**Билет 1**

**1. Обработка и подготовка данных для анализа. Анализ валидности и релевантности исходных данных.**

**Релевантность маркетинговой информации** — это ее соответствие целям, задачам, параметрам и ограничениям исследования. Она должна обладать следующими характеристиками:

* — достоверностью;
* — актуальностью;
* — целенаправленностью;
* — простотой;
* — однозначностью;
* — сопоставимостью;
* — наглядностью;
* — преемственностью.

**Первичная обработка данных для анализа**

Для проведения анализа полученных в ходе полевых исследований данных необходимо создать первичные формы сводных таблиц. Но до внесения в них собранной информации необходимо просмотреть все заполненные анкеты и провести предварительную отбраковку имеющих различные дефекты.

При осуществлении преобразования исследователь старается найти зависимость среди собранных данных и достичь наиболее высокого уровня обобщения.

Преобразование данных имеет следующие функции:

* — обобщение;
* — определение концепции;
* — коммуникация, т.е. перевод результатов статистического анализа на понятный для управленца язык;
* — экстраполяция, т.е. определение достоверности полученных результатов.

При первичной обработке данных для повышения качества последующего анализа важно осуществление процедуры статистического выравнивания данных, состоящей из взвешивания, переопределения переменных и преобразования шкал.

***Метод статистического выравнивания данных*** заключается в приписывании каждому респонденту весового коэффициента, отражающего относительную важность учета его высказывания по сравнению с высказываниями других респондентов.

***Переопределение переменных*** связано с созданием новых или модификацией существующих переменных в соответствии с целями исследования. Наиболее встречающимися типами преобразований являются:

— укрупнение шкалы

— обобщение сведений из нескольких источников информации

— преобразование ответов респондентов в значения нулей и единиц

***Преобразование шкал*** используется для обеспечения сопоставимости оценок разных параметров и подготовки данных, пригодных для анализа.

**2. Относительная частота и ее устойчивость. Статистическое определение вероятности.**

**Относительной частотой события** называют отношение числа испытаний, в которых событие появилось, к общему числу фактически произведенных испытаний. Таким образом, относительная частота события А определяется формулой

https://studme.org/htm/img/33/3614/22.png

где т — число появлений события; п — общее число испытаний.

Сопоставляя определения вероятности и относительной частоты, заключаем: вероятность вычисляют до опыта, а относительную частоту — после опыта.

Длительные наблюдения показывают, что если в одинаковых условиях производят опыты, в каждом из которых число испытаний достаточно велико, то относительная частота обнаруживает **свойство устойчивости**. Это свойство состоит в том, ***что в различных опытах относительная частота изменяется мало (тем меньше, чем больше произведено испытаний), колеблясь около некоторого постоянного числа****.* Оказалось, что это постоянное число есть вероятность появления события.

***Статистической вероятностью****события А называется относительная частота* (*частость*) *появления этого события в п произведенных испытаниях*, т.е. https://studme.org/htm/img/33/3613/7.png

где *Р(Л)****—*** статистическая вероятность события *A; w(A)****—*** относительная частота (частость) события *Ат****—*** число испытаний, в которых появилось событие *А;п* — общее число испытаний.

Статистическое определение вероятности, как и понятия и методы теории вероятностей в целом, применимы не к любым событиям с неопределенным исходом, которые в житейской практике считаются случайными, а только к тем из них, которые обладают определенными свойствами

1. Рассматриваемые события должны быть *исходами только тех испытаний, которые могут быть воспроизведены неограниченное число раз при одном и том же комплексе условий.*

2. События должны обладать так называемой *статистической устойчивостью*, или *устойчивостью относительных частот*.

3. *Число испытаний*, в результате которых появляется событие Л, *должно быть достаточно велико*.

**3. При государственном регулировании курс некоторой национальной валюты 𝑋 оказался равномерно распределен в коридоре [11.5; 38.6] (в евро). Найдите вероятность 𝑃 того, что текущий курс данной валюты превысит 14.6 д.е В поле ответа введите значение полученной вероятности 𝑃.**

В любые ячейки вводим значения а=11,5; b=38.6; x1=14.6; x2=38.6

Формула:

=(14,6 -38,6)/(38,6-11,5)= 0,885609

Ответ: 0,885609

**Билет 2**

1. Специфика предмета и ограничения эмпирических исследований в экономике и управлении.

Эмпирическая совокупность сведений дает первичную информацию о новых знаниях и многих свойствах исследуемых объектов, что является исходной базой для научного исследования. *Эмпирические методы* основаны, как правило, на использовании способов и приемов опытного исследования, позволяющих получить фактическую информацию о СУ. Особое место среди них занимают базовые методы, которые сравнительно часто используются в практической исследовательской деятельности.

К основным в данной группе можно отнести такие методы исследования, как:

* Наблюдение
* изучение первичной документации
* сравнение
* измерение
* нормативный
* эксперимент

Следует отметить, что методы наблюдения и изучения первичной документации по своей сущности можно было бы отнести к логико-интуитивным методам. Однако, учитывая тяготение их к эмпирическим методам (так как они основываются, как правило, на фактической информации, на опыте функционирования СУ), применительно к исследованию управляющих подсистем социально- экономических систем эти способы более целесообразно рассматривать именно в данной классификационной группе методов.

2. Что представляет собой корреляционный анализ?

Корреляционный анализ дает возможность *точной количественной оценки степени согласованности изменений* (варьирования) двух и более признаков. Степень согласованности изменений характеризует теснота связи — абсолютная величина коэффициента корреляции.

Наличие корреляции между двумя результатами, в сущности, означает, что при изменении одного результата другой также изменяется — таким образом, между результатами существует, выявляется связь. Если значение некоторой величины может изменяться, то такую величину называют *переменной.* Корреляция между двумя переменными может быть положительной или отрицательной. *Положительной корреляцией* называется такая связь между переменными, когда значения обеих переменных возрастают или убывают пропорционально: с уменьшением (увеличением) одной уменьшается (увеличивается) другая.

Важно отметить другое: *корреляция еще не означает наличия причинно- следственной связи.* Наличие корреляции говорит о том, что между двумя переменными существует связь, но не о том, что одна из переменных является причиной, а другая — следствием. Существование причинно-следственной связи устанавливается другими методами.

Конечно, статическая связь между признаками — это необходимое, но не достаточное условие причинно-следственной связи между ними. Утверждение о том, что явление *А* есть причина явления *В,* справедливо, если одновременно выполняются три условия:

* явления *А* и *В* статистически связаны;
* *А* происходит раньше *В;*
* отсутствует альтернативная интерпретация появления *В* помимо *А* (другими словами, отсутствует общая причина *С* совместной изменчивости *А* и *В).*

Таким образом, применение корреляционного метода позволяет обосновать наличие только *статистической связи —* одного из трех признаков причинно-следственной связи

3. Акции двух компаний 𝐴 и 𝐵 имеют цены 𝑋 и 𝑌, распределенные по нормальному закону с параметрами соответственно 𝑚𝐴=4, 𝑉𝑎𝑟(𝑋)=0.7 и 𝑚𝐵=13, 𝑉𝑎𝑟(𝑌)=0.55. При этом коэффициент корреляции между ценами 𝜌(𝑋,𝑌)=−0.27. Найдите математическое ожидание и дисперсию цены портфеля, состоящего из 2 акций компании 𝐴 и 8 акций компании B. Решить в Excel

1. Вычисляем мат ожидание в любые ячейки, вводим значения mx= 4; my=13

Формулы:

В ячейку вводим формулу: = 2\*4+ 8\*13 =112

1. Вычисляем дисперсию вводим значение 𝑉𝑎𝑟(𝑋)=0.7, 𝑉𝑎𝑟(𝑌)=0.55, 𝜌(𝑋,𝑌)=−0.27

Формулы:

В ячейку вводим формулу: =-0,27\*Корень(0,7\*0,55) = -0,167531

Дисперсия

В ячейку вводим формулу: 2^2\*0.7+8^2\*0.55+2\*2\*8\* -0.167531=32.639021

Ответ: 32.639021

**Билет 3**

1. Классическое определение вероятности.

Понятие вероятности вводится для того, чтобы выражать на языке чисел степень возможности наступления тех или иных событий.

**За единицу принимают вероятность достоверного события, а вероятность невозможного события считают равной нулю. Тогда вероятность***Р***любого события***А***удовлетворяет неравенству**

https://studme.org/htm/img/33/3514/157.png

События называются **равновозможными,**если ни одно из них не является более возможным, чем другое.

События *А*, *В*, С, ..., *К* называются **единственно возможными,**если в результате опыта (испытания) одно из них обязательно наступит. Например, подбрасывание монеты.

*Классическое определение вероятности* формулируется так: вероятностью события А называется отношение числа *k* элементарных исходов, благоприятствующих этому событию, к общему числу элементарных исходов испытания *п*, если они равновозможны, несовместны и единственно возможны. Вероятность *Р(А)* обозначают следующим образом: https://studme.org/htm/img/33/3514/158.png

При использовании формулы (6.3) нахождение вероятности сводится к нахождению числа всех возможных событий в данном опыте и числа событий, благоприятствующих событию А.

2. Плотность распределения вероятностей непрерывной случайной величины и ее свойства.

