Введение в машинное обучение для Java-разработчиков

Лекция 4

Ермилов Сергей

education

Содержание

- Основные понятия статистики
- Статистические гипотезы
- Статистические критерии
- Критерий Стьюдента
- A/B тестирование

Основные понятия статистики

Выборка

Генеральная совокупность — множество всех объектов или событий исследуемых в рамках поставленной задачи.

Исследовать все элементы генеральной совокупности невозможно, поэтому исследуется только часть из них.

Выборка - элементы генеральной совокупности которые мы непосредственно исследуем.

С какими проблемами мы можем столкнуться используя подобный подход?

Статистика

Выборка, хорошо отражающая свойства генеральной совокупности называется **репрезентативной**.

Обычно исследуют некоторую числовую характеристику генеральной совокупности.

Статистика - любое численное значение вычисленное на основе элементов выборки.

Примеры статистик

Выборочное среднее:

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Несмещенная выборочная дисперсия:

$$S^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N - 1}$$

Медиана:

$$Median = Sorted(X)[N//2]$$

Мода:

наиболее часто встречающееся значение в выборке

Статистические гипотезы

Понятие статистической гипотезы

Статистическая гипотеза — предположение о свойствах и характеристиках исследуемых генеральных совокупностей.

Пример статистической гипотезы: Средний возраст людей в России - 40 лет.

Понятие статистической гипотезы

Пример статистической гипотезы: Средний возраст людей в России - 40 лет.

Если возраст людей в выборке сильно больше или сильно меньше, то эта гипотеза неверна.

Вы опрашиваете своих друзей, получаете средний возраст - 25 лет.

Понятие статистической гипотезы

Пример статистической гипотезы: Средний возраст людей в России - 40 лет.

Если возраст людей в выборке сильно больше или сильно меньше, то эта гипотеза неверна.

Вы опрашиваете своих друзей, получаете средний возраст - 25 лет.

Полученная выборка может быть **нерепрезентативна** относительно генеральной совокупности - всего населения России.

Репрезентативность выборки

Понятие статистической гипотезы

Пример

Вы решили проверить гипотезу о том, что монета является симметричной.

Понятие статистической гипотезы

Пример

Вы решили проверить гипотезу о том, что монета является симметричной.

Чтобы сделать это, вы подбрасываете её 10 раз.

Понятие статистической гипотезы

Пример

Вы решили проверить гипотезу о том, что монета является симметричной.

Чтобы сделать это, вы подбрасываете её 10 раз.

Если “орел” выпал 4-6 раз из них, то с гипотезой можно согласиться.

Понятие статистической гипотезы

Пример

Вы решили проверить гипотезу о том, что монета является симметричной.

Чтобы сделать это, вы подбрасываете её 10 раз.

Если “орел” выпал 4-6 раз из них, то с гипотезой можно согласиться.

Но, если он выпал 10 или 0 раз, то она выглядит сомнительной, так как вероятность этого равна 2^{-10}

Понятие статистической гипотезы

Пример

Вы решили проверить гипотезу о том, что монета является симметричной.

Чтобы сделать это, вы подбрасываете её 10 раз.

Если “орел” выпал 4-6 раз из них, то с гипотезой можно согласиться.

Но, если он выпал 10 или 0 раз, то она выглядит сомнительной, так как вероятность этого равна 2^{-10}.

Следовательно при проверке гипотезы мы должны ориентироваться на некоторый *порог статистической значимости* или *статистический критерий*.

Статистическая гипотеза

Основная или нулевая гипотеза H_0 - это гипотеза, которой мы придерживаемся, пока наблюдения не заставят признать обратное. Ей всегда сопутствует **альтернативная гипотеза** H_1 .

Замечания:

- В статистике нельзя доказать гипотезу, можно только ее опровергнуть.
- Гипотеза порождает следствия - опровергаем следствие, опровергаем гипотезу.
- Если данные согласуются со следствием, нельзя утверждать что гипотеза верна.
- Опровержение гипотезы всегда проводится с некоторой *вероятностью* .

