Соковнин Игорь Леонидович, факультет Искусственный интелект

Практическое задание ко уроку: Урок 2. Анализ данных и проверка статистических гипотез

1. **В чём различие между зависимыми и независимыми выборками?**

Независимые выборки характеризуются тем, что вероятность отбора любого испытуемого одной выборки не зависит от отбора любого из испытуемых другой выборки.

Независимые выборки - сравниваются две разные группы, например мужчины и женщины, молодые и пожилые и т.д

Зависимые выборки характеризуются тем, что каждому испытуемому одной выборки поставлен в соответствие по определенному критерию испытуемый из другой выборки.

Наиболее типичный пример зависимых выборок — повторное измерение свойства (свойств) на одной и той же выборке после воздействия (ситуация «до-после»). В этом случае выборки (одна — до экспериментального воздействия, другая — после экспериментального воздействия) зависимы в максимально возможной степени, так как они включают одних и тех же испытуемых. Могут быть и более слабые варианты зависимости. Например, мужья — одна выборка, их жены — другая выборка (при исследовании, например, их предпочтений). Или дети 5—7 лет — одна выборка, а их братья-или сестры — другая выборка.

Библиотеки Python для Data Science: продолжение.
Урок 2. Анализ данных и проверка статистических гипотез.

2. Когда применяются параметрические статистические критерии, а когда — их непараметрические аналоги?

Для формальной проверки статистических гипотез существуют различные статистические критерии. Их можно разделить на две большие группы: **параметрические** (сравнение средних значений) и **непараметрические** (сравнение рангов значений, измеренных с помощью порядковых шкал).

Параметрические критерии основаны на том, что распределение данных известно. То есть, при применении какого-нибудь параметрического критерия нужно всегда следить за тем, что главное допущение критерия – тип распределения – выполняется.

Группа статистических критериев, которые включают в расчет параметры вероятностного распределения признака (средние и дисперсии).

- Т-критерий Стьюдента
- Критерий Фишера
- Принцип максимального правдоподобия
- Критерий Романовского

В основе применения параметрических критериев сравнения лежит целый набор допущений, которым должны удовлетворять исследовательские данные (например, форма распределения выборочных статистик, равенство дисперсий, метрическая шкала зависимой переменной) для того, чтобы соответствующий критерий можно было использовать.

Как правило, многие параметрические критерии предполагают нормальность распределения данных. Во многом это связано с тем, что нормальное распределение широко распространено. Кроме того, часто все, что мы можем сказать о распределении данных, это то, является ли оно нормальным или нет, потому что задача определения типа распределения довольно сложна и существующие формальные тесты могут определить лишь общий класс распределения или показать, “между какими” распределениями находится интересующее нас распределение.

Однако, часто характеристика, подлежащая сравнению, бывает измерена в порядковой шкале. Последнее делает проверку допущений параметрических критериев бессмысленной, по причине невозможности осуществления большинства математических операций с порядковыми шкалами.
Для таких случаев существуют непараметрические аналоги параметрических критериев, не требующие соблюдения каких-либо допущений:

Непараметрические критерии исходят из того, что распределение данных незвестно. Поэтому при использовании этих критериев часто действия производятся не с самими значениями в выборке/выборках, а с их рангами.

То, что при применении тех или иных критериев нужно думать о распределении данных, не всегда означает, что перед их использованием нужно обязательно проверять распределение данных на нормальность. Иногда формальный критерий может показать, что гипотезу о нормальном распределении нужно отвергнуть, но распределение интересующего нас показателя может быть очень близким к нормальному. Поэтому главное исходить из формы распределения и тщательно анализировать данные на качественном уровне: правда ли, что показатель слишком часто принимает минимальное или максимальное значение (распределение скошено вправо или влево), правда ли, что из теоретических знаний об исследуемом показателе следует, что его распределение похоже на нормальное?

Группа статистических критериев, которые не включают в расчёт параметры и основаны на оперировании частотами или рангами.

- Q-критерий Розенбаума

- U-критерий Манна — Уитни
- Критерий Уилкоксона
- Критерий Пирсона
- Критерий Колмогорова — Смирнова