*Непрерывной случайной величиной***называется функция***%,***заданная на вероятностном пространстве, значения которой могут быть любыми на некотором конечном или бесконечном промежутке.**

Непрерывную случайную величину можно также задать, используя другую функцию, которую называют плотностью распределения или плотностью вероятности (иногда ее называют дифференциальной функцией).

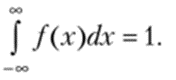
*Плотностью распределения* вероятностей непрерывной случайной величины *X* называют функцию /(*х*) — первую производную от функции распределения *F{x)*

https://studme.org/htm/img/33/3614/397.png

Из этого определения следует, что функция распределения является первообразной для плотности распределения.

### **Свойства плотности распределения.**

Плотность распределения обладает следующими свойствами.

* Свойство 1. *Плотность распределения — неотрицательная функция.*https://studme.org/htm/img/33/3614/417.png
* Свойство 2. *Несобственный интеграл от функции плотности распределения пределах от −∞ до +∞ равен 1:*  
  https://studme.org/htm/img/33/3614/418.png
* Свойство 3. *Любая неотрицательная функция*https://studme.org/htm/img/33/3614/417.png*, удовлетворяющая равенству* *является функцией плотность распределения некоторой непрерывной случайной величины.*
* Свойство 4. Величина вероятности на промежутке в случае непрерывной функции распределения не зависит от концов промежутка.

3. Время 𝑋 (в днях) реализации некоторого типового контракта распределено по показательному закону с параметром 𝜆 = 0.22. Найдите вероятность 𝑃 того, что случайно выбранный типовой контракт будет реализован за время от 2 до 8 дней включительно. В поле ответа введите значение полученной вероятности 𝑃. Решить в Excel

В любые ячейки вводим значения 𝜆 = 0.22; x1=2; x2=8

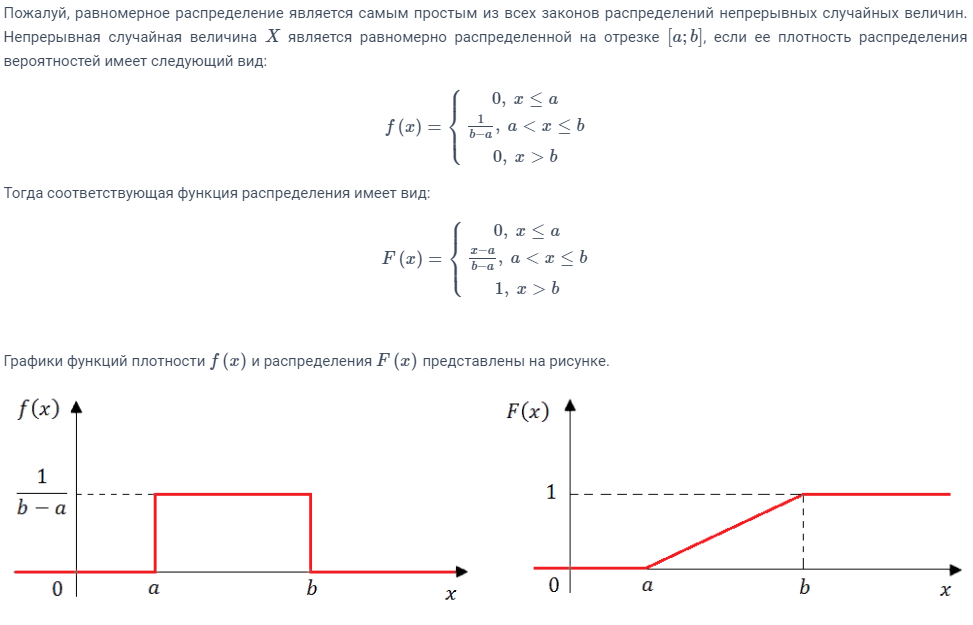
Формулы:

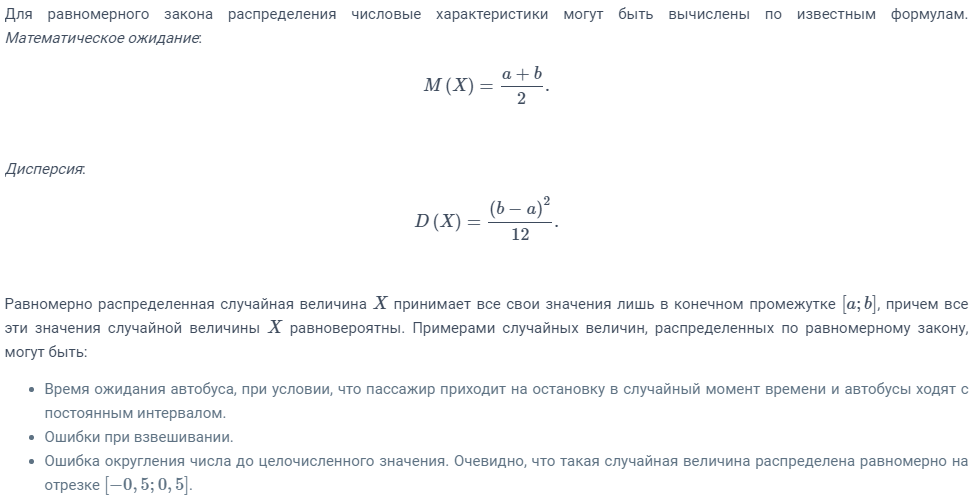
1. =ЭКСП.РАСПР (2;0,22;1)= 0,355964
2. =ЭКСП.РАСПР(8;0,22;1)=0,827955
3. =0,827955-0,355964=0,471992

Ответ: 0,471992

**Билет 4**

1. Закон равномерного распределения.





2. Принципы работы с большими данными.

**Большие данные (**[**англ.**](https://ru.wikipedia.org/wiki/Английский_язык)**big data)** — серия подходов, инструментов и методов обработки структурированных и [неструктурированных данных](https://ru.wikipedia.org/w/index.php?title=Неструктурированные_данные&amp;action=edit&amp;redlink=1) огромных объёмов и значительного многообразия для получения воспринимаемых человеком результатов, эффективных в условиях непрерывного прироста, распределения по многочисленным узлам [вычислительной сети](https://ru.wikipedia.org/wiki/Компьютерная_сеть).

Исходя из определения **Big Data**, можно сформулировать основные принципы работы с такими данными:

1. Горизонтальная масштабируемость. Поскольку данных может быть сколь угодно много – любая система, которая подразумевает обработку больших данных, должна быть расширяемой.

2. Отказоустойчивость. Принцип горизонтальной масштабируемости подразумевает, что машин в кластере может быть много.

3. Локальность данных. В больших распределённых системах данные распределены по большому количеству машин.

3. Банк совершил 3000 транзакций по кредитным картам. Вероятность P того, что транзакция будет ошибочной, равна 0.001. Найдите вероятность того, что банк совершит не более 3 ошибочных транзакций. Решить в Excel  
В любые ячейки вводим значения n = 3000; p=0.001; k1=0; k2=3

Формулы:

=БИНОМ.РАСП(3;3000;0,001;1) = 0,647232

Ответ: 0,647232

**Билет 5**

1. Вероятность появления хотя бы одного события.

*Вероятность появления хотя бы одного из событий А , А*2,..., *Ап, независимых в совокупности, равна разности между единицей и произведением вероятностей противоположных событий*

https://studme.org/htm/img/33/3614/103.png

Доказательство. Обозначим через *А* событие, состоящее в появлении хотя бы одного из событий *Av Ау* ..., *Ап.* События *А* и АхА2 ... Ли (ни одно из событий не наступило) противоположны, следовательно, сумма их вероятностей равна единице:

https://studme.org/htm/img/33/3614/104.png

Отсюда, пользуясь теоремой умножения, получим https://studme.org/htm/img/33/3614/105.png или

https://studme.org/htm/img/33/3614/106.png

Частный случай. *Если события Ар А.„* ..., *Ап имеют одинаковую вероятность, равную р, то вероятность появления хотя бы одного из этих событий*

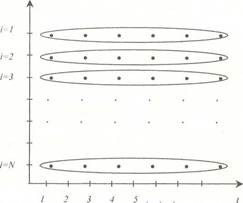
https://studme.org/htm/img/33/3614/107.png

2. Методы формирования панельных данных, сбалансированная и несбалансированная панель данных.

Панельные данные состоят из повторных наблюдений одних и тех же выборочных единиц, которые осуществляются в последовательные периоды времени. В качестве объектов наблюдения могут выступать индивиды, домашние хозяйства, фирмы, страны и т.д.

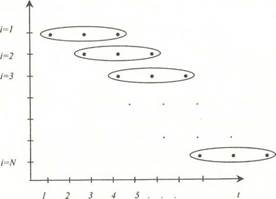
Панельные данные совмещают в себе как пространственные данные, так и временные ряды и сочетают достоинства каждого их этих видов данных. Это позволяет строить более адекватные и содержательные модели для изучения истинной причинно-следственной связи между различными переменными.

Панель бывает сбалансированная и несбалансированная. Если данные присутствуют по всем объектам за все периоды времени, то панель называется сбалансированной (рисунок 1).



**Рисунок 1 - Сбалансированная панель**

Достаточно часто из-за технических, организационных или иных причин в некоторые периоды времени не удается собрать сведения для всех объектов, включенных в выборку первоначально. Чтобы сохранить репрезентативность, отсутствующие объекты приходится заменять другими. В результате получим несбалансированную панель (рисунок 2).



