Корреляционный и регрессионный анализ.

Существуют две категории связей между признаками:

1) **Функциональные** - каждому значению одной переменной величины соответствует одно вполне определенное значение другой переменной (высота столба ртути соответствует определённой температуре);

2) **Корреляционные** - (статистические) - численному значению одной переменной соответствует много значений другой переменной (одному росту соответствует множество значений веса).

Если есть результаты наблюдения, то первый шаг в анализе процесса состоит в построении различного рода графиков, с помощью которых можно было бы исследовать его основные характеристики. Наиболее простую иллюстрацию парных наблюдений даёт график (диаграмма) рассеяния.

Графики дают первую наглядную информацию о наличии связи между переменными величинами. Поэтому возникает потребность в количественном измерении корреляции. Одним из способов является вычисление коэффициента корреляции.

Коэффициент корреляции-это число, показывающее степень зависимости одной переменной величины от другой.

Свойства коэффициента корреляции:

1. r - число; лежащее в интервале от **–1** до **+1**

   (-1≤ r ≤ 1).

2. если **r=±1,** то точки лежат на одной прямой, следовательно, зависимость между x и y – функциональная

3. если **(0< r <0.5)-** то зависимость между переменными слабая.

4. если **(0.5≤r< 0.7)-** то зависимость между переменными средняя

5. если **r≥ 0.7** существует **сильная** линейная зависимость между переменными.

Рассчитывают коэффициент корреляции по формуле:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Коэффициент корреляции указывает лишь на степень связи в вариации двух переменных величин, т.е. дает меру тесноты этой связи, но не дает возможность судить о том, как количественно меняется одна величина по мере изменения другой. На этот вопрос позволяет ответить другой метод определения связи между вариационными признаками - метод регрессии.

Зависимость между биологическими признаками может быть самой разнообразной. В большем числе случаев эмпирические регрессии выражаются простыми уравнениями линейной регрессии:

$y = ax + b$

Формулы для вычисления коэффициентов **a** и **b:**

$$a = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad b = \frac{\sum y_i \sum x_i^2 - \sum x_i \cdot \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Задача:

В анализах крови определяли: Х-содержание гемоглобина(%), У-оседание эритроцитов крови за 2 часа(мм). Построить график рассеяния. Найти уравнение регрессии. Найти коэффициент корреляции.

X	**77**	80	82	79	84	75	82	79	87	87	87	90	97	96	92
Y	32	33	33	34	34	34	34	35	36	37	37	38	40	40	40

x_i	y_i	$x_i-x_{ср}$	$y_i-y_{ср}$	$(x_i-x_{ср})*(y_i-y_{ср})$	$(x_i-x_{ср})^2$	$(y_i-y_{ср})^2$	x_i^2	x_i*y_i
77	32	-7,9	-3,8	30,02	62,41	14,44	5929	2464
80	33	-4,9	-2,8	13,72	24,01	7,84	6400	2640
82	33	-2,9	-2,8	8,12	8,41	7,84	6724	2706
79	34	-5,9	-1,8	10,62	34,81	3,24	6241	2686
84	34	-0,9	-1,8	1,62	0,81	3,24	7056	2856
75	34	-9,9	-1,8	17,82	98,01	3,24	5625	2550
82	34	-2,9	-1,8	5,22	8,41	3,24	6724	2788
79	35	-5,9	-0,8	4,72	34,81	0,64	6241	2765
87	36	2,1	0,2	0,42	4,41	0,04	7569	3132
87	37	2,1	1,2	2,52	4,41	1,44	7569	3219
87	37	2,1	1,2	2,52	4,41	1,44	7569	3219
90	38	5,1	2,2	11,22	26,01	4,84	8100	3420
97	40	12,1	4,2	50,82	146,41	17,64	9409	3880
96	40	11,1	4,2	46,62	123,21	17,64	9216	3840
92	40	7,1	4,2	29,82	50,41	17,64	8464	3680
1274	**537**			**235,8**	**630,95**	**104,4**	**108836**	**45845**
84,9	**36**							

Ход решения задачи:

1. Находят средние значения первой и второй переменной($\overline{X}_i$ $\overline{Y}_i$ **).**
2. Находят разность между каждым значением случайной величины и средним значением для переменной Х и**Y** **$(X_i-X_{ср})$ и $(Y_i-Y_{ср})$.**
3. Находят произведение полученных разностей **$(X_i-X_{ср}) * (Y_i-Y_{ср})$** для каждого значения случайной величины Х и**Y**.
4. Возводят в квадрат полученные разности **$(X_i-X_{ср})^2$ и $(Y_i-Y_{ср})^2$**

5. Суммируют значения полученных квадратов разностей и получают суммы: $\sum(X_i-X_{ср})^2$, $\sum(Y_i-Y_{ср})^2$ и $\sum(X_i-X_{ср}) * (Y_i-Y_{ср})$
6. Подставляют полученные суммы в формулу коэффициента корреляции и рассчитывают его значение.

