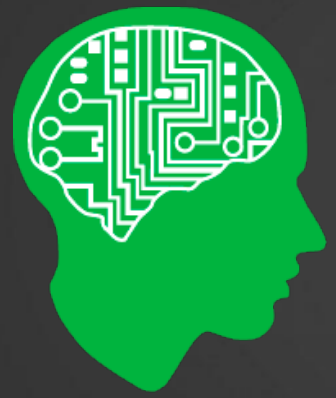
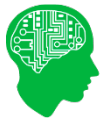
SkillFactory



SkillFactory

Описательные статистики и графики



SkillFactory

Описательные статистики

23, 45, 23, 44, 34, 56, 54, 12, 11, 44, 44, 31, 4,
30, 20, 49, 38, 48, 38, 40, 36, 41, 33, 47, 32

Меры центра

- Среднее
- Медиана
- Мода

Меры разброса

- Стандартное отклонение
- Размах
- Межквартильный размах
  (Квартили)

Описательные статистики

23, 45, 23, 44, 34, 56, 54, 12, 11, 44, 44, 31, 4,
30, 20, 49, 38, 48, 38, 40, 36, 41, 33, 47, 32

Меры центра

- Среднее = 35
- Медиана = 38
- Мода = 44

Меры разброса

- Стандартное отклонение = 13.5
- Размах = 52
- Межквартильный размах:

от 26.5 (1й квартиль) до 44.5 (3й квартиль)

Немного терминов

Случайная величина: $X =$ кол-во контрактов

Наблюдение: $x_5 = 41$

Генеральная совокупность: все менеджеры компании, размер $\mathrm{N} = 1000$

Выборка: случайно выбранные сотрудники, размер $\mathrm{n} = 25$

Важно различать выборку и генеральную совокупность!

Меры центра

- Среднее
- Мода
- Медиана

Меры центра. Среднее и мода

Выборочное среднее: $\bar{x} = \frac{\sum x_i}{n}$

Истинное среднее: $\mu = \frac{\sum x_i}{N}$

23, 45, 23, 44, 34, 56, 54, 12, 11, 44, 44, 31, 4, 30,
20, 49, 38, 48, 38, 40, 36, 41, 33, 47, 32

$$\bar{x} = \frac{23 + 45 + \cdots + 32}{25} = 35.08$$

Мода – наиболее частое наблюдение

Мода = 44 (частота=3)

Меры центра. Медиана

Медиана – серединное значение, отделяет 50% наименьших наблюдений от 50% наибольших наблюдений.

1. Упорядочить наблюдения по возрастанию
2. Найти «середину»
3. Медиана – серединное значение

Меры центра. Медиана

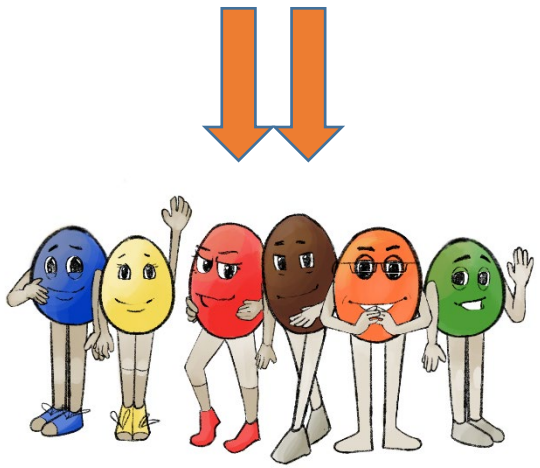
n **нечетное**: 23, 23, 34, 44, 45;

$$k = \frac{n+1}{2} = 3$$

$$\text{медиана} = x_3 = 34$$

Меры центра. Медиана

n **четное**: 23, 23, 34, 44, 45, 56;



k между $\frac{n}{2} = 3$ и $\frac{n}{2} + 1 = 4$

$$\text{медиана} = \frac{x_3 + x_4}{2} = \frac{34 + 44}{2} = 39$$

Меры центра. Медиана

23, 45, 23, 44, 34, 56, 54, 12, 11, 44, 44, 31, 4, 30,
20, 49, 38, 48, 38, 40, 36, 41, 33, 47, 32

1. **Упорядочить**:

4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36, 38, 38,
40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56

2. **Найти «середину»:** n=25, нечетное.

$$k = \frac{n+1}{2} = \frac{25+1}{2} = 13$$

3.**Медиана – серединное значение**

$$\text{медиана} = x_{13} = 38$$

Среднее или медиана?

