4. Корреляционный анализ.

Корреляционный анализ - раздел математической статистики. *Основной задачей корреляционного анализа является выявление статистической зависимости между случайными переменными путём оценок различных коэффициентов корреляции.*

При **функциональной** зависимости между величинами $y=f(x)$, которую изучает математический анализ, каждому значению независимой переменной x соответствует определённое значение величины y.

В теории вероятностей и математической статистике изучается, как правило, **стохастическая** зависимость между случайными величинами, когда одному и тому значению x может соответствовать в зависимости от случая различные значения величины y. При стохастической зависимости величины не связаны функционально, но как случайные величины связаны совместным распределением вероятности.

Методы корреляционного анализа дают хорошие результаты в том случае, когда данные эксперимента можно считать выбранными из совокупности, распределённой по многомерному нормальному закону.

При изучении по выборке корреляционной зависимости двух случайных величин, сначала на координатной плоскости изображают все выборочные точки (x_1, y_1), (x_2, y_2),…, (x_n, y_n). Это изображение называют **корреляционным полем**.

Затем составляют корреляционную таблицу:

Возможные значения	y_1 … y_j … y_m	Всего
X_1	n_{11} … n_{1j}… n_{1m}	$n_{1\bullet}$
…	.....	…
x_i	n_{i1} … n_{ij}… n_{im}	$n_{i\bullet}$
…	.....	…
x_l	n_{l1} … n_{lj}… n_{lm}	$n_{l\bullet}$
Всего	$n_{\bullet 1}$ … $n_{\bullet j}$… $n_{\bullet m}$	n

где n_{ij}- частота, с которой пара (x_i , y_j) встретилась в выборке;

для непрерывных распределений в качестве x_i и y_j берут середины интервалов группировки.

Коэффициент корреляции между случайными переменными X и Y определяется как

$$r = \frac{M(X \cdot Y) - M(X) \cdot M(Y)}{\sigma(X) \cdot \sigma(Y)}.$$

Его оценкой является выборочный коэффициент корреляции, который можно вычислить как:

$$\hat{r} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{S_X \cdot S_Y}$$

,

После вычисления выборочного коэффициента корреляции проверяют гипотезу о значимости связи, так как выборочный коэффициент корреляции, как правило, не совпадает с теоретическим и может не равняться нулю из – за отбора переменных в выборку.

Обычно проверяется основная гипотеза об отсутствии корреляционной связи: $H_0: r = 0$ против альтернативы $H_1: r \neq 0$.

В случае справедливости основной гипотезы статистика

$$St = \frac{\hat{r} \cdot \sqrt{n-2}}{\sqrt{1-\hat{r}^2}}$$

имеет распределение Стьюдента с $n-2$ степенями свободы.

Считают, что выборочный коэффициент корреляции значимо отличается от нуля, если значение статистики

$$|St| > t_{1-\alpha,n-2} ,$$

где $t_{1-\alpha,n-2}$ - критическое значение распределения Стьюдента, определённое на уровне значимости α при числе степеней свободы, равном $n-2$ (т.е. квантиль уровня $\frac{1-\alpha}{2}$ распределения Стьюдента с $n-2$ степенями свободы).

Для значимого коэффициента корреляции можно найти доверительный интервал. При его построении используют Z- преобразование Фишера:

$$z = \frac{1}{2}\ln\frac{1+\hat{r}}{1-\hat{r}} .$$

Величина z распределена асимптотически нормально. Таким образом, доверительный интервал для коэффициента корреляции имеет вид:

$$P\left\{ r \in \left[th(z - \frac{\varphi_{1-\alpha}}{\sqrt{n-3}}); th(z + \frac{\varphi_{1-\alpha}}{\sqrt{n-3}}) \right] \right\} = 1-\alpha,$$

где $th\, z = \frac{e^z - e^{-z}}{e^z + e^{-z}}$ (гиперболический тангенс);

$\varphi_{1-\alpha}$ - критическая граница нормального распределения (т.е. квантиль уровня $\frac{1-\alpha}{2}$ стандартно нормального закона распределения).

