4.2 Тема. Вычисление коэффициента корреляции для больших выборок

Цель. Знакомство с методами вычисления коэффициента корреляции в больших выборках

Методы вычисления и определения коэффициента корреляции на малых и больших выборках для количественных признаков заметно отличаются друг от друга. Формой организации выборочного материала, когда n>30, является корреляционная решетка, в которой разноска вариантов ведется с учетом обоих признаков (x,y). По степени тесноты и рассеивания вариантов по классам корреляционной решетки можно заранее судить о том, будет ли величина коэффициента корреляции большой, средней или малой.

Если варианты образуют узкий эллипс, то связь будет большой. Если варианты расположены по большинству клеток, то связь отсутствует или незначительная.

Для составления корреляционной решетки по каждому из коррелирующих признаков необходимо: 1) наметить величину классовых промежутков и границы классов, установить количество классов; 2) нанести классы одного признака по крайней левой грани корреляционной решетки сверху вниз, а классы второго признака – по верхней строчке, слева направо. Классы разделяются линиями. Горизонтальные и вертикальные линии пересекут друг друга и образуют клетки или ячейки корреляционной решетки.

Для точного учета коэффициента корреляции между двумя количественными признаками при большой выборке используется следующая формула:

$$r = \frac{\sum pa_x a_y - (b_x b_y n)}{\sigma_x \sigma_y n} \quad (38)$$

При вычислении коэффициента корреляции среднее квадратическое отклонение берется в величине классовых промежутков, вычисленное по формуле:

$$\sigma = \pm\sqrt{\frac{\sum pa^2}{n} - b^2} \quad (39)$$

Чтобы вычислить коэффициент корреляции, необходимо к корреляционной решетке добавить четыре графы по горизонтали и четыре

по вертикали. Одна вертикальная и одна горизонтальная графы делаются вдоль классов вариант; они служат для записи отклонений классов от условной средней (а). Остальные три графы по горизонтали и три по вертикали строятся внизу решетки (для ряда у) и с правой ее стороны (для ряда х). В них записываются: частоты (р), произведение частот на отклонения (ра) и произведение частот на квадраты отклонений (pa^2). В заголовке вертикальных граф пишут последовательно: p_x, p_xa_x, $p_xa_x^2$, а в заголовке горизонтальных граф (слева) p_y, p_ya_y, $p_ya_y^2$.

Далее производится вычисление и заполнение добавленных граф.

Пример. Необходимо вычислить коэффициент корреляции между живой массой (х) и обхватом груди (у) у коров красной горбатовской породы по данным таблицы 4.2.1.

Таблица 4.2.1 Живая масса и обхват груди у коров красной горбатовской породы

№	х, кг	у, см	№	х, кг	у, см	№	х, кг	у, см	№	х, кг	у, см
1	489	184	26	440	185	51	445	180	76	450	180
2	467	186	27	524	194	52	420	179	77	450	181
3	462	185	28	447	179	53	491	191	78	387	171
4	441	182	29	430	179	54	450	181	79	374	171
5	473	186	30	485	187	55	406	172	80	360	167
6	491	190	31	440	173	56	417	174	81	545	191
7	545	196	32	488	187	57	434	177	82	454	184
8	433	183	33	439	180	58	432	177	83	467	186
9	488	191	34	445	180	59	505	193	84	454	184
10	539	196	35	504	191	60	534	195	85	441	178
11	440	182	36	550	203	61	473	188	86	434	180
12	475	186	37	495	190	62	441	179	87	519	192
13	411	178	38	536	192	63	556	197	88	488	187
14	488	189	39	426	186	64	486	185	89	441	178
15	426	177	40	388	169	65	535	196	90	456	180
16	390	172	41	407	176	66	460	183	91	400	173
17	482	192	42	425	184	67	469	180	92	420	178
18	391	174	43	390	172	68	421	180	93	403	172
19	470	185	44	418	179	69	520	195	94	390	170
20	421	180	45	465	189	70	445	183	95	442	179
21	429	176	46	391	174	71	384	173	96	445	177
22	439	180	47	365	167	72	488	182	97	429	176
23	442	183	48	383	172	73	500	190	98	425	176
24	490	189	49	427	183	74	432	178	99	457	182
25	426	177	50	448	186	75	475	185	100	493	185

Лимиты обхвата груди равны 167-203 см, а лимиты живой массы – 365-559 кг. Величина классового промежутка для первого признака будет равняться: $K_y = \frac{203-167}{12} \approx 4{,}0$.

