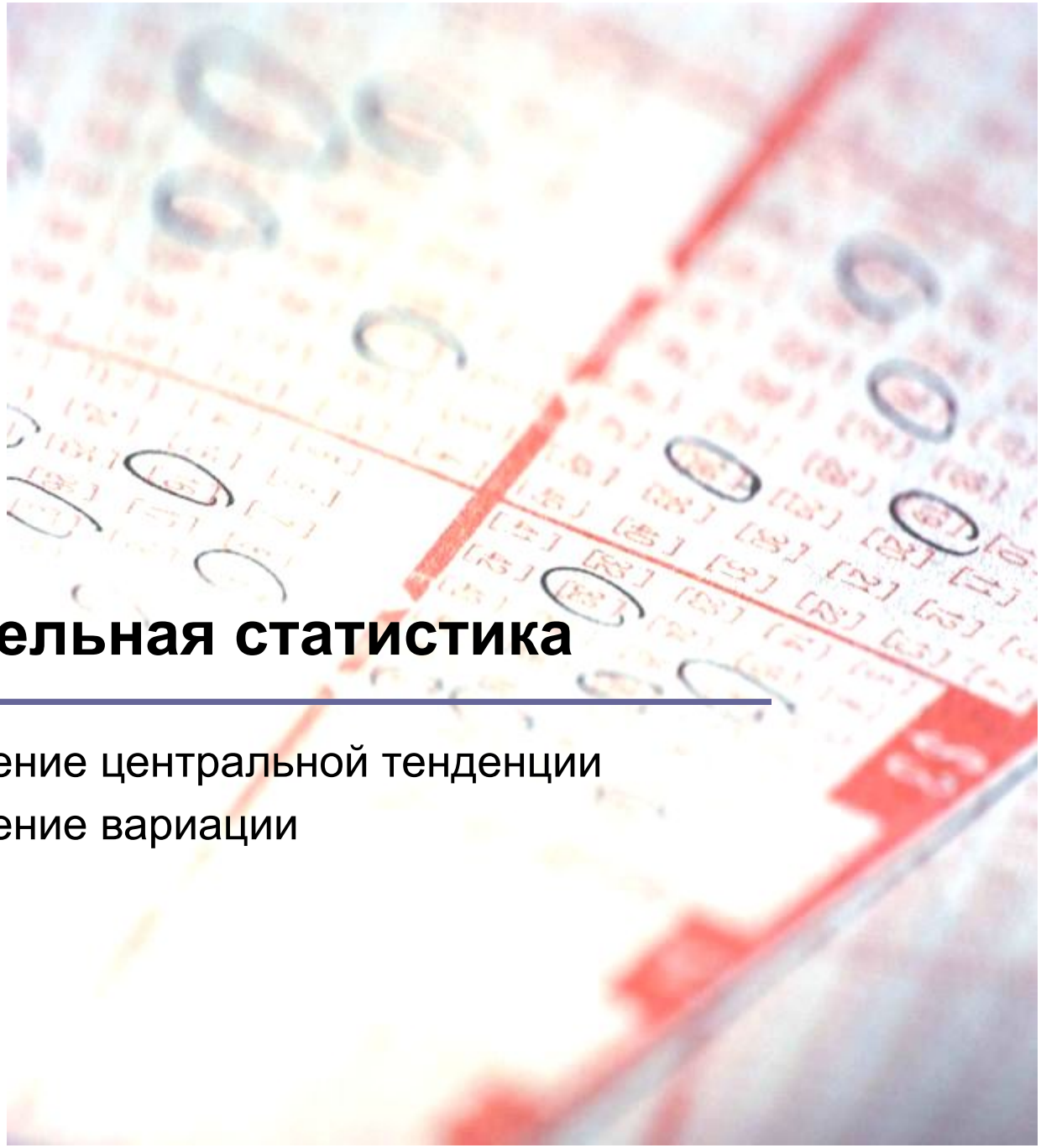


Тема 3.
Описательная статистика

3.1. Измерение центральной тенденции

3.2. Измерение вариации

Математическая статистика

Математическая статистика - область науки, изучающая случайные явления, разрабатывающая математические методы систематизации, обработки и использования статистических данных для научных и практических выводов.

Составными частями математической статистики являются:

(1) описание данных,

(2) статистическое оценивание

(3) проверка статистических гипотез.

