



ФИНАНСОВЫЙ  
УНИВЕРСИТЕТ  
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ

## Введение

В этом разделе курса студенты изучают основные понятия статистики, такие как среднее значение, медиана, мода, дисперсия и стандартное отклонение, а также учатся применять эти понятия для анализа данных в Python.

Кроме того, студенты узнают о различных типах распределений данных, таких как нормальное распределение, распределение Пуассона и распределение Стюдента, и научатся использовать эти знания для анализа данных в Python.

Студенты также узнают о понятии уровня значимости и уровня доверия, которые используются при проверке статистических гипотез, а также о том, как интерпретировать результаты тестов на проверку гипотез.

В целом, раздел "Статистика и тестирование гипотез" в курсе "Python и SQL для анализа данных" помогает студентам научиться применять методы статистического анализа и проверки гипотез в Python, что позволяет им делать более точные выводы на основе данных.

## Статистика. Определение и основные характеристики

Статистика - это наука, которая изучает количественные данные и использует методы анализа данных для извлечения информации и выводов из них.

Распределение данных в статистике обозначает способ, каким данные распределены по различным значениям. Оно описывает, как часто наблюдается каждое значение в наборе данных.

Распределение может быть описано различными параметрами, такими как среднее значение, медиана, мода, дисперсия, стандартное отклонение и другие. Распределение может быть симметричным или асимметричным, унимодальным или мультимодальным.

Примерами распределений данных в статистике являются нормальное распределение, равномерное распределение, биномиальное распределение и другие. Распределение данных имеет важное значение при принятии решений, проведении статистических тестов и анализе данных в различных областях, таких как наука, экономика и социология.

Рассмотрим подробнее.

## Характеристики распределений

1. **Среднее значение (или математическое ожидание) (mean)** - это сумма всех значений выборки, деленная на их количество. Оно характеризует центр распределения данных.

Для расчёта среднего значения в Python можно использовать функцию `mean()` из модуля `numpy`:

```
import numpy as np  
data = [2, 4, 6, 8, 10]
```

```
mean = np.mean(data)
print(mean) # 6.0
```

2. **Медиана (median)** - это значение, которое делит выборку на две равные части: 50% значений находятся выше медианы, а 50% - ниже. Медиана не зависит от выбросов и является более устойчивой мерой центральной тенденции, чем среднее значение.

Для расчета медианы в Python можно использовать функцию `median()` из модуля `numpy`:

```
import numpy as np
data = [2, 4, 6, 8, 10]
median = np.median(data)
print(median) # 6.0
```

3. **Мода (mode)** - это значение, которое встречается наиболее часто в выборке. Мода может быть не единственной, если несколько значений встречаются одинаково часто.

Для расчета моды в Python можно использовать функцию `mode()` из модуля `scipy.stats`:

```
from scipy.stats import mode
data = [2, 4, 6, 8, 10, 10]
mode = mode(data)
print(mode) # ModeResult(mode=array([10]), count=array([2]))
```

4. **Дисперсия (variance)** - это мера разброса значений относительно среднего значения. Она рассчитывается как среднее значение квадратов отклонений каждого значения от среднего значения.

Для расчета дисперсии в Python можно использовать функцию `var()` из модуля `numpy`:

```
import numpy as np
data = [2, 4, 6, 8, 10]
variance = np.var(data)
print(variance) # 8.0
```

5. **Стандартное отклонение (standard deviation)** - это корень из дисперсии. Оно также характеризует разброс значений относительно среднего значения, но выражается в тех же единицах измерения, что и исходные данные.

Для расчета стандартного отклонения в Python можно использовать функцию `std()` из модуля `numpy`:

```
import numpy as np
data = [2, 4, 6, 8, 10]
```



```
std = np.std(data)
print(std) # 2.8284271247461903
```

### **Статистическая гипотеза и тестирование гипотез**

Статистическая гипотеза - это предположение о распределении случайной величины или о связи между двумя или более величинами, которое постулируется для проверки с помощью методов статистического анализа. Статистическая гипотеза обычно формулируется в виде нулевой гипотезы  $H_0$ , которая утверждает отсутствие эффекта или различий между группами, и гипотезы  $H_1$  (альтернативной гипотезы), которая предполагает наличие эффекта или различия между группами.

