К блоку заданий IV. Исследование статистических связей.

Результаты эксперимента $(x_1, y_1), \ldots, (x_n, y_n)$ можно трактовать как независимые реализации случайного вектора (ξ, η). Необходимо проверить гипотезу независимости компонент этого вектора, оценить степень связности величин, а также построить наилучший прогноз одной величины по значениям другой величины.

Данные для обработки: столбец R (лист 2) – измерения ξ (X), столбец S (лист 2) – измерения η(Y)

IV.1. Проверить гипотезу независимости по критерию сопряжённости хи-квадрат

– начальная точка, шаг и количество групп (с учётом бесконечных) в строке 15 (лист 1),

– уровень значимости α: строка 15 (лист 1).

IV.2. Построить оценку наилучшего линейного прогноза одной сл.в. по значениям другой сл.в., привести график линии регрессии и эллипса рассеяния в поле данных

– направление прогнозирования (Y через X или X через Y): строка 17 (лист 1)

IV.3. Проверить гипотезу независимости в ситуации, когда можно считать распределение вектора (ξ, η) нормальным

– уровень значимости α: строка 16 (лист 1)

Теоретические основы

Определения. Две случайные величины ξ, η называются *независимыми*, если для любых измеримых подмножеств A, B из пространств значений этих сл. в. выполняется равенство

$$\mathbf{P}\{\xi \in A \cap \eta \in B\} = \mathbf{P}\{\xi \in A\}\,\mathbf{P}\{\eta \in B\}.$$

Коэффициентом корреляции между случайными величинами ξ, η называется величина

$$\rho = \frac{\sigma_{\xi\eta}}{\sigma_\xi\,\sigma_\eta},$$

где σ_ξ, σ_η – стандартные отклонения ξ, η соответственно,

$$\sigma_{\xi\eta} = \mathbf{E}\left[(\xi - \mu_\xi)(\eta - \mu_\eta)\right] = \mathbf{E}[\xi\,\eta] - \mu_\xi\,\mu_\xi$$

– так называемый коэффициент ковариации, μ_ξ, μ_η – математические ожидания ξ, η соответственно.

Если случайные величины независимы, то коэффициент корреляции равен $\rho = 0$.

Уравнение линейной регрессии η на ξ есть линейная функция $y^*(x) = \beta \cdot x + \alpha$, для которой достигается минимум (среди всех линейных функций) среднеквадратической ошибки линейного прогноза значений сл.в. η по значениям сл.в. ξ:

$$\mathbf{E}[\eta - (b \cdot \xi + a)]^2 = \min_{c,d} \mathbf{E}[\eta - (d \cdot \xi + c)]^2.$$

Если дисперсии ξ, η конечны, то уравнение регрессии можно записать в виде:

$$y^*(x) = \rho \frac{\sigma_\eta}{\sigma_\xi}\left(x - \mu_\xi\right) + \mu_\eta = \beta\,x + \alpha,$$

$$\beta = \rho \frac{\sigma_\eta}{\sigma_\xi}, \qquad \alpha = \mu_\eta - \beta\,\mu_\xi,$$

где x – переменная, принимающая возможные значения сл.в. ξ, y^* – прогноз η при $\xi = x$.

Дисперсия ошибки прогноза (*остаточная дисперсия*) равна $\sigma_\eta^2(1 - \rho^2)$.

Замечание I. Для прогноза ξ по значению $\eta = y$ строится уравнение регрессии ξ на η:

$$x^*(y) = \rho \frac{\sigma_\xi}{\sigma_\eta}\left(y - \mu_\eta\right) + \mu_\xi.$$

Замечание II. На основе выборочных данных мы можем найти только оценку линейной регрессии, если заменим неизвестные параметры их соответствующими оценками.

Эллипс рассеяния представляет собой геометрическую характеристику изменчивости и зависимости случайных величин. Кроме области изменения случайного вектора (в основном), эллипс рассеяния показывает характер зависимости случайных величин. Из всего разнообразия геометрических фигур эллипс выбран по следующим причинам. Во-первых, эллипс имеет удобное аналитическое представление. Во-вторых, реальные данные показывают, что основная масса реализаций вектора визуально располагается внутри фигуры, подобной эллипсу. Наконец, у наиболее популярного двумерного нормального распределения линии постоянства функции плотности образованы именно эллипсами.