Переменные

Данные (data) представляют собой результаты наблюдений, испытаний, накапливаемые с целью последующего изучения и анализа.

Переменная, признак (variable) - это некоторая общая для всех изучаемых объектов характеристика или свойство, конкретные проявления которого могут меняться от объекта к объекту.

Проявления признака называют **значениями, показателями, альтернативами, градациями.**

Распределение переменной (distribution of the variable) - совокупность различных значений, которые переменная принимает для различных изучаемых объектов.

Определения

Генеральная совокупность (population) - вся интересующая исследователя совокупность изучаемых объектов.

- **Параметры (parameters)** - числовые характеристики генеральной совокупности.

Выборка, выборочная совокупность (sample) - некоторая, обычно небольшая, часть генеральной совокупности, отбираемая специальным образом и исследуемая с целью получения выводов о свойствах генеральной совокупности.

- **Статистики (statistics)** - числовые характеристики выборки.

- **Гипотеза (hypothesis)** - предположение относительно параметров генеральной совокупности, которое подлежит проверке на основе анализа выборки.

Измерение явлений

Измерение (measurement) означает присвоение чисел характеристикам изучаемых объектов, явлений согласно некоторому правилу.

Шкала (scale) есть правило или алгоритм, в соответствии с которым изучаемым объектам, явлениям присваиваются числа.

Типы данных

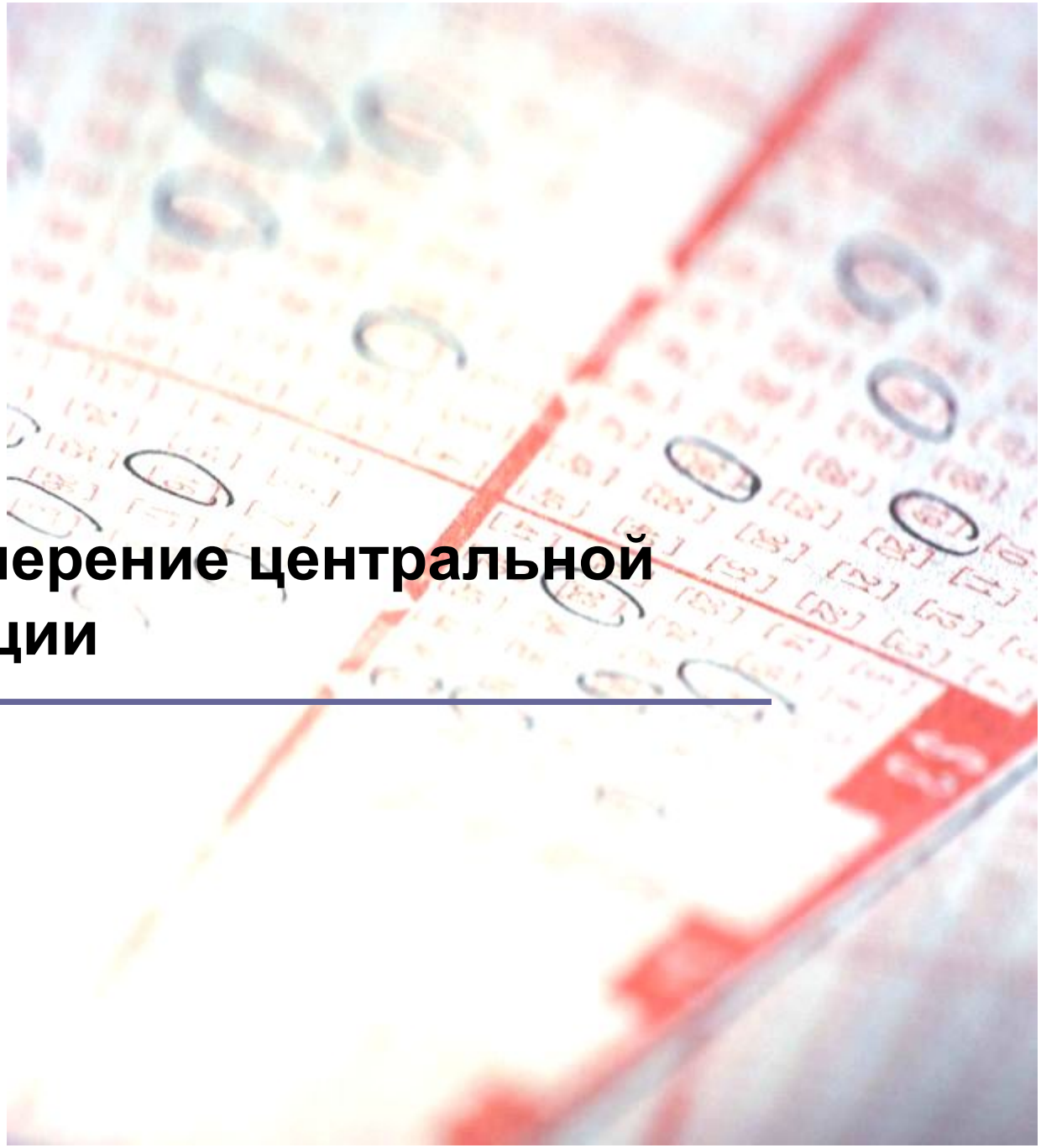
Дискретные данные (discrete data) представляют собой отдельные значения признака, общее число которых конечно либо если бесконечно, то является счетным, т.е. может быть подсчитано натуральными числами от одного до бесконечности.

