Тема: Элементы теории корреляции

Объекты ряда генеральных совокупностей обладают несколькими подлежащими изучению признаками X, У, ..., которые можно интерпретировать как систему взаимосвязанных величин. Примерами могут служить: масса животного и количество гемоглабина в крови, рост мужчины и объем грудной клетки, увеличение рабочих мест в помещении и уровень заболеваемости вирусными инфекциями, количество вводимого препарата и концентрация его в крови и т.д.

Очевидно, что между этими величинами существует связь, но она не может быть строгой фукциональной зависимостью, так как на изменение одной из величин влияет не только изменение второй величины, но и другие факторы. В таких случаях говорят, что две величины связаны *стохастической* (т.е. случайной) зависимостью. Мы будем изучать частный случай стохастической зависимости – *корреляционную зависимость*.

ОПРЕДЕЛЕНИЕ: Зависимость случайных величин называют *стохастической*, если на изменение одной из них влияет не только изменение второй величины, но и другие факторы.

ОПРЕДЕЛЕНИЕ: Зависимость случайных величин называют *статистической,* если изменения одной из них приводит к изменению закона распределения другой.

ОПРЕДЕЛЕНИЕ: Если изменение одной из случайных величин влечет изменение среднего другой случайной величины, то статистическую зависимость называют *корреляционной.*

Примерами *корреляционной зависимости* являются связи между:

- массой тела и ростом;
- дозой ионизирующего излучения и числом мутаций;
- пигментом волос человека и цветом глаз;
- показателями уровня жизни населения и процентом смертности;
- количеством пропущенных студентами лекций и оценкой на экзамене и т.д.

Именно корреляционные зависимости наиболее часто встречаются в природе в силу взаимовлияния и тесного переплетения огромного множества самых различных факторов, определяющих значения изучаемых показателей.

Результаты наблюдения, проведенные над тем или иным биологическим объктом по корреляционно связанным признакам У и X можно изобразить точками на плоскости, построив систему прямоугольных координат. В результате получается некая диаграмма рассеяния, позволяющая судить о форме и тесноте связи между варьирующими признаками.

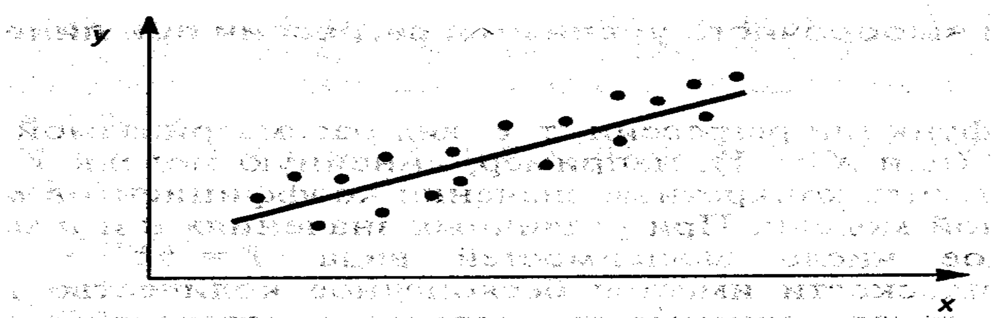
Если эту связь можно будет апроксимировать некоторой кривой, то можно будет прогнозировать изменение одного из параметров при целенаправленном изменении другого параметра.

Корреляционную зависимость Y от X можно описать с помощью уравнения вида

$$\overline{y_x} = f(x) \qquad (1)$$

где $\overline{y_x}$ - **условное среднее** величины Y, соответствующее значению x величины X, а $f(x)$ - некоторая функция. Уравнение (1) называется **выборочным уравнением регрессии** Y на X.

Рис.1. Линейная регрессия значима. Модель $Y = a + bX$.



Функцию $f(x)$ называют **выборочной регрессией** Y на X, а ее график – **выборочной линией регрессии** Y на X.

Совершенно аналогично **выборочным уравнением регрессии** X на Y является уравнение $\bar{x}_y = \varphi(y)$.

В зависимости от вида уравнения регрессии и формы соответствующей линии регрессии определяют форму корреляционной зависимости между рассматриваемыми величинами – **линейной, квадратической, показательной, экспоненциальной.**

Важнейшим является вопрос выбора вида функции регрессии $f(x)$ [или $\varphi(y)$], например линейная или нелинейная (показательная, логарифмическая и т.д.)