Определение. Пусть (μ_ξ, μ_η) – вектор математических ожиданий (ξ, η), $(\sigma_\xi^2, \sigma_\eta^2)$ – дисперсии (ξ, η), ρ – коэффициент корреляции. Эллипс относительно переменных (x, y), определяемый уравнением

$$\frac{(x-\mu_\xi)^2}{\sigma_\xi^2} - 2\rho\frac{(x-\mu_\xi)}{\sigma_\xi}\frac{(y-\mu_\eta)}{\sigma_\eta} + \frac{(y-\mu_\eta)^2}{\sigma_\eta^2} = 4(1-\rho^2)$$

называется *эллипсом рассеяния* случайного вектора (ξ, η).

Справедлива следующая

Теорема. а) Пусть $|\rho| < 1$, тогда эллипс рассеяния – единственный эллипс, равномерное распределение внутри которого имеет одинаковые с (ξ, η) математические ожидания (μ_ξ, μ_η), дисперсии $(\sigma_\xi^2, \sigma_\eta^2)$ и коэффициент корреляции ρ.

б) Если вектор (ξ, η) имеет двумерное нормальное распределение, то вероятность того, что он примет значение внутри своего эллипса рассеяния равна $0{,}865$.

в) Линии регрессии проходят через точки касания с эллипсом прямых, параллельных соответствующим осям координат.

г) Оси эллипса рассеяния параллельны осям координат лишь в случае, когда коэффициент корреляции $\rho = 0$, т.е. компоненты вектора не коррелируют.

д) Площадь эллипса рассеяния $4\pi\sqrt{(1-\rho^2)}\sigma_\xi\sigma_\eta$. Другими словами, вектор имеет малую область изменения не только при малых значениях дисперсий, но и при коэффициенте корреляции, близком к 1.

IV.1. Критерий сопряжённости хи-квадрат

а) Области значений признаков разбиваются соответственно на r и s интервалов $(-\infty; X_1], (X_1; X_2], \dots, (X_{r-2}; X_{r-1}], (X_{r-1}; \infty)$ и $(-\infty; Y_1], (Y_1; Y_2], \dots, (Y_{s-2}; Y_{s-1}], (Y_{s-1}; \infty)$;

б) подсчитываются частоты (количества)

$$n_{kj} = \#\Big((x_i, y_i) \in (X_{k-1}; X_k] \times \big(Y_{j-1}; Y_j\big]\Big)$$

попаданий всех пар выборочных данных в каждую двумерную ячейку $(X_{k-1}; X_k] \times \big(Y_{j-1}; Y_j\big]$;

в) заполняется таблица частот

ξ / η	$\le X_1$	X_2	…	X_{r-1}	$> X_{r-1}$	Всего
$> Y_{s-1}$	n_{s1}	n_{s2}	…	$n_{s(r-1)}$	n_{sr}	n_{s*}
Y_{s-1}	$n_{(s-1)1}$	…	…	…	$n_{(s-1)r}$	$n_{(s-1)*}$
…	…	…	n_{kj}	…	…	…
Y_2	n_{21}	…	…	…	n_{2r}	n_{2*}
$\le Y_1$	n_{11}	n_{12}	…	$n_{1(r-1)}$	n_{1r}	n_{1*}
Всего	n_{*1}	n_{*2}	…	$n_{*(r-1)}$	n_{*r}	$n = n_{**}$

где в строке и столбце «Всего» вычисляются суммы n_{k*}, n_{*j} соответствующих строк и столбцов – сумма чисел в строке «Всего» совпадает с суммой чисел в столбце «Всего» и равна общему объёму выборки n;

г) вычисляется статистика критерия сопряжённости

$$\mathbb{X}^2 = n \sum_{k=1}^{r} \sum_{j=1}^{s} \frac{\left(\frac{n_{kj}}{n} - \frac{n_{*j}}{n}\frac{n_{k*}}{n}\right)^2}{\frac{n_{*j}}{n}\frac{n_{k*}}{n}} = \frac{1}{n} \sum_{k=1}^{r} \sum_{j=1}^{s} \frac{\left(n\, n_{kj} - n_{*j}\, n_{k*}\right)^2}{n_{*j}\, n_{k*}}.$$

д) Известно, что если случайные величины ξ, η независимы, то значение $\mathbb{X}^2$ представляет собой реализацию сл.в. с распределением, приближённо описываемым распределением хи-квадрат с $\nu = (r-1)(s-1)$ степенями свободы.

е) Применить общую схему построения критерия, ориентируясь на то, что при справедливости нулевой гипотезы независимости ожидаются малые значения χ^2 (объяснить почему?).

IV.3. Эллипс рассеяния.

