Анализ данных с помощью статистических методов

Данные есть некоторый факт;
то, на чём основан вывод или
любая интеллектуальная система.

Websteres New World Dictionary

Основная задача анализа данных

Исследователя-прикладника часто интересует вопрос, имеется ли зависимость между измеряемыми характеристиками (признаками). Например, лингвист может захотеть выяснить, есть ли связь между длиной слова и частотой его встречаемости или зависит ли количество синонимов слова от его древности.

В отличие от зависимостей, описываемых с помощью функций, математическая статистика изучает зависимости между случайными величинами. Наблюдения рассматриваются как значения некоторого признака у объектов, случайно выбранных из «необъятной» генеральной совокупности. Насколько большими могут быть отклонения выборочных характеристик от соответствующих характеристик генеральной совокупности определяется законами случайности (в частности, центральной предельной теоремой). Слишком большое отклонение свидетельствуют о неадекватности используемой модели, о наличии некоторого систематического эффекта («сигнала»). Основная задача — выделить «сигнал» из случайного «шума».

Таблица «Объекты – признаки»

Как правило, имеющиеся у исследователя данные можно представить в виде таблицы вида «Объекты — признаки»: для каждого объекта приведены значения ряда его признаков.

	Признак 1	Признак 2	Признак 3	Признак 4
Объект 1	7	1,35	глагол	235
Объект 2	15	2,44	существительное	
Объект 3	9	0,67	наречие	554

Например, объектами могут быть слова некоторого языка, а признаками — некоторые лингвистические характеристики: длина слова, часть речи, количество синонимов и т. п.

В некоторых ячейках таблицы могут отсутствовать значения признака, т. е. таблицы могут содержать пропуски.

Типы шкал

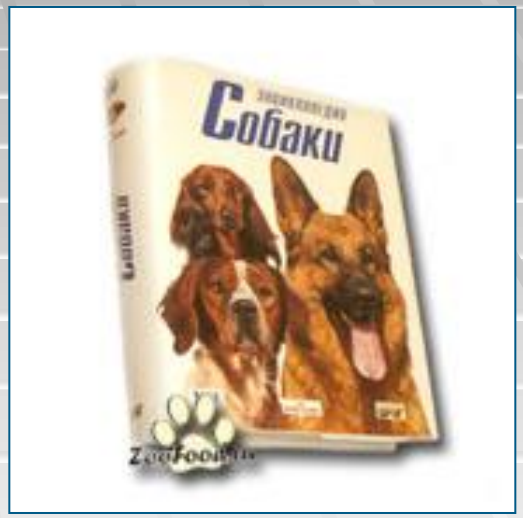
Количественная (интервальная)

В количественной шкале представляются действительные числа (со знаками после запятой).

Порядковая

В порядковой шкале числа используются для установления порядка между объектами.

Номинальная (шкала наименований)



