### **Основные статистические тесты и проверка гипотез**

#### **Повторение пройденного**

* Доверительный интервал
* Правило трех сигм
* Виды распределений: дискретное и непрерывное
* Скалярное произведение векторов и проецирование вектора
* Линейная регрессия и классификационный анализ

##### **Доверительный интервал**

Доверительный интервал — термин, используемый в математической статистике при интервальной оценке статистических параметров, более предпочтительной при небольшом объеме выборки, чем точечная.

Доверительным называют интервал, который покрывает неизвестный параметр с заданной надежностью.

Доверительным называется интервал, в который попадают измеренные в эксперименте значения, соответствующие доверительной вероятности.

##### **Правило трех сигм**

Правило, утверждающее, что вероятность того, что случайная величина отклонится от своего математического ожидания более чем на три среднеквадратических отклонения, практически равна нулю. Правило справедливо только для случайных величин, распределенных по нормальному закону.

##### **Виды распределений: дискретное и непрерывное**

**Дискретные распределения**

Моделируют вероятность наступления дискретных событий, например, выпадение герба или решки (распределение Бернулли), число выпадений герба или решки при многократном бросании монеты (биномиальное распределение), выпадения определенного числа очков при бросании игральной кости (полиномиальное или мультиномиальное распределение).

Примеры дискретных величин самые разнообразные: число телефонных звонков за день, количество перевезенных пассажиров, количество дефектов в партии продукции, количество распавшихся атомов за определенный промежуток, число квантов света, попавших на сетчатку глаза и множество других в физике, технике, биологии, медицине, экономике, транспорте, телефонии.

**Непрерывные распределения**

Моделируют вероятность наступления непрерывных событий, например, длительность телефонного звонка, момент наступления телефонного звонка, уровень шума в сети, расход электроэнергии за день, местоположение дефектов в микросхеме, количество осадков за месяц, биржевые данные, расстояние, пройденное молекулой газа до следующего столкновения и т.д.

Задавая распределение, вы строите вероятностную модель и затем оцениваете вероятности наступления более сложных событий, например, вероятность того, что количество прерванных звонков в сети превысит норму, вероятность того, что производственный процесс выйдет за допустимые пределы, количество бракованных деталей будет критическим, вероятность возникновения взрыва при ядерной реакции и т. д.

##### **Скалярное произведение векторов и проецирование вектора**

**Скалярным произведением двух векторов** называется действительное число, равное произведению длин умножаемых векторов на косинус угла между ними.

**Проекция вектора** — это скалярная величина (число), равная длине геометрической проекции вектора, если направление оси и геометрической проекции совпадают; или число, противоположное длине геометрической проекции вектора, если направления геометрической проекции и оси — противоположные.

##### **Линейная регрессия**

**Линейная регрессия** предполагает линейную или прямую зависимость между входными переменными (X) и единственной выходной переменной (y).

Более конкретно, этот выход (y) может быть вычислен из линейной комбинации входных переменных (X). Когда имеется одна входная переменная, метод называется простой линейной регрессией.

В простой линейной регрессии мы можем использовать статистику обучающих данных для оценки коэффициентов, требуемых моделью для прогнозирования новых данных.

##### **Логистическая регрессия**

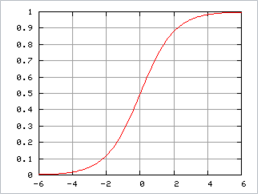
**Логистическая регрессия** — это статистическая модель, используемая для прогнозирования вероятности возникновения некоторого события.

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится так называемая зависимая переменная y, принимающая лишь одно из двух значений — как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество независимых переменных (также называемых признаками, предикторами или регрессорами) — вещественных x1, x2, ... ,xN, на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной.

#### **Сигмоида**

**Сигмоида** — это гладкая монотонная возрастающая нелинейная функция, имеющая форму буквы «S», которая часто применяется для «сглаживания» значений некоторой величины.

