Тема 3.
Описательная статистика

3.1. Измерение центральной тенденции
3.2. Измерение вариации
3.3. Исследовательский анализ данных

Цели

После того, как мы познакомились с основными способами представления данных, изучим числовые характеристики, которые позволяют анализировать выборку и делать некоторые выводы.

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

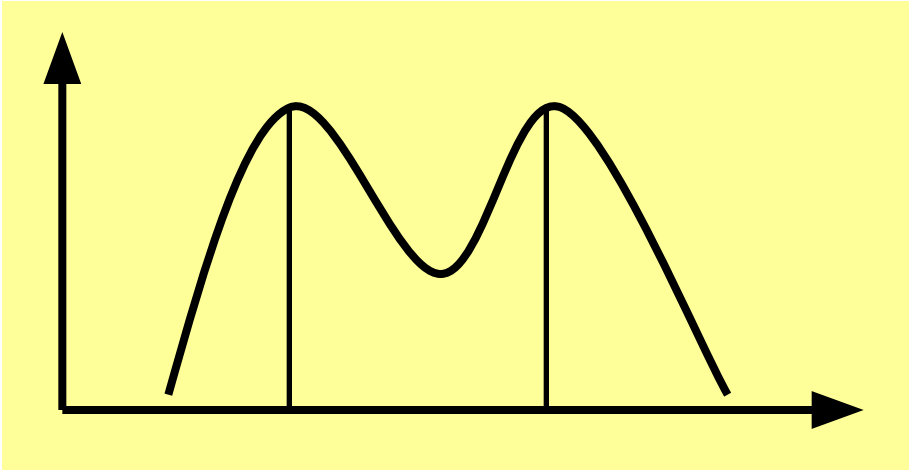
Мода – наиболее часто встречающееся значение в выборке, наборе данных. Обозначается ***Mo***.

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.

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**.

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:

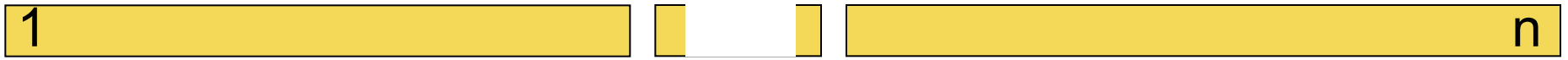
1,5 1,5 3,5 3,5 5 7 7 7 9

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:



Если n четно, медиана находится как среднее арифметическое двух соседних серединных элементов:

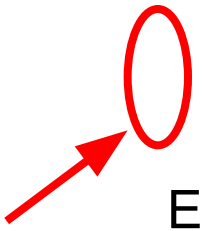
Пример вычисления медианы

Для набора данных из семи чисел:

6 1 3 7 1 7 3

После упорядочения получим вариационный ряд:

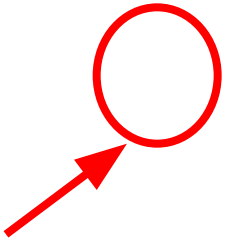
1 1 3 3 6 7 7



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:

где = (всех значений выборки

n = объем выборки

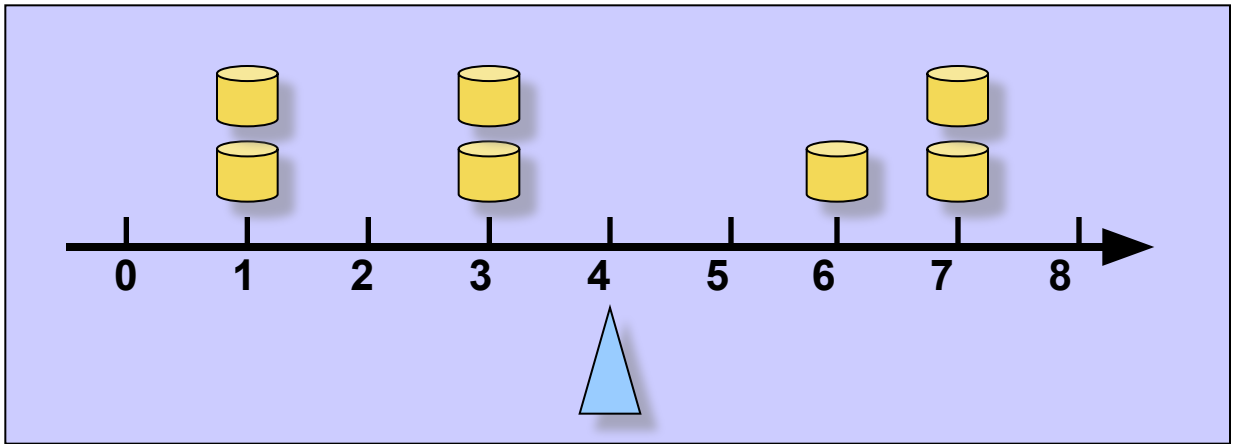
Индекс суммирования в статистической литературе часто опускается.