На практике вид функции регрессии можно определить построив на координатной плоскости множество точек, соответствующих всем имеющимся парам наблюдений (x, y).

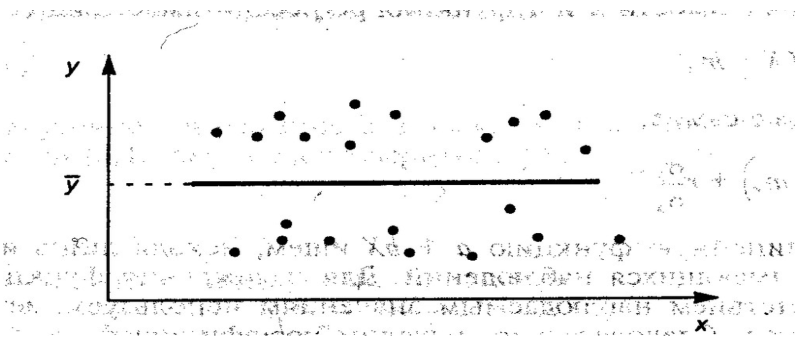


Рис. 2. Линейная регрессия незначима. Модель $Y = \bar{Y}$.

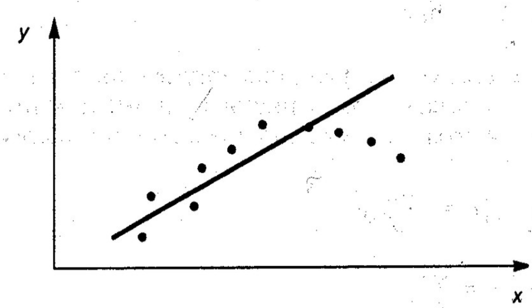


Рис. 3. Нелинейная модель $(y = ax^2 + bx + c)$.

Например, на рис.1. видна тенденция роста значений Y с ростом X, при этом средние значения Y располагается визуально на прямой. Имеет смысл использовать линейную модель (вид зависимости Y от X принято называть моделью) зависимости Y от X.

На рис.2. средние значения Y не зависят от x, следовательно линейная регрессия незначима (функция регрессии постоянна и равна $\bar{y}$).

На рис. 3. прослеживается тенденция нелинейности модели.

Примеры прямолинейной зависимости:

- увеличение количество потребляемого йода и снижение показателя заболеваемости зобом,
- увеличение стажа рабочего и повышение производительности.

Примеры криволинейной зависимости:

- с увеличением осадков – увеличивается урожай, но это происходит до определенного предела осадков. После критической точки осадки уже оказываются излишними, почва заболачивается и урожай снижается,
- связь между дозой хлора, примененной для обеззараживания воды и количеством бактерий в 1 мл. воды. С увеличением дозы хлора количество бактерий в воде снижается, но по достижению критической точки количество бактерий будет оставаться постоянным (или совсем отсутствовать), как бы мы не увеличивали дозу хлора.

Линейная регрессия

Выбрав вид функции регрессии, т.е. вид рассматриваемой модели зависимости Y от X (или X от У), например, линейную модель $\bar{y}_x = a + bx$, необходимо определить конкретные значения коэффициентов модели.

При различных значениях a и b можно построить бесконечное число зависимостей вида $\bar{y}_x = a + bx$ т.е на координатной плоскости имеется бесконечное количество прямых, нам же необходима такая зависимость, которая соответствует наблюдаемым значениям наилучшим образом. Таким образом, задача сводится к подбору наилучших коэффициентов.

Метод наименьших квадратов (МНК)

Линейную функцию $a + bx$ ищем, исходя лишь из некоторого количества имеющихся наблюдений. Для нахождения функции с наилучшим соответствием наблюдаемым значениям используем **метод наименьших квадратов.**

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

Рис.4. Пояснение к оценке коэффициентов методом наименьших квадратов

Обозначим: Y_i - значение, вычисленное по уравнению $Y_i = a + bx_i$

y_i - измеренное значение,

$\varepsilon_i = y_i - Y_i$ - разность между измеренными и вычисленными по уравнению значениям,

$\varepsilon_i = y_i - a - bx_i$.

