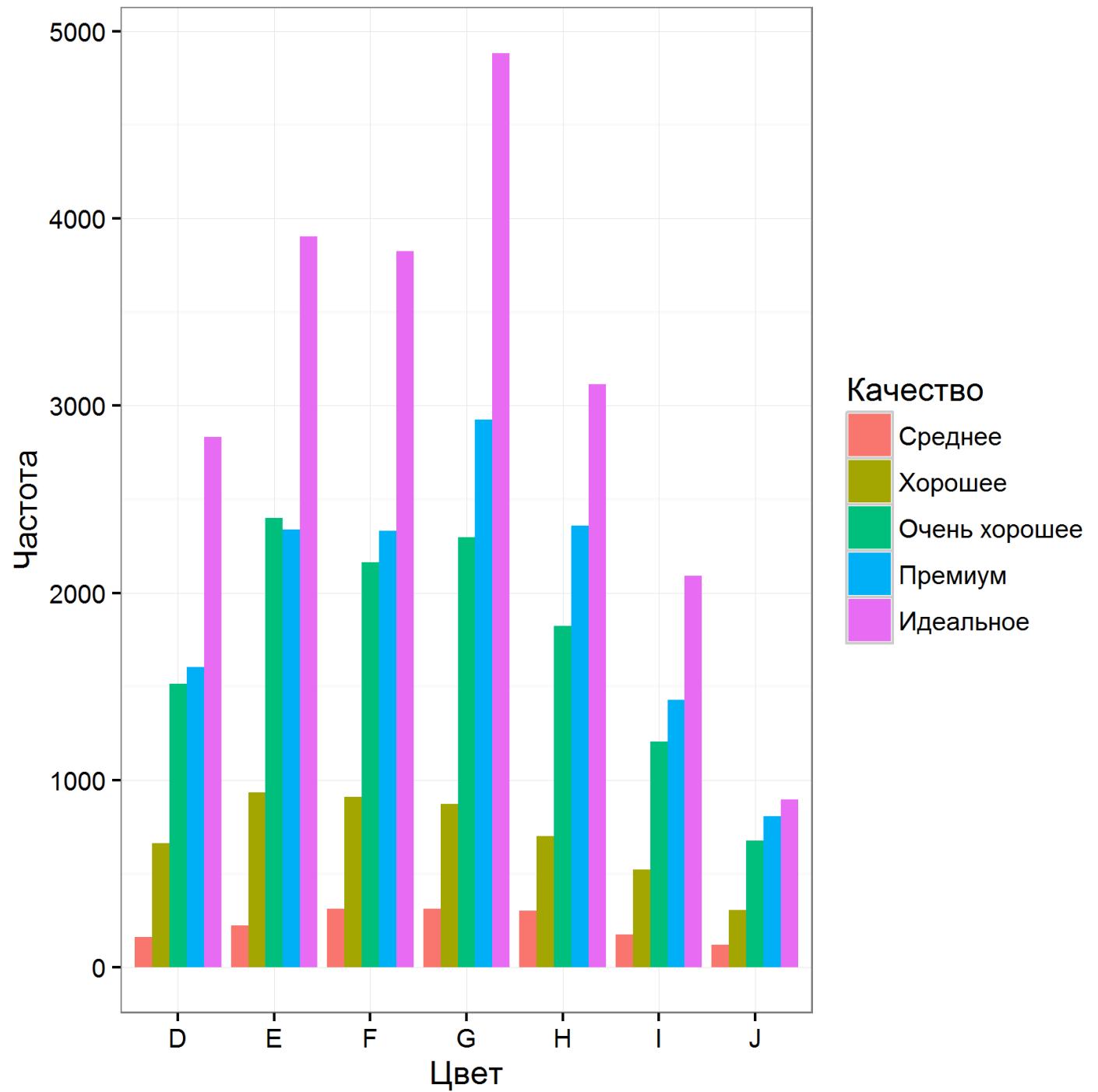
Stacked box chart





Поменяем местами цвет и качество
огранки





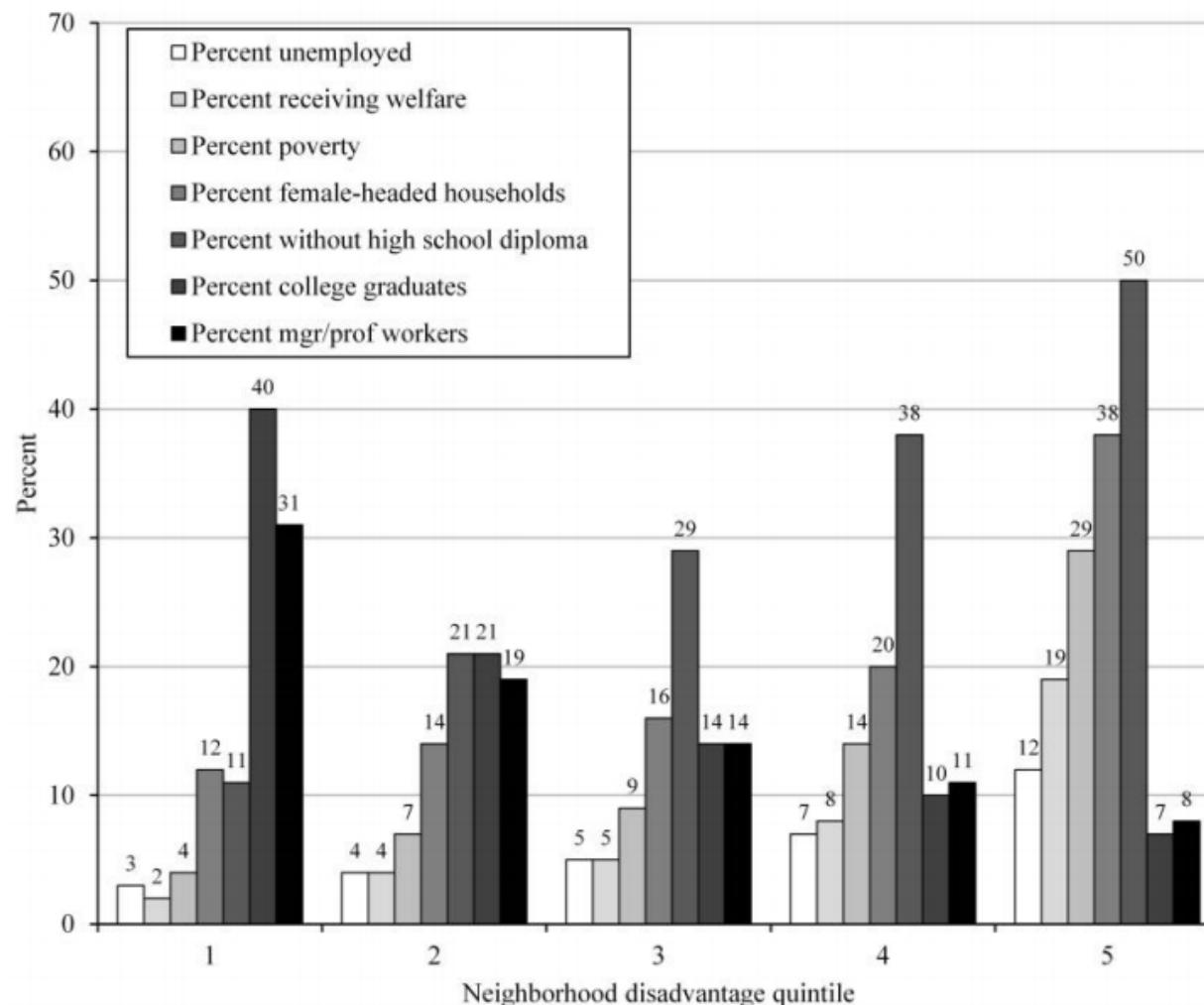


Fig. A1.— Neighborhood socioeconomic characteristics by disadvantage index quintile

Источник: Geoffrey T. Wodtke, Felix Elwert, David J. Harding. Neighborhood Effect Heterogeneity by Family Income and Developmental Period, *American Journal of Sociology*, Volume 121, Number 4 (January 2016), pp. 1168–1222.

Основные типы графиков

- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд, line chart)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

Динамика курса доллара США к рублю (USD, ЦБ РФ)



Дата	Курс	Изменение
28.10.16	63,0399	0,7802 ↑
27.10.16	62,2597	0,2117 ↑
26.10.16	62,0480	-0,1869 ↓
25.10.16	62,2349	-0,2150 ↓
22.10.16	62,4499	0,0305 ↑
21.10.16	62,4194	-0,1647 ↓
20.10.16	62,5841	-0,3059 ↓
19.10.16	62,8900	-0,2610 ↓
18.10.16	63,1510	0,1576 ↑
15.10.16	62,9934	-0,3531 ↓

Источник: <https://news.yandex.ru/quotes/1.html>

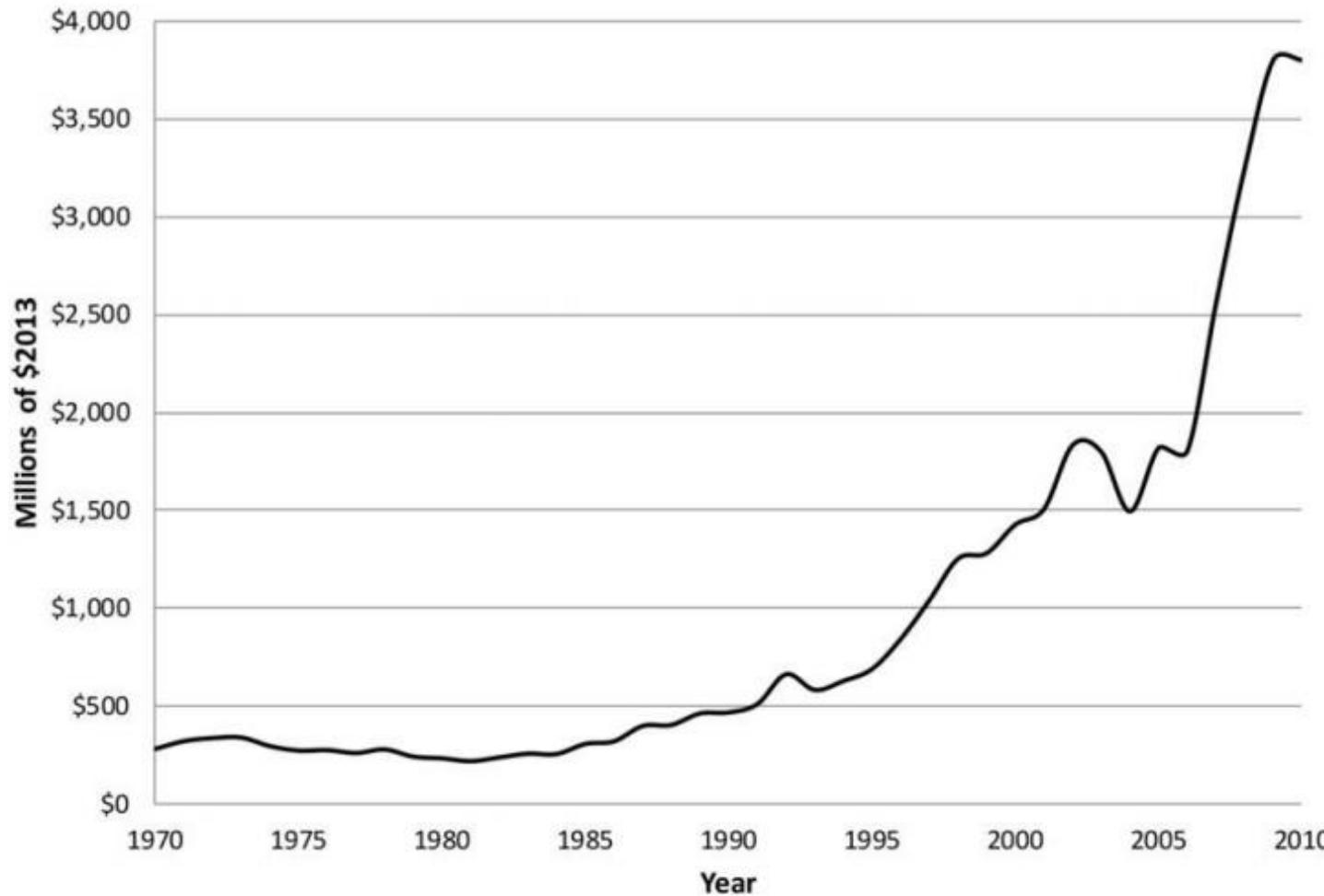


FIG. 1.—Border Patrol budget in millions of 2013 dollars

1567

Источник: Douglas S. Massey, Jorge Durand, Karen A. Pren. Why Border Enforcement Backfired, *American Journal of Sociology*, Volume 121, Number 5 (March 2016), pp. 1557–1600.

American Journal of Sociology

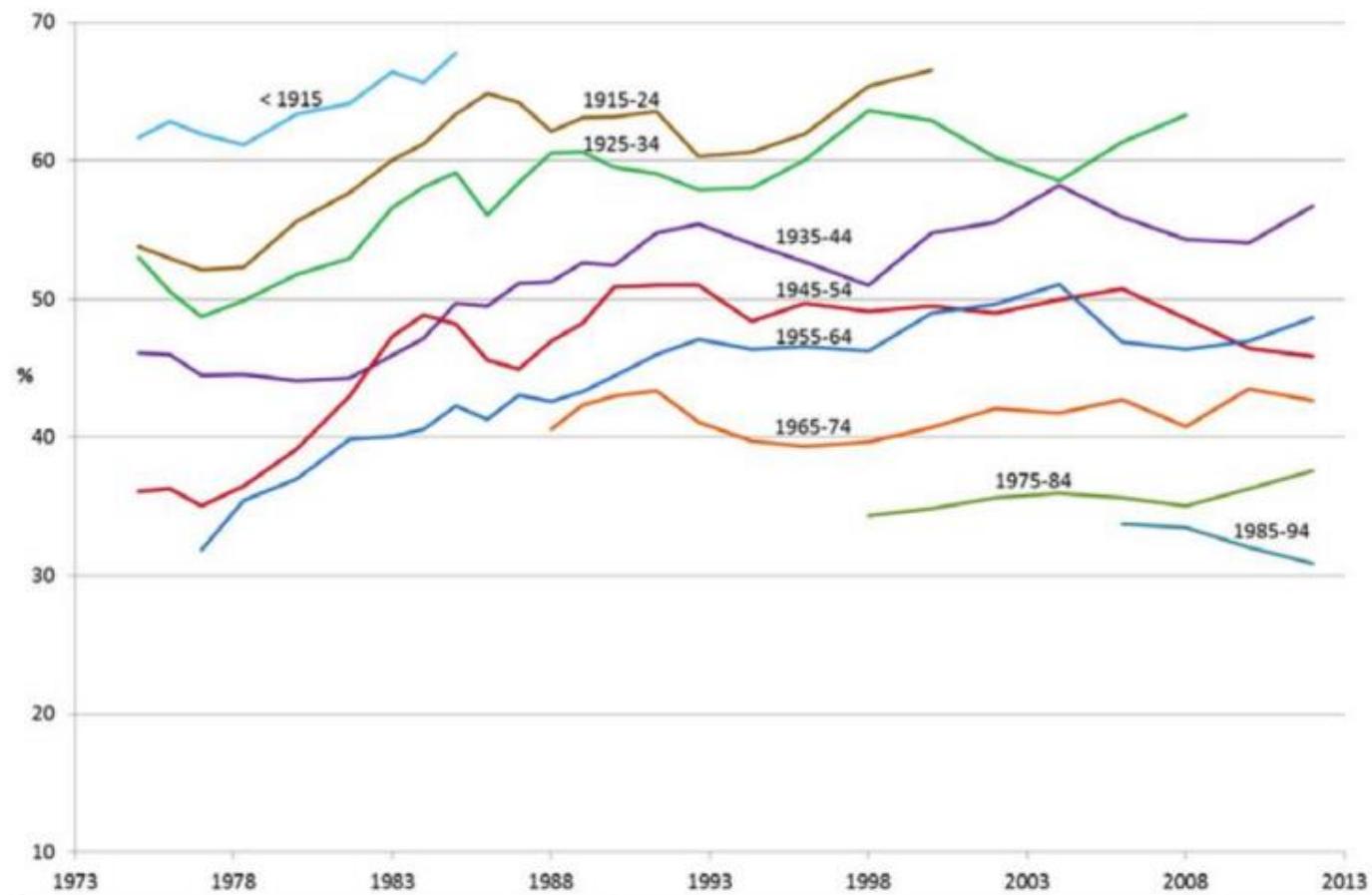


FIG. 8.—Strong or somewhat strong religious affiliation by decade of birth, United States, 1974–2014. Data are from the General Social Survey, 1974–2014. Includes respondents age 20–84 born in the United States. Three-survey moving average.

Источник: David Voas, Mark Chaves. Is the United States a Counterexample to the Secularization Thesis?
American Journal of Sociology, Volume 121, Number 5 (March 2016), pp. 1517–1556.

