ПРАКТИЧЕСКАЯ ЧАСТЬ 4 МОДУЛЯ КУРСА «МАТЕМАТИЧЕСКАЯ СТАТИСТИКА»

Статистические тесты

1. Основные понятия

Статистические тесты используются для проверки соответствия между гипотезой и данными.

Прежде чем переходить к самим тестам, рассмотрим подробнее понятие «гипотезы» на примере следующей задачи.

Есть группа людей, которые принимали новые таблетки для улучшения иммунной системы. Прежде чем проводить тест, необходимо выдвинуть гипотезу. Результаты теста покажут, подтверждается она или нет. Статистические гипотезы обычно делятся на две: нулевую и альтернативную.

Нулевая гипотеза (H_0) – утверждение о параметре генеральной совокупности (параметрах генеральных совокупностей) или распределении, которое необходимо проверить. Как правило нулевая гипотеза утверждает, что между переменными нет статистически значимой разницы или связи. В нашем случае нулевая гипотеза звучит так: «После принятия таблеток иммунитет в выборке не изменился».

Альтернативная гипотеза (H_A, H_1) – утверждение, противоположное нулевой гипотезе. То есть если нулевая гипотеза отвергает статистическую значимость, то альтернативная, наоборот, утверждает, что между переменными существует статистически значимая разница или связь. Важно отметить, что эта гипотеза выдвигается, но не проверяется в ходе статистических тестов. Для нашего примера альтернативная гипотеза будет такой: «После принятия таблеток иммунитет в выборке значимо изменился».

При проведении статистического теста аналитик обычно начинает с предположения, что нулевая гипотеза верна, и пытается опровергнуть её. Если полученные данные не согласуются с нулевой гипотезой, аналитик принимает альтернативную гипотезу. При этом даже когда результат подтверждает нулевую гипотезу, это не обязательно означает, что она верна, — всегда существует вероятность ошибки.

	H_0 верна	H_0 неверна
H_0 отвергается	Ошибка I рода	+
H_0 не отвергается	+	Ошибка II рода

Для описания ошибки I рода существует специальная величина – уровень значимости (α) – вероятность отвергнуть верную нулевую гипотезу.

Нулевая гипотеза всегда проверяется на определенном уровне значимости. Например, если мы проверяем нулевую гипотезу на уровне значимости 5%, это означает, что если мы будем проводить аналогичные исследования 100 раз и проверять на основе имеющихся данных интересующую нас нулевую гипотезу, в 5 случаях из 100 мы отвергнем нулевую гипотезу, хотя она будет верной.

Уровень доверия – вероятность не отвергнуть верную нулевую гипотезу, он противоположен уровню значимости. Имеет место следующее соотношение: $\alpha = 1 - \gamma$. То есть проверить нулевую гипотезу на уровне значимости 5% и проверить нулевую гипотезу на уровне доверия 95% – это одно и то же.

Результат теста должен быть статистически значимым — то есть он с высокой вероятностью не случаен и отражает реальную ситуацию. Статистическую значимость в аналитике данных обозначают показателем p_{value}. При проведении статистического тестирования задача аналитика — найти p_{value} и сравнить его с α. Если p_{value} меньше, нулевую гипотезу отклоняют. Если больше — для отклонения нулевой гипотезы нет оснований. Но это не значит, что она верна. Просто нет оснований.

Выбор подходящего статистического теста — непростая задача. Которая зависит от многих свойств исследуемой предметной области:

- Масштабирование данных. Одним из свойств тестов является масштаб данных, который может быть интервальным (числовым), порядковым или номинальным (категориальным).
- Допущения в структуре данных. Существует две группы статистических тестов: параметрические и непараметрические. Параметрические тесты предполагают, что данные подчиняются определённому распределению-шаблону, обычно нормальному распределению, в то время как непараметрические тесты не делают предположений о распределении. Преимущество непараметрических тестов в том, что они более устойчивы к некорректному поведению данных, например к выбросам. Их недостатком является меньшая точность статистической оценки.
- Тип данных. Некоторые тесты выполняют одномерный анализ на основе одной выборки с одной переменной. Другие сравнивают две или более парные (зависимые) или непарные (независимые) выборки. Также существуют статистические тесты, которые выполняют анализ взаимосвязи между несколькими переменными.
- Количество выборок
- Точность. Тест может быть точным (при котором, если нулевая гипотеза верна, то все допущения, сделанные при выводе распределения статистики критерия, выполняются) или асимптотическим (для него часто предполагается, что размер выборки n может расти бесконечно; свойства оценок и тестов затем

оцениваются в пределе $n \to \infty$). Например, для точного теста с уровнем значимости 5% при повторении на многих выборках эта точность сохраняется. А на асимптотическом тесте желаемая частота ошибок I рода поддерживается только приблизительно (т. е. тест может отклонять > 5% случаев), в то время как это приближение может быть сделано настолько близким к 5%, насколько это необходимо, если сделать размер выборки достаточно большим.

Вернемся к нашему примеру с таблетками. Пусть у нас есть гипотеза, что у людей, принявших таблетки, нет побочных эффектов с вероятностью 0.8. Мы опросили 100 человек из выборки и выяснили, что доля респондентов, которые «заметили» побочные эффекты, равна 0.65. Можно ли сразу по таким результатам опроса сделать однозначный вывод, что доля людей с побочными эффектами не равна 0.8 (ведь $0.65 \neq 0.8$)? Нельзя.

Во-первых, оценки параметра (в данном случае доли), полученные по одной выборке, могут отличаться от истинного значения параметра генеральной совокупности. Поэтому из того факта, что доля людей с побочными эффектами равна 0.65, не следует, что доля таких людей по всей выборке обязательно равна 0.65.

Во-вторых, нам неизвестно, какая разница между выборочной долей и долей, заявленной в гипотезе, считается «маленькой», то есть достаточной для того, чтобы не отвергнуть нулевую гипотезу. В нашем примере доля людей с побочными эффектами в выборке равна 0.65, мы можем считать, что 0.65 сильно отличается от 0.8, поэтому нам следует отвергнуть нулевую гипотезу. А что было бы, если бы выборочная доля была бы 0.7? Или 0.75? Сделали бы мы тогда вывод, что доля людей с побочными эффектами по всей совокупности не равна 0.75? Непонятно, потому что неизвестно, что считать сильным отличием, а что просто списывать на неточность оценок, получаемых по выборке.

Для того, чтобы понять, являются ли различия между значением в гипотезе и полученным по выборке, действительно существенными или эти различия – просто следствие того, что оценки по выборке мы получаем с некоторой погрешностью, требуется формальная проверка гипотез. Для разных видов вопросов существуют свои статистические критерии, позволяющие проверять соответствующие им нулевые гипотезы.

Статистический критерий – правило, которое позволяет делать вывод о том, стоит ли на основе имеющихся данных отвергать нулевую гипотезу или нет. Обычно для критерия определяется соответствующая ему статистика – функция от наблюдений, которая имеет свое распределение. Для того чтобы понять, действительно ли разница между значением параметра в гипотезе и значением оценки, полученной по выборке, является существенной, необходимо сравнить два показателя: наблюдаемое значение статистики и критическое значение статистики.

Наблюдаемое значение статистики – значение статистики, которое получается по выборке, на основе имеющихся данных.

Критическое значение – пороговое значение статистики, которое ожидается в случае, если нулевая гипотеза верна. Критическое значение статистики отделяет область типичных значений статистики от критической области – области редких значений статистики при условии, что нулевая гипотеза верна. Область типичных значений – область не-отвержения нулевой гипотезы, критическая область – область отвержения нулевой гипотезы.