Статистическая гипотеза может возникнуть из предварительных наблюдений или теоретических соображений, а также может быть сформулирована на основе результатов предыдущих исследований, проведенных другими исследователями.

Процесс реализации статистической гипотезы включает сбор данных, подготовку выборки, выбор статистических тестов, проведение статистического анализа и оценку уровня значимости и/или доверия. Уровень значимости представляет собой вероятность ошибочно отвергнуть нулевую гипотезу, когда она на самом деле верна, а уровень доверия - вероятность правильно принять альтернативную гипотезу, когда она на самом деле верна.

Статистическая гипотеза связана с распределением вероятностей, поскольку для проверки гипотезы часто используются статистические тесты, которые основываются на распределении. Распределение вероятностей может быть нормальным, t-распределением, F-распределением или распределением хи-квадрат. Выбор распределения зависит от предположений о свойствах выборки и смысла статистического теста.

Статистическая гипотеза может быть полезна при исследовании различных феноменов в науке, медицине, экономике и других областях. Она позволяет сделать выводы на основе данных, полученных из выборки, и принимать решения на основе статистических методов анализа.

Статистические гипотезы используются для проверки предположений о совокупностях (например, о процессах, связях между переменными, различиях между группами и т.д.). Они позволяют определить, насколько вероятно, что наблюдаемые различия или зависимости между переменными обусловлены случайными факторами или же являются статистически значимыми.

Проверка статистических гипотез является неотъемлемой частью многих исследований и экспериментов в различных областях науки, от медицины до экономики. Это позволяет сделать выводы на основе статистических данных и принимать важные решения, например, о том, нужно ли внедрять изменения или остановить эксперимент.

Также статистические гипотезы позволяют уточнить или опровергнуть предыдущие исследования, а также сделать выводы о том, какие

дополнительные исследования можно провести, чтобы более точно определить зависимости между переменными или процессами.

Таким образом, статистическая гипотеза – это предположение о виде распределения некоторой случайной величины или о связи между двумя или более случайными величинами. Она может быть как подтверждена, так и опровергнута на основе результатов статистического анализа.

Статистическая гипотеза может возникнуть в результате наблюдений, экспериментов или теоретических размышлений. Когда она формулируется, выдвигается нулевая и альтернативная гипотезы. Нулевая гипотеза заключается в том, что утверждаемый эффект отсутствует, а альтернативная – что эффект существует.

Процесс реализации статистической гипотезы состоит из нескольких этапов. Сначала необходимо определить тип данных и выбрать статистический критерий, соответствующий выбранному типу данных.

Затем нужно собрать данные и проанализировать их. Если нулевая гипотеза отвергается, то мы можем принять альтернативную гипотезу.

Тестирование гипотез – это процесс исследования, в ходе которого проверяется определенное утверждение или предположение (гипотеза) на основе данных и экспериментов.

Процесс тестирования гипотез может быть представлен следующим образом:

1. Формулирование гипотезы – важно определить, какую гипотезу необходимо проверить и с какой целью.

2. Определение уровня значимости – это уровень, который позволяет судить о том, насколько вероятно, что результаты эксперимента будут случайными. Обычно, используются уровни значимости 0.05 и 0.01.

Уровень значимости, с другой стороны, представляет собой вероятность ошибочного отклонения нулевой гипотезы. Если уровень значимости равен 0,05 (или 5%), то это означает, что есть шанс 5% того, что мы случайно отклоним нулевую гипотезу, когда на самом деле она верна.