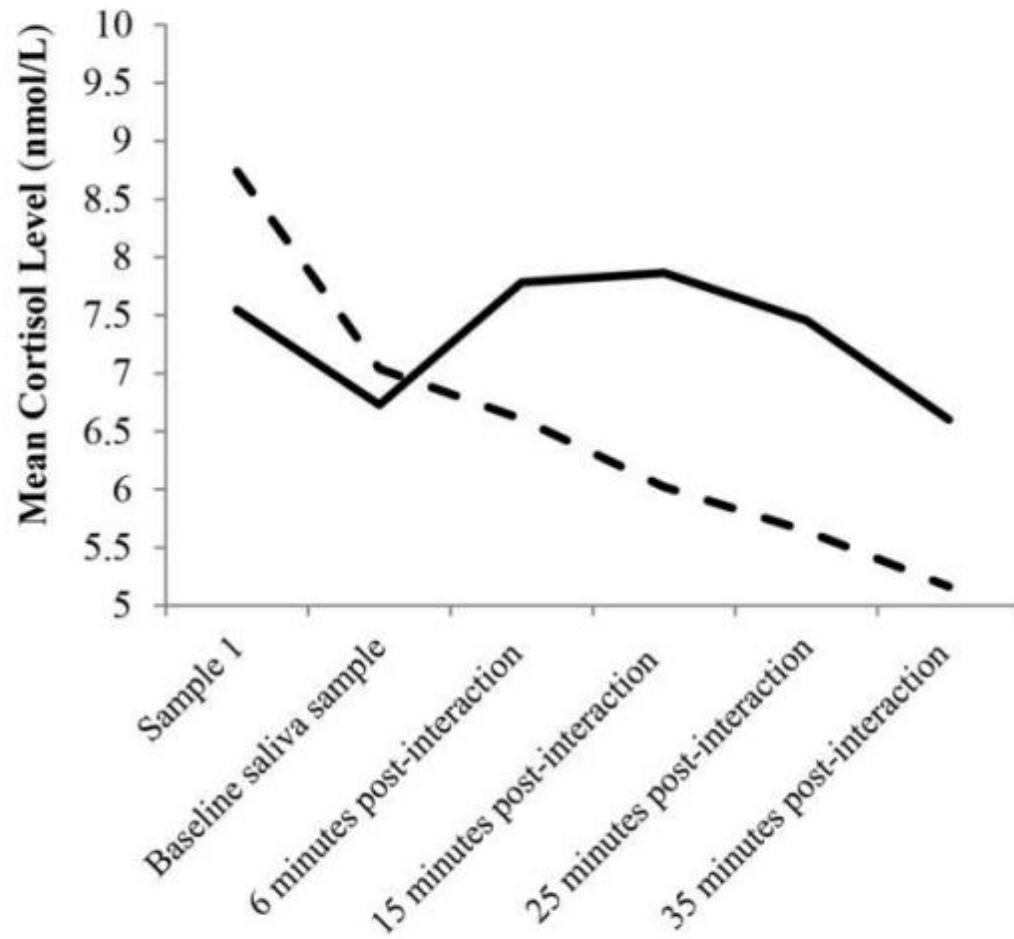


FIG. 3.—Mean level of cortisol in experiment 1. Solid line represents the experimental condition: gendered social exclusion and minority status by sex. Dashed line represents the control condition: gender-neutral social inclusion and majority status by sex.

Источник: Catherine J. Taylor. "Relational by Nature"? Men and Women Do Not Differ in Physiological Response to Social Stressors Faced by Token Women. *American Journal of Sociology*, Volume 121, Number 1 (July 2016), pp. 49–89

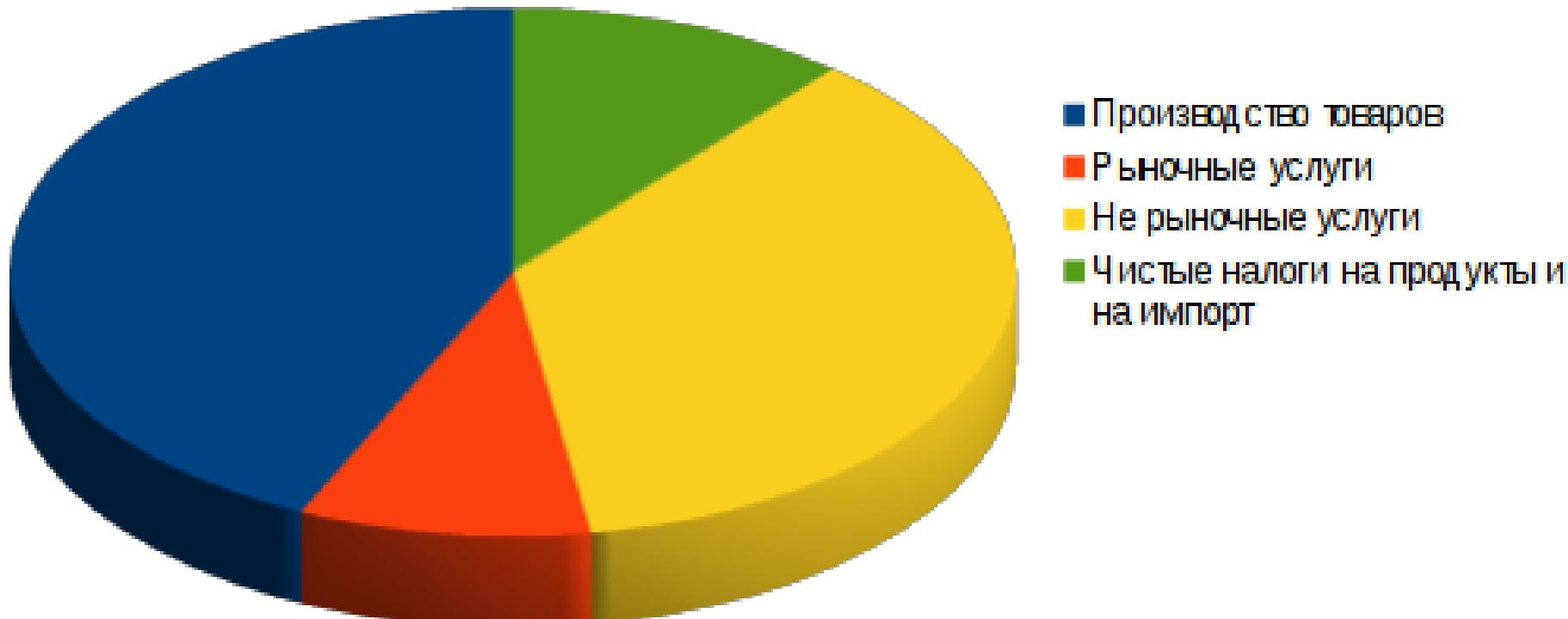
Основные типы графиков

- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

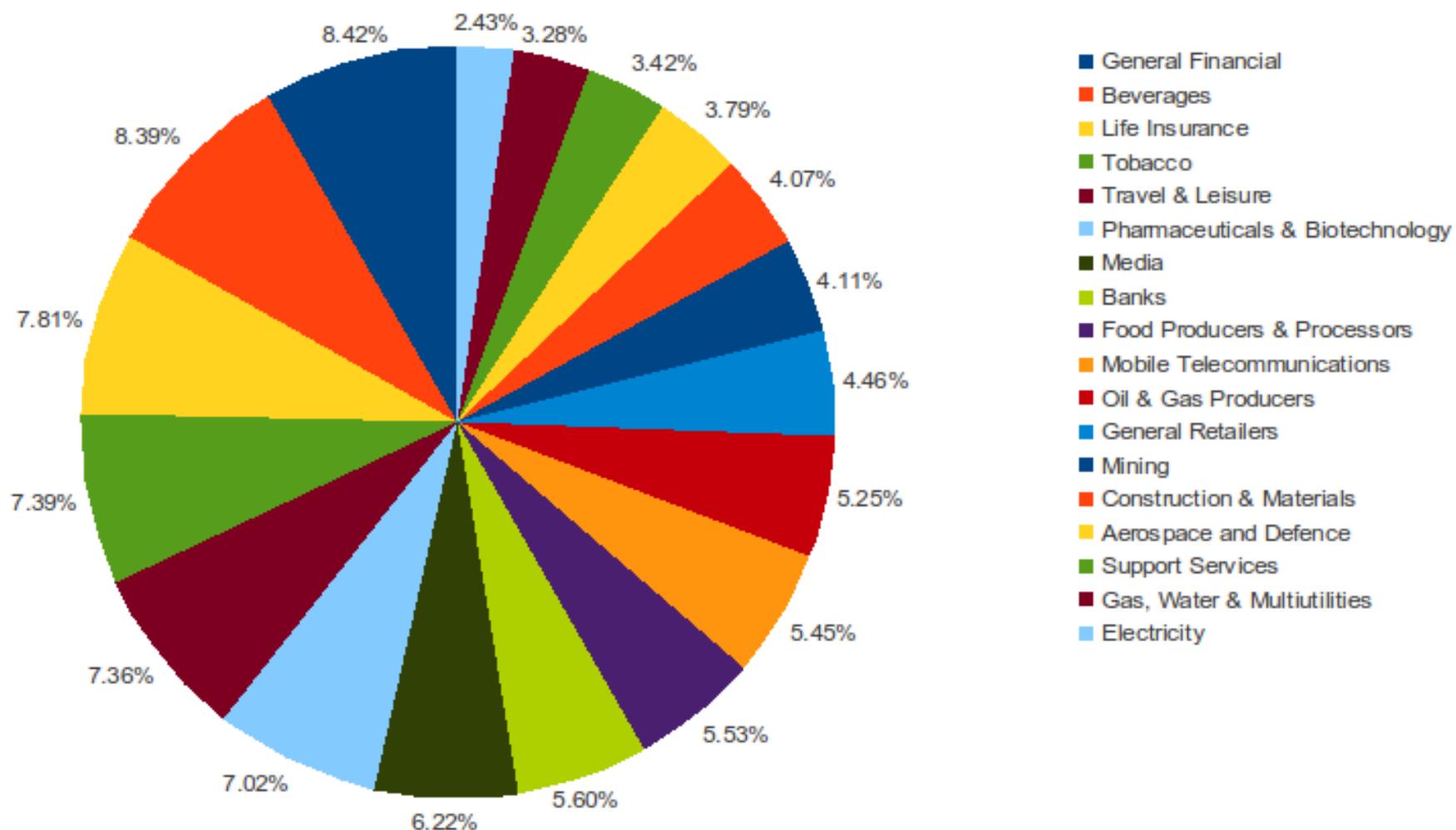
Типичная круговая диаграмма

Состав ВВП с точки зрения производства за 2004 год

текущие цены, млрд. рублей



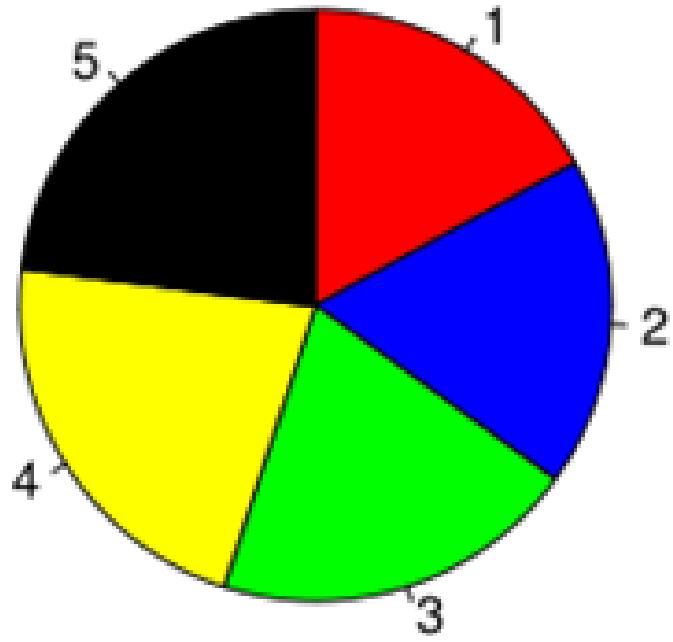
Sector Weightings



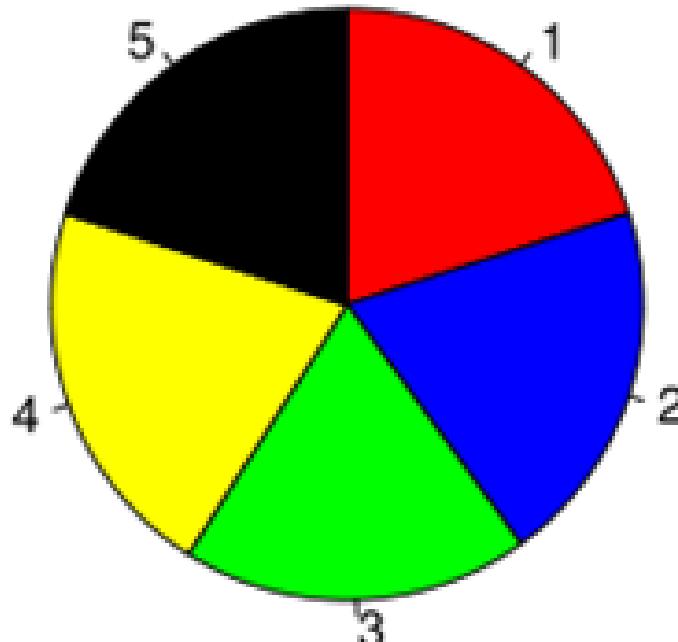
Никогда не используйте круговые
диаграммы.

Однаковы ли «куски пирога» на диаграммах?

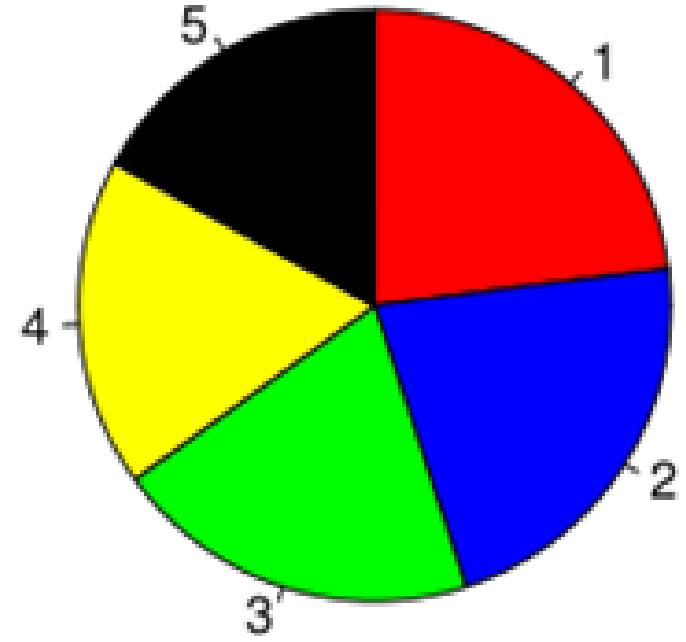
A

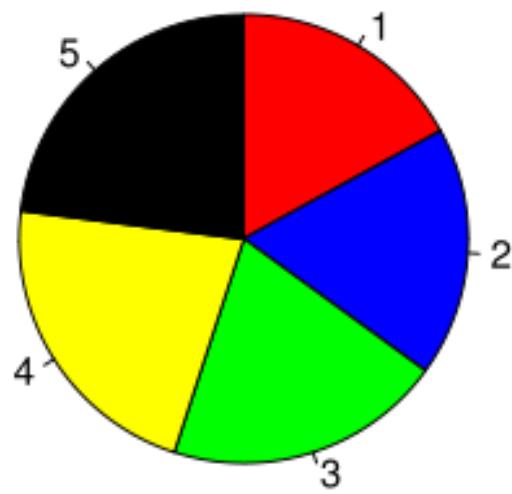
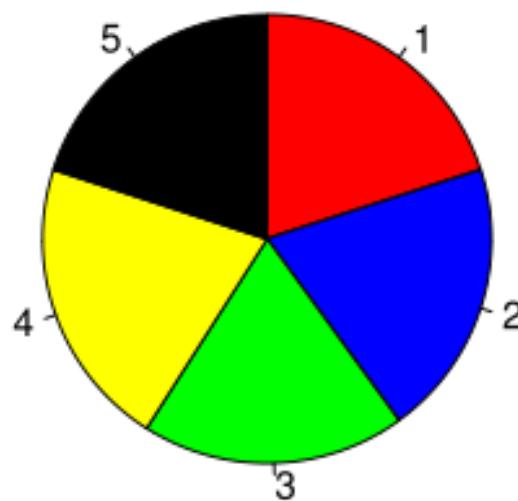
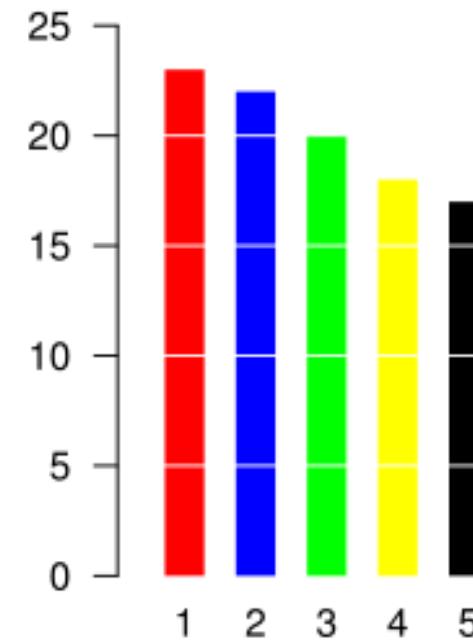
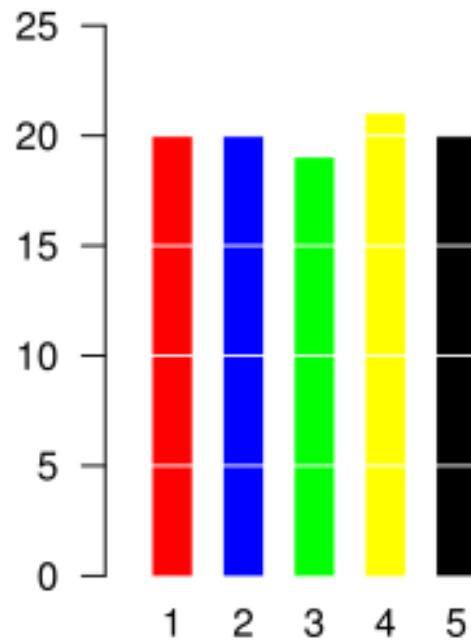
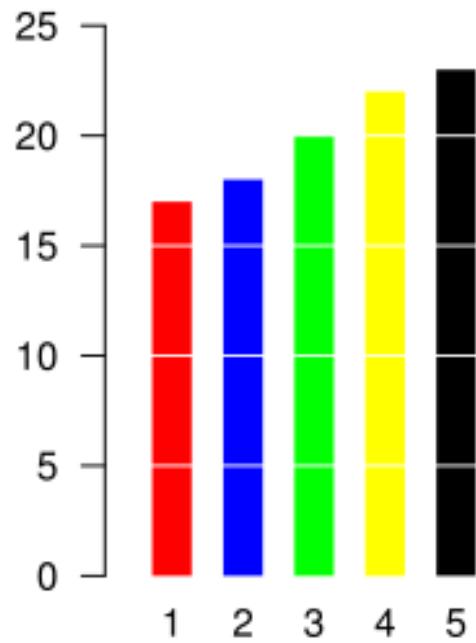
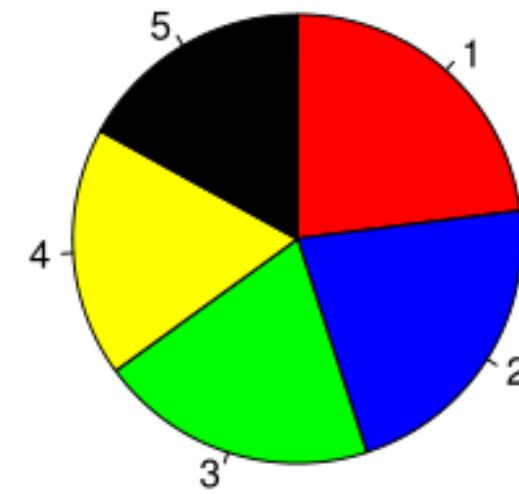


