

Тема: Статистическое изучение связи между явлениями

1. Виды связей в статистике

Важной задачей статистики является изучение статистических закономерностей, знание которых дает основу для предсказания и управления социально экономическими процессами.

Прежде чем приступить к изучению связи между явлениями необходимо выяснить вид связи между факторным и результативным признаками.

Различают два вида связи:

1- Функциональная связь

когда каждому значению величины факторного признака соответствует только одно значение результативного.

При этом зависимость проявляется с одинаковой силой у всех единиц совокупности. Ее можно представить уравнением

$$y = f(x),$$

где y – результативный признак, x – факторный, $f(x)$ – известная функция

2 - Статистическую связь (стахостическую зависимость)

если причинная зависимость проявляется не в каждом отдельном случае, а в общем, среднем при большом числе наблюдений, то такая зависимость называется **стохастической**.

Частным случаем стохастической связи является **корреляционная связь**, при которой изменение среднего значения результативного признака обусловлено изменением факторных признаков.

Ее можно представить в виде:

$$y = f(x) + \varepsilon,$$

где

$f(x)$ – известная функция, а

ε – часть результативного признака, определяемая неучтенными и неконтролируемыми признаками

По направлению связи выделяют связь прямую (направление изменения результативного признака совпадает с направлением признака-фактора) и обратную

По аналитическому выражению (форме) выделяют связи прямолинейные (или просто линейные) и нелинейные (криволинейные).

Например, параболы, гиперболы: степенная, показательная, экспонентная и т.д.

Для корреляционных связей различают:

связь между одним признаком-фактором и результативным признаком (при абстрагировании влияния других факторов) она называется **парной корреляцией**;
и связь между несколькими факторными признаками и результативным (многофакторная связь) - **множественная корреляция**.

По степени тесноты связи различают количественные критерии оценки тесноты связи.

Величина коэффициента корреляции	Характер связи
До $ \pm 0,3 $	практически отсутствует
$ \pm 0,3 $ до $ \pm 0,5 $	слабая
$ \pm 0,5 $ до $ \pm 0,7 $	умеренная
$ \pm 0,7 $ до $ \pm 1,0 $	сильная

Взаимосвязи между общественными явлениями, установленные на основе теоретического анализа, могут быть изучены, измерены и количественно выражены с помощью различных статистических методов.

Для исследования функциональных связей применяется **балансовый** метод - характеризует зависимость между источниками формирования ресурсов и их использованием

$$O_n + П = В + O_k ,$$

где: O_n , O_k – остатки на начало и конец; $П$, $В$ – поступление и выбытие и индексный метод.

Для изучения корреляционных связей используются:

для связей между **атрибутивными** признаками –

- метод взаимной сопряженности,
- ассоциации
- контингенции

для **количественно** варьирующих признаков –

- метод параллельных рядов,
- ранговой корреляции,
- графический,
- аналитических группировок,
- корреляционно-регрессионный анализ.

2. Параметрические методы изучения связи

При статистических исследованиях корреляционных связей одной из главных задач является определение формы корреляционной связи, т.е. построение модели связи.

Под формой связи понимают тенденцию, которая проявляется в изменении результативного признака в связи с изменением признака-фактора.

Построение и анализ корреляционной модели связи осуществляются с помощью корреляционно-регрессионного анализа, который заключается в построении и анализе статистической модели в виде уравнения регрессии (уравнения корреляционной связи), приближенно выражающей зависимость результативного признака от одного или нескольких признаков-факторов.

Он состоит из следующих этапов:

Предварительного априорного анализа;

Сбора информации и ее первичной обработки;

Построения модели (уравнения регрессии);

Оценки и анализа модели

Все этапы взаимосвязаны между собой, границы их часто переплетаются и носят условный характер.

Выбор формы связи решается на основе теоретического анализа существа изучаемых явлений и исследования эмпирических данных.

Эмпирическое исследование формы связи включает построение графиков корреляционных полей, эмпирических линий регрессии, а также анализ параллельных рядов.

Рассмотрим однофакторную регрессию.

Зависимости могут быть линейными и нелинейными.