3. Определение альтернативной гипотезы – это гипотеза, которая будет проверяться в рамках эксперимента. Она формулируется так, чтобы она была обратной к нулевой гипотезе.

4. Определение нулевой гипотезы – это гипотеза, которую мы стремимся опровергнуть. Обычно она высокоуровневая и звучит, как утверждение, что никакой разницы между двумя группами нет.

Уровень значимости – это вероятность того, что нулевая гипотеза будет отклонена, даже если на самом деле она верна. Уровень доверия – это вероятность того, что мы правильно принимаем гипотезу.

Уровень значимости и уровень доверия в общем случае являются взаимодополняющими понятиями и используются для оценки статистической значимости.

Уровень доверия и уровень значимости являются двумя ключевыми понятиями при тестировании гипотез в статистике.

Уровень значимости указывается на уровне 0.05 или 0.01, что означает, что есть шансы 5% или 1%, соответственно, сделать ошибку в отклонении нулевой гипотезы.

Уровень доверия обычно выбирается на уровне 95% или 99% и показывает, с какой вероятностью мы можем быть уверены в правильности принятого решения.

Уровень доверия определяет вероятность правильного выбора статистической гипотезы. То есть, если уровень доверия равен 95%, то это означает, что в 95% случаев мы будем правильно отклонять или не отклонять нулевую гипотезу. Это можно считать вероятностью, что наш выбор является правильным.

Когда мы проводим тестирование гипотез, мы выбираем уровень доверия и уровень значимости на основе наших потребностей. Если мы хотим получить высокую уверенность в нашем выборе гипотезы, то мы можем выбрать уровень доверия 95%. Однако, если мы ошибочно отклоним нулевую гипотезу слишком часто, мы можем выбрать более консервативный уровень значимости, например, 1%.

Кроме того, уровень значимости и уровень доверия связаны между собой. Если мы увеличиваем уровень доверия, то мы должны уменьшать уровень значимости, и наоборот. Это связано с тем, что, увеличивая вероятность правильного выбора гипотезы, мы уменьшаем вероятность ошибочного выбора.

5. Сбор данных и проведение эксперимента - данные собираются с помощью различных методов, например, опросы, анкеты, исследования, наблюдения или использование других научных методов. Важно строго соблюдать план и протокол эксперимента для того, чтобы данные были чистыми и корректными.

6. Анализ результатов - после проведения эксперимента данные обрабатываются и анализируются. Часто используются различные статистические методы, например, t-тест, анализ дисперсии и корреляционный анализ.

7. Принятие решения - на основе анализа результатов принимается решение об имеющихся данных и о том, какие выводы можно сделать о проверяемой гипотезе. Если нулевая гипотеза не отвергнута или подтверждается альтернативная гипотеза, можно сделать вывод о том, что предположение не соответствует реальности.

Существует три основных типа статистических тестов на проверку гипотез в анализе данных: Т-тест, z-тест и хи-квадрат тест.

Рассмотрим каждый из них более подробно и опишем их реализацию на Python.

### 1. Т-тест:

Т-тест - это тест на проверку гипотезы о равенстве средних значений двух выборок. Он используется, когда мы имеем две выборки, и мы

хотим выяснить, есть ли статистически значимая разница между их средними значениями.

Реализация на Python:  
Для реализации t-теста на Python мы можем использовать библиотеку scipy.stats. Например:

```
```python
import scipy.stats as stats
# создаем две выборки
sample1 = [1, 2, 3, 4, 5]
sample2 = [2, 4, 6, 8, 10]
# применяем t-тест
t_statistic, p_value = stats.ttest_ind(sample1, sample2)
# выводим результаты
print("t-statistic:", t_statistic)
print("p-value:", p_value)
```
```

## 2. Z-тест:

Z-тест - это тест на проверку гипотезы о равенстве средних значений двух выборок, который используется, когда мы имеем большой объем выборок (обычно более 30 наблюдений) и известную дисперсию.

