1. ЛАБОРАТОРНАЯ РАБОТА №1. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Цель работы – освоение методики проведения корреляционного анализа данных.

1.1 Теоретические сведения

Одной из важнейших задач статистики является изучение объективно существующих связей между явлениями. Формы проявления существующих взаимосвязей весьма разнообразны, однако принято различать их два основных вида: ***функциональные*** и ***статистические*** связи.

Функциональной называют такую связь, при которой определённому значению факторного признака соответствует одно и только одно значение результативного. Такая связь возможна при условии, что на поведение одного результативного признака влияет только один факторный признак и никакие другие.

Обычно функциональные связи устанавливаются на основе основных физических законов и широко используются при построении аналитических моделей (например, зависимость тока от напряжения в элементе электрической цепи).

Однако при функционировании многих сложных систем часто проявляются ***статистические*** связи, при которых строго определённому значению факторного признака ставится в соответствие множество значений результативного. Для описания статистических связей устанавливается один или несколько определяющих (учтенных) факторных признаков.

Строгое различие между функциональной и статистической связью устанавливает их математическая формулировка.

Любую функциональную связь можно представить уравнением:

$$y_i = f(x_i), \tag{1.1}$$

где y_i – результативный признак ($i = 1, \ldots, n$);

$f(x_i)$ – функция связи результативного и факторного признаков;

x_i – факторный признак ($i = 1, \ldots, n$).

Статистическая связь представляется уравнением следующего вида:

$$\tilde{y}_i = f(x_i) + \varepsilon_i, \tag{1.2}$$

где $\tilde{y}_i$ – расчетное значение результативного признака;

$f(x_i)$ – часть значения результативного признака, сформировавшегося под действием учтенных факторов;

ε_i – часть значения результативного признака, возникающая вследствие действия неконтролируемых факторов или ошибок измерения.

Корреляционная (статистическая) связь появляется, когда одному и тому же значению аргумента (независимой переменной) соответствует ряд значений функции (зависимой переменной). Тогда связь обнаруживается в виде тенденции изменения средних значений функции в зависимости от изменений аргумента. Этим корреляционная связь отличается от функциональной, которая возникает в случае, когда заданному значению аргумента соответствует вполне определенное значение функции. По сути, корреляционная связь является неполной, так как зависимость между функцией и аргументом в каждом конкретном случае подвержена влиянию со стороны других факторов (зачастую носящих переменный характер).

Основные задачи корреляционного анализа – это определение и выражение формы аналитической зависимости результативного признака y от факторных признаков x_i.

Отличительная черта корреляционного анализа – измерение тесноты связи между y и x.

Основные числовые характеристики - коэффициент корреляции и корреляционное отношение.

Выделяют следующие этапы корреляционного анализа:

1. Выявление наличия взаимосвязи между признаками;
2. Определение формы связи;
3. Определение силы (тесноты) и направления связи.

Простейшим визуальным способом выявить наличие взаимосвязи между количественными переменными является построение диаграммы рассеяния, которая представляет собой график, на котором по горизонтальной оси (x) откладывается одна переменная, а по вертикальной (y) другая. Каждому объекту на диаграмме соответствует точка, координаты которой равняются значениям пары выбранных для анализа переменных.

Пример диаграммы рассеяния представлен на рисунке 1.1.

