**ЛЕКЦИЯ 2 ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ**

**СТАТИСТИКИ**

*1 Средние величины*

*2 Общие свойства средних величин*

*3 Виды средних величин*

*4 Мода. Медиана. Квантили, квартили, децили, перцентили.*

5 Разнообразие значений признака

## 1 Средние величины

Из всех групповых свойств наибольшее теоретическое и практическое значение имеет **средний уровень, измеряемый средней величиной признака**.

Средняя величина признака – понятие очень глубокое, появившееся в науке и **практике только на определенном этапе развития** человеческого мышления. Всякая средняя величина обладает тремя основными свойствами: срединным положением, абстрактностью (отвлечение от реально существующего разнообразия) и единством суммарного действия.

**Средняя величина признака** определяется различными способами в зависимости от объектов наблюдения, изучаемых признаков и целей исследования. Поэтому имеется не одна, **а несколько средних**: средняя арифметическая, средняя геометрическая, средняя квадратическая, средняя гармоническая, мода, медиана.

**Основной показатель – средняя величина** – широко используется и в науке, и в практике. При изучении растений, животных, микроорганизмов и человека **расчет средних показателей** составляет основу обработки первичных материалов.

**Средние размеры особей** служат для характеристики видов, разновидностей, сортов, пород и других биологических групп; средние показатели физиологических процессов характеризуют интенсивность различных сторон обмена, силу действия биологических агентов и медицинских препаратов.

В производстве средние показатели используются для оценки работы отдельных специалистов, хозяйств, областей.

Средняя величина какого-нибудь признака определяется для того, чтобы получить характеристику **этого признака для всей изучаемой группы** в целом.

В зависимости от объекта наблюдения и от поставленных целей используются в биологии не одна, а несколько средних величин: **средняя арифметическая, средняя геометрическая, средняя квадратическая, средняя гармоническая**. Кроме того, для характеристики биологических групп иногда употребляются **мода и медиана**.

## 2 Общие свойства средних величин

## Для правильного использования средних величин необходимо знать свойства этих показателей: срединное расположение, абстрактность и единство суммарного действия.

По своему численному значению все средние величины занимают промежуточное положение между минимальным и максимальным значениями признака. При этом наименьшую величину имеет средняя гармоническая (H), а наибольшую – средняя квадратическая (S), что можно представить следующей схемой:



**Средняя признака** показывает, какую величину имел бы каждый из представителей изучаемой группы, если бы все они были одинаковыми и суммарное их действие было такое же, как и фактических неусредненных значений этой группы. При использовании средних величин предполагается, что пока они применяются, разнородная группа заменена однородной группой, в которой все значения признака одинаковы и равны средней величине.

Например, если имеется пять значений признака: 1; 4; 5; 5; 5 со средней величиной  = 4, то при использовании этой средней предполагается, что разнородная группа заменена на однородную с одинаковыми значениями: 4; 4; 4; 4; 4.

Эта особенность средних величин лежит в основе таких обычных производственных выражений, как «от каждой коровы получено по 3000 л молока», «с каждого гектара получено по 500 ц свеклы», «с каждого улья получено по 80 кг меда.

Коровы дают, конечно, различные удои, на разных участках получен разный урожай. И все же для производственной характеристики хозяйства и, особенно, для плановых расчетов оказалось удобным условно принять, что все коровы дали или будут давать одинаковый удой, равный средней величине этого признака для данного стада и года («от каждой коровы»), или, что с каждого гектара получен один и тот же урожай, равный среднему урожаю с общей площади («с каждого гектара»).

**Заменить разнородную группу однородной** можно только путем отвлечения от тех различий, которые существуют **в действительности**. Только абстрагируясь от имеющихся индивидуальных разнообразных значений, можно дать требуемую характеристику группы одним числом – *средней величиной признака*. **В этом смысле всякая средняя величина есть, прежде всего, абстрактная величина, которая часто в действительности не существует, а иногда и не может существовать**.

Если средний вес особей какого-нибудь вида в определенных условиях равен 40,57 кг, то существование такого экземпляра возможно, но крайне мала вероятность того, что какая-нибудь определенная особь будет весить точно 40,57 кг.

