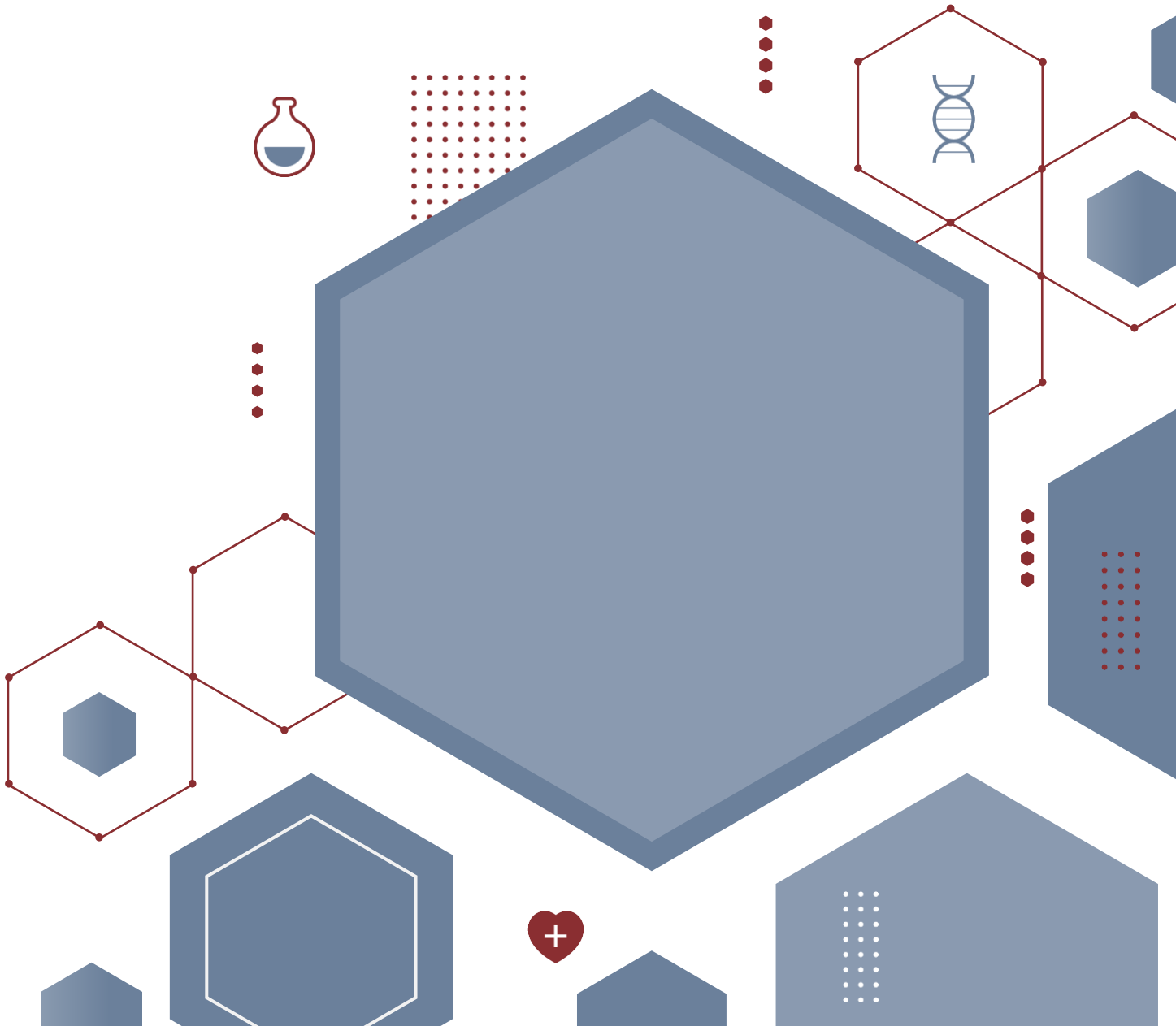
Казанский
Государственный
Медицинский
Университет

Описательная статистика.

