Проверка статистических гипотез

Версия 4 Часть 1

Аббакумов Вадим Леонардович

Определение

- Статистическая гипотеза – утверждение о свойствах распределения вероятностей случайной величины *(или случайного вектора).*

- Гипотеза нуждается в проверке.
- Проверка основывается на результатах эксперимента, на наблюдениях.

Напоминание

- Что такое функция распределения?
- Что такое плотность распределения?

Когда проверяют статистические гипотезы?

A/B тесты

Sir Ronald A. Fisher in the 1920s

at the Rothamsted Agricultural Experimental Station

Amazon, Bing, Facebook, Google, LinkedIn и Yahoo!

10000 A/B тестов каждый год

X5 Retail Group (Пятерочка, Перекресток, Карусель)

Медицина - covid-19

135 производителей. Кто первый?

Randomized double blind placebo control studies

Что сравниваем:

- Доля участников с антителами
- Процент заболевших
- Тяжесть болезни (смертность)
- После вакцины пытаемся заразить

Web: 42 оттенка синего

Google: дополнительно 200M$ в год

https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers

2009 Marissa Meyer против Doug Bowman, ведущего дизайнера

Цвет ссылки — синий, но какой именно?

Докажи!

Рамочка — 3, 4 или 5 пикселей?

Когда проверяют статистические гипотезы?

Что лучше: получить 1000 рублей сегодня или 1100 рублей завтра?

Что лучше: получить 1000 рублей через 30 дней или 1100 рублей через 31 день?

Как возникает потребность в проверке статистических гипотез?

Увеличивает ли продажи ограничение отпуска товара в одни руки?

Например, не больше 5 единиц

Раздел 1

Зачем проверяют статистические гипотезы

Обсудим наиболее важные статистические гипотезы.

1. Гипотеза согласия.

- Обозначим $F_X(t)$ функцию распределения случайной величины X.
- Пусть $F_0(t)$ - некоторая заданная функция распределения.
- Гипотеза : функции распределения совпадают, то есть $F_X(t) = F_0(t)$

- Кому и когда приходится проверять гипотезу согласия?

Пример гипотезы согласия

- Гипотеза о нормальности распределения

$$F_0(t) = \frac{1}{\sqrt{2\pi\sigma_X^2}} \int_{-\infty}^{t} \exp\left(-\frac{(s-a_X)^2}{2\sigma_X^2}\right) ds$$

чаще всего расширяем гипотезу

Когда проверяем гипотезу нормальности,

допустимы любые значения
математического ожидания и дисперсии

кроме критерия Колмогорова-Смирнова

Немецкая марка и нормальное распределение

Почему гипотеза нормальности важна?

- 1. Нормальное распределение часто встречается

  (вспомним центральную предельную теорему).

$$\sqrt{n}\cdot\frac{\bar{X}_n-\mu}{\sqrt{var}(X)}\rightarrow N(0,1)$$

Почему гипотеза нормальности важна?

- 2. Когда распределение нормальное, экономим деньги:
- если
- А) распределение можно считать нормальным и
- Б) задана необходимая погрешность результата,
- то потребуется меньше наблюдений.
- Например, опросим меньше покупателей.

Пример гипотезы согласия 2

- Гипотеза об экспоненциальности распределения.
- В этом случае функция распределения

$$F_0(t) = \begin{cases} 1 - e^{-\lambda t}, t > 0 \\ 0 \quad\quad , t < 0 \end{cases}$$

Почему важна гипотеза экспоненциальности?

- Экспоненциальное распределение часто встречается, когда изучается «время ожидания».

Например,

- Время до аварии (нужно для расчета страховой премии).
- Время обслуживания покупателя кассиром (нужно для определения числа работающих касс в супермаркете).
- Время до поломки изделия (нужно для планирования расходов на гарантийный ремонт).

2. Гипотеза однородности.

- Обозначим $F_X(t)$ функцию распределения случайной величины X.
- Обозначим $F_Y(t)$ функцию распределения случайной величины Y
- Гипотеза : функции распределения совпадают

$$F_X(t) = F_Y(t)$$

- Кому и когда приходится проверять гипотезу согласия?

Например,

- Распределение продаж до рекламной акции и после нее.
- Если распределение продаж не изменилось, то улучшения нет.

- Может сравниваться распределение покупателей по возрасту. Например, если реклама была нацелена на конкретный сегмент, например, на молодых мам.

3. Гипотеза независимости.

- Гипотеза : случайные величины X и Y независимы
- Кому и когда приходится проверять гипотезу независимости?

Если зависимы, то X влияет на Y

Не верно

Нужны аргументы из предметной области

Идет дискуссия, что добавить к гипотезе,

чтобы можно было доказывать причинность, casuality

Например,

- Если возраст покупателей и объем покупки зависимы, то возраст надо учитывать при сегментации покупателей.

- Иногда зависимость бывает неочевидной.
- Длина волос и рост людей – зависимые переменные.

Вопрос:

- наличие балкона влияет на цену квартиры?

- Швеция, 40-70-е годы

  Число аистов и рождаемость

- Ложная регрессия

На шаг дальше…

- В эконометрике редко интересен сам факт зависимости. Обычно идут дальше, пытаются описать зависимость.

- Подобные задачи решаются, в частности, методами регрессионного анализа и машинного обучения.
- Эти методы – следующая тема.

4. Гипотезы о параметре распределения.

- Чаще всего распределение случайной величины не важно.
- Важна лишь одна характеристика распределения.

Если анализируются продажи магазина, то в первую очередь интересно…

Математическое ожидание

- Так как математическое ожидание – вероятностная модель для среднего значения.
- В данном случае для средних продаж.

Если сравниваются средние:

- Гипотеза. Математические ожидания случайных величин X и Y равны.

$$EX = EY$$

Если сравниваются медианы:

- Гипотеза: Медианы случайных величин X и Y равны.

$$Med(X) = med(Y)$$

Типичные значения (центры распределения)

Подмена задачи

Сравнение столбцов подменяем сравнением типичных значений

Остается решить,
когда лучше сравнивать средние значения,
а когда лучше сравнивать медианы

Пример 1

- В обычных условиях зафиксирован некоторый уровень продаж. Затем была проведена рекламная акция.
- Чтобы оценить результат, нужно выяснить, было ли существенное увеличение продаж. В частности, окупились ли затраты на рекламу.

Основная проблема:

Увеличение продаж могло быть вызвано случайными факторами.

- Продажи все время меняются, случайным образом отклоняются от заданного значения.
- Статистически значимое отклонение должно превышать эти случайные отклонения.

Пример 2

- Разработан новый варианта упаковки товара.
- Требуется проверить предположение, что товар в новой упаковке имеет в данном регионе больший уровень продаж, чем вариант в старой упаковке.

Пример 3

Эффект от скидок?

(Там разные метрики)

Основные условия применения статистических тестов

- Вопрос должен касаться какой-либо характеристики массового явления.
- Характеристика меняется случайным образом от наблюдения к наблюдению.
- Вопрос должен быть относительно простым и четко сформулированным

Строительство танкеров на адмиралтейских верфях

Продажи пива

Пользователи интернета