**64 Понятие корреляции, коэффициент корреляции Пирсона, Спирмена, Кендела.**

Корреляция - это статистическая мера, которая показывает, насколько сильно и в каком направлении две или более переменных связаны между собой. Более конкретно, корреляция измеряет степень, в которой изменения в одной переменной сопутствуют изменениям в другой переменной.

Коэффициент корреляции Пирсона - это один из наиболее распространенных коэффициентов корреляции. Он измеряет линейную связь между двумя непрерывными переменными. Значение коэффициента корреляции Пирсона находится в диапазоне от -1 до 1, где отрицательное значение указывает на обратную (отрицательную) корреляцию, положительное - на прямую (положительную) корреляцию, а значение 0 указывает на отсутствие корреляции.

Коэффициент корреляции Спирмена - это другой распространенный коэффициент корреляции. Он измеряет не только линейную, но и монотонную связь между двумя переменными, то есть связь, которая может быть нелинейной.

Коэффициент корреляции Кендела - это еще один тип коэффициента корреляции, который измеряет монотонную связь между двумя переменными.Коэффициент корреляции Кендела чувствителен к выбросам в данных, но не чувствителен к тому, являются ли данные нормально распределенными.

**65 Каковы факторы, влияющие на коэффициент корреляции?**

Наличие выбросов в данных: Выбросы в данных могут сильно исказить результаты коэффициента корреляции, поскольку они могут создавать ложные связи между переменными.

Размер выборки: Обычно коэффициент корреляции более точен при большом размере выборки, поскольку большая выборка обеспечивает более представительную выборку.

Характеристики данных: Коэффициент корреляции может быть чувствителен к типу данных, который используется для измерения переменных.

Тип используемого коэффициента корреляции: Коэффициент корреляции может изменяться в зависимости от того, какой тип коэффициента корреляции используется (например, Пирсона, Спирмена, Кендела и т.д.).

Вид связи между переменными: Тип связи между переменными также может повлиять на коэффициент корреляции. Например, если связь между переменными нелинейна, то коэффициент корреляции Пирсона может быть менее точным, чем коэффициент корреляции Спирмена или Кендела.

Выборка: Случайность в выборке может сильно влиять на результаты коэффициента корреляции. Повторное измерение на разных выборках может дать различные результаты коэффициента корреляции.

**66 Назовите виды связи между переменными при корреляции.**

Линейная связь: это наиболее распространенный вид связи, при котором изменения одной переменной линейно связаны с изменениями другой переменной.

Нелинейная связь: это связь, которая не может быть описана линейной функцией, а может иметь различные формы, такие как квадратичная, экспоненциальная, логарифмическая и т.д.

Положительная связь: это связь, при которой обе переменные изменяются в одном направлении. То есть, если одна переменная увеличивается, то и другая переменная также увеличивается.

Отрицательная связь: это связь, при которой обе переменные изменяются в противоположных направлениях. То есть, если одна переменная увеличивается, то другая переменная уменьшается.

Нулевая связь: это связь, при которой между переменными нет никакой корреляционной связи, то есть изменение одной переменной не связано с изменением другой переменной.

Спуриозная связь: это связь, которая появляется случайно и не имеет никакого реального значения. Спуриозная связь может возникать, например, при использовании неправильных методов статистического анализа или при наличии выбросов в данных.

**67 Что такое регрессионный анализ, какие задачи DM можно проводить с его помощью?**

Регрессионный анализ - это статистический метод, который используется для оценки связи между зависимой переменной и одной или несколькими независимыми переменными. Главная цель регрессионного анализа - определить, какие из независимых переменных наиболее сильно влияют на зависимую переменную и каким образом.

DM (Data Mining) специалисты могут проводить следующие задачи с помощью регрессионного анализа:

Прогнозирование: регрессионный анализ позволяет строить модели, которые могут быть использованы для прогнозирования значений зависимой переменной на основе независимых переменных.

Определение влияния: регрессионный анализ может помочь определить, какие из независимых переменных наиболее сильно влияют на зависимую переменную.

Оценка эффективности: регрессионный анализ может быть использован для оценки эффективности программ или мероприятий.

Идентификация выбросов: регрессионный анализ может помочь идентифицировать выбросы или наблюдения, которые нарушают предположения модели.

Изучение взаимодействия: регрессионный анализ может помочь изучить взаимодействие между независимыми переменными.

**68 Какие способы визуализации корреляции были изучены в курсе Big Data?**

Диаграмма рассеяния (scatter plot) - показывает распределение значений двух переменных и их взаимосвязь. Если переменные сильно коррелируют, то точки на диаграмме будут располагаться близко к прямой линии.

Коррелограмма (correlogram) - это матрица диаграмм рассеяния для всех пар переменных. Это позволяет быстро определить, какие пары переменных имеют сильную корреляцию.

Тепловая карта (heatmap) - показывает коэффициенты корреляции между всеми парами переменных в виде цветовой карты. Это позволяет быстро определить, какие пары переменных имеют сильную корреляцию и какая это связь (положительная или отрицательная).

Линейный график (line plot) - показывает изменение значения переменной во времени. Если переменная коррелирует с другой переменной, то изменение одной переменной может быть связано с изменением другой переменной в течение времени.