Если в совхозе среднее количество ягнят, полученных на одну овцематку, **равнялось 1,7 ягненка**, то такое число живых ягнят вообще не может существовать в действительности, тем не менее, эта средняя имеет вполне определенное производственное значение, например при сравнении этого совхоза с другим, где аналогичная средняя равна 1,2.

##

## 3 Виды средних величин

## 3.1 Средняя арифметическая

**Средняя арифметическая** – это сумма всех известных вариант, разделённых на общее количество вариант:



Пример

Три параллельных определения содержания гемоглобина в крови у одного и того же животного в одно и то же время, проведенные тремя разными лаборантами, дали такие результаты: 75; 80; 70. Наиболее вероятное содержание будет равно средней арифметической из параллельных проб:



*Свойства средней арифметической:*

а) *сумма центральных отклонений равна нулю*:

,

**Пример**

Для значений 1; 4; 5; 5; 5 средняя арифметическая μ = 4. Центральные отклонения будут следующие:

1–4 = –3, 4–4 = 0, 5–4 = +1, 5–4 = +1, 5–4 = +1,

а сумма центральных отклонений: –3+0+1+1+1 = 0.

Это свойство средней арифметической используется для проверки правильности ее расчета: если оказалась неравной нулю, значит, допущена ошибка в вычислениях.

б) *если каждое значение признака увеличить или уменьшить на (в) постоянную величину a, то средняя арифметическая из измененных вариантов будет равна средней арифметической из первоначальных вариантов, увеличенных (или уменьшенных) на (в) величину a*:

,

**Примеры**

- если в разбираемом примере к каждой из первоначальных вариантов 1; 4; 5; 5; 5 **прибавить** 3, то для полученных величин 4; 7; 8; 8; 8 среднее μ = 7 на 3 больше первоначальной средней μ = 4. Если в этой группе из каждого значения **вычесть**, например, 1, то для уменьшенных значений 0; 3; 4; 4; 4 средняя μ = 3 будет на 1 меньше первоначальной средней μ = 4.

- если каждое значение **умножить** на постоянное число *a*, то средняя арифметическая из измененных вариантов будет точно в *a* раз больше первоначальной средней арифметической. Если в разбираемом примере все значения 1; 4; 5; 5; 5 умножить на 10, то для полученных увеличенных вариантов (10; 40; 50; 50; 50) средняя арифметическая μ = 40 ровно в 10 раз больше той, которая получена для неувеличенных вариантов (μ= 4).

- если *a* равно дробному числу, то каждое значение, а также и каждая средняя будут уменьшены во столько же раз. Если в разбираемом примере все значения умножить на 1/5, то средняя арифметическая из уменьшенных вариантов (0,2; 0,8; 1; 1; 1) μ = 0,8 в 5 раз меньше средней арифметической, полученной для неизменных значений (μ = 4).

## 3.2 Средний ранг (непараметрическая средняя)

Средний ранг определяется для таких признаков, для которых еще не найдены способы количественного измерения. По степени проявления таких признаков объекты могут быть ранжированы, т. е. расположены в порядке усиления (или ослабления) выраженности признака. Порядковый номер объекта в таком ряду называется его рангом.

Пример

В зверосовхозе, разводящем голубых норок, получено от двух самцов и одной и той же группы самок 20 щенков с различной окраской меха: от почти белого до темно-голубого. Требовалось выяснить, какой из производителей дает в потомстве более темную окраску меха. Затруднением при этом служит то обстоятельство, что нет способа числового измерения интенсивности окраски волоса у норок. Все потомки оцениваемых производителей были распределены в ранжированный ряд в порядке усиления серого цвета, причем при каждом порядковом номере (ранге) такого ряда был поставлен номер отца (I, II; таблица 1).

Таблица 1 – Ранжированный ряд

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Ранг | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| Отец | I | II | I | II | I | I | II | I | I | I | II | I | II | II | II | II | I | II | II | II |

На основе такого ряда можно рассчитать средние ранги окраски в потомстве каждого производителя и по этим показателям сравнить их:

;



Вывод

Второй производитель дает в потомстве более темную окраску меха.

