Задания

1. **Предварительный статистический анализ данных.**
Проверка предположения: имеет место случайная выборка из нормального распределения.
Указание:
1) использовать выборку значений переменной (номер **E**) для одного вида ириса (номер **C**);
2) использовать выборку с засорением (переменная **G**) для вида ириса (**C**), аномальное наблюдение внести вручную (пять средних значений).

2. Корреляционный и регрессионный анализ данных
1) Корреляционный и регрессионный анализ однородных данных
Указание:
1) использовать выборку значений переменных (**F**) без засорений для одного вида ириса (**C**), в регрессионной модели зависимой является первая переменная;
2) Исследовать эффекты засорений на результаты корреляционного и регрессионного анализ, использовать выборку значений переменных (**F**) с засорением (переменная **G**) для одного вида ириса (**C**),

3. Предварительный, корреляционный и регрессионный анализ неоднородных данных
1) Оценка влияния неоднородности выборки на вероятностные свойства данных: закон распределения, корреляционные и регрессионные зависимости.
Указание: использовать выборки значений переменной (номер **E**) для нескольких видов ирисов (**D**);
2) Сравнительный анализ результатов для однородной и неоднородной выборок.

4. Дискриминантный анализ неоднородных данных
Указание: использовать выборку из смеси распределений (**D**) для всех переменных;
для обучения и экзамена использовать выборки из различных классов в пропорциях 80% и 20 %, либо процедуру кросс-валидации в тех же пропорциях; использовать алгоритмы: ЛДА, КДА, деревья решений CART и др., провести сравнительный анализ результатов.

5. Кластерный анализ неоднородных данных
Указание: использовать неклассифицированную выборку из смеси распределений (**D**) для всех переменных;
использовать алгоритмы: К-средних и иерархический кластерный анализ, провести сравнительный анализ результатов.

Общие указания.

- Для выполнения заданий использовать: пакеты STATISTICA, IBM SPSS Statistics и пакеты Python.
- Ожидаемые результаты по заданиям: анализ дескриптивных статистик, графический анализ, результаты статистической проверки гипотез с комментариями и выводами по каждому проведенному исследованию и итоговые заключение по решаемой задаче.
- Примеры отчетов прилагаются.

Данные Фишера по ирисам **(IRISDAT)**

– выборка значений основных видовых характеристик цветков ирисов на котором Рональд Фишер в 1936 году продемонстрировал работу разработанного им метода *дискриминантного анализа*[1].
Данные собраны американским ботаником Эдгаром Андерсоном[2], занимавшимся селекцией новых видов ирисов. Этот набор данных стал классическим, и часто используется в литературе для иллюстрации работы различных статистических алгоритмов[3].

1. Fisher, R.A. (1936). «The Use of Multiple Measurements in Taxonomic Problems». *Annals of Eugenics* **7**: 179–188.
2. Edgar Anderson (1935). «The irises of the Gaspé Peninsula». *Bulletin of the American Iris Society* **59**: 2–5.
3. UCI Machine Learning Repository: Iris Data Set

Данные Фишера – выборка объемом 150 наблюдений (значений характеристик), по 50 наблюдений для трех видов ирисов:

- Ирис махровый (*Iris Setosa*),
- Ирис виргинский (*Iris Virginica*),
- Ирис разноцветный (*Iris Versicolor*).

Setosa **Virginica** **Versicolor**



Посредством скрещивания двух видов (*Setosa* и *Versicolor*) получен ирис *Virginica.*

Задача анализа: является ли полученный тип ириса *Virginica* новым видом или его можно считать незначительной модификацией исходных видов.

Эта задача решается с помощью статистических методов анализа однородности и классификации выборки на три класса в пространстве 4 основных признаков.

Измеряемые признаки (характеристики) в сантиметрах:

1. Длина чашелистика (*sepal length*);
2. Ширина чашелистика (*sepal width*);
3. Длина лепестка (*petal length*);
4. Ширина лепестка (*petal width*).