Непрерывные данные (continuous data) могут принимать любое значение в некотором интервале.

Измерительные шкалы (С. Стивенс)

- номинативная, или номинальная, или шкала наименований (в том числе дихотомическая)
- порядковая, или ранговая, или ординальная шкала
- интервальная, или шкала равных интервалов
- шкала равных отношений или реляционная шкала

Шкала	Особенности	Пример
Номинальная	Содержит только категории, данные не могут упорядочиваться	Хобби студента. Только название.
Дихотомическая	Содержит две категории	Пол студента. Третьего не дано, если не рассматривать исключения.
Порядковая	Категории могут упорядочиваться, но разности не имеют смысла	Место на соревнованиях. Лучше результат - выше место.
Интервальная	Разности между значениями могут быть вычислены, но нет отношений	Температура студента. У больного выше на 1-2°C
Относительная	Имеется точка отсчета, возможны отношения между значениями	Рост студента. Один в 1,2 раза выше другого



3.1. Измерение центральной тенденции

Мода
Медиана
Среднее

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.

Зачем?

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

Мода

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

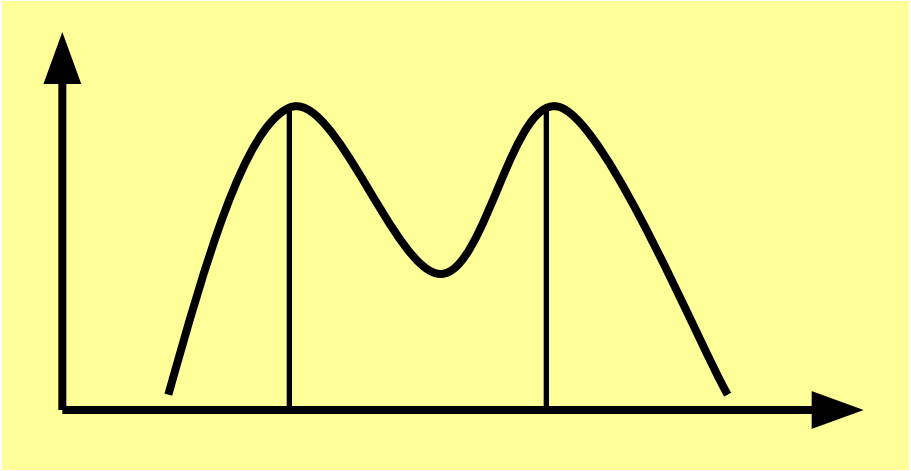
Для данных, расположенных в таблице частот, мода определяется как значение, имеющее наибольшую частоту.

Если наибольшую частоту имеет два соседних значения выборки, то мода определяется как среднее арифметическое этих значений.

Выборка: 5,4 1,2 0,48 1,2 0,48 Мода = (0,48+1,2)/2 = 0,84

Одна ли мода?

Если наибольшую частоту имеет два несоседних 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**.

Свойства моды

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

2. Мода совпадает с точкой наибольшей плотности данных.

3. Мода может иметь несколько значений.

