Лекция 4. Статистический анализ

Основные вопросы:

1. Описательные статистики
- Меры центральной тенденции
- Меры изменчивости
2. Категорийные данные. Особенности анализа категорийных данных
3. Статистические тесты для категорийных переменных
- Биномиальный тест.
- Хи–квадрат Пирсона.
- Точный критерий Фишера
4. Сравнение двух групп
5. Корреляция и линейная регрессия
6. Метод наименьших квадратов
7. Дисперсионный анализ
8. Двухфакторный дисперсионный анализ
9. Инструменты статистического анализа данных

Слово «статистика» происходит от латинского слова status, которое означает «состояние, положение вещей». Статистика изучает численность отдельных групп населения страны и её регионов, производство и потребление разнообразных видов продукции, перевозку грузов и пассажиров различными видами транспорта, природные ресурсы и т. п. Результаты статистических исследований широко используются для практических и научных выводов.

Такие характеристики, как среднее арифметическое, размах и мода, медиана, квартили, квантили и др. находят применение в статистике — науке, которая занимается получением, обработкой и анализом количественных данных о разнообразных массовых явлениях, происходящих в природе и обществе.

Основных характеристик две: *центр* и *разброс*.

Меры *центральной* тенденции: мода, медиана, среднее арифметическое.

Меры *изменчивости*: размах, дисперсия, стандартное отклонение.

Меры центральной тенденции

Средним арифметическим ряда чисел называется частное от деления суммы этих чисел на число слагаемых.

Среднее арифметическое представляет собой то значение величины, которое получается, когда сумма всех наблюдаемых значений мысленно распределяется поровну между единицами наблюдения. Например, вычислив среднее арифметическое удоев молока, полученных за сутки на ферме от всех коров, мы найдём тот удой, который получили бы на ферме в эти сутки от одной коровы, если бы все коровы давали одинаковое количество молока, т. е. найдём среднесуточный удой молока на ферме от одной коровы. Аналогично находят среднюю урожайность пшеницы с 1 га в районе, среднюю выработку рабочего бригады за смену и т. п.

Среднее арифметическое не всегда является лучшим показателем *типичности*.

Выброс – наблюдение в анализируемых данных, значение которого сильно отличается от других. Его часто описывают как значение в данных, которое как-будто

бы происходит из другой генеральной совокупности или выпадает из интервала типичных значений выборки. Предположим, вы исследуете учебную успеваемость в выборке или генеральной совокупности, и почти все испытуемые проучились от 11 до 16 лет (11 лет – окончание средней школы, 16 лет – оконченное высшее образование). Однако у одного из испытуемых значение этой переменной равно 0 (то есть он формально не получил никакого образования), а у другого – 26 (что предполагает много лет обучения после получения высшего образования). Вы, наверное, посчитаете эти два случая выбросами, поскольку их значения сильно отличаются от остальных данных в выборке или генеральной совокупности.

Обнаружение и анализ выбросов – это важный предварительный этап во многих видах анализа, потому что наличие даже одного или двух выбросов может кардинальным образом исказить значения некоторых обычных статистик, таких как среднее.

Кроме того, важно найти выбросы, потому что иногда они могут быть вызваны ошибками при вводе данных. В предыдущем примере первое, что стоит проверить, – это правильно ли были записаны значения; может оказаться, что правильные числа - это 10 и 16, соответственно. Второе, что стоит изучить, – это принадлежит ли данное наблюдение к исследуемой генеральной совокупности. Например, не относится ли 0 к продолжительности обучения ребенка, тогда как данные должны были содержать только информацию о взрослых?

Чтобы избавиться от выбросов, иногда применяют следующий метод: убирают по 5–10 % самых больших и самых маленьких данных и уже от оставшихся считают среднее. Получившийся показатель называют усеченным (или урезанным) средним.

Модой ряда чисел называется число, которое встречается в данном ряду чаще других.

Ряд чисел может иметь более одной моды, а может не иметь моды совсем. Моду ряда данных обычно находят, когда хотят выявить некоторый типичный показатель. Например, если изучаются данные о размерах мужских сорочек, проданных в определённый день в универмаге, то удобно воспользоваться таким показателем, как мода, который характеризует размер, пользующийся наибольшим спросом. Среднее арифметическое в этом случае не даёт полезной информации. Мода является наиболее приемлемым показателем при выявлении расфасовки некоторого товара, которой отдают предпочтение покупатели, цены на товар данного вида, распространённой на рынке, и т. п.

Среднее арифметическое ряда чисел может не совпадать ни с одним из чисел ряда, а мода, если она существует, обязательно совпадает с двумя или более числами ряда. Кроме того, в отличие от среднего арифметического понятие «мода» относится не только к числовым данным.

Медианой упорядоченного ряда чисел *с нечётным числом членов* называется число, записанное посередине, а **медианой** упорядоченного ряда чисел *с чётным числом членов* называется среднее арифметическое двух чисел, записанных посередине. **Медианой** произвольного ряда чисел называется медиана соответствующего упорядоченного ряда.

Если в упорядоченном числовом ряду содержится $2n–1$ членов, то медианой ряда является n–й член, так как $n–1$ членов стоит до n–го члена и $n–1$ членов – после n–го члена. Если в упорядоченном числовом ряду содержится $2n$ членов, то медианой является среднее арифметическое членов, стоящих на n–м и $n + 1$–м местах.

Вообще среднее арифметическое зависит от значений всех членов в упорядоченном ряду данных, в том числе и от значений крайних членов, которые часто бывают наименее характерными для рассматриваемой совокупности данных. Поэтому при анализе данных сведения о среднем арифметическом часто дополняются указанием медианы.

Такие показатели, как среднее арифметическое, мода и медиана, по–разному характеризуют данные, полученные в результате наблюдений. Поэтому на практике при анализе данных в зависимости от конкретной ситуации используют какой–либо из этих показателей, либо два из них, либо даже все три.

Пример: определение наиболее типичной зарплаты в нашей стране можно осуществлять по двум показателям–среднему арифметическому и медиане. Первая определяется как количество денег, деленное на количество людей, а второе – как зарплата человека, стоящего ровно посередине между самым бедным и самым богатым. Как правило, эти значения различаются – средняя зарплата выше медианной. И чем это различие больше, тем выше социальное неравенство в обществе.

Выбор меры центральной тенденции для анализа данных

Если распределение симметрично, унимодально (имеет одну моду) и не имеет заметных выбросов, то можно использовать любую меру центральной тенденции. Все меры дадут приблизительно одинаковые значение.

Если распределение ассиметрично, или имеет несколько мод, или имеет заметные выбросы, то использование в анализе данных среднего значения может привести к некорректным результатам. В таком случае лучше использовать медиану или моду для того, чтобы охарактеризовать данные с точки зрения степени выраженности некоторого количественного признака.

Меры изменчивости

Размахом ряда чисел называется разность между наибольшим и наименьшим из этих чисел.

Размах ряда находят, когда хотят определить, как велик разброс данных в ряду. Пусть, например, в течение суток отмечали каждый час температуру воздуха в городе. Для полученного ряда данных полезно не только вычислить среднее арифметическое, показывающее, какова среднесуточная температура, но и найти размах ряда, характеризующий колебание температуры воздуха в течение этих суток.

Однако, как и среднее арифметическое, эта мера очень чувствительна к выбросам. И, чтобы избежать искажений, мы должны отсечь 25% самых больших значений и 25% самых маленьких и найти размах для оставшихся. Эта мера называется **межквартильным размахом.**

Стандартное отклонение измеряет "средний" разброс значений переменной относительно ее среднего арифметического в тех же единицах измерения, что и сама переменная.

Чтобы понять, какое отклонение является наиболее типичным, можно найти среднее значение по этим отклонениям (сложить все отклонения и поделить их на количество данных).

Однако если так сделать, получим 0, поскольку одни отклонения являются положительными (когда значение больше среднего), другие – отрицательными (когда значение меньше среднего). Необходимо избавиться от знака.

2 способа:

1. взять модуль от отклонений;
2. возвести отклонения в квадрат.

Если найти среднее от квадратов отклонений, получим то, что называется дисперсией.

Дисперсия – отношение суммы квадратов отклонений от среднего арифметического к величине (n–1), где n – общее число измерений.

Для генеральной совокупности: $D = \frac{\sum(x_i - \bar{x})}{n}$.

Для выборки: $D = \frac{\sum(x_i - \bar{x})}{n-1}$.

Квадрат в формуле нахождения дисперсии делает её неудобной для оценки. Например, если мы измеряли размер чего-то в сантиметрах, то дисперсия имеет размерность в квадратных сантиметрах.

Среднеквадратическое отклонение – положительное значение квадратного корня из дисперсии; измеряет "средний" разброс значений переменной относительно ее среднего арифметического в тех же единицах измерения, что и сама переменная.

Среднеквадратическое отклонение и дисперсия неустойчивы к выбросам, как и среднее арифметическое.

Среднее значение и среднеквадратическое отклонение часто используются совместно для описания той или иной группы данных. Например, большинство (а именно около 68 %) данных находится в пределе одного среднеквадратического отклонения от среднего. Эти данные обладают так называемым нормальным размером. Оставшиеся (32%) либо очень большие, либо очень маленькие. Такое распределение признака называется **нормальным.**

Квантиль – значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если вероятность задана в процентах, то квантиль называется **процентилем** или **перцентилем**.