Пример вычисления среднего

Вычислим среднее для выборки из семи значений:

1 1 3 3 6 7 7

Получим:



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:

Среднее для сгруппированных данных

Среднее для сгруппированных данных вычисляется по формуле:

где сех значений выборки

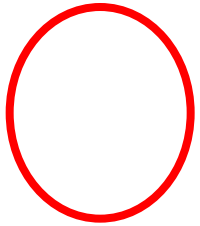
а частот, равна объему выборки

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

Пример вычисления среднего

Имеются результаты экзамена. Найти среднее значение.

x	f	f·x
0	1	0
1	2	2
2	6	12
3	12	36
4	3	12
5	1	5
	25	67



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

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)

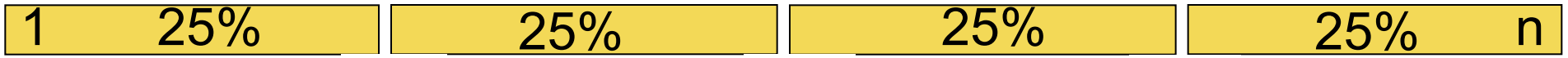
Размах – разность между наибольшим значением набора данных и наименьшим.

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)

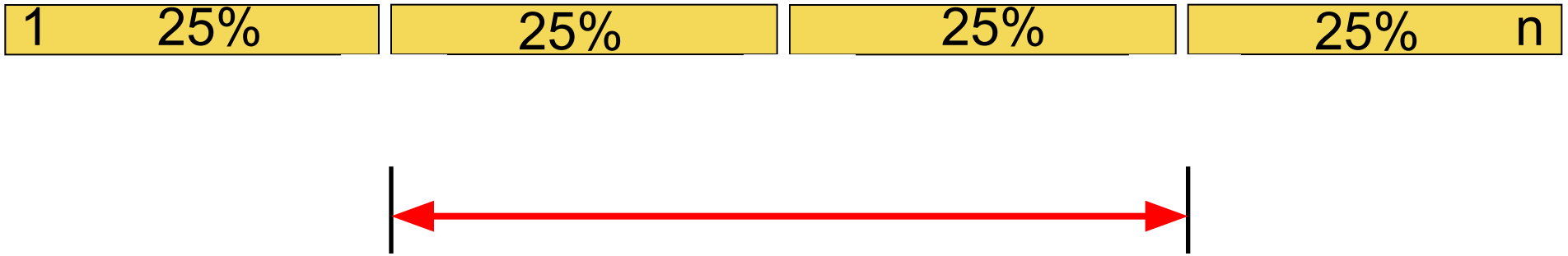
Под квартилями понимаются значения, которые делят вариационный ряд на четыре равные части:



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:

Свойства квартильного размаха



Если при вычислении размаха используются только наибольшее и наименьшее значения признака, а распределение данных между ними полностью игнорируется,

то при вычислении квартильного размаха игнорируются «крайние» данные, расположенные за пределами первого и третьего квартилей.

Между Q_1 и Q_3 расположены 50% всех данных.

Нахождение квартилей

Ранг нижнего квартиля:

Ранг верхнего квартиля:

Коробковая диаграмма (Boxplot)

Диаграмма, основывающаяся на вычислении и построении пяти характеристик. Удобна для анализа данных и используется очень часто.

Минимум Медиана Максимум

Рассмотрим на семинаре.

Процентили

Процентили это характеристики набора данных, которые выражают ранги элементов в процентах от 0% до 100%.

Процентили:

- Минимальное значение 0%
- Первый квартиль 25%
- Медиана 50%
- Третий квартиль 75%
- Наибольшее значение 100%

Процентили разбивают наборы количественных и порядковых данных на определенные части.

Дисперсия

Дисперсия выборки – среднее арифметическое квадратов отклонений значений выборки от их среднего.

