Системы и технологии интеллектуальной обработки данных

Сухорукова Ирина Геннадьевна

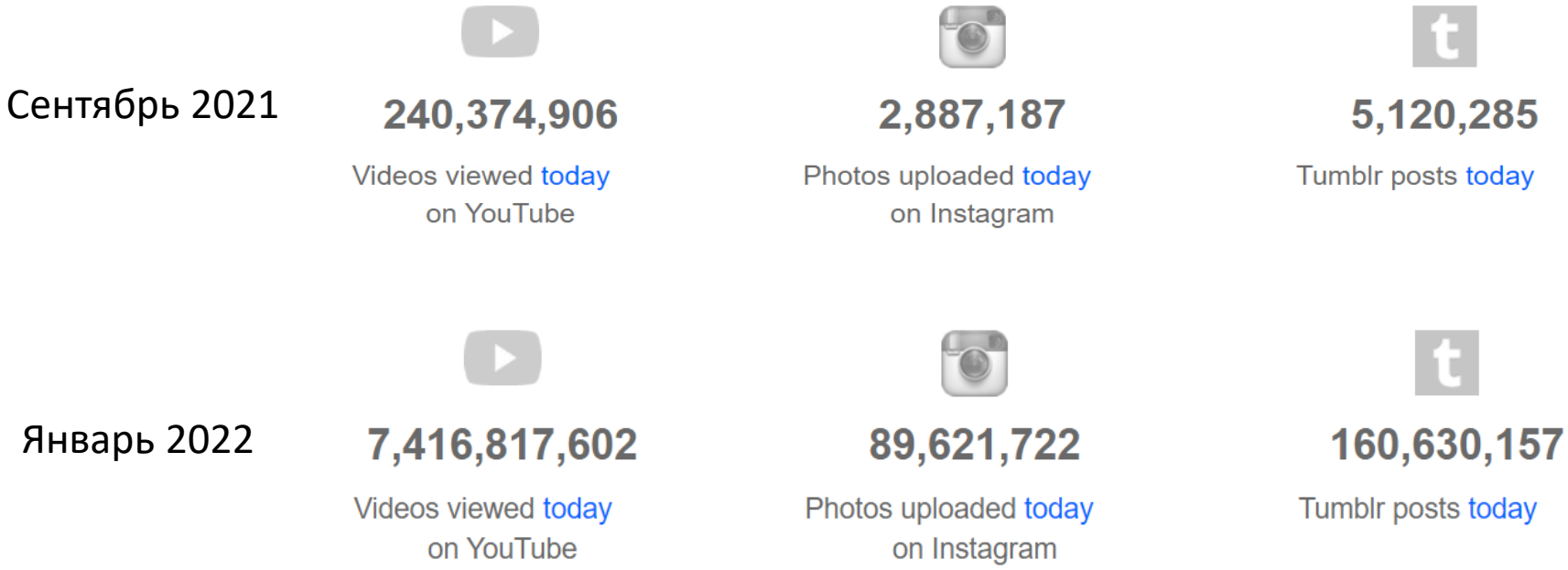
ст. преподаватель кафедры программной инженерии

Контакты: ауд.408 к.1

Лабораторные занятия – Python, Jupyter Notebook, (PyCharm)

В 2020 году общий объём сгенерированных данных **составил 64,2 зеттабайта***. Сообщается, что для дальнейшего использования было сохранено менее 2% информации. Основная часть данных имела временный характер. (интересно, что в 2016г. на 2020г. **прогнозировали 44 зеттабайта**)

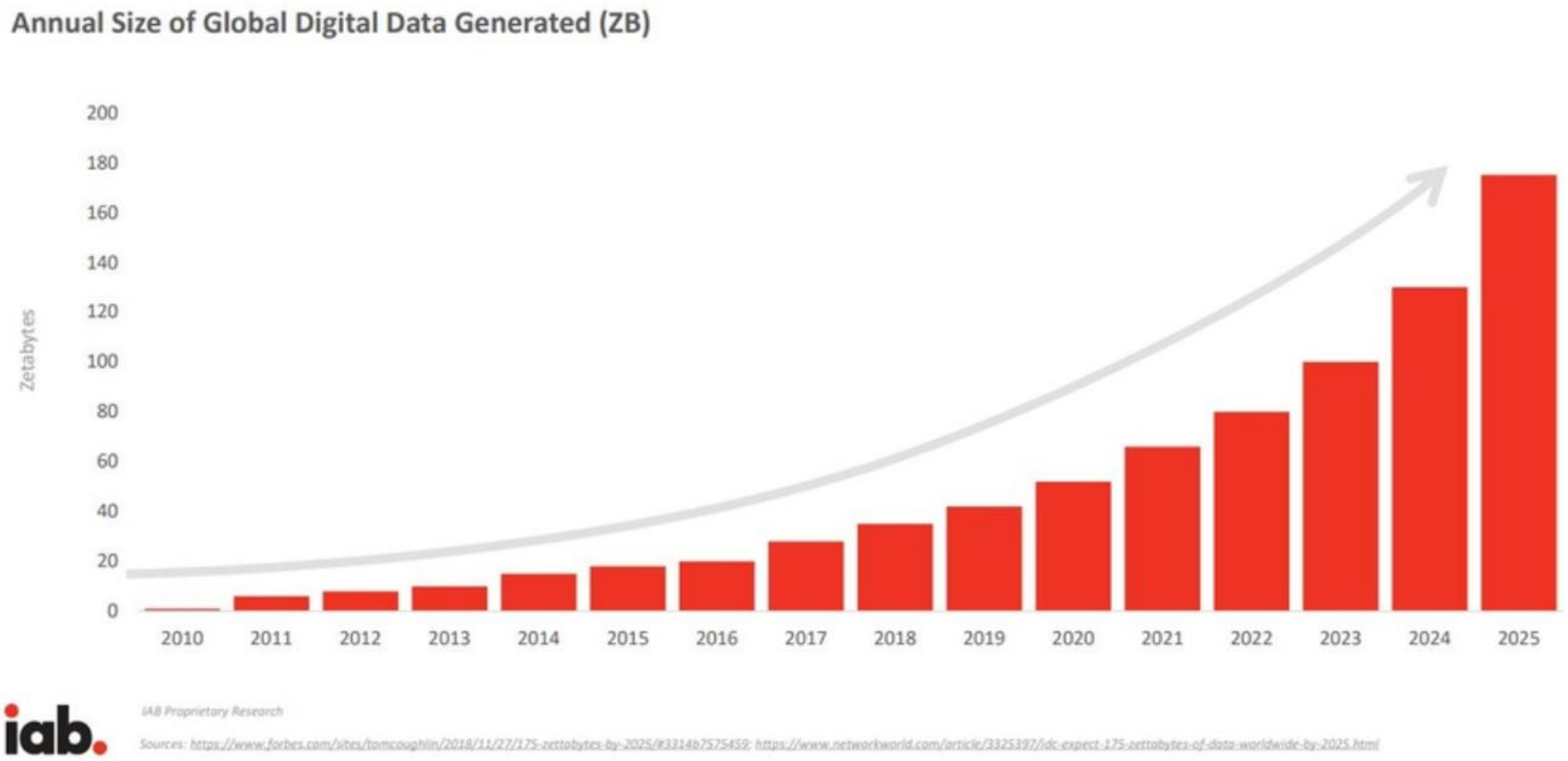
Сайт internetlivestats.com показывает, что происходит в интернете в режиме реального времени



Из-за огромного количества информации очень малая ее часть будет когда-либо увидена человеческим глазом. **Единственная надежда понять и найти что-то полезное в этом океане информации — широкое применение методов Data Mining.**

* 1 зеттабайт — это миллиард терабайтов

Объём генерируемых цифровых данных в мире



[Сколько данных в мире было создано в 2020 году | Наука | Селдон Новости (myseldon.com)]

Интеллектуальный анализ данных

В узком смысле это попытка адекватного русского перевода термина **Data Mining**, который ввёл в обиход Григорий Пятецкий-Шапиро в 1992 году.

Английское словосочетание «data mining» не имеет устоявшегося перевода на русский язык, в русском языке как правило используется термин **интеллектуальный анализ данных.** Более полным и точным является словосочетание «обнаружение знаний в базах данных» (*knowledge discovery in databases, KDD*).

*В настоящее время **data mining** является частью большего понятия – **Big data**, которое помимо обработки данных включает в себя их сбор и хранение.*

Data mining —это автоматизированный поиск данных, основанный на анализе огромных массивов информации.

