Казанский государственный аграрный университет

Лекционный курс для самостоятельного изучения по дисциплине «Эконометрика»

доцент кафедры экономики и информационных технологий
Газетдинов Ш.М.

Лекция 4: Тема: **Проверка статистических гипотез**

Статистической гипотезой *H* называется предположение относительно параметров или вида распределения случайной величины.

Нулевой (основной) называют выдвинутую гипотезу H_0, а **конкурирующей** (альтернативной)- гипотезу H_1, которая противоречит нулевой.

Проверку статистической гипотезы выполняют на основе данных выборки. Так как выборка имеет ограниченный объём, то появляется возможность принятия ошибочного решения.

Определение: вероятность α того, что будет отвергнута правильная нулевая гипотеза, называется **уровнем значимости.**

Выбор, например *5%* - го уровня значимости означает, что в пяти случаях из ста правильная гипотеза H_0 будет отвергнута. Стремление к уменьшению α ведет в то же время к уменьшению вероятности отвергнуть гипотезу, когда она является ложной.

Статистическим критерием называется случайная величина, которая служит для проверки нулевой гипотезы. В качестве статистического критерия выбирается такая случайная величина, например *t*, точное или примерное, распределение которой известно.

Наблюдаемым значением *t* называется значение критерия, вычисленное по данным выборки.

Множество значений критерия *t* разбивают на две непересекающееся области: **критическую** и **область принятия гипотезы.**

Критической областью называется совокупность значений критерия, при которых гипотеза H_0 отвергается. Различают *одностороннюю* и *многостороннюю* критические области.

Областью принятия гипотезы называется совокупность значений критерия, при которых гипотеза H_0 принимается.

Критическими точками $t_{кр}$ называются точки, отделяющие критическую область и область принятия гипотезы. Критические точки $t_{кр}$ определяются по таблицам известного распределения выбранного критерия *t* при заданном уровне значимости и числе степеней свободы.

Сравнивая наблюдаемое значение критерия с критическими точками, можно принять или отвергнуть нулевую гипотезу.

Лекция 5: **Тема: Корреляционный и регрессионный анализы**

Взаимосвязь между случайными величинами, называется **корреляционным анализом**. Он применяется тогда, когда данные наблюдений являются выбранными из генеральной совокупности, подчиняющейся многомерному нормальному закону

распределения. Это условие обеспечивает линейный характер связи между изучаемыми величинами, что делает правомерным применение в качестве инструментов количественной оценки тесноты связей парного, частного и множественного коэффициентов корреляции. Корреляционный анализ состоит в количественном определении тесноты связи между двумя величинами (при парной связи) и между результативным и множеством факторных признаков (при множественной связи). Основным понятием, используемым в корреляционном анализе, является понятие корреляции.

Корреляция – это статистическая зависимость между случайными величинами, когда изменение одной случайной величины приводит к изменению математического ожидания другой. Корреляция бывает парная, частная и множественная.

Парная корреляция – это связь между двумя признаками (результативным и факторным или между двумя факторными).

Частная корреляция – это связь между результативным и одним из факторных признаков или между двумя факторными признаками, при фиксированных значениях остальных факторных признаков.

Множественная корреляция – это связь между результативным и двумя или более факторными признаками, включенными в исследование.

В корреляционном анализе теснота связи количественно оценивается с помощью соответствующих коэффициентов корреляции. Построение соответствующих коэффициентов корреляции основано на сумме произведений отдельных значений признаков (переменных) x_i и y_i от их средних значений $\bar{x}$ *и* $\bar{y}$: $\sum(x_i - \bar{x}) \cdot (y_i - \bar{y})$. Эта величина, деленная на количество наблюдений в выборке n, называется **выборочным коэффициентом ковариации**:

$$\mathrm{cov}(x, y) = \frac{1}{n}\sum(x_i - \bar{x}) \cdot (y_i - \bar{y}),$$

или

$$\mathrm{cov}(x, y) = \overline{xy} - \bar{x} \cdot \bar{y}.$$

Выборочный коэффициент ковариации характеризует *статистическую меру взаимосвязи* между двумя признаками. Коэффициент ковариации является размерной величиной, размерность которой равняется произведению размерностей признаков, участвующих при его конструировании. Но, несмотря на свою недостаточную информативность, этот коэффициент является достаточно важным, потому что он входит во многие формулы корреляционного и регрессионного анализов. Поэтому, далее подробно остановимся на нем и его свойствах.