Алгоритм проверки гипотез:

1. Сформулировать нулевую гипотезу (H_0).
2. Сформулировать альтернативную гипотезу (H_A).
3. Выбрать критерий, необходимый для проверки нулевой гипотезы.
4. Определить критическое значение статистики и критическую область.
5. Определить наблюдаемое значение статистики.
6. Сравнить наблюдаемое и критическое значения. Если наблюдаемое значение (по модулю) больше критического значения статистики – попадаем в критическую область, следовательно, нулевую гипотезу необходимо отвергнуть.
7. Сделать статистический и содержательный вывод.

Важно всегда указывать уровень значимости (уровень доверия), на котором проверяется гипотеза, так как без этого уточнения выводы о нулевой гипотезе не имеют большого смысла: на одном уровне значимости гипотеза может быть отвергнута, а при выборе другого уровня значимости – нет. Желательно также прописывать, что выводы делаются на имеющихся данных, так как мы можем отвечать только за те результаты, которые получили по той выборке / выборкам, которые у нас есть, а не за «истинность» выводов вообще.

По результатам проверки статистической гипотезы мы никогда не делаем вывод о том, что нулевая гипотеза верна / должна быть принята. Вопрос об истинности нулевой гипотезы – содержательный вопрос, и если он и проверяется статистически, то с помощью более продвинутых методов и в рамках специально продуманного дизайна исследования. Всё, что мы можем решить по итогам проверки: отвергнуть нулевую гипотезу или нет. Как из того, что события не независимы, автоматически не следует их зависимость, так и из того, что нулевая гипотеза не отвергается, не следует, что она принимается.

Пример статистического вывода: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) есть основания/нет оснований отвергнуть нулевую гипотезу в пользу альтернативы.

Пример содержательного вывода: среднее количество людей с побочными эффектами превышает 50%.

С точки зрения охвата объекта исследования, статистический анализ можно разделить на два вида: сплошной и выборочный. Сплошной анализ предполагает изучение генеральной совокупности данных, то есть всего явления во всем его многообразии без распространения выводов на другие элементы, не входящие в анализируемую совокупность. В противовес сплошному придумали выборочное наблюдение. Название метода точно отражает его суть: из генеральной совокупности отбирается и анализируется только часть данных, а выводы распространяют на всю генеральную совокупность. Отбор данных происходит таким образом, чтобы выборка была репрезентативной, то есть, сохранила внутреннюю структуру и закономерности генеральной совокупности.

То, какой анализ мы проводим – практически сплошной или выборочный – существенно влияет на такую величину как дисперсия.

Дисперсия, как и доля или средняя арифметическая, также меняет свое значение от выборки к выборке, но здесь есть интересная особенность. Дисперсия ведь рассчитывается от средней величины, а она в свою очередь, тоже рассчитывается по выборке, то есть является ошибочной. Как же это обстоятельство влияет на саму дисперсию?

Если бы мы знали истинную среднюю величину (по генеральной совокупности), то ошибка дисперсии была бы связана только с нерепрезентативностью, то есть с тем, что данные в выборке оказались бы ближе или дальше от средней, чем в целом по генеральной совокупности. При этом при многократном повторении данные стремились бы к своему реальному расположению относительно средней. Выборочный показатель, который при многократном повторении выборки стремится к своему теоретическому значению, называется несмещенной оценкой. Почему оценкой? Потому что мы не знаем реальное значение показателя (по генеральной совокупности), и с помощью выборочного наблюдения пытаемся его оценить. Оценка показателя – это есть его характеристика, рассчитанная по выборке.

Теперь смотрим внимательно на выборочную среднюю. Выборочная средняя – это несмещенная оценка математического ожидания, так как средняя из выборочных средних стремится к своему теоретическому значению по генеральной совокупности. Где она расположена? Правильно, в центре выборки! Средняя всегда находится в центре значений, по которым рассчитана – на то она и средняя. А раз выборочная средняя находится в центре выборки, то из этого следует, что сумма квадратов расстояний от каждого значения выборки до выборочной средней всегда меньше, чем до любой другой точки, в том числе и до генеральной средней. Это ключевой момент. А раз так, то дисперсия в каждой выборке будет занижена. Средняя из заниженных дисперсий также даст заниженное значение. То есть при многократном повторении эксперимента выборочная дисперсия не будет стремиться к своему истинному значению (как выборочная средняя), а будет смещена относительно истинного значения по генеральной совокупности.

Несмещенность оценки – одна из важных характеристик статистического показателя. Смещенная оценка показателя заранее говорит о тенденции к ошибке. Поэтому показатели стараются оценивать таким образом, чтобы их оценки были несмещенными (как у средней арифметической). Чтобы решить проблему смещенности выборочной дисперсии, в ее расчет вносят корректировку – умножают на n/(n-1), либо сразу при расчете в знаменатель ставят не n, а n-1.

Выборочная смещенная дисперсия:

$$s^2 = \frac{\sum_{i=1}^{n}(X - \bar{X})^2}{n}$$

Выборочная несмещенная дисперсия:

$$s_0^2 = \frac{\sum_{i=1}^{n}(X - \bar{X})^2}{n - 1}$$

Под выборочной дисперсией понимают, как правило, именно несмещенный вариант.

Теперь посмотрим на практическую сторону отличия смещенной и несмещенной дисперсии. Соотношение между выборочной и генеральной дисперсией составляет n/n-1. Несложно догадаться, что с ростом n (объема выборки) данное выражение стремится к 1, то есть разница между значениями выборочной и генеральной дисперсиями уменьшается. При переходе к среднеквадратичном отклонению по выборке (корень из выборочной дисперсии) разница становится еще меньше. Таким образом, эффект смещенной дисперсии проявляется в небольших выборках. В больших выборках можно использовать генеральную дисперсию.

2. z-тест

z-тест основан на бесконечной делимости нормального распределения. Этот тест используется для проверки равенства матожидания выборки нормально распределённых величин некоторому значению. Значение дисперсии должно быть известно.

Пусть у нас имеется выборка объёмом n независимых нормально распределенных величин X_i из генеральной совокупности со стандартным отклонением σ. Выдвинем гипотезу, что среднее значение равно μ. Тогда величина:

$$z = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

будет иметь стандартное нормальное распределение (среднее = 0, отклонение = 1). Сравнивая полученное значение z с квантилями стандартного распределения, можно принимать или отклонять гипотезу с требуемым уровнем значимости.

Задача 1. Директор одной из школ заявил, что учащиеся, обучающиеся в этой школе, более умны, чем в средней школе. При подсчете показателей IQ 50 учащихся среднее значение оказалось равным 110. Средний показатель IQ населения равен 100, а стандартное отклонение - 15. Укажите, является ли заявление директора корректным.

Решение.

Сначала мы определяем нулевую гипотезу и альтернативную гипотезу:

H_0: $\mu_x = 100$

H_A: $\mu_x > 100$

Теперь укажем уровень значимости (в данном случае примем самый частый):

$\alpha = 0.05$

Затем вычислим z-оценку:

$\bar{x}$ (возможное среднее значение) = 110

μ (среднее значение) = 100

σ (стандартное отклонение) = 15

n (размер выборки) = 50

z-оценка: $(110 - 100)/(15/\sqrt{50}) \approx 4.71$

Теперь посмотрим на z-таблицу. Для значения α =0.05 z-оценка для правостороннего теста составляет 1,645. Нужно отметить, что в этом тесте наша область отклонения расположена в крайней правой части распределения. То есть здесь наша нулевая гипотеза заключается в том, что заявленное значение меньше или равно среднему значению популяции.

