Измерение центральной тенденции

Мода
Медиана
Среднее

Измерение центральной тенденции (measure of central tendency) состоит в выборе одного числа, которое **наилучшим образом** описывает все значения признака из набора данных. Такое число называют центром, типическим значением для набора данных, мерой центральной тенденции.

Зачем?

- Получим информацию о распределении признака в сжатой форме.
- Сможем сравнить между собой два набора данных (две выборки).
- Минус: ведет к потере информации по сравнению с распределением частот.

Мода – наиболее часто встречающееся значение в выборке, наборе данных. Обозначается ***Mo***.

Выборка: 5,4 1,2 0,42 1,2 0,48 Мода=**1,2**

Для данных, расположенных в таблице частот, мода определяется как значение, имеющее наибольшую частоту.

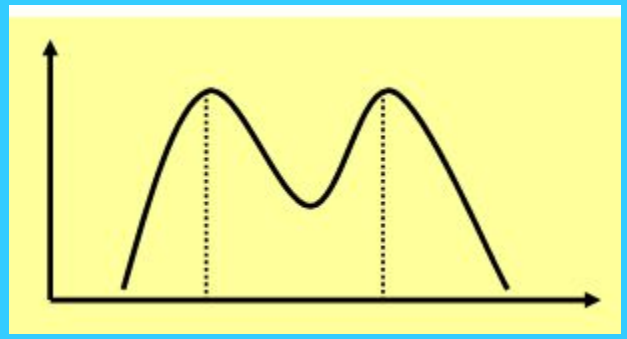
Таблица частот для числа посетителей гипермаркета

X	0	1	2	3	4	5	6	7
m	1	4	3	5	**6**	5	3	3

Mo=4

Одна ли мода?

Если наибольшую частоту имеет два значения выборки, выборочное распределение называется **бимодальным**.



Если наибольшую частоту имеет более двух значений выборки, выборочное распределение называется **мультимодальным**.

Если ни одно из значений не повторяется, **мода отсутствует**.

Вариационный ряд

Вариационный ряд - упорядоченные данные, расположенные в порядке возрастания значения признака

Пример. Набор данных:

6 1 3 7 1 7 3

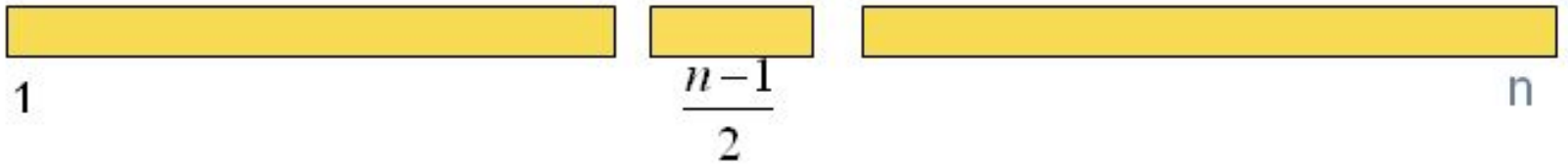
После упорядочения получим вариационный ряд:

1 1 3 3 6 7 7

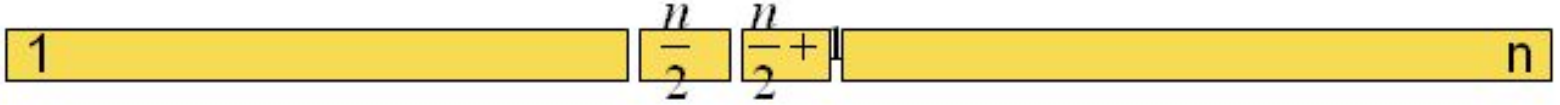
Медиана

Медиана есть значение серединный элемент вариационного ряда.

Для набора из n значений, если n нечетно, средний элемент имеет номер: $\frac{n-1}{2}$



Если n четно, медиана находится как среднее арифметическое двух соседних серединных элементов: $\frac{n}{2}$ и $\frac{n}{2}+1$



Пример вычисления медианы

Для набора данных из семи чисел:

6 1 3 7 1 7 3

После упорядочения получим вариационный ряд:

1 1 3 3 6 7 7

Медиана есть средний элемент. Его номер четвертый.

