

# Корреляция

Математика приводит нас к дверям истины,  
но самих дверей не открывает.

*В. Ф. Одоевский*

# Изучение связей признаков

После того, как (с помощью методов из темы 6) объекты разбиты на однородные группы (кластеры), возникает задача изучения взаимосвязей признаков внутри отдельного кластера.

На практике чаще всего встречаются следующие два вида зависимостей: а) объекты образуют “облако” эллиптического типа (рис. 1а), б) объекты располагаются в окрестности некото-

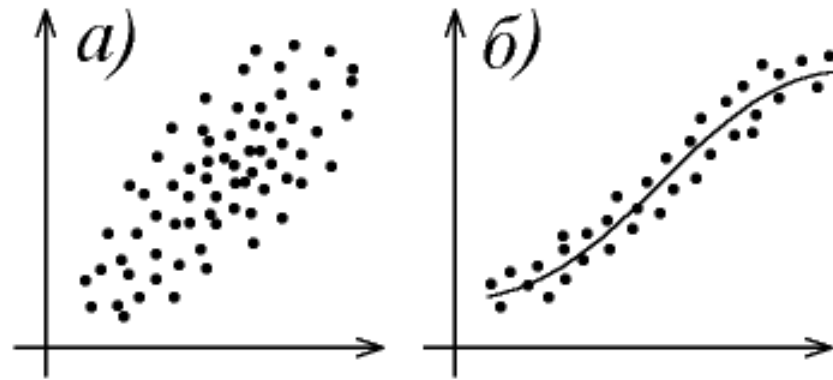


Рис. 1

рой кривой (поверхности) (рис. 1б). В случае а) оба признака являются “полноценными” случайными величинами, и изучению подлежит уровень зависимости (корреляции) между ними. Случай б) соответствует “функциональной” зависимости между признаками, испорченной шумом. Зависимости первого вида изучаются в этой теме методами *корреляционного анализа*. Методы, позволяющие во втором случае построить интересующую исследователя кривую (поверхность), относятся к так называемому *регрессионному анализу*, обсуждаемому в теме 8.

# Коэффициент корреляции

Если величины  $\xi$  и  $\eta$  не являются независимыми, то интерес представляет мера зависимости между ними. Простейшей такой мерой служит *коэффициент корреляции*

$$\rho(\xi, \eta) = \mathbf{cov}(\xi, \eta) / \sqrt{\mathbf{D}\xi \mathbf{D}\eta},$$

где *ковариация*  $\mathbf{cov}(\xi, \eta)$  определяется формулой

$$\mathbf{cov}(\xi, \eta) = \mathbf{M}(\xi - \mathbf{M}\xi)(\eta - \mathbf{M}\eta).$$

Важнейшим свойством коэффициента корреляции является его *инвариантность по отношению к линейным преобразованиям* случайных величин. Точнее: при  $ac > 0$

$$\rho(a\xi + b, c\eta + d) = \rho(\xi, \eta).$$

Известно, что для любых случайных величин  $|\rho(\xi, \eta)| \leq 1$ , причем для независимых величин  $\rho(\xi, \eta) = 0$ . Если  $\rho(\xi, \eta) = 1$ , то  $\eta = a\xi + b$ , где  $a > 0$ . Соответственно, если  $\rho(\xi, \eta) = -1$ , то  $\eta = a\xi + b$ , где  $a < 0$ . Поэтому коэффициент корреляции можно считать мерой линейной связи между  $\xi$  и  $\eta$ .

# Выборочная корреляция

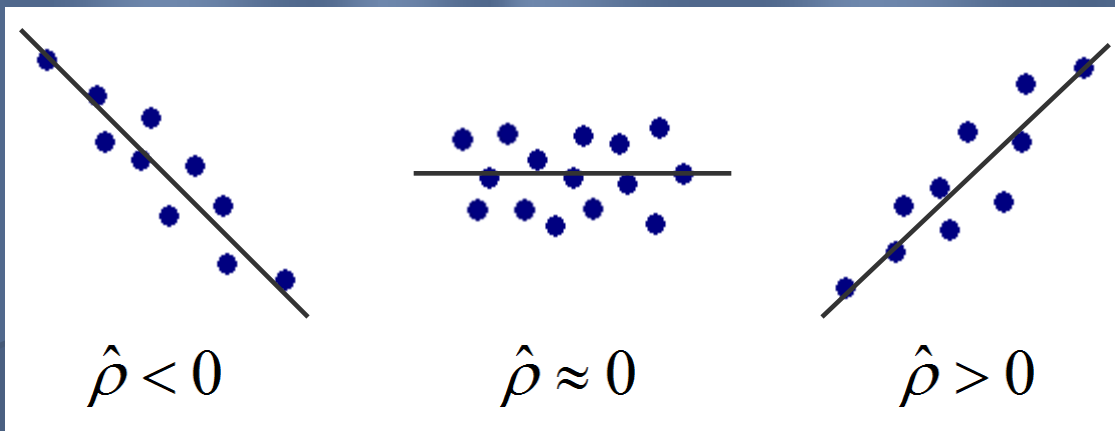
Естественной оценкой для ковариации  $\mathbf{cov}(\xi, \eta)$  является *выборочная ковариация*

$$\hat{\sigma} = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta}) = \frac{1}{n} \sum_{i=1}^n \xi_i \eta_i - \bar{\xi} \bar{\eta}.$$

Оценить коэффициент корреляции  $\rho(\xi, \eta)$  можно с помощью *выборочного коэффициента корреляции Пирсона*

$$\hat{\rho} = \frac{\hat{\sigma}}{S_{\xi} S_{\eta}}, \quad \text{где } S_{\xi}^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2, \quad S_{\eta}^2 = \frac{1}{n} \sum_{i=1}^n (\eta_i - \bar{\eta})^2.$$

Здесь  $n$ , а не  $n - 1$ , чтобы обеспечить выполнение неравенства  $|\hat{\rho}| \leq 1$

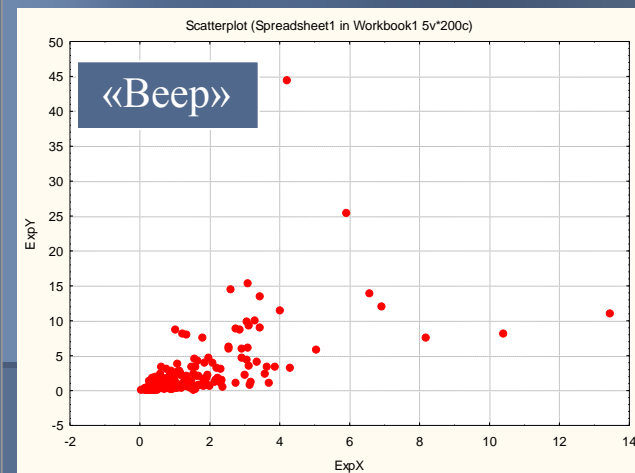
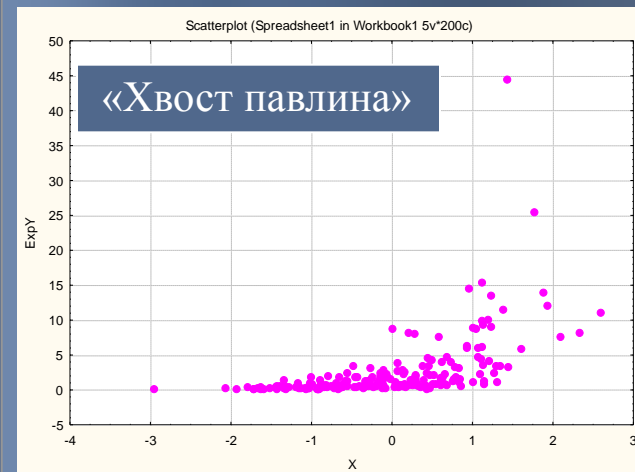
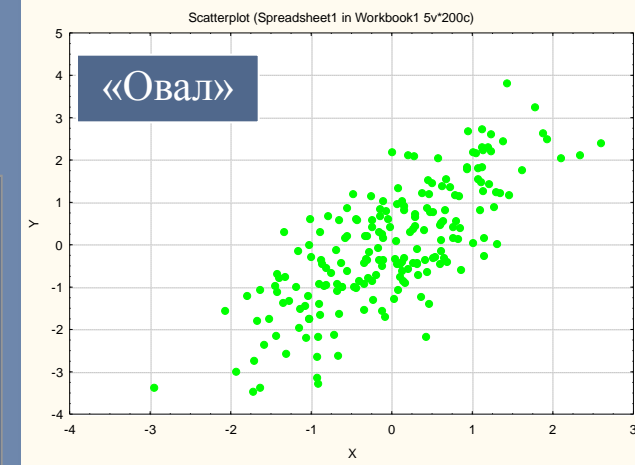