**3.3 Средняя взвешенная арифметическая**

Обычно, чтобы рассчитать среднюю арифметическую, складывают все значения признака и полученную сумму делят на число вариантов. В этом случае каждое значение, входя в сумму, увеличивает ее на полную свою величину. Но не всегда это возможно. Иногда значения признака должны входить в сумму с неодинаковой поправкой. Эта поправка, выраженная определенным множителем, называется математическим весом значения.

Средняя, рассчитанная для значений признака с неодинаковыми весами, называется взвешенной средней. Взвешенная средняя арифметическая рассчитывается по следующей формуле:

, (6.4)

где:

Xi – значение признака, варианта;

p – математический вес усредняемого значения.

Чтобы рассчитать взвешенную среднюю арифметическую, необходимо каждое значение признака помножить на его вес, все эти произведения сложить и полученную сумму разделить на сумму весов.

Пример

Имеются результаты двух исследований длины хоботка пчел: в одном случае получена средняя длина хоботка 6,6 мм, в другом – 6,0 мм. Требуется получить общую среднюю, причем известно, что в первом исследовании были измерены хоботки у 100 пчел, во втором – у 20.

В данном случае значениями признака являются средние μ1 = 6,6 и μ2 = 6,0 мм; их весами – численности групп n1 = 100 и n2 = 20. Взвешенная средняя арифметическая рассчитывается следующим образом:

.

**3.4 Средняя квадратическая**

Средняя квадратическая вычисляется по формуле:

,

Она равна корню квадратному из суммы квадратов данных, деленной на их число.

Например, если имеется пять вариантов: **1, 4, 5, 5, 5**, то средняя квадратическая:

.

Употребляется средняя квадратическая при расчете средних радиусов окружностей.

##

## 3.5 Средняя геометрическая

Чтобы получить среднюю геометрическую для группы с n данными, нужно все варианты перемножить и из полученного произведения извлечь корень n-й степени:

,

где:

G – средняя геометрическая; n – число значений; ΠXn – произведение вариантов.

Пример

Средняя геометрическая из 12 и 3 будет равна: .

Если число значений больше двух, то извлечение корня n-й степени затруднительно, поэтому обычно значение средней геометрической находят путем логарифмирования величин, входящих в основную формулу:

.

##

## 3.6 Средняя гармоническая

Средняя гармоническая рассчитывается по формуле

.

Для пяти вариантов: 1, 4, 5, 5 средняя гармоническая:

.

Применяется средняя гармоническая при усреднении меняющихся скоростей.

Пример

Почтовые голуби одной станции к месту кормежки летят со скоростью 50 км/час, а в обратном направлении – со скоростью 40 км/час. Если ничего больше неизвестно и требуется выяснить среднюю скорость полета для обоих направлений (расстояния, очевидно, равны), то сделать это можно, рассчитав простую среднюю гармоническую для двух значений 50 и 40:

.

**4 Мода. Медиана. Квантили (квартили, децили, перцентили)**

**4.1 Мода**

***Мода****,* или***модус*** *–* это такая варианта или класс распределения вариант, который в исследуемой группе особей встречается наиболее часто. В качестве примера рассмотрим распределение, представленное в таблице.

Таблица 2 – Пример распределения

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Классы  | 100– 119 | 120– 139 | 140– 159 | 160– 179 | 180– 199 | 200– 219 | 220– 239 | 240– 259 | 260– 279 | 280– 299 | 300– 319 |
| Частоты  | 2 | 20 | 60 | 160 | 250 | 240 | 180 | 70 | 15 | 2 | 1 |

В этом распределении наиболее многочисленным является пятый класс (180–199) с частотой 250. Это модальный класс.

В качестве первого приближения можно принять за моду средину **модального** **класса, т. е. 190**.

Более точное значение моды можно получить по формуле:

,

где: М0 – мода; Wα – начало модального класса; k – величина классового промежутка; f1 – частота класса, предшествующего модальному; f2 – частота модального класса; f3 – частота класса, следующего за модальным.

