Нижегородский государственный университет им. Н.И. Лобачевского

Институт информационных технологий, математики и механики

«Наглядный вероятностно-статистический анализ данных»

Лекция 1

Роль математической статистики и теории вероятностей при анализе данных, средства статистического анализа данных

Пройдакова Екатерина Вадимовна,
доцент кафедры ТВиАД ИИТММ

1. ПОНЯТИЕ ЭКСПЕРИМЕНТА ПРИ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОМ АНАЛИЗЕ

1.1. Примеры реальных экспериментов

В результате наблюдения за реальным процессом (экспериментом, опытом, испытанием) формируется набор данных определенной структуры. Рассмотрим примеры реальных экспериментов.

Пример 1. Галилей с Пизанской башни наблюдал за свободным падением тел, одинаковых по форме, размеру и различных по массе (чугунные, деревянные и т. д.). Он нашел адекватную математическую модель, которая связывает высоту падения H и время падения t формулой $H(t) = gt^2/2$, где g — ускорение свободного падения.

1.1. Примеры реальных экспериментов

Пример 2. Пусть в данном резервуаре поддерживается постоянный объем некоторого количества идеального газа. По определению идеальный газ удовлетворяет следующим двум условиям:

1) объем, приходящийся на молекулы газа, много меньше объема резервуара,
2) радиус взаимодействия двух молекул значительно меньше среднего расстояния между ними.

Тогда формула $U = a + bP$ дает адекватную математическую модель вычисления температуры U по заданному значению давления P, где a и b некоторые постоянные, зависящие от вида газа.

1.1. Примеры реальных экспериментов

Пример 3. Температура $\mathbf{U(x, y, z, t)}$ в точке $\mathbf{(x, y, z)}$ твердого тела в момент $\mathbf{t > 0}$ удовлетворяет однородному дифференциальному уравнению в частных производных параболического вида:

$$k^2\left\{\frac{\partial^2 U(x,y,z,t)}{\partial x^2}+\frac{\partial^2 U(x,y,z,t)}{\partial y^2}+\frac{\partial^2 U(x,y,z,t)}{\partial z^2}\right\}-\frac{\partial^2 U(x,y,z,t)}{\partial t^2}=0,$$

где $U(x, y, z, 0) = \varphi(x, y, z)$ — известная функция, определяющая температуру такой системы в момент $t = 0$, и k — коэффициент теплопроводности твердого тела (рис. 1.1).

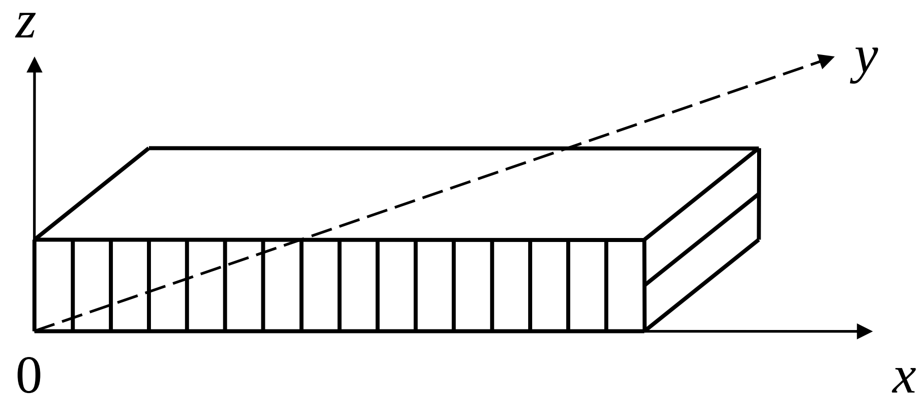


Рис. 1.1

1.2. Формализация понятия эксперимента

В теории вероятностей считается, что любой **реальный эксперимент** E **задается** следующими множествами:

- множеством $\mathfrak{I} = \{A, B, C, A_1, A_2, \ldots\}$ **всех его возможных исходов;**

- множеством $\Sigma = \{u_1, u_2, \ldots, u_s, \ldots\}$**, условий его проведения,** здесь $u_1, u_2, \ldots u_s$ - основные условия, а остальные условия (случайные факторы) нам точно неизвестны.

Множества Σ и $\mathfrak{I}$ могут содержать конечное, счетное (бесконечное), несчетное число элементов. Что определяется содержанием реального эксперимента и целями его исследования.

1.2. Формализация понятия эксперимента

***Пример 4*.** С помощью некоторого механизма один раз подбрасываются две симметричные игральные кости, изготовленные из идентичного материала, на некоторую поверхность стола. Грани костей занумерованы цифрами от 1 до 6. Определяются числа выпавших очков на каждой из костей.

В данном эксперименте множество условий проведения $\boldsymbol{\Sigma = \{u_1, u_2, \ldots, u_5, \ldots\}}$, выделяются пять основных условий ($s = 5$):

$\boldsymbol{u_1}$ — две симметричные игральные кости;

$\boldsymbol{u_2}$ — заданный механизм подбрасывания,

$\boldsymbol{u_3}$ — количество бросков равно единице;

$\boldsymbol{u_4}$ — поверхность стола;

$\boldsymbol{u_5}$ — условие фиксации выпавших очков на каждой из костей (например, достаточная освещенность поверхности стола).

1.2. Формализация понятия эксперимента

Очень часто эксперимент может быть представлен как совокупность других экспериментов E_t, $t \in T$ (примеры 1 и 3). Здесь T — некоторое упорядоченное множество, содержащее конечное, счетное (бесконечное), или несчетное число элементов t. Будем t интерпретировать как время, а T — как **промежуток времени**. В таком случае эксперимент E_t ***называется эволюционным*** *и* определяется комплексом условий Σ_t и множеством допустимых исходов $\mathfrak{I}_t$.