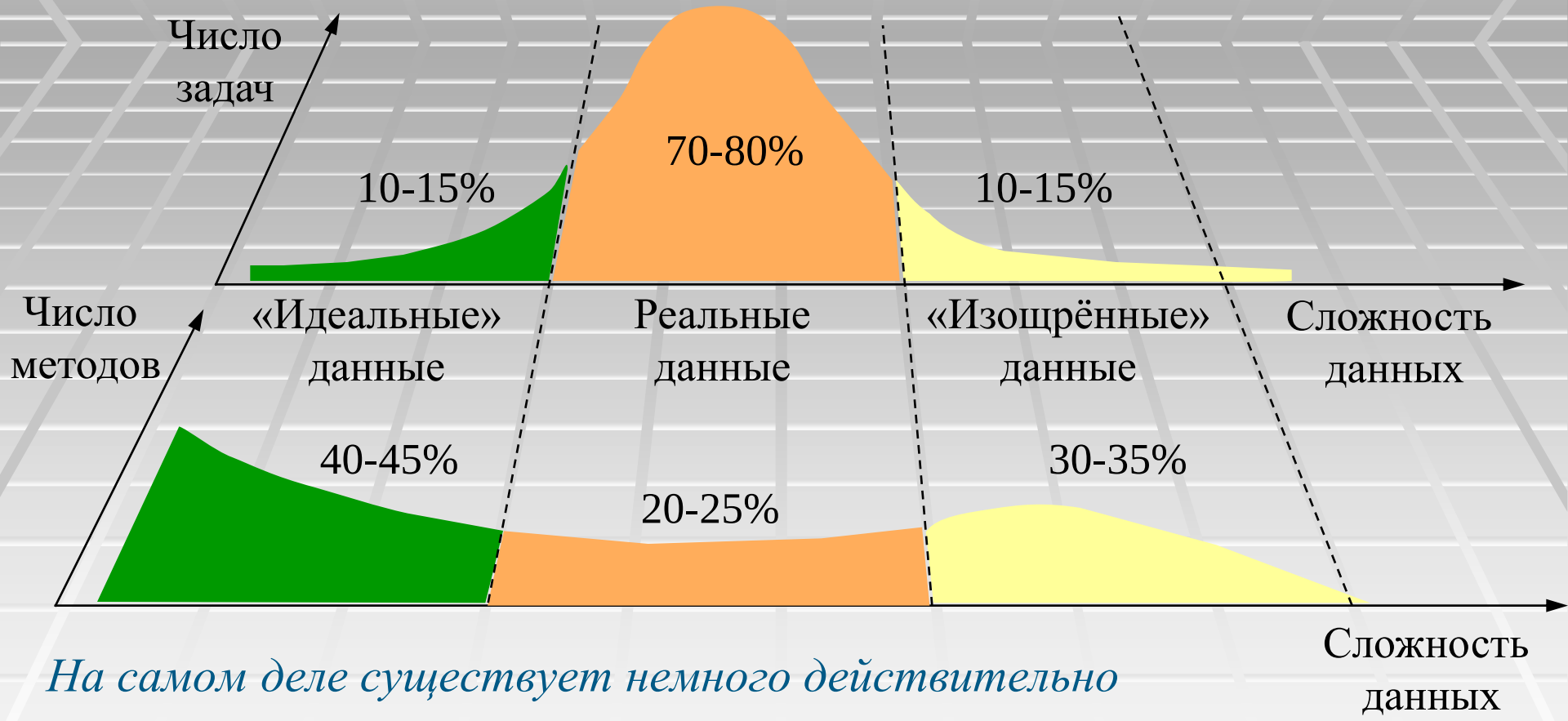
В номинальной шкале числа используются лишь как метки (номера категорий).

Последовательность действий

Наука не начинается с фактов, она начинается с выявления проблемы и веры в возможность её решения.

М. и И. Голдстейн «Как мы познаём»

Несоответствие задач и методов



На самом деле существует немного действительно отличающихся друг от друга и устойчиво работающих на реальных данных математических методов.

В. В. Александров, А. И. Алексеев, Н. Д. Горский

Типичные задачи, решаемые с помощью статистических методов

1) Агрегирование, описание и представление данных

2) Выявление различия между выборками (группами)

3) Анализ повторных наблюдений

4) Классификация (сегментация) многомерных данных

5) Обнаружение зависимостей между признаками

6) Предсказание некоторого показателя на основе других

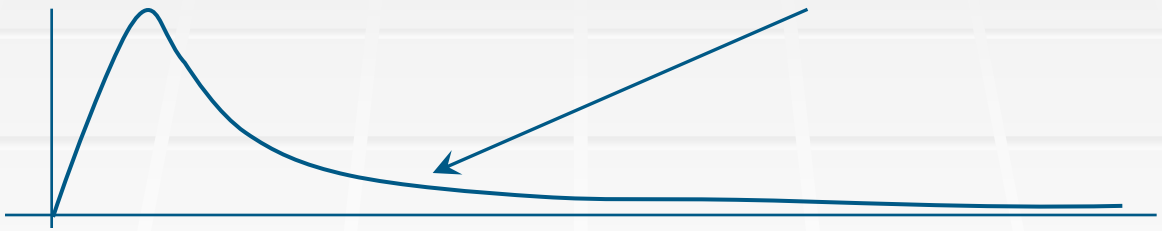
Базовые методы описательной статистики

Вычисление средних значений и стандартных отклонений

Построение диаграмм размахов и
выявление нетипичных наблюдений («выбросов»)

Построение гистограмм для выяснения формы распределения признаков

Преобразования признаков (как правило, применяется логарифмирование признаков, у которых распределения имеют тяжёлые правые «хвосты»)



Оценки положения и масштаба

Имеющиеся наблюдения $X_1, X_2, \ldots, X_n$ рассматриваются как выборка из генеральной совокупности с неизвестной функцией распределения $F(x)$. Под «центром» области типичных значений наблюдений обычно понимают математическое ожидания распределения, естественной оценкой которого служит *выборочное среднее*

$$\overline{X} = (X_1 + \ldots + X_n)/n.$$

Степень рассеяния наблюдений на практике, как правило, измеряют с помощью дисперсии распределения σ^2. Оценкой для параметра σ служит *выборочное стандартное отклонение*

Почему $n-1$, а не n?