Если связь между признаками y и x криволинейная и описывается уравнением **параболы второго порядка**:

$$\overline{y}_x = a_0 + a_1 x + a_2 x^2$$

то система нормальных уравнений имеет вид:

$$\begin{cases} na_0 + a_1 \sum x + a_2 \sum x^2 = \sum y \\ a_0 \sum x + a_1 \sum x^2 + a_2 \sum x^3 = \sum yx \\ a_0 \sum x^2 + a_1 \sum x^3 + a_2 \sum x^4 = \sum yx^2 \end{cases}$$

Если уравнение **гиперболы** вида:

$$\overline{Y_x} = a_0 + \frac{a_1}{x}$$

Система нормальных уравнений:

$$\begin{cases} na_0 + a_1 \sum \frac{1}{x} = \sum y \\ a_0 \sum \frac{1}{x} + a_1 \sum \frac{1}{x^2} = \sum \frac{y}{x} \end{cases}$$

Наиболее часто для определения формы корреляционной связи используют уравнение прямой.

$$y_x = a_0 + a_1 x,$$

где **x** **результативного признака y** от факторного показателя **x**

3. Линейная форма связи и оценка ее параметров.

Уравнение связи называется уравнением регрессии, а анализ, производимый с его помощью, называется регрессионным.

После установления вида функции для модели связи определяются параметры уравнения

$$(a_0, a_1)$$

В уравнениях регрессии параметр a_0 показывает усредненное влияние на результативный признак неучетных (не выделенных для исследования) факторов;

параметр a_1 (а в уравнении параболы и a_2) – это коэффициент регрессии, который показывает, насколько изменяется в среднем значение результативного признака при изменении факторного на единицу его собственного изменения

Оценивание неизвестных параметров производится методом наименьших квадратов (МНК), который дает систему нормальных уравнений:

$$\begin{cases} na_0 + a_1 \sum x = \sum y \\ a_0 \sum x + a_1 \sum x^2 = \sum xy \end{cases}$$

решая которые находятся неизвестные параметры

$$a_0 = \frac{\sum y \sum x^2 - \sum xy \sum x}{n \sum x^2 - \sum x \sum x}; \quad a_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \sum x \sum x}$$

4. Проверка адекватности регрессионной модели.

Среднеквадратическое отклонение результативного признака y_i от выровненных $y_x = \sigma_{ост}^2$ – остаточная дисперсия	$\sigma_{ост} = \sqrt{\frac{\sum (y_x - y_i)^2}{n}}$
Среднеквадратическое отклонение факторного признака x_i от средней $\bar{x}$	$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$
Общая дисперсия	$\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2}{n}$
Среднеквадратическое отклонение модельных значений от средней.	$\sigma_{y_x} = \sqrt{\frac{\sum (y_x - \bar{y})^2}{n}}$
Межгрупповая дисперсия измеряет вариацию групповых средних $\bar{x}_i$ вокруг общей средней. Она измеряет вариацию, обусловленную признаком, положенным в основу группировки.	$\delta^2 = \frac{\sum (\bar{x}_i - \bar{x})^2 f_i}{\sum f_i}$

Для проверки **значимости коэффициентов** линейной регрессии $y_x = a_0 + a_1 x$
при $n < 30$ используют t-критерий Стьюдента.

Для этого вычисляют расчетные значения t – критерия для параметра

$$a_0 : t_{a_0} = |a_0| \frac{\sqrt{n-2}}{\sigma_{ост}};$$

$$a_1 : t_{a_1} = |a_1| \frac{\sigma_x \sqrt{n-2}}{\sigma_{ост}}$$

Полученные значения t_{a_0}, t_{a_1} сравнивают с критическими $t_{кр}$, которые определяют по таблице Стьюдента при заданном *уровне значимости α (обычно $\alpha=0.05$)* и *числом степеней свободы $\nu=n-m=n-2$* .

Параметр признается значимым, если $t_{расч} > t_{кр}$.

Теснота корреляционной связи между x и y может быть измерена **империческим корреляционным отношением**

$$\eta_{\varepsilon} = \sqrt{\delta^2 / \sigma_y^2}$$

Чем ближе оно к 1, тем теснее связь.
При 0 связи нет.

Теснота корреляционной связи между x и y при заданной зависимости определяется **индексом корреляции**

$$R = \sqrt{\frac{\sigma_{\hat{y}}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}}$$

Чем ближе R к 1, тем теснее связь. При $R=0$ связи нет.

Величину R^2 называют **коэффициентом детерминации**.

Коэффициент детерминации характеризует, какая часть общей вариации **y** объясняется изучаемым фактором **x**.