В случае, когда множество T состоит из **одного элемента** (пример 2) эксперимент будем ***называть статическим*** $E_t = E$.

1.3. Группы реальных экспериментов

Все реальные эксперименты делятся на три группы.

К **первой группе** будем относить **детерминированные эволюционные эксперименты**, когда **по любому** $t_0 \in T$ **и** Σ_{t0} **однозначно определяется исход эксперимента** E_t **как при** $t \geq t_0$**, так и при** $t < t_0$. Другими словами, весь будущий ход, включая и настоящее, и все прошлое детерминированного эволюционного эксперимента однозначно определяются условием Σ_{t0} эксперимента E_{t0} в настоящее время $t = t_0$.

Подобная ситуация описана в примерах 1, 2, при этом в примерах 2, промежуток T состоит из одной точки.

1.3. Группы реальных экспериментов

Ко **второй группе** относятся **полудетерминированные эволюционные эксперименты**, если **по любому $t_0 \in T$ и Σ_{t0} однозначно определяется исход эксперимента E_t только при $t \geq t_0$.** Иначе говоря, весь будущий ход полудетерминированного эволюционного эксперимента полностью определяется его исходом или результатом в настоящее время $t = t_0$, а его прошлое - не определяется.

Распределение тепла в нагретом стержне (пример 3) является полудетерминированным процессом. Такого рода процессы протекают во времени и изучаются в электродинамике Максвелла, в теории колебаний, в квантовой механике и т. д.

1.3. Группы реальных экспериментов

К **третьей группе** относятся **эволюционные эксперименты**, если **при его повторении** практически в одних и тех же условиях он может **давать различные, но вполне определенные результаты из множества** $\mathfrak{I}$. Такой эксперимент называется **случайным.**

В данном случае множества Σ_t, $t \in T$ и $\mathfrak{I}_t$, $t \in T$ определяют неоднозначно исход эксперимента E. Поэтому **результат случайного эксперимента предсказать невозможно**. Случайные эксперименты описаны в примерах 4-5.

1.4. Примеры случайных экспериментов

Пример 5. Страховая компания разделяет клиентов по трем классам риска: 1 класс – малый риск, 2 класс – средний, 3 класс – большой риск. Известен процентный состав клиентов компании по классам риска и шансы наступления страхового случая для каждой группы риска. Результат данного эксперимента - количество клиентов, получивших выплаты за период страхования по классам риска.

2. СТАТИСТИЧЕСКИ УСТОЙЧИВЫЕ ЭКСПЕРИМЕНТЫ

2.1. Понятие статистической устойчивости

Среди всех случайных экспериментов выделяется класс **статистически устойчивых экспериментов**, с помощью следующих двух ограничений.

1. Статистически устойчивый эксперимент E можно проводить или наблюдать любое конечное число раз при одних и тех же Σ и $\mathfrak{I}$.

2. Пусть $\mu(A, N)$ есть число наступлений результата $A \in \mathfrak{I}$ за N испытаний эксперимента E. Относительная частота $\mu(A, N)/N$ наблюдения исхода $A \in \mathfrak{I}$ за N испытаний эксперимента E колеблется около некоторого постоянного числа $\mathbf{P}(A)$ при неограниченном увеличении N. Это **свойство статистической устойчивости** должно выполняться для любого $A \in \mathfrak{I}$.

2.1. Понятие статистической устойчивости

Рассмотрим понятие статистической устойчивости на классическом примере подбрасывания симметричной монеты. Естествоиспытатель Бюффон бросал монету $N = \mathbf{4040}$ раз. Герб (исход A) появился **2048** раз. Тогда отношение $\boldsymbol{\mu(A, N)/N \approx 0{,}507}$. Пирсон бросал монету $N = \mathbf{24000}$ раз, и при этом герб выпал **12012** раз и отношение $\boldsymbol{\mu(A, N)/N = 0{,}5005}$.

При единичном проведении такого эксперимента невозможно предсказать его результат. Однако **при большом числе повторов**, обнаруживается **статистическая устойчивость в поведении относительной частоты** $\mu(A, N)/N$ выпадения герба, а именно ее малое **колебание около постоянного числа** $\mathbf{P}(A) = \mathbf{0{,}5}$.

2.1. Понятие статистической устойчивости

На рис 1.2 представлен график для наблюдаемой последовательности вида: *B*, *A*, *A*, *A*, *A*, *B*, *B*, *A*, *B*, *B*, *A*, *B*, *A*, *B*, *B*, *A*, *B*, *A*, *B*, *A*, *B*, *A*, *A*, *B*, *A*, *B*, *B*, *A*, *A*, *B*, *B*, *A*, *A*, *B*, *B*, *A*, *A*, *B*, *B*, *A*, ...

Здесь *A* — выпадение герба, *B* — выпадение решетки. Из графика видно, что соединяющая черные точки ломаная линия колеблется относительно прямой $\mu(A, N)/N = 1/2$ с размахом, уменьшающимся при увеличении *N*.