**Рисунок 2 - Несбалансированная панель**

3. Завод отправил на базу 1000 доброкачественных изделий. Вероятность P того, что в пути изделие повредится, равна 0.0015. Найдите вероятность того, что на базу поступят 4 некачественных изделия. Решить в Excel  
В любые ячейки вводим значения n = 1000; p=0.0015; k1=0; k2=4

Формулы:

=БИНОМ.РАСП(4;1000;0,0015;1) = 0,981512

ОТВЕТ: 0,981512

**Билет 6**

1. Функция распределения вероятностей случайной величины и её свойства.

Для количественного описания распределения вероятностей удобно воспользоваться не вероятностью события *X* = х, а вероятностью события *X <* х, где *х -* некоторая текущая переменная. Вероятность этого события, очевидно, зависит от *х* и является некоторой функцией от *х.* Эта функция называется **функция распределения**случайной величины *X* и обозначается F(x):

https://studme.org/htm/img/33/3731/114.png

Функция распределения - самая универсальная характеристика СВ. Она существует как для дискретных, так и непрерывных СВ.

*Общие свойства интегральной функции распределения*:

* 1. Функция распределения *F(x)* неубывающая функция своего аргумента, то есть при *х2 >* х, F(x2) > /г(х1).
* 2. На минус бесконечности функция распределения равна нулю: F(-oo) = 0.
* 3. На плюс бесконечности функция распределения равна единице: **F(+oo) = l.**
* 4. P(x,<x<x2) = F(x2)-F(x1).</x<x

2. Какой уровень статистической значимости при тестировании гипотез является в экономической науке общепринятым? В чем заключается компромисс в его выборе?

Статистической гипотезой будем называть любое предположение о законе распределения случайной величины или о параметрах этого закона. Выделяют две виды гипотез: формулируемая гипотеза называется нулевой (основной) или гипотезой Н0. Также обязательно формулируют альтернативную (конкурирующую) гипотезу Н1. Обе гипотезы одновременно быть ложными или истинными быть не могут.

Уровень значимости должен стремиться к нулю, а мощность критерия к 1. Эти два требования вступают в противоречие. Сделать равными нулю обе ошибки при проверке гипотез невозможно. С практической точки зрения обычно наиболее критической является ошибка первого рода, а не второго.

В статистических пакетах обычно выдается не заданный нами уровень значимости, а накопленная вероятность наблюдения уровня статистического критерия (p-значение) при принятии нулевой гипотезы. Если p-величина меньше выбранного аналитиком критического уровня накопленной вероятности, то нулевая гипотеза отвергается.

Уровень р ≤ 0,05 часто используется в качестве критерия установления статистиче-ской значимости в экономической науке. Он означает, что с вероятностью 95 % можно утверждать: исследуемое событие произошло неслучайно, то есть связано с какой-то системой. Если p-значение наблюдаемого результата меньше заранее определенного порога, который мы называем уровнем значимости, наблюдаемый результат очень маловероятен, если нулевая гипотеза верна. Поэтому, когда мы отвергаем нулевую гипотезу, это равносильно подтверждению того, что эксперимент обнаружил связь между интересующими переменными.

При заданном объеме выборки вероятность совершения ошибки первого рода можно уменьшить, снижая уровень значимости a. Однако при этом увеличивается вероятность ошибки второго рода b, т.е. снижается мощность критерия. Выбор уровня значимости требует компромисса между заданной значимостью и мощностью. Увеличивая уровень значимости, мы увеличиваем шансы отвергнуть нулевую гипотезу, что является нашей конечной целью, а с другой стороны мы также увеличиваем и вероятность ошибки первого рода.  
3. Разработать имитационную модель обслуживания клиентов ООО «Зебра» Вероятность обнаружения дефекта при каждой проверке бракованного изделия равна 0.8. Найдите вероятность 𝑃 того, что после проверки из 400 бракованных изделий будет выявлено 310 дефектов. Решить в Excel

В любые ячейки вводим значения n = 400; p=0,8; k0=310

Формула:

=БИНОМ.РАСП(310;400;0,8;0) = 0,022354

**Билет 7**

1. Закон больших чисел. Неравенство Чебышева.

**Закон больших чисел** — основной закон в теории [вероятностей](https://smart-lab.ru/finansoviy-slovar/вероятность), который утверждает, что среднее значение случайных величин из заданного распределения близко к теоретическому среднему значению ([математическое ожидание](https://smart-lab.ru/finansoviy-slovar/математическое%20ожидание)) этого распределения.  
  
Названием «закон больших чисел» объединена группа теорем, устанавливающих устойчивость средних результатов большого количества случайных явлений и объясняющих причину этой устойчивости.

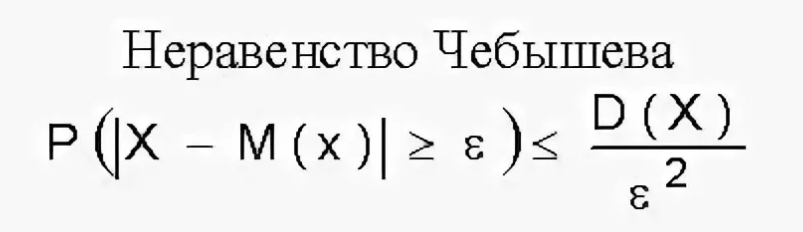
* Теорема Бернулли
* Теорема Пуассона
* Центральная предельная теорема
* Теорема Ляпунова
* Теорема Чебышева

Таким образом, под законом больших чисел в теории вероятностей понимается ряд теорем, в каждой из которых устанавливается факт приближения средних характеристик большого числа опытов к некоторым определенным постоянным.

В основе доказательства этих теорем лежит важное неравенство, установленное в 1845 г. П.Л. Чебышевым.

Неравенство Чебышева лежит в основе качественных и количественных утверждений закона больших чисел.

Оно определяет верхнюю границу вероятности того, что отклонение значения случайной величины от ее математического ожидания больше некоторого заданного числа. Замечательно, что неравенство Чебышева дает оценку вероятности события для случайной величины, распределение которой неизвестно, известны лишь ее математическое ожидание и дисперсия.



2. Числовые характеристики дискретных случайных величин и их свойства.

Числовые характеристики дискретной случайной величины.

Основными характеристиками ДСВ являются **математическое ожидание, дисперсия и среднеквадратическое отклонение**.

**Математическим ожиданием** дискретной случайной величины называют сумму произведений всех ее возможных значений на их вероятности:



Свойства математического ожидания:

1. Математическое ожидание постоянной величины равно самой постоянной:

2. Постоянную можно выносить за знак математического ожидания

3. Математическое ожидание произведения взаимно независимых случайных величин равно произведению математических ожиданий сомножителей:

4. Математическое ожидание суммы случайных величин равно сумме математических ожиданий слагаемых:

(для разности аналогично)

**Дисперсией случайной величины** Х называют математическое ожидание квадрата отклонения случайной величины от ее математического ожидания

Дисперсию удобно вычислять по формуле:

Свойства дисперсии:

1. Дисперсия постоянной равна нулю: 

2. Постоянный множитель можно выносить за знак дисперсии, предварительно возведя его в квадрат:

3. Дисперсия суммы (разности) независимых случайных величин равна сумме дисперсий слагаемых:

4.  

**Средним квадратическим отклонением** случайной величины называют квадратный корень из дисперсии:



Свойства среднеквадратического отклонения

1. Среднее квадратическое отклонение имеет всегда положительную или равную нулю величину: 0 ≤ σ.

2. Среднее квадратическое отклонение для заданного множества неотрицательных чисел лежит между минимальным и максимальным значением линейного отклонения от среднего значения этого множества.

3. Кроме того квадратическое отклонение подчиняется неравенству о средних, то есть для любого множества чисел оно не меньше среднего линейного отклонения: δ ≤ σ

3. Вероятность выпуска бракованного изделия равна 0.24. Найдите вероятность P того, что среди 200 выпущенных изделий более 160 изделий окажется без брака. Решить в Excel

В любые ячейки вводим значения n = 200; p=1-0.24=0,76; k1=161; k2=200

Вычислим бином.распред k1=161-1=160(иначе результат не будет учитываться) и k2=200

=БИНОМ.РАСП(120-1;200;0,76;1) = 0,922669

=БИНОМ.РАСП(200;200;0,76;1) = 1

Вычисляем вероятность

=1-0,922669=0,077331

**Билет 8**

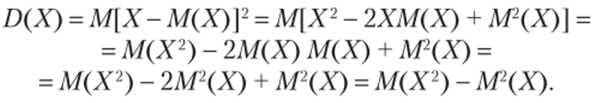
1. Формула для вычисления дисперсий.

Для вычисления дисперсии часто бывает удобно пользоваться следующей теоремой.

**Теорема.** *Дисперсия равна разности между математическим ожиданием квадрата случайной величины X и квадратом ее математического ожидания:*

https://studme.org/htm/img/33/3614/272.png

**Доказательство.** Математическое ожидание *М(Х*) есть постоянная величина, следовательно, 2*М(Х)* и *М2(Х)* есть также постоянные величины. Приняв это во внимание и пользуясь свойствами математического ожидания (постоянный множитель можно вынести за знак математического ожидания, математическое ожидание суммы равно сумме математических ожиданий слагаемых), упростим формулу, выражающую определение дисперсии:



Итак,

https://studme.org/htm/img/33/3614/274.png

Квадратная скобка введена в запись формулы для удобства ее запоминания.