Если набор данных включает восемь чисел:

1 1 3 3 6 7 7 9

Тогда медиана равна (3+6)/2=4,5

Среднее значение

Выборочное среднее будем называть среднее арифметическое выборки, то есть сумму всех значений выборки, деленную на ее объем.

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Пример. **Покупателей гипермаркета попросили ответить на вопрос сколько денег в среднем они тратят при одном посещении гипермаркета. Было опрошено 1000 человек. Найти оценку математического ожидания случайной величины X – количества денег, которые тратит покупатель при посещении гипермаркета.**

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = 1085{,}5$$ **рубля**

=СРЗНАЧ(A1:A1000)

Среднее - еще не значит «лучшее»

Пример. В деревне 50 жителей. Среди них 49 человек – крестьяне с месячным доходом в 1 тыс.рублей, а один житель – зажиточный владелец строительной фирмы, с месячным доходом 451 тыс.рублей.

Среднее равно 10 тыс. рублей.

Однако, вряд ли можно утверждать, что это число адекватно представляет доход жителей деревни.

В этом случае, более разумно взять в качестве меры центральной тенденции моду или медиану (обе равны 1 тыс. рублей).

Измерение вариации

Размах

Квартильный размах

Дисперсия

Стандартное отклонение

Постановка задачи

Рассмотрим три вариационных ряда:

а) 999, 1000, 1001
б) 900, 1000, 1100
в) 1, 1000, 1999

Во всех трёх случаях среднее равно 1000.

Однако, в случае в) значения признака «разбросаны» вокруг среднего сильнее, чем в б); а в случае б) – сильнее, чем в случае а).

Как выразить степень разброса (вариации, measure of variation) одним числом?

Размах (Range)

Размах – разность между наибольшим значением набора данных и наименьшим.

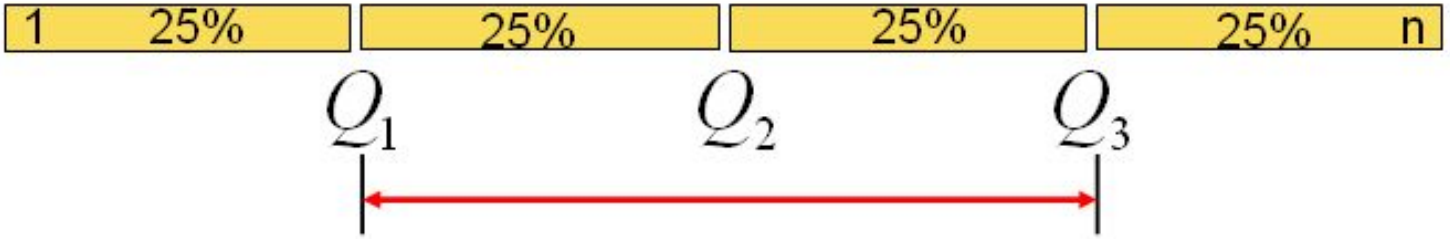
$$R = x_{\max} - x_{\min}$$

Пример: Для набора данных 27, 8, 3, 12, 10, 26, 6, 19
размах равен R = 27 – 3 = 24.

Размах – очень простая мера вариации, но очень «грубая».