Здесь 4,71 > 1,645, поэтому мы отвергаем нулевую гипотезу. Если бы статистика z-теста была меньше z-оценки, то мы не отклонили бы нулевую гипотезу.

Статистический вывод: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) есть основания отвергнуть нулевую гипотезу в пользу альтернативы.

Содержательный вывод: уровень IQ учеников в данной школе может превышать средний показатель IQ населения.

Решение задачи в Python

```
import numpy as np
import scipy.stats as stats

sample_mean = 110
population_mean = 100
population_std = 15
sample_size = 50
alpha = 0.05
# Расчет z-оценки
z_score = (sample_mean-population_mean)/(population_std/np.sqrt(50))
print('z-оценка :',z_score)

# Принятие решения на основе критического значения z
# Расчет критической z-оценки
z_critical = stats.norm.ppf(1-alpha) # Возвращает 95%-ный интервал значимости для теста с одним концом для стандартного нормального распределения, для двустороннего теста было бы 1-alpha/2
print('Критическая z-оценка :',z_critical)
if z_score >  z_critical:
   print("Отклонить нулевую гипотезу")
else:
  print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
```

```
# P-Value : вероятность получить значение меньше, чем z-оценка
p_value = 1-stats.norm.cdf(z_score)
print('p-value :',p_value)
if p_value <  alpha:
   print("Отклонить нулевую гипотезу")
else:
  print("Нельзя отклонить нулевую гипотезу")
```

Задача 2. В настоящее время проводятся исследования в рамках кампании по повышению осведомленности о психическом здоровье. Используя данные всех практикующих врачей общей практики по всей стране, было подсчитано количество пациентов, страдающих депрессией, в процентах от общего числа пациентов за последние 15 лет. Среднее значение составило 21,9 %, а стандартное отклонение - 7,5 %. В районе города N были собраны данные о 35 практикующих терапевтах, и за последние пятнадцать лет была зафиксирована доля пациентов с диагнозом депрессии. Среднее значение составило 24,1 %. Отличается ли доля людей, страдающих депрессией, в районе города N от среднего показателя по стране?

Решение

Нулевая гипотеза заключается в том, что доля людей, страдающих от депрессии в районе города N, ничем не отличается от аналогичной доли во всей стране, тогда как альтернативная гипотеза заключается в том, что доля людей, страдающих от депрессии в районе города N, отличается от среднего показателя по стране. У нас здесь есть двухсторонний тест.

```
import numpy as np
import scipy.stats as stats

# H0: mean(x) = 21.9
# HA: mean(x) ≠ 21.9

sample_mean = 24.1
population_mean = 21.9
population_std = 7.5
sample_size = 35
alpha = 0.05
# Расчет z-оценки
z_score = (sample_mean-population_mean)/(population_std/np.sqrt(sample_size))
print('z-оценка :',z_score)

# Принятие решения на основе критического значения z
# Расчет критической z-оценки
z_critical = stats.norm.ppf(1-alpha/2) # Возвращает 95%-ный интервал значимости для двустороннего теста, от 2.5% до 97,5%
print('Критическая z-оценка :',z_critical)
if z_score >  z_critical:
   print("Отклонить нулевую гипотезу")
else:
  print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
# P-Value : вероятность получить значение меньше, чем z-оценка, и больше, чем -z-оценка
p_value = 1-stats.norm.cdf(z_score) + stats.norm.cdf(-z_score)
print('p-value :',p_value)
if p_value <  alpha:
   print("Отклонить нулевую гипотезу")
else:
  print("Нельзя отклонить нулевую гипотезу")
```

Здесь критическая оценка меньше расчетного значения, 1.73 < 1.96, поэтому мы не можем отклонить нулевую гипотезу.

Статистический вывод: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) нет оснований отвергнуть нулевую гипотезу в пользу альтернативы.

Содержательный вывод: доля людей, страдающих от депрессии в районе города N, может никак не отличаться от аналогичной доли во всей стране.

Задача 3. К соревнованиям готовятся две группы студентов: группа А и группа В. Группа А занималась на офлайн-занятиях, в то время как группа В занималась на онлайн-занятиях. После экзамена подсчитывается оценка каждого студента. Теперь мы хотим определить, какие занятия лучше - онлайн или офлайн.

Группа А: Размер выборки = 50, среднее значение = 75, стандартное отклонение = 10.

Группа В: Размер выборки = 60, среднее значение = 80, стандартное отклонение = 12.

Предполагая 5%-ный уровень значимости, выполните z-тест из двух выборок, чтобы определить, есть ли существенная разница между онлайн- и оффлайн-занятия.

Решение.

Часто нам нужно сравнить средние значения двух выборок, и мы используем z-тест для случая, когда мы знаем дисперсию генеральной совокупности. Существует два типа z-тестов:

- Парный (связанный) z-тест — сравнение двух наборов результатов одинакового размера, если они связаны (если вы тестируете одну и ту же группу участников дважды или если две ваши группы похожи):

$$z = \frac{\bar{d} - D}{\sqrt{\frac{\sigma_d^2}{n}}}$$

где d – среднее значение различий между выборками, D – это предполагаемое среднее значение разностей (обычно оно равно нулю), n – это размер выборки и σ^2_d – это совокупная дисперсия различий.

- Независимый (несвязанный) z-тест — когда между группами нет связи (разные независимые группы):

$$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

где $\bar{x}_1$ и $\bar{x}_2$ – это выборочные средние значения, n_1 и n_2 – это размеры выборок и σ_1 и σ_2 – это дисперсии выборок.

В нашем случае обе группы независимы друг от друга, поэтому выбираем независимый z-тест.

Нулевая гипотеза: нет существенной разницы между средним баллом между онлайн- и офлайн-уроками, то есть $\mu_1 - \mu_2 = 0$. Альтернативная гипотеза: существует значительная разница в средних баллах между онлайн-и офлайн-уроками, то есть $\mu_1 - \mu_2 \neq 0$.

```
import numpy as np
import scipy.stats as stats

# Группа А (офлайн-уроки)
n1 = 50
x1 = 75
s1 = 10

# Группа В (онлайн-уроки)
n2 = 60
x2 = 80
s2 = 12

# Нулевая гипотеза: = mu_1-mu_2 = 0
D = 0

alpha = 0.05

# Расчет z-оценки
z_score = ((x1 - x2) - D) / np.sqrt((s1**2 / n1) + (s2**2 / n2))
print('Z-Score:', np.abs(z_score))

# Принятие решения на основе критического значения z
z_critical = stats.norm.ppf(1 - alpha/2)
print('Critical Z-Score:',z_critical)
if np.abs(z_score) > z_critical:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

```
# Принятие решения на основе p-значения
p_value = 2 * (1 - stats.norm.cdf(np.abs(z_score)))
print('P-Value :',p_value)
if p_value < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

Статистический вывод: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) есть основания отвергнуть нулевую гипотезу в пользу альтернативы.

Содержательный вывод: существует существенная разница между онлайн- и оффлайн-занятиями.

3. t-тест

z-тесты — это статистический способ проверки гипотезы, когда выполняется одно из условий:

- известна дисперсия генеральной совокупности σ^2
- неизвестна дисперсия генеральной совокупности, но размер нашей выборки велик, n≥30; в этом случае мы используем выборочную дисперсию в качестве оценки дисперсии генеральной совокупности.

Если у нас размер выборки меньше 30 и не знаем дисперсию генеральной совокупности, тогда мы должны использовать t-тест.

Вот некоторые дополнительные условия для использования этого типа теста:

- Данные должны быть нормально распределены;
- Все точки данных должны быть независимыми;
- Для каждой выборки отклонения должны быть равны.