B



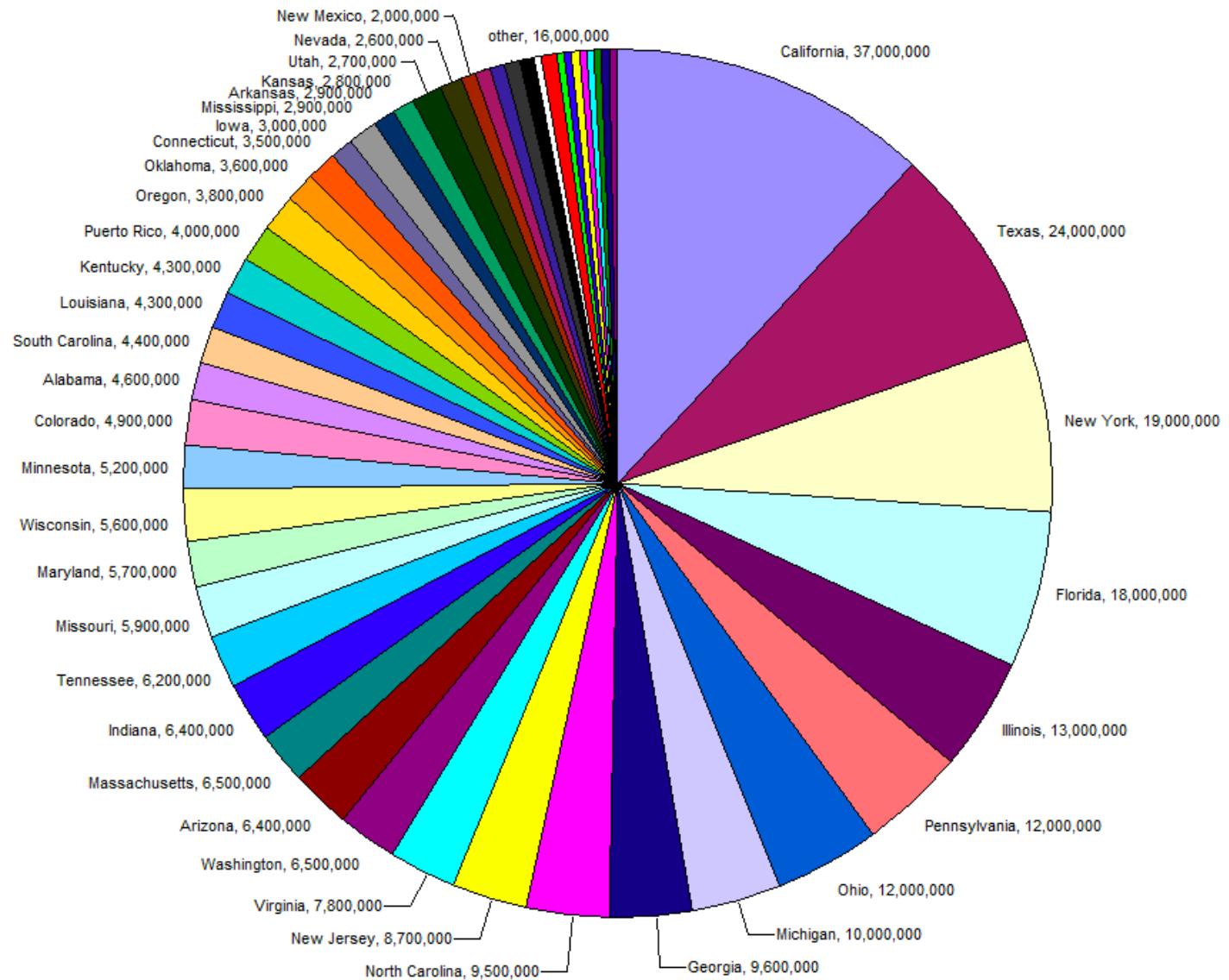
C



A**B****C**

“In a sense, it might be construed as an insult to a man's intelligence to show him a pie chart.”

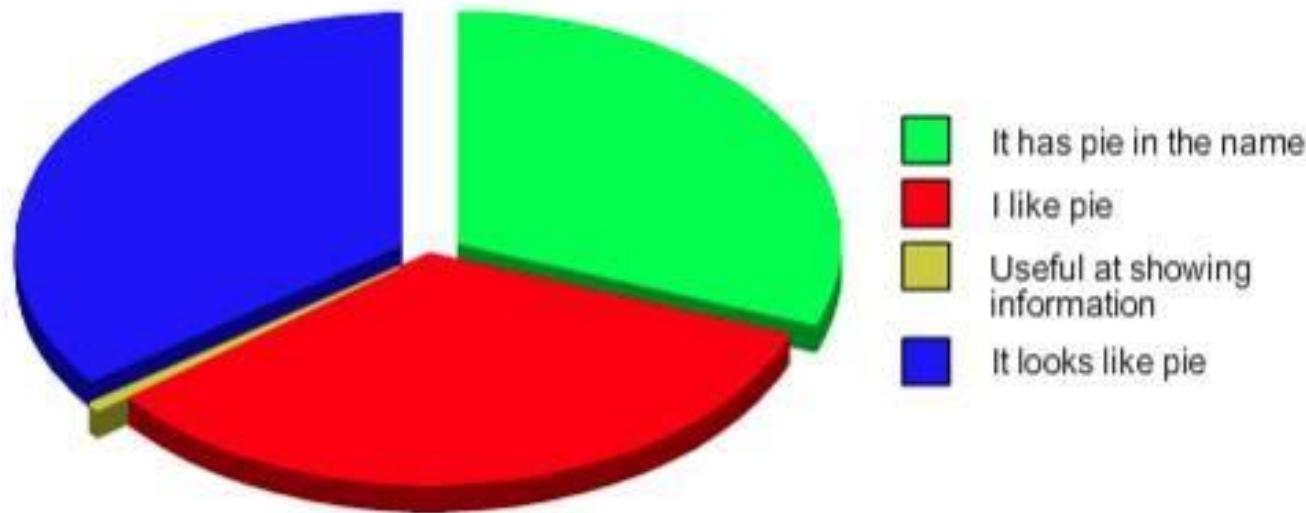
K.G. Karsten, Charts and Graphs (1923)



Никогда не используйте круговые
диаграммы.

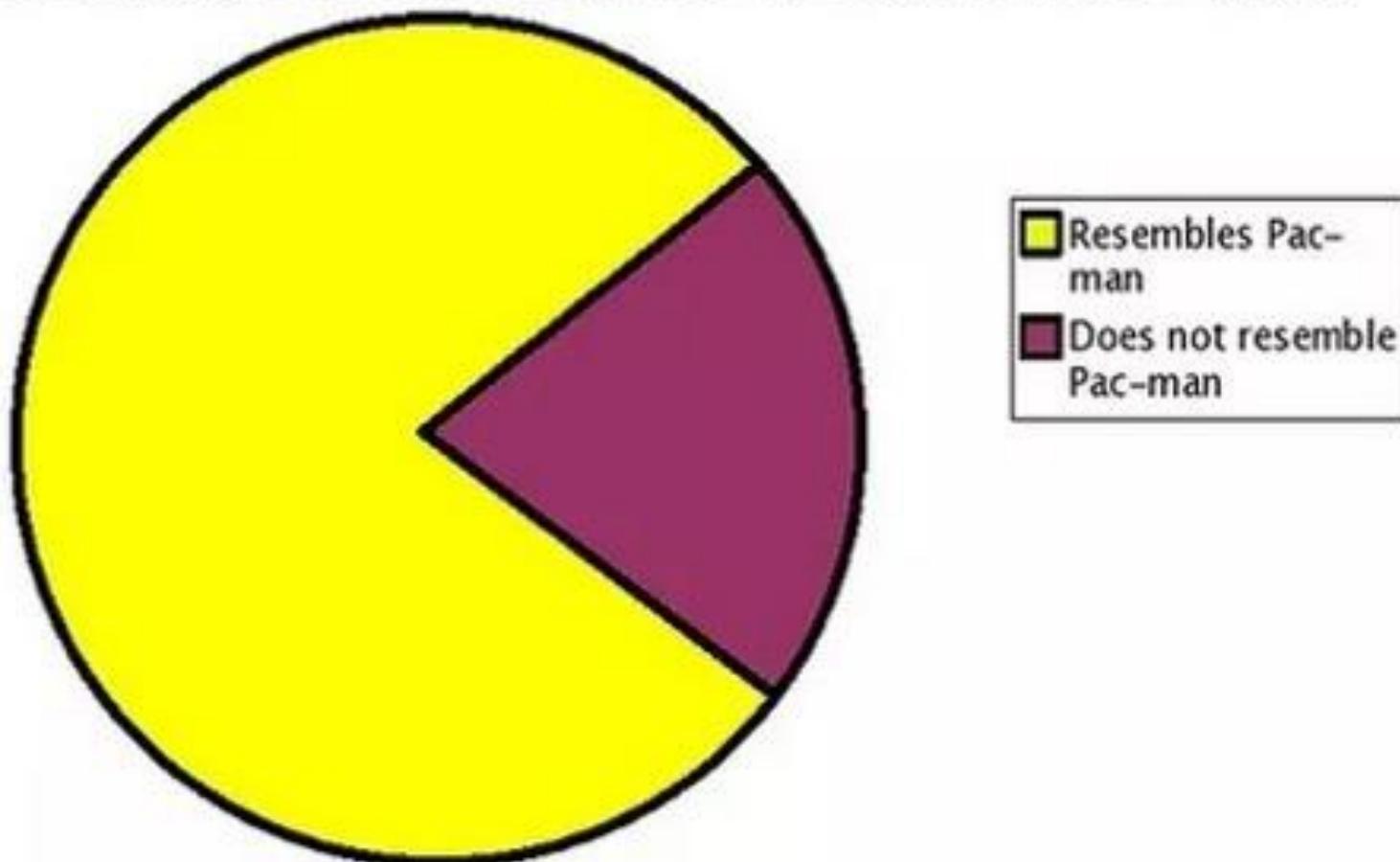
...только если не хотите пошутить.

Why I like pie charts

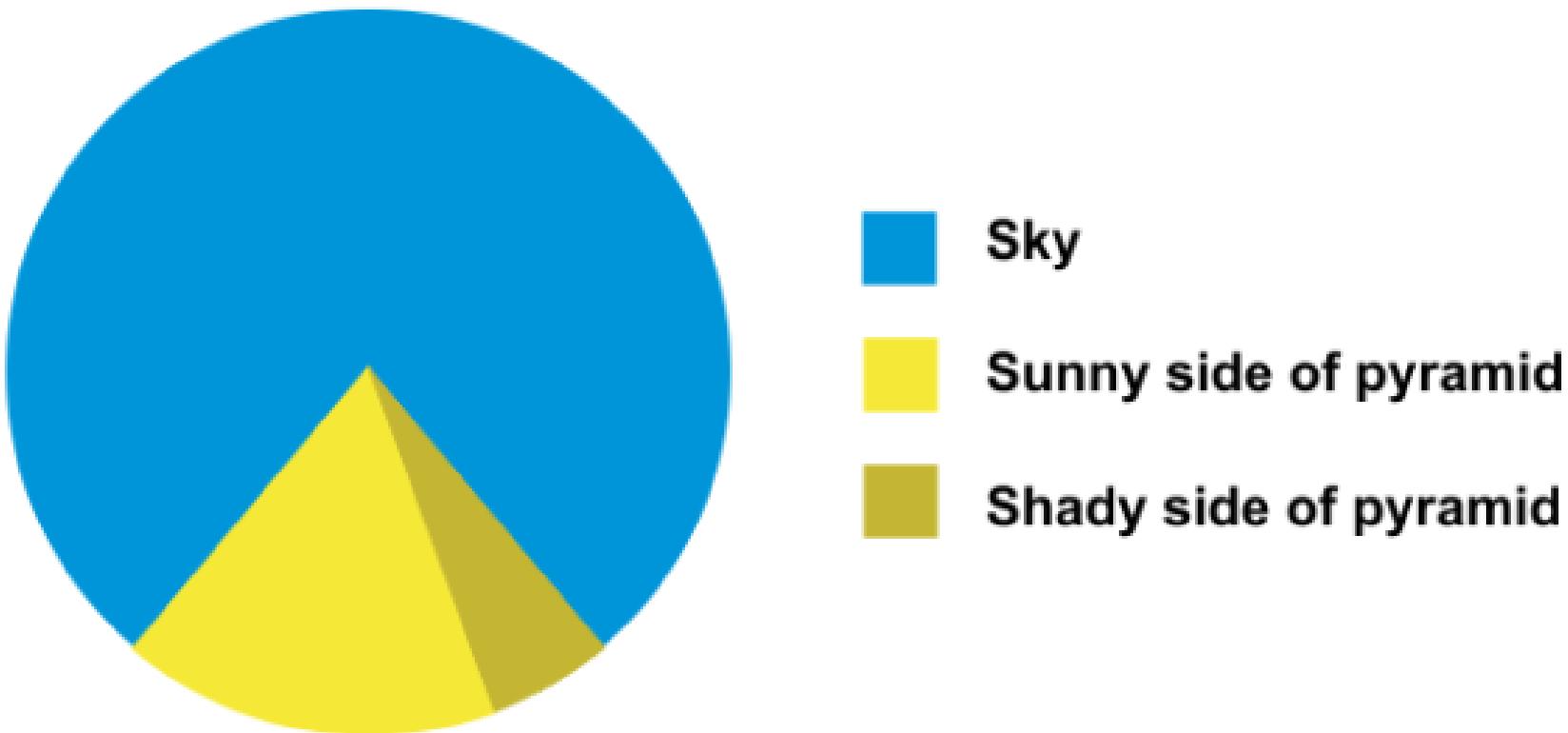


...только если не хотите пошутить.

Percentage of Chart Which Resembles Pac-man



...только если не хотите пошутить.



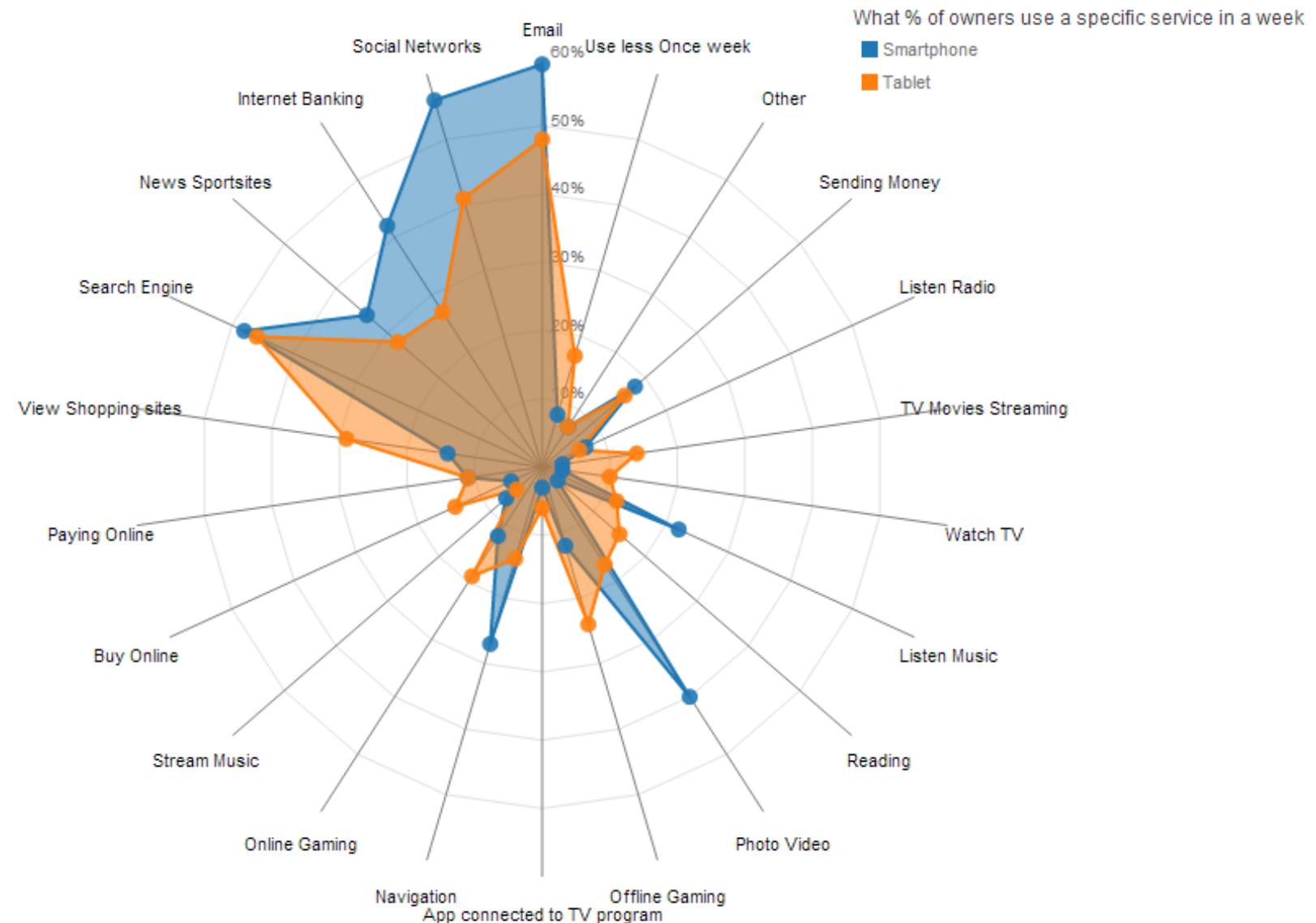
Единственная правильная круговая диаграмма



Основные типы графиков

- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

Радиальная диаграмма (radar chart)



Основные типы графиков

- Диаграмма рассеяния (scatterplot)
- Диаграмма распределения / гистограмма (histogram)
- «Ящик с усами» / диаграмма размаха (boxplot, box-and-whiskers)
- Столбиковая диаграмма (bar chart)
- Линейная диаграмма (временной ряд)
- Круговая диаграмма, или диаграмма-пирог (pie chart)
- Радиальная диаграмма

Ошибки при создании графиков

Ошибки при создании графиков

- Лишние графические элементы (**chartjunk**): штриховка, 3D, тени, заливка...