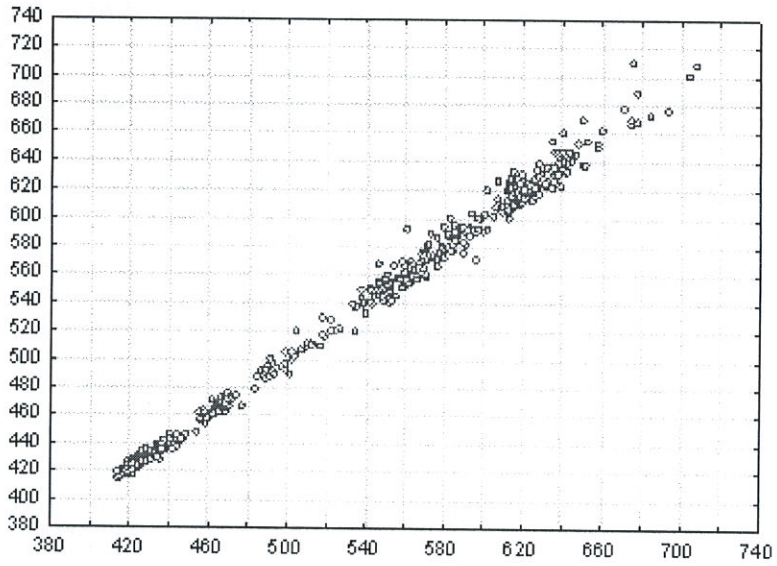

Рисунок 1.1 – Диаграмма рассеяния случайной величины

Поскольку наиболее простой формой зависимости в математике является прямая, то в корреляционном и регрессионном анализе наиболее популярны линейные модели.

Тесноту и направление парной линейной корреляционной связи измеряют с помощью линейного коэффициента корреляции r_{xy}:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\cdot(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\overline{x \cdot y} - \overline{x}\cdot\overline{y}}{\sigma_x \cdot \sigma_y}, \qquad (1.3)$$

где n – количество наблюдений;

x_i, y_i – данные наблюдений;

$\overline{x}, \overline{y}$ – средние значения переменных x и y;

$\sigma_x = \sqrt{D_x} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sqrt{\overline{x^2} - \overline{x}^2}$ – среднеквадратическое отклонение переменной x;

$\sigma_y = \sqrt{D_y} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2} = \sqrt{\overline{y^2} - \overline{y}^2}$ – среднеквадратическое отклонение переменной y.

Коэффициент парной корреляции r_{xy} принимает значения в диапазоне от –1 до +1.

Положительные значения коэффициента корреляции свидетельствуют о положительной связи между признаками, отрицательные – об отрицательной связи, рисунок 1.2.

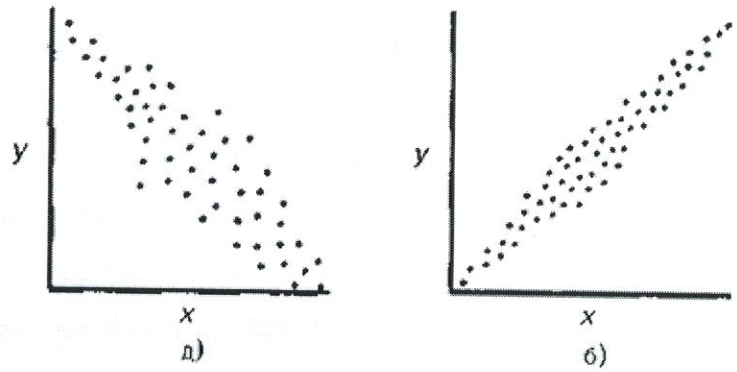

Рисунок 1.2 – Корреляционная связь между переменными
а) – отрицательная; б) – положительная

Если $r_{xy} = 1$, то между двумя переменными существует функциональная положительная линейная связь, т.е. на диаграмме рассеяния соответствующие точки лежат строго на одной прямой с положительным наклоном.

Если $r_{xy} = -1$, то между двумя переменными существует функциональная отрицательная линейная зависимость, т.е. на диаграмме рассеяния соответствующие точки лежат на одной прямой с отрицательным наклоном. При $r_{xy} = 0$ рассматриваемые переменные линейно независимы.

На рисунке 1.3 представлены специально сгенерированные формы зависимостей с соответствующими им коэффициентами корреляции.