# Логарифмирование

- Прежде, чем вычислять коэффициент корреляции Пирсона, необходимо прологарифмировать признаки, имеющие тяжёлый правый «хвост»
- Диаграмма **справа вверху** демонстрирует случай, когда логарифмирование не требуется, **справа в центре** — когда надо прологарифмировать признак по оси  $Y$ , **справа внизу** — когда необходимо прологарифмировать оба признака
- Альтернативой логарифмированию служит использование рангового **коэффициента корреляции Спирмена**

$$\hat{\rho}_S = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (R_i - S_i)^2$$

- Этот коэффициент инвариантен относительно преобразования наблюдений с помощью любой возрастающей функции



# Практическое задание 1

## (одна из ловушек корреляционного анализа)

В файле CarSales.txt содержатся, в частности, данные о числе миль, проезжаемых на одном галлоне [3,78 л] бензина (признак Miles\_gal) и о размерах продаж Sales (в тыс. штук). Есть ли значимая корреляционная связь между этими признаками?

- 1) Импортируйте файл CarSales.txt в RStudio, запишите значения признаков Miles\_gal и Sales в переменные x и y соответственно
- 2) Постройте гистограмму для y и убедитесь, что его распределение очень сильно перекошено вправо
- 3) Поэтому прологарифмируйте y и ещё раз построьте гистограмму
- 4) Постройте диаграмму рассеяния признаков x и y командой plot
- 5) Замените 2 выделяющихся наблюдения на пропуски NA (примените ifelse) и ещё раз построьте диаграмму рассеяния признаков x и y
- 6) Вычислите коэффициент корреляции **Пирсона** функцией cor с аргументом use="pairwise.complete.obs" и проверьте значимость корреляционной связи признаков x и y с помощью функции cor.test

# Продолжение исследования

- 7) Признак `Type` из файла `CarSales.txt`, содержащий сведения о типе автомобиля (0 — легковая машина, 1 — грузовик) запишите в вектор `t`
- 8) С помощью команды `ifelse` замените в `t` значения 0 на 4 (синий цвет) и значения 1 на 2 (красный цвет)
- 9) Объедините `x` и `y` в матрицу `m` командой `cbind` (от слов `column` и `bind`)
- 10) Постройте диаграмму рассеяния для `m` с раскраской в цвета из `t`
- 11) Проверьте значимость корреляционной связи для легковых машин и грузовиков в отдельности (используйте команды `subset` и `cor.test`)
- 12) Для легковых машин вычислите ранговый коэффициент корреляции **Спирмена** и определите фактический уровень значимости (*p-value*) отличия его от нуля (`cor.test` с аргументом `method="spearman"`)

# Степень корреляционной связи

## Неправильная классификация

- $\hat{\rho}_s = 0,3$  — ~~слабая~~
- $\hat{\rho}_s = 0,5$  — значимая
- $\hat{\rho}_s = 0,7$  — сильная
- $\hat{\rho}_s > 0,9$  — очень сильная

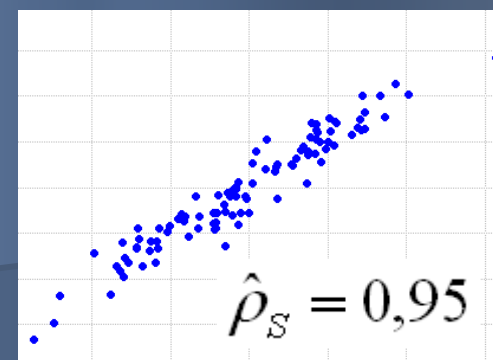
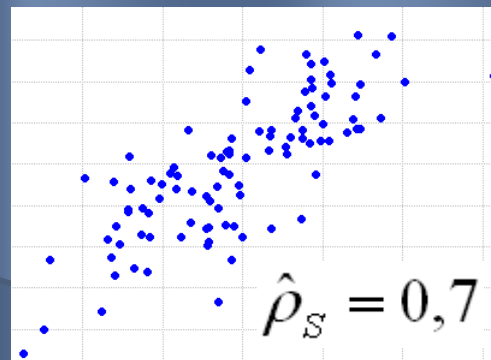
$$\hat{\rho}_s \sqrt{n} \rightarrow Z \sim N(0, 1)$$

При этом не учитывается размер выборки  $n$ . Например, при  $n = 20$  значимым будет коэффициент корреляции 0,45. При  $n = 100$  значимым окажется даже коэффициент корреляции 0,2.

## Правильная классификация

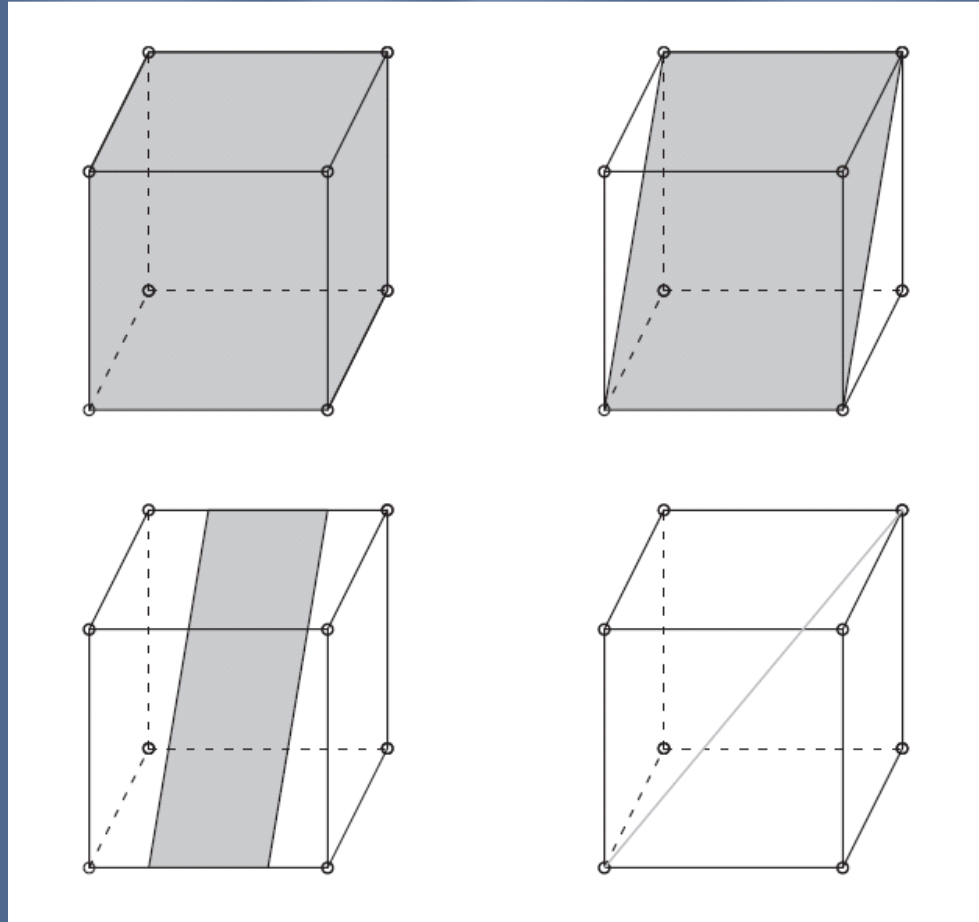
- $p \geq 0,05$  — незначимая
- $p < 0,05$  — значимая
- $p < 0,05$  и  $\hat{\rho}_s = 0,7$  — сильная
- $p < 0,05$  и  $\hat{\rho}_s > 0,9$  — очень сильная (при этом один признак фактически дублирует другой)