Ошибки первого и второго рода

	α-порог H_0	β-порог H_1
H_0	TP	FP
H_1	FN	TN

TP (True Positive) — нулевая гипотеза верно принята.

TN (True Negative) — нулевая гипотеза верно отвергнута.

FP (False Positive) — ошибка первого рода — нулевая гипотеза неверно отвергнута.

FN (False Negative) — ошибка второго рода — нулевая гипотеза неверно принята.

Ошибки первого и второго рода

Пример

Задача классификации “спам - не спам”. Вопрос - “является ли письмо спамом?”

TP — письмо верно помечено как “спам”.

TN — письмо не является спамом.

Соответственно:

FP (type I error) — письмо не является спамом, но было отправлено в корзину.

FN (type II error) — письмо является спамом, но попало во входящие.

Уровень значимости

Низкое значение уровня значимости α уменьшает вероятность совершить ошибку первого рода, но увеличивают вероятность совершить ошибку второго рода.

При α = 0 мы принимаем, что монетка симметрична, даже если 100 раз из 100 выпал “орел”.

Уровень значимости

На практике уровень значимости α чаще принимается за 0.01 (1%), 0.05 (5%) или 0.1 (10%) в зависимости от допустимости ошибки первого и второго рода.

Значение (1 - α) называют *уровень доверия* или *доверительной вероятностью*. Для вышеприведенных значений уровня значимости она равна соответственно 0.99, 0.95 и 0.90.

Минимальный уровень значимости — это минимальное значение α, при котором основная гипотеза ещё отвергается.

Мощность статистического критерия

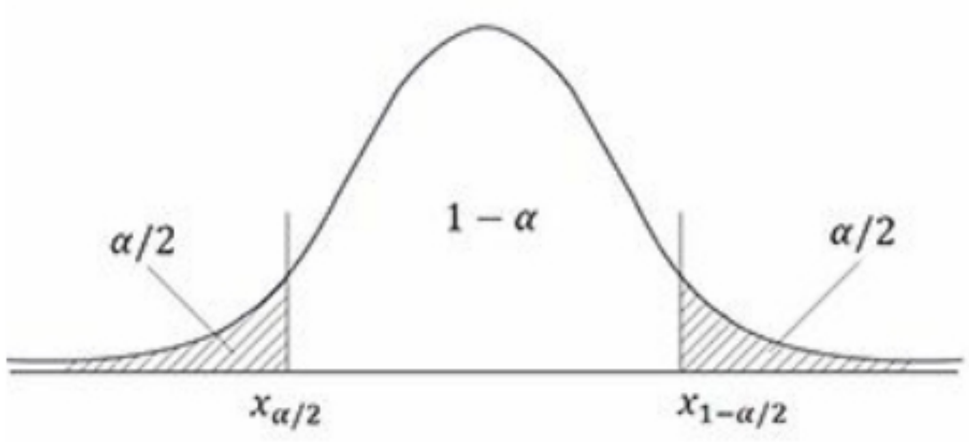
β - вероятность ошибки второго рода.

Значение (1 - β) называют *мощностью статистического критерия*.

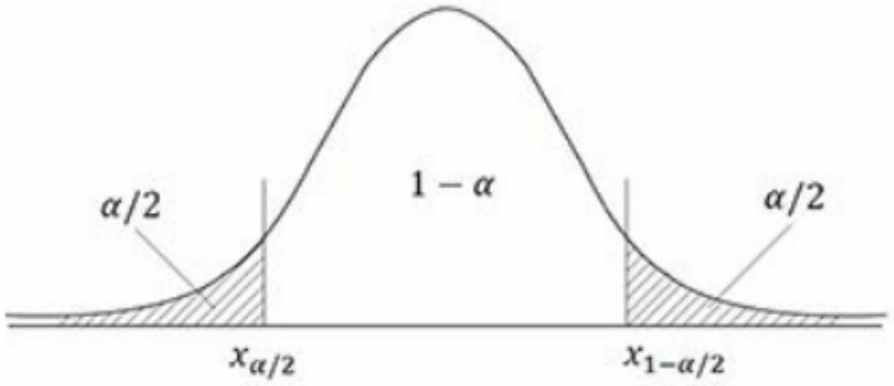
Статистика и p-value

На основе полученной выборки мы получаем *статистику*.