Вычисляем по формуле:

Знаменатель делает оценку дисперсии несмещенной. Будет объяснено позже.

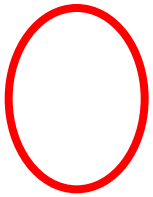
Подсчет дисперсии в таблице

Дисперсию удобно рассчитывать при помощи таблицы.

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.



Вторая формула для дисперсии

Дисперсия вычисляет также по равносильной формуле:

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.

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!!!

Дисперсия для сгруппированных данных

Дисперсия для сгруппированных данных вычисляется по формуле:

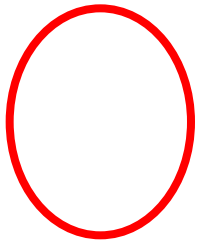
Вычисления удобно проводить при помощи таблицы или с помощью программных средств.

Пример вычисления дисперсии

Период				
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:



Стандартное отклонение

Стандартное отклонение вычисляется как корень из дисперсии:

Стандартное отклонение имеет исключительную важность для описания распределения данных.

Неравенство Чебышева

Для интерпретации стандартного отклонения применяют неравенство Чебышева. В терминах статистического исследования оно имеет следующую трактовку.

В любой совокупности доля значений, попадающих в интервал

будет равна, по крайней мере,

где k - любое число, большее 1.

Интерпретация стандартного отклонения

Исходя из приведенного выше, можно утверждать, что на интервале с границами

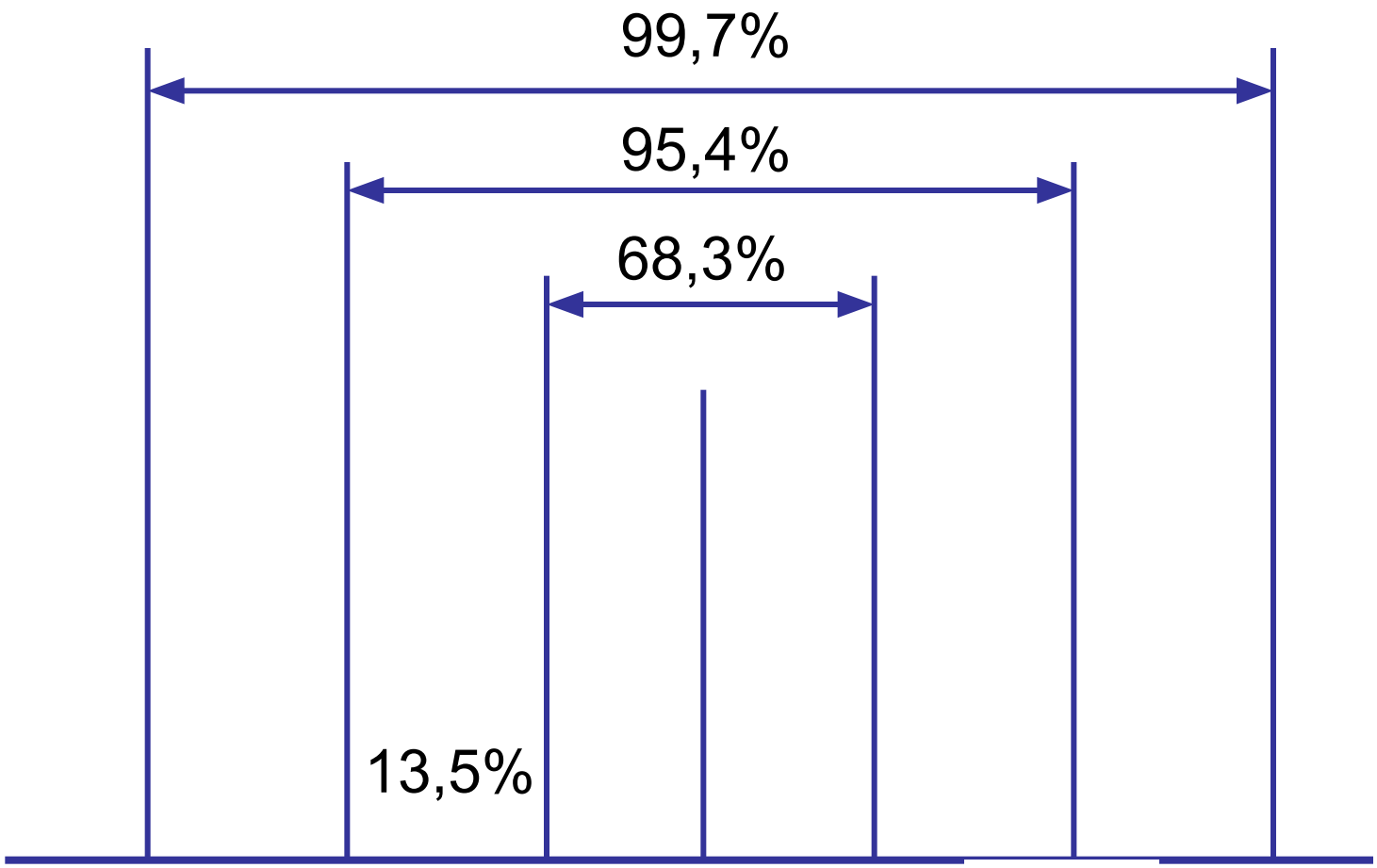
содержится, по крайней мере, 3/4 всех данных (75%).

