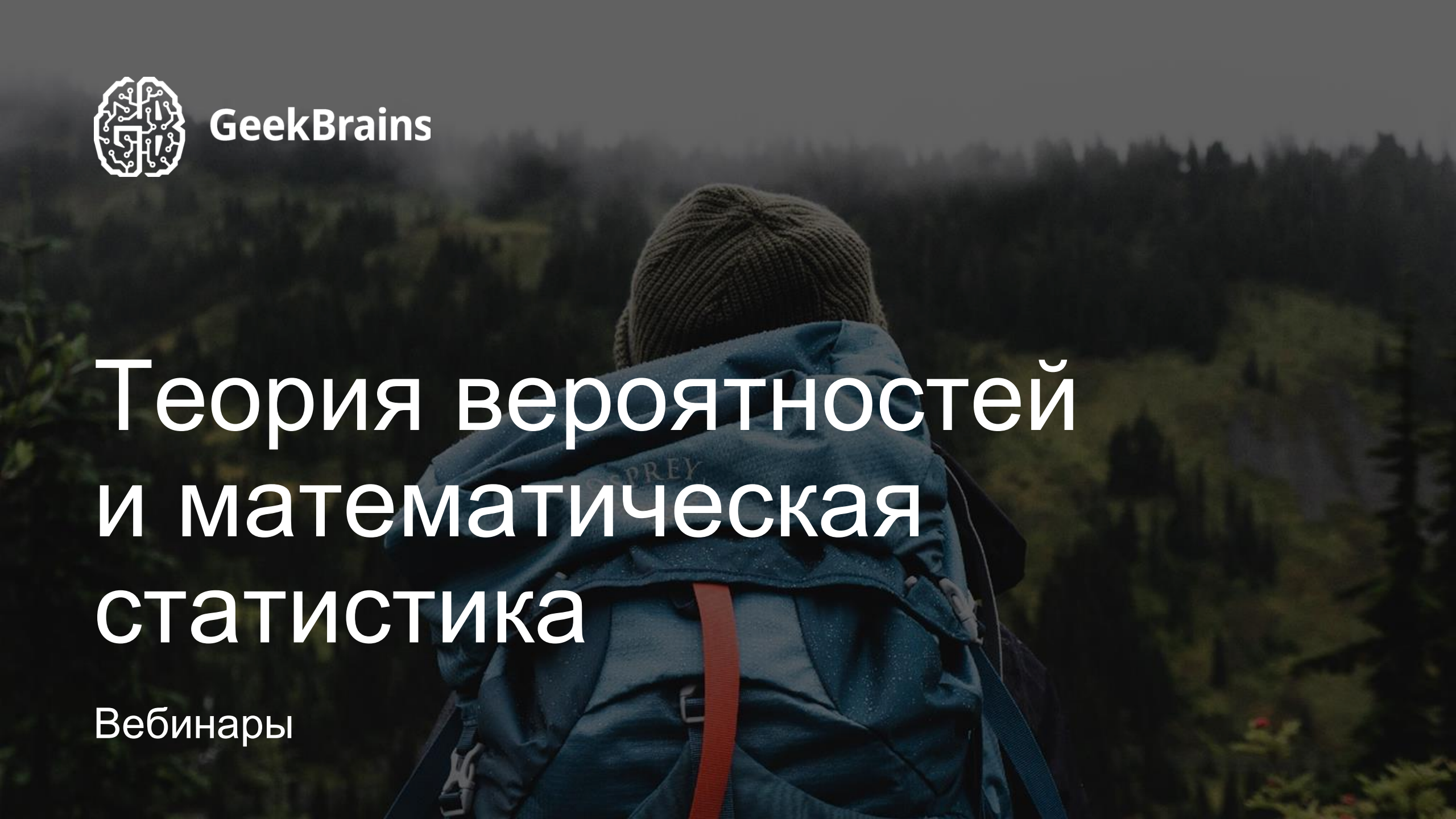
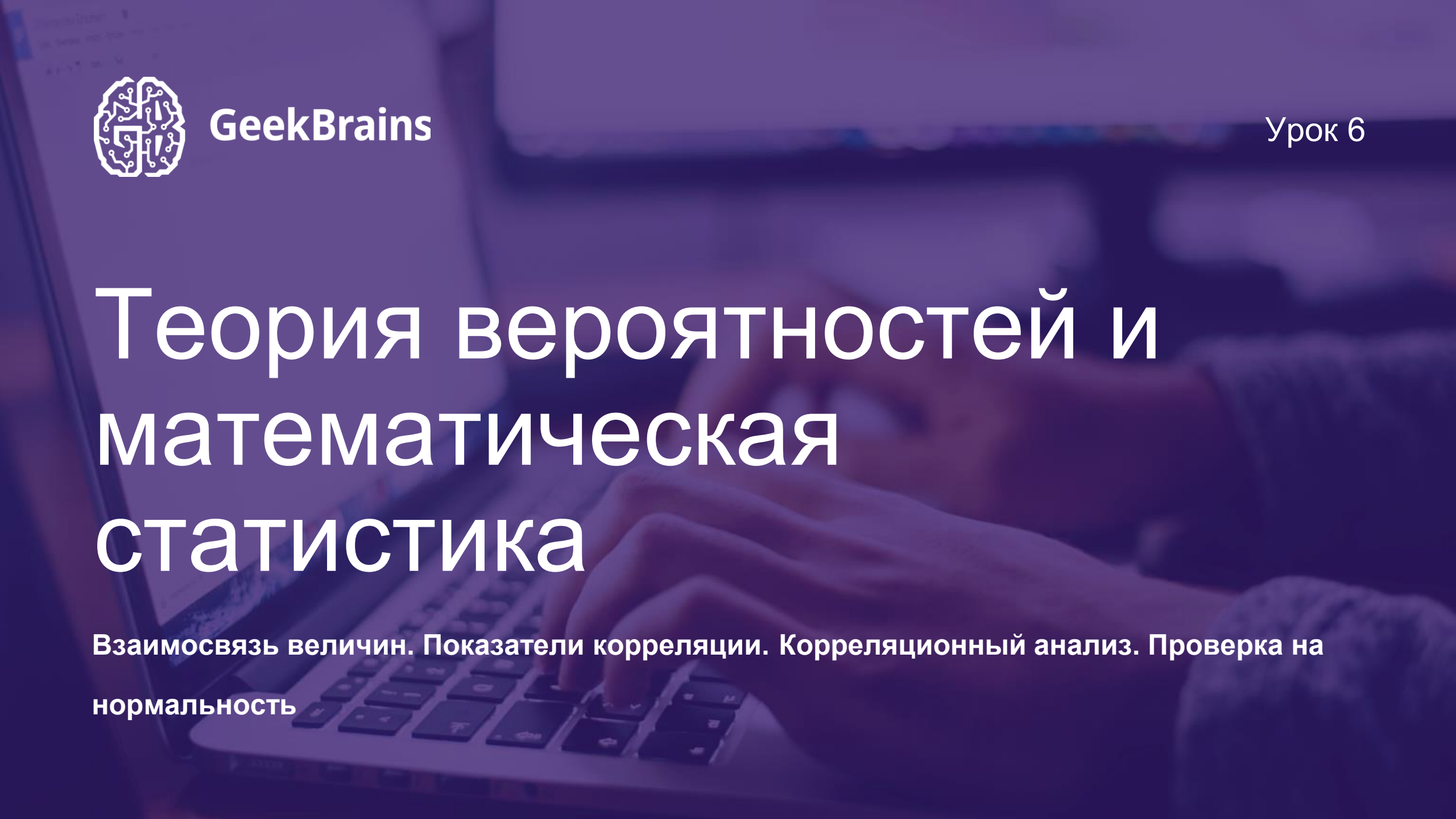