2. Причинность, критерий «при прочих равных» в эмпирических исследованиях. Критерии статистической значимости. Ошибки первого и второго рода.

Принцип **«при прочих равных условиях»** предполагает, что с целью установления влияния выбранной независимой переменной принимается допущение о неизменности всех других переменных модели за исключением исследуемой переменной.

**Критерии значимости** — это критерии, с помощью которых проверяют гипотезы об абсолютных значениях параметров или о соотношениях между ними для генеральных совокупностей с известной (с точностью до параметров) функцией распределения вероятностей.

Можно выделить несколько основных групп задач, для решения которых применяются статистические критерии:

* выявление различий в уровне исследуемого признака;
* выявление различий в распределении признака;
* выявление степени согласованности изменений нескольких признаков;
* анализ изменений признака под влиянием контролируемых условий.

Согласно статистическим гипотез статистические критерии делятся на *параметрические* и *непараметрические.*

*Параметрические критерии* позволяют прямо оценить уровень основных параметров генеральных совокупностей, разницы средних и различия в дисперсиях.

*Непараметрические критерии* позволяют решить некоторые важные задачи, которые сопровождают исследования и анализ.

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Поскольку проверку производят статистическими методами, ее называют *статистической.* В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т.е. могут быть допущены ошибки двух родов.

***Ошибка первого рода*** состоит в том, что будет отвергнута правильная гипотеза.

***Ошибка второго рода****состоит в том, что будет принята неправильная гипотеза.*

3. Известно, что новый гаджет приобретает 25% покупателей магазина. Найдите вероятность P того, что из 600 пришедших покупателей число купивших гаджет будет заключено на отрезке [120, 200]. Решить в Excel

В любые ячейки вводим значения n = 600; p=0.25; k1=120; k2=200

Вычислим бином.распред k1=120-1=119(иначе результат не будет учитываться) и k2=200

=БИНОМ.РАСП(120-1;600;0,25;1) = 0,001642

=БИНОМ.РАСП(200;600;0,25;1) = 0,999998

Вычисляем вероятность

=0,999998-0,001642=0,998356

**Билет 9**

1. Управление данными в пакетах прикладных программ

**Краткий обзор прикладных программ**

Этот класс программных продуктов наиболее представителен, что обусловлено широким применением средств компьютерной техники во всех сферах человеческой деятельности.

***ППП общего назначения***, к которым относятся:

* текстовые процессоры (редакторы), предназначенные для подготовки текстовых документов.
* табличные процессоры (редакторы или электронные таблицы), предназначенные для вычислений хранящихся в таблице данных, создания деловой графики;
* настольные системы управления базами данных (СУБД). СУБД позволяют организовать запросы на поиск данных по какому-либо признаку, оформлять отчеты, выполнять расчеты;
* средства презентационной графики, позволяющие создавать и показывать изображения на экране монитора (например, диаграммы и графики), готовить и редактировать слайд-фильмы, мультфильмы, видеофильмы;
* интегральные пакеты, представляющие собой набор из нескольких программных продуктов, функционально дополняющих друг друга, поддерживающие единые информационные технологии.

***Проблемно-ориентированные ППП****,* к которым относятся пакеты автоматизированного бухгалтерского учета: финансовой деятельности предприятия; управления персоналом или кадрового учета; управления производством; а также банковские информационные системы и др.

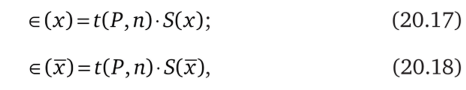
***Методе-ориентированные ППП.*** Программы этого типа ориентированы на математические, статистические и другие методы решения задач независимо от предметной области и функций информационных систем.

***ППП автоматизированного проектирования.*** Программы этого типа служат для поддержания работы конструкторов и технологов, связанной с разработкой чертежей, схем, диаграмм, графическим моделированием и конструированием, созданием библиотек стандартных элементов чертежей и их многократным использованием, созданием демонстрационных иллюстраций и мультфильмов.

2. Доверительные интервалы.

**Доверительный интервал** — это интервал, построенный с помощью случайной [выборки](http://www.machinelearning.ru/wiki/index.php?title=Выборка) из распределения с неизвестным параметром, такой, что он содержит данный параметр с заданной вероятностью.

***Доверительным интервалом*** называется интервал значений, в который попадает истинное значение измеряемой величины с заданной доверительной вероятностью. Для определения доверительного интервала случайной погрешности единичного *х* и среднего х результатов используется критерий Стьюдента t(P, п):



где t(P, п) — критерий Стьюдента (принимается из таблиц в зависимости от доверительной вероятности Р и числа наблюдений п).

3. Дискретная случайная величина 𝑋, описывающая в некоторой модели величину убытков, принимает целые значения 2, 5, 6, 9, 10 с равной вероятностью. Найдите среднюю величину убытков 𝐸(𝑋) и вероятность 𝑃(𝑋<𝐸(𝑋)). Решить в Excel

В любые ячейки вводим значения x1=2; x2=5; x3 =6; x4 =9; x5 =10

И так же значения p1=1/5; p2=1/5; p3 =1/5; p4 =1/5; p5 =1/5

=СУММПРОИЗВ(2:10;1/5:1/5)=6,4

Находим вероятность 𝑃(𝑋<𝐸(𝑋)) записываем <6.4 в любую ячейку и используем формула СУММЕСЛИ

=СУММЕСЛИ(2:10; <6.4; 1/5:1/5)=0,6

**Билет 10**

1. Этические аспекты проведения эмпирического исследования.

Методы и критерии сбора исходных данных. Источники данных. Открытые и платные базы данных.

*При проведении эмпирических исследований полезно учитывать ряд важных этических ограничений, которые являются ориентирами в рефлексии и регулировании профессионального поведения.*

*- личная ответственность за проводимую диагностику и полученные в исследовании выводы и результаты;*

*- конфиденциальность, нераспространение диагностической информации без прямого согласия самого испытуемого;*

*- научная обоснованность применяемых исследовательских методов и методик*

*- максимальная объективность в рамках интерпретации результатов, выводов;*

*- профессиональная компетентность (самоуважение и самоограничение);*

*- ненанесение ущерба;*

*- обеспечение прав людей, привлекаемых к экспериментальным действиям:*

К вопросу идентификации систем с точки зрения методов сбора исходной информации наметилось несколько подходов.

Во втором случае (***методы пассивной идентификации)*** исследователь наблюдает (фиксирует) интересующие его параметры, не вмешиваясь в процесс функционирования социально-экономического объекта.

В третьем подходе ***(активно-пассивный метод)*** наблюдение за объектом осуществляется пассивно, но для идентификации модели данные отбираются исходя из целей исследования, производственных критериев и используемого математического инструментария, т.е. активно.

Четвертый поход ***(экспертный метод)*** основан на эвристической классификации широкого спектра входных сигналов (в том числе предполагается вид и оцениваются параметры шума – ошибок наблюдений).

**Источник данных** – это выборка данных, систематизированных по определенным критериям. Критерии выборки задаются с помощью запроса к базе данных (на языке SQL). Таким образом, источники данных представляют собой набор атрибутов из одной или нескольких таблиц (типов объектов) Системы.

В настоящее время в открытом доступе есть большое количество баз данных, содержащих самые разнообразные сведения. Предполагается, что предоставление свободного доступа к отдельным данным может способствовать повышению качества государственного, регионального и муниципального управления. Принцип открытости получил отдельное название – «открытые данные»

Наиболее полную информацию о хозяйствующих субъектах предоставляют платные базы данных.  Платные базы данных могут быть государственными или частными.   
2. Что представляют собой ошибки первого и второго рода?

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки. Поскольку проверку производят статистическими методами, ее называют *статистической.* В итоге статистической проверки гипотезы в двух случаях может быть принято неправильное решение, т.е. могут быть допущены ошибки двух родов.

***Ошибка первого рода*** состоит в том, что будет отвергнута правильная гипотеза.

***Ошибка второго рода*** состоит в том, что будет принята неправильная гипотеза.

Подчеркнем, что последствия этих ошибок могут оказаться весьма различными.

Замечание 1. Правильное решение может быть принято также в двух случаях:

* 1) гипотеза принимается, причем и в действительности она правильная;
* 2) гипотеза отвергается, причем и в действительности она неверна.

Замечание 2. Вероятность совершить ошибку первого рода принято обозначать через а; ее называют *уровнем значимости.* Наиболее часто уровень значимости принимают равным 0,05 или 0,01.

3. Дискретная случайная величина 𝑋, описывающая прибыль фирмы в млн. рублей, принимает все целые значения в диапазоне: −6,−5,…,0,…,10,11. Найдите математическое ожидание

прибыли фирмы E(X), если известно, что возможные значения 𝑋 равновероятны. Определите, с какой вероятностью прибыль фирмы окажется отрицательной, т.е. фактически фирма будет терпеть убыток. Решить в Excel

В ячейки вводим минимальное значение: Xmin = -6; Xmax=11

Вычисляем вероятность:

=1(11-(-6)+1)=0,055556

Вычисляем математическое ожидание дискретной случайной Х:

=0,055556\*(-6+11)\*(11-(-6)+1)/2=2,5

Рассчитываем искомое значение:

=ABS(-6)/(11-(-6)+1)=0.333333

**Билет 11**

1. Задачи Data Mining.