Фактически t-тест является аналогом z-теста для случая, когда дисперсия или стандартное отклонение выборки неизвестно и должно быть оценено на основании самой выборки.

Рассмотрим пример проверки равенства матожидания нормальной выборки некоторому значению: пусть нам дана выборка X_i нормальных величин объёмом n из некоторой генеральной совокупности, выдвинем и проверим гипотезу о том, что матожидание этой совокупности равно m.

Рассчитаем дисперсию выборки $S = \frac{\sum_{i=1}^{n} X_i^2}{n-1}$. Эта величина будет иметь распределение хи-квадрат (χ^2 – это распределение описывает сумму n квадратов случайных величин X_i, каждая из которых распределена по нормальному закону). Тогда величина $t = \frac{\sum_{i=1}^{n} X_i - m}{n-1} : \frac{S}{\sqrt{n}}$ будет иметь распределение Стьюдента $T_{n-1}(x)$ с n-1 степенью свободы.

Первые два шага для t-теста такие же, как для z-теста, поскольку мы должны определить нулевую и альтернативную гипотезы.

После мы рассчитываем тестовую статистику по приведённой ниже формуле. Как видите, изменения незначительны. Теперь она обозначена t, а дисперсия генеральной совокупности заменена дисперсией выборки s = √S:

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{n}}}$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1}(\sum x_i^2 - n\bar{x}^2)}$$

В конце используем t-таблицы, чтобы найти критические значения при выбранном уровне значимости и правильном количестве степеней свободы n-1, и сравниваем нашу t-статистику. Если наша t-статистика больше критического значения, у нас есть значимый результат на выбранном уровне, то есть вы можете отвергнуть нулевую гипотезу.

Аналогично z-тесту для двух выборок существуют t-тесты для двух выборок:

- Парный (связанный) t-критерий — сравнение двух наборов результатов, которые связаны между собой (например, когда вы дважды тестируете одну и ту же группу участников или две ваши группы похожи друг на друга):

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

где d – это среднее значение разностей между выборками, s – стандартное отклонение разностей. Число степеней свободы для этого случая n-1.

- Независимый (несвязанный) t-критерий — между группами нет связи (разные независимые группы):

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

где $\bar{x}_1$ и $\bar{x}_2$ являются средними значениями из двух выборок, n_1 и n_2 – это размеры выборок, s_1 и s_2 – это выборочные стандартные отклонения. Для этого случая число степеней свободы $n_1 + n_2 - 2$. Для того, чтобы t-критерий объединённой дисперсии был подходящим, вы должны исходить из предположения, что две выборки взяты из одной совокупности и имеют одинаковую дисперсию, что проверяется F-тестом.

В случае же если отклонения выборок не равны, необходимо посчитать t-критерий для приведенный степени свободы ν:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 + 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 + 1}} - 2$$

Задача 1. Психолог хочет проверить, отличаются ли умственные способности людей утром и днём. У него есть группа из десяти студентов. Утром они проходят тест на умственные способности, а днём — другой тест с другими вопросами, но такой же сложности. Результаты тестов приведены в следующей таблице.

Утро	13	20	11	9	7	14	12	13	10	17
День	14	19	13	11	10	17	9	16	10	19

Выполните проверку гипотезы, чтобы выяснить, есть ли разница в работоспособности утром и днём.

Решение

Тип проверки гипотезы, который нам нужно использовать, — это t-критерий для двух выборок, поскольку стандартные отклонения генеральной совокупности неизвестны, а размер выборки невелик. Нам нужно использовать парный t-критерий, поскольку в утренней и дневной выборках участвуют одни и те же участники.

Сначала мы определяем нулевую гипотезу и альтернативную гипотезу:

H_0: $\mu_{утро} = \mu_{день}$

H_A: $\mu_{утро} \neq \mu_{день}$

Это двусторонний тест, так как мы просто хотим узнать, есть ли разница, но не уточняем, лучше ли результаты во второй половине дня или утром.

Теперь укажем уровень значимости (в данном случае примем самый частый):

$\alpha = 0.05$

Затем выполним расчеты:

Утро	13	20	11	9	7	14	12	13	10	17
День	14	19	13	11	10	17	9	16	10	19
d	1	-1	2	2	3	3	-3	3	0	2
<d>	1.2									
s	1.989									
$t_{выборки}$	1.908									
$t_{крит}(n=9)$ $\alpha = 0.05$	2.262									

Статистический вывод: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) нет оснований отвергнуть нулевую гипотезу в пользу альтернативы.

Содержательный вывод: Результаты теста в среднем были немного выше во второй половине дня по сравнению с утренним временем, однако это не подтверждает гипотезу о том, что результаты тестов утром и днём всегда различаются.

```
import numpy as np
import scipy.stats as stats
```

```
x1 = [13, 20, 11, 9, 7, 14, 12, 13, 10, 17]
x2 = [14, 19, 13, 11, 10, 17, 9, 16, 10, 19]
alpha = 0.05
n = len(x1)
x = np.zeros(0)
for i in range (n):
    x = np.append(x, x2[i]-x1[i])
d = x.mean()
S = x.var()*n/(n-1)
t_score = d*np.sqrt(n/S)
print('t-score:', np.abs(t_score))

# Принятие решения на основе критического значения t
t_critical = stats.t.ppf(1 - alpha/2, df = n-1)
print('Critical t-Score:',t_critical)
if np.abs(t_score) > t_critical:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
p_value = (1 - stats.t.cdf(np.abs(t_score), df = n-1)) * 2
print('P-Value :',p_value)
if p_value < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

Или еще проще:

```
import scipy.stats as stats
x1 = [13, 20, 11, 9, 7, 14, 12, 13, 10, 17]
x2 = [14, 19, 13, 11, 10, 17, 9, 16, 10, 19]
alpha = 0.05
t_score, p_value = stats.ttest_rel(x1, x2)
if p_value < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

Задача 2.

Перед вами стоит задача проверить, отличаются ли показатели успеваемости у двух групп школьников. Одна группа состоит из детей из полных семей, а другая — из неполных. Вы хотите узнать, получают ли дети из полных семей более высокие баллы, чем дети из неполных семей. Таблица результатов приведена ниже.

Полные	18	13	12	17	7	11	9	15	16
Неполные	5	12	9	11	15	10	12		

Решение

Тип проверки гипотезы, который нам нужно использовать, — это t-критерий для двух выборок, поскольку стандартные отклонения генеральной совокупности неизвестны, а размер выборки невелик. Нам нужно использовать независимый t-критерий, поскольку в выборках представлены разные ученики.

Сначала мы определяем нулевую гипотезу и альтернативную гипотезу:

H_0: $\mu_1 = \mu_2$

H_A: $\mu_1 > \mu_2$

Это односторонний тест, так как мы указали «направление».

```
import scipy.stats as stats
x1 = [18, 13, 12, 17, 7, 11, 9, 15, 16]
x2 = [5, 12, 9, 11, 15, 10, 12]
alpha = 0.05
```

```
t_score, p_value = stats.ttest_ind(x1, x2)
print(t_score, p_value)
if p_value < alpha: # у нас это не выполняется
   print("Отклонить нулевую гипотезу")
else:
   print("Нельзя отклонить нулевую гипотезу")
```

Статистический вывод: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) нет оснований отвергнуть нулевую гипотезу в пользу альтернативы.

Содержательный вывод: Результаты теста в среднем были немного выше у детей из полных семей по сравнению детьми из неполных семей, однако это не подтверждает гипотезу о том, что результаты тестов зависят от данного аспекта.