В широком смысле это современная концепция анализа данных, предполагающая, что:

- ✓ данные могут быть неточными, неполными (содержать пропуски), противоречивыми, разнородными, косвенными, и при этом иметь гигантские объёмы; поэтому понимание данных в конкретных приложениях требует значительных интеллектуальных усилий;
- ✓ сами алгоритмы анализа данных могут обладать «элементами интеллекта», в частности, способностью обучаться по прецедентам, то есть делать общие выводы на основе частных наблюдений; разработка таких алгоритмов также требует значительных интеллектуальных усилий;
- ✓ процессы переработки сырых данных в информацию, а информации в знания уже не могут быть выполнены по старинке вручную, и требуют нетривиальной автоматизации.

Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

Знания должны быть новые, ранее неизвестные. Затраченные усилия на открытие знаний, которые уже известны пользователю, не окупаются. Поэтому ценность представляют именно новые, ранее неизвестные знания.

Знания должны быть нетривиальны. Результаты анализа должны отражать неочевидные, неожиданные закономерности в данных, составляющие так называемые скрытые знания. Результаты, которые могли бы быть получены более простыми способами (например, визуальным просмотром), не оправдывают привлечение мощных методов Data Mining.

Знания должны быть практически полезны. Найденные знания должны быть применимы, в том числе и на новых данных, с достаточно высокой степенью достоверности. Полезность заключается в том, чтобы эти знания могли принести определенную выгоду при их применении.

Знания должны быть доступны для понимания человеку. Найденные закономерности должны быть логически объяснимы, в противном случае существует вероятность, что они являются случайными. Кроме того, обнаруженные знания должны быть представлены в понятном для человека виде.

Области практического применения Data Mining

Технология Data Mining используется практически во всех сферах деятельности человека, где накоплены ретроспективные данные.

- ✓ БАНКОВСКОЕ ДЕЛО
- ✓ ТЕЛЕКОММУНИКАЦИИ
- ✓ МАРКЕТИНГ
- ✓ ИНТЕРНЕТ-ТЕХНОЛОГИИ
- ✓ ТОРГОВЛЯ
- ✓ МЕДИЦИНА …
- ✓ ФАРМАЦЕВТИКА…
- ✓ МОЛЕКУЛЯРНАЯ ГЕНЕТИКА и ГЕННАЯ ИНЖЕНЕРИЯ…

…..

Области практического применения Data Mining

Сейчас технология Data Mining используется практически во всех сферах деятельности человека, где накоплены ретроспективные данные.

БАНКОВСКОЕ ДЕЛО: анализ кредитоспособности клиента, привлечение новых клиентов, мошенничество с карточками, сегментация.

ТЕЛЕКОММУНИКАЦИИ: удержание клиента, выявления определенных групп клиентов, и разработка наборов услуг, наиболее привлекательных именно для них.

МАРКЕТИНГ: В сфере маркетинга Data Mining находит очень широкое применение. Основные вопросы маркетинга "Что продается?", "Как продается?", "Кто является потребителем?”

ИНТЕРНЕТ-ТЕХНОЛОГИИ: формирование рекомендательных систем, планирование маркетинговой политики в соответствии с обнаруженными интересами и потребностями клиентов.

ТОРГОВЛЯ: анализ рыночных корзин с целью регулирования предложениями, выделение групп потребителей со схожими стереотипами поведения, т.е. сегментирование рынка.

МЕДИЦИНА …

ФАРМАЦЕВТИКА…

МОЛЕКУЛЯРНАЯ ГЕНЕТИКА и ГЕННАЯ ИНЖЕНЕРИЯ…

…..

Data Mining носит мультидисциплинарный характер, поскольку включает в себя элементы численных методов, математической статистики и теории вероятностей, теории информации и математической логики, искусственного интеллекта и машинного обучения.

Data mining основывается на 3-х понятиях:

- ✓ **Математическая статистика** – является основой большинства технологий, используемых для data mining, например, кластерный анализ, регрессионный анализ, дискриминирующий анализ и пр.;
- ✓ **Искусственный интеллект** – воспроизведение нейронной сети мышления человека в цифровом виде;
- ✓ **Машинное обучение** – совокупность статистики и искусственного интеллекта, способствующая пониманию компьютерами данных, которые они обрабатывают для выбора наиболее подходящего метода или методов анализа.


Машинное обучение
Классическое обучение
Без учителя
С учителем
Кластеризация
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Классификация
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Поиск правил
Euclat
Apriori
FP-Growth
Регрессия
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Уменьшение размерности (обобщение)
t-SNE
PCA
LSA
SVD
LDA
Обучение с подкреплением
Генетический Алгоритм
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
Ансамблевые методы
Стекинг
Беггинг
Random Forest
Бустинг
AdaBoost
XGBoost
LightGBM
CatBoost
Нейросети

Ваши доклады

Структура доклада

1. Датасет отличный от примера из прошлого курса (Kaggle, . . .)
2. Подготовка данных (preprocessing)
3. Применение метода ML (train)
4. Оценка метода с помощью метрик качества
5. Выводы

Описательная статистика в анализе данных

Стати́стика — отрасль знаний, **наука**, в которой излагаются **общие вопросы сбора, измерения, анализа** массовых статистических (количественных или качественных) данных и их **сравнение**; **изучение количественной стороны** массовых явлений в числовой форме.

Слово «статистика» происходит от латинского **status** — состояние и положение дел.

В науку термин «статистика» ввёл немецкий учёный Готфрид Ахенвалль в 1746 году, предложив заменить название курса «Государствоведение», преподававшегося в университетах Германии, на «Статистику», положив тем самым начало развитию статистики как науки и учебной дисциплины.

Несмотря на это, статистический учёт вёлся намного раньше: *проводились переписи населения в Древнем Китае, осуществлялось сравнение военного потенциала государств, вёлся учёт имущества граждан в Древнем Риме и тому подобное.*

https://habr.com/ru/company/piter/blog/320772/

ph_piter 30 января 2017 в 10:37

Разница между статистикой и наукой о данных

Блог компании Издательский дом «Питер» , Data Mining *, Алгоритмы *, Big Data *, R *

Мнение: Статистикам весь тренд, связанный с наукой о данных, кажется слегка высокомерным. . . эта сфера деятельности весьма пересекается с той работой, которой статистики занимаются уже не одно десятилетие.

“Думаю, data-scientist – распиаренный синоним для «специалист по статистике»” – заявил *Нейт Сильвер** в 2013 году на лекции в Joint Statistical Meeting.

Брэд Шлумич специалист по data science в Twitch: “**Статистика – важнейшая составляющая науки о данных**. У нас в Twitch команда data science обладает тремя компетенциями: статистика, программирование и понимание продукта. **Мы никогда не взяли бы на работу человека, слабо ориентирующегося в статистике**.”

“Некоторые считают, что наука о данных – это всего лишь прикладная статистика, но мы – определенно не просто статистики. . . Гораздо эффективнее работать, если все одинаково понимают смысл продукта, решают, какие параметры важнее, понимают с точки зрения программиста, как реализовать трекинг, и с точки зрения статистика – как делать анализ. Не понимая, как люди будут пользоваться продуктом, и каковы цели компании, можно исказить весь анализ данных. Задача data scientist'a – держать в голове сразу всю эту информацию и знать, к каким данным обратиться, чтобы ответить на любой нечетко определенный вопрос.

