

# Первичные описательные статистики

# Задача

**Возраст педагогических работников** (в годах):

18; 38; 40; 28; 29; 26; 38; 34; 22; 28; 30;  
22; 23; 35; 33; 27; 24; 30; 32; 49; 37; 28;  
25; 29; 26; 31; 24; 29; 27; 32; 25; 29; 29;  
52; 58; 44; 39; 57; 19; 25.

Насколько молод коллектив?

# Меры центральной тенденции

- **Мода (Mo)** - значение, которое чаще других встречается в выборке.
- Если все значения встречаются одинаково часто — мода отсутствует
- Если два соседних значения имеют одинаковую частоту — мода между ними
- Выборка считается **бимодальной**, если два несмежных значения имеют наибольшую частоту

# Меры центральной тенденции: **Мода**

**В интервальном  
вариационном ряду:**

- 1) Данные уже сгруппированы в интервалы
- 2) Найти интервал с максимальной частотой — модальный
- 3) Считать моду по формуле:  
 $X_{mo}$  — нижняя граница модального интервала;  
 $h$  — ширина интервала;  
 $m$  — частоты модального, премодального и постмодального интервалов

$$M_o = X_{mo} + h * (m_{mo} - m_{mo-1}) / ((m_{mo} - m_{mo-1}) + (m_{mo} - m_{mo+1}))$$

**В безинтервальном  
вариационном ряду:**

- 1) Установить соответствие между значениями  $X$  и их частотой
- 2) Самое частое значение,

или

$$M_o = X_i$$

При условии  $m_{x_i} > \forall m_{x \neq x_i}$

# Меры центральной тенденции

- **Медиана ( $Md$ )** - значение признака, которое делит ранжированное множество данных пополам так, что одна половина оказывается меньше медианы, а другая — больше
- Если объем выборки — нечетное число, то медиана...
- Если объем выборки четное число, то медиана...

# Меры центральной тенденции: Медиана

В интервальном  
вариационном ряду:

- 1) Если данные уже сгруппированы в интервалы,
- 2) Найти медианный интервал, в котором накопленная относительная частота пересекает отметку в 50%
- 3) Считать медиану по формуле:

$X_{me}$  - нижняя граница модального интервала;

$N$  - объем выборки;

$M_{me-1}$  - накопленная частота интервала перед медианным

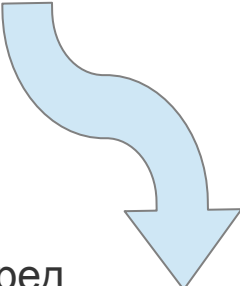
$h$  - ширина интервала;

$m_{me}$  - частота медианного интервала

В безинтервальном  
вариационном ряду:

- 1) Расположить все значения по возрастанию
- 2) Медианой будет значение, находящееся в точном центре ряда.

$Me = X_i$  при условии  
 $i = (N+1)/2$


$$Me = X_{me} + \frac{h * (N/2 - M_{me-1})}{m_{me}}$$

# Меры центральной тенденции

**Среднее арифметическое** - частное от деления всех значений ( $X_i$ ) на их количество ( $N$ )

$$\bar{X} = \frac{\sum X_i}{N}$$

**Свойства среднего:**

- 1) если к каждому значению прибавить число  $C$ , то среднее тоже увеличится на число  $C$ ;
- 2) если каждое значение умножить на  $C$ , то среднее увеличится в  $C$  раз

# Выбор меры центральной тенденции

*«Средняя температура по больнице?»*

- **Мода и медиана** «не чувствительны» к выбросам (на них не влияет отдельное большое или малое значение);
- **Мода** нестабильна в малых выборках;
- **Среднее** содержит погрешности на малых выборках с несимметричным распределением
- Для характеристики малой выборки **выбирайте медиану!**