На интервале с границами

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:

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?

Ответ. Коэффициенты вариации равны. Вариация одинакова.

3.3. Исследовательский анализ данных

Выбросы

Вид распределения

Разделы исследовательского анализа данных

Исследовательский анализ данных - Exploratory Data Analysis (EDA) представляет собой применение статистических методов для представления, упорядочения данных и понимания их важнейших характеристик.

Основными разделами анализа являются:

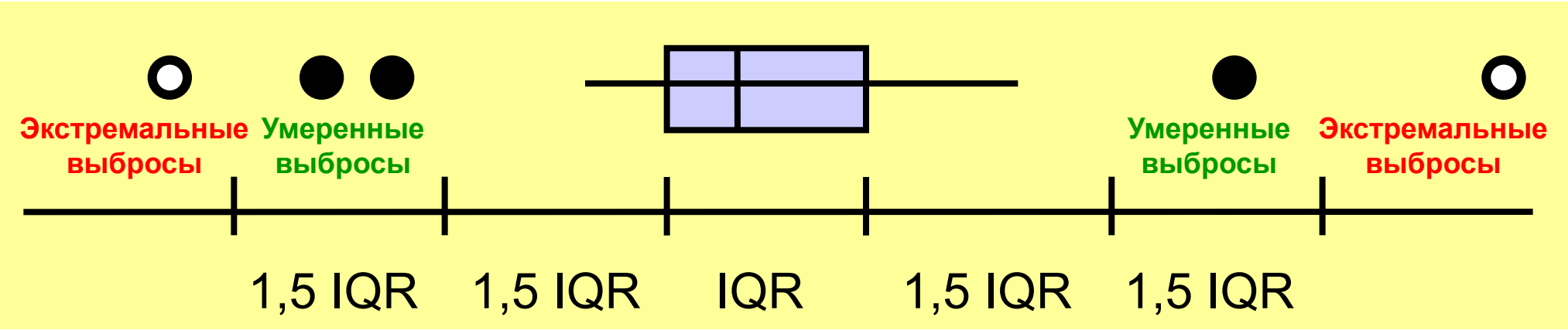
1. **Центральная тенденция**. Вычисление и анализ среднего, моды, медианы.
2. **Стандартное отклонение**. Нахождение дисперсии, стандартного отклонения.
3. **Квартили**. Минимум, максимум, размах, нахождение квартилей.
4. **Выбросы**. Нахождение и анализ выбросов.
5. **Форма распределения**. Асимметрия и куртозис.

Выбросы

Расширенная коробковая диаграмма строится с анализом выбросов. Для этого необходимо знать разброс квартилей IQR.

Умеренные выбросы изображаются темными точками и удалены ниже первой квартили или выше третьей от 1,5 IQR, но не более 3 IQR.

Экстремальные выбросы изображаются светлыми точками и удалены ниже первой квартили или выше третьей более 3 IQR.



Асимметрия (Skewness)

Если распределение симметрично, **асимметрия равна нулю**. В этом случае совпадают значения моды, медианы и среднего арифметического.

Если одно или несколько значений существенно превышают остальные, имеется **положительная асимметрия**. Средняя больше моды и медианы.

Если одно или несколько значений существенно меньше остальных, имеется **отрицательная асимметрия**. Средняя меньше моды и медианы.

