Теория вероятностей и математическая статистика

Вебинар 6

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

Многомерный статистический анализ

Часто в статистике анализируют *многомерные* наблюдения, т.е. имеющие несколько признаков.

Для этого применяют многомерный статистический анализ. Особенно часто его используют, когда нужно:

- изучить зависимость между признаками и их влияние на некоторую переменную,
- классифицировать объекты с множеством признаков,
- понизить размерность пространства признаков (если их слишком много и нет возможности отсеять часть).

Корреляция

Корреляция

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

Коэффициент корреляции принимает значения из отрезка $[-1, 1]$.

Корреляция

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

Коэффициент корреляции принимает значения из отрезка $[-1, 1]$.

Если коэффициент корреляции близок к 1, то между величинами наблюдается прямая связь: увеличение одной величины сопровождается увеличением другой, а уменьшение одной — уменьшением другой.

Если же коэффициент корреляции близок к -1, то между величинами есть обратная корреляционная связь: увеличение одной величины сопровождается уменьшением другой и наоборот.

Коэффициент корреляции, равный 0, говорит о том, что между величинами нет связи, то есть величины изменяются независимо друг от друга.

Взаимосвязь величин

Если две величины коррелируют, это может свидетельствовать о наличии статистической связи между ними.

Однако, говорить о ней мы можем только для величин из одной выборки. Корреляция величин в одной выборке не гарантирует того, что подобная связь встретится и в другой выборке и должна будет иметь такую же природу.

Высокая корреляция

Высокая корреляция между величинами не может быть интерпретирована как наличие причинно-следственной связи между ними.

Например, если рассмотреть данные о пожарах в городе, можно увидеть, что между материальными потерями, вызванными пожаром, и количеством пожарных, которые принимали участие в его тушении, есть сильная корреляция. При этом ложным будет вывод о том, что большое количество пожарных, присутствующих на пожаре, приводит к увеличению ущерба от него.

Высокая корреляция

Высокая корреляция между величинами не может быть интерпретирована как наличие причинно-следственной связи между ними.

Например, если рассмотреть данные о пожарах в городе, можно увидеть, что между материальными потерями, вызванными пожаром, и количеством пожарных, которые принимали участие в его тушении, есть сильная корреляция. При этом ложным будет вывод о том, что большое количество пожарных, присутствующих на пожаре, приводит к увеличению ущерба от него.

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.

Например, наступление зимы может быть причиной и роста заболеваемости простудой, и повышения расходов на отопление. Эти две величины (число заболевших и расходы на отопление) имеют высокую корреляцию между собой, хотя они друг на друга напрямую не влияют.

Низкая корреляция

Напротив, отсутствие корреляции между двумя величинами еще не говорит о том, что между показателями нет связи.

Вполне возможно, что между признаками есть нелинейная зависимость, которую не может уловить используемый коэффициент корреляции.

Показатели корреляции

Ковариация

Ковариация — мера линейной зависимости случайных величин. Её формула похожа на формулу дисперсии (*variance*). Формула ковариации случайных величин X и Y:

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

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

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

Здесь X, Y — выборки размера n.

Коэффициент корреляции Пирсона

В качестве числовой характеристики зависимости случайных величин используют *коэффициент корреляции Пирсона*:

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

Здесь σ_X, σ_Y — несмещённые оценки средних квадратических отклонений.

Использование коэффициента Пирсона

Плюсы:

- Использует много информации (средние и отклонения выборок),
- Позволяет проводить тесты на значимость корреляции: статистика

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

имеет распределение Стьюдента с $n-2$ степенями свободы.

Минусы:

- Выборки должны иметь нормальное распределение,
- Измеряет уровень *линейной зависимости*.

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.

Однако, в отличие от линейной зависимости, при ранговой зависимости не требуется чтобы степень увеличения или уменьшения двух значений были линейно зависимы.

Высокое значение ранговой корреляции означает, что если отсортировать два массива по возрастанию первого, то второй также будет возрастать.

