Меры центральной тенденции

Цель:

- Изучение числовых характеристик, позволяющих анализировать выборку и делать некоторые выводы

Постановка задачи

- 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.

Что получим

- Получим информацию о распределении признака в сжатой форме
- Сможем сравнить между собой два набора данных (две выборки)

- **Минус:** выбор центра ведет к потере информации по сравнению с распределением частот.

Мода (Mode)

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

Обозначается ***Мо.***

Пример моды

Выборка: 5 4 1 2 4 3 1 2 4 8 3 6 4 1

варианты	частоты
1	3
2	2
3	2
4	4
5	1
6	1
8	1

Мода=4

Наиболее часто встречающееся значение

Мода в таблице частот

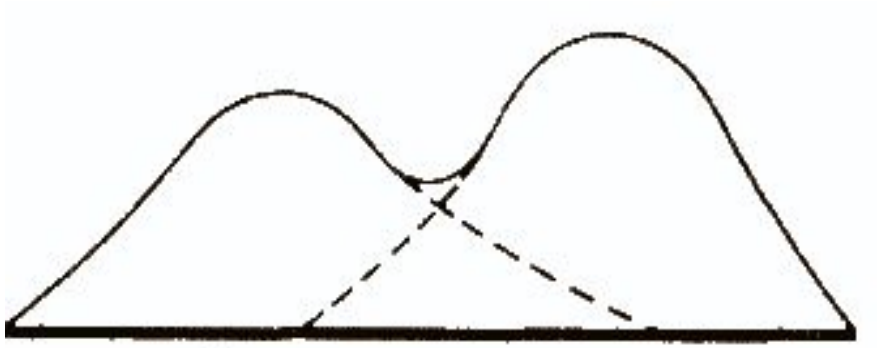
Для данных, расположенных в таблице частот, мода определяется как значение, имеющее наибольшую частоту

Категории	f
Демократы	41
Коммунисты	23
Либералы	22
Любители пива	5
Зеленые	12
Всего	103

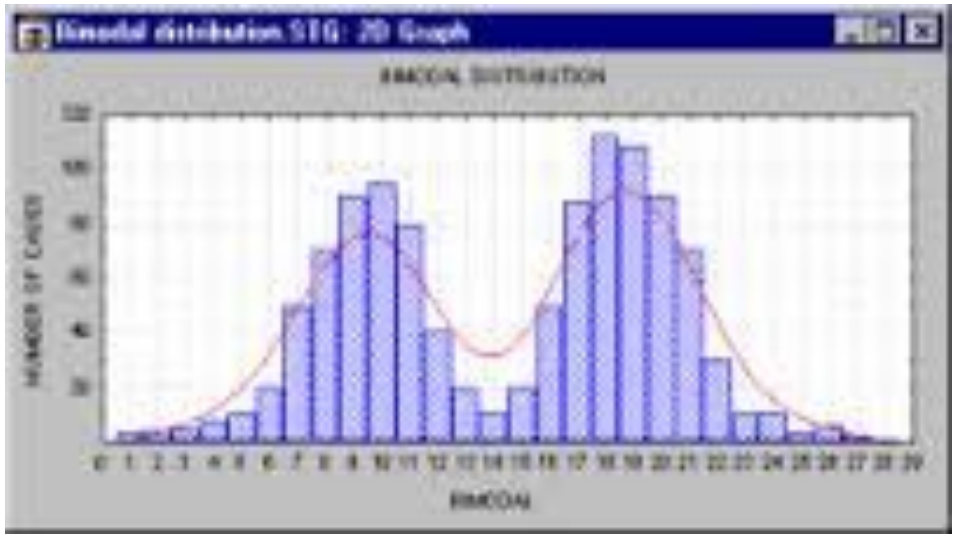
Одна ли мода?