Если объём выборки n достаточно велик, то можно пользоваться более простой формулой:

$$P\left\{ r \in \left[\hat{r} - \left(1-\hat{r}^2\right) \cdot \frac{\varphi_{1-\alpha}}{\sqrt{n}}; \hat{r} + \left(1-\hat{r}^2\right) \cdot \frac{\varphi_{1-\alpha}}{\sqrt{n}} \right] \right\} = 1-\alpha .$$

Из курса теории вероятностей известно, что коэффициент корреляции позволяет судить лишь о наличие **линейной зависимости** между случайными величинами. Однако часто возникает необходимость в

показателе интенсивности связи **в любой форме зависимости**. Для этой цели применяют корреляционное отношение

$$\eta_{Y|X} = 1 - \frac{\sigma^2_{Y|X}}{\sigma^2_Y},$$

где $\sigma^2_{Y|X} = M\left[Y - M\left(Y \mid X\right)\right]^2$ - условная дисперсия случайной величины ***Y***.

В математической статистике в качестве оценки корреляционного отношения используют эмпирическое (или выборочное) корреляционное отношение:

$$\hat{\eta}_{Y|X} = \frac{1}{S_Y}\sqrt{\frac{1}{n}\sum_{i=1}^{l}\left(\bar{y}_i - \bar{y}\right)^2 \cdot n_{i\bullet}},$$

где $\bar{y}_i = \frac{1}{n_{i\bullet}}\sum_{j=1}^{m} y_j \cdot n_{ij}$ (групповые средние, т.е. средние значения y, вычисленные при условии, что $x=x_i$);

под корнем стоит межгрупповая дисперсия.

Корреляционное отношение (как теоретическое, так и эмпирическое) обладает рядом важных свойств:

1) $0 \le \eta \le 1$;
2) Если $\eta=1$, то между переменными существует функциональная зависимость;
3) Если $\eta=0$, то корреляционная связь отсутствует;
4) $\hat{\eta}_{Y|X} \neq \hat{\eta}_{X|Y}$, то есть при вычислении корреляционного отношения существенно, какую переменную считать независимой, а какую – зависимой;
5) В случае линейной зависимости коэффициент корреляции равен корреляционному отношению: $r=\eta$.

В силу последнего свойства величину $\hat{\eta}^2_{Y|X} - \hat{r}^2$ используют в качестве индикатора отклонения зависимости от линейной.

При проверке линейности связи пары признаков (т.е. двух случайных величин) учитывают, что статистика

$$F = \frac{\left(\hat{\eta}^2 - \hat{r}^2\right)\cdot\left(n-m\right)}{\left(1-\hat{\eta}^2\right)\cdot\left(m-2\right)}$$

имеет распределение Фишера – Снедекора с ($m-2;n-m$) степенями свободы (здесь n – объём выборки, m – число интервалов группировки).

Пример. Исследовать корреляционную зависимость между суточной выработкой продукции (Y тонн) и величиной основных производственных фондов (X млн.руб.). Данные уже сгруппированы, в качестве значений x_i и y_j приведены середины интервалов.

	y_1=9	y_2=13	y_3=17	y_4=21	y_5=25	Всего ($n_{i\bullet}$)	Групповые средние ($\bar{y}_i$)

x_1 =22.5	2	1	-	-	-	3	10.3
x_2 =27.5	3	6	4	-	-	13	13.3
x_3 =32.5	-	3	11	7	-	21	17.8
x_4 =37.5	-	1	2	6	2	11	20.3
x_5 =42.5	-	-	-	1	1	2	23.0
Всего ($n_{\bullet j}$)	5	11	17	14	3	50 (=***n***)	
Групповые средние ($\bar{x}_j$)	25.5	29.3	31.9	35.4.	39.2		

Используя данные, приведённые в таблице, находим:

1) выборочные средние $\bar{x} = 32{,}1$ (млн. руб.), $\bar{y} = 16{,}92$ (тонн);

2) выборочные дисперсии $S_X^2 = 21{,}84$; $S_Y^2 = 18{,}2336$

3) эмпирический коэффициент корреляции $\hat{\boldsymbol{r}} = \dfrac{14{,}768}{\sqrt{21{,}84 \cdot 18{,}2336}} = 0{,}740.$

Проверим на уровне α=0,05 значимость полученного коэффициента корреляции. Значение статистики

$$St = \frac{\hat{r} \cdot \sqrt{n-2}}{\sqrt{1-\hat{r}^2}} = \frac{0{,}74 \cdot \sqrt{50-2}}{\sqrt{1-0{,}74^2}} = 7{,}62\,.$$

По таблице находим критическое значение распределения Стьюдента $t_{0{,}95;48}$=2,01. Сравнивая полученное значение статистики и критическое значение распределения, можно сделать вывод, что коэффициент корреляции значимо отличается от нуля.