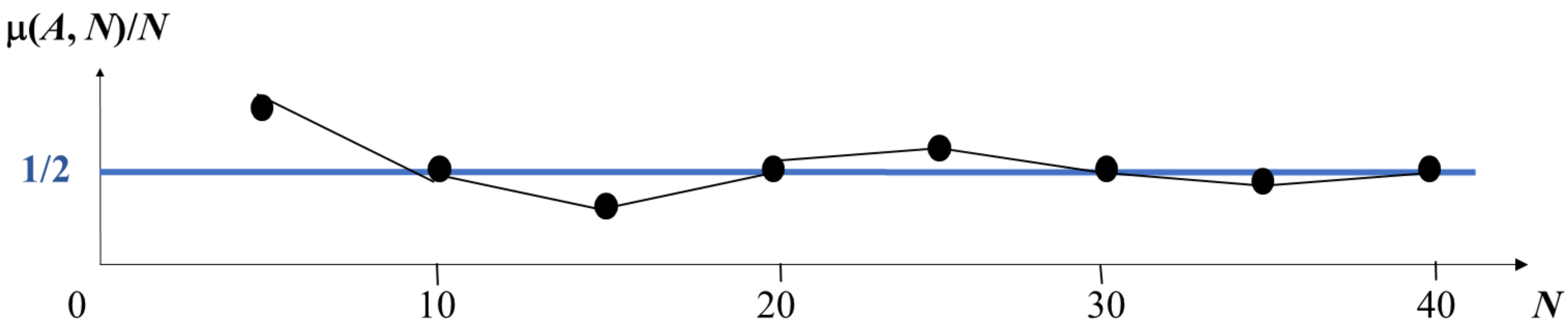


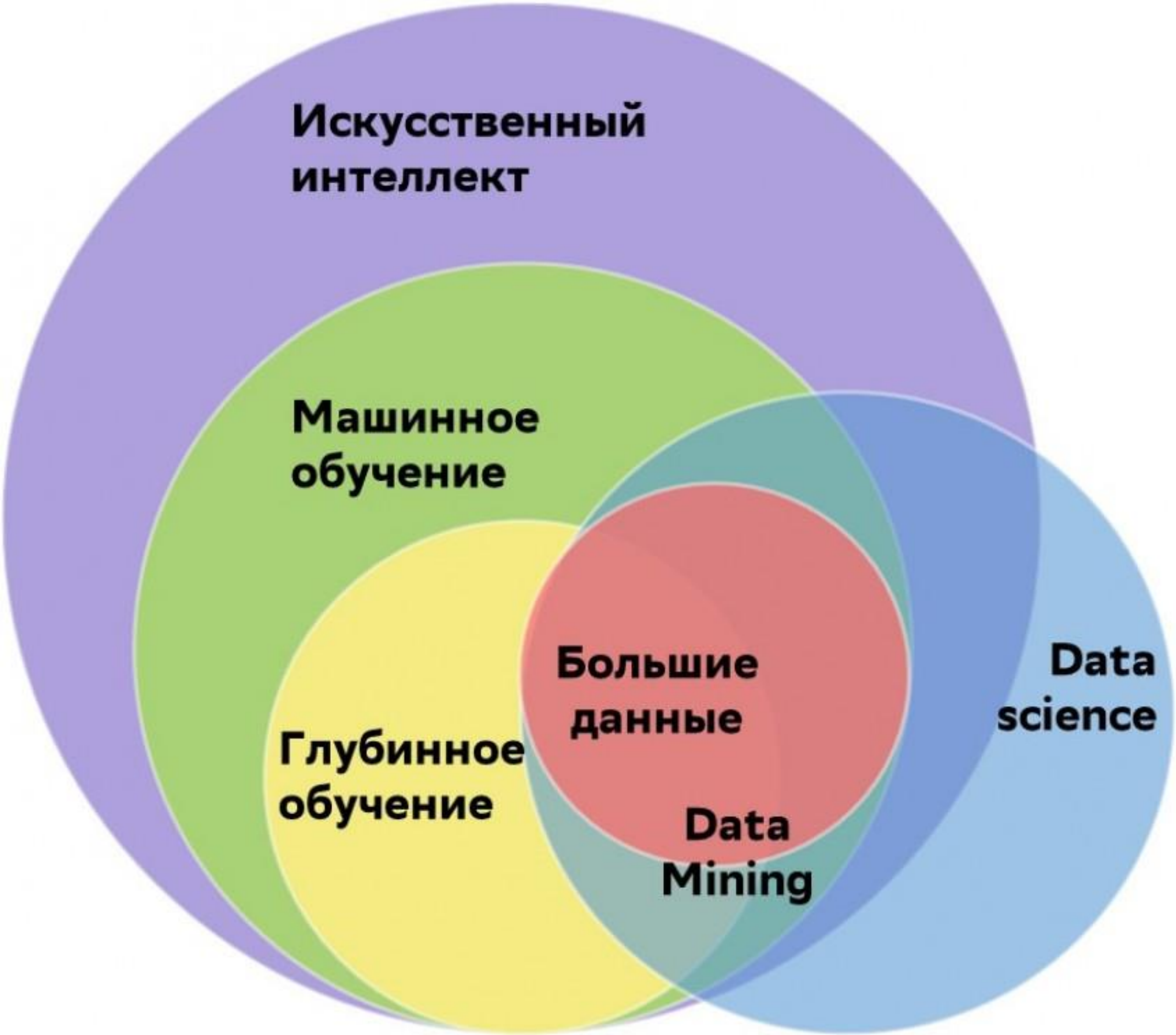
Рис.1.2

2.1. Понятие статистической устойчивости

Свойство **статистической устойчивости проявляется часто при анализе данных большого объема (Data science).** Например, характеристиках деятельность предприятия таких как средняя выработка одного работающего, средний процент брака, средний расход сырья, материалов и т.д.

При принятии решения используют именно средние показатели, **опираясь на свойство статистической устойчивости**, хотя в индивидуальном проявлении эти показатели могут колебаться в достаточно широких пределах.