В некоторых распределениях встречаются **два** или **три модальных** класса. Иногда это может быть следствием того, что в изучаемую группу попал разнородный материал, относящийся к разным категориям (более крупной и менее крупной) по изучаемому признаку.

**4.2 Медиана**

***Медиана*** – это такое значение признака, которое разделяет всю группу на две равные части: одна часть имеет значения признака меньшее, чем медиана, а другая – большее.

Например, если имеется группа из 9 значений признака; 1, 2, 3, 4, 5, 6, 7, 8, 9, то медианой этой группы будет 5.

Для многочисленных групп медиану можно рассчитать по формуле:

,

где: Ме – медиана; Wα – начало того класса, в котором находится медиана; k – величина классового промежутка; n – общее число данных в группе;  – сумма частот классов (начиная с меньшего), предшествующих классу, в котором находится медиана; f – частота класса, в котором находится медиана.

## 4.3 Квантили, квартили, децили, перцентили

***Квантиль*** – это конкретная варианта совокупности, отсекающая в пределах вариационного ряда определенную часть его членов. Медиана делит распределение на две равные части. Таким же образом распределение можно разделить тремя ***квартилями****,* чтобы по­лучить 4 области равного размера или вероятности, или девятью ***децилями****,* чтобы получить 10 областей равного размера или вероятности, или 99 ***перцентилями***чтобы получить 100 областей равного размера или вероятности. При этом 50-й перцентиль является медианой и вместе с 25-м и 75-м перцентилями даст нам квартили.

На практике обычно используют перцентили. Причем перцентили Р25 и Р75 соответствуют первому и третьему квартилям, между которыми содержится 50% элементов выборки.

**5 Разнообразие значений признака**

Всякая группа состоит из **особей** или **объектов**, отличающихся друг от друга по каждому из признаков. Различия эти иногда очень велики, иногда они почти незаметны, но они всегда имеются в группе, так как невозможно найти даже двух абсолютно одинаковых особей. Это второе основное свойство всякой группы – состоять из неодинаковых объектов по любому признаку – точнее всего определяется термином разнообразие признака в группе.

Часто этому групповому свойству даются другие названия: изменчивость, рассеяние, колеблемость, вариабильность, разброс.

## 5.1 Стандартное (среднеквадратическое) отклонение

Степень разнообразия особей в группе по изучаемому признаку измеряется несколькими показателями (**лимиты и размах, стандартное отклонение, коэффициент вариации, квартили, децили, перцентили**), из которых наибольшее значение имеет стандартное, или среднеквадратичное отклонение:

; 

σ – стандартное отклонение; x2 – сумма квадратов центральных отклонений, т. е. квадратов разностей между каждым значением и средней арифметической; Xi – значение признака у каждого объекта в группе; μ – средняя арифметическая признака для данной группы; n–1 – число степеней свободы, равное числу объектов в группе без одного.

**Число степеней свободы**. Число степеней свободы равно числу элементов свободного разнообразия в группе. Оно равно числу всех имеющихся элементов изучения без числа ограничений разнообразия.

Например, для исследования требуется взять 3 объекта с любым развитием изучаемого признака. В данном случае величина признака не имеет никаких ограничений, поэтому число степеней свободы будет *n=3–0=3*.

Если же требуется взять 3 числа с условием, что сумма их должна быть равна определенной величине, например: 100, то первое число может быть любой величины: 80, 800 и т. д., второе число также может быть выбрано свободно без всяких ограничений, например 10, 1269 и т. д., то третье число может иметь только одно значение – такое, чтобы оно вместе с двумя предыдущими составило бы в сумме 100. Если два первых числа были 80 и 10, то третье должно быть 10; если два первых числа 800 и 1269, то третье должно быть отрицательным: –1969 (800+1269–1969 = 100).

В данном случае, при одном ограничении (сумма чисел должна быть равна 100), **два числа выбираются свободно, а третье не имеет свободы выбора: для трех чисел имеются две степени свободы df = 3 – 1 = 2.**

Для *n* значений при *k* ограничениях имеется *df=n–k* степеней свободы.