Построим доверительный интервал, используя Z – преобразование Фишера.

$$\boldsymbol{z} = \frac{1}{2}\ln\frac{1+\hat{\boldsymbol{r}}}{1-\hat{\boldsymbol{r}}} = \frac{1}{2}\ln\frac{1+0{,}74}{1-0{,}74} = 0{,}9505.$$

По таблице находим $\varphi_{0{,}95} = 1{,}96$. Следовательно, доверительный интервал для теоретического коэффициента корреляции имеет вид:

$$\left[th(0{,}9505 - \frac{1{,}96}{\sqrt{50-3}}); th(0{,}9505 + \frac{1{,}96}{\sqrt{50-3}}) \right].$$

Таким образом,

$$P\{r \in [0{,}581; 0{,}844]\} = 0{,}95.$$

Далее вычислим эмпирическое корреляционное отношение

$$\hat{\boldsymbol{\eta}}_{\boldsymbol{Y|X}} = \sqrt{\frac{10{,}36}{18{,}23}} = 0{,}754.$$

Полученное значение близко к выборочному коэффициенту корреляции $\hat{r} = 0{,}740$, поэтому можно предположить, что зависимость между переменными близка к линейной.

Для проверки последней гипотезы, учитывая, что количество интервалов группировки *m*=5, вычислим значение статистики

$$F = \frac{\left(\hat{\eta}^2 - \hat{r}^2\right)\cdot(n-m)}{\left(1-\hat{\eta}^2\right)\cdot(m-2)} = \frac{(0{,}754^2 - 0{,}740^2)(50-5)}{(1-0{,}754^2)(5-2)} = \frac{0{,}94122}{1{,}294452} = 0{,}727.$$

Табличное значение $F_{0,95(3;45)} = 2{,}57$, следовательно, связь можно считать линейной.

В случае, когда изучаются не количественные признаки, а качественные, обычные меры зависимости не годятся.

Однако если удаётся как-то упорядочить изучаемые объекты в отношении некоторого признака, то есть приписать им порядковые номера – **ранги** (по два номера в соответствии с двумя признаками), то в качестве выборочной характеристики связи можно воспользоваться **ранговым коэффициентом корреляции**:

$$\rho = 1 - \frac{6\cdot\sum_{i=1}^{n} d_i^2}{n\cdot(n^2-1)},$$

где d_i - разность рангов по обоим признакам для каждого объекта.

По степени уклонения ρ от нуля можно сделать некоторое заключение о степени зависимости качественных признаков. Проверка гипотезы независимости признаков при небольшом объёме выборки производится с помощью специальных таблиц, а при n>10 для вычисления критических значений выборочных коэффициентов пользуются тем, что эти величины распределены приближенно нормально.

5. Введение в регрессионный анализ.

Первоначально термин «регрессия» был употреблён Ф. Гальтоном (1886) в теории наследственности в следующем специальном смысле. «Возвратом к среднему состоянию» (regression to mediocrity) было названо явление, состоящее в том, что дети тех родителей, рост которых превышает среднее значение на *a* единиц, имеют в среднем рост, превышающий среднее значение, меньше, чем на *a* единиц.

Регрессионная зависимость является частным случаем стохастической зависимости и подразумевает зависимость среднего значения величины ***Y*** от другой случайной величины ***X*** (одномерной или многомерной).

Регрессионная зависимость ***Y*** от ***X*** проявляется в изменении средних значений ***Y*** при изменении ***X***, хотя при каждом фиксированном значении ***X=x*** величина ***Y*** остаётся случайной величиной с определённым распределением.

Регрессия случайной величины ***Y*** по ***X*** – это условное математическое ожидание ***Y***, вычисленное при условии, что случайная величина ***X*** приняла значение, равное ***x***:

$$y(x)=M(Y|X=x).$$

В математической статистике имеют дело с оценками соответствующих вероятностных характеристик, поэтому в качестве оценки условного математического ожидания принимают условное среднее.

Если при каждом значении $x = x_i$ наблюдается n_i значений $y_1^{(i)}, y_2^{(i)}, ..., y_{n_i}^{(i)}$ величины y , то зависимость средних арифметических

$$\overline{y}^{(i)} = \frac{y_1^{(i)} + y_2^{(i)} + ... + y_{n_i}^{(i)}}{n_i}$$

от x_i и является регрессией в статистическом понимании этого термина.