Пусть данные наблюдений признаков *x, y* представлены в виде точечного графика – *диаграммы рассеяния наблюдений* (рис. 1).

Точка $(\bar{x}, \bar{y})$ на диаграмме является центром рассеяния признаков *x, y*.

Вертикальная и горизонтальная прямые, проведенные через точку $(\bar{x}, \bar{y})$, разделяют диаграмму на четыре области.

Наблюдения в областях I, III дают положительный вклад в ковариацию, а в областях II, IV – отрицательную.

Если положительные вклады преобладают над отрицательными, то ковариация будет *положительной*, в противном случае она будет *отрицательной.* Положительной

ковариации отвечает прямо пропорциональная связь, а отрицательной – обратно пропорциональная связь.

При положительной (прямо пропорциональной) связи с увеличением одного признака другой признак в среднем также увеличивается, и наоборот при отрицательной (обратно пропорциональной) связи.

Заметим, что $cox(x, x) = \frac{1}{n} \cdot \sum (x_i - \bar{x})^2 = \text{var}(x)$.

Свойства выборочной ковариации:

1. *cov(x, u + v) = cov(x, u) +cov(x, v).*
2. *cov(x, a) = 0, если a = const.*
3. *cov (x, b·u) = b cov (x,u), если b = const.*
4. *cov (u, v) = cov (v, u).*
5. *var(u, v) = var(u) + var(v) + 2·cov(u, v).*

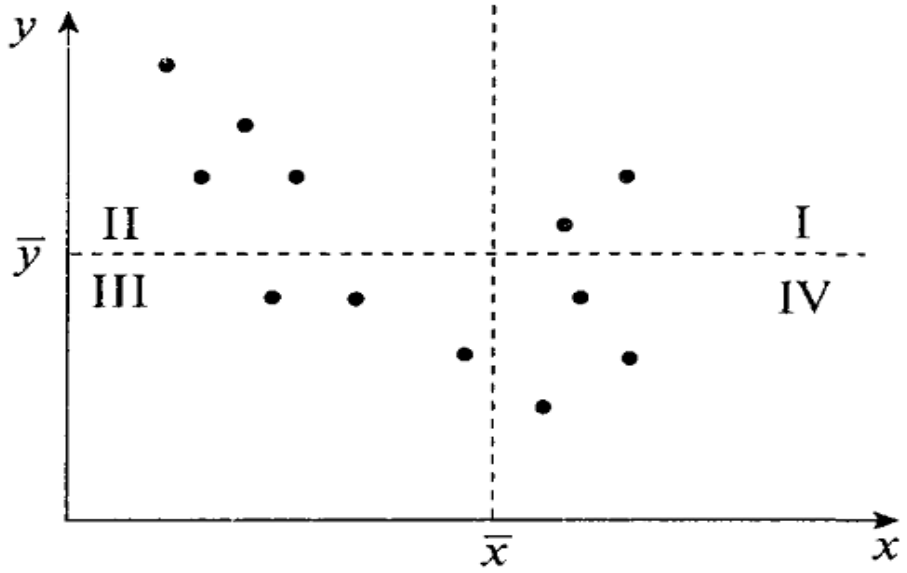

Рис. 1 – Диаграмма рассеивания наблюдений

Доказательство вышеперечисленных свойств вытекает из определения ковариации. Так, например:

$$\text{cov}(x, a) = \frac{1}{n} \sum (x_i - \bar{x}) \cdot (a - a) = 0.$$

Пример 1. Найти значение выборочного коэффициента ковариации между значениями *x* и *y*, по данным приведенным в таблице 1.