## Диаграммы рассеяния



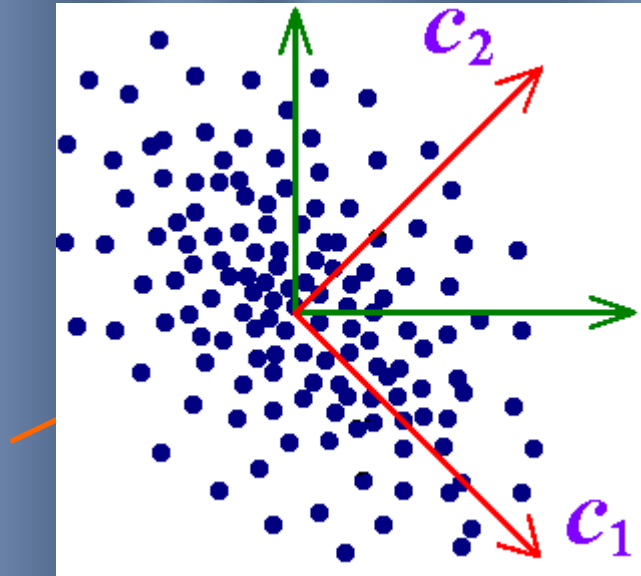


# Идея метода главных компонент



«Облако» точек, координатами которых служат значения  $m$  признаков для каждого из  $n$  объектов, может в действительности располагаться вблизи (содержаться в тонком слое вокруг) некоторой гиперплоскости размерности **намного меньше  $m$** .

# Определение главных компонент (principal components)



- Проекции объектов на первую главную компоненту  $c_1$  имеют **наибольшую выборочную дисперсию** среди дисперсий проекций на всевозможные направления в  $R^m$
- При  $j \geq 2$  можно показать, что  $c_j$  — направление с наибольшей выборочной дисперсией проекций объектов среди направлений, **ортогональных** векторам  $c_1, \dots, c_{j-1}$

# Приведение ковариационной матрицы главным осям

Обозначим через  $\hat{\Sigma}$  матрицу размерности  $m \times m$ , элементами которой являются попарные выборочные ковариации  $\hat{\sigma}_{ij}$  признаков  $X_1, \dots, X_m$ , т. е.  $\hat{\sigma}_{ij} = \mathbf{cov}(X_i, X_j)$ . Она называется *выборочной ковариационной матрицей*. Из линейной алгебры известно, что найдётся *ортогональная* матрица  $\mathbf{C}$  ( $\mathbf{C}^T \mathbf{C} = \mathbf{E}$ ), приводящая к  $\hat{\Sigma}$  главным осям:

$$\mathbf{C}^T \hat{\Sigma} \mathbf{C} = \mathbf{\Lambda}. \quad (1)$$

Здесь  $\mathbf{\Lambda}$  — *диагональная* матрица с неотрицательными числами  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m$  на главной диагонали. Эти числа являются выборочными дисперсиями проекций объектов на *главные оси*  $c_1, \dots, c_m$  (столбцы матрицы  $\mathbf{C}$ ). При этом выполняется равенство

$$\lambda_1 + \dots + \lambda_m = S^2(X_1) + \dots + S^2(X_m).$$

Представляет интерес *относительная доля разброса*

$$\gamma_k = (\lambda_1 + \dots + \lambda_k) / (\lambda_1 + \dots + \lambda_m),$$

приходящаяся на первые  $k$  главных осей (компонент).

$\lambda_1$	0	0	0
0	$\lambda_2$	0	0
0	0	...	0
0	0	0	$\lambda_m$

# Главные оси в матричной записи

Если признаки (столбцы матрицы данных)  $\mathbf{X}$  центрированы, то выборочная ковариационная матрица  $\hat{\Sigma}$  представляется в виде

$$\hat{\Sigma} = \frac{1}{n} \mathbf{X}^T \mathbf{X}. \quad (2)$$

Из формулы (1), приведённой на предыдущем слайде, используя свойства матричного умножения, получаем представление

$$\frac{1}{n} \mathbf{C}^T \mathbf{X}^T \mathbf{X} \mathbf{C} = \frac{1}{n} (\mathbf{X} \mathbf{C})^T \mathbf{X} \mathbf{C} = \mathbf{\Lambda}. \quad (3)$$

Матрица  $\tilde{\mathbf{X}} = \mathbf{X} \mathbf{C}$ , рассматриваемая по строкам, является записью координат объектов ( $m$ -мерных точек) в новом базисе, состоящем из главных компонент  $\mathbf{c}_1, \dots, \mathbf{c}_m$ . Столбец матрицы  $\tilde{\mathbf{X}}$  под номером  $j$  представляет собой  $n$ -мерный вектор, состоящий из длин проекций  $m$ -мерных точек на ось  $\mathbf{c}_j$ . Его можно интерпретировать как выборку значений нового признака  $\tilde{X}_j$ .

# Собственные значения и векторы

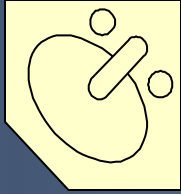
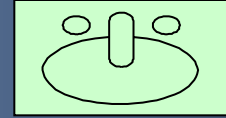
Сравнивая формулы (2) и (3), видим, что диагональная матрица  $\Lambda$  является выборочной ковариационной матрицей для выборок  $\tilde{X}_1, \dots, \tilde{X}_m$ . Поэтому векторы  $\tilde{X}_1, \dots, \tilde{X}_m$  ортогональны, и выборочные дисперсии  $S^2(\tilde{X}_j) = \lambda_j$ ,  $j = 1, \dots, m$ .

Умножая обе части формулы (1) слева на ортогональную матрицу  $C$  и учитывая, что  $CC^T = E$ , выводим формулу

$$\hat{\Sigma}C = C\Lambda.$$

Она означает, что столбец  $c_j$  матрицы  $C$  удовлетворяет условию  $\hat{\Sigma}c_j = \lambda_j c_j$ , т. е.  $c_j$  является собственным вектором ковариационной матрицы  $\hat{\Sigma}$  с собственным значением  $\lambda_j$ .

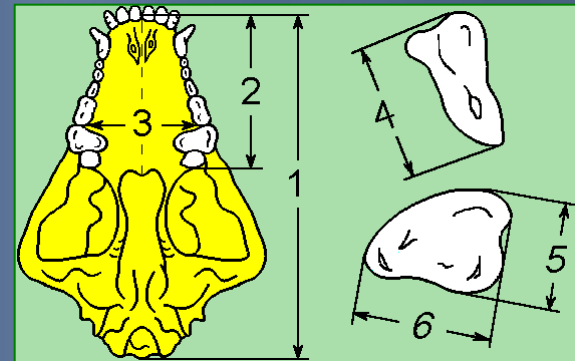
# Исследование черепов



## Уменьшение размерности пространства признаков

В файле **Dog\_Wolf.txt** содержатся (стандартизованные) данные о размерах челюстей и зубов собак и волков:

- 1 – длина черепа,
- 2 – длина верхней челюсти,
- 3 – ширина верхней челюсти,
- 4 – длина верхнего карнизора,
- 5 – длина первого верхнего моляра,
- 6 – ширина первого верхнего моляра.