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

S^2 — несмещённая оценка для σ^2: $MS^2 = \sigma^2$

Практическое задание 1

1) Моделируйте выборку z размера $n = 100$ стандартно нормально распределённых случайных чисел с помощью функции rnorm (название функции — сокращение от англ. слов random и normal)

2) Для выборки z вычислите выборочное среднее $\overline{X}$ и выборочное стандартное отклонение S с помощью функций mean и sd (standard deviation). Насколько сильно отклонились оценки $\overline{X}$ и S от математического ожидания и дисперсии соответственно?

---

3) Для распределения с плотностью $f_X(x) = 1 - |x|$ на отрезке $[-1, 1]$ найдите функцию распределения $F_X(x)$.
(Имейте в виду, что функция распределения задается разными формулами на разных промежутках числовой оси.)

---

4) Найдите обратную функцию к функции распределения $F_X(x)$.
(Имейте в виду, что обратная функция задается разными формулами на отрезках [0, 1/2] и [1/2, 1].)

Практическое задание 2

1) Моделируйте с помощью метода обратной функции выборку x размера $n = 10\,000$ из «треугольного» распределения с плотностью $f_X(x) = 1 - |x|$ на отрезке [–1, 1], встретившегося в практическом задании 1 (используйте функцию ifelse и функцию квадратного корня sqrt — сокращение для square root)

2) Вычислите для выборки x выборочное среднее $\bar{X}$ и выборочное стандартное отклонение S.

3) Найдите теоретически, к какому действительному числу будет приближаться значение sd(x) при увеличении размера выборки n, и вычислите его с точностью 7 знаков после запятой.

Вариационный ряд, медиана, квартили

Вариационный ряд образуют наблюдения $X_1, X_2, \ldots, X_n$, расположенные в порядке возрастания: $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$.

В частности, $X_{(1)} = \min\{X_1, \ldots, X_n\}$, $X_{(n)} = \max\{X_1, \ldots, X_n\}$.

Величина $R = X_{(n)} - X_{(1)}$ называется *размахом выборки.*

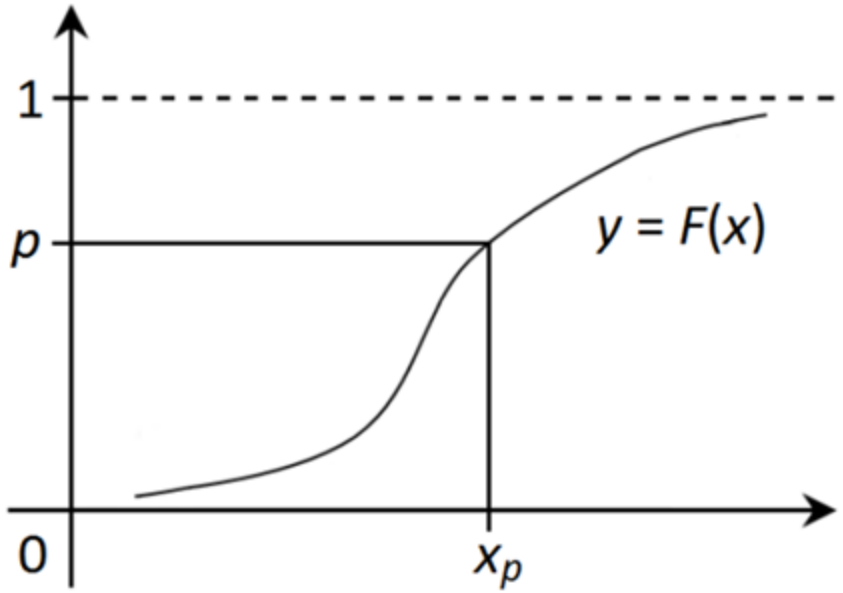
$X_{(2)} = ?$
$(n = 3)$

Выборочная медиана MED определяется как значение, стоящее в середине вариационного ряда (в случае, если n — чётное число, MED — полусумма двух средних членов ряда).

Она, наряду с выборочным средним $\overline{X}$, служит оценкой «центра» распределения, но намного более устойчива к выделяющимся наблюдениям (*«выбросам»*).