В **методе наименьших квадратов** требуется, чтобы ε_i, разность между измеренными y_i и вычисленными по уравнению значениям Y_i, была минимальной. Следовательно, находимо подобрать коэффициенты a и b так, чтобы сумма квадратов отклонений наблюдаемых значений от значений на прямой линии регрессии оказалась наименьшей:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - Y_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min.$$

Это условие достигается если параметры a и b будут вычислены по формулам :

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

$$a = \bar{y} - b\bar{x} \qquad (3)$$

b называют **коэффициентом регрессии**; a называют **свободным членом** уравнения регрессии.

Полученная прямая является оценкой для теоретической линии регрессии. Имеем

$$Y = a + bx = \bar{y} - b\bar{x} + bx = \bar{y} + b(x - \bar{x}).$$

Итак, $Y = \bar{y} + b(x - \bar{x})$ является *уравнением линейной регрессии.*

Регрессия может быть прямой $(b > 0)$ *и обратной* $(b < 0)$.

ОПРЕДЕЛЕНИЕ: ***Обратная регрессия*** означает, что при росте одного параметра, значения другого параметра уменьшаются.

ОПРЕДЕЛЕНИЕ: ***Прямая регрессия*** означает, что при росте одного параметра, значения другого параметра тоже увеличиваются.

Пример 1. Заданному уровню потребления пресной воды на санитарно – бытовые нужды X в л/чел. сутки в n населенных пунктах соответствует множество значений уровня общей заболеваемости Y в %. При этом отмечается, что с ростом X наблюдается уменьшение Y. Это – **обратная**, **отрицательная** корреляционная связь. (Рис. 5)

Пример 2. Возрастание уровня инфекционной заболеваемости Y в % при увеличении плотности рабочих мест в производственном помещении X, чел. – является примером **прямой**, **положительной** корреляционной связи. (Рис. 6)

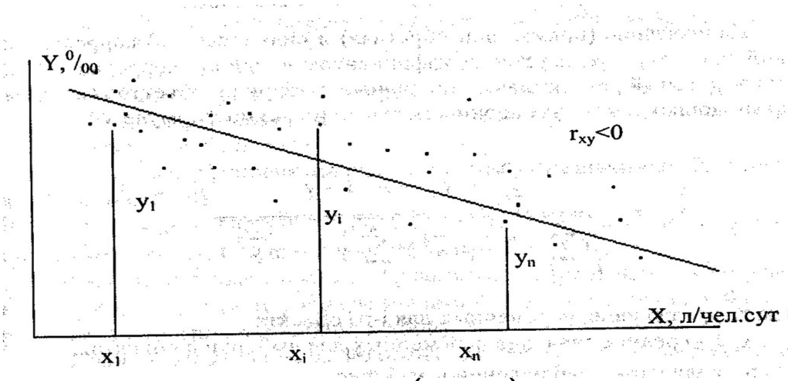


Рис. 5. Поле наблюдений $(i = \overline{1, n})$ *при обратной корреляционной связи между фактором* X *и параметром* Y

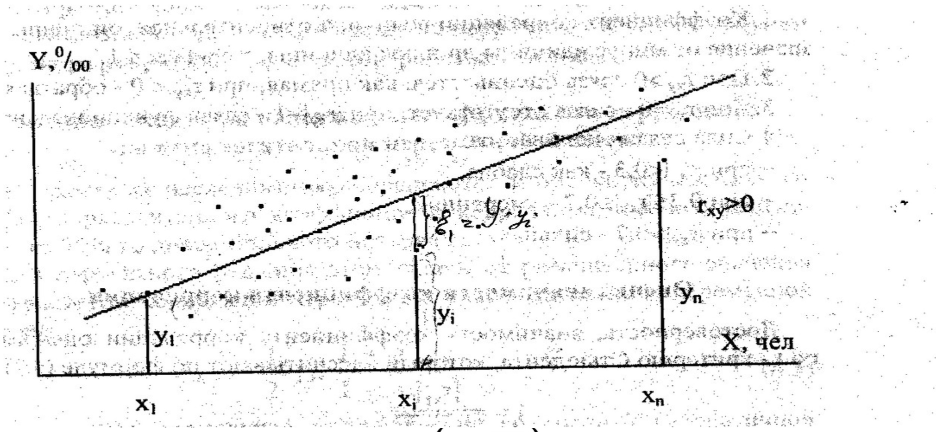


Рис. 6. Поле наблюдений $(i = \overline{1, n})$ *при прямой корреляционной связи между фактором* X *и параметром* Y

Проверка гипотезы о значимости коэффициента регрессии.