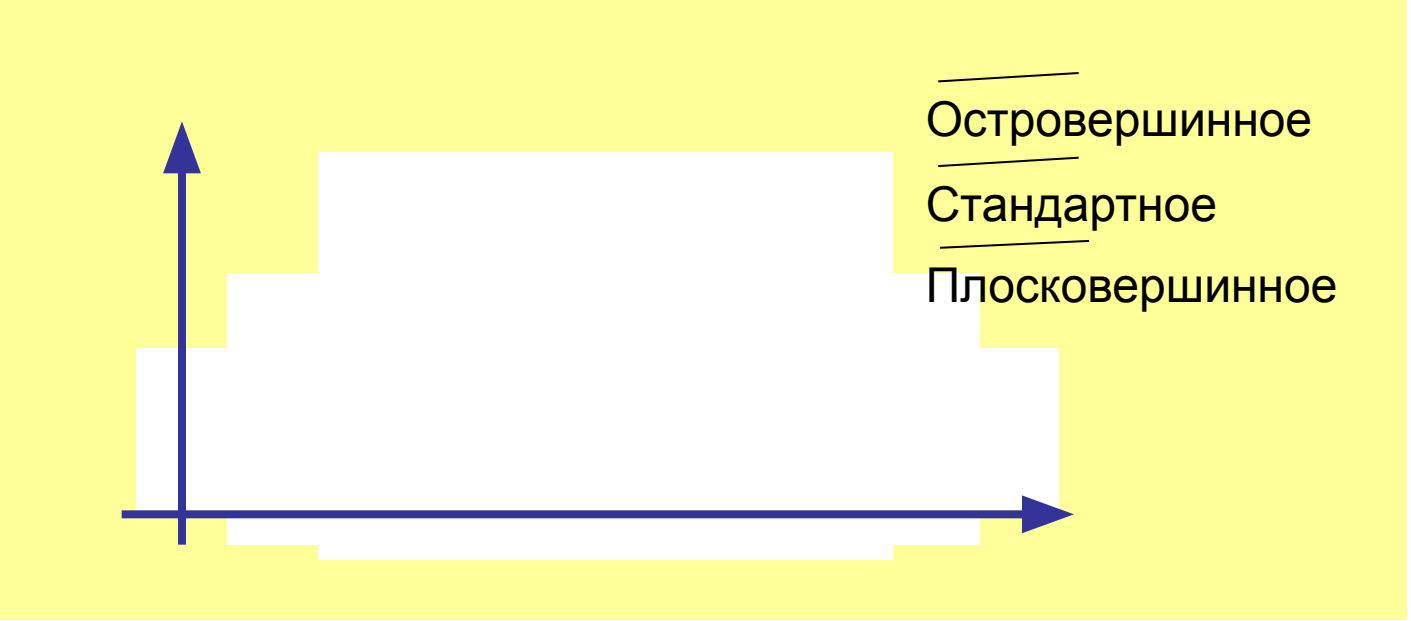
Коэффициент асимметрии

Коэффициент асимметрии находится по следующей формуле:

Изменяется в пределах от -3 до 3. Положителен при положительной асимметрии, отрицателен при отрицательной. Равен нулю, если асимметрия отсутствует.

Куртозис (Kurtosis)

Под **куртозисом** понимается крутость кривой распределения, которая определяется сопоставлением кривой с кривой стандартного нормального распределения.



Понятия и термины

Центральная тенденция
Мода, медиана, среднее
Вариационный ряд, ранжирование

Вариация, разброс
Размах
Квартиль, квартильный размах
Дисперсия, стандартное отклонение
Неравенство Чебышева
Коэффициент вариации

Выбросы
Асимметрия, коэффициент асимметрии
Куртозис

Задание на 5 минут (1)

Напишите своими словами, что такое визуализация данных. Назовите известные вам способы визуализации.

Задание на 5 минут (2)

В чем состоит **отличие размаха от квартильного размаха?** Определение не нужно. Нужно только отличие.

Задание. Актеры и актрисы

Имеются данные о возрасте актеров и актрис, в котором они были удостоены Оскара.

Актеры:

	32	37	36	32	51	53	33	61	35	45	55	39	76	37	42	40	32	60	38
56	48	48	40	43	62	43	42	44	41	56	39	46	31	47	45	60	46	40	36

Актрисы:

	50	44	35	80	26	28	41	21	61	38	49	33	74	30	33	41	31	35	41
42	37	26	34	34	35	26	61	60	34	24	30	37	31	27	39	34	26	25	33

Постройте коробковую диаграмму и сравните данные.