2.2. Место Data science в сфере искусственного интеллекта



Искусственный интеллект включает в себя множество областей математики и информационных технологий, а также биологии, физики и других наук, важное место в этой сфере занимает и **Data science**

2.2.Связь Data science и ИИ

Машинное обучение

Класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Глубинное обучение

Иногда называют «глубокое обучение» (Deep learning). Подобласть машинного обучения, где в качестве решающих алгоритмов используются нейронные сети.

Data Science

Это концепция объединения теории вероятностей, математической статистики, машинного обучения и связанных с ними методов для анализа и прогнозирования реальных статистически устойчивых экспериментов (явлений).

Data Mining

Широкое понятие, означающее извлечение знаний из данных. Одно из важнейших назначений методов data mining состоит в наглядном представлении результатов вычислений (визуализация).

2.2.Связь Data science и ИИ

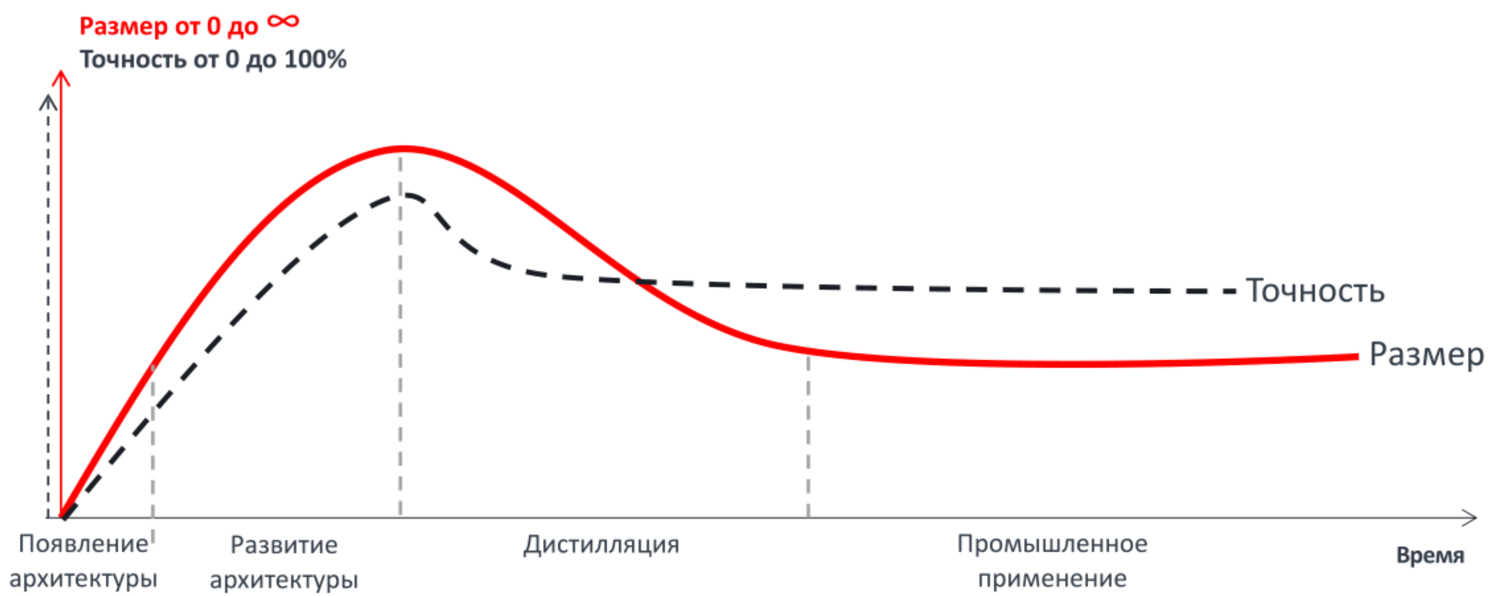
<u>Большие данные</u>

Это набор подходов и методов, разработанных для анализа данных огромных объемов. В широком смысле о «больших данных» говорят как о социально-экономическом феномене, связанном с появлением технологических возможностей анализировать огромные массивы данных и вытекающих из этого трансформационных последствий

«Зеленые» тренды

Экологическая повестка привела к вопросу «нужно ли тратить столько ресурсов на работу больших моделей»

Как следствие - невозможность внедрять большие модели в промышленное производство



3. ПРИЛОЖЕНИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

3.1. Связь теории вероятностей и математической статистики

Основным методом изучения **в теории вероятностей** является построение математических моделей (вероятностных моделей) и изучение их на абстрактном уровне, не прибегая к эксперименту.

В математической статистике исследования, наоборот, связаны с конкретными экспериментальными данными, и идут от практики к предполагаемой гипотезе и к ее проверке.

Приемы и способы научного анализа экспериментальных данных, относящихся к массовым явлениям, с целью определения обобщающих эти данные характеристик и выявления статистических закономерностей составляют **предмет математической статистики**.

3.1. Связь теории вероятностей и математической статистики

Математические модели случайных статистически устойчивых экспериментов в теории вероятностей основываются на понятии **вероятностной модели** $(\Omega, \mathscr{F}, \mathbf{P}(\bullet))$.

Каждая статистическая модель $\{(\Omega, \mathscr{F}, \mathbf{P}(\bullet)) \colon \mathbf{P}(\bullet) \in \wp\}$ описывает такие ситуации, когда в вероятностной модели изучаемого эксперимента имеется некоторая неопределенность в задании вероятностной функции $\mathbf{P}(\bullet)$.