Теоретические квантили и квартили

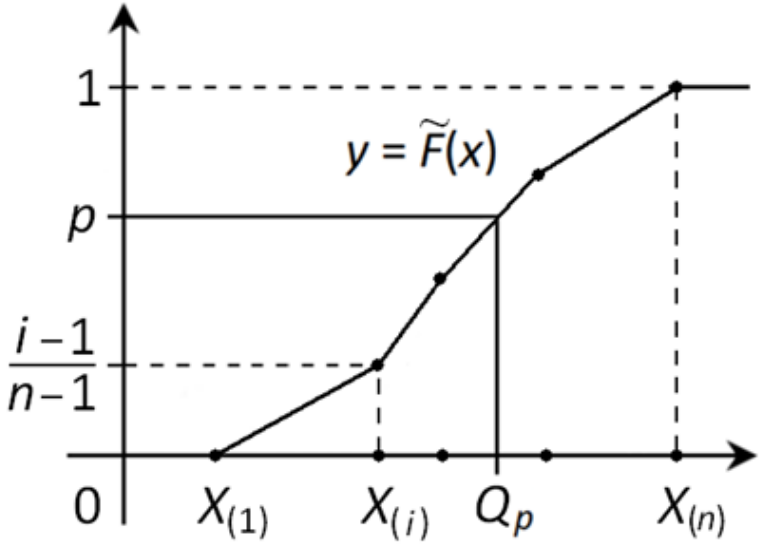
Определение. Пусть $0 < p < 1$. Значение аргумента функции распределения $F(x)$, при котором она достигает уровня p, называется *p-квантилью* и обозначается через x_p.



Иначе говоря, $x_p = F^{-1}(p)$. Частные случаи p-квантилей при $p = 1/4, 1/2, 3/4$ называются, соответственно, *нижней квартилью*, *медианой* и *верхней квартилью* распределения.

Выборочные квантили и квартили

Каким образом можно по выборке оценить p-квантиль x_p? Определим для неё оценку Q_p. Отметим на плоскости точки с координатами $(X_{(i)}, (i-1)/(n-1))$, где $i = 1, 2, \ldots, n$. Последовательно соединив их отрезками, получим ломаную, задающую график функции $\widetilde{F}(x)$.



Для произвольного числа p из интервала (0, 1) положим $Q_p = \widetilde{F}^{-1}(p)$. Оценка $Q_{1/2}$ совпадает с MED. Оценки $Q_{1/4}$ и $Q_{3/4}$ называются *нижней* и *верхней выборочными квартилями*.

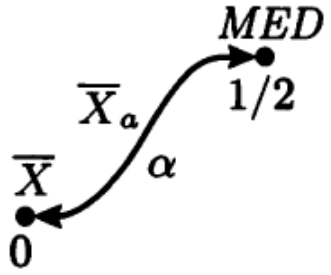
Урезанное (усечённое) среднее

Наряду с выборочным среднем и выборочной медианой для оценки «центра» области типичных значений наблюдений используется *урезанное (усечённое) среднее.*

Определение. Пусть $0 < \alpha < 1/2$, $k = [\alpha n]$, где $[\cdot]$ — *целая часть* числа, а n — объем выборки. *Усеченным средним* порядка α называется

$$\overline{X}_\alpha = \frac{1}{n-2k}\left(X_{(k+1)} + \ldots + X_{(n-k)}\right),$$

где $X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}$ — вариационный ряд.



На практике усечённое среднее применяется, например, для уменьшения необъективности судейства в фигурном катании, прыжках в воду и командной гимнастике.

Практическое задание 3

1) Для ранее моделированной выборки x вычислите минимум, максимум, медиану с помощью функций min, max, median

2) Вычислите вектор v, содержащий следующие 5 характеристик: минимум, нижнюю квартиль, медиану, верхнюю квартиль, максимум с помощью функции quantile без аргументов

3) Вычислите *межквартильный размах* — разность верхней и нижней квартилей и убедитесь, что вычисленное значение совпадает с результатом вызова функции IQR (interquartile range)

4) Подсчитайте усечённое среднее (trimmed mean) порядка $\alpha = 0{,}1$ с помощью функции mean, задав нужное значение параметра trim

