# **Основы анализа данных** Обычно на начальном этапе статистического анализа вычисляют описательную статистику, рассчитывают корреляции и строят линейную регрессию.

1. Описательная статистика позволяет собрать и просуммировать численные данные для придания им более удобной для понимания формы.

Описательная статистика непрерывно связана с такими характеристиками, как среднее, стандартная ошибка и отклонение, мода, медиана, минимум и максимум, сумма, счёт, асимметричность, интервал, эксцесс и дисперсия выборки.

Центральная тенденция – мера, которая лучшим образом описывает какой-либо признак всего набора. Характеристиками этой меры являются среднее значение и медиана. Первая характеристика рассчитывается как среднее арифметическое значение, а вторая – середина всей выборки, которая делит её на две равные части.

Свойства среднего значения:

* не допускаются пропуски в данных;
* вычисляется только для числовых переменных и дихотомических шкал;
* для одного набора данных только одно среднее значение.

Важное условие для расчета медианы – упорядоченная выборка.

Существуют некоторые характеристики вариации данных: одними из самых простых являются минимум и максимум; размах – разница значений минимума и максимума, дисперсией является среднее арифметическое квадратов отклонений от среднего; стандартное отклонение рассчитывается как квадратный корень из дисперсии; асимметрия – отклонение распределения от симметричного; выбросы – это такие данные, которые резко отличаются от всего набора; эксцесс – острота или сглаженность распределения относительно нормального.

1. Корреляционный анализ

Количественная оценка взаимосвязей пары наборов данных называется корреляционным анализом, эта характеристика может помочь определить ассоциативность наборов по величине. Коэффициент корреляции может варьироваться в диапазоне от -1 до +1.

Всякая зависимость между переменными имеет следующие свойства:

* величина;
* надёжность.

Если величина зависимости велика, то велика и зависимость между переменными.

Свойство надёжности показывает, может ли быть подобная зависимость между другими наборами данных.

1. Регрессионный анализ

С помощью этого вида анализа возможно просмотреть форму и характер зависимости между переменными. Регрессионный анализ используется в задачах классификации и прогнозирования.

Различный характер и форма зависимости переменных определяет тип регрессии: положительная и отрицательная линейная регрессия, положительная равноускоренная и равнозамедленная возрастающая регрессия, отрицательная равноускоренная и равнозамедленная регрессия.

Такие типы регрессии зачастую можно встретить в смешанной форме.

Регрессионный анализ помогает оценивать, насколько зависимы переменные при помощи вычисления одного значения на основе нескольких известных.

Уравнение регрессии:

*(1)*

a – свободный член

b – коэффициент регрессии

Остаток – отклонение наблюдения от предсказанного значения.

R – квадрат – мера определённости, характеризующая качество регрессионной прямой, которая находится в пределе от 0 до 1 включительно.