ГРУППИРОВКА ДАННЫХ

Числа округляют следующим образом: если за последней сохраняемой цифрой следуют цифры О, 1, 2, 3, 4, они отбрасываются (округление с недостатком); если же за последней сохраняемой цифрой следуют цифры 5, 6, 7, 8 и 9, то последняя сохраняемая цифра увеличивается на единицу (округление с избытком). Например, числа 45,346; 8,644; 9,425; 3,585 и 3,575 округляются до двух десятичных знаков так: 45,35; 8,64; 9,43; 3,59 и 3,58.

Многие исследователи считают более точным такое правило: если за последней сохраняемой цифрой следует цифра 5 (с нулями или без них, следующими за цифрой), то округление осуществляется с недостатком при условии, что сохраняемая цифра четная. Если же сохраняемая цифра нечетная, то округление осуществляется с избытком. Например, числа 3,585 и 3,575 округляются до двух десятичных знаков таким образом: 3,58 и 3,58.

Обработка начинается с упорядочения или систематизации собранных данных. Процесс систематизации результатов массовых наблюдений, объединения их в относительно однородные группы по некоторому признаку называется группировкой.

Наиболее распространенной формой группировки являются статистические таблицы; они бывают простыми и сложными. К простым относятся, например, четырехпольные таблицы, применяемые при альтернативной группировке, когда одна группа вариант противопоставляется другой; например здоровые - больным, высокие - низким и т. д. В качестве примера такой группировки могут служить результаты обследований 265 учащихся младших классов на состояние нёбных миндалин (табл. 1).

Таблица 1.

Школьные классы	Обнаружено детей		Всего
	здоровых	больных	
Третьи и четвертые	63	92	155
Пятые и шестые	71	39	110
Всего	134	131	265

К сложным относятся многопольные таблицы, применяемые при изучении корреляционной зависимости и при выяснении причинно-следственных отношений между варьирующими признаками. Примером корреляционной таблицы служат классические данные Гальтона, показывающие наличие положительной зависьмости между ростом родителей и ростом их детей (табл.2).

Таблица 2.

Рост родителей, дюймы	Рост детей, дюймы								Всего
	60,7	62,7	64,7	66,7	68,7	70,7	72,7	74,7	
74							4		4
72			1	4	11	17	20	6	62
70	1	2	21	48	83	66	22	8	251
68	1	15	56	130	148	69	11		430
66	1	15	19	56	41	11	1		144
64	2	7	10	14	4				37
Всего	5	39	107	255	387	163	58	14	928

В качестве примера группировки, применяемой при выяснении причинно-следственных отношений между признаками, приведены данные, полученные в Научно-исследовательском институте имени В. В. Докучаева при испытании гречихи сорта «Богатырь» на урожайность в зависимости от предшественников (табл. 3).

Таблица 3.

Предшественники	Урожай гречихи по повторностям, ц/га			Средний урожай
	1	2	3	
Горох раннезеленый	23,7	20,1	20,5	21,4
Чечевица	23,6	25,1	21,1	23,2
Чина степная № 21	26,7	23,2	23,8	24,6
Ячмень	26,0	24,9	25,3	25,4

Из табл. 3 ясно, что в данных условиях лучшим предшественником для гречихи является ячмень.

Особую форму группировки представляют статистические ряды. Статистическим называется ряд числовых значений признака, расположенных в определенном порядке. В зависимости от того, какие признаки изучаются, статистические ряды делят на атрибутивные, вариационные, ряды динамики и регрессии, а также ряды ранжированных значений признаков и ряды накопленных частот, являющихся производными вариационных рядов. Примером атрибутивного ряда могут служить данные, показывающие зависимость между содержанием гемоглобина Hb в крови и высотой организации позвоночных животных (рис. 1).

Класс животных	Рыбы	Амфибии	Рептилии	Птицы	Млекопитающие
Количество Hb, г/кг массы тела	1,6	2,9	3,8	11,2	11,7

Рис. 1.