GeekBrains

Теория вероятностей и математическая статистика

Вебинары



GeekBrains

Урок 6

Теория вероятностей и математическая статистика

Взаимосвязь величин. Показатели корреляции. Корреляционный анализ. Проверка на нормальность

На этом уроке мы изучим:

1. Что такое корреляция.
2. Коэффициент корреляции.
3. Взаимосвязь величин.
4. Ковариацию.
5. Ограничения корреляционного анализа.
6. Проверку на нормальность

Курсовой проект

1. По желанию
2. Тема на выбор студента на основе пройденного материала
3. Общее направление – исследование данных с imdb
   - Разведочный анализ (EDA – exploratory data analysis)
   - Проверка статистической гипотезы
   - Корреляционный анализ
   - Регрессионный анализ
   - Дисперсионный анализ
4. Можно взять свои данные
5. Дедлайн – неделя после окончания 8 урока

Корреляция

Корреляция — математический показатель, по которому можно судить, есть ли статистическая взаимосвязь между двумя и более случайными величинами

1. Принимает значения из отрезка [−1, 1]
2. Если коэффициент корреляции близок к 1 – прямая связь
3. Если коэффициент корреляции близок к -1 – обратная связь
4. Если коэффициент корреляции равен 0 – между величинами нет связи

Корреляция величин в одной выборке не гарантирует того, что подобная связь встретится и в другой выборке и должна будет иметь такую же природу.