Показателем тесноты линейной связи является **линейный коэффициент корреляции**

$$r = \frac{\overline{xy} - \bar{x} \bar{y}}{\sigma_x \sigma_y} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n \sigma_x \sigma_y} = \frac{\left(\sum xy - \frac{\sum x \sum y}{n} \right)}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}}$$

$$(-1 \leq r \leq 1).$$

Величину r^2 называют **линейным коэффициентом детерминации**.

Для оценки **значимости коэффициента корреляции** r используют t -критерий Стьюдента.

Для этого вычисляют расчетные значения t – критерия

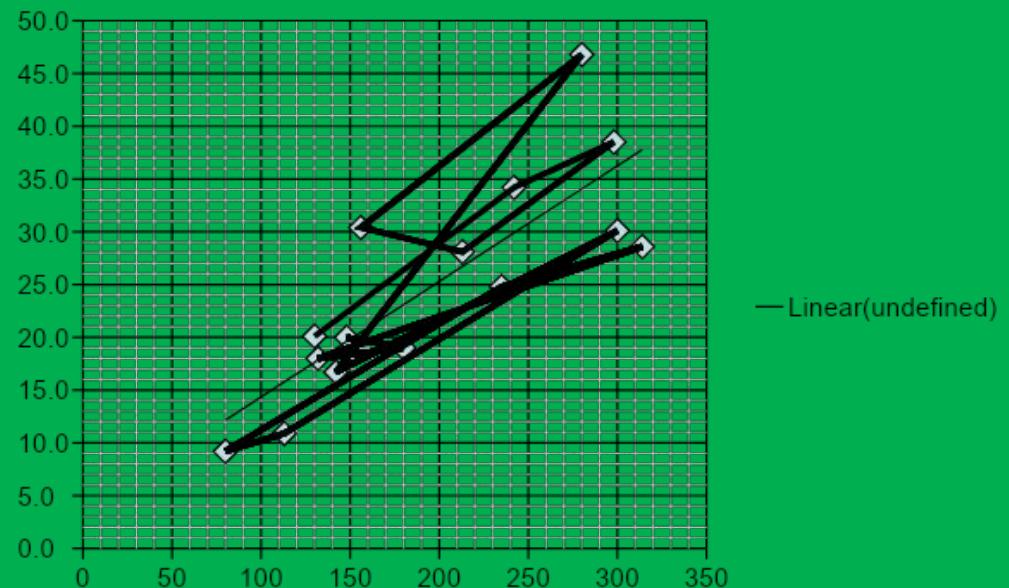
$$t_{расч} = |r| \sqrt{\frac{n-2}{1-r^2}}$$

где $n-2$ – число степеней свободы при заданном уровне значимости α (обычно $\alpha=0.05$). Расчетное значение $t_{расч}$ сравнивают с табличными значениями t -критерия $t_{кр}$, которые определяют по таблице Стьюдента при заданном уровне значимости α

№ магазина	товарооборо т, (млн. р.)	издержки обращения, (млн. р.)
1	148	20,4
2	180	19,2
3	132	18,9
4	314	28,6
5	235	24,8
6	80	9,2
7	113	10,9
8	300	30,1
9	142	16,7
10	280	46,8
11	156	30,4
12	213	28,1
13	298	38,5
14	242	34,2
15	130	20,1

Пример 1

Рассчитаем парный коэффициент корреляции между объёмом товарооборота и издержками обращения для магазинов на основе линейного уравнения



Линейная форма связи выражена уравнением

№ магазина	товарооборот y	издержки обращения x
1	148	20,4
2	180	19,2
3	132	18,9
4	314	28,6
5	235	24,8
6	80	9,2
7	113	10,9
8	300	30,1
9	142	16,7
10	280	46,8
11	156	30,4
12	213	28,1
13	298	38,5
14	242	34,2
15	130	20,1

$$y_x = a_0 + a_1 x,$$

Введем обозначение

Далее необходимо рассчитать параметры
уравнения для чего вводим дополнительные
колонки промежуточных расчетов