Лакман Ирина Александровна

Казань, 2025



КГМУ

План лекции

- Анализ описательных статистик
- Бокс-плоты
- Восполнение пропусков
- Виды аномалий
- Методы обнаружения аномалий в R
- Методы импутации

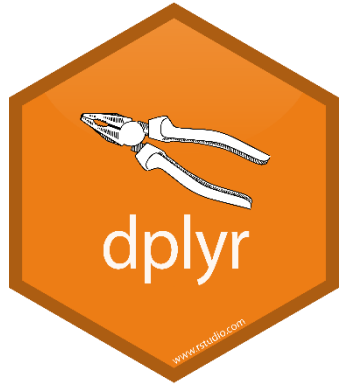
Анализ описательных статистик

Описательный (дескриптивный) анализ данных – нужен для предварительной оценки распределения признаков, для того чтобы в дальнейшем выбрать корректный метод обработки информации.

Описательная статистика (descriptive statistics) или разведочный анализ данных – это статистические методы обработки данных, их систематизации, расчета основных описательных статистик:

Для числовых признаков – показателей центра распределения, показателей вариации, показателей асимметрии и эксцесса распределения.

Для качественных характеристик – абсолютной и относительной (в %) частоты встречаемости признака.



Статистическое распределение признака

Определение: Совокупность значений признака с указанием числа их различных значений называется **распределение признака.**

Определение: Наблюдаемые значения признака X_i называют **вариантами**, а последовательность вариант, записанных в возрастающем порядке – **вариационным рядом.** Число наблюдений n_i (попавших в *i*-ую группу) называют **частотами**, а их отношение к объему выборки $\frac{n_i}{n} = W_i$ называют **относительными частотами** (**частость**),

$$\sum_{i=1}^{k} n_i = n \qquad \sum_{i=1}^{k} W_i = 1$$

где n - общее число наблюдений, k – число групп в совокупности.

Определение: **Накопленная частота (кумулятивная)** S_i характеризует объем совокупности со значением вариантов, не превышающих X_i :

$S_1 = n_1$; $S_2 = n_1 + n_2$;

$S_3 = n_1 + n_2 + n_3$; и т.д.

Определение: **Плотность частот** это частота, приходящаяся на единицу интервала,

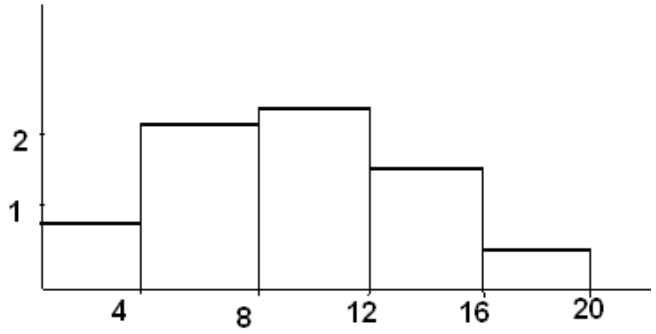
если все дины интервалов одинаковые h, то плотность частот, $\frac{n_i}{h} = q_i$

Если все длины интервалов разные h_i , то плотность частот $\frac{n_i}{h_i} = q_i$

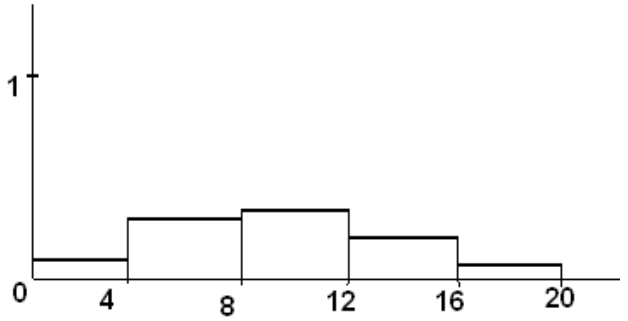
Статистическое распределение признака

Для наглядности статистического распределения в случае дискретного распределения признака ***X*** строят **полигон** (ломанная, где длина ***X*** откладывается на оси абсцисс, а на оси ординат соответствующие им частоты n_i).

В случае непрерывного распределения признака ***X*** строят гистограммы. Для построения гистограммы все наблюдаемые значения признака разбивают на несколько ***i*** частичных интервалов длиной ***h***, и для каждого интервала сумму частот вариант попавших в ***i*** интервал отмечают по оси ординат. Гистограммой распределения частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиной ***h***, а высоты равны отношению $\frac{n_i}{h}$ (плотность частот).