---

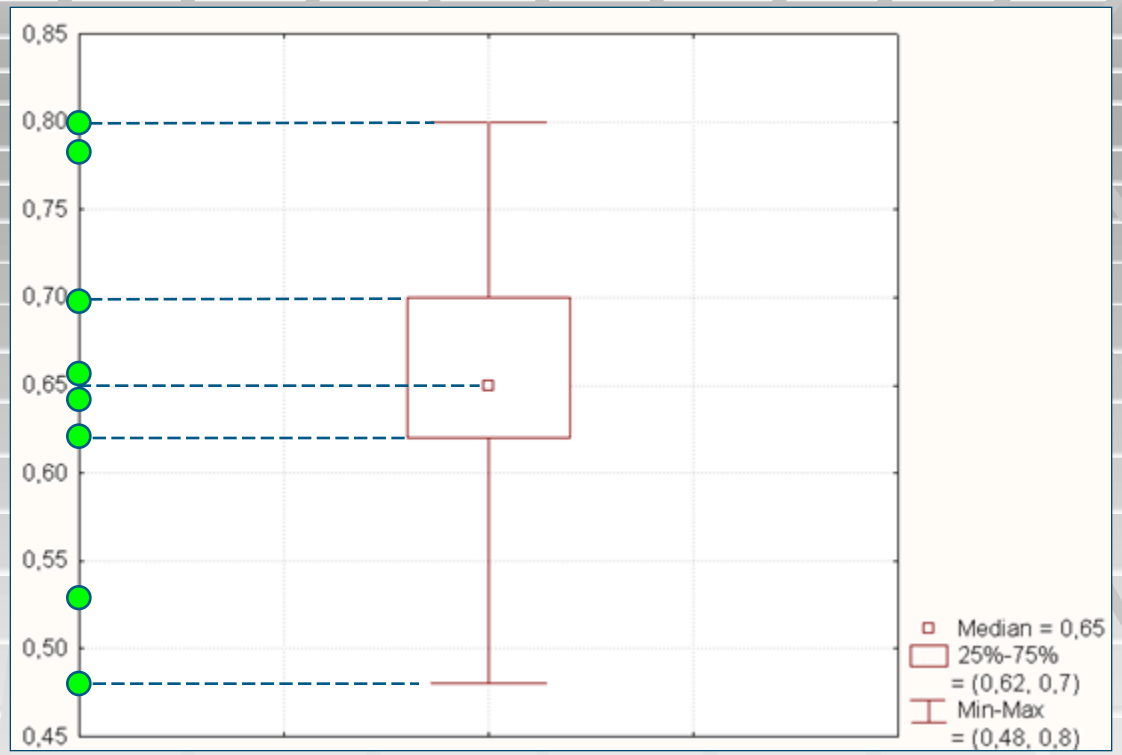
5) Найдите теоретические значения, к которым будут стремиться все перечисленные выше выборочные функции при увеличении размера выборки и сравните их с вычисленными оценками

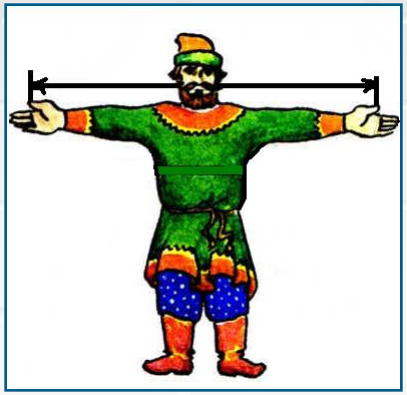
---

6) Приведите пример выборки размера 5 (нарисуйте точки на числовой оси), показывающий, что $\overline{X}_\alpha$ не обязательно находится внутри отрезка, границами которого являются оценки $\overline{X}$ и MED

Упрощённая диаграмма размахов

- Верхний и нижний «усы» — соответствуют наибольшему значению и наименьшему значению среди всех наблюдений
- Верхняя граница прямоугольника — верхняя выборочная квартиль (верхняя граница области типичных значений)
- Нижняя граница прямоугольника — нижняя выборочная квартиль (нижняя граница области типичных значений)
- Маленький прямоугольник внутри большого — выборочная медиана («центр» распределения)





В межквартильном диапазоне (устойчивой к «выбросам» области типичных значений) содержится 50% всех наблюдений.