Величина классового промежутка для второго признака будет:

$$Kx = \frac{559-365}{12} \approx 20.$$

Исходя из полученных цифр, начало границ первого класса по обхвату груди будет 166, а всего классов – 10. Начало границ первого класса по живой массе 360, а всего классов 10. Необходимо начертить корреляционную решетку.

Таблица 4.2.2 Форма корреляционной решетки

x-живая масса	y-обхват груди									
	166-169	170-173	174-177	178-181	182-185	186-189	190-193	194-197	198-201	202-205
360-379										
380-399										
400-419										
420-439										
440-459										
460-479										
480-499										
500-519										
520-539										
540-559										

После того как начерчена корреляционная решетка производится заполнение ячеек этой решетки частотами. Делается это знаками так же, как и при разноске вариант по классам вариационного ряда, только в данном случае по каждой особи обращается внимание не на один показатель, а на два. После разноски всех вариант корреляционная решетка имеет такой вид (таблица 4.2.3).

Таблица 4.2.3 Заполнение корреляционной решетки

x/y	166-169	170-173	174-177	178-181	182-185	186-189	190-193	194-197	198-201	202-205
360-379	· ·	·								

380-399	·		· ·							
400-419		· · ·	· ·	· ·						
420-439					· ·					
440-459			·	· ·						
460-479				·	· · · ·					
480-499					· · ·					
500-519							· · · ·			
520-539							·			
540-559							·	· · ·		·

Полученное расположение вариант по ячейкам корреляционной решетки указывает, что между живой массой и обхватом груди существует прямая связь, так как варианты расположились слева, вниз, направо, а это показывает, что с увеличением обхвата груди увеличивается живая масса. Сосредоточение вариант вдоль одной линии (в узком овале) указывает, что между этими признаками существует большая взаимозависимость. Если бы варианты располагались по линии, идущей из левого нижнего в правый верхний угол, то это указывало бы на отрицательную (обратную) связь.

Однако часто варианты располагаются по ячейкам корреляционной решетки разбросанно, и тогда на взгляд трудно определить характер и степень связи, поэтому лучше выражать эту связь конкретной числовой величиной, для чего и вычисляется коэффициент корреляции.

К начерченной корреляционной решетке добавляем четыре графы по горизонтали и четыре по вертикали. Далее производится вычисление и заполнение добавленных граф; ход вычисления будет понятен при рассмотрении таблицы 4.2.4.

Таблица 4.2.4 Вычисление коэффициента корреляции

X	a	У										p_x	$p_x a_x$	$p_x a_x^2$
		166-169	170-173	174-177	178-181	182-185	186-189	190-193	194-197	198-201	202-205			
		-3	-2	-1	**0**	1	2	3	4	5	6			

360-379	-4	2	1	**I**					II			3	-12	48
380-399	-3	1	6	2								9	-27	81
400-419	-2		3	2	**2**							7	-14	28
420-439	-1			7	**10**	3						20	-20	20
440-459	**0**			**1**	**12**	**8**	**1**					**22**	**0**	**0**
460-479	1				**1**	4	6					11	11	11
480-499	2					3	5	5	1			14	28	56
500-519	3			III				4	IV			4	12	36
520-539	4							1	5			6	24	96
540-559	5							1	2		1	4	20	100
p_y		3	10	12	**25**	18	12	11	8	0	1	100	22	476
$p_y a_y$		-9	-20	-12	**0**	18	24	33	32	0	6	72		
$P_y a^2_y$		27	40	12	**0**	18	48	99	128	0	36	408		

Для нахождения $\Sigma p a_x a_y$ необходимо корреляционную решетку разделить жирными линиями, идущими вдоль классов с нулевыми отклонениями, на четыре квадранта, затем произвести перемножение отклонения a_x на a_y по каждому классу, имеющему частоты, и произведения записать в ячейке, находящейся на пересечении этих классов. Произведения $a_x a_y$ следует перемножить на частоты соответствующей ячейки, в результате чего и будет найдена величина $p a_x a_y$. Подсчет $\Sigma p a_x a_y$ производится отдельно по каждому квадранту.