Задание. Актеры и актрисы. Решение

Задание. Актеры и актрисы. Решение (2)

Всего 39 значений.

Характеристика	Актеры	Актрисы
Минимум	31	21
Первая квартиль	37	30
Медиана	43	34
Третья квартиль	51	41
Максимум	76	80

Задание. Актеры и актрисы. Решение (3)

После построения сокращенной коробковой диаграммы, строим полную.

Несколько значений оказалось выбросами.

Например, актер 76 лет умеренный выброс.

Поскольку для актрис размах квартилей меньше, 80 и 74 года составили экстремальный выброс.

На семинарских занятиях…

Вычислять моду, медиану, среднее

Строить вариационный ряд и ранжировать

Вычислять размах, квартили, квартильный размах

Вычислять дисперсию, стандартное отклонение

Оценивать размещение данных при помощи неравенства Чебышева

Вычислять коэффициент вариации и сравнивать два набора данных

Определять выбросы

Описывать вид распределения, вычислять коэффициент асимметрии

Задачи

Найдите моду, медиану, среднее

Здесь приведено количество запросов, полученных Международной Финансовой Организацией в июле: 18, 12, 25, 16, 27, 32, 25, 15, 23, 22, 37, 16, 25, 19, 16, 25, 19, 16, 29, 38, 29, 30, 21.

Восемнадцать студентов получили следующие экзаменационные оценки за сочинение: 78, 62, 98, 90, 88, 73, 79, 86, 81, 84, 93, 97, 63, 59, 78, 82, 87, 93.

При тестировании 108 студентов коллежа были выявлены следующие показатели IQ:

IQ	**Частота**
90 - 98	6
99 -107	22
108-116	43
117-125	28
126-134	9

Задачи

Найти дисперсию и стандартное отклонение

Были отобраны пятнадцать студентов. Им был задан вопрос: «Сколько времени каждый студент тратит на подготовку к экзамену по статистике?» Их ответы записаны ниже (в часах): 8, 6, 3, 0, 0, 5, 9, 2, 1, 3, 7, 10, 0, 3, 6.

Стаж работы 75 служащих универмага:

1-5	21
6-10	25
11-15	15
16-20	0
21-25	8
26-30	6

Задачи

ПРИМЕР. Средняя цена зданий, расположенных в некотором районе, равна $50000, а стандартное отклонение - $10000. Найдите ценовой диапазон, в котором окажется, по крайней мере, 75% зданий.

Решение. В теореме Чебышева говорится, что 3/4 или 75% всех данных попадают в предел двух стандартных отклонений от среднего. Следовательно,

$50000 + 2 · ($10000) = $50000 + $20000 = $70000

и

$50000 – 2 · ($10000) = $50000 - $20000 = $30000

Следовательно, по крайней мере, 75% всех домов будет иметь ценовой диапазон от $30000 до $70000.

Задачи

Используя теорему Чебышева, решите следующие задачи для распределения со средним 80 и стандартным отклонением 10:

а. Какой процент данных попадет в интервал от 60 до 100?

б. Какой процент данных попадет в интервал с 65 и до 95?

Заработная плата простых служащих, работающих в ресторанах большого города, составляет в среднем $5,02 в час со стандартным отклонением $0,09. Используя теорему Чебышева, найдите диапазон, в котором расположено, по крайней мере, 75% данных.

Задачи

Средний бал на экзамене по английскому языку равен 85, со стандартным отклонением 5, а средний балл по истории - 110, со стандартным отклонением – 8. По какому предмету оценки более изменчивы?

Средний возраст бухгалтеров в корпорации «Три Реки» - 26 со стандартным отклонением 6, а среднее жалованье составляет $31000 со стандартным отклонением $4000. Сравните вариацию возраста и дохода.

Задачи

Следующее распределение показывает вес 18-летних парней. Построив график, найдите приблизительные значения веса, соответствующие следующим процентилям: 9, 45, 75, 20, 60.

Вес (фунты)	**Частота**
120,5-131,5	12
131,5-142,5	16
142,5-153,5	24
153,5-164,5	48
164,5-175,5	62
175,5-186,5	21
186,5-197,5	17