На основе гипотезы мы знаем как распределены значения статистики, и с какой вероятностью статистика может принимать то или иное значение. Если вероятность полученного значения статистики очень мала, то мы отвергаем нулевую гипотезу.



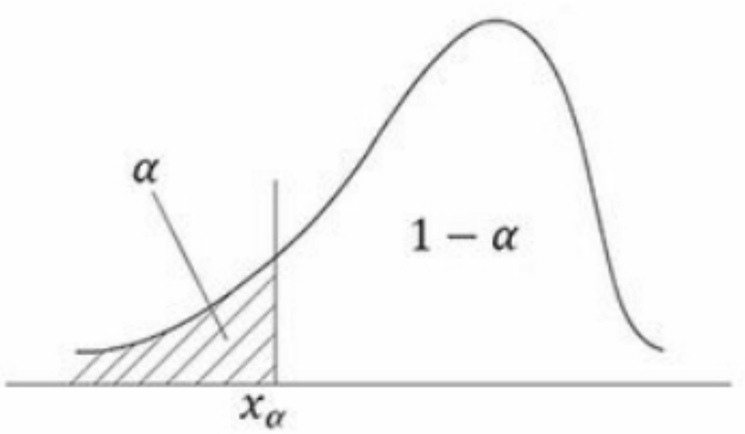
Критическая область



Если мы построим график вероятности выпадения “орла” и “решки” (биномиальное распределение), то сможем отобразить некоторое *критическое значение* α, после которого начинается *критическая область* вероятности, в которой гипотезу о симметричности монеты можем считать отвергнутой. Но, так как у нас таких значений два: чаще выпадает “орел” или чаще выпадает “решка”, то и критических областей будет две.

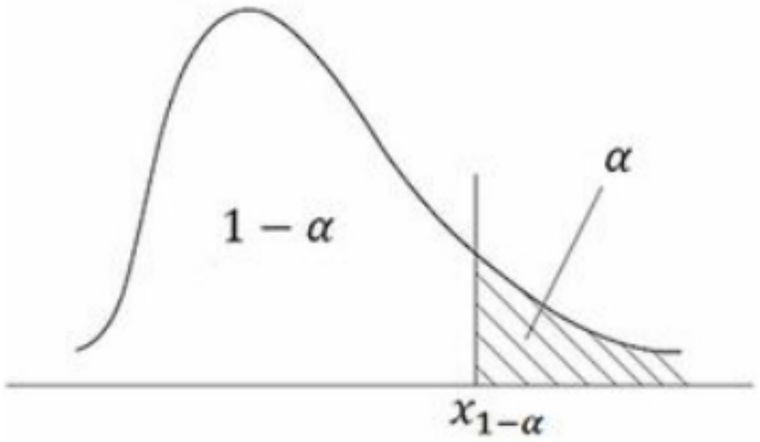
Это случай *двусторонней критической области*.

Критическая область



По тому же принципу бывают *левосторонняя* и *правосторонняя* критические области.

Пример гипотезы, требующей одностороннего критерия. Средний рост женщин в России больше 160 см.



Статистический тест

Процедура проверки гипотезы состоит из нескольких этапов:

1. Сформулировать основную и альтернативную гипотезы и задать уровень значимости α.
2. Найти критические значения и построить критическую область.
3. Вычислить по выборке значение статистики и посмотреть, попало ли оно в критическую область.
4. Сделать вывод. Если значение попало в критическую область, то основная гипотеза отвергается, в противном случае, не отвергается.