Квартили (Quartile)

Под квартилями понимаются значения, которые делят вариационный ряд на четыре равные части:

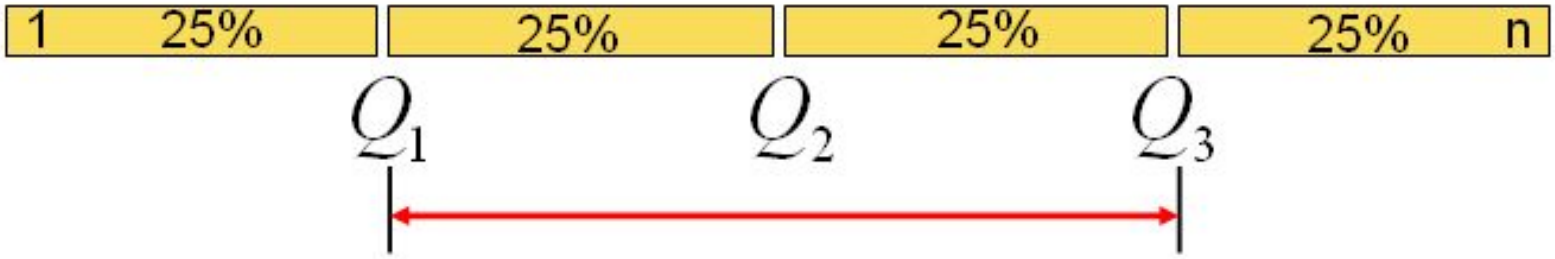


Ниже первого квартиля расположено 25% всех данных. Между первым и вторым квартилем также расположено 25% данных. Второй квартиль совпадает с медианой.

Размах квартилей (InterQuartile Range) вычисляется по формуле:

$$IQR = Q_3 - Q_1$$

Свойства квартильного размаха



Между Q1 и Q3 расположены 50% всех данных.

Оценка дисперсии

Оценкой дисперсии $DX = M(X - MX)^2$

является выборочная дисперсия

$$\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

σ_X **называется среднеквадратичным отклонением**

Подсчет дисперсии в таблице

Дисперсию удобно рассчитывать при помощи таблицы.

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	2 – 5 = -3	9
3	3 – 5 = -2	4
6	6 – 5 = 1	1
9	9 – 5 = 4	16
20		30

В первом столбце выборка. Второй и третий столбцы для вычислений.

Сумма третьего столбца есть сумма квадратов отклонений значений выборки от среднего.

$$\bar{x} = \frac{\sum x}{n} = \frac{20}{4} = 5$$

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{30}{4-1} = 10$$

Оценка по выборке математического ожидания и дисперсии

Пример. **Покупателей гипермаркета попросили ответить на вопрос сколько денег в среднем они тратят при одном посещении гипермаркета. Было опрошено 1000 человек. Найти оценку дисперсии случайной величины X – количества денег, которые тратит покупатель при посещении гипермаркета.**

Оценка по выборке математического ожидания и дисперсии

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
960	-125,5	15750
500	-585,5	342810
1250	164,5	27060
2410	1324,5	1754300
350	-735,5	540960
1120	34,5	1190
1820	734,5	539490
400	-685,5	469910
1050	-35,5	1260
1570	484,5	234740
860	-225,5	50850

$$\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\bar{X} = 1085,5$$

Оценка по выборке математического ожидания и дисперсии

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
960	-125,5	15750
500	-585,5	342810
1250	164,5	27060
2410	1324,5	1754300
350	-735,5	540960
1120	34,5	1190
1820	734,5	539490
400	-685,5	469910
1050	-35,5	1260
1570	484,5	234740
860	-225,5	50850

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 =$$

$$= 329963400$$

Оценка по выборке математического ожидания и дисперсии

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
960	-125,5	15750
500	-585,5	342810
1250	164,5	27060
2410	1324,5	1754300
350	-735,5	540960
1120	34,5	1190
1820	734,5	539490
400	-685,5	469910
1050	-35,5	1260
1570	484,5	234740
860	-225,5	50850

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 329963400$$

$$\sigma_X^2 = \frac{1}{999} \cdot 329963400 = 330293,69$$

Оценка по выборке математического ожидания и дисперсии

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
960	-125,5	15750
500	-585,5	342810
1250	164,5	27060
2410	1324,5	1754300
350	-735,5	540960
1120	34,5	1190
1820	734,5	539490
400	-685,5	469910
1050	-35,5	1260
1570	484,5	234740
860	-225,5	50850

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 329963400$$

$$\sigma_X^2 = \frac{1}{999} \cdot 329963400 = 330293,69$$

=ДИСП(A1:A1000)