Гистограмма распределения плотности частот признака



Гистограмма распределения относительных частот признака

Показатели центра распределения

Определение: Средняя величина $\overline{X}$ характеризует типичный уровень признака в совокупности

$$\overline{X} = \frac{(x_1 + x_2 + \ldots + x_N)}{N}$$

где *N* – объем совокупности.

Или на основе частот и частностей

$$\overline{X} = \frac{(x_1 n_1 + x_2 n_2 + \ldots + x_k n_k)}{N} = \sum_{i=1}^{k} x_i \cdot W_i$$

k – число групп в совокупности.

При интервальном распределении признака, в качестве X_i берутся середины интервалов.

Определение: Модой называется наиболее часто наблюдаемая величина признака, обозначается M_O. Для дискретного ряда мода равна максимальной частоте, для интервального вариационного ряда определяется модальный интервал по наибольшей частоте, а мода определяется как:

$$M_o = x_0 + i \cdot \frac{(n_{Mo} - n_{Mo-1})}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})}$$

Здесь $n_{Mo}, n_{Mo-1}, n_{Mo+1}$ - частоты модального, предмодального и постмодального интервалов; а X_0 и *i* – нижняя граница и величина модельного интервала.

Показатели центра распределения

Определение: **Медианой** называется значение наблюдения, которое находиться в середине ранжированного ряда данных, т.е. наблюдение, занимающее серединное значение, обозначается как $\boldsymbol{M_e}$.

Для интервального вариационного ряда определяется: медианный интервал серединный интервал ряда, а медиана определяется как:

$$M_e = x_0 + i \cdot \frac{1/2 \sum_{i=1}^{k} n_i - S_{Me-1}}{n_{Me}}$$

Здесь $\boldsymbol{n_{Me}}$ - частота медианного интервала; $\boldsymbol{S_{Me-1}}$ - кумулятивная частота предмедианного интервала; а $\boldsymbol{X_0}$ и $\boldsymbol{i}$ – нижняя граница и величина медианного интервала.

Порядковые статистики

Определение: **Квартили** *Q* - это значения вариантов признака, которые делят вариационный ряд по объему на четыре равные части.

Первый и третий квартиль: $$Q_1 = x_{Q_1} + i \cdot \frac{0{,}25\sum_{i=1}^{k} n_i - S_{Q_1-1}}{n_{Q_1}}$$ $$Q_3 = x_{Q_3} + i \cdot \frac{0{,}75\sum_{i=1}^{k} n_i - S_{Q_3-1}}{n_{Q_3}}$$

Второй квартиль равен медиане.

Определение: **Децили** *D* - это значения вариантов признака, которые делят вариационный ряд по объему на десять равных частей, а **процентили** *P* соответственно на 100.

Первая дециль: $$D_1 = x_{D_1} + i \cdot \frac{0{,}1\sum_{i=1}^{k} n_i - S_{D_1-1}}{n_{D_1}}$$

Первая процентиль

$$P_1 = x_{P_1} + i \cdot \frac{0{,}01\sum_{i=1}^{k} n_i - S_{P_1-1}}{n_{P_1}}$$

Показатели вариации

Определение: **Дисперсией** называют среднее арифметическое квадратов отклонений признака совокупности от ее среднего значения.

Для несгруппированных данных:

$$D^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

Для сгруппированных данных:

$$D^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 \cdot n_i}{\sum_{i=1}^{k} n_i}$$

Используют также более простую формулу: $D^2 = \overline{x^2} - \overline{x}^2$

Определение: **Вариационный размах** R показывает насколько велико различие между единицами совокупности. $R = x_{\max} - x_{\min}$

Для нетипичных значений признака используют **квартильный размах**: $R_Q = Q_3 - Q_1$

Показатели вариации

Определение: **Среднеквадратичное отклонение** определяется соответственно для сгруппированных и несгруппированных данных по формуле: $\sigma = \sqrt{D^2}$.

Среднеквадратичное отклонение дает неискаженное представление о б отклонении, в отличие от дисперсии.