Примером такого рода зависимости служит, в частности, зависимость средних диаметров северных сосен от их высот (объём выборки = 624).

	Высота (м)													
	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Средний диаметр (см)	18,5	18,6	17,7	20	22,9	25	27,2	30,1	32,7	38,3	40	41,8	49,5	43,5

Если число наблюдений, соответствующее некоторым значения ***X*** недостаточно велико, то такой метод может привести к ненадёжным результатам.

Уравнение ***y=y(x)***, в котором ***x*** играет роль «независимой» переменной, называют **уравнением регрессии**, а соответствующий график – **линией** или **кривой регрессии**.

Линия регрессии может быть приближенно восстановлена по достаточно обширной корреляционной таблице: за приближенное значение ***y(x)*** принимают среднее из тех наблюдённых значений ***Y*** , которым соответствует значение ***X=x***.

Для выяснения вопроса, насколько хорошо регрессия передаёт изменение ***Y*** при изменении ***X***, используется условная дисперсия ***Y*** при данном значении ***X=x*** – **дисперсия *Y* относительно линии регрессии** (мера рассеяния относительно линии регрессии):

$$\sigma^2(x) = M\left[Y - M\left(Y \mid X = x\right)\right]^2 = M\left[Y - y(x)\right]^2.$$

При точной функциональной зависимости величина ***Y*** при данном ***X=x*** принимает лишь одно определённое значение, то есть рассеяние вокруг линии регрессии равно нулю. Таким образом, если $\sigma^2(x) = 0$ при всех значения ***x***, то можно с достоверностью утверждать, что ***Y*** и ***X*** связаны строгой функциональной зависимостью. Если $\sigma^2(x) \neq 0$ ни при каком значении ***x*** и ***y(x)*** не зависит от ***x***, то говорят, что регрессия ***Y*** по ***X*** отсутствует.

Наиболее простым является тот случай, когда регрессия ***Y*** по ***X*** линейна:

$$y(x) = \beta_0 + \beta_1 \cdot x,$$

где числа β_0 и β_1 называют коэффициентами регрессии.

Коэффициенты линейной регрессии вычисляют по формулам:

$$\beta_0 = m_Y - r\frac{\sigma_Y}{\sigma_X}m_X, \qquad \beta_1 = r\frac{\sigma_Y}{\sigma_X} \quad .$$

Если двумерное распределение $\boldsymbol{Y}$ и $\boldsymbol{X}$ нормально, то линия регрессии $\boldsymbol{Y}$ по $\boldsymbol{X}$ (так же как и $\boldsymbol{X}$ по $\boldsymbol{Y}$) является прямой с уравнением

$$y = m_Y + r\frac{\sigma_Y}{\sigma_X}(x - m_X).$$

В этом случае корреляционное отношение совпадает с коэффициентом корреляции и условная дисперсия не зависит от $\boldsymbol{x}$ (является постоянной величиной):

$$\sigma^2(x) = \sigma_Y^2\left(1 - r^2\right).$$

Следовательно, коэффициент корреляции полностью определяет степень концентрации распределения вблизи линии регрессии.

Если регрессия $\boldsymbol{Y}$ по $\boldsymbol{X}$ отлична от линейной, то уравнение

$$y = m_Y + r\frac{\sigma_Y}{\sigma_X}(x - m_X).$$

является **линейным приближением** истинного уравнения регрессии.

Коэффициенты регрессии обычно неизвестны, и их оценивают по выборочным данным:

$$\hat{\beta}_1 == \frac{\sum_{i=1}^{n} x_i \cdot y_i - n \cdot \bar{x} \cdot \bar{y}}{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2} \quad ; \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \quad .$$

Линейная функция

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

определяет **эмпирическую линию регрессии**, которая служит статистической оценкой неизвестной истинной линии регрессии.

Рассеяние вокруг линии регрессии можно оценить, используя эмпирическую среднюю дисперсию относительно линии регрессии:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{\hat{e}=1}^{n}\left[y_k - \hat{y}(x_k)\right]^2$$

Этот метод, в предположении нормальной распределённости результатов наблюдений, даёт, в некотором смысле, оптимальные результаты и позволяет проводить экстраполяцию (прогнозирование) значений величины $\boldsymbol{Y}$ по имеющимся значениям величины $\boldsymbol{X}$.