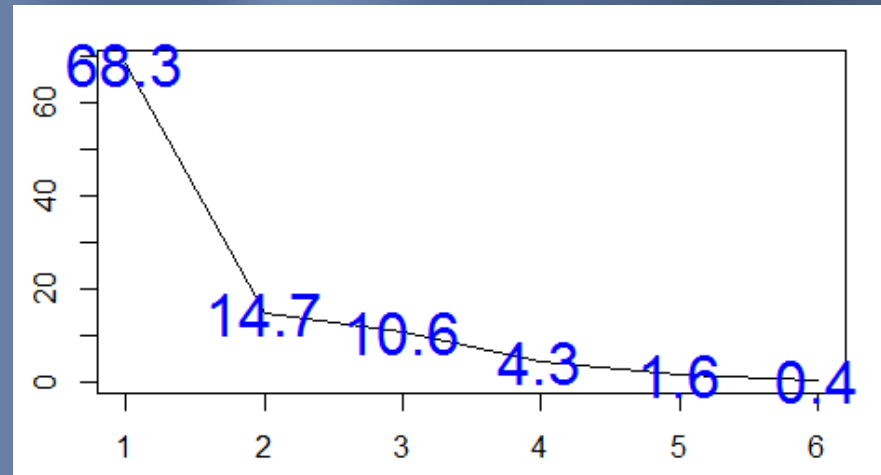
- 1) Импортируйте файл **Dog\_Wolf.txt** под именем **dw** (**Внимание:** чтобы правильно ввести заголовки столбцов при импортировании, выберите для опции **Heading** значение **Yes**)
- 2) Удалите последний столбец из таблиц, найдите с помощью функции **prcomp** корни из собственных значений  $\lambda_1 - \lambda_6$  и главные оси  $c_1 - c_6$ , запишите результат функции **prcomp** в **p**

# Вычисление главных компонент

С помощью команды `summary(p)` узнаем, что на первые 3 главные оси приходится

$$68,3\% + 14,7\% + 10,6\% \approx 93,7\%$$

общего рассеяния объектов.



Наглядную иллюстрацию уменьшения степени важности главных компонент даёт приведённая справа диаграмма «Осыпь». С её помощью можно визуально установить момент смены быстрого убывания функции на медленное и, тем самым, определить, сколько главных компонент отбросить, считая разброс по ним «шумом».

Собственные векторы определяются однозначно с точностью до одновременной смены знака у всех компонент вектора. Поэтому собственные векторы задают не направления, а прямые линии — новые координатные оси.

# Интерпретация главных компонент (новых признаков)

	1 X1	2 X2	3 X3	4 X4	5 X5	6 X6
1	-2,78	-2,83	2,46	-1,37	-2,29	-1,22
2	-1,86	-2,16	0,38	-0,41	-0,07	-0,13
3	-1,28	-1,30	-0,17	-1,22	-1,48	-0,37
4	-0,62	-0,83	0,05	-0,60	-0,74	-0,82
5	-1,61	-1,70	-1,92	-1,52	-1,63	-1,55
6	-1,50	-1,10	-1,59	-1,37	-1,18	-0,86
7	-0,07	0,17	-0,83	-0,14	-0,22	-0,00
8	-0,99	-0,63	-1,70	-1,45	-1,11	-1,02
9	0,26	0,50	-0,83	-0,21	0,00	-0,53
10	0,59	1,10	-1,15	0,06	0,67	0,08
11	-0,80	-0,63	-2,25	-0,68	-1,03	-1,34
12	0,26	0,37	-0,83	-0,52	-0,07	-0,33
13	0,55	0,70	-0,28	-0,48	0,00	-0,49
14	0,40	0,44	-0,06	-0,21	0,08	0,36
15	0,40	0,37	0,27	-0,56	-0,22	-0,17
16	0,00	0,24	-0,50	-0,10	-0,14	-0,17
17	0,84	1,04	0,60	0,56	-0,07	0,40

## Собственные векторы (направления главных осей)

Eigenvectors of correlation matrix (Dog_Wolf) Active variables only						
Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
X1	-0,430827	0,228448	-0,530015	-0,108190	0,046871	0,683653
X2	-0,430515	0,380380	-0,391634	-0,008844	-0,202594	-0,689542
X3	-0,228137	-0,888248	-0,374899	0,022127	-0,003609	-0,133852
X4	-0,439023	-0,065996	0,397174	0,525108	-0,581671	0,176284
X5	-0,460070	0,020866	0,273272	0,307189	0,781509	-0,090008
X6	-0,415443	-0,096679	0,439013	-0,785890	-0,087453	-0,007520

Чтобы найти координату проекции 6-мерной точки на **первую** главную ось, надо скалярно умножить соответствующую строку таблицы данных на вектор **Factor 1**, задающий направление этой оси.

## Интерпретация осей

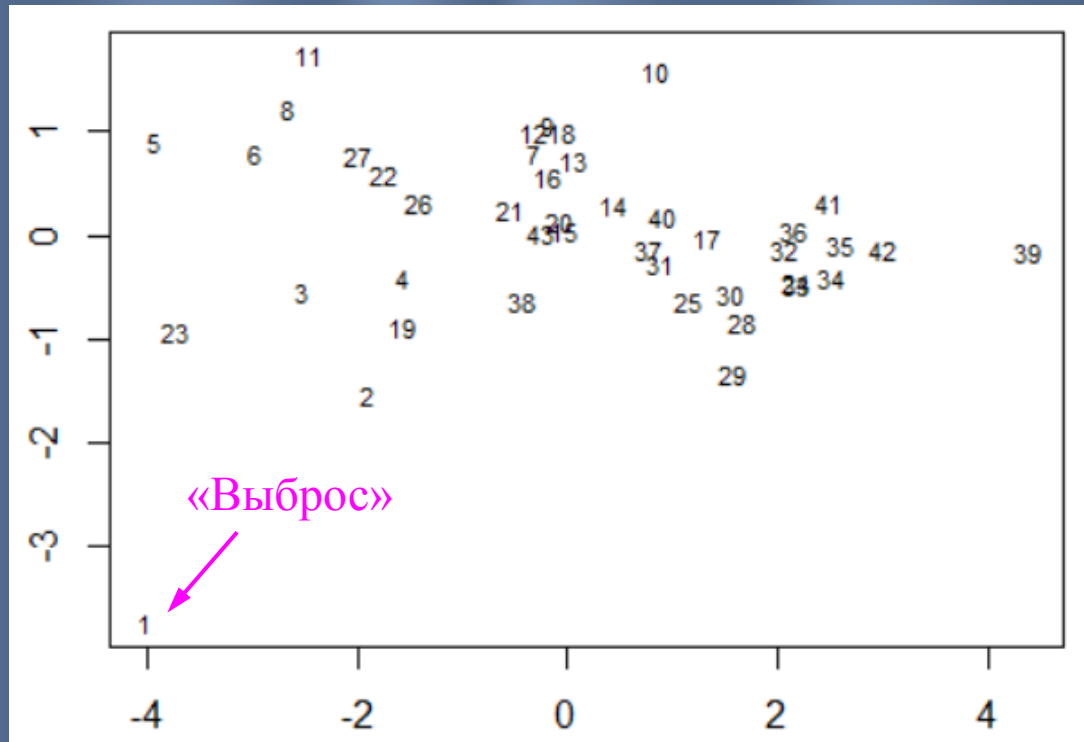
1-я главная ось — «общий размер»,  
2-я главная ось — «ширина верхней челюсти» (старый признак номер 3),  
3-я главная ось — «относительная (по сравнению с размерами черепа) величина зубов».