**Классификация (****Classification) -** Наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных - классы; по этим признакам новый объект можно отнести к тому или иному классу.

**Кластеризация (Clustering)** **-** Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

**Ассоциация (Associations) -** В ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

**Последовательность (****Sequence), или последовательная****ассоциация (****sequential****association)** **-** Последовательность позволяет найти временные закономерности между транзакциями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, связанными во времени.

**Прогнозирование (****Forecasting) -** В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей.

**Определение отклонений или выбросов (****Deviation****Detection),****анализ отклонений или выбросов -** Цель решения данной задачи - обнаружение и анализ данных, наиболее отличающихся от общего множества данных, выявление так называемых нехарактерных шаблонов.

**Оценивание (****Estimation) -** Задача оценивания сводится к предсказанию непрерывных значений признака.

**Анализ связей (Link****Analysis)** - задача нахождения зависимостей в наборе данных.

**Визуализация (****Visualization,****Graph Mining) -** В результате визуализации создается графический образ анализируемых данных. Для решения задачи визуализации используются графические методы, показывающие наличие закономерностей в данных.

**Подведение итогов (****Summarization)** - задача, цель которой - описание конкретных групп объектов из анализируемого набора данных.

2. Регрессия: линейная регрессия и обобщенные линейные модели.

**Регре́ссия** — [односторонняя](https://ru.wikipedia.org/wiki/Односторонняя_функция) [стохастическая](https://ru.wikipedia.org/wiki/Стохастичность) зависимость, устанавливающая соответствие между случайными переменными, то есть [математическое выражение](https://ru.wikipedia.org/wiki/Математическое_выражение), отражающее связь между зависимой переменной у и независимыми переменными х при условии, что это выражение будет иметь [статистическую значимость](https://ru.wikipedia.org/wiki/Статистическая_значимость).

**Линейная регрессия** ([англ.](https://ru.wikipedia.org/wiki/Английский_язык) Linear regression) — используемая в [статистике](https://ru.wikipedia.org/wiki/Статистика) [регрессионная модель](https://ru.wikipedia.org/wiki/Регрессионная_модель) зависимости одной (объясняемой, зависимой) переменной от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) с [линейной функцией](https://ru.wikipedia.org/wiki/Линейная_функция) зависимости.

Модель линейной регрессии является часто используемой и наиболее изученной в [эконометрике](https://ru.wikipedia.org/wiki/Эконометрика). А именно изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели. Предельные (асимптотические) свойства оценок нелинейных моделей также выводятся исходя из аппроксимации последних линейными моделями.

**Обобщенные линейные модели**

Обобщенная линейная модель расширяет общую линейную модель, связывая зависимую переменную с факторами и ковариатами посредством задаваемой функции. Более того, модель допускает наличие у зависимой переменной распределения, отличающегося от нормального.

3. Клиент может обратиться за получением кредита в один из трёх банков (1, 2, 3). Вероятности обращения в каждый банк зависят от их удаленности и равны соответственно 0.3,0.3 и 0.4. Вероятности того, что данному клиенту будет отказано в выдаче кредита равны 0.3 – в банке "1", 0.3 – в банке "2" и 0.05 – в банке "3". Найдите вероятность P того, что при первом же обращении кредит клиентом получен. Указать, в каком из банков (1,2,3) с наибольшей вероятностью он был получен. Дискретная случайная величина 𝑋, описывающая прибыль фирмы в млн. рублей, принимает все целые значения в диапазоне: −6,−5,…,0,…,10,11. Найдите математическое ожидание прибыли фирмы E(X), если известно, что возможные значения 𝑋 равновероятны. Определите, с какой вероятностью прибыль фирмы окажется отрицательной, т.е. фактически фирма будет терпеть убыток. . Решить в Excel

В любые ячейки вводим значения H1=0.3; H2=0.3; H3=0.4

В ячейки вводим условия вероятности отказа от покупки (P(A/H1)=1-0.3=0,7; P(A/H2)=1-0.3=0,7;P(A/H3)=1-0.05=0,95

Вычисляем:

=СУММПРОИЗ(0,3:0,4;0,7:0,95)=0,8

Находим все условия вероятности:

=0,3\*0,7/0,8= 0,2625

=0,3\*0,7/0,8= 0,2625

=0,4\*0,95/0,8=0,475

Ответ:3 банк

**Билет 12**

1. Применение технологий Big Data, Machine learning и Deep learning для решения бизнес-задач.

В последнее время особый интерес в сфере анализа данных вызывают такие направления исследований, как «большие данные» (Big Data) и «глубокое обучение» (Deep Learning). Передовые разработки в сфере искусственного интеллекта поражают воображение. Монетизация больших данных, возможность автоматизированного [управления корпоративной репутации в интернете (SERM)](https://www.bigdataschool.ru/bigdata/big-data-machine-learning-serm-и-отзывы.html), [таргетированная реклам](https://www.bigdataschool.ru/bigdata/big-data-реклама-и-маркетинг.html" \t "_blank)а и [персонализированный маркетинг](https://www.bigdataschool.ru/bigdata/big-data-соцсети-маркетинг.html) -все эти приемы актуальны не только для гигантских корпораций, типа интернет-ритейлера Ozon.ru . Технологии Big Data и Machine Learning также применимы в малом и среднем бизнесе любой прикладной отрасли – там, где необходимо собирать и анализировать большие объемы разноформатных сведений. Информация из ваших корпоративных систем, электронной почты, файловых хранилищ, бумажных записей, архивов и социальных сетей – все это источники данных для моделей и алгоритмов машинного обучения, которые позволят вам привлечь новых клиентов и удержать существующих .

Сегодня эти технологии превращаются из экзотических понятий в прикладные инструменты для решения практических задач, однако их внедрение еще достаточно дорогое мероприятие (порядка нескольких тысяч долларов) для малого и среднего бизнеса. Тем не менее, растущий спрос на такие продукты провоцирует появление новых программных решений и сервисов.   
2. Применение OLAP при решении аналитических задач.

Концепция многомерного представления данных в виде OLAP-кубов ориентирована на решение следующих типовых задач:

* — многомерный анализ данных в режиме реального времени;
* — ведение отчетности;
* — интеллектуальный анализ данных.

Благодаря многомерной модели данных пользователи ХД имеют следующие возможности:

* — оперативный анализ больших объемов данных по любым категориям и показателям на любом уровне детализации;
* — динамическое конструирование аналитических отчетов;
* — мониторинг и прогнозирование ключевых показателей;
* — выявление скрытых закономерностей в данных и пр.

Системы, реализующие OLAP-технологию на основе ХД, называют ***OLAP-системами.***

Основное назначение OLAP-систем — выполнение произвольных информационных запросов к ХД и поддержка аналитической деятельности пользователей.

OLAP- системы обеспечивают:

* — многомерное представление данных, включая полную поддержку иерархий и множественных иерархий (это — ключевое требование OLAP);
* — возможность обращаться к любой нужной информации независимо от ее объема и места хранения в ХД;
* — предоставление пользователю результатов анализа за приемлемое время (обычно не более 5 с), пусть даже ценой менее детального анализа;
* — возможность осуществления любого логического и статистического анализа, сохранения в ХД его результатов в доступном для конечного пользователя виде;
* — многопользовательский доступ к данным с поддержкой соответствующих механизмов блокировок и средств авторизованного доступа.

Перечисленные возможности OLAP-систем ориентированы на проведение ***многомерного анализа данных*** ХД, под которым понимают рассмотрение данных одновременно с различных точек зрения, т. е. по нескольким измерениям гиперкуба.

3. Владелец магазина звукозаписей решил при маркетинговом исследовании классифицировать по возрасту потенциальных покупателей как старшеклассников, студентов и людей более старшего возраста и выяснил, что в контингенте его покупателей эти категории представлены в пропорциях 40%,30% и 30%. Кроме того, было обнаружено, что покупки совершают 10% пришедших в магазин старшеклассников, 50% студентов и 35% людей более старшего возраста.  
Какова вероятность 𝑃1 того, что случайно выбранный посетитель магазина не совершит покупку? Если известно, что случайно выбранный посетитель магазина не совершил покупку, какова вероятность 𝑃2 того, что он оказался старшеклассником? Решить в Excel

В любые ячейки вводим значения H1=0.4; H2=0.3; H3=0.3

В ячейки вводим условия вероятности отказа от покупки (P(A/H1)=1-0.1=0,9; P(A/H2)=1-0.5=0,5;P(A/H3)=1-0.35=0,65

Формулы:

В ячейку вводим следующую формулу:

Находим значение полной вероятности P1

=СУММПРОИЗВ(0,4:0,3;0,9:0,65)=0,705000

Находим искомую вероятность P2=P(H1/A)

=0,4\*0,9/0,705000=0,510638

Вычисляем оставшиеся вероятности: P=P(H2/A); P=P(H3/A)

P=P(H2/A)

=0,3\*0,5/0,705000=0,212766

P=P(H3/A)

=0,3\*0,65/0,705000=0,276596

Вычисляем Сумму

=СУММ(0,510638; 0,212766; 0,276596)=1,000000

**Билет 13**

1. Кластеризация: K-means, нейронные сети Кохонена.

### **K-means**

**Кластеризация** – это разбиение элементов некоторого множества на группы на основе их схожести. К наиболее простым и эффективным алгоритмам кластеризации относится k-means (k-средних). Он состоит из четырёх шагов:

1. Задаётся число кластеров *k*, которое должно быть сформировано из объектов исходной выборки.