Для колеблиемости данных вводят также:

Коэффициент осцилляции: $V_R = \frac{R}{\overline{x}} \cdot 100\%$

Коэффициент вариации: $V_\sigma = \frac{\sigma}{\overline{x}} \cdot 100\%$

Квартильный коэффициент вариации:

$$V_Q = \frac{Q_3 - Q_1}{2Me}$$

Децильный коэффициент вариации:

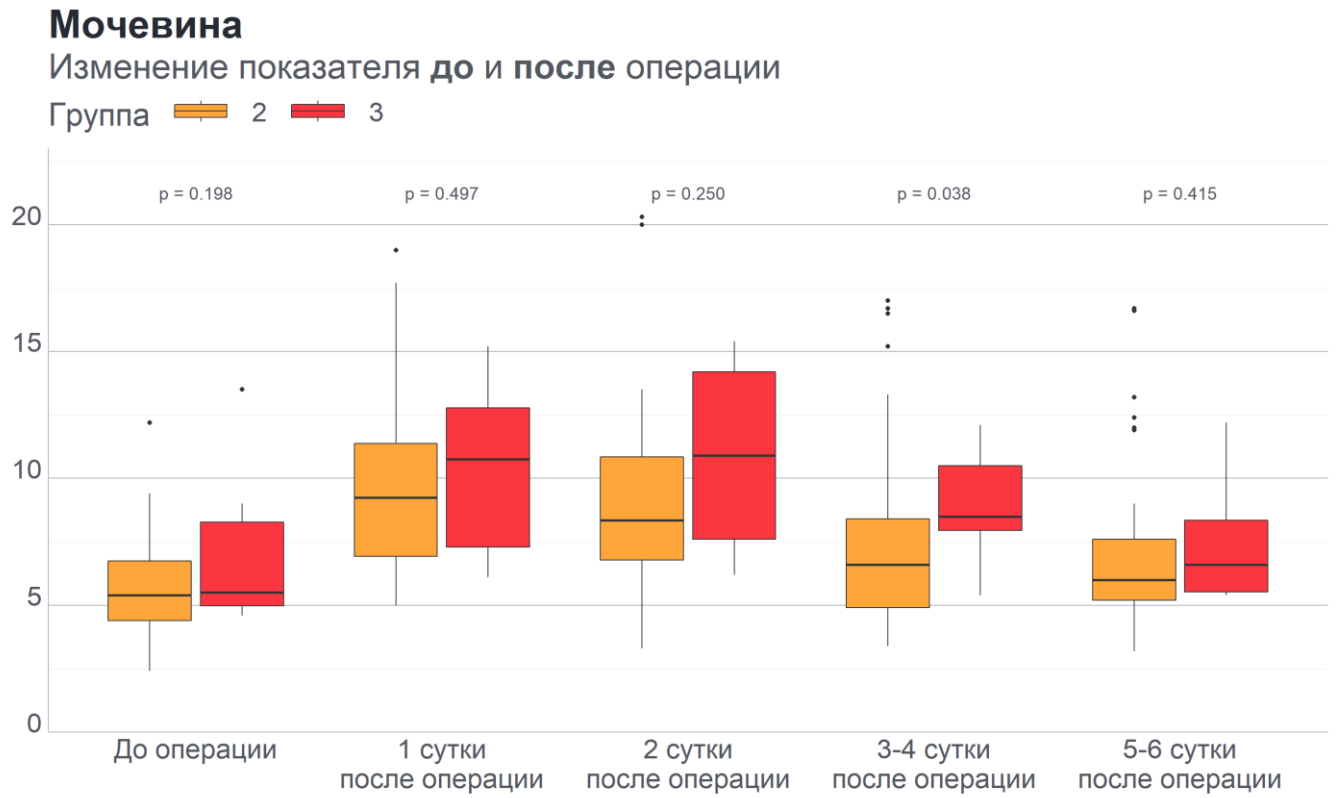
$$V_D = \frac{D_9}{D_1}$$

Бокс-плоты (ящики с усами)

Это диаграммы, на которых наглядно видно распределение признака.