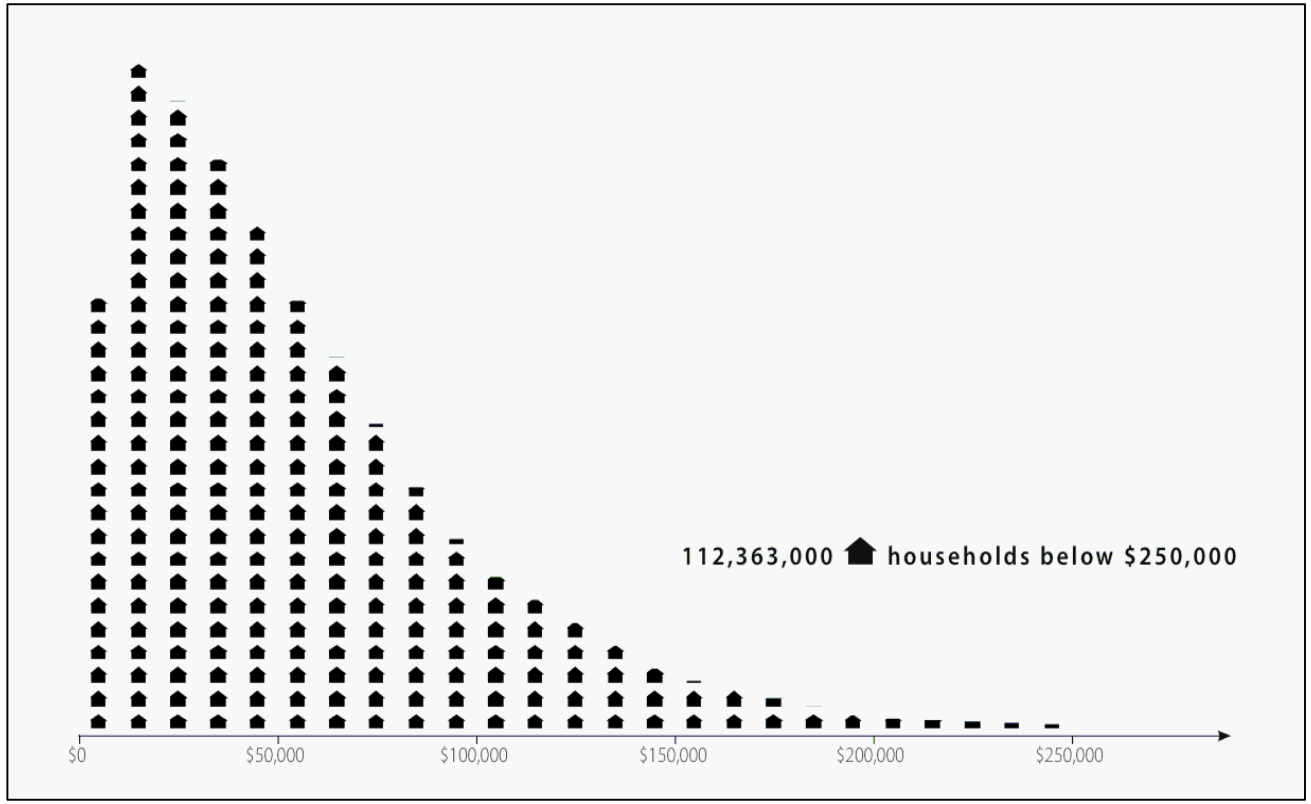
** Нейт Сильвер (Nate Silver) - тот самый человек, который верно спрогнозировал итоги голосования на президентских выборах 2008 года в 49 из 50 штатов США. В 2012 году у него получилось уже 50 из 50.*

Базовые знания статистики крайне полезны в повседневной жизни.

Например, в 2005 году британские СМИ писали о том, что средний уровень дохода населения снизился на 0,2 % по сравнению с предыдущим годом. Некоторые политики даже использовали этот факт, критикуя действующее правительство.

Однако, важно понимать, что среднее арифметическое — хороший показатель, когда наш признак имеет симметричное распределение (богатых столько же, сколько бедных). Реальное же распределение доходов имеет скорее следующий вид:

Распределение имеет явно выраженную асимметрию: очень состоятельных людей заметно меньше, чем представителей среднего класса. Это приводит к тому, что **в данном случае банкротство одного из миллионеров может значительно повлиять на этот показатель.**
Гораздо **информативнее использовать значение медианы для описания таких данных.** И, как ни удивительно, медиана дохода в 2005 году в Великобритании, в отличие от среднего значения, продолжила свой рост.



Основы статистики: просто о сложных формулах / Хабр (habr.com)

Крылатая фраза: *Существует три вида лжи: ложь, наглая ложь и статистика*

Таблица 2.3. Количество выпадений каждой цифры (Kansas Pick 3 Lottery, 15 марта 1997 года)

Цифра	Количество выпадений
0	485
1	468
2	513
3	491
4	484
5	480
6	487
7	482
8	475
9	474

Рис. 2.1. Столбиковая диаграмма, показывающая количество выпадений каждой цифры

Таблица 2.4. Процент выпадений каждой цифры

Цифра	Количество выпадений	Процент выпадений
0	485	10,0% = 485/4 839
1	468	9,7% = 486/4 839
2	513	10,6% = 513/4 839
3	491	10,1% = 491/4 839
4	484	10,0% = 484/4 839
5	480	9,9% = 480/4 839
6	487	10,0% = 487/4 839
7	482	10,0% = 482/4 839
8	475	9,8% = 475/4 839
9	474	9,8% = 474/4 839

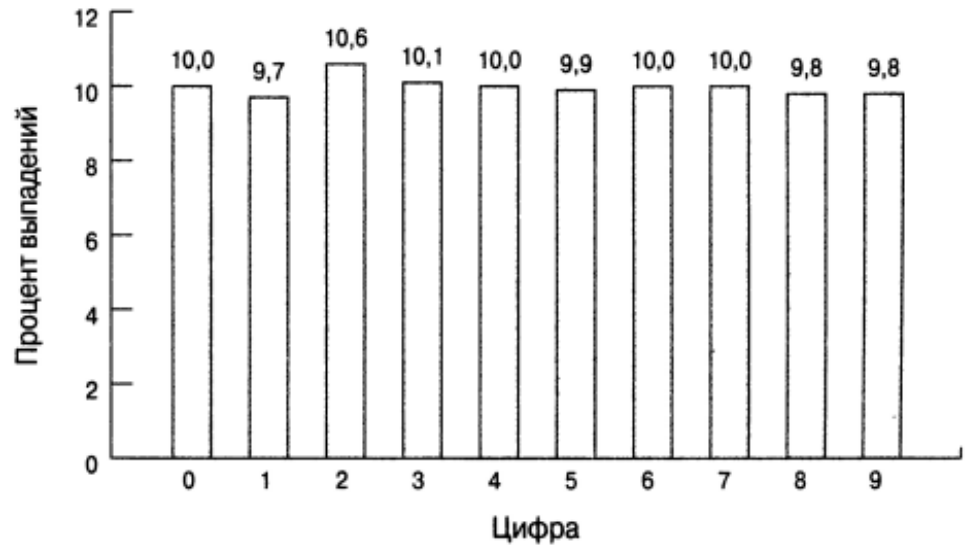


Рис. 2.2. Столбиковая диаграмма, показывающая процентное отношение количества выпадений каждой цифры

Генеральная совокупность

Суммарная численность объектов наблюдения, обладающих определенным набором признаков, ограниченная в пространстве и времени.

Выборка (Выборочная совокупность)

Часть объектов из генеральной совокупности, отобранных для изучения, с тем чтобы сделать заключение о всей генеральной совокупности. Для того чтобы заключение, полученное путем изучения выборки, можно было распространить на всю генеральную совокупность, выборка должна обладать свойством репрезентативности.