4. Мода может существовать для всех типов данных. Единственная мера, которая работает в номинальной шкале!

Вариационный ряд

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

Пример. 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 упорядочения. Ранги присваиваются от 1 до последнего номера в наборе данных. Если несколько соседних элементов равны, им присваивается одинаковый ранг, равный среднему арифметическому.

Пример. Имеем 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

Например, значение 6 имеет ранг 5.

Медиана

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.

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:

$$\frac{n}{2} \quad \frac{n}{2}+1$$

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

Свойства медианы

1. Сильно отличающиеся от остальных данных крайние значения не влияют на величину медианы.

2. Значение медианы является единственным для каждого набора данных.

3. Медиана может быть определена не из полного набора данных. Достаточно знать их расположение, общее число и несколько значений, расположенных в середине вариационного ряда.

4. Медиана может быть определена для числовых данных и данных, измеряемых порядковой шкалой. Для порядковой шкалы в случае четного количества элементов оба серединных значения объявляются медианой.

Среднее значение

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

Формула:

$$\bar{x} = \frac{\sum x_i}{n}$$

где $\sum x_i$ = сумма всех значений выборки

n = объем выборки

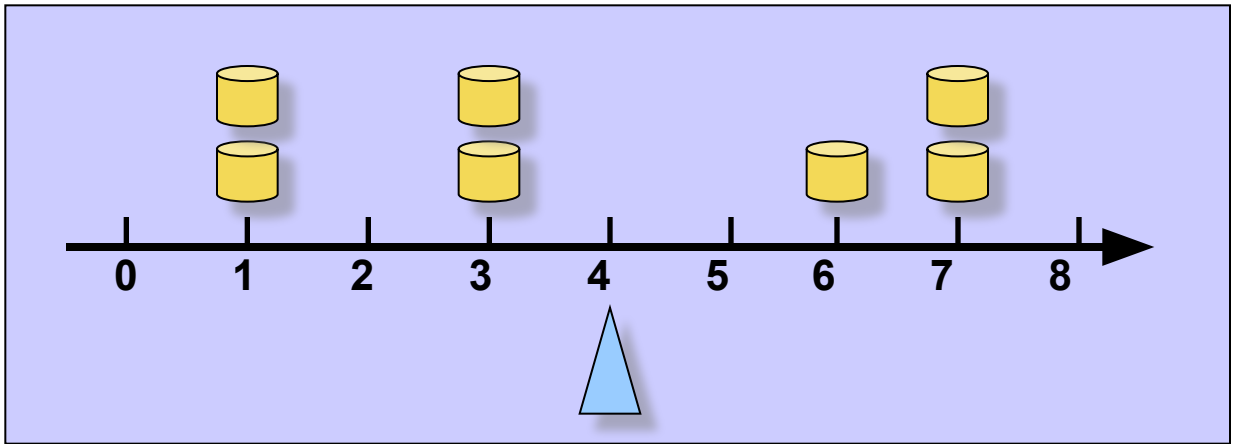
Индекс суммирования в статистической литературе часто опускается.

Пример вычисления среднего

Вычислим среднее для выборки из семи значений:

1 1 3 3 6 7 7

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



Среднее значение является «точкой равновесия».

Свойства среднего

1. Вычисляется только в числовых шкалах.

2. При ее вычислении необходимо использовать все данные.

3. Имеется для каждого набора данных только одно значение средней.

4. Средняя есть единственная мера центральной тенденции, для которой сумма отклонений каждого значения от нее равна нулю:

$$\sum (x_i - \bar{x}) = 0$$

Среднее для сгруппированных данных

Среднее для сгруппированных данных вычисляется по формуле:

$$\overline{x} = \frac{\sum x_i \cdot f_i}{\sum f_i}$$

где $\sum x_i \cdot f_i$ = сумма всех значений выборки

$\sum f_i$ = сумма частот, равна объему выборки

Если данные сгруппированы по интервалам, в качестве значения выбирается середина интервала.

Пример вычисления среднего

Имеются результаты экзамена. Найти среднее значение.

x_i	f_i	$x_i \cdot f_i$
0	1	0
1	2	2
2	6	12
3	12	36
4	3	12
5	1	5
	25	67

$$\overline{x} = \frac{\sum x_i \cdot f_i}{\sum f_i} = \frac{67}{25} = 2{,}68$$

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).