# Предназначение метода главных компонент

- **Уменьшение размерности** пространства признаков за счёт перехода к новым признакам — главным компонентам, получаемым путем агрегирования (взвешенного суммирования) старых признаков
- **Содержательная интерпретация** главных компонент, позволяющая осмыслить взаимосвязи признаков в новых терминах, глубже понять их суть
- **Визуализация данных** в результате их проецирования на пространство, порожденное первыми двумя (тремя) главными компонентами, при котором, как правило, сохраняются основные черты многомерной конфигурации: форма «облака», его однородность или же, напротив, выявляются «выбросы», небольшие или менее изолированные кластеры

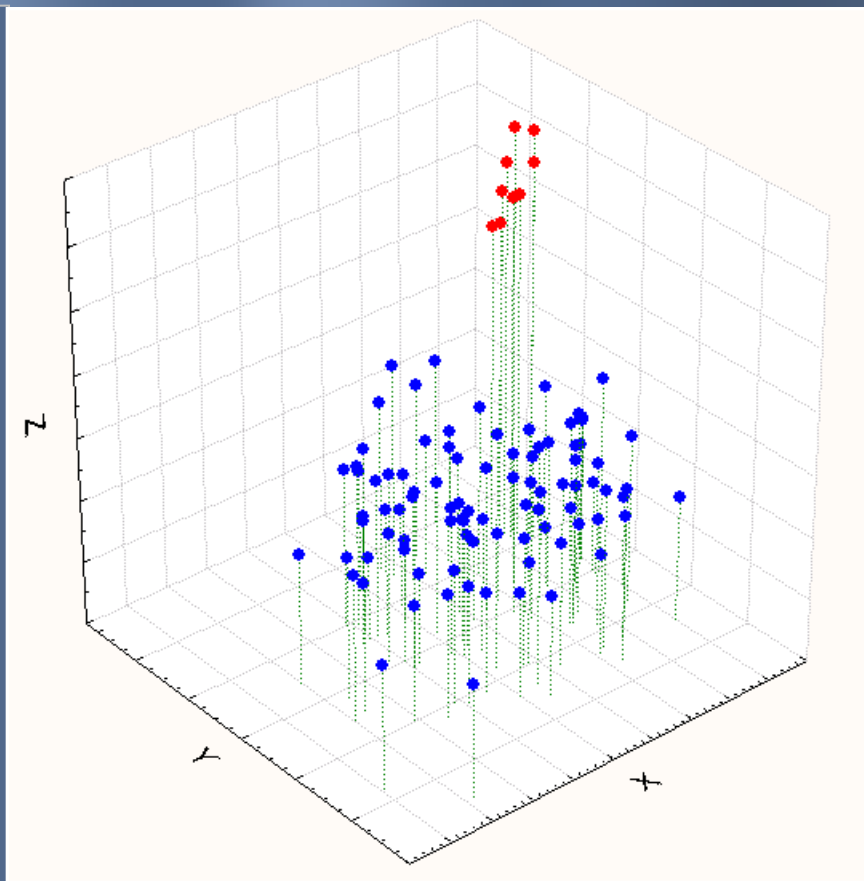
# Диаграмма рассеяния на плоскости двух первых главных компонент



Преобразуйте таблицу в матрицу: `m=as.matrix(dw[, -7])` и постройте диаграмму рассеяния объектов на плоскости двух первых главных компонент, используя матрицу из собственных векторов `prcomp(m)$rotation` и операцию матричного умножения `%*%`

# Опасности агрегирования признаков до кластеризации

- Возможная потеря некоторых кластеров в результате наложения их проекций в пространстве главных компонент
- Существенное изменение направлений самих главных осей под влиянием отдельных резко выделяющихся наблюдений или групп «выбросов»



# Нелинейные методы визуализации

Если “облако” точек не похоже на многомерный эллипсоид (его конфигурация существенно нелинейна), то точки, находящиеся на разных концах конфигурации, могут при ортогональном проецировании накладываться друг на друга.

Для анализа подобных сложных конфигураций применяются *нелинейные методы понижения размерности*. Они заключаются в минимизации некоторой функции  $F$ , выражающей суммарное расхождение между заданными различиями объектов  $\delta_{ij}$  и расстояниями  $d_{ij}$  между образами объектов в подпространстве небольшой размерности.

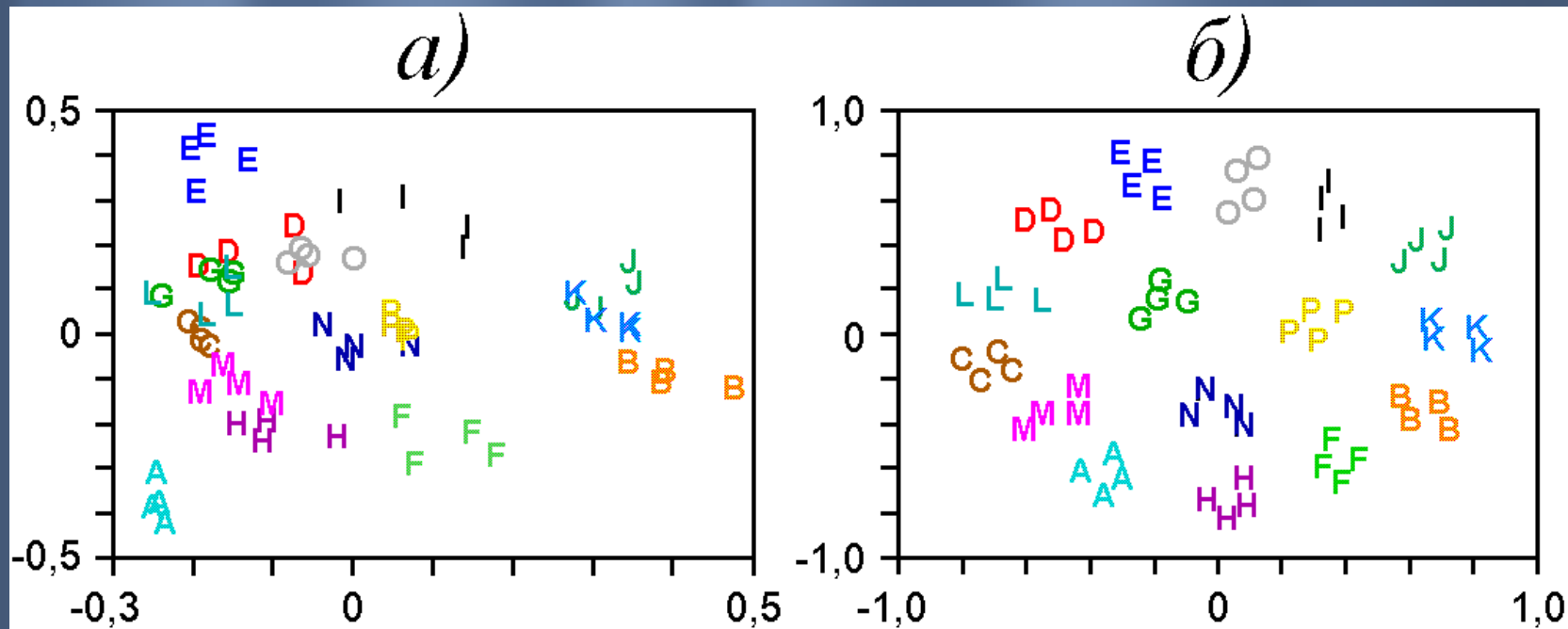
Широкое распространение получил метод **Сэммона**, у которого:

$$F = \frac{1}{C} \sum_{i < j} (d_{ij} - \delta_{ij})^2 / \delta_{ij}, \quad \text{где } C = \sum_{i < j} \delta_{ij}.$$

Предложен  
Дж. Сэммоном  
в 1969 г.

Этот метод обладает свойством более точно передавать небольшие различия и менее точно — большие, так как при отображении больших расстояний допустимы бóльшие ошибки. (Деление на константу  $C$  нужно для инвариантности функции  $F$  к изменению масштаба различий  $\delta_{ij}$ .)

# Проецирование симплекса



- На рисунке *а)* изображены проекции на плоскость двух первых главных компонент 16 групп точек по 4 точки в группе. Каждая группа генерировалась с помощью датчика случайных чисел вблизи одной из 16 вершин *правильного симплекса*. Видим, что проецирование **не позволяет** сохранить разделение на группы.
- На рисунке *б)* изображен результат применения *метода Сэммона* к этим же данным. Значение *F* удалось уменьшить с 0,44 до 0,13. Видим, что разделение точек на группы сохраняется при изображении многомерной конфигурации на плоскости.