Репрезентативность выборки

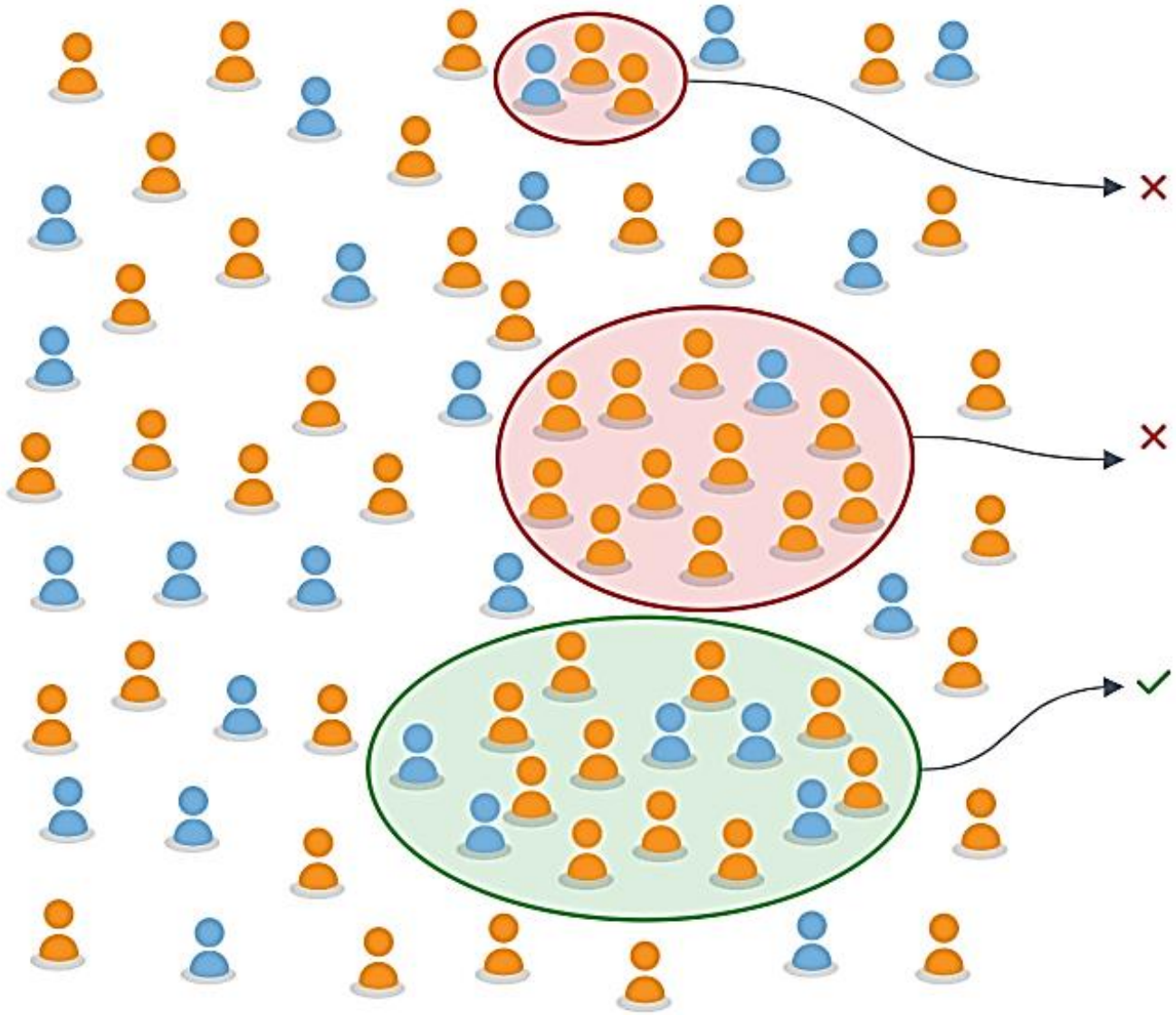
Свойство выборки корректно отражать генеральную совокупность.

Примеры:

- ✓ Выборка, целиком состоящая из горожан, владеющих автомобилем, не репрезентирует все население города.
- ✓ Выборка только из женщин не репрезентирует все население.

Генеральная совокупность включает

- 1/3 и - 2/3



Слишком маленькая выборка

Нерепрезентативная выборка

Репрезентативная выборка

Набор данных и их атрибутов

По горизонтали таблицы располагаются *атрибуты* объекта или его признаки. *По* вертикали таблицы - *объекты*.

Объект описывается как набор атрибутов.

Объект также известен как *запись*, случай, пример, строка таблицы и т.д.

Атрибут - свойство, характеризующее *объект*.

Например: цвет глаз, возраст.

Атрибут также называют переменной, полем таблицы, характеристикой.

Переменная (variable) - свойство или характеристика, общая для всех изучаемых *объектов*, проявление которой может изменяться от *объекта* к *объекту*.

Значение (value) переменной является проявлением признака.

	Атрибуты			
	Код клиента	Возраст	Семейное положение	Доход
	1	18	Single	125
	2	22	Married	100
	3	30	Single	70
	4	32	Married	120
Объекты	5	24	Divorced	95
	6	25	Married	60
	7	32	Divorced	220
	8	19	Single	85

	Атрибуты			
Объекты	Код клиента	Возраст	Семейное положение	Доход
	1	18	Single	125
	2	22	Married	100
	3	30	Single	70
	4	32	Married	120
	5	24	Divorced	95
	6	25	Married	60
	7	32	Divorced	220
	8	19	Single	85

Количественные переменные:

- Дискретные данные являются значениями признака, общее число которых конечно либо бесконечно, но может быть подсчитано при помощи натуральных чисел от одного до бесконечности.
  Пример дискретных данных: 10, 15, 25 мин, количество детей . . .
- Непрерывные данные - данные, значения которых могут принимать какое угодно значение в некотором интервале. Измерение непрерывных данных предполагает большую точность.
  Пример непрерывных данных: температура, высота, вес, длина и т.д.

Качественные (номинативные) переменные

Такие переменные используются для разделения наших испытуемых или наблюдений на группы.

Например, мы можем сказать, что все участники эксперимента женского пола будут обозначены цифрой 1, а все участники мужского пола - цифрой 2 соответственно.

Таким образом, в случае номинативных переменных за цифрами не стоит никакого математического смысла. В данном случае цифры используются как маркеры различных смысловых групп, в отличие от количественных переменных.

Ранговые переменные

Представьте, что у нас есть информация о марафонском забеге: кто прибежал в каком порядке. Мы можем сказать, что испытуемый с рангом 1 быстрее, выше, сильнее испытуемого с рангом 5. Но вот насколько или во сколько он опережает этого испытуемого мы сказать не можем. Единственной возможной математической операцией является сравнение - кто быстрее, а кто медленнее.

Data Science Job Salaries

Data | Code (173) | Discussion (10) | Metadata

844 | New Notebook

ds_salaries.csv (36.96 kB)

Detail | Compact | Column

10 of 12 columns

job_title Job Title	salary Salary	salary_currency Salary Currency	salary_in_usd Salary in USD	employee_reside... Employee Residence	rer Remo
Data Scientist 24% Data Engineer 22% Other (332) 55%	4000 … 30.4m	USD 66% EUR 16% Other (114) 19%	2859 … 600k	US 55% GB 7% Other (231) 38%	0
Data Engineer	80000	USD	80000	US	100
Director of Data Science	250000	USD	250000	US	0
BI Data Analyst	55000	USD	55000	US	50
Data Architect	150000	USD	150000	US	100
Data Architect	170000	USD	170000	US	100

Описательная статистика — это описание наборов данных.

Описательная статистика использует два основных подхода:

- ✓ **Количественный подход**, который описывает общие численные характеристики данных.
- ✓ **Визуальный подход**, который иллюстрирует данные с помощью диаграмм, графиков, гистограмм и прочих графических образов.

Метрики описательной статистики:

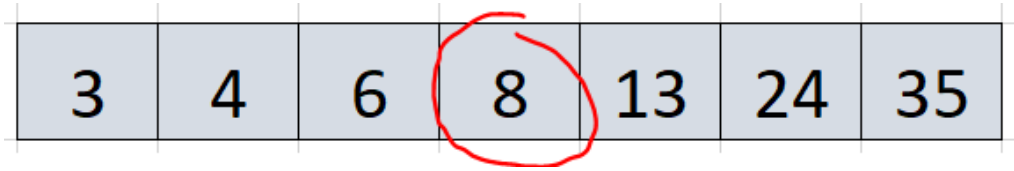
- ✓ **Метрики центрального положения**, которые говорят вам о центрах концентрации данных, таких как *среднее*, *медиана* и *мода*.

- ✓ **Метрики оценки вариативности** данных, которые говорят о разбросе значений, такие как *дисперсия* и *стандартное* ***отклонение***.

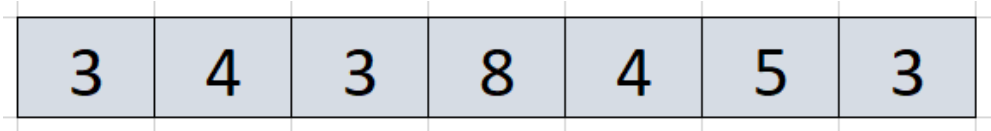
Метрики центрального положения

Среднее (mean)
Сумма всех значений, деленная на количество значений или среднее арифметическое.

Медиана (median)
Середина в отсортированных данных.



Мода (mode)



мода=3

Значение, которое встречается наиболее часто.

Мода часто употребляется для текстовых данных.
Например: цвета автомобилей — белый, чёрный, синий металлик, белый, синий металлик, белый. Какая мода?

Выброс (outlier)
Значение данных, которое сильно отличается от большинства данных.

Метрики оценки вариативности

Размах (range) Разница между самым большим и самым малым значениями в наборе данных.