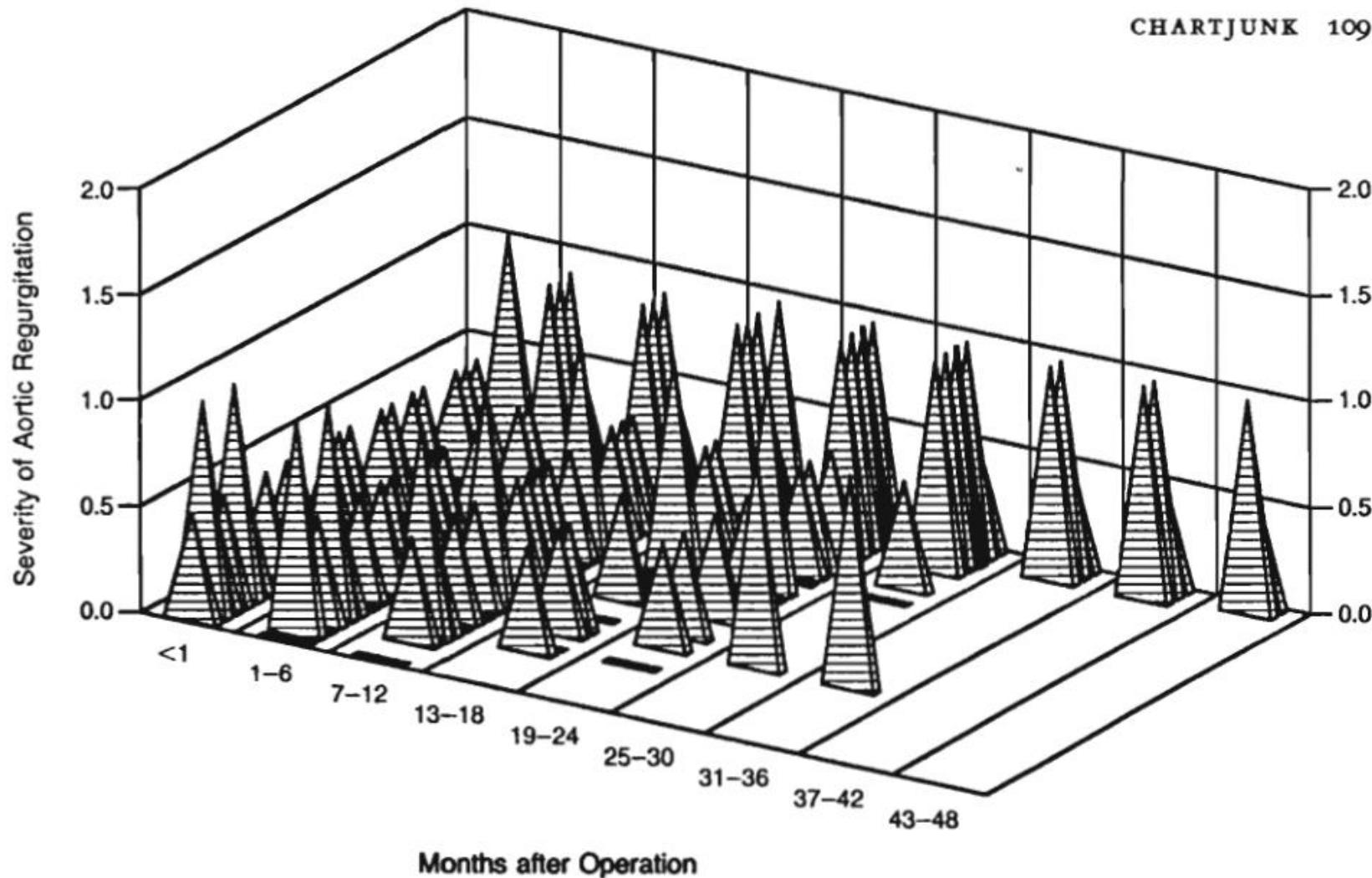


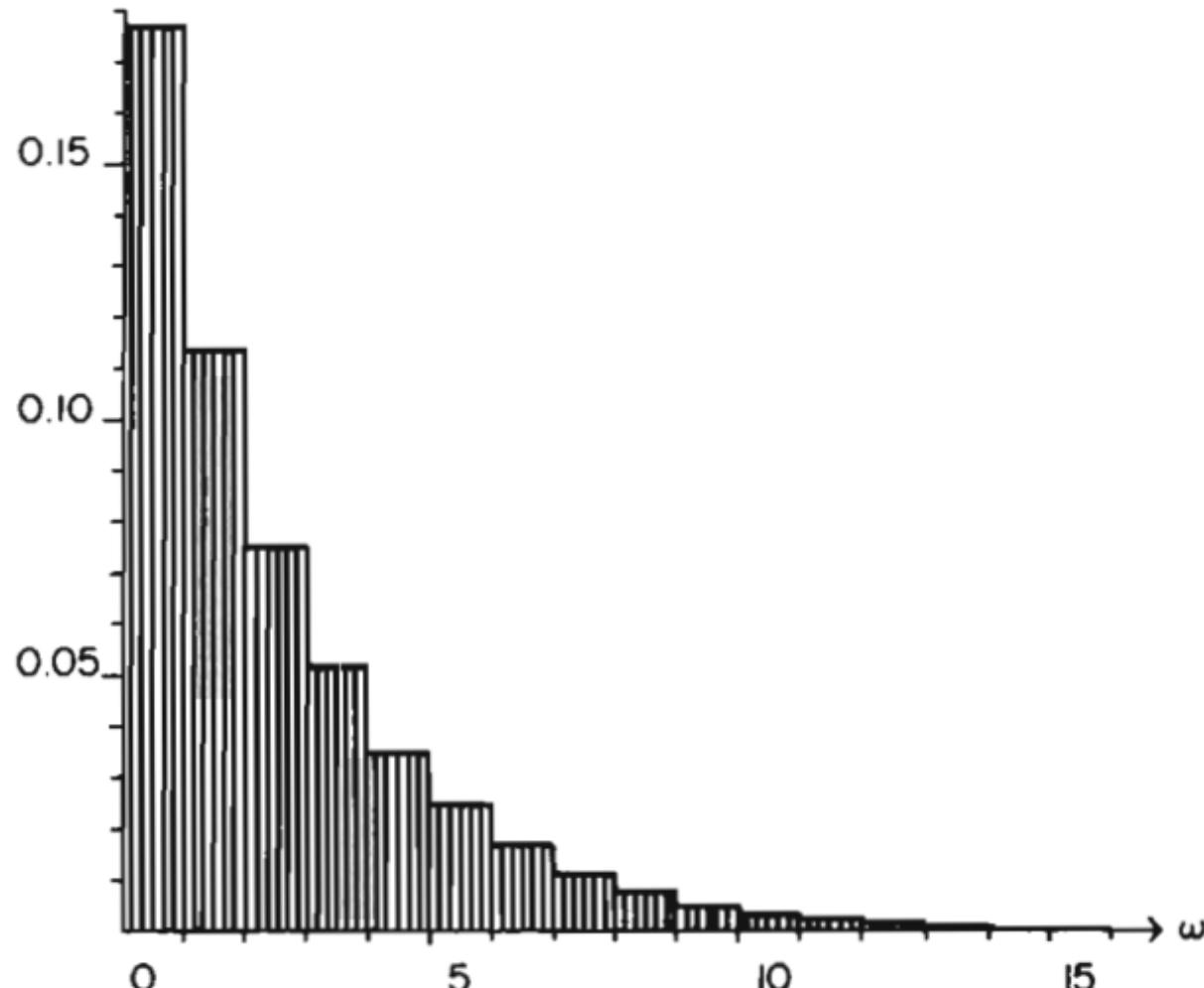
Figure 2. Serial Echocardiographic Assessments of the Severity of Regurgitation in the Pulmonary Autograft in 31 Patients. The numerical grades were assigned according to the severity of regurgitation, as follows: 0, none; 0.5, trivial; 1.0 to 1.5, mild; 2.0, moderate; and 3.0, severe.

Nicholas T. Kouchoukos, *et al.*,
"Replacement of the Aortic Root with
a Pulmonary Autograft in Children and
Young Adults with Aortic-Valve Disease,"
The New England Journal of Medicine, 330
(January 6, 1994), p. 4.

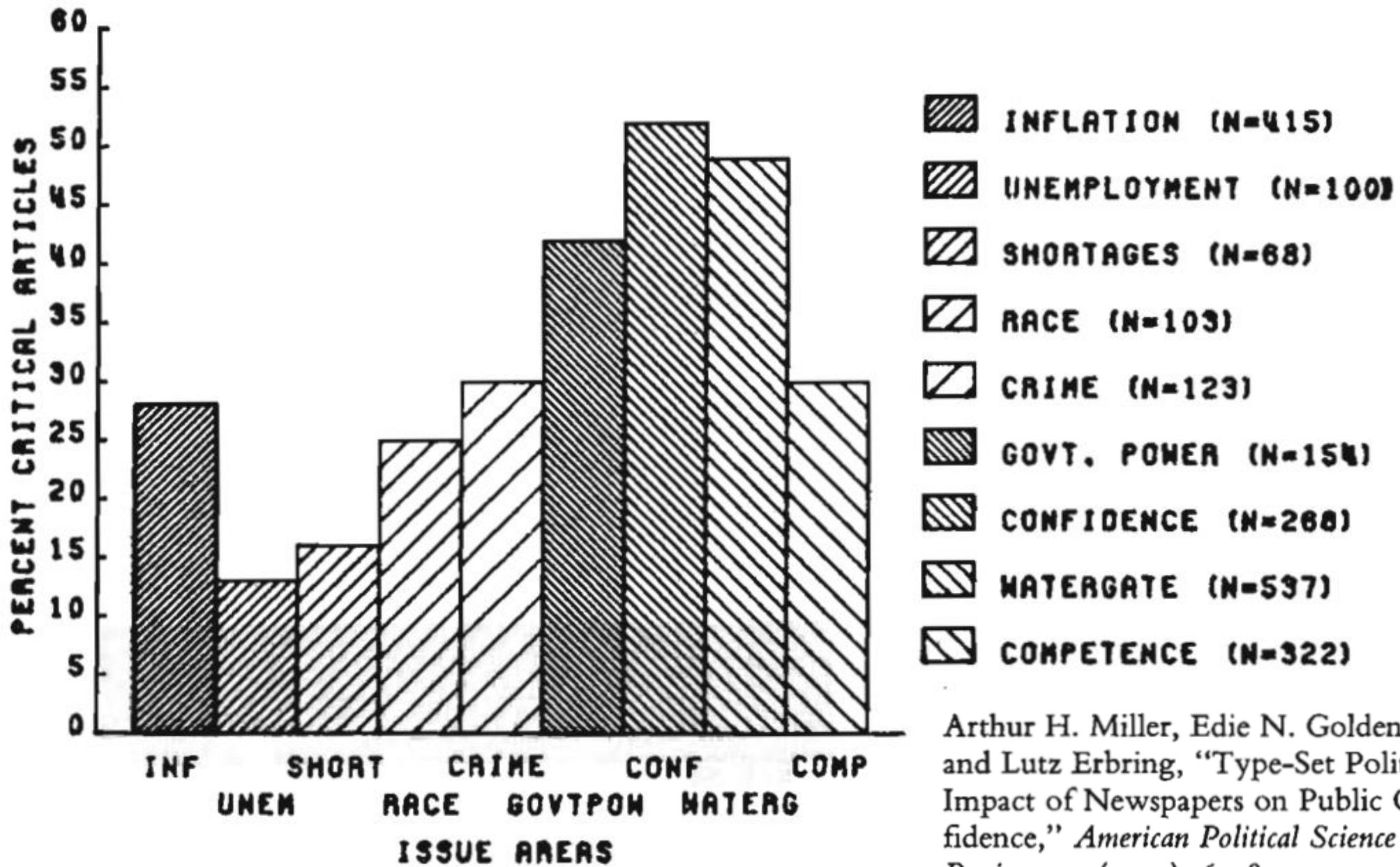
Источник: Edward Tufte. Visual Display of Quantitative Information, 2007. Chapter 13: Chart junk. P. 109.

*A. Average Probabilities of W from $N(1,1)$
with $n = 10$*

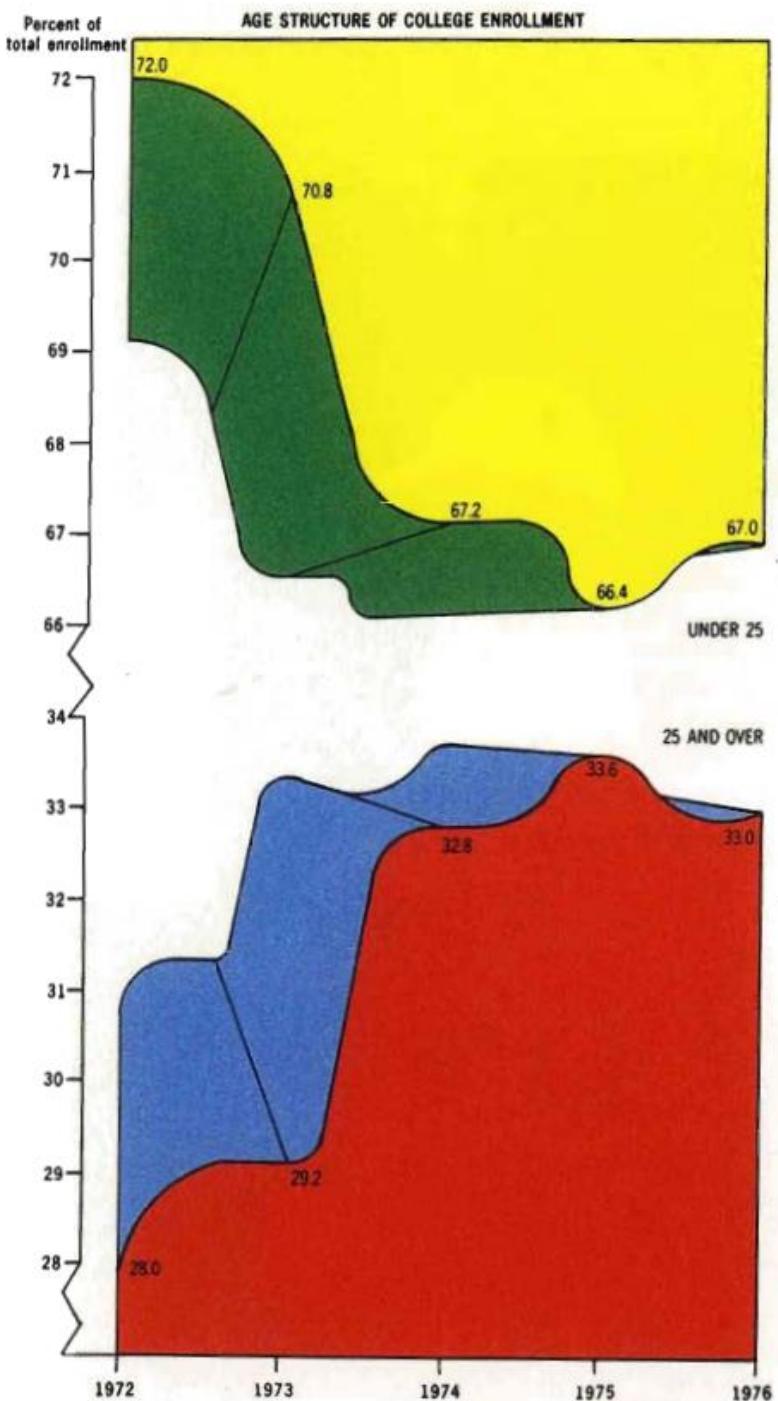
AVERAGE PROBABILITY



"JASA Style Sheet," *Journal of the American Statistical Association*, 71 (March 1976), 260–261.



Arthur H. Miller, Edie N. Goldenberg,
and Lutz Erbring, "Type-Set Politics:
Impact of Newspapers on Public Con-
fidence," *American Political Science
Review*, 73 (1979), 67-84.



Источник: Edward Tufte.
Visual Display of Quantitative Information, 2007. Chapter 13: Chart junk. P. 112.