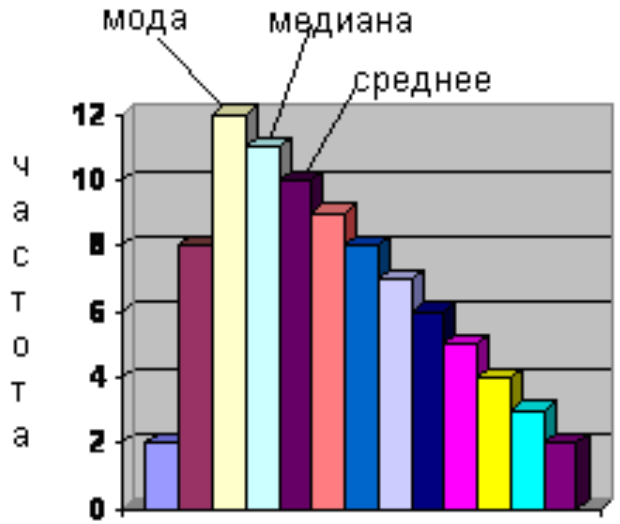
Варианты: границы – первая и третья квартили в середине – медиана

границы – среднее ±стандартное отклонение.



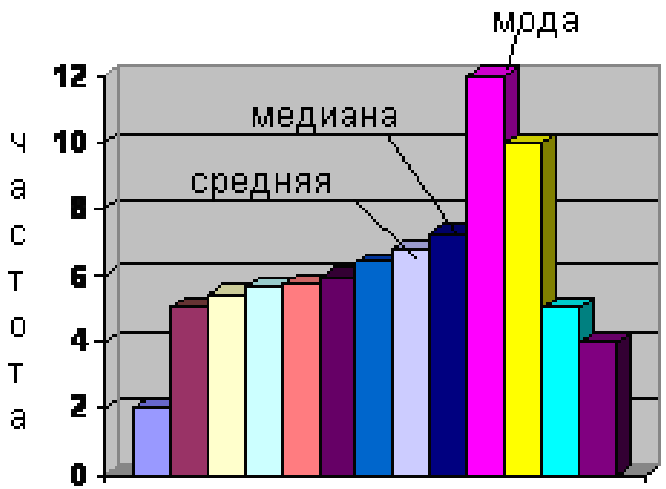
Асимметрия

Асимметрия может быть как положительной, так и отрицательной. Когда асимметрии нет, то говорят, что сдвиг в рассеянии данных отсутствует.



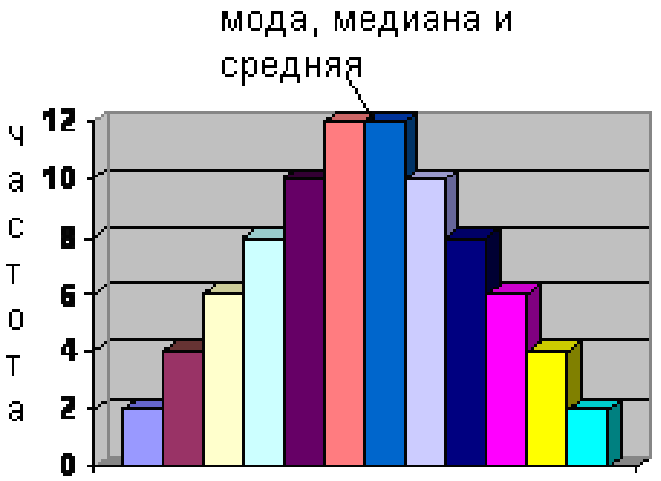
положительная асимметрия

(средняя > медиана)



отрицательная асимметрия

(средняя < медиана)



Симметрично (сдвига нет)

Показатели асимметрии

Определение: Коэффициент асимметрии показывает, есть ли смещение (скошенность) в рассеянии данных.

Коэффициент асимметрии определяется:

$$КА = \frac{\frac{\sum(x - \overline{X})^3}{n-1}}{\left(\sqrt{\frac{\sum(x-\overline{X})^2}{n-1}}\right)^3}$$

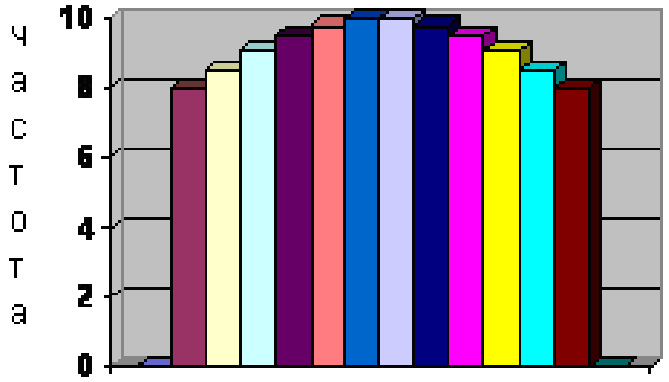
Коэффициент асимметрии Спирмена определяется как:

$$КА_{спир} = \frac{3(средняя\_арифметическая - медиана)}{среднее\_квадратическое\_оклонение}$$