Отклонения (deviations) Разница между наблюдаемыми значениями и оценкой центрального положения. Еще называют: *ошибки, остатки*.

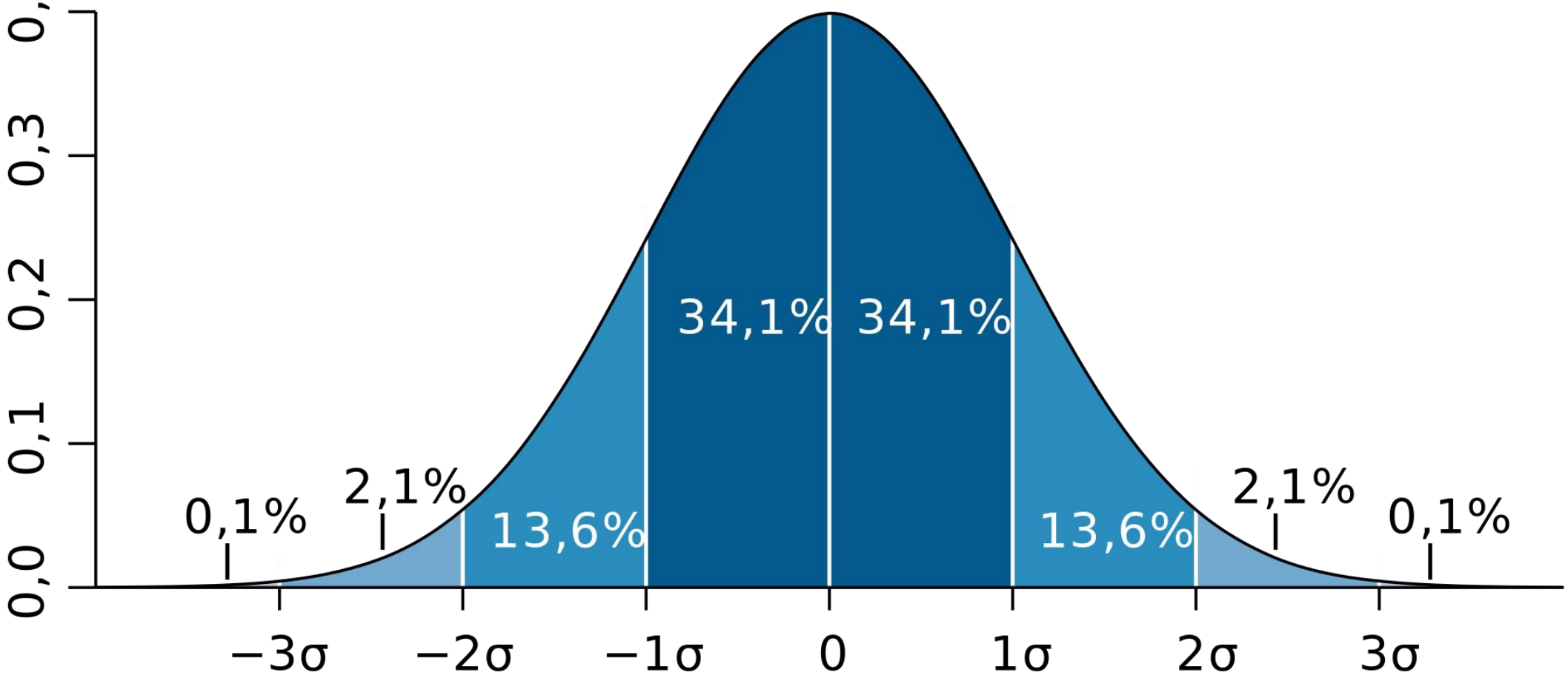
Дисперсия (variance) Сумма квадратических отклонений от среднего, деленная на n – 1, где n — число значений данных. Еще называют : *среднеквадратическое отклонение, среднеквадратическая ошибка*.

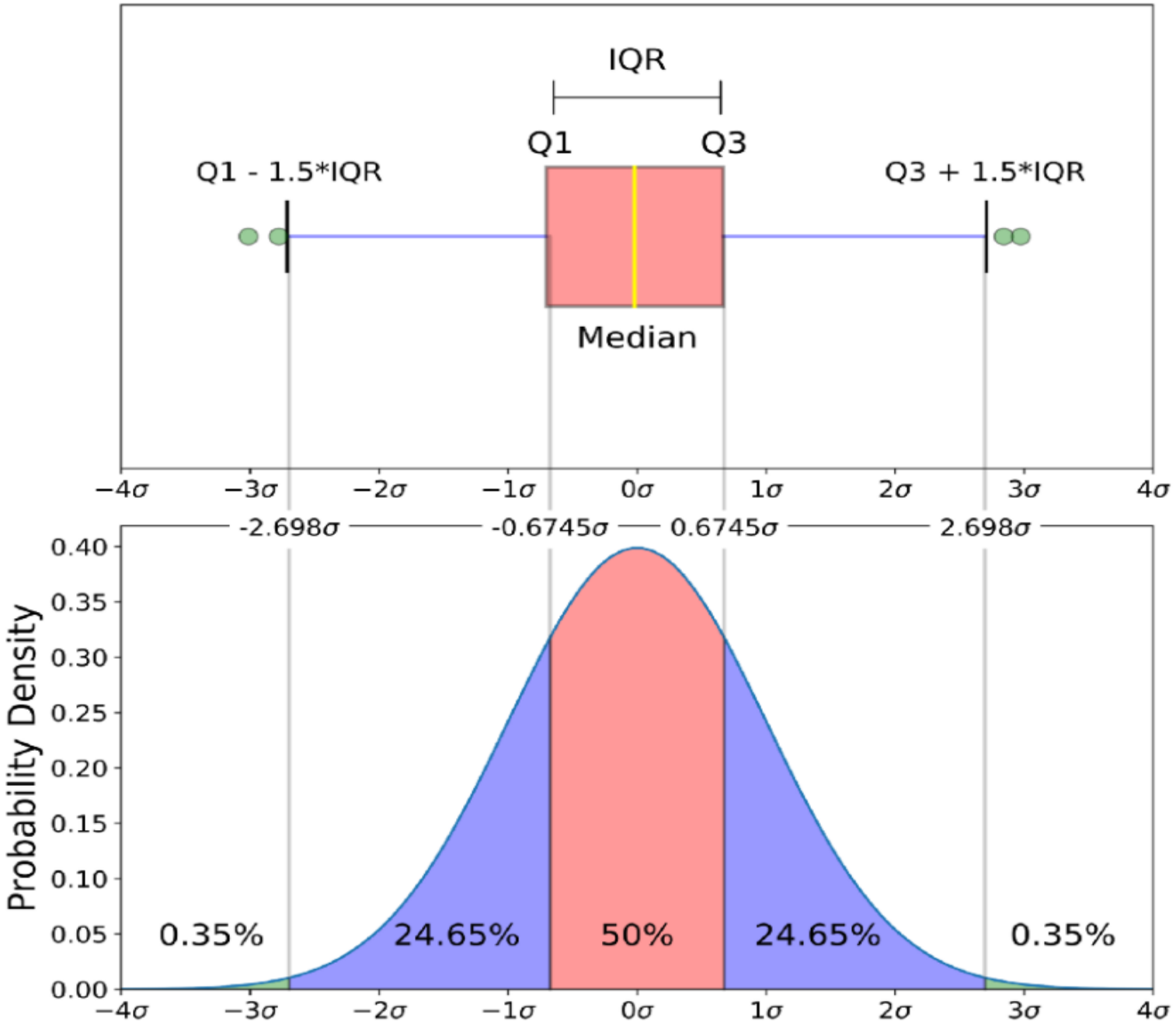
Стандартное отклонение (standard deviation) Квадратный корень из дисперсии.

Процентиль – например, **75**-й процентиль – это число, ниже которого находится **75**% всех наблюдений.