Статистический тест

Статистический тест не показывает истинность или ложность гипотезы, а только нашу уверенность в ней на основе имеющихся данных.

Гипотезу либо отвергают, либо не отвергают. Невозможность отвергнуть гипотезу не означает, что она верна и её стоит придерживаться.

Статистические критерии

Статистические критерии

Для нахождения критического значения используют **статистические критерии**. Критерии делятся на группы. Подходящий критерий выбирается исходя из нулевой гипотезы и дополнительных предположений. Например, о характере распределения исследуемой величины.

Статистические критерии

Критерии согласия проверяют, согласуется ли заданная выборка с заданным фиксированным распределением, с заданным параметрическим семейством распределений, или с другой выборкой.

Примеры критериев согласия:

- Критерий Колмогорова-Смирнова
- Критерий хи-квадрат (Пирсона)
- Критерий омега-квадрат (фон Мизеса)

Статистические критерии

Критерии сдвига являются специальным случаем двухвыборочных критериев согласия. Проверяется гипотеза сдвига, согласно которой распределения двух выборок имеют одинаковую форму и отличаются только сдвигом на константу.

- Критерий Стьюдента
- Критерий Уилкоксона-Манна-Уитни

Статистические критерии

Критерии нормальности — это выделенный частный случай критериев согласия. Нормально распределённые величины часто встречаются в прикладных задачах, что обусловлено действием ЦПТ. Если про выборки заранее известно, что они подчиняются нормальному распределению, то к ним становится возможно применять более мощные параметрические критерии. Проверка нормальность часто выполняется на первом шаге анализа выборки, чтобы решить, использовать далее параметрические методы или непараметрические.

- Критерий Шапиро-Уилка
- Критерий асимметрии и эксцесса

Статистические критерии

Критерии однородности предназначены для проверки нулевой гипотезы о том, что две (или несколько) выборки взяты из одного распределения, либо их распределения имеют одинаковые значения математического ожидания, дисперсии, или других параметров.

Критерии симметричности позволяют проверить симметричность распределения.

Критерии тренда и случайности предназначены для проверки нулевой гипотезы об отсутствии зависимости между выборочными данными и номером наблюдения в выборке. Применяются в анализе временных рядов.

Критерии выбросов

Критерии дисперсионного анализа

Критерии корреляционного анализа

Критерии регрессионного анализа

Критерий Стьюдента

Критерий Стьюдента

Самый распространенный класс методик для проверки гипотез базируется на **t-критерии Стьюдента**. Используется для проверки равенства средних значений (мат. ожидания) в двух выборках, имеющих распределение Стьюдента.

Критерий Стьюдента

Допустим пациенты, которые приняли новый препарат, выздоравливали, в среднем, за 14 дней.

Пациенты контрольной группы, принимавшие плацебо, выздоравливали, в среднем, за 18 дней.

Является ли разница в 4 дня показателем, того что препарат действительно работает или различия являются статистической погрешностью?

Критерий Стьюдента

Если выборка подчиняется нормальному распределению и не содержит выбросов, то ошибки этой выборки тоже будут подчиняться закону нормального распределения (или закону t-распределения Стьюдента, когда размер выборки меньше 30 наблюдений.)

В этом случае в качестве критического значения мы можем взять некоторый множитель на стандартное отклонение выборки. Диапазон между критическими значениями будет называться *доверительным интервалом*.

Если средние по выборкам находятся за пределами доверительного интервала — это означает, что такая разница получена между ними не случайно.