Диаграмма размахов с «выбросами»

UBV (Up Boundary Value) — верхняя выборочная квартиль

LBV (Low Boundary Value) — нижняя выборочная квартиль

Верхний и нижний «усы» ограничивают диапазон невыделяющихся значений

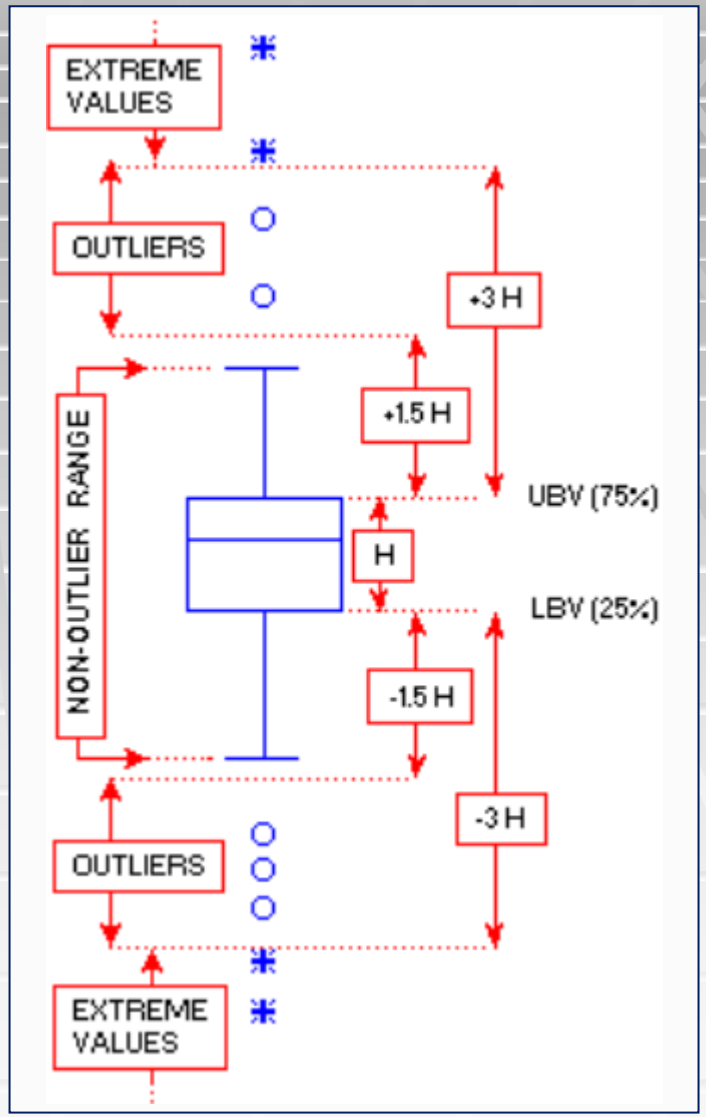
H = UBV − LBV — межквартильный размах

Горизонтальная линия внутри межквартильного диапазона — выборочная медиана

Кружки — возможные «выбросы»

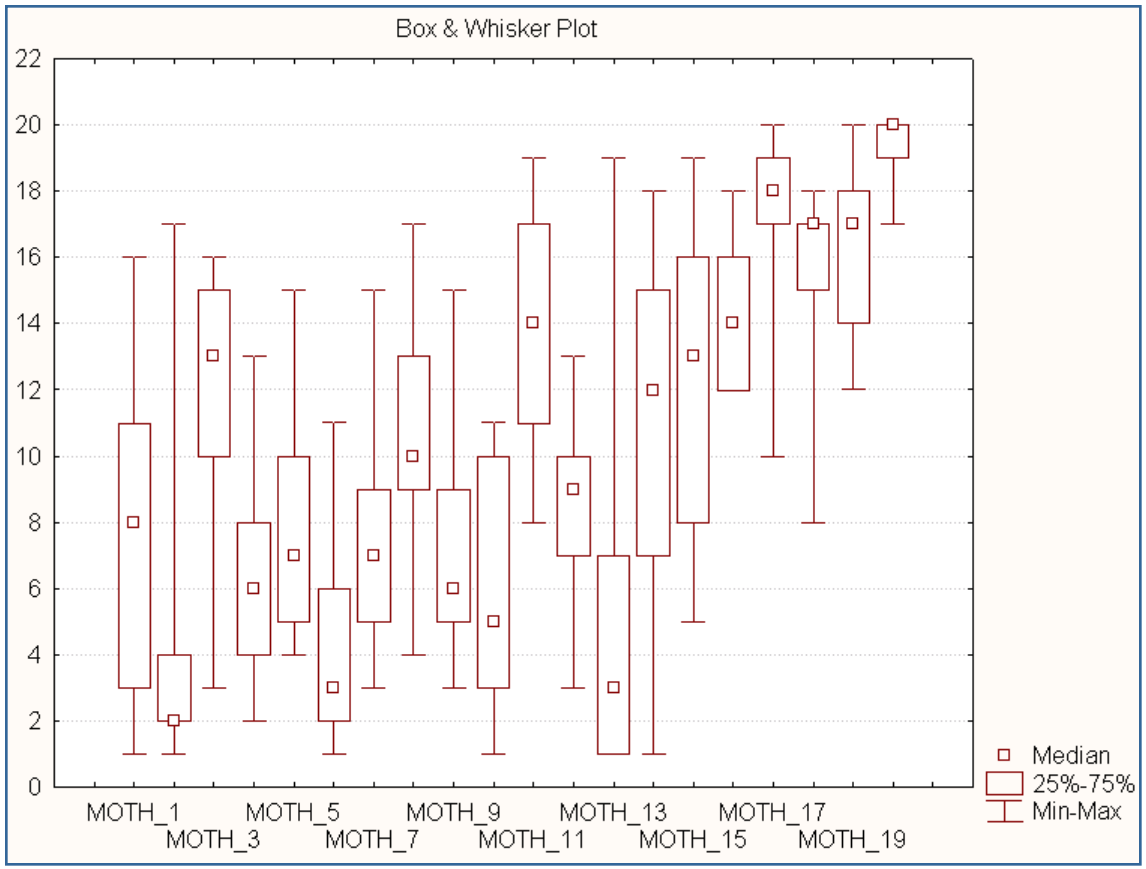
Звёздочки — экстремальные значения или безусловные «выбросы».