Нормальное распределение

δ — среднеквадратическое отклонение

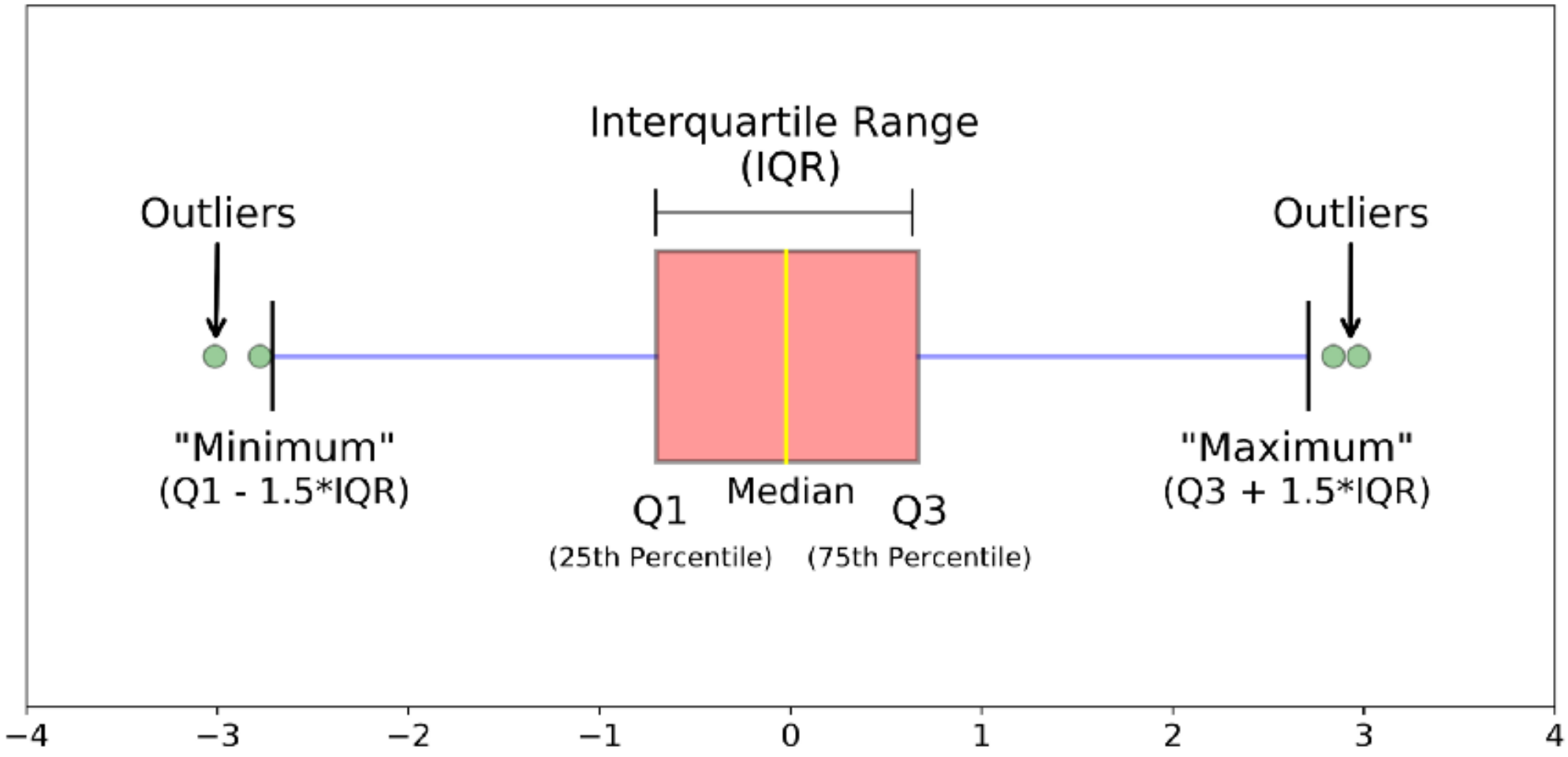




box plot

Ящик с усами

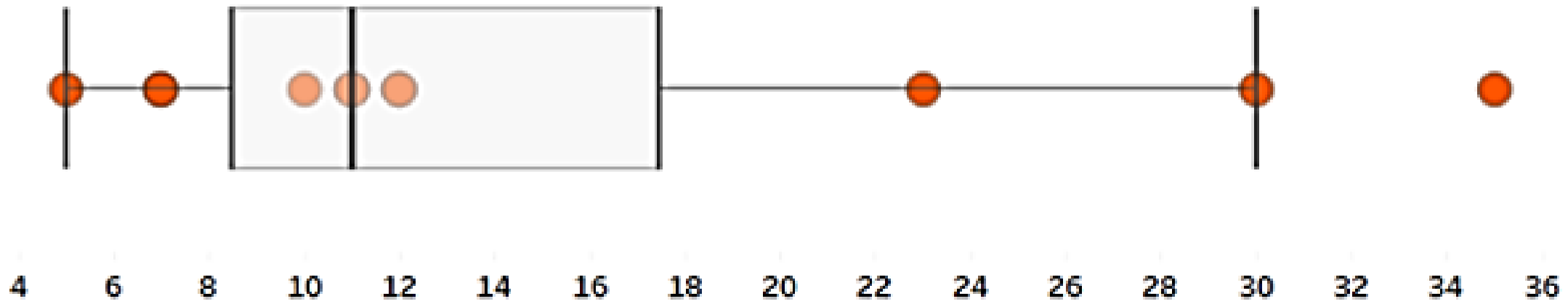
Диаграмма размаха



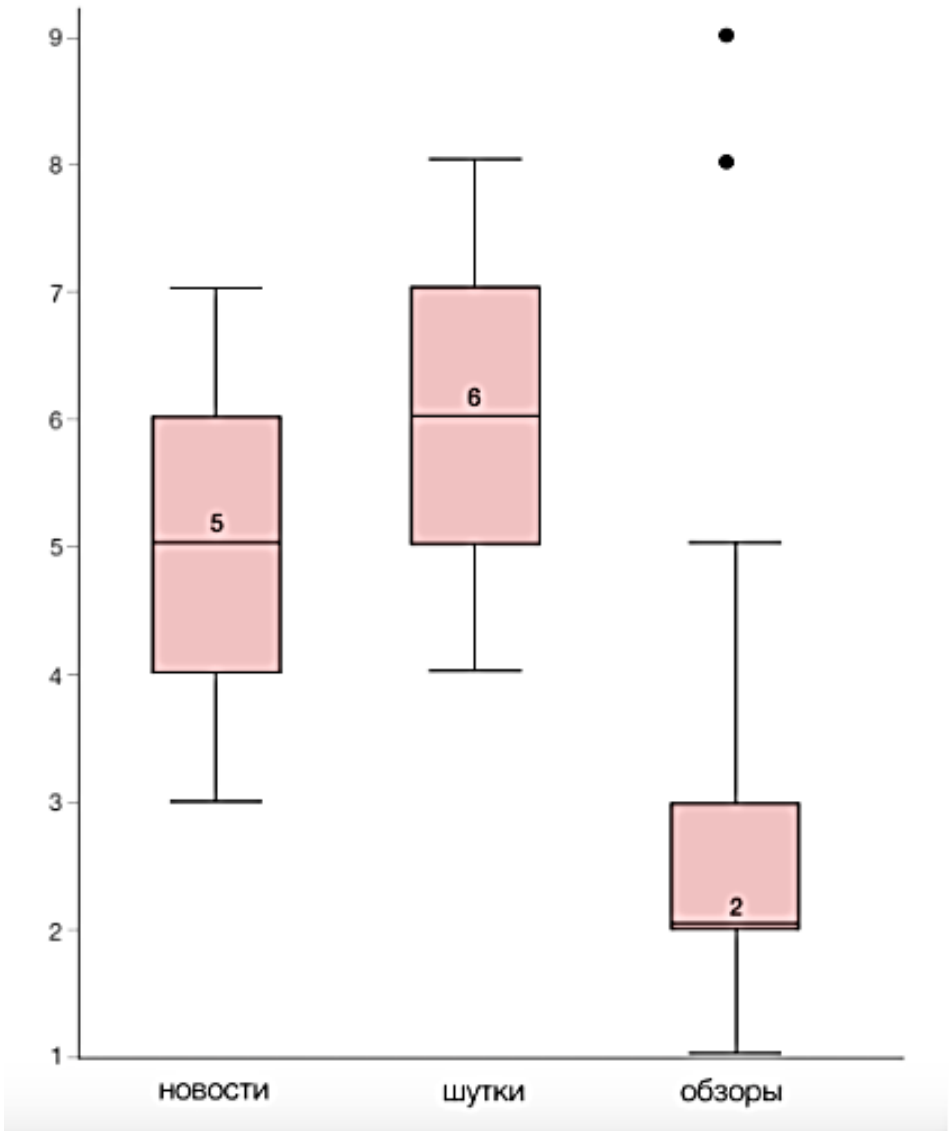
IQR - размах

Q1 – 25 перцентиль (1й квартиль)