Реализация на Python:  
Для реализации z-теста на Python мы также можем использовать библиотеку scipy.stats. Например:

```
```python
import scipy.stats as stats
# создаем две выборки
sample1 = [1, 2, 3, 4, 5]
sample2 = [2, 4, 6, 8, 10]
# применяем z-тест
z_statistic, p_value = stats.zscore(sample1, sample2)
# выводим результаты
print("z-statistic:", z_statistic)
print("p-value:", p_value)
```
```

## 3. Хи-квадрат

тест:

Хи-квадрат тест - это тест на проверку гипотезы о том, что две выборки имеют одинаковое распределение. Он используется, когда мы хотим выяснить, есть ли статистически значимая разница между распределениями двух выборок.

Реализация на Python:  
Для реализации хи-квадрат теста на Python мы можем использовать библиотеку scipy.stats. Например:

```

```python
import                scipy.stats                as                stats
#                    создаем                две                выборки
sample1                =                [1,                2,                3,                4,                5]
sample2                =                [2,                4,                6,                8,                10]
#                    применяем                хи-квадрат                тест
chi_statistic,    p_value    =    stats.chisquare(sample1,    sample2)
#                    выводим                результаты
print("chi-statistic:",    chi_statistic)
print("p-value:",    p_value)
```

```

В целом, эти тесты помогают анализировать данные и выявлять статистически значимые различия между выборками. Реализация этих тестов на Python делает анализ данных более эффективным

## Распределения и их типы

Распределение данных в статистике - это способ, которым данные могут быть описаны и проанализированы с помощью графиков и числовых показателей.

Оно показывает, как часто определенные значения или наборы значений появляются в данных. Распределение данных может быть одномерным, когда оно отображает распределение одной переменной или многомерным, когда он описывает взаимосвязи между двумя или более переменными.

Существует множество разных типов распределений данных, но наиболее распространенные из них - нормальное распределение, равномерное распределение и распределение Пуассона.

Анализ распределения данных позволяет более полно понять характеристики набора данных и может помочь исследователям выявить скрытые закономерности и тенденции в данных, что может быть полезно для различных областей, например, экономики, физики, социологии, биологии и т.д. Инструментами, которые используют для анализа распределения данных и их описания обычно являются статистические показатели, такие как среднее значение, стандартное отклонение, дисперсия, медиана и мода.

Как было замечено ранее, в статистике существует несколько распределений, но конечное число распределений не определено, так как многие распределения могут быть модифицированы и адаптированы под конкретный случай. Вот некоторые из наиболее распространенных распределений, используемых в статистике:

1. Нормальное (Гауссово) распределение
2. Биномиальное распределение
3. Равномерное распределение
4. Экспоненциальное распределение
5. Логнормальное распределение
6. Гамма-распределение



- 7. Бета-распределение
- 8. Пуассоновское распределение
- 9. Геометрическое распределение
- 10. Кривая Лоренца

Это только несколько примеров. Распределений в статистике существует множество, каждый из которых имеет свои особенности и применения в конкретных ситуациях.

Рассмотрим подробнее основные из них:

1. Нормальное распределение (распределение Гаусса) - это самое известное распределение, которое применяется во многих областях исследований, таких как экономика, биология, физика и т.д. Нормальное распределение в большинстве случаев описывает распределение случайных величин. Это распределение, в котором большинство значений находятся вокруг среднего значения, а меньшая часть значений находится дальше от среднего значения.

Погрешность зависит от стандартного отклонения. Данное распределение является основным в статистике, так как многие другие распределения могут быть приближены к нормальному.

Оно распределение имеет форму колокола с высокой вершиной в центре и обратно пропорциональной симметрией вокруг этой вершины.

Это распределение применяется, когда данные признака имеют нормальное распределение, то есть среднее значение и стандартное отклонение характеризуют данные.

Среднее значение такого распределения равно математическому ожиданию ( $\mu$ ), а его погрешность определяется стандартным отклонением ( $\sigma$ ).