Не всегда можно утверждать, что предполагаемая линейная зависимость действительно имет место.

Построив модель, описывающую изменения величин, необходимо определить верна ли она. В регрессионном анализе проверяют гипотезы о значимости свободного члена a и о значимости коэффициента регрессии .

1. Определяем гипотезы H_0 и H_1:
   H_0: =0 (между величинами нет линейной зависимости),
   H_1: $\neq 0$.
2. Зададим уровень значимости α.
3. Статистика критерия.

$$F = \frac{b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{s^2} \qquad s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 2}$$

, где

4. Критические точки и критическая область. Статистика F имеет распределение Фишера с 1 и (n-2) степенями свободы. $F_{\alpha,1,n-2}$

5. Если $|F_{набл}| < F_{\alpha,1,n-2}$, то H_0 отвергается, т.е. можно сделать вывод, что линейная зависимость значима.

Если $|F_{набл}| > F_{\alpha,1,n-2}$, то у нас нет оснований отвергать H_0, т.е. можно сделать вывод, что линейная зависимость – незначима или что данные нельзя описать моделью линейной регрессии.

Корреляционный анализ.

Для достаточно полного описания особенностей корреляционной зависимости между величинами недостаточно определить форму этой зависимости и в случае линейной зависимости описать ее вид по величине коэффициента регрессии. Необходимо так же оценить тесноту связи.

Корреляционный анализ экспериментальных данных для двух случайных величин заключает в себе следующие основные приемы:

1. Вычисление выборочных коэффициентов корреляции.
2. Составление корреляционной таблицы.
3. Проверка статистической гипотезы значимости связи.

Линейная корреляция

ОПРЕДЕЛЕНИЕ: Корреляционная зависимость между случайными величинами X и Y называется ***линейной корреляцией***, если обе функции регрессии $f(x)$ и $\phi(y)$ являются линейными. В этом случае обе линии регрессии являются прямыми; они называются **прямыми регрессии.**

Выборочный коэффициент корреляции.

Например, корреляционная зависимость возраста Y учеников средней школы от года X их обучения в школе является, как правило, более тесной, чем аналогичная зависимость возраста студентов высшего учебного заведения от года обучения, поскольку среди студентов одного и того же года обучения в вузе обычно наблюдается больший разброс в возрасте, чем у школьников одного и того же класса.

Для оценки тесноты линейных корреляционных зависимостей между величинами X и Y по результатам выборочных наблюдений вводится понятие **выборочного коэффициента линейной корреляции,** определяемого формулой

$$r_B = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (4)$$

Следует отметить, что основной смысл выборочного коэффициента линейной корреляции r_B состоит в том, что он представляет собой эмпирическую (т.е. найденную по результатам наблюдений над величинами X и Y) оценку соответствующего генерального коэффициента линейной корреляции r:

$$r = r_B \qquad (6)$$

Принимая во внимание формулы

$$\frac{\sigma_X}{\sigma_Y} = \frac{\sqrt{\overline{X^2} - \overline{X}^2}}{\sqrt{\overline{Y^2} - \overline{Y}^2}}, \qquad r_в = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)}}$$

видим, что выборочные уравнения линейной регрессии Y на X имеют вид

$$Y - \overline{Y} = r_B \frac{\sigma_Y}{\sigma_X}(X - \overline{X}) \qquad (7)$$

где $r_B \frac{\sigma_Y}{\sigma_X} = b$.