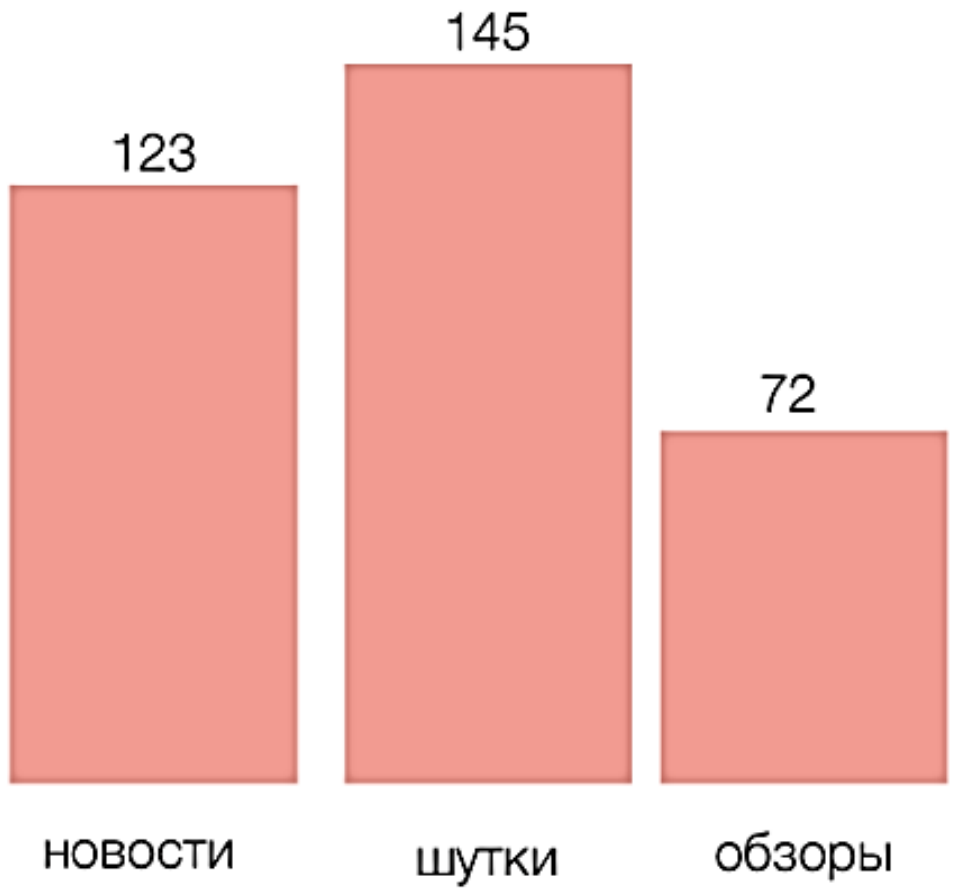
Q3 – 75 перцентиль (3й квартиль)

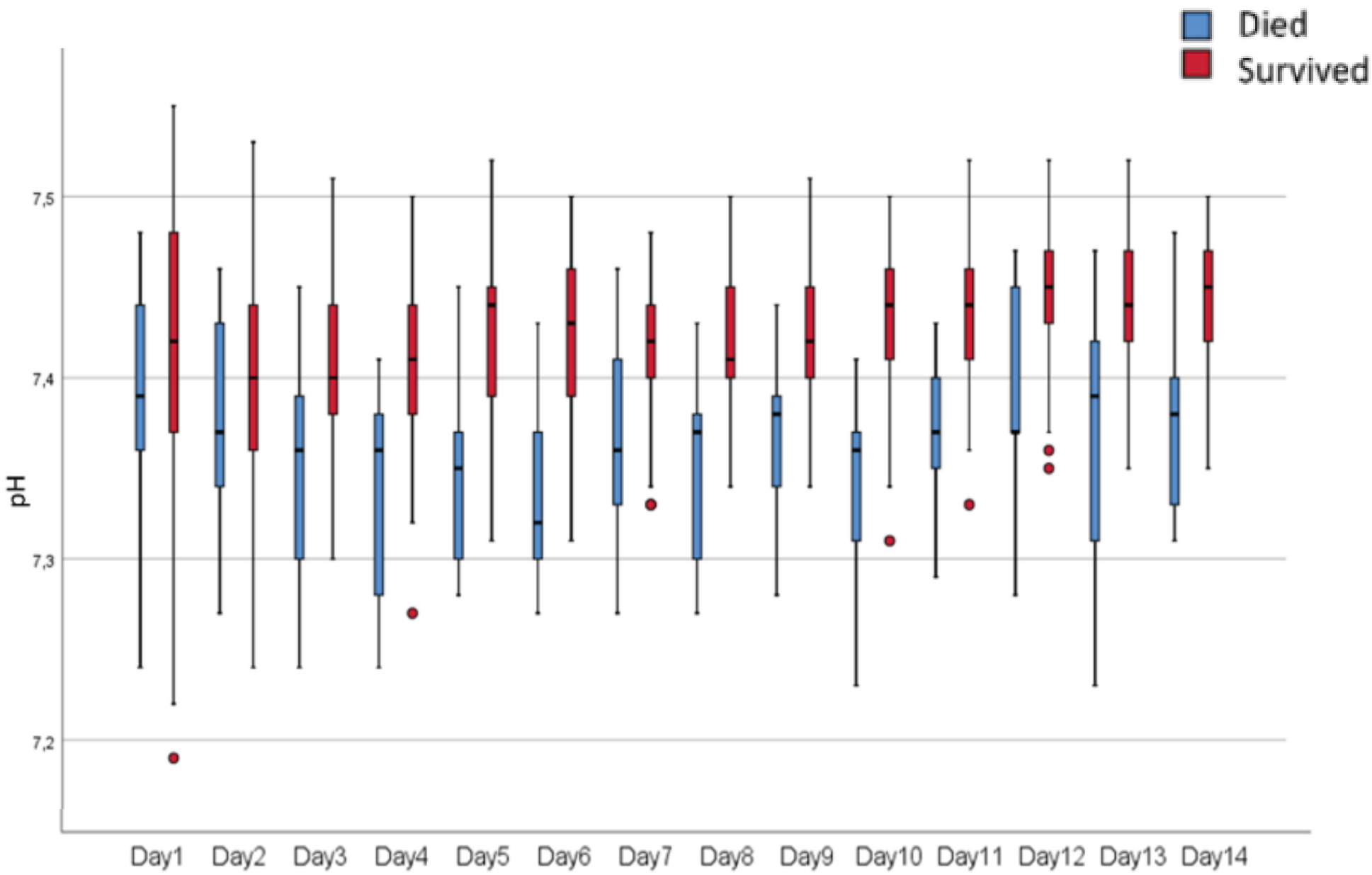

4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36

Распределение лайков по категориям



Сумма лайков по категориям





В немецком исследовании 2021г. было обнаружено, что pH крови позволяет прогнозировать вероятность смерти от ковида. Среди госпитализированных в отделении интенсивной терапии пациентов умерли в основном те, у кого был низкий уровень pH.

Пример

In [1]:
```
import pandas as pd
```

In [2]:
```
titanic = pd.read_csv("data/titanic.csv")
```

Как рассчитать сводную статистику?

Агрегированная статистика

Какой средний возраст пассажиров Титаника?

In [6]:
```
titanic["Age"].mean()
```

Out[6]:
```
29.69911764705882
```

Каков средний возраст и стоимость билета пассажиров «Титаника»?