Например, фраза «для развитых стран 95–процентиль продолжительности жизни составляет 100 лет» означает, что ожидается, что 95 % людей не доживут до 100 лет.

Межквартильный размах о котором мы уже говорили, ещё называется и**нтерквартильным** в терминах квартилей определяется как разность между третьим и первым квартилями.

Анализ категорийных данных

Категорийные (категориальные, номинативные) данные – качественные характеристики объектов, измеренные в шкале наименований. Эти данные не могут быть упорядочены в пространстве. Например, если анализировать давление крови с использованием категорий (низкое, высокое, нормальное, гипертензия и др.).

Категорийные данные не подлежат измерению. Можно лишь оценить то, как часто они встречаются.

Анализ категорийных данных

Первое, что мы можем сделать, это построить сводную таблицу для того, чтобы узнать сколько раз встречаются категорийные данные.

Сводная таблица (кросс-таблица) – таблица, в которой разные группы данных упорядочены по строкам и по столбцам: одна группа данных соответствует строкам, другая – столбцам, на их пересечении – информация, объединяющая их. Сводные таблицы содержат частоту встречаемости категорийных данных.

Проверка статистических гипотез

Гипотеза – предположение, утверждение, догадка, предполагающее доказательство.

Структурная гипотеза – гипотеза о структуре изучаемого объекта («Два самых важных фактора выбора покупателя: цена и качество»).

Объяснительная гипотеза – гипотеза о причинно-следственных связях в исследуемых объектах («Увеличение рекламы влияет на доходы компании»).

Научная гипотеза – гипотеза, удовлетворяющая научному методу (объясняющая все факты, которые она призвана объяснить), принципиально проверяема (может быть проверена критическим экспериментом), логически не противоречива и не противоречит ранее установленным фактам, приложима к возможно более широкому кругу явлений.

Статистическая гипотеза (обозначается буквой H) – утверждение относительно неизвестного параметра (или параметров) генеральной совокупности, проверяемое на выборке.

Пример. H: mean=30 – неизвестное арифметическое среднее данной совокупности равно 30. Такое утверждение либо справедливо, либо нет.

Статистический вывод – утверждение, сделанное о параметрах генеральной совокупности, основываясь на результатах изучения выборочной совокупности.

Рассмотрим ещё одно важное понятие: **Статистическое оценивание** – в выборке найти показатель, максимально близкий к оцениваемому параметру (*точечное оценивание*), или интервал, в котором лежит этот параметр (*интервальное оценивание*) с большой вероятностью.

Проверка гипотез

Что делает исследователь?

- формулирует утверждение о параметрах генеральной совокупности (гипотезу),
- проводит исследование по выборочной совокупности данных и получает результаты,
- оценивает степень соответствия результатов сформулированной гипотезе,
- принимает решение о ложности или истинности гипотезы.

Гипотезы различают **по содержанию**. Это

- ❑ гипотезы о законах распределения

Пример. «Заработная плата сотрудников одного предприятия на одной должности имеет нормальное распределение».

- ❑ гипотезы о параметрах распределения

Пример. «Средние размеры деталей, производимых на однотипных параллельно работающих станках, не различаются между собой».

Есть ещё классификация гипоте по направленности и содержанию:

- ❑ направленные
- ✓ нулевые
- ✓ альтернативные
- ❑ ненаправленные
- ✓ нулевые

альтернативные

Нулевая гипотеза H_0 – гипотеза об отсутствии различий.
Альтернативная гипотеза H_1 – гипотеза о значимости различий.
Примеры.
Направленные гипотезы:
Нулевая гипотеза H_0: X не превышает У
Альтернативная гипотеза H_1: X превышает У
Ненаправленные гипотезы:
H_0: X не отличается от У
H_1: X отличается от У

Если необходимо проверить, правда ли в группе А произошли более сильные изменения, чем в группе В, необходимо сформулировать направленную статистическую гипотезу.

Если необходимо проверить, различаются ли формы распределения исследуемого признака в разных группах, формулируют ненаправленную статистическую гипотезу.

Проверяются статистические гипотезы путем расчета статистических критериев.

Статистический критерий – правило, которое обеспечивает надежное принятие истинной и отклонение ложной гипотезы.

Нулевая гипотеза опровергается или подтверждается в зависимости от соотношения критического и эмпирического значений критерия (если $\chi^2_{эмп} > \chi^2_{крит}$ то нулевая гипотеза H_0 отвергается).

Алгоритм проверки статистических гипотез

1. Сформулировать нулевую и альтернативную гипотезы.
2. Выбрать подходящий статистический критерий (k).
3. Рассчитать по данным выборки эмпирическое значение $k_{эмп}$.
4. Определить критическое значение критерия $k_{крит}$ на основании объема выборки, числа степеней свободы, p-уровня значимости.
5. Сравнить эмпирический и критический значения критерия.
6. Если $k_{эмп} > k_{крит}$, то отвергнуть нулевую гипотезу (исключения составляют: критерий Манна-Уитни, критерий Т-Уилкоксона, критерий знаков).

Статистические критерии делятся параметрические и непараметрические

- *параметрические* (в формулу расчета включают параметры распределения: средние и дисперсии (например, t-критерий Стьюдента, F-критерий и др.))
- *непараметрические* (в формулу расчета не включают параметры распределения, основаны на оперировании рангами или частотами (критерий Уилкоксона, критерий Розенбаума и др.)).

p-уровень значимости (уровень статистической значимости) – вероятность того, случайные различия были признаны достоверными, существенными.

$p \leq 0,05$ означает утверждение, что различия достоверны или существенны на 5%-ном уровне значимости, т.е. вероятность того, что различия незначимы (несущественны, недостоверны) – не больше 5%.

- **Низший** уровень статистической значимости часто берут равным 5% ($p \leq 0,05$);
- **достаточный** – 1% ($p \leq 0,01$);
- **высший** – 0,1% ($p \leq 0,001$).

Важно знать при выборе статистического критерия для проверки гипотезы:

– о чем гипотеза, какой тип статистической задачи необходимо решать;

– шкалы измерения данных;

– размеры выборки (или выборок);

– можно ли выбранный критерий применить к неравным по объему выборкам, если исследование осуществляется на более, чем на одной выборке;

– мощность критерия;

– возможности, которыми располагает исследователь для расчета критерия.

Мощность статистического критерия – способность критерия выявлять различия, если они есть.

Проверяя статистическую гипотезу, исследователь рискует принять неправильное решение.

Ошибка I рода: нулевая верная гипотеза была отклонена.

Вероятность ошибки I рода обозначается буквой α. Вероятность правильно принятого решения: $(1-\alpha)$. Чем меньше вероятность ошибки, тем больше вероятность правильного решения.

Ошибка II рода: приняли нулевую гипотезу в то время, как она неверна.

Вероятность ошибки II рода обозначается буквой β. Вероятность правильного решения $(1 - \beta)$ – мощность статистического критерия.

Основные случаи проверки гипотез о параметрах генеральной совокупности:

- гипотезы о средних
  – о равенстве среднего определенному значению
  – о значимости различия между средними
- гипотезы о дисперсиях
  – о равенстве дисперсий определенному значению
  – о значимости различия дисперсий двух совокупностей
- гипотезы о коэффициентах корреляции
  – о равенстве коэффициентов корреляции определенному значению
  – о значимости различия коэффициентов корреляции двух совокупностей
- гипотезы о долях признака
  – о равенстве доли признака определенному значению
  – о значимости различия долей признака в двух совокупностях
- гипотезы о независимости признаков в корреляционной таблице

Наиболее частый метод, применяемый в статистическом анализе – сравнение средних значений двух выборок. При таком сравнении предполагается, что исследуемые выборки подчинены *нормальному распределению*. Если условие нормальности не выполнено, используются непараметрические тесты.

Тестовые ситуации при сравнении средних значений выборок:

1. Сравнение двух зависимых выборок: t-тест Стьюдента для зависимых выборок.
2. Сравнение двух независимых выборок: t-тест Стьюдента для независимых выборок.
3. Сравнение более двух зависимых выборок: однофакторный дисперсионный анализ с повторными измерениями.
4. Сравнение более двух независимых выборок: однофакторный дисперсионный анализ.

Чтобы понять, насколько они отличаются друг от друга, необходимы, так называемые, **меры различий для несвязанных выборок**.

t-критерий Стьюдента для несвязанных выборок оценивает, насколько различаются их средние размеры.

Чтобы рассчитать **t-критерий Стьюдента**, необходимо из средней одной совокупности вычесть среднюю другой и поделить их на **стандартную ошибку** этой разности. Последняя вычисляется на основе стандартных отклонений размеров обоих совокупностей и нужна для приведения t-критерия к нужной размерности.

Чем больше t-критерий, тем с большей уверенностью можно утверждать, что в среднем одна совокупность отличается от другой.

Сравнение двух групп. Критерий t-Стьюдента

- ✓ Форма t-распределения зависит от числа степеней свободы выборки (числа параметров, которые могут изменяться).
- ✓ В случае t-распределения основной эффект на число степеней свободы оказывает размер выборки, и у тестов для более крупных выборок в целом больше степеней свободы, чем в случае небольших выборок.
- ✓ t-распределение непрерывное и симметричное.

Причины использования t-распределения при проверке различий в средних:

1. работа с совокупностью, которая, возможно, распределена нормально,
2. неизвестное стандартное отклонение генеральной совокупности (когда вместо стандартного отклонения генеральной совокупности приходится использовать стандартное отклонение выборки).