Критерий Стьюдента

ctrl
24
18
13
14
15
26
22
11
17
20

Среднее арифметическое:

$$X_{mean} = \frac{24 + 18 + 13 + \ldots + 20}{10} = 18$$

Дисперсия:

$$D(X_c) = \frac{(24 - 18)^2 + \ldots + (20 - 18)^2}{9} = 24.4$$

Стандартное отклонение:

$$\sigma_c = \sqrt{D(X_c)} = 4.94$$

Стандартная ошибка среднего:

$$\mu_c = \frac{\sigma}{\sqrt{n}} = 1.562$$

Критерий Стьюдента

ctrl
24
18
13
14
15
26
22
11
17
20

Предельная ошибка $\Delta = \mu \cdot t$, где t – уровень доверия:

$$\Delta_c = 2*1.562$$

Доверительный интервал x ± μ для уровня доверия 95% (риск ошибки 5%):

[14.876; 21.124]

Критерий Стьюдента

ctrl	test
24	13
18	14
13	16
14	17
15	12
26	13
22	15
11	11
17	15
20	14

Предельная ошибка $\Delta = \mu \cdot t$, где t – уровень доверия:

$$\Delta_c = 1.564 \cdot 2 = 3.128$$

Доверительный интервал $x \pm \mu$ для уровня доверия 95% (риск ошибки 5%):

$$[18 - 3.128; 18 + 3.128]$$

Критерий Стьюдента

ctrl	test
24	13
18	14
13	16
14	17
15	12
26	13
22	15
11	11
17	15
20	14

Среднее арифметическое:

$$X_{mean} = \frac{13 + 13 + 16 + \ldots + 14}{10} = 14$$

Дисперсия:

$$D(X_t) = \frac{(13 - 14)^2 + \ldots + (14 - 14)^2}{9} = 3.33$$

Стандартное отклонение:

$$\sigma_c = \sqrt{D(X_c)} = 1.83$$

Стандартная ошибка среднего:

$$\mu_c = \frac{\sigma}{\sqrt{n}} = 0.577$$

Критерий Стьюдента

ctrl	test
24	13
18	14
13	16
14	17
15	12
26	13
22	15
11	11
17	15
20	14

t-критерий Стьюдента:

$$t = \frac{(X_c - X_t)}{\sqrt{\mu_c^2 + \mu_t^2}}$$

Для распределения Стьюдента при n = 10:

n	Доверительная вероятность			
	0.9	0.95	0.99	0.999
10	1.812461	2.228138	3.169272	4.586893

Для уровня значимости α = 0.05:

???

Мы делаем вывод о том, что наши выборки независимы. А значит мы не можем отвергнуть гипотезу о действенности препарата???

Критерий Стьюдента

Условия применения:

1) нормальность распределения признака в обеих группах (на самом деле это не так, достаточно нормальности среднего)
2) равенство дисперсий двух сравниваемых групп.

Если данные не отвечают этим критериям, то применяется *U-критерий Манна-Уитни* — непараметрический тест, в котором для расчета используются не исходные данные, а их ранговые позиции.

Если групп больше двух, подойдет *критерий Краскела-Уоллиса*.

Если выборок две и они зависимые применяется ранговый Т-критерий Уилкоксона.

A/B тестирование

A/B-тестирование (или *сплит-тестирование*) — маркетинговый метод, который используется для оценки эффективности веб-страниц и управления ими.

При A/B-тестировании сравнивают страницы A и B, имеющие разные элементы дизайна или функционала (например, цвета кнопки заказа товара). На каждую страницу случайным образом запускают 50% аудитории сайта и затем сравнивают, какая страница показывает наибольший процент конверсии.

За нулевую гипотезу берется предположение, что конверсия на странице B не отличается от конверсии на странице A. Соответственно, обратное утверждение берется за альтернативную гипотезу.

A/B тестирование

1. Определить цель эксперимента и метрики
2. Выбрать аудиторию
3. Рассчитать длительность эксперимента
4. Провести A/A тест
5. Провести A/B тест
6. Проанализировать результаты эксперимента
7. Выкатить фичу на прод

В сервисах обычно идут тысячи экспериментов и анализ десятков тысяч метрик
(могут быть ложные срабатывания)