Высокая и низкая корреляция

1. Высокая корреляция между величинами не может быть интерпретирована как наличие причинно-следственной связи между ними
2. Высокая корреляция двух величин может свидетельствовать о том, что у них есть общая причина, несмотря на то, что прямого взаимодействия между двумя коррелирующими величинами нет
3. Напротив, отсутствие корреляции между двумя величинами еще не говорит о том, что между показателями нет связи (возможно, ее не может уловить используемый коэффициент корреляции)

Показатели корреляции

1. Ковариация — мера линейной зависимости случайных величин

$$\operatorname{cov}(X,Y) = M\left((X - M(X)) \cdot (Y - M(Y))\right)$$

Оценка ковариации бывает смещённой и несмещённой. Несмещённую оценку можно посчитать следующим образом:

$$\sigma_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X}) \cdot (y_i - \overline{Y})$$

Показатели корреляции

2. Коэффициент корреляции Пирсона использует в качестве числовой характеристики зависимости случайных величин:

$$r_{XY} = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y}$$

Здесь σX, σY — несмещённые оценки средних квадратических отклонений

Коэффициент Пирсона

1. Преимущества:
   - Использует много информации (средние и отклонения выборок)
   - Позволяет проводить тесты на значимость корреляции – статистика имеет распределение Стьюдента с n − 2 степенями свободы

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

2. Недостатки:
   - Выборки должны иметь нормальное распределение
   - Измеряет уровень линейной зависимости

Показатели корреляции

3. Ранговая корреляция

Помимо линейной зависимости существует также понятие ранговой (или порядковой) зависимости. Это тип зависимости, при котором увеличение значения одной случайной величины соответствует увеличению второй, а уменьшение первой — уменьшению второй

При ранговой зависимости не требуется чтобы степень увеличения или уменьшения двух значений были линейно зависимы

Коэффициент Кендалла и Спирмена

Коэффициент ранговой корреляции Кендалла

Рассмотрим две выборки X и Y, не имеющие повторов. Две пары (x_i , y_i) и (x_j , y_j) называются согласованными, если $x_i < x_j$ и $y_i < y_j$, или наоборот $x_i > x_j$ и $y_i > y_j$. В противном случае они называются несогласованными

P — число всех согласованных комбинаций из двух пар, а Q — число всех несогласованных комбинаций двух пар

Коэффициент корреляции Кендалла:

$$\tau = \frac{P - Q}{P + Q}$$

Коэффициент корреляции Кендалла

1. Преимущества:
   - Не требует нормальности распределений
   - Порядковая зависимость является обобщением линейной

2. Недостатки:
   - Использует меньше информации, чем коэффициент Пирсона (соответствие значений между парами элементов)
   - Прямое проведение тестов на значимость корреляции малореально