Три меры и тип шкалы

Три меры меры центральной тенденции накладывают ограничения на тип шкалы, в которой измеряется переменная.

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.

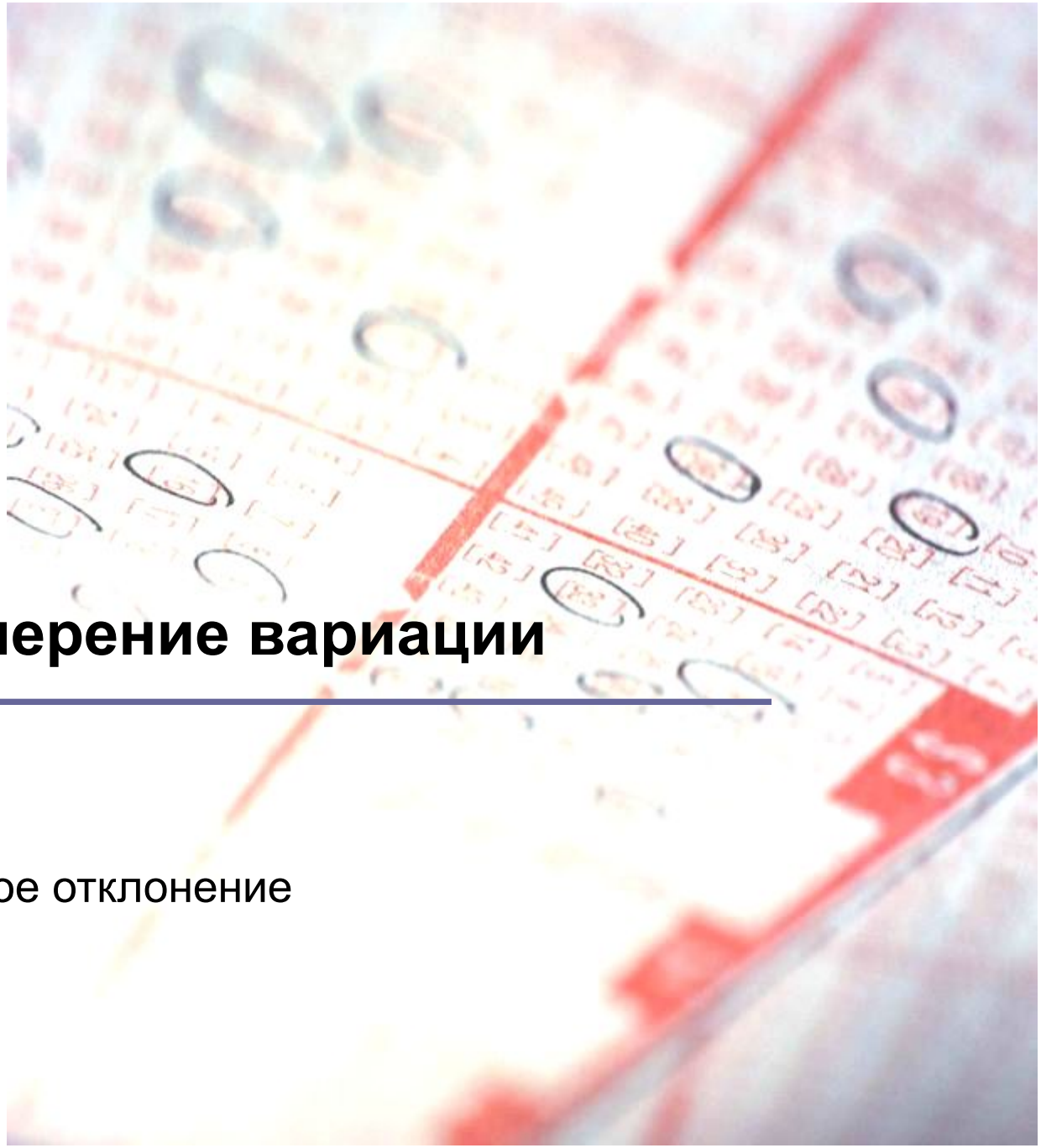
Какое типическое значение наилучшее?

1. «Наилучшее значение» - это такое значение, что для случайно взятого элемента выборки вероятность того, что переменная примет именно это значение, будет максимальной. **Мода**.