Задача математической статистики состоит в том, чтобы уменьшить эту неопределенность, **анализируя результаты наблюдений (данные)**.

3.2. Приложения теории вероятностей и математической статистики

Одной из сфер **приложения** теории вероятностей и математической статистики является **экономика**.

При исследовании экономических явлений применяется **кластерный, регрессионный, корреляционный анализ ит.д**.

Например, **с помощью методов кластерного анализа** осуществляется ***сегментация*** рынков конкурентов и потребителей; ***разбиение*** персонала ***на группы; выявляются схожие*** производственные ***процессы***; ***проводится классификация*** пациентов, препаратов, методов лечения; формируется ***ранжирование*** предприятий по отраслям и т.д.

3.2. Приложения теории вероятностей и математической статистики

Методы корреляционно-регрессионного анализа позволяют решаются **задачи прогнозирования**, как на уровне отдельного предприятия, так и в масштабах субъекта (например РФ) в целом.

Например, ***предсказание развития процессов*** в реальном секторе экономики (изменение индексов промышленного производства, объемов продукции сельского хозяйства, оборота розничной торговли, потребительских цен и т.д);

предсказание динамики в финансово-банковской сфере за определенный период;

прогноз изменения ситуации в социальной сфере (в том числе прогнозирование демографической ситуации).

3.3. Инструменты для анализа данных

Для анализа данных разработана масса инструментов и средств, которые в разной степени эффективно реализуют **общие алгоритмы и методы теории вероятностей и математической статистики**.

Например, готовые аналитические пакеты **SPSS, STATISTICA** и даже стандартное офисное приложение **MS Excel**.

Различные языки программирования, используемые для анализа данных, например **R, Python.**

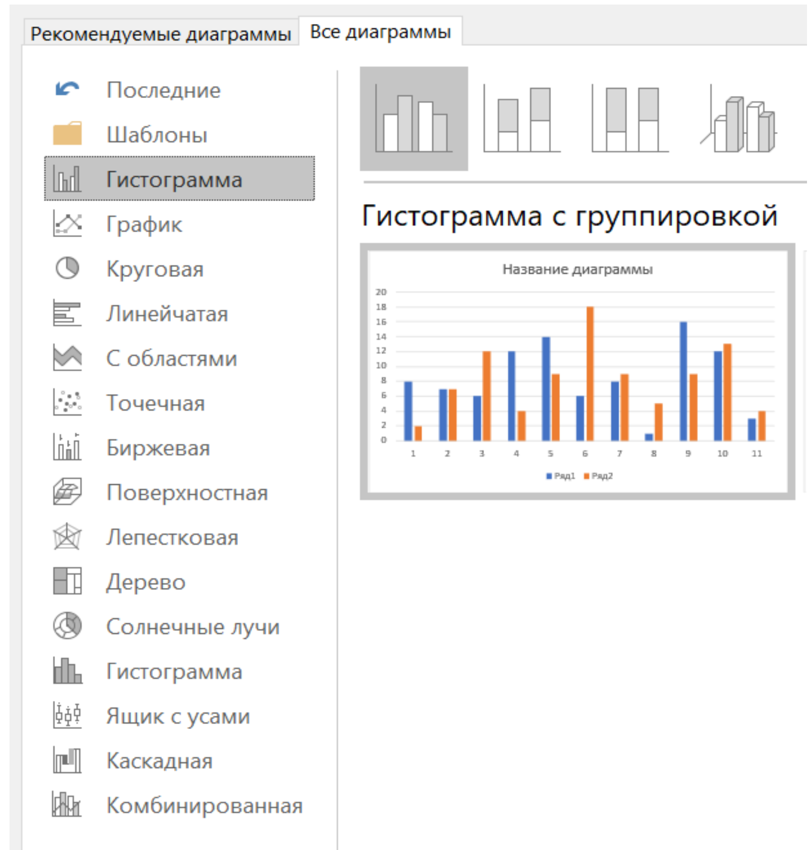
*В данном курсе при рассмотрении примеров анализа данных будет в основном использоваться **Python** и иногда **MS Excel***

3.3. Инструменты для анализа данных

Функции MS Excel для анализа данных представлены в библиотеках:

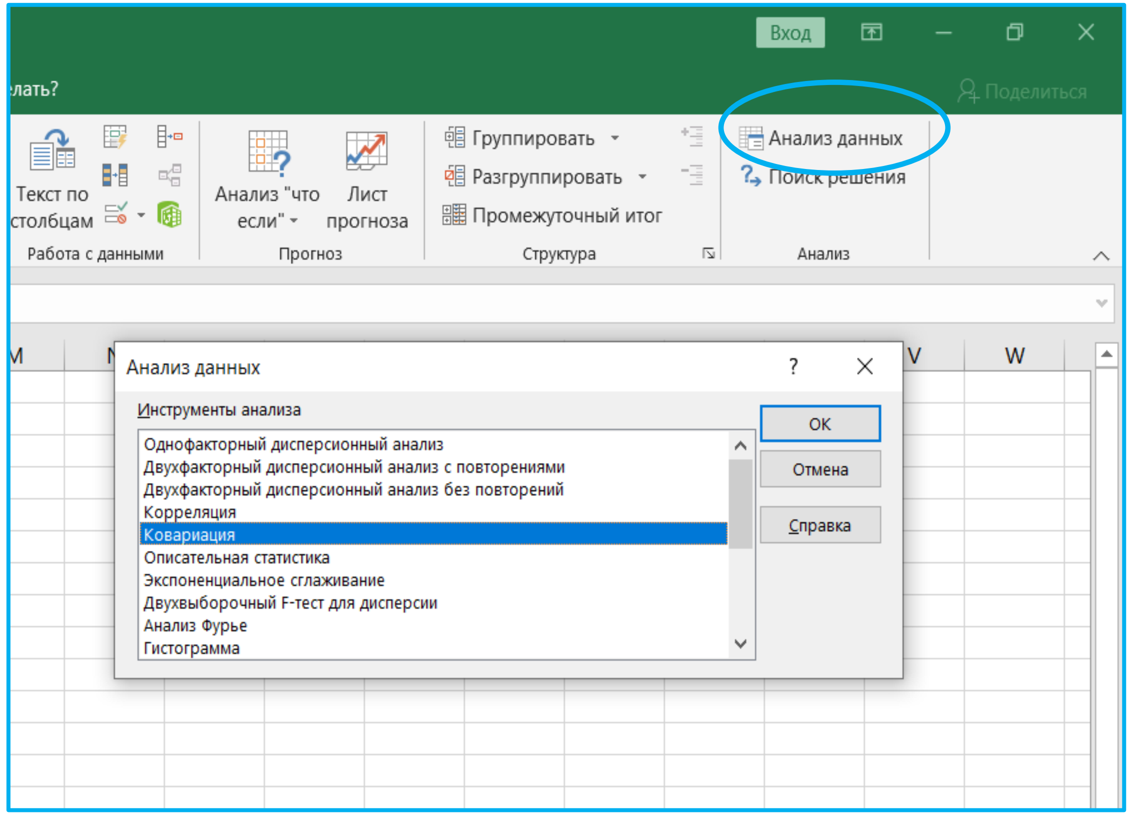
- Статистические,
- Математические,
- Логические.

Графических возможности включают построение **пятнадцати типов** различных двух- и трехмерных диаграмм, а также шаблоны для их настройки



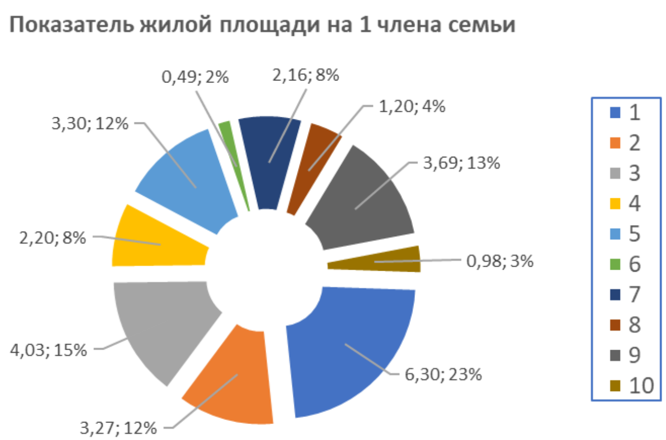
3.3. Инструменты для анализа данных

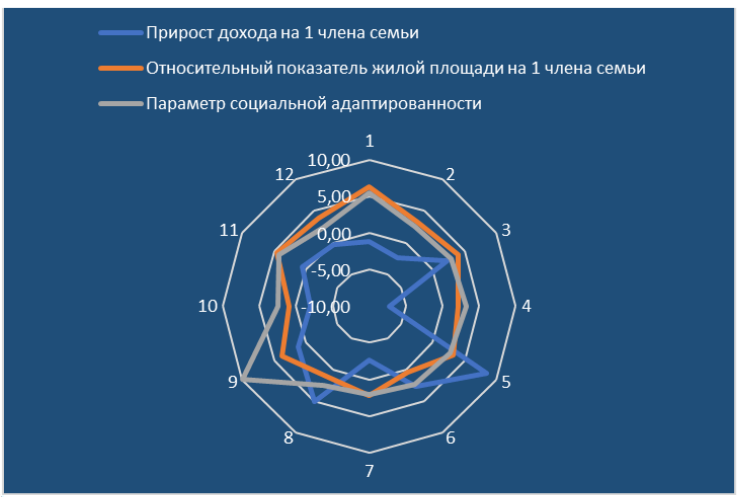
Основные методы статистического анализа в **MS Excel** представлены в надстройке **«Анализ данных»**

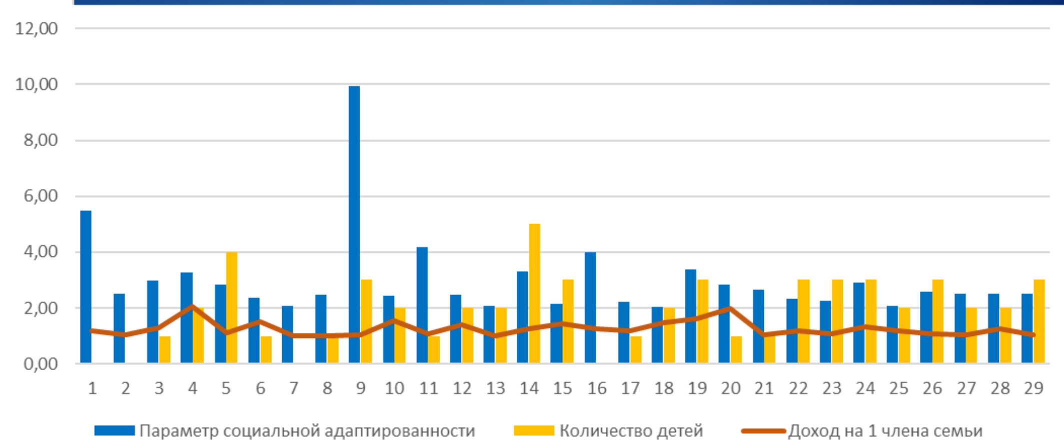


3.3. Инструменты для анализа данных

Графические возможности **MS Excel** для анализа данных







3.3. Инструменты для анализа данных Python

Основные библиотеки **Python,** используемые для анализа данных**:**

- Matplotlib,
- Seaborn,
- NumPy,
- SciPy,
- Pandas,
- Scikit-learn.