# Применение метода Сэммона

- 
- Импортируйте файл `Simpl.txt` в RStudio
  - Спроецируйте данные из `Simpl` на плоскость двух первых главных компонент:

```
p=prcomp(Simpl); biplot(p)
```

- Нажмите кнопку `Zoom` и максимизируйте окно диаграммы
- Постройте матричную диаграмму рассеяния объектов для первых трёх главных осей:

```
pairs(p$x[,1:3], gap=0)
```

- Подключите пакет `MASS`, поставив «галку» на вкладке `Packages`
- Подготовьте нижнетреугольную матрицу расстояний между объектами:

```
d=dist(Simpl)
```

- Примените метод Сэммона: `s=sammon(d)`
- Постройте диаграмму рассеяния образов объектов на плоскости:

```
plot(s$points, type="n"); text (s$points, labels=1:nrow(Simpl))
```

# Связь суффиксов с литературными стилями

Words such as *goodness* and *sharpness* can be analyzed as consisting of a stem, *good*, *sharp*, and an affix, the suffix *-ness*. Some affixes are used in many words, *-ness* is an example. Other affixes occur only in a limited number of words, for instance, the *-th* in *warmth* and *strength*. The extent to which affixes are used and available for the creation of new words is referred to as the productivity of the affix. Baayen (1994) addressed the question of the extent to which the productivity of an affix is codetermined by stylistic factors. Do different kinds of texts favor the use of different kinds of affixes?

The data set `affixProductivity` lists, for 44 texts with varying authors and genres, a productivity index for 27 derivational affixes. The 44 texts represent four different text types: religious texts (e.g. the *Book of Mormon*, coded B), books written for children (e.g. *Alice's Adventures in Wonderland*, coded C), literary texts (e.g. novels by Austen, Conrad, James, coded L), and other texts (including officialse from the US government accounting office), coded O. The classification codes are given in the column labeled `Registers`:

	ian	ful	y	ness	able	ly	Registers
Mormon	0	0.1887	0.5660	2.0755	0.0000	2.2642	B
Austen	0	1.2891	1.5654	1.6575	1.0129	6.2615	L
Carroll	0	0.2717	1.0870	0.2717	0.4076	6.3859	C
Gao	0	0.3306	1.9835	0.8264	0.8264	4.4628	O



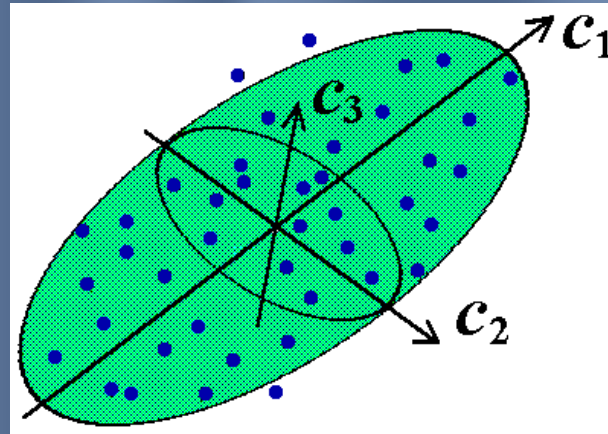
# Практическое задание 2

- 1) Импортируйте файл `AffixProd.txt` в Rstudio под именем `a`. Запишите таблицу `a` без первого и трёх последних столбцов в таблицу `d`
- 2) Примените к таблице `d` метод главных компонент и выясните насколько можно понизить размерность
- 3) Найдите наибольшие по абсолютной величине компоненты трёх первых собственных векторов и выясните, каким именно суффиксам соответствуют первые 3 главные оси (при выводе векторов на экран округлите числа до 3-х знаков после запятой командой `round`)
- 4) Запишите 4 буквы из `AffixProd$Registers` в вектор `s` и выясните, для каких стилей найденные 3 суффикса имеют наибольший средний индекс продуктивности (используйте цикл, команду `subset`, функции `mean` и `arg.max`)



# Эллипсоид рассеяния

Он позволяет выделить «ядро» облака точек: обобщает понятие межквартильного диапазона



Также с его помощью можно выявлять многомерные «выбросы»

- *Эллипсоидом рассеяния* распределения случайного вектора  $\xi$  называется  $m$ -мерный эллипсоид вида

$$x^T \Sigma_{\xi}^{-1} x \leq C.$$

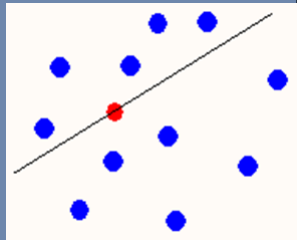
- Осями эллипсоида рассеяния служат главные компоненты матрицы ковариаций  $\Sigma_{\xi}$ . Длины его полуосей пропорциональны квадратным корням из собственных значений  $\lambda_j$  матрицы  $\Sigma_{\xi}$ .

# Диаграмма Bag Plot

В двумерном случае альтернативой эллипсоида рассеяния, содержащего 50% наблюдений, служит устойчивая к «выбросам» диаграмма «Bag Plot». Её построение базируется на понятии «глубины» точки относительно двумерного «облака», которое обобщает ранг наблюдения в одномерной выборке.

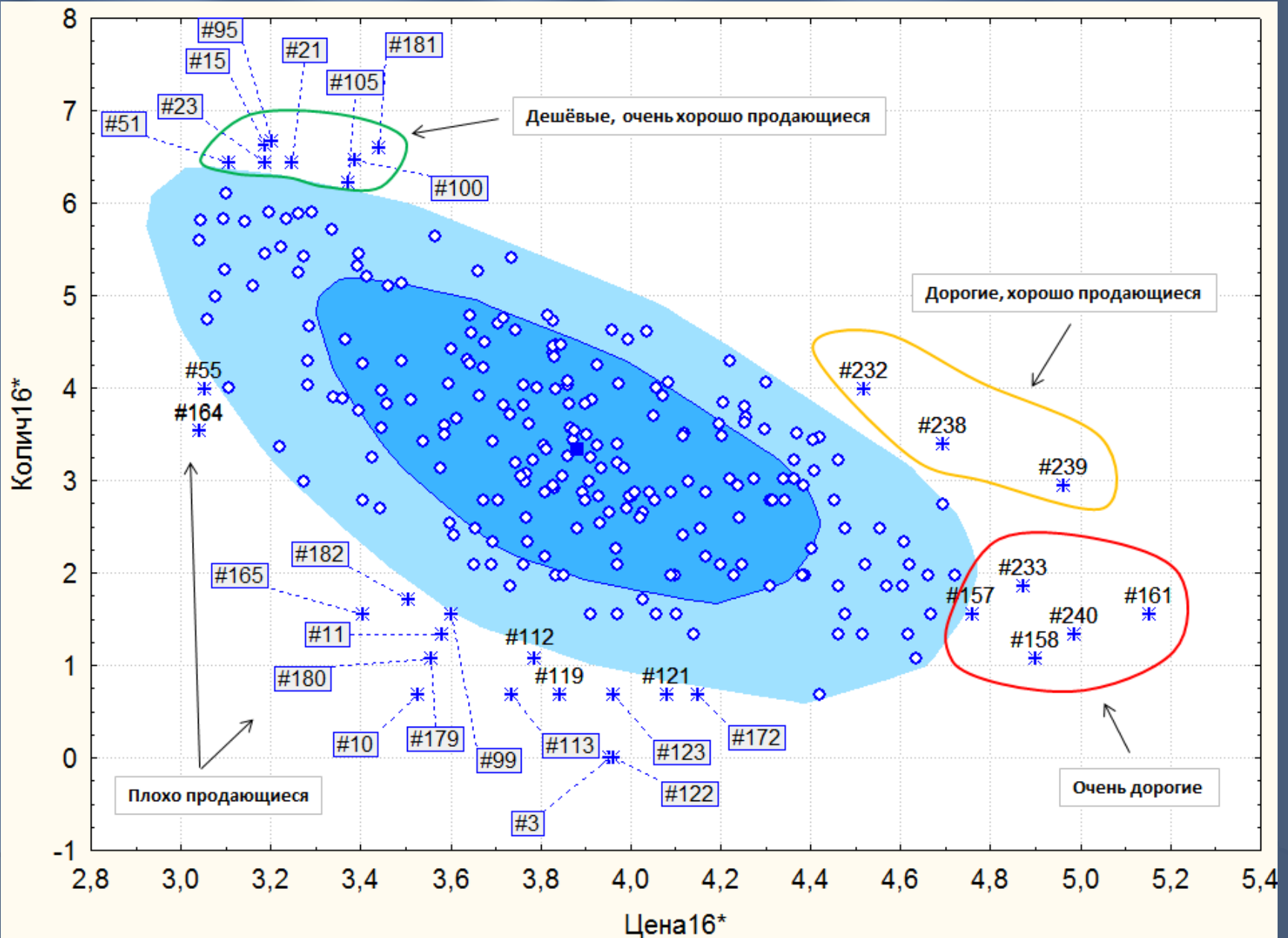
Сначала рассмотрим в одномерном случае выборку  $\{1, 2, 3, 4, 5\}$ . «Глубиной» отдельного наблюдения назовём минимум из двух чисел: количества элементов выборки, оказавшихся слева от наблюдения, количества элементов выборки, оказавшихся справа от наблюдения. Например, наблюдение 2 имеет «глубину»  $\min\{2, 4\} = 2$ . Тогда выборочную медиану *MED* можно альтернативным образом определить как наблюдение, имеющее наибольшую «глубину».