n	y	x					
1	148	20,4					
2	180	19,2					
				y^2	x^2	$x*y$	y_x
4	314	28,6	1	21904	416,16	3019,2	
5	235	24,8	2	32400	368,64	3456	
6	80	9,2	3	17424	357,21	2494,8	
7	113	10,9	4	98596	817,96	8980,4	
8	300	30,1	5	55225	615,04	5828	
9	142	16,7	6	6400	84,64	736	
10	280	46,8	7	12769	118,81	1231,7	
11	156	30,4	8	90000	906,01	9030	
12	213	28,1	9	20164	278,89	2371,4	
13	298	38,5	10	78400	2190,24	13104	
14	242	34,2	11	24336	924,16	4742,4	
15	130	20,1	12	45369	789,61	5985,3	
сумма	2963	375	13	88804	1482,25	11473	
			14	58564	1169,64	8276,4	
			15	16900	404,01	2613	
				667255	10866,26	83127,6	

$$a_0 = \frac{\sum y \sum x^2 - \sum xy \sum x}{n \sum x^2 - \sum x \sum x}$$

$$a_0 = \frac{2963 * 10866,26 - 83127,6 * 375}{15 * 10866,26 - 375 * 375}$$

= 45 тыс. руб.

$$a_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \sum x \sum x}$$

$$a_1 = \frac{15 * 83127,6 - 2963 * 375}{15 * 10866,26 - 375 * 375}$$

= 6,1 тыс. руб.

Далее рассчитываем теоретические уровни ряда подставив в уравнение рассчитанные данные

$$y_x = a_0 + a_1 x,$$

$$y_{20} = 45 + 6,1 * 20 = 167 \text{ тыс. руб.}$$

$$y_{19} = 45 + 6,1 * 19 = 160,9 \text{ тыс. руб.}$$

Делаем проверку расчетов по сумме теоретических уровней

$$r = \frac{\left(\sum xy - \frac{\sum x \sum y}{n} \right)}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n} \right] \left[\sum y^2 - \frac{(\sum y)^2}{n} \right]}} = 0,817$$

Связь между товарооборотом прямая, сильная. При увеличении издержек обращения на 1 тыс. руб. товарооборот возрастает на 6,1 тыс. руб.

Непараметрические методы оценки корреляции связи

Для измерения тесноты связи согласованного варьирования атрибутивных варьирующих признаков применяются разные показатели

1) Наиболее общим из них является

коэффициент взаимной сопряженности Чупрова А.А.

Он применяется для измерения связи между варьированием двух атрибутивных признаков, когда это варьирование образует несколько (три и более) групп

$$\varphi^2 = \frac{\sum \sum f_i^2 / \sum f_i}{\sum f_i} - 1$$

$$K_{\text{ч}} = \sqrt{\frac{\varphi^2}{\sqrt{(m_1 - 1)(m_2 - 1)}}}$$

m – число групп по каждому признаку.

Он изменяется от 0 до 1, но уже при значении 0,3 можно говорить о наличии тесной связи между вариацией изучаемых признаков.

Если вариация обоих альтернативных признаков ограничена двумя группами, то коэффициент взаимной сопряженности может быть исчислен значительно проще – через

коэффициент ассоциации Д. Юла и коэффициент контингенции К. Пирсона (K_k).

Для этого исходные данные сводятся в комбинационную четырехклеточную таблицу (таблица четырех полей).

Признак (работа)	Да (муж.)	Нет (жен.)
Да	A	B
Нет	C	D

$$K_a = \frac{ad - bc}{ad + bc},$$

$$K_k = \frac{ad - bc}{\sqrt{(a + b)(b + d)(a + c)(c + d)}}$$

где: a, b, c, d – числа таблицы.

Значения коэффициентов лежат в интервале $-1 < K_k < 1$, чем ближе к 1,-1, тем сильнее связь. (если показатель отсутствует, то заменяют его единицей для K_k)

Для определения **тесноты связи** как между количественными, так и между качественными признаками (если их можно проранжировать или упорядочить) можно использовать

Коэффициент Фохнера и коэффициент ранговой корреляции Спирмена

$$K_{\Phi} = \frac{C - H}{C + H}$$

С - число совпадений знаков значений признака от средних арифметических, Н – число несовпадений.

От -1 до +1. Если +1 – имеется полностью согласованная прямая изменчивость, если 0 – изменчивость полностью несогласуется, если -1 - имеется полная обратная согласованная изменчивость.

$$P = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

Знач. Р	0.1-0.3	0.3-0.5	0.5-0.7	0.7-0.9	0.9-1
Теснота связи	слабая	умеренная	заметная	высокая	Весьма высокая

[illegible]