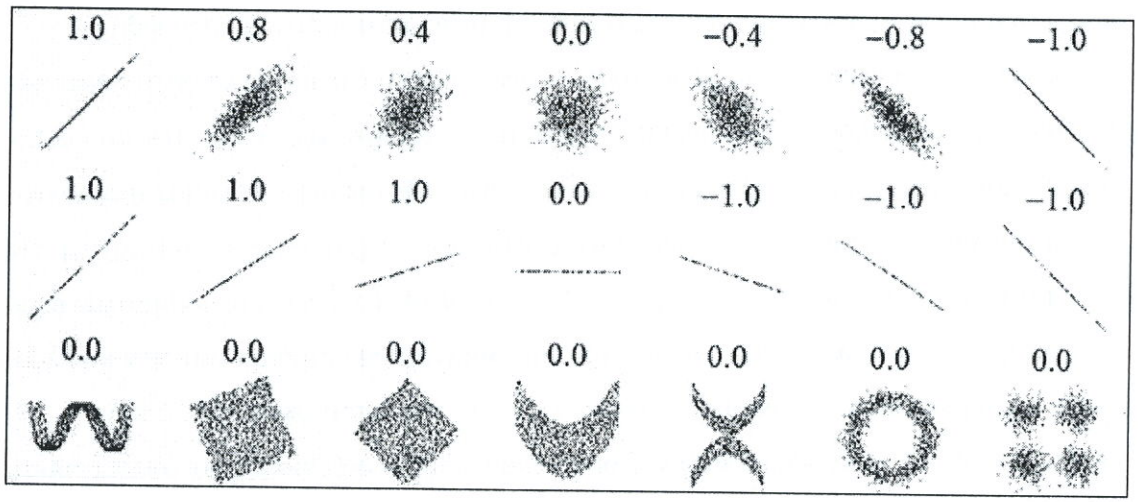

Рисунок 1.3 – Коэффициенты корреляции для разных зависимостей

Для количественной оценки тесноты связи между переменными часто используют шкалу Чеддока, таблица 1.1.

Таблица 1.1 – Оценочная таблица степени тесноты связи между переменными

Количественная мера тесноты связи, r_{xy}	Качественная характеристика силы связи
0,1-0,3	Слабая
0,3-0,5	Умеренная
0,5-0,7	Заметная
0,7-0,9	Высокая
0,9-0,99	Весьма высокая

Если коэффициент корреляции вычислен на основе выборочных данных, то не исключено, что его ненулевое значение является не отражением действительной связи между признаками, а просто получено в результате специфики данной выборки (тогда как в генеральной совокупности коэффициент корреляции равен нулю, т.е. линейной связи между признаками нет).

Для установления связи между переменными необходима проверка коэффициента корреляции на значимость.

Значимость линейного коэффициента корреляции устанавливается на основе t-критерия Стьюдента, согласно которому выдвигается нулевая гипотеза об отсутствии связи между факторным и результативным признаками (H_0: $r_{xy} = 0$). Для проверки нулевой гипотезы нужно рассчитать t-статистику $t_р$ и сравнить ее с табличным значением $t_т$, определяемым с использованием таблицы П1.1 приложения по заданным уровню значимости α и числу степеней свободы k. Если $t_р > t_т$, то гипотеза H_0 отвергается с вероятностью ошибки меньше чем $\alpha \cdot 100\%$. Это свидетельствует о значимости линейного коэффициента корреляции r_{xy} и статистической существенности зависимости между факторным и результативным признаками.

Для вычисления t-критерия Стьюдента используют выражение:

$$t_р = \frac{|r_{xy}| \cdot \sqrt{k}}{\sqrt{1 - r_{xy}^2}}, \qquad (1.4)$$

где $k = n - 2$ для малой выборки;

Под уровнем значимости α понимается вероятность отвергнуть верную гипотезу. Уровень значимости обычно принимается равным $\alpha = 0{,}05$ или $\alpha = 0{,}01$.

Если значение коэффициента корреляции вычислено по выборочным данным, то для оценки его значения в генеральной совокупности, нужно произвести его интервальную оценку: построить доверительный интервал.

Интервальная оценка для коэффициента корреляции (доверительный интервал) определяется по выражению:

$$\left(r_{xy} - t_{\text{т}} \frac{1 - r_{xy}^{\ 2}}{\sqrt{n}}; r_{xy} + t_{\text{т}} \frac{1 - r_{xy}^{\ 2}}{\sqrt{n}} \right), \tag{1.5}$$

где $t_{\text{т}}$ – табличные значения t-критерия Стьюдента, определяемые по таблице П1.1 в зависимости от k степеней свободы и заданного уровня значимости α.

После построения доверительного интервала коэффициента корреляции, делается проверка на попадание нуля в этот интервал.

Если ноль попадет в доверительный интервал, с высокой вероятностью можно считать, что в генеральной совокупности связь между переменными отсутствует. В этом случае коэффициент корреляции является статистически незначимым.