Среднее:

- Чувствительно к нетипичным значениям (выбросам) и смещенным распределениям
- математически удобно

Медиана:

- Типичные значение даже для смещенных
- Иногда неудобно в подсчете

Квартили

Делят данные на 4 равные части:

x_{MIN} = 4, 11, 12, 20, 23, 23,
Q_1
30, 31, 32, 33, 34, 36,
Q_2,
38, 40, 41, 44, 44, 44,
Q_3
45, 47, 48, 49, 54, 56= x_{MAX}

Нижний квартиль Q_1 - отделяет наименьшие 25% наблюдений от остальных 75%

Верхний квартиль Q_3 - отделяет наибольшие 25% наблюдений от остальных 75%

Второй квартиль $Q_2 = \text{медиана}$

Квартили

1. Упорядочить по возрастанию

1. Найти медиану, разделить данные на 2 части относительно нее

1. Найти медиану в нижней части = Q1

1. Найти медиану в верхней части = Q3

Квартили

4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36,

медиана=38,

38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, 56

$n_1 = n_2 = 12 -$ нечетное, k между 6 и 7

$$Q_1 = \frac{x_6 + x_7}{2} = \frac{23 + 30}{2} = 26.5$$

$$Q_3 = \frac{44 + 45}{2} = 44.5$$

Меры разброса

- Размах
- Межквартильный размах
- Стандартное отклонение

Меры разброса. Размах

Размах - разность между максимальным и минимальным значением

$$\text{Размах} = \boldsymbol{Range} = \boldsymbol{x_{MAX}} = \boldsymbol{x_{MIN}}$$

4, 11, 12, 20, 23, 23, 30, 31, 32, 33, 34, 36, 38, 38, 40, 41, 44, 44, 44, 45, 47, 48, 49, 54, **56**

$$Range = 56 - 4 = 52$$

Меры разброса. Межквартильный размах

Межквартильный размах
(Interquartile Range, IQR) - разность между третим и первым квратилем:

$$IQR = Q_3 - Q_1$$

$$Q_1 = 26.5, \qquad Q_3 = 44.5$$

$$IQR = Q_3 - Q_1 = 44.5 - 26.5 = 18$$

Меры разброса. Стандартное отклонение

Стандартное отклонение - мера типичного (стандартного) отклонения наблюдения от среднего.

Истинное стандартное отклонение σ :

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

Истинная дисперсия: $Var(X) = D(X) = \sigma^2$

Выборочное стандартное отклонение S :

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

Меры разброса. Стандартное отклонение

23, 45, 23, 44, 34, 56, 54, 12, 11, 44, 44, 31, 4,
30, 20, 49, 38, 48, 38, 40, 36, 41, 33, 47, 32

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} =$$

$$= \sqrt{\frac{(23 - 38.05)^2 + (45 - 38.05)^2 + \cdots + (32 - 38.05)^2}{25 - 1}} =$$

$$= 13.49$$

Важно!

Var(X)≥ 0, $\sigma \geq 0$

Нетипичные наблюдения. Выбросы

Выброс – наблюдение, существенно отличающееся от остальных в выборке.

$$x_i < Q_1 - 1.5 \cdot IQR$$
$$x_i > Q_3 + 1.5 \cdot IQR$$

Что делать?

- Можно удалить перед подсчетом статистик
- Отдельно сообщить в отчете

Графики

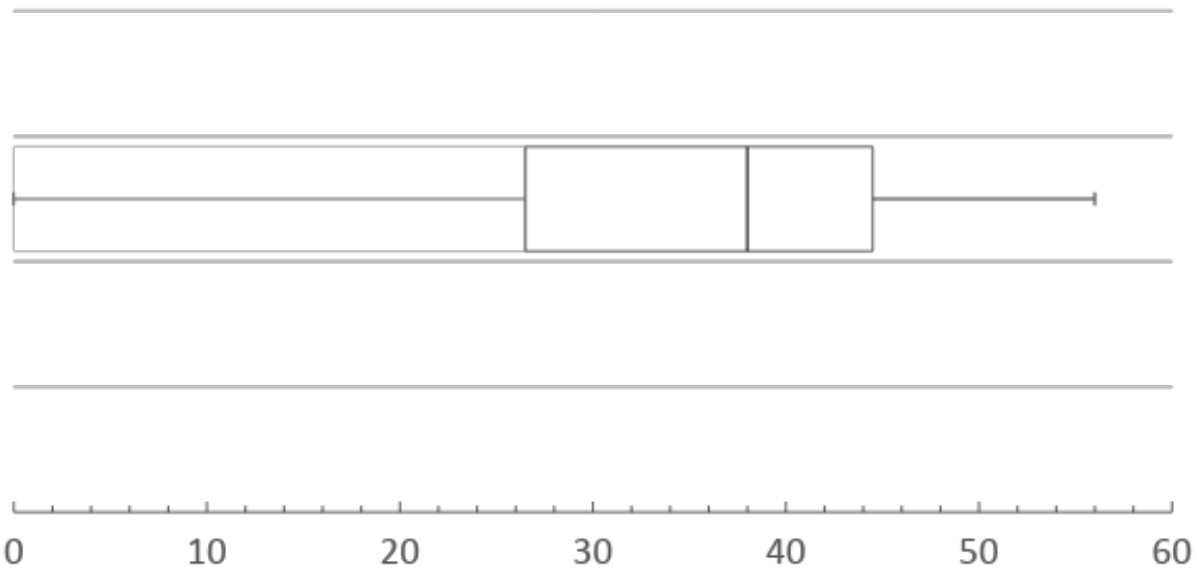
«You cannot do anything that you can't picture yourself doing»

(Unknown)

SKILLFACTORY

Графики: boxplot

Показывает положение $x_{MIN}, Q_1, Q_2, Q_3, x_{MAX}$



Графики: гистограмма

Ось X – значения признака, разбитые на интервалы

Ось Y - (относительная) частота значений в этом интервале