Таким образом, нормальное распределение (или распределение Гаусса) является одним из наиболее известных и широко используемых типов распределений. Оно характеризуется симметрией вокруг среднего значения и плотностью вероятности, которая имеет колоколообразную форму. Нормальное распределение используется для моделирования большинства естественных явлений, таких как рост, вес, интеллект, и т.д. Закон больших чисел и центральная предельная теорема позволяют описывать нормально распределенные данные. Среднее значение и стандартное отклонение являются основными мерами распределения, которые могут использоваться для оценки погрешности.

2. Равномерное распределение - это распределение, при котором каждый элемент выборки имеет одинаковую вероятность попадания в любой интервал. Погрешность для этого распределения зависит от длины интервалов и количества элементов в выборке.

В этом распределении вероятность каждого значения одинакова.

Оно применяется, когда вероятности всех значений равны между собой, а погрешность рассчитывается по

формуле  $(b-a)/\sqrt{12}$ , где  $b$  и  $a$  – верхний и нижний пределы равномерного распределения.

Таким образом, равномерное распределение характеризуется постоянной плотностью вероятности в заданном диапазоне значений. Это означает, что вероятность появления любого значения в этом диапазоне равна другому. Равномерное распределение используется, когда необходимо случайным образом выбирать значения в заданном диапазоне. Иногда равномерное распределение обычно используется как "наивное" распределение для моделирования данных, когда нет других особенностей.

3. Распределение Пуассона - это распределение, которое используется для моделирования случайных событий, которые происходят с постоянной частотой.

Данное распределение используется для описания числа независимых случайных событий за фиксированный промежуток времени или пространства. Погрешность для данного распределения зависит от математического ожидания.

Оно описывает вероятность наступления событий с фиксированной частотой в течение заданного времени. Например, число звонков в колл-центре за день, число аварий на дороге за час и т.д. Расчет погрешности такого распределения включает в себя среднеквадратическое отклонение и математическое ожидание, которые связаны формулой погрешности  $= \sqrt{\mu}$ .

Таким образом, пуассоновское распределение используется для моделирования числа успехов в определенном промежутке времени или пространстве, когда интересующие нас явления случайны и независимы друг от друга. Например, распределение Пуассона может использоваться для моделирования количества запросов на сервер во время высокой загрузки.

4. Экспоненциальное распределение - это распределение, используемое для описания времени между двумя последовательными событиями в процессе Пуассона. Погрешность зависит от параметра распределения.

Это распределение вероятностей длительности времени между двумя последовательными событиями в пуассоновском потоке событий. Оно применяется, когда вы ищете вероятность ожидания некоторого события в течение определенного времени. Погрешность рассчитывается по формуле  $(1/\lambda)$ , где  $\lambda$  - интенсивность потока.

Таким образом, экспоненциальное распределение характеризуется тем, что оно обладает ограниченным равномерным хвостом на левой стороне и крутым падением на правой. Оно часто используется для моделирования времени между двумя событиями, таких как интервал между звонками на горячей линии. Среднее время между событиями (интенсивность) уникально определяет показатель распределения, который также используется для оценки погрешности.

5. Распределение Стьюдента - это распределение, которое используется для оценки параметров выборки, когда размер выборки мал.

Распределение Стьюдента используется как альтернатива t-тесту или Z-тесту при вычислении статистических показателей для небольших выборок ( $< 30$ ).

Ключевым параметром этого распределения является количество степеней свободы (df), которое определяется как  $n - 1$ , где  $n$  - размер выборки.

Распределение Стьюдента имеет менее строгое нормальное распределение, что делает его менее точным, но более практичным в случаях, когда у нас есть небольшой объем данных. Распределение Стьюдента имеет более широкие и толстые хвосты, что делает его более чувствительным к выбросам.

Распределение Стьюдента используется для определения доверительных интервалов, критерия Фишера, анализа дисперсии (ANOVA), корреляционного анализа.