Если в данных есть точки выброса, необходимо проверить данные на однородность и нормальность распределения. Для выяснения есть ли в данных выбросы часто используют диаграмму размаха – боксплот.

Проверка данных на нормальность распределения (критерий Шапиро-Уилка)

Нулевая гипотеза: случайная величина распределена нормально.

Альтернативная гипотеза: случайная величина не распределена нормально.

Если значение **p-value>0,05**, тогда принимаем нулевую гипотезу. Если нет, тогда принимаем альтернативную.

Проверка гемогенности дисперсии (критерий Бартлетта)

Нулевая гипотеза: данные однородные.

Альтернативная гипотеза: данные неоднородные.

Тест Бартлетта наиболее чувствителен к отклонениям от нормальности (это не

то же самое, что равенство дисперсий), так что его следует применять, только если есть уверенность в примерно нормальном распределении совокупностей, из которых взяты выборки.

После проверки нормальности распределения и однородности данных, можно перейти к построению **t-критерия Стьюдента**.

Одновыборочный t-тест предназначен для проверки равенства математического ожидания нормально распределенной случайной величины (для которой известна лишь выборка) некоторому заданному значению в предположении, что дисперсия не известна.

Нулевая гипотеза:$M(X) = \mu$

Двухвыборочный тест служит для сравнения математических ожиданий нормально распределенных случайных величин в предположении, что их дисперсии равны, хотя и не известны.

Нулевая гипотеза: $M(X) - M(Y) = \mu$

Поскольку формула t-критерия включает в себя средние значения, то этот критерий будет давать неадекватные результаты при наличии выбросов. Чтобы этого избежать, можно либо исключить выбросы из анализа, либо воспользоваться непараметрическим **U-критерием Манна-Уитни**.

Чтобы рассчитать критерий Манна-Уитни для двух выборок, необходимо выстроить все данные в один ряд в порядке возрастания и назначить им ранги. Большему значению достанется первый ранг, а самому маленькому – последний. После этого считают суммы рангов отдельно для каждой выборки. Общая логика такова: чем сильнее будут различаться эти суммы, тем больше различаются данные в выборках.

Наконец, проводим некоторые преобразования (которые в основном сводятся к поправкам на количество данных в выборках) и получаем критерий Манна-Уитни, по которому судим, в действительности ли данные выборки отличаются.

Тест Уилкоксона (Манна-Уитни) используется в случае:

- невыполнения одного из требований к t-критерию Стьюдента
- когда не так много данных
- не очень хорошего распределения

Тест хи–квадрат Пирсона

Критерий хи–квадрат – самый распространенный способ изучения связей между двумя и более категорийными переменными. Тест включает расчет статистики хи–квадрат и ее сравнение с распределением хи–квадрат. Среди нескольких типов критерия хи–квадрат, тест хи–квадрат Пирсона – один из самых простых.

Есть три типа критериев хи–квадрат.

1. ***Критерий независимости хи–квадрат***.

Допустим, у нас есть 2 переменные. Критерий проверяет нулевую гипотезу о независимости этих переменных друг от друга (об отсутствии связи между ними). Альтернативная гипотеза: переменные зависимы, связаны между собой.

Пример: в случайной выборке взрослых есть данные о курении и наличии диагноза рака легких (см.табл.).

Таблица. Курение и рак легких

	Рак легких диагностирован	Рак легких не диагностирован
Курят	60	300
Не курят	10	390

Если посмотреть на данные, не вычисляя никаких критериев, бросится в глаза, что есть связь между курением и раком лёгких. Но наши предположения могут быть обманчивыми. Проведем тест хи–квадрат на независимость.

Гипотезы:

H_0: курение и рак легких независимы,

H_1: курение и рак легких зависимы.

Критерий хи–квадрат основан на разнице между наблюдаемыми и ожидаемыми значениями в каждой из ячеек таблицы сопряженности. Наблюдаемые значения – те значения, которые были получены из данных по выборке, ожидаемые значения – те значения, которые мы ожидаем увидеть в случае независимости этих переменных.

Результаты теста: p–значение равно 3.645e-11. Это много меньше 0,05 ==> отвергается нулевая гипотеза об отсутствии связи между курением и раком легких. Т.е. велика вероятность возникновение рака лёгких у курильщиков!

2. ***Критерий равенства пропорций хи–квадрат*** проверяет другую гипотезу.

Критерий равенства пропорций используется с данными, взятыми из ***нескольких независимых выборок***, а **нулевая гипотеза** состоит в том, что распределение какой–то переменной одинаково во всех генеральных совокупностях.

3. ***Критерий согласия хи–квадрат*** используют для проверки гипотезы о том, что распределение некоторой категорийной переменной в генеральной совокупности совпадает с заданным распределением, а альтернативная гипотеза гласит, что распределение этой переменной иное, не предполагаемое.

Замечания и ограничения по применению критерия Пирсона χ^2

• Объем выборки должен быть достаточно большим (>30).

• Теоретическая частота для каждой клетки таблицы должна быть больше 5.

• Выбранные разряды должны охватывать весь диапазон вариативности признака, а группировка на разряды должна быть одинаковой во всех сопоставляемых распределениях.

• Разряды должны быть не перекрещивающимися, т.е. если наблюдение отнесено к одному разряду, то оно не может быть отнесено к другому.

• Сумма наблюдений по разрядам должна быть равна общему числу наблюдений.

• Если сравниваются два эмпирических распределения, то теоретические частоты рассчитываются как отношение произведения итогов по строке на итог по столбцу к общему числу наблюдений.

1. Точный критерий Фишера.

Точный тест Фишера – это непараметрический критерий, аналогичный тесту хи–квадрат, но его можно применять с небольшим количеством данных или в случае разреженного распределения данных, которые не подходят под требования хи–квадрата. **Тест Фишера** основан на гипергеометрическом распределении и рассчитывает точную вероятность наблюдения такого распределения, как в данных, или более экстремального, отсюда в названии и появилось слово «точный». Это не

асимптотический тест, так что он не ограничен правилами о разреженности, которые относятся к тесту хи–квадрат. Обычно для расчета теста Фишера используют компьютерные программы, особенно для таблиц большего размера, чем 2×2, из-за громоздкости расчетов. Рассмотрим пример с таблицей 2×2. Будем исследовать связь между употреблением уличного наркотика и внезапной остановкой сердца у молодых людей. Т.к. наркотик незаконный и новый, а остановки сердца достаточно редко встречаются у молодых людей, достаточно данных для того, чтобы провести тест хи–квадрат, собрать не удалось, В табл. 7.3 приведены данные для анализа.

Таблица. Точный тест Фишера: расчет связи между употреблением уличного наркотика и внезапной остановкой сердца у молодежи

	Остановка сердца	Нет остановки сердца	Сумма
Употребляли наркотик	7	2	9
Не употребляли наркотик	5	6	11
Сумма	12	8	20

Гипотезы:

H_0: риск внезапной остановки сердца у употребляющих и не употребляющих наркотик одинаковый.

H_1: риск внезапной остановки сердца у употребляющих наркотик выше.

Точный тест Фишера рассчитывает вероятность получить результат не менее экстремальный, чем тот, который был найден в исследовании. Более экстремальный результат в данном случае – такой, в котором отличие в частоте внезапной остановки сердца у употребляющих и не употребляющих наркотик больше, чем в наших данных (таб.).

Алгоритмы расчета **теста Фишера** включены практически во все статистические пакеты.

Корреляция

В естественных науках чаще всего сталкиваются со строгими (функциональными) зависимостями (каждому значению одной переменной соответствует единственное значение другой. Но в большинстве случаев между переменными таких зависимостей нет. К примеру, нет строгой зависимости между экономическими переменными: ценой и спросом, доходом и потреблением, стажем работы и производительностью труда.

В самых разных областях знания возникает *задача определения зависимости между случайными величинами,* являющимися признаками одних и тех же объектов:

- между ростом и весом человека;
- между затратами компании на рекламу и доходом от продаж;
- между силой сигнала на входе и выходе технического устройства;
- между уровнем инфляции и безработицей и т.д.

Причины отсутствия строгой зависимости переменных:

- не учитывается целый ряд факторов, влияющих на конкретную переменную при анализе влияния её на другую;
- влияние может быть не прямым, а проявляться через цепочку других факторов;
- многие такие воздействия носят случайный характер.

В таких случаях говорят не о функциональных, а о *корреляционных*, либо статистических зависимостях.

Статистической называют зависимость, при которой изменение одной из

величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой. Такую статистическую зависимость называют ***корреляционной***.

Корреляции зачастую рассчитывают в разведочной фазе исследовательского проекта. Они помогают увидеть, как связаны друг с другом различные непрерывные переменные. Для исследования этих связей также строят диаграммы рассеяния. Некоторые корреляции интересны сами по себе, их логично использовать как отдельные величины, и их можно проверять на достоверность.

Корреляция – мера наблюдаемой связи. Сама по себе корреляция никак не может выявить причину. Многие переменные в реальном мире сильно коррелируют друг с другом, но эти связи могут объясняться случаем, влиянием других переменных или другими неизвестными причинами. Даже если между величинами есть причинно–следственная связь, она может работать в другую сторону, чем мы предполагаем. Поэтому даже самая сильная корреляция сама по себе не может свидетельствовать о причинно–следственной связи; она может быть подтверждена только с помощью постановки эксперимента.