In [7]:
```
titanic[["Age", "Fare"]].median()
```

Out[7]:
```
Age     28.0000
Fare    14.4542
dtype: float64
```

Функция **pandas.DataFrame.describe** рассчитывает параметры описательной статистики

In [8]:
```
titanic[["Age", "Fare"]].describe()
```

Out[8]:

	Age	Fare
count	714.000000	891.000000
mean	29.699118	32.204208
std	14.526497	49.693429
min	0.420000	0.000000
25%	20.125000	7.910400
50%	28.000000	14.454200
75%	38.000000	31.000000
max	80.000000	512.329200

Вместо предопределенной статистики можно определить конкретные комбинации агрегированной статистики для заданных столбцов с помощью **DataFrame.agg()метода**:

In [10]:
```
titanic.agg(
    {
        "Age": ["min", "max", "median"],
        "Fare": ["min", "max", "median", "mean"],
    }
)
```

Out[10]:

	Age	Fare
min	0.42	0.000000
max	80.00	512.329200
median	28.00	14.454200

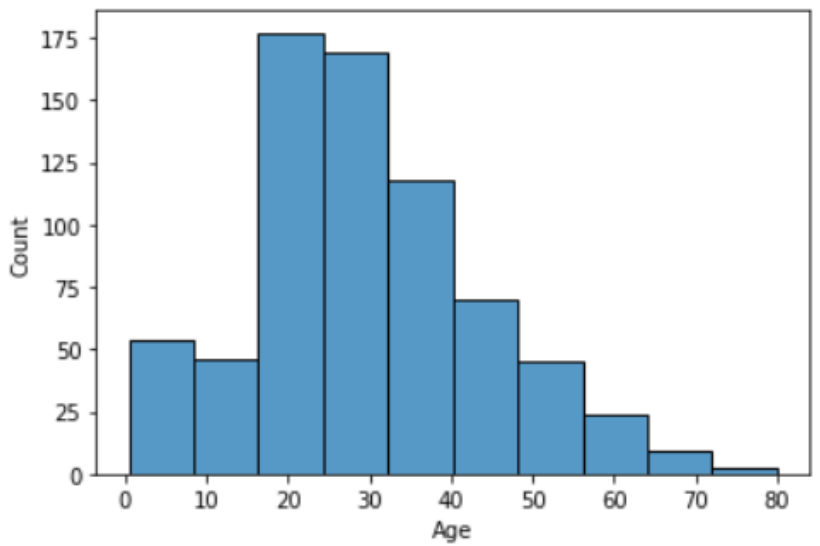
Процентиль — мера, в которой процентное значение общих значений равно этой мере или меньше ее.

In [3]:
```
titanic["Age"].describe(percentiles=[0.05, 0.25, 0.75, 0.95])
```

Out[3]:
```
count    714.000000
mean      29.699118
std       14.526497
min        0.420000
5%         4.000000
25%       20.125000
50%       28.000000
75%       38.000000
95%       56.000000
max       80.000000
Name: Age, dtype: float64
```

In [4]:
```
import seaborn as sns
sns.histplot(data=titanic["Age"],bins=10)
```

Out[4]:
```
<AxesSubplot:xlabel='Age', ylabel='Count'>
```



Группировки

Каков средний возраст пассажиров Титаника мужчинами и женщинами?

In [11]:
```
titanic[["Sex", "Age"]].groupby("Sex").mean()
```

Out[11]:

Sex	Age
female	27.915709
male	30.726645

Если не указывать столбцы, то mean-метод применяется к каждому столбцу, содержащему числовые данные:

In [12]:
```
titanic.groupby("Sex").mean()
```

Out[12]:

Sex	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
female	431.028662	0.742038	2.159236	27.915709	0.694268	0.649682	44.479818
male	454.147314	0.188908	2.389948	30.726645	0.429809	0.235702	25.523893

Группировки

Какова средняя цена билета для каждой комбинации пола и класса салона?

In [16]:
```
titanic.groupby(["Sex", "Pclass"])["Fare"].mean()
```

Out[16]:
```
Sex     Pclass
female  1         106.125798
        2          21.970121
        3          16.118810
male    1          67.226127
        2          19.741782
        3          12.661633
Name: Fare, dtype: float64
```

Расчет количества записей по категориям

Метод **value_counts()** подсчитывает количество записей для каждой категории в колонке.

Какое количество пассажиров в салоне каждого класса?

In [17]:
```
titanic["Pclass"].value_counts()
```

Out[17]:
```
3    491
1    216
2    184
Name: Pclass, dtype: int64
```

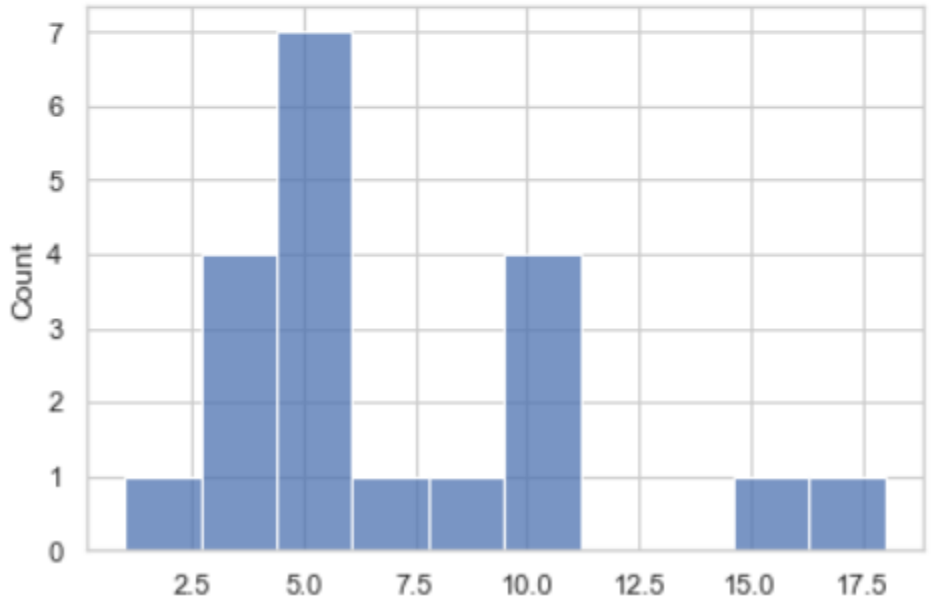
Гистограмма

In [35]:
```
s = pd.Series([1,3,5,11,10,3,6,5,6,6,7,8, 4,6,15,18,6,4,11,10])
```

Seaborn — это библиотека для создания статистических графиков на Python. Она основывается на matplotlib и тесно взаимодействует со структурами данных pandas.

In [47]:
```
import seaborn as sns
sns.histplot(data=s,bins=10)
```

Out[47]:
```
<AxesSubplot:ylabel='Count'>
```



```
In [36]: s.describe()

Out[36]: count    20.000000
         mean      7.250000
         std       4.191156
         min       1.000000
         25%       4.750000
         50%       6.000000
         75%      10.000000
         max      18.000000
         dtype: float64

In [34]: s.median()

Out[34]: 6.0

In [48]: sns.boxplot(x=s)

Out[48]: <AxesSubplot:>
```

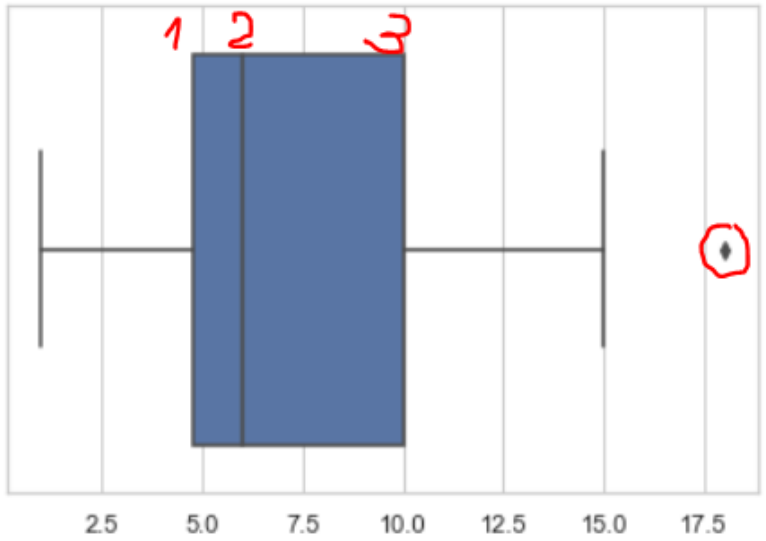


График box-plot

Центром ящика является медиана наших данных или второй квартиль, верхняя граница = 3-й квартиль, а нижняя граница = 1-й квартиль.

Почему некоторые точки на графике отображены отдельно?

Если мы посчитаем разность между 3-м и 1-м квартилем - это межквартильный размах (мера изменчивости).

Чем выше межквартильный размах, тем больше вариативность нашего признака.

Отложим мысленно 1,5 межквартильного размаха вверх и вниз от 1-го и 3-го квартилей. Те значения признака, которые последними принадлежат этому промежутку и будут границами усов.

Точки, которые превосходят полтора межквартильного размаха - наносятся на график отдельно.