6. Гамма-распределение - это распределение, которое широко используется для моделирования времени между пользовательскими событиями в компьютерных системах. Погрешность зависит от параметров распределения.

7. Биномиальное распределение:

Биномиальное распределение используется, когда необходимо знать вероятность нахождения конкретного количества успешных событий в серии испытаний Бернулли (событий с двумя возможными исходами). Биномиальное распределение основано на двух параметрах - количестве испытаний ( $n$ ) и вероятности успешного исхода в каждом испытании ( $p$ ). Оно может использоваться, например, для предсказания количества продаж в определенные дни.

8. Геометрическое распределение - это распределение, которое описывает вероятность того, что первое событие произойдет на определенном шаге, при условии, что все события независимы и имеют одинаковую вероятность.

Геометрическое распределение используется для моделирования времени до первого успеха в определенной серии испытаний Бернулли (событий с двумя возможными исходами). Оно также может быть использовано для описания отказов в процессах эксплуатации, когда первый отказ происходит в результате выполнения большого числа испытаний или обслуживания. Геометрическое распределение характеризуется одним параметром - вероятностью успеха в одном испытании ( $p$ ). Вероятность успеха и вероятность неудачи обычно равны. Среднее значение это  $1/p$ , стандартное отклонение - корень из  $(1-p)/p^2$ .

Применение различных типов распределений имеет преимущества и недостатки - выбор распределения зависит от предполагаемых свойств данных и конкретного вопроса, который требуется исследовать.

Каждое распределение имеет свои особенности, формулы расчета и погрешности. Однако, существуют общие статистические законы, которые могут быть применены к разным распределениям: закон больших чисел, центральная предельная теорема и др. Эти законы помогают предсказать поведение данных и снизить погрешность при их анализе.

Основные законы, связанные с этими распределениями:

1. Закон больших чисел - он говорит, что с увеличением числа испытаний среднее значение выборки будет стремиться к среднему значению генеральной совокупности.

2. Центральная предельная теорема - она утверждает, что при достаточно большом размере выборки распределение выборочных средних близко к нормальному распределению.

3. Закон распределения случайной величины - он говорит о том, что сумма вероятностей всех исходов в случайном эксперименте равна 1.

Расчеты для всех этих распределений могут быть выполнены при помощи статистических программных пакетов, таких как Python, R, SPSS, MATLAB и других.

Тестирование гипотез - это процесс проверки статистических гипотез о том, что некоторые параметры данных имеют определенные значения. В этом разделе курса студенты узнают о различных типах тестов на проверку гипотез, таких как t-тест, z-тест и хи-квадрат тест, и научатся применять эти тесты в Python.

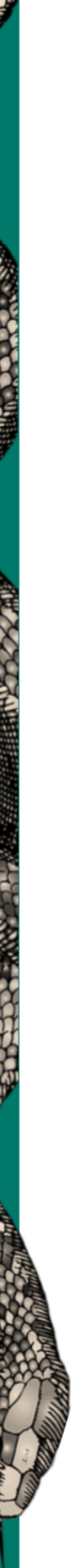
Уровень значимости - это вероятность ошибки первого рода, то есть отвержения верной гипотезы.

Уровень доверия - это вероятность того, что результаты теста правильные.