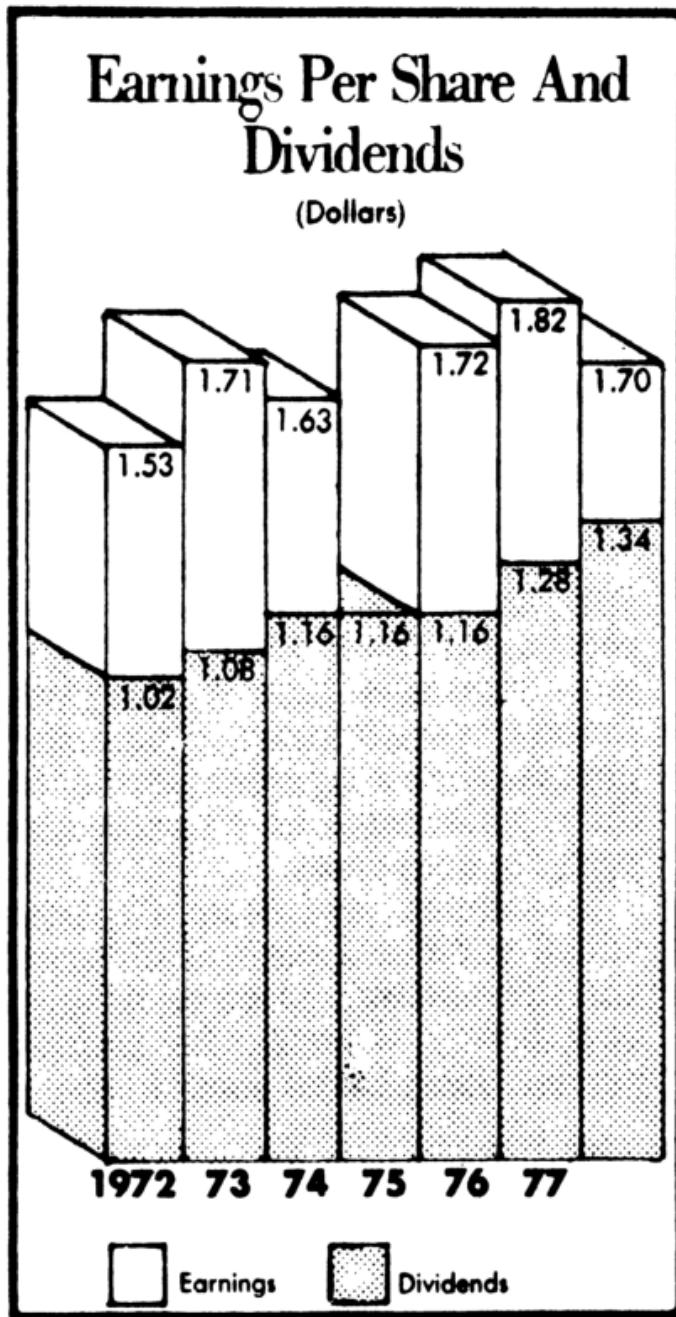


Figure 23. An extra dimension confuses even the grapher
(© 1979, The Washington Post).

Источник: Howard Wainer.
How to Display Data Badly. The American Statistician , Vol. 38, No. 2 (May, 1984), pp. 137-147.

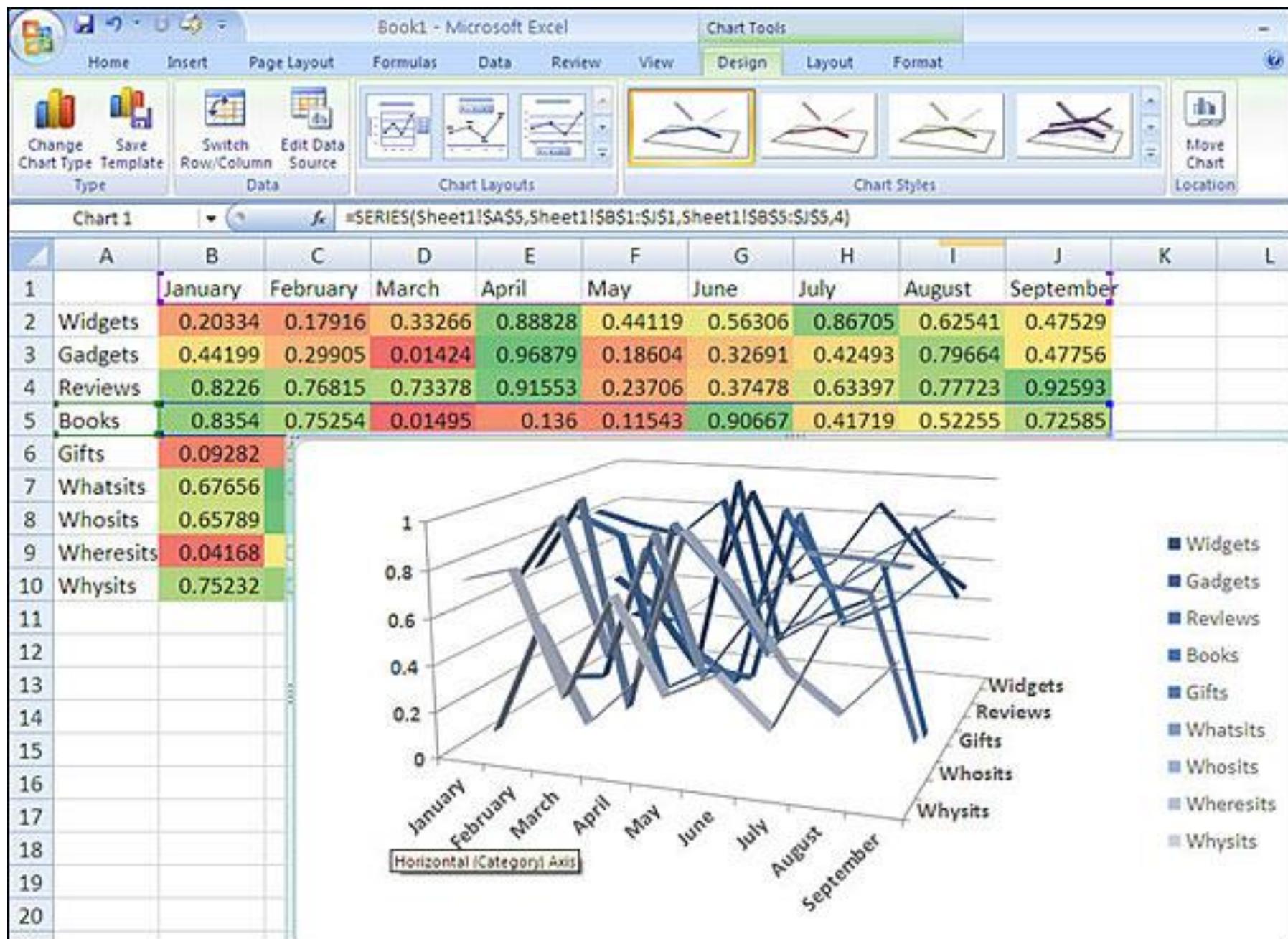
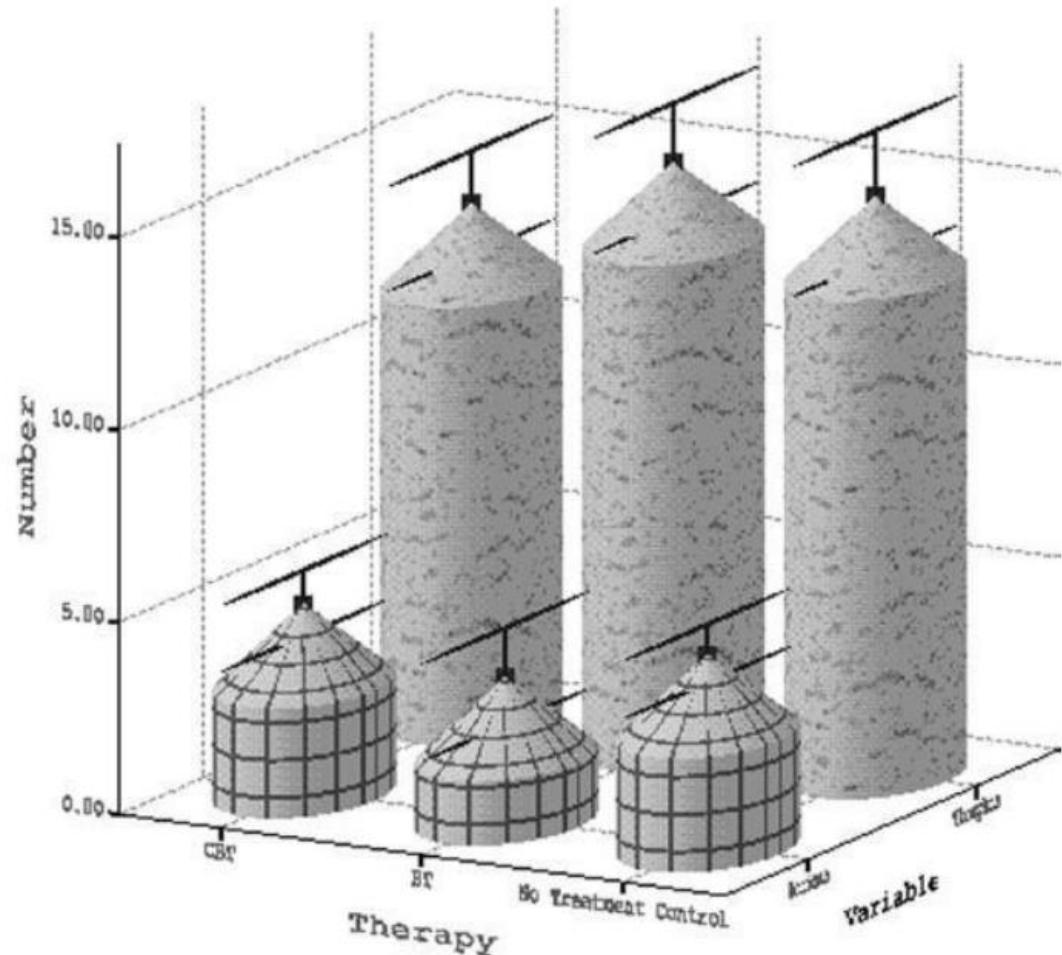


График курильщика

Error Bars show 95.0 % CI of Mean

Bars show Means



Источник: Field et al. Discovering Statistics Using R. 2012. P. 118-119.

График курильщика

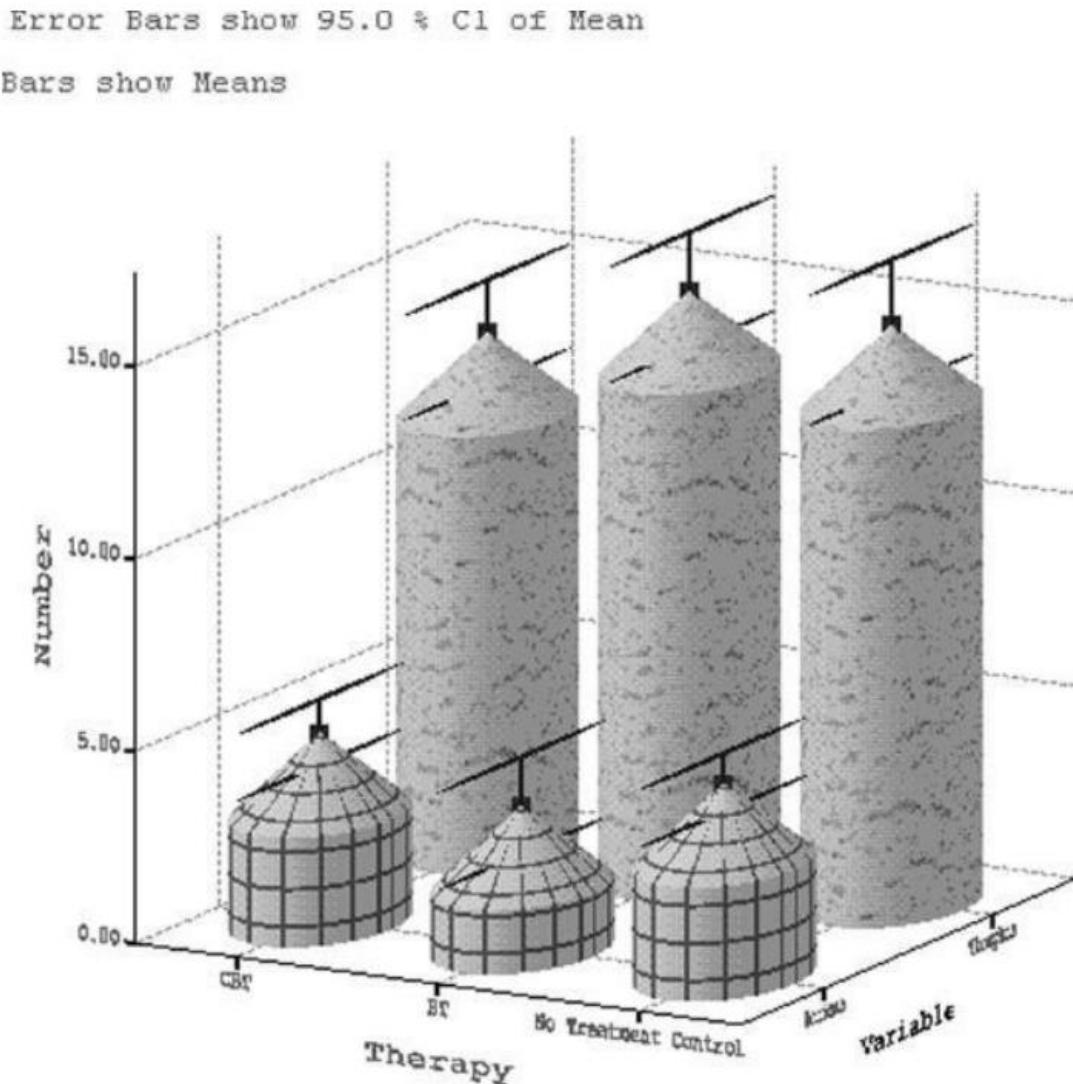
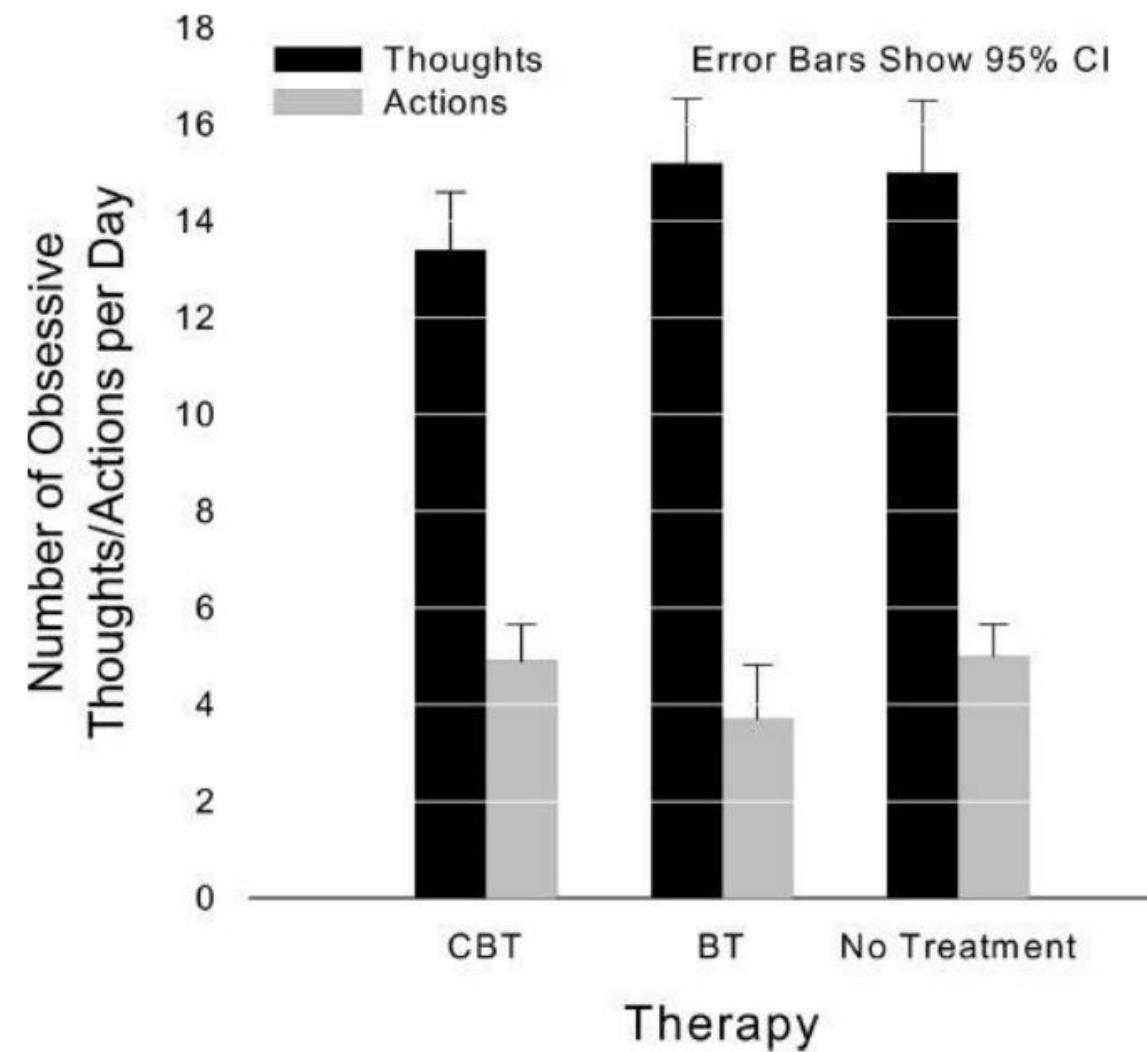


График здорового учёного



Источник: Field et al. Discovering Statistics Using R. 2012. P. 118-119.