В двумерном случае «глубиной» точки называется наименьшее число наблюдений, находящееся в полуплоскости, ограниченной произвольной прямой линией, проходящей через точку (см. рисунок). Двумерным обобщением *MED* служит **медиана Тьюки** (*Tukey's median*) — точка с наибольшей «глубиной».



«Bag» — выпуклый многоугольник, содержащий 50% точек, имеющих наибольшую «глубину». «Fence» — это «Bag», пропорционально раздутый относительно медианы Тьюки в 1,5 раза (по умолчанию).

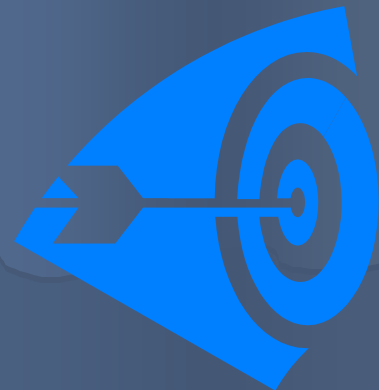
# Сегментация 245 товаров с помощью Bag Plot



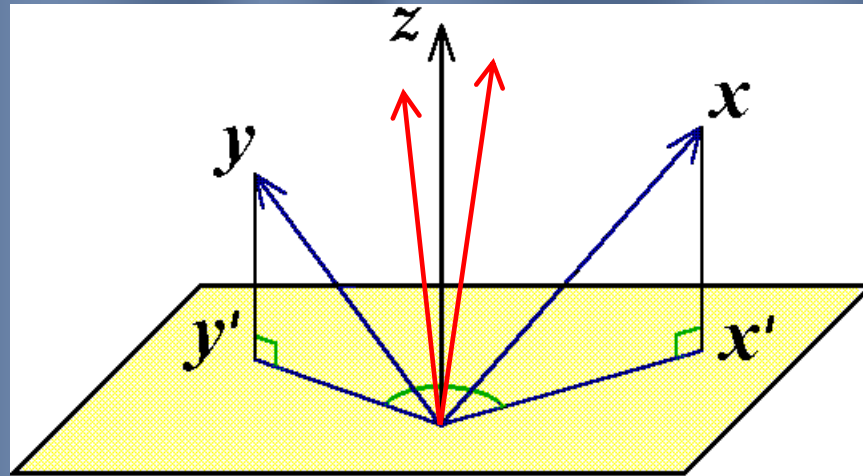
# Многомерные выбросы

**Эксперимент  
можно считать удавшимся,  
если нужно отбросить  
не более 50% сделанных измерений,  
чтобы достичь соответствия  
с теорией.**

*Следствие  
из законов Мейерса*



# Частная корреляция



*Частным коэффициентом корреляции* между случайными величинами  $X$  и  $Y$  при исключении влияния случайной величины  $Z$  называется

$$\rho_{XY|Z} = \frac{\rho(X,Y) - \rho(X,Z)\rho(Y,Z)}{\sqrt{(1 - \rho^2(X,Z))(1 - \rho^2(Y,Z))}}.$$

# Ложная корреляция



Пусть  $Y$  — *ущерб от пожара*,  $X$  — *количество пожарных*, принимавших участие в тушении пожара. Если попытаться выявить корреляционную зависимость  $Y$  от  $X$ , то она будет положительной и значимой. Однако понятно, что на самом деле причиной больших значений обоих признаков служит  $Z$  — некоторая мера сложности пожара (скажем, *площадь, охваченная пожаром*).

*Ложной корреляцией* называют ошибочное понимание корреляционной связи как *причинно-следственной*. В результате у исследователя возникает представление, что, оказывая влияние на один из признаков, можно изменить другой. Однако на самом деле это не всегда так.



Correlation does not imply causation

# Дети и аисты



Если подсчитать корреляцию между  $Y$  — *количеством детей*, ежегодно рождавшихся в 19 веке в Голландии и  $X$  — *количеством прилетавших аистов*, то она окажется значимой. Можно ли на основе этого статистического результата заключить, что детей приносят аисты?

Рассмотрим проблему на содержательном уровне. Аисты появляются там, где им удобно вить гнезда; излюбленным же местом их гнездовья являются высокие дымовые трубы, какие строили в голландских сельских домах. По традиции новая семья строила себе новый дом — появлялись новые трубы и, естественно, рождались дети.

Таким образом, и увеличение числа гнезд аистов, и увеличение числа детей являются следствиями одной причины  $Z$  — *образования новых семей*.



# СВЯЗЬ НОМИНАЛЬНЫХ ПРИЗНАКОВ



Рассмотрим задачу выявления статистической связи для сгруппированных (разбитых на категории) данных. Входной информацией служит *таблица сопряженности* двух признаков  $\|\nu_{ij}\|_{n \times m}$ . Числа  $\nu_{ij}$  представляют собой количества объектов, имеющих по первому признаку категорию  $i$ , а по второму признаку категорию  $j$ ,  $N = \sum \nu_{ij}$  — общее число наблюдений.

Проверяется гипотеза независимости признаков  $H$ , заключающаяся в том, что совместное распределение признаков есть произведение их одномерных (*маргинальных*) распределений:

$$H: p_{ij} = r_i s_j, \quad \text{где } r_i = \sum_{j=1}^m p_{ij}, \quad s_j = \sum_{i=1}^n p_{ij}.$$

Для проверки гипотезы  $H$  обычно применяется *критерий хи-квадрат*, статистика которого имеет вид

$$X^2 = N \sum_{i=1}^n \sum_{j=1}^m \frac{(\nu_{ij} - n_i m_j / N)^2}{n_i m_j}, \quad \text{где } n_i = \sum_{j=1}^m \nu_{ij}, \quad m_j = \sum_{i=1}^n \nu_{ij}.$$

При справедливости  $H$  статистика  $X^2$  приближенно распределена по закону хи-квадрат с  $(n-1)(m-1)$  степенями свободы.



# Критерий отношения правдоподобий

Частоты  $\hat{r}_i = n_i/N$  и  $\hat{s}_j = m_j/N$  служат оценками для неизвестных вероятностей  $r_i$  и  $s_j$ . Обозначим через  $\hat{\mu}_{ij} = N\hat{r}_i\hat{s}_j = n_im_j/N$  — *ожидаемое количество наблюдений* в  $(i, j)$ -й ячейке таблицы сопряжённости. Тогда статистику критерия хи-квадрат можно записать в виде

$$X^2 = \sum_{i,j} \frac{(v_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

Наряду с критерием хи-квадрат используется также критерий *отношения правдоподобий*, статистикой которого служит величина

$$G^2 = 2 \sum_{i,j} v_{ij} \ln \frac{v_{ij}}{\hat{\mu}_{ij}}.$$

Статистика  $G^2$  пропорциональна *взаимной информации*  $I$  между распределением  $\hat{p}_{ij} = v_{ij}/N$  и распределением  $\hat{q}_{ij} = \hat{r}_i\hat{s}_j$ . Она, так же, как и статистика  $X^2$ , имеет в качестве предельного распределения  $\chi^2_{(n-1)(m-1)}$ .

