Задания теоретического плана

Задание №1. Визуализировать закон больших чисел (на примере нахождения двойного интеграла по конечной области (можно взять квадрат [0;1]*[0;1), функцию от которой берется интеграл выберете самостоятельно)

Задание №2. Для геометрического распределения с параметром \theta «показать» свойства такие оценок как: несмещенность, состоятельность. Оценки выбрать из соответствующих задач с семинара.

Задание №3. Сгенерировать выборку из нормального распределения (разных объёмов 50, 500 и 10000) (с параметрами на Ваш выбор). И для каждого параметра постройте доверительный интервал (при этом считайте, что другой параметр либо известен, либо неизвестен), то есть должно получится 4 доверительных интервала. Что происходит с длиной доверительного интервала для разных объемов выборок?

Задания практического плана

Следует рассматривать признаки: experience level , salary in usd, remote ratio и job title (рассматривать только data analyst, data engineer, data scientist).

Во всех заданиях ниже могут помочь библиотеки numpy, scipy.stats, pandas, matplotlib и seaborn (другие библиотеки также не воспрещается использовать).

Задание №1. Для каждого из выбранных признаков выведите описательные статистики по всей выборке и по каждой должности (data analyst, data engineer, data scientist) в отдельности (для тех признаков, для которых это действие имеет смысл).

Задание №2. Для каждого из выбранных признаков постройте гистограмму (для каждой должности в отдельности). Потом визуально сравните их между собой, есть ли смысл проверять гипотезу об одинаковой распределенности данных признаков для разных должностей, если да, то проверьте это с помощью критерия хи-квадрат.

Задание №3. Для каждого из выбранных признаков постройте box-plot (для каждой должности в отдельности). Что можно сказать о выбросах/аномалиях в данных? Попытайтесь как-то объяснить с чем связано наличие выбросов/аномалий (если они есть).

Задание №4. Для признака с наибольшим числом выбросов вычислите и сравните между собой выборочное среднее и выборочную медиану.

Задание №5. Что можно сказать о распределениях каждого признака? Здесь нужно посмотреть отдельно по должностям и по всей выборке целиком. Выдвините гипотезы о распределениях признаков и проверьте их (с помощью известных Вам критериев - хи-квадрат и Колмогорова-Смирнова).

Задание №6. Для осмысленных (на Ваш взгляд) пар признаков постройте scatter-plot. Сделайте выводы о зависимости этих признаков. Найдите коэффициенты корреляции Пирсона между ними. Найдите ранговые коэффициенты корреляции между ними. Для тех пар признаков в которых наблюдается (визуально) независимость - проверьте это с помощью критериях хи-квадрат;