2. «Наилучшее значение» - это такое значение, что сумма абсолютных отклонений значений переменной от данного будет наименьшей. **Медиана**.

3. «Наилучшее значение» - это такое значение, что сумма квадратов отклонений значений переменной от данного будет наименьшей. **Среднее**.

В зависимости от данных каждое из трех значений может стать наилучшим.



3.2. Измерение вариации

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».

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.

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

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

Вторая формула для дисперсии

Дисперсия вычисляет также по равносильной формуле:

$$s^2 = \frac{n \cdot \sum x_i^2 - \left(\sum x_i\right)^2}{n \cdot (n-1)}$$

Считается, что эта формула более пригодна для практических вычислений при ручном счете и при использовании электронных таблиц.

Подсчет дисперсии в таблице

Пример вычисления дисперсии по второй формуле. В таблице рассчитываются лишь квадраты значений.

x	x_i^2
2	4
3	9
6	36
9	81
20	130

В первом столбце выборка. Во втором – квадраты значений. Сумма второго столбца есть сумма квадратов значений.

Не требуется вычислять среднее!!!

$$s^2 = \frac{n \cdot \sum x_i^2 - \left(\sum x_i\right)^2}{n \cdot (n-1)} =$$

$$= \frac{4 \cdot 130 - (20)^2}{4 \cdot (4-1)} = 10$$

Дисперсия для сгруппированных данных

Дисперсия для сгруппированных данных вычисляется по формуле:

$$s^2 = \frac{n \cdot \sum (f_i \cdot x_i^2) - \left[\sum (f_i \cdot x_i)\right]^2}{n \cdot (n-1)}$$

Вычисления удобно проводить при помощи таблицы или с помощью программных средств.

Пример вычисления дисперсии

Период	f_i	x	$x_i \cdot f_i$	$x_i^2 \cdot f_i$
2–4	2	3	6	18
5–7	5	6	30	180
8–10	10	9	90	810
11–13	4	12	48	576
14–16	2	15	30	450
20	23	45	204	2034

Рассчитаем дисперсию для сгруппированных данных, используя таблицу. В первом столбце – возраст службы, во втором – количество респондентов.

Используя вычисления в таблице, получим:

$$s^2 = \frac{n \cdot \sum (x_i^2 \cdot f_i) - \left[\sum (x_i \cdot f_i)\right]^2}{n \cdot (n-1)} = \frac{23 \cdot 2034 - (204)^2}{23 \cdot (23-1)} = 10{,}2$$

Стандартное отклонение

Стандартное отклонение вычисляется как корень из дисперсии:

$$s = \sqrt{s^2}$$

Стандартное отклонение имеет исключительную важность для описания распределения данных.

Интерпретация стандартного отклонения

На интервале с границами

$$\bar{x} \pm 2s$$

содержится, по крайней мере, 3/4 всех данных (75%).

