**ПСМО ФКН ВШЭ, 3 курс, 2 модуль**

# Задание 3. Дисперсионный анализ.

**Прикладная статистика в машинном обучении, осень 2019**

Время выдачи задания: 18 ноября.

Срок сдачи: **2 декабря (понедельник), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x/PYTHON 3.x.

## Правила сдачи

**Инструкция по отправке:**

1. Решения задач следует присылать единым файлом формата `pdf`, набранным в LaTeX, либо в составе ipython-тетрадки в форматах `ipynb` и `html` (присылайте оба формата, т.к. AnyTask из-за высокой загрузки иногда не рендерит тетрадки в формате `ipynb` – а если мы не увидим ваши задачи, мы их не проверим). Отправляйте практические задачи в виде отдельных файлов (ipython-тетрадок или исходных файлов с кодом на языке `python`).

**Оценивание и штрафы:**

1. Максимально допустимая оценка за работу над основными задачами – 10 баллов.

2. Бонусные баллы (см. конец домашнего задания) и влияют на освобождение от задач на экзамене.

3. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.

4. Задание выполняется каждым студентом индивидуально и независимо от других студентов. «Похожие» решения считаются плагиатом и все студенты (в том числе те, у кого списали) не могут получить за него больше 0 баллов, причем обнуляются и бонусные баллы. Если вы нашли решение какого-то из заданий (или его часть) в открытом источнике, необходимо указать ссылку на этот источник в отдельном блоке в конце вашей работы (скорее всего вы будете не единственным, кто это нашел, поэтому чтобы исключить подозрение в плагиате, необходима ссылка на источник).

# Основные задачи

1. (2 балла) Найдите максимум статистики Краскелла-Уоллеса.

2. (2 балла) Получите данные из файла `figure_skating.csv`, это результаты женского фигурного катания Олимпиады-2014 года в Сочи. Сравните перцентильное и квантильное гауссовы преобразования, а также преобразование Бокса-Кокса. Для этого проведите на полученных данных тест Шапиро-Вилкса и постройте для каждого преобразования график изначальных оценок против преобразованных (QQ график). Сделайте вывод: какое преобразование вы бы предпочли для дальнейшего анализа?

3. (3 балла) Существуют две основные альтернативы классическому параметрическому дисперсионному анализу Фишера:

   - анализ Уэлча;
   - анализ Брауна-Форсайта (Brown-Forsythe).

   Для построения теста Уэлча при выборочной дисперсии $s_i^2$ в группе $i$ с $n_i$ событиями, при общем количестве групп $r$, вводятся веса $w_i = \frac{n_i}{s_i^2}$. После чего производится подсчёт следующей характеристики:

$$SSTR = \sum_{i=1}^{r} w_i (\bar{Y}_{i\cdot} - \bar{Y}_w)^2,$$

   где $\bar{Y}_w$ - взвешенное с $w_i$ среднее средних в группе $i$; $\bar{Y}_{i\cdot}$ – среднее в группе $i$. Кроме того вводится

$$\Lambda = \frac{3 \sum\limits_{i=1}^{r} \frac{\left(1 - \frac{w_i}{\sum\limits_{i=1}^{r} w_i}\right)^2}{n_i - 1}}{r^2 - 1}.$$

В этом случае, F-статистика строится:

$$F_w = \frac{SSTR/(r-1)}{1 - 2\Lambda(r-2)/3} \sim F(r-1; 1/\Lambda).$$

Сравните:

- Классический дисперсионный анализ (ANOVA).
- Дисперсионный анализ с использованием метода Уэлча.
- Дисперсионный анализ с использованием теста Краскела-Уоллиса.

Рассмотрите случай трёх выборок. Сделайте выводы о регионах применимости $F$-теста, теста Уэлча, теста Краскела-Уоллиса для однофакторного анализа.

Подсказка:
Возможные эксперименты для рассмотрения:

(a) Сбалансированные выборки, набранные из $\mathcal{N}(0; 1)$, для $n = 5, 10, 20, 100$.

(b) Сбалансированные выборки, набранных из $\mathcal{N}(0; \sigma)$, для $n = 5, 10, 20, 100$. При этом $\sigma$ для выборок – случайная тройка из элементов $1, 2, 3, 4$.

(c) Несбалансированные выборки, набранных из $\mathcal{N}(0; 1)$, для $n$ из набора $5, 10, 20, 100$.

(d) Сбалансированные выборки, со смещённым средним нормального распределения (средние могут принимать значения $0, 1, 2$)

(e) Эксперименты 1-4 для логнормального распределения.

Возможные переменные для вывода (критическое значение $\alpha = 0.05$):

- Эмпиричесткая ошибка 1-го рода.
- Мощность теста.

4. (3 балла) Исследовательская лаборатория исследует новое лекарство для лечения сенной лихорадки. В эксперименте участвует 36 волонтёров. В тесте варьировалась доза двух(2) ингредиентов (A и B), каждый из которых давался в одной из возможных доз: маленькая, средняя и большая доза. 4 волонтёра для каждой из 9 возможных концентраций подбирались случайно. Результаты эксперимента находятся в файле `fever.table`.

   (a) Используя двухфакторную модель оцените среднее, когда фактор A равен 3, а фактор B равен 2.

   (b) Используя QQ график, проверьте нормальность полученных данных. Сделайте вывод.

   (c) Проверьте наличие взаимодействия факторов A и B, используя графический способ. Сделайте вывод.

   (d) Проверьте наличие взаимодействия факторов A и B, используя F-тест с критическим значением $\alpha = 0.05$.

   (e) Проверьте наличие эффекта для каждого из факторов.