Таблица 1

i	1	2	3	4	5	Среднее значение
x	2	4	6	8	10	6
y	3	6	9	12	15	9
x·y	6	24	54	96	150	66

Используя формулу для вычисления значения выборочного коэффициента ковариации получаем:

$$\text{cov}(x, y) = \overline{xy} - \bar{x} \cdot \bar{y} = 66 - 6 \cdot 9 = 12.$$

Лекция 6: Тема : Выборочный коэффициент парной корреляции

Выборочный коэффициент парной корреляции между переменными *X* и *Y* определяемый по выборке из *n* наблюдений вычисляется по формуле:

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - (\overline{x})^2} \cdot \sqrt{\overline{y^2} - (\overline{y})^2}} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x) \cdot \mathrm{var}(y)}}.$$

Выборочный коэффициент парной корреляции дает количественную оценку тесноты линейной связи между переменными *x* и *y*. Он является безразмерной величиной и изменяется в диапазоне $-1 \le r_{xy} \le 1$. Если $r_{xy} = 1$ это означает, что между переменными *X* и *Y* существует прямо пропорциональная линейная функциональная зависимость, если $r_{xy} = -1$ это означает, что между переменными *X* и *Y* существует обратно пропорциональная линейная функциональная зависимость. Если $r_{xy} = 0$, то это означает, что между переменными *X* и *Y* линейной зависимости нет (хотя нелинейная зависимость может существовать), в этом случае говорят, что переменные *X* и *Y* не коррелированы. В случае, когда $-1 < r_{xy} < 1$, говорят что переменные *X* и *Y* стохастически (вероятностно) линейно связаны.

На рис. 2 отражен геометрический смысл коэффициента парной корреляции. На рис. 2, *а* и *б* случайные величины *X* и *Y* коррелированы ($r_{xy} > 0$ или $r_{xy} < 0$), на рис. 2 (*в, г*) – не коррелированы ($r_{xy} = 0$). Если $r_{xy} = 0$, случайные величины могут быть как зависимыми (см. рис. 2, *в*), так и независимыми (см. рис. 2, *г*).