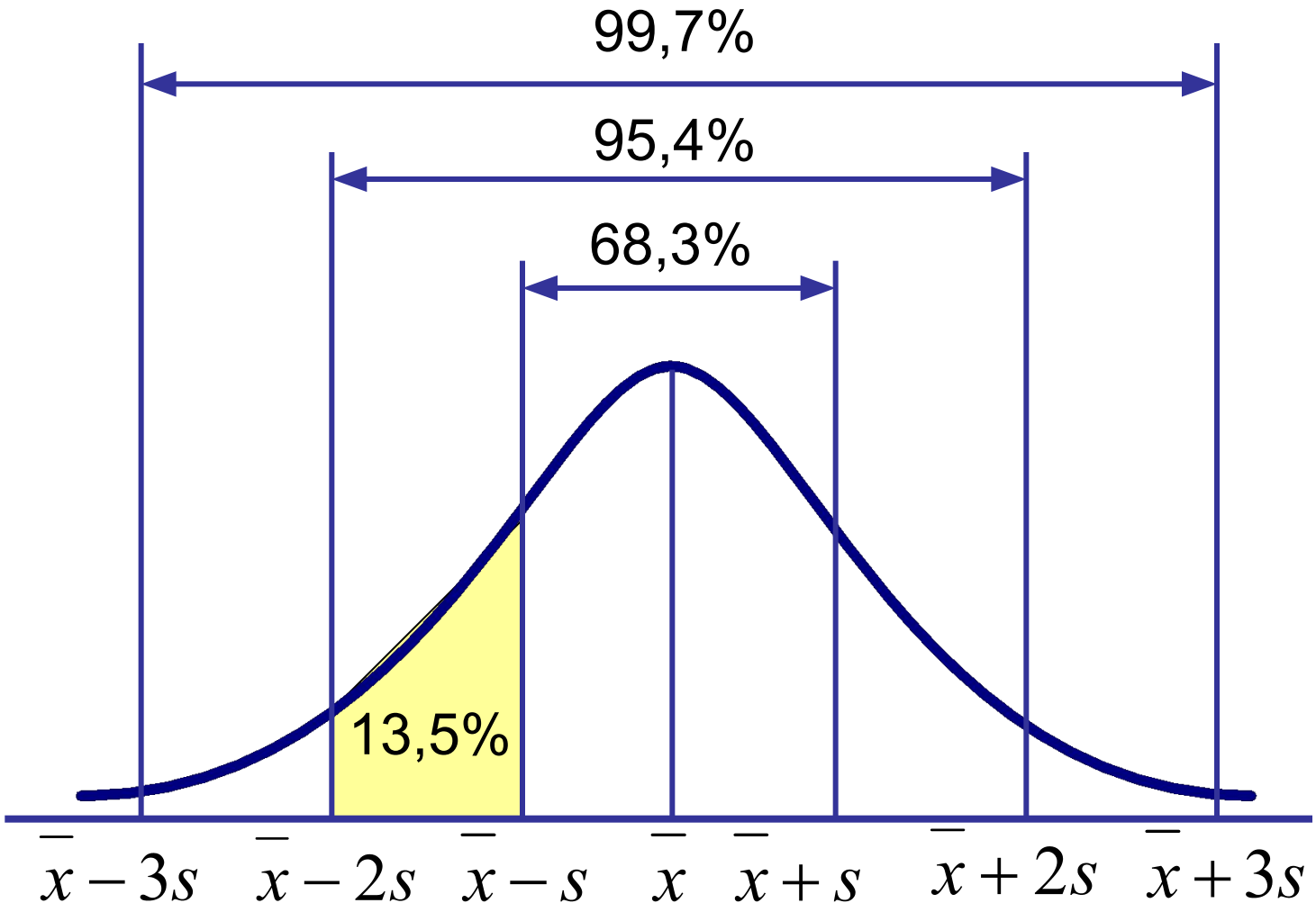
На интервале с границами

$$\bar{x} \pm 3s$$

содержится, по крайней мере, 8/9 всех данных (89,9%).

Это выполнено для любого распределения!!!

Стандартное отклонение для нормального закона



Коэффициент вариации

Коэффициент вариации вычисляется как отношение стандартного отклонения к среднему:

$$V = s / \bar{x}$$

Коэффициент вариации полезен, если:

1. Сравниваются несколько совокупностей, измеряемых в разных величинах.

2. Сравниваются совокупности, измеряемые в одинаковых величинах, но имеющие сильно отличающиеся средние.

Пример для коэффициента вариации

Какие данные имеют большую вариацию:

имеющие стандартное отклонение 20 при среднем 200 или
имеющие стандартное отклонение 3 при среднем 30?

$$V = s / \overline{x} = 20 / 200 = 0{,}1$$

$$V = s / \overline{x} = 3 / 30 = 0{,}1$$

Ответ. Коэффициенты вариации равны. Вариация одинакова.

По величине коэффициента вариации можно судить о степени вариации признаков совокупности. Чем больше его величина, тем больше разброс значений вокруг средней, тем менее однородна совокупность по своему составу и тем менее представительна средняя.

Если коэффициент вариации

меньше 10%, то изменчивость вариационного ряда незначительна, совокупность однородная, среднее значение – типичное для данной совокупности;

от 10% до 20%, то изменчивость вариационного ряда средняя, совокупность относительно однородная, среднее значение – часто встречающееся в данной совокупности;

больше 20% и меньше 33% то изменчивость вариационного ряда является значительной, совокупность переходная, среднее значение – редко встречающееся в данной совокупности;

превышает 33%, то совокупность неоднородна и необходимо исключить из рассмотрения самые большие и самые маленькие значения.