То же можно сказать о выборочном уравнений линейной регрессии X на Y

$$X - \overline{X} = r_B \frac{\sigma_X}{\sigma_Y}(Y - \overline{Y}) \qquad (5)$$

Основные свойства выборочного коэффициента линейной корреляции:

1. Коэффициент корреляции двух величин, не связанных линейной корреляционной зависимостью, равен нулю.
2. Коэффициент корреляции двух величин, связанных линейной корреляционной зависимостью, положителен в случае прямой зависимости и отрицателен в случае обратной зависимости.
3. Абсолютная величина коэффициента корреляции двух величин, связанных линейной корреляционной зависимостью, удовлетворяет неравенству $0 < |r| < 1$.
4. Чем ближе $|r|$ к 1, тем теснее прямолинейная корреляция между величинами Y, X.

По своему характеру корреляционная связь может быть **прямой** и **обратной,** а по силе – **сильной, средней, слабой.** Кроме того, связь может **отсутствовать** или быть **полной**.

Сила и характер связи между параметрами

Сила связи	Характер связи	
	Прямая (+)	Обратная (-)
Полная	1	-1
Сильная	От 0,7 до 1	От -0,7 до -1
Средняя	От 0,699 до 0,3	От -0,699 до -0,3
Слабая	От 0,299 до 0	От -0,299 до 0
Связь отсутствует	0	0

Проверка гипотезы о значимости коэффициента корреляции.

О статистической взаимосвязи говорят, что она существует или отсутствует, имеет направление и характеризуется силой.

Если в результате исследования нулевая гипотеза не отвергается, то «взаимосвязи нет» . В случае, когда нулевая гипотеза отклоняется говорят о существовании связи исследуемых случайных величин.

1. Сформулируем гипотезы H_0 и H_1:

H_0: r =0 (корреляции нет),

H_1: r ≠0.

2. Зададим уровень значимости α.
3. Статистика критерия $t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n - 2}$

4. $t_{\alpha,n-2}$. t-статистика, имеющая распределение Стьюдента с (n-2) степенями свободы.

5. При $|t| \geq t_{\alpha,n-2}$, H_0 отвергается. Это значит, что между параметрами существует значимая корреляция. При $|t| < t_{\alpha,n-2}$, H_0 принимается.

Пример. Даны значения *x* и *y*.

x_i	-2	0	1	2	4
y_i	0,5	1	1,5	2	3

1) найти выборочное уравнение регрессии *y* от *x*;
2) построить график регрессии;
3) вычислить коэффициент корреляции;

4) определить силу и характер корреляционной связи.

Решение. n=5

1) а) Считая, что зависимость между X и Y линейная ($y = a + bx$) вычислим методом наименьших квадратов коэффициент регрессии и свободный член a.

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

$$\bar{x} = \frac{-2+0+1+2+4}{5} = \frac{5}{5} = 1, \quad \bar{y} = \frac{0.5+1+1.5+2+3}{5} = \frac{8}{5} = 1.6$$

$$\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y}) = (-2-1)(0.5-1.6) + (0-1)(1-1.6) + (1-1)(1.5-1.6) + (2-1)(2-1.6) +$$
$$+ (4-1)(3-1.6) = 3.3 + 0.6 + 0 + 0.4 + 4.2 = 8.5$$

$$\sum_{i=1}^{5}(x_i - \bar{x})^2 = (-2-1)^2 + (0-1)^2 + (1-1)^2 + (2-1)^2 + (4-1)^2 = 9+1+0+1+9 = 20$$

$$b = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2} = \frac{8.5}{20} = 0.425$$. Так как $b = 0.425 > 0$, то регрессия прямая.

$a = \bar{y} - b\bar{x} = 1.6 - 0.425 * 1 = 1.175$.

Тогда уравнение будет иметь вид $y = a + bx \Rightarrow y = 1{,}175 + 0{,}425x$.

в) $b = \dfrac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}$ $\qquad a = \dfrac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2}$

$$\bar{x} = \frac{-2+0+1+2+4}{5} = \frac{5}{5} = 1, \quad \bar{y} = \frac{0.5+1+1.5+2+3}{5} = \frac{8}{5} = 1.6$$

$$\overline{xy} = \frac{-2\cdot 0.5 + 0\cdot 1 + 1\cdot 1.5 + 2\cdot 2 + 4\cdot 3}{5} = \frac{16.5}{5} = 3.3$$

$$\overline{x^2} = \frac{(-2)^2 + 0^2 + 1^2 + 2^2 + 4^2}{5} = \frac{25}{5} = 5$$

$$b = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{3.3 - 1\cdot 1.6}{5 - 1^2} = \frac{3.3 - 1.6}{4} = 0.425, \quad a = \frac{\overline{x^2}\,\bar{y} - \bar{x}\,\overline{xy}}{\overline{x^2} - \bar{x}^2} = \frac{5\cdot 1.6 - 1\cdot 3.3}{5 - 1^2} = 1.175$$

$y = a + bx \Rightarrow y = 1{,}175 + 0{,}425x$.