2. Случайным образом выбирается *k* записей, которые будут служить начальными центрами кластеров. Начальные точки, из которых потом вырастает кластер, часто называют «семенами».

3. Для каждой записи исходной выборки определяется ближайший к ней центр кластера.

4. Производится вычисление *центроидов* - центров тяжести кластеров.

Остановка алгоритма производится, когда границы кластеров и расположение центроидов перестают изменяться, то есть на каждой итерации в каждом кластере остаётся один и тот же набор записей. Алгоритм k-means обычно находит набор стабильных кластеров за несколько десятков итераций.

### **Сети Кохонена**

Сети Кохонена представляют собой разновидность самоорганизующихся карт признаков, которые, в свою очередь, являются специальным типом нейронных сетей.

Основная цель сетей Кохонена - преобразование сложных многомерных данных в более простую структуру малой размерности. Таким образом, они хорошо подходят для кластерного анализа, когда требуется обнаружить скрытые закономерности в больших массивах данных.

Сеть Кохонена состоит из узлов, которые объединяются в кластеры. Наиболее близкие узлы соответствуют похожим объектам, а удалённые друг от друга - непохожим.

Алгоритм Кохонена включает следующие шаги:

* Инициализация.
* Возбуждение.
* Конкуренция.
* Объединение.
* Подстройка.
* Коррекция.

2. Информационный подход к моделированию.

При использовании в бизнесе традиционного аналитического подхода неизбежно возникнут проблемы из-за несоответствия между методами анализа и реальностью, которую они призваны отражать. Существуют трудности, связанные с формализацией бизнес-процессов.

Поэтому в последние годы получил распространение информационный подход к моделированию, ориентированный на использование данных. При информационном подходе реальный объект рассматривается как «черный ящик», имеющий ряд входов и выходов, между которыми моделируются некоторые связи. Иными словами, известна только структура модели (например, нейронная сеть, линейная регрессия), а сами параметры модели «подстраиваются» под данные, которые описывают поведение объекта.

Для корректировки параметров модели используется обратная связь — отклонение результата моделирования от действительности, а процесс настройки модели часто носит итеративный (то есть цикличный) характер.

Таким образом, при информационном подходе отправной точкой являются данные, характеризующие исследуемый объект, и модель «подстраивается» под действительность. В информационном моделировании важную роль играют консолидация данных, их очистка и обогащение.

Инструментальной поддержкой процесса построения моделей на основе информационного подхода выступают современные технологии анализа данных KDD и Data Mining, а средством построения прикладных решений в области анализа — аналитические платформы.

3. В ящике содержатся 9 деталей, изготовленных на первом заводе, 7 деталей – на втором заводе и 6 деталей – на третьем заводе. Вероятности изготовления брака на заводах с номерами 1, 2 и 3 соответственно равны 0.07, 0.08 и 0.07. Найдите вероятность 𝑃1 того, что извлеченная наудачу деталь окажется качественной. Чему равна вероятность 𝑃2 того, что ее изготовил второй завод. Решить в Excel

В любые ячейки вводим значения 1завод = 9; 2 завод = 7; 3 завод=6;

Рассчитываем сколько у нас всего деталей по формуле:

=Сумм(9;7;6)=22

Далее рассмотрим гипотезы о происхождении детали.

=9/$22=0,409091

=7/$22=0,318182

=6/$22=0,272727

Выводи общую сумму

=СУММ(0,409091; 0,318182; 0,272727)=1

В любы ячейки вводим P(A/H1)=1-0.07=0.93; P(A/H2)=1-0.08 = 0.92; P(A/H3)=1-0.097=0.3

Находим P1 вероятность

=СУММПРОИЗВ(0,409091: 0,272727; 0.93: 0.93)=0,926818

Находим P2 вероятность

P2=P(H2/A)

=0,318182\*0.92/0,926818=0.315841

P(H1/A)

=0,409091\*0.93/0,926818=0.410495

P(H3/A)

=0,272727\*0.93/0,926818=0.273664

**Билет 14**

1. Регрессия: линейная регрессия и обобщенные линейные модели.

**Регре́ссия** — [односторонняя](https://ru.wikipedia.org/wiki/Односторонняя_функция) [стохастическая](https://ru.wikipedia.org/wiki/Стохастичность) зависимость, устанавливающая соответствие между случайными переменными, то есть [математическое выражение](https://ru.wikipedia.org/wiki/Математическое_выражение), отражающее связь между зависимой переменной у и независимыми переменными х при условии, что это выражение будет иметь [статистическую значимость](https://ru.wikipedia.org/wiki/Статистическая_значимость).

**Линейная регрессия** ([англ.](https://ru.wikipedia.org/wiki/Английский_язык) Linear regression) — используемая в [статистике](https://ru.wikipedia.org/wiki/Статистика) [регрессионная модель](https://ru.wikipedia.org/wiki/Регрессионная_модель) зависимости одной (объясняемой, зависимой) переменной от другой или нескольких других переменных (факторов, регрессоров, независимых переменных) с [линейной функцией](https://ru.wikipedia.org/wiki/Линейная_функция) зависимости.

Модель линейной регрессии является часто используемой и наиболее изученной в [эконометрике](https://ru.wikipedia.org/wiki/Эконометрика). А именно изучены свойства оценок параметров, получаемых различными методами при предположениях о вероятностных характеристиках факторов, и случайных ошибок модели. Предельные (асимптотические) свойства оценок нелинейных моделей также выводятся исходя из аппроксимации последних линейными моделями.

**Обобщенные линейные модели**

Обобщенная линейная модель расширяет общую линейную модель, связывая зависимую переменную с факторами и ковариатами посредством задаваемой функции. Более того, модель допускает наличие у зависимой переменной распределения, отличающегося от нормального.

2. Охарактеризуйте понятия одностороннего и двустороннего критериев значимости. Приведите примеры их использования в эмпирических исследованиях.

Критерии значимости, служащие для проверки **двусторонних гипотез**, называются **двусторонними**, а для односторонних гипотез — **односторонними**. Выбор **односторонней или двусторонней гипотезы находится за пределами формальных статистических методов и полностью зависит от целей исследования**.

Пусть цель исследования в том, чтобы выявить различие параметров двух генеральных совокупностей, которые соответствуют различным ее естественным условиям (условия жизни, возраст испытуемых и т. п.). Зачастую неизвестно, в какой из совокупностей рассматриваемый параметр будет больше, а какой меньше. Например, если сравнивают средние оценки учащихся в контрольной и экспериментальной группах, то заранее неизвестно, в какой группе средняя оценка будет больше. В этом случае нулевая гипотеза состоит в том, что средние равны между собой, а цель исследования — доказать обратное, т.е. наличие выявить различие между средними. При этом допускается, что различие может быть любого знака. Такие гипотезы называются **двусторонними**.

Риск одностороннего критерия в том, что он может назвать значимой переменную, которая не является значимой на самом деле. **Односторонний критерий** – это шанс назвать вашу переменную значимой, когда двусторонний критерий не срабатывает.

Единственное условие, при котором вы можете его применять, это если гипотеза о значимости переменной и ее важности в уравнении регрессии была построена до того, как EViews разочаровал Вас. Это вопрос этики, поскольку тесты созданы для проверки гипотезы, придумывание гипотез по результатам теста – порочный круг.

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel  
  
**Билет 15**

1. Нейронные сети. Многослойный персептрон.

**Нейросети** — это последовательность нейронов, соединенных между собой синапсами. Нейрон — это вычислительная единица, которая получает информацию, производит над ней простые вычисления и передает ее дальше.

**Нейронная сеть** – одно из направлений искусственного интеллекта, цель которого смоделировать аналитические механизмы, осуществляемые человеческим мозгом. Задачи, которые решает типичная нейросеть – классификация, предсказание и распознавание. Нейросети способны самостоятельно обучаться и развиваться, строя свой опыт на совершенных ошибках.

Также нейронные сети способны не только анализировать входящую информацию, но и воспроизводить ее из своей памяти.

**Многослойный****персептрон (****MLP)** - нейронная сеть прямого распространения сигнала (без обратных связей), в которой входной сигнал преобразуется в выходной, проходя последовательно через несколько слоев.

Первый из таких слоев называют входным, последний - выходным. Эти слои содержат так называемые вырожденные нейроны и иногда в количестве слоев не учитываются. Кроме входного и выходного слоев, в многослойном персептроне есть один или несколько промежуточных слоев, которые называют скрытыми.

В этой модели персептрона должен быть хотя бы один скрытый слой. Присутствие нескольких таких слоев оправдано лишь в случае использования нелинейных функций активации.

2. Инновации и технологии RapidMiner.

**RapidMiner** — это мощная и многопользовательская платформа, она служит для создания, передачи и обслуживания наукоемких данных. Платформа RapidMiner предлагает больше функций, чем любое другое визуальное решение, плюс она открыта и расширяема для поддержки всех потребностей научных данных.

Унифицированная платформа RapidMiner ускоряет создание полных аналитических рабочих процессов — от подготовки данных до моделирования до развертывания бизнеса — в единой среде, значительно повышая эффективность и сокращая время, необходимое для проектов в области данных.