Оценка по выборке математического ожидания и дисперсии

X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
960	-125,5	15750
500	-585,5	342810
1250	164,5	27060
2410	1324,5	1754300
350	-735,5	540960
1120	34,5	1190
1820	734,5	539490
400	-685,5	469910
1050	-35,5	1260
1570	484,5	234740
860	-225,5	50850

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 329963400$$

$$\sigma_X^2 = \frac{1}{999} \cdot 329963400 = 330293,69$$

=ДИСП(A1:A1000)

$$\sigma_X = \sqrt{330293,69} = 574,7$$

В файле flat представлены данные о ценах на однокомнатные квартиры (тыс. USD), выставлявшихся на продажу в Москве.

1. Вычислите среднее с помощью функции СРЗНАЧ

2. Постройте вариационный ряд выборки и вычислите по нему медиану.

Отсортируем Выборку – это и есть вариационный ряд

N	Price
1	28
2	28
3	28
4	28
5	29
6	30
7	30
8	30
9	30
10	30
11	30
12	31
13	31
14	31
15	32
16	32

2. Постройте вариационный ряд выборки и вычислите по нему медиану.

Отсортируем Выборку – это и есть вариационный ряд

N	Price
1	28
2	28
3	28
4	28
5	29
6	30
7	30
8	30
9	30
10	30
11	30
12	31
13	31
14	31
15	32
16	32

n=69 – нечетно, медиану ищем под номером (69+1)/2=35

2. Постройте вариационный ряд выборки и вычислите по нему медиану.

Отсортируем Выборку – это и есть вариационный ряд

31	36	
32	37	
33	37	
34	37	
35	37	медиана
36	37	
37	37	
38	38	
39	39	

n=69 – нечетно, медиану ищем под номером (69+1)/2=35

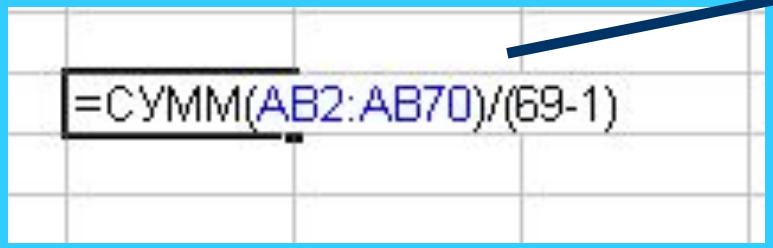
3. Вычислить медиану с помощью функции МЕДИАНА, сравните результаты.

4. Вычислите размах выборки.

5. Вычислить дисперсию с помощью функции ДИСП и по формуле дисперсии.

5. Вычислить дисперсию с помощью функции ДИСП и по формуле дисперсии.

		N	Price	X-Xcp	(X-Xcp)^2
Среднее	39,65	1	28	-11,65	135,77
		2	28	-11,65	135,77
		3	28	-11,65	135,77
		4	28	-11,65	135,77
		5	29	-10,65	113,47
		6	30	-9,65	93,16
		7	30	-9,65	93,16
		8	30	-9,65	93,16
		9	30	-9,65	93,16
		10	30	-9,65	93,16
		11	30	-9,65	93,16
		12	31	-8,65	74,86



5. Вычислить стандартное отклонение с помощью функции СТАНДОТКЛОН и по формуле стандартного отклонения.

6. Вычислить стандартное отклонение с помощью функции СТАНДОТКЛОН и по формуле стандартного отклонения.

7. Вычислить нижний и верхний квартиль с помощью функции КВАРТИЛЬ.
В качестве второго аргумента функции указать 1 для нижнего квартиля и 3 для верхнего. А какая величина будет вычислена, если указать в качестве второго аргумента 2?

8. Дайте экономическую интерпретацию квартилям.

9. Вычислить среднее, медиану, дисперсию стандартное отклонение, нижний и верхний квартили с помощью команды *Сервис – Анализ данных – описательная статистика).*