

2. ISO/IEC 14496-10:2022 Information technology — Coding of audio-visual objects Part 10: Advanced video coding [Электронный ресурс]. – Режим доступа: <https://www.iso.org/standard/83529.html> (дата обращения: 02.11.2024).

3. Онкович Г. В. Медиаобразование: изменение в тематических предпочтениях - управляемый и прогнозируемый [Электронный ресурс]. – Режим доступа: <https://cyberleninka.ru/article/n/mediaobrazovanie-izmenenie-v-tematicheskikh-predpochteniyah-upravlyaemyy-i-prognoziруemyy-protsess> (дата обращения: 07.11.2024).

УДК 004.773.2

А.Х. Оразгелдиев

Институт телекоммуникаций и информатики
Ашхабад Туркменистан

НАУКА О ДАННЫХ

***Аннотация.** Одним из фундаментальных аспектов анализа данных являются статистические гипотезы и методы их проверки. В этой статье мы подробно рассмотрим основные понятия, процессы и применение статистических гипотез и критериев в Data Science.*

А.Н. Orazgeldiev

Institute of Telecommunications and Informatics
Ashgabat Turkmenistan

DATA SCIENCE

***Abstract.** One of the fundamental aspects of data analysis is statistical hypotheses and methods of their verification. In this article, we will take a detailed look at the basic concepts, processes, and applications of statistical hypotheses and criteria in Data Science.*

Data Science — это область, которая стремительно развивается и охватывает множество дисциплин, включая статистику, машинное обучение и анализ данных. Одним из фундаментальных аспектов анализа данных являются статистические гипотезы и методы их проверки. В этой статье мы подробно рассмотрим основные понятия, процессы и применение статистических гипотез и критериев в Data Science.

Пусть любое надмножество изучается относительно цели.

Предположения, сделанные о неизвестном распределении этой выборки или неизвестных параметрах известного распределения, называются статистическими гипотезами. Предположение о том, что неизвестный параметр имеет определенное значение, называется простым предположением. Если параметр принимает значения из некоторого множества, то он называется сложной гипотезой. Например, если

$$f(x, a, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

если это функция плотности нормального закона, то предположение " $a = 0, \sigma = 1$ " является простым, а предположение " $a = a_0$ " сложным, поскольку в последнем предположении параметр σ может принимать любое неотрицательное значение.

Задача проверки статистических гипотез ставится следующим образом. Пусть $f(x; \theta)$ не зависит от главного множества функционала плотности

$$x = (x_1, x_2, \dots, x_n) \quad (1)$$

выбор будет сделан. Параметр θ называется гипотезой $H_0: \theta \in A$ (A – некоторое множество). Эта гипотеза называется первичной (или нулевой) гипотезой. Наряду с основной гипотезой H_0 рассматривают также конкурирующую или альтернативную гипотезу $H_1: \theta \notin A$. Например:

- Нулевая гипотеза (H_0): Утверждает, что нет эффекта, связи или различия. Например, "новая реклама не увеличивает продажи".

- Альтернативная гипотеза (H_1): Утверждает, что эффект, связь или различие существует. Например, "новая реклама увеличивает продажи".

Одной из важных задач является определение согласия выбора (1) с основной гипотезой H_0 . Они используют случайную величину с известным распределением в качестве критерия проверки основной гипотезы. После выбора определенного критерия он делит множество значений на две непересекающиеся части. В одной из этих частей основная гипотеза принимается, а в другой отвергается. Часть основной гипотезы, которая отвергается, называется критической массой. Точки, разделяющие допустимое и критическое значения оценки, называются критическими точками и обозначаются k_{kr} .

Если закон распределения случайной величины, выбранной в качестве критерия, не представляет интереса, то обозначим его K .

$$K < k_{kr}, K > k_{kr}, (k_{kr} > 0) \quad K < k_{kr.1} \text{ we } K > k_{kr.2} \quad (k_{kr.1} < k_{kr.2})$$

множества, определяемые неравенствами, называются соответственно левосторонними, правосторонними и двусторонними критическими множествами. (1) вероятность того, что выборка попадет в критическую массу установить с помощью S :

$$P(x \in S) = P_{\theta}(S) = \int_S f(x; \theta) dx \quad (2)$$

Они выбирают критическую массу (2) так, чтобы вероятность была наименьшей. Теперь рассмотрим некоторые примеры:

1. Проверка средних значений:
 - H_0 : Средний вес мужчин и женщин в группе равен ($\mu_1 = \mu_2$).
 - H_1 : Средний вес мужчин и женщин в группе не равен ($\mu_1 \neq \mu_2$).
2. Проверка пропорций:
 - H_0 : Доля клиентов, предпочитающих продукт А, равна 50% ($p = 0.5$).
 - H_1 : Доля клиентов, предпочитающих продукт А, не равна 50% ($p \neq 0.5$).