Связь

В повседневной жизни вокруг нас огромное количество переменных, которые кажутся связанными друг с другом, и обнаружение этих связей – основная задача науки. Нет ничего сложного в понимании взаимосвязей между величинами. Люди думают в терминах связей и часто ассоциируют с ними причинно–следственные взаимодействия. Спортсмены, которые тратят сотни часов на тренировки, делают это, потому что уверены, что такой подход приведет их к успеху. Родители, наставляющие детей питаться здоровой пищей, а не фастфудом, делают это, поскольку думают, что есть связь между здоровьем и рационом питания. Людям свойственно замечать, что какие–то события вроде бы происходят одновременно, и соответственно верить, что одно становится причиной появления другого. Часто некоторые наши здравые мысли поддерживаются экспериментальными данными, иногда – нет. Как ученые, мы должны научиться задавать себе вопрос: является ли кажущаяся связь действительной, и, если да, то есть ли в ней причинноследственные взаимоотношения?!

Примеры *неверных выводов*, основанных на наблюдениях:

• Есть сильная связь между результатом теста по грамматике русского языка и ростом человека, что можно объяснить тем, что у больших людей и мозг больше, поэтому они могут запомнить больше правил грамматики.

• Уровень рождаемости в регионе сильно связан с числом аистов, проживающих на данной территории, так что, очевидно, что именно аисты приносят детей.

Верные объяснения:

• Тест проводили у школьников, не учитывая их возраст. Вероятно, что чем старше школьник, тем выше его рост, и тем больше он знает правил русского языка. Таким образом, наблюдаемая связь между результатами теста и ростом обусловлена третьей переменной, возрастом.

• Рождаемость всегда была выше не в городах, а в сельской местности, и аисты чаще встречаются в сельской местности, так что связь можно объяснить влиянием другого фактора, типа местности.

Стоит отметить, что даже если отсутствует логичная причина связи двух

переменных, она может обнаружиться между ними совершенно случайно. Это важно в исследовании очень больших выборок, где даже очень слабая корреляция может быть статистически значимой, даже не имея практического значения. И наоборот, даже в случаях сильных связей между переменными (курение и рак легких), на уровне отдельных случаев эта связь может проявляться по–разному. Кто–то курит на протяжении всей жизни и не заболевает, а кто–то никогда в своей жизни не курил и получит рак легких.

Корреляции помогают увидеть характер связи между двумя непрерывными переменными. Исследуют такую связь построением диаграммы рассеяния, в которой независимую переменную откладывают по оси x, зависимую – по оси y. Когда нет данных по поводу взаимоотношений переменных, то ось не имеет значения. Каждому значению выборки соответствует точка на графике с координатами (x, y). Диаграммы рассеяния помогают почувствовать свойства связи между переменными: форму (линейная, квадратичная и др.), направление (положительное, отрицательное), силу (сильная, слабая). С помощью диаграмм рассеяния можно получить общее впечатление о разбросе данных, увидеть выбросы, если они имеются.

Полезно создавать матрицу диаграмм рассеяния – множество таких диаграмм, на которых мы легко можем увидеть связи между парами переменных.

Существуют **два подхода** к оценке силы корреляции.

1. Первый опирается только на абсолютную величину коэффициента корреляции

Таблица 8.1 Оценка силы корреляции по абсолютной величине коэффициента корреляции

Абс. величина коэффициента корреляции |r|	**Сила корреляции**
$\|r\| \geq 0{,}70$	сильная
$0{,}50 \leq \|r\| \leq 0{,}69$	средняя
$0{,}30 \leq \|r\| \leq 0{,}49$	умеренная
$0{,}20 \leq \|r\| \leq 0{,}29$	слабая
$\|r\| \leq 0{,}19$	очень слабая

Для реальных данных мы не ожидаем, что график будет идеальной прямой, даже если имеется довольно сильная линейная зависимость.

1. Второй ориентирован на уровень значимости данного коэффициента корреляции при данном объеме выборки (см. таб. На слайде).

Таблица 8.2 Оценка силы корреляции по уровню статистической значимости

Уровень статистической значимости коэффициента корреляции	**Оценка силы корреляции**
$p \leq 0{,}01$	высокая значимая корреляция
$0{,}01 < p \leq 0{,}05$	значимая корреляция
$0{,}05 < p \leq 0{,}10$	тенденция достоверной связи
коэффициент корреляции не достигает уровня статистической значимости	незначимая корреляция

В соответствии со вторым подходом, чем больше объем выборки, на которой

изучалась корреляция между переменными, тем меньший по абсолютной величине коэффициент корреляции признается показателем достоверности корреляционной связи.

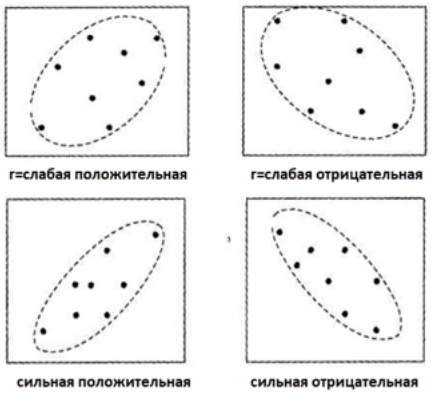


Рис. Схематическое изображение силы и направления корреляции

Наличие значимой корреляции между переменными дает основание не отвергать гипотезу о причинно–следственной связи между переменными, но не служит подтверждением такой гипотезы.

Причинность и корреляция

Т.е. наличие корреляции между переменными не означает, что между ними существует причинно–следственная связь. Возможно, либо одна из переменных является частичной причиной другой, либо обе они являются следствием каких–то общих причин. Наличие значимой корреляции между переменными дает нам основание не отвергать гипотезу о причинно–следственной связи между переменными, но не может служить подтверждением такой гипотезы. Изучая корреляционную зависимость между переменными, исследователь, интерпретируя значения коэффициентов корреляции, должен учитывать сущность, природу этих переменных.

В качестве мер корреляции используются различные коэффициенты. Выбор коэффициента зависит от типа данных, которые обрабатываются. Широко распространенными в маркетинговых, социологических, психолого–педагогических и других исследованиях являются коэффициенты ассоциации и контингенции, коэффициент хи–квадрат Пирсона, коэффициент ранговой корреляции Спирмена, коэффициент ранговой корреляции Кенделла, коэффициент Пирсона–Браве для непрерывных шкал.

Коэффициент корреляции Пирсона (линейный коэффициент корреляции) – мера связи для двух непрерывных или характеризующих отношения переменных.

Обозначается как r для выборки и ρ для генеральной совокупности. Коэффициенты корреляции могут принимать значения от –1 до +1. 0 (нуль) свидетельствует об отсутствии связи между переменными.

Большие абсолютные значения показывают более сильную связь (если никакая из переменных не является константой, см. рис. 8.3). Всегда необходимо строить график для исследуемых данных, поскольку значение коэффициента корреляции может вводить в заблуждение, если на самом деле связь нелинейная.

Примеры диаграмм рассеяния данных с разной величиной r:

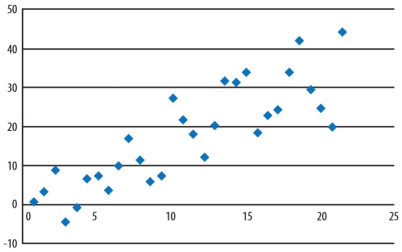


Рис. Диаграмма рассеяния (r = 0.84)

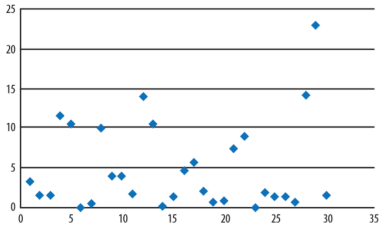


Рис. Диаграмма рассеяния (r = 0.09)

Частая *нулевая гипотеза* для корреляционного анализа: переменные не связаны (r = 0).

Альтернативная гипотеза: r ≠ 0.

Формула для коэффициента корреляции Пирсона:

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

SS_x– это сумма квадратов отклонений x,

SS_y – это сумма квадратов отклонений y,

SS_{xy} – это сумма квадратов отклонений x и y.

Чтобы найти сумму квадратов x:

1. Найти отклонение (из каждого значения x вычесть среднее по всем значениям x).
2. Возвести каждое отклонение в квадрат.
3. Найти сумма квадратов отклонений.

Формула суммы квадратов отклонений:

$$SS_x = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

x_i – отдельное значение x,

$\bar{x}$ – среднее по всем значениям x

n – объем выборки.

Или:

$$SS_x = \sum_{i=1}^{n} {x_i}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$$

Расчет суммы квадратов отклонений x и y:

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Или:

$$SS_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}$$

Формула для проверки статистической значимости коэффициента корреляции Пирсона:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Формула рассчитывает статистику для проверки значимости отличия наших результатов от 0, Эта статистика имеет t–распределение с (n – 2) степенями свободы; степени свободы – статистический термин, характеризующий число величин, которые могут меняться в определенной ситуации. Это также число, которое необходимо знать, чтобы использовать правильное t–распределение для оценки результатов.

Регрессионный анализ

Еще один метод определения связи между двумя варьирующими переменными – *регрессионный анализ*. В отличии от корреляции, он способен объяснить, как переменная предсказывает другую переменную, так как позволяет увидеть взаимосвязь между количественными переменными.