При построении эллипса рассеяния нужно решить уравнение, определяющее эллипс, относительно переменной y. Если обозначить $u = \frac{x-\mu_\xi}{\sigma_\xi}, v = \frac{y-\mu_\eta}{\sigma_\eta}$, то уравнение перепишется в виде $v^2 - 2\rho v u + v^2 = 4(1-\rho^2)$. Решая это уравнение школьными методами, получим две ветви эллипса $v = \rho u \pm \sqrt{(1-\rho^2)(4-u^2)}, |u| \le 2$. Т.о., уравнения ветвей эллипса рассеяния:

$$y_{12} = \mu_\eta + \rho \frac{\sigma_\eta}{\sigma_\xi}(x - \mu_\xi) \pm \frac{\sigma_\eta}{\sigma_\xi}\sqrt{1-\rho^2}\sqrt{4\sigma_\xi^2 - (x-\mu_\xi)^2},$$

при $x \in \left[\mu_\xi - 2\sigma_\xi; \mu_\xi + 2\sigma_\xi\right]$.

IV.2. Критерий независимости компонент двумерного случайного вектора

а) По выборочным данным $(x_1, y_1), \dots, (x_n, y_n)$ вычисляется коэффициент корреляции

$$R = \frac{1}{n\, S_x S_y} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

где $\overline{x}, \overline{y}$ – соответствующие выборочные средние, S_x, S_y – выборочные стандартные отклонения (на основе смещённой выборочной дисперсии);

б) находится преобразование Стьюдента для выборочного коэффициента корреляции

$$t = \sqrt{n-2}\,\frac{R}{\sqrt{1-R^2}}.$$

в) Известно, что, если выборка получена из двумерного нормального закона, значение t представляет собой реализацию случайной величины с распределением Стьюдента с $\nu = n - 2$ степенями свободы.

г) Далее применить общую схему построения статистического критерия, ориентируясь на то, что при справедливости нулевой гипотезы независимости истинный коэффициент корреляции ρ равен нулю.

Пример отчёта.

Имярек Джон Карлович гр. 09-0101 (ZadanMS50)

Задание IV.1. Проверка независимости по критерию хи-квадрат

1. Описание физической, биологической, медицинской, … задачи.
2. Условия проведения эксперимента …
3. Описание вероятностной модели наблюдений …
4. Ожидания экспериментатора. Нулевая гипотеза H_0: ... при альтернативе H_1: ....
5. Уровень значимости $\alpha = \cdots$.
6. Применяемый критерий, тестовая статистика, процесс вычисления статистики. Вид критической области.
7. Функция распределение тестовой статистики …
8. Критическая константа C_α находится из уравнения
   a. ... , т.е. – равна …-квантили распределения …
   b. Воспользовавшись …, нашли, что $C_\alpha = \cdots$.
   Окончательный вид критической области …
9.
   a. По представленным данным найдено

	78,55	81,55	84,55	87,55	>87,55	Σ
>123,55	0	0	2	2	3	7
123,55	0	1	7	13	2	23
119,55	1	6	23	12	0	42
115,55	0	5	12	0	0	17
111,55	2	4	0	0	0	6
Σ	3	16	44	27	5	$n = 95$

Статистика $\mathbb{X}^2$	75,75
степени свободы	16
2.5%-я критическая область	$\mathbb{X}^2$ ...
Гипотеза независимости	…
с критическим уровнем значимости	$p\text{-}val < 0.001$