Строки таблицы данных, содержащие экстремальные значения, как правило, рекомендуется исключить из дальнейшего статистического анализа.



Назначение диаграммы размахов

Диаграмма размахов (Box & Whisker Plot) предназначена для обнаружения возможных (Outliers) и безусловных (Extreme Values) «выбросов», а также для быстрого сравнения распределений нескольких выборок



Практическое задание 4

1) Моделируйте методом обратной функции выборку p размера $n = 50$ из *закона Парето* с функцией распределения

$$F(x) = 1 - 1/x^2 \text{ при } x > 1$$

2) Постройте диаграмму размахов для p с помощью функции boxplot, выясните, есть ли среди наблюдений «выбросы» (OUTLIERS или EXTREME VALUES)

---

3) Вычислите порог b, отделяющий «выбросы» от остальных наблюдений (используйте функции quantile и IQR)

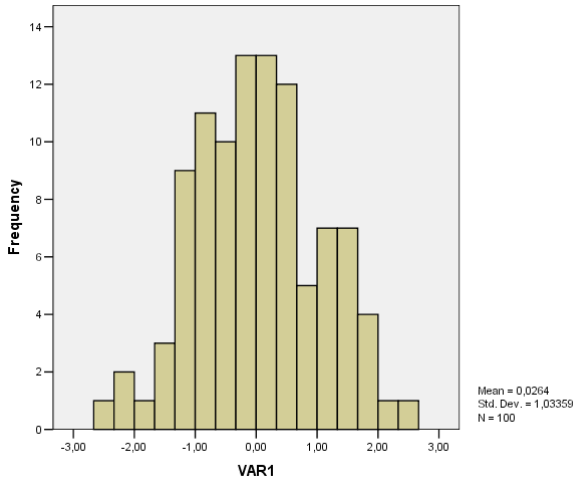
4) Если есть «выбросы», то исключите их, заменив на значения NA с помощью команды ifelse (в языке R пропущенные данные обозначаются символом NA, который происходит от англ. слов not available — нет в наличии), запишите результат в вектор q, постройте для него диаграмму размахов

---

5) Подсчитайте количество исключённых наблюдений с помощью функций sum и is.na (последняя возвращает TRUE (1) или FALSE (0) в зависимости от того, является значение пропуском или нет)

Гистограмма

Гистограмма (Histogram) — диаграмма в виде ряда столбиков. По ней можно судить о форме распределения признака. Диапазон значений признака разбивается на некоторое число равных по длине интервалов. Подсчитываются количества наблюдений, попавших в каждый интервал. Над интервалами строятся столбики, высоты которых пропорциональны подсчитанным количествам. При малой длине интервалов и большой выборке гистограмма служит оценкой плотности.

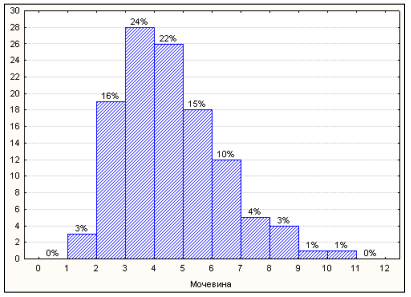


Термин был впервые использован К. Пирсоном в 1895 г.