Если наибольшую частоту имеют два значения выборки,

выборочное распределение называется **бимодальным.**



Бимодальное распределение



Бимодальное распределение

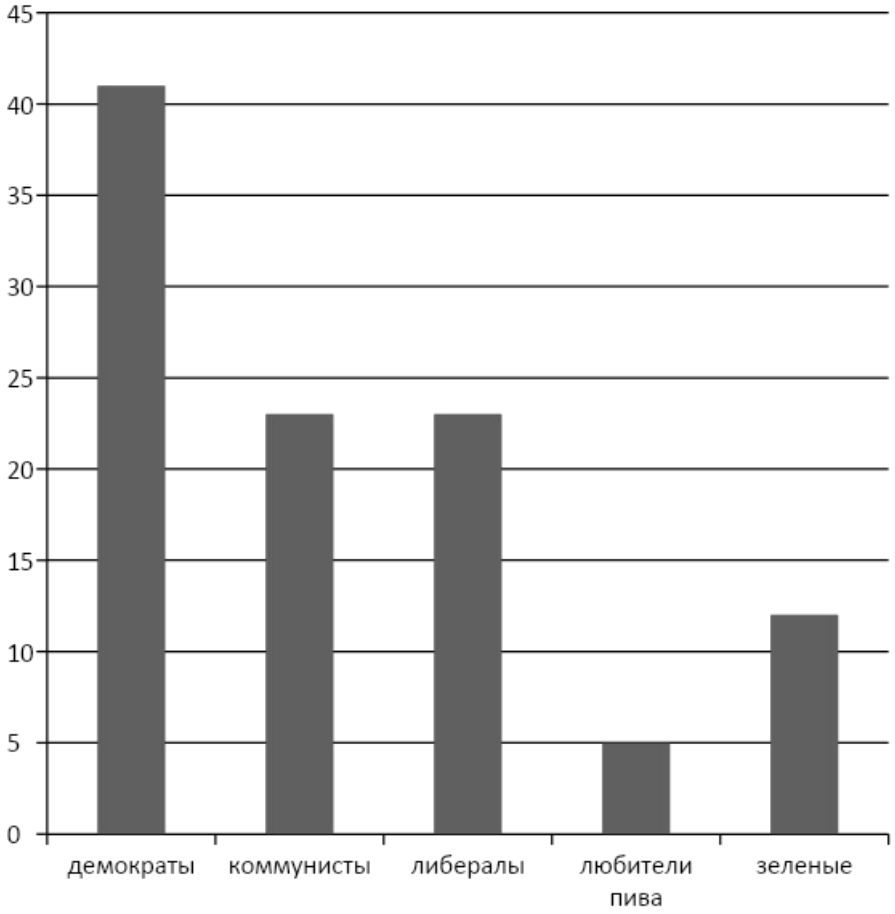
Два значения имеют наибольшую частоту, равную 23.

Две моды!

Категории	f
Демократы	41
Коммунисты	23
Либералы	23
Любители пива	5
Зеленые	12
Всего	103

На гистограмме

Два значения имеют наибольшую частоту, равную 23.



А если нет моды или больше двух?

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

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

Свойства моды

Наличие одного или двух крайних значений, сильно отличающихся от остальных, не влияет на значение моды.

Мода совпадает с точкой наибольшей плотности данных.

Мода может иметь несколько значений.

Мода может существовать для всех типов данных.

Единственная мера, которая работает в номинальной шкале!

Медиана

Вариационный ряд

- **Вариационный ряд - упорядоченные данные, расположенные в** порядке возрастания значения признака, либо в порядке убывания.

- Назван так, поскольку содержит **варианты значений признака.**

Пример вариационного ряда

Набор данных:

6 1 3 7 1 7 3

После упорядочения получим вариационный ряд:

1 1 3 3 6 7 7

В порядке убывания получим другой вариационный ряд:

7 7 6 3 3 1 1

Ранжирование

- Ранжирование означает присвоение числам рангов.
- Ранжирование данных производится после построения вариационного ряда (упорядочения).
- Ранги присваиваются от 1 до последнего номера в наборе данных.