* Хороший GUI. По сути, каждый функциональный блок собран в кубик.
* Есть хорошие инструменты подготовки данных.
* Расширяемость.
* Дружит с Hadoop, причём как с чистым, так и с коммерческими реализациями.
* Архитектурно данные снаружи.
* Кроме IDE есть ещё сервер. Rapid Miner Studio создаёт процессы, а на сервере их можно публиковать.
* Сервер умеет сразу строить минимальные отчёты.
* Быстрое развитие.

Если сравнивать RapidMiner c другими программами, то у RM гораздо шире функциональные возможности по обработке, банально больше узлов.

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel  
  
**Билет 16**

1. Нейронные сети. Метод обратного распространения ошибки.

**Нейросети** — это последовательность нейронов, соединенных между собой синапсами. Нейрон — это вычислительная единица, которая получает информацию, производит над ней простые вычисления и передает ее дальше.

**Нейронная сеть** – одно из направлений искусственного интеллекта, цель которого смоделировать аналитические механизмы, осуществляемые человеческим мозгом. Задачи, которые решает типичная нейросеть – классификация, предсказание и распознавание. Нейросети способны самостоятельно обучаться и развиваться, строя свой опыт на совершенных ошибках.

Также нейронные сети способны не только анализировать входящую информацию, но и воспроизводить ее из своей памяти.

**Метод обратного распространения ошибки** — один из известных методов, используемых для глубокого обучения нейронных сетей прямого распространения (такие сети ещё называют многослойными персептронами).

**Метод обратного распространения ошибки (Back****propagation, backprop)** - алгоритм обучения многослойных персептронов, основанный на вычислении градиента функции ошибок.

**Преимущества и недостатки метода**

К плюсам можно отнести простоту в реализации и устойчивость к выбросам и аномалиям в данных, и это основные преимущества. Но есть и минусы:

* неопределенно долгий процесс обучения;
* вероятность «паралича сети»;
* алгоритм уязвим к попаданию в локальные минимумы функции ошибки.

2. Детализация и транспонирование, кросс-диаграммы.

Существует два типа вычисляемых показателей: на основе других показателей (вычисление) и на основе исходных данных (вычисление на основе детализации) .

Расчет вычисляемых показателей на основе исходных данных происходит одновременно с расчетом стандартных показателей. При этом в функции расчета доступны значения полей текущей записи исходного набора и значения измерений.

Может быть выполнен полный поворот или транспонирование таблицы. При выполнении этой операции все поля-строки переходят в область полей столбцов, а все поля-столбцы наоборот в поля-строки. В отличие от частичного поворота транспонирование не затрачивает значительных вычислительных ресурсов для пересчета таблицы.

**Кросс-диаграмма**

Кросс-диаграмма представляет собой график заданного типа, построенный на основе куба. Основное отличие кросс-диаграммы от обычной диаграммы в том, что она однозначно соответствует текущему состоянию куба и при любых ее изменениях изменяется соответственно.

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel  
  
**Билет 17**

1. Дайте определение понятиям: пространственные данные, временные ряды, панельные данные. В чем заключается смысл подобного подразделения анализируемых данных?

***Пространственные данные*** – значения переменных, относящихся к однотипным объектам в один и тот же фиксированный момент времени. Они позволяют сравнить значение признака, измеренное на разных объектах исследования, сравнить эти объекты по степени проявления в них того или иного свойства и разделить все объекты в соответствии с этим на группы и категории.

К пространственным данным можно отнести:

* данные по курсам покупки и продажи определенной валюты в разных обменных пунктах города на конкретный день;
* данные по курсам покупки и продажи валюты различных стран в определенном обменном пункте города на конкретный день:
* температуру воздуха в разных городах России в определенный день.

Под ***временным рядом* *(динамическим рядом*)** подразумевается последовательность значений (наблюдений) некоторого признака (случайной величины) в последовательные моменты времени.

При практическом анализе временных рядов на основании эмпирических данных необходимо сделать выводы о свойствах этого ряда, механизме случайного процесса, порождающего рассматриваемый ряд. При этом обычно ставятся следующие цели:

* Описание основных характеристик, особенностей временного ряда.
* Подбор статистической модели, описывающей временной ряд.
* Прогнозирование будущих значений на основе имеющихся наблюдений, относящихся к прошлому.

Панельные данные состоят из повторных наблюдений одних и тех же выборочных единиц, которые осуществляются в последовательные периоды времени. В качестве объектов наблюдения могут выступать индивиды, домашние хозяйства, фирмы, страны и т.д.

Панельные данные совмещают в себе как пространственные данные, так и временные ряды и сочетают достоинства каждого их этих видов данных. Выделяют следующие преимущества использования панельных данных.

* Панельные данные позволяют учитывать индивидуальную неоднородность.
* Панельные данные содержат большое число наблюдений
* Панельные данные предоставляют возможность изучать динамику изменений индивидуальных характеристик единиц совокупности.
* Панельные данные лучше способны идентифицировать и измерить эффекты, которые просто не определяемы только во временных рядах или только в пространственных данных.
* Панельные данные позволяют конструировать и тестировать более сложные поведенческие модели
* Панельные данные позволяют избежать смещения, связанного с агрегированием данных
* Панельные данные макроуровня имеют более длинные временные ряды

2. Охарактеризуйте понятия одностороннего и двустороннего критериев значимости.

Критерии значимости, служащие для проверки **двусторонних гипотез**, называются **двусторонними**, а для односторонних гипотез — **односторонними**. Выбор **односторонней или двусторонней гипотезы находится за пределами формальных статистических методов и полностью зависит от целей исследования**.

Пусть цель исследования в том, чтобы выявить различие параметров двух генеральных совокупностей, которые соответствуют различным ее естественным условиям (условия жизни, возраст испытуемых и т. п.). Зачастую неизвестно, в какой из совокупностей рассматриваемый параметр будет больше, а какой меньше. Например, если сравнивают средние оценки учащихся в контрольной и экспериментальной группах, то заранее неизвестно, в какой группе средняя оценка будет больше. В этом случае нулевая гипотеза состоит в том, что средние равны между собой, а цель исследования — доказать обратное, т.е. наличие выявить различие между средними. При этом допускается, что различие может быть любого знака. Такие гипотезы называются **двусторонними**.

Риск одностороннего критерия в том, что он может назвать значимой переменную, которая не является значимой на самом деле. **Односторонний критерий** – это шанс назвать вашу переменную значимой, когда двусторонний критерий не срабатывает.

Единственное условие, при котором вы можете его применять, это если гипотеза о значимости переменной и ее важности в уравнении регрессии была построена до того, как EViews разочаровал Вас. Это вопрос этики, поскольку тесты созданы для проверки гипотезы, придумывание гипотез по результатам теста – порочный круг.

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel  
  
**Билет 18**

1. Многомерное представление данных.

# **Многомерная модель данных.**

В многомерной модели данные представляются в виде гиперкубов, используя которые, можно получать различные срезы при аналитической обработке данных. Оси гиперкуба содержат параметры, а ячейки — зависящие от них агрегатные данные. Вдоль каждой оси данные могут быть организованы в виде иерархии, представляющей различные уровни их детализации. Благодаря такой модели данных пользователи могут формулировать сложные запросы, генерировать отчеты, получать подмножества данных.

Осями многомерной системы координат служат основные атрибуты анализируемого бизнес-процесса. На пересечениях осей измерений находятся данные, количественно характеризующие процесс, — меры.

Многомерная модель называется *полукубической*, если предметная область может быть отображена в виде нескольких гиперкубов с различной размерностью и разными измерениями в качестве граней. Если же модель может состоять из одного гиперкуба или нескольких гиперкубов с одинаковой размерностью и совпадающими измерениями, то она называется *гиперкубической.*

Достоинством такой модели является удобство и эффективность аналитической обработки данных, изменяющихся во времени. Однако она является громоздкой для простых задач оперативной обработки данных.

Типичным представителем баз данных, реализующих многомерную модель, является Oracle Express Server

2. Библиотека Anaconda, инструменты семейства Jupiter.

Conda — это менеджер пакетов с открытым кодом и система управления средой, которая работает на Windows, macOS и Linux.

Anaconda — это дистрибутивы Python и R. Он предоставляет все необходимое для решения задач по анализу и обработке данных (с применимостью к Python).

Anaconda — это набор бинарных систем, включающий в себя Scipy, Numpy, Pandas и их зависимости.

**Scipy** — это пакет статистического анализа.

**Numpy** — это пакет числовых вычислений.

**Pandas**— уровень абстракции данных для объединения и преобразования данных.

**Anaconda полезна тем, что объединяет все это в единую систему**.

Jupyter Notebook – это крайне удобный инструмент для создания красивых аналитических отчетов, так как он позволяет хранить вместе код, изображения, комментарии, формулы и графики

JupyterLab поддерживает отображение и редактирование множества форматов данных: изображений, CSV, JSON, Markdown, PDF, Vega, Vega-Lite [и так далее](_blank). Для быстрой навигации по документам в JupyterLab есть [настраиваемые горячие клавиши](_blank), а также возможность использования [стандартных сочетаний](_blank) из vim, emacs и Sublime Text.

JupyterLab может быть дополнен при помощи расширений, позволяющих настроить под себя различные части JupyterLab, включая темы, редакторы файлов и дополнительные компоненты.

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel

**Билет 19**

1. Относительная частота и ее устойчивость. Статистическое определение вероятности.