### Тест по Разделу 3

1. Что такое статистика?
  - a. **Статистика - наука, изучающая количественные данные и использующая методы анализа данных для извлечения информации и выводов из них.**
  - b. Статистика – это наука, изучающая уровни значимости и уровни доверия, которые используются при проверке статистических гипотез
  - c. Статистика - наука, изучающая только качественные данные и **не** использующая методы анализа данных для извлечения информации и выводов из них.
  - d. Все варианты верны
2. Что такое среднее значение?
  - a. **Среднее значение - это сумма всех значений в выборке, деленная на количество наблюдений.**
  - b. Среднее значение - это произведение всех значений в выборке, деленное на количество наблюдений.
  - c. Среднее значение - это разность значений в выборке, деленное на количество наблюдений.
  - d. Нет верного ответа
3. Что такое медиана?
  - a. **Медиана - это значение, которое разделяет выборку на две равные части.**
  - b. Медиана - это сумма всех значений в выборке, деленная на количество наблюдений.
  - c. Медиана - это значение, которое встречается наиболее часто в выборке.
  - d. Все ответы верны.
4. Что такое мода?
  - a. **Мода - это значение, которое встречается наиболее часто в выборке.**
  - b. Мода - это сумма всех значений в выборке, деленная на количество наблюдений.
  - c. Мода - это произведение всех значений в выборке, деленная на количество наблюдений.
  - d. Нет правильного ответа
5. Что такое дисперсия?
  - a. **Дисперсия - это мера разброса данных относительно их среднего значения.**
  - b. Дисперсия - это сумма всех значений в выборке, деленная на количество наблюдений.
  - c. Дисперсия - это значение, которое разделяет выборку на две равные части.
  - d. Нет правильного ответа
6. Что такое стандартное отклонение?

- 
- a. **Стандартное отклонение - это квадратный корень из дисперсии.**
  - b. Стандартное отклонение - это квадрат дисперсии.
  - c. Стандартное отклонение - это сумма всех значений в выборке, деленная на количество наблюдений.
  - d. Нет правильного ответа
7. Что такое нормальное распределение?
- a. **Нормальное распределение - это распределение, в котором большинство значений находятся вокруг среднего значения, а меньшая часть значений находится дальше от среднего значения.**
  - b. Нормальное распределение - это распределение, которое используется для моделирования случайных событий, которые происходят с постоянной частотой.
  - c. Нормальное распределение - это распределение, которое используется для оценки параметров выборки, когда размер выборки мал.
  - d. Нет правильного ответа
8. Что такое распределение Пуассона?
- a. **Распределение Пуассона - это распределение, которое используется для моделирования случайных событий, которые происходят с постоянной частотой.**
  - b. Распределение Пуассона - это распределение, в котором большинство значений находятся вокруг среднего значения, а меньшая часть значений находится дальше от среднего значения.
  - c. Распределение Пуассона - это распределение, которое используется для оценки параметров выборки, когда размер выборки мал.
  - d. Нет правильного ответа
9. Что такое распределение Стьюдента?
- a. Распределение Стьюдента - это распределение, которое используется для моделирования случайных событий, которые происходят с постоянной частотой.
  - b. Распределение Стьюдента - это распределение, в котором большинство значений находятся вокруг среднего значения, а меньшая часть значений находится дальше от среднего значения.
  - c. **Распределение Стьюдента - это распределение, которое используется для оценки параметров выборки, когда размер выборки мал.**
  - d. Нет правильного ответа
10. Что такое тестирование гипотез?

- a. **Тестирование гипотез - это процесс проверки статистических гипотез о том, что некоторые параметры данных имеют определенные значения.**
- b. Тестирование гипотез - это процесс проверки статистических гипотез о том, что некоторые параметры данных имеют определенные значения.
- c. Тестирование гипотез - это процесс проверки статистических гипотез о том, что некоторые параметры данных имеют определенные значения.
- d. Нет правильного ответа.

11. Что такое уровень значимости?

- a. **Уровень значимости - это вероятность ошибки первого рода, то есть отвержения верной гипотезы.**
- b. Уровень значимости - это вероятность ошибки второго рода, то есть отвержения верной гипотезы.
- c. Уровень значимости - это показатель верной гипотезы.
- d. Нет правильного ответа

12. Что такое уровень доверия?

- a. **Уровень доверия - это вероятность того, что результаты теста правильные.**
- b. Уровень доверия - это вероятность ошибки первого рода, то есть отвержения верной гипотезы.
- c. Уровень доверия – это вероятность того, что результаты теста неправильные.
- d. Верного ответа нет.