Пример ранжирования

Есть упорядоченный набор данных из 9 чисел:

1 1 3 3 6 7 7 7 14

Нумеруем от 1 до 9:

1 2 3 4 5 6 7 8 9

А теперь находим ранги:

1,5 1,5 3,5 3,5 5 7 7 7 9

Если несколько соседних элементов равны, при ранжировании им присваивается одинаковый ранг, равный среднему арифметическому первоначальных рангов.

Медиана (Median)

- Медиана есть значение серединного элемента для набора данных.
- Обозначается ***Me.***
- Для нахождения медианы требуется составить вариационный ряд, то есть расположить все значения признака в порядке возрастания или убывания.
- Медиана расположена в середине вариационного ряда.

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.

Пример вычисления медианы

Если набор данных включает восемь чисел:

1 1 3 3 6 7 7 9

Тогда медиана равна (3+6)/2=4,5

Свойства медианы

Сильно отличающиеся от остальных данных крайние значения

не влияют на величину медианы.

Значение медианы является единственным для каждого набора

данных.

Медиана может быть определена не из полного набора данных.

Достаточно знать их расположение, общее число и несколько

значений, расположенных в середине вариационного ряда.

Медиана может быть определена для числовых и

Виды средних величин

Средняя арифметическая(mean) - применяется, если варианты возрастают (убывают) в арифметической прогрессии.

$$\bar{x} = \frac{\sum x_i}{n} \qquad \bar{x} = \frac{\sum (x_i \times p)}{n}$$

x - средняя арифметическая;

x_i - варианта;

p - частота встречаемости варианты;

n - число наблюдений

Виды средних величин

- **Средняя геометрическая** - вычисляется, если варианты возрастают (убывают) в геометрической прогрессии

$$\bar{x}_g = \sqrt[n]{x_1 x_2 x_3 \ldots x_n}$$

На практике используют логарифмированную формулу:

$$\log \bar{x}_g = 1/n(\log x_1 + \log x_2 + \log x_3 + \ldots \log x_n)$$

Пример вычисления среднего арифметического

Вычислим среднее для выборки из семи значений:

1 1 3 3 6 7 7

Получим: $\bar{X} = \frac{1+1+3+3+6+7+7}{7} = 28/7 = 4$

Среднее является «точкой равновесия»

Свойства среднего

Вычисляется только в числовых шкалах.

При вычислении необходимо использовать все данные.

Для каждого набора данных имеется только одно среднее.

Среднее есть единственная мера центральной тенденции, для которого сумма отклонений каждого значения равна нулю:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = \sum d_i = 0$$

Пример вычисления среднего для сгруппированных данных

Имеются результаты экзамена. Найти среднее значение

x	f	fx
2	*6*	*12*
3	*12*	*36*
4	*18*	*72*
5	*15*	*75*
	51	**195**

$$\bar{X} = \frac{\Sigma fx}{\Sigma f} = \frac{195}{51} = 3{,}82$$

Среднее для интервальных частот

интервал	частота	середина	произведение
	f	m	fm
0-99	11	49,5	544,5
100-199	12	149,5	1794,0
200-299	14	249,5	3493,0
300-399	1	349,5	349,5
400-499	2	449,5	899,0
всего	Σ =40		Σ =7080,0

Для каждого интервального распределения надо выбрать представителя каждого интервала - середину

Среднее для интервального распределения

Среднее для интервального распределения вычисляется по формуле:

$$\bar{X} = \frac{\Sigma (fm)}{\Sigma f}$$

где **Σ (fm)** = сумма произведений частоты на середину

Σ f = сумма частот, равна объему выборки

m *= середина интервалов*

Среднее - еще не значит «лучшее»

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).

Меры и шкалы

Шкала, по которой измеряется переменная, накладывает ограничения на выбор меры центральной тенденции.

Типическое значение	Номинальные данные	Порядковые данные	Интервальные данные
Мода			
Медиана			
Среднее			

Среднее для дихотомической шкалы

Среднее может также применяться и для переменной,

измеренной в дихотомической шкале.

Если два значения признака кодируются 0 и 1, то среднее указывает долю (относительную частоту) единиц в выборке.