**Относительной частотой события** называют отношение числа испытаний, в которых событие появилось, к общему числу фактически произведенных испытаний. Таким образом, относительная частота события А определяется формулой

https://studme.org/htm/img/33/3614/22.png

где т — число появлений события; п — общее число испытаний.

Сопоставляя определения вероятности и относительной частоты, заключаем: вероятность вычисляют до опыта, а относительную частоту — после опыта.

Длительные наблюдения показали, что если в одинаковых условиях производят опыты, в каждом из которых число испытаний достаточно велико, то относительная частота обнаруживает **свойство устойчивости**. Это свойство состоит в том, ***что в различных опытах относительная частота изменяется мало (тем меньше, чем больше произведено испытаний), колеблясь около некоторого постоянного числа****.* Оказалось, что это постоянное число есть вероятность появления события.

***Статистической вероятностью****события А называется относительная частота* (*частость*) *появления этого события в п произведенных испытаниях*, т.е. https://studme.org/htm/img/33/3613/7.png

где *Р(Л)****—*** статистическая вероятность события *A; w(A)****—*** относительная частота (частость) события *Ат****—*** число испытаний, в которых появилось событие *А;п* — общее число испытаний.

Статистическое определение вероятности, как и понятия и методы теории вероятностей в целом, применимы не к любым событиям с неопределенным исходом, которые в житейской практике считаются случайными, а только к тем из них, которые обладают определенными свойствами

1. Рассматриваемые события должны быть *исходами только тех испытаний, которые могут быть воспроизведены неограниченное число раз при одном и том же комплексе условий.*

2. События должны обладать так называемой *статистической устойчивостью*, или *устойчивостью относительных частот*.

3. *Число испытаний*, в результате которых появляется событие Л, *должно быть достаточно велико*.

2. Обработка и подготовка данных для анализа. Анализ валидности и релевантности исходных данных.

**Релевантность маркетинговой информации** — это ее соответствие целям, задачам, параметрам и ограничениям исследования. Она должна обладать следующими характеристиками:

* — достоверностью;
* — актуальностью;
* — целенаправленностью;
* — простотой;
* — однозначностью;
* — сопоставимостью;
* — наглядностью;
* — преемственностью.

**Первичная обработка данных для анализа**

Для проведения анализа полученных в ходе полевых исследований данных необходимо создать первичные формы сводных таблиц. Но до внесения в них собранной информации необходимо просмотреть все заполненные анкеты и провести предварительную отбраковку имеющих различные дефекты.

При осуществлении преобразования исследователь старается найти зависимость среди собранных данных и достичь наиболее высокого уровня обобщения.

Преобразование данных имеет следующие функции:

* — обобщение;
* — определение концепции;
* — коммуникация, т.е. перевод результатов статистического анализа на понятный для управленца язык;
* — экстраполяция, т.е. определение достоверности полученных результатов.

При первичной обработке данных для повышения качества последующего анализа важно осуществление процедуры статистического выравнивания данных, состоящей из взвешивания, переопределения переменных и преобразования шкал.

***Метод статистического выравнивания данных*** заключается в приписывании каждому респонденту весового коэффициента, отражающего относительную важность учета его высказывания по сравнению с высказываниями других респондентов.

***Переопределение переменных*** связано с созданием новых или модификацией существующих переменных в соответствии с целями исследования. Наиболее встречающимися типами преобразований являются:

— укрупнение шкалы

— обобщение сведений из нескольких источников информации

— преобразование ответов респондентов в значения нулей и единиц

***Преобразование шкал*** используется для обеспечения сопоставимости оценок разных параметров и подготовки данных, пригодных для анализа.

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel  
  
**Билет 20**

1. Функция распределения вероятностей случайной величины и её свойства.

Для количественного описания распределения вероятностей удобно воспользоваться не вероятностью события *X* = х, а вероятностью события *X <* х, где *х -* некоторая текущая переменная. Вероятность этого события, очевидно, зависит от *х* и является некоторой функцией от *х.* Эта функция называется **функция распределения**случайной величины *X* и обозначается F(x):

https://studme.org/htm/img/33/3731/114.png

Функция распределения - самая универсальная характеристика СВ. Она существует как для дискретных, так и непрерывных СВ.

*Общие свойства интегральной функции распределения*:

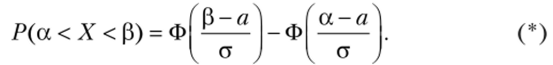
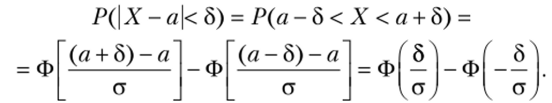
* 1. Функция распределения *F(x)* неубывающая функция своего аргумента, то есть при *х2 >* х, F(x2) > /г(х1).
* 2. На минус бесконечности функция распределения равна нулю: F(-oo) = 0.
* 3. На плюс бесконечности функция распределения равна единице: **F(+oo) = l.**
* 4. P(x,<x<x2) = F(x2)-F(x1).</x<x

2. Вычисление вероятности заданного отклонения.

Часто требуется вычислить вероятность того, что отклонение нормально распределенной случайной величины *X* по абсолютной величине меньше заданного положительного числа 8, т.е. требуется найти вероятность осуществления неравенства | *X - а*< 8.

Заменим это неравенство равносильным ему двойным неравенством

https://studme.org/htm/img/33/3614/485.png

Пользуясь формулой  получим  Приняв во внимание равенство

https://studme.org/htm/img/33/3614/487.png

(функция Лапласа — нечетная), окончательно имеем https://studme.org/htm/img/33/3614/488.png

Замечание. Очевидно, события, состоящие в осуществлении неравенств |Х-а|<8и|Х-а|>8, — противоположные. Поэтому если вероятность осуществления неравенства | *X - а*< 8 равна *р,* то вероятность неравенства | *X* - *а*> 8 равна 1 - *р.*

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel  
  
**Билет 21**

1. Способы анализа данных, технология Data Mining.

**Основные способы анализа данных**

Анализ собранной информации -- самый увлекательный этап исследования. На этом этапе проверяется, насколько верны были исходные предположения, даются ответы на заданные вопросы и выявляются новые проблемы. Методологический инструмент анализа данных -- гипотезы, которые были сформулированы в программе, и те, что возникают по мере их проверки и отвержения уже в процессе анализа собранных данных.

Методы анализа данных в качественном исследовании и метода анализа данных в количественном исследовании

* **Метод группировки данных**
* **Типологизация**
* **Классификация**

Способы анализа полученных данных в качественных исследованиях:

**Описание случаев**. В общем описание можно определить как перечисление всех интересующих исследователя характеристики объекта анализа (т.е. события).

**Анализ содержания**: взяв список кодированных категорий, каждый сегмент записанных данных относят к одной из категорий.

**Обобщения идей** из собранных свидетельств также является способом анализа данных в качественном исследовании.

**Data Mining (добыча данных, интеллектуальный анализ данных, глубинный анализ данных или просто майнинг данных)** — это процесс, используемый компаниями для превращения необработанных больших данных в полезную информацию.  
  
**Задачи Data Mining**  
Модели интеллектуального анализа данных применяются для нескольких типов задач:

* прогнозирование
* риск и вероятность
* рекомендации
* поиск последовательностей
* группирование

Интеллектуальный анализ данных в основном используется отраслями, обслуживающими потребителей, в том числе в сфере розничной торговли, в финансах и маркетинге.  
  
2. Фреймворки машинного обучения и моделирования нейронных сетей.

Популярность ИИ-технологий растёт, а значит, растёт и спрос на них. Это приводит к увеличению всего сообщества разработчиков и к появлению фреймворков ИИ, упрощающих обучение и работу.

Фреймворк — это набор инструментов и стандартных реализаций для обеспечения более быстрой разработки какого-либо программного продукта.

Если перейти во вселенную метафор, то фреймворк — это как мастерская, в которой вы можете сесть и, например, собрать автомобиль или стул.

Представляем следующую подборку:

**TensorFlow** :

Отличный фреймворк для создания нейронных сетей, которые будут работать в продакшене.

* Берёт на себя оптимизацию ресурсов для вычислений.
* Огромное комьюнити.
* За счёт популярности выше вероятность, что проблему, подобную вашей, уже решили.

## PyTorch

[PyTorch](https://pytorch.org/) — это среда машинного обучения на языке Python с открытым исходным кодом, обеспечивающая тензорные вычисления с GPU-ускорением. Она была разработана компанией Facebook.

Имеет множество модульных элементов, которые легко комбинировать.

* Легко писать собственные типы слоев и работать на GPU.
* Имеет широкий выбор предварительно обученных моделей.

[Keras](https://keras.io/) — открытая среда глубокого обучения, написанная на Python. Она была разработана инженером из Google Франсуа Шолле и представлена в марте 2015 года.

[Darknet](https://pjreddie.com/darknet/) — это фреймворк с открытым исходным кодом, написанный на языке C.

Он быстрый, лёгкий и удобный в использовании.

Так как фреймворк написан на C и не имеет другого API, то в случае, когда требования платформы или собственные предпочтения заставят обратиться к другому языку программирования, вам придётся дополнительно заморочиться над его интеграцией. К тому же он распространяется только в формате исходного кода, и процесс компиляции на некоторых платформах может быть несколько проблематичным.

3. Рассчитать для параметры уравнения регрессии, которые обеспечат оптимальное расположение линии регрессии среди множества точек поля корреляции. Построить регрессионную линейную функцию и экспериментальных данных. Решить в Excel