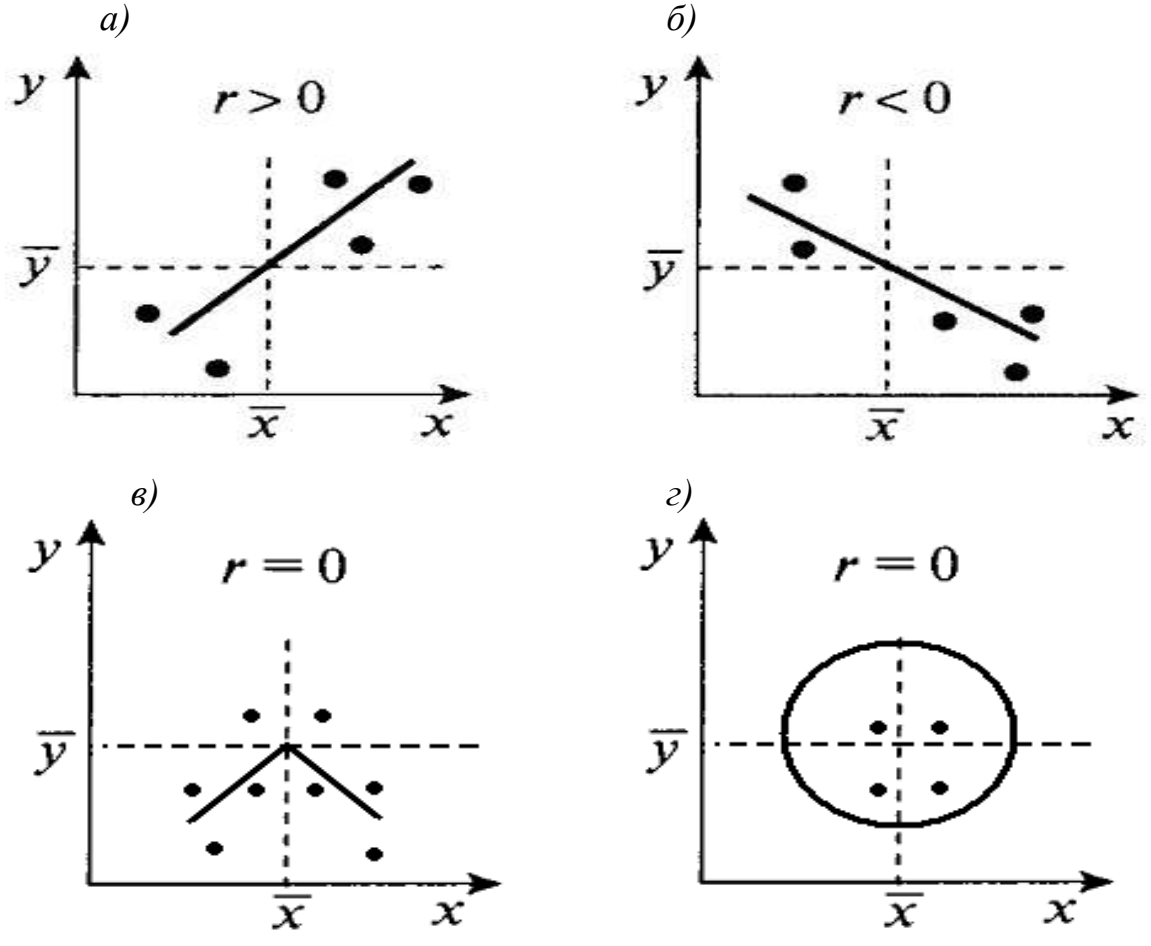

Рис. 2 (а, б, в, г)– Виды связи коэффициента парной корреляции

Выборочный коэффициент парной корреляции является случайной величиной.

Как уже было указано выше, коэффициент парной корреляции r_{xy} служит для количественной оценки тесноты (силы) линейной связи между признаками. Для быстрой и качественной оценки тесноты линейной связи между выборочными значениями признаков с помощью коэффициента парной корреляции имеется специальная шкала, которую предложил английский статистик **Чеддок**. Она позволяет весь диапазон изменения значений модуля коэффициента парной корреляции, разбить на интервалы, для которых можно указать, пусть и достаточно приблизительно, степень тесноты линейной связи. Эта шкала приводится ниже, как видно из нее с возрастанием $|r_{xy}|$ корреляционная связь становится более тесной:

Шкала Чеддока

Значения $\|r_{xy}\|$	0,0-0,1	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-1,0
Характеристика силы связи	Очень слабая	слабая	умеренная	заметная	высокая	весьма высокая

Лекция 7: Тема: Проверка гипотезы о корреляции случайных величин.

Пусть по данным выборки объема n получен выборочный коэффициент корреляции $r_{xy} \neq 0$. Требуется проверить гипотезу о равенстве нулю истинного значения коэффициента корреляции т.е.

$$\begin{cases} H_0 : \rho_{xy} = 0, \\ H_1 : \rho_{xy} \neq 0. \end{cases}$$

В качестве критерия проверки гипотезы H_0 принимается случайная величина:

$$t_{pac} = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1-r_{xy}^2}}.$$

Величина t_{pac} при справедливости гипотезы H_0 имеет распределение Стьюдента (t-статистика) с $\nu = n - 2$ степенями свободы.