Ошибки при создании графиков

- Лишние графические элементы (**chartjunk**): штриховка, 3D, тени, заливка...
- **Неправильный масштаб осей**

Public and Private Elementary Schools
Selected Years 1929-1970

□ Public
■ Private

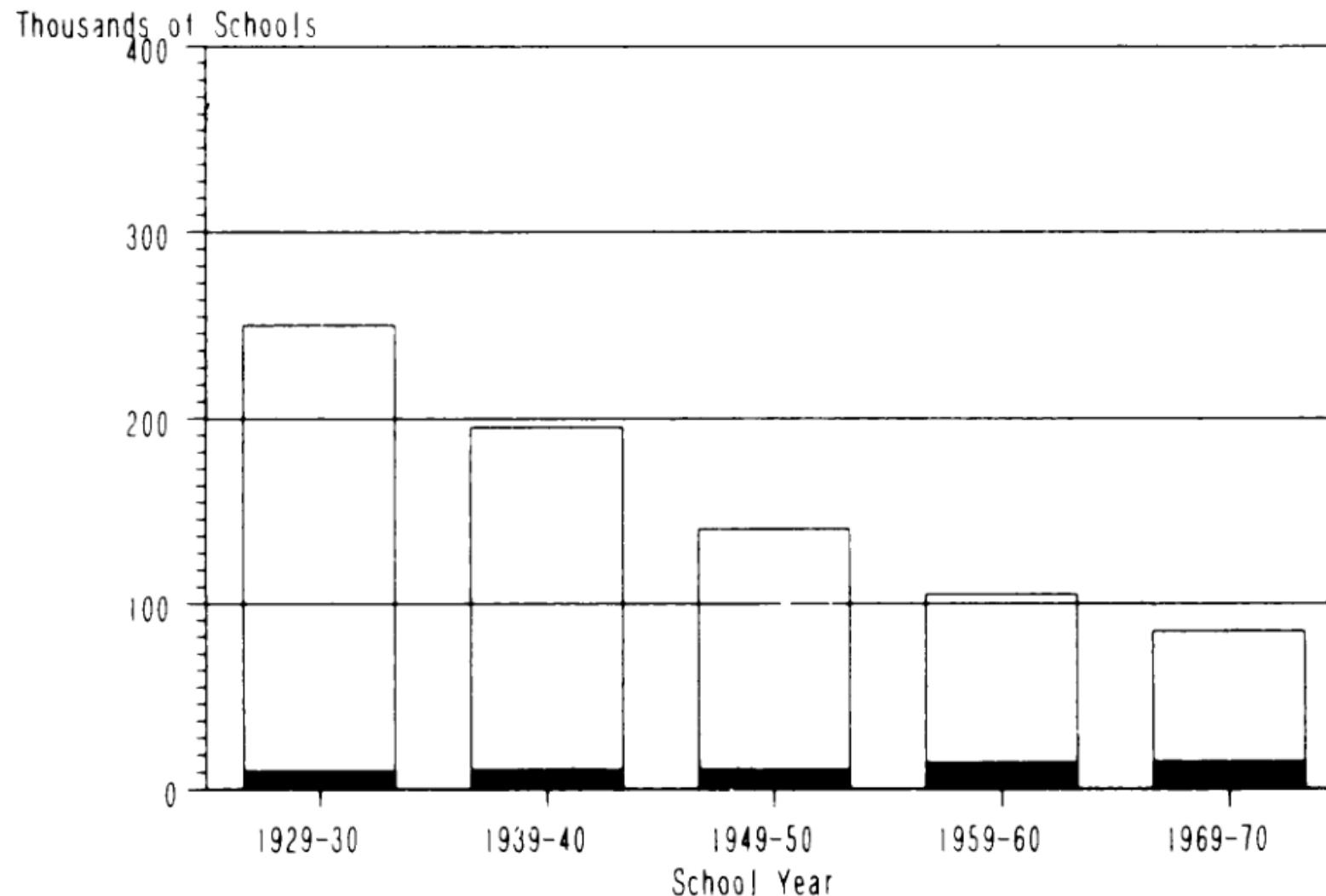
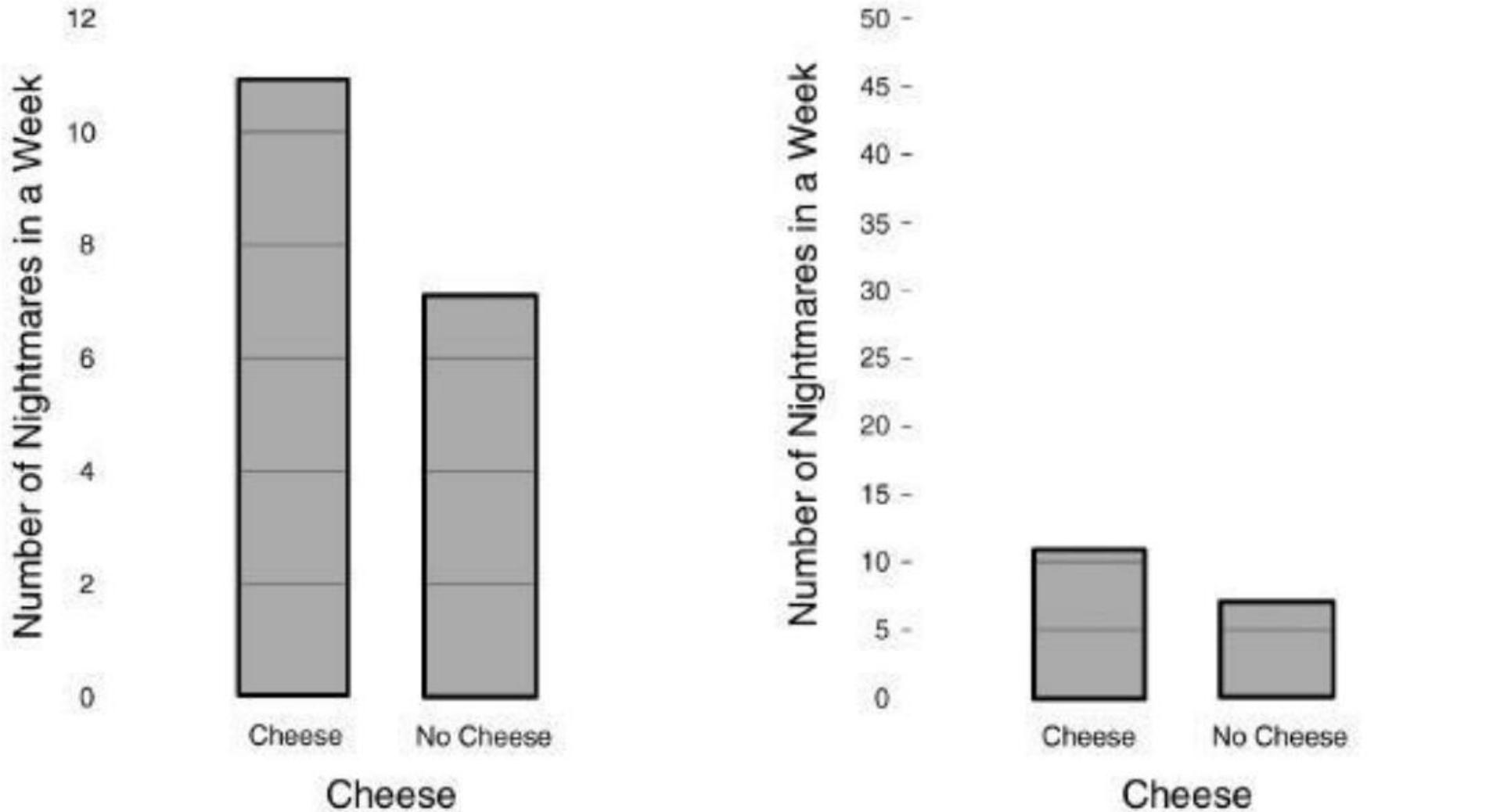


Figure 4. Hiding the data in the scale (from SI3).

Источник: Howard Wainer.
How to Display Data Badly. The American Statistician , Vol. 38, No. 2 (May, 1984), pp. 137-147.

Вызывает ли сыр перед сном кошмары?



Источник: Field et al. Discovering Statistics Using R. 2012. P. 120.

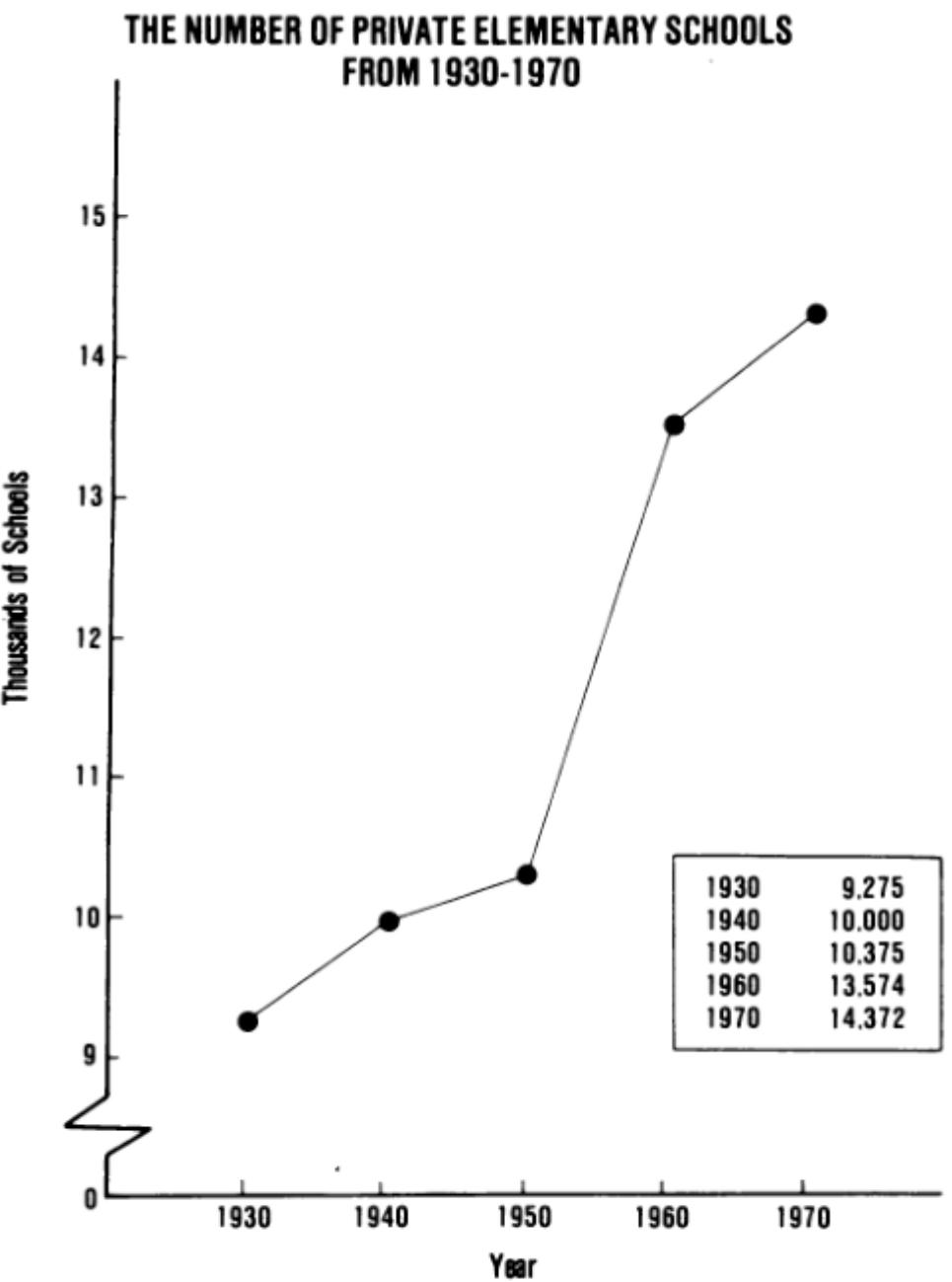


Figure 5. Expanding the scale and showing the data in Figure 4 (from SI3).

Источник: Howard Wainer.
How to Display Data Badly. The American Statistician , Vol. 38, No. 2 (May, 1984), pp. 137-147.

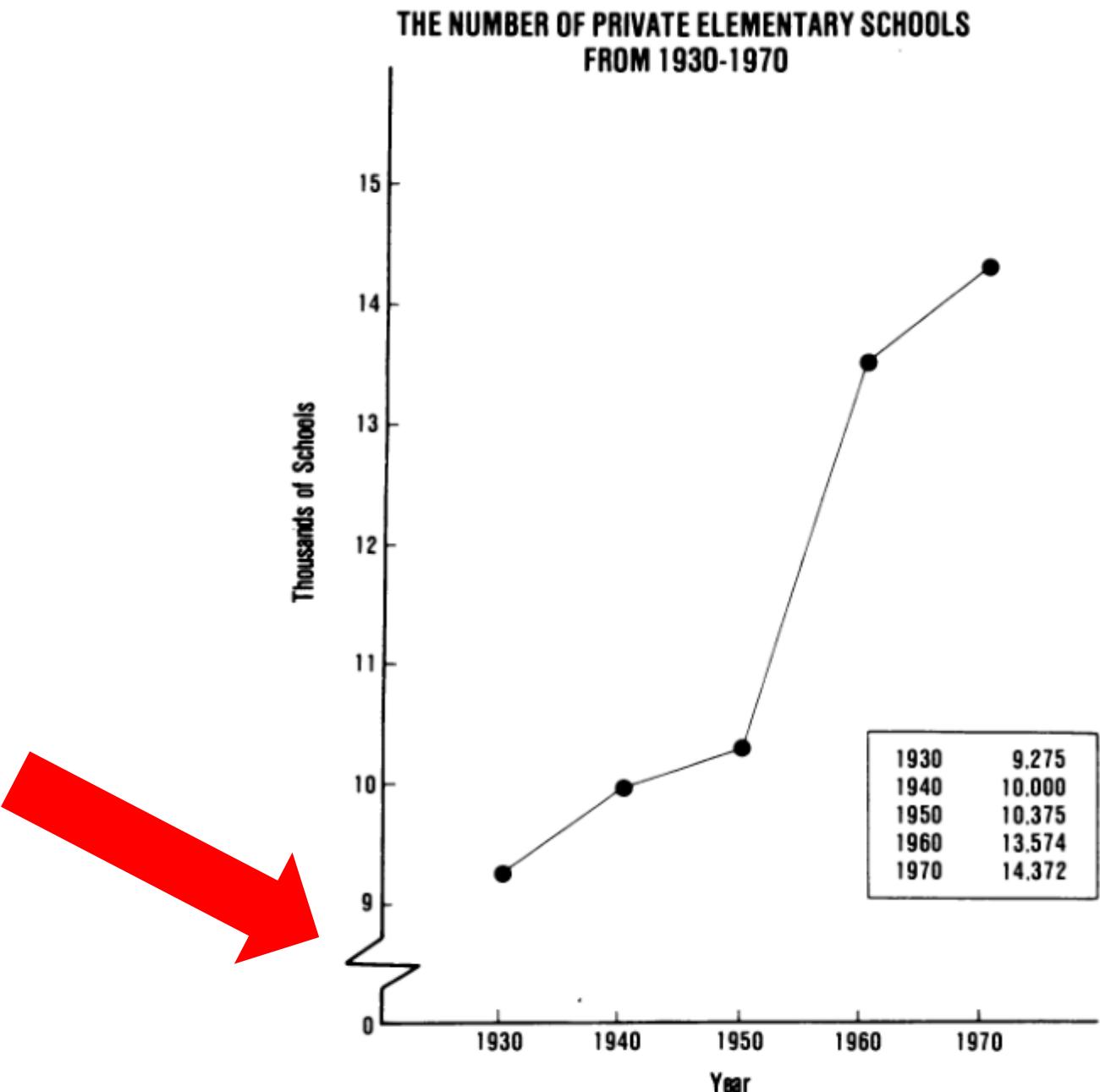


Figure 5. Expanding the scale and showing the data in Figure 4 (from SI3).

Источник: Howard Wainer.
How to Display Data Badly. The American Statistician , Vol. 38, No. 2 (May, 1984), pp. 137-147.

Ошибки при создании графиков

- Лишние графические элементы (**chartjunk**): штриховка, 3D, тени, заливка...
- Неправильный масштаб осей
- Отсутствующие/непонятные подписи и легенда

Ошибки при создании графиков

- Лишние графические элементы (**chartjunk**): штриховка, 3D, тени, заливка...
- Неправильный масштаб осей
- Отсутствующие/непонятные подписи и легенда
- Использование круговых диаграмм

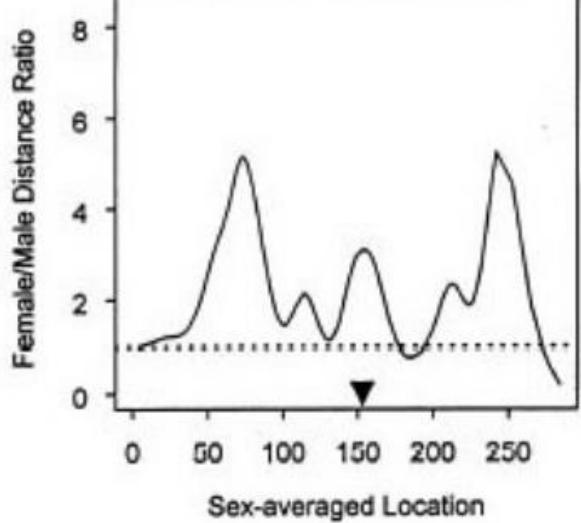


Карл Броман: топ-10 худших научных графиков

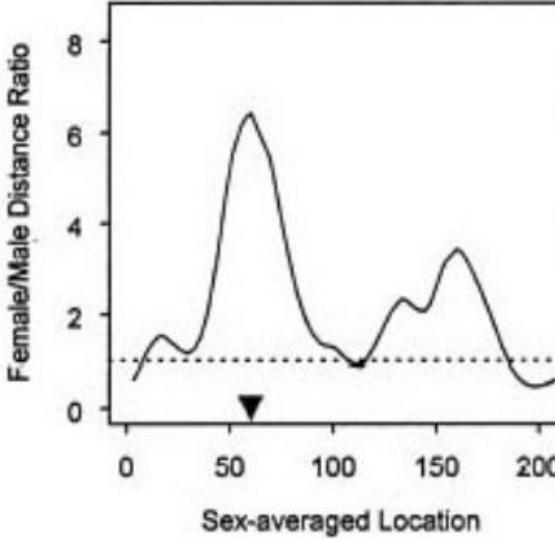
WWW: https://www.biostat.wisc.edu/~kbroman/topten_worstgraphs/