Если ноль не попал в доверительный интервал, то с высокой вероятностью в генеральной совокупности не может быть нулевого значения коэффициента корреляции, что означает: связь между переменными существует. В таком случае коэффициент корреляции является статистически значимым.

Надо иметь в виду, что сам по себе коэффициент корреляции не имеет содержательной интерпретации. Однако его квадрат, называемый коэффициентом детерминации (обычно обозначается R^2 и выражается в %), имеет простой математический смысл – это показатель того, насколько изменения зависимого признака объясняются изменениями независимого. Для парной линейной зависимости коэффициент детерминации можно рассчитать по выражению:

$$R^2 = r_{xy}^{\ 2} \cdot 100\% \tag{1.6}$$

Величина коэффициента детерминации показывает, какая доля общей дисперсии (вариации) результативного признака (y) объясняется влиянием факторного (x).

1.2. Методические указания

По данным таблицы 1.2, для соответствующего варианта задания необходимо:

1. Построить диаграмму рассеяния
2. Вычислить линейный коэффициент парной корреляции r_{xy}.
3. Определить направление и вид связи в рядах данных.
4. Оценить степень тесноты связи между переменными.
5. Проверить значимость коэффициента корреляции r_{xy} при заданном уровне значимости α.

6. Построить доверительный интервал для значимого линейного коэффициента корреляции r_{xy}.
7. Определить коэффициент детерминации.

Результаты вычислений рекомендуется оформить в виде таблицы 1.3.

Таблица 1.2 – Таблица исходных данных

№ вар.	α		Результаты наблюдений														
			1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,05	X	1	3	4	8	7	9	15	14	12	20	16	11	2	5	6
		Y	17	27	38	96	76	103	139	154	87	185	181	112	28	41	61
2	0,1	X	2	4	5	8	6	12	11	18	25	20	14	7	9	10	6
		Y	23	40	62	66	68	82	80	129	202	120	121	55	72	80	55
3	0,01	X	14	16	21	15	2	7	9	6	5	19	8	1	14	10	7
		Y	70	40	23	73	114	94	92	99	99	42	87	116	56	87	80
4	0,05	X	7	9	3	12	15	14	10	8	20	18	8	2	6	5	4
		Y	103	121	114	110	80	87	101	125	66	53	101	122	118	110	119
5	0,1	X	2	5	6	8	12	14	18	20	16	17	9	8	5	4	3
		Y	18	32	45	53	88	77	117	113	96	113	63	56	34	28	26
6	0,01	X	3	4	6	9	11	15	17	6	8	21	25	1	14	16	10
		Y	25	31	51	66	80	97	98	50	53	118	149	15	94	117	68
7	0,05	X	18	16	15	12	10	11	9	7	1	3	6	5	4	2	20
		Y	28	40	32	38	50	61	58	71	80	70	70	69	65	75	6
8	0,1	X	15	16	11	12	10	13	9	8	2	5	7	5	9	2	18
		Y	46	21	41	48	58	36	61	62	75	72	67	65	57	74	28
9	0,05	X	16	15	11	14	10	10	9	8	3	5	7	8	9	1	17
		Y	142	135	103	160	110	117	86	59	29	44	76	69	83	14	181
10	0,05	X	10	11	15	12	14	10	3	8	2	5	20	18	9	2	17
		Y	40	30	20	36	20	31	51	44	56	51	12	18	35	56	19

Таблица 1.3 – Таблица для расчета коэффициента корреляции

n	X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$
1 2 3 4 5 .... n							
Сумма			-	-			
Среднее			-	-	-	-	-

1.3. Требования к отчету

Отчет по лабораторной работе должен содержать:

1. Цель работы
2. Результаты проведенных вычислений
3. Выводы

1.4. Контрольные вопросы

1. В чем заключается основное отличие между функциональной и статистической связью между переменными ?
2. Основные задачи корреляционного анализа данных
3. Как определяется и что характеризует коэффициент детерминации
4. Как вычисляется линейный коэффициент парной корреляции r_{xy} ?
5. Как осуществляется оценка статистической значимости линейного коэффициента парной корреляции r_{xy} ?
6. Что называется уровнем значимости ?
7. Как строится доверительный интервал для линейного коэффициента парной корреляции?