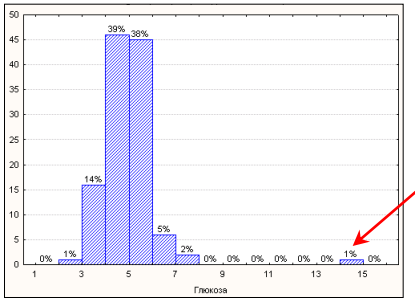
Назначение гистограммы

С помощью гистограммы можно обнаружить следующие особенности распределения признака:

а) отсутствие симметрии распределения признака:

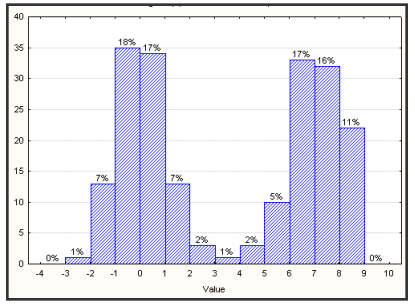


б) наличие выделяющихся значений («*выбросов*»):



«выброс»

в) отсутствие однородности данных по какому-то признаку:



Приведём три правила выбора числа интервалов гистограммы из книги Venables W. N., Ripley B. D. «Modern Applied Statistics with S» (Springer, 2002), с. 112.

В R её сопровождает пакет MASS

The standard histogram function is `hist(x, ...)` which plots a conventional histogram. More control is available via the extra arguments; `probability = T` gives a plot of unit total area rather than of cell counts.

The argument `nclass` of `hist` suggests the number of bins, and `breaks` specifies the breakpoints between bins. One problem is that `nclass` is only a suggestion, and it is often exceeded. Another is that the definition of the bins that is of the form $[x_0, x_1], (x_1, x_2], \ldots$, not the convention most people prefer.

The default for `nclass` is $\lceil \log_2 n + 1 \rceil$. This is known as Sturges' formula, corresponding to a bin width of $\text{range}(x)/(\log_2 n + 1)$, based on a histogram of a normal distribution (Scott, 1992, p. 48). Note that outliers may inflate the range dramatically and so increase the bin width in the centre of the distribution. Two rules based on compromises between the bias and variance of the histogram for a reference normal distribution are to choose bin width as

Чему равен предел H при $n \to \infty$ для выборки из закона $N(0, 1)$?

$$h = 3.5Sn^{-1/3}$$

$$h = 2Hn^{-1/3}$$

Сколько интервалов получается для $N(0, 1)$ при $n = 100$ и 1000?

due to Scott (1979) and Freedman and Diaconis (1981), respectively. Here S is the estimated standard deviation and H the inter-quartile range. The Freedman–Diaconis formula is immune to outliers, and chooses rather smaller bins than the Scott formula.

breaks = "S"

breaks = "FD"

Сегодня это действительно слишком просто: вы можете подойти к компьютеру и практически без знания того, что вы делаете, создавать разумное и бессмыслицу с поистине изумительной быстротой.

Дж. Бокс

Домашнее задание

В файле Prefix-ver.txt в столбце LogFrequency содержатся логарифмы частоты встречаемости 985 голландских слов с приставкой ver.

1) В правом верхнем окне RStudio щёлкните по Import Dataset, выберите пункт From Text (base)…, импортируйте файл Prefix-ver.txt под кратким именем d (новое имя d надо указать в поле Name во время импортирования файла)

2) Постройте гистограмму для признака LogFrequency с помощью функции hist (чтобы закрасить гистограмму серым цветом, установите аргумент функции col=8)

3) Ещё раз постройте гистограмму, задав количество интервалов гистограммы примерно в 2 раза больше, чем предлагается по умолчанию

4) Сколько истинных (не объясняемых случайностью) «горбов» (локальных максимумов) имеет распределение признака?

Продолжение домашнего задания

5) Отберите с помощью функции subset в новую таблицу s строки из таблицы d с ненулевыми значениями признака LogFrequency и значениями opaque признака SemanticClass
(opaque — трудные для понимания, несоставные слова;
transparent — семантически простые, составные слова)

6) Вычислите для отобранных значений признака LogFrequency следующие характеристики:

а) выборочное среднее,

б) усечённое среднее с параметром $\alpha = 0{,}05$,

в) межквартильный размах.