Полезные функции для t-теста выборок с равными стандартными отклонениями:

- *scipy.stats.ttest_rel* – t-тест для двух связанных выборок
- *scipy.stats.ttest_ind* – t-тест для двух независимых выборок
- *scipy.stats.ttest_ind_from_stats* - t-тест для определения средних значений двух независимых выборок из описательной статистики (проверка нулевой гипотезы о том, что две независимые выборки имеют одинаковые средние (ожидаемые) значения)
- *scipy.stats.ttest_1samp* – t-тест для среднего значения выборки (проверка нулевой гипотезы о том, что ожидаемое значение (среднее значение) выборки независимых наблюдений равно заданному среднему значению по совокупности).

4. F-тест

При анализе данных обычно сравнивают средние значения выборок при проверке гипотез. Однако возможно наличие двух идентичных средних значений для двух разных выборок/групп, в то время как их дисперсии могут сильно различаться. Статистический тест, используемый для сравнения дисперсий, называется F-тест (или критерий отношения дисперсий). Он сравнивает две дисперсии, чтобы проверить, происходят ли они из одной и той же совокупности.

Формула коэффициента дисперсии выглядит следующим образом:

$$F = \frac{\text{Оценка большей дисперсии}}{\text{Оценка меньшей дисперсии}}$$

Пусть имеются две независимые выборки x_i и y_i нормально распределенных данных объёмами n_x и n_y соответственно. Выдвинем гипотезу о равенстве дисперсий выборок и проверим её статистически.

Рассчитаем величину $F = \frac{n_y \sum_{i=1}^{n_x} x_i^2}{n_x \sum_{i=1}^{n_y} y_i^2}$. Суммы квадратов нормальных величин имеют распределение хи-квадрат, в свою очередь величина F будет иметь распределение Фишера со степенями свободы (n_x-1, n_y-1). Как и для t-критерия, существует таблица F-распределения. Она организована в соответствии со степенями свободы двух оценок дисперсии.

Мы сравниваем вычисленную F-статистику с критическим значением из таблицы значимости для F-распределения. Если статистика больше критического значения, то мы делаем вывод, что недостаточно доказательств того, что две выборки взяты из одной и той же совокупности значений. Следовательно, мы принимаем гипотезу о том, что дисперсии выборок значительно отличаются. Если статистика меньше критического значения, то мы принимаем нулевую гипотезу о том, что две дисперсии выборок равны.

F-критерий — это односторонний критерий, поскольку он определяет, больше ли числитель знаменателя.

Задача 1. В настоящее время проводится клиническое исследование, чтобы выяснить, насколько эффективны современные методы лечения для снижения уровня тревожности у людей, страдающих ОКР. Двум группам по 8 пациентов с ОКР назначаются два различных метода лечения. Одной группе назначается когнитивно-поведенческая терапия (КПТ), в то время как другая группа получает курс антидепрессантов. Перед началом лечения баллы пациентов по шкале оценки тревожности Гамильтона (HAM-A) были занесены в таблицу ниже. После восьми недель лечения их баллы HAM-A регистрируются снова, а затем вычисляется разница между баллами до и после лечения, чтобы увидеть, насколько эффективны методы лечения. Один из клинических статистиков из группы, проводившей клиническое исследование, обеспокоен тем, что у двух разных групп изначально могут быть разные показатели HAM-A, и считает, что это может повлиять на результаты исследования. Проведите тест, чтобы понять, стоит ли клиническому статистику беспокоиться.

CBT	16	22	29	28	17	19	20	19
SSRI	23	19	25	26	25	24	18	23

Решение

Наши гипотезы следующие:

H_0: дисперсии выборок существенно не отличаются.

H_A: дисперсии выборок существенно отличаются.

Сначала нам нужно вычислить различия между двумя группами:

$$S = \frac{\sum_{i=1}^{n} X_i^2}{n-1}$$

Получим: $S_1 = 23.357$, $S_2 = 8.411$.

Теперь вычислим F-статистику: $F = 2.777$.

При уровне значимости 0.05 $F_{крит}(7, 7) = 3.79$, F-статистика меньше; это означает, что дисперсии выборок незначительно отличаются друг от друга. То есть мы принимаем нулевую гипотезу.

В рамках психологических исследований мы бы сообщали о наших результатах следующим образом:

«Средний балл для группы SSRI немного выше, чем в группе когнитивно-поведенческой терапии. Однако мы обнаружили, что дисперсии двух групп не являются статистически значимыми».

```
import scipy.stats as stats
import numpy as np
x1 = np.array([16, 22, 29, 28, 17, 19, 20, 19])
x2 = np.array([23, 19, 25, 26, 25, 24, 18, 23])
alpha = 0.05
n = len(x1)
S1 = x1.var()*n/(n-1)
S2 = x2.var()*n/(n-1)
F_score = S1/S2
print('F-score:', np.abs(F_score))

# Принятие решения на основе критического значения F
F_critical = stats.f.ppf(1 - alpha, dfn = n-1, dfd = n-1)
print('Critical F-Score:', F_critical)
if np.abs(F_score) > F_critical:
   print("Отклонить нулевую гипотезу")
else:
   print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
p_value = 1 - stats.f.cdf(np.abs(F_score), dfn = n-1, dfd = n-1)
print('P-Value :',p_value)
if p_value < alpha:
   print("Отклонить нулевую гипотезу")
else:
   print("Нельзя отклонить нулевую гипотезу")
```

Задача 2. Имеются две независимые выборки измерений точности размеров деталей обуви, полученных на двух прессах X и У. Требуется проверить гипотезу, что точность обработки одинакова (то есть что дисперсии выборок равны).

Пресс 1- X	6,63	6,64	4,56	9,73	11,56	14,99	14,77	6,33	4,61	5,73
Пресс 2 - У	5,05	5,84	5,74	6,44	7,09	9,82	9,11	7,50	2,89	6,55

Решение

```
import scipy.stats as stats
import numpy as np
x1 = np.array([6.63, 6.64, 4.56, 9.73, 11.56, 14.99, 14.77, 6.33, 4.61, 5.73])
x2 = np.array([5.05, 5.84, 5.74, 6.44, 7.09, 9.82, 9.11, 7.50, 2.89, 6.55])
alpha = 0.05
n = len(x1)
S1 = x1.var()*n/(n-1)
S2 = x2.var()*n/(n-1)
F_score = S1/S2
print('F-score:', np.abs(F_score))
```

```
# Принятие решения на основе критического значения F
F_critical = stats.f.ppf(1 - alpha, dfn = n-1, dfd = n-1)
print('Critical F-Score:', F_critical)
if np.abs(F_score) > F_critical:
   print("Отклонить нулевую гипотезу")
else:
   print("Нельзя отклонить нулевую гипотезу")
```

5. χ^2-тест

Критерий хи-квадрат (χ^2-критерий) — это статистический критерий проверки гипотезы, используемый при анализе таблиц сопряжённости при больших размерах выборки. В отличие от предыдущих статистик, χ^2-критерий является непараметрическим (нет изначального предположения, что данные распределены нормально) и может работать даже для категориальных признаков.

Существует два важных применения критерия хи-квадрат: тест на соответствие и тест на наличие связи.

- Тесты на соответствие. Мы используем χ^2-тесты для проверки соответствия данных гипотетической модели, например, соответствует ли количество посетителей аттракциона на ярмарке распределению Пуассона.

Сначала нужно определить нулевую и альтернативную гипотезы.

H_0: Наши данные распределены равномерно.

H_A:Наши данные не распределены равномерно.

Равномерное распределение можно заменить любым другим распределением вероятностей.

После сбора данных мы вычисляем значение хи-квадрат:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

где O = наблюдаемые частоты, E = ожидаемые частоты.

