# Basic statistics – Homework

**Одним из критериев оценки решений, является наличие комментариев, описывающих процесс анализа и полученные выводы.**

## Описательные статистики

Задача:

Прежде чем применять методы дисперсионного анализа или строить регрессионные модели необходимо познакомиться с исходными данными для того, чтобы выбрать наиболее подходящий метод. Проведите описательный анализ исходных данных <https://kb.epam.com/download/attachments/570727359/data.tsv?version=1&modificationDate=1539594667371&api=v2> (немного слов о самой задаче  https://kb.epam.com/download/attachments/570727359/Description.docx?version=1&modificationDate=1539594706349&api=v2).1) (Подсчет и визуализация описательных статистик, поиск и отсечение выбросов, решение проблемы с пропусками, взаимная корреляция.)

## Стратификация (обратите внимание на презентацию в материалах)

Задача:

Предположим, что генеральная совокупность - смесь 3х нормально распределенных с.в., и можем разбить ее на 3 страты. Доли страт в генеральной совокупности разные.

Необходимо сравнить свойства точечных оценок мат.ожидания и дисперсии, полученных на основе выборок с помощью простого случайного выбора из всей генеральной совокупности, пропорционального разбиения на подсэмплы и оптимального разбиения на подсэмплы.

Рассмотреть следующие случаи:

А) все страты имеют одинаковые мат. ожидания и дисперсии

Б) страты имеют разные мат.ожидания, но одинаковые дисперсии

В) страты имеют разные и мат. ожидания и дисперсии.

Объем объединенной выборки 40 и 500.

Проверка производится моделированием (1000 повторов)

Доли страт и параметры нормальных распределений выбираете самостоятельно.

## Проверка статистических гипотез

Критерий SPRT <https://en.wikipedia.org/wiki/Sequential_probability_ratio_test>, две простых гипотезы (средние берете из моделирования)

Задача:

Создайте синтетические данные полученные путем смешивания выборок из нормально распределеных генеральных совокупностей с разными параметрами. Применить к сгенерированным данным последовательный критерий Вальда силы (α′,β′), используя приближенные границы, для проверки гипотезы H0:a=a0 при альтернативе H1:a=a1 , считая σ0 известным. Какой объем выборок потребовался в каждом случае (следует проанализировать поведение критерия при разных долях компонент в смеси)? Выясните, не менее скольки наблюдений в среднем необходимо провести, если вы будете использовать последовательный критерий силы (α′,β′), в случае справедливости нулевой и альтернативной гипотез.

## Корреляция:

Задача:

1) Есть ли взаимосвязь между размером мозга интеллектом (следует применить критерий <http://statistica.ru/theory/znachimost-koeffitsienta-korrelyatsii-doveritelnyy-interval/> для совместной выборки мужчин и женщин, а также для каждой из выборок в отдельности). Данные: <https://kb.epam.com/download/attachments/570727359/brain.tsv?version=1&modificationDate=1539594861927&api=v2>

2) Есть ли различия в выборе авто между мужчинами и женщинами (здесь следует применить <http://www.machinelearning.ru/wiki/index.php?title=%D0%A2%D0%B0%D0%B1%D0%BB%D0%B8%D1%86%D0%B0_%D1%81%D0%BE%D0%BF%D1%80%D1%8F%D0%B6%D0%B5%D0%BD%D0%BD%D0%BE%D1%81%D1%82%D0%B8>). Данные : <https://kb.epam.com/download/attachments/570727359/carprefs.tsv?version=1&modificationDate=1539594893979&api=v2>