# Меры изменчивости

- **Размах (P)** – интервал между максимальным и минимальным значениями признака  
выборка: {1, 2, 3, 4, 5, 6, 7, 7, 8, 9}  
Размах=8 N=10

$$P = X_{\text{мах}} - X_{\text{мин}}$$

# Меры изменчивости

- **Среднее абсолютное отклонение (mad)** – это среднеарифметическое разницы (по абсолютной величине) между каждым значением в выборке и ее средним

$$\text{mad} = \frac{\sum d}{N}$$

- где  $d = |x_i - M|$  - модуль расстояния;
- $M$  – среднее или медиана выборки;
- $x_i$  – конкретное значение;
- $N$  – объем выборки

# Меры изменчивости

- **Дисперсия ( $S^2$ )** — мера изменчивости, пропорциональная сумме квадратов отклонений значений от среднего

$$S^2 = \frac{\sum d^2}{N}, \text{ для больших выборок}$$

$$S^2 = \frac{\sum d^2}{N - 1}, \text{ для малых выборок (>30чел)}$$

# Свойства дисперсии

- Если все значения равны друг другу, дисперсия равна 0 (нет рассеяния признака);
- Если ко всем значениям прибавить число  $C$ , это не меняет дисперсию;
- Увеличение всех значений в  $C$  раз увеличивает дисперсию в  $C^2$  раз
- Применима только для данных метрических шкал! (т.к. является мерой расстояния)

# Меры изменчивости

- **Стандартное отклонение ( $s$ ) или ( $S_n$ )** — мера изменчивости, являющаяся положительным значением квадратного корня из дисперсии

- Для больших выборок  $\sqrt{S^2} = \sqrt{\frac{\sum d^2}{N}}$

- Для малых выборок  $S_n = \sqrt{\frac{\sum d^2}{N - 1}}$

- Всегда выражается в исходных единицах признака, в отличие от дисперсии

# Асимметрия и эксцесс

Асимметрия и эксцесс характеризуют распределение признака в выборке, являются 3 и 4 моментами среднего

Показатели **асимметрии** и **эксцесса**.

$$A = \frac{\frac{1}{n} * \sum_{i=1}^n (x - \bar{X})^3}{s^3}$$
$$E = \frac{\frac{1}{n} * \sum_{i=1}^n (x - \bar{X})^4}{s^4} - 3$$

Свойства **асимметрии** и **эксцесса**:

- Если  $A > 0$  существенно, то среднее  $>$  медианы  $>$  моды и наоборот, при отрицательной асимметрии  $M_0 > M_e > M$
- Если  $E > 0$  существенно, то распределение выборки островершинное (большее количество людей набирает близкие к моде баллы); а при  $E < 0$  распределение плосковершинное — т.е. больше людей «рассеяны» от центра

# Меры положения

- **Квантиль** — точка на числовой оси измеренного признака, которая делит всю совокупность измерений на две группы с известным соотношением численности.
- **Квартили** — 3 точки — значения признака, которые делят сортированное по возрастанию множество значений на 4 равных интервала (по 25% выборки в каждом). 2-й квартиль — это медиана.
- **Процентили** - 99 точек - значений признака....  
(аналогично делят на отрезки по 1%)
- См. накопленные относительные частоты, чтобы понять, каким квантилем является конкретное значение

# Какие описательные статистики можно применять...

НА ШКАЛЕ НАИМЕНОВАНИЙ?

НА РАНГОВОЙ ШКАЛЕ?

НА ШКАЛЕ ИНТЕРВАЛОВ?

НА ШКАЛЕ РАВНЫХ ОТНОШЕНИЙ?



**Метрика** — функция, вводящая понятие **расстояния** между двумя элементами **a** и **b** множества **A**

**Расстояние** — числовая функция  $R(a, b)$ , удовлетворяющая следующим условиям:

- (1)  $R(a, b) \geq 0$ , причем  $R(a, b) = 0$  тогда и только тогда, когда  $a = b$ ;
- (2)  $R(a, b) = R(b, a)$ ;
- (3)  $R(a, b) + R(b, c) \geq R(a, c)$ , «правило треугольника».

Введение метрики делит шкалы на неметрические и метрические.