Затем мы сравниваем эту статистику с χ^2-таблицей соответствующих степеней свободы ν = (количеству категорий) минус (количество оцениваемых параметров) минус 1. Если значение хи-квадрат не превышает критического значения на определенном уровне значимости, тогда нулевая гипотеза будет принята, то есть данные соответствуют гипотетической модели. Если она превышает критическое значение, нулевая гипотеза должна быть отвергнута.

Задача 1. Количество несчастных случаев в день на фабрике регистрировалось в течение трёх месяцев; результаты приведены в таблице ниже.

Количество в день	0	1	2	3	4	>=5
Частота	44	33	10	4	1	0

Предложите распределение, которое могло бы соответствовать этим данным, и проверьте, подходит ли оно.

Решение.

Поскольку мы рассматриваем количество несчастных случаев за определённый промежуток времени (день) и нет фиксированного ограничения на количество несчастных случаев, которые могут произойти, а также сами по себе несчастные случаи – это редкие дискретные события, то подходящим распределением будет распределение Пуассона.

H_0: Наши данные распределены по Пуассону.

H_A:Наши данные не распределены по Пуассону.

Чтобы вычислить нашу тестовую статистику, нам нужно вычислить ожидаемые значения/частоты на основе распределения Пуассона. Мы используем формулу: $P(x = r) = \frac{\lambda^r e^{-\lambda}}{r!}$ – чтобы найти ожидаемые вероятности, а затем умножить их на общий размер выборки (92), чтобы получить соответствующие ожидаемые частоты. Прежде чем использовать эту формулу, нам нужно оценить λ. Для распределения Пуассона, λ равно среднему значению.

λ = (0*44 + 1*33 + 2*10 + 3*4 + 4*1 + 5*0) / 92 = 0.75

Теперь у нас есть λ, и мы можем рассчитать ожидаемые вероятности. Причем для 5 или более категорий мы можем просто сложить все остальные вероятности и вычесть из 1, поскольку все распределение вероятностей должно быть равно 1.

После мы можем рассчитать нашу тестовую статистику.

Количество в день	0	1	2	3	4	>=5
Частота (O)	44	33	10	4	1	0
P	0.47237	0.35427	0.13285	0.03321	0.00623	0.00106

Ожидаемая частота (E)	43.45804	32.59284	12.2222	3.05532	0.57316	0.09752
$(O - E)^2 / E$	0.00676	0.00509	0.40403	0.29209	0.31787	0.09752
χ^2	1.12336					

Степени свободы равны (количеству категорий) минус (количество оцениваемых параметров) минус 1 = 6 – 1 – 1 = 4.

Критическое значение χ^2 из таблицы при $\nu = 4$ при $\alpha = 0.05$ равно 9.488. Поскольку 1,12336<9,488, нет доказательств против нулевой гипотезы H_0. Мы не можем ее отвергнуть. Тогда количество несчастных случаев, зарегистрированных за день на фабрике, подчиняется распределению Пуассона.

```
import scipy.stats as stats
import numpy as np
X = np.array([44, 33, 10, 4, 1, 0])
Y = np.array([43.45772286, 32.59329214, 12.22248455, 3.055621138, 0.572928963, 0.097950343])
res = stats.chisquare(X, Y)
chi2_score, p_value = res.statistic, res.pvalue
print(chi2_score, p_value)
nu = X.shape[0]-1-1
alpha = 0.05

# Принятие решения на основе критического значения chi
chi2_critical = stats.chi2.ppf(1 - alpha, df=nu)
if np.abs(chi2_score) > chi2_critical:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
# p_value = 1 - stats.chi2.cdf(np.abs(chi2_score), df = nu)
if p_value < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

Задача 2. В компании работает 500 продавцов, на каждого из которых может поступить жалоба. За последний месяц на 275 продавцов жалоб не поступало, на 150 поступило по одной жалобе, на 50 – по две жалобы, на остальных – три или более жалоб. С помощью критерия хи-квадрат проверьте гипотезу о том, что количество жалоб на продавца есть случайная величина, подчиняющаяся распределению Пуассона со средним значением одна жалоба в месяц. Уровень значимости считать равным 0.05. Округлите значения до второго знака после запятой, даже если после запятой два нуля.

Решение.

```
import scipy.stats as stats
import numpy as np
import math
X = np.array([275, 150, 50, 25])
temp = X/X.sum()
# Предположим, что это Пуассон с lambda=1
l = 1
Y = np.zeros(X.shape[0])
for i in range(X.shape[0]):
    Y[i] = 500 * l**i * math.exp(-l) / math.factorial(i)
print(Y)
```

```
import scipy.stats as stats
import numpy as np
X = np.array([275, 150, 50, 25])
Y = np.array([186, 187, 95, 32])
# Обращаю внимание, что величины из Y не совсем совпадают с теми, которые получаются при вычислении; это особенность функции chisquare (чтобы суммы по обеим выборкам не отличались больше, чем $10^{-8}$ степени), поэтому в итоге формируем числа методом «подгониана» :)
res = stats.chisquare(X, Y)
```

```
chi2_score, p_value = res.statistic, res.pvalue
print(chi2_score, p_value)
nu = X.shape[0]-1-1
alpha = 0.05

# Принятие решения на основе критического значения chi
chi2_critical = stats.chi2.ppf(1 - alpha, df=nu)
if np.abs(chi2_score) > chi2_critical:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
# p_value = 1 - stats.chi2.cdf(np.abs(chi2_score), df = nu)
if p_value < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

- Тест на наличие связи/независимости. Например, есть ли связь между оценками, выставленными студентами преподавателям их курсов, и средними баллами на экзаменах по этим курсам. Здесь используются качественные данные – категории, к которым относятся преподаватели в соответствии с оценками студентов. Нулевая гипотеза состоит в том, что связи нет, и χ^2-тест проверяет эту гипотезу.

Вам нужно будет использовать χ^2-тест как проверку независимости, когда данные представлены в виде таблицы сопряжённости.

Метод очень похож на тест на соответствие требованиям.

Сначала мы формируем наши гипотезы. Независимо от того, что представляют собой категориальные переменные, если вы проводите тестирование, чтобы увидеть, являются ли они независимыми или нет (связанными), гипотезы будут следующими:

H_0: Нет никакой связи между категориальными переменными.

H_A: Есть связь между категориальными переменными.

Затем мы вычисляем тестовую статистику, которая совпадает с тестовой статистикой для проверки соответствия.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Для этого теста нам не нужно беспокоиться о распределении вероятностей, так как мы просто проверяем независимость. Мы можем вычислить ожидаемые частоты с помощью следующей формулы:

$$E = \frac{\text{сумма по строкам} \times \text{сумма по столбцам}}{\text{общий размер выборки}}$$

Как только у нас будет тестовая статистика, мы сможем найти критическое значение. Соответствующие степени свободы для использования определяются следующим образом: ν = (количество строк − 1)×(количество столбцов − 1).

Если тестовая статистика больше критического значения, мы отвергаем H_0 в пользу H_A. Это эквивалентно утверждению, что есть достаточно доказательств того, что категориальные переменные связаны или не являются независимыми. В противном случае мы принимаем H_0.

Задача 3. В следующей таблице приведены данные о количестве дней больничного, которые в прошлом году брали руководители и рядовые сотрудники универмага «Джеймс Льюис».

	Рабочие	Управляющие	Всего
0-10 дней	47	10	57
11-20 дней	35	13	48
21 или более дней	33	27	60
Всего	115	50	165

Существует ли связь между типом сотрудника и количеством дней отпуска по болезни?

Решение.

Наши гипотезы:

H_0: Нет связи между типом сотрудника и количеством дней больничного.

H_A: Есть связь между типом сотрудника и количеством дней больничного.