Однако у статистики  $G^2$  сходимость более медленная. Кроме того, в таблице сопряжённости недопустимы нулевые  $v_{ij}$  и рекомендуется иметь  $\hat{\mu}_{ij} \geq 5$ .

# Синтаксис дательного падежа

В файле [Dative.txt](#) содержатся признаки, относящиеся к 903 английским глаголам из примера, приведённого в книге Р. Байен «Анализ лингвистических данных» на с. 4.

Bresnan *et al.* (2007) studied the dative alternation in English in the three-million-word Switchboard collection of recorded telephone conversations and in the Treebank *Wall Street Journal* collection of news and financial reportage. In English, the recipient can be realized either as an NP (*Mary gave John the book*) or as a PP (*Mary gave the book to John*). Bresnan and colleagues were interested in predicting the realization of the recipient (as NP or PP) from a wide range of potential explanatory variables, such as the animacy, the length in words, and the pronominality of the theme and the recipient.

RealizationOfRec	Verb	AnimacyOfRec	AnimacyOfTheme	LogLenOfTheme
NP	feed	animate	inanimate	2.6390573
NP	give	animate	inanimate	1.0986123
NP	give	animate	inanimate	2.5649494
NP	give	animate	inanimate	1.6094379
NP	offer	animate	inanimate	1.0986123

# Практическое задание 3

- 1) Импортируйте файл `Dative.txt` в программу RStudio
- 2) Постройте перекрёстную таблицу `t` для признаков `RealizationOfRec` и `AnimacyOfRec` с помощью команды `table` и выведите таблицу `t` на экран
- 3) Вычислите по таблице `t` частоты использования конструкции NP для одушевлённых и для неодушевлённых получателей
- 4) Проверьте критерием хи-квадрат (`chisq.test`) независимость синтаксической формы дательного падежа от одушевлённости получателя

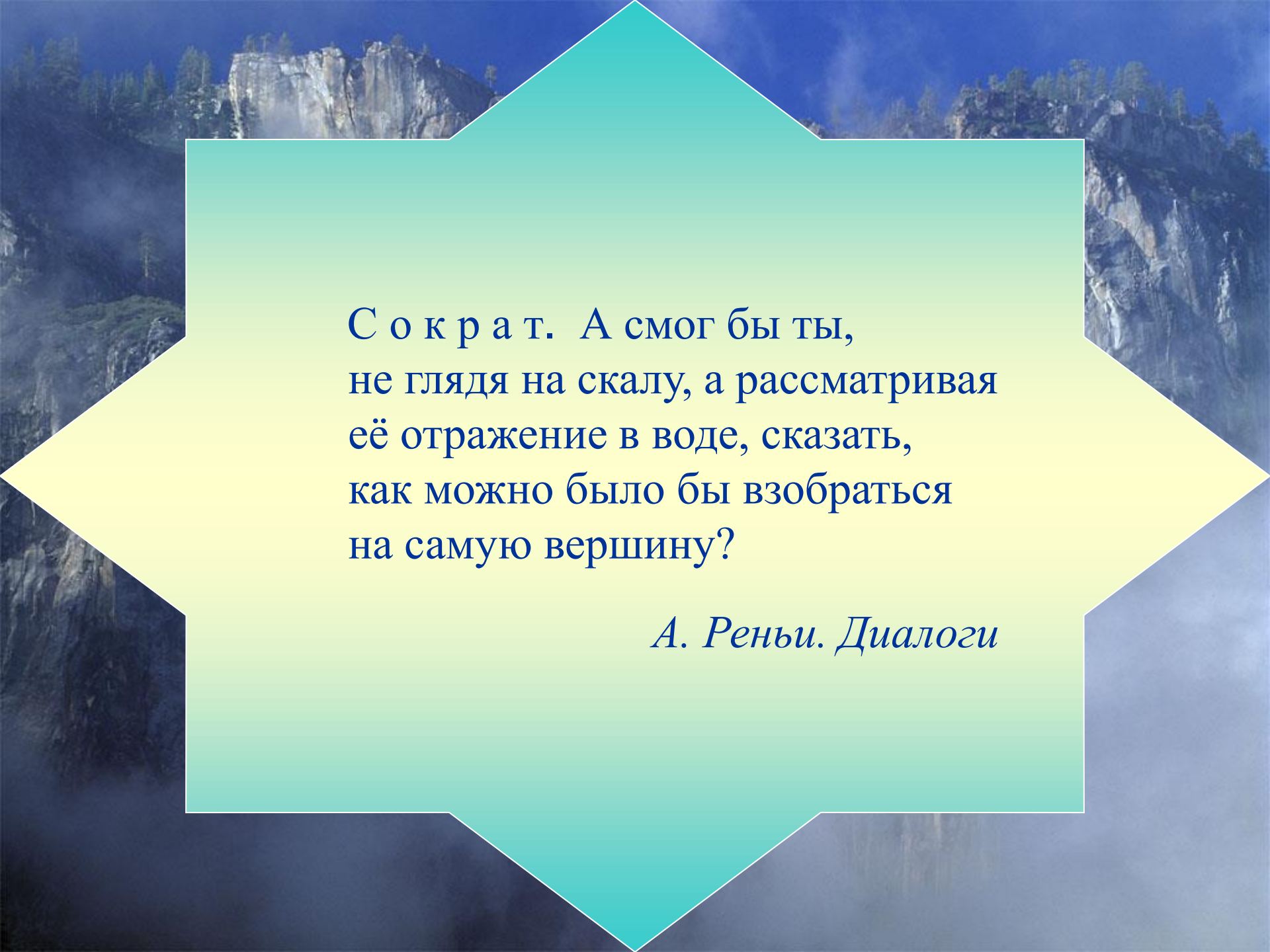
---

Рассмотрим встроенный в пакет `MASS` пример — таблицу данных `caith` о соответствии цвета глаз и цвета волос жителей шотландского графства `Caithness`. Эта местность особенно интересна тем, что её жители являются потомками скандинавов, кельтов и англо-саксонцев.

- 5) Активируйте пакет `MASS` и выведите на экран таблицу `caith`
- 6) Проверьте гипотезу независимости цвета глаз и цвета волос

# Главное в теме

- Когда в таблице имеется много зависимых признаков, то можно попытаться их агрегировать, т. е. перейти от них к небольшому числу новых признаков по возможности без большой потери информации. **Основная цель** — уменьшение размерности пространства признаков [это важно, скажем, для последующего регрессионного анализа — построения точной и надёжной модели прогнозирования некоторого признака на основе других]
- **Главные компоненты** задают направления наибольшей вытянутости многомерного «облака» координат объектов (строк таблицы данных). Проекции объектов на эти направления можно считать новыми признаками
- При проецировании на плоскость, порождённую двумя первыми главными компонентами, как правило, сохраняются основные черты многомерной конфигурации: форма «облака», его однородность или же, напротив, выявляются «выбросы» или небольшие, менее изолированные кластеры
- **Коэффициент частной корреляции** используется для проверки того, что корреляционная связь между двумя признаками вызвана их сильной зависимостью от третьего признака



С о к р а т. А смог бы ты,  
не глядя на скалу, а рассматривая  
её отражение в воде, сказать,  
как можно было бы взобраться  
на самую вершину?

*А. Реньи. Диалоги*

# Домашнее задание

В файле Children.txt содержится таблица, к которой представлены результаты социологического обследования о связи между доходом семей и количеством детей в них. Признак А означает количество детей и принимает значения

«0», «1», «2», «3», «4 и более».

Признак В указывает, какому из диапазонов

«0 – 1000», «1000 – 2000», «2000 – 3000», «больше 3000»

принадлежит доход семьи (в шведских кронах).

- 1) Импортируйте файл Children.txt в Rstudio
- 2) Проверьте гипотезу независимости признаков А и В
- 3) Постройте столбиковую диаграмму для всей таблицы. Визуально сравните распределения количества детей в семьях с разным доходом
- 4) Сформулируйте вывод о форме зависимости