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} = \frac{235}{\sqrt{631\cdot 104}} = 0.92$$

7. Делают вывод: R=0,92 – зависимость сильная, прямопропорциональная.
8. Для построения линии регрессии рассчитывают коэффициенты a и b. Для этого находят суммы: $\sum Xi^2$ и $\sum Xi*Yi$.

$$a = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \qquad b = \frac{\sum y_i \sum x_i^2 - \sum x_i \cdot \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$a = \frac{15\cdot 45845 - 1274\cdot 537}{15\cdot 108836 - 1274^2} = 0.37 \qquad b = \frac{537\cdot 108836 - 1274\cdot 45845}{15\cdot 108836 - 1274^2} = 4.06$$

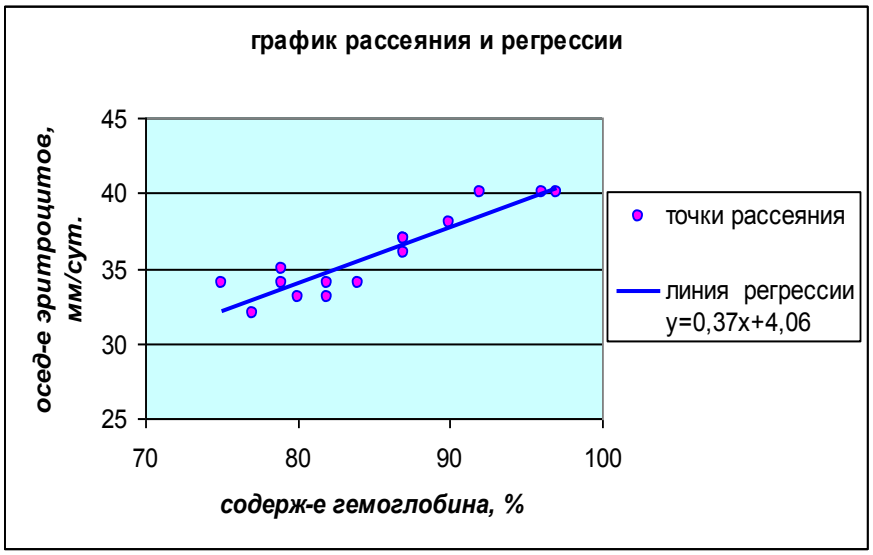
9. Строят уравнение регрессии: **y=ax+b**

Y=0,37x+4,06

X_1=77 Y_1=32.55

X_2=90 Y_2=37.36

10. Строят график:



Ранговая корреляция.

Из непараметрических показателей связи наиболее широкое применение нашел коэффициент корреляции рангов.

Для вычисления обычного коэффициента корреляции необходимо, чтобы исходные данные были выражены достаточно точно. Однако это далеко не всегда возможно. Существуют такие количественные признаки, которые с трудом поддаются точной оценке. Кроме того, распределение одного или обоих коррелирующих признаков может быть очень неравномерным или

неправильным. Эти трудности можно обойти, если применить оценку вариант по каждому признаку порядковыми номерами от меньших значений к большим (или наоборот). Порядковый номер по каждому признаку является его рангом. Отсюда название этого метода - определение коэффициента ранговой корреляции. Формула для его вычисления**:**

$$r = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2 - 1)}$$

X_i и Y_I -ранги по первому и второму признаку;

n- число пар коррелирующих величин.

Задача:

Имеются данные о суточной потребности в белках у восьмилетних девочек. Определить коэффициент корреляции рангов между весом девочек (X) и суточной потребностью у них в белках (Y).

X(кг)	**20**	**22**	**23**	**25**	**26**	**27**	**28**
Y(г)	62	66	62	75	75	78	82

Решение: Используют формулу коэффициента ранговой корреляции **:**

$$r = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2 - 1)}$$

Для решения задачи составляют таблицу:

Ранг X_i	**Вес** (в порядке возрастания)	**Вес**	**потр в белках**	**потр в белках** (в порядке возрастания)	**Ранг Y_i**	**X_i-Y_i**	**$(X_i-Y_i)^2$**
1	20	***20***	***62***	62	1,5	-0,5	0,25
2	22	***22***	***66***	62	1,5	-1	1
3	23	***23***	***62***	66	3	1,5	2,25
4	25	***25***	***75***	75	4,5	-0,5	0,25
5	26	***26***	***75***	75	4,5	0,5	0,25
6	27	***27***	***78***	78	6	0	0
7	28	***28***	***82***	82	7	0	0
							$\sum$ =4

Ход решения задачи:

1. Выстраивают данные задачи в порядке возрастания.
2. Ранжируют полученные ряды (нумеруют)*.
3. Находят разность рангов для каждой пары чисел.
4. Возводят в квадрат полученную разность рангов.
5. Находят сумму квадратов разности рангов.
6. Находят коэффициент корреляции рангов по формуле:

$$r = 1 - \frac{6\sum (x_i - y_i)^2}{n(n^2 - 1)}$$

R=1-6·4/ 7(7^2–1)=1-0,07=0,93

Вывод: **R= 0,93** – связь между весом девочек и суточной потребностью белка сильная, прямопропорциональная.

Примечание:

* Правила ранжирования

1. Меньшему значению начисляется меньший ранг.
   Наименьшему значению начисляется ранг 1.
   Наибольшему значению начисляется ранг, соответствующий количеству ранжируемых значений. Например, если n=7, то наибольшее значение получит ранг 7, за возможным исключением для тех случаев, которые предусмотрены правилом 2.
2. В случае, если несколько значений равны, им начисляется ранг, представляющий собой среднее значение из тех рангов, которые они получили бы, если бы не были равны.
3. Общая сумма рангов должна совпадать с расчётной, определяемой по формуле:

$\sum (R_i) = \frac{N*(N+1)}{2}$ где N-общее количество наблюдений.