   b. Критический уровень значимости p-value вычислялся по формуле

$$p\text{-}val = \cdots = 9.6 \times 10^{-10}.$$

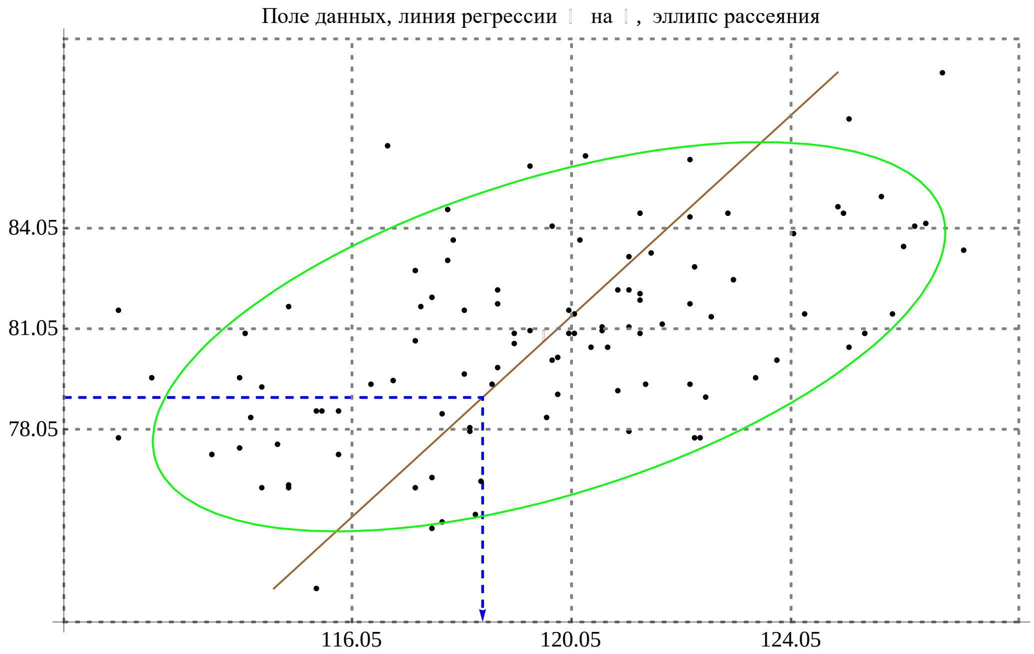
Т.к. p-val ..., следует считать наблюдения зависимыми.

Замечание I. Если располагать ячейки слева-направо и снизу-вверх (как у меня), то можно увидеть направление зависимости; у меня – с ростом одной характеристики (X) наблюдается тенденция к увеличению другой характеристики (Y).

Задание IV.2. Наилучший линейный прогноз, эллипс рассеяния

1. По результатам независимых измерений рентабельности и доходности $n = 95$ предприятий найти оценки коэффициентов линейной среднеквадратической регрессии доходности (ξ) на рентабельность (η); представить график линии регрессии в поле всех данных; найти прогноз значения доходности при значении рентабельности $\eta = 79$; дать оценку точности прогноза, изобразить эллипс рассеяния.
2. Условия проведения эксперимента …
3. По представленным данным найдено

Коэффициент линейной регрессии	$\beta = 0{,}667$
Уравнение регрессии ξ на η	$x = 0{,}667\, y + 65{,}732$
Прогноз при $y = 79$	$x = 118.43$
Коэфф.корреляции	$R = 0{,}537$
Стандарт.отклонение наблюдений прочности	$S_x = 3{,}611$
Оценка стандарт. ошибки прогноза	3,046



Вывод. При таком невысоком значении коэффициента корреляции ($R = 0{,}537$; стандартная ошибка прогноза равна $3{,}05$) прогностические качества линии регрессии очень низкие.

Замечание (для исполнения, но не для копипастирования в отчёт). Линии сетки в поле данных (серые, пунктирные) совпадают с границами ячеек задания IV.1. Такое представление помогает видеть правильность заполнения ячеек. Красный квадрат – центр данных, синяя линия – процесс нахождения прогноза с помощью линии регрессии (можно без стрелки).

Задание IV.3. Проверка независимости по выборочному коэффициенту корреляции

1. Описание физической, биологической, медицинской, … задачи.
2. Условия проведения эксперимента и наблюдаемый сл.вектор …
3. Описание вероятностной модели наблюдений …
4. Ожидания экспериментатора. Нулевая гипотеза H_0: ... при альтернативе H_1: ....
5. Уровень значимости $\alpha = \cdots$.
6. Применяемый критерий, тестовая статистика, процесс вычисления статистики. Вид критической области.
7. Функция распределение тестовой статистики …
8. Критическая константа C_α находится из уравнения
    a. ... , т.е. – равна …-квантили распределения …
    b. Воспользовавшись …, нашли, что $C_\alpha = \cdots$.
    Окончательный вид критической области …
9.
    a. По представленным данным найдено

	x	y
Среднее,	119,64	80,81
Дисперсия, s^2	13,039	8,440
Стандарт.отклонение s	3,611	2,905
Объём выборки, n	103	103
Коэффициент корреляции, R		0.537
Преобразование Стьюдента, t		6.392
5%-я критическая область		$\|t\| > 1.66$
Гипотеза независимости		…
с критическим уровнем значимости		p-val< 0.0001

    b. Критический уровень значимости p-value вычислялся по формуле

$$\text{p-val} = \cdots = 2.6 \times 10^{-9}.$$

Т.к. p-va... , следует считать отклонение выборочного коэффициента корреляции от нуля статистически … значимым.