Сравнивая наблюдаемое значение критерия t_{pac} с критическим значением $t_{кр}$, определяемым по таблице распределения Стьюдента (приложение 1) по заданному уровню значимости α и по числу степеней свободы n, получим, что:

- если $|t_{pac}| < t_{кр}$, то H_0 принимается, т.е. нет значимой линейной связи между переменными,
- если $|t_{pac}| > t_{кр}$, то H_0 отвергается, т.е. имеется значимая линейная связь между переменными.

Пример 2. Вычислить значение коэффициента парной корреляции между средней заработной платой Y (тыс.руб.) и среднедушевым прожиточным минимумом X (тыс.руб.) по данным нескольких регионов страны (данные приведены в таблице 2).

Для повышения наглядности расчетов промежуточные результаты добавлены, в ту же таблицу 2.

Таблица 2

Региона	x	y	x^2	$x \cdot y$	y^2
1	7	13	49	91	169
2	8	14	64	112	196
3	8	13	64	104	169
4	7	15	49	105	225
5	8	16	64	128	256
6	10	19	100	190	361
7	6	13	36	78	169
8	8	15	64	120	225
9	7	15	49	105	225
10	8	16	64	128	256
Итого	77,000	149,000	603,000	1161,000	2251,000
Среднее	7,700	14,900	60,300	116,100	225,100
	$\bar{x}$	$\bar{y}$	$\overline{x^2}$	$\overline{x \cdot y}$	$\overline{y^2}$

С использованием данных таблицы 2 получаем:

var (x) = $\overline{x^2} - (\bar{x})^2$ *= 60,300 – 59,290 = 1,010;*

var (y) = $\overline{y^2} - (\bar{y})^2$ *= 225,100 – 222,010 = 3,090;*

cov (x, y) = $\overline{xy} - \bar{x} \cdot \bar{y}$ *= 116,100 – 114,73 = 1,370;*

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x) \cdot \text{var}(y)}} = \frac{1{,}370}{\sqrt{1{,}010 \cdot 3{,}090}} = 0{,}775.$$

Проверим значимость выборочного коэффициента парной корреляции.

Наблюдаемое значение критерия есть

$$t_{рас} = \frac{r_{xy} \cdot \sqrt{n - p - 1}}{\sqrt{1 - r_{xy}^2}} = \frac{0{,}775 \cdot \sqrt{8}}{\sqrt{1 - 0{,}601}} = 3{,}474.$$

При $\alpha = 0{,}05$, $v = n - p - 1$ по таблице (приложение 1) находим $t_{кр} = 2{,}301$.

Поскольку $|t_{рас}| = 3{,}474 > t_{кр} = 2{,}301$, то гипотеза H_0 отвергается, т.е. имеется значимая линейная зависимость между выборочными значениями переменных *X* и *Y*.

ПРИЛОЖЕНИЕ 1

Значения *t* - критерия Стьюдента при *5%* - ном уровне значимости

Число степеней свободы,	Значения t-критерия	Число степеней свободы,	Значения t-критерия	Число степеней свободы,	Значения t-критерия
1	12.71	11	2.201	21	2.080
2	4.303	12	2.179	22	2.074
3	3.182	13	2.160	23	2.069
4	2.776	14	2.145	24	2.064
5	2.571	15	2.131	25	2.060
6	2.447	16	2.120	26	2.056
7	2.365	17	2.110	27	2.052
8	2.306	18	2.101	28	2.048
9	2.262	19	2.093	29	2.045
10	2.228	20	2.086	30	2.042
				∞	1.960

Задачи для самостоятельного решения

1. Вычислить значение коэффициента парной корреляции между качеством почв (*X* - баллы) и урожайностью зерновых (Y - ц/га) в нескольких регионах страны, данные по которым приведены в таблице 3. Так же построить диаграмму рассеяния и сделать предположение о характере связи.

Таблица 3

x	45	72	50	48	52	60	90	65	70	95
y	15	20	22	18	20	22	26	24	26	32