3. Python: Библиотеки для статистического анализа

1. **Библиотеки Matplotlib** предоставляет следующие возможности:

   - Построение линейных графиков, различного вида диаграмм (столбчатых, точечных, круговых, спектральных, коробчатых).

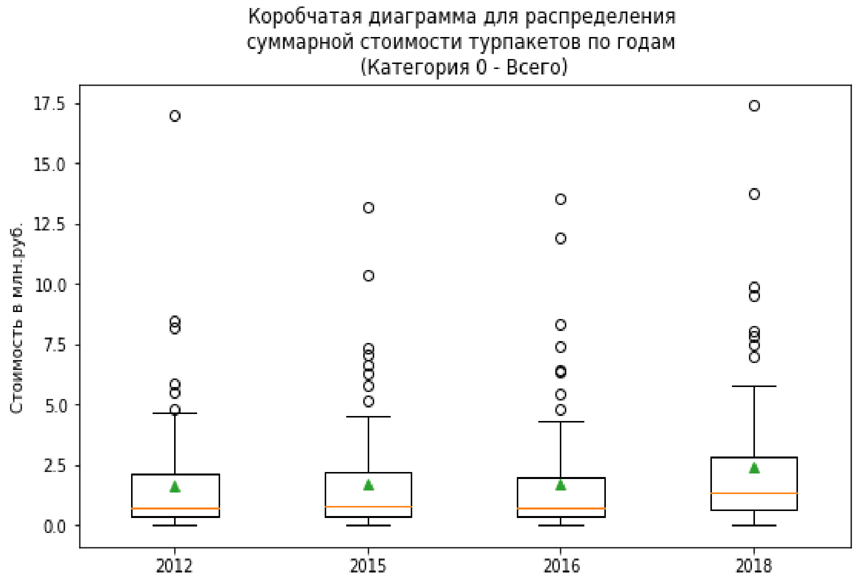
   - Большое количество поддерживаемых форматов изображений: PNG, JPEG, PDF и др.

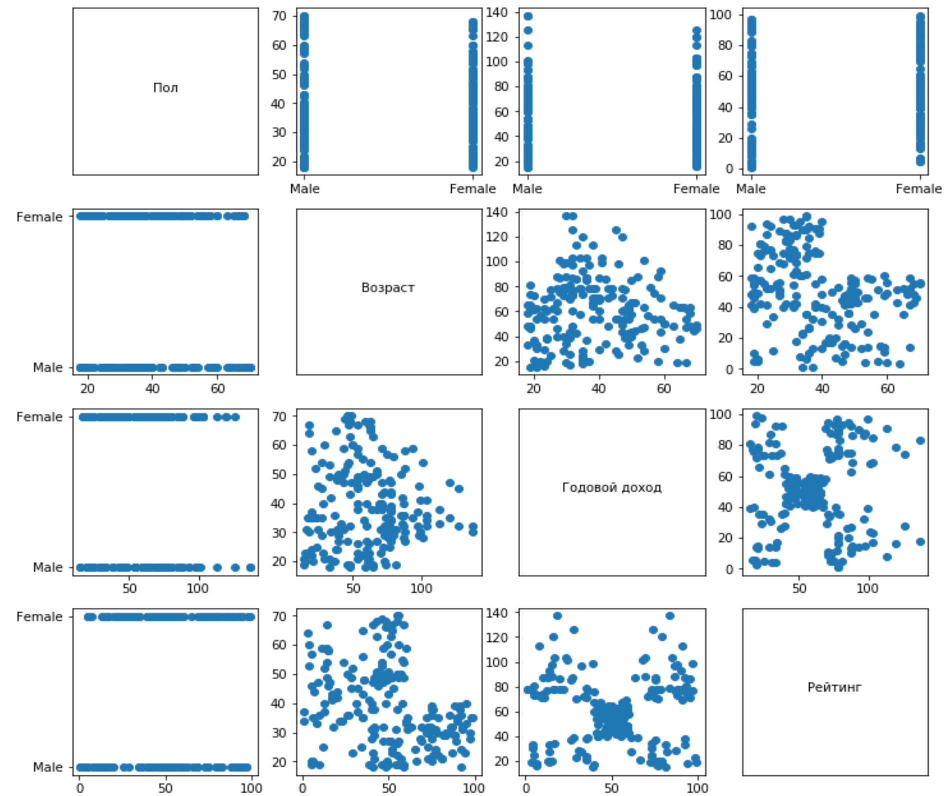
2. **Seaborn** — это **библиотека** для создания статистических графиков, основывается на Matplotlib и тесно взаимодействует со структурами данных Pandas.

   - Графический анализ данных.