Оценивать *y* по *x* можно имея данные о функциональном отношении связи между переменными *x* и *y*.

Переменная, которую необходимо оценить – зависимая переменная (*y*), то, что измеряется в эксперименте: правильность ответа, скорость реакции и т.д. Переменная в фокусе интереса, которую используют для оценки зависимой переменной называется независимой (фактором, предикатором) (*x*). Главная задача – узнать, имеет ли предиктор какое-то отношение к зависимой переменной. Приведет ли изменение предиктора к изменению в зависимой переменной? То есть выразить корреляционную зависимости в виде функциональных отношений.

Примеры задач оценивания *y* по *x*:

1. Предсказание объема продаж (*y*) по рекламным затратам (*x*).
2. Прогнозирование текучести кадров (*y*) в зависимости от уровня удовлетворенности работой (*x*) и др.

Пример. Необходимо предсказать средний балл оценок студентов, полученных во время сессии (*y*) в зависимости времени, потраченного на просмотр в период подготовки к экзаменам сериалов (*x*). Для этого:

1) узнать, сколько часов в неделю n студентов потратили на просмотр сериалов во время подготовки к сессии;
2) измерить средний балл оценок, полученных теми же n студентами во время сессии;
3) вывести уравнение, которое связывает *y* и *x*;
4) использовать уравнение среднего балла оценок, полученных в сессию (*y*) теми студентами, про которых известно сколько времени они потратили на просмотр сериалов в период подготовки к сессии (*x*).

На практике на значение исследуемой величины влияет множество факторов. Для простоты будем считать, что основное влияние оказывает один из них, соответственно анализ будем называть *однофакторным.*

Будем считать, что оба признака, зависимость между которыми мы выявляем,

могут быть представлены как значения вещественных переменных. Предположим, что нам известны результаты n измерений. Каждое измерение i ($i = \overline{1,n}$) даёт пару чисел (x_i, y_i)– значения двух признаков измеряемого объекта (например, затраты на рекламу – доход, рост – вес и др.).

Построим таблицу, каждая строка в которой соответствует одному объекту (наблюдению). Признак, который может быть измерен (x), является фактором (предиктором), а прогнозируемая переменная (y) – переменная отклика.

Цель исследования – построить (линейную) функцию (регрессионную модель), которая по известному значению фактора (x) позволит прогнозировать значение переменной отклика (y).

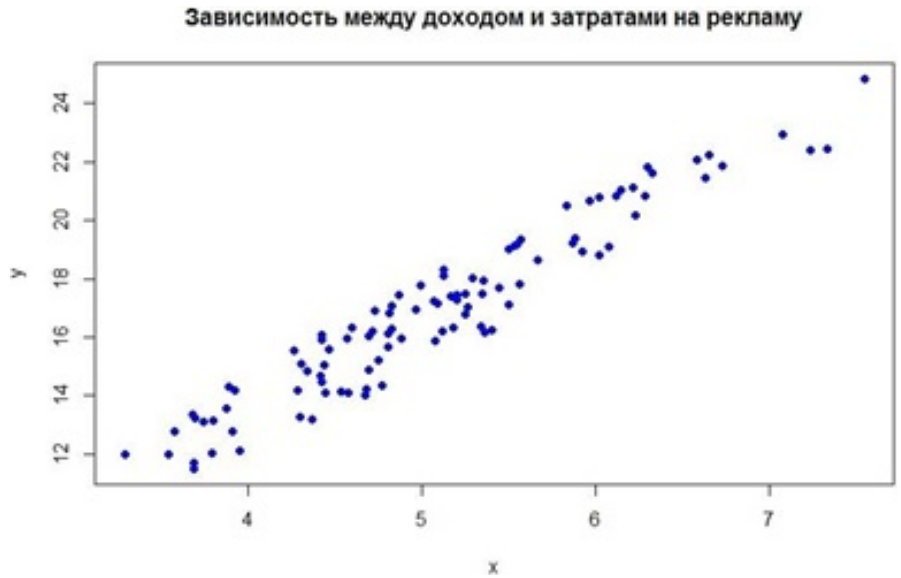
Простейшая форма функциональной зависимости: **линейная**. Для линейной регрессии: найти такую прямую (линию предсказания или по-другому, **линию тренда**), сумма квадратов отклонений от которой точек диаграммы рассеивания минимальна (сумма квадратов ошибки оценки минимальна). **Уравнение** такой прямой – **уравнение регрессии**

№ наблюдения, i	Значения фактора, x_i	Значения переменной отклика, y_i
1	x_1	y_1
…		
i	x_i	y_i
…		
n	x_n	y_n

Пример

Построим систему координат, где по оси абсцисс будем откладывать значения фактора (x), по оси ординат – значения переменной отклика (y). Таким образом, каждому наблюдению (т.е. каждой паре) (x_i, y_i) ($i = \overline{1,n}$) соответствует точка на координатной плоскости. Если бы зависимость между изучаемыми признаками была линейной, то все эти точки лежали бы на одной прямой. Однако из–за наличия случайной компоненты («шума») точки разбросаны по координатной плоскости.

Пример: ось абсцисс – затраты на рекламу, ось ординат – объём продаж, зафиксированный через заданное время после проведения рекламной кампании.



Задача: найти такую линейную функцию, которая наилучшим образом отражает

зависимость переменной отклика y (объёма продаж) от фактора x (затрат на рекламу).

Эта задача называется *задачей однофакторной линейной регрессии*.

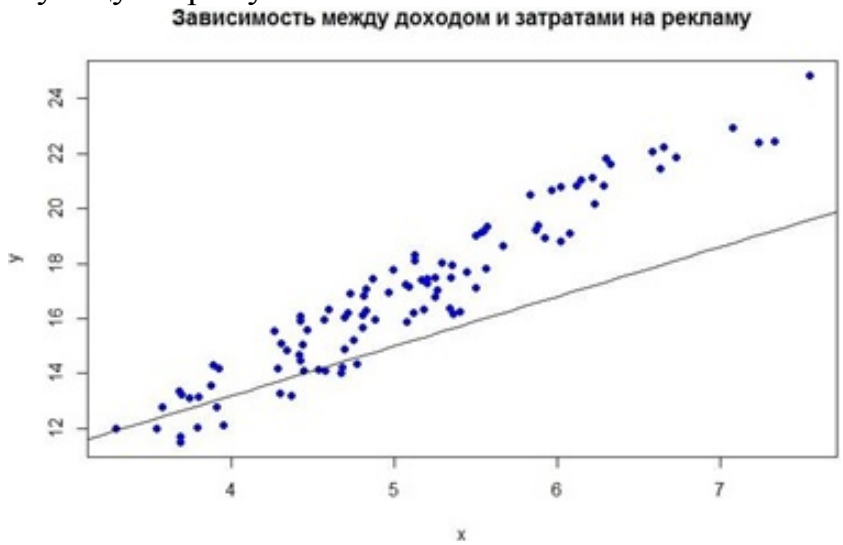
Математическая постановка задачи нахождения уравнения регрессии

Линейная функция одной переменной x имеет вид:

$$y = a + bx \qquad (1)$$

где b – тангенс угла наклона графика функции к оси OX, a – ордината точки пересечения этой прямой с осью OY. Задача: найти такие значения переменных a и b, при которых прямая (1) наилучшим образом проходит через облако точек (x_i, y_i) $(i = \overline{1, n})$.

Поясним геометрически смысл задачи. Зафиксируем произвольные значения a , b и построим соответствующую прямую:



Видно, что построенная «наугад» прямая, не лучшая для данного облака точек.

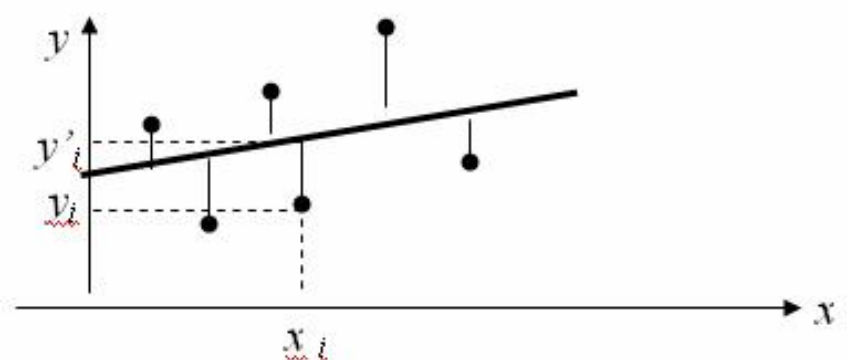
При фиксированных a и b «ожидаемое» значение y при $x = x_i$, составляет $a + bx_i$, $i = \overline{1, n}$ (т.е. точка $(x_i, a + bx_i)$ лежит на построенной прямой). Но фактическое значение переменной y при $x = x_i$ составляет y_i, т.е. «ошибка» составляет $\varepsilon_i = a + bx_i - y_i$. Наша задача: найти значения a и b, минимизирующих сумму квадратов ошибок (чтобы сумма квадратов отклонений наблюдаемых значений от значений на прямой линии регрессии оказалась наименьшей):

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a + bx_i)^2 \to min \qquad (2)$$

Метод наименьших квадратов (МНК)

Метод наименьших квадратов – принцип поиска коэффициентов регрессии путём минимизации суммы квадратов отклонений между реальными значениями признака и прогнозируемыми согласно предполагаемой форме зависимости (в нашем случае – линейной).