13. Что такое t-тест?

- a. **T-тест - это тест на проверку гипотезы о равенстве средних значений двух выборок.**
- b. T-тест - это тест на проверку гипотезы о равенстве соответствующих значений двух выборок.
- c. T-тест - это тест на проверку гипотезы о неравенстве средних значений двух выборок.
- d. Нет правильного ответа

14. Как реализовать t-тест на Python?

- a. **Для реализации t-теста на Python можно использовать библиотеку `scipy.stats`**
- b. Для реализации t-теста на Python можно использовать библиотеку `matplotlib`
- c. Для реализации t-теста на Python можно использовать библиотеку `pumpy`
- d. Нет правильного ответа

15. Что такое z-тест?

- a. **Z-тест - это тест на проверку гипотезы о равенстве средних значений двух выборок.**
  - b. Z-тест - это тест на проверку гипотезы о равенстве средних значений двух выборок.
  - c. T-тест - это тест на проверку гипотезы о неравенстве средних значений двух выборок.
  - d. Нет правильного ответа.
16. Как реализовать z-тест на Python?
- a. Для реализации z-теста на Python мы также можем использовать библиотеку `scipy.stats`. Например:

```
```python
import scipy.stats as stats
# создаем две выборки
sample1 = [1, 2, 3, 4, 5]
sample2 = [2, 4, 6, 8, 10]
# применяем z-тест
z_statistic, p_value = stats.zscore(sample1, sample2)
# выводим результаты
print(&quot;z-statistic:&quot;, z_statistic)
print(&quot;p-value:&quot;, p_value)
```
```
  - b. Для реализации z-теста на Python мы также можем использовать библиотеку `scipy.stats`. Например:

```
```python
import scipy
# создаем две выборки
sample1 = [1, 2, 3, 4, 5]
sample2 = [2, 4, 6, 8, 10]
# применяем z-тест
z_statistic, p_value = stats.zscore(sample1, sample2)
# выводим результаты
print(&quot;z-statistic:&quot;, z_statistic)
print(&quot;p-value:&quot;, p_value)
```
```
  - c. Для реализации z-теста на Python мы также можем использовать библиотеку `scipy.stats`. Например:

```
```python
import scipy.stats as stats
# создаем две выборки
sample1 = [1, 2, 3, 4, 5]
sample2 = [2, 4, 6, 8, 10]
# применяем z-тест
z_statistic, p_value = stats.zscore
# выводим результаты
print(&quot;z-statistic:&quot;, z_statistic)
```



```
print(&quot;p-value:&quot;, p_value)
...
```

d. Правильного ответа нет.

17. Что такое Хи-квадрат тест?

- a. **Хи-квадрат тест - это тест на проверку гипотезы о том, что две выборки имеют одинаковое распределение.**
- b. Хи-квадрат тест - это тест на проверку гипотезы о равенстве средних значений двух выборок, который используется, когда мы имеем большой объем выборок (обычно более 30 наблюдений) и известную дисперсию.
- c. Хи-квадрат тест - это тест на проверку гипотезы о равенстве средних значений двух выборок.
- d. Нет правильного ответа.

18. Когда используется z-тест?

- a. **Когда мы имеем большой объем выборок (обычно более 30 наблюдений) и известную дисперсию.**
- b. Когда мы имеем маленький объем выборок.
- c. Нет правильного ответа
- d. Все варианты верны

19. Что применяется, когда мы хотим выяснить, есть ли статистически значимая разница между распределениями двух выборок?

- a. **Хи-квадрат тестирование**
- b. z-тест
- c. t-тест
- d. нет правильного ответа

20. Какая величина характеризует разброс значений относительно среднего значения, но выражается в тех же единицах измерения, что и исходные данные?

- a. **Стандартное отклонение**
- b. Дисперсия
- c. Математическое ожидание
- d. Нет правильного ответа