1 тема) Никитин Роман ПИ18-2

Анализ качества данных 5 баллов

репрезентативность данных

обработка некачественных данных (с выбросами и пропусками)

робастность (устойчивость к выбросам)

1) Что такое качество данных?

**А) Качество данных - обобщённое понятие, отражающее степень их пригодности к решению определённой задачи.**

Б) Качество данных - понятие, содержащее в себе критерии оценки качества данных.

В) Качество данных – возможность достоверно донести данные, информацию до конечного потребителя.

Г) Все вышеперечисленные варианты ответов.

2) Какие из нижеперечисленных проблем можно назвать основными, снижающими качество данных?

А) Пропущенные значения;

Б) Дубликаты;

В) Противоречия;

Г) Аномальные значения и выбросы;

Д) Шум;

Е) Отсутствие полноты данных;

Ё) Нарушения целостности данных;

Ж) Некорректные форматы и представления данных;

И) Фиктивные значения;

К) Ошибки ввода данных;

Л) Нарушения структуры.

**М) Все вышеперечисленные варианты.**

3) В какие два этапа борются со снижением качества данных?

А) Профайлинг— исследование данных с целью выявления проблем и выработку стратегии их решения.

Б) Очистка — применение различных методов для разрешения обнаруженных проблем: восстановление пропущенных значений, редактирование аномалий, обработка дубликатов и противоречий и т.д.

В) Аналитика - операции мысленного или реального расчленения целого на составные части

4) Какие из перечисленных проблем являются критичными?( пропущенные значения и нарушения структуры)

5) Какие из перечисленных проблем не являются критичными? (дубликаты, противоречия, шумы)

А) Пропущенные значения;

Б) Дубликаты;

В) Противоречия;

Г) Аномальные значения и выбросы;

Д) Шум;

Е) Отсутствие полноты данных;

Ё) Нарушения целостности данных;

Ж) Некорректные форматы и представления данных;

И) Фиктивные значения;

К) Ошибки ввода данных;

Л) Нарушения структуры.

6) Что такое робастность?

**А) свойство**[**статистического метода**](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B9_%D0%BC%D0%B5%D1%82%D0%BE%D0%B4)**, характеризующее независимость влияния на результат исследования различного рода**[**выбросов**](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D1%80%D0%BE%D1%81_(%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0))**, устойчивости к помехам.**

Б)  [распределение вероятностей](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9), которое в одномерном случае задаётся функцией [плотности вероятности](https://ru.wikipedia.org/wiki/%D0%9F%D0%BB%D0%BE%D1%82%D0%BD%D0%BE%D1%81%D1%82%D1%8C_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B8), совпадающей с [функцией Гаусса](https://ru.wikipedia.org/wiki/%D0%93%D0%B0%D1%83%D1%81%D1%81%D0%BE%D0%B2%D0%B0_%D1%84%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F)

7) Что такое Выбросоустойчивый (робастный) метод?

А) **Метод, направленный на выявление выбросов, снижение их влияния или исключение их из**[**выборки**](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0)**.**

Б) Метод, направленный на выявление выбросов, изучение их влияния и создания отдельной выборки выбросов для более точного анализа.

8) Перечислите два основных направления ограничения влияния неоднородностей, либо его исключения.

Группировка данных без удаления отдельных наблюдений (для снижения возможности порчи выборки отдельными выпадами). После чего с достаточной степенью уверенности допустимо использование [классических методов статистики](https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B8%D0%B5_%D0%BC%D0%B5%D1%82%D0%BE%D0%B4%D1%8B).

Отслеживание выбросов непосредственно в процессе анализа. Например, для определения параметров [закона распределения](https://ru.wikipedia.org/wiki/%D0%A0%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B2%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1%82%D0%B5%D0%B9) возможно использование [итерационной процедуры](https://ru.wikipedia.org/wiki/%D0%A7%D0%B8%D1%81%D0%BB%D0%B5%D0%BD%D0%BD%D0%BE%D0%B5_%D1%80%D0%B5%D1%88%D0%B5%D0%BD%D0%B8%D0%B5_%D1%83%D1%80%D0%B0%D0%B2%D0%BD%D0%B5%D0%BD%D0%B8%D0%B9) с усечёнными или [th](https://ru.wikipedia.org/w/index.php?title=Th-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B8&action=edit&redlink=1" \o "Th-оценки (страница отсутствует))-[сниженными M-оценками](https://ru.wikipedia.org/w/index.php?title=%D0%A1%D0%BD%D0%B8%D0%B6%D0%B5%D0%BD%D0%BD%D1%8B%D0%B5_%D0%9C-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B8&action=edit&redlink=1).