Проверка гипотезы о значимости коэффициента регрессии.

1. Определяем гипотезы H_0 и H_1:
   H_0: =0 (между величинами нет линейной зависимости),
   H_1: ≠0.
2. Зададим уровень значимости α=0,05.
3. Статистика критерия.

$$F = \frac{b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{s^2}, \quad \text{где} \quad s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-2}$$

$$\sum_{i=1}^{5}(y_i - \bar{y})^2 = (0.5-1.6)^2 + (1-1.6)^2 + (1.5-1.6)^2 + (2-1.6)^2 + (4-1.6)^2 = 3.7$$

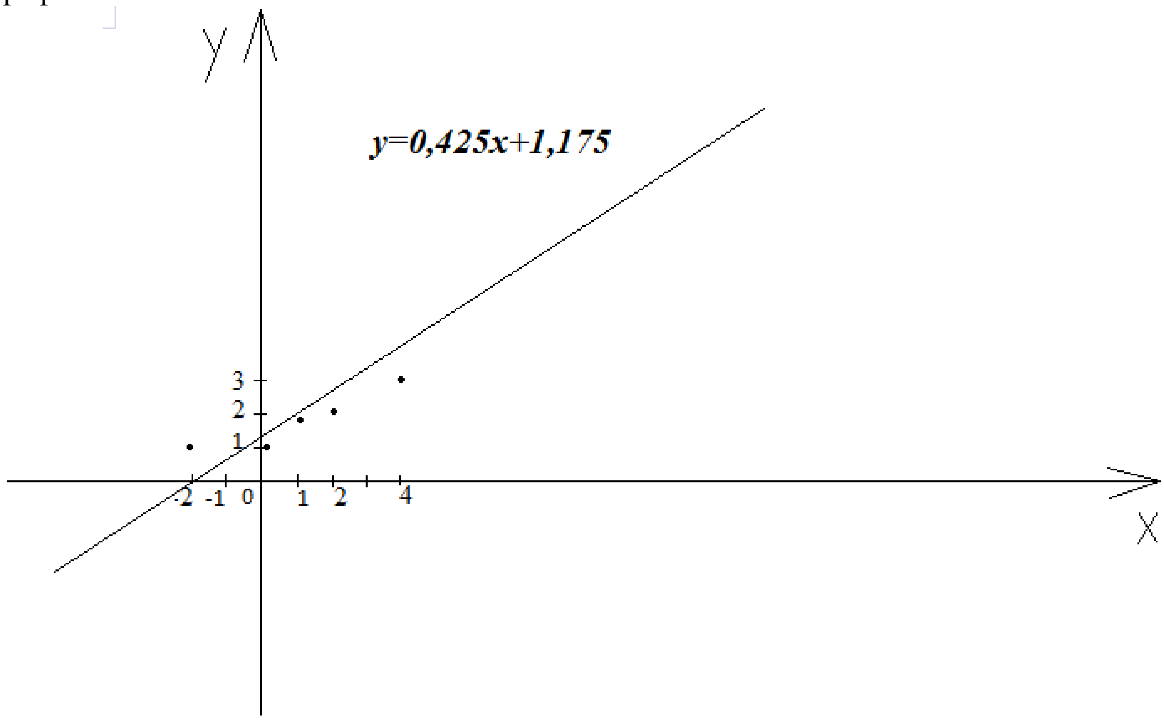
$$S^2 = \frac{\sum_{i=1}^{5}(y_i - \bar{y})^2}{n-2} = \frac{3.7}{5-2} = 1.23$$

$$F_{набл}=\frac{b^2\sum_{i=1}^{n}(x_i-\bar{x})^2}{S^2}=\frac{0.425^2\cdot 20}{1.23}=2.94$$

4. Критические точки и критическая область. Статистика F имеет распределение Фишера с 1 и (n-2) степенями свободы. $F_{\alpha,1,n-2}$. $F_{кр}=F_{кр}(\alpha,1.n-2)$. $F_{крит}=F_{крит}(0.05,1,3)=10.13$

5. $|F_{набл}|=2{,}94<F_{\alpha,1,n-2}=10{,}13$, то H_0 отвергается, т.е. можно сделать вывод, что линейная зависимость значима.