10

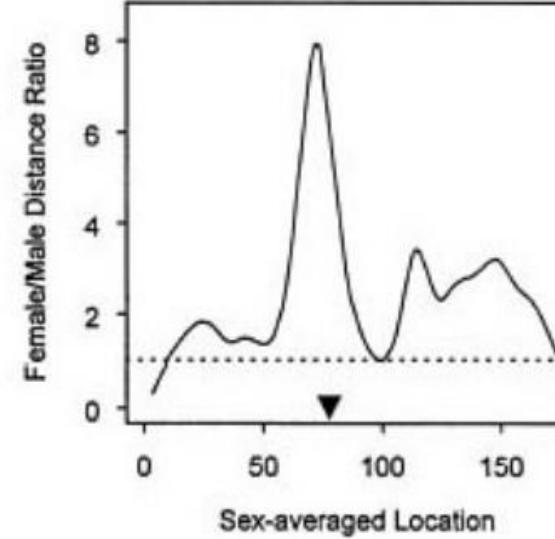
Chromosome 1



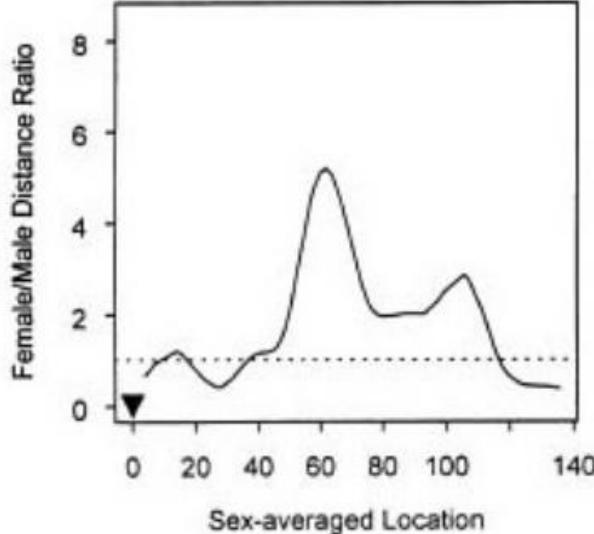
Chromosome 4



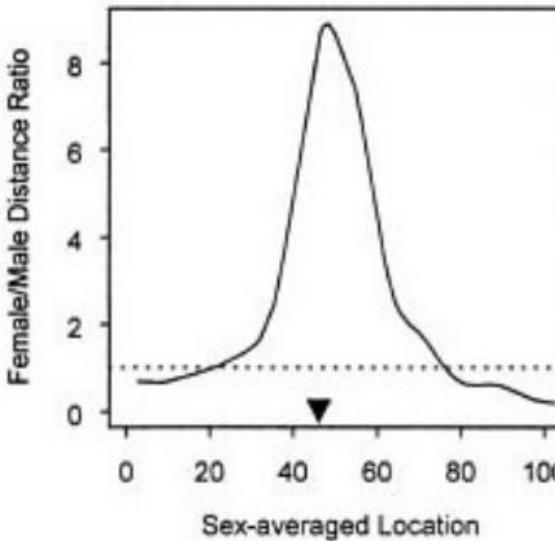
Chromosome 7



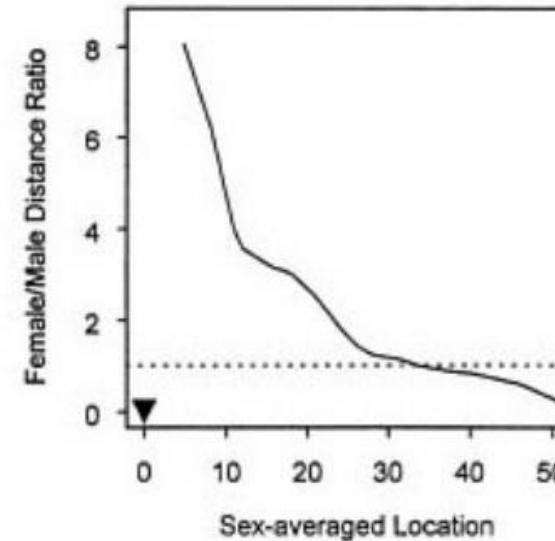
Chromosome 14

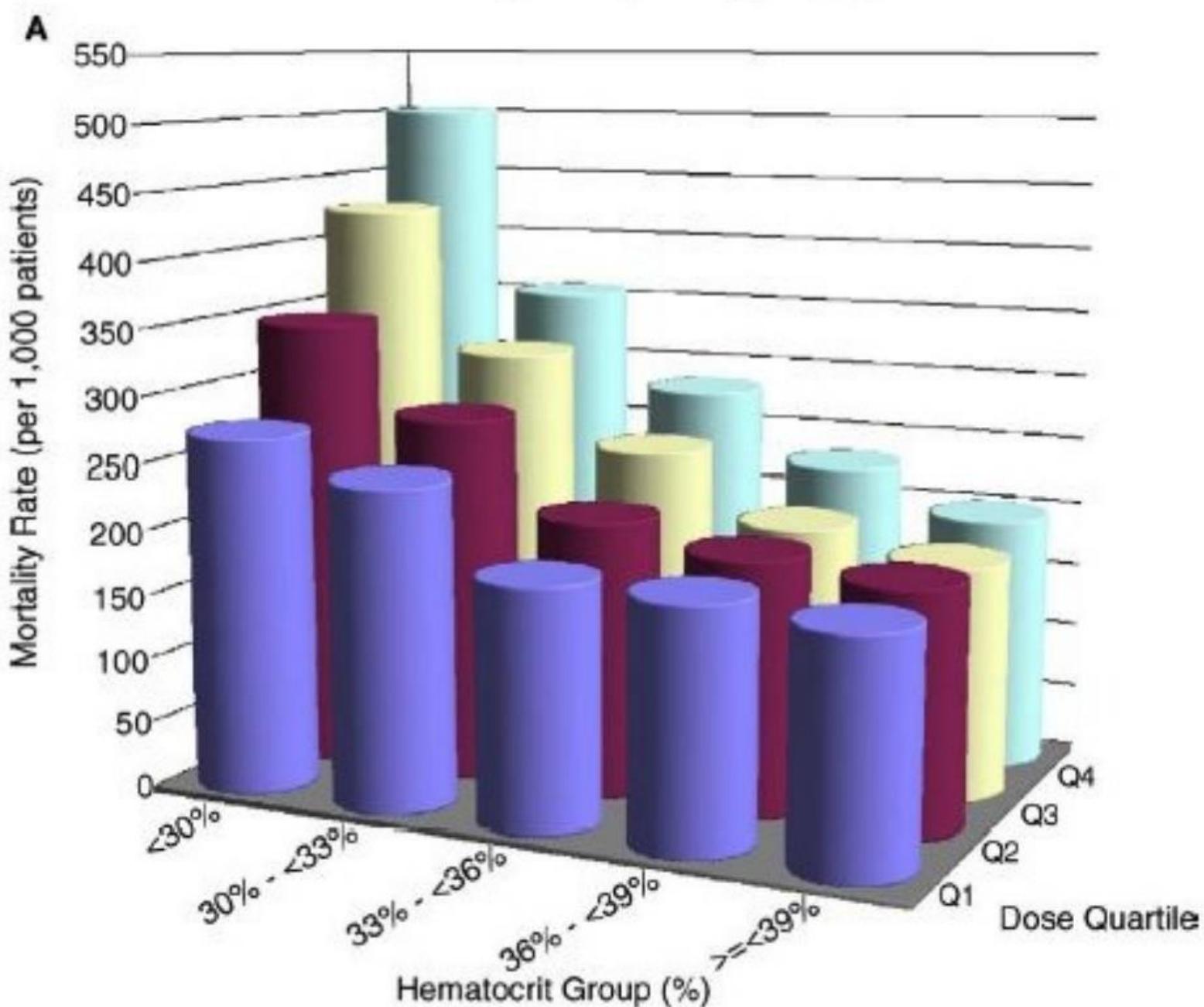


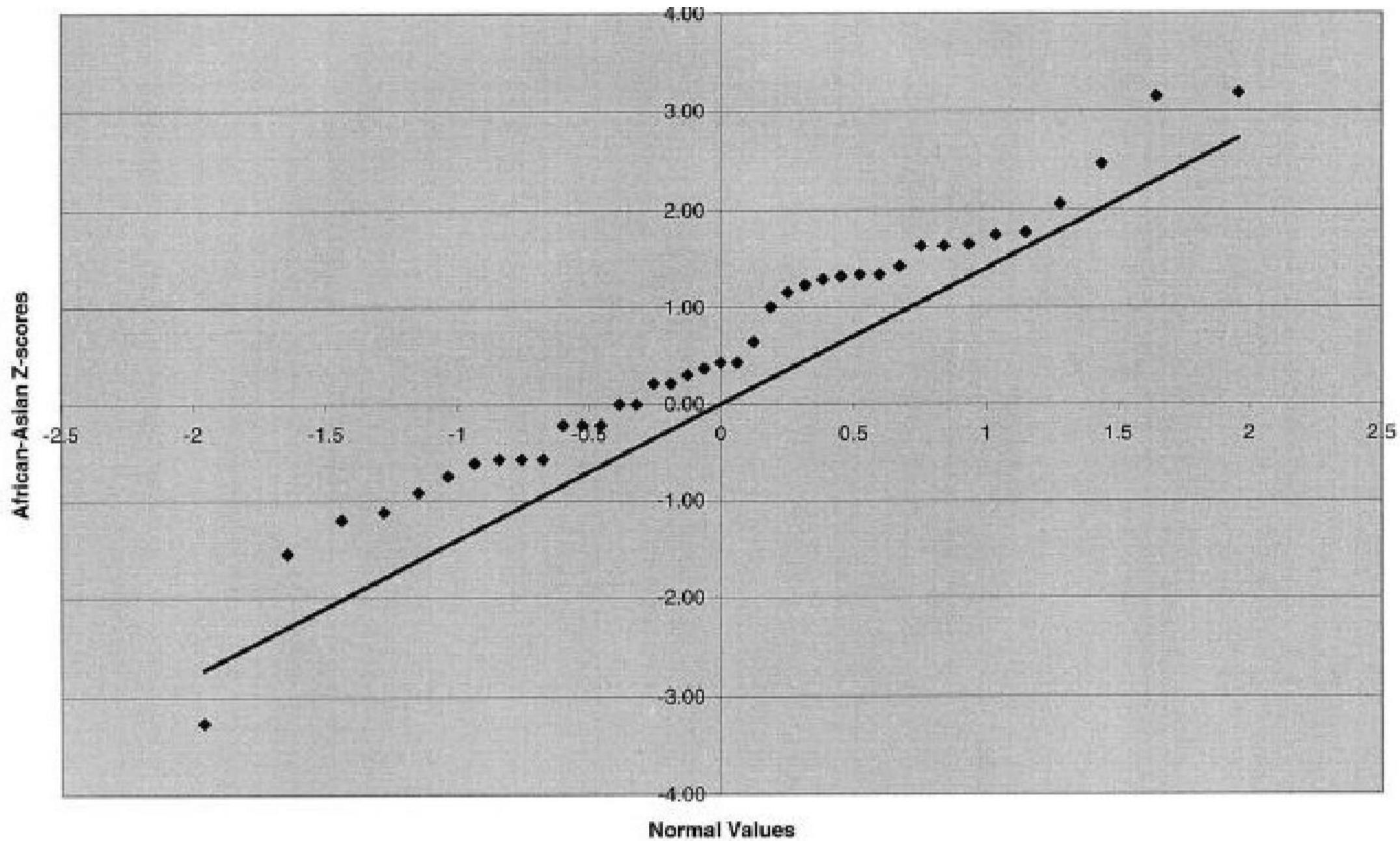
Chromosome 19

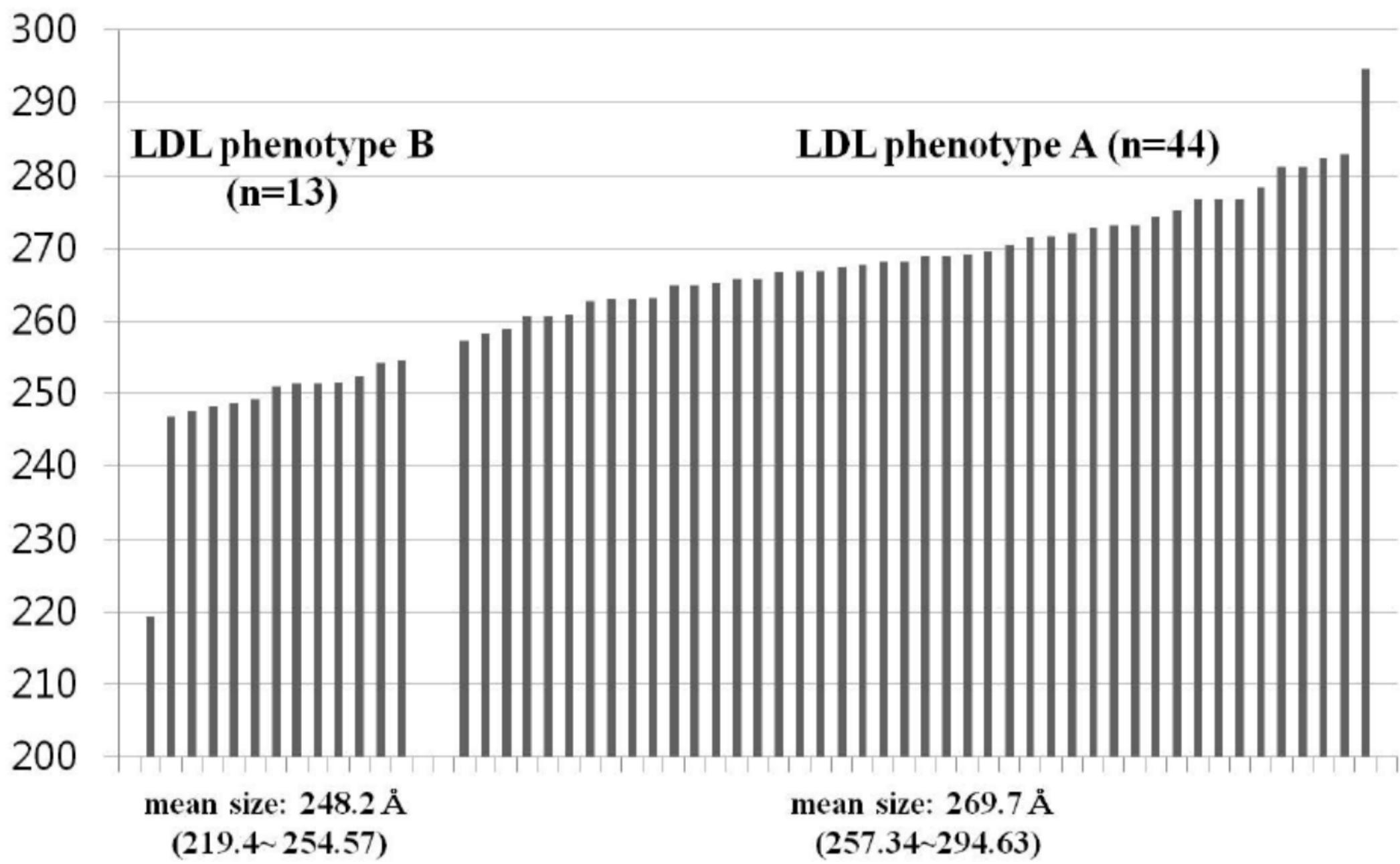


Chromosome 21

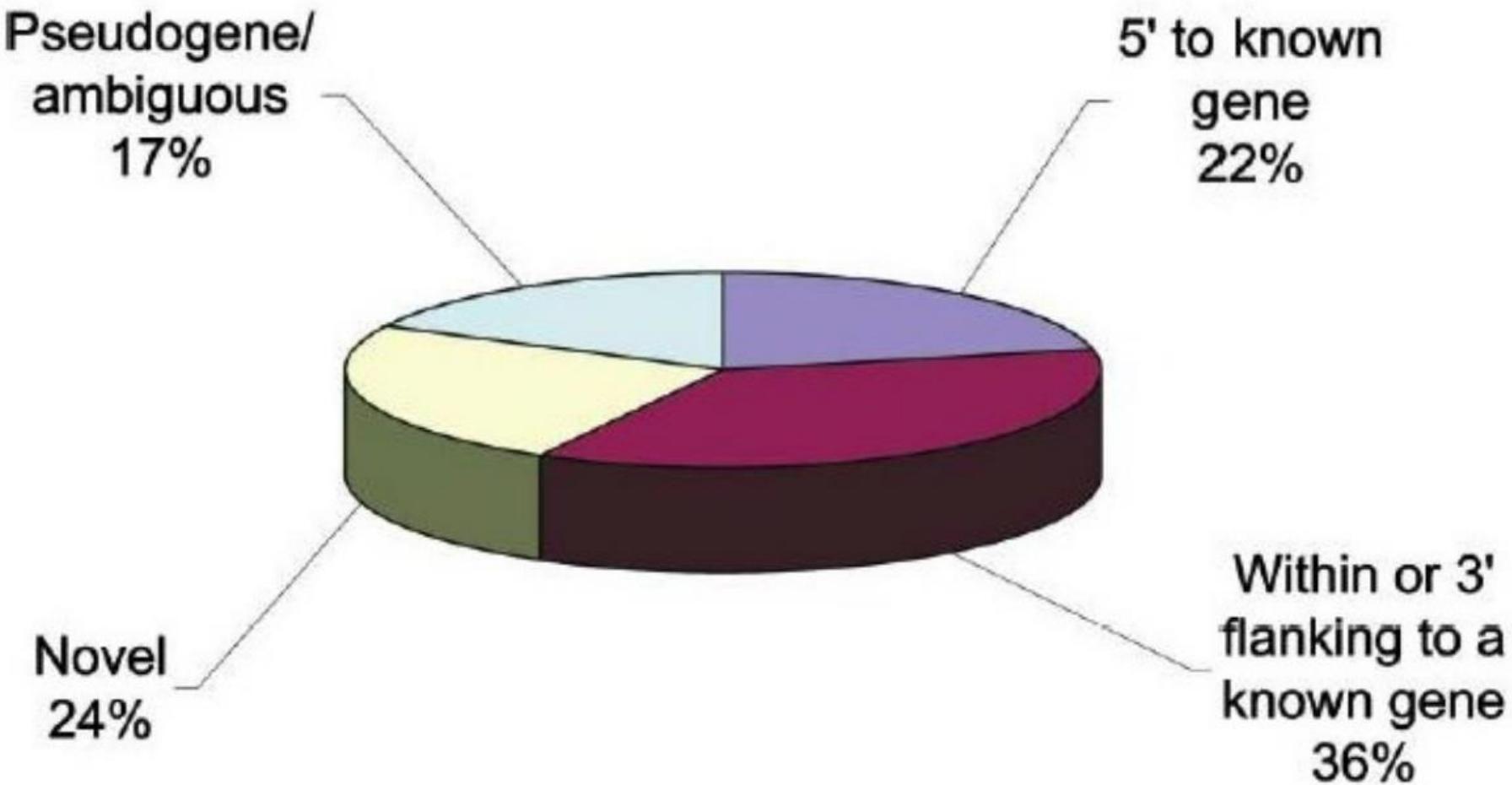








Distribution of All TFBS Regions



866 Total TFBS Regions

luciferase activity

100
75