Вариационным рядом или рядом распределения называют двойной ряд чисел, показывающий, каким образом числовые значения признака связаны с их повторяемостью в данной статистической совокупности. Например, из урожая картофеля, собранного на одной из опытных делянок, случайным способом, т. е. наугад, отобрано 25 клубней, в которых подсчитывали число глазков. Результаты подсчета оказались следующие: 6, 9, 5, 7, 10, 8,9, 10,8, 11, 9, 12,9,8, 10, 11,9, 10,8, 10,7,9, 11,9, 10. Чтобы разобраться в этих данных, расположим их в ряд (в порядке регистрации результатов наблюдений) с учетом повторяемости вариант в этой совокупности (рис. 2).

Вариаиты x_i	6	9	5	7	10	8	11	12
Число иариант f_i	1	7	1	2	6	4	3	1

Рис. 2.

На рис. 2 представлен пример вариационного ряда. Числа, показывающие, сколько раз отдельные варианты встречаются в данной совокупности, называются частотами или весами вариант и обозначаются строчной буквой латинского алфавита *f*. Общая сумма частот вариационного ряда равна объему данной совокупности (формула 1).

$$\sum_{i=1}^{k} f_i = n$$

, где n – общее число наблюдений. [1]

Частоты (веса) выражают не только абсолютными, но и относительными числами - в долях единицы или в процентах от общей численности вариант, составляющих данную совокупность. В таких случаях веса называют относительными частотами или частостями. Общая сумма частостей равна единице, т. е. $\Sigma f_i/n=l$, или $\Sigma(f_i/n)100=l00\%$, если частоты выражены в процентах от общего числа наблюдений n.

Распределение исходных данных в вариационные ряды дает следующие преимущества:

1. ускорение работы при вычислении по вариационному ряду обобщающих числовых характеристик - средней величины и показателей вариации;
2. выявление закономерности варьирования учитываемого признака.

Второе преимущество достигается построением ряда распределения по ранжированным значениям признака.

Под ранжированием (от франц. ranger - выстраивать в ряд по ранжиру, т. е. по росту) понимают расположение членов ряда в возрастающем (или убывающем) порядке. Так, в данном случае результаты наблюдений следует распределить как на рис. 3.

Вариаиты x_i	5	6	7	8	9	10	11	12
Частоты f_i	1	1	2	4	7	6	3	1

Рис. 3.

Этот упорядоченный ряд распределения в равной мере удовлетворяет достижению и первой, и второй целей. Он хорошо обозрим и наилучшим образом иллюстрирует закономерность варьирования признака.

В зависимости от того, как варьирует признак – дискретно или непрерывно, в широком или узком диапазоне, - статистическая совокупность распределяется в безынтервальный или интервальный вариационные ряды. В первом случае частоты относятся непосредственно к

ранжированным значениям признака, которые приобретают положение отдельных групп или классов вариационного ряда, во втором - подсчитывают частоты, относящиеся к отдельным промежуткам или интервалам (от-до), на которые разбивается общая вариация признака в пределах от минимальной до максимальной варианты данной совокупности. Эти промежутки, или классовые интервалы, могут быть равными и не равными по ширине. Отсюда различают равно- и неравноинтервальные вариационные ряды. Примером неравноинтервального ряда распределения могут служить данные А. Ф. Ковшарь, показывающие зависимость между числом стай сизых голубей и количеством особей в стае в гнездовой (с 15 марта по 15 августа) и послегнездовой (с 15 августа по 15 марта) периоды их жизни (табл.4).

Таблица 4.

Число особей в стае	Чнсло встреч (частота)				Плотность распределения			
	в гнездовой пернод		в остальное время года		в гнездовой период		в остальное время года	
	абсолютные значения	значения в процентах	абсолютные значения	значения в процентах	абсолютная	относитель-ная	абсолютная	относитель-ная
Одиночки	6	18,20	1	1,70	6,00	18,20	1,00	1,70
2—5	19	57,60	9	15,25	6,33	19,20	3,00	5,08
5—10	4	12,10	4	6,78	0,80	2,42	0,80	1,36
10—20	2	6,10	12	20,34	0,20	0,61	1,20	2,03
20—30	1	3,00	13	22,03	0,10	0,30	1,30	2,20
30—50	1	3,00	11	18,65	0,05	0,15	0,55	0,93
50—100	0	0,00	9	15,25	0,00	0,00	0,18	0,31
Всего встреч	33	100	59	100	—	—	—	—