2) график



3) **Выборочный коэффициент корреляции**

$$r_B=\frac{\overline{xy}-\bar{x}\cdot\bar{y}}{\sigma_x\sigma_y}=\frac{3.3-1\cdot 1.6}{2*0{,}86}=\frac{1.7}{1.72}=0.99$$

$$\sigma_x=\sqrt{\sigma_x^2}=\sqrt{\overline{x^2}-(\bar{x})^2}=\sqrt{5-1^2}=2 \qquad \sigma_y=\sqrt{\sigma_y^2}=\sqrt{\overline{y^2}-(\bar{y})^2}=\sqrt{3.3-1.6^2}=\sqrt{0{,}74}=0{,}86$$

или $r_B=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\dfrac{8.5}{\sqrt{20\cdot 3.7}}=0.99$

4) Так как $r_B=0.99$, то корреляционная связь по своему характеру **прямая**, а по силе – **сильная**.

Проверка гипотезы о значимости коэффициента корреляции.

1. Сформулируем гипотезы H_0 и H_1:

H_0: r =0 (корреляции нет),

H_1: r ≠0.

2. Зададим уровень значимости α=0,05.
3. Статистика критерия $t_{набл}=r_T\sqrt{\dfrac{n-2}{1-r_T^2}}=0{,}99\sqrt{\dfrac{5-2}{1-0{,}99^2}}=0{,}99\sqrt{\dfrac{3}{0{,}0199}}=12{,}16$

4. $t_{крит}(\alpha,df)=t_{крит}(0.05,3)=3.18$-t-статистика, имеющая распределение Стьюдента с (n-2) степенями свободы.

6. $|t_{набл}|\ge t_{(n-2),\alpha}\Rightarrow|t_{набл}|\ge t_{(5-2),0.05}\Rightarrow 12.16>3.18$, то H_0 отвергается. Это значит, что между параметрами существует значимая корреляция.

Задания.

1. Даны показатели охвата населения прививками X (%) и заболеваемости брюшным тифом Y (в %).

Районы	A	B	C	D	E	F	G	H	I
X	14,7	13,4	9,6	8,1	5,5	5,2	4,4	4,4	4,0
Y	1,4	1,4	2,3	2,1	6,2	6,9	8,6	10,8	11,0

1) найти выборочное уравнение регрессии y от x;
2) построить график регрессии;
3) вычислить коэффициент корреляции;
4) определить силу и характер корреляционной связи.

2. У окуня озера Баторино измерены длина головы X и длина грудного плавника $У$:

X 66 61 67 73 51 59 48 47 58 44 41 54 52 47 51 45
$У$ 38 31 36 43 29 33 28 25 36 26 21 30 28 27 28 26

Проведите корреляционно-регрессионный анализ полученных данных.

3. Надо было установить, есть ли корреляция между высотой головы X и длиной 3-го членика усика $У$ у Drosophila funebris. Для этого с ромощью окуляр-микрометра получены следующие данные по X и $У$ (в делениях окуляр-микрометра):

X 15 16 15 16 17 18 17 17 15 16 15 17 13 14 17 16 15 16 16 15 18 17 14 15
$У$ 29 31 33 32 33 36 35 35 35 33 31 35 30 31 35 33 32 33 33 30 34 34 31 33

Проведите корреляционно-регрессионный анализ полученных данных.

Литература.

1. Ю.В. Морозов. Основы высшей математики и статистики, М., «Медицина», 2001г.
2. И.В. Павлушков и др. Основы высшей математики и математической статистики, М., Издательский дом ГЭОТАР-МЕД, 2003г.
3. В.Е. Гмурман Теория вероятностей и математическая статистика М., «Высшая школа», 2003г.
4. В.Е. Гмурман Руководство к решению задач по теории вероятностей и математической статистики, М., «Высшая школа», 2003г.