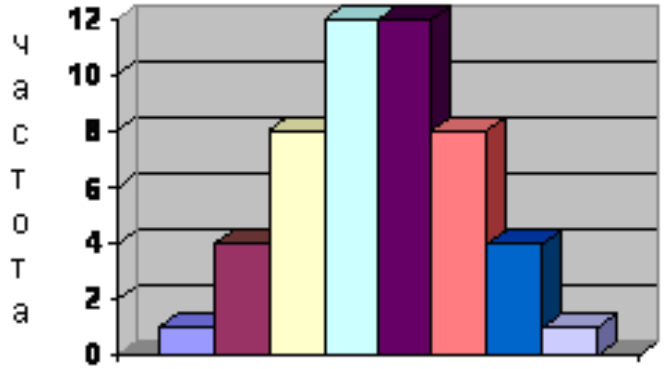
Коэффициент асимметрии является моментом третьего порядка

Эксцесс

Определение: **Показатель эксцесса** описывает «пиковость» распределения частот. Распределения, имеющие более выраженный пик, называются островершинными. Те же распределения, у которых степень вытянутости вдоль оси ординат меньше, называют плосковершинными.



Плосковершинное

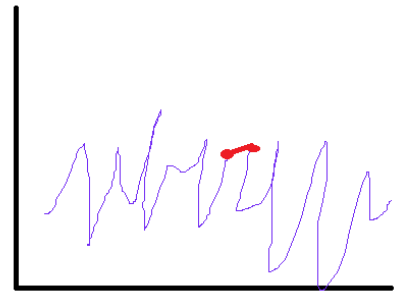


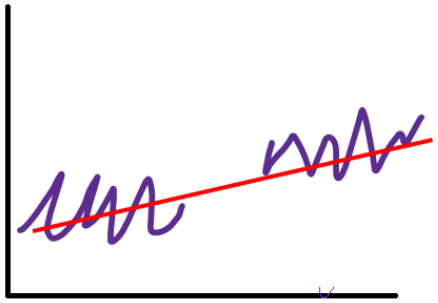
Островершинное

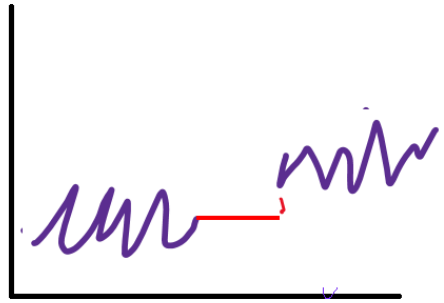
Коэффициент эксцесса определяется по формуле:

$$КЭ = \frac{\frac{\sum(x - \overline{X})^4}{n-1}}{\left(\sqrt{\frac{\sum(x-\overline{X})^2}{n-1}}\right)^4}$$

Восполнение пропусков







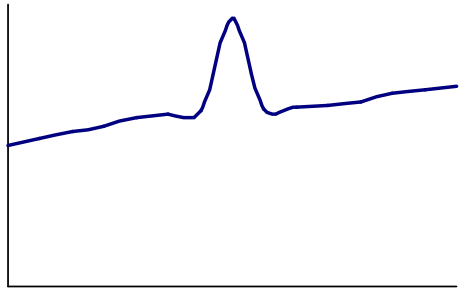
Пропуски почти всегда встречаются в данных!!!

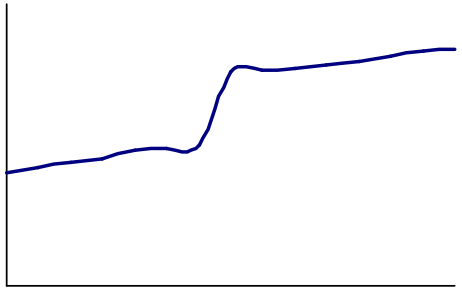
Если пропуски в кросс-секционных данных, то наблюдения с пропущенными данными можно удалить, либо заменить средним значением или медианой (в случае если признак числовой).