Таким образом, задача метода наименьших квадратов: подобрать прямую линию, которая ближе всего расположена к точкам корреляционного поля. Согласно МНК, линия выбирается так, чтобы сумма квадратов расстояний по вертикали между точками корреляционного поля и этой линией была бы минимальной.



Математическая запись данной задачи:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - a + bx_i)^2 \to min$$

Значения y_i и x_i $(i = \overline{1,n})$ – данные наблюдений. В целевой функции задачи S они представляют собой константы. Переменными в данной функции являются искомые оценки параметров $\tilde{a}$ и $\tilde{b}$. Чтобы найти минимум целевой функции двух переменных необходимо вычислить частные производные данной функции по каждому из параметров, приравнять их нулю:

$$\frac{\partial S}{\partial \tilde{a}} = 0, \frac{\partial S}{\partial \tilde{b}} = 0$$

В результате получим систему линейных уравнений:

$$\begin{cases} \sum_{i=1}^{n} y_i = \tilde{a} \cdot n + \tilde{b} \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i \cdot x_i = \tilde{a} \sum_{i=1}^{n} x_i + \tilde{b} \sum_{i=1}^{n} {x_i}^2 \end{cases}$$

Далее вводятся обозначения:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ , \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \ , \qquad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i\, y_i \ , \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2$$

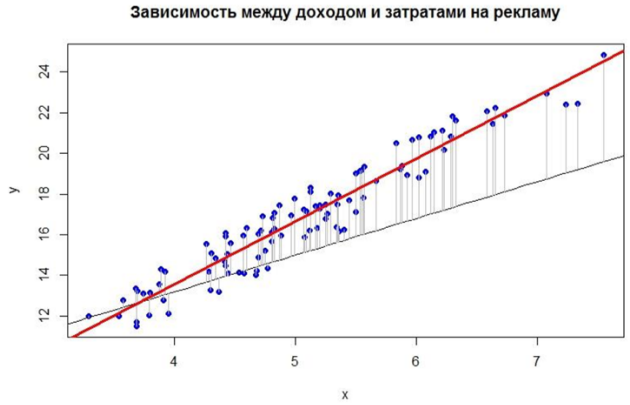
Решая данную систему, найдем искомые оценки параметров:

$$\tilde{b} = \frac{\overline{x} \cdot \overline{y} - \overline{xy}}{\overline{x}^2 - \overline{x^2}}$$

$$\tilde{a} = \overline{y} - \tilde{b}\overline{x}$$

Правильность подсчета параметров уравнения регрессии можно проверить сравнением сумм $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \tilde{y}_i$ (расхождение возможно из–за округления расчетов).

Учитывая свойства функции S нетрудно показать, что это решение является точкой минимума функции. Иными словами, значения $\tilde{a}$ и $\tilde{b}$ обеспечивают получение наилучшей линейной функции, отражающей зависимость переменной отклика *y* от фактора *x*. График этой линейной зависимости называется прямой регрессии (*y* на *x*). На приведённом ниже рисунке эта прямая имеет красный цвет.



Найденная линейная функция позволяет прогнозировать значение зависимого признака (y) по заданным значениям независимого фактора (x).

Дисперсионный анализ

Регрессия и дисперсионный анализ (ANOVA) – два статистических метода, использующих общую линейную модель (GLM), в основе которых лежит предположение о том, что зависимая переменная представляет собой функцию от одной или более независимых переменных.

Дисперсия – характеристика рассеивания данных вокруг их среднего значения.

Дисперсионный анализ (ANOVA) – статистическая процедура, используемая для сравнения средних значений определенной переменной в двух и более независимых группах.

Основная статистика в дисперсионном анализе – **F-отношение**, используемое для выявления статистической значимости различий между группами.

В дисперсионном анализе теоретическое распределение F-отношения не является нормальным, оно подчиняется распределению Фишера. Изменение формы распределения Фишера в зависимости от числа степеней свободы (от количества групп и общего числа наблюдений), показано на рисунках 8.1-8.3. F-значение всегда является положительным, потому вероятность отклонения рассчитывается только в правую сторону.

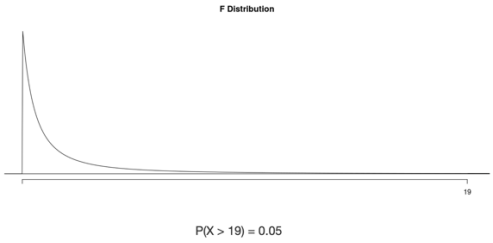


Рис. 8.1. F-распределение для числа групп – 2 и числа степеней свободы – 2

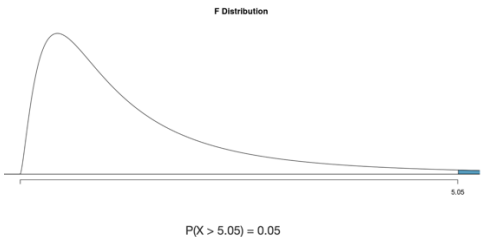


Рис. 8.2. F-распределение для числа групп – 5 и числа степеней свободы – 5

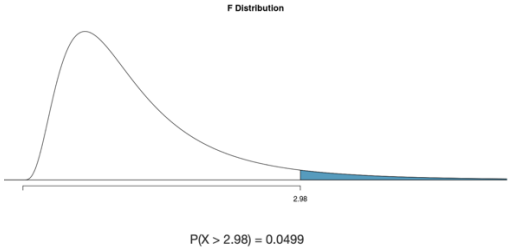


Рис. 8.3. F-распределение для числа групп – 10 и числа степеней свободы – 10

В дисперсионном анализе рассматривается отношение двух дисперсий: **межгрупповой** и **внутригрупповой**.

Общая сумма квадратов SST (общая изменчивость данных) **–** показатель, характеризующий степень изменчивости данных без учета разделения их на группы. Вычисляется общая сумма квадратов следующим образом:

- для каждого наблюдения рассчитывается насколько оно отклонится от среднего значения,
- складывается сумма квадратов полученных отклонений.

Общая сумма квадратов SST получена из двух источников: **межгрупповая сумма квадратов** SSB (характеристика, показывающая насколько групповые средние отклоняются от общего среднего) и **внутригрупповая сумма квадратов** SSW (сумма квадратов отклонений от среднего внутри каждой из групп).

Межгрупповая дисперсия MSB, объяснённая влиянием фактора, характеризует рассеивание значений между градациями (группами) вокруг средней всех данных.

Внутригрупповая дисперсия MSW, необъяснённая, характеризует рассеивание данных внутри градаций фактора (групп) вокруг средних значений этих групп.

Отношение межгрупповой и внутригрупповой дисперсий – **фактическое отношение Фишера**. Его сравнивают с **критическим значением отношения Фишера**. В случае, когда фактическое отношение Фишера превышает критическое, то средние классов градации различны, а исследуемый фактор оказывает существенное влияние на изменение данных. В обратном случае: средние классов градации друг от друга не отличаются, а фактор не оказывает существенного влияния на изменение данных.

Целью дисперсионного анализа является исследование наличия/отсутствия существенного влияния некоторого количественного/качественного фактора на изменения исследуемого признака.

Фактор, предположительно имеющий/не имеющий существенное влияние, делят на группы и на основе исследования значимости средних в наборах данных, соответствующих группам фактора, выясняют одинаково ли влияние фактора.

Пример 1. Исследование зависимости прибыли предприятия от типа используемого сырья. В данном случае группы – типы сырья.

Пример 2. Исследование зависимости себестоимости выпуска единицы продукции от размера предприятия. Здесь группы – величины предприятий (малое, среднее, большое).

Минимальное число групп в дисперсионном анализе – две. Группы могут быть количественные и качественные.

В дисперсионном анализе вычисляется удельный вес суммарного воздействия одного/нескольких факторов. Насколько влияние фактора существенно, исследуется с помощью гипотез:

Нулевая гипотеза H_0 утверждает, что все *a* классов градации имеют одинаковые значения средних: $\mu_1 = \mu_2 = ... = \mu_a$.

Альтернативная гипотеза H_1: не все классы градации имеют одно значение средних.

Статистический комплекс в дисперсионном анализе – таблица эмпирических данных. Статистический комплекс называется **однородным** (гомогенным), если во всех классах градаций одинаковое число вариантов, и **разнородным** (гетерогенным) – если число вариантов разное.

Однофакторный дисперсионный анализ

При формировании групп для сравнения в однофакторном дисперсионном анализе используется только одна переменная (фактор).

Пример. Исследуется эффективность работы нового станка по обработке металлов с помощью дисперсионного анализа. Сравнение проводится с работой старого станка, который уже используется в производстве. В данном исследовании фактор – используемый станок. У него два уровня: новый, старый станки.

В дисперсионном анализе фактор может иметь более двух уровней.

Однофакторный дисперсионный анализ с двумя уровнями аналогичен t-критерию. Нулевая гипотеза обычно говорит о равенстве средних двух групп, альтернативная – о различии средних (двусторонний тест) или различии в определенном направлении (односторонний тест).

Основные условия проведения дисперсионного анализа:

1. Зависимая переменная должна быть непрерывной, неограниченной/изменяющейся в широком интервале и представлена интервальными/характеризующими отношения данными; факторы должны быть дихотомическими/категориальными.
2. Каждое значение зависимой переменной не должно зависеть от других ее значений.