3. Python: Библиотеки для статистического анализа

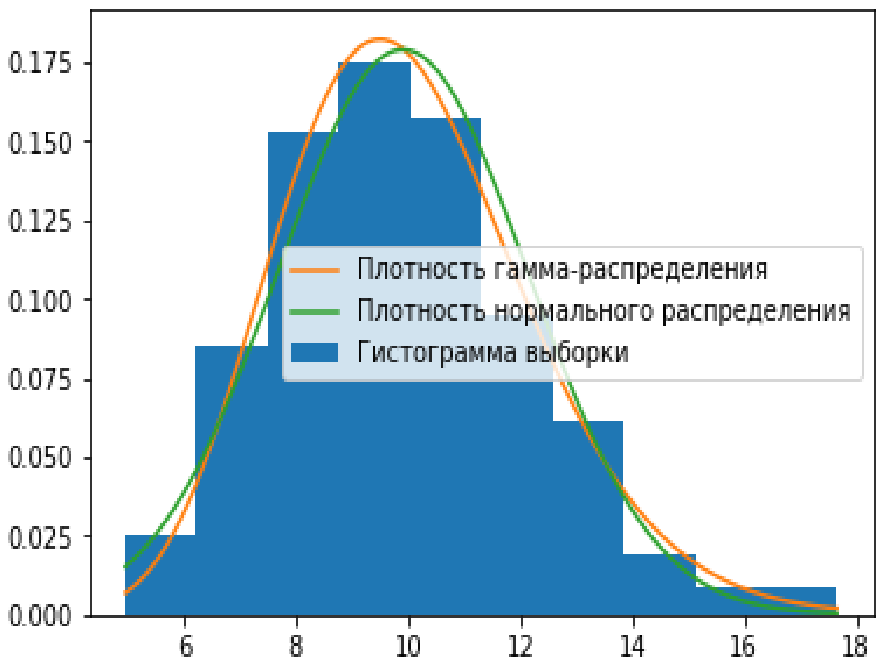
Примеры построенных диаграмм и графиков:

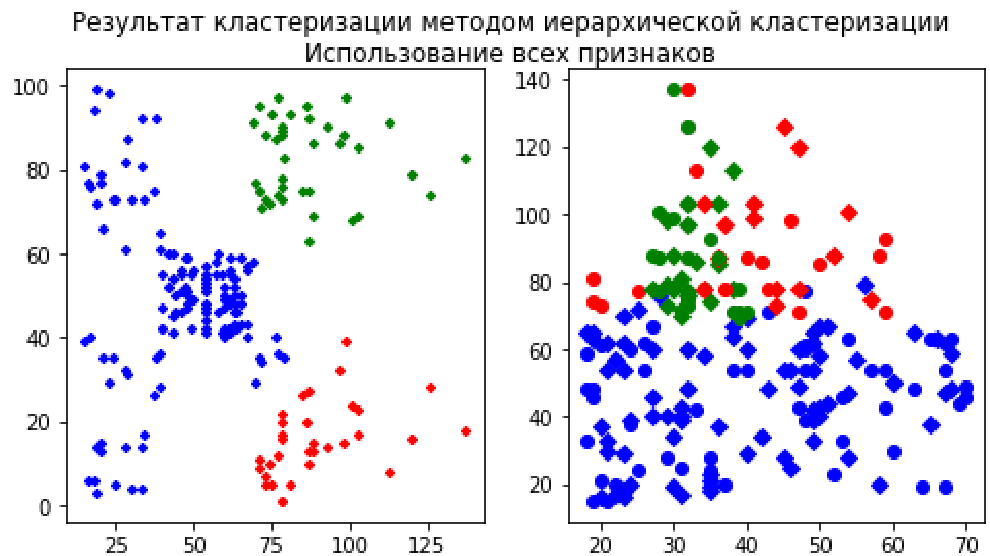




3. Python: Библиотеки для статистического анализа

Примеры построенных диаграмм и графиков:





3. Python: Библиотеки для статистического анализа

3. Библиотека NumPy

- Многомерные массивы и матрицы, функции работы с ними.

4. Библиотека SciPy

- Методы линейной алгебры: решение систем линейных уравнений, поиск собственных векторов и значений, разложений матриц.
- Методы теории вероятностей и математическая статистика: случайные величины, их распределения вероятностей, статистические числовые характеристики, проверка гипотез, подсчет статистик и др.
- Интегральное исчисление, преобразование Фурье

3. Python: Библиотеки для статистического анализа

4. Библиотека Pandas:

- Сбор, «очистка» и загрузка данных.
- Специальные структуры данных, переформатирование данных, сводные таблицы.
- Анализ данных: группировка, агрегирование, фильтрация.

5. Библиотека Scikit-learn:

- Алгоритмы анализа данных и машинного обучения: регрессия, классификация, понижение размерности, детектирование аномалий, выделение признаков.
- Большое количество методов кластеризации: метод K-средних, K ближайших соседей, нейронные сети, деревья решений и др.

4. ЗАКЛЮЧЕНИЕ

4. Заключение

- Теория вероятностей и математическая статистика две науки, методы которых лежат в основе практически любого анализа данных.
- Явления, рассматриваемые в обеих дисциплинах очень сложны. Только лишь в массовой совокупности наблюдений проявляются их общие закономерности. Выявление таких закономерностей невозможна без работы с большими объемами данных.
- <u>**Data Science**</u> – одно из важнейших приложений теории вероятностей и математической статистики. Существует множество готовых программных средств и языков программирования, предназначенных для работы в сфере анализа данных
- В предлагаемом курсе при рассмотрении практических примеров анализа данных предпочтение отдается языку Python.

Литература

1. Федоткин М.А. Основы прикладной теории вероятностей и статистики. — М.: Высшая школа. 2006. - 168 с.

2. Грас. Дж. Data Science. Наука о данных с нуля: Пер. с англ. – СПб.: БХВ-Петербург, 2017. – 336 с.

3. Лагутин М. Б. Наглядная математическая статистика: учебное пособие. — 2-е изд., испр. — М. : БИНОМ. Лаборатория знаний, 2009. — 472 с.