Приступая к построению равноинтервального вариационного ряда, важно правильно наметить ширину классового интервала. Дело в том, что грубая группировка (когда устанавливают очень широкие классовые интервалы) искажает типичные черты варьированння и ведет к снижению точности числовых характеристик ряда. При выборе чрезмерно узких интервалов точность обобщающих числовых характеристик повышается, но ряд получается слишком растянутым и не дает четкой картины варьирования.

Для получения хорошо обозримого вариационного ряда и обеспечения достаточной точности вычисляемых по нему числовых характеристик следует разбить вариацию признака (в пределах от минимальной до максимальной варианты) на такое число групп или классов, которое удовлетворяло бы обоим требованиям. Эту задачу решают делением размаха варьирования признака на число групп или классов, намечаемых при построении вариационного ряда.

$$\lambda = \frac{x_{max} - x_{min}}{K}$$, [2]

где λ - величина классового интервала; Xmax, Xmin - максимальная и минимальная варианты совокупности; К - число классов, на которые следует разбить вариацию признака.

Число классов (К) можно приблизительно наметить, пользуясь табл. 5.

Таблица 5.

Число наблюдений n (от — до)	Число классов K
25—40	5—6
40—60	6—8
60—100	7—10
100—200	8—12
>200	10—15

Более точно величину К можно определить по формуле Стерджеса:

К= 1+3,32 lg n. [3]

При наличии в совокупности большого числа членов (n> 100) можно использовать формулу К=5 Ig n. [4]

Вопрос о том, распределять ли собранные данные в интервальный или безынтервальный ряд, решают в зависимости от характера и размаха варьирования признака. Если признак варьирует дискретно и слабо, т. е. в узких границах (величина λ оказывается равной единице или может быть приравнена к единице), данные распределяются в безынтервальный вариационный ряд. Если же признак варьирует в широких границах, то независимо от того, как он варьирует - дискретно или непрерывно, по данным строят интервальный вариационный ряд.

Техника построения вариационных рядов.

1. Определить минимальную X_{min} и максимальную X_{max} варианты.
2. Определить величину классового интервала λ. Если окажется, что λ= 1, собранный материал распределяется в безынтервальный вариационный ряд; если же λ≠1, исходные данные необходимо распределять в интервальный ряд. При этом точность величины классового интервала должна соответствовать точности, принятой при измерении признака. Например, жирномолочность коров (n=60), содержащихся на ферме, варьирует от 3,21 до 4,55%. В таком случае классовый интервал устанавливается следующим образом:

$$\lambda = \frac{4,55 - 3,21}{1 + 3,32 \lg 60} = \frac{1,34}{1 + 5,90} = 0,194 \approx 0,19.$$

Если точность измерения данного признака ограничить десятыми долями единицы, величина классового интервала окажется следующей:

$$\lambda = \frac{4,6 - 3,2}{6,9} = 0,2.$$

В обоих случаях результаты наблюдений должны распределяться в интервальный вариационный ряд.

3. Определить границы первого классового интервала. На данном этапе необходимо соблюсти следующее правило, минимальная варианта совокупности должна попадать примерно в середину первого классового интервала. Выполнение этого требования гарантирует построение вариационного ряда, наиболее полно отвечающего природе изучаемого явления, а, следовательно, и наименьшие потери информации о точности вычисляемых статистических характеристик ряда. Этому требованию удовлетворяет формула 3.

$$x_н = x_{min} - \lambda/2 \quad [5]$$

где x_H - нижняя граница первого классового интервала; x_{min} – минимальная варианта исследуемой совокупности; λ – величина классового интервала.

Так, при x_{min} =3,21 и λ=0,19 нижняя граница первого класса x_H =3,21-0,19/2=3,115≈3,12. Прибавив к этой величине λ=0,19, определяем верхнюю границу первого класса: 3,12+0,19=3,31. Затем находим верхнюю границу второго класса: 3,31+0,19=3,50 и т. д., пока не получим интервал, в который попадает максимальная варианта совокупности (x_{max} =4,55).