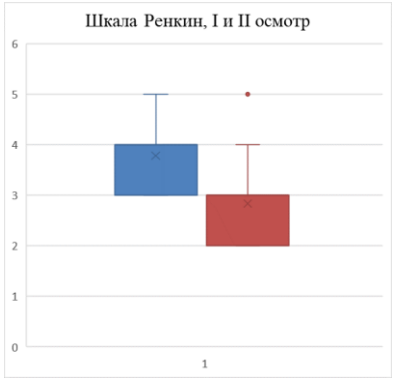
Методы восполнения пропусков во временных рядах

1. Замена средним соседних (справа и слева) от пропуска значений (плохо подходит, если подряд много пропущенных значений)
2. Замена средним нескольких соседних (справа и слева) от пропуска значений
3. Замена медианой нескольких соседних (справа и слева) от пропуска значений
4. Замена пропущенных значений линейной функцией по времени (трендом) – метод может исказить тип процесса и привести в итоге к плохому прогнозу
5. Замена пропущенных значений последним значением (протяжка) – метод используется чаще всего
6. Замена нулем (в случае только, когда это согласуется с смыслом временного ряда)

Виды аномалий







Аномальные наблюдения встречаются как в кросс-секционных данных, так и во временных рядах.

Как определять аномалии:

1. На основе правила 3σ: рассчитывается среднее значение m и стандартное отклонение σ, все что выходит за интервал m±3σ – считается аномалией. Данный метод подходит для нормально распределённых признаков
2. Для ненормально распределенных признаков:

2.1 метод 3 IQR: все что выходит за интервал трех межквартильных размахов считается аномалией

2.2 все что соответствует 1, 2,5 или 5 крайним перцентилям – аномалия

Аномалии из кросс-секционных данных обычно отбрасывают или выделяют в отдельный признак.

Во временных рядах аномалии бывают 2-х типов:

1) отклонения, носящие объективный характер и не оказывающие влияние на дальнейший ход развития процесса: из временного ряда удаляют и восполняют пропуск

2) отклонения, носящие объективный характер и оказывающие влияние на дальнейший ход развития процесса: с ними ничего не делают, учитывают в моделях АВР

Методы обнаружения аномалий в R

AnomalyDetection — пакет в R для обнаружения аномалий, который является надежным со статистической точки зрения.

Метод IQR() основан на межквартильном размахе. Определяется межквартильный размах Q1-Q3 и это значение умножается на 3, все что выходит за границы утроенного IQR считается выбросом.

Метод GESD: Generalized Extreme Studentized Deviate Test. Проводит итеративную оценку теста обобщенного экстремального стьюдентизированного отклонения GESD на наличие аномалий: на каждом шаге определяя самую большое аномальное значение и откидывая его из рассмотрения, оценка проводится до тех пор, пока согласно тесту в данных помалюй наблюдаться не будет.

Метод GESD имеет лучшие характеристики по точности выявления множественных аномалий (выбросов) в данных, но уступает по скорости вычислений методу IQR, за счет итеративного пересчета.

Методы импутации

Импутация это процесс восполнения данных.

Для замены аномалий пропусками NA можно использовать команду ifelse():

data2 <- data %>% mutate(y = ifelse(y < 3 | y > 20, NA, y))

Для того чтобы с помощью библиотеки dplyr перекодировать значения переменной на основе формулы:

LHS (условие) ~ RHS (результат), в которой условия проверяются на основе логических выражений нужно использовать функцию case_when()

Метод "polyreg" используется для импутации (восстановления) пропущенных значений для неупорядоченных категориальных переменных.

Существуют методы импутации, которые основаны на замене пропусков синтетическими наблюдениями. Для методов синтезирования используются различные инструменты машинного обучения.

Так, например, пакет MissForest – проводит импутацию на основе алгоритма случайного леса.

Пакет **MICE** применяет для восстановления пропусков метод цепных уравнений.

Метод искусственной импутации не подходит, когда много пропущенных, в том числе подряд, значений.

Спасибо за внимание!