Доверительный интервал для прогнозируемого значения $\boldsymbol{Y}$ имеет вид:

$$\left[\hat{y}(x) - t_p\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}; \hat{y}(x) + t_p\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}\right],$$

где t_p - критическая граница распределения Стьюдента с ***(n-2)*** степенями свободы, соответствующая уровню ***p***.

Пример. В магазине постельных принадлежностей в течение 5 дней подсчитывали число покупок простыней ***X*** и подушек ***Y***.

x_i	10	20	25	28	30
y_i	4	8	7	12	14

Найти выборочной коэффициент корреляции и выборочное уравнение линейной регрессии.

Транспонируем и расширим таблицу для упрощения подсчетов:

	x_i	y_i	x_i^2	$x_i \cdot y_i$	y_i^2
i = 1	10	4	100	40	16
i = 2	20	8	400	160	64
i = 3	25	7	625	175	49
i = 4	28	12	784	336	144
i = 5	30	14	900	420	196
Всего	**113**	**45**	**2809**	**1131**	**469**

Сначала вычислим выборочные средние:

$$\bar{x} = \frac{1}{n} \cdot \sum_{k=1}^{n} x_k = \frac{113}{5} = 22{,}6 \qquad \bar{y} = \frac{1}{n} \cdot \sum_{k=1}^{n} y_k = \frac{45}{5} = 9.$$

Находим значение выборочного коэффициента корреляции:

$$\hat{r} = \frac{\sum_{k=1}^{n} x_k \cdot y_k - n \cdot \bar{x} \cdot \bar{y}}{\sqrt{\left(\sum_{k=1}^{n} x_k^2 - n \cdot \bar{x}^2\right) \cdot \left(\sum_{k=1}^{n} y_k^2 - n \cdot \bar{y}^2\right)}} = \frac{1131 - 5 \cdot 22{,}6 \cdot 9}{\sqrt{\left(2809 - 5 \cdot 22{,}6^2\right) \cdot \left(469 - 5 \cdot 9^2\right)}} = 0{,}89.$$

Посчитаем выборочные коэффициенты линейной регрессии

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{n} x_k \cdot y_k - n \cdot \bar{x} \cdot \bar{y}}{\sum_{k=1}^{n} x_k^2 - n \cdot \bar{x}^2} = \frac{1131 - 5 \cdot 22{,}6 \cdot 9}{2809 - 5 \cdot 22{,}6^2} = 0{,}447 .$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} = 9 - 0{,}447 \cdot 22{,}6 = -1{,}1 .$$

Отсюда выборочное уравнение линейной регрессии имеет вид:

$$\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x .$$

Подставляя вычисленные значения, получим:

$$\hat{y}(x) = 0{,}447 \cdot x - 1{,}1 .$$

Построим доверительный интервал для прогнозируемого значения ***Y.***

Сначала вычислим среднее отклонение вокруг линии регрессии:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{\hat{e}=1}^{n}\left[y_k - \hat{y}(x_k)\right]^2 =$$

$$= \frac{1}{3}\left(\left[5{,}1 - 0{,}447 \cdot 10\right]^2 + \left[9{,}1 - 0{,}447 \cdot 20\right]^2 + \left(8{,}1 - 0{,}447 \cdot 25\right)^2\right) +$$

$$+ \frac{1}{3}\left(\left(13{,}1 - 0{,}447 \cdot 28\right)^2 + \left(15{,}1 - 0{,}447 \cdot 30\right)^2\right) = 4{,}36 .$$

Отсюда $\hat{\sigma} \approx 2{,}1$.

Зададимся уровнем значимости p=0,1 ,тогда критическая граница $t_p = 2{,}35$ и доверительный интервал имеет вид:

$$\left[\hat{y}(x) - t_p\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2}}; \hat{y}(x) + t_p\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}^2}}\right] =$$

$$= \left[0{,}447\cdot x - 1{,}1 - 2{,}35\cdot 2{,}1\cdot\sqrt{\frac{1}{5} + \frac{(x-22{,}6)^2}{2809 - 5\cdot 22{,}6^2}}; 0{,}447\cdot x - 1{,}1 + 2{,}35\cdot 2{,}1\cdot\sqrt{\frac{1}{5} + \frac{(x-22{,}6)^2}{2809 - 5\cdot 22{,}6^2}}\right]$$

$$\left[0{,}447\cdot x - 1{,}1 - 0{,}3\sqrt{(x-22{,}6)^2 + 51}; 0{,}447\cdot x - 1{,}1 + 0{,}3\sqrt{(x-22{,}6)^2 + 51}\right]$$