Исключения: рассматривается временная зависимость или значения были измерены у объектов, которые объединены в группы (члены одной семьи, учащиеся в одном классе) и это повлияло на зависимую переменную.

3. В каждой группе непрерывная переменная имеет приблизительно нормальное распределение. Нормальность распределения можно проверить, используя гистограмму («на глаз») или статистические тесты на нормальность.
4. Дисперсии изучаемых групп должны быть приблизительно одинаковыми. Проверить похожесть дисперсий можно с помощью теста Левина, в котором нулевая гипотеза гласит, что дисперсия однородна, и если результат теста Левина статистически не значим (при применении критерия $\alpha < 0,05$), то дисперсии достаточно похожи.

Некоторые условия проведения дисперсионного анализа могут нарушаться, например, F-статистика надежна в случае, когда распределение непрерывной переменной отлично от нормального, а размеры групп одинаковы. Одинаковый размер обеспечивает и устойчивость F-статистики к нарушениям однородности дисперсии. А нарушение условия независимости может сильно исказить результаты.

Однофакторный дисперсионный анализ основан на том, что общая сумма квадратов SST получена из двух компонент: межгрупповой суммы квадратов SSB и внутригрупповой суммы квадратов SSW:

$$SST = SSB + SSW$$

Обозначим n_i – число значений в i-ой группе, m - общее число групп (градаций фактора), то общее число значений (наблюдений) вычисляется по формуле: $\sum_{i=1}^{m} n_i = n$.

Среднее значение всех наблюдений (общее среднее наблюдений) вычисляется по формуле:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} x_{ij}}{n}$$

Слайд 54

Найдём общее число квадратов отклонений:

$$SST = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{\bar{x}}\right)^2$$

Общее число степеней свободы:

$$dF_{SST} = n - 1$$

Среднее наблюдений в каждой градации фактора (средние значения наблюдений внутри групп):

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} x_j}{n_i}, i = \overline{1, m}$$

Межгрупповая сумма квадратов SSB (сумма квадратов отклонений, объяснённая влиянием фактора):

$$SSB = \sum_{i=1}^{m} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

Число степеней свободы объяснённой дисперсии:

$$dF_{SSB} = m - 1$$

Внутригрупповая сумма квадратов SSW (необъяснённая сумма квадратов отклонений, сумма квадратов отклонений ошибки):

$$SSW = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}_i\right)^2$$

Внутригрупповая сумма квадратов также находится через дисперсии групп:

$$SSW = \sum_{i=1}^{m}(n_i - 1)\, D_i\ ,$$

где D_i – дисперсия группы.

Число степеней свободы необъяснённой дисперсии:

$$dF_{SSW} = n - m$$

Межгрупповой средний квадрат (межгрупповая дисперсия):

$$MSB = \frac{SSB}{dF_{SSB}}$$

Внутригрупповой средний квадрат (внутригрупповая дисперсия):

$$MSW = \frac{SSW}{dF_{SSW}}$$

Провести однофакторный дисперсионный анализ данных статистического комплекса:

- вычислить фактическое отношение Фишера:

$$F = \frac{MSB}{MSW}$$

- сравнить фактическое отношение Фишера с критическим значением Фишера $F(\alpha, dF_{SSB}, dF_{SSW}) = F_{\alpha;dF_{SSB};dF_{SSW}}$.

F-распределение характеризуется двумя значениями числа степеней свободы. Первый индекс – число степеней свободы для числителя (dF_{SSB}), второй – для знаменателя (dF_{SSW}).

При нахождении критического значения F-распределения используют таблицу. В столбце данной таблицы указано число степеней свободы числителя, в строке — число степеней свободы для знаменателя.

Таблица 8.1. F-распределение для уровня значимости 0,05

df2/df1	1	2	3	4	5	6	7	8	9	10
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472

Пример 3.

Группа	Группа	Группа
1	3	5
2	4	6
3	5	7

Сравниваем 3 группы, в каждой из которых по 3 значения.

Нулевая гипотеза: в генеральной совокупности нет значимых различий между средними, все средние трёх групп равны друг другу. Альтернативная гипотеза: хотя бы пара средних значимо различается между собой.

H₀: $\mu_1 = \mu_2 = \mu_3$

H₁: $\mu_1 \neq \mu_2 = \mu_3$ или $\mu_1 = \mu_2 \neq \mu_3$ или $\mu_1 \neq \mu_2 \neq \mu_3$

Вычислим среднее значение всех наблюдений:

$$\bar{\bar{x}} = \frac{1+2+3+3+4+5+5+6+7}{9} = 4$$

Вычислим общую сумму квадратов:

$$SST = (1-4)^2 + (2-4)^2 + (3-4)^2 + (3-4)^2 + (4-4)^2 + (5-4)^2 + \\ +(5-4)^2 + (6-4)^2 + (7-4)^2 = 30$$

Степени свободы для общей суммы квадратов:

$$dF_{SST} = n - 1 = 9 - 1 = 8$$

Вычислим средние значения внутри каждой из групп:

$$\bar{x}_1 = \frac{1+2+3}{3} = 2$$
$$\bar{x}_2 = \frac{3+4+5}{3} = 2$$
$$\bar{x}_3 = \frac{5+6+7}{3} = 2$$

Внутригрупповая сумма квадратов:

$$SSW = (1-2)^2 + (2-2)^2 + (3-2)^2 + (3-2)^2 + (4-2)^2 + (5-2)^2 + \\ +(5-2)^2 + (6-2)^2 + (7-2)^2 = 6$$

Степени свободы для внутригрупповой суммы квадратов:

$$dF_{SSW} = n - m = 9 - 3 = 6$$

Межгрупповая сумма квадратов:

$$SSB = 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$$

Степени свободы для межгрупповой суммы квадратов:

$$dF_{SSB} = m - 1 = 3 - 1 = 2$$

SST =30

SSB=24 SSW=6

Получили, что большая часть общей изменчивости обеспечивается благодаря межгрупповой сумме квадратов, значит группы значительно различаются между собой.

Межгрупповая дисперсия:

$$MS_B = \frac{SSB}{dF_{SSB}} = \frac{24}{2} = 12$$

Внутригрупповая дисперсия:

$$MS_W = \frac{SSW}{dF_{SSW}} = \frac{6}{6} = 1$$

Вычислим F-значение:

$$F = \frac{MS_B}{MS_W} = \frac{12}{1} = 12$$

Критическое значение отношения Фишера:

$F_{0,05;\ 2;6} = 5,1$

Так как фактическое отношение Фишера меньше критического:

$$F = 12 > 5,1 = F_{\ 0,05;\ 2;6}$$

можно сделать вывод, что есть существенные различия между группами.

Пост-хок тесты

После проведения дисперсионного анализа получаем данные о том, значимо ли влияние изучаемого фактора на данные: различаются ли между группами средние значения зависимой переменной. Однако результаты анализа не дают ответа на вопрос: благодаря каким различиям это влияние оказалось значимым?

Для решения данной задачи предназначены пост-хок тесты.

Свойства пост-хок тестов

- post-hoc тесты применяются когда влияние фактора значимо;
- тесты делают поправку для снижения вероятности ошибки I рода;
- они учитывают величину различий между средними значениями и количество сравниваемых между собой пар;

- тесты отличаются по степени консервативности (разумный компромисс – пост-хок тест Тьюки).

Пост-хок тест Тьюки

- строго контролирует значимость критерия α (0.05)
- одновременно проверяет все парные гипотезы;
- чувствителен к неравенству дисперсий;
- если размер групп имеет сильные различия, работает плохо.

Двухфакторный дисперсионный анализ без повторений

Двухфакторный дисперсионный анализ применяется для проверки возможной зависимости результативного признака от двух факторов.

Пусть *m* – число градаций первого фактора и k– число градаций второго фактора.

Двухфакторный дисперсионный анализ основан на том, что общая сумма квадратов SST получена из трёх компонент: объяснённой влиянием фактора *A* суммы квадратов отклонений SSB_A, объяснённой влиянием фактора *B* суммы квадратов отклонений SSB_B и необъяснённой суммы квадратов отклонений (суммы квадратов отклонений ошибки):

$$SST = SSB_A + SSB_B + SSW$$

На слайде формулы, по которым вычисляются

Среднее значение всех наблюдений (общее среднее наблюдений)

Общее число квадратов отклонений:

Общее число степеней свободы:

Среднее наблюдений в каждой градации первого фактора:

Среднее наблюдений в каждой градации второго фактора:

Межгрупповая сумма квадратов SSB_A (сумма квадратов отклонений, объяснённая влиянием первого фактора):

Число степеней свободы дисперсии, объяснённой влиянием первого фактора:

Межгрупповая сумма квадратов SSB (сумма квадратов отклонений, объяснённая влиянием второго фактора):

Число степеней свободы дисперсии, объяснённой влиянием второго фактора:

Необъяснённая сумма квадратов отклонений (сумма квадратов отклонений ошибки):

Число степеней свободы необъяснённой дисперсии:

Дисперсия, объяснённая влиянием первого фактора:

Дисперсия, объяснённая влиянием второго фактора:

Необъяснённая дисперсия (дисперсия ошибки):

В случае независимости факторов, выдвигаются две нулевые и две альтернативные гипотезы:

- для первого фактора:

Нулевая гипотеза H_0: $\mu_{1A} = \mu_{2A} = ... = \mu_{mA}$,

Альтернативная гипотеза H_1: не все μ_{iA} равны;

- для второго фактора:

Нулевая гипотеза H_0: $\mu_{1B} = \mu_{2B} = ... = \mu_{kB}$,

Альтернативная гипотеза H_1: не все μ_{jB} равны.