При вычислении средней арифметической никаких ограничений величины значений признака не имеется. Поэтому число элементов, образующих среднюю арифметическую, равно числу вариантов.

При вычислении стандартного отклонения имеется одно ограничение – сигма вычисляется для определенной группы, имеющей определенную среднюю арифметическую. Поэтому разнообразие элементов, образующих стандартное отклонение, ограничено этим одним условием и в данном случае число степеней свободы равно числу вариантов без одного df = n–1.

**Стандартное отклонение служит основным показателем разнообразия значений признака в группе**. Используется сигма и как самостоятельный показатель, и как основа для построения многих других показателей статистики: коэффициента вариации, ошибок репрезентативности, различных показателей распределения, коэффициентов корреляции и регрессии, элементов дисперсионного анализа, формул регрессии. Следует иметь в виду, что формула сигмы с числом степеней свободы в знаменателе подкоренного количества применяется только для выборок:



Сигма генеральной совокупности вычисляется по формуле:

,

где:  – генеральная сигма;  – генеральная средняя; N – объем генеральной совокупности

Пример

На сельскохозяйственной выставке сравниваются экспонаты двух хозяйств, представивших лучшие экземпляры тыквы со своих огородов. Первое хозяйство представило 6 тыкв, весивших 33, 37, 32, 38, 34, 36 кг, второе представило 5 тыкв, весивших 33, 37, 34, 36, 35 кг.

Так как средний вес экспонатов оказался одинаковым у обоих экспонатов (μ1 = 35, μ2 = 35), было решено провести сравнение стандартности тыкв по среднему стандартному отклонению.

В данном случае сравнивались не выборки, а генеральные совокупности, так как оценка проводилась по выставочным экземплярам, которые целиком исчерпывали всю требующуюся в данном случае информацию.

Дисперсии и средние стандартные отклонения (сигмы):

  (а не );

  (а не ).

; 

Коэффициенты вариации:

; .

Оказалось, что второй совхоз представил более стандартную партию тыкв.

Так как сравнение в данном случае проводилось по генеральным параметрам, то разность сигм заведомо достоверна и не нуждается в определении достоверности. Обычные методы определения достоверности разности в данном случае не нужны и неприменимы, так как сравниваемые группы выбирались не так, как это требуется при организации выборок (не рандомизированно).

Сравнение стандартности двух партий тыкв характеризует более благоприятно второе хозяйство, которое смогло подобрать более выровненную группу экспонатов.

**5.2 Коэффициент вариации**

Стандартное отклонение – величина именованная, выраженная в тех же единицах измерения, как и средняя арифметическая.

Поэтому для сравнения разных признаков, выраженных в разных единицах измерения, используется не абсолютное, а относительное значение среднего квадратического отклонения в форме коэффициента вариации:

,

где: V – коэффициент вариации; σ – стандартное отклонение; μ – средняя арифметическая.

Коэффициент вариации есть сигма, выраженная в процентах от средней арифметической. Этот показатель неименованный, поэтому он и пригоден для сравнения разных признаков или одного и того же признака, но в группах с резко различной средней величиной признака.

Пример

Если в стаде коров сводные показатели по удою за первую лактацию равны μ1 = 3000 кг, σ1 = 600 кг, а по живому весу μ2 = 400 кг, σ2 = 20 кг, то коэффициенты вариации будут иметь следующую величину:

по удою ;

по живому весу 

Сопоставление коэффициентов вариации указывает на сильное разнообразие первотелок изученного стада по удою и слабое разнообразие их по живому весу.

**5.3 Лимиты и размах**

Для быстрой и примерной оценки степени разнообразия часто применяются простейшие показатели:

lim = {min ÷ max} – лимиты, т. е. наименьшее и наибольшее значения признака,

p = (max – min) – размах, или разность между лимитами.

Для группы данных 1, 2, 3, 4, 5 лимиты и размах могут быть обозначены так:

lim = 1 ÷ 5.

Иногда характеристика разнообразия группы в форме лимитов имеет столь большое производственное значение (например, при упаковке яблок, помидор и т. д., при оценке партии беконных тушек), что кладется в основу денежной оценки продукта.