4. Далее необходимо распределить по намеченным классовым интервалам все варианты совокупности, т. е. определить частоты каждого класса. Здесь возникает вопрос: в какие классы относить варианты, которые по своей величине совпадают с верхней границей одного и нижней границей другого, соседнего класса? Например, в какой класс следует отнести варианту 3,31- в первый или во второй? Этот вопрос решается по-разному. Можно помещать в один и тот же класс варианты, которые больше нижней, но меньше или равны его верхней границе, т. е. по принципу «от и до включительно». Чаще, однако, поступают таким образом: верхние границы классов уменьшают на величину, равную точности, принятой при измерении признака, чем и достигается необходимое разграничение классов.

5. Следующий шаг ведет к замене классовых интервалов на их центральные или срединные значения. В результате интервальный вариационный ряд превращается в безынтервальный ряд. Необходимость такой замены вызывается тем, что обобщающие числовые характеристики (средняя, дисперсия и др.) вычисляются по безынтервальным рядам. Срединные значения классовых интервалов x_i, как это следует из формулы (5), отстоят от их нижних границ x_H на величину, равную половине классового интервала.

Наиболее точно центральную величину классового интервала можно получить по формуле 6.

$$x_i = \frac{1}{2}(x_н + x_к) \quad [6]$$

где x_K - конечная точка интервала, равная x_{H+1} - ε, т. е. началу следующего класса, уменьшенному на точность измерения признака.

Середины классов приобретают значения отдельных вариант и называются классовыми вариантами в отличие от конретныlx вариант, составляющих данную совокупность.

Задание 1.

На свиноферме зарегистрировано 64 опороса. Количество поросят, полученных от каждой свиноматки, варьировало следующим образом (рис. 4).

8	10	6	10	8	5	11	7	10	6	9	7	8	7	9	11	8	9	10
8	7	8	11	8	7	10	8	8	5	11	8	10	12	7	5	7	9	7
10	5	8	9	7	12	8	9	6	7	8	7	11	8	6	7	9	10	

Рис. 4.

По виду варьирования признака выбрать вид вариационного ряда: интервальный и безынтервальный, определить λ и построить соответствующий вариационный ряд.

Задание 2.

На основании многолетних клинических наблюдений, проводившихся в Сухумском питомнике обезьян, составлена следующая выборка, включающая 100 анализов на содержание кальция (мг%) в сыворотке крови низших обезьян (павианов-гамадрилов) – рис. 5.

3,6	12,9	12,3	9,9	12,7	11,7	10,8	10,4	10,9	10,2
4,7	10,4	11,6	11,7	12,1	10,9	12,1	9,2	10,7	11,5
3,1	10,9	12,0	11,1	13,5	11,2	13,5	10,1	14,0	10,0
11,6	12,4	11,9	11,4	12,8	11,4	10,9	12,7	13,8	13,2
11,9	10,8	11,0	12,6	10,0	10,3	12,7	11,7	12,1	13,8
12,2	11,9	11,6	10,6	11,1	10,7	12,3	11,5	11,2	11,5
12,7	10,5	11,2	11,9	9,7	13,0	9,6	12,5	11,6	9,0
11,5	12,3	12,8	12,6	12,8	12,5	12,8	11,4	12,5	12,3
14,5	12,3	12,6	11,7	12,2	12,3	11,6	12,0	13,5	12,5
11,6	11,9	12,0	11,4	14,7	11,3	13,2	14,3	13,2	14,2

Рис. 5.

По виду варьирования признака выбрать вид вариационного ряда: интервальный и безынтервальный, определить λ и построить соответствующий вариационный ряд.

Задание 3.

В результате учета яйценоскости 80 кур, содержащихся на птицеферме, было установлено, что признак варьирует от 208 до 250 яиц, полученных от несушки за 1 год.

По виду варьирования признака выбрать вид вариационного ряда: интервальный и безынтервальный, определить λ и построить соответствующий вариационный ряд.