Нам нужно рассчитать ожидаемые частоты, прежде чем мы сможем рассчитать тестовую статистику.

Наблюдаемые частоты (O)	Ожидаемые частоты (E)	$(O - E)^2 / E$
47	57×115/165 = 39.7273	1.3314
10	57×50/165 = 17.2727	3.0622
35	48×115/165 = 33.4545	0.0714
13	48×50/165 = 14.5455	0.1642
33	60×115/165 = 41.8182	1.8595
27	60×50/165 = 18.1818	4.2768
		$\chi^2 = 10.7655$

Нам нужно сравнить χ^2 с критическим значением для $(3 - 1) \times (2 - 1) = 2$ степеней свободы. На уровне значимости 0.05 критическое значение равно 5.991.

Поскольку 10,765 > 5,991, мы можем сделать вывод, что есть очень убедительные доказательства того, что существует связь между количеством дней больничного и типом сотрудника. Мы принимаем H_A.

```
import scipy.stats as stats
import numpy as np
X = np.array([[47, 10], [35, 13], [33, 27]])
res = stats.chi2_contingency(X)
chi2_score, p_value = res.statistic, res.pvalue
nu = (X.shape[0]-1)*(X.shape[1]-1)
alpha = 0.05

# Принятие решения на основе критического значения chi2
chi2_critical = stats.chi2.ppf(1 - alpha, df=nu)
if np.abs(chi2_score) > chi2_critical:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
# p_value = 1 - stats.chi2.cdf(np.abs(chi2_score), df = nu)
if p_value < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

ANOVA

Дисперсионный анализ (ANalysis Of VAriance, ANOVA) — это набор статистических моделей и связанных с ними процедур оценки (таких как «вариация» внутри и между группами), используемых для анализа различий между средними значениями. Дисперсионный анализ основан на законе общей дисперсии, согласно которому наблюдаемая дисперсия конкретной переменной делится на компоненты, относящиеся к различным источникам вариации. В своей простейшей форме дисперсионный анализ предоставляет статистическую проверку того, равны ли два или более средних значения генеральной совокупности, и, следовательно, обобщает t-критерий для более чем двух средних значений. Другими словами, дисперсионный анализ используется для проверки разницы между двумя или более средними значениями.

Перед тем как приступить к применению дисперсионного анализа, который предназначен для минимизации риска неправильной оценки ошибки I рода в случае множественных сравнений необходимо убедиться в соблюдении ряда условий:

- Количественный непрерывный тип данных, дискретные данные менее желательны.
- Независимые между собой выборки.
- Нормальное распределение признака в статистических совокупностях, из которых извлечены выборки.
- Равенство (скедастичность) дисперсий изучаемого признака в статистических совокупностях, из которых извлечены выборки (проверяется с помощью критерия Levene).
- Независимые наблюдения в каждой из выборок.

Типы ANOVA:

- Однофакторный ANOVA – это метод статистического анализа данных, который используется для определения наличия статистически значимых различий между двумя или более группами по одной независимой переменной. Входными данными для однофакторного ANOVA являются значения зависимой переменной и групповой фактор, на основе которых проводится анализ. Фактор может быть любой номинальной или порядковой переменной, которая разделяет выборку на группы (в простом случае, это может быть пол, возраст, уровень образования и т.д.). Зависимая переменная – это та переменная, которую мы хотим сравнить в различных группах. Однофакторный ANOVA проверяет нулевую гипотезу о том, что среднее значение зависимой переменной одинаково во всех группах.
- Двухфакторный ANOVA – это метод статистического анализа данных, который позволяет определить наличие статистически значимых различий между группами по двум независимым переменным (факторам). Такой подход позволяет оценить влияние каждой независимой переменной на зависимую переменную, а также выявить возможное взаимодействие между факторами. В случае значимых различий, производится дополнительный анализ, чтобы установить, между какими группами существуют различия.
- Многовариантный ANOVA — это статистический метод, который используется для анализа различий между группами (факторами) и влияния различных переменных (факторов) на исследуемую зависимую переменную. Он позволяет выявить, есть ли статистически значимое влияние одного или нескольких факторов на зависимую переменную, и определить, какие из факторов оказывают наибольшее влияние. Многовариантный ANOVA может использоваться для анализа различных типов данных, включая непрерывные, дискретные и категориальные переменные. Он также может рассчитываться для различных уровней взаимодействия между факторами, что позволяет учитывать сложные взаимодействия между переменными. Основная идея многовариантного ANOVA заключается в том, что общее количество изменений в зависимой переменной разделяется на две части: изменения, связанные с факторами, и изменения, которые не могут быть объяснены факторами (остаток). Факторы могут быть любого типа, но обычно они бывают двух типов: факторы, которые могут быть контролируемыми или экспериментальными (например, воздействие на здоровье человека разных типов диет), и факторы, которые являются неконтролируемыми или наблюдаемыми (например, пол, возраст, образование).

Преимущества ANOVA:

- Эффективность многогруппового сравнения: упрощает одновременное сравнение нескольких групп, повышая эффективность, особенно в ситуациях, когда задействовано более двух групп.
- Простота сравнения дисперсий: позволяет легко интерпретировать различия в дисперсиях между группами, способствуя чёткому пониманию наблюдаемых закономерностей.
- Универсальность в различных дисциплинах: демонстрация широкой применимости в различных областях, включая социальные, естественные и технические науки.

Ограничения ANOVA:

- Чувствительность к допущениям: F-критерий очень чувствителен к определённым допущениям, таким как однородность дисперсии и нормальность, которые могут повлиять на точность результатов теста.
- Ограниченный охват при сравнении групп: F-критерий предназначен для сравнения дисперсий между группами, что делает его менее подходящим для анализа за пределами этой конкретной области.

- Проблемы с интерпретацией: F-критерий не позволяет выявить конкретные пары групп с различными дисперсиями. Необходима тщательная интерпретация, и для более детального понимания групповых различий часто требуются дополнительные постфакторные тесты.

1. Однофакторный ANOVA

Формулы однофакторного ANOVA:

- «Объяснимая» дисперсия (вариативность между группами), где $\overline{Y_i}$ – среднее значение в i-ой выборке, n_i – количество наблюдений в i-ой выборке, $\overline{Y}$ – среднее значение по всем выборкам, K – количество выборок, $K-1$ – число степеней свободы «объяснимой» дисперсии:

$$S_b = \frac{1}{K-1}\sum_{i=1}^{K} n_i(\overline{Y_i} - \overline{Y})^2$$

- «Необъяснимая» дисперсия (внутригрупповая изменчивость) где Y_{ij} – j-ое наблюдение в i-ой выборке, N – общий размер выборок $N-K$ – число степеней свободы:

$$S_e = \frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_i})^2$$

- Сумма «объяснимой» и «необъяснимой» дисперсии дает полную дисперсию наблюдений.
- Отношение «объяснимой» дисперсии к «необъяснимой» подчиняется F-распределению:

$$\mathrm{F}(K-1, N-K) = \frac{S_b}{S_e}$$

Если F-статистика < критического значения F:

- Не удается отклонить нулевую гипотезу
- Отклонить альтернативную гипотезу
- Существенных различий между средними значениями выборки нет
- Наблюдаемые различия между средними значениями выборки могут быть разумно вызваны самой случайностью
- Результат не является статистически значимым

Если F-статистика > критического значения F:

- Принять альтернативную гипотезу
- Отклонить нулевую гипотезу
- Между средними значениями по выборке наблюдаются существенные различия
- Наблюдаемые различия между средними значениями выборки не могли быть разумно вызваны самой случайностью
- Результат статистически значим

Обратите внимание, что когда есть только две группы для одностороннего ANOVA, то $F=t^2$, где t – статистика Стьюдента (t-критерий).

Задача 1. В течение 5 недель на трёх группах крыс тестируются три разных вида корма. Цель состоит в том, чтобы проверить разницу в среднем весе (в граммах) крыс за неделю.

Корм 1	8	12	18	8	6	11
Корм 2	6	9	7	10	9	7
Корм 3	11	8	7	11	7	7

Решение.

Примем гипотезы:

H_0: μ1= μ2=μ3

H_A: Средние значения не равны

Вычислим средние значения по выборкам: $X_1 = 10.5$, $X_2 = 8$, $X_3 = 8.5$.

Вычислим общее среднее значение: $X = 9$.

Вычислим объяснимую дисперсию: $S_b = 10.5$

Вычислим необъяснимую дисперсию: $S_e = 8.2$

Вычислим F-критерий: $F \approx 1.28$

Критическое значение F: $F_{крит}(2, 15, \alpha = 0.05) = 3.68$.

Значит, нельзя отклонить нулевую гипотезу и существенных различий между средними значениями выборок нет.