**5.3 Нормированное отклонение**

Обычно степень развития признака определяется путем его измерения и выражается определенным именованным числом: 3 кг веса, 15 см длины, 20 зацепок на крыле у пчел, 4% жира в молоке, 700 г привеса в сутки и др. Этот основной способ характеристики признаков оказывается недостаточным, когда требуется еще и оценить полученное значение, т. е. определить, можно ли его считать значительным или, наоборот, недостаточным, или находящимся в норме.

Задачи подобного рода возникают очень часто как в научных работах, так и в производственных условиях.

Для решения таких задач применяется общий принцип: ценность особи определяется ее отношением к группе, что может быть измерено особым показателем – ***нормированнным отклонением***, вычисляемым по формуле:

,

где:  – нормированное отклонение; Xi – результат непосредственного измерения признака; μ – средняя арифметическая соответствующей группы, из которой взята изучаемая особь; σ – стандартное отклонение этого признака в группе.

Нормированное отклонение показывает, на сколько сигм отклоняется значение признака от средней для соответствующей группы.

Пример

Если мальчик 14 лет имеет длину тела 159 см, обхват груди 64 см, емкость легких 2070 см3, обхват головы 52 см и вес 40 кг, то по этим цифрам невозможно дать сравнительную характеристику развития разных частей тела и определить общий тип физического развития ученика.

Для этой цели необходимо сравнить индивидуальные показатели со средними характеристиками мальчиков этого возраста по среднему уровню каждого признака и его разнообразию в форме нормированного отклонения.

##

## 5.4 Проверка выпадов (артефактов)

Нормированное отклонение помогает определить выпады, или артефакты, т. е. такие записанные значения признака, которые резко отличаются от всех других значений признака в группе. Проверка артефактов должна проводиться всегда перед началом обработки полученных первичных данных. Если подтвердится, что резко выделяющееся значение действительно не может относиться к объектам данной группы, и попало в записи вследствие ошибок внимания, следует такой артефакт исключить из обработки.

Проверка артефактов может производиться по критерию, равному нормированному отклонению выпада:

,

где: Т – критерий выпада;** – выделяющееся значение признака (или очень большое или очень малое); μ, σ – средняя и сигма, рассчитанные для группы, включающей артефакт; Tst – стандартные значения критерия выпадов, определяемых по таблице.

Таблица 3 – Стандартные значения критерия выпадов (Tst)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| n | Tst | n | Tst | n | Tst | n | Tst |
| 2 | 2,0 | 16 – 20 | 2,4 | 47 – 66 | 2,8 | 125 – 174 | 3,2 |
| 3 – 4 | 2,1 | 21 – 28 | 2,5 | 67 – 84 | 2,9 | 175 – 349 | 3,3 |
| 5 – 9 | 2,2 | 29 – 34 | 2,6 | 85 – 104 | 3,0 | 350 – 599 | 3,4 |
| 10 – 15 | 2,3 | 35 – 46 | 2,7 | 105 – 124 | 3,1 | 600 – 1500 | 3,5 |

Если Т ≥ Tst, то анализируемое значение признака является артефактом. Альтернатива Т < Tst не позволяет исключить из анализа значение признака.

Табулированные данные таблицы 3 можно аппроксимировать следующей функцией: Tst = 0,287⋅ln(n) + 1,714

Пример

Данные: 1, 2, 3, 10; n = 4, μ = 4, σ = 4, ; 10 еще не может считаться выпадом.

Данные: 1, 2, 2, 3, 3, 4, 4, 5, 21; n = 9, μ = 5, σ = 6,1, ;

21 может считаться выпадом и должна быть исключена из обработки.

##

## 5.5 Скошенность (асимметрия) и крутизна (эксцесс) кривой распределения

Для больших выборок (n > 100) вычисляют еще два статистических показателя.

Скошенность кривой называется ***асимметрией***:



Правосторонняя асимметрия – отрицательна, левосторонняя – положительна.

|  |  |
| --- | --- |
|  |  |
| асимметрия | эксцесс |

Отклонение крутизны называют ***эксцессом***:



Эксцесс положителен при островершинной кривой, отрицателен при плосковершинной.