50
25
0

□ p53^{+/+}
■ p53^{-/-}

untt

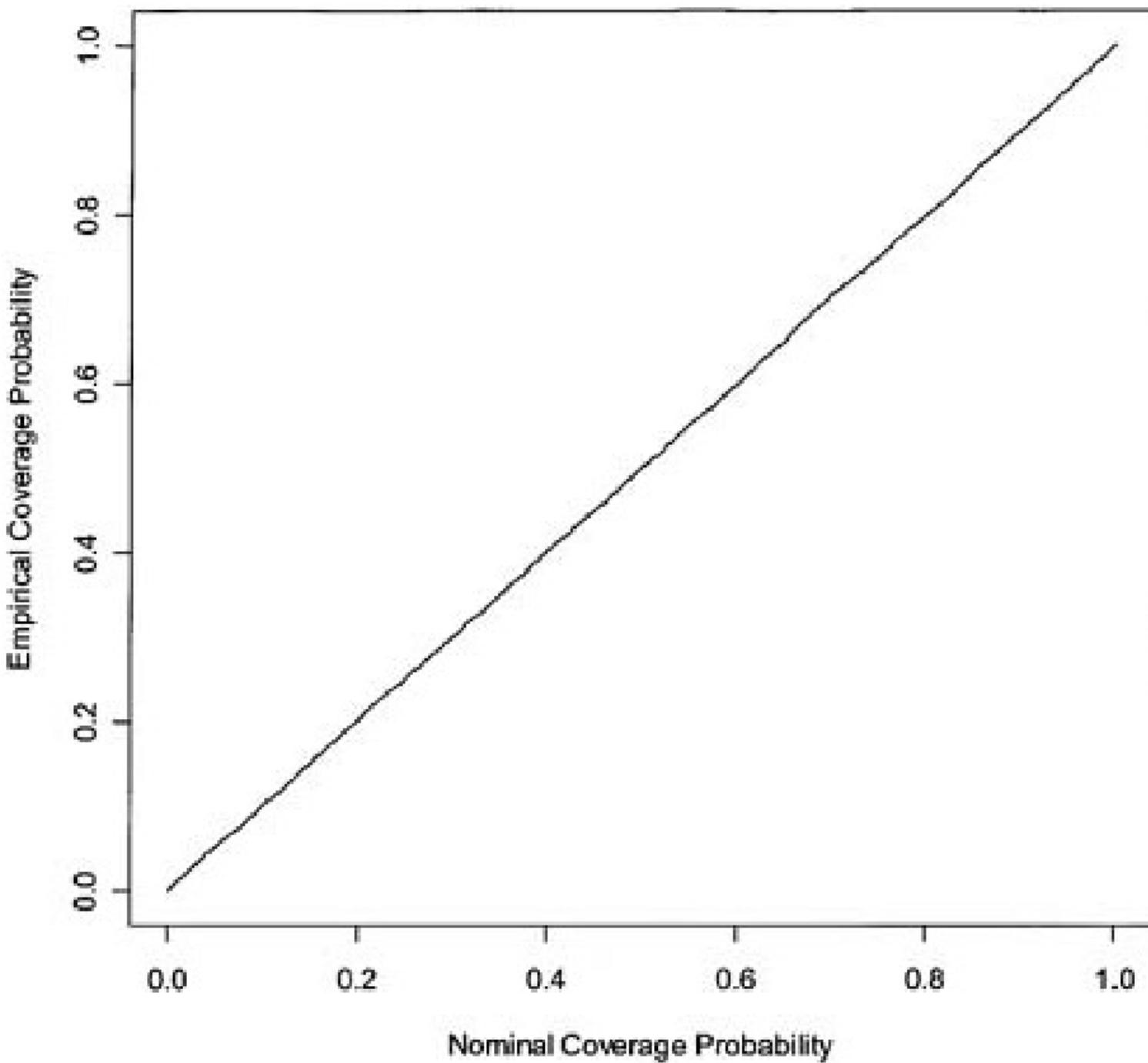
IFN

dsRNA

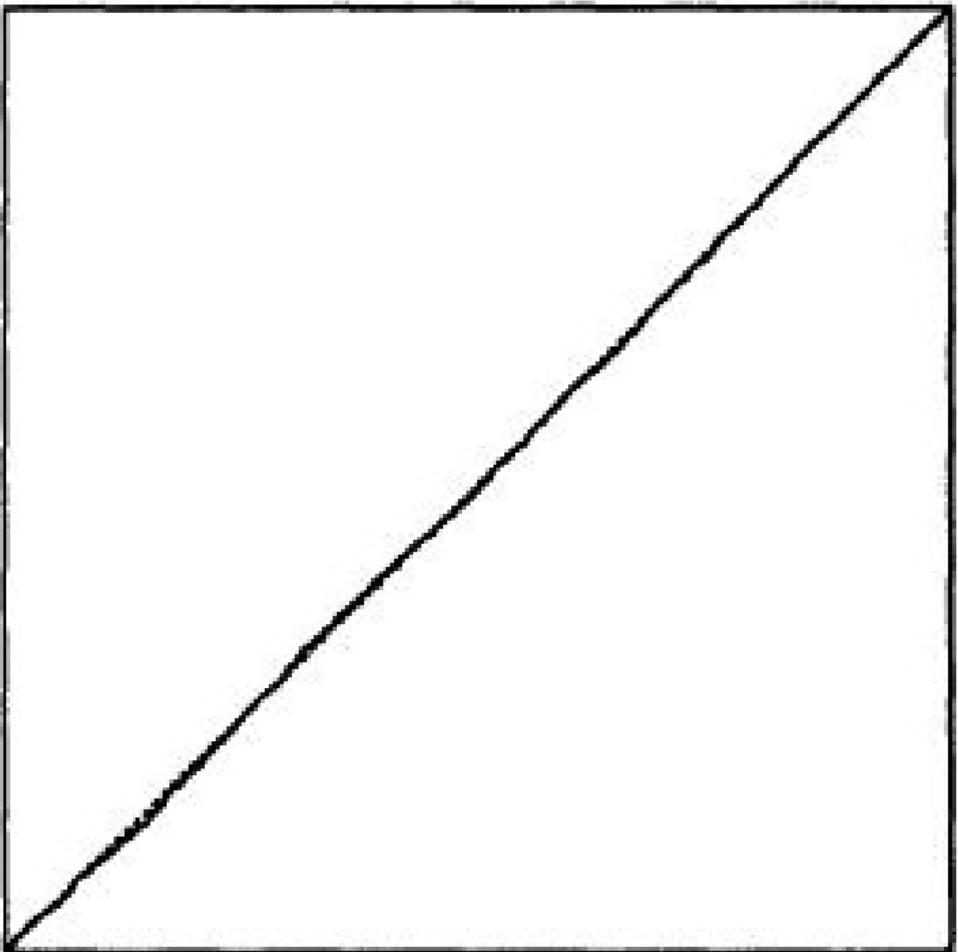
SV

5

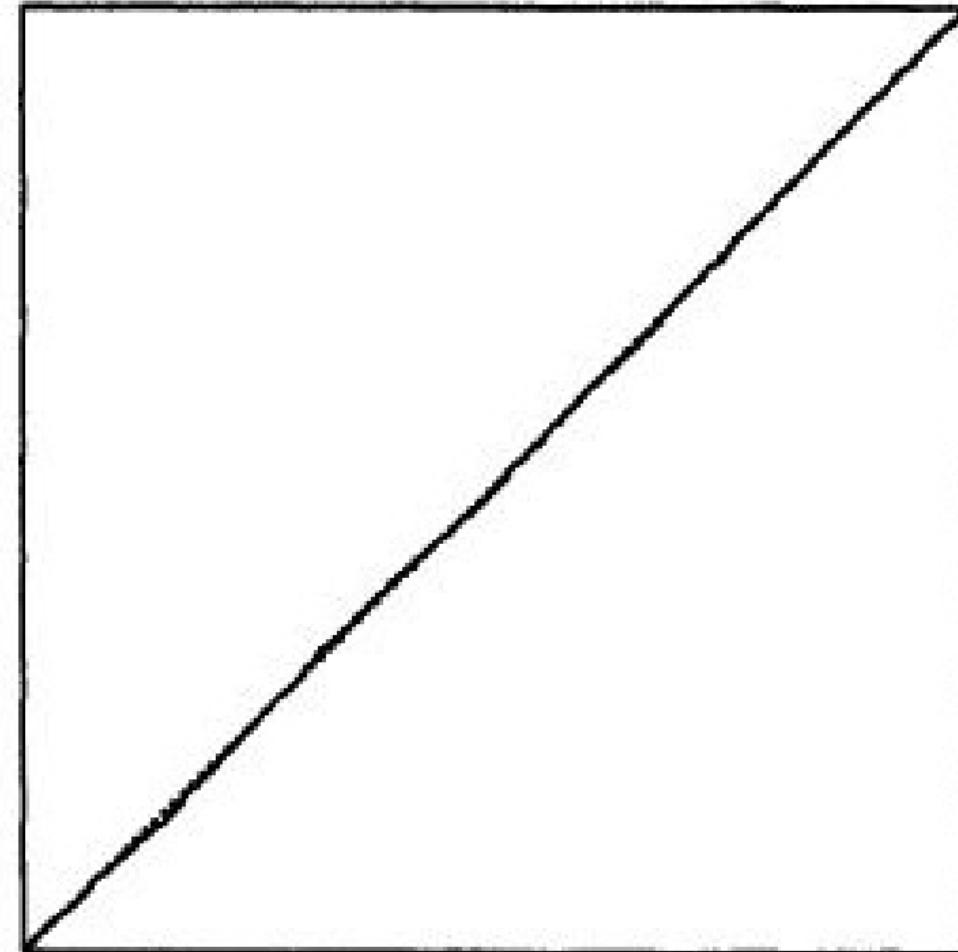
4



3

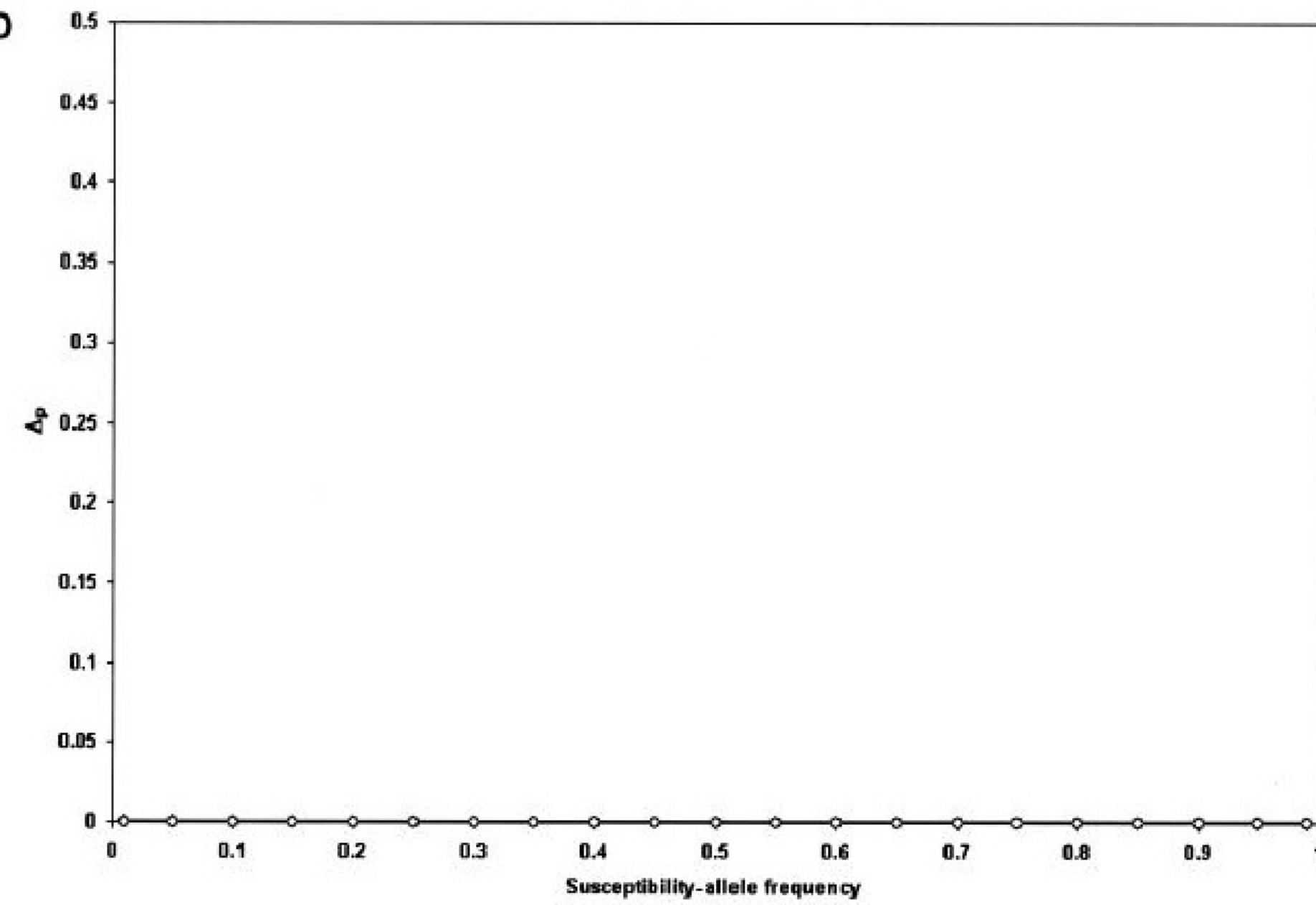


(a)



(b)

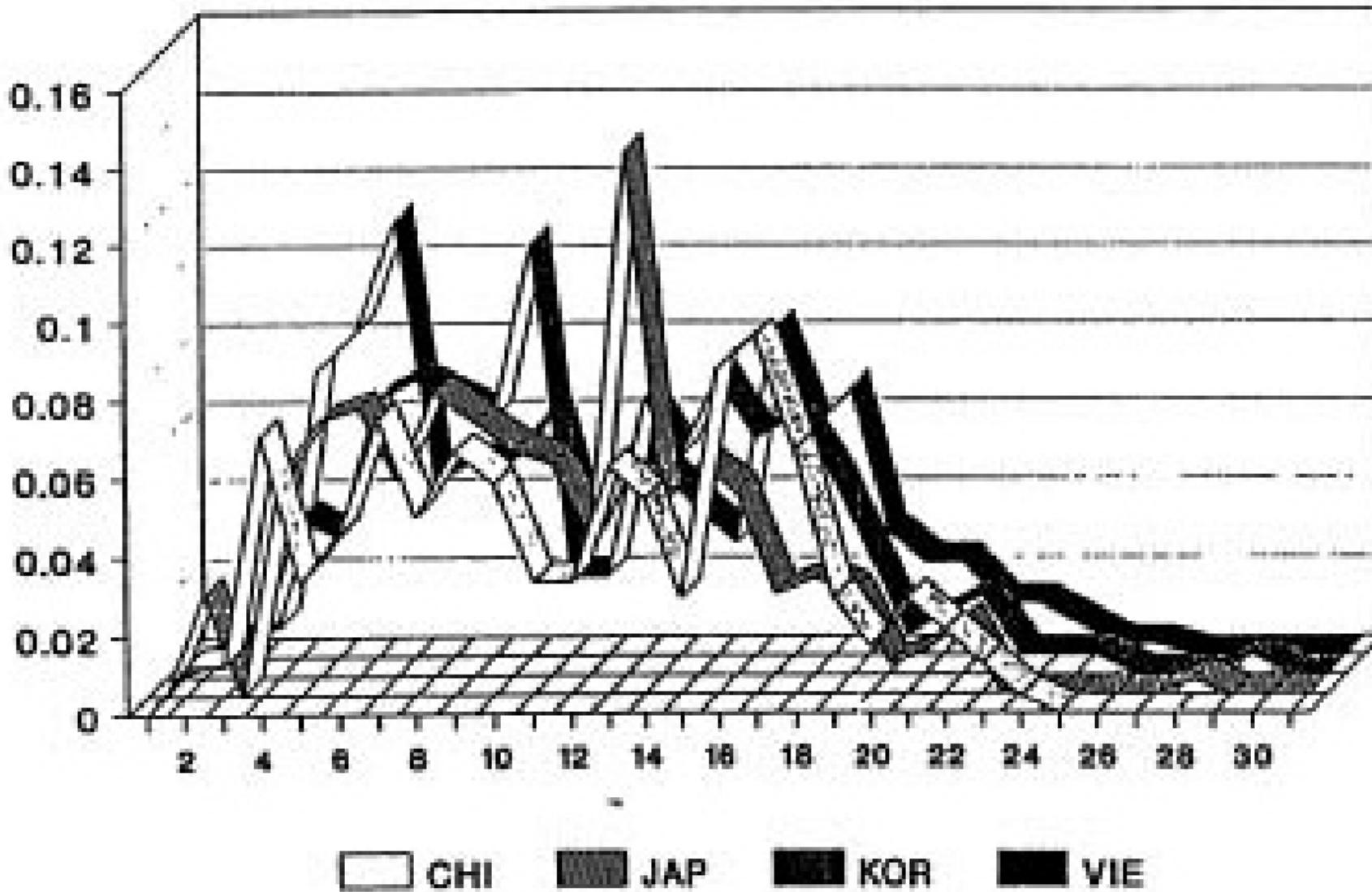
2



1

B

BINNED FREQUENCY DATA - D10S28
CHINESE, JAPANESE, KOREAN, VIETNAMESE



Спасибо! Вопросы?

Ссылка на презентацию:

alexeyknorre.ru/courses/datavis2016/datavis-2.pdf

Алексей Кнорре

Email: aknorre@eu.spb.ru

WWW: alexeyknorre.ru