9) Что такое выбросы?

*А)***это фактически имевшие место события, вызванные исключительными условиями;**

*Б)* это, как правило, ошибки или фиктивные значения.

*10)* Что такое экстремальные значения?

*А)* это фактически имевшие место события, вызванные исключительными условиями;

*Б)* **это, как правило, ошибки или фиктивные значения.**

11) Какая основная задача редактирования выбросов?

**Во многих случаях сводится к задаче заполнения пропусков – выброс заменяется пропуском, и к нему применяются те или иные методы.**

12) Причиной выброса может быть:

А) Ошибка измерения.

Б) Необычная природы входных данных.

В) Выбросы могут быть и частью распределения — так, в [нормальном распределении](https://ru.wikipedia.org/wiki/%D0%9D%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%BE%D0%B5_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5) каждое 22-е измерение будет выходить из «[двух сигм](https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B0%D0%B2%D0%B8%D0%BB%D0%BE_%D1%82%D1%80%D1%91%D1%85_%D1%81%D0%B8%D0%B3%D0%BC)», и каждое 370-е — из трёх.

Г) **Все варианты ответа верны.**

13) Как определить выброс?

А) Простейшие способы основаны на межквартильном расстоянии — например, всё, что не попадает в диапазон считается выбросами.

Б) Критерий Шовене

В) Критерий Граббса

Г) Критерий Пирса.

Д) Критерий Диксона

Е) **Все варианты**

14) Что такое репрезентативность?

А) **Соответствие характеристик**[**выборки**](https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D0%B1%D0%BE%D1%80%D0%BA%D0%B0)**характеристикам**[**популяции**](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BF%D1%83%D0%BB%D1%8F%D1%86%D0%B8%D1%8F)**или**[**генеральной совокупности**](https://ru.wikipedia.org/wiki/%D0%93%D0%B5%D0%BD%D0%B5%D1%80%D0%B0%D0%BB%D1%8C%D0%BD%D0%B0%D1%8F_%D1%81%D0%BE%D0%B2%D0%BE%D0%BA%D1%83%D0%BF%D0%BD%D0%BE%D1%81%D1%82%D1%8C)**в целом**

Б) Способность перепрезентовать по-иному.

В) Оба варианта

15) Что такое Репрезентативная выборка?

**Выборка конечного объёма, обладающая всеми свойствами исходной популяции, значимыми с точки зрения задач исследования.**

16) Назовите необходимое условие построения репрезентативной выборки.

**Равная вероятность включения в нее каждого элемента генеральной совокупности.**

17) Какие бывают ошибки репрезентативности?

А) Случайные

Б) Систематические

В) **Оба варианта верны.**

18) Что такое Случайная ошибка репрезентативности?

***А) это отклонения характеристик выборочной совокупности от генеральных распределений*. Они носят вероятностный характер, а значит, по законам вероятности могут "гаситься" сами собой. Например, определяя какую-то характеристику выборки, мы будем получать отклонения этой характеристики то в одну, то в другую сторону от истинного значения приблизительно с одинаковой частотой, а значит, средняя ошибки будет приближаться к нулю. Эти отклонения можно оценить, но нельзя предотвратить, но они не так сильно влияют на показатели.**

*Б)* это отклонения, смещения распределения признака в одну сторону либо преувеличения, либо преуменьшения истинного значения характеристик генеральной совокупности. Сама собой эта ошибка не гасится, поэтому очень важно выявлять причины образования подобных ошибок.

19) Что такое систематическая ошибка репрезентативности?

18) Что такое Случайная ошибка репрезентативности?

*А) это отклонения характеристик выборочной совокупности от генеральных распределений*. Они носят вероятностный характер, а значит, по законам вероятности могут "гаситься" сами собой. Например, определяя какую-то характеристику выборки, мы будем получать отклонения этой характеристики то в одну, то в другую сторону от истинного значения приблизительно с одинаковой частотой, а значит, средняя ошибки будет приближаться к нулю. Эти отклонения можно оценить, но нельзя предотвратить, но они не так сильно влияют на показатели.

***Б)* это отклонения, смещения распределения признака в одну сторону либо преувеличения, либо преуменьшения истинного значения характеристик генеральной совокупности. Сама собой эта ошибка не гасится, поэтому очень важно выявлять причины образования подобных ошибок.**

20) Что такое данные с точки зрения анализа данных?

**А) зарегистрированная информация; представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, или обработки человеком или с помощью автоматических средств.**

Б) Внешние качества, которыми обладает человек. Красота, фигура, цвет кожи и тд.