Популярными коэффициентами ранговой корреляции являются коэффициент Кендалла и коэффициент Спирмана. Мы здесь рассмотрим первый из них.

Коэффициент ранговой корреляции Кендалла

Допустим, $(x_1, y_1), \ldots, (x_m, y_m)$ — все пары значений двух выборок. 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*.

Пусть P — число всех согласованных комбинаций из двух пар, а Q — число всех несогласованных комбинаций двух пар. *Коэффициент корреляции Кендалла*:

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

Замечание. Такое определение коэффициента Кендалла возможно только если выборки X и Y не имеют повторов. Существуют уточнённые версии коэффициента Кендалла, допускающие повторы, но мы их здесь не приводим.

Использование коэффициента Кендалла

Плюсы:

- Не требует нормальности распределений,
- Порядковая зависимость является обобщением линейной.

Минусы:

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

Проверка на нормальность

Проверка на нормальность

Ранее мы неоднократно отмечали, что применимость того или иного метода сильно зависит от того, является ли распределение в нашей выборке нормальным.

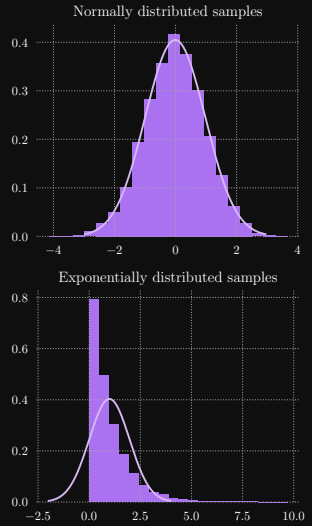
Методы проверки на нормальность делятся на 3 класса:

1. Графические методы:
   - Гистограмма,
   - Q-Q кривая,
2. Методы на основании правил разброса (стандартное отклонение, 2 сигмы, 3 сигмы),
3. Статистические методы:
   - Колмогорова-Смирнова,
   - Шапиро-Уилка,
   - ...

Проверка на нормальность. Графические методы

Как следует из названия, графическе методы используют для проверки на нормальность различные графики и диаграммы.

Например, по выборке можно построить гистограмму и оценить, насколько она «похожа» на гистограмму нормального распределения.

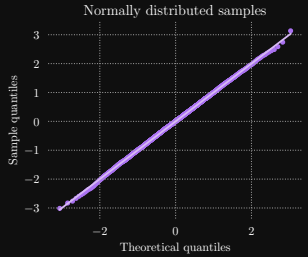


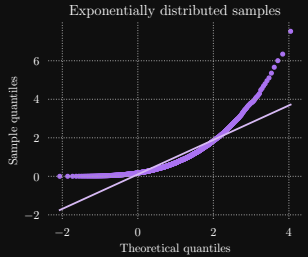
Проверка на нормальность. Графические методы

Другой способ графически оценить нормальность выборки — *Q-Q кривая* (или *кривая квантиль-квантиль*):

1. По данной выборке считаем выборочные среднее μ и среднее квадратическое отклонение σ.
2. Для каждого значения $\alpha \in (0, 1)$ 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 $f(x) = x$. Оценить близость получившихся точек к данной прямой можно с помощью *парной регрессии* (об этом — на следующем занятии).





Проверка на нормальность. Методы на основании правил разброса

Ещё один способ оценить нормальность выборки — известные нам правила разброса для нормального распределения:

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

Данные правила должны приблизительно выполняться для выборки из нормального распределения.

Проверка на нормальность. Статистические методы

Наконец, существует несколько статистических методов, позволяющих «по-честному» проверить гипотезу о нормальности распределения. Например, часто для таких задач используются метод Колмогорова-Смирнова или метод Шапиро-Уилка.

Однако, как правило, такие методы основаны на использовании статистик, имеющих в предположении верности нулевой гипотезы крайне сложные распределения, квантили которых невозможно посчитать напрямую, а только лишь с помощью приближённых методов.

На следующем занятии

Линейная регрессия. Однофакторный дисперсионный анализ. A/B-тестирование