Процесс проверки гипотез

1. Формулировка гипотез

Первый шаг — четкая формулировка нулевой и альтернативной гипотез. Это критически важный этап, поскольку от правильности формулировки зависит корректность всех последующих шагов.

2. Сбор данных

Следующий этап включает сбор данных, которые будут использоваться для проверки гипотез. Данные могут быть получены из различных источников: эксперименты, опросы, базы данных и т. д. Важно, чтобы данные были репрезентативными и соответствовали условиям проверки гипотез.

3. Выбор статистического критерия

Выбор статистического теста зависит от типа данных и задач. Вот несколько распространённых тестов:

t-тест

- Когда использовать: Для сравнения средних значений двух независимых групп.

- Предположения: Нормальное распределение, однородность дисперсий (при использовании двух выборочного t-теста).

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

где $\bar{X}$ — среднее, s — стандартное отклонение, n — размер выборки.

ANOVA (дисперсионный анализ)

- Когда использовать: Для сравнения средних значений более чем двух групп.

-Предположения: Нормальное распределение, однородность дисперсий.

$$F = \frac{\text{вариация внутри групп}}{\text{вариация между группами}}$$

Хи-квадрат тест используется для анализа зависимостей между категориальными переменными. Наблюдения должны быть независимыми.

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

где O — наблюдаемое значение, E — ожидаемое значение.

На этом этапе используется выбранный статистический тест для расчета соответствующего статистического значения (t , F , χ^2 или Z) и p -значения. Сравните p -значение с заранее установленным уровнем значимости (α , обычно 0.05). Если $p < \alpha$, нулевая гипотеза отвергается.

На основании полученных данных необходимо сделать выводы. Важно не только объявить, что гипотеза отвергнута или принята, но и интерпретировать, что это значит в контексте исследования.

A/B-тестирование — это метод, который широко используется в маркетинге и продуктовой аналитике для сравнения двух версий одного продукта. Например, можно протестировать две версии веб-страницы, чтобы определить, какая из них лучше конвертирует пользователей. Здесь H_0 может утверждать, что конверсии равны, а H_1 — что одна из версий лучше другой.

Моделирование и прогнозирование. При построении предсказательных моделей важно проверять предположения о данных. Например, в регрессионном анализе необходимо проверять, является ли связь между переменными статистически значимой.

Временные ряды часто анализируются с целью выявления трендов и аномалий. Статистические тесты могут помочь определить, были ли изменения значительными или случайными.

На этапе предварительного анализа данных (EDA) важно проверять различные предположения о данных. Например, можно проверить нормальность распределения данных с помощью теста Шапиро-Уилка или теста Колмогорова-Смирнова.

Статистические гипотезы и критерии играют критическую роль в Data Science, позволяя исследователям и аналитикам делать обоснованные выводы на основе данных. Понимание этих концепций и методов их применения позволяет эффективно анализировать данные, оптимизировать процессы и принимать более обоснованные решения. Развитие навыков в этой области открывает новые возможности для карьерного роста и профессионального развития в мире Data Science.

Список использованных источников

1. "Data Science. Наука о данных с нуля" - Джоэл Грас.
2. "Теоретический минимум по Big Data. Все, что нужно знать о больших данных" - Анналин Ын.
3. "Основы Data Science и Big Data. Python и наука о данных" - Дэви Силен, Арно Мейсман, Мохамед Али.
4. "Data Science. Наука о данных" - Джон Келлехер, Брендан Тирни.

УДК 001.895:338.24

Л. С. Познякова

Белорусский государственный технологический университет,
г. Минск, Республика Беларусь

О НЕКОТОРЫХ МЕЖНАЦИОНАЛЬНЫХ ЦИФРОВЫХ ПЛАТФОРМАХ ОТКРЫТЫХ ИННОВАЦИЙ

Аннотация. Рассмотрены особенности функционирования межнациональных платформ открытых инноваций и приведены примеры наиболее известных из них. Материалы подготовлены при финансовой поддержке БРФФИ (НИР «Развитие инновационной деятельности на основе сетевого взаимодействия в условиях цифровой экономики», договор № Г23М-050 от 02.05.2023).

L. S. Poznyakova

Belarussian State Technological University,
Minsk, Republic of Belarus

THE EVOLUTION OF DIGITAL INNOVATION PLATFORMS

Abstract. The features of the functioning of transnational open innovation platforms are considered and examples of the most famous of them are given. The materials were