$\Sigma p a_x a_y$= (I квадрант = 24+9+8+36+12+6+4+7=106; II квадрант = 0; III квадрант = -3; IV квадрант = 4+6+12+20+30+36+12+15+8+80+40+30=293) = 396.

Вычисление величин b, b^2, σ для обоих признаков производится обычно (как в вариационном ряде).

$$b_x = \frac{72}{100} = 0{,}72; \qquad b_y = \frac{22}{100} = 0{,}22.$$

$b^2_x = 0{,}72^2 = 0{,}51$; $b^2_y = 0{,}22^2 = 0{,}05$.

$$\sigma_x = \pm\sqrt{\frac{408}{100} - 0{,}51} = \pm 1{,}89;$$

$$\sigma_y = \pm\sqrt{\frac{476}{100} - 0{,}005} = \pm 2{,}17.$$

$$r = \frac{396 - (0{,}72 \cdot 0{,}22 \cdot 100)}{1{,}89 \cdot 2{,}17 \cdot 100} = \frac{380{,}2}{410{,}6} = +0{,}93.$$

Полученный коэффициент корреляции +0,93 близок к 1, что указывает на очень большую положительную связь между живой массой и обхватом груди у коров красной горбатовской породы.

Коэффициент корреляции выборочного исследования, как и все выборки, имеет свои ошибки. Ошибка коэффициента корреляции для малочисленной выборки (n>100) вычисляется по формуле:

$$m_r = \pm\frac{1-r^2}{\sqrt{n}} \qquad (40)$$

Ошибка коэффициента корреляции для малочисленной выборки:

$$m_r = \pm\frac{1-r^2}{\sqrt{n}-2} \qquad (41)$$

Критерий достоверности корреляции (t_r) вычисляется по формуле:

$$t_r = \frac{r}{m_r} \qquad (42)$$

Достоверность корреляции определяется по таблице Стьюдента (табл.2.9.1) с учетом числа степеней свободы (ν). Для t_r число степеней свободы равно:

$$\nu = n - 2.$$

Коэффициент корреляции достовернее, если t_r вычисленное равняется или больше табличного ($t_r > t_{st}$).

Для нашего примера:

$$m_r = \pm\frac{1-0{,}93^2}{100} = \pm 0{,}0135,$$

$$t_r = \frac{0{,}93}{0{,}0135} = 68{,}9.$$

Здесь t_r превышает табличное значение t при всех уровнях вероятности: $P_{0,95}$ = 1,6; $P_{0,99}$ = 2,0; $P_{0,999}$ = 3,4. Такой критерий называется высоким и он свидетельствует о большой достоверности корреляции.

Задание 1. Вычислить коэффициент корреляции между живой массой матерей (х) и живой массой телят (у) при рождении

х	У	х	у	Х	у
438	38	513	47	468	29
502	41	439	45	492	45
456	37	487	39	398	42
380	20	395	28	415	29
479	45	493	50	438	33
500	48	480	49	450	42
405	26	475	44	395	30
463	48	390	23	423	23

412	28	453	35	485	48
483	45	487	38	426	32
446	44	413	31	487	46

Задание 2. Вычислить коэффициент корреляции между высотой в холке (х) и обхватом груди (у) у кобыл русской рысистой породы

х	У	х	у	Х	у
161	176	149	178	156	170
160	175	155	180	156	176
150	167	150	169	154	172
256	170	156	175	154	176
164	187	152	167	152	172
157	180	155	178	147	160
157	172	157	180	155	179
156	178	149	164	152	174
159	178	154	173	155	171
155	164	155	181	158	175
166	182	150	168	154	172
152	178	158	181	160	183
155	172	155	170	160	185
155	179	156	170	154	180
154	175	160	180	152	164
152	163	164	184	151	164
159	175	155	182	155	180
152	165	148	166	155	171
157	180	160	175	157	171
160	186	155	170	154	169
154	175	159	181	152	169
152	187	150	171	163	190
158	182	152	173	153	173
149	160	149	165	163	186
154	180	155	172	155	168

Контрольные вопросы.

1. Как вычисляют достоверность коэффициента корреляции?
2. В каких случаях используется показатель корреляционного отношения и чем он характеризуется?
3. На что указывает знак при коэффициенте корреляции?
4. Что следует понимать под отрицательной корреляцией? Приведите примеры отрицательной корреляции.