Провести двухфакторный дисперсионный анализ данных статистического комплекса:

- вычислить фактическое отношение Фишера для определения влияния первого фактора:

$$F = \frac{MS_A}{MS_W}$$

- сравнить фактическое отношение Фишера с критическим отношением Фишера $F(\alpha, dF_{SSB_A}, dF_{SSW})$
- вычислить фактическое отношение Фишера для определения влияния второго фактора:

$$F = \frac{MS_B}{MS_W}$$

- сравнить фактическое отношение Фишера с критическим отношением Фишера $F(\alpha, dF_{SSB_B}, dF_{SSW})$

Если фактическое отношение Фишера больше критического отношения Фишера, то следует отклонить нулевую гипотезу с уровнем значимости α, что означает: фактор существенно влияет на данные.

Есть ещё **двухфакторный дисперсионный анализ с повторениями**, но мы рассматривать подробно его не будем.

Инструменты статистического анализа данных

Пакеты стат-анализа данных обычно представляют собой десктопные приложения, в которых вычисления происходят локально. Данные загружаются в виде электронных таблиц. Как правило, есть несложный визуальный ETL, как в **Tableau**. Есть встроенный язык программирования для автоматизации действий.

Плюсы:

- Очень богатые возможности для статистического анализа. Справка этих пакетов успешно конкурирует с учебниками по прикладному анализу данных. Сами статистические функции тщательно протестированы, в отличие от общедоступных статистических калькуляторов в интернете.
- Хорошие графические возможности.
- Внимание к деталям, что важно для научных исследований.
- С данными можно работать офлайн

Минусы:

- Высокий порог входа. Вы должны понимать, что делать, какой именно статистический критерий использовать. Обязательно требуются базовые знания математической статистики.
- Коммерческие продукты стоят дорого.

Продуктивность выполняемой работы тесно связана с используемыми инструментами. Программные пакеты, применяемые для статистического анализа, следует относить к математическим программам.

Как правило, первые шаги в статистике молодые ученые делают в табличных процессорах, причем подавляющее большинство использует MS Excel. Второй по популярности табличный процессор на сегодняшний день - Calc из офисного пакета OpenOffice.org. Полноценная статобработка больших данных в данных процессорах невозможна.

Язык программирования R имеет широкие возможности для статистической обработки данных, в том числе и для работы с графикой, а оконный интерфейс можно установить как дополнительное приложение. Следует помнить, что все промежуточные данные при работе с этим языком, хранятся не во временных файлах, а непосредственно в оперативной памяти. Эту особенность необходимо иметь в виду при обработке очень больших объемов информации: R будет использовать значительную часть оперативной памяти компьютера.

Система SAS (компания SAS Institute). Область применения SAS –научные исследования, бизнес аналитика и т. д.

Система состоит из модулей, каждый из которых выполняет определенный круг задач. Наиболее часто при статобработке используются модули BASE и STAT. В системе SAS реализован собственный язык программирования, который по своему синтаксису ближе к бэйсику и не похож на R. Система позволяет загружать данные из внешних файлов или же вводить их непосредственно в окно терминала. Работая с использованием SAS можно проводить статистическую обработку данных разного уровня сложности, в соответствии с поставленными задачами. Взаимодействие с программой возможно как в консольном режиме, так и через графический интерфейс, который представляет собой графическую оболочку для упрощенного ввода команд языка программирования SAS.

Stata (корпорация StataCorp). Приложение может работать на операционных системах семейства Windows, в MasOS и Linux. Ввод данных здесь возможен как путем загрузки из внешних файлов, так и с использованием встроенного табличного редактора, который довольно прост, но позволяет выполнять все необходимые манипуляции с таблицами.

Statistica или **SPSS Statistics**. Обе программы являются настоящими «монстрами» по своим вычислительным возможностям. Statistica разрабатывается компанией StatSoft. На сегодняшний день последней версией является Statistica 9. Программа SPSS относительно недавно стала принадлежать компании IBM и сменила название на PASW (Predictive Analytics SoftWare) Statistics. Обе программы снабжены великолепным графическим интерфейсом, а также имеют встроенный язык программирования и возможность интеграции с языком статистических вычислений R.

Почти безграничные возможности в статобработке, предоставляемые данными инструментами, требуют от компьютера больших ресурсов. Так, для работы SPSS необходимо не менее 1 Гб оперативной памяти. Операционные системы, в которых можно запускать SPSS: Windows, MacOS и Linux. Statistica же разработана только под Windows, что несколько уменьшает число ее пользователей.

В программах есть все наиболее востребованные статистические методы: частотный анализ, расчет статистических характеристик, таблиц сопряженности, корреляций, построения графиков, t-тесты и большое количество непараметрических критериев, многомерный линейный регрессионный анализ, дискриминантный анализ, факторный анализ, кластерный анализ, дисперсионный анализ, анализ надежности, многомерное шкалирование и ряд других. Вызов этих статистических процедур делается с помощью выбора из меню соответствующих окон и внесения в них необходимых настроек. Все типы анализа разбиты по группам, что помогает быстро ориентироваться в интерфейсе приложений.

Системы STATISTICA и SPSS обладают широкими графическими возможностями. Они включают в себя большое количество разнообразных категорий и

типов графиков, в том числе научные, деловые, трехмерные и двухмерные графики в различных системах координат, специализированные статистические графики — гистограммы, матричные, категоризованные графики и др.

Аналитическая платформа **Polymatica** (рус. Полиматика) от компании Полиматика Рус предназначена для аналитики больших массивов данных в любой предметной области. Высокая скорость взаимодействия обеспечивается за счёт технологий In-Memory и GPU, а также собственной технологии Мультисфер для хранения и сжатия данных.

IQPLATFORM – это цифровая аналитическая платформа, позволяет выполнять продвинутую аналитику на базе больших объёмов информации, синтез новых знаний и мониторинг и контроль информационных объектов. Программный продукт IQPLATFORM (рус. АйКьюПлатформ) от компании Айкумен ИБС (англ. IQMen - Business Intelligence) предназначен для сбора, обработки, хранения и глубинного анализа больших объёмов структурированных и неструктурированных данных из различных типов источников. Программное обеспечение IQPLATFORM первоочерёдно ориентировано на средние и крупные предприятия кредитно-финансовой сферы, промышленности и ТЭК, а также на организации государственного сектора и некоммерческие организации.

M-Brain Intelligence Plaza – это ИТ-платформа для управления потоками информации о рынках и конкурентах для отделов аналитики, продаж, маркетинга, менеджмента. Хранение в облаке, структурирование и внутрикорпоративная рассылка информации по темам, как: отрасли, компании, и др. Программный продукт M–Brain Intelligence Plaza (рус. М-Брэйн Интеллидженс Плаза) от компании M–Brain предназначен для хранения в едином месте, интеллектуального структурирования и распределения деловой информации из различных источников компании. В системе Intelligence Plaza может обрабатываться любой контент, включая внутрикорпоративные данные от разных поставщиков или из внешних источников, контент о бизнес-среде по любым отраслям, в форме кратких деловых сводок, аннотаций и резюме, создаваемых аналитиками M-Brain.

Anaconda – это платформа управления пакетами приложений анализа данных (для языков Python и R) с открытым исходным кодом. Система позволяет специалистам по обработке данных быстро разворачивать проекты машинного обучения, предоставляя необходимую информацию для лиц, принимающих решения.

Программный продукт Anaconda (рус. Анаконда) от одноимённой компании предназначен для управления приложениями анализа данных, основанных на моделях машинного обучения, современных алгоритмах интеллектуального анализа данных и статистических методах анализа. Плафтформа подходит для работы как студентов, так и опытных специалистов (имеются разные варианты инсталляции). Продукт подходит также для предприятий, которым требуется обеспечить полный жизненный цикл машинного обучения и платформу для поддержки искусственного интеллекта.

TIBCO Data Science – это комплексная аналитическая платформа, позволяющая применять полный комплекс современных аналитических методов над деловыми данными компании.

Программный продукт TIBCO Data Science (рус. ТИБКО Дата Сайнс) от компании TIBCO предназначен для анализа деловых данных в компаниях различных размеров и организациях. Данное программное обеспечение позволяет организациям внедрять подходы управления на основе анализа данных по всей организации, предоставляя

гибкие возможности разработки и развертывания аналитических моделей.

Как основной инструмент в программе Data Science используются распределенные аналитические сервисы. Эти сервисы используют конвейеры, созданные в системе, и позволяют перенести вычисления в системы потокового анализа больших данных, такие как Spark. Кроме того, распределенные сервисы обеспечивают управление проектами, совместную работу, планирование, управление моделями и управление ими.

Аналитическое программное обеспечение TIBCO Data Science – это единая платформа, объединяющая и усиливающая возможности многих других программных продуктов TIBCO: TIBCO Statistica, TIBCO Spotfire (ранее Alpine Data) и компании TIBCO Enterprise Runtime for R (TERR).

Есть ещё Minitab, MatLab, Octave, GenStat, JMP, Analyse-it, отечественная разработка STADIA и множество других программ.