Часто под сигмоидой понимают логистическую функцию



#### **Градиентный спуск**

**Градиентный спуск** — метод нахождения локального экстремума (минимума или максимума) функции с помощью движения вдоль градиента. Для минимизации функции в направлении градиента используются методы одномерной оптимизации, например, метод золотого сечения. Также можно искать не наилучшую точку в направлении градиента, а какую-либо лучше текущей.

#### **Квантиль**

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

Интеркварти́ льным размахом называется разность между третьим и первым квартилями, то есть x\_{0,75}-x\_{0,25}. Интерквартильный размах является характеристикой разброса распределения величины и является аналогом дисперсии. Вместе, медиана и интерквартильный размах могут быть использованы вместо математического ожидания и дисперсии в случае распределений с большими выбросами, либо при невозможности

вычисления последних.

#### **Генеральная совокупность и выборка**

**Генеральная совокупность** – это совокупность всех мысленно возможных объектов данного вида, над которыми проводятся наблюдения с целью получения конкретных значений определенной случайной величины.

Генеральная совокупность может быть конечной или бесконечной в зависимости от того, конечна или бесконечна совокупность составляющих ее объектов.

**Выборкой (выборочной совокупностью)** называется совокупность случайно отобранных объектов из генеральной совокупности.

Выборка должна быть репрезентативной (представительной), то есть ее объекты должны достаточно хорошо отражать свойства генеральной совокупности.

#### **Статистические гипотезы**

**Статистическая гипотеза** — предположение о виде распределения и свойствах случайной величины, которое можно подтвердить или опровергнуть применением статистических методов к данным выборки.

**Нулевая гипотеза** — гипотеза о сходстве

**Альтернативная гипотеза, конкурирующая** — гипотеза о различиях

#### **Статистически значимость**

Величину или значение называют статистически значимой, если мала вероятность случайного возникновения этой или еще более крайних величин. Здесь под крайностью понимается степень отклонения тестовой статистики

Статистическая значимость (значение Р) - расчетная вероятность ошибки первого рода, которая рассчитывается с помощью различных статистических критериев.

#### **A/B тестирование**

A/B-тестирование — метод маркетингового исследования, суть которого заключается в том, что контрольная группа элементов сравнивается с набором тестовых групп, в которых один или несколько показателей были изменены, для того, чтобы выяснить, какие из изменений улучшают целевой показатель.

#### **Виды статистических критериев**

**Критерии согласия** - проверка на согласие подразумевает проверку предположения о том, что исследуемая случайная величина подчиняется предполагаемому закону.

**Параметрические критерии** -группа статистических критериев, которые включают в расчет параметры вероятностного распределения признака

(средние и дисперсии).

**Непараметрические критерии** - группа статистических критериев, которые не включают в расчет параметры вероятностного распределения

и основаны на оперировании частотами или рангами.

#### **Параметрическая — непараметрическая гипотеза**

**Параметрические критерии** - группа статистических критериев, которые

включают в расчет параметры вероятностного распределения признака (средние и дисперсии):

* t-критерий Стьюдента
* Критерий Фишера
* Критерий отношения правдоподобия
* Критерий Романовского

**Непараметрические критерии** —группа статистических критериев, которые не включают в расчет параметры вероятностного распределения и основаны на оперировании частотами или рангами:

* Q-критерий Розенбаума
* U-критерий Манна — Уитни
* Критерий Уилкоксона
* Критерий Пирсона
* Критерий Колмогорова — Смирнова

#### **Распределение Стьюдента**

**Распределение Стьюдента** (t-распределение) в теории вероятностей — это однопараметрическое семейство абсолютно непрерывных распределений.

Распределение Стьюдента может быть использовано для оценки того, насколько вероятно, что истинное среднее находится в каком-либо заданном диапазоне.

#### **Тест Фишера**

Тест Фишера — тест статистической значимости, используемый в анализе при работе с выборками малых размеров.

Относится к точным тестам значимости, поскольку не использует приближения большой выборки (асимптотики при размере выборки стремящемся к бесконечности).

#### **Критерий Манна-Уитни**

U-критерий Манна — Уитни — статистический критерий, используемый для оценки различий между двумя независимыми выборками по уровню какого-либо признака, измеренного количественно.

Позволяет выявлять различия в значении параметра между малыми выборками.