Пример:

1, 0, 0, 0, 1, 1, 1, 1, 1, 0

Среднее равно 0,6. То есть 60% значений выборки принимают значение, равное единице.

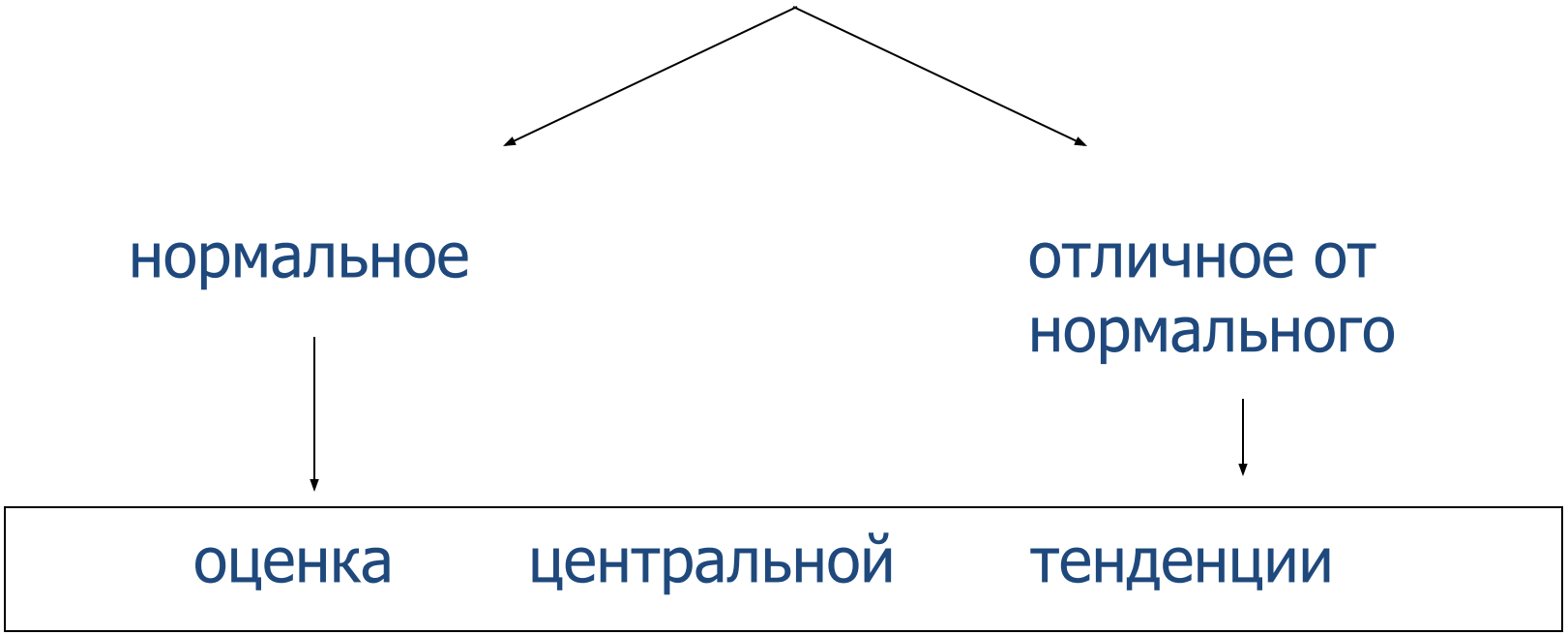
Какое типическое значение наилучшее?

1. «Наилучшее значение» - это такое, которое имеет наибольшую вероятность быть выбранным → **Мода**

2. «Наилучшее значение» - это такое значение, для которого сумма абсолютных отклонений значений переменной от типического будет наименьшей → **Медиана**

3. «Наилучшее значение» - это такое значение, для которого сумма квадратов отклонений значений переменной от типического будет наименьшей → **Среднее**

Вид распределения



Какое типическое значение наилучшее?

В зависимости от данных каждое из трех

значений может стать наилучшим!