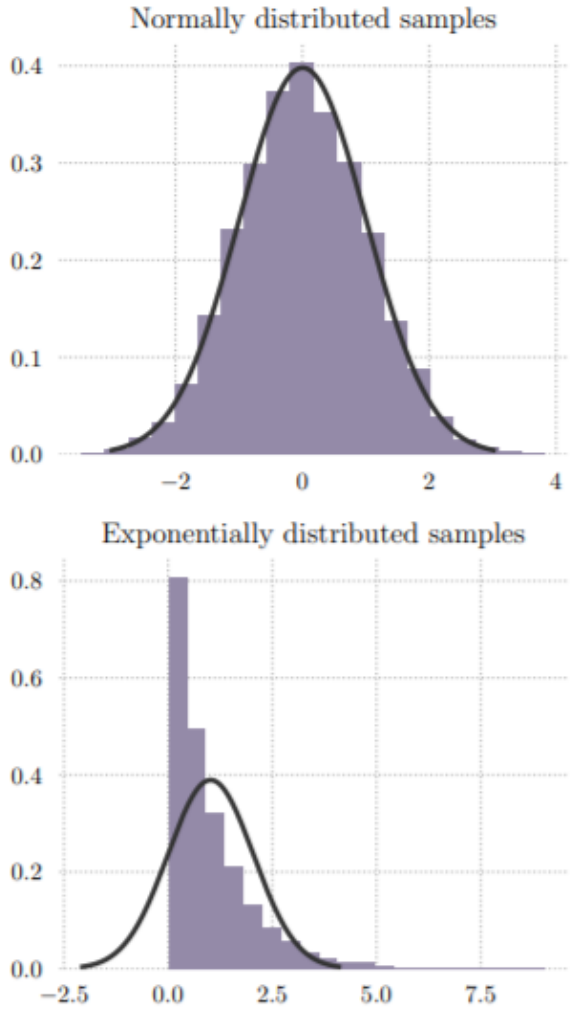
Проверка на нормальность

Методы проверки на нормальность:

1. Графические
   - Гистограмма
   - Q-Q кривая
2. На основании правил разброса
3. Статистические методы
   - Колмогорова-Смирнова
   - Шапиро-Уилка
   - Критерий согласия Пирсона

Графические методы

1. Гистограмма: по выборке можно строим гистограмму и оценить, насколько она «похожа» на гистограмму нормального распределения



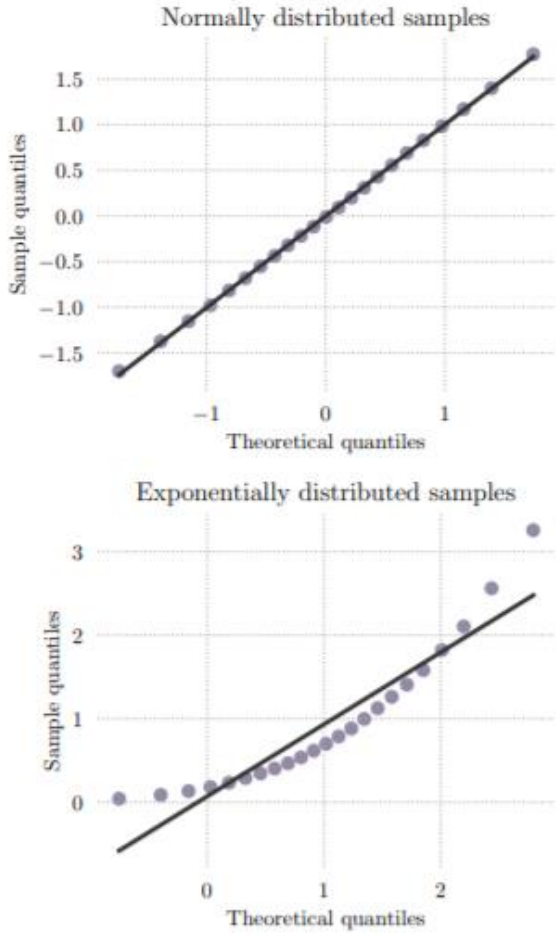
Графические методы

1. Гистограмма: по выборке можно строим гистограмму и оценить, насколько она «похожа» на гистограмму нормального распределения
2. Q-Q кривая (или кривая квантиль-квантиль):

По данной выборке считаем выборочные среднее μ и среднее квадратическое отклонение σ.

Для каждого значения α ∈ (0, 1) откладываем по оси x квантиль порядка α для нормального распределения с параметрами μ, σ, а по оси y — выборочный квантиль порядка α.

Получившийся набор точек должен лежать на прямой f(x) = x



Метод на основании правил разброса

1. Вероятность попасть в интервал от $\mu - \sigma$ до $\mu + \sigma$ равна 0.68
2. В интервал от $\mu - 2\sigma$ до $\mu + 2\sigma$ — 0.95
3. В интервал от $\mu - 3\sigma$ до $\mu + 3\sigma$ — 0.997

Данные правила должны приблизительно выполняться для выборки из нормального распределения.

Итоги

1. Что такое корреляция.
2. Коэффициент корреляции.
3. Взаимосвязь величин.
4. Ковариация.
5. Ограничения корреляционного анализа.
6. Как проверить нормальность распределения