```
import numpy as np
import pandas as pd
import scipy.stats as stats
X1 = np.array([8, 12, 18, 8, 6, 11])
```

```
X2 = np.array([6, 9, 7, 10, 9, 7])
X3 = np.array([11, 8, 7, 11, 7, 7])
df = pd.DataFrame()
alpha = 0.05
df['X1'] = X1
df['X2'] = X2
df['X3'] = X3
X = np.array([df['X1'].mean(), df['X2'].mean(),df['X3'].mean()])
mean_all = X.mean()
df_b = len(df.columns) - 1
S_b = 0
for i in X:
    S_b = S_b + len(df)*(i - mean_all)**2
S_b = S_b/df_b
df_e = len(df.columns)*(len(df)-1)
S_e = 0
for i in df.columns:
    df[i] = (df[i]-df[i].mean())**2
    S_e = S_e + df[i].sum()
S_e = S_e/df_e
F_score = S_b/S_e
print('F-Score:', F_score)

# Принятие решения на основе критического значения F
F_critical = stats.f.ppf(1 - alpha, dfn = df_b, dfd = df_e)
print('Critical F-Score:', F_critical)
if np.abs(F_score) > F_critical:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")

# Принятие решения на основе p-значения
p_value = 1 - stats.f.cdf(np.abs(F_score), dfn = df_b, dfd = df_e)
print('P-Value :',p_value)
if p_value < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

Или с использованием встроенной функции:

```
import numpy as np
import scipy.stats as stats
X1 = np.array([8, 12, 18, 8, 6, 11])
X2 = np.array([6, 9, 7, 10, 9, 7])
X3 = np.array([11, 8, 7, 11, 7, 7])
res = stats.f_oneway(X1, X2, X3)
alpha = 0.05
if res.pvalue < alpha:
    print("Отклонить нулевую гипотезу")
else:
    print("Нельзя отклонить нулевую гипотезу")
```

Задача 2. Имеется сводка по трем экспериментальным группам о воздействии некоторого препарата. Проверьте наличие существенных различий между средними значениями групп.

№	n	mean	sd
1	30	50.26	10.45
2	30	45.32	12.76
3	30	53.67	11.47

Решение.

Вычислим необъяснимую дисперсию:

$$\sigma_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y_i})^2$$

$$S_e = \frac{1}{N - K}\sum_{i=1}^{K}(n_i - 1)\sigma_i^2$$

В нашем случае: $S_e = (10.45^2 + 12.76^2 + 11.47^2)/3 = 134.527$.
Вычислим объяснимую дисперсию: $S_b = ((-0.51)^2 + 4.43^2 + (-3.92)^2) * 30 / 2 = 528.771$.
Вычислим F-критерий: $F \approx 3.93$.
Критическое значение F: $F_{крит}(2, 87, \alpha = 0.05) = 3.1$.
Значит, между средними значениями по выборке наблюдаются существенные различия.

Важно отметить дальнейшие действия после завершения ANOVA:

- Если различия есть, нужно использовать методы множественного сравнения (группы сравнивают попарно) – критерии Тьюки (наиболее распространенный и рекомендуемый, проверяет только парные гипотезы), Ньюмена-Кейлса (наименее строгий), Шеффе (проверяет и парные, и комплексные гипотезы), Даннетта (для сравнения с заранее выбранной контрольной группой).
- Если различий нет, мы НЕ ИМЕЕМ ПРАВА ПРЕДПРИНИМАТЬ ДАЛЬНЕЙШИЙ АНАЛИЗ!

2. Двухфакторный ANOVA

Двухфакторный ANOVA является расширением однофакторного дисперсионного анализа, который изучает влияние двух различных категориальных независимых переменных на одну непрерывную зависимую переменную. Двухфакторный дисперсионный анализ направлен не только на оценку основного эффекта каждой независимой переменной, но и на выявление взаимодействия между ними:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

где $i \in \{1..I\}, I$ – количество данных для первого фактора, $j \in \{1..J\}, J$ – количество данных для второго фактора, $k \in \{1..n_{ij}\}, n_{ij}$ – количество уникальных пар (i,j) (вариантов объединения), Y_{ijk} – наблюдаемое значение, μ – общее среднее по генеральной совокупности, α_i – аддитивный основной эффект (отклонение) уровня i от первого фактора, β_j – аддитивный основной эффект уровня j от второго фактора, γ_{ij} – неаддитивное отклонение, обусловленное совместным влиянием уровней факторов, $\varepsilon_{ijk} \sim N(0, \sigma)$ – независимая ошибка (случайная погрешность).

Задача 1. Выполните анализ влияния различных диет и режимов тренировок на снижение веса. Независимыми переменными являются тип диеты и режим тренировок, а зависимой переменной - потеря веса.

Diet	Workout	WeightLoss
A	Low	3
A	Medium	4
A	High	5
A	Low	3.2
B	Medium	5
B	High	6
B	Low	5.2
B	Medium	6
C	High	5.5
C	Low	4
C	Medium	5.5
C	High	6.2

```
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

data = {
    'Diet': ['A', 'A', 'A', 'A', 'B', 'B', 'B', 'B', 'C', 'C', 'C', 'C'],
    'Workout': ['Low', 'Medium', 'High', 'Low', 'Medium', 'High', 'Low', 'Medium', 'High', 'Low', 'Medium',
'High'],
    'WeightLoss': [3, 4, 5, 3.2, 5, 6, 5.2, 6, 5.5, 4, 5.5, 6.2]
}
```

```
df = pd.DataFrame(data)
model = ols('WeightLoss ~ Diet + Workout + Diet:Workout', data=df).fit()
anova_results = anova_lm(model, typ=2)
print(anova_results)
```

	sum_sq	df	F	PR(>F)
Diet	4.676333	2.0	9.169281	0.052715
Workout	4.521333	2.0	8.865359	0.055050
Diet:Workout	0.603667	4.0	0.591830	0.694337
Residual	0.765000	3.0	NaN	NaN

Столбик PR(>F) – это p-value. Если он меньше, чем α (0.05), то данный фактор оказывает значительное влияние на целевой параметр.

Вы можете настроить таблицу ANOVA, указав различные типы сумм квадратов. В statsmodels, вы можете использовать type параметр для указания типа теста ANOVA:

- Тип 1 : Последовательные суммы квадратов.
- Тип 2 : Частичные суммы квадратов.
- Тип 3 : Предельные суммы квадратов.

Подробнее о stats.anova: https://www.geeksforgeeks.org/how-to-obtain-anova-table-with-statsmodels/